MEGA

Molecular Evolutionary Genetics Analysis

Version 1.01

Sudhir Kumar, Koichiro Tamura, and Masatoshi Nei

Institute of Molecular Evolutionary Genetics
The Pennsylvania State University
University Park, PA 16802
USA

MEGA is distributed with a nominal fee to defray the cost of producing the user manual, the diskette(s), and the mailing and handling expenses (see order form). However, for anyone who is unable to pay the fee for some reason (e.g., lack of hard currencies in some countries), it will be provided free of charge after receiving a letter of explanation. MEGA will not be sent by electronic-mail because the accompanying manual cannot be included in this case. To obtain an order form, contact Joyce White or the authors at the address given below.

> **Although utmost care has been taken to ensure the correctness of the software, the software is provided "as is" without any warranty of any kind. In no event shall the authors and their employers be liable for any damages, including but not limited to special, consequential, or other damages. Authors specifically disclaim all other warranties, expressed or implied, including but not limited to the determination of suitability of this product for a specific purpose, use, or application.**

Suggested Citation:

Sudhir Kumar, Koichiro Tamura, and Masatoshi Nei. 1993. MEGA: Molecular Evolutionary Genetics Analysis, version 1.01. The Pennsylvania State University, University Park, PA 16802.

Distribution:

Institute of Molecular Evolutionary Genetics
328 Mueller Laboratory
The Pennsylvania State University
University Park, PA 16802, USA

Telephone: 814-863-7334 (not for technical assistance)
Fax: 814-863-7336
E-mail: imeg@psuvm.psu.edu
imeg@psuvm

MEGA version 1.0	F1	F2	F3	F4		F5	F6	F7	F8		F9	F10
	Help	Save File	Open File	Data Presentation		Browse	Next Window	Compute Distances	Construct Tree		Zoom Window	Main Menu
Alt	*Using Help*	*Save File As*	*Close File*	*Close Data*	*Alt*		*Previous Window*	*Std Error Test*	*Bootstrap Test*	*Alt*	*Re-size Window*	

Make a photocopy of this page and cut out the template for your keyboard.

Preface

Currently, many computer programs are available for estimating evolutionary distances and reconstructing phylogenetic trees from molecular data. However, most of them are written for specific methods and cannot be interconnected easily because of their inflexible input and output file formats. MEGA presents an interactive, user-friendly platform for estimating evolutionary distances, reconstructing phylogenetic trees, and computing basic statistical quantities that are of evolutionary interest. MEGA has been developed specifically for use on IBM and IBM-compatible personal computers.

MEGA is designed to facilitate extensive sequence data analysis from an evolutionary perspective using a single program package. At the same time, the overlap between the methods implemented in MEGA and those in other existing evolutionary analysis programs has been consciously avoided. This is reflected in the exclusion of the maximum likelihood method (PHYLIP) and in the absence of extensive options for the maximum parsimony method (PAUP and MacClade). Limitations on the memory size and relatively slower speeds of desktop computers (and the presence of many commercial and non-commercial programs) prompted the decision not to include sequence alignment methods in MEGA.

In this manual, chapter 1 (*Getting Started*) explains the procedures for installing and running MEGA and provides information on obtaining technical support. In the following chapter (*Input Data and Formats*), various data and input file formats are discussed. This chapter also elaborates on in-memory data editing features available in MEGA.

In chapter 3, a brief explanation of various statistical quantities that are useful for studying the evolutionary change of DNA and amino acid sequences is presented. Chapter 4 (*Distance Estimation*) describes most of the important statistical methods currently used for estimating evolutionary distances. This chapter also explains the handling of alignment gaps and missing data in the computation of evolutionary distances. Chapter 5 (*Phylogenetic Inference*) presents descriptions of different tree-building methods and related issues.

Basic features of interactive user-interface, sequence data presentation, phylogenetic-tree editing, context-sensitive help, and text-file editing and browsing are discussed in chapter 6 (*User-interface*). Chapter 7 (*Walk Through MEGA*) presents a tutorial on using MEGA for data analysis. Chapter 8 (*Command Reference*) explains the use of menus and commands present in the user-interface. Descriptions of cryptic errors and some possible remedies are provided in chapter 9 (*Error Messages*).

There are five Appendices included in this manual. Appendix A provides a list of computational and editing functions available in MEGA, whereas Appendix B gives a list of common questions concerning the use of MEGA and their answers. Appendices C, D, and E provide other information useful to the users of MEGA.

The MEGA project was initiated by the suggestion of M.N. However, the entire set of computer programs in MEGA was written by S.K. and K.T. S.K. is responsible for designing the layout of the programs and writing most parts of the programs, whereas K.T. is responsible for developing the tree-editing algorithm and major portions of the branch-and-bound and heuristic search algorithms for maximum parsimony. M.N. is responsible for choosing the statistical methods that are included in MEGA. He is also responsible for writing chapters 4 and 5 in this manual, though the algorithms for the maximum parsimony method in chapter 5 were developed by K.T. and S.K. The other chapters were written by S.K. and edited by M.N.

Many friends and colleagues provided encouragement and assistance in the development of MEGA and during the preparation of this manual. Test versions of MEGA were used in the classroom teaching by Masatoshi Nei and Austin Hughes at the Pennsylvania State University and by Marcy Uyenoyama at Duke University. These provided excellent opportunities to test the programs, and the students drew our attention to many inconsistencies in the user-interface. Initial drafts of certain chapters of this manual were critically read by Carla Hass, Tatsuya Ota, Naoko Takezaki, Marcy Uyenoyama, and Thomas Whittam. We also appreciate comments and help from Kyung-Eui Han, Blair Hedges, Thomas Lopez, Paul Mosquin, Arun Roychoudhury, Claudia Russo, Andrey Rzhetsky, Steve Schaeffer, and Tanya Sitnikova. We sincerely acknowledge everyone else who assisted in some form during this endeavor and whose name we fail to mention. The development of MEGA was partially supported by research grants from the National Institutes of Health and the National Science Foundation to M.N.

August, 1993

Sudhir Kumar
Koichiro Tamura
Masatoshi Nei

Contents

1

Getting Started

This chapter presents information on the hardware and software requirements for MEGA, installation instructions, and availability of technical support and future updates.

1.1 Hardware and Software

MEGA, version 1.0, is written and compiled in the Borland C++ and Applications Framework, version 3.1. It runs on all IBM and IBM-compatible personal computers, including PCs, XTs, ATs, PS/2s, laptops, and notebooks with most color and monochrome monitors. This program requires 640KB RAM memory and DOS (operating system) version 3.3 or later. MEGA can also be run on OS/2 and Microsoft Windows using DOS application capabilities. MEGA requires a hard disk, but extended and expanded memories, graphics adapters, and math co-processors are not necessary. However, the availability of a math-chip will enhance the speed and performance of MEGA.

The user-interface in MEGA responds to the keyboard as well as to the mouse. Although a mouse is not essential, one of the following types of mouse is recommended for MEGA:

1. Microsoft mouse version 6.1 or later or any true compatible mouse.
2. Logitech mouse version 3.4 or later.
3. Mouse systems' PC mouse version 6.22 or later.
4. IMSI mouse version 6.11 or later.

1.2 Installing MEGA

An automatic installation program, INSTALL, is provided with MEGA. You **MUST** install MEGA from the master diskette(s) distributed by the authors. Failure to do so may cause unexpected results. **DO NOT** simply copy the files from the MEGA master diskette(s) to the computer.

TO INSTALL MEGA:

1. Insert MEGA disk #1 into an external drive (example A:).
2. Type **A:** and press **Enter**.
3. Type **INSTALL** and press **Enter**.
4. Follow the instructions on the screen, if any.

Installation of MEGA will automatically create a **C:\MEGA** directory and the program files will be installed in this directory. MEGA **MUST NOT** be installed or moved to other drives and/or other directories.

1.3 Running MEGA

To run MEGA, go to **C:\MEGA** directory; type MEGA; and press **Enter**. To use MEGA from any other directory, add **C:\MEGA** to the **PATH** command in the **AUTOEXEC.BAT** file of the computer. Please consult your DOS manuals about the modification of the **PATH** command.

1.4 README File

The README file contains current information, including changes not listed in this manual. You should take a careful look at this file if it is present in the **C:\MEGA** directory.

1.5 Technical Support and Updates

Only registered users may request technical support, including information on programming errors and future improvements of MEGA. If you do not have a registration number in your name (as indicated on the original diskette(s) sent by the authors), you must register by writing to the authors.

If MEGA does not run properly, please refer to the manual **before contacting the authors.** Re-install the MEGA program from the master diskettes carefully following the instructions given in section **1.2**. If the problem persists, you **MUST** send the following information for technical assistance to the address given on the inside page of the front cover by a letter, e-mail, or fax. No telephone enquiry will be accepted.

1. Your name, address, telephone number, and e-mail address, if any.
2. MEGA version number, and your registration number.
3. Model numbers of the computer, monitor, printer, and mouse, if available.
4. Operating system and version number.
5. Copy of the input data file.
6. Exact sequence of events that led to the problem.

1.6 For Classroom Teaching

Because MEGA includes many statistical methods for the study of molecular evolution and because it has an interactive user-interface, it is suitable for classroom teaching. If you are interested in using MEGA for your classroom, please contact the authors to make suitable arrangements.

1.7 Source Code Availability

This manual provides information on most features of MEGA, and we discourage requests for the source code of the complete program or of any part. However, any suggestion that may improve the analysis of molecular evolutionary data will be welcomed.

2

Input Data and Formats

Input file formats for different kinds of data are discussed in this chapter. In addition, the use of in-memory data editing options is explained. Note that there is no limit on the amount of molecular sequence or distance matrix data that can be analyzed in MEGA; the size of data set is constrained only by the computer memory available.

2.1 MEGA Format

Either sequence data or distance data can be entered in MEGA as ASCII-text files. These data must be organized in a format specific to MEGA. These input file formats are consistent and flexible, and they include options for writing extensive comments in the data file.

2.1.1 Key Words

Every data file must contain the key words **#MEGA** and **TITLE**. These key words can be written in any combination of lower- and upper-case letters.

#MEGA This key word indicates that the data file is prepared for analysis using MEGA. It must be present on the **very first line** in the data file.

TITLE The word **TITLE** must be written on the **second line**. It may be followed by some description of data on the same line. This description is written in all the output files containing results. If the specified description exceeds 128 characters in length, the additional characters are ignored.

After the MEGA format identifier (#MEGA) and the title (TITLE), the data should follow. Comments may be written on one or more lines right after the TITLE line and before the data (see examples in sections **2.2** and **2.3**).

2.1.2 OTU Labels

Distance matrices as well as sequence data may come from species, populations, or individuals. These evolutionary entities are designated as OTUs (Operational Taxonomic Units). Each OTU must have an identification tag, i.e., an OTU label. In the input files prepared for use in MEGA, these labels should be written according to the following conventions.

'#' Sign Every OTU label must be written on a new line, and a '#' sign must proceed the label. OTU labels cannot be longer than *40 characters*; extra characters are disregarded. OTU labels are not required to be unique, but identical labels may result in ambiguities.

Forbidden Characters The '#' sign, *blanks*, and *tabs* cannot be a part of an OTU label. For multiple word labels, an underscore can be used to represent a *blank space*. All underscores are converted into *blank spaces*, and subsequent displays of the OTU label show this change. For example, *E._coli* becomes *E. coli.*

2.2 Sequence Input Formats

The sequence data must consist of two or more sequences of **equal length.** All sequences must be aligned (MEGA does not include an alignment program) and should be arranged either in interleaved (block-wise) or in noninterleaved (continuous) format (see below).

Nucleotide or amino acid sequences should be written in IUPAC single-letter codes. In this system, A, T(U), C, and G represent the four different nucleotides, and all alphabets <u>except</u> B, J, O, U, X, and Z represent the twenty different amino acids (see Table 2.1). However, the use of N (and n) for ambiguous nucleotides and X (and x) for ambiguous amino acid residues must be avoided. Sequences can be written in any combination of upper- and lower-case letters. Special symbols for alignment gaps, missing data, and identical sites can also be included in the sequences.

Special Symbols *Blank spaces* and *Tabs* are frequently used to format data files, so they are simply ignored by MEGA. Unique ASCII characters, except alphabets and '*', can be used as special symbols for alignment gaps, missing-information sites, and identical sites. Frequently used symbols for identical sites, alignment gaps, and missing-information sites are '.', '-', and '?', respectively.

Table 2.1 IUPAC single-letter codes used in MEGA.

Symbols	Name	Remarks
DNA and RNA		
A	Adenine	Purine
G	Guanine	Purine
C	Cytosine	Pyrimidine
T	Thymine	Pyrimidine
U	Uracil	Pyrimidine
Amino Acids		
A	Alanine	Ala
C	Cysteine	Cys
D	Aspartic acid	Asp
E	Glutamic acid	Glu
F	Phenylalanine	Phe
G	Glycine	Gly
H	Histidine	His
I	Isoleucine	Ile
K	Lysine	Lys
L	Leucine	Leu
M	Methionine	Met
N	Asparagine	Asn
P	Proline	Pro
Q	Glutamine	Gln
R	Arginine	Arg
S	Serine	Ser
T	Threonine	Thr
V	Valine	Val
W	Tryptophan	Trp
Y	Tyrosine	Tyr

Noninterleaved Format In the noninterleaved format, the complete sequence for an OTU is written on one or more lines following its label as shown in the following example.

```
#mega
TITLE: Noninterleaved sequence data

#mouse       AATTTTTACCCCGGGGGG
             AGGGGGGACCCCGGGGGG
#human       AACCCTTACCCCGGGGGG
             AGGGGGGACCCCGGGGGG
#cat         AATTTTTACAAAGGGGGG
             AGGGGGGACCCCGGGGGG
```

In noninterleaved format there are alternate ways of writing the OTU label and the sequence:

```
  (a)  #mouse   AATTTTTACCCCGGGGGG
  (b)  #mouse
       AATTTTTACCCCGGGGGG
```

Interleaved Format

In contrast to the noninterleaved format, interleaved sequences are arranged in blocks consisting of homologous sites for all OTUs. The sequences for all the OTUs must be present in the same order in every block, and these sequences should be written on the consecutive lines in each of the blocks. Sequence blocks should be separated from each other by at least one blank line.

```
#mega
TITLE: Interleaved sequence data

#mouse       AATTTTTACCCCGGGGGG
#human       AACCCTTACCCCGGGGGG
#cat         AATTTTTACCCCGGGGGG

#mouse       AGGGGGGACCCCGG
#human       AGGGGGGACCCCGG
#cat         AGGGGGGACAAAGG
```

Comments

Comments can be placed after the **TITLE** line and before the data as well as within the sequences. Comments included inside a sequence must be contained within a pair of double quotation marks.

```
Format type →  #mega
Title       →  TITLE: 2 exons from gene XYZ
Comments    →     Authors: James R. and Ray S., 1987
                  Sequencing procedure: PCR

Sequences &
Comments    →  #cat           ATTCCCGGCCG"intron  10"ACCC
               #rat           ATTCCCGGGGG"intron of length 8" ACCC
               #rabbit        GTTCCCGGGAA"no introns" ACCC
```

2.3 Distance Input Formats

There are $m(m-1)/2$ pairwise distances for m OTUs. These distances can be arranged either in the lower-left or in the upper-right triangular matrix.

Following the key word **#MEGA** on the first line and the **TITLE** on the second line, all OTU labels should be written on consecutive lines. OTU labels should be prefixed with the '#' mark and should be written according to the conventions described in section **2.1.2**. This list should be separated from the following distance matrix by at least one blank line.

```
Format type→   #mega
      Title→   Title:   Upper-right triangular matrix
    Blank 1→

OTU names on   #one
consecutive    #two
      lines    #three
               #four
               #five
    Blank 2→

     one vs.   1.0   2.0   3.0   4.0
others, etc.         3.0   2.5   4.6
                           1.3   3.6
                                 4.2
end of file→
```

In this example, blank line 1 is optional, but blank line 2 is required. The two alternate distance matrix formats are:

Lower-left matrix:

$$\begin{array}{llll} d_{12} & & & \\ d_{13} & d_{23} & & \\ d_{14} & d_{24} & d_{34} & \\ d_{15} & d_{25} & d_{35} & d_{45} \end{array}$$

Upper-right matrix:

$$\begin{array}{llll} d_{12} & d_{13} & d_{14} & d_{15} \\ & d_{23} & d_{24} & d_{25} \\ & & d_{34} & d_{35} \\ & & & d_{45} \end{array}$$

Comments In data files containing distance matrices, comments can only be placed after the **TITLE** line and before the OTU labels.

2.4 Editing Sequence Data

Input sequence data consist of two or more aligned sequences of equal length. In MEGA, any subset of this sequence data can be selected for analysis using options available in the *Data* menu. *Select OTUs* and *Select Sites/Codons* commands are used to choose a desired subset of data. This subset is referred to as the current data, and it is maintained until it is modified.

Selecting Mode for Analysis The *Select Mode* command is used to select the protein-coding or noncoding mode for nucleotide sequences. The coding mode provides codon-by-codon and site-by-site analyses, whereas the non-coding mode provides only site-by-site analysis.

Selecting OTUs

By default, all OTUs are included in the current data. Some of these OTUs can be removed by using the *Data*|*Select OTUs* command. These OTUs will stay deleted until the *Select OTUs* command is used again.

Selecting Sites or Codons

Options for selecting domains as well as individual sites or codons are provided in MEGA. To start with, all the sites (codons) are included in the current data. With the *Domains* option, up to 10 nonoverlapping domains of sites (or codons) can be chosen. Individual sites (or codons) are chosen by using the *Individual* command.

The options for including alignment gaps and missing information sites and the choice of nucleotide positions in codons provide a second level of data editing. These options are prompted every time before the analysis begins (see section **4.5**), and they only affect the current analysis.

Choosing Sites in Codons

Any combination of first, second, and third nucleotide positions in the codons can be chosen if the nucleotide sequences are used in the protein coding mode.

Excluding Missing-Information Sites and Alignment Gaps

In distance computation, alignment gaps and missing-information sites can be treated in two different ways. One is to eliminate all these gap and missing-information sites from all the sequences. The other is to ignore only the gap and missing-information sites that are involved in a particular pairwise comparison. These options are usually prompted before distance calculation and tree reconstruction. Detailed discussions on this topic are presented in the chapters on *Distance Estimation* and *Phylogenetic Inference*.

2.5 Editing Distance Data

A set of OTUs can be selected in distance matrix data by using the *Select OTUs* command from the *Data* menu. The distance matrix is reduced automatically by removing rows and columns corresponding to the excluded OTUs.

3

Basic Sequence Statistics

In the study of molecular evolution it is often necessary to know some basic statistical quantities such as nucleotide frequencies, codon frequencies, and transition/transversion ratios. The statistical quantities that can be computed by MEGA are discussed in this chapter.

3.1 Nucleotide and Amino Acid Compositions

The relative frequencies of the four nucleotides (nucleotide composition) or of the twenty amino acid residues (amino acid composition) can be computed for a specific sequence or for all the sequences used.

Example 3.1 Nucleotide composition of HLA sequences.

```
-------- Nucleotide Composition -------
All values in per cent (%) except Totals

                   A    T    C    G     Total

     HLA-A2     20.8 15.2 29.8 34.2       822
     HLA-A3     20.4 14.7 30.2 34.7       822
     HLA-A11    20.6 14.1 30.5 34.8       822
     HLA-AW24   20.9 14.6 30.2 34.3       822
     HLA-AW68   20.7 14.8 30.2 34.3       822
     All        20.7 14.7 30.2 34.5      4110
```

For coding regions of DNA, three additional tables are presented for the nucleotide compositions at first, second, and third codon positions. From these tables the G+C content can easily be computed. The amino acid composition can also be presented in a similar tabular form.

3.2 Codon Usage

There are 64 (4^3) possible codons that code for 20 amino acids (and stop signals), so an amino acid may be encoded by several codons (e.g., serine is encoded by six codons in nuclear genes). It is therefore interesting to know the codon usage for each amino acid. In MEGA the numbers of the 64 codons used in a gene can be computed either for a specific sequence or for all sequences examined. Four different genetic codes are included; the "universal" code and the mammalian, *Drosophila,* and yeast mitochondrial genetic codes.

MEGA is also capable of computing Sharp *et al.*'s (1986) relative synonymous codon usage (RSCU). RSCU is the observed frequency of a codon divided by its expected frequency under the assumption of equal codon usage. That is,

$$RSCU_{ij}=\frac{X_{ij}}{\sum\limits_{j=1}^{n_i} X_{ij}/n_i}\ . \qquad (3.1)$$

Here, X_{ij} is the number of occurrences of the *j*th codon for the *i*th amino acid, and n_i is the number (from one to six) of alternative codons for the *i*th amino acid. This index is useful for knowing the codons that are used more often or less often than expected under the assumption of equal usage.

Example 3.2 Codon frequencies and RSCU values for HLA-A2.

```
--------- Codon Usage --------
Codon Usage Table for HLA-A2
Frequency of codons and relative synonymous codon usage (RSCU)

TTT (F)  0 (0.00)     ...     TGT (C)   0 (0.00)
TTC (F)  8 (2.00)     ...     TGC (C)   4 (2.00)
TTA (L)  0 (0.00)     ...     TGA (*)   0 (0.00)
TTG (L)  2 (0.71)     ...     TGG (W)  10 (1.00)
.
.
.
GTT (V)  0 (0.00)     ...     GGT (G)   3 (0.60)
GTC (V)  2 (0.50)     ...     GGC (G)   7 (1.40)
GTA (V)  0 (0.00)     ...     GGA (G)   2 (0.40)
GTG (V) 14 (3.50)     ...     GGG (G)   8 (1.60)

Total codons scored: 274
'*' indicates a stop codon.
RSCU is given in parentheses.
```

3.3 Nucleotide Pair Frequencies

When two nucleotide sequences are compared, the frequencies of 10 different types of nucleotide pairs can be computed. In MEGA these frequencies are tabulated in the following form.

Example 3.3 Nucleotide pair frequencies for alleles of the HLA-A locus.

```
------- Observed nucleotide pair frequencies -------

n:  total number of nucleotides compared
ns: number of transitional differences
nv: number of transversional differences
nd: ns+nv (total number of nucleotide differences)

                     Tran-    Trans-           Identical
                     sition   version            pair

                     AG TC   AT AC TG CG    AA  TT  CC  GG   ns/nv   nd   n
HLA-A2 vs. HLA-A3    11  5    2  2  5  8   162 117 239 271    0.94   33  822
HLA-A2 vs. HLA-A11   11  8    3  4  4 10   161 113 237 271    0.90   40  822
HLA-A2 vs. HLA-AW24  13 11    3  1  5 15   163 113 233 265    1.00   48  822
HLA-A2 vs. HLA-AW68   3  2    2  2  5 11   167 119 239 272    0.25   25  822
```

3.4 Alignment Gap Frequencies

The observed numbers of alignment gaps of different lengths (sites) are useful for studying the distribution of insertions/deletions and for deciding whether all sites containing gaps should be deleted (see section **4.5**). In MEGA, the numbers of gaps of length 1 to 10 can be computed either for each sequence or for all sequences. The numbers of gaps longer than 10 sites are pooled together with the number of gaps of length 10.

Example 3.4 Alignment gap frequencies for HLA sequences.

```
-------- Alignment Gap Frequencies ------

All entries in the table are the observed number of occurrences

           1     2     3     ...   ≥10   Total

HLA-A2     0     0     0     ...   1     1
HLA-A3     0     0     0     ...   1     1
HLA-A11    0     0     0     ...   1     1
HLA-AW24   0     0     0     ...   1     1
HLA-AW68   0     0     0     ...   1     1
All        0     0     0     ...   5     5
```

3.5 Variable Regions of Sequences

It is well known that some regions of DNA or amino acid sequences are more variable than others. For example, the control region of mammalian mitochondrial DNA has two hypervariable segments (Kocher and Wilson 1991). One way of detecting such variable regions is to examine the number of variable sites in different segments of the DNA. In MEGA, the numbers of variable sites in overlapping and nonoverlapping segments of equal size can be computed for any segment size (window size). In the output, the numbers of variable sites in overlapping (sliding window) or nonoverlapping segments of a specified size are given along with a histogram.

Example 3.5 Nonoverlapping windows for HLA-A sequence data.

```
--------- Variability --------
Total number of variable sites: 71
Numbers of variable sites in nonoverlapping segments of size 100

Location

  1-100 |  6 | ▪▪▪▪▪▪
101-200 |  5 | ▪▪▪▪▪
201-300 | 19 | ▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪
301-400 | 10 | ▪▪▪▪▪▪▪▪▪▪
401-500 |  7 | ▪▪▪▪▪▪▪
501-600 | 13 | ▪▪▪▪▪▪▪▪▪▪▪▪▪
601-700 |  5 | ▪▪▪▪▪
701-800 |  5 | ▪▪▪▪▪
801-    |  1 | ▪
```

4

Distance Estimation

The evolutionary distance between a pair of sequences is usually measured by the number of nucleotide or amino acid substitutions between them. Evolutionary distances are fundamental for the study of molecular evolution and are useful for phylogenetic reconstruction and estimation of divergence times. In MEGA, most of the widely used methods for distance estimation for nucleotide and amino acid sequences are included. In the following, they are presented in three sections: *nucleotide substitutions*, *synonymous-nonsynonymous substitutions*, and *amino acid substitutions*. For advice in the use of these methods, see *Guidelines for Choosing Distance Measures* in section **4.4**. The treatment of alignment gaps and missing-information sites in distance computation is explained in section **4.5**.

4.1 Nucleotide Substitutions

The evolutionary distances that are computed from DNA sequence data are primarily estimates of the number of nucleotide substitutions per site (d) between two sequences. There are many methods for estimating evolutionary distances, depending on the pattern of nucleotide substitutions (see Nei 1987, Gojobori *et al.* 1990, Saccone *et al.* 1990, and others). Here we have included only methods that are relatively simple and frequently used by molecular evolutionists. Two methods, i.e., the Tamura and Tamura-Nei methods, are new and their utility has not been well tested, but they are included here because they seem to be useful for analyzing mitochondrial DNA data, which are now often used for phylogenetic inference. In the following we first present the simplest method and then discuss gradually more complicated ones.

p-distance

This distance is merely the proportion (p) of nucleotide sites at which the two sequences compared are different. This is obtained by dividing the number of nucleotide differences (n_d) by the total number of nucleotides compared (n). Thus,

$$p = n_d/n. \tag{4.1}$$

The variance of p is given by

$$V(p) = [p(1 - p)]/n. \tag{4.2}$$

The p-distance is approximately equal to the number of nucleotide substitutions per site (d) only when it is small, say $p < 0.1$. However, the computation of this distance is simple, and for constructing phylogenetic trees it gives essentially the same results as the more complicated distance measures mentioned below, as long as all pairwise distances are small. Actually, when the rate of nucleotide substitution is the same for all evolutionary lineages, the p-distance gives the correct topology slightly more often than the Jukes-Cantor and Kimura distances mentioned below, because it has a smaller variance (Saitou and Nei 1987, Saitou and Imanishi 1989, Schöniger and von Haeseler 1993, Tajima and Takezaki, 1994). Of course, for estimating the divergence times of two sequences, this is not a good measure.

Under certain circumstances, one may want to compute the proportion of sites with transitional and transversional nucleotide differences. In MEGA, the proportions of transitional differences (P) and transversional differences (Q) are computed by

$$P = n_s/n \text{ and } Q = n_v/n, \tag{4.3}$$

respectively, where n_s and n_v are the numbers of transitional and transversional differences between the two sequences, with $n_s + n_v = n_d$. The variances of P and Q are computed by equations analogous to (4.2). In addition, the ratio of transitional to transversional differences (R_d) and its variance are given by

$$R_d = P/Q, \tag{4.4}$$

$$V(R_d) = [c_1^2 P + c_2^2 Q - (c_1 P + c_2 Q)^2]/n, \tag{4.5}$$

where $c_1 = 1/Q$ and $c_2 = - P/Q^2$.

Jukes-Cantor distance

This method (Jukes and Cantor 1969) was developed under the assumption that the rate of nucleotide substitution is the same for all pairs of the four nucleotides A, T, C, and G (see Table 4.1), and it gives a maximum likelihood estimate of the number of nucleotide substitutions (d) between two sequences. It is given by

$$\hat{d} = - {}^3/_4 \log_e (1 - {}^4/_3\, p), \tag{4.6}$$

where p is computed by using equation (4.1).

Table 4.1 Models of nucleotide substitution.

Nucleotides		Mutant		
Original	A	T	C	G
A. Jukes-Cantor model				
A	-	λ	λ	λ
T	λ	-	λ	λ
C	λ	λ	-	λ
G	λ	λ	λ	-
	λ is the rate of substitution.			
B. Tajima-Nei model				
A	-	β	γ	δ
T	α	-	γ	δ
C	α	β	-	δ
G	α	β	γ	-
	α, β, γ, and δ are the rates of substitution.			
C. Kimura 2-parameter model				
A	-	β	β	α
T	β	-	α	β
C	β	α	-	β
G	α	β	β	-
	α and β are the rates of transitional and transversional substitution, respectively.			
D. Tamura model				
A	-	$(1 - \Theta)\beta$	$\Theta\beta$	$\Theta\alpha$
T	$(1 - \Theta)\beta$	-	$\Theta\alpha$	$\Theta\beta$
C	$(1 - \Theta)\beta$	$(1 - \Theta)\alpha$	-	$\Theta\beta$
G	$(1 - \Theta)\alpha$	$(1 - \Theta)\beta$	$\Theta\beta$	-
	α and β are the rates of transitional and transversional substitution, respectively, and Θ is the G+C content.			
E. Hasegawa *et al.* model				
A	-	$g_T\beta$	$g_C\beta$	$g_G\alpha$
T	$g_A\beta$	-	$g_C\alpha$	$g_G\beta$
C	$g_A\beta$	$g_T\alpha$	-	$g_G\beta$
G	$g_A\alpha$	$g_T\beta$	$g_C\beta$	-
	α and β are the rates of transitional and transversional substitution, respectively, and g_i denotes the nucleotide frequencies (i=A,T,C,G).			
F. Tamura-Nei model				
A	-	$g_T\beta$	$g_C\beta$	$g_G\alpha_1$
T	$g_A\beta$	-	$g_C\alpha_2$	$g_G\beta$
C	$g_A\beta$	$g_T\alpha_2$	-	$g_G\beta$
G	$g_A\alpha_1$	$g_T\beta$	$g_C\beta$	-
	α_1 and α_2 are the rates of transitional substitution between purines and between pyrimidines, respectively; β is the rate of transversional substitution; and g_i denotes the nucleotide frequencies (i=A,T,C,G).			

The variance of this estimate is given by

$$V(\hat{d}) = p(1 - p)/[(1 - 4/3\,p)^2 n] \tag{4.7}$$

(Kimura and Ohta 1972).

The Jukes-Cantor distance can be computed if $p<0.75$; otherwise it is not applicable because the argument of the logarithm becomes negative. This distance gives a good estimate of the number of nucleotide substitutions if (1) the frequency of each nucleotide is close to 0.25, (2) there is no transition/transversion bias (i.e., the transition/transversion ratio is nearly equal to 0.5), and (3) d is not very large (say $d<1.0$). However, when the number of nucleotides examined is small, say $n<100$, the Jukes-Cantor distance tends to give overestimates of the true number of nucleotide substitutions (Tajima 1993).

Tajima-Nei distance

In real data, nucleotide frequencies often deviate substantially from 0.25. In this case the Tajima-Nei distance (Tajima and Nei 1984) gives a better estimate of the number of nucleotide substitutions than the Jukes-Cantor distance. This estimator is based on the equal-input model of Nei and Tajima (1981) (see Table 4.1). The estimator ($\hat{d}$) and its variance [V($\hat{d}$)] are given by the following equations.

$$\hat{d} = - b \log_e (1 - p/b), \tag{4.8}$$

$$V(\hat{d}) = p(1 - p)/[(1 - p/b)^2 n], \tag{4.9}$$

where

$$b = \frac{1}{2}(1 - \sum_{i=1}^{4} g_i^2 + p^2/c)\,,$$

$$c = \sum_{i=1}^{3}\sum_{j=i+1}^{4} \frac{x_{ij}}{2g_i g_j}\,.$$

Here, g_i and g_j are the frequencies of the ith and jth nucleotides, respectively (i,j=A,T,C,G), and x_{ij} is the relative frequency of nucleotide pair i and j.

Computer simulations have shown that this estimate is quite robust and is applicable to a wide variety of cases unless the number of nucleotide substitutions is very large, say more than 1.0 per site.

Kimura 2-parameter distance

In actual sequence data the rate of transitional nucleotide substitution is often higher than that of transversional substitution. This is particularly so for animal mitochondrial DNA (Brown *et al.* 1982). In this case, the Jukes-Cantor distance is expected to give an underestimate of d unless d is quite small, say $d<0.1$. A maximum likelihood estimate of d for this case is given by Kimura's (1980) 2-parameter method (Table 4.1). This estimate and its variance are given by

$$\hat{d} = -\tfrac{1}{2}\log_e(1 - 2P - Q) - \tfrac{1}{4}\log_e(1 - 2Q), \quad (4.10)$$

$$V(\hat{d}) = [c_1^2P + c_3^2Q - (c_1P + c_3Q)^2]/n, \quad (4.11)$$

where $c_1 = 1/(1 - 2P - Q)$, $c_2 = 1/(1 - 2Q)$, and $c_3 = \tfrac{1}{2}(c_1 + c_2)$.

With Kimura's model, it is possible to compute the numbers of transitional ($\hat{s}$) and transversional ($\hat{v}$) nucleotide substitutions per site and their variances.

<u>Transitional substitutions</u>:

$$\hat{s} = -\tfrac{1}{2}\log_e(1 - 2P - Q) + \tfrac{1}{4}\log_e(1 - 2Q), \quad (4.12)$$

$$V(\hat{s}) = [c_1^2P + c_4^2Q - (c_1P + c_4Q)^2]/n, \quad (4.13)$$

where $c_4 = \tfrac{1}{2}(c_1 - c_2)$.

<u>Transversional substitutions</u>:

$$\hat{v} = -\tfrac{1}{2}\log_e(1 - 2Q), \quad (4.14)$$

$$V(\hat{v}) = c_2^2Q(1 - Q)/n. \quad (4.15)$$

<u>Transition/transversion ratio</u> ($R = \hat{s}/\hat{v}$):

The ratio (R) of the number of transitional substitutions ($\hat{s}$) to that of transversional substitutions ($\hat{v}$) is called the transition/transversion ratio. Note that R is different from R_d defined earlier. In the present case, R and its variance [$V(R)$] are given by

$$R = \log_e(1 - 2P - Q)/\log_e(1 - 2Q) - \tfrac{1}{2}, \quad (4.16)$$

$$V(R) = [c_5^2P + c_6^2Q - (c_5P + c_6Q)^2]/n, \quad (4.17)$$

where $c_5 = -2c_1/\log_e(1 - 2Q)$ and $c_6 = \tfrac{1}{2}[c_5 + 4c_2\log_e(1 - 2P - Q)/(\log_e(1 - 2Q))^2]$.

Tamura distance

Kimura's 2-parameter distance is based on the assumption that the nucleotide frequencies are all equal to 0.25 throughout the evolutionary process. In practice, however, this assumption rarely holds. In particular, the G+C content of *Drosophila* mitochondrial DNA is much lower than 0.5. Tamura (1992) developed a maximum likelihood estimator of d, which is suitable for this case (Table 4.1). The estimator and its variance are given by

$$\hat{d} = -2\Theta(1-\Theta)\log_e(1 - P/(2\Theta(1-\Theta)) - Q) - \tfrac{1}{2}(1 - 2\Theta(1-\Theta))\log_e(1 - 2Q), \tag{4.18}$$

$$V(\hat{d}) = [c_1^2P + c_3^2Q - (c_1P + c_3Q)^2]/n, \tag{4.19}$$

where $c_1 = 1/(1 - P/(2\Theta(1-\Theta)) - Q)$, $c_2 = 1/(1 - 2Q)$, $c_3 = 2\Theta(1-\Theta)(c_1 - c_2) + c_2$, and Θ is the G+C content.

The estimates of the numbers of transitional ($\hat{S}$) and transversional ($\hat{V}$) substitutions per site are obtained by the following equations.

<u>Transitional substitutions</u>:

$$\hat{S} = -2\Theta(1-\Theta)\log_e(1 - P/(2\Theta(1-\Theta)) - Q) + \Theta(1-\Theta)\log_e(1 - 2Q), \tag{4.20}$$

$$V(\hat{S}) = [c_1^2P + c_4^2Q - (c_1P + c_4Q)^2]/n, \tag{4.21}$$

where $c_4 = 2\Theta(1-\Theta)(c_1 - c_2)$.

<u>Transversional substitutions</u>:

$$\hat{V} = -\tfrac{1}{2}\log_e(1 - 2Q), \tag{4.22}$$

$$V(\hat{V}) = c_2^2Q(1 - Q)/n. \tag{4.23}$$

<u>Transition/transversion ratio</u> ($R = \hat{S}/\hat{V}$):

$$R = 4\Theta(1-\Theta)\log_e(1 - P/(2\Theta(1-\Theta)) - Q)/\log_e(1 - 2Q) - 2\Theta(1-\Theta), \tag{4.24}$$

$$V(R) = [c_5^2P + c_6^2Q - (c_5P + c_6Q)^2]/n, \tag{4.25}$$

where $c_5 = -2c_1/\log_e(1 - 2Q)$ and $c_6 = 2\Theta(1-\Theta)[c_5 + 4c_2\log_e(1 - P/(2\Theta(1-\Theta)) - Q)/(\log_e(1 - 2Q))^2]$.

In MEGA, the average G+C content for the pair of sequences compared is used for Θ. Therefore, different pairwise comparisons may have different values of Θ. There are other ways of computing Θ, but the distance estimates obtained are usually very similar.

Tamura-Nei distance

One of the useful mathematical models for analyzing mitochondrial DNA is that of Hasegawa *et al.*'s (1985). This model (Table 4.1) has been used for phylogenetic inference by the maximum likelihood method. However, no analytical formula for estimating d has been derived for this model.

Tamura and Nei (1993) noted that model F in Table 4.1 is more realistic than model E. In model E, $\alpha_1 = \alpha_2$ is assumed, but actual data indicates that the rates of transitional substitution between purines (A and G) and between pyrimidines (T and C) are often different. For model F, Tamura and Nei (1993) derived the following formula for estimating d.

$$\begin{aligned}\hat{d} = &- \frac{2g_A g_G}{g_R}\log_e(1 - \frac{g_R}{2g_A g_G}P_1 - \frac{1}{2g_R}Q)\\ &- \frac{2g_T g_C}{g_Y}\log_e(1 - \frac{g_Y}{2g_T g_C}P_2 - \frac{1}{2g_Y}Q)\\ &- 2(g_R g_Y - \frac{g_A g_G g_Y}{g_R} - \frac{g_T g_C g_R}{g_Y})\log_e(1 - \frac{1}{2g_R g_Y}Q),\end{aligned} \qquad (4.26)$$

where P_1 and P_2 are the proportions of transitional differences between A and G and between T and C, respectively, and Q is the proportion of transversional differences.

They also derived the variance of $\hat{d}$, but we are not going to present it here because it is somewhat complicated. The computation of the variance is included in the computer program.

The estimates of the numbers of transitional ($\hat{S}$) and transversional ($\hat{V}$) substitutions per site are obtained by the following equations.

Transitional substitutions:

$$
\begin{aligned}
\hat{S} = & - \frac{2g_A g_G}{g_R} \log_e\left(1 - \frac{g_R}{2g_A g_G} P_1 - \frac{1}{2g_R} Q\right) \\
& - \frac{2g_T g_C}{g_Y} \log_e\left(1 - \frac{g_Y}{2g_T g_C} P_2 - \frac{1}{2g_Y} Q\right) \\
& + 2\left(\frac{g_A g_G g_Y}{g_R} + \frac{g_T g_C g_R}{g_Y}\right) \log_e\left(1 - \frac{1}{2g_R g_Y} Q\right) .
\end{aligned}
\tag{4.27}
$$

Transversional substitutions:

$$
\hat{V} = - 2g_R g_Y \log_e\left(1 - \frac{Q}{2g_R g_Y}\right) .
\tag{4.28}
$$

The transition/transversion ratio is given by $R = \hat{S}/\hat{V}$. The computation of this ratio and its variance as well as the variances of $\hat{S}$ and $\hat{V}$ is included in MEGA.

Gamma distances

In the distance measures discussed so far, the rate of nucleotide substitution is assumed to be the same for all nucleotide sites. In actual data this assumption rarely holds, and the rate varies from site to site. Statistical analyses of the distribution of the number of substitutions at different sites have suggested that the rate varies approximately according to the gamma distribution (Uzzell and Corbin 1971, Kocher and Wilson 1991, Tamura and Nei 1993, Wakeley 1993). The gamma distribution can be specified by the parameter a, which is the inverse of the coefficient of variation of the substitution rate (λ). The smaller the parameter a, the higher the extent of variation in λ. In one hypervariable segment of the control region of mitochondrial DNA, a has been estimated to be 0.47 (Wakeley 1993), whereas Uzzell and Corbin (1971) obtained $a=2$ for amino acid sequence data for cytochrome c.

In the following gamma distances the rate of nucleotide substitution is assumed to follow the gamma distribution specified by parameter a. They are due to Jin and Nei (1990) and Tamura and Nei (1993). The default option of MEGA assumes $a=1.0$ for nucleotide substitution, except for the gamma distance for the Tamura-Nei model. When a is small ($a<1$) and the number of nucleotides examined is small ($n \leq 100$), the following formula tends to give underestimates of the true number of nucleotide substitutions (Rzhetsky and Nei 1994). It is therefore important to use a large number of nucleotides.

Gamma distance for the Jukes-Cantor model:

When the rate of substitution in the Jukes-Cantor model varies with the gamma distribution, the gamma distance and its variance are given by

$$\hat{d} = {}^{3}/_{4}\, a[(1 - {}^{4}/_{3}\, p)^{-1/a} - 1], \tag{4.29}$$

$$V(\hat{d}) = p(1 - p)[(1 - {}^{4}/_{3}\, p)^{-2(1/a + 1)}]/n. \tag{4.30}$$

Gamma distance for the Kimura 2-parameter model:

In this case the gamma distance and its variance are given by

$$\hat{d} = (a/2)[(1 - 2P - Q)^{-1/a} + \tfrac{1}{2}(1 - 2Q)^{-1/a} - {}^{3}/_{2}], \tag{4.31}$$

$$V(\hat{d}) = (c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2)/n, \tag{4.32}$$

where $c_1 = (1 - 2P - Q)^{-(1/a + 1)}$, $c_2 = (1 - 2Q)^{-(1/a + 1)}$, $c_3 = \tfrac{1}{2}(c_1 + c_2)$, and P and Q are the same as those of the Kimura 2-parameter model.

Transitional substitutions:

$$\hat{s} = (a/2)[(1- 2P - Q)^{-1/a} - \tfrac{1}{2}(1 - 2Q)^{-1/a} -\tfrac{1}{2}], \tag{4.33}$$

$$V(\hat{s}) = (c_1^2 P + c_4^2 Q - (c_1 P + c_4 Q)^2)/n, \tag{4.34}$$

where $c_4 = \tfrac{1}{2}(c_1 - c_2)$.

Transversional substitutions:

$$\hat{v} = (a/2)[(1 - 2Q)^{-1/a} -1], \tag{4.35}$$

$$V(\hat{v}) = c_2^2 Q(1 - Q)/n. \tag{4.36}$$

Transition/Transversion ratio ($R = \hat{s}/\hat{v}$):

$$R = [(1- 2P - Q)^{-1/a} - \tfrac{1}{2}(1 - 2Q)^{-1/a} - \tfrac{1}{2}]/[(1 - 2Q)^{-1/a} - 1]. \tag{4.37}$$

The formula for the variance of R is rather complicated and is not presented here, but it is computed in MEGA.

Gamma distance for the Tamura-Nei model:

In the control region of mammalian mitochondrial DNA, the rate of nucleotide substitution is known to vary extensively from site to site, and there is a strong transition/transversion bias. The gamma distance for the Tamura-Nei model was developed primarily for the sequence data for this region. There are two hypervariable segments (5′ and 3′ segments), and the middle section is highly conserved. Using human data, Kocher and Wilson (1990) and Tamura and Nei (1993) estimated that a is about 0.11 for the entire control region, whereas Wakeley (1993) obtained a=0.47 for the 5′ hypervariable segment. Since most investigators use only the 5′ hypervariable segment, we have decided to use a=0.5 for the default option of MEGA. The gamma distance for the Tamura-Nei model is given by

$$\begin{aligned}\hat{d} = 2a[&\frac{g_A g_G}{g_R}(1 - \frac{g_R}{2g_A g_G}P_1 - \frac{1}{2g_R}Q)^{-\frac{1}{a}} \\ &+ \frac{g_T g_C}{g_Y}(1 - \frac{g_Y}{2g_T g_C}P_2 - \frac{1}{2g_Y}Q)^{-\frac{1}{a}} \\ &+ (g_R g_Y - \frac{g_A g_G g_Y}{g_R} - \frac{g_T g_C g_R}{g_Y})(1 - \frac{1}{2g_R g_Y}Q)^{-\frac{1}{a}} \\ &- g_A g_G - g_T g_C - g_R g_Y\,] \,.\end{aligned} \tag{4.38}$$

A formula for the variance of this estimate is available (Tamura and Nei 1993) but is not reproduced here. It is incorporated in the computer program.

Tamura and Nei also derived formulas for the estimates of the average numbers of transitional ($\hat{S}$) and transversional ($\hat{V}$) nucleotide substitutions per site, their variances, and the variance of the $\hat{S}/\hat{V}$ ratio. These formulas are incorporated in MEGA.

4.2 Synonymous and Nonsynonymous Substitutions

Nucleotide substitutions in coding genes can be subdivided into two classes, i.e., synonymous and nonsynonymous substitutions. Synonymous (or silent) substitutions are the nucleotide substitutions that do not result in amino acid changes, whereas nonsynonymous substitutions are those that change amino acids. The former substitutions are likely to be subject to little purifying selection except in lower organisms (however, see Britten 1993), while a majority of nonsynonymous changes are eliminated by purifying selection. Therefore, the rate of synonymous substitution is usually higher than that of nonsynonymous substitution (Miyata *et al.* 1980, Kimura 1983). Under certain conditions, however, nonsynonymous substitution may be accelerated by positive Darwinian selection (Hughes and Nei 1988, Lee and Vacquier 1992, and others). It is therefore interesting to examine the number of synonymous substitutions per synonymous site and the number of nonsynonymous substitutions per nonsynonymous site.

There are several methods for estimating these numbers (Miyata and Yasunaga 1980, Li *et al.* 1985, and others). In MEGA, however, we have included the simple method given by Nei and Gojobori (1986), since all methods give essentially the same results unless there are strong transition/transversion and G+C content biases. In Nei and Gojobori's (1986) method the numbers of synonymous (S) and nonsynonymous (N) sites are first computed. Here synonymous and nonsynonymous sites are the sites at which synonymous and nonsynonymous substitutions potentially occur, respectively (see Nei 1987, Pp. 73-76 for the method of computation). The sum of S and N is equal to the total number of nucleotides, n, and N is usually much larger than S. The numbers of synonymous (S_d) and nonsynonymous (N_d) substitutions that have occurred between two sequences are then computed by considering all pathways of nucleotide substitution between each pair of codons compared.

Using these quantities, we can compute the proportion of synonymous (p_S) and nonsynonymous (p_N) nucleotide differences per synonymous and nonsynonymous site, respectively. They are

$$p_S = S_d/S, \tag{4.39}$$

$$p_N = N_d/N, \tag{4.40}$$

with variances

$$V(p_S) = p_S(1 - p_S)/S, \tag{4.41}$$

$$V(p_N) = p_N(1 - p_N)/N. \tag{4.42}$$

Approximate estimates of the number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) can be obtained by applying the Jukes-Cantor formula.

$$\hat{d}_S = - {}^3/_4 \log_e(1 - {}^4/_3\, p_S), \tag{4.43}$$

$$\hat{d}_N = - {}^3/_4 \log_e(1 - {}^4/_3\, p_N), \tag{4.44}$$

with variances

$$V(\hat{d}_S) = p_S(1 - p_S)/[(1 - {}^4/_3\, p_S)^2\, S], \tag{4.45}$$

$$V(\hat{d}_N) = p_N(1 - p_N)/[(1 - {}^4/_3\, p_N)^2\, N]. \tag{4.46}$$

Computer simulations have shown that the above equations give good estimates if there is no transition/transversion bias (Nei and Gojobori 1986). However, when this bias is large, d_S tends to underestimate the true number of substitutions (Kondo *et al.* 1993). Li (1993) and Pamilo and Bianchi (1993) developed a method to take care of this problem. It is also possible to extend Nei and Gojobori's method to this case. We plan to include these methods in future versions of MEGA.

It should be noted that $\hat{d}_S$ and $\hat{d}_N$ are not reliable when p_S and p_N are large, say greater than 0.4, because their variances are large. In this case one may use p_S and p_N directly, particularly for studying positive Darwinian selection (Tanaka and Nei 1989).

In the study of adaptive evolution at the nucleotide level it is often necessary to compare the average values of d_S and d_N or p_S and p_N for a group of related sequences (e.g., Hughes and Nei 1988). In this case we have to know the variances of average d_S and d_N or average p_S and p_N. These variances can be computed by Nei and Jin's (1989) method, and this computation is implemented in MEGA.

Once these values are computed, the statistical significance of the difference (d) between average d_S and d_N or average p_S and p_N can be tested by the t-test with an infinite degrees of freedom. That is, t is given by

$$t = d/s(d), \tag{4.47}$$

where $s(d)$ is the standard error of d and is given by $[V(\bar{d}_S) + V(\bar{d}_N)]^{1/2}$ or by $[V(\bar{p}_S) + V(\bar{p}_N)]^{1/2}$. Here, $V(\bar{d}_S)$, $V(\bar{d}_N)$, $V(\bar{p}_S)$, and $V(\bar{p}_N)$ are the variances of average d_S, d_N, p_S, and p_N , respectively.

4.3 Amino Acid Substitutions

The methods for estimating the number of amino acid substitutions are similar to those for estimating the number of nucleotide substitutions except that there are 20 different states for the former rather than four states. The distance measures presented below can be computed either from amino acid sequences or from the coding regions of nucleotide sequences. In MEGA nucleotide sequences are translated into amino acid sequences by using one of the four genetic code tables ("universal" code and mammalian, *Drosophila*, and yeast mitochondrial genetic codes). Presence of a stop codon aborts the translation process and produces an error message. The treatment of missing nucleotides (or amino acids) and alignment gaps is discussed in the following section.

p-distance

As in the case for nucleotide sequences, the p-distance is merely the proportion of different amino acids between two sequences compared. Therefore, the statistical properties of this distance are the same as those of the p-distance for nucleotide sequence data.

$$p = n_d/n, \tag{4.48}$$

$$V(p) = p(1 - p)/n \tag{4.49}$$

Here n_d and n are the number of amino acid differences and the total number of amino acids compared, respectively.

Poisson-correction distance

This distance is for estimating the number of amino acid substitutions per site under the assumption that the number of amino acid substitutions at each site follows the Poisson distribution. This estimator ($\hat{d}$) and its variance are given by

$$\hat{d} = -\log_e(1 - p), \tag{4.50}$$

$$V(\hat{d}) = p/[(1 - p)n], \tag{4.51}$$

where p is estimated by equation (4.48).

Gamma distance

This distance is an estimate of the number of amino acid substitutions per site under the assumption that the rate of amino acid substitution varies from site to site and follows the gamma distribution with parameter a. This distance and its variance can easily be computed from Nei *et al.*'s (1976) work.

$$\hat{d} = a[(1 - p)^{-1/a} - 1], \tag{4.52}$$

$$V(\hat{d}) = p[(1 - p)^{-(1 + 2/a)}]/n. \tag{4.53}$$

In the default option of MEGA, a=2 is used. When a=2 is used, $\hat{d}$ is close to Dayhoff's (1978) PAM distance per site (0.01 PAM) (Tatsuya Ota, personal communication).

4.4 Guidelines for Choosing Distance Measures

In the above three sections, we have discussed various distance measures considering different situations. In general, a complex mathematical model fits data better than a simple one. However, a complex model requires the estimation of many parameters, and this increases the variance of the estimate of d. Theoretically, it is possible to choose a distance measure most appropriate for a given set of data by using certain statistical criteria. Such statistical criteria are now under investigation (Bulmer

1991, Goldman 1993, Tamura 1994), but it seems to be premature to include these model-selection methods in this version of MEGA. Without such model-selection methods, it is possible to write some guidelines for choosing distance measures for the purpose of phylogenetic inference (modified from Nei 1991). They are as follows:

(1) When the Jukes-Cantor estimate of the number of nucleotide substitutions per site (*d*) between different sequences is about 0.05 or less ($d \leq 0.05$), use the Jukes-Cantor distance whether there is a transition/transversion bias or not or whether the substitution rate (λ) varies with nucleotide site or not. In this case, the Kimura distance or the gamma distance gives essentially the same value as the Jukes-Cantor distance. One may also use the *p*-distance for constructing a topology.

(2) When $0.05 < d < 0.3$, use the Jukes-Cantor distance unless the transition/transversion ratio (R) is high, say $R > 2$. When this ratio is high and the number of nucleotides examined is large, use the Kimura distance or the gamma distances for Kimura's 2-parameter model.

(3) When $0.3 < d < 1$ and there is evidence that λ varies extensively with site, use gamma distances. In general, one may choose different gamma distances, estimating *a* from data.

(4) When $0.3 < d < 1$ and the frequencies of the four nucleotides (A,T,C,G) deviate substantially from equality but there is no strong transition/transversion bias, use the Tajima-Nei distance. When there are strong transition/transversion and G+C content biases, use the Tamura or Tamura-Nei distance.

(5) When $d > 1$ for many pairs of sequences, the phylogenetic tree estimated is not reliable for a number of reasons (e.g., large standard errors of *d*'s and sequence alignment errors). We therefore suggest that these sets of data should not be used. In this case one may eliminate the portion(s) of the gene that evolves very fast and use only the remaining region(s) as is often done in studies of the evolution of different kingdoms or phyla using ribosomal RNA genes. If a coding region of DNA is examined, we suggest that amino acid sequences rather than DNA sequences be used. One may also use a different gene that evolves more slowly.

In the study of evolution of multigene families, it is often necessary to examine phylogenetic relationships (topologies) of distantly related sequences with $d > 1$. In this case the nucleotide or amino acid *p*-distance is helpful because this distance has a smaller variance and it generates no inapplicable cases (e.g., Burke *et al.* 1993).

(6) When a phylogenetic tree is constructed from the coding regions of a gene, the distinction between synonymous (d_S) and nonsynonymous (d_N) substitutions may be helpful because the rate of synonymous substitutions is usually much higher

than that of nonsynonymous substitution. When relatively closely related species with $d_S < 1$ are studied for a large number of codons, one may use d_S for constructing a tree. This procedure is expected to reduce the effect of variation in substitution rate among different sites, because synonymous substitutions are apparently largely neutral in higher organisms. However, when relatively distantly related species are studied, d_N or amino acid distances should be used.

(7) As a general rule, if two distance measures give similar distance values for a set of data, use the simpler one because it has a smaller variance. When the rate of nucleotide substitution is the same for all evolutionary lineages and the number of nucleotides used is relatively small, the *p* or Jukes-Cantor distance seems to give a correct tree more often than the Kimura distance even if there is some extent of transition/transversion bias (Schöniger and von Haeseler 1993, Tajima and Takezaki, 1994). When the substitution rate varies with evolutionary lineage, however, this is not the case.

Note that the above guidelines are for constructing phylogenetic trees. For estimating evolutionary times or for testing the reliability of branch lengths, unbiased estimators are better than biased estimators, though unbiased estimators may vary with the data set used.

4.5 Alignment Gaps and Sites with Missing Information

Gaps are often inserted during the alignment of homologous regions of sequences and represent deletions or insertions (indels). These gaps introduce some complications in distance estimation. Furthermore, sites with missing information can sometimes occur because of experimental difficulties, and they create the same problems as that for gaps. In the following discussion both of these sites are treated in the same way.

In MEGA, gap sites are ignored in distance estimation, but there are two different ways to treat these sites. One way to deal with this problem is to delete all of these sites from data analysis. This option, which is called the *Complete-Deletion* option in MEGA, is generally desirable because different regions of DNA or amino acid sequences often evolve under different evolutionary forces. However, if the number of nucleotides involved in a gap is small and gaps are distributed more or less at random, one may compute a distance for each pair of sequences ignoring only those gaps that are involved in the comparison. This option is called the *Pairwise-Deletion* option. Table 4.2 illustrates the effect of these options on distance estimation with the following three sequences:

```
                 1        10        20
seq1             A-AC-GGAT-AGGA-ATAAA
seq2             AT-CC?GATAA?GAAAAC-A
seq3             ATTCC-GA?TACGATA-AGA    Total sites = 20.
```

Here, the alignment gaps are indicated with a hyphen (-) and the missing information sites are denoted by a question mark (?).

Table 4.2 Complete-Deletion and Pairwise-Deletion options

Option	Sequence Data	Differences/Comparisons		
		(1,2)	(1,3)	(2,3)
Complete-Deletion	1. A C GA A GA A A A	1/10	0/10	1/10
	2. A C GA A GA A C A			
	3. A C GA A GA A A A			
Pairwise-Deletion	1. A-AC-GGAT-AGGA-ATAAA	2/12	3/13	3/14
	2. AT-CC?GATAA?GAAAAC-A			
	3. ATTCC-GA?TACGATA-AGA			

In Table 4.2, the number of sites compared varies with pairwise comparison in the Pairwise-Deletion option, but it remains the same for all pairwise comparisons in the Complete-Deletion option. In this particular data set, more information can be obtained by using the Pairwise-Deletion option. In practice, however, different regions of nucleotide or amino acid sequences often evolve differently. In this case the Complete-Deletion option is preferable.

5

Phylogenetic Inference

Reconstruction of the evolutionary history of genes and species is one of the most important subjects in the current study of molecular evolution. If reliable phylogenies are produced, they will shed light on the sequence of evolutionary events that generated present day diversity of genes and species and help us to understand the mechanisms of evolution as well as the history of organisms.

5.1 Phylogenetic Trees

Species Trees and Gene Trees

Reconstruction of phylogenetic trees is a statistical problem (Cavalli-Sforza and Edwards 1967), and a tree reconstructed is an estimate of a true tree with a given topology and given branch lengths. Biologists are often interested in knowing the history of species (or population) splitting and divergence times after each splitting event. When these historical events are expressed in terms of a phylogenetic tree, this tree is called a *species (or population) tree*. It is usually very difficult to know the true species tree for any group of organisms, but it is possible to infer the species tree by examining the evolutionary relationships of genes from the organisms involved. A phylogenetic tree based on a gene (nucleotide or amino acid sequences) is called a *gene tree*. A gene tree may not agree with the species tree, because (1) nucleotide or amino acid substitution is subject to stochastic errors and (2) a gene tree is affected by sampling errors of polymorphic alleles that existed in the ancestral populations (Tajima 1983, Nei 1986, Neigel and Avise 1986, Pamilo and Nei 1988). The disagreement of gene trees and species trees may also occur when there are two or more copies of the same gene in the genome (Goodman *et al.* 1982). In general, one has to study many genes to infer a species tree.

Of course, gene trees are not studied just for inferring a species tree. In many cases, they are interesting for their own right. When one wants to know the evolutionary relationships of genes belonging to a multigene family or of polymorphic alleles within and between species, one must study gene trees.

Rooted vs. Unrooted Trees

Phylogenetic relationships of genes or organisms are usually presented in a treelike form with a root, as those in figure 5.1(A). This type of tree is called a *rooted tree*. It is also possible to draw a tree without a root, like those in figure 5.1(B). This type of tree is called an *unrooted tree*. The branching pattern of a tree is called a *topology*. There are many possible rooted and unrooted trees for a given number of species (m). In the case of $m = 4$, there are 15 possible rooted trees and 3 possible unrooted trees, as shown in figure 5.1. The number of possible trees rapidly increases with increasing m. In general, the number of bifurcating rooted trees for m species is given by

$$1 \cdot 3 \cdot 5 \cdot 7 \cdot \ldots \cdot (2m - 3) = (2m - 3)!/[2^{m-2}(m - 2)!] \qquad (5.1)$$

for $m \geq 2$ (Cavalli-Sforza and Edwards 1967). This indicates that when $m = 10$, the number is 34,459,425. Only one of these trees is the true tree. The number of bifurcating unrooted trees for m species is given by replacing m by $m - 1$ in equation (5.1). This becomes more than two million for $m = 10$. In many cases, of course, a majority of these possible trees can be excluded because of obviously unlikely genetic relationships or of other biological information. Nevertheless, it is a very difficult task to find the true phylogenetic tree from observed data on extant species when m is large.

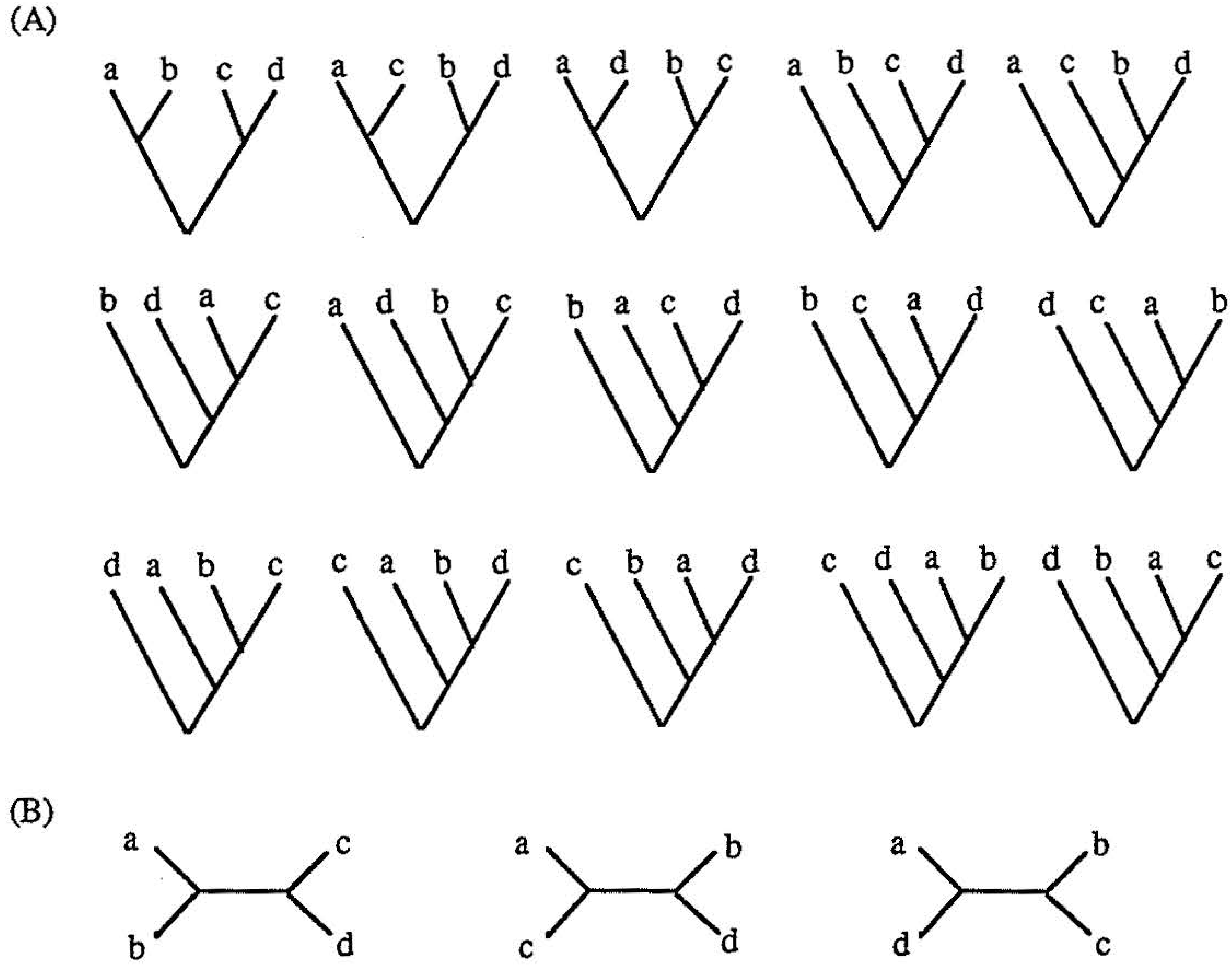


Fig 5.1 Fifteen possible rooted trees (A) and three possible unrooted trees (B) for four species.

5.2 Tree-Building Methods

There are numerous methods for constructing phylogenetic trees from molecular data (see Felsenstein 1988, Miyamoto and Cracraft 1991). They can be classified into *distance methods* and *discrete-character methods.* In distance methods, a pairwise evolutionary distance is computed for all species or OTUs to be studied, and a phylogenetic tree is constructed by certain principles and algorithms. In discrete-character methods, data with discrete character states such as nucleotide states in DNA sequences are used, and a tree is constructed by considering the evolutionary relationships of OTUs or DNA sequences at each character or nucleotide position.

It should be noted that some types of molecular data (e.g., DNA hybridization data) exist only as distance data. Therefore, phylogenetic trees for these data can be constructed only by distance methods. By contrast, discrete-character data can usually be converted into distance data. Therefore, they can be analyzed either by distance methods or by discrete-character methods. Some authors (e.g., Farris 1981, Penny 1982) have argued that distance methods are inherently inferior to discrete-character methods (e.g., parsimony methods), but their arguments are apparently based on misconceptions of distance methods (Felsenstein 1986, Nei 1987). Actually, some distance methods can be superior to discrete character methods in obtaining the correct tree, depending on the situation.

Recent computer simulations (see Nei 1991) have shown that one of the most efficient distance methods in recovering the correct topology is the neighbor-joining method proposed by Saitou and Nei (1987). Empirical studies have also shown that their method generally gives reasonable trees. Therefore, we decided to include this method in MEGA. Another distance method included in MEGA is the unweighted pair-group method with arithmetic means (UPGMA; Sneath and Sokal 1973). We included this method because of its simplicity and utility under certain circumstances. There are several other popular distance methods such as Fitch and Margoliash's (1967) method, but they are not included in MEGA because they are available in PHYLIP (Felsenstein 1993).

There are two major groups of discrete character methods, i.e., maximum parsimony methods and maximum likelihood methods. [The compatibility (Le Quesne 1969, Estabrook *et al.* 1975) and evolutionary parsimony (Lake 1987) methods are also discrete-character methods, but they are rarely used.] The first group of methods are included in PAUP (Swofford 1993), whereas the second are available in PHYLIP (Felsenstein 1993). Therefore, there seems to be no need to include these methods in MEGA. However, we have developed new algorithms for the maximum parsimony method for molecular data, and these algorithms seem to be quite efficient in obtaining maximum parsimony trees. We have therefore included these new algorithms in MEGA.

5.3 UPGMA

This method was originally proposed for taxonomic purposes, but it is possible to use it for tree building if we assume that the rate of nucleotide or amino acid substitution is the same for all evolutionary lineages. Computer simulations have shown that when the molecular clock works and the evolutionary distance is large for all pairs of OTUs, it recovers the correct tree with a reasonably high probability (Tateno *et al.* 1982; Sourdis and Krimbas 1987). One interesting aspect of this method is that it produces a simple tree that mimics a species tree, the branch lengths for two OTUs being the same after their separation. It is therefore appealing to biologists who are interested in constructing species trees. If this method is applied to distance data computed from many genes with large numbers of nucleotides, it is expected to give a reasonably good tree. At present, however, many investigators use relatively short sequences for phylogenetic construction, and the molecular clock often fails to work for DNA sequences. Therefore, one should be cautious about UPGMA trees. Because of the assumption of a constant rate of evolution, this method produces a rooted tree, though it is possible to remove the root for certain purposes (see section **5.6.2**).

Algorithm

In this method, it is important to use a distance measure that is linearly related to evolutionary time. Once such distances are computed for all pairs of OTUs, they can be presented in the following matrix form.

$$\begin{array}{llll}
\text{OTU} & 1 & 2 & 3 \\
2 & d_{12} & & \\
3 & d_{13} & d_{23} & \\
4 & d_{14} & d_{24} & d_{34}
\end{array}$$

(In MEGA, the distance values can be presented either below or above the diagonal.) Here, d_{ij} stands for the distance between the i-th and j-th OTUs. Clustering of OTUs starts with the two OTUs with the smallest distance, and more distantly related OTUs are gradually added to the cluster. Suppose that the distance (d_{34}) between OTUs 3 and 4 is smallest among all distance values in the above matrix. These two OTUs are then clustered with a branch point located at distance ½d_{34}. Here, we have assumed that the lengths of the branches leading from the branch point to OTUs 3 and 4 are the same. OTUs 3 and 4 are then combined into a single OTU (34). New distances between this combined OTU and the other OTUs are then calculated:

$$\begin{array}{lll}
\text{OTU} & 1 & 2 \\
2 & d_{12} & \\
(34) & d_{1(34)} & d_{2(34)}
\end{array}$$

Here, $d_{1(34)}$ and $d_{2(34)}$ are given by ½($d_{13} + d_{14}$) and ½($d_{23} + d_{24}$), respectively. We again search for the smallest value in the new distance matrix. Suppose that $d_{2(34)}$ is smallest. OTU 2 then joins the (34) cluster with a branch point located at distance $d_{2(34)}$

/2. In this case, OTU 1 is the last to be clustered. The branch point at which this last OTU joins the others is $\frac{1}{2}d_{1(234)} = (d_{12} + d_{13} + d_{14})/(3 \times 2)$. If $d_{1(34)}$ is smallest among the three distance values above, OTU 1 joins the (34) cluster first and then OTU 2. On the other hand, if d_{12} is smaller than any of $d_{1(34)}$ and $d_{2(34)}$, OTUs 1 and 2 are clustered, and then the two clusters (12) and (34) are joined into the final single family. When more than four OTUs are involved, the above procedure is continued until all OTUs are clustered into a single family.

5.4 Neighbor-Joining (NJ) Method

This method (Saitou and Nei 1987) is a simplified version of the minimum evolution (ME) method (Saitou and Imanishi 1989, Rzhetsky and Nei 1992). In the ME method, distance measures that correct for multiple hits at the same sites are used, and a topology showing the smallest value of the sum (S) of all branches ($2m - 3$ branches for a bifurcating tree with m OTUs) is chosen as an estimate of the correct tree. Rzhetsky and Nei (1993) have shown that when unbiased estimates of evolutionary distances are used, the true tree (topology) always gives the smallest expected value of S. Therefore, the minimum evolution method has a solid theoretical foundation.

However, construction of a minimum evolution tree is time-consuming, because in principle the S values for all topologies have to be evaluated. The number of possible topologies (unrooted trees) rapidly increases with the number of OTUs. Therefore, it becomes very difficult to examine all topologies, though there are some ways to exclude all unlikely trees (Rzhetsky and Nei 1992, 1993).

In the case of the NJ method, the S value is not computed for all or many topologies, but the examination of different topologies is imbedded in the algorithm, so that only one final tree is produced. Since the algorithm of the NJ method is somewhat complicated, we shall not present it here. If the user of this program is interested in the algorithm, he or she should refer to Saitou and Nei's (1987) original paper and Studier and Keppler's (1988) slight modification. This method produces an unrooted tree, and it usually requires an outgroup OTU to find the root. In the absence of outgroup OTU's, the root is sometimes given at the midpoint of the longest route connecting two OTU's in the tree under the assumption of a constant rate of evolution. In MEGA, this practice is used unless outgroup OTUs are specified.

As mentioned above, the NJ tree is usually the same as the ME tree when the number of OTUs is small. However, if this number is large and the extent of sequence divergence is small, the topological difference between the NJ and ME trees can be substantial (Rzhetsky and Nei 1993). In this case the ME tree is obviously preferable, though the difference in S between the NJ and ME trees is usually statistically nonsignificant. In MEGA, we have not included the program for obtaining ME trees, because it requires a large amount of computer memory. A computer program (METREE) for this method is available (see Appendix E).

5.5 Maximum Parsimony (MP) Method

Maximum parsimony (MP) methods were originally developed for morphological characters, and there are many different versions (Sober 1988, Maddison and Maddison 1992, Swofford 1993). In MEGA we consider only the method that is appropriate for nucleotide sequence data, i.e., the method where evolutionary change is assumed to occur between any pair of the four nucleotides (Fitch 1971). [This is a special case of Eck and Dayhoff's (1966) method, where evolutionary change is allowed to occur between any pair of the 20 different kinds of amino acids.] In this method it is possible to give different weights to different types of substitutions (e.g., transitions and transversions, Sankoff and Cedergren 1983, Williams and Fitch 1990), but this type of modified parsimony methods will not be considered here.

The MP method is not always a *consistent estimator* of the true tree. [A tree-building method is said to be a consistent estimator if it gives the correct tree (topology) when an infinitely large number of nucleotides are used.] Felsenstein (1978) showed that the MP method is an inconsistent estimator when the evolutionary rate varies extensively with evolutionary lineage. Inconsistency of the MP method is known to occur even in the case of constant rate if the true tree has very short interior branches (Hendy and Penny 1989, DeBry 1992). Furthermore, even when the MP method is a consistent estimator, its efficiency of obtaining the correct tree seems to be generally lower than that of the neighbor-joining and maximum likelihood methods (Saitou and Imanishi 1989, Tateno *et al.* 1994). However, when (1) the extent of sequence divergence is small ($d<0.1$), (2) the rate of nucleotide substitution is more or less constant, (3) there are no strong transition/transversion and G+C content biases, (4) the number of nucleotides examined is very large (more than a few thousand nucleotides), and (5) a small number of sequences are used, it seems to be a good method for estimating the true tree (Sourdis and Nei 1988, Nei 1991). Furthermore, unlike the distance or maximum likelihood method, this method is capable of using information on insertions/deletions.

For constructing an MP tree, only nucleotide sites at which there are at least two different kinds of nucleotides, each represented at least twice, are used. These sites are called *parsimony-informative sites*. Other variable sites are not used for constructing an MP tree, though they are informative for distance and maximum-likelihood methods.

In the MP method, the nucleotides of ancestral sequences are inferred at each nucleotide site for a given tree topology, and the minimum number of substitutions that are required to explain the nucleotide differences is counted. The sum of this number over all parsimony-informative sites of the sequences for a given topology is called the *number of steps* or the *tree length*. The tree length is then computed for all possible topologies, and the topology that shows the smallest tree length is chosen as the final tree (*maximum parsimony tree*). In practice, there may be two or more topologies that show the smallest tree length. These topologies are called *equally parsimonious trees*. The MP method is intended to find unrooted trees, and its primary goal is to determine the topology of a tree. Although it is possible to estimate branch lengths under certain assumptions (Fitch 1971, Maddison and Maddison 1992, Swofford 1993), the estimates for molecular data are usually poor unless the extent of sequence divergence is very small. Therefore, we shall not consider the estimation of branch lengths of MP trees in MEGA.

When the number of OTUs (m) is small, say $m < 10$, it is possible to examine all possible trees and determine the MP tree, though it can be very time-consuming. This type of search for an MP tree is called the *exhaustive search*. This method is not included in MEGA, because it is found in PAUP. As mentioned earlier, however, the number of topologies rapidly increases as m increases. Therefore, it is virtually impossible to examine all topologies if m is larger than 10.

There are two ways of dealing with this problem. One is to use the *branch-and-bound method* (Hendy and Penny 1982). In this method, the trees which obviously have a tree length longer than that of a previously examined tree are all ignored, and the MP tree is determined by evaluating the tree lengths for a group of trees that potentially have shorter tree lengths. This method guarantees finding of all MP trees, though it is not an exhaustive search. However, even this method becomes very time-consuming if m is 20 or larger. In this case one has to use another approach called the *heuristic search* method. In this method only a small proportion of all possible trees is examined, and there is no guarantee that the MP tree will be found. Nevertheless, it is possible to enhance the probability of obtaining the MP tree.

5.5.1 Branch-and-Bound Search

In the branch-and-bound method the search for an MP tree starts with a core tree of three OTUs, which has only one unrooted tree [Fig. 5.2 (A)]. Other OTUs are added to this core tree one by one according to a certain rule, and the tree length is computed at each stage of OTU addition. If the addition of an OTU to a particular branch of a core tree results in a tree length greater than a predetermined upperbound of tree length (L_U), this topology and all the subsequent topologies that can be generated by adding more OTUs to this core tree will be ignored from further consideration.

In our branch-and-bound algorithm, the initial core tree of three OTUs is chosen such that the length (L) of the tree is largest (or approximately largest) among all possible 3-OTU trees. This is to make this L closer to the length (L_M) of the MP tree so that we can reach the MP tree faster. To obtain this initial tree, we first compute the nucleotide differences for all possible pairs of OTUs and choose the pair that shows the largest number of nucleotide differences. We then make a tree of three OTUs using this pair of OTUs and one of the remaining OTUs. For this tree, we compute the tree length using the maximum parsimony principle. This process is repeated for all the remaining OTUs, and a tree of three OTUs that shows the largest L value is chosen as the initial core tree.

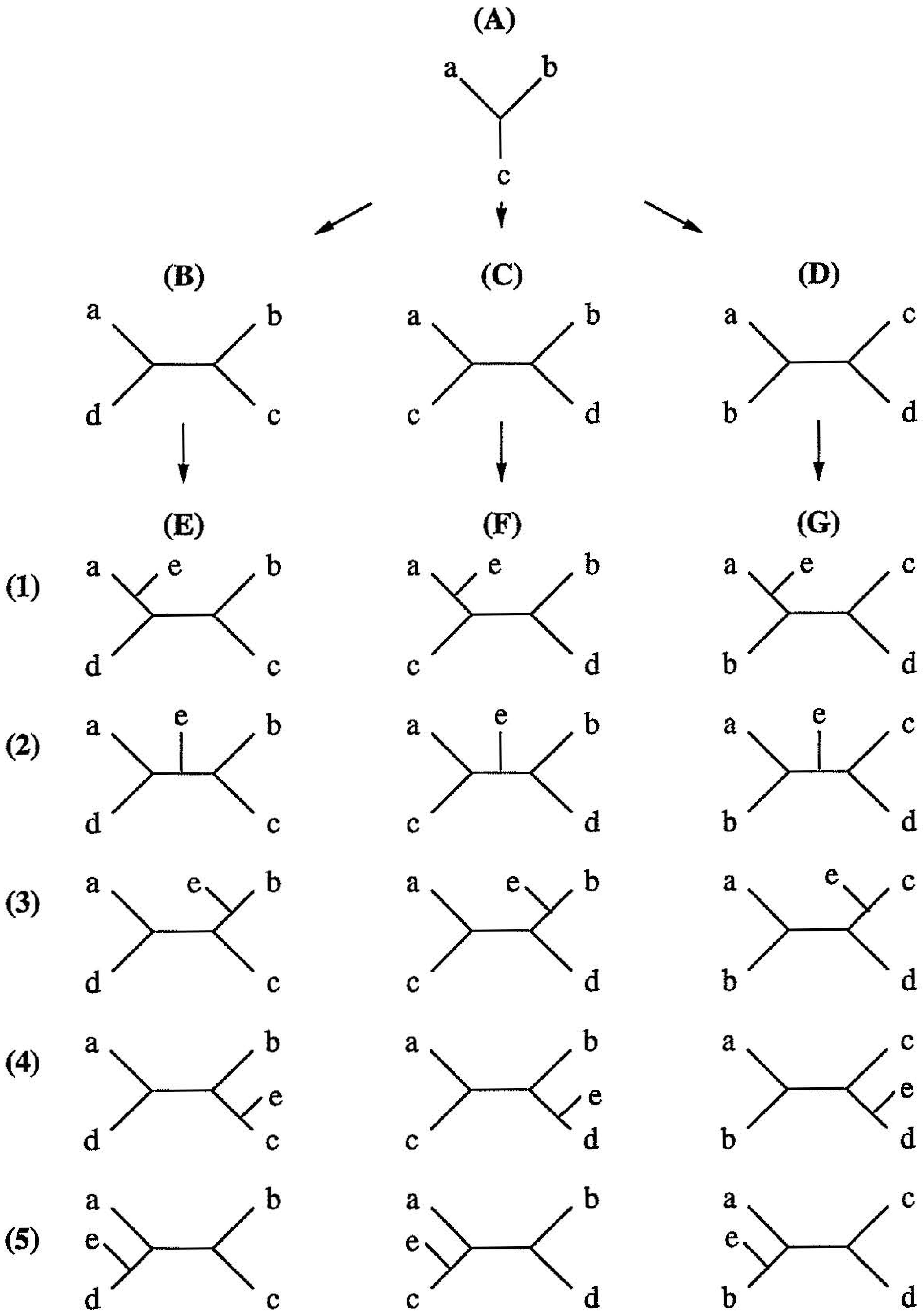


Fig 5.2 Diagrams showing the procedure of the branch-and-bound and heuristic searches.

Order of OTU addition

The next step is to determine the order of OTU addition that makes the search for the MP tree faster. Our algorithm for this step is as follows. We add one of the remaining OTUs to one of the three branches of the core tree and compute the tree length by the MP procedure. We repeat the same computation for the two remaining branches and record the minimum value of the three tree lengths. We apply the same procedure for all remaining OTUs. We then find the OTU that shows the maximum value of the minimum tree lengths. This OTU is the first OTU to be added to the initial core tree. We call this procedure the maximum-of-the-minimum-values algorithm or simply the *max-mini algorithm*. To find the next OTU, we apply this max-mini algorithm for the remaining OTUs using the parsimony tree for the first four OTUs as the next core tree. In this case, of course, the number of minimum tree lengths to be computed for each OTU is five, because a 4-OTU tree has five branches. We can then find an OTU that shows the maximum of the minimum tree lengths. This OTU will be the second OTU to be added to the initial core tree of three OTUs. This process is repeated until the addition order of all OTUs is determined. Since the maximum of the minimum values is closer to L_M than some other value (e.g., the minimum of the minimum values), this order of OTU addition is expected to speed up the search for an MP tree.

Searching for MP tree(s)

Once the initial core tree and the order of OTU addition are determined, we are in a position to search for an MP tree. Before applying our algorithm for finding the MP tree, however, we must have a predetermined upperbound of tree length, i.e., L_U for a *temporary MP tree*. This value is a temporary minimum number of substitutions, which is likely to be slightly larger than the real minimum number, L_M. We determine this value by running our heuristic search program with search factor equal to 0 (see next section).

Let us now explain our algorithm with the diagrams in Fig. 5.2. We start with the initial core tree in diagram (A). In this example of five OTUs, OTUs *a, b,* and *c* form the initial core tree, and OTUs *d* and *e* are added in this order. There are three ways of adding *d* to the core tree [trees (B), (C), and (D)]. We first compute the tree length (L) for tree (B). If this L is greater than L_U, we ignore all the subsequent trees that are generated by adding OTU *e* to this tree [five trees given in column (E)]. If $L \leq L_U$, we add *e* to each of the five branches of tree (B) to form five trees with five OTUs. We again compute L for each of these five trees and find a tree (or trees), which shows the smallest L value. If this L is greater than L_U, then we move on to tree (C). However, if the L is equal to L_U, we save the tree with this L as another potential MP tree and move on to tree (C). On the other hand, if the smallest L is smaller than L_U, the tree with this L will become the next temporary MP tree, and the L_U is now replaced by this new L value. We then move to tree (C).

We apply the same procedure to tree (C) and the trees generated by adding *e* to tree (C). If all these trees are examined, we then move to tree (D) and its descendant trees. Since we adjust L_U whenever we find a tree with an L smaller than the previous L_U, we are assured to find the MP tree. Of course, there may be two or more equally parsimonious trees, but all these trees are identified by the present method. The same algorithm can be used for the case where the number of OTUs (m) is greater than 5. This algorithm saves computer time considerably, because many trees need not be examined if L_U is sufficiently close to the tree length (L_M) of the true MP tree. However, even this method becomes time-consuming if $m \geq 20$.

5.5.2 Heuristic Search

The algorithm of our heuristic search is somewhat similar to that of the branch-and-bound method mentioned above. We start with an initial core tree of three OTUs that is determined in the same way as before. The order of OTU addition is also determined in a similar fashion except for the following. In the case of the branch-and-bound method, we computed the minimum numbers of substitutions for all OTUs for each core tree (each step of addition) and then chose the OTU that showed the maximum value among all the minimum values (max-mini algorithm). In the case of the heuristic search we choose the minimum of all the minimum values, because we are not going to do a semi-exhaustive search as in the case of the branch-and-bound method. We call this procedure the minimum-of-the-minimum-values algorithm or simply the *mini-mini algorithm.*

The algorithm of the search for MP trees is also similar to that of the branch-and-bound method. Let us again consider Fig. 5.2 to explain this algorithm. As before, we start with the core tree (A) and first connect OTU d to branch a to produce tree (B). We then compute the tree length (L) of this tree. We call this the *local upperbound* (L_1) for the first OTU addition and keep this value for future use. We then connect OTU e to branch a of tree (B) to produce tree E(1). We again compute the L value of this tree and call it the local upperbound (L_2) for the second OTU addition. If there is another OTU (f) to be added, we connect this OTU to branch a of tree E(1) and obtain tree E(1, 1) in Fig. 5.3. If f is the last OTU to be added, we now compute the L value not only for tree E(1, 1) but also for all other six trees that can be derived from E(1) (see Fig. 5.3). We then choose the tree that shows the smallest L value among the seven trees and call it a temporary MP tree with tree length L_U.

The next step of search is to go back to tree E(2) in Fig. 5.2 and compute the L value. If this L is greater than L_2, we neglect all the trees that can be generated by adding f to this tree. If $L = L_2$, we compute L for all the descendant trees. If any of the trees shows an L equal to L_U, the tree is saved as another potential MP tree. If there is any tree showing an $L < L_U$, this tree is now considered as a new temporary MP tree, and the previous L_U is replaced by this L. By contrast, if E(2) shows an $L < L_2$, L_2 is replaced by this L. The L values for all descendant trees are then computed, and a new potential MP tree or a new temporary MP tree is searched for. This procedure is applied to the remaining three trees [E(3), E(4), and E(5)] of five OTUs, and the temporary MP tree (or trees) that shows the smallest L value among the 35 (= 5 × 7) trees derived

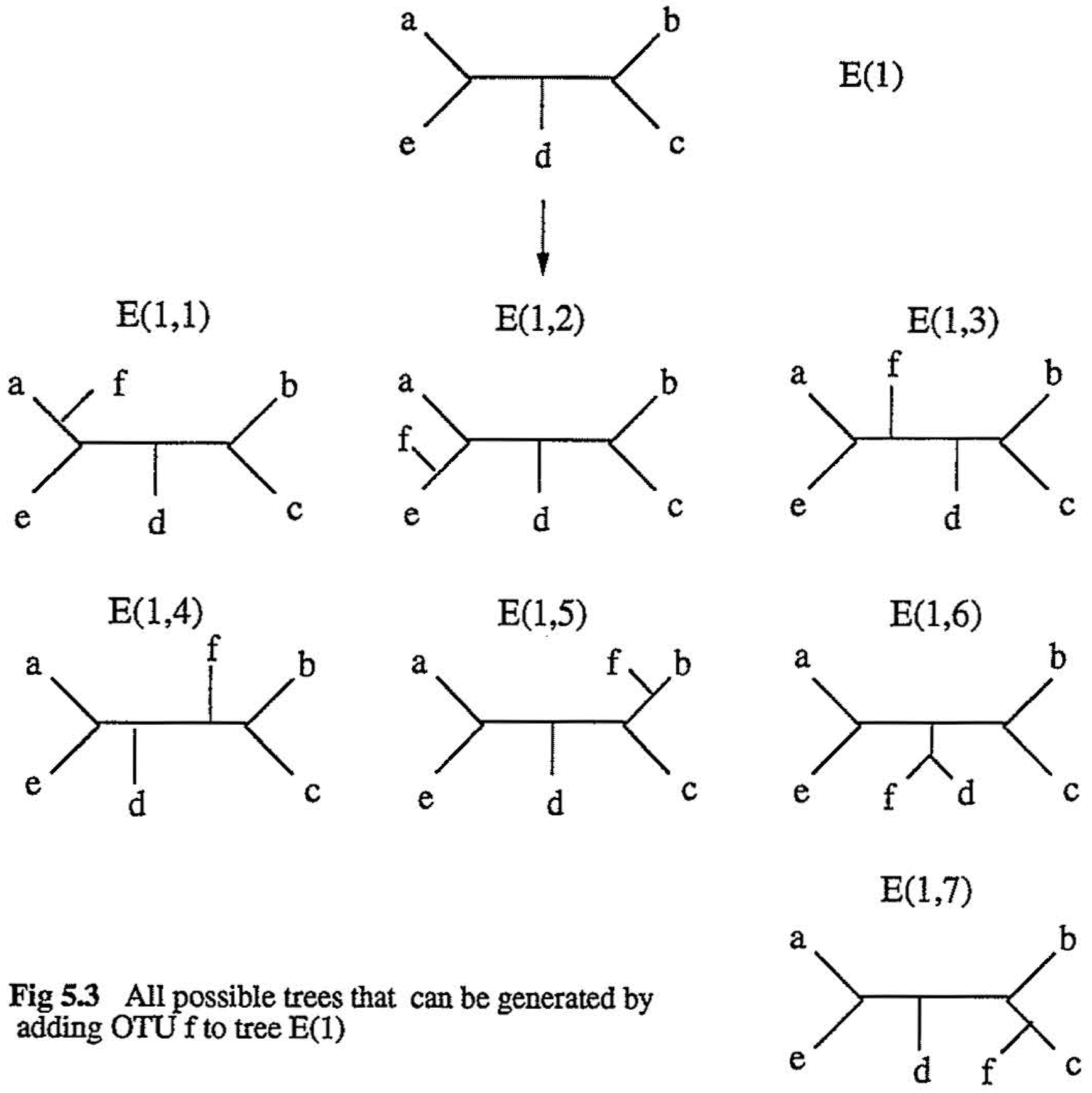
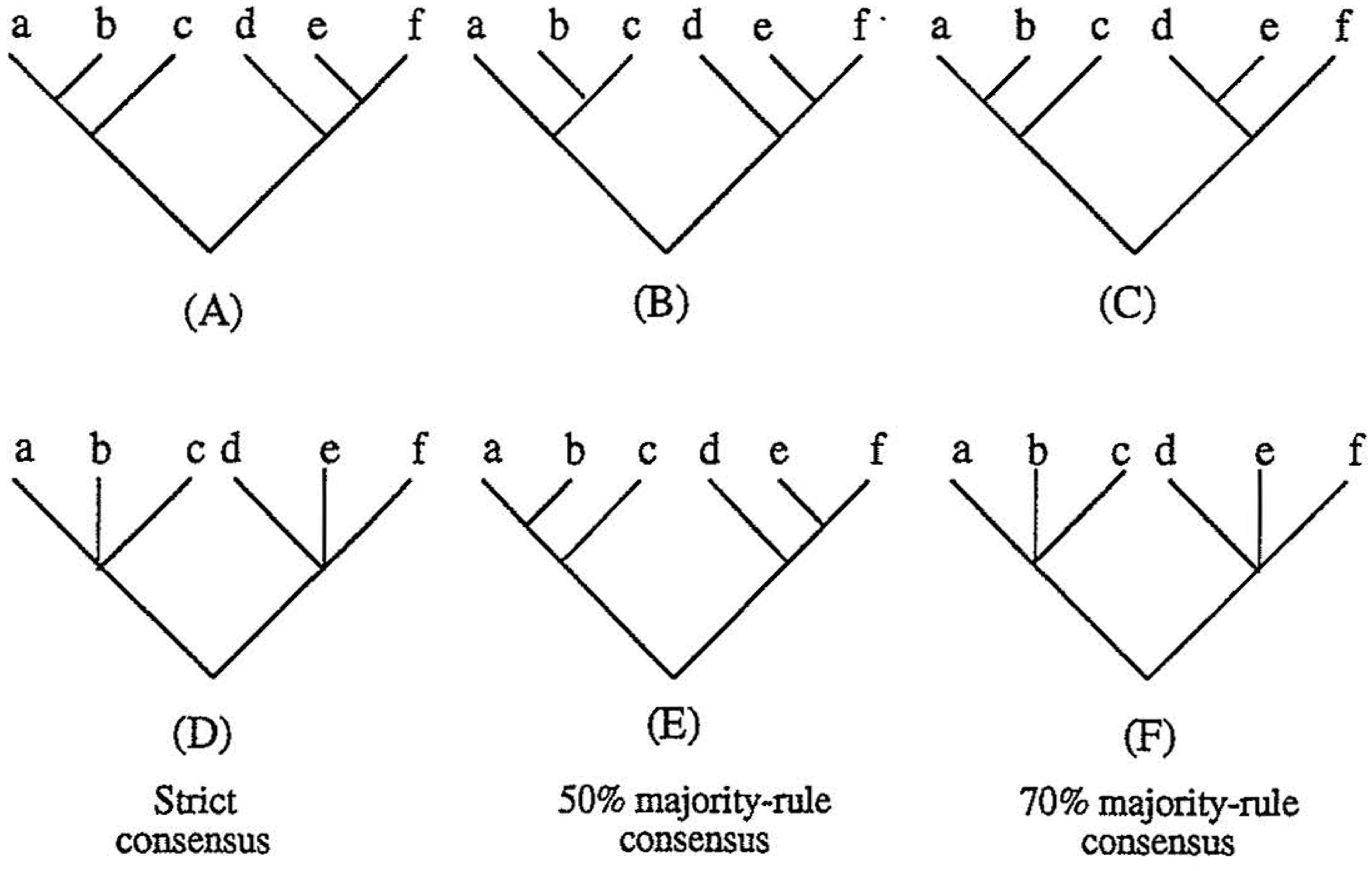


Fig 5.3 All possible trees that can be generated by adding OTU f to tree E(1)



Strict consensus

50% majority-rule consensus

70% majority-rule consensus

Fig. 5.4 Examples of consensus trees.

from tree (B) is determined.

If the above computation is completed, we now move to tree (C) in Fig. 5.2 and apply the same procedure to all trees that can be derived from this tree. Thus, if tree F(1) shows an $L > L_2$, all seven trees derived from this tree will be ignored. However, if $L = L_2$, all the descendant trees are examined for their L values. If F(1) shows an $L<L_2$, L_2 is now replaced by this L, and we use this new L_2 for the subsequent search of an MP tree. We then compute L for all the trees derived from F(1) by adding OTU *f*. If there is any tree with $L = L_U$, it will be saved as another potential MP tree. If there is a tree with $L < L_U$, it becomes a new temporary MP tree, and L_U is replaced by this L. The same procedure is applied to trees F(2), F(3), F(4), and F(5) and their descendant trees. Similarly, the same procedure is applied to tree (D) and its descendant trees. If this is completed, we have the final MP tree or trees determined.

When there are more than six OTUs, essentially the same algorithm is used. The only difference is that there are many steps of OTU addition and that at each step of OTU addition the local upperbound (L_1, L_2, L_3, ..., and L_{m-4}) is computed, where m is the number of sequences. L_1, L_2, L_3, ..., and L_{m-4} are then used to determine whether a group of descendant trees should be ignored or not in later computations.

In this algorithm, many trees which are unlikely to have a small L value are ignored from computation of their L values, and thus the algorithm speeds up the search for the MP tree. However, the final tree or trees obtained by this algorithm may not be the true MP tree(s), because the upperbounds of the L values used here are local bounds rather than the global upperbound as used in the branch-and-bound method and the tree with the global minimum value of L may not have been obtained.

There is a way to improve the efficiency of finding the MP tree. It is to increase the local upperbound at each step of OTU addition. If the local upperbound is large, the number of trees to be examined automatically increases. In the above algorithm, the local upperbound at the *i*-th step of OTU addition was L_i except for the first step. We now increase L_i by x_i, so that the upperbound is given by $L'_i = L_i + x_i$. If x_i is very large for all *i*'s, every topology will be examined. In this case, however, the computational time will be prohibitively large. We call x_i a *search factor*.

We are not yet sure about the optimal values of x_i's to obtain the MP tree most efficiently. Intuitively, one might argue that a rather large value of x_i be given for the first few steps and a small value of x_i be given for the subsequent steps. In the above explanation of our algorithm, we examined all three topologies [tree (B), (C), and (D)] at the first step of OTU addition. This corresponds to $x_1 = \infty$. In MEGA the user can use two different values of x_i, one for the steps before a certain specified step, which we call the *transition step*, and the other for all the remaining steps from and after the transition step. Our limited experience, however, has not necessarily supported the intuitive argument mentioned above. We have therefore decided to use $x_i = 2$ for all steps in the default option of MEGA. When search factors are large, excessive computer time may be required to complete the heuristic search. The user then should adjust the values of search factors.

Previously we mentioned that the first temporary MP tree for the branch-and-bound method is determined by the heuristic search. In MEGA, this tree is determined by using the search factors $x_1 = x_2 = \ldots = x_{m-4} = 0$.

It should be noted that there is no mathematical proof that the MP tree is the best estimator of the true tree. On the contrary, the MP tree is often an inconsistent estimator, as mentioned earlier. Therefore, it would be unwise to spend too much time for finding the true MP tree. When m is large, some parts of the MP tree (or any other tree) are likely to be incorrect. In this case a sub-maximum parsimony tree may serve the purpose of the investigator as well as the true MP tree does.

5.5.3 Alignment Gaps and Sites with Missing Information

In the MP method, information on alignment gaps caused by insertions/deletions (indels) may be used for phylogenetic inference. In this case one gap or indel is treated as an additional character state, i.e., the fifth state for nucleotide sequences. In MEGA, this option is included. If the gaps are not given the fifth character state, they are disregarded in the computation of tree lengths. In MEGA, sites with missing information may be included in the data, but they are never used in computing tree lengths. Of course, these sites can be eliminated from the phylogenetic analysis from the beginning. Particularly when the alignment gaps are long and the sequence alignment is questionable, this option is recommended.

5.5.4 Consensus Trees

The MP method often produces many equally parsimonious trees. In this case, it is difficult to present all the trees for publication. One way to solve this problem is to make a composite tree that includes all the trees. Such a composite tree is called a *consensus tree.*

There are several different types of consensus trees (Swofford 1993), but the most commonly used are the *strict consensus* and *majority-rule consensus trees*. Let us explain these trees using the examples given in Fig. 5.4. Suppose that trees (A), (B), and (C) are three equally parsimonious trees obtained by the MP method. In a strict consensus tree any conflicting branching patterns for a set of OTUs among the rival trees are resolved by forming a multifurcating branching pattern. Thus, the strict consensus tree for trees (A), (B), and (C) are given by tree (D). Among the majority-rule consensus trees, the most commonly used is the 50% majority-rule consensus tree. In this tree a branching pattern that occurs with a frequency of >50% is adopted. In the present sample, the branching pattern ((*ab*)*c*) for OTUs *a, b,* and *c* occurs two times among the three rival trees, so this pattern is adopted. Similarly, branching pattern ((*de*)*f*) occurs two times for the other cluster. Therefore, the 50% majority-rule consensus tree is given by tree (D). It is possible to change the majority-rule percentage to any value. For example, if we use 70%, none of the branching patterns of the two 3-OTU clusters reaches 70%. Therefore, the 70% majority-rule consensus tree [tree (F)] becomes identical with the strict consensus tree. Note that the 100% majority-rule consensus tree is always identical with the strict consensus tree.

5.6 Statistical Tests of a Tree Obtained

There are two different types of methods for testing the reliability of a tree obtained. One is to test the topological difference between the tree and its closely related tree by using certain quantity such as the likelihood value in the maximum likelihood method (Kishino and Hasegawa 1989) and the sum of all branch lengths in the minimum evolution method (Rzhetsky and Nei 1992). This type of test is supposed to examine the reliability of every interior branch of the tree, and it is generally a conservative test. The procedure of the test is usually quite complicated and requires a large amount of computer memory. We have therefore decided not to include it in MEGA. A computer program for the test for minimum-evolution trees is available separately (see Appendix E).

The other type of test is to examine the reliability of each interior branch whether it is significantly different from 0 or not. If a particular interior branch is not significantly different from 0, we cannot exclude the possibility of trifurcation of the branches associated or even the other types of bifurcating trees that can be generated by changing the splitting order of the three branches involved. There are two different ways of testing the reliability of an interior branch. One way is to compute the standard error of the interior branch and test the deviation of the branch length from 0, and the other is to use the bootstrap test (Efron 1982, Felsenstein 1985). These tests are included in MEGA.

5.6.1 NJ Trees

Since the statistical properties of NJ trees are better understood than those of UPGMA and MP trees, let us first discuss the test of these trees. In MEGA, the standard error test of NJ trees is conducted following Rzhetsky and Nei's (1992, 1993) method. That is, once an NJ tree is obtained by the Saitou-Nei algorithm, the branch lengths of the tree are re-estimated by using the ordinary least squares method, and the standard errors of the estimates are computed. Let b and $s(b)$ be an estimate of an interior branch length and its standard error, respectively. The statistical significance of b from 0 is then tested by the t-test [$t = b/s(b)$] with degrees of freedom ∞. In the standard statistical test, a null hypothesis is tested by computing the probability of Type I error (α;significance level). In MEGA, however, the complement of this probability $(1 - \alpha)$ is computed. We call this the *confidence probability (CP)*. Therefore, the reliability of a branch length is high when CP is high. Usually, if $CP \geq 0.95$ or 0.99, the branch length is considered to be statistically significant.

In bootstrap tests, the same number of nucleotides as the actual number used for constructing the NJ tree are randomly sampled with replacement from the original sequence data, and an NJ tree is produced from this set of resampled nucleotide data. The topology of the tree is then compared with the original NJ tree. Any interior branch of the NJ tree that gives the same partition of sequences as that of the bootstrap tree (see Penny and Hendy 1985, Rzhetsky and Nei 1992 for partition of sequences) is given value 1 (identity value), whereas other interior branches are given 0. This process is repeated several hundred times, and the percentage of times each interior branch of the NJ tree receives identity value 1 is computed. We call this the *bootstrap confidence level (BCL).* Note that this test is different from that included in PHYLIP, where a bootstrap consensus tree is constructed.

The statistical properties of the bootstrap test are complicated and are not well understood (Zharkikh and Li 1992a, b, Felsenstein and Kishino 1993, Hillis and Bull 1993). When the test is applied to an NJ tree, however, the interpretation of the test results is simpler. If (1) every site of the DNA sequence evolves in the same way, (2) the distance measure used is an unbiased estimator of the number of nucleotide substitutions, and (3) the numbers of sequences and nucleotides used are sufficiently large, the null hypothesis of the bootstrap test is that the length of each interior branch is 0, and the *BCL* of a branch approximately measures the probability of the branch length being different from 0 at least when the $BCL > 0.9$. Computer simulations have shown that the *BCL* is indeed very close to the equivalent probability (*CP*) determined by the standard error test mentioned above when $BCL > 0.9$ (T. Sitnikova, unpublished results). Of course, for this interpretation to be correct, the original sequence data should contain a substantial number of nucleotides. If this number is small and happens to be a biased sample from a long sequence that is under investigation, bootstrap resampling will never be able to correct the bias (Nei 1991, Zharkikh and Li 1992a).

The bootstrap test for NJ trees is known to be conservative if an unbiased distance measure is used (M. Nei and S. Kumar in Nei and Rzhetsky 1991). However, if biased distance measures are used, this test may lead to an incorrect conclusion and an incorrect topology may receive a high *BCL* value. Therefore, it is important to use proper distance measures for this test.

5.6.2 UPGMA Trees

Nei *et al.* (1985) developed a method for computing the standard errors of interior branch lengths for a UPGMA tree under the assumption of constant rate of evolution. However, their method is not easy to apply when the number of sequences is large. Furthermore, if the rate of nucleotide substitution is not constant, their test may give an erroneous conclusion. Therefore, we have not included it in MEGA.

Instead, we have included the bootstrap test. In this test, the root of UPGMA trees is eliminated, and both the original and bootstrap trees are treated as unrooted trees.

The procedure of the test is the same as that for NJ trees, and each interior branch of the original UPGMA tree receives the *BCL* value.

If the rate of nucleotide substitution is constant for all evolutionary lineages and the assumptions mentioned for UPGMA are all satisfied, the *BCL* is again expected to give the probability of each interior branch length being different from 0 when *BCL* is high. This is because UPGMA gives unbiased estimates of branch lengths under this condition if the topology obtained is correct (Chakraborty 1977).

In practice, however, the rate of nucleotide substitution is not necessarily constant, and there is increasing evidence that the rate often varies from lineage to lineage. Furthermore, the probability of obtaining the correct topology by UPGMA is generally lower than that by the NJ method even if the rate of nucleotide substitution is constant (Saitou and Nei 1987). In these cases, the bootstrap test may lead to an incorrect conclusion. Particularly when the rate of nucleotide substitution varies with evolutionary lineage, an incorrect branching pattern may receive a high bootstrap value. Therefore, one should be cautious about the bootstrap test of UPGMA trees.

5.6.3 MP Trees

It is very difficult to develop a solid statistical test for MP trees, because the stochastic nature of nucleotide substitution is not taken into account in obtaining these trees. Although it is possible to estimate branch lengths, the estimates are usually biased downward. Therefore, the standard error test cannot be applied.

As mentioned earlier, the MP method may give a reasonably good tree under certain conditions. In this case, it is meaningful to conduct a bootstrap test. Felsenstein (1985) proposed a bootstrap test for an MP tree, but his test is different from ours. While we are interested in testing the accuracy of an MP tree obtained, his test is for examining the accuracy of a bootstrap consensus tree. We have initiated implementation of our test for an MP tree in MEGA but decided not to include it in the present version, because it is still in a preliminary stage. We plan to include it in the next version. Here, we briefly describe the strategy and algorithm of our bootstrap test.

We first note that if the MP method produces a large number of equally parsimonious trees for a given set of sequence data (e.g., Hedges *et al.* 1992), there is no need to conduct a statistical test, because in this case we cannot choose the best tree or a few best trees anyway. Second, if the number of OTUs is so large that we have to use the heuristic search, the bootstrap test is not very meaningful, because we are not sure whether the tree obtained is the MP tree. In MEGA, therefore, we will consider only the case where the number of OTUs is relatively small (say $m < 20$) and the number of equally parsimonious trees is one or a few.

In our approach, we first determine all MP trees for a given set of data using the branch-and-bound method. If there is only one global MP tree, we keep this tree and examine the reliability of the tree by using the bootstrap test. In this case, it is possible that in a particular bootstrap replication two or more equally parsimonious trees will appear. We then compare each of these trees with the global MP tree obtained from the entire data set and determine the identity value 1 or 0 for a given interior branch of the global MP tree compared with each of the bootstrap MP trees. For each interior branch of the global MP tree, the sum of identify values over all bootstrap MP trees is then divided by the number of the latter trees for this particular bootstrap replication. Thus, if all the bootstrap MP trees for a particular bootstrap replication show the same partition of sequences as that of the MP tree for a given interior branch, this branch receives value 1 for this replication. This procedure is repeated for all bootstrap replications. We can then compute the *BCL* values for all interior branches of the global MP tree.

When there are several global MP trees obtained from the entire data set, we construct a strict consensus tree for them and regard it as a single global MP tree. We can then apply the same procedure as that mentioned above and compute the *BCL* values for all interior branches of the global MP tree.

In the case of the MP method, the *BCL* value of an interior branch is unlikely to be equal to the probability that the length of the branch is different from 0 even when the sequence data favorable for the MP method are used (see section 5.5). However, a *BCL* value higher than 95 percent probably gives some confidence of the branching pattern associated with the branch [see Zharkikh and Li (1992a) for the special case of four sequences].

However, when the sequence data do not satisfy the condition required for the MP method and this method gives an inconsistent tree, a bootstrap test may give a false confidence for the tree obtained (Zharkikh and Li 1992b). That is, even an incorrect branching pattern may receive a *BCL* value of 100 percent if the number of nucleotides examined is large. Therefore, the user of this test should always be cautious about the interpretation of the *BCL* values. We suggest that when a tree topology is estimated by the MP method, the branch lengths of the topology should also be estimated by some other method such as the NJ, ME, and maximum likelihood methods. Information on branch lengths will give some idea about the accuracy of the bootstrap test for an MP tree. Note also that if the number of informative sites used is small, bootstrap tests may give erroneous conclusions.

5.6.4 Condensed Trees

When a phylogenetic tree has low *CP* or *BCL* values for several interior branches, it is often useful to produce a multifurcating tree by assuming that all such interior branches have a branch length equal to 0. We call this multifurcating tree a *condensed tree*. In MEGA, this condensed tree can be produced for any level of *CP* or *BCL* value. For example, if there are several branches with *CP* or *BCL* values of less than 50%, a condensed tree with the 50% *CP* or *BCL* level will have a multifurcating tree with all these branch lengths reduced to 0.

Since the branches of low significance are eliminated to form a condensed tree, this tree gives emphasis on reliable portions of branching patterns. However, this tree has one drawback. That is, since some branches are reduced to 0, it is difficult to draw a tree with proper branch lengths for the remaining portion. We have therefore decided to give our attention only to the topology. Thus, the branch lengths of a condensed tree in MEGA are not proportional to the number of nucleotide or amino acid substitutions.

Note that condensed trees are different from consensus trees mentioned earlier, though they may look similar in practice. A consensus tree is produced from many equally parsimonious trees, whereas a condensed tree is merely a simplified version of a tree. A condensed tree can be produced for any type of tree (NJ, ME, UPGMA, MP, or maximum-likelihood tree).

5.6.5 General Comments on Statistical Tests

It should be noted that any statistical test of topological differences or branch lengths depends on a number of assumptions, which are not always satisfied by actual data, and that when the assumptions are not satisfied an incorrect tree may be statistically supported even when the NJ, ME, or the maximum-likelihood method are used (Tateno *et al.* 1994). Therefore, when the pattern of nucleotide or amino acid substitutions for the data set used is complicated, one should be cautious about the interpretation of the results of statistical tests. As a general rule, it is safe not to trust results based on a relatively small number of nucleotides even if every interior branch of an estimated tree is significant at the 95% *CP* or *BCL*. These results should be confirmed by increasing the number of nucleotides if possible.

Particularly for establishing evolutionary relationships of different organisms, it is important to examine a large number of nucleotides from many different genes, because different genes may be subject to different evolutionary forces. Furthermore, if a large number of nucleotides are used, there is no need to use a sophisticated and time-consuming tree-building method. A simple method like the NJ method usually gives the same tree as that obtained by time-consuming statistical methods.

6

User-Interface

This chapter gives the description of various building blocks of the user-interface and phylogenetic-tree editor, sequence data presentation, context-sensitive help, and multiple-file editor and browser. In the following discussion we refer to many standard special keys, such as function keys **F1** to **F10** and **Alt**, **Esc**, **Tab**, **Shift**, **Ctrl**, and **Enter** keys. If you are not familiar with these keys, please locate them on your keyboard before proceeding further.

6.1 Screen

The computer monitor is the screen. It has three components in MEGA (Fig. 6.1) the menubar at the top, the status line on the bottom, and the desktop window in the middle. A clock follows the menubar on the top-right corner and the heap view that displays the amount of computer memory available for analysis is displayed directly below this clock. The rest of the bottom line to the left of the heap view is the status line.

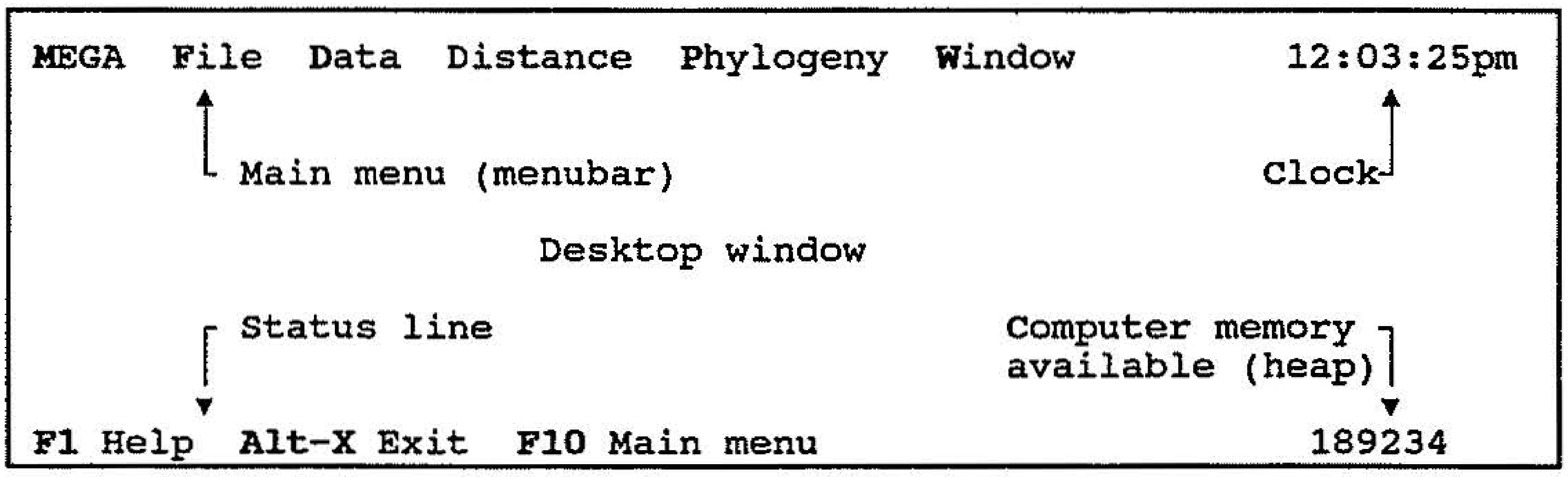

MEGA File Data Distance Phylogeny Window
12:03:25pm
Main menu (menubar)
Clock
Desktop window
Status line
Computer memory available (heap)
F1 Help Alt-X Exit F10 Main menu
189234

Fig. 6.1 General desktop pattern.

6.1.1 Menubar, Desktop, and Status Line

The main menu (menubar) appears at the top of the screen and has six pull-down menus. Every menu contains a list of commands (Fig. 6.2). If a command is followed by an ellipsis (...), its selection displays a dialog box. If an arrow (▸) appears after a command, then a submenu is displayed. Commands without any ellipsis (...) or arrow (▸) indicate that some action will follow subsequent to their selection. The menus and the commands can be chosen with the keyboard as well as with the mouse.

Keyboard Function key **F10** activates the menubar. A sliding bar that moves back and forth with arrow keys will appear. A menu with highlighted bar can be selected by pressing **Enter**. Pressing **Alt+highlighted letter** of the menu also shows the corresponding menu.

Mouse The left mouse button is for clicking and dragging, unless the right mouse button is chosen from the mouse control panel. Click on the menubar to display the desired menu.

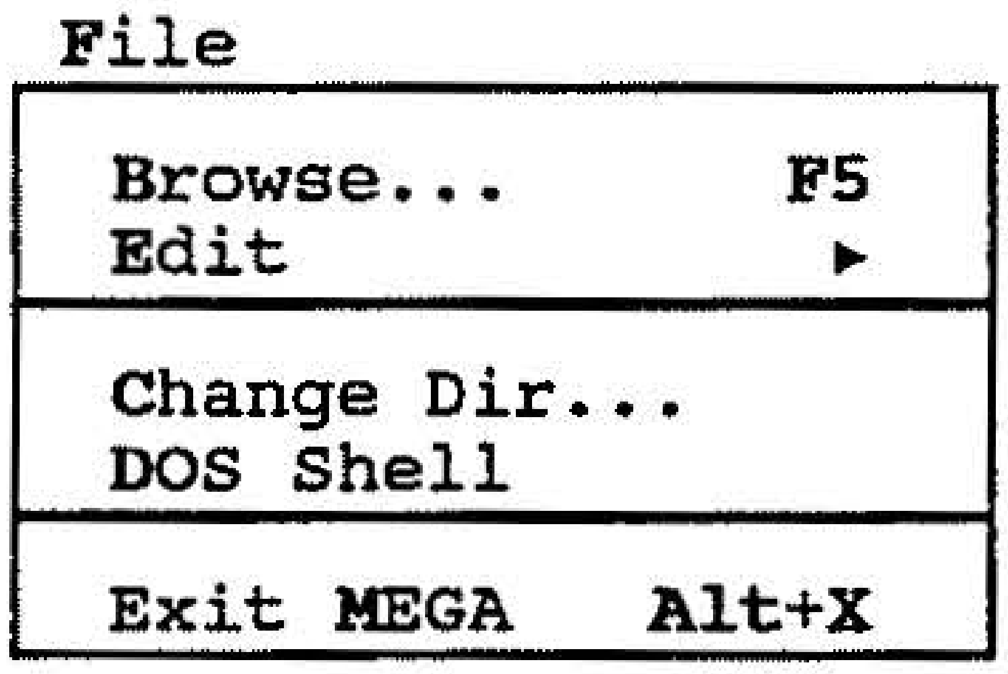


Fig. 6.2 A typical pull-down menu.

Enabled and Disabled Commands Enabled commands are displayed in bright shade and color, and one of the character from the command name is highlighted. Disabled commands are not selectable and appear in light shade of gray. Highlighting and gray shading may not be visible on old monochrome monitors, and the only way to know that the command is enabled is to try and select it.

The region between the menubar and the status line is the desktop. Desktop is a window without any borders. All the windows (file-editor, file-browser, data presentation, etc.) are displayed on this desktop.

A dynamic status line is given at the bottom of the screen. It provides hints, short-cuts, and additional options. Short-cut commands are activated either by clicking with a mouse or by pressing the highlighted letter of the command name. One line hint on menu commands and dialog box items is also available on the status line.

6.1.2 Hot-Keys and Short-Cuts

Many frequently used commands can be activated by pressing hot-keys (e.g., hot-key for help is **F1**). These hot-keys produce an action only if the corresponding command is enabled. Some hot-keys are shown with a '+' sign between them to indicate that these keys should be pressed simultaneously.

Table 6.1 General hot-keys

Key(s)	Menu item	Function
F1	(none)	Display context-sensitive help.
Alt+F1	*MEGA|Using Help*	How to use help.
F2	*File|Edit|Save*	Save current file.
Alt+F2	*File|Edit|Save As*	Save current file as another file.
F3	*File|Open File*	Open a file for editing.
Alt+F3	*Windows|Close*	Close current file and quits.
F4	*Data|Data Presentation*	Display sequence data.
Alt+F4	*Data|Close Data*	Close active data file.
F5	*File|Browse*	Open text-file for viewing only.
F6	*Window|Next*	Go to next window.
Alt+F6	*Window|Previous*	Go to previous window.
F7	*Distance|Compute Distances*	Compute pairwise distances.
Alt+F7	*Phylogeny|Std. Error Test*	Conduct branch length test.
F8	*Phylogeny|Construct Tree(s)*	Construct phylogenetic tree(s).
Alt+F8	*Phylogeny|Bootstrap Test*	Conduct bootstrap test on phylogenies.
F9	*Window|Zoom*	Expand the current window.
Alt+F9	*Window|Re-size/Move*	Re-size or move the window.
F10	(none)	Go to menubar.

Table 6.2 Main menu hot-keys

Key(s)	Item	Function
Alt+M	MEGA menu	Takes you to MEGA menu.
Alt+F	File menu	Takes you to File menu.
Alt+D	Data menu	Takes you to Data menu.
Alt+T	Distance menu	Takes you to Distance menu.
Alt+P	Phylogeny menu	Takes you to Phylogeny menu.
Alt+W	Window menu	Takes you to Window menu.
Alt+X	*File|Exit MEGA*	Exit MEGA to end the session.

Table 6.3 Windows hot-keys

Key(s)	Menu item	Function
Alt+F3	*Window|Close*	Close the currently active window.
F6	*Window|Next*	Go to next window.
Alt+F6	*Window|Previous*	Go to previous window.
F9	*Window|Zoom*	Expand the current window.
Alt+F9	*Window|Re-size/Move*	Re-size and move the current window.

6.2 Windows and Dialog Boxes

A window is a rectangular area on the screen that may have a frame. Some windows can be opened, closed, re-sized, or overlapped (e.g., file editor window; Fig. 6.3), whereas others are frameless and cannot be re-sized (e.g., dialog boxes).

If many overlapping windows are opened at the same time, the active window is on the top with a double line border. On the top border of this frame exist a close box icon on the left, a title bar in the center, and a zoom icon on the right corner. The re-size corner is located at the bottom-right corner at the intersection of horizontal and vertical scroll bars.

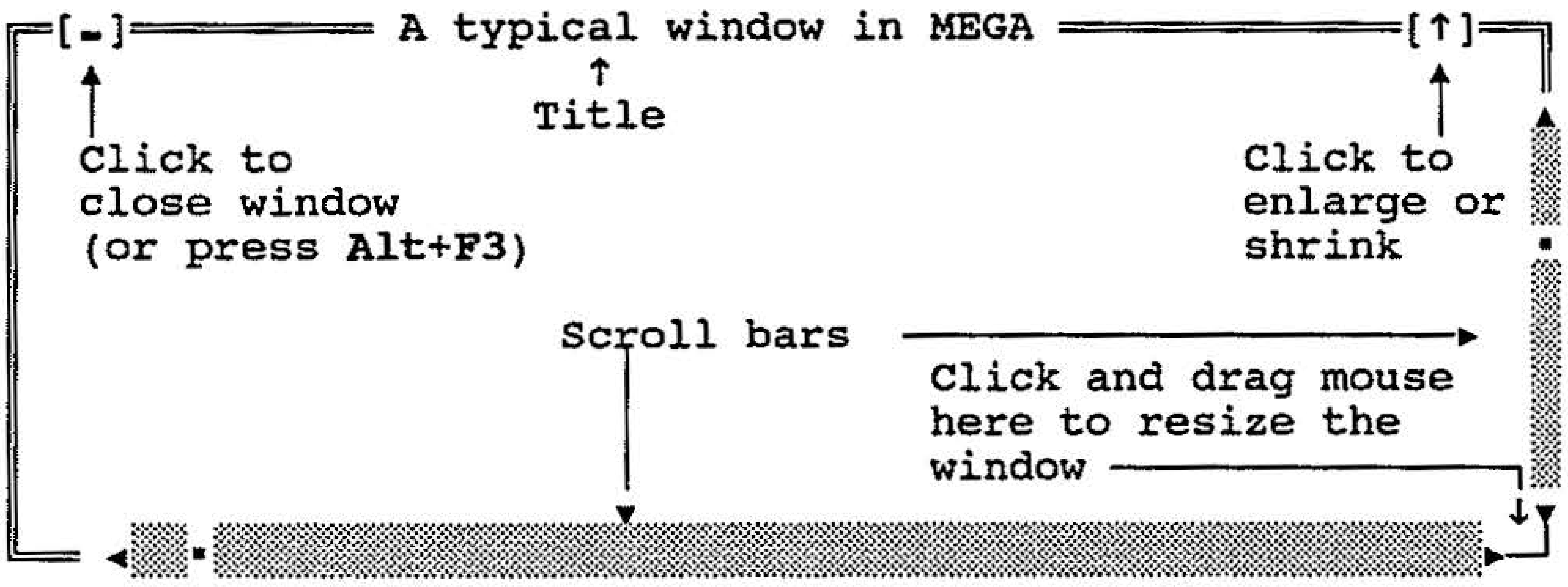
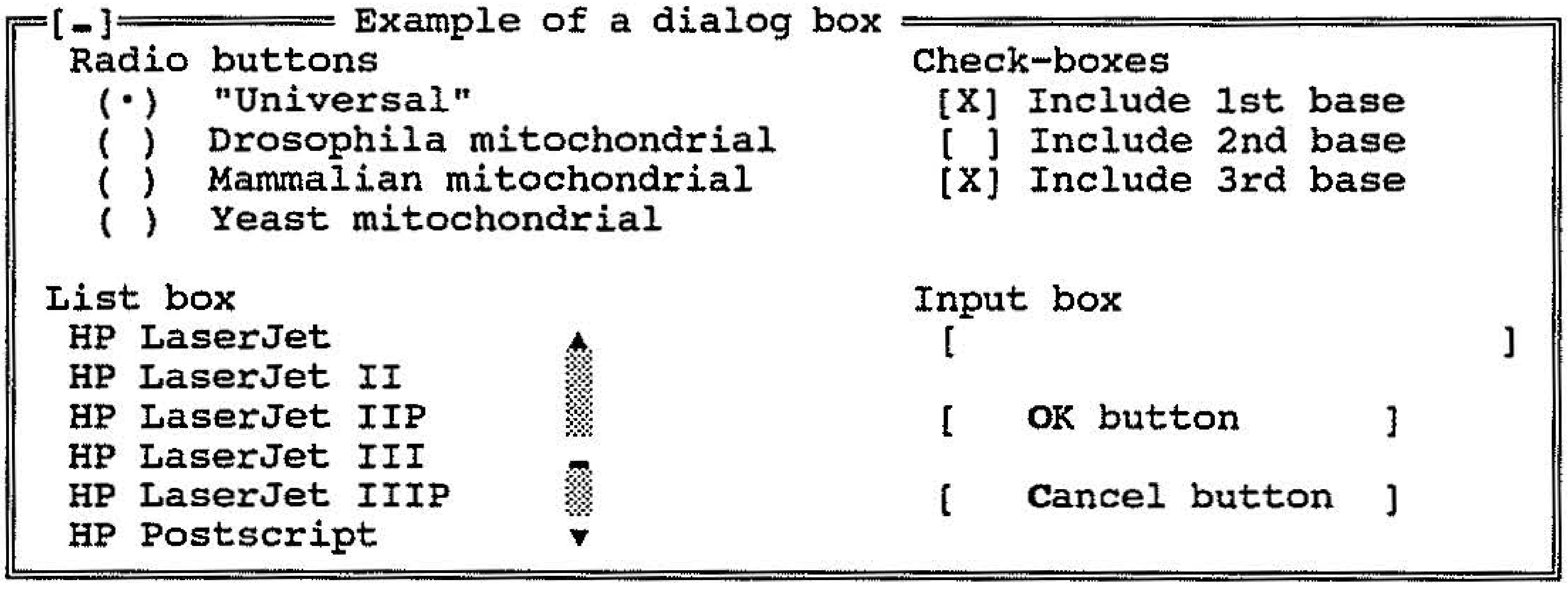

[■]
A typical window in MEGA
[↑]
Title
Click to close window (or press Alt+F3)
Click to enlarge or shrink
Scroll bars
Click and drag mouse here to resize the window

Fig. 6.3 A typical window.

Dialog boxes are used to select various options at different stages of analysis. Up to five kinds of elements may comprise a dialog box: action- and radio-buttons; and check-, input-, and list-boxes. Dialog boxes cannot be re-sized, but they can be moved around on the screen. In a dialog box, the **Esc** key aborts the analysis, and the **Enter** key is pressed to accept all displayed options and initiate subsequent actions. Key that provide movement between different groups and within groups themselves are listed in Table 6.4.

Table 6.4 Dialog box keys

Key(s)	Description
↑,↓,→,←	Move within options in a group.
Tab	Cycle through the groups in clockwise direction.
Shift+Tab	Cycle through groups counter-clockwise.
Enter	Select all the setting and proceed.
Esc	Cancel the box (no action follows).
Spacebar	Select/un-select a check box.
Alt+Letter	Select the button with that highlighted letter.



```
[■]══════ Example of a dialog box ══════════════════
 Radio buttons                          Check-boxes
  (•)  "Universal"                       [X] Include 1st base
  ( )  Drosophila mitochondrial          [ ] Include 2nd base
  ( )  Mammalian mitochondrial           [X] Include 3rd base
  ( )  Yeast mitochondrial

 List box                               Input box
  HP LaserJet              ▲             [                      ]
  HP LaserJet II
  HP LaserJet IIP                        [   OK button      ]
  HP LaserJet III
  HP LaserJet IIIP                       [   Cancel button  ]
  HP Postscript            ▼
```

Fig. 6.4 A dialog box.

Action-buttons

In Fig. 6.4, **OK** and **Cancel** are the action-buttons. In a dialog box, you may find many such buttons (e.g., the **Preview** button in the *Print Tree* dialog box). Using these action buttons, you issue commands. In every dialog box, there is a default action button that is tied to the **Enter** key. Whenever you press **Enter**, the action associated with the default button will occur. Usually the **OK** button is a default button. The **OK** button is used to accept all options selected in the dialog box. You can also issue the *OK* command by either pressing **Alt+O** or clicking on it. Clicking on **Cancel** will abort any further action. The **Enter** and **Esc** keys are short-cuts for the *OK* and *Cancel* commands.

Radio-buttons

Radio-buttons have round parentheses ( ) before them. Presence of a dot in the parentheses, (•), indicates that the radio-button is selected. Only one radio-button can be selected at any time. Related radio-buttons are clustered in a group and are listed under a common label. The **Up** (↑) and **Down** (↓) arrow keys provide movement within groups of radio buttons. You may also click on a radio-button to select it. (If you have a monochrome monitor, the focussed check-box is indicated by the chevron symbol, ».)

Check-boxes

Check-boxes have square brackets [ ] before them. An X in the square brackets, [X], indicates that the check-box is selected, and many check-boxes may be selected at one time. Many related check-boxes are clustered in a group and are listed under a common label. The **Up** (↑) and **Down** (↓) arrow keys provide movement within such groups. (If you have a monochrome monitor, the focussed check-box is indicated by the chevron symbol, ».) To unselect a check box, click with the mouse or press **Spacebar** on it.

Input-boxes

An input box allows the user to type-in text. Most of the editing keys (**Home**, **End**, ↑, ↓, →, ←, **Delete**, etc.) work in these input boxes. If the line displayed on the screen is not long enough to show all the text, the arrowheads (► and ◄) appear at the end of the line. Some input boxes have a down arrow icon [↓] at their right-end indicating an associated history list. The input file name box is one such box. Pressing down arrow key (↓) in the input box opens the history list with all the text strings that were entered at the input line before. Use arrow keys to move in this list and select any line of text by pressing the **Enter** key. Press **Esc** to come out of the history list without making any selection.

List-boxes

In a list box, a list of variable-length strings is displayed that can be scrolled by using arrow keys and selected by pressing **Enter**. List boxes may contain lists in more than one columns (e.g., file list box in Input file name dialog box).

6.3 File Name Dialog Box

Files to be edited, browsed, or analyzed are specified in the *File Name* dialog box. It contains an input box, a file list, a file information panel, *Cancel* and *Open* buttons, and a history list attached to the *Name* input box. Movement between different elements in this dialog box is provided with **Tab** and **Shift+Tab** keys.

The name of the file to be opened or loaded (or the file-name mask to be used as a filter for the files list box, for example, *.*) is entered in the *Name* input box. A history list, [↓], is attached to this box that retains all the file names typed in the *Name* box before. The *File list* box shows the names of all the files in current directory that match the file-name mask in the *Name* input box. Present below the list box is a file information panel that displays the path name, file name, date, time, and size of the selected file.

The Open button selects the current file for use; Cancel rejects the current file but does not aborts the current operation. **Esc** cancels the dialog box.

6.4 Context-Sensitive Help Box

The on-line context-sensitive help is invoked with the **F1** key. It brings up a *Help* dialog box (Fig. 6.5) that instantly displays the relevant information about the current command or option.

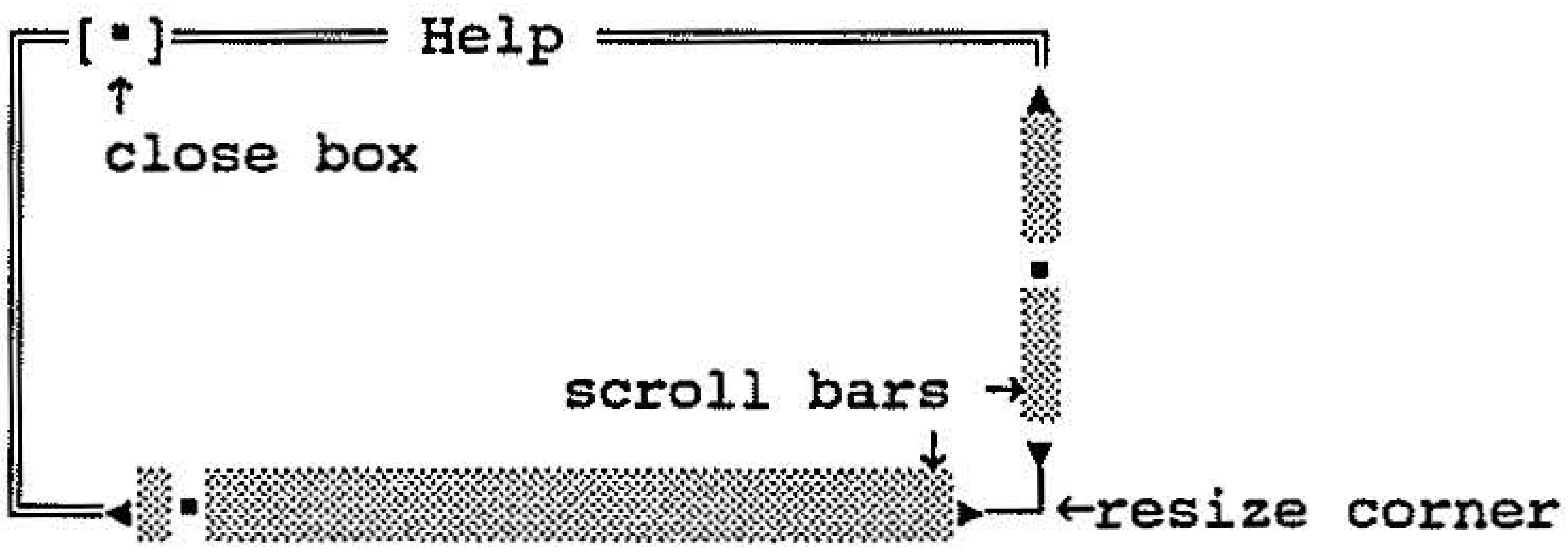


Fig. 6.5 Context-sensitive help box.

In this help window, two kinds of information are present:

Help text Help for the current item is written like a text file. Simple cursor movement keys help navigation in the window. Vertical and horizontal scroll bars can be used if the mouse is installed on the system.

Additional help Some words are highlighted in the help window indicating that further help is available. This cross-reference help is selected by pressing the **Tab** key to get to the desired highlighted word and then by pressing **Enter**. (If you're using a mouse, click on the highlighted word to retrieve further information.)

Table 6.5 Help box keys

Key(s)	Description
↑/↓,→/←,PgUp/PgDn	Movement in the box.
Tab	Cycle through cross-reference help clockwise.
Shift+Tab	Cycle through cross-reference help counter-clockwise.
Enter	Open cross-reference help.
Esc	Exit help.

6.5 Text-File Browser

Selection of the file browsing option (with **F5** or *File|Browse* command) brings up a *File Name* box that prompts for the file to be opened. The selected file is displayed in the read-only mode. Any number of files can be opened at the same time for browsing depending on the memory size of the computer. Maximum characters per line in the file browser is 1023.

Key(s)	Response
F5	Open file for browsing.
↑/↓	Move up/down one line.
→/←	Move right/left one column.
PgDn/PgUp	Move one page down/up.
Home	View left most column.
End	View right most column.
Ctrl+PgUp	View top of file.
Ctrl+PgDn	View bottom of file.
Alt+F3	Close the current file window.

6.6 Text-File Editor

A simple text-editor is included in MEGA. Many files can be edited simultaneously in this editor. It is provided for revising output files and editing small data files (up to **32KB**). In this editor, the functions for saving edited files, transferring and copying blocks, and finding and replacing text strings are available. Editor shows only upto 256 characters in any line.

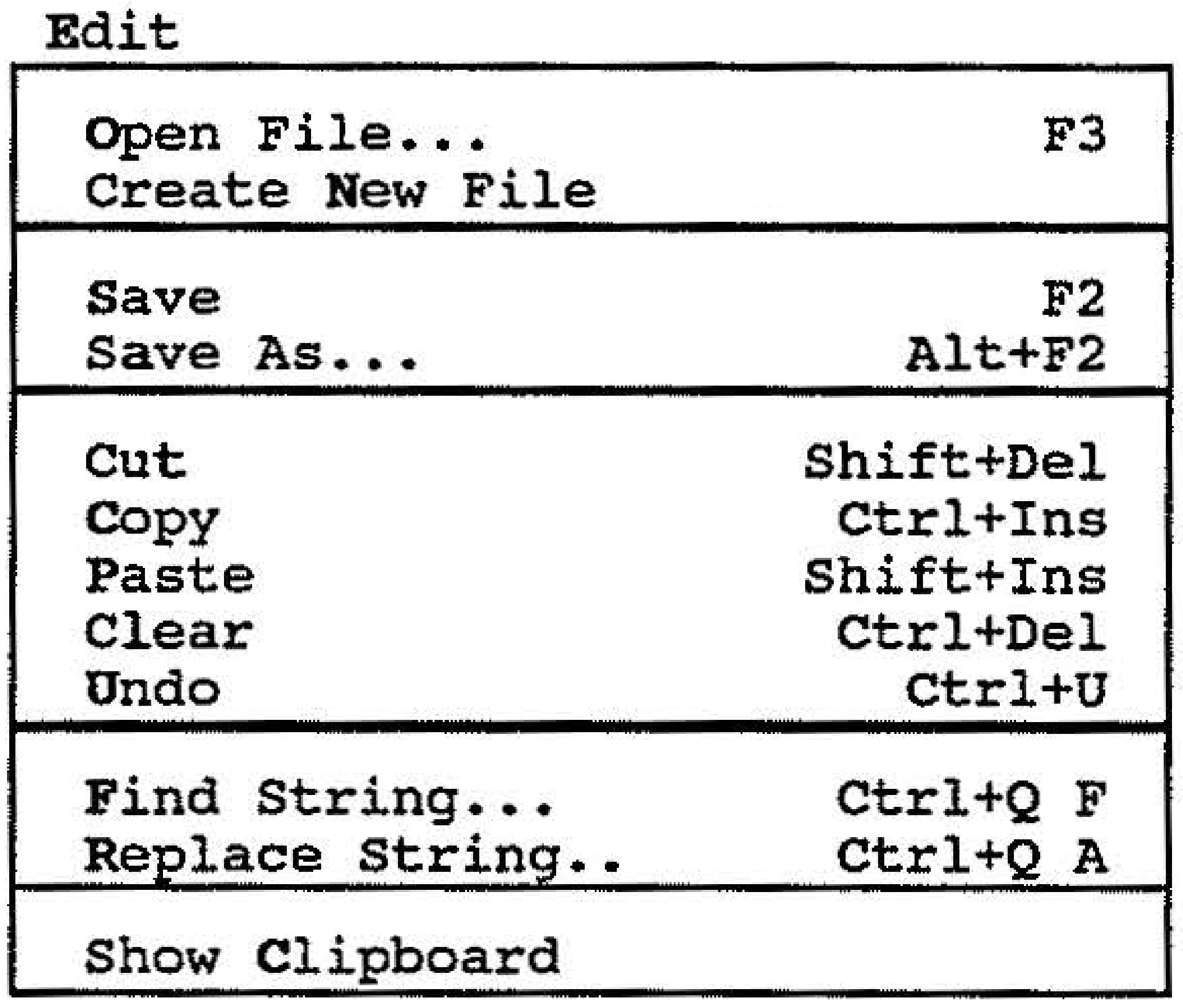


Fig. 6.6 Editor options menu.

File commands An existing file is opened for editing with the *File|Edit|Open File* command (**F3**). The *File Name* input box inquires about the file name. (New files can be created with the *File|Edit|New File* command.) Once a file is opened, its contents can be edited and then saved with the *Save* and *Save As* commands from the *File|Edit* menu. Files can be closed either with **Alt+F3** or by clicking on the upper-left close icon of its window.

Key(s)	Command
F3	Open an existing file to edit.
Alt+F3	Close and quit edited file.
F2	Save edited file, do not exit.
Alt+F2	Save file as different file, (do not exit).

Cursor commands

Standard cursor movement keys such as ↑/↓, →/←, **PgUp/PgDn**, and **Home/End**, have normal meaning in this editor. The **Ctrl** key, when used in conjunction with any of these keys, accelerates the movement of the cursor in the file. For example, **Ctrl+PgDn** travels to the end of the file, whereas **PgDn** travels ahead by one page only. Similarly, the character right key, →, when coupled with the **Ctrl** key moves to the next word instead of the next character.

Cursor movement keys

Key(s)	Command
←/→	Character left/right.
Ctrl+←/→	Word left/right.
↑/↓	Line up/down.
Home	Beginning of the line.
End	End of line.
PgUp/PgDn	Previous/next page.
Ctrl+PgUp	Top of the file.
Ctrl+PgDn	End of file.

Text editing commands

Text is entered in the file in two modes: insert and typeover. The **Ins** key alternates this mode. The cursor is blocked (█) in the typeover mode, whereas in the insert mode it is more like an underscore (_). A list of commands for deleting characters, words, and lines is given below.

Key(s)	Command
Del	Delete current character.
Backspace	Delete previous character.
Ctrl+Y	Delete current line.
Ctrl+Q Y	Delete to the end of the line.
Ctrl+T	Delete current word.
Ins	Insert mode on/off.

Block commands

A block of text is any continuous text, from a single character to hundreds of lines, that is selected (highlighted) on the screen. At any time, only one block may be selected in a file. Blocks may be marked with the keyboard or with the mouse.

Selecting text by using keyboard:

Hold the shift key down (keep pressed), and press one of the keys that moves the cursor; the text starts becoming highlighted.

Selecting text by using mouse:

Click and hold the mouse button at the place of the origin of the text to be marked and drag it to the end of the text to be selected holding the mouse down.

Mark block commands

Key(s)	Command
Shift+←/→	Left/right one character.
Shift+↑/↓	Same column on previous/next line.
Shift+End	End of line.
Shift+Home	Beginning of line.
Shift+PgUp/PgDn	One page up/down.
Shift+Ctrl+←/→	Left/right one word.
Shift+Ctrl+PgDn	End of file.
Shift+Ctrl+PgUp	Beginning of file.

Block-text manipulation

As soon as some text is marked, the block manipulation commands in the *File|Edit* menu become available. Marked blocks of text can be deleted, moved, and copied to the same file or to other opened files for editing.

All block transfers use a special window called the clipboard. In fact, the clipboard is an edit window that is hidden from the user. The contents of the clipboard can be examined with the *File|Edit|Show Clipboard* command. For any block-transfer operation, the text-block is first stored in the clipboard. For example, whenever any text-block is to be copied, it is first marked as a block and then copied to the clipboard with command *File|Edit|Copy*. This text may be pasted (retrieved) in any file with the *File|Edit|Paste* command. There are short-cuts for these operations that are explained below.

Copy a block

Ctrl+Ins, and then **Shift+Ins**

Ctrl+Ins copies the selected block to the clipboard. Position the cursor where you want to insert the text and then press **Shift+Ins** to paste the text-block there.

Copy text

Ctrl+Ins

Copies the selected text to the clipboard.

Cut text **Shift+Del**

Copies the selected text to the clipboard and deletes it from the file.

Clear block **Ctrl+Del**

Deletes a block from the file. It can not be recovered.

Move a block **Shift+Del**, and then **Shift+Ins**

Shift+Del copies the selected text to the clipboard and removes it from the current position. Position the cursor where you want the text to be moved and press **Shift+Ins** to copy the block from the clipboard to that position.

Paste from clipboard **Shift+Ins**

It copies the contents of the clipboard to the current cursor position.

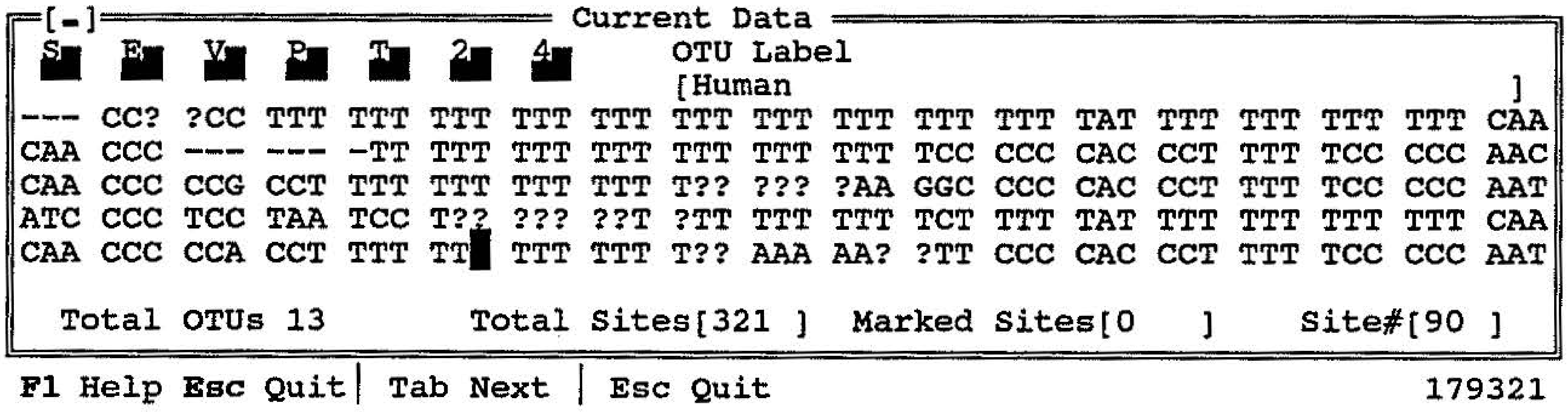
Text Search Commands The text-search and text-search-and-replace commands are used for searching and replacing patterns of characters. These commands search for the desired string of characters, with case-sensitive and whole-word-only options, from the current position of the cursor till its first occurrence. In the text-search-and-replace command, *Replace String*, the string to be searched and the replacement string can be specified. All occurrences of the search string may be replaced with/without confirmation.

Key(s)	**Commands**
Ctrl+Q F	Find a text string.
Ctrl+Q A	Find and replace a text string.

6.7 Sequence Data Presentation

Selection of the *Data|Data Presentation* command displays the currently used sequence data on the screen in the "Current Data" window. In this window, nucleotide sequences can be translated into amino acid sequences, and both can be displayed on the screen. They can be written in files for PAUP (Swofford, 1993), PHYLIP (Felsenstein, 1993), and other formats. The variable, parsimony-informative, and two- and fourfold redundant sites can be highlighted on the screen. In addition, this window contains a command for computation of various statistical quantities for molecular data (see chapter 3).

The 'Current Data' window (Fig. 6.7) contains three elements: a list of command buttons on the upper-left corner, the sequence data in the middle, and four dynamic views (*OTU Label*, *Site#*, *Total Sites*, and *Marked Sites*) that show the current position of the cursor and display other important attributes. The command buttons at the top of the window provide all the options (Table 6.7). These buttons are toggles and their effects are reversible. Inapplicable buttons at any stage are automatically disabled.



```
[-]================== Current Data ========================================
 S   E   V   P   T   2   4          OTU Label
                                    [Human                                      ]
--- CC? ?CC TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TAT TTT TTT TTT TTT CAA
CAA CCC --- --- -TT TTT TTT TTT TTT TTT TTT TCC CCC CAC CCT TTT TCC CCC AAC
CAA CCC CCG CCT TTT TTT TTT TTT T?? ??? ?AA GGC CCC CAC CCT TTT TCC CCC AAT
ATC CCC TCC TAA TCC T?? ??? ??T ?TT TTT TTT TCT TTT TAT TTT TTT TTT TTT CAA
CAA CCC CCA CCT TTT TT█ TTT TTT T?? AAA AA? ?TT CCC CAC CCT TTT TCC CCC AAT

  Total OTUs 13       Total Sites[321 ]  Marked Sites[0    ]    Site#[90 ]

F1 Help Esc Quit| Tab Next | Esc Quit                                  179321
```

Fig 6.7 Sequence Data Presentation window

Table 6.6 Cursor movement keys

Key(s)	Description
↑/↓	Go to previous/next sequence.
←/→	Go to previous/next site.
Ctrl+←/→	Go to previous/next cluster of sites.
PgUp/PgDn	Go to previous/next screen of sequences.
Home	Go to top left corner of the screen.
Home,Home	Go to first site in the first sequence.
End	Go to bottom right corner of screen.
End,End	Go to the last site in the last sequence.
Tab	Cycle forward in the data display window.
Shift+Tab	Cycle backward in the data display window.
Esc	Quit data display.

Table 6.7 Data display commands

Key	Command	Description
V	Variable	Highlight all variable sites. The number of variable sites is shown in the *Marked Sites* display.
P	Parsimony-informative	Highlight all parsimony-informative sites. The number of these sites is shown in the *Marked Sites* view.
T	Translate	Translate protein-coding nucleotide sequences. Presence of '*' indicates a stop codon. The number of amino acids is shown in the *Total Sites* view. Make sure that the correct genetic code table is used.
2	Twofold	Highlight all common twofold redundant sites in protein-coding nucleotide sequences. Their number is displayed in the *Marked*

		Sites view. Make sure that the appropriate genetic code table is used.
4	Fourfold	Highlight all common fourfold redundant sites in protein-coding nucleotide sequences. Their number is displayed in the *Marked Sites* view. Make sure that the appropriate genetic code table is used.
S	Statistics	Compute various statistical quantities for the sequences (see chapter 3).
E	Export	Write sequence data (and its subsets) to files in MEGA, NEXUS (PAUP, etc.), PHYLIP, and publication formats. A dialog box prompts for various options that determine the layout of the output data file. For information on individual options, press **F1** in the dialog box.

Table 6.8 Cursor navigation keys

Keys(s)	**View mode**	**Edit mode**
↑/↓	Go up/down one row.	Go to upper/lower descendant.
←	Go left one column.	Go to the ancestor.
→	Go right one column.	Go to lower descendant.
Shift+←	Go left fast.	Scroll screen to left.
Shift+→	Go right fast.	Scroll screen to right.
Shift+↑	Go to previous page.	Scroll screen up.
Shift+↓	Go to next page.	Scroll screen down.
PgUp	Go to previous page.	Go to sister branch.
PgDn	Go to next page.	Go to sister branch.
Home	Go to upper left corner of the tree.	Go to the left most branch on the tree.
End	Go to lower right corner of the tree.	
Esc	Quit tree display.	Exit tree display.
F1	Help on view mode.	Help on edit mode.

6.8 Phylogenetic-Tree Editor

The phylogenetic-tree editor displays the trees constructed using the *Construct Tree(s), Bootstrap Test*, and *Standard Error Test* commands from the *Phylogeny* menu. It facilitates root relocation, tree re-sizing, and branch flipping and swapping on the screen. The edited tree can be stored in text- or graphics-files, printed as graphic images on a wide range of printers, and previewed in the graphic mode on the screen with the

EGA, VGA, and Hercules graphics adapters.

In this editor, the tree is displayed on the screen with approximate branch lengths because computer screens are only 80 columns wide. The accuracy of branch lengths displayed can be improved by re-sizing the tree with the *Expand* command. An expanded tree is spread beyond the margins of the screen.

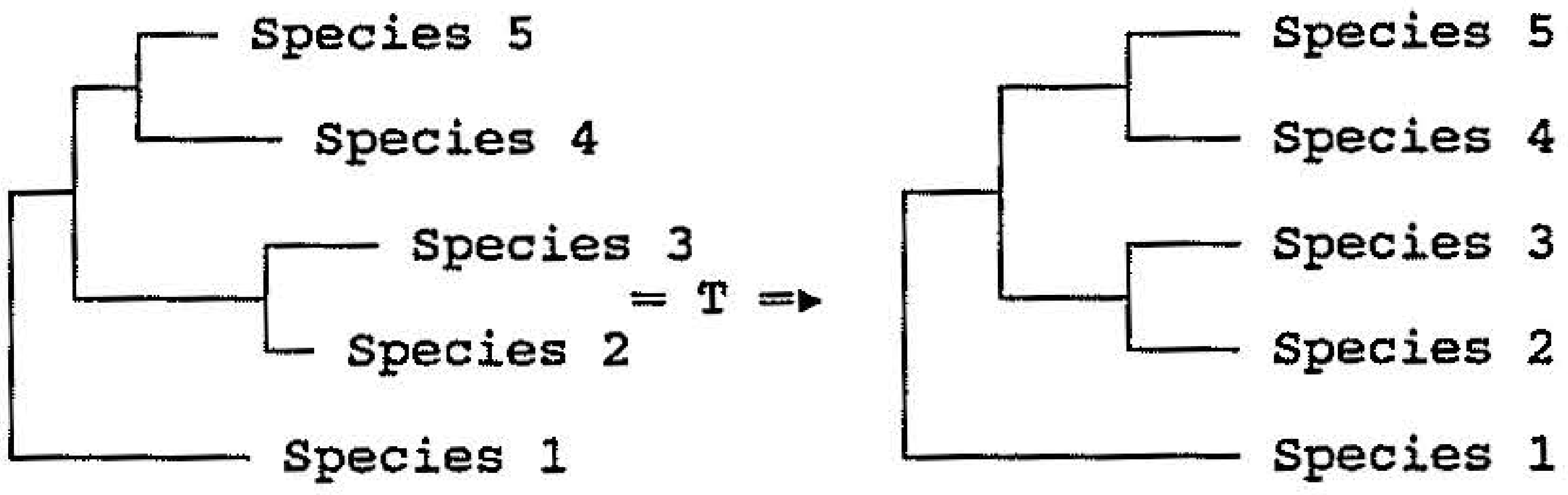
By default, a tree is displayed in the view mode where the tree is presented like a text-file. You can invoke the tree-editing mode by pressing the **E** key. A blocked cursor will appear at the left side of the screen. This cursor can be moved with cursor movement keys (Table 6.8). The branch where the cursor is resting is referred to as the focused branch and tree-editing commands can be used. In the edit mode branches can be swapped and flipped, and the tree can be re-rooted. The *cut* and *paste* operations are not allowed in this editor because they alter the reconstructed branching pattern.

Table 6.9 Commands and effects.

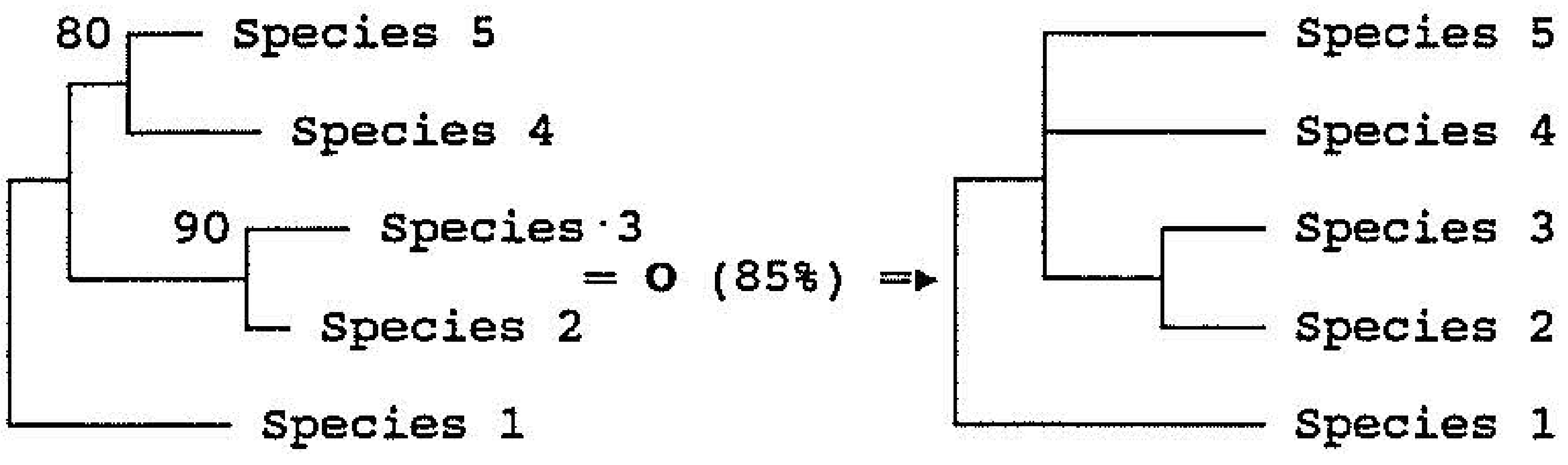
Command	Key	Description
Edit	E	Change to edit/view mode.
Print	P	Print tree.
Topology	T	Display branching pattern only.
Cut-Off	O	Use a cut-off level and display branching pattern only.
Contract	C/-	Reduces the size of an expanded tree.
Expand	X/+	Magnifies the tree.
Mirror	M	Displays mirror image of the complete tree.
Root	R	Roots the tree on focused branch.
Flip	F	Display mirror image of a sub-tree.
Swap	S	Swap the descendent branches.

Topology

By default, MEGA displays a tree using the branch length estimates, if available. Using the *Topology* command, you can display this tree without any branch lengths. To return back to the original tree, press 'T' again.



Cut-Off The *Cut-off* command is available whenever branch length estimates and additional information such as BCL or CP are available for the displayed NJ or UPGMA tree. In this case, the *cut-off* command can be used to collapse branches that have BCL (or CP) values smaller than the specified cut-off level. Only the branching pattern of the collapsed or condensed tree is displayed on the screen. Note that this condensed tree is not a consensus tree. To return back to the original tree, press **O** again.

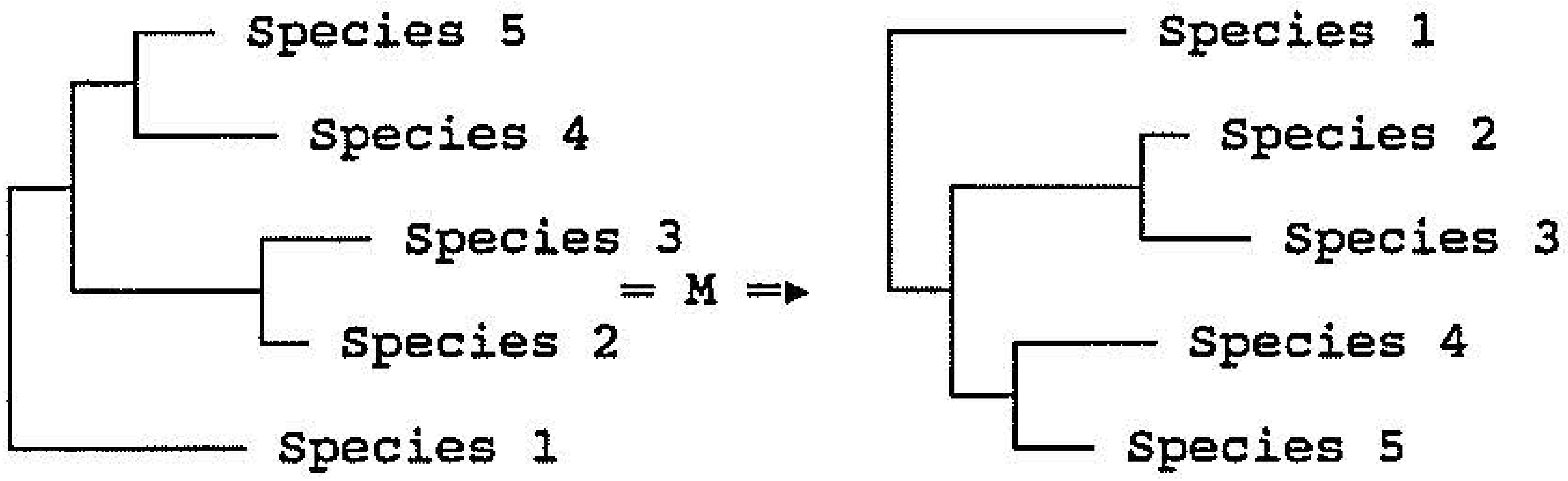


Original tree 85% condensed tree

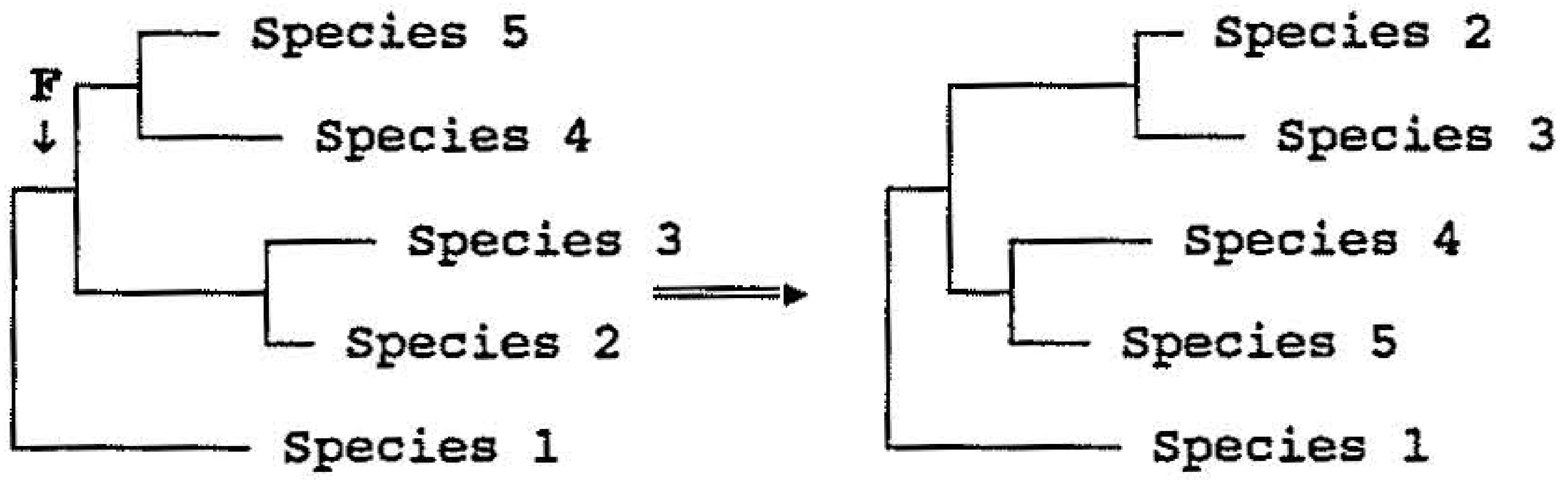
The *Cut-off* command is also available if two or more MP trees are produced. In this case, this command provides a way of constructing consensus trees. To begin with, a majority-rule (>50% frequency) consensus tree is displayed. Using the *cut-off* command, you can construct majority-rule consensus trees at any level >50%.

Root Trees can be re-rooted for neighbor-joining and maximum parsimony trees only. To relocate the root, position the blocked cursor with the cursor navigation keys on the desired branch and press '**R**'. The tree is re-rooted instantly. If the cursor is positioned on a branch that immediately follows the root, its branch length is reduced by half, and the length of its sister branch increases appropriately.

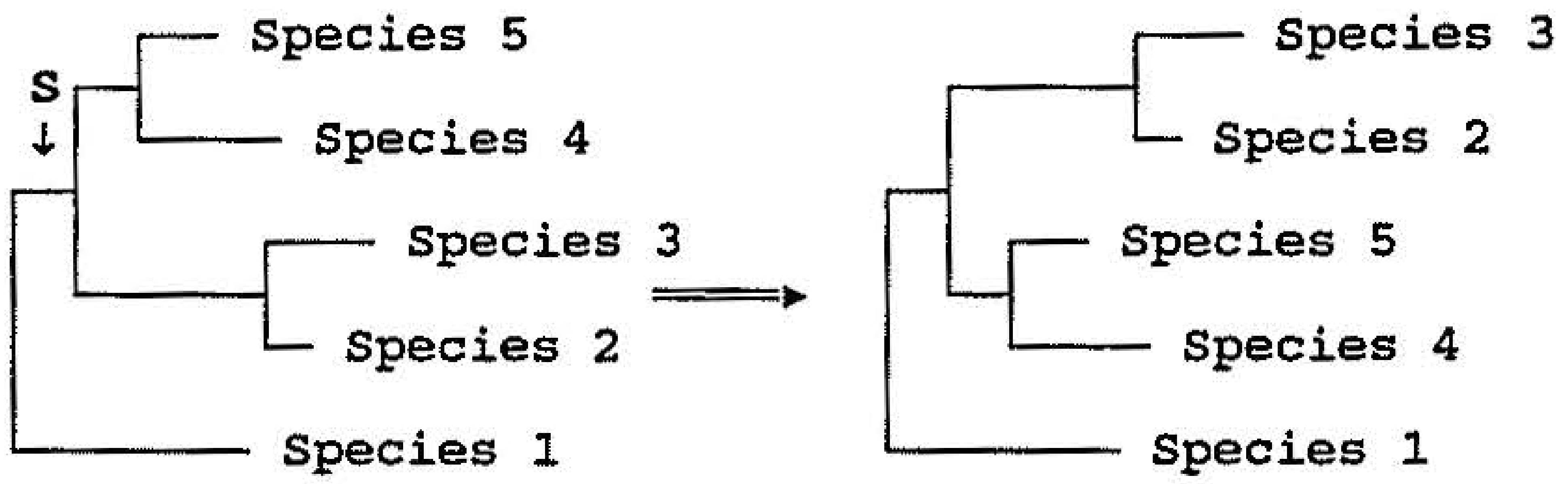
Mirror It has a mirror like effect on the tree. The tree is drawn such that the OTUs previously on the top goes to the bottom, and the one at the bottom jumps to the top.



Flip This command has a mirror like effect on part of the tree, i.e., the right-hand side of the focused branch.



Swap It swaps the position of two descendant branches of the focused branch. The structures of the sub-trees hinged on the descendant branches remain unaffected.



6.9 Printing Trees

MEGA provides the printing of phylogenetic trees in graphics and text formats from the phylogenetic tree editor. These trees can be printed on various dot matrix, laser, PaintJet, and PostScript printers and saved in graphics and text files. The trees can be previewed on the screen before printing if a graphics adapter is available. The options available in the tree printing function are described below.

Output devices Printers, files, and the computer monitors are all output devices. MEGA allows you to output the tree to any of these devices in graphics and text formats.

Printers

MEGA supports most of the popular printers. A list of these printers is provided in the **Appendix D**. Even if your printer is not given in this list, it may still be supported. Most printers emulate one of the industry standard printers. Some printers require special DIP switch settings in order to provide emulation, while others automatically perform emulation. Refer to you printer documentation for specific emulation instructions.

Text Files

In this option the phylogenetic tree is written as an ASCII text-file. The branch lengths on this tree are only approximate, and this tree is a true copy of the tree shown on the screen. This text file can be imported in word processors such as *WordPerfect* for printing and modification.

Graphics Screen

The graphic image of the tree shown on the screen in the text mode can be visualized with the *Preview* command. This command works only when you have a graphics adaptor.

Size The size of the output tree can be altered by adjusting the size of the page, the number of OTUs per page, and the orientation of the page.

Page size

Page size specifies the portion of the page to use for printing. The valid options are Full, Half, and Quarter page.

Number of OTUs per page

This command provides a way to print trees in multiple pages. You cannot print less than 5 OTUs per page. Example: If your tree contains 10 OTUs and you ask for 20 OTUs per page, this command will produce a tree in half a page.

Orientation

Page orientation specifies the layout of the page as Landscape or Portrait. By default, the Portrait orientation is chosen.

Style You can select fonts, printing resolution, etc. to modify your printouts.

Fonts

One of the three fonts can be selected for printing. These fonts are Small, Simplex, and Script.

Resolution

The graphics resolution depends on the printer that you use. MEGA provides three resolutions: low, medium, and high. Not all printers have all these resolutions. If you specify a resolution that exceeds your printer's capability, printing will be done using the highest resolution available. You should refrain from using the *high* option because they print rather slowly.

OTU Alignment

With this command, you can place all OTU labels on the right-hand side of the tree. By default, all the labels are written immediately after the corresponding exterior branch.

Scale bar

With this command a scale bar can be written at the end of the printed tree.

Writing information

Using these options, you can print branch lengths, *BCL*, or *CP* values on the tree. If the selected option is not applicable, it is ignored at the time of printing.

Legend

With this option, a legend can be appended to the tree printed. The legend cannot be longer than 80 characters. On the printed tree, the legend will be truncated if the font size is not small enough to accommodate all the 80 characters.

7

Walk through MEGA

This chapter provides a tutorial for using MEGA through 6 examples. The data files for these examples can be found in the C:\MEGA\EXAMPLES directory. In these example files data are deliberately written in different input formats. We recommend that these examples be studied in the order presented because the techniques introduced in previous examples are used in the following ones.

In the following discussion, **highlighted** words indicate the keys to be pressed on the keyboard. If two keys are required to be pressed simultaneously, they are shown with a + sign between them (e.g., **Alt+F3** means that the **Alt** and **F3** keys should be pressed simultaneously). Italicized letters are used to mark the commands available in menus, submenus, and other options as they appear on the computer screen at various times.

In every example, we discuss many procedures to introduce the techniques of analysis. For ease of reference in later examples, these procedures are arranged in steps that are numbered in the Ex*u.v.w* format, where *u* is the example number, *v* is the procedure number in the *u*th example, and *w* is the step number in the procedure *v*. For instance, Ex1.3.2 refers to the 2nd step of the 3rd procedure in example 1.

7.1 Constructing Trees from Distance Data

This example introduces procedures for changing the default directory, selecting options from menus, opening files in the read-only mode, activating a distance data file for analysis, and building trees from the distance data.

Ex1.0.1 Go to the C:\MEGA directory, type MEGA on the C:\MEGA> DOS prompt, and press **Enter**.

Ex1.0.2 A *Welcome* box appears on the screen that displays the current version of MEGA and the names and addresses of the authors.

Ex1.0.3 Press **Enter** to remove this box from the screen.

Since all the example files are located in the C:\MEGA\EXAMPLES directory, we first set C:\MEGA\EXAMPLES as the current working directory.

Ex1.1.1 Press **F10** to go to the main menu. A sliding bar will then appear at the top of the screen. Using the arrow keys (→,←), go to the *File* option and press **Enter**. The *File* menu unfolds.

Ex1.1.2 Using the down arrow key (↓), go to the *Change Dir* option, and press **Enter**. The *Change Directory* dialog box appears. Type C:\MEGA\EXAMPLES, and press **Enter**. Now with the **Tab** key, go to the **OK** button and press **Enter**. This sets C:\MEGA\EXAMPLES as the current directory.

In this example we will use the data present in the HUMDIST.MEG file. Let us examine the content of this file before proceeding further. Since we do not intend to edit this file, we will use the file browsing command.

Ex1.2.1 The *Browse* command is present in the *File* menu. So, first unfold the *File* menu (follow Ex1.1.1) and then use the down arrow key (↓) to go to the *Browse* option and press **Enter**. The *File Name* dialog box appears. Type HUMDIST.MEG, and press **Enter**.

Ex1.2.2 The HUMDIST.MEG is displayed in a window with a double line border. (Note the presence of two icons on the top border of this window. They are for use with the mouse.) Examine the contents of the file, and note the presence of the #mega format specifier, a title, OTU names, and the lower-left triangular distance matrix.

Ex1.2.3 Now close this file before proceeding for analysis by selecting the *Window*| *Close* command.

A data file must be activated before the analysis can be performed. (Remember that opening a file for browsing or editing is different from activating it for analysis.) Let us activate the HUMDIST.MEG data file now.

Ex1.3.1 Press **F10** to go to the main menu. Using the arrow keys, move the slide bar to the *Data* menu and press **Enter**. The *Data* menu contains many commands. Use arrow keys to go to the *Open Data* command and press **Enter**. A submenu with four options will appear.

Ex1.3.2 Of the four options available, choose the *Distance* option, and press **Enter**. A *File Name* dialog box will appear. Type HUMDIST.MEG, and press **Enter**.

Ex1.3.3 This produces a *Input Data* dialog box that inquires about the input distance data format. Using the **Tab** key select the *Lower-left triangular-matrix* option, and press **Enter**. The message "Reading input data. Please Wait!" will appear.

Ex1.3.4 Since this data file does not contain any errors, no error messages are flashed. Do you see any change on the screen? A box labelled *Current Data* appears that contains information about the input data just activated for analysis. A *Selections* box also appears on the screen that informs you regarding the current analysis methods chosen.

Let us make a phylogenetic tree from the distance data. For this purpose, you select a tree building method first, and use the *Construct Tree(s)* command.

Ex1.4.1 The *Phylogeny* menu contains the *Neighbor-joining* command. Select this command, and press **Enter**.

Ex1.4.2 Choose the *Construct Tree(s)* option from the *Phylogeny* menu, and press **Enter**. The message "The tree is being reconstructed. Please Wait!" is displayed.

Ex1.4.3 A neighbor-joining tree is displayed on the screen instantly. Examine the tree on the screen, and press **Esc** key to remove it.

With this, let us end this session of MEGA.

Ex1.5.1 Go to the *Data* menu, select the *Close Data* command, and press **Enter**. The program inquires if data are to be inactivated. Press **Enter**, the *Current Data* and *Selections* boxes disappear from the screen.

Ex1.5.2 To exit MEGA, press **Alt+X** or select the *Exit MEGA* command from the *File* menu.

7.2 Computing Statistical Quantities for Nucleotide Sequences

In this exercise the use of the *Data|Data Presentation* command for computing various statistical quantities of nucleotide sequences is illustrated. In addition, short-cuts for frequently used commands, method of accessing on-line helps, and the reason why some commands are enabled and others are disabled are explained.

Ex2.0.1 Go to the C:\MEGA directory first, type MEGA on the C:\MEGA> DOS prompt, and press **Enter**. Press **Enter** in the *Welcome* box that appears on the screen.

Now, set C:\MEGA\EXAMPLES as the default directory (see Ex1.1.1 - Ex1.1.2).

Let us examine the contents of the file DROSOADH.MEG by using the hot-key for the *File|Browse* command.

Ex2.1.1 Press **F5**. This brings up a *File Name* dialog box. Type DROSOADH.MEG, and press **Enter**. The distance file will appear on the screen in a double line bordered window. Press **F1**, the help key, to activate the help, and after a quick glance at the help text, press **Esc** or click on the icon ([■]) on the top left corner to put the help window away.

Ex2.1.2 Examination of the DROSOADH.MEG file reveals the presence of the #mega format specifier, a title, OTU names, and the interleaved sequence data.

Ex2.1.3 Let us close this file by pressing **Alt+F3** (short-cut for the *Window|Close* command).

Before activating the data file DROSOADH.MEG for analysis, let us try the *Data|Close Data* command displayed in the light shade of gray. Can you select this command? No, because no data file is currently active. Isn't the *Open Data* command displayed in a brighter color? The *Open Data* command is enabled, but the *Close Data* command is disabled and is not selectable.

For studying statistical quantities of the data present in the file DROSOADH.MEG, we first activate it.

Ex2.2.1 Select the *Data|Open Data* command, and choose *DNA* from the resulting menu. Type DROSOADH.MEG in the *File Name* box and press **Enter**.

Ex2.2.2 A dialog box appears where the noninterleaved (continuous) format is selected. Use the **Tab** and arrow keys to choose the interleaved format. Everything seems alright. Press **Enter** or click on the **OK** button.

Ex2.2.3 The message "Reading input data. Please Wait!" appears. Soon after, the program inquires whether the data are protein-coding or not. Press **Y** to select the *Protein-coding* mode. For the genetic code table to be used, select the *"Universal"* option and press **Enter**. The *Current Data* and the *Selection* boxes appear on the screen.

Now examine the *Data* menu again. The *Close Data* command is enabled, displayed in a bright color, and the *Open Data* command is disabled (try to select any data type in its submenu). The *Close Data* command is enabled because some data are active, whereas the *Open Data* command is disabled because it is not possible to activate more than one data set at any one time in MEGA.

Let us take a look at the data by using the *Data Presentation* command and compute some basic statistics for these data.

Ex2.3.1 Select the *Data|Data Presentation* command. The message "Sequence data in preparation. Please wait!" appears. The sequences are then displayed on the screen.

Ex2.3.2 DNA sequences are displayed on the screen with the cursor on the first site of the first sequence. Use the right arrow (→) and left arrow (←) keys to move from site to site and note a change in the *Site#* display in the bottom-right corner. Use the up (↑) and down (↓) arrow keys to move between OTUs and note changes in the *OTU Name* view on the top panel. The *Total Sites* view on the bottom panel displays the sequence length at all times and the *Marked Sites* displays 0 because no special site attributes are marked yet.

Ex2.3.3 To highlight variable sites, press **V** (or click on the button marked V). All sites that are variable are highlighted, and the number in the *Marked Sites* display changes. Press **V** again. The sites return to the normal color and *Marked Sites* display shows 0 again.

Ex2.3.4 Now to highlight parsimony-informative, and 2- and 4-fold redundant sites. (Read about these buttons by pressing help key **F1**.)

Ex2.3.5 To compute the nucleotide base frequencies, nucleotide pair frequencies, and the codon usage bias, we use the Statistics command. Press **S** (or click on the S button). From the dialog box, select the *All OTUs* option for nucleotide frequencies, nucleotide pair frequencies, and codon usage and the *Overlapping* option for the *Variability* by using the **Tab** and arrow keys, and press **Enter**. Type C:\NUCSTAT.OUT in the *Output File* box when the program asks for an output file name.

Ex2.3.6 To examine the statistical quantities computed in the previous example, press **Esc** to remove the *Sequence Data* window. Then use the *File|Edit|Open File* command (**F3**) to see the C:\NUCSTAT.DAT file.

Ex2.3.7 Now again display the sequence data on the screen by using the *Data|Data Presentation* command (or use the hot-key **F4**).

Ex2.3.8 Since the data are in the protein-coding mode, they can be translated into amino acid sequences. To do this, press **T**. The DNA sequences are now replaced by the amino acid sequences. Note that the commands for highlighting 2- and 4-fold redundant sites are no longer enabled.

Ex2.3.9 Now use the Statistics command to compute the amino acid frequencies. For the output file name, type C:\AMINOSTAT.OUT. Before examining the output from this operation, press **T** to restore the nucleotide sequences to the screen.

Ex2.3.10 As usual, press **Esc** to remove the displayed data and use **F5** to examine this file.

To inactivate the currently used data and exit MEGA, press **Alt+X**. You simply come out of MEGA. Did you realize that we did not inactivate the data file before exiting MEGA? You don't need to do it because it is automatically done by the program.

7.3 Estimating Evolutionary Distances from Nucleotide Sequences

In this example we compute various distances for the *Adh* sequences from 11 *Drosophila* species (Thomas and Hunt 1993). We used this data in the previous example to study various sequence statistics. In addition, you will be see how these distances can be written in a file in various formats through options for page size, precision, and relative placement of distances and their standard errors.

Ex3.0.1 Go to the C:\MEGA directory first, type MEGA on the C:\MEGA> DOS prompt, and press **Enter**. Now again press **Enter** in the *Welcome* box that appears on the screen.

As usual, set C:\MEGA\EXAMPLES as the default directory using the *File|Change Dir* command. Now activate the data file DROSOADH.MEG using the instructions given in Ex2.2.1 - Ex2.2.3.

The computation of distances from nucleotide sequences is a two step process. First you need to select an appropriate distance estimation method in the *Distance* menu, and the distances are then computed by using the *Compute Distances* command that is also available in the *Distance* menu.

Now look at the *Current Data* box present at the lower right corner of the screen. It indicates that the data are being used in the coding mode. At this time, go to the *Distance* menu (**Alt+T**), and note that all distance estimation methods in submenus *Nucleotide*, *Syn-nonsynonymous*, and *Amino Acid* are displayed in a bright shade (enabled commands). If you are analyzing noncoding sequences, only the *Nucleotide* submenu will contain enabled commands, and the *Syn-nonsynonymous* and *Amino Acid* submenus will contain disabled commands.

Let us begin by computing the proportion of nucleotide differences between each pair of *Adh* sequences.

Ex3.1.1 Select the *Distance|Nucleotide* command (**Alt+T,N**). From the submenu, select the *p-distance*. This produces a box with four options (to learn about these options, press **F1**). Just press **Enter** to select the default option.

Ex3.1.2 Look at the *Selection* box on the screen. It shows that you have chosen the *p*-distance.

Ex3.1.3 Now select the *Distance | Compute Distances* command. This command will produce a dialog box with many options. At this moment, just press **Enter** to accept all default options.

Ex3.1.4 The message "Pairwise distances are being estimated. Please wait!" appears. Once all the distances are computed, the program requests a file name to output these distances. For now just type **C:\PDIST.OUT**.

Ex3.1.5 Use the *File | Browse* command to examine the distance output file.

Now you know how to compute distances. So let us compute distances using some other methods and compare them with each other.

Ex3.2.1 Select the *Distance | Nucleotide* command. From the submenu, select the *Jukes-Cantor Distance*. Now select the *Distance | Compute Distances* command. Just press **Enter** to accept all the default options in the resultant dialog box. Once the distances are computed, supply **C:\JCDIST.OUT** as the file name to write the distances estimated.

Ex3.2.2 Follow the steps Ex3.1.1 - Ex3.1.3 and compute the *Tamura Distance*. For the file name, type **C:\TAMDIST.OUT**.

Ex3.2.3 By this time you have three files containing the distances estimated by three different methods. You can now compare these distances on the screen by pressing the hot-key **F5** three times for the three files created above.

Ex3.2.4 For an easy comparison, use the *Window | Tile* command to arrange multiple files on the screen.

Ex3.2.5 Now remove all these files from the screen by pressing **Alt+F3** three times.

The file **DROSOADH.MEG** contains nucleotide sequence data, and we have computed nucleotide distances from these data. Let us now compute the proportion of amino acid differences. Note that MEGA will automatically translate the nucleotide sequences into amino acid sequences using the selected genetic code table.

Ex3.3.1 Select the *Distance | Amino Acid* command (**Alt+T,A**). From the submenu, select the *p-distance*.

Ex3.3.2 Look at the *Selection* box on the screen. It shows that you have chosen the amino acid *p*-distance.

Ex3.3.3 Now select the *Distance | Compute Distances* command. This command will produce a dialog box with many options. Use the **Tab** key and go the *Estimate* option in this dialog box and select *Distances and SE's*. In this dialog box, note that the *Write Distances* and *Write Standard Errors* options show different selections. This means that the distances and their standard errors will be written on the opposite sides of the output matrix. In any case, just press **Enter** to accept the settings.

Ex3.3.4 The message "Pairwise distances are being estimated. Please wait!" appear. Once all the distances are computed, the program requests a file name to output these distances. For now just type-in **C:\PAMINO.OUT**.

Ex3.3.5 Use the *File | Browse* command to examine the distance output file. In contrast to previous files, this file contains both the distances and their standard errors.

In the previous steps, we chose the default option where distances and their standard errors were written on the opposite sides of a matrix, and the distance matrix was fragmented in many parts because it did not fit on one page. Let us write these estimates in the distance ± standard error format in one single matrix.

Ex3.4.1 Look at the *Selection* box on the screen. It shows that you have chosen the amino acid *p*-distance. So we do not need to choose the distance estimation method again.

Ex3.4.2 Select the *Distance | Compute Distances* command. This command will produce a dialog box where *Distances and SE's* option is already selected. (MEGA remembers your previous selections.) Now go to the *Write Standard Errors* option with the help of the **Tab** key and use arrow keys to choose the *Upper-right matrix*. At this time, the *Write Distances* and *Write Standard Errors* options show the same selection. This means that the distances and the standard errors will be written on the same side of the matrix. To write the complete matrix on one page, go to the *Page size* option by using the **Tab** key and specify a page size of 1000. Large page sizes ensure that the distance matrix will not be fragmented. Finally, just press **Enter** to accept the settings.

Ex3.4.3 The message "Pairwise distances are being estimated. Please wait!" appear. Once all the distances are computed, the program requests a file name to output these distances. For now just type-in **C:\PAMINO.OUT**. Program will enquire whether you want to overwrite the file. Press **Enter** to say *Yes*.

Ex3.4.4 Use the *File|Browse* command to examine the distance output file. In contrast to the previous files, this file contains both the distances and their standard errors in the desired format. Now close this file.

Let us inactivate the currently used data set and end the current session of MEGA by pressing the hot-key **Alt+X**.

7.4 Constructing Trees and Selecting OTUs from Nucleotide Sequences

The CRAB.MEG file contains nucleotide sequences for the large subunit mitochondrial rRNA gene from different crab species (Cunningham *et al.* 1992). Since the rRNA gene is transcribed but not translated, it is in the category of non-coding genes. Let us use this data file to illustrate the procedures of building trees and in-memory sequence data editing using the commands present in the *Data* and *Phylogeny* menus.

Ex4.0.1 Go to the C:\MEGA directory first, type MEGA on the C:\MEGA> DOS prompt, and press **Enter**. Now again press **Enter** in the *Welcome* box that appears on the screen.

Now set C:\MEGA\EXAMPLES as the default directory using the *File|Change Dir* command, and examine the contents of the CRAB.MEG file (use hot-key **F5**). In this data file, note the comments starting on the third line. The comments indicate that the data are in the noninterleaved format and that '?' and '-' are used to designate missing-information and alignment gap sites. Close this file using **Alt+F3**.

Let us activate the crab sequence data for analysis.

Ex4.1.1 Select the *Data|Open Data* command and choose the *DNA* option. In the *File Name* dialog box, type CRAB.MEG and press **Enter**.

Ex4.1.2 A dialog box will appear. Use the **Tab** key to move around in the dialog box but do not change anything. In this box the noninterleaved format is selected and the symbols used for missing-information data, identical sites, and alignment gap are '?', '.', and '-', respectively. So everything is fine. Just **Enter** (or click on the **OK** button).

Ex4.1.3 A status report box informs that the data are being read. At this stage, the program inquires whether the nucleotide sequence data are from a coding or noncoding gene. Select the noncoding mode by pressing the N key. The *Current Data* and the *Selection* windows appear on the screen.

The use of *Data|Data Presentation* command was introduced in the second example. As an exercise, you may try to examine this data set on the screen by using that command. Just press **F4**, the hot-key for the *Data Presentation* command, and you will see the data on the screen. For help, press **F1** anytime.

Let us start by building a neighbor-joining tree. For this purpose, we need to specify a distance estimation method in the *Distances* menu and a tree building method in the *Phylogeny* menu. The *Phylogeny | Construct Tree(s)* command is then used for tree building.

Ex4.2.1 Select the *Distance | Nucleotide* command. Choose the *Jukes-Cantor Distance* from the resultant submenu.

Ex4.2.2 To use the neighbor-joining method for tree building, select the *Phylogeny | Neighbor-Joining* command.

Ex4.2.3 Invoke the *Phylogeny | Construct Tree(s)* command. This brings a status report box with a message. The neighbor-joining tree will soon appear on the screen.

Ex4.2.4 At this moment, you are automatically put into the phylogenetic-tree editor. This editor provides operations in two modes: view mode and edit mode. The edit mode can be recognized by the presence of a blinking cursor. By default you are placed in the view mode. Press **E** to enter the edit mode (a blinking cursor will appear).

Ex4.2.5 Use the arrow keys (↑,↓,→,←) to move to different branches on the tree and note the change in the branch length in the lower-left corner corresponding to the focused branch. Now, position your cursor on the far left corner of the screen.

Ex4.2.6 At this time the cursor assumes a triangular shape instead of the diamond (♦). Press **M**, the mirror image of the original tree is displayed instantly. Press **M** again, the tree reverts to its original shape.

Ex4.2.7 Press the **Up** arrow key (↑) just once. The cursor moves upwards to the next branch. Press **F**, the flip command. A mirror like effect is produced on the sub-tree anchored on the currently focused branch.

Ex4.2.8 The *Topology* command is to display just the branching pattern of the tree. Press **T**, the *Topology* command, the branching pattern (without actual branch lengths) is displayed on the screen. Press **T** again, the actual NJ tree reappears.

Ex4.2.9 Press **F1** to examine the help for tree editor. Use the **Tab** key to get to the highlighted word *Swap* and press **Enter**. You will see information about the *Swap* command. This can be used for more commands. Press **Esc** to exit help.

Ex4.2.10 **DO NOT** remove the tree from the screen. We shall use it for illustrating how a tree can be printed.

At this moment, we have the NJ tree on the screen. In MEGA, you can print this tree by using a printer. Let us see how.

Ex4.3.1 You can print a tree in two ways. First, a tree can be written as an ASCII-text file. In this case, an exact replica of the tree displayed on the screen is written in the desired file. Since the NJ and UPGMA trees are shown with approximate branch lengths, this output does not reflect true branch lengths. By contrast, if you have a printer attached to your computer, you can print the tree with exact branch lengths.

Ex4.3.2 Press **P**, the Print command. A dialog box with two options appears. If there is no printer attached to your computer, select the *ASCII-Text file* output option using the **Tab** and arrow keys, and then press **Enter**. For the output file name, type **C:\TREE.NJ**. If you have a printer attached to your computer, select the *Printer* option and press **Enter**. An dialog box appears on the screen. In this dialog box many options are available. (Press **F1** to learn about them.)

Ex4.3.3 Do not change anything in this dialog box, and just select the *Preview* command using the **Tab** key. A graphic image of the tree will be displayed on the screen. Press **Enter**, and you are back to the option box. Now go to *Write information* option, and select the *Branch lengths*. Again select the *Preview* command (you may press **Alt+V**). The tree is now drawn with branch lengths. Press **Enter** to come out of the graphics image.

Ex4.3.4 To print the tree with a printer, select an appropriate printer using the *Printer* command.

Ex4.3.5 Press **Enter** (or click on **OK**) to print the tree on the selected printer.

Ex4.3.6 Press **Esc** to exit the phylogenetic-tree editor.

In MEGA, you can also construct maximum parsimony trees. Let us construct a maximum parsimony tree(s) by using the *branch-and-bound search* option.

Ex4.4.1 Select the *Phylogeny|Maximum Parsimony* command. In the resultant submenu, choose the *Branch-and-Bound Search* option.

Ex4.4.2 Invoke the *Phylogeny|Construct Tree(s)* command, and press **Enter** to accept default options in the dialog box produced. This brings a status report box. An MP tree appears on the screen as soon as the search is completed.

Ex4.4.3 Note that no branch lengths are given for an MP tree in MEGA. Also that the *Topology* command is disabled because in this case only the branching pattern is available.

Ex4.4.4 Now print this tree (See Ex4.3.1 - 4.3.5). You do not have to specify the printer name again because MEGA remembers your selection.

Ex4.4.5 Press **Esc** to exit the phylogenetic tree editor.

Ex4.4.6 Compare the NJ and MP trees. For this data set, the branching pattern of these two trees is identical.

As an exercise, use the *Heuristic Search* for finding the MP tree. In this example, you will find the same tree as that obtained by the branch-and-bound method if you use the default option (search factor equal to 2 for all steps of OTU addition). However, the computational time will be much shorter. Actually, in this example even a search factor equal to 0 will recover the MP tree.

We will now examine how some data editing features work in MEGA. For noncoding sequence data, OTUs as well as sites can be selected for analysis. Let us remove the first OTU from the current data set.

Ex4.5.1 Select the *Data|Select OTUs* command. A *Select OTUs* list dialog box is displayed.

Ex4.5.2 All the OTU labels are checked (✓) in this box. This indicates that all OTUs are included in the current active data subset. To remove the first OTU from the data, press the **Del** key (or double click on the first OTU). The first OTU is no longer checked. Press **Enter**.

Ex4.5.3 Note a change in the *Used OTUs* entry in the *Current Data* window. The number of OTUs used for analysis has been reduced by one.

Ex4.5.4 Again use the *Data|Data Presentation* command (**F4**) to see the changes made.

Now, construct a neighbor-joining tree from this data set (Ex4.2.3) that contains 12 OTUs instead of 13.

Let us inactivate the currently used data set and end the current session of MEGA by pressing the hot-key **Alt+X**.

7.5 Tests of the Reliability of a Tree Obtained

In this example, we will conduct two different tests using mitochondrial 12S rRNA gene sequences from 12 flightless birds (ratites) and one related species (Cooper *et al.* 1992) and learn how to construct a condensed tree.

Ex5.0.1 Go to the C:\MEGA directory first, type MEGA on the C:\MEGA> dos prompt, and press **Enter**. Now again press **Enter** in the *Welcome* box that appears on the screen.

Set C:\MEGA\EXAMPLES as the default directory and browse through the file RATITE.MEG.

Activate the data present in the RATITE.MEG file by using the *Data|Open Command* and using the default options. This gene does not code for a protein so choose the noncoding mode.

Let us start with the bootstrap test for the neighbor-joining tree. For this purpose, we need to specify a distance estimation method in the *Distances* menu and a tree building method in the *Phylogeny* menu. The *Phylogeny|Bootstrap Test* command is then used for performing a bootstrap test.

Ex5.1.1 Select the *Distance|Nucleotide* command. Choose the *Jukes-Cantor Distance* from the resultant menu.

Ex5.1.2 To use the neighbor-joining method for tree building, select the *Phylogeny|Neighbor-Joining* command.

Ex5.1.3 Invoke the *Phylogeny|Bootstrap Test* command. This produces a dialog box with many options. Just press **Enter**. The program will ask about the filename to store some information from bootstrap test. Just press **Enter** at this time. The test begins, and you can see its progress on the screen. The neighbor-joining tree with bootstrap confidence limits (*BCL*) appears on the screen in the phylogenetic tree editor.

Ex5.1.4 Press **E** to go to the Edit mode. A blinking cursor will appear.

Ex5.1.5 Use arrow keys (↑, ↓, →, ←) to move to different branches on the tree and note the change in the branch length and *BCL* values in the lower-left corner.

Ex5.1.6 Let us make a condensed tree. For this purpose, we will use the *Cut-Off* command. Press **O**, and you will be asked about a cut-off level. Type 70 in the box and press **Enter**. The condensed tree is produced on the screen. This tree shows all the branches that are supported at *BCL*≥70%. Press **O** again, and the actual NJ tree will reappear.

Ex5.1.7 Print this tree to the printer (see Ex4.3.1 - Ex4.3.6) with *BCL* values selected in the *Write information* option in the tree printing dialog box.

Ex5.1.8 Press **Esc** to exit the tree editor.

For neighbor-joining trees, it is possible to conduct the standard error test for every interior branch by using the *Phylogeny|Standard Error Test* command. In MEGA this test is available for the *p*-distance, Jukes-Cantor distance, and Kimura's 2-parameter ($s+v$) distance for nucleotide sequences. Since we did the above analysis for the Jukes-Cantor distance, we will use the same distance estimation method to compare the results from the bootstrap and standard error tests. Since the *Selections* box shows that Jukes-Cantor distance and NJ tree making method are already selected. We just have to invoke the *Phylogeny|Standard Error Test* command.

Ex5.2.1 Go to the *Phylogeny* menu and select the *Standard Error Test* command. This produces a dialog box that shows that the *Complete-Deletion* option will be used for missing-information and alignment gap sites. Press **Enter** to start the test, and you will see its progress on the screen.

Ex5.2.2 The neighbor-joining tree with confidence probabilities (*CP*) from the standard error test of branch lengths is displayed on the screen.

Ex5.2.3 Compare the *CP* values on this tree with the *BCL* values of the tree that you printed in the previous procedure.

Now exit MEGA using the **Alt+X** command.

7.6 Test of Positive Selection

In this example, various analyses of protein-coding nucleotide sequences for five alleles from the human HLA-A locus (Nei and Hughes 1991) are presented.

Ex6.0.1 Go to the C:\MEGA directory first, type MEGA on the C:\MEGA> DOS prompt, and press **Enter**. Now again press **Enter** in the *Welcome* box that appears on the screen.

Set C:\MEGA\EXAMPLES as the default directory and browse through the file HUMHLA.MEG. In this file, sequences are arranged in the interleaved (block-wise) format. Note that the antigen recognition sites (ARS) are marked in comments.

For analyzing the data present in file HUMHLA.MEG, we first activate the data.

Ex6.1.1 Select the *Data|Open Data* command, and choose *DNA* from the resulting menu. Type HUMHLA.MEG in the *File Name* box and press **Enter**.

Ex6.1.2 A dialog box appears where the noninterleaved (continuous) format is selected. Use the **Tab** and arrow keys to choose the interleaved format.

Everything seems alright. Press **Enter** or click on the **OK** button.

Ex6.1.3 The message "Reading the input data file. Please Wait!" appears. Soon after, the program inquires whether the data is protein-coding or not. Press the **Y** key to select the *Protein-coding* option. For the genetic code table to be used, select the *"Universal"* option, and press **Enter**. The *Current Data* and the *Selection* boxes appear on the screen.

Now to study positive Darwinian selection for HLA-A alleles, we need to select all codons that are involved in the antigen recognition sites. These codons are shown with a plus sign (+) in the HUMHLA.MEG data file. For this, we need to use the *Select Sites/Codons* command.

Ex6.2.1 Select the *Data|Select Sites/Codons* command and choose the *Individual* option. A *Select Codons* box appears with a list of codon numbers.

Ex6.2.2 By default all the codons are checked (✓) in this list indicating that all of them are included in the currently active data set. To remove any codon from the data, press the **Del** key (or double click). Press **Del** on all numbers except 5, 7, 9, 22, 24, 26, 57, 58, 59, 61-77, 80-82, 84, 95, 97, 99, 114, 116, 143, 145-147, 149-152, 154-159, 161-163, 165-167, 169, and 171. Now press **Enter**.

Ex6.2.3 Note a change in the *Used Codons* entry in the *Current Data* window. This number must be 57. If it is not, go back to Ex6.2.1 and check.

Ex6.2.4 Now use the *Data|Data Presentation* command to see the selected data subset. Here you can check if the correct codons are included in the data set or not.

Let us compute the synonymous and nonsynonymous distances appropriate for studying positive Darwinian selection in this set of antigen recognition codons. For this, you must first specify the distance measure and then use the *Compute Distances* command.

Ex6.3.1 Select the *Distance|Syn-Nonsynonymous* command. Choose the *Jukes-Cantor Correction* from the resultant menu. A dialog box appears. Select the *Synonymous* option.

Ex6.3.2 Now select the *Distance|Compute Distances* command. In the dialog box, select the *Distance and SE's* option and also select the *Compute overall mean* option. We need to do this to obtain all the pairwise synonymous distances and the average synonymous distance and the standard error of this average. (Please read the manual to find the meanings of different options or use the **F1** key to get help.) Now press **Enter**. "Pairwise distances are being estimated. Please Wait" appears.

Ex6.3.3 Once the distances are calculated, an output file name with correct path is required to save the distances. Type C:\SYN.DAT and press **Enter**. Distances are output to this file. You may use the file browsing command to examine this file. (The average synonymous distance and its standard error should be 0.0618 and 0.0262, respectively.)

Ex6.3.4 Now we need to compute the average nonsynonymous distance and its standard error. For this purpose, we repeat the process shown in Ex6.3.1 - Ex6.3.3 but for nonsynonymous distances this time. That is, select the *Distance | Syn-Nonsynonymous* option, choose the *Jukes-Cantor Correction* from the resultant menu, and select *Nonsynonymous* option from the dialog box.

Ex6.3.5 Now select the *Distance | Compute Distances* command. A dialog box appears. In this dialog box, the *Distance and SE's* and *Compute overall mean* options are already selected. Now press **Enter**. "Pairwise distances are being estimated. Please Wait" appears.

Ex6.3.6 Once the distances are calculated, an output file name with correct path is required to save the distances. Type C:\NONSYN.DAT and press **Enter**. Distances are output to this file. You may use the file browsing command to examine this file. (The average nonsynonymous difference and its error should be 0.1373 and 0.0231, respectively.)

Ex6.3.7 Now we have estimated the average synonymous and average nonsynonymous substitutions per site and the standard errors of these estimates. To conduct the test, refer to section **4.2** (equation 4.47). The difference in synonymous and nonsynonymous substitutions should come out to be significant at the 5% level.

Now exit MEGA using the **Alt+X** command.

8

Command Reference

In this chapter, all the menus and commands that are available in the user-interface will be discussed. Many of these commands are tied to the hot-keys displayed next to them. These hot-keys sometimes require two or more keys. Such combinations are shown with a + sign to indicate that these keys should be pressed simultaneously.

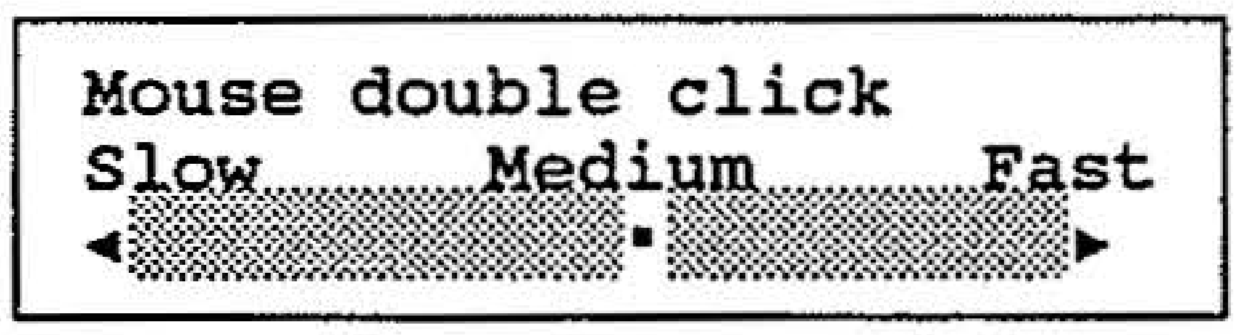
MEGA menu
Alt+M

About MEGA — It opens a window with information about the current version of MEGA copyright notice, and the name and addresses of the authors for correspondence.

Reference — It gives a way of citing MEGA in the "Literature Cited" section of research articles.

Mouse — The Mouse command brings up a dialog box for selection of various mouse control options, including:

- how fast a double-click is, and
- which mouse button (right or left) is active.



Mouse double click
Slow Medium Fast

[X] Reverse mouse buttons

The *Mouse-double-click* slider bar adjusts the double-click speed of the mouse. The *Reverse mouse button* makes the right most mouse button active instead of the default left most button.

Calculator — It opens a simple four function calculator that can be operated with the keyboard as well as the mouse.

Calendar — The *Calendar* displays the current month, highlighting today's date. The next and previous months are viewed with the '+' and '-' keys. Clicking the mouse on ▲ and ▼ icons also changes the month.

Thank You — A *Thank You* note for help in development of MEGA is included here.

Using Help **Alt+F1** — Help on the on-line context-sensitive help is provided here.

File menu
Alt+F

Browse **F5** — The *Browse* command brings up a *File Name* box where the file to be opened is specified. This file is displayed on the screen in the read-only mode. Use the cursor movement keys to move around in the file or use the mouse on the scroll bars (see chapter 6 for details).

Edit — This editor supports most of the basic editing functions, including: saving file, manipulating text-blocks, and finding and replacing text-strings (see chapter 6 for more details).

Open File **F3** — It displays a *File Name* dialog box for selecting the file to edit. The file is opened for editing in a new window. Many files can be opened for editing simultaneously in this editor.

Create New File — This command creates a new file with a temporary file name **UNTITLED** for editing. MEGA will require a file name whenever the file is saved.

Save File **F2** — The currently edited file is saved on the disk with this command. If an **UNTITLED** file is saved, then a dialog box will prompt for the file name to store the text.

Save As **Alt+F2** — A *File Name* box prompts for a name to save the current file with a new file name.

Cut **Shift+Del** — *Cut* removes the selected text from the current file and places it in the clipboard. A *Paste* operation will retrieve this text when desired. Text can be pasted many times and to many files.

Copy
Ctrl+Ins
Copy keeps the selected text intact and places its copy in the clipboard for the *Paste* command that can retrieve this text in any desired editor window.

Paste
Shift+Ins
With this command, the most recently selected text in the clipboard is inserted into the current file at the cursor position.

Clear
Ctrl+Del
Removes the selected text from the current document. The cleared text is not retrievable.

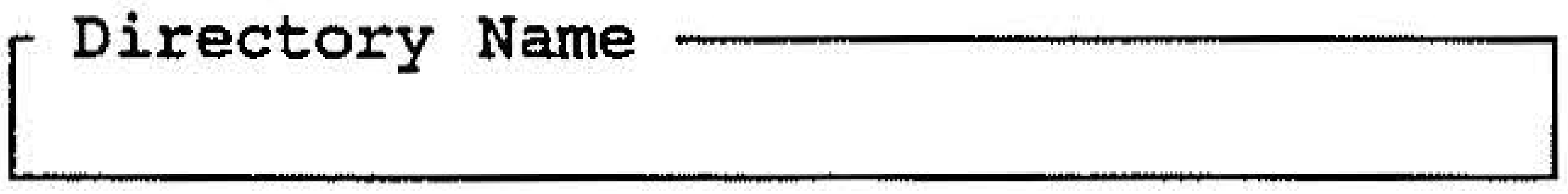
Undo
Ctrl+U
Undo reverses the last editing command and restores the text. It works only on the last modified line.

Find String
Ctrl+Q F
This command displays the *Find Text* dialog box in which the text to be searched is entered.

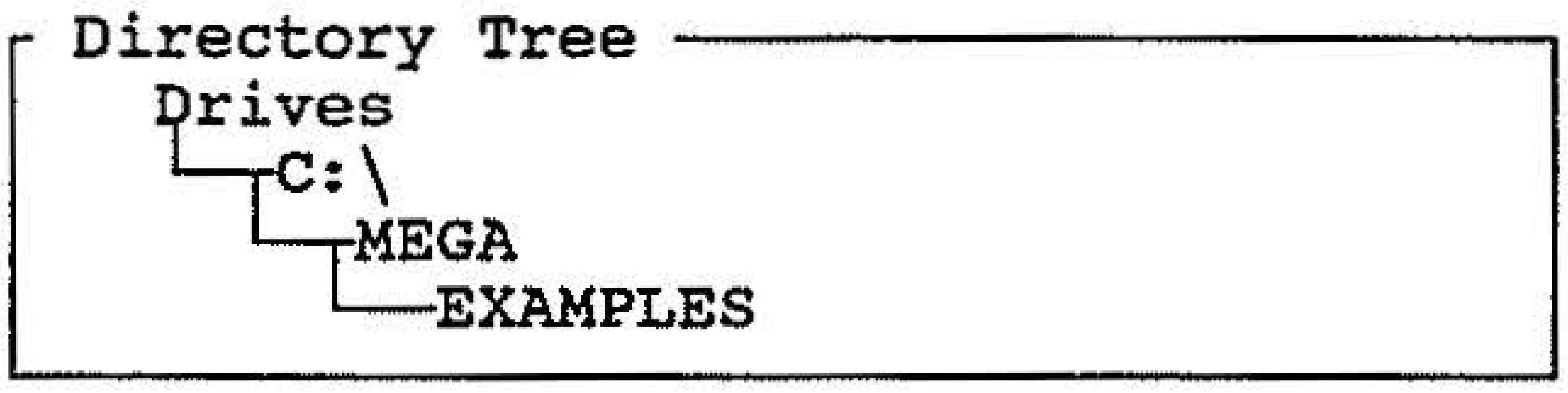
Replace String
Ctrl+Q A
It displays the *Find and Replace* box where the text-string to be searched and replaced is entered along with the replacement string.

Show Clipboard
Shows the contents of the clipboard and the currently selected text block for *Paste* operations.

Change Dir
The *Change Directory* dialog box consists of a directory input box, directory tree list box, and three buttons: *OK*, *Chdir*, and *Revert*.

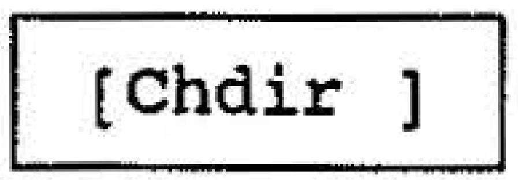


The path of the new directory is typed in the *Directory Name* input box.



```
Drives
└─┬C:\
  └─┬MEGA
    └──EXAMPLES
```

The *Directory Tree* displays a tree of directories.



The *Chdir* button changes the current directory once you select or type in a directory name. The *Revert* button lets you go back to the previous directory, as long as you have not exited the dialog box.
The *Directory Tree* list box provides navigation throught the directory structure with selection bar and **Alt+C** (short cut for *Chdir* command). If you are using the keyboard, press **Enter** to choose the changes made. Press **Esc** to cancel the changes made.

DOS Shell

With the *DOS Shell* command, you can leave the program temporarily to perform DOS commands and run other programs. To return to MEGA, type EXIT at the DOS prompt.

Exit MEGA
Alt+X

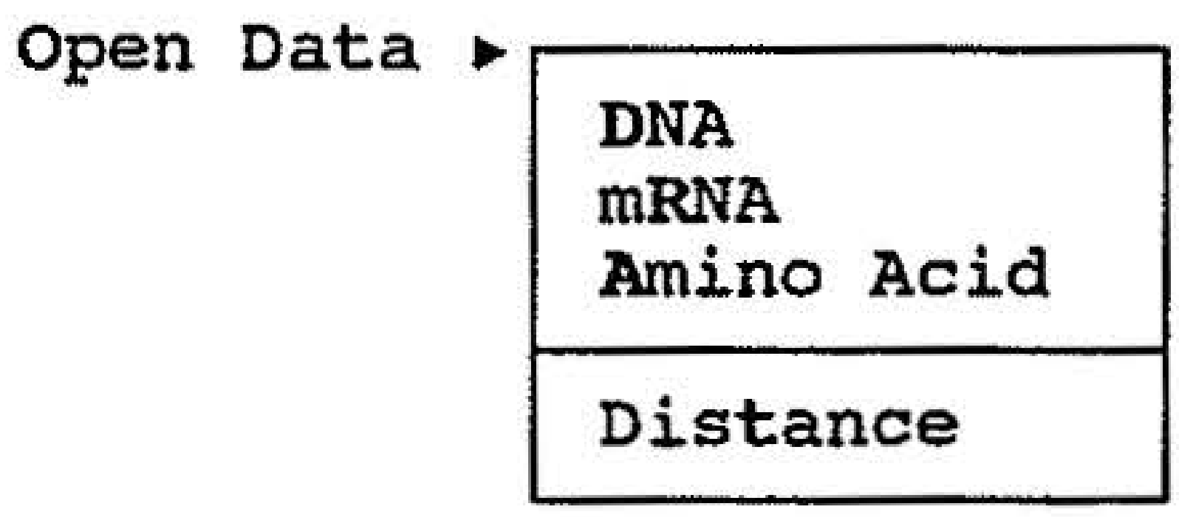
This command terminates the current session of MEGA. It deletes all options and cleans all the temporary files from the disk.

Data menu
Alt+D

Activation of the data file is the first step in data analysis. As soon as a data set becomes active, relevant in-memory data editing options are enabled depending on the input data. Virtually any subset of the original data can be selected through options for selecting OTUs and sites (or codons). A detailed description of in-memory editing options is given in chapter 2. Selected subset data can be examined with the *Data Presentation* command.

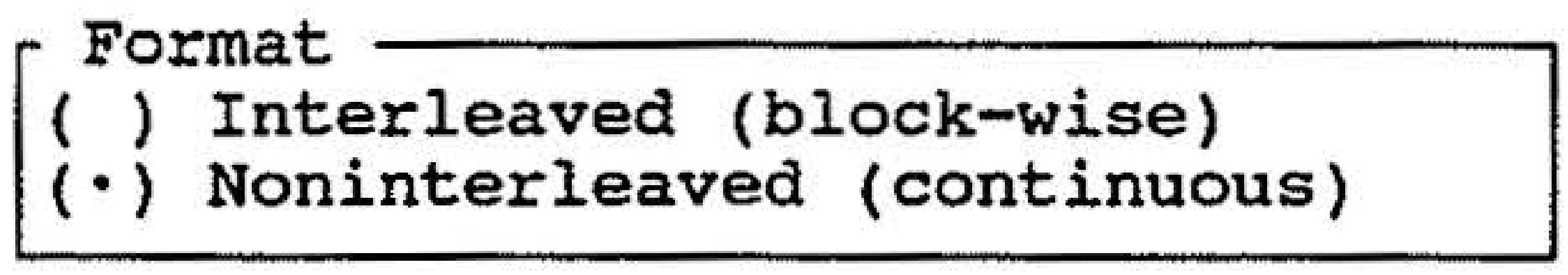
Open Data

This command produces a submenu with four options. This command is enabled only if no other data set is active.



Once the type of data is selected, you will be asked to specify the input file name. A dialog box will then appear to query about various input attributes of the data present in the input data file.

For sequence data



```
Format
( ) Interleaved (block-wise)
(•) Noninterleaved (continuous)
```

A discussion on the difference between interleaved and noninterleaved is given in chapter 2.

Alignment Gap	[-]
Missing Information	[?]
Identical Site	[.]

In the first two input boxes, the symbols for alignment gaps and missing-information sites are specified. In the *Identical site* input box, the first sequence specifies the character. The alignment gap, missing-information sites, and identical sites symbols must be unique (see chapter 2).

For distance data

Upon selection of the *Distance* option, the program inquires about the format of the distance matrix: upper-right or lower-left triangular matrix.

Close Data
Alt+F4

It inactivates the current data set. Before doing so, it reconfirms your action. To save in-memory data editing, use *Export* command from the *Data Presentation* option.

Select OTUs

The OTUs can be deleted (or re-inserted) with this option obviating any need for the modification of original input data file. The *Select OTUs* command brings up a list of OTU labels. In this list, some labels are checked (✓) and others are not. Presence of a ✓ mark indicates that the OTU is included in the current data subset. You can delete or insert an OTU by using the **Del** or **Ins** keys. To have the changes made, press **Enter** (click on **OK**) or abort the current changes with the **Esc** key (clicking on **Cancel**).

Select Mode

This command is available for DNA and mRNA sequences only. It brings up a submenu where protein-coding or non-coding mode of analysis can be chosen.

```
Select Mode ▸  ┌──────────────────┐
               │ Protein Coding   │
               │ Noncoding        │
               └──────────────────┘
```

If DNA sequences code for proteins, the protein-coding mode must be used; otherwise, the noncoding mode is appropriate. In the coding mode analysis can be performed site-by-site as well as codon-by-codon; the noncoding mode allows only the site-by-site analysis.

Select Sites/Codons

Desired sites/codons can be selected with this option.

```
Select Sites/Codons ►  ┌──────────────┐
                       │ All          │
                       │ Domains...   │
                       │ Individual...│
                       └──────────────┘
```

All

All inserts all the sites (or codons depending on the mode) into the data subset for analysis. It is selected by default whenever a new data file is opened.

Domains

With this command a subset of original data containing up to 10 disjoint domains can be selected. This option can be useful if multiple exons exist for the gene studied.

The *Domain* command queries about the number of domains in the sequence first. You may enter any value from 1 to 10 (for more than 10 domains use *Individual* option). The location of domains in terms of their position in the input sequences is then entered in the next dialog box. All the domains must be non-overlapping and should be specified in the order from left to right.

Individual

This option is useful if the sites (or codons) to be analyzed are spread over the length of the sequence. Selection of this option produces a list of numbers corresponding to sites (or codons). Check (✓) mark before a site (or codon) number indicates its inclusion in the data subset. Use the **Ins** and **Del** keys to insert and delete the desired sites and codons. The changes made can be accepted with the **Enter** key and discarded with the **Esc** key.

Select Outgroups

Outgroups are selected with this option. None of the methods in MEGA is sensitive to outgroups, but they are used when displaying phylogenetic trees on the screen. A list box containing all current OTU labels is displayed on selection of this option. In this list, checked (✓) labels are designated as outgroups. Use the **Ins** and **Del** keys to include and exclude OTUs from the set of outgroups. To have the changes made, press **Enter** or click on **OK**; to abort, press **Esc** key or click on *Cancel*.

Edit OTU Labels

Often OTU labels present in the input file are required to be altered for publication purposes. This can be done in MEGA without affecting the input data file by using the *Edit OTU Labels* command. This command displays a list box dialog that contains a list of all the OTU labels. To edit an OTU label: press **Spacebar** (or double click) on the OTU label and modify the existing label in the input line box produced, and press

Enter to make the change. Labels of many OTUs can be edited at the same time. All the changes made can be accepted by pressing **Enter** finally or discarded by pressing **Esc.** MEGA does not require unique OTU labels, but you may find it difficult to identify OTUs individually.

Restore OTU Labels

OTUs labels are converted back to their original names as specified in the input data file.

Data Presentation **F4**

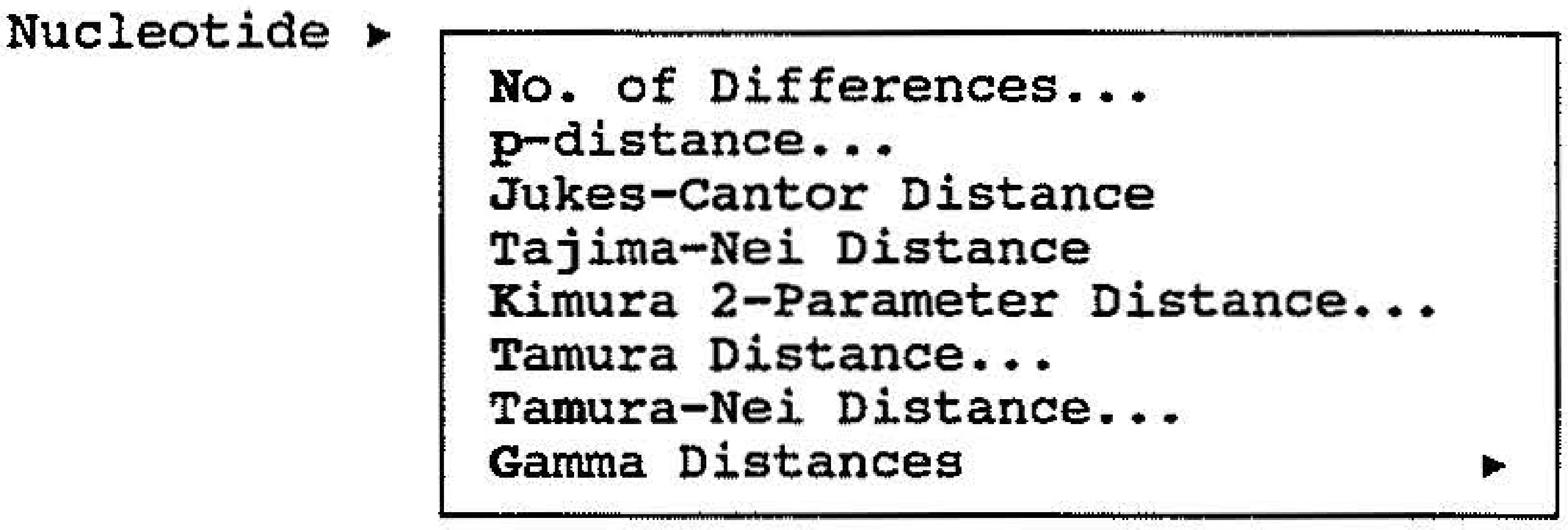
This module provides an assortment of useful functions. It contains commands for exporting current data subsets to files in various formats; highlighting variable, parsimony-informative, and two- and fourfold redundant sites; translating nucleotide sequences; and computing sequence statistics such as base compositions, codon usage, alignment gap frequencies, and the variability in sliding windows. Selection of *Data|Display Sequence Data* command displays "Current Data" window that is described in detail in chapter 6.

Distance menu **Alt+T**

Commands on this menu are used to select a distance measure from three submenus: *Nucleotide*, *Syn-Nonsynonymous*, and *Amino Acid*. Distances are computed and saved to files with *Compute Distances* command.

Nucleotide

This command is available only for nucleotide sequences. It presents a choice among seven distance measures. Selection of a distance measure does not automatically initiate the distance computation; it only selects the distance measure to be used. The following submenu appears on selection of the *Nucleotide* command:



```
Nucleotide ▸   No. of Differences...
               p-distance...
               Jukes-Cantor Distance
               Tajima-Nei Distance
               Kimura 2-Parameter Distance...
               Tamura Distance...
               Tamura-Nei Distance...
               Gamma Distances              ▸
```

For some distance measures, which are shown in the above menu with ellipsis (...), it is possible to estimate the numbers of transitional and transversional substitutions or differences separately. In this case, a dialog box appears and inquires whether transitions, transversions, all changes, or the transition/transversion ratio is to be estimated.

Syn-Nonsynonymous

This command is available for protein-coding nucleotide sequences only. It brings up a submenu containing three methods. After selecting a distance measure, you will be required to specify whether the synonymous or the nonsynonymous distances are to be computed. (See discussion in chapter 4.)

```
Syn-Nonsynonymous ▸  ┌──────────────────────────┐
                     │ No. of Differences       │
                     │ p-distance               │
                     │ Jukes-Cantor Correction  │
                     └──────────────────────────┘
```

Amino Acid

This command is enabled for amino acid sequences and protein-coding nucleotide sequences. Nucleotide sequences are automatically translated into amino acid sequences by using the selected genetic code table.

```
Amino Acid ▸  ┌────────────────────────────┐
              │ No. of Differences         │
              │ p-distance                 │
              │ Poisson Correction         │
              │ Gamma Distance             │
              └────────────────────────────┘
```

Genetic Code Table

This command produces a dialog box with four radio-buttons. A genetic code table can be selected by double clicking (or by using arrow keys) on the desired code table. The changes made are discarded by pressing **Esc** (or by clicking on **Cancel**); and accepted by pressing **Enter** (or clicking on **OK**).

```
┌────────────────────────────────┐
│ (•)  "Universal"               │
│ ( )  Drosophila mitochondrial  │
│ ( )  Mammalian mitochondrial   │
│ ( )  Yeast mitochondrial       │
└────────────────────────────────┘
```

Compute Distances
F7

This command is for estimating pairwise distances. It should only be used if you want the distances are to be written into a text file. It is not necessary to compute distances using this command for estimating phylogenetic trees, because the distances are obtained automatically by the *Construct Tree(s)* command.

Selection of this command brings up a dialog box with options such as precision, page width, codon positions, etc.

```
┌─ Gaps/Missing sites ──────┐
│ (•) Complete-Deletion     │
│ ( ) Pairwise-Deletion     │
└───────────────────────────┘
```

You may choose to exclude sites containing alignment gaps and missing-information sites either before the distance calculation begins or include them such that they are ignored only during the pairwise comparisons (see section **4.5**). Presence of • in ( ) means that the option is selected.

```
┌ Codon Positions ──────┐
│ [X] Include 1st base  │
│ [X] Include 2nd base  │
│ [ ] Include 3rd base  │
└───────────────────────┘
```

In the *Codon Positions* box you may select any combination of 1st, 2nd, and 3rd codon positions for analysis. Obviously choice of codon positions is available only if protein coding nucleotide sequence data is used to compute *Nucleotide* distances. You may change the inclusion status of any codon position by either pressing **Spacebar** or clicking on the desired codon position.

```
┌─ Estimate ───────────────┐
│ (•) Distances only       │
│ ( ) Distances and SE's   │
└──────────────────────────┘
```

Using this box, you may calculate distances only or both distances and their standard errors.

```
┌─ Write Distances ──────────┐
│ (•) Upper-right matrix     │
│ ( ) Lower-left matrix      │
└────────────────────────────┘
```

As you may notice, the *Write Distances* and *Write Standard Errors* boxes are similar. Distances and standard errors can be saved in a file either on the same side of the matrix or on the opposite sides.

Precision [4]
Page Width [80]

The *Precision* refers to the number of decimal places that are printed in the output, whereas the *Page width* is the width of the page. If the complete distance matrix does not fit in one page, it is printed in interleaved format. If you need to have a complete set of distances in one matrix, use a very large value for the page width (e.g., 1000 or so).

Phylogeny menu **Alt+P**

This menu contains options for selecting tree-building methods and conducting test of their interior branch lengths.

UPGMA

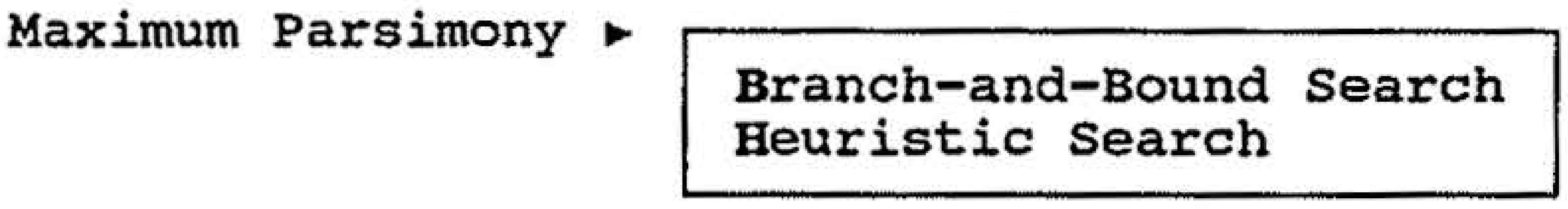
This command is used to select the **Unweighted Pair Group Method** with Arithmetic means (UPGMA) for phylogenetic reconstruction (see chapter 5 for details.)

Neighbor-Joining

The neighbor-joining (NJ) method is selected with this command (see chapter 5 for details).

Maximum Parsimony

In this method you have two options: the branch-and-bound search and the heuristic search (see chapter 5 for details).



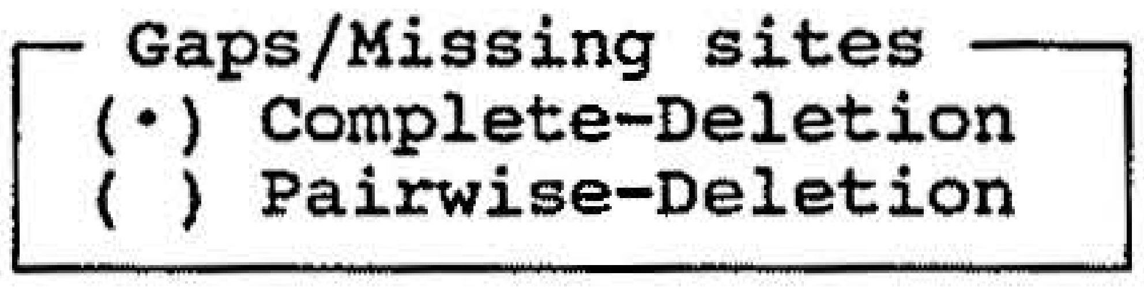
```
Maximum Parsimony ►
                      Branch-and-Bound Search
                      Heuristic Search
```

Construct Tree(s)
F8

This command initiates the reconstruction of a phylogeny by using the selected tree-building method and the distance measure (if applicable). It automatically computes the distances, if required, and presents the tree on the screen in the phylogenetic-tree editor (see chapter 6).

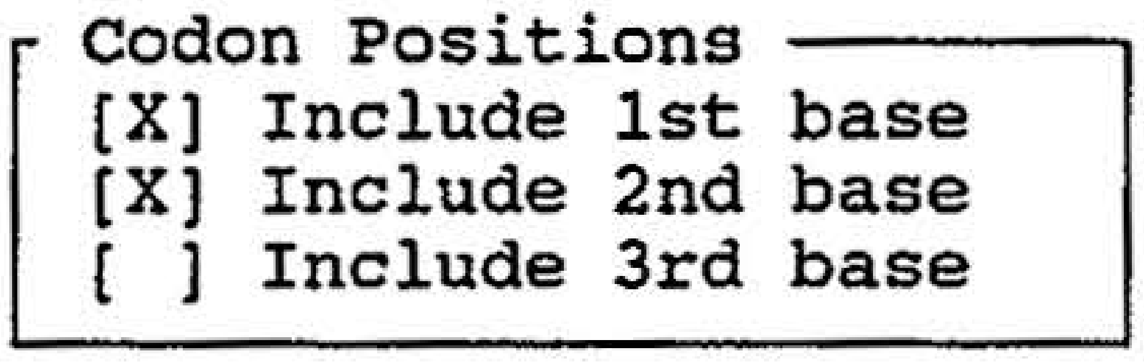
For distance matrix methods

If a distance matrix method (i.e., UPGMA or neighbor-joining) is used, a dialog box with the following options appears.



```
Gaps/Missing sites
(•) Complete-Deletion
( ) Pairwise-Deletion
```

It allows you to include/exclude sites containing alignment gaps and missing data. Presence of • in the parentheses ( ) indicates the selected option.



```
Codon Positions
[X] Include 1st base
[X] Include 2nd base
[ ] Include 3rd base
```

The codon positions box allows you to choose any combination of 1st, 2nd, and 3rd codon positions for analysis. Obviously, the choice of codon positions is available only if the coding nucleotide sequence is used. You can change the inclusion status of three codon positions by either pressing the **Spacebar** on the desired codon position or clicking on it.

For Maximum Parsimony methods

If a maximum parsimony method is used for constructing phylogenetic trees, some of the following options appear depending on whether the branch-and-bound or the heuristic search is selected.

```
┌─ Gaps/Missing sites ──────────────────────────────┐
│ [ ] Include sites containing alignment gaps        │
│ [ ] Include sites containing missing data          │
└────────────────────────────────────────────────────┘
```

It allows you to include/exclude sites containing alignment gaps and missing-information sites. Presence of X in the square brackets [ ] indicates the selected option.

```
[ ] Use alignment gap as an additional state
```

With this option, you can include gap symbol as an additional state (i.e., fifth state in the nucleotide sequences). This option has no effect if all sites containing alignment gaps are eliminated from the data.

```
┌ Codon Positions ──────┐
│ [X] Include 1st base   │
│ [X] Include 2nd base   │
│ [X] Include 3rd base   │
└────────────────────────┘
```

The codon positions box allows you to choose any combination of 1st, 2nd, and 3rd codon positions for analysis. Obviously, the choice of codon positions is available only if the coding nucleotide sequence is used. You can change the inclusion status of three codon positions independently by either pressing the **Spacebar** on the desired codon position or clicking on it.

```
Search factor          [ 2  ]
Transition step        [ 1  ]
Maximum trees          [100 ]
```

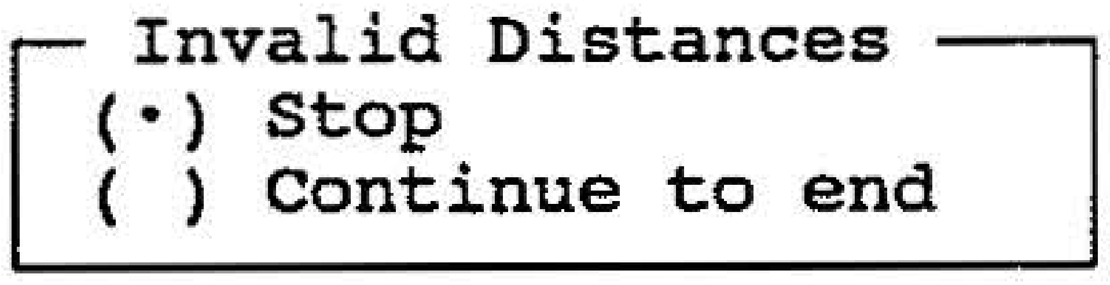
These two options come up only if the heuristic search is requested. By default, a search factor of 2 is selected that is applied from the first step of OTU addition to the core tree in the search procedure (see chapter 5). The *Maximum trees* option becomes useful if you expect that there are many MP trees.

Bootstrap Test
Alt+F8

The bootstrap test is used for assessing the reliability of a tree obtained (see chapter 5). MEGA provides a bootstrap test for UPGMA and the NJ method only. In response to the *Bootstrap Test* command, a dialog box appears with various options. Many options in this dialog box are similar to the ones described above for the *Construct Tree(s)* command. Therefore, only the options that are not described before are discussed below.

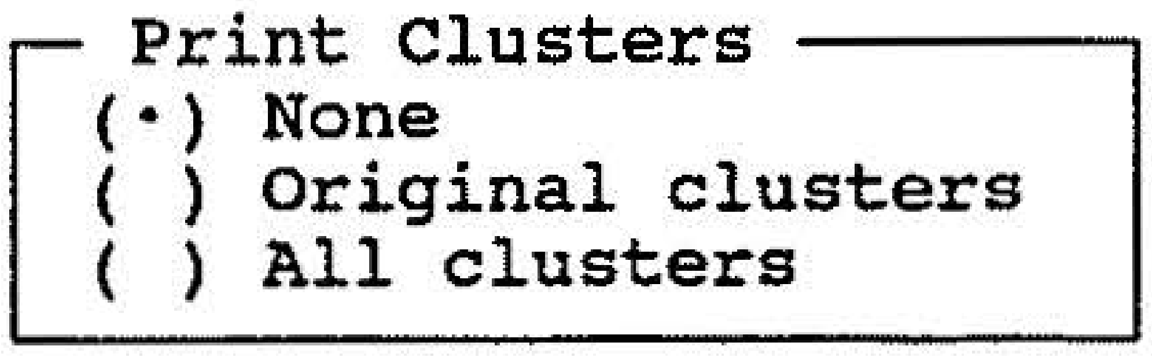
```
Replications       [500  ]
Random seed        [785  ]
```

With the *Replications* option, you can specify the number of replications desired for the bootstrap test. The default value for this option is 500. Since the bootstrap test requires the generation of random numbers for the resampling of data, MEGA generates pseudorandom numbers through a series of three linear congruential generators with effectively infinite period for most practical purposes. You initialize this random number generator by entering an arbitrary number in the *Random seed* option. Example, 785.



```
Invalid Distances
(•) Stop
( ) Continue to end
```

In distance matrix methods, the distances are calculated under a model that provides a distance estimation formula. Some of these formulas are not applicable outside a certain range of observed differences. If a bootstrap replication generates data for which some pairwise distances cannot be computed, the UPGMA or NJ tree cannot be constructed for that replication. Using the *Invalid Distances* option, you can choose either to abort the bootstrap test as soon as an invalid distance is encountered or to continue the bootstrap test by neglecting all replications generating invalid distances.



```
Print Clusters
(•) None
( ) Original clusters
( ) All clusters
```

The *Print Clusters* option allows you to write the frequency of either all the groups (clusters) in the starting tree that is constructed using all the data (*Original clusters*) or all the clusters generated in the bootstrap procedure (*All clusters*), including the original clusters, to a specified file. The clusters are written in binary format by using 1 for presence of an OTU in a cluster and 0 for absence. This information is written to the file that was specified before the bootstrap test was started.

Standard Error Test
Alt+F7

This command is to conduct the standard error test for interior branch lengths of a neighbor-joining tree. This option is available only for nucleotide sequence data whenever *p*-distance, Jukes-Cantor distance, or Kimura's 2-parameter distance is used (see chapter 5).

In this test, MEGA removes all sites containing alignment gaps and missing-information sites (*Complete-Deletion* option) from the sequence data. If the nucleotide sequence data is used in the coding mode, the choice of codon positions is available.

Window menu
Alt+W

The Window menu contains commands to close, move, and perform other window-management commands. Most of the windows in this program have all the standard window elements, including scroll bars, a close box, and zoom icons.

Resize/Move
Alt+F9

Choose this command to change the size and position of the active window.

Size

If you press **Shift** with an arrow key, the size of the active window will be altered. Once you've adjusted its size or position, press **Enter**. If a window has a re-size corner, you can drag that corner to resize the window.

Move

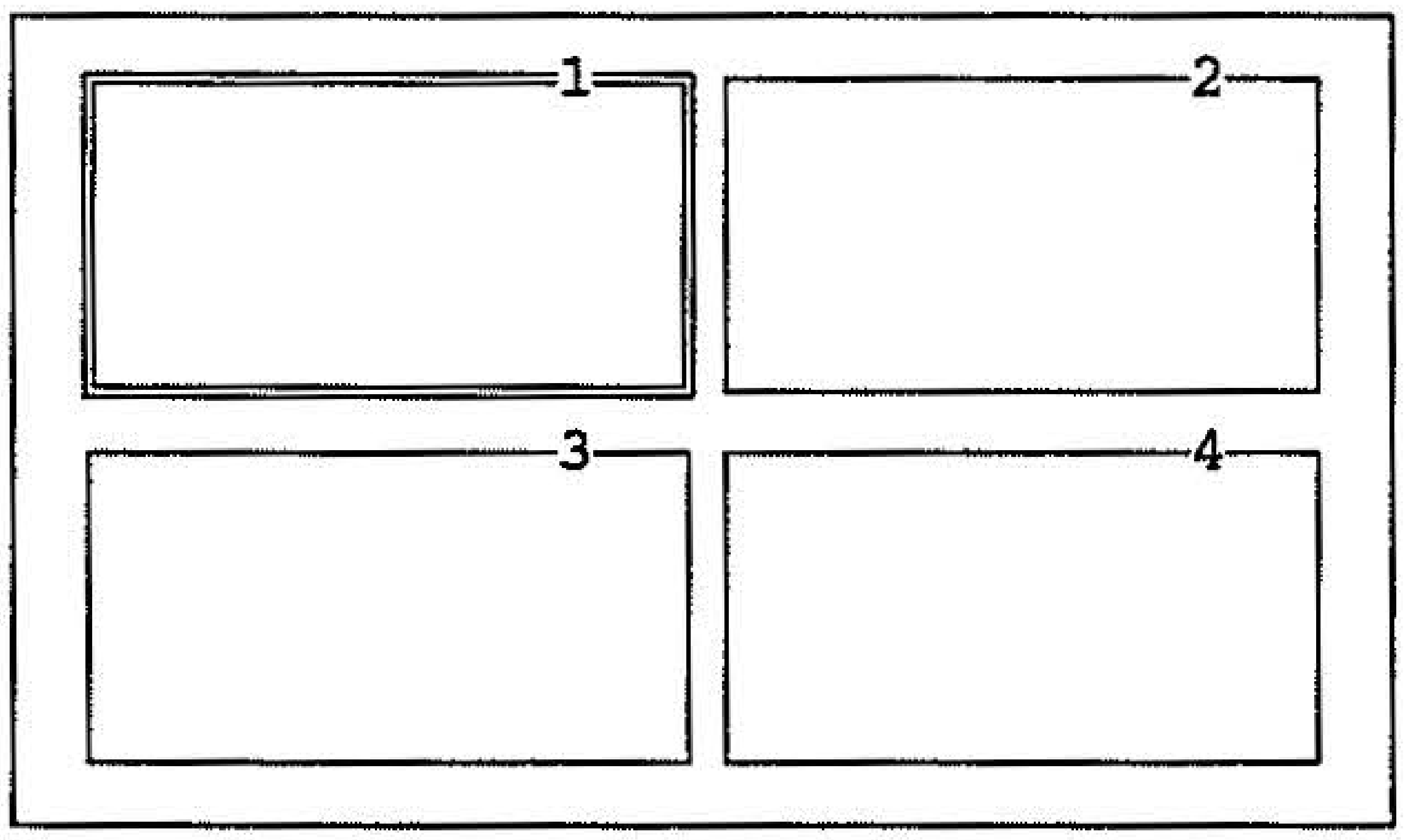
When the *Window|Re-size/Move* is chosen, the active window moves in response to the arrow keys. Press **Enter** after the window has been moved. With the mouse, a window is moved by dragging its title bar with the mouse.

Zoom
F9

Choose *Zoom* to resize the active window to the maximum size. If a window is already zoomed, this command restores it to its previous size. You may also double-click on the zoom-icon on the window's title bar to zoom or un-zoom.

Tile

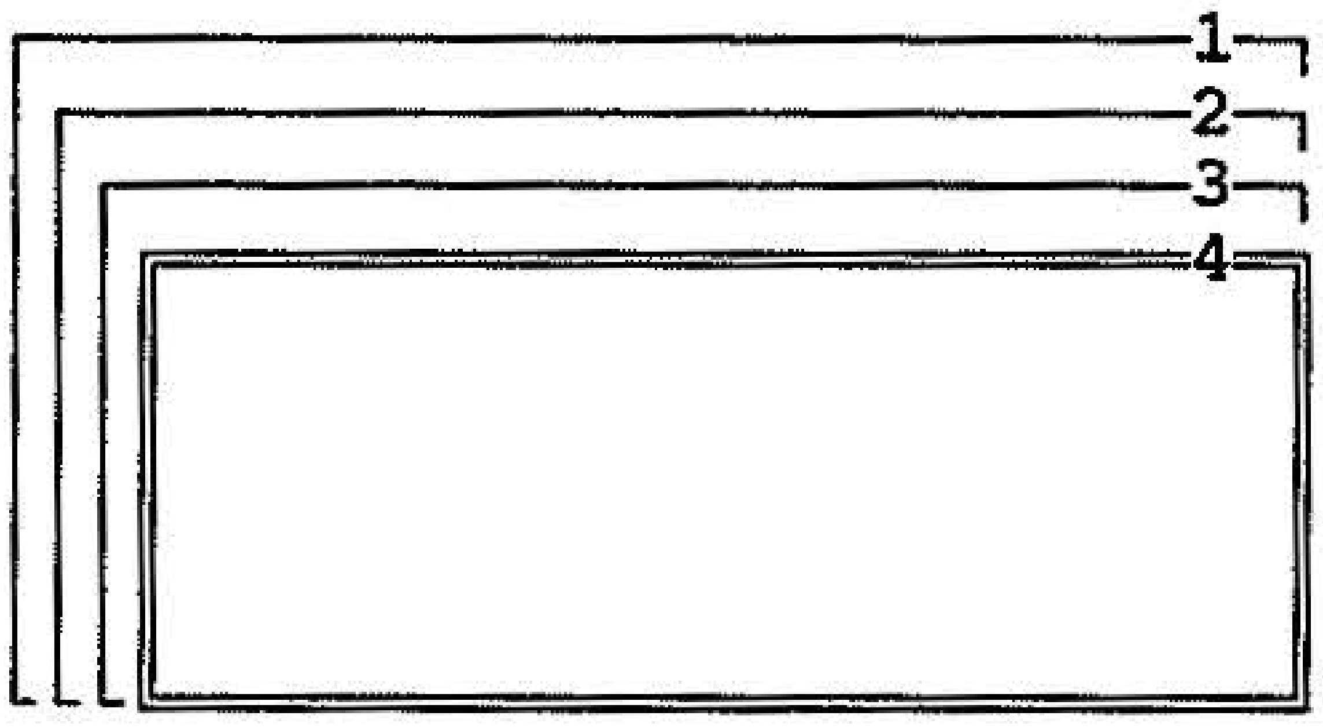
The *Tile* command arranges all the windows on the screen in the following manner:



Tiled Windows

Cascade

The *Window|Cascade command* stacks all the windows on desktop as shown below.



Cascaded Windows

Next
F6
Choose *Next* to cycle forwards through the windows on the desktop.

Previous
Alt+F6
Previous cycles backwards through the windows on the desktop.

Close
Alt+F3
Choose *Close* to remove the active window. The Close window icon on the upper right corner can also be used to close any window.

9

Error Messages

Errors are reported with a unique three digit identification number, a brief description, and, if applicable, the place of error that gives the line and the column position in the input data file. If MEGA is not installed from the master diskette(s) (i.e., copied from someone else's computer), only the error numbers may be displayed, you must request MEGA from the authors to obtain the master diskette(s).

Abnormal program termination

If memory is insufficient to run the MEGA program, this error will occur. Please see error number 001. If the problem persists, contact the authors.

001 RAM memory is exhausted

IBM personal computers run under the DOS operating system that allows only 640KB of RAM. DOS, other device drivers, and the MEGA program occupy much of this memory, and a small amount of memory is left for use during analysis. Look at the lower-right corner of your screen. If you think that the amount of memory shown there is enough, try the same command again. If the same message appears again and the amount of memory indicated is less than 200KB, try some of the following remedies.

1. If you have a version of DOS earlier than 5.0, upgrade it to DOS 5.0. Load DOS in high memory area to free extra 100KB memory. Details for loading DOS to high memory area are found in most of the DOS reference manuals.

2. Memory resident programs such as virus scanners and DOS clocks use a large amount of RAM memory. Removing these programs will give you additional memory (30-50KB). The procedure to remove such

memory resident programs is described in the DOS reference manual.

3. Device drivers for printing, networking, etc., take up some RAM memory, and computers running network software do not have much free RAM. You may free some memory by removing such programs.

002 Specified input data file is not found

Check the spellings of the input file name. If it is correct, then check that you are in the correct working directory. If not, use the *File|Change Dir* command to go to the appropriate directory.

003 Unexpected end of input data file

While reading input data file, more information was expected to follow, but the input file suddenly ended. Examine the contents of the data file with the *File|Browse* command and check the file. Does your input file contain non-ASCII characters?

004 Temporary file could not be created

MEGA creates many intermediate files to protect the original data. Make sure that your hard disk has at least 1MB of free space to store these files. If not, clean up your hard disk to free at least 1MB.

005 Numbers fall outside the valid range

All integer and real numbers entered in MEGA must be inside the predefined range for a particular option. If you have reasons to believe that some of these bounds are not justified, write to us about the reasons so that the range can be modified.

006 Invalid value

An invalid character or number has been entered.

007 User terminated the process

This message informs that the user terminated the process.

011 "TITLE" is not found in input data file

Check if your data file contains keyword "Title" on the second line. All input data files must include the *Title* keyword on the second line following the mega format specifier, *#mega*. (Read chapter 2 for more information.)

012 First OTU label must have preceding # sign

In MEGA, every OTU label should be prefixed with # sign. Check the beginning of the first OTU label. (Read chapter 2 for examples of input file.)

013 OTU label did not end

A blank space, tab, or a new-line should separate the OTU label and the sequence. Also, make sure that only one OTU label is written on any line.

014 Only one sequence is detected

Only one sequence is found in the input file. It may be because of the # sign missing in the second OTU label or because of the absence of a blank, a tab, or a newline. If you have only one sequence to analyze and you are interested in using *Data|Data Presentation* utilities, then simply duplicate the sequence in the input data file.

015 Invalid character encountered in data file

A character that is neither a valid nucleotide base (or amino acid residue) nor a special character for missing, identical, and alignment gap symbols is present. Examine the character at the place of error indicated. (For more information, consult chapter 2.)

016 No identical-site symbol is permitted in the first sequence

The first sequence should not contain any identical-site symbols because this symbol is used in the following sequences and is resolved in reference to the homologous site in the first sequence.

017 Last sequence seems incomplete

In the input data file all the sequences must be aligned and of equal length. Check if the last sequence in the file is of a different length.

018 Sequences are of unequal lengths

In the input data file, all the sequences must be aligned such that they are of equal length. Check the sequence data at the place of error.

019 Chosen data type is not implemented

Please consult the authors with appropriate information given in chapter 1.

020 Symbols for missing-information-, gap- and identical-site are not unique

Missing-information, gap, and identical sites symbols must be unique.

021 End of comment is missing

Comments are written like quotations within a pair of double quotes (e.g., "this is a comment"). Check if a double quote (") is missing.

022 Incomplete sequence encountered

All sequences must be of equal length in the input data file. Check data close to the place of error specified.

023 Vertical tabs are not permitted

Remove all vertical tabs from your data file.

024 Corresponding OTU labels must be identical in different data blocks

In interleaved sequences, all the OTUs must be presented in every block in the same order and with the same OTU labels. Check the OTU labels and their order in the input data file.

025 The number of OTUs is different in different blocks of interleaved sequences

In interleaved sequences, all the OTUs must be present in every block in the same order (see example in chapter 2). Note that blocks of interleaved data must be separated by at least one blank line and that the sequence data for different OTUs must be present on consecutive lines without any blank lines between them in every block.

026 One of the distance values read is invalid

The distance values in the input file must be positive or 0. The presence of negative values and other non-numeric character will result in errors. **Do not** use the scientific format for real numbers (1.24E2, etc.).

027 Input data must contain at least two OTUs

MEGA is designed for comparing different OTUs. So, there must be at least two entities for comparison. If you have only one sequence to analyze and you are interested in using *Data|Data Presentation* utilities, simply duplicate the sequence in the input data file.

028 New data file cannot be activated

Please report it to the authors with all the relevant information (chapter 1).

029 #mega format specifier missing

The very first line in the data file must contain *#mega* format specifier. This identifier is required to indicate that the data file is prepared for MEGA.

030 Data file contains too many OTUs

This error message can appear because of several reasons. First, the upper limit of the number of OTUs that can be read by MEGA is 500, so the data file should not contain more than 500 OTUs. If the sequence data in the file is in the interleaved format and MEGA attempt to read the data file using the *noninterleaved* option, this error may occur. Also, if your data file has no blank line between the blocks of interleaved sequences, this error will occur.

031 Error occurred during distance calculation

Please report this error to the authors with the information requested in chapter 1.

032 No distance type is selected, so distances cannot be calculated

Distance calculation is a two step process—first, a distance calculation method is selected, and then the *Distances|Compute Distances* command is invoked. For reconstructing phylogenies, select a distance estimation method, select a tree-building method, and call the *Phylogeny|Make Tree(s)* or

Phylogeny|Bootstrap command.

033 Failure in estimating distances

Distances are calculated under a model that provides distance estimation formula. Most formulas are only applicable for a certain range of observed difference. For instance, if the proportion (p) of nucleotide differences between two sequences ≥ 0.75, the argument in the log term of Jukes-Cantor's distance becomes ≤ 0, and the distance estimate is no longer obtainable. In MEGA, such invalid distances are shown with an '*'.

After computation of distances, the tree-building process is aborted. In this case, you may use the *Distance|Compute Distances* command and output distances to a file to identify OTUs that produce invalid distances. Remove these OTUs from the data set with the *Data|Select OTUs* command.

034 Program lost some important information required for internal use

Please report to the authors with appropriate information.

035 Stop codon(s) encountered in protein-coding sequences

Coding sequence should not contain any stop codons. This error may be caused by the use of an incorrect genetic code table. Select the appropriate genetic code table with the *Distance|Genetic Code Table* command and examine the sequence data with the *Data|Data Presentation* command in translation mode.

041 Phylogenetic trees cannot be reconstructed

Please consult the authors.

042 Less than 3 OTUs detected during phylogenetic reconstruction

Phylogenetic reconstruction is meaningless for a data set with only two OTUs.

043 No bootstrap replications specified

The message is obvious.

044 Bootstrap test is not available for the specified tree building method

MEGA does not provide the bootstrap test for the MP method.

045 None of the bootstrap replications produced valid results

This means that your data set is not appropriate for a bootstrap test.

051 No site/codon in data subset

Inspect the current data subset using the *Data|Data Presentation* command. Use the *Data|Select Sites/Codons* command to include desired data to the current data subset for analysis.

052 No OTU found in the currently active data set

Somehow all the OTUs from the data subset have been deleted. Use the *Data|Select OTUs* command to include some OTUs.

053 An error occurred during preparation of data for analysis

This error usually occurs due to the shortage of memory to store the data. Please refer to error number 001.

054 Protein-coding or noncoding mode are not specified

Please report this error to the authors.

055 No parsimony informative sites found in the sequence data

In the current data set none of the sites is informative for constructing an MP tree.

056 The data contain just one site (codon)

It is not possible to conduct bootstrap test if there is only one data site (codon).

057 Number of parsimony informative sites are less than number of OTUs

MEGA does not attempt to build a parsimony tree if the number of parsimony informative sites are less than the number of OTUs, because it is impossible to resolve the phylogeny in this case and many MP trees exist. Choose only representative OTUs, and delete others.

061 Device drivers may be missing

For printing phylogenetic trees on printers or for previewing them in graphic environment on the screen, MEGA comes with its own device drivers. This message indicates that it failed to find them in the appropriate directory. If you copied MEGA from someone else's computer, or if MEGA is not installed properly from the master diskette(s), this problem may arise. Please re-install MEGA from master diskette(s) properly. If the problem persists, contact the authors.

062 Tree drawing initialization error
063 Tree drawing error

Before a phylogenetic tree is drawn for previewing and printing, it is drafted in the vector format on an intermediate file. Apparently this file could not be initialized in the present case. Please report the error to the authors. Also see error number 081.

064 Trees cannot be previewed since there is no graphics capability

The user-interface is implemented in text-mode, but the tree image can be previewed before printing if some graphic capability is available. The EGA, VGA, and Hercules monitors are supported for this purpose.

071 Tree drawing file is missing

Before a phylogenetic tree is drawn for previewing and printing, it is drafted in the vector format on an intermediate file. This message indicates that this file has been lost. Try to use the print command again. If this does not help, report the error to the authors.

072 Printer cannot be opened for printing

Printers are usually connected to computers through PRN, LPT1, LPT2 or LPT3 ports. MEGA uses PRN to send data to the printers. Somehow this PRN port could not be opened for printing. Check printer connections.

073 I/O read error during tree printing
074 I/O write error during tree printing
075 Temporary file error

No free space on the disk. Please refer to error numbers 004 and 081.

076 Drawing file is corrupted

Before a phylogenetic tree is drawn for previewing and printing, it is drafted in the vector format on an intermediate file. Apparently this file has been corrupted. Try to use the print command again. If this does not help, report the error to the authors.

077 Unknown error occurred
078 Unknown error occurred
079 Unknown error occurred

The cause of the error is unknown, but the errors are detected by one of the error checking routines. You may see error number 081. If the problem persists, please contact the authors.

080 Drawing file error

Refer to error number 076.

081 Out of memory during tree drawing

Not enough memory is available for drawing the tree. Tree printing routines require at least 120KB (and more) of memory for drawing and printing. Look at the lower right corner of your screen. If the amount of memory available is less than 120KB, refer to the error number 001.

082 Cannot write to the printer

This is an input/output error. Use a different printer to print the phylogenetic tree. Also, report this error to the authors.

083 Device drivers are corrupted

For printing phylogenetic trees on various kinds of printers and previewing them in graphic environment, MEGA comes with many device drivers. This message indicates that the desired device driver has been corrupted. This problem can be solved by reinstalling MEGA to the computer using the master diskettes.

084 Printing aborted by the user
The **Esc** key was pressed to abort tree printing.

085 Bad drawing file
See error number 076.

086 Tree cannot be output in PCX format
The PCX file format is for storing the tree in a file that may be modified in the PC PaintBrush (Microsoft Windows) program afterwards. An input-output error occurred while writing this file. Please report your finding to the authors.

087 EPS output function failed
The EPS output is used to create a PostScript file or to send a PostScript file to a PostScript printer for printing. An input-output error occurred during this process. Please report your finding to the authors.

088 Polygon can not be drawn due to insufficient memory
089 Graphics mode cannot be initialized
Please report it to the authors.

090 Specified device driver is missing
091 Desired font file is missing
To remove this error, reinstall MEGA from the master diskette(s). If the problem persists, contact the authors.

092 Printer is either off line or paper is out
The printer is not responding to the program. Please check the connections.

093 Font file seems to be corrupted
To remove this error, reinstall MEGA from the master diskette(s). If the problem persists, contact the authors.

Appendix A: Functions in MEGA

The following list shows various computational and editing functions available in MEGA.

Input

Input data:
- DNA sequences
- RNA sequences
- Amino acid sequences
- Distance matrices

Input formats:
- Interleaved sequences
- Non-interleaved sequences
- Upper-triangular distance matrix
- Lower-triangular distance matrix

Choice of:
- Alignment gap symbol
- Missing-information site symbol
- Identical site symbol

In-memory data editing features

Selection:
- Desired OTUs
- Domains of sequences
- Individual sites and codons
- Codon positions
- Exclude/include missing information sites
- Exclude/include alignment gap sites

Edit OTU labels
Restore OTU labels

Sequence data presentation

Highlight:
- Variable sites
- Parsimony-informative sites

- Two-fold redundant sites
- Four-fold redundant sites

Translate:

- Translation of nucleotide sequences into amino acid sequences

Output:

- Formats:
  - MEGA
  - PAUP
  - PHYLIP
  - Publication
- Data subsets:
  - Only variable sites
  - Only parsimony-informative sites
  - Amino acid sequences translated
  - Codon positions
- Sequence statistics:
  - Nucleotide and amino acid frequencies
  - Nucleotide pair frequencies in pairwise comparisons
  - Insertion-deletion frequencies
  - Codon usage frequencies
  - Relative synonymous codon usage (RSCU) values
  - Variable sites in overlapping segments
  - Variable sites in nonoverlapping segments

Distance estimation

Nucleotide substitutions

- Quantities:
  - Number of nucleotide differences
  - Nucleotide substitutions
  - Transitional substitutions
  - Transversional substitutions
  - Transition/transversion ratio
- Distance measures:
  - *p*-distance
  - Jukes-Cantor distance
  - Kimura 2-parameter distance
  - Tajima-Nei distance
  - Tamura distance
  - Tamura-Nei distance
  - Gamma distances
    - Jukes-Cantor model
    - Kimura 2-parameter model

Tamura-Nei model

Synonymous-nonsynonymous substitutions
- Genetic code tables:
  - "Universal"
  - Mammalian mitochondrial
  - *Drosophila* mitochondrial
  - Yeast mitochondrial
- Computation:
  - Synonymous substitutions
  - Nonsynonymous substitutions
  - Average distances for all pairwise comparisons and standard errors

Amino acid substitutions
- Distance measurers:
  - Number of amino acid differences
  - *p*-distance
  - Poisson-correction distance
  - Gamma distance

Distance output:
- Control on:
  - Page size
  - Precision for distance output
  - Distance ± standard error formats

Tree building and test

Methods:
- Neighbor-joining (NJ)
- UPGMA
- Maximum parsimony (MP):
  - Branch-and-bound search
  - Heuristic search

Statistical Tests:
- Bootstrap test:
  - Neighbor-joining
  - UPGMA
- Branch length test:
  - Neighbor-joining

Phylogeny editing:
- Tree re-rooting
- Swapping and flipping branches

Consensus tree
Condensed tree

Phylogeny printing:
Various printers
Multiple page printouts
Choice of fonts
Choice of orientation
Choice of page size
Tree preview

General functions

File browsing
File editing
Exiting to DOS temporarily
Context-sensitive Helps
Error messages

Appendix B: Common Questions and Answers

The following is a list of questions that are commonly asked by MEGA users. These questions are given in different categories to allow easy reference. This list will be updated as new questions arise. A user can obtain an updated list of these questions and answers either by writing to the authors or from the bionet.molbio.evolution newsgroup.

General Questions

Does MEGA use extended memory?

No, MEGA does not use any extended or expanded memory. Therefore, it can be used on any computer that has basic 640KB memory.

MEGA frequently crashes on my computer. Why?

This may be due to lack of memory to run MEGA. Please refer to the *Abnormal Program Termination* section in chapter 9. It is also possible that your computer is not a true IBM-compatible.

I am stuck, how can I get help?

MEGA has a sophisticated on-line context sensitive help system where help is available by pressing the *F1* key. This manual also discusses statistical methods included in MEGA, and you should go through it at least once. For further technical assistance, refer to section **1.7**.

Input Data

Does MEGA accept distance data files as input data?

Yes. See section 7.1.

The Input formats of MEGA differ from those of other existing programs. Why?

The programs such as PAUP and MacClade can be used for analyzing morphological data. Therefore, their input file formats are more complicated. MEGA is designed exclusively for studying molecular evolution, so that simplified input file formats are used.

Can MEGA read amino acid sequences written in three-letter amino acid codes?

No, we plan to include this option in a later version of MEGA.

Why can't I use 'n' (or 'N') to designate missing-information sites?

Since MEGA provides automatic translation of coding nucleotide sequences, conflicts arise in the definition of valid symbol for missing-information sites and one letter code for the amino acid asparagine.

How can I insert blanks in the OTU labels?

Please refer to section **2.1.2**.

How long can a comment be?

A comment can be as long as you want, and it may run on several lines.

Where can I write comments in a distance input file?

In data files containing distance matrices, comments can only be placed after the TITLE line and before the OTU labels.

Why can't MEGA read its own distance output files directly?

For making phylogenetic trees from sequence data, MEGA automatically computes the distance selected and builds a tree. In this process, the accuracy of computation is maintained up to 15 places of decimal. If you print out distances computed in a file and feed them back to MEGA, the accuracy of this computation is reduced to 6 or 8 places of decimal. So, we do not recommend such a procedure.

Data Editing

Can I select OTUs from my sequence data for analysis?

Yes, use the *Data|Select OTUs* command.

How do I analyze different domains (exons) separately in a sequence data?

MEGA lets you choose domains of sites (and codons) for analysis. To do so, use the *Data|Select Sites/Codons* command. Also, see the tutorial to learn more about in-memory data editing options in MEGA.

I want to extract four-fold redundant sites. How can I do this?

The *Data|Data Presentation* command displays the sequence data on the screen. In this data set you can highlight all common four-fold redundant sites by pressing key **4** if you are using the data in coding mode. Then note down the site number of each of these sites. Once you have done this, exit the data presentation window by pressing **Esc**. Now go to the *Data|Select Mode* command and choose the Noncoding mode. After this, use the *Data|Select Sites/Codons* command with the *Individual* command. A list of site numbers will be produced. In this list, press the **Del** key on all numbers except those that correspond to the four-fold redundant sites. Now go to the *Data|Data Presentation* command and use the *Export* command to output these data.

Will the in-memory data editing command modify my original input file?

No, your input data file will remain intact.

Distance Calculations

Why can't I select synonymous-nonsynonymous and amino acid distances for my nucleotide sequence data?

For nucleotide sequences, syn-nonsynonymous and amino acid distances can only be computed for protein coding DNA sequences. So, if your sequence data comes from a protein-coding gene, use the *Data|Select Mode* command to choose the *Protein-Coding* mode. You will then be able to select these distances.

Why can't I use the Compute Distances command?

The *Compute Distances* command is enabled only if a distance estimation method has been selected. So, choose an appropriate distance calculation method before trying the *Compute Distances* command.

Does MEGA compute the parameter a for Gamma distance?

No. It must be estimated by some other methods (e.g., Tamura and Nei 1993, Wakeley 1993).

How can I select desired positions in codons?

If your nucleotide sequences code for a protein, use the *Data|Select Mode* command to choose the *Protein-Coding* mode. You will then be able to select codon positions from the options box when you estimate nucleotide substitutions.

Does MEGA report the presence of stop codons in the coding region? How can I eliminate this error?

Check if you have selected an appropriate genetic code table by using the *Distances|Genetic Code Table* command. If not, choose the correct one. If this is not the problem, use the *Data|Data Presentation* command. This will display all the sequence data on the screen. In this window, press the key **T**. The nucleotide sequences will then be translated into amino acid sequences. If you see any '*' (stop codon symbol) characters in your sequence data except at the very end, there is a problem in the input data file. Check your data file and make sure that the data have been written properly. By contrast, if a '*' appears at the end of the sequence, as is expected for the end of a coding sequence, exclude the last codon using the *Select Sites/Codons* command from the *Data* menu.

I want to write the distances and standard errors in the distance ± standard error format. How can I do that?

This can be done by simply choosing the same side for printing the distances and standard errors in the options box that appears in response to the *Distance|Compute Distances* command.

My distance matrix is fragmented. However, I want to write one complete distance matrix?

Specify a large number (such as 1000 or more) for the page size. This will ensure the output of the complete distance matrix in one block.

Why can't MEGA read its own distance output files directly?

Please see the Input Data section above.

What does '' mean in the distance output file?*

The presence of '*' indicates that it was not possible to compute the distance for the given pair of sequences. If you are computing the transition/transversion ratio, the '*' symbol will appear whenever the number of transversions estimated is 0.

Tree Building

Why can't I use the Construct Tree(s) command?

The *Construct Tree(s)* command is enabled only if a tree-making method is selected. Therefore, choose a tree reconstruction method before trying the *Construct Tree(s)* command.

How can I select the desired positions in codons?

If your nucleotide sequences code for a protein, use the *Data|Select Mode* command to choose the *Protein-Coding* mode. You will then be able to select the codon positions from the option box when you construct trees.

Does MEGA report the presence of stop codons in the coding region? How can I eliminate this error?

See the **Distance Calculations** section above for this error.

How can I build trees if invalid distances occur?

You have to remove the OTUs that produce invalid distances. In this case, first you have to identify the OTUs that cause invalid distances. This can be done by using the *Distance|Compute Distances* command. In the output file invalid distances will be marked with the * symbol. Using *Data|Select OTUs*, remove all these OTUs from the data set and then use the *Phylogeny|Construct Tree(s)* command. It is also possible to eliminate invalid distances by computing other distance measures such as p-distance.

How do I conduct the standard error test?

This test can be conducted only for the branch lengths of an NJ tree. The *Standard Error Test* command is enabled only if you have selected the NJ tree building method. You are also required to select one of the three distances: *p*-distance ($s+v$), Jukes-Cantor distance, and Kimura's 2-parameter distance ($s+v$).

Branch length estimates are not given for the maximum parsimony trees. Why?
See section **5.5**.

Is the tree length of an MP tree obtained by MEGA different from that obtained by other programs?
In MEGA, the tree length of an MP tree is computed by using the parsimony-informative sites only. Other programs sometimes compute the tree length using all the variable sites, including non parsimony-informative sites. Try and use the *exclude non-parsimony-informative sites* option in those programs to compare the tree length.

How can I use a bootstrap test for an MP tree in MEGA?
MEGA does not provide the bootstrap test if maximum parsimony is used to construct a phylogenetic tree. Please use other programs for this purpose.

Is there any special algorithm for using outgroups in tree building?
No, but the phylogenetic tree editor uses the outgroups for rooting unrooted NJ and MP trees.

I want to get a listing of all the nodes in the tree and the branch lengths (and other information such as the standard errors of the branch lengths) to input in other tree-editors such as the one provided in the DISPAN program. How can I get that?
It is possible to get the listing of such a file from MEGA, but you have to be cautious and responsible for any mishandling that you may have. In the tree window, press the key **Z** and you will be asked for a filename to save this file.

What do the Search Factor and Transition Step options mean in the heuristic search of MP trees?
See section **5.5.2**.

What is the difference between a consensus tree and a condensed tree?
See section **5.6.4**.

Appendix C: Mathematical Notations and Abbreviations

λ	= Rate of substitution per site.
Θ	= G+C content
a	= Inverse of the coefficient of variation of a gamma distribution.
ASCII	= American Standard Code for Information Exchange
BCL	= Bootstrap confidence level
CP	= Confidence probability
d	= Number of nucleotide or amino acid substitutions per site.
$\hat{d}$	= Estimate of d.
g_i	= Nucleotide frequencies (i=A,T,C,G).
g_R	= Frequency of purines (A,G).
g_Y	= Frequency of pyrimidines (C,T).
IUPAC	= International Union of Pure and Applied Chemistry
L	= Tree length
L_U	= Tree length of a temporary MP tree
L_M	= Tree length of the MP tree
m	= Number of sequences (or OTUs)
n	= Number of nucleotides or amino acids compared.
n_s	= Number of transitional differences.
n_v	= Number of transversional differences.
n_d	= Total number of nucleotide or amino acid differences.
N	= Number of nonsynonymous sites in a sequence.
N_d	= Number of nonsynonymous differences.
OTU	= Operational Taxonomic Unit
p	= Proportion of nucleotide or amino acid differences.
P	= Proportion of transitional differences.
P_1	= Proportion of transitional differences between A and G.
P_2	= Proportion of transitional differences between T and C.
p_S	= Proportion of synonymous nucleotide differences.
p_N	= Proportion of nonsynonymous nucleotide differences.
Q	= Proportion of transversional differences.
R	= Transition/transversion ratio.
RSCU	= Relative Synonymous Codon Usage.
S	= Number of synonymous sites in a sequence.
S_d	= Number of synonymous differences.
$s(x)$	= Standard error of x.
$\hat{s}$	= Number of transitions per site.
$\hat{v}$	= Number of transversions per site.
t	= t-statistic
T	= Time of divergence for the pair of sequences compared.
$V(x)$	= Variance of the estimated parameter x.
x_i	= Search factor

Appendix D: Printers Supported

The following is a list of printers that can be used to print phylogenetic trees. If your printer does not appear in the list, check the printer manual and select a compatible printer. We have not tested the printing routines in MEGA for all these printers, but we believe that they will work. If you find difficulties in using a listed printer, please contact us at the address given on the inside page of the front cover.

Acer LP76
AEG Olympia NP 136SE
AEG Olympia NP 80-24E
AEG Olympia NP 80SE
ALPS Allegro 500
ALPS Allegro 500XT
ALPS ASP1600
Apple Person LaserWriter NT
Apple Person LaserWriter NT PostScript
Bezier BP4040
Bezier BP4040 PostScript
Brother HL-4
Brother HL-4PS
Brother HL-4PS PostScript
Brother HL-8
Brother HL-8e
Brother HL-8V
Brother M-1309
Brother M-1324
Brother M-1909
Brother M-1924L
Bull Compuprint 4/22
Bull Compuprint 4/23
Bull Compuprint 4/24
Bull Compuprint 4/40
Bull Compuprint 4/43
Bull Compuprint 4/54
Bull Compuprint 4/68
Bull Compuprint 970
C-Tech C-510
C-Tech C-515
C-Tech C-610 Plus
C-Tech C-610C Plus
C-Tech C-645
C-Tech ProWriter C-240
C-Tech ProWriter C-245
C. Itoh ProWriter CI-4
CalComp ColorMaster Plus 6603 PS
Cannon BJ-300 Bubble Jet
Cannon BJ-330 Bubble Jet
CIE CI-250 LXP
CIE CI-5000
Citizen 200 GX
Citizen 200 GX Fifteen
Citizen GSX-130
Citizen GSX-140
Citizen GSX-140 Plus
Citizen GSX-145
Citizen MSP-10
Citizen MSP-15
Citizen MSP-15 Wide Carriage
Citizen PN48
Dataproducts 9030
Dataproducts 9044
Dataproducts LZR 2450D
Dataproducts LZR 650
Dataproducts LZR 960
Dataproducts LZR 960 PostScript
Datasouth Performax
Datasouth XL-300
DEClaser 1150
DEClaser 1150 PostScript
DEClaser 2150
DEClaser 2250
EiconLaser
Epson DFX-8000
Epson EPL-6000
Epson EPL-7000
Epson EPL-7500
Epson EPL-7500
Epson FX
Epson FX Wide Carriage

Epson FX-850
Epson LQ Low Resolution
Epson LQ Low Resolution Wide Carriage
Epson LQ Wide Carriage
Epson LQ-1010
Epson LQ-1050
Epson LQ-200
Epson LQ-2500
Epson LQ-850
Epson LQ-860
Epson MX
Epson MX Wide Carriage
Everex Abaton LaserScript
Everex Abaton LaserScript PostScript
Everex LaserScript LX
Everex LaserScript LX PostScript
Facit B1200
Facit P6060
Fortis DM3215
Fujitsu DL3600
Fujitsu DL4600
Fujitsu RX7100 S/2
Fujitsu RX7100 S/2 PostScript
Fujitsu RX7100PS Plus
GCC BLP II
GCC BLP IIS
General Parametrics Spectra*Star 43
Genicom 3840
Genicom 3840E
HP DeskJet
HP DeskJet 500
HP DeskJet 500C
HP DeskJet Plus
HP LaserJet
HP LaserJet II
HP LaserJet IID
HP LaserJet III
HP LaserJet III PostScript
HP LaserJet IIID
HP LaserJet IIIP
HP LaserJet IIISi
HP LaserJet IIP
HP LaserJet IV PostScript
HP LaserJet Plus
HP PaintJet
HP PaintJet/XL
HP Rugged Writer
HP ThinkJet
IBM ExecJet
IBM LaserPrinter 10
IBM LaserPrinter 10L
IBM LaserPrinter 4019
IBM LaserPrinter 5e
IBM LaserPrinter 6
IBM LaserPrinter E
IBM PP Series II 2380
IBM PP Series II 2381
IBM PP Series II 2390
IBM PP Series II 2391
Kodak Ektaplus 7008
Kodak Ektaplus 7016 PS
Kodak Ektaplus 7016 PS PostScript
Kyocera F Series
Kyocera F-5000A
LaserMaster LM 1000
LaserMaster TrueTech 1000
LaserMaster TrueTech 1200
LaserMaster TrueTech 1200 PostScript
LaserMaster TrueTech 800/4
Mannesmann Tally 130/24
Mannesmann Tally 130/9
Mannesmann Tally 131/24
Mannesmann Tally 131/9
Mannesmann Tally 906PS
Mannesmann Tally 906PS PostScript
Mannesmann Tally MT661
Mannesmann Tally MT735
Mannesmann Tally MT82
Mannesmann Tally MT911 PS
Mannesmann Tally MT911 PS PostScript
Microtek TrueLaser
Microtek TrueLaser PostScript
Mitek 130T
Mitek 130T PostScript
NCR 6417-0101
NCR 6421-0201
NCR 6435
NCR 6436-0310
NCR 6436-0501
NCR 6436-0501 PostScript
NEC Colormate PS Model 40
NEC Pinwriter P3200
NEC Pinwriter P3300
NEC Pinwriter P5200

NEC Pinwriter P5300
NEC Pinwriter P6200
NEC Pinwriter P6300
NEC Pinwriter P9300
NEC Silentwriter
NEC Silentwriter2 290
NEC Silentwriter2 290 PostScript
NEC Silentwriter2 90
NEC Silentwriter2 90 PostScript
NEC Silentwriter2 990
NEC Silentwriter2 990 PostScript
NewGen Turbo PS/360
NewGen Turbo PS/360 PostScript
NewGen Turbo PS/480
NewGen Turbo PS/480 PostScript
Oce Graphics G5241-PS
Okidata Microline 380
Okidata Microline 390 Plus
Okidata Microline 391 Plus
Okidata Microline 393 Plus
Okidata Okilaser 400
Okidata Okilaser 820
Okidata Okilaser 840
Okidata Okilaser 840 PostScript
Okidata OL830
Okidata OL830 PostScript
Olivetti DM309E
Olivetti DM600S
Olivetti PG306
Output Duraline
Packard Bell BP9500
Panasonic KX-P1123
Panasonic KX-P1124
Panasonic KX-P1124i
Panasonic KX-P1180
Panasonic KX-P1191
Panasonic KX-P1624
Panasonic KX-P1695
Panasonic KX-P2624
Panasonic KX-P4420
Panasonic KX-P4450i
Panasonic KX-P4455
Panasonic KX-P4455 LaserPartner
Panasonic KX-P4455 LaserPartner PS
PCPI Laser Image 1030
Printware 720 IQ Professional II
Printware Pro-III
QMS ColorScript 100 Model 10p
QMS PS-2000
QMS PS-2000 PostScript
QMS PS-2210
QMS PS-2210 PostScript
QMS PS-410
QMS PS-410 PostScript
QMS PS-810 Turbo
QMS PS-810 Turbo PostScript
QMS PS-815
QMS PS-815 MR
QMS PS-815 MR PostScript
QMS PS-815 PostScript
QMS PS-820 Turbo
QMS PS-820 Turbo PostScript
QMS PS-825
QMS PS-825 PostScript
Qume CrystalPrint Express
Qume CrystalPrint Express PostScript
Qume CrystalPrint Publisher II
Qume CrystalPrint Publisher II PostScript
Qume CrystalPrint Series II
Qume CrystalPrint Super Series II
Seiko ColorPoint PSX Model 14
Seiko ColorPoint PSX Model 4
Seikosha BP 5780
Seikosha SL-90
Seikosha SP-2000
Sharp JX-9500
Sharp JX-9500H
Sharp JX-9500PS
Sharp JX-9500PS PostScript
Sharp JX-9700
Star LaserPrinter 4
Star LaserPrinter 4 StarScript
Star LaserPrinter 4 StarScript PostScript
Star LaserPrinter 8 II
Star NX-1020 Rainbow
Star NX-1500
Star NX-2410
Star NX-2415
Star NX-2420 Multi-font
Star NX-2420 Rainbow
Star Starjet SJ-48
Star XB-2420 Multi-font
Star XB-2425 Multi-font
Star XR-1000

Star XR-1020 Multi-Font
Star XR-1520 Multi-Font
Tandy DMP 135
Tandy DMP 136
Tandy DMP 2130
Tandy DMP 240
Tandy LP 950
Tektronics Phaser II PXi
Tektronics Phaser III PXi
TI 8930
TI MicroLaser
TI MicroLaser PS35
TI Microlaser XL PS35
Toshiba PageLaser 6

Appendix E: Other Computer Programs Available

The following computer programs for evolutionary studies are available free of charge from the Institute of Molecular Evolutionary Genetics, The Pennsylvania State University. They are written by the current or former associates of M. Nei for specific purposes. All the programs are for IBM and IBM-compatible personal computers. A request should be sent to M. Nei along with an appropriate DOS-formatted floppy diskette(s) as indicated in the following description.

DISPAN: Genetic *Dis*tance and *P*hylogenetic *An*alysis, version 1.1, 1993 (T. Ota). This program written in C is for computing genetic distances from gene frequency data (Nei, *Amer. Naturalist* 106:283-292, 1972; Nei *et al.*, *J. Mol. Evol.* 19:153-170, 1983) and for constructing phylogenetic trees (UPGMA and NJ trees). Bootstrap tests are available, and trees can be edited for publication. Send a formatted 720KB diskette.

METREE, version 1.2, 1993 (A. Rzhetsky and M. Nei). This computer program is written in C. It computes minimum evolution trees from DNA and amino acid sequence data and tests the statistical significance of topological differences and of the branch lengths of the minimum evolution tree (Rzhetsky and Nei 1992, 1993). Different distance measures may be used. Send a 720KB diskette.

RESTDATA, version 1.0, 1994 (T. Ota). This is written in C and is for estimating the number of nucleotide substitutions per site between two DNA sequences from restriction-site and restriction-fragment data (Nei and Li, *Proc. Natl. Acad. Sci. USA* 76:5269-5273, 1979; Nei and Tajima, *Genetics* 105:207-217, 1983) and for constructing UPGMA and NJ trees. Send a 360KB diskette. (This program will be available after January 1, 1994)

RESTSITE, version 1.2, 1991 (J. C. Miller). This C program is for estimating the average number of nucleotide substitutions per site within and between populations for the case where a large number of individuals are examined for many restriction enzymes (Nei and Miller, *Genetics* 125:873-879, 1990) and for constructing phylogenetic trees (UPGMA and NJ trees). Both restriction-site and restriction-fragment data can be analyzed. Send a 1.44MB diskette.

SEND, version 1.0, 1989 (L. Jin). It is a Microsoft FORTRAN program that estimates the average number of nucleotide substitutions per site within and between populations and their standard errors (Nei and Jin, *Mol. Biol. Evol.* 6:290-300, 1989). Both DNA sequence and restriction-site data can be analyzed. Send a 720KB diskette.

Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution* 9:678-687.

Tamura, K. 1994. Model selection in the estimation of the number of nucleotide substitutions. *Molecular Biology and Evolution* 11:154-157.

Tamura, K. and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* 10:512-526.

Tanaka, T. and M. Nei. 1989. Positive Darwinian selection observed at the variable-region genes of immunoglobulins. *Molecular Biology and Evolution* 6:447-459.

Tateno, Y., M. Nei, and F. Tajima. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. *Journal of Molecular Evolution* 18:387-404.

Tateno, Y., N. Takezaki, and M. Nei. 1994. Relative efficiencies of the maximum likelihood, neighbor-joining, and maximum parsimony methods when substitution rate varies with site. *Molecular Biology and Evolution* 11:261-277.

Thomas, R. H. and J. A. Hunt. 1993. Phylogenetic relationships in *Drosophila*: A conflict between molecular and morphological data. *Molecular Biology and Evolution* 10:362-374.

Uzzell, T. and K. W. Corbin. 1971. Fitting discrete probability distribution to evolutionary events. *Science* 172:1089-1096.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes, and A. C. Wilson. 1991. African populations and the evolution of human mitochondrial DNA. *Science* 253:1503-1507.

Wakeley, J. 1993. Substitution rate variation among sites in hypervariable region I of human mitochondrial DNA. *Journal of Molecular Evolution* 37:613-623.

Williams, P. L. and W. M. Fitch. 1990. Phylogeny determination using dynamically weighted parsimony method. In R. F. Doolittle, ed., *Methods in Enzymology, Vol. 183, Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences,* pp. 615-626. Academic Press, New York.

Zharkikh, A. and W.-H. Li 1992a. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Molecular Biology and Evolution* 9:1119-1147.

Zharkikh, A. and W.-H. Li. 1992b. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock. *Journal of Molecular Evolution* 35:356-366.

Britten, R. J. 1993. Forbidden synonymous substitutions in coding regions. *Molecular Biology and Evolution* 10:205-220.

Brown, W. M., E. M. Prager, A. Wang, and A. C. Wilson. 1982. Mitochondrial DNA sequences of primates: Tempo and mode of evolution. *Journal of Molecular Evolution* 18:225-239.

Bulmer, M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Molecular Biology and Evolution* 8:868-883.

Burke, W. D., D. G. Eickbush, Y. Xiong, J. Jacubczak, and T. H. Eickbush. 1993. Sequence relationship of retrotransposable elements R1 and R2 within and between divergent insect species. *Molecular Biology and Evolution* 10:163-185.

Cavalli-Sforza, L. L. and A. W. F. Edwards. 1967. Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics* 19:233-257.

Chakraborty, R. 1977. Estimation of time of divergence from phylogenetic studies. *Canadian Journal of Genetics and Cytology* 19:217-223.

Cooper, A., C. Maourer-Chauviré, G. K. Chambers, A. von Haeseler, A. C. Wilson, and S. Pääbo. 1992. Independent origins of New Zealand moas and kiwis. *Proceedings of the National Academy of Sciences, USA* 89:8741-8744.

Cunningham, C. W., N. W. Blackstone, and L. W. Buss. 1992. Evolution of king crabs from hermit crab ancestors. *Nature* 355:539-542.

Dayhoff, M. O. 1978. Survey of new data and computer methods of analysis. In M. O. Dayhoff, ed., *Atlas of Protein Sequence and Structure*, vol. 5, supp. 3, pp. 29, National Biomedical Research Foundation, Silver Springs, Maryland.

DeBry, R. W. 1992. The consistency of several phylogeny-inference methods under varying evolutionary rates. *Molecular Biology and Evolution* 9:537-551.

Eck, R. V. and M. O. Dayhoff. 1966. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Springs, Maryland.

Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38, SIAM, Philadelphia.

Estabrook, G. F., C. S. Johnson, and F. R. McMorris. 1975. An idealized concept of the true cladistic character. *Mathematical Biosciences* 23:263-272.

Farris, J. S. 1981. Distance data in phylogenetic analysis. In V. A. Funk and D. R. Brooks, eds., *Advances in Cladistics. Proceedings of the First Meeting of the Willi Hennig Society*, pp. 3-23. New York Botanical Garden, Bronx.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27:401-410.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791.

Felsenstein, J. 1986. Distance Methods: Reply to Farris. *Cladistics* 2:130-143.

Felsenstein, J. 1988. Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics* 22:521-565.

Felsenstein, J. 1993. Phylogeny Inference Package (PHYLIP). Version 3.5. University of Washington, Seattle.

Felsenstein, J. and H. Kishino. 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Systematic Biology* 42:193-200.

Fitch, W. M. 1971. Towards defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology* 20:406-416.

Fitch, W. M. and E. Margoliash. 1967. Construction of phylogenetic trees. *Science* 155:279-284.

Gojobori, T., E. N. Moriyama, and M. Kimura. 1990. Statistical methods for estimating sequence divergence. In R. F. Doolittle, ed., *Methods in Enzymology, Vol. 183, Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, pp. 531-550. Academic Press, New York.

Goldman, N. 1993. Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* 36:182-198.

Goodman, M., A. E. Romero-Herrera, H. Dene, J. Czelusniak, and R. E. Tashian. 1982. Amino acid sequence evidence on the phylogeny of primates and other eutherians. In M. Goodman, ed., *Macromolecular sequences in systematic and evolutionary biology*, pp. 115-191. Plenum Press, New York and London.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22:160-174.

Hedges, S. B., S. Kumar, K. Tamura, and M. Stoneking. 1992. Human origins and analysis of mitochondrial DNA sequences. *Science* 255:737-739.

Hendy, M. D. and D. Penny. 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences* 59:277-290.

Hendy, M. D. and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Zoology* 38:297-309.

Hillis, D. M. and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* 42:182-192.

Hughes, A. L. and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167-170.

Jin, L. and M. Nei. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular Biology and Evolution* 7:82-102.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. In H. N. Munro, ed., *Mammalian Protein Metabolism*, pp. 21-132, Academic Press, New York.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111-120.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, Massachusetts.

Kimura, M. and T. Ohta. 1972. On the stochastic model for estimation of mutational distance between homologous proteins. *Journal of Molecular Evolution* 2:87-90.

Kishino, H. and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution* 29:170-179.

Kocher, T. D. and A. C. Wilson. 1991. Sequence evolution of mitochondrial DNA in humans and chimpanzees: Control region and a protein-coding region. In S. Osawa and T. Honjo, eds., *Evolution of Life*, pp. 391-413. Spring-Verlag, New York.

Kondo, R., S. Horai, Y. Satta, and N. Takahata. 1993. Evolution of hominoid mitochondrial DNA with special reference to the silent substitution rate over the genome. *Journal of Molecular Evolution* 36:517-531.

Lake, J. A. 1987. A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Molecular Biology and Evolution* 4:167-191.

Lee, Y. H. and V. D. Vacquier. 1992. The divergence of species-specific abalone sperm lysins is promoted by positive Darwinian selection. *Biological Bulletin* 182:97-104.

Le Quesne, W. J. 1969. A method of selection of characters in numerical taxonomy. *Systematic Zoology* 18:201-205.

Li, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* 36:96-99.

Li, W.-H., C.-I. Wu, and C.-C. Luo. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* 2:150-174.

Maddison, W. P. and D. R. Maddison. 1992. MacClade: Analysis of phylogeny and character evolution. Version 3. Sinauer Associates, Sunderland, Massachusetts.

Miyamoto, M. M. and J. Cracraft. 1991. *Phylogenetic Analysis of DNA Sequences*. Oxford University Press, New York.

Miyata, T. and T. Yasunaga. 1980. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *Journal of Molecular Evolution* 16:23-36.

Miyata, T., T. Yasunaga, and T. Nishida. 1980. Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proceedings of National Academy of Sciences, USA* 77:7328-7332.

Nei, M. 1986. Stochastic errors in DNA evolution and molecular phylogeny. In H. Gershowitz, D. L. Rucknagel, and R. E. Tashian, eds., *Evolutionary Perspectives and the New Genetics*. pp. 133-147. Alan R. Liss, New York.

Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M. 1991. Relative efficiencies of different tree making methods for molecular data. In M. M. Miyamoto and J. L. Cracraft, eds., *Recent Advances in Phylogenetic Studies of DNA Sequences*, pp. 90-128. Oxford University Press, Oxford.

Nei, M. and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* 3:418-426.

Nei, M. and A. L. Hughes. 1991. Polymorphism and evolution of the major histocompatibility complex loci in mammals. In R. K. Selander, A. G. Clark, and T. S. Whittam, eds., *Evolution at the Molecular Level*, pp. 222-247. Sinauer Associates, Sunderland, Massachusetts.

Nei, M. and L. Jin. 1989. Variances of the average numbers of nucleotide substitutions within and between populations. *Molecular Biology and Evolution* 6:290-300.

Nei, M. and A. Y. Rzhetsky. 1991. Reconstruction of phylogenetic trees and evolution of major histocompatibility complex genes. In J. Klein and D. Klein, eds., *Evolution of MHC Genes*, pp. 13-27. Springer-Verlang, Heidelberg.

Nei, M. and F. Tajima. 1981. DNA polymorphism detectable by restriction endonucleases.

Genetics 97:145-163.

Nei, M., R. Chakraborty, and P. A. Fuerst. 1976. Infinite allele model with varying mutation rate. *Proceedings of National Academy of Sciences, USA* 73:4164-4168.

Nei, M., J. C. Stephens, and N. Saitou. 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. *Molecular Biology and Evolution* 2:66-85.

Neigel, J. E. and A. C. Avise. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In S. Karlin and E. Nevo, eds., *Evolutionary Processes and Theory*, pp. 515-534. Academic Press, New York.

Pamilo, P. and N. O. Bianchi. 1993. Evolution of the *Zfx* and *Zfy* genes: Rates and interdependence between the genes. *Molecular Biology and Evolution* 10:271-281.

Pamilo, P. and M. Nei. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* 5:568-583.

Penny, D. 1982. Towards a basis for classification: Incompleteness of distance measures, incompatibility analysis, and phenetic classification. *Journal of Theoretical Biology* 96:129-142.

Penny, D. and M. D. Hendy. 1985. The use of tree comparison metrics. *Systematic Zoology* 34:75-82.

Rzhetsky, A. and M. Nei. 1992. A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution* 9:945-967.

Rzhetsky, A. and M. Nei. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution* 10:1073-1095.

Rzhetsky, A. and M. Nei. 1994. Unbiased estimates of the number of nucleotide substitutions when substitution rate varies among different sites. *Journal of Molecular Evolution* 38:295-299.

Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53:131-147

Saccone, C., C. Lanave, G. Pesole, and G. Preparata. 1990. Influence of base composition on quantitative estimates of gene evolution. In R. F. Doolittle, ed., *Methods in Enzymology, Vol. 183, Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, pp. 570-598. Academic Press, New York.

Saitou, N. and M. Imanishi. 1989. Relative efficiencies of the Fitch-Margolish, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree reconstruction in obtaining the correct tree. *Molecular Biology and Evolution* 6:514-525.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.

Sankoff, D. and R. J. Cedergren. 1983. Simultaneous comparison of three or more sequences related by a tree. In D. Sankoff and J. B. Kruskal, eds., *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison,* pp. 253-263. Addison-Wesley, Reading, Massachusetts.

Schöniger, M. and A. von Haeseler. 1993. A simple method to improve the reliability of tree reconstructions. *Molecular Biology and Evolution* 10:471-483.

Sharp, P. M., T. M. F. Tuohy, and K. R. Mosurski. 1986. Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research* 14:5125-5143.

Sneath, P. H. A. and R. R. Sokal. 1973. *Numerical Taxonomy*. Freeman, San Francisco.

Sober, E. 1988. *Reconstructing the Past.* MIT Press, Cambridge, Massachusetts.

Sourdis, J. and C. Krimbas. 1987. Accuracy of phylogenetic trees estimated from DNA sequence data. *Molecular Biology and Evolution* 4:159-166.

Sourdis, J. and M. Nei. 1988. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. *Molecular Biology and Evolution* 5:298-311.

Studier, J. A. and K. L. Keppler. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution* 5:729-731.

Swofford, D. L. 1993. Phylogenetic Analysis Using Parsimony (PAUP), Version 3.1.1. University of Illinois, Champaign.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437-460.

Tajima, F. 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution* 10:677-688.

Tajima, F. and M. Nei. 1984. Estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution* 1:269-285.

Tajima, F. and N. Takezaki. 1994. Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. *Molecular Biology and Evolution* 11:278-286.

Index