

技术引领医学转化
专业创造行业口碑



吉康医学
GENECOME MEDICAL

PHYLOGENETIC TREE
CONSTRUCTION

技术手册 进化树构建

BEIJING

GENECOME CO.,LTD.

北京吉康医学科技有限公司

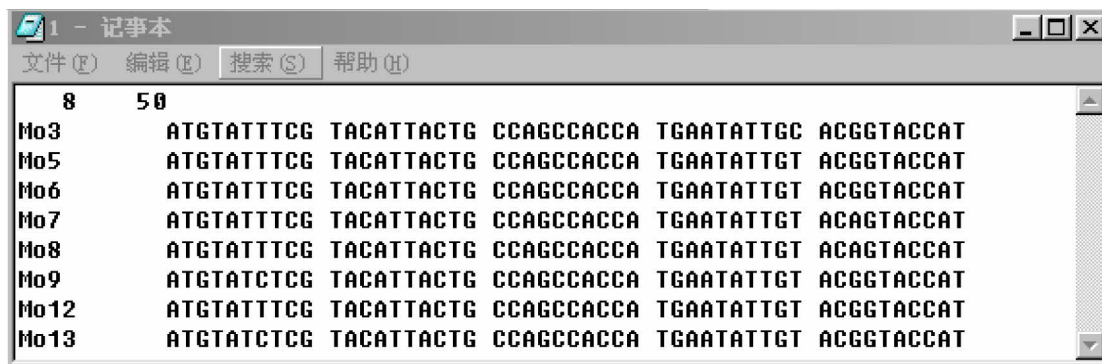
www.geneome.cn

下面介绍几个软件的使用。首先是 **PHYLIP**。其是多个软件的压缩包，下载后双击则自动解压。当你解压后就发现 **PHYLIP** 的功能极其强大，主要包括五个方面的功能软件：**i**，DNA 和蛋白质序列数据的分析软件。**ii**，序列数据转变成距离数据后，对距离数据分析的软件。**iii**，对基因频率和连续的元素分析的软件。**iv**，把序列的每个碱基/氨基酸独立看待（碱基/氨基酸只有 0 和 1 的状态）时，对序列进行分析的软件。**v**，按照 **DOLLO** 简约性算法对序列进行分析的软件。**vi**，绘制和修改进化树的软件。在此，我主要对前两种功能软件进行说明。

我们现在有几个序列如下：

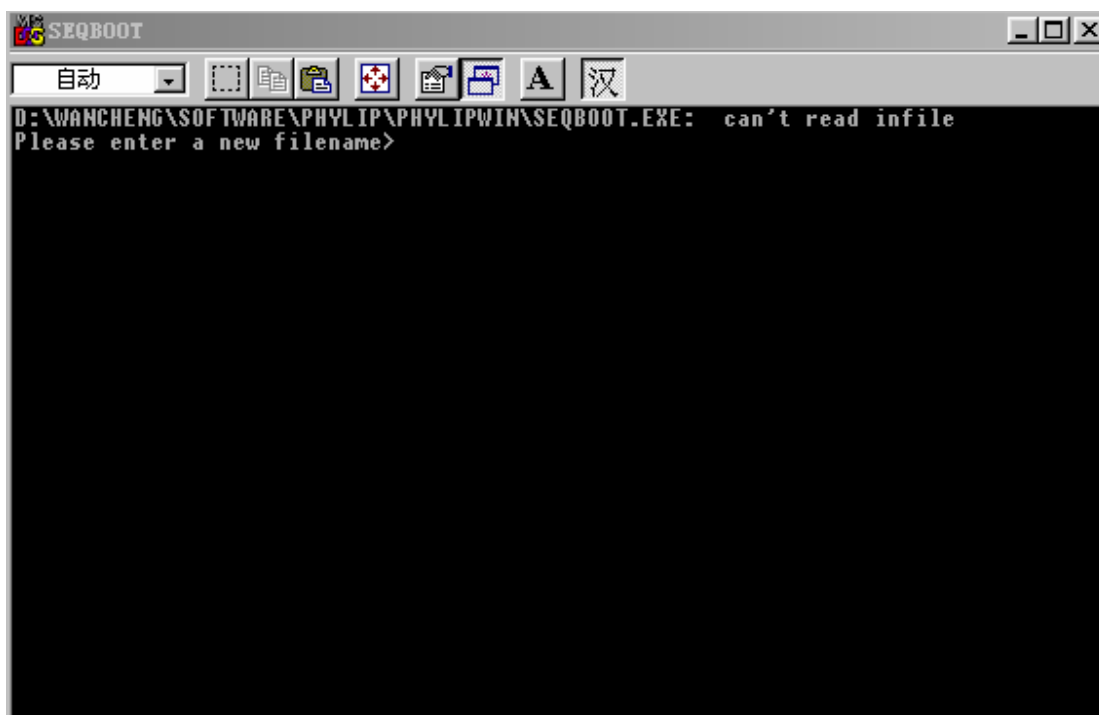
```
Mo3      ATGTATTTTCGTACATTACTGCCAGCCACCATGAATATTGCACGGTACCAT
Mo5      ATGTATTTTCGTACATTACTGCCAGCCACCATGAATATTGTACGGTACCAT
Mo6      ATGTATTTTCGTACATTACTGCCAGCCACCATGAATATTGTACGGTACCAT
Mo7      ATGTATTTTCGTACATTACTGCCAGCCACCATGAATATTGTACAGTACCAT
Mo8      ATGTATTTTCGTACATTACTGCCAGCCACCATGAATATTGTACAGTACCAT
Mo9      ATGTATCTCGTACATTACTGCCAGCCACCATGAATATTGTACGGTACCAT
Mo12     ATGTATTTTCGTACATTACTG CCAGCCACCATGAATATTGTACGGTACCAT
Mo13     ATGTATCTCGTACATTACTGCCAGCCACCATGAATATTGTACGGTACCAT
```

要对这 8 个序列进行进化树分析，按照上面的步骤，首先用 **CLUSTALX** 排列序列，输出格式为 ***.PHY**。用记事本打开如下图：

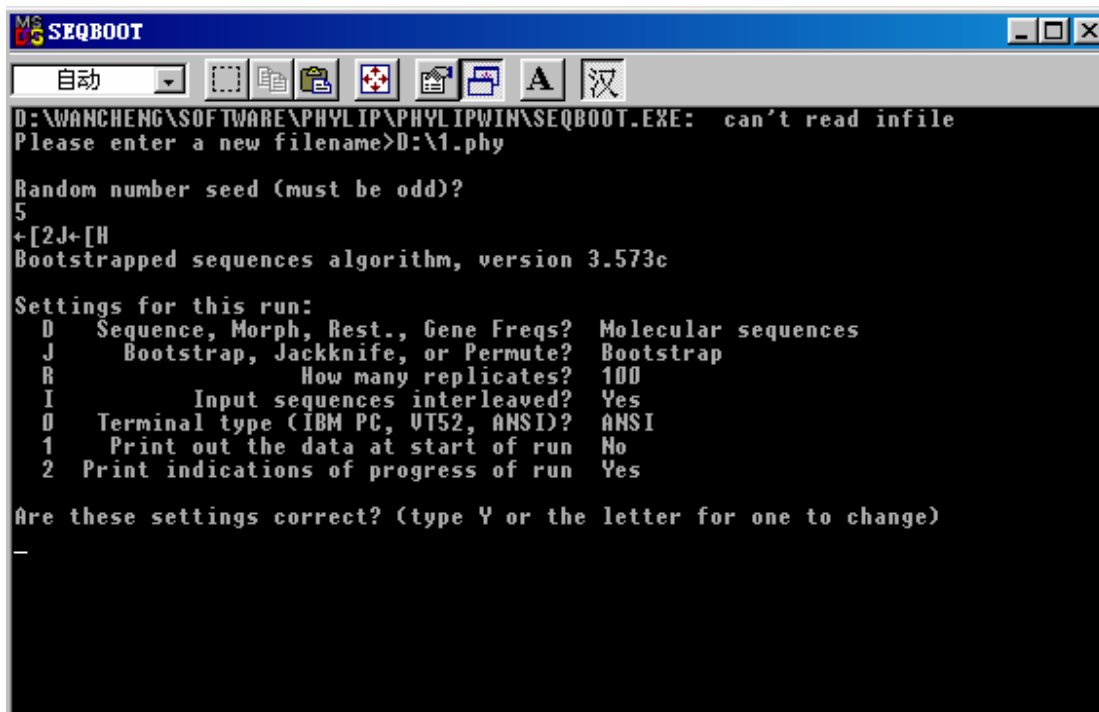


图中的 8 和 50 分别表示 8 个序列和每个序列有 50 个碱基。然后，打开软件

SEQBOOT, 如下图:



按路径输入刚才生成的 *.PHY 文件, 并在 Random number seed (must be odd) ? 的下面输入一个 $4N+1$ 的数字后, 屏幕显示如下:



图中的 D、J、R、I、O、1、2 代表可选择的选项, 键入这些字母, 程序的条件就会发生改变。D 选项无须改变。J 选项有三种条件可以选择, 分别是 Bootstrap、

Jackknife 和 Permute。文章上面提到用 Bootstrapping 法对进化树进行评估，所谓 Bootstrapping 法就是从整个序列的碱基（氨基酸）中任意选取一半，剩下的一半序列随机补齐组成一个新的序列。这样，一个序列就可以变成了许多序列。一个多序列组也就可以变成许多个多序列组。根据某种算法（最大简约性法、最大可能性法、除权配对法或邻位相连法）每个多序列组都可以生成一个进化树。将生成的许多进化树进行比较，按照多数规则（majority-rule）我们就会得到一个最为逼真的进化树。Jackknife 则是另外一种随机选取序列的方法。它与 Bootstrap 法的区别是不将剩下的一半序列补齐，只生成一个缩短了一半的新序列。Permute 是另外一种取样方法，其目的与 Bootstrap 和 Jackknife 法不同，这里不再介绍。R 选项让使用者输入 replicate 的数目。所谓 replicate 就是用 Bootstrap 法生成的一个多序列组。根据多序列中所含的序列的数目的不同可以选取不同的 replicate。当我们设置好条件后，键入 Y 按回车。得到一个文件 outfile Outfile 用记事本打开如下：

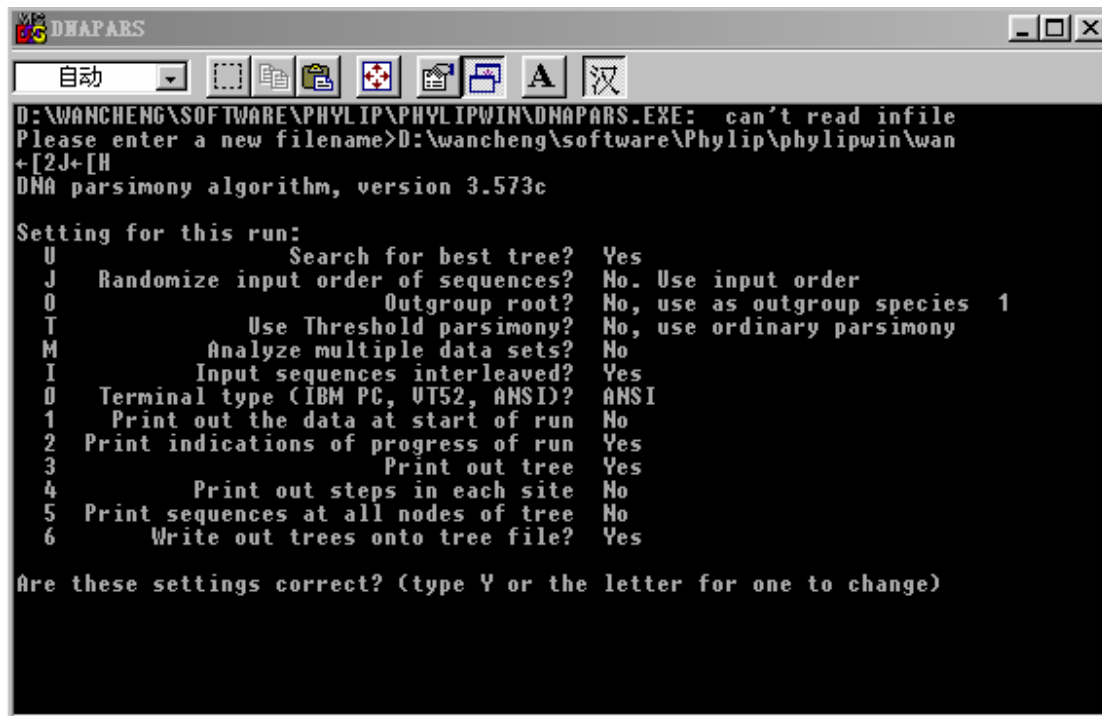
```

8      50
Mo3      AATGTATTTG TATTAACCCC CAAACCCCAA TAATTTGGGC CAACTCTTTT
Mo5      AATGTATTTG TATTAACCCC CAAACCCCAA TAATTTGGGT TAACTCTTTT
Mo6      AATGTATTTG TATTAACCCC CAAACCCCAA TAATTTGGGT TAACTCTTTT
Mo7      AATGTATTTG TATTAACCCC CAAACCCCAA TAATTTGGGT TAACTCTTTT
Mo8      AATGTATTTG TATTAACCCC CAAACCCCAA TAATTTGGGT TAACTCTTTT
Mo9      AATGTATTTG TATTAACCCC CAAACCCCAA TAATTTGGGT TAACTCTTTT
Mo12     AATGTATTTG TATTAACCCC CAAACCCCAA TAATTTGGGT TAACTCTTTT
Mo13     AATGTATTTG TATTAACCCC CAAACCCCAA TAATTTGGGT TAACTCTTTT
8      50
Mo3      TTGTTTAAAA CATTAAACTT TGACGGGATA TGGGCAACCC GGGACCAAA
Mo5      TTGTTTAAAA CATTAAACTT TGACGGGATA TGGGTAACCC GGGACCAAA
Mo6      TTGTTTAAAA CATTAAACTT TGACGGGATA TGGGTAACCC GGGACCAAA
Mo7      TTGTTTAAAA CATTAAACTT TGACGGGATA TGGGTAACCC GGGACCAAA
Mo8      TTGTTTAAAA CATTAAACTT TGACGGGATA TGGGTAACCC GGGACCAAA
Mo9      TTGCTTAAAA CATTAAACTT TGACGGGATA TGGGTAACCC GGGACCAAA
Mo12     TTGTTTAAAA CATTAAACTT TGACGGGATA TGGGTAACCC GGGACCAAA
Mo13     TTGCTTAAAA CATTAAACTT TGACGGGATA TGGGTAACCC GGGACCAAA
8      50
Mo3      ATGGGTTTTT CGGAATTAC CTCCAAGCA TAAATAATTT GGGTACCCTT
Mo5      ATGGGTTTTT CGGAATTAC CTCCAAGCA TAAATAATTT GGGTACCCTT
Mo6      ATGGGTTTTT CGGAATTAC CTCCAAGCA TAAATAATTT GGGTACCCTT
Mo7      ATGGGTTTTT CGGAATTAC CTCCAAGCA TAAATAATTT GGGTACCCTT
Mo8      ATGGGTTTTT CGGAATTAC CTCCAAGCA TAAATAATTT GGGTACCCTT

```

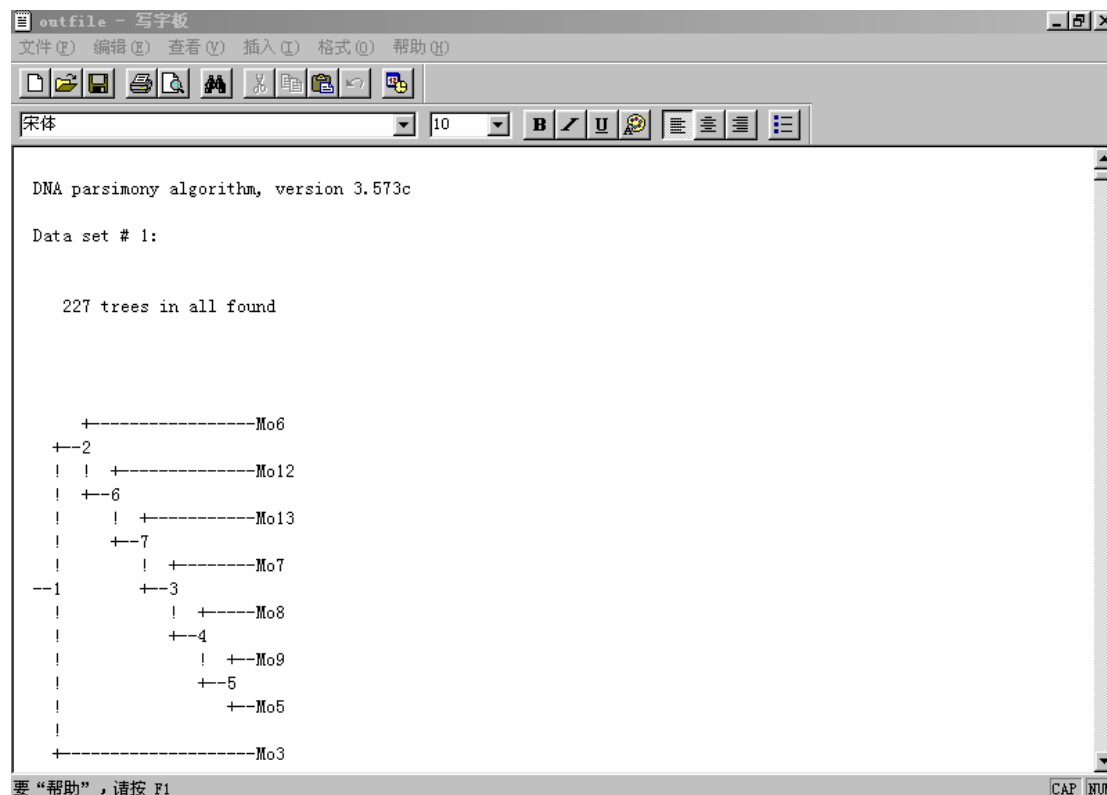
这个文件包括了 100 个 replicate。

打开 DNAPARS（最大简约性法）或 DNAML（最大可能性法）软件。将刚才生成的 outfile 文件更名后输入。如下图：



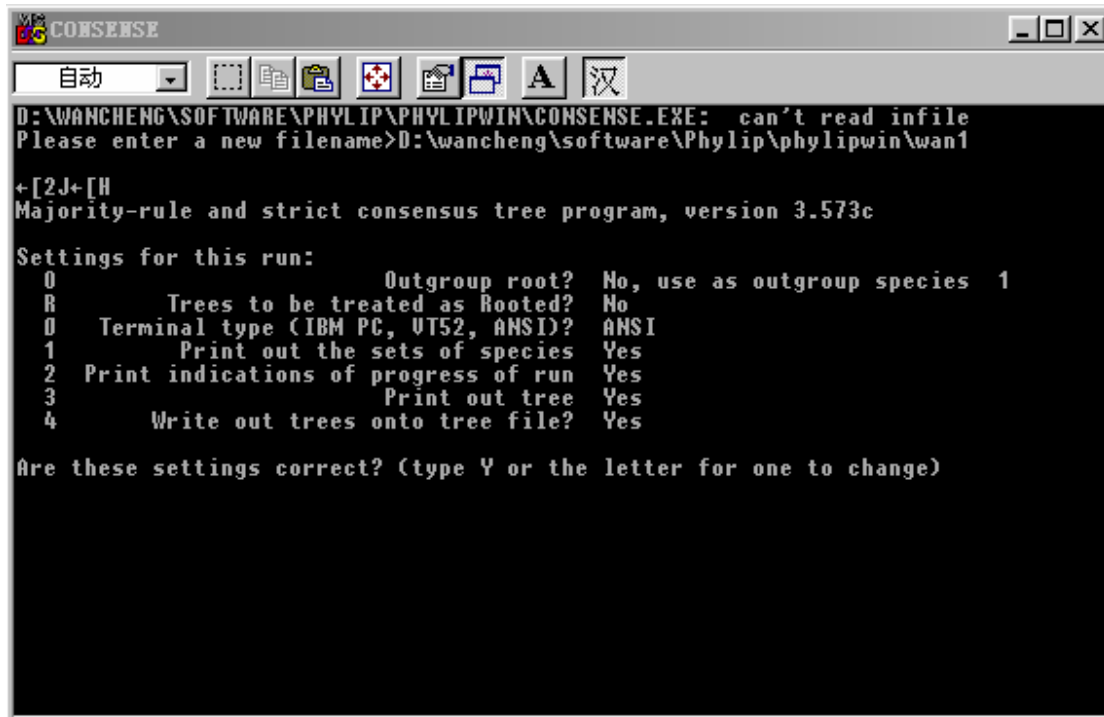
选项 O 是让使用者设定一个序列作为 outgroup。一般选择一个亲缘关系与所分析序列组很接近的序列作为 outgroup (本例子不选 outgroup)，outgroup 选择的好坏将直接影响到最后的进化树的好坏。选项 M 是输入刚才设置的 replicate 的数目。设置好条件后，键入 Y 按回车。生成两个文件 outfile 和 treefile。

Outfile 打开如下图：

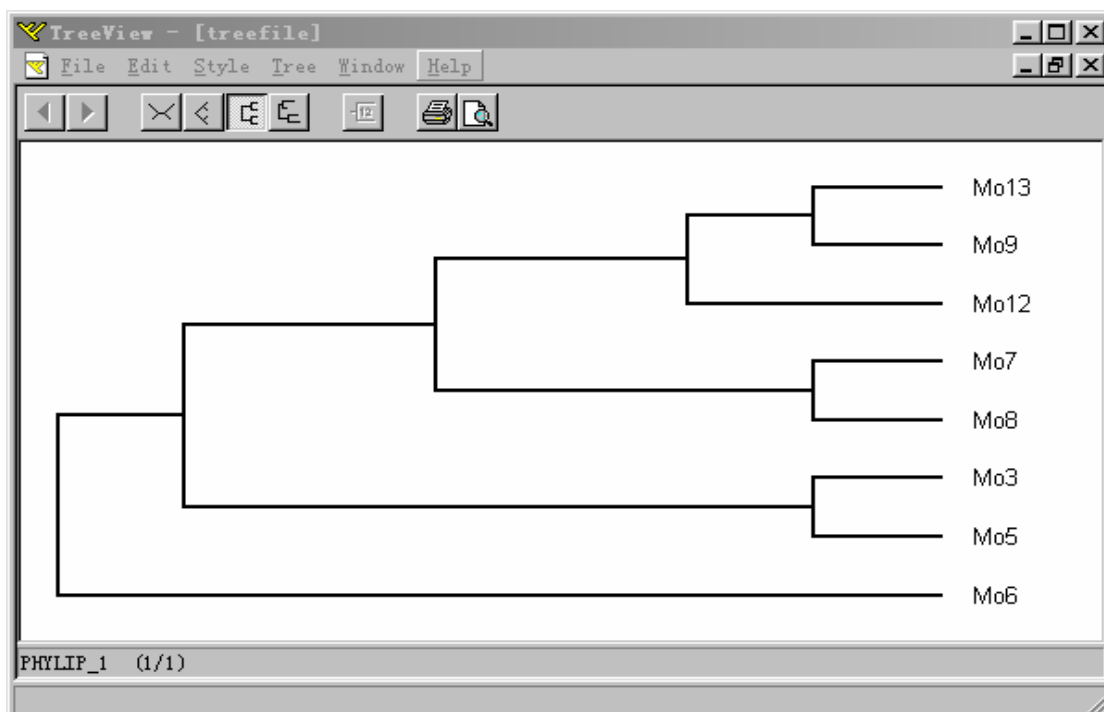


该文件包括了 227 个进化树。Treefile 可以用 TREEVIEW 软件打开同样包含了这 227 个进化树。

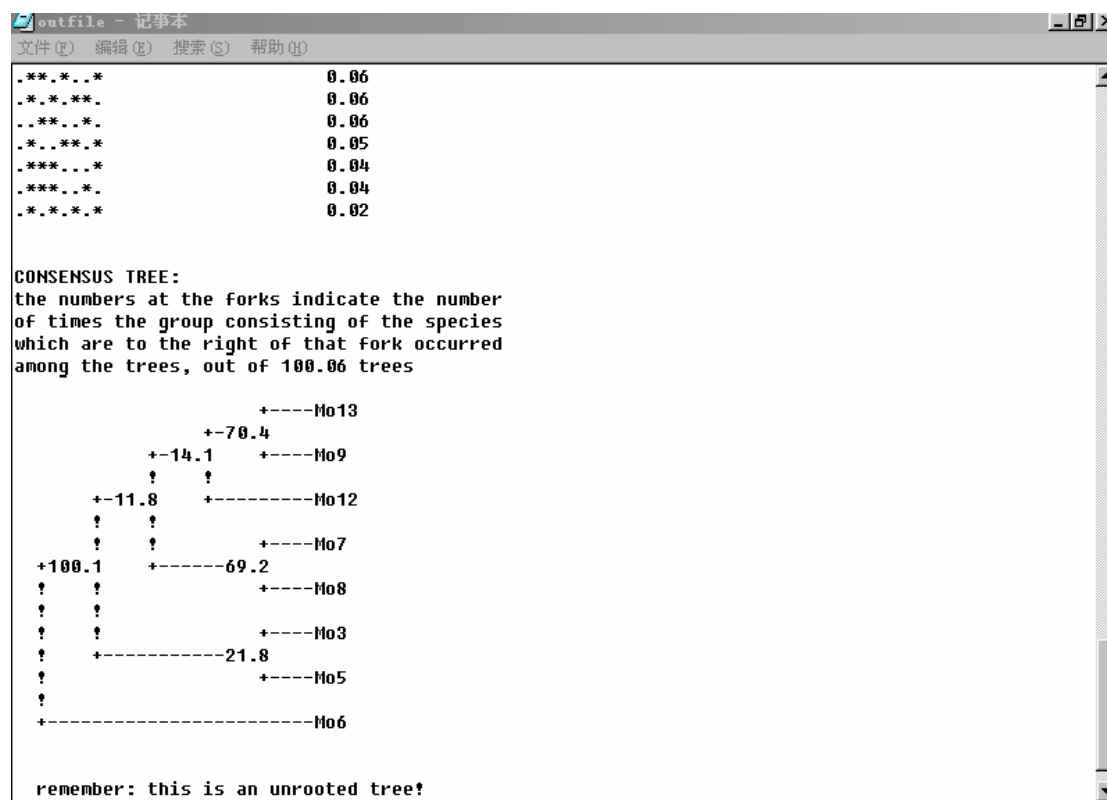
打开 CONSENSE 软件，将刚才生成的 treefile 文件更名后输入。如下图：



键入 Y 按回车。生成两个文件 outfile 和 treefile。Treefile 用 TREEVIEW 打开，如下图：



Outfile 打开如下图:



```
outfile - 记事本
文件(F) 编辑(E) 搜索(S) 帮助(H)

.**.**.*          0.06
.**.**.*          0.06
.**.**.*          0.06
.**.**.*          0.05
.**.**.*          0.04
.**.**.*          0.04
.**.**.*          0.02

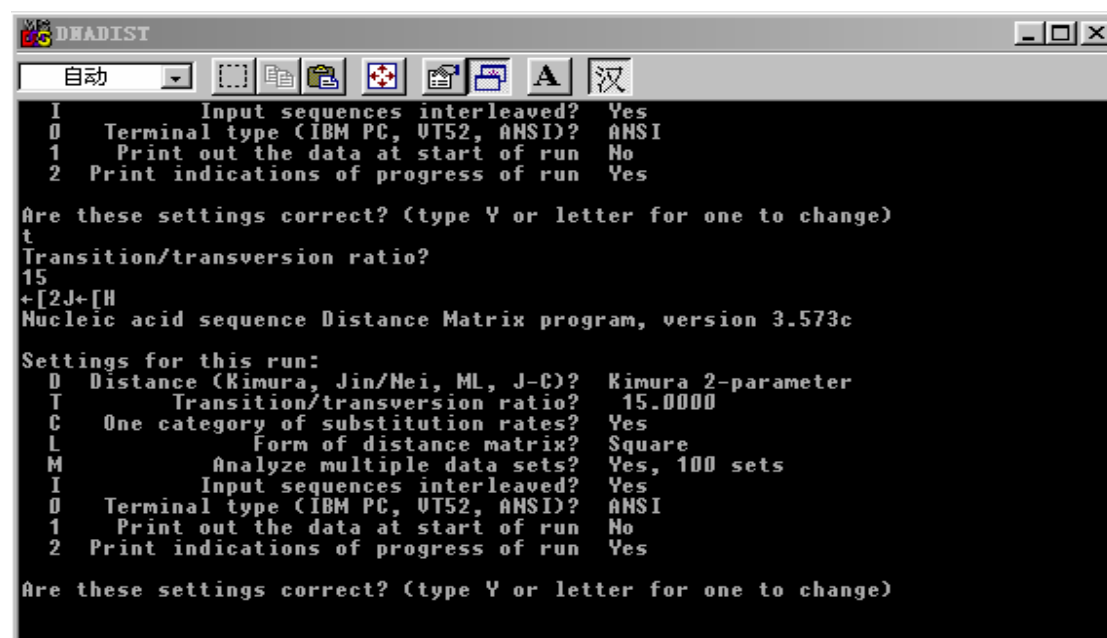
CONSENSUS TREE:
the numbers at the forks indicate the number
of times the group consisting of the species
which are to the right of that fork occurred
among the trees, out of 100.06 trees

          +----Mo13
        +-70.4
      +---14.1 +----Mo9
      |         |
      |         +----Mo12
      |         |
      |         |
    +-11.8     +----Mo7
      |         |
      |         +----Mo8
    +100.1     +----Mo3
      |         |
      |         +----Mo5
      |         +----Mo6
      |         |
      |         +----Mo6

remember: this is an unrooted tree!
```

我们看出两个树是同样的。但在 outfile 的树上的数字表示该枝条的 Bootstrap 支持率（除以 100.6）。到现在，8 个序列的进化树分析（最大简约法）已经完成。

如果要用邻位相连法对这 8 个序列进行分析的话，也首先执行 SEQBOOT 软件将这 8 个序列变成 100 个 replicate。然后，打开 DNADIST 软件，把 SEQBOOT 生成的文件输入，如下图：



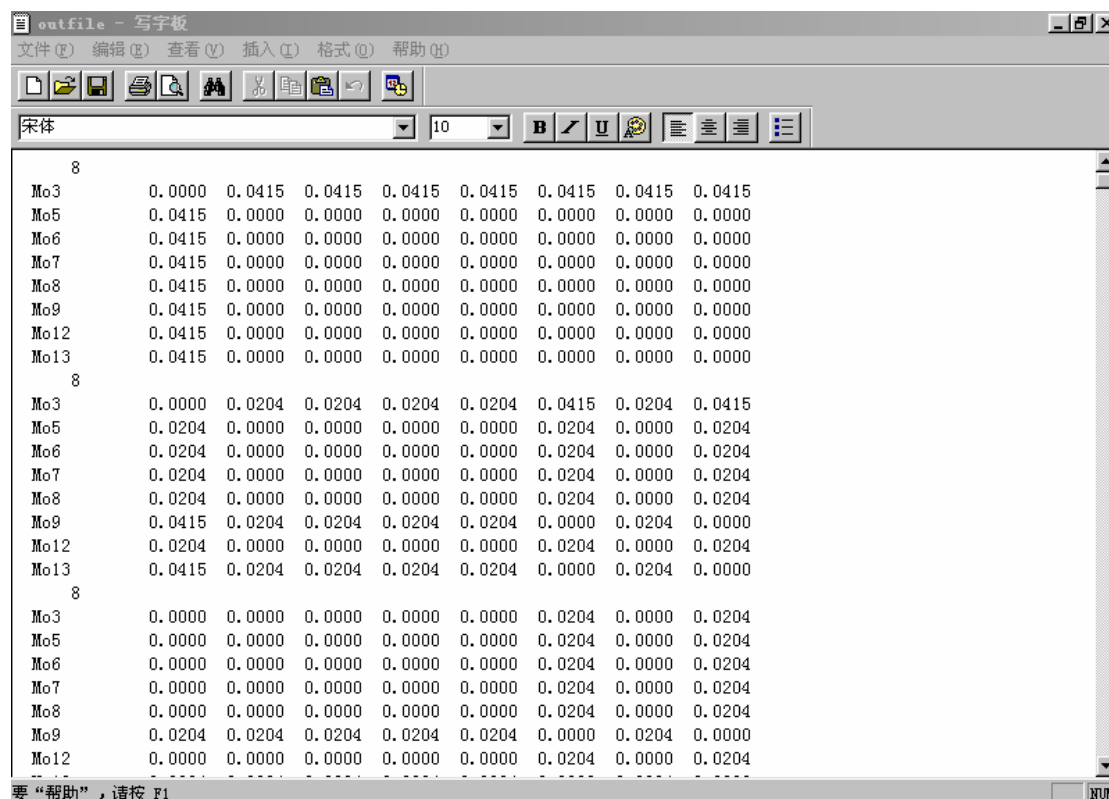
```
DNADIST
自动
I Input sequences interleaved? Yes
O Terminal type (IBM PC, UT52, ANSI)? ANSI
1 Print out the data at start of run No
2 Print indications of progress of run Yes

Are these settings correct? (type Y or letter for one to change)
t
Transition/transversion ratio?
15
+ [2] + [H]
Nucleic acid sequence Distance Matrix program, version 3.573c

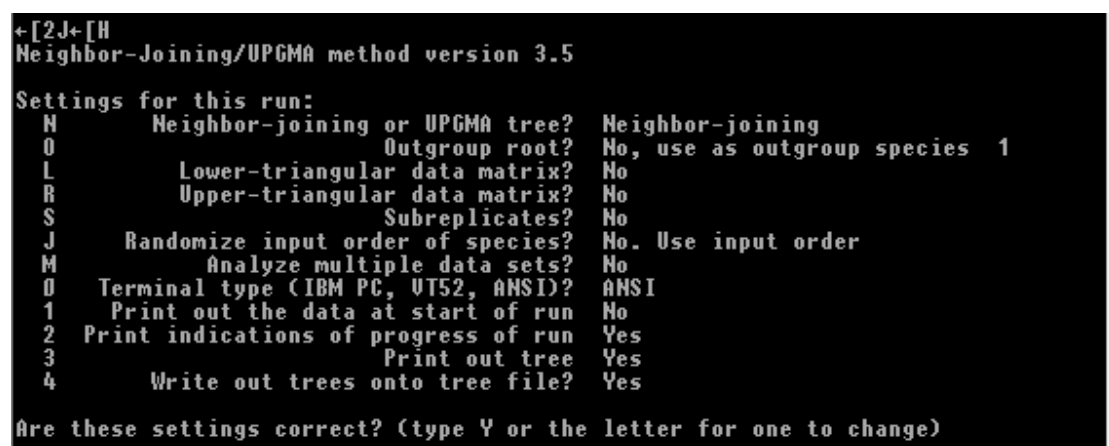
Settings for this run:
D Distance (Kimura, Jin/Nei, ML, J-C)? Kimura 2-parameter
T Transition/transversion ratio? 15.0000
C One category of substitution rates? Yes
L Form of distance matrix? Square
M Analyze multiple data sets? Yes, 100 sets
I Input sequences interleaved? Yes
O Terminal type (IBM PC, UT52, ANSI)? ANSI
1 Print out the data at start of run No
2 Print indications of progress of run Yes

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

选项 D 有四种距离模式可以选择，分别是 Kimura 2-parameter、Jin/Nei、Maximum-likelihood 和 Jukes-Cantor。选项 T 一般键入一个 15-30 之间的数字。选项 M 键入 100。运行后生成文件如下图：



这个文件包含了与输入文件相同的 100 个 replicate，只不过每个 replicate 是以两两序列的进化距离来表示。文件中的每个 replicate 都省略了第一排的 Mo3 Mo5 Mo6 Mo7 Mo8 Mo9 Mo12 Mo13。以这个输出文件为输入文件，执行 NEIGHBOR 软件。如下图：

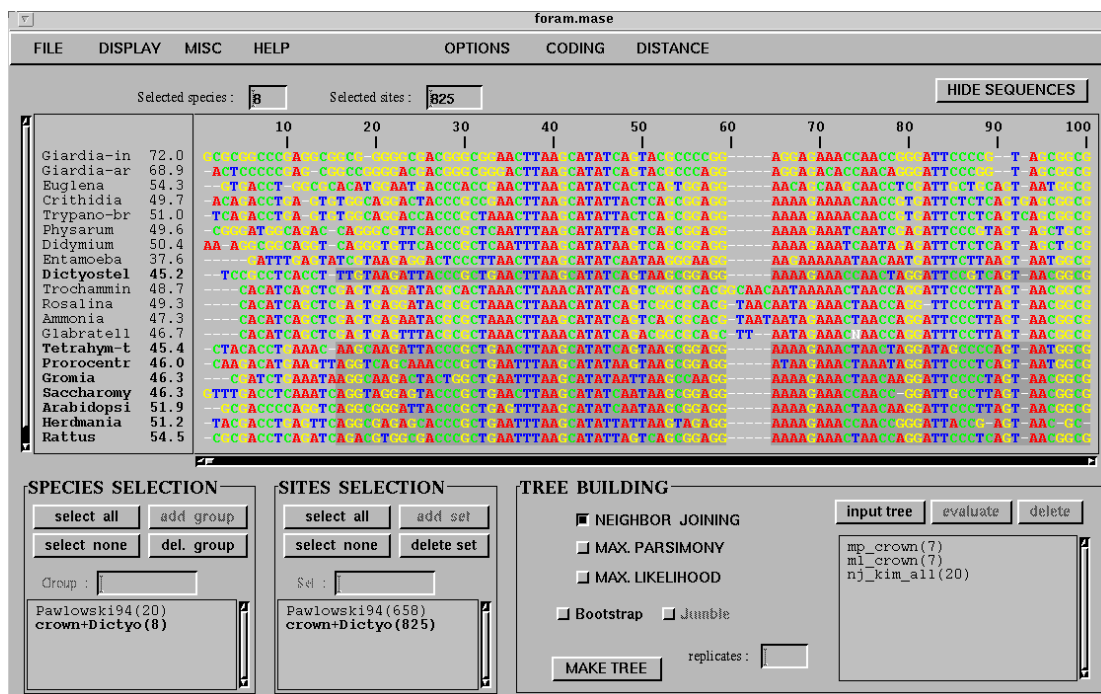


选项 M 键入 100。生成两个文件 outfile 和 treefile 用记事本和 TREEVIEW 打开后，发现这两个文件都含有 100 个进化树。再将 treefile 文件更名后输入

CONSENSE 软件，又得到两个文件 outfile 和 treefile，这就是最后的结果。以上是对 DNA 序列的分析，如果要对蛋白质序列进行分析，PROTDIST、PROTPARS 等软件。其他软件的使用可以参照 PHYLIP 的 documents。

下面介绍 PUZZLE 软件。它是用最大可能性的方法来构建进化树的一个软件，并且对树进行 bootstrap 评估。该软件搜寻进化树时用的算法是 quartet puzzling，这个算法相对较快，但如要分析的序列较多时，也相当耗时。另有 LINUX 版，运行起来相对较快。PUZZLE 的输入格式为 PHYLIP INTERLEAVED。CLUSTAL 可以生成此格式文件。PUZZLE 的界面与 PHYLIP 类似，也是 MS-DOS 下的软件。

PHYLO-WIN 是 LINUX 下的一个软件。界面友好，极易操作。该界面如下图：



Puzzle: <http://www.tree-puzzle.de>

Phylo-win: <http://www.evolution.bmc.uu.se>

Phylip、Treeview and Clustalx: <http://biosoft.yeah.net>