




系统发生分析小结

Caas07f2a1-a4

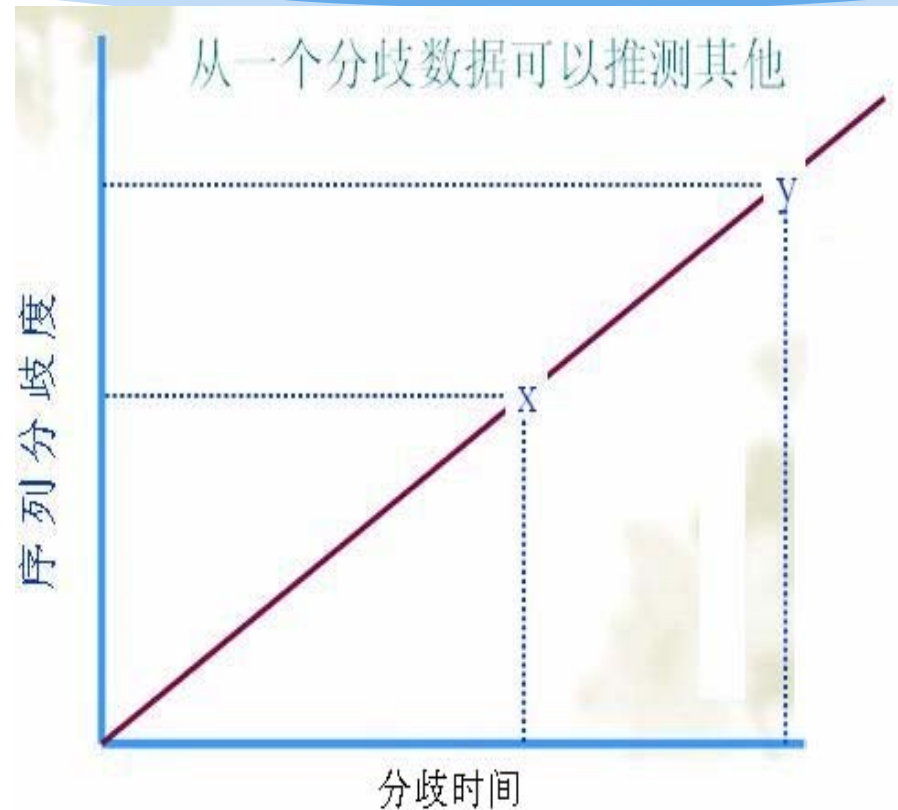
黄拔严 于琳琳 李康 韦永龙

Introduction

- ❖ 系统发生（**phylogeny**）是指生物形成或进化的历史
- ❖ 系统发生学（**phylogenetics**）研究物种之间的进化关系，其基本思想是比较物种的特征，并认为特征相似的物种在遗传学上接近。
- ❖ 系统发生研究的结果往往以系统发生树（**phylogenetic tree**）表示，用它来描述物种之间的进化关系

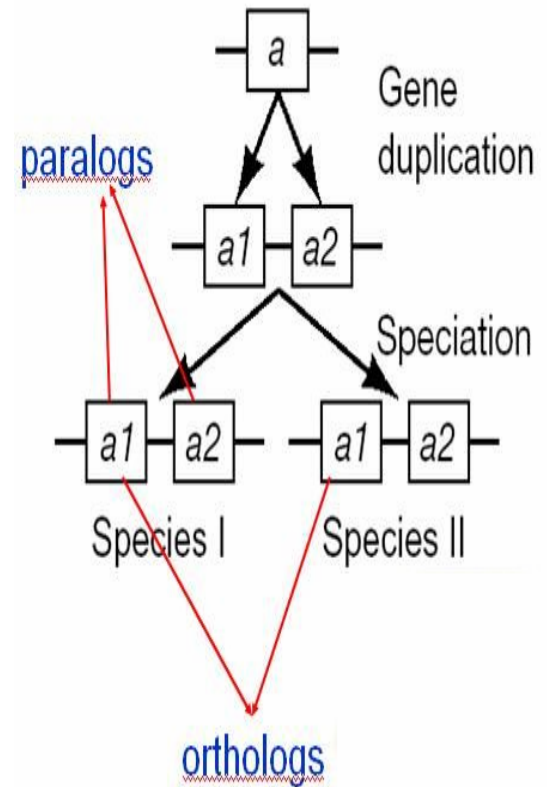
- 
- ❖ 蛋白质与核酸中序列与结构上保留有遗传的痕迹，可用于系统发生关系的研究
 - ❖ 分子系统发生分析通过比较生物分子序列，比较序列之间的关系，构造系统发生树，进而阐明各个物种的进化关系。

- ❖ 系统发生分析一般是建立在分子钟基础上的
- ❖ 分子钟:分子序列进化是按照一恒定速率进行的,所以积累突变的数量和进化时间成一定比例,基于这个假说,发生树上的树枝长度可以用来估算基因分离的时间。



直系同源与旁系同源

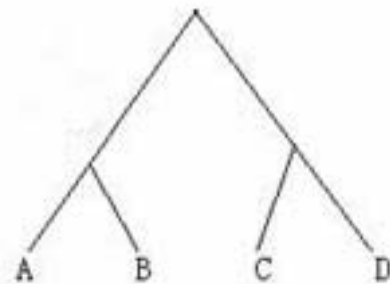
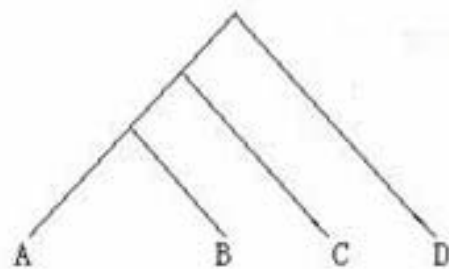
- ❖ 直系同源(**orthologs**):同源的基因是由于共同的祖先基因进化而产生的;
- ❖ 旁系同源(**paralogs**):同源的基因是由于基因复制产生的。



用于分子进化分析中的序列必须是直系同源的，才能真实反映进化过程

系统发生树

- ❖ 系统发生树就是一个用来表示一组对象之间的进化关系的树形结构。
- ❖ 系统进化树分有根(**rooted**)和无根(**unrooted**)树。有根树反映了树上物种或基因的时间顺序，而无根树只反映分类单元之间的距离而不涉及谁是谁的祖先问题。



有根树


进化树的构建

基本思想:

物种体内同功能生物分子（如蛋白质或核酸分子）的相似程度越高，则物种的亲缘关系越近。

具体步骤:

- 选择“特征分子”，原则是：a. 各个物种都有的同源分子，b. 进化速率适当；
- 对这些同源分子的序列进行多序列比对(multi-sequences alignment), 截取比对的最好的区域作为物种的代表序列；

- 
3. 按某种方法，算出代表序列两两之间的差异度，
 4. 基于这些差异度，绘制系统发生树
 5. 对系统发生树进行可信度检验(bootstrap)

选择特征分子


- ❖ 既可以用核酸序列又可以用蛋白序列
- ❖ 用核酸序列还是蛋白序列主要取决于序列的**性质**和研究的**目的**
- ❖ 对于具有很近亲缘关系的生物来说，选择核酸序列研究要比选择蛋白序列更快的推断出结果
- ❖ 在大多数情况下，通过蛋白质序列研究要比用核酸来研究要好，因为蛋白质序列含有更多相对保守的序列

- ❖ 由于蛋白质序列由**20**个氨基酸组成，而核酸序列是由**4**种核酸组成，因此蛋白质序列的比对比DNA序列的比对更灵敏。

大多数情况下以蛋白质为基础的发生树比以**DNA**为基础的发生树更恰当。

序列比对

- ❖ 只有正确的比对结果才会能推出正确的系统发生。错误的比对结果会导致最后发生树在分类上的错误，甚至是整个树的错误
- ❖ 多序列比对的结果应该进行检验并找出一个最合理的结果。序列自动比对的结果通常会存在错误，应该进行进一步的编辑或是进行提炼



❖ 对这些同源分子的序列进行多序列比对(multi-sequences alignment), 截取比对的最好的区域作为物种的代表序列

方法

- ❖ 根据所处理数据的类型，可以将系统发生树的构建方法大致分为两大类：

基于距离的构建方法

UPGMA（unweighted pair group method with arithmetic mean，平均连接聚类法）、ME（Minimum Evolution，最小进化法）和NJ（Neighbor-Joining，邻接法）

- 基于特征的构建方法

最大简约法（MP法），最大似然法（ML法），进化简约法（EP法），相容性方法等


Neighbor-Joining Method (NJ法/邻接法)

- ❖ 不需要关于分子钟的假设
- ❖ 基本思想：进行类的合并时，不仅要求待合并的类是相近的，而且要求待合并的类远离其他的类

最大简约法 (MP)

- ❖ 解释一个过程的最好理论是所需假设数目最少的那一个。对所有可能的拓扑结构进行计算，并计算出所需替代数最小的那个拓扑结构，作为最优树。
- ❖ 最大简约法不需要在处理核苷酸或者氨基酸替代的时候引入假设（替代模型）。此外，最大简约法对于分析某些特殊的分子数据如插入、缺失等序列有用。
- ❖ 在分析的序列位点上没有回复突变或平行突变，且被检验的序列位点数很大的时候，最大简约法能够推导获得一个很好的进化树。
- ❖ 在分析序列上存在较多的回复突变或平行突变，而被检验的序列位点数又比较少的时候，最大简约法可能会给出一个不合理的或者错误的进化树推导结果

适用于近缘序列

- 
- ❖ 适用于: 物种(序列)相似程度很高的情况。
 - ❖ 优点: 找到的一定是最优的树(结构), 能推测“祖先”序列。
 - ❖ 缺点: 当物种(序列)的数目较大时($N > 13$), 计算时间太长, 所以, 可行性很差。

最大似然法 (ML)

- ❖ 选取一个特定的替代模型来分析给定的一组序列数据，使得获得的每一个拓扑结构的似然率都为最大值，然后再挑出其中似然率最大的拓扑结构作为最优树。
- ❖ 最大似然法的建树过程是个很费时的过程，因为在分析过程中有很大的计算量，每个步骤都要考虑内部节点的所有可能性。
- ❖ 最大似然法也是一个比较成熟的参数估计的统计学方法，具有很好的统计学理论基础，在当样本量很大的时候，似然法可以获得参数统计的最小方差。
- ❖ 只要使用了一个合理的、正确的替代**模型**，最大似然法可以推导出一个很好的进化树结果。

对于模型的巨大依赖性是最大似然法的特征

一般情况下，若有合适模型，**ML**的效果较好；近缘序列，一般使用**MP**（基于的假设少）；远缘序列，一般使用**NJ**或**ML**

系统发生树的可靠性

- ❖ 用截然不同的距离矩阵法与简约法分析一个数据集，如果能够产生相似的系统发生树，这样的树可以认为是可靠的

进化树的可信度检验

常用的三种方法：

1. **The bootstrap**
2. **Delete-half-jackknifing**
3. **Permuting species within characters**

The bootstrap

方法: 对“列”进行“有放回地”重取样。

S1: AACCAAC
S2: AACCCC
S3: ACCAAC
S4: CCACCA
S5: CCAAAC

S1: ACCCAC
S2: ACCCCC
S3: CCCCAC
S4: CAAACA
S5: CAAAAC

S1: AAAACC
S2: AACCCC
S3: ACAACC
S4: CCCCAA
S5: CCAACC

S1: AAAAAC
S2: AACCCC
S3: CCAAAC
S4: CCCCCA
S5: CCAAAC

.....



在任何一组新的序列中：

- ❖ 序列的长度和原始的长度一样；
- ❖ 某些“列”可能被使用多次，而某些“列”则可能没用到。

Delete-half-jackknifing

从一组(**set**)原始序列中，“无放回地”随机抽取一半的“列”，形成一组组新的序列。

- ❖ 在新的序列组中，序列的长度是原来的原来的一半。
- ❖ 在一组新的序列中，每一“列”最多出现一次。
- ❖ 这种方法的思想 and 效果都和**bootstrap**类似。

Permuting species within characters

方法：对“列”进行“序列改变”，结果产生的序列组表面上看起来和原来一样，但本来所含的分类关系 (taxonomic structure) 信息已被破坏。

- ❖ 如果这样做并不明显改变那些分类相关的统计量(如树的总分支长度)，则认为原来的序列组不含有意义的分类关系信息。
- ❖ 如果这种统计量明显变大，则可认为原来的序列组含有明显的分类关系信息，以此为基础构建的树是有意义的。
- ❖ 这种方法的思路和前两种完全不同，类似于“反证法”。



具体做法(以PHYML包为例)步骤:

1. 用“seqboot”程序来产生新的序列组(一般地, 100 到1000组);
2. 运行“dnapars”产生相应数目的“树”;
- 2'. 也可先用“dnadist”将上述序列组变为一个个的距离矩阵, 然后用“neighbor”构建相应数目的“树”;
3. 最后运行“consense”, 得出一棵“一致性”的“树”, 其各结点上带有bootstrap值。

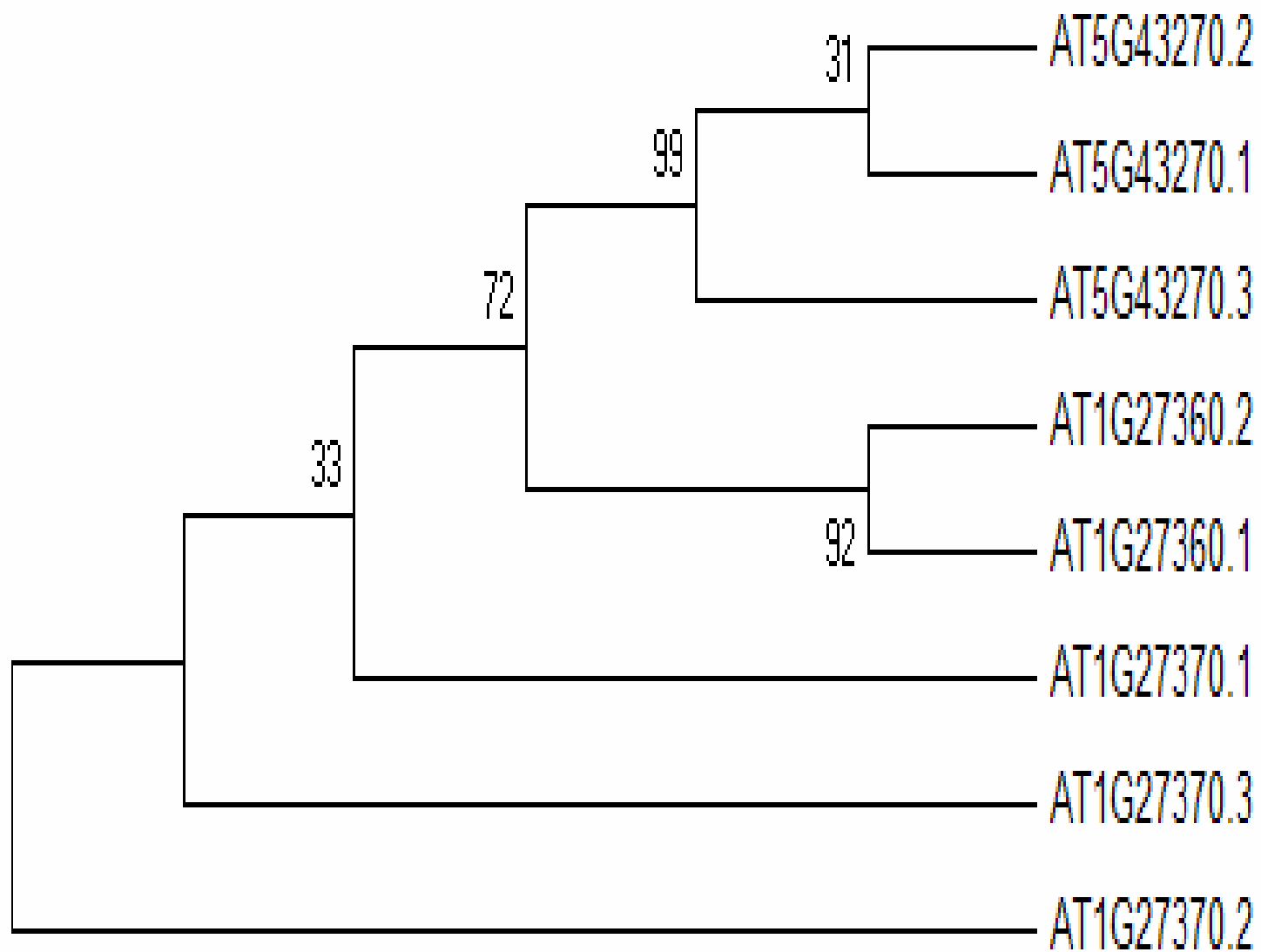
Bootstrapping algorithm, version 3.66

Settings for this run:

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

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







分子进化与系统发育分析软件


软件名称	网址	说明
PHYLIP	http://evolution.genetics.washington.edu/phylip/software.html	目前发布最广，用户最多的通用系统树构建软件，由美国华盛顿大学Felsenstein开发，可免费下载，适用绝大多数操作系统
PAUP	ftp://onyx.si.edu/paup	国际上最通用的系统树构建软件之一，美国simthsonian institute开发，仅适用Apple-Macintosh和UNIX操作系统
Tree of Life	http://phylogeny.arizona.edu/tree/program/program.html	美国University of Arizona建立的系统发育方面网站
MEGA	http://bioinfo.weizmann.ac.il/databases/info/mega.sof	美国宾西法尼亚州立大学MasatoshiNei开发的分子进化遗传学软件
MOLPHY	ftp://ftpsunmh.ism.ac.jp/pub/molphy	日本国立统计数理研究所开发，最大似然法构树
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html	英国University college London开发，最大似然法构树和分子进化模型
PUZZLE	ftp://fx.zi.biologie.uni-muenchen.de/pub/puzzle	应用quarter puzzling方法(一种最大简约法)构建系统树
TreeView	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html	英国University of Glasgow开发
phylogeny	http://www.ebi.ac.uk/biocat/phylogeny.html	欧洲生物信息研究所(EBI)的系统发育分析软件

- ❖ 构建NJ树，可以用PHYLP或者MEGA。MEGA是Nei开发的方法并设计的图形化的软件，使用非常方便，推荐使用。虽然多序列比对工具ClustalW/X也自带了一个NJ的建树程序，但是该程序只有p-distance模型，而且构建的树不够准确，一般不用来构建进化树。
- ❖ 构建MP树，最好的工具是PAUP，但该程序属于商业软件，并不对科研学术免费。MEGA和PHYLP也可以用来构建MP树。
- ❖ 构建ML树可以使用PHYML，速度较快。也可使用Tree-puzzle，该程序做蛋白质序列的进化树效果比较好。ML还可以使用PAUP、PHYLP（或BioEdit）来构建。

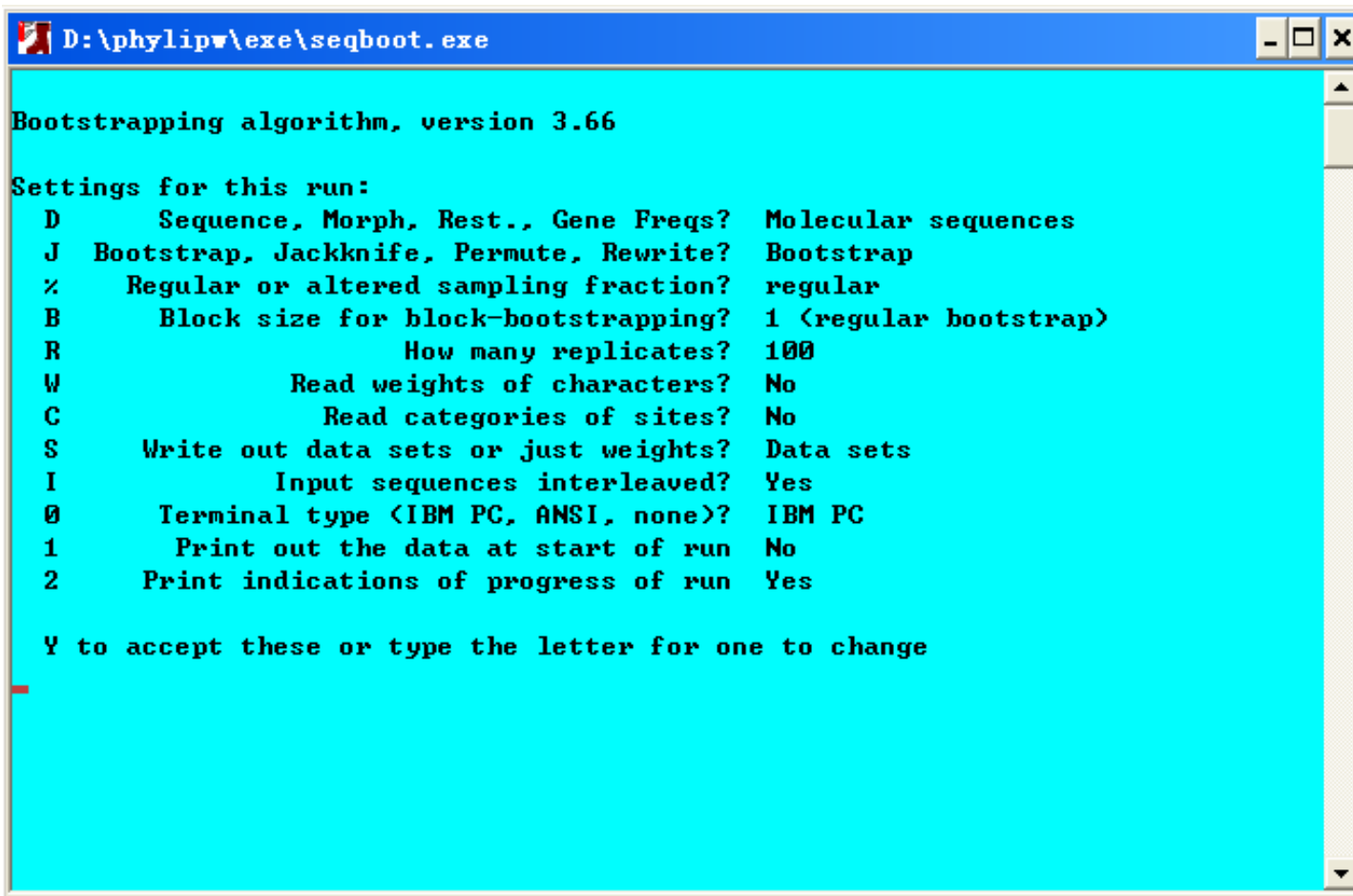


利用Phylip构建进化树使用示例

- ❖ **Phylip**是一个免费的系统发生(phylogenetics)分析软件包，由华盛顿大学遗传学系开发。**Phylip**主要包括以下几个程序组：分子序列组、距离矩阵组、基因频率组、离散字符组、进化树绘制组。
- ❖ 根据分析数据，选择适当的程序
- ❖ 选择适当的分析方法：若分析的是**DNA**数据，可以选择简约法（**DNAPARS**）、似然法（**DNAML**，**DNAMLK**）、距离法（**DNADIST**）等。
- ❖ 进行分析：选择好程序后，执行，读入分析数据，选择适当的参数，进行分析，结果自动保存为**outfile**、**outtree**。

- 
- ❖ 通过**clustalw**比对获得的蛋白序 列推测进化树。
注意：更改输出文件的默认格式，打开输出
PHYLIP格式选项

❖ 用seqboot创造抽样数据（一般100-1000组），运行seqboot，输入上次得到的.phy文件（按路径）



```
D:\phylipw\exe\seqboot.exe

Bootstrapping algorithm, version 3.66

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

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

J选项有三种条件可以选择，分别是Bootstrap、Jackknife和Permute。

MP法

运行protpars.exe, 选择需要改变的参数



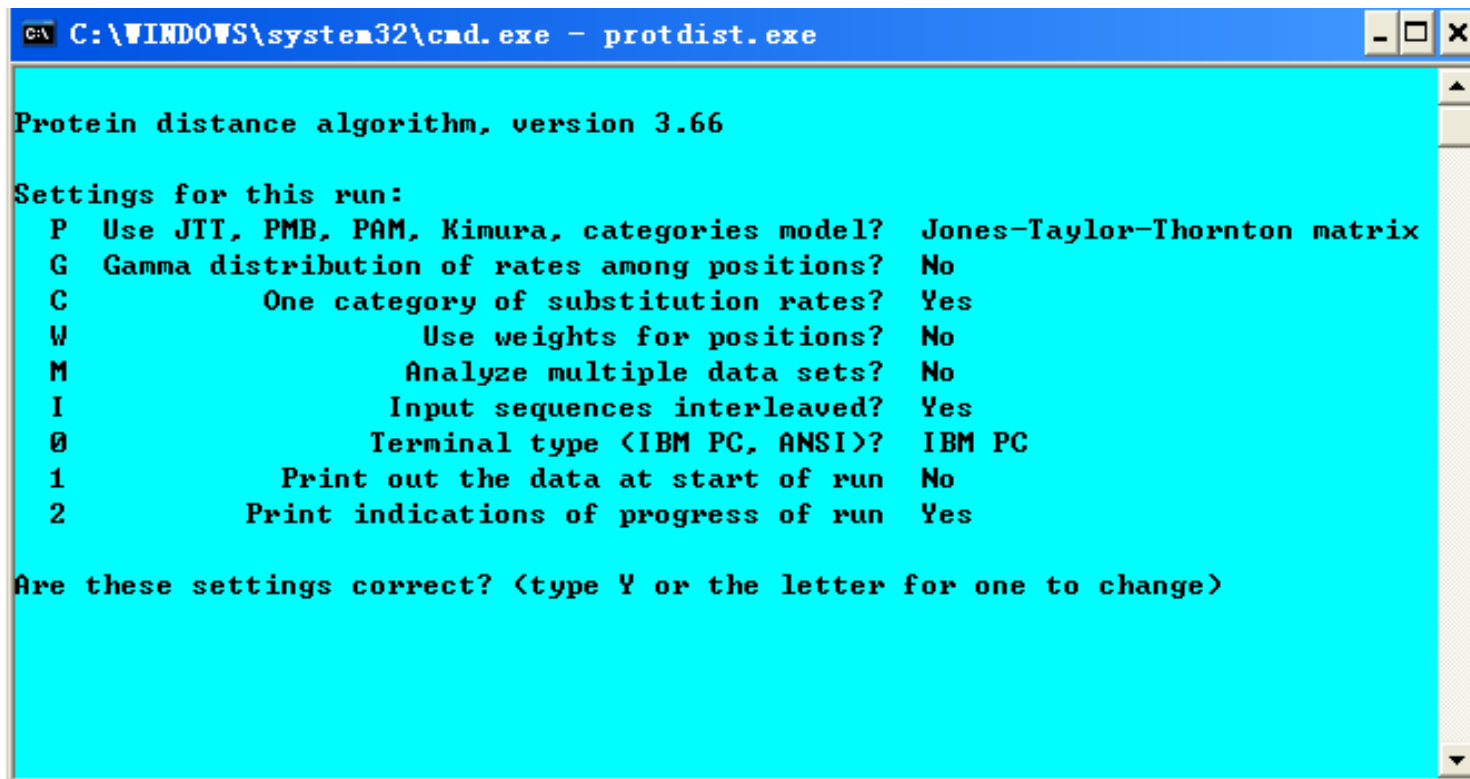
```
D:\phylip\exe\protpars.exe
Protein parsimony algorithm, version 3.66
Setting for this run:
U          Search for best tree?  Yes
J  Randomize input order of sequences?  No. Use input order
O          Outgroup root?  No, use as outgroup species 1
T          Use Threshold parsimony?  No, use ordinary parsimony
C          Use which genetic code?  Universal
W          Sites weighted?  No
M          Analyze multiple data sets?  No
I          Input sequences interleaved?  Yes
@  Terminal type (IBM PC, ANSI, none)?  IBM PC
1  Print out the data at start of run  No
2  Print indications of progress of run  Yes
3          Print out tree  Yes
4          Print out steps in each site  No
5  Print sequences at all nodes of tree  No
6          Write out trees onto tree file?  Yes

Are these settings correct? (type Y or the letter for one to change)
m
Multiple data sets or multiple weights? (type D or W)
d
How many data sets?
100
Random number seed (must be odd)?
17
Number of times to jumble?
7
```

运行CONSENSE, 获得最优树, 结果可用 treeview 查看

NJ法

❖ 运行protdist.exe ,



```
C:\WINDOWS\system32\cmd.exe - protdist.exe

Protein distance algorithm, version 3.66

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

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

然后再运行neighbor

```
C:\WINDOWS\system32\cmd.exe - neighbor.exe

Neighbor-Joining/UPGMA method version 3.66

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

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

运行CONSENSE，获得最优树，结果可用treeview查看



❖ 也可利用**MEGA3.1**构建进化树

❖ **MEGA3.1**是一个关于序列分析以及比较统计的工具包，其中包括距离建树法和**MP**建树法，可自动或手动进行序列比对、推断进化树、估算分子进化率、进行进化假设测验，还能联机的**Web**数据库检索。

❖ 使用（略）

总结

- ❖ 在进行系统发生的推断分析中，最重要的因素不是进行系统发生推断所采用的方法，而是输入数据的质量。
- ❖ 很难准确地建立一个发生树
- ❖ 一定要根据序列信息的特点及目的选择适当的方法与分析软件

**Take time to play with
it**

reference

- ❖ Jin xiong. Essential Bioinformatics. 2006, Cambridge University Press.
- ❖ 孙啸, 陆祖宏等. 生物信息学基础. 2006, 清华大学出版社
- ❖ <http://www.genecool.com/bbs/>
- ❖ <http://evolution.genetics.washington.edu/phylip.html/>



Thank You !

