系统发育分析简介

- 王海秀、谷伟红、张晓娜、徐汇洋
（按姓氏笔画排名）

基础知识

具体步骤

实例分析

基础知识简介

- 有根树和无根树

系统发育树可分为有根树和无根树，有根树是有方向的树，具有一个唯一的根节点，代表树中所有物种的共同祖先；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。

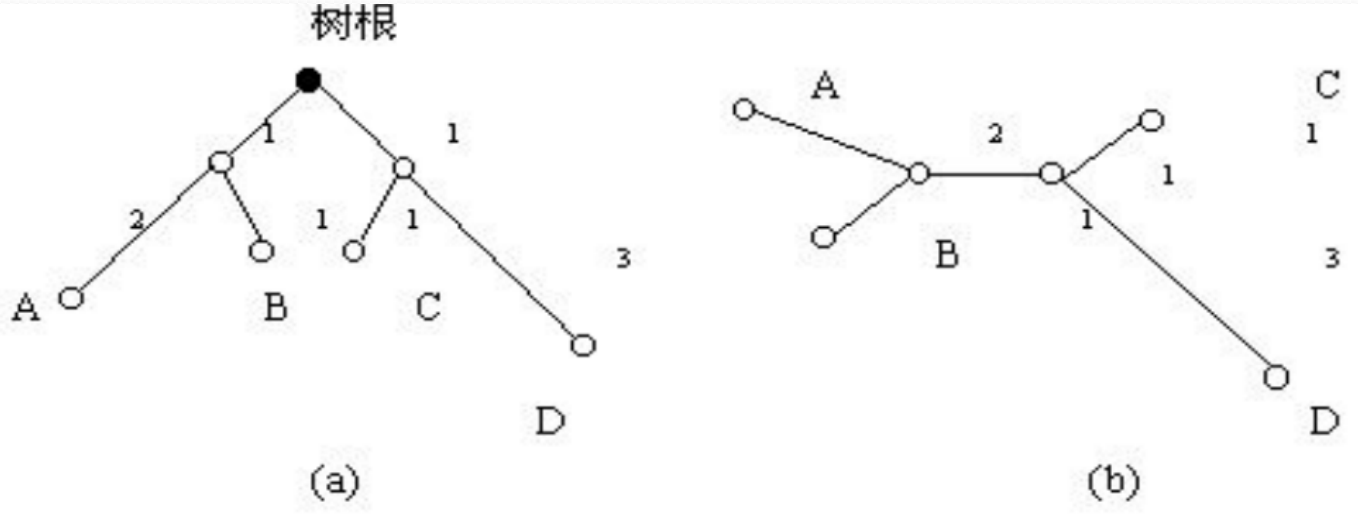


图 6.1 系统发生树。 (a)有根树； (b)无根树。

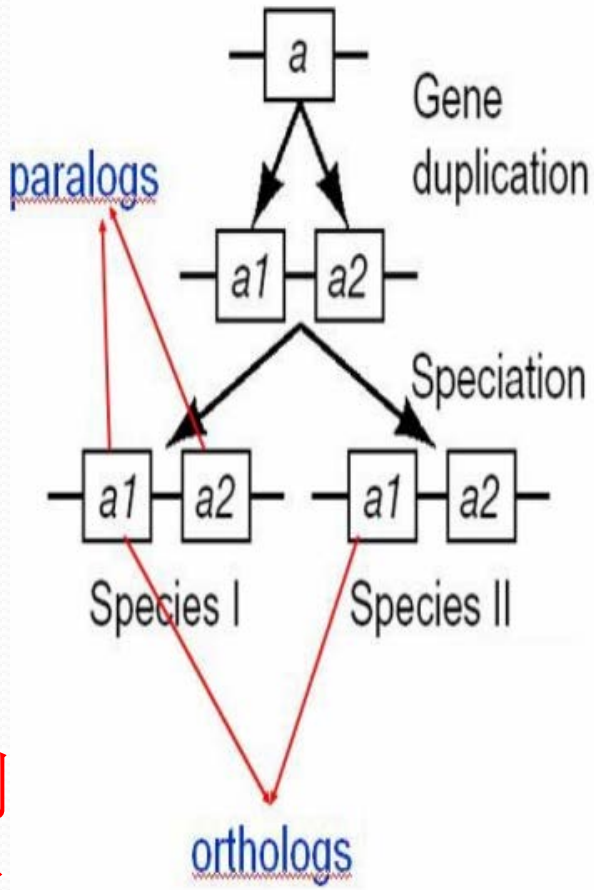
基础知识简介

- 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。

注意：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。



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**, 截取比对的最好的区域作为物种的代表序列;
- 按某种**方法**，算出代表序列两两之间的差异度;
- 基于这些差异度，**绘制系统发生树**;
- 对系统发生树进行**可信度检验**。

具体步骤

- **选择特征分子**

- 既可以用核酸序列又可以用蛋白序列
- 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

具体步骤

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

具体步骤

- **建树方法**
- 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和基于特征的构建方法。

分类	名称	简称
Distance Matrix methods(DM)	平均连接聚类法	UPGMA
	最小进化法	ME
	邻接法	NJ
characters	最大简约法	MP
	最大似然法	ML
	进化简约法	EP

- 常用方法的基本特征

名称	基本特征	适用范围	优点	缺点
邻接法	不需要分子钟假设，是基于最小进化原理，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。	远缘序列，进化距离不大，信息位点少的短序列	假设少，树的构建相对准确，，计算速度快，只得一颗树，可以分析较多的序列，运行速度优于最大简约法	序列上的所有位点等同对待，且所分析的序列的进化距离不能太大
最大简约法	基于进化过程中碱基替代数目最少这一假说，不需要替代模型，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	近缘序列 物种序列的数目≤12	善于分析某些特殊的分子数据如插入、缺失等序列有用。	只适于序列数目N≤12。存在较多回复突变或平行突变时，结果较差。变异大的序列会出现长枝吸引而导致建树错误。
最大似然法	依赖于某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。	特定的替代的模型，远缘序列	很好的统计学基础，大样本时似然法可以获得参数统计的最小方差，在进化模型确定的情况下，ML法是与进化事实吻合最好的建树算法	所有可能的系统发育树都计算似然函数，计算量大，耗时时间长。依赖于合适的替代模型，

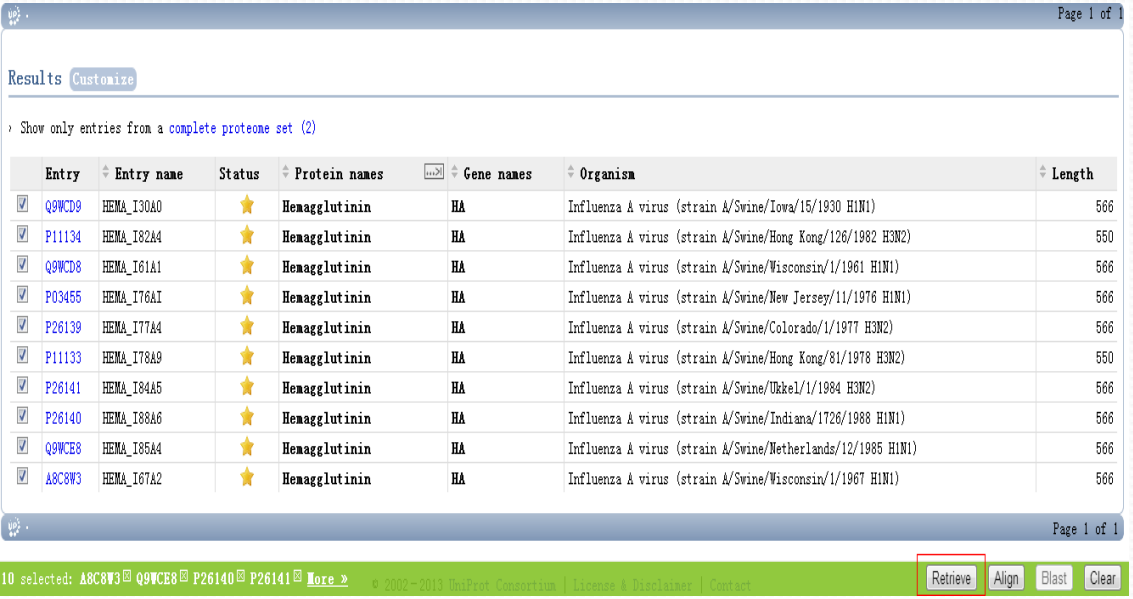
具体步骤

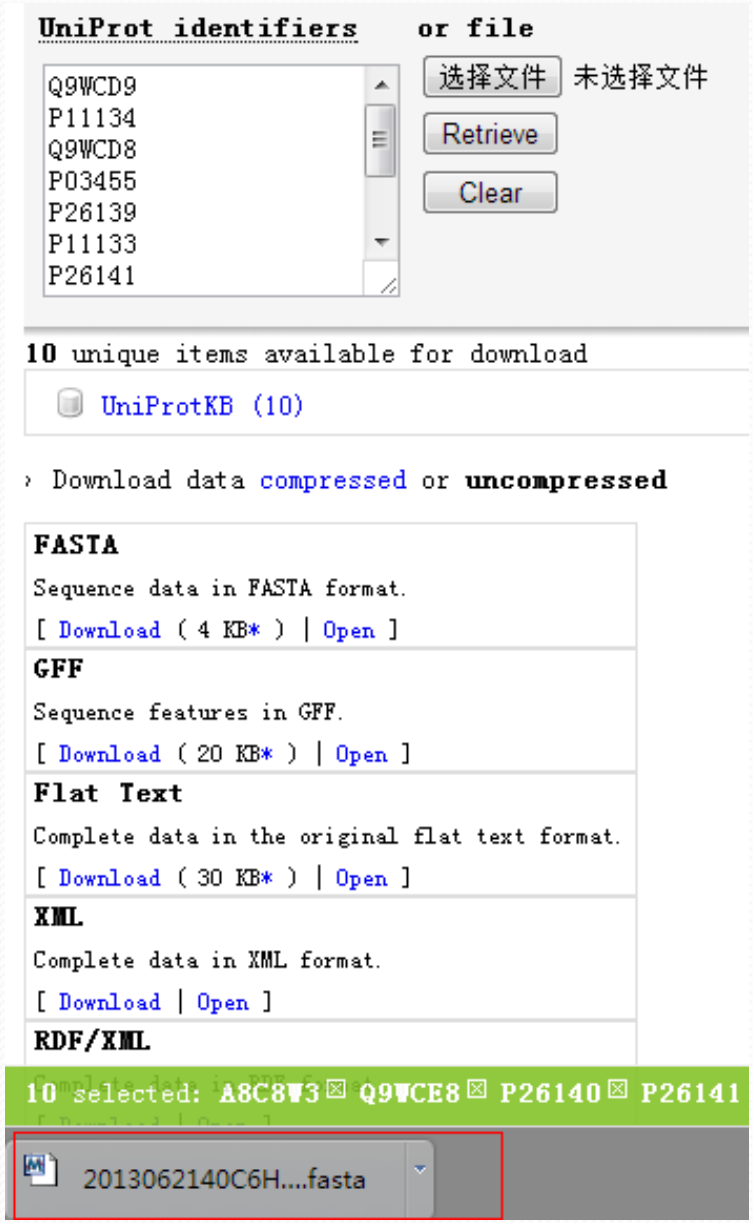
- **可信度检验**
- 常用的三种方法：
- 1. The bootstrap
- 2. Delete-half-jackknifing
- 3. Permuting species within characters
- 注： Bootstrap选项一般都要选择，当 Bootstrap 的值 >70，一般都认为构建的进化树较为可靠。对于进化树的构建，如果对理论的了解并不深入，则推荐使用缺省的参数，并启用Bootstrap 检验。一般情况下，使用两种不同的方法构建进化树，如果得到的进化树基本一致，结果较为可靠。

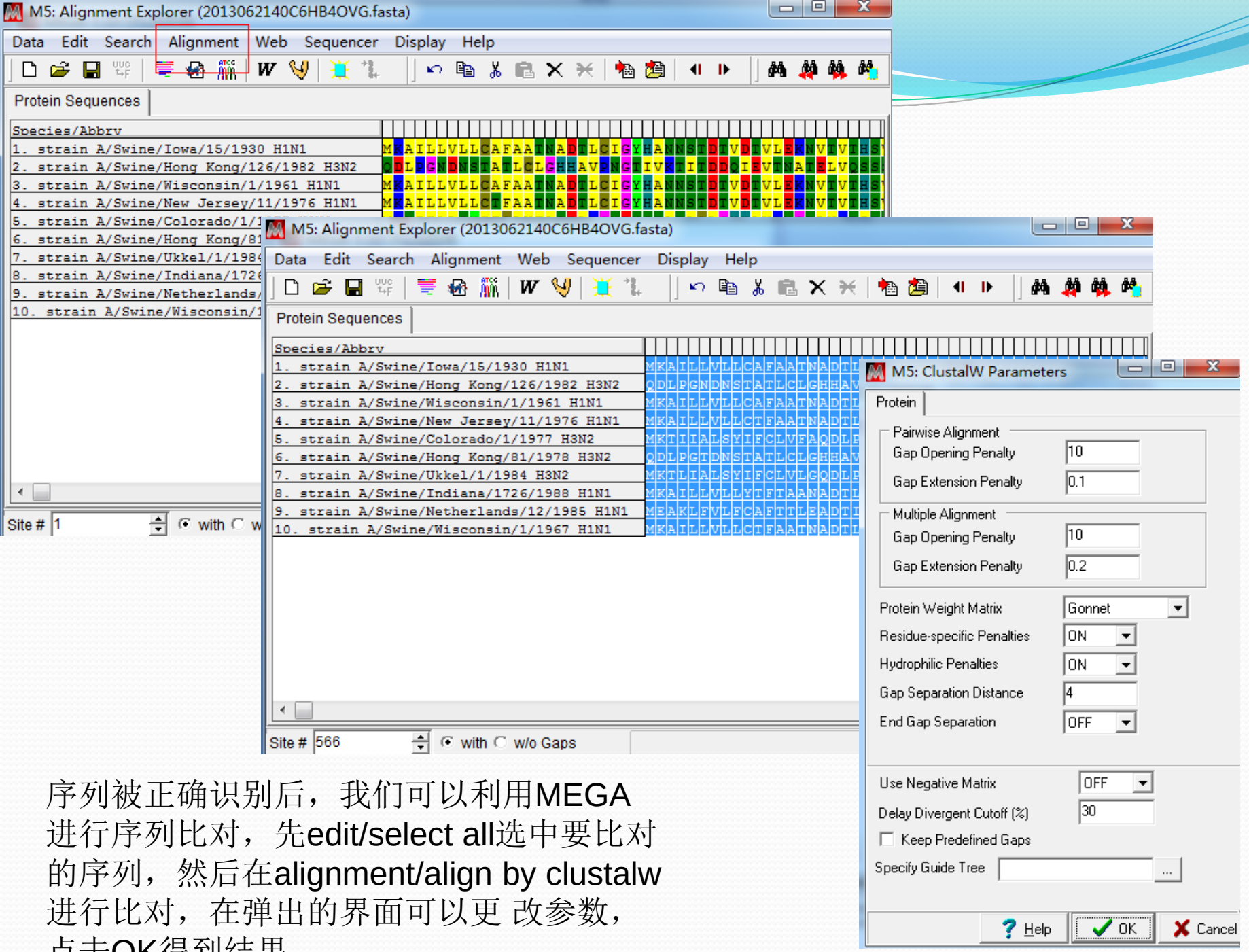
系统进化树构建常用软件汇集

软件名称	说　明
PHYLIP	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	国际上最通用的系统树构建软件之一，美国simthsonion institute 开发，仅适用Apple-Macintosh和 UNIX 操作系统
MEGA	美国宾西法尼亚州立大学MasatoshiNei 开发的分子进化遗传学软件，图形化、集成的进化分析工具，不包括ML
MOLPHY	日本国立统计数理研究所开发，最大似然法构树
PAML	英国 University college London 开发，最大似然法构树和分子进化模型
PUZZLE	应用 quarter puzzling 方法(一种最大简约法)构建系统树
TreeView	英国 University of Glasgow开发，进化树显示工具

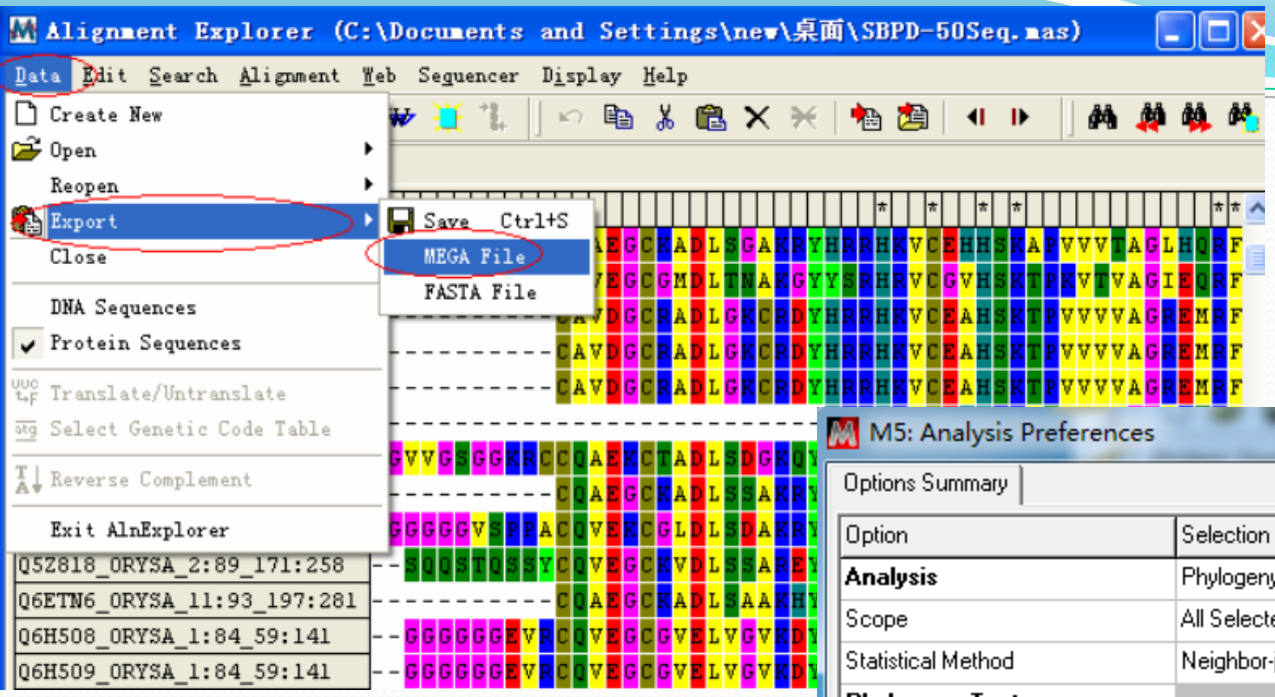
实例分析

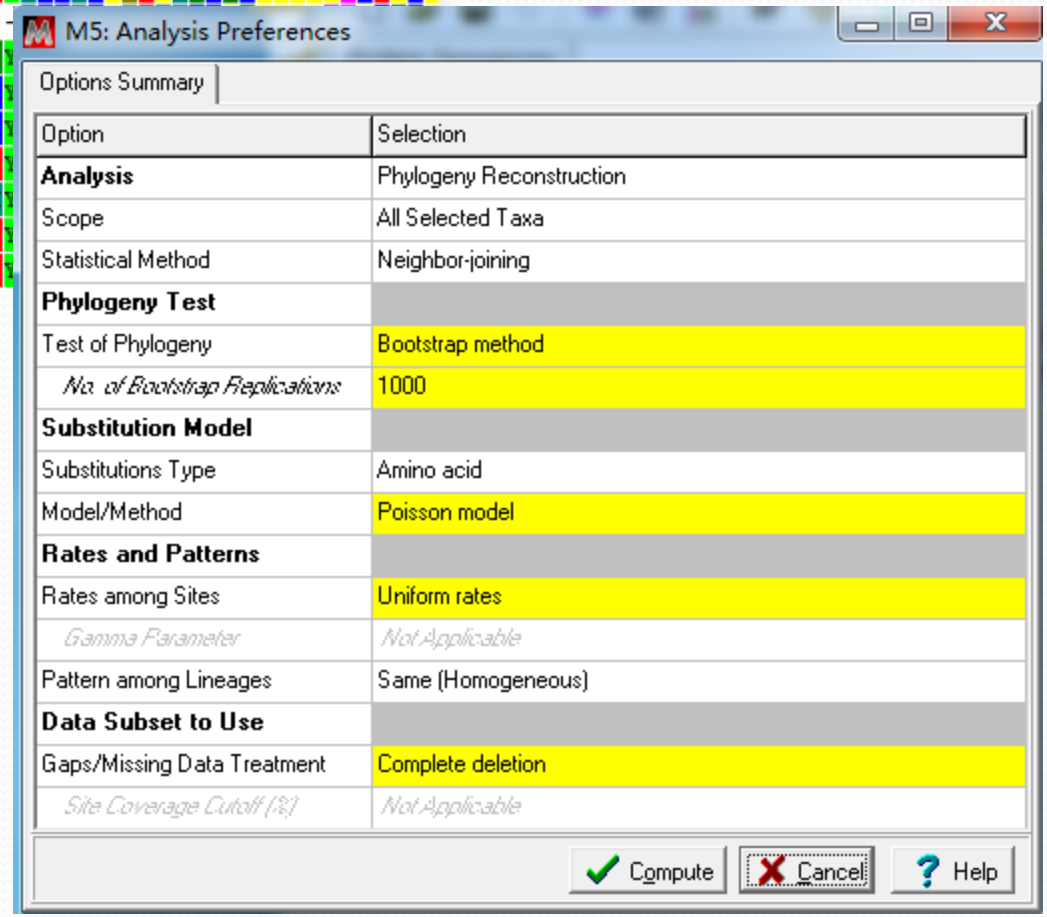

Page 1 of 1
Results Customize
› Show only entries from a complete proteome set (2)
Entry | Entry name | Status | Protein names | Gene names | Organism | Length
Q9WCD9 HEMA_I30A0 Hemagglutinin HA Influenza A virus (strain A/Swine/Iowa/15/1930 H1N1) 566
P11134 HEMA_I82A4 Hemagglutinin HA Influenza A virus (strain A/Swine/Hong Kong/126/1982 H3N2) 550
Q9WCD8 HEMA_I61A1 Hemagglutinin HA Influenza A virus (strain A/Swine/Wisconsin/1/1961 H1N1) 566
P03455 HEMA_I76AI Hemagglutinin HA Influenza A virus (strain A/Swine/New Jersey/11/1976 H1N1) 566
P26139 HEMA_I77A4 Hemagglutinin HA Influenza A virus (strain A/Swine/Colorado/1/1977 H3N2) 566
P11133 HEMA_I78A9 Hemagglutinin HA Influenza A virus (strain A/Swine/Hong Kong/81/1978 H3N2) 550
P26141 HEMA_I84A5 Hemagglutinin HA Influenza A virus (strain A/Swine/Ukkel/1/1984 H3N2) 566
P26140 HEMA_I88A6 Hemagglutinin HA Influenza A virus (strain A/Swine/Indiana/1726/1988 H1N1) 566
Q9WCE8 HEMA_I85A4 Hemagglutinin HA Influenza A virus (strain A/Swine/Netherlands/12/1985 H1N1) 566
A8C8W3 HEMA_I67A2 Hemagglutinin HA Influenza A virus (strain A/Swine/Wisconsin/1/1967 H1N1) 566
Page 1 of 1
10 selected: A8C8W3 Q9WCE8 P26140 P26141 More »
© 2002-2013 UniProt Consortium | License & Disclaimer | Contact
Retrieve
Align
Blast
Clear

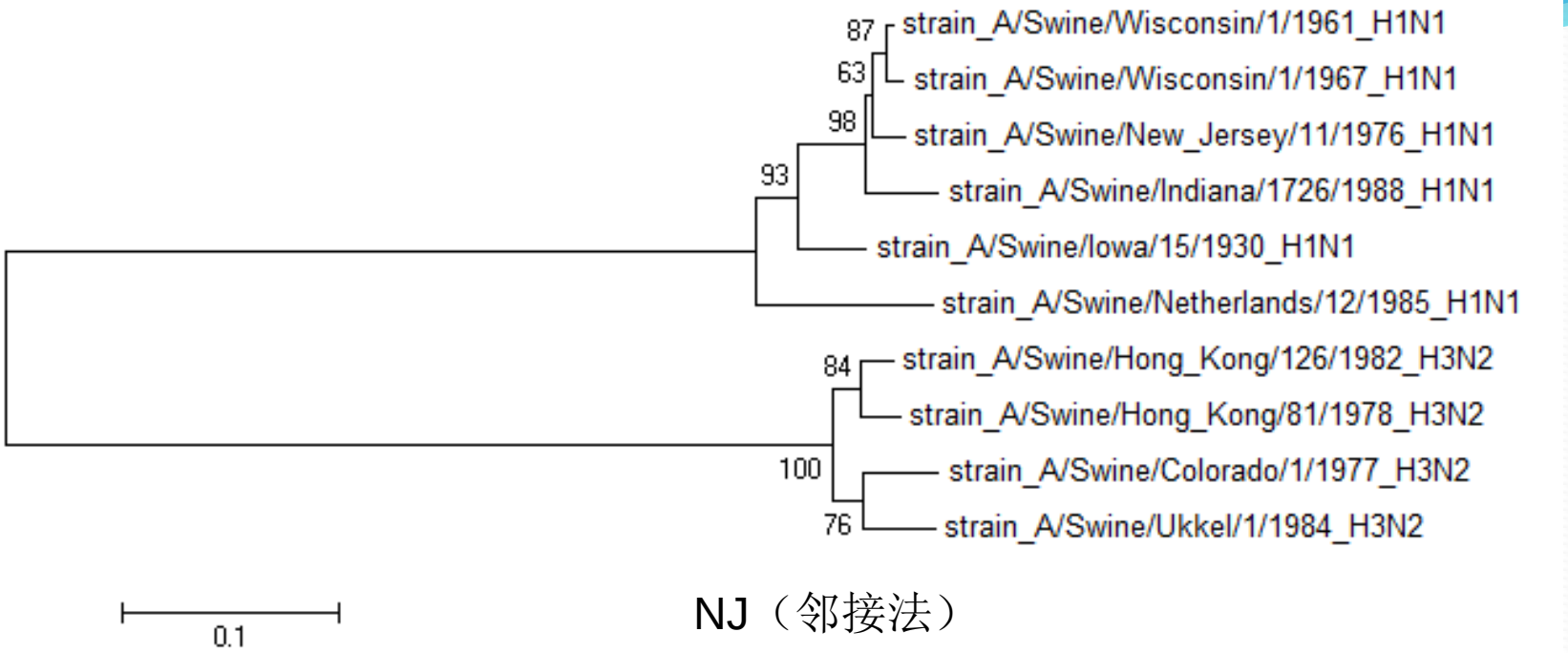

UniProt identifiers
or file
Q9WCD9
P11134
Q9WCD8
P03455
P26139
P11133
P26141
选择文件
未选择文件
Retrieve
Clear
10 unique items available for download
UniProtKB (10)
› Download data compressed or uncompressed
FASTA
Sequence data in FASTA format.
[Download (4 KB*) | Open]
GFF
Sequence features in GFF.
[Download (20 KB*) | Open]
Flat Text
Complete data in the original flat text format.
[Download (30 KB*) | Open]
XML
Complete data in XML format.
[Download | Open]
RDF/XML
10 selected: A8C8W3 Q9WCE8 P26140 P26141
2013062140C6H....fasta



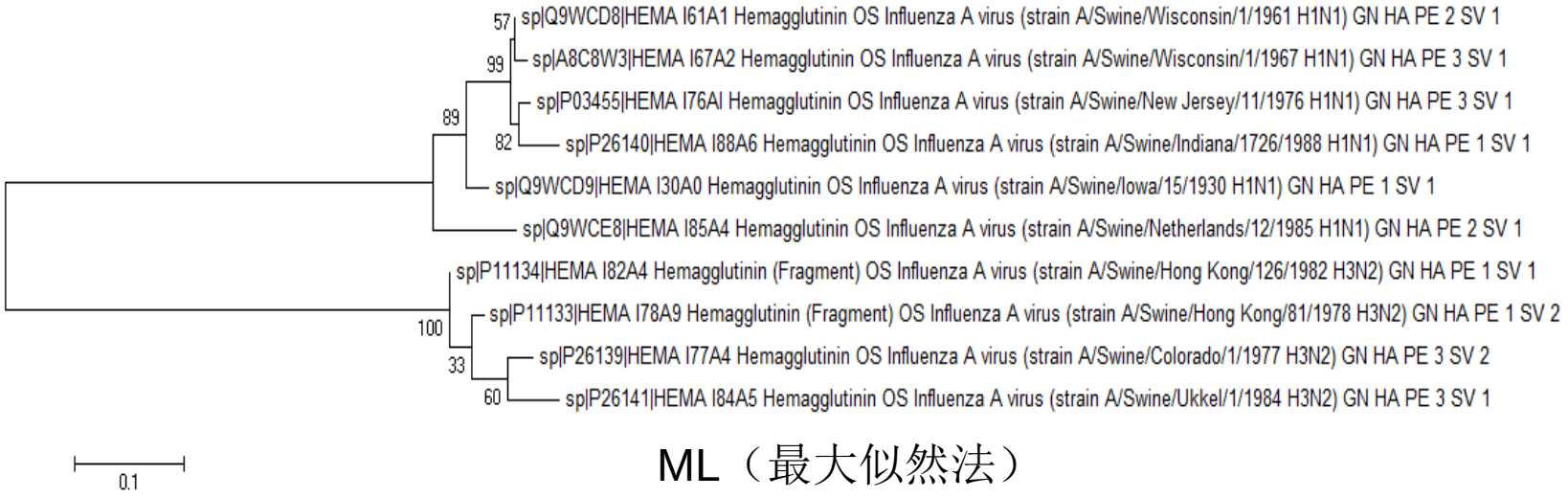
序列被正确识别后，我们可以利用MEGA进行序列比对，先edit/select all选中要比对的序列，然后在alignment/align by clustalw进行比对，在弹出的界面可以更 改参数，点击OK得到结果。


Alignment Explorer (C:\Documents and Settings\new\桌面\SBPD-50Seq.mas)
Data Edit Search Alignment Web Sequencer Display Help
Create New
Open
Reopen
Export
Close
DNA Sequences
Protein Sequences
Translate/Untranslate
Select Genetic Code Table
Reverse Complement
Exit AlnExplorer
Save Ctrl+S
MEGA File
FASTA File
Q5Z818_ORYSA_2:89_171:258
Q6ETN6_ORYSA_11:93_197:281
Q6H508_ORYSA_1:84_59:141
Q6H509_ORYSA_1:84_59:141


M5: Analysis Preferences
Options Summary
Option
Selection
Analysis
Phylogeny Reconstruction
Scope
All Selected Taxa
Statistical Method
Neighbor-joining
Phylogeny Test
Test of Phylogeny
Bootstrap method
No. of Bootstrap Replications
1000
Substitution Model
Substitutions Type
Amino acid
Model/Method
Poisson model
Rates and Patterns
Rates among Sites
Uniform rates
Gamma Parameter
Not Applicable
Pattern among Lineages
Same (Homogeneous)
Data Subset to Use
Gaps/Missing Data Treatment
Complete deletion
Site Coverage Cutoff (%)
Not Applicable
Compute
Cancel
Help



NJ（邻接法）



ML（最大似然法）

注意事项

1. 到目前为止，在进行系统发育分析中，最重要的因素不是采用的建树方法，而是输入数据的质量。数据选择和序列比对都非常重要。因为即使是最复杂的系统发育分析方法都不能校正输入数据的错误。

2. 从尽可能多的角度观察数据。使用三种主要方法的每一个，然后比较它们所建立的进化树的一致性。

3. 外类群对于分析的影响是相当的。使用无可争议的同源物种作为外类群，这个外类群要足够近，以提供足够的信息，但又不能太近以至于和树中的种类相混。

4. 有时候程序可以给出不同的进化树，仅仅是因为序列出现在输入文件的顺序不同。