安徽省教育厅自然科学研究重点项目资助成果

博士论丛 BO SHI CONG LUN

生物序列数值化表征模型的矩阵分解方法及其应用

余宏杰 著

The Approaches to Matrices Factorization for Biological Sequences Numerical Characterization Model and Their Applications

中国科学技术大学出版社

生物序列数值化表征模型的矩阵分解方法及其应用

The Approaches to Matrices Factorization for Biological Sequences Numerical Characterization Model and Their Applications

余宏杰 著

中国科学技术大学出版社

内 容 简 介

本书以生物序列的数值化表征模型所涉及的矩阵分解为核心，以序列的特征信息提取为主要目标，在非序列比对(Alignment-free)的框架下，分别提出了针对DNA/蛋白质序列、基因组序列等的若干个不同的特征信息抽取模型，并将所抽取的特征信息应用于序列的相似度分析. 本书取材广泛，内容新颖，理论与应用紧密结合. 书中所介绍的生物序列的建模方法、矩阵分解抽取其特征信息的研究策略，可供读者在解决实际问题时予以借鉴.

本书适合生物信息学、图像处理、信号处理等领域有关科研人员参考使用.

图书在版编目(CIP)数据

生物序列数值化表征模型的矩阵分解方法及其应用/余宏杰著. —合肥：中国科学技术大学出版社，2014.6

ISBN 978-7-312-03454-1

Ⅰ. 生… Ⅱ. 余… Ⅲ. 生物分析—序列—数据模型—生物信息论 Ⅳ. Q5

中国版本图书馆CIP数据核字(2014)第110630号

出版 中国科学技术大学出版社
安徽省合肥市金寨路96号，230026
http://press.ustc.edu.cn
印刷 安徽省瑞隆印务有限公司
发行 中国科学技术大学出版社
经销 全国新华书店
开本 710 mm× 1000 mm 1/16
印张 13
字数 255千
版次 2014年6月第1版
印次 2014年6月第1次印刷
定价 36.00元

前　言

当前正处于后基因组时代，日益积累的海量数据亟待分析解释，生物信息学便应运而生.其研究内容十分丰富，而其中的序列相似度分析尤为重要.这必然会涉及生物序列的表征方式，其中数值化表征模型会运用到矩阵分解技术等.针对现有的一些序列数值化表征方法普遍存在的不足之处，本书分别从算法设计、数据应用等不同角度，提出若干种有效的用于表征序列的模型.通过与相关研究成果从理论和实验结果上分别加以比较，以期验证所提出算法的有效性.而矩阵分析法显示出的特有的、行之有效的影响力，已渗透到相关领域，并取得较好的应用效果.本书旨在系统地介绍如何从生物序列中抽取特征信息的矩阵模型构建及矩阵分解，以期为相关领域问题的研究提供借鉴.

本书内容安排如下：

(1) 生物序列的图形化表示，为我们提供了一个可供研究序列的可视化工具.为了直观地比较不同的 DNA 序列，提出一种新的特征信息抽取模型，可对序列作图形化表示，并作序列之间的相似度分析.引入变换将每条 DNA 序列用近邻核苷酸矩阵(NNM)来表示.再基于近似联合对角化(AJD)，从每条 DNA 序列变换所得的 NNM 矩阵中抽取特征值作为表征向量(EVV)，视每个 EVV 向量为各自所对应序列的数值描述子(descriptor).基于表征向量 EVV 可得 DNA 序列的二维表征图形.此外，利用 k-均值法将这些表征各条序列的曲线图聚为若干个合理的子类.利用所得向量计算成对距离(pairwise distance)，分析原始序列之间的相似度，从而同步、联合地从多重序列中抽取更多的信息，而非孤立地分析各条序列.

(2) 为了比较不同的基因组序列，提出了新的非比对序列比较方法：考虑到序列具有“序”这一本质属性，基于 16 种不同类型的 2-mer，即双核苷酸(dinucleotides)，定义一种复合变换，能将每条基因组序列转换成 $16\times(L-1)$ 的特征矩阵 $\boldsymbol{M}$.此外，还发现上述变换具有“保序”的特性.由矩阵分析理论，对矩阵 $\boldsymbol{M}$ 施以奇异值分解，来导出 16 维的向量用以描述每条基因组序列，用于对 20 条真哺乳亚纲线粒体基因组序列作相似度分析.

(3) 为解决基因组序列维数较高，直接在低维空间数值表征很困难的问题，书中还提出了具有“保距”特性的基因组序列的非比对模型.先将基因组序列转换成

$16\times(L-1)$的稀疏矩阵 $\boldsymbol{M}$，对所得矩阵 $\boldsymbol{M}$ 施以奇异值分解，便得 16 维“特征值”向量 $\boldsymbol{F}$ 用以表征每条基因组序列. 通过主成分分析(PCA)，将所得的前几个主元用于序列之间的比较. 从理论上证明了：① 模型属于保距变换；② 16 元组向量与最近邻的双核苷酸数目密切相关. 利用“特征值”向量 $\boldsymbol{F}$ 构建了各组哺乳动物基因组序列系统树图. 此外，由主成分分析所得的前两个主元绘制物种的二维“Map图”，用以表征所涉物种间的亲缘关系. 分析结果符合已知的哺乳动物谱系关系，揭示了线粒体基因组以及全基因组序列均能很好地将不同物种区分开来.

(4) 基于所有各种近邻氨基酸(AAA)的分布情况，可将每条蛋白质序列映射成 $400\times(L-1)$的矩阵 $\boldsymbol{M}$. 对 $\boldsymbol{M}$ 施行奇异值分解，从而可得从原始蛋白质序列抽取出归一化的数学描述符 D，其维数为 400. 所得的 400 维归一化“特征”向量(NFV)便于对蛋白质序列作定量分析. 运用蛋白质序列的归一化表示形式，遴选两个典型数据集作相似度分析.

(5) 由于计算开销大的原因，传统的多重序列比对(MSA)不再适合基因组规模上的序列比较. 本书还提出了改进的 K-mer 法：将序列分成若干段，同时将每一段转换成相应的 K-mer. 该算法的关键在于确定出距离测度 d，K 值以及段数 s 的最优组合(d^*，s^*，K^*). 基于从寻优分成的 s^* 个片段的序列转化而来串联在一起的“特征”向量，运用所提出的分段 K-mer 模型(即 s-K-mer)，获得 34 条哺乳动物线粒体基因组序列的系统树状图.

(6) 比较多重基因组序列时，不仅要考虑全局相似性，还须考虑局部相似性. 从信号处理的角度，本书还提出了拟用于基因组序列比较的新算法：先将各条基因组序列分成若干个片段，每段同时转换成相应的基于 K-mer 的向量，此过程可以视为将多重基因组信号经过虚拟传感器(VM)混合后的数值输出，实现了将长度迥异的原始序列转换为等长的向量. 随后，利用基于独立分量分析法(ICA)的变换，可将上述混合输出的向量组向独立主成分投影，由此经过投影抽取器(PE)捕获得到其投影向量；并从理论上严格证明了复合变换具有保距特性. 此外，作为改进，引入双层 VM-PE 模型，以提高相似度分析的性能. 而且经过层级 VM-PE 模型(HVMPE)，大大降低了数据的维度. 利用所提出的 HVMPE 模型，运用于两个线粒体基因组序列数据集作相似度分析.

本书能够顺利地出版，得益于安徽省教育厅自然科学研究重点项目的资助，项目名称为：保序映射算法在序列相似性分析中的应用研究(KJ2013A076). 最后，作者真诚而郑重地感谢安徽科技学院科研处以及中国科学技术大学出版社对本书出版所给予的热忱支持和帮助，在此一并致谢.

余宏杰

2014 年 1 月谨识于中都

此为试读，需要完整PDF请访问：www.ertongbook.com

目　录

1.1.2 两种基本的生物序列

根据上述定义得知，生物信息学研究的基本对象是：核酸和蛋白质序列. 下面就对这两种序列进行简单的介绍.

1. 核酸序列

生物体中核酸分子有两类：脱氧核糖核酸(deoxyribonucleic acid)，简记为DNA；核糖核酸(ribonucleic acid)，简记为RNA. DNA主要分布在细胞核内，少量存在于线粒体中，是生物遗传信息的携带者. RNA大部分存在于细胞质内，小部分分布在细胞核内，与蛋白质的合成有密切关系. DNA分子中有四种碱基，表1.1列出了核酸序列链碱基缩略字符表示及含义.

表1.1 核酸序列链碱基缩略字符表示及含义

符 号	含 义	名 称	符 号	含 义	名 称
a	a	腺嘌呤	k	g或t/u	酮基
g	g	鸟嘌呤	b	g或c或t/u	非a
c	c	胞嘧啶	d	a或g或t/u	非c
t	t	胸腺嘧啶	h	a或c或t/u	非g
r	g或a	嘌呤	v	a或g或c	非t,非u
y	t/u或c	嘧啶	n	a或g或c或t/u，未知，或其他	任何
m	a或c	氨基	u	u	尿嘧啶

实验发现，在任何DNA的组成中，腺嘌呤A和胸腺嘧啶T的含量相同，鸟嘌呤G和胞嘧啶C的含量相同，嘌呤碱基的含量等于嘧啶碱基的含量. 1953年，James Watson与Francis Crick，根据DNA分子的X射线衍射数据，提出了DNA结构的双螺旋模型：每个DNA分子包含两条链，每条链都是由A,C,G和T这四种核苷酸重复组成的线性多聚体. 一条链的碱基与另一条链的碱基配对，A始终和T配对，G始终与C配对. 这种配对称为互补(complement)，即A与T互补，C与G互补. 通常用碱基对(base pair)来表示DNA的长度.

2. 蛋白质序列

生命体的大部分物质是各种各样的蛋白质. 蛋白质有很多种类，例如生化反应中的催化剂酶、构成组织的结构蛋白等. 从结构上说，蛋白质是由氨基酸组成的线性多聚体，氨基酸之间由肽键相连. 表1.2中列出了最常见的构成蛋白质的20种

氨基酸的名称及其缩写.每个氨基酸有一个中心碳原子 C_α,连接到 C_α 上的蛋白质中的肽键由氨基酸 A_i 的羧基碳原子与氨基酸 A_{i+1} 的氮原子连接而成,并且脱去一分子水,故蛋白质序列链中的氨基酸被称为残基(residue).氨基酸之间的肽键构成了蛋白质的骨架,这个骨架由重复的单元— N — C_α—(CO)—构成,每个 C_α 有一个侧链,一端是一个氨基,另一端是一个羧基.习惯上,蛋白质序列的书写顺序是从氨基(N 端)到羧基(C 端).氨基酸序列又称为蛋白质的一级结构;多肽链中氨基酸间形成的氢键使线性链转变为二级结构.二级结构包括规则的 α 螺旋、β 折叠和非螺旋非折叠的环.

表 1.2 蛋白质一级序列中的氨基酸缩略符

符 号	缩 写	含 义	符 号	缩 写	含 义
Ala	A	丙氨酸	Pro	P	脯氨酸
Cys	C	半胱氨酸	Gln	Q	谷氨酰胺
Asp	D	天冬氨酸	Arg	R	精氨酸
Glu	E	谷氨酸	Ser	S	丝氨酸
Phe	F	苯丙氨酸	Thr	T	苏氨酸
Gly	G	甘氨酸	Val	V	缬氨酸
His	H	组氨酸	Trp	W	色氨酸
Ile	I	异亮氨酸	Tyr	Y	酪氨酸
Lys	K	赖氨酸	Asx	B	天冬氨酸或天冬酰胺
Leu	L	亮氨酸	Glx	Z	谷氨酸或谷氨酰胺
Met	M	蛋氨酸	Xaa	X	未知或其他
Asn	N	天冬酰酸			

蛋白质是生命体赖以生存的营养要素,是细胞组织的重要组成部分,几乎所有的生物过程都与蛋白质发生某种联系.根据蛋白质序列的排列顺序和序列信息确定蛋白质的功能是生物学研究的重点.它的主要研究方法可分为两大类:① 利用实际实验的方法来预测,包括 X 光绕射和核磁共振;② 利用理论计算的方法,包括同源建模法、折叠识别法以及从头预测法三种.虽然用实验的方法较为准确,但花费的时间长,而且很多蛋白质难以结晶,因而实验结果也受到技术和设备上的制约;相对而言,用理论计算的方法则可以避免这些缺点,所以发展基于蛋白质序列对结构和功能进行预测的模型很有必要.

此为试读,需要完整PDF请访问: www.ertongbook.com

1.2 生物序列比对概述

1. 生物序列比对的意义

各种不同类型的生物序列之间的比较是生物序列分析的核心问题之一，其最终目的是：寻找、确定不同生物序列的保守区域及变化规律，并由此发现它们的功能、结构特征以及区别所在. 或者说，从核酸以及氨基酸的层次去分析不同生物序列的相同点和不同点，以期能够推测它们的结构、功能以及进化上的联系. 最常用的比较方法是序列比对，它为两个或更多个序列的残基之间的相互关系提供了一个非常明确的图谱.

序列比对(sequence alignment)是通过在序列中搜索一系列单个性状或性状模式来比较两个(双序列比对)或更多个(多重序列比对)序列的方法. 将两个序列写成两行来进行对准(align). 相似或同一性状的(残基或碱基)置于同一列，非同一性状要么放在同一列作为一个错配，要么在另一个序列上对应一个间隔. 在一个最优排列中，非同一性状和间隔的放置，应尽可能地使同一或者相似的性状垂直对齐.

通过残基与残基之间的比对，可以发现：某些位置的氨基酸残基，相对于其他位置的残基具有较高的保守性，这个信息揭示了某些残基对于一个蛋白质的结构和功能是极为重要的. 这些保守的残基，对于保持蛋白质的结构与功能至关重要. 因此，序列比对是从已知获得未知的一个十分有用的方法，比如：将一个新的蛋白质同其他已被深入研究过的蛋白质加以比较，可以推断出此未知蛋白质的结构与功能的某些性质. 不过，仅通过比较分析来推断还不够，结论还须经过实验加以验证.

另一方面，序列比对可用来判断两条序列间的相似性(similarity)，从而判定二者是否同源(homology). 相似性和同源性虽然在某种程度上具有一致性，但二者是完全不同的两个概念. 相似性是指一种很直接的数量关系，比如：部分相同、相似的百分比或其他一些合适的度量；而同源性是指从一些数据中推断出的两个基因在进化上曾具有共同祖先的结论，它是质的判断. 基因之间要么同源，要么不同源.

通常来说，若两条序列相似性很高，则二者往往具有同源关系. 当然，也有可能两条序列的相似度虽很高，但它们并非同源序列. 二者有可能是通过两条不同的进化路径独立获得相同的功能，这在进化中称为趋同(convergence)，这样的序列称为同功序列.

序列比对算法有两种：一是双重比对(pairwise alignment)，即只比较两条序

列；二是多重序列比对(multiple sequence alignment，MSA)，即两条以上的序列同时进行比较. 两条序列比对的实现方法有：① 点阵分析；② 动态规划(dynamic programming，DP)算法；③ 词或 K 串方法(如程序 BLAST：basic local alignment search tool).

2. 序列聚类

序列聚类旨在将序列数据集划分成若干个簇(cluster)，使得每个簇中的序列之间尽可能相似，而不同簇中的序列之间尽可能不相似. 以蛋白质为例，随着蛋白质序列数据的增长，通过实验来确定蛋白质性质的速度，远远赶不上蛋白质序列测序的速度，日益积累的海量蛋白质序列亟待确定其功能. 为了预测一个未知性质的蛋白质的功能，可将其与已知生化性质的蛋白质加以比较，根据序列的相似性将其分配到不同蛋白质家族中. 为进一步分析相似蛋白质在功能上的差异，通过聚类分析，可将每个家族的蛋白质聚成不同的子家族(subfamily)，各个子家族内的蛋白质彼此间具有功能上的相关性，容易用实验进行分析. 假若一个未知功能的蛋白质属于某个子家族，并且子家族内的蛋白质功能已知，则有很大把握推断出此蛋白质也具有这种功能. 相反，倘若一个蛋白质新近被发现具有某种生物功能，则此功能也可推广到所有的子家族中的其他序列上. 另一方面，隶属于同一家族的蛋白质，通常彼此起源于某个共同的祖先. 把蛋白质序列聚成相关的类，将有助于进一步地分析蛋白质的进化关系.

另外，在序列数据库中，数据冗余是个非常普遍的问题，这些冗余的数据通常难以提供额外的信息. 对于诸多应用来说，只需考虑它们的代表序列即可. 例如，国家生物技术信息中心(national center for biotechnology information，NCBI)使用 BLAST 将数据库中的相同序列加以合并，构建出一个非冗蛋白质数据库(non-redundant protein database，NR). 通过对蛋白质数据库进行聚类，将大于某个相似度(比如 90%)的序列聚成一个类，然后选出其中的一个蛋白质作为代表序列，将会大大地减小数据库的规模. 使用这种约减后的数据库进行搜索会节省时间，且不会降低识别敏感性，甚至有可能提高对远亲蛋白质的识别.

1.3 基于序列比对的系统发育树构建方法

1.3.1 分子进化研究的基本方法

对于进化研究，主要通过构建系统发育过程有助于通过物种间隐含的种系关

系揭示进化动力的实质.

表型的(phenetic)和遗传的(cladistic)数据有着明显差异.早在 1973 年,Sneath 和 Sokal 就已将表型性关系定义为根据物体一组表型性状所获得的相似性,而遗传性关系含有祖先的信息,因而可用于研究进化的途径.这两种关系可用系统进化树(phylogenetic tree)或树状图(dendrogram)来表示.表型分枝图(phenogram)和进化分枝图(cladogram)两个术语已分别用于表示根据表型性的和遗传性的关系所建立的关系树.进化分枝图可以显示事件或类群间的进化时间,而表型分枝图则不需要时间概念.文献中,更多的是使用"系统进化树"一词来表示进化的途径,另外还有系统发育树(phylogenetic tree)、物种树(species tree)、基因树(gene tree)等一些相同或含义略有差异的名称.

系统进化树分有根(rooted)树和无根(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.用于构建系统进化树的数据有两种类型:一是特征数据(character data),它提供了基因、个体、群体或物种的信息;二是距离数据(distance data)或相似性数据(similarity data),它涉及的则是成对基因、个体、群体或物种的信息.距离数据可由特征数据计算获得,但反过来则不行.这些数据可由矩阵的形式来加以表达.距离矩阵(distance matrix)是在计算得到的距离数据基础上获得的,距离的计算总体上是要依据一定的遗传模型,并能够表示出两个分类单位间的变化量.系统进化树的构建质量依赖于距离估算的准确性.

以 Bioedit-Mega 建树法为例,其主要步骤简单介绍如下:

(1) 将所测得的序列在 NCBI 上进行比对,此处不再赘述;

(2) 选取序列保存为 text 格式;

(3) 运行 Bioedit,使用其中的 Clustal W 进行比对;

(4) 运用 Mega 4 建树,首先将前面的文件转化为 mega 格式,然后进行激活,最后进行 N-J 建树.

1.3.2 构建系统进化树的详细步骤

1.3.2.1 建树前的准备工作

1. 利用 BLAST 获取序列

BLAST 是目前常用的数据库搜索程序,意思是"基本局部相似性比对搜索工

具”. 国际著名生物信息中心均提供了基于 Web 的 BLAST 服务器. BLAST 算法的基本思路：首先，找出被检测的序列与目标序列之间相似性程度最高的片段；然后，作为内核向两端延伸，以便找出尽可能长得相似的序列片段.

首先，登录到提供 BLAST 服务的常用网站，比如：国内的 CBI、美国的 NCBI、欧洲的 EBI 和日本的 DDBJ 等. 这些网站提供的 BLAST 服务在界面上相差无几，但所使用的算法、程序有所差异. 它们都有一个大的文本框，用于粘贴需要搜索的被检测序列. 将序列依照 FASTA 格式（即：第一行为说明行，是以“>”符号开始的，后面接着是序列的名称、说明等；其中，“>”是格式规定必需的，名称以及说明等均可以是任意形式，换行之后便是序列）粘贴到那个大的文本框处，选择完合适的 BLAST 程序和数据库，便可以开始搜索了. 若是 DNA 序列，通常多数选择以 BLASTN 来搜索 DNA 数据库.

以 NCBI 为例，先登录 NCBI 主页，依次通过：点击 BLAST→点击 Nucleotide →nucleotide BLAST（blastn）→在 Search 文本框中粘贴被检测的序列→点击 BLAST! →点击 Format 等步骤的操作，最后得到 BLAST 的结果.

关于 BLASTN 的结果（参数意义）分析如下：

> gi | 28171832 | gb | AY155203. 1 | Nocardia sp. ATCC 49872 16S ribosomal RNA gene, complete sequence

Score = 2020 bits (1019), Expect = 0.0

Identities = 1382/1497 (92%), Gaps = 8/1497(0%), Strand = Plus/Plus

其中，主要指标的含义如下.

Score：是提交的被检测序列和搜索出的目标序列之间的匹配结果的分值，该值越高说明二者越相似.

Expect：是比对的期望值. 比对效果越好，则期望值越小；一般地，在核酸层次上的比对，当期望值小于 10^{-10} 时，便可认为此时的比对效果很好了，多数情况下为 0.

Identities：是提交的序列和参比序列的相似性，上述示例指的是 1 497 个核苷酸中，两条序列之间有 1 382 个相同.

Gaps：一般翻译成空位，指的是未能匹配成功的碱基数目.

Strand：指序列链的方向. Plus/Minus 意味着提交的序列和参比序列是反向互补的，而 Plus/Plus 则表明二者皆为正向.

2. FASTA——序列的格式

缘于 EMBL 和 GenBank 的序列数据格式较为复杂，所以为了分析的方便，出现了十分简单的 FASTA 数据格式. FASTA 格式又称为 Pearson 格式，该种序

列格式约定：序列的标题行须以大于号“>”开头，而下一行开始为具体的序列. 一般地，每行的字符数不宜超过 60 或 80 个，以便于程序的处理. 各条核酸和蛋白质序列只需按照该格式连续列出即可. 其中，“>”为 Clustal X 默认的序列输入格式，必不可少. 随后可以是种属名称，亦可是序列在 Genbank 中的登录号(Accession No.)，允许自定义编号也可以，不过需要注意的是名字不能太长，通常是由英文字母和数字组成的，一开始的几个字母最好不能相同，因为有时 Clustal X 程序只将前几位字符默认为该序列的名称；回车换行后的内容便是序列. 将被检测的序列和搜索到的同源序列(目标结果)，以 FASTA 格式编辑成为一个文本文件(例如：C:\temp\test.txt)，即可导入 Clustal X 等程序进行后续的比对建树.

1.3.2.2 构建系统树的相关软件和操作步骤

进化树构建的主要步骤是：序列比对→建立位点取代模型→建立进化树→进化树的评估. 鉴于以上对于构建系统树的评价，以下主要介绍针对 N-J 算法的系统树构建的相关软件及其相应的操作步骤.

1. 利用 Clustal X 构建 N-J 系统树的过程

(1) 打开 Clustal X 程序，载入源文件.

(2) 进行序列比对.

(3) 掐头去尾：将开始区域和末尾处长短不同的序列剪切整齐. 此处，由于测序引物不尽相同，所以比对之后序列参差不齐. 一般来说，要“掐头去尾”，以避免因序列前后参差不齐而错误地增加了序列间的差异. 剪切后的文件存为 ALN 格式.

(4) 重新载入剪切后的序列.

(5) 输出树的参数选项.

(6) 绘制 N-J 树.

(7) 树形视图.

通常需要对进化树进行适当地编辑，这时首先要编辑-复制至 PowerPoint 上，然后再复制至 Word 上，以便进行图片编辑. 如果直接复制至 Word 上则会显示乱码，而且进化树不能被正确地显示出来.

2. 利用 Mega 建树的过程

Clustal X 虽然可以构建系统树，但其输出的结果比较粗放，故现在一般很少用它构树，而 Mega 因为操作简单，输出结果比较美观，故而被很多研究者用以作为构建系统树的首选. 主要步骤如下：

(1) 首先用 Clustal X 软件作序列比对,剪切后生成 C:\temp\test.aln 文件(同上);

(2) 再打开编辑程序 BioEdit,将目标文件格式转化(另存)为 FASTA 格式;

(3) 再打开 Mega 程序,转化为 mega 格式并激活目标文件,经过计算,最终得到结果;

(4) Image-Copy to Clipboard→粘贴至 Word 文档进行编辑.

另外,Subtree 按钮中还提供了多个命令,可以用来对生成的进化树进行编辑,Mega 窗口左侧提供了很多快捷键方便使用;View 中则给出了多个树型的模式.下面只介绍几种最常用的.

Subtree-Swap:任意两个相邻的分支互换位置;

Flip:所选分支翻转 180 度;

Compress/Expand:合并/展开多个分支;

Root:定义外群;

View Topology:只显示树的拓扑结构;

Tree/Branch Style:多种树型间的转换;

Options:关于树的诸多方面的改动选项.

3. 利用 Treecon 建树的过程

首先,打开 Clustal X→File-Load→File→Save Sequence as …(Format→Phylip;Save from residue→1 to 末尾;Save sequence as).

其次,打开 Treecon 程序:

(1) 进行估计(Distance estimation);

(2) 树拓扑的推断(Infer tree topology):

点击 Infer tree topology→Start inferring tree topology→Method→Neighbor→joining,Bootstrap analysis→Yes,OK;

(3) 确定无根树的根节点(Root unrooted trees):

点击 Root unrooted trees→Start rooting unrooted trees,Outgroup opition→single sequence(forced),Bootstrap analysis→Yes,OK;

(4) 绘制系统树(Draw phylogenetic tree):

点击 Draw phylogenetic tree,File→Open→(new) tree,Show→Bootstrap values/Distance scale.

最后,File→Copy,粘贴至 Word 文档,编辑.

注意 Treecon 的操作过程似乎看起来比 Mega 的过程要烦琐些,且运算速度明显不及 Mega.但是如果参数选择一样,则用它构建出来的系统树和 Mega 构建

此为试读,需要完整PDF请访问: www.ertongbook.com