

<<理解生物信息学>>

图书基本信息

书名：<<理解生物信息学>>

13位ISBN编号：9787030328328

10位ISBN编号：7030328329

出版时间：2012-1

出版时间：科学出版社

作者：Robert F?Weaver 著，李亦学 等译

页数：588

字数：1263000

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：<http://www.tushu007.com>

## <<理解生物信息学>>

### 内容概要

本书是一本集生物信息学专业参考书和教材于一体的书，共分为7部分：基础知识、序列联配、进化过程、基因组特征、二级结构、蛋白质三级结构、细胞和组织，以及附录和字符表等。每部分由不同章节构成，大多数章节可以被归为应用章节或理论章节。因此在每部分开始时，都有应用章节，描述了特定研究领域较实用的方面。理论章节则紧随其后，解释了其科学、理论基础以及在已有应用中所使用的技术。本书还提供了思维导图、流程图、扩展阅读等其他书不常见的内容，以供读者能够在每一章、每一节开始时对整体内容有所把握，并能够了解更多扩展知识、发展技能的参考文献。

## &lt;&lt;理解生物信息学&gt;&gt;

## 作者简介

Robert

F·Weaver出生于美国堪萨斯州的首府托皮卡市，在弗吉尼亚州的阿灵顿地区长大。

1964年在俄亥俄州的奥斯勒学院获得化学学士学位。

1969年在杜克大学获得生物化学专业博士学位，此后他在加州大学旧金山分校从事了两年的博士后研究工作，师从William

J?Rutter教授研究真核生物RNA聚合酶的结构。

1971年他受聘于堪萨斯大学，任生物化学助理教授，后晋升为副教授，并于1981年任教授。

自1984年以来，Robert

F?Weaver一直担任生物化学系的系主任，1995年开始担任文理学院副院长。

李亦学，研究员，博士生导师。

1982年2月毕业于新疆大学，获物理学学士学位，1987年10月毕业于新疆大学，获理论物理学硕士学位，1996年10月毕业于德国海德堡大学理论物理研究所，获理论物理博士学位。

1996年11月至1997年3月在德国斯图加特大学第三计算机应用研究所从事计算数学博士后研究。

1997年4月至2000年6月在欧洲分子生物学实验室（EMBL）从事生物物理博士后研究。

2000年7月回国，现任中科院上海生命科学研究院生物信息中心主任，2002年7月任上海生物信息技术研究中心主任，2006年12月任中科院系统生物学重点实验室副主任。

李亦学研究员主要研究方向为生物信息学，2000年回国后，先后主持和承担了国家“九五”863计划生物技术领域《生物信息学数据库开发和建设》重大项目。

作为首席科学家主持了中科院《生物信息学重大基础理论与应用》重大研究项目；国家973《重大疾病相关蛋白质组学

## &lt;&lt;理解生物信息学&gt;&gt;

## 书籍目录

译者序

前言

给读者的短笺

致谢名单

第1部分 基础知识

第1章 核酸的世界

1.1 DNA和RNA的结构

DNA分子是由4种不同类型的碱基组成的线性多聚体

两条互补 DNA链通过碱基配对形成双螺旋

RNA分子通常为单链结构，但在某些情况下可形成碱基配对结构

1.2 DNA？

RNA和蛋白质：中心法则

DNA是信息载体，而RNA则是信使

信使RNA根据遗传密码翻译产生蛋白质

翻译过程涉及了含DNA和RNA的核糖体的转移

1.3 基因结构和基因调控

特定的定位序列能和RNA聚合酶结合，并识别转录起始点

真核生物中的转录起始信号远比细菌中复杂得多

真核生物 mRNA转录物在翻译前需经历一系列修饰

翻译的调控

1.4 生命与进化之树

主要生命形式的基本特征

突变可以改变核苷酸序列

总结

名词解释

扩展阅读

第2章 蛋白质结构

2.1 初级结构和二级结构

我们可从多个不同水平考察蛋白质结构

氨基酸是蛋白质的组成单位

侧链决定了氨基酸化学和物理特性的不同

蛋白质链中的氨基酸通过肽键共价连接

蛋白质的二级结构由 螺旋？

链构成

在蛋白质结构中已发现了几种不同类型的 折叠片

螺旋和链通过转角？

发夹结构和环连接

2.2 对生物信息学的启发

某些氨基酸倾向于形成特定的结构单元

从进化角度帮助序列分析

蛋白质结构的计算和可视化

2.3 蛋白质通过折叠形成紧凑的结构

蛋白质的三级结构是通过多肽链的路径来定义的

蛋白质折叠的稳定状态是能量最低的状态

很多蛋白质是由多个亚基组成的

## <<理解生物信息学>>

总结

名词解释

扩展阅读

第3章 数据库的处理

3.1 数据库的结构

平面文件数据库以文本文件的方式存储数据

关系数据库广泛应用于存储生物信息

XML的灵活性可以确定定制的数据分类

一些用于生物数据的其他数据库结构

数据库可以通过本地访问或通过互联网相互链接

3.2 数据库类型

数据库中不仅仅是数据

原始数据和衍生数据

我们如何定义和链接事物的重要性：本体

3.3 数据库搜索

序列数据库

芯片数据库

蛋白质相互作用数据库

结构数据库

3.4 数据质量

冗余性对一些应用特别重要

自动化方法可用于检查数据的一致性

初步的分析和注释通常是自动化完成的

为了产生高质量的注释经常需要人为干预

数据库更新和条目注释版本号的重要性

总结

名词解释

扩展阅读

第2部分 序列联配

第4章 产生和分析序列联配

4.1 序列联配的原理

联配是在两个或更多序列的相同区域寻找最大相似性的任务

联配可以揭示序列间的同源性

比较蛋白质序列比核酸序列更容易检测同源性

4.2 联配分值

一个联配的质量是通过给予一个量化的分值来衡量的

量化两个序列间的相似性的最简单的方法是百分数

基于一致度的点图可以可视化地评价相似性

真正的匹配不必相同

最低一致度比可以被接受为具有显著性

对于打分联配有许多不同的方法

4.3 替代矩阵

使用替代矩阵对每个排列后的序列位点分配一个单独的值

PAM 替代矩阵使用密切相关的蛋白质序列集的替代频率

BLOSUM 替代矩阵使用了局部高度保守区域序列的突变数据

替代矩阵的选择取决于要解决的问题

4.4 插入空缺

## &lt;&lt;理解生物信息学&gt;&gt;

在序列插入空缺以达到和另一条序列的相似度最大, 需要罚分制度

动态规划算法可以决定引入最优空缺

#### 4.5 联配类型

对于不同情况采用不同类型的联配

多重序列联配能同时比较一些相似序列

有几种不同的技术可构造多重联配

多重联配可以提高低相似性序列联配的精确度

ClustalW 可以对 DNA和蛋白质序列进行全局联配

通过合并一些局部联配可以构建多重联配

增加新信息可以改进联配

#### 4.6 检索数据库

已开发了快速而准确的搜索算法

FASTA格式是一个基于较短的相同片段

匹配的快速的数据库搜索方法

BLAST的基础在于发现非常相似的短片段

对不同的问题采用不同版本的BLAST和FASTA

PSI?BLAST基于配置文件的数据搜索

SSEARCH 是一个严格的联配方法

#### 4.7 搜索核酸或蛋白质序列

可直接使用或翻译后的 DNA或 RNA序列

必须测试数据库的匹配质量, 以确保其不可能是偶然发生

选择一个适当的卷值的阈值有助于限制数据库搜索

低复杂度区域可以将同源性搜索复杂化

不同的数据库可以用来解决具体问题

#### 4.8 蛋白质序列模体或模式

建立数据库的模式需要专业知识

BLOCKS数据库包含自动编译的保守蛋白质序列的多重联配的较短序列模块

#### 4.9 使用模式和模体搜索

可以在PROSITE数据库中搜索蛋白质的模式和模体

基于模式的PHI?BLAST程序同时搜索同源性和模体匹配

可以使用PRATT从多条序列产生模式

PRINTS数据库包括了指纹图谱, 描述一个蛋白质家族的一些保守模体

Pfam数据库定义了蛋白质家族的表达谱

#### 4.10 模式和蛋白质功能

可以搜索蛋白质上特定的功能位点

序列比较不是唯一分析蛋白质序列的途径

总结

名词解释

扩展阅读

### 第5章 序列比对及数据库搜索

#### 5.1 替换矩阵和打分

联配分值用于衡量公共进化祖先的似然性

PAM (MDM)替代打分矩阵用于探索蛋白质进化起源

BLOSUM 矩阵用于寻找保守的蛋白质区域

用于核苷酸联配的打分矩阵需由相似的方式得到

替换打分矩阵必须适用于特定的联配问题

插入空缺的打分相对替换而言使用了更为启发式的方法

## &lt;&lt;理解生物信息学&gt;&gt;

## 5.2 动态规划算法

使用改进后的 Needleman?Wunsch算法构建全局最优联配

对动态规划算法的简单改进就能用于局部序列联配

不计算完整的矩阵，牺牲精确度提高时间效率

## 5.3 索引技术和近似算法

后缀树定位和独特及重复序列的位置

散列索引是一种技术，列出了所有k的起始位置元组 ( $k^?$ tuples)

FASTA算法使用哈希算法和快速链接进行数据库搜索

BLAST算法利用了有限状态自动机

直接比较核酸序列和蛋白质序列，需要对BLAST和FASTA进行特殊的调整

## 5.4 联配分值的显著性

有空缺局部联配的统计可以按相似的算法进行

## 5.5 联配全基因组序列

有效索引和扫描全基因组序列对高等生物序列比对至关重要

密切关联的物种基因组之间复杂进化关系需要创新的联配算法

总结

名词解释

扩展阅读

## 第6章 模式？

序列和多序列比对

## 6.1 序列和序列标记

位置特异性分数矩阵是得分矩阵的扩展

解决构建PSSM时数据缺失问题的方法

PSI?BLAST是一个序列数据库检索程序

将序列表现为序列标记

## 6.2 谱式隐马尔可夫模型

用于序列比对的 HMM 的基本结构

利用联配序列建立 HMM 参数

利用谱式 HMM 给序列打分：最大可能路径以及所有路径的总和

利用未联配序列评估 HMM 参数

## 6.3 序列联配

利用联配比较两个PSSM

联配谱式 HMM

## 6.4 利用序列递增 (gradual sequence addition) 的多序列比对

序列添加的顺序是基于评估合并联配错误可能性而决定的

许多不同的打分策略用于建立多序列联配

多序列联配是利用向导树以及谱式方法构建的，且可能进一步改进

## 6.5 其他获得多序列联配的方法

多序列联配程序 DIALIGN联配无间隙的区段

利用遗传算法的SAGA多序列联配方法

## 6.6 序列模式发现

在多序列联配中查找模式：eMOTIF和AACCC

序列中共有模式的概率查询：Gibbs和MEME

总结

名词解释

扩展阅读

## 第3部分 进化过程

## &lt;&lt;理解生物信息学&gt;&gt;

## 第7章 重现进化历史

## 7.1 系统发生树的结构和解释

系统发生树重建进化关系

用几种方式描述树的拓扑结构

一致树和可信树报告拓扑结构的比较结果

## 7.2 分子进化及其结果

大多数相关序列有许多变异了几次的位置

可接受突变速率对所有类型的碱基替换通常是不相同的

密码子不同位置有不同的突变速率

只应该用直系同源基因构建物种系统发生树

基因组大区域变化是常见的

## 7.3 系统发生树构建

核糖体小亚基rRNA序列非常适用于重建物种的进化

构树方法的选择在某种程度上依赖于数据集的大小和质量

在使用这些方法时必需选择一个进化模型

所有的系统发生分析必须以精确的多序列比对开始

16SRNA序列的一个小数据集的系统发生分析

为酶家族建立基因树有助于发现酶功能的进化

总结

名词解释

扩展阅读

## 第8章 构建系统发生树

## 8.1 进化模型和进化距离的计算

一个简单但不精确衡量进化距离的是种距离

Poisson校正距离考虑了同一位点上的多次突变

Gamma校正距离考虑了不同的序列位点上突变速率的差异

Jukes-Cantor模型再现了核苷酸序列进化的一些基本特征

更复杂的模型区分不同类型突变的相对频率

在DNA序列上存在核苷酸的偏好

蛋白质序列的进化模型和用于序列联配的替代矩阵密切相关

## 8.2 产生系统发生树

聚类方法基于进化距离产生一个系统发育树

UPGMA方法假定一个恒定的分子钟，并产生一个等距树

Fitch-Margoliash方法产生一个无根的加性树

邻接法：此方法涉及最小进化的概念

通常使用逐步增加和星形分解方法用以产生一棵起始树用于进一步的探索，这不是最终树

## 8.3 产生多种树的拓扑结构

分枝限界法大大提高了搜索树的拓扑结构的效率

可以通过对一个现存树做一系列细小的变化以优化树拓扑结构

寻找根给出了系统发生树在时间上的方向

## 8.4 评价树的拓扑结构

可使用基于进化距离的函数以评价树

加权简约法寻找具有突变最少的树

使用简约法可以采用不同的方式对突变作加权

可以使用最大似然法用以评估树

四重奏迷惑 (quartet-puzzling) 方法在标准执行中也包括了最大似然法

贝叶斯方法也可用于重建系统发生树



## <<理解生物信息学>>

### 8.5 评估树的特征和比较树的可靠性

即使是完善的数据和方法也会出现长枝吸引的问题

可以检验内部分枝测试树的拓扑结构

用于比较两棵或两棵以上的树的检验方法

总结

名词解释

扩展阅读

### 第4部分 基因组特征

### 第9章 揭示基因组特征

#### 9.1 基因组序列的初步分析

将整个基因组序列分割开来简化基因检测

结构 RNA基因和重复序列在进一步分析中可以排除

同源性可以用于原核和真核基因的鉴定

#### 9.2 原核基因组中的基因预测

#### 9.3 真核基因组中的基因预测

外显子和内含子的预测程序使用了多种方法

基因预测必须要保持正确的阅读框

有些程序只利用查询序列和外显子模型来预测外显子

有些程序只利用查询序列和基因模型来预测外显子

可以利用基因模型和序列相似性来预测基因

相关物种的基因组可以用来帮助基因预测

#### 9.4 剪接位点的预测

剪接位点可以由专门的程序独立地鉴定

#### 9.5 启动子区域的预测

原核启动子有较好定义的基序

真核启动子一般要比原核启动子复杂

有许多启动子的在线预测工具

启动子预测结果并不十分清晰

#### 9.6 证实预测结果

有多种计算基因预测准确率的方法

翻译预测的外显子可以证实预测的准确性

构建蛋白质和鉴定同源基因

#### 9.7 基因组注释

基因组注释是基因组分析中的最后一步

GO(geneontology)提供了一套基因注释的标准词汇表

#### 9.8 大基因组比较

总结

名词解释

扩展阅读

### 第10章 基因检测和基因组注释理论章节

#### 10.1 利用决策树检测功能 RNA分子

利用tRNAscan算法检测tRNA基因

检测真核生物基因组中的tRNA基因

#### 10.2 原核生物基因检测中有用的特征

#### 10.3 原核生物基因检测的算法

GeneMark利用了非均匀马尔可夫链(inhomogeneous Markov chains)和双密码子(dicodon)统计

## &lt;&lt;理解生物信息学&gt;&gt;

GLIMMER利用了编码概率的差值马尔科夫模型

ORPHEUS利用了同源性?

密码子统计和核糖体结合位点

GeneMark.hmm 利用精确状态持续隐马尔可夫模型

EcoParse是一个 HMM 基因模型

10.4 真核生物基因检测中用到的特征

真核生物基因与原核生物基因的差异

内含子?

外显子和剪切位点

转录因子的启动子序列和结合位点

10.5 预测真核生物基因信号

检测核心启动子结合信号是很多真核生物基因预测方法的关键元素

为了定位核心启动子序列信号而设计的一类模型

利用序列一般性质预测启动子区域可以去掉相当数量的假阳性结论

预测真核生物转录和翻译起始位点

转录和翻译终止信号给出基因完整定义

10.6 预测外显子和内含子

可以利用普遍序列性质 (generalsequence property)来识别

剪切位点预测

可以通过序列模式与碱基统计相结合预测剪切位点

GenScan将加权矩阵和决策树整合以定位剪切位点

GeneSplicer利用一阶马尔可夫链预测剪切位点

NetPlantGene整合内含子和外显子的神经网络模型以预测剪切位点

其他特征可能也可以用于剪切位点预测

利用特定方法识别起始和终止外显子

利用数据库中的同源区域可以定义外显子

10.7 完整真核生物基因模型

10.8 预测独立基因之余

功能注释

通过比较相关基因组,可以减少难以确定的预测

基因检测方法的评估和再评估

总结

名词解释 308 oxviiiio

扩展阅读

第5部分 二级结构

第11章 从序列中获得二级结构

11.1 预测方法的类型

基于规则的统计方法使残基形成一个特定二级结构成为可能

最近邻法是结合了有关蛋白质结构额外信息的统计方法

主要利用神经网络及隐马尔可夫方法进行二级结构预测的机器学习方法

11.2 训练和测试数据库

确定蛋白质二级结构的几种方法

11.3 预测程序准确性评估

Q3 衡量个别残基分配的精度

二级结构的预测不应该期望达到100%的残基精度

Sov值衡量全元素的预测精度

CAFASP/CASP: 无偏的和随时可用的蛋白质预测评估

## &lt;&lt;理解生物信息学&gt;&gt;

## 11.4 统计和基于知识的方法

GOR方法用作信息论方法

Zpred程序包括了同源序列和残基保守信息的多重联配

使用多个序列信息提高整体预测精度

最近邻法：使用多个非同源序列

PREDATOR是一种综合了统计和基于知识的程序，其中包括了最近邻法

## 11.5 二级结构预测的神经网络方法

评估神经网络预测的可靠性

基于网络的神经网络二级结构预测程序的几个例子

PROF：蛋白质预测

PSIPRED

Jnet：使用序列比对的几种可选描述

## 11.6 一些需要特殊预测方法的二级结构

跨膜蛋白

量化膜环境的属性

## 11.7 跨膜蛋白结构的预测

多螺旋膜蛋白

选择预测跨膜螺旋的预测程序

统计方法

基于知识的预测

蛋白质家族的进化信息改善了预测结果

神经网络在跨膜预测中的应用

使用隐马尔可夫模型预测跨膜螺旋

比较结果：选择哪个

如果提交一个非跨膜蛋白给跨膜预测程序会发生什么

含 链的跨膜结构的预测

## 11.8 卷曲螺旋结构

COILS预测程序

PAIRCOIL和 MULTICOIL是COILS算法的扩展

拉上亮氨酸拉链：一个特殊的卷曲螺旋

## 11.9 RNA二级结构预测

总结

名词解释

扩展阅读

## 第12章 二级结构预测

## 12.1 定义二级结构和预测精度

蛋白质二级结构指定定义不同给出结果也不同

对二级结构的预测精度存在着几种不同的测度

## 12.2 二级结构预测基于残基的偏好性

每个结构状态存在着氨基酸的倾向，这可以在指定时作为残基偏好性

最简单的预测方法是基于在一个序列窗口中的平均残基偏好性

残基偏好性由附近的序列所调控

通过从同源序列得到的信息可以大为改善预测

## 12.3 近邻方法是基于序列片段的相似性

发现相似序列的短片段具有相似的结构

使用了几种序列相似性的测度用以寻找近邻片段

使用近邻片段结构的加权平均用以预测

## &lt;&lt;理解生物信息学&gt;&gt;

已发展了近邻方法用于预测那些较易发生错误折叠的区域

12.4 神经网络已经被成功应用于二级结构预测

分层前馈神经网络可以将序列转变为结构预测 378 oxio

包括同源序列信息将改善神经网络的预测正确度

更复杂的神经网络已应用于预测二级结构和其他一些结构特点

12.5 隐马尔可夫模型已应用在结构预测中

发现 HMM 方法对膜蛋白特别有效

使用 HMM, 也可以成功地预测非膜蛋白的二级结构

12.6 可以预测结构特征的一般数据分类技术

支持向量机已成功用于蛋白质结构预测

Discriminates ?

SOM 和其他一些方法

总结

名词解释

扩展阅读

第6部分 蛋白质三级结构

第13章 蛋白质结构预测

13.1 势能函数和力场

蛋白质的构象可以在势能面上观察到

构象能量可以用简单的数学函数来描述

相似的力场可以用来表示平均环境中的构象能量

势能函数可以用来评估构建的结构

能量最小化可以用来搜索建模结构和确定局部能量最小值

分子动力学和模拟退火可以用来搜索全局能量最小值

13.2 用折叠识别法预测蛋白质结构

在没有同源蛋白的情况下预测蛋白质结构折叠

非冗余蛋白质折叠数据库在穿线法中的应用

穿线法中采用的两种不同的打分机制

动态规划方法搜索目标序列与已知折叠匹配的最佳方案

评估折叠识别可信度的方法

穿线法实例: 网柱黏菌中的C2结构域

13.3 同源建模原理

目标序列与模板序列相关性越大, 同源建模的结果越好

关键序列一致性取决于整个序列的长度

针对目前可建模的大批量序列的同源建模已经实现自动化

建模所基于的一系列假设

13.4 同源建模的步骤

在PDB数据库中寻找目标蛋白质的同源结构

目标和模板蛋白序列的精确比对对于成功建模是必不可少的

蛋白质的结构保守区域最先建模

进入下一阶段前需检验建模的核心结构是否存在不适之处

序列重新比对和重新建模可能会提高建模结构的准确性

插入和缺失序列通常建模成环区域

不同氨基酸侧链的建模主要通过旋转异构体数据库来实现

采用能量最小化来消除结构错误

分子动力学可以用来搜索可移动的loop区域可能采取的构象

检查模型的准确性

## &lt;&lt;理解生物信息学&gt;&gt;

同源建模的可信度

13.5 自动化同源建模

MODELLER通过适当的蛋白质结构约束条件来建模

COMPOSER使用基于片段的建模方法来自动化生成相应的模型

网络中可用于比较建模的自动化方法

结构预测结果的评价

13.6 PI3蛋白激酶p110 的同源建模

SwissPdbViewer能够用于手工或者半手工建模

同时做序列比对？

核心结构建模和侧链建模

柔性区域(loop)通过数据库中可能的结构建模

SwissPdbViewer软件可以实现能量最小化和质量评估

MolIDE是一个可下载的半自动的建模软件包

基于网络的自动化建模(以p110 激酶为例)

构建一个功能上相似但是序列不相似的蛋白oxo质：mTOR

从序列生成一个多结构域三维结构

总结

名词解释

扩展阅读

第14章 结构 功能关系分析

14.1 功能保守性

发挥功能的区域通常结构上是保守的

相似的生物学功能存在于具有不同折叠模式的蛋白质上

折叠数据库确定了结构上相似的蛋白质而无论其功能

14.2 结构比较方法

找到蛋白质的结构域可以帮助结构比较

结构比较能够揭示序列比较不能辨别的保守功能

CE方法通过匹配蛋白质片段把两个蛋白质叠合到一起

向量叠合搜索工具(vectoralignmentsearch tool, VAST)能够叠合二级结构

DALI确定蛋白质结构的叠合方式,但是并不保持片段之间的相对顺序

FATCAT在刚性的片段之间引入了旋转

14.3 找到结合位点

高度保守的？

带电荷的或者疏水的表面是相互作用位点的标志

通过表面性质寻找蛋白质 蛋白质的相互作用位点

通过计算蛋白质的表面,可以找到那些可能是结合位点的裂缝和洞

通过分析氨基酸的保守性可以确定结合位点

14.4 分子对接方法和程序

当同源蛋白和类似的小分子复合物的结构已知的时候,可以作简单的分子对接

一些专用的分子对接程序可以自动地把配体对接到蛋白质结构上去

通过打分函数来确定最可能的对接结果

DOCK软件采用半刚性的方法,通过分析配体和结合位点形状和化学性质的互补来做对接

片段对接方法可以通过预测结合位点处的原子类型和功能基团确定可能的底物

GOLD是一个柔性的对接程序,它使用遗传算法

结合位点的水分子也应该考虑

总结

名词解释

## &lt;&lt;理解生物信息学&gt;&gt;

## 扩展阅读

## 第7部分 细胞和组织

## 第15章 蛋白质谱和基因表达分析

## 15.1 大规模基因表达分析

大量不同基因的表达可同时被 DNA 芯片检测

基因表达芯片主要用于检测基因在不同条件下的表达差异

基因表达系列分析也被用于研究基因表达的总体模式

数字差异显示：应用生物信息学和统计学来检测不同组织中基因的差异表达

推动不同地方和不同实验的数据整合

分析基因表达微阵列数据的最简单方法是层次聚类分析

基于自组织映射网络的技术可被用于分析微阵列数据

自组织树算法 (SOTA) 自上而下地对类别进行连续分割

基因表达数据的聚类结果是进一步研究的工具

## 15.2 大规模蛋白质表达分析

二维凝胶电泳是分离细胞内各种蛋白质的一种方法

检测二维凝胶中显示的表达水平

二维凝胶能发现不同样本间的蛋白质表达差异

用聚类方法识别具有相似表达模式的蛋白质位点

主成分分析 (PCA) 是分析微阵列和二维凝胶数据除聚类分析之外的又一选择

跟踪一组蛋白质位点在一系列样本间的差异

数据库和在线工具可用来辅助二维凝胶数据的解释

蛋白质微阵列芯片能同时检测大量不同蛋白质的存在或活性

可用质谱来鉴定已经由二维凝胶或其他技术分离和纯化的蛋白质

对质谱进行蛋白质鉴定的程序可从网上免费获得

质谱能用于检测蛋白质浓度

## 总结

## 名词解释

## 扩展阅读

## 第16章 聚类方法和统计学概念

## 16.1 分析表达数据之前的准备工作

数据标准化用于去除实验中的系统误差

表达水平通常用比值表示并取对数转换后再分析

有时在数据转换后再进行标准化不无裨益

主成分分析用于合并被分析对象的某些属性

## 16.2 聚类分析的先决条件是定义所有数据点之间的距离

欧氏距离在日常生活中广泛使用

Pearson 相关系数表征的距离能衡量表达响应的形状相似性

Mahalanobis 距离综合考虑表达响应之间的变异性 and 相关性

## 16.3 聚类方法能鉴定出内部相似且彼此间不同的表达模式

层次聚类对数据生成一组彼此关联的备选划分方案

均值聚类将数据分成预先指定数目的类群，但不能确定类群间彼此的远近关系

自组织图 (SOM) 采用神经网络算法将数据聚类成预先指定数目的类群

进化聚类算法用选择？

重组和突变等概念来搜索问题的可能最优解

自组织树算法 (SOTA) 确定所需要的聚类数目

双向聚类可鉴定出在部分样本中呈现相似表达模式的一组基因

聚类类群的合理性可由其他方法独立验证

## &lt;&lt;理解生物信息学&gt;&gt;

## 16.4 统计分析可量化观测到的差异表达的显著性水平

标检验能用于估计两个表达水平之间差异的显著性

非参数检验用于规避对数据采样方式做假定

对差异表达的多重假设检验需要采取特殊的技术来控制错误率

## 16.5 基因和蛋白质表达数据能用于样本分类

有许多可选手段能用于样本分类

支持向量机是另一种能生成分类器的有监督学习算法

总结

名词解释

扩展阅读

## 第17章 系统生物学

## 17.1 什么是系统

系统大于部分之和

生物学系统是有生命的网络

数据库是网络构建的有效起点

构建模型需要比网络更加丰富的信息

构建模型的三种可行的方法

动力学模型并非系统生物学研究的唯一途径

## 17.2 模型的结构

控制环路是生物学系统的必要组成部分

网络中的相互作用可以被表述为简单的微分方程

## 17.3 生物学系统的鲁棒性

鲁棒性是生物体复杂性的一个独特属性

模块性在鲁棒性中扮演重要角色

系统中的冗余性能够提供鲁棒性

生命系统可以通过双稳态开关实现从一个状态到另一个状态的转换

## 17.4 存储和运行系统模型

特定的程序使得系统模拟更加便捷

标准化的系统描述有助于存储和再利用

总结

名词解释

扩展阅读

## 附录A

概率论?

熵和信息

互斥事件

发生两个事件

两个随机变量的发生

贝叶斯分析 554 贝叶斯定理

参数值的推导

扩展阅读

## 附录B分子能量函数

用力场计算分子内部和分子间相互作用的能量

成键项

非成键项

势能在穿线法中的使用

平均力的势能

## <<理解生物信息学>>

与溶剂效应相关的势能项

扩展阅读

附录C 功能优化

全搜索 (full search)方法

动态规划和分支界限法 (branch and bound)

局部最优 (localoptimization)

下降单纯形 (downhillsimplex)法

最速下降 (steepestdescent)法

共轭梯度 (conjugategradient)法

使用二阶导数的方法

热力学模拟和全局优化

蒙特卡罗和遗传算法

分子动力学

模拟退火

总结

扩展阅读

字符表

索引

彩图



## &lt;&lt;理解生物信息学&gt;&gt;

## 编辑推荐

《理解生物信息学》学习效果：每章开篇都有一个学习效果列表，它总结了该章所涉及的主题，可作为一个反馈清单。

思维导图：每一章都含有一个思维导图，这是《理解生物信息学》一个特别的教学特征，它确保每个学生都能看到并记住一些特定应用中所必需的步骤。

偶尔地，思维导图的两个独立方面也可能有着重要的关联。

流程图：每一章的每个小节都有一个流程图以帮助读者记忆该小节所涵盖的主题。作为示例，下面给出了第5章的一个流程图，其中在本节将要解释的概念用深灰色框标注，且相互间用箭头连接起来。

例如，两种主要类型的最优联配：局部和全局将在本章的这一节描述。

那些已在之前小节描述过的概念用浅灰色框标注，这样我们就很容易了解本节涉及的主题和已介绍过的主题间的联系。

例如，构建联配需要为空缺打分的方法和为替换打分的方法，两者都已经在这一章描述过了。

通过这种方式，整章涉及的主要概念以及相互间的关系就能渐渐地被构架出来。

插图：每一章都配有插图。

插图的配置是经过充分考虑的，以保证既简单易懂又与本书其他章节保持连贯一致。

扩展阅读：在这么一个快速发展的学科中，我们不可能在这本有限的《理解生物信息学》囊括现有的所有知识，更不用说将来的发展了。

因此在每章的结尾我们都列了一些研究文献和专业著作的参考文献以帮助读者进一步扩展知识、发展技能。

我们根据不同主题收集文章，使得扩展阅读中每节都与这一章相应小节的内容相对应。

我们希望这能帮助读者以最快的速度找到他们感兴趣的扩展材料。

字符表：生物信息学需要使用很多符号，对还不了解生物信息的人来说，许多符号都是不熟悉的。

为了帮助读者了解本书适用的符号，我们在《理解生物信息学》后面给出了引用的每个符号、它的定义以及它在本书最常出现的位置的列表。

名词解释：在文中，所有技术术语在第一次出现时都用黑体显示，且在名词解释中列出其相应的解释。

此外，每个在名词解释中的术语都会出现在索引中，这样读者就能很快获得详细介绍这一术语的相应页码。

《理解生物信息学》设计成可以进行交叉参考，以尽可能帮助读者阅读。

图版：《理解生物信息学》所有的英文原图都可以在GarlandScience网站上下载。

插图文件以.zip格式保存，其中每个.zip文件对应一章。

每张图都可以从相应的.zip文件中以.jpg的格式解压出来。

更多材料：GarlandScience的网站还包括一些与《理解生物信息学》主题相关的额外的材料。7个部分中任何一部分都对应一个.pdf文件，它通过一系列与这些章节内容相关的有用的网址链接，能链接到一些有用的数据库、文件格式定义、免费的程序以及允许数据在线分析的服务器上。

此外，在阐述分析方法时所用到的数据也会被提供。

这就允许读者对同一数据重新进行分析，重现《理解生物信息学》所显示的结果，并尝试其他技术。

<<理解生物信息学>>

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:<http://www.tushu007.com>