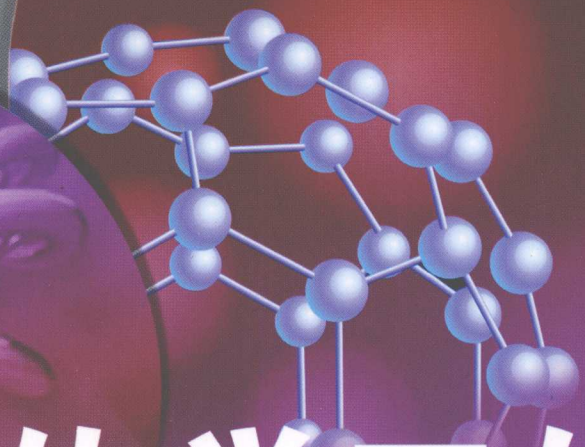
生物信息学数据分析丛书



遗传学工作者的生物信息学

（原书第二版）

BIOINFORMATICS FOR GENETICISTS

A bioinformatics primer for the analysis of genetic data

（英）迈克尔 R.巴恩斯 著

丁 卫 李慎涛 廖晓萍 主译

科学出版社

www.sciencep.com

遗传学工作者的生物信息学

（原书第二版）

[redacted]

生物信息学数据分析丛书

Bioinformatics for Geneticists

A bioinformatics primer for the analysis of genetic data

遗传学工作者的生物信息学

（原书第二版）

科学出版社

北京

图字：01-2007-5840 号

内 容 简 介

本书由五部分共 19 章组成，第一部分主要介绍了遗传学工作者所面临的生物信息学挑战以及遗传数据的操作和管理；第二部分主要介绍了以人类单体型图谱计划（HapMap）、人类基因组学和比较基因组学等为代表的多元化数据；第三部分主要介绍了用于遗传学研究设计和分析的生物信息学策略和手段；第四部分重点介绍了基因分析与疾病的关联及其代表案例；第五部分介绍了利用数据库界面进行全面生物信息学系统分析的流程，其中覆盖了微阵列等一些非常实用的技术，并且就药物遗传学等前沿领域展开了前瞻性的论述。

本书适合统计学和群体遗传学专业，以及具有分子遗传学和医学遗传学背景的高等院校师生和科研人员阅读，对从事人类及模式生物研究的广大实验室研究人员、临床研究人员以及实验室负责人都有较大的参考意义。

图书在版编目(CIP)数据

遗传学工作者的生物信息学（原书第二版）/（英）巴恩斯（Barnes, M. R.）著；丁卫，李慎涛，廖晓萍主译. —北京：科学出版社，2009
（生物信息学数据分析丛书）
书名原文：Bioinformatics for Geneticists: A bioinformatics primer for the analysis of genetic data
ISBN 978-7-03-025490-0

Ⅰ. 遗… Ⅱ. ①巴…②丁…③李…④廖… Ⅲ. 遗传学-生物信息论 Ⅳ. Q3

中国版本图书馆 CIP 数据核字（2009）第 157371 号

责任编辑：李 悦 刘 晶/责任校对：钟 洋
责任印制：钱玉芬/封面设计：耕者设计工作室

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

新蕾印刷厂 印刷
科学出版社发行 各地新华书店经销

*

2009 年 10 月第 一 版 开本：787×1092 1/16
2009 年 10 月第一次印刷 印张：29
印数：1—3 000 字数：652 000

定价：88.00 元
（如有印装质量问题，我社负责调换）

本书译校者名单

主　　译

丁　卫　　首都医科大学
李慎涛　　首都医科大学
廖晓萍　　华南农业大学

其他译校者（按姓氏拼音排序）

郭　玲　　首都医科大学
贾书勤　　首都医科大学
李　京　　首都师范大学
马静云　　华南农业大学
王　炜　　首都医科大学
王一松　　首都医科大学
温铭杰　　首都医科大学
薛冠华　　首都医科大学附属儿童医院
闫韶飞　　首都医科大学
张凤兰　　首都医科大学
张力青　　首都医科大学
郑少鹏　　首都医科大学
朱俊萍　　首都医科大学

中 译 本 序

诞生于20世纪80年代的生物信息学是伴随基因组研究而产生的一门新学科，它综合运用数理科学、计算机科学和生物学工具，对基因组相关信息进行获取、加工、储存、分配、分析和解释。具体地说，生物信息学是把基因组DNA序列信息分析作为源头，找到基因组序列中代表蛋白质和RNA基因的编码区；同时，阐明非编码区的信息实质，破译隐藏在DNA序列中的遗传语文规律。

进入21世纪以来，基因组研究得到了深入与广泛的发展，生物信息学也成为归纳、整理与基因组遗传语文信息释放及其调控相关的转录组、蛋白组、代谢组等组学信息，应用系统生物学的方法认识生物体代谢、发育、分化、进化以及疾患发生规律的不可或缺的工具。

由丁卫等医学工作者翻译的迈克尔R. 巴恩斯著的《遗传学工作者的生物信息学》是一本专门为遗传学工作者了解生物信息学的概念、掌握生物信息学的方法并应用于实际问题而写的书。其内容翔实且系统，既有基本概念、算法、软件与数据库，也有对新进展的介绍，为遗传学工作者学习与应用生物信息学提供了很好的机会。

生物信息学是一门很有用的学问，所有与基因组相关的研究都离不开它。生物信息学也是一门发展十分迅速的学科，不紧密跟踪就不能及时、全面地掌握它。希望不断有更多、更新的生物信息学的著作出现，为广大科技工作者、教育工作者提供丰富的资源。

2009年8月

译 者 序

自 20 世纪 80 年代后期以来，随着计算机、互联网和数据库技术的发展，生物信息学（bioinformatics）作为一门新兴的交叉学科取得了令人振奋的成就。人类基因组计划的完成以及各种高通量分析手段的不断完善，标志着一个信息时代的来临。日益增长的海量数据，常常使得人们对相应的分析和理解陷入困惑，在实际应用中，又陷入大海捞针般的困境。因此，人们对生物信息学的期待和依赖日益显著。

在生物信息学的学习和应用过程中，由于学科本身的动态复杂性和应用的广泛性、多元性，人们常常会感到无从下手。已出版的生物信息学专著，在数量和种类上有限，而较为全面和经典之作更为稀少，远远不能满足人们的需要。本书深入浅出、实用性强，实例丰富且图文并茂，是一本难得的生物信息学佳作。我们相信本书的出版会对国内生物信息学的普及和提高起到巨大作用。

本书名为《遗传学工作者的生物信息学》，是一本特别适合遗传学工作者但又绝非仅限于此的专著。著名的肿瘤遗传学家 Volgelstein 曾经指出，现代医学的研究应当以遗传学为核心和基础。事实上，遗传学已经渗透到生命科学的各个研究领域，并且在疾病诊断、治疗和预防等应用中也发挥着越来越重要的作用。生物信息学分析作为一种强有力的技术手段，在从实验设计到结果分析等各个层面发挥着不可替代的作用，既能够启迪研究人员设计阶段的预判以少走弯路，也可以从结果分析中挖掘大量的有用信息，起到事半功倍的作用。可以说，没有生物信息学就没有人类基因组等大型的国际合作项目。

本书由五部分共 19 章组成。第一部分主要介绍了遗传学工作者所面临的生物信息学挑战以及遗传学数据的操作和管理；第二部分主要介绍了以人类单体型图谱计划（HapMap）、人类基因组学和比较基因组学等为代表的多元化数据；第三部分主要介绍了用于遗传学研究设计和分析的生物信息学策略和手段；第四部分重点介绍了基因分析与疾病的关联及其代表案例；第五部分介绍了利用数据库界面进行全面生物信息学系统分析的流程，其中覆盖了微阵列等一些非常实用的技术，并且就药物遗传学等前沿领域展开了前瞻性的论述。虽然我国目前在生物信息学领域的研究与国际先进水平仍有差距，但由于该新兴学科的历史较短，毕竟相差并不遥远，希望本书热情的读者能够和译者一道接受挑战、奋起直追，为我国生物信息学的发展和繁荣尽自己微薄之力。

本书的多位译者分别完成了各个章节的初始翻译工作，为此付出了艰苦努力，李慎涛、丁卫负责了全书的审校工作。此外，还要特别感谢科学出版社的编辑为出版本书付出的辛勤劳动。

本书的译者均为从事生物医学科研和教学的一线人员，对生物信息学有着较为深刻的理解。然而鉴于生物信息学的前沿性和综合性，加之其研究动态的更新速度较快、译

校人员的水平有限，书中难免存在一些错误或失当之处，真诚希望广大读者给予批评指正。

丁 卫
首都医科大学生物化学与分子生物学系教授
2009 年 1 月于北京

原 书 序

尽管出现时间相对较短，生物信息学似乎始终被认为是一个不同寻常的、多学科综合性领域。15 年前，序列数据依然贫乏，当时计算机的处理能力仅相当于如今“比萨饼盒”样超级计算机的极小部分，然而生物信息学已经深入到许多方面的议题中，数据库的发展、序列的排列、蛋白质结构的预测、编码区和启动子区位点的鉴别、RNA 的折叠和进化树的构建等，都包含在早期生物信息学工作者的职业范围内。为了解决这些问题，生物信息学从统计学、数学、物理学、计算机科学，当然还有分子生物学中分离出来。今天依然可以预期，生物信息学仍将反映出其创立时广泛的学科基础，并且选择性地集合了各专业的科学工作者。

鉴于其本质的多样性，很难对生物信息学作为一门学科的范畴进行划分，乃至试图硬性划定该领域的界限也是徒劳无益的。即便是当今，如果有人刻意在生物信息学范围内编排一系列涉及广泛的研究领域，那么很可能将与其共享基础的生物学科——遗传学排除在外，这是颇具讽刺意味的。一方面，这似乎难以置信，因为这些领域与统计方法学拥有高度相同的背景，依赖于高效的计算机算法、快速增长的生物学数据和与分子生物学共同的原理；而另一方面，这也完全可以理解，生物信息学的很大一部分在过去的数年中致力于包括人类在内的一系列基因组测序工作。在很多情况下，这些测序项目的重点在于构建一个单一代表性序列，即保守序列，这是一个与核心的遗传学原理及个体变异差异完全不同的概念。尽管彼此之间的关注日益增加，在拥有一些明显区别的同时，遗传学和生物信息学仍然保持着各自的特征。

遗传学工作者离不开生物信息学，尤其对于那些需要发现和了解影响复杂表型的遗传学家更是如此。在人类遗传学领域，这种需求已变得十分明确，以至于多数大型实验室都拥有一至两位生物信息学专家以帮助其他实验室成员完成与计算相关的事务。这些专家需要为令人生畏的各种各样的应用程序提供技术支持，来源于服务对象的典型咨询包括：如何找到进入互联网的说明；如何解析复杂的数据库架构；如何在并行的计算机组中优化高强度的数值算法。这类人才，虽然在某种程度上仍很缺乏，但对于一个实验室的成功非常必要。

随着不断涌现的序列数据、表达信息和对结构的详细描述，以及即将到来的大规模、多人种的基因型和单体型数据，遗传学实验室工作转型的必要性已超越对生物信息学人才的单纯依赖。对实验室每一个人来说，具有一定程度的理解和应用生物信息学工具软件的能力是十分必要的。所幸的是，生物信息学家现在已经非常成功地开发出界面友好的软件提供给实验型的科学家以完成复杂的统计分析，而正是他们负责数据的采集并且理所当然最了解用于分析的信息。为进行更深入的分析，生物信息学家已经创建了许多精妙的软件用于结果的显示，同时整合了一系列有用的注释信息，如染色体特征、序列标名、疾病的相关性和物种的比对等。随着这些工具的免费提供与持续进行的开

发，有效运用遗传学和基因信息的基因图谱计划的项目必然会比不充分具备这些条件的项目取得更大的成功。简而言之，不能运用生物信息学的遗传学工作者们将越来越被擅长此术的其他遗传学工作者捷足先登。

在遗传学领域更为广泛地认识生物信息学的必要性是该书关注的重点，通过对标题的快速浏览便可以明确感受到。同样明确的是，遗传学工作者是该书所面向的主要读者群。也许有人仍然会问："这本书的读者对象是遗传学工作者中的哪一些特殊群体呢?"鉴于生物信息学的软件和核心计算方法与统计学和群体遗传学领域享有共同的关注点，因此统计学专家似乎应被看做是可能的读者。但是，就该书的初衷而言，比起统计学家，更侧重于那些具有广阔分子遗传和医学遗传背景的科研人员，包括人类和模式生物的研究等，其内容的获取应该对实验室研究人员、临床研究人员乃至实验室负责人敞开。从计算层面上讲，人们仅需要基本的计算机技能便可以通晓大部分的知识；从生物学层面上讲，对所述问题的理解要求人们对遗传学研究基本熟悉，并且具有对遗传学与基因组学信息内在价值的认知。

该书涵盖的生物信息学内容实质上反映出该领域的多样性。为了使这些广泛的内容获得一定的有序性，著者着重于计算机软件和现有数据库的有效使用。这种侧重意味着对统计学原理、数学算法以及数据库组建的描述留给了其他书籍加以介绍，有意绕过这些内容旨在强化对广为使用的软件的介绍，而非生物信息学方法和工具的开发。

这里描述的许多生物信息学工具，其背后数据的变化和扩展非常快，相应地，软件工具和数据库本身趋于不断的动态改变（有些令人难以承受）。这种流动的结果使得学会应用现有的程序不一定能保证在将来使用这些软件时获得技巧，因而人们不能期待该书（或在同期其他生物信息学教科书中）所涉及的工具和数据具有长久的延续性。尽管如此通过学习该书，更有效地运用现有的工具，遗传学工作者还是能从现有的技术中获益，并使更多的生物信息学家加入到精彩的遗传学研究之中。将生物信息学展现给遗传学工作者是非常关键的第一步，该书将对同类领域进行整合，并且对影响复合表型的、令人无奈且难以捉摸的基因加以描述。

Lon R. Cardon
牛津大学人类遗传学维康信托中心
生物信息学教授

前　言

如果你从事人类遗传学工作的时间足够长，那么我说“从基因座到基因座”而不是“从基因到基因”，你便会认识到或许你可能永远得不到一个基因，但是你将不会为此而使工作停滞不前。

Peter A. Holmans

对遗传学实验结果的合理解释在任何层面上来说都是一种挑战，而通过撰写一部关于生物信息学应用的书籍以实现这一目标在某种程度上似乎是一种颇为自负的举措。关于生物信息学究竟由何组成在个人观点上见仁见智。就本书而言，生物信息学主要阐述生物学功能，而我们首要关注的疾病遗传学是认识生物学功能异常的基础。基于这种考虑，不妨将生物信息学当作是促进对遗传学认知的工具。

自本书第一版以来，以这种方式考虑问题的理由变得更加引人注目。人类疾病遗传学正快速成为高通量的行为，这意味着当前对遗传学实验的解释实质上是对数以百万计数据点的理解，这也显示出对以遗传学为重点的生物信息学的需要。简而言之，我们需要处理和分析此等规模数据的信息学，同时我们也需要解释人类整体生命系统的生物学。

没有各章节作者的杰出贡献，本书是难以完成的，我真切地感受到正是他们的帮助才使本书在生物学和信息学方面名副其实。我要把最诚挚的感谢给予 Ian C. Gray，他在第一版中与我合作，并在这一版中提供了极为有益的意见和支持。如果没有 John & Wiley 出版公司难得的富于奉献精神的专业小组，大家在此所见的令人振奋的科学内容将不复存在，正是他们始终保持着各项工作按部就班，他们是 Joan Marsh、Andrea Baier、Fiona Woods、Kate Pamphilon 和 Emilie McDonough 等。除了撰写本书，我拥有一份全职工作，我非常感谢 GSK 的 Philippe Sanseau 和 David Searls，他们给予我的时间以及激励和支持使我最终完成了这项工作。最后，我还要感谢我妻子 Aruna，感谢她一直以来的关爱、支持、鼓励和精益求精的付出。没有她，我将不会有毅力有条不紊地来完成这本巨著。

Michael R. Barnes

2006 年 8 月于英国 Harlow 市

原书作者

Catherine A. Ball Department of Biochemistry, Stanford University Medical School, Stanford, CA, USA

Aruna Bansal Discovery and Pipeline Genetics, GlaxoSmithKline Pharmaceuticals, Third Avenue, Harlow, Essex, UK

Michael R. Barnes Bioinformatics, GlaxoSmithKline Pharmaceuticals, Third Avenue, Harlow, Essex, UK

Bryan J. Barratt Research and Development Genetics, AstraZeneca, Alderley Park, Macclesfield, Cheshire, UK

Matthew J. Betts Structural and Computational Biology Programme, EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Diana Blaydon Centre for Cutaneous Research, Institute of Cell and Molecular Science, Queen Mary's School of Medicine and Dentistry, Whitechapel, London, UK

Karl W. Broman Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA

Ellen M. Brown Discovery Informatics, AstraZeneca, Alderley Park, Macclesfield, Cheshire, UK

James R. Brown Bioinformatics, GlaxoSmithKline Pharmaceuticals, Upper Providence, PA, USA

Elissa J. Chesler Oak Ridge National Laboratory, Biosciences Division, Oak Ridge, TN, USA

Richard R. Copley Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

Barry Dancis Bioinformatics, GlaxoSmithKline Pharmaceuticals Upper Providence, PA, USA

Steve Deharo Bioinformatics, GlaxoSmithKline Pharmaceuticals, Third Avenue, Harlow, Essex, UK

Paul S. Derwent Bioinformatics, GlaxoSmithKline Pharmaceuticals, Third Avenue, Harlow, Essex, UK

Ian C. Gray Paradigm Therapeutics (S) Pte Ltd, 10 Biopolis Way, Singapore 138670

Joel Greshock Translational Medicine, Clinical Pharmacology Division, GlaxoSmithKline Pharmaceuticals, Upper Merion, PA, USA

Simon C. Heath Centre National de Genotypage, Evry Cedex, France

David P. Kelsell Centre for Cutaneous Research, Institute of Cell and Molecular Science, Queen Mary's School of Medicine and Dentistry, Whitechapel, London, UK

Ralph McGinnis Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Charles A. Mein Genome Centre, Queen Mary's School of Medicine and Dentistry, Charterhouse Square, London, UK

Mary Plumpton Bioinformatics, GlaxoSmithKline Pharmaceuticals, Stevenage, Hertfordshire, UK

Robert B. Russell Structural and Computational Biology Programme, EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Philippe Sanseau Bioinformatics, GlaxoSmithKline Pharmaceuticals, Stevenage, Hertfordshire, UK

Colin A. M. Semple Bioinformatics, MRC Human Genetics Unit, Edinburgh EH4 2XU, UK

Gavin Sherlock Department of Genetics, Stanford University Medical School, Stanford, CA, USA

Christopher Southan Global Compound Sciences, AstraZeneca R&D, Mölndal, Sweden

Martin S. Taylor Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

Magnus Ulvsbäck Molecular Pharmacology, AstraZeneca R&D, Mölndal, Sweden

Charlotte Vignal Discovery and Pipeline Genetics, GlaxoSmithKline Pharmaceuticals, Third Avenue, Harlow, Essex, UK

Chaolin Zhang Department of Biomedical Engineering, State University of New York at Stony Brook, NY, USA

Michael Q. Zhang Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

Xiaoyue Zhao Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

生物信息学词汇表

BLAST（**B**asic **L**ocal **A**lignment **S**earch **T**ool） 是基本的局部比对搜索工具，也是鉴定数据库中的序列与某个已知查询序列的匹配的工具，用统计学分析来判断每种匹配的意义。匹配的序列可能会与查询序列同源或相关。有以下几种版本的 BLAST。

（1）BLASTP 将一个氨基酸查询序列与一个蛋白质序列数据库进行比较。

（2）BLASTN 将一个核苷酸查询序列与一个核苷酸序列数据库进行比较。

（3）BLASTX 将一个以所有可读框翻译成的核苷酸查询序列与一个蛋白质序列数据库进行比较。

（4）BLASTTN 将一个蛋白质查询序列与核苷酸序列数据库（动态地以所有可读框翻译成蛋白质序列）进行比较。

（5）BLASTTX 将一个核苷酸查询序列的 6 种可读框翻译的蛋白质序列与核苷酸数据库的 6 种可读框翻译的蛋白质序列进行比较。

BLAT（**B**last-**L**ike **A**lignment **T**ool） BLAT 表面上似乎与 BLAST 一样，也是一种检测与一个已知查询序列匹配的子序列的工具，然而，BLAT 和 BLAST 有许多差异，BLAT 是由 UCSC 开发的，它通过在内存中保留一个全基因组的索引来搜索人类基因组，索引由所有不重叠的 11mer（除了重复序列以外）组成。对人类基因组的一个 BLAT 搜索会迅速发现相似性在 95%以上、长度至少为 40 个碱基的序列，它会丢失更趋异或更短的序列比对（有关此工具的详细介绍，见 UCSC FAQ:http://genome.ucsc.edu/FAQ.html）。

CDS（**C**oding **S**equence） 编码序列。

Contig map 重叠群图谱。描述代表一个完整基因组片段或染色体片段的重叠的（连续的）克隆的一种图谱。

DAS（**D**istributated **A**nnotation **S**ystem） 分布式注释系统。一种浏览和共享互联网上基因组序列注释的协议，使用户能够搜索和比较几种来源的注释。Ensembl 提供 DAS 参考服务器，为大量专门的人类基因组注释提供入口（有关细节见 http://www.ensembl.org/das/）。

Data mining 数据挖掘。查询数据库以便满足某种假说（“top-down”数据挖掘），或查询一个数据库以便基于严格的统计学相关性建立新的假说（“bottom-up”数据挖掘）。

Domain（protein） 结构域（蛋白质）。在单一蛋白质序列内的具有特定生物学意义的一个区域。然而，结构域也可以被定义为蛋白质三维结构内的一个区域，它包括几个不同蛋白质序列的区域，能够完成某一特定的功能。一种结构域分类是一组具有相同确定的特性或特征的结构域。

Electronic PCR（ePCR） 电子 PCR。一种类似于实验室 PCR 的计算机模拟过程，使用两条引物来定位一个序列的特性（如单核苷酸多态性）。为了验证位置，两条引物必

须在跨度为一定距离的相同邻近区域内进行定位，有效地产生一种电子 PCR 产物。

Expressed sequence tag (EST) 一种从表达的基因（源自一个 cDNA 文库）得来的短序列。可使用存储大量 EST 的数据库来测定不同转录物在 cDNA 文库和制备 cDNA 文库的组织中的相对丰度。EST 也可用作相应 cDNA 和基因鉴定、克隆和全长测序的一种物理标签。

FASTA format FASTA 格式。FASTA（Fast-All）最初被设计用于 Lipman 和 Pearson 的序列算法，是一种最简单和最被认可的序列格式，采用的形式是：简单的标题，在前面加一个大于号（>），序列位于下一行，如>sequence _ id gataggctgagcgatgcgatgctagctagctagc。

Golden path 用于人类基因组第一个和后续的拼接结果的术语。

Hidden Markov model (HMM) 隐马尔可夫模型。一种联合的统计学模型，用于有序的变量顺序。在马尔可夫链中随机干扰变量的结果（这样原始变量被“隐藏”），由此，马尔可夫链具有在每一步都选择 HMM“状态”的不连续的变量，受干扰的值可以是连续的，是 HMM 的“结果”。HHH 相当于一种结合的混合模型，这种状态的联合分布就是马尔可夫链。在生物信息学中，HHH 很有价值，因为 HHH 能够用未比对的或无权重的输入序列训练搜索或比对算法，并允许使用位置依赖性的评分参数（如空位罚分），这样能够更准确地对进化事件对序列家庭的作用进行建模。

Homology（strict） 同源性（严格意义上的）。享有共同进化祖先（通用）的两种或多种生物种系、系统或分子，或者两种或多种基因和蛋白质序列享有显著性程度的相似性，通常根据表现在其长度内的特征（以 DNA 为例）或保守性替换（针对蛋白质而言）的量值来测定。序列的同源性检索往往通过查询一个 DNA 或蛋白质序列以发现与之高度近似的基因或基因产物，从而明辨其待查基因的祖先、继承状况以及可能的功能。

in silico “电算化”（生物学概念，文字含义是由计算机所中介的）。利用计算机去模拟、处理或分析某个生物学实验。

NCBI National Center for Biotechnology Information 的缩写，位于美国的华盛顿区。

Open reading frame (ORF) 可读框。任意一段潜在编码蛋白质的 DNA。可读框起始于一个起始密码子并且终止于一个终止密码子。其内部不存在终止密码子。对可读框的鉴定是判断其该区段有可能作为一个功能性基因组分的首要指征。

Orthologue/paralogue 直系同源物/旁系同源物。旁系同源物与基因组内基因的复制相关。直系同源物在进化中保持了相同的功能，而旁系同源物会衍生出新的功能，即使与其起源基因相关。

Perl（practical extraction and report language） Perl 语言在一定的水平上相对比较直白，而这有助于其作为生物计算中主要语言的发展。

Relational database 关系型数据库。指遵循 E. F. Coddis′ 11 规范的数据库，一系列对数据进行组织和系统化的数学与逻辑步骤被植入软件系统，使数据易于提取、更新和扩充。关系型数据库管理系统（RDBMS）对数据在数据库中的存储包括一个或多个由行列组成的表格。行对应的是记录（条目），而列对应的是记录的属性（字段）。RDBMS

采用结构化检索语言（SQL）进行数据定义、数据管理以及数据操作和提取。关系型和对象关系型数据库广泛地用于生物信息学进行序列及其他生物学数据的存储。

Secondary structure (protein) 二级结构（蛋白质）。蛋白质多肽骨架的组织结构，如α螺旋或β折叠片层，其出现可由氢键造成。

Sequence tagged site (STS) 标签序列位点。位于已知染色体上的独特序列，可以经PCR扩增而获得。STS可以作为基因组定位和克隆的物理标记。

Single nucleotide polymorphism (SNP) 单核苷酸多态性。一个DNA序列上的变异导致了一个核苷酸被另外一个所置换。

Structured query language (SQL) 结构化检索语言。一种用于构建数据库检索和对关系型数据库进行更新及其他维护的程序设计语言。SQL不是一种成熟的语言，不能创建单独运行的应用程序，但功能强大，足以在其他数据库程序中建立交互式的脚本。

Substitution matrix 置换矩阵。在序列水平的蛋白质进化模型，可用来开发并产生出一整套用途广泛的置换矩阵，经常被称为Dayoff、MDM（mutation data matrix，突变数据矩阵）、BLOSUM或PAM（percent accepted mutation，突变承受百分率）矩阵。其由密切相关的序列经全局排列比对得出。较大进化差距的矩阵可以根据较小者推断。

Tertiary structure (protein) 三级结构（蛋白质）。通过蛋白质分子侧链间的相互作用造成的蛋白质链的折叠，包括半胱氨酸残基之间二硫键的形成。

UCSC（University of California，Santa Cruz） 一个杰出的基因组浏览器。

UTR（untranslated region） 非翻译区。mRNA转录物的非编码区域，可位于可读框两翼的任一端。

目　录