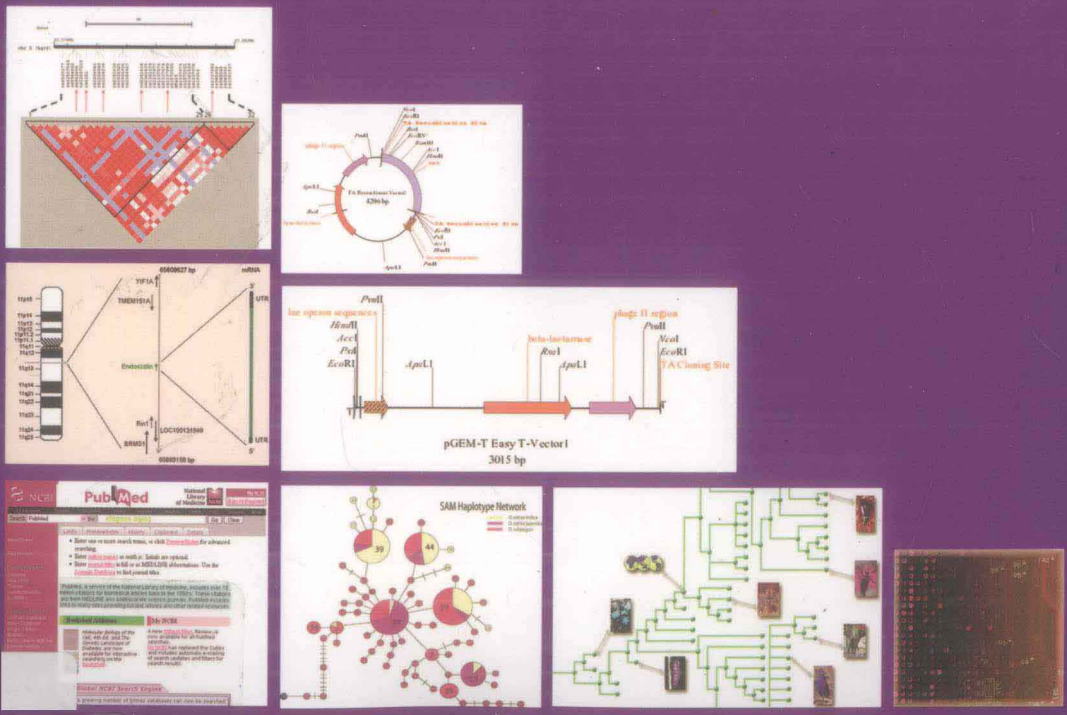
高等院校动植物类本科专业实验指导系列教材（之十）

Bioinformatics Experiment Manual

生物信息学 实验指导

主 编：于 浩 张 晶 张 英
主 审：欧阳红生



[illegible]

Bioinformatics Experiment Manual

生物信息学 [illegible]

[illegible]

高等院校动植物类本科专业实验指导系列教材(之十)

生物信息学实验指导

(第一版)

主　编　于　浩　张　晶　张　英
主　审　欧阳红生

吉林大学出版社

图书在版编目(CIP)数据

生物信息学实验指导/于浩，张晶，张英主编.—长春：吉林大学出版社，2009.9
（高等院校动植物类本科专业实验指导系列教材之十）
ISBN 978-7-5601-4421-4

Ⅰ.生… Ⅱ.①于… ②张… ③张… Ⅲ.生物信息论-实验-高等学校-教学参考资料 Ⅳ.Q811.4-33

中国版本图书馆 CIP 数据核字(2010)第 012606 号

书　名：高等院校动植物类本科专业实验指导系列教材（之十）
　　生物信息学实验指导
主　编：于　浩　张　晶　张　英

责任编辑、责任校对：矫正　　封面设计：杨　举
吉林大学出版社出版、发行　　吉林省金山印务有限公司　印刷
开本：787×1092　毫米 1/16　　2009 年 9 月　第 1 版
印张：16.5　　字数：316 千字　　2009 年 9月　第 1 次印刷
ISBN 978-7-5601-4421-4　　定价：31.00 元

社址：长春市明德路 421 号　邮编：130021
发行部电话：0431-88499826
网址：http://www.jlup.com.cn
E-mail:jlup@mail.jlu.edu.cn

高等院校动植物类本科专业实验指导系列教材

编写委员会

编审人员

主　编：于　浩　张　晶　张　英

副主编：王铁东　李　莉

主　审：欧阳红生

编　者：刘松财　赵志辉　逄大欣　宋　宇

王大成　焦虎平　任林柱　丁　镌

王　莹　梁　洋　饶家辉　王宏娟

序

培养学生实践动手能力和创新能力，是高等学校人才培养的主要目标之一，是本科教学质量与教学改革工程的重要内容。而大力加强实验教学，建设一批具有科学性、系统性、先进性和可操作性的实验教材，是不断提高实验教学水平和人才培养质量的有效保障。

吉林大学农学部历来重视通过实验教学培养学生的动手技能和创新能力。目前，在加强实验教学条件建设的同时，为适应人才培养目标和教学内容改革，加强实验教材建设，现以本校为主体，联合相关院校，编写了这套《高等院校动植物类本科专业实验指导系列教材》，涵盖了动物类专业、植物类专业和食品类专业等实验课程，计划出书20余部，与高水平实验教学示范中心建设相匹配，从而使实验教材建设规范化、配套化、系列化，进一步规范实验教学，对相关专业实验教学起到示范和带动作用。这套实验教材有三个比较突出的特点：

一是系统性。丛书涵盖了高等院校动植物类的动物医学、动物科学、生物技术、农学、园艺、植物保护、农业资源与环境、食品科学与工程、食品质量与安全等专业主要学科基础和专业必修课程，与每门课程的理论教材相配合，完善了教材体系建设。每本实验指导既单独成册、自成体系，同时又按专业分类规划、成型配套。这种实验教材编写方式，在其它学科专业领域有过成功范例，但在动植物类专业尚不多见。

二是实用性。参加丛书编写的教师，既有具有较高学术造诣的专家学者，又有长期从事实验教学的行家里手，均具有较强的教学内容选择和把握能力，在编写过程中注重了简洁明快，宜学宜用。每本教材对实验关键仪器设备的使用方法、注意事项给予了介绍，对每个项目的实验目的、材料、方法进行了说明，对实验内容、原理、操作、仪器设备的使用等进行了规范，加强了实验准备、基本规范、标准操作、参数测定、数据合成、误差分析、实验报告写作等训练，书中图例丰富，示范方法准确，着力强化基本实验操作能力的规范培养。丛书适用于全日制动植物类专业的本科生及研究生实验教学，也可作为相关专业科研人员的参考书

和技术人员的培训教材。

三是创新性。教材依据动植物类各专业实验课程的教学基本要求，融合专业改革和课程改革成果，结合理论教学的需要和实验条件的改进，以广受认可的高水平专业理论教材为蓝本，有计划地调整实验内容，对经典实验项目进行了改造，引入了本专业最新相关科研成果和国外高水平教材内容。在编写体例上，每本教材将实验项目划分成了演示性实验、验证性实验、综合性实验、设计性实验和研究性实验等类型，分章节安排编写，部分课程的综合性、设计性实验项目所占比例达到了30%以上，并安排了一定数量的由学生自主完成的综合性实验项目，引导学生自主设计、自主实验，加强了学生科学研究能力和团队协作精神培养，推进学生自主学习、合作学习、研究性学习。

系列化出版这样一套动植物类专业实验教学指导教材，在高等农业教育中还属于一个尝试。相信这套系列实验指导教材的出版和推广应用，能为提高学生的实践动手能力，为创新型人才培养起到应有的推动作用。

二〇〇九年五月二十八日

前　言

生物信息学(Bioinformatics)是以信息学的方法处理分析生物数据的一门学科。它通过对分子生物学实验数据的获取、加工、存储、检索与分析，进而达到揭示这些数据所蕴含的生物学意义的目的。随着人类基因组计划(HGP)的不断推进，生物信息学已经成为当今生命科学和自然科学的核心领域和最具活力的前沿领域之一。

当前，国内许多高校都面向本科生开设生物信息学课程，甚至有的高校开设了生物信息学(生物信息技术)的本科生专业。由于生物信息学属于交叉学科，不同的专业在面向授课时，总是从专业特性介绍生物信息学，使得国内没有统一的规范教材，基于上述考虑，作者在从事生物信息学教学的基础上，并结合自己在生物信息学领域的研究认识，编写一本以介绍生物信息学领域基础知识和概况的、适合理工农医等院系本科生、研究生通识教育的教材《生物信息学实验指导》。

本书由吉林大学畜牧兽医学院欧阳红生教授精心审稿和修改。

编　者

2009 年 7 月

目 录

第一部分
生物信息学理论及网络指南

第二部分
生物信息学软件的使用说明

第一部分 生物信息学理论及网络指南

第一章 引言

1. 生物信息学概念

生物信息学(Bioinformatics)是指信息技术和计算机科学在分子生物学领域中的应用。生物信息学一词是乌特勒支大学的 Paulien Hogeweg 教授于 1979 年在系统生物学中进行信息过程研究时创造出来的。自 20 世纪 80 年代末,生物信息学已经初步在基因组学和遗传学,特别是在涉及大规模 DNA 测序的基因组学研究领域得到应用。

生物信息学属于交叉学科,有许多不同的定义,生物信息学广义的概念是指应用信息科学的方法和技术,研究生物体系和生物过程中信息的存贮、信息的内涵和信息的传递,研究和分析生物体细胞、组织、器官的生理、病理、药理过程中的各种生物信息,或者也可以说成是生命科学中的信息科学。生物信息学狭义的概念是指应用信息科学的理论、方法和技术,管理、分析和利用生物分子数据。通过收集、组织、管理生物分子数据,使研究人员能够迅速地获得和方便地使用相关信息;通过处理、分析、挖掘生物分子数据,得到深层次的生物学知识,加深对生物世界的认识;在生物学、医学的研究和应用中,利用生物分子数据及其分析结果,可以大大提高研究和开发的科学性及效率,如根据基因功能分析结果来检测与疾病相关的基因,根据蛋白质分析结果进行新药设计。一般提到的"生物信息学"是就指这个狭义的概念,更准确地说,应该是分子生物信息学(Molecular Bioinformatics)。

现在,生物信息学需要的是在数据库、比对算法、计算和统计技术、解决在管理和分析生物数据产生问题的方法和理论方面的创新和进步,在过去的几十年

里，随着基因组学和其它分子相关的研究的快速发展，产生大量的分子信息数据，而应用数学和计算科学将有利于对生物过程的了解，生物信息学的主要功能是 DNA、蛋白质的序列分析，比对 DNA 和蛋白质序列间的差异，以及蛋白质三级结构的构建等。

生物信息学的产生是由于计算机技术和分子生物学在 40－50 年代的快速发展，但最为直接的因素是人类基因组计划的实施。其实，早在 20 世纪 50 年代生物信息学就已经形成萌芽，20 世纪 70 年代就已经产生生物信息学的基本思想，但是生物信息学的真正发展则是在 20 世纪的 90 年代，在人类基因组计划的推动下，生物信息学才得以迅猛发展。人类基因组计划产生的生物分子数据是生物信息学的源泉，而人类基因组计划所需要解决的问题则是生物信息学发展的动力。

2. 生物分子信息

生物信息学是在系统生物学的研究过程中产生的，生物体是一个复杂的系统，生命过程是一个极端复杂的过程。生物体同时也是一个信息系统，该系统控制着生物的遗传、生长和发育。所有的信息都存贮在生物体内的遗传物质中。如果能解读遗传物质的信息内涵，并能通过信息的流动过程了解生命的指令是十分必要的，而在生命科学的研究中，人们已经逐渐认识到，不仅需要用物理、化学和生物学方法研究生命的物质基础、能量转换、代谢过程等，还需要用信息科学方法研究生命信息特别是遗传信息的组织、复制、传递、表达及其作用，否则难以理解生命的工作机制，难以揭示生命的奥秘。

从信息学的角度来看，生物分子是生物信息的载体，生物信息学主要研究两种载体，即 DNA 分子和蛋白质分子。生物分子至少携带着三种信息，即遗传信息、与功能相关的结构信息、进化信息。

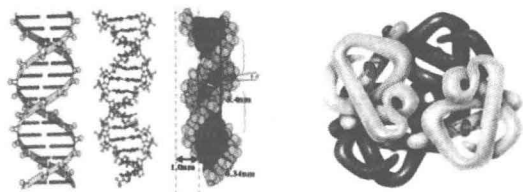


图 1-1 生物信息学研究的数据对象

DNA是遗传信息的载体。DNA的核苷酸序列上存储着蛋白质的氨基酸序列编码信息，存储着基因表达调控的信息，存储着遗传信息。遗传信息存储在DNA四种字符组成的序列中，生物体生长发育的本质就是遗传信息的传递和表达。因此，可以说DNA序列包含着最基本的生命信息。存储在DNA中的信息使无活力的分子组织成有功能的活细胞，进而构成能进行新陈代谢、生长和繁殖的生物体。人们已经认识到遗传信息的载体主要是DNA[在少数情况下核糖核酸(RNA)也充当遗传信息的载体]，控制生物体性状的基因是一系列DNA片段。一方面，DNA通过自我复制，在生物体的繁衍过程中传递遗传信息。另一方面，基因通过转录和翻译，使遗传信息在生物个体中得以表达，并使后代表现出与亲代相似的生物性状。在基因表达过程中，基因上的遗传信息首先通过转录从DNA传到RNA，然后再通过翻译从RNA传递到蛋白质。基因控制着蛋白质的合成，从基因的DNA序列到蛋白质序列存在着一种明确的对应关系，而这种对应关系就是我们所知道的第一遗传密码。

蛋白质分子在生物体内执行着各项重要任务，如生化反应的催化、营养物质的输运、信号的识别与传递等。蛋白质的功能多种多样，但是必须注意一点，即蛋白质功能取决于蛋白质的空间结构。要了解和掌握蛋白质的功能必须首先分析蛋白质的结构，对于其它生物大分子也一样。因此，蛋白质结构是一种重要的生物分子信息。然而，蛋白质结构决定于蛋白质的序列(这是目前基本共认的假设)，蛋白质结构的信息隐含在蛋白质序列之中。

作为信息的载体，DNA分子和蛋白质分子都打上了进化的烙印。通过比较相似的蛋白质序列，如肌红蛋白和血红蛋白，可以发现由于基因复制而产生的分子进化证据。比较来自于不同种属的同源蛋白质，即直系同源蛋白质，可以分析蛋白质甚至种属之间的系统发生关系，推测它们共同的祖先蛋白质。生物分子信息具体表现为DNA序列数据、蛋白质序列数据、生物分子结构数据、生物分子功能数据等。序列数据、结构数据是非常直观的，但是功能数据却是多变复杂的，如关于蛋白质功能的定性描述、蛋白质之间的相互作用描述、基因表达数据、代谢路径、调控网络等。在所有类型的数据中，序列是最基本的数据，而且也是目前最多的数据。

对生物分子数据及其关系的概括见下图。遗传信息从DNA序列向蛋白质序列的传递是人类已经基本了解的第一部遗传密码，然而蛋白质序列与蛋白质

结构也存在着一定的对应关系，蛋白质序列决定蛋白质结构，因此有人将从蛋白质序列到蛋白质结构的关系称为第二部遗传密码。

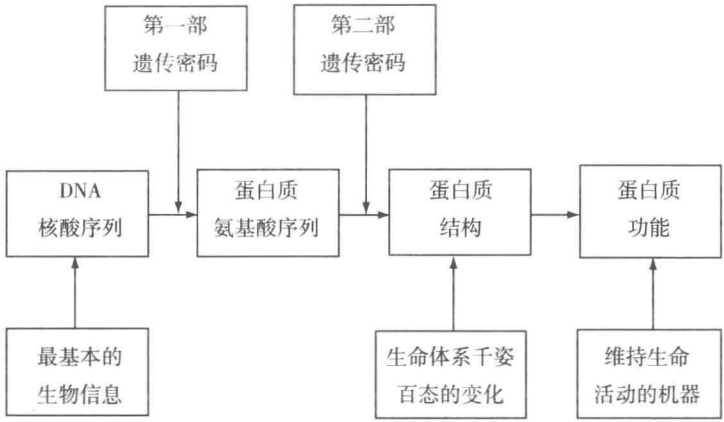


图 1-2 生命过程中的遗传语言

第一部遗传密码已被破译，但是，对于密码究竟处于 DNA 序列的哪些区域还了解得不全面，对密码的转录过程还不清楚，对大多数 DNA 非编码区域的功能还知之甚少，对 DNA 遗传语言还有待于进一步探索。对于第二部密码，目前则只能用统计学的方法进行分析。无论是第一部遗传密码，还是第二部遗传密码，都隐藏在大量的生物分子数据之中。生物分子数据是宝藏，生物信息数据库是金矿，等待我们去挖掘和利用。

与一般信息相比，生物分子信息具有明显的特征。首先，生物分子信息数据量大，例如 DNA 序列以千兆碱基(Giga base，Gb)为单位。随着信息处理技术进入现代生物学研究领域，随着互联网在全球的贯通，各种生物信息学数据库迅速发展，生物分子数据积累速度成倍增长。其次，生物分子信息复杂，既有生物分子序列的信息，又有结构和功能的信息，既有生命本质信息，如基因，又有生命表象信息，如基因表达信息。生物分子信息另一个重要的特征是，生物分子信息之间存在着密切的联系，例如，基因序列与蛋白质序列之间的关系，生物分子序列与结构之间的关系，结构与功能之间的关系，基因变异与疾病之间的关系。

对于生物分子信息，靠人工难以完成数据处理和分析的任务，更谈不上发现隐藏在这些信息之中的内在规律。同时，对于生物分子信息，仅靠某一学科的专家，也无法进行分析研究，因此，在生物信息学研究领域中，要求生物学家、数学

家和计算机科学工作者协力合作，发展新的分子生物学计算理论和方法，运用先进的计算机技术收集、集成和分析处理生物信息。

3. 生物信息学主要研究内容

生物信息学作为一门新的交叉学科，其研究范畴是以基因组 DNA 序列的信息分析作为出发点，分析基因组结构，寻找或发现新基因，分析基因调控信息，并在此基础上研究基因的功能，研究基因的产物即蛋白质，模拟和预测蛋白质的空间结构，分析蛋白质的性质，其结果将为基于靶分子结构的药物分子设计和蛋白质分子改性设计提供依据。当前，生物信息学已在理论生物学领域占有了核心的地位。

生物信息学主要有以下几个方面的研究内容。

(1)序列分析

自从 1977 年噬菌体 Φ-X174 基因组被测通以来，数以万计的生物体的 DNA 序列已被解码并存储在数据库上。对这些序列信息进行分析，以确定基因编码多肽(蛋白质)、RNA 基因、调控序列、结构域和重复的序列。比较不同物种之间同源基因的相似之处可以揭示同源蛋白质的功能，或物种之间的进化关系(即构建分子系统树)。基因组计划开展以来，DNA 数据量激增，人工手段分析 DNA 序列已变得不切实际。

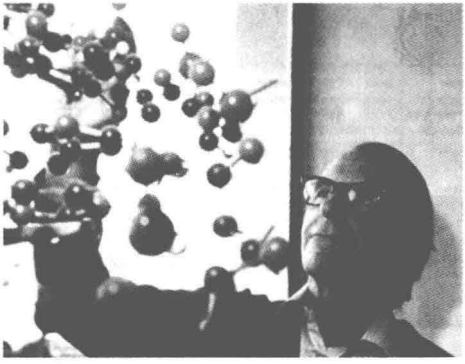


图 1-3 英国生物化学家 Frederick Sanger 测通了噬菌体 Φ-X174 全基因组

今天，像 BLAST 这样的计算机程序，每天被数以万计的人使用搜索含有数百亿核苷酸的各色生物体的基因组。这些程序可以用于比对含有突变(交换，删

除或插入基地)的 DNA 序列,以确定哪些序列是相关的,但不完全相同的。这个比对算法更大的改进时是用在测序过程本身。所谓的鸟枪测序技术[例如基因组研究所(The Institute for Genomic Research)就是使用这种方法测通了第一个细菌的基因组,即流感嗜血杆菌],鸟枪测序技术不会对整个染色体按顺序测序,而是产生了成千上万的小 DNA 片段的序列(35 至 900 个核苷酸)。这些片段的两端重叠,当这些小片段 DNA 序列排列得当,可以用来重建完整的基因组。鸟枪法测序速度快,但对于相当复杂的大型基因组,其碎片组装任务可是相当艰巨的。比如人类基因组,它组装基因组片段可能需要多处理器、大内存的计算机运行很多天,以及由此产生的重叠群通常包含许多空白要在以后填补。今天,几乎所有的基因组测序都要选择鸟枪法测序,因此基因组组装算法成了生物信息学研究的重要领域。

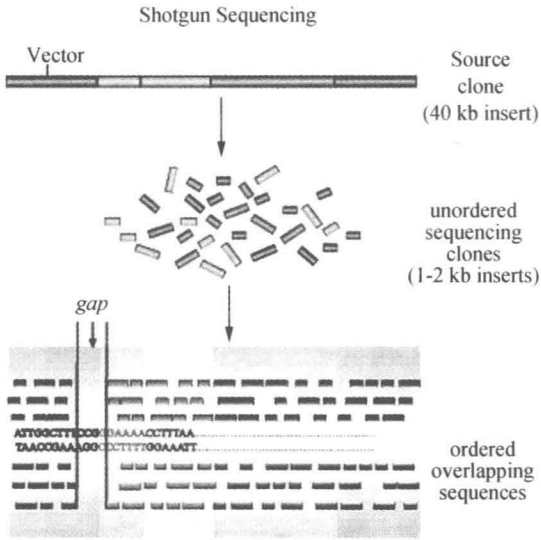


图 1-4 鸟枪法测序过程

生物信息学的另一种序列分析是"annotation-注释",涉及到使用计算生物学的方法在基因组上搜索的蛋白质编码基因、RNA 基因、以及其它的功能基因组序列。基因组内不是所有的核苷酸都是基因。在高等生物中,在基因组 DNA 的大部分区域表现不出任何明显的功能。然而这种所谓的"垃圾 DNA",可能包含无法识别的功能元件。生物信息学有助于在基因组和蛋白质组的鸿沟上建立桥