“十二五”国家重点出版物出版规划项目
北京理工大学教育基金会·教授文库

生物信息处理技术与方法

Biological Information Processing Techniques and Methods

罗森林　潘丽敏　马　俊　编著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

“十二五”国家重点出版物出版规划项目
北京理工大学教育基金会·教授文库

生物信息处理技术与方法

Biological Information Processing Techniques and Methods

罗森林　潘丽敏　马　俊　编著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

内 容 提 要

本书共分8章，主要内容包括生物信息处理知识基础、数据处理方法基础、序列比对方法、系统发生树构建方法、基因芯片数据处理方法、RNA结构预测方法、蛋白质结构预测方法、生物分子网络构建方法等。

本书可用作计算机科学与技术、生命信息工程、软件工程、通信与信息系统等相关学科、专业的教材，也可作为参考书直接使用，同时也可供科研人员参考和有兴趣者自学使用。

图书在版编目（CIP）数据

生物信息处理技术与方法/罗森林，潘丽敏，马俊编著．—北京：北京理工大学出版社，2015.1

ISBN 978-7-5640-8314-4

Ⅰ.①生… Ⅱ.①罗…②潘…③马… Ⅲ.①生物信息论-信息处理 Ⅳ.①Q811.4

中国版本图书馆CIP数据核字（2015）第209153号

出版发行／北京理工大学出版社有限责任公司
社　　址／北京市海淀区中关村南大街5号
邮　　编／100081
电　　话／（010）68914775（总编室）
　　82562903（教材售后服务热线）
　　68948351（其他图书服务热线）
网　　址／http：／／www.bitpress.com.cn
经　　销／全国各地新华书店
印　　刷／保定市中画美凯印刷有限公司
开　　本／710毫米×1000毫米　1/16
印　　张／20.75
字　　数／368千字
版　　次／2015年1月第1版　2015年1月第1次印刷
定　　价／56.00元

责任编辑／王玲玲
文案编辑／王玲玲
责任校对／周瑞红
责任印制／马振武

前 言

生物信息数据的快速增长迫切需要生物信息处理技术与方法的有效应用和发展，生物信息处理涉及内容非常广泛，学科间相互交叉，互融关系复杂，本书梳理了生物信处理技术与方法的知识点，注重领域内核心思想、原理、方法的论述，并融入国内外最新研究进展，内容力求系统、全面、先进。在讨论技术与方法的同时，引入应用实例以强调其具体应用方法，使理论联系实际，有利于技术与方法的快速掌握和有效运用。

本书经过长期酝酿并总结多年的教学、应用经验认真构架而成，以便于学生充分利用生物信息处理技术。全书共分 8 章，各章的主要内容安排如下：

第 1 章为绪论。内容包括生物信息处理的产生背景和意义、生物信息处理知识基础、生物信息处理发展简史和现状、生物信息处理数据库及技术工具、技术难点与发展趋势等。

第 2 章为数据处理方法基础。内容包括概率论基础、数据分类分析、数据聚类分析、关联规则发现、隐马尔科夫模型、高维数据处理等。

第 3 章为序列比对方法。内容包括序列比对知识基础、双序列比对、多序列比对、应用实例分析等。

第 4 章为系统发生树构建方法。内容包括系统发生树知识基础、基于距离的构建方法、基于离散特征的构建方法、Quartet 方法、应用实例分析等。

第 5 章为基因芯片数据处理方法。内容包括基因芯片知识基础、基因芯片数据预处理、基因芯片数据聚类分析、基因芯片数据分类分析、应用实例分析等。

第 6 章为 RNA 结构预测方法。内容包括 RNA 知识基础、比较序列分析方法、动态规划算法、组合优化算法、启发式算法、应用实例分析等。

第 7 章为蛋白质结构预测方法。内容包括蛋白质结构知识基础、蛋白质二级

结构预测、蛋白质三级结构预测、应用实例分析等。

第 8 章为生物分子网络构建方法。内容包括生物分子网络知识基础、基因调控网络构建方法、蛋白质互作网络构建方法、应用实例分析等。

本书由罗森林、潘丽敏、马俊共同撰写，其中第 3～5 章的基础技术与方法部分主要由潘丽敏负责撰写，第 6～8 章的基础技术与方法部分主要由马俊负责撰写，其余部分主要由罗森林负责撰写。罗森林负责整书的章节设计、内容规划和统稿工作。

在本书的编写过程中，得到了北京理工大学杨煜祥、刘畅、明道福老师以及陈功、郭峰、郭伟东、李金玉、刘盈盈、刘峥等同学多方面的帮助，在此一并表示衷心的感谢。

由于时间有限，加之笔者能力范围的限制，书中疏漏之处敬请广大师生批评指正，以使本书日渐完善。谢谢！

罗森林

目 录

第 1 章

绪论

1.1 产生背景和意义

生物信息处理的研究目标是通过处理复杂的 DNA、RNA、蛋白质等生物数据，揭示基因组信息结构的复杂性及遗传语言的根本规律，解读人类基因组全部 DNA 序列，认识人类自身，揭示遗传、发育和进化的联系等。生物信息处理丰富和发展了现有的物理学、生物学、化学、数学、计算机科学、信息科学和系统科学的理论和方法，推动了学科群的进步，成为自然科学中多学科交叉的有活力、有影响的新领域。相对于其他日渐成熟的学科，对生物信息处理技术与方法的研究仍然处于初始阶段，随着生物技术的快速发展，新的数据、新的需求的不断涌现，生物信息处理的理论与技术将会快速发展。

1.1.1 产生与兴起

生物信息处理的实质，就是利用计算机科学和网络技术来解决生物学问题，是由生物学对大量数据处理和分析的需求而引发的，它的诞生和发展是应时所需，是历史的必然。20 世纪，尤其是 20 世纪末期，生物科学技术迅猛发展，无论是从数量上还是从质量上，都极大地丰富了生物科学的数据资源。数据资源的急剧膨胀使人们不得不考虑寻求一种强有力的工具去组织它们，以利于对已知生物学知识的储存和进一步加工利用。

1970—1980 年，随着生物化学技术的发展，许多生物分子序列数据产生，促使一部分计算机科学家应用计算机技术解决生物学问题，特别是与生物分子序列相关的问题，并提出了一系列著名的序列比较算法。1980 年以后，一批生物信息服务机构和生物信息数据库被创建出来。2001 年，人类基因组工程测序的完成，使生物信息处理达到了一个高潮。由于 DNA 自动测序技术的快速发展，DNA 数据库中的核酸序列公共数据量以每天 10^6 bp 的速度增长，生物信息数据迅速膨胀，从积累数据向解释数据的时代转变，同时，数据的大量积累也蕴含着潜在的突破性发现的可能。

生物信息处理是建立在分子生物学的基础上的，因此，要学习生物信息处理，就必须先对分子生物学的发展有所了解。对生物细胞的生物大分子的结构与功能的研究很早就已经开始了。1866 年，孟德尔在实验的基础上提出了假设——基因以生物成分存在。1871 年，Miescher 从死的白细胞核中分离出脱氧核糖核酸（DNA）。1944 年，Avery 和 Mc Carty 证明了 DNA 是生命器官的遗传物质，在此之前，人们认为染色体蛋白质携带基因，而 DNA 则是一个次要的角色。同年，Chargaff 发现了著名的 Chargaff 规律，即 DNA 中鸟嘌呤的量与胞嘧啶的量相等，腺嘌呤的量与胸腺嘧啶的量相等。与此同时，Wilkins 与 Franklin 用 X 射线衍射技术测定了 DNA 纤维的结构。1953 年，James Watson 和 Francis Crick 在《科学》上推测出 DNA 的三维结构——双螺旋，即 DNA 以磷酸糖链形成双股螺旋，脱氧核糖上的碱基按 Chargaff 规律构成双股磷酸糖链之间的碱基对。这个模型表明 DNA 具有自身互补的结构，根据碱基对原则，DNA 中储存的遗传信息可以精确地进行复制。他们的理论奠定了分子生物学的基础。DNA 双螺旋模型已经预示了 DNA 复制的规则。1956 年，Kornberg 从大肠杆菌中分离出 DNA 聚合酶Ⅰ（DNA polymerase Ⅰ），它能使 4 种 dNTP 连接成 DNA。DNA 的复制需要一个 DNA 作为模板。Meselson 与 Stahl 于 1958 年用实验方法证明了 DNA 复制是一种半保留复制。1954 年，Crick 提出了遗传信息传递的规律：DNA 是合成 RNA 的模板，RNA 又是合成蛋白质的模板，这个规律称为中心法则（Central Dogma），其对分子生物学和生物信息处理的发展都起到了极其重要的指导作用。此后，经过 Nirenberg 和 Matthai 的努力研究，编码 20 氨基酸的遗传密码得到破译，限制性内切酶的发现和重组 DNA 的克隆奠定了基因工程的技术基础。

正是由于分子生物学的研究对生命科学的发展有巨大的推动作用，生物信息处理理论与技术的出现也就成了一种必然。该领域的主要课题是研究如何通过对 DNA 序列的统计计算分析，以便更加深入地理解 DNA 序列、结构、演化及其与生物功能之间的关系，其研究课题涉及分子生物学、分子演化及结构生物学、统计学及计算机科学等许多领域。生物信息处理是内涵非常丰富的学科，其核心是

基因组信息学，包括基因组信息的获取、处理、存储、分配和解释。基因组信息学的关键是“读懂”基因组的核苷酸顺序，即全部基因在染色体上的确切位置以及各 DNA 片段的功能。同时，在发现了新基因信息之后，对其进行蛋白质空间结构模拟和预测，然后依据特定蛋白质的功能进行药物设计。了解基因表达的调控机理也是生物信息处理的重要内容，根据生物分子在基因调控中的作用，描述人类疾病诊断、治疗的内在规律。生物信息处理理论与技术的研究目标是揭示“基因组信息结构的复杂性及遗传语言的根本规律”，解释生命的遗传语言，其已成为整个生命科学的重要组成部分，成为生命科学研究的前沿。

1.1.2 研究的意义

生物信息处理的研究任重而道远，如同门捷列夫在分析化学元素的性质数据时发现了元素周期表一样，生物信息处理也要通过对生物学数据的分析研究，归纳总结出生物系统生长、演化的规律。

生命科学研究最重要的突破莫过于对生物遗传基因物质（DNA）的测定，主要研究集中在对生物学数据的收集、整理、筛选、编辑、管理、显示、利用（计算、模拟）等。主要研究方向包括序列比对、基因识别（解码）、基因重组、蛋白质结构预测、基因表达、蛋白质反应的预测以及建立演化模型等。

生物遗传基因载体物质的发现和成功测定打开了生物信息处理研究的大门。以人为对象，研究 30 亿个碱基哪一段究竟代表什么意思，哪一段管理生物个体中哪一部分、执行哪一个功能，被称为基因识别与解码。研究同一种基因在不同个体中的微小不同（序列比对），以及所导致的不同形态（基因表达），还有它们的运动变化规律（建立演化模型），都是非常有意义也非常有趣味的工作。这些工作，可以帮助解释遗传现象，防治遗传疾病。计算机技术强大的计算能力与数学统计理论的结合，也使得在一组序列中找出父-子演化传承关系成为可能。生物信息处理研究工作不仅需要计算机技术的知识、统计学的知识和信息学的知识，同时也需要分子生物学的知识。由此可见，生物信息处理是一个多学科结合的交叉研究领域，需要各学科的合作和共同努力。

总之，生物信息处理是一门研究生物和生物相关系统中信息内容与信息流向的综合系统科学。只有通过生物信息处理相关理论与技术，人们才能从众多分散的生物学观测数据中获得对生命运行机制的系统理解。从工具的角度来讲，生物信息处理几乎是今后所有生物（医药）研究开发所必需的工具。对生物信息处理的研究不仅具有重大的科学意义，还具有巨大的经济效益。

1.2 知识基础

1.2.1 基本概念

1. 生物信息处理的定义

生物信息是指决定生物体性状特征的信息。它包含三个层次：储存在 DNA 线性分子中的一维信息，即遗传密码包含的遗传信息；储存在蛋白质分子中的三维信息，即由 DNA 分子决定的肽链经折叠呈现生物学功能的蛋白质三维结构；储存在 DNA、蛋白质等各类物质分子中的按时间、空间的特定程序相互作用的网络系统的四维结构。

数据处理是对数据进行采集、存储、检索、加工、变换和传输。数据是事实、概念或指令的一种表达形式，可由人工或自动化装置进行处理。数据的形式可以是数字、文字、图形或声音等。数据经过解释并赋予一定的意义之后，便成为信息。数据处理的基本目的是从大量的、可能是杂乱无章的、难以理解的数据中抽取并推导出某些特定的、有价值、有意义的信息。数据处理技术的发展及其应用的广度和深度，极大地影响着人类社会发展的进程。数据处理离不开软件的支持，数据处理软件包括：用以书写处理程序的各种程序设计语言及其编译程序，管理数据的文件系统和数据库系统，以及各种数据处理方法的应用软件包。为了保证数据安全可靠，还有一整套数据安全保密的技术等。

技术方法是人们在技术实践过程中所利用的各种方法、程序、规则、技巧的总称，它帮助人们解决“做什么”、“怎样做”以及“怎样做得更好”的问题。人们在技术活动中利用技术知识和经验，选择适宜的技术方法或创造出全新的方法，以完成设定的技术目标。

生物信息处理技术与方法涵盖了生物信息、数据处理和技术方法等多方面的内容，其主要目的是，用计算机科学、信息技术以及数学理论来处理生物学问题，主要内容包括：生物学数据的获取、存储、处理、管理和可视化，基因遗传和物理图谱的处理，核苷酸和氨基酸序列分析，新基因的发现和蛋白质结构的预测等。

生物信息学是研究生物信息的采集、处理、存储、传播、分析和解释等各方面的一门学科，综合利用生物学、计算机科学和信息技术以揭示大量而复杂的生物数据所包含的生物学奥秘。生物信息学属于典型的交叉学科，而生物信息处理是进行该学科研究的主要理论与技术方法，侧重于利用数学模型和计算仿真技术对生物学问题进行研究。通常，生物信息处理的研究可以划分成两个阶段，第一阶段是数据挖掘和知识发现，即从大量的实验数据中提取隐藏的模式，然后形成

假设；第二个阶段是建立数学模型，利用计算机模拟来检验各种假设，为进一步的实验研究提供预测结果和指导建议。生物学处理的特点就在于两个研究阶段的不可分割性，它既不是单纯的生物信息学研究，也不是纯粹的生物数学理论研究，更不是简单的计算机技术应用研究。

2. 生物信息处理的特点

概括起来讲，生物信息处理有以下几方面的特点：

（1）交叉性

生物信息处理和生物学的其他分支一样，有一个共同的目标，就是揭示生命的奥秘，探索生命现象中的规律，为人类创造更美好的生活。然而，它的研究手段完全不同于传统生物学实验，而是从大量不连贯的生物学实验数据中发现有用的生物学信息，这离不开现代信息技术、计算机技术和数学。生物信息处理并非生物学或信息科学的一个简单分支，而是多学科的有机交叉。

（2）复杂性

生物数据的海量性和生物系统本身的复杂性都对生物信息处理研究提出了挑战。仅人类基因组就产生一部几十亿字符的“天书”，而这几十亿字符是四个字母的重复，没有语法，也没有标点符号。如何读懂这部“天书”，以现有的计算技术，仍然是一个无法解决的难题。

（3）广泛性

生物学数据每天以千万的数量级呈爆炸式增长，除了数量上的增长，生物学的研究范围也绝不局限于人类基因组计划，各种植物和动物的基因组研究相继展开。随着人类基因组计划的顺利进行，蛋白质组、人类基因组多样性计划、比较基因组、环境基因组和药物基因组的研究也相继被提出来。

（4）前沿性

生物信息处理用最先进的信息技术和数理技术研究生命本质，帮助人们逐步认识生命的起源、进化、遗传和发育的本质，破译隐藏在 DNA 序列中的遗传语言，揭示人体生理和病理的分子基础，为人类疾病的预测、诊断、预防和治疗提供最合理、最有效的方法和途径。

1.2.2 生物信息数据特点

生物信息不仅包括基因组信息，如基因的 DNA 序列、染色体定位等，也包括基因产物——蛋白质或 RNA 的结构和功能及各物种间的进化关系等其他信息资源。就数据分析而言，生物信息数据的特点包括：

1. 高通量和大数据量

人类基因组计划（HGP）产生了很多高通量技术，如一次基因表达谱芯片实

验可以获得数万个基因表达数据，一次大规模基因组测序可以获得数亿个序列数据。人类基因组由 3×10^9 碱基对组成，各种模式的生物基因组序列、蛋白质序列源源不断地产生，在此基础上，还可以产生数倍的二次数据。基因组的基因表达数据因时间、环境不同而不同，基因表达数据的数据量将很快超过基因序列的数据量。生物信息以指数级快速增长，远远超出传统分析方法的处理能力。

2. 多类型

生物信息数据包括了 DNA 序列、蛋白质序列、蛋白质各级空间结构数据、基因表达、代谢途径和文献等，各种数据的特性不同，存储方式不同，这给数据集成、共享、分析都带来很多困难。目前的数据库管理系统并不适合生物信息中生物序列数据的存储和检索。

3. 异构性

生物信息数据的异构性包括结构上的异构、语义上的异构和系统实现上的异构三大类。结构上的异构指同一个数据采用不同的数据模型或不同的数据结构来表示；语义上的异构指同一个术语在不同的地方代表不同的含义，或同一个含义用不同的术语来表示；系统实现上的异构指生物数据有的是以文本形式组织的，有的是以关系表的形式组织的。生物数据以各种形式存储于网络上的数据源中，即使同一数据，也有不同的存储形式和存储内容，难以满足共享、交流、集成、综合分析的要求。

4. 网络性和动态性

生物信息数据的网络性，一方面是指生物数据大部分存在于互联网中，数据库分布在不同的研究机构、不同的地理区域和不同的服务器系统上，具有自治的特点。这些数据库通过网络实现互连，进行数据的存取。如三大核酸序列数据库访问以及目前序列常规分析比对均需通过互联网完成。另一方面是指数据之间本身就相互作用、相互关联，如基因调控网络、代谢网络以及不同种类数据之间的相互作用网络。

动态性一方面是指数据随研究的深入而不断被更新，如瑞士日内瓦大学的 SwissProt 数据库，每日更新文件，一段时间后会有更新汇总文件和新的版本发布。另一方面指数据之间相互作用、相互关联的动态关系。

5. 高维

在一个平面或关系数据库中，记录中的每一个字段代表一维。很多生物信息数据具有高维特征，例如表达谱数据所分析的情形的个数，可以构成几十维数据；而在序列数据分析中，往往将一个单位（如碱基、氨基酸）当作一维，这样数据就会有几十维甚至上百维。

6. 序列数据

序列数据是目前生物数据中数据量最大的基础数据，其特点有：所用符号集合很小，例如 DNA 序列仅由 A、C、T、G 四个字符构成；序列长短差别很大，有的只有几十个字符，而有的会达到 1 M 的长度；总量巨大且增加迅速。序列数据的存储、分析都不同于典型的数据类型的处理。

1.2.3 主要研究内容

1. 基因组学研究

基因组表示一个生物体所有遗传信息的总和。一个生物体基因组所包含的信息决定了该生物体的生长、发育、繁殖和消亡等几乎所有的生命现象。研究基因组的学科称为基因组学，根据研究重点的不同，基因组学可以分为序列基因组学、结构基因组学、功能基因组学与比较基因组学。2001 年 2 月 12 日，人类基因组的精细图谱被公布在《自然》和《科学》上。 2002 年，在中国上海召开的第 7 次国际人类基因组大会，标志着一个关键性的转折，即国际基因研究正从大规模基因组测序转向与基因诊断和基因治疗息息相关的功能基因组学领域。

以下简要介绍一些基因组学的研究重点：

（1）序列比对

序列比对的基本任务是比较两个或两个以上符号序列的相似性或不相似性。从生物学的角度来看，该问题包含以下几个意义：从相互重叠的序列片断中重构 DNA 的完整序列；在各种试验条件下，由探测数据决定物理和基因图存储；遍历和比较数据库中的 DNA 序列；比较两个或多个序列的相似性；在数据库中搜索相关序列和子序列，寻找核苷酸的连续产生模式；找出蛋白质和 DNA 序列中的信息成分。序列比对考虑了 DNA 序列的生物学特性，如序列局部发生的插入、删除和替代，序列目标函数的获得，序列之间突变集最小距离加权和或最大相似性和。序列比对常采用动态规划算法和启发式的方法，动态规划算法在序列长度较小时适用，而对于海量基因序列，如人的 DNA 序列（高达 10^9bp），就需要采用启发式的方法进行比对。

（2）基因识别，非编码区分析

基因识别的基本任务是，给定基因组序列后，正确识别基因的范围，并在基因组序列中精确定位。非编码区由内含子组成，一般在形成蛋白质后被丢弃，但在实验中，如果去除非编码区，则不能完成基因的复制。显然，DNA 序列作为一种遗传语言，既存在于编码区，又隐含在非编码序列中。目前没有一般性的指导方法用以分析非编码区 DNA 序列，侦测非编码区的方法包括测量非编码区密码子的频率、一阶和二阶马尔科夫链、ORF（Open Reading Frames）、启动子识别、

HMM（Hidden Markov Model）、GENSCAN 和 Splice Alignment 等。

（3）分子进化和比较基因组学

分子进化是根据不同物种间同一基因序列的异同来研究生物的进化，并构建进化树。既可以用 DNA 序列也可以用其编码的氨基酸序列来构建，还可以通过相关蛋白质的结构比对来研究分子进化，但其前提假定是，相似种族在基因上具有相似性。通过比对，可以在基因组层面上发现哪些是不同种族所共同的，哪些是不同的。早期研究方法常以外在的因素，如大小、肤色、肢体的数量等，作为进化的依据。近年来，随着较多模式生物基因组测序任务的完成，人们可从整个基因组的角度来研究分子进化。在匹配不同种族的基因时，一般须处理三种情况：Orthologous——不同种族，相同功能的基因；Paralogous——相同种族，不同功能的基因；Xenologs——有机体间采用其他方式传递的基因，如被病毒注入的基因。这一领域常用方法是构造进化树，通过基于特征的方法（即 DNA 序列或蛋白质中的氨基酸的碱基的特定位置）、基于距离（对齐的分数）的方法和一些传统的聚类方法（如 UPGMA）来实现。

（4）序列重叠群装配

在测量人类基因时采用了短枪方法，要求把大量的较短的序列全体构成重叠群，并逐步将其拼接起来，形成序列更长的重叠群，直至得到完整序列。这个过程称为重叠群装配。从算法角度来看，序列的重叠群装配是一个 NP——完全问题。

2. 蛋白质组学研究

蛋白质组是指一个基因组、一种生物或一种细胞组织所表达的整套蛋白质；而有关蛋白质组的研究称为蛋白质组学。蛋白质组学的核心内容包括蛋白质组研究体系的建立、完善以及与重要生物学问题有关的功能蛋白质组的研究两个部分；而蛋白质信息学则涉及蛋白质数据库的建立、相关软件的开发与应用，并进而开展重要蛋白质的结构预测、三维结构和动态结构的研究，在蛋白质组水平上深入探索其作用模式、功能机理、调节控制及其在蛋白质群体内或与相关生物大分子间的相互作用。

3. 生物芯片

生物芯片主要是根据分子间特异性相互作用的原理，将生命科学领域中不连续的分析过程集成于芯片表面，构建微流体生物化学分析系统，以实现对细胞、蛋白质、核酸、糖类及其他生物组分的准确、快速、大信息量的检测。按照芯片上固定的生物大分子的不同，可以将生物芯片划分为基因芯片、DNA 芯片、PNA 芯片、蛋白质芯片和芯片实验室等。而从其功能的角度来划分，生物芯片又可分为测序芯片、表达芯片和比较基因组杂交（CGH）芯片。生物芯片可以广泛应用

于基因差异表达分析、DNA 测序、基因突变及多态性扫描、基因组 DNA 突变及染色体变异检测、肿瘤与传染病的诊断、环保监测、药物筛选、食品监督、商品检验、司法鉴定和军事等多方面。

4. 生物计算机

生物计算机是以生物界处理问题的方式为模型的计算机，目前主要有生物分子或超分子芯片、自动机模型、仿生算法、生物化学反应算法等几种类型。DNA 计算机是一种生物化学反应计算机，它是计算机科学与分子生物学相互结合、相互渗透而产生的新兴交叉研究领域。DNA 计算机基本设想是，以 DNA 碱基序列作为信息编码的载体，利用现代分子生物学技术，在试管内控制酶作用下的 DNA 序列反应，以实现运算，即以反应前的 DNA 序列作为输入的数据，以反应后的 DNA 序列作为运算的结果。DNA 计算机的重要特点是信息容量的巨大性与密集性以及处理操作的高度并行性，通过强力搜索策略迅速得出正确答案，从而使其运算速度大大超过常规计算机的速度。DNA 计算机的许多方面都还很不成熟，主要表现在构造的现实性、计算潜力、运算过程中的错误问题以及人机界面。无论如何，生物计算机的提出开拓了人们的视野，启发人们用算法的观念来研究生命，向众多相关领域提出了挑战。

5. 生物学数据库

随着大量生物学实验数据的积累，多种生物学数据库也相继形成，它们各自按照一定的目标收集和处理生物学实验数据，并提供相关的数据查询和数据处理的服务。现阶段，数据库的类型几乎涵盖了生命科学的各个领域。国际上主要的核酸序列数据库有 GenBank、EMBL、DDJB，蛋白质序列数据库有 SwissProt、PID、OWL、ISSD，蛋白质片段数据库有 PROSITE、BLOCKS、PRINTS，三维结构数据库有 PDB、NDB、BisMagResBank、CCSD，与蛋白质结构有关的数据库还有 SCOP、CATH、FSSP，与基因组有关的数据库还有 ESTdb、OMIM、GDB、GSDB，文献数据库有 Medline、Uncover。另外，一些公司开发了商业数据库，如 MDL。一些生物计算中心将多个数据库整合在一起提供综合服务，如 EBI 的 SRS 包括了核酸序列数据库、蛋白质序列数据库、三维结构数据库等 30 多个数据库及 ClustalW、PROSITESEARCH 等强有力的搜索工具，便于用户进行多个数据库的多种查询。生物学数据库除了在种类和数量上有急剧增长外，其复杂程度也在不断增加，但是，数据库的管理和使用却越来越便捷，目前，大多数数据库都具有自动投送数据、在线查询、在线计算和空间结构的可视化浏览等多种功能。成立于 1997 年 3 月的北京大学生物信息中心所建立的数据库和服务项目在国内是最多的，我国对数据库的研究起步很晚，因此有两点特别重要：一是构建我国自己的数据库；二是与国际常用数据库的有效连接和及时更新。