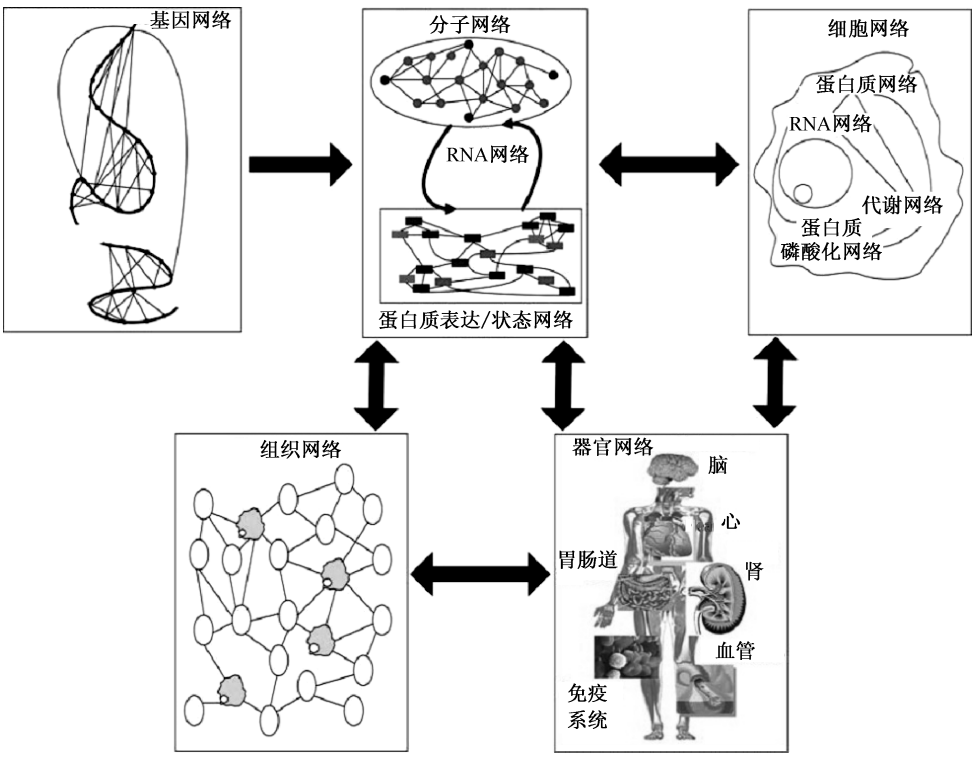
第 1 章　生物信息学简介

21 世纪是生命科学的时代，也是信息科学的时代，作为生命科学与信息科学碰撞的产物，生物信息学一经兴起，就开始蓬勃发展，成为世人瞩目的焦点。从基因到分子，从细胞到组织，从器官到个体，生命的信息不断流动，周而复始，生生不息。作为生命信息的传递中心，人们不禁猜想：21 世纪的生物信息学将发挥怎样的作用？生物信息学能否给人类带来革命性的重大发现呢？



1.1 引言

20 世纪，生命科学得到了飞速发展，生理学、细胞生物学、分子生物学等学科的发展使人们从器官、组织、细胞、生物大分子等各个层次认识了生命的物质基础。由于分子生物学研究的不断深入和实验技术的快速发展，使得人们可以从分子层面鉴定和测量生物系统中的所有生物大分子。首先，人类为了更深入地了解和认识自身，制定了宏伟的人类基因组计划。随着人类基因组计划的完成和多种生物基因组测序计划的开展，已产生了海量的全基因组序列信息。之后，为了考察不同个体之间的基因组差异，相继开展了人类单倍型图谱计划和千人基因组计划，使得基因组数据成倍增长。而转录组学和蛋白质组学研究产出的数据比基因组序列更加复杂，因为这些数据不像基因组序列是静态的，而是随着时间、条件、样本等一直在发生改变，对任意时刻和条件下的测量都会产生海量的转录表达和蛋白质表达数据。这些生物分子数据从基因、转录物、蛋白质和代谢物等各个层面揭示了生命的特征，隐藏着人类目前尚不知道的生物学知识。而且生物系统并非生物大分子的简单堆积，生物体的生长发育是生命信息控制之下的复杂而有序的过程，牵涉到生物信息的组织、传递和表达。因此，为了充分利用各种生物学数据，通过数据分析、处理，揭示这些数据的内涵，人们开始尝试用信息科学的方法和技术来认识和分析生命信息。**生物信息学**(bioinformatics)就是在这种数据大爆炸背景下出现的一门新兴学科，它是由生物学、应用数学、计算机科学相互交叉所形成的学科，是当今生命科学和自然科学的重大前沿领域之一，也是 21 世纪自然科学的核心领域之一。

生物信息学有许多不同的定义。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储、信息的内涵和信息的传递，研究和分析生物体细胞、组织及器官的生理、病理、药理过程中的各种生物信息，或者可称为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。一般提到的“生物信息学”就是指这个狭义的概念，更准确地说，应该是**分子生物信息学**(molecular bioinformatics)。

生物信息学是一门交叉学科，它包含了生物信息的获取、处理、存储、分发、分析和解释等在内的所有方面，综合运用数学、计算机科学和生物学的各种工具，以阐明和理解大量数据所包含的生物学意义(见图 1.1)。生物信息学研究的目的是解决生物数据分析和管理中的理论和实践问题，以创建和改进数据库、算法、计算和统计分析技术为研究内容，为理解生物过程提供基础。其特点是计算机技术的集中应用和开发。其研究重点主要体现在**基因组学**(genomics)和**蛋白质组学**(proteomics)两个方面，即从核酸和蛋白质序列出发，分析序列中表达的结构功能的生物信息。生物信息学对于生物学研究具有重要意义，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、

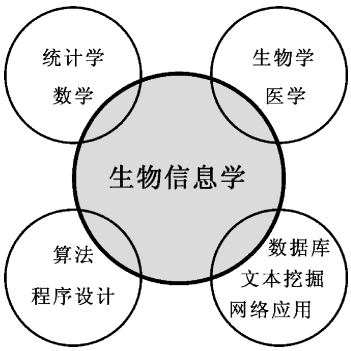


图 1.1 生物信息学涉及多个学科的交叉

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。

1.2　生物信息学的发展历史

生物信息学的发展大致经历了三个阶段。

第一个阶段是前基因组时代。这一阶段以各种算法法则的建立、生物数据库的建立及 DNA 和蛋白质序列分析为主要工作。这一阶段，著名的 Needleman-Wunsch 和 Smith-Waterman 序列比对算法先后发表；国际上的三个核酸序列数据库（EMBL、GenBank 和 DDBJ）相继建立并提供序列服务。

第二个阶段是基因组时代。这一阶段以各种基因组测序计划、网络数据库系统的建立和基因识别为主要工作，以人类基因组计划和各种模式生物基因组测序为代表，大规模测序全面铺开。

第三个阶段是后基因组时代。这一阶段的主要工作是进行大规模基因组分析、蛋白质组分析及其他各种组学研究。随着人类基因组计划和各种基因组计划测序的完成，以及新基因的发现，系统了解基因组内所有基因的生物功能成为后基因组时代的研究重点。生物信息学进入了功能基因组时代。

1.2.1　生物信息学的诞生

生物信息学的基础是分子生物学。因此，要了解生物信息学，首先必须简单了解分子生物学的发展。早在 19 世纪，人们已经知道蛋白质在生命活动中的作用。1883 年，Curtius 首先提出蛋白质线性一级结构的假设。1933 年，Tiselius 首次通过电泳将溶液中的蛋白质分离出来。在 20 世纪 50 年代前后，通过实验已经测定了一些蛋白质的序列，如 1947 年测出短杆菌的五肽结构，1951 年重构胰岛素的 30 个氨基酸。几乎同一时期，科学家认识到 DNA 是遗传物质。1949 年，研究人员发现了 DNA 链中 A＝T 和 G＝C 的规律。1951 年，Pauling 和 Corey 提出蛋白质的 α 螺旋和 β 折叠结构。1953 年 Watson 和 Crick 根据 Franklin 和 Wilkins 得到的 X 射线衍射数据提出 DNA 的双螺旋结构模型，揭开了分子生物学研究的序幕。在其后的 20 年中，科学家们逐步认识了从 DNA 到蛋白质的编码过程，掌握了三联密码子的本质。1961 年，Jacob 和 Monod 发现大肠杆菌的 lac 操纵子中存在调控元件，证实非编码序列并不是垃圾序列。1962 年，Khesin 等发现噬菌体中的基因转录表达具有定时调节机制。20 世纪60 年代出现了通用的核酸测序技术，70 年代中期开始进行基因组规模的测序工作。正是由于分子生物学研究对于生命科学发展的巨大推动作用，生物信息学的出现也成为了一种必然。

早在 20 世纪 50 年代，生物信息学就已经开始孕育。1956 年，在美国田纳西州的加特林堡镇召开了首次“生物学中的信息理论研讨会”。20 世纪 60 年代，一些计算生物学家开始进行相关研究，虽然没有具体地提出生物信息学的概念，但是开展了许多生物信息搜集和分析方面的工作。在这一时期，生物大分子携带信息成为分子生物学的重要理论，生物分子信息在概念上将生物学和计算机科学联系起来。大量的生物分子序列成为丰富的信息源，相关

或者同源蛋白质序列之间的相似性引起了人们的注意。1962 年，Zucherkandl 和 Pauling 研究了序列变化与进化之间的关系，开创了一个新的领域——分子进化。随后，通过序列比对确定序列的功能及序列分类关系，成为序列分析的主要工作。氨基酸序列的收集也是这个时期的一项重要工作，1967 年，Dayhoff 研制出蛋白质序列图集，该图集后来演变为著名的蛋白质信息源 PIR。20 世纪 60 年代是生物信息学形成雏形的阶段。

然而，就生物信息学发展而言，它仍是一门相当年轻的学科。一般认为，生物信息学的真正开端是 20 世纪 70 年代。从 20 世纪 70 年代初期到 80 年代初期，出现了一系列著名的序列比对方法和生物信息分析方法。1970 年，Needleman 和 Wunsch 提出了著名的全局优化算法。同年，Gibbs 和 McIntyre 提出了矩阵打点作图法。Dayhoff 提出的基于点突变模型的 PAM 矩阵是第一个广泛使用的比较氨基酸相似性的打分矩阵，它大幅提高了序列比较算法的性能。1972 年，Gatlin将信息论引入序列分析，证实自然的生物分子序列是高度非随机的。1977 年，出现了将 DNA 序列翻译成蛋白质序列的算法。1975 年，继第一批 RNA(tRNA)序列发表之后，Pipas 和 McMahon 首先提出运用计算机技术预测 RNA 的二级结构。1978 年，Gingeras 等研制出核酸序列中限制性酶切位点的识别软件。这一时期，随着生物化学技术的发展，产生了许多生物分子序列数据，而数学统计方法和计算机技术也得到较快的发展，于是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。他们开始研究生物分子序列，研究如何根据序列推测结构和功能。这时，生物信息学开始崭露头角。

1.2.2 生物信息学的兴起

20 世纪 80 年代以后，出现了一批生物信息服务机构和生物信息数据库。1982 年，核酸数据库 GenBank 第 3 版公开发布。1986 年，日本核酸序列数据库 DDBJ 诞生。1986 年，出现蛋白质数据库 SWISS-PROT。1988 年，美国国家卫生研究所和美国国家图书馆成立国家生物技术信息中心(NCBI)。同年，成立欧洲分子生物学网络(EMBnet)，该网络专门发布各种生物数据库。

20 世纪 90 年代，科学家们开始大规模的基因组研究。1986 年，出现基因组学概念，即研究基因组的作图、测序和分析。1990 年，国际人类基因组计划启动，该计划被誉为生命科学的“阿波罗登月计划”。1995 年，第一个细菌基因组被完全测序。1996 年，酵母基因组被完全测序。1996 年，美国昂飞公司生产出第一块 DNA 芯片。1998 年，第一个多细胞生物——线虫的基因组被完全测序。1999 年，果蝇的基因组被完全测序。2000 年 6 月 24 日，人类基因组计划协作组中 6 个国家的研究机构在全球同一时间宣布已完成人类基因组的工作框架图。与此同时，生物信息学在人类基因组计划的促动之下迅速发展。

2001 年 2 月，人类基因组计划测序工作的完成，使生物信息学走向了一个高潮。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 比特的速度增长，生物信息迅速膨胀成数据的海洋。毫无疑问，人们正从一个积累数据的时代转向一个解释数据的时代，数据量的巨大积累往往蕴含着潜在突破性发现的可能，生物信息学正是在这一前提下产生的交叉学科。当时，生物信息学的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揭示基因组的核苷酸顺序，即全部基因在染色体上的确切位置及各 DNA 片段的功能；同时，在发现新基因信息之后进行蛋

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等实际应用研究。了解基因表达的调控机理也是生物信息学的重要内容，根据生物分子在基因调控中的作用，描述人类疾病的诊断、治疗的内在规律。它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。这时，生物信息学已经成为整个生命科学发展的重要组成部分，成为生命科学研究的前沿。

1.2.3　生物信息学的蓬勃发展

伴随着 21 世纪的到来，生命科学的重点由 20 世纪的实验分析和数据积累转移到数据分析及其指导下的实验验证。分子生物学家使用还原论的方法，将生物系统逐级分解、还原，以理解遗传、进化、发育和疾病等基本过程。但是，这些研究集中在识别基因及认识它们的表达产物的功能，而生物系统的功能蕴藏在系统的整体结构和各种组分的相互作用中，即使知道了所有成分的结构和功能，也不足以解释复杂的生物系统。因此，生物学的研究开始由分解转向为整合。

生物体是由大量结构和功能不同的元件组成的复杂系统，并由这些元件选择性和非线性地相互作用，产生复杂的功能和行为。由于生物体的复杂性和大量过程的非线性动力学特征，需要建立多层次的组学技术平台，研究和鉴别生物体内所有分子及其功能和相互作用。1994 年，澳大利亚麦考瑞大学的 Wilkins 和 Williams 首先提出了蛋白质组(proteome)的概念。蛋白质组学的发展使人们对生物系统中所有蛋白质的组成和相互作用关系有了更深入的了解。之后，出现了一系列组学(omics)，如转录组学、蛋白质组学、代谢组学和相互作用组学等，多组学的高通量方法为研究生物系统提供了大量的数据。数据处理、模型构建和理论分析等算法的发展，则为生物系统模拟提供了强有力的计算工具。在基因组学、蛋白质组学等新型大学科发展的基础上，孕育了系统生物学。系统生物学的主要任务是尽可能地获得每个层次的信息并将它们进行整合，模拟复杂的生物系统行为，解释生物系统背后的运行机制。系统生物学的发展，使生命科学的研究模式发生了深刻变化。它改变了传统生物学研究以小型实验室为基础和“单干”的研究模式，也促进了更大范围和更高层次上的学科交叉和国际合作，如人类基因组计划、人类单倍型图谱计划、人类表观基因组计划等。

生命科学正在经历一个从分析还原思维到系统整合思维的转变。人们所寻求的强有力的数据处理分析工具成为了生命科学研究的关键。同时，以数据处理分析为本质的计算机科学技术和网络技术获得了迅猛的发展，计算机技术和网络技术日益渗透到生命科学的方方面面，崭新的、拥有巨大发展潜力的生物信息学正在坚定而如火如荼地发展和成熟起来。可以说，历史必然性地选择了生物信息学——生命科学与计算科学的融合体作为下一代生命科学研究的重要工具。

1.3　生物信息学的研究内容

在短短十几年间，生物信息学已经形成了多个研究方向(见图 1.2)，以下简要介绍其中的一些研究重点。

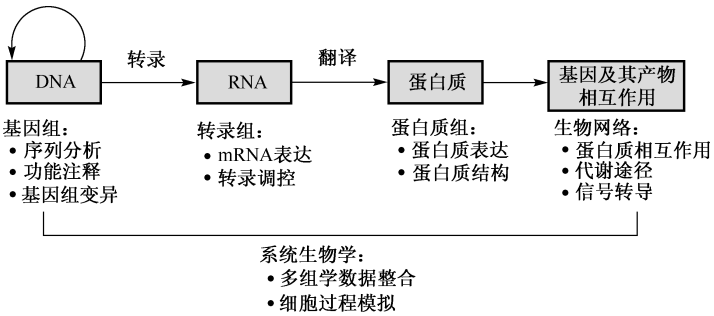


图 1.2 生物信息学的研究内容

1.3.1 基因组学研究

1. 基因组注释

基因组注释(genome annotation)是指利用生物信息学方法和工具，对基因组所有基因的生物学功能进行高通量注释，其研究内容包括基因识别和基因功能注释两个方面。

基因识别的核心是确定全基因组序列中所有基因的确切位置，也可以认为是基因组的结构注释。基因识别的一个常用方法是同源比较，通过两两比对或多序列比对，了解基因家族特性，并预测新基因的功能。例如，对于一个家族中所有相关蛋白的多重序列比对，有助于理解这些蛋白中的系统发育关系，揭示蛋白进化过程。进一步，通过研究多序列比对中高度保守的区域，可以对蛋白质的结构进行预测，并推断这些保守区域对于维持三维结构的重要性。

为了实现自动化的序列比对，研究人员开发了一系列序列比对算法和软件，如双序列比对的 BLAST 和多序列比对的 Clustal。不同的算法得到的比对结果往往不尽相同。当待比对的序列较多时，计算复杂度会大大增加。因此，如何针对特定的问题设计合适的比对算法，并且在计算速度和最佳比对效果之间达到一种平衡，仍是生物信息学要解决的研究课题。

2. 进化生物学

进化生物学研究物种的起源和演化，基因组测序获得的海量数据为从分子水平研究进化论提供了数据基础，从而大大推进了进化生物学的发展。利用生物信息学方法研究进化生物学的优势在于：通过度量 DNA 序列的改变，可以研究众多生物体、生物物种之间的进化关系；通过整个基因组的比对，能够研究更为复杂的进化论课题，如基因复制、基因横向迁移等；能够为种群进化建立复杂的计算模型，以便预测种群随时间的演化；同时，保存了大量物种的遗传信息。

3. 基因组变异

遗传信息变异是所有基因组的共同特征。不同个体、群体在疾病易感性、对环境致病因子的反应性和其他性状上的差别，都与基因组序列中的变异有关。在最低的层次上，单个核苷酸位点发生了点变异，就形成了通常所说的单核苷酸多态性。发现单核苷酸多态性位点并构建其相关数据库，是基因组研究走向应用的重要步骤。在较高的层次上，大的染色体片段经历了复制、横向迁移、逆转、调换、删除和插入等过程。在最高的层次上，整个基因组会经历杂交、倍交、内共生等变异，并迅速产生新的物种。

研究人类基因组变异是理解群体和个体间疾病易感性和其他生物学性状差异的遗传学基础，有助于了解基因变异与性状的关系，发现基因与疾病易感性之间的关联，从而预测发病风险，发展基于群体和个体遗传学特点的医学。而基因组变异的发现、基因组差异性的比较，以及单个核苷酸多态性位点与疾病易感性的关联分析，都需要生物信息学方法的支持。

1.3.2　转录组数据分析

1. 基因表达数据的分析与处理

转录组学研究细胞在某一功能状态下所有基因的表达情况，是了解生命活动动态的重要手段。通过对大规模基因表达数据的分析和处理，可以了解基因表达的时空规律，探索基因的功能和表达调控网络，提供疾病发病机理的信息。目前，已有多种生物学技术可以用于测量基因的表达，如 DNA 微阵列、基因表达序列分析、大规模平行信号测序等。不同于以往的少数几个生物分子信息的数据处理，现在通过转录组学技术通常可产出成千上万个基因的表达数据，数据处理量大幅度增加，数据之间的关系也更加复杂。因此，对于高维数、高噪声、强耦合的基因表达数据的分析和处理方法，成为生物信息学发展的一个重要方向。

目前，用于基因表达数据处理的方法主要包括相关分析、降维方法、聚类分析和判别分析等。通过主成分分析等降维方法，可以在多维数据集合中确定关键变量的特点，分析在不同条件下基因响应的规律和特征。聚类分析则将表达模式相似的基因聚为一类，在此基础上寻找相关基因，分析基因的功能。虽然聚类方法是基因表达数据分析的基础，但是此类方法只能找出基因之间简单的线性关系，要发现基因之间复杂的非线性关系则需要发展新的分析方法。

2. 基因表达调控分析

基因表达调控是指当细胞受到外信号刺激之后，其内部发生的一系列反应过程。生物信息技术可以用于分析基因表达调控的各个步骤。对于一个生物体，人们可以用生物芯片技术观察细胞在不同外界刺激、不同细胞周期或不同状态下的响应情况，并利用聚类算法分析这些基因表达数据，以寻找表达相似的基因或样本，了解基因的转录调控模式。进一步，还可以探索基因的转录调控网络，发现基因在环境或药物作用下表达模式的变化，阐明各基因之间的调节作用。

在基因调控网络分析方面，研究人员已经开展了大量有意义的工作，建立了一系列基因调控网络的数学模型，如布尔网络模型、线性关系网络模型、微分方程模型、互信息相关网络模型等。在此基础上，还研究了部分基因调控网络的动力学性质。但是由于问题的复杂性，目前还只能构建小规模的基因调控网络，对于预测网络的可靠性也缺乏有效的评估。如何整合更多的生物学证据，构建大规模、精确的基因调控网络是生物信息学研究的一个重要课题。

1.3.3　蛋白质组学分析

1. 蛋白质组学表达分析

基因组对生命体的整体控制必须通过它所表达的全部蛋白质来执行。由于基因芯片技术只能反映从基因组到 RNA 的转录水平的表达情况，而从 RNA 到蛋白质还要经历许多中间环节，因此仅凭基因芯片技术还不能揭示生物功能的具体执行者——蛋白质的整体表达

状况。为了测量基因组所有蛋白质产物的表达水平，研究人员发展出一系列蛋白质组学技术，主要包括二维凝胶电泳技术和质谱技术。通过二维凝胶电泳技术可以获得某一时间截面上蛋白质组的表达情况，通过质谱技术则可以得到所有蛋白质的序列组成。质谱技术往往能够产出海量的蛋白质表达数据，而对这些数据的分析和利用则要借助于生物信息学方法，如通过搜索数据库的方法鉴定蛋白质组分，对每种蛋白质的多少进行定量研究，通过质量控制的方法提高数据的可信性。这就涉及大量的统计分析和数据处理工作，并且导致了更多新的问题涌现，例如如何有效地存储海量的质谱数据？如何快速地进行蛋白质鉴定和定量分析？如何提高质谱数据的质量和覆盖度？解答这些问题都有待生物信息学方法的进一步发展。

2.蛋白质功能与结构预测

随着基因组和蛋白质组研究的开展，许多新蛋白的序列得以揭示，但是要想了解它们的功能，只有一级结构——氨基酸序列还远远不够。蛋白质通过其三维结构来执行功能，而且蛋白质的三维结构通常是动态的，在行使功能的过程中其结构会发生相应的改变。因此，获得这些新蛋白的完整、精确和动态的三维结构，就成为摆在人们面前的紧迫任务。目前，除了通过 X 射线衍射晶体结构分析、多维核磁共振波谱分析和扫描电子显微镜二维晶体及三维重构等实验技术得到蛋白质三维结构，通过生物信息学方法预测蛋白质结构是一种非常重要的研究手段。

用于蛋白质高级结构预测的方法大多为启发式方法，其中最常用的是同源建模技术。同源性是生物信息学中的一个重要概念。在基因组的研究中，同源性被用于分析基因的功能：若两个基因同源，则它们的功能可能相近；在蛋白质结构的研究中，同源性被用于寻找在形成蛋白质结构和蛋白质反应中起关键作用的蛋白质片段。利用同源建模的技术，可以从蛋白质的已知结构预测与其同源的蛋白质的三维结构。目前，蛋白质结构预测方法的总体准确率不高，而且计算比较复杂，其改进一方面依赖于蛋白质结构稳定性相关理论研究的深入，另一方面也有待计算方法的进一步发展。

1.3.4 生物网络分析

近年来，各种生物网络理论的研究及通过构建生物网络进行基因功能挖掘的研究，正逐渐成为生物信息学领域的研究热点。要了解细胞的整体状态，就必须依据人们的现有知识去重新构建复杂的生物学网络并进行相关分析，在基因组水平上阐释基因的活动规律。这从根本上改变了传统生物学的思维方式，形成了一种新的全局方法。

1.蛋白质-蛋白质相互作用的研究

蛋白质之间的相互作用存在于生物体每个细胞的生命活动过程中，它们相互交叉形成网络，构成细胞中的一系列重要生理活动的基础。研究蛋白质之间相互作用的方式和程度，将有助于蛋白质功能的分析、疾病致病机理的阐明和治疗。因此，确定蛋白质之间相互作用关系并绘制相互作用图谱已成为蛋白质组学研究的热点。近年来，随着蛋白质组学研究技术的不断发展，蛋白质之间相互作用研究的新方法不断出现，除了常用的免疫共沉淀、酵母双杂交、噬菌体展示、荧光共振能量转移等技术，一些全新的方法及对原有技术的改进方法也不断涌现。随着技术的进步，研究人员已经发现了很多大规模的蛋白质相互作用数据集，

但它们还存在假阳性较高、覆盖度不够等问题，仍有大量的蛋白质相互作用没有被揭示。而生物信息学方法综合蛋白质之间的同源性、蛋白质的序列特征、结构特征及基因表达关联等多种生物学证据，既可以对蛋白质相互作用进行可靠性验证，也可以对未知的蛋白质相互作用进行挖掘。

2. 生物网络的构建与分析

生物网络的构建主要包括两个方面：一方面是构建代谢和调控网络，如 KEGG 数据库已经整理了跨物种的代谢网络图，并在积极完善各种调控网络图；另一方面是构建基因表达调控网络。基因表达存在组织特异性、细胞周期特异性和外界信号的影响特异性，这些特异性都是由细胞内复杂而有序的调控机制来实现的。基因表达数据的研究为构建复杂的表达调控网络提供了基础。蛋白质-蛋白质相互作用、蛋白质- DNA 相互作用等数据，则可用于构建大规模的分子相互作用网络。进一步，有必要整合各种网络信息与已有的生物学知识，从整体网络结构来研究基因及其产物的相互作用，提取基因的功能信息，这种研究思路更符合细胞的生命本质。

对于已构建的生物网络，则可借助于图论等网络分析方法对网络属性进行研究。目前，已发现生物网络具有无尺度性质、小世界属性、高聚集性和鲁棒性等，它们有利于保持生物学重要功能的稳定。同时，研究人员已着手研究与条件相关的动态生物网络，以便更深入地揭示生物网络内部的运行规律。

1.3.5　系统生物学研究

传统生物学独立地检测单个基因或者蛋白质。与之不同的是，系统生物学同时研究多个水平上的生物信息(DNA、mRNA、蛋白质、蛋白质复合体、生物通路及生物网络)之间复杂的相互作用，从而理解它们如何共同发挥作用。

图 1.3 给出了系统生物学研究的两种典型策略。由分子生物学实验和生物信息学获得的分子属性是构建各种网络模型的基础。图中给出了系统生物学中常用的 3 种模型：计量模型、调控模型和动力学模型。自底向上的系统生物学(bottom-up systems biology)从分子属性出发来构建模型以预测系统属性，并进行实验验证和模型修正。相反，自顶向下的系统生物学(top-down systems biology)是系统数据驱动的。从实验数据去发现和提炼现有的模型，使之能够更好地描述实验数据。通过这种方式，可以识别未知的分子相互作用和机制。经典的自底向上的系统生物学多采用动力学模型，而自顶向下的系统生物学多采用调控模型来分析数据。网络中各结点代表酶、调控因子或代谢物，实线代表化学反应，虚线代表调控关系。

1. 生物系统的建模与仿真

系统生物学研究的一个主要任务是生物系统的建模与仿真。其目标是：在已知的生物学知识和定量数据的基础上，利用各种建模工具建立生物系统的描述模型，以尽可能精确地模拟系统的行为。进一步，基于分子网络的定量描述模型，可以进行细胞过程的模拟，动态观测细胞中各种分子随时间和空间的改变，研究生物系统的运行机制，并预测其在各种刺激下可能的响应情况。例如，虚拟细胞是指通过数学计算和分析，对细胞的结构和功能进行分析、整合和应用，模拟和再现细胞的生命现象。通过该项研究，有望从单个细胞开始，建立一个能够模拟人体系统运行过程中所有生化反应的虚拟人体。

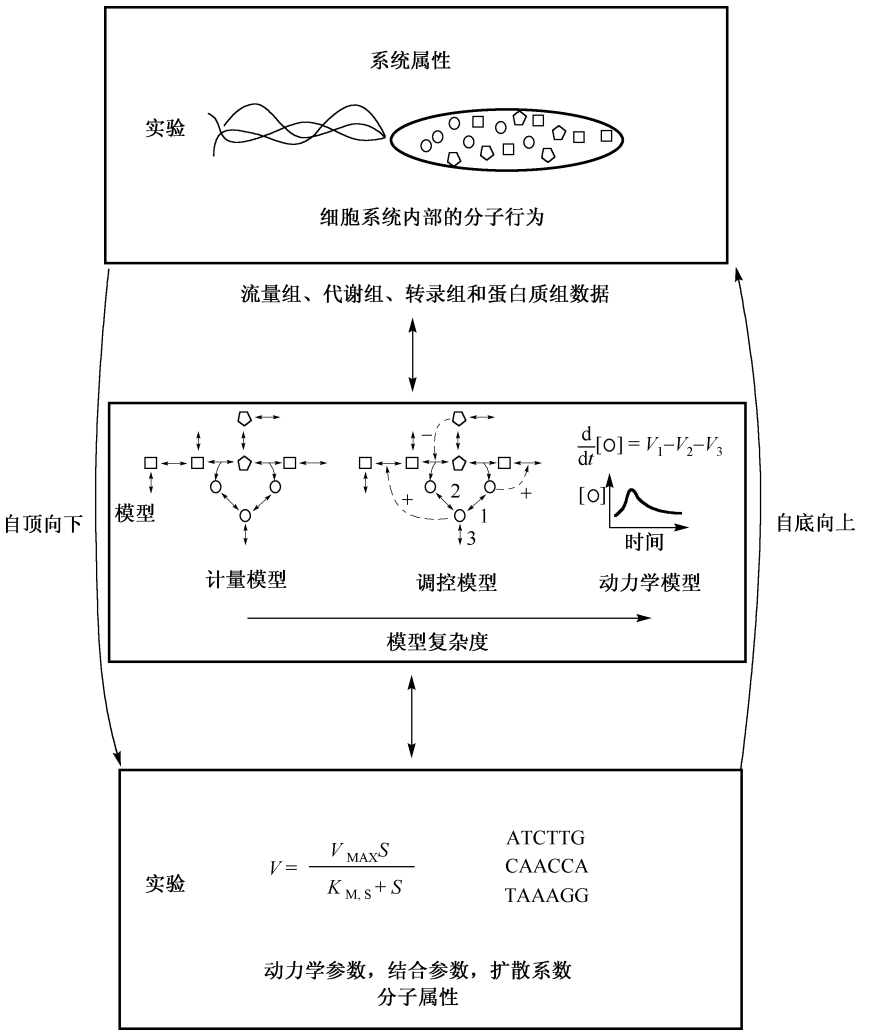


图 1.3 系统生物学的自底向下和自底向上的研究策略[11]

2. 多组学数据的整合

由于实验技术的全面发展，获取高通量的组学数据变得更加容易和成本低廉，它们提供了细胞中几乎所有成员和相互作用的综合描述。这些组学数据之间既相互关联又各有侧重，如何综合分析多组学数据，根据组学数据之间的相似性和互补性，挖掘生物过程的新观点，成为系统生物学领域的重要课题。组学数据整合就是要对来自不同组学的数据源进行归一化处理、比较分析，建立不同组学数据之间的关系，综合多组学数据对生物过程进行全面深入的阐释。组学数据整合的任务可以归纳为如下 3 个层次：

① 对两个组学数据之间进行比较分析，挖掘数据之间的相关性和差异性；

② 给定三个或多个组学数据，挖掘它们之间的内在关系；

③ 针对现有的所有组学数据，发展通用的数据整合方法和软件，进行大规模的、系统的数据整合。

1.3.6　医学相关研究

1. 药物靶标筛选与验证

现代药物研发通常以药物靶标为基础进行有针对性的药物设计，而生物信息学为药物靶标基因的发现和验证提供了有力的工具。目前，人们已经构建了多种数据库用于存储疾病相关的生物信息，通过分析不同组织在正常/疾病状态下的基因表达的差异，可以获得疾病特异的药物靶标。另外，还可以根据蛋白质功能区和三维结构预测对药物靶标进行鉴定，以便了解所研究蛋白质的属性，预测其是否适用于药物作用。

2. 基于结构的药物设计

合理药物设计的目标是：依据药物发现过程所揭示的药物作用靶标，即受体，参考其内源性配基和天然药物的化学结构特征，寻找和设计合理的药物分子，以发现既能选择性地作用于靶标，又具有药理活性的先导化合物。药物设计中最基本的原理是“锁和钥匙”原理，即药物在体内与特定的靶标作用，并引起靶标分子的结构和功能的变化。利用生物信息学方法可以进行计算机辅助的药物设计，开发多种药物设计工具。

实际上，生物信息学的研究内容远不止于此，随着更多实验数据的产出和生物学理论的发展，生物信息学的研究范畴还在不断扩展。其总体任务是：运用数学理论成果对生物体进行完整、系统的数学模型描述，使人类能够从一个更明确的角度和一个更易于操作的途径来认识和控制自身及其他所有生命体。

1.4　生物信息学的研究资源

1.4.1　研究机构

1. 国际著名的生物信息中心

表 1.1 列出了国际上比较著名的生物信息学研究机构，其中最著名的是**美国国家生物技术信息中心**(National Center for Biotechnology Information，NCBI)，由美国国立医学图书馆于 1988 年 11 月 4 日建立。NCBI 下属的不仅有分子生物学数据库，还有相关的检索系统和工具。

表 1.1　国际上比较著名的生物信息学研究机构

机　构	所在国家	全　称	网　址
NCBI	美国	National Center for Biotechnology Information	http://www.ncbi.nlm.nih.gov/
EBI	欧洲	European Bioinformatics Institute	http://www.ebi.ac.uk/
HGMP	英国	Human Genome Mapping Project Resource Centre	http://www.hgmp.mrc.ac.uk/
ExPASy	瑞士	Expert of Protein Analysis System	http://au.expasy.org/
CMBI	荷兰	Centre of Molecular and Biomolecule	http://www2.cmbi.ru.nl/
ANGIS	澳大利亚	National Genome Information Service	http://www.angis.org.au/
NIG	日本	National Institute of Genetics	http://www.nig.ac.jp/english/index.html
BIC	新加坡	National Bioinformatics Centre	http://www.bic.nus.edu.sg

NCBI 的下属数据库包括：GenBank 数据库(Nucleotide)、三维蛋白质结构的分子模型数据库(MMDB)、在线人类孟德尔遗传数据库(OMIM)、生物门类数据库(Toxonomy)和文献数据库(Pubmed)。NCBI 的检索系统有两个体系，一个是 Entrez 数据库检索系统，可以查询核酸序列、蛋白质序列、蛋白质三维结构、种系序列数据及文献数据等，另一个是 BLAST(Basic Local Alignment Search Tool)相似性检索系统，提供序列比对等工具。

2. 国内部分生物信息学和生物医学信息服务器

国内也越来越重视生物信息学。在一些院士和教授的带领下，许多研究团队在各自领域取得了一定成绩，并在国际上占有一席之地。表 1.2 列出了国内比较著名的生物信息学研究中心，如北京大学的罗静初和顾孝诚教授在生物信息学网站建设方面，中科院生物物理所的陈润生院士在表达序列标签拼接方面及基因组演化方面，天津大学的张春霆院士在 DNA 序列的几何学分析方面，以及中科院理论物理所郝柏林院士、清华大学的李衍达院士和孙之荣教授、内蒙古大学的罗辽复教授、上海的丁达夫教授等，都做出了卓有成效的工作。北京大学于 1997 年 3 月成立了生物信息学中心，中科院上海生命科学研究院也于 2000 年 3 月成立了生物信息学中心，分别维护着国内两个专业水平相对较高的生物信息学网站。

表 1.2 国内比较著名的生物信息学研究中心

机　构	网　址
北京大学生物信息中心	http://www.cbi.pku.edu.cn
中国生物医学大数据中心	http://www.biosino.org/
北京大学化学与分子工程学院	http://www.chem.pku.edu.cn
北京大学医学信息学中心	http://medic.bjmu.edu.cn
中国科学院微生物研究所	http://www.im.cas.cn
天津大学生物信息中心	http://tubic.tju.edu.cn
中国科学院计算技术研究所前瞻研究实验室生物信息学研究组	http://www.bioinfo.org.cn/
华大基因	http://www.genomics.cn/

还有一些专门的生物类网站和论坛，包含了生物信息学各方面的资源和软件，如生物谷、丁香通、生物秀等。

1.4.2 数据库

数据库是生物信息学研究的重要基础，各种数据库几乎覆盖了生物科学的各领域。随着人类基因组计划的完成和多种组学研究的开展，已积累海量的生物信息，并以不同组织形式构成许多数据库。国际上已建立许多公共生物分子数据库，大部分数据库是公开和免费的，并可通过互联网访问。随着研究的深入，公共数据库越来越成为世界各地生物学家的重要给养。1994 年起，国际知名期刊 *Nucleic Acid Research*(核酸研究)将每年的第一期刊物作为分子生物学数据库专刊，专门综述当前的在线分子生物学数据库资源(http://nar.oxfordjournals.org/)。

按照构建方式，数据库可分为一级数据库和二级数据库。一级数据库要求数据库中至少有一项信息来自直接的实验数据，通常收录生物大分子序列和结构，提供相关注释信息，内容比较全面、稳定，有持续更新。国际上著名的一级核酸数据库有 Genbank 数据库、EMBL 核酸库和 DDBJ 库等；蛋白质序列数据库有 SWISS-PROT 和 PIR 等；蛋白质结构库有 PDB 等。而二级数据库是在一级数据库、实验数据和理论分析的基础上针对特定目标衍生而来的，是对生物学知识和信息的进一步整理。根据不同的构建方法，二级数据库包括：

① 文献挖掘数据库，如 PubMeth 和癌症甲基化数据库等；

② 对多个数据库进行整合得到的数据库，如 International protein index；

③ 由预测、建模工具整理得到的实验数据集，如 PSORTb。

目前，已建立的二级生物学数据库非常多，它们因针对不同的研究内容和需要而各具特色，如人类基因组图谱库 GDB、转录因子和结合位点库 TRANSFAC、蛋白质结构家族分类库 SCOP 等。

按照包含的内容，数据库可以分为核酸数据库、RNA 数据库、蛋白质数据库和生物通路数据库等。蛋白质数据库还可以细分为蛋白质序列数据库、蛋白质组学数据库、蛋白质序列模体数据库和蛋白质结构数据库等。这些数据库由专门的机构建立和维护，负责收集、组织、管理和发布生物分子数据，并提供数据检索和分析工具，向生物学研究人员提供大量有用的信息，最大限度地满足研究和应用的需要。

1. 核酸数据库

基因组数据量非常庞大，有组织地收集和管理这些数据是开展各项研究工作的前提。为了便于研究人员共享这些数据，及时得到最新的实验数据结果，也为了保证基因组数据的一致性和完整性，世界各国政府相继建立了专门的机构来搜索和管理这些数据，还有一些企业提供商业的生物信息服务。其中最权威的三大国际核酸数据库为 GenBank、EMBL 和 DDBJ。1979 年，美国洛斯阿拉莫斯国家实验室开通了基因库 GenBank，它包含了已知核酸序列和蛋白质序列，以及相关文献著作和生物学注释。现在 GenBank 改由美国国家生物技术信息中心(NCBI)管理维护。1982 年，欧洲分子生物学实验室建立了 EMBL 数据库，并随后建立了欧洲生物网(EMBNet)，1994 年后该数据库改由欧洲生物信息学研究所(EBI)管理。1984 年，日本着手建立国家级的核酸数据库 DDBJ，1987 年正式开始服务。目前，绝大部分核酸和蛋白质数据是在美国、欧洲和日本产生的。为了保证数据的完整性，以上三方共同组成了 DDBJ/EMBL/GeneBank 国际核酸序列数据库。根据数据交换协议，三大数据库包含的数据内容基本一致，仅在数据格式上略有区别。

此外，还有一些专门的模式生物基因组数据库(见表 1.3)，如线虫基因组数据库 AceDB、酿酒酵母基因组数据库 SGD 等。这些数据库除了收录基因组数据资源，还收录分子生物学及遗传学等大量信息，为相关研究提供了共享和交流信息的平台。

表 1.3　模式生物基因组数据库

数　据　库	网　　址
UCSC Genome Browser	http://genome.ucsc.edu/
Ensembl Genome Browser	http://www.ensembl.org/
NCBI Map Viewer	http://www.ncbi.nlm.nih.gov/mapview
大肠杆菌基因组数据库 EcoGene	http://ecogene.org/
酿酒酵母基因组数据库 SGD	http://www.yeastgenome.org/
疟原虫基因组数据库 PlasmoDB	http://plasmodb.org/
线虫基因组数据库 AceDB	http://www.acedb.org/
果蝇基因组数据库 FlyBase	http://www.fruitfly.org
斑马鱼基因组数据库 ZFIN	http://zfin.org/
小鼠基因组数据库 MGI	http://www.informatics.jax.org
拟南芥基因组数据库 TAIR	http://www.arabidopsis.org
水稻基因组数据库 BGI-RIS	http://rice2.genomics.org.cn

2. 蛋白质数据库

与蛋白质相关的数据库集中了蛋白质的各种格式化的知识，除了文献描述，数据库成为了主要的知识表示、存储和交换来源。这里以蛋白质的不同属性为分类标准，介绍蛋白质相关的数据库资源。蛋白质相关数据库的常见类型如图 1.4 所示。

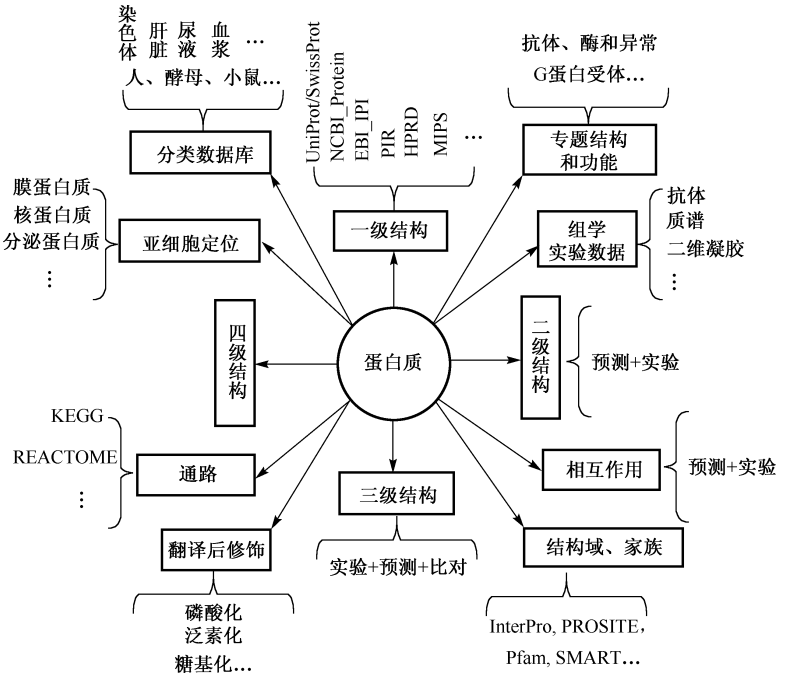


图 1.4 蛋白质相关数据库

1)蛋白质序列数据库

① SWISS-PROT/TrEMBL。SWISS-PROT(http://www.expasy.ch/sprot/)由瑞士生物信息学研究所(SIB)和欧洲生物信息学研究所(EBI)共同维护。与同类数据库相比，SWISS-PROT 是高度注释的(包括蛋白质功能描述、结构域信息、转录后修饰、变异等)，冗余程度最低，与其他数据库的整合程度最高。TrEMBL 是 SWISS-PROT 的补充，包含所有的 EMBL 核苷酸的翻译产物，采用与 SWISS-PROT 库完全一致的格式。但由于 TrEMBL 是经计算机翻译所得的，序列错误率较高且存在较大的冗余度，因此未整合进 SWISS-PROT。

② PIR。PIR(Protein Information Resource，http://pir.georgetown.edu/)是一个应用较为广泛的、经注释的、非冗余蛋白质序列数据库。

③ NCBInr。NCBInr(http://www.ncbi.nlm.nih.gov)是一个非冗余的蛋白质数据库，它由 NCBI 搜集并建立，以供搜索工具 BLAST 和 Entrez 所用。

④ OWL。OWL(Composite Protein Sequence Database，混合蛋白质数据库，http://www.bioinf.man.ac.uk/dbbrowser/OWL/index.php)是一个非冗余蛋白质序列数据库，由 4 个公用的一级资源组成：SWISS-PROT、PIR、Genbank 和 NRL-3D。

2)蛋白质组数据库

① AAindex(氨基酸索引数据库，http://www.genome.ad.jp/dbget/)

② GELBANK

③ Predictome

④ Proteome Analysis Database

⑤ REBASE(http://rebase.neb.com/rebase/rebase.html)

⑥ SWISS-2DPAGE(http://www.expasy.org/ch2d/)

⑦ YPL.db

3)蛋白质序列模体数据库

① Blocks(http://blocks.fhcrc.org)

② CDD

③ CluSTr

④ InterPro(http://www.ebi.ac.uk/interpro/)

⑤ Pfam(http://xfam.org)

⑥ PROSITE(http://www.expasy.org/prosite)

4)蛋白质二级结构数据库

① DSSP。蛋白质二级结构数据库 DSSP(Database of Secondary Structure of Protein，http://swift.cmbi.ru.nl/gv/dssp/)是一个关于蛋白质二级结构归属的数据库。

② PredictProtein。PredictProtein(http://www.predictprotein.org)是蛋白质结构预测服务器，可根据要求的方法对所提交的蛋白质序列给出蛋白质多重序列比对结果，预测二级结构、残基可溶性、跨膜螺旋位置、折叠拓扑类型等。

③ SCOP。蛋白质结构分类数据库 SCOP(http://scop.mrc-lmb.cam.ac.uk/scop/)详细描述了已知的蛋白质结构之间的关系。该数据库基于若干层次对结构进行分类，包括家族(描述相近的进化关系，要求归属的蛋白质序列相似度大于 30%)、超家族(描述远源的进化关系，即具有相似的结构和功能，但序列相似性较低的一类蛋白质)和折叠类(包括全 α、全 β、α/β、α+β 和多结构域等，用于描述二级结构单元的排列及拓扑结构)。

5)蛋白质三维结构和相关数据库

① PDB 数据库。PDB(http://www.rcsb.org/pdb/)由美国 Brookhaven 国家实验室建立，是国际上重要的生物大分子结构数据库。该数据库收录了通过 X 射线衍射晶体结构分析、核磁共振等实验手段测定的生物大分子的三维结构，主要是蛋白质的三维结构，也包括了部分核酸、糖类、蛋白质与核酸复合体三维结构。

PDB 中的每条记录包含显式序列(explicit sequence)和隐式序列(implicit sequence)信息。隐式序列即为立体化学数据，包括每个原子的名称和原子的三维坐标。由于 PDB 的主要信息是三维结构，如果直接将三维结构信息以文本形式返回给用户，那么用户将难以读懂这些信息。实用的方法是通过分子模型软件，以图形方式显示三维结构。互联网上有许多可以利用的分子模型软件，如 RasMol、CHIME 和 MolPOV 等，这些软件能够以各种模型显示出生物大分子的三维结构，如结构骨架模型、棒状模型、球棒模型、空间填充模型和带状模型等。此外，PDB 还说明了蛋白质某些特定部位的二级结构类型，如 α 螺旋和 β 折叠等。

② CPHmodels。CPHmodels(http://www.cbs.dtu.dk/services/CPHmodels/)是采用同源建模来预测蛋白质三级结构的一个网络服务器，同时也采用了以预测距离为基础的线程(threading)算法。

③ MMDB 数据库。MMDB(http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml)数据库收录了所有经实验测定的蛋白质三维结构。

3. 基因表达数据库

目前,收集和存储基因表达数据的最有影响的数据库是微阵列数据仓库(GEO)、微阵列公共知识库(ArrayExpress)和斯坦福微阵列数据库(SMD)。

1)微阵列数据仓库 GEO

GEO(http://www.ncbi.nlm.nih.gov/geo)是由 NCBI 于 2000 年开发的基因表达和杂交芯片数据仓库,提供了来自不同物种的基因表达数据的在线资源。截至 2017 年 12 月,GEO 数据库中已存储了 4348 个数据集,包括 17 951 个平台(platform)、2 300 211 个样本(sample)和 92 421 个系列(series)。其中,平台是关于物理反应物的信息;样本是关于待检测的样本信息和使用单个平台产生的数据;系列是关于样本集的信息,反映样本间的相关性和组织。

2)微阵列公共知识库 ArrayExpress

ArrayExpress(http://www.ebi.ac.uk/arrayexpress/)是基于基因表达数据的芯片公共知识库,包含多个基因表达数据集和与实验相关的原始图像。ArrayExpress提供一个简单的基于网页的数据查询界面,并直接与 Expression Profiler 数据分析工具相连,可以表达数据聚类和其他类型的网页数据挖掘。另外,ArrayExpress 中的数据可与所有由 EBI 维护的在线数据库相链接,方便进行交叉查询和注释分析。

3)斯坦福微阵列数据库 SMD

SMD(http://smd.princeton.edu/)是一个使用 Oracle 作为管理软件的关系数据库。该数据库存储基因芯片实验的原始数据、归一化数据和对应的图像文件,另外还提供数据获取、分析和可视化的界面,包括层次聚类、自组织映射和缺失值归纳等方法。

除了以上 3 个综合性的基因表达数据仓库,还有如下一些专门的基因表达数据库:

- YMD (Yale Microarray Database, http://medicine.yale.edu/keck/ymd)
- ArrayDB
- BodyMap
- ExpressDB
- HuGE Index(Human Gene Expression Index)

这些数据库收集的数据往往具有物种特异性,使用比较方便。

4. 生物通路数据库

Google 索引的在线资源中心 Pathguide(http://www.pathguide.org/)介绍了 702 个途径和相互作用数据库的概况(截至 2018 年 3 月),提供了大部分生物学通路的索引。该网站收录了蛋白质-蛋白质相互作用、代谢途径、信号转导通路、表达途径、转录因子/基因调控网络、蛋白-复合体相互作用、基因相互作用网络等类型的数据库,可查看其基本表示格式和是否免费开放等信息,以帮助用户选择合适的数据库进行数据搜索。下面对其中常用的一些生物通路数据库进行介绍。

1)蛋白质相互作用数据库

随着高通量的蛋白质相互作用检测技术的发展，已经揭示出了多种模式生物，包括酵母、线虫、果蝇和人的大规模相互作用网络。例如，2005 年，*Nature*和 *Cell* 期刊上分别发表文章，报道了人的大规模蛋白质相互作用数据集，分别包括 2800 对和 3186 对相互作用。通过分析和比较，可以发现这两组数据集的一些特点。两个数据集都由酵母双杂交方法获得，再经过独立实验验证，以保证其可靠性。但是两组数据交叉很少，原因可能是这两个数据集中存在大量的假阴性，也可能是相互作用的数据规模远比这两个数据集的规模大得多。

同时，采用生物信息学预测方法也得到了大量的蛋白质相互作用数据，其规模大概是由实验方法产出的相互作用数据的两倍左右。因此，人们发展了多个数据库，收集并整理相关的蛋白质相互作用数据(见表 1.4)，以下简单介绍其中 3 种。

DIP 数据库收集了经实验验证的蛋白质相互作用数据。数据库包括 3 个部分：蛋白质信息、相互作用信息和检测相互作用的实验技术。用户可以根据蛋白质、生物物种、蛋白质超家族、关键词、实验技术或引用文献来查询 DIP 数据库。

BIND 全称为 Biomolecular Interaction Network Database，即生物分子相互作用网络数据库，主要提供蛋白相互作用信息，现已整合到 BOND 数据库中。

Pathway Commons 数据库(http://www.pathwaycommons.org/)是一个蛋白质相互作用的整合数据库，目前已整合的数据源包括 BioGRID、Cancer Cell Map、HPRD、HumanCyc、IMID、IntAct、MINT 和 NCI/Nature Pathway Interaction Database。

表 1.4　常用的蛋白质相互作用数据库[16]

数据库	网　　址	实验手段	预测方法	数据验证	描　　述
DIP	http://dip.doe-mbi.ucla.edu	+	−	+	收集由实验确定的蛋白质相互作用
MINT	http://mint.bio.uniroma2.it/	+	−	−	从文献中提取经实验检测获得的蛋白质相互作用
MIPS	http://mips.gsf.de	+	−	−	酵母特异，同时包括基因相互作用信息
BioGRID	http://thebiogrid.org/	+	−	−	对来自 BIND、MIPS 等多个基因组规模的数据集进行编辑，酵母特异
STRING	http://string-db.org/	−	+	−	基于基因邻接、系统发育谱和结构域融合方法预测的蛋白质相互作用

2)代谢途径数据库

通用型代谢途径数据库以统一的数据格式记录已知的代谢相关信息，适合作为非物种特异相关研究的代谢数据来源。目前，常用的通用型综合数据库包括：由日本京都大学生物信息学中心开发和维护的京都基因和基因组百科全书 KEGG，由斯坦福国际生物信息研究小组开发和维护的通路/基因组数据库 BioCyc 及代谢通路百科全书 MetaCyc，由韩国科学与技术高级研究所开发的整合了 KEGG 和 BioCyc 的数据库系统 BioSilico 等。表 1.5 列出了常用的综合代谢数据库的相关信息。

表 1.5 常用的综合代谢数据库

数据库	网 址	主要收录的数据信息
KEGG	http://www.genome.ad.jp/kegg	代谢通路图谱、酶列表、化合物列表、基因组图谱和注释、基因的同源性与生物系统功能的层次关系等
BioCyc	http://biocyc.org/	分物种存储的通路和基因组数据，包括多个模式生物的数据库，如 EcoCyc、AraCyc 和 YeastCyc
MetaCyc	http://metacyc.org/	多个物种的综合代谢通路信息，包括化合物、基因、酶及酶促反应等
BioSilico	http://biosilico.kaist.ac.kr/	整合多个数据库的酶、化合物和代谢通路信息，提供对多个代谢数据库的访问

这里特别介绍一下京都基因和基因组百科全书(Kyoto Encyclopedia of Gene and Genomes，KEGG)。它是系统分析基因功能、联系基因组信息和功能信息的知识库，由基因(GENES)、通路(PATHWAY)和配体(LIGAND)三个子库组成。基因组信息存储在 GENES 数据库里，包括完整和部分测序的基因组序列；更高级的功能信息存储在 PATHWAY 数据库里，包括图解的细胞生化过程(如代谢、膜转运、信号转导和细胞周期)，还包括同系保守的子通路等信息；LIGAND 库包含了关于化学物质、酶分子和酶反应等信息。KEGG 提供了 Java 的图形工具来访问基因组图谱，可以比较基因组图谱和操作表达图谱，还免费提供了其他序列比较、图形比较和通路计算的工具。

3)信号转导通路数据库

信号转导通路数据库的资源非常丰富，常用的数据库包括 Biocarta、KEGG 和 Reactome 等。其中，Biocarta 是目前覆盖范围最广的信号转导通路数据库，包含了大量的通路细节知识，方便进行单个分子的查询，但是单个通路规模较小，不提供批量下载。KEGG 和 Reactome 作为经典的信号转导通路数据库，建立时间较早，图示清楚，下载方便，但与 Biocarta 相比包含的通路数据不够全面。STKE 数据库由通路专家收集整理，包括通用的细胞信号数据和部分组织细胞中特殊的信号过程，其内容较为详细，但是包含的通路数目较少。AfCS(The Alliance for Cellular Signalling)数据库以信号分子为基础，提供其参与的相互作用及信号通路图，包含了细胞信号转导联军项目最新的研究成果。Pathway Interaction Database(PID)专门收集人的信号转导通路，包含了大量由文献挖掘得到的信号转导通路，并从 Biocarta 和 Reactome 中导入了大部分信号转导通路，适于人的信号转导通路分析。此外，AMAZE 数据库提供了一个面向对象的平台，整合来自于代谢、细胞信号和基因调控通路的生物条目和相互作用信息。

尽管上述数据库包含了大量有关生物通路的有用知识，但它们都是以静态的连接图形式来描述通路的，难以实现信号转导通路的定量分析。因此，部分研究人员收集了小规模的定量实验结果，构建了一些定量信号转导通路数据库，如 DOQCS 和 SigPath 等。DOQCS 数据库专门收集具有定量信息的信号转导通路，包括反应方程、底物浓度和速率常数等，并对这些模型提供了注释信息。表 1.6 列出了常用的定性和定量信号转导通路数据库的网址和简单描述。

表 1.6　常用的信号转导通路数据库

数据库	网　　址	描　　述
Biocarta	http://www.biocarta.com	信号转导通路图片及注释数据库
KEGG	http://www.genome.ad.jp/kegg	各种细胞过程中分子相互作用的图表
Reactome	http://www.reactome.org	生物核心通路及反应的挖掘知识库
PID	http://pid.nci.nih.gov	由其他数据库导入及文献挖掘获得的人信号转导通路数据库
STKE	http://stke.sciencemag.org	参与信号转导的分子及其相互作用关系的信息
AfCS	http://www.signaling-gateway.org	参与信号转导通路的蛋白质相互作用和信号转导通路图
AMAZE	http://www.amaze.ulb.ac.be	对细胞过程的相关信息进行表示、管理、注释和分析
BIND	http://www.bind.ca	提供参与通路的分子序列和相互作用信息
DOQCS	http://doqcs.ncbs.res.in	细胞信号转导通路的量化数据库，提供反应参数及注释信息
SigPath	http://sigpath.org	提供细胞信号转导通路的量化信息

4)转录因子数据库

TRRD 和 TRANSFAC 是两个转录因子数据库。

在不断积累的真核生物基因调控区结构和功能信息的基础上，研究人员构建了转录调控区数据库 TRRD。TRRD 条目包含了特定基因的各种结构和功能特性，如转录因子结合位点、启动子、增强子、静默子及基因表达调控模式等。TRRD 包括 5 个相关的数据表：TRRDGENES(包含所有 TRRD 数据库基因的基本信息和调控单元信息)，TRRDSITES(包括调控因子结合位点的具体信息)，TRRDFACTORS(包括 TRRD 中与各位点结合的调控因子的具体信息)，TRRDEXP(包括对基因表达模式的具体描述)，TRRDBIB(包括所有注释涉及的参考文献)。TRRD 主页提供对这些数据表的检索服务(http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/)。

TRANSFAC 数据库是关于转录因子、结合位点和 DNA 结合谱的数据库。它由位点、基因、因子、类别、阵列、细胞、方法和参考文献等数据表构成。此外，还有几个与 TRANSFAC 密切相关的扩展库：PATHODB 库收集可能导致病态的、突变的转录因子和结合位点；S/MART 库收集与染色体结构变化相关的蛋白质因子和位点的信息；TRANSPATH 库用于描述与转录因子调控相关的信号转导网络；CYTOMER 库体现人类转录因子在各个器官、细胞类型、生理系统和发育时期的表达状况。TRANSFAC 及其相关数据库既可以免费下载，也可以通过网页进行检索和查询(http://www.gene-regulation.com/pub/databases.html)。

5.其他数据库

此外，还有很多用于实现特定功能的数据库，如蛋白质功能注释数据库、蛋白质同源信息数据库、基因突变数据库、疾病相关数据库等，下面介绍两个较为重要的数据库：基因/蛋白质功能注释数据库 GO 和文献索引数据库 PubMed。

蛋白质功能注释的一个通用标准是**基因本体**(Gene Ontology，GO)，它提供了一种等级化、结构化、动态和限定的词汇表，用于描述基因或者蛋白质具有的生物学功能、参与的生物学进程及亚细胞定位信息。目前，GO 已广泛应用于模式生物的蛋白质功能注释，并且成为事实上的功能注释标准。大部分蛋白质的已知功能注释信息由 GO 协会(GO consortium)提供(http://www.geneontology.org/)。

PubMed 是 NCBI 维护的文献引用数据库，提供对 MEDLINE 和 Pre-MEDLINE 等文献数据库的引用查询，以及对大量网络科学类电子期刊的链接。利用 Entrez 系统可对 PubMed 进行检索(http://www.ncbi.nlm.nih.gov/pubmed/)。

除了以上提及的数据库，还有许多专门的生物信息数据库，涉及生物学研究的各个层面和领域，由于篇幅所限无法一一详述。国内也有一些大数据库的镜像站点和自己开发的有特色的数据库，如欧洲分子生物学网络组织 EMBNet 的中国结点——北京大学分子生物信息镜像系统。

6.数据库的选择与查询

随着各类数据库的增长，一系列问题也随之产生：这些数据库是否具有相同的格式？哪一个最精确？哪一个更新最快？哪一个最全面？使用者应该如何选用？以蛋白质数据库为例，PDB 数据库以蛋白质三维结构信息为特征；PIR 数据库包含的数据信息最全面，但是其中的解释说明相对贫乏；SWISS-PROT 数据库的组织结构非常好，并对每个条目进行了详尽的说明，但是它所覆盖的序列比 PIR 数据库少。

通常来说，依据人们对待查询序列所掌握的信息和查询的目的，选择多个数据库进行查询并比较它们的结果，是更为合理的策略。

1.4.3 文献资源

通过阅读文献可以了解生物信息学相关领域的研究现状，发现待研究的问题，收集相关数据并建立有效的分析策略。同时文献的更新程度较快，体现了最新的研究动态和目前的技术水平。因此阅读文献是做研究的基本功，也是日常工作之一。

1.文献的获取

获取文献有多种方法，如通过谷歌进行学术搜索、在文献数据库中查询、向作者索要、论坛求助等。其中最重要的文献资源来自生物医学文献数据库 PubMed，它收录了生物医学相关领域的超过 2700 万篇文献(截至 2017 年 12 月)，以网页形式提供查询，并且很大一部分文献可以免费获得全文。PubMed 数据库的高级搜索界面如图 1.5 所示。同时，很多高校的图书馆购买了一定的文献资源，可通过校内网络方便地下载全文。如国防科技大学图书馆(http://library.nudt.edu.cn/)购买了 *Nature* 期刊、Elsevier 出版社和 Springer 出版社的期刊数据库，并可下载中国知网、万方和维普科技收录的大部分中文文献和优秀硕博论文。

2.文献的类型

各种期刊中收录的文献主要包括如下 5 种类型。

① 综述(review)，这类文献系统总结某个主题的已有研究成果，分析现有方法的问题，并展望未来的发展趋势。

② 技术报告(technical report)，主要对某项研究或技术方法的过程进行描述，给出研究进展、技术现状、研究结果或问题。

③ 评述(comments)，即对某项研究或者观点的评论。

④ 验证研究(validation studies)，即验证某种方法或某项实验结果。

⑤ 其他，如数据库、工具和研究策略介绍等。