第一章　概论

当今世界，科学技术的发展日新月异。其中，生命科学的进展尤为引人注目。进入分子水平以来，人们发现在生物化学、分子生物学、免疫学以及遗传学领域的研究中有大量的数据资料需要处理。于是，随着计算机技术、网络通讯的飞速发展，产生了一门新兴的学科——生物信息学。它首先利用电子计算机技术，对在分子生物学等学科的研究中产生出来的大量原始数据进行收集、整理和管理；其次，对各种数据进行对比、分析、归纳并建立计算模型，以期更好地解释数据，并进行结构、功能的预测以及仿真，等等(图 1-1)。它的出现极大地推动了分子生物学的发展，在人类基因组计划的研究中发挥了重要的作用。这门学科在生物学、医学领域有着十分广泛的应用。其中的一些大型生物学数据库包含了众多的生物学信息资源，人们可以很方便地从国际互联网上寻找

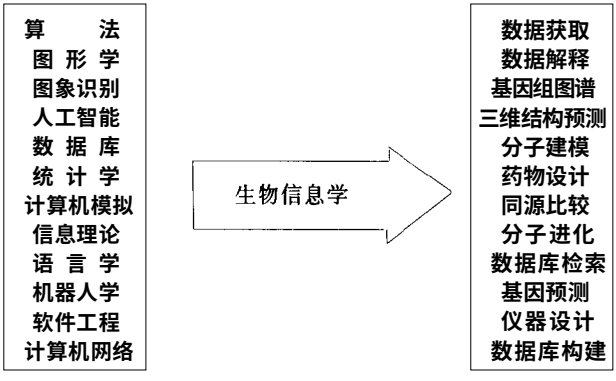

图 1-1　生物信息学是计算机科学、数学和分子生物学之间的桥梁

所需的资料和处理工具。这不仅方便了研究思想和资料的交流，减少了许多重复性的工作，而且也提供了一种崭新的思维方式和科研工作方法。近年来，互联网的高速发展为人们共享数据资源、合作研究提供了网络这一物质基础。越来越多的生物学、医学、药学工作者认识到生物信息学的重要性和实用性，其良好的发展前景业已显现。

第一节　生物信息学及其与生物学的关系

近十多年来，生命科学在分子水平上进行了广泛而深入地研究。随之而来的是大量的数据结果需要处理。特别是生物化学、分子生物学及遗传学的研究，各种各样的有关生物分子的原始实验数据，数量十分庞大。因此，利用计算机技术处理数据十分必要。另外，众多的学科，诸如结构生物学、酶学、细胞生物学、生理学、病理学、神经生物学等等，从不同角度的研究结果，可经过计算机的分类、组织和构建，形成具有生物学意义的新的研究结果。这些新的结果是对生命体的细胞结构和功能更为本质的反映。在这样的情形下，生物信息学应运而生。

生物信息学（Bioinformatics）的萌生可以追溯到 1956 年，那时还是计算机的初创期。在美国田纳西州的 Gatlinburg，曾召开过首次“生物学中的信息理论讨论会”，这拉开了生物信息学的序幕。随着二十世纪八、九十年代计算机技术的迅猛发展，它才同时获得自身的快速成长。无论从理论上来讲，还是从现实情况来看，生物信息学都还是一门相当年轻的学科，它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。它的诞生和发展是应时所需，是历史的必然，并且已经悄然渗透到生命科学的每一个角落。以至于在整个科学界意识到它的存在之前，相关学科的研究者就已经离不开它了。

二十世纪末期，生命科学技术的迅猛发展，无论从数量上还是在质量上，都极大地丰富了生命科学的数据资源。数据资源的急剧膨胀首先迫使人们不得不考虑寻求一种强有力的工具，在有效地组织数据的同时，有利于对已知生物学知识的储存和进一步地加工利用。在大量多样化的生物学数据资源中，必然蕴含着许多重要的生物学规律。这些规律是我们解决许多生命之谜的关键所在。然而，继续沿用传统手段以人脑来分析如此庞杂的数据是不可能的。人们同样需要寻求一种强有力的工具去协助人脑完成这些分析工作。可以说，伴随着二十一世纪的到来，生命科学的重点和潜在的突破点已经由上个世纪的试验分析和数据积累，转移到数据分析及其指导下的实验验证上来。生命科学也正在经历着一个从分析还原思维到系统整合思维的转变。

那么，我们所寻求的那种强有力的数据处理分析工具，就成为未来生命科学的关键所在；伴随着生命科学这一需求的加剧，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突飞猛进的发展，而自然地成为生命科学家的必然选择。计算机科学技术和网络技术正日益渗透到生命科学的方方面面，一门崭新的、拥有巨大发展潜力的生物信息学也就悄然而坚定地发展起来了。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——作为新一代生物科学研究的重要工具。

生物信息学(Bioinformatics)这一名词的由来，还要从八十年代末期说起。美国佛罗里达州立大学超级计算机计算研究所的林华安博士认识到将计算机科学与生物学结合起来的重要意义，遂开始留意为这一新的领域构思一个合适的名称。考虑到与佛罗里达州立大学大型计算机计算研究所的关系，起初，他使用的是“CompBio”。当时，这一机构支持由他主办的一系列“生物信息学”的会议；之后，他又将其改为兼具法国风情的“bioinformatique”。因其拼写看起来似乎有些古怪，不久，他便进一步把它更改为“bio-informatics(或 bio/informatics)”但由于当时的电子邮件系

统与今日不同,该名称中的‘一’或‘／’符号经常会引起许多系统问题。于是，林博士又将其去除。今天，我们所看到的“bioinformatics ”就这样正式诞生了。林华安博士也因此赢得了“生物信息学之父’的美誉。

一、生物信息学的定义

生物信息学主要是由分子生物学与信息学、计算机技术、数学、物理学等学科交叉结合的产物。对于这样一门年轻的边缘科学,不同的学者对它的定义不尽相同(见附录二)。有不严格的定义称之为：分子生物学与计算生物学的交叉学科。国外学者一般认为，它是对现代分子生物学和生物化学技术带来的不断增加的复杂的资料进行分析、组织并使之系统化的一门科学。也有人认为，生物信息学应含有生物系统内信息链的内容，它主要指的是贮存于 DNA 或 RNA 中的信息，表现为核苷酸的序列并能通过翻译表达出重要的生命大分子——蛋白质。对这部分内容的研究,无疑是生物信息学在应用上的一个很重要的方面。我们认为生物信息学的含义是基于计算机和互联网的应用和信息科学的知识方法对生物信息进行收集、整理、分析研究、处理和应用的一门交叉学科。今天已经认识到：一项研究欲更加深刻地反映生物的本质规律，需要用到这门新兴的学科。例如，基因密码的含义与相对应的生物机体生理特点之间的关系、人脑的研究、基因与意识及心理行为的关系、系统遗传学家对各物种之间内在关系的研究等等，这类研究均需在计算机软件技术、各种不同类型的生物学数据库的辅助下完成。又如，结构生物信息学对靶蛋白质活性位点精细结构的描述可为新药的模拟设计提供良好的基础。总之，生物系统的复杂性需要生物学方法与计算技术的结合。所以，生物信息学是一门建立、管理并运用生物信息数据库研究生命现象，并最终模拟出生命有机体复杂性的科学。

一门学科的建立除了有应用上的需求外，还应当有相应的理论支持。信息学理论的发展是其重要的支柱之一。此外，计算机凭

借其强大的运算分析功能介入到生物学的研究中，使研究手段、工具方法迈上了新台阶。美国学者 H. Rashidi 和 L. K. Buehler 就认为生物信息学是建立在这样一个假设的基础上的：即基因结构、基因在基因组中的排列位置、蛋白质的功能以及在机体中引起能量代谢、繁殖和构成诸如身材、体型等蛋白质的相互作用之间存在着一个分等级的关系。而对其相互关联的研究，使人们意识到计算方法的介入为此提供了一个良好的平台。

二、生物学的发展与生物信息学

二十世纪初，人们用有机化学的方法研究三大物质的代谢途径，研究酶的组成及生理作用，等等。那时候，生物化学家没有分子生物学、基因的知识，并不知道核酸是生命的遗传单位。他们的研究是对各种实验现象的观测和记录。而时至今日，人们已经可以在电脑前完成基因测序、基因筛选、计算机识别蛋白质功能、计算机模拟蛋白质三维结构以及新药设计等工作，发展出计算和实验方法相结合的新的生物学研究模式。下面试举一例，来说明生物信息学在这一新模式中的用途。

基因是生命的遗传单位。在复制时，保持基因中分子信息的严密性和准确性是十分重要的。我们在研究某个基因突变与肿瘤发生的关系时，该基因的克隆是首先应完成的，因为这是获得核酸序列及寻找调节因子的第一步。首先，将我们需要的 DNA 片段从有关的基因组中分离出来，然后将这段基因插入到一个载体 DNA 中，从而制成重组 DNA。按生物进化的观点，所有生命体在遗传上是有密切的相关性的，所以人类基因在其他动物体或微生物体内操纵复制是完全有可能的。由此，人们将上述重组 DNA 置入细菌体内繁殖，从而达到基因克隆的目的并可复制出大量的基因拷贝。应用这种方法复制基因、扩增 DNA 简单而有效，而且避免了为纯化 DNA 或蛋白质而需获取大量的人体组织标本的过程。

基因克隆完成后，即可对基因测序。通过基因序列可预测其

相应蛋白质的结构和功能。这些工作如今已可在计算机辅助下完成。而重组 DNA 又可用来合成相应的蛋白质，对后者进行生物化学的检测分析，以进一步明确其结构、功能及在致病过程中的作用。对肿瘤基因及遗传性疾病的研究中，为了明确该致病基因在基因组中的定位，常常需要获得携带有突变基因的个体样本及正常人的样本。在实践中，对一定数量的两种样本的对比分析可运用相应的计算工具。这种工具是按医学研究的目的而建立起来的生物学数据库，并经过不断地调整编辑而成。毫无疑问，这一编辑过程也促进了人们对疾病本质及其遗传本质的理解。

二十世纪八十年代后期，计算机技术进入快速发展时期，此后的互联网以更高的速度在全球铺展开来。与此同时，一项庞大的人类基因组计划及其他的生命体基因组研究业已全面展开。这些均是在生物信息学形成和发展中具有决定性的事件。在基因组计划中，人们更关注基因的核酸序列。在获取基因序列并揭示其中的生物信息的过程中，生物信息学是重要的分析工具。

当前，分子生物学与生物信息学的结合愈发紧密。后者为前者提供了新的研究手段和方法，如今已在基因克隆、核酸测序、基因定位等方面有着广泛地应用。在人类基因组研究及后基因组的研究工作中，效率是经常会被提及的因素之一。在 DNA 的序列研究中，任何一种计算方法都比实验分析要迅速和廉价，且计算分析为实验分析提供了互补的预测性信息。现有的算法在利用已知的生物学知识的基础上，已经完成了不少工作。可以预见：在未来，对生物学有了更深入的研究之后，计算分析学家和实验生物学家会有更频繁更深入的合作，这一领域会有更为显著的进步。以计算和实验相结合的新生物学，已快步向我们走来。

三、基因组学、蛋白质组学与生物信息学

如今，基因组学、蛋白质组学已是生命科学研究中最重要的内容之一。传统上，要获得一个基因序列，需要 mRNA 的分离或检测蛋白质的氨基酸序列。其后，通过诸如蛋白电泳等检测手段，

探寻该基因及其相关蛋白质与生物体的发生、发育、老化及疾病发生之间的联系。基因的组成、表达等与生物体的生理功能相关的信息，在阅读理解基因图谱时，有重要的意义。现代基因组计划有两方面的工作：其一，基因组结构是与一定的生理功能相关的。所以，人们希望通过研究一个生命体的全部基因组序列，以帮助了解其生物学特点；其二，高等生物含有大量的非编码 DNA，直到最近人们才对其功能及存在的意义有了初步的了解。在过去，人们采用功能性的检测方法不能获得非编码 DNA 的序列信息，而未来的研究会进一步加深对它的认识。

科学家的研究是从 DNA 开始的，但在生物体内它只是遗传信息的载体。因为 DNA 在体外是不能自我复制的，所以核酸并不是单独完成遗传使命的，DNA 上所携带的遗传信息必须首先被解读。在细胞中，这一工作是由一些蛋白质承担的，诸如存在于胞核或胞浆中的蛋白质及酶等等，都是解读遗传信息的工具，而且在胚胎发生的早期就起着重要的作用。尽管基因成对出现在染色体上，但表达的只有一条，它来自父方或源于母方。所以，并不仅仅只有 DNA 序列是遗传信息。染色体的结构， DNA 与其表达的蛋白质间的相互关系以及其构型组合也是信息的一部分。一种影响或决定子代中母系抑或是父系的基因激活的表达机制叫遗传印迹（Genetic Imprinting） 。要搞清楚这一现象及其基因剂量效应（一个基因能表达多少蛋白质是与其相关基因有联系的），就必须研究发生在细胞内的整个遗传过程的时空调控。因此，有人认为人类基因组计划一旦完成，接下来的工作将由分子生物学实验室转入由电子仪器组成的实验室。在那里，基因的各种信息会轻松地获得，这有利于确定基因在发育、衰老及疾病发生等过程中的作用。

蛋白质组学是基因组计划完成后或同时开展的重要研究领域之一。要明确各种器官、组织、细胞以及正常和疾病组织中多种蛋白质表达谱的变化，即蛋白质的大规模识别和定性，需要有强

有力的分析工具。2D凝胶电泳是十分有用的方法，在2D凝胶电泳的结果分析过程中离不开生物信息学，后者使这一分析过程自动化。本书有专门章节介绍生物信息学在蛋白组学中的应用。

当前，数据库内容的增加及变化都十分迅速和频繁。因为每天都可能有新提交的数据加入其中，所以数据库的目录可能每天都在更新。这更有利于研究机构的获取和利用。有资料显示：在1998年4月包括83个物种的基因组计划已完成了21个物种的测序，其中大部分为微生物。这一工作采用了自动克隆和聚合酶链式反应（PCR）完成DNA的扩增及测序。这些方法可以把由盲法产生的随机DNA片段重建为无间沟（gap）的连续序列，并最终得到全部基因组的所有碱基序列。在进行基因组计划的过程中，每天都有大量的信息出现，并进入到数据库中。

基因组研究院(The Institute for Genomic Research，TIGR　)创建于1992年，是一个非赢利性的研究院。它位于美国马里兰州的Rockville，与美国国立卫生院(NIH)、约翰一霍普金斯大学、马里兰大学及其它研究所、生物技术公司毗邻。占地12公顷，有50000平方英尺的实验室及办公区域。该研究院有大型的DNA测序实验室和与生物信息学、生物化学和分子生物学相关的现代化设备。它从一开始就利用网络成长起来。对于许多科学家来说，TIGR使他们开始真正认识了基因组计划，并获得了一种新的大批量测序的方法。TIGR的工作促进了测序程序及数据分析的发展。类似TIGR这样的研究院和组织还有一些，他们在网络上提供的信息数量之大令人惊讶。TIGR的研究对象是病毒、真菌、致病菌、原生质、真核生物以及人类的基因组，并对基因产物的功能、结构进行比较分析（摘自 http://www.tigr.org/about/)。

TIGR率先发展了大量复制DNA的必需技术，以及EST(Expressed Sequence Tag，表达序列标签）测序计划。EST文件提交格式简单、易于快速处理，因此每天提交进入数据库的数量级可以达到数千个，峰值期可达每周100，000个提交量。这种测

序形式已在一定程度上影响了生命科学界的工作方式。因为现今大多数的期刊已不再刊登完整的序列数据，而只标明其序列在数据库中的序号。研究者在公开发表文章时要向公共数据库提交其研究结果，这已成为一条准则。而且，有些大型基因研究中心规定新发现序列的公开应先于论文的发表。这些情况都使得相关数据库的内容呈指数级上升，同时也使数据信息的整理、分析、利用显得十分重要。

最初，生物信息学就是为来自不同国家、不同研究组织之间的信息交流服务的，是他们相互合作、信息共享的方式之一。随着数据库的集中合并及网络交流的迅猛发展，它不仅成为了业内人士主要的交流方式，而且很快转变为一门独立的学科。特别是人类基因组计划的实施，更多的愈来愈强大的克隆和测序技术不断出现，直接促进了生物信息学的发展。在这项庞大的计划中，国际间的合作成为了必然。由此而来，出现了拥有各种数据库的公立或私立的组织，他们的数据库为整个基因组测序、基因定位以及在细胞或分子水平上寻找 DNA 序列信息与结构功能的关系提供了实用的工具和便捷的服务。

其后，工商企业界及金融投资者逐渐认识到生物信息的处理和出售极具潜在的利润价值。他们的介入使得数据库的建立、完善有了充足的资金来源，并且在其未来发展中扮演着重要的角色。潜在的利润和商机，又促进了各种基因研究工作的深入并使其竞争日趋激烈。

四、国内生物信息学现状及展望

国际上，欧美等国家在生物信息学的研究和应用方面已经有了较长时间的积累。国内对生物信息学领域也越来越重视，在一些著名院士和教授的带领下，在各自领域取得了一定的成绩。2001 年 4 月在北京召开了中国首届生物信息学大会，参会人员遍及全国 10 多个省市，共 600 余人。此次会议较为全面地回顾了我国生物信息学研究的现状。2001 年 10 月，国家发展计划委员会宣布，

我国将在中国科学院建设“生物信息系统国家研究中心”，形成有国际竞争能力的基因组学、蛋白质组学和生物信息学的整体技术平台。这将会推进我国生物信息技术的发展。下面将国内部分单位的生物信息学发展状况做一简单的介绍。

1. 中国科学院基因组信息学中心生物信息学平台

中国科学院基因组信息学中心设有专门的生物信息室，配备有由国家智能计算机研究开发中心研制的曙光 3000 型大型计算机。这是目前国内性能最高、运算速度最快的超级服务器。该系统峰值浮点运算速度为每秒 4032 亿次，内存总量为 168GB，磁盘总容量为 3.63TB。它具有先进的体系结构，丰富而完善的软件系统和一大批行业应用软件。该生物信息学平台负责的项目包括：人类基因组计划中国部分完成图、嗜热菌基因组、螺旋藻基因组、超级杂交水稻基因组工作框架图和中华民族基因组及疾病相关基因的多态性研究等。

2. 北京大学生物信息学服务器

北京大学生物信息学服务器是在罗静初和顾孝诚教授领导下建立的，由北京大学附属的分子设计实验室和物理化学研究所维护。它是国内第一家生物信息学网站，设有多个国外著名分子生物学数据库的镜像站点，如：Protein Data Bank (PDB)、Structural Classification of Protein (SCOP)、Protein Information Resources (PIR)、SWISS-PROT、ENZYME、PROSITE、BLOCKS 等，在国内查询这些数据库亦非常方便快捷。他们开展的项目包括蛋白质结构预测、以结构为基础的药物设计、蛋白建模和设计等方向的研究。

3. 联众研究院生物信息分析平台

该生物信息分析平台隶属于上海复旦大学，他们建立了自已的 EST 数据库、Cluster 数据库 (UniGene 和全长基因数据库，并从国外引进了 GenBank、SWISS-PROT、EMBL、OMIM、UniGene 等数据库。每天能开展对约 1500 个序列进行公开数据库

的查询、全长基因的识别、全长基因编码蛋白的结构与功能预测、部分全长基因的染色体定位等方面的工作。对外提供的生物信息学服务包括：引物设计、核酸一级结构、同源核酸序列数据库搜索分析、ORF 预测、氨基酸组成、理化特性的分析、蛋白质功能域分析、基因或蛋白家族分析、基因组 DNA 的外显子区域预测、ESTs 与基因组序列比较、蛋白质亲水性分析、蛋白质跨膜区预测、信号肽预测、序列抗原性分析、二级结构预测等。他们开发了中文环境的软件 Biolink，可以用于计算蛋白质等电点分析、蛋白质二级结构的分析预测、一条或多条序列在一个或多个核酸或蛋白序列库中进行同源搜索、DNA 限制性内切酶图谱分析、识别基因 ORF 编码区、预测蛋白质在细胞内定位、分析蛋白质的一些理化性质（如：亲／疏水性、跨膜片断等）。具有识别跨膜螺旋区、分析氨基酸的组成、PCR 引物和杂交探针设计等功能。

4. 中国人民解放军总医院神经信息中心

该中心成立于 2001 年 9 月。人类脑研究计划是继人类基因组计划之后又一国际性的重大科研项目，其核心是神经生物信息学。科学界认为该计划比基因组计划规模更大，囊括了更加广泛的内容。人类脑研究计划的目标是提供先进的信息学工具，使神经科学家和信息学家能够将脑的结构和功能研究结果联系起来，建立数据库，进行搜索、比较分析、合成和整合，绘制出脑功能、结构和神经网络图谱。中国人民解放军总医院神经信息中心的主要任务是：建立神经信息工作平台，为开展神经信息学研究提供必要的条件。其中包括：在国内 6 大城市 11 个研究单位开通神经信息电子网络，进行网上信息交流和科研协作；与国际神经信息电子网络接轨，引进和推广全球性“人类脑计划”的科研成果；开展神经信息科研工作，组织全国性脑研究计划；拟成立相关工作组，以代表中国加入全球神经信息学工作组织，参与全球人类计划的研究工作；提供神经信息服务。

虽然国内生物信息学的发展非常快，但总体来讲与国际水平

差距还比较大。一方面表现为相对于国内生物医药科学的研究与开发，对生物信息学的研究和服务的需求滞后；另一方面是，真正开展生物信息学服务的公司相对较少。仅有的几家科研机构主要开展生物信息学的理论研究，而声称提供生物信息学服务的公司所提供的服务也仅局限于简单的计算机辅助分子生物学实验设计，而且服务体系并不完善，这就与欧美发达国家有了较大的差距。

生物信息学的产业特点是投资少、见效快、效益大，适合我国的现实条件。如果从互联网上源源不断地采集数据，然后进行分析、归类与重组，发现新线索、新现象和新规律，用以指导实验工作的设计，这是一条既快又省的科研线路，可以避免不必要的重复，提高我国生命科学的研究水平。其关键在于加速培养一批在数学、物理、计算机科学和分子生物学方面均有造诣的跨学科青年人才。如能充分发挥现有人才的潜力，进一步培养大批生物信息学的专业人员，才能迎接 21 世纪的挑战。

第二节　计算机在生物学及医学领域的应用

一、生物学、医学与计算机

众所周知，技术的进步对科学的发展起到了重要的促进作用。在最近的二十年，这一趋势更加明显。例如，纳米，这一奇妙世界的物理尺度，也是生命分子本身各种组成部分的尺度。纳米技术是一种新近发展起来的对单个分子进行操作的技术，现已成为一个时髦的研究领域。该技术显示出：将以单分子机械装置为目标，促进医疗和电子技术的微型化。又如，材料科学把化学与生物化学有机地结合在一起。在生命科学领域，生物样品的准备过程中发展起来的荧光染色技术，引起了细胞生物学和 DNA操作技术的革命（例如：Affymetrix 公司的 DNA 芯片技术）可视化的脑检测技术又使脑科学的研究进入了一个新的层次。化学和生

物学研究方法的结合又开拓出一些新的研究领域（如：组织工程），给二者的发展注入了活力。同样的，计算机技术在生物医学领域也有很广泛的应用。

应用数学及计算机科学是现代生物学的重要研究工具。假如没有计算机对信息的存储和读取，没有数据装载和统计分析，没有计算机模拟系统，就不可能产生现代分子生物学。可见，计算机在这一领域发挥了重要的作用。从软件设计、 PC 机的应用到互联网的交流，都充分利用了计算机技术。而且，几乎在所有的科研活动中，它都发挥了日趋重要的作用。在未来，这一作用将更为明显。

在基础医学研究中，实验就是在特定的时间内通过一系列的技术手段检验一种观点或得到新发现的过程，其结果如何并非出自偶然，而是由事物的必然性决定的。现在认为，计算机工具在实验的设计、执行和分析研究过程中可以起到核心的作用。计算机的介入并没有改变科学思想本身，也没有改变围绕着科学发现与错误模型之间的假设与争论，但却改变了科学研究的核心内容——具体实验的本身。计算机可以计数培养皿中的细胞，测量显微镜下的各种组织切片中的细胞核大小，记录对慢性疼痛敏感的神经元的电活动，读取电泳凝胶上可记录于 光底片上的 序列。其他不可缺少计算机的实验室设备有：流式细胞仪、远程病理诊断系统、 芯片分析仪、DNA 凝胶成像系统等等。计算机可以帮助人们快速而精确地记录许多重要数据，明显提高了实验的精确性。

当然，实验的精确性并不完全取决于计算机，而主要是实验仪器的质量。如：电镜下精确切割冰冻细胞样品、把微玻璃管插入干细胞转移细胞核以培育转基因鼠、或者测量脑组织中单个神经元的电活动，都需要设计制造出高质量的合金。计算机在生命科学研究中的作用依然是控制、运算、数据分析及存储。计算机的数字化记录及其易于复制的特点，极大地提高了生物数据的存

储 量 。

临床医学是计算机技术的另一个受益者。两种成功的无创诊断技术——核磁共振成像和超声波检查，均是有效地利用了所有物质中特异性原子或分子的物理特性而成像的，计算机技术在其中起到了重要的作用。另外，计算机专家系统的发展和使用，将会使常规医疗变得愈来愈方便，且治疗的成功率也愈来愈高。在临床科研及临床诊治中，精密的检验和治疗仪器的使用都离不开计算机。现代医学借助了许多具有分析功能的仪器及新颖的实用医疗操作技术辅助医生诊治疾病。例如：用来监测血糖水平的生物传感器、用作微血管介入技术的导管等等，都与计算机技术相关。目前，国外医学领域的科研医疗与计算机的联系已十分密切，甚至已呈现出较强的依赖性。尤其有代表性的是“千年虫”问题，在 1999 年底给发达国家的医学界也带来了不少麻烦和干扰。

生物信息学是计算机与生物学紧密结合的产物。人类基因组计划旨在绘制一幅染色体图谱。也正是由于这一计划中产生的大量序列信息，大大地促进了生物信息学的发展。而神经生物学正在绘制由大脑解剖及其细胞组分构成的图谱。大脑是一个十分复杂的器官，对它的研究意义重大。但利用传统技术研究多年，进展有限。在神经生物学家、认知科学家及心理学家的携手合作下，一门新兴的边缘学科——神经生物信息学出现了。这一领域实际上也属于生物信息学的范畴。这是研究大脑与神经元结构和功能的新途径，是理解以网络形式存在的复杂的神经系统的新途径。

二、计算机算法

在预计计算机解决某一问题需要多少时间时，这里面其实包含了两部分内容：其一，计算机自身计算时所需要的时间；其二，分析指令的时间，这是人为干预占用的时间。所以，现在的许多自动程序软件以培养计算机的自主决定能力为目标，以避免人的操作占用了较多的时间。这样更有利于计算机在短时间内处理大量的数据。计算机是以其特殊的专有系统作为其运行的基础，这

一系统可完成一项或多项计算量很大的任务。过去，许多需要人们操作干预完成的工作，现在已经可由具有学习功能的神经网络（neural networks，NNs）作出正确的答案。神经网络尽管发展前景可观，但仍难以成功地解决如何对符号和记忆进行控制的问题，并且无法训练 NNs 产生数据之外的信息。一旦建立了一种计算方法，并且成功地运行后，计算机可不停地重复这一过程。并且，在不断地输入及输出的过程中，通过反馈及反馈环作用进行调整，以适应预先设定的要求。除了神经网络算法外，还有许多种著名的算法，如 HMM（Hidden Markov Model ）等等应用于生物信息学。在人类基因组计划的序列拼接中，生物信息学发挥了很重要的作用，目前 DNA 自动测序仪的每个反应只能测序大约 500bp。如何将这些序列片段拼接成完整的 DNA 顺序，就成为测序后的一项重要工作。传统的测序技术通常是将克隆进行亚克隆，并对亚克隆进行排序，这些工作需要大量的人力物力。现在，生物信息学提供了自动而高速地拼接序列的算法，即根据 Lander-Waterman 模型利用鸟枪法进行测序，再将大量随机测序的片段用计算机进行自动拼接。这种技术不仅避免了亚克隆排序所需的大量繁琐的工作，还使序列具有一定的冗余性以保证序列中每个碱基的准确性。

可见，计算机算法所表现出的优越性是无可置疑的，但对其不足也应有充分的了解。比如说，一个文字处理程序因可使书写变得轻而易举而大受欢迎，成为必要的文案书写工具。尽管编辑一个文本所需的时间不多，但是由于文本和图表排列很容易改变，反而使得纸张的浪费增多了。人们更愿意看到打印在纸上的文本，它似乎比荧光屏上的文件更可靠些，而且符合相当一批人的阅读习惯。但是，计算机的拼写监测器是缺乏语言分析能力的。其中的一个难题是如何校正文本的内容，如果输入的内容与原文不符但拼写无误，计算机就无法识别这种版面上的错误。而人脑之所以可识别这种区别，是因为它与计算机的工作机制不同。虽然校

对、分析数据以及随后的内容解释仍需依靠计算机的辅助，但这是在人脑的严格控制下进行的。计算机算法亦有相似的情况，其运行结果亦需研究者给出必要的分析判断。

计算机是出色的辅助工具，其算法可用以解决数字性问题，控制和指导仪器的运行，编辑处理文字信息，进行检索并寻找数据间内在的联系，建立数据库，等等。其中，后三项作用对生物信息学而言，至关重要。

三、不同类型计算机的功用

1. PC 机（Personal Computer）具有多方面的功能，可进行文字处理、表单分析、文件显示、互联网应用、甚至控制实验仪器。例如，登陆网站 http://www.axon.com 后，可以控制 AXON 的膜片钳(pCLAMP)。这是一种广泛应用于电生理学中，可以控制和测量神经元电活动的应用软件。还可进行离子成像和分析液浓度成像。PC 机在生物医学领域有多方面的用途，完成诸如：在基因组计划中可以分析 DNA 芯片的杂交信号、功能性神经外科手术中微电极的导向分析、诊断监测运动障碍（如帕金森氏症）等许多工作。PC 机和局域网（奔腾处理器、NT 工作站）的应用十分广泛，而且运算速度快、能力强大。这使得科学家无需使用超级计算机就能开展诸如构建分子结构和多序列对齐的工作。一般说来，机控的实验仪器有可选择的计算机界面，便可满足研究人员用于不同的实验目的。据估计，世界上仅有 1%的计算机微处理器用于台式 PC 机，其余的 99%则装备于其它的工业产品中，如飞机、供暖系统、实验装置、安全设施等等。它们通常是具有特定功能的软硬件结合的，且无需指令程序的芯片。

科学研究需要使用具有多种内置处理器的仪器，如气相色谱仪、电子天平、分光光度仪等等。例如，分光光度仪可在不同的波长处读取液体的吸光度，同时还可即时测量受测液体化学成份的变化。又如液相色谱仪，它的自动分离功能可根据分子的大小及溶解度的不同，将混合物分离成单一的组分。上述这类仪器均

由内置的微处理器直接控制，不必通过外接计算机另行控制。人们只需在类似 ATM （自动取款机）的小视窗上输入代码或提示指令，即可完成工作。近二十年来，这些微处理器从简单的控制回路发展成为今日可以贮存大量的数据和图象文件的芯片，正逐步取代记录信息的纸张和胶片的作用。

2. 超级计算机（Supercomputers）这类计算机可以完成需大量运算的任务，并且具有准确的工作记忆和超大容量的存储能力。它是网络服务商的主要的工作对象，大都应用 UNIX 操作系统，且已经服务于许多学术团体。鉴于其功能和价格，它所提供的服务已成为学术界设备共享的范例。例如，圣迭戈超级计算机中心(San Diego Super Computer Center，SDSC)(http://www.sdsc.edu/)是一家提供公共操作服务的公司，在学术界的应用十分广泛。SDSC 可提供并支持广泛的计算资源，目前其系列产品包括 CRAY C90 超级计算机、CRAY T3E 超级计算机、高级视像实验室、档案存储系统，等等。在美国，这些服务可供学术研究者及学生们使用。并且在费用均分的协议下，亦可供国内外的商业及政府人员使用。目前，已有超过 240 个机构的 5100 多位研究人员将这些服务平台用于科研中。

高速链接和并行运算

1997 年 6 月 20 日，匹兹堡超级计算机中心与德国斯图加特大学在大西洋两岸通过高速研究网络将超级计算机相互连接起来。这是首次将高速电信网用于跨洋运算。

作为国际高性能网络的预期模型，这项计划将匹兹堡的 512 位处理器 CRAY T3E 计算机与斯图加特高能计算中心的一台 512 位处理器 T3E 机相连。在不同的地点连接两台或更多的超级计算机，进行同一项运算任务称为宏观运算(Metacomputing)。这种连接实际上产生了 1024 位程序系统，其运算性能之高在理论上可达到每秒 6750 亿次。这项研究计划的实施是以两地间有高速越洋链接的一系列研究网络为基础的。近几年建立的这种网络，