生物信息学概论

陈新　　生命科学学院

2001 年 10 月

（一）、概述

生物信息学是在数学、计算机科学和生命科学的基础上形成的一门新型交叉学科，是指为理解各种数据的生物学意义，运用数学、计算机科学与生物学手段进行生物信息的收集、加工、储存、传播、分析与解析的科学。近年来随着快速序列测定、基因重组、基因芯片，多维核磁共振等技术的应用，生物学实验数据呈爆炸趋势增长，同时计算机和国际互联网络的发展使对大规模数据的贮存、处理和传输成为可能。作为一门新的学科领域，它是将基因组 DNA 序列信息分析作为源头，在获得了蛋白质编码区的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必要的药物设计。它由相互依赖、相互渗透的两个研究领域组成，即构筑现代生物学所必需的信息基础研究，以及旨在解析基本生物学问题的基于计算机技术的基础生物学研究。因此，在基因组研究时代，基因组信息学、蛋白质的结构模拟以及药物设计必将有机的结合在一起，它们是生物信息学的三个重要组成部分。

生物信息学更多的具备研究领域的特征，而非一套完整的科学概念和原理，因而具有独特的开放性和应用途径的多样性等特征。作为建立于应用数学、计算机科学、生物学、物理学、化学等多学科加查结合基础之上的生物信息学，无疑具有坚实的理论基础和广泛的应用前景。

本文将对生物信息学的发展、主要研究领域、任务及方法进行简要的介绍

（二）、生物信息学发展

1. 生物信息学的诞生和发展

生物信息学(Bioinformatics)就其萌生而言，是一门有“较长历史”的学科，因为早在计算机初创期的 1956 年就已经在美国田纳西州的 Gatlinburg 召开过首次“生物学中的信息理论讨论会”。而就其发展而言，却是一门相当年轻的学科， 因为继 20 余年的沉默之后，只有伴随着八九十年代计算机技术的迅猛发展，它才得以巨大发展。

二十世纪，尤其是末期，生命科学技术的迅猛发展，无论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首先迫使我们不得不考虑寻求一种强有力的工具去组织他们，以利于对已知生物学知识的储存和进一步加工利用。大量多样化的生物学数据资源中必然蕴含着大量重要的生物学规律，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实在是太勉为其难了！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，生物科学的重点和潜在的突破点已经由二十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试验验证上来，生物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物科学的关键所在；似乎是上帝的恩赐，伴随着生物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同样获得了突飞猛进的进展，自然就成为生物科学家的必然选择，计算机科学技术和网络技术日益渗透到生物科学的方方面面，一门崭新的、正是如火如荼的、拥有巨大发展潜力的生物信息学也就悄然而坚定地发展和成熟起来了！可以说，历史必然性的选择了生物信息学——生物科学与计算科学的融合体——作为下一代生物科学研究的重要工具。

关于生物信息学（Bioinformatics）这一名词是从何而来的，我这里想多说两句。据说八十年代末期，有个叫林华安博士的认识到将计算机科学与生物学结合起来的重要意义，开始留意要为这一领域构思一个合适的名称。起初，考虑到与将要支持他主办一系列生物信息学会议的佛罗里达州立大学超型计算机计算研究所的关系，他使用的是“CompBio”；之后，又将其更改为兼具法国风情的“bioinformatique”，看起来似乎有些古怪。因此不久，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-或/符号经常会引起许多系统问题，于是林博士将其去除，今天我们所看到的“bioinformatics”就正式诞生了。

2. 生物信息学的国内外现状

二十一世纪是生命科学的世纪，其里程碑就是即将完成的、历时13年、耗资数十亿的著名的人类基因组计划(Human Genome Project，HGP)，因为该计划的完成将为最终揭示人体构造之迷奠定坚实的数据

基础；而应人类基因组计划和生物科学迅猛发展的要求而迅速兴起的生物信息学则历史性地成为下一世纪生命科学浪潮中的热门学科。

通俗来讲，基因组是由四种不同的脱氧核糖核苷酸(A、T、C 和 G)按照特定的编码规则串联成的脱氧核糖核苷酸串（DNA），其中蕴藏着生物体中所有的结构信息和控制信息，因此，基因组可以说就是生物体内的控制中心，其中的功能单位可以转录为核糖核苷酸序列（RNA），有的就以 RNA 的形式发挥生物功能，有的则进一步被翻译成为各种蛋白质而行使生物体构建和生命调控功能。因此，基因组是一本完整地讲述人体构造和运转情况的指南，有了它，就可以揭开有关人体生长、发育、衰老、患病和死亡的秘密，因而危害人类健康的 5000 多种遗传病以及与遗传密切相关的癌症、心血管疾病、关节炎、糖尿病、高血压、阿尔茨海默氏症以及多发性硬化症和精神病等，就都可以得到诊断和治疗。

人类基因组计划就是要测出人类基因组的全部脱氧核糖核苷酸序列（估计其中编码有约十万多个蛋白质基因），进而弄清楚其中所有功能单位的组织结构形式以及调节机制，并绘制成直观图谱，该计划实现之后更深入的工作就是要弄清楚基因组所编码的所有蛋白质的表达情况，最终达到从整体系统水平上认识人体构造与功能并帮助制定有效治疗策略和开发有效治疗药物的目的。除此以外，还要对其它几个属于不同生物进化期的模式生物的基因组进行测序，如酵母、果蝇、蠕虫和小鼠等，利用这些模式生物可以进行很多在人体内不可能进行的实验研究，是我们了解人类基因组功能的重要工具。所有这些工作都涉及到大量数据的处理工作，而且数据量也在以科学史上前所未有的高速度增长着，所有这些情况表明，生物学已不再是仅仅基于试验观察的科学，仅靠传统的研究手段是无济于事的，理论和计算将越来越发挥巨大作用，数学、物理、计算机科学将日益渗透到生物学研究中来，海量的数据必须通过生物信息学的手段进行收集、分析和整理后，才能成为有用的信息和知识，才能再加以传播应用，也就是说，只有经过生物信息学手段的分析处理，我们才能获得对基因组的正确理解，因此可以说生物信息学兴盛于人类基因组计划，因为人类基因组计划首次为生物信息学创造了施展身手的巨大空间；当然，生物信息学并不局限于人类基因组工程，它已经深入到生命科学的方方面面。

国外一直非常重视生物信息学的发展，各种专业研究机构和公司如雨后春笋般涌现出来，生物科技公司和制药工业内部的生物信息学部门的数量也与日俱增。但由于对生物信息学的需求是如此迅猛，即

使是象美国这样的发达国家也面临着供不应求、人才匮乏的局面。

尽管在许多大学和研究机构已经各自成立了自己的生物信息学部门或中心，1999 年 6 月 3 日，美国国家卫生研究院（NIH）的专家委员会还是建议，迅速在大学和研究机构中建立 20 个生物计算中心，给予每个中心每年 800 万美元的支持，从事有关研究和人才培养，该建议可能在 2001 年开始实施。

近来，英国鉴于国内对生物信息学专业人才日益迫切的需求，所有主要的研究资助机构[医学研究委员会（MRC，Medical Reasearch Council）、生物技术和生物科学研究委员会、工程学和物理科学研究委员会（EPSRC，Engineering and Physical Sciences Research Council）、粒子物理和天文学研究委员会（PPARC，Particle and Astronomy Research Council）和 Wellcome Trust]不仅已经达成共识，认为应该高度优先地满足对生物信息学技术的需求，而且已经实现了对生物信息学人才培养的大力资助。

事实上，欧美等发达国家在生物信息方面已有较长时间的积累。

从数据库的角度来讲，早在 60 年代，美国就建立了手工搜集数据的蛋白质数据库。美国洛斯阿拉莫斯国家实验室 1979 年就已经建立起 genBank 数据库，欧洲分子生物学实验室 1982 年就已经提供核酸序列数据库 EMBL 的服务，日本也于 1984 年着手建立国家级的核酸序列数据库 DDBJ 并于 1987 年开始提供服务。

从专业机构的角度来讲，美国于 1988 年在国会的支持下成立了国家生物技术信息中心（NCBI），其目的是进行计算分子生物学的基础研究，构建和散布分子生物学数据库；欧洲于 1993 年 3 月就着手建立欧洲生物信息学研究所（EBI），日本也于 1995 年 4 月组建了自己的信息生物学中心（CIB）。

从数据分析技术的角度来讲，早在 1962 年，Zuckerkandl 和 Pauling 就将序列变异分析与其演化关系联系起来，从而开辟了分子演化的崭新研究领域；1964 年，Davies 开创了蛋白质结构预测的研究；1970 年，Needleman 和 Wunsch 发表了广受重视的两序列比较算法 ；1974 年，Ratner 首先运用理论方法对分子遗传调控系统进行处理分析；1975 年，Pipas 和 McMahon 首先提出运用计算机技术预测 RNA 二级结构；随着 1976 年之后大量生物学数据分析技术的涌现，Science 于 1980 年第 209 卷就已经发表了关于计算分子生物学的综

述；正如我们现在所看到的那样，在八九十年代，生物学数据分析技术在国外更是获得了突飞猛进的发展。

从专业出版业来看，由于没有专业领域专门的期刊，起初的专业文献都散落在各种其他领域的期刊中，到了 1970 年，出现了 Computer Methods and Programs in Biomedicine 这本相关期刊，到 1985 年 4 月，就有了第一种生物信息学专业期刊——Computer Application in the Biosciences；现在，我们可以看到的专业期刊已经很多了，包括书面期刊和网上期刊两种，如 Bioinformatics (formerly Computer Applications in the Biosciences) 、Acta Biotheoretica 、Bio Informatics Technology & Systems 、Bioinform Newsletter 、Briefings in Bioinformatics 和 Journal of Computational Biology 等。

从网络资源来看，国外互联网上的生物信息学网点非常繁多，大到代表国家级研究机构的、小到代表专业实验室的都有，大型机构的网点一般提供相关新闻、数据库服务和软件在线服务，小型科研机构一般是介绍自己的研究成果，有的还提供自己设计的算法的在线服务，总体而言，基本都是面向生物信息学专业人士，各种分析方法虽然很全面，但却分散在不同的网点，分析结果也需专业人士来解读。

目前，绝大部分的核酸和蛋白质数据库由美国、欧洲和日本的 3 家数据库系统产生；他们共同组成了 DDBJ/EMBL/GenBank 国际核酸序列数据库，每天交换数据，同步更新。其他一些国家，如德国、法国、意大利、瑞士、澳大利亚、丹麦和以色列等，在分享网络共享资源的同时，也分别建有自己的生物信息学机构、二级或更高级的具有各自特色的专业数据库以及自己的分析技术，服务于本国生物（医学）研究和开发，有些服务也开放于全世界。

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成绩，有的在国际上还占有一席之地，如北京大学在生物信息学网站建设方面、中科院生物物理所在 EST 序列拼接方面以及在基因组演化方面、天津大学在 DNA 序列的几何学分析方面、中科院理论物理所、清华大学、内蒙古大学等等……；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……，但从全国总体上来看与国际水平差距很大。

从生物信息学诞生的历史必然性来看，当世界各地的那些科学家开始认识到计算机的重要性并着手尝试利用计算机来组织、储备和分析生物学的观测资料的时候，生物信息学就已经开始了最初的萌芽；而随着计算机时代的到来，随着计算机技术广泛的介入生物学领域，生物信息学就可谓是遍地开花了。当然，限于历史局限性，那些科学家当时也许并没有意识到生物信息学已经在他们手里诞生了……

有人总结了一个 生物信息学大事记:

1956 年 10 月 29-31 日

“生物学中的信息理论讨论会”于美国田纳西州的 Gatlinburg 召开。

1958 年

由 H.P. Yockey 编辑的《生物学中的信息理论讨论会》由纽约 Pergamon 出版社出版。

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 等人发现噬菌体中的基因转录表达具有定时调节机制。俄文的 Biokhimia 第 27 卷。

1962 年

J. C. Kendrew 和 M. F. Perutz 因阐明“肌红蛋白与血红蛋白的晶体结构”而获得诺贝尔化学奖

1962 年

F. H. C Crick 和 J. D. Kendrew 因提出“DNA 分子双螺旋结构模型”而获得诺贝尔生理与医学奖

1962 年

Zuckerkandl 和 Pauling 将序列变异与其演化关系联系起来，从而开辟了分子演化的崭新研究领域。

Kasha 和 Pullman 的 Horizons in Biochemistry 一书。

1964 年

蛋白质结构预测的研究由 Davies 的工作开始。

J.Mol.Biol 第 9 卷。

1970 年

期刊 Computer Methods and Programs in Biomedicine 诞生。

1970 年

Needleman 和 Wunsch 发表了广受重视的两序列比较算法。

1970 年

Gibbs 和 McIntyre 发表了单序列分析方法——矩阵打点作图法，用于寻找单条序列中的重复片断，从而推测其功能。

Eur.J.Biochem 第 16 卷。

1972 年

Gatlin 在序列比较中引入信息理论，首次得到证明自然序列具有高度非随机性的定量证据。

1972 年

蛋白质序列数据库出现。

Dayhoff 的 Atlas of Protein Sequence and Structure 一书。

1974 年

Ratner 首先对分子遗传调控系统进行理论处理。

Prog.Theor.Biol.第 3 卷。

1975 年

继第一批小 RNA（tRNA）序列的发表之后，Pipas 和 McMahon 首先提出运用计算机技术预测 RNA 二级结构。

Proc.Natl.Acad.Sci.USA 第 72 卷。

1976 年

Fiers 等人测得第一个基因组全序列——RNA 噬菌体 MS2 的全部 RNA 序列；许多序列分析的算法开始涌现。

1977 年

将 DNA 序列翻译成蛋白质序列的算法出现。

Korn 等, Proc.Natl.Acad.Sci.USA 第 74 卷; Mccallum 等, J.Mol.Biol.第 116 卷。

1978 年

核酸序列数据库出现,收录有发表的 5S 和 5.8S 核醣体 RNA 序列。

Erdmann, Nucl.Acids.Res.第 5 卷。

1978 年

核酸序列中限制性酶切位点的计算机预测软件出现。

Fuch 等, Gene 第 4 卷; Gingeras 等, Nucl.Acids.Res.第 5 卷。

1980 年

Science 第 209 卷发表 Gingeras 和 Roberts 关于计算分子生物学的综述: Steps towards a programmed analysis of nucleic acid sequences。

Science 第 209 卷。

1982 年

A. Klug 因研究“病毒空间构象”而获得诺贝尔化学奖

1985 年 4 月

生物信息学专业期刊——Computer Application in the Biosciences 创刊。

1986 年

日本核酸序列数据库 DDBJ 诞生。

1986 年 7 月 21 日

A.Bairoch 创建蛋白质数据库 SWISS-PROT。

1988 年

R. Huber 因研究“紫色细菌光合反应中心三维结构”获得诺贝尔化学奖

1988 年

在美国国会的支持下, 美国国家生物技术信息中心(NCBI)成立,

该中心隶属于美国国家图书馆，其目的是进行计算分子生物学的基础研究，构建和散布分子生物学数据库。

1988 年 3 月

A.Bairoch 创建 PROSITE 数据库。该数据库收录由实验证实的已知的蛋白质序列中具有生物学重要意义的位点和序列模式，因此可以用来判断一个新蛋白可能具有的功能及其家族归属。

1990 年 04 月 10-13 日

第一届国际电泳、超级计算和人类基因组会议在美国佛罗里达州会议中心举行。尽管会议的名称并没有出现生物信息学这一名词，实际上生物信息学却是会议的主要部分。

1990 年 10 月

国际人类基因组计划启动，被誉为生命科学的“阿波罗登月计划”。

1992 年

Henikoff. S 和 Henikoff. J. G 在序列比较算法中引入之后被广泛应用的 BLOSUM 矩阵。

1993 年

中国人类基因组计划在国家自然科学基金资助下启动。

1993 年 3 月

欧洲生物信息学研究所（EBI）获准成立。

1993 年 8 月 1 日

专业蛋白质分析系统网络服务器诞生。

1994 年

国际生物信息学系列会议由 Cambridge Healthtech 研究所接管，并走向商业化和联机化

1994 年

澳大利亚 Macquarie 大学的 Marc Wilkins 和 Keith Williams 首先提出蛋白质组的概念（Proteome）。

1995 年 7 月的 Electrophoresis。

1994 年 06 月 01-04 日

第三届国际生物信息学和基因组研究会议在佛罗里达州会议中心举行。

1995 年

由于生物信息学日益普及，Cambridge Healthtech 研究所决定将国际生物信息学系列会议改成年会形式。

1995 年 4 月

日本信息生物学中心（CIB）成立。

1996 年

在教育部和科技部的支持下，中国北京大学蛋白质工程和植物遗传学工程国家实验室加入欧洲分子生物学网络（EMBnet）。

1997 年

澳大利亚生命科学研究学院的生物信息学研究组成立。研究分子生物信息学以及生物分化信息学。

1997 年 3 月

中国北京大学生物信息学中心成立。

1998 年

亚太生物信息学网络（APBioNet）成立。

1998 年

中国人类基因组研究北方中心（北京）和南方中心（上海）成立。

1998 年底

人类完成第一个多细胞生物——线虫的基因组全序列测定。

1998 年 2 月

生物信息学专业期刊——Comput.Appl.Biosci. 更名为 Bioinformatics。

1998 年 3 月 30 日

瑞士生物信息学研究所（SIB）成立。

1998 年 5 月

美国塞莱拉遗传公司成立，目标是到 2001 年绘制出完整的人体基因图谱，与国际人类基因组计划展开竞争。

1999 年

Prusiner 因发现引发疯牛病的朊病毒而获得诺贝尔生理/医学奖

1999 年 9 月

中国获准加入人类基因组计划（负责测定人类基因组全部序列的 1%—3 号染色体上的 3000 万个碱基对）成为第六个国际人类基因组计划参与国，也是参与该计划的唯一发展中国家。

1999 年 12 月 1 日

国际人类基因组计划联合研究小组宣布人类第一次获得一对完整人染色体——第 22 对染色体——的遗传序列。

2000 年 3 月

中国科学院上海生命科学研究院生物信息中心成立。

2000 年 3 月 14 日

美总统克林顿和英首相布莱尔针对某些私营生物技术公司为商业利益而试图为自己的研究成果申请专利而发表联合声明，呼吁公开人类基因组研究成果。

2000 年 4 月

我国科学家按照国际人类基因组计划的部署，完成了 1%人类基因组的工作框架图。

2000 年 5 月 8 日

德、日等国科学家宣布，他们已基本完成人体第 21 对染色体的测序工作。

2000 年 6 月 26 日

6 国合作、公众支持的国际人类基因组计划协作组在全球同一时间宣布已完成人类生命的蓝图——人类基因组的工作框架图。

。

（三）、生物信息学的主要研究内容

围绕着生物信息学的三个重要组成部分——基因组信息学、蛋白质的结构模拟以及药物设计，生物信息学的研究内容可分为:

一、 基因组相关信息的收集、储存、管理与提供

随着当前互联网提供的大量重要的生物学数据库及相关服务器，有关基因组相关数据库的发展相应受到研究者的广泛关注: 建立基因组信息的评估与检测系统、数据标准化、进行基因组信息的可视化和专家系统的研究、次级及专业数据库的发展、以因特网为基础的基因组信息学传输网络。

二、 新基因的发现、鉴定

新基因的确认、鉴定将为更好的了解与其相关的生理功能或疾病的本质提供依据，从而为新药的开发、设计奠定基础。目前利用 EST 序列信息寻找新基因成为国际基因争夺战的热点。包括通过计算分析从 EST（Expressed Sequence Tags）序列库中拼接出完整的新基因编码区，也就是通俗所说的“电子克隆”; 通过计算分析从基因组 DNA 序列中确定新基因编码区，经过多年的积累，已经形成许多分析方法，如根据编码区具有的独特序列特征、根据编码区与非编码区在碱基组成上的差异、根据高维分布的统计方法、根据神经网络方法、根据分形方法和根据密码学方法等。

****BLAST 简介

生物信息学计算的核心是序列比较，这包括同一个序列内不同片段的比较，以及两个或多个序列的对比。比较的内容，从序列的组分变化、寻找特殊的字段，到序列间字母的对应。比较的主要目的在于阐明序列之间的同源关系，以及从已知序列预测新序列的结构和功能。

两个或多个符号序列按字母比较，尽可能确切地反映它们之间的相似和相异，称为序列的比对或联配（alignment）.

我们先讨论序列联配算法所涉及的一些主要概念。核酸和蛋白质序列联配的前提是， 假定两个序列来自同一个祖先(“同源”)，它们在演化过程中由于变异的积累而成为不同的序列。作为符号序列看待，点变异包括字母的代换(substitution)、删除(deletion)和插入(insertion); 插入和删除统称为“插删”(indel). 两个序列联配时，往往要插入空位(gap)，以达到总体上更好的排列效果。 每当第一次

插入空位时， 要计一定的“罚分”(penalty)；连续插入空位时常按比例经以稍小的罚分，因此， 计算一组连续空位罚分的公式是 p=a+b xn，其中 n 是连续空位总数。两个常数 a 和 b 的值，与所比较的是核酸还是蛋白质序列有关，而且要同打分矩阵的选择和数值范围适应。

两个核酸序列的联配较为简单。序列中一个嘌呤被嘧啶代换或反之， 称为颠换(transversion)； 嘌呤或嘧啶互换称为置换(transition). 20 种氨基酸之间代换，远比核苷酸复杂。残基代换所引起的后果，与它们的具体物理化学性质有关。因此，对各种代换的效果，要有所估计，计算出各种打分矩阵。

生物序列中有重要功能的片段，往往比较保守，既变异的速率很低。序列的其他部分可能具有较高的变异速率，在演化过程中变得面目全非。例如，真核生物的 DNA 序列中，往往是比较少，比较短的保守片段，被甚为丰富的高变异区淹没。如果片面强调整体比对，可能会漏掉真正的同源序列。良好的局域比对往往会更有效的揭示同源关系。

绝大多数序列比对是针对蛋白质的。提交一条蛋白质序列，直接同蛋白质库里所有的序列对比，不需对序列再做什么变换。如果要把这条蛋白质序列，同数据库里的 DNA 序列比较，那就要把后者翻译成“蛋白质”。对于双链 DNA 的每个单链，因为翻译起始点的不同，要按照 3 个读框去翻译，一共得到 6 条供比较用的“蛋白质”序列。提交一条 DNA 序列，去同核酸数据库中的序列做比较，当然也无需交换。如果要同蛋白质序列库做比较，所提交的 DNA 序列也得按照 6 个读框翻译出来。像 BLAST 和 FASTA 这类通用程序，序列的变换都包含在其功能之内，用户只需提出要求。一般地说，翻译成蛋白质序列再进行比对，结果比较灵敏。

接下来介绍以下半经验的直观算法。假定已经选择好打分矩阵、空位罚分等参数，要求把一条给定的核酸或蛋白质序列，同数据库中所有现存序列进行联配，找出最相似的哪些序列，这是远非平庸的计算课题。如果进一步允许在联配时插入空位，计算难度就会空前增大。这首先是因为插入空位的位置和数目有大量可能的组合，一切靠“穷学”法挑出最佳方案的企图，都会超出现在和可以设想的未来的计算机能力。

从 20 世纪 80 年代以来，人们发展了一些半经验的直观算法，它

们可以相当快地给出较好的结果，但不能保证所得结果是最优的。BLAST 和 FASTA 就是很成功的实例。下面具体说说 BLAST。BLAST 是目前常用的数据库搜索程序，它是 Basic Local Alignment Search Tool 的缩写，意为“基本局部相似性比对搜索工具”[Altschul, 1990, 1997]。国际著名生物信息中心都提供基于网络的 BLAST 服务器。BLAST 算法的基本思路是首先找出检测序列和目标序列之间相似性最高的片段，并作为内核向两端延伸，以找出尽可能长的片段。BLAST 程序之所以使用广泛，主要因为其运行速度比 FastA 等其它数据库搜索工具快，而改进后的 BLAST 程序允许空位的插入。

我们极其简单地说明一下 BLAST 算法的基本思想。(如图 1)首先，BLAST 事先为数据库里的全部序列作了“索引”。它首先规定了一个字母串长度(在 FASTA 中相应参数为 WORD 或 ktup)，对 DNA 序列是 11，蛋白质序列是 6。把每个序列所含的此种串的类型作为索引。提交一个新序列时，也先对它做索引。只有索引类型兼容的库中序列才用来做比较。这样就大为减少了搜索工作量。其次，从局域联配得分最高的片段开始，向左右两端延伸，直到一端到头或总计分下降超过事先设置的值。然后再把这样得到的结果作比较，选出统计上最显著的哪些，排队输出。

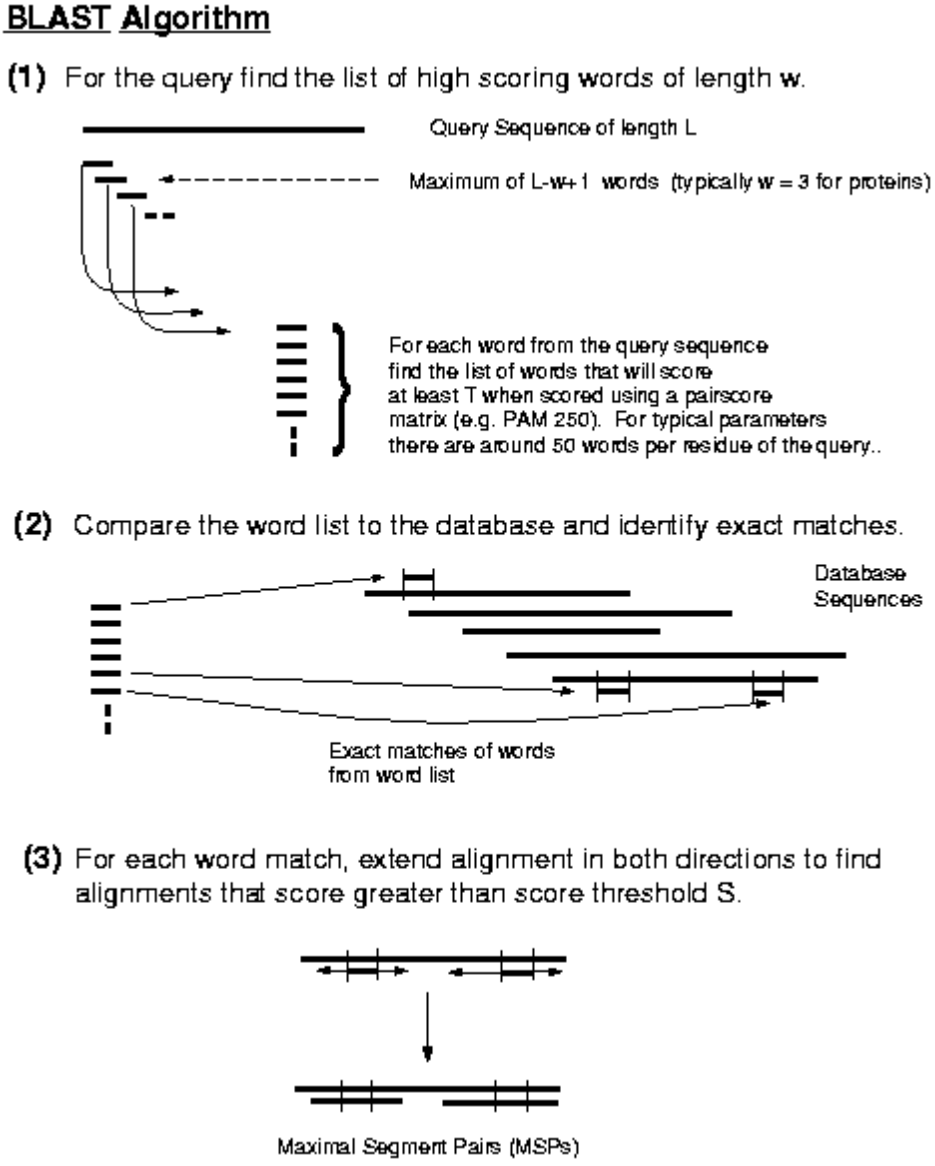

图 1. BLAST 的算法

BLAST 软件包实际上是综合在一起的一组程序，不仅可用于直接对蛋白质序列数据库和核酸序列数据库进行搜索，而且可以将检测序列翻译成蛋白质或将数据库翻译成蛋白质后再进行搜索，以提高搜索结果的灵敏度(表 1)。位置特异性叠代 BLAST (Position-Specific Iterated BLAST，简称 PSI-BLAST)则是对蛋白质序列数据库进行搜索的改进，其主要思想是通过多次叠代找出最佳结果。具体做法是利用第一次搜索结果构建位置特异性分数矩阵，并用于第二次的搜索，第二次搜索结果用于第三次搜索，依此类推，直到找出最佳搜索结果。此外，BLAST 不仅可用于检测序列对数据库的搜索，还可用于两个序列之间的比对。

表 1 BLAST 程序检测序列和数据库类型

程序名	检测序列	数据库类型	方　　法
Blastp	蛋白质	蛋白质	用检测序列蛋白质搜索蛋白质序列数据库
Blastn	核酸	核酸	用检测序列核酸搜索核酸序列数据库
Blastx	核酸	蛋白质	将核酸序列按 6 条链翻译成蛋白质序列后搜索蛋白质序列数据库
Tblastn	蛋白质	核酸	用检测序列蛋白质搜索由核酸序列数据库按 6 条链翻译成的蛋白质序列数据库
Tblastx	核酸	核酸	将核酸序列按 6 条链翻译成蛋白质序列后搜索由核酸序列数据库按 6 条链翻译成的蛋白质序列数据库

BLAST 程序是免费软件，可以从美国国家生物技术信息中心 NCBI 等文件下载服务器上获得，安装在本地计算机上，包括 UNIX 系统和 WINDOS 系统的各种版本。但必须有 BLAST 格式的数据库，可以从 NCBI 下载，也可以利用该系统提供的格式转换工具由其它格式的核酸或蛋白质序列数据库经转换后得到。对核酸序列数据库而言，不论用哪种方式，都需要很大的磁盘空间；而程序运行时，需要有较大的内存和较快的运算速度，才能用于日益增长的核酸序列数据库。对一般用户来说，目前常用的办法是通过 NCBI、EBI 等国际著名生物信息中心的 BLAST 服务器进行搜索。北京大学生物信息中心也提供了 BLAST 数据库搜索服务。需要说明的是，各生物信息中心 BLAST 用户界面有所不同，所提供的数据库也可能不完全相同，使用前最好先进行适当的选择。欧洲生物信息研究所 BLAST 服务器的用户界面(图 2)比较简洁，提供的数据库和参数均很多，用户可以根据不同要求，选择不同的数据库和各种参数。一般情况下，可以先按照系统给定的缺省参数进行初步搜索，对结果进行分析后再适当调整参数，如改变相似性矩阵、增加或减少空位罚分值、调节检测序列滑动窗口大小等。对于核酸序列和数据库，一般选择重复序列屏蔽功能，而对于蛋白质序列，通常不必选择重复序列屏蔽功能。

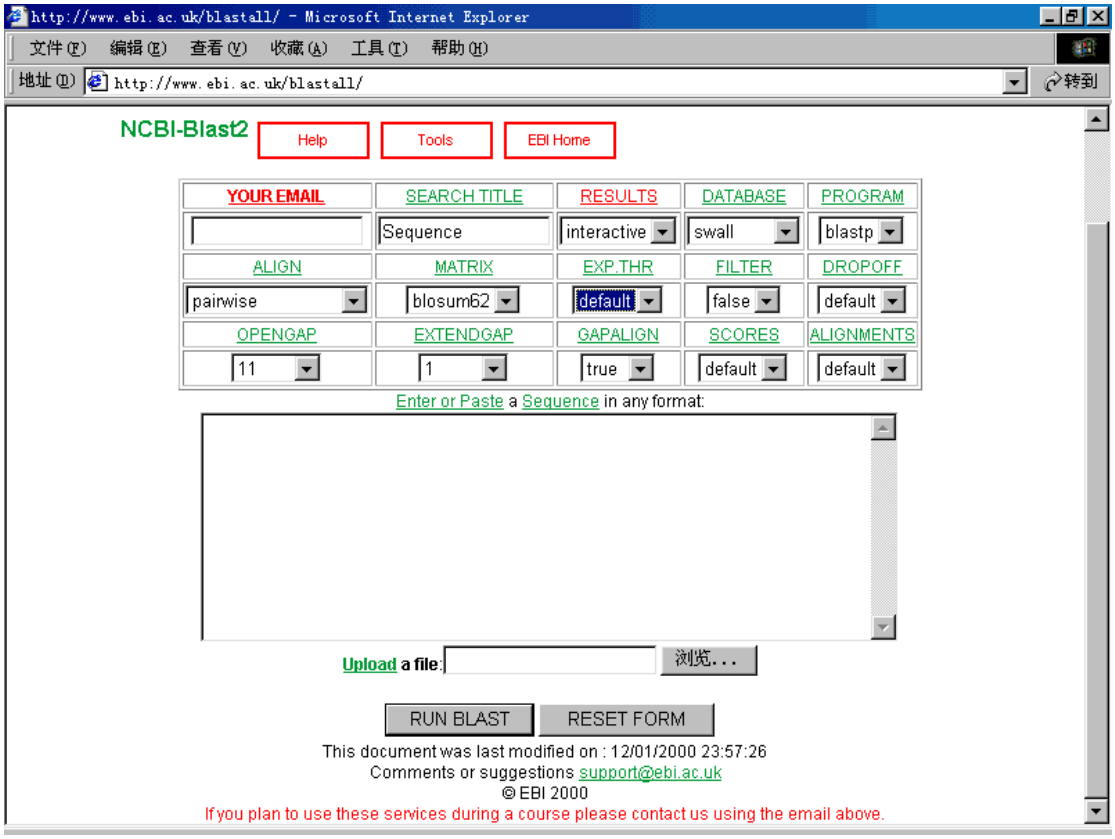

图 2. 欧洲生物信息学研究所的 BLAST 服务器的用户界面

图 3 是 BLAST 程序运行结果实例。这里，检测序列是与细胞凋亡有关的人自噬基因氨基酸序列，通过欧洲生物信息学研究所的 BLAST 服务器对包括 SwissProt 和 TrEMBL 数据库在内的蛋白质数据库进行搜索。输出结果中包括程序名称、版本号以及文献引用出处，以及检索序列的名称、数据库名称；结果中列出搜索到可能的同源序列的条目，以及它们在数据库中的编号和简要说明。每个条目后面给出相似性分数值 Score 和期望频率值 E，以相似性分数值大小为序排列，分数越高，相似性越大。而 E 值则表示随机匹配的可能性，E 值越大，随机匹配的可能性也越大。最后给出检测序列和目标序列同源性比对(图中只给出检测序列和一个目标序列的比对)。

图 3. BLAST 程序搜索结果实例

```
BLASTP 2.0.14 [Jun-29-2000]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro
A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman
(1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein
database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= O60875 (275 letters)

Database:  swall    531,777 sequences;  171,451,799  total
letters



Score     E
Sequences      producing      significant      alignments:
(bits)  Value
SWALL:O60875    O60875    APOPTOSIS    SPECIFIC    PROTEIN
(DJ134E15.2) ...    585  e-166
SWALL:Q9W3R7         Q9W3R7         CG1643         PROTEIN.
244  7e-64
SWALL:CAC03432 CAC03432 DJ354M18.1 (APG5 (AUTOPHAGY 5, S.
CE...    167  1e-40
SWALL:BAB10516 BAB10516 APG5 (AUTOPHAGY 5)-LIKE PROTEIN.
139  2e-32
SWALL:O74971 O74971 APOPTOSIS SPECIFIC PROTEIN HOMOLOGUE.
98  1e-19
SWALL:APG5_YEAST    Q12380    AUTOPHAGY    PROTEIN    APG5.
68  9e-11
SWALL:O01683 O01683 SIMILAR TO SINGLE-STRAND RECOGNITION
PRO...     33  3.2
SWALL:P90646     P90646     ABC     PROTEIN     (FRAGMENT).
33  4.2
SWALL:Q9M9A3                Q9M9A3                F27J15.20.
32  5.5
SWALL:O13957   O13957   HYPOTHETICAL   37.4   KDA   PROTEIN
C23H4.16C ...     32  9.5



>SWALL:Q9W3R7 Q9W3R7 CG1643 PROTEIN.

          Length = 306

 Score =  244 bits (617), Expect = 7e-64
```

三、 非编码区信息结构分析

非蛋白编码区约占人类基因组的 95%，其生物学意义目前尚不是很清楚，但从演化观点来看，其中必然蕴含着重要的生物学功能，由于它们并不编码蛋白，一般认为，它们的生物学功能可能体现在对基因表达的时空调控上。因此寻找非编码区的编码特征、信息调节及表达规律无疑将是未来相当长时间内的热点课题。对非蛋白编码区进行生物学意义分析的策略有两种，一种是基于已有的已经为实验证实的所有功能已知的 DNA 元件的序列特征，预测非蛋白编码区中可能含有的功能已知的 DNA 元件，从而预测其可能的生物学功能，并通过实验进行验证；另一种则是通过数理理论直接探索非蛋白编码区的新的未知的序列特征，并从理论上预测其可能的信息含义，最后同样通过实验验证。

四、 生物进化的研究

尽管已经在分子演化方面取得了许多重要的成就，但仅仅依靠某些基因或者分子的演化现象，就想阐明物种整体的演化历史似乎不太可靠。例如，智人与黑猩猩之间有 98%-99%的结构基因和蛋白质是相同的，然而表型上却具有如此巨大的差异，这就不能不使我们联想到形形色色千差万别的建筑楼群，它们的外观如此不同，但基础的部件组成却是几乎一样的，差别就在于这些基础部件的组织方式不同，这就提示我们基因组整体组织方式而不仅仅是个别基因在研究物种演化历史中的重要作用。由于基因组是物种所有遗传信息的储藏库，从根本上决定着物种个体的发育和生理，因此，从基因组整体结构组织和整体功能调节网络方面，结合相应的生理表征现象，进行基因组整体的演化研究，将是揭示物种真实演化历史的最佳途径。随着分子生物学的发展，特别是生物信息学时代的到来，使得分子进化的研究具备了极好的时机。

五、 完整基因组的比较研究

在后基因组时代，生物信息学家面对的不仅是序列和基因，更多的则是越来越多的完整基因组，有完整基因组研究导致的比较基因组学必将为后基因组研究开辟新的领域。

六、 基因组信息分析方法研究

为了更有效的发展生物信息学，发展相应的分析方法及手段是至关重要的。因此，发展有效的能够支持大尺度作图与测序需要的软件和数据库以及若干数据库工具，改进现有的理论分析方法，无疑将更好、更快的引导生物信息学的发展。

七、 大规模基因功能表达谱的分析

目前，基因组的研究已经从结构基因组逐渐过渡到功能基因组，因此获得基因的功能表达谱将存在于人类基因祖上的静的基因图谱，向时间、空间维上展开将是新阶段基因组研究的核心。为了获得基因表达的功能谱，国际上在核酸和蛋白质两个层次上均发展了新技术。在核酸层次上的新技术是 DNA 芯片，在蛋白质层次上则是二维凝胶电泳和测序质谱技术。

八、 蛋白质分子空间结构预测、模拟和分子设计

基因组和蛋白质组研究的迅猛发展，使许多新蛋白序列涌现出来，然而要想了解它们的功能，只有氨基酸序列是远远不够的，因为蛋白质的功能是通过其三维高级结构来执行的，而且蛋白质三维结构也不一定是静态的，在行使功能的过程中其结构也会相应的有所改变。因此，得到这些新蛋白的完整、精确和动态的三维结构就成为摆在我们面前的紧迫任务。目前除了通过诸如 X 射线晶体结构分析、多维核磁共振（NMR）波谱分析和电子显微镜二维晶体三维重构（电子晶体学，EC）等物理方法得到蛋白质三维结构之外，另外一种广泛使用的方法就是通过计算机辅助预测的方法，目前，一般认为蛋白质的折叠类型只有数百到数千种，远远小于蛋白质所具有的自由度数目，而且蛋白质的折叠类型与其氨基酸序列具有相关性，这样就有可能直接从蛋白质的氨基酸序列通过计算机辅助方法预测出蛋白质的三维结构。

我们知道，蛋白质分子是由 20 种不同的氨基酸通过共价键连接而成的线性多肽链，每一种蛋白质在天然条件下都有自己特定的空间结构。但以一定氨基酸顺序排列的多肽链是如何形成有一定空间结构的蛋白质分子的，也就是蛋白质结构的预测，仍是没有完全解决的问题。这里主要介绍一下蛋白质分子模拟与蛋白质结构预测的方法及新进展。

1. 蛋白质分子模型的建立与显示

分子模拟技术是利用计算机建立原子水平的分子模型来模拟分子的结构与行为，进而模拟分子体系的各种物理与化学性质。利用分子模拟技术结合计算机图形技术可以更形象、更直观地研究蛋白质等生物大分子的结构，蛋白质的空间结构的更清晰的表述和研究对揭示蛋白质的结构和功能的关系、总结蛋白质结构的规律、预测蛋白质肽链折叠和蛋白质结构等，都是有力的帮助和促进。

当前的分子模拟技术主要借助于先进的计算机图形工作站，通过友好的图形环境，使用者可利用鼠标极为方便地建立多肽、蛋白分子的初始模型。同时，也可以对已经被测定的生物大分子的三维结构进行显示，并对这些结构进行灵活方便的平移、旋转、放大及缩小等操作，分子模型的建立为下一步进行的分子模拟以及了解结构与功能的关系打下了基础。

2. 蛋白质结构预测

1）蛋白质结构预测的方法

蛋白质结构预测的目的是利用已知的一级序列来构建出蛋白质的立体结构模型，对蛋白质进行结构预测需要具体问题具体分析，在不同的已知条件下对于不同的蛋白质采取不同的策略，目前预测蛋白质空间结构的方法可以分为两大类:

a.分子动力学方法

这类方法采用分子力学、分子动力学的方法，根据物理化学的基本原理，从理论上计算蛋白质分子的空间结构，这类理论计算方法依据一个基本热力学假定: 一个蛋白质分子的溶液中的天然构象相应于热力学上最稳定、自由能最低的构象，但这一方法目前存在着三个主要问题，首先，用以描述蛋白质一溶剂系统工程力场和能量函数还和于半定量阶段；其次，数学上还没有有效方法解决能量极小化问题；第三，目前并没有证据证明蛋白质的天然构象就是全局自由能最小的构象。

b.基于知识的预测方法

这类方法通过对已知空间结构的蛋白质进行研究和分析，找出蛋白质一级结构和空间之间的联系，总结出一定的规律并建立一些经验规则。这类方法已经被成功地应用于同源蛋白质空间结构预测的研究。然而对于同源性低的和非同源蛋白质分子来说，由于受二级结构预测精度的限制，这种方法只取得了非常有限的成功。

通过对大量已知空间结构的蛋白质分子的研究和分析，发现一条多肽链可能采取的构象的数目是相当大的，但在蛋白质分子中由二级结构预测是解决从蛋白质的一级结构预测其空间结构这一问题的关键步骤，现有的预测方法都假定蛋白质的二级结构主要由邻近残基因的短程相互作用所决定的，然后通过对一些已知空间结构的蛋白质分子进行分析、归纳，制定出一套预测规则并根据这些规则对其他已知或未知结构的蛋白质分子的二级结构进行预测，目前常用的方法有：基于单残基统计的 Chou-Fasman 方法，基于信息论和统计的 Garnier 方法，Lim 方法，人工神经网络方法等。据一些检验结果，上述几种方法的预测率分别为 50%、56%、59%和 64%，而现在一般认为二级结构的预测准确率如果达到 80%的话，我们就可以基本准确地预测一个蛋白质分子的三维空间结构，因此进一步提高蛋白质二级结构预测的精度是当务之急。

2） 蛋白质二级结构预测新进展

由于蛋白质二级结构预测方法中第一类方法在数学上遇到难以解决的多重极小值问题，而第二类方法又受到预测精度的限制，近年来一些科学家提出了一种预测蛋白质三维结构的新策略，这类方法被称为 Threading 方法或折叠类型识别方法，这一方法的基本思想是假定被预测蛋白质的折叠类型与某一已知结构的蛋白质的折叠类型相同，这样，蛋白质结构预测的问题就转变为在已知空间结构的蛋白质中，选取一种被预测序列最可能采取的折叠类型，从而大大减少了预测蛋白质结构的难度，这一方法已经成功地预测了一些蛋白质的空间结构。除了 Threading 方法外，近年来，国际上一些研究组还发展了一些从蛋白质的一级结构直接预测蛋白质空间结构的新方法。这些方法的基本思想是将基于知识的方法与计算化学以及统计物理学的方法相结合，采用简化的蛋白质模型和根据已知结构的蛋白质所导出的平均势场，从理论上计算蛋白质的空间结构。这些方法不仅可以从蛋白质的一级结构直接预测蛋白质的三维结构，而且可以在计算机上模拟

蛋白质分子折叠的全过程。目前，还有一些新方法如遗传算法、模拟退火、多维统计、模糊集合论方法等在蛋白质结构预测中的应用也正在研究中。通过对一些简单蛋白质分子的模拟研究，这些新方法已经显示出很强有力的生命力，许多权威人士推测，随着这些新方法的进一步改进和完善，在今后 10 年内，蛋白质折叠这一分子生物学中的难题将有望得到解决。

3、蛋白质分子模拟软件

随着分子模拟技术的飞速发展，逐步形成了一些商品化的软件。应用于生物大分子领域的商品化分子模拟软件的主要有美国 MSI 公司的 Insight II 软件和 Quanta 软件，以及 Tripos 公司的 Sybyl 软件；在国内，北京大学物理化学研究所也开发了一套“北京大学蛋白质分子设计系统”。这些商品化软件在不断的变化和发展中，有些软件模块，每年都更新版本，不断完善这些软件的功能。

九、 药物设计

近年来随着结构生物学的发展，相当数量的蛋白质以及一些核酸、多糖的三维结构获得精确测定，基于生物大分子结构知识的药物设计成为当前的热点。生物信息学的研究不仅可提供生物大分子空间结构的信息，还能提供电子结构的信息，如能级、表面电荷分布、分子轨道相互作用等以及动力学行为的信息，如生物化学反应中的能量变化、电荷转移、构象变化等。理论模拟还可研究包括生物分子及其周围环境的复杂体系和生物分子的量子效应。传统的药物研制主要是从大量的天然产物，如动物、植物、微生物和合成有机、无机化合物以及矿物中进行筛选。得到一个可供临床使和的药物要耗费大量的时间与金钱。近年来由于生物信息学的发展，相当数量的蛋白质以及一些核酸、糖类三维结构已被人们精确测定，使得基于蛋白质和核酸结构的药物设计成为可能。这种设计途经与蛋白质工程有着深刻和广泛的联系。

1、蛋白质改性和分子设计

原则上，任何有功能的蛋白质都可以作为蛋白质工程的改造对象，

但实际上选择目标时往往要考虑以下几个方面：改性对象有没有测出空间结构；结构和生物功能的联系是否明确；所选对象的重要性；是否易于进行分子设计和最后的基因工程生产。当前的分子设计主要以能显示图形图象的计算机为工具，在了解了需要改造的蛋白质的性能及其相应的结构基础后，采用基于物理学原理的各种模拟方法以和基于蛋白质分子结构知识的模型构建方法，提出蛋白质改性的设计方案。这一方案将提供改性后的蛋白质哪些部分的氨基酸顺序与天然蛋白质不同，这些新的序列可以导致怎样的空间结构和电子结构的变化，从而能赋予新蛋白质什么样的特性。

2、基于生物大分子结构的药物设计

要了解层白质的功能找到其致病的分子基础，只有氨基酸顺序是不够的，必须知道它们的三维结构。要设计药物治疗这些疾患也需要了解蛋白质的三维结构，目前的 X 射线晶体学技术、多维核磁共振波谱学技术等测定蛋白质空间结构的方法还不能很好的满足研究需要。因此，生物信息学中的理论模拟与结构预测就显示了重要性。模拟的结果对于在分子、亚分子和电子结构层次上了解生命现象的基本过程具有重要意义，为天然生物大分子的改性和基于受体结构的药物分子了设计提供的依据。

1） 基于受体结构的药物筛选

药物的治疗作用主要是通过药物与受体的相互作用。在生物体中受体多半是生物大分子，像蛋白质和核酸，而以蛋白质居多。如果人们了解了受体蛋白的结构，就可以根据其结构来研究药物是怎样改变它的构象、进而产生治疗作用的。设想各种可能的药物结构来模拟这种相互作用应有助于找到较好的药物，这是当前药物筛选中的一个合理的和有效的途径。目前已有很多生物大分子作为药物设计的受体模型，例如：基于酶结构的药物设计，基于抗体结构的药物设计，基于致癌、抑癌基因表达产物的药物设计，基于细胞表面受体结构的药物设计，基于转录因子结构的药物设计。近年来随着人类基因组计划的进一步进行、化合物合成技术的进步和一些先进技术的使用已使受体药物筛选发展成为高通量筛选。利用生物信息学技术所建立的化合物库是筛选化合物的重要来源。

2） 药物设计与分子模拟

近年来，分子图形学在药物研究的广泛应用，使得以结构为基础的药物设计与计算机分子模拟密不可分。

i. 分子模拟在药物设计中的应用

分子模拟可以应用于药物开发的众多环节中，它除了可以用于确定生物大分子三维结构外，在寻找先导化合物，优化先导化合物方面更能大显身手。例如，从三维结构数据库中搜索能嵌合到靶点的分子的方法已经用于先导化合物的发现；以结构为基础的计算机全新药物设计已经用于设计先导化合物。

ii. 先导化合物的发现

a、三维数据库搜索

三维数据库搜索发现，先导化合物就是利用计算机在含有大量化合物三维结构的数据库中，搜索能与生物大分子靶点匹配的化合物，或者搜索能与结合药效团相符的化合物，利用这种方法，可以快速地确定具有假定药效因的化合物。这种搜索也常常提供一些意想不到的化合物，或发现老药的新用途，目前实现这类搜索的软件很多如 MSI 的 Catalyst 软件包等。在已知靶点结构的情况下，采用 DOCK 程序，能对搜索到的化合物再次进行优化筛选，减少候选化合物的数目，以便采购或合成以及进行生物活性测试。

b、全新药物设计

新的先导化合物可以在工作站上通过配体与靶点相互作用而设计，靶点表面的分子图形显示有助于完成这一工作。但是在没有构建和优化假想结构的情况下，很难用这种方法设计出全新的结构。采用全自动计算机全新药物设计可以有效地完成这项任务，全新药物设计又称为从头设计，它根据受体部位的形状、性质要求，让计算机自动构建出形状、性质互补的新分子，熟知的全新药物设计软件有 Ludi,Leapfrg 等。这种方法能产生大量的独特结构，采用全新药物设计先导化合物已不乏成功之例。

iii. 先导化合物的优化

通过优化先导化合物来提高其与靶点的匹配性。小范围匹配性的改进，通常可通过在高分辨图形工作站上简单地观察化合物的嵌合情况而实现。在这种情况下采用网状形式来显示溶剂化表面往往有助于问题的解决，而表面情况、化学匹配情况可用颜色显示出来。当嵌合的配体结构受到制约或具有高度柔韧性时，优化的一个目标是适当减

小制约或增加分子的刚性。这些变化所带来的影响，可通过计算机对起始化合物和修饰物的构象分析进行预测。分子力学适应于这项研究，但在力场参数缺乏的情况下，在工作站上也能方便地进行半经验或从头计算的量子力学计算。通过对先导化合物与靶点的嵌合结构的分子模拟研究后，可将优化选择在（1）分子与活剂接触部分的效价中性部位（2）分子中与靶点结合不重要的部位。

3、药物设计中理论方法

在药物的分子设计中，理论方法起了非常重要的作用。从原理上说它涉及了原子分子物理，凝聚态物理、量子力学和统计物理等物理学科在处理多粒子问题时的成就和进展，也应用了量子化学、计算化学和计算技术的原理和方法。但是对生物分子这样大的体系，为了计算的可行，目前还只能作各种的必要近似。在大多数分子设计的计算中所使用的物理模型是经验势函数，即把体系的总相互作用看做是组成该体系两两原子相互作用的总和；原子与原子之间的相互作用又简化为各种简单相互作用，如键伸缩能、键角扭曲能、二面角扭曲能、范德华相互作用、静电能等的累加，基于这样简单的模型通常使用的理论方法有：

1）、分子动力学方法（MD）

2）、Monte-Carlo 方法（MC）

3）、自由能微方法（FEP）

4）、模拟退火技术等。

在分子设计的某些合适情况下，也可使用较为精确的物理模型，包括：量子化学从以计算（ab initio）方法，半经验的量子化学计算方法以及密度泛函理论等。近几年来逐渐发展了量子力学与分子动力学相接合的方法，分子力学和分子动力学的力场不能描述键的形成和断裂、无法研究反应过程，量子力学方法可以研究这一切，如键的形成与断裂、电子转移等，但耗费巨大。把两者结合起来就能克服它们的困难。随着计算机和分子图形学技术的进步，使得考虑生物活性分子与受体结合时的三维结构性质为特征的 3D-QSAR（三维定量构效关系）方法取得了较大的发展，其中通过研究药物与受体非共价相互

作用时的静电场和立体场来探寻药物分子性质的比较分子力场分析法（CoMFA）应用得最广且最为成功。

（四）、展望

综上所述，我们不难看出，作为计算机科学和分子生物学等形成的交叉科学，生物信息学已经成为基因组研究中必不可少的有力研究手段

（1）理论研究。任何学科的发展都离不开基础理论的研究，生物信息学也不例外。它对许多学科都提出了巨大的挑战。这些学科包括分子进化遗传学、群体遗传学、统计生物学、基因组学以及计算机科学和应用数学的相关学科。如果基础理论研究得不到应有的发展，生物信息学的发展将受到严重的阻碍。

（2）软件的重用和说明。现在虽然已经开发出大量的软件工具，但是大多数软件缺乏技术细节的描达，使得新软件编制时不能很好地利用已有的软件资源，不得不从头开始，造成各种软件都有自己的输入输出格式，相互之间互不通用。同时，大量软件的出现带来一个新问题，即生物学家面对数量众多的软件无从选择。这两个问题的解决需要对各种软件的功能特性和技术细节进行详尽的介绍，并进行比较。这样的话，新软件的编制者可以避免一些编程的重复劳动，甚至直接利用已有的程序模块，并且可以编制已有软件输出格式的接口，统一输入输出的格式，用户也可以方便地选择合适的软件。

（3）集成数据库。公共数据库与因特网相连，为世界各地的科学家提供快速高效的服务，因而成为获取生物学数据的最佳媒介。目前，国际上著名的公共数据库有 Genebank、EMBL、DDBJ、Swiss-Port、PIR、PDB 等。

（4）生物数据的质量监控。监控已有的生物数据究竟具有多大的可信度，对于物理图谱的构建工作将有十分重大的意义。