生物信息学及其研究现状*

田 云, 卢向阳, 彭丽莎, 徐 锋

(湖南农业大学 理学院, 中国湖南 长沙 410128)

摘 要: 生物信息学是采用计算机技术和信息论方法研究生命科学中各种生物信息的表达、采集、储存、传递、检索、分析和解读的科学, 是现代生命科学与信息科学、计算机科学、数学、统计学、物理学、化学等学科相互渗透而高度交叉形成的一门新兴前沿学科. 对生物信息学的起源、研究内容、研究热点及应用等进行了综述, 并展望了其今后的发展前景.

关键词: 生物信息学; 蛋白质组学; 生物芯片; 生物计算机; 生物学数据库; 比较基因组学; 功能基因组学
中图分类号: Q811 文献标识码: A 文章编号: 1007-7847(2002) S1-0153-06

Bioinformatics and Its Research Current Status

TIAN Yun, LU Xiang-yang, PENG Li-sha, XU Feng

(*College of Sciences, Hunan Agriculture University, Changsha* 410128, *Hunan, China*)

Abstract: Bioinformatics is an interdisciple science developed by the interaction of modern biology, informatics, computer science, mathematics, statistics, physics and chemistry. It studies the collection, storage, transference, search, analysis and translation of various biological information. The main research areas and recent accomplishments of bioinformatics are briefly introduced.

Key words: bioinformatics; proteomics; biochip; bio-computer; bio-databases; comparative genomics; functional genomics

(*Life Science Research*, 2002, 6(Suppl): 153~ 158)

1 生物信息学的起源

从 20 世纪 90 年代以来, 随着各种生物基因组测序计划的展开与分子结构测定技术的突破以及 Internet 的普及, 无数的生物学数据如雨后春笋般迅速涌现. 2001 年 2 月 12 日, 美国 Celera 公司与美国国家人类基因组计划分别在 Science 和 Nature 上公布了人类基因组的精细图谱及其初步分析结果[1,2]. 今年 4 月 5 日出版的 Science 杂志又把水稻基因组的序列框架图公布出来[3,4]. 8 月 23 日出版的 Science 杂志公布了河豚的全基因组序列[5]. 到目前为止, 已经测出了上百种生物体的完整基因组序列. 如何分析这些从实验过程中获得的大量原始数据, 并从中获得与生物结构、功能相关的有用信息是当前困扰理论生物学家的一个棘手问题. 生物信息学(Bioinformatics)就是在此背景下发展起来的综合运用生物学、数学、统计学、物理学、化学、信息科学以及计算机科学等诸多学科的理论方法而形成的一门崭新交叉学科[6].

生物信息学这一名词在 1991 年左右才在文献中出现, 而事实上, 生物信息学的存在已经有 30 多年的历史了, 只是最初人们称之为基因组信息学[7]. 虽然基因组信息量在生物总信息量中占有极大的比重, 但是, 生物信息并不仅限于基因组信息, 生物信息学也并不等同于基因组信息学. 目

* 收稿日期: 2002-11-06

作者简介: 田云(1979-), 男, 湖南沅江人, 硕士研究生, 从事生物化学与分子生物学研究. E-mail: tianyun@hunau.net

前,我们普遍认为生物信息学是把基因组 DNA 序列信息分析作为源头,破译隐藏在 DNA 序列中的遗传语言,找到代表蛋白质和 DNA 基因的编码区,特别是阐明非编码区的实质,从而认识生物有机体代谢、发育、分化和进化的规律;同时在发现了新基因信息之后进行蛋白质空间结构的模拟和预测,然后依据特定蛋白质的功能进行必要的药物设计.因此,现代生物信息学主要包括 3 个重要内容,它们分别是基因组信息学、蛋白质的结构模拟以及药物设计.其中,基因组信息学是它的源头和基础.

生物信息学的研究目标是揭示"基因组信息结构的复杂性及遗传语言的根本规律".它是本世纪自然科学与技术科学领域中"基因组"、"信息结构"和"复杂性"这 3 大科学问题的有机结合.该项目的实施不仅有助于人们认识遗传语言,认识人类本身,而且必将有助于揭示"信息结构"和"复杂性"的深刻内涵,以及遗传、发育和进化的联系,大大丰富和发展现有的物理学、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,从而推动整个学科群的发展,成为自然科学中多学科交叉的有活力的、有影响的新领域.事实上,生物信息学是一门理论与实践并重的学科.

2 主要研究内容

2.1 基因组学研究

基因组(genome)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.关于基因组的研究称之为基因组学(Genomics),基因组学根据研究重点的不同可以分为序列基因组学(Sequence genomics)、结构基因组学(Structural genomics)、功能基因组学(Functional genomics)与比较基因组学(Comparative genomics).人类基因组计划(HGP)于 1990 年正式启动,计划在 15 年内提供 30 亿美元,至 2005 年完成人类基因组全部序列的测定.2001 年 2 月 12 日,人类基因组的精细图谱已经被公布在 Science 和 Nature 杂志上[1,2].近年来,各种模式生物的基因组测序工作也已完成.2002 年 4 月 14 日~ 17 日在中国上海召开了第 7 次国际人类基因组大会,本届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.

2.2 蛋白质组学研究

蛋白质组(proteome)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(Proteomics).蛋白质组学的核心内容包括蛋白质组研究体系的建立、完善和与重要生物学问题有关的功能蛋白质组研究两个部分.而蛋白质信息学则涉及蛋白质数据库的建立、相关软件的应用与开发,并进而开展重要蛋白质的结构预测、三维结构和动态结构研究,在蛋白质组水平上深入探索其作用模式、功能机理、调节控制及其与蛋白质群体内或与相关生物大分子间的相互作用.最近,发展了几种计算机识别蛋白质功能的新方法,这些方法的原理是根据相同特征的蛋白质之间具有功能上的关联或直接作用.如系统发生模式、mRNA 表达模式和结构域融合模式等[8].今年,我国已启动了重大疾病蛋白质组的研究计划,将我国蛋白质组的研究推向一个新的阶段.

2.3 生物芯片

生物芯片(Biochip)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、糖类及其它生物组分的准确、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(gene chip)或 DNA 芯片(DNA chip)、PNA 芯片(PNA chip)、蛋白质芯片(protein chip)和芯片实验室(Lab-on-a-chip)等.而从其功能的角度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、司法鉴定和军事等各个方面.生物芯片的整个检测与分析技术环节都属于生物信息学的研究领域.

2.4 生物计算机

生物计算机(Bio-computer)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(DNA computer)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,自它出现以来短短 9 年时间就取得了较大的进展.其基本设想是: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,作为

实现运算的过程, 即以反应前的 DNA 序列作为输入的数据, 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的速度. DNA 计算机毕竟还只是一种理论设想, 许多方面都还很不成熟, 主要表现在构造的现实性及计算潜力、运算过程中的错误问题与人机界面等. 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, 向众多的领域提出了挑战[9, 10].

2.5　生物学数据库

近年来随着大量生物学实验数据的积累, 无数的生物学数据库(Biology databases) (见表 1) 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、数据处理的服务等. 现阶段, 数据库的类型几乎覆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 等, 蛋白质序列数据库有 SWISS-PROT、PID、OWL、ISSD 等, 蛋白质片段数据库有 PROSITE、BLOCKS、PRINTS 等, 三维结构数据库有 PDB、NDB、BisMagResBank、CCSD 等, 与蛋白质结构有关的数据库还有 SCOP、CATH、FSSP 等, 与基因组有关的数据库还有 ESTdb、OMIM、GDB、GSDB 等, 文献数据库有 Medline、Uncover 等. 还有一些公司开发了商业数据库, 如 MDL 等. 另外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(Sequence Retrieval System) 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, 这样用户可以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, 不过, 数据库的管理和使用却越来越简捷, 现在大多数数据库能实现自动投送数据、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(CBI) 所建立的数据库和服务项目在国内是最多的, 我国在数据库的研究中起步很晚, 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.

表 1　国际互联网上一些重要的生物信息学数据库

Fig. 1　Some important biological information databases on the Internet

数据库名	国际互联网址(URL)	数据库内容
EMBL	http://www.ebi.ac.uk/embl/	基因组数据核酸序列
GenBank	http://www.ncbi.nlm.nih.gov/	基因组数据核酸序列
DDBJ	http://www.ddbj.nig.ac.jp/	基因组数据核酸序列
GDB	http://www.gdb.org/	人类基因及基因组图谱
HuGeMap	http://www.infobiogen.fr/services/Hugemap/	人类基因组遗传和物理图谱
PIR	http://pir.georgetown.edu/	蛋白质序列
SWISS-PROT	http://www.ebi.ac.uk/swissprot/	蛋白质序列
PROSITE	http://www.expasy.ch/prosite/	蛋白质功能位点
PDB	http://www.rcsb.org/pdb/	蛋白质三维空间结构
SCOP	http://scop.mre-lmb.cam.ac.uk/scop/	蛋白质结构
COG	http://www.ncbi.nlm.nih.gov/COG/	蛋白质直系同源簇数据库
KEGG	http://www.genome.ad.jp/kegg/	功能数据库
DIP	http://dip.doe-mbi.ucla.edu/	蛋白质相互作用数据库
ASDB	http://cbcg.nersc.gov/asdb/	可变剪接数据库
TRRD	http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4/	转录调控区数据库
TRANSFAC	http://transfac.gdf.de/TRANSFAC/	转录因子数据库
GOBASE	http://megasun.bch.umontreal.ca/gobase/	细胞器基因组数据库
AtDB	http://genome-www.stanford.edu/Arabidopsis/	拟南芥基因组数据库
INE	http://www.staff.or.jp/giot/INE.html/	水稻基因组数据库
SGD	http://genome-www.stanford.edu/Saccharomyces/	酵母基因组数据库
DBCat	http://www.infobiogen.fr/services/dbcat/	生物信息数据库目录数据库
CBI	http://cbi.pku.edu.cn/	北京大学生物信息中心

2.6 分子进化及生物应用软件的研究

分子进化钟的发现与中性理论的提出, 极大的推动了分子进化的研究, 并建立了一套依赖于核酸、蛋白质序列信息的理论方法. 从各种基因结构与成分的进化、密码子的使用到进化树的构建等, 各种理论上和实验上的课题都有待生物信息学家的研究[11].

预测生物大分子的空间结构需要大量的生物计算, 计算内容包括序列的分析比较、分子结构及其可视化、基因的模式识别等等. 现在虽然已经开发出了大量的生物学应用软件, 但大多数软件缺乏技术细节的描述, 使生物学家面对数量众多的软件却无从选择. 所有这些问题的解决需要新软件的编制者统一输入输出格式, 这样用户才可以方便地选择合适的软件.

3 当前研究热点

3.1 新知识的发现

数据库中的知识发现(Knowledge Discovery in Database, KDD) 是生物信息学的一个新的研究方向. 所谓 KDD, 就是一个挖掘数据库中有效的、新颖的、潜在有用的和最终可理解模式的复杂过程.

3.1.1 新基因和新 SNPs 的发现与鉴定

各种生物有机体的基因组工作草图相继完成, 因此, 目前的当务之急是从复杂的基因序列中发现新基因, 发现一个新的基因就能了解与其相关的生理功能或疾病的本质, 从而为新药的开发、设计奠定基础. 使用 KDD 的过程是目前发现新基因的重要手段.

利用 EST 数据库(dbEST) 发现新基因: 该方法也被称为基因的电脑克隆, EST 序列(Expressed Sequence Tags) 是从基因表达的短 e-DNA 序列, 它们携带着完整基因某些片段的信息. 现在, GenBank 的 EST 数据库中人类 EST 序列将达到 400 万条, 它大约覆盖了人类基因的 90% 以上, 由于 EST 序列中包括了大量未发现的人类基因的信息, 如何利用这些信息发现新基因成了近几年的重要研究课题.

从基因组 DNA 序列中预测新 ORF: 从基因组 DNA 预测新基因, 是发现新基因的另一个重要途径. 它除依据同源性与包含已知的数据库进行比较外, 经典的方法还分为两类, 一类是基于编码区所具有的独特信号, 比如起始密码子、终止密码子等; 另一类是基于编码区的碱基组成. 近年来, 随着生物信息学的发展, 出现了一批确定编码区的新方法, 如考虑高维分布的统计方法、神经网络方法、分形方法等, 这些方法都侧重于生物数据的处理, 模型的建立, 以及用计算机技术、数学方法等手段去探索已知数据的规律. 另外, 将密码学方法用于识别编码区, 也取得了较好的效果.

当人类找到了自己的基因之后, 要解决的问题将是不同的人之间的基因差别, 这就是我们通常所说的 SNPs(单核苷酸多态性). 构建 SNPs 及其相关数据库是基因组研究走向应用的重要步骤. 这主要是因为 SNPs 将提供一个强有力的工具, 用于高危群体的发现、疾病相关基因的鉴定、药物的设计和测试以及生物学的基础研究等. 1998 年, 国际上已经开展了以 EST 为主发现新 SNPs 的研究. 由于我国是一个多民族大融合的国家, 因此, 开展中华民族 SNPs 的研究也显得至关重要.

3.1.2 非编码区的结构、功能分析

非编码区("JUNK" DNA) 在人类基因组中占有很大一部分(约 98%). 研究表明"JUNK" 是许多生命过程富有活力的不同类型的 DNA 复合体, 现在对于它的作用人们还不清楚, 但从生物进化的观点来看, 这部分序列必定具有重要的生物功能. 当前的认识是, 它们与基因在四维时空的表达调控有关. 另外, 生物信息学家现在还要寻找新的非三联体密码子的编码方式. 寻找这些区域的编码特征、信息调节与表达是未来相当长时间内的热点课题, 而 KDD 过程是在这一热点中比较容易取得突破的方法.

3.1.3 完整基因组的比较研究

随着完整基因组的数据越来越多, 人们开始在基因组水平对若干重大生物学问题如生命的起源和进化等进行分析研究. 举例来说, 鼠和人的基因组大小相似, 都含有约 30 亿碱基对, 基因的数目也类似, 且大部分都具有同源性, 可是鼠和人的差异却如此之大? 同样, 科学家估计不同人种间基因组的差别仅为 0.1%, 人猿间的差别约为 1%, 但是它们之间的表型差异却十分显著. 这种差异不仅要从基因、DNA 序列找原因, 更应在整个基因组、染色体组织上的差异来找原因. 这一工作也就开创了比较基因组学. 科学家通过对几个完整基因组的比较, 统计出维持生命活动所需要的最少基因的个数为 250 个左右. 同样, 人们通过对人和鼠的基因组进行比较发现, 尽管两者基因

组的大小和基因数目类似, 但基因组的组织差别却很大. 例如存在于鼠 1 号染色体的基因已分布到人的 1、2、5、6、8、13、18 号 7 个染色体上; 存在于鼠 16 号染色体上的基因分布到人的 3、8、12、16、21、22 号 6 个染色体上[12]. 我国在基因组的序列测定及信息分析等方面已经跻身世界前列, 并获得了大量的原始实验数据, 这些数据将为我国在比较基因组学这一领域的研究提供最直接的素材.

3.1.4　基因功能表达谱的分析

虽然人们已经获得了许多生物有机体的完整基因图谱, 但我们还不知道这些基因是如何发挥它们的功能的, 或者说它们是如何按照特定的时间、空间进行基因表达以及表达的数量是多少. 许多实验研究表明, 在不同的组织中表达基因的数目差别是很大的, 其中脑中基因表达的数目最多, 约有上万个, 有的组织则只有几十或几百个基因表达. 不知道每种组织中表达基因的数目以及每个基因的表达量, 就无法从分子水平了解该组织在生命活动中的功能. 并且同一组织在不同的个体生长发育阶段表达基因的种类、数量也是不同的, 有些基因是在幼年时期表达, 有些则是在中年阶段表达, 更有一些需要到老年时期才能表达. 不知道伴随着生物的生长发育, 基因表达状况的变更, 也就无法确切的说明生命的过程. 特别是现在关于干细胞功能表达谱的研究, 不仅与医疗实践关系密切, 而且也是研究发育生物学、进化生物学极好的模型. 这些结果都说明人们需要从基因组的静的基因图谱向时间、空间上展开, 也就是说在不同时间、不同组织中基因的表达谱研究, 即我们通常所说的功能基因组研究.

为了得到基因的表达谱, 国际上在核酸和蛋白质两个层次上都发展了新技术. 这就是在核酸层次上的基因芯片(DNA 芯片) 技术和在蛋白质层次上的大规模蛋白质分离和序列鉴定技术, 即蛋白质组技术. 无论是哪个方面的技术的发展, 都强烈地依赖于生物信息学的理论、技术和数据库. 如 Jaccoud 等以水稻基因组为模型, 用 cDNA 微点阵分析 DNA 的多态性, 采用差异排列技术高通量的分析基因组中来自不同器官的总 DNA, 然后通过数据库分类比较产生各个组织的遗传指纹, 从而可以确定提供 DNA 的器官和组织. 下一步功能基因组的研究将朝着复杂系统的方向发展, 也就是探讨生物系统中各部分、各层次的相互作用, 即基因表达的网络问题, 从而进入系统生物学的领域.

3.2　蛋白质的结构、功能预测

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需要采取不同的策略. 目前, 预测蛋白质空间结构的方法主要分为两类: 一类是分子动力学方法; 另一类是基于知识的预测方法(又称为同源建模) , 该方法主要是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结构之间的关系, 总结出一定的规律并建立一些经验规则. 该方法已经成功的应用于同源蛋白质空间结构的预测研究. 这类方法中目前常用的主要有基于单基因的 Chou-Fasman 方法, 基于信息论和统计学的 Garnier 方法、Lim 方法、人工神经网络方法等. 近年来, 科学家提出了一种预测蛋白质空间结构的新策略, 称之为 Threading 方法或折叠类型识别方法. 另外, 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之中. 但是, 所有这些方法都存在其严重的缺陷, 可以想象, 如果该过程一旦取得突破, 将是生物学中一个里碑程, 即第二套生物学遗传密码破译.

当人们获得含有能够编码蛋白质的完整 DNA 序列后, 就需要分析所表达的蛋白质的功能, 尤其是一些与已知 DNA 序列无同源性或同源性很低的. 这时, 我们可以利用生物信息学技术, 通过与已知蛋白质相比较来判定未知蛋白质的功能, 此外, 蛋白质的一些理化性质(如疏水性、跨膜螺旋等) 也可以由序列直接计算得到, 其主要依据是以下两个方面: 一是所获得的序列是否与已知蛋白质结构相似; 二是所获得的序列是否含有特殊蛋白质家族或功能的保守残基. 目前, 我们主要利用 BLAST (basic local alignment search tool) 和 FASTA 工具将已知序列与蛋白质序列库中的序列进行同源性比较来进行蛋白质功能的预测.

生物信息学对蛋白质结构和功能预测的研究都是建立在已经获得的分子生物学知识和原始实验数据基础之上的, 它是对以往理论知识和实验结果的充分而有效的运用并做合理的推论, 因此, 很有可能存在着差错, 仍然需要进一步的实验室工作来进行验证和补充.

3.3 药物设计

要了解蛋白质的功能找到其致病的分子机理,只有氨基酸的顺序是不够的,还必须知道它们的空间结构.要设计药物对这些疾病进行治疗更需要了解这些蛋白质的空间结构.目前,一些常规的方法如 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的需要.这时,生物信息学中的理论模拟与结构预测就显示了其重要性,基于生物大分子结构知识的药物设计也成了当前药物研究的一个热点,它是根据药物分子与大分子之间作用的互补原理,在受体结构的基础上反过来设计药物分子.对于药物设计除了前面所说的蛋白质结构预测的方法之外,还常应用三维定量构象关系(3D-QSAR)的方法和虚拟受体的方法,这些方法也都取得比较明显的效果.

4 展望

由于生物信息学对科学上和商业上都具有非同一般的重要性,因此,它的成果不仅对相关基础学科起到巨大的推动作用,还将对农业、医学、环境、卫生、食品等产业产生巨大的影响.所以,各国政府机构和商业组织都纷纷投资生物信息学的研究,欧美各国及日本都相继成立了生物信息数据中心.我国在生物信息学方面的研究起步较晚,但也已经显露出了蓬勃发展的势头,清华大学、北京大学、军事医学科学院、中科院生物物理研究所、中科院上海生命科学研究院等大学和科研单位都开展了生物信息学的研究,为揭示生命的奥秘和保护国家的"基因资源"而努力.总之,生物信息学积极倡导的全球范围的资源共享将对整个人类社会的发展产生深远的影响,其研究领域和应用范围也将得到进一步的拓展.

参考文献(References):

[1] VENTER J C, ADAMS M D, MYERS E W, *et al*. The sequence of the human genome[J]. Science, 2001, 291: 1304-1351.

[2] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome[J]. Nature, 2001, 409: 860-921.

[3] YU J, HU S N, WANG J, *et al*. A draft sequence of the rice genome(*Oryza sativa L. ssp. Indica*)[J]. Science, 2002, 296: 79-92.

[4] GEFF S A, RICKE D, LAN T H, *et al*. A draft sequence of the rice genome(*Oryza sativa L. ssp. Japonica*)[J]. Science, 2002, 296: 92-100.

[5] SAMUEL A, JARROD C, ELIA S, *et al*. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*[J]. Science, 2002, 297: 1301-1310.

[6] 田云, 卢向阳. 生物信息学[J]. 生物学杂志(TIAN Y, LU X Y. Bioinformatics[J]. Journal of Biology), 2002, 19(3): 11-12.

[7] 孙言伟, 邹立君. 生物信息学的研究进展[J]. 中华医学图书情报杂志(SUN Y W, ZOU L J. Bioinformatics: advances in research[J]. Chin J Med Libr Info Sci), 2002, 11(4): 1-3.

[8] 刘秀艳, 滕 胜. 应用计算机识别蛋白质功能[J]. 生命的化学(LIU X Y, TENG S. Current advances in protein function assigned by computational methods[J]. Chemistry of Life), 2000, 20(3): 109-102.

[9] 陈惟昌, 陈志华, 邱红霞, 等. DNA 计算机的研究和展望[J]. 生物化学与生物物理进展(CHEN W C, CHEN Z H, QIU H X, *et al*. Progress in DNA computer[J]. Prog Biochem Biophys), 2001, 28(2): 156-159.

[10] 孟大志, 曹海萍. DNA 计算与生物数学[J]. 生物物理学报(MENG D Z, CAO H P. DNA computing and biological mathematics[J]. Acta Biophysica Sinica), 2002, 18(2): 163-173.

[11] ARVIND K B, TERRANCE E M. Evolutionary analysis by whole-genome comparisons[J]. Journal of Bacteriology, 2002, 184(8): 2260-2272.

[12] RICHARD J M, MARK D A, EUGENE W M, *et al*. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome[J]. Science, 2002, 296: 1661-1671.