

医学科学研究中的生物信息学应用

来茂德

(浙江大学医学院病理学与病理生理学系、环境基因组学研究中心,浙江 杭州 310031)

[摘要] 生物信息学是研究生物信息处理(采集、管理和分析应用),并从中提取生物学新知识的一门科学,它连接生物数据和医学科学研究。生物信息学在基因组学研究中发挥了不可替代的作用,并为解决生物系统的复杂问题提供了可能。对生物信息学的基本概念,同源性和相似性的区别作了介绍,并且对生物信息学在新基因发现、蛋白鉴定和生物芯片结果分析中的应用作了概述。

[关键词] 信息科学;生物学;专业;医学;生物信息学;同源性;相似性;基因组学;蛋白质组学;生物芯片

[中图分类号] Q 811.4, R 318.04 **[文献标识码]** A **[文章编号]** 1008-9292(2004)02-0091-04

人类基因组计划的成功实施使生命科学进入了信息时代。基因组学、蛋白质组学和生物芯片技术的发展,使得与生命科学相关的数据量呈线性高速增长。对这些数据全面、正确的解读,为阐明生命的本质提供了可能。连接生物数据与医学科学研究的是生物信息学(bioinformatics)。应用生物信息学研究方法分析生物数据,提出与疾病发生、发展相关的基因或基因群,再进行实验验证,是一条高效的研究途径。

1 生物信息学的基本概念

研究生物信息处理(采集、管理和分析应用),并从中提取生物学新知识的一门科学称生物信息学。生物信息包括多种类型的数据,如核酸和蛋白质序列,蛋白质二级结构和三级结构的数据等。由实验获得的核酸蛋白序列和三维结构数据等构成初级数据,由此构建的数据库称初级数据库。由初级数据分析得来的诸如二级结构、疏水位点、结构域(domain),由核酸序列翻译来的蛋白质以及预测的二级三级结构,称为二级数据。创新算法和软件是生物信息学持续发展的基础,高通量生物学研究方法和平台技术是验证生物信息学研究结果的关键技术。因此,生物信息学是一门新兴的交叉学科,涉及生物学、数学和信息科学等学科领域。生物信息学伴随基因组研究而产生,并随基因组及相关研究发展而拓展。

“同源性 homology”与“相似性 similarity or analogy”是生物信息学的两个最基本的概念,且常常混淆。同源性是指来自同一祖先,在进化起源上同一。在生物进化过程中由于趋异(divergence)性事件的作用而产生,如为序列片段称为同源序列,如为蛋白质称同源蛋白。相似性指的是由不同祖先来源,由于趋同进化(convergent evolution)而形成的共同结构或功能特征。因此相似性就是简单比较两者的相同程度^[1]。

同源性和相似性是两个不同的概念,相互之间没有直接的等同关系。相似的并不一定同源,因为在进化过程中,来源不同的基因或序列由于不同的突变而趋同。同源一般表现为相似,但同源的并不一定比非同源的相似程度高。

同源中分直系同源(orthology)和旁系同源(paralogy)。直系同源是指在不同物种中来自同一祖先的基因,其执行相同功能的基因或蛋白质,如哺乳动物的胰岛素。旁系同源指在同一生物体内执行不同但相关功能的基因或蛋白质,其基因有共同起源,如同是起源于珠蛋白的 α 珠蛋白, β 珠蛋白和肌红蛋白。直系同源物允

收稿日期:2004-01-06 修回日期:2004-02-06

基金项目:国家自然科学基金(30371605, 30370636, 30070343)资助项目

作者简介:来茂德(1960—),男,博士,病理学教授,主任医师,博士生导师,主要从事分子病理学研究。E-mail: lmp@zju.edu.cn

许对跨物种间的关系进行研究,而通过基因复制产生的旁系同源物则可能有助于研究一些基本的进化机制,因为复制基因可以产生分离的不同进化通路,并且通过变异和适应而进化得到新的特征(<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html>)。有关直系同源和旁系同源的概念用得比较乱且有不少争论。也有文献认为直系同源物可以有不同的功能,旁系同源物并不一定限于同一物种^[2,3]。

2 生物信息学研究的主要医学应用

生物信息学研究涉及两个层次,一是对海量数据的收集、整理和服务,也就是管好这些数据;二是从中发现新的规律,也就是用好这些数据(<http://www.bioinfo.org.cn/course/crs1.html>)。根据遗传学的中心法则,生物学数据涉及 DNA、RNA 和蛋白质数据及相关技术产生的数据。以下主要述及目前与医学基础研究相关的应用。

2.1 新基因的发现 疾病的发生发展与特异基因的改变有关。鉴定与疾病相关的基因是科学家在积极探索的一个方向。通过 mRNA 差异显示、抑制性差减杂交、cDNA 微阵列、基因表达系列分析(SAGE)等方法,获得了大量差异表达的序列片段,需要进一步鉴定这些序列片段是什么。通常用 BLAST 工具或 FASTA 工具进行序列的相似性搜索(表 1)。我们实验室用免费的 Linux 操作系统和低价位的 PC 机为基础,建立了高通量的 EST 分析平台。借助 Phrep/Phrap/consed 系列软件及自行编译 Perl 程序,利用常用的核酸序列数据库,实现了大批量差异基因片段从测序峰图到核酸序列的转换和序列的拼接,及序列比对等系列过程的全自动化分析(图 1)。应用这个平台我们成功地分析了 300 多条结肠肿瘤中的差异表达序列,取得了很好的结果^[4]。同时也可据此原理建立蛋白质组学研究结果的分析平台。

除此之外,还有在全基因组中寻找符合基因结构的新基因,现有软件可以应用,如 Genefinder, GeneScan 等可以预测新基因的存在。

通过上述自动化分析,我们可以获得是否有价值进行下一步验证的信息:①是不是新基因?②是不是与已知的基因有同源性或相似性?③是否属于一个已知的多基因家族?④该基因编码的蛋白质将发挥什么功能?通过进一步的实验,验证生物信息学所得的结果。

表 1 BLAST 程序及运用范围
Table 1 A set of Blast programs and their applications

程序	数据库	比较内容	运用
Blastn	核酸	核酸序列与核酸数据库比较	可能找到具有远源进化关系的匹配序列
Blastp	蛋白质	氨基酸序列与蛋白数据库比较	适合寻找分值较高的匹配
Blastx	蛋白质	核酸序列的 6 种翻译序列与蛋白质数据库比较	适合新 DNA 序列与 EST 序列的分析
tBlastn	核酸 (翻译)	蛋白质与核酸数据库的有 6 种翻译序列的比较	适合寻找数据库中尚未标记的编码区
tblastx	核酸 (翻译)	核酸序列的 6 种翻译序列与核酸数据库的 6 种翻译序列比较	适合分析 EST 序列比较

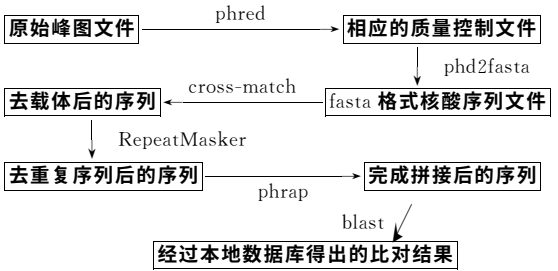


图 1 整合程序后本地平台的工作流程图
Fig. 1 Flowchart of the local platform built from a bunch of programs

基因是由多种成分所组成的,如外显子、内含子、启动子、增强子等。对这些成分的充分了解,有助于进一步预测编码蛋白并对其进行功能研究。对启动子的深入分析有助于了解该基因表达的调控机制^[5]。蛋白质的分子量、等电点、二级结构、三级结构、四级结构、膜蛋白的跨膜区段、酶的活性位点,以及蛋白质之间相互作用等结构和功能信息,都可以从序列信息借助于生物信息学工具来获得。当然,各种预测方法都是建立在现有的生物学数据和已知的生物学知识之上的,但在不同模型和算法基础上建立

的不同分析程序有其一定的适用范围,以及相应的限制条件,因此,最好对同一生物学问题尽量多用几种分析程序,综合各种方法获得结果。由于知识的加速更新,例如网页的刷新、撤换,新数据库的不断建立,这就要求我们不断地学习摸索适用于自身专业与实际条件的生物信息学方法。

2.2 从实验数据中鉴定蛋白质 作为蛋白质组学研究的核心技术,双向电泳(2-DE)根据蛋白质的分子量和等电点分离蛋白质。严格控制实验条件,2-DE胶上的蛋白质点足以用于鉴定一些蛋白质。SWISS-2DPAGE提供了很多标准凝胶图像及检测胶上蛋白质是否移位的工具,通过与已知细胞或组织蛋白质图谱比较,可以鉴定一些标志性蛋白。然而,事实上因为蛋白质样品自身的可变性,样品制备的不可重复性,及2-DE不可能分离样品中的所有蛋白质,所以很难用该法鉴定蛋白质。

应用质谱技术可获得的高精确肽质量或氨基酸序列,结合一系列数据和分析工具来鉴定蛋白质。蛋白质的功能预测起始于序列。如果蛋白质是已知的,可通过序列数据库及一些主要的文献来决定其功能。如果蛋白质是未知的,则需要通过分析相关蛋白质来获得与其功能相关的一些线索。BLAST分析获得一系列紧密相关的蛋白质,其中一些蛋白质被高度注释。如果全长序列匹配,即可预测相似的功能,如果只有短片断序列匹配,则表明这些片断是保守的蛋白域,可能预示着蛋白质的一些功能。这些分析功能的实现也可以通过工具整合搭建平台,像EST一样实现快速高通量分析。

蛋白质组学最终将产生远远超过DNA序列数据库中存储的数据。对人类与模式动物蛋白质的完整分类及对蛋白质功能的注释,将推动蛋白质组学的飞速发展。蛋白质组学的目标是全面了解蛋白质,但因为每一种细胞的蛋白质表达、修饰、与其它蛋白质的相互作用是不同的,所以蛋白质组学分析要比基因组学复杂得多,对蛋白质组的探索将永远依赖于新的数据处理资源。

2.3 基因芯片数据的处理分析 生物芯片数据的生物信息学分析,成为基因表达(mRNA

和蛋白质)研究临床疾病的分子机制、疾病诊断和药物筛选的重要领域。基因芯片利用生物样品的cDNA与芯片上的探针杂交,获得高密度杂交点阵图像,用图像分析软件,提取各杂交点图像吸光度值、面积和吸光度比值等荧光信号数据,并转化成基因表达矩阵(gene expression matrix),进行定量分析。基因表达矩阵是用来描述基因芯片基因表达数据的矩阵,一般用行来表示基因,列表示各个不同的样本(如不同的组织标本,处理方法以及发展阶段)。因此,每一个格子表示一个样本在某一个基因的表达水平。基因表达矩阵的建立是基因芯片数据分析的基础。首先,采用图像处理分析软件自动识别原始的杂交信号点,提取各个杂交点的信号强度如吸光度,扣除背景信号水平后作为杂交点信号净值。因为芯片实验中的系统误差,如样本差异、荧光标记效率以及检出率等的存在,因此在数据分析之前需对原始信号进行标准化处理(normalization)后才能进行分析^[6]。一般标准化的方法采用调整标准化系数使平均比值(ratio)为1,如管家基因法和整体平均值法以及密度依赖法^[7]等。

对基因表达矩阵分析的目的是探讨潜在的生物学信息和规律。有两类基本方法。第一类是差异分析。目前一般用ratio值(cy3/cy5的比值)分析,当某样本间的ratio值在0.5~2之间表示不存在表达差异。但是因为对同一样本的多个重复样本分析时存在变异,因此可以用显著性检验的方法,如t检验、方差分析等方法进行处理。另一类方法是聚类分析。聚类分析是将性质相同或相近的基因、样本归于同一类,而差别较大的归于另一类。用聚类分析方法分析基因芯片数据的原理,是基于具有相同表达谱的基因具有相似的生物学功能的假设,因此可以从表达谱相似的已知基因的功能推测未知基因的功能。通常采用监督分析和非监督分析两种策略^[6]。监督分析通过已知表达谱的信息,如功能分类、疾病状态等参数建立分类标准,以此来预测未知基因的功能。监督学习技术包括SVM(supervised machine-learning techniques)、logistic回归、神经网络以及LDA(linear discriminate analysis)等方法。而非监督分析方法是通过对表达谱的归

类,寻找相关基因或有相关性的样本。在分析技术上常用相关系数或欧式(euclidean)距离等相似距离来比较数据间的关系,常用的方法包括层次式聚类(hierarchical clustering)、自组织作图(self-organizing maps)、K-means 聚类等。通过聚类可以分析在不同条件下功能相关或共表达的基因,在疾病分类、诊断、疗效和预后评估方面有很大的应用价值^[8]。但是,因为分析的假设本身并不完善,所以,还需不断探索新的方法来分析基因芯片数据。

上面仅简述了医学科学实验研究目前最常遇到的三个生物信息学问题。通过综合多方面的信息来探究各种功能,并对复杂的生物学系统采用一种更全面的认识,是生物信息学今后很重要的一个发展方向。就生理功能的实现来说,蛋白质需要在彼此相关的网络中才能发挥其作用。因此,我们不仅要考虑代谢通路、信号转导系统等模块的信息,还要考虑由这些模块链接成网络的综合信息。就疾病来讲,大多数疾病是由多基因决定的,这些基因与环境的相互作用形成疾病的表型。因此个体的基因型与疾病表型之间的关系是极其复杂的,这种复杂关系的破解要借助不断完善的生物信息学工具的帮助。人类对疾病的易感性、药物的反应性等的差异与 SNP 有关,人类基因组单倍体型图的绘制是人类基因组计划走向应用的重要步骤,但 SNP 数据与疾病关系的阐明有赖于生物信息学技术的应用。现在作为一名分子生物学者,不具备一些基本的生物信息学技能已几乎难以胜任;同样作为医学科学研究者不掌握基本的生物信息学知识,也不可能进入分子医学研究的前沿。实验室的每一项技术,从简单的克隆、PCR 到基因表达的分析,都需要在计算机上进行数据处理,虽然对医学科学研究工作者来说

重要的不是创造算法和编制软件,但了解 DNA 和蛋白质分析工具的基本原理是需要的。医学工作者加强与不同学科专家之间的沟通、交流与合作,可以促进解决复杂问题工具的诞生,从而促进医学的发展。

References:

- [1] ATTWOOD T K. Genomics; The Babel of Bioinformatics [J]. *Science*, 2000, 290(5491): 471—473.
- [2] JENSEN R A. Orthologs and paralogs—we need to get it right [J]. *Genome Biology*, 2001, 2 (8): interactions1002. 1—1002. 3.
- [3] GOGRTN J P, OLENDZENSKI L. Orthologs, paralogs and genome comparisons [J]. *Curr Opin Genet Dev*, 1999, 9: 630—636.
- [4] GU Xue-mei, ZHANG Hao, LAI Mao-de (谷雪梅, 张昊, 来茂德). Bioinformatics analysis to the differentially expressed genes of normal mucosa and carcinoma of colon [J]. *Journal of Zhejiang University: Medical Sciences* (浙江大学学报: 医学版), 2004, (2): 95—101. (in Chinese)
- [5] LIN Jie, ZHU Yi-min, LAI Mao-de (林洁, 朱益民, 来茂德). Structural features of GR6 gene and its expression in colorectal neoplasm [J]. *Journal of Zhejiang University: Medical Sciences* (浙江大学学报: 医学版), 2004, (2): 102—107. (in Chinese)
- [6] LEUNY Y E, CAVULIERI D. Fundmentals of cDNA microarray data analysis [J]. *Trend Genetics*, 2003, 19 (11): 649—659.
- [7] WANG Yong-yu, ZHANG You-yi (王永煜, 张幼怡). Analysis and treatment of gene chip data [J]. *Prog Biochem Biophy*, 2003, 30(2): 321—323. (in Chinese)
- [8] BROWN M P, GRUNDY W W, LIN D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines [J]. *Proc Natl Acad Sci USA*, 2000, 97(1): 262—267.

[责任编辑 张荣连]