新技术讲座

生物信息学在新基因全长 cDNA 序列分析及功能预测中的应用*

张成岗** 贺福初

(军事医学科学院放射医学研究所，北京 100850)

摘要 差异表达基因的检测与分析已成为研究具有差异的生物学表型的常规策略. 对通过实验所获得的差异基因片段进行生物信息学分析，主要包括基于国际互联网的序列相似性分析、片段重叠群分析和全长 cDNA 序列分析，以及如何构建局域网并采用本地服务器进行规模化的数据分析，从而为研究人员提供可参考的生物信息学数据分析方案.

关键词 生物信息学，全长 cDNA，功能预测，PC 机，Linux 操作系统

学科分类号 Q786

人类基因组计划已逐渐从“结构基因组计划”迈入“功能基因组计划”的时代. 全基因组序列的解读，并不能使人类对编码基因这一层次有更明确的认识. 因此，cDNA 测序计划就成为人们了解编码基因结构与功能的关键所在[1,2]. 通常，人们极其关注不同模型处理下同种组织中基因表达的差异，进而关注这些差异表达的基因与相应生物表型之间的关联[3]. 能否发现真正的差异表达基因，一个共同的问题就是所采用的数据分析技术及其分析能力能否有效、快速地揭示大量序列数据中所蕴含的生物信息. 因此，本文拟针对性地提出生物信息学如何用于中、小规模差异表达基因片段的分析策略，以加速相关的研究进展，并提高生物信息学在我国的应用水平.

从事生物信息学数据分析无非是利用一定的硬件、软件条件和一定的数据库环境，对用户的数据进行理性分析的过程. 计算机的普及、计算技术的大力发展以及国际互联网相关技术的进步，使得我们很容易找到适合于自己的资源，从而可构筑各种有效的数据分析平台. 此处假设用户通过测序获得了 50 条差异基因片段. 现在的问题是，用户如何快速地判断这些基因片段对应于什么基因？这些基因片段之间的关系如何？是否是同一基因不同部分的片段？甚至于，从公司返回的这些序列中有没有载体序列？如果有，如何去除？这些基因的染色体定位及基因组结构如何？其所编码的蛋白质的功能如何分析？以下就这些问题展开具体讨论.

1 依赖互联网和 Windows 的数据分析方案

目前大多数实验室拥有一台以上安装有 Windows 操作系统的 PC 机. 在这种环境下，对大量基因片段的分析主要依赖于本机所安装的生物信息学软件，或互联网上的数据分析资源. 对于序列性质的初步判断，一个最简单也最直接的分析策略是，将这些序列利用 Blast 等软件进行序列相似性比对，即可快速判定这些序列对应于何种基因[4]. 通常人们习惯于直接联网到美国国家生物技术信息中心 (National Center for Biotechnology Information, NCBI)，使用 Blast 服务器进行序列比对 (http://www.ncbi.nlm.nih.gov/BLAST/)，其优点在于能够使用最新版本的 Blast 程序和最新的数据库资源. 然而，这种方式的最大缺点是每次只能对一条序列进行比对分析. 如果用户有 50 条序列，那么就需要重复 50 次“序列复制→序列粘贴→序列提交→结果保存”的过程，异常繁琐. 比较理想的解决方案是，用户可直接使用国内北京大学生物信息中心的 Blast 服务器 (http://blast.cbi.pku.edu.cn/)，以这种方式用

*国家高技术“863”计划、国家海外青年学者合作研究基金 (30128010)、国家自然科学基金 (30100049, 39900041, 39900074) 与北京市自然科学基金 (7002030)、军事医学科学院科技创新启动基金 (0102001, 9905105) 部分资助.

**通讯联系人.

Tel：010-66931590，E-mail：zhangcg@nic.bmi.ac.cn

收稿日期：2002-05-28，接受日期：2002-08-06

户一次可将所有序列以 FASTA 格式提交服务器进行分析，网络浏览器可一次性返回所有序列的相似性比对结果. 此时，用户应当注意参数的选择. 如果选择的相似性比对阈值较低，或允许显示的相似性序列数量较多，均将导致输出结果过于庞大. 笔者建议设参数值为“Expect = 0.01，Description = 10，Alignments = 10，Color schema = Color schema 4”. 这样可保证真正相似的序列不至于漏检，又不至于使输出结果文件过于庞大，而且查询序列与数据库中对应的相似序列的差异位点之处，还能够以彩色方式显示出来.

核酸序列的相似性分析有几个层次，可依次按照 mRNA/cDNA、EST、基因组 DNA 等进行. 首先，需要判断用户所分析的序列是否与国际基因数据库（GenBank/EMBL/DDBJ）非冗余数据库（non-redundant database，nr 数据库）中的序列有相似性. nr 数据库中收录了除 EST、STS、GSS 及第 0、1 或 2 期 HTGS（高通量基因组序列）的所有序列. 一般来说，如果用户查询的序列和 nr 库中的序列有高度相似性，则可由此推断用户序列的性质. 如果无法从 nr 库找到相似序列，用户就需要判断是否能够从 EST 数据库中找到相似序列. 这样，即使用户的查询序列不能找到相似的 cDNA 序列，也有可能找到别人已经测定过的 EST 序列. 如果从 EST 数据库中还找不到相似序列，用户可直接对相应物种的基因组序列进行查询判定用户序列的性质.

然而用户很快就会发现，在进行序列相似性比对分析时，nr 数据库中可能会返回所查询的序列含有克隆时所采用载体序列的信息. 这对序列分析无疑是一个打击. 因此，对所查询的序列进行载体序列的判别应当成为序列分析的第一步. 测序公司一般不会去除测序结果中的载体序列，除非用户声明有此需求. 当然，载体序列的鉴定也可通过联网到 NCBI 进行分析，在 BlastN 数据分析时选择 Vector database（载体序列数据库）即可，或直接使用 NCBI 所提供的网址（http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html）即可判知. 但该结果只是告诉用户所提交的序列那部分与何种载体序列具有较高的相似性，并不直接输出去除载体后的序列. 后者必须通过一定的软件加工才能实现，例如 Sequencher™.

没有载体序列污染的（即干净的序列）可随后用于序列之间相互关系的鉴定. 例如，用户所提交的 50 条序列中很可能只对应少数几种或十几种基因，即在这 50 条序列之间存在“冗余性”. 对冗余序列的分析需要依赖于片段重叠群分析技术，即“contig analysis”. 这种分析的目的是鉴定冗余序列是否能够以及如何形成一条共有序列（consensus sequence），要求该共有序列能够覆盖所有的子序列，而这条共有序列就有可能对应于较长甚至是全长 cDNA 序列. 这一分析过程必须通过特定的软件才能实现. 此处，笔者推荐使用 Sequencher™软件（http://www.genecodes.com/）进行分析. 首先是需要去除序列（也可以是原始的测序峰图文件）中位于末端的低质量序列区域. 随后，用户可使用该软件删除载体序列. 此前用户需输入所使用的载体序列. 这两步操作之后，序列就可进行“contig”分析. 在 Sequencher™软件的操作界面中，用户可直接选择“Assemble automatically”，即可对所选择的序列进行片段重叠群分析. 分析结果中，该软件将告知用户哪些序列片段可形成一个完整的“contig”，而且提供具体的对齐情况. 用户藉此可判断本次差异显示实验结果中序列的重复情况. 进一步，用户可将这些基因片段所对应的 cDNA 序列也一并输入到 Sequencher™软件后，再进行片段重叠群分析，从而使用户能够判断这些序列片段在 cDNA 上的分布区域.

通常，所测定的差异基因片段往往只对应于 cDNA 序列的一段. nr 数据库中所收录的 cDNA 序列也有相当一部分不是全长 cDNA 序列. 因此，随着序列分析的深入进行，用户将发现全长 cDNA 序列对于进行后续研究是十分重要的. 常规方法是或通过实验研究获得全长 cDNA 序列[5]，或直接基于 EST/cDNA 数据库进行电子序列延伸. 如果尚未构建本地化的电子序列延伸系统，就必须依赖于网络资源进行分析（例如：http://www.hgmp.mrc.ac.uk/ESTBlast/），但这些资源一般只对注册用户提供服务. 一个折衷的办法是可考虑采用基于 UniGene 数据库进行电子序列延伸. 新版本的 Blast 软件已增加了查询 UniGene 数据库的功能，并可下载该 UniGene 记录所对应的全部 mRNA/cDNA 序列和 EST 序列. 随后，用户可将所下载的全部序列用 Sequencher™软件进行拼接分析，得到较长的 contig 序列. 这样就实现了对所分析序列片段的一轮电子序列延伸分析. 该过程可不断重复，直至无法获得更长的序列，从而可获得较长的对应于用户原始序列的 cDNA 序列. 用户可据此设计引物并采

用 RT-PCR 技术对其进行鉴定. 不过，由于此过程要求用户对计算机操作、网络文件传输等比较熟悉，而且该过程较为耗时，一般不建议用户进行此种操作. 倘若有规模较大的核酸序列电子拼接的需求，建议采用自行构建相关数据分析平台的策略进行[6].

基因的染色体定位以及基因组结构也是核酸序列的常规分析内容之一. 可通过将研究的 cDNA 序列，利用 NCBI 的 Blast 服务器查询其所对应的基因组序列而获得（例如，查询人基因组序列：http://www.ncbi.nlm.nih.gov/genome/seq/page.cgi? F=HsBlast.html&&ORG=Hs). 浏览器将返回该基因片段所对应的基因组序列信息，并在页面中提供一个按钮“Genome View”. 点击此按钮，即可出现相关的基因组定位信息. 需要注意的是，Blast 软件无法准确判断 cDNA 序列中外显子和内含子的边界（exon/intron boundaries），因此需要采用其他软件如 Sim4（http://pbil.univ-lyon1.fr/sim4.html)进行分析[7]，从而可获得较为准确的基因组结构. 作者实验室目前已开发了基于 sim4 软件的 EST/cDNA 序列结构自动分析系统，最近将向公众开放.

当然，获得较长 cDNA 序列并经实验验证之后，就应对其进行更进一步的分析. 这些分析一般包括开放阅读框架（open reading frame，ORF）查询、编码蛋白质的相似性分析及结构功能域分析. ORF 分析可直接联网进行(http://www.ncbi.nlm.nih.gov/gorf/gorf.html). 蛋白质序列的相似性分析可通过 Blast 软件实现(http://www.ncbi.nlm.nih.gov/BLAST/). 蛋白质的结构功能域分析，一般建议采用 InterPro 服务器(http://www.ebi.ac.uk/interpro/scan.html)[8]，或 SMART 服务器(http://smart.embl-heidelberg.de/)[9] 进行分析. 这些服务器均能够可视化地返回分析结果，用户可藉此预测蛋白质的功能.

以上介绍了在没有专用的生物信息学数据分析平台时进行基因功能分析的策略，可为普通实验室所参考. 然而，如果需要经常进行序列分析，或者是分析的规模较大，这种方法就会花费大量人力和物力. 此时，我们推荐在实验室内部构建专用的生物信息学数据分析平台，把一些重复性的、可程序化的过程直接交由计算机完成，用户的主要精力就可放在对分析结果进行后续分析及实验设计上.

2 自行构建生物信息学数据分析平台

生物信息学固然是一门新兴学科，但是真正操作起来也不是十分困难. 就构建序列分析平台而言，一台普通的具有流行配置的 PC 机就可完成上述序列分析工作. 当然，有条件的实验室可将所有的 PC 机连接成局域网，选择一台安装有 Linux 操作系统的高性能微机作为服务器，并配备多种生物信息学分析软件和数据库.

软件的选择取决于所拟进行的任务. 就本文描述的序列分析工作而言，以下软件是必需的：NCBI 所开发的独立运行的 Blast 软件(ftp://ftp.ncbi.nih.gov/blast/executables/blast.linux.tar.Z)，可用于本地化的序列相似性分析. 也可下载用于局域网中提供网络界面比对服务的 Blast 软件(ftp://ftp.ncbi.nih.gov/blast/server/wwwblast-Jan.11.2002/wwwblast.Linux.tar.gz). web Blast 程序能够为用户提供网络界面的批量序列相似性分析服务，类似于 NCBI Blast 的 web 界面. 另外，Washington University 所开发的 Phred/Phrap/Consed 系列软件可用于从测序峰图中读取序列、载体序列鉴定、片段组装、重复序列鉴定甚至单核苷酸多态性分析（single nucleotide polymorphism，SNP）鉴定等，也是构建本地化数据分析平台的必选[10]. 其他一些软件也在可选之列，例如，用于多序列比对的 ClustalW/ClustalX 软件(ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw/clustalw1.82.UNIX.tar.gz)[11]，用于鉴定外显子/内含子边界区域的 sim4 软件(http://globin.cse.psu.edu/dist/sim4/sim4.tar.gz)[7]. 这些软件通常已经足以应付基本的序列分析任务了. 其他所需要的数据分析软件可从相关的服务器中查找下载，例如欧洲生物信息学研究所的网站(ftp://ftp.ebi.ac.uk/pub/software/unix/)等.

构建本地化的数据分析平台，另一必须考虑的是数据库的选择. 为了方便用户的需求，NCBI 提供了可供 Blast 软件直接进行序列相似性比较的数据库(ftp://ftp.ncbi.nih.gov/blast/db/)，目前已有 8.9 GB 的容量. 值得一提的是，用户也可构建自己的核酸和蛋白质序列数据库，并用 Blast 软件进行序列分析. 这是构建本地化序列分析平台的一大优点. 而且，这种数据库也可被挂接到网络版 Blast 软件中，从而可被局域网内其他用户所检索[12]. 关于如何将这些软件、数据库有机地组织起来使本地化

的数据分析发挥最大效率，可参考有关此方面的一系列论著[13~15].

3 讨 论

本文重点介绍了核酸序列分析的常规策略，对于进行差异基因表达研究后续的数据分析具有一定的参考价值. 同时，本文所推荐的基于国际互联网的分析策略和构建本地化的生物信息学数据分析平台，均不需要很多投资就可实现. 而且，国内的一些服务器中也有可供免费下载的大量软件和数据库资源，因此，对于从事较多关于新基因研究的实验室而言，构筑一个强有力的本地化的生物信息学数据分析平台是一个较为理想的选择，可以大大加速相关研究的进展.

参 考 文 献

1 Hu R M, Han Z G, Song H D, *et al*. Gene expression profiling in the human hypothalamus-pituitary-adrenal axis and full-length cDNA cloning. Proc Natl Acad Sci USA, 2000, **97** (17): 9543~9548

2 Yu Y T, Zhang C G, He F C, *et al*. Gene expression profiling in human fetal liver and identification of tissue- and developmental-stage-specific genes through compiled expression profiles and efficient cloning of full-length cDNAs. Genome Res, 2001, **11** (8): 1392~1403

3 王吉村，药立波，赵忠良. 筛选差异表达基因和蛋白质的方法进展. 生物化学与生物物理进展，2001，**28** (1)：33~36
Wang J C, Yao L B, Zhao Z L. Prog Biochem Biophys, 2001, **28** (1): 33~36

4 Altschul S F, Madden T L, Schäffer A A, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997, **25** (17): 3389~3402

5 Zhang C G, Wang C L, Deng M Y, *et al*. Full-length cDNA cloning of human neuroglobin and tissue expression of rat neuroglobin. Biochem Biophys Res Commun, 2002, **290** (5): 1411~1419

6 张成岗，孙焕东，贺福初，等. 基于PC/Linux的核酸序列电子延伸系统的构建及其应用，遗传，2002，**24** (1)：50~54
Zhang C G, Sun H D, He F C, *et al*. Hereditas, 2002, **24** (1): 50~54

7 Florea L, Hartzell G, Zhang Z, *et al*. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res, 1998, **8** (9): 967~974

8 Zdobnov E M, Apweiler R. InterProScan-an integration platform for the signature- recognition methods in InterPro. Bioinformatics, 2001, **17** (9): 847~848

9 Schultz J, Richard R, Copley T, *et al*. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res, 2000, **28** (1): 231~234

10 张成岗，欧阳曙光，贺福初，等. 基于PC/Linux的核酸序列分析系统的构建及其应用. 生物化学与生物物理进展，2001，**28** (2)：263~266
Zhang C G, Ouyang S G, He F C, *et al*. Prog Biochem Biophys, 2001, **28** (2): 263~266

11 Thompson J D, Higgins D G, Gibson T J. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res, 1994, **22** (22): 4673~4680

12 张成岗，张利达，贺福初，等. 序列同源性分析软件Blast的WEB界面构建及其应用. 生物化学与生物物理进展，2001，**28** (6)：916~918
Zhang C G, Zhang L D, He F C, *et al*. Prog Biochem Biophys, 2001, **28** (6): 916~918

13 Baxevanis A D, Ouellette B F F. 李衍达，等译. 生物信息学：基因和蛋白质分析的实用指南. 北京：清华大学出版社，2000. 133~174
Baxevanis A D, Ouellette B F F. translated by Li Y D, *et al*. Bioinformatics: A Practical Guide to The Analysis of Genes and Proteins. Beijing: Qinghua University Press, 2000. 133~174

14 Attwood T K, Parry-Smith D J. 罗静初，等译. 生物信息学概论. 北京：北京大学出版社，2002. 97~176
Attwood T K, Parry-Smith D J. translated by Luo J C, *et al*. Introduction to Bioinformatics. Beijing: Beijing University Press, 2002. 97~176

15 张成岗，贺福初. 生物信息学. 北京：科学出版社，2002. 64~109
Zhang C G, He F C. Bioinformatics: Methods and Protocols. Beijing: Science Press, 2002. 64~109

Application of Bioinformatics in Full- length cDNA Sequence Analysis and Function Prediction of Novel Genes*

ZHANG Cheng- Gang**, HE Fu- Chu

(Department of Genomics and Proteomics, Beijing Institute of Radiation Medicine, Beijing, 100850, China)

Abstract Detection and analysis of differentially- displayed genes is a routine strategy for the study of different biological phenotype. A data- mining flowchart on the bioinformatic analysis of these genes is described, which includes sequence similarity analysis based on internet contig assembly and full- length cDNA sequence obtaining. Moreover, the strategy for how to construct a local bioinformatic platform for large- scale analysis on novel genes is also introduced. All of this information will accelerate the progress on novel genes cloning and

function prediction.

Key words bioinformatics, full - length cDNA sequence, function prediction, personal computer, Linuxoperating system

* This work was partially supported by grants from State 863 High Technology R & D Project of China, National Science Fund for Distinguished Young Scholars (30128010), The National Natural Science Foundation of China (30100049, 39900041 and 39900074), The Natural Scienec Foundation of Beijing (7002030), Initiative Foundation for Scientific and Technological Innovation of Academic Military Medical Science (0102001 and 9905105).

** Corresponding author. Tel: 86-10-66931590, E-mail: zhangcg@nic.bmi.ac.cn

Received: May 28, 2002 Accepted: August 6, 2002

第三届亚洲视觉科学会议（ACV2003）第一轮通知

由中国科学院视觉信息加工重点实验室、神经科学研究所和西南眼科医院联合主办的第三届亚洲视觉科学会议（ACV2003）将于2003年11月21日～25日在重庆举行.

大会组织委员会 主席：李朝义（中国）；副主席：Keiji Uchikawa（日本），Chan-Sup Chung（韩国），赫荣乔（中国）

大会学术委员会 主席：王书荣（中国）；副主席：Makoto Ichikawa（日本），Choongkil Lee（韩国）

大会执行委员会 主席：李书章（中国）；副主席：李兵（中国），谢汉平（中国）

会议主题：1) Visual Neuroscience; 2) Visual Perception; 3) Depth and Spatial Vision; 4) Color Vision; 5) Visual Attention; 6) Eye Movements and Visuo-Motor Coordination; 7) Mathematical Models on Vision; 8) Retina; 9) Object Recognition; 10) Clinical Vision Studies; 11) Neural Imaging of Visual System; 12) Vision and Other Modalities.

重要时间 1）会议回执：2003年3月30日；2）论文摘要截止日期：2003年6月30日

有关会议地点、注册费、论文摘要格式等信息请详见第二轮通知. 第二轮通知将于2003年4月底前登出，届时将通过E-mail或邮寄发出，并请见会议网址：www.ibp.ac.cn/acv2003

第三届亚洲视觉科学会议（ACV2003）回执

（2003年11月21日～25日，重庆）

姓　名	性别：　职务：
工作单位	
通讯地址	
邮政编码	
电　话	传真：
E-mail	
□	我拟参加第三届亚洲视觉科学会议，请寄第二轮通知
□	我拟提交会议论文摘要

会议回执寄至：

100101 北京朝阳区大屯路15号中国生物物理学会　魏舜仪

电话：010-64889894/64889872，传真：010-64871293，E-mail：acv2003@sun5.ibp.ac.cn