DOI:10.13471/j.cnki.j.sun.yat-sen.univ(med.sci).2007.0079

·信息研究·

基于基因组学的日本血吸虫序列分析平台的构建及应用

刘稳升[1], 胡 斐[2], 吕志跃[2], 郑焕钦[2], 袁衡新[1], 吴忠道[2]
(中山大学中山医学院 1. 生物医学工程系, 2. 寄生虫学教研室, 广东 广州 510080)

摘 要:【目的】构建基于基因组学的日本血吸虫序列分析平台。【方法】基于工作站和 Linux 操作系统, 利用公开的日本血吸虫或相关资源, 收集有关生物信息学序列分析工具, 并自行编写相关软件实现多种资源的有机整合。【结果】构建了本地化的日本血吸虫序列分析平台, 在此基础上, 自主设计了日本血吸虫 EST 序列的电子延伸系统、基于日本血吸虫 EST 数据库的同源全长 cDNA 序列检索系统和序列自动注释系统等。【结论】所构建的序列分析平台不仅可以完成常规的序列分析工作, 还可成功地完成 EST 序列的电子延伸, 同源全长 cDNA 序列的检索以及序列的自动注释, 它可对大规模日本血吸虫序列进行快速、方便、高效的分析, 为充分利用血吸虫基因组开展疫苗研究和新药开发提供了有效的技术手段。

关键词: 日本血吸虫; 基因组; 序列分析系统
中图分类号: R383.2 **文献标识码:** A **文章编号:** 1672-3554(2007)03-0348-05

Construction and Application of Sequence Analysis Platform Based on Genomics of Schistosoma japonicum

LIU Wen-sheng[1], HU Fei[2], LU Zhi-yue[2], ZHENG Huan-qin[2], YUAN Heng-xin[1], WU Zhong-dao[2]
(1. Department of Biomedical Engineering, 2. Department of Parasitology, Zhongshan School of Medicine, SUN Yat-sen University, Guangzhou 510080, China)

Abstract: 【Objective】To construct a sequence analysis platform for the sequences of Schistosoma japonicum based on genome. 【Methods】Based on workstation and Linux operation system, using public resources of Schistosoma japonicum or related data, collecting some related bioinformatics tools for sequences analysis, and integrating organically some resources by programming at the mean time. 【Results】A platform was constructed successfully. Founded on this platform, we have developed the in silico EST elongating system of Schistosoma japonicum, homological full-cDNA sequence retrieval system based on EST database of Schistosoma japonicum, and auto sequence annotation system et al. 【Conclusion】This system can not only complete routine jobs of sequences analysis, but also complete in silico EST elongating, homological full-cDNA sequence retrieval and auto sequence annotation successfully. It is proven that the sequence analysis platform is available for utilizing to analyze the sequence data in high-throughput, rapidly, conveniently and efficiently, and subsequently, and help throwing new light on vaccine candidates and drug targets for Schistosomiasis japonica.

Key words: Schistosoma japonicum; genome; sequence analysis system
[J SUN Yat-sen Univ(Med Sci), 2007, 28(3):348-352]

继人类基因组计划(Human Genome Project, HGP)**之后, 血吸虫基因组计划**(Schistosome Genome Project, SGP)**于** 1992 **年正式启动, 并得到了美国基因组研究所**(The Institute for Genomic Research, TIGR)**的技术支持**, 1994 **年** WHO/UNDP/TDR **和多个国家的实验室参与了此项计划, 其最终目标是找到新的候选疫苗分子和治疗药物以及完全的基因组图谱。**1999 **年, 在** WHO/Word Bank/UNDP **血吸虫基因组网络会议上, 确定了** WHO **资助** SGP **的主要内容, 包括曼氏血吸虫新基因的发**

收稿日期: 2006-12-04
基金项目: 国家自然科学基金资助项目(NO. 30271165); 教育部博士基金资助项目[(2005)216 号]
作者简介: 刘稳升(1981-), 男, 湖南益阳人, 硕士生; 吴忠道, 教授, 博士生导师, 通讯作者, E-mail:wuzhd@mail.sysu.edu.cn

现、日本血吸虫新基因的发现、物理图谱的制作、cDNA 微型矩阵的制作、线粒体基因的测序以及生物信息学的研究, 也确定了未来 80%的 TDR 基因组资助奖金将用于后基因组计划。经过多年的努力, 在 cDNA 文库、EST 序列分析、基因差异表达的研究、基因组文库和基因组图谱、血吸虫信息学和数据库管理等多方面取得了重要进展[1-3]。目前, WHO 血吸虫基因组协作组是最广泛的国际合作组织之一, 其成员国美国、英国、埃及、巴西等多参与曼氏血吸虫基因研究计划, 而中国重点开展日本血吸虫基因组的研究工作, 研究单位包括中国南方基因组中心, 中国疾病预防控制中心寄生虫病预防控制所和中山大学中山医学院等。最近, 我国在日本血吸虫功能基因组和蛋白质组的工作取得了阶段性重大成果[4]。2006 年 5 月 17 日, 上海市生命科学与生物技术数据中心向全世界公布了由我国自主测序的日本血吸虫基因组工作框架图, 共计 300 多万条的 DNA 序列[5], 为全球科学工作者提供服务(http://lifecenter.sgst.cn/)。由于血吸虫基因组计划的实施, 综合利用现阶段的大量生物信息, 并从中快速挖掘有用信息, 应用于血吸虫病的免疫诊断、疫苗研发和新药靶点筛选等已成为可能。为此, 本课题组在前期工作基础上, 开展了本项研究工作, 旨在有力推动我国血吸虫病的科学研究和防治技术创新。

1　材料与方法

1.1　工作站的配置

1.1.1　硬件配置　工作站为 DELL Precision workstation 670, 处理器为双 Intel® Xeon™, 主频 2.80 GHz, 内存为 DDR DRAM 2.0 G, 硬盘为 SATA 146 G。

1.1.2　软件配置　操作系统为 RedHat Enterprise Linux AS 4; 应用软件包括 Apach 网络服务器, MySQL 网络数据库系统, Perl 和 PHP 等相关软件。

1.2　核酸及蛋白质序列数据库

通过互联网访问公开数据库, 在许可的条件下下载数据信息, 并进行相关的解压缩、安装, 构建序列分析的本地化数据库。现将本实验室收集的主要序列分析数据库列表如表 1。

表 1　本地化序列分析平台主要数据库

Table 1　Main databases used in the local sequence analysis platform

Databases	Version or update time	Sources	Number
Schistosoma japonicum cDNA	2006.08.20	http://www.ncbi.nlm.nih.gov/	9,771
Schistosoma japonicum protein	2006.09.20	http://www.ncbi.nlm.nih.gov/	9,934
Schistosoma japonicum EST	2006.09.20	http://www.ncbi.nlm.nih.gov/	97,526
Schistosoma japonicum genome sequence	2006.07.08	ftp://lifecenter.sgst.cn/pub/sjr_release2/	2,916,497
nr	2006.08.16	ftp:// ftp.ncbi.nih.gov/blast/db/	3,879,234
nt	2006.08.16	ftp:// ftp.ncbi.nih.gov/blast/db/	4,309,818
uniprot	2006.09.25	ftp:// ftp.expasy.org/databases/uniprot/	3,429,309
RepeatMasker Libraries	Ver.11.05	E-MAIL: jola@girinst.org	/
GO[6]	2006.09.25	http://www.godatabase.org/dev	99,714
InterPro	13.0	ftp://ftp.ebi.ac.uk/pub/databases/interpro/iprscan/	/

nr : non-redundant protein database ; nt : non-redundant nucleotide database ; uniprot : universal protein resource; GO: The gene ontology

1.3　核酸及蛋白质序列分析工具

通过互联网, 下载相关基于 GNU 通用许可证(GPL)的, 可免费使用的生物信息学分析软件或通过与软件作者联系, 取得相关的软件, 应用于学术研究。本实验平台所安装的软件主要包括综合序列分析软件包 Staden package[7], EMBOSS[8]; 两序列比对分析软件包 BLAST [9], FASTA [10]; 多序列比对软件包 ClustalW[11], DIALIGN-T[12]; 序列组装拼接软件 Phrap, CAP3 [13]; 重复序列屏蔽工具 RepeatMasker[14]; 基因预测软件 GenScan[15]; 序列进化分析软件 Phylip [16]; 综合蛋白序列分析平台 InterProScan[17]。

1.4　自主设计三大序列分析系统

1.4.1　日本血吸虫 EST 序列的电子延伸系统　构

建该系统所需的生物信息学相关软件包括：同源性检索的 Blast 软件包，序列类组装拼接的 Phrap 软件包或 CAP3 拼接程序，数据库采用日本血吸虫数据库。系统设计方法数据流图如图 1。

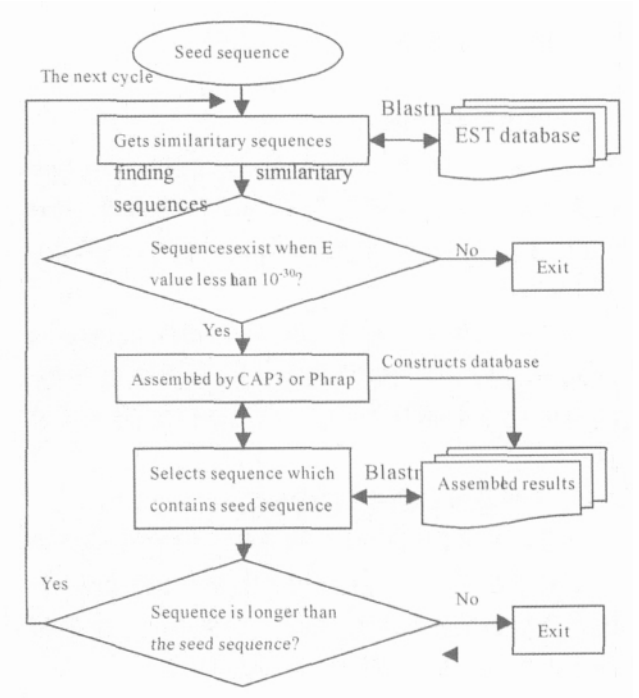

图 1 EST 电子延伸系统数据流图

Fig.1 In silico EST elongating system data flow

1.4.2 基于日本血吸虫 EST 数据库的同源全长 cDNA 序列检索系统 构建该系统所需的生物信息学相关软件包括：电子延伸系统的所需软件，如有需要还可以对待分析的序列进行预处理，包括载体序列的去除（Phrap 软件包中 cross- match）和重复序列的屏蔽（RepeatMasker）。系统设计数据流图如图 2。

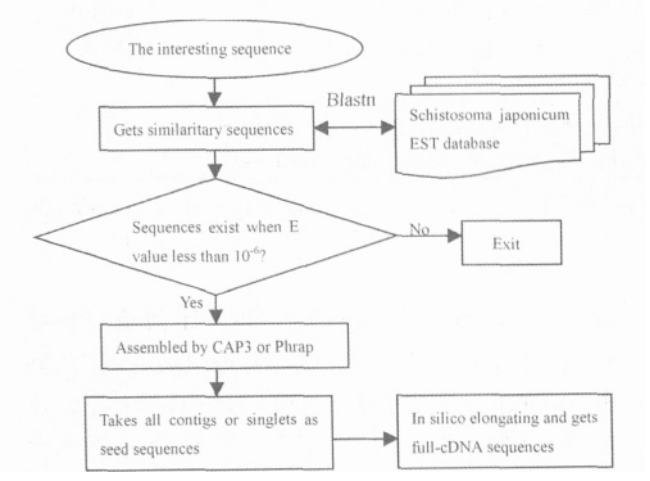

图 2 同源全长 cDNA 序列检索系统数据流图

Fig.2 Homological full-cDNA sequence retrieval system data flow

1.4.3 序列自动注释系统的构建 构建该系统所需的生物信息学相关软件包括：Blast 软件包，InterProScan 软件包（包含 HMMPfam, scanregexp, blastprodom 等 15 个数据库检索或序列分析的程序），相关的数据库包括非冗余蛋白质序列数据库（nr），非冗余核酸序列数据库（nt），以及 Interpro 平台数据库 [17,18]（包括 PROSITE、PRINTS、Pfam、Gene3D、PANTHER、ProDom 和 Smart 等 10 多个数据库的内容，涵盖了 UniprotK 数据库中蛋白质序列的 77.6%）（http://www.ebi.ac.uk/interpro/release_notes.html）。系统设计数据流图如图 3。

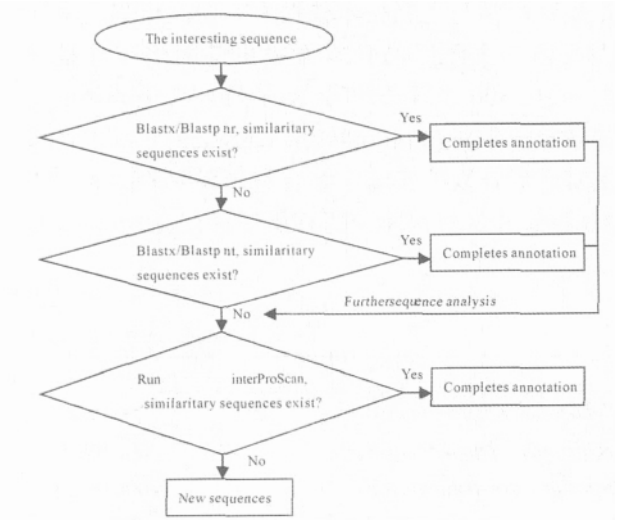

图 3 序列自动注释系统数据流图

Fig.3 Auto sequence annotation system data flow

2 结 果

2.1 日本血吸虫序列分析平台实现的基本功能

由表 1 的数据资源和所列的生物信息学分析工具搭建的序列分析平台，可完成大多数日本血吸虫序列分析的常规任务（具体可参考相关安装软件的使用文档,说明书或参考文献）。包括数据库的相似性搜索（包括非冗余核酸和蛋白质数据库，以及日本血吸虫 EST, 基因组等数据库）、序列的比对（两两比对或多重比对）、蛋白质或核酸 motif, pattern 等的分析、分子进化树的构建、密码子使用频率分析、B 淋巴细胞抗原表位的预测、PCR 引物

的设计、蛋白质二级结构的预测、蛋白质信号肽的预测、计算蛋白质的等电点和相对分子量、酶切位点的识别、绘制质粒图谱、测序峰图的阅读、序列片段的组装拼接，开放读码框的预测或基因结构的预测、GO 分类等。

2.2　日本血吸虫 EST 序列的电子延伸系统的组成与应用

此系统主要由相似性序列检索模块、序列拼接模块、种子序列挑选模块三部分组成，其中相似性序列检索模块完成日本血吸虫 EST 数据库中相似性序列的检索，序列拼接模完成相似性 EST 序列的拼接工作，种子序列挑选模块则是挑选出序列拼接模块结果中生成的包含有种子序列的 contigs。最后系统判断是否对种子序列进行了一次有效的延伸。通过对本实验组获得的 552 条 EST 进行电子延伸和自动分析，显示 487 条得到不同程序的延伸，其中延伸比例超过 100%的有 251 条，延伸前 EST 平均长度 552 bp(中位数)，延伸后序列平均长度 931 bp (中位数)，延伸效率为 168.66%。将延伸后序列对非冗余蛋白质序列数据库(nr)进行检索，可得到更多参考信息[19-21]。

2.3　基于日本血吸虫 EST 数据库的同源 cDNA 序列检索系统的组成与应用

此系统主要由相似性序列检索模块、序列拼接模块、文件拆分模块、电子延伸模块四部分构成。其中相似性序列检索模块主要完成日本血吸虫 EST 数据库中与感兴趣序列相似的 EST 的检索，序列拼接模完成相似性 EST 序列的电子拼接，文件拆分模块完成对电子拼接后生成的 contigs 和 singlets 文件拆分为单序列文件。电子延伸模块则对上面生成的单序列文件中的序列作为种子序列，进行电子延伸，得到假想的全长 cDNA。以已知的人类信号转导分子序列 2226 条作为种子序列，对所建的系统进行测试，同时得到相关的生物学信息。分析结果显示：待分析的 2226 条人类信号转导分子 cDNA 序列中，有 202 条序列可以在血吸虫的 EST 数据库中可以找到同源性的序列，将所得同源性序列电子拼接后所得的序列进行电子延伸后，大部分可延伸一次，也有的可以延伸 10 多次，最终得到假想的全长 cDNA 515 条。将它们对所建的日本血吸虫 cDNA 数据库进行检索，以 E 值 10^{-20} 为标准，416 条可找到相似性序列，其中 217 条与库中的相关序列相似性极高，E 值为 0.00。另外有 99 条可能蕴含未知的基因，有待进一步研究。目前，本课题组正在对可能有意义的基因进行生物学验证。

2.4　序列自动注释系统的组成与应用

此系统主要由 nr 检索模块、nt 检索模块以及 InterPro 检索平台三部分构成。其 nr 检索模块主要完成对待分析序列在非冗余蛋白序列数据库中相似性序列的检索，nt 检索模块主要完成对待分析序列在非冗余核酸序列数据库中相似性序列的检索，InterPro 检索平台则是由 InterProScan 程序完成待分析序列在 InterPro 数据库中的相似性检索，三者都将相关信息赋与待分析序列，进行序列的注释。对所建立的系统，本课题组已经过初步测试，对每一感兴趣的序列进行注释，均可在非冗余蛋白质序列数据库，非冗余核酸序列数据库，InterPro 数据库的检索中得到对应的结果文件，其结果显示，系统运行状态良好。在此基础上，我们还利用上面获得的 515 条日本血吸虫假想全长 cDNA 序列作为待注释的序列，进行结构和功能的分析，旨在获得其相关生物学信息，为血吸虫病疫苗候选分子及药物作用靶标的筛选提供进行生物学实验的依据。

3　讨　论

HGP 初，美国科学家 Venter 等提出了用 EST 技术来进行基因组研究的设想，称为 EST 计划。Adams[22]等第一次应用 EST 技术从人脑组织中获得了 337 个未知基因，在此后的一年里，Adams[23]等继续发展了 EST 技术，开始大规模 EST 自动化测序，获得 2375 条 EST 序列，并将此技术从结构基因组学扩展到功能基因组学和比较基因组学。此后，利用 EST 进行基因的研究得到了飞速的发展，但 EST 自身的局限(仅代表一个完整基因的一部分，同时公共数据库中 EST 测序中存在错误等)及 cDNA 全长序列难以获得等问题，成了制约新基因发现的瓶颈。从生物信息学原理出发，基于公共数据库中的 EST 序列进行电子延伸，已成为众多研究者关注的焦点。目前虽可网上进行 EST 的电子延伸服务，但在使用过程中存在数据安全，响应不及时等问题，因此，本实验组开展了本地化电子延伸系统的构建工作。初步应用本平台已对本实验组获得的 552 条 EST 进行电子延伸，其中 487

条得到了不同程序的延伸。

同源全长 cDNA 序列检索系统为本课题组所特有,它可针对特定的序列,在日本血吸虫 EST 数据库中进行虚拟 cDNA 的查找,对发现日本血吸虫相对应的基因起到了积极的作用,对 2226 条人类信号转导分子序列对日本血吸虫的 EST 数据库进行同源全长 cDNA 序列的检索,其中有 202 条序列可以找到同源性的序列,同时得到假想的全长 cDNA 515 条,并对其进行注释,为日本血吸虫功能基因组及疫苗研究提供了有价值的信息。

本课题组利用公开软件及公共的生物信息学数据资源,建立本地化的日本血吸虫序列分析平台,作为连接生物信息学和实验生物学的桥梁。通过反复的测试和大量数据分析,结果表明,利用公开资源建立的基于基因组学的日本血吸虫序列分析平台已构建成功。

参考方献:

[1] LOVERDE P T, HIRAI H, MERRICK J M, et al. Schistosoma mansoni genome project: an update [J]. Parasitol Int,2004,53(2):183-192.

[2] 袁竹青,吴忠道. 寄生虫基因组:现状及未来展望[J]. 中国热带医学杂志,2003,3(1):121-124.

[3] 袁竹青,吴忠道. 寄生虫基因组计划的启动[J]. 中国热带医学杂志,2003,3(2):230-232.

[4] LIU F, LU J, HU W, et al. New perspectives on host-parasite interplay by comparative transcriptomic and proteomic analyses of Schistosoma japonicum[J]. PLoS Pathog,2006,2(4):e29.

[5] 曹 玲.我科学家公布日本血吸虫‘基因天书”[J]. 发明与创新, 2006, 7:28.

[6] HARRIS M A, CLARK J, IRELAND A, et al. The Gene Ontology (GO) database and informatics resource [J]. Nucleic Acids Res,2004,32(Database issue):D258-261.

[7] STADEN R.The Staden sequence analysis package [J]. Mol Biotechnol,1996,5(3):233-41.

[8] RICE P, LONGDEN I, BLEASBY A. EMBOSS: the European Molecular Biology Open Software Suite [J]. Trends Genet,2000,16(6):276-277.

[9] ALTSCHUL, STEPHEN F, THOMAS L, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. Nucleic Acids Res,1997, 25:3389-3402.

[10] MILLER P L, NADKARNI P M, CARRIERO N M. Parallel computation and FASTA: confronting the problem of parallel database search for a fast sequence comparison algorithm[J]. Comput Appl Biosci, 1991, 7(1):71-78.

[11] THOMPSON, J D, HIGGINS D G, GIBSON, T J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice [J]. Nucleic Acids Research, 1994, 22(22):4673-4680.

[12] SUBRAMANIAN A R, WEYER-MENKHOFF J, KAUFMANN M, et al. DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment [J]. BMC Bioinformatics, 2005, 6:66.

[13] HUANG X, MADAN A. CAP3: A DNA sequence assembly program[J]. Genome Res,1999,9(9):868-877.

[14] BEDELL J A, KORF I, GISH W. MaskerAid: a performance enhancement to RepeatMasker [J]. Bioinformatics, 2000, 16(11):1040-1041.

[15] BURGE C, KARLIN S. Prediction of complete gene structures in human genomic DNA[J]. J Mol Biol,1997, 268(1):78-94.

[16] RETIEF J D. Phylogenetic analysis using PHYLIP [J]. Methods Mol Biol,2000,132:243-58.

[17] ZDOBNOV E M, APWEILER R. InterProScan-an integration platform for the signature-recognition methods in InterPro[J]. Bioinformatics,2001,17(9): 847-848.

[18] APWEILER R, ATTWOOD T K, BAIROCH A, et al. InterPro, progress and status in 2005[J]. Nucleic Acids Res,2005,33(Database issue):D201-205.

[19] 刘翰腾, 吴忠道, 邹赛德, 等.日本血吸虫 EST 序列的电子延伸及结果分析 [J]. 中国寄生虫学与寄生虫病杂志,2004,22(4):218-222.

[20] 邵 筱, 余新炳, 吴忠道, 等. 日本血吸虫(大陆株)成虫表达序列标签的获取及电子延伸 [J]. 中国地方病学杂志, 2004, 23(04):297-299.

[21] 邵 筱, 吴忠道, 刘翰腾, 等. 应用 EST 和电子克隆策略研究血吸虫表达基因谱 [J]. 基础医学与临床, 2005, 25(7):602-606.

[22] ADAMS M D, KELLEY J M, GOCAYNE J D, et al. Complementary DNA sequencing: expressed sequence tags and human genome project[J]. Science, 1991, 252(5013):1651-1656.

[23] ADAMS M D, DUBNICK M, KERLAVAGE A R, et al. Sequence identification of 2,375 human brain genes [J]. Nature,1992,355(6361):632-634.

(编辑 刘清海)