文章编号:1000-7423(2004)-04-0218-05

【论著】

日本血吸虫 EST 序列的电子延伸及结果分析

刘翰腾[1*]，吴忠道[2]，邹赛德[1]，邵筱[2]

【摘要】 **目的** 建立日本血吸虫表达序列标签（SjEST）序列自动分析系统，筛选新基因、分析其表达谱。**方法** 建立本地化日本血吸虫专业数据库，整合序列同源性比较软件（BLAST）及片段整合分析软件（PHRAP），运用生物信息学策略编写程序控制 EST 自动延伸。建立本地化蛋白库，对延伸结果进行蛋白库同源性分析。实现对成批数据大规模自动分析，筛选出可能的新基因全长 cDNA 序列，并进行基因表达谱分析。 **结果** 延伸系统规则有效，延伸结果序列与原始序列高度同源。对 552 条 EST 进行自动分析，487 条得到不同程度的延长，其中有 104 条 EST 原始序列检索无同源性，但经过延伸后获得高度同源性的联配序列信息。根据延伸结果尝试分析基因表达谱，筛选出 27 个可能新基因。 **结论** 建立了本地化的有效的 SjEST 序列自动分析系统。该系统为新基因的筛选及基因表达谱的分析提供了重要的参考信息。

【关键词】 **日本血吸虫；表达的序列标记；电子序列延伸；序列分析；计算生物学**

中图分类号：R383.24 文献标识码：A

The *in Silico* Elongation and Analysis of the EST from *Schistosoma japonicum*

LIU Han-teng, WU Zhong-dao, ZOU Sai-de, SHAO Xiao

(*Computer Centre, the School of Pre-Clinical Medicine, Sun Yat-Sen University, Guangzhou* **510089**, *China*)

【**Abstract**】 **Objective** To construct a platform for in silico elongation and batch analysis of *Schistosoma japonicum* (Sj) ESTs, acquire the potential novel genes and research the expression profile of the genes. **Methods** On the basis of Linux operating system and local ESTs database of Sj, the BLAST and PHRAP softwares were used to construct a program to achieve the elongation of ESTs. Stand-alone BLAST search against the nr database helped analyze the elongated sequence. After finishing the batch analysis script, the platform was used to research the Sj gene expression profile and acquire the potential novel genes. **Results** The platform showed satisfactory efficiency and fidelity. 487 elongated sequences obtained from 552 and 307 elongated sequences showed high homology within the nr database downloaded from NCBI. Furthermore, 104 elongated sequences displayed significant homology but showed no homology before elongated. 27 potential novel genes were filtered out. **Conclusion** An effective platform for Sj ESTs data mining was accomplished and further information on the potential novel genes was acquired.

【**Key words**】 *Schistosoma japonicum*; Expressed sequence tags (EST); In silico elongation; Sequence analysis; Computational biology

Supported by the National Natural Science Foundation of China (No. 30070683)

表达序列标签（expressed sequence tag，EST）是对随机选取的 cDNA 末端测序获得的部分片段，它提供了表达基因相关的“标记”，因此为发现新基因、揭示基因表达以及调节信息等提供了一种快速、有效的途径[1]。研究者用 cDNA 文库大规模测序的策略可获得大量的 EST 序列，而全长 cDNA 序列的获得则成了新基因发现的瓶颈，制约着后续基因表达及蛋白功能的研究。因而充分利用公共数据库的 EST 序列或较长的 cDNA 序列对新获得的 EST 序列进行电子延伸，再通过同源性比较获得如全长、可读框、模序、功能预测等信息，能获得更多的序列分析信息。

世界卫生组织（WHO）于 1995 年 1 月发起了全球科学家携手共同开展血吸虫基因组计划的倡议，同年 Franco 等[2]运用 EST 策略快速发现 154 个新的曼氏血吸虫基因，从此 EST 策略被广泛应用于血吸虫基因组研究。1999 年联合国（WHO）开发计划署（UNDP）世界银行血吸虫基因组网络年会确定，日本血吸虫新基因的发现是 WHO 主要研究内容之一。截止 2003 年 7 月 GenBank 中日本血吸虫 EST 数据已达45 900条，解决上述新基因发现的瓶颈问题已成为血吸虫基因组主要的研究内容之一。本文尝试利用现有生物信息软件资源、公众数据库以及已有日本血吸虫 EST 数据，构建基于 Linux 平台的 EST 自动分析

基金项目：国家自然科学基金(No. 30070683)

作者单位：中山大学基础医学院 1 计算机中心，2 寄生虫学教研室，广州 510089； * 现在单位：广州市卫生局信息鉴定和评估中心，广州 510080

体系，完成序列延伸、同源检索分析等功能。并利用系统运行结果分析日本血吸虫（大陆株）成虫的基因表达谱，筛选可能的新基因。

材料与方法

1 原始数据

552 条日本血吸虫 EST 待分析序列，取自日本血吸虫（大陆株）成虫 cDNA 文库[3]。

2 电子序列延伸系统构建策略

① 利用序列同源性比较软件（例如 BLAST 软件[4]）将待进行电子延伸的序列（种子序列）对指定数据库进行同源性检索；② 按一定的标准从数据库里挑选符合要求的全部相关序列（本系统采用标准：BLASTN 同源性比较时 e-value＜1e-30）；③ 对筛出的序列群进行片段整合分析即 contig 分析，本系统使用片段组装程序（phragment assembly program, phrap)[5,6]软件；④ 从 contig 分析结果（含有 1 个至多个的 contig 序列）选择合适的 contig 序列，作为新生种子序列；⑤ 重复上述步骤，直至新生序列不能被进一步延伸为止。

3 系统环境及软件环境

硬件条件：曙光 X230XP，CPU（2.4 GHz，双 CPU），1 GB 内存，120 GB SCSI 硬盘。操作系统：采用 Red Hat Linux 8.0 版本[7]。软件：序列同源比较软件 blast for Linux（ftp: //ftp. ncbi. nih. gov/blast/executables/blast. linux. tar. z）blastall 2.2.5 版本。用于 contig 分析的 PHRAP 软件，通过 E-mail 联系软件作者获取（Phil Green: phg@u. Washington. edu）。数据库：从 NCBI 下载日本血吸虫全部 EST 作为本地的 EST 数据库（共下载45 900条，下载时间 2003-05-09）；从 NCBI 网站下载非冗余蛋白质序列数据库即 nr 蛋白库，为氨基酸序列数据库包括：非冗余 GenBank 编码序列（coding sequence, CDS）的翻译序列，PDB，SwissProt，PIR，PRF 等（2003-11-07 更新）。

4 程序设计发行

序列延伸部分：细化序列延伸策略，将上述 ①～③ 3 个步骤组织到一个程序段中；上述 ④ 步骤需要经常测试修改以选择最佳 contig 作为下一轮延伸的种子序列，故单独做成一个程序段；上述 ⑤ 步骤控制循环的执行作为主线程序段。

序列延伸结果分析部分：进行核酸序列对本地 nr 蛋白库的 BLASTX 检索，实现批量序列对库检索；BLASTX 结果的解读，生成联配分值（score）超过 100 的匹配蛋白序列的报表（对单一查询序列），生成与延伸序列最佳匹配的蛋白序列的报表（对成批查询序列）。

5 系统效能评价

通过原始序列与延伸后结果两两对比，评价延伸保真性、延伸效率；通过原始数据与延伸后结果对蛋白库同源检索信息的前后比较，评价该系统的价值。

6 分析血吸虫基因表达谱、筛选可能新基因

根据延伸后结果对蛋白质库比对结果作表达谱分析，将比对联配结果中未出现相关蛋白序列而出现其他物种高同源性蛋白质的 EST 序列认为是尚未发现的日本血吸虫新基因。

结　果

1 序列分析系统

该系统由两部分组成：即序列延伸部分及序列结果分析部分。考虑系统需要调用多次控制台指令，且 Shell[8]脚本程序开发的高效性，易移植性，本体系由一系列的 bash（bourne again shell）脚本程序构成。

2 序列延伸部分

延伸部分有 3 个程序段：elong. sh、selectseq. sh、autoelong. sh。其中 elong. sh 完成对种子序列的一轮电子延伸；selectseq. sh 选出上一轮延伸结果中较合适的 contig 作为下一轮的新生种子序列；autoelong. sh 完成对上述两个程序段的循环调用，直至最后的新生序列和上一轮延伸序列的长度变化在 5 个碱基范围内。

以下简要说明 3 个程序段里的主要指令调用：

① elong. sh 脚本中关于 BLAST、PHRAP 程序调用的语句：

blastall-p blastn-m 8-b 500-d/ncbi/db/schi-J -i $2 -o $2-blastn. out-e 1e-30

使用 BLAST 软件将种子序列 $2 对日本血吸虫 EST 数据库进行同源检索，取 e-value＞1e-30 结果格式为表格式，输出结果到文件 $-blastn. out。

phrap $2-ests＞phrap. out

使用 PHRAP 软件对种子序列和上面 BLAST 检索所得结果序列的整合列表进行 contig 分析，生成结果文件 $2-ests. contigs。

② selectseq. sh 脚本从 *-ests. contigs 中选取合

适 contig 的主要指令语句：

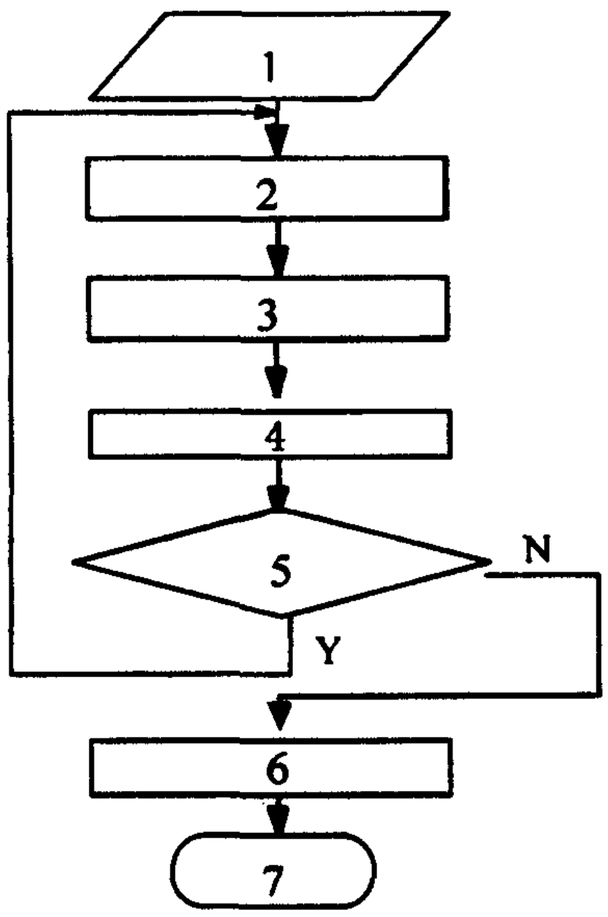
formatdb-p F-o T-i $1-n contigdb

将得到的 contigs 格式化成自制的数据库 contigdb。

blastall-p blastn-d contigdb-m 8-i $2-o ${2}-contig. out

用原始种子序列 $2 对自制数据库进行同源检索，输出报表格式结果文件 $2-contig. out。从结果文件中按 score 从高到低的顺序提取 contig 序列，判断新生 contig 是否较原始序列长，条件符合则输出为下一轮的种子序列。

③ autoelong. sh 脚本循环调用 elong. sh，selectseq. sh；判别规则：若上述 selectseq. sh 产生的新生种子序列较上一轮的延伸序列短，则退出循环（图 1）。



1 读入原始 EST 序列 2 将输入序列作为待延伸序列 3 调用 elong. sh，得到拼接结果 *. contigs 4 调用 selectseq. sh，选取与原始序列最相似 *. contig 序列 5 Contig 是否长于待延伸序列 6 删除临时文件，打包过程文件，报告延伸结果 7 终止

1 Input the original sequence 2 Treat input sequence as seed sequence 3 Call elong. sh, finish segment assembly, output the result in the file *. Contigs 4 Call selectseq. sh, select the contig from *. Contigs, which is the most homological one comparing with the original sequence 5 While the contig length longer than original one, loop to step 2 6 Delete the temporary files, compress the process files, output the elongated sequence 7 Exit

图 1 Autoelong. sh 程序流程图

Fig. 1 The flow chart of the autoelong. sh script program

3 序列结果分析部分

3.1 Blastthedir. sh 自动完成所有子目录内核酸序列的对库检索（本体系每个 EST 及其延伸结果被分别置于单独的子目录）。 主要程序语句为：

blastall-p blastx-m 8-b 500-d/ncbi/db/nr-031027/nr-i ${firstfile} -o report. first 及 blastall-p blastx-m 8-b 500-d /ncbi/db/nr-031027/nr-i ${lastfile} -o report. last。

将延伸前后序列对 nr 蛋白库进行同源检索生成 report. * 的结果文件；循环调用上述指令可自动完成所有子目录的联配检索。

callana. sh；tableout. sh 完成 BLASTX 结果文件的二次分析，对联配结果进行筛选，按课题组的要求，生成 score 值＞100 有意义联配序列的结果报表；包含延伸前后序列长度、最高匹配序列名称、score 值，汇总匹配数目的字段的汇总报表。一系列过程可以全部整合在一个程序里，输入批量的待分析 EST 序列（FASTA 格式），自动完成序列延伸、延伸后结果对蛋白库同源检索分析、BLAST 结果二次分析、统计报表输出。该流程还可以继续向下扩展以满足不同的需要。

3.2 EST 数据及其延伸结果分析 利用该系统对本课题组获得的 552 条 EST 进行电子延伸和自动分析，结果显示 487 条 EST 序列得到不同程度的延伸；其中延伸比例超过 100% 的有 251 条；延伸前 EST 长度均值 553 bp（中位数，下面均值量亦同），延伸后序列长度均值 1 008 bp，两值比出延伸效率为 182.3%。

为检验电子延伸的可靠性，将上述可被延伸的 487 条 EST 与其延伸结果序列分别进行两序列联配比较（bl2seq）。汇总结果显示，487 条 EST 原始长度的均值为 553 bp，而 487 条 EST 与其各自延伸序列作 bl2seq 比较时重叠区长度的均值为 477 bp，重叠区一致性（identity）的均值为 99.7%，联配分值（score）的均值为 902.50。表明延伸序列与原始序列是高度同源的，因而保证了用延伸后序列进行生物信息学分析时，不会遗漏与原始序列有关的同源性信息。

与 nr 蛋白库联配比对结果显示，延伸后序列提供更多可参考的联配序列信息。将延伸前后两组序列对蛋白库进行同源检索 BLASTX，取两组检索结果中 score 值最高的序列联配数据进行前后对比：直接比较两组联配数据中匹配长度、一致性（identity）、score 值的均值（中位数），可见延伸后序列与目标蛋白序列之间具有更好的匹配性(表 1)。

按 BLASTX 分值 0～100 为界限分为 3 组序列，从表 2 比较可看出延伸后序列对蛋白库能检索出更多有意义的匹配序列（score≥100 者）。

为检测延伸前后对库比对联配最高分值之间是否存在明显差异，选用 wilcoxon 符号秩检验；建立单侧

检验假设；用统计软件包 SPSS 对前后两组 552 对 scores 值进行分析，求得检验统计量 $Z=-14.45$，对应的单侧概率为 $P<0.01$；故在 $\alpha=0.01$ 水平上拒绝零假设，可认为延伸后联配最高 score 中位数高于延伸前组别。

3.3 日本血吸虫基因表达谱 将序列延伸后结果对 nr 蛋白库进行 BLASTX 比较，过滤出其中 scores 值 >100 的序列，共 316 条；将分值最高联配的蛋白序列作为该 EST 的已知序列或同源序列；对上述联配序列进行功能归类（表 3），结果显示：42 个序列与核糖体相关蛋白同源，9 个序列与线粒体相关蛋白同源，108 个序列同源于各种酶类，157 个序列同源于其他各种功能蛋白。同时发现有 27 个基因是多拷贝，其中有 4 个基因在此文库中是高度冗余的（拷贝数>5），分别是：细胞色素 C 氧化酶亚单位Ⅳ、卵壳蛋白、组织蛋白酶 B 样的丝氨酸蛋白酶和动力蛋白轻链。

在这 316 条延伸后序列的对库联配 score>100 的结果中，有 240 条的联配蛋白列表包含日本血吸虫相关蛋白；其余的 76 条最高联配蛋白中除去线粒体、核糖体相关蛋白和未知蛋白，剩余的 27 条可能代表血吸虫物种尚未发现的新基因。如名称 JAYG0050 的 EST 序列，延伸后提示与旋毛虫原肌球蛋白高度同源，可能编码免疫相关的蛋白；还有一些序列可能编码各种酶如蛋白激酶 C、carbonic anhydrase Ⅱ、组织蛋白酶 L 等。

表 1 延伸前后两组联配数据均值（中位数）比较表
Table 1 Comparison of the median between two gathering groups of sequences

552 组联配数据的均值 Median between two gathering groups of sequences	相似度 Identity(%)	匹配长度 Alignment length(rps)	联配分值 Score
原始序列组 Original sequences	48.4	82	65.08
延伸后序列组 Sequences after elongated	51.7	129	129

表 2 延伸前后两组按比对后联配最高分值分组汇总比较表
Table 2 Comparison of scores in two gathering groups of sequences

原始序列对库检索最高分值序列分组 Grouping of highest scores from original sequences' blastx result	延伸后序列对库检索最高分值序列分组(条) Grouping of highest scores from elongated sequences' blastx result			合计 Total
	0	1~99	≥100	
0	13	14	15	42
1~99	14	191	89	294
≥100	0	4	212	216
合计 Total	27	209	316	552

表 3 延伸后序列分析所得表达谱
Table 3 Expression profile of sequences after in silico elongation

功能归类 Category	该类序列条数 Number of match	所占比率 Ratio (%)
未知序列(无同源蛋白) No match	236	42.8
核糖体相关的 Ribosomal interrelated	42	7.6
线粒体相关的 Mitochondrial interrelated	9	1.6
酶类 Enzymes interrelated	108	19.6
信号和调控蛋白 Regulatory and signal proteins	32	5.8
结构蛋白 Structural proteins	38	6.9
其他蛋白 Other proteins	87	15.8
合计 Total	552	100

讨 论

本研究完成了基于本地日本血吸虫 EST 数据库的电子延伸并建立序列自动分析系统。应用该系统对日本血吸虫 EST 序列进行电子延伸，并比较延伸前后序列，显示电子延伸环节是有效的（平均延伸效率 182.3%）、保真的（bl2seq 比较得 score 均值 902.5）。对 nr 蛋白库同源检索结果分析，可看出延伸后序列可获得更多有价值的联配序列，如 104 条原始 EST 序列检索不到高同源性的蛋白序列，但经延伸后序列

则可检索到高同源性的蛋白序列；批量数据分析对于基因表达谱分析、新基因筛选具有重要的参考意义。

目前 NCBI 或美国基因组研究所（The Institute for Genome Research，TIGR）已有多个物种的基因索引数据库，但仍未分出针对日本血吸虫物种的基因索引数据库。因而在线分析 EST 序列的时候只能对全物种的 EST 库作 BLASTN 比较，再将联配意义显著的目标序列列表输入给在线拼接程序如 Cap3 或其他 assembler，等待运行结果，将结果序列再对库进行 BLASTN 比较，如此循环。不仅分析周期长，手工操作烦琐，而且非血吸虫物种的同源性 EST 也进入拼接程序的列表中，可能会对拼接结果造成一定程度的误导。针对专业数据库的本地化分析，克服了上述的种种弊端，尤其是高通量的数据分析服务，大大加速了新基因筛选，分析表达谱的周期。

根据最高同源性蛋白功能推测 EST 的功能并据此描述日本血吸虫大陆株成虫表达谱，结果显示编码各种酶的相关基因占有很大比重（108/316）；而分析那些冗余序列可见 4 个高度冗余序列中有两个属于酶（即细胞色素 C 氧化酶亚单位Ⅳ和组织蛋白酶 B 样的丝氨酸蛋白酶），并且大部分冗余序列具有看家基因功能（如核酸合成、氨基酸合成和一般代谢调控），这与其他物种表达谱的无差别。

对序列自动分析结果进行筛选得到的 27 个可能的新基因编码的蛋白作用于重要的生理过程和免疫环节，DAD 样蛋白质（DAD-like protein）、原肌球蛋白、组织蛋白酶 B、纤维蛋白、翻译后调控瘤蛋白（Translated controlled tumor protein，TCTP）等。因而，延伸后序列的对库联配搜索有助于发现更多编码免疫相关蛋白的基因，为进一步研究这些基因提供了新的有价值信息。

本系统为 Sj EST 批量数据的处理及新基因的筛选提供了有效的生物信息学工具，为充分利用已有的血吸虫基因数据加快血吸虫病疫苗的研究和新药的靶点的发现提供了有效的手段，并为开发其他物种 EST 数据分析平台提供借鉴。

参 考 文 献

[1] Adams MD, Kelley JM, Gocayne JD, *et al*. Complementary DNA sequencing: expressed sequence tags and human genome project[J]. Science, 1991,252:1651-1656.

[2] Franco GR, Adams MD, Soares MB, *et al*. Identification of new *Schistosoma mansoni* genes by the EST strategy using a directional cDNA library[J]. Gene, 1995,152:141-147.

[3] 吴忠道,余新炳,徐劲,等. 日本血吸虫(大陆株)成虫基因表达谱的研究[J]. 中山医科大学学报, 2002,23:401-404.

[4] Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. Nucleic Acids Res, 1997,25:3389-3402.

[5] http://bozeman.genome.washington.edu/phrap.docs/phrap.html.

[6] Ewing B, Green P. Base-calling of automated sequencer traces using phred. Ⅱ. Error probabilities[J]. Genome Res, 1998,8:186-194.

[7] Arman Danesh 著. 邱仲潘等译. Red Hat Linux7 从入门到精通[M]. 北京:电子工业出版社, 2001.1-562.

[8] Sriranga Veeraraghavan 著. 卢涛译. 精通 shell 编程[M]. 第 2 版. 北京:人民邮电出版社, 2003.1-350.

（收稿日期:2003-12-18 编辑：富秀兰）

文章编号:1000-7423(2004)-04-0222-01

【病例报告】

妊娠并发骨盆棘球蚴囊肿一例

陈洪俊

中图分类号：R532.32 文献标识码：D

棘球蚴病又称包虫病。作者在阿尔及利亚工作期间，曾诊治妊娠妇女并发骨盆棘球蚴囊肿 1 例，报告如下。患者女性，28 岁，孕妇。1 年前曾做过肝棘球蚴囊肿摘除手术。此次住院是因为骨盆有一肿块并与妊娠子宫相连。妇科检查证实子宫妊娠 5 月余，子宫颈被道格拉斯凹陷处一肿块压迫，该肿块质地硬、固定。盆腔“B”超检查发现在道格拉斯凹陷处有一液态肿块，由 6 cm×5 cm 和 7 cm×8 cm 两叶组成。剖腹手术证实患者已怀孕 5 月余，并发现在道格拉斯凹陷处有 2 个囊肿。手术摘除后经病理学检查证实为棘球蚴囊肿。手术后患者妊娠过程正常，并于术后 4 个月正常分娩。

妊娠并发骨盆棘球蚴囊肿病例十分罕见。一般认为骨盆棘球蚴囊肿是肝脾棘球蚴囊肿破裂后的并发症，也有学者认为它的形成是因为血液传播所致。骨盆“B”超检查可对棘球蚴囊肿与妊娠情况一目了然，根据患者既往病史及“B”超检查即可确诊。骨盆棘球蚴囊肿可能产生的并发症有感染、早产、难产、因分娩用力造成囊肿破裂而导致的过敏性休克等，有时是致命的。妊娠期骨盆棘球蚴囊肿，治疗手段主要是在超声波显示下穿刺抽液及外科手术摘除，棘球蚴囊肿合并妊娠穿刺危险性较大，特别是棘球蚴囊肿靠近胎儿时。作者认为，外科手术摘除则比较安全，但选择外科手术时间与分娩方式，可考虑：① 如果在妊娠初期发现囊肿，应选择在孕期 6 个月左右实施棘球蚴囊肿摘除术，并尽量使妊娠继续。本例即在 5 个月顺利实施手术，术后妊娠继续。② 如果在妊娠末期确诊骨盆棘球蚴囊肿，如病情允许，可在妊娠足月时采用剖腹产，同时进行骨盆棘球蚴囊肿摘除术，否则要提前中止妊娠，行急诊剖腹产术加棘球蚴囊肿摘除术。

（收稿日期：2004-04-06 编辑：富秀兰）

作者单位：湖北省荆州市第一人民医院普外科，荆州 434000