文章编号:1001 - 2486(2006)04 - 0037 - 06

适合可变剪接研究的转录组序列分析策略*

王正志[1],李稚锋[1,2],杭兴宜[2],毛逸清[3],骆志刚[4],赵东升[3],张成岗[2]
(1. 国防科技大学 机电工程与自动化学院,湖南 长沙 410073;
2. 军事医学科学院 放射与辐射医学研究所,北京 100850;
3. 军事医学科学院 卫生勤务与医学情报研究所,北京 100850;
4. 国防科技大学 并行与分布处理国防科技重点实验室,湖南 长沙 410073)

摘 要:规模化基因表达实验所产生的大量与生物组织特定时空状态相关的 cDNA 和表达序列标签(EST)等信息可用于新基因的发现、基因表达模式分析和基因组的注释,从而可为转录组研究提供实验设计和结果分析的参考标准。真核基因可变剪接的普遍性及其在机体生理与病理过程中的重要作用,使得可变剪接的系统分析已成为功能基因组研究中的热点之一。在面临海量表达数据的指数增长和不断有新的基因组获得测序的情况下,实现转录组序列分析的规模化、自动化计算迫在眉睫。讨论不同转录组分析系统中的数据分析算法及其计算需求,并提出适用于大规模可变剪接分析的策略。

关键词:转录组;EST 聚类;EST 装配;可变剪接;高性能计算
中图分类号:TP393 文献标识码:A

The Strategy of Transcriptome Analysis for Alternative Splicing Research

WANG Zheng-zhi[1], LI Zhi-feng[1,2], HANG Xing-yi[2], MAO Yi-qing[3], LUO Zhi-gang[4], ZHAO Dong-sheng[3], ZHANG Cheng-gang[2]
(1. College of Mechatronics Engineering and Automation, National Univ. of Defense Technology, Changsha 410073, China;
2. Institute of Radiation Medicine, Academy of Military Medical Sciences, Beijing 100850, China;
3. Institute of Health Administration and Medical Information, Academy of Military Medical Sciences, Beijing 100850, China;
4. National Lab of Parallel and Distributed Processing, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Experiments on transcriptome analysis have resulted huge genes expression data related with specific temporal and spatial information. These data can be used in new genes identification, analysis of genes express patterns and the annotation of genomes, which may provide the reference standard for experiment design and result analysis of transcriptome experiments. Since the alternative splicing of eukaryotic genes have found to be universal and play an important role in physiology and pathology, systematic analysis of alternative splicing is becoming a new hotspot of functional genome research. Facing the immense and exponential increase of experimental express data and more new genomes getting sequenced, there is exigent of the strategy which can handle transcriptome sequences in large scale and automatic way. We elucidate the algorithms, the computing requirements and programs in different transcriptome sequences analysis systems and propose a strategy more suitable for large scale analysis of alternative splicing.

Key words: transcriptome; EST clustering; EST assembly; alternative splicing; high performance computing

大规模基因表达分析是转录组研究的核心,它能够直接揭示参与特定生命过程的基因及其状态的变化,其发展导致蕴含基因表达信息的表达序列标签(EST)数据快速增长。最早由美国 NCBI 于 1992 年创建的 EST 数据库(dbEST, http://www.ncbi.nlm.nih.gov/dbEST/)的记录数量已达 3289 万条,其中来自于人的已超过 760 万条。基因的数量、表达模式以及可能的功能一直是转录组研究需要回答的问题,而近年来可变剪接普遍性的发现使得这个问题更加复杂[1]。由此,数据分析方法的效率和可变剪接的分

* 收稿日期:2006 - 03 - 21
基金项目:国家并行与分布处理国防重点实验室基金资助项目(51484050304JB4401);军事医学科学院科技创新启动基金资助项目(0401001);0402013)
作者简介:王正志(1945—),男,教授,博士。

析已成为当前转录组研究策略需要考虑的两个重要方面。本文综述不同转录组序列分析系统中的数据分析算法和相应的计算需求,并提出适用于大规模可变剪接分析的策略。

1 目前转录组序列分析的基本路线

以往建立较早、可用于转录组序列辅助分析的系统有 NCBI 的 UniGene 数据库(http://www.ncbi.nlm.nih.gov/UniGene),TIGR 的 TIGR Gene Indices(http://www.tigr.org/tdb/tgi/)以及 SANBI 的 STACK(http://www.sanbi.ac.za/Dbases.html)。这三个系统的数据分析方法可概括为三个步骤:聚类前预处理(pre-processing)、聚类(clustering)、装配(assembly)[2]。目前大部分转录组序列分析系统的 EST 聚类都是基于序列之间存在相似的重叠部分。对重叠的宽度和相似度设定不同阈值将产生不同的聚类效果。由于 EST 是基因表达序列的片段,因此通常需要装配产生更完整的转录本的一致序列。Bouck 等曾比较了以上三个系统在利用的数据集、聚类参数、问题序列处理和可变剪接转录信息这些方面的差别,指出它们侧重提供的信息不同,各有利弊[3]。UniGene 利用的数据除了 dbEST,还包括 GenBank 中的 mRNA 序列,聚类标准相对宽松,使得剪接变体可以被包括在同一类中,虽然它本身不进行装配,但其结果易于进行可变剪接的分析;TIGR Gene Indices 聚类后进行装配,区分剪接变体,获得更长的假想共有序列(TCs, tentative consensus sequences);STACK 只利用 dbEST 的数据,先将 EST 数据按组织和状态信息进行分类,然后再进行聚类、装配,可以分析不同组织或不同状态下表达的剪接变体形式。

基于相似性的 EST 聚类方法,最早是采用两两比对的方法分析序列之间的相似性,由于存在计算效率方面的需求,出现了多种改进方法以及不同方法的并行化实现[4-12]。然而单纯利用表达序列之间的相似性进行聚类,面临的问题是相似基因容易被错误地聚类,基因组污染序列、嵌合序列(不同基因来源序列的嵌合)、嵌套序列(基因组序列重叠,但非同一基因)也会导致错误聚类和错误装配,影响可变剪接分析的结果[13]。直接装配获得的一致序列之间的比对可以进行可变剪接分析,然而,没有基因组信息将很难了解基因的剪接结构和可变剪接的调控信息。随着基因组数据的完善,有必要将基因组信息纳入 EST 聚类和 EST 装配过程。以下针对转录组序列分析各阶段的常用算法进行评述,并指出其中的优化策略。

2 聚类前预处理

EST 测序策略为随机选择 cDNA 克隆,从克隆插入的一端或两端开始的单遍测序。EST 通常为 300~700 个碱基长度,由于只进行单遍测序,测序错误率为 1%~3%。除了测序错误,EST 序列存在的问题还包括载体序列的污染、非基因组源序列的污染、基因组序列的污染、以及人工嵌合序列。因此在挖掘 EST 数据中的生物学意义时,需要了解这些问题对聚类产生的潜在影响,并建立可靠的分析策略。

基因组序列的污染和人工嵌合序列的污染难以与真实序列相区别,需要在聚类过程中根据一定的准则进行过滤。载体序列、非基因组源序列、重复序列和低复杂度序列片段会对基于相似性的聚类产生严重影响,因此需要预先对这些序列进行掩盖标记(mask)。方法是将所有序列与载体序列库(ftp://ftp.ncbi.nih.gov/pub/UniVec/)、已知重复序列库 Repbase[14]和可能的非基因组源序列(通常是线粒体和核糖体的 DNA 序列)进行比对,因此需要进行 $n_{mask} \times N_{est}$ 次的序列比较运算,其中 n_{mask} 是要标记的模板序列数,而 N_{est} 是待标记处理的序列个数。常用程序为 RepeatMasker (http://www.repeatmasker.org),它可以基于库文件进行标记处理,并能检测低复杂度序列片段。它的比对工具采用敏感的 cross_match,计算量很大。

快速预处理可以使用 RepeatMasker 的加速程序 MaskerAid。MaskerAid 利用 WU BLAST,速度是原始 RepeatMasker 的 30 倍[15]。但在配备 2GHz CPU 的服务器上处理整个人类基因组仍需要一周,处理全部 EST 数据以及 GenBank 中的 mRNA 序列则需要更长的时间。因此,对于海量数据集的预处理,应使用配备有足够内存和高速磁盘的高性能计算机系统(SMP 或机群)。

3 EST 聚类

对 EST 序列进行合理聚类是进行转录组分析的重要环节。“合理聚类”要求尽可能减少 EST 的聚

类错误。EST 聚类的错误类型可简单分为两种,第一类错误是属于同一基因的序列而未能聚为一类,即假阴性 ;第二类错误是将不属于同一基因的序列聚在一起,即假阳性。对于基于相似性的聚类方法,由于 EST 对基因的覆盖具有 3 '端和 5 '端的偏好性,容易产生第一类错误。因此通常借助长度更长的 mRNA 序列,并在聚类结果的基础上考虑表示同一克隆对的信息,将属于同一基因但没有重叠的序列聚为一类。而相似基因、嵌合基因和嵌套基因的存在容易使聚类产生第二类错误,利用基因组信息可以筛查出这些可疑序列。

3.1 基于相似性的 EST 聚类

基于相似性的 EST 聚类,需要针对全部序列两两之间进行相似性分析,理论上需要进行约 N_{est}^2 次的比对运算,因此在表达序列呈指数增长的情况下,通过并行化以及其他改进策略以提高聚类的计算效率是十分必要的。UniGene、TIGR Gene Indices 利用两两比对方法进行相似重叠分析,最初采用 BLAST 工具,目前则使用 megaBLAST 进行比对。megaBLAST 是 NCBI 开发的适合只有非常小的差异的序列之间的比对程序,比 BLAST 速度提高 30 倍[4]。TIGR Gene Indices 使用 megaBLAST 的改进版本 mgBLAST。另外 BLAST 的并行版本也得到了发展,其中 mpiBLAST[5] 具有超线性的加速比。STACK 采用 d2 _ Cluster[6] 进行初始聚类,其中 d2 算法为基于词匹配快速确定序列差异的方法,比两两比对方法分析相似重叠的速度稍快,目前 SANBI 也推出了 d2 _ Cluster 的并行化版本[7]。计算复杂度与序列数量的关系小于二次方的改进方法包括基于散列表的方法,如 UIcluster[8]、CLU[9]、PECT[10] 和基于后缀树结构的方法,如 xsact[11]、PaCE[12],基于后缀树结构的方法计算复杂度理想情况能达到 $O(N_{est})$。这些方法都有并行策略的实现,或者直接只推出并行版本。

由于难以有已知的正确结果来作评价参考,因此常常用 Jaccard 指数(Jaccard index = $\frac{a}{a+b+c}$)来评价不同方法的聚类相似性,其中 a 为在两种聚类方法中都聚在一个类当中的序列对个数,b 为在第一个聚类方法中被聚在一类而在第二个聚类方法中不在一类的序列对个数,c 为在第一个聚类方法中不在一类而在第二个聚类方法中被聚在一类的序列对个数。小规模的 EST 数据分析表明[11],xsact 与 d2 _ cluster 以及 BLAST 的聚类结果比较接近,而离 UIcluster 较远,但均离实际的 UniGene 结果较远,这是因为 UniGene 中还有来自 GenBank 的 mRNA 数据,以它们的聚类结果为种子类,会丢弃连接两个类的 EST。不同数据库的大规模聚类相似性分析还未见报道,一般建议在实际使用时参考多个数据库的结果[3]。UniGene 通过 mRNA 来提高聚类的可靠性,但同源基因 mRNA 的错误聚类同样会造成表达序列的过度聚类,Frank 等用 PECT 对 Glycine max 的 mRNA 序列进行聚类,对照 UniGene 的结果和已知实验的分析表明了这一点[16]。同源基因的过度聚类仍然是基于相似性聚类方法难以解决的问题。

3.2 基于基因组定位的 EST 聚类

自从基因组计划开展以来,越来越多物种的基因组序列被测定。利用表达序列与基因组的比对,可以获得表达序列在基因组上的定位,而利用考虑真核基因剪接结构特征的比对工具,可进一步获得基因的外显子 – 内含子结构,以及基因转录加工的调控信息。相应地,剪接比对工具也得到快速发展。此方面的经典工具为 sim4[17]。BLAT[18] 从计算策略上有很大改进,在保证敏感性水平的同时,速度提高至少 500 倍,它除了单机版方式,还有服务器/客户端方式,可以快速处理大批量表达数据与整个基因组的比对[19]。我们曾在超级刀片计算机系统上,建立多个 BLAT 服务器/客户端节点(7 × 8),在 7.5 小时内完成了 500 多万条人类 UniGene 序列与基因组的比对,利用单机版则至少需要一个月[20]。

这些工具为利用标准基因组信息进行表达序列质量的评价提供了有力的支持。我们曾利用多种工具对常用的人类表达序列参考数据集 RefSeq 进行与基因组的剪接比对分析,结果表明其中 5% 的序列不能与基因组良好匹配,其中有多样性的原因,也有可能是基因转座嵌合的原因,但也不能排除是表达序列错误的原因[21]。而 Murray 等[22] 利用 BLAT 对 dbEST 中的全部人类序列进行基因组的比对分析发现,5.5% 的序列存在载体污染,其中污染可能嵌合在序列中间,4.4% 的序列在基因组上有多个可能的定位,与重复基因和相似基因相关,以基因组位置重叠关系进行的聚类,80% 的 RefSeq 可与唯一的类对

应,而只有25%的RefSeq与唯一的UniGene类对应,说明基于相似性聚类的结果中第一类错误也比较严重。

根据表达序列与基因组的比对信息,能有效筛查可能与同源基因、嵌合基因、多样性以及序列错误相关的序列,其聚类结果应另做特殊处理。简单根据基因组位置重叠进行聚类还可能过度聚类嵌套基因,因此要求聚类的序列两两之间至少要有一剪接位点位置相同。由于序列质量原因,这个位置约束可以有6bp左右的放宽。而没有剪接的序列可与该位置的聚类结果进行映射式的聚类,完全重叠的则可加入聚类,无法完全重叠的不能排除基因组序列污染的原因。

4 EST装配

从转录组角度而言,只有通过EST装配过程才能获得有意义的转录组序列,有的数据库通过EST装配进一步聚类,区分初始聚类中的不同剪接变体或是同源基因。不同的装配方法对这个问题有不同的解决效果。

4.1 以传统基因组序列装配方法进行的EST装配

基因组序列片段装配的主要软件有Phrap(http://bozeman.mbt.washington.edu/phrap.docs/phrap.html)、TIGR Assembler[23]、CAP3[24]等。STACK采用Phrap,因为Phrap可以利用序列测序的质量文件,但实际上很多EST序列提交时没有附带测序的质量文件。CAP3在产生一致序列时采用了多序列比对,因此产生的一致序列质量较高,被TIGR Gene Indices系统选用[25]。这些方法的基本思想是采用传统的"重叠-排列-生成共有序列"(Overlap-Layout-Consensus)策略。有重叠的序列才有可能进行装配。装配的重叠分析方法与初始聚类方法是相似的,但是装配算法还要考虑重叠关系的具体排列,而且这些排列关系需保存在内存中,因此"排列"成为算法的内存瓶颈,使得装配的适用规模受限,因此初始聚类还是必要的。EST聚类后装配的计算复杂度估计为 $c = k\sum_{i=2}^{m}(n_{clu})_i i^2$,其中 m 是最大的类包含的序列数,而 $(n_{clu})_i$ 表示序列数目为 i 的类的个数。EST数据仍在急剧增加,每一类的序列数势必会大量增加,因此利用传统的基因组装配程序进行EST的装配在效率上还需进一步优化。

4.2 基于图论的EST装配

传统基因组序列装配方法考虑的问题是将一组序列装配成唯一的一致序列,剩余无法装配的序列之间再进行装配或报告为问题序列。因此对于可能存在可变剪接的EST装配,传统方法对装配顺序敏感,有可能结果正确,也有可能产生截短的甚至错误的装配,如图1所示。基因组序列装配还有一种基于图论的方法[26],将序列装配问题转化为de Bruijin图中的欧拉路径问题,具有多项式的求解算法,同样的方法可以应用在EST的装配上。因为图的连接逻辑上与外显子-内含子结构对应,因此也称为剪

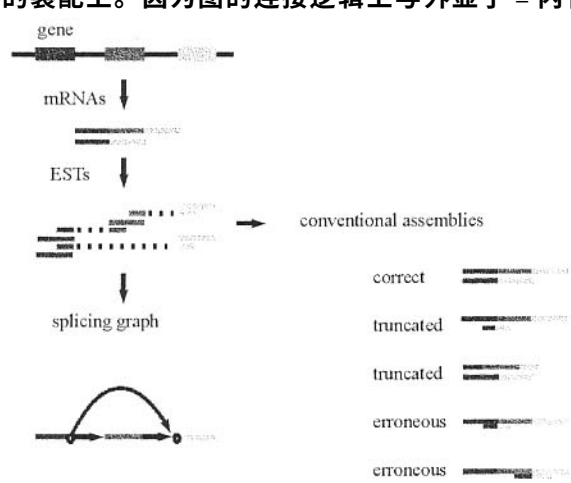

图1 可变剪接基因的EST装配问题

Fig.1 EST assembly problem of alternative splicing gene

接图(splicing graph)方法[27]。一个转录则对应剪接图中从源点(source,起始片段)到汇点(sink,终止片段)的所有路径,遍历所有的路径则产生所有可能的转录变体。

4.3 利用基因组信息的 EST 装配

直接利用表达序列构造剪接图的问题是序列质量和多样性问题,会使得 de Bruijin 图分析变得困难,常常需要通过多序列比对来消除序列差异。而借助剪接图的概念,利用表达序列与基因组的剪接比对结果可以获得外显子片段之间的连接关系,直接构造剪接图。以这种方法分析可变剪接的数据库如 ASG[28]、ECgene[13],而早期的根据 EST 信息进行基因预测的 TAP 方法其实也是类似的原理,被用于我国家蚕的可变剪接数据库中[29]。这些方法使用的剪接比对程序都是 sim4。Xing 等在 UniGene 聚类结果的基础上,根据偏序图多序列比对(POA)方法,同时考虑一组 EST 与基因组的比对情况,消除单个 EST 与基因组剪接比对容易产生的剪接边界和末端外显子的错误问题,根据比对结果构造剪接图,形成可变剪接及其产物的数据库 ASP[30]。

这些图论方法都具有很好的计算效率,POA 算法的计算复杂度与序列个数的关系是线性的,其它方法中剪接比对计算复杂度与序列长度的关系是超线性的,但小于二次方,而且表达序列长度都很短,另外剪接比对过程还具有并发性,根据剪接比对结果构造剪接图的计算复杂度与序列个数的关系是线性的,而且数据得到很好的压缩,对内存的要求不高。

5 总 结

在功能基因组时代,转录组研究成为揭示基因系统工作机制的重要手段,产生的数据也与日俱增。在没有基因组信息可利用的时候,基于相似性的聚类还是很有必要的,例如在对于基因组特别大的植物的转录组分析过程中,PaCE 对于 PlantGDB[31] 的构建起到了很大的作用。文中对这方面的方法发展做了介绍和评价。其所涉及的难以解决的同源家族问题、嵌套基因和嵌合基因问题,是否只对应小部分具有特殊性的基因,还是一种确实存在的普遍的生命现象,还有待生命机制的探索。而目前的结果已经为转录组的功能和机制研究提供了丰富的信息。

早期可变剪接分析是通过比较同一基因的表达序列之间是否有相对的插入或删除,但这只能提示可变剪接事件的发生,无法分析可变剪接如何改变基因的内含子 - 外显子结构,更无法分析可能相关的调节机制,而且容易与多样性差异混淆。在具有高质量的标准基因组数据后,利用表达序列与基因组的剪接比对结果,一方面可以对序列的质量和基因的复杂情况做一个筛查,区分嵌套基因,屏蔽可能有错误的序列,对存在同源家族的基因,可能发生嵌合的基因,以及具有多样性的基因做特殊考虑;另一方面可以获得基因的整体剪接结构,以及基因转录加工(包括剪接加工)的基因组环境的调控信息。

综上所述,大规模可变剪接分析需要基于基因组信息建立高效、可靠的分析策略。首先,收集尽可能多的 EST 和 mRNA 数据,对其进行载体去除和多嘧啶尾去除等必要的预处理。这一阶段应利用高性能计算的支持加快分析流程。第二阶段,利用 BLAT 或 BLAST 等工具进行基于基因组的定位聚类。利用高性能计算环境,实现 BLAT 和 BLAST 的并行计算可显著加速这一过程。第三阶段,利用剪接比对方法或 POA 方法对聚类序列和相应局部基因组序列进行剪接结构的分析,并利用共享某一剪接位点的传递关系进一步聚类。其中剪接比对过程可根据其并发性进行计算优化。第四阶段,根据剪接比对结果进行剪接图分析,揭示聚类的可变剪接,并产生一致序列。最后,分析初始聚类中不发生剪接的序列与该类各剪接图中外显子的对应关系,完全属于某一外显子的则可以被聚类,而无法归类的,即出现延伸外显子或连接多个外显子的情况,倾向于考虑是基因组污染,除非内含子保留事件有更多的支持证据。这样的策略既可作为大规模可变剪接分析的策略,也可作为全转录组的分析策略,中间的各个阶段采用的方法及参数可以根据实际的严格性要求进行选择,分析结果的表示及其可靠性评价也应根据转录组分析系统的应用目的进行设计。

参考文献：

[1] Modrek B，Lee C. A Genomic Vew of Alternative Splicing [J]. Nat Genet. 2002，30(1)：13 – 19.

[2] 丁克越，沈岩. ESTs数据分析及ESTs数据系统[J]. 国外医学分子生物学分册. 2002，24(2)：113 – 117.

[3] Bouck J，Yu W，Gibbs R，et al. Comparison of Gene Indexing Databases [J]. Trends Genet. 1999，15(4)：159 – 162.

[4] Zhang Z，Schwartz S，Wagner L，et al. A Greedy Algorithm for Aligning DNA Sequences [J]. J Comput Biol. 2000，7(1 – 2)：203 – 214.

[5] Darling A E，Carey L，Feng W C. The Design，Implementation，and Evaluation of mpiBLAST [R]. In：ClusterWorld Conference & Expo and the 4th International Conference on Linux Clusters：The HPC Revolution，2003，LA – UR 03：2862.

[6] Burke J，Davison D，Hide W. d2 _ cluster：a Validated Method for Clustering EST and Full-length cDNA Sequences [J]. Genome Res. 1999，9(11)：1135 – 1142.

[7] Carpenter J E，Christoffels A，Weinbach Y，et al. Assessment of the Parallelization Approach of d2 _ cluster for High-Performance Sequence Clustering [J]. J Comput Chem，2002，23(7)：1 ~ 3.

[8] Trivedi N，Bischof J，Davis S，et al. Parallel Creation of Non-redundant Gene Indices from Partial mRNA Transcripts [J]. Future Generation Computer Systems. 2002，18(6)：863 – 870.

[9] Ptitsyn A，Hide W. CLU：a New Algorithm for EST Clustering [J]. BMC Bioinformatics 2005，6(Suppl 2)：S3.

[10] Mudhireddy R，Ercal F，Frank R. Parallel Hash-based EST Clustering Algorithm for Gene Sequencing [J]. DNA Cell Biol 2004，23(10)：615 – 623.

[11] Malde K，Coward E，Jonassen I. Fast Sequence Clustering Using a Suffix Array Algorithm [J]. Bioinformatics 2003，19(10)：1221 – 1226.

[12] Kalyanaraman A，Aluru S，Kothari S，et al. Efficient Clustering of Large EST Data Sets on Parallel Computers [J]. Nucleic Acids Res，2003，31(11)：2963 – 2974.

[13] Kim N，Shin S，Lee S. ECgene：Genome-based EST Clustering and Gene Modeling for Alternative Splicing [J]. Genome Res. 2005，15(4)：566 – 576.

[14] Jurka J. Repbase Update：a Database and an Electronic Journal of Repetitive Elements [J]. Trends Genet，2000，16(9)：418 – 420.

[15] Bedell J A，Korf I，Gish W. MaskerAid：a Performance Enhancement to RepeatMasker [J]. Bioinformatics. 2000，16(11)：1040 – 1041.

[16] Frank R L，Ercal F. Evaluation of Glycine max mRNA Clusters [J]. BMC Bioinformatics 2005，6(Suppl 2)：S7.

[17] Florea L，Hartzell G，Zhang Z，et al. A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence [J]. Genome Res. 1998，8(9)：967 – 974.

[18] Kent W. Blat-the Blast-Like Alignment Tool [J]. Genome Research，2002，12(4)：656 – 664.

[19] 杭兴宜，赵东升，等. 序列比对程序BLAT在转录组数据分析中的应用[J]. 生物信息学，2005，3(2)：85 – 88.

[20] 赵东升，杭兴宜，李稚锋，等. 军事医学科学院生物医学超级计算中心的计算资源与应用[J]. 军事医学科学院院刊. 2005，29(4)：363 – 367.

[21] 李稚锋，李玉鉴，赵东升，等. 基于RefSeq数据库的人类标准转录数据集的构建[J]. 遗传，2006，28(3)：329 – 333.

[22] Murray C G，Larsson T P，Hill T，et al. Evaluation of EST-data Using the Genome Assembly [J]. Biochem Biophys Res Commun. 2005，331(4)：1566 – 1576.

[23] Sutton G，White O，Adams M，Kerlavage A. TIGR Assembler：a New Tool for Assembling Large Shotgun Sequencing Projects [J]. Genome Sci Technol. 1996，1(1)：9 – 19.

[24] Huang X，Madan A. CAP3：a DNA Sequence Assembly Program [J]. Genome Res. 1999，9(9)：868 – 877.

[25] Liang F，Holt I，Pertea G，et al. An Optimized Protocol for Analysis of EST Sequences [J]. Nucleic Acids Res. 2000，28(18)：3657 – 3665.

[26] Pevzner P A，Tang H，Waterman M S. An Eulerian Path Approach to DNA Fragment Assembly [J]. Proc Natl Acad Sci U S A. 2001，98(17)：9748 – 9753.

[27] Heber S，Alekseyev M，Sze S H，et al. Splicing Graphs and EST Assembly Problem [J]. Bioinformatics. 2002，18(Suppl1)：S181 – 188.

[28] Leipzig J，Pevzner P，Heber S. The Alternative Splicing Gallery (ASG)：Bridging the Gap Between Genome and Transcriptome [J]. Nucleic Acids Res. 2004，32(13)：3977 – 3983.

[29] Zha X F，Xia Q Y，Zhao P，et al. Detection and Analysis of Alternative Splicing in the Silkworm by Aligning Expressed Sequence Tags with the Genomic Sequence [J]. Insect Mol Biol. 2005，14(2)：113 – 119.

[30] Xing Y，Resch A，Lee C. The Multiassembly Problem：Reconstructing Multiple Transcript Isoforms from EST Fragment Mixtures [J]. Genome Res. 2004，14(3)：426 – 441.

[31] Dong Q，Schlueter S D，Brendel V. PlantGDB，Plant Genome Database and Analysis Tools [J]. Nucleic Acids Res. 2004，32(Database issue)：D354 – 359.