

基因序列的真实 EST 匹配识别与外显子区鉴定

周艳红* 井 辉* 李延恩 刘怀兰

(华中科技大学生物信息技术研究中心, 武汉 430074. * 同等贡献. E-mail: yhzhou@hust.edu.cn)

摘要 在对部分已知基因序列与 EST 数据库比对搜索的匹配结果进行深入分析的基础上, 总结出了匹配程度检验、Gap 检验、包含性检验、长度检验等多种措施对 EST 匹配结果的真假进行判别, 并研制出了相应的分析软件 EDSAcl.0, 用于从基因序列的众多 EST 匹配中尽可能地筛选出真正相关的匹配。在此基础上, 该软件可进一步鉴定出基因序列中的外显子区域。用标准测试数据集 HMR195 中的人类基因序列对其性能进行测试时, EDSAcl.0 鉴定出的蛋白编码区在核苷酸水平上的专一性 Sp 达到了 0.997, 敏感度 Sn 达到了 0.88, 优于对国际同类软件 TAP 的对比测试结果。EDSAcl.0 已提供网上服务 (<http://infosci.hust.edu.cn>)。

关键词 基因序列 EST 序列比对 真实匹配 外显子识别

EST是从随机选择的cDNA克隆中进行 5'或 3'端一次测序获得的cDNA序列片段, 长度一般在 400~600 bp左右, 代表了一个完整基因的一部分。从 1991 年EST的概念被提出以来^[1], EST序列数据的积累异常迅速^[2]。截至 2004 年 6 月 25 日, NCBI的dbEST数据库中已收录EST序列 22165266 条, 其中人类 (*Homo sapiens*) 的有 5643076 条 (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html)。

巨量积累的EST序列是一种宝贵的生物信息资源, 在寻找新基因^[1,3,4]、疾病相关基因^[5]、药物靶点^[6]以及识别选择性剪接^[7,8]与SNP位点^[9]等方面均具有重要应用价值, 而正确确定与基因序列真正相关的EST是有效开展这些应用研究的基础。然而, EST序列的“错误倾向”性^[10]以及基因组序列的复杂性^[11]等使得基因序列相关与无关的EST难以有效区分, 从而制约了巨量EST数据潜力的充分发挥。例如, 在基因注释方面, 利用EST帮助识别基因结构得到了广泛的研究^[12~14]。然而, 迄今的研究报道表明, 利用EST辅助识别基因的效果尚不理想, 主要表现在: () 对于将EST信息与从头预测(*ab initio*)方法等结合起来进行基因结构预测的软件, 加入EST后的预测效果并不理想, 往往表现为预测敏感度(Sn)有所提高, 但预测专一性(Sp)则降低很多^[14~18]; () 对于仅利用DNA序列与EST数据库的比对搜索结果进行基因结构预测的软件, 其典型代表为EbEST^[19], TAP^[20]以及ESTGenes^[21], 其中, EbEST同样存在预测专一性 Sp 低的问题^[19], 而TAP和ESTGenes通过对EST匹配进行较严格的筛选而获得了较高的预测专一性

Sp 。然而, 严格的筛选措施又容易导致预测敏感度 Sn 的下降^[20,21], 同样难以充分发挥EST数据应有的潜力。造成上述结果的主要因素有EST序列的“错误倾向”性、基因组序列的复杂性、以及高度冗余的EST序列可能代表的基因选择性剪接表达情况。在EST序列的“错误倾向”性方面, 除了EST一次测序带来的约 3% 的错误率^[3]外, 还有由于在构建cDNA文库过程中可能引入的种种污染而产生的非正常转录序列, 如载体序列、接头序列、嵌合基因序列、基因间隔区序列、pre-mRNA序列等^[10]。而基因组序列的复杂性主要是指由于基因组中广泛存在的转座成分、repeats^[11]以及基因家族之间的motifs等, 导致基因组序列间的局部相似。EST的“错误倾向”性和基因组序列间的局部相似性使得常用的序列比对技术难以将某基因序列相关与无关的EST有效区分开来, 即基因序列容易被匹配上一些非其产生的假阳性EST序列, 从而造成在推断外显子和基因结构时的预测专一性 Sp 的显著下降。此外, 由于基因组中可能广泛存在着选择性剪接, 而高度冗余的EST数据可能在较大程度上反映了这种选择性剪接情况, 从而使得利用EST数据识别外显子和基因结构这一问题变得更加复杂^[20]。显然, 如何利用EST数据有效鉴别基因的选择性剪接也首先有赖于能否将基因序列真正相关与无关的EST匹配有效区分开。因此, 为了更好地利用EST数据, 迫切需要解决如何有效识别与基因序列真正相关的EST数据。

本研究在对部分已知基因序列与EST数据库的Blastn比对搜索结果进行深入分析的基础上, 总结出了匹配程度检验、Gap检验、包含性检验、长度检验

等多种措施对EST匹配结果的真假进行判别,并研制出了相应的分析软件EDSAc1.0,用于从基因序列的众多EST匹配中尽可能地筛选出真正相关的匹配。在此基础上,该软件可进一步鉴定出基因序列中的外显子区域。用标准测试数据集HMR195^[22]中的人类基因序列对其性能进行测试时,EDSAc1.0鉴定出的蛋白编码区在核苷酸水平上的专一性Sp达到了0.997,敏感度Sn达到了0.88,优于对国际同类软件TAP的对比测试结果。

1 判别基因序列的真实 EST 匹配的措施

为分析基因序列与公共数据库中 EST 序列的匹配情况,以提炼用于判别基因序列的真实 EST 匹配的措施,本研究首先从 Kulp 和 Reese 所收集并被广泛用于分析人类基因结构特征的数据集中(http://www.fruitfly.org/seq_tools/datasets/Human/)随机挑选出 50 条人类基因序列,然后利用序列比对软件 Blastn 搜索 NCBI 的 dbEST 数据库。比对搜索的结果表明,数据库中的 EST 序列与这些已知基因序列的匹配情况可分为“真匹配”、“假匹配”、“多段匹配”、“单段匹配”、以及“冲突”等多种类型(图1)。其中,“真匹配”是指与某一 DNA 序列匹配的 EST 确实来源于该序列;否则,则为“假匹配”。“多段匹配”是指某一 EST 与 DNA 序列的匹配结果由多个分离的片段组成(当 EST 跨越基因内含子区域时将出现这种情况);否则,则称为“单段匹配”。“冲突”是指对于与某一基因序列匹配的多条 EST,其匹配结果显示的基因外显子-内含子结构不一致(选择性剪接或假匹配均可能导致“冲突”)。



图1 DNA 序列与 EST 序列的部分匹配情况

本研究在对上述比对搜索结果进行深入分析的基础上,提炼总结出了下述匹配程度检验、Gap 检验、包含性检验、长度检验等判别基因序列的真实 EST 匹配的措施。

() 匹配程度检验。EST 序列与 DNA 序列的匹配程度(identity)高低是判断一段 EST 序列是否为“真

匹配”的重要依据。理论上,“真匹配”EST 的匹配程度应该为 100%,但由于 EST 序列中平均存在约 3% 的测序错误^[3],因此,“真匹配”EST 的匹配程度容易降至 97% 左右。而且,由于 DNA 序列中存在的低复杂度区域(low-complexity sequence)以及数据库中存在的部分测序质量较低的 EST 序列等因素的影响,某些“真匹配”EST 的匹配程度可能更低。因此,若选取较高的匹配程度阈值(如 97%)作为判别“真匹配”EST 的标准,势必会丢失部分“真匹配”EST 而使其无法发挥作用。但选取太低的阈值又难以将“真匹配”与“假匹配”EST 有效区分开。为解决这一矛盾,本研究提出了一种基于缓冲机制的匹配程度检验方法。该方法通过设置主匹配程度阈值(比如 98%)和辅匹配程度阈值(比如 85%),将低于主匹配程度阈值但又高于辅匹配程度阈值的 EST 匹配先保留起来,待匹配程度高于主匹配程度阈值的 EST 片段“搭成”基因外显子区的基本骨架之后,再将这些保留的 EST 片段插入进来,并配合后续措施进一步分析这些保留的 EST 片段是否为“真匹配”。

() Gap 检验。对于“多段匹配”EST,各匹配片段在 EST 序列上的编号必须前后连续才可能是“真匹配”(图2(a))。因为,如果一条 EST 横跨了一个或多个

EST: 1 .. 104	DNA: 501 .. 604
EST: 100 .. 294	DNA: 1245 .. 1439
EST: 295 .. 497	DNA: 1950 .. 2152
EST: 494 .. 557	DNA: 2234 .. 2298
(a)	
EST: 235 .. 314	DNA: 1387 .. 1466
EST: 429 .. 473	DNA: 2426 .. 2470
EST: 506 .. 658	DNA: 2698 .. 2850
EST: 670 .. 736	DNA: 3211 .. 3277
EST: 824 .. 944	DNA: 3711 .. 3828
(b)	
EST: 424 .. 19	DNA: 3450 .. 3853
EST: 544 .. 422	DNA: 3186 .. 3308
EST: 621 .. 543	DNA: 2874 .. 2952
EST: 650 .. 621	DNA: 2708 .. 2737
EST: 112 .. 54	DNA: 2070 .. 2145
EST: 106 .. 62	DNA: 1180 .. 1237
(c)	

图2 多段匹配的几种情况

每幅子图代表一条 EST 与 DNA 序列的匹配结果,每一行代表一个匹配片段,“EST”表示该匹配片段在 EST 序列上的位置,“DNA”表示该片段匹配在 DNA 序列上的位置。(a) 各匹配片段在 EST 序列上的编号前后连续(考虑到 Blastn 并不能准确地界定出段与段之间的边界,允许相邻片段之间有几个碱基的重叠或间隔);(b) 匹配片段在 EST 序列上的编号前后不连续;(c) 某些参与匹配的 EST 片段恰好是其他匹配片段的子序列

内含子区域, 那么当其与来源 DNA 序列相匹配时, EST 的前一匹配片段与后一匹配片段应该是连续的, 且这些匹配片段显示的也就是一个一个的外显子区域. 若“多段匹配”EST 的前后匹配片段之间不连续, 则该 EST 很可能是由于序列的局部相似而导致的“假匹配”(图2(b)). 这种利用匹配片段间的连续性判别“真匹配”EST 的措施称为“Gap 检验”.

() 包含性检验. 上述假“多段匹配”情况一般是完完全全的假, 即参与匹配的 EST 并非来源于与之匹配的 DNA 序列, 这样的假“多段匹配”片段很容易用“Gap 检验”舍去. 但还有一些特殊的“多段匹配”则是半“真”半“假”, 即在 EST 的多个匹配片段中, 一些 EST 匹配片段的确来源于所匹配的局部 DNA 序列, 但另外一些 EST 匹配片段又不是来源于所匹配的局部 DNA 序列, 后者的匹配长度一般较前者短得多, 且往往恰好是前者的子序列(图2(c)). 这种现象很可能是由于序列的局部相似性引起的, 即参与匹配的 EST 确实来自于与之匹配的 DNA 序列, 但其部分碱基由于序列相似而匹配到了这段 DNA 的其他区域. 有时, 这样的“假匹配”的匹配程度还可能比较高, 用“匹配程度检验”无法将其舍去. 由于这些“假匹配”EST 片段往往是“真匹配”EST 片段的子序列, 本研究采用“包含性检验”方法对其进行处理, 即若 2 个 EST 匹配片段之间呈现包含关系, 则舍去短匹配片段而保留长匹配片段.

() 长度检验. 对已知基因序列与 EST 数据库的比对搜索结果的分析显示, “单段匹配”是“假匹配”的可能性要比“多段匹配”大. 这可能与基因组序列的特性有关, 如广泛存在的逆转录转座成份^[11]、局部相似序列等都可能造成假的“单段匹配”; 同时, EST 序列中的一些错误也可能导致假的“单段匹配”的发生, 如 pre-mRNA、基因间隔序列的掺入^[10]等等. 由于“Gap 检验”对“单段匹配”无能为力, 因此本研究对“单段匹配”情况设置了长度检验措施, 即“单段匹配”为“真匹配”的匹配长度应不低于 EST 测序总长度的某一比例.

2 EDSAc1.0 软件的处理流程

基于上述判别“真匹配”EST 的各种措施, 研制出了相应的分析软件 EDSAc1.0, 用于从基因序列的众多 EST 匹配中筛选出真正相关的匹配, 而且, 该软件还可在此基础上进一步鉴定出基因序列中的外显子区域. EDSAc1.0 软件的处理流程如图3所示.

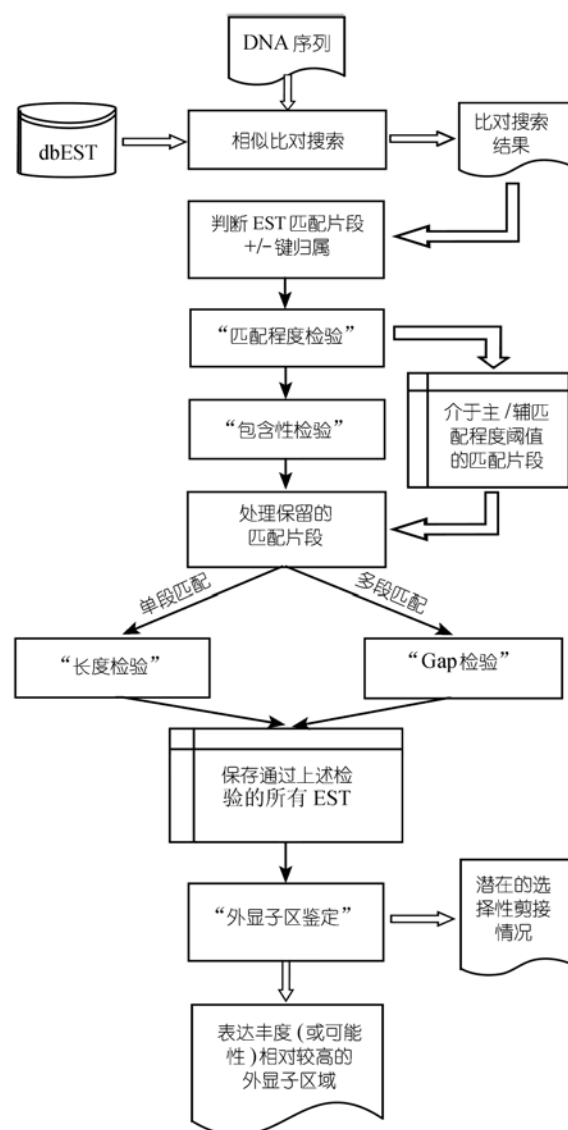


图3 EDSAc1.0 软件的处理流程

() 相似比对搜索. 采用核酸序列比对工具 Blastn 对所提交的 DNA 序列在 dbEST 数据库中进行相似比对搜索(Blastn 和 dbEST 下载到本地运行).

() 判断 EST 匹配片段的正负链归属. 利用 EST 的测序端(5'或 3')以及 Blastn 输出匹配结果的“Strand”项, 可以判断该 EST 匹配片段是匹配在 DNA 的正链还是负链, 方法如下: 对于 5'EST, 若“Strand = Plus/Plus”, 则匹配在 DNA 的正链, 若“Strand = Plus/Minus”, 则匹配在 DNA 的负链; 对于 3'EST, 若“Strand = Plus/Plus”, 则匹配在 DNA 的负链, 若“Strand = Plus/Minus”, 则匹配在 DNA 的正链.

() 匹配程度检验. EDSAc1.0 中, 主匹配程

度阈值的默认值为 98%, 而辅匹配程度阈值为 85%。之所以主匹配程度阈值取得比较高, 是为了保证较高的专一性 S_p , 而辅匹配程度阈值取得比较低, 则是为了尽可能获得比较高的敏感度 S_n 。对于高于主匹配程度阈值的 EST 片段, 进一步进行下一步的包含性检验与外显子区骨架搭建; 对于低于主匹配程度阈值但高于辅匹配程度阈值的 EST 匹配片段, 则先保留; 而对于低于辅匹配程度阈值的 EST 匹配片段, 则舍去。

() 包含性检验与外显子区骨架搭建。对于高于主匹配程度阈值的 EST 片段, 由“包含性检验”进一步剔除一些由序列相似性等引起的短的“假匹配”片段后, 搭建出外显子区的基本骨架轮廓。

() 处理保留的匹配片段。对于上述匹配程度介于主、辅匹配程度阈值之间的保留匹配片段, 将其插入到上一步搭建成的外显子区骨架之中, 并舍去在插入时与已有外显子区骨架结构矛盾的 EST 片段, 从而兼顾敏感度 S_n 与专一性 S_p 。

至此, EDSAc1.0 完成了对一条 EST 所有匹配片段的筛拣工作, 形成了一条由 EST 匹配片段顺序连接构成的“链”。对这条“链”还要进一步进行下述检验以判别其是否为“真匹配”。

() 长度检验或 Gap 检验。对于上述由 EST 片段构成的“链”, 若“链”是单段的, 则进一步进行“长度检验”。理论上, 真的“单段匹配”的匹配长度应该等于相应 EST 测序的总长度, 但考虑到 EST 序列的两端测序质量相对偏低, 比对过程中容易被舍去, 因此, 实际中这一比例要低于 1。本研究中取经验值 0.8 为长度检验阈值; 若“链”是多段的, 则进行 Gap 检验。在进行 Gap 检验时, 有多种“Gap”情况需要分别考虑, 总体上可以分为等长 Gap 和不等长 Gap 两类。顾名思义, 等长 Gap 就是 EST 上的缺口与 DNA 上的缺口大致等长。一般来说, 这可能是由于 Blastn 对低复杂度序列的过滤引起的, 对于这样的 Gap, 可以连同其两端的片段合成为更长的一段。而不等长 Gap 则是 EST 上的缺口长度与 DNA 上的缺口长度相差很大, 这又可以分为两类, 第 1 类是 EST 上的缺口比 DNA 上的短; 第 2 类则正好相反, 是 DNA 上的缺口比 EST 上的短。前一类情况比较普遍, 很可能意味着这个匹配是假的, 但也可能是低复杂度序列出现在某个外显子的边缘区域, 或者是低复杂度序列占据了一个较短外显子的整个区域, 导致 Blastn 将其过滤掉了; 而

后一类情况则较少见, 可能是由于 DNA 序列的多态性或者 cDNA 文库的污染造成的。在 EDSAc1.0 中, 这两种不等长的情况都被舍去。

至此, EDSAc1.0 完成了对一条 EST 是否为“真匹配”的判别。重复上述过程对所有经 Blastn 比对搜索得到的 EST 匹配进行处理后, 输出所提交 DNA 序列的所有“真匹配”EST, 并进一步进行下面的外显子区域鉴定。

() 外显子区域鉴定。由于 EST 只是 cDNA 序列的一部分, 所以, 需要将相互重叠的多条“真匹配”EST 拼接起来并鉴定出基因序列的外显子区域。然而, 由于选择性剪接的广泛存在以及来源于 genomic DNA 或 pre-mRNA 的错误 EST 和部分未能去除的“假匹配”EST 等的影响, 经上述处理过程得到的各“真匹配”EST 所显示的基因外显子-内含子结构可能在某些区域发生“冲突”(图1), 如何处理这种“冲突”是外显子区域鉴定过程中的关键问题。EDSAc1.0 中采用了一种“分段-赋分”的机制, 可在“冲突”情况下鉴定出表达丰度(或可能性)相对较高的外显子区域, 而将表达丰度(或可能性)相对较低的情况作为潜在的选择性剪接区域给出。

基于“分段-赋分”机制的外显子区域鉴定过程包括: (1) 取各“真匹配”EST 在 DNA 序列上的覆盖区域(包括匹配区域和所跨越的内含子区域)的并集, 得到 DNA 序列上的所有 EST 覆盖区域; (2) 根据全部 EST 匹配显示的所有外显子-内含子边界, 将 DNA 上的 EST 覆盖区域划分成一系列小区段; (3) 根据每条 EST 的匹配情况对各小区段赋分, 即对于任一 EST, 在其覆盖范围内, 若与某一小区段匹配, 则增加该小区段的 match 分值, 否则, 增加该小区段的 mismatch 分值(各小区段的 match 与 mismatch 的初值均为零; 对于“多段匹配”EST, match 与 mismatch 的增加值设为 1, 而“单段匹配”情况下则经验性地取为 0.4。这是由于受局部相似性、逆转录转座成份以及序列污染等的影响, “单段匹配”的可信度相对而言没有“多段匹配”的高); (4) 根据各小区段的 match 和 mismatch 分值来推断其类型: 若 $match > 0$ 且 $mismatch = 0$, 则该小区段为外显子区; 若 $match = 0$ 且 $mismatch > 0$, 则该小区段为内含子区; 若 $match > mismatch > 0$, 则该小区段为“冲突”区域, 且其作为外显子区的表达丰度(或可能性)相对较高; 若 $mismatch > match > 0$, 则该小区段也为“冲突”区域, 且其作为内含子区的

表达丰度(或可能性)相对较高. 这种处理方法是基于基因的表达式度高低与对应存在的EST数量正相关的思想, 虽然均一化(normalization)和消减扣除(subtraction)等实验技术^[23]使得EST的数量与基因的表达丰度不再成比例, 但对EST的简单计数还是能够用来粗略地估计一个基因的不同剪接变体表达量的相对多少^[20,24,25].

需要说明的是, 由于 Blastn 在比对时并不能非常准确地界定出外显子的边界, 因此, EDSAc1.0 中利用外显子边界的 GT-AG 双核苷酸规则对外显子边界进行了校正. 此外, EDSAc1.0 尚未对基因翻译起止位点进行确定, 所得到的基因外显子区域既包括编码(CDS)区, 也包含两侧的 UTR 区.

3 EDSAc1.0 软件的测试结果及其分析

为评价EDSAc1.0 软件识别“真匹配”EST和鉴定基因外显子区域的效果, 本研究采用HMR195数据集^[22](<http://www.cs.ubc.ca/~rogic/evaluation/dataset.html>)对其进行测试. 该数据集由Rogic等人^[22]于2001年仔细挑选出的195条基因序列(其中103条属于人类, 92条属于鼠类)组成, 用于对基因结构预测系统的性能进行测试与评价. 该数据集在构建时已尽可能地排除了存在选择性剪接的基因序列, 因此, 采用该数据集可尽量避开目前还难以有效测试与评价的选择性剪接问题.

用HMR195 数据集中的人类基因序列作为测试数据, 用NCBI的dbEST库中的人类EST(release 022004, 共 5472005 条人类EST)作为EST数据源, 采用评价基因结构预测性能的标准指标^[20,26,27](包括核苷酸水平上的敏感度Sn和专一性Sp, 外显子水平上的敏感度Sn, 专一性Sp, 完全丢失外显子比率ME和完全预测错误外显子比率WE)对EDSAc1.0 软件进行测试, 其结果如表1所示. 值得指出的是, 由于该测试数据集中没有UTR区外显子的注释信息而只有CDS区的注释信息, 故表1中给出的是CDS区的预测效果. 此外, 为便于比较分析, 表1中同时给出了对

国际同类软件中预测精度较高的TAP进行对比测试的结果以及直接用Blastn得到的判别结果(即对Blastn的输出结果仅仅进行正负链归属判断而不经EDSAc1.0 处理). 其中, 对TAP的测试结果是通过将测试数据提交到其网络服务站点(<http://sapiens.wustl.edu/~zkan/TAP/>)得到的. 对于上述测试数据, EDSAc1.0 和Blastn均可对其所有 103 条人类基因序列进行分析, 但TAP只能得到其中 101 条的分析结果(AF082802 和U96846 因故无法预测). 为便于比较, 表1中给出的均是对这101条基因序列进行测试的结果(其中Sn, Sp, ME和WE的计算是按照文献^[26]中提到的“by gene”方式进行的, 也可参见文献^[20]. 为了衡量各个程序对不同基因序列预测时的精度稳定性, 表1中还给出了Sn及Sp相应的标准差). 对于Blastn, 因其在比对时并不考虑外显子边界的具体特征而难以准确确定外显子边界, 故在外显子水平上对其进行评价的意义不大.

由表1可知, 对上述101条基因序列采用 Blastn 对 dbEST 数据库进行序列比对搜索, 共得到 362976 个 EST 匹配片段(“-b”参数取 2500). 当根据这些 EST 匹配片段识别基因序列中的编码区域时, 核苷酸水平上的敏感度 Sn 可达到 0.926, 但专一性 Sp 只有 0.579. 比较高的预测敏感度在一定程度上表明, 目前积累的 EST 数据已覆盖了人类基因的绝大部分编码区域, 因而其有效利用非常重要. 另一方面, 如果上述 101 条基因序列的选择性剪接情况确实很少(因 HMR195 在构建时已尽可能地排除了存在选择性剪接的基因序列), 那么, 不足 0.6 的预测专一性则表明, 上述众多 EST 匹配片段中存在大量的假阳性匹配, 也就是说, 常用的序列比对技术难以将基因序列相关与无关的 EST 有效区分开来.

当经过 EDSAc1.0 的分析处理后, 共得到 61238 个“真匹配”EST 片段, 其数量仅为相同参数下 Blastn 所得结果的 1/6. 当根据这些“真匹配”EST 片段识别基因序列中表达丰度(或可能性)相对较高的编码区

表 1 对 HMR195 数据集中的 101 条人类基因序列进行测试的结果

软件名称	认定为真匹配的 EST 片段总数	核苷酸水平精度		外显子水平精度			
		Sn	Sp	Sn	Sp	ME	WE
EDSAc1.0	61238	0.880 ± 0.268	0.997 ± 0.016	0.779 ± 0.338	0.858 ± 0.265	0.096	0.003
TAP	—	0.795 ± 0.348	0.984 ± 0.075	0.714 ± 0.402	0.825 ± 0.316	0.163	0.005
Blastn (“-b”取 2500)	362976	0.926 ± 0.169	0.579 ± 0.323	—	—	—	—

域时,在得到预测敏感度 S_n 为 0.88 的同时,预测专一性 S_p 高达 0.997.

相对于直接使用 Blastn 时的结果,EDSAc1.0 的预测敏感度有一定的下降(从 0.926 下降到 0.88). 导致这一结果的原因可能来自 3 方面: () EDSAac1.0 在排除假阳性 EST 匹配的同时,可能丢失了一些“真匹配”EST,这是 EDSAac1.0 需要进一步研究解决的一个问题;() 上述 101 条基因序列中仍然存在选择性剪接情况,由于 EDSAac1.0 对其鉴定的结果与注释结果不一致而导致预测敏感度“下降”(后面的分析结果确实证实了这种情况的存在); () 由于 Blastn 比对搜索结果中存在大量的假阳性 EST 匹配,其中有些假阳性 EST 可能恰好匹配在基因序列的外显子区,从而导致预测敏感度的“非正常”提高,也就是说,与所测试的基因序列真正相关的 EST 所覆盖的基因编码区域实际上不到 92.6%,因而 EDSAac1.0 丢失“真匹配”EST 的实际程度比表1所示结果要低.

从表1还可以看出,对于基因外显子区域的鉴定,EDSAc1.0 的各项测试指标均优于 TAP. 在预测敏感度 S_n 方面,无论在核苷酸水平还是外显子水平,EDSAc1.0 相对 TAP 均有较显著提高,分别为 0.88 : 0.795 和 0.779 : 0.714; 而在预测专一性 S_p 方面,尽管 TAP 已达到了较高的精度,但 EDSAac1.0 在核苷酸水平和外显子水平上相对 TAP 仍有一定的提高,分别为 0.997 : 0.984 和 0.858 : 0.825; 在完全丢失外显子比率 ME 和完全预测错误外显子比率 WE 方面,EDSAc1.0 的性能相对 TAP 也有一定的改善,分别为 0.096 : 0.163 和 0.003 : 0.005. 此外,从 S_n 和 S_p 的标准差可以看出 EDSAac1.0 对不同基因序列预测时的精度稳定性也要好于 TAP.

在运行效率方面,EDSAc1.0 主要取决于 Blastn 对 EST 数据库进行比对的快慢,相比而言,其后处理过程只占用很少的时间. 在 PC 机上运行时(CPU: P4 1.8 GHz, 内存: 256 M),EDSAc1.0 分析一条 8000 bp 的人类 DNA 序列约耗时 5 min(其中 99% 以上为 Blastn 的运行时间). 目前 EDSAac1.0 正在向高性能服务器平台移植,届时其处理速度将可加快 30 倍以上.

以上测试结果及其分析表明,EDSAc1.0 能有效排除与基因序列无关的假阳性 EST 匹配并鉴定出基因序列中的外显子区域. 然而,EDSAc1.0 仍然存在以下需要进一步研究解决的问题.

首先是潜在选择性剪接情况的进一步研究. 对于 HMR195 中的 103 条人类基因序列,尽管 EDSAac1.0 鉴定出的表达丰度(或可能性)较大的基因外显子区域与注释结果吻合得很好,但 EDSAac1.0 也同时输出了为数不少的潜在选择性剪接区域(“冲突”区域). 在所有 103 条基因序列中,共有 75 条基因的 590 个局部区域被判定为潜在选择性剪接区域(未考虑 UTR),这似乎与 HMR195 在构建时已尽可能排除了存在选择性剪接的基因序列存在一定的矛盾. 为了搞清楚这么多的“冲突”到底意味着选择性剪接的广泛存在还是由 EDSAac1.0 未能排除的假阳性 EST 匹配所引起,我们进行了以下的初步分析. 在上述 590 个“冲突”区域中,仅有 6 个区域的 EDSAac1.0 鉴定结果与注释结果不一致. 对于这 6 个“冲突”区域中的 3 个,我们已在 GenBank 找到了相应的 mRNA 支持 EDSAac1.0 的鉴定结果(这 3 个区域分别是 AF049259 的 4312 ~ 4337, AF068624 的 17468 ~ 17578 和 AF071596 的 596 ~ 706). 这充分说明,尽管 HMR195 在构建时试图排除存在选择性剪接的基因序列,但实际上 HMR195 中的基因序列仍然存在甚至可能广泛存在选择性剪接情况(包括不正常的剪接变体). 相关研究已表明,人类基因中存在的选择性剪接情况相当普遍^[7,8,11],而且不少错误的剪接会引发人类疾病包括癌症^[28,29],由于 dbEST 中有超过 50% 的 EST 来自于癌症或肿瘤组织及细胞系^[30],因而选择性剪接变体在 dbEST 中应该还是比较普遍的. 因此,可以推测,EDSAc1.0 鉴定出的“冲突”区域很可能有相当部分是由选择性剪接(包括不正常的剪接情况)引起的,但由于问题的复杂性,更深入的分析还有待进一步的研究.

EDSAc1.0 有待进一步研究解决的另 1 个问题是有效确定基因边界并区分基因编码区外显子与 UTR 区外显子. 为解决这一问题,需要进一步研究 UTR 区与编码区的特征差异、以及综合考虑基因转录与翻译起止位点信号等其他信息.

另外,需要研究如何有效利用相关物种的 EST 数据以进一步改善 EDSAac1.0 的性能. 从表1中 Blastn 的测试结果可推知,对于 dbEST 中数据量最多的物种——人类,其 EST 序列也只覆盖了基因序列的约 90% 左右. 虽然 EST 测序技术的进步和 EST 数据量的继续增大将会使覆盖率进一步提高,但利用其他物种尤其是近缘关系物种已有的 EST 数据可能是目前

弥补 EST 覆盖率不足的一个途径. 为此, 本研究以 HMR195 中的 92 条鼠类基因为测试数据, 对利用 dbEST 库中的人类 EST 数据识别鼠类基因的效果进行了初步测试, 共识别出其中的 11 条基因序列(数据未列出), 在核苷酸水平上的敏感度 S_n 达到了 0.586, 而专一性 S_p 则高达 0.998. 这在一定程度说明, 利用相关物种 EST 对于提高 EST 覆盖率、从而进一步改善 EDSAc1.0 的性能是可行的.

4 结论

由于 EST 的“错误倾向”性以及基因组序列的复杂性等, 常用的序列比对技术难以将基因序列相关与无关的 EST 匹配有效区分开来, 导致巨量 EST 数据的潜力难以充分发挥. 本研究提出的判别“真匹配”EST 的措施和研制出的 EDSAc1.0 软件, 能有效识别与基因序列相关的 EST 匹配序列, 可为利用 EST 数据进行基因注释以及开展其他方面的应用研究(包括寻找疾病相关基因、识别选择性剪接与 SNP 位点等)提供帮助.

本研究的进一步工作主要包括: 真假 EST 特征的进一步分析; 基因选择性剪接的有效识别; 基因边界的有效确定; 基因编码区与 UTR 区的有效区分以及相关物种 EST 的有效利用.

致谢 本工作为国家自然科学基金(批准号: 90203011, 30370354)和湖北省自然科学基金(批准号: 2002AC014)资助项目.

参 考 文 献

- Adams M D, Kelley J M, Gocayne J D, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 1991, 252(5013): 1651~1656
- Benson D A, Karsch-Mizrachi I, Lipman D J, et al. GenBank: Update. *Nucleic Acids Res*, 2004, 32(Database issue): D23~D26
- Hillier L D, Lennon G, Becker M, et al. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res*, 1996, 6(9): 807~828
- Boguski M S, Tolstoshev C M, Bassett D E Jr. Gene discovery in dbEST. *Science*, 1994, 265(5181): 1993~1994
- Vitt U, Gietzen D, Stevens K, et al. Identification of candidate disease genes by EST alignments, synteny, and expression and verification of Ensembl genes on rat chromosome 1q43-54. *Genome Res*, 2004, 14(4): 640~650[DOI]
- Lizotte-Waniewski M, Tawe W, Guiliano D B, et al. Identification of potential vaccine and drug target candidates by expressed sequence tag analysis and immunoscreening of *Onchocerca volvulus* larval cDNA libraries. *Infect Immun*, 2000, 68(6): 3491~3501[DOI]
- Brett D, Hanke J, Lehmann G, et al. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett*, 2000, 474(1): 83~86[DOI]
- Mironov A A, Fickett J W, Gelfand M S. Frequent alternative splicing of human genes. *Genome Res*, 1999, 9(12): 1288~1293[DOI]
- Buetow K H, Edmonson M N, Cassidy A B. Reliable identification of large numbers of candidate SNPs from public EST data. *Nat Genet*, 1999, 21(3): 323~325[DOI]
- Sorek R, Safer H M. A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res*, 2003, 31(3): 1067~1074[DOI]
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 2001, 409(6822): 860~921
- Zhang M Q. Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet*, 2002, 3(9): 698~709[DOI]
- Mathe C, Sagot M F, Schiex T, et al. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*, 2002, 30(19): 4103~4117[DOI]
- Reese M G, Hartzel G, Harris N L, et al. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res*, 2000, 10(4): 483~501[DOI]
- Salamov A A, Solovyev V V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res*, 2000, 10(4): 516~522[DOI]
- Ashburner M. A biologist's view of the *Drosophila* genome annotation assessment project. *Genome Res*, 2000, 10(4): 391~393[DOI]
- Krogh A. Using database matches with HMMGene for automated gene detection in *Drosophila*. *Genome Res*, 2000, 10(4): 523~528[DOI]
- Reese M G, Kulp D, Tammanna H, et al. Genie--gene finding in *Drosophila melanogaster*. *Genome Res*, 2000, 10(4): 529~538[DOI]
- Jiang J, Jacob H J. EbEST: An automated tool using expressed sequence tags to delineate gene structure. *Genome Res*, 1998, 8(3): 268~275
- Kan Z, Rouchka E C, Gish W R, et al. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res*, 2001, 11(5): 889~900[DOI]
- Eyras E, Caccamo M, Curwen V, et al. ESTGenes: Alternative splicing from ESTs in Ensembl. *Genome Res*, 2004, 14(5): 976~987[DOI]
- Rogic S, Mackworth A K, Ouellette F B. Evaluation of gene-finding programs on mammalian sequences. *Genome Res*, 2001, 11(5): 817~832[DOI]
- Bonaldo M F, Lennon G, Soares M B. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res*, 1996, 6(9): 791~806
- Rudd S. Expressed sequence tags: Alternative or complement to whole genome sequences? *Trends Plant Sci*, 2003, 8(7): 321~329[DOI]
- Ewing R M, Ben Kahla A, Poirot O, et al. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res*, 1999, 9(10): 950~959[DOI]
- Burset M, Guigo R. Evaluation of gene structure prediction programs. *Genomics*, 1996, 34(3): 353~367[DOI]
- 周艳红, 杨雷, 王卉 等. 基于多级优化的真核生物基因结构预测. *科学通报*, 2004, 49(2): 140~145
- Hui L, Zhang X, Wu X, et al. Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogene*, 2004, 23(17): 3013~3023[DOI]
- Xu Q, Lee C. Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res*, 2003, 31(19): 5635~5643[DOI]
- Baranova A V, Lobashev A V, Ivanov D V, et al. In silico screening for tumour-specific expressed sequences in human genome. *FEBS Lett*, 2001, 508(1): 143~148[DOI]

(2004-07-16 收稿, 2004-10-29 收修改稿)