基于 RefSeq 数据库的人类标准转录数据集的构建

李稚锋[1,2],李玉鉴[3],赵东升[4],杭兴宜[1],王正志[2],骆志刚[5],张成岗[1]

(1.军事医学科学院放射与辐射医学研究所,北京 100850;
2. 国防科技大学机电工程与自动化学院,长沙 410073;3. 北京工业大学计算机学院,北京 100822;
4. 军事医学科学院卫生勤务与医学情报研究所,北京 100850;5. 国防科技大学并行与分布处理国防科技重点实验室,长沙 410073)

摘 要:美国国家生物信息技术中心(NCBI)提供了具有生物意义上的非冗余的基因和蛋白质序列的 RefSeq 参考序列数据库。然而,由于基因普遍存在的多态性以及不同实验室对于序列测定的质量控制存在差异等原因,已发现 RefSeq 数据库可能存在部分质量问题。文章基于"中心法则"提出"标准转录数据集"的概念,以人类基因和基因组序列为例,利用 BLAT、Sim4 和自行设计的 EIparser 等基因结构解析程序分析了 RefSeq 人类基因转录数据(2005-4-18)与目前所公布的人类标准基因组(2005-4-20)的对应关系。对于有实验证据支持的标记为 NM_和 NR_的记录,多种程序分析结果表明,其与标准基因组完全相对应的记录为 9 771 个;符合多个程序修订标准的记录有 10 943 个;而与标准基因组有较大差异的记录为 203 个,多种程序分析结果不一致的记录为 2 676 个,提示研究人员在使用此非标准转录组数据时,必须考虑到其存在非标准转录的原因甚至存在错误的可能性。此文为基于标准、高质量转录数据集的生物信息学数据分析、分子生物学实验设计、基因多样性和遗传变异分析等提供了重要的参考标准。相关结果可通过 http://biocompute.bmi.ac.cn/transcriptome/index.htm 访问。

关键词:RefSeq 数据库;转录组;质量控制;人类标准转录数据集

中图分类号:Q754 **文献标识码**:A **文章编号**:0253－9772(2006)03－0329－05

Construction of Standard Human Transcript Dataset Based on RefSeq and Human Genome Sequence Database

LI Zhi-Feng[1,2], LI Yu-Jian[3], ZHAO Dong-Sheng[4], HANG Xing-Yi[1],
WANG Zheng-Zhi[2], LUO Zhi-Gang[5], ZHANG Cheng-Gang[1]

(1. *Beijing Institute of Radiation Medicine, Beijing* 100850, *China*; 2. *College of Mechatronics Engineering and Automation, National University. of Defense Technology, Changsha* 410073, *China*; 3. *College of Computer Science & Technology, Beijing University of Technology, Beijing* 100822 *China*; 4. *Beijing Institute of Health Administration and Medical Information, Beijing* 100850, *China*; 5. *National Lab of Parallel and Distributed Processing, National University of Defense Technology, Changsha* 410073, *China*)

Abstract: The NCBI Reference Sequence (RefSeq) database aimed to provide a biologically non-redundant collection

收稿日期:2005－03－16;**修回日期**:2005－07－12

基金项目:国家重点基础研究发展计划(973 计划)(编号:2003CB715900)、国家高技术研究发展计划(863 计划)(编号:2002AA234021)、并行与分布处理国防科技重点实验室基金(编号:51484050304JB4401)、中国教育网格(ChinaGrid)生物信息学网格项目联合资助[Supported by National Basic Research Project (973 Program) (No. 2003CB715900); National High Technology Research and Development Program of China (863 Program) (No. 2002AA234021); Research Fund from National Lab of Parallel and Distributed Processing (No. 51484050304JB4401); and Bioinformatics Grid Project of China Education Grid(ChinaGrid)]

作者简介:李稚锋(1977—),女,江西奉新人,博士生,研究方向:生物信息学,E-mail:lizf@bmi.ac.cn

通讯作者:张成岗(1970—),男,陕西白水人,博士,研究员,研究方向:分子生物学、生物信息学、神经生物学。Tel:010-66931590;Fax:010-68169574;E-mail:zhangcg@bmi.ac.cn

致 谢:本文数据的计算在军事医学科学院生物医学超级计算中心的万亿次星盈超级刀片计算机上完成,网络数据服务由军事医学科学院网络信息中心提供,王小磊制作维护,特此致谢。

of DNA, RNA, and protein sequences and to promote the research on genes and proteins of human beings and other species. However, because of widely distributed polymorphisms and different quality control of experiments in individual laboratories, there are potential problems need to be identified in the RefSeq database. Regarding which, we herein define the concept, standard transcript, based on the Central Dogmas of Biology that each standard transcript should be perfectly mapped to the standard genomic DNA sequence at the exon level. A large scale analysis for mapping all of the RefSeq records of human being (2005-4-18) to the officially released human genome sequence database (2005-4-20) was further performed using BLAT, Sim4 and a homemade program, Elparser, which was especially designed for this purpose. The standard transcripts based on the RefSeq database were obtained according to the alignment with standard human genome database. There are 9 771 RefSeq records of human being labeled with "NM_" and "NR_" could be perfectly mapped to human genome sequences, while other 10 943 records could be considered as standard transcripts after reasonable revision by comparing with the genome sequences according to all of the three methods. Moreover, the left 203 unrevisable records and 2 676 inconsistent records reported by the above programs could not be considered as standard transcripts and should be checked critically before using because of potential errors in them. Our study has thus provided a reference standard dataset of human beings with high quality for further bioinformatic and experimental analysis such as polymorphism and mutation of human genes. The reference standard dataset based on above criteria could be retrieved from http://biocompute.bmi.ac.cn/transcriptome/index.htm.

Key words: RefSeq database; transcriptome; quality control; database of standard transcript sequences of human

人类基因组计划的完成意味着人类第一次对自身在基因组水平有了全面的认识,但是人类基因转录组的研究却并未同步完成,这是因为基因组相对是稳定的,而转录组则是动态变化的。为此,美国国家生物信息技术中心(NCBI)所发行的 RefSeq 数据库综合了 GenBank 数据库中基因原始数据的信息,意图准确链接染色体定位、转录形式和蛋白质产物等信息,将多物种的大量数据整合到包括序列、遗传、表达、功能信息的单一、一致的框架体系中[1],以期对生物信息学理论分析与生物医学实验研究提供重要参考。但在具体应用时发现,RefSeq 并不能代表标准的转录组数据集,而是存在一定的缺陷,如张德礼等曾报道的错误类型[2],包括单个碱基或片段的缺失、插入等,对于直接利用该数据集进行实验设计、基因预测、基因表达调控元件的挖掘、剪接位点分析等均会带来一定误差,但是并未见到 NCBI 发布能够和基因组数据完全一致的 RefSeq 数据集。当前转录组研究与基因可变剪接的研究迫切呼唤能够代表标准基因组序列的基因转录数据集,因此从"中心法则"出发,我们提出并定义"标准转录数据集(standard transcription dataset)"的概念,是指和标准基因组序列相一致、能够代表其转录信息的基因转录数据的集合。由于标准的转录数据集对于理论分析和实验研究均十分重要,因此,本文拟以人类 RefSeq 数据库为出发点进行详细分析,在此基础上提供基于人类 RefSeq 数据库的标准转录数据集。

1 材料和方法

1.1 数据来源

本研究所采用的基因组数据来源于 NCBI 2005 年 4 月 20 日发布的测序完成并拼接好的人类基因组数据 (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/human_genomic.gz),其中包括 24 条标记为 NC_的完整基因组序列。转录数据采用 NCBI RefSeq 中 2005 年 4 月 18 日发布的人类基因转录数据,(ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/human.rna.fna.gz),记录的数量为 29 176 个,其中有实验数据支持的 NM_和 NR_记录个数分别为 23 389 和 204。

1.2 策略与流程

通常用来进行 cDNA 与基因组序列比对以预测基因结构的程序是 Sim4[3] 和 BLAT[4]。但由于现有程序难以满足方便读取表达序列相比于标准基因组的差异以及判定剪接边界是否合理的要求,因此我们根据剪接比对的原理设计了新的剪接比对程序 Elparser(可通过 E-mail 索取),可详细地显示全部比对信息,同时提供在内含子/外显子边界处延伸了基因组的序列情况,可了解剪接位点处的匹配情况,尽量满足 GT-AG 等常见剪接规则,但不强制边界类型以避免破坏最优匹配,也可以动态调节字长,但不

允许小于 5 bp 的匹配且严格要求短外显子的剪接边界,同时可产生根据比对情况获得修订序列,报告不匹配的情况,并对不匹配、缺失或插入发生在剪接边界的 5 bp 以内的情况进行标记。由于大多数 RefSeq 记录来源于实验,因此先用 EMBOSS 软件包中的 trimest 程序去除所有 RefSeq 记录的 3′末端多腺苷序列[5]。为了加速计算过程,首先利用 BLAT 的服务器/客户端方式快速比对获得基因组定位信息,输出选用 psl 格式[4,6],并选取匹配分值最高的比对去除 BLAT 获得的很多短的匹配。往定位边界两端分别延伸 10 kb 获得定位的基因组序列,再利用 BLAT 命令行方式、Sim4 和 Elparser 分别进行分析。根据 Elparser 的比对结果进行分类,同时考察与 Sim4、BLAT 比对结果的差别。对于与基因组不完全匹配的编码基因记录,根据差异报告序列和 RefSeq 标注的 CDS 位置,考察 RefSeq 与标准基因组的差异是否会影响 RefSeq 序列关联的蛋白产物。同时,由于 RefSeq 数据量庞大(将近 3 万个记录),因此我们采用万亿次超级计算机通过并行计算方式实现该任务[7]。

2　结　果

2.1　剪接比对、分类与比较

根据 BLAT 客户端/服务器方式的比对结果获得定位信息,截取相应的基因组片段,利用 BLAT 命令行方式、Sim4 和 Elparser 分别对有实验证据支持的 NM_和 NR_类数据进行分析,结果见表 1。

表 1　Elparser,BLAT 和 Sim4 对 23593 个 Refseq NM_和 NR_记录的分类结果

Table 1　Types of 23593 RefSeq records labeled with "NM_"and "NR_" classified by Elparser, BLAT and Sim4

类型 Types	Elparser	BLAT	Sim4	Overlap
类型Ⅰ:完全匹配 Type Ⅰ: Perfect match	10 619	10 567	9 818	9 771
类型Ⅱ:可修订匹配 Type Ⅱ: Revisable match	11 368	12 688	12 863	10 943
类型Ⅲ:不可修订匹配 Type Ⅲ: Unrevisable match	1 596	328	902	203

注:Type Ⅰ(完全匹配)对 Elparser,Blat 和 Sim4 而言均指 100% 匹配。Elparser 的 Type Ⅱ(可修订匹配)指整体匹配率大于 95%,且没有匹配中断,以及剪接 5 bp 内没有不匹配;BLAT 的 Type Ⅱ(可修订匹配)指整体匹配率大于 95%,且没有大于 11 bp 的连续插入;Sim4 的 Type Ⅱ(可修订匹配)指每个外显子匹配率大于 90%,且没有匹配中断,头尾缺失的序列长度不超过全长的 5%。非各自的 Type Ⅰ 和 Type Ⅱ 的记录则属于相应的 Type Ⅲ(不可修订匹配)。

Note: A RefSeq record of Type Ⅰ (perfect match) for each of Elparser, Blat and Sim4 must have an overall matching ratio of 100%. A RefSeq record of Type Ⅱ (revisable match) for Elparser must have an overall matching ratio above 95% without gaps between exons and without mismatch and insertion/deletion less than 5 bp away from any splicing sites. A RefSeq record of Type Ⅱ (revisable match) for Blat must have an overall matching ratio above 95% with no continuous insertions over 11 bp. A RefSeq record of Type Ⅱ (revisable match) for Sim4 must have the matching ratio above 90% for every exon allowing no gaps between exons and missing no head or tail with more than 5% of its whole length. For any RefSeq record that could not be considered as a member of Type Ⅰ or Type Ⅱ, it is classified into Type Ⅲ (unrevisable match).

表 2　Elparser,BLAT 和 Sim4 对 RefseqNM_ 记录分类结果的交叉比较

Table 2　Cross comparison of different types of RefSeq records labeled with "NM_" by using Elparser, BLAT and Sim4

		完全匹配 Perfect match(Type Ⅰ)			可修订匹配 Revisable match(Type Ⅱ)			不可修订匹配 Unrevisable match(Type Ⅲ)		
	程序 Program	Elparser	Blat	Sim4	Elparser	Blat	Sim4	Elparser	Blat	Sim4
完全匹配	Elparser	10 477	10 402	9 673		73	639		2	165
Perfect match	Blat	10 402	10 424	9 638	10		625	12		161
(Type Ⅰ)	Sim4	9 673	9 638	9 680	4	42		3	3	
可修订匹配	Elparser		10	4	11 316	11 273	10 931		33	381
Revisable match	Blat	73		42	11 273	12 632	12 063	1 286		527
(Type Ⅱ)	Sim4	639	625		10 931	12 063	12 802	1 232	114	
不可修订匹配	Elparser		12	3		1 286	1 232	1 586	288	351
Unrevisable match	Blat	2		0	33		114	288	323	209
(Type Ⅲ)	Sim4	165	161		381	527		351	209	897

利用 NM_类的数据进行程序的交叉比较的结果见表 2。Elparser 与 BLAT 的差异主要在于 BLAT 会找到一些过短的匹配同时又丢失一些短匹配,这和 BLAT 内在的装配策略有关。Elparser 与 Sim4

的主要差别在于 Sim4 过于强调剪接边界而不能良好匹配,以及它内在对于内含子和外显子长度的限定使得它会丢失一些匹配,Elparser 在保证最大匹配率的同时优先考虑常用剪接规则,而且由于 Elparser 能够显示更多的基因组序列,便于人工判读合理的剪接边界。对于 BLAT 和 Sim4,剪接边界附近是否存在不匹配或缺失或插入的信息都不易获得,因此无法将剪接边界受影响的数据排除出第二类。而 Elparser 将发生匹配中断,不匹配或缺失或插入的位置在剪接边界 5 bp 以内的比对归为第三类,认为这种情况对基因结构存在重要影响,不宜进行标准化修订,因此 Elparser 的第三类记录偏多,但根据表 1 三类程序同样归类的记录达到 20 917,只有 11.3% 的记录有不同的分类结果,因此可以采用 Elparser 的分类结果,同时对照其他程序的情况,对 RefSeq 转录数据进行标准化修订。

2.2 非标准转录数据的特征

根据 Elparser 的分类标准,第三类 RefSeq 记录与基因组差异较大或者基因结构存在疑问。详细分析发现少数是因为相应的基因组尚未测序完全,如 NM_001585;还有少数记录是因为序列基因组跨度太长,未能被 BLAT 完整定位,如 NM_001094,改成向两侧各延伸 2 Mb 后 Elparser 能够完全匹配;一些记录存在较长的缺失或插入的,一种原因是因为多样性,如 NM_002319 具有 20 个连续的 TG 重复,而标准基因组只有 8 个,另一种原因可能是特殊转录方式或序列污染,如 NM_153827 和基因组相比多出近 250 bp,单独分析这一段仍找不到合适的匹配,NM_000222 和基因组比较多出 100 bp 左右,单独分析这一段发现和其上游的转录序列重复。还有一大部分记录与基因组差异较小但是有较短的匹配中断,或者不匹配、缺失、插入发生在剪接边界 5 bp 以内,这些差异很可能对剪接机制造成影响,对确定基因结构带来问题。因此,对于第三类数据的使用尤其要采取审慎的态度。

每个 RefSeq 记录用 Reviewed, Validated, Provisional, Predicted 等标明该记录的可信程度,其中 Reviewed 是通过专家组校对审核确信而且有实验数据支持的记录。当前版本中含有 Reviewed 标记的记录有 10 586 个,而属于第一类的只有 6 195 个,特别是属于第三类的有 512 个。这一结果说明 RefSeq 经过专家组校对审查确信的序列并非都是标准转录,如 RefSeq release 7(2004 年 9 月 12 日)版本的人类转录数据中 NM_000074 和 NM_000534 都是带有具有 Reviewed 标记,它们各自的版本为 NM_000074.1 和 NM_000534.2。根据本文的流程分析以及 Elparser 的分类标准,它们是属于的第三类记录的,其中 NM_000074 有 32 个连续的 AC 重复,而标准基因组只有 26 个;NM_000534 和基因组比较多出 50 bp 以上的一段序列,进一步分析提示是自身重复,在本文分析的版本(RefSeq 2005-4-18)中,NCBI RefSeq 已对这两个记录进行了相应的修订,为 NM_000074.2 和 NM_000534.3,比对结果分别为第二类和第一类,而归为第二类的 NM_000074 和基因组只有末端一个碱基的差别。因此这一结果说明 RefSeq 经过专家组校对审查确信的序列也并非都是标准转录,基于此类记录进行的各类分析也存在潜在问题,同时也说明我们的分析流程是合理有效的。

遗传变异以及疾病造成的转录差异会通过蛋白产物的差异表现出来,从而导致非标准转录序列相关的蛋白产物可能也是非标准的。我们进一步分析 RefSeq 的 NM_类记录与标准基因组的差异是否会对开放阅读框架产生影响,以判断 RefSeq 序列关联的蛋白产物的标准性。根据 Elparser 程序报告的差异序列对照 RefSeq 对 CDS 的标注分析差异对于 CDS 区域和阅读框相位的影响,统计结果见表 3。23 389 个 NM_记录中有 7 540 个与基因组的差异可能影响其蛋白产物。该分析表明与标准转录对应的蛋白序列也有其修订的必要性。该工作目前正在进行中。

表 3 与基因组的差异影响编码区的 RefSeq 记录数量统计

Table 3 Number of RefSeq records labeled with "NM_"whose CDS was modified by difference comparison with the "standard" genome

RefSeq NM_类型记录 RefSeq NM_Record	总数 All	阅读框相位移动 Frame shift	阅读框相位内插入或删除 Frame indel	不匹配 Unmatch	差异在编码区外 Out CDS
类型Ⅱ:可修订匹配 Type Ⅱ: Revisable match	11 316	332	83	5 927	4 974
类型Ⅲ:不可修订匹配 Type Ⅲ: Unrevisable match	1 586	325	62	811	388

3 讨 论

NCBI 的 RefSeq 数据库对于转录组的研究有着极大的促进作用,但其中存在的质量问题、非普遍的多样性问题以及可能是疾病导致的异常状态,严重影响了其作为正常转录机理研究的参考标准性。因此我们提出标准转录的概念,即与基因组完全匹配的转录,并基于人类 RefSeq 数据库构建了人类标准转录数据集,为分子生物学实验设计以及对标准数据集比较依赖的生物信息学研究提供了像标准基因组一样的统一参考标准,这个标准将对基因组学、转录组学、剪接组学乃至蛋白质组学的研究具有重要的促进意义。

根据"中心法则",转录后序列应与基因组有很好的匹配。实际情况下,基因所发生的突变可包括点突变、插入、删除或插入转座重复元件,以及微卫星的扩展或收缩,这些突变可能严重影响基因的表达调控或者功能。通过剪接比对全面考察 RefSeq 与基因组的对应关系,Elparser 比对结果能够给出详细的比对情况,尽可能避免不符合生物学意义的短外显子,以及可能影响表达调控的变异,同时产生与标准基因组数据对应的标准转录数据集。根据 Elparser 的分类几乎一半的 RefSeq 记录还是和标准基因组吻合的,另一半的大部分与基因组的差异很小,可以归结为多样性或突变的原因,但该类记录其多样性或突变可能改变相应蛋白产物,因此基于 RefSeq 数据库进行基因结构与功能研究时应考虑这些差别。对于另外 1 000 多个与基因组匹配情况较差的 RefSeq 记录,在排除基因组没有测序完全的原因外,有些基因表现出微卫星的扩展或收缩,有些存在较长的插入但该片段无法在目前的基因组中找到或者是自身重复的片段,推测其可能是发生变异或者病理条件下所得到的结果,或者可能是不常见的转录加工方式甚至测序错误,应考虑采用实验方法进一步验证。目前有很多的方法利用 EST 与基因组比对构建转录图谱以及鉴定可变剪接,本文的分析也可以揭示这类方法要考虑的数据质量问题。所有基于 RefSeq 的研究应根据本文研究结果进行重新分析以修正以前的结论。对于强烈依赖于训练数据集的准确性的理论分析研究例如基因结构的识别、剪接边界的特征分析等而言,本研究也为此提供了很好的参考数据。详情可访问 http://biocompute.bmi.ac.cn/transcriptome/index.htm。

参 考 文 献(References):

[1] Kim Pruitt, Tatiana Tatusova, James Ostell. The reference sequence (RefSeq) project. NCBI handbook chapter 18. http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/handbook/.

[2] ZHANG De-Li, JI Liang, LI Yan-Da. Analysis, Identification and correction of some errors of model refseqs appeared in NCBI human gene database by in silico cloning and experimental verification of novel human genes. *Acta Genetica Sinica*, 2004, 31(5): 431～443.
张德礼, 季　梁, 李衍达. 通过新基因计算机识别与实验确认对 NCBI 人类基因数据库一些模式参考序列错误的分析与纠正. 遗传学报, 2004, 31(5): 431～443.

[3] Florea L, Hartzell G, Zhang Z, Rubin G M, Miller W. A computer program for aligning a cDNA sequence with a genomic sequence. *Genome Res*, 1998, 8(9): 967～974.

[4] Kent W J. BLAT-The BLAST-Like alignment tool. *Genome Res*, 2002, 12(4): 656～664.

[5] Rice P, Longden I, Bleasby A. EMBOSS: The european molecular biology open software suite. *Trends in Genetics*, 2000, 16(6): 276～277.

[6] HANG Xing-Yi, ZHAO Dong-Sheng, ZHANG Cheng-Gang. The application and evaluation of Blat in transcriptome analysis. *China Journal of Bioinformatics*, 2005, 3(2), 85～88.
杭兴宜, 赵东升, 张成岗. 序列比对程序 Blat 在转录组数据分析中的应用. 生物信息学, 2005, 3(2), 85～88.

[7] ZHAO Dong-Sheng, HANG Xing-Yi, Li Zhi-Feng, ZHANG Cheng-Gang. Computing resource of AMMS biomedicine super-computing center and its applications. *Bulletin of the Academy of Military Medical Sciences*, 2005, 29(4): 363～367.
赵东升, 杭兴宜, 李稚锋, 张成岗. 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. 军事医学科学院院刊, 2005, 29(4): 363～367.