ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 4,July 2018,pp. 637-645
DOI:10.16337/j.1004-9037.2018.04.008

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:+86-025-84892742

基于 RNA-seq 的基因训练集构建方法

段荣静[1,2]　刘金定[1]

(1. 南京农业大学领域知识关联研究中心,南京,210095; 2. 南京农业大学生物信息学中心,南京,210095)

摘　要: 针对基因组新测序物种缺乏高质量的基因结构用于从头预测软件训练的现状,本文提出了一种以新测序物种自身 RNA-seq 组装为基础的可靠基因训练集构建方法(Building reliable training gene set,BRTGS)。该方法利用 RNA-seq 组装获得大量初始基因结构,然后根据蛋白同源证据筛选具有正确且编码区相对完整的基因结构,最后综合利用 RNA-seq 组装结构和蛋白同源证据统计信息确定的基因起始密码子和终止密码子位置,从而获得基因完整的编码结构。实验结果表明,该方法不仅可为各种组装水平的基因组构建高质量的基因训练集,而且从头预测软件在这些基因集上训练后能够获得很好的预测性能。

关键词: 生物信息学; 基因结构;RNA-seq;蛋白同源;训练基因集
中图分类号: TN713;Q751　　**文献标志码:** A

Construction Method of Gene Training Set Based on RNA-seq

Duan Rongjing[1,2], Liu Jinding[1]

(1. Research Center for Correlation of Domain Knowledge, Nanjing Agricultural University, Nanjing, 210095, China; 2. Bioinformatics Center, Nanjing Agricultural University, Nanjing, 210095, China)

Abstract: There are no extant high-quality gene structures for newly sequenced genomes to train ab initio gene prediction algorithms. In the study, we present the building reliable training gene set (BRTGS) computational method for building reliable training gene set from RNA-seq assembly. Firstly, the initial gene structures are obtained from RNA-seq assembly. Then, the gene structures with complete and correct coding region are identified with the alignments of transcripts against homology protein. Finally, the sites of start and stop codon are determined according to the homology evidences and RNA-seq assembly structures. Experimental results show that BRTGS can build high-quality of training gene set for various genomes and ab initio algorithms trained on the gene sets can obtain good prediction performance.
Key words: bioinformatics; gene structure; RNA-seq; protein homology; training gene set

引　　言

从头基因预测软件(下简称预测软件)常用于识别基因组上的编码基因结构,在基因组注释中发挥着非常重要的作用。预测软件使用前必须优化其模型参数使之适应被测序物种。除了几个非监督预测

基金项目: 国家自然科学基金(31301691)资助项目;教育部中央高校基本业务费(KYZ201667,KJQN201430)资助项目。
收稿日期: 2014-05-09;**修订日期:** 2016-10-14

软件外[1]，大多数预测软件都属于监督训练软件[2-5]，其训练过程必须建立在足够数量的可靠基因结构上。对于新测序的基因组而言，其研究基础薄弱，往往缺乏足够数量的编码基因结构用于训练，这大大限制了预测软件的使用。因此，构建高质量的基因训练集是一个非常有价值的工作。

为新测序物种构建基因训练集（下简称训练集）是一个非常耗时、费力的过程，有时甚至比基因组测序本身还繁琐。构建训练集不仅需要克隆全长基因，而且还需要专家人工矫正基因结构。另外，由于高表达和高保守的基因容易被发现和验证，所以人工克隆方法构建的训练集通常会出现样本不平衡的现象。考虑到人工准备基因训练集的繁琐和缺陷，很多基因组注释项目直接用其他近缘物种基因对预测软件进行训练，或者直接使用其他物种模型参数跳过预测软件训练环节。预测软件性能依赖物种特异性的模型参数，而物种间的差异性又是必然存在的，因此这种做法必然会导致预测性能下降，尤其在亲缘关系较远物种之间更是如此。为此，研究人员提出在基因组上识别保守基因结构用于构建训练集的方法[6-8]。这些方法定义了真核生物共同拥有的核心蛋白家族并提供了识别这些蛋白对应基因结构的计算方法。这些方法存在两个缺陷：(1)真核生物核心基因过于保守，导致训练集不能特异优化被测序物种的模型参数；(2)测序基因组不完整导致基因组上核心基因丢失，从而引发训练基因数量不足的现象。RNA-seq 是快速获得基因表达数据的有效方法，已成为基因组注释项目中必不可少的步骤。理论上，RNA-seq 可以获得全部表达的基因序列，而且通过组装读段可以获得全长序列[9-11]。将测序读段比对到基因组上，可以获得大量外显子内含子位置信息，进一步组装将得到外显子-内含子相互串联的基因结构。但是由于测序和组装错误的存在，导致获得的基因结构质量不能满足构建训练集的要求[12]。另一方面，基于 RNA-seq 组装的基因结构无法直接给出编码区的结构以及起始密码子和终止密码子位置信息，因此进一步降低了组装结果的可用性。

针对基因组新测序物种基因匮乏，难以满足预测软件训练需要的现状，本文提出了一种以新测序物种自身 RNA-seq 组装为基础，结合蛋白同源证据的可靠基因训练集构建方法（Building reliable training gene set，BRTGS）。该方法利用蛋白同源证据排除不完整和错误的 RNA-seq 组装结构，利用 RNA-seq 组装结构弥补蛋白同源证据难以获得同源区外基因结构以及起始密码子和终止密码子位置的缺陷。将该方法应用于果蝇、人、拟南芥和水稻等几种模式生物基因组上，实验结果表明该方法不仅可以为各种组装水平的基因组构建足够数量的高质量训练基因，而且训练后的预测软件可以获得很好的预测性能。

1 从头基因预测软件训练以及相关数据基础

1.1 从头基因预测软件训练

从头预测软件训练是指利用物种已有的基因结构信息优化其模型参数，使之适应该物种。虽然预测软件训练效果会随着训练基因的数量增加而逐步提高，但实践表明当训练基因数量超过 1 000 h 时，训练效果的提升不再明显。因此，构建基因训练集时更注重基因质量而不是数量。高质量的训练基因不仅要具有完整的外显子-内含子串联结构，而且要有正确的起始密码子和终止密码子位置信息（图 1）。RNA-seq 组装可以快速获得大量基因外显子-内含子串联结构信息，但是这些结构信息不仅存在组装错误，而且缺少起始密码子和终止密码子位置信息，因此不能直接用于构建基因训练集。

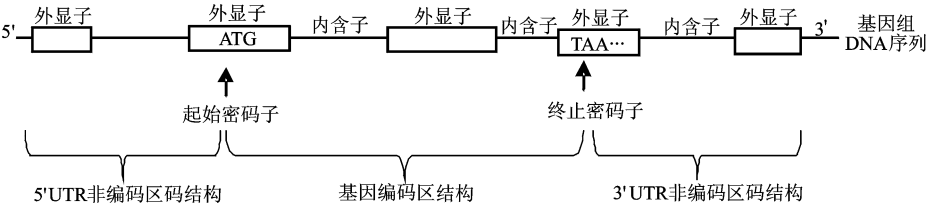

图 1 真核基因结构

Fig. 1 Structure of eukaryotic genes

1.2 RNA-seq 组装中的编码基因结构

目前公共数据库中存在大量低 N50 基因组,甚至存在 N50 为 10K 的基因组。从头预测软件训练一般需 200～1 000 个可靠的基因结构。为了模拟各种组装水平的基因组,果蝇、人、拟南芥和水稻等几种模式生物基因组被打碎随机抽取序列构成 N50 分别为 2K,5K,10K,20K,40K,80K,160K,320K,640K 和 1M 的片段化基因组(这些基因组在下文被分别引用为 G2K,G5K,G10K,G20K,G40K,G80K,G160K,G320K,G640K 和 G1M)。在低组装水平的基因组中大量基因组序列没有被组装到一起,导致跨度较大的基因被分散在多个基因组序列中而被重复计数,因此随着 N50 下降,基因总数不断提高(图 2(a)),但完整基因数量不断下降(图 2(b))。在 4 个物种中,人类基因平均跨度最大,所以基因被碎片化的程度最高。在 G2K 基因组中,人类基因总数上升比例最高达到了 460%,而完整基因数量下降最多达到 42%。从数字上看,在 G2K 基因组中,4 个物种分别保留了 8 228,10 003,16 349 和 22 431 个完整基因,这表明组装水平较低的基因组中仍然保留着大量完整基因资源可用于构建基因训练集。

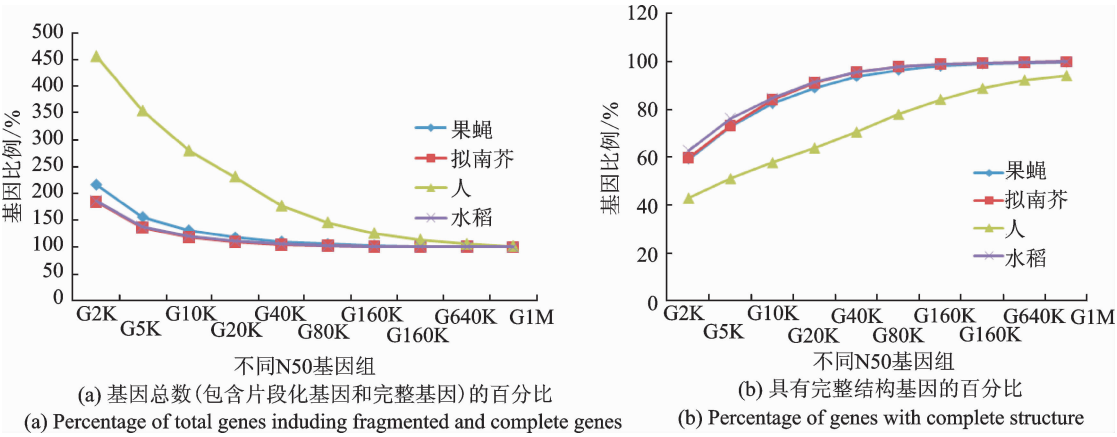

图 2 N50 对保留在片段化基因组中基因数量的影响

Fig. 2 Influence of genome N50 on the numbers of genes retained in fragmented genomes

虽然低水平组装的基因组中仍然存在大量完整基因结构,但需要确定 RNA-seq 组装结果是否能够提供足够数量的正确外显子-内含子串联结构(简称基因结构)。鉴于生物学实验通常 3 个重复的要求,本文随机地为每个物种准备了 3 个 RNA-seq 数据。首先将每个 RNA-seq 数据比对到基因组上,然后组装比对结果获得基因结构。一般而言,将多个 RNA-seq 组装结果合并将得到更多基因结构(基因结构数量并不总是随着 RNA-seq 数量增加而增长,当 RNA-seq 测序总量足够饱和时,基因结构数将不再增长),因此,本文进一步利用 CuffMerge 将 3 个 RNA-seq 组装结果合并在一起产生 1 个非冗余组装结果,这样每个基因组上得到 4 个 RNA-seq 组装结果。由于绝大多数预测软件只预测编码区部分,所以本文只统计编码区部分的结构被完整且正确组装的数量结果如表 1 所示。第 1 个样品 RNA-seq 数据在 G2K 基因组上组装获得正确编码基因结构数量最少,组合 3 个 RNA-seq 数据在 G1M 上组装获得正确编码基因结构数量最多。由于这两种情况分别代表了最差组装结果和最优组装结果,因此本文后续研究只关注这两种 RNA-seq 组装结果。在 G2K 基因组上,第 1 个样品 RNA-seq 组装结果分别为 4 个物种提供了 2 482,6 627,5 649 和 7 928 个正确编码基因结构。将 3 个样品 RNA-seq 组装结果合并起来,为 G2K 基因组分别提供 4 933, 10 941, 8 764 和 11 846 个正确编码基因结构。对于 G1M 基因组而言,无论是第 1 个样品 RNA-seq 还是组合 RNA-seq 都组装出非常多的正确编码基因结构。果蝇的第 1 个样品 RNA-seq 数据在 G1M 上组装的正确编码基因结构数量最少为 3 866 个,人类的组合 RNA-seq 数据在 G1M 上组装的正确编码基因结构数量最多达到 19 535 个。预测软件的训练只需几百个正确编码

基因结构，从 RNA-sq 组装结果上看，即使在低组装水平的基因组（G2K）上采用单个 RNA-seq 数据，也能组装出近2 500个正确编码基因结构，这表明基于 RNA-seq 的组装结果可为预测软件训练提供充足的编码基因结构。

表 1 RNA-seq 组装的完整编码基因结构的数量

Tab. 1 Number of complete gene coding structures assembled correctly by RNA-seq

RNA-Seq	果蝇		拟南芥		水稻		人	
	G2K	G1M	G2K	G1M	G2K	G1M	G2K	G1M
样品 1	2 482	3 866	6 627	11 366	5 649	11 352	7 928	12 971
样品 2	3 215	4 876	8 482	14 600	5 748	11 468	9 053	14 781
样品 3	3 006	4 598	7 899	13 502	6 361	12 642	9 133	14 966
合并	4 933	10 176	10 941	19 071	8 764	17 753	11 846	19 535

注：每个物种的样品 1、样品 2 和样品 3 对应的具体 RNA-seq 见 3.1 节实验数据。

2 基于 RNA-seq 组装的基因训练集构建方法

2.1 包含完整编码区的 RNA-seq 转录本识别

首先用 Tophat 将 RNA-seq 短读段序列比对到基因组上，然后用 Cufflinks 对比对结果进行组装获得基因结构，最后利用 CuffMerge 对每个基因组的 3 个 RNA-seq 组装结果进行合并获得非冗余基因结构。根据 RNA-seq 组装的基因结构获得对应的转录本序列后，用局部比对搜索工具 X（Basic local alignment search tool X, BLASTX）将转录本序列比对到参考蛋白数据库中（E-value 设为 1e−10），只保留前 50 个洞（Gap）百分比小于 3%且对齐区域覆盖参考蛋白序列 75%以上的比对结果用于后续分析。如果一个基因结构符合上述比对结果，那么该基因结构将被认为可能包含完整编码区。

2.2 起始密码子和终止密码子位置范围计算

正确识别起始密码子和终止密码子位置对提升训练集质量具有重要的意义，为此 BRTGS 进一步利用同源参考蛋白序列初步确定起始密码子和终止密码子位置范围，以提高后续起始密码子和终止密码子的预测可靠性，结果如图 3 所示。转录本上的每个比对结果都可以推导一对起始密码子和终止密码子位置。在转录本对齐起始位置向 5 端延伸一定长度可以推导出起始密码子位置，这个长度为参考蛋白序列上 N 端非对齐氨基酸长度的 3 倍。同样，在转录本对齐终止位置向 3 端延伸一定长度可以推导出终止密码子位置。一个转录本所有比对结果推导出的起始密码子和终止密码子位置收集在一起就构成起始密码子和终止密码子位置范围。

2.3 转录本起始密码子和终止密码子位置确定

根据起始密码子和终止密码子确定方法，可将包含完整编码结构的转录本（Complete coding transcripts, CCTs）分为以下 4 类：

（1）CCT I：如果最佳比对结果的对齐区域完全覆盖参考蛋白序列，对齐区域 5 端的 3 个碱基为起始密码子，3 端向后 3 个碱基为终止密码子，那么这个转录本属于 CCT I，转录本上的对齐区域为编码区。

（2）CCT II：如果最佳比对结果的对齐区域覆盖参考蛋白序列的 N 端，对齐区域 5 端的 3 个碱基为起始密码子，3 端向后出现的第 1 个终止密码子位于终止密码子范围内，那么这个转录本属于 CCT II。转录本的编码区从对齐区域开始到推导的终止密码子上游位置结束。

（3）CCT III：如果最佳比对结果的对齐区域覆盖参考蛋白序列的 C 端，3 端向后 3 个碱基正好是终止密码子，用 ATGpr[13] 计算推导出的起始密码子位于起始密码子位置范围内，那么这个转录本属于

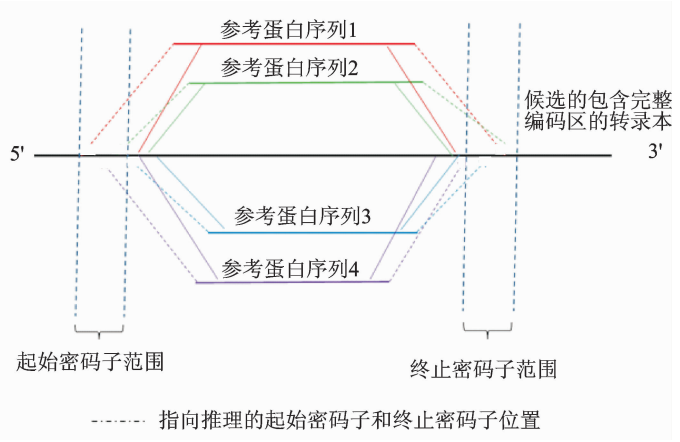

图 3 计算起始密码子和终止密码子范围

Fig. 3 Calculation of site range of start codon and stop codon

CCT III。转录本的编码区为从 ATGpr 计算的起始密码子位置开始到终止密码子上游位置结束。

(4)CCT IV:如果最佳比对结果的对齐区域既不能覆盖参考蛋白序列的 N 端,也不能覆盖 C 端,但是用 ATGpr 计算推导出的起始密码子位于起始密码子位置范围内,下游第 1 个终止密码子位于终止密码子范围内,那么这个转录本属于 CCT IV。转录本的编码区为从 ATGpr 计算的密码子位置开始到推导的终止密码子上游位置结束。

在确定起始密码子和终止密码位置后,再次计算转录本编码区的长度,如果编码区长度大于最佳比对的参考蛋白序列长度 130%,那么这个转录本将被淘汰。

2.4 同源聚类后排序生成基因训练集

在构建的 4 类 CCTs 中,CCT I 的起始密码子和终止密码同时具有表达、同源和预测证据支持,其可靠性最高;CCT II 的起始密码子同时有转录、同源和预测证据支持,终止密码子只有表达和预测证据支持,其可靠性次之;CCT III 的起始密码子只有表达和预测证据支持,终止密码子同时有表达、同源和预测证据支持,其可靠性再次;CCT IV 的起始密码子和终止密码子都只有表达和预测证据支持,其可靠性在 4 类 CCTs 中最低。构建基因训练集不仅要考虑基因结构可靠性,而且还要考虑基因的同源性,应尽量避免使用过多同源基因构造训练集,以防止产生训练偏好性。图 4 给出了根据可靠性和同源性排序生成基因训练集的示意图,由图 4 可知,BRTGS 对 4 类 CCTs 进行排序处理的过程为:

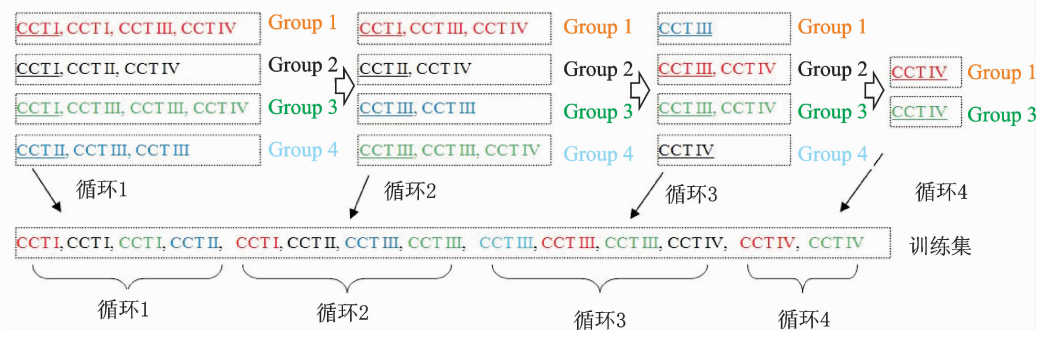

图 4 根据可靠性和同源性排序生成基因训练集示意图

Fig. 4 Demonstration of building training gene set according to reliability and homology

(1)用 OrthoMcl[14] 对所有的 CCT s 进行同源聚类,将同源的 CCTs 聚类到同 1 个同源分组中,然后对每个同源分组内的 CCTs 按照"CCT I > CCT II > CCT III > CCT IV"进行可靠性排序,以确保在每个同源分组中高可靠的 CCTs 被优先选择用于构建训练集。

(2)对同源分组按照包含 CCTs 类别进行排序。排序后包含 CCT I 的同源分组将排在最前面,如果同源分组中没有 CCT I,那么包含 CCT II 的同源分组排在最前面,以此类推。如图 4 所示,在第 1 轮排序中 Group 4 不包含 CCT I 被排到最后;从 4 个 Group 中取走第 1 个 CCT 后进入第 2 轮排序,Group 1 具有 CCT I 被排在最前,Group 4 和 Group 3 都没有 CCT I 和 II,所以被排在最后;再次从 4 个 Group 中取走第 1 个 CCT 后进入第 3 轮排序,Group 2 不包含 CCT III 被排到最后。

(3)根据步骤(2)对同源分组排序的结果,依次将每个同源分组中的第 1 个 CCT 移到训练集中。

(4)重复步骤(2,3),直到每个分组中的 CCT 都转移到训练集中。

经过上述排序处理后,训练集中的同源基因将被打散。由于训练时,BRTGS 只取训练集中前 1 000 个基因用于训练,因此不仅可以避免同一同源分组中的 CCTs 被过度采用的现象,而且确保每个同源分组中高可靠性 CCTs 被优先采用。

3 实验与结果分析

3.1 实验数据

果蝇、拟南芥、人和水稻 4 个模式物种基因组数据来自于 Ensemble(http://asia.ensembl.org),版本号分别为 BDGP6.31,TAIR10.31,GRCh38.p2 和 IGSP-1.0.31。为了模拟不同组装水平的基因组,按照 N50 定义将每个物种基因组随机打碎。在产生低于 N50 值的序列时,如果序列长度小于 100 bp 时,那么这个序列将会被丢弃。在产生大于 N50 值的序列时,最长序列不超过 1 Mb(G640K 和 G1M 的最长序列分别为 2 Mb 和 8 Mb)。4 个物种 RNA-seq 数据都来源于 NCBI 的 SRA(http://www.ncbi.nlm.nih.gov/sra)数据库,访问编号分别为"SRR3091999,SRR3091976,SRR3138705","ERR754089,SRR1792926,ERR1109342","SRR3184282,SRR3184286,SRR3184279"和"SRR2156305, SRR2047477,SRR2048540"。每个物种的 3 个 RNA-seq 数据都分别被依次标为样品 1、样品 2 和样品 3。RNA-seq 比对和组装分别采用 Tophat 和 Cufflinks 实现,参数使用默认值[15-17]。参考蛋白序列来自 NCBI-RefSeq 数据库,并去除掉来自这 4 个物种的蛋白序列。

3.2 实验方案

为了评估 BRTGS 的性能,G2K 上最差 RNA-seq 组装结果和 G1M 上最佳 RNA-seq 组装结果(合并了 3 个 RNA-Seq 的组装结果)分别用于构建基因训练集。作为比较,核心真核基因定位方法(Core eukaryotic genes mapping approach, CEGMA)(v2.5)、通用单拷贝直系同源基因检测(Benchmarking universal single-copy orthologs, BUSCO)(v1.22)和转录本解码器(TransDecoder)(3.0)也用于构建基因训练集。CEGMA 和 BUSCO 利用保守的核心基因构建训练基因集,因此直接用在 G2K 和 G1M 基因组上。TransDecoder 会为每个转录本序列报道多个编码区,这里只保留标志为"Complete"且最长的结果作为基因编码结构。本文将对 4 种方法产生的训练基因数量和质量进行比较。

预测软件的训练效果也是构建基因训练集优劣的重要评价指标。本文用训练集中的前 1 000 个基因(如果不足,则训练集中的全部基因都用于训练)训练 Augustus(v3.2.1),并用其训练后的预测性能评估训练集的训练效果。测试集来自基因组上和训练集不重叠的参考基因。本文将用训练后 Augustus 在核酸水平、外显子水平以及基因水平上的预测性能评估训练效果,训练和评估过程完全按照 Augustus 操作手册进行。

3.3 实验结果分析

首先对 4 种方法构建训练基因数量情况进行比较,结果如表 2 所示。总体上 CEGMA 和 BUSCO

构建的基因数量明显偏少,主要原因是 CEGMA 和 BUSCO 构建基因训练集被限制在其定义的真核(或者某个谱系)核心基因集上。这些核心基因数量本来就偏少,在低水平组装的基因组上构建的训练基因则进一步减少。BRTGS 和 TransDecoder 构建训练基因数量偏多,主要原因是 BRTGS 和 TransDecoder 构建基因训练集来自于本物种 RNA-seq 测序数据。RNA-seq 测序深度高,其组装结果基本可以覆盖基因组上全部基因(除了少数极低丰度表达或不表达的基因外),因此,构建的训练基因数量明显比 CEGMA 和 BUSCO 多。和 TransDecoder 相比,BRTGS 进一步采用蛋白同源证据检验,因此构建的基因数量比较少。尽管如此,BRTGS 在每个基因组上仍然构建了超过 1 600 个基因的训练集。BRTGS 构建的训练基因并没有完全来自 CCT I,这说明其构建的基因训练集并不完全来自跨物种保守基因。在拟南芥和人的基因组上大约 64%和 61%来自 CCT I,在果蝇和水稻上只有 26%和 39%来自 CCT I。此外,在 BRTG 构建的基因集中,只有不足 5%的基因被同源聚类到一起,因此训练集不具有同源偏好性。

表 2 4 种方法在不同基因组上构建的训练基因数量

Tab. 2 Numbers of training genes built by four methods on different genomes

物种	基因组	RNA-seq	CEGMA	BUSCO	Trans Decoder	BRTGS					
						总数	CCT I/%	CCT II/%	CCT III/%	CCT IV/%	同源聚类分组数
果蝇	G2K	样品 1	417	1 493	4 022	1 650	28	23	21	28	1 584
	G1M	组合样	452	2 480	8 230	2 573	25	23	21	31	2 436
拟南芥	G2K	样品 1	411	108	8 557	4 550	64	20	12	4	4 301
	G1M	组合样	453	151	15 519	11 085	64	20	12	5	10 397
人	G2K	样品 1	300	659	7 456	2 825	61	19	16	4	2 717
	G1M	组合样	383	1 262	13 543	6 638	61	19	13	7	6 365
水稻	G2K	样品 1	359	100	10 576	3 534	38	16	32	15	3 364
	G1M	组合样	445	145	15 948	8 835	40	16	30	14	8 322

其次,对 4 种方法构建训练基因质量进行比较,结果如表 3 所示。以 4 个物种基因组注释的基因结构为参考,统计 4 种方法构建的基因训练集的准确性。在外显子水平上,当编码外显子两侧边界和参考基因外显子完全重叠时,该外显子结构被认为正确,否则为错误。在基因水平上,构成基因编码结构的外显子和内含子与参考基因完全一致时,该基因结构被认为正确,否则为错误。准确率为构建训练集中正确外显子(或基因)数与外显子(或基因)总数的比值。在所有测试的基因组上,BRTGS 构建的基因结构可靠性最高。BRTGS 平均外显子准确性达到 93%,分别比 CEGMA,BUSCO 和 TransDecoder 高 16%,24%和 20%;平均基因结构准确性为 77%,分别比 CEGMA,BUSCO 和 TransDecoder 高 40%,40% 和 31%。BRTGS 构建的基因结构具有最高的可靠性,主要原因是其利用了表达、同源和预测 3 个方面证据确保基因结构的完整性和正确性。而 CEGMA 和 BUSCO 只利用跨物种保守基因的蛋白同源证据构建基因训练集,难以准确捕捉同源区域外的外显子结构以及起始密码子和终止密码子位置。虽然 TransDecoder 也利用了 RNA-seq 组装数据(即 RNA-seq 表达组装证据),但没有对组装结构正确性以及完整性进行检查,因此基因结构的可靠性也比 BRTGS 低。

最后,用训练效果(即预测软件训练后的预测性能)对 4 种方法构建的训练集进行比较,结果如表 4 所示。在每个测试的基因组上,基于 BRTGS 基因集训练的 Augustus,在核酸、外显子和基因水平上获得的灵敏度和特异度都比其他 3 种方法高(详细数据未列出)。基因层面的灵敏度和特异度是检查训练

效果的最重要指标。用 BRTGS 基因集训练的 Augustus，平均基因灵敏度为 33%，分别比 CEGMA，BUSCO 和 TransDecoder 高 9%，12% 和 9%；平均基因特异度为 22%，分别比 CEGMA，BUSCO 和 TransDecoder 高 6%，8%和 4%。总体上，BRTGS 和 TransDecoder 构建基因集的训练效果比 CEGMA 和 BUSCO 好。BRTGS 构建的基因训练集具有最好的训练效果主要得益于其充足的训练基因数量和可靠的基因结构。CEGMA 和 BUSCO 构建基因集的训练效果差的原因来自两个方面：(1)训练基因数量偏少、可靠性偏低；(2)在跨物种高度保守的核心基因上训练，难以获得基因组新测序物种自身特性。

表 3 4 种方法构建的基因训练集质量

Tab. 3 Quality of training genes built by four methods

物种	Genome	RNA-seq	外显子的准确率/%				基因结构的准确率/%			
			CEGMA	BUSCO	TransDecoder	BRTGS	CEGMA	BUSCO	TransDecoder	BRTGS
果蝇	G2K	样品 1	81	91	57	94	48	70	59	87
	G1M	合并样	85	89	77	97	69	72	57	89
拟南芥	G2K	样品 1	86	59	82	97	41	34	55	90
	G1M	合并样	89	59	89	98	51	30	60	92
人	G2K	样品 1	54	72	75	92	10	31	41	65
	G1M	合并样	65	72	79	94	11	30	42	67
水稻	G2K	样品 1	75	53	57	86	33	20	26	64
	G1M	合并样	79	55	63	86	36	16	34	66

表 4 基于 4 种方法基因训练集的预测软件平均预测性能

Tab. 4 Average prediction performance based on training sets built by four methods

训练集	灵敏度/%				特异度/%			
	CEGMA	BUSCO	TransDecoder	BRTGS	CEGMA	BUSCO	TransDecoder	BRTGS
核酸(Nucleotide)	77	78	82	84	68	66	69	72
外显子(Exon)	55	53	63	67	49	49	55	61
基因(Gene)	24	21	24	33	16	14	18	22

4 结束语

长期以来，为从头预测软件准备训练基因集需要专家人工矫正基因结构，这是个耗时、费力的工作。高通量测序技术能够快速获得大量基因表达序列，尤其是三代测序技术的应用能够产生更长甚至全长 mRNA 序列，这为构造基因训练集进一步提供了十分有价值的序列数据。然而从头预测软件训练不仅需要基因序列数据，而且还要基因元件(如起始密码子、终止密码子、内含子以及外显子等)在基因组上的位置信息。本文提出了一种利用物种自身 RNA-seq 组装，结合同源证据构建基因训练集的方法，该方法克服了单独利用蛋白同源证据或 RNA-seq 表达证据构建基因训练集的缺陷，提高了构建基因训练集的质量。通过比较分析，该方法构建的基因训练集不仅具有更高质量的基因结构，而且具有很好的训练效果。值得注意的是，由于基因编码区决定了基因功能，因此本文提出的方法只关注训练基因编码区结构的准确性。事实上，大量研究表明，基因的非编码区(Untranslated region，UTR)也非常重要，比如 microRNA 经常结合在 3UTR 上实现对基因的调控。下一步研究重点是进一步优化本文提出的方法，

利用 RNA-seq 测序数据尤其是三代测序长序列数据为基因训练集补充结构完整的 UTR。

参考文献:

[1] Lomsadze A, Burns P D, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm[J]. Nucleic Acids Research, 2014, 42(15): 119.

[2] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA[J]. Journal of Molecular Biology, 1997, 268(1): 78-94.

[3] Parra G, Blanco E, Guigo R. GeneID in Drosophila[J]. Genome Research, 2000, 10(4): 511-515.

[4] Korf I. Gene finding in novel genomes[J]. BMC Bioinformatics, 2004, 5: 59.

[5] Keller O, Odronitz F, Stanke M, et al. Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species[J]. BMC Bioinformatics, 2008, 9: 278.

[6] Parra G, Bradnam K, Ning Z, et al. Assessing the gene space in draft genomes[J]. Nucleic Acids Research, 2009, 37(1): 289-297.

[7] Simao F A, Waterhouse R M, Ioannidis P, et al. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs[J]. Bioinformatics, 2015, 31(19): 3210-3212.

[8] 马玉韬,车进,关欣,等. 加窗窄通带滤波器蛋白质编码区预测算法[J]. 数据采集与处理,2013,28(2):129-135.
Ma Yutao, Che Jin, Guan Xin, et al. Prediction algorithm for protein coding regions based on windowed narrow pass-band filter [J]. Journal of Data Acquisition and Processing, 2013, 28(2):129-135.

[9] Birol I, Jackman S D, Nielsen C B, et al. De novo transcriptome assembly with ABySS[J]. Bioinformatics, 2009, 25(21): 2872-2877.

[10] Trapnell C, Williams B A, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation[J]. Nature Biotechnology,2010, 28(5): 511-515.

[11] Mezlini A M, Smith E J, Fiume M, et al. iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data[J]. Genome Research, 2013, 23(3): 519-529.

[12] Steijger T, Abril J F, Engstrom P G, et al. Assessment of transcript reconstruction methods for RNA-seq[J]. Nature Methods, 2013, 10(12): 1177-1184.

[13] Salamov A A, Nishikawa T, Swindells M B. Assessing protein coding region integrity in cDNA sequencing projects[J]. Bioinformatics, 1998, 14(5): 384-390.

[14] Li L, Stoeckert C J, Jr, Roos D S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes[J]. Genome Research, 2003,13 (9): 2178-2189.

[15] Trapnell C, Pachter L, Salzberg S L. TopHat: Discovering splice junctions with RNA-Seq[J]. Bioinformatics, 2009, 25(9): 1105-1111.

[16] Trapnell C, Williams B A, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation[J]. Nat Biotechnol, 2010, 28(5):511-U174.

[17] Trapnell C, Roberts A, Goff L,et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks[J]. Nat Protoc, 2014, 9(10):2513-2513.

作者简介:

段荣静(1980-),女,助理研究员,研究方向:生物信息学,E-mail: duangri@njau.edu.cn。

刘金定(1978-),男,博士,副教授,研究方向:生物信息学与数据挖掘,E-mail: liujd@njau.edu.cn。

(编辑:王静)