Doc-Start BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1–3

BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS

Katharina J. Hoff [1]*, Simone Lange [1], Alexandre Lomsadze [3], Mark Borodovsky [2,3,4]* and Mario Stanke [1]

[1]Ernst Moritz Arndt Universität Greifswald, Institute for Mathematics and Computer Science,
Walther-Rathenau-Straße 47, 17487 Greifswald, Germany
[2]School of Computational Science and Engineering
[3]Joint Georgia Tech and Emory University Wallace H Coulter Department of Biomedical Engineering, Atlanta, GA 30332, USA
[4]Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT
Motivation:

Gene finding in eukaryotic genomes is notoriously difficult to automate. The task is to design a work flow with a minimal set of tools that would reach state-of-the-art performance across a wide range of species. GeneMark-ET is a gene prediction tool that incorporates RNA-Seq data into unsupervised training and subsequently generates *ab initio* gene predictions. AUGUSTUS is a gene finder that usually requires supervised training and uses information from RNA-Seq reads in the prediction step. Complementary strengths of GeneMark-ET and AUGUSTUS provided motivation for designing a new combined tool for automatic gene prediction.

Results: We present BRAKER1, a pipeline for unsupervised RNA-Seq-based genome annotation that combines the advantages of GeneMark-ET and AUGUSTUS. As input, BRAKER1 requires a genome assembly file and a file in bam-format with spliced alignments of RNA-Seq reads to the genome. First, GeneMark-ET performs iterative training and generates initial gene structures. Second, AUGUSTUS uses predicted genes for training and then integrates RNA-Seq read information into final gene predictions. In our experiments, we observed that BRAKER1 was more accurate than MAKER2 when it is using RNA-Seq as sole source for training and prediction. BRAKER1 does not require pre-trained parameters or a separate expert-prepared training step.

Availability: BRAKER1 is available for download at http://bioinf.uni-greifswald.de/bioinf/downloads/ and http://exon.gatech.edu/.

Contact: katharina.hoff@uni-greifswald.de & borodovsky@gatech.edu

1 INTRODUCTION

Transcriptome sequencing data, RNA-Seq reads, aligned to a genome sequence have great potential to improve the accuracy of structural genome annotation: spliced alignments may indicate intron positions, and coverage increase and decrease along genomic sequence may indicate locations of exon-noncoding region borders. Nevertheless, RNA-Seq coverage alone is no reliable indicator of protein coding regions (Hoff and Stanke, 2015).

The prediction of protein coding regions in genomes is often accomplished by tools that use statistical models. Some gene prediction tools can additionally use RNA-Seq to improve prediction accuracy.

Statistical models used in gene prediction usually require a training step to identify species specific parameters. For many tools, including AUGUSTUS, the training has to be performed on a set of example genes. In the past, training sets have often been produced with help of expressed sequence tags (ESTs) or protein data, and sometimes have been subject to validation by experts. With the rapidly growing number of novel sequenced genomes, this approach becomes infeasible. Fast and fully automated training methods, ideally using the nowadays often available RNA-Seq data, can provide significant advantages.

In principle, RNA-Seq reads can be assembled into longer contigs; such contigs can be used similarly to EST data both in training of gene finders and in the prediction step. One of the tools that follow this idea is the MAKER2 pipeline (Holt and Yandell, 2011). However, the RNA-Seq Genome Annotation Assessment Project (RGASP) (Steijger *et al.*, 2013) has shown that transcriptome assembly is prone to errors. To avoid transferring assembly errors into gene prediction, it is advantageous to use the transcript information contained in unassembled mapped reads.

We have developed BRAKER1, a pipeline that combines the complementary strengths of two gene prediction tools: GeneMark-ET (Lomsadze *et al.*, 2014) incorporates unassembled RNA-Seq reads into unsupervised training and subsequently generates *ab initio* gene predictions. A subset of genes predicted by GeneMark-ET are used to train AUGUSTUS (Stanke *et al.*, 2008). AUGUSTUS lacks an unsupervised training procedure and requires a good training set. Additionally, AUGUSTUS incorporates information derived from mapped unassembled RNA-Seq reads into the prediction step; in RGASP, AUGUSTUS was one of the most accurate tools for predicting protein coding genes with RNA-Seq

*to whom correspondence should be addressed support. We report accuracy results for BRAKER1 on four model organisms and compare to the accuracy of MAKER2. Recently, Testa et al. published CodingQuarry, a pipeline for RNA-Seq assembly supported training and gene prediction, but recommend its application only to fungi (Testa *et al.*, 2015). Therefore, we include CodingQuarry into the comparison on *Schizosaccharomyces pombe*.

2 METHODS

BRAKER1 is implemented in Perl and requires two input files: an RNA-Seq alignment file in `bam`-format, and a corresponding genome file in `fasta`-format. Spliced alignment information is extracted from the RNA-Seq file and stored in `GFF`-format. GeneMark-ET uses the genome file and the spliced alignment `GFF`-file for RNA-Seq supported unsupervised training. After training, GeneMark-ET creates an *ab initio* gene set. Those gene structures that have support by RNA-Seq alignments in all introns are selected for automated training of AUGUSTUS. After training, AUGUSTUS predicts genes in the input genome file using spliced alignment information from RNA-Seq as extrinsic evidence. The pipeline is illustrated in Suppl. Figure 2.1.

In order to access prediction accuracy, nuclear genomes, reference annotations and RNA-Seq libraries for four model organisms *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Schizosaccharomyces pombe* were retrieved from databases specified in Suppl. Materials.

Presence of repetitive sequences and mobile elements (transposable elements, TEs) is a characteristic feature of eukaryotic genomes. Repetitive sequences create challenges for automatic gene finders both at parameter estimation step and gene prediction step. The size and quality of the training set generated by GeneMark-ET for AUGUSTUS (multi-exon genes with so called anchored introns, the introns predicted ab initio and also supported by RNA-Seq read mapping) is not significantly affected by TEs masking since TEs have no anchored introns. However, at the prediction step TEs can corrupt gene prediction. For this reason, soft masking of genomic sequence is recommended before execution of BRAKER1. In this publication we used RepeatModeler to generate repeat library and RepeatMasker to mask sequence [Smit, A.F.A. and Hubley, R. (2008-2015) RepeatModeler Open-1.0, `http://www.repeatmasker.org/`].

3 RESULTS AND DISCUSSION

When comparing BRAKER1 to MAKER2 (details on the MAKER2 run are described in Suppl. Materials), BRAKER1 gains on average ∼15 percent points in accuracy on gene level [1] (see Table 1). We should remind that in these runs of BRAKER1 and MAKER2 we use only RNA-Seq information as the source of external evidence.

Notably, Reid *et al.* (2014) developed a pipeline, SnowyOwl, with GeneMark-ES (Ter-Hovhannisyan *et al.*, 2008) and AUGUSTUS to predict genes in fungal genomes. SnowyOwl attempts to improve prediction accuracy by selecting a gene variant with the highest homology score from a set of predicted gene variants in the same locus. This pipeline requires protein database information as additional external resource. We did not include SnowyOwl into comparisons since it cannot work without protein information.

Yet another recently developed automatic pipeline for fungal genome annotation utilizing RNA-Seq data is CodingQuarry (Testa *et al.*, 2015). Tests of CodingQuarry on the *S. pombe* genome demonstrated that it makes an improvement in comparison to MAKER2, however, BRAKER1 is on average ∼4% more accurate on gene level than CodingQuarry (Table 1).

In attempt to elucidate the roles and contributions of separate gene finding tools in BRAKER1 and MAKER2 as well as the role of repeat masking and incorporation of RNA-Seq information, we show the following results: values of *ab initio* accuracies of GeneMark-ET and AUGUSTUS for repeat masked and unmasked genomes are provided in Tables 1.1 and 1.2, respectively. These two tables show the BRAKER1 accuracies as well.

Given that BRAKER1 uses AUGUSTUS trained on genes most reliably predicted by GeneMark-ET, and since AUGUSTUS incorporates RNA-Seq into the prediction step, we expect to see an increase in accuracy when comparing BRAKER1 (the "hints supported" AUGUSTUS) with GeneMark-ET and with *ab initio* AUGUSTUS. This is the case for *A. thaliana*, *C. elegans* and *D. melanogaster*; on the fungus *S. pombe*, GeneMark-ET shows even higher accuracy than the current formal output of BRAKER1 (see Suppl. Tables 1.1 and 1.2).

Repeat masking on genome scale is an optional pre-processing step for running BRAKER1; still, taking this step does not significantly affect prediction accuracy (Suppl. Table 1.2).

To quantify the accuracy that MAKER2 gains by combining predictions from SNAP, AUGUSTUS and GeneMark-ES, from masking and from RNA-Seq information, we show the *ab initio* accuracies of the three gene-finders on unmasked genomes (Suppl. Table 1.3). These results show that the unsupervised training of GeneMark-ES allows to get accuracy close to or even better (*S. pombe*) than the one achieved by the MAKER2 training with utilization of RNA-Seq information.

Interestingly, we have observed (Suppl. Table 1.4) that the prediction accuracy of *ab initio* AUGUSTUS fully automatically trained by BRAKER1 is in most cases few percent lower than the *ab initio* accuracy of AUGUSTUS utilizing the packaged parameter files (obtained by supervised training). However, after adding RNA-Seq information, prediction accuracy of BRAKER1 (Suppl. Table 1.2) clearly exceeds accuracy of *ab initio* predictions made by "expert trained" AUGUSTUS.

In summary, we have observed that when the transcript data (RNA-Seq) is used as a sole source of evidence, BRAKER1 predicts genes more accurately than MAKER2 and CodingQuarry. The gain of accuracy is due to i/ use of GeneMark-ET and generation of accurate training sets for AUGUSTUS as well as ii/ use of hints originated from mapping of RNA-Seq reads that AUGUSTUS incorporates in the final gene prediction step.

In contrast to running MAKER2, running BRAKER1 is a "one step process", meaning that after starting it once, it will execute training and prediction in fully automated mode without manual command execution.

The example running time of BRAKER1 is ∼17.5 hours on a single CPU for training and prediction on *D. melanogaster*; running time can be improved by use of parallel processors.

ACKNOWLEDGEMENT

We would like to thank Mark Yandell and Carson Holt for valuable advice on running MAKER2.

[1] A *gene* may have several *transcripts* both in the reference and in the predicted gene set. When computing *transcript* level accuracy, each transcript variant is counted as a TP/FP/FN on its own. When computing *gene* level accuracy, a predicted gene is counted as TP if at least one of the predicted gene's transcripts matches correctly a reference transcript. For details, see documentation of the EVAL package (Keibler and Brent, , 2003)

Table 1. Gene prediction accuracy of of BRAKER1 and MAKER2 (both pipelines used repeat masking) as assessed by comparison with annotation of the genomes of four model organisms. In all cases, RNA-Seq was the only source of extrinsic evidence. For the fungus *S. pombe*, we also assessed the accuracy of gene predictions made by CodingQuarry.

	Arabidopsis thaliana		*Caenorhabditis elegans*		*Drosophila melanogaster*		*Schizosaccharomyces pombe*		
	BRAKER1	MAKER2	BRAKER1	MAKER2	BRAKER1	MAKER2	BRAKER1	MAKER2	CodingQuarry
Gene sensitivity	**64.4**	51.3	**55.0**	41.0	**67.6**	58.0	77.4	42.8	**79.7**
Gene specificity	52.0	**52.5**	**55.2**	30.8	**61.1**	47.9	**80.5**	68.7	72.6
Transcript sensitivity	**55.0**	43.5	**43.0**	31.3	**50.2**	42.3	77.4	42.8	**79.7**
Transcript specificity	50.9	**52.5**	**53.2**	30.8	**59.9**	47.9	**76.5**	68.7	72.6
Exon sensitivity	**82.9**	76.1	**80.2**	69.4	**73.3**	64.9	**83.2**	50.1	79.6
Exon specificity	**79.0**	76.1	**85.3**	62.3	**67.3**	55.0	**83.2**	71.4	81.7

Funding: This work is supported in part by the US National Institutes of Health grant HG000783 to MB.

REFERENCES

Hoff, K.J. and Stanke, M. (2015) Current Methods for Automated Annotation of Protein-Coding Genes, *Current Opinion in Insect Science*, doi:10.1016/j.cois.2015.02.008.

Holt, C. and Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects, *BMC Bioinformatics*, **12**:491.

Keibler, E. and Brent, M.R. (2003) Eval: a software package for analysis of genome annotations, *BMC Bioinformatics* **4**:50.

Lomsadze, A. and Burns, P.D. and Borodovsky, M. (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm, *Nucleic Acids Research*, doi:10.1093/nar/gku557.

Reid, I. and O'Toole, N. and Zabaney, O. and Nourzadesh, R. and Dahdouli, M. and Abdellateef, M. and Gordon, P.M.K. and Soh, J. and Butler, G. and Sensen, C. W. and Tsang, A. (2014) SnowyOwl: accurate prediction of fungal genes by using RNA-Seq and homology information to select among ab initio models, *BMC Bioinformatics* **15**:229.

Stanke, M. and Diekhans, M. and Baertsch, R. and Haussler, D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding, *Bioinformatics*, **24**(5), 637.

Steijger,T. and Abril,J.F. and Engström,P.G. and Kokocinski,F. and The RGASP Consortium, Hubard,T.J. and Guigo,R. and Harrow, J. and Bertone, P. (2013) Assessment of transcript reconstruction methods for RNA-seq, *Nature Methods*, doi:10.1038/nmeth.271.

Ter-Hovhannisyan, V. and Lomsadze, A. and Chernoff, Y. and Borodovsky, M. (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training, *Genome Research*, **18**(12):1979-90.

Testa, A.C. and Hane, J.K. and Ellwood, S.R. and Oliver R.P. (2015) CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts, *BMC Genomics* **16**:170.