Doc-Start BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1–2

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

Katharina J. Hoff[1,*], Simone Lange[1], Alexandre Lomsadze[2], Mark Borodovsky[2,3,4,5] 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]Center for Bioinformatics and Computational Genomics, Georgia Institute of Technology, Atlanta, GA 30332, USA
[4]Department of Biological and Medical Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia
[5]Joint Georgia Tech and Emory University Wallace H Coulter Department of Biomedical Engineering, Atlanta, GA 30332, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT
Motivation:
GeneMark-ET is a gene prediction tool that incorporates unassembled RNA-Seq reads into unsupervised training and subsequently generates *ab initio* gene predictions. AUGUSTUS is a gene finder that usually requires supervised training and uses information form unassembled RNA-Seq reads in the prediction step.
Results: We present BRAKER1, a pipeline for unsupervised RNA-Seq-based genome annotation that combines the advantages of GeneMark-ET and AUGUSTUS. BRAKER1 requires an RNA-Seq read alignment file and a genome file as input. First, GeneMark-ET performs iternative 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 training step.
Availability: BRAKER1 is available for download at `http://bioinf.uni-greifswald.de/downloads/` and `http://exon.gatech.edu/.`.
Contact: katharina.hoff@uni-greifswald.de

1 INTRODUCTION

Transcriptome sequencing data (RNA-Seq) that has been aligned to a genome sequence has great potential to improve the accuracy of structural genome annotation: spliced alignments may indicate intron positions, and coverage increase and decrease may give information about exon-noncoding region borders. Nevertheless, RNA-Seq alone does not indicate the presence or absence of protein coding regions.

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

The statistical models of gene prediction tools usually require a training step to adjust parameters to the genomic properties of individual species. For many tools, including AUGUSTUS, the training step has to be performed on an initially existing example gene set. Training gene sets have in the past often been compiled on the basis of expressed sequence tags (ESTs) or protein data, and have been subject to validation by experts. With the rapidly growing number or novel sequences genomes, this becomes infeasible. Fast and fully automated methods for training gene prediction tools, ideally using the nowadays often available RNA-Seq data, are urgently needed.

In principle, RNA-Seq reads can be assembled into longer contigs, and such contigs can be used similarly to EST data in training of gene finders and for 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 therefore advantagous to use the information from unassembled mapped reads.

We have developed BRAKER1, a pipeline that combines the advantages of two gene prediction tools: GeneMark-ET [Lomsadze *et al.*, 2014] is a gene prediction tool that incorporates unassembled RNA-Seq reads into unsupervised training and subsequently generates *ab initio* gene predictions. Genes predicted by GeneMark-ET are subsequently used to train AUGUSTUS [Stanke *et al.*, 2008]. AUGUSTUS is a gene prediction tool that lacks an unsupervised training procedure but incorporates unassembled RNA-Seq reads into the prediction step; AUGUSTUS was one of the most accurate

[*] to whom correspondence should be addressed

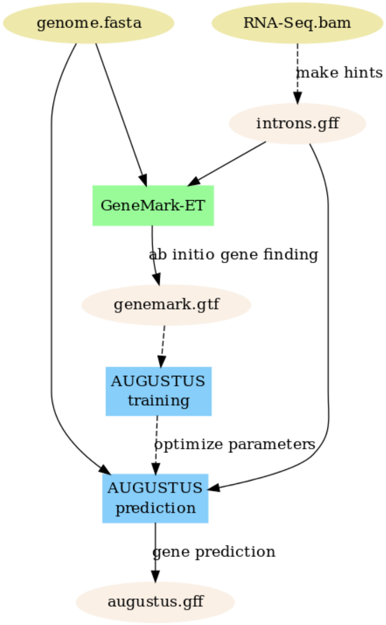

Fig. 1. Schematic view of the BRAKER1 pipeline.

tools for predicting protein coding genes with RNA-Seq in RGASP. We report accuracy results for BRAKER1 on four model organisms and compare to the accuracy of MAKER2. Recently, CodingQuarry [Testa *et al*., 2015], a pipeline particularly for RNA-Seq assembly supported training and gene prediction in fungi was published. We compare the results of BRAKER1 on *Schizosaccharomyces pombe* to CodingQuarry.

2 PIPELINE DESCRIPTION

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 intput genome file using spliced alignment information from RNA-Seq as extrinsic evidence. The pipeline is illustrated in figure 1.

3 TEST DATA

In order to demonstrate prediction accuracy, genomes, reference annotations and RNA-Seq libraries were retrieved for four model organisms from the respective databases: for *Arabidopsis thaliana*, TAIR 10 was downloaded from http://arabidopsis.org/; for *Caenorhabditis elegans*, WS240 was downloaded from http://www.wormbase.org/; for *Drosophila melanogaster*, R5 was downloaded from http://flybase.org/; for *Schizosaccharomyces pombe*, ASM294v2.23 was downloaded from http://www.pombase.org/. The following RNA-Seq libraries were retrieved from the short read archive at NCBI: SRR934391 (for *A. thaliana*); SRR065719 (for *C. elegans*); SRR023505, SRR023546, SRR023608, SRR026433, SRR027108 (for *D. melanogaster*); SRR097898, SRR097899, SRR097900, SRR097902, SRR097903, SRR097905, SRR097906, SRR097907, SRR097908, SRR097909, SRR097912, SRR097915, SRR097917, SRR097921, SRR097922, SRR097925, SRR402833 (for *S. pombe*).

4 RESULTS AND DISCUSSION

Since BRAKER1 uses AUGUSTUS being trained on GeneMark-ET, and since AUGUSTUS incorporates RNA-Seq into the prediction step, we expect to see an increase in accuracy when comparing AUGUSTUS to GeneMark-ET. This is the case for *A. thaliana*, *C. elegans* and *D. melanogaster* (see table 1). On the fungus *S. pombe*, GeneMark-ET is superior to AUGUSTUS in the current version of BRAKER1. This is due to the fact that compared to all other test species, the fungus has fewer introns and many intron-less genes. AUGUSTUS parameters can be adapted to improve prediction accuracy in such species, but we here report results with the BRAKER1 default settings. Table 1 also shows that Transcript sensitivity and specificity of AUGUSTUS is higher because AUGUSTUS reports alternative tanscripts that are supported by RNA-Seq data.

For the comparison to MAKER2, we followed the tutorial for training and prediction at

5 CONCLUSION

ACKNOWLEDGEMENT

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

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

REFERENCES

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.

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.

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.

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

esta, 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.

Table 1. Accuracy results of BRAKER1 and MAKER2 in genomes of four model organisms. For BRAKER1, accuracy is shown for the GeneMark-ET *ab initio* predictions as well as for the AUGUSTUS predictions with hints from RNA-Seq.

	Arabidopsis thaliana			*Caenorhabditis elegans*			*Drosophila melanogaster*			*Schizosaccharomyces pombe*			
	BRAKER1-GeneMarkET	BRAKER1-AUGUSTUS	MAKER2	BRAKER1-GeneMarkET	BRAKER1-AUGUSTUS	MAKER2	BRAKER1-GeneMarkET	BRAKER1-AUGUSTUS	MAKER2	BRAKER1-GeneMarkET	BRAKER1-AUGUSTUS	MAKER2	CodingQuarry
Gene sensitivity	53.9	63.2	51.3	43.0	55.1	41.0	58.5	70.2	58.0	80.0	77.3	42.7	79.7
Gene specificity	46.1	51.3	52.5	41.7	56.1	30.8	49.9	59.0	46.9	84.9	81.2	68.6	72.6
Transcript sensitivity	45.4	53.9	43.5	32.9	43.2	31.3	42.3	52.0	42.3	80.0	77.3	42.7	79.7
Transcript specificity	46.1	50.0	52.5	41.7	54.0	30.8	49.9	57.8	47.9	84.9	77.4	68.6	72.6
Exon sensitivity	81.1	83.0	76.1	79.9	80.9	69.4	68.5	75.1	64.9	85.2	84.2	50.1	79.6
Exon specificity	72.4	78.5	76.1	78.2	85.4	62.3	57.9	66.2	55.0	89.0	82.6	71.4	81.7