Many genome sequencing projects are accompanied by transcriptome sequencing. The resulting RNA-Seq data is often assembled to aid structural genome annotation. However, as the RGASP [1] competition has shown the RNA-Seq assemblies contain errors and, as a result, the training of RNA-Seq-based gene finders can be involved and the prediction of protein-coding genes is still error prone. Therefore, there is a clear need for new easily applicable and accurate methods. Recently developed GeneMark-ET [2] is a gene prediction tool that incorporates unassembled RNA-Seq reads into unsupervised training and subsequently generates gene predictions as an ab initio gene prediction tool. AUGUSTUS [3] is a gene finder that usually requires supervised training; according to the RGASP results AUGUSTUS was one of the most accurate gene finders that uses RNA-Seq read information as extrinsic evidence in the prediction step. We saw a good potential in bypassing the RNA-Seq assembly step and developing a new method that would use mapped to genome RNA-Seq reads both in unsupervised automatic training and in gene prediction.

Here, 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 (in bam format) and a genome file as input. 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 as extrinsic evidence into final gene predictions. In our experiments we observed that BRAKER1 was more accurate than MAKER2 when it is using assembled RNA-Seq as sole source of extrinsic evidence. BRAKER1 does not require pre-trained parameters or a separate manually curated training step. BRAKER1 is available for download at http://bioinf.uni-greifswald.de/augustus/downloads/index.php and http://exon.gatech.edu/.

References:


Funding

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