BRAKER2: Incorporating Protein Homology Information into Gene Prediction with GeneMark-EP and AUGUSTUS

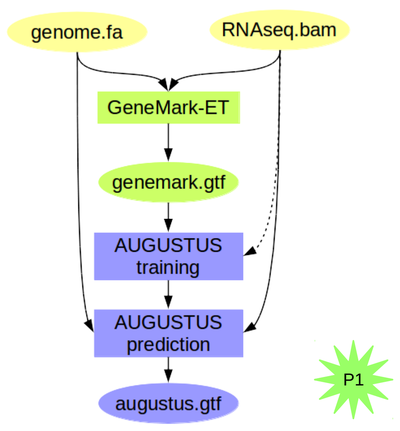
Katharina J. Hoff[1], Alexandre Lomsadze[2], Mario Stanke[1] and Mark Borodovsky[2]

1) Institute for Mathematics and Computer Science, University of Greifswald, Greifswald, GERMANY
2) Joint Georgia Tech and Emory Wallace H Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, U.S.A.
Contact: katharina.hoff@uni-greifswald.de

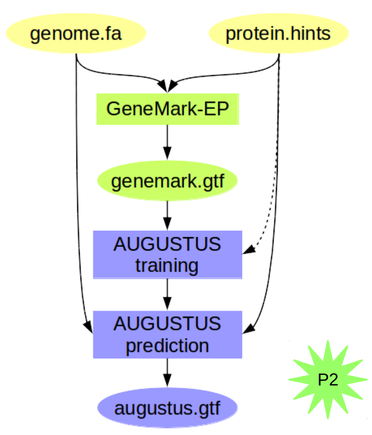
Abstract

The rapidly growing number of sequenced genomes requires fully automated methods for accurate gene structure annotation. With this goal in mind, we have developed BRAKER1 [1], a combination of GeneMark-ET [2] and AUGUSTUS [3], that uses genomic and RNAseq data to automatically generate full gene structure annotations in novel genomes. BRAKER2 is an extension of this earlier developed tool which allows for the integration of protein homology information.
If no annotation of a closely related species is available, a database of protein clusters (that must not contain proteins of very closely related species) can be used instead. In this case, BRAKER2 can execute self-training GeneMark-EP supported by protein alignments generated with ProSplign [4], trains AUGUSTUS on the basis of GeneMark-EP predictions and finds genes with protein homology information with AUGUSTUS.
In presence of the proteome of a very closely related species, BRAKER2 can align the proteome with GenomeThreader [6], trains AUGUSTUS on the basis of spliced alignments and predicts genes with AUGUSTUS using protein homology information.
BRAKER2 is available for download at http://bioinf.uni-greifswald.de/bioinf/braker and http://exon.gatech.edu.

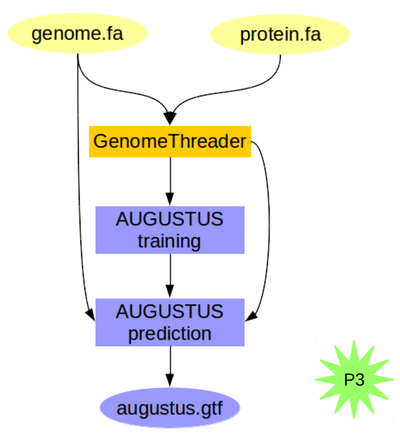
RNA-Seq



Proteins (remote)



Proteins (close)



Conclusions

- **BRAKER2** trained on protein data employing **GeneMark-EP (P2)** achieves high accuracy even in absence of close relatives in the protein database.
- Spliced alignments of proteins against genome with GenomeThreader loose sensivity with increasing distance between target and reference species.
- If the proteome of a very closely related species is available, **BRAKER2 (P3)** trained on spliced alignments from **GenomeThreader** is superior to **BRAKER1 (P1)** and **BRAKER2 (P2)** with **GeneMark-EP**.

Results

Table: Accuracy of BRAKER in *D. melanogaster*. Expert trained AUGUSTUS (AUG) *ab initio*, self-trained GeneMark-ES (GM-ES) *ab initio*, BRAKER1 GeneMark-ET (GM-ET) and AUGUSTUS, BRAKER2 with inNog except for all *Drosophila* species GeneMark-EP (GM-EP) and AUGUSTUS, BRAKER2 with inNog with *D. grimshawi, D. virilis* and *D. willistoni* and without all other *Drosohpila species* (w_gvw), BRAKER2 with inNog with *D. grimshawi, D. virilis, D. willistoni, D. pseudoobscura* and *D. ananassae* and without all other *Drosohpila* species (w_gvwpa).

	ab inito	*ab initio*	**BRAKER1 (P1)** RNAseq		**BRAKER2 (P2)** noDrosophila		**BRAKER2 (P2)** w_gvw		**BRAKER2 (P2)** w_gvwpa	
	AUG	**GM-ES**	**GM-ET**	**AUG**	**GM-EP**	**AUG**	**GM-EP**	**AUG**	**GM-EP**	**AUG**
Gene Sens.	54.7	45.1	48.9	62.8	49.2	55.2	49.7	60.7	49.7	60.6
Gene Spec.	60.5	62.2	49.2	62.4	49.9	58.4	49.8	62.7	49.8	62.3
Trans. Sens.	37.7	31.4	34.0	44.7	34.3	39.7	34.5	43.1	34.6	42.9
Trans. Spec.	60.5	46.2	49.2	59.2	49.9	54.5	49.8	57.3	49.8	55.8
Exon Sens.	71.2	66.1	67.6	77.3	67.2	69.3	67.6	73.2	67.7	73.8
Exon Spec.	83.0	70.9	74.0	80.6	75.2	81.3	74.8	82.3	74.8	81.2

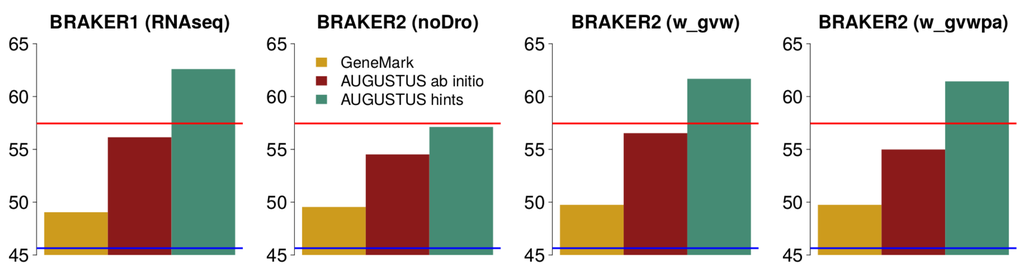
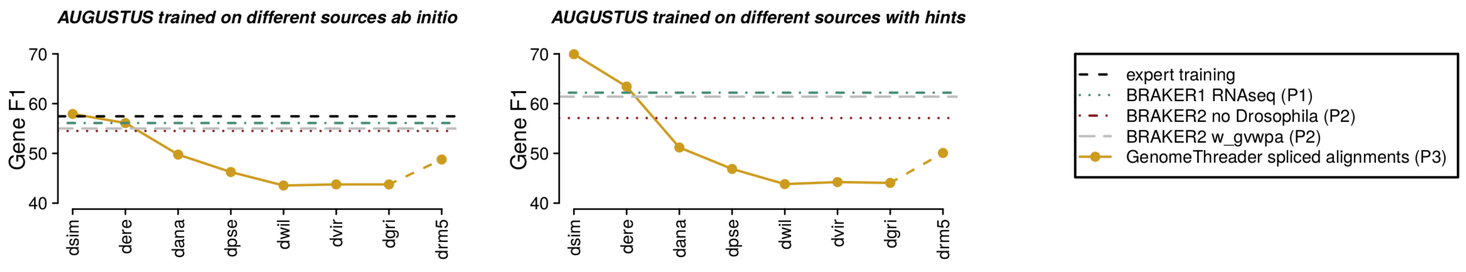


Figure 1: Gene F1 (P1/P2) of sensitivity and specificity from GeneMark *ab initio* predictions, AUGUSTUS *ab initio* predictions, AUGUSTUS predictions with hints; GeneMark-ES (blue), AUGUSTUS expert trained (red).

Figure 2: GenomeThreader trained AUGUSTUS (P3) vs. BRAKER1 with RNAseq (P1) & BRAKER2 trained on ProSplign alignments against inNog with GeneMark-EP (P2).



Data: A case study on *D. melanogaster*

- **Genome:** dmel-all-chromosome-r6.18.fasta
- **RNAseq (P1):** SRR023505, SRR023546, SRR023608, SRR026433, SRR027108 aligned against softmasked Genome with STAR [6]
- **Reference annotation**: dmel-no-analysis-r6.18.gff
- **Proteomes for running BRAKER2 (P2):** EggNog (inNog)n excluding
  I. all *Drosophila* species **(noDro)**
  II. *Drosophila* species except for *D. grimshawi, D. virilis, D. willistoni* **(w_gvw)**
  III. *Drosophila* species except for *D. grimshawi, D. virilis, D. willistoni, D. pseudoobscura, D. ananassae* **(w_gvwpa)**
  aligned with ProSplign [4].
- **Proteomes for BRAKER2 with spliced Alignments of close relatives (P3):** Flybase proteomes of

dsim *D. simulans*
dere *D. erecta*
dana *D. ana-nassae*
dpse *D. pseudo-obscura*
dwil *D. willistoni*
dvir *D. virilis*
dgri *D.grimshawi*
drm5dana, dpse, dwil, dvir, dgri

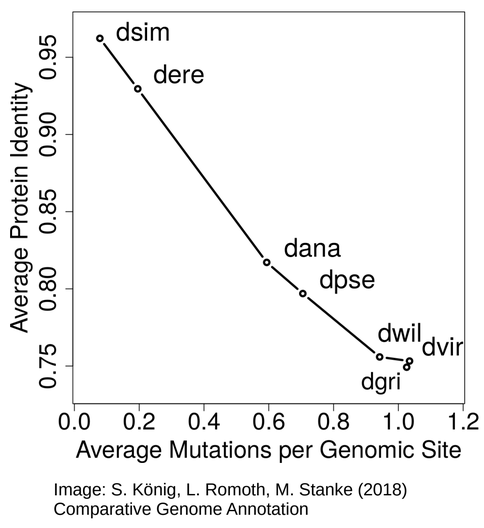


Image: S. König, L. Romoth, M. Stanke (2018) Comparative Genome Annotation

aligned with GenomeThreader.

References

[1] Hoff, Katharina J., et al. "BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS." Bioinformatics 32.5 (2015): 767-769.
[2] Lomsadze, Alexandre, Paul D. Burns, and Mark Borodovsky. "Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm." Nucleic acids research 42.15 (2014): e119-e119.
[3] Stanke, Mario, et al. "Using native and syntenically mapped cDNA alignments to improve de novo gene finding." Bioinformatics 24.5 (2008): 637-644.
[4] Kiryutin, Boris, Alexandre Souvorov, and Tatiana Tatusova. "ProSplign–Protein to Genomic Alignment Tool." In Proc. 11th Annual International Conference in Research in Computational Molecular Biology, San Francisco, USA. 2007.
[5] Gremme, Gordon. "Computational Gene Structure Prediction." (2013) PhD thesis University of Hamburg, Germany.
[6] Dobin, Alexander, et al. "STAR: ultrafast universal RNA-seq aligner." Bioinformatics 29.1 (2013): 15-21.