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

A pipeline for fully automated training and prediction

Plant and Animal Genomes XXVI, January 14th 2018

Katharina J. Hoff, Alexandre Lomsadze, Mario Stanke, Mark Borodovsky

Presenting author: katharina.hoff@uni-greifswald.de
Contents

1 Gene prediction

2 BRAKER1: RNAseq

3 BRAKER2: proteins
   Short evolutionary distance
   Long evolutionary distance

4 Summary

5 References
Structural genome annotation problem

Input

- genome assembly
- extrinsic evidence, e.g. from RNAseq, protein database

Output

- protein-coding genes: exon-intron structures (.gff)

Example (from Chr I in C. elegans)
BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS

Katharina J. Hoff, Simone Lange, Alexandre Lomsadze, Mark Borodovsky, Mario Stanke

Bioinformatics, Volume 32, Issue 5, 1 March 2016, Pages 767–769,
https://doi.org/10.1093/bioinformatics/btv661

• >4000 downloads
• 73 citations since 2016 (Google Scholar)
BRAKER1: RNAseq integration

- **BRAKER1**: RNAseq integration
- **BRAKER2**: proteins
  - Short evolutionary distance
  - Long evolutionary distance

**Summary**

**References**
BRAKER2: Part I - proteins of closely related species

Gene prediction

BRAKER1: RNAseq
BRAKER2: proteins

Short evolutionary distance
Long evolutionary distance

Summary

References
Drosophila melanogaster and relatives

For a given species,

- the average number of mutations per genomic site was computed from alignments of ortholog gene sequences (including introns).
- the protein identity was computed as average of identity values of the best exonerate hit found for each protein of this species against the *D. melanogaster* genome.
Increasing evolutionary distance leads to decreasing gene prediction accuracy of AUGUSTUS

**AUGUSTUS ab initio prediction**

Gene prediction

- BRAKER2: proteins
- BRAKER1: RNAseq
- AUGUSTUS: ab initio prediction

**Graph**

- BRAKER2 GenomeThreader training
- Expert training
- BRAKER1 RNAseq training

**Summary**

AUGUSTUS ab initio prediction

**References**
Increasing evolutionary distance leads to decreasing gene prediction accuracy of AUGUSTUS

AUGUSTUS prediction with training set hints

- BRAKER2 GenomeThreader training
- BRAKER1 RNAseq training

Gene F1

- dsim
- dere
- dana
- dpse
- dwil
- dvir
- dgri
- drm5

Short evolutionary distance
Long evolutionary distance

Summary
References
Increasing evolutionary distance leads to decreasing gene prediction accuracy of AUGUSTUS

With increasing distance between query protein and target genome, spliced alignments become

- less sensitive while keeping a constant level of specificity (e.g. GenomeThreader),
- or both less sensitive and less specific (e.g. Exonerate).

Therefore, training AUGUSTUS on spliced alignments is suitable upon availability of a very closely related query species, only!
BRAKER2: Part II - proteins of more remote species

“Standard mapping approach”: proteins to genome

- genome.fa
- proteins.fa

GenomeThreader

CDS, introns, starts, stops (protein.hints)

→ works well for closely related species, only
**BRAKER2: Part II - proteins of more remote species**

**GeneMark-EP protein mapping pipeline**

1. **genome.fa** → **GeneMark-ES** → **genemark.gtf** → **predicted proteins** → **BlastP** → **“hits”**

   - For each “hit”:
     - **predicted gene** → **nucleotide sequence (seed)** → **ProSplign** → **intron (protein.hints)**

   - **GeneMark-EP** → **genemark.gtf** → **AUGUSTUS training** → **AUGUSTUS prediction** → **augustus.gtf** → **braker.pl**

**Summary**

**References**
### Insect portion of EggNOG (inNOG) excluding *Drosophila* species

<table>
<thead>
<tr>
<th>Protein database for gene prediction in <em>D. melanogaster</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Short evolutionary distance</strong></td>
</tr>
<tr>
<td><strong>Long evolutionary distance</strong></td>
</tr>
<tr>
<td>Gene prediction</td>
</tr>
<tr>
<td>BRAKER1: RNAseq</td>
</tr>
<tr>
<td>BRAKER2: proteins</td>
</tr>
<tr>
<td><strong>Summary</strong></td>
</tr>
<tr>
<td><strong>References</strong></td>
</tr>
<tr>
<td><strong>Protein database for gene prediction in <em>D. melanogaster</em></strong></td>
</tr>
<tr>
<td><strong>Insect portion of EggNOG (inNOG) excluding <em>Drosophila</em> species</strong></td>
</tr>
<tr>
<td><strong>Acyrthosiphon pisum</strong></td>
</tr>
<tr>
<td><strong>Aedes aegypti</strong></td>
</tr>
<tr>
<td><strong>Anopheles darlingi</strong></td>
</tr>
<tr>
<td><strong>Anopheles gambiae</strong></td>
</tr>
<tr>
<td><strong>Apis mellifera</strong></td>
</tr>
<tr>
<td><strong>Atta cephalotes</strong></td>
</tr>
<tr>
<td><strong>Bombyx mori</strong></td>
</tr>
<tr>
<td><strong>Culex quinquefasciatus</strong></td>
</tr>
<tr>
<td><strong>Danaus plexippus</strong></td>
</tr>
<tr>
<td><strong>Heliconius melpomene</strong></td>
</tr>
<tr>
<td><strong>Nasonia vitripennis</strong></td>
</tr>
<tr>
<td><strong>Pediculus humanus</strong></td>
</tr>
<tr>
<td><strong>Tribolium castaneum</strong></td>
</tr>
</tbody>
</table>

Katharina J. Hoff, Alexandre Lomsadze, Mario Stanke, Mark Borodovsky
Intron recovery from protein mapping

Protein mapping with no *Drosophila* EggNOG (inNOG)

- 30,996 introns predicted
- 21,843 matched introns in CDS part of the annotated genes

Mapping of proteins from remote species recovers \(\sim 45\%\) of introns with specificity of \(\sim 70\%\).
Intron recovery from protein mapping

Protein mapping with some *Drosophila* species present as external evidence

- **no_Dro**: no *Drosophila* species
- **w_gvw**: with *D. grimshawi, D. virilis, D. willistoni*
- **w_gvwpa**: with *D. grimshawi, D. virilis, D. willistoni, D. pseudoobscura, D. ananassae*

→ more introns were detected
→ performance of protein mapping with addition of 5 fly proteomes came closer to performance with RNAseq external evidence
Accuracy of GeneMark-EX with different sources of evidence

- results are on **softmasked genome** (strongly recommended!)

![Exon prediction accuracy](chart)

<table>
<thead>
<tr>
<th></th>
<th>ES</th>
<th>EP-no_Dro</th>
<th>ET-RNAseq</th>
<th>Ideal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>60</td>
<td>65</td>
<td>70</td>
<td>75</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>70</td>
<td>75</td>
<td>80</td>
</tr>
</tbody>
</table>

![Introns in CDS prediction accuracy](chart)

<table>
<thead>
<tr>
<th></th>
<th>ES</th>
<th>EP-no_Dro</th>
<th>ET-RNAseq</th>
<th>Ideal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>60</td>
<td>65</td>
<td>70</td>
<td>75</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>70</td>
<td>75</td>
<td>80</td>
</tr>
</tbody>
</table>

- GeneMark-EP and GeneMark-ET outperformed GeneMark-ES
- GeneMark-EP with “remote” proteins was comparable with GeneMark-ET
- GeneMark-EP and GeneMark-ET were close to the best possible performance: compared to training with “ideal” introns
Accuracy of BRAKER2

Gene prediction accuracy (F1)

- BRAKER1 (RNAseq)
- BRAKER2 (no_Dro)
- BRAKER2 (w_gvw)
- BRAKER2 (w_gvwpa)

Legend:
- GeneMark
- AUGUSTUS ab initio
- AUGUSTUS with hints
- expert trained AUGUSTUS ab initio
- GeneMark-ES (ab initio and self training)
BRAKER2 is a novel fully automatic pipeline which makes gene prediction in eukaryotic genomes with RNAseq or protein external evidence.

Training in BRAKER2 is done by GeneMark-EX which particularly can use remote proteins as external evidence.

Prediction in BRAKER2 is done by AUGUSTUS using RNAseq or proteins as hints.
Ongoing & future work

- Optimization of evidence integration in BRAKER2
- Combining RNAseq and protein information
- UTR training & integration of RNAseq coverage information
References


BRAKER2 is available for download at

- http://bioinf.uni-greifswald.de
- http://exon.gatech.edu
State of the art: BRAKER with RNAseq & proteins

Close homology

Gene prediction
BRAKER1: RNAseq
BRAKER2: proteins
Short evolutionary distance
Long evolutionary distance

Summary
References
State of the art: BRAKER with RNAseq & proteins

AUGUSTUS ab initio prediction

Gene F1

BRAKER2 GenomeThreader training
BRAKER2 GenomeThreader & RNAseq training
expert training
BRAKER1 RNAseq training

Gene prediction
BRAKER1: RNAseq
BRAKER2: proteins
Short evolutionary distance
Long evolutionary distance

Summary
References
State of the art: BRAKER with RNAseq & proteins

AUGUSTUS prediction with training set hints

- BRAKER2 GenomeThreader training
- BRAKER2 GenomeThreader & RNAseq training
- BRAKER1 RNAseq training

Gene prediction
BRAKER1: RNAseq
BRAKER2: proteins
Short evolutionary distance
Long evolutionary distance

Summary
References
State of the art: BRAKER with RNAseq & proteins

Remote homology

Gene prediction
BRAKER1: RNAseq
BRAKER2: proteins
Short evolutionary distance
Long evolutionary distance

Summary
References
State of the art: BRAKER with RNAseq & proteins