Improving homology-based gene prediction using intron position conservation and RNA-seq data

Jens Keilwagen[1], Frank Hartung[1], Michael Paulini[2], and Jan Grau[3]

[1]*Institute for Biosafety in Plant Biotechnology, Julius Kühn-Institut, Quedlinburg, Germany*
[2]*EMBL-EBI, Hinxton, United Kingdom*
[3]*Institute of Computer Science, Martin Luther University Halle–Wittenberg, Germany*
jens.keilwagen@julius-kuehn.de

In the era of next generation sequencing, new genomes are sequenced and assembled rapidly. Annotations are added to those genomes mostly based on RNA-seq data and computational predictions. Homology-based approaches predict genes or transcripts in newly sequenced target genomes based on the similarity to known genes from well annotated reference genomes. Most programs including BLAST assume a conservation of the encoded amino acid sequence but do not utilize the known gene structure defined by exons and introns. However, the gene structure of intron-containing orthologous genes is highly conserved throughout the whole plant or animal kingdom and to a smaller extent even across kingdoms.

Here, we present a Gene Model Mapper approach called GeMoMa that exploits the conservation of the encoded amino acid sequence and the gene structure from a reference species to predict gene models in a target genome [1]. Instead of searching for the complete coding sequence or amino acid sequence, GeMoMa utilizes tblastn to search for the amino acid sequences encoded by the (partially) coding exons and subsequently build gene models from these hits. Recently, we extended GeMoMa allowing for utilizing RNA-seq data to define splice sites and to improve the prediction accurray.

In contrast to purely transcriptomics-based gene predictions, GeMoMa is capable of predicting lowly or specifically transcribed genes. By design, GeMoMa automatically provides information about putative homologous gene pairs and allows for transferring information about gene function.

We assess the performance of GeMoMa and compare it with state-of-the-art competitors on plant, animal, and fungi genomes. Subsequently, we predict gene models for four nematodes species and compare them with the official annotation from Wormbase proposing hundreds to thousands of high-quality new or re-annotations per species.

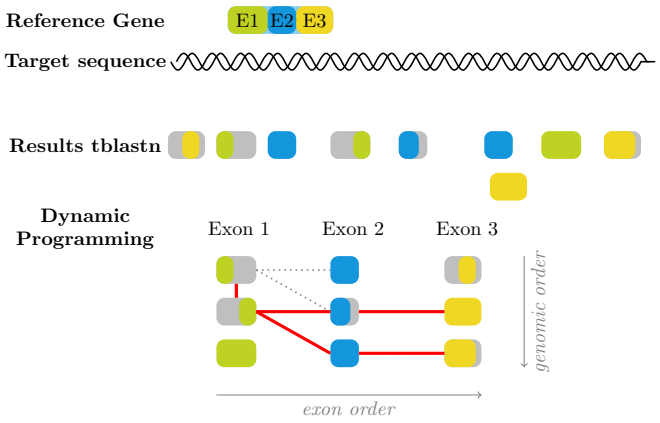

Figure 1: Illustration of the GeMoMa algorithm. GeMoMa utilizes tblastn for searching for the amino acid sequences encoded by the (partially) coding exons. Colored boxes indicate tblastn hits. Grey parts of boxes indicate partial hits. Subsequently, these results are stitched together to build a gene model in the target species using a dynamic programming approach. Thereby, the orientation and the distance of the blast hits as well as potential splice sites are used to build potential gene models indicated by lines between the boxes. Solutions with a high sum of raw tblastn scores are returned as predictions as indicated in red.

References

[1] Jens Keilwagen, Michael Wenk, Jessica L. Erickson, Martin H. Schattat, Jan Grau, and Frank Hartung. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research*, 44(9):e89, 2016.