A Fully Automatic AUGUSTUS Pipeline for Eukaryotic Genome Annotation Based on ESTs

Xin Xing, Mario Stanke

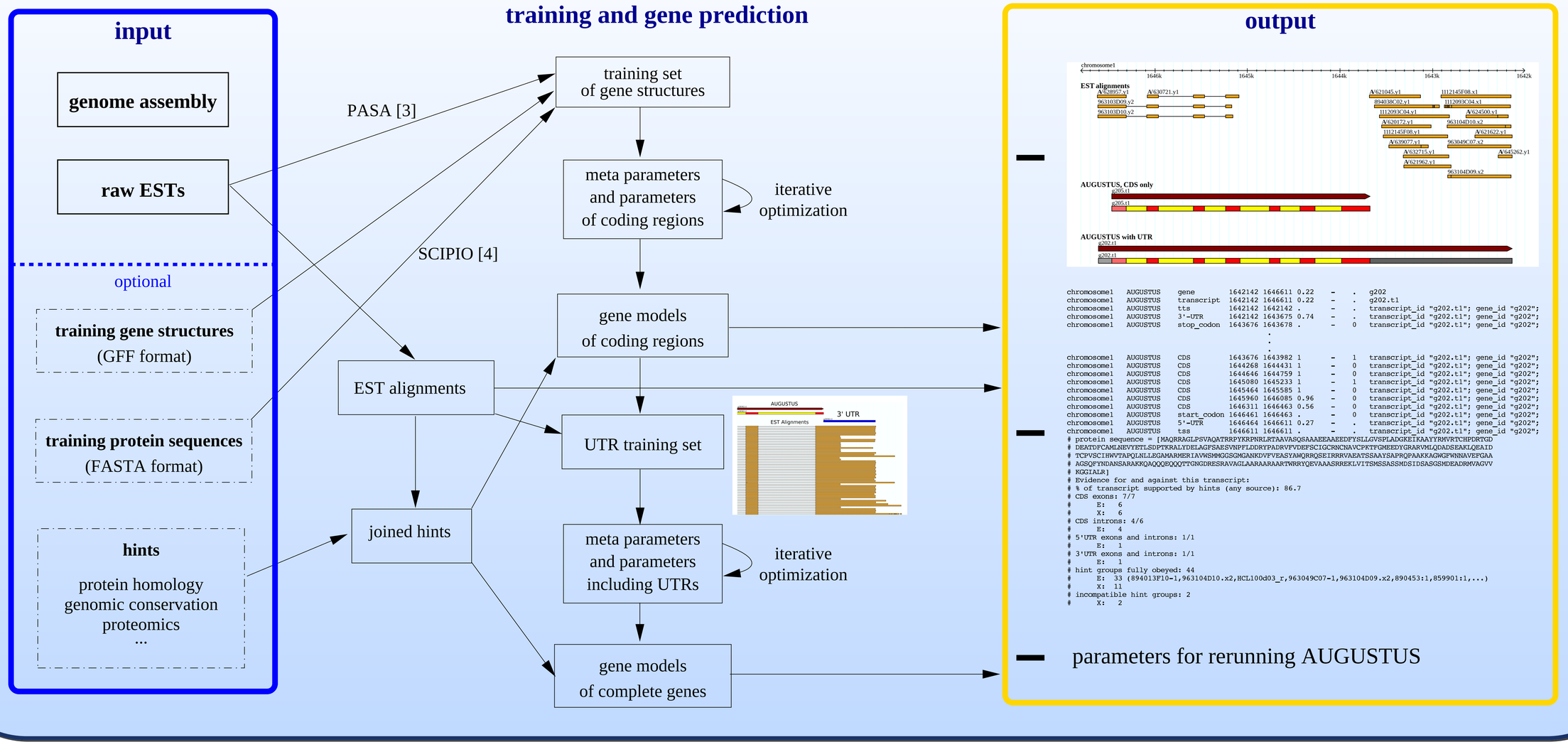
Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik, Universität Göttingen, Germany

Contact: mario@gobics.de

Introduction

Eukaryotic gene finders require for competitive performance that their parameters for detecting biological signals and for distinguishing coding and non-coding regions are reestimated for new species. We are introducing a fully automatic structural annotation pipeline that trains a gene finder solely based on the sequence data – the genome and transcript sequences – and then predicts the protein-coding genes genome wide. Its core is the gene prediction program AUGUSTUS [1] that is based on a generalized Hidden Markov Model and performed among the three best transcript-based gene finders at the most recent independent assessment of gene finders, nGASP [2].

Pipeline for Structural Annotation of New Genomes



Hints - Extrinsic Evidence

Where available, extrinsic evidence about the location and structure of protein-coding genes can be incorporated in the predictions from sources such as

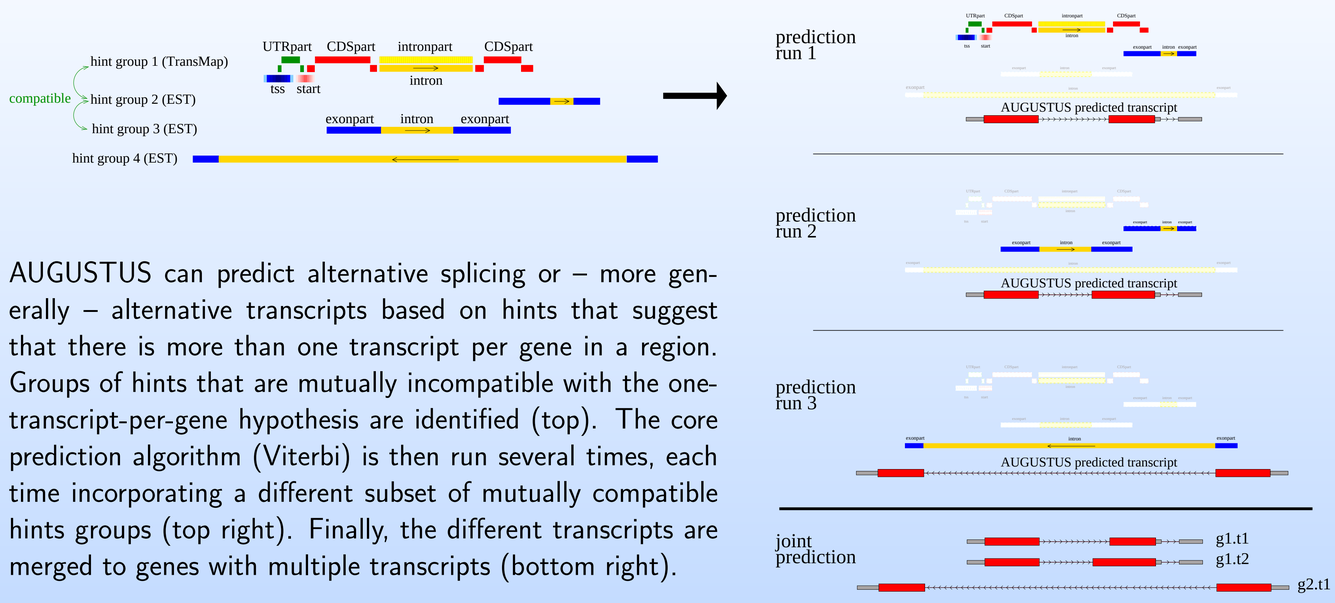
- EST/454/mRNA alignments
- genome-genome comparisons
- annotation of related species using a syntenic alignment
- protein alignments
- peptide mass spectrometry.

However, the pipeline also predicts genes in regions where extrinsic evidence is missing. A *Hint* is a piece of local extrinsic information about the gene structure. Hints are read in GFF format and can be used to locally enforce or suggest a gene structure. Hints can be given individual strengths. Different sources of hints can be given different priorities. E.g. if alien EST alignments should be considered *unless* an alignment of a native EST suggests a different transcript structure.

Some Genome Projects Using AUGUSTUS

Aedes aegypti, **Science**, 2007
Brugia malayi, **Science**, 2007
Tribolium castaneum, **Nature**, 2008
Papaya, **Nature**, 2008
Schistosoma mansoni, **Nature**, in press
Coprinus cinereus
Nasonia vitripennis
Amphimedon queenslandica
Chlamydomonas reinhardtii
Pea Aphid

Alternative Splicing



AUGUSTUS can predict alternative splicing or – more generally – alternative transcripts based on hints that suggest that there is more than one transcript per gene in a region. Groups of hints that are mutually incompatible with the one-transcript-per-gene hypothesis are identified (top). The core prediction algorithm (Viterbi) is then run several times, each time incorporating a different subset of mutually compatible hints groups (top right). Finally, the different transcripts are merged to genes with multiple transcripts (bottom right).

References

[1] Mario Stanke, Mark Diekhans, Robert Baertsch, and David Haussler. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24:5, 2008

[2] Avril Coghlan, Tristan J. Fiedler, Sheldon J. McKay, Paul Flicek, Todd W. Harris, Darin Blasiar, the nGASP Consortium, and Lincoln D. Stein. nGASP - the nematode genome annotation assessment project. *BMC Bioinformatics*, 9(549), 2008.

[3] Brian J. Haas, Arthur L. Delcher, Stephen M. Mount, Jennifer R. Wortman, Roger K. Smith, Linda I. Hannick, Rama Maiti, Catherine M. Ronning, Douglas B. Rusch, Christopher D. Town, Steven L. Salzberg and Owen White. Improving the Arabidopsis genome annotation using maximal transcipt alignment assemblies. *Nucleic Acids Res*, 31:19, 2003

[4] Oliver Keller, Florian Odronitz, Mario Stanke, Martin Kollmar, and Stephan Waack. Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics*, 9:278, 2008.