Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W465–W467***
doi:10.1093/nar/gki458

AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints

Mario Stanke* and Burkhard Morgenstern

Universität Göttingen, Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik, Goldschmidtstraße 1, 37077 Göttingen, Germany

Received February 14, 2005; Revised April 4, 2005; Accepted April 13, 2005

ABSTRACT

We present a WWW server for AUGUSTUS, a software for gene prediction in eukaryotic genomic sequences that is based on a generalized hidden Markov model, a probabilistic model of a sequence and its gene structure. The web server allows the user to impose constraints on the predicted gene structure. A constraint can specify the position of a splice site, a translation initiation site or a stop codon. Furthermore, it is possible to specify the position of known exons and intervals that are known to be exonic or intronic sequence. The number of constraints is arbitrary and constraints can be combined in order to pin down larger parts of the predicted gene structure. The result then is the most likely gene structure that complies with all given user constraints, if such a gene structure exists. The specification of constraints is useful when part of the gene structure is known, e.g. by expressed sequence tag or protein sequence alignments, or if the user wants to change the default prediction. The web interface and the downloadable stand-alone program are available free of charge at http://augustus.gobics.de/submission.

INTRODUCTION

The first step in genome annotation is to find all genes in a given genomic sequence. As experimental validation of gene structures is usually too costly, the development of gene finding programs is an important field in biological sequence analysis. For eukaryotes, this problem is far from trivial, since eukaryotic genes usually contain large introns. A large number of gene finding programs have been proposed since the 1980s (1–5). Such tools are routinely used for automatic genome annotation; despite considerable effort in the Bioinformatics community, the performance of existing gene prediction tools is still unsatisfactory.

The most reliable non-experimental method of annotation is considered to be the manual correction by experienced annotators of *ab initio* predictions in the presence of expressed sequence tag (EST) and protein alignments for the region under study. Recently, an automatic procedure has been developed for combining the diverse predictions of several *ab initio* gene finders with the EST and protein homology information to one gene structure (6).

However, despite all efforts to automate gene prediction there is still a need for tools that allow the user to decide on a part of the gene structure. Suppose, for example, that there is evidence for alternative splicing, such that an exon is included in the transcript in one splice variant but excluded in another splice variant, then the user may want to enforce the alternatively spliced exon in one prediction and enforce an intron at that position in another prediction. Another example, where constraints are useful is the case when one intron is confirmed by an RT–PCR experiment but the rest of the gene is not. In addition, a user may want to assume that a certain base is protein coding, e.g. when a single nucleotide polymorphism is correlated with the appearance of a certain phenotype.

This paper presents a web tool for accurate gene prediction under user-specified constraints. To our knowledge, the only other gene prediction server with a constraint option is that of HMMgene (1). It allows the upload of constraints similar to the ones presented here but dies when the constraints do not conform to the model. It is also only trained for human and *Caenorhabditis elegans*, which restricts its use.

METHODS

AUGUSTUS is based on a generalized hidden Markov model (GHMM), which defines probability distributions for the various sections of genomic sequences. Introns, exons, intergenic regions, etc. correspond to states in the model and each state is thought to create DNA sequences with certain pre-defined emission probabilities. Similar to other HMM-based gene finders, AUGUSTUS finds an optimal parse of a given

*To whom correspondence should be addressed. Tel: +49 551 3914926; Fax: +49 551 3914929; Email: mstanke@gwdg.de genomic sequence, i.e. a segmentation of the sequences into states that is most likely according to the underlying statistical model. We probabilistically model the sequence around the splice sites, the sequence of the branch point region, the bases before the translation start, the coding regions and non-coding regions, the first coding bases of a gene, the length distribution of single exons, initial exons, internal exons, terminal exons, intergenic regions, the distribution of the number of exons per gene and the length distribution of introns.

The performance of AUGUSTUS has been extensively evaluated on sequence data from human and Drosophila (7,8) (http://webdoc.sub.gwdg.de/diss/2004/stanke/). These studies showed that, especially for long input sequences, the accuracy of our program is superior to that of existing *ab initio* gene finding approaches. To make our tool available to the research community, we have set up a WWW server at GOBICS (Göttingen Bioinformatics Compute Server) (9).

AUGUSTUS may be forced to predict an exon, an intron, a splice site, a translation start or a translation end point at a certain position in the sequence. An arbitrary number of such constraints is allowed and supported types of constraints are given in Table 1.

With the term gene structure, we refer to a segmentation of the input sequence into any meaningful sequence of exons, introns and intergenic regions. This includes the possibility of having no genes at all or of having multiple genes. AUGUSTUS tries to predict a gene structure that

(i) is (biologically) consistent in the following sense:
  (a) No exon contains an in-frame stop codon.
  (b) The splice sites obey the gt–ag consensus. All complete genes start with atg and end with a stop codon.
  (c) Each gene ends before the next gene starts.
  (d) The lengths of single exons and introns exceed a species-dependent minimal length.
(ii) That obeys all given constraints.

Among all gene structures that are consistent and that obey all constraints, AUGUSTUS finds the most likely gene structure. A constraint may contradict the biological consistency. For example, an exonpart constraint may be impossible to realize because there is no containing open reading frame with allowed exon boundaries. If no consistent gene structure is possible, which obeys all constraints, then some constraints are ignored. Also, if two or more constraints contradict each other, then AUGUSTUS obeys only that constraint that fits better to the model. Figure 1 illustrates the concept. Further examples are on the page http://augustus.gobics.de/help.

Table 1. The types of constraints that can be imposed by the user on the predicted gene structure

Constraint type	Meaning
Start	The translation initiation site (requires an atg in the sequence)
Stop	The translation end point (requires a stop codon)
Ass	Acceptor (3′) splice site (requires ag consensus)
Dss	Donor (5′) splice site (requires gt consensus)
Exonpart	An interval or a single position that is coding i.e. it is contained in an exon
Exon	An interval that is a complete exon
Intronpart	An interval or a single position that is contained in an intron

The constraints can refer to either strand. Exon and exonpart constraints may optionally specify a reading frame.

DESCRIPTION OF THE WEB SERVER

Input

The AUGUSTUS web server allows to upload a DNA sequence in FASTA format or multiple sequences in multiple FASTA format or to paste a sequence into the web form. The maximal total length of the sequences submitted to the server is 3 million base pairs. Currently, AUGUSTUS has four species-specific parameter sets that can be chosen at the web site: *Homo sapiens*, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Brugia malayi*. Parameter sets for further species are in preparation. The optional constraints may either be uploaded as a file or entered directly into the web form in 'General Feature Format' (GFF). Examples and a detailed description are available at http://augustus.gobics.de/help.

Furthermore, there are two global options for the predicted gene structure. First, the user can restrict the predicted gene structure to contain either exactly one complete gene, or any number of complete genes, or at least one complete gene, or,

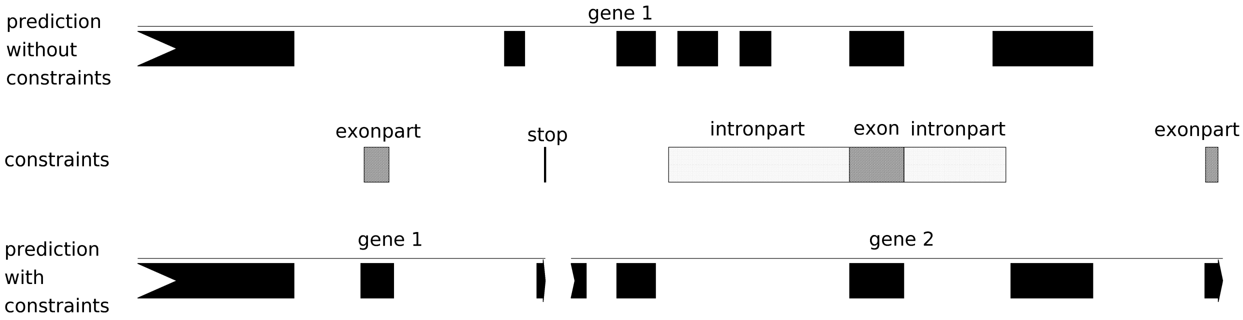

Figure 1. A contrived example for user constraints on the predicted gene structure. The top line shows the prediction of AUGUSTUS on a sequence of 5000 bp when no constraints are input. It predicted an incomplete gene with seven exons. The middle line shows six constraints: three constraints that enforce coding regions, two constraints that enforce intronic regions and one constraint that enforces the translation stop of a gene. The third line shows the prediction of AUGUSTUS under these constraints. This set of constraints is satisfiable and thus the prediction is consistent with all constraints.

by default, any number of genes, which may be partial at the boundaries of the sequence. Second, the user may suspend the above consistency requirement that each gene ends before the next gene starts. Then, the genes are predicted independently on both strands and genes on different strands may overlap or may be nested.

Output

The prediction consists of the protein coding parts of the genes as well as the amino acid sequences of the predicted genes. AUGUSTUS outputs its results both in graphical and in text format. The results page of the web server shows for each sequence a clickable thumbnail and links to images in pdf and postscript format. The pictures are generated with the program gff2ps (10) from the text output. The text output is in the General Feature Format proposed by Richard Durbin and David Haussler. The Sanger Institute lists at http://www.sanger.ac.uk /Software/formats/GFF a large number of tools that work with the GFF. In this format, the results contain for each exon one line with data fields separated by a TAB character. These data fields include the start and end position of the exon, a name for the sequence, a name for the gene and whether it is on the forward or reverse strand.

ACKNOWLEDGEMENT

The authors thank Oliver Keller for testing and giving helpful comments on the user interface. The Open Access publication charges for this article will be paid from the annual group budget.

Conflict of interest statement. None declared.

REFERENCES

1. Krogh,A. (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 179–186.
2. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Comput. Biol.*, **268**, 78–94.
3. Reese,M.G., Kulp,D., Tammana,H. and Haussler,D. (2000) Genie—gene finding in *Drosophila melanogaster*. *Genome Res.*, **10**, 529–538.
4. Parra,G., Enrique,B. and Guigó,R. (2000) GeneID in Drosophila. *Genome Res.*, **10**, 511–515.
5. Korf,I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, S1–S9.
6. Allen,J.E., Pertea,M. and Salzberg,S.L. (2004) Computational gene prediction using multiple sources of evidence. *Genome Res.*, **14**, 142–148.
7. Stanke,M. (2003) Gene prediction with a hidden Markov model. PhD Thesis, Universität Göttingen, Germany.
8. Stanke,M. and Waack,S. (2003) Gene prediction with a hidden Markov model and new intron submodel. *Bioinformatics*, **19** (Suppl. 2), ii215–ii225.
9. Stanke,M., Steinkamp,R., Waack,S. and Morgenstern,B. (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.*, **32**, W309–W312.
10. Abril,J.F. and Guigó,R. (2000) gff2ps: visualizing genomic annotations. *Bioinformatics*, **16**, 743–744.