DOI: 10.1093/nar/gkh386

AGenDA: gene prediction by cross-species sequence comparison

Leila Taher[1], Oliver Rinner[2,3], Saurabh Garg[1], Alexander Sczyrba[4] and Burkhard Morgenstern[5,*]

[1]International Graduate School for Bioinformatics and Genome Research, University of Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany, [2]GSF Research Center, MIPS/Institute of Bioinformatics, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany, [3]Brain Research Institute, ETH Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland, [4]Faculty of Technology, Research Group in Practical Computer Science, University of Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany and [5]Institute of Microbiology and Genetics, University of Göttingen, Goldschmidtstrasse 1, 37077 Göttingen, Germany

Received March 5, 2004; Revised and Accepted March 24, 2004

ABSTRACT

Automatic gene prediction is one of the major challenges in computational sequence analysis. Traditional approaches to gene finding rely on statistical models derived from previously known genes. By contrast, a new class of comparative methods relies on comparing genomic sequences from evolutionary related organisms to each other. These methods are based on the concept of phylogenetic footprinting: they exploit the fact that functionally important regions in genomic sequences are usually more conserved than non-functional regions. We created a WWW-based software program for homology-based gene prediction at BiBiServ (Bielefeld Bioinformatics Server). Our tool takes pairs of evolutionary related genomic sequences as input data, e.g. from human and mouse. The server runs CHAOS and DIALIGN to create an alignment of the input sequences and subsequently searches for conserved splicing signals and start/stop codons near regions of local sequence conservation. Genes are predicted based on local homology information and splice signals. The server returns predicted genes together with a graphical representation of the underlying alignment. The program is available at http://bibiserv.TechFak.Uni-Bielefeld.DE/agenda/.

INTRODUCTION

Accurate prediction of gene structures in raw genomic sequence data is a first and critical step in genome annotation. Thus, with the huge amount of data produced by sequencing projects, the development of high-quality gene-prediction tools has become a priority in bioinformatics. For eukaryotic organisms, gene prediction is particularly challenging, since protein-coding exons are usually separated by introns of variable length. Most traditional methods for gene prediction are *intrinsic* methods; they use hidden Markov models (HMMs) or other stochastic approaches describing the statistical composition of introns, exons, etc. Such models are trained with already known genes from the same or a closely related species; the most popular of these tools is *GenScan* (1).

Despite considerable efforts since the 1980s, the reliability of standard gene-finding methods remains limited. While most tools produce good results for short input sequences containing not much more than a single gene, their performance drops dramatically when applied to longer input sequences (2). In this situation, most standard methods tend to predict far too many genes. Substantial progress has recently been made in the field of HMM-based gene prediction. Stanke introduced a novel model for intron length distribution that reflects the real length distribution much more accurately than standard methods do (3). The gene-prediction program AUGUSTUS (4) is based on this new model. Systematic program evaluation demonstrated that AUGUSTUS performs significantly better than other intrinsic methods (4). Nevertheless, there is a common limitation for all intrinsic approaches: they depend on statistical models derived from already known genes. As a consequence, they often have difficulties predicting genes with new or unusual features.

GENE PREDICTION BY CROSS-SPECIES SEQUENCE ALIGNMENT

With the increasing number of completely or partially sequenced genomes, an alternative approach to gene prediction has

*To whom correspondence should be addressed. Tel: +49 551 39 14628; Fax: +49 551 39 14929; Email: bmorgen@gwdg.de

been proposed. It is possible to identify genes by comparing evolutionary related genomic sequences to each other. This idea is based on the phylogenetic footprinting principle (5): during evolution, functional regions of genomic sequences tend to be more conserved than non-functional regions. Therefore, local sequence similarity usually indicates biological function. In particular, regions of strong sequence conservation often correspond to protein-coding exons (6). Several new gene-finding approaches use homology information from cross-species alignments of genomic sequences (7–15). The program AGenDA (*A*lignment-based *Gene*-*D*etection *A*lgorithm), developed by Rinner and Morgenstern, is based on pair-wise human/mouse alignments created by CHAOS (16) and DIALIGN (17). It searches for conserved splice sites around local sequence similarities in order to identify candidate exons from which complete gene models are then constructed.

THE AGenDA WWW SERVER AT BiBiServ

To make AGenDA available to the genome-research community, we developed a WWW server (18) that automatically performs the following steps:

(i) RepeatMasker (http://repeatmasker.genome.washington.edu/) masks low-complexity regions in the input sequences.

Figure 1. AGenDA submission form. The user can upload two syntenic genomic sequences in FASTA format. Parameters can be adjusted to the organisms from which the input sequences come, and a threshold value can be specified for local sequence similarities that are considered for gene prediction. Single-gene models or multiple-gene models can be considered and genes can be predicted on the Watson strand only or on both the Watson and Crick strands.

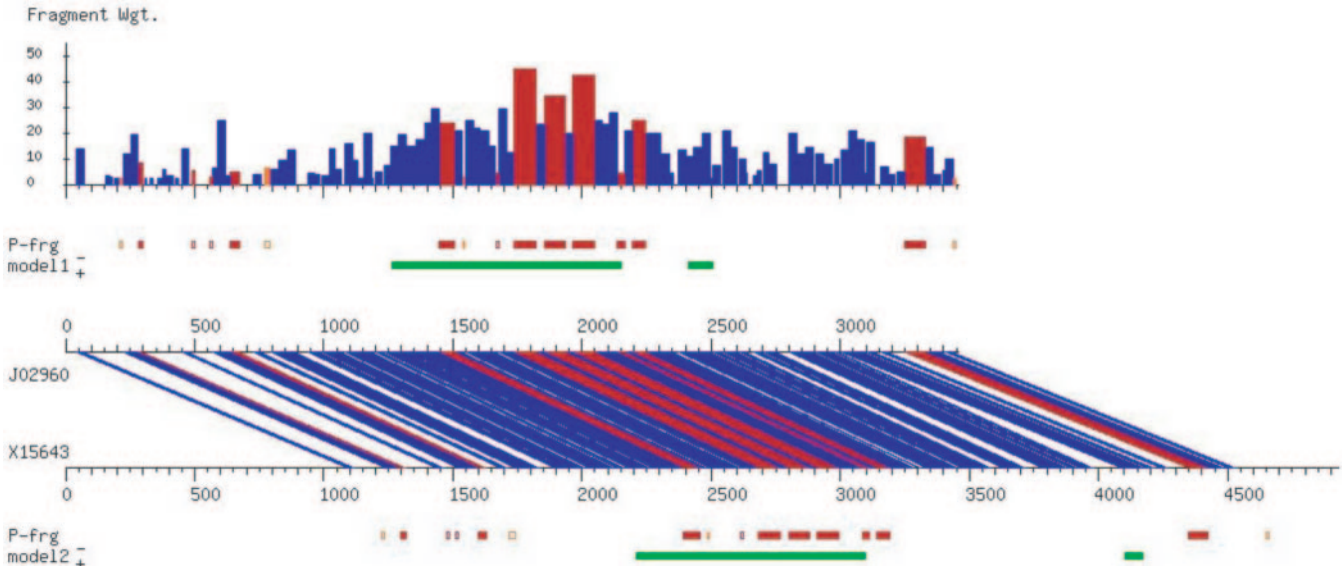

Figure 2. AGenDA program output for a pair of genomic sequences from human and mouse. Red and blue bars between the sequences represent the alignment calculated by DIALIGN using CHAOS anchors. Red bars correspond to similarities at the peptide level, while blue bars indicate similarity at the nucleotide level. Vertical bars on the top line represent the same similarities, their height representing the similarity calculated by DIALIGN ('Fragment Wgt.'). Green bars represent the gene model calculated by AGenDA.

(ii) CHAOS is used to obtain a chain of high-homology regions. Local alignments returned by CHAOS are used as anchor points to reduce the search space and running time for the subsequent alignment procedure.
(iii) DIALIGN calculates an alignment of the input sequences based on the set of anchor points created in the previous step.
(iv) AGenDA produces a list of candidate exons, using as input both the sequences and the DIALIGN output file. These candidate exons are scored based on the degree of local sequence similarity and other criteria as described in (13) (see Figure 1).
(v) An optimal gene model is built from the list of candidate exons using a fragment-chaining algorithm (19).
(vi) A graphical representation of the CHAOS/DIALIGN alignment and the identified gene models is produced.

Both DIALIGN and AGenDA have a range of parameters and options that can be set by the user as appropriate for their specific data. Our web server enables the user to adjust the following parameters (Figure 1):

(i) *Threshold value*. A threshold can be specified to impose a lower bound on the scores of the local sequence similarities returned by DIALIGN. This way, low-scoring random similarities can be filtered out. Note that this threshold does not affect the alignment procedure but is applied after the alignment has been carried out.
(ii) *Similarity level*. The user can select between two different levels of sequence similarity. For genomic sequences, DIALIGN can calculate sequence similarity at the nucleotide level by considering nucleotide matches and at the peptide level by translating DNA segments to peptide segments according to the genetic code and comparing the implied peptide segments. The user can decide if only peptide similarity or both types of similarity are used for gene prediction (Figure 2).
(iii) *Single-gene or multi-gene output*. The user can decide if only a single gene is returned or if multiple genes are allowed as output.
(iv) *DNA strand*. It is possible to select the strand for the gene prediction: genes can be searched (a) on the Watson strand only or (b) on both the Watson and Crick strands.
(v) *Input species*. The parameters used for RepeatMasker can be adjusted to different species.
(vi) *Alignment iteration steps*. For large genomic sequence data, DIALIGN can be run iteratively, such that in a first iteration step strong sequence similarities are returned while in subsequent steps additional, weaker similarities between those already identified homologies are considered. Details of this procedure are explained in (6). The user can decide if only strong similarities returned in the first iteration step are used for gene prediction or if weaker similarities from subsequent iteration steps are considered as well. Up to three iterations can be performed.

For all these options, default values are offered. These values performed best in our experience, but the user is free to use alternative parameter settings. Finally, the output is returned to the user via e-mail. The e-mail contains information about the predicted gene structure, as well as hyperlinks to three different WWW pages. These include the complete lists of candidate exons considered for gene modelling (one list for each of the two input sequences) and a graphical representation of predicted genes along with the underlying alignments.

We want to emphasize that AGenDA has been optimized for input sequences from primates and rodents. For human/mouse data, its prediction accuracy is comparable to GenScan, the most widely used gene-finding tool for vertebrates. It is, of course, possible to apply comparative gene-finding approaches to different species at varying evolutionary distances. Some authors have suggested, for example, that comparison of primates with cold-blooded vertebrates or even invertebrates

might be more suitable for gene-finding purposes (12). The user is free to submit sequences from arbitrary species to the server. In this case, however, one should keep in mind that the pre-selected default parameters have been optimized for human/mouse comparison. Thus, one cannot expect to obtain the same quality of results if these parameters are applied to other species. The user is therefore encouraged to experiment with varying parameter settings if input species other than human and mouse are submitted.

ACKNOWLEDGEMENTS

We would like to thank Michael Brudno and Serafim Batzoglou for their help with the CHAOS software and Henning Mersch and Jan Krüger for their help at BiBiServ. Mario Stanke made helpful comments on the manuscript.

REFERENCES

1. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
2. Guigó,R., Agarwal,P., Abril,J.F., Burset,M. and Fickett,J.W. (2002) An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.*, **10**, 1631–1642.
3. Stanke,M. (2004) Gene Prediction with a Hidden Markov Model. PhD Thesis, Universität Göttingen, Germany.
4. Stanke,M. and Waack,S. (2003) Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics (ECCB 2003 special issue)*, **19**, ii215–ii225.
5. Tagle,D., Koop,B., Goodman,M., Slightom,J., Hess,D. and Jones,R. (1888) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*): nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.*, **203**, 439–455.
6. Morgenstern,B., Rinner,O., Abdeddaïm,S., Haase,D., Mayer,K., Dress,A. and Mewes,H.-W. (2002) Exon discovery by genomic sequence alignment. *Bioinformatics*, **18**, 777–787.
7. Bafna,V. and Huson,D.H. (2000) The conserved exon method for gene finding. *Bioinformatics*, **16**, 190–202.
8. Batzoglou,S., Pachter,L., Mesirov,J.P., Berger,B. and Lander,E.S. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, **10**, 950–958.
9. Wiehe,T., Gebauer-Jung,S., Mitchell-Olds,T. and Guigó,R. (2001) SGP-1: Prediction and validation of homologous genes based on sequence alignments. *Genome Res.*, **11**, 1574–1583.
10. Parra,G., Agarwal,P., Abril,J.F., Wiehe,T., Fickett,J.W. and Guigó,R. (2003) Comparative gene prediction in human and mouse. *Genome Res.*, **13**, 108–117.
11. Korf,I., Flicek,P., Duan,D. and Brent,M.R. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**(Suppl. 1), S140–S148.
12. Novichkov,P.S., Gelfand,M. and Mironov,A. (2001) Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics*, **17**, 1011–1018.
13. Rinner,O. and Morgenstern,B. (2002) AGenDA: gene prediction by comparative sequence analysis. *In Silico Biol.*, **2**, 195–205.
14. Blayo,P., Rouzé,P. and Sagot,M.-F. (2003) Orphan gene finding—an exon assembly approach. *Theoret. Comput. Sci.*, **290**, 1407–1431.
15. Meyer,M. and Durbin,R. (2002) Comparative ab initio prediction of gene structures using pair HMMS. *Bioinformatics*, **18**, 1309–1318.
16. Brudno,M., Chapman,M., Göttgens,B., Batzoglou,S. and Morgenstern,B. (2003) Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, **4**, 66.
17. Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.
18. Taher,L., Rinner,O., Gargh,S., Sczyrba,A., Brudno,M., Batzoglou,S. and Morgenstern,B. (2003) AGenDA: homology-based gene prediction. *Bioinformatics*, **19**, 1575–1577.
19. Gusfield, D. (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge, UK.