BIOINFORMATICS APPLICATIONS NOTE
Vol. 19 no. 12 2003, pages 1575–1577
DOI: 10.1093/bioinformatics/btg181

AGenDA: homology-based gene prediction

Leila Taher[1,*], Oliver Rinner[2], Saurabh Garg[1],
Alexander Sczyrba[3], Michael Brudno[4], Serafim Batzoglou[4] and
Burkhard Morgenstern[1, 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]Faculty of Technology, Research Group in Practical Computer Science, University of Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany, [4]Computer Science Department, Stanford University, Stanford, CA 94305, USA and [5]University of Göttingen, Institute of Microbiology and Genetics, Goldschmidtstr. 1, 37077 Göttingen, Germany

Received on August 19, 2002; revised on November 30, 2002; accepted on February 14, 2003

ABSTRACT

Summary: We present a www server for homology-based gene prediction. The user enters a pair of evolutionary related genomic sequences, for example from human and mouse. Our software system uses CHAOS and DIALIGN to calculate an alignment of the input sequences and then searches for conserved splicing signals and start/stop codons around regions of local sequence similarity. This way, candidate exons are identified that are used, in turn, to calculate optimal gene models. The server returns the constructed gene model by email, together with a graphical representation of the underlying genomic alignment.

Availability: http://bibiserv.TechFak.Uni-Bielefeld.DE/agenda/

Contact: ltaher@TechFak.Uni-Bielefeld.DE

A primary goal of large-scale sequencing projects is to identify all genes in a given organism. Consequently, the problem of computational gene-prediction has become one of the most active areas of research in Bioinformatics, see Mathé *et al.* (2002) for a comprehensive review. Despite these efforts, the reliability of the current gene-finding methods is limited (Guigó *et al.*, 2000). With the massive genomic data that are now available, a new approach to gene prediction has been proposed: it is possible to identify genes by comparing un-characterized genomic sequences from evolutionary related species, e.g. from human and mouse, to each other (Bafna and Huson, 2000; Batzoglou *et al.*, 2000; Wiehe *et al.*, 2001; Korf *et al.*, 2001; Novichkov *et al.*, 2001; Blayo *et al.*, 2002; Meyer and Durbin, 2002). The idea behind these homology-based methods is simple: during evolution, functional elements such as genes and regulatory sites tend to be more conserved than non-functional sequences; therefore, local sequence simliarity usually indicates biological function. One problem with traditional gene-prediction approaches is that they rely heavily on information derived from already known genes of the same or a closely related species. Thus, they succeed only where such information is available, and they are unable to detect genes with different properties. By contrast, the new comparative approaches rely more on sequence conservation and less on features of previously-known genes, and therefore are more likely to identify genes with new features and different statistical composition.

Rinner and Morgenstern (2002) recently proposed a homology-based gene-finding program called AGenDA (Alignment-based Gene-Detection Algorithm). The program takes a DIALIGN alignment (Morgenstern, 1999) of two genomic sequences as input and searches for conserved splice sites around peaks of local sequence similarity. This way, *candidate exons* are identified from which a gene model is constructed. It has been shown that, for sequence data from human and mouse, the accuracy of AGenDA is comparable to GenScan (Burge and Karlin, 1997) which is generally considered the best gene-finding program for vertebrates. To make AGenDA available to the genome-research community, we developed a www server that automatically performs the following steps: First, RepeatMasker (http://repeatmasker.genome.washington.edu/) is applied to the input sequences in order to mask low-complexity regions. Next, CHAOS (Brudno and Morgenstern, 2002, http://www.stanford.edu/∼brudno/chaos/) and DIALIGN are used to compute an alignment of the sequences. Here, CHAOS identifies *anchor points* to improve the running

[*]To whom correspondence should be addressed.

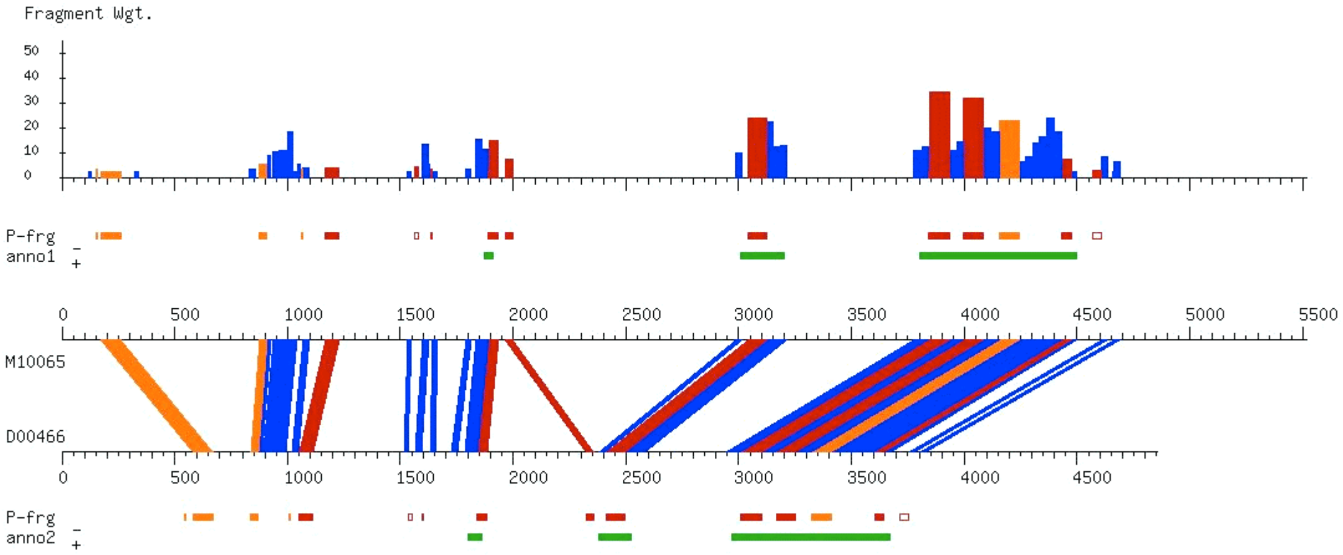

Fig. 1. Graphical program output of AGenDA representing the genomic alignment calculated by CHAOS and DIALIGN together with the identified gene model. Blue bars between the input sequences represent sequence similarity at the *nucleotide level* (N-fragments) while red and orange bars represent similarity at the *peptide level* (P-fragments) on the posititve and reverse strand, respectively. Vertical bars on the top line indicate the degree of similarity of the local sequence similarities (Fragment Wgt.), green bars show the gene model calculated by AGenDA.

time of DIALIGN. In a fourth step, AGenDA constructs a gene model based on sequence similarities found by DIALIGN. Finally, the resulting gene model is returned to the user by email, together with hyperlinks to www pages with additional information. These pages contain a complete list of the *candidate exons* considered for gene modelling as well as a graphical representation of the output gene model and the underlying alignment as shown in Figure 1.

Several parameters can be adjusted by the user: (a) a threshold value can be applied to local sequence similarities returned by DIALIGN such that only high-scoring similarities are considered for gene modelling. (b) The current version of DIALIGN applies an iterative procedure for alignment of large genomic sequences. In a first step, strong similarities are identified, in subsequent steps regions between those similarities are reconsidered and weaker similarities are added to the alignment. At the AGenDA server, it is possible to restrict the program to considering similarities that were found in the first step, i.e. to relatively strong homologies. (c) The new version of DIALIGN distinguishes between similarities at *nucleotide level* (N-fragments) and similarities at the *peptide level* (P-fragments). It is possible to exclude N-fragments and to consider only P-fragments for gene finding. In addition, options are available for (d) finding *multiple* genes in the input sequences and for (e) including genes located on the *reverse* strand. For all these options, default values are suggested that performed well in our experience, but the user is free to try out different parameter settings.

We use DIALIGN in our gene-finding approach because it has been shown that local sequence similarities detected by DIALIGN in eukaryotic genomic sequences are well correlated with protein-coding exons (Morgenstern *et al.*, 2002). In this regard, DIALIGN turned out to be superior to other alignment methods that have been tested. However, DIALIGN is slower than alternative programs for genomic alignment and, for this reason, AGenDA is currently restricted to data sets of moderate size. At present, our server accepts input sequences of up to 200 kb in length.

Finally, it should be mentioned that the results of any sequence comparison strongly depend on the evolutionary distance between the compared species. AGenDA has been optimized for input sequences from human and mouse. Novichkov *et al.* (2001) suggest that wider distances—such as between primates and cold-blooded vertebrates or invertebrates—may be more suitable for gene-prediction. It should be worthwile to apply AGenDA to other species and to explore the potential and the limitations of our method.

REFERENCES

Bafna,V. and Huson,D.H. (2000) The conserved exon method for gene finding. In *Proc. ISMB*, **8**, 3–12.

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.*, **7**, 950–958.

Blayo,P., Rouzé,P. and Sagot,M.-F. (2002) Orphan gene finding—an exon assembly approach. *Theo. Comput. Sci.*, **290**, 1407–1431.

Brudno,M. and Morgenstern,B. (2002) Fast and sensitive alignment of large genomic sequences. In *Proceedings IEEE Computer Society Bioinformatics Conference*. pp. 138–147. http://www.stanford.edu/∼brudno/chaos/

Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

Guigó,R., Agarwal,P., Abril,J.F., Burset,M. and Fickett,J.W. (2000) An assessment of gene prediction accuracy in large DNA sequences. *Genone Res.*, **10**, 1631–1642.

Korf,I., Flicek,P., Duan,D. and Brent,M.R. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**, S140–S148.

Mathé,C., Sagot,M.-F., Schiex,T. and Rouzé,P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucl. Acids. Res.*, **30**, 4103–4117.

Meyer,I.M. and Durbin,R. (2002) Comparative *ab initio* prediction of gene structures using pair HMMs. *Bioinformatics*, **18**, 1309–1318.

Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.

Morgenstern,B., Rinner,O., Abdeddaïm,S., Haase,D., Mayer,K., Dress,A. and Mewes,H.-W. (2002) Exon prediction by comparative sequence analysis. *Bioinformatics*, **18**, 777–787.

Novichkov,P.S., Gelfand,M.S. and Mironov,A.A. (2001) Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics*, **17**, 1011–1018.

Rinner,O. and Morgenstern,B. (2002) AGenDA: Gene prediction by comparative sequence analysis. *In Silico Biol.*, **2**, 195–205. http://www.bioinfo.de/isb/2002/02/0018/

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.