



Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters

Vladimir B. Bajic*, Seng Hong Seah, Allen Chong, Guanglan Zhang, Judice L. Y. Koh and Vladimir Brusic

BIC-KRDL, Kent Ridge Digital Laboratories, 21, Heng Mui Keng Terrace, Singapore 119613

Received on April 26, 2001; revised on June 28 and August 6, 2001; accepted on August 8, 2001

ABSTRACT

Summary: Dragon Promoter Finder (DPF) locates RNA polymerase II promoters in DNA sequences of vertebrates by predicting Transcription Start Site (TSS) positions. DPF's algorithm uses sensors for three functional regions (promoters, exons and introns) and an Artificial Neural Network (ANN). Results on a large and diverse evaluation set indicate that DPF exhibits a superior predicting ability for TSS location compared to three other promoter-finding programs.

Availability: <http://sdmc.krdl.org.sg:8080/promoter/>

Contact: bajicv@krdl.org.sg

INTRODUCTION

Promoters are functional regions responsible for the initiation and regulation of DNA transcription. Protein-coding genes in eukaryotes are transcribed by the RNA polymerase II. Developing promoter recognition algorithms is a challenging problem since the understanding of transcriptional processes is incomplete (Weinzierl, 1999; Pedersen *et al.*, 1999; Fickett and Hatzigeorgiou, 1997). A major deficiency of available Transcription Start Site (TSS) finding programs is the very high number of False Positive (FP) predictions for any significant level of True Positive (TP) recognition (Fickett and Hatzigeorgiou, 1997; Reese *et al.*, 2000).

The Dragon Promoter Finders (DPF) algorithm was designed to address the problem of low prediction specificity. We tested the predictive performance of DPF over a broad sensitivity range (up to 66%). The DPF was found to perform significantly better than three other promoter finding programs: NNPP2.1 (Reese *et al.*, 1996), Promoter2.0 (Knudsen, 1999), and PromoterInspector (Scherf *et al.*, 2000). These programs were selected because they are accessible through www and allow for the analysis of long sequences.

*To whom correspondence should be addressed.

ALGORITHM OUTLINE

DPFs algorithm identifies TSS positions using five independent promoter recognition models. Each model has been optimized for a different pre-defined sensitivity/specificity level. Because of the underlying non-linearity, a single model cannot provide optimal performance over a broad range of sensitivity/specificity. However, all models of the DPF algorithm have the same basic structure. To our knowledge, this is the first application of the multi-model concept for promoter recognition algorithms. Each model uses a data window that slides along the DNA sequence. Based on the competition of three sensors, DPF predicts TSS presence for each data window. The three sensors recognize the promoter, exon, and intron regions, respectively. The use of three sensors for promoter recognition is, in a way, similar to the approach reported by Levy *et al.* (1998). Sensor models are position weight matrices of pentamers and therefore, contain information on the positional distribution of pentamers. Each sensor uses only those pentamers that contribute most significantly to the separation of promoter regions from other functional regions. Sensor outputs serve as inputs to the ANN system which, in turn, predicts the presence of TSS. DPFs predictions are strand-specific.

DATASETS AND PREDICTION ACCURACY

DPF was trained on the vertebrate promoters from the Eukaryotic Promoter Database Ver. 65 (Périer *et al.*, 2000), and on randomly selected 800 human exon and 4000 human intron sequences from GenBank Rel. 121 (Benson *et al.*, 2000). A comprehensive evaluation set was compiled from human and viral sequences used by other researchers in training gene-finding and analysis programs (Genie, Reese *et al.*, 1999; GENESCAN, Burge and Karlin, 1997; NetGene, Brunak *et al.*, 1991) and sequences used in testing promoter recognition programs (Mache *et al.*, 1996; Reese and Eeckman, 1999). The

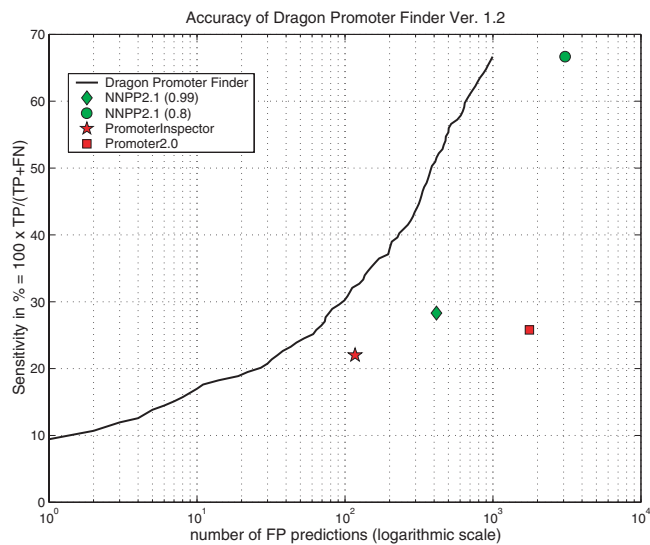


Fig. 1. Comparison results of the four promoter recognition programs achieved on the evaluation set.

Table 1. Preliminary results of the promoter prediction on human chromosome 22

Setting	DPF Ver. 1.2		Promoter Inspector	
	Chr22 known genes	Evaluation set	Chr22 known genes	Evaluation set
Very high accuracy	Se 20.06% FP/TP 0.75	22% 1	45% 1.975	22% 3.3429
High accuracy	Se 30.67% FP/TP 2.798	30% 2.06	Sensitivity in %	
Medium accuracy	Se 60.177% FP/TP 3.5637	40% 3.64	Se = 100 × TP/(TP + FN)	

training and evaluation sets of DPF were mutually exclusive. The evaluation set had a significant diversity (159 TSS) and a considerable cumulative sequence length (1.15 Mbp).

The accuracy of DPF, assessed on the evaluation set, appeared superior to that of the other three promoter recognition systems (Figure 1). Results of further testing with DPF on the annotated known genes of human chromosome 22, Rel. 2.3 (<http://www.sanger.ac.uk/HGP/Chr22/>) was consistent with those obtained on the evaluation set (Table 1). Surprisingly, results obtained by PromoterInspector for these two data sets did not demonstrate such consistency (Scherf *et al.*, 2001). The FP/TP measure used in Table 1 is explained at the DPF web site.

CONCLUSION

Since DPF can analyze over 11 000 bp per second (PIII 833 MHz laptop running Linux 7.1) it can be used easily for promoter search in large contigs of anonymous DNA. Further details are available online at the DPF web site.

REFERENCES

- Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.
- Burge,C. and Karlin,S. (1997) Prediction of complete gene structure in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
- Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, **220**, 49–65.
- Fickett,J.W. and Hatzigeorgiou,A.G. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.
- Knudsen,S. (1999) Promoter 2.0: for the recognition of Pol II promoter sequences. *Bioinformatics*, **15**, 356–361.
- Levy,S., Compagnoni,L., Myers,E.W. and Stormo,G.D. (1998) Xlandscape: the graphical display of word frequencies in sequences. *Bioinformatics*, **14**, 74–80.
- Mache,N., Reczko,M. and Hatzigeorgiou,A. (1996) Multistate time-delay neural networks for the recognition of POL II promoter sequences. *Ismb96*, St Louis, <http://www.informatik.uni-stuttgart.de/ipvr/bv/personen/mache>.
- Pedersen,A.G., Baldi,P., Chauvin,Y. and Brunak,S. (1999) The biology of eukaryotic promoter prediction—a review. *Comput. Chem.*, **23**, 191–207.
- Périer,R.C., Praz,V., Junier,T., Bonnard,C. and Bucher,P. (2000) The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res.*, **28**, 302–303.
- Reese,M.G. and Eeckman,F.H. (1999) Time-delay neural network for eukaryotic promoter prediction, unpublished.
- Reese,M.G., Harris,N.L. and Eeckman,F.H. (1996) Large scale sequencing specific neural networks for promoter and splice site recognition. In Hunter,L. and Klein,T.E. (eds), *Biocomputing: Proceedings of the 1996 Pacific Symposium, 2–7 January, 1996*. World Scientific, Singapore, http://www.fruitfly.org/seq_tools/promoter.html.
- Reese,M., Kulp,D., Gentles,A. and Ohler,U. (1999) http://www.fruitfly.org/seq_tools/datasets/Human.
- Reese,M.G., Hartzell,G., Harris,N.L., Ohler,U., Abril,J.F. and Lewis,S.E. (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.*, **10**, 483–501.
- Scherf,M., Klingenhoff,A. and Werner,T. (2000) Highly specific localisation of promoter regions in large genomic sequences by Promoter Inspector: a novel context analysis approach. *J. Mol. Biol.*, **297**, 599–606.
- Scherf,M., Klingenhoff,A., Frech,K., Quandt,K., Schneider,R., Grote,K., Frisch,M., Gailus-Durner,V., Seidel,A., Brack-Werner,R. and Werner,T. (2001) First pass annotation of promoters on human chromosome 22. *Genome Res.*, **11**, 333–340.
- Weinzierl,R.O.J. (1999) *Mechanism of Gene Expression*. Imperial College Press, London.