

Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes

Vladimir B. Bajic* and Seng Hong Seah¹

Knowledge Extraction Laboratory and ¹Discovery Systems Laboratory, Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613

Received February 15, 2003; Revised and Accepted April 3, 2003

ABSTRACT

Recognition of gene starts is a difficult and yet unsolved problem. We present a program, Dragon Gene Start Finder (DGSF), which assesses the gene start in mammalian genomes and predicts a region which should overlap with the first exon of the gene or be in its proximity. The program has been rigorously tested on human chromosomes 4, 21 and 22, and in a strand specific search achieves an overall sensitivity of ~65% and a positive predictive value of ~78%. The sensitivity for the CpG-island related promoters is >88%. DGSF is free for academic and non-profit users at http://sdmc.lit.org.sg/promoter/dragonGSF1_0/genestart.htm; the download version of the program integrated within the TRANSPLOER™ package can be obtained from Biobase GmbH, at <http://www.biobase.de/>.

INTRODUCTION

Prediction of gene starts is an important issue (1), since the promoter region around the 5' end of a gene contains crucial regulatory modules to activate a gene under different conditions and timing (2). In prediction of the 5' ends of genes using the *ab initio* search, general gene recognition programs such as Genscan and Fgenesh (1,3) are not very efficient as these programs mainly focus on the recognition of coding exons which, sometimes, can be very far from the gene's promoter (4,5). Specialized programs devoted to promoter search are more successful for this type of problem. Previous results and efficiency of promoter prediction (reviewed in 6,7) revealed that this problem requires far more efficient solutions. The greatest problem was that the increased sensitivity of prediction programs was inevitably accompanied by a high number of false positive (FP) predictions which rendered the programs unusable for large-scale analyses. The first breakthrough was made in 2000 with the appearance of PromoterInspector (8) which set up standards for promoter prediction accuracy. This was the first program that has made a significant proportion of true positive (TP) predictions with an

acceptable level of FP predictions (9,10). Following PromoterInspector, a number of new generation promoter prediction programs appeared (11–18), most of which claimed similar or better accuracy. Analysis of these programs reveals that, generally, the most efficient were those (11–13,17) that relied on either the concept of CpG-islands (19–23) or those which exploited G + C richness around the promoter region (16,18). CpG islands seem to be the most prominent marker of genes in mammalian genomes (23) which justifies their use in the annotation of the human genome (24,25). CpG islands are also suggested as the global genomic signal to enhance promoter predictions (26). In this article we present a system, Dragon Gene Start Finder (DGSF), which exploits the concept of CpG islands for the prediction of promoters.

OUTLINE OF THE IMPLEMENTED ALGORITHM

Full details of the implemented recognition algorithm are provided on-line at the program's web page 'System model'. Here we present the outline of the algorithm implementation (Supporting Material 1, available as Supplementary Material) to enable users to have a clear picture of the general concepts utilized and the resulting program constraints. DGSF predicts a region that ideally overlaps with the first exon of the gene or is in its proximity. It also assesses the gene start. DGSF combines three systems to achieve these goals: (i) the Dragon Promoter Finder (DPF) system (14,15,18); (ii) the system which estimates the presence of the CpG islands; (iii) the system which combines information obtained from (i) and (ii) into the final predictions using sensor fusion methods, data preprocessing and an ANN (artificial neural network). DPF searches for transcription start sites (TSSs) on both strands of the query DNA sequence. When a CpG-island is detected, every prediction of DPF within [–3700, +3700] relative to the midpoint of the detected CpG-island is assessed to find out if the combination of that TSS prediction and the CpG-island is the one which suggests the start of the gene. The best combination is selected by the system. For each detected CpG island the combination algorithm selects at most one predicted TSS location such that, when combined with the other input data, the ANN produces the highest score above the user-selected threshold. This threshold is the only parameter which the user can change to influence the program's performance.

*To whom correspondence should be addressed. Tel: +65 68748800; Fax: +65 67748056; Email: bajicv@i2r.a-star.edu.sg

There is no guarantee that the algorithm will identify the TSS prediction closest to the actual TSS location. The bias in selected TSS locations is removed based on a statistical analysis (see formulae for predicted region in Supporting Material 2, available as Supplementary Material). The corrected TSS positions are denoted as gene starts with the 'GS' identifier in the report file.

The ANN is a four-layer network which uses 'logsig' transfer functions (27) for the hidden and output layers. It has 10 neurons in the first hidden layer, 15 neurons in the second hidden layer and one output layer neuron. It is trained by the optimized backpropagation algorithm (28) with weight decay and for optimal separation of the 'correct' versus 'wrong' combination of the found CpG island and related TSS predictions of DPF.

WEB SITE DESCRIPTION

The DGSF portal allows users to input the sequence in FASTA, GenBank, EMBL, IG, GCG or plain formats by either pasting it into the input box or by reading it from a file. A snapshot of the program's portal is shown in Supporting Material 3, available as Supplementary Material. Sequences of <100 000 nt in length can be submitted for an analysis. By default, the system will make predictions on both strands and the predictions are strand specific. First, the predictions on the '+' strand will be presented, then predictions on the '-' strand. If the query sequence has regions or nucleotides which are not A, C, G or T, but other letters from the IUPAC ambiguity code set (29), such as R, M, N, etc., the system skips such areas. It will also report any such segment composed of 100 or more successive ambiguous nucleotides. This handling of insufficiently defined regions in the query sequence aims to help wet-lab biologists in sequence manipulation and the analysis of results. The user can change this minimum length of gaps in the DNA sequence for reporting purposes. Predictions of DGSF are CpG island related, i.e. the assessed gene starts are contained within the detected CpG island. The program allows for the two modes of operation: interactive and non-interactive. The interactive mode is suitable for wet-lab biologists to have a more detailed analysis of the predicted promoter/gene start region. Due to its simple report format the non-interactive mode is suitable for an inspection of the presence of the CpG island related promoters. For convenience, users have a test sequence available (human CST3 gene for cystatin C, HSCST3G, Accession number X52255) at the program's web page 'Test sequence'. In the non-interactive mode only the report page will be presented to the user in a separate window (Supporting Material 4, a report page for the test sequence, available as Supplementary Material). For the same test sequence in the interactive mode, predictions are shown as in Supporting Material 5, available as Supplementary Material. Using radio button fields at the report page, the user can select which of several possible predictions he wants to examine further. The user can select a region around the assessed gene start and subject that segment to analysis by the MATCHTM program (30) which uses the TRANSFACTM database version 6.1 (30). Transcription factor binding sites (TFBSs) found with MATCHTM in the selected region are presented to the user

through a graphical display with all pertinent information appended (the sample report page for the first selected 'GS' position of the test sequence is shown in Supporting Material 6, available as Supplementary Material). This can help the user in identifying potentially important TFBSs in the immediate neighborhood of the gene start. In using MATCHTM, users can select matrix profiles optimized for the minimum sum of FP and false negative (FN) predictions of TFBSs (default), or they can manually select the thresholds for the matrix score and core score as provided by MATCHTM. For details on how MATCHTM operates and the meaning of these thresholds, users should consult the Biobase web site.

REPORT PAGE

A report page is displayed using two letter identifiers SQ, PO, GS, ST, TH and UD associated with the information presented. SQ stands for the sequence identifier if any; PO stands for the region that is expected to overlap or to be in proximity of the first exon; in the interactive mode it is shown in red color for easy observation; GS stands for the assessed gene start (note that the position indicated by GS is not the same as the prediction of DPF); in the interactive mode it is shown in blue color for easy observation; ST stands for the strand where the gene is; TH stands for the threshold used in predictions; UD stands for the bounds of the DNA segments consisting of 100 or more successive nucleotides which are not fully defined. The user can change the minimum number of the undefined nucleotides in gaps to be reported.

In the sample report file (Supporting Materials 4–6, available as Supplementary Material), the test sequence is analyzed. The single stranded sequence has 7292 nt. A gene start is on the '+' strand at 1010 (the 5' end of mRNA falls on 1010 '+' strand). The gene start is predicted by the system at position 992. The region that should overlap with the first exon is from [862, 1604] counted from the 5' end of the submitted sequence. The system also made one prediction on the '-' strand.

SUGGESTIONS FOR EFFICIENT PROGRAM USE

DGSF is designed under the assumption that there is no evidence for gene presence in the query DNA sequence. Its primary aim is to produce an approximate assessment of the locations of gene starts and promoters in such a blind search. DGSF assesses the CpG island presence in the sequence and makes predictions of gene starts within the CpG islands. Sequences which are >100 000 nt would need to be segmented. The default threshold is 0.994. The recommended range for the threshold is [0.99, 1]. By increasing the threshold, the user reduces the number of expected FP predictions, but also reduces the sensitivity. By decreasing the threshold the sensitivity should increase at the expense of more FP predictions. At the default threshold DGSF detects more than 88% of the CpG island related promoters, which is a very high sensitivity for this promoter class. Note that if promoters in the sequence are not CpG island related, then no matter how much the threshold is reduced, the system will not make predictions. If the analysis does not produce predictions, then either there are no CpG island-related promoters in the sequence, or the

existing CpG island promoters require a lower threshold for detection. If the reduced threshold does not result in positive predictions, then it is likely that the sequence does not contain CpG island related promoters.

Because the assessment of the gene start is only approximate, a more detailed search for TSSs, including alternative TSSs, can be done by DPF (14,15,18), or some other programs efficient in locating TSSs. We suggest a search in the range [-4000, +4000] relative to the location of the assessed gene start. Since the total number of predictions that DGSF makes at the default threshold on the whole non-masked human (36 080 predictions) and mouse (<45 000 predictions) genomes is relatively small, this search for TSSs is, in fact, very focused.

PERFORMANCE

Details regarding the program's performance and comparisons with several other programs, are presented at the program's web site pages 'Accuracy' and 'Hum. chr. 4,21,22'. In these analyses DGSF achieves favorable results. To train and test the DGSF system we used a large dataset and procedure explained at the program's web site. We evaluated the trained system on whole human chromosomes 4, 21 and 22. These chromosomes contain ~500 000 000 nt used in a two stranded analysis. Chromosome 4 is the most G + C-poor (38%), chromosome 21 has a G + C content of 41%, just below the average for the whole human genome (42%), while chromosome 22 is the second G + C richest chromosome (48%) and the most well annotated. No sequences from these chromosomes were used in the training of DGSF. The predictions were matched to the gene starts obtained by mapping full-length cDNA sequences from DBTSS (31,32).

DGSF achieves an overall sensitivity of 65% with a positive predictive value (ppv) of ~78%, while in the category of CpG island related promoters it achieves sensitivity >88%. The criteria used are explained at the program's web page 'Accuracy'. To the best knowledge of the authors, the only program that achieves higher overall sensitivity with an acceptable level of FP predictions is FirstEF (13).

BENEFITS OF USING DGSF WITH REGARDS TO OTHER SIMILAR SOFTWARE

DGSF makes strand specific predictions, assessment of gene starts, allows for several different formats for query sequences, allows for the analysis of promoter region content based on TRANSFAC database and reports on gaps in the query sequence. Our comments refer to PromoterInspector (8), CpGProD (17), CpG-promoter (11), Eponine (16) and FirstEF (13). Benefits of using DGSF over these programs are as follows. Firstly, none of these programs allows for a more detailed analysis of the predicted promoter region (based on the TRANSFAC or any other transcription factor database), nor do they report gaps in the sequence. Secondly, all these programs, with the exception of PromoterInspector, allow only FASTA format of the query sequences. Additionally, PromoterInspector, CpGProD and CpG-Promoter do not make strand specific predictions and do not predict gene starts. PromoterInspector has very limited web access and CpGProD

requires that the query sequence be previously masked by RepeatMasker (A.F.A. Smit and P. Green, unpublished data).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to Yutaka Suzuki for providing the full-length cDNA sequences from DBTSS (Sugano Laboratory) mapped to genomic contigs and used in evaluation of DGSF's performance.

REFERENCES

1. Stormo,G.D. (2000) Gene-finding approaches for eukaryotes. *Genome Res.*, **10**, 394–397.
2. Weinzierl,R.O.J. (1999) *Mechanism of Gene Expression*. Imperial College Press, London.
3. Burset,M. and Guigo,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.
4. Suzuki,Y., Yamashita,R., Nakai,K. and Sugano,S. (2002) DBTSS: DataBase of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
5. Suzuki,Y., Taira,H., Tsunoda,T., Mizushima-Sugano,J., Sese,J., Hata,H., Ota,T., Isogai,T., Tanaka,T. and Morishita,S. (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.*, **2**, 388–393.
6. Fickett,J.W. and Hatzigeorgiou,A.G. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.
7. Prestridge,D.S. (2000) Computer software for eukaryotic promoter analysis. Review. *Methods Mol. Biol.*, **130**, 265–295.
8. Scherf,M., Klingenhoff,A. and Werner,T. (2000) Highly specific localisation of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.*, **297**, 599–606.
9. Scherf,M., Klingenhoff,A., Frech,K., Quandt,K., Schneider,R., Grote,K., Frisch,M., Gailus-Durner,V., Seidel,A., Brack-Werner,R. *et al.*, (2001) First pass annotation of promoters on human chromosome 22. *Genome Res.*, **11**, 333–340.
10. Werner,T. (2002) Finding and decrypting of promoters contributes to elucidation of gene functions. *In Silico Biol.*, **2**, 23.
11. Ioshikhes,I.P. and Zhang,M.Q. (2000) Large-scale human promoter mapping using CpG islands. *Nature Genet.*, **26**, 61–63.
12. Hannehalli,S. and Levy,S. (2001) Promoter prediction in the human genome. *Bioinformatics*, **17**, S90–S96.
13. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nature Genet.*, **29**, 412–417.
14. Bajic,V.B., Seah,S.H., Chong,A., Zhang,G., Koh,J.L.Y. and Brusica,V. (2002) Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. *Bioinformatics*, **18**, 198–199.
15. Bajic,V.B., Chong,A., Seah,S.H. and Brusica,V. (2002) Intelligent system for vertebrate promoter recognition. *IEEE Intell. Sys. Mag.*, **17**, 64–70.
16. Down,T.A. and Hubbard,T.J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, **12**, 458–461.
17. Ponger,L. and Mouchiroud,D. (2002) CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, **18**, 631–633.
18. Bajic,V.B., Seah,S.H., Chong,A., Krishnan,S.P.T., Koh,J.L.Y. and Brusica,V. (2003) Computer model for recognition of functional transcription start sites in polymerase II promoters of vertebrates. *J. Mol. Graph. Model.*, **21**, 323–332, Available online 12 November 2002.
19. Bird,A.P., Taggart,M.H., Nichollas,R.D. and Higgs,D.R. (1986) Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. *EMBO J.*, **6**, 999–1004.

20. Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
21. Cross, S.H. and Bird, A.P. (1995) CpG islands and genes. *Curr. Opin. Genet. Dev.*, **5**, 309–314.
22. Cross, S.H., Clark, V.H. and Bird, A.P. (1999) Isolation of CpG islands from large genomic clones. *Nucleic Acids Res.*, **27**, 2099–2107.
23. Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.
24. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
25. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
26. Pedersen, A.G., Baldi, P., Chauvin, Y. and Brunak, S. (1999) The biology of eukaryotic promoter prediction—a review. *Comp. Chem.*, **23**, 191–207.
27. Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.
28. Sha, D. and Bajic, V.B. (2002) On-line hybrid learning algorithm for MLP in identification problems. *Comp. Elect. Eng., An Int. J.*, **28**, 587–598.
29. IUB Nomenclature Committee (1985) *Eur. J. Biochem.*, **150**, 1–5.
30. Matys, V., Fricke, E., Geffers, R., Gossling, E., Hanbrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
31. Suzuki, Y., Yamashita, R., Nakai, K. and Sugano, S. (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
32. Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A. and Sugano, S. (1997) Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene*, **200**, 149–156.