BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 4 2002
Pages 631–633

CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences

Loïc Ponger and Dominique Mouchiroud*

Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558 - Université Claude Bernard, 43, Bd du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

Received on September 7, 2001; revised on October 23, 2001; accepted on November 7, 2001

ABSTRACT

Results: CpGProD is an application for identifying mammalian promoter regions associated with CpG islands in large genomic sequences. Although it is strictly dedicated to this particular promoter class corresponding to ≈50% of the genes, CpGProD exhibits a higher sensitivity and specificity than other tools used for promoter prediction. Notably, CpGProD uses different parameters according to species (human, mouse) studied. Moreover, CpGProD predicts the promoter orientation on the DNA strand.

Availability: http://pbil.univ-lyon1.fr/software/cpgprod.html

Supplementary information: http://pbil.univ-lyon1.fr/software/cpgprod.html

Contact: ponger@biomserv.univ-lyon1.fr

INTRODUCTION

A number of promoter detection programs attempting to recognize functional sequences (TATA, CAAT, transcription factor binding site, . . .) or to identify the oligonucleotide frequencies specific for promoters exist (for a review see Fickett and Hatzigeorgiou, 1997), but, excepting the recently developed programs PromoterInspector and CpG_promoter (Scherf *et al.*, 2000; Ioshikhes and Zhang, 2000), their specificity is often too low to be used for annotation of large genomic sequences.

In vertebrata, there is a particular class of promoters colocalized with an atypical structure, the CpG Islands (CGIs). In vertebrate genomes, the CpG dinucleotide is often methylated and is depleted at 25% of the expected frequency. The CGIs are stretches of DNA escaping methylation and characterized by a high G + C content and a high frequency of CpG dinucleotides relative to the bulk DNA (Bird, 1986). 50–60% of the human genes exhibit a CGI over the Transcription Start Site (TSS) but not all the CGIs are associated with promoter regions (Larsen *et al.*, 1992). The CGIs associated with promoters (start CGIs) can be, *a priori*, identified from their structural characteristics (greater size, higher G + C content and CpGo/e ratio than no-start CGI; Ioshikhes and Zhang, 2000; Ponger *et al.*, 2001).

This paper presents CpGProD, a mammalian-specific software to identify the TSS associated with CGIs.

METHODS

The CpGProD method can be divided into two steps. Firstly, CpGProD searches for all CGIs located in the submitted sequences. Secondly, CpGProD identifies the start CGIs and predicts the orientation of these potential promoters. CpGProD was trained and tested by using a human and a mouse dataset composed by genes with a known TSS.

Datasets

The human and the mouse coding protein sequences were extracted from HOVERGEN (release 114, October 1999, Duret *et al.*, 1994). HOVERGEN corresponds to GenBank sequences from all vertebrate species with some additional data allowing extraction of non-coding sequences. The TSS annotations were obtained from the mRNA descriptions available in the features (partial mRNA were not considered). For each gene, we extracted a sequence composed by the 5′ non-coding region, the exons, the introns and the 3′ non-coding region. Sequences with less than 500 nt (CGIs′ length) upstream and downstream the TSS were excluded. The sequence dataset is composed by 755 human and 147 mouse genes with a known TSS (32.8 and 2.4 Mb for human and mouse datasets respectively). CpGProD was used to find the CGIs over these sequences. Partial CGIs, that is CGIs overlapping one extremity of the sequences, were excluded. CGIs located over the TSS were classified as start CGI whereas other CGIs were classified as no-start CGIs. The CGI dataset is composed by 818 human CGIs and 163 mouse CGIs. These CGIs datasets were divided into two halves: the first half of each dataset was used to train CpGProD to identify start

*To whom correspondence should be addressed.

CGIs and the second half was used to test CpGProD. Moreover the sequences and the CGIs used in the dataset of Scherf *et al.* (2000) and Ioshikhes and Zhang (2000) were excluded from the training part of the datasets.

CpG island search

In order to enhance the specificity, the sequences have to be primarily processed by RepeatMasker (Smit and Green, unpublished) to exclude potential noise due to some repeat elements exhibiting a structure similar to CGIs whereas they are often methylated (Ponger *et al.*, 2001). Moreover, to eliminate small CGIs corresponding generally to no-start CGIs, CpGProD uses a CGI definition more stringent than that proposed by Gardiner-Garden and Frommer (1987). CGIs are defined as DNA regions longer than 500 nucleotides (instead 200 bp), with a moving average G + C frequency above 0.5 and a moving average CpG observed/expected (CpGo/e) ratio greater than 0.6. Moving average value for the G + C frequency and the CpGo/e ratio are calculated for each sequence by using a 500 nucleotides window moving along the sequence in steps of 1 nt. Overlapping windows with a G + C frequency greater than 0.5 and a CpGo/e ratio greater than 0.6 were grouped to form the CGIs. Considering these parameters, 56% of the human genes and 52% of the rodent genes in the sequence dataset exhibit a start CGI. The percentage observed for human genes is similar to the result of Larsen *et al.* (1992) who used a threshold of 200 bp, indicating that the sensitivity is not decreased.

Start CpG island identification

A first score corresponding to the probability to be over the TSS (start-p) is calculated from the length, the G + C content and the CpGo/e ratio of each CGI. A second score is calculated from the AT-skew and the GC-skew values which are two parameters quantifying a compositional bias between the plus and the minus DNA strands (Lobry, 1996) and exhibiting different values according to the strand of the corresponding gene (L.Ponger, unpublished data). A strand (plus or minus) and a probability to be over this predicted strand (strand-p) are determined from this score. These two relations were determined by using a generalized linear model (McCullagh and Nelder, 1989) with the first half of the CGI dataset. Since, the CGI structure seems to be conserved in all studied mammals (pig, bovine, human) except in mouse and rat (Cuadrado *et al.*, 2001; Matsuo *et al.*, 1993), we used two datasets, one composed by human CGIs and one composed by rodent CGIs.

IMPLEMENTATION

CpGProD is implemented in C language. It is available either via a web server, useful for small datasets, or as a standalone application for larger datasets (for Solaris, Windows, Linux, SGI and MacOS). The output gives the structural characteristics (length, G + C frequency and CpGo/e ratio), the start-p value, the strand and the strand-p value of each detected CGI. Moreover, a graph representing CGIs over the sequences is drawn.

RESULTS AND DISCUSSION

The main result of CpGProD is a start-p value corresponding to the predicted probability to be a start CGI. The sensitivity and the specificity of CpGProD depend on the minimal start-p threshold chosen to predict promoter. CpGProD was tested by using the second part of the CGI datasets that was not used during the training step (cf. web site, Table 1). In the human dataset, if all the detected CGIs are considered as promoters, CpGProD finds a CGI over 56% of the TSSs with specificity about 0.39. If we consider as promoters only the CGIs with a start-p value greater than 0.3, the sensitivity decreases to 27% whereas the specificity increases to 0.51. For both species, the sensitivity decreases and the specificity increases while the threshold value increases indicating that the start-p value is correlated with the probability to be a start CGI. Concerning the orientation of the promoters, 70% of the human and 73% of the rodent predictions are correct. These percentages increase with the start-p threshold (cf. web site, Table 1).

CpGProD was compared with CpG_promoter and PromoterInspector by using three different datasets since these programs cannot be used on our data: the former needs a commercial license (for Splus), online access to the latter is strongly restricted. Thus, CpGProD was tested on a dataset composed by 19 human genes with a start CGI and already used to test CpG_promoter (cf. web site, Table 1). The results show that CpGProD exhibits a higher sensitivity (0.74 versus 0.62) and a higher specificity (0.87 versus 0.62) than CpG_promoter (cf. web site, Table 1). Another test was made by using two datasets previously used for PromoterInspector. The first is composed by 35 human and mouse genes with TSS annotations (cf. web site, Table 1) whereas the second is composed by 545 genes located over the chromosome 22 (cf. web site, Table 2; Dunham *et al.*, 1999). For this latter dataset we used the same method as that used by Scherf *et al.* (2001) with PromoterInspector: all the predictions located in the range $-2000:+500$ around the 5′ extremity of a known gene or in the range $-6000:+500$ around the 5′ extremity of a predicted gene were considered as a true positive promoter region. The results show that CpGProD exhibits a higher sensitivity (0.38 versus 0.33 for the chromosome 22) and a higher specificity (0.62 versus 0.40 for the chromosome 22) than PromoterInspector (cf. web site, Tables 1 and 2).

The differences observed between CpG_promoter and CpGProD can be explained by the method used to

search the CGIs. With CpGProD, repeated sequences and small CGIs are not considered, thus increasing the specificity of the start CGIs detection. Contrary to PromoterInspector, CpGProD is strictly dedicated to CGI associated promoters and is more efficient for this class of promoters. This difference between PromoterInspector and CpGProD confirms the results of Hannenhalli and Levy (2001) showing that CGIs are the best signal to detect promoter regions. CpGProD was also applied to the Human Genome Project data (cf. web site, Table 2). The results indicate that 27% of the gene starts are localized in a CGI exhibiting a start-*p* value greater than 0.3. We observe a difference of sensitivity between the known and the predicted genes (41 and 23% respectively) probably due to inaccuracy in location of 5′ extremity of predicted genes. It could be useful for gene annotation to determine if all the CGIs with a start-*p* value greater than 0.3 can be associated with a gene.

To date, although relatively simple, CpGProD is the most efficient tool dedicated to the detection of CGI associated promoters in mammalian sequences. In sequence annotation, CpGProD should be used as a first step, before using other promoter prediction software exhibiting a lower specificity but able to localize more accurately the core promoter and the TSS.

REFERENCES

Bird,A.P. (1986) CpG rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.

Cuadrado,M., Sacristan,M. and Antequera,F. (2001) Species-specific organization of CpG island promoter at mammalian homologuous genes. *EMBO Rep.*, **2**, 586–592.

Dunham,I. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.

Duret,L., Mouchiroud,D. and Gouy,M. (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.*, **22**, 2360–2365.

Fickett,J.W. and Hatzigeorgiou,A.G. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.

Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.

Hannenhalli,S. and Levy,S. (2001) Promoter prediction in the human genome. *Bioinformatics*, **17**, S90–S96.

Ioshikhes,I.P. and Zhang,M.Q. (2000) Large-scale human promoter mapping using CpG islands. *Nature Genet.*, **26**, 61–63.

Larsen,F., Gundersen,G., Lopez,R. and Prydz,H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.

Lobry,J.R. (1996) Asymmetric substitution patterns in two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**, 660–665.

Matsuo,K., Clay,O., Takahashi,T., Silke,J. and Schaffner,W. (1993) Evidence for erosion of mouse CpG islands during mammalian evolution. *Somat. Cell Mol. Genet.*, **19**, 543–555.

McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models*. Chapman and Hall, London.

Ponger,L., Duret,L. and Mouchiroud,D. (2001) Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res.*, **11**, 1854–1860.

Scherf,M., Klingenhoff,A. and Werner,T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: 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.