BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 4 2004, pages 583–585
DOI: 10.1093/bioinformatics/btg460

AntiHunter: searching BLAST output for EST antisense transcripts

Giovanni Lavorgna[1,*], Luca Sessa[1,2], Alessandro Guffanti[3], Lelio Lassandro[3] and Giorgio Casari[1]

[1]Istituto Scientifico H. S. Raffaele, Via Olgettina 60, 20132 Milan, Italy, [2]Dulbecco Telethon Institute IGB CNR Via Pietro Castellino 111, 80131 Naples, Italy and [3]IFOM-FIRC Institute of Molecular Oncology, Via Adamello 16, 20139 Milan, Italy

Received on August 22, 2003; accepted on October 6, 2003
Advance Access publication January 22, 2004

ABSTRACT

Summary: AntiHunter is a new web-based tool for the identification of expressed sequence tag (EST) antisense transcripts from BLAST output. In order to perform an analysis, user is required to input a genomic sequence plus an associated list of transcript names and coordinates of the genomic region (i.e. genome annotation). After masking the repeated regions (if any), program will perform a BLASTN search of the input sequence versus the selected EST database, reporting by Email the EST entries that reveal a putative antisense transcript with respect to the user supplied list.

Availability: AntiHunter is currently available through a web interface at http://bio.ifom-firc.it/ANTIHUNTER/.

Contact: giovanni.lavorgna@hsr.it

Recent experimental work suggests a functional role for mRNA antisense (AS) transcripts at a surprising variety of levels in gene regulation, including genomic imprinting, RNA interference, translational regulation, alternative splicing, X-inactivation and RNA editing (reviewed in Vanhee-Brossollet and Vaquero, 1998). We have developed a software tool, AntiHunter, aimed at facilitating the *in-silico* identification of potential AS expressed sequence tag (EST) transcripts within a given genomic region of interest. Program will take as input a genomic sequence and a list of annotated transcripts of the genomic regions. This list includes transcript names, their beginning and ending positions plus their strand occurrence. Then, it will perform the following tasks:

— Run the RepeatMasker (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) program on the genomic sequence in order to filter out repeated sequences.

— Perform a BLASTN search of the resulting sequence versus a selected EST database.

— Parse the BLASTN output looking for AS EST with respect to the annotated genes. If any match is found, other information (such as the length of the spanned genomic region of the EST AS transcript, the sequence of the actual splicing sites plus some flanking sequences, etc.) is added to the output as well.

— Report the results to the user by Email.

In the past months, five papers have reported the computational identification of transcripts with the potential for sense–antisense pairing (Fahey *et al*., 2002; Lehner *et al*., 2002; Shendure and Church, 2002; Yelin *et al*., 2003; Kiyosawa *et al*., 2003). The first two studies refrained from using EST databases because of the uncertainties regarding the correct orientation of the ESTs. Advances in algorithm design allowed to Shendure and Church (2002) and, to a greater extent, to the 'Antisensor' algorithm developed by Yelin *et al*. (2003) to overcome the search background associated with problematic EST strand annotation. We employed advances developed in these studies to lower the noise of our search. In particular, besides using the database annotation, the program gains independent information on EST strand source by looking (i) at the splice junctions of the genomic region matching a spliced EST and (ii) at the presence of a PolyA tail in 3′ annotated ESTs. Only EST showing at least one of these independent evidences for strand source are considered further for potential sense–antisense pairing. Moreover, since oligo(dT)-priming can also take place on internal PolyA stretches within an unspliced transcript, the algorithm identifies such genomic PolyA stretches and disregards the relative PolyA information obtained from the EST sequence. More details on the implementation of these methods can be found on the AntiHunter documentation page (http://bio.ifom-firc.it/ANTIHUNTER/ah_help.new.htm).

The Antisensor algorithm, thanks to its gene clustering approach, can also gain information on strand source by tentatively translating the AS transcript (Yelin *et al*., 2003). Moreover, since much of the work is precomputed, web

*To whom correspondence should be addressed.

```
  AI370586  H. sapiens   0e+00  3'  482    7 - 482  0.99  1 n  n  y  P  66333 - 66816    483  FGF2 + not reported
  AI492978  H. sapiens   0e+00  3'  347   13 - 347  0.97  1 n  n  y  P  66333 - 66667    334  FGF2 + kidney
  BQ774001  H. sapiens   0e+00  3'  799   12 - 798  0.98  4 y  y  y  P  66333 - 91288  24955  FGF2 + Human Chondrosarcoma Cell Line
  BU619706  H. sapiens   0e+00  3'  805   12 - 797  0.98  4 y  y  y  P  66333 - 91288  24955  FGF2 + Cell Line
  BM970451  H. sapiens   0e+00  3'  686   12 - 686  0.98  3 y  y  y  P  66333 - 86294  19961  FGF2 + Lung
  BF739878  H. sapiens   0e+00  3'  630    2 - 630  1.00  2 y  y  y  P  66333 - 71358   5025  FGF2 + kidney
  AI803148  H. sapiens   0e+00  3'  559    9 - 547  0.96  1 n  n  y  P  66333 - 66873    540  FGF2 + not reported
  AA628770  H. sapiens   0e+00  3'  553    8 - 553  0.99  1 n  n  y  P  66333 - 66879    546  FGF2 + not reported
  AI337876  H. sapiens   0e+00  3'  604   34 - 604  0.95  1 n  n  y  P  66344 - 66914    570  FGF2 + uterus
  AI074587  H. sapiens   0e+00  3'  558    6 - 558  0.99  2 y  y  y  P  66416 - 71366   4950  FGF2 + Liver and Spleen
  CA392430  H. sapiens   0e+00  5'  716    1 - 716  1.00  4 y  y  y  M  66490 - 91369  24879  FGF2 + RPE/choroid
  BQ434362  H. sapiens  1e-157  5'  933    1 - 909  0.97  5 y  y  y  M  66511 - 96198  29687  FGF2 + retinoblastoma
  BE897694  H. sapiens  1e-152  5'  692   84 - 561  0.69  3 y  y  y  M  66546 - 86312  19766  FGF2 + melanotic melanoma
  BF239070  H. sapiens  1e-106  5'  733    5 - 715  0.97  5 y  y  y  M  66609 - 96584  29975  FGF2 + from chronic myelogenous leukemia
 HSAA14389  H. sapiens  1e-134  5'  448    3 - 419  0.93  3 y  y  y  M  66616 - 86314  19698  FGF2 + Pooled human melanocyte, fetal heart
  BE894167  H. sapiens  1e-107  5'  708    1 - 708  1.00  5 y  y  y  M  66695 - 96196  29501  FGF2 + melanotic melanoma
  BG386740  H. sapiens  1e-135  5' 1048    2 - 777  0.74  5 y  y  y  M  66704 - 96272  29568  FGF2 + adenocarcinoma cell line
  CB128685  H. sapiens  1e-145  5'  616    4 - 616  1.00  5 y  y  y  M  66887 - 96293  29406  FGF2 + Liver
```

Fig. 1. Detection of EST AS transcripts by the AntiHunter program (excerpt from program output). A 96 700 bp sequence from the human chromosome 4 (chr4:124140673–124237372 region from the UCSC genome browser, April 2003 freeze), containing the FGF-2 gene at coordinates 397-71910, was used as a query to AntiHunter. As a result, program returned several AS ESTs, whose accession numbers are shown in the first column. These ESTs correspond to the exons of NUDT6, a known AS transcript to the FGF-2 gene (Li and Murphy, 2000). Following columns in the output indicate, respectively, the EST organism, the BLAST match significance (*E*-value), the EST strand annotation and length, the beginning and ending EST matching position and the relative fraction of the encompassed region over EST length, the number of EST sub-matches in BLAST output, whether has been the EST spliced ('y' or 'n'), whether have been found canonical GT and AG splicing consensi on genomic sequence ('y' or 'n'), whether is there a lack of an annotated overlapping gene for this match ('y' or 'n'), the Plus/Plus or Plus/Minus alignment orientation on BLAST output ('P' or 'M'), the beginning and ending genomic matching position and the relative length of the encompassed region, the annotated sequence for which the antisense match was detected and its strand, the EST tissue or organ source (first line only).

response from the Antisensor takes only a few minutes, whereas AntiHunter might take even a few hours to complete the job. On the plus side for AntiHunter there is a greater flexibility, i.e. whereas the Antisensor is currently limited to human sequences, AntiHunter can be used, in principle, to analyse genomic regions from any species for which EST and genomic data are available. Also, AntiHunter can tolerate a variable number of distance bases between an annotated gene and an antisense transcript. This can be useful for detecting AS transcripts to genes with only partially characterized 5′ and/or 3′ ends. It can also facilitate the detection of transcribed gene regulatory regions that originate from intergenic regions and that contribute to regulation of their neighbour genes (Bae *et al*., 2002; Drewell *et al*., 2002; Rank *et al*., 2002).

The accuracy of AntiHunter was tested using genes, which had been shown previously to possess AS transcripts. Fifteen genomic regions, containing overlapping transcriptional units in mammalian genomes described previously in literature, were used as input to the program. As a result, program correctly determined the presence of EST AS transcripts in 14 out of 15 cases. In the missing case, given by the human distal-less homeobox protein 1 (DLX1) gene, the presence of AS transcripts was not detected because of the lack of double-checked EST entries in the dbEST (for details on program performance, see the program documentation page at http://bio.ifom-firc.it/ANTIHUNTER/ah_help.new.htm#Bugs). An excerpt from AntiHunter output is shown on Figure 1. A 96 700 bp sequence from the human chromosome 4 (chr4:124140673–124237372 region from the UCSC genome browser, April 2003 freeze), containing the FGF-2 gene at coordinates 397-71910, was used as a query to AntiHunter. As a result, program returned several AS ESTs, whose accession numbers are shown in the first column of Figure 1, corresponding to the exons of NUDT6 [nudix (nucleoside diphosphate linked moiety) X-type motif 6], a known AS transcript to the FGF-2 gene (Li and Murphy, 2000). Intriguingly, AntiHunter was also able to identify, despite the fact that it was used to query a human sequence, the presence of EST antisense transcripts from other species than human (i.e. *Rattus norvegicus*, *Mus musculus* and *Bos taurus*), unravelling the possible evolutionary conservation of the phenomenon (data not shown).

In conclusion, AntiHunter is a new tool capable of performing an *in-silico* search for putative EST AS transcripts. It can effectively use relatively raw, but frequently updated, material, such as the EST sequences and provide useful preliminary results for guiding the design of further experimental analysis. However, due to the fact that EST data can be still inaccurate in many aspects, it is strongly recommended that, whenever possible, user should verify the results by 'wet-biology' methods.

ACKNOWLEDGEMENTS

We are grateful to Steve Scherer, Layla Parker-Katiraee, Lucilla Luzi, Valerio Orlando, Alessandro Brozzi and Stefano Confalonieri for helpful comments and suggestions and to Gyorgy Simon for setting up the IFOM computer facility. L.S. was supported by grants Telethon Foundation and Volkswagenstiftung I/77 996 to Valerio Orlando. A.G., S.C.

and L.L. are supported by the Italian Foundation for Cancer research (FIRC).

REFERENCES

Bae,E., Calhoun,V.C., Levine,M., Lewis,E.B. and Drewell,R.A. (2002) Characterization of the intergenic RNA profile at abdominal-A and Abdominal-B in the *Drosophila* bithorax complex. *Proc. Natl Acad. Sci. USA*, **99**, 16847–16852.

Drewell,R.A., Bae,E., Burr,J. and Lewis,E.B. (2002) Transcription defines the embryonic domains of cis-regulatory activity at the *Drosophila* bithorax complex. *Proc. Natl Acad. Sci., USA*, **24**, 16853–16858.

Fahey,M.E., Moore,T.F. and Higgins,D.G. (2002) Overlapping antisense transcripts in the human genome. *Comp. Funct. Genomics*, **3**, 244–253.

Kiyosawa,H., Yamanaka,I., Osata,N., Kondo,S., Hayashizaki,Y., RIKENGER GROUP, GSL members (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.*, **13**, 1324–1334.

Lehner,B., Williams,G., Campbell,R.D. and Sanderson,C.M. (2002) Antisense transcripts in the human genome. *Trends Genet.*, **18**, 63–65.

Li,A.W. and Murphy,P.R. (2000) Expression of alternatively spliced FGF-2 antisense RNA transcripts in the central nervous system: regulation of FGF-2 mRNA translation. *Mol. Cell. Endocrinol.*, **170**, 233–242.

Rank,G., Prestel,M. and Paro,R. (2002) Transcription through intergenic chromosomal memory elements of the *Drosophila* bithorax complex correlates with an epigenetic switch. *Mol. Cell. Biol.*, **22**, 8026–8034.

Shendure,J. and Church,G.M. (2002) Computational discovery of sense–antisense transcription in the human and mouse genomes. *Genome Biol.*, **3**, 1–14.

Vanhee-Brossollet,C. and Vaquero,C. (1998) Do natural antisense transcripts make sense in eukaryotes? *Gene*, **211**, 1–9.

Yelin,R., Dahary,D., Sorek,R., Levanon,E.Y., Goldstein,O., Shoshan,A., Diber,A., Biton,S., Tamir,Y., Khosravi,R. *et al.* (2003) Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.*, **21**, 379–386.