AntiHunter: searching BLAST output for EST antisense transcripts

Giovanni Lavorgna¹,∗, Luca Sessa¹,², Alessandro Guffanti³, Lelio Lassandro³ and Giorgio Casari¹

¹Istituto Scientifico H. S. Raffaele, Via Olgettina 60, 20132 Milan, Italy, ²Dulbecco Telethon Institute IGB CNR Via Pietro Castellino 111, 80131 Naples, Italy and ³IFOM-FIRC Institute of Molecular Oncology, Via Adamiello 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.
The accuracy of AntiHunter was tested using genes, which had been previously shown 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. 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, the annotated sequence for which the antisense match was detected and its strand, the EST 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).

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-Katirae, Lucilla Luzi, Valerio Orlando, Alessandro Brozzi and Stefano Conflonieri for helpful comments and suggestions and to Gyorgy Simon for setting up the IFOM computer facility. L.S. was supported by grants Teledon 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


