Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W665–W668***
doi:10.1093/nar/gki448

AntiHunter 2.0: increased speed and sensitivity in searching BLAST output for EST antisense transcripts

Giovanni Lavorgna[1,*], Riccardo Triunfo[2], Federico Santoni[2], Ugo Orfanelli[1], Sara Noci[1], Alessandro Bulfone[1,3], Gianluigi Zanetti[2] and Giorgio Casari[1]

[1]DIBIT–San Raffaele Scientific Institute, Via Olgettina 58, 20132 Milan, Italy, [2]CRS4–Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna and [3]Bio-flag Srl, Parco Scientifico e Tecnologico POLARIS, Loc. Pixina Manna, 09010 Pula (CA), Italy

Received February 14, 2005; Revised and Accepted March 31, 2005

ABSTRACT

An increasing number of eukaryotic and prokaryotic genes are being found to have natural antisense transcripts (NATs). There is also growing evidence to suggest that antisense transcription could play a key role in many human diseases. Consequently, there have been several recent attempts to set up computational procedures aimed at identifying novel NATs. Our group has developed the AntiHunter program for the identification of expressed sequence tag (EST) antisense transcripts from BLAST output. In order to perform an analysis, the program requires a genomic sequence plus an associated list of transcript names and coordinates of the genomic region. After masking the repeated regions, the program carries out a BLASTN search of this sequence in the selected EST database, reporting via email the EST entries that reveal an antisense transcript according to the user-supplied list. Here, we present the newly developed version 2.0 of the AntiHunter tool. Several improvements have been added to this version of the program in order to increase its ability to detect a larger number of antisense ESTs. As a result, AntiHunter can now detect, on average, >45% more antisense ESTs with little or no increase in the percentage of the false positives. We also raised the maximum query size to 3 Mb (previously 1 Mb). Moreover, we found that a reasonable trade-off between the program search sensitivity and the maximum allowed size of the input-query sequence could be obtained by querying the database with the MEGABLAST program, rather than by using the BLAST one. We now offer this new opportunity to users, i.e. if choosing the MEGABLAST option, users can input a query sequence up to 30 Mb long, thus considerably improving the possibility to analyze longer query regions. The AntiHunter tool is freely available at http://bioinfo.crs4.it/AH2.0.

INTRODUCTION

Several examples of natural antisense transcripts (NATs) have been reported in prokaryotes and viruses, where they are found to regulate gene expression by affecting mRNA transcription, processing and translation (1). A growing number of endogenous antisense RNA transcripts have also been found in many eukaryotic organisms during recent years, with experimental evidence suggesting a functional role for them at a surprising variety of levels in gene regulation, including transcriptional interference (2), genomic imprinting (3,4), RNA interference (5), translational regulation (6), alternative splicing (7), X-inactivation (8) and RNA editing (9). To facilitate the *in silico* search of potential antisense transcripts, we have recently developed a software tool, AntiHunter, aimed at facilitating the identification of antisense expressed sequence tag (EST) transcripts within a given genomic region of interest (10,11).

We report here the availability of a new AntiHunter release. Thanks to an improved algorithm, this version of the program can detect a significantly larger number of antisense ESTs, while keeping to a minimum the number of false positives. Moreover, because of improvements in the underlying processing pipeline, it is now possible to analyze longer query sequences with significantly shorter response times.

SOFTWARE UPGRADE

The AntiHunter-processing pipeline uses a genomic sequence and a list of annotated transcripts of the genomic region as

*To whom correspondence should be addressed. Tel: +39 02 2643 4776; Fax: +39 02 2643 4767; Email: giovanni.lavorgna@hsr.it

input. This list includes transcript names, their beginning and ending positions plus their strand occurrence. Subsequently, it first runs the RepeatMasker program on the genomic sequence in order to filter out repeated sequences and then performs a BLASTN search of the resulting sequence in the selected EST database. Finally, it parses the BLAST output looking for antisense EST with respect to the annotated genes and reports the results to the user by email.

In order to gain information about the EST sequencing strand, the program uses the database annotation, i.e. 5′ or 3′, reported in the actual EST entry. However, since only a fraction of ESTs, 74% (18 111 572 versus 24 481 418, as of January 2005), possesses such information, a significant percentage of ESTs cannot be used in AntiHunter searches. For this reason, we tried to incorporate into our program a procedure that was able to take advantage of the information contained in these un-annotated ESTs. We found that the main reason for the missing EST strand annotation was that these ESTs belonged to random-primed and non-directionally cloned libraries (G. Lavorgna, unpublished data). Therefore, it was conceivable to attempt recovery of the missing information about the EST sequencing strand by looking at the splicing consensi located in proximity to the edge of the alignment of each EST exon with the genomic sequence. Splice donor and acceptor sites are GT–AG for the vast majority of introns (12). Thus, in AntiHunter 2.0, un-annotated ESTs spanning an intron are recognized quite reliably by the presence of these consensus sites, or their reverse-complementary sequence CT–AC, at the intron's border. It should be noted that the above procedure was already implemented in the previous AntiHunter version, but it was meant only to double check the source of the sequencing strand of already annotated EST and was not used to attempt an *ab initio* strand prediction.

Furthermore, in AntiHunter 2.0, we have made the width of the region that searched for the presence of splicing sites from the edge of the alignment, previously set to 5, user settable. By changing this parameter, the user can compensate for BLAST reported alignments whose edges go past the biologically correct one. This allows AntiHunter 2.0 to detect antisense transcripts missed by the previous version of the program. As an example, a detail of the BLAST alignment between a query genomic sequence from the MYCN locus (coordinates: chr2:16024168–16039977 from the release hg17 of the UCSC genome browser) versus the human EST AA609982 (subject sequence) is shown in Figure 1A.

The identified alignment goes beyond the intron/exon border: 11 genomic bases, shown in uppercase, are, indeed, spuriously aligned to the EST, making it difficult to identify the correct splicing sites. The specialized programs SIM4 (http://pbil.univ-lyon1.fr/sim4.php) correctly detects the alignment boundaries, as shown in Figure 1B. In AntiHunter 2.0, the splicing consensi are correctly identified by AntiHunter when using a value >11 for the width of the region searched for splicing consensi (parameter 'Bases_Searched_For_Splicing_Consensi'). However, it should also be noted that high value for this parameter will also increase the chance of detecting artifactual splicing sites.

The new capabilities of AntiHunter 2.0 were tested using known examples. In particular, we used the same test set we used to benchmark AntiHunter capabilities (http://bio.ifom-firc.it/ANTIHUNTER/ah_help.new.htm#function). It consists of 15 genomic regions, previously described in the literature, containing overlapping transcriptional units in mammalian genomes. It was previously shown that AntiHunter was able to pick up antisense transcripts from 14 out of these 15 loci (10). This time, we also measured the number of antisense

A

```
Score =  252 bits (127), Expect = 5e-63
Identities = 133/135 (98%)
Strand = Plus / Plus


Query: 7527 cgggggggagtaatggcttctgcgaaaagaaattccctcggctctagaagatctgtctgtg 7586
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 171  cgggggggagtaatggcttctgcgaaaagaaattccctcggctctagaagatctgtctgtg 230


Query: 7587 tttgagctgtcggagagccggtgcgtccccaccccaggctggggttcttctccaaagggt 7646
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 231  tttgagctgtcggagagccggtgcgtccccaccccaggctggggttcttctccaaagggt 290


Query: 7647 gcccCTGGAGGAAGA 7661
            ||||| ||| |||||
Sbjct: 291  gcccccggacgaaga 305
```

B

```
290  TGCCC         CCGGACGAAGATGACTTCTACTTCGGCGGCCCCGAC
     |||||<<<...<<<|||||||||||||||||||||||||||||||||||
7646 TGCCCCTG...TACCCGGACGAAGATGACTTCTACTTCGGCGGCCCCGAC
```

Figure 1. Parameterizing the value of constant 'Bases_Searched_For_Splicing_Consensi' in AntiHunter. The constant 'Bases_Searched_For_Splicing_Consensi' determines the number of bases located upstream and downstream of the edge of a BLAST alignment between a genomic and an EST sequence that are searched for in the presence of splicing consensi. It used to be set to a fixed value of 5 in the AntiHunter program. This low value made unfeasible the detection of alignments like those shown in (**A**), where up to 11 spurious bases (shown in boldface uppercase) are added at the edge of the alignment between a query genomic sequence from MYCN locus (coordinates: chr2:16024168-16039977 from the release hg17 of the UCSC genome browser) and the EST AA609982. The specialized programs SIM4 (http://pbil.univ-lyon1.fr/sim4.php) correctly detects the alignment boundaries of the alignment, as shown in (**B**). In AntiHunter 2.0, this hard-coded constant value has been parameterized, allowing the user to experiment with it: the splicing consensi are indeed correctly identified by AntiHunter when using a value >11.

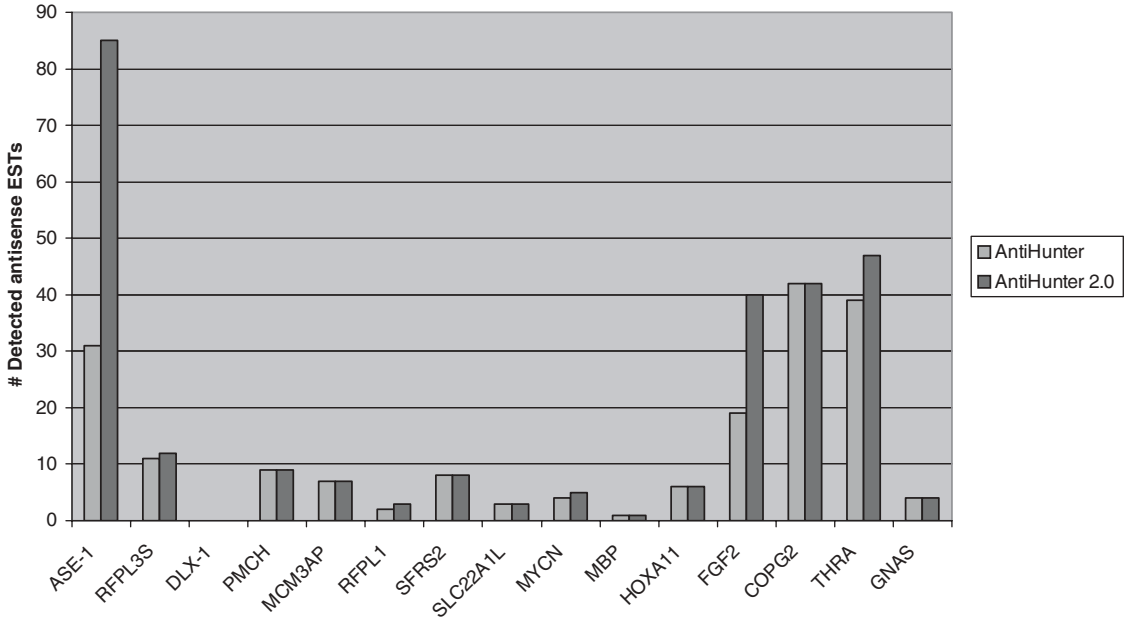

Figure 2. Benchmarking the performance of AntiHunter 2.0. The capability of AntiHunter 2.0 to detect EST antisense transcripts was compared with that of AntiHunter on a test case of 15 genomic regions, containing overlapping transcriptional units previously described in literature in mammalian genomes (for details see http://bioinfo.crs4.it/AH2.0/ah_help.new.html). As a result, AntiHunter 2.0 detected a significantly larger number, 272 versus 186, of antisense ESTs than the previous version of the program. The newly detected ESTs belonged to six different genomic loci (ASE-1, RFPL3S, RFPL1, MYCN, FGF2 and THRA).

transcripts reported by both versions of the program. As shown in Figure 2, AntiHunter 2.0 detected a total of 272 ESTs versus 186 ESTs detected by AntiHunter, yielding an increase in search sensitivity of >45%. The newly identified ESTs were found in 6 loci. In all cases except one, the newly reported ESTs were not reported before because they had no database annotation about their sequencing strand. In the remaining case, EST AA609982 from MYCN locus was identified because the search was run using the parameter 'Bases_Searched_For_Splicing_Consensi' set to a value of 20. It should also be noted that general-purpose resources that are not specially designed for this purpose could have been used in antisense transcription detection of these examples. Among these resources are the popular online genome browsers, such as the University of California Santa Cruz (UCSC, http://genome.ucsc.edu/) and Ensembl (http://www.ensembl.org/) browsers, in which the user can view a specific genomic locus with all cDNAs and ESTs aligned to it. The direction of full-length cDNAs and spliced ESTs is usually shown so that the user can determine whether there is an antisense overlap with a chosen gene. However, this choice can be rather problematic in the case of genomic regions larger than a few hundred kilobases, since details on the transcriptional orientation of the ESTs will be less easily discernible, especially in the case of extensively transcribed regions (i.e. regions with a large number of associated ESTs). In addition, these browsers do not present several key orientation parameters, such as poly(A) sequences or sites and database annotation. Therefore, they should be mainly used to view the results of AntiHunter in their genomic context. In this sense, they are an important complementary resource.

Since the newly implemented procedure had somewhat loosened the stringency of AntiHunter searches, we attempted to measure the resulting search background. To do this, we analyzed each of the newly found ESTs to determine if it was localized within no more than 1 kb of the previously identified ones. As a result, all the new ESTs fell within this range (data not shown), thus supporting the idea that they belonged to the same transcriptional unit as the old ones and that none of them was, indeed, artifactual.

SYSTEM UPGRADE

The underlying AntiHunter pipeline was upgraded in order to efficiently manage larger sizes of the query sequence and/or to offer quicker response times. This was mainly achieved by replacing the BLAST program with mpiBLAST, a freely available open source parallelization of NCBI BLAST, which permits BLAST queries to be processed on many nodes simultaneously (http://mpiblast.lanl.gov/). Up to eight nodes (16 CPUs) have been allocated for AntiHunter usage on the new system, depending on the estimate of the needed computing power. In addition, BLAST EST databases have been pre-split in 4, 8 and 16 pieces in order to save CPU cycles at running time. It is expected that BLAST search speed will scale up quasi-linearly with the number of the used CPUs on the new system, with the main reason for not having a completely linear scaling of the search speed being the time spent in rejoining the results from the split databases, an operation not-parallelized in mpiBLAST. Because of these system improvements, AntiHunter 2.0 can now process queries up to 3 Mb long, thus tripling the limit of the previous version.

As a further speed-up, the AntiHunter 2.0 interface now allows the user to disable the pre-processing of the input sequence with the RepeatMasker program (A. F. A. Smit, R. Hubleyand P. Green; RepeatMasker Open-3.0, 1996–2004, http://www.repeatmasker.org), a time-consuming step,

especially in case of large genomic sequences. We added this option since it is now possible to use web resources, like the UCSC genome browser (http://genome.ucsc.edu/), to obtain pre-masked genomic regions belonging to the whole genome of several organisms. These pre-processed regions can be used as input to AntiHunter; similarly, the user might have already masked once his/her sequence; therefore, no need for further masking.

AntiHunter searches, especially when performing intra-species comparisons, are expected to deal with sequences that differ only slightly as a result of sequencing or other similar errors. For this reason, several queries could be, in principle, handled by MEGABLAST, a less sensitive but up to 10 times faster program (13). Moreover, MEGABLAST is also able to efficiently handle much longer DNA sequences than the BLASTN program (http://www.ncbi.nlm.nih.gov/blast/megablast.shtml). In AntiHunter 2.0, we offer the possibility to analyze queries up to 30 Mb long using the MEGABLAST program. Our tests indicated a reasonable trade-off between the loss in search sensitivity and the increased size of the query sequence. We applied, in fact, MEGABLAST/AntiHunter to the identification of antisense transcripts in the same test set of Figure 2. Programs were able to detect 245 out of 272 antisense transcripts, with <10% loss in search sensitivity with regard to the mpiBLAST version (see Supplementary Material).

CONCLUSIONS

Identifying antisense transcripts embedded within genomes and understanding their function is a formidable and challenging task. We present a completely re-developed AntiHunter 2.0 web server, for high-throughput detection of antisense transcripts. By combining advances in the algorithm design and in the underlying server system, AntiHunter 2.0 maximizes the probability of identifying functional antisense transcripts. The novel features added to AntiHunter 2.0 also make this tool very powerful for identifying antisense molecules in long genomic intervals.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Stephen Altschul for advice on the usage of BLAST programs, Robert Hubley for providing an updated RepeatMasker module and two anonymous reviewers for constructive criticisms. This work was performed under the auspices of MIUR (Ministero dell'Istruzione, dell' Universita' e della Ricerca) with a research grant to G.L. (RBNE01N2ZE_004). Funding to pay the Open Access publication charges for this article was provided by MIUR.

Conflict of interest statement. None declared.

REFERENCES

1. Wagner,E.G., Altuvia,S. and Romby,P. (2002) Antisense RNAs in bacteria and their genetic elements. *Adv. Genet.*, **46**, 361–398.
2. Prescott,E.M. and Proudfoot,N.J. (2002) Transcriptional collision between convergent genes in budding yeast. *Proc. Natl Acad. Sci. USA*, **99**, 8796–8801.
3. Moore,T., Constancia,M., Zubair,M., Bailleul,B., Feil,R., Sasaki,H. and Reik,W. (1997) Multiple imprinted sense and antisense transcripts, differential methylation and tandem repeats in a putative imprinting control region upstream of mouse Igf2. *Proc. Natl Acad. Sci. USA*, **9**, 12509–12514.
4. Sleutels,F., Zwart,R. and Barlow,D.P. (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, **415**, 810–813.
5. Billy,E., Brondani,V., Zhang,H., Muller,U. and Filipowicz,W. (2001) Specific interference with gene expression induced by long, double-stranded RNA in mouse embryonal teratocarcinoma cell lines. *Proc. Natl Acad. Sci. USA*, **98**, 14428–14433.
6. 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.
7. Munroe,S.H. and Lazar,M.A. (1991) Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA. *J. Biol. Chem.*, **266**, 22083–22086.
8. Lee,J.T., Davidow,L.S. and Warshawsky,D. (1999) Tsix, a gene antisense to Xist at the X-inactivation centre. *Nature Genet.*, **21**, 400–404.
9. Kumar,M. and Carmichael,G.G. (1997) Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. *Proc. Natl Acad. Sci. USA*, **94**, 3542–3547.
10. Lavorgna,G., Dahary,D., Lehner,B., Sorek,R., Sanderson,C.M. and Casari,G. (2004) In search of antisense. *Trends Biochem. Sci.*, **29**, 88–94.
11. Lavorgna,G., Sessa,L., Guffanti,A., Lassandro,L. and Casari,G. (2004) AntiHunter: searching BLAST output for EST antisense transcripts. *Bioinformatics*, **20**, 583–585.
12. International Human Genome Sequencing Consortium (2001), Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
13. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.