BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 4 2006, pages 497–499
doi:10.1093/bioinformatics/bti754

Sequence analysis

GenoMiner: a tool for genome-wide search of coding and non-coding conserved sequence tags

Tiziana Castrignanò[1], Paolo D'Onorio De Meo[1], Giorgio Grillo[2], Sabino Liuni[2], Flavio Mignone[3], Ivano Giuseppe Talamo[1] and Graziano Pesole[2,3,*]

[1]Consorzio Interuniversitario per le Applicazioni di Supercalcolo per Universitàe Ricerca, CASPUR, Rome, Italy, [2]Istituto Tecnologie Biomediche, Sede di Bari, Consiglio Nazionale delle Ricerche, Italy and [3]University of Milan, Dipartimento di Scienze Biomolecolari e Biotecnologie, via Celoria 26, Milan 20133, Italy

Received on June 27, 2005; revised on September 14, 2005; accepted on October 28, 2005
Advance Access publication November 2, 2005
Associate Editor: Thomas Lengauer

ABSTRACT

Summary: GenoMiner is a software tool that searches for regions of similarity between user-submitted genome or transcript sequences and user-specified whole genome assemblies. The program then identifies conserved sequence tags (CSTs) in these homologous regions and provides a prediction of their coding or non-coding nature. The analysis is carried out through three steps: (1) definition of sequence regions homologous to the query sequence in the selected target genomes by a fast BLAT alignment; (2) identification of CSTs by a more sensitive BLAST-like alignment between the query and the homologous regions in the target genomes and (3) assessment of the coding or non-coding nature of detected CSTs through the computation of a suitable coding potential score. GenoMiner allows the user to search the query sequence against a number of vertebrate genome assemblies in a single run providing a user-friendly graphical output.

Availability: http://www.caspur.it/GenoMiner/. GenoMiner software and documentation is available from the authors upon request.

Contact: graziano.pesole@unimi.it

The increasing amount of nucleotide sequence data produced by large-scale sequencing projects aimed at deciphering the genome and transcriptome of a wide range of organisms requires the concurrent development of adequate bioinformatics tools for their functional annotation.

The search for statistically significant local alignments between a query sequence and a target database is by far the most heavily used approach in comparative sequence analysis (Miller *et al*., 2004). In particular, the alignment of a query sequence (e.g. mRNA) and a number of complete genome assemblies can prove extremely helpful in the definition of gene structure, for detecting homologous genome regions and for gene annotation.

Currently available tools, implemented in the major genome browsers such as UCSC (Karolchik *et al*., 2003) and Ensembl (Hubbard *et al*., 2005) efficiently construct alignments between query sequences and a single genome assembly. The algorithm presented here, named GenoMiner, is a novel implementation of the CSTminer software (Mignone *et al*., 2003; Castrignano *et al*., 2004). CSTminer was designed to identify conserved sequence tags (CSTs) (i.e. conserved sequences longer than 30 bp and expected occurrence $<10^{-5}$) through the comparison of two user-specified homologous sequences and to assess their coding or noncoding nature through the computation of a coding potential score (CPS) based on the evaluation of the peculiar evolutionary dynamics of protein coding sequences at both the nucleotide and amino acid levels. GenoMiner, that does not require previous knowledge of homologous sequences, has been designed to rapidly and efficiently detect CSTs from the comparison of a user-submitted query sequence and multiple genome assemblies by performing several simultaneous alignments of the first against the latter. It also assigns a CPS to each detected CST as described previously (Castrignano *et al*., 2004).

GenoMiner analysis is partitioned into three steps: (1) the best local alignments between the query sequence and the target genomes are identified by BLAT (Kent, 2002) and those alignments define one or more homologous regions in the selected target genomes; (2) CSTs are detected though a BLAST-like alignment of the query sequence and the homologous regions delimited previously using sensitive parameters; (3) a CPS is assigned to each detected CST as described in Castrignanò *et al*. (2004).

GenoMiner is available as a web tool where the user can paste or upload a query sequence in Fasta format with no length limitation and select one or more target genomes. Presently, seven vertebrate genomes collected in the Ensembl database are available for GenoMiner analysis. In order to increase the computational efficiency, the searches against genome assemblies, optionally masked for repetitive sequences, are spread across multiple distributed BLAT servers. BLAT hits obtained against each selected genome are then sorted according to the overlap degree between the query and the target sequences, and only the best hit for each genome is shown in the output, likely corresponding to the orthologous genome region. However, the user can ask for a full output showing all hits, independently of the overlapping degree, by clicking on the 'show all' link. The user can also browse the results and select the ones he judges more reliable by flagging the relevant check-box at the right-side of each panel. By pressing the 'show selected' link only flagged panels are shown in the output.

*To whom correspondence should be addressed.

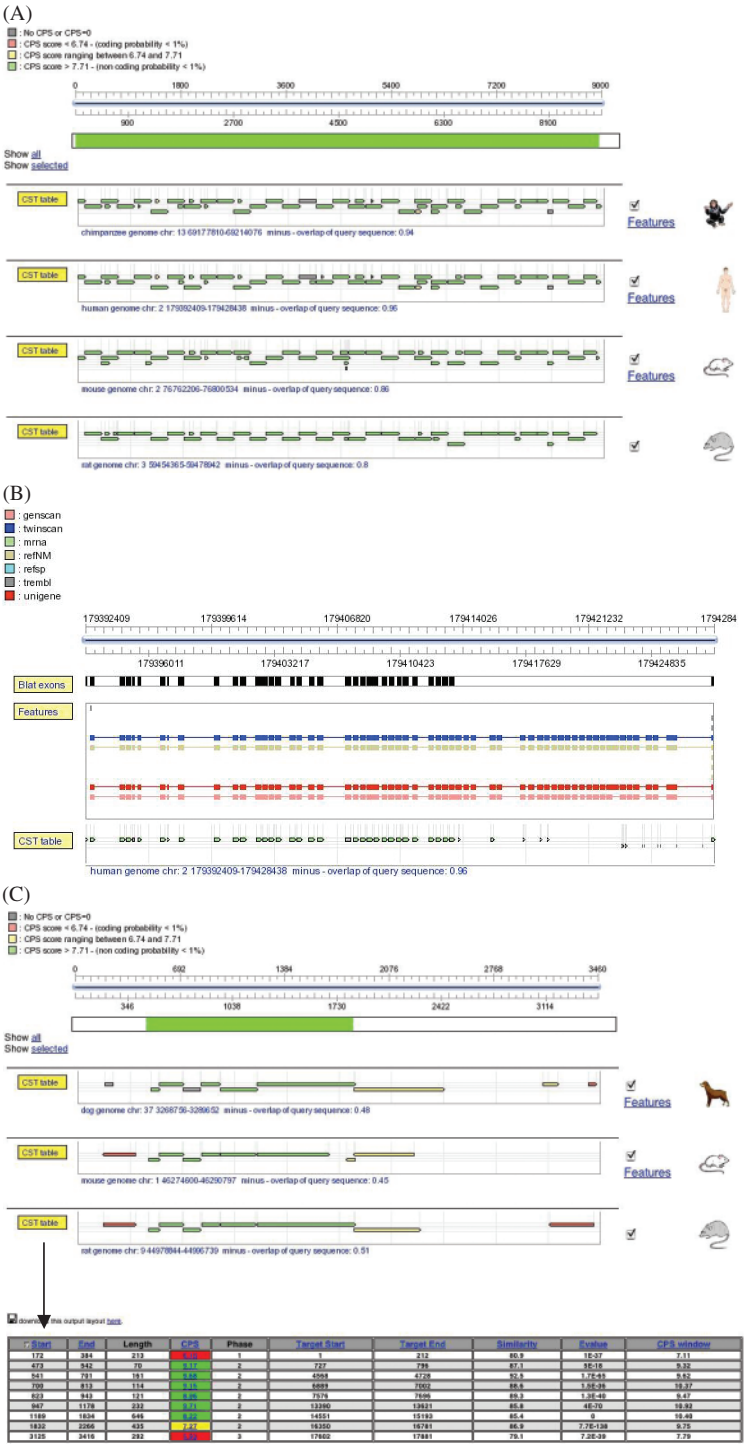

Fig. 1. (**A**) Snapshot of GenoMiner output obtained by aligning the first 9000 nt of dog mRNA XM_535982.1 with human, chimp, rat and mouse genomes. Below the scaling bar referring to the query sequence a summary plot of the predicted coding region is shown in green colour. Panels for each analyzed organism follow and CSTs identified are schematically shown and labeled in the relevant colour. Below each genome panel the chromosome number, the absolute coordinates of the matched genome region and the overlap percentage are reported. When the cursor is placed on a CST a tooltip appears showing relevant summary information (CPS, start, end and frame) and a popup window is invoked by clicking on the CST which shows more detailed information on the CST, including the pairwise alignment with the target genome. A link at the bottom of the popup window opens a plot of the CPS along the CST using a sliding window of 60 nt. (**B**) Clicking on the 'Features' link a new window opens showing a number of features (e.g. mRNA, proteins, Unigene, Genscan and Twinscan) mapped on the relevant genome region with CSTs displayed according to their genome coordinates. (**C**) GenoMiner output obtained by aligning human mRNA NM_014585 with mouse rat and dog genomes. The lower panel can be obtained by clicking on the 'CST Table' where a summary table of all detected CSTs is shown.

GenoMiner output consists of a graphical representation of the alignment between the query sequence and the CSTs detected in the target genomes. The user can visualize the full list of CSTs detected in each target genome and their overlapping annotated features, i.e. matches to known or predicted genes (Fig. 1B). A zoom facility has been also implemented to enlarge the output of selected regions, particularly useful in the case of long query sequences.

To allow rapid visual discrimination, CSTs belonging to the coding or non-coding class are shown in green or red, respectively. A minimum amount of sequence divergence, specified as 5%, is needed for reliable CPS computation. CSTs are labelled grey if aligned sequence blocks are identical or diverge <5%.

The GenoMiner application may be particularly useful for functional annotation of novel genomic or transcript sequences when their homologous counterparts in other genomes are not known. Novel genes or gene isoforms may be predicted when clustered coding CSTs are detected. For transcript sequences the coding region (CDS) can be determined with high reliability. This may prove powerful in larger-scale transcriptome analyses (Okazaki *et al*., 2002) and particularly for those transcripts with long 5′-UTRs which contain one or more upstream AUG, or which contain sequencing errors inducing frame-shifts.

Two sample applications of GenoMiner are shown in Figure 1. Figure 1A shows the alignment between the RefSeq dog mRNA XM_535982.1 (region 1–9000) and human, chimp, rat and mouse genomes. Despite the CDS annotation for this mRNA spanning positions 7115–19 495 a strong coding CPS is also observed in the annotated 5′-UTR in all genomes considered. Indeed, considering the features mapped on the matching human region (Fig. 1B) it is clear that the dog transcript falls in the coding region of the *ttn* gene (Entrez geneid: 7273) thus supporting the reliability of GenoMiner prediction. Indeed, a careful investigation of the dog mRNA entry revealed that the wrong CDS annotation was likely due to a frame-shift G-insertion (GGG instead of GG) at position 6228, that when removed restored the full ORF.

Figure 1C shows the alignment between the human mRNA coding for SLC40A1 protein and mouse, rat and dog genomes. In this case the annotated coding region (positions 431–2146) fits very well with mapped coding CSTs whereas non-coding CSTs are present in the 5′-UTRs. A further support for the prediction of the coding region in the transcript under investigation is given when multiple consecutive coding-labelled CSTs share the same coding-frame or phase. Indeed, in the case of SLC40A1 transcript all coding CSTs share the same phase (Fig. 1C, bottom panel). On the other hand, in the case of *ttn* transcript a phase shift can be observed in correspondence of the frame-shift (data not shown).

The GenoMiner web tool has been implemented on a pool of four 4-processor servers distributing BLAT searches. PHP scripts are used for internet submission and to plot the dynamic web results using GD libraries (version 2.0.28). A queue system is administrated by a Perl script querying a MySQL database (version 5.0.3-0).

ACKNOWLEDGEMENTS

We thank David Horner for valuable comments on the manuscript. This work was supported by Fondo Italiano Ricerca di Base (FIRB) projects ‘Bioinformatica per la Genomica e la Protomica’ and ‘Laboratorio Italiano di Bioinformatica’, MIUR Cluster C03/2000-CEGB, PON 2000–2006 Progetto BIG, D.D. 9/10/2002 n.1105 obiettivo 1, EU STREP project TRANSCODE and by AIRC.

Conflict of Interest: none declared.

REFERENCES

Castrignano,T. *et al*. (2004) CSTminer: a web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison. *Nucleic Acids Res*., **32** (Web Server issue), W624–W627.

Hubbard,T. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33** (Database issue), D447–D453.

Karolchik,D. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.

Mignone,F. *et al*. (2003) Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. *Nucleic Acids Res*., **31**, 4639–4645.

Miller,W. *et al*. (2004) Comparative genomics. *Annu. Rev. Genomics Hum. Genet*., **5**, 15–56.

Okazaki,Y. *et al*. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.