BIOINFORMATICS APPLICATIONS NOTE

Vol. 20 no. 16 2004, pages 2848–2850
doi:10.1093/bioinformatics/bth287

RRE: a tool for the extraction of non-coding regions surrounding annotated genes from genomic datasets

F. Lazzarato[1], G. Franceschinis[1], M. Botta[2], F. Cordero[3] and R. A. Calogero[3,*]

[1]Department of Informatics, University of Piemonte Orientale 'Amedeo Avogadro', Spalto Marengo 33, Alessandria, Italy, [2]Department of Informatics University of Torino, Corso Svizzera 185-10149, Torino, Italy and [3]Department of Clinical and Biological Sciences University of Torino, Regione Gonzole 10, 10043 Orbassano (TO), Italy

Received on November 3, 2003; revised on March 3, 2004; accepted on April 17, 2004
Advance Access publication April 29, 2004

ABSTRACT

Summary: RRE allows the extraction of non-coding regions surrounding a coding sequence [i.e. gene upstream region, 5′-untranslated region (5′-UTR), introns, 3′-UTR, downstream region] from annotated genomic datasets available at NCBI.

Availability: RRE parser and web-based interface are accessible at http://www.bioinformatica.unito.it/bioinformatics/rre/rre.html

Contact: raffaele.calogero@unito.it

INTRODUCTION

High-throughput techniques (e.g microarrays, SAGE, etc.) now enable biologists to obtain a vast amount of gene function data. Since the genomes sequenced are rapidly increasing in number, biologists studying the functional behaviour of transcription factors or regulative mechanisms tightly bound to transcription need an easy access to potential regulative regions annotated in sequenced genome datasets to investigate their composition and look for correlations with transcription profile experiments (Sudarsanam *et al*., 2002). This information can be easily extracted on a single-gene basis at NCBI (www.ncbi.nih.gov) and ENSEMBL (www.ensembl.org) databases. However, users wishing to investigate the regulative characteristics of hundreds of differentially expressed genes must write their own extraction program/script, extract potential promoter regions from annotated genomic databases using, e.g. the TRASER database (http://genome-www6.stanford.edu/cgi-bin/Traser/traser, only *Homo sapiens* data), or search not-redundant databases with specifically dedicated tools, such as INCLUSIVE (Thijs *et al*., 2002) and PEG (Zhang and Zhang, 2001).

RRE (Retrieval of Regulative Regions) is a JAVA application for the extraction of any genomic region [i.e. gene upstream region, 5′-untranslated region (5′-UTR), introns, 3′-UTR, gene downstream region] surrounding annotated coding sequence and its upload on a MySql database.

SYSTEMS AND METHODS

Data extraction

Data are routinely downloaded from NCBI with an automatic robot based on CURL (curl.haxx.se). This is a command line tool for transferring files with URL syntax, supporting FTP, FTPS, HTTP, HTTPS, GOPHER, TELNET, DICT, FILE and LDAP. For each chromosome of a sequenced genome, the downloader retrieves files with 'fa' (FASTA file) and 'gbs' (GBS file) extensions containing contig sequences in FASTA format and gene annotations, respectively. A JAVA application (RRE parser) then extracts from the GBS files the feature keys GENE, CDS and mRNA, together with the official gene symbol, product description and location on the direct or complement strand. This information is combined with the chromosome number and the DNA contig accession number extracted from the GBS files. Gene size, first-transcribed nucleotide and coding sequence location in the FASTA files are easily mapped with the feature keys GENE, mRNA and CDS, respectively. Starting from this information, potential upstream regulative regions (upstream) are extracted together with 5′-UTR, introns/exons located outside and inside the CDS, and 3′-UTR and DNA sequences located downstream from the GENE end. Furthermore, upstream and downstream region size can be defined by the user (10 kb upstream and 1 kb downstream are the default values in our implementation); if the intergenic spacer is smaller than the size defined by the user, a size preference rule is applied. In this preference rule the user-defined size of the upstream region prevails

[*]To whom correspondence should be addressed.

over the downstream region. Therefore, if an intergenic region between gene A and B is smaller than the user-defined size of the upstream region, all the intergenic sequence will be assigned as 'upstream from gene B' and the downstream region will be considered as 'missing area' in the RRE log file. The extracted data can be saved in FASTA, HTML, XML format or used to populate a MySql database. A log file containing information related to the data extraction is also generated.

Web implementation

A Spitfire server (spitfire.web.cern.ch) providing a grid-enabled middleware service for access to relational databases is used to access to the MySql DBMS populated by the parser, and the data generated by the parser are integrated in the DBMS with Locus Link ID and orthologues annotation. (http://www.ncbi.nlm.nih.gov/HomoloGene/). The output of an RRE query can be obtained as an HTML table by including in each row genomic contig ID, gene official symbol, gene description, Locuslink ID, chromosomal location, range of extracted nucleotides, cytogenetic locus, extracted sequence and organism taxonomy ID. In the HTML output, each sequence feature is hyperlinked to the corresponding NCBI database. An output in FASTA format can also be obtained. In this case, annotations are associated with each retrieved sequence as: >locuslink ID|Official symbol|Cytogenetic locus|contig ID|Taxonomy ID|range of extracted nucleotides|sequence feature (e.g. 5′-UTR, upstream, first intron, etc.)| gene description. Two modified FASTA formats (Melina and Consensus) can be used to facilitate the submission of RRE output data to analysis by online tools capable of discovering conserved motifs in a set of sequences. These modified FASTA files are suitable as input for Melina (Poluliakh *et al*., 2003), Patsearch (Grillo *et al*., 2003) and CONSENSUS programs (Hertz and Stormo, 1999). The standard FASTA output of RRE, on the other hand, is suitable for the direct submission of data to MotifSampler (Thijs *et al*., 2001) and other multiple sequence alignment tools (e.g.ClustalW, Higgins *et al*., 1994; pipmaker, http://bio.cse.psu.edu/cgi-bin/pipmaker?basic).

We are currently maintaining up-to-date *H.sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Plasmodium falciparum* and *Saccharomyces cerevisiae* datasets. A web interface to the RRE database is accessible at http://www6.unito.it:8443 if a valid certificate is installed in the browser (instructions for obtaining a certificate are provided at https://www.bioinformatica.unito.it/bioinformatics/rre/rre.html

DISCUSSION

The core of our application is the RRE parser. Since this is a JAVA application (see RRE web page for downloading/installation instructions), it is portable and particularly suitable for users wishing to generate genome-wide, specific sequence-feature datasets (i.e. putative promoters, first non-coding exon, all introns of a specific chromosome or contig, etc.). The tool is easy to use and quite efficient. On a Pentium III 2 GHz running Linux, extraction of all gene features of the annotated human genome takes < 30 min.

The RRE database, like other web-based interfaces (http://www.ensembl.org/EnsMart/; http://bio.ifom-firc.it/PROM_MACHINE/index.html; http://genome-www6.stanford.edu/cgi-bin/Traser/traser), is used to retrieve annotated putative regulative regions (i.e. putative promoters, 5′-UTRs, 3′-UTRs) as well as the non-coding regions linked to orthologue annotations. Extraction of sequence features related to orthologues is of assistance in the identification of conserved transcriptional elements, since phylogenetic footprinting dramatically improves the predictive selectivity of bioinformatics approaches to the analysis of promoter sequences (Lenhard *et al*., 2003). Couples of orthologous putative promoter sequences extracted using RRE queries can be directly used as input for the ConSite tool developed by Lenhard (http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite/). Some queries available at the RRE database, indeed, are not answered by other tools (e.g. retrieval of introns/exons located outside the coding sequence, extraction of the sequence features of genes located in the neighbourhood of a selected gene, retrieval of the sequence features of genes located within a specific contig, etc.). Furthermore, access to RRE data via Spitfire has the double advantage of securing sensitive data and avoiding unauthorized access by defining access policies for authorized users (identified by an X509 certificate). In conclusion, the RRE parser/database is a valuable tool for the rapid generation and updation of genome-wide datasets of non-coding regions and their association with specific groups of genes (e.g. co-regulated genes derived from microarray experiments), as well as for the quick and easy generation of broad collections of specific sequence features for comparative genomic analysis (Lane *et al*., 2001; Alvarez *et al*., 2002).

FUTURE DEVELOPMENTS

To enlarge the range of application of the RRE database, we have begun to integrate data from microarray transcriptional studies related to cancer. Extraction of sequence features related to genes containing estrogen (ER)-responsive elements (Cicatiello *et al*., 2004; Frasor *et al*., 2003) is already implemented in RRE, as well as the extraction of features related to genes differentially expressed between $ER^{+}$and ER^{-} breast cancers (Sotiriou *et al*., 2003; van 't Veer *et al*., 2002). Lung, colon and ovarium microarray cancer data will soon be implemented. We are also finishing the implementation of queries allowing the retrieval, starting from a user-defined list of genes annotated on human, mouse or rat genomes, of the location and significance score for all the putative transcription elements annotated, using TRANSFAC alignment matrices and PATSER program (Hertz and Stormo, 1999), in the region −2000/+500 with respect to the first transcribed nucleotide. Implementation of analysis tools to identify groups of transcription elements (motifs) co-evolving within phylogenetically related organisms will also be addressed.

ACKNOWLEDGEMENTS

F.L. is the recipient of a fellowship supported by a PRIN 2002053274 grant. This work was also partially supported by PRIN 2001057147, FIRB RBAU01JTHS and RBNE0157EH grants from the Italian Ministry for the Universities and Scientific and Technical Research (MIUR).

REFERENCES

Alvarez Martinez,C.E., Binato,R., Gonzalez,S., Pereira,M., Robert,B. and Abdelhay,E. (2002) Characterization of a Smad motif similar to *Drosophila* mad in the mouse Msx 1 promoter. *Biochem. Biophys. Res. Commun.*, **291**, 655–662.

Cicatiello,L., Scafoglio,C., Altucci,L., Cancemi,M., Natoli,G., Facchiano,A., Iazzetti,G., Calogero,R., Biglia,N., De Bortoli,M. *et al*. (2004) A genomic view of estrogen actions in human breast cancer cells by expression profiling of the hormone responsive transcriptome. *Mol. Endocrinol.*, **32**, 719–775.

Frasor,J., Danes,J.M., Komm,B., Chang,K.C., Lyttle,C.R. and Katzenellenbogen,B.S. (2003) Profiling of estrogen up- and down-regulated gene expression in human breast cancer cells: insights into gene networks and pathways underlying estrogenic control of proliferation and cell phenotype. *Endocrinology*, **144**, 4562–4574.

Grillo,G., Licciulli,F., Liuni,S., Sbisa,E. and Pesole,G. (2003) PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res.*, **31**, 3608–3612.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Higgins,D., Thompson,J., Gibson,T., Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Lane,R.P., Cutforth,T., Young,J., Athanasiou,M., Friedman,C., Rowen,L., Evans,G., Axel,R., Hood,L. and Trask,B.J. (2001) Genomic analysis of orthologous mouse and human olfactory receptor loci. *Proc. Natl Acad. Sci., USA*, **98**, 7390–7395.

Lenhard,B., Sandelin,A., Mendoza,L., Engstrom,P., Jareborg,N. and Wasserman,W.W. (2003) Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.*, **2**, 13.

Poluliakh,N., Takagi,T. and Nakai,K. (2003) MELINA: motif extraction from promoter regions of potentially co-regulated genes. *Bioinformatics*, **19**, 423–424.

Sotiriou,C., Neo,S.-Y., McShane,L.M., Korn,E.L., Long,P.M., Jazaeri,A., Martiat,P., Fox,S.B., Harris,A.L. and Liu,E.T. (2003) Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl Acad. Sci., USA*, **100**, 10393–10398.

Sudarsanam,P., Pilpel,Y. and Church,G.M. (2002) Genome-wide co-occurrence of promoter elements reveals a *cis*-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae. Genome Res.*, **12**, 1723–1731.

Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouzé,P. and Moreau,Y. (2001) A higher order background model improves the detection of regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113–1122.

Thijs,G., Moreau,Y., De Smet,F., Mathys,J., Lescot,M., Rombauts,S., Rouze,P., De Moor,B. and Marchal,K. (2002) INCLUSive: Integrated Clustering, upstream sequence retrieval and motif sampling. *Bioinformatics*, **18**, 331–332.

van 't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Zhang,T. and Zhang,M. (2001) Promoter extraction from genebank (PEG): automatic extraction of eukaryotic promoter sequences in large sets of genes. *Bioinformatics*, **17**, 1232–1233.