BIOINFORMATICS APPLICATIONS NOTE

Vol. 19 no. 12 2003, pages 1587–1588
DOI: 10.1093/bioinformatics/btg196

ESTWeb: bioinformatics services for EST sequencing projects

Apuã C. M. Paquola, Milton Y. Nishyiama Jr, Eduardo M. Reis, Aline M. da Silva and Sergio Verjovski-Almeida*

Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, 05508-900, São Paulo, SP, Brazil

Received on September 11, 2002; revised on November 3, 2002; accepted on March 2, 2003

ABSTRACT

Summary: ESTWeb is an internet based software package designed for uniform data processing and storage for large-scale EST sequencing projects. The package provides for: (a) reception of sequencing chromatograms; (b) sequence processing such as base-calling, vector screening, comparison with public databases; (c) storage of data and analysis in a relational database, (d) generation of a graphical report of individual sequence quality; and (e) issuing of reports with statistics of productivity and redundancy. The software facilitates real-time monitoring and evaluation of EST sequence acquisition progress along an EST sequencing project.

Availability: http://bioinfo.iq.usp.br/estweb

Contact: verjo@iq.usp.br

Supplementary information: http://bioinfo.iq.usp.br/estweb

Expressed sequence tags (ESTs) (Adams *et al.*, 1993) have proved to be an extremely valuable resource for high-throughput gene discovery. As the sequence of the human genome approaches completion it is becoming apparent that identification of genes can only be made with confidence by mapping ESTs and full-length mRNA sequences onto the genomic sequence (Lander *et al.*, 2001; Venter *et al.*, 2001) . This brings renewed importance to the acquisition of large sets of EST sequences (Camargo *et al.*, 2001) obtained from mRNA extracted from different tissues and at different stages of the cell life cycle, in order to cover the complete transcriptome of an organism. Large-scale EST sequencing projects are usually carried out by consortia of laboratories, implying a need for bioinformatics tools to facilitate uniform data processing, data exchange among labs and to store project data at a centralized site. ESTWeb is a software package designed for this purpose.

EST PROCESSING, STORAGE AND REPORTS

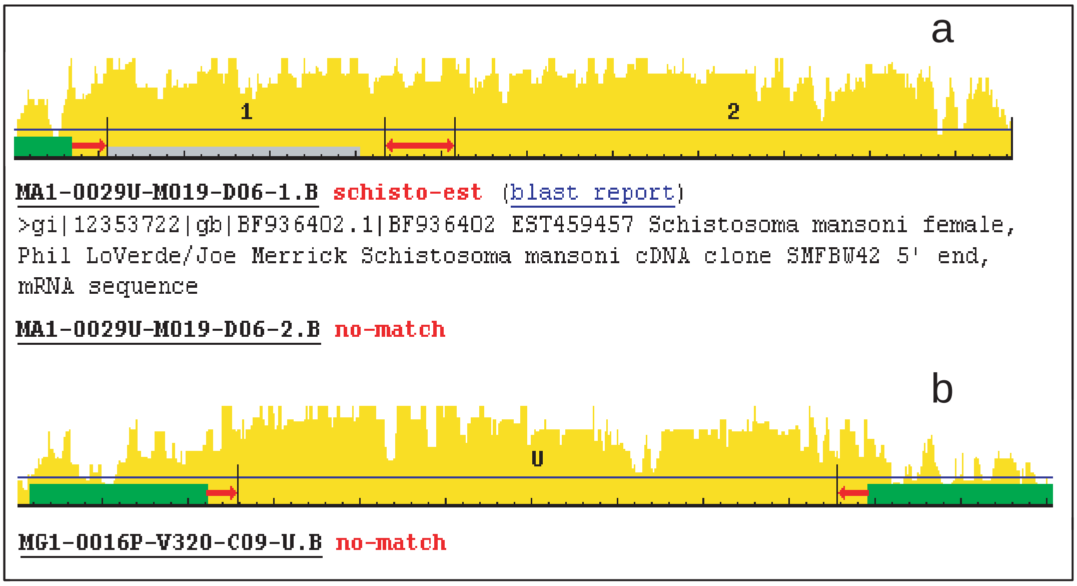
The user uploads a zipfile containing typically 96 chromatograms generated in an automated DNA-sequencer and originated from a single library. Reads are base-called with phred (Ewing *et al.*, 1998) and cloning vectors and adaptors (or primers) are identified with cross-match (http://www.phrap.org/). A graphical report is generated (Fig. 1). Segments not identified as vectors or primers, having at least 75 bases within a window of 100 bases with phred quality higher than a specified threshold are considered candidate cDNA inserts. Low-quality ends are trimmed by using the same 100-bases searching window together with an algorithm that anchors the accepted cDNA ends to primers if they are present within the sequence (see Supplementary information). The EST processing pipeline allows for two chimeric clones within one single read to be correctly handled and split, if they include either a stretch of poly-A sequence, a primer or an adaptor between the two chimeric inserts (Fig. 1a).

Sequences are then compared, with BLASTN (Altschul *et al.*, 1990), to a pre-categorized, project-specific database. Matching sequences are indicated in the graphical report as gray bars (Fig. 1). Each EST sequence along with all annotations is stored in the relational database. An e-mail with redundancy statistics and graphical quality reports is generated and is sent to the user. A schematic overview of ESTWeb dataflow is provided as supplementary information. A project progress report web page with cumulative sequence acquisition statistics is updated every day. The outputs of ESTWeb are trimmed and screened EST sequences. In a gene index project, these sequences must be further assembled into contigs and annotated for possible function. These tasks are not covered by the current version of ESTWeb.

ESTWeb database is designed to store a comprehensive set of information related to cDNA generation, such as mRNA sources and isolation protocols, primers used in reverse transcription reactions (Adams *et al.*, 1993) and in low-stringency RT-PCR (Dias-Neto *et al.*, 2000), PCR protocols and cloning vectors used in the project. With the above information, the database provides a link between the sequence and the physical clone. These clones can be used in many downstream applications such as cDNA

*To whom correspondence should be addressed.

Downloaded from bioinformatics.oxfordjournals.org by guest on September 15, 2011



MA1-0029U-M019-D06-1.B schisto-est (blast report)
>gi|12353722|gb|BF936402.1|BF936402 EST459457 Schistosoma mansoni female,
Phil LoVerde/Joe Merrick Schistosoma mansoni cDNA clone SMFBW42 5' end,
mRNA sequence

MA1-0029U-M019-D06-2.B no-match

MG1-0016P-V320-C09-U.B no-match

Fig. 1. Graphical reports for three EST sequences extracted from two sequence reads. Examples are taken from the '*Schistosoma mansoni* EST Genome Project'. (http://bioinfo.iq.usp.br/schisto/) where cDNA is generated with 18 to 20-mer primers and the low-stringency RT-PCR technique (Dias-Neto *et al*., 2000). (a) Two chimeric inserts (1) and (2) were identified by ESTWeb within a single EST read; sequence was split and marked as reads MA1-0029U-M019-D06-1.B and MA1-0029U-M019-D06-2.B. (b) An unique insert (U) is identified as read MG1-0016P-V320-C09-U.B. Minor ticks above the horizontal line mark 10 bp, major ticks mark 50 bp. Yellow vertical bars represent phred quality score for each base; the scale range is 0 to 50 (*Y*-axis not shown). A continuous horizontal line is at the quality threshold of phred score 15. Cloning vectors are shown as green bars and primers are indicated by red arrows. Gray bar indicates the matching region of the best hit obtained by a BLAST search against a public database; accession number(s) and annotation(s) of the match(es) are shown below the graph.

microarray construction and large scale analysis of gene expression.

IMPLEMENTATION

ESTWeb runs on Unix machines with Apache-CGI-Perl environment and uses PostgreSQL database server. In order to prevent server crashes, the EST processing program is not immediately started from the web script, but is instead put in an execution queue that limits the number of programs running simultaneously. This queue is managed by GNU Queue program and makes ESTWeb extremely stable even on average size PC computers.

APPLICATIONS

ESTWeb was developed in our laboratory for 'The *Schistosoma mansoni* EST Genome Project' which involves a network of nine laboratories in the State of São Paulo (http://bioinfo.iq.usp.br/schisto/). It runs on a Linux system installed on a 933 MHz Pentium-III with 512 MB RAM and 70 GB disk. Subsequently, ESTWeb was put into the present general format and is being used in another sequencing project in the laboratory of Dr. Suely L. Gomes (from our Department) that is generating ESTs from the fungus *Blastocladiella emersonii*.

ACKNOWLEDGEMENTS

We thank FAPESP, Fundação de Amparo a Pesquisa do Estado de São Paulo and CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil, for funding.

REFERENCES

Adams,M.D., Soares,M.B. *et al.* (1993) Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nature Genet.*, **4**, 373–380.

Altschul,S.F., Gish,W. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Camargo,A.A., Samaia,H.P. *et al.* (2001) The contribution of 700 000 'ORF sequence tags' to the definition of the human transcriptome. *Proc. Natl Acad. Sci. USA*, **98**, 12103–12108.

Dias-Neto,E., Garcia-Correa,R. *et al.* (2000) Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl Acad. Sci. USA*, **97**, 3491–3496.

Ewing,B., Hillier,L. *et al.* (1998) Base-calling of automated sequencer traces using Phred I: Accuracy assessment. *Genome Res.*, **8**, 175–185.

Lander,E.S., Linton,L.M. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Venter,J.C., Adams,M.D. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Downloaded from bioinformatics.oxfordjournals.org by guest on September 15, 2011