PHOREST: a web-based tool for comparative analyses of expressed sequence tag data

Ahrén, Dag; Troein, Carl; Johansson, Tomas; Tunlid, Anders

Published in:
Molecular Ecology Notes

DOI:
10.1111/j.1471-8286.2004.00613.x

2004

Molecular Ecology Notes (2004) **4**, 311–314 doi: 10.1111/j.1471-8286.2004.00613.x

PROGRAM NOTE

PHOREST: a web-based tool for comparative analyses of expressed sequence tag data

DAG AHREN,*‡ CARL TROEIN,† TOMAS JOHANSSON* and ANDERS TUNLID*
**Department of Microbial Ecology, Ecology Building and †Department of Theoretical Physics, Sölvegatan 14A, Lund University, SE-223 62 Lund, Sweden*

Abstract

Comparative analysis of expressed sequence tags is becoming an important tool in molecular ecology for comparing gene expression in organisms grown in certain environments. Additionally, expressed sequence tag database information can be used for the construction of DNA microarrays and for the detection of single nucleotide polymorphisms. For such applications, we present PHOREST, a web-based tool for managing, analysing and comparing various collections of expressed sequence tags. It is written in PHP (PHP: Hypertext Preprocessor) and runs on UNIX, Microsoft Windows and Macintosh (Mac OS X) platforms.

Keywords: comparative analysis, expressed sequence tags, software

Received 14 September 2003; revision received 3 December 2003; accepted 9 January 2004

Expressed sequence tag (EST) analyses are based on single-pass, large-scale DNA sequencing of reverse-transcribed messenger RNA (cDNAs) and have become a widely used approach for transcriptome analyses within functional genomics. Genes expressed under given growth conditions or at various developmental stages are characterized by EST sequencing and, due to the redundancies of transcripts, the expression levels can be inferred (i.e. transcript profile) (Qutob *et al*. 2000; Davey *et al*. 2001). There are currently more than 600 different species with ESTs publicly available through dbEST (Boguski *et al*. 1993). Within molecular ecology, EST database information can be used for the construction of DNA microarrays and for the detection of single nucleotide polymorphisms (Picoult-Newberg *et al*. 1999; Gibson 2002; Oleksiak *et al*. 2002).

In this study we present PHOREST, which is a novel web-based tool for comparative studies across multiple EST projects (EST libraries). The tool is designed to facilitate the storage, handling and mining of EST sequence data. Projects can easily be shared between multiple users via an intranet or the Internet and can be made public with PHOREST features by adding a guest user. PHOREST differs from other software packages specializing in EST data analysis, such as RED (Everitt *et al*. 2002), XGI (Waugh *et al*. 2000), STACKPACK (Miller *et al*. 1999) and PIPEONLINE (Ayoubi *et al*. 2002), in that the comparative approach is well developed and that management is possible through a web interface, requiring a minimum of computational skills. PHOREST can handle many EST projects simultaneously and compare the redundancy levels (i.e. the frequencies of a given transcript in the different EST projects). PHOREST also greatly facilitates the exchange and collaboration possibilities between researchers by using the web-based approach. By having restrictions on various viewing and editing options the security of the data is ensured. The system also allows for several different databases running on the same PHOREST installation. The users will only see the databases to which they have access. This makes it possible to have PHOREST as a service available to many research groups from a central server. The data analysed and maintained through PHOREST can be used for various research areas, for example to generate a uniset for construction of DNA microarrays. In addition, the results can be exported in a number of formats to facilitate external analysis of the PHOREST data. For example, it is possible to export the annotations from PHOREST into a pathological format (Karp *et al*. 2002) which can then be directly used for metabolic pathway reconstruction.

Correspondence: Dag Ahren. ‡Present address: European Bioinformatic Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Fax: + 44 1223494471; E-mail: ahren@ebi.ac.uk

PHOREST main view

To compare various collections of EST sequences, PHOREST can handle information from more than one EST database and each database can contain information from several EST projects (corresponding to one biological sample, cDNA library). After logging into PHOREST, the databases and EST projects to which the user has access are displayed. Each database has a statistics page calculated by PHOREST which provides information on the number of clones uploaded into each EST project, the number of contigs after assembly, the number and relative fraction of singletons, orphans and low-, medium- and high-score clones after assembly and BLAST similarity search. The page is of use both for EST projects where sequencing is not yet finished (validation of quality) as well as computing summary statistics once the EST sequencing is finished.

The main page of the PHOREST web interface shows a list of EST clones from one or several of the EST projects in the selected database (Fig. 1, no. 1). The top frame contains a search option that can be used for filtering and displaying subsets of the EST clones or contigs (Fig. 1, no. 3). For example, it is possible to search and then display ESTs with specific clone names, sizes (sequenced base pairs), specific score values (from BLAST searches), GenBank descriptions, cluster sizes and redundancy levels. The search profile can be reviewed and modified using the 'Advanced search' option. The results can either be listed as individual clones ('Listed by clones') or as contigs ('One clone per cluster') (Fig. 1, no. 7). The contig list provides a uniset of unique transcripts within the selected project(s). For the analysis of a subset of the ESTs/contigs there is a function for selection that can be activated by using the flag hyperlink (Fig. 1, no. 6). Once a subset of clones/contigs has been flagged it can be further analysed, exported in FASTA format or as a tab-delimited table suitable for Microsoft EXCEL and other programs. The table includes information on the transcript profiles in the various projects, BLAST scores, the GenBank description lines and Accession nos as well as results from functional annotations. In addition, users of PHOREST can

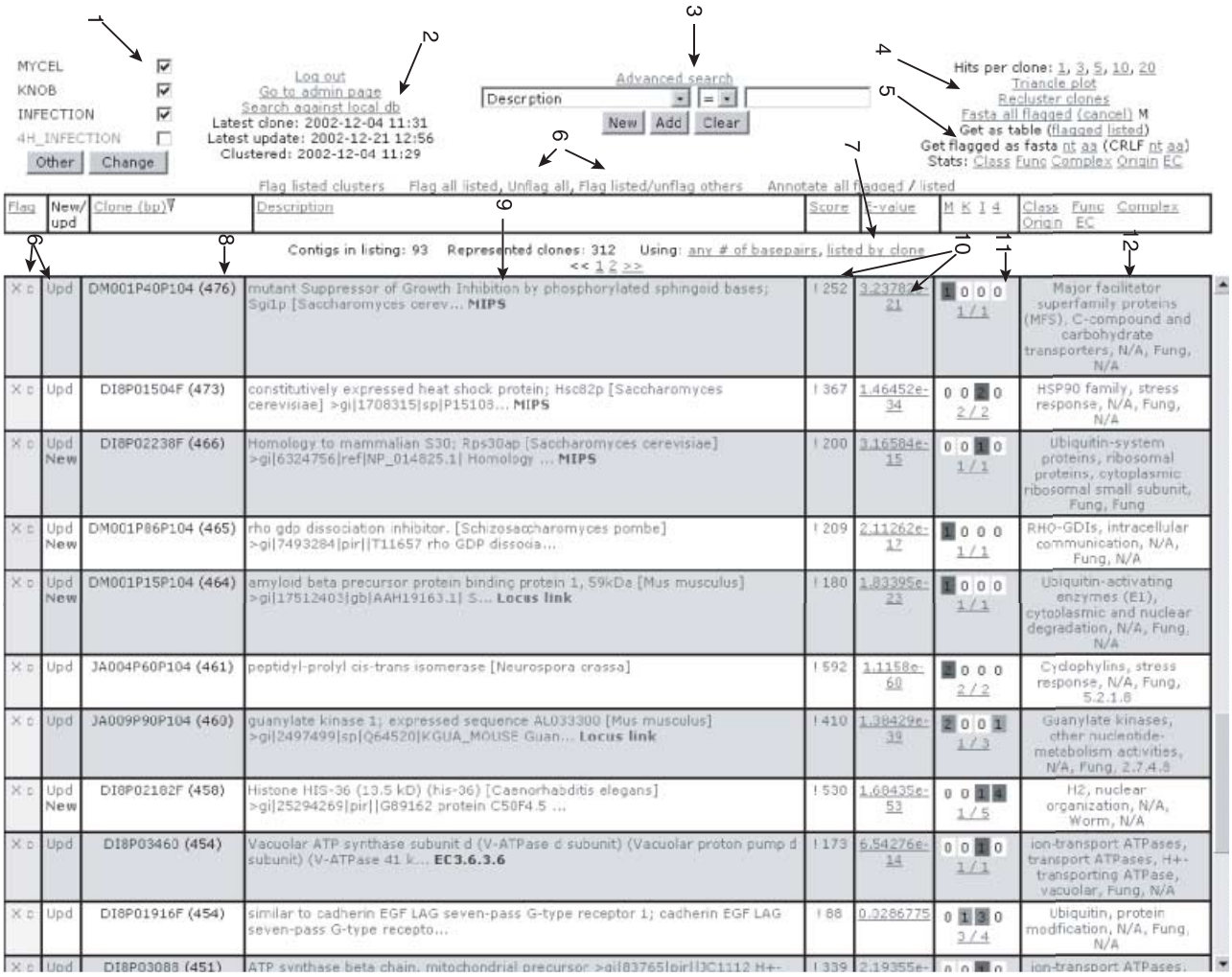

Fig. 1 The PHOREST web interface. See text for explanation.

perform BLAST (Altschul *et al*. 1990) searches against any locally stored sequence database (both public such as nr. from GenBank and custom databases) (Fig. 1, no. 2). The PHOREST software will detect whether it is a protein or nucleotide database and run the appropriate BLAST program.

In the lower frame, either all EST clones ('Listed by clones') or only one clone per cluster (contig) ('One clone per cluster') can be listed (Fig. 1, no. 7) as an html table. The columns of the table contain information on the clone name and the number of sequenced base pairs (in parentheses) (Fig. 1, no. 8), descriptions from homology searches (Fig. 1, no. 9) and the similarity score and the expected frequency value of the top hit from the BLAST search (Fig. 1, no. 10) (Altschul *et al*. 1990). The EST clones can be sorted based on the information in any of these columns. The EST sequences are automatically reblasted as new updates of the GenBank nonredundant (nr) database (10) are released. In addition, clones can be manually selected for updates (Fig. 1, no. 6). New top hits are indicated by red text ('New') and sequences with new top hits can be listed separately by asking for 'Only new' sequences in the search function available in the top frame (Fig. 1, no. 3).

Clustering and annotation

Results from the clustering analyses using the CONTIG ASSEMBLY PROGRAM (CAP) (Huang 1992) are shown as the number of clones in the contig (Fig. 1, no. 11). The column with the clustering data also provides a link to a page showing the alignment of the sequences in the cluster (Fig. 1, no. 11).

The ESTs are annotated according to the four categories used by MIPS (Mewes *et al*. 1997): protein classes, functional classes, protein complex classes and EC numbers. A fifth category (origin) is used to describe the type of sample (tissue, developmental stage or species) (Fig. 1, no. 12). Linked to this column is a page that gives a list of BLAST hits for the six possible open reading frames of the EST clone. The description column provides links to the GenBank nr database, the ExPaSy ENZYME database, LocusLink and the MIPS yeast genome database (Burks *et al*. 1985; Mewes *et al*. 1997; Bairoch 2000; Wheeler *et al*. 2000). Links are automatically displayed when the description field in GenBank contains an EC number or when the EST sequence displays similarity to sequences in *Saccharomyces cerevisiae* or LocusLink organisms. Simultaneous examination of ESTs assembled into a given contig can reduce the time needed for annotation. In addition, the annotation process can be speeded up by allowing multiple users to have full access to the EST projects.

Graphical display

In PHOREST, all or a subset of ESTs from one or more projects can be displayed using various graphical interfaces. The global distribution of contigs among projects can be viewed in a triangle diagram. Based on normalized redundancies, assembled contigs are distributed along the axes of a triangle. Contigs at the corners of the triangle represent transcripts unique to one of the EST projects, whereas contigs found at the centre are equally redundant in all projects. Clicking anywhere within the triangle brings up a table of all contigs with corresponding distributions and these can be used for further analysis. The distribution of ESTs or contigs in different functional categories can be presented in tables, pie charts or bar diagrams. The functional distribution of ESTs and contigs in different projects is presented to provide a comparative overview of multiple EST projects.

Administration of PHOREST

In order to ensure the security of the information in the PHOREST database a number of security measures is in place. One person has the administration rights of the system and will have total control. The PHOREST administrator can add new users as well as change users' access rights to the different databases so that a user can only see the databases that he/she has access rights to. In addition, the PHOREST administrator can also add new databases and EST projects by simply filling in a form on the administration page. The input file required is a FASTA file for each EST project. Users with full access rights to a specific database can upload new sequences from the administration page.

System requirements and availability

PHOREST only requires freely available programs. These include the MYSQL relational database (www. mysql.com) and the web server APACHE (www. apache.org). To cluster the EST sequences, CAP (Huang 1992) is used and the homology searches are performed using BLAST standalone (Burks *et al*. 1985). PHOREST is written in PHP and runs unchanged on UNIX, Microsoft Windows and Macintosh (Mac OS X) platforms. As most administrative tasks are handled through the administration page in PHOREST, no advanced computer knowledge is required (i.e. no UNIX or programming skills).

PHOREST is freely available for academic use from the corresponding author. A manual and PHOREST demonstration database are available on the web (http://www.biol.lu.se/phorest/).

Implementation

We have used PHOREST to analyse EST data generated from parasitic and symbiotic fungi (Tunlid & Ahrén 2001; Johansson *et al*. 2003). Both studies involve comparative analyses of ESTs from a number of different growth stages.

More than 22 000 EST sequences in five databases containing 13 separate EST projects are currently being handled by PHOREST on a Linux computer (2 × 1.8 GHz, 1 GB RAM). Two other research groups are currently testing the PHOREST software for their EST projects. The fact that several groups have actively used PHOREST in research for several years has ensured that PHOREST is now a stable software tool with a minimum of bugs.

Acknowledgements

This work was supported by grants from the Swedish Research Council. We would like to thank Karl Söderström and BioBridge Computing AB for valuable discussions

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Ayoubi P, Jin X, Leite S *et al.* (2002) PipeOnline 2.0: automated EST processing and functional data sorting. *Nucleic Acids Research*, **30**, 4761–4769.

Bairoch A (2000) The ENZYME database in 2000. *Nucleic Acids Research*, **28**, 304–305.

Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST — database for 'expressed sequence tags'. *Nature Genetics*, **4**, 332–333.

Burks C, Fickett JW, Goad WB *et al.* (1985) The GenBank nucleic acid sequence database. *Computer Applications in the Biosciences*, **1**, 225–233.

Davey GC, Caplice NC, Martin SA, Powell R (2001) A survey of genes expressed in the Atlantic salmon (*Salmo salar*) as identified by expressed sequence tags. *Gene*, **263**, 121–130.

Everitt R, Minnema SE, Wride MA *et al.* (2002) RED: the analysis, management and dissemination of expressed sequence tags. *Bioinformatics*, **18**, 1692–1693.

Gibson G (2002) Microarrays in ecology and evolution: a preview. *Molecular Ecology*, **11**, 17–24.

Huang X (1992) A contig assembly program based on sensitive detection of fragment overlaps. *Genomics*, **14**, 18–25.

Johansson T, Le Quéré AL, Ahrén D *et al.* (2003) Transcriptional responses of *Paxillus involutus* and *Betula pendula* during formation of ectomycorrhizal root tissue. *Molecular Plant Microbe Interactions*, **17**, 212–215.

Karp PD, Paley S, Romero P (2002) The Pathway Tools software. *Bioinformatics*, **18** (Suppl. 1), S225–S232.

Mewes HW, Albermann K, Heumann K, Liebl S, Pfeiffer F (1997) MIPS: a database for protein sequences, homology data and yeast genome informations. *Nucleic Acids Research*, **25**, 28–30.

Miller RT, Christoffels AG, Gopalakrishnan C *et al.* (1999) A comprehensive approach to clustering of expressed human gene sequence: The Sequence Tag Alignment and Consensus Knowledgebase. *Genome Research*, **9**, 1143–1155.

Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. *Nature Genetics*, **32**, 261–266.

Picoult-Newberg L, Ideker TE, Pohl MG *et al.* (1999) Mining SNPs from EST databases. *Genome Research*, **9**, 167–174.

Qutob D, Hraber PT, Sobral BW, Gijzen M (2000) Comparative analysis of expressed sequences in *Phytophthora sojae*. *Plant Physiology*, **123**, 243–254.

Tunlid A, Ahrén D (2001) Application of genomics to the improvement of nematode pathogenic fungi. In: *NATO Advanced Research Workshop Enhancing Biocontrol Agents and Handling Risks* (eds Vurro M, Gressel J, Butt T, Harmann GB, Pilgeram A, St Leger RJ, Muss DL), pp. 193–200. IOS Press, Amsterdam.

Waugh M, Hraber P, Weller J *et al.* (2000) The phytophthora genome initiative database: informatics and analysis for distributed pathogenomic research. *Nucleic Acids Research*, **28**, 87–90.

Wheeler DL, Chappey C, Lash AE *et al.* (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **28**, 10–14.