Nucleic Acids Research, 2005, Vol. 33, Database issue ***D619–D621***
doi:10.1093/nar/gki119

CR-EST: a resource for crop ESTs

C. Künne, M. Lange, T. Funke, H. Miehe, T. Thiel, I. Grosse and U. Scholz*

Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, 06466 Gatersleben, Germany

Received August 15, 2004; Revised and Accepted October 20, 2004

ABSTRACT

The crop expressed sequence tag database, CR-EST (http://pgrc.ipk-gatersleben.de/cr-est/), is a publicly available online resource providing access to sequence, classification, clustering and annotation data of crop EST projects. CR-EST currently holds more than 200 000 sequences derived from 41 cDNA libraries of four species: barley, wheat, pea and potato. The barley section comprises approximately one-third of all publicly available ESTs. CR-EST deploys an automatic EST preparation pipeline that includes the identification of chimeric clones in order to transparently display the data quality. Sequences are clustered in species-specific projects to currently generate a non-redundant set of ~22 600 consensus sequences and ~17 200 singletons, which form the basis of the provided set of unigenes. A web application allows the user to compute BLAST alignments of query sequences against the CR-EST database, query data from Gene Ontology and metabolic pathway annotations and query sequence similarities from stored BLAST results. CR-EST also features interactive JAVA-based tools, allowing the visualization of open reading frames and the explorative analysis of Gene Ontology mappings applied to ESTs.

INTRODUCTION

Draft sequences of the *Arabidopsis* genome (1) and the rice genome (2,3) provide the basis for comprehensive gene indices from dicotyledonous and monocotyledonous plants. The avalanche of structural and functional genomic information available from plant genome projects makes it a challenge to annotate gene functions. For plants whose genomes are too large to be completely sequenced within the next years, expressed sequence tags (ESTs) (4) are being sequenced with the goal of having direct access to the transcribed genome. The technique of sequencing ESTs based on cDNA libraries allows powerful analyses in the field of molecular biology, such as the detection of genes in incompletely sequenced genomes.

Worldwide, a huge set of ESTs from diverse species is being sequenced and stored in the dbEST division of the GenBank database (http://www.ncbi.nih.gov/dbEST). Large sets of ESTs from barley, wheat, pea and potato are being sequenced at the Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, within different GABI and in-house research projects (http://www.gabi.de/). More than 381 000 barley ESTs are currently available as public domain data (http://srs.ebi.ac.uk/, July 2004), and ~35% of them were sequenced on the basis of cDNA libraries produced at IPK.

Important tasks that follow the process of EST sequencing are the reduction of redundancy in the set of ESTs, the functional annotation of represented genes, the selection of representative cDNA clones for the preparation of subsequent experiments, such as array expression experiments, and the presentation of those data in the World Wide Web. Some parts of these tasks are performed by a project called GABI primary database at the German Resource Center for Genome Research in Berlin (http://gabi.rzpd.de/). One goal of the CR-EST database is to extend the functionality of the primary database and to integrate more comprehensive aspects, such as clustering results, multiple alignments or the construction of complex search queries. Besides the common approach of providing data stocks of ESTs, such as structured databases or flat file collections, the goal of CR-EST is to provide an information system. Analysis tools are developed and directly embedded in the web front-end of CR-EST, using day to day experiences of biologists as the basis.

DATA PREPARATION, EST PROCESSING AND DATA QUALITY

The flexible data structure of the CR-EST database enables the management of multiple species. Among the covered species, barley comprises the majority of the EST sequences in the CR-EST database. Hence, we will focus on barley in the remainder of this paper. The barley sequencing (5,6) of 34 barley cDNA libraries containing 135 249 clones results in 114 435 5′-ESTs and 72 846 3′-ESTs. Their library size ranges from 300 to 15 000 ESTs. The average sequence length is 516 bp. A detailed statistics can be found in Supplementary Table 1 or

*To whom correspondence should be addressed. Tel: +49 39482 5 513; Fax: +49 39482 5 595; Email: scholz@ipk-gatersleben.de

directly queried from the database (http://pgrc.ipk-gatersleben.de/cr-est/liball.php).

EST processing

EST processing is a fundamental step in order to obtain high-quality sequence reads from raw sequencer trace data. CR-EST uses a semi-automated pipeline for EST processing, utilizing publicly available software in conjunction with in-house developed Perl scripts. PHRED (7,8) is used for base-calling the chromatogram trace files. Trace files are converted into FASTA files after bad sequence reads are removed by using the trim_fasta option together with a trim_cutoff parameter value of 0.01. Vector-derived sequences as well as user-specified adaptor sequences are screened with CROSS_MATCH (part of the PHRAP package) using a minscore parameter value of 20 and a minmatch parameter value of 12. Masked sequences are removed from the tails of the sequences.

In order to trim poly(A) tails, sequences are scanned from their ends for a series of at least five consecutive adenine nucleotides within a window of 50 bp. If such a poly(A) run is found, the trailing sequence, including the poly(A) run, is considered as a poly(A) tail and will be removed. This step is repeated recursively until no poly(A) run is found. The trimming of poly(T) tails is done analogously. Sequences shorter than 100 bp are discarded, and sequences longer than 700 bp are clipped at their 3′ ends. After this processing step, the average EST length is 516 bp (Supplementary Table 1).

To reduce redundancy, all ESTs are clustered using stackPACK (http://www.egenetics.com/) (9). Relevant clustering project information, including consensus sequences, multiple alignments and EST membership, are extracted from the stackPACK database.

Chimeric ESTs occur during the cDNA library construction as a result of recombination events. They may cause problems for subsequent analysis steps, such as the generation of a representative unigene set or the functional annotation of ESTs. Hence, the detection of chimeric ESTs is a crucial step of the EST processing pipeline, and CR-EST uses the following two-step approach to identify chimeric ESTs:

i. Detection by genome comparison. ESTs are aligned to genomic DNA of the appropriate model organism using the spliced alignment program GENESEQER (10). The rice genome is used for the monocotyledonous crop plants barley and wheat, and the *Arabidopis thaliana* genome (both obtained from ftp://ftp.tigr.org/pub/) is used for the dicotyledonous crop plants pea and potato. When dispersed alignments for a given EST are found such that two subsequences of this EST align to different places in the genomic DNA, this particular EST is considered to be chimeric.
ii. Manual inspection of large clusters. Chimeric ESTs tend to join two distinct sub-clusters of ESTs. Hence, all ESTs that join two sub-clusters of a stackPACK clustering project are removed, and the stackPACK clustering step and manual inspection step is repeated recursively.

DATABASE DESIGN

Following EST processing, the need arises to efficiently store and access the resulting data. First, the cDNA library information, EST sequences and the consensus sequences resulting from clustering, including information about their multiple alignments and their EST memberships, need to be stored in the CR-EST system. Second, it also needs to contain integrated information about BLASTX (11) results of both ESTs and consensus sequences against the non-redundant protein database (NRPEP, ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz) and about functional annotations including metabolic pathway and Gene Ontology (http://www.geneontology.org/) (12) information. Third, user management and data access restrictions need to be considered.

This requirement analysis leads to a complex Entity Relationship (ER) database schema (13) for CR-EST. The database implementation comprises 27 tables representing data entities and relationships listed above. The database schema can be found in Supplementary Figure 1. Different mechanisms are applied to loading data, including automatic imports using the Oracle SQL*Loader and SQL or Perl import scripts. Read-only access to stored information is ensured by database views on specific species.

The CR-EST system is structured using a 3-tier architecture. Web clients interact with a web server that hosts the CR-EST application, which in turn interacts with the database on another server. The database is implemented on the relational database management system Oracle9i® (http://www.oracle. com/).

SEARCHING AND DATA ACCESS IN CR-EST

CR-EST is designed to provide an explorative database access using a set of front-end tools. In addition, a flat-file archive of FASTA formatted EST sequences and consensus sequences is available at http://pgrc.ipk-gatersleben.de/cr-est/files/, and all public ESTs are available at the EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/).

The CR-EST web application allows the user to query information from the stored data, to visualize and to integrate them, as well as to use online tools for further analysis and processing. The applications are based on HTML pages dynamically generated by PHP scripts and JAVA Web Start (http://java. sun.com/products/javawebstart/).

Sequence Search

Major applications are different database search forms. The sequence form allows the user to query EST or consensus sequences. The search for EST memberships and alignments within a consensus sequence originating from a clustering process is offered, as well as the search for open reading frames. Another search form allows complex database queries involving BLASTX results of ESTs and consensus sequences against the NRPEP database to retrieve their stored BLASTX hits or the complete text document. In addition, a BLAST module provides DNA sequence homology searches using the WU-BLAST 2.0 programs BLASTN, TBLASTN and TBLASTX against the available ESTs in CR-EST.

Annotation Search

Several tools are offered to build catalogs of functional EST annotations, including tools that map ESTs to metabolic functions in KEGG (14) or that map ESTs to functional terms in the Gene Ontology Database (12).

Both tasks are accomplished by a GraphMining approach in integrated databases (15). The basic idea of this automatic functional categorization is to (i) take BLASTX hits for an EST, (ii) extract the database identifiers, (iii) query them in an integrated database and (iv) pick the pre-computed data link graphs to spawn a transitive connection to the data of the destination table, enabling a comparative analysis of EST functions.

With respect to metabolic functional groups, 8607 out of 111 090 in-house sequenced barley ESTs could be annotated to show enzymatic activity. Of these ESTs, 21% could be assigned to functional categories of carbohydrate, 19% to amino acid and 17% to energy metabolism.

Further features

In general, the user has a guest status, which allows the use of all applications that operate on unrestricted data. Registered user accounts grant extended access to restricted data, and the application menu with additional specialized queries is generated depending on the user login. The login form also allows to change between different species. In addition to the major search features, a list of cDNA library statistics link to detailed descriptions. Further statistics give an overview of the species-specific database content and the cluster results from clustering projects.

Every view of detailed sequence-related information is cross-linked with external data, if available. This also includes links to the main public sequence information resources, such as the SRS installation at the EBI (http://srs.ebi.ac.uk/) or the ENTREZ system at the NCBI (http://www.ncbi.nlm.nih.gov/entrez/).

FUTURE OF CR-EST

Our goal is to provide a comprehensive information system for storing and analyzing data resulting from crop plant EST sequencing projects. Currently, only data from in-house EST sequencing projects are provided, but due to the flexible data schema it is possible to integrate publicly available ESTs and specialized clustering projects. The data schema is organized to integrate the results of external unigene selection tools. Another extension will be the provision of links for downloading sequence quality files in the web application. Thus, the user has the opportunity to interpret the quality of a sequence and all corresponding generated data, e.g. computed consensus sequences, on their own.

Furthermore, the ESTs stored within CR-EST build a useful resource for 'genomeless genomics' analyses. This EST information resource is a starting point for further molecular biological investigations, such as the setup of array experiments for expression analysis. We develop a data warehouse to have an integrated view on this various data. Within this project, we focus on the development of databases and algorithms that can be combined with advanced experimental technologies to identify functional genetic elements in molecular sequences and pathways, which in turn control and regulate gene expression. The plant data warehouse presently developed at IPK combines genotypic, phenotypic, taxonomic and expression data from both, IPK-internal and publicly available data sources. Hence, it serves as one example for the ongoing activities to combine data integration, data analysis and data modeling based on data warehouse concepts.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank L. Altschmied, S. Biemelt, A. Graner, T. Münch, R. Radchuk, R. Schnee, P. Schweizer, R. Sigmund, N. Sreenivasulu, N. Stein, S. Weise, W. Weschke and H. Zhang for discussions and technical assistance, and the German Ministry of Education and Research for financial support (BIC-GH 0312706A).

REFERENCES

1. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
2. Goff,S.A., Ricke,D., Lan,T.H., Presting,G., Wang,R., Dunn,M., Glazebrook,J., Sessions,A., Oeller,P., Varma,H.*et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92–100.
3. Yu,J., Hu,S., Wang,J., Wong,G.K., Li,Z., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.
4. Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B., Moreno,R.F. *et al*. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.
5. Zhang,H., Sreenivasulu,N., Weschke,W., Stein,N., Rudd,S., Radchuk,V., Potokina,E., Scholz,U., Schweizer,P, Zierold,U. *et al*. (2004) Large-scale analysis of the barley transcriptome based on expressed sequence tags (ESTs). *Plant J*., **40**, 276–290.
6. Michalek,W., Weschke,W., Pleissner,K.-P. and Graner,A. (2002) EST analysis in barley defines a unigene set comprising 4000 genes. *Theor. Appl. Genet*., **104**, 97–103.
7. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*., **8**, 175–185.
8. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*., **8**, 186–194.
9. Christoffels,A., van Gelder,A., Greyling,G., Miller,R., Hide,T. and Hide,W. (2001) STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Res*., **29**, 234–238.
10. Usuka,J., Zhu,W. and Brendel,V. (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, **16**, 203–211.
11. Altschul,S.F., Gish,W., Miller,W, Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
12. The Gene Ontology Consortium (2001) Creating the Gene Ontology Resource: design and implementation. *Genome Res*., **11**, 1425–1433.
13. Chen,P.P. (1976) The Entity-Relationship Model—towards a unified view of data. *ACM Trans. Database Syst*., **1**, 9–36.
14. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, 277–280.
15. Balko,S., Lange,M., Schnee,R. and Scholz,U. (2004) BioDataServer: an applied molecular biological data integration service. In Rahm,E. (ed.), *Proceedings of the Data Integration in the Life Sciences: First International Workshop* (*DILS 2004*), Leipzig, Germany, LNCS **2994**, March 25–26, 2004. Springer-Verlag, Heideberg, Germany, pp. 140–155.