D344–D347 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki121

The Diatom EST Database

Uma Maheswari[1], Anton Montsant[2,3], Johannes Goll[1,4], S. Krishnasamy[5], K. R. Rajyashri[1], Villoo Morawala Patell[1] and Chris Bowler[2,3,*]

[1]Avestha Gengraine Technologies Pvt. Ltd, 'Discoverer', 9th Floor, International Technology Park, Whitefield Road, Bangalore 560066, India, [2]Laboratory of Molecular Plant Biology, Stazione Zoologica 'Anton Dohrn', Villa Comunale, I-80121 Naples, Italy, [3]Organismes Photosynthétiques et Environnement, ENS/CNRS FRE 2433, Ecole Normale Supérieure, 46 rue d'Ulm, 75230 Paris, France, [4]Weihenstephan University of Applied Sciences, Am Hofgarten 4, 85354 Freising, Germany, [5]Bioinformatics Center, School of Biotechnology, Madurai Kamaraj University, Palkalai Nagar, Madurai 625021, India

Received August 16, 2004; Revised and Accepted October 21, 2004

ABSTRACT

The Diatom EST database provides integrated access to expressed sequence tag (EST) data from two eukaryotic microalgae of the class Bacillariophyceae, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. The database currently contains sequences of close to 30 000 ESTs organized into PtDB, the *P.tricornutum* EST database, and TpDB, the *T.pseudonana* EST database. The EST sequences were clustered and assembled into a non-redundant set for each organism, and these non-redundant sequences were then subjected to automated annotation using similarity searches against protein and domain databases. EST sequences, clusters of contiguous sequences, their annotation and analysis with reference to the publicly available databases, and a codon usage table derived from a subset of sequences from PtDB and TpDB can all be accessed in the Diatom EST Database. The underlying RDBMS enables queries over the raw and annotated EST data and retrieval of information through a user-friendly web interface, with options to perform keyword and BLAST searches. The EST data can also be retrieved based on Pfam domains, Cluster of Orthologous Groups (COG) and Gene Ontologies (GO) assigned to them by similarity searches. The Database is available at http://avesthagen.sznbowler.com.

INTRODUCTION

Diatoms (Bacillariophyceae) are brown algae with a wide distribution and abundance in the world's water bodies, and are thought to be responsible for around one-fifth of global primary productivity. Being such important players in the global ecosystem, their ecology and physiology have been the focus of research for decades. More recently, the intricate siliceous bioarchitecture of diatom cell walls has attracted the interest of nanotechnologists. Understanding the information within diatom genomes is therefore likely to lead to dissection of the molecular mechanisms controlling bioinorganic pattern formation in these organisms and is fundamental for understanding their ecological success (1,2).

As part of a general effort to study diatom biology at a molecular level, large-scale sequencing projects are being undertaken (2,3) (http://genomic.jpi-psf.org/thaps1.home.html). This rapidly growing body of sequence information requires accurate gene annotation as well as dedicated platforms for storage, processing and curation, and must be available for immediate data retrieval at any time.

CONSTRUCTION OF THE DATABASE

Raw data and core analyses

PtDB contains expressed sequence tags (ESTs) derived from *Phaeodactylum tricornutum* Bohlin clone CCMP632 (Provasoli-Guillard National Center for Culture of Marine Phytoplankton, Bigelow, ME). The RNA used for cDNA generation was isolated from exponentially growing cells (2). The cDNA library was created in a Uni-Zap XR vector (Stratagene) using oligo dT primers and directionally inserted into EcoRI–XhoI sites of pBluescript. 5′ end sequences (12 136) were generated using the T3 primer. PTSS0001–PTSS0997 have been described previously (2); PTAM00001–PTAM01131 were generated by MWG Biotech (Ebersberg, Germany) and PTMM00001–PTMM10008 were obtained from Avesthagen (Bangalore, India). TpDB contains ESTs derived from *Thalassiosira pseudonana* clone CCMP1335 (Provasoli-Guillard National Center for Culture of Marine

*To whom correspondence should be addressed. Tel: +33 144323525; Fax: +33 144323935; Email: cbowler@biologie.ens.fr

Phytoplankton), from an exponentially growing culture in ASW medium. The cDNA library was created in the pZERO-2 vector (Invitrogen) using oligo dT primers and was not directionally inserted. A total of 6500 clones were sequenced from both ends and were denoted with an .x or .y extension in the clone ID based on the direction of sequencing. In some cases poor-quality runs were repeated, giving rise to .x2 and .x3 extensions etc. until 15 174 sequences were obtained.

Prior to annotation, the sequences were subjected to quality checking and vector clipping using the Trimest, Trimseq and Vectorstrip programs of EMBOSS (European Molecular Biology Open Software Suite). The vector data were provided interactively to Vectorstrip and all sequences with a maximum mismatch level of 10% were detected and removed. As the *T.pseudonana* ESTs were generated from both ends, assembling was done using the consensus sequence rather than the individual ESTs when overlap was detected, which occurred for 1056 pairs of ESTs. Such complete cDNA sequences are labelled with the same ID as the individual ESTs, but without any extension.

All sequences longer than 50 nt were then subjected to clustering using the Contig Assembling Program (CAP3) (4) to detect sequence redundancy. Sequences with >95% identity over a region longer than 30 nt were clustered, yielding 1243 contig assemblies for *P.tricornutum* and 832 contigs for *T.pseudonana*. Contigs were given a unique contig ID consisting of a prefix C and a four-digit number, assigned in descending order of number of ESTs in each contig. This helps to organize the ESTs based on the level of redundancy. The longest sequence from each assembly was then selected and pooled with the singletons (i.e. ESTs that did not fall into any cluster) to form the non-redundant set. PtDB contains 5108 non-redundant sequences and TpDB contains 5444 (Table 1). These sequences were then subjected to automated annotation, which comprised searches against the NCBI (5) non-redundant protein database using BLASTX and against protein domain databases, CDD (6) and COG databases using RPS-BLAST (Figure 1). The results of all similarity searches were parsed and stored in MySQL tables.

Expressed sequence tag analysis—full-length clones and function assignments

In order to identify putative full-length sequences for function assignment, the EST sequences were grouped into nine alignment classes (Figure 2) based on the subject coverage (CovS) and identities of the BLASTX results. The subject coverage was calculated as follows (7):

$$\mathrm{CovS} = 100\mathrm{Hlen}/\mathrm{Slen},$$

where Hlen is defined as length of the HSPs (high-scoring segment pairs) and Slen is defined as subject sequence length. CovS is an indicator of the extent to which the query sequence matches the target protein sequence.

Table 1. Number of sequences in PtDB and TpDB

Sequence type	Number in PtDB	Number in TpDB
Raw ESTs	12 136	15 174
Contigs	1243	832
Singletons	3865	4612
Non-redundant set	5108	5444

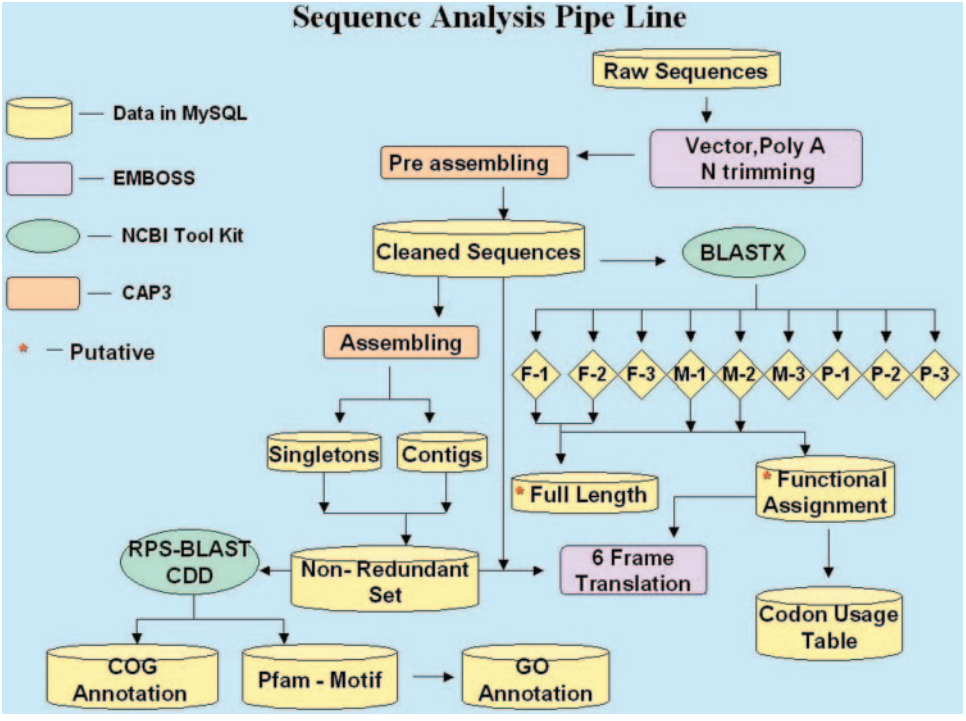

Figure 1. EST analysis overview.

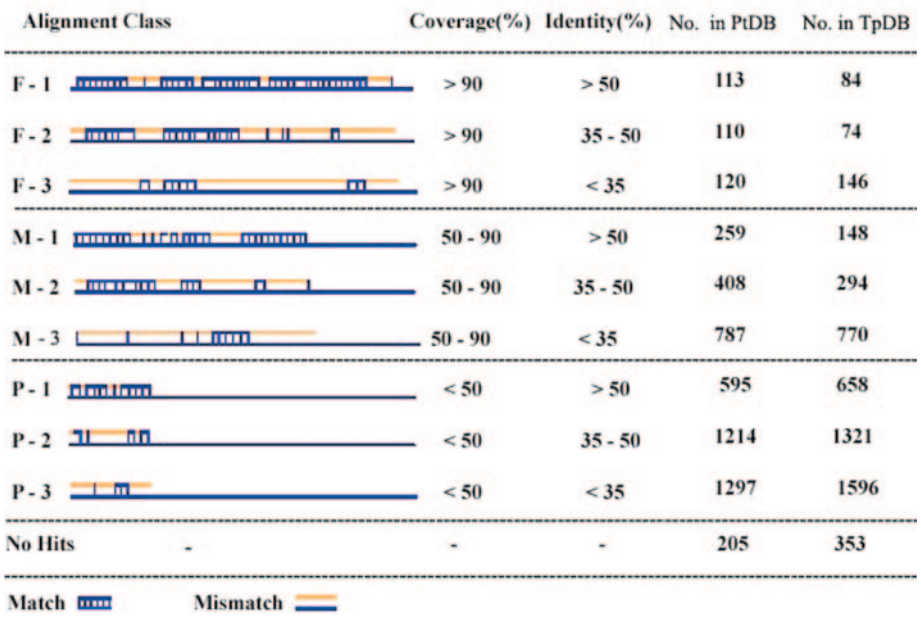

Figure 2. Classification of non-redundant sequences based on BLASTX alignment coverage and identity percentages.

For sequences falling in F-1, F-2, M-1 and M-2 alignment classes (see Figure 2), the protein coding frame and the putative function were assigned based on the BLASTX description. For this subset of ESTs, in the six-frame translation output, the start and stop codons and the assigned frame are highlighted so that the user can detect complete open reading frames easily. The F-1 and F-2 categories were considered to comprise putative full-length clones. These sequences were aligned using CLUSTAL W (8) with their corresponding 10 most similar GenBank protein database entries obtained from the BLASTX results. The alignment output is linked to the web interface for quick reference.

We used RPS (Reverse PSI) BLAST (6) to identify the COG (Cluster of Orthologous Groups) (9) to which each sequence in the non-redundant collection could be assigned. This allows the sequences to be classified into one of the groups shown in Figure 3.

Furthermore, to support the functional assignments and classifications, a motif search was made among all non-redundant sequences using RPS BLAST against the Pfam database (10). Motifs were assigned to ESTs in which a Pfam domain was detected with an *E*-value < 0.05. The corresponding Gene Ontology (GO) description (11) was also assigned to the non-redundant sequences based on the dbxref table in the GO database (MySQL format).

Codon usage tables were created for each organism using the subsets of non-redundant ESTs falling in the alignment classes F-1, F-2, M-1 and M-2 (which amount to 859 sequences for *P.tricornutum* and 465 for *T.pseudonana*). The coordinates delimiting the coding region of these ESTs were obtained from the BLASTX output and the codon usage table was created using the Cusp program of EMBOSS.

Database architecture

The Diatom EST database is based on Linux Red Hat 9.0 and was developed with MySQL 4.0 as a backend with a web interface using PHP4. Bioperl and Perl Scripts were used to parse and fill the data into the database (Figure 1).

SEARCHING THE DATABASE

The database can be accessed through a web interface, and querying can be done using the View and Search options. The View option facilitates listing of the raw ESTs, contigs, singletons and non-redundant sequences. The ESTs are also listed based on their COG and GO assignments. Search options include simple searches by organism name, keyword, accession number or sequence ID. BLAST, BLASTN, TBLASTN and TBLASTX searches can also be performed against the Diatom EST Databases (PtDB, TpDB or both). An Advanced Search option provides additional possibilities such as the use of boolean terms (AND, OR and NOT) to search for a keyword/organism pair, defined alignment class, subject coverage (CovS), percentage identity and *E*-value. The search output contains information about the EST and its contig and functional annotations, sorted by *E*-value and sequence ID.

FUTURE DIRECTIONS

In the future, we hope that the Diatom EST Database will incorporate data from additional species as EST and genome-sequencing projects for diatoms (and other algae) are performed. Orthology will be assigned according to eukaryotic orthologous groups (KOGs) as soon as they are made available at the NCBI. Apart from the sequence and related data within the currently available database structure, gene expression data, including from microarray studies, could also be included. The database could also be integrated with a genome browser, where available, and enhanced functional annotation could be mined from other cluster and pathway databases. The server will be periodically upgraded for faster access to the growing body of data.

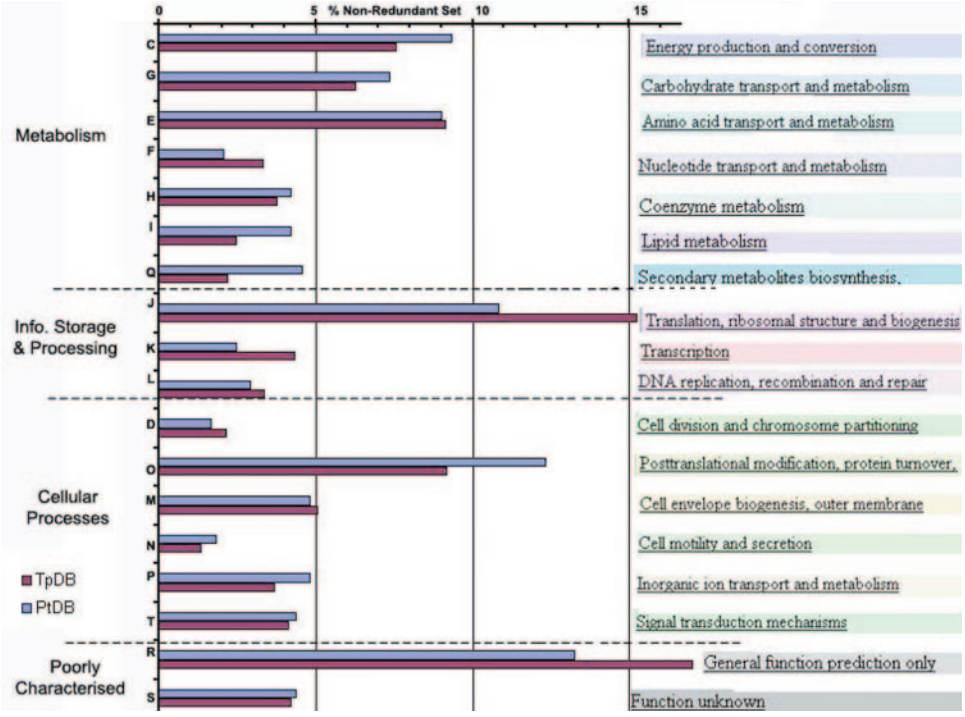

Figure 3. Classification of the non-redundant set into COG functional categories.

AVAILABILITY

The Diatom EST Database is freely available on the web at http://avesthagen.sznbowler.com. The *P.tricornutum* ESTs have been submitted to the NCBI dbEST (GenBank accession numbers CD374840–CD384835 and BI306757–BI307753). Requests for bulk queries and to house EST data from other diatoms should be addressed to C. Bowler.

ACKNOWLEDGEMENTS

We are grateful to Kala Thangalakshmi, Savita Shrivastava and Santha Kumar for managing the server and the software and for their help in web interface creation, to Kamel Jabbari and Dhruvdev Vyas for their help and suggestions and to Ullas PV for suggestions on preliminary sequence analysis. We are also grateful to the sequencing team of Avestha Gengraine Technologies. The *T.pseudonana* ESTs were a kind gift from Mark Hildebrand and Diego Martinez. Partial funding for the Diatom EST database was from the EU-funded Margenes project to C.B. (QLRT-2001-01226).

REFERENCES

1. Falciatore,A. and Bowler,C. (2002) Revealing the molecular secrets of marine diatoms. *Annu. Rev. Plant Biol.*, **53**, 109–130.
2. Scala,S., Carels,N., Falciatore,A., Chiusano,M.L. and Bowler,C. (2002) Genome properties of the diatom *Phaeodactylum tricornutum*. *Plant Physiol.*, **129**, 993–1002.
3. Armbrust,E.V., Berges,J.B., Bowler,C., Green,B.R., Martinez,D., Putnam,N.H., Zhou,S., Allen,A.E., Apt,K.E., Bechner,M. *et al.* (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science*, **306**, 79–86.
4. Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
5. 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.
6. Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, 383–387.
7. Xing,L. and Brendel,V. (2001) Multi-query sequence BLAST output examination with MuSeqBox. *Bioinformatics*, **17**, 744–745.
8. Thompson,J.D., Higgins,D.J. 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.
9. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
10. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
11. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.