Nucleic Acids Research, 2002, Vol. 30 No. 21 4761–4769

PipeOnline 2.0: automated EST processing and functional data sorting

Patricia Ayoubi, Xiaojing Jin, Saul Leite[1], Xianghui Liu, Jeson Martajaja[1], Abdurashid Abduraham, Qiaolan Wan, Wei Yan, Eduardo Misawa[1] and Rolf A. Prade*

Department of Microbiology and Molecular Genetics and [1]School of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, OK 74078, USA

Received May 6, 2002; Revised August 20, 2002; Accepted September 3, 2002

ABSTRACT

Expressed sequence tags (ESTs) are generated and deposited in the public domain, as redundant, un-annotated, single-pass reactions, with virtually no biological content. PipeOnline automatically analyses and transforms large collections of raw DNA-sequence data from chromatograms or FASTA files by calling the quality of bases, screening and removing vector sequences, assembling and rewriting consensus sequences of redundant input files into a unigene EST data set and finally through translation, amino acid sequence similarity searches, annotation of public databases and functional data. PipeOnline generates an annotated database, retaining the processed unigene sequence, clone/file history, alignments with similar sequences, and proposed functional classification, if available. Functional annotation is automatic and based on a novel method that relies on homology of amino acid sequence multiplicity within GenBank records. Records are examined through a function ordered browser or keyword queries with automated export of results. PipeOnline offers customization for individual projects (MyPipeOnline), automated updating and alert service. PipeOnline is available at http://stress-genomics.org.

INTRODUCTION

Large expressed sequence tag (EST) projects have accumulated over 11.3 million un-annotated files from more than 5318 cDNA libraries (as of January 1, 2002) currently deposited in the public domain. While processing and annotation of DNA sequence data is possible using commercially available software, they typically read single input files, produce single output files and require extensive manual intervention by the user. Thus, most commercial software has limited use in processing large-scale sequence data (1). While software such as PHRED (2), PHRAP (http://bozeman.mbt.washington.edu/phrap.docs/phred.html) and BLASTALL (3) are designed to process large-scale DNA sequence data, assignment of metabolic or biological function remains a process for which specific software is not easily accessible. Commercial bioinformatics efforts such as Electric Genetics Paracel (Paracel, Pasadena, CA) and Lion Biosciences (Lion Biosciences AG, Heidelberg, Germany) systems are costly customer-specific enterprise software services.

Often, the term EST relays the notion that a large number of single read DNA sequencing (end)-reaction data files, randomly chosen from a specified cDNA clone library, has been produced. Given that ESTs are derived from a population of mRNA molecules, the entire data set portrays a digital expression profile (4,5) of the condition from which the RNA was harvested (6). Thus, deciphering the DNA sequence and assigning biochemical function results in a functional, cellular and metabolic profile for a given physiological condition or morphological state. Because ESTs projects produce large amounts of individual DNA sequence files that must be individually edited and functionally categorized, via homologous sequence matching, automated annotation is essential. Implementation of automated annotation creates a number of problems falling within two classes: first, the stringency of the homologous match that defines a biological function is not a constant applicable to all cases and secondly, annotation of biochemical function for public DNA protein sequences comprises uncontrolled vocabulary and contains mostly free text.

Efforts to assemble, organize and annotate raw DNA sequence data have been developed employing diverse strategies such as: clustering of ESTs as in STACK (7), indexing as in TIGR Gene Indices (8,9) and UNIGENE, Merck Gene Index databases (10,11), cross-referencing as in TIGR Orthologous Gene Alignment database (12) and a combination of automated and manual curation as in MendelGFDb (13).

Here, we describe PipeOnline, which is a series of integrated and fully automated programs used to process large quantities of raw DNA sequence data files and associate function at the gene level. PipeOnline accepts input raw DNA sequence trace files for base calling, vector editing, and homology driven assembly into a unigene set and generation of consensus contigs. The unigene contig set is compared with public annotated databases using BLASTX similarity

*To whom correspondence should be addressed. Tel: +1 405 744 7522; Fax: +1 405 744 6790; Email: prade@okstate.edu
Present address:
Patricia Ayoubi, Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK 74078, USA

matching algorithms. DNA-sequence to protein-function relations are established and a MySQL database is created. PipeOnline links functional information gathered from the BLASTX similarity searches of input unigene-contig sequences to a Metabolic Pathways Database (MPW)-based functional dictionary (14), to generate a functional overview of a given input EST collection. Moreover, PipeOnline users can browse a functional overview, query and export records from a database, and relate edited nucleotide-, protein-sequences and similarity alignments for a given collection of input files.

MATERIALS AND METHODS

Execution of PipeOnline begins from a web-interface where a user enters the specified directory on the PipeOnline server containing raw data files, a database name and an e-mail address. PipeOnline executes a series of programs to assess quality, then edits and assembles the input DNA sequence information into a non-redundant data set. This non-redundant data set (unigene contigs) is used as the input for homology-driven public database searches, homology-based functional sorting and creation of a MySQL database. DNA sequences can be submitted to PipeOnline in Applied Biosystems Inc. (ABI) format, Standard Chromatogram Format (SCF) or FASTA format.

Processing of DNA data files

The first module in PipeOnline is the base-caller PHRED (2), which takes input chromatogram files in ABI format, or SCF generated from automated sequencing efforts, and converts them to bases and quality indices. The output sequence and base quality index files are written in standard FASTA format. The program CROSSMATCH (http://bozeman.mbt.washington.edu/phrap.docs/phrap.html) is used to compare FASTA file sequences generated by PHRED to vector sequences using the Smith–Waterman–Gotoh algorithm (15,16). Vector masked sequence and base quality index files are used by the assembly engine PHRAP (http://bozeman.mbt.washington.edu/phrap.docs/phred.html) to construct consensus sequences (contigs) from these input files while eliminating vector masked regions from the output of all formed contigs. A Perl script (removex.pl) is then used to remove the masked sequences (Xs) corresponding to vector from the singleton sequence files.

Protein database searches

BLASTALL (NCBI) is used for batch execution of BLASTX (17) to search for sequence similarity between processed DNA-sequence outputs and public non-redundant amino acid databases. In each case, an expectation value cutoff of 5×10^{-1} is used and the top five most significant alignments (based on the expectation values) are stored in the PipeOnline database.

Functional sorting

Automated functional sorting of each record in a database is based on the MPW functional dictionary (14) obtained from what is there (WIT) (18). Alternate enzyme names (synonyms) for the standard enzyme names contained within the MPW functional dictionary were obtained from SWISS-PROT (19). All matches between the MPW functional dictionary and NCBI protein records were stored in a table containing NCBI gene index numbers and matching MPW functional scheme to generate an NCBI protein-function dictionary. All tables used for PipeOnline functional sorting were generated, then stored and queried locally from a MySQL database server.

PipeOnline UNIX scripts

The script PipeOnline.sql can be executed from a web interface. Each program module within the script (Table 1) can be updated, replaced or removed from PipeOnline.sql permitting customization and updating of PipeOnline. PipeOnline can also be customized to individual users' needs through modification of program module arguments added to the command lines for execution of PHRED, CROSSMATCH, PHRAP or BLASTALL. Execution of PipeOnline.sql generates predefined directory architecture for locating input trace or DNA sequence files on the server while intermediate and subsequent output files produced by PipeOnline.sql are stored for the database owner.

Database field descriptions

Each PipeOnline database is stored in a MySQL relational database server and contains all information collected from the individual program modules (Table 2). This includes the name and sequence of original input sequences (clones), contigs and singlets generated from assembly and the total length of each DNA sequence. Fields regarding the top BLASTX alignments contain extensive information including the NCBI record description, NCBI gene index number, source organism name (where available), bit score, *E*-value, high-scoring segment pairs (HSP), percent identity, percent similarity and up to five amino acid sequence alignments. Estimation of function is inferred from the amino acid sequence alignment descriptions and query of the local NCBI functional dictionary and a functional overview is generated for each PipeOnline database.

Database queries

PipeOnline database records can be accessed by keyword searches via CGI query of databases. Keyword searches include any portion of the protein alignment description field as found in the NCBI non-redundant protein record (including gene index numbers or gene, enzyme or source organism name). Records can also be queried by contig or clone name. For each term query, users can enter a desired bit score, HSP or *E*-value (obtained from the stored protein alignments) to restrict or expand record retrieval based on a user-defined statistical significance threshold. Results from a term query are presented in table format containing the list of hits and pertinent information such as the description and gene index number of the top alignments, number of corresponding clones forming each contig and contig name, which is hyperlinked to database individual records. PipeOnline databases can also be queried using integrated BLASTN, TBLASTN or TBLASTX search programs. These results are displayed in a typical BLAST result format and each match links to the corresponding PipeOnline record.

Browsing

GeneBrowser allows users to navigate through a cell-structured functional overview of all records contained within a

Table 1. PipeOnline program modules

Module	Description	Reference/URL
Modules and scripts		
PHRED	Base-calling and generation of quality values from trace files	Ewing *et al.* (2)
CROSSMATCH	Vector screening	Green http://www.phrap.org
PHRAP	Sequence assembly	Green http://www.phrap.org
BLASTALL	Local alignment of sequences against local non-redundant sequences	NCBI ftp://ftp.ncbi.nlm.nih.gov/blast
Pipeonline.sql	UNIX script linking PipeOnline modules	This work
CreateTables.sh	Generates PipeOnline MySQL relational database from the BLASTX outputs	This work
createBrowse_Table.pl	Functional assessment and sorting of records based on closest homolog gene index numbers	This work
Query, report and export		
querydb.pl	CGI-script allowing boolean-type query and retrieval of records in a PipeOnline database	This work
querySequence.pl	CGI-script allowing BLAST query and retrieval of records in a PipeOnline database	This work
browsegene.pl	CGI-script to display a functional overview of PipeOnline records for browsing by function	This work
annota_dbEST.pl	CGI-script allowing batch annotation of PipeOnline EST records for submission to NCBI	This work
query_ex.pl	CGI-script allowing generation of portable text file of records in a PipeOnline database	This work

PipeOnline is composed of several program modules linked by a single UNIX script termed PipeOnline. The modules used for processing trace files, vector editing, assembling sequences and BLASTX comparison are public programs. A local copy of GenBank is used for running local BLASTX and is updated daily (see Fig. 1). Generation of a relational MySQL database and functional sorting of BLASTX outputs are accomplished using novel programs and a Web-based interface.

Table 2. PipeOnline database field descriptions

Field	Description
DNA sequence	Base calling, vector editing and contig assembly
Contig name	Unigene contig sequence name
Sequence length	Length of contig nucleotide sequence (bp)
Clones forming contig	Hyperlink to list of raw sequences (clones) used to generate contig
Nucleotide sequence	Complete sequence of contig
Amino acid sequence alignment	BLASTX comparison
Top alignment	Description of the most significant protein alignment
Organism	Name of organism source for the most significant protein alignment
Gene index number	Gene index number of the most significant protein alignment
Bit score	Bit score of the most significant protein alignment
HSP	HSP of the most significant protein alignment
E-value	*E*-value of the most significant protein alignment
% Identity	Percent identity of the most significant protein alignment
% Similarity	Percent similarity of the most significant protein alignment
% Gaps	Percent of gaps in the most significant protein alignment
Most significant alignment	Protein sequence alignment for the most significant protein alignment
NCBI NR release date	Release date of protein sequences used for original BLASTX comparison
Total hits	Total number of alignments found by BLASTX comparison
Additional alignments	Hyperlinks to additional alignments
Functional assignment	Functional assignment based on amino acid sequence homology
Level I	Major functional class
Levels II–IX	General functional subclass
Lowest level	Enzyme/protein name (standardized functional definitions)

Fields included in PipeOnline records can be categorized into three groups: sequence data, protein alignment data and functional classification. Sequence data include the contig sequence and clones making up that contig. Data from the top five alignments are stored and each record is matched to the NCBI protein-function dictionary to generate the functional classification.

database. The overview is an outline of metabolic and cellular processes, which includes links to individual records at its lowest level. For each overview level, the total number of records contained within is displayed enabling metabolic reconstruction from randomly sequenced DNA fragments. Due to overlap between biological pathways and protein/enzyme functions, enzyme and protein names often appear in multiple categories and GeneBrowser reflects this redundancy.

Export of data

The table of results generated from individual queries can be downloaded from a web browser in tab delimited text format. In addition, the DNA sequence(s) retrieved from a search or using GeneBrowser can be downloaded in FASTA format and saved locally as a text file. These export features allow retrieval of data for further statistical analysis and interpretation. The export of sequences in FASTA format permits further analysis such as primer design or BLAST analysis against PipeOnline or other private and public databases.

Display

The record display can be divided into three basic components: (i) contig and clone sequence information, (ii) top protein alignment(s) information and (iii) the lowest level(s) of functional classification all displayed in a table format. Basic contig sequence information includes the contig

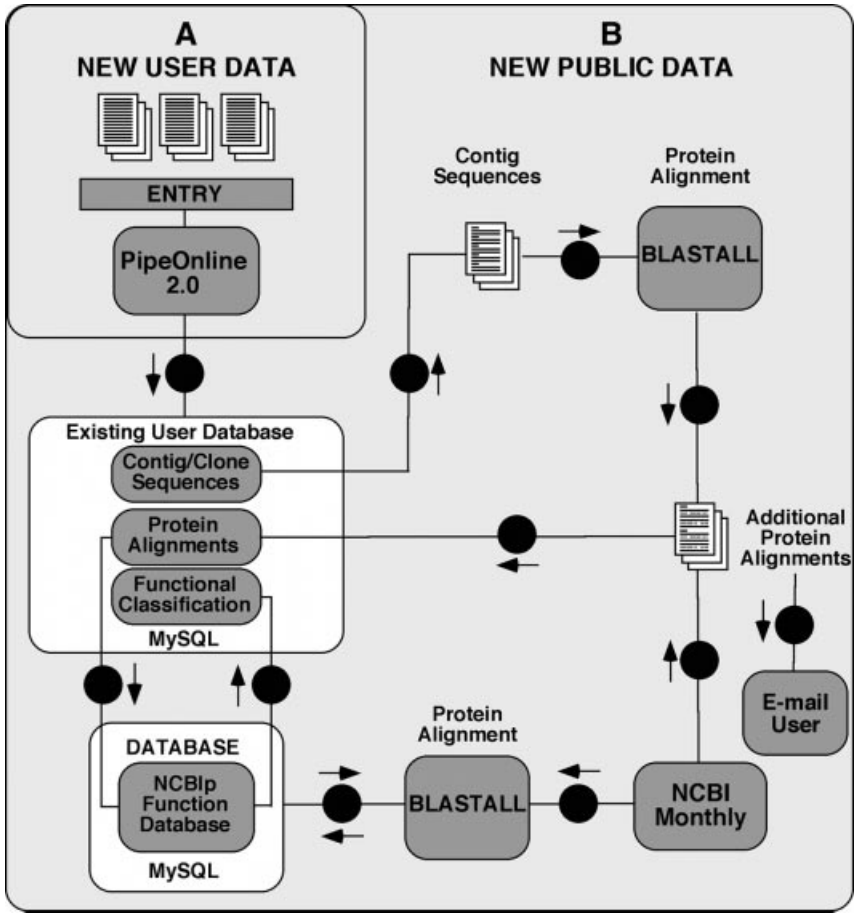

Figure 1. PipeOnline updating schemes. Two types of database updating are currently supported by PipeOnline: (**A**) addition of new DNA sequence records, ABI, SCF or FASTA, to an existing database by database owners; and (**B**) update of existing database records with new information from public databases. Automated updating occurs with each monthly update to the protein database from NCBI. New protein sequence entries are first matched with and added to the NCBI protein-function database. Locally maintained databases are automatically compared with the monthly updates, new matches added to the database and e-mail notification is sent to database owners.

DNA sequence and sequence length. In addition, the number of clones assembled to form each contig sequence is given with hyperlinks provided to each original clone sequence. The BLASTX results for the most significant hit are displayed with the amino acid sequence alignment. Records are also hyperlinked to GeneBrowser and NCBI protein records.

Updating

To reflect the most current sequence data and protein alignment information available from constantly changing public databases, automated updating schemes are implemented within PipeOnline (Fig. 1). The NCBI protein-function dictionary table is updated monthly using the NCBI monthly updates employing the same matching programs as before. Gene index numbers of each successfully classified entry from the NCBI monthly updates are then added to the PipeOnline NCBI protein-function dictionary. The NCBI monthly updates are also used to update PipeOnline databases by BLASTX analysis of contigs providing the most current data possible for each database. This is accomplished through automated monthly execution of BLASTX comparison between contig sequences in PipeOnline databases and the current NCBI monthly update. Up to five additional alignments (obtained from query of the NCBI monthly updates) can be added to PipeOnline databases and functional assignment is estimated as described above. When, following regular monthly updating, more than 10 hits are found by BLASTX analysis of a given contig sequence, the top 10 alignments (based on the *E*-value of the BLASTX protein alignments) are stored in the database. Following monthly updates to a database, the database owner is notified of all changes via e-mail. Database owners can also add new sequence files to an existing PipeOnline database using the PipeOnline.sql script, which updates PHRAP assemblies in existing databases (Fig. 1). New and updated contig sequences are identified and subjected to BLASTX comparison. The corresponding data are parsed, functionally sorted and all results are added to the existing PipeOnline database.

Annotation

The information stored in PipeOnline is useful for annotation of new sequence data. PipeOnline Annotator is a companion export program offering a batch procedure designed to facilitate database owners in large-scale data submission to

GenBank. Using the web-interactive PipeOnline Annotator, database owners enter the publication, library, author information and sequence type information common to all records while the sequences and putative identification for each record are extracted from the PipeOnline database. The putative identification is derived from the top protein alignment. As with queries, GeneBrowser and all other export features, the user can set a threshold for assignment of the putative identification.

Availability and system requirements

PipeOnline is available freely to academic researchers for research purposes. PipeOnline continues to be developed and improved, thus no installation scripts are available. However, extensive program and installation documentation will be available upon request. PipeOnline requests should be submitted to badmin@bioinfo.okstate.edu. Current documentation includes descriptions and source codes of PipeOnline user scripts, automated scripts and required UNIX setting. Briefly, a UNIX platform running a Web server and CGI, Perl version 5.005_02 or later and MySQL version 3.22.22 or later are required. In addition, PHRED, CROSSMATCH, PHRAP, BLAST suite of programs and NCBI non-redundant protein database must be installed locally. Connectivity with desktop computers for uploading trace or DNA sequence files by FTP to the UNIX is also required.

RESULTS AND DISCUSSION

PipeOnline overview

Figure 2 shows a summarized description of operations involved in processing raw input files through the PipeOnline environment and creation of a database. Program modules and essential operation modules are described in Table 1. Programs, scripts and related computational methods are described in detail in Materials and Methods.

The major problem in automatically processing DNA sequence files and annotation of gene-function through similarity matching of data deposited in the public domain is two-fold: first, to define the degree of similarity required to identify functionally equivalent proteins with confidence and secondly, information relating to function depends on annotated protein records with no quality control within the annotation.

The first problem of homology versus divergence is determined by the user through setting HSP, *E*-value or bit score thresholds. The second problem, far more complex and important to address, is resolved in PipeOnline through a sorting algorithm, which is independent of free text descriptions of public protein records.

Functional assignment and sorting

Functional dictionary. A combination of matching programs were used to match NCBI protein records to a MPW-based functional dictionary (14) and generate a pre-classified functional dictionary of NCBI protein records using the complete non-redundant amino acid database from NCBI (Fig. 3). The MPW database contains Enzyme Commission (EC) numbers and standard enzyme/protein names (standard functional definitions) organized into six basic functional classes. Classes are sub-divided into two to nine subclasses totaling 4155 unique functional classes with 2727 unique standard functional definitions at the lowest level.

To generate an NCBI protein-function dictionary, EC numbers and keywords from the NCBI record descriptions were extracted and matched to the MPW dictionary standard functional definitions, while words with no functional meaning (putative, hypothetical, intergenic region, etc.) were filtered out. To match a broader range of enzyme names found in NCBI descriptions, an enzyme synonym lookup table (19) was used.

Each successfully classified NCBI protein record was stored in the new functional dictionary by the gene index (GI) numbers with matching standard functional definition(s) from the MPW functional dictionary. These word-matching algorithms successfully classified 26.8% (117 100 of 437 749 records) of all NCBI non-redundant protein database records based on record descriptions.

A more rigorous search program (the BLAST amino acid sequence search program BLASTP) was used to match records where descriptions were not suitable for word matching (i.e. descriptions containing no identifiable functional meaning, where conventional nomenclature schemes were not followed or in the case of misspellings). For this, each NCBI record classified by word matching (see above) was obtained and significant protein alignments were identified using BLASTP comparisons against the complete NCBI non-redundant protein database while applying an expectation value of 1×10^{-40} (20) to search for related protein sequences independent of NCBI descriptions. This level of comparison stringency resulted in matches with bit scores greater than 160 and HSP greater than 400 providing confidence in applying functional classification to each match. Combined, these approaches successfully classified 41% of all NCBI protein records resulting in a more extensive NCBI protein-function dictionary (Table 3).

Functional assignment and sorting of PipeOnline records. Following protein BLASTX database searches for each contig sequence, functional assignment of records in a PipeOnline database is accomplished by matching GI numbers from top protein alignments to the NCBI protein-functional dictionary. Each successful match is added to PipeOnline records with matching standard definition(s) from the MPW functional dictionary (Fig. 3). In cases where NCBI GI numbers from records match more than one functional definition within the dictionary, each matching definition is recorded. The combination of word-matching and homology-based matching (BLASTP) as described here is able to group and provide functional classification for similar DNA sequences. This feature enhances characterization of large sets of sequence data and enables comparisons between databases obtained from a wide range of organisms.

Traditionally, the user-controlled descriptions found in databases such as NCBI have not supported automated prediction of gene function (21). The lack of a common vocabulary, typographical errors, missing information and use of synonyms in sequence annotations perpetuates this problem and imposes limitations for automated integration of public database records with ontological resources (22). Through extensive use of EC numbers and enzyme name matching

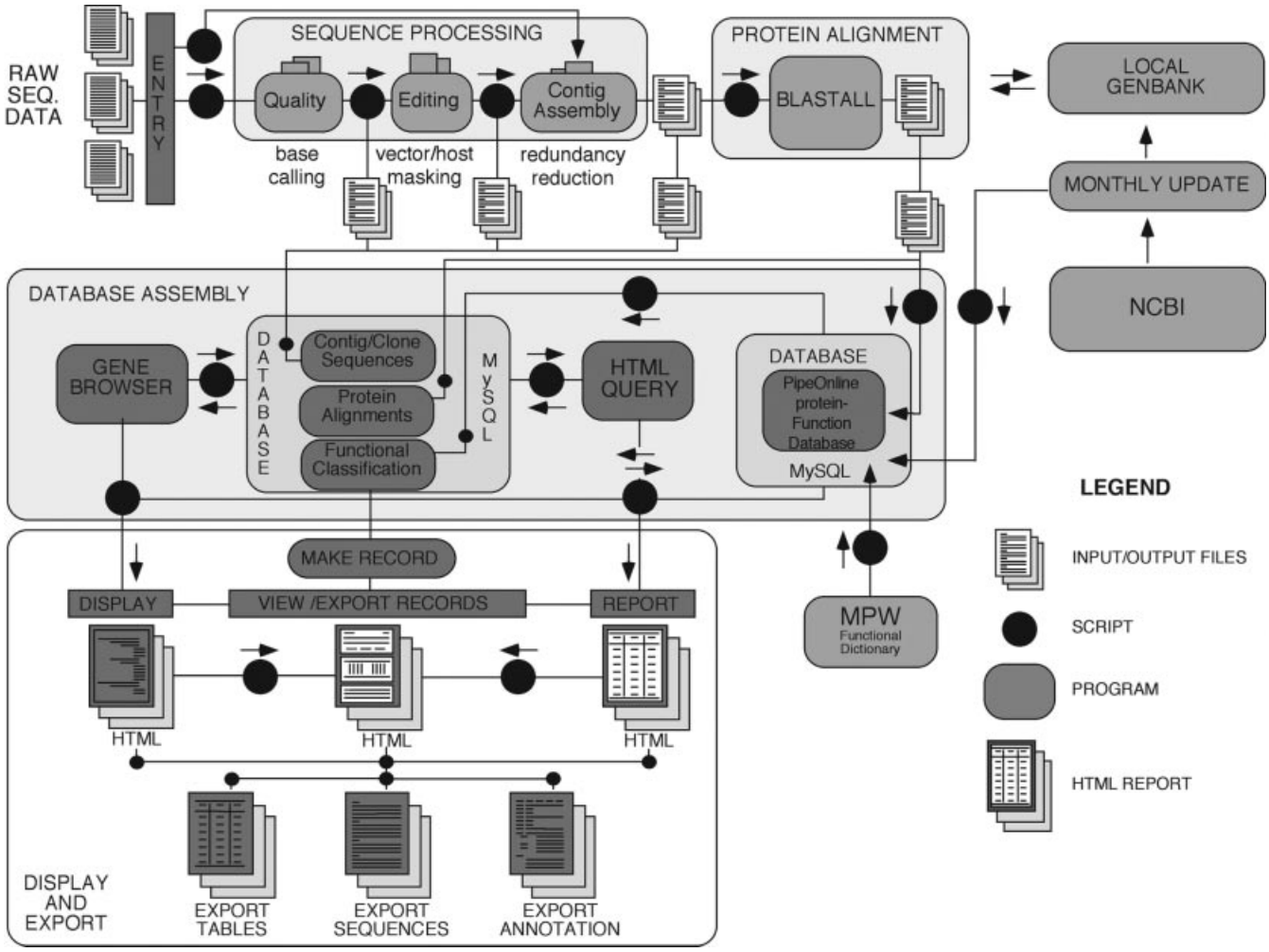

Figure 2. Schematic overview of modules used in PipeOnline for large-scale sequence processing. Users upload chromatograms or FASTA DNA sequence files. Each chromatogram file is converted to DNA sequence, edited for vector sequences and assembled into contigs. Following local BLASTX comparisons with GenBank using NCBI BLASTALL, each output is parsed and a MySQL relational database containing each sequence record and blast output is assembled. This portable database is linked to Web forms permitting keyword searching and browsing of records with links to public databases with export features. BLASTX results are sorted by biological function with an output generated in a Web-browsable form (GeneBrowser). Each step is an independent component module linked together with UNIX scripts, and can easily be updated or replaced as needed.

broadened with the introduction of synonyms and BLASTP protein sequence alignments, we have successfully matched a significant portion (41%) of the NCBI protein database to a functional dictionary providing standard functional definitions (protein and enzyme names).

To test the accuracy of automated functional assignment by the PipeOnline algorithm, we compared the comprehensive yeast genome database (CYGD) catalog of *Saccharomyces cerevisiae* ORFs reported by Munich Information Center for Protein Sequences (MIPS) with the same ORF collection annotated with PipeOnline. The difficulty with such a functional comparison is the difference in organization of catalogs and descriptors used by CYGD-MIPS and PipeOnline. However, Figure 4 shows that distribution of ORFs according to functional classification by CYGD-MIPS or PipeOnline follows a close correlation strongly suggesting that automated and manual functional annotation produce similar results. After thorough visual inspection it appears that disagreements are mainly caused by organizational differences between the two dictionaries where certain functional categories are omitted in one classification schema.

Example processing of model organism ESTs

PipeOnline is best illustrated in assembly of EST sequences from several model organisms. We chose to obtain EST sequences from *Arabidopsis thaliana*, *Aspergillus nidulans*, *Dunaliella selina*, *Hordeum vulgare*, *Mesembryanthemum crystallinum*, *Oryza sativa, S.cerevisiae* and *Selaginella* sp. as model organisms.

For each model organism, PipeOnline was used for automated processing of 30 783 total input raw sequence files resulting in 16 090 unique DNA sequences in eight different databases (Table 4). BLASTX comparison of each contig sequence revealed 8377 or an average of 61% with hits in GenBank (minimum HSP score of 100). Of these, function was assigned to 2879 (35%) contigs, using PipeOnline.

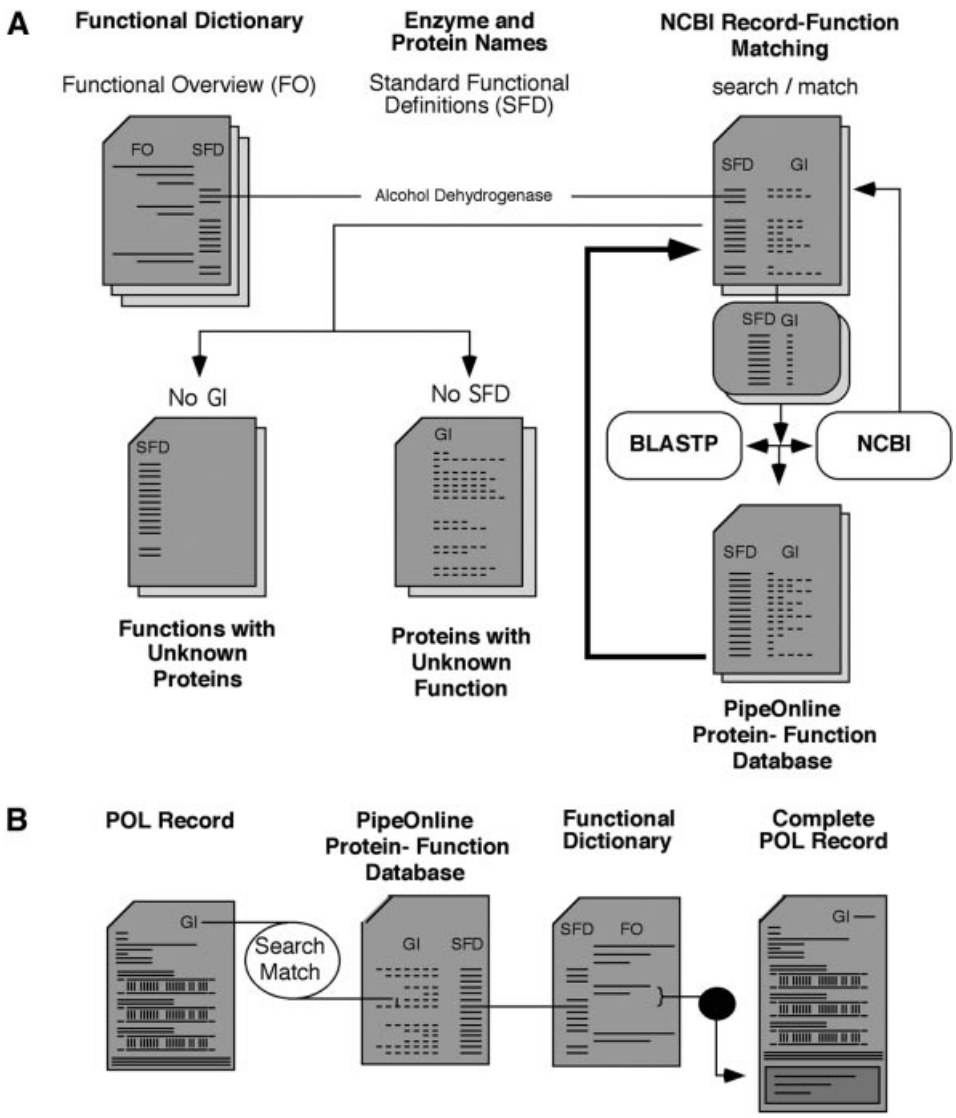

Figure 3. Generation of a functional dictionary of the NCBI protein database and functional sorting of PipeOnline records. (**A**) A MPW functional dictionary was used to correlate standard functional definitions with records from the NCBI protein database using a combination of word-based and protein alignment matching algorithms to generate an NCBI protein-function database. (**B**) This NCBI protein-function database is used as a lookup table for functional assignment of protein alignment descriptions found in PipeOnline records to provide functional classification of records and a functional overview for each database.

Table 3. Composition of the PipeOnline functional dictionary

Functional categorization levels	Total	Classified (%)	Unclassified (%)
MPW functional dictionary			
Functional overview			
Classes and subclasses	4155	–	–
Standard functional definitions	4885	2727 (55.8)	2158 (44.2)
Synonym dictionary			
SWISS-PROT enzyme synonyms	3920	–	–
Total enzyme/protein names	8805	5442 (61.8)	3363 (38.2)
PipeOnline protein-function dictionary			
GenBank records	437 749	179 477 (41.0)	258 272 (59.0)
Classification method			
Word-by-word matching	–	117 102 (26.8)	–
BLASTP	–	62 375 (14.2)	–

The MPW database contains 4885 standard functional definitions of which 2727 are divided (redundantly) into 4155 functional categories. The addition of 3920 synonyms from SWISS-PROT provided 8805 extended functional definitions. Using these extended functional definitions, 117 102 protein records from NCBI were successfully matched to the MPW database. BLASTP protein sequences alignment (expectation value of 1×10^{-40}) of these records matched 62 375 additional records for a total of 179 477 records.

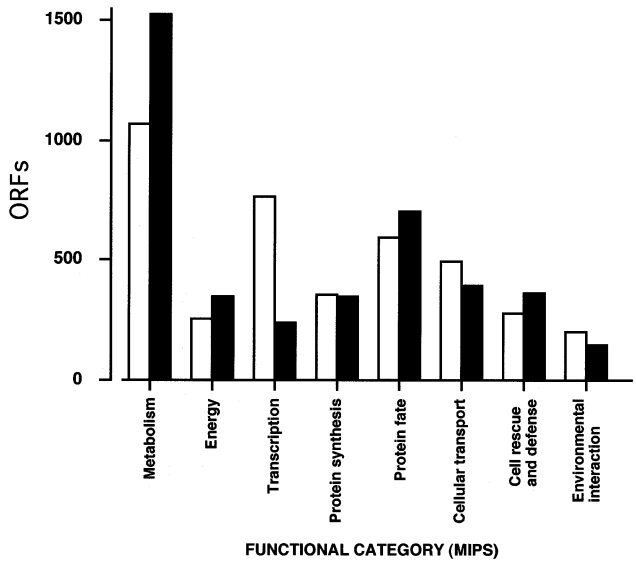

Figure 4. Distribution of ORF functional assignments of *S.cerevisiae* by two independent functional classification methods. Yeast ORFs automatically annotated with PipeOnline (closed bars) were compared with the same ORFs manually annotated by the CYGD at MIPS (open bars).

Databases for each model organism can be viewed and queried at http://stress-genomics.org.

Highly expressed genes represented by redundant sequences were assembled into contigs to generate a consensus sequence of the gene or gene fragment. The number of cDNA sequences used to assemble each contig is reported to provide users with an indication of cDNA/mRNA occurrence. PipeOnline also provides a basic framework for comparing expression profiles of cDNA libraries obtained from the same or different organisms in differing physiological states. While providing a repository for the distribution of sequence data, public databases such as GenBank are unable to support this type of expression analysis and functional data mining (23). Processing of cDNA sequence data using PipeOnline significantly increases the biological value of these data by construction of a database containing information collected from a variety of sources, estimation of biological function and the generation of a functional overview for each cDNA library.

Finally, PipeOnline has been developed primarily as a batch analysis tool for processing of raw cDNA sequence data. The resulting PipeOnline database, GeneBrowser, PipeOnline Annotator and Web-integration function together to enable rapid public release of annotated data. While the MPW functional dictionary has been selected for functional sorting of sequence records, the modular design of PipeOnline coupled with MySQL as a database server permits PipeOnline to be expanded to include functional dictionaries (ontology) from a variety of sources. Expansion of the PipeOnline functional dictionary is being considered with the integration of functional databases such as COG (24), EcoCyc and MetaCyc (25), GO (22) and MIPS (26).

ACKNOWLEDGEMENTS

We thank Hans Bohnert, University of Illinois, Ralph Dean, North Carolina State University, Nancy Keller, University of Wisconsin and Bruce Roe, University of Oklahoma for providing access to data and for their valuable contributions and suggestions before publication. We also thank Phil Green for providing PHRED, CROSSMATCH and PHRAP software. This work was supported by a grant from the Plant Genome Research division of the National Science Foundation (grant no. 9813360).

Table 4. Overview of model organism EST PipeOnline ver2.0 databases[a]

Surveyed EST libraries	ESTs		PipeOnline database record with HSP >100					
	All	Unique	Total		Function		No function	
			n	%	*n*	%	*n*	%
Plant models								
Arabidopsis thaliana	3336	2379	1495	63	422	28	1073	72
Hordeum vulgare	576	537	308	57	116	38	192	62
Mesembryanthemum crystallinum	7327	3788	2096	55	659	31	1437	69
Oryza sativa	3816	2553	1224	48	371	30	853	70
Dunaliella salina	2052	1243	622	50	229	37	393	63
Selaginella sp.	1191	995	769	77	251	33	518	67
All plant models	18 298	11 495	6514	–	2048	–	4466	–
Average	3050	1916	1086	58	341	33	744	67
Fungal models								
Aspergillus nidulans[b]	12 485	4595	1863	41	831	45	1254	67
Saccharomyces cerevisiae[b]	2799	1587	1343	85	373	28	1025	76
All fungal models	15 284	6182	3206	–	1204	–	2279	–
Average	7642	3091	1603	63	602	36	1140	72
All biological systems	30 783	16 090	8377	–	2879	–	5720	–
Average	3848	2011	1047	61	360	35	715	69

[a]Based on the assumption that libraries are mRNA populations representative and tags were selected through random sampling.
[b]ESTs downloaded from NCBI dbEST sequence database using batch ENTREZ. These sequences were then assembled and processed using PipeOnline. Each sequence contig was compared to a local copy of the GenBank non-redundant sequence database using BLASTX. The settings used for BLASTX comparisons included the default settings and a 5×10^{-1} expectation value. In each case, the top five hits were saved and entered into the respective databases. PipeOnline databases can be viewed and searched at http://stress-genomics.org.

REFERENCES

1. Wendl,M.C., Dear,S., Hodgson,D. and Hillier,L. (1998) Automated sequence preprocessing in a large-scale sequencing environment. *Genome Res.*, **8**, 975–984.
2. 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.
3. 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.
4. Audic,S. and Claverie,J.M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.
5. Claverie,J.M. (1996) Effective large-scale sequence similarity searches. *Methods Enzymol.*, **266**, 212–227.
6. Ewing,R.M. and Claverie,J.M. (2000) EST databases as multi-conditional gene expression datasets. *Pac. Symp. Biocomput.*, **5**, 430–442.
7. 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.
8. Quackenbush,J., Cho,J., Lee,D., Liang,F., Holt,I., Karamycheva,S., Parvizi,B., Pertea,G., Sultana,R. and White,J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.*, **29**, 159–164.
9. Quackenbush,J., Liang,F., Holt,I., Pertea,G. and Upton,J. (2000) The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.*, **28**, 141–145.
10. Yuan,J., Liu,Y., Wang,Y., Xie,G. and Blevins,R. (2001) Genome analysis with gene-indexing databases. *Pharmacol. Ther.*, **91**, 115–132.
11. Zhuo,D., Zhao,W.D., Wright,F.A., Yang,H.Y., Wang,J.P., Sears,R., Baer,T., Kwon,D.H., Gordon,D., Gibbs,S. *et al.* (2001) Assembly, annotation and integration of UNIGENE clusters into the human genome draft. *Genome Res.*, **11**, 904–918.
12. Lee,Y., Sultana,R., Pertea,G., Cho,J., Karamycheva,S., Tsai,J., Parvizi,B., Cheung,F., Antonescu,V., White,J. *et al.* (2002) Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res.*, **12**, 493–502.
13. Lonsdale,D., Crowe,M., Arnold,B. and Arnold,B.C. (2001) Mendel-GFDb and Mendel-ESTS: databases of plant gene families and ESTs annotated with gene family numbers and gene family names. *Nucleic Acids Res.*, **29**, 120–122.
14. Selkov,E.J., Grechkin,Y., Mikhailova,N. and Salkov,E. (1998) MPW: the Metabolic Pathways Database. *Nucleic Acids Res.*, **26**, 43–45.
15. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
16. Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.
17. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
18. Overbeek,R., Larsen,N., Pusch,G.D., D'Souza,M., Selkov,E.,Jr, Kyrpides,N., Fonstein,M., Maltsev,N. and Selkov,E. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, **28**, 123–125.
19. Bairoch,A. (1999) The ENZYME data bank in 1999. *Nucleic Acids Res.*, **27**, 310–311.
20. Andrade,M.A., Brown,N.P., Leroy,C., Hoersch,S., de Daruvar,A., Reich,C., Franchini,A., Tamames,J., Valencia,A., Ouzounis,C. *et al.* (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15**, 391–412.
21. Karp,P.D. (2000) An ontology for biological function based on molecular interactions. *Bioinformatics*, **16**, 269–285.
22. 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.
23. Tsoka,S. and Ouzounis,C.A. (2000) Recent developments and future directions in computational genomics. *FEBS Lett.*, **480**, 42–48.
24. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
25. Karp,P., Riley,M., Saier,M., Paulsen,I., Paley,S. and Pellegrini-Toole,A. (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res.*, **28**, 56–59.
26. Mewes,H., Frishman,D., Gruber,C., Geiser,B., Haase,D., Kaps,A., Lemcke,K., Mannhaupt,G., Pfeiffer,F., Schuller,C. *et al.* (2000) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **28**, 37–40.