Genetics and Molecular Biology, 30, 3 (suppl), 1024-1029 (2007)
Copyright by the Brazilian Society of Genetics. Printed in Brazil
www.sbg.org.br

Special Case

Bioinformatics for the Citrus EST Project (CitEST)

Marcelo S. Reis[1], Marco A. Takita[1,2], Darío A. Palmieri[1,3] and Marcos A. Machado[1]

[1]Centro APTA Citros Sylvio Moreira, Instituto Agronômico de Campinas, Cordeirópolis, SP, Brazil.
[2]Centro de Recursos Genéticos Vegetais, Campinas, SP, Brazil.
[3]Laboratório de Estudos do Meio Ambiente, Universidade Católica de Salvador, Salvador, BA, Brazil.

Abstract

In this work we describe all the computational environments, pipelines, and web services developed for the CitEST transcriptome project, on which all the annotation researchers relied. We also present a complete list of CitEST libraries and, for each of them, the general features after the *in silico* processing, showing some quantitative information.

Key words: transcriptomics, comparative genomics, clustering, trimming, digital northern.

Received: July 24, 2006; Accepted: March 6, 2007.

Introduction

With more than 300,000 clones sequenced, the Citrus EST Project (CitEST) is a large transcriptome sequencing project, and has the purpose of integrating knowledge on genetic mapping with new data on functional and comparative genomics of Citrus species and related genera. The project focuses on the response of plants under different biotic and abiotic stresses. The scope of the project also included the integration of the functional genomic studies of citrus pathogens whose genomes have been fully sequenced (virus and bacteria) into the scenario of host-pathogen interactions.

The sequencing stage was carried out almost entirely by the Citrus Biotechnology Laboratory (CBL), while the data mining and analysis stages of the project had the collaboration of researchers from several different laboratories, and employed a consortium-based model used since the very first Brazilian genome project, *Xylella fastidiosa* (Simpson *et al.*, 2000). Bioinformatics has a key role in this project, since it is responsible for storing and processing all the EST data, performing the subsequent analyses, and making them available in a web-based, user-friendly data-mining system.

In this paper, we begin by describing the computational resources of the project, and the services related to sequence submission and data storage. After that, the processing of the sequences is presented, and the trimming, clustering, and comparative genomics pipelines are explained in detail, including the software and biological databases used for the procedures. The list of CitEST cDNA libraries is also presented, showing some quantitative information obtained from the *in silico* processing. Finally, we comment on computational resources and key aspects that must be considered during transcriptome projects with a large amount of data. We also present some current works: the improvement of the clustering procedure, and the digital northern software.

Methods and Results

The CitEST website (http://biotecnologia.centrodecitricultura.br) is a provider for several web services, related to sequence submission and data mining. Although the general organization of users, services, data, and programs resemble previous Brazilian consortium-based transcriptome projects, namely the Sugarcane EST (Vettore *et al.*, 2001), and *Schistosoma mansoni* EST (Verjovski-Almeida *et al.*, 2003), CitEST has its own improvements, and some new tools developed by the CBL team. Figure 1 shows a diagram for the general modules in the CitEST system.

Computational systems

The CitEST website, data repositories, web services, and programs are hosted by a Sun V880, with four Sparc 900 Mhz processors, seven 73 Gb hard disks, 8 Gb RAM, and Solaris 8 operating system. Some computer-intensive analyses were performed using a Beowulf, linux-based cluster, composed of nine Dell-Dimension 4700C CPUs,

Send correspondence to Marcelo S. Reis. Centro APTA Citros Sylvio Moreira, Instituto Agronômico de Campinas, Caixa Postal 4, 13490-970 Cordeirópolis, SP, Brazil. E-mail: marcelo@centrodecitricultura.br.

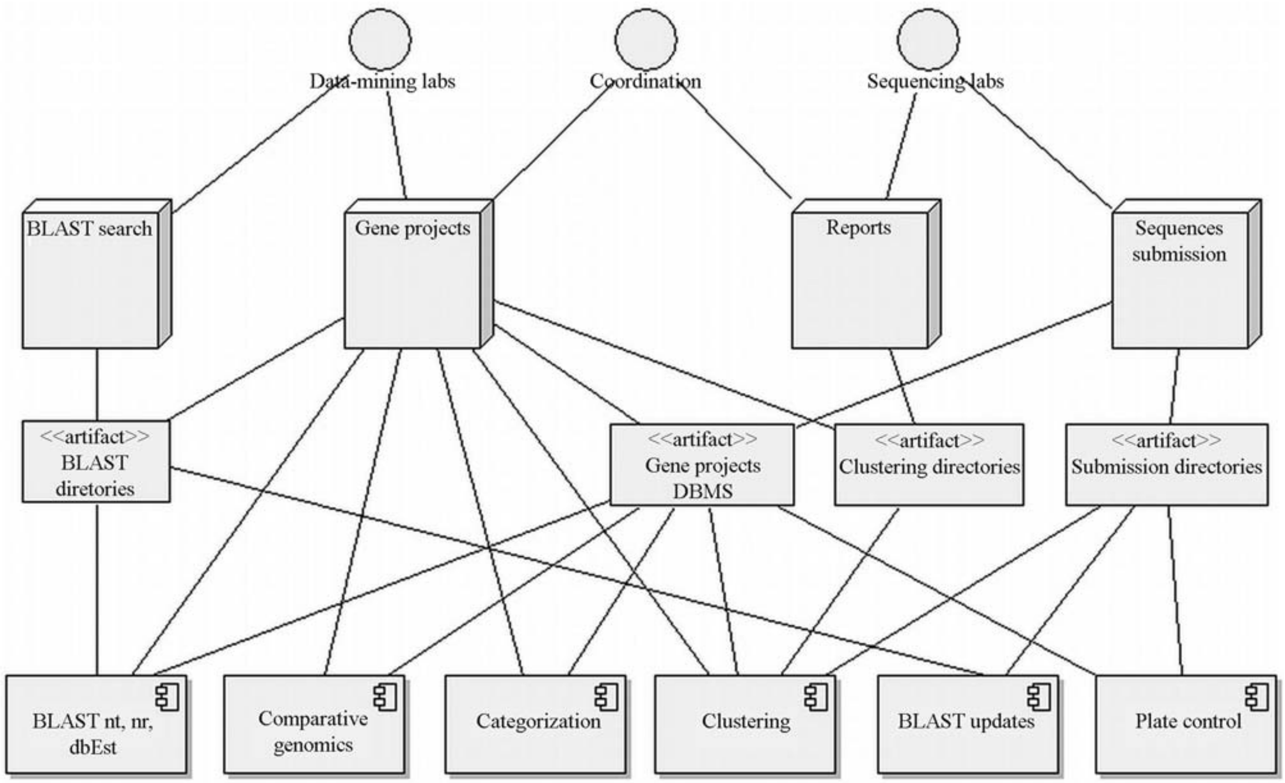

Figure 1 - Diagram of the CitEST transcriptome project systems. There are three main interfaces: one for the Data-mining Laboratories, another for the Coordination, and a third for the Sequencing Laboratories.

with four 3.2GHz Intel CPU, 150 Gb hard disk, and 1 Gb RAM, running a Debian/Linux operating system.

The web server used is Apache, version 2.0.54. The chosen relational database management system was the MySQL, version 5.0. The web services were written in Perl, version 5.8.4.

A remarkable feature is the use of a proper ongoing project manager, the Gene Projects (Carazzolle *et al.*, this issue). Using Gene Projects, the data-mining user works initially on reads, creating a project and adding reads to it. Furthermore, the user may select some or all reads in the project, and run a clusterization. In a cluster-based manager, like the one used in the Sugar Cane EST Project – SUCEST, the data-mining users could work only on already clustered sequences. The CitEST version of Gene Projects had some improvements coded by the CBL team. The improvements will be explained in the "Clustering" section.

cDNA and shotgun libraries

CitEST has a comprehensive list of cDNA and whole-genome shotgun libraries, including seven different *Citrus* species, and one species from related genera: *Poncirus trifoliata.* It also has sequences from a related pathogen, *Phytophthora parasitica* (a fungi). Table 1 shows a complete list of CitEST libraries and their features after the *in silico* processing, grouped by species.

In order to properly store the reads (chromatograms) to the CitEST system, a nomenclature convention was created, including all species, varieties, conditions, sequencing strategies, conditions, plates, the reads, and the sequencing orientation. For instance, a read with the label "CS00-C1-100-001-A01.F" could be described as:

* "CS00": species (*Citrus sinensis*) and variety (Pêra IAC);

- "C1": strategy (cDNA) and tissue (leaves);
- "100": condition (healthy tissue);
- "001": plate;
- "A01": read;
- "F": orientation (forward; R for reverse).

The complete list of CitEST nomenclatures can be found at http://biotecnologia.centrodecitricultura.br/cbl/nomenclature.htm. The list of nomenclatures is comprehensive, covering all the libraries presented in Table 1, and also species and/or conditions not sequenced yet.

Sequence submission

The input data, consisting mainly of chromatograms produced by ABI 3700 and ABI 3730 automatic sequencers (Applied Biosystems), were received through a submission system originally developed by the LBI team (Laboratory for Bioinformatics, Institute of Computing, University of Campinas). The system allows the submission of a zipped

Table 1 - List of CitEST libraries, grouped by species. Valid reads are the remaining reads after the trimming pipeline; sequencing efficiency is the ratio between the valid reads and the submitted reads; assembled sequences are the contigs and the singletons obtained using the clustering pipeline; redundancy is one minus the ratio between the assembled sequences and the valid reads, and it is related to the redundancy of transcripts in a given species.

Species	Strategies	Tissues	Conditions	Number of submitted reads	Number of valid reads	Sequencing efficiency	Assembled sequences	Sequencing redundancy
Citrus aurantifolia	cDNA	Leaf	Healthy greenhouse plant	9,600	8,219	85.61%	4,327	47.35%
Citrus aurantium	cDNA	Leaf	Healthy field plant	9,600	8,439	87.91%	5,607	33.55%
Citrus latifolia	cDNA	Leaf	Healthy greenhouse plant	9,600	5,484	91.21%	4,883	44.23%
Citrus limettiodes	cDNA	Leaf	Infected with CiLV Virus	9,600	8,188	85.29%	4,540	44.55%
Citrus limonia	cDNA	Root	With and without hydric stress	15,455	11,045	71.47%	5,945	46.17%
Citrus reticulata	cDNA	Fruit, leaf	Fruit development stages, infected or not with *Xylella fastidiosa*	59,484	51,679	86.88%	18,873	63.48%
Citrus sinensis	cDNA, and whole genome shotgun	Bark, flower, fruit, leaf	Healthy greenhouse plant, healthy young plant, fruit development stages, infected or not with *Xylella fastidiosa*, infected or not with CiLV	126,660	108,611	85.75%	32,121	70.42%
Citrus sunki	cDNA	Bark	Healthy plant	7,584	5,216	68.78%	2,652	49.15%
Poncirus trifoliata	cDNA	Bark, leaf, seed	Infected or not with *Phytophthora*, infected or not with CTV	38,976	32,637	83.74%	12,873	60.55%
Phytophthora parasitica	cDNA	leaf	within *Poncirus trifoliata* leaves	25,152	17,529	69.69%	5,518	68.52%
All species	-	-	-	311,711	260,319	83.51%	-	-

file containing 96 electropherograms through the web. User authentication is required for accessing the service.

After receiving the plate, the submission system checks that all the chromatograms are in accordance with the nomenclature. If there is a problem at this point, the submitted plate is automatically rejected. Otherwise, the phredPhrap package, version 0.990329 is started, producing sequences and qualities files in fasta format, with the contaminants (adaptor and vector sequences) masked using the cross-match software (included in the phredPhrap package). If either the average base quality of all reads is less than 20 (phred log error probability), or there are more than 20% of masked sequences, the plate is automatically rejected.

For each accepted plate, a HTML submission report is produced and displayed to the user. At this point, the user can access a submission summary, containing information concerning base quality, adaptor and vector masking. It shows both the information for each single read and the average for all 96 reads. Therefore, by analyzing the HTML report, the submission user may accept or reject the submission.

The accepted plates are stored into the submission directories. If there is already a previous version of the plate, it is then replaced by the newer one.

BLAST updates

In the CitEST environment, there are some scheduled BLAST (Altschul *et al.*, 1997) report updates. It uses the blastall implementation of the BLAST algorithm, version 2.0. The system scans for new submitted reads daily. If there is a new read, it is automatically compared with GenBank proteins, and the report results are stored in the BLAST directories, and also in the Gene Projects database management system (DBMS).

Monthly, a scheduled BLAST report is updated for the all CitEST libraries: it retrieves the latest GenBank protein database version from the NCBI website, formats it using the formatdb program (version 2.0), and compares all the CitEST's reads against the updated GenBank.

Trimming

The trimming stage has a major influence on the work performed by the assembler, as well as on the produced contigs and singletons. As unwanted sequences are eliminated (*e.g.,* poly(A), slippages – echoed bases, low base quality, chimeras, vectors, adaptors, and others), the readability of the assembly is improved and, at the same time, the computational clustering process becomes less expensive.

In the CitEST system, the Gene Projects uses the trimming methodology described by Telles and da Silva (2001), and applied in the SUCEST bioinformatics pipeline (Telles *et al.*, 2001). The methodology includes:

- Ribossomal RNA sequences removal through comparison against rRNA from another *Plantae* species, such as 18S rRNA from *Zea mays* (GenBank ID: AF168884), 5.8S rRNA from *Platanus occidentalis* (GenBank AF162215), and 26S rRNA from *Lambertia inermis* (Genbank AF274652). The comparison was performed using the blastall program, version 2.0;
- Vector and adapter sequence masking and removal, using the cross-match program;
- Quality trimming, removing sequences with phred quality below 20 within a 12-sized window;
- Slippage trimming, discarding artifacts (Anon, 1998) with more than 10 echoed bases. This procedure takes into account, for each sequence, its bases' quality;
- Poly A/T trimming, using the cross-match program;
- Cutoff length, removing, at the end of the procedure, any read with less than 100 bases, or with less than 50 bases and phred quality equal to or greater than 20.

In their work, Telles and da Silva (2001) present the default values used for the trimming of the SUCEST cDNA libraries. Nevertheless, some adjustments were needed for the CitEST libraries, especially concerning the cutoff values for echoed bases.

Clustering

In order to reduce the redundancy of the libraries, and also to make the data-mining work easier for the user, clustering functionality is available in Gene Projects. It allows the user to choose sequences from the libraries previously inserted into a user's project, and to start a clusterization on them.

The Gene Projects use the CAP3 assembly tool (Huang and Madan, 1999), with the default parameters for overlapping cutoff length (40 bp), and for overlapping percent identity cutoff (80% of read coverage). The sequences are trimmed before clustering, through the trimming pipeline previously described.

After running CAP3 on the trimmed reads, the generated tentative consensus sequences (contigs) and singletons are stored on the Gene Projects DBMS. After that, the system performs an automatic categorization and comparative genomics, both analyses are implemented by the CBL team for the CitEST transcriptome project. All the automatic categorizations, comparative genomics, and all the biological databases available in the CitEST are listed in the "Internet Resources" section.

Functional categorization

Automatic categorizations are carried out twice in the CitEST Gene Projects: after a new read from a recently-submitted plate, entry on the Gene Projects DBMS; and after a clusterization, requested by the data-mining user, is finished. There are two kinds of categorizations: the functional categories, from the Munich Center for Protein Sequences (MIPS), version 2.0; and the Gene Ontology (G.O.).

The MIPS categorization is gathered through a BLAST comparison of the query sequence (contigs, singletons and reads) against the MIPS' *Arabidopsis thaliana* annotated proteins. This retrieves the MIPS accession number from the best hit (considering a minimum e-value of 10^{-5} and at least 80% identity), which, in turn, retrieves the functional category from the MIPS FunCat table. A web link to the MIPS website, for a given term, brings more information concerning the functional category for the given sequence.

The G.O. Term is gathered through a BLAST comparison of the query sequence against a recent Gene Ontology protein database. The BLAST cutoffs are the same for the MIPS comparison, and for the G.O. Terms they are extracted from the best hit. A web link to AmiGO, for a given term, retrieves further information concerning the query sequence.

Comparative genomics

The identification of sequences (consensi, singletons and reads) was made through the comparison with other organisms, using some public protein databases, such as the GenBank and TrEMBL. To increase the accuracy of the identification, comparisons are also made using the Swiss-Prot, a curated protein database, and the Pfam protein domain database (Finn *et al.*, 2006).

For the comparison using BLAST, the cutoff criteria are the same for automatic categorization: minimum e-value of 10^{-5}, and at least 80% query sequence coverage. HTML BLAST reports are generated as well, and all results are stored on the Gene Projects DBMS.

The comparison against Pfam uses the HMMER implementation for Hidden Markov Model (HMM) profile databases. The sequences are translated into the six possible frames, and all of them are compared with the Pfam domain profiles. The cutoff criteria are: e-value of 10^{-5} or lower and a HMMER score 50 or higher.

Metabolic pathways

Another improvement for the CitEST Gene Projects, made by the CBL team was the addition of a web link to the Metabolic Pathways database from The *Arabidopsis* Information Resource (TAIR). The comparison of the query sequence against the MIPS' *A. thaliana* protein sequences gives the MIPS accession code for the best hit. This code can be used to get, through a table search, the metabolic pathway map code (PWY) at TAIR. Note that this feature is limited, since there are considerably fewer proteins in metabolic pathway maps than annotated proteins. If a protein is present in more than one map, the Gene Projects make web links available for every map the protein is in.

Digital Northern

One of the most important ways to analyze a collection of cDNA libraries is to perform an *in silico* assessment of differentially expressed genes (often called digital northern, or *in silico* hybridization). In the literature, one can find methods to measure the abundance of gene transcripts in cDNA libraries. However, there is no tool available to compare the results obtained with the most important methods and to perform automatic data analysis. Therefore, there was the need for development of a new digital northern tool for the CitEST data-mining users.

The pipeline of the Digital Northern Tool includes the following: the choice of two or more cDNA libraries; automatic generation of tentative consensus sequences (not related to the Gene Projects clustering procedure); the computing of several different statistics for each consensi, namely, the test statistic P-Value (Audic and Claverie, 1997), the entropy among cDNA libraries (Stekel *et al.*, 2000), the relative abundance for each library, and the Fischer Exact Test. Automatic comparisons between the contig and public protein repositories (*e.g.* GenBank, UniProt, KEGG and MIPS) are generated as well. It prints the results onto an HTML table, one contig per row, performing automatic functional categorization (based on MIPS functional categories), showing the results of the comparison against public databases, and creating pictures illustrating graphically the differential expression among the libraries. Figure 2 shows an example of a picture generated by the digital northern tool with some *Citrus sinensis* assembled sequences, where the conditions of the comparison are four different fruit development stages.

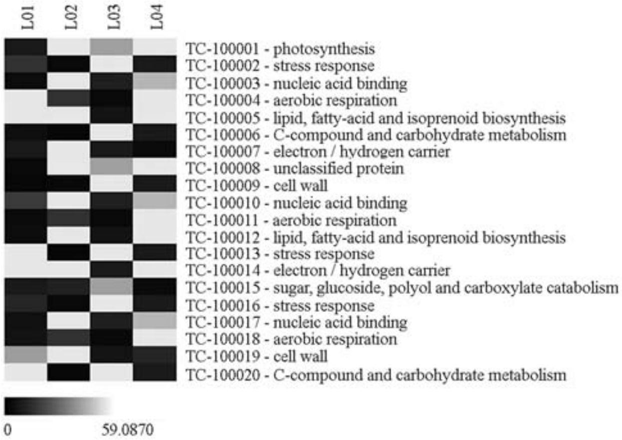

Figure 2 - Assessment of some *Citrus sinensis* differentially expressed genes through Digital Northern. The rows are the tentative consensi (TC), and the columns are the libraries used to perform the clusterization; in this example, four different fruit development stages were used. For a given tentative consensi and column, the intensity of color denotes its relative abundance, taking into account the number of composing reads and the size of the libraries.

Discussion and Outlook

Computational resources

Some key aspects certainly are the computational resources. During the development of CitEST, we noticed three bottlenecks: the constant growth of public sequence repositories, the number of system users, and the size of EST libraries.

The number of sequences hosted by the public repositories (*e.g.* GenBank, UniProt) is increasing exponentially (GenBank website, January 10, 2007). However, the capacity of hard disks, CPU memory, and speed do not follow such growth. Nowadays, the sequence clustering, assembly, and comparison are much more computer-intensive and time-consuming than they were at the time the first Brazilian consortium-based genome projects were created.

Since it is not possible to take sequences out from the assembly process, nor reduce the protein databases for comparison, we recommend the replacement of the most time-consuming tools with their parallel versions. Both of the most important ones for CitEST have parallel versions available: PCAP (Huang *et* al.*,* 2003) for CAP3, and a Beowulf configuration for the blastall implementation of the BLAST algorithm. A parallel comparison with multiple sequences against large protein databases can be done up to 13 times faster than a single node for a cluster with 20 nodes (Mathog, 2003).

Clustering improvements

We are working on a pipeline for the improvement of assembling and clustering procedures. The original version of Gene Projects offers an improvement using BLAST saturation (Carazzolle *et al.*, in this issue), where an iterative heuristic procedure tries to enlarge the consensi through BLAST comparison between the consensi themselves and the remained singletons. Our proposal is to set up an expert system that includes new assembling rules, taking into account biological information from the analyzed sequences, in a similar way to that presented in Chevreux *et al.*, (2004). Thus we expect to increase the quality of the assemblies, and also perform single nucleotide polymorphisms (SNPs) detection.

Digital Northern improvements

A web interface for our Digital Northern Tool is also under development (currently it is a terminal-based tool). Once it is finished, it will be available at the CitEST website, in a first stage as a stand-alone tool; after that, the next step is the integration of the Digital Northern Tool into the Gene Projects.

Apart from the interface improvements, we are working on a gene clusterization, using the information from the relative abundance of transcripts. For that, a modified k-means clustering heuristic algorithm is being developed, using the algorithm presented by Kanungo *et al.* (2002).

Acknowledgments

This work was supported by CNPq and FAPESP. We would also like to thank all the researchers who helped us with system maintenance and improvement, namely Marcelo Carazzolle, Gustavo Astúa-Monge, Luciano Kishi, and Vinicius de Lima.

References

Altschul SF, Madden TL, Schaffer AA, Zhang, J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Anon (1998) Chemistry Guide for Automated DNA Sequencing. Applied Biosystems, Foster City, 242 pp.

Audic S and Claverie JM (1997) The significance of digital gene expression profiles. Genome Res 7:986-995.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T and Suhai S (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 6:1147-1159.

Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, *et al.* (2006) Pfam: Clans, web tools and services. Nucleic Acids Res 34:D247-D251.

Huang X and Madan A (1999) CAP3: A DNA assembly program. Genome Res 9:868-877.

Huang X, Wang J, Aluru S, Yang SP and Hillier L (2003) PCAP: A whole-genome assembly program. Genome Res 13:2164-2170.

Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R and Wu AY (2002) An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24:881-892.

Mathog DR (2003) Parallel BLAST on split databases. Bioinformatics 19:1865-1866.

Simpson AJ, Reinach FC, Arruda P, Abreu FA, Acencio M, Alvarenga R, Alves LM, Araya JE, Baia GS, Baptista CS, *et al.* (2000) The genome sequence of the plant pathogen *Xylella fastidiosa*. Nature 406:151-159.

Stekel DJ, Git Y and Falciani F (2000) The comparison of gene expression from multiple cDNA libraries. Genome Res 12:2055-2061.

Telles GP and da Silva FR (2001) Trimming and clustering sugarcane ESTs. Genet Mol Bio 24:17-23.

Telles GP, Braga MDV, Dias Z, Lin TL, Quitzau JAA, da Silva FR and Meidanis J (2001) Bioinformatics of the sugarcane EST project. Genet Mol Bio 24:9-15.

Verjovski-Almeida S, DeMarco R, Martins EAL, Guimarães PEM, Ojopi EPB, Paquola ACM, Piazza JP, Nishiyama MY, Kitajima JP, Adamson RE, *et al.* (2003) Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. Nature Genet 35:148-157.

Vettore AL, da Silva FR, Kemper EL and Arruda P (2001) The libraries that made SUCEST. Gen Mol Bio 24:1-7.

Internet Resources

AmiGO database, http://www.godatabase.org (January 10, 2007).

Apache web server, http://www.apache.org.

GenBank protein database, http://www.ncbi.nlm.nih.gov/Genbank (December 20, 2006).

Gene Ontologies (GO), http://www.geneontology.org (December 20, 2006).

HMMER Software, http://hmmer.janelia.org.

Kyoto Encyclopedia of Genes and Genomes (KEGG), http://www.genome.jp/kegg (December 20, 2006).

MIPS Functional Categories (FunCat), http://mips.gsf.de/projects/funcat (November 10, 2006).

MySQL relational database management system, http://www.mysql.com.

PhredPhrap Software, http://www.phrap.org.

Practical extraction and report language (Perl), http://www.perl.org.

Protein Families (PFAM), http://www.sanger.ac.uk/Software/PFAM (November 10, 2006).

Swiss-Prot curated protein database, http://www.expasy.org/sprot (December 20, 2006).

The *Arabidopsis* information resource (TAIR), http://www.arabi dopsis.org (January 10, 2007).

Transport classification database (TCDB), http://www.tcdb.org (November 10, 2006).

TrEMBL computer-annotated protein database, http://www.ebi.ac.uk/trembl (December 20, 2006).

Associate Editor: Reinaldo Montrazi Barata