Genetics and Molecular Biology, 30, 3 (suppl), 1030-1036 (2007)

www.sbg.org.br

Research Article

Gene projects: A genome Web tool for ongoing mining and annotation applied to CitEST

Marcelo F. Carazzolle[1], Eduardo F. Formighieri[1], Luciano A. Digiampietri[2], Marcos R.R. Araujo[1], Gustavo G.L. Costa[1] and Gonçalo A.G. Pereira[1]

[1]*Laboratório de Genômica e Expressão, Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, SP, Brazil.*
[2]*Laboratório de Sistemas de Informação, Departamento de Sistemas de Informação, Instituto de Computação, Universidade Estadual de Campinas, Campinas, SP, Brazil.*

Abstract

Genome projects, both genomic DNA and ESTs (cDNA), generate a large amount of information, demanding time and a well-structured bioinformatics laboratory to manage these data. These genome projects use information available in heterogeneous formats from different sources. The amount and heterogeneity of this information, as well as the absence of a world consensus pattern, make the integration of these data a difficult task. At the same time, sub-tasks, such as microarray analyses of these projects, are very complex. This creates a demand for the development of creative solutions for ongoing annotation, thematic projects, microarray experiments, etc. This paper presents Gene Projects, a system developed to integrate all kinds of solutions.

Key words: annotation, genome, web, microarray and clustering.

Received: July 21, 2006; Accepted: May 4, 2007.

Introduction

Large-scale sequencing projects often target either genomic or cDNA sequences. For genomic sequences, the most used sequencing strategy is the Whole Genome Shotgun - WGS (Venter *et al.*, 1998), where the DNA is broken into fragments that are sequenced randomly. The goal is to assemble these fragments to reconstitute the whole genome sequence or, at least, to obtain an ordered list of large genomic regions with interspersed gaps. In the latter case, we say that it is a draft genome. On the other hand, cDNA sequencing projects directly examine the genes expressed under different conditions. The cDNA is synthesized using cellular mRNA as the template, thus it samples the pool of active genes in a specific cellular condition. The fragments of sequenced cDNA are called ESTs (Expressed Sequence Tags), thus, cDNA sequencing projects are often called EST genome projects. The ESTs are often assembled to improve the length and the quality of the ESTs, to obtain the current set of full sequence of cDNAs and to identify expression patterns from different studied conditions.

Send correspondence to Marcelo F. Carazzolle. Laboratório de Genômica e Expressão, Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, Caixa Postal 6109, 13083-970 Campinas, SP, Brazil. E-mail: mcarazzo@lge.ibi.unicamp.br.

Genome projects produce a large amount of reads, which are chromatograms representing sequence fragments. This sequencing stage usually lasts for many months. On the other hand, the available tools for retrieving information from these incoming data require expertise in bioinformatics and a tremendous amount of computational capacity. For this reason, the end user typically has access to this information only after all of the sequencing has been completed. Furthermore, small projects are almost never able to afford their own bioinformatics infrastructure.

At the same time, there are different kinds of information related to genome projects that can be produced and utilized during the process of assembly and annotation of the genome. For example, there is a strong relationship between a genome project and the choice of clones to be analyzed by microarrays, as well as the inverse pathway that is the selection of sequenced clones from microarray analyses. To favor this kind of interaction, it is necessary to have an efficient search mechanism and a robust analysis mechanism, both with the capacity to deal with thousands of clones and minimize the possibility of human errors. Furthermore, the annotation of genes (identification and classification of genes) and other complex activities need the interaction of several kinds of heterogeneous data, databases and tools. Special care must also be taken since there is no worldwide consensus for the annotation nomenclature

used and since there are no guarantees that all previous data were annotated correctly.

A way to minimize these potential errors is for specialists in the different areas involved in the annotation to work together to process each genome. However, this is not always practical because these specialists are spread around the world.

In this context of increasing needs for sophisticated data analysis and user helpful interfaces in genome projects, we developed a system called Gene Projects (Patented - INPI Protocol no. 0057290, in 09.01.2004). The Gene Projects (GP) system can manage small and large genome projects (both, DNA and EST projects) and it allows annotation to occur at the same time as the sequencing, without the need for computer specialists. This process is also able to separate analyses by theme:, a process we called "projects". Furthermore, this provides a mechanism for multiple tools and techniques to interact, facilitates searches of annotated genes, allows Web access and can be used by researchers that do not have a specialized knowledge of computers or bioinformatics.

Projects and Ongoing Annotation

The system is called Gene Projects due to the importance of what we call projects (short for thematic projects). Each user of the system has a specific login and password. With this login he/she can create and manage his/her projects, making it a flexible and powerful way to conduct data mining. A project is a structure inside the system where researchers can develop and organize their thematic studies. The user can add reads through several search mechanisms such as BLAST (Altschul *et al.*, 1997) results, read names, keywords, etc. Once the reads are chosen, the user can assemble them, view and edit the assembly results, improve the quality of the contig and enlarge the contig (Saturation Blast), find ORFs, select sub-sequences of a contig (seeds) for future annotation, obtain the map of the 96 well plates used in the sequencing or of the 384 wells plates used in microarray experiments, among other uses. Through these projects, it is possible to find, select and study annotated genes of interest in order to study genomic functions such as infection related genes or genes that belong to specific metabolic pathways, for example. One of the advantages of this system is that all of these things can be done during the sequencing stage through the gene annotation interface.

System Description and Architecture

Gene Projects is written in perl (http://www.perl.org) and uses standard, open source modules such as CGI.pm (http://stein.cshl.org/WWW/software/CGI), GD.pm (http://www.boutell.com/gd/) and DBI.pm (http://dbi.perl.org/). The system needs the Linux operating system with a Web server (*e.g.*, Apache - www.apache.org) and MySQL server (www.mysql.com) installed.

As a default, Gene Projects also utilizes some programs, databases and directory structure for its execution. Table 1 shows the main required programs.

Figure 1 shows the main characteristics of the system. The Web interface allows users around the world to access our system. The operations carried out by the users are processed locally in the processing module.

GP has three main integrated functions: data mining, annotation and microarray management. Data mining and annotation are strongly related. Data mining processes help the scientists analyze huge amounts of data and thus facilitate the annotation. On the other hand, the annotation process can identify terms to be mined, such as genes of one specific metabolic pathway.

Microarrays can be built with known cDNA (cDNA chip) or with random ESTs (blind chip). Gene Projects fa-

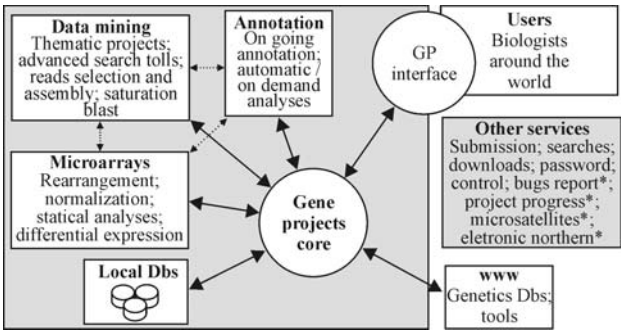

Figure 1 - The Gene Projects architecture. There are three main modules: data mining, annotation and microarrays. The Gene Projects core manages and integrates these modules, the access to local databases, web databases and tools, and the user interface. The asterisks indicate the tools that are not adapted yet for the CitEST Project.

Table 1 - List of programs required to execute Gene Projects.

Program	Description	Reference/URL
Phred	Basecalling and generation of quality values from chromatograms	(Ewing and Green, 1998)
Cross_match	Vector screening and generation of FASTA sequence files with masked vector sequences	http://www.phrap.org
Phrap	Clustering and assembling program for shotgun sequences	http://www.phrap.org
CAP3	Clustering and assembling program for EST sequences	(Huang and Madan, 1999)
Local Blast	Sequence alignment	(Altschul *et al.*, 1997)
Fuzznuc	Pattern search	Emboss package (http:// emboss.sourceforge.net)

cilitates the selection of cDNA of interest for the assemblage of cDNA chips (through data mining and annotation). The selection of ESTs for blind chips is facilitated by the automatic identification of ESTs with a given characteristic of interest, for example, ESTs that presented the same expression pattern. In other words, blind chips can indicate themes for metabolic studies, which can become new *projects* in Gene Projects.

A search tool is one of the most important activities in systems that manage genome projects. Gene Projects has a graphic tool for "Advanced Search" that allows boolean queries, allowing for the easy use of queries over the main fields in the databases. Another important function of the system is the set of tools that allows the assembly of a set of reads for each *project* and the viewing of the results. This process typically needs an experienced end user; however, with the Gene Projects Web interface, the process is very simple.

The main advantage of this architecture is that our servers do the majority of the processing, which makes it unnecessary for external users to have high performance computers with large amounts of memory. The system components that demand the most CPU time are executed as child processes, running independently of the rest of the system This permits the user to disconnect and, when he reconnects, the results will be ready. When there are several requested processes to be executed at the same time, they are stored and executed using a queue scheme. The results are stored on our database and can be queried at any time.

Technological Advantages

Our approach is based on perl script and is available via CGI technology. There are some advantages to this approach:

- Scalability: due to the fact that the internal computer infrastructure of our laboratory is hidden by the Web interface, we have the freedom to alter our processing capabilities without changing any user interface, for example, using a distributive infrastructure, such as a computational Grid, to execute the main processes;
- Availability: all services are available on the Internet and the user only need a browser to make searches and edit data;
- Automatic Updates: each time that a user accesses a GP Web interface, he/she sees the latest version of the system interface and any updates to the internal software are transparent to the user;
- Updated data: whenever an user updates some information, this information is automatically updated in all systems and available to the other users;
- Security: there are levels of user authentication that determine the kind of information each user can read and/or edit.

Practical Results

Gene Projects was originally developed to manage data from the *Crinipellis perniciosa* fungus genome (www.lge.ibi.unicamp.br/vassoura) and it has been used in several projects, such as: the Coffee genome project (www.lge.ibi.unicamp.br/cafe), Citrus (http://biotecnologia.centrodecitricultura.br) and Eucalyptus genome project (www.lge.ibi.unicamp.br/eucalyptus). More projects and details can be obtained at the main site, www.lge.ibi.unicamp.br.

The system was developed to allow users, typically biologists, to make specific searches from a set of sequenced reads. These reads can be filtered through several different criteria, such as quality of sequences and percentage of vectors (contaminants), as well as other criteria. The filtered reads can be processed by several bioinformatics programs and can be compared against most biological databases. All of the results are organized and stored inside the individual projects.

Figure 2 shows some Gene Projects Web interfaces. Figure 2A shows the mechanism that users can select to add reads to projects. The search mechanism is composed of:

- Reads name search: Used when analyzing specific reads or all reads from one plate (for example, to see the quality of the sequenced plate);
- Keyword search: Used when looking for a particular theme/topic (for example, to find reads related with the product of a given gene);
- Blast search: Used when searching for similarity of sequences (for example, to find reads with similarity to the sequence of a given gene);
- Pattern search: Used when searching by domain or repeat regions (for example, to find microsatellites, protein domains or transcription factors).

Assembly

The reads of a given project can be clustered and assembled. CAP3 (Huang and Madan, 1999) is used for EST projects and phrap for Shotgun projects. The results available from this assembly, shown in Figure 2B, are:

- Assembled fasta sequence;
- Assembled visualization;
- Blast results of the assembled sequences against the gene banks NR or NT;
- Search for ORFs inside the sequences;
- Assembled sequence reverse complement;
- Reads that belong to the cluster.

The cluster of interest can be submitted to the automatic annotation process, in which the cluster is compared with several databases like GO (The Gene Ontology Consortium, 2000) and NR (http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.shtml#databases) and information available in the interface is shown in Figure 2C (annotation interface).

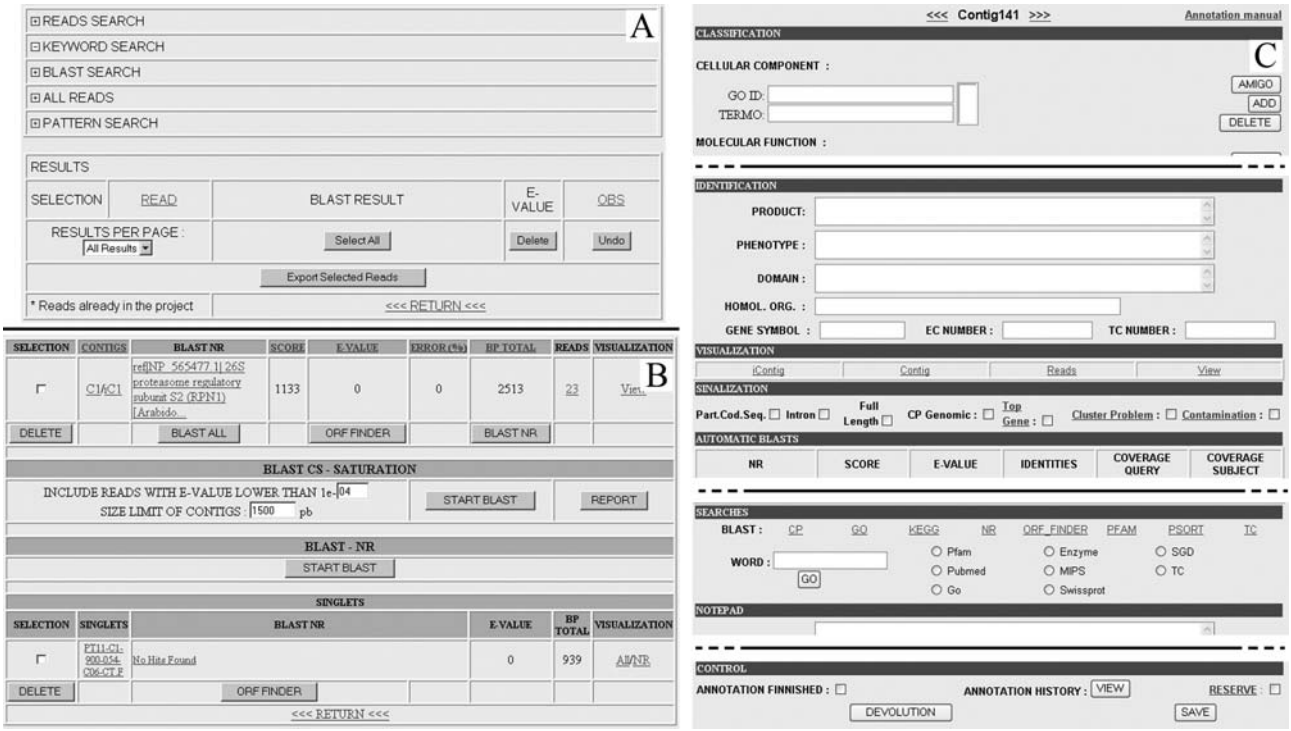

Figure 2 - Some Gene Projects Web interfaces: (A) Reads management - find and import reads; (B) Project assembly - view assembly results and run other analyses, such as ORF Finder or saturation blast; (C) Annotation interface - tools for identification and classification of clusters.

Saturation blast

In the typical data mining process, it is common to have keywords searches and searches based on Blast results. These are useful searches, but they cannot find every related read. To improve the quality and increase the number of reads of a given cluster (increasing the size of the cluster) it is necessary to add the right reads to the project. To facilitate this process, we developed a Blast saturation system. The program uses the original cluster sequence and performs successive comparisons with the data bank reads. It selects the reads with the greatest similarity according to the E-value and performs a new assembly of all reads. Selection of new reads continues until the contig obtained in the assembly reaches the desired minimum size (var2) or there are no more reads fulfilling the request (var1) (Figure 3).

This tool allows that, regardless of the bioinformatics knowledge of the end user, he/she can work with the best assembled contig (following some criteria) using every sequenced available read.

Annotation

There are three classes of users that interact with GP manual annotation interface during a genome project. The first is the *annotator*, which is the person that fills in information about the clusters. The *selector* chooses interesting clusters from his annotator group and reviews the annotation. The last, the *curator*, is a special kind of user that has the ability to review information provided by all annotators from all groups. Each class has tools to register and manage logins and passwords. In the case of a personal project, the owner can select and change his/her user class.

The annotation interface is comprised of eight main sections that have many features to assist the annotators. The sections and some of their features that they provided the user are:

• Classification: here the annotator can view/edit GO terms for the cluster. There is also a direct link from the selected term to the Amigo DAG Viewer (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi). A second classification system, defined by the coordinators of the project can also be inserted in this section;

- Identification: in this section, the annotator enters information about the product, phenotype, domain, homologous organism, gene symbol, Enzyme Commission number and Transporter Classification number;
- Visualization: the annotator can view the sequence of the cluster and its reverse complement, the reads that constitute the cluster, the assembled cluster and its translated sequence in all six frames.
- Flags: here the annotator can flag a particular cluster. There are flags to indicate:
  - Whether the cluster contains the complete coding sequence of a known gene;
  - Whether the clone(s) of a cluster contains the complete coding sequence of a known gene;

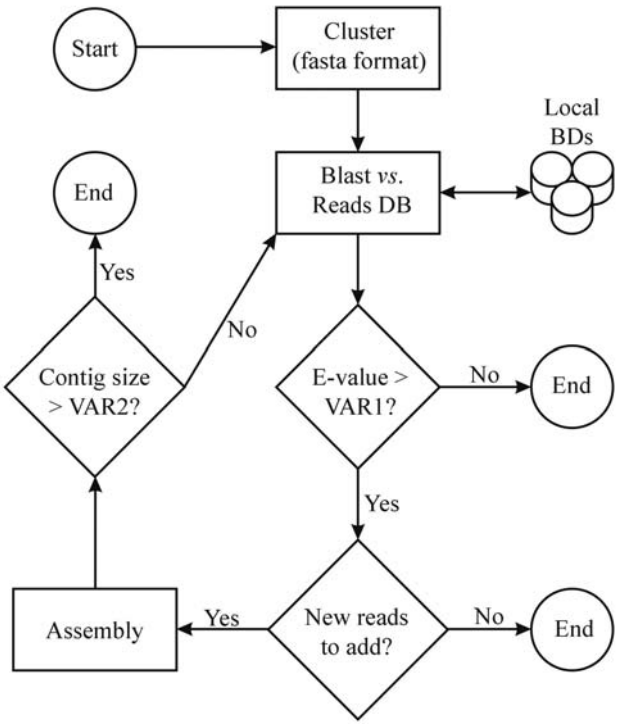

Figure 3 - Flowchart of the Saturation Blast algorithm. Initially, one cluster (contig or singlet) is selected. Using this cluster as a template and with the program BLAST, similar reads are extracted from the database. These reads are added to the thematic project and the program generates a new assembly. This process continues until the minimum defined size is reached.

  - Whether the cluster contains an intron sequence;
  - Whether the cluster has a central role in the specific project;
  - Whether the cluster has assembly problems such as a frame shift or a significant repeated region, for instance;
  - Whether the cluster is a contaminant.
- Pre-processed Blast alignments: here the annotator can view summary data about the alignments to some databases. Optionally, the annotator can visually inspect the alignments by clicking on a link. The list of databases is fully customizable by the project's bioinformatics team in accordance with the specific requirements for the project.
- Easy searches: in this section there are links to a set of web Blast interfaces. These Blast searches, unlike the former ones, are dynamically processed. The web Blast interfaces are loaded with custom parameters and the queried sequence input field is already filled in. There is also an interface to keyword searches in some biological databases. The bioinformatics team can customize the list of sites available for both keyword searches and Blast searches.
- Notepad: in the notepad section there are two input text fields. In the first, the annotator can enter personal notes and relevant information about the cluster that was not entered previously. Any annotator can edit the second, with the guest notepad, even if he/she is not the primary annotator of the cluster.
- Control: In this region, the annotator can designate a finished annotation, save the updates and view the annotation history of the cluster. The history contains all edited operations made to the cluster, including user changes, date and time. The history has such a level of detail that it can be used to reconstruct the annotation database. The annotator can also reserve the cluster for functional analysis. The system offers an additional function to the selector: he can return a cluster that he has selected from his group. For the curator, he may indicate whether the cluster annotation was reviewed or not with the use of a check box.

Microarray analyses

There is another kind of analysis that can be done with Gene Projects, the microarray analyses, before and/or after microarray experiments. The microarray experiment is divided into two steps: pre-processing and pos-processing. In the pre-processing stage it is necessary to map the clones from 96 well plates to 384 well plates and finally, to a spot on the microarray. The pre-processing stage outlines the microarray experiment so that the experimental criteria for retrieving the clones from the 96 well plates and transferring the 384 well plates follows a rule that allows for multi-channel pipetting and robotic spotting of the microarrays. The robotic "spotting" step (www.flexys.com) utilizes a file that is generated in the pre-processing stage.

In the post-processing stage it is necessary to process the color intensity result of each spot, following specific criteria: statistical significance normalization (Yang *et al.*, 2002), validation of the color intensity and gene clustering by expression class (Aittokallio *et al.*, 2003).

Gene Projects executes the following steps in the pre-processing stage:

- 96 well Plates: the arrangement is automatic, the sorting criteria is read by name in alphabetic order. The user can manually rearrangement the order.
- 384 well Plates: this stage is automatic, following the logic given by the mechanic process.
- Microarray spotting: the Gene Projects is able to read the robotic result files, and automatically generate rearrangements in the respective spotting grids. This stage is automatic only when using Flexys - Genomic Solutions robots.

All these stages are registered in the GP interface and there are links to BlastX results. This integration is fundamental to a correct interpretation of experimental results.

Table 2 - Some surveyed systems and their characteristics.

Characteristics	ESTWeb[1]	ESTAP[2]	GENOTRACE[3]	ESTAnnotator[4]	PartiGene[5]	ESTIMA[6]	EST-PAGE[7]	Gene Projects
i. Web interface	X	X		X	X	X	X	X
ii. Managing EST genome projects	X	X		X	X	X	X	X
iii. Managing Shotgun genome projects			X					X
iv. Processing of chromatograms files	X	X	X	X	X		X	X
v. Trimming and filtering sequences	X	X		X	X		X	X
vi. Reads annotation	X	X		X	X		X	X
vii. Clustering and assembly		X	X	X	X			X
viii. Clusters annotation		X	X	X	X			X
ix. Search tools in annotations data	X	X	X		X	X	X	X
x. Microarray analysis								X
xi. Thematic projects environment						X		X
xii. Working in genome projects in progress (sequencing stage)	X	X	X					X
xiii. Data access control (username and password)						X		X

Table references: [1] Paquola *et al.*, 2003; [2]Mao *et al.*, 2003; [3]Berezikov *et al.*, 2002; [4]Hotz-Wagenblatt *et al.*, 2003; [5]Parkinson *et al.*, 2004; [6]Kumar *et al.*, 2004; [7]Matukumalli *et al.*, 2004.

Correlated Works

It is possible to find similar programs to Gene Projects in the literature, particularly after 2002. But, these programs are designed for specific purposes, for example, only microarray analyses or annotation. Table 2 shows a comparison of the functions of GP and other programs. As shown in Table 2, there are many programs that have the capacity to manage genome projects, but typically only on a large scale, which limits their use for small projects especially during the sequencing stage.

Conclusions

Currently, every information system is concerned with user-connection issues, such as usability and interfacing. Systems that deal with genomic projects, which work with large volumes of data, have additional problems with data presentation to users, graphic visualizations, data summaries and connection between heterogeneous data sources. Another important characteristic of genome projects is that the data generation through sequencing is time consuming. Therefore, the systems must be able to utilize incomplete data and build upon this until the end of the sequencing process (to provide what we call “ongoing annotation”).

As shown, Gene Projects is a system, available via the Web, which addresses all of these concerns. It has been used in real genome projects and has produced satisfactory results by integrating, in a transparent way, heterogeneous data and extending the functions of other systems.

Future Work

We will continue to improve GP functions in the future, with particular attention to microarray post processing, such as normalization, statistic significance tests and clustering methods for the identification of expression patterns.

Acknowledgments

The work described in this paper was partially funded by FAPESP, CNPq and CAPES. We acknowledge the contributions of researchers that were or are end users of our system and that have helped us improve the functions, interfaces and the general quality of Gene Projects.

References

Aittokallio T, Kurki M, Nevalainen O, Nikula T, West A and Lahesmaa R (2003) Computational strategies for analyzing data in gene expression microarray experiments. J Bioinform Comput Biol 1:541-586.

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.

Berezikov E, Plasterk RH and Cuppen E (2002) GENOTRACE: cDNA-based local GENOme assembly from TRACE archives. Bioinformatics 18:1396-1397.

Ewing B and Green P (1998) Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186-194.

Hotz-Wagenblatt A, Hankeln T, Ernst P, Glatting KH, Schmidt ER and Suhai S (2003) ESTAnnotator: A tool for high

throughput EST annotation. Nucleic Acids Res 31:3716-3719.

Huang X and Madan A (1999) CAP3: A DNA Sequence Assembly Program. Genome Res 9:868-877.

Kumar CG, LeDuc R, Gong G, Roinishivili L, Lewin HA and Liu L (2004) ESTIMA, a tool for EST management in a multi-project environment. BMC Bioinformatics 5:1-10.

Mao C, Cushman JC, May GD and Weller JW (2003) ESTAP-an automated system for the analysis of EST data. Bioinformatics 19:1720-1722.

Matukumalli LK, Grefenstette JJ, Sonstegard TS and Van Tassell CP (2004) EST-PAGE-managing and analyzing EST data. Bioinformatics 20:286-288.

Paquola AC, Nishyiama MY Jr, Reis EM, da Silva AM and Verjovski-Almeida S (2003) ESTWeb: Bioinformatics services for EST sequencing projects. Bioinformatics 19:1587-1588.

Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A and Blaxter M (2004) PartiGene-constructing partial genomes. Bioinformatics 20:1398-1404.

The Gene Ontology Consortium (2000) Gene ontology: Tool for the unification of biology. Nat Genet 25:25-29.

Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO and Hunkapiller M (1998) Shotgun sequencing of the human genome. Science 280:1540-1542.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP (2002) Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30:15.

Internet Resources

AmiGO is the official tool for searching and browsing the Gene Ontology database, http://amigo.geneontology.org/cgi-bin/amigo/go.cgi (March 5, 2007).

Apache software foundation, www.apache.org (March 5, 2007).

Flexys Genomic Solutions, www.flexys.com (March 5, 2007).

Non-redundant Genbank database, http://www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml#databases (March 5, 2007).

Perl programming language, http://www.perl.org (March 5, 2007).

Perl5 library for writing World Wide Web CGI scripts, http://search.cpan.org/dist/CGI.pm (March 5, 2007).

Standard database interface module for Perl, http://dbi.perl.org (March 5, 2007).

The Citrus genome project, http://biotecnologia.centrodecitricultura.br (March 5, 2007).

The Coffee genome project, www.lge.ibi.unicamp.br/cafe (March 5, 2007).

The Eucalyptus genome project, www.lge.ibi.unicamp.br/eucalyptus (March 5, 2007).

The *Moniliophthora perniciosa* genome project, www.lge.ibi.unicamp.br/vassoura (March 5, 2007).

The open source database, www.mysql.com (March 5, 2007).

Associate Editor: Marco Aurélio Takita