Bioinformatics of the Sugarcane EST Project

Marília D.V. Braga Zanoni Dias Tzy Li Lin

João Meidanis José A.A. Quitzau Felipe R. da Silva

Guilherme P. Telles

Bioinformatics Lab – IC-UNICAMP

CP 6176 CEP 13083-970 – Campinas-SP-Brazil

Abstract

The Sugarcane EST Project – SUCEST – produced 291,904 ESTs in a consortium-based fashion. In this project, the Bioinformatics Lab created a web site that served as a "meeting point" of a network of 76 sequencing and data mining laboratories, receiving, processing, analyzing, and providing services to help explore the sequence data. In this paper we describe the information pathway implemented by the Bioinformatics Lab to support this project, including a brief explanation of the clustering procedure, which resulted in 43,141 clusters.

1 Introduction

The application of expressed sequence tag – EST – technology has proven to be an effective tool for gene discovery [1], mapping [2], and generation of gene expression profiles [3].

EST projects are usually conducted by a single laboratory, which makes the libraries, isolates clones, sequences them, analyzes the data and submits it to GenBank. However, the Sugarcane EST Project – SUCEST – involved 24 sequencing labs, a bioinformatics lab, a coordinating lab, a group for international relationship, and 50 data mining groups scattered all over Brazil working in cooperation. A newborn Brazilian bioinformatics group associated to the project lately, intending to learn from previous experience. Starting early in 1999, in 15 months it generated 291,904 sequences from 260,352 clones of 37 different libraries.

Genomic research in Brazil has been consortium-based since its first project, the complete genome of the phytopathogenic bacterium *Xylella fastidiosa* [4], conducted by the ONSA network. A consortium-based genome project provides a larger number of researchers, technicians and sequencing machines, but demands a much more organized data flow. In SUCEST, the Bioinformatics Lab – LBI – was responsible for receiving data from a network of wet labs, assessing their quality, storing, clustering, and providing many other services on the data. These tasks are described in some detail here, and quantitative figures from the project are given.

2 Computational Systems

For a short time in the beginning of the project, the web site was hosted by a 128 MB memory personal computer running the Linux operating system (distribution Red Hat 6.2), and presently the site resides on a Compaq AlphaServer ES40 with two Alpha 667 MHz processors, 8 GB of RAM, and 384 GB in disks, running OSF-1 version 4.0G. However, the bulk of the project was executed on a Compaq AlphaServer DS20 with two Alpha 500 MHz processors, 4 GB of RAM, and 144 GB in disks, running OSF-1 version 4.0F. This is the system in which most of the tools were developed and we will concentrate on it for the rest of the paper.

The Web engine server was **Apache** [5] version 1.3.9. CGI's and programs were written in Perl [6] version 5.005, and PHP [7] version 3.0.12. The database management system was **MySQL** [8] version 3.22.26a.

The input data consisted of data received through web forms, including ABI 377 chromatograms, and HTML formatted data mining reports.

The base calling and sequence extraction programs used were **phred** [9] version 0.980904.e and **phd2fasta** [9] version 0.990622.d. The comparison programs used are **cross-match** [9] version 0.990319 and **blastall** [10, 11] version 2.0.10. Assembly programs were **phrap** [9] version 0.990319 and **CAP3** [12]. Off-the-shelf scripts were used to provide keyword search in data mining reports [13], database administration and other minor tasks. The software used is either free for academic purposes, or was developed by our team.

3 Methods and Results

From a computational viewpoint, SUCEST may be seen as a big data repository and as a provider of Internet based services for a community of different types of users. Figure 1 shows the major relationships among users, services, data and programs in the project.

Users are members of sequencing labs, which submit chromatograms from clone libraries, members of data mining labs, which perform searches on the project database and publicize their results in data mining reports, and members of the project coordination, which monitor the project status and control plate distribution and validation. Users interact with data through services that add, retrieve, and update the data repositories.

Data include sugarcane ESTs, information about the participants of the project, data mining reports, control data, summaries, and output from programs that perform automated searches in databases, clustering, and categorization of clusters.

Below is a description of what users, data, services and programs in SUCEST are, and how they interact.

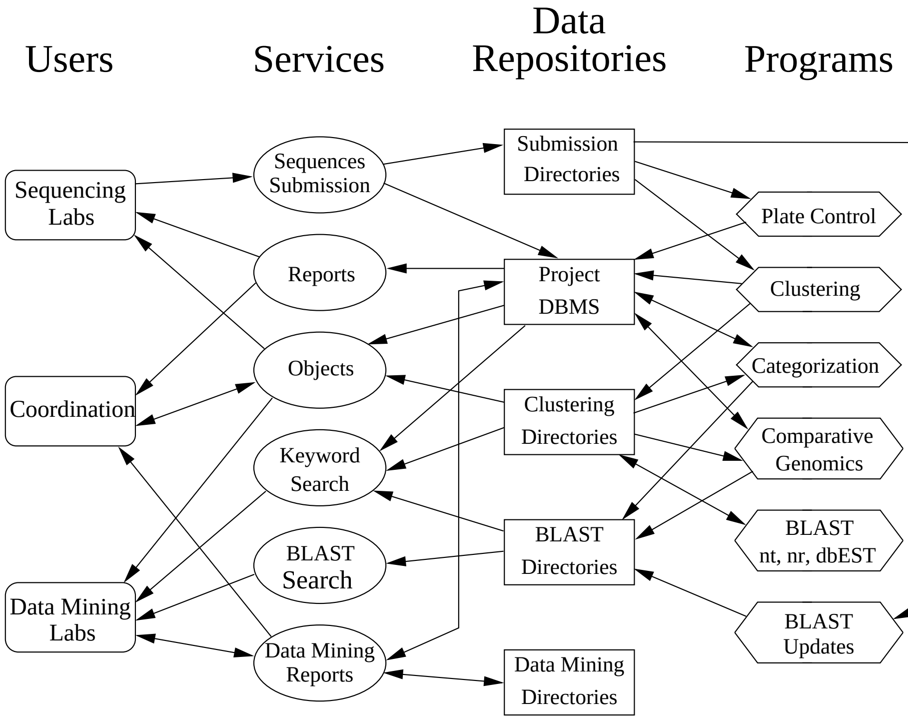

Figure 1: Major relations among users, services, data, and programs in SUCEST. Arrows indicate the flow of information.

3.1 Objects

SUCEST data is stored in two different kinds of repositories: operating system directories and a relational database. The directories hold biological sequence files, results from BLAST and cross-match searches in biological databases, and data mining reports. Biological sequence files include chromatograms, files in fasta format, quality files, and files generated by clustering, categorization, and comparative genomics procedures. The project uses only one relational database, with several interconnected tables that store other biological and management data, such as information on libraries, sequencing plates, labs, and their members. The database also points to data in directories. The major entities (objects) in our database are described below, where we also introduce quantitative figures and details from the pipeline.

Laboratories There are 78 laboratories in SUCEST that belong to one or more of five groups: "DNA Coordination", "Bioinformatics", "Data Mining", "Sequencing", and "International Cooperation". The services and data that a member of a lab can access depend on the group to which the lab belongs. One of the members in a lab is its head, and receives notifications of some of the activities performed by the lab members. A two-letter code identifies each lab.

Members A SUCEST member is a person that belongs to at least one lab. Several members belong to a sequencing lab and also to a data-mining lab. Data about members include name, the labs they belong to, e-mail address, phone numbers, and login and password to grant access to the authorized services. SUCEST had 256 members on March 25, 2001.

Libraries An EST results from the sequencing reaction of one of the ends of a clone. In turn, a clone is a cDNA that was prepared from a particular tissue of the plant in specific conditions. ESTs in SUCEST came from 37 different libraries, prepared from different sugarcane tissues under different conditions [14]. Library name, description, and vector employed are recorded for the libraries. A code composed of two letters and one number was assigned to each library. The letters indicate the tissue from where the library has originated and a consecutive number was assigned for every new library deriving from the same tissue. For example, "LR1" coded for leaf roll with long inserts, "LR2" coded for leaf roll with small inserts. Libraries have three possible status: "TEST", for validating libraries, "START", for libraries released for sequencing, and "STOP", when the coordination decides it is not worth to keep sequencing a distributed library. Of the 37 libraries prepared for the project, 32 were started and 5 were abandoned after the TEST phase. Those not started either produced too much redundancy or very small reads.

Plates Clones are organized in 96-will plates in SUCEST. Every plate holds clones from the same library in a 8×12 grid. Sequencing is done for a whole plate, and this data is sent to LBI for processing and storage. Data for a plate include the library which it came from, and the lab that is authorized to send data from this plate. A plate has a three-digit identification tag, except for control plates (see below for details on control plates), that have the letter "C" and two digits. The SUCEST database records data from 2,771 different plates.

Reads A read is the same as an EST. Reads are extracted (using the program **phred**) from chromatograms submitted by the sequencing labs, and screened for vectors (using the program **cross_match**). All reads are stored in directories as chromatogram files, and also as a pair of text files holding the sequence and its quality. The following attributes are stored in the database for every read:

- the plate and the position in the plate where it came from,
- information about the submission process, such as the date and time it happened,
- the number of vector and non-vector bases with phred quality equal or higher to 20,
- the number of vector and non-vector bases with phred quality lower than 20,
- the start and ending positions for every vector sequence identified in the read, and
- whether it has trace data.

Every read has a name that is the concatenation of its lab name, library name, plate name, position in plate, and direction. The string SCACAD1001A01.g, for example, is the name for the 5' read of the clone in well A01 of plate 001 of library AD1, and sequenced by lab AC. (Note: 5' reads are identified by the .g suffix, while 3' reads use the .b suffix.) Every position in the plate is identified by its row and column names, with row names ranging from A to H, and column names from 01 to 12.

Preparation Sheet Before a lab can sequence and submit a plate, it must submit information about the plate preparation process. There are records in the database for every position where the bacteria did not grow and for the positions from which it was not possible to obtain DNA. Every position marked with a problem will correspond to a sequence without information relevant to the project.

Control Plates A scheme for plate validation is used in the project, as follows. For every set of 12 plates a control plate is built with the 8th column of each. Control plates are sequenced, and information about the reads that match in control and in controlled plates is stored in the database. The idea is to identify tracking and naming errors in the sequencing process. Plates with problems may be fixed and resubmitted by the lab that produced them.

Clusters SUCEST reads are grouped by a clustering procedure that is described below. This clustering procedure creates sets of aligned reads that we call *clusters*. In our database we store the reads that are part of each cluster. Therefore, in addition to being a set of reads, a cluster has an alignment and a consensus sequence. Alignments, consensus sequences, and quality files are stored in cluster directories. A cluster also has a name, which is equal to the name of oldest read in it.

3.2 Services and Programs

Data come in and out of SUCEST's data repository through a set of services. These services are provided by web pages hosted at LBI. Data is also generated within the LBI by programs that are executed either automatically or manually. Brief descriptions on these services and programs are presented below, starting with data retrieval, giving a general overview on how the SUCEST web site is organized and how it works.

Data Retrieval Data is retrieved from SUCEST in units called "objects". SUCEST objects are the same as data entities described above: Library, Reads, Plates, Control Plate, Member, Laboratory, and Cluster. Each object has its own, individual web page, containing information about itself and links to any other object, service, and report directly related to it. Starting from a lab or library object it is possible to reach the web page of any other object.

Some objects point to pages that include data extracted from the projects' directory structure. For instance, one can visualize reads and base qualities in many versions: immediately after submission (but before screening), after screen (but before trimming), and after trimming (see Clustering below). For clusters, it is possible to visualize the reads in a cluster, and their alignments, including the consensus.

An *object search* service was created to allow direct access to any given object. Given the code and the type of object the service delivers its page. For object type Member, it is possible to search by name, email, department, city, and institution.

Besides objects, some reports that summarize data are also available for the project:

- a *summary of submitted reads* gives totals per lab or per library of submitted, payable and clusterizable reads, and
- a *summary of control plates* gives totals of accepted and rejected plates by control.

SQL skilled users may take advantage of a service that allows generic queries to the database. Queries can be typed in a web form and the results are returned in tabular fashion. Entity-relationship diagrams and table descriptions for our databases are available to help users in this task.

Sequence submission Sequence submissions are done by sequencing laboratories only. The process requires accessing the project's web site using a valid login/password pair, and uploading a set of 96 chromatograms – exactly one plate. When an upload finishes, certain pre-requisites are verified: all chromatograms must belong to the same plate, the lab that is trying to submit a plate must be the one authorized to do so, the preparation sheet for that plate must have already been submitted, and the reads must be in accordance with the naming conventions.

If the pre-requisites are satisfied, programs **phred** and **phd2fasta** are used to extract the sequences and their qualities in fasta format from chromatograms. Program **cross-match** is used to mask vector sequences in the reads. These steps take a few minutes, and this time was essentially constant during the entire project, because the analysis done upon submission does not depend on the other reads present in the repositories.

After the submission analysis, a report that summarizes the process and the sequences received is presented to the submitter. At this point the submitter is asked to confirm the submittion or not. If the submission is confirmed, the database is updated and, if there is an older version of the plate, it is replaced. Directories are updated as well. If the submission is not confirmed (e.g., if the submitter is not happy with the quality assessment), the submission is discarded.

The path followed by a read in LBI's pipeline, starting from the submission, is shown in Figure 2. The submission procedure corresponds to the part in the figure starting at "Zip file" and extending through the top line. Other steps in the diagram are performed by programs described in the sequel.

Clustering Clustering of ESTs is important to reduce the amount of sequence data that miners have to look at, and to organize the reads in a less redundant set. In SUCEST, clustering had one additional motivation: the

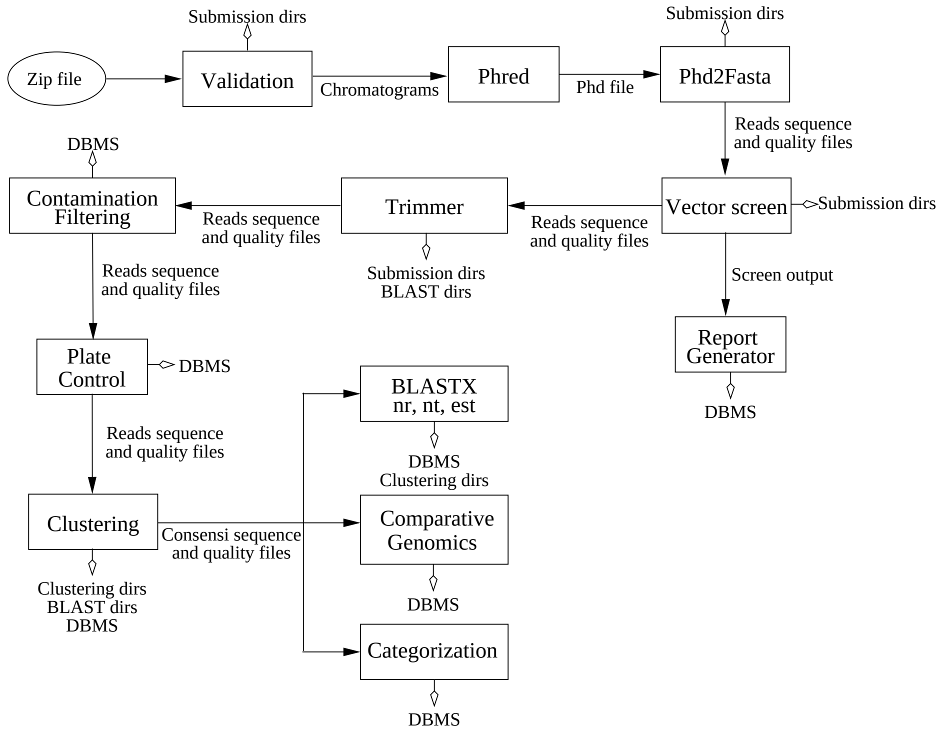

Figure 2: The operations over a read in the SUCEST pipeline. Black arrows linking boxes indicate data that flow from one stage to the next, and white-headed arrows going out of boxes indicate data repository updates.

need for redundancy estimation in the libraries.

Two pivotal decisions were made early on. One is that each cluster should reflect a transcript, rather than a gene, an allele, or other biological entity. The second was that a cluster consists not only of a set of reads, but also of an alignment of these reads.

In this context, our first scheme was to group similar transcripts and to produce consensus sequences using the assembly program **phrap**. This strategy was sufficient in the early stages of the project, but as data accumulated a series of problems forced us to change the scheme, as described below.

To minimize artifacts, reads were trimmed before clustering. This trimming procedure started with vector masking using program **cross_match**. Then some of the poly-A, vector and adaptor regions were removed. A quality trimmer was also applied, removing bases from the ends of the sequence one by one until there were at least 12 bases with phred quality above 15 in a window of 20 bases at the end. Reads were also checked for contamination against *Xylella fastidiosa*, *Xanthomonas citri*, *Escherichia coli* and other potential contaminators in the labs that produced the libraries. BLAST was used to compare the reads and the potential contaminators, and if a match of at least 100 bases and more than 90% of identity occured, the read was marked as a probable contamination. However, marked reads were kept in clustering and subsequent analyses, to allow data miners to decide by themselves whether a specific read was contaminated or not.

Trimmed reads were assembled using phrap with quality and with the following arguments, which made it very stringent:

```
-penalty -15 -bandwidth 14 -minscore 100 -shatter_greedy.
```

Every contig and singlet produced by phrap was taken as a cluster. As new plates came in, our clusters were automatically updated by a program. This daemon updated the database, directories and BLAST results for every cluster that changed. Initially clustering was performed every day, but as the set of sequences grew, the updates became sparser, running once a week. In the final phases of the project, clustering would typically occupy an entire CPU for about 20 hours.

The last assembly done with phrap included 261,609 trimmed reads, and produced 81,223 clusters. However, remarks by several members of the project that the total number of clusters in the database was unreasonably large, that many clusters were malformed, and that some clusters looked like they could be combined, led to changes, described in detail by Telles and da Silva [15].

The new scheme was based on careful testing and evaluation, and consisted of a more elaborate trimming procedure, and of the use of the CAP3 [12] assembler, the same tool used to produce TIGR's gene indices [16]. Trimming in this new procedure included ribossomic RNA removal, a comprehensive removal of poly-A, poly-T, vector and adaptors regions, and improved low-quality-end trimming. CAP3 was fed with 237,954 reads and quality, and produced 43,141 clusters.

Both versions of the clustering are presently accessible through the project web site, with data from both methods available for most services.

Keyword Search Keyword search is a service that allows users to search for a set of keywords in the header lines of every sequence in NCBI's nr, nt and dbEST databases [10] that hits any cluster in SUCEST. To perform a query, the user gives a database name (nr, nt or dbEST), a logical expression of keywords (that may include “or” and “and” connectors), and the maximum e-value required (which is an optional parameter and defaults to **1e-5** $= 10^{-5}$). The service then returns the clusters that have a hit with the expected e-value or better and whose subject heading contain words satisfying the logical expression. The resulting list of clusters is ordered by e-value.

A program was created for keeping BLAST results against nr, nt and dbEST up to date for all SUCEST clusters. A BLAST result against a certain database is considered *outdated* for a SUCEST cluster if the cluster was newer than the result or if the cluster or the database were modified after the last BLAST run. When the program finds outdated BLAST results, it builds a BLAST queue giving priority to older clusters. If the databases are available in different computers, the system is able to improve the processing time, running several BLASTs in parallel, one on each remote server, and taking about 2 or 3 days. Otherwise, it alternates BLAST runs for each blast database in the same machine, taking considerably longer.

Subclustering This service is used to evaluate statistics about subsets of clusters of the clustering, including read frequency by cluster size, total of reads, total of clusters, redundancy, and novelty.

To select the subset of clusters, the user has to indicate the reads that belong to the clusters. Any cluster that contains a read in the selection is included in the evaluation. To indicate the reads, elements from their names should be selected: lab, library, plate, position and direction. More than one element may be selected. Selecting a particular lab, for example, will generate the statistics for the clusters that have at least one read sequenced by that lab.

BLAST Search A BLAST service allows searches against SUCEST reads, reads in their trimmed version, and cluster consensi. These databases were updated automatically in a daily basis, to incorporate new reads and consensi.

Data Mining Report Data mining groups submit HTML formatted reports to the SUCEST site, and update them periodically. Users may access reports through an index page that provides access to the reports of every data mining group. A search by keywords is also available. When a report archive is uploaded, a service takes care of unpacking the files, updating the index page and the searching index.

Information about reports is also kept in SUCEST's database, including the project's name, summary, members, submission date, and submitter name.

Categorization SUCEST clusters were categorized, in an attempt to determine their function, and to aggregate information. Thirty categories where built by a team of SUCEST members, and 32,438 example proteins were assigned to the categories. Categorization was done in two steps:

1. Automatic: a BLAST database was constructed, containing example proteins for the categories. A BLAST search was performed against this database having SUCEST clusters as input. Any cluster was considered to be in a category A, if it hit with some example protein in A, with e-value better or equal to 10^{-10}, and covered 70% or more of the example. A cluster could be in many different categories. This method categorized 36% of the 43,141 clusters.

2. Manual: A web service was built to allow manual annotation, when automatic annotation produced ambiguous categorization or produced no categorization at all. Based on BLAST results against nr, SUCEST members were able to establish a direct relation between a cluster and a category. Manual annotation increased significantly the number of categorized clusters and, as of March 20^{th}, 2001, 60.5% of the clusters were categorized.

Comparative Genomics In an effort to obtain knowledge about sugarcane and its relation to other species, cluster consensi from SUCEST were compared against other organisms. The first organism selected for comparison was the model plant *Arabdopsis thaliana*. Every cluster consensus was BLASTed against *A. thaliana*'s chromossomes, proteins, and ESTs. Clusters that produced no hit against *A. thaliana*, were also BLASTed against ESTs from *Lycopersicon esculentum*, *Glycine max*, *Lotus japonicus*, *Hordeum vul-*

gare, *Oryza sativa*, *Sorghum bicolor*, *Zea mays*, *Triticum aestivum*, and *Medicago truncatula*. Results from these searches were inserted in our database, allowing several queries to determine the distribution of these hits per library, per cluster, or grouped according to many other criteria.

Management These services provide a way for the DNA coordination to input management information into the SUCEST database. This information is used mainly by services that perform checking and summarizing operations. Using the library management services, the DNA Coordination modifies in the system the status of any library and assigns plates to sequencing laboratories. Manual plate approval is also possible for the DNA coordination, via a service that displays control and controlled plates showing which cells match in control and controlled plates.

4 Discussion

A key aspect of the project was the deep interaction between the biological labs and the bioinformatics lab. Through discussion lists or over the phone, the users would give suggestions of new services, and point out quickly problems with the services (broken links, bugs, etc.) This daily, intensive interaction was undoubtedly one of the main reasons for the success of the project.

Clustering started early, and had a dramatic impact along the project.

Reclustering in a regular basis demanded designing and implementing programs to update databases, and BLAST results against nr, nt, and dbEST. It also demanded a lot of CPU time. When another clustering scheme was adopted, the web site had to change to accommodate both versions simultaneously, and to show relations among clusters in different versions. Both bioinformatics and data mining staff experienced some level of overhead to adapt for changes.

Among the lessons learned in the process we point out two important ones:

Avoid changing systems: during this project we had to change the underlying computing system twice. The first time from a PC to a medium-sized server, and then to a bigger server. These changes caused many problems, for instance, programs that used to work in one system would not to work on the other, users have to get used to new URLs, and so on. The migration process proved time-consuming and error-prone. Our advice would be to set up a system that is big enough right from the start, and keep the project there for as long as possible. To minimize the impact of migration, it is important to devise the directory structure in a system-independent way, for instance, put your data in directories that won't collide with system directories, and install your programs in standard places, using execution path variables to assure they will work.

Another important piece of advice is to use software that combines many physical disks into one big volume, say, of a few hundred gigabytes. Most vendors provide such software for a small fee.

Keep reference sequences, not cluster lists: in this project, data accumulated at a fast rate and clustering was redone frequently. Some data mining groups had problems trying to keep up with the frequent updates because they maintened lists of relevant clusters. Each time the clustering was redone, some clusters would dissapear (merge into larger ones), or the read composition of a cluster would change, requiring a lot of manual labor. Our advice would be to use reference sequences from Genbank or other stable sequence database, which will then be used as queries to retrieve the cluster lists via BLAST. Proceeding in this way, the lists could be reconstructed quickly from the reference sequences, using automated methods.

There are many other programs, not presented here, that contributed to the functionality of SUCEST's web site. Some services and programs have already been disabled, for example, sequence submission and plate control, but others, such as keyword search, BLAST and report submission, are still being used by data mining labs, and will be used by international community when the data goes public. This will certainly transform the meeting point of the project's community into the meeting point of a wider group, setting new demands for services and data storage.

5 Acknowledgments

This work was supported by FAPESP, CNPq, and COOPERSUCAR.

6 Summary in Portuguese

O projeto SUCEST (Sugarcane EST Project) produziu 291.904 ESTs de cana-de-açúcar. Nesse projeto, o Laboratório de Bioinformática criou o web site que foi o "ponto de encontro" dos 76 laboratórios de sequenciamento e *data mining* que fizeram parte do consórcio para o projeto. O Laboratório de Bioinformática recebeu, processou, analizou e disponibilizou ferramentas para a exploração dos dados. Neste artigo os dados, serviços e programas implementados pelo LBI para o projeto são descritos, incluindo o procedimento de *clustering* que gerou 43.141 *clusters*.

References

[1] Adams, M., Kelley, J., Gocayne, J., et al. (1991) Complementary DNA sequencing: "expressed sequence tags" and the human genome project, *Science,* **252**, 1651–1656.

[2] Schuler, G. (1997) Pieces of the puzzle: Expressed sequence tags and the catalog of human genes, *Journal of Molecular Medicine,* **75**(10), 694–698.

[3] Boguski, M. and Schuler, G. (1995) ESTablishing a human transcript map, *Nature Genetics,* **10**, 369–371.

[4] Simpson, A., Reinach, F., Arruda, P., et al. (2000) The genome sequence of the plant pathogen *xylella fastidiosa*, *Nature,* **406**, 151–157.

[5] Apache homepage, www.apache.org.

[6] Perl homepage, www.cpan.org.

[7] PHP homepage, www.php.net.

[8] MySQL homepage, www.mysql.com.

[9] The Phred/Phrap/Consed homepage, www.phrap.org.

[10] The National Center for Biotechnology Information (NCBI) homepage, www.ncbi.nlm.nih.gov.

[11] Altschul, S., Madden, T., Schäffer, A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research,* **25**, 3389–3402.

[12] Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program, *Genome Research,* **9**, 868–877.

[13] Weil, B. and Baron, C. Drag 'n' Drop CGI: Enhance Your Web Site Without Programming chapter 12 Addison Wesley (1997).

[14] Vettore, A., daSilva, F., Kemper, E., and Arruda, P. (2001) The libraries that made SUCEST, the Sugarcane EST Project, *GMB,* **This issue**.

[15] Telles, G. and daSilva, F. (2001) Trimming and clustering sugarcane ESTs, *GMB,* **This issue**.

[16] Quackenbush, J., Liang, F., Holt, I., et al. (2000) The TIGR Gene Indices: reconstruction and representation of expressed gene sequences, *Nucleic Acids Research,* **28**(1), 141–145.