SisGen: A CORBA–Based Data Management Program for DNA Sequencing Projects

Georgios J. Pappas Jr.[1,2], Robson P. Miranda[1], Natália F. Martins[2], Roberto C. Togawa[2], and Marcos M.C. Costa[2]

[1] Biotechnology and Genomic Sciences program, Universidade Católica de Brasília
gpappas@bioinformatica.ucb.br
[2] EMBRAPA Recursos Genéticos e Biotecnologia, Brazil
mcosta@cenargen.embrapa.br

Abstract. Biological data deluge has challenged researchers over the last decade. Expressed sequence tag (EST) analyzes provide a rapid and economical means to identify candidate genes, gene expression profiles in different cell conditions, as well as functional annotation of putative gene products. Although EST analysis tools are publicly available there is still a lack of comprehensive data analysis and management programs. This work presents `SisGen`, an integrated software system capable of efficiently managing multi–user genomic projects. `SisGen` is a Java client–server application that uses CORBA as a middleware in a multi–layer architecture. The software integrates data management an annotation pipeline in a rich graphical visualization environment. The architectural design is presented and highlights the advantages in terms of portability, interconnectivity, modularity and user interface that can be achieved with this concept.

1 Introduction

The advent of the genomic era in the last century promoted an exponential increase in sequences in public databases, exceeding by far the capacity to perform experimental analyzes to pinpoint their roles in the cellular milieu. Concomitantly, the use of computers was perceived as pivotal to help transform sequence information into biological knowledge [1].

One type of genomic data that greatly contributed to sequence accumulation was the expressed sequence tags (ESTs; [2]). ESTs are short, single–pass sequences derived from random sequencing of cDNA library clones. Given that their generation is affordable, ESTs rapidly became a popular strategy for gene discovery in eukaryotes.

In the course of EST sequencing projects, data is continually generated by sequencing machines in the form of electropherograms, the starting material for the computational processing cascade aiming to infer biological function. Aside from numerous theoretical and algorithmic difficulties inherent in sequence annotation, a more fundamental problem of data management, processing and integration emerges. Many solutions have been developed over the years to provide

A. Bairoch, S. Cohen-Boulakia, and C. Froidevaux (Eds.): DILS 2008, LNBI 5109, pp. 116–123, 2008.

software systems dealing with the task of managing and annotating EST data, among them ESTWeb [3] and ESTExplorer [4], to cite a few. A critical evaluation several such programs was recently published [5]. A common theme is that they are web–based and coded in scripting languages like PERL or PHP. Despite progress in data organization and visualization, most of the existing systems still lack an integrated and robust approach required for EST data management.

Here we explore the concept of using enterprise–level software architectures to tackle the EST project management problem, which can be modeled as a distributed computing system. In this context, middleware technologies connect software components and provide an integration layer between heterogeneous systems. One of the earliest and most successful middleware architectures is CORBA (Common Object Request Broker Architecture), on top of which many home banking and electronic commerce systems were built.

Several groups recognized the importance of middleware technologies, such as CORBA, to enable the creation of elaborated applications integrating the many data formats and analytical tools present in the bioinformatics field [6]. Using this technology, we present a new software package, called `SisGen` that employs middleware concepts to cope with the data integration and administration problems faced in EST sequencing projects.

2 Methods

All development was geared to adopt free software. The programming language and the ORB were provided by the Java Platform Standard Edition v1.5. The persistence layer was provided by the hibernate framework in conjunction with the relational database server PostgreSQL 8.2. Production servers run the Linux operating system. The clients are platform independent and distributed via Java Web Start technology. Further details can be found at `http://bioinformatics.cenargen.embrapa.br/genoma`.

The data model was created with a project–centric vision, modeling aspects of raw sequence data being sent from different laboratories and providing detailed provenance and accounting. EST projects are hierarchically divided as having multiple cDNA libraries, each containing several plates, which in turn consist of individual reads. There is also provision for version control that permits read resubmission.

Starting from electropherograms, several third–party bioinformatics programs are applied in order to process raw data and provide functional annotation of the sequences. Custom–made wrappers and parsers were created to coordinate execution and integration of the ensuing results to the system. The pre–processing pipeline starts with the base calling program PHRED [7], cloning vector removal with cross_match (`http://www.phrap.org`), repeat masking with RepeatMasker (`http://www.repeatmasker.org`) and quality trimming with Lucy [8]. These steps are executed concurrently with sequence submission and provide real–time feedback to the submitter about the read/plate quality.

The next step in the pipeline is the functional annotation run on demand at server side. It starts with EST clustering using TGICL [9]. The resulting cluster consensi are subjected to several similarity searches using BLAST [10] against a series of databases defined during the project setup. Classification according to Gene Ontology (GO) and Enzyme [11] is inferred by mapping similarity search results against appropriate databases.

Sequence features that can be used as potential molecular markers for genetic studies are also annotated. Single nucleotide polymorphisms (SNPs) are predicted for each EST cluster using PolyBayes [12]. Simple sequence repeats (SSRs) are located in cluster consensi using the program `mreps` [13].

3 Results

3.1 Platform Design

The main objective was to provide an integrated software system capable of efficiently managing multi–user genomic projects, encapsulating several bioinformatics services for the analysis and manipulation of sequence data. Also, some key points were detrimental to the design process, such as (*i*) portability, (*ii*) efficiency and (*iii*) rich graphical user interface (GUI) for easy navigation.

The web based systems currently used by the vast majority of EST management systems often sacrifice design in detriment of simplicity and rapid development. Though successful most of the times, they may face shortcomings in terms of scalability, performance and flexibility. We adopted, instead, a multi–layer architecture using CORBA as middleware to service data between a java client program and the project database. Implementation was made in Java language and tried to adopt design patterns such as business object, data transfer, business delegate and session facade [14]. This promoted code reuse as well as clear separation between the data and the presentation layer. An overview of the system architecture is shown in Fig. 1, and the individual components are detailed below.

This is the piece of software used by the end user to interact with `SisGen`. Instead of a web browser, a custom made graphical user interface (GUI) was created using Java's swing library. A general overview of selected windows is shown in Fig. 2.

Departing from the common solution employing web browsers has some trade-offs though. First there is an increase in time spent designing the GUI. Also there is the versioning problem of how to distribute the client updates to maintain the compatibility with the server. This was effectively solved by using Java Web Start technology, which transparently ensures that the latest version of the application is deployed. However, the GUI programming is really an issue since the majority of the code in `SisGen` is devoted it. Notwithstanding, several benefits arise when using Java GUIs, which include better navigation and management of several windows. Also, there is a gain in flexibility since streamlined graphical components can be created, as seen in Fig. 2.

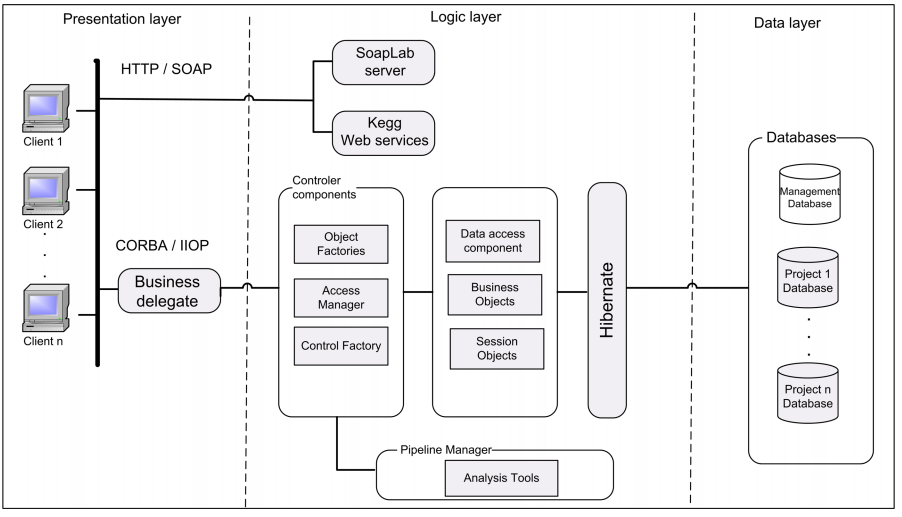

Fig. 1. Diagram of `SisGen` multi–layer architecture using CORBA as a middleware to service data between a client program and a project database. Distribution of presentation, logic and data layers elements.

Fig. 2. Screenshot of several features available on `SisGen` showing clustering population, sequence analysis, chromatogram viewer, blast output analysis tool, plate quality visualization and clone quality map.

The client can navigate through various levels of project and sequence information, querying and gathering data from the server through the coordination of a business delegate [14], that hides client–server remote communication details (Fig. 1) reducing coupling between the presentation and logic layers.

As long as the user queries and loads data, some computation can be carried out at the client side, relieving server communication burden. In order to expand `SisGen` client capabilities, a feature was added permitting remote execution of analytical bioinformatics tools using Soaplab [15]. Soaplab exposes command–line applications as web services using SOAP (Simple Object Access Protocol)

protocol. The modular design allows SisGen client to seamlessly interact with different middleware technologies, aside from its core functionalities mediated by the CORBA server.

Controls the main aspects of application functionality in response to client queries. It provides the unified interfaces to interact with the data layer. A object–relational mapping layer, driven by hibernate framework, hides the inner details of database operations encapsulating them in the object–oriented realm. This not only improves coding but also provides database back end independence. Additionally, one design strategy was to make provision to physical separation of the machines running the database server and the logic layer, improving security and distributing computation. Finally, the logic layer controls annotation pipeline execution, which is shielded from the end user.

For a specific SisGen project there are two main databases. One, the management database, is shared by all projects and contains project and user information details. The other database, on the same server machine, contains the sequence and annotation data itself.

3.2 System Features

Several core aspects of an EST management software are shared by SisGen and other web platforms, like EST data summary statistics, visualization of sequence and associated PHRED quality, project/user management or inspection of BLAST run reports, among others. Additionally, some noticeable features are peculiar to SisGen and are detailed next.

The main use case from a sequencing facility perspective is to transfer raw electropherograms to the central bioinformatics repository. The middleware architecture adopted by SisGen enables a data transfer solution that is efficient and flexible. Directories, individual or compressed files containing electropherograms can be transferred in batch to the server. Real–time feedback permits the monitoring of transfer progress and individual plate quality (Fig. 3a).

A generic sequence alignment and assembly viewer was created, capable of displaying several types of data present in an EST sequencing project. This viewer is integrated in SisGen but it is a completely independent and stand–alone component that can be used to visualize DNA/protein multiple alignments, BLAST results and sequence assembly files (ace format). The viewer is used to inspect the EST clusters and presents several measures of sequence conservation, like sequence logos and entropy plots (Fig. 3b).

As described in the Methods section the annotation pipeline predicts the location of polymorphic sequence sites that could be used as molecular markers. A bayesian inference procedure is used to predict the incidence of SNPs taking in consideration sequence coverage and quality [12]. An example of such SNP discovery process can be found in Fig. 3c.

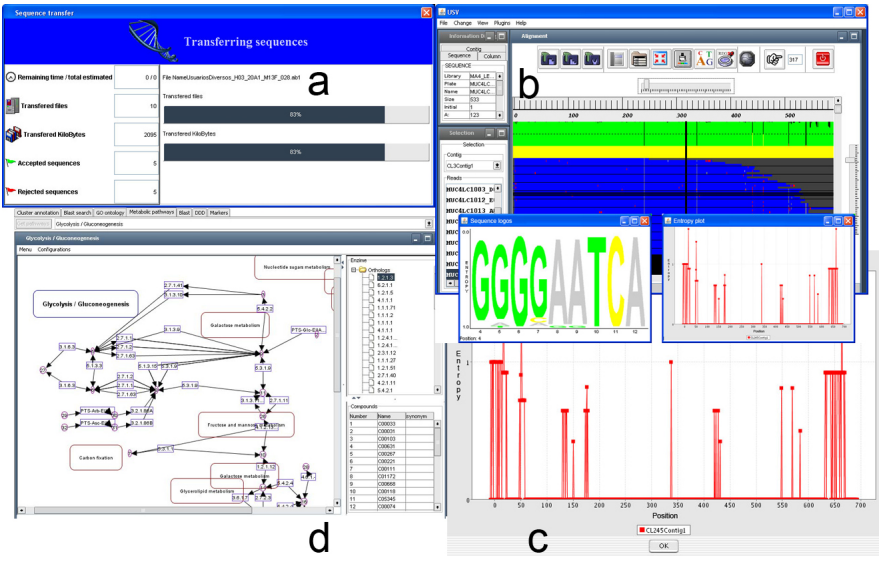

Fig. 3. Screenshots of `SisGen` user interface. a) Shows the file transfer interface; b) The Universal Sequence Viewer with cluster sequence alignment, detailing the sequence logo plug-in and the entropy plot; c) Display of Single Nucleotide Polymorphism (SNP) predictions; d) Metabolic pathway according to KEGG database.

Another type of molecular markers, the SSRs or microsatellites, are also annotated. Primer pairs flanking each microsatellite region are automatically generated. These PCR (polymerase chain reaction) pairs are suited to experimentally verify genetic diversity.

Finally, an electronic–PCR service is provided. The user provides several primer pair sequences and a search is performed to identify which sequences potentially could result in a PCR amplification product. This information can be used to assign gene annotations on markers placed on genetic maps.

Cross references to Enzyme database [11] are made by means of similarity searches. The sequence annotation section of `SisGen` client has an option to visualize ESTs annotated as enzymes inside their corresponding metabolic pathway(s), by performing queries to KEGG database [16]. It is possible to interrogate which ESTs map to a specific metabolic pathway and provide a visual component capable interacting with KEGG and the EST database (Fig. 3d).

3.3 Practical Applications

EST projects of varying complexities are currently being managed by `SisGen` from small to large scale. At one end the project for the plant parasite contains about 2,000 ESTs from one cDNA library [17]. Conversely, the Genolyptus project [18] contains $\approx$ 130,000 sequences from four

eucalyptus species, obtained from more than 20 cDNA libraries, but also including genomic sequences derived from BAC (Bacterial Artificial Chromosome) ends.

4 Discussion

The inherent complexity of genomic sequencing efforts was the main motivation to create a new software for managing EST data. The multi–layer architecture centered on CORBA offers several advantages in the software engineering perspective, that sets it apart from previously reported software solutions devoted to this problem [5]. The main advantages of such design are improved modularity, efficiency and better testing and debugging. Also, the choice to create GUIs instead of using web browsers for the presentation layer, although time–consuming, pays off in terms of added capabilities of the client software to handle the heterogeneous and data–rich environment of genomics.

The core of the software, based on CORBA as the middleware, has some disadvantages tough. Albeit a popular enterprise solution last decade, several issues about CORBA complexity and maintainability were raised [19]. In our experience complexity was not an issue, since we streamlined the code to use only essential CORBA services. Still some CORBA aspects were not satisfactory, like firewall traversal and lack of ORB interoperability. The inclusion of a business delegate in our platform provides an extra level of independence from the middleware technology. In principle, porting to another middleware solution like Java RMI (Remote Method Invocation) or web services would only involve the redesign of the business delegate itself.

5 Conclusion

A new concept of EST management software is presented. It is currently in full production managing dozens of projects. In the future we envision improving data integration, by providing compatibility layers to data models such as the Generation Challenge Program standards for crop data [20] and exposing several data querying modules as BIOMOBY services [21] to enable interoperability with other bioinformatics servers.

References

1. Miller, C.J., Attwood, T.K.: Bioinformatics goes back to the future. Nature Reviews Molecular Cell Biology 4, 157–162 (2003)
2. Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F.: Complementary dna sequencing: expressed sequence tags and human genome project. Science 252, 1651–1656 (1991)
3. Paquola, A.C.M., Nishyiama, M.Y., Reis, E.M., da Silva, A.M., Verjovski-Almeida, S.: ESTWeb: bioinformatics services for EST sequencing projects. bioinformatics 19, 1587–1588 (2003)

4. Nagaraj, S.H., Deshpande, N., Gasser, R.B., Ranganathan, S.: ESTExplorer: an expressed sequence tag (EST) assembly and annotation platform. Nucleic Acids Research 35, W143–147 (2007)
5. Nagaraj, S.H., Gasser, R.B., Ranganathan, S.: A hitchhiker's guide to expressed sequence tag (EST) analysis. Briefings in Bioinformatics 8, 6–21 (2007)
6. Stevens, R., Miller, C.: Wrapping and interoperating bioinformatics resources using CORBA. Briefings in Bioinformatics 1, 9–21 (2000)
7. Ewing, B., Hillier, L., Wendl, M.C., Green, P.: Base–calling of automated sequencer traces using phred. i. accuracy assessment. Genome Research 8, 175–185 (1998)
8. Chou, H.H., Holmes, M.H.: DNA sequence quality trimming and vector removal. Bioinformatics 17, 1093–1104 (2001)
9. Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., Quackenbush, J.: TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19, 651–652 (2003)
10. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25, 3389–3402 (1997)
11. Bairoch, A.: The ENZYME database in 2000. Nucleic Acids Research 28, 304–305 (2000)
12. Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y., Gish, W.R.: A general approach to single–nucleotide polymorphism discovery. Nature Genetics 23, 452–456 (1999)
13. Kolpakov, R., Bana, G., Kucherov, G.: mreps: Efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Research 31, 3672–3678 (2003)
14. Marinescu, F.: Ejb Design Patterns: Advanced Patterns, Processes, and Idioms. John Wiley & Sons, Inc., New York (2002)
15. Senger, M., Rice, P., Oinn, T.: Soaplab - a unified sesame door to analysis tools, 509–513 (2003)
16. Kanehisa, M., Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28, 27–30 (2000)
17. Pappas, G.J., Benabdellah, K., Zingales, B., González, A.: Expressed sequence tags from the plant trypanosomatid Phytomonas serpens. Molecular and Biochemical Parasitology 142, 149–157 (2005)
18. Grattapaglia, D.: Integrating genomics into eucalyptus breeding. Genetics and Molecular Research 3, 369–379 (2004)
19. Henning, M.: The rise and fall of CORBA. Queue 4, 28–34 (2006)
20. Bruskiewich, R., Davenport, G., Hazekamp, T., Metz, T., Ruiz, M., Simon, R., Takeya, M., Lee, J., Senger, M., McLaren, G., Hintum, T.V.: Generation challenge programme (GCP): standards for crop data. Omics 10, 215–219 (2006)
21. Wilkinson, M.D., Links, M.: BioMOBY: an open source biological web services proposal. Briefings in Bioinformatics 3, 331–341 (2002)