Leading the management of ESP-SOL genomic project

Victoria Martín\textsuperscript{1,2,$\S$}, Antonio Muñoz\textsuperscript{1,2,$\S$}, Rafael Fernández\textsuperscript{3} and Oswaldo Trelles\textsuperscript{1,*}

(1)University of Malaga, Computer Architecture Department
(2) Integrated Bioinformatics GN5-node at the National Institute for Bioinformatics, Spain
(3)Experimental Station La Mayora – CSIC
(* correspondingle author ots@ac.uma.es
($\S$)these authors equally contributed to this work

MOTIVATION

Genomic projects usually produce huge quantities of heterogeneous data such as sequencing data, phenotypical characteristics, raw and processed microarray data that must be stored and shared by all the project participants.

Handling and sharing this load of information poses additional challenges in project management tasks. The problem increases when several groups from different laboratories disperse around the world need to share documents or simply use the available information in a confidential way. In this context we propose a general genomic projects management system (pUMA stands for project at the University of Malaga) with the aim of making easier the information handling and the data analysis by the users through a customisable project web-portal. A first prototype has been implemented and currently is being used as an essential tool for the progress of the ESP-SOL project.

BACKGROUND

The development of laboratory technologies has produced that genomic projects produce more and more results related to different areas of research in a short period of time. The basis of any genomic project is to have all the data produced by the research groups in a database, but having all this information in a standard format, being accessible to every members in a easy and protected way, not being dependent of a web administrator and software, and giving the possibility of upload new information directly by the provider group wherever they are in a easy way is also important.

Different approaches have been proposed to deal with project management. LIMS (Laboratory and Information Management System) and EMS (Experiment Management System) software are specifically designed to provide support data flow control and laboratory administration, inventory, project creation and execution, and equipment operation. Other specific software cover specific steps in a project (e.g. BASE [1] and arrayExpress [2]) such as organization, storage and tracking of all data produced in microarray experiments. Finally, it is frequent that each new project develop specific and customised web-portals to deal with this problem, e.g. the SGN network at http://www.sgn.cornell.edu/ and the Planet Project at www.eu-plant-genome.net.

METHODS

The following demands need to be solved in most of the genomic projects:

Content: Every project needs –from one side- to cover the dissemination activities, that must be public accessible and from the other side data-handling and tools sections should only be available to project members. Contents is also related with organization. In this sense, the different areas should be structured in hierarchical way (including protection/access mechanisms) to facilitate localization and exploitation.

Access: It is very important a safe access where confidential information will be only available to the owners and their partners (and not all, because there are also permissions inside the private area). But it is also important to include a free access, to allow the project dissemination.
Data management: The system should provide to organise the data sources in different areas: Transcriptomic, genomics, phenotypical, chemical, proteomics, etc. It is also remarkable that not all genomic project have the same needs so, the data management must be done by modules.

Data analysis tools: In the same way, the project manager should give facilities to use the stored data in integrated and coherent way using external software and tools for different necessities such as statistical data processing, specific algorithms or general visualization tools, etc.

From this requirements we devise a general and customisable software solution for the design, maintenance and exploitation of project data. Regarding data management, pUMA’s approach is based on “Entity-Attribute-Value” database scheme which is appropriate to satisfy storage of data with very different composition and provenance. New fields –as they arise- are defined by the system administrator and organised in areas or zones. Access permissions are defined in unix-like style (owner; group and others). In this way adequate rights are needed to access, modify; and in general, manage each data field inside a given zone.

Figure 1. pUMA core architecture comprises a set of libraries for data handling and web scripts pages that works with different databases and produces the dynamic web-page content. On top of that library a group of applications supports the pUMA functionality. To increase the level of user friendliness and to allow remote access to pUMA all the above applications have front-ends in a web server that are written in PHP and produces dynamic HTML user web pages. Thanks to web scripts and databases the web portal is easily configurable and extensible with minimum administration effort.
A plethora of well-trained software is currently available in the public domain to manage specific datatypes (e.g. BASE, arrayExpress). Instead of duplicate efforts we opted by design a mechanism to integrate this software. In some cases, software is provided with its own permission scheme, thus, a plug-in to directly connect data avoiding double registration has been designed.

A similar situation as for data management occurs in data analysis tools. There is a large number of well-trained software currently available in the public domain for data processing. Some of them are even considered as the standard of fact for specific data-type processing (e.g. R or bioconductor for statistical data processing); and on the other hand, services arise dynamically when new algorithms are implemented and deployed in the form of programs running in a web-server or as desktop applications like PreP for microarray data pre-processing, engene for data processing and ARco for extract information where is not evident through association rules.

RESULTS AND DISCUSSION

pUMA satisfy all the conditions previously described for a web-based open-source application which aim is to manage genomic projects. The system is completely configurable and extensible to cover new project requirements. As a proof of concept, here we report the installation of pUMA to help in the management of the ESP-SOL project (http://chirimoyo.ac.uma.es/espsol).

Anonymous users are allowed to surf around the dissemination information and browse the web portal through a configurable three-level menu. An authenticated user can simply explore the stored data, search information related to the sequencing project, or check whether the chosen data are of our interest with the visualization tools adapted in the portal from R statistics packages, in order to finally download the data to the user local machine to analyse them. Other additional tools are available in the portal for data preprocessing, clustering and knowledge discovery.

For data storing the generic-database has been adapted to morphological and chemical data, sequencing data is stored using a Distributed Annotation System implementation, GBrowse [3] and all the transcriptomic data are stored, organized and retrieved through the BASE system.

pUMA integrates tools for a complete data analysis like transcriptomic data pre-processing (PreP+07), statistical packages from R, functions needed in the projetc such as histograms, correlation and boxplot graphs are implemented. Access to external applications can be provided by an easy click-to-use link system; i.e. data processing tools like engene, connectivity with INB web-services and extraction of association rules using ARco (see http://chirimoyo.ac.uma.es). Other specific tools like volatiles analysis programs can be incorporated into the portal as soon as users demand them.

All the possibilities given to the project users make that pUMA play an important role in the project advance from the early beginnings because its easy organization, the possibility of creating new areas in a very short time, the adaptability to the project requirements and the intuitive way to access, share and safely process project data.

ACKNOWLEDGEMENTS

This work has been partially supported by grant “Identification of genes and molecules associated to tomato fruit quality and participation in the sequencing of euchromatic regions of chr 9: a genomic approach (ESP-SOL)” by the Genoma-España foundation. Antonio Perez Púlido for his great work making possible data visualization and annotation tools in the web-portal.

Authors would like to thanks all the ESP-SOL consortium by contibution in software specifications

REFERENCES


