SEDM, a Expression Data Module for the SOL Genomic Network Database

Aureliano Bombarely[1], Kieron D. Edwards [2], Naama Menda[1], Isaak Tecle[1], Anuradha Pujar[1], Tom York[1], Adri Mills[1], Joseph Gosselin[1], Robert Buels[1], Stephen A. Coates[2] & Lukas Mueller[1]

1- Boyce Thompson Institute for Plant Research, Tower Rd, Ithaca, NY 14853 USA
2- Advanced Technologies (Cambridge) Ltd., 210 Cambridge Science Park, Milton Road, Cambridge, CB4 0WA, UK.

BACKGROUND:

The technological advances during the last decade have changed the way biological research is conducted, transforming biology into an information-based science. Microarrays can analyze the expression of thousands of genes per hybridization, and the new sequencing technologies such as 454 Life Sciences or Solexa can produce millions of DNA sequences per run. The development of biological databases tasked with storing this massive quantities of data in a scalable fashion is a challenge. The SOL Genomics Network (SGN, http://solgenomics.net/) is a clade oriented database dedicated to the biology of the Solanaceae family which includes a large number of closely related and many agronomically important species such as tomato, potato, tobacco, eggplant, pepper, and the ornamental Petunia hybrida [1]. The SGN database stores different types of biological data such as transcript and whole genome sequences, markers, phenotypic data and the annotations associated with them. A new module, called SEDM (SGN Expression Data Module), has recently been created as a modular set of database tables, and application programming interfaces (APIs) to store, manage and present graphical interfaces for expression data produced by different microarrays platforms and the next generation sequencing technologies. The expression data is interlinked with other datatypes in the SGN database, such as curated locus information and phenotypes, for maximum data mining possibilities.

RESULTS:

A SQL and Perl modules for expression data in the SGN database were developed with the following features:

1- Modular design compatible with other biological database systems as Chado [2]. It is composed by three parts (figure 1):

I - SQL Database tables in three set of tables:

+ Metadata, to store information about who, when and how the data was added into the database.
+ Biosource, to store information about the samples and the protocols used for the expression data
+ SEDM, to store expression data and the some analysis of these expression data as correlation analysis and clustering

II - Perl objects to manipulate the data from the database in two levels:

+ 1st: DBIx::Class derived objects to interact with the database (one object per table)
+ 2nd: Object to coordinate the data manipulation between different tables.

III - Mason modules to create graphical interface for the web-browse

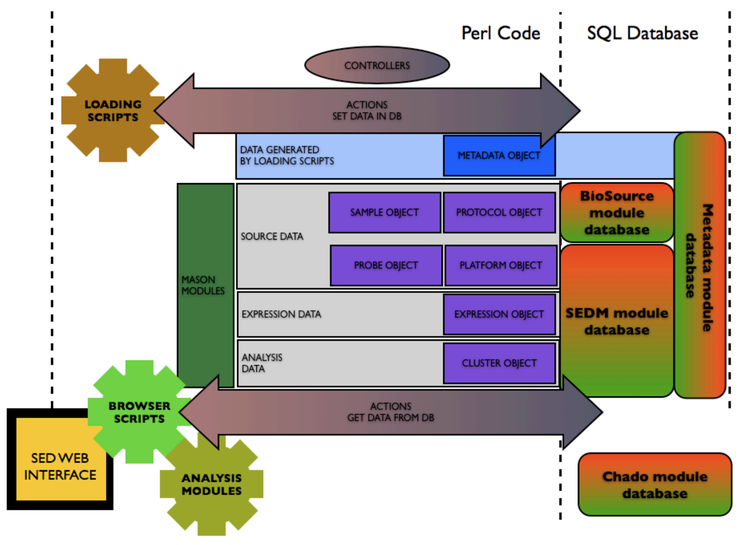

Figure 1: SEDM system components.

2- Expression data are organized by probes. They are associated with SGN sequences such as unigenes (figure 2).

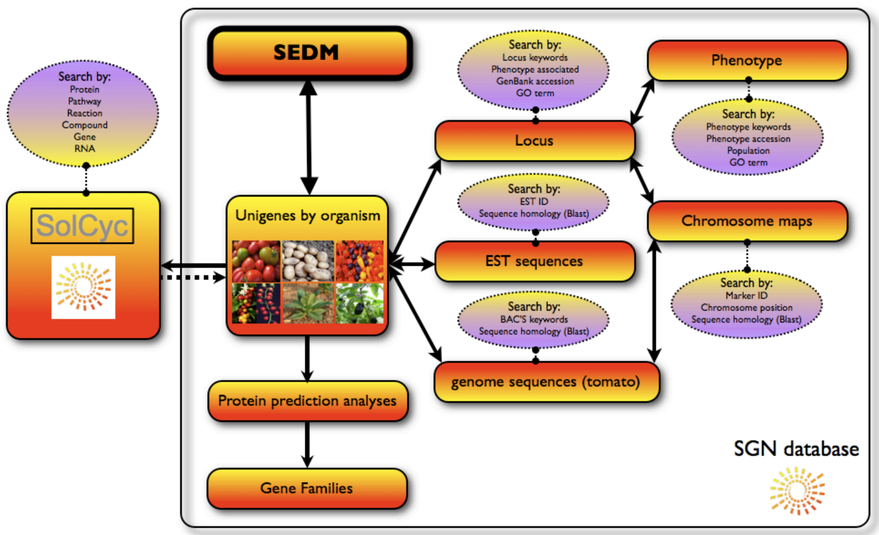

Figure 2: Interrelation of SEDM module with other SGN modules

3- Web interface interrelated with other SGN features as locus or unigenes (figure 3) expanding the ways to search information.

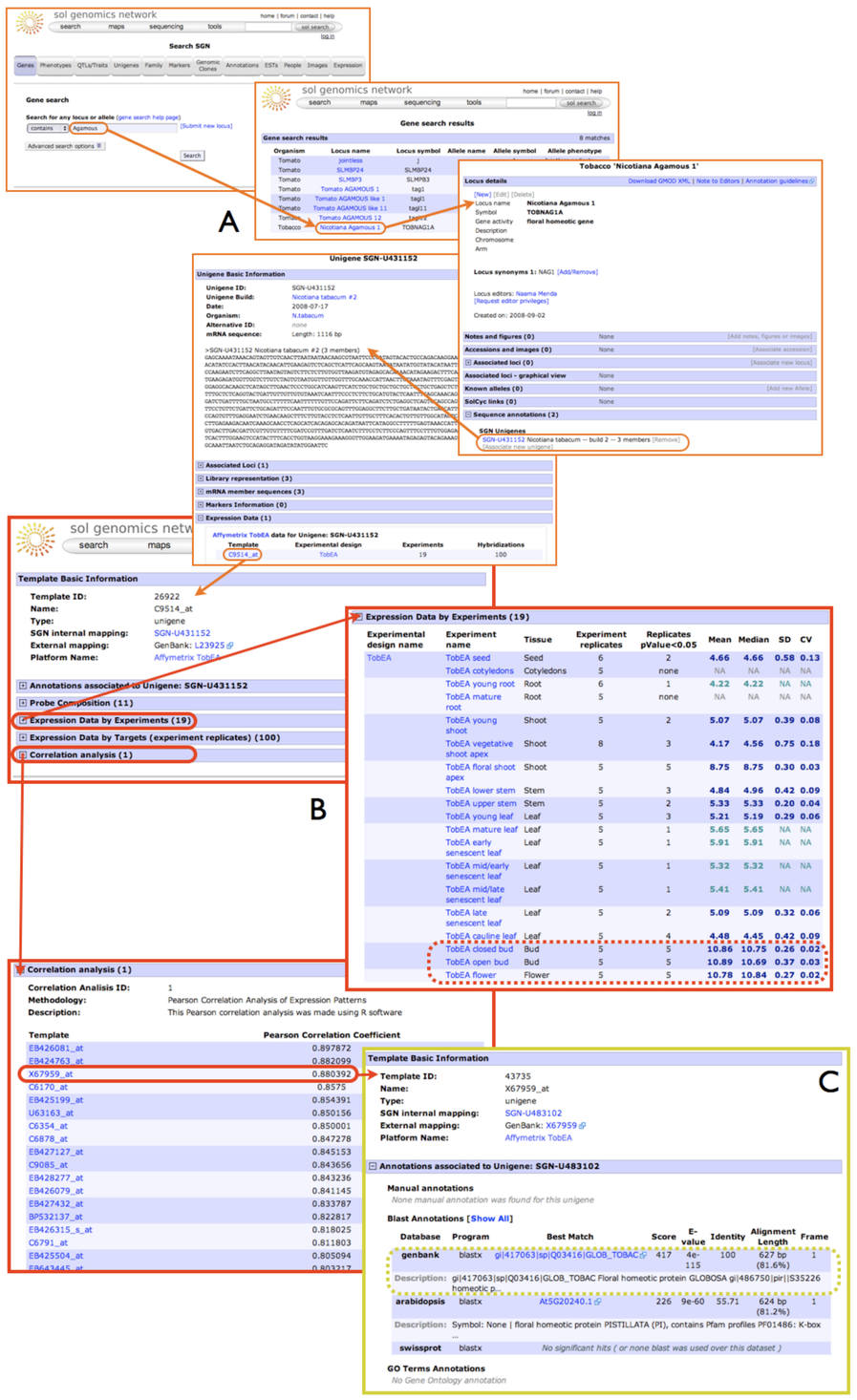

Figure 3: An example of the use of SEDM through web browser. A- Search of a locus based in a term and navigate to unigene page linked with this locus. In the unigene page there are links to some probes associated with it. B- Probe web-page with expression data by experiment and correlation list. C- Probe that have correlation.

CONCLUSION:

A modular database system was developed to store and manage the expression data such as microarray data. A first application of this system has been the storage of the TobEA in the SGN database [3]

REFERENCES:

[1] The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R, Binns J, Lin C, Wright MH, Ahrens R, Wang Y, Herbst EV, Keyder ER, Menda N, Zamir D, Tanksley SD. Plant Physiol. 2005 Jul;138(3):1310-7.
[2] A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Mungall CJ, Emmert DB; FlyBase Consortium. Bioinformatics. 2007 Jul 1;23(13):i337-46.
[3] TobEA: An Atlas of Tobacco Gene Expression from Seed to Senescence. Edwards K.D, Bombarely A, Story G.W., Allen F., Mueller L., Coates S.A. and Jones L. Submitted to BMC Genomics

FUNDING SOURCES:

HOME-PAGE: http://solgenomics.net

CONTACT: sgn-feedback@solgenomics.net