Overview of the BSGC LIMS

Contact Person: John-Marc Chandonia; e-mail: JMChandonia@lbl.gov

A Laboratory Information Management System (LIMS) stores experimental results for targets, as well as bioinformatic predictions and other information useful to experimentalists. After investigating the possibility of adapting a LIMS from another structural genomics center to fit the protocols and procedures used at the BSGC, we determined that it would be more efficient to develop a LIMS locally.

The current LIMS consists of two primary databases, distributed between two BSGC components. The first, the "central LIMS," tracks overall progress on each target and contains data necessary for generation of public reports required by the NIH. These reports (a HTML table designed for human readability and XML tables designed for automatic export to public databases TargetDB and PEPCdb) are automatically generated and archived on our public website on a weekly basis. Except for these public reports, access to the central LIMS is limited to a range of IP addresses associated with the BSGC, and data entry access is further limited to authorized researchers via a password protection mechanism. The central LIMS consists of two parts: a front end web interface written in PHP, which displays data and allows human interaction, and a back end database built using MySQL which stores the data. Each experiment is tracked in a journal, which records the date and time, type of experiment, identity of the experimentalist, and basic results of the experiment. This journal is "write-only;" if an experiment is found to be invalid or repeated, earlier data may be annotated as invalid but not overwritten, allowing a complete history of BSGC experiments to be analyzed at a later data. The central LIMS also collects and stores information and predictions about each target and Mycoplasma ORF in the database, and displays the information to BSGC researchers via the web interface. Examples of such data include homologous proteins, DNA and amino acid sequences, codon usage, potential cloning problems such as nonstandard codons or restriction sites, predicted molecular weight, cellular half-life, and stability, predicted domain architecture, as well as other predicted biophysical properties used in target selection, including coiled coils, low complexity regions, and transmembrane helices. These data are calculated locally using standalone programs, or obtained from public internet resources using automated methods; data from both sources are automatically obtained and imported into the database every time new targets are selected. We automatically design PCR primers for each target, display the primers and predicted melting temperatures on the web site, and automatically create primer order forms at users' request. The primer design procedure is now in its third generation, and is updated with regard to the current procedures used in Component II.

The second database in the LIMS, the "Crystallographic Data Repository," is developed by component VI. This repository contains data collected in crystallographic experiments as well as the results of data analysis tools used in solving structures. The two databases share a common authentication mechanism and a similar web-based interface. When targets are successfully crystallized, data on those targets is automatically exported from the central LIMS to the crystallographic data repository. Processed data collected in crystallographic experiments is also deposited by the Component III experimentalists. This data is processed automatically using a variety of tools. Final coordinates of each solved structure are stored in the crystallographic data repository, and deposited to the PDB. The Crystallographic Data Repository is described in more detail in a separate document.

Our current LIMS is third generation software; each generation was built from the previous one using design principles we learned from prior experience. The first generation LIMS contained many of the features currently in the central LIMS for target status tracking and data reporting. The second generation featured a nearly identical interface, but was based on a complete redesign and rewrite of "middleware" code, which reduces the possibility of introducing bugs when adding new reports or changing procedures; this code also performs basic checks on the data entered to reduce the possibility of human error, such as entering incorrect experimental dates. The third generation integrates more data from other components, including automatic import of annotations provided by component VII. The third generation also introduced a federated approach to our LIMS development: separation of the crystallographic data repository from the central LIMS allowed rapid development and adaptation of the former database to new experimental protocols and analysis tools. However, unlike completely separate databases, the federated approach ensures consistency of data between the databases as well as convenience for the users, such as integrated sessions which allow a user to log in only once to use both databases.