GenDB & SEED: Two Genome Annotation Systems Integrated

A. Goesmann, H. Neuweger, V. Vonstein, R. Overbeek, F. Meyer

GenDB
The flood of sequence data resulting from the large number of current genome projects has increased the need for a flexible, open source genome annotation system. From a user's and from a programmer's point of view, several key features can be identified that a genome annotation system has to provide: From a user's perspective, the quality of the automatic annotation, the range of functions available for manual annotation, and the quality of the user interface play the key role. From a programmer's perspective, the application programmer's interface, the internal data representation, and the extensibility of any system are the most important features. The central design concept of GenDB (Meyer et al. (2003) GenDB--an open source genome annotation system for prokaryote genomes. Nucleic Acids Res. 31(8): 2187-95) is the idea of combining automatic and manual annotation to achieve a very high quality in the fastest possible manner. In order to ensure that no manual effort is lost, annotations are not overwritten. Instead, multiple annotations in chronological order can be added to each region. In general, the latest annotation is presented. The GenDB system represents an effort to build a genome annotation system that not only satisfies the end user's needs, but also enables the programmer to adapt the system and extend it.

The SEED
The SEED genome annotation system is designed to support comparative analysis of genomes. It is based on genomic data available to the public and contains a similarity matrix for all genes that allows for detection and visualisation of chromosomal clusters. The main goal of the SEED system is to provide functions to annotate the growing amount of sequenced genomes by changing the common approach of annotating one genome at a time to annotating homologous genes across several genomes. The genes that contribute to the central metabolism will be annotated using subsystems. They are an informal metabolic reconstruction of partitions of the cellular machinery. One subsystem is defined by the specification of its functional roles, and the inventory of which genes in a specific organism implement the functional roles. The SEED system offers methods to define, manipulate and populate subsystems.

Integration of GenDB and the SEED
SEED relies on predicted locations of chromosomal features (e.g. coding sequences, RNAs) on the genome of an organism and lacks detailed methods to predict, analyze or correct genes on the sequence level. Through the integration with GenDB via the BRIDGE layer (Goesmann et al. (2003) Building a BRIDGE for the integration of heterogeneous data from functional genomics into a platform for systems biology. J. Biotechnol 106(2-3): 157-67) it is now possible to enrich the SEED system with this functionality. A first application of the joined annotation systems is the directed search for uncalled genes detected by gaps in subsystems or chromosomal clusters. Additionally, the automated gene prediction pipeline of GenDB was already used to generate the gene calls for over 50 genomes that have been added to the SEED database.

Current work and planned extensions
In addition to the two genome annotation systems described so far, we are currently working on several other applications. For the analysis of all kinds of sequence reads we have implemented SAMS, a Sequence Analysis and Management System that can be used for read quality control and annotation within whole genome shotgun, EST, and meta-genome shotgun projects. EMMA is a MAGE compliant software platform for transcriptome data analysis including a LIMS component (ArrayLIMS) and our ProDB software package can be used for the large scale analysis of proteome data. Each system can be used as a stand-alone application, but they can also be linked via the BRIDGE integration layer. Our current work is directed towards integrating all systems providing a unified interface with direct and interactive links between each other.

Availability
We have setup a public GenDB demo project containing sample data for S. meliloti that is freely accessible at http://www.cebitec.uni-bielefeld.de/groups/brf/software/gendb_info/demo.html. For local installation on a UNIX/Linux system the software basically requires MySQL, Perl, Apache, and several bioinformatics tools like Glimmer, BLAST, HMMer, InterPro etc. Further information and detailed installation instructions for GenDB, EMMA, ProDB, and other applications developed in our group are available at http://www.cebitec.uni-bielefeld.de/groups/brf/software/. Details about the SEED genome annotation system can be found at http://theseed.uchicago.edu/FIG/index.cgi. In addition to the public demo project we are also offering private and password protected project setup on a collaborative basis for ongoing research projects, including training and support at various levels. For local installations we provide code bundles of all applications on our FTP server at ftp.cebitec.uni-bielefeld.de/pub/software/. All source code is released under the GPL license. For further details and more information about the software please contact fm@cebitec.uni-bielefeld.de.