MAGPIE is a system designed to reside locally at the site of a genome project and actively carry out analysis of genome sequence data as it is generated.[1,2] DNA sequences produced in a sequencing project mature through a series of stages that each require different analysis activities. Even after DNA has been assembled into contiguous fragments and eventually into a single genome, it must be regularly reanalyzed. Any new data in public sequence databases may provide clues to the identity of genes. Over a year, for 2 megabases with 4-fold coverage, MAGPIE will request on the order of 100,000 outputs from remote analysis software, manipulate and manage the output, update the current analysis of the sequence data, and monitor the project sequence data for changes that initiate reanalysis.

In collaboration with Canada's Institute for Marine Biosciences and the Canadian Institute for Advanced Research, MAGPIE is being used to maintain and study comparative views of all open reading frames (ORFs) across fully sequenced genomes (currently 5), nearly completed genomes (currently 2) and 1 genome in progress (Sulfolobus solfataricus). Together, these genomes represent multiple archaeal and bacterial genomes and one eukaryotic genome. This analysis provides the necessary data to assign phylogenetic classifications to each ORF (e.g., "AE" for archaeal and eukaryotic). This data in turn provides the basis for validating and assessing functional annotations according to phylogenetic neighborhood (e.g., selecting the eukaryotic form of a biochemical function over a bacterial form for an "AE" ORF).[3]

Once an automated functional overview has been established, it remains to pinpoint the organisms' exact metabolic pathways and establish how they interact.To this end, the WIT (What Is There) system supports efforts to develop metabolic reconstructions. Such constructions, or models, are based on sequence data, clearly established biochemistry of specific organisms, understanding of the interdependencies of biochemical mechanisms. WIT thus offers a valuable tool for testing current hypotheses about microbial behavior. For example, a reconstruction may begin with a set of established enzymes (enzymes with strong similarities in identified coding regions to existing sequences for which the enzymatic function is known) and putative enzymes (enzymes with weak similarity to sequences of known function). From these initial "hits," within a phylogenetic perspective, we identify an initial set of pathways. This set can be used to generate a set of expected enzymes (enzymes that have not been clearly detected, but that would be expected given the set of hypothesized pathways) and missing enzymes (enzymes that occur in the pathways but for which no sequence has yet been biochemically identified for any organism). Further reasoning identifies tentative connective pathways.

In addition to helping curators develop metabolic reconstructions, WIT lets users examine models curated by experts, follow connections between more than two thousand metabolic diagrams, and compare models (e.g., which of certain genes that are conserved among bacterial genomes are found in higher life). The objective is to set the stage for meaningful simulations of microbial behavior and thus to advance our understanding of microbial biochemistry and genetics.

DOE Contract No. W-31-109-Eng-38 (ANL FWP No. 60427).

References

[1] T. Gaasterland and C. Sensen, Fully Automated Genome Analysis that Reflects User Needs and Preferences --a Detailed Introduction to the MAGPIE System Architecture, Biochemie, 78(4), (accepted).

[2] T. Gaasterland, J. Lobo, N. Maltsev, and G. Chen. Assigning Function to CDS Through Qualified Query Answering. In Proc. 2nd Int. Conf. Intell. Syst. for Mol. Bio., Stanford U. (1994).

[3] T. Gaasterland and E. Selkov. Automatic Reconstruction of Metabolic Structure from Incomplete Genome Sequence Data. In Proc. Int. Conf. Intell. Syst. for Mol. Bio., Cambridge, England (1995).

Database Transformations for Biological Applications

G. Christian Overton, **Susan B. Davidson**,[1] and **Peter Buneman**[1]
Department of Genetics and [1]Department of Computer and Information Science; University of Pennsylvania; Philadelphia, PA 19104
Overton: 215/573-3105, Fax: -3111, *coverton@cbil.humgen.upenn.edu*
Davidson: 215/898-3490, Fax: -0587, *susan@cis.upenn.edu*
Buneman: 215/898-7703, Fax: -0587,*peter@cis.upenn.edu*
http://agave.humgen.upenn.edu/cpl/cplhome.html
http://sdmc.iss.nus.sg/kleisli-stuff/MoreInfo.html

We have implemented a general-purpose query system, Kleisli, that provides access to a variety of "non-standard" data sources (e.g., ACeDB, ASN.1, BLAST), as well as to "standard" relational databases. The system represents a major advance in the ability to integrate the growing number and diversity of biology data sources conveniently and efficiently. It features a uniform query interface, the CPL query language, across heterogeneous data sources, a modular and extensible architecture, and most significantly for dealing with the Internet environment, a programmable optimizer. We have demonstrated the utility of the system in composing and executing queries that were considered difficult, if not unanswerable, without first either building a monolithic database or writing highly application-specific integration code (details and examples available at URL above).

In conjunction with other software developed in our group, we have assembled a toolset that supports a range of data