Findings

We have thus far made the following advances:

A. Communication between BioBIKE and the SEED

The SEED is a database consisting of hundreds of bacterial and phage genomes, with genes connected across genomes by human-curated subsystems. BioBIKE is a graphical interface that combines a general purpose programming language with concepts of molecular biology and connects the two to an underlying database. Our major goal for the first year was to connect the SEED and BioBIKE. Initially we experimented with a SOAP connection (a protocol for web-based interaction), but we found this method to be unacceptably slow. We developed instead a direct connection, where BioBIKE resides on the same machine as the SEED and accesses the same MySQL database.

Since that time, we have been increasing the internal access of BioBIKE to the information within the SEED and building tools to make practical access by users. As a result, it is now possible to exploit the SEED table of precompututed protein comparisons, making comparisons between a protein in SEED organism and all other proteins very fast. For comparison of arbitrary sequences with proteins and genomes of SEED organisms, we have built precomputed Blast databases that permit rapid comparisons over either standard or arbitrary subsets of SEED organisms.

A menu system has been devised to allow easy access within BioBIKE to all SEED organisms, all subsystems and their internal subcategories, and all genes within these categories. This has proven invaluable in annotating large numbers of related genes at one time.

B. Expansion of the number of organisms BioBIKE can handle

Until recently, BioBIKE could consider only a limited number of genomes simultaneously, limited by available memory (on our current machine, the limit was several dozen bacterial-sized genomes). To handle the thousands of genomes available through the SEED database, the way BioBIKE handled organisms and their genes and proteins had to be fundamentally changed. During the first year, we spent a good deal of time planning the change and testing the feasibility of the caching strategy we finally adopted. By the end of the first year, a fully functional version of BioBIKE was established that could make use of an essentially unlimited number of organisms.

A great deal in BioBIKE has had to be modified to absorb the changes that have come from a huge increase in its database and to make the information available to researchers in a form they can understand. During the second year, we have modified the way in which BioBIKE initializes its database to translate the information provided by the SEED into organism, gene, and replicon names more familiar to researchers.

C. Standardization of the BioBIKE interface codebase

During the first year, we transformed the computer code underlying the interface to make use of standard javascript libraries rather than the homegrown lispscript modules we previously used. This major change in the codebase will have several benefits. First, it will enable us to expand accessibility to PHANTOME to all common browsers. At present, users are required to reach the

interface with Firefox. Second, transitions of browsers to new versions will be less apt to cause new bugs in the interface. Third, the code will be more easily maintained.

D. Visualization tools

Researchers and students new to computer programming often have a difficult time making the connection between abstract computational constructs (like arrays) and the objects they represent. Similarly, users often fail to construct an adequate mental model of the genome sequences they are analyzing. We have introduced tools to help such users see more directly what they are working with. At the same time, we expanded the plotting capabilities of BioBIKE to make it easier to create graphs on the spur of the moment. It is now possible for users to explore a graphical circular map of an organism and to exploit the SEED's ability to line up regions of similar genomes, marking orthologs by color.

E. New capabilities, new bugs

We've continued to add to the power made available to users of BioBIKE (and eventually to PhAnToMe). For example, it is now simple for researchers to compare their favorite sequences against all available genomes, with respect to either dinucleotide frequencies or codon usage. At the same time, we have continued to find and remove bugs in the code.

PhAnToMe (including the PhAnToMe BioBIKE instance) moved to a new and faster server Summer 2010.

F. Annotation interface and tools

The major advance during the second year has been the introduction of a graphical annotation interface that allows researchers to inspect what is known about a gene and to add to the knowledge base. The interface was sufficiently complete to use during the Jan/Feb 2011 annotation workshop held in Tucson. It allows users to view and possibly modify information about various aspects of a gene: it's coordinates, sequence (non-modifiable), functional role and associated subsystem, physiological role, biochemical role, and known mutants. All changes in the knowledge base are tagged with an evidentiary tag, along with information provided by human annotators regarding the evidence (e.g. a supporting publication). All fields can be searched within BioBIKE, for example for genes with experimental evidence for a given functional role (the search capability has not at time of writing been implemented).

Most importantly, the annotation facility is easily accessible from within BioBIKE, so that when some characteristic of a gene is discovered by a researcher by analysis within BioBIKE, it is easy to incorporate that knowledge through the annotation page.