Name and brief description of initiative:
NHGRI Genome Informatics Program

Brief description of goals of initiative:
The genome informatics program at NGHRI encourages development and maintenance of databases of genomic and genetic data. This emphasis includes new tools for annotating complex genomes so as to expand their utility. The program also supports the production of robust, exportable software that can be widely shared among different databases in order to facilitate database interoperability. These bioinformatics resources will allow the scientific community efficient access to genomic data, which will enable new types of analyses. The analyses, in turn, will allow for the computer modeling and subsequent experimental validation of the complex pathways and networks that ultimately determine the phenotype of a cell or the causes of many human diseases.

Program contact information: Peter Good, Ph.D. 301-406-7531; goodp@mail.nih.gov

Website address of initiative: http://www.genome.gov/10001735

Brief description of biomedical informatics and computational biology components and their goals:
Model organism databases (MODs) represent one component of the NHGRI genome informatics program that serves to link biological information to the genome sequence. These MODs serve their respective research communities as essential resources. A second component is the UCSC Genome Browser. This resource maps genome features onto available genome sequences including the human genome and provides comparative genomics tools to leverage data from the NHGRI large scale sequencing program. Another component are knowledgebases that provide information about proteins or pathways. UniProt is the world's most comprehensive catalog of information on proteins while Reactome provides a curated resource of core pathways and reactions in human biology. Additional components are consortia that develop standards and ontologies for functional genomics including the Gene Ontology Consortium (GOC), responsible for defining a controlled vocabulary to describe gene productions and the Microarray Gene Expression Data Society (MGED), attempting to establish standards for microarray data annotation and exchange, to facilitate the creation of microarray databases and related software implementing these standards, and to promote the sharing of high quality, well annotated data within the life sciences community.

Brief description of resources and tools available for sharing:
All components of this program are considered research resources and freely provide data and tools for sharing. MODs provide curated information about genes and gene products in the respective organism. The Generic Model Organism Database project includes contributions from the MODs and other members in the genome informatics community (www.gmod.org). This project strives to develop a complete set of

software for creating and administering a model organism database. Included in this effort are software for genome annotation (Apollo), pathway analysis (Pathway Tools), genome browsing (Generic Genome Browser), literature searching (Textpresso), and core schema for a model organism database (Chado).

Brief description of integrative efforts:

Standard ontologies/terminologies: The MODs all contribute to the development of the (GOC). The goal of the GOC is to enable integration of information at different resources by providing a controlled vocabulary to describe gene products. MGED is promoting the development of data standards for microarray experiments and of appropriate software to manipulate the data. As necessary, each MOD will develop appropriate controlled vocabularies to describe organism-specific data.

Interactions with other initiatives: The MODs already collaborate extensively with their respective research community to input data from large scale experimental projects.

Opportunities for collaboration or synergy with the NCBCs: Interactions on development and adoption of terminologies and ontologies; sharing of approaches for developing genomic databases; use of high quality curated data as test sets for NLP and machine learning methods.