

## **Beyond Terascale Biological Computing: GIST and Genomes To Life**

Philip LoCascio, Doug Hyatt, Frank Larimer, Manesh Shah, Inna Vokler, Ed Uberbacher  
Computational Biology Program, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee. <http://compbio.ornl.gov/>

High performance computing played a critical part in the successful completion of the working draft of the human genome, and continues to be necessary to handle the ever-increasing flood of new biological data. We have successfully met the first rounds of such computing for biology with the Genomic Integrated Supercomputing Toolkit (GIST). GIST provides a transparent, fault tolerant interface for the research community to an ever increasing suite of accelerated massively parallel biological applications. It also demonstrates key concepts that can be extended for large-scale computation for the Genomes to Life (GTL) program. GIST and associated data sets are accessible via the WWW interfaces of the Genome Analysis Toolkit, Genome Channel and other ORNL tools, and optionally a command line interface for biologists developing new applications.

With the advent of the DOE Genomes To Life Program, we are actively developing new technologies that will help biologists link large-scale experimentally derived biological data with increasingly sophisticated computational analyses. Our basic approach is to support the mode of operation where computational biology is concerned with transactional tool usage upon data and the construction of “recursive” pipelines of analyses, which ultimately execute on supercomputer resources (via GIST). The central theme here is to organize the libraries of biological tools around the available biological data types, and use the existing methodology of context dependent XML schema to classify both the data types and the biological tools. In this way, it becomes possible for users to (i) automatically detect which tools and which data can be combined in a valid operation, (ii) configure linked sets of analysis steps without detailed knowledge of computing or tools, and (iii) record transactions in an RDBMS to look for dependencies and redundancies. Furthermore, existing user interfaces to biological tools in any browser format can be instantly coupled with the appropriate tool-data combinations and linked transparently to high performance operations.

Where available, existing biomedical community templates are being used as the basis for standards which could be established across the GTL enterprise. With new data types rapidly becoming available, a strategy that reuses existing tools and interfaces, supports new tools, and makes it possible to combine tools to perform novel analyses, will be the most effective way to manage software complexity and development cost. Additionally, where data types are not “naturally” aligned, it will be possible to create “filters” for the most common types of conversions. e.g. FASTA -> Masked FASTA.

This intermediate software layer can serve as an effective conduit between users, with their novel complex biological data sets, and the emerging beyond terascale computing infrastructure. Using this approach it should be possible to create friendly interfaces to new classes of algorithms that are built of both new and old components, with the flexibility necessary to tackle biological problems of increasing complexity. Libraries of software for tasks such as metabolic pathway reconstruction, gene regulatory network modeling and cell modeling can be constructed, supported and utilized by the community if organized and accessed in thus manner.

(Research sponsored by the Office of Biological and Environmental Research, US DOE under contract number DE-AC05-00OR22725 with UT-Battelle, LLC)