Report on the Computer Science Workshop for the Genomes to Life Program[1]

U.S. Department of Energy
Gaithersburg, Maryland
March 6–7, 2002

Executive Summary

On March 6–7, 2002, the U.S. Department of Energy (DOE) sponsored a 2-day workshop on computer science for the Genomes to Life (GTL) program. About 50 researchers from universities, national laboratories, research institutions, DOE, and industry attended. The objective was to bring together experts in computational and experimental biology with researchers in bioinformatics and computer science to address the following objectives:

- Discuss computational-science research issues and approaches,
- Identify key computer science challenges for DOE's GTL program, and
- Develop recommendations about how computer science can contribute to major thrusts in the GTL program.

Each day began with presentations by speakers from government agencies, academia, and industry. The intention was not only to outline the daunting challenges in systems biology but also to inspire workshop participants to formulate a program for advanced biology research. Five breakout groups were established: (1) genome annotation, (2) protein expression and proteomics, (3) technical text mining for biological data, (4) simulation tools for cell networks, and (5) interoperability facilitation. At the end of the day, a representative from each group presented a status report on the state of the art, approaches and obstacles in the specific area, and a list of recommendations for future work. These presentations were followed by an open discussion of cross-cutting issues and next steps.

Issues

- Participants agreed that high-performance computing has fundamentally changed the way biologists do science.
- The use of parallel computing systems has enabled high-throughput genome analysis and even comparative analysis.

[1] This report was produced from the best available notes and does not represent a verbatim or consensus document of the workshop.

- Protein chips have offered a viable technology for proteomics studies.
- Modern search engines are allowing access to unprecedented amounts of biological data.
- Cell models have made possible quantitative predictions of metabolic pathways.
- A number of frameworks and tools are starting to support component-based software development.

Nevertheless, a number of outstanding issues were identified.

Genome annotation. A clear need exists for computing systems that automatically produce genome annotations, including protein-function predictions, at a much higher accuracy than currently possible. Annotation of such other features as operons, promoters, transcription factor binding sites, SNPs, and protein complexes must be automated as well. More effective methodologies are needed to validate function predictions, encode the expertise of human annotation experts, and apply confidence levels at multiple levels of granularity. Systematic revisions of outdated genome annotations are required to correct predictions or generate predication for previously unidentified proteins.

Protein expression and proteomics. The proteome is far more complex than the genome; for example, there are at least 300,000 proteins encoded by only about 30,000 genes. Integrating data and uncovering associated regulatory networks will require new methods for pattern discovery and for assigning confidence measures to the resulting computational models. Furthermore, specialized visualization systems will be essential for displaying protein-interaction networks, mapping data to pathways, and examining computational results from cluster analysis.

Technical text mining for biological data. A huge amount of critical biology literature is simply not able to take advantage of modern search-engine technology. Further research is needed to understand how—and indeed whether—relevant technology from text-data mining and natural language processing can be applied effectively to biology. Tools also must support semantic interoperability; key issues involve lexicography, semantics, syntax, and recovery of information implicit in context.

Simulation tools for cell networks. Today's simulations are limited to subsets of processes in individual cells or simple cellular interactions. For more comprehensive modeling, fundamental research issues must be addressed, including representation of multiple levels of spatial and temporal scales in cellular systems and coupling of modeling and simulation with mathematical analysis and experimental databases.

Facilitating interoperability. Component-based architectures are essential for a cross-disciplinary project such as Genomes to Life. Groups such as the forum on Common Component Architecture (CCA) have been developing standards to support a scalable component-based architecture; their work should be adopted and extended. Equally important is further research in data discovery and analysis; the scale and heterogeneity of GTL data sources will require interoperability within and without the GTL program, extensible schemas, and multimodal representations.

Computer Science Challenges

To address these problems, workshop participants formulated a number of specific challenges that require computer science advances, broadly summarized here by topic. The group recommended development of the following:

Data Representation

- Next-generation genome-annotation system with accuracy equal to or exceeding the best human predictions
- Mechanism for multimodal representation of data

Analysis Tools

- Scalable methods of comparing many genomes
- Tools and analyses to determine how molecular complexes work within the cell
- Techniques for inferring and analyzing regulatory and signaling networks
- Tools to extract patterns in mass spectrometry data sets
- Tools for semantic interoperability

Integration Methods

- Methods for integrating dissimilar mathematical models into complex and integrated overall models
- Tools for semantic interoperability

Visualization

- Tools to display networks and clusters at many levels of detail
- Approaches for interpreting data streams and comparing high-throughput data with simulation output

Models

- High-performance, scalable algorithms for network analyses and cell modeling
- Methods to propagate measures of confidence from diverse data sources to complex models

Validation

- Robust model and simulation-validation techniques
- Methods for assessing the accuracy of genome-annotation systems

Standards

- Good software-engineering practices and standard definitions (e.g., CCA)
- Standard ontology and data-exchange format for encoding complex types of annotation

Databases

- Large repository for microbial and ecological literature relevant to GTL
- Big relational database derived by automatic generation of semantic metadata from the biological literature
- Databases that support automated versioning and identification of data provenance
- Long-term support of public sequence databases

Projects

- Series of challenge evaluations to track the state of the art in text processing, data mining, and annotation, as applied to biology
- Collaboratory pilot project in biology (similar to SciDAC projects)