Vision Statement on Bioinformatics and Computational Chemistry

David Konerding

Lawrence Berkeley National Laboratory

Near-Term Vision

An important first step is to build a multi-tier platform to facilitate application scientists' use of the grid through a visual workflow programming environment. This platform will engender a community of software developers who develop additional functionality across various domains for use by application scientists. The design of the platform will support diverse scientific communities with similar needs such as high-energy physics and astronomy, but will initially target biological investigators. The multi-tier platform will simplify access to grid-enabled services, including complex functions such as security, resource discovery, optimal compute and data migration, and coordination, addressing many of the concerns which have limited application scientist usage of the grid. To enable this platform for grid-enabled bioservices, we will need to carry out basic tool and data development in several areas:

- Schema: schema representing core chemical/biological structures such as DNA and protein sequences, sequence alignments, genes, protein structure, and function must be developed. These schema are necessary for reliable data interchange between applications, user input/output.

- Services: Low-level web service wrappers for legacy applications, including the most frequently used applications such as NCBI BLAST are needed.

- Higher-order Bioservices: Nearly all existing biological workflows are composed of pre-existing biological tools lashed to higher-order business/procedural logic. For example, many existing published protein functional annotation systems apply search tools such as NCBI BLAST and HMMER[1] to biological databases including Pfam[2], Ensembl[3] and GO to annotate proteins using a "guilt by association" technique.

- Low-level Bioprimitives: Many computational biology and chemistry investigators do not wish to develop their own low-level bioprimitives. However, they would like to compose their own workflows based on bioprimitives, or re-use existing workflows such as those described above. One common paradigm is the web portal interface. Although the web portal interface has benefits, it is beset with browser interoperability issues and the difficulty of developing rich user interfaces. Drag and drop and graph representation are both very difficult to implement portably even with Dynamic HTML and XHTML.

Long-term Vision

Machine reasoning over Protein Sequence/Structure/Function Ontology

The biological data community has undergone a tremendous revolution in the past few years; data is being generated at much higher rates than ever before. Several large scale projects, including genome sequencing, structural genomics, microarray and microscopic imaging experiments, challenge traditional means of biological data collection, storage, publication and querying. Coupled with reams of legacy data of poor provenance, annotation, homogeneity and cross database linkage, this tide of data threatens to overwhelm biological scientists, greatly reducing the potential for new discoveries. Data collection and archiving has benefited from rapid increases in storage technology, but methods for applying meaning to captured data and forming useful data linkage across different experiments have not kept pace. Further, in most cases of data linkage across databases, the results rarely have referential integrity, especially when the database cross-linkages evolve as new versions of the databases are released.

Several projects are aiming to solve the problem of applying meaning to data by using industry-standard data meaning representation have emerged. The most visible project is GO which is building a set of consistent descriptions of gene products in different databases. Other projects include the Semantic Moby branch of the BioMoby project, GONG[4], and NCI's Ontology for the Cancer Grid[5]. BioMoby aims to develop open-source biological web services for database access, and Semantic Moby addresses many issues that arise when attempting to perform queries across multiple databases: common syntax, common semantics, and service discovery, by applying the industry standard semantic web technologies RDF[6] and OWL[7]. This design decision allows web services to perform machine reasoning over biological data for which semantic annotations exist. Further, even when direct is-a relationships cannot be applied across databases due to weak linkage evidence, the results can be placed in a probabilistic context using Bayesian Network methods. The visual workflow environment must support semantic networks.

References

[1] http://hmmer.wustl.edu/

[2] HThe Pfam Protein Families DatabaseH http://www.sanger.ac.uk/Software/Pfam/

[3] http://www.ensembl.org/

[4] Hhttp://gong.man.ac.uk/H

[5] http://cabig.nci.nih.gov/

[6] http://www.w3.org/RDF/

[7] http://www.w3.org/TR/owl-ref/