

Project Summary/Abstract

This application addresses broad Challenge Area (06) Enabling Technologies and Specific topic, 06-RR-102* Infrastructure for biomedical knowledge discovery.

New discoveries in biology rely increasingly on the integration of data from disparate sources. For example, the Cancer Genome Atlas generates a wealth of data about several cancers, including sequence information, copy number and epigenomic changes, expression and proteomic profiling, and detailed clinical and pathological information. Integrating all of the data helps researchers gain a better understanding of the biological principles underlying the disease. The research process underlying such discovery is inherently collaborative and iterative. Yet, computer systems do not support sharing, refinement, and repetition adequately. We propose to design, build, and integrate into ongoing biological research a computing infrastructure that provides first-class primitives for managing the integration of heterogeneous, scientific data. At the core of this infrastructure is the concept of *sibling data*: a collection comprising related data items potentially obtained from disparate sources.

Our proposed system is called SIB. SIB allows users to group data, share “live” collections with collaborators, and publish snapshots of the collection for others. Currently, these tasks are independent and cumbersome: a researcher’s work takes place on a local system or on a computing resource, collaborators share via explicit email or wiki’s, and publishers provide project web pages to make polished data and scripts available for others. These generic collaboration techniques are not sufficient for rigorous-but-fast scientific collaboration.

Updates to data sets also lead to challenges and unique opportunities. Reproducing the analyses of others is inhibited by ever-changing online databases: different results could represent a failure in the software or represent a real change in outcome. Perhaps those who collect new basic data can observe how new observations alter the conclusions of prior analyses; such an activity could further motivate fundamental study. We believe such analyses are unnecessarily hampered by a dearth of computing system support for repeatability and integration.

We propose to validate the sibdata approach by integrating it with ongoing biological research. Sibdata are accessible through a filesystem interface: our prototype design incorporates a user space file system. This interface allows existing applications and scripts to be integrated into a sibdata instance and shared with others.

The central research of this proposal is twofold: (1) designing and building a SIB infrastructure that will allow efficient access and aggregation of sibling biomedical data, and (2) integrating SIB into the workflow of biology research, thereby leading to new discoveries. The first goal requires SIB provide location-independent access to sibdata and its source data, an expressive and powerful security architecture that protects confidential data in insecure environments, and provenance data needed to reproducing findings and evaluate the reliability of source data. The second goal requires SIB support sibdata integration tools, multi-dimensional indexing of disparate, heterogeneous biomedical data for answering analytical and statistical queries, and tools for extracting entities and relationships from these data sources.