Technical Components of the GTL Knowledgebase

Data, Metadata, and Information

Modern systems biology is inherently dependent on a variety of data to inform statistical inferences, mathematical modeling, and theoretical work. The GTL Knowledgebase (GKB) thus should provide the appropriate data types, metadata structures, visualization capabilities, and analysis and inference tools to enable critical synergy between computational sciences and more traditional experimental approaches.

The GKB should focus on the acquisition, integration, and accessibility of a rich body of data. Effective use of the knowledgebase will require evolving standards to support emerging research themes. The GKB must incorporate processes to receive, transmit, and update information; it also should contain protocols for documenting and assessing the state and quality of the system and its contents.

In addition to hardware, software, and network capabilities, a broader view of the GTL Knowledgebase clearly reveals the need for sustained support of core personnel with scientific and information technology expertise.

To better understand GKB requirements relating to data and metadata, several critical issues must be addressed, including (1) data and their generation by experimentation or simulation and modeling, (2) the use of metadata for setting the context of data to enable their interpretation, (3) data handling (e.g., archiving, annotation, and maintenance), and (4) quality control and assurance (see Fig. 2.1. GTL Knowledgebase, below, and Box 2.1, Data Stewardship and Availability, p. 20).

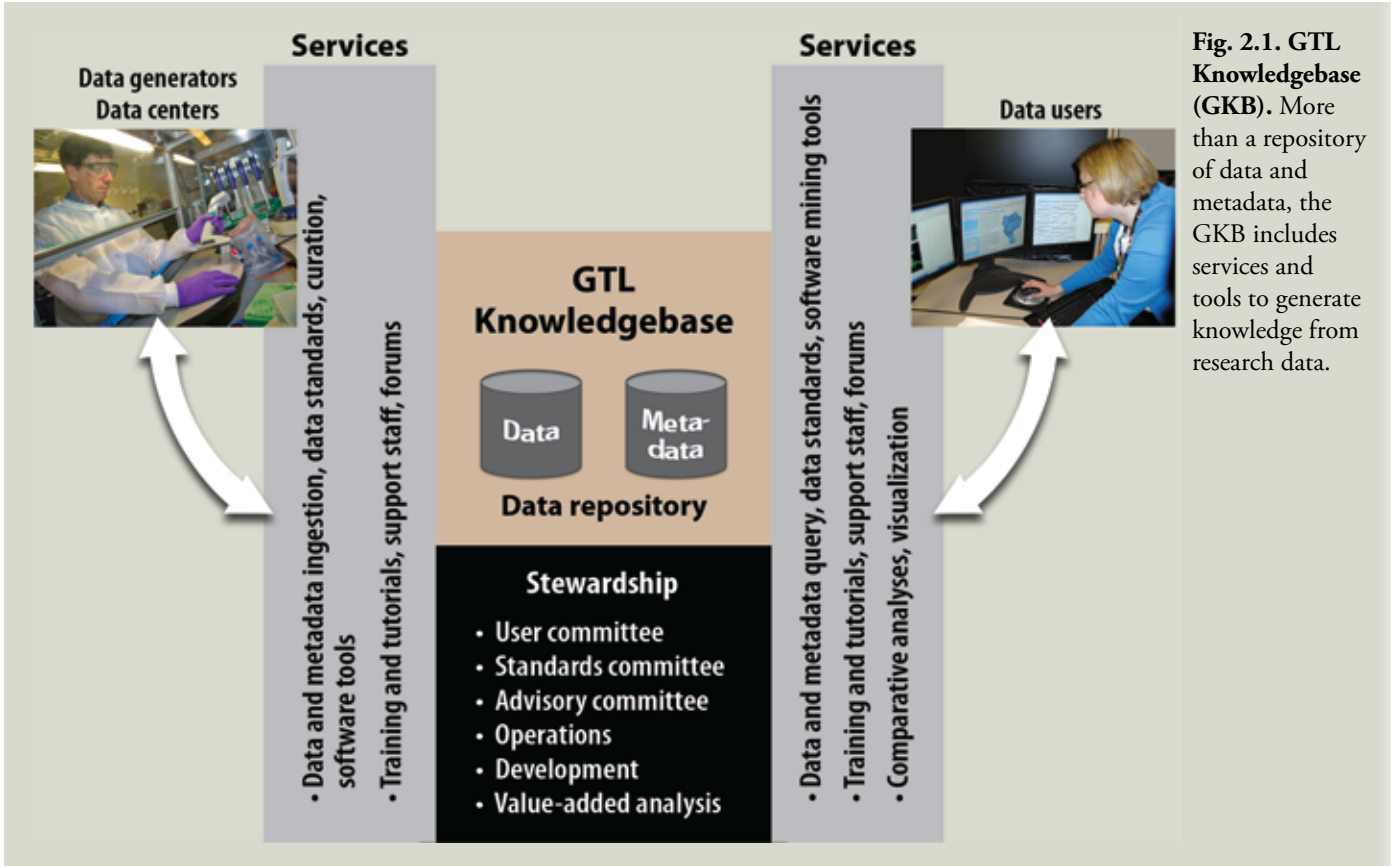

Fig. 2.1. GTL Knowledgebase (GKB). More than a repository of data and metadata, the GKB includes services and tools to generate knowledge from research data.

Box 2.1

Data Stewardship and Availability

Proper stewardship of GTL-generated data will maximize the scientific impact of the program's research investments and will support additional investigations using data-mining activities provided by the GTL Knowledgebase.

- Data submitted to the GKB become public and available to anyone desiring access.
- Regarding data embargoes, the GTL Knowledgebase should be available to the user community for prepublication analysis of experimental or computational data and information. Providing this service would require devising data embargo guidelines that will append the current GTL Information and Data Sharing Policy (see Appendix 1, p. 59, and http://genomicsgtl.energy.gov/datasharing/). In this circumstance, the GKB would serve two functions: integrating publicly accessible data and information and facilitating the analysis of data and information for additional research conclusions.
- The knowledgebase community should develop a reasonable, clear, and extensible embargo policy that can evolve to accommodate the increasing use of nongenomic datasets (e.g., images and simulation outputs).

Data Sources

Findings

The GTL Knowledgebase should support a wide variety of highly complex data from many sources. These data must be comprehensively integrated and structured for analyses and discovery.

- The GKB should gather or link to data from public repositories so users can perform complex queries across information in public systems and across GTL-derived data in the knowledgebase. Public data systems of interest include the Universal Protein Resource (UniProt), Kyoto Encyclopedia of Genes and Genomes (KEGG), National Center for Biotechnology Information (NCBI), and more topically oriented databases (see Appendix 10, List of Web Addresses, p. 139). Effective integration with these and other external data sources is centrally important as institutions make rapid advances relating to many types of relevant data.
- As the primary data repository for GTL-funded projects, the GKB must effectively relate data from such projects to the growing wealth of external data and analytical tools.
- Inferred data, which are the products of modeling activities, comparative analyses, and simulations, are expected to become increasingly important components of the GKB as its use among the scientific community grows.

Recommendations

- **Formal benchmarking.** Although no singularly integrative systems biology knowledgebase currently exists, there are excellent best-in-class particular databases from which to draw examples. The GKB should benchmark data and information standards and, in certain cases, systems interoperability against best-in-class relevant data repositories.
- **Realistic scope and expectations.** The GTL Knowledgebase is an ambitious endeavor, requiring active participation by scientists. For example, using knowledgebase data and services for scientific investigations and then feeding resultant data and knowledge back into the GKB, when coupled with existing data management resources, could constitute 10% to 20% of researchers' efforts. Because of

its scale, the GKB should be developed in phases with consideration to existing, established data management systems. The initial phase should support critical mission-relevant research and foundational science with well-defined needs and should provide resources to facilitate data access and ingestion. Implementation of these features would provide immediate value to the scientific community and would serve as a prototypic template for knowledgebase expansion.

- **Development of a database of critical information.** An extensive list of data entities has been compiled and itemized for capture in the GTL Knowledgebase (see Table 1.2. Critical Datasets and Data Types, beginning on p. 16). Selecting and prioritizing data types for GKB inclusion should be critical first steps in defining system requirements.
  - As part of its early activities, the GKB project should further develop a database of the identified entities and include data types, data volumes, and current format standards. Database development could be facilitated by surveying GTL principal investigators and establishing a website to collect survey data. The database should be reviewed regularly by the scientific community and perhaps be discussed and evaluated during the annual Genomics:GTL Contractor Grantee Workshop.
  - Once database development is under way, data entities should be prioritized in terms of importance to the GTL community and the challenges associated with incorporating the entities and establishing standards for each. The GKB project likely would have a practical limit determined by available funding, which thus will help define the scope of knowledgebase data.

Metadata

The term "metadata" refers to information about data, such as how an experiment was performed, which organism was studied, and what methods were used for data analysis. Because metadata allow scientists to reproduce results, capturing metadata is vital for meaningful knowledgebase use among the scientific community. In many cases, metadata will follow existing community guidelines of minimum standards and ontologies (i.e., structured, controlled vocabularies) set forth by community-driven efforts.

Findings

Metadata management is a core capability that will allow integration of data generated from different technologies. Critical to the success of the GTL Knowledgebase are the following key elements:

- Effective metadata management with common descriptions of data elements across multiple laboratories and investigators.
- Common descriptive language for integrating data from multiple investigators (currently a limiting factor).
- Metadata management tools that allow data generators to easily annotate and describe their data products and to extend metadata ontologies.

For GKB users to make comparisons among data and experimental results, each dataset from an environmental sample must be accompanied by metadata that provide contextual information. Such information would include, for example, the environment from which the sample was collected, methods used in collection and sample processing, types of analyses conducted on a given or nearby sample, and the overall sampling plan.

At the lowest level of information submitted to the GKB—that of data entities—data standards should define the minimum metadata requirements that must accompany all submissions.

- Where appropriate, these standards should be provided with templates and tools to help data generators harvest metadata and plan experiments before data are collected.
- Much (if not most) metadata should be collected before generating experimental data, but this process must be defined and prepared prior to conducting the experiment.

Recommendation

- Proper organization and capture of metadata are essential to providing the data context consumers require. Since collection of metadata often will be viewed as burdensome, effective standards describing required metadata must be established. Plans for enforcing these standards should be introduced early in the GKB project.

Data Handling

Archiving Raw Data

Findings

- Deposition of raw data from various technologies (e.g., imaging or mass spectrometry) into a comprehensive archive such as the GTL Knowledgebase is, in many cases, impractical. For each data-generating technology, the GKB project must determine the level of resultant data that could be captured. For cases in which raw-data capture is impossible or impractical, the knowledgebase should provide references to the sources of such data.
- Many technologies employed in GTL research generate large volumes of raw data that often are processed in complex analysis pipelines to produce final data products useful to the scientific community. For example, mass spectrometry–based proteomic analyses currently can generate terabytes of raw mass spectra that are processed to produce information about the peptides and proteins present in a biological sample.
- Since the GKB would not manage all raw data, the long-term availability of such information cannot be guaranteed. Researchers thus should devise their own local data-preservation strategies at the conclusion of a project that has generated raw data potentially important for future studies.

Recommendation

- The GTL Information and Data Sharing Policy (see Appendix 1, p. 59) should define the responsibilities of data generators in managing their raw data for the lifetime of their projects and in preserving data upon project completion.

Annotating Data

A central goal of the GTL Knowledgebase is supporting annotation, with the objective of achieving improved accuracy by removing inconsistencies and reducing ambiguity. Pursuing this objective will involve coordination with DOE's Joint Genome Institute (JGI; see sidebar, Analysis and Annotation at DOE's Joint Genome Institute, p. 23).

Analysis and Annotation at DOE's Joint Genome Institute

Analysis of DNA sequence at the Department of Energy's (DOE) Joint Genome Institute (JGI) is performed through a combination of centralized data processing and distributed data analysis capabilities. Extensive sequence annotations and analyses are generated for DOE's scientific community by JGI partner labs—including the Hudson Alpha Institute for Biotechnology, Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), Los Alamos National Laboratory (LANL), Oak Ridge National Laboratory (ORNL), and Pacific Northwest National Laboratory (PNNL). Through an extensive data and computing hardware infrastructure (550 terabytes and 1600 processors at the LBNL and LLNL Production Genomics Facilities alone), analysis of genomes from a diverse cross-section of the tree of life includes rich annotation, curation, and comparative genomics studies. Results of these studies—many of which are present in a wide variety of JGI public databases and high-profile scientific publications—underlie the value of genomics to the scientific community.

Annotation of plant genomes is carried out by DOE JGI's Computational Genomics group in collaboration with other researchers at the institute and elsewhere. State-of-the-art methods for gene prediction using ab initio, homology, and expressed sequence tag (EST) data are integrated to produce gene sets. Research efforts include applying new technology ESTs to improve gene predictions and incorporating small-RNA datasets. Comparative analysis of plant genomes is facilitated by Phytozome (http://www.phytozome.net), a hub for plant genomics.

Comprising more than 75% of DOE JGI sequencing capacities, the annotation of eukaryotic microbes is a significant component of the institute's informatic and analytical activities. Annotation and analysis of these genomes are exploited by more than 80% of JGI users and result in a considerable portion of the JGI publications in *Nature* and *Science*. The success of eukaryotic annotation is based on a community annotation program that is unique among genome-sequencing centers and highly valued by user communities.

Experience with prokaryotic genome annotation and comparative genomics is prevalent among DOE JGI partners and is most evident in teams at ORNL and the Production Genomics Facility. The flow of data—from production sequencing to assembly to finishing and annotation—is producing critical information on hundreds of new bacterial and archaeal genomes. Advancements in gene models also are being achieved through manual data curation, comparative and higher-quality functional annotations, and automated metagenome and metatranscriptome analyses. These capabilities are paving the way to new discoveries underpinning DOE missions in bioenergy, carbon cycling and biosequestration, and environmental remediation.

Furthermore, DOE JGI activities and resources significantly support the goals of DOE's Genomics:GTL program (GTL). When embraced, integrated into, and further expanded on by the GTL community, JGI capabilities can help achieve the program's vision to usher biology into a new era of systems sciences characterized by predictive understanding of the interactions of biological systems—both with their environment and each other.

Assigning function to genes and gene products is the classic concept of annotation. However, a substantially broader view is needed to describe the gradual refinement of assertions and inferences. As metabolic reconstructions, regulons, regulatory circuits, dynamic models, and phenotypic measurements and predictions are introduced into the GKB, the notion of annotation and maintenance of annotations extends significantly beyond the curation of protein function. Annotation also involves detection and removal of inconsistencies at higher levels in the biological hierarchy (for example, between phenotypic measurements and hypothesized metabolic reconstructions, such as the systematic approach used by *Shewanella* (see sidebar, *Shewanella* Knowledgebase, p. 24).

The concept of high-quality genomic annotation differs between eukaryotes and prokaryotes, largely because of the difficulties in accurately identifying eukaryotic genes (whether from plants or unicellular eukaryotes). At minimum, high-quality

Shewanella Knowledgebase

The *Shewanella* Knowledgebase is designed to provide a framework for investigators to share, combine, and analyze data. The first version of the database was released to investigators in 2007. Priorities of this knowledgebase include (1) improving coverage and support for "omic" data, such as expression arrays and proteomic, physiological, and biological information; (2) improving data linkage to key investigators and developing procedures to capture their data streams; (3) developing database support for multiple *Shewanella* species and strains by providing, for example, tools for comparative annotation and gene-function editing across multiple species; and (4) strengthening links to other data resources in the scientific community and to reference materials.

Capabilities for investigating multiple species have been implemented, including construction of several ShewCyc pathway databases. Tools for species comparison at pathway and genome levels are available and improving; regulatory data from numerous sources have been integrated. The knowledgebase also includes computational predictions of *Shewanella* regulatory elements collected from published literature and Internet resources, such as Rfam, RibEx, Tractor_DB, RegTransBase, BioCyc, and PromScan. This information was analyzed to identify a set of basic regulatory classes to present in the database. Such regulatory elements include translated coding sequences, DNA regulator-binding sites, sigma-factor binding sites, transcription units, promoters, regulons, stimulons, and RNA regulators. The latter encompasses a diverse class of regulators, including noncoding and small RNAs, different types of terminators, and riboswitches.

Current efforts focus on advancing tools and interfaces for cross-species annotation of multiple *Shewanella* species and on supporting the manual curation of 20 available genomes. Results of such activities will provide a foundation for experimental studies using a comparative approach that can be applied to essentially any group of model organisms. The *Shewanella* Knowledgebase is now equipped with a publication-mining system that includes a list of journals and other sources with links to references, authors, and related knowledgebase projects, as well as a text-search function. Procedures to maintain and update this library are in place.

Data Analysis

The knowledgebase user interface has many intuitive guides for exploring *Shewanella* experimental results. Multiple analytical modules perform one-on-one analysis across diverse biological datasets, supplemented by corresponding visualization capabilities at various data aggregation levels and biological contexts. The user interface also provides a unified set of integration analysis tools that support ShewCyc pathways and pathway-group categories. Future releases include KEGG pathways, TIGR roles, and GO ontologies for exploring data.

Data Visualization

Various data visualization tools display *Shewanella* experimental results. One such tool compares relative expression data at the gene level, while others compare averages or percentages of under- or overexpressed genes in a pathway or pathway group. These data viewers are cross-referenced to Pathway Tools software, which contains reference pathways for multiple *Shewanella* strains.

Web Portal

The *Shewanella* Knowledgebase Web portal is a data and knowledge integration environment enabling investigators to query across *Shewanella* datasets, link to *Shewanella* and other community resources, and visualize data in a cell-system context. The Web portal offers several ways to access and analyze data. Users can download data to their computers in the original format, and various data navigation features enable data exploration on the server. The knowledgebase's infrastructure is coupled with a powerful system-wide search feature that includes *Shewanella* data and publications. The user interface of the knowledgebase is built using a combination of Web 2.0 presentation layer technologies. Its Web portal is built with HTML 4.0, CSS, and Script.aculo.us javascript library. Generally, content is dynamically generated using Java server pages standard tag library.

prokaryotic annotations must include accurately identifying genes, assigning correct functional roles to gene products, and providing estimates of operons. For a growing number of prokaryotic genomes, reasonable estimates of metabolic networks and regulons also can be included in annotations.

In eukaryotes, the process of identifying genes and assigning meaningful descriptions to particular DNA segments (referred to as gene calling) is far more challenging than prokaryotic annotation. Much focus centers on overcoming this difficulty given that gene calls form the foundation for more advanced annotations. High-quality eukaryotic gene calls will need to incorporate cDNA data, including expressed sequence tags (ESTs), which—for some protists with high gene overlap—must be directional for effective use. These gene calls also should include sequence similarity and computational predictions based on the recognition of probable splice sites. As with prokaryotes, once reliable eukaryotic annotations have been established, the next goal is placing gene products in a larger context (e.g., within a metabolic pathway, complex, or nonmetabolic subsystem). Issues relating to cellular location and tissue specificity become important, but many are just beginning to be explored. Rapid progress is anticipated, however, as access to more genomes and expression data increases. This expanded accessibility will enhance opportunities for comparative analysis and will support, in particular, gene calling.

Findings

Accurate annotation of thousands of microbial genomes and a rapidly increasing number of plant genomes is a central goal of the GTL Knowledgebase. Achieving this goal would require the following:

- Incorporation of new empirical data and inferences.
- Detection of inconsistencies across a wide variety of data types.
- Logging of each inconsistency and the change introduced to correct it.
- Collection of such logs as a source of data to streamline annotation.

Recommendations

- The GTL Knowledgebase should support development of tools to refine and expand the concept of annotation. Doing so would establish consistency and remove ambiguity in assigning function across the hierarchy of biological components and systems—from DNA to proteins to pathways and networks.
- This process ultimately must be anchored in the characterization of phenotype, which includes environmental influences. Establishing protocols to control the annotation process will be essential to GKB viability.
- The GTL community also will need to agree on cultural strategies to move beyond "expert owner–based" curation.

Supporting Creation, Storage, and Maintenance of Inferred Data

Inferred data will be produced by comparative analysis, modeling and simulation. Since one central goal of the GKB is to support derivation and validation of inferred data, the project must include standards for defining provenance, attachment of appropriate metadata, and integration with experimental data.

Findings

- Accessing and using models to make and store inferences are emerging capabilities that increasingly more knowledgebase users will employ. The GTL Knowledgebase needs to support not only development of subsystems and models but also access to the models themselves.
- Researchers will extract data entities from the GKB, process them through analysis pipelines and workflows, and generate new entities that then will be submitted to the knowledgebase.
- The GKB needs to capture appropriate metadata and provenance information.

Recommendations

- The knowledgebase community should determine how to manage and control the introduction of inferred data and models.
- In addition to data, the GKB needs to provide software tools to the GTL community and link to other sites containing relevant software of interest. Such tools would include applications that facilitate capturing and recording inferred data and provenance.
- The knowledgebase should encourage open access to applications developed within the GKB framework, thereby connecting the GTL community.

Making Quality Control and Assurance Integral Parts of GKB Data Input, Annotation, and Modeling

Findings

- Quality control (QC) and assurance (QA) are critical aspects of incorporating new data into the GKB. Controlling quality occurs at the following two levels:
  - QA establishes processes to ensure the quality of the overall experimental program that generates data flowing into the knowledgebase.
  - QC screens data to reject faulty data.

These same quality processes can be applied effectively to inferred data and the mechanisms producing such data.

- Existing data providers may have their own protocols to control the quality of their data. However, sources often provide processed data to users as a "black box," meaning little relevant metadata are easily accessible. Such metadata provide information on how data have been processed and normalized, including the tools, parameter values, and versions of resources used.
- Another aspect of quality control involves changes to annotations and models. Such changes often result from supporting evidence discovered through a tedious, manual curation process. Often, however, neither the process itself nor the evidence is propagated in a coupled way with annotation changes.
- Managing data quality through QA processes and QC protocols and retaining and communicating information about such quality require the following:
  - Data quality and information must be established for all data products and should be communicated with all exchanges of such products. This approach

would enable users to efficiently complete evaluations of data products extracted for a specific use.

- Assembling and retaining quality information (e.g., QA processes and QC protocols) in a manner not overwhelming to data generators and consumers are significant topics to be resolved in the GKB design process.

Recommendations

The GTL Knowledgebase should provide a systematic approach for controlling the quality of data flowing into the GKB.

- Data must undergo appropriate QC protocols at their originating source. Although this responsibility for compliance lies with the source, the source also should provide metadata describing the data-processing workflow that can easily be queried, accessed, and summarized by GKB users.
- Establishing QC standards and protocols as they relate to annotation and inference must be the responsibility and an essential component of the GTL Knowledgebase. Adhering to GKB standards and implementing required protocols must be the responsibilities of data producers and users. Minimally, changes to data must be logged and detected conflicts updated and managed appropriately.
- GKB infrastructure should enable users to access and contribute to the evidence behind each act of curation. For example, an assertion of the presence of a given variant of a subsystem should be accompanied by users' ability to relate it directly to phenotypic measurements, expression data, and the functions associated with a set of proteins and to record this ensemble as evidence supporting an annotation change. The real power of data integration is manifest in these capabilities, which represent a major step forward for systems research. As such, they should be integral parts of the GKB process.
- Knowledgebase infrastructure also should provide mechanisms to quantify and record uncertainty (and dependencies) at all levels of analysis and propagate it in a consistent, probabilistic, and Bayesian manner. Doing so would involve, for example, characterizing and quantifying errors and biases across different metagenomic sequencing technologies.

Findings

- Curation is a long-lived process. Knowledgebase design should comprise methods for maintaining this process over the long term.
- Good stewardship of GKB information requires robust, ongoing curation accompanied by a mandatory independent assessment of knowledgebase data.
- In this context, curation includes—as an early step—tests to ensure data are complete, meet minimum reporting requirements, and have no obvious mistakes such as format problems and count errors.
- Testing for consistency will range from manual curation to automated checking.
- Documentation for inferred and assumed data entries must be rigorous.

Using Data Standards

Standards are a mechanism for capturing information in a form easily shared and integrated with other data or data types. Using data standards to capture data entities is the

foundation for comparative analysis and integration of information. These standards also define the minimum level of metadata that must be associated with a data entity. For some data, in the early stages of development, standards may be unavailable or even competing. Establishing and maintaining standards for incoming data thus should be essential components of the GTL Knowledgebase.

Findings

- Contributions of data and information to the GKB will depend on the adoption of existing standards or the development, when necessary, of new standards, and also on defined submission procedures.
- The user community must be involved in both the adoption of existing standards and the development of new standards, so that the standards enable reliable access to data and information deposited in the knowledgebase.
- One advantage of having standards is that they can be used to automate data extraction. For example, standardized metadata and provenance information can be obtained using automated tools and processes that remove the burden of data extraction from users.

The development of data standards often proceeds best within the auspices of international working bodies. The GTL program should establish a strong policy of adopting existing standards and seeking out emerging ones. This policy will have advantages for users, helping them develop internal data standards, and may even lead to the GTL Knowledgebase spearheading efforts to promote community-wide standards.

- The GTL Knowledgebase should use best practices to build consensus on internal consistency of data collection and management.
- Examples of existing and emerging standards to consider for the GKB are the Minimum Information about a Microarray Experiment (MIAME), Minimum Information about a Proteomics Experiment (MIAPE), and Minimum Information about a Genome Sequence (MIGS).
- Examples of organizations actively involved in developing data standards include the Proteomics Standards Initiative (PSI), Genomic Standards Consortium (GSC), Consortium for the Barcode of Life (CBoL), and International Census of Marine Microbes (ICoMM).

For models and algorithms, SBML (an extension of XML) and other markup languages have made substantial inroads in standardizing models in systems biology, at least in the context of dynamical systems modeling (Slepchenko et al. 2003; Hlavacek et al. 2006).

Recommendations

- The GTL Knowledgebase will need a standards committee to define minimum requirements, recommend adoption of community-developed standards, and initiate drafting of GKB standards as the need arises.
- This committee should be empowered to institute standards quickly and thus must establish a clear set of principles and decision-making processes to be followed. The following are operational examples:

- The GTL Knowledgebase must include enough information for skilled practitioners to reproduce any available data. Achieving this goal requires adopting and developing appropriate schemas.
- Data requirements need to address uncertainty propagation, so that all types of output data have a confidence limit, confidence interval, or other uncertainty field.
- Data input tools should be developed to ensure a model or algorithm meets all minimum requirements prior to submission to the knowledgebase.