Data Integration

Data integration is a feature that clearly expands the role of the GTL Knowledgebase (GKB) beyond an archive to a dynamic systems biology resource for progressively increasing scientific understanding. The GKB is envisioned to contain data, such as those described below, on thousands of complete genomes and thousands of metagenomic and transcriptomic samples.

- Data for each complete bacterial and archaeal genome should include estimates of gene function and regulons as well as detailed metabolic reconstructions.
- For each metagenomic sample, the GKB should provide estimates of microbial population and data on metabolic potential.
- For each of the more complex eukaryotic genomes, such as those of plants, protists (including algae), and fungi, data should include detailed estimates of genes and metabolic reconstructions for different tissues (e.g., root versus stem). These data also must cover various stages of development (e.g., in meristems and seeds), some of which have been extremely well elucidated at the molecular level.

To develop accurate and predictive models, the GKB needs to capture additional data for a limited (but increasing) number of organisms. These data include phenotypic, metabolic, expression, protein-protein, and protein-DNA measurements. Incorporating such data in the knowledgebase would advance the development of stoichiometric and regulatory models, leading to improvements in metabolic reconstructions that would be propagated to all genomes in the GKB.

The GTL Knowledgebase should integrate genomic, metabolic, regulatory, and phenotypic estimates under continual revision. Integrating such data would require ongoing curation by the GKB to ensure increased data consistency among a growing body of measurements, which would enhance the predictive capability of models and provide a new resource for the study of organisms.

Core applications within DOE are intended to drive the knowledgebase initiative. These applications typically revolve around macro processes (e.g., the carbon and nitrogen cycles). A key GKB requirement thus would be to couple these flows with modeling of individual organisms and communities of organisms. Seeking insight relating to processes that operate at widely separated temporal and spatial scales will be extremely challenging. Although many studies develop hypotheses for organisms' potential functional roles and interactions with their environment, few studies have tested such hypotheses. Because of system complexity, obtaining these measurements is a major challenge. Nonetheless, empirically determining process rates in complex systems is essential and should include appropriate experimental scaling that allows measured rates to be related to the genetic and regulatory bases for processes.

An example of research for which integrated process and activity rates are needed involves photosynthesis by marine microbes. Little is known about the rates at which different organisms (e.g., cyanobacteria versus the wide variety of protistan primary producers) take up CO_2 or the impact of competition on these organisms' performance. Similarly, much remains to be learned about the subsequent fate of photosynthetically fixed carbon as it is respired by organisms or exported to the deep ocean for long-term storage. Some of these microbes (and consequently their carbon) can descend to the deep ocean on their own; others must be consumed by

larger organisms to sink; and still others are lysed by viruses. As a result, little carbon reaches ocean depths because some is lost at each step of the way as organisms consume or degrade organic carbon and remineralize it to CO_2 through respiration. These mechanisms play central roles in oceanic carbon flow and sequestration. Hence, the GKB should capture and integrate measurements of these activities. A systems biology approach would facilitate such integration and thus enable predictive modeling (Azam and Worden 2004).

To support both the development and use of detailed models to infer properties of organisms and resolve model ambiguities, the GTL Knowledgebase should incorporate increasingly diverse data types. This growing body of measurements should be directly integrated into GKB estimates based on genomic data. To be effective, expression data should be related to metabolic reconstructions and estimates of regulons, and these relationships must undergo continual curation.

As a secondary service, the GTL Knowledgebase also should support archiving a diverse collection of data types. For example, GKB structure must allow heterogeneous structural and dynamic process data to be scalably and consistently captured, reconciled, and related, as well as retrievably stored. The knowledgebase also should include tools that will enable extraction, integration, and modeling to draw inferences and make predictions that can be incorporated into the GKB as additional data.

Defining the Scope of Data Integration

The GKB's data integration effort should have the capacity to access and interrelate thousands of prokaryotic genomes, hundreds of eukaryotic genomes, and thousands of environmental samples. This service also should provide capabilities for integrating associated phenotypes, other high-throughput data, and metadata. Furthermore, data integration needs to support effective comparative analyses, modeling and simulation, and inference of data hierarchies based on an increasing wealth of information.

This ambitious GKB integration strategy poses many challenges, two of which are particularly significant: data quantity and incorporation of numerous data types.

1. The sheer volume of data is imposing and includes both genomic and other information required to support comparative analysis and interpretation. A large, diverse collection of reference genomes clearly will form the essential framework needed to support more focused investigation of the biological communities relevant to DOE mission-application areas.
    a. A primary knowledgebase requirement is the ability to integrate an exponentially growing body of data and to provide reasonably accurate initial annotations, ongoing refinements, metabolic reconstructions, and estimates of regulons.
    b. Inclusion of data and information for thousands of metagenomic samples ultimately may constitute the bulk of the GKB's computational load. The rapid accumulation of a collection of well-annotated reference genomes will represent an essential asset supporting the interpretation of more complex metagenomic data. Central to the overall goals of the knowledgebase is the need to absorb this quickly expanding flow of new data into a framework offering tools for convenient comparative analysis. Similar to the complexity of metagenomes will be the integration of higher plants with distinct tissue types and their linkage to particular environmental locales and responses.

2. The second challenge lies in using the planned integration to support extensive incorporation and reconciliation of numerous types of data. These data range from genes and estimated gene products to metabolic reconstructions and models of regulatory circuitry.
   a. In addition to integrating large numbers of reasonably well annotated genomes, another GKB objective would be to select a limited set of organisms with specific relevance to DOE missions and to develop predictive models of them.
   b. Developing these models will impose consistency among the models, metabolic reconstructions, and experimental data that will form the foundations for biological research in this century.
   c. However, imposing consistency on these elements necessarily implies the ability to make and maintain numerous changes to widely shared and deeply interdependent data.

Today's architectures are capable of supporting the data structures and integrations envisioned for the GTL Knowledgebase. Existing data systems clearly support the feasibility and utility of an ambitious integration effort. None, however, currently addresses the opportunities introduced by recent advances in both microbial modeling and the ability to obtain and analyze metagenomic sequences.

Core Requirements for Data Integration

Improving the Quality of Data Annotation through Continuous, Semiautomatic Curation

Findings

Incorporating data annotations at various scales and resolutions is one objective of the envisioned GTL Knowledgebase. Achieving this goal would require addressing several challenges associated with the expanding scope of annotation.

- Assigning function to genes and gene products is the classic view of annotation.
- A substantially broader concept of this process is emerging, however, in the context of systems biology. This wider view includes annotated models of metabolic pathways and regulons, protein interactions and interaction networks, and three-dimensional protein structures.
- Many annotations—computationally derived from uncertain, noisy, incomplete, and complex data—contain various inconsistencies, ambiguities, and gaps in knowledge.

The infrastructure of GKB's data integration service presents a unique opportunity for improving annotation quality.

- Increasingly, research groups are successfully using integrative approaches to significantly improve the quality of data annotation. For example, the *Shewanella* Federation has demonstrated a systematic approach to detect inconsistencies between phenotypic measurements and hypothesized metabolic reconstructions (see section, Illustration of Use Case Scenario 1: Integrated Approach to Reconstruction of Metabolic and Transcriptional Regulatory Networks in Bacteria, in Appendix 2, p. 74).

- Similarly, incorporating information on three-dimensional protein structure has been highly valuable in annotating hypothetical genes (i.e., those without functional assignments) identified by genomics-based annotation pipelines. For example, of the unannotated proteins in *Halobacter* NRC-1, Bonneau et al. (2004) assigned functions to about half and reconstructed metabolic pathways by combining structural-functional predictions from the Robetta server (see sidebar, Example Analysis and Integration, p. 35) with genomic-context data and a variety of experimental information.
- While highly promising, many of these approaches significantly rely on tedious manual curation.

Recommendations

The GTL Knowledgebase should provide semiautomatic tools to expert curators to help them more efficiently improve the quality of annotations. These tools would support several activities.

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

As the GTL Knowledgebase incorporates increasingly higher levels of data—such as metabolic reconstructions, regulons, regulatory circuits, dynamic models, and phenotypic information—the concept of GKB annotation would need to expand to encompass and maintain these entities. This expansion would require creating various resources and protocols to improve data quality, for example, the following:

- Tools to support consistency and improve confidence, particularly as the scope of data widens.
- Mechanisms to efficiently link existing knowledgebase annotations to emerging and newly published experimental evidence (e.g., from mutagenesis studies or expression profiles) that refines or confirms such annotations.
- Protocols to control annotation.

Facilitating Data Integration through Standards, Controlled Vocabularies, and Ontologies

Findings

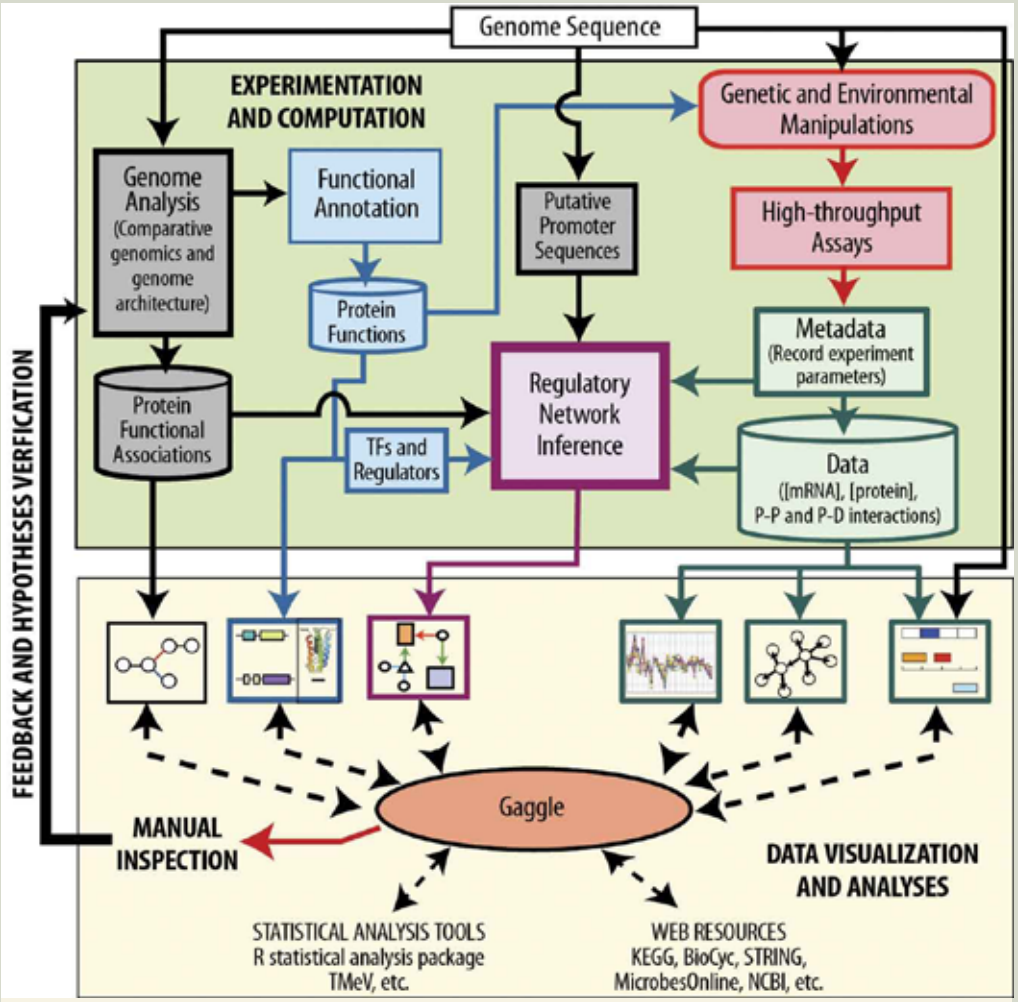
Data integration and model development in systems biology largely are hampered by the lack of semantically consistent naming conventions.

- Although different annotation systems depend on each other, they often use inconsistent definitions, resulting in decreased quality of the systems and their annotations. For example, genome annotation pipelines may use gene-function definitions inconsistent with the controlled vocabularies used by systems that annotate metabolic pathways.
- Such semantic ambiguity and inconsistency probably lead to holes in reconstructed metabolic pathways.

Example Analysis and Integration: The Process of Generating Models of Metabolic and Regulatory Networks

The ability to generate accurate and predictive models of organisms' metabolic and regulatory circuitries represents a substantial advancement in systems biology. Developing such models may be viewed as a process that produces, as a by-product, consistency among protein functions, metabolic reconstructions, and derived models. The need is to have massive data-driven and falsifiable (testable) hypotheses. The "trivial" underlying hypothesis is, "Can a network model represent the available datasets?" The ultimate driver of these models is the need to generate new predicted hypotheses that can be tested in silico and in vivo. Deriving these models requires the following data:

- Annotated genomes (including genes, transcription start sites, and operons).
- Detailed metabolic and regulatory reconstructions.
- Initial estimates of regulons.
- A list of binary associations between proteins, reflecting existing data on protein-protein interactions, relationships inferred from phylogenetic profiles, and co-occurrence information. (The number of data sources providing evidence of protein associations clearly will increase over time.)
- Estimates of transcription factors.



Example Analysis and Integration: The Process of Generating Models of Metabolic and Regulatory Networks. [Source: Adapted with permission from Elsevier. From Bonneau, R., N. Baliga, et al. 2007. "A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell," *Cell* **131**(7),1354–65 (http://www.sciencedirect.com/science/journal/00928674).]

Generating a model of an organism's regulatory circuitry involves designing manipulative experiments that induce genetic or environmental perturbations and recording measurements of the resultant changes through high-throughput assays. These measurements include (at minimum) expression data, protein-DNA binding, protein-protein interactions, and protein modifications. Each perturbation is described in a controlled vocabulary, measurements are recorded and normalized, and the resulting data pairs (i.e., the induced perturbation coupled with the observed outcome) become input for an inference process. This process involves ever-improving algorithms that use pair sets to infer aspects of an organism's regulatory circuitry. Producing an accurate model then requires (1) iteratively examining the derived regulatory circuitry; (2) reconciling it with known phenotypic data; (3) gradually understanding the sources of inconsistency; and (4) changing asserted protein function, metabolic reconstructions, and proposed circuitry to reconcile inconsistencies.

Several existing efforts support semantic normalization and standardization.

- The Open Biomedical Ontologies (OBO) Foundry (http://www.obofoundry.org) has emerged as a framework for community-ontology development that conforms to a set of principles and best practices (see Table 3.1. Open Biomedical Ontologies Foundry, p. 37).
- The Foundry and Gene Ontology (GO; http://www.geneontology.org) include many emerging GTL-relevant ontologies, such as those for plants (PO), environment (ENVO), phenotype (PATO), chemicals (ChEBI), proteins (PRO), and metagenomes (MIMS).

Recommendations

To promote interoperability and facilitate data integration, the GTL Knowledgebase should identify, adopt, and develop common standards, controlled vocabularies, and ontologies.

- Annotation systems that semiautomatically populate their inferences into GKB infrastructure should clearly define and post their underlying ontologies.
- These interdependent annotation systems should be part of a consolidated effort to conduct semantic consistency checks. Proper semantic mapping functions, synonymous lists, and controlled vocabularies need to accompany data and annotations released from and propagated into the GKB.
- For GTL researchers who are not currently served by a vocabulary or associated ontology that describes their area of research, new ontologies might need to be defined or the scope of existing ontologies expanded with input from the user community.

Increasing the Efficacy of Complex Queries from Integrated Data

Findings

Public query engines—such as NCBI's Basic Local Alignment Search Tool (BLAST), which enables analysis of genomic sequences—have significantly increased the throughput of data extraction and made this information routinely accessible to all types of users, including computationally skilled scientists and lay people.

- However, performing similar queries of integrated systems biology data still is in its infancy.
- Despite the potential to enable new discoveries, using integrative systems biology approaches to extract information of interest is a painstaking and time-consuming task that only a few scientists endeavor on their own.

The heterogeneity and complexity of data to be integrated into the GTL Knowledgebase will be substantial.

- Effective use of these data thus will require the GKB to support highly diverse advanced queries that few existing databases have encountered. Box 3.1, Typical Complex Queries, p. 38, lists several examples of such queries.
- Supporting these types of advanced integrative queries in a user-friendly, automatic, and routine manner comparable with BLAST will transform future systems biology studies.

Table 3.1 Open Biomedical Ontologies (OBO) Foundry*

Granularity	Continuant				Concurrent
	Independent		Dependent		
Organ and organism	**Organism** (NCBI taxonomy)	**Anatomical entity** (FMA, CARO)	**Organ function** (FMP, CPRO)	**Phenotypic quality** (PaTO)	**Organism-level process** (GO)
Cell and cellular component	**Cell** (CL)	**Cellular component** (FMA, GO)	**Cellular function** (GO)		**Cellular process** (GO)
Molecule	**Molecule** (ChEBI, SO, RNAO, PRO)		**Molecular function** (GO)		**Molecular process** (GO)

*Aiming to create a suite of orthogonal interoperable reference ontologies to support integration and analysis of biological data, the OBO Foundry ontologies are organized along two dimensions: (1) granularity (from molecules to populations of organisms) and (2) relation to time (a distinction between entities that undergo changes through time and the entities—processes—that *are* such changes). [Source: Adapted by permission from Macmillan Publishers Ltd. From Smith, B., et al. 2007. "The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration," *Nature Biotechnology* **25**(11), 1251–55 (http://www.nature.com/nbt/).]

Recommendations

The GKB should provide easy-to-use interfaces to significantly increase the throughput of predictive inferences resulting from queries of integrative data by lay users.

The knowledgebase should support both "vertical" and "horizontal" queries.

- Vertical queries span data levels (e.g., from correlating climate data and habitats to genes found in different samples).
- Horizontal queries associate equivalent data entities across species, samples, or habitats (e.g., homologous genes between species, community composition across samples, and abundance or enrichment of metabolic pathways across habitats).

So-called canned queries in the GKB should support systems biology modeling tasks performed by a broad community of users. Both generic and model-specific information need to be automatically retrieved in response to relatively simple inputs provided by users. For example, when a user selects an organism to query, the knowledgebase should automatically compute and retrieve (in a structured and downloadable format) relevant information for the specified metabolic model of interest. This information should include the following components:

- A list of proteins (e.g., enzymes and transporters), inferred reactions, and metabolites.
- All associated information and features, including functional assignments (from various sources) and evidence; association with protein families (e.g., phylogenetic profiles); multiple alignments and phylogenetic trees for each family; domains, motifs, and structural features (known or predicted); genomic context (e.g., operons and regulons); functional context (e.g., associated pathways and subsystems); gene expression data (users may choose from integrated or uploaded datasets); proteomic data; associated reactions and metabolites; and other types of data relating to specific genes.
- Clusters (lists) of functionally coupled genes (e.g., stimulons) with a detailed correlational analysis (e.g., linkages between gene expression and pathways or between gene expression and protein levels).

Box 3.1

Typical Complex Queries

1. Which genes in genome X are known to be essential, and which are predicted to be essential? Display the differences.
2. Which of the genes in X have relatively solid annotations? Which are less reliable but have estimates of function, and which are completely uncharacterized?
3. Given genome X, what is a working estimate of the regulons in X?
4. In genome X, list—in a convenient format—the sets of genes believed to be coregulated. Then, given microarray MA, list the sets of expressed genes that agree with existing estimates of regulons and display discrepancies.
5. Which functional roles are used in model M of genome X but are not yet mapped to any specific gene or genes in X?
6. Does model M predict that organism X can sustain growth with just Y as a carbon source based on the organism's genome and other data?
7. Given the phenotype of a metabolic pathway, which genes and gene products are probably active in the steps of the pathway, and which are likely rate limiting?
8. Which transporters are required by model M? Which are mapped to specific genes, and which have been supported by experimental evidence but have not yet been mapped to specific genes?
9. Given metagenomic sample S, what is the existing best estimate of the microbial population (i.e., which operational taxonomic units make up the sample and in what relative abundances)?
10. Given two metagenomic samples, what distinguishes them? Similarly, given two sets of metagenomic samples, what distinguishes one set from the other? Given a set of genomes, which genes are common, and which distinguish one genome from the other?
11. Given a set of genomes, which subsystems are common, and which distinguish one from the other?
12. Given two sets of genomes, G1 and G2, which subsystems distinguish G1 from G2 genomes?
13. Given two different models, M1 and M2, which experimental measurement would help differentiate models? Alternatively, list differing phenotypic predictions based on M1 and M2.
14. Given a dataset, do the data have biological or technical replicates?
15. Given a particular gene, what are all its associated annotations?
16. Integrate and compare proteomic and transcriptomic datasets for the same experimental condition.
17. Conduct visualization analyses of data or model simulations (e.g., onto pathway maps).
18. Query metadata and conduct fuzzy matching of such data.
19. Which proteins have been observed for a particular organism and also across all organisms? How do protein profiles correlate with phylogenetic differences? Determine conservation of post-translational modification across organisms.
20. Obtain upstream sequences for a coding region.
21. What is the location of a protein under a specific condition?
22. Determine conservation of regulation across species.
23. Horizontal gene transfer: Which genes have been horizontally transferred?
24. How many genes relating to photosynthesis and nitrogen fixation are present in metagenomic data?

Streamlining GKB Incorporation of Dynamically Changing Biological Data

Findings

The GTL Knowledgebase should seamlessly incorporate new classes of data and models to meet the demands arising from continuous advances in both the experimental technologies producing data and the informatic methods deriving predictions from such data.

- Knowledgebase integration would involve inputs from two basic categories of data sources.
  - Projects producing initially processed data.
  - Curated information from other public data resources (e.g., UniProt, KEGG, NCBI, and topically oriented databases).

Critical to GKB integration efforts, the first category would be responsible for initial processing of experimental data, which should be normalized and condensed into a form directly incorporable into the knowledgebase.

- The most obvious example of such processing is genome sequence data, which should be incorporated into the GKB as assembled contigs, not raw reads.
- Similarly, microarray data should be normalized by their sources and accompanied by descriptions of the experiments from which they were derived; such inputs would not include images.
- To support modeling efforts, phenotypic data also should be condensed into a form suitable for GKB integration.

Enabling Integrative Capabilities for Data Analysis and Visualization

Findings

Although significant progress has been made in developing bioinformatic tools that derive predictions from individual data types, there is an emerging and critical need for tools that support comparative analysis and visualization of the results. The significance of advances in interface conception and implementation are obvious. Comparative genomic tools such as those available through KEGG, the SEED, the Expert Protein Analysis System (ExPASy), or NCBI provide good examples of integrated and easily accessible capabilities. However, while the ability to visualize data in these resources has advanced, it is far from optimal.

The variety of genomic and comparative genomic tools can be attributed to the availability of such resources on the Web. However, similar capabilities for quantitative proteomics, metabolomics, or transcriptomics are just emerging. Moreover, these tools typically are presented as stand-alone applications, making their adoption by the biological community problematic.

Recommendations

The utility of the GTL Knowledgebase would be enhanced substantially by the availability of easy-to-use, broadly accessible, and predictive visual analytical environments. GKB infrastructure should make visual analytics an integral component of the core tool set for data integration. The following three core activities should be supported to achieve this objective:

- **Adapting existing data analysis tools for ease of use and accessibility.** GKB should pursue a systematic effort to make the evolving set of analysis tools for various omic data accessible through user-friendly knowledgebase portals.
- **Extending GKB infrastructure to add new tools for data analysis and visualization.** The GKB community should develop guidelines allowing developers of data analysis tools to contribute to GKB infrastructure through easy-to-use plug-in interfaces. While Web services offer a mechanism for providing third-party tools to the scientific community, their inherent limitations make it desirable that such tools be easily downloadable and readily accessible as integral parts of GKB infrastructure.
- **Enriching GKB capabilities for integrative data analysis and visualization.** The GKB community should develop new comparative and integrative data analysis and visualization tools for creating more predictive models. Specifically needed are the development of and increased accessibility to tools for relating quantitative proteomic and transcriptomic data, finding conserved evolutionary network motifs, and linking transcriptional and metabolic network models.

Provenance

Researchers conducting laboratory experiments routinely control and record all aspects of their experimentation environment and manipulations for description in scientific publications. Similarly, biologists using the GTL Knowledgebase should be able to capture details of the in silico experimental process used to derive their results. These details, known as provenance data, include experimental information about the datasets used, the software models and tools that processed the data, and the resultant information that eventually is added to the knowledgebase. Provenance data will allow biologists not only to visualize the experimental processes used to reach a particular conclusion but also to potentially reproduce the results of a specific experiment.

Capturing provenance information in large-scale data management repositories can lead to an exponential explosion of GKB data. Knowledgebase designers thus should devise a novel, scalable strategy for provenance capture and visualization that meets specific needs of the GTL community. Various potential solutions exist (see Fig. 3.1. Example of Provenance Browser in Taverna, p. 41), and GKB planners need to evaluate them to design an architectural component that satisfies GTL requirements.

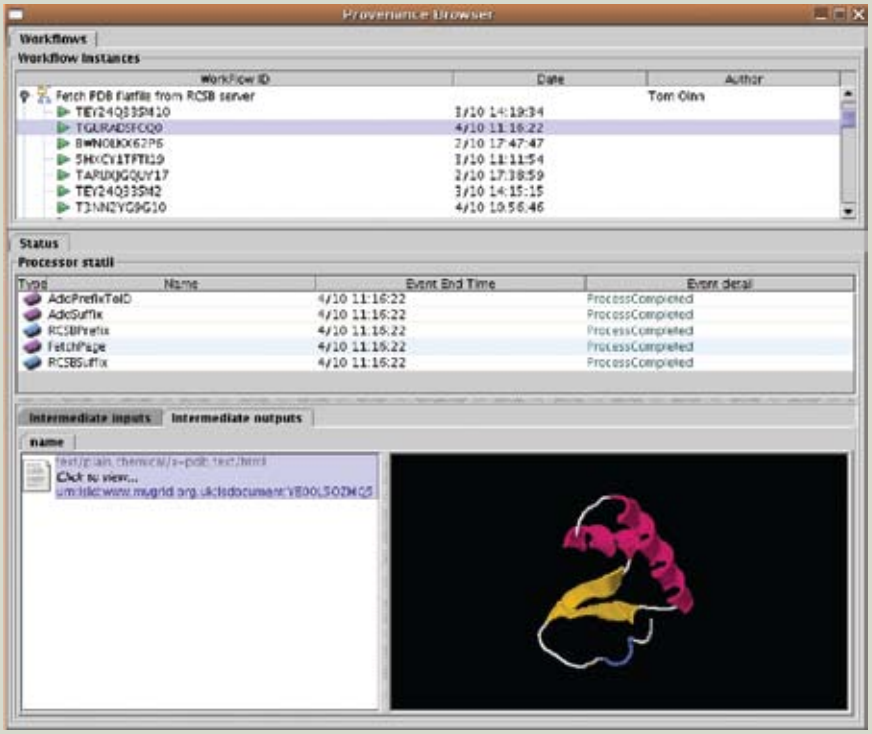


Fig. 3.1. Example of Provenance Browser in Taverna (http://www.taverna.org.uk). This feature provides a way for biologists to view the origins of data.