Report on the 2nd International Workshop on Data Integration in the Life Sciences (DILS'05)

Amarnath Gupta
gupta@sdsc.edu

Bertram Ludäscher
ludaesch@ucdavis.edu

Louiqa Raschid
louiqa@umiacs.umd.edu

Following a successful first international workshop on *Data Integration in the Life Sciences* (DILS) in 2004 in Leipzig, Germany [2], the second DILS workshop was held in San Diego, California from July 20–22, 2005 [1]. This new workshop series reflects the strong interest in an annual event, bringing together biologists and computer scientists conducting research in life science data management and integration, in domains and applications such as molecular biology, biodiversity, drug discovery and personalized medical research. The program committee accepted 15 long papers and 5 short papers from 42 submissions. In addition, DILS also featured 7 posters, 2 keynotes, several reports on ongoing research activities in academia and industry, and a panel on *The Electronic Health Record of the Future: Incorporating Molecular Information*, organized by the AMIA Genomics Working Group.

Setting the theme of the workshop, Shankar Subramaniam, Professor of Bioengineering and Chemistry at UCSD, noted in his keynote *Challenges in Biological Data Integration in the Post-Genome Sequence Era* that "the standard paradigm in biology that deals with '*hypothesis to experimentation (low throughput data) to models*' is being gradually replaced by '*data to hypothesis to models*' and '*experimentation to more data and models*'." Given the complexity and incompleteness of data in biology, he called for robust data repositories that allow interoperable navigation, query and analysis across diverse data, a plug-and-play environment that will facilitate seamless interplay of tools and data and versatile biologist-friendly user interfaces. In the second keynote, *Curated Databases*, Peter Buneman noted that biological data is often created and maintained at a high cost involving extensive human curation, and he explored the relationship between database research and data curation. He identified challenges in annotation, provenance, archiving, publishing and security and emphasized the goal of making sure that curated data is accessible and understandable to future biologists.

The twenty accepted research papers covered a wide spectrum of theoretical and practical research issues including user interfaces, analysis tools, scientific/clinical workflows, ontologies, and data integration techniques. Below we highlight some of the contributions.

User Interfaces and Analysis Tools. S. Cohen-Boulakia, S. Davidson and C. Froidevaux describe *A User-Centric Framework for Accessing Biological Sources and Tools*, based on an analysis of scientists' needs. They presented the BioGuide system that helps biologists choose among a variety of resources, taking into account *queries*, *preferences* and *strategies*. Information resources are connected in a graph; then information requests are processed by finding appropriate paths in this graph using the preferences and strategies. The research studies the semantics of this user-centric framework and its complexity.

The *Hybrid Integration of Molecular-Biological Annotation Data* from public sources is studied by T. Kirsten, H.-H. Do, C. Körner, and E. Rahm. Annotation data is integrated in a virtual (*i.e.*, mediator) approach, coupling SRS with a warehouse of expression data. Their system exploits correspondences between molecular-biological objects in the integration and supports functional analysis using annotations from GeneOntology, Locuslink and Ensembl.

The *BioNavigation* system by Z. Lacroix *et al.* allows scientists to express queries over semantic concepts and labeled relationships. Their ESearch algorithm generates all possible source paths following links and paths through data resources, and then ranks these source paths using source metadata metrics such as the number of result objects in the target, or the evaluation cost of a source path.

N. Tran *et al.* present a BioSigNet-RR, a *Knowledge-Based Integrative Framework for Hypothesis Formation in Biochemical Networks*. Hypothesis formation is modeled as a reasoning process that attempts to explain all concepts that are not completely entailed by the current knowledge of the system, by logically extending the knowledge base. It has been applied to the problem of modeling typically incomplete knowledge about biochemical networks.

In *BioLog: A Browser Based Collaboration and Resource Navigation Assistant*, P. Singh *et al.* describe a system for biomedical researchers that archives access patterns of scientists as they browse PubMed and also allows users to browse group specific archives. It makes recommendations using gene-to-gene, abstract-to-abstract and user-to-user relevance networks that use a combination of collaborative filtering and content based filtering techniques.

Challenging Applications. O. Fiehn *et al.* describe a system for *Setup and Annotation of Metabolomic Experiments*, noting that long-term reusability of metabolomic data needs correct metabolite annotation and consistent biological classification, and that "data without metadata are junk". The system attempts to produce automatic annotation of the results of mass spectrometry experiments, and feeds the annotated information back to a LIMS system for verification and use by scientists.

J. Kang *et al.* address the problem of *Integrating Heterogeneous Microarray Data Sources using Correlation Signatures*. The paper develops a technique to compute the *signature vector* of a gene with other genes in a microarray experiment, and then perform a statistical correlation among these signatures. They also propose an OLAP-inspired structure called the *gene signature cube* to perform global analysis of multiple experiments.

S. Jablonski *et al.* report on *Building a Generic Platform for Medical Screening Applications based on Domain Specific Modeling and Process Orientation*. They propose a process-oriented approach and system for distributed screenings for the early detection and diagnosis of glaucoma disease.

Tools for Legacy Data and Data Integration Solutions. Wrappers are crucial to the task of accessing legacy biological data sources and automatic wrapper generation is an ongoing challenge. K. Sinha *et al.* report on a tool to learn the layout and generate wrappers for flat-file datasets. The tool uses a number of heuristics to determine file delimiters, semiautomatically eliminate incorrectly formulated files and create a parser for the data files. The authors report on experiments with Swissprot, Genbank and Pfam data.

Associating a unique key with an object is crucial to effective data management. G. Neglur *et al.* report on applying the "Unique SMILES" notation to chemical structures and show that the prevalent algorithm relies on symmetry properties of graphs and will fail for certain graphs. Their paper on *Assigning Unique Keys to Chemical Compounds for Data Integration* modifies the algorithm by introducing a step where a "tie-breaking" is performed for formulas having the same tree-expansion. They demonstrate that it creates unique SMILE codes for all their counterexamples.

E. Guérin *et al.* describe GEDAW, an object-oriented gene expression data warehouse that integrates public data on expressed genes, experiment data from microarrays and other relevant data resources, *e.g.*, ontologies such as GO and UMLS. The objective is to combine biological mechanisms and medical knowledge with experiment data, and the project addresses issues in both semantic data integration and analysis of the integrated resources.

The AutoMed tool designed with a hypergraph data model for warehouse-based integration has been in existence for some time. M .Maibaum *et al.* discuss the adaptation of this system to biological data integration. Specifically, the paper discusses the issue of information integration when part of the data to be integrated are structured, and another part has to be clustered based on data values to find equivalent entities.

M. Mahoui *et al.* explore the concept of semantic correspondence for biological objects that are actually similar but are differently represented in their respective data sources. The paper puts forward the notions of *degree of semantic correspondence* and *cardinality of semantic correspondence* as two measures of such correspondence and shows how they can be used for information integration across heterogeneous sources.

There is an increasing number of internet-accessible services for providing, comparing and transforming biological data. A. Ngu *et al.* consider the problem of automatically classifying these sources based on their output and the kind of user interaction they need. They develop the notion of a *service class descriptors*, that captures metadata that would be adequate to describe these services, and present two techniques to automatically construct the service class descriptors, and discuss their applicability and performance.

Integrating private data with large public data banks can be expensive if they are to provide absolute privacy. R. Pon *et al.* provide a technique based on the semi-join operation; it exploits the uncertainty caused by hash collisions and injects noise. This method provides an upper bound on the amount of privacy loss during data integration.

Ontologies and Data Integration. As the number and sizes of ontologies increase, there is a growing need to query repositories of large ontologies. S. Trißl *et al.* consider the problem of querying ontologies that are structured like trees and DAGs, and offers indexing schemes for both cases. They analyze the performance of queries using these index structures for a number of common query patterns in biological ontologies.

Taxonomic identification is key to research in several domains. While biologists identify their data with scientific names, this is insufficient to unambiguously distinguish taxon concepts. A model for the representation of diverse taxonomic concepts is presented by J. Kennedy *et al.*

The INDUS system by D. Caragea *et al.* employs ontologies and inter-ontology mappings to provide a user view of multiple heterogeneous sources that reflects the user's preferred ontology. The authors present an experiment where INDUS is used to learn probabilistic models to predict GO functional classifications; mappings such as EC2GO and MIPS2GO are used in this task.

Ontologies are expected to have significant impact on research in ecology. STELLA is a widely adopted software tool for ecological modeling. C.M. Keet presents a formal mapping between STELLA and ontology elements. This

mapping simplifies ontology development and can support semi-automated techniques for ontology development.

P. Mork *et al*. describe *The Multiple Roles of Ontologies in the BioMediator Data Integration System*. Queries are expressed against concepts in the ontologies; at the same time they may serve as data sources. A system ontology provides metadata about sources, *e.g.*, how often it is updated. Finally, ontologies are used to describe the mapping from data sources to the BioMediator mediated schema.

DILS'06. The third international DILS workshop will be held at European Bioinformatics Institute (EBI) at Hixton, south of Cambridge, UK from July 20–22, 2006; for details see `www2.informatik.hu-berlin.de/dils2006`.

References

[1] B. Ludäscher and L. Raschid, editors. *Data Integration in the Life Sciences, Second InternationalWorkshop, DILS 2005, San Diego, CA, USA, July 20-22, 2005, Proceedings*, volume 3615 of *Lecture Notes in Computer Science*. Springer, 2005.

[2] E. Rahm, editor. *Data Integration in the Life Sciences, First International Workshop, DILS 2004, Leipzig, Germany, March 25-26, 2004, Proceedings*, volume 2994 of *Lecture Notes in Computer Science*. Springer, 2004.