



# **4<sup>th</sup> International Workshop on Data Integration in the Life Sciences**

## **Poster Abstracts**

**June 27-29, 2007**

**University of Pennsylvania  
Philadelphia, PA, USA**

### **Sponsored by**

The School of Engineering and Applied Science at the University of Pennsylvania  
The Penn Genomics Institute  
Microsoft Research  
The Penn Center for Bioinformatics





## Welcome

We are delighted to welcome you to Philadelphia for the 4<sup>th</sup> International Workshop on Data Integration in the Life Sciences. DILS 2007 is the fourth in a workshop series that aims at fostering discussion, exchange, and innovation in research and development in the areas of data integration and data management for the Life Science.

Over the next three days, we will hear from 28 speakers covering a wide spectrum of theoretical and practical issues in the domain of data integration and data management for the life science, including implementation of systems or experience with systems in practice, scientific workflows, annotation in data integration, mapping and matching techniques, and modeling of Life Science data. DILS 2007 also features two keynote talks by Kenneth H. Buetow, National Cancer Institute, and Junhyong Kim, University of Pennsylvania.

We would like to express our gratitude to our sponsors: the School of Engineering and Applied Science at the University of Pennsylvania, the Penn Genomics Institute, Microsoft Research, and the Penn Center for Bioinformatics.

We hope you enjoy your visit to the Penn Campus and surrounding areas, and find the meeting both enjoyable and fruitful for your research.

Susan Davidson, General Chair

Val Tannen, Program Committee co-Chair

Chris Stoeckert, Program Committee co-Chair

Sarah Cohen-Boulakia, Proceeding co-Editor

**Poster 1: Jump-Starting Data Integration in Biomedicine**  
*(Selected for oral presentation)*

*Peter Mork, Jean Stanford, Len Seligman, Jeff Hoyt and Ken Smith*

*The MITRE Corporation*  
*{pmork, jstanford, seligman, jchoyt, kps}@mitre.org*

As evidenced by the ongoing success of DILS workshops, data integration remains a significant challenge for the practice of biomedical research. To date, the community has (necessarily) focused on the post hoc integration of data sources. As an alternative, we offer an approach in which a community first develops a “core” schema that defines attributes central to the community. Each participant extends the core with a protocol-specific “corona.” Our goal is to lower significantly the cost of entry—to provide catalysts for data sharing.

In our demonstration, we show how our tools support the development of integration-ready schemata. Because any schema derived from the core shares attributes with any other, queries formulated using these attributes can be answered by any participant without any additional integration effort. More specifically, we will demonstrate the following:

Two core schemata for biomedicine based on blood chemistry analyses and MRI scans.

A technique for explicitly representing the derivation relationships among schemata. The derivation graph expresses both the core and protocol-specific extensions. We exploit this graph to improve discovery and query services.

A software tool (Galaxy) that allows a researcher to browse the derivation graph or to search for extensions relevant to his experimental protocol. This software encourages the reuse of schemata, which minimizes the need for post hoc integration.

A middleware platform (MRALD) that allows an end-user to generate a query using his preferred local schema. The middleware infers which other sources are relevant, forwards the query to them, and aggregates the results.

These tools benefit two broad categories of users. First, system developers can search for and adapt existing schemata. This approach provides several benefits: 1) the time required to develop a custom schema is minimized, 2) the system can immediately query other sources for core information, and 3) by registering the derivation relationship, the system can also query other sources that adopt the local customizations. In essence, data producers are rewarded for advertising their modeling work by gaining access to a larger set of sources.

Second, a data user can search for a schema against which his query can sensibly be posed. Because the system stores the schema derivation graph, the system automatically forwards the query to all data sources capable of answering it. In this way, users can discover new data resources and the system assures that the user receives answers from as many data sources as possible.

**Poster 2: Discovering relevant sequential associations between heterogeneous features of proteins**  
*(Selected for oral presentation)*

*Bastien Rance (1), Frédérique Lisacek (2), Christine Froidevaux (1)*

*(1) LRI, Univ. Paris-Sud, CNRS UMR 8623; F-91405 Orsay, France*

*(2) Proteome Informatics Group, Swiss Institute of Bioinformatics, Geneva, Switzerland*

Post-genomics leads to an explosion of information in biology. A challenge for bioinformatics is to integrate those data and extract from them new pieces of knowledge useful for biological experts. In particular, discovering unknown associations between features of proteins distributed across databases is of paramount importance for proteomists. For instance, determining the function of a protein may depend on the function and the order of its motifs (i.e., a functional or a well conserved part of the amino acid sequence). The originality of our approach stems from the fact that we do not consider the motifs as independent attributes but take into account the order in which they appear in the protein.

In data mining, classical knowledge extraction relies on the detection of frequent itemsets or association rules in a database containing objects described with properties of interest.

Association rules are expressions of the form  $A \Rightarrow B$ , where  $A$  and  $B$  are disjoint itemsets. Frequent sequential patterns mining was introduced in the case where the data stored in the database was relative to behavioral facts that occurred over time [1]. The method was introduced as a refinement of frequent pattern mining that accommodates ordered items. Our focus is on significant rare data that co-occur in relatively close association with a specific data. In other words, we examine close dependencies between facts that almost always co-occur and consider them as informative (tight associations), even if these facts are not frequent in the database (low support). Using the notions of interestingness measure and nuggets of knowledge [2] which are association rules with possible low support but high quality for the association, we introduced Sequential Nuggets of Knowledge. We designed SNK, a Sequential Nuggets of Knowledge mining algorithm [3]. Key ideas of SNK include: (i) finding association rules with possible low support and high quality, (ii) using interestingness measures and (iii) using efficient prefix growth strategy for building the antecedent of the rules.

In this demonstration, we highlight how SNK can be used to handle the problem of functional annotation in the context of the Phospholipase D (PLD) family of bacterial proteins. In this testbed, each protein is described by attributes coming from different databases: UniProt, Pfam and secondary private data. Motifs and other information such as functional annotation and partial length are considered as attributes. Finally we show that the identification of sequential constraints can lead to a refinement of the functional classification of proteins: a large class grouped upon a single rough criterion can be subdivided into sub-classes upon explicit and informative distinctive traits.

[1] Agrawal,R.,Srikant,R., (1995) Mining sequential patterns, *In Proc. Eleventh International Conference on Data Engineering*, 3-14.

[2] Azé,J., Kodratoff,Y., (2002) A study of the Effect of Noisy Data in Rule Extraction Systems, *Proc. of the Sixteenth European Meeting on Cybernetics and Systems Research (EMCSR'02)* (2) 781-786.

[3] Froidevaux,C.,Lisacek,F.,Rance,B., (2007) Extracting Sequential Nuggets of Knowledge, *In Proc. Database and EXpert systems Applications DEXA'07*, September, to appear.

**Poster 3: Using GO and MeSH Annotations to Find Meaningful Associations**  
*(Selected for oral presentation)*

*Woei-Jyh Lee, Louiqa Raschid, University of Maryland  
and Padmini Srinivasan, University of Iowa*

The vast amounts of knowledge that is being generated by the biological enterprise is captured and represented in a variety of disparate resources. This data is typically annotated with links to concepts from different ontologies. Data objects in one repository are also physically linked to objects in other repositories. The semantics of these physical links is typically not explicit and not accessible to the scientists. Biologists spend countless hours navigating this Web of interconnected resources, following physical links from objects in one repository to objects in another, then following links from the data to annotations and back to the data, trying to aggregate the information that they need. While the annotated data objects and their physical links form a rich knowledge base, few tools allow users to explore the knowledge captured in these richly annotated graphs, and to find meaningful associations.

We develop the methodology that provides users with a set of tools to explore the rich Web of interconnected and annotated objects in multiple repositories, and to identify meaningful associations. Consider a physical link between objects in two repositories, where each of the objects is annotated with controlled vocabulary (CV) terms from two ontologies. Using the associations generated from a background dataset of knowledge we identify associations between pairs of CV terms that are potentially significant and may lead to new knowledge. We develop an approach based on the logarithm of the odds (LOD) to determine a confidence in the associations between pairs of CV terms.

Using a case study of Entrez Gene objects annotated with GO terms linked to PubMed objects annotated with MeSH terms, we demonstrate a use case to explore potentially significant associations. We consider a simple query where the scientist identifies a gene symbol. Based on this dataset that are associated with the Entrez Gene record, the LOD based confidence scores are determined. A threshold for significance can be determined by the scientist based on the range of scores for this dataset. The scientist can then select a threshold LOD score. The system will use this threshold to identify all associations that exceed the score. The interface will order all associations for this GO or MeSH term, based on the LOD score, and display these associations.

**Poster 4: Querying Relevant Provenance information in Scientific Workflow Systems  
with Zoom\*UserViews  
(Selected for oral presentation)**

*Sarah Cohen-Boulakia, Olivier Biton, and Susan Davidson  
University of Pennsylvania*

Scientific workflow systems have become increasingly popular for managing large-scale in-silico experiments where many bioinformatics tasks are chained together. Due to the large amount of data products generated by these experiments and the need for reproducible results, provenance has become of paramount importance. Several workflow systems are therefore starting to provide support for querying provenance. However, the amount of provenance information produced may be overwhelming, so there is a need for abstraction mechanisms to present the most relevant information. The technique we pursue is that of "user views". Since bioinformatics tasks may themselves be complex sub-workflows, the notion of a user view determines what level of granularity the user can see in the workflow. For example, biologists may simply wish a view in which reformatting tasks are hidden and biologically relevant tasks are seen. Thus the user view determines what data products and tasks can be seen and queried when answering questions of provenance.

In this demonstration, we present the ZOOM\*UserView system [1,2], and focus on the module which generates a "user view" based on what tasks the user perceives to be relevant in the workflow specification [3]. We will show how user views can be used to reduce the amount of information returned by provenance queries, while focusing on information the user finds relevant. More information can be found at <http://db.cis.upenn.edu/research/provwf.html>.

[1] Shirley Cohen, Sarah Cohen-Boulakia, and Susan Davidson  
Towards a Model of Provenance and User Views in Scientific Workflows  
DILS'06, Data Integration for the Life Sciences, Springer-Verlag, Lecture Notes in  
Bioinformatics (LNBI), Num. 4075, pp. 264-279. (2006)

[2] Sarah Cohen-Boulakia, Olivier Biton, Shirley Cohen, and Susan Davidson  
Addressing the Provenance Challenge using ZOOM  
Concurrency and Computation: Practice and Experience, Wiley InterScience (2007)

[3] Olivier Biton, Sarah Cohen-Boulakia, Susan B. Davidson, and Carmem S. Hara  
Querying and Managing Provenance through User Views in Scientific Workflow Systems  
University of Pennsylvania, Report number: MS-CIS-07-13 (2007)

**Poster 5: VisTrails: Using Provenance to Streamline Data Exploration**  
(Selected for oral presentation)

*Erik Andersen, Steven P. Callahan, David A. Koop, Emanuele Santos, Carlos E. Scheidegger,  
Huy T. Vo, Juliana Freire and Claudio T. Silva*

*University of Utah*

In this poster presentation, we will give an overview and a demo of VisTrails, a new provenance management system that introduces a set of new technologies to support and streamline exploratory processes through workflows. Whereas workflows have been traditionally used to automate repetitive tasks, for applications that are exploratory in nature, change is the norm. As a scientist generates and evaluates hypotheses about data under study, a series of different, albeit related, workflows are created while the computational task is adjusted in an interactive process.

VisTrails was designed to manage rapidly-evolving, exploratory computational tasks. By automatically capturing detailed history information about the exploration process and explicitly maintaining the relationships among the workflows created, VisTrails not only allows results to be reproduced, but it also enables users to efficiently and effectively navigate through the space of workflows used in an exploration task (e.g., to follow chains of reasoning backward and forward). In addition, this provenance information is used to simplify the creation and maintenance of workflows; to optimize their execution; to provide scalable mechanisms for collaborative exploration of large parameter spaces in a distributed setting; and it serves as the basis of an infrastructure for knowledge sharing and re-use. As an important goal of our project is to produce tools that domain scientists who are not expert programmers can use, VisTrails provides intuitive, point-and-click interfaces that allow users to interact with and query the provenance information, including the ability to visually compare different workflows and their results.

VisTrails has been released under an open-source license and can be downloaded from <http://www.vistrails.org>.

**Poster 6: A Life Science Data Warehouse System to enable Systems Biology  
in Prostate Cancer**  
*(Selected for oral presentation)*

*Bernhard Pfeifer (1), Christian Baumgartner (1), Johannes Aschaber (1), Friedrich Hanser (1),  
Stefan Dreiseitl (1), Robert Modre (2), Günter Schreier (2), and Bernhard Tilg (1)*

*(1) Institute of Biomedical Engineering, University for Health Sciences, Medical Informatics and  
Technology, Austria, bernhard.pfeifer@umit.at*  
*(2) Austrian Research Centers GmbH - ARC, Austria*

The aim of the IMGUS prostate cancer project focuses on the integration of high-throughput technologies to identify molecular signatures allowing the stratification of patients who are susceptible to curative treatment of prostate cancer. Therefore, patient samples are analyzed using the established laboratory platforms for obtaining -omics data. This amount of data is then integrated into the life science integrative data warehouse (LINDA) for performing systems biology driven data mining and modeling approaches.

Biomedical data provided by the IMGUS project are integrated to enable systems biology approaches. The partners of this prostate cancer project are: Department of Urology - Innsbruck Medical University (biobank, probes and phenomic data), Biocrates life sciences GmbH (metabolomics), Institute of Analytical Chemistry and Radiochemistry (proteomics) - University of Innsbruck, German Cancer Research Centre Heidelberg (genomics), Max Planck Institute for Molecular Genomics Berlin (modeling), and the University for Health Sciences, Medical Informatics and Technology - UMIT Austria (IT infrastructure and data warehousing). To enable all partners in integrating and accessing the project relevant data, a web based platform for structured collection of molecular biological data from a clinical study has been established. Furthermore, in order to identify novel prognostic and diagnostic biomarkers for a tailored clinical management, a quality assured data collection of molecular signatures combined with patient-related data is mandatory with the objective of establishing a high-quality data repository. This platform for web-based medical research networks consists of a Good Clinical Practice compliant Electronic Data Capture (EDC) system with basic features like central patient registry, electronic case report forms, user-role-rights and communication management. This data basis is then processed by extract-transform-load (ETL) services using the Talend ETL software available as open source for performing schema adaptation, data cleansing, transformation and subsequent integration into the data warehouse for performing business intelligence services.

For data mining, modeling and all analytical purposes the integrated data stored in the data warehouse need to be accessed. Therefore, ad-hoc queries are generated and executed. The result set is then used to interpret findings and to perform simulations to get a deeper understanding in the investigated biological processes. The implementation of the query builder enables the end-user to extract data of interest without knowing the database schema, attributes or domains. Using meta data hides technical details from the end-user. Therefore, the end-user can focus on his/her question whereas the query builder is a helpful tool for finding relations between the selected entities.

**Poster 7: CleanTAX: A Framework and System for Automated Reasoning with Taxonomies and Articulations**  
(Selected for oral presentation)

David Thau

Dept. of Computer Science  
UC Davis

Sichen Bao

Agricultural & Resource  
Economics, UC Davis

Bertram Ludäscher

Dept. of Computer Science &  
Genome Center, UC Davis

Taxonomies may be seen as specialized variants of ontologies used for classification. In biology, the traditional Linnaean taxonomic system is widely used for naming and classifying life forms across a number of taxonomic ranks (species, genus, family, order, etc.) based on certain properties. In recent years, advanced approaches for capturing taxonomic information have been developed that improve on the traditional name-based system by explicitly modeling taxonomic concepts. In the taxonomic concept approach, names are further qualified by authority, enabling the differentiation between various “versions” and definitions of taxonomic names. To facilitate data integration, scientists and classification experts are creating expert articulations which interrelate concepts from different taxonomies. While such “bridge-articulations” are obviously very valuable, their manual creation and curation is also extremely time-consuming and costly, as there are only few experts who can articulate the necessary inter-taxonomy concept relations. Moreover, it is almost impossible to understand all logical implications that result from combining taxonomies and articulations, resulting in “integrated” taxonomies that may be logically inconsistent without the curator realizing it.

We have developed a knowledge representation framework and system that views taxonomies as sets of first-order constraints over a decidable logic language. Unlike prior approaches, our system allows scientists and knowledge engineers to automatically (i) check the logical consistency of taxonomies in the presence of latent taxonomic assumptions (logical constraints that are often assumed, but not explicitly modeled), (ii) check whether taxonomies by two different experts are logically consistent with an articulation from a third authority, (iii) decide whether any given expert articulations are redundant, and (iv) infer new implied articulations not stated by the taxonomic expert. The CleanTAX system is our prototypical implementation of the above framework. We outline our overall approach, describe the system architecture, and point out some interesting findings from a large-scale analysis of an expert articulation between two floral taxonomies [1, 2].

[1] L. D. Benson. A Treatise on the North American Ranuncul. *American Midland Naturalist*, 40:1–261, 1948.

[2] J. T. Kartesz. Synthesis of North American Flora. *BONAP*, North Carolina Botanical Garden, 2004.

## **Poster 8: Capturing and Using Semantics for Biological Database Schemas**

*Yuan An, Department of Computer Science, University of Toronto  
Thodoros Topaloglou, Department of Mechanical and Industrial  
Engineering, University of Toronto*

Data integration and data exchange are key elements in conducting scientific investigations in biomedical research. Effective data integration and exchange requires good understanding of the semantics of different biological database schemas. Traditionally, it was a hard and error-prone task to figure out what objects and relationships of the subject matter were represented by the symbols and structures in a schema. In this poster, we explore a framework that employs conceptual models (ontologies) to capture the semantics for biological database schemas. The semantics of a biological database schema are expressed in terms of a mapping connecting the schema that describes the underlying data, to a conceptual model (CM) that describes the subject matter. This framework has been defined to aid the schema mapping problem in database integration. Here we explore how this framework applies to a database maintenance scenario. We describe a motivating example from gene expression data management where the database schema evolves over time and data need to be migrated to the newest schema. We then show how the use of a CM and a tool for discovering the semantic mapping from the previous to the new database schema associated to the CM, could help the database administrator to accomplish the data migration task. We also outline new research issues for utilizing the semantics of biological database schemas in terms of conceptual models to effectively integrate and migrate data across disparate biological databases.

## Poster 9: Screening functional genomics data by querying a data-warehouse

Frederic Lemoine<sup>1,5</sup>, Eric Coissac<sup>2,4</sup>, Anne Morgat<sup>3,4</sup>, Bernard Labedan<sup>5</sup>, Christine Froidevaux<sup>1</sup>

<sup>1</sup>LRI, CNRS, UMR 8623, Université Paris Sud 11, France

<sup>2</sup>Université Joseph Fourier, LAPM, Grenoble, France

<sup>3</sup>Swiss-Institute of Bioinformatics, Swiss-Prot group, Geneva

<sup>4</sup>INRIA, Helix project, Grenoble, France

<sup>5</sup>IGM, CNRS, UMR 8621, Université Paris Sud 11, France

This work lies within the scope of data integration in microbiology. Our main objective is to design strategies to query integrated systems for biology. In the field of genome annotation for example, helping the biologists in the task of handling as well their own data as public data (primary or secondary ones) becomes a crucial priority. In that perspective, we are developing a system which integrates several relational sources, while helping annotation or re-annotation of prokaryotic genomes.

The relational data-warehouse that we are designing has several features. We provide the user with the ability to query the system in a transparent way. For this, we propose a warehouse architecture consisting of 1) a conceptual model that gathers the main biological entities present in the data-warehouse and the relations between them 2) a global schema which is the union of the data sources schemes together with links tables 3) mappings from the concepts of the conceptual model into individual sources. Then, we give to the administrator the possibility to describe his/her standalone databases independently of each others in order to define the concepts present in it, and the associations between them. The architecture is flexible as far as new sources can be easily added and corresponding mappings easily built.

Using the mappings, several ways of querying the data-warehouse are possible based on the concepts of the abstract conceptual model and associations between them. Then different query plans are generated, according to the mappings of the concepts into the underlying sources. Moreover, the user can create mixed queries involving both abstract concepts and concepts expressed as views over particular sources, i.e., the user can specify in his query in which sources some concepts instances should be extracted. The system then answers the query, and also proposes all the other ways to get the same kind of results by searching for the data in all the sources which provide instances of the concepts including the sources not mentioned in the query. We exploit the fact that the data within the data-warehouse can be redundant, complementary or contradictory for re-annotation purpose. If the different answers agree, the annotation can be considered as confident. If these answers are complementary, annotation can be refined. Finally, if they disagree, a re-annotation process should be performed.

This work is part of the French national ANR project Microbiogenomics between biologists and computer scientists (see <http://microbiogenomics.u-psud.fr/>) and is related to studies about OBIWarehouse (see <http://www.grenoble.prabi.fr/obiwarehouse>).

## Poster 10: Using the Orchestra system for biological data sharing

*Todd J. Green, Grigoris Karvounarakis, Nicholas E. Taylor, Olivier Biton,  
Zachary G. Ives*

*University of Pennsylvania*

One of the most elusive goals of structured data management has been sharing among large, heterogeneous populations: while data integration and exchange are gradually being adopted by corporations and small (e.g., organism-specific) scientific confederations, little progress has been made in integrating broad varieties of related sources in areas like biology. Yet the need for large-scale sharing of heterogeneous data is increasing: the life sciences are becoming increasingly data-driven as they have attempted to tackle larger questions.

Schemes for data sharing at scale have generally failed in the past because database approaches tend to impose strict global constraints: a single global schema, fully consistent data, and central administration. To sidestep these limitations, data providers typically resort to custom, fairly ad hoc tools: e.g., large databases placed on FTP sites, which users download and convert into their local format using custom Perl scripts. Meanwhile the original data sources continue to be edited. Our research goal is to provide a more principled and general-purpose infrastructure for data sharing with significant gains in terms of freshness, flexibility, functionality, and extensibility. We have defined a model for a declarative, yet extremely flexible, approach to data sharing, called the collaborative data sharing system, or CDSS [1].

In our system each peer has a locally controlled and edited database instance, but wants to ask queries over related data from other peers as well. To achieve this, a peer's updates propagate along schema mappings to the other peers. However, this update exchange is filtered by trust conditions -- expressing what data and sources a peer judges to be authoritative -- which may cause a peer to reject another's updates. In order to support such filtering, updates carry provenance information. This system targets scientific data sharing applications, and its general principles and architecture have been described in [1,2,3]. In this demonstration we will highlight how such features can be used to handle many common problems in bioinformatics data sharing.

[1] Z. G. Ives, N. Khandelwal, A. Kapur, M. Cakir. Orchestra: Rapid, collaborative sharing of dynamic data. In CIDR, 2005.

[2] N. E. Taylor and Z. G. Ives. Reconciling while tolerating disagreement in collaborative data sharing. In SIGMOD, 2006.

[3] T. J. Green, G. Karvounarakis, Z. G. Ives, V. Tannen. Update exchange with mappings and provenance. Submitted for publication, 2007.

## Poster 11: BIPASS: BioInformatics Pipeline Alternative Splicing Services

Maliha Aziz<sup>1</sup>, Zoé Lacroix<sup>1</sup>, Christophe Legendre<sup>1</sup>, Hervé Ménager<sup>1</sup>, Louiqa Raschid<sup>2</sup>,  
and Ben Snyder<sup>2</sup>

<sup>1</sup>Arizona State University, <sup>2</sup>University of Maryland

Keywords: Alternative splicing, Bioinformatics service, Data integration, Computational pipeline

Alternative Splicing (AS) is a major mechanism in eukaryotic cells. Because a single gene may produce multiple mature mRNAs and multiple protein isoforms, AS contributes to an increase in the complexity of the cellular mechanism for the spliceosome and the transcriptome. A “computational” AS analysis/pipeline typically includes the following two steps: (1) alignment of one or more transcripts against genomic sequences and (2) a clustering step which delimits the transcriptive region of a gene. The alignment step can be executed on known transcripts retrieved from different databases which include a collection of transcript sequences such as cDNA, mRNA, EST, etc. Transcripts are often associated with annotations. The clustering step is a complex process which groups transcripts with respect to a locus, to the sequence orientation on the genomic strand, transcripts overlapping, and identifies and annotates exons and introns in the cluster. The accuracy of a cluster may depend on several parameters such as the quality of alignment, the cardinality or the number of transcripts forming the cluster, etc.

We present two BIP Alternative Splicing Services (BIPASS) that integrate transcript and genomic sequences and use custom methods to align sequences, to cluster transcripts and to annotate exons and introns. BIPASS support two key features of importance to AS researchers. Transcripts and genomic sequences of an organism are first processed by the BIP-Align module that aligns the inputs and stores the results in the BIPAS database. The results of alignment are then submitted to the BIP-Splice module which clusters and annotates exons. The results of both modules are stored in the BIPAS instance supported by IBM's WebSphere Information Integrator (WSII). BIPAS-SpliceDB, the first service, provides access to pre-computed clusters generated from transcripts extracted from various public resources including UCSC (genome data), GenBank/Entrez Nucleotide (full-length mRNAs), and dbEST (EST data). In contrast the second service BIPAS-Align&Splice allows scientists to submit their own transcript and genomic sequences and runs online the complete protocol (powered by WSII). Both services are available at <http://bip.umiacs.umd.edu:8080/index.html> and are registered as Web services and BioMOBY services.

## Poster 12: ProtocolDB: a repository for scientific protocols

Zoé Lacroix, Michel Kinsy, Piotr Włodarczyk, Nadia Yacoubi

Arizona State University

Scientific discovery relies on the adequate expression, execution, and analysis of scientific protocols. Although data sets are properly stored, the protocols themselves are often recorded on paper or remain in a digital form developed to implement them. Once the scientist who has implemented the scientific protocol leaves the laboratory, the record of the scientific protocol may be lost. Collected data sets without the description of the process that produced them may become meaningless. Moreover, to support scientific discovery, anyone should be able to reproduce the experiment. Therefore, a detailed description of the protocol is necessary, together with the collected data sets.

A scientific protocol is the process that describes the experimental component of scientific reasoning. Scientific reasoning follows a hypothetico-deductive pattern and is composed of the succession of the expression of a *causal question*, a *hypothesis*, the *predicted results*, the design of an *experiment*, the actual *results* of the experiment, the comparison of the predicted and the experimental results, and the *conclusion*, supportive or not of the hypothesis. Scientific protocols (also called data-analysis pipelines, workflows or dataflows) are complex procedural processes composed of a succession of tasks expressing the way the experiment is conducted. They usually involve a data-gathering stage that may be followed by an analysis stage. A scientific protocol thus describes how the experiment is conducted and records all necessary information to reproduce the experiment.

We present ProtocolDB, a repository to store and reason on scientific protocols. In ProtocolDB are represented two different layers associated with scientific protocols: design and implementation, and the mapping between the protocol design and its implementation(s). Our approach benefits scientists by allowing the archiving of scientific protocols with the collected data sets to constitute a scientific portfolio for the laboratory to query, compare, integrate, and revise scientific protocols.

## Poster 13: Making data integration research transferable to product

*Arnon Rosenthal*

*The MITRE Corporation*

Data integration has been a database and AI research topic for decades, but industrial strength integration systems embody very few of these research results. The DILS community may benefit from understanding the transition barriers experienced by mainstream integration (XML and relational structures), and possible mitigation strategies.

For the business market, commercial tool suites (at >\$200K!) provide GUIs for manual specifications, glue code (e.g., transformations and wrappers), data profilers, rule languages, and code generators. (Query processors, a separate market, will not be addressed here). Database and AI research offer great promise for automated assistance and automation, to reduce the time and skills required of human integrator. We address four transition barriers.

- Much research from each viewpoint (ontology, DB) ignores the other community's achievements and concerns (e.g., standard logics, tool frameworks, attracting purchasers). We describe opportunities for cross-fertilization.
- Researchers often focus on automating "upstream" tasks. Consider schema integration / ontology alignment tools, whose output describes attribute correspondences. A programmer must manually disambiguate (e.g., should join preserve unmatched items?), insert transformations, and weave them into relational/XML queries that end users can run. Each subsequent schema change will again require a programmer downstream. The work process is complex, with modest benefit. Would your programmers buy and use such a tool? *Downstream principle: automate the last remaining manual step before the end user.*
- A researcher's bright idea needs to be added to a system that has critical mass. Today, each vendor produces (slowly, expensively) a tightly coupled suite. Open source (e.g. Red Hat MetaMatrix, or MITRE's Harmony) would speed the incorporation of capabilities outsiders develop, notably those for life sciences. It could reduce the need for government-to fund life sciences suites. They may also reduce barriers to market entry, and thus prices.
- Researchers rightly present results that require preconditions, e.g., handling only target schemas that lack constraints or are acyclic. But then they stop, providing no guidance about messy real systems. Researchers should decompose the *general* problem, using their technique on some parts and leaving a residue simpler than the original task. For example, if a target schema's constraints make full alignment intractable, align to the target structure with an easier constraint set. The residue (consistent schemas, different constraints) is now a simpler and smaller task.

**Acknowledgment:** The author would like to thank Peter Mork and Jean Stanford for suggesting many improvements.

## **Poster 14: TropGENE-DB, a multi-tropical crop information system**

*Chantal HAMELIN, Manuel RUIZ and Xavier ARGOUT*

*UMR DAP, Développement et Amélioration des Plantes, CIRAD, Centre de coopération Internationale en Recherche Agronomique pour le Développement, Avenue Agropolis, 34398 Montpellier Cedex 5, France*

At the Centre de Coopération Internationale en Recherche Agronomique pour le Développement, (CIRAD), a public French research centre which mandate is to contribute to rural development in tropical and subtropical countries through research, experimentation, training operations, transfert of scientific and technical information, primarily in the fields of agriculture, forestry and agrifoods, researchers gather a lot of various data, especially on the genetic, molecular and phenotypic characteristics of tropical plants of important economic interest for many people in developing countries.

A crop information system called TropGENE-DB has been developed as a tool for the researchers to store and query their data. The most common data stored in TropGENE-DB are information on agro-morphological data, parentages, allelic diversity, molecular markers, genetic maps, results of quantitative trait loci analyses, data from physical mapping, sequences, genes, as well as the corresponding references.

TropGENE-DB is organized on a crop basis with currently nine running modules (banana, cocoa, coconut, coffee, cotton, oil palm, rice, rubber tree, sorghum, sugarcane), with plans to create additional modules for taro, yam and citrus.

TropGENE-DB has been developed using the object-oriented AceDB database management system (J. Thierry Mieg and R. Durbin, 1996). It is based on a generic database model with standardized class and tag names. The same object classes were created for all the species allowing easy comparison and interoperability between the different modules.

TropGENE-DB is accessible for consultation via the internet at <http://tropgenedb.cirad.fr>. Web interfaces have been designed to allow quick consultations as well as complex queries. Each crop module has several interfaces to carry out specific requests. Molecular Marker, QTL or Parentage query sections have been created to be used for all crop modules. They are implemented with Perl/ CGI scripts using modules of the AcePerl Application Programming Interface (API) and the AceBrowser generic web interface.

Standard Excel files corresponding to the various types of data that can be submitted have been created to allow standardized and easy data submission. These files and a web form to post the standard data files for their incorporation in TropGENE-DB are available on our internet TropGENE-DB website. Current TropGENE-DB data have been submitted by different CIRAD teams and by scientists from other institutions working on tropical crops. Potential submitters can contact us at the following address [tropgene@cirad.fr](mailto:tropgene@cirad.fr).

TropGENE-DB is being moved to a relational MySQL database which offers more possibilities for future developments since the ACEDB system is no longer maintained.

## **Poster 15: GenMapping Web Query, an integrated and generic Web application connected to multiple genetic and genomic mapping plant data sources**

*Alexis Dereeper, Xavier Argout, Manuel Ruiz*

*UMR DAP, Développement et Amélioration des Plantes, CIRAD, Centre de coopération Internationale en Recherche Agronomique pour le Développement, Avenue Agropolis, 34398 Montpellier Cedex 5, France*

GenMapping Web Query is an integrated and generic Web application connected to multiple genetic and genomic mapping plant data sources.

Interoperability between various data sources is supported by a Java-based middleware called "Pantheon" (1) developed for the Generation Challenge Programme, GCP (2). This middleware specifies a Model-View-Controller architecture and implements a set of public GCP domain models and ontologies (3). We have also developed GCP domain-constrained BioMoby (4) Web services to connect and to integrate the information belonging to the dispersed data sources.

Some of the accessible databases are generic such as GBrowse (5), CMap (6) or Chado (7), whereas others are specific such as TropGeneDB (8), OrygenesDB (9), Gramene (10) or EMBL databases (11). GenMapping Web Query is also linked to generic viewers like CMap for displaying comparative maps, and GBrowse for displaying genome annotations.

GenMapping Web Query can be tested at the URL:

<http://genmapping.cirad.fr/genmapping/webquery/org.generationcp.genmapping.web.servlet>.

Some example of available use cases are:

- get markers, maps, QTL and gene information combining several data sources and several species
- provide link between rice QTL from Gramene and rice sequences annotation from OrygenesDB.

- (1) <http://pantheon.generationcp.org>
- (2) <http://www.generationcp.org>
- (3) <http://pantheon.generationcp.org/demeter/index.html>
- (4) <http://www.biomoby.org/>
- (5) <http://www.gmod.org/wiki/index.php/GBrowse>
- (6) <http://www.gmod.org/wiki/index.php/CMap>
- (7) <http://www.gmod.org/wiki/index.php/Chado>
- (8) <http://tropgenedb.cirad.fr>
- (9) <http://orygenesdb.cirad.fr>
- (10) <http://www.gramene.org>
- (11) <http://www.ebi.ac.uk/cgi-bin/dbfetch#Databases>

## **Poster 16: Efficient Bucketization and Its Role In Data Integration**

*J Kadin, J Blake, C Bult, J Eppig, M Ringwald, J Richardson, and the  
Mouse Genome Informatics Staff*

*Mouse Genome Informatics  
The Jackson Laboratory  
Bar Harbor, ME 04605*

The primary activity of the Mouse Genome Informatics (MGI) project is the acquisition, integration, and dissemination of scientific data on the laboratory mouse. Integration usually involves comparing objects in an incoming dataset with objects already in the MGI database in order to establish correspondences between them. In order to deal with the exploding volume of data, while maintaining the high quality for which MGI is known, we have developed methods and tools that allow large scale automated data processing to be combined with targeted expert curation of "problem cases". At the core of many of our data loads is a process we call "bucketization" in which association graphs between two sets of objects are created and partitioned (into "buckets").

In a typical load, the vast majority of cases end up in a high-confidence bucket and can be loaded with no manual intervention. Other buckets can be assigned to curators for resolution.

As an example, we use bucketization to correlate UniProt protein records with MGI gene records by analyzing correspondences between GenBank sequence IDs associated with both sets of records. The UniProt records that correlate one to one with an MGI gene record form a high confidence "bucket" of associations that the software can load automatically.

We will define bucketization, show how it is used in actual data loads, and describe an efficient implementation. Using appropriate indexing schemes, the algorithm runs in linear time of the size of the two data sets being compared.

URL: [www.informatics.jax.org](http://www.informatics.jax.org)

MGI is supported by NIH Grants: HG000330, HD033745

## **Poster 17: Large-Scale Microarray Analysis Reveals Switch-Like Behavior in Genes within Cell Communication and Adhesion Pathways**

*Adam Ertel and Aydin Tozeren*

*Center for Integrated Bioinformatics, Drexel University, Philadelphia, PA*

Genes expressed at high levels within a small subset of tissue types have been previously described as tissue-selective, with a subset of these genes highly expressed in one and only one tissue described as tissue-specific. Additional descriptions exist for the distribution of gene expression, including maintenance/ housekeeping genes; those expressed constitutively, graded genes; expressed across a continuous range, and binary/bimodal genes; expressed around two discrete levels. The relationships between these distributions of gene expression and aspects of tissue selectivity, however, have not been well described. Our work bridges previously published concepts of tissue-selective gene expression with descriptions of gene expression distribution profiles, focusing on genes with bimodal expression distributions.

We assembled a large collection of Affymetrix MGU74Av2 mouse gene expression data for normal tissue in order to profile gene expression distributions across diverse conditions, as well as identify potential of genes to participate in tissue-specific functions. Approximately one quarter of the genes on the MGU74Av2 array were identified as bimodal, or “switch-like,” by fitting a mixture of two normal curves to the gene expression distribution. The set of bimodal genes was further divided into subsets of genes expressed in the higher mode within one (tissue-specific) or more (tissue-selective) tissue types. Resulting gene sets were tested for functional enrichment among Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Gene Ontology (GO) terms. KEGG Cell Communication and Cell Adhesion molecules pathways, as well as the GO cellular component terms “integral to membrane” and “extracellular space” were significantly enriched for tissue selective genes. A majority of the bimodal genes within these functional categories also had tissue-selective behavior. Results suggest genes with bimodal switch-like control play a large role in determining cell phenotype through interaction with the extracellular environment.

**Poster 18: Interactome-wide Analysis of HIV-1 Host-Pathogen Interactions and Their Disruption of the Native Network Topology**

*William Dampier and Aydin Tozeren*

*Drexel Center for Integrated Bioinformatics*

Interaction networks between genes and proteins enable living cells to process information and respond to stimuli. Pathogenic infection of the host cell disrupts the native network by creating new paths and obliterating existing ones. In order to elucidate significant changes in the network we integrated the NIAID HIV-1 Protein Interaction Database into the Kyoto Encyclopedia of Genes and Genomes pathway database. By measuring the change in nodal-centrality between host and host-pathogen networks we can determine which genes and proteins are the most disrupted within the host network. The nodes can be ranked by the magnitude of the fold-change of their centrality from the native network or by the statistical likelihood of observing that fold-change as compared to random additions of HIV-‘like’ nodes. This provides a set of ranked gene lists which shows statistical over-representation in canonical pathways. These lists can also be used to search for therapeutic targets which lay many steps away from a direct interaction with HIV yet may have therapeutic benefit.