Information and Process Integration: A Life Science Perspective

Organizers
Johann-Christoph Freytag, Humboldt-Universität zu Berlin, Germany,
freytag@dbis.informatik.hu-berlin.de
Thure Etzold, Lion, England, etzold@lionbio.co.uk
CaroleGoble, University of Manchester, England, carole@cs.man.ac.uk
Peter Schwarz, IBM Almaden Research Center, Ca, USA, schwarz@almaden.ibm.com
Rolf Apweiler, EBI, England, Rolf.Apweiler@EBI.ac.uk

January, 29 – 31, 2003

Summary

This workshop brought together scientists and industrial developers and researchers to discuss the challenges of integrating bioinformatics/life science data in a meaningful way. Despite the technological advances many open problems and issues persists and need to be addressed. This workshop focused on the main issues of data nad process integration in the life science domain.

Detailed Agenda for the Workshop

Over the last fifteen years the amount of data in the area of Life Science/Bioinformatics has grown exponentially. This data is stored and is available in an ever increasing number of data collections (also often referred to as databases), each focusing on specific aspects of life science, such as nucleotide or protein sequences, functional motifs, metabolic pathways, specific organisms, or information related to specific diseases. At the same time the bioinformatics community has developed hundreds of tools to visualize, to analyze, and to process that data, with the goal of turning raw data as produced by sequencing machines into knowledge applicable to drug design and to the development of new therapies. Examples include gene prediction, motif recognition, the computation of phylogenetic relationships, and the deduction of pathways from gene expression arrays. However, almost all of these tools use proprietary, non-standard data formats thus making it (almost) impossible to change those or to introduce new tools without recognizing the need for bridging the gap between the existing world of data and processing conventions and new promising approaches.

With the advent of middleware technology, the focus of research and development in data integration has begun to shift. While many previous efforts have addressed the syntactic integration of data collections, the real challenge now, and for years to come, will be the development of new approaches, techniques, methods and algorithms for performing **semantic integration**. What will be needed are systems that bring together data that “belongs” together, making this determination on the basis of both structure **and** meaning. To achieve this goal, current middleware technology will need to be extended so that it can take advantage of ontologies, semantic networks and other metadata (e.g. information about data quality) to gain a deeper understanding of the primary data.

The problems described are present in both academic and research institutions as well as in pharmaceutical, drug design, medical, and health care businesses. Only the use of modern technology promises the users a platform to bring diverse data, information, knowledge, and processing software together to advance science and to satisfy business needs. If the current time necessary for the development of a new drug, which is estimated to be at app. 10 – 15 years, is to be reduced fundamentally, the process from molecular biology evidence to clinical studies has to be highly streamlined, which requires a tight yet flexible intertwining of a multitude of databases and applications.

This seminar should bring together scientists and practitioners from the fields of bioinformatics and information technology, in order to better understand the new challenges as well as existing approaches and relevant technologies. Solutions to the new problems will most likely be driven by extending existing technology (e.g. Object-Relational DBMS) to meet new needs (e.g. federated database management, highly-parallel distributed problem-solving on a grid), emerging tools and standards for managing semi-structured data (e.g. XML, XQuery, XSchema) and process technologies (e.g. CORBA, Java Beans, message-driven workflow using Web Services).

New technology areas such as the onlotogies, the **Semantic Web** and the Grid are highly applicable to a more meaningful integration of data, information, and processes for Life Sciences. It becomes important that mutual understanding in both the research and business world arises to make the necessary advances in bioinformatics. Still, it is time to evaluate the current solutions and approaches to drive future research and development directions by the pressing needs of the bioinformatics/life science community.

The areas to discuss include:

- Achieving semantic integration
    - What are today’s approaches for semantic integration? Are those sufficient for the life science domain?
    - What are the necessary concepts such ontologies that are necessary to perform semantic integration?
    - What are the languages required to specify the various forms of biological and medical knowledge that is required for bioinformatics research? Are relations and attributes really enough?
    - Which knowledge management techniques (personalization, community building, knowledge sharing, text mining) are appropriate to the Life Science area?
    - How to ensure data quality, data consistency, and completeness? How can data quality be compared, assessed, measured, combined?

- Information discovery and publication

  - What is the optimal access form to the various data collections that are important to scientific organization and business in the different life science areas?
  - Can XML be used as the “universal language” for describing the integrated information base? How to capture "navigational access” based on hyper-linked HTML pages performed today in many application areas?
  - Version management for data collections and metadata that change daily/weekly? Are there compression schemes that can reduce the large amount of repeated (redundant) data? How can we efficiently store the relationships between new or changing evidence and new versions of data?
  - How is information described? What are approaches to handle the description of data (metadata)? Which metadata is relevant (schema, ontologies)? How to store and access it? How to keep it current?
  - What is a federated schema if structured and unstructured data are brought together? Which schema integration techniques, federated query and search technologies are applicable?
  - What are possible system structures in a highly dynamic world that constantly changes and that makes constant progress?

- Information processing paradigms
  - Which processing/transaction models are appropriate?
  - How can ontologies and other meta data support more meaningful processing techniques? Are current techniques adequate for distributed query processing? What are new requirements coming from Life Science?
  - How to represent and manage derived data, data quality and data provenance?
  - How do Semantic Web and Grid technologies contribute?
  - Which federated database technologies can be used in which context? Are the trade-offs that provide the bases to decide which approach to choose in a particular situation?

- Information technologies and standardization
  - How to use different technologies like SQL/MED wrappers, J2EE connectors, EAI adapters, and Web Services for virtual or physical integration. Which technology should be used under which circumstances?
  - Which role will database systems, application server, workflow systems, messaging systems, portal servers, etc. play? How do they relate and cooperate?
  - Does Web Database Technology suffice?
  - What is the query/retrieval interface for the future?
  - What must be standardized in the storage, access, and processing for better information integration?
  - What is the minimum in standards one needs for improved “cooperation” and “collaboration” of applications?

- How can XML-based meta data help to improve to understand the semantics of data to perform challenging tasks such as information integration?

As cross fertilization is important, the major goal of the seminar is to bring representatives from the different communities (from research, from vendors, and from users) together for a joint in-depth understanding of the issues, to identify and prioritize the main research items, identify standardization needs, and to discuss demanding questions and open problems in detail. As a major driving force we plan to use case studies coming for life scientists to discuss many of these issues from a user's (i.e. Life Science) perspective.

Short Summary

The result of the workshop showed that integration is still wide open field base on the differences in technology, the expectations by the users, and the kind of problems that biologists and life scientists try to solve. It became apparent that often the integration task is driven by the specifics of the application ("lab protocols' and their mapping onto computer systems). The discussions also made clear that integration must include semantic integration, in particular the meaningful integration of different space and time scales (microseconds vs. millions of years) and the presentation of discrete and continuous data (the former is well understood, the latter is an open area). Another open (biological) issue is the use of measurements which are often not reproducible thus making it difficult to compare and to use. Finally it became apparent that biologists and computer scientists must cooperate much closer to solve the complex problems that exist in life science and are about to appear on the (scientific) horizon.

Abstract of Presentations

Process and Data Integration: and Issues from a Computing Perspective

Carole Goble
University of Manchester, UK
carole@cs.man.ac.uk

Data and process integration has attracted a great deal of industrial and academic effort over the last 30 or so years. As an introduction and scene setting talk to begin the workshop, I cover three points:

(i) Various dimensions of integration questions what are we integrating, in what setting, for what purpose and so on. In particular, addressing the difference between integrating stable resources for well defined protocols in a production environment and that of an experimental scientist having to form ad-hoc and short lived integration that are personalized.

(ii) The "glue" that we use for integration such as better keys, ontologies and schema mapping, and that sometimes the integration model is not known or understood;

(iii) Experiences from my own work in (a) a closed schema based mediation system TAMBIS, and (b) more recently open, loose complex service collaboration framework, in the myGrid project: http://www.mygrid.org.uk.

Seamless Integration of SRS into Relational Database Systems

Thure Etzold
Lion Bioscience Ltd., Cambridge, UK
Thure.Etzold@uk.lionbioscience.com

The SRS (Sequence Retrieval System) approach to data integration has evolved over many years to address the needs of researchers in the life sciences to query, retrieve, and analyze biological data. SRS follows the federation approach to data integration, leaving the underlying data sources in their original formats which are stored as flat files, XML files, or within a RDBMS. SRS is designed to be scalable and supports server installations with more than 200 databases.

SRS has its own query language to express search predicates, but not data retrieval. This is achieved in a second step using the entry set generated by the query. This does not present a limitation, but is considered as being non-intuitive by users familiar with SQL.

A new development in SRS to address this issue uses APIs provided for the RDMS from Oracle and for DiscoveryLink from IBM. These are called gateway and wrapper, respectively, and allow seamless integration of content stored in foreign databases and query systems, when then "behave" as a native table structure.

The gateway ad wrapper implementation for SRS a very similar to each other and support a subset of SQL which includes string comparison, range queries, sorting, and joins. A capability table publishes this SQL subset to the "host RDBMS" which can then complement it so that overall a fully featured interface to SRS is achieved

DiscoveryLink – A Status Report

Peter Schwarz
IBM Almaden Research Center, CA, USA
schwarz@us.ibm.com

My talk focuses on the description of DiscoveryLink, a federated database management system that supports integration of a wide variety of query engines onto a single virtual database. In particular, I describe, I describe the wrapper architecture, which allows the DB2 optimizer and wrappers to specific data sources to cooperatively develop an efficient query execution plan. After some illustrative sample queries I describe how several customers have effectively used DiscoveryLink to answer interesting scientific queries that combine data from multiple sources. This experience has brought several issues to light, including the inappropriateness of SQL for use by scientists, the difficulty of wrapper development, the need to track and present the origins of data values, and the ability to degrade gracefully when sources are unavailable. I close with a description of Clio, a prototype system that addresses the problem if mapping a set of legacy schemas into a known target schema.

UniProt & Integra8

Rolf Apweiler
EMBL-EBI, UK
apweiler@ebi.ac.uk

In my presentation I present how we want to build on SWISS-PROT, TrEMBL, and PIR to create UniProt, the Universal Protein Knowledgebase, which will be the central hub for protein-related information. UniProt and other major life science databases (EMBL, DDBJ, GenBank, InterPro, PDB, ArrayExpress, model organism databases, etc) will be integrated in Integr8, an interoperability layer modeled along the InterPro model: Only core data gets stored in a central database, the bulk of the data resides in the member databases.

Processing Genome Data using Scalable Database Technology

Johann-Christoph Freytag
Humboldt-Universität zu Berlin
freytag@dbis.informatik.hu-berlin.de

My talk provides an overview of our (DBIS) research work in the area of Life Science. Based on current database technology we describe three challenges that we try to tackle using scalable database technology. First, we describe our approach to (physical) data integration by building a “genome warehouse” based on a framework that encompasses flexibility and extensibility. Second, we show how to integrate BLAST as a user defined function (UDF) into IBM’s DB2 database management system. Third, I use the example of alternative splicing to argue that more efforts are necessary to integrate various technologies (such as workflow concepts) to allow scientists in Life Science to access and process relevant data more efficiently in their problem solving efforts. For more information please visit http://www.dbis.informatik.hu-berlin.de/research/bioinformatics.

Workflow in DiscoveryNet

Yike Guo
Department of Computing
Imperial College, UK
yg@doc.ic.ac.uk

This talk presents our research in workflow from the Life Science research in an e-Science project, DiscoveryNet. The presentation focuses on the main issues of design and development of workflow as a uniform mechanism for integrating information, discovery processes, and knowledge results. In particular, we argue that the workflow development plays an important role in organizing distributed, information driven scientific research within the grid environment. We demonstrated the implementation of workflow solutions for various bioinformatics applications. The demonstration shows that a workflow system provides support for composing services, collaborative research discovery, provenance, and, even more importantly, for unifying knowledge discovery. For more information please visit http://www.discovery-on-the.net.

Building Integration Systems to support Scientific Discovery

Zoé Lacroix
Arizona State University, USA
zoe.lacroix@asu.edu

Query-driven approach vs. generic approach:

Life scientists typically follow a query-driven approach. They aim to collect data to answer a scientific question. The collection process is usually expressed by a list of scientific tasks that can be represented by a workflow. In contrast, computer scientists aim to build generic systems that answer multiple queries. In order to develop systems in a timely manner that support users’ needs, the system requirements should focus and capture a narrow set of specific scientific tasks. Such an approach would satisfy a small community of users instead of trying to build a “one-fits-all” system that is likely not to fulfill users’ expectations.

Integrating data and tools:

Traditional approaches are not designed to accommodate appropriately the need for integration of both data and applications. Data warehouses, mediations, and federations are data-centric while CORBA, SOAP, and Web services are application-driven. We need to define new models for integration of data, applications, and processes.

Semantic integration:
Integration of biological resources raises a variety of issues related to semantics including scientific object identification, characterization of resources (Web semantic), and exploitation of multiple query evaluation paths.

Evaluation of biological integration systems:
The design of a biological system should rely on clear system requirements capturing users' needs. These specifications should be used to evaluate the performance of the system and its ability to meet those needs. In addition, they characterize the system, therefore providing potential users some insight on what the system can be used for and how to compare it to other systems. Requirements may be sued for improvements (e.g. for optimizations). Current research should benefit from the development of appropriate performance models, evaluation matrices, cost models, and benchmarks.

Methods for Integration of Databases with common Subject Domains

Maria G. Samsonova
St. Petersburg State Polytechnical University, Russia
samson@spbcas.ru

We propose a method for integrating databases with common subject domains whose key components are: use of the conceptual schema of the knowledge domain and domain-oriented dictionaries, processors of queries in natural language, and a multi-agent architecture to integrate the results of information retrieval from several databases.

The advantages of this method are the likely integration of any set of databases (either stored locally or published on the Web), the possibility of querying the database in any natural language, the use of the conceptual schema (and knowledge) to formulate and process queries, the simplicity of adding additional databases, the easy of access, the adaptivness to changes in the knowledge domain and the user's view of the data, the system robustness, and the possibility to distribute optimally a load among several database mirrors.

Data and Process Integration in Clinical Genome Research and ChemInformatics

Martin Hoffmann
Fraunhofer-Institute, SCAI, St. Augustin, Germany
martin.hofmann@scai.fraunhofer.de

Our group is working closely with medical genome researchers from the University of Bonn. The massive use of functional genomics technology in clinical research results in a flood of data that – although the data has been generated using patient material (samples) – is only poorly linked to phenotype descriptions and patient information. As a first task we face the need to bring existing database systems to a professional state, moreover, during this process we have to adapt the database to a high level conceptual schema representing information and material (sample) flow in clinical research.

For reasons of simplicity and following an opportunistic philosophy, we built on existing technology for data integration and representation whenever possible. That is why we plan to integrate all clinical and molecular databases by using SRS. We can count on hundreds of parsers publicly available for “external knowledge” (public databases). For “local databases” we will be able to reuse parsers, e.g. for expression database such as the ArrayExpress database.

Regarding ChemInformatics we have modified a workflow engine (“TENT”) that has originally been built for engineering simulation purposes. The system will be tailored for the use in ChemInformatics, a compute intensive discipline that will benefit mostly from distributed compute abilities.

Pathport - A Life Science Entrance to the GRID

Stefan Hoops
Virginia Bioinformatics Institute, USA
shoops@vt.edu

PathPort is an extensible collection of viewers and tools, which allow the life scientist to access and disperse data and programs. In addition, it provides the users the means to associate different data models. This is accomplished by using a software bus architecture “ToolBus” on the client side. The server side is based on Web services providing access to data as well as to processing capabilities.

Standard and Ontologies for Microarrays

Ugis Sarkans
Europeans Bioinformatics Institute (EBI), UK
ugis@ebi.ac.uk

The microarray community recognized the need for standards already several years ago. Currently, several problems have been solved: what information has to be exchanged, how to organize the exchange (object model, XML format), common terms (ontologies). Following problems still remain: how to precisely describe data processing, how to generalize data processing (workflows), what infrastructure facilitates the process of building ontologies for the whole community (both on the class and on the instance level), how to model various domains using a layered approach enabling the transfer of experience into new domains and into (standard) queries.

Biotalk: A Portal for Tools and Ontologies for XML-based biological data Exchange

Isabel Rojas
European Media Laboratory, Heidelberg, Germany
Andreas Eberhart
International University, Germany
isabel.rojas@eml.villa-bosch.de, andreas.eberhart@i-u.de

People carrying out a project that requires data integration from publicly available databases normally write their own parsers for those databases. This is a tedious and cumbersome task that is repeated over and over again. We propose the creation of a portal where people can download (and submit) (XML) parsers for the main biological databases or to their own databases, for others to use them. This portal should offer an exchange platform for parsers and schemas, a discussion forum, documentation of schemas and databases, and tools for XML data processing and XML schema transformation. To prove this concept, we carried out a small project where we provided two parsers for two of the existing enzyme databases. Using existing XML tools we managed to create a database with all the data with little effort. The second stage of the project is aimed at offering tools to combine the schema with ontologies in order to support the curation and validation process.

Workflow for Data Mining Genes

Luciano Milanesi
CNR –ITB, Italy
luciano.milanesi@itb.cnr.it

Gene identification in newly discovered DNA sequences is an important problem in current molecular biology studies. Due to recent progress in large scale sequencing projects, gene identification programs have become widely used. The use of these programs can significantly simplify the analysis of newly sequenced DNA especially when applied in combination with experimental methods.

Although good results have been obtained with a variety of computational approaches the problem of gene structure prediction has not yet been completely solved. The analysis of human genes cannot merely be considered as a “linguistic analysis problem” of nucleotide strings because the gene structure is made up of many other important features. These include higher-order chromatin structures, the near random (??) nucleosome positioning along the DNA, the different features of the three dimensional structure of the DNA (or RNA), and the torsion strain (???) of the DNA included by transcription.

A workflow of programs for data mining genes in different organisms will be described. The workflow is able to use related information scattered over numerous databases and web sites.

Open Grid Service Architecture – Data Access & Integration (OGSA-DAI)

Dave Pearson
Oracle UK, UK
dave.pearson@oracle.com

AGSA-DAI is a collaborative program of work including the University of Edinburgh, Manchester, and Newcastle, with industrial participation from IBM and Oracle. It represents a significant contribution on behalf of the UK e-Science Core Program to extend the Grid model to include database interoperability. The scope of work is the definition and development of generic Grid Data Services providing access to and integration of data held in RDBMSs, as well as data in XML repositories. The functional definition will ultimately form the basis of standards/recommendations on data access and integration to the Global Grid Forum (GGF).

The presentation provides an overview on the technology and the e-Science drivers in development of the Grid. It defines the principle concepts of Grid computing, the benefit of the Open Grid Service Architecture in enabling virtualization, discovery, and sharing. It describes the motivation for developing grid data services, and reviews their functional scope based on requirements for data access and integration within the UK e-Science community. The presentation describes the design of the OGSA-DAI architectural framework and the functionality of its components. It also illustrates the behavior of Grid Data Services within the current framework. The functional enhancements in future scheduled releases of the reference implementation are outlined, as well as the direction of planned future work.

The initial OGSA-DAI reference implementation and associated documentation is available for download under an open source license agreement at the following site: http://www.agsa-dai.org.uk.

Integration of biological Resources Information in the Bioinformatics Network Environment

Paolo Romano
National Cancer Research Institute, Genoa, Italy
Paolo.romano@istge.it

The term “biological resources” refers to living biological materials, such as bacterial and fungal cultures, human, animal, and plant cells, isolated genetic materials collected in culture collections.

Over the decades a wealth of information on and around these materials has been accumulated, however information is still dispersed and not always easy to retrieve. Coordinated approaches have been initiated to improve accessibility of biological resource centers, their holdings and related information through the internet/web. Standardization of data handling and data accessibility had to be covered.

The management of internal and public data needs constant improvement, and retrieval of information such as access to bibliographic databanks and to molecular biology databases is needed for keeping abreast of taxonomical developments. Bioinformatics tools, including databases and software, are more and more integrated in a unique distributed environment, so that end users can easily retrieve and elaborate the information they need and software can automatically exchange data on the internet. Coherent and widely agreed data models, controlled vocabularies and ontologies, together with the most innovative telematics standards and tools make this goal feasible within the next years. Integration of biological resource catalogs (BRCs) in this bioinformatics network environment will allow the user a global viewing of all available data and will make the collections available more effectively after a search in molecular biology databases. Vice versa, it will make the retrieval of detailed information from molecular biology databases easier after a search in the catalogues of those collections. This will result in the major consequence that more and more researchers will refer to BRCs and make a wider use of biological materials of certified quality, thus raising the level of today's biological research.

CABRI is a demo project funded by the EU from 1996 to 1999. It has implemented a unified access to culture collection catalogues of participating collections by also guaranteeing a common high level of quality of material and related information. The final achievement of the project is a "one-stop-shop" where researchers can search, analyze, identify, select, and re-order strains of their interest. Currently, it includes more than 100.000 strains and cell lines from 28 collections (http://www.cabri.org).

The integration of databases is mainly a matter of identifying links among existing information and data, and implementing both common formats for data interchange and application driven data interchange software. When databases refer to similar of identical objects, a careful selection of and comparison of data models and structures is also extremely important.

CABRI is a good example of an SRS based system for integrating searchers on the catalogues of European culture collections. The implementation of the CABRI collections' catalogue in an SRS based system has been carried out according to the following three steps: the comparison of all the information that is available in CABRI catalogues, the definition of Minimal and Recommended data sets for each material, the design of the catalogue's production guidelines, including procedures of authentication of data for the creation of flat files. These data sets can be an adequate starting point for the definition of a DTD for a Biological Markup Language.

Following steps will be the set of Web Services according to the current efforts that have led to the definition of standards such as WSDL and UDDI.

Database Support for Microarray Gene Expression Analysis

Erhard Rahm
Universität Leipzig, Germany
rahm@informatik.uni-leipzig.de

Microarrays make it possible to monitor the expression of thousands of genes simultaneously, thus generating large amounts of data. In my talk I present the major requirements for microarray data management. I consider the various kinds of data, data normalization, data and metadata integration, and analysis needs. Next, eight previously developed microrarray databases are comparatively evaluated with respect to the identified requirements. Finally, I present the microarray data warehouse GEWARE that we developed in the bioinformatics center in Leipzig that overcomes limitations of the older approaches.

Service Based Distributed Query Processing on the Grid

Norman Paton
University of Manchester, UK
norm@cs.man.ac.uk

The Grid is emerging as a middleware platform for dynamically discovering, accessing and exploiting distributed computational resources. The Open Grid Services Architecture (OGSA) is combining Grid resources management functionality with the service description, invocation, and coordination facilities of Web Services. How can such infrastructure be applied for information integration in bioinformatics? It is perhaps the case that Grid Services can be exploited directly by bioinformatics applications, but a more likely scenario is that higher-level services will be used to ease the development of high-performance distributed applications on the Grid. As an early step in this direction, we are developing a service-based distributed query processor (DQP) for the Grid. By service-based, we mean (i) that distributed queries range over resources described as services, and (ii) that core Grid Services are used to implement the DQP. When a DQP is developed in a Grid setting, computational resources can be allocated dynamically to support the efficient evaluation of computationally challenging queries. By deploying query optimization and evaluation techniques form parallel databases in the Grid setting, implicit parallelism can be used to provide scalable performance. As many bioinformatics applications combine data access with computationally intensive analysis, we believe that system-supported optimization is a distributed environment should significantly ease the expression of complex requests over biological resources.

OBIEnv: A simple integrated Environment of Dispatcher, Information Server, and Software/Data Updater for the OBIGrid

Tomoyuki Yamamoto
Japan Advanced Institute of Science and Technology (JAIST), Japan
t-yama@jaist.ac.jp

The Gird computing is, in our view, a virtual environment that provides transparency for users. The technology is emerging, but there are still a huge number of problems to be solved. The problems cannot only be found on the technology side, but also in policies and/or on in "psychological" issues regarding the users. We emphasize the importance of real-world experiments, using simple systems that can grow step-by-step. The OBIEnv is an experimental system for this purpose. The functionality is limited, assuming only, short-term, small file-based tasks only (e.g. parallel execution of BLAST search). The system consists of a dispatcher, an information server (called P2P server), and a data updater. The center piece of the system is the P2Pserver using PostgresSQL. The functionality is limited, but (quasi-) real-time update of CPU-usage is implemented which is essential for the Grid. The system grows, but should be based on the "loosely-coupled" principle which allows users to be flexible.

Assigning Experimental Data to Real World Objects

Stephan Heymann
Humboldt-Universität zu Berlin, Germany
heymann@dbis.informatik.hu-berlin.de

It is a permanent task in bioinformatics to properly assign experimental findings to basic item sets of real world objects kept as digital images in life science data collections. There are at least three major categories of data that increasingly add "semantic glue" by filling the space between those objects:

(i) properties and features that characterize *individual* objects;
(ii) facts of certain qualities that constitute *pairwise* relationships between objects;
(iii) subset-specific information for a *group* of objects.

In terms of graph theory, these categories function as structuring principles for an abstract depiction of the basic item set. Graph structures in and across large experimental data sets follow their own "laws". Thus, bioinformatics experiences a need for powerful support by graphical visualization and navigation strategies. The Viator family of graphical networking tools provides a user with aids for the elevation of data inherent inter-object links from large data sets. Such graph structures, when superimposed in the GUI of a data storage system enables the user to intuitively highlight hidden correlations behind heterogeneous data sets.

Protocols

Werner Kriechmaum
IBM Development, Böblingen, Germany
kriechba@de.ibm.com

Protocols (e.g. methods) have a long history in he wet lab where they are used to describe the data collection and the data analysis process. Most current bioinformatics tools are designed for interactive usage, and the protocol how to connect different steps of the complex data analysis exists only in the "head of the user." As can bee seen from the recent special issue of Nature Genetics - that details the steps needed for some of the more common analysis procedures in a "how to" fashion – there is an increasing demand to capture this informal knowledge in protocols to facilitate reproducibility and automation of data analysis in bioinformatics. From an IT point of view, a protocol may be implemented as one algorithm, a pipeline, a workflow, a query plan, a directed graph, a Petri net, …

On Data Quality in Life Science Data Integration

Ulf Leser
Humboldt-Universität zu Berlin, Germany
leser@informatik.hu-berlin.de

Data quality is a frequently mentioned problem of lice science databases. However, data quality issues are rarely studied in a systematic way. The talk presents a number of data quality issues in bioinformatics base on experiences in a variety of projects in life science data integration, data curation, and data quality assurance in high throughput experiments.

The talk is more geared towards showing problems of data quality and stimulating discussions, rather than presenting applicable results. We also question whether it is possible at all to talk about data quality in life science on a satisfactory level of abstraction.

Towards an interactive Toolbox for computational Cell Biology using Grid Resources

Wolfgang Tvarusko
IBioS, DFKZ, Germany
w.tvarusko@dkfz.de

The modern view of life science is holistic, integrative view ranging from single molecules to new forms of individualized medicine. Advanced robotics allow for efficient high throughput and repetition of simple tasks. More sensitive analytical methodologies and their miniaturization enable high experimental densities on even smaller amounts of material. The recent collision of theses major paradigms together with the drive for high parallelization will produce petabytes of data. It is called upon computational biology and especially system biology to bring meaning to the data created.

Detailed analysis using time-lapse DIC microscopy has allowed Gönczy, P. et al (1999) to identify phenotypic signatures characteristic of functional groups of genes, and thus establish a unique functional database cell division proteins. This DIC assay constitutes the core of the screen and should be optimized for detecting even minor deviations from the wild-type development. An automated system for interpreting movies of cell development patterns would therefore have a number of advantages over current manual practice. There would include objectivity, reliability, and repeatability after the ontological descriptions pf these patterns.

In order to come to meaningful biological results in assigning cellular function to genes in BIC arrays, thousand of movies in the range of 20 to 30 megabytes have to be evaluated. The feature extraction in 3D and the following classification represents a bottleneck in compute resources.

Mine-it is an integrative toolbox at the iBioS group for processing pieces of information (objects). With a general library of nodes, it is possible to create a large set of processes by chaining them up in different ways. It is used predominantly on the data mining sector and therefore provides nodes with classification capabilities. Additional nodes with image processing and movie processing capabilities are about to be added.

Since the workflows can be built and called by the *mine-it* integrative toolbox running on the application server, an interface from mine-it to the DKFZ backbone has to be created.

Model-based Mediation and Scientific Workflows

Bertram Ludäscher
San Diego Supercomputer Center
Ludaescher@sdsc.edu

How can a domain scientist run meaningful queries across a set of databases which appear to be completely unrelated to the non-expert? Traditional database integration techniques, e.g. schema integration approaches do not apply since the schemas do not provide any direct means for correlating the different data sources. However, for the scientist, the data are connected since the domain expert knows how to fill the knowledge gaps, e.g. relating data from animal studies with human studies through so-called animal models, or relating data across scales (molecular, cellular, species …).

Model-based Mediation (MBM) is an extension to traditional database mediation techniques in that sources export data at the level of conceptual models (ontologies, integrity constraints, OO models), and the mediator can use this information together with a (semi-) formal representation of the domain scientist's "glue knowledge" (a shared ontology) to run meaningful inter-source queries. The technologies to make this happen involve a mix of query rewriting and reasoning techniques, or more general, database and knowledgebase representation approaches. A third kind of technology needs to be brought into the picture to capture “scientific workflows” such as a promoter identification workflow which e.g. a molecular biologist is interested in.

The domain scientific data “integration” (not in the database sense, but in a broader sense) is a very rich source of technical computer science problems, e.g. query rewriting over sources with limited capabilities in the presence of semantic integrity constraints.