LBNL Report LBNL-52767

Data Management for the Biosciences:

Report of the NSF/NLM Workshop on Data Management for Molecular and Cell Biology

National Library of Medicine
Feb. 2-3, 2003

November 4, 2003

Editors:

H.V. Jagadish
University of Michigan
Ann Arbor, MI
jag@umich.edu

Frank Olken
Lawrence Berkeley National Laboratory
Berkeley, CA
olken@lbl.gov

Abstract

Technological, institutional, economic and budgetary changes over the past decade have transformed the life sciences to become increasingly "data rich". Hundreds of millions of dollars have been (and are being) spent to develop large biological information resources, e.g., the human genome sequence, protein structures, and assembling this information in public databases, e.g, Genbank, PDB. Data management tools to facilitate access and analysis such data are necessary to obtain the full benefits of the investments in collecting these large datasets. Sequence data would be of little use if confined to publication in traditional print media.

Our conclusion is that data management technology has not kept pace with data generation in biology. We believe that further research and development of data management technology is needed to effectively utilize and exploit the large biological datasets which are now becoming available.

This is the report of a workshop held on Feb. 2-3, 2003 at the National Library of Medicine, Bethesda, MD on Data Management Technology for Molecular and Cell Biology. The workshop web site is: `http://www.lbl.gov/~olken/wdmbio/`

The workshop summary report and many of the workshop white papers appeared in the journal *OMICS – A Journal of Integrative Biology*, volume 7, number 1, Spring 2003 published by Mary Ann Liebert, Inc. publishers. Detailed information and citations can be found in Appendix A.

Disclaimer

Writing Committee

The writing committee was responsible for drafting the report subsequent to the workshop.

- Russ Altman (Stanford)
- Susan Davidson (U. Penn.)
- Barbara Eckman, (IBM Life Sciences)
- Michael Gribskov (SDSC, PDB)
- H.V. Jagadish, (Univ. of Michigan)
- David Maier (Oregon Health & Science Univ.)
- Frank Olken (Lawrence Berkeley National Lab)
- Meral Ozsoyoglu (Case Western Reserve Univ.)
- Louiqa Raschid (Univ. of Maryland)
- John C. Wooley (UC San Diego)

Workshop Participants

Russ Altman (Stanford)
Carol Bean (NLM)
Helen Berman (PDB,Rutgers)
Phil Bernstein (Microsoft)
Anthony Bonner (U. Toronto)
Jim Cassatt (NIGMS)
Milton Corn (NLM)
Peter Covitz (NCI)
Judy Cushing (Evergreen)
Susan Davidson (U. Penn.)
Barbara Eckman (IBM)
Jordan Fiedler (Mitre, DARPA)
George Garrity (MSU)
Peter Good (NHGRI)
Michael Gribskov (SDSC, PDB)
Amarnath Gupta (SDSC)
Joachim Hammer (U. Florida)
Chris Hogue (U. Toronto)
John Houghton (DOE OBER)
Mike Huerta (NIMH)
H.V. Jagadish (U. Michigan)
Gary Johnson (DOE OASCR)
Toni Kazic (U. Missouri)
Peter Karp (SRI)
Jessie Kennedy (Napier Univ.)
Eugene Kolker (BIATECH)
Michael Liebman (U. Penn.)
Steve E. Lincoln (Informax)
Zoe Lacroix (Arizona)
Peter Li (Celera)
David Maier (OHSU)
Victor Markowitz (Genelogic)
Mike Marron (NCRR)
Alexa McCray (NLM)
Dan Miranker (UT Austin)
Richard W. Morris (NIDA)
John Norvell (NIGMS)
Frank Olken (LBNL)
Jim Ostell (NCBI)
Meral Ozsoyoglu (Case Western)
Jignesh Patel (U Mich.)
Vijay Pillai (Oracle)
Julia Rice (IBM)
Stott Parker (UCLA)
Louiqa Raschid (U. Maryland)
Peter Schwarz (IBM)
Karen Skinner (NIDA)
Ambuj Singh (UCSB)
Bruno Sobral (Virginia Bioinformatics Instt.)
Sylvia Spengler (NSF)
Chris Stoeckert (U. Penn.)
Gary Strong (NSF)
Bhavani Thuraisingham (NSF)
Dick Tsur
Mark Tuttle (Apelon)
Zhiping Weng (BU)
John Westbrook (Rutgers)
Gio Wiederhold (Stanford)
John Wooley (UC San Diego)
Maria Zemankova (NSF)

Acknowledgements

The organizers would like to thank the National Science Foundation, Directorate for Computer and Information Science and Engineering for being the primary funder of this workshop, providing travel and other financial support, through grant EIA-0239993. The National Library of Medicine generously hosted the workshop and provided us with conference support services. The Department of Energy, Office of Science, provided one of the organizers with support for the organizing and editorial work via the VIMSS GTL Project at LBNL under Contract No. DE-AC03-76SF0009 and via the Synechococcus GTL Project via subcontract from Sandia Corp. IBM Corporation sponsored some of the meals.

Gary Strong at NSF was the program manager for the workshop. Maria Zemankova, through the IDM program in NSF CISE, has been a long-term activist towards making data management researchers pay attention to the life sciences, and thus played a key role in instigating the ideas that led to this workshop. In fact, a breakout group discussion on Bioinformatics at an IDM PIs meeting in May 2002 was a direct precursor to this workshop. We would also like to thank Sylvia Spengler (NSF Biology Directorate) and Bhavani Thuraisingham (NSF CISE) for their constant and continuing support.

We are grateful to Milton Corn (NLM) not only for arranging support from NIH, but also for the magic he worked to resolve many logistical issues as we prepared for the workshop. The NIH conference and audio-visual staff were of great assistance in handling the local arrangements and mechanics of the conference.

We wish to thank Eugene Kolker, the editor of OMICS, who arranged to have the summary report and many white papers published quickly as a special issue of OMICS – A Journal of Integrative Biology, vol. 7, no. 1, Spring 2003, and to have the publisher provide *free* access to the papers on the web. We would also like to thank the publisher of OMICS, Mary Ann Liebert, Inc. for their assistance with the publication and providing free web access to the papers.

Kevin D. Keck provided many helpful comments on drafts of this document.

Contents

Executive Summary

Over the past 15 years we have witnessed a dramatic transformation in the practice of life sciences research. We have already selected many of the proverbial low hanging fruit of dominant mutations and simple diseases. Chronic and more complex – non-monogenic – diseases, as well as efforts to design microbes for engineering needs or to uncover the basis of genetic repair, need the ladder of IT to reach the higher branches in living systems. At the same time, technological improvements in sequencing instrumentation and automated sample preparation have made it possible to create high throughput facilities for DNA sequencing, high throughput combinatorial chemistry for drug screening, high throughput proteomics, high throughput metabolomics, etc. In consequence, what was once a cottage industry marked by scarce expensive data obtained largely by the manual efforts of small groups of graduate students, postdocs, and a few technicians has become industrialized and data-rich, marked by factory scale sequencing organizations (Joint Genome Institute, Whitehead Institute, The Institute for Genomic Research, Celera, etc.).

Such industrialization and the accompanying growth in molecular biology data availability demand similar scale up and specialization in the data management systems that support and exploit this data gathering. Data management tools can interface molecular and cellular data to image and physiological data, which will be important to scale across the levels of living systems and particularly to translate the findings of basic biology to human health care. Similarly, public health depends on ability to integrate query and model with very diverse, very fragmented, non-standard, distributed data sources and databases.

We expect that this explosive growth in the amount and diversity of biological and biochemical data will continue into the 21^{st} century, i.e., that 21st life sciences will be data-rich. Success in the life sciences will hinge critically on the availability of computational and data management tools to analyze, interpret, compare, and manage this abundance of data. Increasingly, much of biology is viewed as an information science, concerned with how cells, organisms, and ecological systems encode and process information in genetics, cellular control, organism development, environmental response, and evolutionary settings.

Instruments, data, and data management systems are complementary goods, i.e., their joint consumption is much more useful than consuming a single commodity at a time. Consider the limited utility of genomic sequence data, if we could only publish such data in books and manually compare the sequences. The availability of data management software that permits the rapid searching of large genomic sequence databases

for similar sequences greatly enhances the utility of such sequence data. Quick sequence comparison routines are not sufficient by themselves: the fact that many (most) of these sequences have been collected into a few databases (e.g., Genbank) greatly simplifies the comparison task.

To obtain the full benefit of the massive public investments in generating biological data will require commensurate investments in effective data management systems and judicious choices of how to assemble and manage shared databases. To turn the vast amounts of new information being generated through scientific experiments into knowledge that can be applied towards better practice in medicine, agriculture, and environmental science, the federal agencies need to encourage a profound, deep partnership between experimental biology and database management.

Orchestrating fruitful interdisciplinary research across biology and data management is not easy. Lack of sufficient interaction between biologists and data management researchers can easily lead to attempts to reinvent well-known data management technologies by bioinformaticists, or sterile pursuits of irrelevant (or misunderstood) problems by data management researchers. Also the time scales of data management research and development are often incompatible with the production requirements of ongoing biological laboratories or community databases. It is striking to note that the major human genome sequencing centers have generally not been major sources of innovative data management technology. The most fruitful endeavors have often come from data management (or computer science) research groups with looser collaborations with biologists.

A sustained program by the federal agencies at the frontier between biology and data management technology will allow us to share the database expertise of the IT community with the large number of experimentalists supported across the federal agencies. There are needs for both fundamental research in database management technologies as well as their applications to biological problems. Funding agencies will have to set up appropriately staffed review panels charged with suitable review criteria for funding such interdisciplinary work. Adequate funding for small, medium and large-scale collaborative research projects as well as including funding within those collaborative projects to train a new generation of database management experts in the labs of IT professionals will be important.

For fastest progress in the biological sciences, we must encourage both the development of content for biological databases as well as data management technology for managing this content. We must recognize that database content development and database technology development are two complementary but quite different endeavors. Funding for the two must come in two different colors so that it is not easily possible to move money from one to the other. Otherwise, the pressing needs of today's content will too frequently triumph over technology's promise of a better tomorrow. Most research-driven companies recognize this tension and fund (at least some of) their research activity from corporate sources rather than through product divisions. In similar fashion, funding agencies should create a supplemental funding program for data management specialists to collaborate with life scientists in developing superior data management technology for life science applications.

We expect, in the foreseeable future, that it will become important to have MDs and

experimental biologists trained in computational methods (just as training in microbiology has now become routine for doctors where it was completely absent only a few years ago). Biology is often an exercise in induction (generalization from many instances), whereas computer science is more often a deductive enterprise, because computer algorithms and systems are usually designed (not evolved) artifacts. Solution to a specific biological data management problem is less of interest to a computer scientist than the generalization of this problem to a class of data management problems, all of which can be solved in one fell swoop through an appropriate computational advance. And rightly so, since this paradigm is significantly more cost-effective in the domains to which it is applicable.

This dichotomy has significant repercussions not just on how we undertake research activities, but also in how we train scientists. Currently, some biological scientists get trained in performing specific computational tasks, such as sequence analysis. Knowing how to select Blast parameters is not a transferable skill, in that it is likely to have little value if a new computational method is devised that is superior to Blast. What we need is training in the underlying principles so that a completely new and different sequence matching technique can be utilized rapidly and effectively. To this end, we need opportunities for people at every level to train themselves in the "other discipline" and work at the interface between data management and biomedical science. We also need support for curriculum development. The funding for such activities has to be ongoing for a substantial period of time – a typical three-year cycle is not enough to see the sort of major changes required.

Workshop Report

Chapter 1

The Need for Information in Biomedical Science

Biomedical research is now information intensive; the volume and diversity of new data sources challenges current database technologies. The development and tuning of database technologies for biology and medicine will maintain and accelerate the current pace of innovation and discovery. There are four main classes of situations in which data management technology is critical to supporting health-related goals. They are:

1. The rapid construction of task-specific databases to manage diverse data for solving a targeted problem.
2. The creation of data systems that assist research efforts by combining multiple sources of data to generate and test new hypotheses, for instance, about diseases and their treatments.
3. The management of databases to accumulate data supporting entire research communities.
4. The creation of databases to support data collection and analysis for clinical (and field) decision support.

To make these situations concrete, we present below examples that indicate both current successes as well as future opportunities and challenges in this regard.

1.1 Rapid construction of task-specific databases

1.1.1 Four Corners Hantavirus Outbreak

Identifying new pathogens used to take months to years. The identification of Legionnaires' disease and AIDS pathogens are cases in point. However, in 1993, when healthy young people in the American southwest began to die from an unknown pathogen, the virus responsible was identified in only one week using a combination of molecular biology and bioinformatics approaches. Traditional immunological approaches were only able to suggest that the virus involved in this "Four Corners" epidemic was distantly related to a family of viruses known as hantaviruses – not enough information to prevent or treat the disease. DNA sequences of related viruses in the hantavirus family were

retrieved from DNA sequence databases, and allowed the design of molecular probes (PCR primers) which were used in the first definitive test for the virus (confirming it as the pathogen) and allowing the determination of the DNA sequence of the new virus. In turn, the DNA sequence allowed the identification of the new virus's closest relatives (in viruses found in Korea), which shared similar animal vectors (rodents) and produced similar symptoms.

Because the Four Corners hantavirus produces symptoms that resemble those of cold or flu before progressing to pulmonary arrest and sudden death, the assay developed based on sequences found in DNA sequence databases was critical in stopping the spread of this epidemic. If this information had not been available – online, well described, and searchable – it might have taken several years and many, many deaths before this pathogen was identified.

In the intervening ten years, electronic data resources have continued to grow, leading to ongoing challenges in building the kind of integrated, online resources needed to attack similar diseases. We have much more information today, but also greater challenges in locating what we need. With the imminent threat of bio-terrorism, every day spent in obtaining the requisite data in response to a new outbreak can make a difference of thousands of lives. The 2003 SARS threat underlines this need.

1.1.2 World Trade Center Victim Identification

After the tragedy of September 11, 2001, the coroner's office in New York City had the task of identifying the remains of victims, so that they could be returned to family members. Existing database systems were built predominantly on the assumption that individual remains would be found and identified on a one-by-one basis. The possibility of more than 3000 victims and tens of thousands of samples was never considered in the design of the initial database system. GeneCodes Inc. has published its experience (still ongoing) in creating a data management system to assist in the identification of remains [SJ03]. This data management system had to be built on very short notice, and had to integrate information from a variety of sources.

There are two sources of DNA in tissues: nuclear and mitochondrial. Each of these sources has a number of attributes that can be measured, and the combination of these attributes tends to be unique for individuals, thus allowing identification. Given a sample of known origin (taken from the personal effects of the victims, and gathered from their families), it can be compared with the profile of attributes gathered from the unknown samples, and matched. In many cases, additional evidence is required, including DNA samples from parents and siblings (who share some, but not all DNA attributes with their relatives), information about where the remains were found, information about what personal effects were used for identification, and the contact information about all the people who are reported as missing.

To manage these data, the investigators built a complex system using cutting-edge database technology and state-of-the-art understanding of how to use genetics and other evidence to identify victims. The resulting tool continues to evolve, but has assisted in the identification of many victims, and the return of their remains to loved ones. Although this database was built under extraordinary circumstances, the need for urgent

assembly and integration of data, and the provision of novel analytic capabilities based on this data occurs routinely in both biomedical research as well as in the delivery of healthcare. When these needs arise, it is too late to perform essential background research in order to support these efforts, and so these needs must be anticipated in order to respond in a timely manner to urgent needs.

1.2 Databases to assist in research

1.2.1 Malaria Studies

The malaria parasite, *Plasmodium Falciparum*, is responsible for nearly 11 million deaths annually of children under the age of five. One of the great scientific achievements of 2002 was the publication of the full genome (the DNA sequence) of both the parasite as well as the mosquito (*Anopheles Gambiae*) that carries it to human victims, and the first public release of the full genome database (PlasmoDB, [Pla03]). For the first time ever, we have the complete triad of genomes involved in this disease (the parasite, the vector mosquito, and the human host). A primary health goal is to develop new drugs to effectively treat (and perhaps eradicate) malaria as a major threat to human health.

The genome database provides the list of the genes that are present in the parasite, but does not organize these genes into the pathways and networks of interaction that could be used to understand the underlying "wiring" of the parasite and how it works. Fortunately, there are other databases, including the MetaCYC database [Met03] of metabolic pathways, that can be used to assemble the genes into the metabolic machine that makes the parasite run. With a clear picture of this machine, we are able to identify vulnerable regions that can be targeted for interference with new drugs. In order to validate these targeted metabolic capabilities, we use other research databases (revealing where and when genes are turned on and off, including micro-array databases and proteomics databases) in order to prioritize the possible targets and assess their likelihood of success. Given a set of genes that would be good targets, we can further filter them by comparing them to human genes in order to help ensure that the new drugs will not be toxic for human use.

In some cases, the gene targets are proteins with known three-dimensional structures (or strong similarity to known structures), stored in the Protein Data Bank (PDB) [PDB03], and in those cases we can explore the detailed atomic structure of these genes, and use databases of existing compounds (such as Chemical Abstracts Service [CAS03]) in order to get a detailed understanding of how a potential drug might actually interact with its target and what modifications might make the drug more potent.

At the end of this pipeline, then, we will have a relatively small set of candidate for further drug development that have been filtered using disparate information sources, each of which provides a unique type of information. The resulting drugs can then be tested experimentally, and the process of drug discovery has begun, taking full advantage of all relevant data sources up front, thus decreasing the time to useful new drugs.

1.3 Long term storage of accumulated data to support discovery

1.3.1 Analysis in Breast Cancer

Breast cancer is a multi-component disease that appears to reflect both genetic and potentially environmental factors. Genetic linkage of mutations in the BRCA1 gene have been associated with high risk for breast cancer and appear to be predominant in women who develop breast cancer pre-menopause. Sporadic breast cancer, which includes 90 percent of all cases, tends to occur post-menopause. Conventional pathologic staging can readily identify cases of both type of breast cancer at seemingly equivalent stages of progression. It is of research interest to be able to evaluate the two disease processes to determine if they are identical. Such differences, if they exist, may be critical for enhancing diagnosis and staging of patients and the development of appropriate diagnostics and biomarkers, therapeutics, treatment options and outcomes.

For instance, suppose tissue samples are obtained from patients' biopsies and staged by pathology as being stage IIA. Some of the patients are known to have a BRCA1 mutation that is most likely linked to their early disease (less than age 40) while the other group is truly post-menopausal (greater than age 60) and not likely to be related to BRCA1 mutations.

1. Comparative Genome Hybridization (CGH) micro-arrays using a BAC (Bacterial Artificial Chromosome) clone library can be used on each sample to identify regions of deletion and amplification at the genomic level.
2. Within the set of BRCA1 patients, regions of common amplification and deletion are noted across the data set.
3. Similarly, within the sporadic breast cancer patients, regions of common amplification and deletion are noted across the data set.
4. The sets of common amplifications and deletions within each patient group are compared to identify those regions that are common across the patient groups.
5. The common regions need to be analyzed to examine genes that are within the chromosomal regions and require expansion of the regions to incorporate flanking regions because the BAC's are not end-sequenced. This requires both algorithmic processing for flanking regions as well as analysis using the physical map.
6. These genes should be further compared to expression array analysis data as well as genotyping data that may exist and the potential presence of single nucleotide polymorphisms (SNP's).
7. Those genes that are identified in these regions require association with molecular pathways and determination of potential interaction among the pathways included. This requires analysis of the graph representation of the pathways for linkages, direct and indirect, among the pathways. This defines the common processes across the patient types within a specific stage of breast cancer.
8. Within each patient group, the additional regions of deletion and amplification require similar analysis to identify genes and potential pathways associated uniquely to that patient group.

9. Potential biomarkers need to be identified based on the pathway associations within each patient group to assess whether the associated pathways are either causative or responsive to the common pathway perturbations. These biomarkers can serve as potential diagnostics for early disease detection as well as yielding information about the possibility that disease progression within these two patients groups is different irrespective of the similarity of staging.

1.4 Long term storage of data to support clinical decision making

1.4.1 Failure-to-Thrive Studies

Every year, thousands of children fail to grow properly. What pediatricians call "failure-to-thrive" (FTT) has many causes, most prominently metabolic disorders The incidence is quite high: for hospitalized children under the age of two, 1 – 5 percent have FTT; and 10 percent of children whose families have medical, social, or economic problems present the syndrome.

Precise diagnosis offers the best hope of treatment but can be notoriously hard: for some mechanisms, less than ten cases world-wide have ever been diagnosed, let alone recorded in the literature For these cases, clinicians send email via one of several different metabolic disease listserves, asking each other if they have seen a similar case, requesting advice on which assays to perform (and asking which laboratories perform those assays), and attempt to form an hypothesis and plan a treatment strategy by discussing the case. In many instances, we have no known therapies. This "email grand rounds" is certainly better than nothing, but we believe that advanced database technology can enable us to do much, much better.

In our vision, clinicians would query a consortium of databases containing information on syndromes, cases, biochemistry, genetics, endrocrinology, physiology, laboratory analyses, treatments, and outcomes, looking for cases similar to the one they have right now in the clinic. As they identified exact, near, or non-matches, these data would be scooped up automatically, fully anonymizing the data to protect patient privacy, so as to continuously record the incidence and causes of FTT. Based on the results of this initial pattern discovery query, the databases would generate sets of possible hypotheses and the results of assays and suggested therapies and their contraindications for each hypothesis would be quantitatively modeled so that the clinician could consider these.

Many of the assays, such as metabolic heavy-atom tracer experiments, require sophisticated mathematical models to interpret and analyze their results, and using these models to explore different ideas and test hypotheses now requires both a firm understanding of the clinical side and the mathematics. Today, designing a treatment regimen relies heavily on clinical intuition and experience, and is very much a trial-and-error process. In the future, the clinician would "test-drive" proposed treatments by simulating them, playing with the generated simulations or designing his or her own through interfaces. As treatment proceeded, the results would be reported back automatically to the database consortium via the local electronic medical record, again taking full care to completely anonymize the data.

1.4.2 HIV Studies

The HIV virus has caused roughly half a million deaths in the United States, and there are 42 million HIV infected persons worldwide [NIA02]. One of the features of this virus that makes it very difficult to treat is its ability to rapidly mutate in order to become resistant to medications. This mutation is manifested by changes in the genome of virus that are captured in the blood stream of individuals. In order to understand which sequence changes explain resistance, we must correlate the sequences found in the bloodstream with the history of exposure to medications, and the history of response to the medications. Thus, HIV researchers have created databases in which the history of medication regimens and their effectiveness for individual patients are stored along side the HIV viral sequences that were present before and after treatment (for example, [HIV03].

Analysis of these data allows researchers to do two things. First, they can look at statistical trends in the data in order to recognize which genetic alterations correlate with resistance to particular drugs. Second, they can use this knowledge to make decisions for individual patients about which drugs are likely to be best, based on the history of drug exposures and the responses of other, similar patients to different drug regimens. In this way, each patient benefits from both detailed understanding of the HIV virus in their own bloodstream, as well as the community-experience with different approaches to treatment.

1.5 Data management needs

Considering the scenarios above, several data-related needs are seen. Perhaps the need that is most immediately evident is the requirement for effective integration of data from multiple data sources. Such data integration is technically difficult for several reasons. First, the technologies on which different databases are based may differ and do not interoperate smoothly. Standards for cross-database communication allow the databases (and their users) to exchange information. Second, the precise naming conventions for many scientific concepts (such as individual genes, proteins, drugs) in fast developing fields are often inconsistent, and so mappings are required between different vocabularies. Third, the precise underlying biological model for the data may be different (scientists view things differently) and so to integrate these data requires a common model of the concepts that are relevant and their allowable relations. This reason is particularly crucial because unstated assumptions may lead to improper use of information that, on the surface, appears to be valid. Fourth, as our understanding of a particular domain improves, not only will data change, but even database structures will evolve. Any users of the data source, including in particular any data integrators, must be able to manage such data source evolution.

When a scientist obtains data from any electronic source, even if there is no data integration involved, many of the questions above remain. How is the data source structured? What is the model and underlying assumptions of the data provider? How did the data provider obtain this data – what is the "provenance" of the data? (For instance, in the World Trade Center example above, where an object was found would be

quite different from where it was at the time of the blast, which could also be different from where it was at the time of the tower collapse. An attribute such as "location" has to be interpreted accordingly.) If the data is a direct report of experimental observations, what were the precise experimental conditions? (For example, a scientist, working on the breast cancer scenario above, who finds an interesting database entry on phenotypes associated with the BRCA1 gene may need to know the age distribution of the population studied – and this information may not be mentioned in a typical terse database entry.) If data is derived from other sources, what was the derivation process? How reliable is the data? Is there a likelihood of error? (For example, this sort of annotation is crucial for the statistical odds the physician needs to associate with hypotheses when faced with a failure-to-thrive diagnosis scenario. Similarly, statistical trends are required in HIV studies to be able to correlate genetic alterations with resistance to particular drugs.)

The types of data representation and queries can also present challenges. The requirements in life sciences are often different from what is typically needed by business data processing for which commercial databases are designed. (For instance, metabolic pathways are important to represent and access in a cancer research scenario. Similarly, 3-dimensional structure representation is required to find ligand docking sites required for drug discovery studies.) New types of queries present an additional set of challenges – commercial databases expect keyword-based or predicate match queries for equality and range predicates; much richer query types are frequently required for biological data. While several sequence similarity tools are in wide use, there are many other types of similarity searches. (For instance, finding similar cases in the failure to thrive scenario above, or structural matching for proteins and drugs.)

To summarize, there are many data management needs to address a wide range of biological and health-related efforts. Some of these are evident as obstacles to scientific progress if not addressed. Others are opportunities for much faster progress, if capitalized upon. While no individual bio-medical research project may have the resources to make the necessary data management advances on its own, the combined need of multiple health-related research efforts makes the development of biological data management technologies a critical element of the national research infrastructure.

Chapter 2

Introduction

To study the issues surrounding the management of data for life sciences, a workshop was held at the NIH campus in Bethesda, MD on Feb 2 and 3, 2003. Approximately sixty experts in a range of related disciplines participated in this event. See the list of participants at the front of this report. This report represents the central recommendations of this group of experts.

Integration and exchange of data within and among organizations is a universally recognized need in bioinformatics and genomics research. We begin this report with a study of some of the obstacles to effective data integration, in the next section. We continue thereafter with a discussion of a few central research challenges in data management, addressing which will make a huge difference to biological science research. While most of these research challenges will require several years of work to be addressed fully, we believe that benefits can start accruing almost immediately from partial solutions that are generated in response to these challenges. We present our analysis of risk and time line associated with these efforts, and conclude with recommendations for specific actions that should be taken now.

Our bottom-line conclusion is that effective information management is crucial to rapid advancement in the life sciences. While there are incredible opportunities at the interface of life sciences and computer science, exploiting these opportunities requires an understanding of the differences between the two fields and the careful crafting of symbiotic mechanisms.

2.1 Overcoming Obstacles to Data Integration

By far the most obvious frustration of a life scientist today is the extreme difficulty in putting together information available from multiple distinct sources. A commonly noted obstacle for integration efforts in bioinformatics is that relevant information is widely distributed, both across the Internet and within individual organizations, and is found in a variety of storage formats, both traditional relational databases and non-traditional data sources (e.g., text data sources in semi-structured text files or XML, and the results of analytic applications such as gene-finding applications or homology searches). This syntactic heterogeneity is currently being addressed by two

main approaches: data warehousing (e.g., [Com], [VCC+03], and data federation (e.g., [HSK+01], [DCB+01], [Won00a], [TKM99]. In practice, a hybrid of the two approaches is generally able to handle most syntactic integration needs.

Arguably an even more critical need in data integration is to overcome semantic heterogeneity, i.e., to identify objects in different databases that represent the same or related biological objects (genes, proteins, etc.), and to resolve the differences in database structures, or schemas, among the related objects [KS99]. The same protein sequence is known by different names or accession numbers (synonyms) in GenBank [BKML+00], and SwissProt [OMG+02]. The same mouse gene (broadly understood) may be identified and represented as a genetic map locus in the Mouse Genome Database (MGD) [BRB+02], as the aggregation of multiple individual exon entries in GenBank, or as a set of EST (Expressed Sequence Tag) sequences in UniGene [WCL+02]. Its product has a single protein entry in SwissProt, and perhaps a structure entry in Protein Data Bank (PDB) [PDB03], which may reflect a slightly different amino acid sequence. Semantic integration also deals with how different data sources are to be linked together. For example, according to documentation at the Jackson Lab web site [JAX03], MGD links to SwissProt through its marker concept, to RATMAP [Gro02] through orthologies, to PubMed [WCL+02] through references, and to GenBank through either markers (for genes) or molecular probes and segments (for anonymous DNA segments). Finally, a schema element with the same names in two different data sources can have different semantics and therefore different data values. For example, retrieving orthologues to the human BRCA1 gene in model organisms from several commonly used web sites yields varying results: GeneCards [RCCPL98] returns the BRCA1 gene in mouse and C. elegans; MGD returns the mouse, rat and dog genes; GDB [TKM99] returns genes from MGD and FlyBase [GCM+97], [Gen97]; and LocusLink [WCL+02] returns only mouse.

In this discussion we consider integration of the results of back-end analysis packages such as BLAST [AGM+90], as well as more traditional data sources. Many scientifically relevant queries involve joins between the input or output of BLAST and other data sources (e.g., GenBank). We have chosen not to address integration of front-end tools such as visualization packages in this report, even though integrating them is an important goal. Their integration obstacles and potential solutions, while overlapping, are not co-extensive with obstacles and solutions for integrating back-end data sources (e.g., declarative optimized query languages). Finally, we consider integration in the context of read-only systems.

2.2 Obstacles to Integration

Four major categories of obstacles currently make integration of biological data difficult: syntactic and semantic issues; issues around the evolution of data sources; sociological factors; and systems issues.

2.2.1 Syntax and Semantics

Biological data sources differ widely, in both their syntax and their semantics: their format, structure, the meaning of their key concepts, and the query capabilities they support. Documentation of the contents and structure of data sources is often missing or incomplete, and often there is a mismatch between the actual database and its documentation, either because the documentation is outdated or because mistakes are made in specifying complex database constraints. There is little to no formal specification of database schemas and semantics, APIs, the semantics of operators, etc. Biological knowledge is often represented only implicitly, in the shared assumptions of the community that produced the data source, and not explicitly via metadata that can be used either by human users or by integration software. Identifiers are often not shared by multiple data sources, leading to the need to discover relationships among objects in multiple data sources, and to maintain synonym tables to map among them. Many biological data sources do not use controlled vocabularies, making them difficult to query. Many data sources require extensive cleansing and transformation before they are optimally useful for querying: for example, GenBank is sequence-centric instead of gene-centric, and contains legacy functional characterizations of sequences that are frequently incorrect. Finally, machine processable documentation of measurement units is often lacking or incomplete – for example concentrations may be specified as ratios (parts per million) without indicating whether they are mass, molar, or volume ratios. Adherence to standard S. I. (metric) measurement units can reduce problems of heterogeneous units.

2.2.2 Evolution of data sources

Biological research is a fast-paced, quickly evolving discipline, and data sources evolve with it: new experimental techniques produce more and different types of data, requiring database structures to change accordingly; applications and queries written to access the original version of the schema must be rewritten to match the new version. Incremental updates to data warehouses (as opposed to wholesale rebuilding of the warehouse from scratch) are difficult to accomplish efficiently, particularly when complex transformations or aggregations are involved. Finally, insufficient attention is paid to data provenance: e.g., where did the characterization of a given GenBank sequence originate? Has an inaccurate legacy annotation been "transitively" propagated to other similar sequences? What is the evidence for this annotation? (See sections below on provenance and on data evolution.)

2.2.3 Sociological issues

Several sociological issues appear to stand in the way of effective integration. The intense competition for ever-dwindling resources in the form of grants produces incentives to data providers to make integration difficult. If a small effort makes it too easy to access its database, the fear is that a large database effort will simply absorb its data, making it harder to argue that the continued existence of the small effort is necessary. Holding back critical information can give the scientist data owners an edge over their

competition; hence, key data items are often missing. Intellectual property (IP) issues become more difficult when data is shared, since it is harder to establish IP claims over ones data if it has been downloaded to other, related databases. Community deposition and curation of data is critically important for biological research, but monitoring and ensuring the quality of data in such a scenario can be very difficult. (No one wants to contribute to databases that do not keep track of data provenance and credit the contributor.)

2.2.4 Systems Issues

Internet sites, in particular academic and government sites, whether central resources, e.g., NCBI [NCB02], EMBL [EMB02], ExPASy [Swi02], or boutique databases, e.g., EpoDB for the erythropoetic cascade [CJSSBO99], SCDb for stem cells [PEI+00], GPCRDB for G-Protein Coupled Receptors [HVC01], are of critical importance to research in biology and bioinformatics. Genomics prides itself on its long tradition of publicly funded, public domain data, including GenBank, the Human Genome Project [HGM03], and the WashU-Merck EST sequencing project [WES95, HLB+96]. But performance on the Internet can be unpredictable, and therefore any integration approach that accesses sources via the Internet inherits this unpredictability. Furthermore, data sources on the Internet lack a common query interface, and there is no single directory of data sources on the Internet that an application can use to automatically identify and access Internet data sources.

2.3 A Continuum of Integration Approaches

As we attempt to address the issues raised above, it should be noted that database integration is not necessarily a monolithic enterprise, but rather comprises a continuum of approaches, from very simple to very complex and powerful. At one end of the continuum is a system that accesses a single source and fetches a single page or entity from that source. Next is a system that accesses multiple sources via relatively limited access methods, e.g., via web services. In the relative center of the continuum are systems that provide declarative, optimized query access over multiple sources that are mutually semantically compatible, i.e., sources whose central concepts (gene, protein, etc.) reflect a common understanding. Above average in difficulty and complexity is a system that provides declarative, optimized query access over multiple semantically heterogeneous sources. At the far end of the continuum is an idealized system that actively identifies data sources of interest, automatically overcomes syntactic and semantic heterogeneities wherever it discovers them, and provides transparent declarative, optimized query access over all of the sources.

The continuum of solutions based on the degree of integrated access, integration across data sources, and semantic heterogeneity, is as follows:

1. Point-to-point object (fetch) from multiple sources and integration across the access methods supported by these sources.
2. Distributed computation access to multiple semantically compatible sources.

3. Distributed declarative optimized computation over multiple semantically heterogeneous sources.

The farther we are able to move along this continuum, the better our ability to advance the science. Existing approaches for data integration fall short of the ultimate goal. We briefly review the advantages and disadvantages of some of the more popular solutions that have been deployed in the biological enterprise and identify some example systems.

Scripts written in Perl or Python are the most common solution. The drawbacks of these solutions are well known and include the difficulty of maintaining and re-using scripts, especially as the underlying data sources evolve. Source evolution is discussed in a later section. More important, scripts provide no support for the incorporation of data management and data analysis tools. As the size of data sets increases, scripts are unable to provide reliable and efficient access.

Data warehousing is also a popular solution to create data repositories for specific tasks. Data warehousing entails the pro-active collection of data from multiple sources into a single site. By replacing query access to many different data sources with queries to a central warehouse, data warehousing permits more rapid and reliable access to warehoused data. Other advantages include the availability of tools for data cleaning; and support for privacy and security. Weaknesses are that data warehousing technology traditionally supports only the relational data model and (R)OLAP ((Relational) Online Analytical Processing) and is arguably not optimally suited for the complex and variable structure of biological data. There is little support to resolve semantic heterogeneity across sources. Further, data warehousing solutions cannot utilize complex search and query processing services, e.g., BLAST or search engines, hosted at remote servers, nor can they explore the increasing number of hyperlinks and annotations that are frequently added by data curators. Finally, the greatest drawback of data warehousing solutions is that data in the warehouse becomes stale and must be refreshed. Data sources such as GenBank or PubMed are constantly being updated. This dynamism can impose substantial burdens to propagate these updates into the data warehouse.

A more recent solution that has been adopted by the biological enterprise includes a variety of architectures for federated access or mediation. Strengths include the ability to provide reliable and efficient access to remote data sources that are accessible over wide area networks. These solutions are more tolerant of semi-structured data since they are typically built on DBMS platforms that are not always limited to relational data models. A major advantage is that they can exploit complex search and computational services hosted at remote servers. As with data warehousing solutions, federation or mediation also does not provide many tools to process complex or semi-structured data. There is little support to resolve semantic heterogeneity across sources. Finally, such solutions may fail when remote servers are inaccessible. Data warehousing and federated database management fundamentally confront the same problems of data integration. Data warehousing does eager evaluation, federated database perform lazy evaluation.

Web services technologies, such as WSDL (Web Service Description Language) [CGMW02] and SOAP (Simple Object Access Protocol – actually an XML-encoded remote procedure call protocol), are gaining increasingly wide adoption in commercial settings. Such web services technologies could provide an infrastructure for standardized

registration and invocation of retrieval (or processing) services across the World Wide Web. These web services technologies are based on widely deployed HTTP servers and XML encodings. They can provide a convenient platform for data federation.

The Semantic Web is an effort by the World Wide Web Consortium, DARPA, and related researchers and developers to move web content from text content directed primarily at human readers toward documents with more rigorous semantic specifications intended for machine processing. This effort encompasses a variety of efforts in the development of knowledge representation languages (e.g., RDF, DAML+OIL, and OWL), development of formal ontologies for various domains (e.g., Gene Ontology (GO)), and related software tools (e.g., description logic based inference engines), and rules systems (e.g., RuleML and related systems). The semantic web efforts are directly concerned with the most difficult question in data integration - understanding, representing, and communicating the semantics (meanings) of data. Ongoing ontology efforts in the biology community (GO, Biopax, ...) are adopting these technologies.

Finally, workflow management systems (WFMS) offer a potentially attractive paradigm for specifying complex (or repetitive) scenarios of biological data analysis involving combined querying of databases and computations over the retrieved data. WFMS targeted at data retrieval and computation are known as "scientific workflow systems", as distinguished from WFMS deployed for managing laboratory activities (usually called LIMS (Laboratory Information Management Systems)). WFMS also offer the prospect of automatically recording information regarding the provenance of derived datasets.

2.4 Content Development Policies

While the most effective solutions can only be developed over several years, through addressing the research questions in the next section, there are short-term steps that could be taken by content providers to facilitate database interoperation, which we address next. In a nutshell, our recommendation is that creators of biological databases improve their operating procedures to utilize best practices for database development and dissemination, and thereby render individual databases into well-behaved citizens in an interoperable database infrastructure for molecular biology. Funding agencies should adopt a set of new review criteria for grants that fund data set creation, where those criteria are aimed at increasing the use of best practices for database development. In addition, funding agencies should develop policies for enforcing use of these best practices through other means in addition to the review process.

A good starting point for best practices guidelines might be the following guidelines adapted from the bioinformatics core guidelines from the GLUE Grant RFP [NIG]:

- What are the data release policies and what are the associated intellectual property issues?
- How will the data be available to the scientific community? Will there be browser access, formats for downloading complete data sets, on-line computational aids, etc.? We recommend a common format, such as XML, WSDL/SOAP.
- What is the nature and structure of the data? Present the plans to date for ontologies, schema, or other data models. Schemas must be well-documented,

and the documentation kept up to date. Explain provisions for documenting measurement units for every data element.

- What is the underlying structure of the database, e.g., relational, object-oriented, etc.?
- What is the mechanism for communication (both computational and human) between the distributed sites and the database managers? Will there be data liaisons? What are the key interacting databases? How will the data be linked?
- How will progress be available to the public, e.g., will lists of the systems being analyzed be available?
- What experience in bioinformatics is available in the group, and what resources can the consortium draw on?

Chapter 3

Data Types and Queries

3.1 Priority Data Types

One of the most striking features of biological data is the great diversity of data types used. Relational DBMS traditionally provide a handful of scalar data types (Booleans, integers, floating point, strings, date-time) and one collection type constructor (sets and perhaps bags). Object-oriented databases provide the capability of constructing a richer type system, but typically ship with a limited set of standard data types similar to that of relational DBMSs. However, nearly all OODBMSs come with a richer set of bulk types, e.g., arrays, sets, and symbol tables. Here we discuss some of the top priority biological data types both for individual data elements and for collection types, directing most of our attention to data types that have not been well supported by conventional DBMSs.

3.1.1 Sequences

The availability of sequence data, e.g., DNA, RNA, amino-acid sequences (proteins), has grown explosively over the past decade with the development of automated sequencing machines and large scale sequencing projects such as the human and mouse genome sequencing projects.

Sequences (DNA, RNA, amino acid) are presently often stored as text strings, but this representation is awkward when we want to annotate sequences, since text strings typically lack addressability at the level of individual letters (nucleotides, or amino acids). Often DNA sequences include not only individual nucleotides, but also gaps, usually with a length (or bounds on length) specification of the gap.

3.1.2 Graphs

Another common type of biological data is a graph, which could be a directed (or undirected) labeled graph, nested graph, or a hypergraph. Examples of this type of data include various biopathways (metabolic pathways, signaling pathways, and gene regulatory networks), genetic maps (partial order graphs (i.e., directed acyclic graphs),

taxonomies (either trees or DAGs), chemical structure graphs, contact graphs (for 3D protein structure), etc. See discussion in [Olk03].

Biologists are interested in performing a variety of comparative analyses and pattern matching queries against biopathways databases. Such queries are the analogs of similar queries in sequence databases. Graphs are easily stored in existing DBMSs, e.g., relational DBMSs. However, many graph queries, e.g., subgraph isomorphism, subgraph homomorphism, and subgraph homeomorphism are difficult (or impossible) to pose and answer efficiently in existing relational DBMSs, which know nothing of graphs. See additional discussion below and in [Olk03].

Sequences can be viewed as linear directed graphs (with nucleotide labels for the nodes). Multiple sequence alignments can then be described as partial order graphs. Such representations have been used as the basis of efficient multiple sequence alignment algorithms [Lee03].

Laboratory protocols (for molecular biology labs, clinical labs, etc.) can be modeled as workflow process model graphs. This could be used to support formal representation and querying of laboratory protocols, and automated support of lab protocols (Laboratory Information Management Systems (LIMS)). Formal representation of laboratory protocols and experimental conditions is also often needed for subsequent data analysis, e.g., of microarray data.

3.1.3 High-Dimensional Data

High-dimensional data sets are of increasing importance in molecular biology. Most of the this data arises from microarray experiments of gene expression. It is not unusual for these experiments to involve thousands (or tens of thousands) of genes and hundreds (or thousands) of experimental conditions and samples. Hence the datasets are arrays of spot intensities over the Cartesian product of genes and samples (e.g., experimental conditions). Often researchers are interested identifying clusters of genes which exhibit similar (or opposite) patterns of gene regulation. Specialized data structures and clustering algorithms are needed to support nearest neighbor, range searching, and clustering queries in high-dimensional spaces.

3.1.4 Shapes

Three dimensional molecular (protein, ligand, complex) structure data is another common data type. Such data includes both shape information (e.g., ball and stick models for protein backbones) and (more generally) scalar and vector field data of charge, hydrophobicity, and other chemical properties which are specified either as functions over the volume of a molecule or complex, or over the surface.

3.1.5 Scalar and Vector Fields

Scalar and vector field data is normally thought of primarily in the context of spatio-temporal applications such as computational fluid dynamics, weather, climate, oceanography and combustion modeling. However, a number of participants of the workshop argued that such data is quite important for molecular and cell biology applications.

Examples include modeling reactant and charge distribution across the volume of a cell, calcium fluxes across the cell surface or cell volume, reactant or protein fluxes across cell membranes, transport across cellular compartments, clinical response to drugs. Efforts in the visualization, computational fluid dynamics, and geographic information systems communities to deal with vector and scalar field data have focused on the development of fiber bundle or vector bundle data models. See the discussion of fiber bundle data models in [Tre03].

3.1.6 Temporal Data

Temporal data of various types (e.g., scalars, vectors, etc.) is also another prominent class of data types when studying the dynamics of biological systems. Examples include cellular response to environmental changes, pathway regulation, dynamics of gene expression levels, protein structure dynamics, developmental biology, and evolutionary biology.

Temporal data is critical for incorporation of the analysis of biological processes that occur over time. This can be applied to a variety of problems, ranging from stages of development of a cell or organism to the impact of aging on establishing the background for accurate disease diagnosis.

Temporal data in biological settings can either be *absolute* or *relative*. Absolute timestamping is common in administrative or long term ecological observational databases - time is recorded relative to an absolute global temporal coordinates such as UTC datetime. Relative timestamping records time relative to some event – e.g., cell division, organism birth, oncogenesis, diagnosis, cold shock, etc. Most implementations of time in the database community have focused on absolute time, whereas relative time is much more commonly used in most biological experiments. In complex settings such as disease progression, there may be many important events against which time is reckoned. The AI community has addressed many of these issues in temporal reasoning research.

3.1.7 Patterns

Much effort has gone into specifying, characterizing, and finding patterns (a.k.a. motifs) in DNA, RNA, and protein sequences. Of particular interest are regulatory sequences in genomic (DNA) sequences. Similar efforts are proceeding with respect to three-dimensional protein structure data, microarray data, pathways data, proteomics data, metabolomics data. In sequences these patterns are often represented as regular expressions or Hidden Markov Models (HMMs) or other types of grammars. Increasingly, biologists are interested in collecting, storing, and querying these patterns. Patterns thus need to be considered as first class data types, with support for storage and querying. Unfortunately, many queries will require testing for equivalence of variously encoded patterns (regular expressions, grammars, HMMs) - often a difficult matter. See additional discussion below under pattern matching.

3.1.8 Constraints

Historically, DBMS systems have provided mechanisms to specify and enforce a variety of logical constraints on the contents or allowable updates of the database, e.g., referential integrity constraints. Most of these constraints are fairly localized in scope. In recent years there has been increasing interest in rule-based systems (e.g., RuleML) for specifying and enforcing more elaborate logical constraints (logic programs).

Biological databases require a variety of constraint specifications, both logical rules, and mathematical constraints (e.g., equations or inequalities) as first class data types in a biological data management system, with the ability to store, enforce, and query such constraints.

Examples of mathematical constraints include various conservation constraints such as mass, momentum and energy conservation. Thus individual chemical reactions in a bio-pathway database must satisfy mass balance for each element. Such constraints are local. In contrast, cycles of reactions in thermodynamic database must satisfy energy conservation constraints. These are non-local (global) constraints. Another example of non-local constraints are the prohibition of cycles in overlap graphs of DNA sequence reads for linear chromosomes, or in the directed graphs of conceptual or biological taxonomies.

3.1.9 Mathematical and Statistical Models

Much of modern biological data analysis is concerned with the specification, development, parameter estimation, and testing (statistical or simulation) of various mathematical and statistical models of biological systems and datasets. Thus far the database community has largely been concerned with storing and querying input data sets, estimated parameters sets, and simulation output datasets. Relatively little attention has been paid to systematic methods of representing, storing, and querying the mathematical and statistical models being used. One would like to have declarative specification of mathematical and statistical models, means of recording bindings of model variables to database contents, and some way of recording the statistical analysis method (or simulation method) used.

There have been two major efforts in the systems biology community aimed at developing markup languages, e.g., SBML [SBM03] and CellML [Cel03], for the declarative specification and exchange of mathematical models of cells. These efforts have been primarily concerned with issues of expressiveness. Software efforts have focused on model entry, syntax validation, and simulation, not model storage and querying. ¿From a database standpoint, these models have largely been viewed simply as XML documents, not mathematical models. Treating models as mathematical objects would likely require some sort of computer algebra tools – e.g., to recognize that two models that use different variable names or measurement units are mathematically equivalent.

Model management systems have long been pursued in the operations research community, mostly in the context of large linear programming and nonlinear mathematical programming models. Similar issues arise in the estimation and testing of statistical models. The need for model management systems arises from the large size and complexity of the models being developed, the computation (simulation) of many versions

of similar models which often differ only in model parameters, and the development (and testing) of large numbers of models. The need for model management in systems biology will grow along with expanding research in the area.

3.1.10 Text

Text data types were seen as important, both for annotations, and permit inclusion and processing of biomedical literature along with other types of data. Existing text support in relational DBMSs was not seen as adequate to encompass the varied requirements of natural language processing. Note that the development of XML databases has seen increased greatly increased attention to (semi-structured) text data types. However, even XML databases have a ways to go before they have good support for text.

3.2 Priority Query Types

Conventional database applications are dominated by equality, range, and equi-join queries. In biological applications we see a much broader set of query types.

3.2.1 Similarity Queries

The single most popular type of query in molecular biology are similarity queries, most commonly sequence similarity queries, e.g., BLAST or Smith-Waterman sequence similarity queries. Such queries can be computed over DNA, RNA, or protein sequences. Similarity queries also arise on graphs (comparison of metabolic pathways), 3D protein structures, time series, high-dimensional data sets (e.g., microarray data), etc.

The popularity of similarity queries in biology arises from evolutionary biology. DNA sequences and hence the mRNA sequences and the proteins that they code for are all subject to random mutation and recombination. Many of these mutations will leave the function undisturbed, or will lead to genes or proteins with similar functions. Hence, it is often very useful to query for similar DNA, RNA or protein sequences in the hopes of finding similar genes or proteins with similar function. Comparative analyses of DNA sequences across diverse organisms (e.g., humans and mice) can identify conserved sequences that are often biologically important.

Some similarity queries, e.g., clustering microarray data, are performed over data (gene expression values) that can be viewed as vectors in high-dimensional spaces. Often, biologists must perform these queries over data which are not coordinate vectors. Common examples of non-coordinate similarity queries include sequences (due to insertion or deletion errors), biopathways graphs, and protein-structure data (3D shapes).

Similarity queries are typically computed with respect to either similarity or distance measures. We need the ability of users to specify which similarity or distance measure is to be used for a particular query. Some distance measures satisfy the triangle inequality, i.e., the direct distance between two points is never more that the distance via an intermediate point. Such distance metrics permit the use of special indices and pruning of the search processes. These techniques work for any distance measure that satisfies the triangle inequality.

Dissimilarity queries (e.g., outlier detection) were also viewed as important.

3.2.2 Pattern Matching Queries

A second class of queries consists of pattern matching queries, i.e., queries which find instances of sequences, etc. which match a specific pattern. On strings these queries involve pattern specifications such as regular expressions, Hidden Markov Models, or chart grammars. Graph pattern queries might involve patterns specified by graph grammars, subgraph homomorphism queries, etc.

One will also want to be able query collections of patterns (motifs). One such query would involve finding all patterns which match a sequence (the inverse of the customary query). Alternatively, one might ask for patterns which are similar to a specified pattern. Pattern similarity might be defined either structurally (akin to sequence similarity) or in terms of the overlap in the sequences matched by the two patterns from a specified database.

3.2.3 Pattern Discovery Queries

A third class of queries involve pattern discovery, elsewhere known as data mining. This includes the detection of frequently occurring patterns in sequences, graphs, 3D structures, etc. Such queries have been extensively treated in relational settings (in the database literature), and in sequences (sometimes in the database literature, mostly in computational biology). Little is known about pattern discovery in graphs. There has been work on 2D shape pattern discovery in the imaging and pattern recognition community, but less is known about shape pattern recognition algorithms in three dimensions, which is important for structural biology.

3.2.4 Spatio-Temporal Queries

Spatio-temporal queries form another important class of biomedical queries. One example use would be spatial genomics – the mapping of gene expression and protein abundance over fine-grained spatial scales (e.g., cellular or sub-cellular resolution) and time scales. One such query could ask for up-regulated genes (or proteins) in a particular anatomical or cellular region in a specified time interval following an experimental intervention (e.g., drug administration). Similarly, one could ask for gene co-regulation in both time and space. Other examples would include spatio-temporal queries of brain activity – from real-time MRI imaging. Note that there are well known problems of spatial registration of variable geometries across time and individuals – due to motion of organs (e.g., heart), microbial cell movement, organism development, or anatomical variation across populations (e.g., brain geometry).

Again we note that the primary market for such technologies are likely to be geographic information systems and computational fluid dynamics applications. In addition to spatio-temporal range queries, and iso-quant surfaces we envision that ultimately one would like to have a query language that supports interpolation and queries based on the vector calculus (e.g., queries that include div, grad, curl operators, line and surface integrals). Thus a query over a vector field of chemical transport data, might ask

to compute the integral of the chemical flux across a specified compartmental or cell surface.

3.2.5 Computational Queries

Classic relational database systems provide simple arithmetic operators over numeric data types, and comparison operators over a somewhat larger set of data types, e.g., date-time, strings, etc. Biological research entails much more complex mathematical and statistical operators. For example, we often commence the analysis of microarray data by computing the correlation matrix of some subset of the data. Queries on the correlation matrix are basis for clustering of genes and genetic regulatory inference.

Many computational biology scenarios involve intermixed database queries and computations. Thus we believe that the ability to specify computational workflows as a type of query would be highly desirable. Some database investigators have also viewed computational workflows as a mechanism for data integration.

3.3 Constraints and Constraint Enforcement

Biological databases utilize rules and constraints extensively, including those reflecting biological, chemical, and physical constraints; logical and temporal constraints; data and model validation constraints; equational, linear programming, and inequality constraints. Classical database constraints such as key and referential integrity constraints, triggers and some consistency checks (to a limited degree) are managed by current database systems. However, utilization of the new kinds of constraints, e.g., linear programming constraints used to model metabolic pathway constraints [EIP01], [PRP+03], [SVC02], need efficient, effective and scalable techiques to model and manage so that they can be incorporated into biological databases. Note that many physical, chemical, and thermodynamic databases have either (or both) localized and global constraints arising from conservation of mass, momentum, energy, and charge. Thus metabolic pathways databases need to satisfy conservation of mass (for each element) across chemical reactions. Such constraints have been used for consistency checking in the Ecocyc pathway database.

3.4 Research Issues

The research issues arising from the diverse data types and queries are the familiar DBMS design issues specialized to these data types and queries:

- What sorts of data structures (e.g., indices) are best suited for these data types and queries?
- What algorithms are efficient for processing such queries?
- How should such queries be expressed in a declarative query language?
- How should query optimization of such queries be done?
- How can we build extensible DBMSs to support diverse data types and queries?

- How can we make such a DBMS scalable for large databases and large numbers of users?

The issues above need to be addressed for individual data types and query types. There is also a need for research on methods of synthesizing individual data types and queries into a composite DBMS. While the issue was contentious, some researchers at the meeting believed that contemporary (e.g., commercial) DBMS have grown unwieldy, and it has become difficult to extend their functionality, e.g., by external research groups, despite mechanisms to support "data blades". Hence, we conclude that software engineering of DBMSs needs to return to the research agenda of the database community. Issues here include better methods of extending the query language to incorporate new data types and queries and better methods of integrating novel query operators and indices into the operation of query optimizers.

There is also a question of boundary definition for biological DBMSs. Where to stop? DBMSs are typically not very friendly development environments. What functions should we not attempt to include in the DBMS?

To facilitate progress in the development of such DBMSs the following sorts of infrastructure would be useful:

- Software testbeds for new access methods and query operators
- Public test data sets for software testing and benchmarking
- Synthetic database and query generators for testing and benchmarking
- Example queries

We note, by way of example, that the image processing and pattern recognition community has been developing a database of graph problems for testing and benchmarking graph-based image processing codes.

Chapter 4

Bio-aware design patterns

How do we use existing data models effectively in the bio-community? Simply placing data modeling tools in the hands of biologists and bioinformaticians doesn't guarantee the creation of a good information model. However, at this point there are a number of extant information models for certain paradigms - e.g. sequence and gene expression databases. Can we identify what is good in these and identify re-useable patterns in them? One outcome of this effort would be a catalog of common design patterns for bioinformatics databases, listing alternative representations for certain kinds of concepts. This pattern book could evaluate patterns in terms of their storage efficiency, complexity of implementation, ease of querying, amenability to integration, and extensibility. Data modeling and schema evaluation tools could use these patterns and offer them as templates or "mix-ins", and ultimately application development tools could become pattern aware.

Some examples of patterns include:

- Similar steps of experimental protocols
- Taxonomies or other hierarchical domains
- Versioning
- Annotation
- Dimensioned value

It is also useful to collect examples of how common biological concepts have been modeled in particular databases, e.g. sequences and variants, mutations, sequence assembly and gene prediction, gene expression, protein interaction, metabolic pathways, protein abundance, and protein structure.

One way of packaging patterns could be with data entry and display widgets, along with standard APIs. A "design wizard" could be used to encode some knowledge of the design process, and give an order in which to make design decisions.

Chapter 5

Provenance, Pedigree, Lineage

Data collections have little utility if humans cannot judge their suitability for the problem at hand. Part of the meaning of data is not its semantics vis a vis the data model, but the process by which it arrived: the "how" and "why" of data. This information is what researchers use to judge the suitability of the data for a particular task. Biological data may come from experiments, either in vivo or in vitro; from computational techniques (*in silico*); from (human) interpretation of primary data; and so forth. As a result, it is difficult to judge how much to trust a particular data item: In Genbank, do you have enough information to trust any entry? Is it the case that you trust some parts of the entry more than others? Similarly, entries in pathways databases (such as KEGG, MetaCyc, WIT, DIP, BIND) combine information from a large diversity of sources, and contain computational information, experimental information, and as well information derived from the literature.

Unfortunately, much of the data available in biological databases today has little or no provenance information associated with it. This lack has been recognized by scientists, and we are making progress in small steps towards correcting it, such as the MIAME (Minimal Information About a Microarray Experiment) standard for data derived from micro-array experiments. However, even MIAME leaves out many details of the experiment that may be crucial to effective data interpretation. Yet there has been resistance to MIAME compliance because of the perceived up-front overhead of recording all this information.

To address these issues, we suggest the following steps:

- Capture metadata at original data entry, automatically, with low overhead (see subsection on Workflow below). Even with ease of recording, coercion (social or software enforced) may be required to ensure that adequate provenance information is recorded with data.
- Develop techniques to record for each single datum, where it comes from and why. This may be non-trivial when facts are not easily modeled as independent discrete units. As data is derived based on other data, keep track of provenance through the derivation process.
- Develop techniques to manage issues of efficiency as these derivation chains grow long and have many branches. In effect, we need to develop a "pathway chart" for data.

As we construct aggregate provenance for a fact derived from two or more basis facts, an important question to answer becomes "Do these two facts have an independent basis?" Effective means to answer such questions must be developed.

Biological data often admits interpretation and annotation. In typical data management scenarios, these annotations are often considered metadata. In the realm of biological data management it is critical to distinguish this information from data provenance information. In fact, it is worth noting that base experimental data (recorded experimental observation), and their associated provenance, should not change, whereas we can have many versions of interpretation and other derived data. Effective means are required to model and manage this sort of data versioning.

Occasionally, for instance to protect patient privacy as required by U.S. HIPAA (Health Insurance Portability and Accountability Act), it may actually be necessary to mask the provenance of data. Unfortunately, data provenance, once lost, can never be recovered. We must develop techniques that permit the recording of full provenance for the data, while masking out appropriate components of provenance, thereby meeting both the highest standards of privacy protection while permitting the maximum possible correlational use of data.

Further discussion of data provenance issues can be found at the web site(s) for the recent Workshop on Data Derivation and Provenance [BF02], [Zha02]. A second workshop on Data Provenance and Annotation will be held in Edinburgh in Dec. 2003 [BBF+03].

Chapter 6

Uncertainty

Biological data has a great deal of inherent uncertainty. Often, when a scientist says "A is a B" they mean "A is probably a B, because there is some (possibly substantial) evidence suggesting that such is the case". For example: Yeast 2-hybrid experiments for protein-protein interaction are known for producing many false positives. Data in GenBank is sometimes erroneously reported (for instance, there may have been only a partial protein recorded), and then is propagated when another scientist runs a Blast search against sequences in GenBank and reports matches against such an erroneous sequence.

For all of these reasons, it is important to recognize uncertainty in data recorded in biological databases. Standard database technology provides no support for uncertainty, since business-oriented commercial databases typically contain data that is certain. When biological data is processed by trained scientists, they "know" which data to believe based on what it says and how it was obtained and therefore may not need support for managing uncertainty. In fact, the issue of uncertainty and error is explicitly dealt with through a manual curation process, based primarily on the exercise of human judgment. As we move to automated processing of large amounts of data, the inability of computers to exercise human judgment can lead to errors that compound in unmanaged fashion (as in the GenBank example above).

To effectively manage biological data in this context, we need management of uncertainty in databases. There are many research challenges in this regard, and we list some central ones here:

There are many different dimensions of uncertainty in biology. We could have contaminants in an experiment, errors in sources from which the data was derived, inherent error in the experimental technique used, honest disagreements in interpretation (where two reasonable scientists may interpret the same data differently), and so on. We must model these sources of uncertainty in biology, and develop appropriate mathematical (statistical) representations of them. Often, scientists are not used to thinking quantitatively about errors and uncertainty. To help them validate data that is being produced, we should define standards for specific experimental processes and create reference data that can be used to calibrate error rates.

We have discussed above the need to preserve provenance information with data. This requirement by itself will only provide the raw information that a human would

need to understand the reliability of the data. Quantitative uncertainty annotation will permit automated processing and aggregation of uncertainty information. We must develop techniques to propagate such uncertainty annotation as we develop derived data sets.

Given the large size of biological data sets, and the expectation that these will grow ever larger with time, we must develop efficient techniques to query and manipulate data with uncertainty annotation. These techniques will include means to retrieve efficiently only data with certainty above some set threshold, in the presence of a great number of less likely data points, also recorded in the database; and a means to compute efficiently the derived uncertainty annotation as multiple large data sets are combined in some computational process to create a new (computationally derived) data set.

Finally, the creation of linkages between data items is a key part of *in silico* biology today. To the extent there is uncertainty in the data, one must expect corresponding uncertainty in the linkages between data. Appropriate technologies must be developed to model such uncertainty in data linkages, through support for "fuzzy pointers" or similar means.

Chapter 7

Workflow and Derived Data

Workflows and data derivations occur at all scales and multiple settings in biological research. At one end, for example, there are "industrial-scale" investigations, such as sequencing of a complete genome, with parallel sample preparation feeding a farm of sequencers and hundreds of computer processors working on sequence assembly. At the other end of the spectrum is an individual researcher at a workstation searching a public repository of protein structures and performing a series of computations and visualizations on the results. While the volume, multiplicity and nature of processing is quite different in the two cases, in both it is valuable to record the sequence of steps that lead up to the results produced. Such a process record has many uses. It helps assure replicability of an experiment or computation. It can form part of the provenance of the results, providing a means by which users can judge the reliability or applicability of data. These records in the aggregate can be used to diagnose quality problems and refine experimental protocols.

The issues identified in this area, however, do depend somewhat on setting and scale. Large "production" laboratories often use a LIMS (Laboratory Information Management System) to track samples and processing steps. A limitation to their use is that they are relatively expensive to purchase and expensive to operate. They often require a "LIMS wrangler" to administer the database, update data entry forms, interface new instruments, and encode standard operation procedures (SOPs) and protocols. Development of automated or assisted LIMS wrangling systems is a research opportunity.

On a smaller scale, electronic laboratory notebooks ("e-notebooks") are now offered by several vendors and provide for flexible representation of laboratory steps, calculations and comments. However, their adoption at the bench-top has been limited, because of the need to stop and interact with a device to record information. Research on less intrusive interfaces, such as voice, touch screen or even barcode sensing could lead to more widespread acceptance and the ability to capture more of the early steps of the processing pipeline electronically. (One advantage to electronic recording of such information is that it can provide opportunities for automated capture of metadata.)

Explicit representation of inputs, parameters and procedures involved in the generation of data products has value at the level of an individual investigator as well as in a large-scale production line. Developing methods to separate a product definition or "product recipe" from the invocation of that recipe on particular inputs and parameters

would yield many advantages, including:

- Developing a data product by incremental manipulation and evaluation of a recipe.
- Allowing suites of related products to be created by editing and rerunning of a recipe.
- Enabling calculations of quality or uncertainty in parallel with production of data.
- Helping biologists take advantage of data and computational grids, where descriptions of computations to be carried out must be submitted to servers that schedule and distribute those computations (rather than being accessed through direct interactive interfaces).

Traditional data access (based on traditional data management technology) requires the specification of the query (e.g., SQL) or some application program that is to be evaluated against the data. Browsing, resource discovery, querying, computation, and complex computational workflow management are distinct and disjoint activities, not readily intermixed. This requirement is a limitation on the process of scientific discovery where the scientist wants the ability to express a workflow of potentially complex operators, each of which may have some domain specific semantics. An example is where the scientist wishes to gather a collection of proteins that has a maximal number of links to certain publications and are associated in a specific database with specific sequences. Thus, the development of a biological query language that can support the user as he or she browses the metadata, links, and query processing services of multiple sources, and allows the user to express a complex workflow and domain specific semantics corresponding to her task is critical. Without such support, the scientist will be hindered by the limitations of current query languages.

Chapter 8

Data Integration

The ultimate goal of data integration, as discussed in Chapter 2, is to be able to accomplish distributed declarative optimized computation over multiple semantically heterogeneous sources. This long term solution will provide support for integration over the contents of heterogeneous multi-modal sources whose data types include text, image, structured data, semi-structured data, graphs, results of computations. It is only at this point that we will have true ease-of-use in a manner that can adequately address the opportunities for life-saving impact described in the target scenarios in Chapter 1.

Several systems have been designed for domain specific integration of biomolecular data. BioKleisli [DOTW97], [BCD+98] and its extensions K2 [DCB+01] and Pizzkell/Kleisli [Won00b] follows a mediation approach and enables queries against integrated data sources. P/FDM [GK03] [KDG96] provides support to access specific capabilities of sources such as SRS [EUA96, EA93]. No semantic knowledge is expressed or utilized in either system. TAMBIS [BBB+98] is primarily concerned with overcoming semantic heterogeneity through the use of ontologies. It provides an integrated view of data sources but offers no ability to explore and exploit alternate identifiers and alternate links (paths). Garlic and its new extension for life science DiscoveryLink [HSK+01] encapsulate the access to specialized search capabilities into wrapper functions. While they provide extensive cost-based optimization to support efficient and seamless data integration, they too are hindered by the lack of knowledge about source capabilities as well as semantic knowledge about relationships among sources and their contents. The OPM multi-database system is based on the Object Protocol Model (OPM) [CM95] and object views [MCKS99]. While OPM provides the ability to evaluate complex queries, it too does not capture knowledge of semantic equivalences of scientific entity instances, links and paths. The Sequence Retrieval System SRS [EV97] applies full text indexing and keyword-based search techniques that are indeed very powerful. However, it is limited in that SRS was not designed to support semantic equivalences. For example, the SRS interface available at EBI offers the powerful capability of retrieving all sequences from EMBL that contain the keyword apoptosis in their description field (DE). However, an SRS query against both EMBL and MEDLINE no longer offers this powerful capability and is limited to full text search on apoptosis; thus the search on both data sources may return large numbers of irrelevant hits.

To summarize, no single existing technology appears to dominate all others for

purposes of biological data integration. Research may result in the development of new technologies, or suitable hybrids of two or more of the technologies described above, with some appropriate cost and benefit tradeoff. While long-term research objectives are important, their accomplishment depends upon progress along some of the other issues mentioned here, such as data provenance, source evolution, and the management of multiple models. As such, we recommend, in addition to the longer-term research, the development of short-term research milestones for data integration, particularly in the form of wrapper toolkits.

8.1 Toolkits for Data Providers to Quickly Wrap Data Sources

It is desirable to develop plug-and-play toolkits that are data provider friendly and can be used to create new data sources that are compliant with the requirements for successful integration, i.e., they provide data, schema and metadata information, the data has been cleaned or transformed as required for integration, etc. In addition to their use in creating new resources, wrapper toolkits may also be of some use in easing the overhead of wrapping existing legacy data sources. Features of toolkits include the following:

- Data transformation and data cleansing.
- Specifying APIs for computations services of data providers, for example, BLAST or search engines.
- Registration of sources, services and data types. Discovery of sources and services based on content, overlap of content, capabilities, as well as quality of their content ([MRV00, MRV01]).
- A default cost model, and the ability to plug in alternative cost models. An automated means of discovering cost-related statistics (metadata) such as table cardinality, column-value distributions, average execution time of sample queries, etc.

8.2 Managing multiple models

Mapping between different data models or data representations is an integral part of any biological database application. For example, there is often external information or archival data that must be imported to augment local computationally or experimentally derived data. Even within a single project, there can be the need for multiple models or representations for the same kind of information, as it moves through various stages, e.g., data entry, data query, data interchange, and data archiving. With Affymetrix gene expression data, for example, data entry may be what Affymetrix produces, data query may be a relational database with some local model, data interchange may use MAGE-ML, and data archive is what some consortium requires.

Issues that must be addressed to manage suites of models related to the same abstract information include the following:

- **Enabling mappings:** Models are typically large and complex, and the process of discovering and specifying how concepts in one relate to another can be laborious. Semi-automated approaches and mapping wizards could help with this task.
- **Verifying mappings:** Models often capture the semantics as well as the syntax of data. As mappings are performed, automatic checks for semantic conflicts between models should be enabled, e.g., a one-to-many relationship in model one being mapped to a one-to-one relationship in the other.
- **Semantic propagation:** As models evolve, changes in one model must "propagate" to the corresponding portions of related models.
- **Making data comparable and self aware:** Much biological data has been developed as a "silo", i.e., as a specialized resource with little attention to being interoperable with other data sources. However, no one data resource is all-encompassing and therefore requires linkage to other sources. One strategy that has emerged for making data comparable is the development of common naming schemes (thesauri, terminologies (controlled vocabularies)) and ontologies.

Data about scientific entities, e.g., genes or proteins, are stored in multiple sources. Each source captures some features (attributes) describing both the structure and function of the scientific entity. Typically, information about a single instance of some scientific entity, e.g., the gene TP53, may be found in multiple data sources. While there is overlap of data among sources, typically these sources are not replicates. Instead, each source references instances in other sources, and each source captures some information about the structure or function of the scientific entities. Under these circumstances, solving the challenge of data integration across multiple data sources, successfully, requires the acquisition of metadata about the contents and overlap of contents among these sources in order to correctly identify and completely characterize the structure and function of an instance of a scientific entity across multiple sources. Currently, neither data warehousing nor federated or mediated solutions provide adequate support to address this issue. Needed are:

- Tools to define and manage semantics.
- Tools for the resolution of semantic heterogeneity.
- Tools for ontology specification.
- Tools for ontology management.
- Scalable tools and methodologies (algebra) to integrate ontologies.

Another effect of data silos is that specialized domain knowledge frequently resides only in the minds of the developers and expert users. As data silos are made available to non-expert users and linked to other types of data sources, this implicit domain knowledge must be explicitly represented. Ontologies can also be used in this situation to enable the "self-awareness" of data. (For example, knowing that crystallographic coordinates are centered on an internal atom whereas protein structure coordinates are in an external coordinate space.) Making implicit knowledge explicit can be used to enable one database to explain itself to another.

If we do not know implicit assumptions then there is the danger of a data item crossing out of the context where these assumptions hold and being misinterpreted.

Despite the emergence of many useful ontologies (see SOFG [SOF03] and the Jan. 2003 issue of Comparative and Functional Genomics [Oli03]) and some technologies for

ontology representation (e.g. DAML-OIL), more work is needed on both models for representing and tools for managing taxonomies and ontologies, and to facilitate the automatic creation of linkages. In particular:

- Coping with clash: overlapping nomenclature and terminology may exit (homonyms). For example, independently discovered proteins in mouse and human may have been given the same name, but are completely different proteins.
- Managing the ontology life cycle: ontologies are initially developed for humans, then used by programs to make connections between data, then used for reasoning.
- Pluralistic versus unitary: multiple ontologies may exist for the same domain. Many of the same issues must be addressed as mapping between multiple models (schemas).

8.3 Evolution of Data Sources

Data sources evolve as knowledge changes and as new experimental techniques produce more data and different characterizations of the data. As a result, both the schemas that describe the data as well as applications and queries written specific to the original version of the schema must be updated. This is difficult to accomplish, particularly when the data types and structures are complex and when the analysis involves complex transformations or aggregations. Keeping up with evolution becomes significantly more difficult if there is a fundamental change in our understanding of the meaning or the characterization of the data.

The value of biomedical data increases as it is combined with other data. Thus, we address the issue of schema evolution within the greater context of data integration. We consider the impact of the evolution of a single source as it relates to data integration from other sources. Data sources may be integrated in the following manners:

- Mapping data values through a controlled vocabulary or ontology.
- Cleaning and transforming the data sources before importing the data into a warehouse.
- Wrapping the data sources with a commonly agreed-upon query language and defining views over the data sources.
- Wrapping the data sources using a commonly agreed-upon API.

Each of these approaches depends on the schemas that define each data source and on the "mapping", "transformation" or "wrapping" that is utilized to integrate them. Thus, if the schema of any one source were to change, it will have a negative effect on any applications or queries that access this source and on the correctness and successful operation of the data integration process. For simplicity, we use the generic term "mapping" to describe all of the above transformations to facilitate data integration.

Sometimes, a schema change is purely an extension that changes neither the format nor the semantics of any of the existing data. In this case, the problem is how to determine quickly that existing mappings can still be relied upon. This schema-matching problem is relatively straightforward, and where existing algorithms [RB01] would likely be quite effective.

Alternately, the schema change may entail a more radical reformatting. For example, it may partition a table into sub-tables or expand a scalar element into a compound structure. In this case, it is often useful to define a view that presents the newly formatted data in the old format, so that existing mappings can be executed without changing them. Today, this process is a manual process. Graphical tools are beginning to appear that simplify this manual process. Research is underway to partially automate this view construction, for example the Clio project at IBM [IBM03, Mil03, HMH01, MHH00, HSK+01] The process involves automated schema matching (to determine portions of the schema that are unchanged), schema analysis (to identify join clauses to combine schema elements) and data mining (to identify parts of the modified structure that store the same values as the old structure). A complementary approach is model management, which contains operators that can update mappings automatically based on changes to schemas that are incident with the mapping [Ber03, SM03].

Finally, the most extreme schema change is one that alters the semantics or our understanding of the data source. For example, one may add a data structure whose content semantically overlaps the existing data source but it is syntactically incompatible with the original representation in the schema. Other changes may require a modification of the actual contents of the source, not just the schema, in order to reconcile the data source with a controlled vocabulary or ontology. Since the data semantics has changed, this modification in effect creates a new data integration problem, and typically will require manual solution by a database designer.

8.4 Performance metrics and quality of service

A key objective of data integration is seamless and efficient access to remote sources. This requirement can be addressed by appropriate performance metrics that affect the quality of service. The first aspect of quality of service is the end-to-end latencies or delays associated with a computational task. The second aspect is support of the task of scientific exploration that reflects domain specific semantic knowledge and the quality of the results. Once the scientist has accomplished a process of discovery he or she is able to formulate a complex computational task to be evaluated across multiple remote data sources. Mediation, data warehousing, and workflow technologies are all suited to support a reliable and efficient computational platform for integration. Further research is needed to support learning the costs of query evaluation in noisy wide area networks (WANs); query evaluation with delayed, bursty or unavailable sources; and cost-based query optimization that can exploit the existence of multiple, alternative data sources and the complex search and query processing services hosted by remote servers.

Challenges to this task include the following: Difficulty of predicting access costs accurately. Learning and other techniques are needed to construct cost models ([GRZZ00], [KLMM03], [NKNV02]). A variety of optimization approaches are needed, e.g., performance targets; alternate sources; adaptive evaluation strategies by ([Hal00], [HFC+00], [ZRV+02] [PMT03]). In many situations, clients, especially automated clients such as crawlers, can overwhelm the computational capability of a data source. There is a need for servers to be able to advertise their service constraints and semi-automated mechanisms to enforce these constraints. An example of server constraints are those

published by NCBI for users of their E-Search utilities, which prohibit automated tools from accessing their servers during peak access periods.

Typically, query optimization with multiple alternate data sources makes the assumption that the results are independent of the particular source or query evaluation plan that is chosen. For biological data sources, while there is significant overlap of sources, few of the sources are exact replicas. As an example, the three sequence data sources, GenBank, DDBJ and EMBL, do not all contain the same data about DNA sequences. There has been some research on query evaluation with incomplete, imprecise or alternate but dissimilar sources, as well as flexible query answering and approximate query answering [DGL00], [NFL03], [Nau01], [FLMS99] Issues include the following: imprecise values or incomplete data sources with missing data or dirty data, unavailable sources, alternate sources and query evaluation plans with dissimilar semantics, e.g., result cardinality may vary or characterization of objects in sources may be different [LNRV03]. As was discussed in a previous section, both data provenance and data curation can vary, and this variability has a significant effect on the quality of the results. A simple example is an archival data source, which can contain obsolete data versus a (human) curated data source.

The challenge is query planning and evaluation for scientific exploration that can exploit domain specific semantics to provide answers that closely match the desired result quality and semantic requirements of the biological scientist or application. For example, a scientist who is exploring some hypothesis will very like be interested in reducing access latencies as (s)he explores multiple alternatives. However, for a validation task, a scientist would probably want to explore the results from all the relevant sources, despite the overlap of their content.

Traditional data access (based on traditional data management technology) requires the specification of a query (e.g., in SQL) or some application program that is to be evaluated against the data. This is a limitation on the process of scientific discovery where the scientist wants the ability to express a workflow of potentially complex operators, each of which may have some domain-specific semantics. An example is where the scientist wishes to gather a collection of proteins that has a maximal number of links to key publications and is associated in a specific database with specific curated sequences. Thus, the development of a biological exploration language that can support the user as he or she browses the metadata, contents and links, and query processing services of multiple sources, allows the user to express a complex workflow, and allows the user to specify their desired result quality and semantic requirements is critical.

Chapter 9

Benchmarking and Prototype Development

9.1 Benchmarking and Evaluation of Existing Approaches and Technologies

Existing technology solutions must be evaluated for their potential benefits for biological data management and integration. More important, the pitfalls and limitations of each solution when considering the specific challenges of biological data sources must be clearly identified. Specific challenges include the following:

- Partial replication of data across multiple autonomous sources;
- The lack of a common unique identifier for instances of the same scientific object across these sources;
- Challenges associated with dirty and incomplete data;
- Evolution over time of both data and schema as scientific knowledge is updated;
- Potential for efficient and seamless access to data from remote sources;
- Potential to exploit semantic knowledge.

This sort of benchmarking and evaluation is particularly important for understanding biological data integration since, over the last decade, there has been much activity in developing architectures and tools for this purpose. While these efforts were not targeted at the biological enterprise, several systems have been built to support biology. We must evaluate the strengths and weaknesses of these systems, paying particular attention to the impediments to successful data integration enumerated previously.

9.2 Prototype Development

We propose a series of prototype systems, corresponding to the series of challenge problems that have been described above. These systems will range from short-term prototypes, built primarily with current off-the-shelf technology, to extensive solutions that require new technology. It is important to invest both in short-term prototypes, which can bring value to biomedical science today, and in long-term prototypes, which can

bring much greater value to biomedical science tomorrow. A focus on one, to the exclusion of the other, will be counterproductive. For instance, consider the problem of data integration. Lightweight prototypes, such as MOBY and DAS, have already made an initial stab at the data integration problem. In the short term, we expect it to be feasible to support point-to-point object (fetch) from multiple sources and integration across the access methods supported by these sources. While such solutions do not solve issues of semantic heterogeneity and semantic mismatch, nevertheless, they can greatly benefit the biological enterprise. In the medium term, in addition to support for query evaluation, prototypes may be able to incorporate tools for schema and data transformation and tools for data cleansing, and thus become of greater use to the scientist. In the longer term, solutions to even harder problems, such as mismatch in nomenclature, can be addressed.

9.3 Time scales and risks for various research topics

We can classify research topics by the time scale over which they can be addressed, and the level of risk involved in addressing the topic. Often, the two attributes (time scale and risk) are correlated, with longer term project often entailing higher levels of risk. We have classified a number of topics below:

- Short term topics: data management support for molecules and molecular complexes – the study of individual molecules and complexes of a few molecules.
- Medium term topics: data management support for reactions, pathways, processes, physiology study of molecular interactions, signaling pathways, small networks (medium risk), biomarkers (medium risk)
- Long Term topics: data management support for large networks (higher risk) – whole cell, systems biology, understanding control mechanisms; for clinical Models – disease models, models of therapies

Our assessment of the required time and degree of scientific risk of this research agenda is given above. We expect a significant number of new queries on molecules and molecular complexes, from sequences to three-dimensional structures changing conformation over time, to appear in the short term. This area builds on an existing, large infrastructure of sequence and structure databases and algorithms, so that the initial problems to be solved are the design, implementation, and optimization of new queries more closely integrated with the databases. We perceive this shortest time frame to also be of relatively low risk.

In the medium term, we believe elementary queries of reactions, pathways, networks, and processes will be successful. Some of the graph-theoretic queries, such as path-finding and certain types of partitions, are of relatively low risk: efficient algorithms for the first, and algorithms for the second that are robust to cycles, are known. Similarly, numerical simulation of small systems of stiff, nonlinear, ordinary differential equations or small PDE systems (the most demanding cases) is for now largely a hand-crafted use of existing numerical packages that would seem amenable to at least partial scale-up and automation. We should also be able to progress from running Perl scripts over large data sets to the use of effective query interfaces made available over the web.

A special instance of a molecule is a biomarker. These serve as more easily detected indicators of a particular physiological state, such as a disease; of an organism, such as a pathogen; or of itself or another molecule, such as a toxin or antigen. Ideal biomarkers are highly sensitive to low concentrations of the detectant; robust to noise in the organism or environment; non-, or minimally, invasive; easily deployed; and cheap. Identifying and evaluating candidate biomarkers is of prime importance for public health, both for more rapid and accurate diagnosis and for early warning of biological and chemical weapons, and we view improving current methods for these tasks as a medium risk effort.

Higher risk is likely to occur either when data are absent and must be acquired or when the performance is poor. For instance, with biological data types and queries, poor performance is most likely to occur in four situations: when the queries are NP-hard, the graphs are inconvenient in topology (e.g., non-planar cyclic), the graphs are large (where "large" could mean as few as a hundred nodes and edges), and the equational systems expand past some threshold of size and complexity. Methods that exploit parallelization may prove relevant in this regard, but parallel graph algorithms for the types of graphs that occur most often in biology are in their infancy. Obviously parallel numerical algorithms and packages may prove useful, but we also expect significant research to be devoted to the development and evaluation of qualitative and semi-qualitative approximate models.

We believe the challenges of scale and structural complexity will be addressed over the longer term, especially in two key areas: the exploration, simulation, and understanding of much larger networks, such as at the level of the whole cell; and predictive models of clinical states, diseases, and therapies. Research in these areas is presently of the highest risk, in part due to our inexperience in integrating data and models that are very large and complex, and which differ in their mathematical structure and assumptions. Nonetheless, this area offers the greatest payoff, and sets the stage for further advances in the decade beyond.

Chapter 10

Related Work

The importance of data management to the biological sciences has been well-recognized in recent years, and there have been several efforts to address this issue. Many of these have dealt with concerns such as intellectual property, standards definition for a particular community, computational training for biological scientists, and such other concerns that are beyond the scope of the current study. We mention below three studies of particular relevance:

10.1 Report of the Workshop on Interconnection of Molecular Biology Databases (WIMBD), Stanford, CA, August 9-12, 1994

The workshop on interconnection of molecular biology databases pointed out the advantages of interconnection and interoperation of databases. Specifically, biological data are more meaningful in context and no single database supplies all context for any datum. For example, we better understand a gene when we know the function of its product, the sequence of the gene and its regulatory regions, the three-dimensional structure of its products, and the functions of evolutionarily related genes. These types of information are scattered across different databases. New biological theories and regularities are derived by generalizing across a multitude of examples, which again are scattered across different databases. Integration of related data enables data validation and consistency checking.

A number of non-technical barriers to interoperation were identified at the workshop:

- Workshop participants expressed strong resistance to standards, in part out of concern that standards stifle creativity, and because significant efforts are often required to modify existing software to conform to standards. Many existing molecular biology databases are not accessible via Internet query; similarly, many biologist users do not have Internet access. The semantic descriptions of many molecular-biology databases are terribly incomplete. Without an understanding of the semantic relationships among databases, interoperation is impossible. Few incentives now favor interoperation; funding and scientific credit often reward

efforts that distinguish themselves according to how they differ from prior work, rather than according to their compatibility with prior work.

Today, nine years later, significant progress has been made with respect to many of these points. Community efforts to develop standards are now prominent – in particular one can note the adoption of the mmCIF [BBM+97] standard for 3-dimensional structural data, the development of MIAME [MGE03] and MAGE [Cov03] standards for micro-array data, and the establishment of the Gene Ontology (GO) Consortium [Con00]. Most molecular biology data resources are now backed by relational database management systems and are accessible via the Internet. Problems identified in 1994 remain important; although formal specification of semantics is now widely recognized as important, many databases still lack such specifications. Cultural and sociological issues that discourage interoperation unfortunately remain largely unchanged.

10.2 Report of the NSF Invitational Workshop on Scientific Database Management, March 1990

This seminal workshop identified seven main issues, many of which remain important to this day. These issues were:

- **Metadata:**
  - Who did what and when.
  - Characteristics of experimental devices and processes.
  - Definitions of (computational) transforms.
  - Documentation and citations.
  - Structure and format descriptions.

  It is imperative that the metadata remain attached to the data for it to be meaningful.
- **Locating Data:** Early in any scientific inquiry, the need to find data becomes critical to the successful outcome of the investigation. Hypotheses need to be corroborated, or perhaps, archived data is to be mined for possible undiscovered properties. It becomes necessary to address questions such as: What data exists and where is it? Is the data relevant to my interests? Do useful data items exist? This need requires a general data browsing capability providing facilities first for locating data sets, and then for scanning them for indications of probable interest.
- **User interfaces:** To manipulate data and produce information, a scientist needs to access data and apply analysis tools in concert. Failure to integrate the data management and analysis environments restricts the productivity of the scientist.
- **More Flexible Representational Structures:** Perhaps the single unifying cry of the workshop was that existing data models are inadequate for science data needs. The relational model has some advantages. Chief among them is that it is well-defined and has solid theoretical underpinnings. And, more pragmatically, it exists within successful commercial products. However, the semantic gap between

the relational model and what scientists need must be addressed. We must seek alternatives such as extending the relational paradigm, object-oriented database technology, extensible tool kits, and logic databases. We must also consider alternatives to the relational model for efficiently supporting temporal, spatial, image, sequences, graph, and other more richly structured data.

- **Appropriate Analysis Operators:** One area of concern noted by most of the participants was the lack of appropriate operators within existing DBMS for manipulating the kinds of data encountered in scientific applications. For example, more flexible comparison operators are necessary when attempting to match DNA sequences or retrieve image data. There was not universal agreement as to where these operators belong – within the DBMS as intrinsic operators or external to the DBMS as utilities or part of an analysis package. The approach used now is to have a commercial DBMS export data for use by external utilities.

- **Standards:** Heterogeneity in data and operational environments is a fact of life. We must find ways to promote consistency within and across scientific disciplines. It is unreasonable to expect all disciplines to converge on some unifying standard, so heterogeneity will continue to be a force to be reckoned with.

- **Standards for Data Citation:** There was strong sentiment that data used in the conduct of an investigation should be cited prominently. A standard citation mechanism would allow other researchers to locate and examine precisely the data used in the investigation. It would also give due credit to the data collectors.

- **GenBank and PDB:** Two of the most visible success stories in terms of data sharing for biomedical research are GenBank and PDB. As new genes are discovered, and the sequence information available to science has exploded, GenBank has become a valuable central repository for information regarding known genes, including their DNA sequences. The Protein Data Bank has structural information regarding proteins and is a central research resource for organic crystallographers and structural biologists. Both of these valuable community resources have been created through a judicious combination of administrative pressure and social goodwill. While both are of great value, it is not hard to see how they could become even more valuable. For instance, much of the data in GenBank is not considered reliable enough so that many scientists create their own curated derivatives. Better tracking of provenance, and techniques to manage reliability, could make GenBank that much more valuable. Similarly, the data in PDB is very useful once a protein of interest has been identified. However, search facilities for structurally similar proteins would greatly enhance the value of PDB to a scientist. In the body of this report we identify several such technological opportunities that can lead to significant benefits for advances in biological science.

10.3 Report of Dagstuhl Seminar on Information and Process Integration: A Life Science Perspective

A seminar [AEF+03] was held at the Schloss Dagstuhl International Conference and Research Center for Computer Science, Dagstuhl, Germany on Jan. 28-31, 2003 on Information and Process Integration: A Life Sciences Perspective, shortly before our workshop.

Considerable emphasis was given to issues of semantic integration of different life sciences data sets, since the organizer believed that the semantic integration issues are much more difficult than syntactic issues.

The Dagstuhl meeting also discussed issues of process integration – encompassing issues of workflow management in life sciences laboratories and data analysis and protocol representation.

There was also discussion at the Dagstuhl meeting concerning integration of temporal and spatio-temporal data from the life sciences.

Chapter 11

Recommendations

11.1 Research Funding

A sustained program by the federal agencies at the frontier between biology and data management technology will allow us to share the database expertise of the IT community with the large number of experimentalists supported across the federal agencies. Funding agencies will have to set up appropriately staffed review panels charged with suitable review criteria for funding such interdisciplinary work. Adequate funding for small, medium and large-scale collaborative research projects as well as including funding within those collaborative projects to train a new generation of database management experts in the labs of IT professional will be important.

For fastest progress in the biological sciences, we must encourage both the development of content for biological databases as well as technology for managing this content. While the direct benefits are typically obtained from the content, it must be recognized that it is the technology that enables delivery of the relevant content, in the right format, at the right time.

Imagine the rate of scientific advance possible if DNA sequences had to be stored in books, in the absence of even the rudimentary data management facilities available today. We must recognize that database content development and database technology development are two complementary but quite different endeavors. Funding for the two must come in two different colors so that it is not too easily possible to move money from one to the other. Otherwise the pressing needs of today's content will too frequently triumph over technology's promise of a better tomorrow. Most research-driven companies recognize this tension and fund (at least some of) their research activity from corporate sources rather than through product divisions. In similar fashion, funding agencies should create a supplemental funding program for data management specialists to collaborate with life scientists in developing superior data management technology for life science applications.

One way to accomplish this end is by providing explicitly earmarked supplemental grants for IT development in association with standard grants for biological science. In this fashion, it will be possible to review proposals for such supplemental funding purely on the basis of the quality of the proposed IT research, yet ensure that it is conducted in close collaboration with the primary funded biological science effort.

To energize and focus research activity at this boundary of two disciplines, it is valuable to define challenge problems that push the boundaries of data management technology and if successful would enable major advances in biomedical science. Creation of test data sets and benchmarks towards this end are worthwhile endeavors in themselves, and should be supported as appropriate and possible.

11.2 Information Sharing Standards

The structural biology community and the genetics community have evolved strong mechanisms over decades to ensure sharing so that the richness of the data can be mined by all. For the rest of the life sciences, we need to accelerate the process. Data sharing approaches will need to be built into the structure of the entire publishing and grant processes. To be successful, the federal program managers will need to work with professional societies and journal editors to develop policies with teeth to enforce requirements on data standards and sharing at the time of grant funding and renewal. Current best practices should be required of anyone developing a data collection of any significance. Hence, we should expect that new data collection projects will use existing, community-based data exchange formats – e.g., some dialect of XML such as SBML – where feasible, rather than idiosyncratic data exchange formats. Database developers should also be expected to provide carefully specified and appropriately documented schemas in machine processable formats, as well accessibility across the web using standards such as SOAP and WSDL.

The sociological barriers to data sharing must be addressed, and technology can sometimes provide paths around some of these barriers. For instance, mechanisms to record provenance of data can make it possible to give credit to a contributor of data. Mechanisms to count data usage and accesses can make it possible to create for a data provider the social equivalent of a citation count for a research paper author. Unfortunately, caching, data warehousing, or the use of derived data will often mask references to the underlying data sets. Some of these issues have begun to be addressed in web caching protocols, wherein the cache manager propagates aggregated reference counts to the original data source. In commercial settings, proper reference counts drive advertising revenues.

11.3 Work force Training

We expect, in the foreseeable future, that it will become important to have MDs and field biologists trained in computational methods (just as training in microbiology has now become routine where it was completely absent only a few years ago). The addition of this computational training is likely to require a significantly greater effort than the addition of microbiology because of fundamental differences in the way knowledge is organized and imparted in computer science and the biological sciences. Biological objects (humans, plants, pathogens, cells, proteins) are enormously complex, but have underlying commonalities that an intelligent practitioner can benefit from. Experience with repeated instances, each slightly different, makes the practitioner that much more

able to deal with the diversity. A surgeon in the operating theater is able to draw upon experience with patients seen in the past to determine what to do with a specific patient with a specific tumor location the surgeon has never seen before. Computational artifacts, on the other hand, while simpler and more controllable than biological objects, are endlessly more diverse. There is no reason to expect two computational artifacts to behave in similar fashion unless they were explicitly designed to be so. A computer scientist will therefore establish a result once, in general, and never consider re-establishing it repeatedly for one instance at a time. Solution to a specific biological data management problem is less of interest to a computer scientist than the generalization of this problem to a class of data management problems, all of which can be solved in one fell swoop through an appropriate computational advance.

This dichotomy has significant repercussions not just on how we undertake research activities, but also in how we train scientists. Currently, some biological scientists get trained in performing specific computational tasks, such as sequence analysis. Knowing how to select Blast parameters is not a transferable skill, in that it is likely to have little value if a new computational method is devised that is superior to Blast. What we need is training in the underlying principles so that a completely new and different sequence matching technique can be utilized rapidly and effectively. To this end, we need opportunities for people at every level to train themselves in the "other discipline" and work at the interface between data management and biomedical science. Potential vehicles for delivering such training include conference tutorials, short courses, and summer schools. We also need support for curriculum development. The funding for such activities has to be ongoing for a substantial period of time – a typical three-year cycle is not enough to see the sort of major changes required.

Bibliography

[AEF+03] R. Apweiler, T. Etzold, J.-C. Freytag, C. Goble, and P. Schwarz. Information and process integration: A life science perspective, report of the dagstuhl seminar 03051. Technical report, Schloss Dagstuhl International Conference and Research Center for Computer Science, Jan. 2003.

[AGM+90] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–10, 1990.

[BBB+98] Patricia G. Baker, Andy Brass, Sean Bechhofer, Carole Goble, Norman Paton, and Robert Stevens. TAMBIS: Transparent access to multiple bioinformatics information sources. In Janice Glasgow, Tim Littlejohn, Francis Major, Richard Lathrop, David Sankoff, and Christoph Sensen, editors, *6th Int. Conf. on Intelligent Systems for Molecular Biology*, pages 25–34, Montreal, Canada, 1998. AAAI Press, Menlo Park.

[BBF+03] Dave Berry, Peter Buneman, Ian Foster, Michael Wilde, WangChiew Tan, and Yannis Ioannides. Data provenance and annotation. `http://www.nesc.ac.uk/esi/events/304/`, 2003.

[BBM+97] P. E. Bourne, H. M. Berman, B. McMahon, K. D. Watenpaugh, J. Westbrook, and P. M. D. Fitzgerald. The macromolecular crystallographic information file (mmcif). In *Methods in Enzymology*, volume 277, pages 571–590. Acadmemic Press, 1997.

[BCD+98] Peter Buneman, Jonathan Crabtree, Susan Davidson, Val Tannen, and Limsoon Wong. Biokleisli. In Stan Letovsky, editor, *Bioinformatics*. Kluwer Academic Publishers, 1998.

[Ber03] Phil A. Bernstein. Applying model management to classical meta data problems. In Mike Stonebraker and David DeWitt, editors, *Proceedings of the Conference on Innovative Data Research (CIDR)*, pages 209–220, January 5-8 2003.

[BF02] Peter Buneman and Ian Foster. Workshop on data provenance and derivation. `http://www-fp.mcs.anl.gov/~foster/provenance/`, Oct. 2002.

[BKML+00] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, Barbara A. Rapp, and David L. Wheeler. Genbank. *Nucleic Acids Research*, 28(1):15–18, 2000.

[BRB+02] JA Blake, JE Richardson, CJ Bult, JA Kadin, and JT Eppig. The mouse genome database (mgd): the model organism database for the laboratory mouse. *Nucleic Acids Research*, 30(1):113–115, 2002.

[CAS03] CAS. CAS: Chemical abstracts service home page. `http://www.cas.org/`, 2003.

[Cel03] CellML.org. What is CellML? `http://www.cellml.org/public/about/what_is_cellml.html`, 2003.

[CGMW02] Roberto Chinnici, Martin Gudgin, Jean-Jacques Moreau, and Sanjiva Weerawarana. Web services description language (wsdl) version 1.2 part 1: Core language. Technical report, World Wide Web Consortium (W3C), June 2002.

[CJSSBO99] Jr. Christian J. Stoeckert, Fidel Salas, Brian Brunk, and G. Christian Overton. Epodb: a prototype database for the analysis of genes expressed during vertebrate erythropoiesis. *Nucleic Acids Research*, 27(1):200–203, 1999.

[CM95] I-Min A. Chen and Victor M. Markowitz. An overview of the object-protocol model (OPM) and OPM data management tools. *Information Systems*, 20(5):393–418, 1995.

[Com] Computational Biology and Informatics Laboratory. Allgenes: a web site providing access to an integrated database of known and predicted human and mouse genes. (version 6.0, 2003). `http://www.allgenes.org`.

[Con00] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[Cov03] Robin Cover. Technology reports: Microarray and gene expression markup language (mage-ml). `http://xml.coverpages.org/mageML.html`, 2003.

[DCB+01] Susan B. Davidson, Jonathan Crabtree, Brian P. Brunk, Jonathan Schug, Val Tannen, G. Christian Overton, and Christian J. Stoeckert Jr. K2/kleisli and gus: Experiments in integrated access to genomic data sources. *IBM Systems Journal*, 40(2):512–531, 2001.

[DGL00] Oliver M. Duschka, Michael R. Genesereth, and Alon Y. Levy. Recursive query plans for data integration. *Journal of Logic Programming*, 43(1):49–73, 2000.

[DOTW97] Susan B. Davidson, G. Christian Overton, Val Tannen, and Limsoon Wong. Biokleisli: A digital library for biomedical researchers. *Int. J. on Digital Libraries*, 1(1):36–53, 1997.

[EA93] T. Etzold and P. Argos. Srs–an indexing and retrieval tool for flat file data libraries. *Computer Applications in Biology (CABIOS)*, 9(1):49–57, 1993.

[EIP01] Jeremy S. Edwards, Rafael U. Ibarra, and Bernhard O. Palsson. In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. *Nature Biotechnology*, 19(2):125–130, February 2001.

[EMB02] EMBL. Embl: European molecular biology laboratory, research in molecular biology. http://www.embl-heidelberg.de/, 2002.

[EUA96] T. Etzold, A. Ulyanov, , and P Argos. Srs: Information retrieval system for molecular biology data banks. In *Methods in Enzymology*, volume 266, pages 114–28. Academic Press, 1996.

[EV97] T. Etzold and G. Verde. Using views for retrieving data from extremely heterogeneous databanks. In *Pacific Symposium on Biocomputing*, 1997.

[FLMS99] Daniela Florescu, Alon Levy, Ioana Manolescu, and Dan Suciu. Query optimization in the presence of limited access patterns. In *Proceedings of SIGMOD, ACM Conference onf the Management of Data*, pages 311–322, 1999.

[GCM+97] WM Gelbart, M Crosby, B Matthews, WP Rindone, J Chillemi, S Russo Twombly, D Emmert, M Ashburner, RA Drysdale, E Whitfield, GH Millburn, A de Grey, T Kaufman, K Matthews, D Gilbert, V Strelets, and C Tolstoshev. Flybase: a drosophila database. the flybase consortium. *Nucleic Acids Research*, 25(1):63–66, 1997.

[Gen97] Genetics Society of America. Flybase: A database of the drosophila genome. `http://flybase.bio.indiana.edu`, 1997.

[GK03] Peter M.D. Gray and Graham J.L. Kemp. An expressive functional data model and query language for bioinformatics data integration. In P. M.D. Gray, L. Kerschberg, P. J.H. King, and A. Poulovassilis, editors, *The Functional Approach to Data Management*, pages 168–188. Springer Verlag, 2003.

[Gro02] RATMAP Group. Ratmap: The rat genome database. `http://ratmap.gen.gu.se/`, 2002.

[GRZZ00] Jean-Robert Gruser, Louiqa Raschid, Vladimir Zadorozhny, and Tao Zhan. Learning response time for WebSources using query feedback and application in query optimization. *VLDB Journal: Very Large Data Bases*, 9(1):18–37, 2000.

[Hal00] Alon Halevy. Special issue on adaptive query processing. *IEEE Data Engineering Bulletin*, 23(2), June 2000.

[HFC+00] J. M. Hellerstein, M. J. Franklin, S. Chandrasekaran, , A. Deshpande, K. Hildrum, S. Madden, V. Raman, and M. Shah. Adaptive query processing: Technology in evolution. *IEEE Data Engineering Bulletin*, 23(2):7–18, 2000.

[HGM03] HGMIS. Human genome project information. `http://www.ornl.gov/TechResources/Human_Genome/home.html`, 2003.

[HIV03] HIVDB. Stanford hiv drug resistance database. `http://hivdb.stanford.edu/`, 2003.

[HLB+96] L.D. Hillier, G. Lennon, M. Beckerand M.F. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello, W. Gish, M. Hawkins, M. Hultman, T. Kucaba, M. Lacy, M. Le, N. Le, E. Mardis, B. Moore, M. Morris, J. Parsons, C. Prange, L. Rifkin, T. Rohlfing, K. Schellenberg, and M. Marra. Generation and analysis of 280,000 human expressed sequence tags. *Genome Research*, 6(9):807–828, September 1996.

[HMH01] Mauricio A. Hernandez, Renee J. Miller, and Laura M. Haas. Clio: A semi-automatic tool for schema mapping. In *Proceedings of ACM SIGMOD Conference on the Management of Data*, page 607. ACM, 2001.

[HSK+01] L. M. Haas, P. M. Schwarz, P. Kodali, E. Kotlar, J. E. Rice, , and W. C. Swope. Discoverylink: A system for integrated access to life sciences data sources. *IBM Systems Journal*, 40(2):489–511, 2001.

[HVC01] Florence Horn, Gerrit Vriend, and Fred E. Cohen. Collecting and harvesting biological data: The gpcrdb & nucleardb databases. *Nucleic Acids Research*, 29(1):346–349, 2001.

[IBM03] IBM Corp. IBM Almaden Research Center, Clio Overview. `http://www.almaden.ibm.com/software/km/clio/index.shtml`, 2003.

[JAX03] JAX. Mouse genome informatics. `http://www.informatics.jax.org/`, 2003.

[KDG96] G.J.L. Kemp, J. Dupont, and P.M.D. Gray. Using the functional data model to integrate distributed biological data sources. In *Proceedings of the Eighth International Conference on Scientific and Statistical Database Systems*, pages 176–185, 1996.

[KLMM03] Craig A. Knoblock, Kristina Lerman, Steven Minton, and Ion Muslea. Accurately and reliably extracting data from the web: A machine learning approach. In Piotr S. Szczepaniak, Javier Sergovia, Janusz Kacprzyk, and Lofti Zadeh, editors, *Intelligent Exploration of the Web*, pages 275–287. Physica-Verlag, 2003.

[KS99] V. Kashyap and A. Sheth. Semantic similarities between objects in multiple databases. In A. Elmagarmid, M. Rusinkiewicz, and A. Sheth, editors,

Management of Heterogeneous and Autonomous Database Systems, The Morgan Kaufmann Series in Data Management Systems, pages 57–89. Morgan Kaufmann, 1999.

[Lee03] Christopher Lee. Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics*, 19(8):999–1008, 2003.

[LNRV03] Zoe Lacroix, Felix Naumann, Louiqa Raschid, and Maria Esther Vidal. Exploring life sciences data sources. In *Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03)*, 2003.

[MCKS99] V.M. Markowitz, I.A. Chen, A.S. Kosky, and E. Szeto. Object-protocol model data management tools '97. In Stan Letovsky, editor, *Bioinformatics, Databases and Systems*, pages 187–199. Kluwer Academic Publishers, 1999.

[Met03] MetaCyc. Metacyc: Metabolic encyclopedia. `http://metacyc.org/`, 2003.

[MGE03] MGED Society. Minimum information about a microarray experiment - miame. `http://www.mged.org/Workgroups/MIAME/miame.html`, 2003.

[MHH00] Renee J. Miller, Laura M. Haas, and Maurcio Hernandez. Schema mapping as query discovery. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, pages 77–88. Morgan Kaufmann, 2000.

[Mil03] Renee J. Miller. Clio. `http://www.cs.toronto.edu/db/clio/`, 2003.

[MRV00] George A. Mihaila, Louiqa Raschid, and María-Esther Vidal. Using quality of data metadata for source selection and ranking. In *Proceedings of the Third Workshop on the Web and Databases (WebDB)*, pages 93–98, 2000.

[MRV01] George A. Mihaila, Louiqa Raschid, and María-Esther Vidal. *Source Selection and Ranking in the WebSemantics Architecture Using Quality of Data Metadata*, volume 55, pages 87–118. Academic Press, July 2001.

[Nau01] Felix Naumann. *Quality-driven Query Answering for Integrated Information Systems*, volume 2261 of *Lecture Notes on Computer Science*. Springer-Verlag, 2001.

[NCB02] NCBI. National Cente for Biotechnology Information. `http://www.ncbi.nlm.nih.gov`, 2002.

[NFL03] Felix Naumann, Johann C. Freytag, and Ulf Leser. Completeness of information sources. In *Workshop on Data Quality in Cooperative Information Systems 2003 (DQCIS)*, 2003.

[NIA02] NIAID. Facts and figures: HIV/AIDS statistics. `http://www.niaid.nih.gov/factsheets/aidsstat.htm`, December 2002.

[NIG] NIGMS. Large-scale collaborative project awards rfa gm-02-007. `http://grants1.nih.gov/grants/guide/rfa-files/RFA-GM-02-007.html`. Also known as Large Scale Glue Grants.

[NKNV02] Zaiqing Nie, Subbarao Kambhampati, Ullas Nambiar, and Sreelakshmi Vaddi. Mining coverage statistics for websource selection in a mediator. In *Proceedings of CIKM*. ACM Press, 2002.

[Oli03] Steve Oliver (editor). Special issue on workshop on ontology for biology. *Comparative and Functional Genomics*, 4(1):1–168, Jan./Feb. 2003.

[Olk03] Frank Olken. Biopathways graph data manager (bdgm). `http://pueblo.lbl.gov/~olken/graphdm/graphdm.htm`, 2003.

[OMG+02] C O'Donovan, MJ Martin, A Gattiker, E Gasteiger, A Bairoch, and R Apweiler. High-quality protein knowledge resource: Swiss-prot and trembl. *Briefings in Bioinformatics*, 3(3):275–284, September 2002.

[PDB03] PDB. The protein data bank. `http://www.rcsb.org/pdb/`, 2003.

[PEI+00] R.L. Phillips, Brunk B. Ernst, R.E., N. Ivanova, M.A. Mahan, J.K. Deanehan, K.A. Moore, G.C. Overton, and I.R. Lemischka. The genetic program of hematopoietic stem cells. *Science*, 288(5741):1635–1640, 2000.

[Pla03] PlasmoDB. PlasmoDB: The plasmodium genome resource. `http://www.plasmodb.org/`, 2003.

[PMT03] Vassilis Papadimos, David Maier, and Kristin Tufte. Distributed query processing and catalogs for peer-to-peer systems. In Mike Stonebraker and David DeWitt, editors, *Proceedings of CIDR 2003, First Biennial Conference on Innovative Data Systems Research*, January 5-8 2003.

[PRP+03] ND Price, JL Reed, JA Papin, I Famili, and BO Palsson. Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices. *Biophysics Journal*, 84:794–804, Feb. 2003.

[RB01] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal: Very Large Data Bases*, 10(4):334–350, 2001.

[RCCPL98] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. Genecards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14(8):656–664, 1998.

[SBM03] SBML. SBML: Systems biology markup language. `http://www.sbml.org`, 2003.

[SJ03] James Sikela and Eric Juengst. Genetic diversity and DNA-based identification: A discussion panel. In *Pacific Symposium on Bicomputing*, 2003.

[SM03] P.A. Bernstein S. Melnik, E. Rahm. Rondo: A programming platform for generic model management. In *Proceedings of ACM SIGMOD 2003*, pages 193–204. ACM, June 2003.

[SOF03] SOFG. SOFG: Standards and ontologies for functional genomiccs. `http://www.sofg.org/`, 2003.

[SVC02] D. Segre, D. Vitkup, and G.M. Church. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences U.S.A.*, 99(23):15112–15117, November 2002.

[Swi02] Swiss Institute of Bioinformatics. Expasy molecular biology server. `http://www.expasy.org`, 2002.

[TKM99] Theodoros Topaloglou, Antony Kosky, and Victor Markowitz. Seamless integration of biological applications within a database framework. In T Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H-W Mewes, and R. Zimmer, editors, *Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB'99)*, pages 272–281, 1999.

[Tre03] Lloyd A. Treinish. A function-based data model for visualization. `http://www.research.ibm.com/people/l/lloydt/dm/function/dm_fn.htm`, 2003.

[VCC+03] Markowitz V.M., J. Campbell, I.A. Chen, A. Kosky, K. Palaniappan, and T. Topaloglou. Integration challenges in gene expression data management - chapter 10. In Lacroix Z. and Critchlow T., editors, *Bioinformatics: Managing Scientific Data*, pages 277–301. Morgan Kauffman Publishers (Elsevier Science), 2003.

[WCL+02] DL Wheeler, DM Church, AE Lash, DD Leipe, TL Madden, JU Pontius, GD Schuler, LM Schriml, TA Tatusova, L Wagner, and BA Rapp. Database resources of the national center for biotechnology information: 2002 update. *Nucleic Acids Research*, 30(1):13–16, 2002.

[WES95] AR Williamson, KO Elliston, and JL Sturchio. The merck gene idex, a public resource for genomics research. *Journal of NIH Research*, 7:61–63, 1995.

[Won00a] Limsoon Wong. The functional guts of the Kleisli query system. *ACM SIGPLAN Notices*, 35(9):1–10, 2000.

[Won00b] Limsoon Wong. Kleisli: Its exchange format, supporting tools, and an application in protein interaction extraction. In *IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE'00)*, pages 21–28, Nov. 2000.

[Zha02] Yong Zhao. Data provenance/derivation workshop position papers and talks. `http://people.cs.uchicago.edu/~yongzh/position_papers.html`, October 2002.

[ZRV+02] Vladimir Zadorozhny, Louiqa Raschid, Maria-Esther Vidal, Tolga Urhan, and Laura Bright. Efficient evaluation of queries in a mediator for web-sources. In *Proceedings of the ACM SIGMOD Conference*, June 2002.

Appendix A

Summary and Whitepapers Publications

The whitepapers and summary report were published in a special issue of the journal *Omics – A Journal of Integrative Biology*, Volume 7, Number 1, Spring 2003. The relevant portion of the table of contents for the issue is reproduced below. The actual papers may be obtained may obtained by any one of:

- Ordering a print copy of the entire issue of the journal from the publisher, Mary Ann Liebert, Inc, OR
- Online access to the articles in .pdf files, by clicking on the online version logo on the right hand side of the OMICS web page http://www.liebertpub.com/omi/. The publisher has graciously made these papers available for free.
- Online access to pre-print versions of the articles from the web site for the workshop. http://www.lbl.gov/õlken/wdmbio/

A.1 Summary Report

F. Olken; H. V. Jagadish. **"Guest Editorial: Data Management for Integrative Biology",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 1

H. V. Jagadish; Frank Olken. **"Database Management for Life Science Research: Summary Report of the Workshop on Data Management for Molecular and Cell Biology at the National Library of Medicine, Bethesda, Maryland, February 2-3, 2003",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 131-137

A.2 Whitepapers

Russ B. Altman. **"Complexities of Managing Biomedical Information",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 127-130

Helen M. Berman; John Westbrook. **"The Need for Dictionaries, Ontologies, and Controlled Vocabularies",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 9-10

Philip A. Bernstein. **"Applying Generic Schema Management to Bioinformatics",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 99-100

Adriane Chapman; Cong Yu; H.V. Jagadish. **"Effective Integration of Protein Data through Better Data Modeling",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 101-102

Peter A. Covitz. **"To Infinity, and Beyond: Uniting the Galaxy of Biological Data",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 21-22

Judy Bayard Cushing. **"Metadata and Semantics: A Computational Challenge for Molecular Biology",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 23-24

Susan B. Davidson. **"Sharing Biomedical Data with Impunity and Ease",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 11-12

J. Dana Eckart; Bruno W. S. Sobral. **"A Life Scientist's Gateway to Distributed Data Management and Computing: The PathPort/ToolBus Framework",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 79-88

Barbara Eckman; Julia Rice; Peter Schwarz. **"Data Management in Molecular and Cell Biology: Vision and Recommendations",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 93-97

George M. Garrity; Catherine Lyons. **"Future-Proofing Biological Nomenclature",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 31-33

Michael Gribskov. **"Challenges in Data Management for Functional Genomics",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 3-5

Amarnath Gupta; Bertram Ludscher. **"The Many Faces of Process Interaction Graphs: A Data Management Perspective",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 105-107

Joachim Hammer; Markus Schneider. **"Going Back to Our Database Roots for Managing Genomic Data",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 117-119

Peter D. Karp. **"What Database Management System(s) Should Be Employed in Bioinformatics Applications?",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 35-36

Toni Kazic; Ed Coe; Mary Polacco; Chi-Ren Shyu. **"Whither Biological Database Research?",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 61-65

Jessie Kennedy. **"Supporting Taxonomic Names in Cell and Molecular Biology Databases",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 13-16

Zo Lacroix. **"Designing Efficient User-Friendly Biological Data Management Systems",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 113-115

Peter Li. **"Biological Data Extinction",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 49-50

Michael N. Liebman. **"Data Management Systems: Science versus Technology?",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 67-69

David Maier. **"Will Database Systems Fail Bioinformatics, Too?",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 71-73

Victor M. Markowitz. **"Data Management Challenges for Molecular and Cell Biology: An Industry Perspective",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 121-122

Daniel P. Miranker. **"Metric-Space Indexes as a Basis for Scalable Biological Databases",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 57-60

Frank Olken. **"Graph Data Management for Molecular Biology",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 75-78

Z. Meral Ozsoyoglu; Joseph H. Nadeau; G. Ozsoyoglu. **"Pathways Database System",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 123-125

D. Stott Parker; Michael M. Gorlick; Christopher J. Lee. **"Evolving from Bioinformatics in-the-Small to Bioinformatics in-the-Large",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 37-48

Jignesh M. Patel. **"The Role of Declarative Querying in Bioinformatics",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 89-91

Louiqa Raschid. **"Data Modeling and Data Management for the Biological Enterprise",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 51-55

Ambuj K. Singh. **"Querying and Mining Biological Databases",** *OMICS*, vol. 7, no. 1, Spring 2003, pages 7-8

Christian J. Stoeckert Jr. **"Common Objects: Think Global, Act Local"**, *OMICS*, vol. 7, no. 1, Spring 2003, pages 103-104

Shalom Tsur. **"A Plea for Normalization of Biosciences Information"**, *OMICS*, vol. 7, no. 1, Spring 2003, pages 109-112

Mark S. Tuttle. **"Explaining Biology to Computers"**, *OMICS*, vol. 7, no. 1, Spring 2003, pages 27-29

Zhiping Weng. **"Managing Biological Sequence and Protein Structure Data"**, *OMICS*, vol. 7, no. 1, Spring 2003, pages 25-26

Gio Wiederhold. **"The Impossibility of Global Consistency"**, *OMICS*, vol. 7, no. 1, Spring 2003, pages 17-20