Life Science Research and Data Management – what can they give each other?

Amarnath Gupta
San Diego Supercomputer Center
University of California San Diego
La Jolla, CA 92130
gupta@sdsc.edu

1 Introduction

"Databases for the life sciences" is not really a newly emerging area in databases. The Nucleotide Sequence Database from the European Molecular Biology Laboratory (EMBL) has been operational at the European Bioinformatics Institute (EBI) since 1980. SWISS-PROT, the classical database containing protein information, was established in 1986. Today, there are over a thousand different life science databases with contents ranging from gene-property data for different organisms to brain image data for patients with neurological disorders.

What then calls for a special issue of the SIGMOD Record devoted to the topic?

We may posit three observations that underscore the need for the database community to take a closer and deeper look at the problems associated with life science databases. The first, reported in more detail by Jagadish and Olken in this issue, is that *biological information is inherently more complex* because it needs to represent objects, interactions and phenomena of the natural world – and yet it has received little attention from the data modeling community ([2] is an interesting exception). Apart from sequence data, there have been few data models to develop efficient representations and operations for biologically relevant data. The second observation is that *scientific questions in life sciences lead to much more complex queries* than the traditional database world is used to. While significant progress has been made in sequence databases, temporal databases, graph databases, recursive query processing, user-defined aggregate computation and so forth, real-world scientific queries often need to cut across all of them, and place new demands on query languages and evaluation. The third observation follows from the problem of information integration across a combination of homegrown, improperly modeled as well as highly sophisticated databases, often having distinctly different data models, laden with unstated assumptions. In many life science applications, information needs to be integrated across databases that are not only heterogeneous, but may have non-overlapping schemas (because the data are from different instruments or scales of resolution), may have contradictory information (because experiments were performed with different assumptions), or may contain data that can only be compared after matching their context expressed in the form of constraints.

Fortunately, the database, data mining and knowledge management communities are gradually becoming more active in investigating these and other life-science inspired problems. Although genomic and proteomic applications dominate the life science data management research today, this is likely to broaden over the next few years and embrace problem areas like disease-specific databases, tissue-specific databases and process-specific databases. In the current issue of the SIGMOD Record, we present some of the research originating from both database and life science application scientists.

2 An Overview of the Special Issue

In selecting the papers submitted to this issue, we have attempted to strike a balance between papers that illustrate how database techniques can be applied in novel ways to life science problems and those that illustrate novel database problems that arise in the course of solving life science application needs. The issue contains a workshop report, six research papers and two problem domain analyses from recently funded groups that describe their approach to the specific life science data management issues they investigate.

The opening paper "Database Management for Life Sciences Research" by H.V. Jagadish and F. Olken is a report that summarizes the findings of a workshop jointly sponsored by NSF-NIH-DARPA held last year. The report identifies several challenge areas that require more research attention, and can be viewed as a high-level requirements exploration that is needed to stimulate more interactions between the two communities. They emphasize that aside from the variety of data types, the treat-

ment of uncertain data, and the integration of database and workflow technologies are necessary to solve life-sized life science problems.

It is now widely recognized that a significant fraction of biological relationships can be modeled as a graph, and a subgraph-searching forms the core of many query processing algorithms. The paper by Liu, Chang and Chao entitled "An Optimal Algorithm for Querying Tree Structures and its Applications in Bioinformatics" considers the problem of searching graph-structured databases such as protein structures and regulatory pathways. They show that with appropriate encoding of the data, the problem transforms into one of searching for root-to-leaf paths in a labeled graph. They show while the general problem is hard, it is polynomially solvable in the case where the labeled graph is a tree.

Terminological databases play a special role in life sciences. These databases often contain taxonomic and partonomic relationships among a large set of terms that constitute the vocabulary of a specific domain. When databases are constructed to describe the properties of biological objects such as genes and gene products, these vocabularies act as attribute-value domains. The Gene Ontology is a 17000-term DAG database that is now the standard vocabulary for molecular processes, biological functions and location of gene products. The article entitled "Using Reasoning to Guide Annotation with Gene Ontology Terms in GOAT" by Bada *et al* describes a tool that treats the ontology as a description logic database, and illustrates how reasoning is used to facilitate proper annotation (value assignment to attributes) for gene products.

The paper entitled "Managing and Analyzing Carbohydrate Data" by Aoki, Ueda and Yamaguchi is a very interesting example demonstrating how data management techniques are finding their use in life sciences. The paper considers a class of carbohydrate molecules called *glycans* that can be modeled as tree structures with typed nodes having ordered children and labeled edges. The paper develops a tree matching algorithm for this class of data by combining a common subtree finding technique with a dynamic programming strategy. Aside from matching, they also present a data mining method where the task is to classify *glycans*. For this task, the authors have developed a probabilistic sibling tree Markov model for the classification algorithm which treats *glycans* as labeled unordered trees.

Unlike tree-like structures such as *glycans*, similarity searching in DNA sequences has received more research attention. Yet, aside from the traditional BLAST method, no other technique has emerged into a standard. The paper entitled "Piers: An Efficient Model for Similarity Search in DNA Sequence Databases" by Cao *et al* presents a new, time-efficient technique that does not need to scan the full sequence database to implement the similarity search. It reduces the search by identifying sequence segments called *piers*, and storing them in a hash table that can remain in main memory. The technique is shown to outperform BLAST in efficiency.

As mentioned in the previous section, not many data models have been developed specifically for life science information. In their paper entitled "The GenAlg Project: Developing a New Integrating Data Model, Language, and Tool for Managing and Querying Genomic Information", Hammer and Schneider present a high-level algebraic framework for genomic information, developed specifically from the biologist's point of view. They first define a conceptual universe (an ontology) of terms and relationships inspired by the extended entity-relationship model. In this framework, an algebra can be defined through abstract data types based on the entities defined in the ontology. The operations in the algebra are high-level and often correspond to biological processes that transform one type of genomic entity into another.

The next paper, entitled "BIO-AJAX: An Extensible Framework for Biological Data Cleaning" by Herbert *et al* deals with the pragmatic problem of data cleaning. An important problem in data cleaning is to reconcile multiple databases by identifying duplicate objects so that their properties can be consolidated. Led by the data warehousing community, most of the work in this domain is done in a relational setting. This papers extends the data reconciliation process to tree-structured databases that abound in life science applications. In the example case, TREEBASE, a phylogenetic tree hosted by the University at Buffalo, is reconciled with TAXONOMY, another tree-structured database from the National Center for Biotechnology Information (NCBI).

The problem of automatic wrapper generation for web-based sources is well-researched but not completely solved yet. Specifically, the maintainability of web wrappers against source updates remains a challenge. The paper entitled "Automatic Composite Wrapper Generation for Semi-Structured Biological Data Based on Table Structure Identification" by Chen, Jamil and Wang describes the *Tag-tree* data model that extracts tabular structures from web pages. The generated wrapper can produce nested tables by following links up to a user-specified depth. The utility of their system, called *Pickup*, is demonstrated over the web site of the Cancer Genome Anatomy Project at NCBI.

The next paper, entitled "A Distributed Database for BioMolecular Images" by Singh, Manjunath and Murphy describes a newly funded scientific project at the University of California Santa Barbara. The scientific objective of the project is to gain a genetic/molecular-level understanding of diseases of the human retina. To achieve this however, information from a variety of image-producing

instruments need to be integrated. The need to integrate image-based information sources leads to new and interesting challenges for data management communities. The paper describes, in significant detail, the different classes of queries the prospective system must accommodate. Expectedly, a large part of the challenge involves extracting features from the various images and assigning scientifically valid semantics to them such that this semantics can be used by the system integrator for complex distributed queries.

The concluding paper of this issue deals with an information integration problem that arises in the context of today's biological sources. "BioFast: Challenges in Exploring Linked Life Science Sources" by Bleiholder *et al* considers integrating information sources where a data instance in one source (like a gene in OMIM) has a direct link to a corresponding data element in another data source (the same gene in GENBANK). Quite possibly, an efficient query plan in such cases will involve direct pointer following, instead of a set-based value join used in a conventional mediator. But to come up with such a query plan, the system must first have a model of the pointer trajectories that weave through the sources. The paper describes the authors' approach in developing parameterized logical links and the link topology that can be exploited to find efficient trajectories for specific queries.

3 Tailpiece

Despite the quality and the breadth of research papers presented in this special issue, there are many other aspects of life-science data management not covered here, perhaps because fewer research groups are addressing these problems. We conclude this introduction with a couple of such examples.

A large fragment of life science data come from experimental records where a subset of attributes serve as control variables and another subset as observed variables. Typically, experimentally recorded data has significant variability. There is now a growing body of research on data quality – models for information quality have been proposed [4, 5] and issues related to quality-aware query processing have been explored [3]. Unfortunately, this body of work has not really carried over to the realm of scientific data. Moreover, one can identify a typical set of queries against such experimental data, for which today's database systems do not provide any native support. They would often include statistical functions and comparisons. For example, a query that uses a group by to produce subsets of data for each binding of the grouping variables might be extended to include a hypothesis testing function on the non-grouping variables. Due to this inadequate support, life science researchers often compute statistics outside the DBMS. We believe there needs to be more investigation on characterizing statistical queries and how to support them more effectively.

We mentioned the growing importance of ontologies for life science information systems. While there is clearly a lot of activity in specifying such ontologies for focused domains [7], there is a need to investigate whether current ontology languages are sufficient for expressing important real-world constraints. We have also witnessed some activities in bringing graph-structured domains within the scope of mainstream data management [1, 6]. Unfortunately, we have seen remarkably few papers on how query evaluation is affected in the presence of such ontologies. Specifically, biological ontologies specify many relationships aside from *is-a* and *part-of*, that come with their own properties and constraints. Can these constraints be used for semantic query optimization? How would a query processor simultaneously utilize the graph-structured character and the relationship constraints of the ontologies?

Clearly, there is ample room for further research.

Ackowledgement. I sincerely thank Ling Liu for coming up with the idea of this special issue of the SIGMOD Record. I also gratefully acknowledge all the reviewers who donated their time.

References

[1] V. Christophides, G. Karvounarakis, D. Plexousakis, M. Scholl, and S. Tourtounis. Optimizing taxonomic semantic web queries using labeling schemes. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(2):207–228, Feb. 2004.

[2] C. M. M. Keet. Biological data and conceptual modelling methods. *Journal of Conceptual Modeling*, October 2003.

[3] U. Leser and F. Naumann. Query planning with information quality bounds. In *Proc. 5th International Conference on Flexible Query Answering Systems (FQAS)*, pages 85–94, 2000.

[4] A. Motro and I. Rakov. Estimating the quality of databases. In *Proc. 3rd International Conference on Flexible Query Answering Systems (FQAS)*, pages 298–307. Lecture Notes in Artificial Intelligence, vol. 1495, 1998.

[5] F. Naumann. From databases to information systems – information quality makes the difference. In *Proc. 6th International Conference on Information Quality (IQ)*, pages 244–260. MIT, 2001.

[6] W. Ng. An extension of the relational data model to incorporate ordered domains. *ACM Transactions on Database Systems*, 26(3):344–383, September 2001.

[7] R. Stevens, C. Goble, I. Horrocks, and S. Bechhofer. Building a bioinformatics ontology using OIL. *IEEE Trans. on Information Technology in Biomedicine*, 6(2):135–141, June 2002.