

Heterogeneous Data and Algorithm Integration in Bioinformatics

ISMB 2002 Tutorial Proposal

Barbara Eckman, Julia Rice, William Swope
IBM Life Sciences

Motivation. The need for data integration is widely acknowledged in bioinformatics and genomics. Data is currently spread across the Internet in a wide variety of formats. Success in most bioinformatics-related activities, from functional characterization of genomic sequence to prioritization of drug targets, requires an integrated view of all relevant data, including the results of back-end analyses such as BLAST. The problems of integration may be addressed using a wide variety of approaches, and integration systems abound in both the academic and commercial sectors. While each approach has strengths and weaknesses, it can be difficult to evaluate which approach suits a particular need best without fully understanding the data integration landscape.

Audience. I/T oriented bioinformaticians and computationally sophisticated biologists. The tutorial will assume a basic familiarity with relational database technology, as well as common biological databases and analysis algorithms. Mackey and Pearson's morning tutorial, "Relational Databases for Biologists", would be a good introduction.

Objectives. We expect that the attendees will come away from the tutorial with the following:

- An understanding of the variety of integration problems and needs in bioinformatics and genomics today
- An understanding of the variety of integration strategies currently available, and their strengths and weaknesses
- The ability to evaluate existing or new integration approaches according to five general categories or "axes"
- An appreciation of the integration problems that have not yet been solved (open problems in the integration field)

We will also offer a very brief introduction to the challenges of integrating bioinformatics and cheminformatics, an increasingly important issue in pharmaceutical research.

Tutorial Outline

Introduction

- Audience and Assumptions
- Objectives of Tutorial
- Motivation of Tutorial
- Obstacles to Integration
- A Motivating Example
- Axes describing the space of solutions

Examples of Integration Solutions

- Browsing Solutions

- Data Warehousing Solutions
- Federated Database Solutions
- More Practice with the Categorization Axes

Advanced Topics

- Query Optimization in Federated Systems
 - Featured Example: IBM's Discovery Link
- Semantic Integration
 - Featured Example: U Manchester's TAMBIS/OIL

Strengths and Weaknesses of the Various Approaches to Integration

- Issues in Evaluating Strengths and Weaknesses
- Strengths and Weaknesses Compared

Open/Tough Problems in Integration

- Query Planning over Web Data Sources: Eckman, Lacroix, Raschid, BIBE 2001 and International Journal of Bioinformatics and Bioengineering 2002
- Efficient Integration and Querying of XML Data Sources: XML Wrapper for Federated DB2 from IBM Research
- Data and Schema Transformation: Clio Project from IBM Research
- Generalized Annotation: Sophia Project from IBM Research
- Data Standardization Efforts

Integrating Bioinformatics and Cheminformatics

- The drug discovery pipeline
- Workflow across the pipeline
- Cheminformatics data sources
- Data integration challenges in chemistry
- Sample queries across bioinformatics and cheminformatics data sources

Acknowledgments

Presenters' Qualifications

All three presenters are members of IBM's Discovery Link team. Discovery Link is IBM's heterogeneous data integration solution for life sciences.

Barbara Eckman, Ph.D.

Bioinformatics and Data Integration:

- Graduate work in Computer Science and Computational Biology at the University of Pennsylvania under Susan Davidson and Tandy Warnow
- 10 years experience in bioinformatics
- Programmer and database expert at Philadelphia Human Genome Center for Chromosome 22, under Chris Overton and David Searls
- Department of Bioinformatics at Merck, co-developer of the Merck Gene Index. See Eckman et al., *Bioinformatics* (1998)
- Assistant Director, Department of Bioinformatics at GlaxoSmithKline (formerly SmithKlineBeecham), where she led a 3-year project in biological database integration. See Eckman, Kosky and Laroco, *Bioinformatics* (2001)

- Currently involved in a research collaboration on query planning and optimization in the integration of biological web data sources. See Eckman, Lacroix and Raschid, *IEEE BioInformatics and BioEngineering (BIBE 2001)*

Teaching experience:

- Teaching Assistant, Trinity College, Hartford, CT 1978-79
- Instructor, Connecticut College, New London, CT 1979-80
- Instructor, University of Pennsylvania, Philadelphia, PA 1981-84

Julia Rice, Ph.D.

- Manager of Scientific and Technical Information Management Group, Science and Technology, IBM Research Division, Almaden Research Center, San Jose, CA
- Member of the IBM Garlic data integration research project team
- Fellow of the American Physical Society (APS)
- Extensive experience with computational quantum chemistry methods development and applications, as well as with scientific and technical software design and implementation
- Author on over 60 publications in physical chemistry

William Swope, Ph.D.

- Research Staff Member, Science and Technology, IBM Research Division, Almaden Research Center, San Jose, CA
- Over 25 years technical experience in computational chemistry methods development and applications to problems such as drug design and development
- Member of the IBM BlueGene science project to study protein folding by computer simulation
- Editor of externally refereed special issues of IBM journals on Deep Computing in the Life Sciences (<http://www.research.ibm.com/journal/sj40-2.html>, <http://www.research.ibm.com/journal/rd45-34.html>)