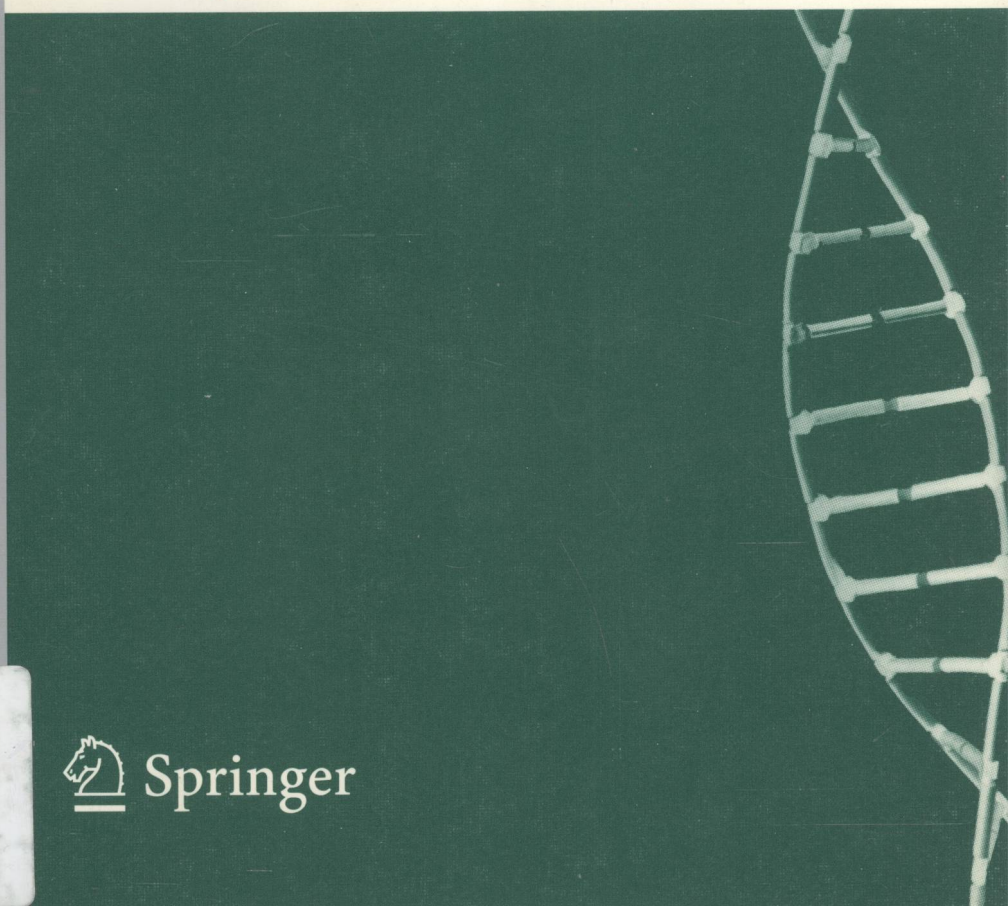
Bertram Ludäscher
Louiqa Raschid (Eds.)

LNBI 3615

Data Integration in the Life Sciences

**Second International Workshop, DILS 2005
San Diego, CA, USA, July 2005
Proceedings**



Springer

Bertram Ludäscher Louiqa Raschid (Eds.)

Data Integration in the Life Sciences

Second International Workshop, DILS 2005
San Diego, CA, USA, July 20-22, 2005
Proceedings

Springer

Volume Editors

Bertram Ludäscher
University of California at Davis, Department of Computer Science
One Shields Ave, Davis, CA 95616, USA
E-mail: ludaesch@ucdavis.edu

Louiqa Raschid
University of Maryland, Department of Computer Science
A.V. Williams Building, College Park, MD 20742, USA
E-mail: louiqa@umiacs.umd.edu

Library of Congress Control Number: 2005928957

CR Subject Classification (1998): H.2, H.3, H.4, J.3

ISSN 0302-9743
ISBN-10 3-540-27967-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-27967-9 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11530084 06/3142 5 4 3 2 1 0

Lecture Notes in Bioinformatics 3615

Edited by S. Istrail, P. Pevzner, and M. Waterman

Subseries of Lecture Notes in Computer Science

Preface

The explosion in the number and size of life science data resources, and the rapid growth in the variety and volume of laboratory data has been fueled by world-wide research activity and the emergence of new technologies. The modeling, management and analysis of this data often requires a comprehensive integration of heterogeneous and typically semistructured data, distributed across many possibly data sources. Recent interoperability standards such as XML and WSDL solve some (easy) problems, but data and process integration often remain time-consuming and error-pone manual tasks. The difficulty of these tasks is compounded by the high degree of semantic heterogeneity across data sources, varying data quality, and other domain-specific application requirements.

DILS 2005 was the 2nd International Workshop on Data Integration in the Life Sciences, following a successful first DILS workshop, March 2004 in Leipzig, Germany. For a specialized workshop, the DILS 2005 call for papers created a large interest (over 50 abstracts and eventually 42 paper submissions; an increase of over 20% over DILS 2004), out of which the international Program Committee selected 15 full papers, as well as 5 short papers, and 8 posters/demonstrations, which are all included in this volume. They cover a wide spectrum of theoretical and practical issues including scientific/clinical workflows, ontologies, tools and systems, and integration techniques. DILS 2005 also featured keynotes by Dr. Peter Buneman, Professor at the School of Informatics, University of Edinburgh, and Dr. Shankar Subramaniam, Professor at the Department of Bioengineering and Chemistry, UC San Diego. The program also included 6 invited presentations and reports on ongoing research activities in academia and industry and a panel organized by the AMIA Geomics Working Group.

The workshop was organized by the San Diego Supercomputer Center (SDSC) and took place July 20–22, 2005 at the University of California, San Diego. Additional sponsors included Microsoft Research, the American Medical Informatics Association (AMIA), the UC Davis Genome Center, and the University of Maryland Center for Bioinformatics and Computational Biology.

As the workshop co-chairs and editors of this volume, we thank all authors who submitted papers and the Program Committee members and external reviewers for their excellent work. Special thanks also go to Amarnath Gupta who served as workshop general chair, and his team, especially Donna Turner, Jon Meyer, and LInda Ferri, all at SDSC. We thank Chani Johnson and the Microsoft CMT Team for the excellent support of their paper management system. Finally, we thank Alfred Hofmann, Erika Siebert-Cole, and the team from Springer for their cooperation and help in putting this volume together.

June 2005 Bertram Ludäscher and Louiqa Raschid

2nd International Workshop on Data Integration in the Life Sciences (DILS)

University of California, San Diego
July 20–22, 2005

DILS 2005 Co-chairs

Amarnath Gupta	(General Chair)	University of California, San Diego, USA
Bertram Ludäscher	(PC Co-chair)	University of California, Davis, USA
Louiqa Raschid	(PC Co-chair)	University of Maryland, USA

Program Committee

Vineet Bafna	University of California, San Diego, USA
Chitta Baral	Arizona State University, USA
Judith Blake	Jackson Laboratory, USA
Shawn Bowers	University of California, Davis, USA
Terence Critchlow	Lawrence Livermore National Laboratory, USA
Alin Deutsch	University of California, San Diego, USA
Barbara Eckman	IBM Life Sciences, USA
Christoph Freytag	Humboldt University, Berlin, Germany
Floris Geerts	University of Edinburgh, UK
Carole Goble	University of Manchester, UK
Amarnath Gupta	University of California, San Diego, USA
Michael Gribskov	Purdue University, USA
Ralf Hofestaedt	University of Bielefeld, Germany
Hasan Jamil	Wayne State University, USA
Matthew Jones	University of California, Santa Barbara, USA
Jessie Kennedy	Napier University, Edinburgh, UK
Zoé Lacroix	Arizona State University, USA
Ulf Leser	Humboldt University Berlin, Germany
Felix Naumann	Humboldt University Berlin, Germany
Frank Olken	Lawrence Berkeley National Laboratory, USA
Jignesh Patel	University of Michigan, USA
Erhard Rahm	University of Leipzig, Germany
Julia Rice	IBM Life Sciences, USA
Peter Tarczy-Hornoch	University of Washington, USA
Limsoon Wong	Institute for Infocomm Research, Singapore
Aidong Zhang	State University of New York at Buffalo, USA

Additional Reviewers

Alexander Bilke	Humboldt University Berlin, Germany
Jens Bleiholder	Humboldt University Berlin, Germany
Hong-Hai Do	University of Leipzig, Germany
Antoon Goderis	University of Manchester, UK
Woo-Chang Hwang	State University of New York at Buffalo, USA
Daxin Jiang	State University of New York at Buffalo, USA
Toralf Kirsten	University of Leipzig, Germany
Peter Li	University of Newcastle, UK
Phillip Lord	University of Manchester, UK
Hervé Ménager	Arizona State University, USA
Peter Mork	University of Washington, USA
HweeHwa Pang	Institute for Infocomm Research, Singapore
Pengjun Pei	State University of New York at Buffalo, USA
Benjamin Prins	University of Bielefeld, Germany
Robert Stevens	University of Manchester, UK
Thoralf Töpel	University of Bielefeld, Germany
Silke Trißl	Humboldt University Berlin, Germany
Chris Wroe	University of Manchester, UK
Xian Xu	State University of New York at Buffalo, USA

Sponsors

Microsoft Research	research.microsoft.com
San Diego Supercomputer Center	www.sdsc.edu
American Medical Informatics Association (AMIA)	www.amia.org
UC Davis Genome Center	genomics.ucdavis.edu
U Maryland Institute for Advanced Computer Studies	www.umiacs.umd.edu
University of California, San Diego	www.ucsd.edu

Organization Committee

Amarnath Gupta	San Diego Supercomputer Center
Jon C. Meyer	San Diego Supercomputer Center
Donna Turner	San Diego Supercomputer Center
Linda Ferri	San Diego Supercomputer Center

Website

For more information on the workshop please visit the workshop website at www.sdsc.edu/dils05.

Lecture Notes in Bioinformatics

Vol. 3615: B. Ludäscher, L. Raschid (Eds.), Data Integration in the Life Sciences. XII, 344 pages. 2005.

Vol. 3500: S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 632 pages. 2005.

Vol. 3388: J. Lagergren (Ed.), Comparative Genomics. VII, 133 pages. 2005.

Vol. 3380: C. Priami (Ed.), Transactions on Computational Systems Biology I. IX, 111 pages. 2005.

Vol. 3370: A. Konagaya, K. Satou (Eds.), Grid Computing in Life Science. X, 188 pages. 2005.

Vol. 3318: E. Eskin, C. Workman (Eds.), Regulatory Genomics. VII, 115 pages. 2005.

Vol. 3240: I. Jonassen, J. Kim (Eds.), Algorithms in Bioinformatics. IX, 476 pages. 2004.

Vol. 3082: V. Danos, V. Schachter (Eds.), Computational Methods in Systems Biology. IX, 280 pages. 2005.

Vol. 2994: E. Rahm (Ed.), Data Integration in the Life Sciences. X, 221 pages. 2004.

Vol. 2983: S. Istrail, M.S. Waterman, A. Clark (Eds.), Computational Methods for SNPs and Haplotype Inference. IX, 153 pages. 2004.

Vol. 2812: G. Benson, R.D.M. Page (Eds.), Algorithms in Bioinformatics. X, 528 pages. 2003.

Vol. 2666: C. Guerra, S. Istrail (Eds.), Mathematical Methods for Protein Structure Analysis and Design. XI, 157 pages. 2003.

Table of Contents

Keynotes

User Applications

Ontologies

Data Integration I–IV

Potpourri

Posters and Demonstrations

Invited Briefings

Challenges in Biological Data Integration in the Post-genome Sequence Era

(Keynote Talk)

Shankar Subramaniam

University of California, San Diego
shankar@sdsc.edu

Abstract. We are witnessing the emergence of the "data rich" era in biology. The myriad data in biology ranging from sequence strings to complex phenotypic and disease-relevant data pose a huge challenge to modern biology. The standard paradigm in biology that deals with "hypothesis to experimentation (low throughput data) to models" is being gradually replaced by "data to hypothesis to models and experimentation to more data and models". And unlike data in physical sciences, that in biological sciences is almost guaranteed to be highly heterogeneous and incomplete. In order to make significant advances in this data rich era, it is essential that there be robust data repositories that allow interoperable navigation, query and analysis across diverse data, a plug-and-play tools environment that will facilitate seamless interplay of tools and data and versatile user interfaces that will allow biologists to visualize and present the results of analysis in the most intuitive and user-friendly manner. This talk will address several of the challenges posed by enormous need for scientific data integration in biology with specific exemplars and strategies. The issues addressed will include:

- Architecture of Data and Knowledge Repositories
- Databases: Flat, Relational and Object-Oriented; what is most appropriate?
- The imminent need for Ontologies in biology
- The Middle Layer: How to design it?
- Applications and integration of applications into the middle layer
- Reduction and Analysis of Data: the largest challenge!
- How to integrate legacy knowledge with data?
- User Interfaces: web browser and beyond

The complex and diverse nature of biology mandates that there is no "one solution fits all" model for the above issues. While there is a need to have similar solutions across multiple disciplines within biology, the dichotomy of having to deal with the context, which is everything in some cases, poses severe design challenges. For example, can a system that describes cellular signaling also describe developmental genetics? Can the ontologies that span different areas (e.g. anatomy, gene and protein data, cellular biology) be compatible and connective? Can the detailed biological knowledge accrued painstakingly over decades be easily integrated with high throughput data? These are only few of the questions that arise in designing and building modern data and knowledge systems in biology.

B. Ludäscher and L. Raschid (Eds.): DILS 2005, LNBI 3615, p. 1, 2005.

Curated Databases

(Keynote Talk)

Peter Buneman

School of Informatics and Digital Curation Centre,
University of Edinburgh
opb@inf.ed.ac.uk

Abstract. Measured in dollars per byte, the cost of data in some biological data sets exceeds that of "big science" data by several orders of magnitude. This somewhat pointless observation does at least underline the fact that biological databases are constructed and maintained with a very great deal human effort—they are *curated.* So what are the issues with curated data, and how well does current database technology serve them?

In this talk I shall describe some of the new challenges to database research that arise from curated databases and what my colleagues and I are doing to tackle them. They include annotation, data provenance, database archiving, data publishing and security. I shall also attempt to summarise the work of the recently formed Digital Curation Centre, which is concerned not only with these database-related issues but also with the larger problems of ensuring that our scientific and scholarly data is understandable not only by current users but is "curated" in the sense that it will be usable in the future.

B. Ludäscher and L. Raschid (Eds.): DILS 2005, LNBI 3615, p. 2, 2005.

A User-Centric Framework for Accessing Biological Sources and Tools*

Sarah Cohen-Boulakia[1], Susan Davidson[2], and Christine Froidevaux[1]

[1] LRI, CNRS UMR 8023, Université Paris-Sud, Orsay, France
{cohen, chris}@lri.fr
[2] Department of Computer and Information Science,
University of Pennsylvania, USA
susan@cis.upenn.edu

Abstract. Biologists face two problems in interpreting their experiments: the integration of their data with information from multiple heterogeneous sources and data analysis with bioinformatics tools. It is difficult for scientists to choose between the numerous sources and tools without assistance. Following a thorough analysis of scientists' needs during the querying process, we found that biologists express *preferences* concerning the sources to be queried and the tools to be used. Interviews also showed that the querying process itself – the *strategy* followed – differs between scientists. In response to these findings, we have introduced a user-centric framework allowing to specify various querying processes. Then we have developed the BioGuide system which helps the scientists to choose suitable sources and tools, find complementary information in sources, and deal with divergent data. It is generic in that it can be adapted by each user to provide answers respecting his/her *preferences*, and obtained following his/her *strategies*.

Availability: http://www.lri.fr/~cohen/bioguide/bioguide.html

1 Introduction

Life sciences are continuously evolving so that the number and size of new sources providing specialized information in biological sciences have increased exponentially in the last few years,[1] as well as the number of tools required to carry out bioinformatics tasks. Scientists are therefore frequently faced with the problem of selecting sources and tools when interpreting their data. The diversity of sources and tools available makes it increasingly difficult to make this selection without assistance.

We firstly introduce a framework allowing to specify various querying processes. Our work was developed following a thorough study of scientists' needs during querying and data management. After interviewing scientists working in

* This work was supported in part by the European Project HKIS IST-2001-38153, the Fulbright Program as well as a Hitachi Chair at INRIA.

[1] See the annual Nucleic Acids Research database issue (January).

B. Ludäscher and L. Raschid (Eds.): DILS 2005, LNBI 3615, pp. 3–18, 2005.

various domains, we found that they expressed *preferences* concerning the sources queried and the tools used. Moreover, this study emphasized the fact that the process of querying itself – the *strategy* – varies from one scientist to another. We have then designed the BioGuide system, which provides scientists with support during the querying process. BioGuide assists the scientist with data searches within sources, providing information concerning the sequences of sources to be consulted and the tools to be used: the *paths* between sources to be followed.

We first describe the method used to assess scientists' requirements, and present the needs identified (section 2). We then describe the notion of strategy (section 3) and the way in which we propose to manage preferences (section 4). Section 5 introduces the formal framework and presents the general architecture of BioGuide, explaining how it provides support for the querying process. The biological significance of the results obtained will be presented in section 6. Section 7 compares our work to previous work and concludes the paper.

2 User Requirements

2.1 Process: Interviews and Questionnaire

We started with a thorough study of user requirements (cf. BioGuide site). We investigated the way in which scientists query sources and perform bioinformatics tasks (in the spirit of [18] and [6]), paying particular attention to determining why biologists query one source rather than another (*preferences*) and identifying the steps of their querying process (*strategies*).

A questionnaire was developed based on lists of user requirements in three kinds of documents: (i) survey articles [11] and reports of workshops on biological source querying (ii) studies on data quality [14], [4], [15] and (iii) studies on user guidance during the querying process, involving BioMediator [12], BioNavigation [9] and DSS [2]. The questionnaire comprised 28 questions and was constructed according to standard guidelines. As an illustration, four questions are provided:

- Choose a particular context from your own area of study and list some biological queries that you frequently make.
- If several sources yield answers for your query, do you access all of them or only few? If you query only a few, how do you proceed?
- In your mind, what is a "high-quality" source/tool?
- When you look for data related to two linked entities (e.g. a gene and the protein it encodes), how do you proceed (sources accessed, way of correlating information, etc.)?

After collecting responses to the questionnaire, we conducted *interviews* according to classical techniques. We sent questionnaires to 20 individuals, including both biologists and bioinformatics specialists. Their research interests fell into three main domains: *studies of diseases*, *functional* and *structural genomics*.

From the questionnaire, we identified 156 common queries. Some had almost identical structures (e.g. the search for genes involved in *breast* or in *bladder* cancer) and we grouped them together, giving a total of 119 distinct queries.

2.2 Transparent Queries and Traceability

In most cases, neither the sources to access nor the tools to be used were specified by the biologists in their queries. Instead, their queries involved only biological **entities** and **relationships** between entities. An example of such queries is *"Return all contigs that map 'close' to marker M on chromosome 19"* which includes the biological entities CONTIG, MARKER and CHROMOSOME and includes the relationships "maps close to" and "(located) on". We conclude that scientists find it very useful not to have to specify the sources and tools that is, to make **transparent** queries [10].

Follow-up interviews showed that scientists want to ask transparent queries while being aware of the **origin of the answers** obtained. They want to know the *why-provenance* [1] that is, which sources and/or which tools have been used to calculate the data they obtain. Traceability is particularly important for verifying results, drawing conclusions and testing biological hypotheses [19].

2.3 Source and Tool Requirements

A more complex step in the querying process is the **assembly** of information between entities. From the sample queries, we observed that relationships between entities are either explicitly **stored** in the sources or **calculated** by a bioinformatics tool. For example, in the query *"Return all contigs that map 'close' to marker M on chromosome 19"*, the fact that Marker M is on chromosome 19 must be *stored* in the data sources queried by the biologist. Conversely, the relationship of "close mapping" can be *calculated* (e.g. using *Blastn*). For each calculated relationship between entities, we also determined which tools were used to achieve it (e.g. *Blastn*) based on the interview information.

Different kinds of **links** between sources may therefore be distinguished: *internal links* (within the same source), *cross-references* (between different sources) and *tool-links*. *Internal links* may be seen as a way of obtaining information on one entity from another entity within the same source. *Cross-references* are hypertext links from an entity in one source to complementary information in another source, and are not necessarily symmetric (e.g. there are an increasing number of specialized sources which crossreference GenBank but are not referenced in return). Finally, *tool-links* are services provided by a source, yielding links with entities in other sources. Each source may provide several different services achieving a given relationship. For example, GenBank provides different tools (e.g. *Blastx*, *tBlastn*) to enable users to carrying out "similarity searches" between the genes of GenBank and proteins of various sources.

It is also clear from interviews that scientists have **preferences** concerning entities in sources and tools. One of the key issues facing bioinformaticians is therefore to help the scientists to evaluate their confidence in sources and tools, and to make use of this confidence in a semi-automatic querying process. We return to this in section 4.