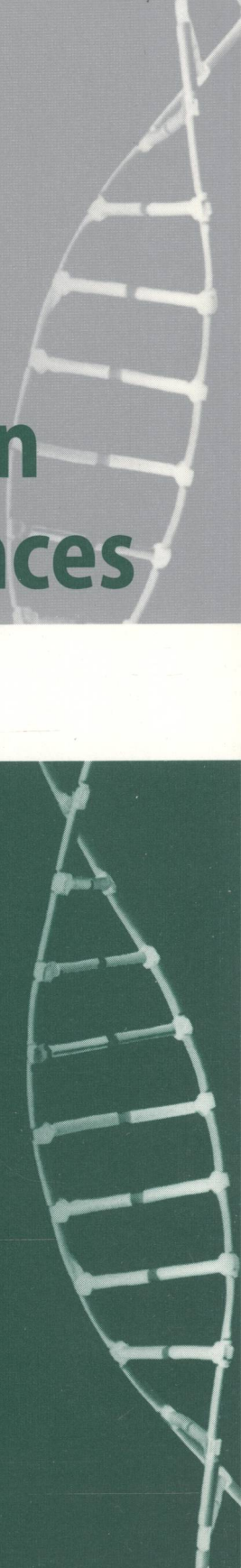
Sarah Cohen-Boulakia
Val Tannen (Eds.)

LNBI 4544

Data Integration in the Life Sciences

4th International Workshop, DILS 2007
Philadelphia, PA, USA, June 2007
Proceedings



Springer

Sarah Cohen-Boulakia Val Tannen (Eds.)

Data Integration in the Life Sciences

4th International Workshop, DILS 2007
Philadelphia, PA, USA, June 27-29, 2007
Proceedings

Springer

E2007003281

Volume Editors

Sarah Cohen-Boulakia
University of Pennsylvania, Department of Computer and Information Science
303 Levine Hall, 3330 Walnut St., Philadelphia, PA 19104, USA
E-mail: sarahcb@seas.upenn.edu

Val Tannen
University of Pennsylvania, Department of Computer and Information Science
570 Levine Hall, 3330 Walnut St., Philadelphia, PA 19104, USA
E-mail: val@cis.upenn.edu

Library of Congress Control Number: 2007928915

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

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743
ISBN-10 3-540-73254-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-73254-9 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12081769 06/3180 5 4 3 2 1 0

Lecture Notes in Bioinformatics 4544

Subseries of Lecture Notes in Computer Science

Preface

Understanding the mechanisms involved in life (e.g., discovering the biological function of a set of proteins, inferring the evolution of a set of species) is becoming increasingly dependent on progress made in mathematics, computer science, and molecular engineering. For the past 30 years, new high-throughput technologies have been developed generating large amounts of data, distributed across many data sources on the Web, with a high degree of semantic heterogeneity and different levels of quality. However, one such dataset is not, by itself, sufficient for scientific discovery. Instead, it must be combined with other data and processed by bioinformatics tools for patterns, similarities, and unusual occurrences to be observed. Both data integration and data mining are thus of paramount importance in life science.

DILS 2007 was the fourth in a workshop series that aims at fostering discussion, exchange, and innovation in research and development in the areas of data integration and data management for the life sciences. Each previous DILS workshop attracted around 100 researchers from all over the world. This year, the number of submitted papers again increased. The Program Committee selected 19 papers out of 52 full submissions. The DILS 2007 papers cover a wide spectrum of theoretical and practical issues including scientific workflows, annotation in data integration, mapping and matching techniques, and modeling of life science data. Among the papers, we distinguished 13 papers presenting research on new models, methods, or algorithms and 6 papers presenting implementation of systems or experience with systems in practice. In addition to the presented papers, DILS 2007 featured two keynote talks by Kenneth H. Buetow, National Cancer Institute, and Junhyong Kim, University of Pennsylvania.

The workshop was held at the University of Pennsylvania, in Philadelphia, USA. It was kindly sponsored by the School of Engineering and Applied Science of the University of Pennsylvania, the Penn Genomics Institute, and Microsoft Research, who also made available their conference management system. As editors of this volume, we thank all the authors who submitted papers, the Program Committee members, and the external reviewers for their excellent work. Special thanks go to Susan Davidson, General Chair, Chris Stoeckert, PC Co-chair, as well as Olivier Biton, Tara Betterbid, and Howard Bilowsky. Finally, we are grateful for the cooperation and help of Springer in putting this volume together.

June 2007

Sarah Cohen-Boulakia
Val Tannen

Organization

Executive Committee

General Chair

Susan Davisdon, University of Pennsylvania, USA

Program Chairs

Chris Stoeckert, University of Pennsylvania, USA
Val Tannen, University of Pennsylvania, USA

Program Committee

Judith Blake	Jackson Laboratory, USA
Sarah Cohen-Boulakia	University of Pennsylvania, USA
Marie-Dominique Devignes	LORIA, Nancy, France
Barbara Eckman	IBM
Christine Froidevaux	LRI, University of Paris-Sud XI, France
Cesare Furlanello	ITC-irst, Trento, Italy
Jim French	University of Virginia, USA
Floris Geerts	University of Edinburgh, UK
Amarnath Gupta	University of California San Diego, USA
Ela Hunt	ETH Zurich, Switzerland
Jacob Koehler	Rothamsted Research, UK
Anthony Kosky	Axiope Inc.
Hilmar Lapp	NESCENT
Ulf Leser	Humboldt-Universität zu Berlin, Germany
Bertram Ludäscher	University of California Davis, USA
Victor Markowitz	Lawrence Berkeley Labs
Peter Mork	MITRE
Tom Oinn	European Bioinformatics Institute, UK
Meral Ozsoyoglu	Case Western Reserve University, USA
John Quackenbush	Harvard, USA
Louiqa Raschid	University of Maryland, USA
Fritz Roth	Harvard Medical School, USA
Susanna-Assunta Sansone	European Bioinformatics Institute, UK
Kai-Uwe Sattler	Technical University of Ilmena, Germany
Chris Stoeckert	University of Pennsylvania, USA
Val Tannen	University of Pennsylvania, USA
Olga Troyanskaya	Princeton University, USA

External Reviewers

Jérôme Azé
Jana Bauckmann
Julie Bernauer
Shawn Bowers
William Bug
Amy Chen
Adnan Derti
Francis Gibbons
Philip Groth
Timothy McPhillips
Joe Mellor
Krishna Palaniappan
Norbert Podhorszki
Murat Tasan
Weidong Tian
Silke Trißl
Daniel Zinn

Sponsorship Chair

Howard Bilofsky, University of Pennsylvania

Sponsoring Institutions

School of Engineering and Applied Science at the University of Pennsylvania
http://www.seas.upenn.edu/

Penn Genomics Institute
http://www.genomics.upenn.edu/

Microsoft Research
http://research.microsoft.com/

Web Site and Publicity Chairs

Olivier Biton	University of Pennsylvania
Sarah Cohen-Boulakia	University of Pennsylvania
DILS 2007 Web site	http://dils07.cis.upenn.edu/

Lecture Notes in Bioinformatics

Vol. 4544: S. Cohen-Boulakia, V. Tannen (Eds.), Data Integration in the Life Sciences. XI, 282 pages. 2007.

Vol. 4463: I. Măndoiu, A. Zelikovsky (Eds.), Bioinformatics Research and Applications. XV, 653 pages. 2007.

Vol. 4453: T. Speed, H. Huang (Eds.), Research in Computational Molecular Biology. XVI, 550 pages. 2007.

Vol. 4414: S. Hochreiter, R. Wagner (Eds.), Bioinformatics Research and Development. XVI, 482 pages. 2007.

Vol. 4366: K. Tuyls, R. Westra, Y. Saeys, A. Nowé (Eds.), Knowledge Discovery and Emergent Complexity in Bioinformatics. IX, 183 pages. 2007.

Vol. 4360: W. Dubitzky, A. Schuster, P.M.A. Sloot, M. Schroeder, M. Romberg (Eds.), Distributed, High-Performance and Grid Computing in Computational Biology. X, 192 pages. 2007.

Vol. 4345: N. Maglaveras, I. Chouvarda, V. Koutkias, R. Brause (Eds.), Biological and Medical Data Analysis. XIII, 496 pages. 2006.

Vol. 4316: M.M. Dalkilic, S. Kim, J. Yang (Eds.), Data Mining and Bioinformatics. VIII, 197 pages. 2006.

Vol. 4230: C. Priami, A. Ingólfsdóttir, B. Mishra, H.R. Nielson (Eds.), Transactions on Computational Systems Biology VII. VII, 185 pages. 2006.

Vol. 4220: C. Priami, G. Plotkin (Eds.), Transactions on Computational Systems Biology VI. VII, 247 pages. 2006.

Vol. 4216: M.R. Berthold, R.C. Glen, I. Fischer (Eds.), Computational Life Sciences II. XIII, 269 pages. 2006.

Vol. 4210: C. Priami (Ed.), Computational Methods in Systems Biology. X, 323 pages. 2006.

Vol. 4205: G. Bourque, N. El-Mabrouk (Eds.), Comparative Genomics. X, 231 pages. 2006.

Vol. 4175: P. Bücher, B.M.E. Moret (Eds.), Algorithms in Bioinformatics. XII, 402 pages. 2006.

Vol. 4146: J.C. Rajapakse, L. Wong, R. Acharya (Eds.), Pattern Recognition in Bioinformatics. XIV, 186 pages. 2006.

Vol. 4115: D.-S. Huang, K. Li, G.W. Irwin (Eds.), Computational Intelligence and Bioinformatics, Part III. XXI, 803 pages. 2006.

Vol. 4075: U. Leser, F. Naumann, B. Eckman (Eds.), Data Integration in the Life Sciences. XI, 298 pages. 2006.

Vol. 4070: C. Priami, X. Hu, Y. Pan, T.Y. Lin (Eds.), Transactions on Computational Systems Biology V. IX, 129 pages. 2006.

Vol. 4023: E. Eskin, T. Ideker, B. Raphael, C. Workman (Eds.), Systems Biology and Regulatory Genomics. X, 259 pages. 2007.

Vol. 3939: C. Priami, L. Cardelli, S. Emmott (Eds.), Transactions on Computational Systems Biology IV. VII, 141 pages. 2006.

Vol. 3916: J. Li, Q. Yang, A.-H. Tan (Eds.), Data Mining for Biomedical Applications. VIII, 155 pages. 2006.

Vol. 3909: A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 612 pages. 2006.

Vol. 3886: E.G. Bremer, J. Hakenberg, E.-H.(S.) Han, D. Berrar, W. Dubitzky (Eds.), Knowledge Discovery in Life Science Literature. XIV, 147 pages. 2006.

Vol. 3745: J.L. Oliveira, V. Maojo, F. Martín-Sánchez, A.S. Pereira (Eds.), Biological and Medical Data Analysis. XII, 422 pages. 2005.

Vol. 3737: C. Priami, E. Merelli, P. Gonzalez, A. Omicini (Eds.), Transactions on Computational Systems Biology III. VII, 169 pages. 2005.

Vol. 3695: M.R. Berthold, R.C. Glen, K. Diederichs, O. Kohlbacher, I. Fischer (Eds.), Computational Life Sciences. XI, 277 pages. 2005.

Vol. 3692: R. Casadio, G. Myers (Eds.), Algorithms in Bioinformatics. X, 436 pages. 2005.

Vol. 3680: C. Priami, A. Zelikovsky (Eds.), Transactions on Computational Systems Biology II. IX, 153 pages. 2005.

Vol. 3678: A. McLysaght, D.H. Huson (Eds.), Comparative Genomics. VIII, 167 pages. 2005.

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

Vol. 3594: J.C. Setubal, S. Verjovski-Almeida (Eds.), Advances in Bioinformatics and Computational Biology. XIV, 258 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.

¥484.00元

Table of Contents

Keynote Presentations

New Architectures and Experience on Using Systems

Managing and Designing Scientific Workflows

Mapping and Matching Techniques

Modeling of Life Science Data

Annotation in Data Integration

Enabling the Molecular Medicine Revolution Through Network-Centric Biomedicine

(Keynote Presentation)

Kenneth H. Buetow

National Cancer Institute
2115 East Jefferson Street
Suite 6000, MSC 8505
Bethesda, MD 20892, USA
buetowk@nih.gov

To deliver on the promise of next generation treatment and prevention strategies in cancer, we must address its multiple dimensions. The full complement of the diverse fields of modern biomedicine are engaged in the assault on this complexity. These disciplines are armed with the latest tools of technology, generating mountains of data. Each surpasses the next in their unprecedented and novel view of the fundamental nature of cancer. Each contributes a vital thread of insight. Information technology provides a promising loom on which the threads of insight can be woven.

Bioinformatics facilitates the electronic representation, redistribution, and integration of biomedical data. It makes information accessible both within and between the allied fields of cancer research. It weaves the disparate threads of research information into a rich tapestry of biomedical knowledge. Bioinformatics is increasingly inseparable from the conduct of research within each discipline. The linear nature of science is being transformed into a spiral with bioinformatics joining the loose ends and facilitating progressive cycles of hypothesis generation and knowledge creation.

To facilitate the rapid deployment of bioinformatics infrastructure into the cancer research community the National Cancer Institute (NCI) is undertaking the cancer Biomedical Informatics Grid, or caBIG^{TM}. caBIG^{TM}, is a voluntary virtual informatics infrastructure that connects data, research tools, scientists, and organizations to leverage their combined strengths and expertise in an open environment with common standards and shared tools. Effectively forming a World Wide Web of cancer research, caBIG^{TM} promises to speed progress in all aspects of cancer research and care including etiologic research, prevention, early detection, and treatment by breaking down technical and collaborative barriers.

Researchers in all disciplines have struggled with the integration of biomedical informatics tools and data; the caBIG^{TM} program demonstrates this important capability in the well-defined and critical area of cancer research, by planning for, developing, and deploying technologies which have wide applicability outside the cancer community. Built on the principles of open source, open access, open development, and federation, caBIG^{TM} infrastructure and tools are open

S. Cohen-Boulakia and V. Tannen (Eds.): DILS 2007, LNBI 4544, pp. 1–2, 2007.

and readily available to all who could benefit from the information accessible through its shared environment. caBIG^{TM} partners are developing or providing standards-based biomedical research applications, infrastructure, and data sets. The implementation of common standards and a unifying architecture ensures interoperability of tools, facilitating collaboration, data sharing, and streamlining research activities across organizations and disciplines.

The caBIG^{TM} effort has recognized that in addition to new infrastructure new information models are required to capture the complexity of cancer. While the biomedical community continues to harvest the benefits of genome views of biologic information, it has been clear from the founding of genetics that biology acts through complex networks of interacting genes. Information models and a new generation of analytic tools that utilizing these networks are key to translating discover to practical intervention.

Phyl-O'Data (POD) from Tree of Life: Integration Challenges from Yellow Slimy Things to Black Crunchy Stuff

(Keynote Presentation)

Junhyong Kim

Department of Biology
Penn Center for Bioinformatics
Penn Genomics Institute
415 S. University Ave.
Philadelphia, PA 19104, USA
junhyong@sas.upenn.edu

1 Background

The AToL (Assembling the Tree of Life) is a large-scale collaborative research effort sponsored by the National Science Foundation to reconstruct the evolutionary origins of all living things. Currently 31 projects involving 150+ PIs are underway generating novel data including studies of bacteria, microbial eukaryotes, vertebrates, flowering plants and many more. Modern large-scale data collection efforts require fundamental infrastructure support for archiving data, organizing data into structured information (e.g., data models and ontologies), and disseminating data to the broader community. Furthermore, distributed data collection efforts require coordination and integration of the heterogeneous data resources. In this talk, I first introduce the general background of the phylogenetic estimation problem followed by an introduction to the associated data modeling, data integration, and workflow challenges.

2 Phylogeny Estimation and Its Utility

While ideas about genealogical reconstruction have been around since Darwin, quantitative algorithmic approaches to the problem have been developed only in the last 50 years. The basic structure of the problem involves considering all possible tree graph structures compatible with an organismal genealogy and measuring their fits to observed data by various objective functions. There are now many algorithms based on various inferential principles including maximum information, maximum likelihood, Bayesian posterior, etc. Many flavors of the phylogeny reconstruction problem have been shown to be NP-hard and there is a considerable body of literature on associated computational and mathematical problems. Phylogenetic methods provide the temporal history of biological diversity and have been used in many applications. For example: to track the

S. Cohen-Boulakia and V. Tannen (Eds.): DILS 2007, LNBI 4544, pp. 3–5, 2007.

history of infectious diseases; to reconstruct ancestral molecules; to reveal functional patterns in comparative genomics; and even in criminal cases, to infer the relatedness of biological criminal evidence. But, a grand challenge for phylogenetic research is to reconstruct the history of all extent organismal life-the so-called Tree of Life.

3 Problems and Challenges

In modern times, much of phylogenetic estimation is carried out using molecular sequences, some explicitly gathered for phylogenetic research; others, systematically collected and deposited in public databases. There are many data management problems associated with using molecular sequence data even from public databases-which are not discussed here. But, for the frontline researcher the problem starts at the stage of actually collecting the biological material for experimentation. That is, the animal must be captured, preserved, measured, and recorded along with associated metadata (e.g., capture location). Such specimens must be physically archived (called voucher specimen) and identified if possible and given a name. All of these activities are sometimes called *alpha-taxonomy.*

Once a specimen has been obtained and named then it (or presumed identical specimens) must be measured for relevant traits including extracting molecular sequences. This action of obtaining "relevant measurements" involve gathering characteristics that will be broadly comparable amongst different varieties of organisms-thus require a prior data model of what is or is not relevant. Once the relevant measurements are recorded, the next important step is deriving an equivalence relationship between measurements on different organisms such that the measurements are considered to be evolutionarily comparable to each other. This activity is called "establishing homology relationships" and is a critical prelude to further analysis. An example of such homologizing activity is the alignment of molecular sequences whereby equivalent relations of individual sequence letters are established. This activity of defining relevant characters and homologizing their assembly is called *Systematics* and the final product of this activity is the "data matrix" that encapsulates the data model of relevant measurements and the relational maps of sets of such measurements. Biologists widely disagree on details of such matrices-for example, whether a particular measurement should be described as present or absent; thus, these matrices are best seen as a "data view" of the primary objects. It is common in the literature and in public databases to have available only the fixed data matrix. Given the complicated and uncertain ontology of such matrices, **a critical challenge is to endow the phylogenetic matrices with their provenance information as well as to provide a facility to change the "views" as biologists' assumptions change.**

Notwithstanding the fundamental problems described above, there are continuing activities to collect empirical measurements and generate data matrices and place them into a structured information source. For example, there are current database efforts within the AToL projects where all projects have sub-aims

targeting data storage, access, and sharing; and, a small number of projects have been funded to develop domain specific data models and analysis tools such as databases for 3D morphological data, web-based information storage and dissemination, and mining molecular databases for phylogenetic information. As a simple fact the magnitude of the AToL efforts is insufficient to meet the real-world needs of the AToL projects that will become critical as each empirical project matures. More importantly, there is little or no coordination between these efforts and there is a critical need to enable the integration of distributed, heterogeneous, and changing data sources; provide for reliable data archiving and maintenance of data provenance; and, help manage the complex data collection and analysis processes. Many of the projects are already very mature and domain-specific problems, cultural problems, and legacy problems make it difficult to develop a single solution to the problems. Therefore, another **critical challenge is to provide ways to post-hoc integrate the extremely dispersed and heterogeneous phylogenetic data sources in a scalable manner.**

The ultimate end product of phylogenetic reconstruction is the tree graph depicting the genealogical history and associated data. The associated data is usually mapped to substructures within the tree graph-say the nodes of the graph, usually from a secondary analysis (so-called post-analysis). For example, once the phylogeny is known, there are algorithms available for reconstructing the measurements of putative ancestors; thus, we may assign data matrices to interior nodes of the tree. In actual practice, the researcher may try many different algorithms to reconstruct the tree, each algorithm may generate multiple trees (e.g., because of equivalent optimality score), and given some preliminary tree estimate one may want to modify the data matrix (try a different view) and re-estimate the tree. Furthermore, there is often a large battery of post-analysis routines that involve other calculations including calculations on substructures of estimated tree. As is typical in complicated data analysis, the total analysis may involve a large number of steps, some steps recursive, cyclic, or branching. Thus, **a final challenge is to develop a workflow for phylogenetic analysis that automatically tracks analysis flow and helps manage the complexity in such a way that is useful to the primary researcher and helps other researchers recapitulate analyses carried out by third parties.**