Executive Summary of the Dagstuhl Workshop "Managing and Mining Genome Information: Frontiers in Bioinformatics"

Jacek Blazewicz
Institute of Bioorganic Chemistry
Polish Academy of Sciences
Poland
blazewic@put.poznan.pl

Martin Vingron,
MPI für molekulare Genetik, Berlin
Germany
vingron@molgen.mpg.de

Johann-Christoph Freytag
Humboldt-Universität zu Berlin
Germany
freytag@dbis.informatik.hu-berlin.de

Abstract

This report summarizes the important aspects of the workshop on "Managing and Mining Genome Information: Frontiers in Bioinformatics" which took place October 31st until November 4th, 2005. Twenty five Participants came from six different countries representing various "branches" of the bioinformatics community. The presentations ranged from describing highly theoretical models to presenting prototypes or systems for managing and mining data in bioinformatics.

1. Goals and Structure of the Workshop

Bioinformatics has evolved at the interface of biology (especially molecular biology), mathematics, and computer science. Its main goal is to develop mathematical models of biological phenomena, especially at a molecular level. The models are then used to construct algorithmic methods for the analysis of biomolecular sequences, structures, and more recently functional data and gene networks. In particular, it has become clear during the last years that only a small fraction of the human genome encodes proteins. This leaves the non-coding DNA responsible for the regulatory functions, i.e. the encoding when and where a gene becomes active. However, while algorithms development has constituted an emphasis of bioinformatics for a long time, data used to be stored in flat files with little importance

Dagstuhl Seminar Proceedings 05441
Managing and Mining Genome Information: Frontiers in Bioinformatics
http://drops.dagstuhl.de/opus/volltexte/2006/475

attributed to data base and knowledge management issues. Under the pressure of incoming data, this is currently changing and increasing efforts are dedicated to knowledge management in combination with data analysis in molecular biology and genome research.

Bioinformatics in Dagstuhl Since 1992, several Dagstuhl Seminars on Molecular Bioinformatics took place. Some of these dealt with Bioinformatics in general, while others focused on the more specific topic of Metabolic Pathways. These seminars successfully brought together computer scientists and applied mathematicians with biochemists and molecular biologists in order to discuss possibilities of cooperation in the growing field of analysis of biomolecular sequences and structures. The seminars were seen to be extremely fruitful by the participants. Many contacts were established that provided the basis for future cooperation. In particular, a number of application problems could be formulated providing computer scientists with a sound basis for further work. Many German computer scientists who participated in these seminars have since taken on the challenge of working in bioinformatics .

This workshop understands itself in the tradition of the earlier Dagstuhl Bioinformatics meetings. At the same time the scientific focus shall be shifted to the interplay between biological data management and analysis methods for biological data. This particular area has not constituted the focus of a Dagstuhl Bioinformatics Seminar before. It is of particular relevance today due to the rapid developments of novel high-throughput methods for probing gene function (“functional genomics”). These methods have produced a data deluge unprecedented in the life sciences. Traditional as well a newly developed analysis methods need to applied to these data. In this situation it is of prime importance to interlink data management and analysis methods in the most efficient and flexible manner.

Scientific focus It is the very goal of bioinformatics to provide for the increase of biological knowledge based on the existing biological data. The current situation, however, is characterized by data sets of different types and quality being stored in numerous data bases all over the world. Queries, analysis questions, and their resulting algorithms differ from data type to data type. A moderate number of researchers from computer science as well as from biology are actively involved in

developing new approaches to the handling of biological data in order to remedy this situation. Initiatives like DAS, the Distributed Annotation Server, the development of SRS, the Sequence Retrieval System, or IBM's DiscoveryLink are witness to this development. There exists, however, a large number of open questions in this field. Some prominent issues in data integration and intelligent data processing for managing and analyzing life science data are:

- Processing life science data is more than querying databases. What is need are "richer", semantically more meaningful concepts to organize the complex task of lice scientists working on specific problems (drug discovery, protein/protein interaction etc.). To approach these issues we need
    a. "Intelligent queries", i.e. queries that do more than to access data;
    b. The use of semantic knowledge, i.e. ontologies (are those sufficient)
    c. Use of standards regarding data/knowledge representation. What are these emerging standards? How to apply/use them?
- How to deal with redundancies in the many data collections: many data collections overlap in structure and content. However, based on different ways to manipulate/operate on the data, those might not always evolve consistent from a global point of view.
- Data reliability. How to deal with contradictions in different or integrated data collections. How to deal with automatically derived data which inevitably contain errors?
- How to deal with evolving data? Determine the identity of existing/new objects, how to keep them, how to make sure that those identities are not changed.
- How to deal with evolving annotations? Annotation is typically constantly added to existing data. Are annotations "first class objects', and if not, how to convert them to entities that can be process and queries as any other data.
- New concepts to represent data in time and space that are "universal", i.e. independent of positions observed with individual organisms.
- Efficient analysis across data-sets, e.g. identifying patterns in sequences where there is additional class information available elsewhere.
- Large-scale data mining in functional genomics data of different types, like DNA microarray data, proteomics data, protein-DNA binding data.

- Computing and supplying biological criteria like, e.g., phylogeny (orthology, paralogy of genes) as query criteria.

2. Participants

The following researchers participated in the workshop. The title of their talk is given in parathesis.

- Jacek Blazewicz, Poznan University of Technology (SBH - State of the Art and Some Perspectives)
- Stefan Bleuler, ETH Zürich (Biclustering of Multiple Gene Expression Data Sets)
- Nadia Brauner, IMAG - Grenoble
- David Corne, University of Reading (Two-Phase EA/k-NN for Feature Selection and Classification in Cancer Microarray Datasets)
- Clarisse Dhaenens, Université de Lille (Mining DNA Micro-array data with association rules)
- Barbara Eckman, IBM Life Sciences - West Chester (Graph Data Management for Molecular and Cell Biology)
- Piotr Formanowicz , Poznan University of Technology, Multistage Isothermic Sequencing by Hybridization
- Johann Christoph Freytag, Humboldt-Universität zu Berlin (Exploring overlapping data sources)
- Krzysztof Fujarewicz, Silesian University of Technology, Gliwice (Hybrid systems identification - application to cell signaling pathways)
- Anna Gambin, University of Warsaw (Medical diagnosis by mass spectrometry - computational methods)
- David Roger Gilbert, University of Glasgow (Modeling the kinetic behavior of the MAPK cascade: negative feedback amplifier characteristics)
- Misha Kapushesky, EBI – Cambridge (ArrayExpress etc.: Microarray Informatics @ EBI
- Marta Kasprzak, Poznan University of Technology (Computational complexity of the Simplified Partial Digest Problem)
- Ulf Leser, Humboldt-Universität zu Berlin (A Query Language for Biological Networks)
- Piotr Lukasiak, Poznan University of Technology (Tabu search strategy in HP protein model)
- Steffen Neumann, IPB – Halle (What the plant does: Mass Spectrometry and Bioinformatics for Metabolomics)
- Sven Rahmann, Universität Bielefeld (Weighted HMMs and Applications to Fragment Statistics for Peptide Mass Fingerprinting)
- Dietrich Rebholz-Schuhmann, EBI – Cambridge (Facts from Text – information extraction online)
- Alexander Schliep , MPI für Molekulare Genetik, Berlin (Mining Heterogeneous Data with Mixture Models)
- Florian Sohler, Universität München (Identifying active transcription factors from expression data using Pathway Queries)
- Andrea Splendiani, Institut Pasteur – Paris (Ontology based data analysis)
- Marta Szachniuk, Poznan University of Technology (On some computational problems arising in RNA 3D structure determination process with NMR)
- Jerzy Tiuryn , University of Warsaw (Distribution of paralog families in genomes)

- Martin Vingron , MPI für Molekulare Genetik, Berlin (Statistics for detecting overrepresentation of genes in Gene Ontology categories)

- Ralf Zimmer , Universität München (Expert knowledge without the expert - Automatic derivation of network contexts from expression data and the biomedical literature; Petri Nets for Bioinformatics)

3. Workshop Schedule

The five days were filled with discussions and formal presentations. The organizers decided on purpose to leave more room for informal discussions to support the emergence of new problems, topics etc. that could be important for the development of the field in the future.

The workshop was organized around the following schedule:

Monday (Chair: Christoph Freytag)

9:00 Welcome – *J.C. Freytag*

9:30 SBH – state of the art and some perspectives – *Jacek Błażewicz*

10:00 Coffee break

10:30 Multistage isothermic SBH – *Piotr Formanowicz*

11:00 Computational complexity of the Simplified Partial Digest Problem – *Marta Kasprzak*

12:00 Lunch

13:00 Hiking due to excellent weather

19:00 Identifying active transcription factors using pathways queries – *Florian Sohler*

19:30 Graph queries over pathways – *Barbara Eckman*

20:00 Pathway query language – *Ulf Leser*

Tuesday (Chair: Jerzy Tiuryn)

9:00 Tabu search strategy in HP model – *Piotr Łukasiak*

9:30 On some computational problems arising RNA 3D structure determination by NMR – *Marta Szachniuk*

10:00 Coffee break

10:30 Overrepresentation of Gene Ontology in sets of genes – *Martin Vingron*

11:00 Deriving Concepts from Literature & Data – *Ralf Zimmer*

11:30 Information extraction from scientific text – *Dietrich Rebholz*

12:00 Lunch

16:00 Mass spectrometry & bioinformatics for metabolomics – *Steffen Neumann*

16:30 Statistics for mass spectrometry – *Sven Rahmann*

17:00 Medical diagnosis by Mass Spectrometry – *Anna Gambin*

19:30 Bio knowledge – my ideas and approaches - *David Gilbert*

20:00 Discussion: Can we represent biological knowledge?

Wednesday (*Martin Vingron*)

9:00 ArrayExpress – *Misha Kapushesky*

10:00 Coffee break

10:15 Ontology based data analysis – *Andreas Plendiani*

10:45 Signaling pathway – *David Gilbert*

12:00 Lunch

13:00 Trip to Trier

19:30 Hybrid Systems Identification – Application to cell signaling pathways – *Fujarewicz Krzysztof*

20:00 Live Demo – *Misha Kapushesky*

Thursday (*Jacek Blazewicz & Ralf Zimmer*)

9:00 "Surprise" – *David Corne*

9:30 Information extraction from scientific text – *Dietrich Rebholz*

10:00 Coffee break

10:30 Biclustering of Gene Expression Data – *Stefan Bleuler*

11:00 Mining heterogeneous data with mixture models - *Alexander Schliep*

12:00 Lunch

13:00 Discussions

16:00 Mining micro array data by association rules – *Clarisse Dhaenens*

16:30 Discussion in Data Mining (based on the three talks)

17:00 Evolution of paralog gene families – *Jerzy Tiuryn*

19:30 Outcome of working group on pathway data – *Ulf Leser*

19:45 Comparing overlapping data sources OR an alternative approach to data cleansing – *Johann-Christoph Freytag*

Friday

9:00 Discussion: Representing data on processes – *Ralf Zimmer, David Gilbert*

9:30 Graph Indexing – *Ulf Leser*

10:00 Coffee break

10:30 Wrap up: Results and open questions

12:00 Lunch & departure

4. Workshop Résumé

Based on the initial proposal the workshop discussed many of the topics mentioned, some of the ones discussed were not mentioned in the initial proposal. In particular, one of the discussion sessions focused on how to represent data/information on dynamic aspects of biological “knowledge”. Ralf Zimmer and David Gilbert presented various formalisms for handling such information. Based on discussions during the workshop on how to represent graph oriented data Ulf Leser showed how graphs could be represented in relational DBMS and how queries – using an extended SQL syntax – solve important graph problems using existing database technology.