R. Hofestädt

Bielefeld University, Faculty of Technology
Bioinformatics Department
Bielefeld, Germany

Synopsis

Bioinformatics

The papers selected for this section describe important topics of Bioinformatics. Methods of Molecular Biology are important for the future of Medicine. Therefore, methods of computer science, e.g. database systems, information systems and analysis tools, have to be developed and implemented. In Germany, the Ministry of Science recently started new programs to build Bioinformatic Centres at five universities and five research centres. The main reason to support this new research topic is the fact that the exponential growth of molecular data can only be handled using methods of computer science. The selected five papers in this section of the Yearbook 2002 demonstrate the current situation of the electronical infrastructure of Molecular Biology. The paper by Berman et al. describes the Protein Data Bank (PDB). Today more than 300 molecular databases are available via the Internet and PDB represents the knowledge of analysed proteins. Database systems that represent information about genes, gene regulation processes, metabolic reactions, enzymes, signal pathways etc. also exist. Beyond molecular database systems, different information systems are available that also allow information fusion to solve specific questions and problems. In this sense, the paper by Kolaskar and Naik shows an information system for identifying viruses. All these database systems only represent molecular data selected from published experimental results. For future aspects, modelling and simulation must be integrated and will help analyse the molecular mechanisms. Data modelling is one important aspect addressed by Paton et al.. Their paper presents a collection of data models for genomic data. However, to understand the metabolic behaviour of cell simulation, tools for metabolic processes have to be implemented. Rzhetsky et al. present a new model for the analysis of gene controlled metabolic networks. Modelling and simulation is the backbone for the implementation of the virtual cell. We are far from implementing the first version of this vision today. However, one important application of Bioinformatics is Medicine. The last paper, by Miller, represents a current discussion of the future of Bioinformatics in the field of Health Informatics. This paper is based on a presentation made to the Symposium of the American College of Medical Informatics in 2000.

Bioinformatics

Based on the Human Genome Project, the new interdisciplinary subject of Bioinformatics has become an important research topic during the last decade. Methods of Molecular Biology allow automatic sequencing and synthesis of nucleic and amino acids. Based on this technology robots able to sequence small genomes in one month's time are developed. The automatic assembly and annotation of the sequence data can only be done using methods of computer science. This is one of the main reasons for the success of this new research topic. Today, beside the genome and protein sequence data, a new domain of data is arising – the so-called proteomic project, which allows the identification of protein profiles. The molecular data is stored in database systems available via the Internet. Based on that data, different questions can be solved by implementing specific analysis tools. Regarding the DNA sequence, we are looking for tools that will predict DNA-functional units. Today, we call this topic "From the Sequence to the Function" or "Post Genomics". The main application area of this new research topic is Molecular Medicine. Therefore, Bioinformatics has also become an important topic of Medical Informatics. Regarding current definitions of Bioinformatics, we can see two different views: The German definition of Bioinformatics is a global definition. On the one hand, the application of the methods and concepts of computer science in biology represent the main focus. This is also the common definition. On the other hand, looking onto the history of computer science, we can identify important innovations coming from the analysis of molecular mechanisms. Regarding this aspect, we can distinguish between direct and indirect innovations. The implementation of neuronal networks, genetic algorithms or DNA Computing methods try to solve severe problems using molecular mechanisms. We can call this research topic the biological paradigm of computing. Moreover, the definition of formal systems, like the cellular automaton (J. v. Neumann), the finite state automaton (H. Kleene) or L-systems (A. Lindenmayer), is based on the idea of the implementation of analysis tools for modelling neural networks.

The common definition of Bioinformatics addresses the application of methods and concepts of computer science in the field of biology. Bioinformatics currently stresses three main topics. The first major topic is sequence analysis or genome informatics. Its basic tasks are: assembling sequence fragments, automatic annotation, and implementation of database systems, like EMBL, TRANSFAC, PIR, GENBANK, KEGG etc.. The sequence alignment problem still represents the kernel of sequence analysis tools. Their development and implementation represents the second aspect of sequence analysis. Nevertheless, sequence analysis is not a new topic. It was, and still is, a topic of Theoretical Biology or Computational Biology. Protein Design is the second current major research topic of Bioinformatics. The first task is to implement specific database systems that represent knowledge about the proteins. Today many different systems, like PIR or SWISSPROT, are available. The main idea of this topic still is to develop and implement a model, that will allow the automatic calculation of the 3D structure, including the prediction of the molecular behaviour of this protein. Until now, molecular modelling has been unsuccessful. Protein design is also not a new research topic. Its roots can be found in Biophysics, Pharmaco Kinetics and Theoretical Biology. The third current major Bioinformatics topic is Metabolic Engineering, which was defined by J. Bailey. Its goal is to analyze and synthesize metabolic processes. The basic molecular information of metabolic pathways is stored in database systems, like KEGG, WIT, etc.. Models and specific algorithms, based on the molecular knowledge represented by these database and information systems, allow the implementation of analysis tools.

Prospects for the future - Virtual Cell

The idea of Metabolic Engineering represents the basic idea of the Virtual Cell. Using molecular data and molecular knowledge, the implementation of specific models allows the implementation of simulation tools. Behind the algorithmic analysis of molecular data, modelling and simulation methods and concepts allow the analysis and synthesis of complex gene controlled metabolic networks. The current and available knowledge and data of Molecular Biology is still rudimentary. Furthermore, the experimental data available in molecular databases have a high error rate, while biological knowledge has a high rate of uncertainty. Therefore, only modelling and simulation methods will suffice to discuss arising important questions. Such formal descriptions can be used to specify of a simulation environment. Therefore, modelling and simulation can be interpreted as the basic step for implementing virtual worlds that allow virtual experiments.

The papers of this section show parts of the electronic infrastructure of Molecular Biology and the application of molecular data to model and simulate metabolic processes. The concepts available in literature are based on specific questions, such as the gene regulation process phenomena, or the biochemical process control. To solve current questions, we must implement integrative models which can be used to implement the virtual cell. If we take a look at the Internet, we can see that only online representations of cellular illustrations, taken directly from books, are available today (http://www.life.uiuc.edu/plantbio/cell/). The state of the art methods and concepts for the implementation of a virtual cell have been documented by the seminars organized at Dagstuhl, including a summer school focusing on: Modelling and Simulation of Gene Regulation and Metabolic Pathways

- http://www-bm.cs.uni-magdeburg.de/iti_bm/ibss/
- http://www-bm.cs.uni-magdeburg.de/iti_bm/dagstuhl/
- http://www-bm.cs.uni-magdeburg.de/iti_bm/dagstuhl2001/

Based on these events, MIT Press will publish a book by the end of 2001. One chapter of this book will include a description of M. Tomita's E-Cell system, which represents the first implementation of the virtual cell. His work represents a specific software solution and cannot be used globally (www.e-cell.org). Many new virtual cell projects are following the E-Cell project. However, it will take much time to implement a useful and powerful virtual cell. Rudimentary knowledge is one problem confronting the implementation of such systems. Furthermore, data and information are still missing. We are not yet able to understand the quantitative behaviour of simple metabolic processes.

Benefits

Bioinformatics will present the electronical infrastructure of Molecular Biology and will support drug design, molecular diagnosis and gene therapy. Based on molecular methods, modelling tools will be implemented that allow computer-supported design of new drugs. Information systems, in combination with methods of artificial intelligence, will support molecular diseases detection. Therefore, first information and expert systems for the detection of metabolic diseases as well as tools for the analysis of genotype/phenotype correlations have already been implemented. Diagnosis and therapy of metabolic diseases will be supported by database systems, knowledge based systems and molecular expert systems. Therefore, it is important to integrate Bioinformatics into the Medical Informatics curriculum. Thus, the future of Molecular Medicine is correlated with the future of Molecular Biology. Gene therapy is only one example, that demonstrates that methods of Bioinformatics are important. On the one hand, the analysis of all genes is supported by Bioinformatics methods. On the other hand, gene therapy is based on the idea of gene transfer methods. The molecular effect of the transfer of one or more genes into another organism must be tested and simulated using Bioinformatics methods, which allow the implementation of hypothetical worlds. The first task is to understand the gene regulatory mechanism, while the second task is to control the molecular effect of the gene product. The later can be discussed using the molecular information stored in the molecular database systems. Identification of negative side effects can only occur with help of available molecular data from different database systems, used in combination with modelling and simulation methods. Beside this scientific approach, simple methods of information fusion of molecular data can be used and implemented in modern health care systems today. These tools will support expert systems or simple information systems, that are able to monitor patient data, for example a specific drug therapy.

Barriers

Bioinformatics methods in use are database systems, information systems, analysis tools, modelling, and simulation. Evaluation processes of molecular database systems show that most database systems represent much incorrect and/or junk data. The mistakes are caused by false experimental and/or published data. Moreover, the copy process from the selected papers to the database entry, which is done by humans, also shows a high error rate. However, there are many scientists saying that most of the molecular database systems represent junk data. Althoug some tests have shown that the error rate is very high, it will not be easy to solve this problem in the future. Until now no efforts to implement software tools, that will reduce this error rate can be seen in the area of Bioinformatics. Another problem is that we are not able to implement analysis tools for many of the open questions. We need clear definitions and specifications to develop analysis tools. This is not the case in the field of Biology. For example the fundamental term „homology of sequences“ has hundreds of definitions. For this reason, so many different alignment algorithms exist. The other reason is the high complexity of time and space of most of these problems. Complexity is the main argument against the implementation of the virtual cell within the next decades. Finally, the main barriers come directly from Molecular Biology. Today, it seems as though we will never understand basic molecular mechanisms, such as the fundamental process of gene regulation.

Prospects for the near future

Information systems for scientists, patients and doctors, to represent the basic knowledge of Molecular Biology are available already. Moreover, their data is growing exponentially. Today, molecular data is available via the Internet and can be used to support therapy, diagnosis and drug design. The molecular diagnosis of metabolic diseases is a current research topic. Thousands of metabolic diseases are known and about 500 relevant inborn errors are discussed in the literature. Based on medical data of inborn errors, the German Human Genome Project initiated a project, to discuss the actual benefits of molecular information fusion in combination with modelling and simulation methods. Databases such as METAGENE, KEGG, TRANSFAC and MDCave will be integrated into this project. Using gene regulation and metabolic processes modelling, the analysis process will be supported. This running project shows that genotype/phenotype correlations can be identified using current molecular data and knowledge.

References:

1. Attwood TK, Parry-Smith DJ. Introduction to Bioinformatics. Prentice-Hall, Hempstead; 1999
2. Bailey J. Toward a Science of Metabolic Engineering. Science 1991;252:1668-74.
3. Collado-Vides J, Hofestädt R. Gene Regulation and Metabolism. MIT Press. In press 2001
4. Frenkel K. The Human Genome Project and Informatics. Commun ACM 1991; 11:41-51.
5. Goldberg DE. Genetic Algorithms in Search, Optimization and Machine Learning. Amsterdam Bonn Singapore: Addison-Wesley; 1989.
6. Heijne G v. Sequence Analysis in Molecular Biology. San Diego New York London Toronto: Academic Press; 1987.
7. Hofestädt R, Collado-Vides J, Löffler M, Mavrovouniotis M. Modelling and Simulation of Metabolic Pathways, Gene Regulation and Cell Differentiation. Bioessays 1996;18:333-5.
8. Hofestädt R. Computer Science and Biology. Biosystems 1997;43:69-71.
9. Hofestädt R, Lengauer T, Löffler M, Schomburg D, editors. Bioinformatics. Heidelberg: Springer; 1997.
10. Hofestädt R, Scholz U. Information Processing for the Analysis of Metabolic Pathways and Inborn Errors. Biosystems 1998;47:91-102.
11. Kanehisa M. Post-Genome Informatics. Oxford: Oxford University Press; 2000.
12. Pevzner P. Computational Molecular Biology: An Algorithmic Approach. Cambridge: MIT Press; 2000.
13. Setubal J, Meidanis J. Introduction to Computational Molecular Biology. Boston: PWS Publishing Company; 1997
14. Waterman M. Introduction to Computational Biology. Boston: Chapman Hall; 1992.

www-glossary

EMBL – Gene sequences
http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html

GEPASI – Biochemical simulation
http://gepasi.dbs.aber.ac.uk/softw/gepasi.html

KEGG – Metabolic pathways
http://genome.ad.jp

GENBANK – Gene sequences
http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html

HGMD – Human Gene Mutation Database
http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html

MatInspector – Promoter detection
http://genomatix.gsf.de

MDCave – Molecular data of inborn errors
http://mdcave.genophen.de

METAGENE – Medical data of inborn errors
http://www.metagene.de

OMIM – Online Mendelian Inheritance in Man
http://www.ncbi.nlm.nih.gov/Omim/

PIR – Protein information
http://pir.georgetown.edu/

PDB – Protein Data Bank
http://www.rcsb.org/pdb/

SRS – Database integration
http://srs6.ebi.ac.uk

SWISSPROT – Protein information
http://www.expasy.ch/sprot/

TRANSFAC – Gene regulation
http://transfac.gbf.de/

WIT – Metabolic pathways
http://wit.mcs.anl.gov/WIT2/

Address of the author:
Ralf Hofestädt
Bielefeld University, Faculty of Technology
Bioinformatics Department
P.O. Box 100 131
D-33501 Bielefeld, Germany
E-Mail: hofestae@techfak.uni-bielefeld.de