Bioinformatics' Challenges to Computer Science

Mario Cannataro[1], Mathilde Romberg[2], Joakim Sundnes[3],
and Rodrigo Weber dos Santos[4]

[1] University Magna Graecia, Catanzaro, Italy
cannataro@unicz.it
[2] Research Centre Julich, Germany
m.romberg@fz-juelich.de
[3] Simula Research Laboratory, Norway
sundnes@simula.no
[4] Federal University of Juiz de Fora, Brazil
rodrigo.weber@ufjf.edu.br

Abstract. The workshop Bioinformatics' Challenges to Computer Science covers the topics of data management and integration, modelling and simulation of biological systems and data visualization and image processing. This short paper describes the requirements Bioinformatics has towards computer science, summarizes the papers accepted for the workshop and gives a brief outlook on future developments.

Keywords: Bioinformatics, Data Management and Integration, Modelling and Simulation of Biological Systems, Data Visualization.

1 Bioinformatics - An Overview

Bioinformatics[1], a link between biology and computer science, involves the design and development of advanced algorithms and computational platforms to solve problems in biology and medicine. It also deals with methods for acquiring, storing, retrieving, analyzing, and integrating biological data obtained by experiments, or by querying databases.

Bioinformatics is providing the foundation for fast and reliable analysis of biological and medical data. Genomics, transcriptomics, proteomics, epidemiological, clinical and text mining applications have made essential progress through using bioinformatics tools. Although standard tools are widely offered through the Web, they are no longer sufficient to cope with the increasing demands of the complex analysis and simulation tasks of today's emerging fields of biotechnology research. Moreover, emerging life sciences applications need to use in a coordinated way bioinformatics tools, biological data banks, and patient's clinical data, which require seamless integration, privacy preservation and controlled sharing[2,3,4]. Therefore, new challenges to computer science arise from the sheer problem scale, the huge amounts of data to be integrated and the computing power necessary to analyze large data sets or to simulate complex biological systems.

M. Bubak et al. (Eds.): ICCS 2008, Part III, LNCS 5103, pp. 67–69, 2008.

2 Goals

The aim of the workshop was to bring together scientists from computer and life sciences to discuss future directions of bioinformatics algorithms, applications, and data management. The discussion evolves from the basic building blocks for Bioinformatics applications which are: data sources (e.g. experimental datasets, local and public biological databases); software tools providing specialized services (e.g. searching of protein sequences in protein databases, sequence alignment, biophysical simulations, data classification, etc.); and high level description of the goals and requirements of applications and results produced in past executions. From a computational point of view, bioinformatics applications bring a wide range of challenges and a huge demand for computing power. The challenging issues involve the large number of involved datasets, the size of the datasets, the complexity inherent in the data analysis and simulations, the heterogeneous nature of the data, and the need for a secure infrastructure for processing private data. From another perspective, emerging high performance computer architectures may offer huge computational resources in the trade of specific development of new algorithms for bioinformatics. These are the cases of Grid and Web services, as well as of multicore architectures that demand bioinformatics algorithms to be specifically tailored to these new computational frameworks.

3 Workshop Summary

The papers in the present workshop address a number of the requirements and challenges mentioned above, with special emphasis on data management and integration, modelling and simulation, and data visualization and image processing. A first group of papers discusses data management and integration issues in genomics and proteomics as well as in biomedical applications, including management of provenance data. In particular, Swain *et al.* present a data warehouse and repository for results from protein folding and unfolding simulations together with optimization strategies for the data warehouse that exploit grid tools. The ViroLab virtual laboratory is a common framework discussed by the papers of Balis *et al.* and Assel *et al.*. The first one uses ViroLab to investigate a data provenance model, whereas the second one discusses data integration and security issues for medical data. Mackiewicz *et al.* deal with the analysis of genetic code data and the minimization of prediction errors.

A second group of papers discusses modelling and simulation in systems biology and biosciences. In particular, Cui *et al.* demonstrate a computational model for calcium signalling networks. Cardiac modelling for understanding Chagas' Disease is the focus of Mendoça Costa *et al.* whereas Vega *et al.* present an adaptive algorithm for identifying transcription factor binding regions. The use of DNA microarray expression data for gene selection is exploited by Maciejewski to optimize feature selection for class prediction. Cannataro *et al.* discuss a meta tool for the prediction of possible protein combinations in protein to protein interaction networks.

The third group of papers discusses visualization of genomics data and biomedical images, as well as metadata extraction for e-learning purposes. In particular, Jakubowska *et al.* present a new technique for improving the visual representation of data in genomic browsers. Miranda Teixeira *et al.* discuss methods for automatic segmentation of cardiac Magnetic Resonance Images and Bułat *et al.* deal with algorithms for computer navigation systems assisting bronchoscope positioning. Kononowicz and Wiśniowski exploit the MPEG-7 meta data standard for medical multimedia objects to support e-learning.

The results and current research trends presented in the papers reassure the importance of especially data and meta data management and of computational modelling. In addition, the workshop includes discussion slots on topics not covered by the accepted papers, like bioinformatics middleware for future applications, full semantic integration of biological data banks, and utilization of new visualization techniques.

4 Conclusions and Outlook

Data integration and data management are under research for years but the complexity associated to biological processes and structures demand further investigations and the development of new methods and tools for bioinformatics. The simulation of complex biological phenomena becomes more feasible as the amount of information provided by new experimental techniques and the computing power rapidly increases. On the other hand, the development and use of complex computational models demand new implementations that better exploit the new computer architectures. Research on distributed and Grid computing for bioinformatics applications as well as related workflow modeling will enable emerging life sciences applications to use in a coordinated way bioinformatics tools, biological data banks, and patient's clinical data, that requires seamless integration, privacy preservation and controlled sharing.

Acknowledgements. We would like to thank the members of the program committee for the invaluable assistance on reviewing the papers and the organizers of ICCS for promoting this workshop.

References

1. Cohen, J.: Bioinformatics - an introduction for computer scientists. ACM Computing Surveys 36, 122–158 (2004)
2. Talbi, E.G., Zomaya, A.Y. (eds.): Grid Computing for Bioinformatics and Computational Biology. Wiley, Chichester (2008)
3. Cannataro, M. (ed.): Computational Grid Technologies for Life Sciences, Biomedicine and Healthcare. Information Science Reference. Hershey (in preparation, 2008)
4. Aloisio, G., Breton, V., Mirto, M., Murli, A., Solomonides, T.: Special section: Life science grids for biomedicine and bioinformatics. Future Generation Computer Systems 23, 367–370 (2007)