PERGAMON

Information Systems 28 (2003) 241–242

www.elsevier.com/locate/infosys

Preface

Special issue on data management in bioinformatics

Bioinformatics is the science of storing, extracting, organizing, analyzing, and interpreting biological data. Its importance has increased with the technology of DNA sequencing, microarrays, as well as the widespread understanding that genes and proteins act in networks. This in turn renders high performance data analysis algorithms central to various bioinformatics tasks.

The research topics in this area include data mining and warehousing as applied to biology, data types and modeling needed for biological analysis, interactive exploration and visualization for biology, and new indexing and search structures with applications to biology. This special issue on Data Management in Bioinformatics provides a collection of papers on recent advances in this area. It contains six articles selected from eighteen manuscripts submitted to the special issue.

The first paper, "Cancer classification using gene expression data," by Ying Lu and Jiawei Han, surveys various cancer classification methods and evaluates them based on computation time, classification accuracy and biological significance. The paper also discusses problems relevant to gene selection and gene contamination, and addresses issues related to significance and errors of a cancer classifier.

The second paper, "Mooshka: A system for the management of multidimensional gene expression data in situ," by Andrei Pisarev, Ekaterina Poustelnikova, Maria Samsonova and Peter Baumann, discusses the application of the array DBMS, RasDaMan, to the management of gene expression data in situ. Their system, called Mooshka, is able to facilitate processing and analysis of the information on the segmentation of gene expression, and is open for inclusion of new data and query types.

The third paper, "Genomic data modeling," by Jake Yue Chen and John V. Carlis, outlines the challenges in representing biological data, namely, the inherent complexity of biological data, domain knowledge barrier, the evolution of domain knowledge and the lack of expert data modeling skills. The authors propose several techniques for modeling genomic data, and summarize their experience into practical lessons and guidelines useful for novice biology data modelers.

The fourth paper, "Design and implementation of a string database query language," by Gosta Grahne, Raul Hakli, Matti Nykanen, Hellis Tamm and Esko Ukkonen, presents the language Alignment Declarations designed for string querying and restructuring. This language extends the capabilities of existing database query languages by allowing the user to define database predicates that express structural properties of strings (e.g. containment of certain patterns) or relations between several strings (e.g. similarity measures). Since strings are common in molecular biology, the proposed language has potential applications to querying DNA and protein sequences.

The fifth paper, "Information technology challenges of biodiversity and ecosystems informatics," by David Maier, Eric Landis, Judy Cushing, Anne Frondorf, John L. Schnase and Avi Silberschatz, provides an overview of the emerging field of biodiversity and ecosystem informatics. This paper demonstrates how the demands of biodiversity and ecosystem research can advance one's understanding and use of information technologies.

PII: S0306-4379(02)00046-7

Finally, the sixth paper, "The building of BODHI, a bio-diversity database system," by Srikanta J. Bedathur, Jayant R. Haritsa and Uday S. Sen, reports on the authors' experiences in building BODHI, an object-oriented database system for storing plant bio-diversity information. This paper describes various indexing techniques and the rule-based query optimizer employed by BODHI, and analyzes the system's performance on a representative set of queries.

It has been a pleasure for us to act as guest co-editors for this special issue. We would like to thank the authors of all the papers submitted to the special issue, and the external referees who helped us in the reviewing process.

Mohammed J. Zaki
Department of Computer Science
Rensselaer Polytechnic Institute, Troy,
NY 12180, USA
E-mail address: zaki@cs.rpi.edu

Jason T.L. Wang
Data and Knowledge Engineering Laboratory
Department of Computer Science
College of Computing Sciences
New Jersey Institute of Technology
University Heights, Newark, NJ 07102, USA
E-mail address: wangj@njit.edu