Artificial Intelligence in Medicine (2004) **31**, iii–iv

ELSEVIER

Artificial Intelligence in Medicine

http://www.intl.elsevierhealth.com/journals/aiim

GUEST EDITORIAL

Data mining in genomics and proteomics

Studies in genomics and proteomics are contributing to major improvements in the field of medicine. Genomics is generating biomedical information that is very relevant to clinical diagnosis and therapy. Proteomics is providing researchers with important data concerning the structure and function of proteins. All this information undoubtedly benefits fields like drug design and the protein industry. However, the amount of new information is so great that data mining techniques are essential in order to obtain knowledge from the experimental data.

This special issue focuses on data mining techniques that transform genomic and proteomic information into knowledge. The six papers included in the issue cover different data mining methods such as Bayesian networks, decision trees, feature subset selection, hierarchical clustering, *k*-NN, logistic regression, naïve Bayes, neural networks, principal components analysis, and rule induction.

The paper by Inza, Larrañaga, Blanco, and Cerroloza contributes to the field of gene selection in DNA microarray data sets. It makes an empirical comparison between filter and wrapper approaches in four classification paradigms (decision trees, *k*-NN, naïve Bayes, and rule induction) for colon and leukemia data sets. Most of the genes selected by the proposed filter and wrapper procedures appear on the lists of relevant genes detected by previous studies.

Randall, Baldi, and Villarreal's paper introduces a new resource containing structural information of poxvirus proteins: the Poxvirus Proteomics Database (PPDB). Poxviruses and the like are of special interest nowadays because they can be used as bioterrorist weapons. In the PPDB, the authors developed bioinformatics structure prediction tools, based on bi-directional recurrent neural networks and on a genomic scale, that render results in a format accessible to the public. The current version of the system contains both predicted and experimentally determined information about protein structure features, such as secondary structure and relative solvent accessibility, or tertiary structure and homology information.

In the paper by Robles, Larrañaga, Peña, Menasalvas, Pérez, Herves, and Wasilewska a new multi-classifier that combines some of the best secondary structure prediction methods is presented. The multi-classifier is based on a seminaïve Bayes model, learned from a wrapper approach, that combines the results provided by JPred, SSPro, PHD, PSIPRED, PROF, and SAM-TO2. The results obtained by the multi-classifier over nine different data sets show the superiority of the proposed approach.

In the next paper, Walker, Smith, Liu, Famili, Valdés, Liu, and Lach address the problem of dealing with microarray data that come from two known classes (Alzheimer and normal). They applied three separate data mining techniques—hierarchical clustering and two statistical tests—to discover genes associated with Alzheimer's disease. The 67 genes identified in this study included 17 genes that are already known to be associated with Alzheimer's and other neurological diseases. Twenty known genes, not previously related to the disease, have been identified, besides 30 uncharacterized expressed sequence tags.

Weber, Vinterbo, and Ohno-Machado investigate several algorithms to select gene classification markers. The authors test these algorithms through logistic regression, and demonstrate, using ten different data sets, that a conditionally univariate algorithm is a viable way of quickly determining a set of gene expression levels that can work as disease markers. The paper proves that the classification performance of logistic regression differs only slightly from the performance of more sophisticated algorithms applied in previous studies, and that the gene selection in the logistic regression paradigm is reasonable.

Finally, the main topic of the paper by Yoo and Cooper is modeling the expected value of experimentation—intervention and observation—to discover causal pathways in gene expression data. The system introduced, named GEEVE, helps biologists to discover gene-regulation pathways, recommending which experiments to perform and the number of measurements to include in the

0933-3657/$ – see front matter
doi:10.1016/j.artmed.2004.03.001

experimental design, and providing a Bayesian analysis that combines prior knowledge with the results of recent microarray experiments to derive posterior probabilities of gene-regulation relationships.

Acknowledgements

We would like to thank all the reviewers that have made this high-quality special issue possible: Elena Alvarez-Buylla, Instituto de Biología, UNAM, México. David Bowtell, Peter MacCallum Cancer Institute, Australia. Juan P. Caraça-Valente, Universidad Politécnica de Madrid, Spain. Chris Ding, Lawrence Berkeley National Laboratory, USA. Joaquín Dopazo, Centro Nacional de Investigaciones Oncológicas (CNIO), Spain. Vladimir Filkov, State University of New York at Stony Brook, USA. Jane Fridlyand, UCSF Cancer Center, USA. Georg K. Gerber, Massachusetts Institute of Technology, USA. Trond Hellem Bø, University of Bergen, Norway. Misha Kapushesky, European Bioinformatics Institute, UK. Javed Khan, National Cancer Institute, USA. Elliot J. Lefkowitz, University of Alabama, USA. Wentian Li, North Shore LIJ Research Institute, USA. Walter J. Lukiw, Louisiana State University, USA. Juan M. Marín, Universidad Rey Juan Carlos, Spain. Socorro Millán, Universidad del Valle, Colombia. Ursula Pieper, University of California, USA. Eran Segal, Stanford University, USA. Guillermo Soler-Espiauba, Hospital de León, Spain. Simon Tong, Stanford University, USA. Chris Upton, University of Victoria, Canada. Chen-Hsiang Yeang, Massachusetts Institute of Technology, USA.

[*]Pedro Larrañaga
Department of Computer Science and Artificial Intelligence, Universidad del País Vasco Aptdo. 649, San Sebastian, Donostia 20080, Spain

Ernestina Menasalvas
Department of Computer Science Universidad Politécnica de Madrid, Spain

José M. Peña
Víctor Robles
Department of Architecture and Technology Universidad Politécnica de Madrid, Spain

[*]Corresponding author.
E-mail address: ccplamup@si.ehu.es (P. Larrañaga).