Pacific Symposium on Biocomputing 5:362-363 (2000)

Data Mining and Knowledge Discovery in Molecular Databases

Janice Glasgow
Department of Computing and Information Science
Queen's University
Kingston, Canada
janice@qucis.queensu.ca

Igor Jurisica
Faculty of Information Science
University of Toronto
Toronto, Canada
jurisica@fis.utoronto.ca

Raymond Ng
Department of Computer Science
University of British Columbia
Vancouver, Canada
rng@cs.ubc.ca

The development and growth of molecular databases over the last decade has brought with it a growing problem to the biocomputing community. Our ability to analyze, summarize and extract information from these databases has lagged far behind our ability to collect and store data. As well, traditional methods for handling data (either automated or manual) cannot be effectively applied because of the volume and complexity of these emerging databases.

Knowledge discovery generally refers to the process of identifying valid, novel and understandable patterns. Knowledge discovery from large databases, often called data mining, refers to the application of the discovery process on large databases or datasets. The discovery process can be broken into several steps, including: developing an understanding of the application domain; creating a target data set; data cleaning and preprocessing; finding useful features with which to represent the data; searching for patterns of interest; and interpreting and consolidating discovered patterns.

Research in molecular data mining and knowledge discovery has several important application areas, including protein structure prediction and drug design. For example, techniques for inverse protein folding require the extraction of useful information concerning the relationship between sequence and structure from protein databases. One use of data mining in the drug discovery process is to find common attributes or structural features for molecules with similar function.

Four manuscripts were accepted for publication and oral presentation in this session of PSB 2000. These papers represent different approaches to data mining as well as a variety of application areas in molecular biology. Nakaya, Hishigaki and Morishita apply data mining techniques to determine combinations of marker loci with respect to oral glucose tolerance. This was achieved by constructing a conjunctive rule loci to formalize genotype judgement. Their method was tested on the OLETF model rat. A novel statistical approach that uses stochastic segment models of eukaryotic promoter regions is presented by Ohler, Stemmer and Harbeck. They demonstrate that a five-state segment model improves over their previous approach which modeled the region as a whole. Whelan and Glasgow extend previous work in relational instance based learning to define an approach to data mining that is applied to the identification of amino acid residues in medium resolution electron density maps. The problem of finding similar sequence motifs is addressed by Wareham, Jiang, Zhang and Trendall. They approach this problem using a modified Gibbs Sampling heuristic and apply their approach to both simulated and real datasets.

Data mining and knowledge discovery have been topics considered at many artificial intelligence, database and statistical conferences. However, PSB offers the opportunity to focus on the application of these techniques in the area of molecular biology. Given the response to the call for papers for the session, and the quality of the papers submitted, it is obvious that it is an area of high interest for the community and there remains much research to be carried out in this emerging and exciting field.

Acknowledgements

The session co-chairs are grateful to the reviewers for their careful comments and insightful suggestions: Alan Ableson, Dianne Cook, Aris Floratos, Michael Gribskov, Satoru Miyano, Rebecca Parsons, Isidore Rigoutsos, Steven Salzburg, Jason Wang, Ken Whelan.