Pacific Symposium on Biocomputing 4:124-125 (1999)

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

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 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; data mining to search 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.

Six manuscripts were accepted for oral presentation in this session of PSB '99. These papers represent different approaches to data mining as well as a variety of application areas in molecular biology. Giegerich, Haase and Rehmsmeier propose a software approach to the reversible change of conformation in an RNA molecule. This work applies a clustering approach that incorporates an energy barrier distance in order to predict such structural

switching. The work of Savoie, Kamikawaji, Sasazuki and Kuhara involves the understanding of sequence motifs that affect T cell activation. They incorporate a technique referred to as the Bonsai algorithm to cluster data and produce rules that differentiate between positive and negative amino acid sequences. Murakami and Takagi incorporate a clustering algorithm, similar to the k-means clustering method, in order to detect new motifs in 5' splice sites of mRNA. A probablistic approach to multiple sequence alignment is presented by Lazareva and Haussler; the approach is a special case of a profile HMM. Williams' paper focuses on the retrieval aspect of data mining. He demonstrates that unselective filtering reduces the effectiveness of retrieval and proposes a novel technique for filtering that incorporates a stop-list of frequently occurring subsequences. Finally, the paper by Su, Cook and Holder is concerned with the discovery of structural regularities in protein sequences. Using a system called SUBDUE, they obtain secondary structure patterns for a variety of proteins in the PDB.

In addition to the oral presentations, several high-quality papers were selected for publication and poster and/or panel presentation. The main objectives of the panel were: 1) to determine to what degree existing data mining techniques for alphanumeric and relational data can be applied to the domain of molecular biology, and 2) to identify and define problems and issues in data mining and knowledge discovery that must be addressed to successfully mine sequence and structure databases and to increase our knowledge and understanding of the underlying relationships between sequence, structure and function.

Data mining and knowledge discovery have been topics considered at many AI, database and statistical conferences. However, this is one of the first opportunities 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.

Acknolwedgements

The session co-chairs are grateful to the reviewers for their careful comments and insightful suggestions: Alan Ableson, Dianne Cook, Paul Kearney, Steven Salzburg, Evan Steeg, Michael Gribskov, Tao Jiang, Aleksandar Milosavljevic, Jason Wang, Satoru Miyano, Rebecca Parsons, Igor Jurisica, Kim Baxter, Jean Gerster, Peter Karp, Kyuseok Shim, Dan Suciu, Edwin Knorr.