Prior Knowledge for Discovery Proteomics Data Interpretation

Vanathi Gopalakrishnan

Keywords: Proteomics, Mass Spectra, Data Mining, Task-specific Domain Knowledge

1 Introduction.

Discovery proteomics usually refers to the search for potential biomarkers of disease from clinical samples such as blood serum, plasma, urine or cerebrospinal fluid (CSF). These samples are typically analyzed using high-throughput Mass Spectral (MS) technologies yielding high-dimensional data that needs interpretation in order to identify a small number of protein/peptide biomarkers predictive of disease states. This research provides an analysis of a discovery process leading to identification of potential biomarkers of a neurodegenerative disease called Amyotrophic Lateral Sclerosis (ALS). Task-specific domain knowledge is identified that can be used as priors to guide the search. Ongoing research involves the identification of such prior knowledge, and its use within rule learning and probabilistic machine learning systems to enhance the accuracy of biomarker panels predictive of disease states.

2 Background.

The use of high-throughput MS technologies such as SELDI-TOF (Surface Enhanced Laser Desorption Ionization Time of Flight) is increasingly becoming common for disease-specific biomarker discovery. In [1], we describe a case study involving the use of a rule-learning program (RL-Wrap) to the analysis of proteomic MS data from cerebrospinal (CSF) fluid samples of patients with ALS, a rapidly neurodegenerative disease for which early diagnosis is fuzzy, and there is no current cure. Three biomarkers found by the program were identified using Peptide Mass Fingerprinting (PMF) and tandem mass spectral analyses (MS-MS). Two of the larger proteins were

---

1 Department of Biomedical Informatics, University of Pittsburgh, PA. E-mail: vanathi@cbmi.pitt.edu
further validated using immunoblotting and immunohistochemistry on a separate patient cohort. The biomarker discovery process adopted in this case study is depicted in Figure 1 from [1].

3 Prior Knowledge Incorporation.

In another task related to proteins, namely that of macromolecular crystallization which is needed for three dimensional structure determination by X-Ray crystallography, we have utilized methods by which prior knowledge could be effectively encoded and utilized by a learning system [2]. In our ongoing research (Figure 2), we aim to systematize the task-specific prior knowledge for several sub-tasks depicted in Figure 1. Some of this involves the development and maintenance of (1) lists of proteins normally found in various clinical samples such as CSF or plasma; (2) relationships between proteins as found in pathways; and (3) the encoding of specific domain knowledge such as bond fragmentation patterns in MS data.

References
