Poster I-7

Analyzing high-throughput biological data: complexities, issues, challenges and some lessons learned

Authors:
Fazel Famili (*IIT*)
Sieu Phan (*IIT*)
Ziying Liu (*IIT*)
Junjun Ouyang (*IIT*)
Youlian Pan (*IIT*)
Alan J. Barton (*IIT*)

Short Abstract: This poster consists of three parts. In part one, we provide an overview of data mining in genomics and the BioMine project. In part two we describe some of our case studies using the BioMiner data mining software that we have built. Part three includes our experiences in this project.

Long Abstract:
Knowledge discovery is the process of developing strategies to discover useful and (ideally all) previously unknown knowledge from historical or real-time data. Applied to the analysis of high-throughput biological applications, knowledge discovery processes will help in various research and development activities. Examples of some of the efforts reported, including our own on genomics microarrays, are: (i) studying data quality for possible anomalous or questionable expressions of certain genes or experiments, (ii) investigating gene response to different treatments, or special conditions, (iii) identifying relationships between genes, between proteins and their functions based on time-series or other forms of high throughput genomics/proteomics profiles, and (iv) discovering models for clinical diagnosis/classifications based on expression profiles.
This poster consists of three parts. In part one, we provide an overview of data mining in genomics and the BioMine project. In part two we describe some of our case studies using the BioMiner data mining software that we have built in this project. These are all cases in which real genomics data sets have been used for tasks such as gene function identification and gene response analysis. Some of these applications are targeted towards biomarker identifications in which one would attempt to develop models for prediction and accurate diagnosis of certain diseases, such as cancer. The poster will include some of the scientific results from our case studies. We will describe a few examples explaining complexities and challenges in dealing with real data from genomics field. These include: (i) access to sufficient amount of data, (ii) lack of proper standards, which means dealing with multiple platforms and overlapping genes, (iii) production-specific factors such as issues related to probes, and spots, (iv) rapid evolution of unit identifiers, such as gene's or protein's ids, (v) availability of relevant clinical data, and (vi) proper validation of the results (e.g. gene response models, identified genes).
In the last part of this poster, we share our experiences gained over the last 5 years and describe our future plans. Some of the lessons learned are: (i) complete understanding of the domain/problem is extremely important, (ii) interaction with domain experts is essential, (iii)

proper understanding of the data, data pre-processing and the right data analysis strategy are the key issues for success. Examples are: normalization, data selection/reduction strategy, data re-representation and defining the right data mining processes/strategies and use of the most suitable methods/algorithms for knowledge discovery. We think that the knowledge discovery of biological data is at a point where there is a potential for BioIntelligence applications where one can validate, summarize, structure, enhance, and maintain all this knowledge, that in many cases will be useful to a number of researchers and practitioners in the field. We introduce BioIntelligence in this poster and present what we are planning to do in this research direction.