Workshop on Data Mining in Computational Biology and Bioinformatics

ABSTRACTS

DAY I:

12.45 pm > Mohammed Alshalalfa (University of Calgary- Canada) **Introduction to Bioinformatics: Integration of wetlab and computational techniques**

This talk will start with a brief historical coverage to highlight how the interest in analyzing the cell and its internal systems has developed over time leading to the completion of the genome project. The speaker will also describe the invention of the microarray technology which has produced huge amounts of data that are beyond human capabilities for manual analysis. Hence, researchers realized the need for sophisticated computational techniques to eliminate noise and to reduce redundancy in the data for better analysis and discovery.

01.30 pm > Reda Alhajj (University of Calgary & Global University Lebanon)

Introduction to Data Mining: Building Association Rules Models to Serve Bioinformatics Applications

Market basket analysis has been the main motivation for developing association rules mining methods. However, it is a general model that could fit any application which can be modeled in terms of baskets and items. The speaker will give the background of association rules mining and will describe how the model could be adapted for knowledge discovery in bioinformatics applications.

02.00 pm > Tansel Özyer (TOBB University- Turkey) **Clustering Techniques and their impact on Gene Expression Data Analysis**

Clustering has been successfully used to serve a wide range of applications ranging from image processing to web mining. The speaker will cover the background necessary to understand the basic concepts and will describe the main clustering techniques with concentration on multi-objective genetic algorithms based clustering. The talk will also discuss the importance of clustering for analyzing gene expression data.

02:30 pm > Kievan Kianmehr (University of Calgary- Canada) **The Importance of Classification for Bioinformatics Applications**

This talk will concentrate on classification as an attractive technique that fits well bioinformatics applications. While clustering is unsupervised learning, classification is supervised learning because the classes/categories are known in advance and the main target of the process is to consider some existing samples to build a model capable of classifying new samples into the known classes. The speaker will describe some basic model construction techniques are will comment on their successful application for classifying gene expression data.

DAY II:

09:30 am > Reda Alhajj (University of Calgary- Canada & Global Univ): **A multiagent based approach for Motif Discovery and Translation Initiation Site Prediction**

Motif discovery and translation initiation site (TIS) prediction recently received considerable interest from both computational biologists and computer scientists. Identifying motifs is greatly significant for understanding the mechanism behind regulating gene expressions. At least equally important is TIS prediction which leads to the proper protein formulation. This talk highlights the importance of multi-agents in discoveries motifs and TIS. Such an approach is attractive because both problems are complex and require multiple perspectives to be well covered for robust techniques that lead to consistent discoveries. For motif discovery our method is based on the following observation: if some elements are conserved, then these elements may be part of a conserved motif. Further, the proposed approach is based on the divide and conquer concept, where we divide each DNA sequence into four subsequences, one subsequence per each of the four letters, representatives of the nucleotides, namely {A, C, G, T}. For TIS prediction, we apply different perspectives to discover all possible TIS occurrences and then decide on the real TIS by voting and negotiation.

10.00 am > Mohammed Alshalalfa (University of Calgary- Canada): **A multiagent based approach to identify disease biomarkers: the cancer case**

This talk addresses an important and vital problem within the general area of disease recognition, namely identifying disease biomarker genes. Given the complexity of this domain, the basic idea tacked in this paper is employing multiple agents to handle this problem. Though the developed methodology is general enough to be applied to any other domain, we concentrate on identifying cancer biomarkers in this paper. Our approach is mainly based on detecting the minimum set of genes that could successfully identify cancer samples. Multiple agents are involved in the process. After each agents applies its own rules and reports candidate cancer biomarkers, the agents negotiate to agree on the actual

biomarkers. The latter process may require further investigation of the characteristics of each of the reported genes because some of them may have the same functionality and the target is a compromise of the best representative of each functionality set. A degree of confidence in each candidate biomarker influences the negotiation process. The conducted experiments reported very encouraging results with high classification rate; none of the involved agents could alone achieve a close success rate.

10.30 am > Kievan Kianmehr (University of Calgary- Canada): **Feature Reduction as Effective Approach to Identify Important Genes**

In this research project, we take advantage of using fuzzy classifier rules to capture the correlations between genes. The main motivation to conduct this study is that a fuzzy classifier rule is essentially an “if-then" rule that contains linguistic terms to represent the feature values. This representation of a rule that demonstrates the correlations among the genes is very simple to understand and interpret for domain experts. In our proposed gene selection procedure, instead of measuring the effectiveness of every single gene for building the classifier model, we incorporate the impotence of a gene correlation with other existing genes in the process of gene selection. That is, we reject a gene if it is not in a significant correlation with other genes in the dataset. Furthermore, in order to improve the reliability of our approach, we repeat the process several times in our experiments, and the genes reported as the result are the genes selected in most experiments. We report test results on several datasets and analyze the achieved results from biological perspective.

11.00 am > Dr. Asmaa Hamze (Global University- Lebanon) **Engineered and Natural Mutations in Tissue Inhibitors of Metalloproteinases**

Tissue inhibitors of metalloproteinases (TIMPs) comprise a family of proteins that modulate the turnover of the extracellular matrix by regulating the activities of endopeptidases that catalyze its degradation, especially the matrix metalloproteinases (MMP). Generally, TIMPs are broad-spectrum tight-binding inhibitors of MMPs with individual differences in specificity. Our research involves engineering TIMPs mutants to make them more selective

for certain MMPs to target related disease mechanisms. In order to study how certain mutations of amino acids affect function of TIMPs, we engineer the mutants by site-directed mutagenesis. Thereafter, we express the recombinant protein and perform fluorescence kinetic assays. The values measured in such assays are analyzed and compared to those of wild-type TIMP, thereby determining relative activity of these inhibitors.
The other aspect of our research focuses on TIMP-3 mutations that occur naturally in Sorsby's Fundus Dystrophy patients. In order to determine how such mutations affect the protein properties and ultimately lead to disease, we have chosen to imitate one of the TIMP-3 mutants found in a severe form of Sorsby's Fundus Dystrophy. We expressed the recombinant form of the protein in bacteria to produce large amounts for biochemical characterization. We also cloned the mutant into a mammalian expression vector to study effects on human retinal cells.

11.30 am > break

12:00 noon >. Mohammed Alshalalfa (University of Calgary- Canada) **Construction and analysis of Gene Regulatory Networks**

Due to the complex structure and scale of gene regulatory networks, we support the argument that combination of multiple types of biological data to derive satisfactory network structures is necessary to understand the regulatory mechanisms of cellular systems. In this talk, we will describe a simple but effective method of combining two types of biological data, namely microarray and transcription factor (TF) binding data, to construct gene regulatory networks. The proposed algorithm is based on and extends the well-known PC algorithm [23]. Further, we developed a method for measuring the significance of the interactions between the genes and the TFs. The reported test results on both synthetic and real data sets demonstrate the applicability and effectiveness of the proposed approach; we will also report the results of some comparative analysis that highlights the power of the proposed approach.

12:30 pm Tansel Özyer (TOBB Univ- Turkey) **Discovering Accurate and Interesting Classification Rules Using Genetic Algorithm and its Application for the Analysis of Clinical and Gene Expression Data**

Discovering accurate and interesting classification rules is a significant task in the post-processing stage of a data mining process. Therefore, an optimization problem exists between the accuracy and the interesting metrics for post-processing rule sets. To achieve a balance, in this talk, we describe two major post-processing tasks. In the first task, we use a genetic algorithm (GA) to find the best combination of rules that maximizes the predictive accuracy on the sample training set. Thus we obtain the maximized accuracy. In the second task, we rank the rules by assigning objective rule interestingness measures (or weights) for the rules in the rule set. Henceforth, we describe a pruning strategy using a GA to find the best combination of interesting rules with the maximized (or greater) accuracy. The application of this framework to clinical and gene expression data produced excellent results.

01.00 pm > Mohammad Salah (Cairo University- Egypt) **Development of a DNA - Fingerprinting Assay for Animal Species Identification of Meat using Polymerase Chain Reaction (PCR)**

Now in Egypt, an assay for forensic identification of different animal species using Polymerase Chain Reaction (PCR) analysis has been developed. PCR analysis allowed the discrimination between meat tissues of donkey and the most used species for human consumption based on targeting a conserved region of the mitochondrial cytochrome *b* gene (*Cyt b*). Using bioinformatics tools, GenBank databases were searched for Cytochrome *b* gene sequences (1140 bases) of donkey (*Equus asinus*), Cattle (*Bos taurus*), Buffalo (*Bubalus bubalis*), Camel (*Camelus dromedaries*), and Sheep (*Ovis aries*). Multiple alignment of different mitochondrial Cytochrome *b* gene sequences using T-COFFEE (http://www.ch.embnet.org/) revealed significant identity between cattle and buffalo DNA, so, a common single primer set for both species was designed. All primers were designed using *Primer 3* software (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_ www.cgi) and

genomic DNA was extracted from raw meat samples using Phenol: Chloroform: Isoamyl standard method. The good quality of extracted genomic DNA increased the efficiency of primers to generate specific fragments of 105, 307, 391 and 510 base pairs length for donkey, cattle and buffalo, camel, and sheep, respectively. Such rapid, simple and highly specific polymerase chain reaction will be a potentially reliable technique for accurate forensic identification of meat origins, additionally; this PCR-based approach opens a new avenue to veterinary forensic medicine on the molecular way of animal-species identification.

01.30 pm > Lunch

02.45 pm > Mohammed Alshalalfa (University of Calgary- Canada): **The path to the 1000$ human genome project?**

After the completion of the human genome project in 2001, which had cost more than 3 billion dollars, it became now important to sequence DNA to understand the variation among organisms; and more importantly, differences among humans. Any two humans are more than 99 percent the same at the genetic level. However, it is important to understand the small fraction of genetic material that varies among people because it can help explain individual differences in susceptibility to disease, response to drugs or reaction to environmental factors. The 1000 Genomes Project, launched in January 2008, is an international research effort to establish by far the most detailed catalogue of human genetic variation. The big obstacle in this project is the cost of sequencing; Re-sequencing a human genome with Sanger method would today cost ~$10 million. This talk will go over the most recent developed sequencing technologies which will allow sequencing the human genome with less than 1000$

03:15 pm > Kievan Kianmehr (University of Calgary- Canada) **Fuzzy discretization and Its Application for Gene Expression Data Analysis**

This talk presents a novel classification approach that integrates fuzzy class association rules and support vector machines. A fuzzy discretization technique based on fuzzy c-means clustering algorithm is employed to transform the training set, particularly quantitative attributes, to a format appropriate for association rule mining. A hill-climbing procedure is adapted for automatic thresholds adjustment and fuzzy class association rules are mined accordingly. The compatibility between the generated rules and fuzzy patterns is considered to construct a set of feature vectors, which are used to generate a classifier. The reported test results show that compatibility rule-based feature vectors present a highly-qualified source of discrimination knowledge that can substantially impact the prediction power of the final classifier. In order to evaluate the applicability of the proposed method, we show how this method provide biologists with an accurate and more understandable classifier model compared to other machine learning techniques.

03:45 pm > Mohammad Salah (Cairo University- Egypt) **Tumor Suppressor Gene p53 as a Universal Marker for Forensic Human and Non-human Species Identification**

Polymerase Chain Reaction (PCR) assay is one of the most powerful tools used for differentiation between animal species. So, in order to develop a universal primer for different animal identification, bioinformatics tools were applied.

Multiple alignments of sequences of different chromosomal and mitochondrial markers from some animal species revealed no significant conservation between aligned sets. Consequently, different species-specific primers were developed for animal species identification [Moawad and Aref 2006]. This study investigated and tested the possibility of tumor suppressor protein *p53* to be as a potential marker for developing a universal DNA-based PCR assay to allow discrimination between different animal species in singleplex reaction. Such discrimination assay may have important applications in the forensic science, agriculture, quarantine and customs fields. DNA samples were extracted from blood of five different human being as well from five animal individuals within the same species

followed by PCR amplification and agarose gel electrophoresis. DNA amplicons bound to ethidium bromide was transilluminated using UV and results were imaged by gel documentation system. Tumor suppressor protein *p53* gene showed very specific and discriminatory results since this marker produced greatest fragment size differences between human and non-human species studied. Sample differentiation for different species was possible following *p53* amplification, suggesting that this gene could be used as a potential animal species identifier.

04:15 pm > Reda Alhajj (University of Calgary- Canada & Global University): **Social Network Construction and Analysis: Discovering Communities of Genes/Proteins**

The foundations of social network research were established in 1930s. However, it was only in the 1970s that researchers started to study social networks from sociological point of view by mainly analyzing communities of humans and animals. Recently, an interest in the discovery of social communities has received more attention, in particular since the appearance of web-based communities. This has led to a shift into the automated analysis of social networks using data mining and machine learning techniques which is very promising and expected to dominate several aspects of daily life. In this talk, we consider genes as actors of a social network, a research area that has not yet received attention in the literature of social network mining and analysis. Even though our research project covers both genes and proteins, we concentrate in this talk on genes; we first try to describe the gene expression data and how genes interactions can be realized as a social network. Then we describe how data mining techniques could reveal important information by identifying disease biomarkers from the social communities of genes. This is possible because of the way genes interact and form communities that are anticipated to have certain effects on the different processes that take place within an organism. Gene communities both contribute to the development of an organism by coding proteins, and on the other hand gene communities cause serious diseases.

04.45 pm > Closing