14

Data Mining Applications in the Post-Genomic Era

Eugenia G. Giannopoulou[1] and Sophia Kossida[2]
[1]*University of Peloponnese, Department of Computer Science and Technology,*
[2]*Biomedical Research Foundation of the Academy of Athens,*
Greece

1. Introduction

The post-genomic era involves the experimental and computational efforts that aim to address the challenge of clarifying and understanding the function of the genes and their products. In particular, functional genomics intends to exploit the great wealth of data produced nowadays by high-throughput methods, in order to describe gene and protein functions and their interactions.

Proteomics, playing a significant role in this endeavour by complementing other functional genomics approaches, encompasses the large-scale analysis of complex mixtures, including the identification and quantification of proteins expressed under different conditions, the determination of their properties, modifications and functions. Although the term proteomics was initially defined as the large-scale study of the functions of all expressed proteins within an organism, it now also evokes the set of all protein isoforms and modifications as well as the interactions between them. In other words, proteomics expanded to the point that it now integrates all information that could be characterized "post-genomic" (Tyers & Mann, 2003). Therefore, the area of proteomics can be broadly divided into two subcategories: protein expression mapping and protein interaction mapping (Palzkill, 2002).

The protein expression mapping area uses separation and identification methods, such as 2-Dimensional Gel Electrophoresis (2DGE) and Mass Spectrometry (MS) respectively, to perform differential analysis of proteome expression levels (e.g., normal versus diseased cells, diseased versus treated cells and so on). Also known as *differential proteomics,* its basic aim is to discover reliable biomarkers for different biological states, to be used for diagnostic or therapeutic purposes, by identifying proteins that are up- or down-regulated or modified in a disease-specific manner (Monteoliva & Albar, 2004).

The protein-protein interaction mapping (also known as functional proteomics) includes the determination of protein interactions that exist in a proteome and aims at inferring unknown functions of specific proteins. Through these interactions, proteins carry out most of the processes in cells, such as gene regulation, intracellular communication and others. By exploiting protein interactions, it is feasible for the researcher to deduce the unknown function of a protein from the known functions of its interaction partners (i.e., proteins that participate in the same interaction). For example, if protein A with unknown function is found to participate in an interaction with proteins B and C, which are known to be

Source: Data Mining in Medical and Biological Research, Book edited by: Eugenia G. Giannopoulou,
ISBN 978-953-7619-30-5, pp. 320, December 2008, I-Tech, Vienna, Austria

involved in cellular process X, then protein A could be inferred to play a part in X as well. Thus, protein-protein interaction maps of the cell assist in the comprehension of its biology.
High-throughput technologies are widely used in proteomics, in order to achieve the analysis of thousands of proteins. These technologies generate high-dimensional proteomics data which require the application of different data mining approaches for efficient and accurate analysis of the proteomics results. More specifically, the application of data mining techniques on large-scale proteomics data sets can assist in many ways the data interpretation; it can reveal protein-protein interactions, improve the protein identification, evaluate the experimental methods used and facilitate the diagnosis and biomarker discovery, to name a few.
This chapter aims to: (a) familiarize the user with the most commonly used proteomics analysis methods, (b) present data mining techniques that have been used in the broad field of proteomics and (c) demonstrate indicative examples that highlight the importance of these methods in biological research.
In this chapter we will proceed as follows. In Section 2, we familiarize the reader with the typical proteomics workflow, introduce proteomics definitions and explain the benefits of applying mass spectrometry-based proteomics. In section 3, we discuss basic data mining methods, their characteristics and application to proteomics studies. Section 4 presents outstanding application paradigms, which confirm that data mining approaches are needed at all levels of proteomics analysis to provide a wealth of information. Finally, in the conclusion section, we summarize the chapter and discuss possible future research directions in this field.

2. Proteomics basics

2.1 Mass spectrometry-based proteomics workflow

A proteomic analysis includes two basic steps: application of a separation method (e.g., 2-Dimensional Gel Electrophoresis (2DGE), Liquid Chromatography (LC)), followed by a mass spectrometry (MS)-based identification method (Liebler, 2002). As far as the separation step is concerned, it is important to note that although two-dimensional electrophoresis (2DGE) is the separation technique most frequently used in proteomics, lately Liquid Chromatography (LC) is gaining momentum due to its ability to detect low abundance proteins and peptides (Garbis et al., 2005; Neverova & Van Eyk, 2004).
More specifically, 2DGE is applied to a complex protein mixture in order to separate its proteins in the highest possible degree. The protein mixture is inserted into a polyacrylamide gel and is resolved in two dimensions, the isoelectric point (pI) (i.e., the pH at which a particular molecule carries no net electrical charge) and the molecular weight (MW). As a result, proteins move to certain places in the polyacrylamide gel and after staining, they form the well-known gel spots. The outcome of this technique, a 2D-gel image, is then subjected to image analysis so as to detect and extract the spots from the gel. At this point, several statistical and quantitative methods (e.g., Mann-Whitney test, Student's t-test, volume fold factor criterion) are applied in order to perform differential expression analysis and detect the spots that discriminate the biological states.
After the extraction of the gel spots and their enzymatic digestion, the resulting peptides are inserted into a mass spectrometer, which produces spectra of their mass-to-charge (m/z) ratio. Using *peptide mass fingerprinting* (PMF) method, the information that the spectra carry for every peptide mass that appears in them (i.e., the m/z and the intensity), is used from

software tools to achieve protein identification. In particular, search engines (e.g., Mascot, Sequest) compare and match the measured masses with already known theoretical masses in protein databases and provide identification results which show the protein most likely contained in each gel spot.

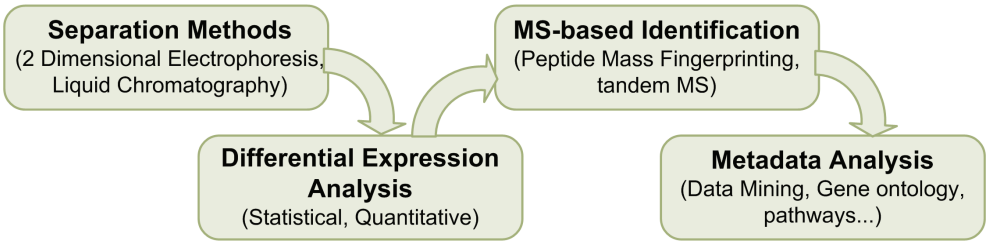

Fig. 1. Steps of a typical MS-based proteomics workflow.

In the case of LC coupled to tandem MS (LC-MS/MS), the mixture of many different proteins is digested to yield peptides, which are then resolved into fractions by two- or multi- dimensional liquid chromatography. The peptide fractions can then be processed by *tandem Mass Spectrometry* (MS/MS) to generate amino acid sequence information. This amino acid sequence can be used for protein database searching to identify the protein of interest (Liebler, 2002). As it is obvious, the most important benefit of the tandem mass spectrometry compared to peptide mass fingerprinting, apart from increasing selectivity and specificity in protein identification, is that the amino acid sequence information of the peptides is more precise for the protein identification than the peptides masses.

Differential proteomics is a powerful approach since it offers the ability to measure the relative abundance of proteins between two or more different biological states. A variety of quantitation approaches have been recently introduced, most of which use methods of stable isotope labeling (e.g., iTRAQ, ICAT, SILAC) that determine accurately, through the mass spectrometer, the relative changes in the two (or more) states of the proteome (Flory et al., 2002; Griffin et al., 2003; Gygi et al., 1999; Zieske, 2006).

The MS-based identification and quantitation is not the last step in a proteomic analysis workflow. Meta-data analysis follows and includes several methods that assist in the interpretation of the high-dimensional proteomics results. Several data mining and statistical methods (e.g., clustering, classification, ANOVA and so on) have been used extensively to facilitate the proteomics results interpretation (Beer et al., 2004; Bensmail et al., 2005; Cannataro et al., 2007; Hilario et al., 2006; Ventoura et al., 2008). Moreover, matching the identified proteins to their corresponding Gene Ontology annotations (i.e., molecular function, biological process, cellular component) is a task that places the protein results in a global perspective (Halligan et al., 2007). Lately, it has also become of immense importance to retrieve pathway information from databases (Krishnamurthy et al., 2003), in order to discover the biological pathways and molecular networks, that the proteins participate in. In summary, meta-data analysis includes using different methods and tools which focus on inferring protein-protein interactions, understanding how proteins are organized into biological networks and generally speaking, perceiving proteomics data from a "systems biology" point of view. Figure 1 summarizes the steps of a proteomics workflow, as described above.

2.2 Protein-protein interactions discovery

The proteome-wide scale discovery of physical interactions among proteins, plays a key role in the functioning of cells, since it assists in understanding the function of each protein. The integration and visualization of such interactions in protein networks, provides interesting hints for the unknown functions of proteins (Schwikowski et al., 2000) and new insights into the mechanism of a biological process by offering detailed information for the binding partners within a complex.

There have been developed many high-throughput experimental methods that aim at detecting thousands of protein interactions per study. The most commonly used techniques are the yeast 2-hybrid (Y2H) (Uetz et al., 2000; Ito et al., 2001) and the tandem affinity precipitation (TAP) (Gavin et al., 2002) combined with mass spectrometry or tandem mass spectrometry (Shevchenko et al., 1996).

The detection of interactions using the yeast two-hybrid system is based on the reconstruction of transcription factors by exploiting the modular properties of site-specific transcriptional activators (Fields & Song, 1989). For example, hybrid proteins composed of a DNA binding domain fused with protein A and a transcriptional activation domain fused with protein B, are produced in yeast. If proteins A and B interact, it reconstitutes the transcription factor and leads to the expression of a reporter gene. The main drawback of this method is that it allows the detection of interactions in the nucleus of the cell only.

The TAP method combined with mass spectrometry, a powerful approach for the comprehensive analysis of protein–protein interactions, involves identifying the components of protein complexes. First, the complexes are isolated from cells using affinity-based methods, which demand the identification of at least one protein in the complex. Then, after this protein has been tagged with an affinity handle, it can be over-expressed in cells and affinity purified so that its interaction partners co-purify. The complex is then subjected to a typical proteomics analysis using mass spectrometry, so that individual proteins can be identified.

Computational methods have also been developed and used widely, in order to infer protein-protein interactions, not from physical proteins binding, but indirectly from properties that are related to interacting protein pairs. Some of the most known computational methods for discovering protein interactions are the domain fusion or Rosetta Stone method (Marcotte et al., 1999), the phylogenetic profiles (Pellegrini et al., 1999), the correlated expression of gene pairs (Grigoriev, 2001; Deng et al., 2003) and the gene neighbor method (OverBeek et al., 1999).

The domain fusion method is based on the observation that some pairs of interacting proteins have homologs in another organism that are fused into a single protein chain. For instance, the interacting proteins A and B in the fly genome might be found as a single longer protein C in the worm genome. If such proteins or protein domains unrelated in the fly, are fused together in worm, it suggests that they are likely to function or interact together in the fly. The fused protein C is called Rosetta Stone Sequence (Marcotte et al., 1999; Ng et al., 2003). Thus, this computational method entails searching through genomic sequences for two proteins, A and B, which in some other species are expressed as a fused protein, A-B.

Phylogenetic profiles encode patterns of presence or absence of genes across genomes, and are used to assign functional relationships to non-homologous pairs of proteins (Pellegrini et al., 1999). This method is based on the hypothesis that proteins which are functionally linked (i.e., participate in a common structural complex or biochemical pathway) evolve in a

correlated way and, thus, they have homologs in the same subset of organisms. In other words, it is very unlikely that two proteins would always be both present (or absent) to a new species unless they were functionally linked. Thus, if homologs to a pair of proteins are found in the same subset of organisms, the proteins are functionally linked.
The correlated gene expression methods (Grigoriev, 2001; Deng et al., 2003) detect protein-protein interactions based on the assumption that genes with correlated gene expression levels are more likely to encode interacting proteins.
Last but not least, the idea behind the gene neighbor method (Overbeek et al., 1999) is that if the genes encoding proteins A and B are neighbours on the chromosomes of several genomes, then A and B could participate in the same interaction or be involved in a similar function.

3. Data mining in proteomics

3.1 Classification

Classification is a data analysis task in which individual items are placed into groups based on one or more quantitative characteristics inherent in the items (also called "variables") and is based on a training set of previously labeled items. This means that classification is a supervised technique, since the prediction results fall in classes which are known beforehand. The algorithms used for classification are numerous. Here, we mention only the k-nearest-neighbour (k-NN) and the Support Vector Machines (SVMs), two algorithms that have been extensively used in proteomics.
The k-nearest-neighbour algorithm is amongst the simplest and mostly used algorithms for classification. In k-NN, an object is assigned to the class most common amongst its k nearest neighbours, where k is a typically small positive integer. If k = 1, then the object is simply assigned to the class of its nearest neighbor. In binary classification problems (i.e., two classes), it is helpful to choose k to be an odd number as this avoids tied votes. The neighbours are taken from a set of objects for which the correct classification is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. In order to identify neighbours, the objects are represented by position vectors in a multidimensional feature space. In k-NN several distance measures can be used, such as the Euclidean, Manhattan or Mahalanobis distance. In proteomics, k-NN has been used to classify mass spectra data (Taguchi et al., 2007), as well as jointly with a genetic algorithm approach, first introduced for microarray data, in order to discover biomarkers of chemical exposure and disease (Li et al., 2004).
The Support Vector Machines define a hyperplane (i.e., a linear decision boundary) that separates the classes under study. The hyperplane maximizes the distance (also called margin) between the two sample groups. By using the margin optimization only a small set of data points, called support vectors are critical for the separation, while the dimensions unnecessary for the separation of the classes are penalized. Thus, the problem of model overfit can be handled using SVMs. As a result, the SVM is a robust classification technique which is suitable for datasets with many features and a relatively small sample size, such as the microarrays and the proteomics datasets. Applications of SVM-based classifiers in proteomics include identifying significant differences between patients and healthy individuals from mass spectra (Willingale et al., 2006), as well as producing prediction models for the classification or annotation of biological function of novel protein sequences (Yang & Chou, 2004).

Other classification methods are the decision trees as well as the artificial neural networks. An extensive and more detailed description of many classification algorithms used in proteomics is provided in (Mavroudi et al., 2007).

3.2 Clustering

Clustering is the unsupervised (i.e., not available class labelling of the training patterns) classification of objects into different groups. Clustering techniques have found many applications in many fields, such as machine learning, pattern recognition, image analysis, bioinformatics and so on. In order to perform a clustering analysis task, one should follow the steps of (a) features selection, (b) choosing the appropriate proximity measures and clustering criteria, (c) running the clustering algorithm, (d) validating and interpreting the results. During the first step, features must be properly selected so as to encode as much information as possible regarding the dataset and the given problem, and possibly undergo preprocessing. In the next step, the measure which quantifies the similarity or dissimilarity of feature vectors is chosen and the clustering criteria of a specific algorithm are selected. Once the results of the clustering algorithm have been obtained, it is important to validate their correctness by comparing the scores of different results using validity indices (e.g., Silhouette, Dunn, Daves-Bouldin). The results interpretation is of great importance since an expert in the application field integrates the clustering results with other experimental evidence in order to draw the right conclusions.

According to the final representation of the results, data can be clustered either in a hierarchical or in a non-hierarchical way. Hierarchical clustering is a widely used data analysis method which tries to reveal the underlying structure of a dataset at many levels, based on the idea of building iteratively a tree by successively merging groups of data points, starting from the most similar ones. In other words, hierarchical clustering algorithms produce a hierarchy of nested clusters. They require as input a metric for measuring the distance between data points, and a linkage method, which defines how the distance between two already formed clusters is estimated. These methods are appropriate for the recovery of both elongated and compact clusters. Hierarchical clustering is very popular due to its simplicity and has been applied in various scientific fields, including functional genomics using DNA micro-arrays (Eisen et al., 1998; Heyer et al., 1999).

A very representative example of a non-hierarchical clustering algorithm is k-means. K-means is a greedy iterative algorithm, which assigns each data vector to the cluster the center of which (also called "centroid") is nearest to it. Unlike hierarchical clustering algorithms, k-means requires as input the number of clusters (k) to be formed. Furthermore, the user should specify a distance metric, a centroid initialization method and a stopping criterion. As a data partitioning algorithm, it has also found many applications in genomics (Tavazoie et al., 1999). The reasons for its popularity are its implementation simplicity, performance scalability, convergence speed and adaptability to sparse data.

Clustering algorithms have been widely used in functional genomics (Bolshakova et al., 2005) in order to group genes based on their relative expression levels in a sample. In proteomics, clustering methods have been recently introduced for LC-MS/MS spectral analysis (Beer et al., 2004), as it has been observed that they can contribute to peptide identification, comparison of peptide mixtures, prediction of retention time and so on. Moreover, as a recent review indicates (Hilario et al., 2006), for data extracted from 2D gels, clustering the peptide mass fingerprinting spectra can divulge the similarity of spots without even knowing their protein identity.

3.3 Association rules

The aim of association rules mining is to reveal underlying interactions in large sets of data items. This data mining method was initially used in "market basket analysis" for discovering regularities between products in large scale transaction data recorded in supermarkets. The output from this analysis consists of association rules, which describe groups of items that are frequently purchased together.
In general, an association rule is of the form of:

$$\{X\} => \{Y\} \quad (1)$$

where Y, represents the items that consist the Left Hand Side (LHS) of the equation, and X represents the items included in the Right Hand Side (RHS) of it. A rule states that whenever the LHS items are present in a transaction, the RHS items are likely to be present as well.
To evaluate the importance of an association rule, "interestingness measures" have been established and used. The *support* of rule (1) is the probability of a transaction in the dataset to contain both X and Y. In other words, support describes how frequently the rule occurs among transactions. The *confidence* of rule (1) shows its accuracy and is defined as the number of transactions with both the X and Y items, divided by the number of transactions with X items. Confidence shows the number of transactions in which the rule is correct, relative to the number of transactions in which it is applicable. An interesting rule must at least have support and confidence values greater than the user-specified minimum thresholds. *Leverage* is another measure which shows the percentage of additional cases covered by both the X and Y, above those expected if X and Y were independent of each other, and represents the unexpectedness of the rule. *Coverage* is the proportion of the transactions in the dataset which have the X items. Finally, the *lift* is a measure of the association's importance, which is independent of coverage, and is the confidence divided by the proportion of all transactions which have the Y items.
A number of approaches and methods have been proposed for association rules extraction, the main idea of which is based on the concept of frequent itemsets (i.e., sets of values in the same tuple). Some well known algorithms are Apriori (Agrawal & Srikant, 1994), Eclat (Zaki, 2000) and FP-Growth (Han et al., 2000).
Data mining based on association rules has also been applied in biomedical research. For instance, in the medical domain, association mining has been used to discover rules that relate patient symptoms, diagnosis and procedures performed on patients (Doddi et al., 2001), as well as to detect hidden and previously unknown patterns on large public health datasets, which can provide surveillance warnings (Giannopoulou et al., 2007), to name a few. Association mining has been also applied to the analysis of gene expression data, in order to reveal biologically relevant associations among different genes or between environmental effects and gene expression (Creighton & Hanash, 2003), as well as to proteomics, where it is important to discover rules that relate protein properties (e.g., functional annotation, sequence motifs) to protein-protein interactions (Kotlyar et al., 2006; Oyama et al., 2002).

4. Application examples

4.1 Sample classification from protein mass spectra

The study described in (Tibshirani et al., 2004) suggests a novel algorithm for pattern classification from protein mass spectra, which is a slight variation of the "nearest centroid"

classiffication. In particular, when applied to spectra from both diseased and healthy patients, the proposed "Peak Probability Contrast" (PPC) technique provides a list of all common peaks among the spectra, their statistical significance, and their relative importance in discriminating between the two groups. Compared to other statistical approaches for class prediction, this method performs as well or better than several methods that require the full spectra, rather than just labeled peaks. The algorithm consists of six sequential steps, shown in Figure 2.

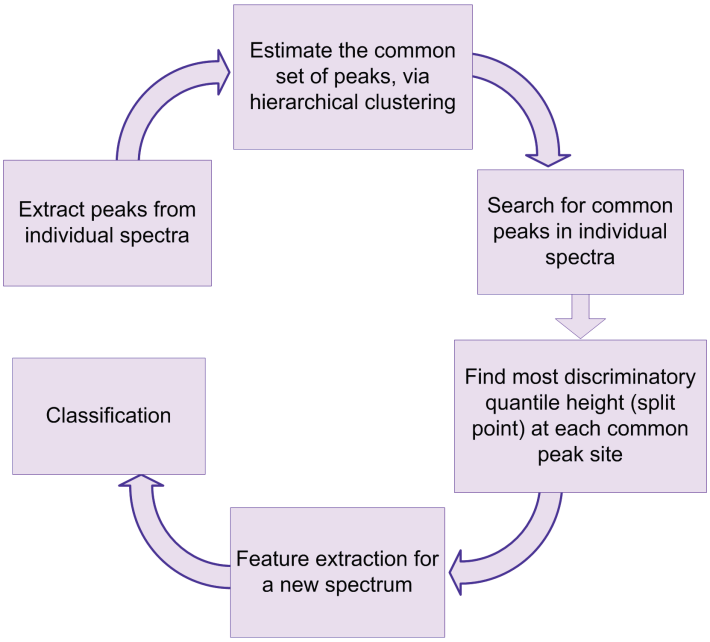

Fig. 2. Flow chart of the Peak Probability Contrast classification analysis. Figure adapted from (Tibshirani et al., 2004).

The step of peak extraction looks for mass-to-charge ratio (m/z) values the intensity of which is higher than that at the ±s m/z value surrounding it, and higher than the estimated average background at that m/z.

The next step estimates the common set of peaks using complete linkage hierarchical clustering. The clustering is one dimensional, using the distance along the log m/z axis. The idea is that tight clusters should represent the same biological peak that has been horizontally shifted in different spectra. Then the mean position (centroid) of each cluster is extracted, to represent the "consensus" position for that peak across all spectra.

Searching for common peaks in individual spectra is the step to follow. In particular, given the list of common peaks from clustering in previous step, the individual spectra are searched in order to record whether each spectrum exhibits each of these common peaks. A peak in the individual spectra is considered to be one of the common peaks if its center lies within a specific distance from the estimated center position of the common peak. If it is present, the height of the individual peak in the spectrum is also recorded.

From the previous steps, the spectrum peak heights are estimated for all observations and for all m/z values. As mentioned before, these heights are the centroids from the hierarchical clustering of all individual spectra peaks. If there is no peak at a specific m/z value, then the corresponding height is zero. During this step, the algorithm "cuts" each peak height at some quantile, in such a way so as to maximally discriminate between the healthy and normal samples in the training set.
The final step involves the class prediction of new mass spectra. A spectrum from a new patient has a binary feature vector of peak heights, with a component equal to one if the spectrum has a peak above the cutpoint height at that m/z value, and zero otherwise. Then, this binary profile is compared to each of the probability centroid vectors of the two classes (e.g., health vs. cancer) and classified to the class that is closest in overall squared distance (or some other metric).
The application of this method, so as to find a relative small number of peak clusters for class prediction, is expected to facilitate the identification of biologically significant and relevant proteins for specific biological states, such as tumor development and progression.

4.2 Clustering mass spectra peak-lists

In the study presented in (Ventoura et al., 2007) clustering algorithms are applied to proteomics data, in an attempt to group proteins based on their spectral similarities. Moreover, clustering validation methods are used to find the clustering method which most faithfully captures the underlying distribution of the samples. This work also shows that the application of clustering algorithms in proteomics can assist in (a) identifying peak features responsible for categorizing samples, (b) formulate hypotheses on the possible function and role of unidentified proteins and (c) reveal proteins which act jointly as biomarkers in a concrete biological state.
The proteomics data on which clustering is performed are the mass spectra peak-lists (not the raw mass spectra) which derive from a mass spectrometer. A mass spectrum peak-list is the intensity as a function of mass-to-charge ratio (m/z) profile of a sample (e.g., a protein spot) that has undergone mass spectrometry analysis. In order to apply cluster analysis, these peak-lists are represented as vectors in a multidimensional space, where each vector element is a feature of a specific mass (e.g., its intensity) or a group of masses. To deal with the high dimensionality of the generated peak-list vectors mass "bins" (i.e., contiguous non-overlapping regions in the m/z axis) can be defined before analyzing the samples of an experiment. The process of binning performs dimensionality reduction by grouping consecutive masses and selecting a representative feature of those masses for each group (e.g., mean, log, maximum intensity value). Moreover, one can preprocess the peak-lists vectors by performing scaling or normalization.
The suggested clustering algorithms for these data are the hierarchical as well as the k-means clustering. For a better comprehension of the clustering results several visualization methods are also exploited (i.e., dendrograms, heatmaps and cluster sets). In the clustering results that derive from this method, not only well separated protein clusters can be easily discerned, but also the spectral bins that are most influential in partitioning the proteins into clusters (Figure 3).
Furthermore, the presented method offers the option of integrating the identification results for the proteins – members of each cluster, as well as their Gene Ontology annotation. By exploiting both the identification and the Gene Ontology classification information for most

proteins in each cluster, one can attempt to infer the role of unidentified proteins. This can be based on the already known functions of the proteins which are identified with high confidence and are found to be close to unidentified proteins in the same cluster.

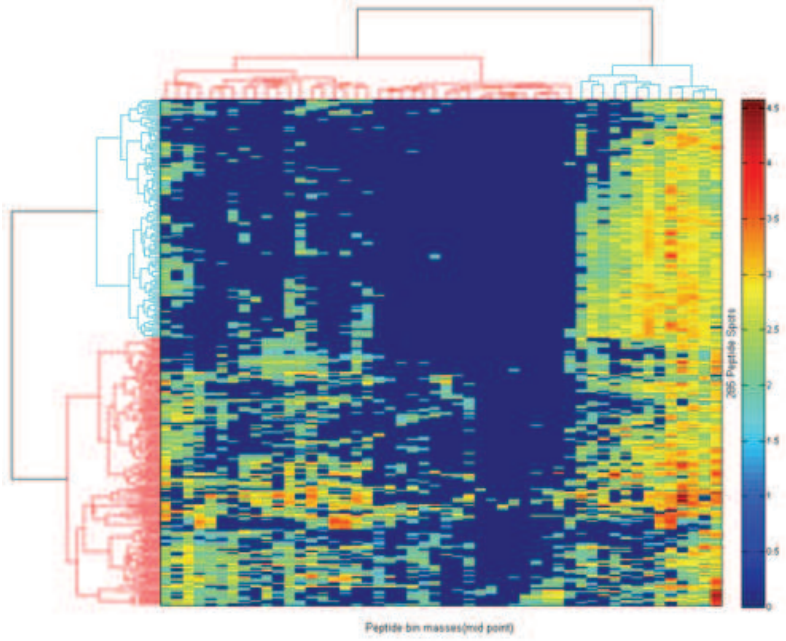

Fig. 3. Heatmap visualization of a hierarchical clustering result. Two well separated protein clusters (horizontal dendrogram) and two well separated bin clusters (vertical dendrogram) can be observed at the top-level. Figure adapted from (Ventoura et al., 2007).

4.3 Protein-protein interactions prediction using association rules

The work by (Kotlyar et al., 2006) is a very recent attempt that uses association rules not only to discover protein-protein interactions, but also to predict whether a given pair of proteins interacts. Predicting interactions with association mining can be viewed as a classification problem where the RHS part of the rule consists of a single item only, the class variable. After the application of association mining, the rules are ranked according to a measure of "interestingness" (e.g., confidence, support) and used for prediction as follows: a given protein pair is predicted to interact if its attributes include the LHS items of any rule. The presented approach is based on the idea that both direct and indirect evidence (e.g., data coming from experimental and computational methods) could be used to predict interactions reliably and on a proteome-wide scale. In particular, datasets that consist of interacting and non-interacting protein pairs annotated with different types of evidence are first constructed. Then, with the help of association rules, patterns that discriminate the interacting and the non-interacting proteins are detected. Lastly, using these patterns the prediction of interactions is achieved, assigning a confidence level to each interaction. The three steps followed in this approach to predict protein-protein interactions will be further described (Figure 4).

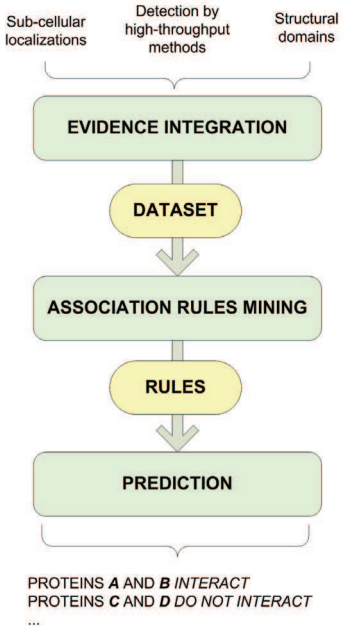

Fig. 4. Predicting protein-protein interactions using association rules. The reliability of interactions is increased if different types of evidence (direct and indirect) are jointly considered.

The first step in the suggested approach refers to the creation of datasets which integrate evidence from (a) high-throughput interaction detection methods, (b) gene expression microarrays and (c) protein annotation projects. This step seems to be of immense importance since previous studies have shown that by integrating data the reliability of protein-protein interactions can be improved (Goldberg & Roth, 2003; Zhang et al., 2004). Thus, a dataset is generated from protein pairs annotated with various attributes (e.g., detection by high-throughput methods, sub-cellular localizations, structural domains and so on). Since the aim of this work is to be able to decide if protein pairs interact (classification problem of 2 classes), the datasets deliberately include both protein pairs that represent true interactions, as well as randomly chosen non-interacting protein pairs.

During the association mining step, the Extended FP-Growth algorithm is chosen for creating association rules, due to its ability to succeed short run times in these large datasets. It is important to note that in the protein-protein interaction prediction problem, the vast majority of the rules are associated with non-interacting protein pairs that may not be very informative, and that significant rules may have very low support relative to the size of the dataset. These observations can be explained by the very low ratio of interacting to non-interacting protein pairs that has been observed in organisms (e.g., 1 interaction pair to 600 non-interacting pairs in the yeast).

After the rules are determined, they are ranked based on confidence. Rules having the same confidence are ranked by support. To predict if a given protein pair interacts, its attributes should match the LHS items of any rule. Then, the confidence of this prediction is the confidence of the highest ranked rule that is matched.

To conclude, with this approach, different types of evidence for interaction are integrated in order to create rules that act as a classifier for new interaction pairs. Thus, association mining is used to search thoroughly in large datasets for predictive patterns. However, to evaluate the performance of this method and strengthen its applicability, it is important to incorporate additional evidence, perform testing and validation using already known interactions from specific organisms and compare the results to those of other interaction detection methods.

5. Conclusion

Proteomics, the large-scale study and analysis of proteins, is a field of powerful techniques which offer significant experimental knowledge to experts in drug design and clinical applications (Fountoulakis & Kossida 2006; Simpson et al., 2008; Ge et al., 2008). Several studies and reviews available in the literature (Bachi & Bonaldi, 2008; Feng et al., 2008) also indicate that proteomics is of great value and significance to the analysis of complex biological model systems and to systems biology (i.e., the systems-level understanding of correlations among molecular components).

Using data-mining techniques in the large volumes of data obtained either by high-throughput differential expression proteomics analyses or by large-scale protein interaction experiments, serves as a powerful and promising mechanism for extracting useful knowledge and reaching interesting biological conclusions. This research area is rapidly growing and enriched with new applications which focus on detecting previously unknown protein functions and relations. However, the future directions should concentrate on developing novel methods and algorithms so as to improve the proteomics mining results in terms of validity, scientific soundness and verification.

6. References

Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. *Proceedings of the 20th International Conference on Very Large Data Bases,* pp. 487-499, ISBN 1-55860-153-8, Santiago Chile, September 1994, Morgan Kaufmann

Bachi, A. & Bonaldi, T. (2008). Quantitative proteomics as a new piece of the systems biology puzzle. *J Proteomics*. (July 2008)

Beer, I.; Barnea, E.; Ziv, T. & Admon, A. (2004). Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics*, Vol., 4, (February 2004) 950– 960

Bensmail, H.; Golek, J.; Moody, M. M.; Semmes J. O. & Haoudi A. (2005) A novel approach for clustering proteomics data using Bayesian fast Fourier transform. *Bioinformatics*, Vol., 21, (March 2005) 2210-2224

Bolshakova, N.; Azuaje, F. & Cunningham, P. (2005). An Integrated Tool for Microarray Data Clustering and Cluster Validity Assessment. *Bioinformatics*, Vol., 21, (February 2005) 451–455

Cannataro, M.; Guzzi, P. H.; Mazza, T.; Tradigo, G. & Veltri P. (2007). Using Ontologies for preprocessing and mining spectra on the Grid. *Future Generation Computer Systems*, Vol., 23, (January 2007) 66-60

Creighton, C. & Hanash, S. (2003). Mining gene expression databases for association rules. *Bioinformatics*, Vol., 19, (January 2003) 79–86

Deng, M.; Sun, F. & Chen, T. (2003). Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac Symp Biocomput*, 140–151

Doddi, S.; Marathe, A.; Ravi, S. S. & Torney, D. C. (2001). Discovery of association rules in medical data. *Med Inform Internet Med*, Vol., 26, (January/March 2001) 25–33.

Eisen, M. B.; Spellman, P. T.; Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, Vol., 65, (December 1998) 14863–14868

Feng, X.; Liu, X; Luo, Q. & Liu, B. F. (2008). Mass spectrometry in systems biology: An overview. *Mass Spectrom Rev.* (July 2008)

Fields, S. & Song, O. (1989). A novel genetic system to detect protein–protein interactions. *Nature*, Vol., 340, (July 1989) 245–246

Flory, M. R.; Griffin, T. J.; Martin, D. & Aebersold, R. (2002). Advances in quantitative proteomics using stable isotope tags. *Trends in Biotechnology*, Vol., 20, (December 2002) 23-29

Fountoulakis, M. & Kossida, S. (2006). Proteomics-driven progress in neurodegeneration research. *Electrophoresis*, Vol., 27, (April 2006) 1556-1573

Garbis, S.; Lubec, G. & Fountoulakis M. (2005). Limitations of current proteomics technologies. *Journal of Chromatography A*, Vol., 1077, (May 2005) 1 – 18

Gavin, A. C.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A. M.; Cruciat, C. M.; Remor, M.; Hofert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M. A.; Copley, R. R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster B.; Neubauer, G. & Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, Vol., 415, (January 2002) 141–147

Ge, X.; Wakim, B. & Sem, D. S. (2008). Chemical Proteomics-Based Drug Design: Target and Antitarget Fishing with a Catechol-Rhodanine Privileged Scaffold for NAD(P)(H) Binding Proteins. *J Med Chem.* (July 2008)

Giannopoulou, E. G.; Kemerlis, V. P.; Polemis, M.; Papaparaskevas, J.; Vatopoulos, A. C. & Vazirgiannis, M. (2007). A Large Scale Data Mining Approach to Antibiotic Resistance Surveillance. *Proceedings of 20th IEEE International Symposium on Computer Based Medical Systems*, pp. 439-444, ISBN 0-7695-2905-4, Maribor Slovenia, June 2007, IEEE Computer Society

Goldberg, D. S. & Roth, F. P. (2003). Assessing experimentally derived interactions in a smallworld. *Proc Natl Acad Sci*, Vol., 100, (April 2003) 4372–4376

Griffin, T. J.; Lock, C. M.; Li, X.; Patel, A.; Chervetsova, I.; Lee, H.; Wright, M. E.; Ranish, J. A.; Chen, S. S. & Aebersold, R. (2003). Abundance ratio-dependent proteomic analysis by mass spectrometry. *Analytical Chemistry*, Vol., 75, (January 2003) 867–874

Grigoriev, A. (2001). A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeas Saccharomyces cerevisiae. *Nucleic Acid Res*, Vol., 29, (July 2001) 3513-3519

Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H. & Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, Vol., 17, (October 1999) 994-999

Halligan, B. D.; Mirza, S. P.; Pellitteri-Hahn, M. C.; Olivier, M. & Greene, A. S. (2007). Visualizing Quantitative Proteomics Datasets using Treemaps. *Proceedings of 11th International Conference Information Visualization*, pp. 527-534, ISBN 0-7695-2900-3, Zurich Switzerland, July 2007, IEEE Computer Society

Han, J.; Pei, J. & Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 1–12, ISBN 1-58113-218-2, Dallas Texas, May 2000, ACM

Heyer, L. J.; Kruglyak, S. & Yooseph S. (1999). Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome Research*, Vol., 9, (November 1999) 1106–1115

Hilario, M.; Kalousis, A.; Pellegrini, C. & Muller, M. (2006). Processing and Classification of Protein Mass Spectra. *Mass Spectrometry Reviews*, Vol., 25, (February 2006) 409– 449

Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci*, Vol., 98, (March 2001) 4569–4574

Kotlyar, M. & Jurisica, I. (2006). Predicting protein-protein interactions by association mining. *Inf Syst Front*, Vol., 8, (February 2006) 37-47

Krishnamurthy, L.; Nadeau, J. H.; Ozsoyoglu, G.; Ozsoyoglu, Z. M.; Schaeffer, G.; Tasan, M. & Xu, W. (2003). Pathways Database System: An Integrated System for Biological Pathways. *Bioinformatics*, Vol., 19, (May 2003) 930-937

Li, L.; Umbach, D. M.; Terry, P. & Taylor, J. A. (2004). Application of the GA/KNN method to SELDI proteomics data. *Bioinformatics*, Vol., 20, (July 2004) 1638-40

Liebler, D. C. (2002). *Introduction to Proteomics*, Humana Press, ISBN 978-089603-991-9, Totowa New Jersey USA

Marcotte, E. M.; Pellegrini, M.; Ng, H. L.; Rice, D. W.; Yeates, T. O.; & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science,* Vol., 285, (July 1999) 751-753

Mavroudi, S.; Papadimitriou, S.; Kossida, S.; Likothanassis, S. D. & Vlahou, A. (2007). Computational Methods and Algorithms for Mass-Spectrometry Based Differential Proteomics. *Current Proteomics*, Vol., 4, (December 2007) 223-234

Monteoliva, L. & Albar, J. P. (2004). Differential proteomics: An overview of gel and non-gel based approaches. *Briefings in Functional Genomics and Proteomics*, Vol., 3, 2004, (November 2004) 220-239

Neverova, I. & Van Eyk, J. E. (2004). Role of chromatographic techniques in proteomic analysis. *Journal of Chromatography B*, Vol., 815, (December 2004) 51 – 63

Ng, S.K.; Zhang, Z.; Tan, S. H. & Lin, K. (2003). Interdom: A database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res.*, Vol., 31, (October 2002) 251–254

Overbeek, R.; Fonstein, M.; D'Souza, M.; Pusch, G. D. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci*, Vol., 96, (March 1999) 2896-2901

Oyama, T.; Kitano, K.; Satou, K. & Ito, T. (2002). Extraction of knowledge on protein-protein interaction by association rule discovery. *Bioinformatics,* Vol., 18, (May 2002) 705–714

Palzkill, T. (2002). *Proteomics,* Kluwer Academic Publishers, ISBN 0-792-37565-3, USA

Pellegrini, M.; Marcotte, E. M.; Thompson, M. J.; Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci*., Vol., 96, (April 1999) 4285-4288.

Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H. & Mann, M. (1996). Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc Natl Acad Sci,* Vol., 93, (December 1996) 14440–14445

Simpson, R. J.; Bernhard, O. K.; Greening, D. W. & Moritz, R. L. (2008). Proteomics-driven cancer biomarker discovery: looking to the future. *Curr Opin Chem Biol.,* Vol., 12, (February 2008) 72-77

Taguchi, F.; Solomon, B.; Gregorc, V.; Roder, H.; Gray, R.; Kasahara, K.; Nishio, M.; Brahmer, J.; Spreafico, A.; Ludovini, V.; Massion, P. P.; Dziadziuszko, R.; Schiller, J.; Grigorieva, J.; Tsypin, M.; Hunsucker, S. W.; Caprioli, R.; Duncan, M. W., Hirsch, F. R.; Bunn, P. A. & Carbone D. P. (2007). Mass spectrometry to classify non-small-cell lung cancer patients for clinical outcome after treatment with epidermal growth factor receptor tyrosine kinase inhibitors: a multicohort cross-institutional study. *J Natl Cancer Inst.*, Vol., 99, (June 2007) 838-846

Tavazoie, S.; Hughes, J. D.; Campbell, M. J.; Cho, R. J. & Church G. M. (1999). Systematic determination of genetic network architecture. *Nat Genet*, Vol., 22 (July 1999) 281–285

Tibshirani, R.; Hastiey, T.; Narasimhanz, B.; Soltys, S., Shi, G.; Koong, A. & Le, Q. (2004). Sample classification from protein mass spectrometry, by "peak probability contrasts". *Bioinformatics*, Vol., 22, (November 2004) 3034-3044

Tyers, M. & Mann, M. (2003). From genomics to proteomics. *Nature,* Vol., 422, (March 2003) 193-197

Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T. A.; Judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S. & Rothberg, J. M. (2000). A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*, Vol., 403, (February 2000) 623–627

Ventoura, S.; Giannopoulou, E. G. and Manolakos, E. S. (2008). ProtCV: A Tool for Extracting, Visualizing and Validating Protein Clusters Using Mass Spectra Peak-Lists. *Proceedings of 21st IEEE International Symposium on Computer Based Medical Systems*, pp. 221-223, ISBN 978-0-7695-3165-6, Jyvaskyla Finland, June 2008, IEEE Computer Society

Willingale, R.; Jones, D. J.; Lamb, J. H.; Quinn, P.; Farmer, P. B. & Ng, L. L. (2006). Searching for biomarkers of heart failure in the mass spectra of blood plasma. *Proteomics*, Vol., 6, (November 2006) 5903-14

Yang, Z. R. & Chou, K, C. (2004). Bio-support vector machines for computational proteomics. *Bioinformatics*. Vol., 20, (March 2004) 735-741

Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, Vol., 12, (May/June 2000), 372-390

Zieske, L. R. (2006). A perspective on the use of iTRAQ reagent technology for protein complex and profiling studies. *Journal of Experimental Botany*, Vol., 57, (March 2006) 1501-1508

Zhang, L.V.; Wong, S. L.; King, O. D. & Roth, F. P. (2004). Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, Vol., 5, (April 2004) 38

Data Mining in Medical and Biological Research

Edited by Eugenia G. Giannopoulou

ISBN 978-953-7619-30-5
Hard cover, 320 pages
Publisher InTech
Published online 01, November, 2008
Published in print edition November, 2008

This book intends to bring together the most recent advances and applications of data mining research in the promising areas of medicine and biology from around the world. It consists of seventeen chapters, twelve related to medical research and five focused on the biological domain, which describe interesting applications, motivating progress and worthwhile results. We hope that the readers will benefit from this book and consider it as an excellent way to keep pace with the vast and diverse advances of new research efforts.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Eugenia G. Giannopoulou and Sophia Kossida (2008). Data Mining Applications in the Post-Genomic Era, Data Mining in Medical and Biological Research, Eugenia G. Giannopoulou (Ed.), ISBN: 978-953-7619-30-5, InTech, Available from: http://www.intechopen.com/books/data_mining_in_medical_and_biological_research/data_mining_applications_in_the_post-genomic_era

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821