Chapter 1

Introduction

1.1 Motivation and Goals

The increasing availability and decreasing cost of high-throughput (HT) technologies coupled with the availability of computational tools and data form a basis for a shift to a more integrated approach in analyzing biological processes. Classical statistical analysis techniques were designed to analyze data characterized by a single data source distinguished by a much higher ratio of subjects in comparison to the number of observations arising from each subject. In contrast, bioinformatics and systems biology often involve high-throughput data characterized by an abundance of observations spawned from a relatively small sample of subjects. Additionally, the complexity of these systems under analysis coupled with the need to integrate inconsistent (noisy) data often violates many of the assumptions of classical analytic techniques. My primary focus has been based upon the identification of relationships amongst noisy, inconsistent data within the context of providing a more integrated approach to analyzing biological processes. The approaches I developed identify subsets of data that maintain robust analytic relationships obscured by the standard methodologies.

My initial work was within the field of metabolomics and focused on providing a digital data representation through automated alignment of Nuclear Magnetic Resonance (NMR) Spectra. The longer-term goal of this work was to open up new avenues for analysis and

integration of metabolomic data and aid their incorporation into larger integrative analysis frameworks. Specifically, the transformation of the spectrum representation from points to peaks which reduces the inconsistency within a spectrum by focusing directly on the component of analysis. The algorithm reduces each spectrum from thousands of points to hundreds of consistent peaks for final analysis. Moreover, the automated alignment of the NMR spectra served as a means of further noise reduction, increasing the likelihood that the peaks within each spectrum would be fruitful with regards to the final result. The noise reduction provided by this transformation and alignment process greatly simplified data complexity and enabled further application of other means of statistical analysis of NMR spectra.

From this, my focus shifted to developing data mining methods to identify relationships that exist within subsets of inconsistent, imperfect data. My research deliberately focused upon data where traditional means of analysis proved to be futile, to identify association between response and explanatory variables as data sources are integrated over a common set of subjects. Specifically, the toxicological associations between animal study endpoints (response variables) and high-throughput/high-content bioassays (explanatory variables) as perturbed by the same potentially toxic chemicals (subjects). The methods I employed use pattern identification approaches to identify subsets of potentially toxic chemicals that perturbed sets of animal endpoints and bioassays in a consistent manner. These methods have been enhanced to allow for the incorporation of user-defined amounts of fuzziness into the results and to enable the identification of statistically significant results based upon user defined metrics (no p-value required). Furthermore, the methods can be employed upon larger, more dense datasets through targeted analysis and can be used in the integration of three or more datasets.

1.2 Brief Overview of Existing Methods

1.2.1 NMR Spectra Noise Reduction Methods

As discussed in detail in Chapter 2 within the field of metabolomics, the standard way to reduce the noise in spectra prior to analysis is through binning, a procedure that involves dividing the spectra into small windows and taking the area under the curve for each window as the final intensity (Gartland et al., 1991; Anthony et al., 1994). Ideally, these windows will be large enough to encompass peak drift and to reduce the number of points that represent a spectrum, but not so large as to include many peaks in a single bin. The latter consequence is unavoidable in crowded spectra and thus there is the potential for significant loss of information when binning, for example by including peaks belonging to multiple compounds within a single bin. Alternatives to binning typically involve some form of peak alignment procedure. Several algorithms have also been recently developed to align peaks in sets of NMR spectra Wu et al. (2006); Kim et al. (2006); Torgrip et al. (2003); Veselkov et al. (2009); Savorani et al. (2010).

Current advanced NMR alignment methods such as fuzzy warping (Wu et al., 2006), Bayesian alignment (Kim et al., 2006), Recursive Segment-Wise Peak Alignment (Veselkov et al., 2009), peak alignment by FFT (Wong et al., 2005; Savorani et al., 2010) and peak alignment using reduced set mapping without recursive target update (Torgrip et al., 2003), are based on the use of a template spectrum to help align a set of spectra. Choosing a template typically involves either selecting a single sample spectrum that appears most like the others as determined by some measure of similarity, creating an ”average” spectrum, or by choosing a reference spectrum not contained within the sample. All remaining sample spectra are then aligned to this selected template using some form of pairwise alignment algorithm. A significant problem with the template approach is that there can be a great amount of variability between any two spectra. Part of this difference arises due to the previously described chemical shift variation. In addition, significant differences arise due to the existence of disparate

groups within the data; for instance, inter-group variation between control and treated groups, subpopulation differences within these groups, etc. There may often be a priori knowledge of general subgroups, but one of the goals of metabolomics is to discover new subgroups such as different types of responders in drug or toxicity studies; by definition, templates for such groups are not known beforehand. Thus in such cases, the use of a template can significantly complicate downstream analyses. Further discussion of existing methodologies and comparison of these methodologies to our own can be found in Chapter 2.

1.2.2 Identifying Association in Inconsistent, Noisy Data

Clustering is a fundamental method of unsupervised learning that partitions data in a way as to highlight meaningful relationships by exploring how data groups based upon similarity. Given a two-dimension data matrix, 2-D hierarchical clustering can be used to consider both columns and rows of the data when looking for meaningful relationships within the data. 2-D hierarchical clustering is not ideal in inconsistent, noisy data because the methodology considers the entire record (all the data in a given row and for a given column) when partitioning the data into meaningful groups. Similarly to 2-D hierarchical clustering, biclustering is able concurrently partition data by both rows and columns. Unlike 2-D hierarchical clustering, biclustering is able to consider submatrices, or subsets of the data; thus, using biclustering is a better method than hierarchical clustering to identify meaningful relationships within inconsistent data. Computationally, biclustering works best on sparse data matrices or when heuristics are used to limit the exhaustive enumeration of all possible submatrices. This is because biclustering solutions employ algorithms with computational complexity of NP-complete, meaning they have no known polynomial time algorithms and in the worst case their runtimes are exponential. van Uitert et al. (2008) demonstrate the use of biclustering on high-throughput data when they employ their method of biclustering on sparse binary genomic data to identify interacting transcription factors. Another example is DiMaggio et al. (2010) use of biclustering on inconsistent data to explore the use of logistic regression to identify predictive association

between sets of explanatory variables and a response variables. Methods of biclustering most directly compare to our methodology because they focus upon analysis of subsets of the data.

Other methods of determining association across multiple datasets with inconsistent data typically involve a bayesian framework. Specifically, these methods tend to weight the data based on its usefulness in the underlying mathematical model of association as was demonstrated by Webb-Robertson et al. (2009) using metabolomic data. The primary motivation of the study by DiMaggio et al. (2010) was to identify relationships between explanatory and response variables that could be used in prediction; whereas, the motivation of the Webb-Robertson et al. (2009) methodology was to identify relationships that provided the most significant differences between classes based upon integrated metabolomic data. Zhang et al. has developed data mining methods to identify significant relationships that existed between sets of explanatory and response variables for categorical data (Zhang et al., 2010b,a). Similarly, van Uitert et al. (2008) developed a method that was used to determine an association between two sets of genomic data to identify clusters with novel associations between the datasets. Although not focused on the relationship between explanatory and response variables, Reif et al. (2010) developed a measure that integrates multiple sources of toxicological data together to prioritize toxicological risk. Unlike DiMaggio et al., the methodology of Zhang et al. is able to integrate together the search for relationships with significance testing of discovered relationships. DiMaggio and Webb-Robertson both use methods that are more suited for integrating data from multiple data sources where a high degree of inconsistency (noise) exists between the data sources. Additionally DiMaggio, Webb-Robertson, and Reif's methodologies are more suitable for handling numeric data as compared to the methods that Zhang et al. employ which involve pairwise association between categorical data. The methodology of van Uitert et al. addresses some degree of inconsistency within the data, but unlike the other methods, its primary goal is the discovery of novel associations identified through integration with little regard for finding all associations or assigning statistical significance to the results. Similarly to van Uitert et al., Reif's methodology does not provide statistical significance to

indicate the importance of its results. However, their methodology does provide a ranking of toxicological risk based upon multiple data sources. As discussed above there are multiple methods of integrating inconsistent data, but the biclustering methodology (like van Uitert et al. (2008)) is most similar to our methods because they both focus upon analysis of subsets of data to deal with inconsistency.

1.3 Approach and Innovations

1.3.1 Methods to Enhance NMR Spectra Analysis

Our novel approach for the alignment of NMR spectra is based on the creation of a consensus spectrum alignment through integration of pairwise spectrum comparisons (referred to as PCANS hereafter - Progressive Consensus Alignment of Nmr Spectra). To our knowledge, this is the first such consensus approach applied to the alignment of NMR spectra and the only approach that transforms spectra from points to peaks prior to alignment as opposed to using the entire spectrum. This approach has several advantages that include the ability to align spectra with significant amounts of noise in chemical shift position, peak height and peak width. By using peaks as the basis for alignment we maintain the maximally informative set of information existing within a set of spectra. As a result, the existence of subgroups within a set of spectra can be identified since group-specific peaks are maintained in the final alignment.

We characterize the performance of this approach by aligning simulated NMR spectra which have been provided with user-defined amounts of chemical shift variation as well as inter-group differences as would be observed in control-treatment applications. Moreover, we demonstrate how our method provides better performance than either a template-based alignment or binning. Finally, we further evaluate this approach in the alignment of real mouse urine spectra and demonstrate its ability to improve downstream statistical analyses such as PCA and OPLS models commonly used in metabolomics analyses.

1.3.2 Methods to Enhance Association Identification

The data mining methods implemented focus on data where traditional methods of predictive modeling failed to identify useful relationships because they considered the entire data record. In contrast our approach, similar to biclustering, identifies relationships amongst subsets of the data. Our methods differ from the biclustering and prediction scheme of DiMaggio et al. (2010) by allowing one to incorporate group identification and association in a more streamlined framework. Moreover, our methods exhaustively explore the inclusion of multiple response variables with regards to association with the explanatory variables, while the work of DiMaggio et al. considers each response variable separately. Our methods more fully explore all possible enumerations of the subsets of data that specifically support the desired association; in our case the association between response and explanatory variables. Our methods are more similar to those employed by Zhang et al. (2010b,a) with regards to incorporating association finding and significance into a streamlined framework. However, unlike Zhang, our methods focus on subsets of the data (Zhang et al., 2010b,a). While our algorithms are similar to the methodology of van Uitert et al. (2008) as in they are applied to sparse inconsistent binary data; they differ from this work in that they provide a measure of statistical significance for the results. Additionally they provide the full complement of results for a given threshold, and can be modified to integrate more than a pair of datasets. Our methods differ from all three (Zhang, DiMaggio, Webb-Robertson) by allowing one to incorporate fuzziness (allowable zeros) given specific restrictions (described later). Finally, our algorithm is able to be applied to the mining of larger datasets by constraining the search space through requiring a minimum number of pre-specified features in the output through the use of seed nodes.

1.3.3 Thesis

Classical statistical analyses are not robust in identifying relationships within data in the presence of inconsistencies or noise. By focusing analysis on subsets of data with internal

consistency, I develop methods that show improved identification of relationships as evidenced by the relevance of the generated results.

The methods I develop focus on two areas of research, NMR spectra analysis and data mining for association within inconsistent data. The contributions to improving NMR spectral analysis are discussed first. The data mining for association follows because these methods can be directly applied to NMR spectral analysis to improve the relevance of the results.

1.3.4 Contributions to Enhance NMR Spectra Analysis

To address these problems of inconsistency between NMR spectra when performing metabolomic type analysis, our methods transform and align the peaks of each spectrum. This reduces the analysis to a small subset of well-aligned aligned peaks as opposed to attempting to quantify and analyze all the unaligned points of each spectrum. Treating each spectrum as hundreds of aligned peaks as opposed to thousands of unaligned points enhances the analysis we can perform and enables us to use more sophisticated means of analysis as is discussed in detail in Chapter 2.

Innovations made with regards to NMR spectra analysis are the following:

- Spectra are transformed (subset) to a collection of peaks with properties of location, height and width instead of a collection of points
  - Reduces spectrum to relevant information
  - Reduces complexity of alignment and analysis
  - Reduction allows for more sophisticated analysis
- Alignment algorithm that employs consensus as opposed to template alignment
  - Improves quality of alignment by preventing misalignment of peaks not found within the template
  - Consensus alignment can be incorporated into any pairwise alignment scheme

  - Removes need to identify all peaks within sample spectra for template formation
  - Same amount of computation as template alignment when coupled with pairwise alignment schemes

1.3.5 Contributions to Enhance Association Identification

With the integration of datasets over a common set of observations, inconsistencies are addressed by identifying subsets of data that most strongly support the desired association between the datasets. For this problem in particular, the methods developed focus on deficiencies in current methodologies by identifying consistent relationships within noisy data. Once these subsets are identified, the methodology establishes a statistical framework under which the significance of the subsets can be ranked and their strength of association can be determined. This approach of exploring relationships within subsets of the data is meant to be used when traditional means of analysis fail to produce adequate results due to inconsistency within the data. Furthermore, this methodology is meant to be used as an exploratory tool to find underlying relationships that were obscured with traditional means of analysis.

Innovations made with regards to the discovery and prioritization of subsets:

- Determination of Subsets with Closed/Approximate Itemset Mining
  - Means to target analysis on certain relationships with use of *seed nodes*
    - Full enumeration of desired relationships based upon frequency criterion
    - Exploration of larger, more dense data through targeted analysis
    - Ability to focus analysis on multivariate associations (2+ response variables)
  - Incorporation of Fuzziness into subsets
    - For larger, more dense datasets
    - Use of statistic to provide relevance of results
- Establish Metric of Importance

- Strength of association and statistical relevance
  - Phi Coefficient used to rank and provide statistical relevance for closed / approximate itemsets applied to identify association between explanatory and response variables
  - Established techniques to enable use of bootstrap methodology on larger datasets with higher support thresholds to facilitate use of *any* metric to quantify association
- Integration of 3+ Datasets with bootstrap methodology's use of multiple metrics

1.4 Dissertation Outline

In chapter 2, I describe my initial work on PCANS in the field of metabolomics in detail. Beginning with background and motivation, describing the methodology, results on simulated and real data, our conclusions and future directions. Chapters 3 through 5 focus on my work developing data mining techniques to mine for association with inconsistent, noisy datasets. Chapter 3 describes in detail the background and related work. Chapter 4 describes using closed/approximate itemset mining in conjunction with the phi coefficient to discover significant subsets within data. Chapter 5 describes in detail mining for association with a bootstrap methodology where statistical significance is no longer dependent upon a metric with an associated p-value. Chapter 6 presents the conclusions and future directions of my thesis research.