Analysis of Immunosignaturing Case Studies

by

Muskan Kukreja

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved August 2012 by the
Graduate Supervisory Committee:

Stephen Albert Johnston, Chair
Valentin Dinu
Phillip Stafford

ARIZONA STATE UNIVERSITY

December 2012

ABSTRACT

Immunosignaturing is a technology that allows the humoral immune response to be observed through the binding of antibodies to random sequence peptides. The immunosignaturing microarray is based on complex mixtures of antibodies binding to arrays of random sequence peptides in a multiplexed fashion. There are computational and statistical challenges to the analysis of immunosignaturing data. The overall aim of my dissertation is to develop novel computational and statistical methods for immunosignaturing data to access its potential for diagnostics and drug discovery.

Firstly, I discovered that a classification algorithm Naive Bayes which leverages the biological independence of the probes on our array in such a way as to gather more information outperforms other classification algorithms due to speed and accuracy.

Secondly, using this classifier, I then tested the specificity and sensitivity of immunosignaturing platform for its ability to resolve four different diseases (pancreatic cancer, pancreatitis, type 2 diabetes and panIN) that target the same organ (pancreas). These diseases were separated with >90% specificity from controls and from each other.

Thirdly, I observed that the immunosignature of type 2 diabetes and cardio vascular complications are unique, consistent, and reproducible and can be separated by 100% accuracy from controls. But when these two complications arise in the same person, the resultant immunosignature is quite different in that of individuals with only one disease.

I developed a method to trace back from informative random peptides in disease signatures to the potential antigen(s). Hence, I built a decipher system to trace random peptides in type 1 diabetes immunosignature to known antigens. Immunosignaturing, unlike the ELISA, has the ability to not only detect the presence of response but also absence of response during a disease. I observed, not only higher but also lower peptides intensities can be mapped to antigens in type 1 diabetes.

To study immunosignaturing potential for population diagnostics, I studied effect of age, gender and geographical location on immunosignaturing data. For its potential to be a health monitoring technology, I proposed a single metric Coefficient of Variation that has shown potential to change significantly when a person enters a disease state.

ACKNOWLEDGMENTS

I am highly grateful to Stephen Albert Johnston without which none of my projects would have been possible. Also, I am highly thankful to Phillip Stafford, who guided me in all my projects about informatics and statistics analysis. I would also like to acknowledge my committee member Dr. Valentin Dinu who provided invaluable suggestions regarding computational data analysis for each project. I really appreciate the time and devotion all these people gave me. I also want to thank Dr. Lucus Restrepo, Dr Zibinew, Dr Bart Legutki, Dr Rebecca Halperin for providing there valuable immunosignaturing datasets and useful discussions without which my computational analysis would not have been possible. In addition, I feel highly privileged to work with the immunosignaturing core in the Center for Innovations in Medicine (CIM) of the Biodesign Institute at ASU. I am also thankful for all the administrative support I obtained in both CIM and Biological Design Graduate Program: Pattie Madjidi, Penny Gwynne, Preston Hunter, Kevin Brown, Maria Halin, Lauren Dimpsey, Dr Joann Williams. Last but not least, I would like to thank my colleagues that have help during my course of PhD by providing help computationally and doing wet lab experiments: Sakshi Arora, Pradeep Kanwar, Sushovan De, Dr HoJoon Lee, Kurt Whittemore, Krupa Navalkar, Dr Nidhi Gupta, Dr Chris Delhnet and lastly to family members Bhumi Sabarwal, Sakshi Arora, Sourabh Kukreja for their invaluable love and support.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

Out of the many economic problems that we are facing today, health care expenditure is clearly the one that draws the most attention. When compared to total US energy spending, total US health care spending is clearly higher and brings more impact on the both social and financial life of the people. Figure 1-1 shows the total U.S energy spending (Administration 2009) versus total U.S. health care spending (US Department of Health and Human Services 2010)

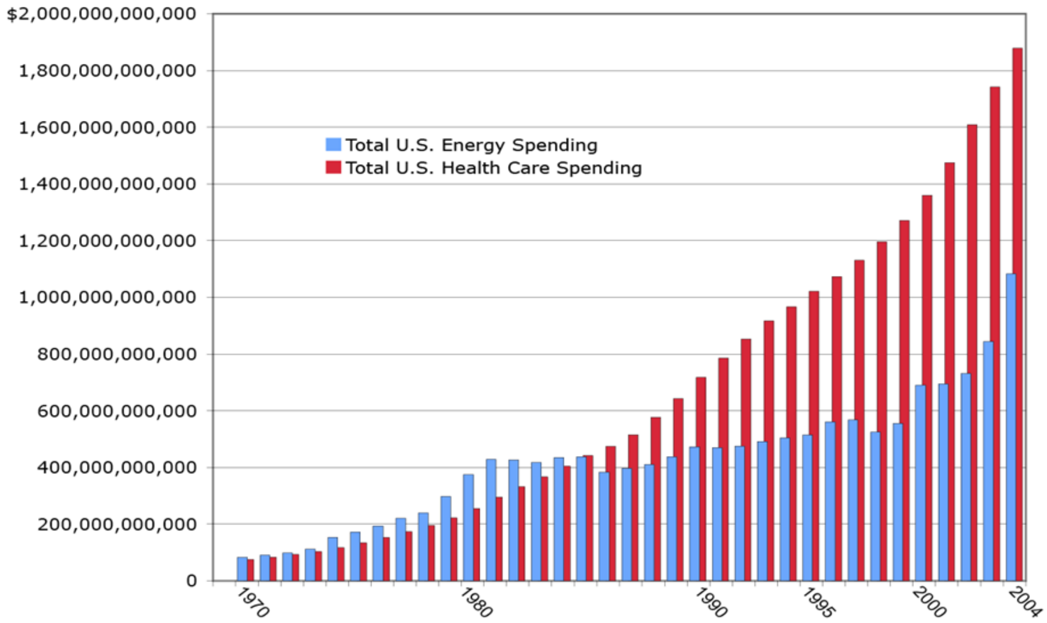


Figure 1-1 U.S Energy Vs. Healthcare Spending from 1970-2004

With more than 2.5 trillion dollars spent this financial year, the US health care expenditures continue to rise at a faster rate. With the current rate of increase in the expenditure, the total U.S spending would reach to 4 trillion dollars in 2015 (US Department of Health and Human Services 2010). **Figure 1-2** shows the total U.S Healthcare expenditure projections.

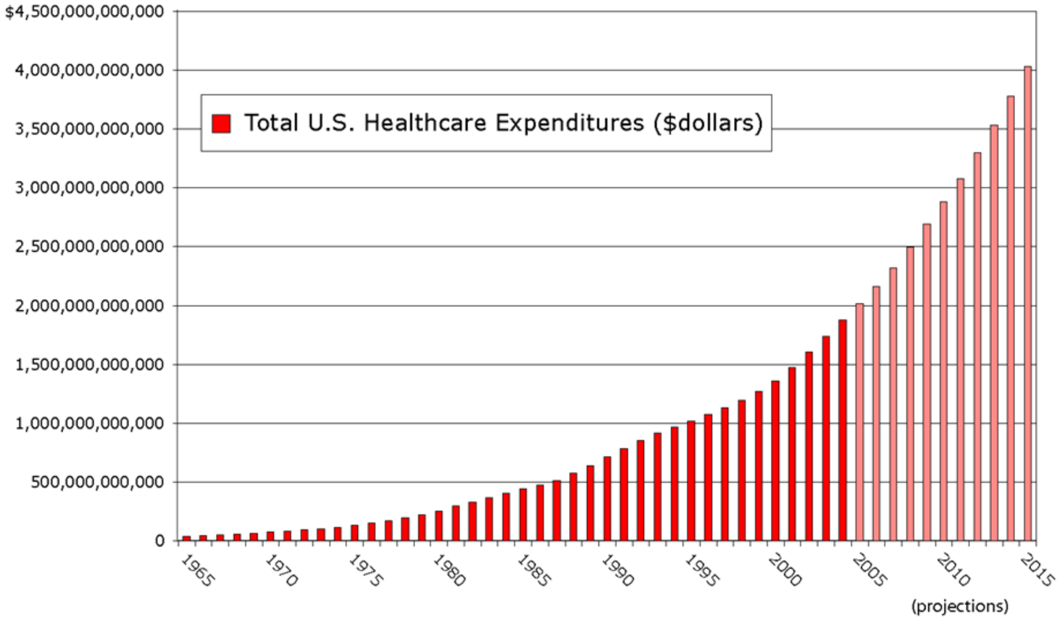


Figure 1-2 Total U.S. Healthcare expenditures projections ($dollars)

The current scenario of post symptomatic medicine is based on treating a subject with a disease after diagnosis. With more than 2.5 trillion dollar expenditure, 88% of the cost of post symptomatic medicine goes into patient care ie taking care of sick people. 10% of the total expenditure goes into drug development and only 2% goes into diagnostics (US Department of Health and Human Services 2010). The primary reason for the expensive health care is often late diagnosis which means current drugs have to encounter a biological system which is already out of order. A paradigm shift is required to revolutionize health care which should be based on pre-symptomatic diagnosis rather than the current scenario of post symptomatic diagnosis. Towards the goal of an effective diagnostic, biological molecules are continuously being sought for their potential to be biomarkers for diseases.

1.1 BIOMARKERS

In theory, a given biomarker molecule can serve as a proxy for detecting and diagnosing disease and hence could be the most effective means of measuring drug efficacy and improving patient health (Weston and Hood 2004). One of the most ubiquitous technologies used for biomarker identification is mass spectrometry (Li et al. 2005; L. Ackermann, E. Hale, and L. Duffin 2006). It has been widely used to search for diagnostic markers, and the high sensitivity has made it useful for identifying informative biomarker molecules that associate with disease. This process of reducing biomarkers down to a single or few best candidates occasionally leads to overtraining, where highly precise biomarkers that work well in small cohorts become harder to correlate with large and diverse test populations (Kiehntopf, Siegmund, and Deufel 2007). It is becoming increasingly apparent that utilizing higher numbers of biomarkers simultaneously can relieve some of this ‘low-feature-number’ classification problem. Unfortunately, some attempts at using mass spectrometry to identify disease-associated mass spectrogram signatures have lead to skepticism about this concept (Chapman 2002). One of the major drawbacks of serum based biomarkers is dilution. The ability to detect small concentration has been tested and numerous issues with reproducibility and sensitivity have arisen (Barbosa et al. 2005; Diamandis and van der Merwe 2005). Were there a candidate biomarker that was abundant, unaffected by age, sex, race, or genetic factors, different between healthy and sick persons and physically stable, the problem would become simpler. One such candidate is the immunoglobulin molecule.

1.2 ANTIBODIES AS BIOMARKERS

The immune system periodically monitors and performs surveillance against any foreign antigen or abnormal self activity. Although, invasive antigen concentration might be low during the early course of infection or disease, a B-cell has the potential to effectively recognize the antigen by producing antibodies and rapidly clonally expanding leading to amplified signal. At a given time point, an individual's antibody repertoire consists of nearly 10^{10} different antibodies. The antibodies which are near to germline antibody clones are low affinity while with matured clones are of high affinity (Dotan et al. 2006). Taken together, this repertoire effectively determines an individual's immunological status. This status is an indicative measure of individual immunization history or exposure to inflammation, infection or a chronic disease. Being exposed to any of the above changes, ones immunological status can be reflected by the rich information content present by different antibody profiles. Such rich information if deciphered could potentially be used for disease diagnostics even pre-symptomatically, vaccine evaluation and drug discovery.

Preventive medicine relies on pre-symptomatic, high specificity detection of a disease. Towards this goal, there has been a spur to develop novel and efficient biomarkers of diseases that can be captured at an early onset for effective diagnosis and treatment. Although more than 100,000 biomarkers have been reported currently in the literature (Kurian et al. 2009) but only 43 are FDA approved (Amur et al. 2008). Clearly, novel technologies are required to improve current early biomarker discovery methods. But there is a significant challenge involved in identifying biomarkers at early stages. One of the challenges involves the blood dilution problem. At an early stage of disease, concentrations of potential biomarker molecules are low which poses a clear identification problem for technologies. To overcome this limitation, one of the best ways out is to amplify the response of the biomarker. An individual immunological status at any given time point can be represented by its antibody profile. Also, antibodies can solve blood dilution problem. Abnormalities in immune system at early stage, activates B cells which can produce 5,000 to 20,000 antibodies/min (Cenci and Sitia 2007). Moreover the cell itself replicates every 70 hour (Cooperman et al. 2004) with lifespan up to 100 days (Forster and Rajewsky 1990) which leads to ~ 10^{11} amplification of signals from an antigen in a week. Antibodies are very stable allowing ease of sample processing, and convenience of using historically archived serum samples (Geijersstam et al. 1998) and also through saliva at extreme temperatures (Chase, Johnston, and Legutki 2012)

Towards the goal of using antibodies as effective biomarkers, the primary concern is to know if antibodies are generated in diseases other than general inflammation, infection or vaccination particularly at an early stage. Literature review clearly suggests that the humoral immunity actively participates specifically and early in autoimmunity like T1D (Bonifacio et al. 2000; Imagawa et al. 2000) and arthritis (Schellekens et al. 2000; Thurlings et al. 2006). In T1D autoantibodies against Insulin, IA-2 and GAD-65 are used as promising clinical biomarkers for early diagnostics. Taken together, the presence of three autoantibodies can predict up to 75% risk of having T1D (Hawa et al. 2000).

Antibody response are also being observed in cardiovascular disease, Alzheimer's disease and even different types of cancer (Ada and Jones 1986; Brichory et al. 2001; Brydak and Machala 2000; Cox et al. 1994; DiFronzo et al. 2002; Hooks et al. 1979; Lennon, Lindstrom, and Seybold 1976; Sreekumar et al. 2004; Stockert et al. 1998; Wilder 1995; Restrepo et al. 2011) . Also, these antibodies have been observed to be present at long before the disease symptoms start to show up.

1.3 ANTIBODY PROFILING TECHNOLOGIES

Use of antibodies as biomarker shows both potential and promise to identify disease at an early stage. Given ~10^{10} antibodies that an individual has, clearly there is a requirement for a high throughput technology which could profile such a huge amount of information. Modern technologies are being developed, focusing on maximizing targets that antibodies can recognize with significant differential affinity. Current technologies have to maintain a balance in performance among various key factors (Bacarese-Hamilton, Gray, and Crisanti 2003; Anderson and LaBaer 2005). The summary of some existing technologies to profile antibodies are given below.

1.3.1 PROTEIN BASED ASSAYS

These assays are primarily focused to span maximum coverage of antigen tested. In these assays, protein are either directly spotted on the surface or prepared in-vitro on the surface in a form of protein microarray. Whole protein arrays, on one hand have the capability to detect antibodies raised against conformational epitopes but rely on appropriate folding mechanism for spotted protein.

1. **Spotted Protein Arrays**: These assays consist of recombinant protein library selected as candidates by dissecting entire proteome via sequencing a pathogen of interest or selected immunogenic proteins from large proteome pathogens (Bacarese-Hamilton, Gray, and Crisanti 2003; Fang, Frutos, and Lahiri 2002; Mattoon et al. 2005). Complete human proteomes could also be spotted for autoantibody screening against autoimmunity (Robinson et al. 2002). The limitations of these technologies are they require prior knowledge of antigens and inefficient to probe membrane proteins.
2. **Nucleic Acid Programmable Protein Arrays (NAPPA):** It is a promising technology which spots DNA encoding protein as probes while transcription-translation happens dynamically on the surface. One of the advantages of such technology is the formation of protein dynamically on the surface which preserves the structure of the protein and also enables membrane protein to be spotted more efficiently than traditional assays (Anderson and LaBaer 2005; Ramachandran et al. 2008). The main limitation is lack of efficient discovery since it requires prior knowledge of antigens tested. It is technically demanding and has not become generally useful.
3. **Cell Lysate Arrays**: This type protein array is a high throughput method to discover protein abundance or its modification state. These have wide dynamic range and multiplexed detection to quantify and compare multiple targets (Sheehan et al. 2005; Hall, Ptacek, and Snyder 2007). It consists of

complex samples such as tissue lysates that are spotted on the array and are exposed to antibodies. The main limitation lies in lack of reproducibility and identification of reactive antigen due to multiplexing.

4. **ELISA**: It is one of the standard clinical assays to detect antibodies, the main limitation is that this assay is not high throughput and requires large amount of sera. This assay requires a prior knowledge of antigen and is not effective for antigen discovery.
5. **Luminex Assays**: These assays require fewer amounts of sera compared to ELISA. Here, proteins are coupled to color coded beads, and using secondary labeled antibody binding is detected. This technology is high throughput and enables multiplexing due to wide availability of color coded beads, through which a number of antigens can be measured in parallel (Burbelo, Ching et al. 2010). The main limitation of these assays lies in tedious set up, expensive machinery and large amount of antigen.
6. **Luciferase Immunoprecipitation (LIPS)**: It is one of the solution phase immunoprecipitation assays that do not require use of radiolabeled antigen unlike traditional immunoprecipitation assays (Burbelo, Bren et al. 2007; Burbelo, Ching et al. 2007). This assay has greater performance measures and have been widely use to detect autoantibodies to autoimmune diseases especially herpes simplex virus (Burbelo et al. 2009) .The main limitation of this assay is to generate recombinant proteins with a luciferase tag which makes it more feasible for small proteomes. This assay might not be effective

for large scale screening but will be more useful when high specificity and sensitivity is required.

1.3.2 PEPTIDE BASED ASSAYS

This is one of the indirect approaches compared to protein based assays to detect target antibodies through their cross reactivities. Overlapping peptide for a proteome assays may increase the performance while reducing the labor and the cost that is involved in direct protein spotting based technologies. Several reports suggest that small peptides have the potential to mimic linear and conformational mimotopes (Legutki et al. 2010). Antibodies have been known to bind the exact sequence on the antigen (epitope) and also bind to sequence similar to the original sequence (mimitope). Hence spotting peptides on a surface have been widely used to profile antibodies from a system. Some of the approaches are as follows

1) **Phage Display:** It is one of the prominent methods to detect epitopes or mimotopes of antibody targets using cDNA fragments or random sequence peptides library respectively. Through this technology, many novel peptides sequences have been discovered that are cross reactive with a particular antibody (Derda et al. 2011; Meloen, Puijk, and Slootstra 2000). This technology typically consists of equivalent number of peptides sequences to total number of antibodies ~10^{10}. The main limitation of this technology lies in the effort that requires multiple rounds of planning, isolation and finally sequencing of phage display.

2) **Epitope Peptide Array:** Pathogen specific peptide microarrays have been widely used for epitope recognition patterns (Reineke 2009). For a small

proteome, whole proteome sequence can be dissected into small overlapping peptides and for large proteomes, certain portion of bioinformatically or clinical predicted epitopes can be spotted (Bialek, Swistowski, and Frank 2003). These peptides have a potential for immuno diagnostic application or epitope discovery (Uttamchandani and Yao 2008).

3) **Random Peptide Microarray:** It is based on using random peptides as mimotopes by constructing libraries. These can provide a quick and simultaneous measure of antibody binders and discovering diagnostic biosignatures specific to disease or vaccination. One such approach is immunosignaturing which utilizes random sequence peptide microarray to profile the humoral immune response (Boltz et al. 2009; Legutki et al. 2010; Halperin, Stafford, and Johnston 2011; Restrepo et al. 2011).

1.4 IMMUNOSIGNATURING

The Center for Innovations in Medicine at Biodesign Institute has developed a mimotopes-based immuno-diagnostic technology known as immunosignaturing. It combines the high throughput potential of random phage display libraries with the speed of the microarray. Antibody profiles generated during infection, chronic disease or vaccination can be uniquely captured by this technology, and which is informative for disease classification. The random peptide microarray consists of 10k, 20 residue peptides of random amino acid sequences, with a C-terminal linker of Gly-Ser-Cys-COOH. 19 of the amino acids (except cysteine) were selected by custom software completely at random at each of the first 17 positions followed by a constant GSC as the C-terminus linker (Legutki et al. 2010). Each slide is first treated with amino-silane and activated with sulfo-SMCC, this produces a maleimide activated surface designed to react with the cysteine terminal peptides. Each peptide is spotted in duplicate and fluorescent fiducials are applied asymmetrically with Alexa dye labeled peptides for quality assurance. Serum or plasma are diluted 1:500 in incubation buffer (3% BSA, 1XPBS, 0.05% Tween 20) and applied to the array after which secondary IgG antibody is added at 1nM. At the final stage, tertiary fluorescent labeled antibody are applied which can be detected through a scanner producing a signature of an individual. A cartoon of the immunosignaturing is shown in **Figure** 1-3

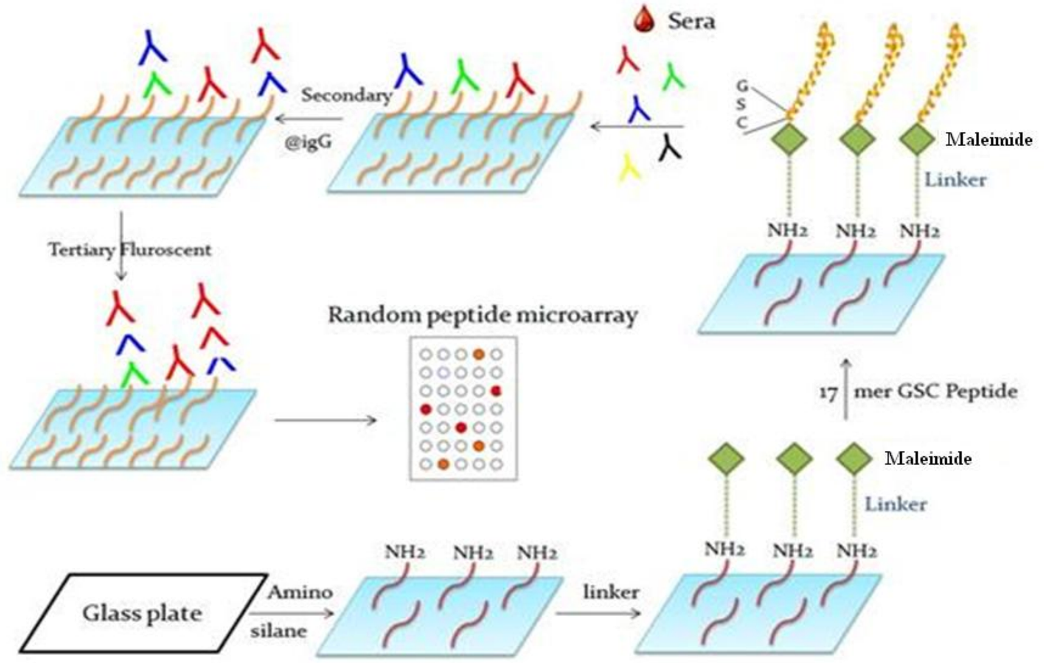


Figure 1-3 Schematic representation of how immunosignaturing works. Glass slide is preprocessed with amino silane, then peptides are spotted with the help of linker. Antibody profiles of are captured when sera/plasma/saliva IgG antibodies binding to these peptide

These arrays show promise in diagnostic applications for several diseases. Recently the physical characterization of the immunosignaturing technology demonstrated that how using of antibodies as biomarkers can solve a problem for early diagnosis (Stafford et al. 2012). The first proof of principle study on influenza showed that immunosignatures are reproducible, stable over time and can discern mice with influenza infection and immunization (Legutki et al. 2010). It has also been demonstrated that these arrays captured the complexities of the humoral immune response (Stafford et al. 2011). One of the advantages of the immunosignature is that it relies on antibodies which are relatively stable and can be accessed through sera, plasma or even saliva and can be used in immuno diagnostic applications (Chase, Johnston, and Legutki 2012). The immunosignaturing technique has also been successfully applied to distinguish Alzheimers disease (Restrepo et al. 2011), pancreas disease including pancreatic cancer, pancreatitis, panIN and type 2 diabetes (Kukreja, Johnston, and Stafford 2012) , brain cancer and several types of cancer (unpublished data). The immunosignaturing clearly outperforms when compared to other clinical approaches in processing large number of samples inexpensively. Average slide cost about 67$ to process compared to several thousand dollars other biomarker discovery methods. The immunosignaturing are now been done on a silicon surface with more than 300,000 feature on a surface. It is 30X increase in peptide coverage compared to our existing arrays. Clearly there is a lot of information that comes with this new technology and this poses a computational and statistical challenging in analyzing this novel microarray technology. Several algorithms and statistical methods have been developed for microarray data analysis in the literature including the areas in image processing of features, normalization, data transformations, feature selection methods, classification, clustering and dimensional reduction methods for microarray data analysis. With respect to immunosignaturing microarrays, novel method of feature identification using segmentation have been developed for spot analysis (Yang, Stafford, and Kim 2011) and identification of latent factors for diseases through immunosignaturing data from structural equation and mixture modeling (Brown et al. 2011). Clearly, there is lot of requirement of informatics and statistical analysis that needs to be done on this technology. The next section, reviews the current existing methods for data analysis for microarrays.

1.5 CHALLENGES IN MICROARRAY DATA ANALYSIS

Microarrays in general have been extremely useful for high throughput screening in variety of applications including sequencing (Schena et al. 1998), SNP detection (Wang et al. 1998) and also for purely computational purposes of DNA computing (Kari and Landweber 2000). With the information content that the modern microarray brings to the table, it also increases the demand of computational and statistical methods to harness this information.

One of the advantages of microarrays is that they allow high throughput inspection of biological features simultaneously. But this brings a huge challenge in data analysis due to sources of variation at each level, quantification of each feature into a single number etc. Some of the main challenges include

1. **Noise**: Microarrays including immunosignaturing are noisy in nature irrespective of scrutiny in performing experiments. The random stochastic noise is introduced inevitably at almost every step (Schuchhardt et al. 2000). This includes noise in preparation of probes of mRNA or peptides, surface chemistry, humidity, target volume, spotting methodology (pin type), slide to slide variation, hybridization parameters (time, temperature, incubation period), non specific background, artifacts of contamination, scanning, segmentation, and towards the end of process in quantification of spot signal. The main challenge arises when features across different conditions are compared to determine if a particular one (gene, peptide) are different due to condition or random noise. Noise is one of the inevitable features that microarrays have to deal with although replication and randomization help to determine the true cause of variation.
2. **Normalization**: Systematic differences often occur in microarray data due to various sources of variation. These differences might be in overall intensity, or specific trends due to artifacts of secondary or tertiary dyes. Normalization is required at almost every type of microarray analysis to remove these differences but consensus of which specific normalization method to use has not been achieved concretely.
3. **Experimental Design**: It is one of the critical steps that are often ignored during a microarray experiment. In order to get a true response of effect of input variable to the output variable, sensitivity analysis should be done to observe the true cause of the output variable. Complete random design,

randomized block design involving blocking factors might be one of the features for good experimental design.

4. **High feature problem**: Microarrays being a high throughput technology, examines thousands of feature (genes, peptides) in parallel to determine differential pattern in two or more conditions. This can be challenge due in indentifying true real features among less relevant features. Overall error rate should be keep in mind by also taking into account multiple testing corrections methods to calculate overall alpha rate. Using multiple testing procedures, one can ensure to find only relevant differential features, else with 10,000 features one can expect few peptides by random chance. To make sure, features are not just selected by chance, further clustering analysis should be performed without using prior knowledge of classes.
5. **Level of significance**: To obtain differential expression profiles of hundreds of features over fewer samples, traditional classical methods like chi square test may not be valid. Challenges arise in choosing appropriate statistical measures to discover expression patterns above a threshold error rate.
6. **Biological factors:** Although microarray technology provides a plethora of information of expression profiles of genes or peptides that are expressed or bind in different conditions, but there are many biological factors that might mislead the interpretation of this data. For gene expression microarrays, expression of genes may not correspond to the amount which determines the phenotype. For peptides and protein

microarrays, conformation, folding and PTM are biological factors which affect the signal of expression profiles. Validation of the observed features by different assays and taking orthogonal measurements should be supplemented by microarray experiments.

7. **Quality Control:** Data analysis at each step of the process is required in order to ensure quality of data that is generated through the microarray. Certain threshold metrics need to be used that are specific for each technology to discard or accept the microarray data depending upon interference of noise with the signals.

There are so many ways that microarray data can be influenced, since at each step of the procedure is complicated (Fujita et al. 2006). Hence, a protocol and standardization are made for storing large amount of data and to make it accessible to users (Brazma 2009). Some of the standards developed for quality assurances of microarray are normalization methods, ontologies for annotation, MAGE: a data exchange format and MIAME: minimum information about a microarray for making inference.

Microarrays are extremely helpful in answering the high throughput biological research hypothesis, but it is a long way from the formation of hypothesis to reaching the right inference. Figure 1-4 shows the layout for step by step procedures in microarray data analysis.

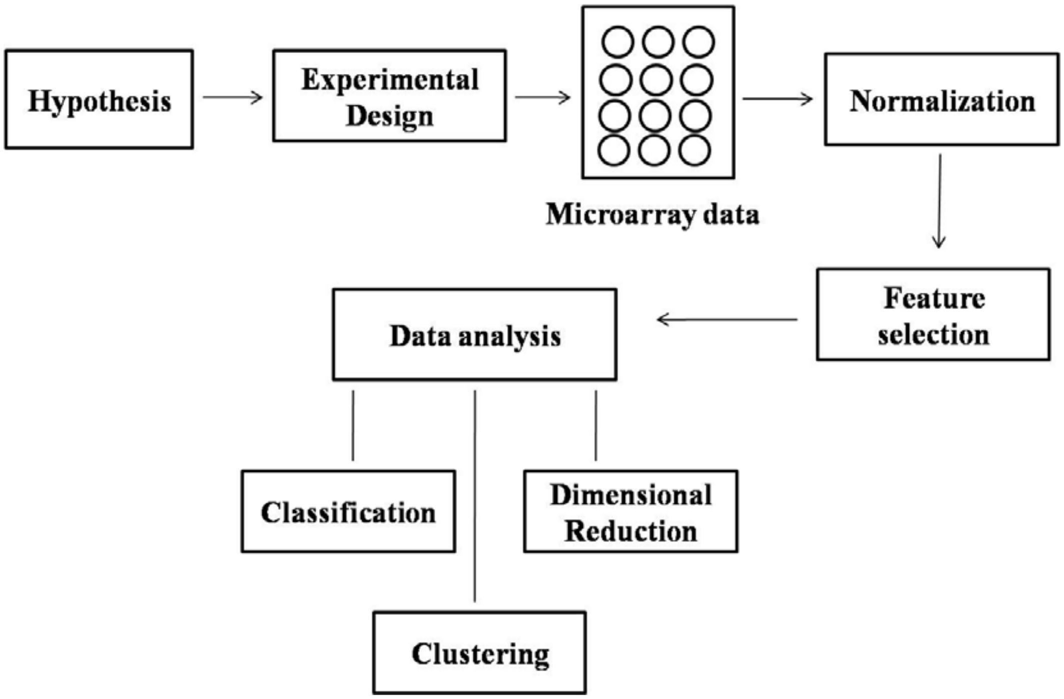


Figure 1-4: Step by Step layout for microarray data analysis

1.6 DATA PRE-PROCESSING AND NORMALIZATION

Data pre-processing and normalization are the most important step before any data analysis methods. Preprocessing basically extracts and transforms the data into more meaningful form. The most common method of pre-processing is log transform (Yang et al. 2002). It provides values that are more easily interpretable. Measuring absolute differences in the features among different conditions might lead to false inferences. Secondly, log transformation makes the distribution symmetrical and close to normal distribution (Long et al. 2001), this eases the process of applying classical statistical methods which assumes the data distribution is normal. Due to the noise associated with the microarray data, it is always advisable to do repeated measures of a sample and take an average of a sample with its technical replicate. A sample with less than <0.7 Person correlation coefficient should be discarded based on our experience. Along with combining the samples with the technical replicate, extreme caution must be given to the outliers. Outliers might be a cause of bad mechanical problem or it can be the actual sample differences which may be very heterogeneous in the population. Hence outliers should be treated with extreme caution and should not be removed from the analysis without any prior knowledge of the cause of the outliers. The primary use of microarrays is to find differences among the conditions and extract the relevant differential features associated with the difference. Hence the primary requirement before the data analysis is to normalize the data to remove bias for personnel, experimental and technology variation. There have been many normalization methods reported in the literature which can be useful for various microarray technologies. Normalization per median, in which each slide/sample is normalized to its median, is most common in peptide microarrays. Other normalization such as normalization per feature (gene or peptide) is also useful when samples of different batches or time points are compared for the analysis. Hence it is extremely important to transform the data and normalize the data depending on the hypothesis and microarray technology before proceeding towards the data analysis methods.

1.7 EXPERIMENTAL DESIGN

Experimental design is the most crucial and often neglected area in microarray experiments. For good experimental data, it is important that the experiments are carefully designed. An experiment should be well thought and designed to make purposeful changes to the input variables to observe reasons for changes in the output response (Montgomery 2009). Lack of designed experiments might lead to misleading inference about the data and hence it is very important to provide data that includes major source of variation. Data analysis is performed after the experiments are performed and hence data analyst and statisticians have often no or less control about the source of the data. In a well designed experiment, the key thing is to identify the factor which contributes to the noise.

Some of the key guidelines for experimental design are as follows

1. **Research Hypothesis**: Stating the experimental question in detail and with specification is extremely useful before designing the experiment efficiently. Doing a literature search on the similar research questions might give an idea about effective design and possible outcomes.
2. **Microarray Technology**: It is extremely useful to make an appropriate choice of microarray technology that can address the research question specifications. Choosing between cDNA/oligo arrays, commercial/custom arrays, brand specific, which features to spot (gene, real space peptides, random sequence peptides, proteins etc) can be difficult but extremely vital to address for the research question.
3. **Factors of interests**: The primary vital task in microarray experiments is to identify major inputs or factors that affect the output of the experiment for example testing efficacy of drug on given subjects, finding differential

features (gene, peptides) among conditions of interests. At the same time, it is also important to identify nuisance factors that would adulterate the output of the experiment and hence necessary methods (replication, randomizations) can be performed to rectify the effect of nuisance factors.

4. **Threshold statistics**: Once the research hypothesis is formed, one of the prime decisions is choose the optimal value of type 1 and type 2 errors. Depending on the specificity and sensitivity level required to make the inference, its useful to decide when to reject the true null hypothesis and vice versa. Balanced has to achieved between power (rejecting null hypothesis when it is false) and alpha rate (not rejecting null hypothesis when it is false). Typically in microarray experiments power is chosen to be >90% and alpha rate to be less than 5%.
5. **Data analysis**: Recording the data accurately is as important as data analysis and effective tools, computational and statistical methods should be chosen to make right and meaningful inference of the research hypothesis.

The basic principles involved in the experimental design are:

1.7.1 REPLICATION

It is one of the key methods to remove experimental error by doing performing a parameter analysis more than once. The prime purpose for doing the replication is to test whether the observed differences in the data are significant. Replication at various levels in the microarray experiments enable the data quality and hence strengthen the inference drawn from the data. Typically in the microarray experiments, replication should be performed starting from spot level where every feature (gene or peptide) should be spotted randomly across the slide to remove any bias of location on the intensity of the spot. Also, replication should be performed at the chip/slide level where a single biological sample is run of multiple slides under the same condition to remove any slide bias. This is often referred as technical replicate and any sample with Pearson correlation of less than 0.8 to its technical replicate should be discarded. On the top of that, microarray experiments should also include biological replicates to remove individual bias and personal variation that may adulterate the inference. Pooling of sample is often recommended to remove any individual biases but it is sometimes criticized to dilute the actual signal coming from individual and hence pooling should be used with caution in microarray experiments.

1.7.2 RANDOMIZATION

It is a technique often used in microarray experiments to opt for random choices for factors that are not of interests. These factors are often referred as nuisance factors which might influence the outcome of a microarray experiment. Spotting features randomly across the slide allows removing any biases due to location, choosing slides and running the sample condition randomly allows removing biases of slide or environment. Randomization may be sometimes be difficult to incorporate at various levels but often recommended to draw meaningful inferences.

1.7.3 BLOCKING

A block is a subset of homogenous experimental conditions that are created for keeping the nuisance factor constant and then allowing the factor of interest to vary (Montgomery 2009). It allows eliminating the variability due to differences in homogenous blocks, this block can be age, sex, geographical location, print run batch etc. Both the randomization and blocking aim to reduce the nuisance factors with the difference that blocking can only be applied for factors which are under control. If the nuisance factors are not under control, randomization is the only solution.

1.8 FEATURES SELECTION IN MICROARRAYS

Microarray data expression values are often compared to a fixed number or between two or more samples expression values. In order to make statistical significant inference of the differential pattern, various tests are performed. Some of the key statistical methods are as follows:

1.8.1 ONE SAMPLE T-TEST

It is performed when an observed expression values of a feature (gene or peptide) e.g. vector ā is compared to a known fixed value to see if the observed expressional value is significantly different from the fixed expression value c. The test statistics used here is student t-distribution, where X' is average of vector ā, c is known fixed value, s is the standard deviation of vector ā and n is the number. of observations in vector ā

$$t = \frac{X' - c}{\frac{s}{\sqrt{n}}}$$

From this equation t-values are calculated and compared to a t-table with df = n -1 to obtain the significance level (p value) and the hypothesis of no of difference between expression values of feature of interest to a fixed value is rejected if the p value is <0.05. The basic assumption made here is vector ā values follow normal distribution (Schena et al. 1995).

Paired testing

It's often of interest to find out which features change significantly when a sample is exposed to a certain condition like temperature, environment etc. These are known as simultaneous tests, before and after tests or matched tests. In this case, statistical test is required to test if there is a significant difference between a feature before and after the test. For such experiments, one sample t-test is performed where X' is replaced with difference between the mean Xd'.

1.8.2 CHI SQUARE TEST FOR EXPRESSION VARIANCE

Often variance is calculated for expression values of feature of interest (gene, peptide) multiple times through different technology to access if one technology provides significantly more uniform distribution of a feature of interest. In this case, chi square test statistics is calculated by

$$\chi 2 = \frac{(n - 1)s^2}{\sigma^2}$$

Here s is the standard deviation of the sample and σ is the standard deviation of the population, if test statistics is larger than the critical value, null hypothesis of no significant difference between the variance is concluded (Schena et al. 1995).

1.8.3 TWO SAMPLE TEST FOR MEANS

Microarray expression values of one or more feature(s) (gene or peptide) for two conditions (disease vs normal) are often compared to find if there is a significant difference between the mean of a feature in each condition. If the interest is in the higher /lower value of a feature in one condition then one tail/two tail test is performed respectively.

The test statistics used here is student t-distribution where X1' and X2' represent the average of a feature in condition 1 and 2 respectively. The term s1 and s2 denotes the standard deviation of a feature in condition 1 and 2 respectively. The term u1 and u2 represent the expected mean under the null hypothesis for condition 1 and 2 respectively. If the variance of a feature is not significantly different in two conditions, sp^2 (pooled variance is used) or else separate variances are used for t-statistics calculation, n1 and n2 denotes the sample size for condition 1 and 2 respectively.

For equal variance

$$sp^2 = \frac{(n1-1)s1^2 + (n1-1)s2^2}{n1+n2-2}$$

$$t = \frac{X1' - X2' - u1 - u2}{sp * \sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

If the t-statistic is higher than the critical value at a defined significant level (95%) with degree of freedom v = n1+ n2 - 2 then the null hypothesis of mean of feature is equal in condition 1 and 2 is rejected (Schena et al. 1995).

For unequal variance

$$t = \frac{X1' - X2' - \ u1 - u2}{sp * \sqrt{\frac{s1^2}{n1} + \frac{s2^2}{n2}}}$$

$$v = \frac{\left(\frac{s1^2}{n1} + \frac{s2^2}{n2}\right)^2}{\frac{\frac{s1^{2^2}}{n1}}{n1 - 1} + \frac{\frac{s2^{2^2}}{n2}}{n2 - 1}}$$

If the t-statistic is higher than the critical value at a defined significant level (95%) with the degree of freedom υ, then the null hypothesis of equality of the mean of the feature in 2 conditions is rejected.

1.8.4 ONE WAY ANOVA

A feature of interest is often compared in multiple conditions to test if there is a significant difference among the means of a feature. For this, F-test or 1-way ANOVA is used. The basic assumption of this model is that k samples are independent and all the populations are same variance with normal distribution. There are two sources of variation when the multiple groups are involved. SS (treat) is variation due to difference in the means of the groups and SS (error) is variation due to random variation. The mean square MS (treat) and MS (error) is calculated for the test statistics to find out if these variations are significantly different by taking the ratio of sum of squares to the degree of freedom. The degree of freedom for treatment group is r-1, where r is the number of groups and for error is n-r where n is total number of samples. The F test basically tests if the MS (treat) is significantly higher than MS (error). If the obtained F value is then higher than the critical value of F (1-α, r-1, n-r), then the null hypothesis of equality of the means in different conditions is rejected and we conclude that there is a significant among the means of a feature of interest in multiple conditions.

1.8.5 TWO WAY ANOVA

1-way ANOVA methods investigate data which is influenced by only one factor. Often a feature (peptide or gene) might be affected by two factors which may or may be not independent. For such a study, 2-way ANOVA is performed. The overall goal for such a study is to see if the mean of a feature is different in treatment A conditions and also in treatment B conditions. Such analysis is often complicated if there is interaction between treatment A and treatment B. Hence a test for interaction is performed where the mean square due to interactions is compared to mean square error and if F value obtained is greater than the critical value then we conclude that there is a significant interaction between the two treatments. If there is no significant interaction between the two treatments, then individual F-tests are performed for effect of factor A and factor B separately on a particular feature of interest. This is done by taking ratio of mean square for MS (treat A) and MS (error) and also ratio of MS (treat B) and MS (error) to calculate F statistics specific to the hypothesis for treatment A and treatment B respectively.

1.8.6 MULTIPLE COMPARISONS

Typically in microarray experiments, large numbers of features are simultaneously tested for significance in various conditions. In that case there is always a chance of an inflated overall alpha level. For each feature tested against the alpha level, the overall probability to obtain false positives increases. For example, if each feature (gene or peptide) is tested against 5% type 1 error level, and simultaneously 10,000 features are tested, the overall probability of type 1 error increases significantly. To maintain the overall probability of type 1 error as 5%, several corrections are suggested including Bonferroni, Benjamini and Hochberg false discovery rates (Hochberg and Benjamini 1990).

1.9 DATA ANALYSIS METHODS IN MICROARRAYS

Once the optimal number of differential features is selected after choosing the appropriate feature selection method and correction of multiple testing, data becomes ready for analysis. With the few selected features for different conditions, various analysis and visualization tools have been developed. Some of the key visualization tools are

1. **Box plots**: It graphically represents several descriptive statistics of a given data sample. It gives the visual representation of a particular sample or class 25th, 50th (median) and 75th percentiles along with outlier points.
2. **Histogram:** It is a graph showing frequency distribution of values with horizontal axis showing the range of values and vertical axis showing the

distribution. This representation is useful to determine the distribution of data (normal, left/right skewed, bimodal, uniform etc).

3. **Scatter plot :** It is the simplest tool to represent two classes on horizontal and vertical classes respectively. This graph is useful to visualize how feature intensity differs in two conditions and thus a range of features can be selected from scatter plot fulfilling certain conditions (2X, >2X fold etc)
4. **Line graph:** It is useful in time series analysis or comparing certain features of interest over different samples. It gives a visual representation of trajectory of features over selected samples.

1.9.1 CLASSIFICATION METHODS FOR MICROARRAYS

Once the optimal number of differential features is selected by the feature selection method, a classifier is needed to build a model that can classify different classes of interests based on the features selected. Once a model is built on the training data, this model is tested for its performance against the test data. Model performance can be calculated by various metrics including accuracy, sensitivity, specificity and area under ROC curve. There have been algorithms described in the literature specific for different types of microarray data. A more complete review is presented in chapter 3. Linear discriminant analysis is one of the traditional methods for studying gene expression data. But certain technologies like protein arrays and peptide arrays have a different mechanism of binding of targets to the probe due to which traditional methods may not work effectively. Due to this reason, a classifier should be chosen with caution with respect to the microarray technology involved.

1.9.2 CLUSTERING METHODS FOR MICROARRAYS

Clustering analysis is the most frequently used multivariate technique to analyze different types of microarrays. This technique is most appropriate when no prior knowledge regarding the data is known or unsupervised learning is required. In recent years the clustering analysis of biological data in unsupervised setting has caught the attention of many researchers that has resulted to use clustering analysis to analyze their microarray data (Yeh et al. 2009; Jupiter and VanBuren 2008). Clustering aims to group certain objects based on a similarity measure called distance. The distance between two n-dimensional vectors is calculated by a distance metric. Some of the frequent distance metrics are Euclidean, Manhattan, Chebchev, Correlation, Mahalannobis and Minkowski distance (de la Fuente, Brazhnik, and Mendes 2002). Prior to using a distance metrics, it is extremely important to normalize (per slide or per feature) and transform (z-transform or log transform) the data to remove any experimental bias. The important question before the clustering methods is choice of various distance measures mentioned above. Every distance metrics have their own limitations and advantages; hence it is extremely vital to examine the characteristics of feature (gene or peptide) and samples. Some of the rule of thumbs involves using Euclidean and Manhattan metrics when the clustering features are different while the clustering samples are same. For the opposite case, correlation distance metrics is more appropriate. There are many clustering methods developed and used for microarray expression data. One of the simplest and most commonly used methods is k-means clustering which require prior knowledge of the number of clusters (k) (Spadone et al. 2012; Yu et al. 2012). The algorithm randomly chooses k points as centers and in every iteration, updates the centers based on inter and intra clusters distances. Towards the end of the iterations, k clusters are formed. This method is quite robust, but is vulnerable to the initial selection of k points and the number of iterations. Hence these two parameters should be varied to obtain a stable set of k clusters. This method is often used to cluster samples based on their microarray expression signatures. Apart from clustering samples into separate clusters, hierarchical clustering is considered to be more appropriate for microarray expression data (Eisen et al. 1998; Heyer, Kruglyak, and Yooseph 1999). Here, a dendrogram is constructed based on a distance metrics between either samples or features (gene or peptide) or even both. One classical representation of hierarchical clustering in microarray expression data is heat map. Here, samples and features are clustered hierarchically in x axis and y axis respectively and the spot intensity for the corresponding feature and sample are represented as gradient of colors where red color denotes high binding, yellow is average binding and blue as low binding. Heat maps help to visualize the binding of selected features over samples in the experiment. It gives the picture of how samples are clustered together in axis as a tree and how features are clustered together in y axis.

1.9.3 DIMENSIONAL REDUCTION PROCEDURES

One of the challenges in the microarray experiments is the large number of dimensions. Every experiment has at least 10 samples in each class, and every sample is run for 10,000 features (genes or peptides) or more. This leads to the data explosion for microarrays. For each sample having 10,000 data points or more, it becomes a challenge to visualize. A natural approach is to try to reduce the number of dimensions by eliminating the irrelevant dimensions. A common approach in this regard is constructing fewer dimensions that account for most variation in the data. This is the approach used by **Principal Component Analysis (PCA)**. It basically works by calculating a new system of dimension or coordinates, which are in linear combination of all the other variables in such a way to incorporate maximum variance. The direction of the coordinates are eigenvectors of the covariance matrix of the patterns. The eigenvalue with the largest absolute value will signify maximum variation along its eigenvector. The projections of the data points in the new dimensional systems are called the principal components of the data. By projecting the n-dimensional input vector into a space of 2 or 3 dimensions, dimensional reduction is achieved. Apart from its usefulness, it has some serious limitations of only relying on the first order statistical characteristics (variance) of the data (Eisen et al. 1998; Hilsenbeck et al. 1999). Another approach in the dimensional reduction procedure is **Independent Component Analysis (ICA)** (Bell and Sejnowski 1995), which considers higher order statistical dependencies (kurtosis, skewness) for separating a multivariate microarray data into additive subcomponents assuming the mutual statistical independence. The prime difference between the two approaches is that in PCA, the directions of the new axes are perpendicular to each other, while in ICA the new axes are not necessarily perpendicular to each other. ICA has been found to be more effective for solving blind source separation and has the potential to separate the multiplexed signals from individual features (genes or peptides).

CHAPTER 2 COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS FOR IMMUNOSIGNATURING DATA

2.1 ABSTRACT

High-throughput technologies such as DNA, RNA, protein, antibody and peptide microarrays are often used to examine differences across drug treatments, diseases, transgenic animals, and others. Typically one trains a classification system by gathering large amounts of probe-level data, selecting informative features, and classifies test samples using a small number of features. As new microarrays are invented, classification systems that worked well for other array types may not be ideal. Expression microarrays, the most prevalent array type, have been used for years to help develop classification algorithms. Many biological assumptions are built into classifiers that were designed for these types of data. One of the more problematic is the assumption of independence, both at the probe level and again at the biological level. Probes for RNA transcripts are designed to bind single transcripts. At the biological level, many genes have dependencies across transcriptional pathways where co-regulation of transcriptional units may make many genes appear as being completely dependent. Thus, algorithms that perform well for gene expression data may not be suitable when other technologies with different binding characteristics exist. The immunosignaturing microarray is based on complex mixtures of antibodies binding to arrays of random sequence peptides. It relies on many-to-many binding of antibodies to the random sequence peptides. Each peptide can bind

multiple antibodies and each antibody can bind multiple peptides. This technology has been shown to be highly reproducible and appears promising for diagnosing a variety of disease states. However, it is not clear what is the optimal classification algorithm for analyzing this new type of data.

We characterized several classification algorithms to analyze immunosignaturing data. We selected several datasets that range from easy to difficult to classify, from simple monoclonal binding to complex binding patterns in asthma patients. We then classified the biological samples using 17 different classification algorithms.

Using a wide variety of assessment criteria, we found the 'Naïve Bayes' far more useful than other widely used methods due to its simplicity, robustness, speed and accuracy.

The 'Naïve Bayes' algorithm appears to accommodate the complex patterns hidden within multilayered immunosignaturing microarray data due to its fundamental mathematical properties.

2.2 INTRODUCTION

Serological diagnostics have received increasing scrutiny recently (Haab 2003; Whiteaker et al. 2007) due to their potential to measure antibodies rather than low-abundance biomarker molecules. Antibodies avoid the biomarker dilution problem and are recruited rapidly following infection, chronic, or autoimmune episodes, or exposure to cancer cells. Serological diagnostics using antibodies have the potential to reduce medical costs and may be one of the few methods that allow for true pre-symptomatic detection of disease. For this reason, our group has pursued immunosignaturing for its ability to detect the diseases early and with a low false positive rate. The platform consists of a peptide microarray with either 10,000 or 330,000 peptides per assay. This microarray is available for standard mathematical analysis, but for a variety of reasons, certain methods of classification enable the best accuracy (Reimer, Reineke, and Schneider-Mergener 2002) (Merbl et al. 2009). Classification methods differ in their ability to handle high or low numbers of features, the feature selection method, and the features' combined contribution to a linear, polynomial, or complex discrimination threshold. Expression microarrays are quite ubiquitous and relevant to many biological studies, and have been used often when studying classification methods. However, immunosignaturing microarrays may require that we change our underlying assumptions as we determine the suitability of a particular classifier.

In order to establish the question of classification suitability, we examine a basic classification algorithm, Linear Discriminant Analysis (LDA). LDA is widely used in analyzing biomedical data in order to classify two or more disease classes (Braga-Neto and Dougherty 2004) (Stemke-Hale et al. 2005) (Sima et al. 2005) (Braga-Neto and Dougherty 2004). One of the most commonly used high-throughput analytical methods is the gene expression microarray. Probes on an expression microarray are designed to bind to a single transcript, splice variant or methylation variant of that transcript. These one-on-one interactions provide relative transcript numbers and cumulatively help to define high-level biological pathways. LDA uses these data to define biologically relevant classes based on the contribution of differentially expressed genes. This method often uses statistically identified features (gene transcripts) that are different from one condition to another. LDA can leverage coordinated gene expression to make predictions based on a fundamental biological process. The advantage of this method is that relatively few features are required to make sweeping predictions. When features change sporadically or asynchronously, the discriminator predictions are adversely affected. This causes low sensitivity in exchange for occasionally higher discrimination. Tree-based methods use far more features to obtain a less biased but less sensitive view of the data. These methods can partition effects even if the effect sizes vary considerably. This approach can be more useful than frequentist approaches where it is important to maintain partitions in discreet groups.

Immunosignaturing has its foundations in both phage display and peptide microarrays. Most phage display methods that use random-sequence libraries also use fairly short peptides, on the order of 8-11 amino acids (Cwirla et al. 1990). Epitope microarrays use peptides in the same size range, but typically far fewer total peptides, on the order of hundreds to thousands (Nahtman et al. 2007). Each of these methods assumes that a single antibody binds to a single peptide, which is either detected by selection (phage display) or by fluorescent secondary antibody (epitope microarray). Immunosignaturing uses long 20-mer random-sequence peptides that have potentially 7 or more possible linear epitopes per peptide. Although immunosignaturing must make do with only 10,000 to

~300,000 peptides, the information content derived from partial binding makes these data useful in ways quite different from phage display (Boltz et al. 2009) (Brown et al. 2011) (Halperin, Stafford, and Johnston 2011) (Legutki et al. 2010) (Restrepo et al. 2011).

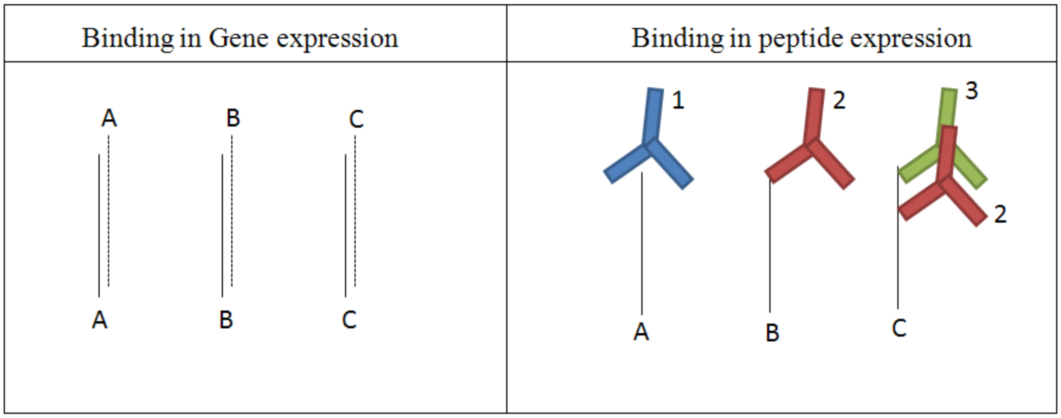


Figure 2-1 One-to-one correspondence found in gene expression microarrays is not observed for the immunosignaturing arrays.

The complexity in analysis arises from the many-to-many relationship between peptide and antibody **Figure 2-1**. This relationship imposes a particular challenge for classification because a simple one-to-one relationship between probe and target, idiomatic for gene expression microarrays, allows a coherent contribution of many genes that behave coordinately based on biological stimuli. That idiom is broken for immunosignaturing microarrays, where each peptide may bind a number of different antibodies and every antibody might bind a number of peptides. Unless disease-specific antibodies find similar groups of peptides across individuals, very little useful information is available to the classifier. The aim of this work is to assess the performance of various classification algorithms on immunosignaturing data.

We have considered 17 diverse data mining classification methods. For feature selection, we used a simple t-test when we examined two classes, and a fixed-effects 1-way ANOVA for multiple classes with no *post-hoc* stratification. We have assessed these algorithms' ability to handle increasing numbers of features by providing four different sets of peptides with increasing p-value cutoff. The four levels include from 10 (minimum) to >1000 (maximum) peptides. Each algorithm is thus tested under conditions that highlight either synergistic or antagonistic effects as the feature numbers increase.

2.3 MATERIALS AND METHODS

2.3.1 TECHNOLOGY

A peptide microarray described previously (Boltz et al. 2009) (Brown et al. 2011) (Halperin, Stafford, and Johnston 2011) (Legutki et al. 2010) (Restrepo et al. 2011) was used to provide data for analysis. Two different sets of 10,000 random peptide sequences are tested. The two peptide sets are non-overlapping and are known as CIM10Kv1 and CIM10Kv2.

2.3.2 SAMPLE PROCESSING

Samples consist of sera, plasma or saliva – each produces a suitable distribution of signals upon detection with an anti-human secondary IgG-specific antibody. Samples are added to the microarray at 1:500 dilutions in sample buffer (1xPBS, 0.5% Tween20, 0.5% Bovine Serum Albumin (Sigma, St. Louis, MO)), IgG antibodies are detected through a biotinylated secondary anti human IgG antibody (Novus anti-human IgG (H+ L), Littleton, CO), which binds the primary antibody. Fluorescently labeled streptavidin is used to label the secondary antibodies and the slide is scanned with an Agilent 'C' laser scanner in single-color mode. 16-bit images are processed using GenePix Pro 8, which provides the tabular information for each peptide in a continuous value ranging from 0-65,000. Four unique data sets have been used in this analysis, 2 run on the CIM10Kv1 and 2 on the CIM10Kv2. Each individual sample was run in duplicate; replicates with >0.8 Pearson correlation coefficient were considered for analysis.

2.3.3 DATASETS

Center for Innovations in Medicine, Biodesign Institute, Arizona State University has an existing IRB 0912004625, which allows analysis of blinded samples from collaborators.

a.) **T1D data set**: This dataset contains 80 sera samples (41 controls and 39 T1D children ages 6 to 13). These samples were tested on the CIM10kV1microarrays.

b.) **Alzheimer's disease data set**: This dataset contains 23 samples (12 controls and 11 Alzheimer's disease subjects). These were tested on the CIM10kV2 microarrays.

c.) **Antibodies dataset**: This dataset contains 50 samples and has 5 groups of monoclonal antibodies, arbitrarily arranged. All monoclonal were raised in mouse, and use the same secondary detection antibody. Samples were run on the CIM10kV1 microarrays.

d.) **Asthma dataset**: This dataset consists of 47 unique samples containing serum from patients in 4 distinct classes corresponding to the household environment.

Condition A consists of 12 control subjects who had no environmental stimuli. Condition B consists of 12 subjects who had stimuli but no asthma-related symptoms. Condition C consists of 11 subjects who had no stimuli but with clinical asthma. Condition D consists of 12 subjects who have both stimuli and clinical asthma. Samples were tested on the CIM V2 10K microarrays. Asthma datasets were been analyzed by considering all four conditions using ANOVA in order to study the combined effect of stimuli and asthma on subjects and then by considering pair wise comparison of condition A vs. B, A vs. C, and B vs. D.

2.3.4 DATA PREPROCESSING, NORMALIZATION AND FEATURE SELECTION

The 16-bit tiff images from the scanned microarrays were imported into GenePix Pro 6.0 (Molecular Devices, Santa Clara, CA). Raw tabular data were imported into Agilent's GeneSpring 7.3.1 (Agilent, Santa Clara, CA). Data were median normalized per array and $\log_{10}$ transformed. For feature selection we used Welch-corrected T-test with multiple tested (FWER=5%). For multiple groups (Antibody and Asthma datasets) we used 1-way fixed-effects ANOVA.

2.3.5 DATA MINING CLASSIFICATION ALGORITHMS

Four distinct peptide features are chosen for the comparison study. For each analysis, peptides are selected by t-test or ANOVA across biological classes, with 4 different p-value cutoffs. Cutoffs were selected to obtain roughly equivalent sized feature sets to assess the ability of each algorithm to process sparse to rich feature sets. Once the significant features were collected, data was imported into WEKA (Hall et al. 2009) for classification. The algorithms themselves spanned a wide variety of classifiers including Bayesian, regression based methods, meta-analysis, clustering, and tree based approaches.

We obtained accuracy from each analysis type using leave-one-out cross-validation. We obtained a list of t-test or ANOVA-selected peptides at each stringency level. The highest stringency uses peptides with p-values in the range of 10^{-5} to 10^{-10} and contains the least ‘noise’. The less-stringent second set uses p-values approximately 10-fold higher than the most stringent. The third contains the top 200 peptides and the forth contains ~1000 peptides at $p<0.05$. Although different numbers of peptides are used for each dataset, each peptide set yields the same general ability to distinguish the cognate classes. The WEKA default setting of parameters were used for every algorithm to avoid bias and over fitting. These default parameters are taken from the cited papers listed below for each algorithm. Brief details of default parameters and algorithms are listed

1. **Naïve Bayes**: Probabilistic classifier based on Bayes theorem. Numeric estimator precision values are chosen based on analysis of the training data. In the present study, normal distribution was used for numeric attributes rather than kernel estimator (John and Langley 1995).
2. **Bayes net**: Probabilistic graphical model that represents random variables and conditional dependencies in the form of a directed acyclic graph. A Simple Estimator algorithm has been used for finding conditional probability tables for Bayes net. A K2 search algorithm was used to search network structure(Friedman, Geiger, and Goldszmidt 1997) (Yu and Chen 2005).

3. **Logistic Regression (Logistic R.)**: A generalized linear model that uses logistic curve modeling to fit the probabilistic occurrence of an event(Friedman, Hastie, and Tibshirani 1998). The Quasi-Newton method is used to search for optimization. 1x10^8 has been used for ridge values in the log likelihood calculation (Cessie and Houwelingen 1992).
4. **Simple Logistic**: Classifier for building linear logistic regression models. For fitting the logistic model 'LogitBoost', simple regression functions are used. Automatic attribute selection is obtained by cross validation of the optimal number of 'LogitBoost' iterations (Landwehr, Hall, and Frank 2005). Heuristic stop parameter is set at 50. The number of maximum iterations for LogitBoost has been set to 500.
5. **Support Vector Machines (SVM)**: A non-probabilistic binary linear classifier that constructs one or more hyper planes to be can be used for classification. For training support vector classes, John Platt's sequential minimal optimization algorithm was used which replaces all missing values (Platt 1998). Here multiclass problems are used using pair-wise classification. The complexity parameter is set to 1. Epsilon for round off error is set to 1x10^{*-12}. PolyKernel is the set to be kernel. The tolerance parameter is set to 0.001 (Hastie and Tibshirani 1998; Keerthi et al. 2001).
6. **Multilayer Perceptron (MLP):** A supervised learning technique with a feed forward artificial neural network through back-propagation that can classify non-linearly separable data (Chaudhuri and Bhattacharya 2000; Gardner and Dorling 1998). The learning rate is set to 0.3 and momentum

applied during updating weights is set to 0.2. The validation threshold use to terminate the validation testing is set to 20.

7. **K nearest neighbors (KNN):** Instance based learning or lazy learning which trains the classifier function locally by majority note of its neighboring data points. Linear NN Search algorithm is used for search algorithm (Aha, Kibler, and Albert 1991; Weinberger, Blitzer, and Saul 2006). K is set to 3.
8. **K Star:** Instance based classifier that uses similarity function from the training set to classify test set. Missing values are averaged by column entropy curves and global blending parameter is set to 20 (Cleary and Trigg 1995).
9. **Attribute Selected Classifier (ASC):** 'Cfs subset' evaluator is used during the attribute selection phase to reduce the dimension of training and test data. The 'BestFit' search method is invoked after which J48 tree classifier is used (Hall 1998).
10. **Classification via clustering (K means)**: Simple k means clustering method is used where k is set to the number of classes in the data set (Hartigan 1985). Euclidean distance was used for evaluation with 500 iterations.
11. **Classification via Regression (M5P)**: Regression is a method used to evaluate the relationship between dependent and independent variables through an empirically determined function. The M5P base classifier is used which combines conventional decision tree with the possibility of

linear regression at the nodes. The minimum number of instances per leaf node is set to 4 (Quinlan 1992).

12. **Linear Discriminant Analysis (LDA)**: Prevalent classification technique that identifies the combination of features that best characterizes classes through linear relationships. Prior probabilities are set to uniform and the model as homoscedastic.
13. **Hyper Pipes**: Simple, fast classifier that counts internally defined attributes for all samples and compares the number of instances of each attribute per sample. Classification is based on simple counts. Works well when there are many attributes (Ian H. Witten 2011).
14. **VFI**: Voting feature interval classifier is a simple heuristic attribute-weighting scheme. Intervals are constructed for numeric attributes. For each feature per interval, class counts are recorded and classification is done by voting. Higher weight is assigned to more confident intervals. The strength of the bias towards more confident features is set to 0 (G¸venir and «akIr 2010).
15. **J48**: Java implementation of C4.5 algorithm. Based on the Hunt's algorithm, pruning takes place by replacing internal node with a leaf node. Top-down decision tree/voting algorithm (Salzberg 1994). 0.25 is used for the confidence factor. No Laplace method for tree smoothing (Quinlan 1996).
16. **Random Trees**: A tree is grown from data that has K randomly chosen attributes at each node. It does not perform pruning. K-value (log$_2$

(number of attributes) + 1) is set at zero. There is no depth restriction. The minimum total weight per leaf is set to 1 (Ian H. Witten 2011).

17. **Random Forest (R. Forest)**: Like Random Tree, the algorithm constructs a forest of random trees (Breiman 2001) with locations of attributes chosen at random. It uses an ensemble of unprune decision trees by a bootstrap sample using training data. There is no restriction on the depth of the tree; number of tress used is 100.

2.3.6 TIME PERFORMANCE

CPU time was calculated for every algorithm at the four different significance levels. This time was measured on a standard PC (Intel dual core, 2.2 GHz 3 Gb RAM) that was completely dedicated to WEKA. To measure CPU time, open source jar files from WEKA were imported to Eclipse where the function 'time ()' was invoked prior to running the classification including the time required for cross validation. Most Windows 7 services were switched off; the times reported were an average of 5 different measurements.

2.4 RESULTS

2.4.1 OVERALL PERFORMANCE MEASURE OF CLASSIFICATION ALGORITHMS OVER ALL DATASETS

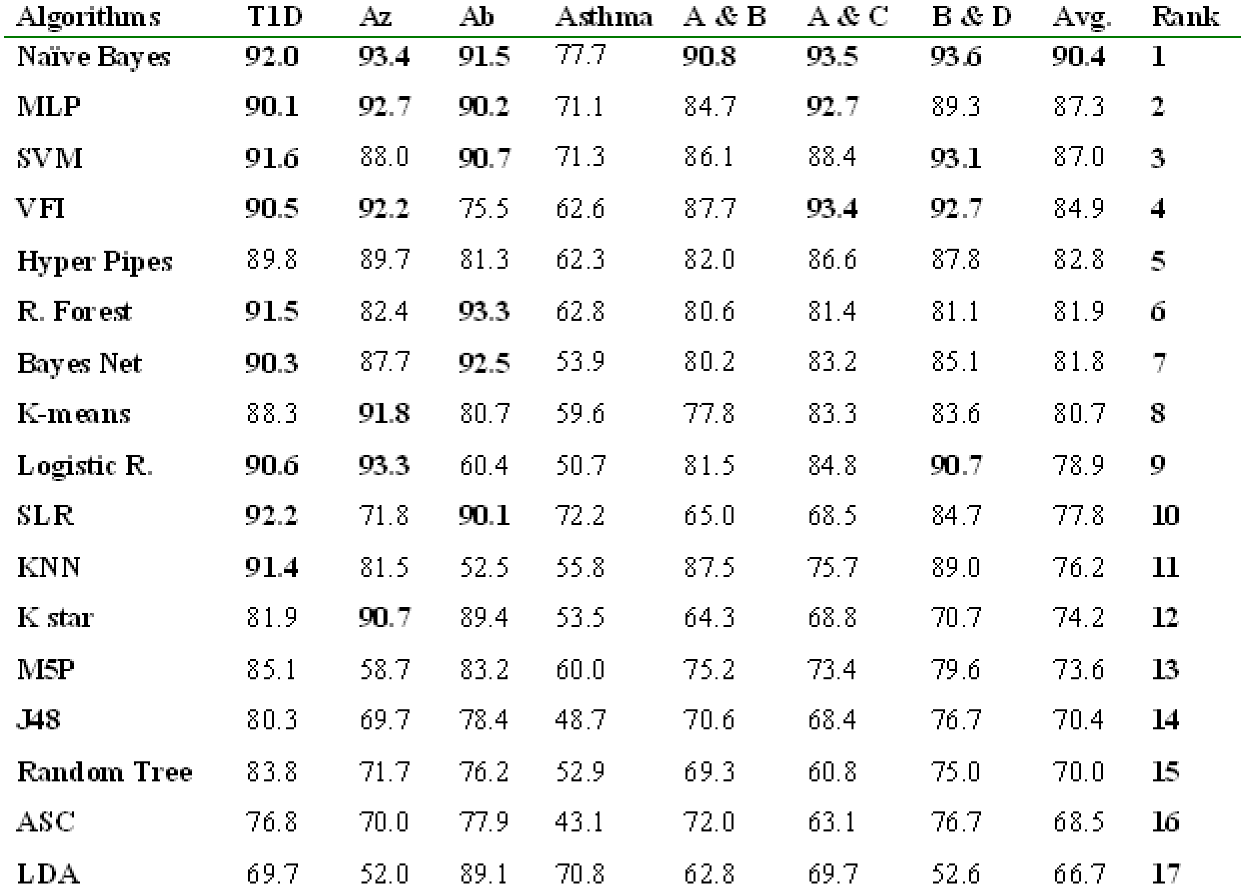
For each dataset, accuracies are measured at four levels (top 10, 50,200, 1000 peptides) at various levels of significance. Overall average performance measure is calculated for each algorithm for a given data set.

Table 2-1 shows the overall average percentage score for each algorithm calculated by averaging accuracy, specificity, sensitivity and area under ROC curve under all levels of significance. Scores >90% are marked in bold. The MLP algorithm did not finish due to huge memory requirements on last level of significance and is averaged based on first three levels of significance. For type 1 diabetes, Alzheimer's and antibodies dataset, >6 algorithms scored >90% average score. Overall, Naïve Bayes had the highest average score (90.4%) and was always among top 3 algorithms among all datasets.

2.4.2 PERFORMANCE MEASURE OF CLASSIFICATION ALGORITHMS AT DIFFERENT LEVELS OF SIGNIFICANCE OVER ALL DATA SETS

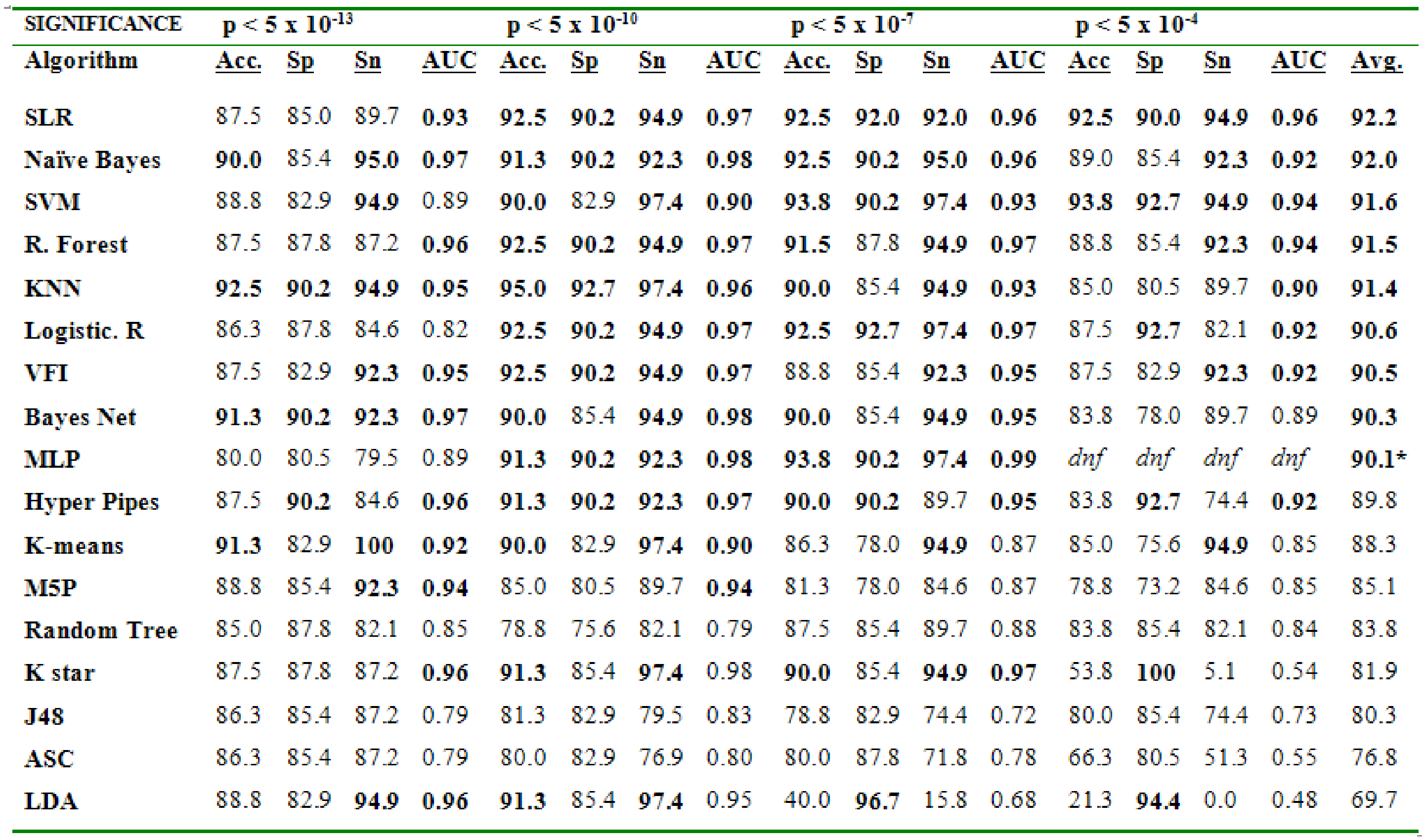
For each data set, different levels of significance are chosen to measure the performance accuracy of each algorithm. These levels contain approximately equal number of peptides for each data set. The first level contains 10 peptides selected from the t-test (lowest p value) and hence contains the least noise. Next, approximately 50 peptides, 200 peptides and 1000 peptides were chosen for the other three levels. **Table 2:2-8** shows 4 different performance measures (accuracy, specificity, sensitivity and area under ROC curve) at different levels of significance over 7 datasets.

Table 2-1: Overall performance measures of classification algorithms on datasets.



Algorithms	T1D	Az	Ab	Asthma	A & B	A & C	B & D	Avg.	Rank
Naïve Bayes	**92.0**	**93.4**	**91.5**	77.7	**90.8**	**93.5**	**93.6**	**90.4**	**1**
MLP	**90.1**	**92.7**	**90.2**	71.1	84.7	**92.7**	89.3	87.3	**2**
SVM	**91.6**	88.0	**90.7**	71.3	86.1	88.4	**93.1**	87.0	**3**
VFI	**90.5**	**92.2**	75.5	62.6	87.7	**93.4**	**92.7**	84.9	**4**
Hyper Pipes	89.8	89.7	81.3	62.3	82.0	86.6	87.8	82.8	**5**
R. Forest	**91.5**	82.4	**93.3**	62.8	80.6	81.4	81.1	81.9	**6**
Bayes Net	**90.3**	87.7	**92.5**	53.9	80.2	83.2	85.1	81.8	**7**
K-means	88.3	**91.8**	80.7	59.6	77.8	83.3	83.6	80.7	**8**
Logistic R.	**90.6**	**93.3**	60.4	50.7	81.5	84.8	**90.7**	78.9	**9**
SLR	**92.2**	71.8	**90.1**	72.2	65.0	68.5	84.7	77.8	**10**
KNN	**91.4**	81.5	52.5	55.8	87.5	75.7	89.0	76.2	**11**
K star	81.9	**90.7**	89.4	53.5	64.3	68.8	70.7	74.2	**12**
M5P	85.1	58.7	83.2	60.0	75.2	73.4	79.6	73.6	**13**
J48	80.3	69.7	78.4	48.7	70.6	68.4	76.7	70.4	**14**
Random Tree	83.8	71.7	76.2	52.9	69.3	60.8	75.0	70.0	**15**
ASC	76.8	70.0	77.9	43.1	72.0	63.1	76.7	68.5	**16**
LDA	69.7	52.0	89.1	70.8	62.8	69.7	52.6	66.7	**17**

Table 2-2 Performance measures of data mining algorithm at different levels of significance over T1D dataset

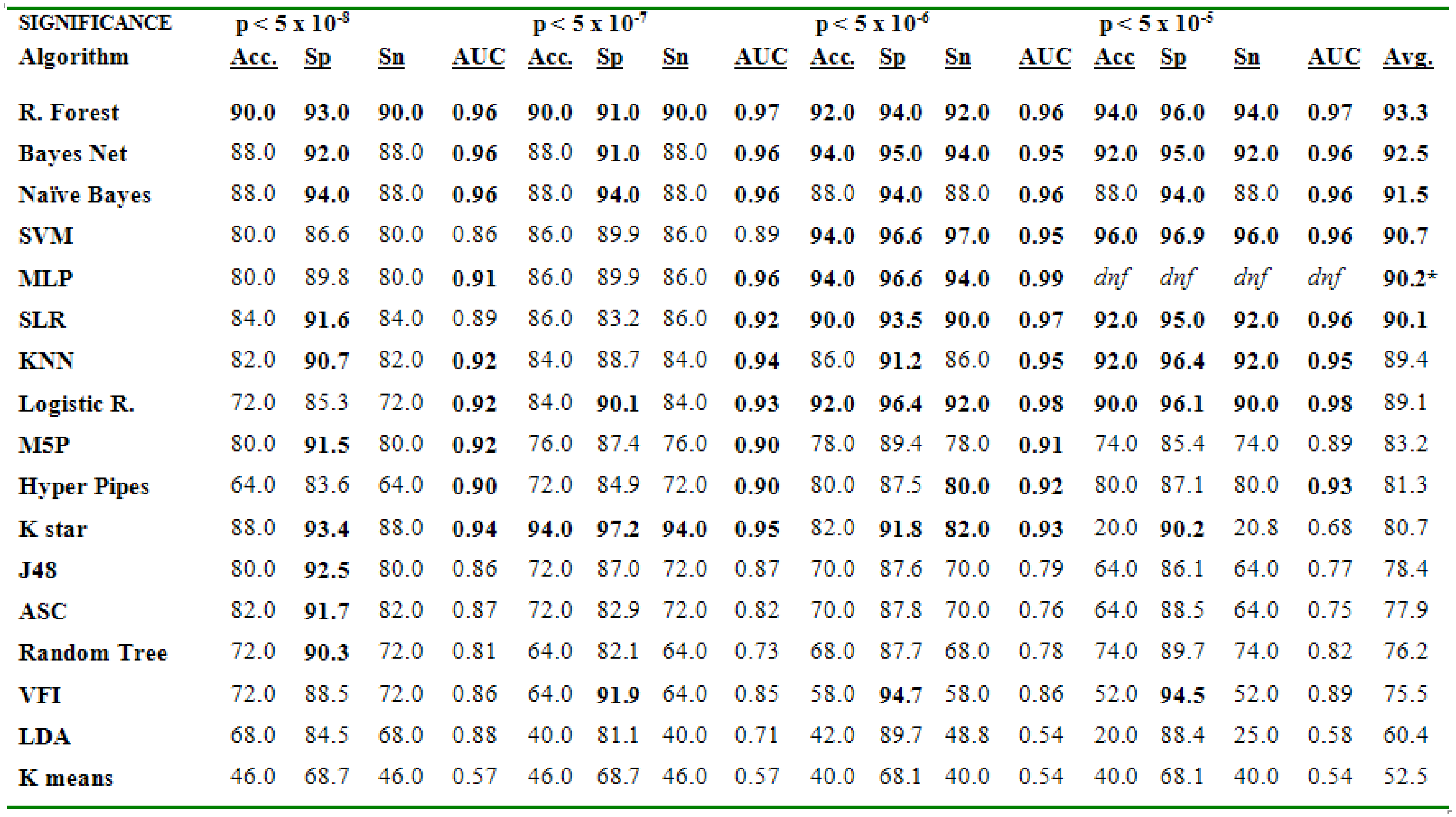


SIGNIFICANCE	$p < 5 \times 10^{-13}$				$p < 5 \times 10^{-10}$				$p < 5 \times 10^{-7}$				$p < 5 \times 10^{-4}$				
Algorithm	**Acc.**	**Sp**	**Sn**	**AUC**	**Acc.**	**Sp**	**Sn**	**AUC**	**Acc.**	**Sp**	**Sn**	**AUC**	**Acc**	**Sp**	**Sn**	**AUC**	**Avg.**
SLR	87.5	85.0	89.7	**0.93**	**92.5**	**90.2**	**94.9**	**0.97**	**92.5**	**92.0**	**92.0**	**0.96**	**92.5**	**90.0**	**94.9**	**0.96**	**92.2**
Naïve Bayes	**90.0**	85.4	**95.0**	**0.97**	**91.3**	**90.2**	**92.3**	**0.98**	**92.5**	**90.2**	**95.0**	**0.96**	89.0	85.4	**92.3**	**0.92**	**92.0**
SVM	88.8	82.9	**94.9**	0.89	**90.0**	82.9	**97.4**	**0.90**	**93.8**	**90.2**	**97.4**	**0.93**	**93.8**	**92.7**	**94.9**	**0.94**	**91.6**
R. Forest	87.5	87.8	87.2	**0.96**	**92.5**	**90.2**	**94.9**	**0.97**	**91.5**	87.8	**94.9**	**0.97**	88.8	85.4	**92.3**	**0.94**	**91.5**
KNN	**92.5**	**90.2**	**94.9**	**0.95**	**95.0**	**92.7**	**97.4**	**0.96**	**90.0**	85.4	**94.9**	**0.93**	85.0	80.5	89.7	**0.90**	**91.4**
Logistic. R	86.3	87.8	84.6	0.82	**92.5**	**90.2**	**94.9**	**0.97**	**92.5**	**92.7**	**97.4**	**0.97**	87.5	**92.7**	82.1	**0.92**	**90.6**
VFI	87.5	82.9	**92.3**	**0.95**	**92.5**	**90.2**	**94.9**	**0.97**	88.8	85.4	**92.3**	**0.95**	87.5	82.9	**92.3**	**0.92**	**90.5**
Bayes Net	**91.3**	**90.2**	**92.3**	**0.97**	**90.0**	85.4	**94.9**	**0.98**	**90.0**	85.4	**94.9**	**0.95**	83.8	78.0	89.7	0.89	**90.3**
MLP	80.0	80.5	79.5	0.89	**91.3**	**90.2**	**92.3**	**0.98**	**93.8**	**90.2**	**97.4**	**0.99**	*dnf*	*dnf*	*dnf*	*dnf*	**90.1***
Hyper Pipes	87.5	**90.2**	84.6	**0.96**	**91.3**	**90.2**	**92.3**	**0.97**	**90.0**	**90.2**	89.7	**0.95**	83.8	**92.7**	74.4	**0.92**	89.8
K-means	**91.3**	82.9	**100**	**0.92**	**90.0**	82.9	**97.4**	**0.90**	86.3	78.0	**94.9**	0.87	85.0	75.6	**94.9**	0.85	88.3
M5P	88.8	85.4	**92.3**	**0.94**	85.0	80.5	89.7	**0.94**	81.3	78.0	84.6	0.87	78.8	73.2	84.6	0.85	85.1
Random Tree	85.0	87.8	82.1	0.85	78.8	75.6	82.1	0.79	87.5	85.4	89.7	0.88	83.8	85.4	82.1	0.84	83.8
K star	87.5	87.8	87.2	**0.96**	**91.3**	85.4	**97.4**	0.98	**90.0**	85.4	**94.9**	**0.97**	53.8	**100**	5.1	0.54	81.9
J48	86.3	85.4	87.2	0.79	81.3	82.9	79.5	0.83	78.8	82.9	74.4	0.72	80.0	85.4	74.4	0.73	80.3
ASC	86.3	85.4	87.2	0.79	80.0	82.9	76.9	0.80	80.0	87.8	71.8	0.78	66.3	80.5	51.3	0.55	76.8
LDA	88.8	82.9	**94.9**	**0.96**	**91.3**	85.4	**97.4**	0.95	40.0	**96.7**	15.8	0.68	21.3	**94.4**	0.0	0.48	69.7

Table 2-3: Performance measures of data mining algorithm at different levels of significance over Alzheimer's dataset

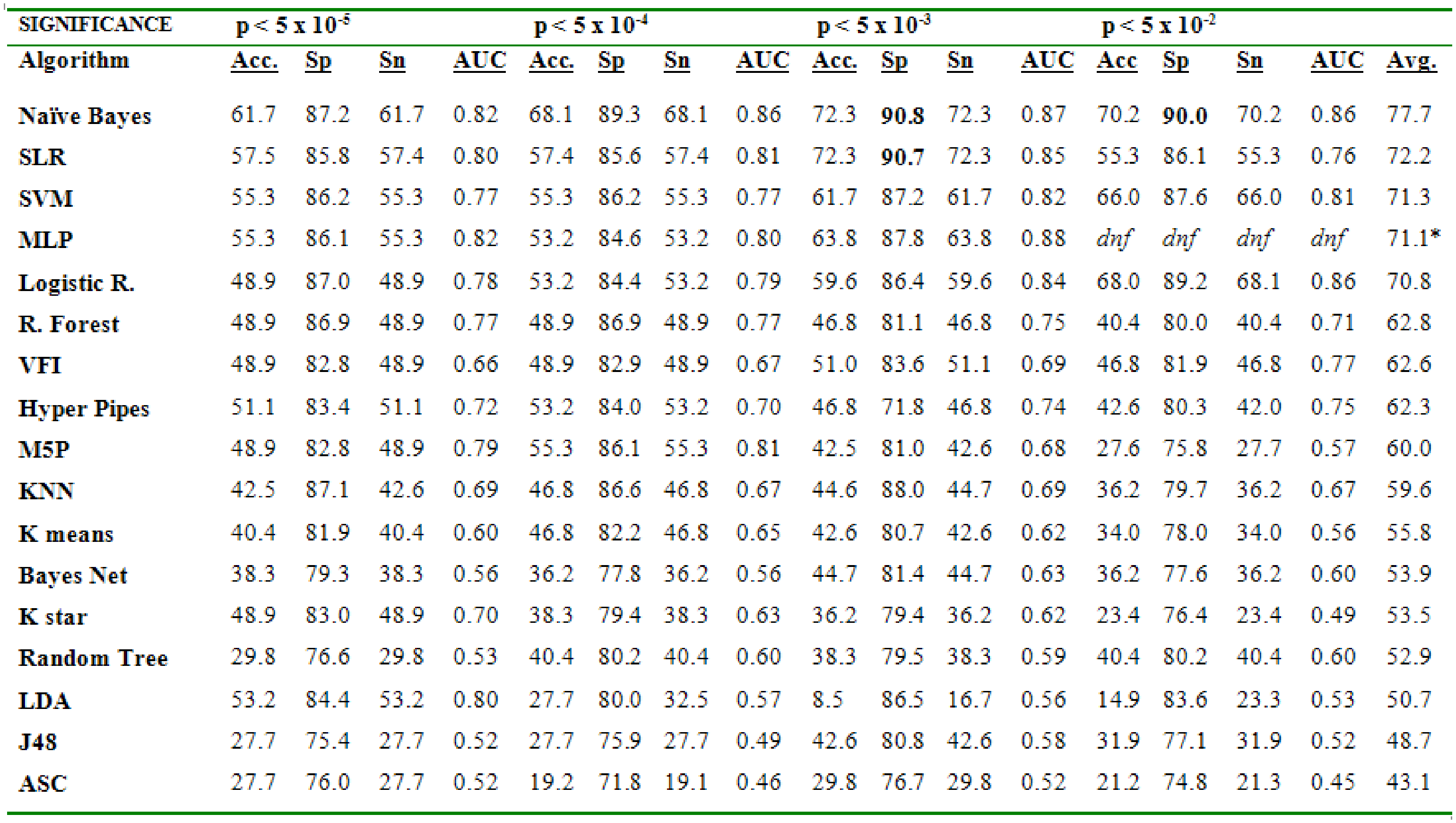
SIGNIFICANCE	$p < 5 \times 10^{-5}$				$p < 5 \times 10^{-4}$				$p < 5 \times 10^{-3}$				$p < 5 \times 10^{-2}$				
Algorithm	**Acc.**	**Sp**	**Sn**	**AUC**	**Acc.**	**Sp**	**Sn**	**AUC**	**Acc.**	**Sp**	**Sn**	**AUC**	**Acc**	**Sp**	**Sn**	**AUC**	**Avg.**
Naïve Bayes	**100**	**100**	**100**	**1.00**	**91.3**	82.0	**100**	**0.96**	**91.3**	82.0	**100**	**0.96**	86.5	**91.0**	84.0	**0.94**	**93.4**
Logistic. R	**95.0**	**90.0**	**100**	**0.99**	**95.7**	**90.0**	**100**	**0.97**	**91.3**	**90.0**	**91.7**	**0.90**	**91.3**	**90.0**	**91.7**	**0.90**	**93.3**
MLP	**91.3**	**90.9**	**91.7**	**0.97**	**95.6**	**90.9**	**100**	**0.97**	87.0	**90.9**	83.3	**0.97**	*dnf*	*dnf*	*dnf*	*dnf*	**92.7***
VFI	**91.3**	**90.9**	**91.7**	0.87	**95.7**	**90.9**	**100**	**0.92**	**91.3**	81.8	**100**	0.89	**91.3**	81.8	**100**	**1.00**	**92.2**
KNN	**91.3**	**90.9**	**91.7**	**0.93**	**95.6**	**90.9**	**100**	**0.93**	86.9	**90.9**	83.3	**0.95**	**91.3**	**90.9**	**91.7**	**0.92**	**91.8**
K-means	82.6	**100**	66.7	0.83	**91.3**	**90.9**	**100**	**0.91**	**95.7**	**90.9**	**100**	**0.96**	**91.3**	81.8	**100**	**0.90**	**90.7**
Hyper Pipes	**91.3**	81.8	**100**	**0.98**	**95.7**	**90.9**	**100**	**0.97**	**91.3**	81.8	**100**	**0.95**	73.9	81.8	66.7	**0.90**	89.7
SVM	87.0	**90.9**	83.3	0.87	**95.7**	**90.9**	**100**	**0.95**	82.6	81.8	83.3	0.83	87.0	81.8	**91.7**	0.87	88.0
Bayes Net	**91.3**	81.8	**100**	**0.96**	**91.3**	**90.9**	**91.7**	**0.95**	87.0	81.8	**91.7**	0.86	78.3	81.8	75.0	0.84	87.7
R. Forest	86.9	81.8	**91.7**	**0.94**	82.6	81.8	83.3	**0.93**	73.9	72.7	75.0	0.89	72.6	81.8	75.0	0.84	82.4
K star	**95.7**	**90.9**	**100**	**0.98**	**91.3**	**90.9**	**91.7**	**0.94**	78.2	81.8	75.0	0.86	56.5	18.2	**91.7**	0.64	81.5
SLR	86.9	81.8	**91.7**	**0.96**	73.9	72.7	75.0	0.82	60.9	63.6	58.3	0.80	52.2	54.5	50.0	0.69	71.8
Random Tree	78.3	72.7	83.3	0.78	60.9	54.5	66.7	0.61	73.9	63.6	83.3	0.74	73.9	81.8	66.7	0.74	71.7
ASC	73.9	63.6	83.3	0.61	68.9	63.6	58.3	0.56	73.9	81.8	66.7	0.75	78.2	63.9	**91.7**	0.61	70.0
J48	73.9	63.6	83.3	0.61	60.9	63.6	58.3	0.56	73.9	81.8	70.0	0.75	78.3	63.6	**91.7**	0.61	69.7
M5P	69.5	54.5	83.3	0.80	52.2	45.5	58.3	0.73	56.5	45.5	66.7	0.43	56.5	36.4	75.0	0.44	58.7
LDA	69.6	72.7	66.7	0.81	34.8	40.0	75.0	0.45	34.8	0.0	**100**	0.30	30.4	**100**	0.0	0.52	52.0

Table 2-4: Performance measures of data mining algorithm at different levels of significance over Antibodies dataset



SIGNIFICANCE	**$p < 5 \times 10^{-8}$**				**$p < 5 \times 10^{-7}$**				**$p < 5 \times 10^{-6}$**				**$p < 5 \times 10^{-5}$**				
Algorithm	**Acc.**	**Sp**	**Sn**	**AUC**	**Acc.**	**Sp**	**Sn**	**AUC**	**Acc.**	**Sp**	**Sn**	**AUC**	**Acc**	**Sp**	**Sn**	**AUC**	**Avg.**
R. Forest	**90.0**	**93.0**	**90.0**	**0.96**	**90.0**	**91.0**	**90.0**	**0.97**	**92.0**	**94.0**	**92.0**	**0.96**	**94.0**	**96.0**	**94.0**	**0.97**	**93.3**
Bayes Net	88.0	**92.0**	88.0	**0.96**	88.0	**91.0**	88.0	**0.96**	**94.0**	**95.0**	**94.0**	**0.95**	**92.0**	**95.0**	**92.0**	**0.96**	**92.5**
Naïve Bayes	88.0	**94.0**	88.0	**0.96**	88.0	**94.0**	88.0	**0.96**	88.0	**94.0**	88.0	**0.96**	88.0	**94.0**	88.0	**0.96**	**91.5**
SVM	80.0	86.6	80.0	0.86	86.0	89.9	86.0	0.89	**94.0**	**96.6**	**97.0**	**0.95**	**96.0**	**96.9**	**96.0**	**0.96**	**90.7**
MLP	80.0	89.8	80.0	**0.91**	86.0	89.9	86.0	**0.96**	**94.0**	**96.6**	**94.0**	**0.99**	*dnf*	*dnf*	*dnf*	*dnf*	**90.2***
SLR	84.0	**91.6**	84.0	0.89	86.0	83.2	86.0	**0.92**	**90.0**	**93.5**	**90.0**	**0.97**	**92.0**	**95.0**	**92.0**	**0.96**	**90.1**
KNN	82.0	**90.7**	82.0	**0.92**	84.0	88.7	84.0	**0.94**	86.0	**91.2**	86.0	**0.95**	**92.0**	**96.4**	**92.0**	**0.95**	89.4
Logistic R.	72.0	85.3	72.0	**0.92**	84.0	**90.1**	84.0	**0.93**	**92.0**	**96.4**	**92.0**	**0.98**	**90.0**	**96.1**	**90.0**	**0.98**	89.1
M5P	80.0	**91.5**	80.0	**0.92**	76.0	87.4	76.0	**0.90**	78.0	89.4	78.0	**0.91**	74.0	85.4	74.0	0.89	83.2
Hyper Pipes	64.0	83.6	64.0	**0.90**	72.0	84.9	72.0	**0.90**	80.0	87.5	**80.0**	**0.92**	80.0	87.1	80.0	**0.93**	81.3
K star	88.0	**93.4**	88.0	**0.94**	**94.0**	**97.2**	**94.0**	**0.95**	82.0	**91.8**	**82.0**	**0.93**	20.0	**90.2**	20.8	0.68	80.7
J48	80.0	**92.5**	80.0	0.86	72.0	87.0	72.0	0.87	70.0	87.6	70.0	0.79	64.0	86.1	64.0	0.77	78.4
ASC	82.0	**91.7**	82.0	0.87	72.0	82.9	72.0	0.82	70.0	87.8	70.0	0.76	64.0	88.5	64.0	0.75	77.9
Random Tree	72.0	**90.3**	72.0	0.81	64.0	82.1	64.0	0.73	68.0	87.7	68.0	0.78	74.0	89.7	74.0	0.82	76.2
VFI	72.0	88.5	72.0	0.86	64.0	**91.9**	64.0	0.85	58.0	**94.7**	58.0	0.86	52.0	**94.5**	52.0	0.89	75.5
LDA	68.0	84.5	68.0	0.88	40.0	81.1	40.0	0.71	42.0	89.7	48.8	0.54	20.0	88.4	25.0	0.58	60.4
K means	46.0	68.7	46.0	0.57	46.0	68.7	46.0	0.57	40.0	68.1	40.0	0.54	40.0	68.1	40.0	0.54	52.5

Table 2-5: Performance measures of data mining algorithm at different levels of significance over Asthma dataset 4 classes.

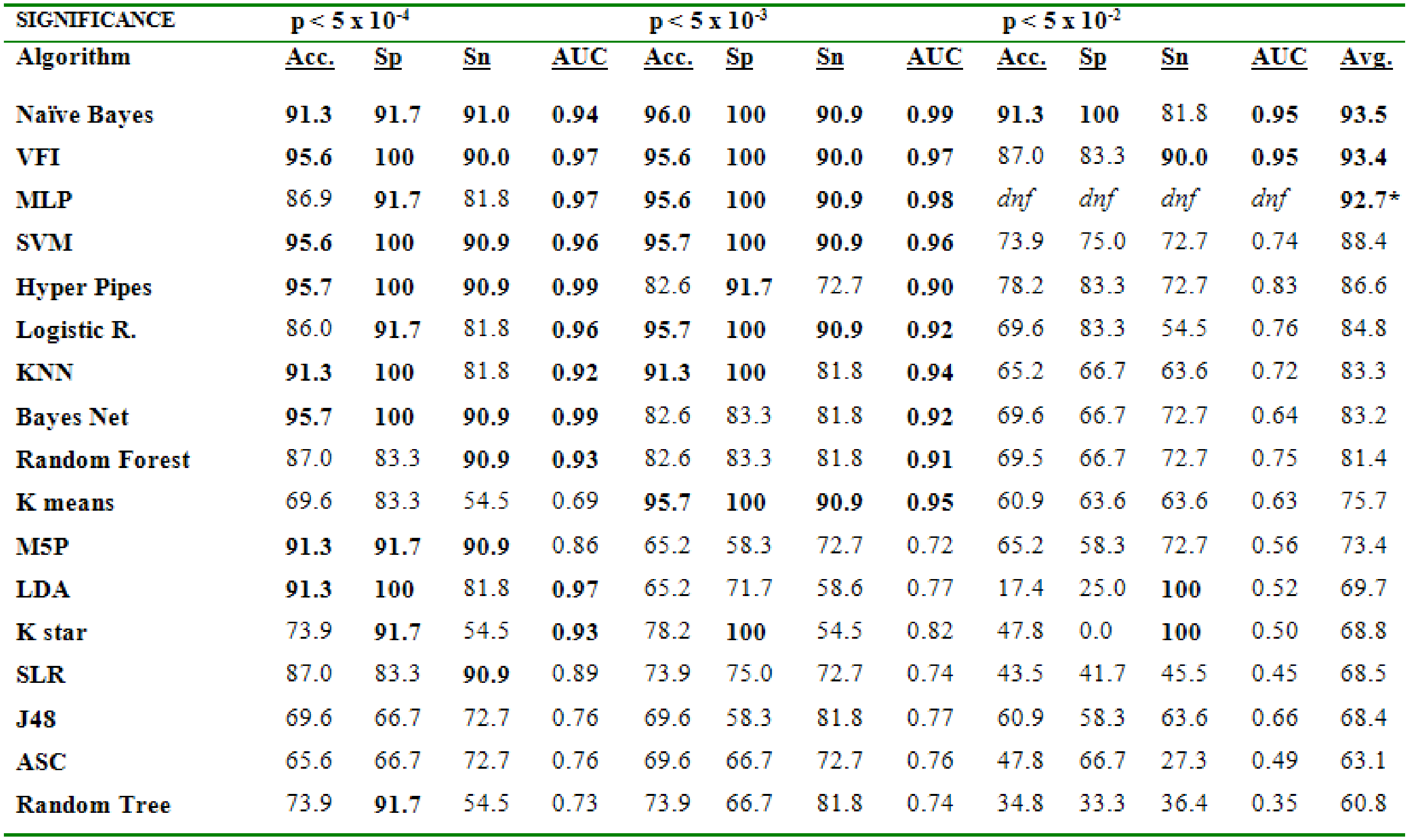


SIGNIFICANCE	$p < 5 x 10^{-5}$				$p < 5 x 10^{-4}$				$p < 5 x 10^{-3}$				$p < 5 x 10^{-2}$				
Algorithm	Acc.	Sp	Sn	AUC	Acc.	Sp	Sn	AUC	Acc.	Sp	Sn	AUC	Acc	Sp	Sn	AUC	Avg.
Naïve Bayes	61.7	87.2	61.7	0.82	68.1	89.3	68.1	0.86	72.3	**90.8**	72.3	0.87	70.2	**90.0**	70.2	0.86	77.7
SLR	57.5	85.8	57.4	0.80	57.4	85.6	57.4	0.81	72.3	**90.7**	72.3	0.85	55.3	86.1	55.3	0.76	72.2
SVM	55.3	86.2	55.3	0.77	55.3	86.2	55.3	0.77	61.7	87.2	61.7	0.82	66.0	87.6	66.0	0.81	71.3
MLP	55.3	86.1	55.3	0.82	53.2	84.6	53.2	0.80	63.8	87.8	63.8	0.88	*dnf*	*dnf*	*dnf*	*dnf*	71.1*
Logistic R.	48.9	87.0	48.9	0.78	53.2	84.4	53.2	0.79	59.6	86.4	59.6	0.84	68.0	89.2	68.1	0.86	70.8
R. Forest	48.9	86.9	48.9	0.77	48.9	86.9	48.9	0.77	46.8	81.1	46.8	0.75	40.4	80.0	40.4	0.71	62.8
VFI	48.9	82.8	48.9	0.66	48.9	82.9	48.9	0.67	51.0	83.6	51.1	0.69	46.8	81.9	46.8	0.77	62.6
Hyper Pipes	51.1	83.4	51.1	0.72	53.2	84.0	53.2	0.70	46.8	71.8	46.8	0.74	42.6	80.3	42.0	0.75	62.3
M5P	48.9	82.8	48.9	0.79	55.3	86.1	55.3	0.81	42.5	81.0	42.6	0.68	27.6	75.8	27.7	0.57	60.0
KNN	42.5	87.1	42.6	0.69	46.8	86.6	46.8	0.67	44.6	88.0	44.7	0.69	36.2	79.7	36.2	0.67	59.6
K means	40.4	81.9	40.4	0.60	46.8	82.2	46.8	0.65	42.6	80.7	42.6	0.62	34.0	78.0	34.0	0.56	55.8
Bayes Net	38.3	79.3	38.3	0.56	36.2	77.8	36.2	0.56	44.7	81.4	44.7	0.63	36.2	77.6	36.2	0.60	53.9
K star	48.9	83.0	48.9	0.70	38.3	79.4	38.3	0.63	36.2	79.4	36.2	0.62	23.4	76.4	23.4	0.49	53.5
Random Tree	29.8	76.6	29.8	0.53	40.4	80.2	40.4	0.60	38.3	79.5	38.3	0.59	40.4	80.2	40.4	0.60	52.9
LDA	53.2	84.4	53.2	0.80	27.7	80.0	32.5	0.57	8.5	86.5	16.7	0.56	14.9	83.6	23.3	0.53	50.7
J48	27.7	75.4	27.7	0.52	27.7	75.9	27.7	0.49	42.6	80.8	42.6	0.58	31.9	77.1	31.9	0.52	48.7
ASC	27.7	76.0	27.7	0.52	19.2	71.8	19.1	0.46	29.8	76.7	29.8	0.52	21.2	74.8	21.3	0.45	43.1

Table 2-6: Performance measures of data mining algorithm at different levels of significance on A & B conditions

SIGNIFICANCE	$p < 5 x 10^{-4}$				$p < 5 x 10^{-3}$				$p < 5 x 10^{-2}$				
Algorithm	**Acc.**	**Sp**	**Sn**	**AUC**	**Acc.**	**Sp**	**Sn**	**AUC**	**Acc.**	**Sp**	**Sn**	**AUC**	**Avg.**
Naïve Bayes	87.5	83.3	**91.7**	0.84	**91.7**	83.3	**100**	**0.97**	**91.7**	83.3	**100**	**0.96**	**90.8**
VFI	79.2	75.0	83.3	**0.93**	**91.7**	83.3	**100**	**0.95**	87.5	75.0	**100**	**0.90**	87.7
K means	87.5	83.3	**91.7**	0.88	**91.7**	83.3	**100**	**0.92**	83.3	75.0	**91.7**	0.83	87.5
SVM	83.3	83.3	83.3	0.83	87.5	**91.7**	83.3	0.87	87.5	83.3	**91.7**	0.88	86.1
MLP	79.2	83.3	75.0	0.70	**91.7**	**91.7**	**91.7**	**0.95**	*dnf*	*dnf*	*dnf*	*dnf*	84.7*
Hyper Pipes	83.3	75.0	**91.7**	**0.91**	83.3	83.3	83.3	**0.93**	70.8	83.3	58.3	0.88	82.0
Logistic R.	66.7	83.3	50.0	0.76	**95.8**	**91.7**	**100**	**0.92**	79.2	83.3	75.0	0.85	81.5
Random Forest	79.2	83.3	75.0	**0.91**	79.2	75.0	83.3	0.86	79.2	75.0	83.3	0.78	80.6
Bayes Net	83.3	75.0	**91.7**	0.87	83.3	83.3	83.3	0.83	75.0	75.0	75.0	0.67	80.2
KNN	75.0	83.3	66.7	0.85	75.0	**91.7**	58.3	**0.90**	75.0	**91.7**	58.3	0.84	77.8
M5P	75.0	83.3	66.7	0.74	75.0	75.0	75.0	0.79	75.0	75.0	75.0	0.74	75.2
ASC	62.5	66.7	58.3	0.65	79.2	83.3	75.0	0.85	70.8	75.0	66.7	0.76	72.0
J48	62.5	66.7	58.3	0.65	79.2	83.3	75.0	0.85	66.7	75.0	58.3	0.72	70.6
Random Tree	70.8	75.0	66.7	0.70	70.8	75.0	66.7	0.70	66.7	66.7	66.7	0.67	69.3
SLR	70.8	75.0	66.7	0.80	66.7	75.0	58.3	0.77	50.0	50.0	50.0	0.60	65.0
K star	66.7	**91.7**	41.7	0.83	58.3	**100**	46.7	0.83	50.0	0.0	**100**	0.50	64.3
LDA	79.2	83.3	75.0	0.84	61.2	64.5	54.5	0.52	29.2	14.3	**100**	0.56	62.8

Table 2-7: Performance measures of data mining algorithm at different levels of significance on A & C conditions



SIGNIFICANCE	$p < 5 \text{ x } 10^{-4}$				$p < 5 \text{ x } 10^{-3}$				$p < 5 \text{ x } 10^{-2}$				
Algorithm	**Acc.**	**Sp**	**Sn**	**AUC**	**Acc.**	**Sp**	**Sn**	**AUC**	**Acc.**	**Sp**	**Sn**	**AUC**	**Avg.**
Naïve Bayes	**91.3**	**91.7**	**91.0**	**0.94**	**96.0**	**100**	**90.9**	**0.99**	**91.3**	**100**	81.8	**0.95**	**93.5**
VFI	**95.6**	**100**	**90.0**	**0.97**	**95.6**	**100**	**90.0**	**0.97**	87.0	83.3	**90.0**	**0.95**	**93.4**
MLP	86.9	**91.7**	81.8	**0.97**	**95.6**	**100**	**90.9**	**0.98**	*dnf*	*dnf*	*dnf*	*dnf*	**92.7***
SVM	**95.6**	**100**	**90.9**	**0.96**	**95.7**	**100**	**90.9**	**0.96**	73.9	75.0	72.7	0.74	88.4
Hyper Pipes	**95.7**	**100**	**90.9**	**0.99**	82.6	**91.7**	72.7	**0.90**	78.2	83.3	72.7	0.83	86.6
Logistic R.	86.0	**91.7**	81.8	**0.96**	**95.7**	**100**	**90.9**	**0.92**	69.6	83.3	54.5	0.76	84.8
KNN	**91.3**	**100**	81.8	**0.92**	**91.3**	**100**	81.8	**0.94**	65.2	66.7	63.6	0.72	83.3
Bayes Net	**95.7**	**100**	**90.9**	**0.99**	82.6	83.3	81.8	**0.92**	69.6	66.7	72.7	0.64	83.2
Random Forest	87.0	83.3	**90.9**	**0.93**	82.6	83.3	81.8	**0.91**	69.5	66.7	72.7	0.75	81.4
K means	69.6	83.3	54.5	0.69	**95.7**	**100**	**90.9**	**0.95**	60.9	63.6	63.6	0.63	75.7
M5P	**91.3**	**91.7**	**90.9**	0.86	65.2	58.3	72.7	0.72	65.2	58.3	72.7	0.56	73.4
LDA	**91.3**	**100**	81.8	**0.97**	65.2	71.7	58.6	0.77	17.4	25.0	**100**	0.52	69.7
K star	73.9	**91.7**	54.5	**0.93**	78.2	**100**	54.5	0.82	47.8	0.0	**100**	0.50	68.8
SLR	87.0	83.3	**90.9**	0.89	73.9	75.0	72.7	0.74	43.5	41.7	45.5	0.45	68.5
J48	69.6	66.7	72.7	0.76	69.6	58.3	81.8	0.77	60.9	58.3	63.6	0.66	68.4
ASC	65.6	66.7	72.7	0.76	69.6	66.7	72.7	0.76	47.8	66.7	27.3	0.49	63.1
Random Tree	73.9	**91.7**	54.5	0.73	73.9	66.7	81.8	0.74	34.8	33.3	36.4	0.35	60.8

Table 2-8: Performance measures of data mining algorithm at different levels of significance on B & D conditions

SIGNIFICANCE	$p < 5 \times 10^{-4}$				$p < 5 \times 10^{-3}$				$p < 5 \times 10^{-2}$				
Algorithm	Acc.	Sp	Sn	AUC	Acc.	Sp	Sn	AUC	Acc.	Sp	Sn	AUC	Avg.
Naïve Bayes	**91.7**	**100**	83.3	**0.95**	**91.7**	**91.7**	**91.7**	**0.92**	**95.8**	**91.7**	**100**	**0.98**	**93.6**
SVM	**91.7**	**100**	83.3	**0.92**	**91.7**	**91.7**	**91.7**	**0.92**	**95.8**	**100**	**91.7**	**0.96**	**93.1**
VFI	87.5	**100**	75.0	**0.93**	**91.7**	**100**	83.3	**0.94**	**95.8**	**100**	**91.7**	**1.00**	**92.7**
Logistic R.	79.1	83.3	75.0	**0.92**	**100**	**100**	**100**	**1.00**	87.5	**91.7**	83.3	**0.97**	**90.7**
MLP	87.5	**91.7**	83.3	**0.94**	87.5	83.3	**91.7**	**0.96**	*dnf*	*dnf*	*dnf*	*dnf*	89.3*
K means	87.5	**91.7**	83.3	0.88	**91.4**	**91.7**	**91.7**	**0.92**	87.5	83.3	**91.7**	0.88	89.0
Hyper Pipes	87.5	83.3	**91.7**	0.89	**91.7**	**91.7**	**91.7**	0.87	83.3	75.0	**91.7**	**0.90**	87.8
Bayes Net	83.3	83.3	83.3	0.89	87.5	**91.7**	83.3	0.86	83.3	83.3	83.3	0.84	85.1
SLR	83.3	83.3	83.3	0.88	79.2	66.7	**91.7**	**0.90**	87.5	**100**	75.0	0.89	84.7
KNN	79.2	75.0	83.3	0.80	83.3	83.3	83.3	0.83	87.5	**91.7**	83.3	**0.90**	83.6
Random Forest	83.3	83.3	83.3	0.83	79.2	83.3	75.0	0.84	79.2	83.3	75.0	0.81	81.1
M5P	87.5	**91.7**	83.3	0.88	79.2	83.3	75.0	0.73	75.0	83.3	66.7	0.69	79.6
ASC	**91.7**	**100**	83.3	0.83	75.0	83.3	66.7	0.61	70.8	75.0	66.7	0.64	76.7
J48	**91.7**	**100**	83.3	0.83	75.0	83.3	66.7	0.61	70.8	75.0	66.7	0.64	76.7
Random Tree	83.3	**91.7**	75.0	0.83	70.8	66.7	75.0	0.71	70.8	66.7	75.0	0.71	75.0
K star	70.8	66.7	75.0	0.83	79.2	75.0	83.3	0.82	58.3	**100**	16.7	0.58	70.7
LDA	62.5	72.3	60.9	0.75	50.0	65.0	48.0	0.71	20.8	42.6	18.6	0.45	52.6

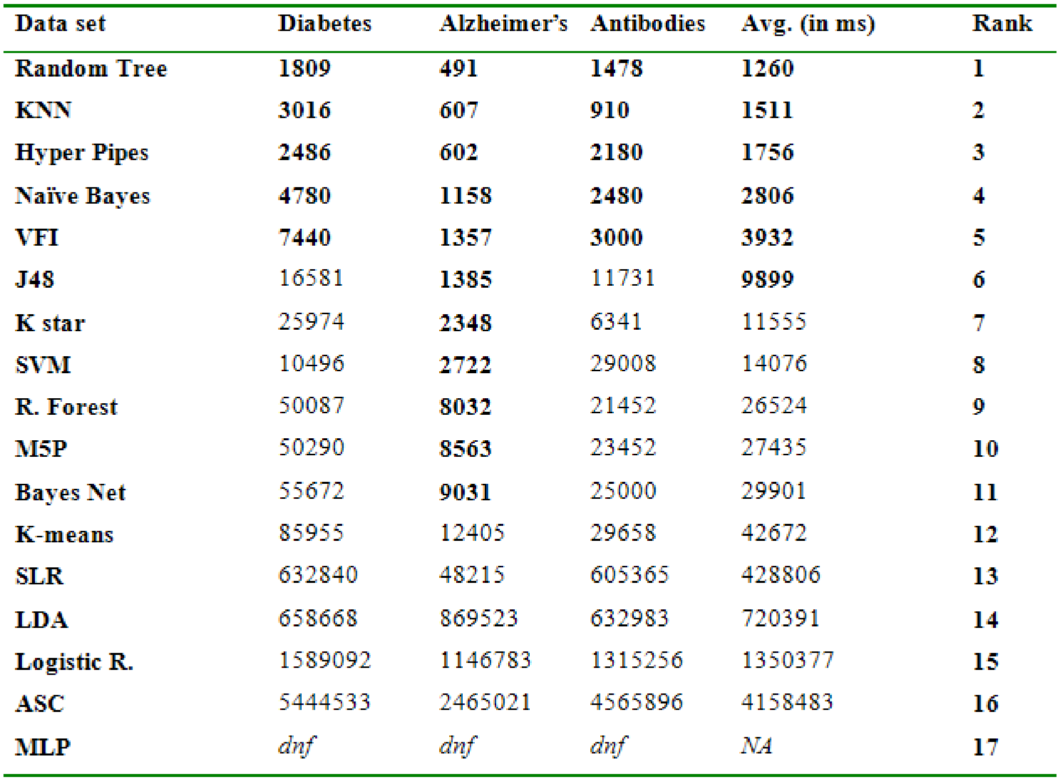
For the Asthma dataset, we considered all conditions A-D together, then performed the pair-wise comparisons of condition A and B, condition A and C, and condition B and D at three different levels of significance. Measures >90% are marked in bold. For the diabetes dataset, 9 algorithms achieved >90% score. For Alzheimer's and the Antibodies dataset, 6 algorithms achieved >90% score. Naïve Bayes scored 100% in all 4 measures at the first level of significance in the Alzheimer's dataset and scored 91.5% average score on the Antibodies dataset. For the Asthma datasets, the highest score was <80%. Only Naïve Bayes had >90% specificity for more than one level of significance. For two conditions in Asthma datasets, Naïve Bayes and VFI scored >90% average score. Acc: Accuracy, Sp: Specificity, Sn: Sensitivity, AUC: Area under ROC curve, Avg: Average score in % for each algorithms, *dnf:* "Did Not Finish", * denotes Avg. from 3 significance levels. Measures >90% are marked in bold.

2.4.3 COMPARATIVE ANALYSIS OF WORST TIME PERFORMANCE OF CLASSIFICATION ALGORITHMS OVER DATA SETS

The amount of time taken by each algorithm to build the model and perform cross validation was measured.

Table 2-9 shows the time in milliseconds for each algorithm at the lowest level of significance when the number of peptides nears 1000. Random Tree was the fastest, at ~1000 milliseconds (average) to complete the task, while MLP was the worst which did not finish due to high memory requirements. Random tree, Hyper Pipes, Naïve Bayes, VFI and KNN were the five fastest algorithms; each took less than ~4000 milliseconds to complete classification of >1,000 peptides. Logistic Regression and Attribute Selected Classifier, MLP were among the slowest algorithms taking more than 20 minutes to perform classification of >1,000 peptides. The absolute ranking for every algorithm was consistent per dataset; only three datasets have been considered to measure time performance. Random Tree, KNN, Hyper Pipes and VFI were among the fastest. MLP were among the slowest with dnf: “Did not finish”. Time measurements less than 10 seconds are marked in bold.

Table 2-9: Worst case time performance (in ms) of classification algorithms



Data set	Diabetes	Alzheimer’s	Antibodies	Avg. (in ms)	Rank
Random Tree	**1809**	**491**	**1478**	**1260**	1
KNN	**3016**	**607**	**910**	**1511**	2
Hyper Pipes	**2486**	**602**	**2180**	**1756**	3
Naïve Bayes	**4780**	**1158**	**2480**	**2806**	4
VFI	**7440**	**1357**	**3000**	**3932**	5
J48	16581	**1385**	11731	**9899**	6
K star	25974	**2348**	6341	11555	7
SVM	10496	**2722**	29008	14076	8
R. Forest	50087	**8032**	21452	26524	9
M5P	50290	**8563**	23452	27435	10
Bayes Net	55672	**9031**	25000	29901	11
K-means	85955	12405	29658	42672	12
SLR	632840	48215	605365	428806	13
LDA	658668	869523	632983	720391	14
Logistic R.	1589092	1146783	1315256	1350377	15
ASC	5444533	2465021	4565896	4158483	16
MLP	*dnf*	*dnf*	*dnf*	*NA*	17

2.4.4 COMPARATIVE ANALYSIS OF TIME PERFORMANCE OF CLASSIFICATION ALGORITHMS AT DIFFERENT LEVELS OF SIGNIFICANCE OVER THREE DATA SETS

For each level of significance, time was measured for each algorithm to build the model and for cross validation. At the highest level of significance (about 10 peptides), each algorithm was fast enough to complete the task in under 25 seconds. Execution times increased as the level of significance was lowered due to the higher number of features and increased difficulty in constructing the model.

Table 2-10 shows classification algorithms time performance at various levels of significance.

Table 2-10: Time performance (in ms) of classification algorithms on datasets

	Diabetes dataset				**Alzheimer's dataset**				**Antibodies dataset**			
p value <	5x10^{-13}	5x10^{-10}	5x10^{-7}	5x10^{-4}	5x10^{-5}	5x10^{-4}	5x10^{-3}	5x10^{-2}	5x10^{-8}	5x10^{-7}	5x10^{-6}	5x10^{-5}
R. Tree	**337**	**408**	**571**	**1809**	**184**	**200**	**218**	**491**	**250**	**265**	**608**	**1478**
KNN	**265**	**333**	**585**	**3016**	**130**	**156**	**239**	**607**	**187**	**234**	**414**	**910**
Hyper Pipes	**226**	**274**	**630**	**2486**	**119**	**259**	**423**	**602**	**281**	**312**	**736**	**2180**
Naïve Bayes	**250**	**456**	**1120**	**4780**	**182**	**340**	**500**	**1158**	**265**	**362**	**892**	**2480**
VFI	**299**	**561**	**1384**	**7440**	**187**	**337**	**623**	**1357**	**280**	**368**	**1379**	**3000**
J48	**415**	**833**	**3718**	16581	**166**	**256**	**712**	**1385**	**468**	**880**	**3011**	11731
K star	**468**	**1387**	**4150**	25974	**187**	**260**	**666**	**2349**	**299**	**562**	**2340**	**6341**
SVM	**3313**	**3635**	**5304**	10496	**1054**	**1108**	**1389**	**2722**	18297	18372	23712	29009
R. Forest	**5717**	11889	18254	50087	**952**	**1852**	**4843**	**8032**	**5004**	**6749**	13848	21452
M5P	**701**	**2583**	**7717**	50290	**290**	**524**	**2324**	**8563**	**2632**	**4711**	12033	23452
Bayes Net	**718**	**2087**	**5653**	55672	**334**	**662**	**4996**	**9031**	**733**	**1140**	**3394**	25000
K means	**2618**	**6651**	11876	85955	**593**	**1123**	**7212**	12405	**850**	**908**	**3442**	29658
SLR	11215	26380	79308	632840	**1330**	**3413**	22625	48215	17389	20649	89107	605365
LDA	**683**	**1044**	**7994**	658668	**402**	**699**	35568	869523	**1512**	**2018**	17373	632983
Logistic R.	**1204**	**2592**	24687	1589092	**629**	**1651**	48659	1146783	**1654**	**9379**	255103	1315256
ASC	**864**	**3504**	32836	5444533	**518**	**1859**	36849	2465021	**1217**	**1763**	25496	4565896
MLP	23759	314076	4572305	*dnf*	**2057**	30342	2789485	*dnf*	22916	156905	3277395	*dnf*

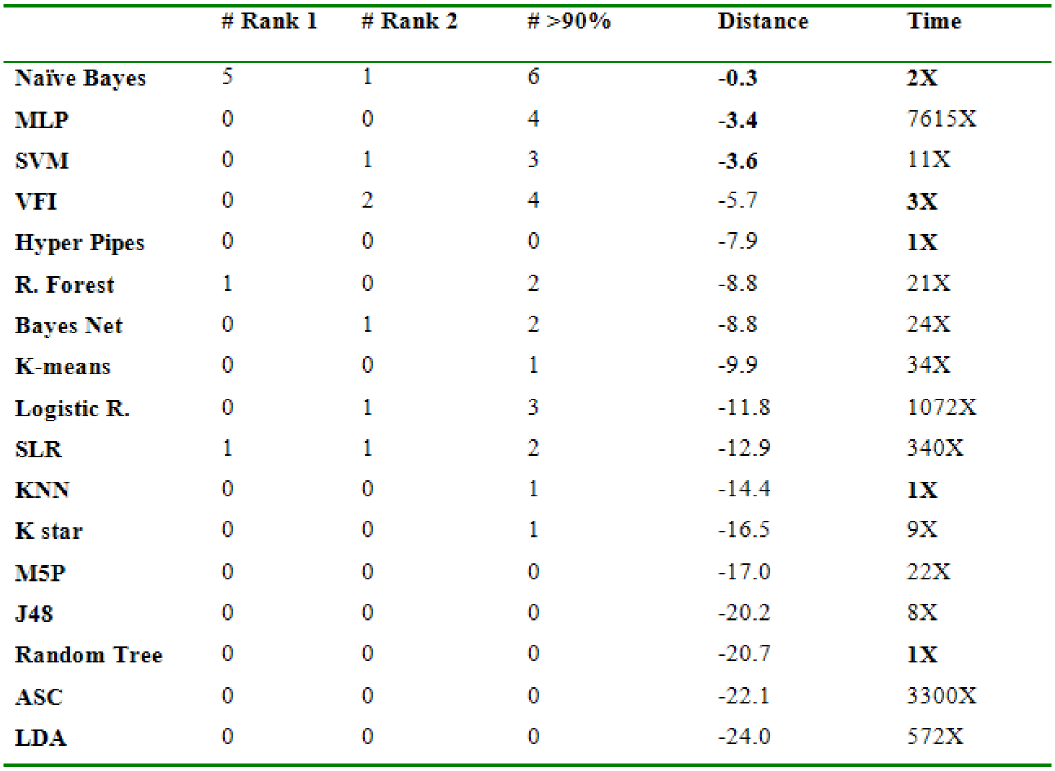
2.4.5 RESULTS SUMMARY

We have explored several disparate classifiers using a relatively new type of microarray data: immunosignaturing data. The tested algorithms come from a broad family of approaches to classify data. We chose algorithms from Bayesian, regression, trees, multivariate and meta analysis and we believe we have sampled sufficiently that the results are relevant. From

Table 2-1 we found that Naïve Bayes had a higher average performance than all other algorithms tested. Naïve Bayes achieved > 90% average for 2 classes datasets where there is a clear distinction between two classes. For the multi-class the Antibodies dataset, where there is a clear difference between different types of antibodies, Naïve Bayes scored 88% average accuracy and was ranked third, close to the 93.3 % accuracy of random forest. On the asthma dataset, containing four classes, none of the algorithms were able to achieve more than 75% accuracy. This matches the biological interpretation very well. Naïve Bayes outperformed all algorithms for speed and accuracy, achieving 77.7% average score overall. Naïve Bayes was one of the top five fastest algorithms, ~500 times faster than the logistic regression. A summary of the all algorithms performance measures and time is given in below and described in **Table 2-11**. Distance metrics have been defined to access performance measures for all algorithms compared to the highest scoring algorithm on a given dataset. #Rank 1, Rank 2: No. of times algorithm ranked 1st and 2nd on 7 datasets, #>90%: No. of times algorithm scored overall average score >90% on 7 datasets, Distance: magnitude an algorithm trails behind on average from the Rank 1 for the datasets

(5% or less distance are marked in bold). Time: performance slower with respective to fastest algorithm. Time performances slower by 5 folds to fastest algorithm are marked in bold.

Table 2-11: Summary of performance and time measures of classification algorithms



	# Rank 1	# Rank 2	# >90%	Distance	Time
Naïve Bayes	5	1	6	**-0.3**	**2X**
MLP	0	0	4	**-3.4**	7615X
SVM	0	1	3	**-3.6**	11X
VFI	0	2	4	-5.7	**3X**
Hyper Pipes	0	0	0	-7.9	**1X**
R. Forest	1	0	2	-8.8	21X
Bayes Net	0	1	2	-8.8	24X
K-means	0	0	1	-9.9	34X
Logistic R.	0	1	3	-11.8	1072X
SLR	1	1	2	-12.9	340X
KNN	0	0	1	-14.4	**1X**
K star	0	0	1	-16.5	9X
M5P	0	0	0	-17.0	22X
J48	0	0	0	-20.2	8X
Random Tree	0	0	0	-20.7	**1X**
ASC	0	0	0	-22.1	3300X
LDA	0	0	0	-24.0	572X

Summary of all classification algorithms are given below

1. **Naïve Bayes:** Naïve Bayes performed best overall with > 90% overall average score. It was always among the top 3 algorithms in all 7 comparisons. It ranked first 5 out 7 times when comparing all datasets. It was on an average just 0.3% behind the rank 1 algorithm in overall comparison. It is 2X slower than the fastest algorithm due to its mathematical properties. It would be feasible to perform large-scale

classification studies using Naïve Bayes.

2. **Multilayer Perceptron (MLP):** It ranked second with overall score of 87.3% and was very close to SVM. The overall score is biased since MLP did not finish for levels containing ~1000 peptides and hence scored was averaged from just the three levels. It was the slowest algorithm and infeasible to perform large-scale classification.
3. **Support Vector Machines (SVM):** Although it ranked third, it was not significantly different from the MLP in terms of performance measures. It was 700X faster than MLP and achieved >90% measured accuracy 3 times. Both MLP and SVM were <5% behind the rank 1 algorithm on average.
4. **VFI**: VFI ranked fourth in overall performance measures and was the among top 5 fastest algorithms due to its voting method. Four times it obtained >90% average overall accuracy and ranked 2nd twice.
5. **Hyper Pipes**: Hyper pipes ranked fifth overall in performance measures and was among the fastest of the tested algorithms, likely due to its inherently simplistic ranking method. It was <8% from first place 6 times.
6. **Random Forest**: Random forest ranked sixth in overall performance measures and performed better on datasets having multiple classes (Antibodies and Asthma). It was 21 times slower than the fastest algorithm due to bootstrapping.
7. **Bayes net**: Ranked in the middle for overall accuracy and time. It scored >90% overall measures twice. It was slower than the Naïve Bayes due to

construction of networks in the form of an acyclic graph and it is relatively inefficient compared to Naïve Bayes due to the change in network topology during assessment of probability.

8. **K means**: K-means ranked eighth in overall performance measures and was 34X slower than the fastest algorithm in time performance due to the multiple iterations required to form clusters. It performed far better for 2 classes compared to multiple classes because guaranteed convergence, scalability and linear separation boundaries are more easily maintained.
9. **Logistic Regression**: Logistic regression ranked ninth in overall accuracy. It was >90% three times. It was among the worst in time performance, being ~1000 times slower than the fastest algorithm as it needs to regress on high number of features. It is efficient for small numbers of features and sample sizes > 400.
10. **Simple Logistic**: It ranked tenth in overall performance measures and ranked first on the diabetes dataset. It ranked second in multiclass Asthma dataset. It was slow in time performance due to LogitBoost iterations.
11. **K nearest neighbors**: It performed well on the 2 classes dataset but did not perform as well for multi class datasets. It was >90% performance for only rather difficult diabetes dataset. This may be related to evenly defined but diffuse clusters related to the subtle differences between the asthma patients.
12. **K star**: It performed >90% for only the diabetes dataset and was 9 times slower than the fastest algorithm. This algorithm may also be sensitive to

the even and diffuse clusters described by this dataset.

13. **M5P**: It did not perform well on either time performance or accuracy. It never achieved >90% average score and was 22 times slower than the fastest algorithm due to formation of comprehensive linear model for every interior node of the unpruned tree.
14. **J48**: Top 5 fastest algorithm due to rapid construction of trees. It was >20% behind from the rank 1 algorithm on an average; its lower performance may possibly be due to formation of empty/insignificant branches which often leads to overtraining.
15. **Random Trees**: It was the fastest algorithm since it builds trees of height log(k) where k is the number of attributes, however it achieves poor accuracy since it performs no pruning.
16. **Attribute Selected Classifier (ASC):** One of the slowest algorithms as it had to evaluate attributes prior to classification. It underperformed in performance measures due to the C4.5 classifier limitations that prevent overtraining.
17. **Linear Discriminant Analysis (LDA):** Its performance accuracy decreased as the number of features increased due to its inability to deal with highly variant data. It was slow (>500X slower than the fastest algorithm) since it tries to optimize class distinctions but the variance covariance matrix increases dramatically as the number of features increased.

2.5 DISCUSSION

The comparisons provided in this article provide a glimpse into how existing classification algorithms handle data with intrinsically different properties than traditional microarray expression data. Immunosignaturing provides a means to quantify the dispersion of serum (or saliva) antibodies that result from disease or other immune challenge. Unlike most phage display or other panning experiments, fewer but longer random-sequence peptides are used. Rather than converging to relatively few sequences, the immunosignaturing microarray provides data on the binding affinity of *all* 10,000 peptides with high precision. Classifiers in the open-source program WEKA were used to determine whether any algorithm stood out as being particularly well suited for these data. The 17 classifiers, which were tested, are readily available and represent some of the most widely used classification methods in biology. However, they also represent classifiers that are diverse at the most fundamental levels. Tree methods, regression, and clustering are inherently different; the grouping methods are quite varied and top-down or bottom-up paradigms address data structures in substantially different ways. Given this, we present and interpret the results from our tests, which we believe will be applicable to any dataset with target-probe interactions similar to immunosignaturing microarrays.

From the comparisons above, Naïve Bayes was the superior analysis method in all aspects. Naïve Bayes assumes a feature independent model, which may account for its superior performance. It relies on the degree of correlation of the attributes in the dataset; for immunosignaturing, the number of attributes can be quite large.

In gene expression data, where genes are connected by gene regulatory networks, there is a direct and significant correlation between hub genes and dependent genes. This relationship affects the performance of Naïve Bayes by limiting its efficiency through multiple containers of similarly - connected features (Hedenfalk et al. 2001; Li, Zhang, and Ogihara 2004; Liu, Li, and Wong 2002). In peptide-antibody arrays, where the signals that arise from the peptides are multiplexed signals of many antibodies attaching to many peptides, there is no direct correlation between peptides, but there is a general trend. Moreover, there is a competition of antibodies attaching to a single peptide, which makes it difficult for multiple mimotopes to show significant correlation with each other. Thus, the 10,000 random peptides have no direct relationships to each other each contributes partially to defining the disease state. This makes the immunosignaturing technology a better fit for the assumption of strong feature independence employed by the Naïve Bayes technique, and the fact that reproducible data can be had at intensity values down to 1 standard deviation above background enables enormous numbers of informative, precise, and independent features. Presence or absence of a few high- or low-binding peptides on the microarray will not impact the binding affinity for any other peptide, since the kinetics ensures that the antibody pool is not limiting. This is important when building microarrays with >300,000 features per physical assay, as in our newest microarray. More than 90% of the peptides on either microarray demonstrate normal distribution for binding signals. This is important since feature selection methods used in this analysis (t-test and one way ANOVA) and the Naïve Bayes classifier all assume normal distribution of features.

The Naïve Bayes approach requires relatively little training data, which makes it a very good fit for the biomarker field. The sample sizes usually range from N=20-100 for the training set. Naïve Bayes has other advantages as well: it can train well on a small but high feature data set and still yield good prediction accuracy on a large test set. Any microarray with more than a few thousand probes succumbs to the issue of dimensionality. Since Naïve Bayes independently estimates each distribution instead of calculating a covariance or correlation matrix, it escapes relatively unharmed from problems of dimensionality.

The data used here for evaluating the algorithms were generated using an array with 10,000 different features, almost all of which contribute signal. We have arrays with >300,000 peptides per assay (current microarrays are available from www.peptidearraycore.com) which should provide for less sharing between peptide and antibody, effectively spreading out antibodies over the peptides with more specificity. This presumably will allow resolving antibody populations with finer detail. This expansion may require a classification method that is robust to noise, irrelevant attributes and redundancy. Naïve Bayes has an outstanding edge in this regard as it is robust to noisy data since such data points are averaged out when estimating conditional probabilities. It can also handle missing values by ignoring them during model building and classification. It is highly robust to irrelevant and redundant attributes because if Y_i is irrelevant then P (Class|Y_i) becomes uniformly distributed. This is due to that fact that the class conditional probability for X_i has no significant impact on the overall computation of posterior probability**.** Naïve Bayes will arrive at a correct classification as long as the correct classes are even slightly more predictable than the alternative. Here, class probabilities need not be estimated very well, which corresponds to the practical reality of immunosignaturing: signals are multiplexed due to competition, affinity, and other technological limitation of spotting, background and other biochemical effects that exist between antibody and mimotope.

2.5.1 TIME EFFICIENCY

As the immunosignaturing technology is increasingly used for large-scale experiments, it will result in an explosion of data. We need an algorithm that is accurate and can process enormous amounts of data with low memory overhead and fast enough for model building and evaluation. One aims for next-generation immunosignaturing microarrays is to monitor the health status of a large population on an on-going basis. The number of selected attributes will no longer be limited in such a scenario. For risk evaluation, complex patterns must be normalized against themselves at regular intervals. This time analysis would require a conditional probabilistic argument along with the capacity of accurately predicting the risk with low computational cost. The slope of Naïve Bayes on time performance scale is extremely small, allowing it to process a large number of attributes.

2.6 CONCLUSION

Immunosignaturing is a novel approach which aims to detect complex patterns of antibodies produced in acute or chronic disease. This complex pattern is obtained using random peptide microarrays where 10,000 random peptides are exposed to antibodies in sera/plasma/saliva. Antibody binding to the peptides is not one-to-one but a more complicated and multiplexed process. The quantity and appearance of this data appears numerically, distributionally, and statistically the same as gene expression microarray data, but is fundamentally quite different. The relationships between attributes and functionality of those attributes are not the same. Hence, traditional classification algorithms used in gene expression data might be suboptimal for analyzing immunosignaturing results. We investigated 17 different kinds of classification algorithm spanning Bayesian, regression, tree based approaches and meta-analysis and compared their leave-one-out cross-validated accuracy values using various numbers of features. We found that the Naïve Bayes classification algorithm outperforms the majority of the classification algorithms in classification accuracy and in time performance, which is not the case for expression microarrays (Stafford and Brun 2007). We also discussed its assumptions, simplicity, and fitness for immunosignaturing data. More than most, these data provide access to the information found in antibodies. Deconvoluting this information was a barrier to using antibodies as biomarkers. Pairing immunosignaturing with Naïve Bayes classification may open up the immune system to a more systematic analysis of disease.

2.7 COMPETING INTERESTS

US Patent Compound Arrays for Sample profiling: 61218890

US Patent ‘Naïve Bayes Classification for Immunosignaturing M12-104L

2.8 ACKNOWLEDGEMENTS

This chapter is published in BMC 2012 (doi:10.1186/1471-2105-13-139). We are thankful to Dr. Abner Notkins, NIH, National Institute of Dental and Craniofacial Research for providing T1D sample, Dr Lucus Restrepo for providing Alzheimer’s dataset. Dr Bart Legutki and Dr Rebecca Halperin for providing Antibodies dataset, University of Arizona, department of Pharmacy and Pharmacology (Serine Lau, Donata Vercelli, Marilyn Halonen) for providing Asthma samples, Peptide Microarray Core (Dr Zbigniew Cichacz) for providing Asthma datasets, Pradeep Kanwar for implementing the time function in JAVA and Valentin Dinu for invaluable discussion regarding algorithm selection. This work was supported by an Innovator Award from the DoD Breast Cancer Program to SAJ.

CHAPTER 3 APPLICATION OF IMMUNOSIGNATURE TECHNOLOGY TO RESOLVE PROFILES OF CLOSELY RELATED PANCREAS DISEASES

3.1 ABSTRACT

Immunosignaturing is a technology that allows the humoral immune response to be observed through the binding of antibodies to random sequence peptides. Profiles of the antibody repertoire produced during infection or during long-term chronic disease have proven to be informative for disease classification. An important unanswered question relative to this technology is whether different diseases that target the same organ and result in similar early phenotypes have similar or distinguishable immunosignatures. This question is of clinical relevance when considering patients who present with similar symptoms early during their disease. The pancreas is one such organ; disease that affects this organ can cause the patient both broad and acute distress, with little to distinguish the disease source. If the cause were made clear without biopsy, and could be accomplished during routine monitoring, earlier intervention could improve health. Pancreatic cancer, chronic or acute pancreatitis, diabetes mellitus, hepatitis B or C infection, and other diseases can deeply affect the function of the pancreas, complicating diagnosis. We tested the immunosignaturing platform for its ability to resolve four different diseases that target the same organ; pancreatic cancer, pre-pancreatic cancer (panIN), type II diabetes and acute pancreatitis. These diseases were separated with >90% specificity from controls and from each other. We also describe a mathematical method that allows identification of 3 distinct components of an immunosignature: disease specific, ‘housekeeping’ and patient specific variation. The first component is useful in diagnosing disease, the second for baseline for the technology and third for monitoring changes in a healthy individual over time.

Keywords: Immunosignature, immune profile, random peptide microarray, microarray proteomics, pancreas disease, pancreatic cancer, type II diabetes, panIN, pancreatitis

3.2 INTRODUCTION

In theory, a given biomarker molecule can serve as a proxy for detecting and diagnosing disease and could be the most effective means of measuring drug efficacy and improving patient health (Weston and Hood 2004). One of the more ubiquitous technologies used for biomarker identification is mass spectrometry (L. Ackermann, E. Hale, and L. Duffin 2006; Lamont et al. 2006; Li et al. 2005). It has been widely used to search for diagnostics biomarkers, and the high sensitivity has made it useful for identifying informative biomarker molecules that associate with disease. This process of reducing biomarkers down to a single or few best candidates occasionally leads to overtraining, where highly precise biomarkers that work well in small cohorts become harder to correlate with large and diverse test populations (Kiehntopf, Siegmund, and Deufel 2007). It is becoming increasingly apparent that utilizing higher numbers of biomarkers simultaneously can relieve some of this ‘low-feature-number’ classification problem. Unfortunately, some attempts at using mass spectrometry to identify disease-associated mass spectrogram signatures have lead to skepticism about this concept (Chapman 2002; Davies 2000)

One of the major drawbacks of serum-based biomarkers is the dilution in the blood volume. The ability to detect small concentrations of protein or other biological compounds reproducibly has been tested and numerous issues with reproducibility and sensitivity have arisen (Barbosa et al. 2005; Diamandis and van der Merwe 2005; Elias et al. 2005). Were there a candidate biomarker that was abundant, unaffected by age, sex, race, or genetic factors, different between healthy and sick persons and physically stable, the problem would become simpler. One such candidate is immunoglobulin molecules. Antibodies are amplified during an illness so dilution is less of a problem, they are differentially abundant between healthy and ill person, they are stable and are relatively unaffected by genetic factors. The humoral immune response can distinguish non-self antigens, modified self-antigens in the case of autoimmune disease, and neo-antigens in the case of many cancers (Ada and Jones 1986; Brichory et al. 2001; Brydak and Machala 2000; Cox et al. 1994; DiFronzo et al. 2002; Hooks et al. 1979; Lennon, Lindstrom, and Seybold 1976; Sreekumar et al. 2004; Stockert et al. 1998; Wilder 1995).

In order to visualize changes in the antibody repertoire *en masse*, we developed a system we call 'immunosignaturing' (Boltz et al. 2009; Brown et al. 2011; Halperin, Stafford, and Johnston 2011; Legutki et al. 2010; Restrepo et al. 2011). We capture and display the complexities of humoral immunity using a microarray of random-sequence peptides. The system works for any isotype and has detected autoimmune disease, cancer, infectious disease, and chronic disease. The microarray is commercially printed to reduce variability and cost; technical reproducibility between replicate arrays averages 0.95 but is often >0.99. While we have seen clear distinctions between disease and healthy controls, we had not tested the idea that immunosignatures might be quite similar if a general inflammation response is raised for a particular target organ, though the primary disease might be quite different. We tested four different diseases that each affects the pancreas, leading to similar acute symptoms, but leading to substantially different late-stage symptoms (Dugernier et al. 2003; Fineberg et al. 2005; Orchekowski et al. 2005). Clinically, this would aid patients who present with similar early symptoms. If the immunosignatures revealed distinctions regardless of the common symptoms, it would enhance early intervention and could improve patient health. Is a general inflammation response driving the early humoral immune response in pancreatic disease or are antibody profiles distinct enough to predict disease. We examined patients with pancreatic cancer, pancreatitis, a pre-pancreatic cancer condition known as panIN, and type II diabetes.

Pancreatic cancer refers to a malignant neoplasm of the pancreas. About 95% of pancreatic tumors arise within the exocrine component of pancreas (Hruban et al. 2001; Li et al. 2004). Pancreatitis is inflammation of the pancreas due to ectopic or restricted activation of enzymes (Saluja and Steer 1999). PanIN stands for Pancreatic Intraepithelial Neoplasia and is the initial stage of pancreatic cancer (Hruban et al. 2000), also considered a non-carcinomic dysplasia. Type II diabetes is a chronic condition in which body has insulin resistance and deficiency resulting in high glucose level in the body (Katsilambros et al. 2006). There has been no complete survey of pancreatic diseases in the context of humoral immunity, but there is increasing evidence that patients with one pancreas disease have higher risk of a subsequent pancreas disease due to shared pathology and immunological involvement including autoimmunity (Ekbom et al. 1994; Huxley et al. 2005; Lowenfels et al. 1993; Deshpande et al. 2005; Inoue et al. 2006; Okazaki and Chiba 2002)

An immunosignature is the cumulative information from selected random-sequence peptides that bind differentially to antibodies from healthy controls vs. disease patients. Peptides are selected using statistical measures (t-test or ANOVA). Each signature, whether at a single time point from multiple patients with the same disease or from a single patient across multiple time points, can be considered a vector. This vector has three major components: 1) the disease component, 2) the unchanged component and 3) the personal variation component. **Figure 3-1** (A) illustrates the three components, (B) explains the three components over the array. The (3,4) array is laid out in order to explain the three components. The block in red explains the personal variation component since all peptides in the array have different binding. The block in black (2, 1), (2, 2), (3, 1), (3, 2) explains the disease component (uniqueness). These peptides are different in disease groups vs. controls. The third component (normal component) is explained in the blue block (4, 1), (4, 2) and also whole of 3rd column peptides which is consistent in the all the groups.

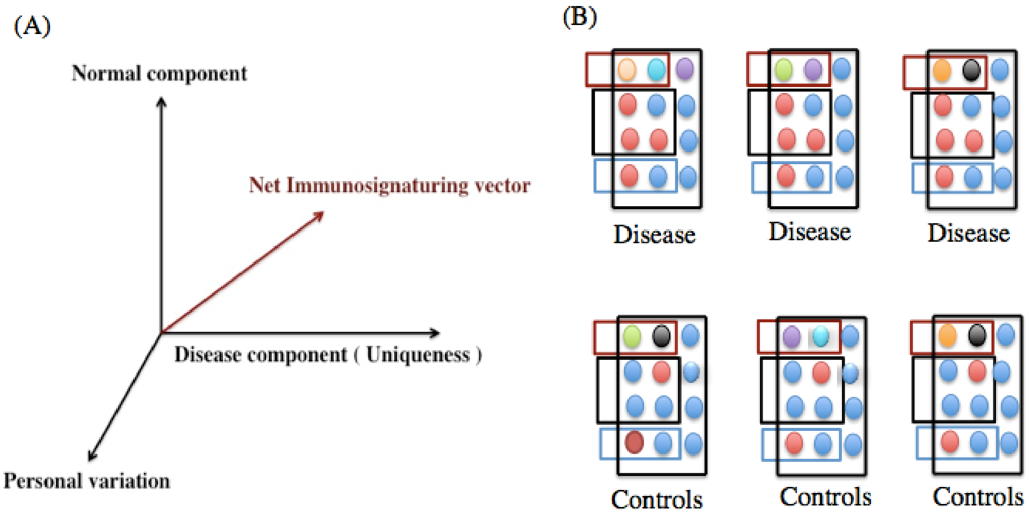


Figure 3-1: (A) shows Immunosignaturing vector comprising of three major components. (B) explains these components in (3*4) array. Red block showing the personal variation component, black block showing the disease component while the blue block showing the normal

The first component consists of peptides that show a relative ‘up’ or ‘down’ response during the course of disease compared to healthy controls. A simple t-test with multiple testing corrections applied can identify peptides that are reproducibly higher or lower in patients vs. controls. Typically, biomarkers are missing in healthy controls and begin to appear in patients with a given disease. In immunosignaturing, signals can be either higher *or* lower between disease and control; this is not typical for the biomarker paradigm.

The second component represents peptides that do not change between disease and healthy individuals. These antibodies are not activated during disease, and may simply be circulating or basal level antibodies produced against a common infection or vaccination. This component helps quantify the part of the immunosignature that does vary during the course of disease, helping to establish a baseline of variance and dynamic range.

The third component is personal variation and signifies the behavior of an individual's own immune system. This component is necessary when establishing a baseline for a patient over time. These three components are extracted mathematically from a given immunosignature. We present these three components in the context of our analysis of four pancreas diseases.

3.3 MATERIALS AND METHODS

3.3.1 MICROARRAY

The CIM 10K array is a 2-up microarray containing 10,000 random-sequence 20-mer peptides attached via a maleimide reaction to the NH_3 terminal sulfur of cysteine, creating a covalent attachment (Boltz et al. 2009; Brown et al. 2011; Halperin, Stafford, and Johnston 2011; Legutki et al. 2010; Restrepo et al. 2011). The CIM 10K microarray is available to the public at www.peptidemicroarraycore.com.

3.3.2 SAMPLES PROCESSING

Plasma samples from patients and healthy controls were stored at $-80^{O}C$ until needed. Samples were aliquoted and refrozen at $-20^{O}C$. Samples were diluted at 1:500 in sample buffer (1xPBS, 0.5% Tween20, 0.5% Bovine Serum Albumin (Sigma, St. Louis, MO)) and exposed to the array according to the protocol in (Legutki et al. 2010). Antibodies were detected with 5nm Alexafluor 647-labeled streptavidin (Invitrogen, Carlsbad, CA), which bound 5nM biotinylated anti-human secondary antibody (Novus anti-human IgG (H+ L), Littleton, CO). Microarrays were scanned and converted to tabular data as in (Legutki et al. 2010). Median foreground signal was used as the value which best-represented binding of antibody to peptide.

3.3.3 SAMPLES

Center for Innovations in Medicine, Biodesign Institute, Arizona State University has an existing IRB 0912004625, which allows analysis of blinded samples from collaborators.

1) Type II diabetes: 17 plasma samples which had poorly controlled type II diabetes with no history of CHF (Congestive Heart Failure) and MF (Myocardial Infraction).

2) Pancreatic cancer: This set contains 13 plasma samples from patients with ductal adenocarcinoma of the pancreas.

3) Pancreatitis: This set contains 10 plasma samples of patients with refractory pancreatitis.

4) PanIN: This set contains 5 plasma samples. Samples were obtained from a single family with history of pancreatic cancer. Samples were diagnosed with a pre-stage of pancreatic cancer.

5) Common Controls: This set contain 16 plasma samples from the diabetes study.

3.3.4 DATA ANALYSIS

The raw tabular data were imported to GeneSpring 7.3.1 (Agilent, Santa Clara, CA). Data were median normalized per array and $\log_{10}$ transformed. Feature selection used t-test with family-wise Multiple Error correction of 5% (FWER=5%). For multiple groups we used 1-way fixed-effects ANOVA, FWER=5%. *All p-values presented are after FWER correction*. The three components were selected as follows: component 1 (disease component) was selected by using t-test. Component 2 (unchanged component) was selected by ANOVA (FWER = 5%) on all samples including controls and disease, these peptides are the ones which were not selected by ANOVA signifying no significant change over samples excluding those peptides that were selected for component 1. Component 3 (unchanged component) are those peptides that passed ANOVA (FWER= 5%) on all samples including disease and controls. For classification, Naïve Bayes and leave one out cross-validation were used. Classification was performed in open source JAVA software WEKA (Hall et al. 2009).

3.4 RESULTS

3.4.1 ANALYSIS OF IMMUNOSIGNATURING VECTORS

10 samples of pancreatitis (Kijanka et al.), 5 samples of panIN (PN), 17 samples of type II diabetes (T2D), 13 samples of pancreatic cancer (Kijanka et al.) and 16 samples of healthy controls were run in duplicate on the 10K peptide microarrays. Technical replicates with Pearson's correlation coefficient <0.90

were discarded. For each disease, the three components listed in **Table 3-1** were determined as a number of peptides at a given p-value (p<0.05, FWER= 5%).

Features that comprise each of the three immunosignaturing components were identified at an adjusted p<0.05 and are presented in **Table 3-1**. The disease components of pancreatic cancer and panIN contribute from 10-20% (lowest to highest) to the net immunosignaturing vector, while the disease component of type II diabetes and pancreatitis contributes little (< 3%) to the net vector. The unchanged components within type II diabetes, pancreatitis and panIN contribute 17-23% (average=19%) to the net vector while that of pancreatic cancer contributes <5%. The personal variation component comprises most of the immunosignaturing net vector in pancreatic cancer, with 60-80% (average=76%).

Table 3-1: Distribution of three components for pancreas disease:

# peptides	Type 2 diabetes	Pancreatic cancer	Pancreatitis	PanIN
Disease component	92	1058	258	1696
Unchanged component	1700	536	2235	2041
Personal variation	8208	8406	7507	6263

3.4.2 IMMUNOSIGNATURING OF EACH PANCREAS RELATED DISEASE VS COMMON CONTROLS

Each of the diseases tested were subjected to a test/training analysis consisting of feature selection (component 1) followed by classification using Naïve Bayes and leave-one-out cross-validation. Naïve Bayes treats features as completely independent sources of information, which has advantages for a system like immunosignaturing, less so for expression or SNP microarrays where there is a biological connection across features.

We did a Welsh t-test between each pancreas disease and common control with multiple testing corrections and the null hypothesis was for 92 peptides with p<5e-2 for type 2 diabetes, 244 peptides with p<5e-3 for pancreatic cancer, 258 peptides with p<5e-2 for pancreatitis and 233 peptides with p<5e-4 for panIN.

3.4.3 IMMUNOSIGNATURING OF TYPE II DIABETES AND CONTROLS

For type 2 diabetes, Figure 3-2 shows the most informative peptides among the 90 selected features that were upregulated in the tested disease state. The principal component analysis shows that ~70% of the variance is explained by the first two components.

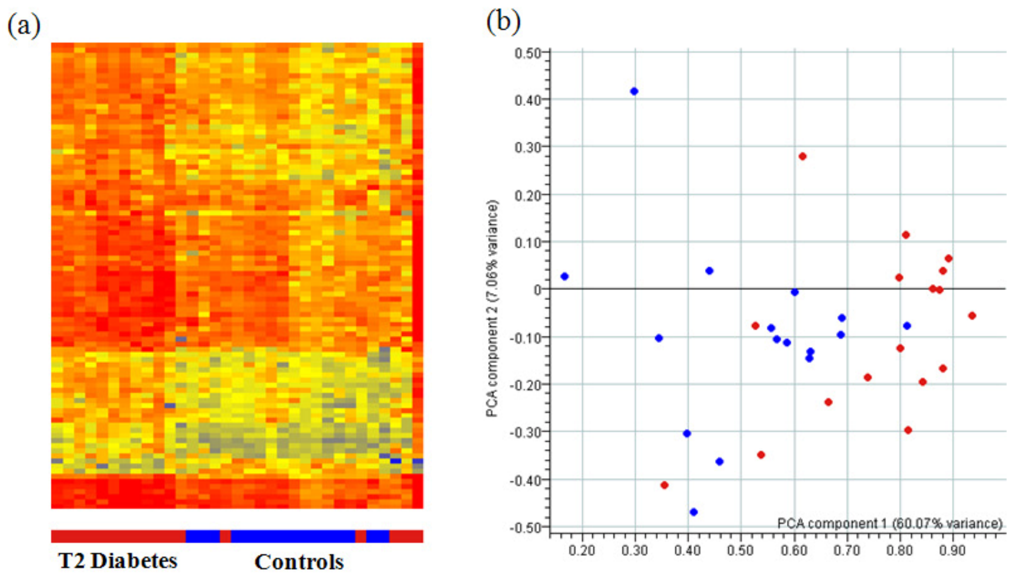


Figure 3-2 Heatmap and PCA analysis of 90 peptides for type II diabetes and controls

The performance from the peptides that compose the disease immunosignaturing component is shown in **Table 3-2**, where specificity > 93%.

Table 3-2 Performance measure and classification table of type II diabetes and controls

Accuracy	87.88 %
Specificity	93.75 %
Sensitivity	82.4 %

Predicted ->	T2 Diabetes	Controls
T2 Diabetes	14	3
Controls	1	15

3.4.4 IMMUNOSIGNATURING OF PANCREATIC CANCER AND CONTROLS

For pancreatic cancer, the heatmaps in **Figure 3-3** show that approximately half of the informative peptides in the 244 show high binding response and other half shows low binding response. In principal component analysis, about 65 % of variance is explained. The two groups are well-separated by these two components. More than 90% performance measures were obtained for this set.

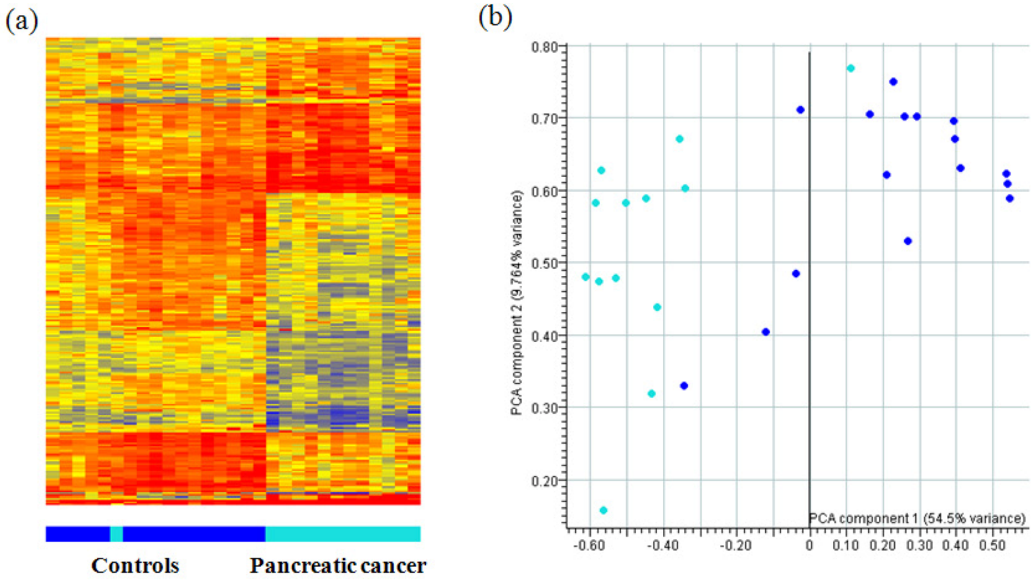


Figure 3-3: Heatmap and PCA analysis of pancreatic cancer and controls

The performance from the peptides that compose the disease immunosignaturing component is shown in Table 3-3, where specificity > 93%.

Table 3-3: Performance measure and classification table of pancreatic cancer and controls

Accuracy	93.10 %
Specificity	93.8 %
Sensitivity	92.3 %

Predicted ->	Pancreatic	Controls
Pancreatic	12	1
Controls	1	15

3.4.5 IMMUNOSIGNATURING OF PANCREATITIS AND CONTROLS

For pancreatitis, about 25% of the top 262 peptides showed high binding response while the rest showed lower binding response. For the principal component analysis, 60% of the variance is explained by the first two components and the two groups are well separated. The performance accuracy is >95% with 100% specificity.

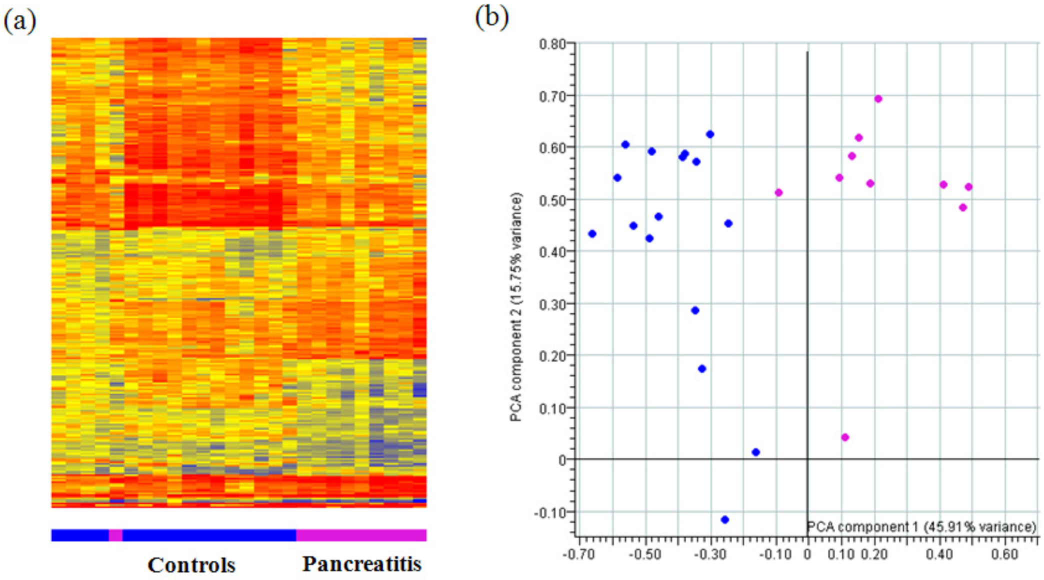


Figure 3-4: Heatmap and PCA analysis of pancreatitis and controls

The performance from the peptides that compose the disease immunosignaturing component is shown in **Table 3-4**, where specificity is 100%.

Table 3-4: Performance measure and classification table of pancreatitis and controls

Accuracy	96.15 %
Specificity	100 %
Sensitivity	90 %

Predicted ->	Pancreatitis	Controls
Pancreatitis	9	1
Controls	0	16

3.4.6 IMMUNOSIGNATURING OF PANIN AND CONTROLS

For panIN, the heatmap shows that much of the top 233 peptides showed lower binding response while some showed higher binding response. In principal component analysis about 70% of the variance is explained by two components and the groups are very well separated. The performance accuracy is more than 95% with 100% specificity

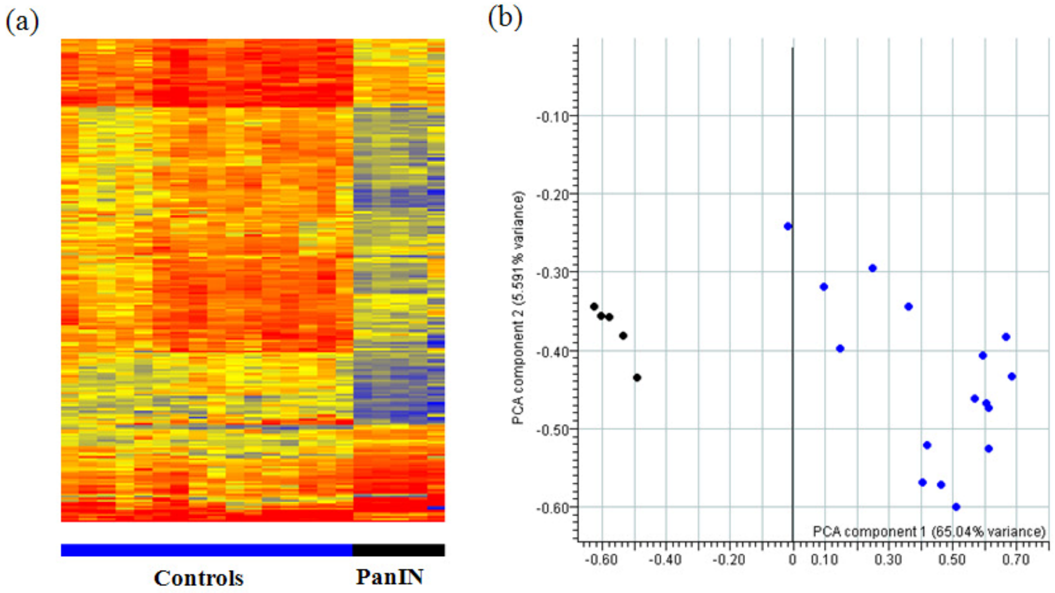
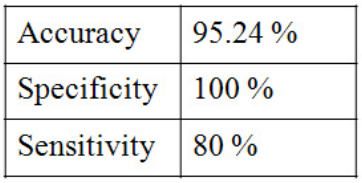


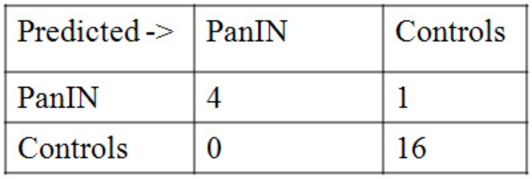
Figure 3-5: Heatmap and PCA analysis of panIN and controls

The performance from the peptides that compose the disease immunosignaturing component is shown in **Table 3-5**, where specificity is 100%.

Table 3-5: Performance measure and classification table of panIN and controls



Accuracy	95.24 %
Specificity	100 %
Sensitivity	80 %



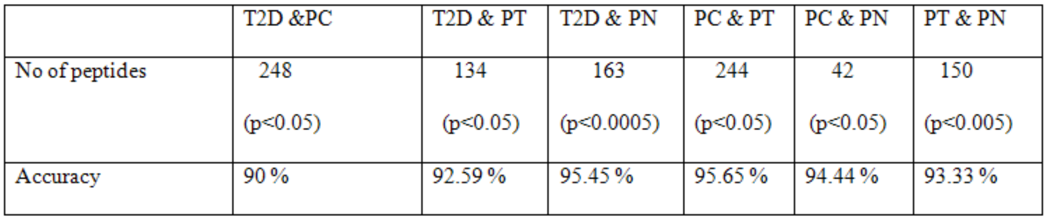
Predicted ->	PanIN	Controls
PanIN	4	1
Controls	0	16

We established the primary features that distinguish disease vs. control, along with the p-value cutoff, classification accuracy, specificity and sensitivity using Naïve Bayes error and leave one out cross validation. We then asked of the peptides that are changed between control and each disease, how many are up or down compared to controls. Of note, detecting informative signals less than normal is a feature not easily done for ELISA-type assays. For type II diabetes, most (~90%) of the peptides intensities were high compared to controls while for panIN, the same percentage were lower. For pancreatic cancer and pancreatitis, between 50 and 60 % were lower compared to controls while 40 to 50% of the selected peptides were higher. Given the initial question of similarity between diseases that affect the same target organ, and how much of an immunosignature is derived from a general inflammation response, we asked how many of these peptides were in common, and how well could we distinguish the diseases from each other.

3.4.7 IMMUNOSIGNATURING OF PANCREAS DISEASE WITH EACH OTHER

Features that distinguish each disease from every other were identified using corrected t-test and presented in **Table 3-6**. The accuracy of Naïve Bayes classification using leave-one-out cross validation is shown for each comparison is shown along with the p-value cutoff for each comparison.

Table 3-6: Classification between each pancreas disease



	T2D &PC	T2D & PT	T2D & PN	PC & PT	PC & PN	PT & PN
No of peptides	248 (p<0.05)	134 (p<0.05)	163 (p<0.0005)	244 (p<0.05)	42 (p<0.05)	150 (p<0.005)
Accuracy	90 %	92.59 %	95.45 %	95.65 %	94.44 %	93.33 %

Notably at $p<0.05$, the differences between type 2 diabetes, pancreatic cancer and pancreatitis are small (< 2% of the peptides are distinct) while the difference between pancreatic cancer and panIN is <1%. Between type 2 diabetes and panIN the differences were greater, at 15% of the peptides. Between pancreatitis and panIN, 11% but we were able to achieve > 90% classification accuracy for each disease comparison. These are the differences between net immunosignatures of pancreas related diseases. These differences come from various sources, it can come for either of the disease component or it can come from other components. To get more insight into the differences, the source of the difference in each comparison is calculated and shown in **Figure 3-6.**

TheVenn diagram shows three overlapping sets. The left set shows the number of peptides that are 95% significantly different in the two diseases. Similarly, the right and the bottom set show peptides that are 95% significantly different in each disease and its common controls. As described earlier, peptides which are significantly different in a disease with respect to the common controls are the ones which make the disease component or uniqueness. Differences between disease and the common controls at 95% confidence interval, varies by a huge amount. These differences comprises of uniqueness component. It is 92 peptides for type 2 diabetes, 258 for pancreatitis, 1032 for pancreatic cancer and lastly 1696 for panIN.

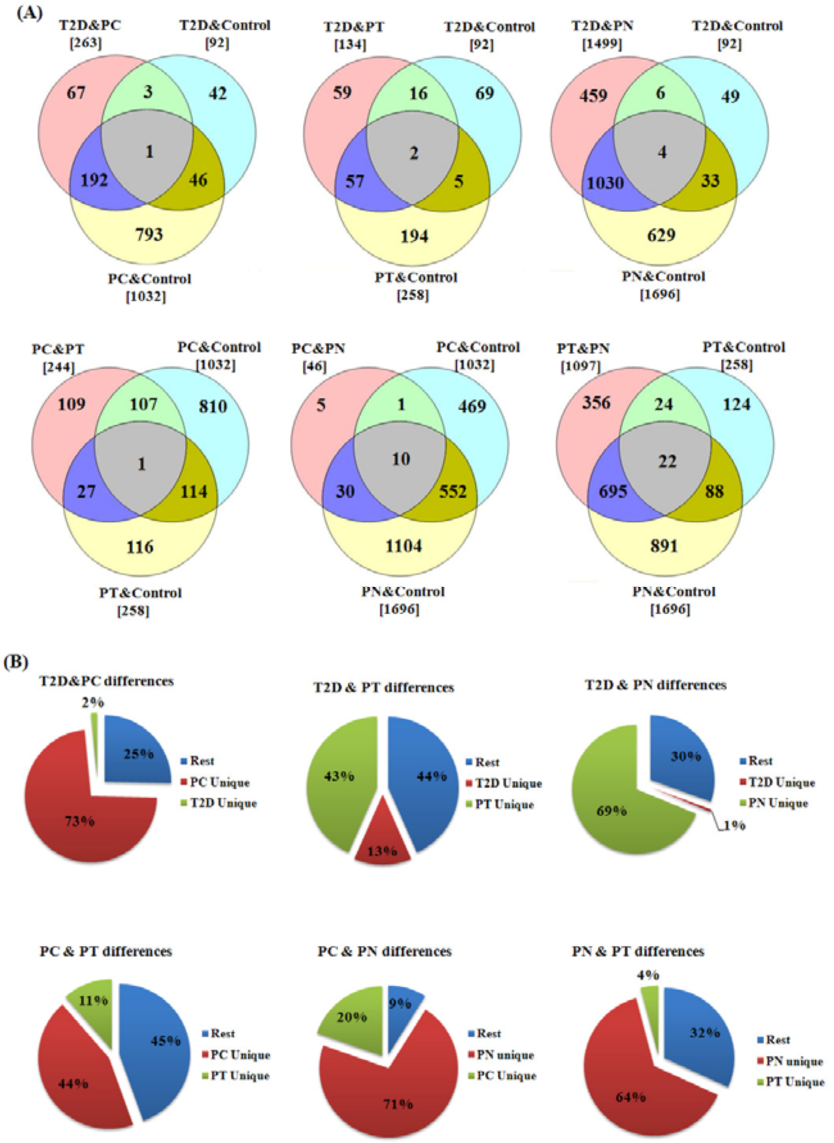


Figure 3-6: The source of differences in net immunosignatures of pancreas related disease at 95% confidence interval. (A) shows the Venn diagram of difference between 2 pancreas diseases and their difference from common controls. (B) summarizes the each component

In the 2.62% difference between the net Immunosignature of type 2 diabetes and pancreatic cancer, only 2% is type 2 diabetes unique, and majority (73%) is pancreatic cancer unique. Hence the small difference that type 2 diabetes and pancreatic cancer have, it is due to peptides that are highly specific to pancreatic cancer. In the 1.34% difference type 2 diabetes and pancreatitis, only 13% is due to type 2 diabetes unique signatures and 43% is due to pancreatitis signature. Hence, overall the differences between type 2 diabetes and pancreatitis is primarily due to the peptides, which are specific to pancreatitis. In the significantly large difference (15%) between type 2 diabetes and panIN, only 1% is due to type 2 diabetes unique signatures and about 69% is due to peptides, which are specific to panIN. Hence it signifies, the differences between the type 2 diabetes and panIN is largely due to the peptides, which are specific to panIN. In the 2.44 % difference between pancreatic cancer and pancreatitis, about 11% difference is to pancreatitis signature and 44% is due to peptides unique to pancreatic cancer. Hence the majority of the difference between pancreatic cancer and pancreatitis is due to other components and pancreatic cancer unique signature. In an extremely small difference (0.46%) between pancreatic cancer and panIN, 71% of the difference is to the peptides specific to panIN and about 20% of the difference is due to pancreatic cancer. Hence the small difference between the pancreatic cancer and panIN is due to panIN unique signature. Lastly the significantly large difference between the pancreatitis and panIN signature is to peptides specific to panIN (64%) and a very small difference (4%) is contributed by pancreatitis unique signature.

Here we quantified the differences between each pancreas disease and respective component contribution in each difference. Type 2 diabetes unique signatures contribute significantly less in the difference from the other pancreas disease. Pancreatitis uniqueness only dominates when compared to type 2 diabetes. PanIN uniqueness always dominates over the other disease uniqueness. In order to get commonality among these diseases, we compared each pancreas disease signature pairwise.

3.4.8 SIMILARITIES AMONG DISEASE COMPONENT OF PANCREAS DISEASE

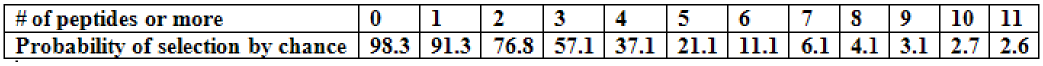
The top 200 features that distinguished each disease component from controls were identified, regardless of the p-value cutoff. Combinatorially, the probability of seeing *r* or more peptides among 200 from each disease component by chance is obtained from equation 1, where *r* is the set of peptides selected to distinguish any given comparison, p is number of peptides selected out of total (200) and n is total no of peptides (10,000). For *r* > 10, this probability is <1%. Probability of *r* or more peptides is given by **Equation 1**.

Equation 1: Probability of selecting r or more peptides by chance

$$\text{Probability} = 1\text{-} \left(\frac{\sum_{i=0}^{r-1} \binom{p}{i} * \binom{n-p}{p-i}}{\binom{n}{p}}\right)$$

This equation can simply be derived from simple permutation and combination. We substituted r=10,000 and p = 200 and varied i=0 from 11 to come to more than 95% significance level. **Table 3-7** shows the probability of selection of r or more peptides by chance for this case. Hence the probability of having 11 or more peptides in common is just 2.6%, which is greater than 97% significance level. So if we see more than 11 peptides in common, its highly likely that these two diseases share something in common which is not by chance.

Table 3-7: Probability of selection of r or more peptides in common while selecting 200 peptides twice from pool of 10,000.



# of peptides or more	0	1	2	3	4	5	6	7	8	9	10	11
Probability of selection by chance	98.3	91.3	76.8	57.1	37.1	21.1	11.1	6.1	4.1	3.1	2.7	2.6

Figure 3-7 (D) shows that when we compared two unrelated disease like Alzheimer's unique versus H1N1, valley fever, tularemia, and influenza there are about 1-10 peptides, which are in common, which would occur by chance. Also when we compared T1D that happens in kids and flu in mice, we got 3 peptides in common among 200. Hence seeing 10 peptides or less in common out of 200 can be purely accidental. Hence no significant conclusions can be drawn in these cases. (C) shows the binding of peptides among 200 in each disease uniqueness. This can also be seen from heatmaps for every pancreas disease. For type 2 diabetes, 195 out of 200 peptides were high binders. 74 peptides are high binders in pancreatic cancer and 61 peptides are high binders in pancreatitis. In panIN only 28 peptides that are high binders. (A) and (B) shows the overlap between two diseases uniqueness. Among the 200 peptides for each disease, 22 peptides were common in type 2 diabetes unique and pancreatic cancer unique. Out of 22 peptides, 21 were high binders and 1 was low binders compare to expression in

common controls. Having such peptides in common is much above the significance level and reveals there is some common signature in two diseases. 22 peptides in common that are high binders in both the diseases signify that these peptides can be mimitopes of the same antigen. Comparison of type 2 diabetes unique and panIN unique reveals that there is one peptide in common which may be highly due to chance as seen from the Eq. 3.3.1. Also type 2 diabetes and pancreatitis have only 8 peptides in common which is likely due to chance. There are 50 peptides that are common between pancreatic cancer and pancreatitis out of which most of the peptides are low binders. Also in pancreatic cancer and panIN, 50 peptides are in common out of which only 9 peptides are up regulated and rest are down regulated. In pancreatitis and panIN, there are 24 peptides is that in common and all of these are low binders. This analysis shows that, type 2 diabetes only has significant common signature with pancreatic cancer and not with any other pancreas disease, hence in (B) the Venn diagram shows 17 peptides which were common in pancreatic cancer, panIN and pancreatitis, all of which were low binders and not present in type 2 diabetes. This analysis shows significant commonality in some of the pancreas related diseases. Also epidemiology supports the argument of that occurrence of one is correlated with another pancreas disease. In this line, taking this information from a diagnostic level to monitoring the level of pancreas disease, we considered to design a chip of peptides, which can be used to monitor people with pancreas related diseases.

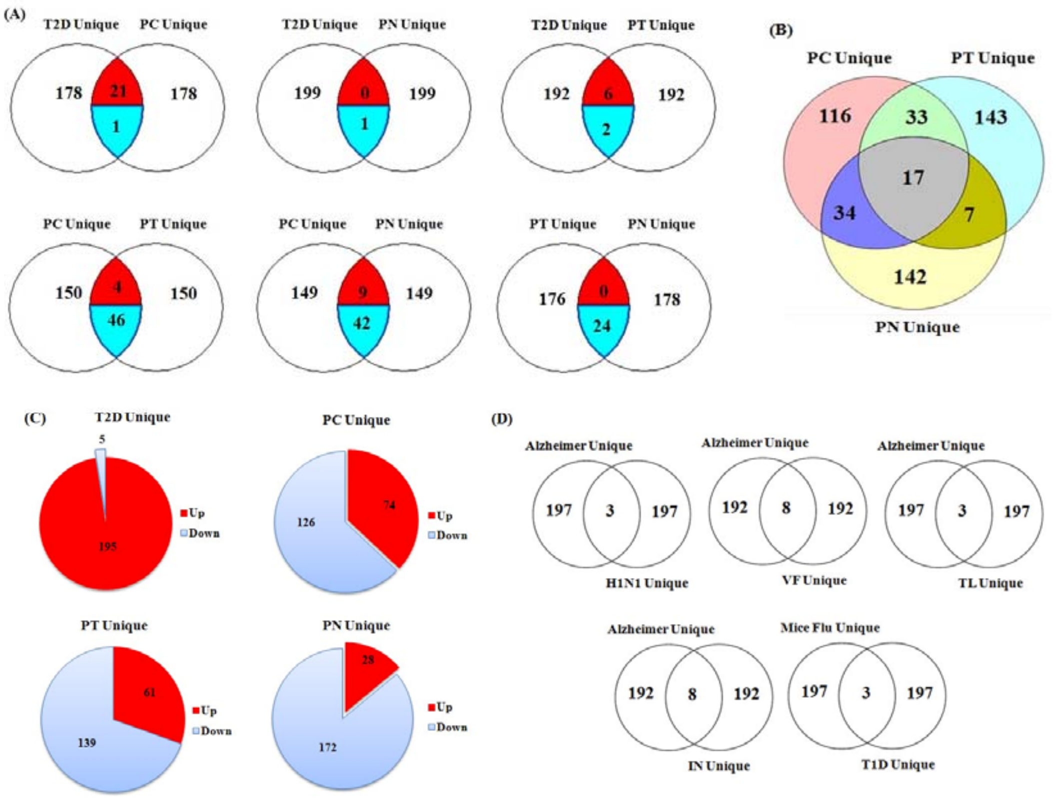


Figure 3-7: shows the Venn diagram of each pairwise comparison for every disease, using 200 peptides selected by t-test. (B) shows the Venn diagram for pancreas diseases

3.4.9 TOWARDS A GENERAL DIAGNOSTIC OF PANCREAS RELATED DISEASE

The pancreas diseases we have considered in this work show both certain unique components and also significant evidence of having some signatures in common. Here, we have considered each disease under one pancreas disease and uncover the potential to identify pancreas disease from common controls. To achieve this, we have considered two methods to select peptides that would be appropriate to perform this task most efficiently. Here, we compared these two approaches and pick the best one that would be suitable for this purpose.

Method A: *Selection of peptides from t-test between pancreas diseases and common controls (2 classes)*

Method B: *Selection of peptides from multiple t-test (4) between each disease and common controls (4 times 2 classes), which is identical to choosing from each disease uniqueness*.

For the first method, we labeled each pancreas disease as disease and performed a t-test and pick top 643 peptides ($p<0.025$). For the second method, we considered top 200 peptides from each disease uniqueness. Since there were some peptides that were in common, we got 668 unique peptides out of 800 (200*4). **Figure 3-8** shows the Venn diagram of peptides selected by these two methods. Only 50% of peptides are in common by these two approaches.

This indicates that the peptides which are informative to classify a super set of diseases are different than that of peptides that are informative for identifying each subset of super set. This has a huge implication in biomarker discovery. In order to compare these two approaches, we compared their performance measures accuracy, specificity and sensitivity over ability to detect pancreas disease from common controls.

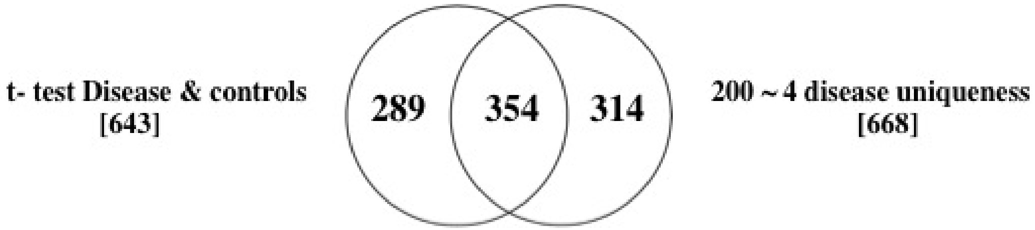
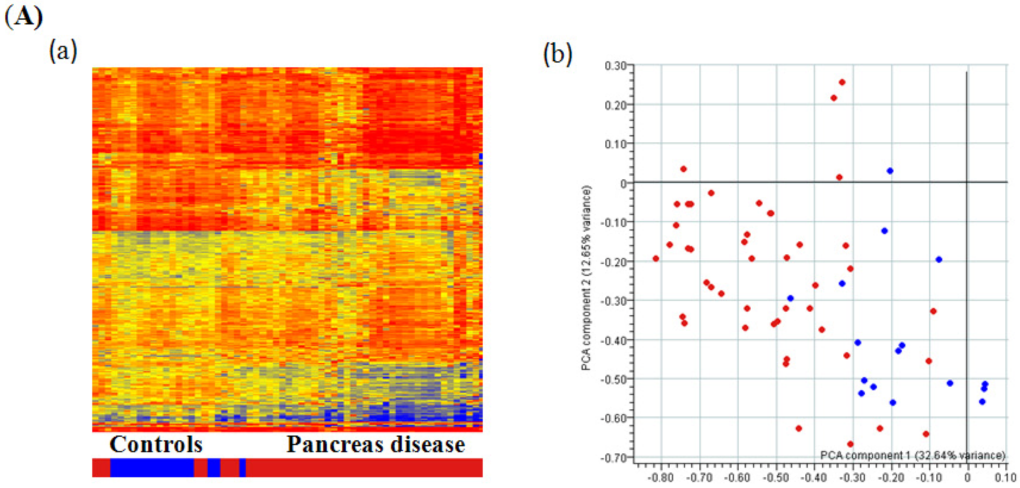


Figure 3-8: 1Venn diagram showing selection of peptides by t-test between pancreas disease and controls (method 1) and iterative t-test between each pancreas disease and common controls (method 2).

Heat map for method 1 Figure 3-9 (A)(a) and method 2 Figure 3-10 (B)(a) shows that in method 1, peptides have high gradient of regulation between pancreas disease and common controls.



(c)

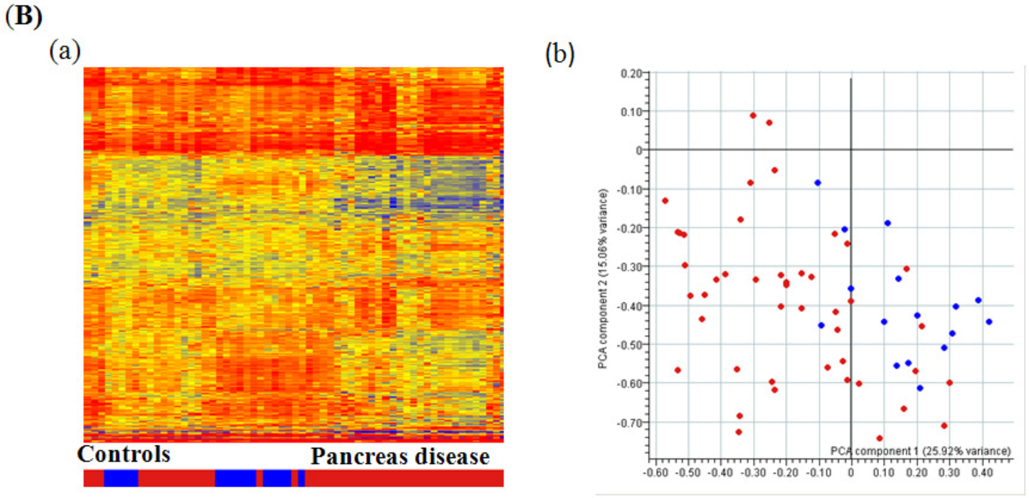
Accuracy	91.8 %
Specificity	93.8 %
Sensitivity	91.8 %

(d)

Predicted ->	Disease	Normal
Disease	41	4
Normal	1	15

Figure 3-9: Heatmap, PCA and performance measures for method 1

The controls are more tightly clustered in method 1 compared to method 2. Principal component analysis of method 1 explains about 45% of the variance in two components while that of method 2 explains about 40% of the variance. The samples are almost equally separated in both the principal component analysis but method 1 is better due to higher percentage of variance explained. Figure 3-9 and Figure 3-10 (c) (d) shows the performance measures and classification table in which method 1 is better than method 2 in all the measures including accuracy, specificity and sensitivity. Hence method 1 is overall better than method 2 in finding the peptides to detect pancreas disease and controls.



(c)

Accuracy	**86.9 %**
Specificity	**75 %**
Sensitivity	**91.1%**

(d)

Predicted ->	**Disease**	**Normal**
Disease	**41**	**4**
Normal	**4**	**12**

Figure 3-10: Heatmap, PCA and performance measures for method 2

3.5 DISCUSSION

We used immunosignaturing to examine different diseases that target the same organ, the pancreas. We tested whether there was a general immunological affect that might render signatures from patients with pancreas disease very similar by looking for overlapping peptides. We found a distinct set of peptides that could classify each of the 4 diseases, but there were also peptides that were common across the 4 diseases. We found that there were different numbers of peptides that were in common across the 4 diseases and different numbers of peptides that were uniquely personal to each patient and different numbers of peptides that were unchanged within the disease class.

Initially we investigated whether each disease was distinct compared to healthy controls. We obtained >95% specificity on average (93.75, 93.8, 100, 100 for type II diabetes, pancreatitis, panIN and pancreatic cancer respectively). Next we tested whether each disease was distinct from each other, and in this case we obtained >90% classification accuracy. We thus show that each disease, although affecting the pancreas to some extent, also has unique immunological characteristics.

We then looked for similarities between each pancreas disease. The disease component (the part of the signature that defines the uniqueness of each disease) was used to examine the commonality between type 2 diabetes and pancreatic cancer. These two diseases share a significant portion of this component, perhaps caused by common immunological stimuli. All the common peptides were up compared to controls, suggesting common antigens. Similarly, there was significant similarity across pancreatic cancer, pancreatitis and panIN. All common peptides were down compared to controls suggesting that there may be some immune suppression in these diseases. The different pancreas diseases have their own unique signatures but also share portions of their 'common component' (component 2). The third component (personal variation) was found to contribute differently across the diseases. This component is important when using immunosignaturing for monitoring health status over time. The patient uniqueness was established as those peptides which differ from person to person, excluding the disease specific (component 1) and housekeeping or 'normal component' (component 2) peptides. We found that a range between 60 and 85% of the total 10,000 different peptides were individual specific (component 3) when examining these 4 diseases.

Finally, in order to establish the potential of this technology for creating a diagnostic, we tested whether mixing peptides specific to a number of diseases would detrimentally affect the classification performance. We chose 643 peptides that differentiate these 4 diseases. These peptides could be printed on a 24-up microarray, which allows much cheaper per-assay cost and far higher throughput. Mixing the most informative peptides for distinguishing each disease from controls and from each other yielded >90% classification accuracy. The fact that a pattern of peptides can be found that reliably distinguishes a disease from unrelated individuals is remarkable, but the presence of at least 3 distinguishable components within that signature lends credence to the fact that antibodies are highly tuned to the health status of an individual.

3.6 ACKNOWLEDGEMENTS

This chapter is published at Journal of Proteomics and Bioinformatics, Special Issue of Microarray Proteomics 2012 (doi:10.4172/jpb.S6-001). We are thankful to Teri Brentnall for providing samples of pancreatic cancer, pancreatitis and panIN and Molecular Biosignatures Analysis Unit at Biodesign Institute, ASU for providing type II diabetes samples and common controls. This work was supported by discretionary funds from SAJ.

CHAPTER 4 FUSION EFFECT OF CARDIO VASCULAR COMPLICATIONS ON TYPE II DIABETES IMMUNOSIGNATURE

4.1 ABSTRACT

Immunosignaturing technology has allowed capturing the humoral immune response through binding of antibodies to random sequence peptides. Profiles of the antibody repertoire produced during any single infection or during long term chronic disease have proven to be unique, reproducible, stable and consistent across the diseased population. Every disease by itself has a unique immunosignature specific to its immunological binding of antibodies during its course. But in practice, individuals are often exposed to multiple diseases thereby potentially confounding the antibody binding to random sequence peptides. An important unanswered question relative to this technology is what is the net profile of the antibody repertoire when two diseases or infections fuse together biologically in a single person. Cardiovascular complications such as CHF and MI are often accompanied with type II diabetes (T2D) or vice versa. In this work, we obtained the immunosignature of subjects suffering from T2D and cardiovascular complications exclusively and then compared it with the immunosignature of subjects having both T2D and cardio vascular complications to observe the net effect of immunosignature when both the complications are fused biologically. We found, that immunosignature of T2D and cardiovascular disease were consistent and we were able to achieve 100% sensitivity when compared with common control group. The immunosignature of biologically fused mix of T2D

and heart complications was moderately consistent and unique, yielding only 81.25% accuracy. When tested for interaction for the two complications as disease factors by 2 way ANOVA, the interaction was highly significant p<0.001 indicating the non orthogonality of two immunosignature. Overall, the signature of the mixture of the two complications was different to the original complications, thereby resulting in a new complication with respect of their immunosignatures.

4.2 INTRODUCTION

Serological diagnostics biomarkers have received increasing scrutiny recently (Kurian et al. 2007) due to their potential to measure antibodies rather than low abundance biomarkers. Using antibodies as biomarkers solves the biomarker dilution problem, moreover these molecules are recruited rapidly following infection, chronic or autoimmune diseases, or even exposure to cancer cells (Ada and Jones 1986; Brichory et al. 2001; Cox et al. 1994; Fineberg et al. 2005). Using antibodies as serological diagnostics biomarkers can reduce medical costs and may be the one of the best methods for early diagnostic. Immunosignaturing is one such technology that aims to detect disease early accurately. The platform consists of a peptide microarray with 10,000 peptides per assay. For every disease that we have tested we have a found a unique, consistent and a reproducible pattern of immunosignature that can be used for predicting new test cases. As a control study, this unique pattern of diseases has been obtained from a tightly controlled study involving subjects with disease of interest and matched controls (Brown et al. 2011; Chase, Johnston, and Legutki 2012; Legutki et al. 2010; Restrepo et al. 2011). But in reality, many subjects often have two or more disease which might be due to correlation, causation or even by chance. Multiple diseases may often interact with each other and change the resultant profile of antibodies produced by the immune system. Since immunosignaturing technology profiles the humoral immune response by profiling the net antibody profile of a disease, it is imperative to know how the net profile/immunosignature would change when two diseases are mixed inside a person. Immunosignatures are obtained through sera or plasma containing the antibodies source. When two antibody sources of disease (sera/plasma) are mixed physically rather than biologically the net immunosignature changes since every antibody source concentration is now diluted to half. So if the two sources have identical antibody profile, the net sera mixture would look exactly the same as the individual ones but if this effect is no additive, or if there is an interaction between the two profiles, then the resultant mixture immunosignature would change. In earlier work on immunosignaturing technology, we have seen that two completely independent immunizations of KLH and PR8 in mice sera mixture immunosignature have the respective KLH specific immunosignature and PR8 specific immunosignature (Stafford et al. 2012). But when two diseases would interact biologically due to correlation or causation, the sera mixture could also change. Another unanswered question related to the biological mixing and sera mixing of immunosignature is how similar would the immunosignatures be?

T2D is a chronic metabolic disorder due to high blood glucose accompanied with insulin resistance and relative insulin deficiency (Burgoyne 1961; Hamilton 1953). One of the primary reasons for the cause of T2D is due to obesity which results in fat deposits (Ozcan et al. 2004). One of the long term complications from high blood sugar can include heart disease and strokes. Cardio vascular complications such as Congestive Heart Failure (CHF) and Mycocardinal infarction (MI) are results of interruption of blood supply to a part of the heart that causes it to die. These is primarily due to occlusion of coronary artery. When taken together, epidemiology studies have shown that T2D and cardiovascular complications are correlated (Haffner et al. 1998).

In this work, we tested the hypothesis that the immunosignature of subjects having both cardiovascular complications (CHF & MI) and T2D (biological mixture) will be the sum of the immunosignatures of subjects having cardiovascular complications and T2D exclusively. We also tested a sera mixture (1:1) of T2D samples and cardiovascular samples resemble the signature of the biological mixture of subjects having both the complications.

4.3 METHODS

4.3.1 MICROARRAY

The CIM 10k V.1 is a 1-up microarray containing 10,000 random sequence 20-mer peptides top and bottom attached via a maleimide reaction to the NH_3 terminal sulfur of cysteine, creating a covalent attachment (Legutki et al. 2010).

4.3.2 SAMPLES PROCESSING

Plasma samples from subjects with T2D, CHF& MI, (T2D + CHF&MI) and were stored at -80 C after which samples were aliquoted and refrozen at -20 C. Samples were diluted at 1:500 in sample buffer containing 1XPBS, 0.5% Tween20, 0.5% Bovine Serum Albumin (Sigma, St. Louis, MO) and then applied to the array using the standard immunosignaturing protocol as mentioned (Legutki et al. 2010). Antibodies were detected with 5nm Alexafluor 647-labeled streptavidin (Invitrogen, Carlsbad, CA). This tertiary antibody bind to 5nM biotinylated anti-human secondary antibody (Novus anti-human IgG (H+L), Littleton, CO). Microarrays were scanned and converted to tabular data after which median foreground signal was used as the value of antibody peptide binding.

4.3.3 SAMPLES

Center for Innovations in Medicine, Biodesign Institute, Arizona State University has an existing IRB 0912004625, which allows analysis of blinded samples from collaborators.

We collected 11 plasma samples of subjects having poorly controlled type 2 diabetes, 7 plasma samples having CHF&MI and 7 plasma samples of subjects having both T2D and CHF & MI related symptoms and 11 plasma samples of healthy controls.

A plasma mixture is obtained by 1:1 dilution of two individual plasma samples of T2D and (CHF&MI) samples.

4.3.4 DATA ANALYSIS

The raw tabular data were imported to GeneSpring 7.3.1 (Agilent, Santa Clara, CA). Data were median normalized per array and then transformed into log $_{(10)}$ scale.

For feature selection (choosing significant peptides in each condition), we performed Welsh t test after multiple testing corrections (FWER=5%) to select peptides that are differential among the classes and then the analysis is performed. Line graph, scatter plot, principal component analysis and classification have been performed for supervised methods. For classification, we used Naïve Bayes with leave one out cross validation method since this algorithm has been known to work best for immunosignaturing data (Kukreja, Johnston, and Stafford 2012). For finding the interaction between the two disease factors T2D and (CHF&MI), balanced 2-way ANOVA is performed after multiple testing corrections (FWER= 5%).

4.4 RESULTS

4.4.1 IMMUNOSIGNATURING OF DISEASE VS CONTROLS

11 plasma samples of subjects having poorly controlled type 2 diabetes, 7 plasma samples having CHF&MI and 7 plasma samples of subjects having both T2D and CHF & MI related symptoms and 11 plasma samples of healthy controls were run in duplicate in CIM 10k V.1 peptide microarrays. Technical replicate with Pearson's correlation <0.90 were discarded. For each disease, approximately 200 features are selected after Welsh t-test with multiple testing corrections.

Table 4-1 shows the classification accuracy; specificity and sensitivity for all the three complications when compared with common controls using Naïve Bayes leave one out methodology. For type 2 diabetes, 178 peptides passed the t-test at p <0.0025 and 93.75% accuracy was achieved. For cardio vascular complications, 239 at p < 0.0005 informative peptides were selected after feature selection method yielding 95% sensitivity. For T2D and (CHF&MI), 234 peptides were selected at p < 0.01 yielding only 81.25% accuracy. Naïve Bayes treats features as completely independent source of information, which has advantages for a system like immunosignaturing where there is no biological connection across features. In order to test how different are each disease from each other, we tested all the three complications for their ability to discern from each other.

Table 4-1: Performance measures for Immunosignaturing of disease vs controls

Conditions	**T2D Vs Controls**	**CHF, MI Vs Controls**	**T2D & CHF, MI Vs Controls**
# of peptides	**178 (p < 0.0025)**	**239 (p < 5e-4)**	**234 (p < 0.01)**
Accuracy	**95 %**	**93.75 %**	**81.25 %**
Specificity	**89.9 %**	**88.9 %**	**71.4 %**
Sensitivity	**100 %**	**100 %**	**88.9 %**

4.4.2 IMMUNOSIGNATURING OF DISEASE VERSUS EACH OTHER

Immunosignaturing of disease versus each other: Features that distinguish each disease from every other were identified using corrected t-test and presented in **Table 4-2**. The accuracy of Naïve Bayes classification using leave one out cross validation is shown for each comparison along with the p-value cutoff for each comparison. Notably at $p < 0.0005$, the difference between T2D and cardio vascular disease were distinct with 100% accuracy; at $p < 0.0005$ the difference between cardio vascular immunosignature and T2D and (CHF & MI) immunosignature were distinct with 92.86 % accuracy. At $p < 0.0025$ the difference between T2D and type 2 diabetes, (CHF&MI) yielded only 83.33% accuracy using Naïve Bayes leave one out cross validation. The three complications were fairly distinct from each other; we then compared how many unique peptides for each complication are similar and whether unique peptides for individual complications are contained in the combined complication of T2D and (CHF, MI). We selected 42 peptides from ANOVA (p-value <0.001) **Figure 4-1** shows principal component analysis showing 65% of the variance.

Table 4-2: Performance accuracy of identifying complications from each other

Conditions	T2D Vs (CHF, MI)	T2D Vs (T2D & (CHF, MI)	(CHF, MI) Vs (T2D & CHF, MI)
# peptides	**148 (p < 5e-4)**	**152 (p < 0.0025)**	**157 (p < 0.0005)**
Accuracy	**100 %**	**83.33 %**	**92.86 %**

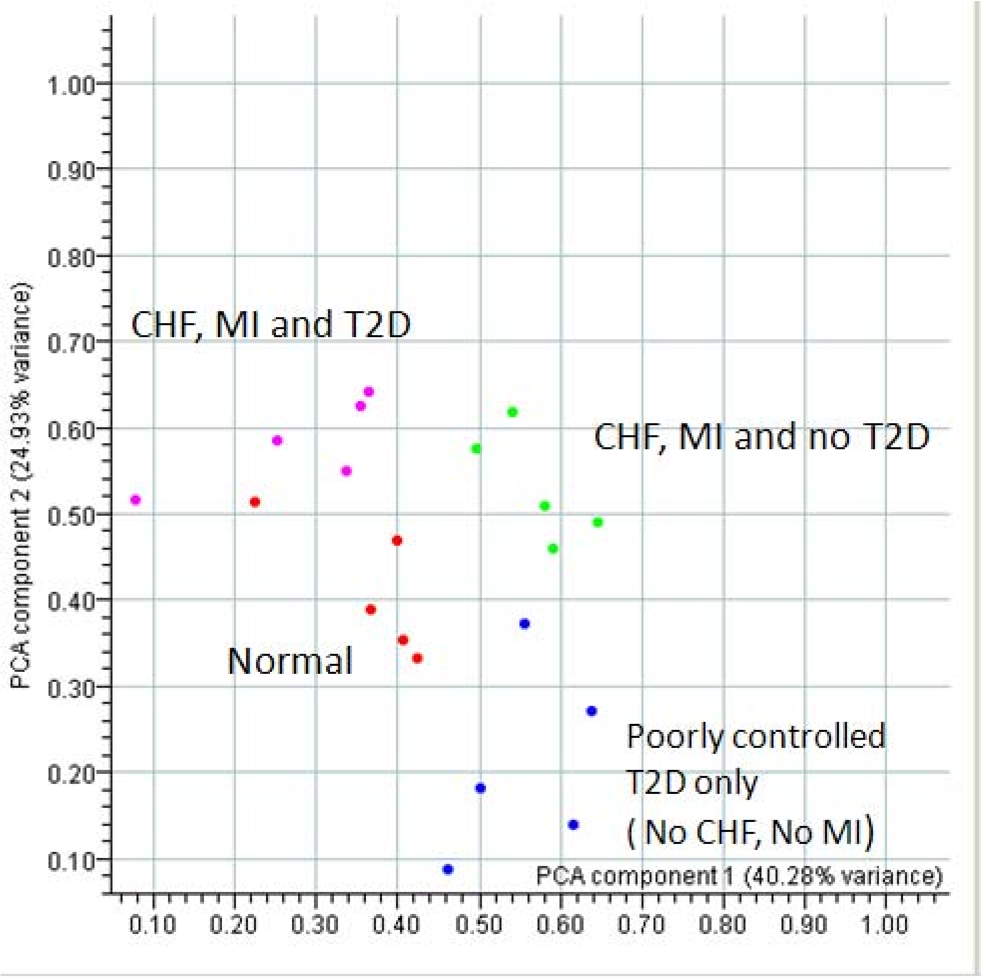


Figure 4-1: PCA showing 42 peptides distinguishing three complications from controls

4.4.3 SIMILARITIES AMONG COMPLICATIONS

The top 200 peptides that distinguished each disease from controls were identified. The probability of seeing r or more peptides by chance common among 200 from each disease is given by equation 1 discussed in chapter 2. For 10 or more peptides in common, this probability is less than 1% indicating non random or significant similarity between the two groups. **Figure 4-2** shows 18 peptides that are common in T2D and (CHF, MI) immunosignature showing a significant similarity in unique peptides that are specific to individual complications. Between T2D & (CHF, MI) and individual complications (T2D and (CHF, MI)) there are significant overlapping peptides (13 and 34) respectively. Overall each complication has significant overlap of peptides indicating similarity in each complication unique immunosignature. The unique peptides for the combined complication of T2D and (CHF, MI) had 82.5% (165) peptides non-overlapped with the two individual complications indicating a new immunosignature that is non-dependent on individual complications. We tested whether there is an interaction between the two complications T2D and (CHF, MI).

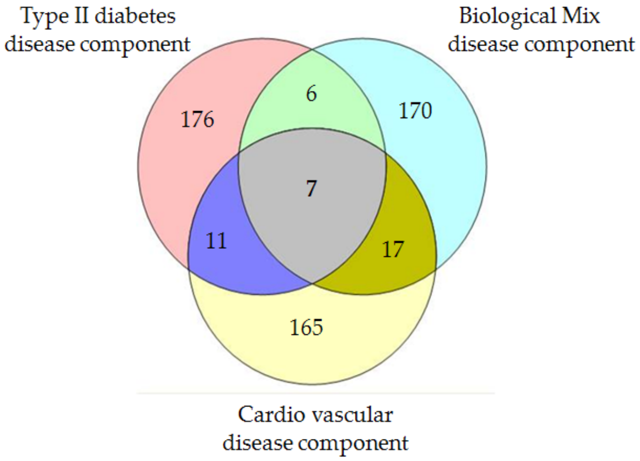


Figure 4-2: Venn diagram showing top 200 peptides by t-test for each complication

4.4.4 INTERACTION BETWEEN DISEASE FACTORS

In order to test whether there is a significant interaction between the two disease factors (T2D and CHF, MI), we performed balanced 2-way ANOVA. Each cell among a total 4 cells contained 7 samples indicating presence/absence of each factors. **Figure 4-3** shows Venn diagram indicating 245 peptides that passed the test at p< 0.001 indicating significant interaction between T2D and (CHF, MI) immunosignatures. We then compared whether the combine complication immunosignature is similar to the physical mixture of individual disease plasma 1:1 mixture sample.

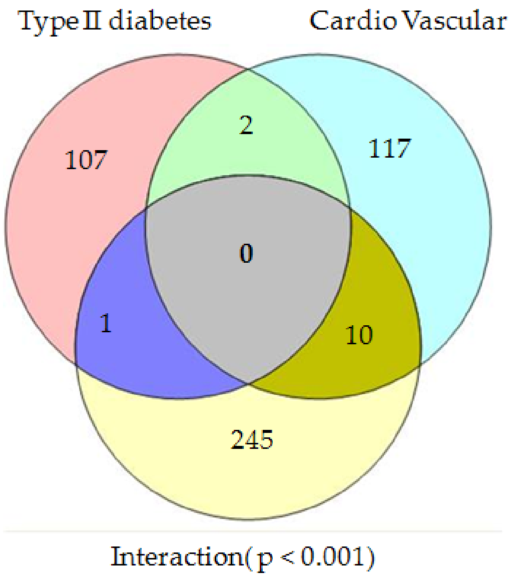


Figure 4-3: Interaction between disease factors T2D and CHF, MI

4.4.5 COMPARISON OF PLASMA MIXTURE AND BIOLOGICAL MIXTURE IMMUNOSIGNATURE

We ran 4 plasma samples of subjects having T2D and (CHF, MI) on the CIM 10 K V.2 array in duplicates (biological mixture). 4 plasma samples each of T2D and CHF, MI are mix 1:1 to create 4 plasma mixture samples. Differential peptides were selected for biological mixture and plasma mixture population compared to healthy controls. **Figure 4-4** A shows a Venn diagram of number of peptides that are common at p<0.05 significant level between biological mixture peptides and plasma mixture peptides. At this significance level, the plasma mixture has >3 times more peptides than that of biological mixture immunosignature, although there are 70 overlapped peptides, the immunosignature of biological mixture and plasma mixture were quite distinct. **Figure 4-4** B shows a principal component analysis for 134 peptides selected from a t-test between the two groups at p<0.025. The two classes can easily be separated by a linear line. **Figure 4-4** C shows a Venn diagram indicating the peptides selected at p<0.01 for T2D and (CHF, MI) immunosignature, and at p<0.05 for plasma mixture immunosignature from the t-test between the individual group and common controls. 67% of the plasma mixture immunosignature non-overlapped with individual complications (112 / 167) indicating even the plasma mixture immunosignature cannot be obtained by individual components immunosignature.

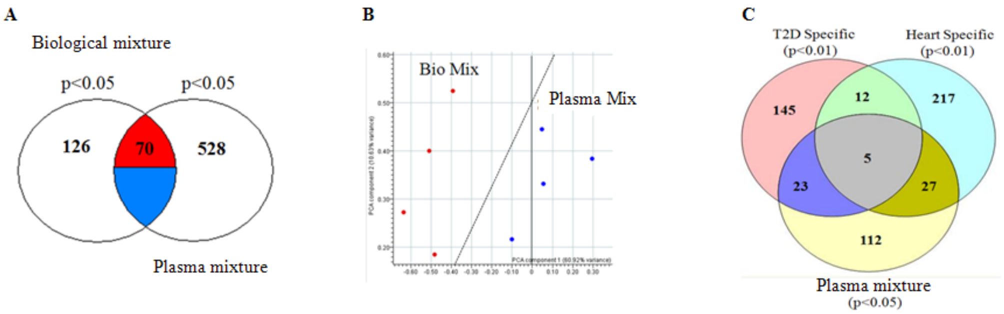


Figure 4-4: Comparison of biological mixture and plasma mixture immunosignature

4.5 CONCLUSION AND DISCUSSION

We tested how the antibody profile interacts when an individual has two correlated disease using immunosignaturing technology. For a single disease course the profile of antibodies binding to random sequence peptides is unique, consistent and reproducible but when an individual suffers from another complication, there may or may not be interaction among the two antibody profiles. It is important to know how the biosignatures changes when a person has multiple diseases so that appropriate diagnostic measures can be undertaken if multiple diseases biosignatures are interactive to reduce false positive during diagnostics. In this work, we undertook two complications which are highly correlated; T2D and (CHF, MI) studied their immunosignature individually at first and then studied subjects having both the complications. Features were selected that differentiate subjects of T2D and (CHF, MI) from healthy controls. The immunosignature was consistent and reproducible and for both the cases classification sensitivity was 100% using Naïve Bayes leave-one-out methodology. The profiles of both T2D and (CHF, MI) when taken on an individual basis and compared to healthy controls produced a distinct immunosignature. But the immunosignature of subjects suffering from both the complications combined (T2D & (CH, MI)) were less consistent and the classification accuracy was lesser (81.25%) in predicting test cases compared to the immunosignature of individual complications. Since there was interaction between the antibody profiles between T2D and (CHF, MI) antibody profile, the net immunosignature of combined complication were less clear from control group immunosignature. The peptides unique for the combined complication had only 17.5 % overlapping with the individual complication indicating a new immunosignature due to interaction effects. We formally tested for interaction between the two disease factors by 2-way ANOVA, significant interaction was found at $p< 0.001$. This interaction clearly suggests that the significant effect of immunosignature in the combine complication is due to fusion of antibody profiles between T2D and (CHF, MI) immunosignature. To test if this biological fusion can be obtained physically by mixing the antibody sources in 1:1, we tested plasma mixtures immunosignature of T2D and (CHF, MI) against subjects having both (T2D & CHF, MI). The plasma mixture immunosignature was completely different from the immunosignature of subjects having combined complications. When we tested for immunosignature of two immunizations in mice KLH and PR8 immunization, the net signature was simply an addition of the individual immunization, this is due to the fact the profiles of antibody of KLH and PR8 do not interact and these two immunizations are not correlated. So we suspect that whenever subjects will have more disease which are non-correlated, their net immunosignature would still contain individual disease immunosignature but not when the subject have two diseases which are correlated. Overall, we conclude that immunosignature for single disease is unique, consistent and reproducible but when two diseases which are interactive or correlated, the net antibody profile of the combine complication changes and cannot be obtained by profiles of individual diseases.

CHAPTER 5 DECIPHER SYSTEM TO TRACE RANDOM PEPTIDES IN IMMUNOSIGNATURE TO KNOWN PROTEINS IN TYPE 1 DIABETES

5.1 ABSTRACT

Infections or chronic diseases are accompanied by a humoral immune response that activates B cells to produce antibodies. Immunosignaturing captures the antibody repertoire of an individual at a given point in time. This is done by applying antibody source of an individual over unbiased probes of random sequence peptides thereby capturing their antibody profile whether that individual is healthy, combating a pathogen or is undergoing an immune response to their own defective cells. Because the peptides used for this process are completely random, it may be that portions of actual epitopes are present within the 17mer peptide sequence. As only 10,000 different peptides are used, it is unlikely that an extended identity exists. However antibodies have a unique way of imposing specificity. We showed an unusual dipeptide inversion is possible when antibodies bind to peptides that carry some sequence similarity to the eliciting antigen. There are likely many more ways that antibodies recognize epitopes in random peptide sequence space. Although there are biochemical methods of using the immunosignaturing peptides to capture disease-specific antibodies, these methods are quite time consuming. It would be beneficial to capture as much sequence information as possible directly from the random peptides themselves. To this end we examined serum from patients with type I diabetes (T1D). There are 3 well-known autoantigens for type 1 diabetes, but others are postulated to

exist. We used an informatics approach to identify eight different protein candidates for T1D that are biologically consistent with the disease using only the 10,000 peptides on our immunosignaturing microarray. If the correspondence of random sequence peptides from immunosignature can be made to the specific epitopes of the antigens, random sequence peptides could potentially be used to discover novel antigens for diagnostics as well as in drug discovery. We tested our ability to bioinformatically relate the peptides in the immunosignature of T1D to known and suspected autoantigens in T1D.

We found immunosignaturing technology can separate T1D from controls with >90% specificity with 679 informative peptides. These differential peptides can be significantly mapped to 8 known and suspected antigens in T1D with <0.0001 false discovery rate and not to known random proteins (negative controls). In the mapping process of T1D immunosignature differential peptides to known and suspected antigens, there were 210 peptides whose intensities were higher and 479 peptides whose intensities were lower compared to controls. We showed that not only high binder peptides are mapped to the 8 antigens but also the low binder peptides are mapped against some of the antigens (3).

Using GUITOPE, an epitope matching tool for mapping random sequence peptides to protein(s) we found significant mapping of our differential random peptides of immunosignature to known and suspected antigens in T1D. In the mapping process, we bioinformatically predicted the parts of 8 autoantigens in T1D where our differential random peptides are aligned (predicted epitopes).

We validated our predicted epitopes by spotting mapped regions of autoantigens in the form of 20-aminoacid non overlapping peptides on our arrays. We found significantly higher binding to our mapped predicted epitopes compared to random regions on the same autoantigens.

We also tested the random peptides intensities which are mapped to IA-2 and GAD-65 antigens in T1D subjects having high/low titers of respective protein. We found that these mapped random peptides showed significantly higher intensities in high titer subjects compared to low titer T1D subjects (p value < 0.001)

5.2 INTRODUCTION

Biomarkers are a promising way for detection and diagnosing disease and could be the most effective means to improve human health (Weston and Hood 2004). Towards efficient detection of biomarkers, hunt of high specificity diagnostics biomarkers for a particular disease is being performed by various methods and technologies. A persistent problem with this approach is that in reducing biomarkers to a few best candidates overtraining leads to fail one of the markers in large diverse populations (Kiehntopf, Siegmund, and Deufel 2007). One of the putative solutions to this problem is to use more biomarkers for diagnosing the disease of interest and that are less affected by genetics factors. We propose immunoglobulin molecules may solve this purpose. During an infection or chronic disease, antibodies are specifically amplified and are present in differentially abundant amount compared to a healthy person. It has been shown that antibodies are produced during body humoral immune response which can distinguish non self antigens, altered antigens in autoimmunity like T1D (Ada and Jones 1986; Brichory et al. 2001; Cox et al. 1994; Sreekumar et al. 2004; Stockert et al. 1998).

In order to observe the differential pattern in the antibody repertoire, we developed a machine-readable platform known as immunosignaturing, which profiles antibodies through random sequence peptides (Brown et al. 2011; Chase, Johnston, and Legutki 2012; Legutki et al. 2010; Restrepo et al. 2011; Stafford et al. 2012). Immunosignaturing works for any isotype and has detected cancer, infectious disease and autoimmune diseases. In order to control variability and cost, the random peptide microarray is commercially printed. Immunosignaturing has shown a high reproducibility (>0.95) between technical replicates.

Immunosignaturing has made clear distinctions between disease and healthy controls, but we had not tested the idea whether the random sequence peptides have any detectable similarity to the real antigens involved in the disease. We tested an autoimmune disease in children. T1D is an autoimmune disorder in which the immune systems produces autoantibodies against self proteins like IA-2, GAD-65, ZnT8, Insulin, ICA-69 (Hawa et al. 2000; Leslie, Atkinson, and Notkins 1999; Urakami et al. 2009; Stayoussef et al. 2011; Lan et al. 1996; Bonifacio et al. 2000). It has also been shown that there is a significant suppression of the immune system in the T1D autoimmunity (Elo et.al 2010). This suppression can be general due to down regulation of genes like the HLA class I and II genes etc to prevent the ongoing autoimmunity (Elo et.al 2010). We hypothesize that there might be some already existing autoantibodies against the self proteins in low concentrations at the normal state and when the autoimmunity is induced due to a specific autoantigen, this might activate the T regulatory cells in a constrained manner, there by turning the B cell off to reducing the already existing autoantibodies in the T1D people relative to healthy people (Colnaghi, Menard, and Porta 1977). ELISA type immune assays have been used to detect increases in autoantibodies to certain antigens, but as far as we know, have not been used to detect decrease in autoantibody levels relative to healthy controls.

The immunosignature technology may be more facile than the current techniques in detecting both up regulation and down regulation of antibodies from the binding of random sequence peptides on our array. **Figure 5-1** explains this phenomenon where a self protein/antigen causes an immune response that can lead to either an activation of B cells to reproduce and produce more antibodies specific to epitopes of an antigen or can regulate T-regulatory cells which can inhibit B cells in producing an already existing antibody against a specific antigen leading to lack of binding of these antibodies on our random sequence peptide array. We have earlier shown that an antibody raised against a particular epitope can also recognize random sequence peptide sharing significant identity or similarity with the epitope sequence (Halperin, Stafford, and Johnston 2011).

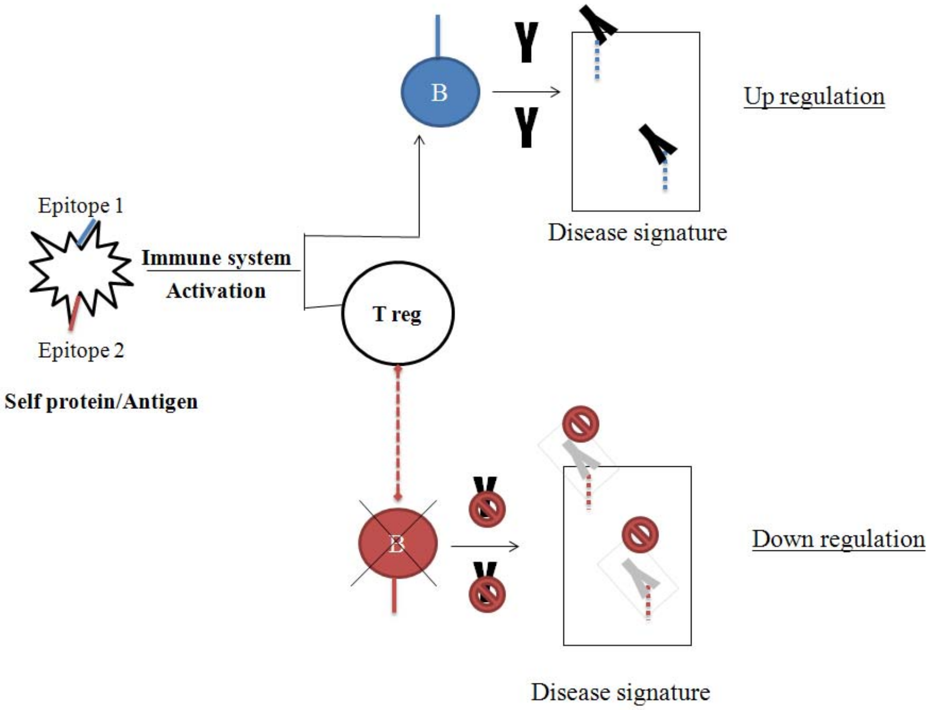


Figure 5-1: Pathway showing how a self protein/antigen can lead to up-regulation and down-regulation of immunosignatures on the random sequence peptide microarray.

In order to test our ability to map random sequence peptides to the real antigens involved in the disease, we attempted to bioinformatically relate the immunosignature of T1D to 8 known antigens. We validated these predicted peptides of by correspondence to Radio Immune Precipitation Assay (R.I.P) titers for IA-2 and GAD-65. We also mapped these predicted peptides onto the real proteins and spotted the predicted peptides of the protein in the form of short non overlapping 20 amino acid peptides onto our array to cross validate our hypothesis.

5.3 METHODS

5.3.1 MICROARRAY

CIM 10k V1 array is a microarray containing 10,000 20 amino acid random sequence peptides. These peptides create a covalent attachment via a maleimide reaction to the NH_3 terminal sulfur of cysteine (Legutki et al. 2010)**.** The CIM 10K microarray is available publically at www.peptidemicroarraycore.com

5.3.2 SAMPLE PROCESSING

Sera samples from T1D and controls were stored at -80oC. Samples were aliquoted and refrozen at -20^{o}C. Samples were diluted at 1:500 for IgG detection detection in sample buffer (1xPBS, 0.5%Tween20, 0.5% Bovine Serum Albumin (Sigma, St. Louis, MO)) and run to the CIM 10K array. Tertiary antibodies were detected with 5nm Alexafluor 555-labeled streptavidin (Invitrogen, Carlsbad, CA). For IgG antibodies detection 5nM biotinylated anti-human secondary antibody (Novus anti-human IgG (H+L), Littleton, CO) is used. After processing the microarray was scanned and converted to tabular data as in (Legutki et al. 2010). Median foreground signal was used as the value best representing binding of antibody to the peptide.

5.3.3 SAMPLES

The Center for Innovations in Medicine, Biodesign Institute, Arizona State University has an existing IRB 0912004625, which allows analysis of blinded samples from collaborators.

T1D and controls: This set contains 80 sera samples (39 diabetic children from age 6 to 13 and 41 samples of age-matched healthy controls). Radio immune precipitation (R.I.P) titers for IA-2 and GAD-65 for 40 samples of T1D were available.

5.3.4 DATA ANALYSIS

The raw data were imported to Gene Spring 7.3.1 (Agilent, Santa Clara, CA) after which it was median normalized per array and transformed to log10 scale. We used Welsh t-test with family-wise multiple error correction (FWER) for feature selection. For multiple factor interaction, we used a balanced 2-way ANOVA with FWER. All p values presented are after FWER correction. We used GUITOPE, an epitope matching tool to find degree of similarity of peptides to given protein sequence with 1000 iterations per comparison (Halperin et al. 2012). The peptides having >1 score per alignment length was initially selected to plot against protein in GUITOPE. The peptides which indicated higher significant matching against the specific protein than the equivalent number of random peptides selected from the library (1000 iterations) were finally selected and their false positive rates were calculated. The false positive rate obtained from GUITOPE for these peptides represents number of times equal number of peptides from the library showed the same level matching. For classification, Naïve Bayes and leave one out cross-validation was used due to its outperformance in immunosignaturing data (Kukreja, Johnston, and Stafford 2012). Classification was performed in open source JAVA software WEKA (Hall et al. 2009).

5.3.5 ANTIGEN SELECTION

Table 5-1 shows 8 suspected T1D antigens and 8 random proteins have been selected for a match against T1D specific peptides. The sequence were downloaded from NCBI.

Table 5-1: List of antigens selected for decipher

Antigens/self protein for Type 1 diabetes	Random Protein
IA-2	ACAA1
IA-2 β	ACOT1
ICA-69	ACLS1
Insulin	ACSBG1
GAD-65	AOX
GAD-67	ALDH2
GLUT-2	CoA
ZnT8	CPT1A

5.4 RESULTS

5.4.1 IMMUNOSIGNATURING OF T1D AND CONTROLS

We ran 40 samples each of T1D and controls and selected peptides that are significantly different in T1D compared to controls. At p<0.0001 false discovery rate, we observed 689 differential peptides between T1D and common controls. **Table 5-2** shows the classification accuracy, specificity and sensitivity for these differential peptides. We were able to achieve >90% specificity using Naïve Bayes classification algorithm leave one out cross validation though immunosignaturing. The distribution of the 679 differential peptides was skewed with only 210 peptides were high binders (peptides whose intensities are higher) and while 479 peptides were low binders (peptides whose intensities are lower) compared to controls. We then asked as to how many peptides bioinformatically can be mapped to the known antigens in T1D.

Table 5-2: Performance measures of differential peptides in Immunosignaturing of type 1 diabetes

	T1D Vs Controls (IgG)
No. of features	689 (p < 0.0001)
Accuracy	88.75 %
Specificity	92.3 %

5.4.2 BIOINFORMATICS DECONVOLUTION OF T1D IMMUNOSIGNATURE

We first considered 210 high binder peptides in T1D, and bioinformatically mapped it to 8 known antigens in T1D using epitope matching tool GUITOPE. It has outperformed current existing epitope matching tools in regard finding not so near matches from complex random sequences to actual epitopes. The protein sequence for 8 known antigens was matched for sequence similarity with 210 peptides using GUITOPE. The number of iterations was set to 1000 which signifies the number of times equal numbers of random peptides are selected from a 10k peptide sequence library to match against given protein(s). Net score is calculated by the GUITOPE output parameter score and alignment length by taking this ratio. Peptides showing significant matching on a region in a protein (epitope) over equal no. of peptides from our 10,000 peptide library at p<0.0001 were selected. On average, we were able to successfully map peptides from 210 to known antigens with 1-2 possible epitopes per antigen with 3-4 peptides on average mapped to one epitope of an antigen. 7 peptides on an average were mapped against the antigen from a high binder set of peptides. **Table 5-3** shows the number of epitopes, peptides per epitope and total peptides mapped per antigen for up regulated peptides from 210 peptides over 8 known antigens in T1D. **Figure 5-2** shows the GUITOPE graphical representation of mapped peptides to the respective proteins. The red line indicates the average score obtained through selection of random peptides, the green peaks shows those regions significantly mapping peptides on a protein region. The number of peaks represents the corresponding epitopes. With the assumption of at least single antibody binding an epitope, at least 14 antibodies correspond to binding of mapped high binder peptides and 210/56*14=52 antibodies correspond to total up signature.

Table 5-3: Showing 210 high binder peptides from immunosignature mapped to 8 known antigens in type 1 diabetes.

Protein Upreg.	# Epitopes	# Peptides-E1	#Peptides E2	Total	FDR
ICA-69	2	5	3	8	<0.0001
GLUT-2	2	3	3	6	<0.0001
GAD-67	2	4	3	7	<0.0001
GAD-65	1	13	NA	13	<0.0001
IA-2 β	2	3	4	7	<0.0001
Insulin	2	3	4	7	<0.0001
ZnT8	2	4	3	7	<0.0001
IA-2	1	1	NA	1	<0.0001
Total	14			56	

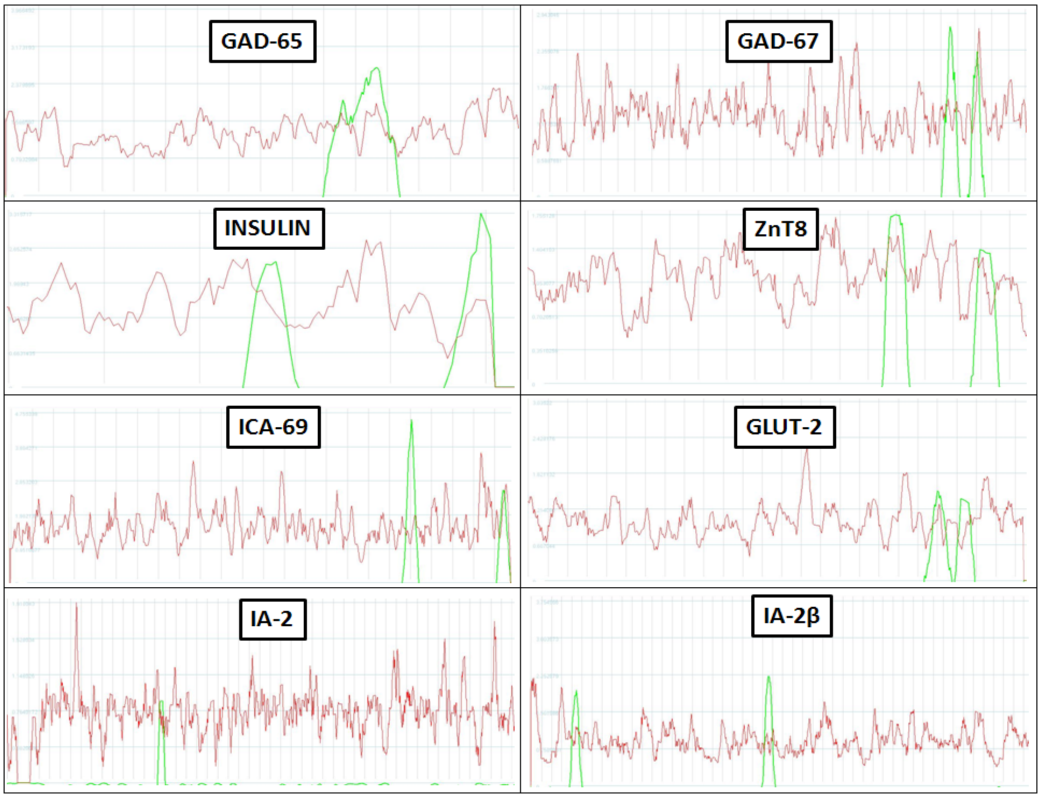


Figure 5-2: GUITOPE analysis: mapping 210 high binder random peptides to 8 known antigens

Next, we considered remaining 479 low binder peptides and mapped it against the known 8 antigens in type 1 diabetes. Our hypothesis behind this is presented in **Figure 5-1**. **Table 5-4** shows the 479 low binder peptides from the T1D immunosignature mapped to 8 known antigens in T1D. The low binder peptides, mapped to only GAD-65, ZnT8 and Insulin, which are the strongest antigens in T1D. The number of mapped peptides in the low binder set was higher compared to high binder set of peptides but the number of epitopes per antigen were similar for both up and down binder peptides. Peptides mapped to the antigens for two cases are indeed not only different since they are coming from two distinct sets, but they also mapped to epitopes that are in different locations on the same protein sequence. Hence, a particular epitope of an antigen can result in high binding of peptides while other epitopes on the same protein can result in low binding of other sets of peptides. **Figure 5-3** shows the GUITOPE graphical representation of mapped peptides to the respective proteins. The red line indicates the average score obtained through selection of random peptides and green peaks show significant mapping of these peptides on a protein region. The number of peaks represents the corresponding epitopes. With the assumption of at least a single antibody binding an epitope, at least 5 antibodies correspond to binding of low binder mapped peptides and 479/81*5=30 antibodies corresponds to total of down immunosignature.

Table 5-4: Showing 479 low binder peptides from immunosignature mapped to 8 known antigens in type 1 diabetes.

Protein Downreg	**# Epitopes**	**# peptides-E1**	**#Peptides E2**	**Total**	**FDR**
ICA-69	-	-	-	-	>0.999
GLUT-2	-	**-**	-	-	>0.999
GAD-67	-	-	-	-	>0.999
GAD-65	2	25	17	42	<0.0001
IA-2 β	-	-	-	-	>0.999
Insulin	2	**9**	14	23	<0.0001
ZnT8	1	16	**-**	16	<0.0001

IA-2	-	-	-	-	>0.999
Total	5			81	

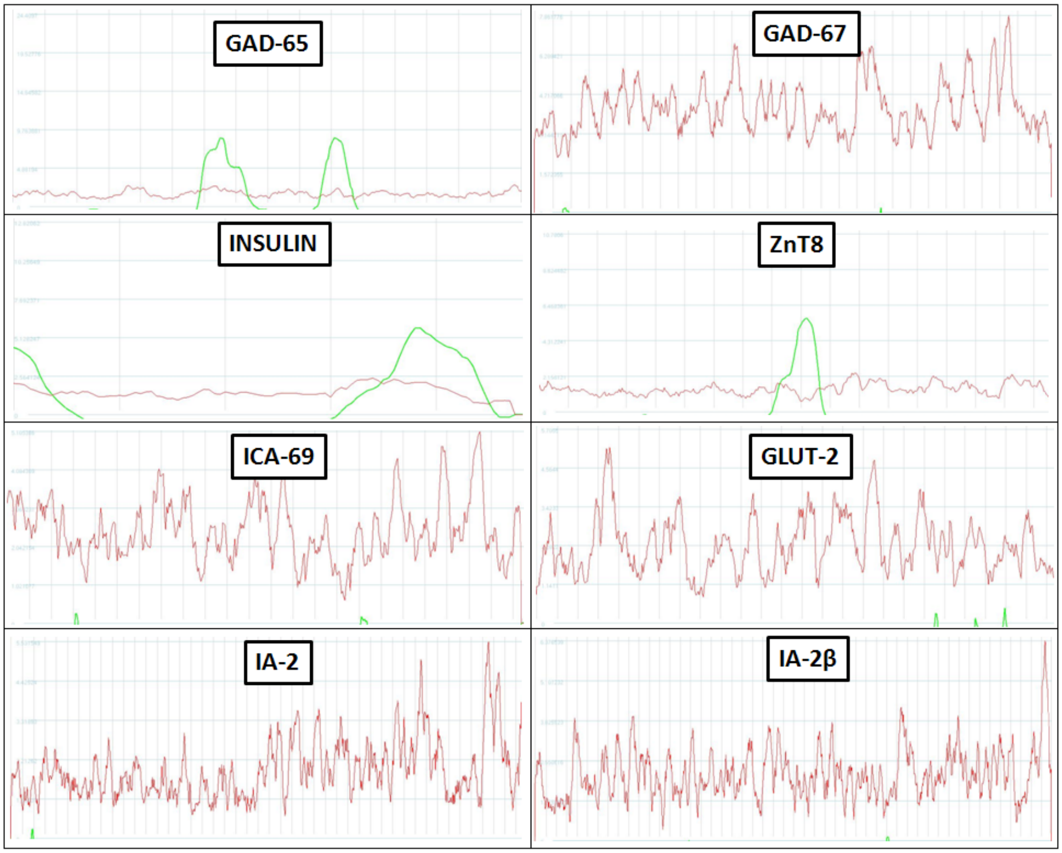


Figure 5-3: GUITOPE analysis: mapping 479 low binder random peptides to 8 known antigens

In order to set up a null control and method verification, we mapped differential peptides (both high binders and low binders) against 8 randomly choosen proteins from the human proteome, upon post selection we found some of the proteins are related to enzymes and none of them are involved in antibody production or T1D pathway. **Table 5-5** shows the number peptides that were mapped against set of 8 random proteins, which validates our method. **Figure 5-4** shows GUITOPE graphical representation of the mapped peptides to respective proteins. The red line indicates the average score obtained through selection of random peptides and green peaks shows significantly mapping of these peptides on a protein region. The number of peaks represents corresponding epitopes. For random peptides, no significant mapping was observed. In order to validate and verify our mapped peptides against known antigens, we did an orthogonal measurement by measuring intensities of mapped random peptides from T1D subjects sera.

Table 5-5: Showing 679 differential peptides from immunosignature of T1D mapped to 8 random protein/antigens.

Protein	**# Epitopes**	**Total**	**FDR**
AOX	-	-	>0.99
ACAA1	-	-	>0.999
ACSBG1	-	-	>0.999
ACOT1	-	-	>0.999
CPT1A	-	-	>0.999
ALDH2	-	-	>0.99
CoA	-	-	>0.99
ACLS1	-	-	>0.99

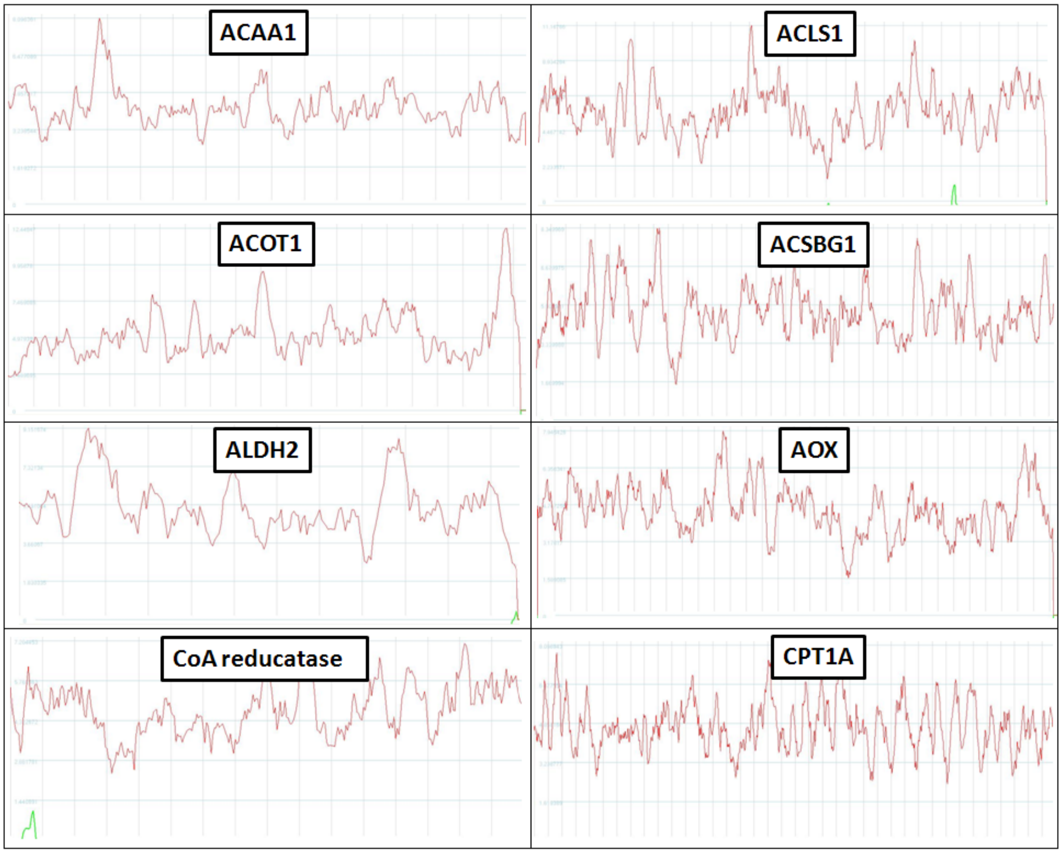


Figure 5-4: GUITOPE analysis of 689 peptides to 8 random proteins

5.4.3 ORTHOGONAL MEASUREMENT

The GAD-65 titers using radio immuno precipitation assay had been independently measured on each of the T1D samples (Burbelo, Hirai et al. 2010). We asked if there is any correspondence between 13 mapped peptides intensities against GAD-65 and R.I.P titers. **Figure 5-5** shows a box plot of average of 13 peptides intensities on high and low titer GAD-65. There was a significant difference between the mean of the two groups (p <0.001).

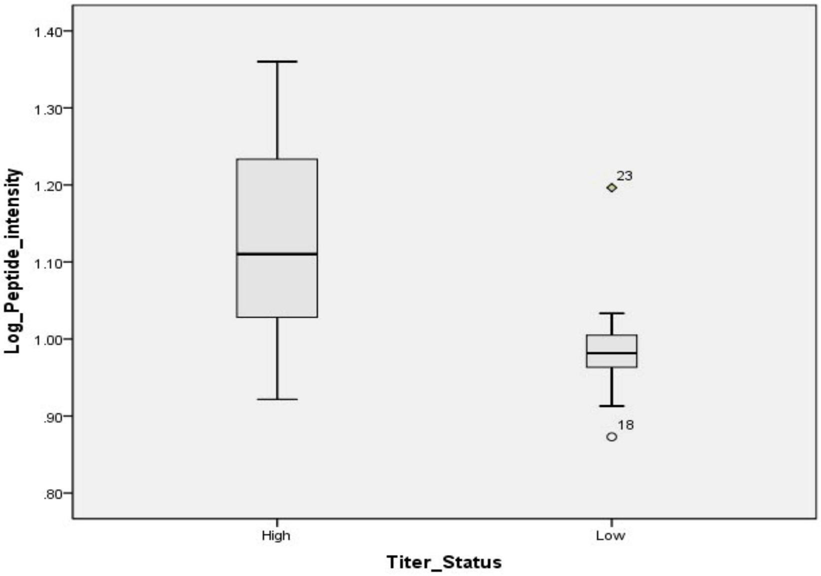


Figure 5-5: Box plot showing average of 13 peptide intensity on 2 groups of T1D samples (high and low GAD-65 R.I.P titer.

5.4.4 VALIDATION OF PREDICTED PEPTIDES

Towards the goal of supporting the assignment of epitopes for the immunosignature, we mapped the random peptides specific to IA-2 and GAD-65 to the real protein and select 17 amino acid non overlapping peptides of IA-2 and GAD-65 proteins and spotted them on our arrays. As a negative control, we selected other non mapped peptides on the same proteins. **Figure 5-6** (A) shows the binding of T1D subjects on bioinformatically mapped peptides and non mapped peptides on GAD-65 protein. The predicted epitopes shows a higher binding for the predicted epitopes compared to the non-predicted ones for each individual T1D sample. **Figure 5-6** (B) shows the average high binding of T1D subjects compared to control subjects on selected GAD-65 peptides. **Figure 5-6** (C) shows the similar pattern for IA-2 protein where only 1 T1D subject show higher binding towards non predicted epitopes of IA-2 while the rest subjects shows that mapped epitopes have high binding than the non mapped for each T1D subject. **Figure 5-6** (D) shows the T1D and controls binding on down regulated peptides of GAD-65, where controls subjects show higher binding than the T1D subjects.

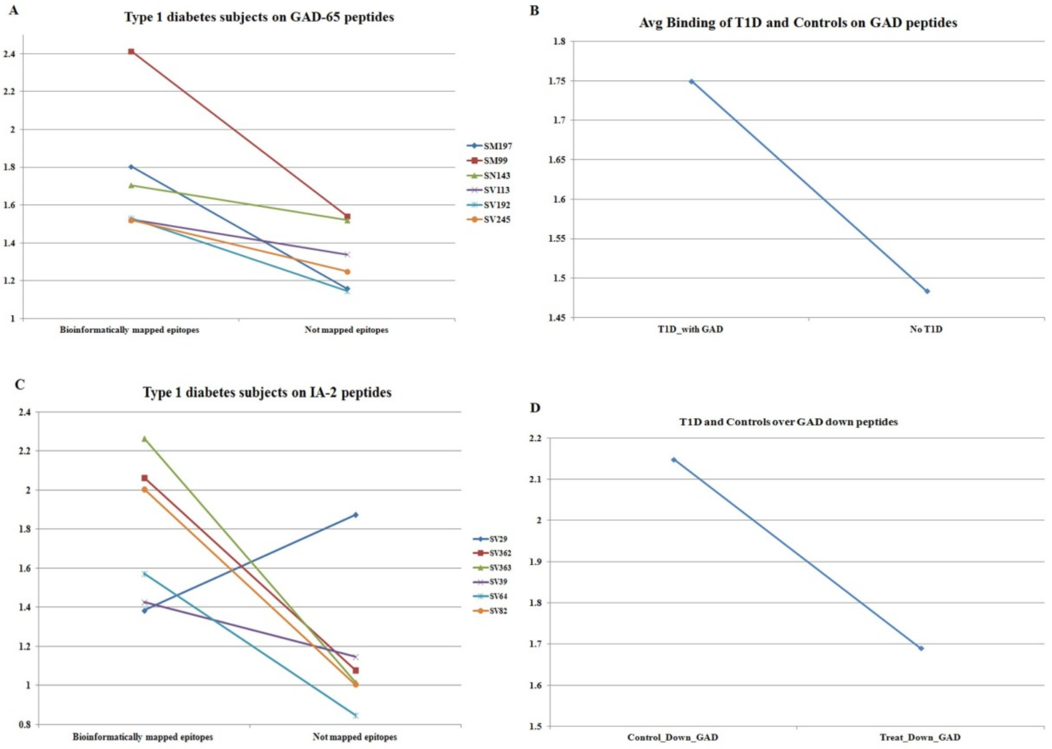


Figure 5-6: Testing of bioinformatically predicted epitopes on GAD-65 and IA-2 protein

5.5 CONCLUSION AND DISCUSSION

Immunosignaturing technology profiles antibodies during infection or chronic disease through interaction with random sequence peptides. Antibodies raised in response to a particular infection are captured by random sequence peptides having sequence similarity with epitopes of the original antigen. We have bioinformatically deciphered the information from random sequence peptides in case of T1D. We ran 80 samples (40 each T1D and controls). We found 679 differential peptides at p<0.001(FWER) that were significantly different in T1D compared to controls. Out of these, 210 peptides were high binders and the rest were low binders. The significant part of the down immunosignature in T1D subjects is consistent with the gene level study done in a systematic differential expression on 520 probes which also had a high significant down regulated gene networks involved in T cell receptor and insulin signaling and antigen presentation (Elo et.al 2010).

We then hypothesized that any antigen can potentially lead to either high or low peptide intensities in an immunosignature. Presence of an antigen during an autoimmune disorder like T1D can illicit immune responses and can either stimulate B cells to produce more antibodies against a particular epitope of self antigen which would lead to high binding of peptides in immunosignature, or it can stimulate T regs, for example, which can shut down B cells and hinder the already present autoantibodies against self protein and will lead to low binding of peptides of the immunosignature. This is based on the fact that the immune system already has autoantibodies present against the self protein but these are too low in concentration to elicit any autoimmune damage or being able to detect through normal ELISA. One of the other plausible reasons of the low binding of peptides in T1D might be a protective defense mechanism in autoimmunity to shut some regulatory biological pathways to prevent subsequent attack.

We used the immunosignaturing technology to examine whether antibody profiling through random sequence peptides have any congruity with clinical serological diagnosis of type 1 diabetes. We found using immunosignature we can separate T1D from controls by accuracy of 88.75%. We found suppression of IgG binding in 85% of the informative peptides that best separate T1D from healthy controls.

We mapped these informative differential peptides from the T1D through immunosignature to the real antigens. Towards this bioinformatics deciphering, we took 8 known antigens in T1D and mapped both high and low binding peptides to these antigens using the epitope matching tool GUITOPE. This tool has outperformed the current tools in finding epitope of the antigens based on sequence on random peptides (Halperin et al. 2012). We calibrated the GUITOPE by selecting the best candidates according to score per unit length to remove false positives. Using the same deciphering method we did not detect any significant matching against 8 random human proteins as a part of negative control.

In the mapping process, 210 high binder peptides were mapped to all 8 suspected antigens with 1 or 2 epitopes per antigen. For some predicted epitopes, there were corresponding minimum of 3 random peptides while for others there were 15 peptides. On total there were 14 predicted epitopes with corresponding 56 mapped peptides out of 210 that were mapped of high binder peptides to 8 autoantigens. If we assume that at least one antibody would bind an epitope of protein, a total of 14 antibodies can be estimated from this data corresponding to the mapped peptides and total of 52 (210/56*14) antibodies for total differential high binder peptides. Similarly while mapping low binder peptides, we found random peptides mapped to only 3 autoantigens with only 5 epitopes predicted in total. But for each predicted epitope in this case, there were significantly higher number of peptides (total of 81) that were aligned with epitopes possibly due to higher number of low binder peptides. Here 5 antibodies can be estimated for low binder mapped random peptides set but together 30 antibodies (479/81*5) can be estimated for down signature of T1D. Combining both the up and down immunosignature, a total of 82 (30+52) antibodies can be estimated from T1D immunosignature.

Various analyses have been done on the antibody titers to 3 of these autoantigens IA-2, GAD-65 and insulin titers level. Clinical observation reveals frequency of IA-2 autoantibodies varies significantly with HLA genotyping and age (Hawa et al. 2000; Leslie, Atkinson, and Notkins 1999). It is known that it is highly persists one year after diagnosis and decreases after (Notkins and Lernmark 2001). We found two subgroupings in T1D subject's immunosignature with respect to their IA-2 titers. The one group consists of low titers and the other consists of medium IA-2 levels and high IA-2 levels of titer. This is consistent with the above observation of the IA-2 high variability in T1D subjects. Being highly variable, immunosignatures of low IA-2 titers are significantly different from medium and high level IA-2 immunosignature while no significant difference exists between medium and high level titers. Large studies determining the frequency of GAD-65 titers in T1D subjects have found that the GAD-65 to be less variable and highly consistent marker (Hawa et al. 2000;

Leslie, Atkinson, and Notkins 1999). This marker is known to be present at high levels consistently in T1D subjects. We observed similar pattern in subgrouping of T1D with respect to their GAD levels. The immunosignature of low and medium level GAD-65 titers are no different but these were significantly different between from high level titers (data shown in supplementary materials).

IA-2 and GAD-65 are biologically distinct in function and their roles related to autoimmunity in type 1 diabetes. Clinically there is no observed interaction between IA-2 and GAD-65 biologically in terms of autoantibodies that are in common (Hawa et al. 2000). Using the immunosignaturing technology, we observed no interaction at 95% significance level using a balanced design two way ANOVA. This shows the consistency between antibody profiling through immunosignature and clinical observation of IA-2 and GAD-65 independence (supplementary materials).

We validated our decipher method by considering samples, which have high and low GAD-65 titer and measured 13 GAD-65 specific peptide intensities. There was a significant difference between the mean (shown in Figure 5-5) indicating an overall higher mean for high GAD-65 titers compared to low ones.

We also validated our predicted random peptides of GAD-65 and IA-2 by mapping these peptides to real proteins. We selected the parts of the GAD-65 and IA-2 protein which were mapped by random sequence peptides, and spotted them on our arrays. We found that epitopes mapped by up regulated random peptides are higher binding in T1D subjects compared to other epitopes on the same protein. We also found the same result for down regulated peptides of GAD-65 where T1D subjects had lower binding compared to the controls, hence confirming our prediction. In this work, we showed how random peptides can be mapped the real antigens but the main potential of this technique can be extended to discover unknown protein to discover suitable drug targets. Differential peptides can be mapped back to the whole proteome and list of the matched antigens can be selected and validated to discover drug targets. Here by we conclude that immunosignaturing has potential for inexpensive diagnosis and be deconvoluted back to the known antigens of a disease and also can be used to discover new ones.

Acknowledgment

We are thankful to Dr. Abner Notkins and Dr Ying Xiong, NIH, National Institute of Dental and Craniofacial Research for providing T1D and controls sample, Dr Rebecca Halperin and Krupa Navalkar, Center for Innovations in Medicine, Biodesign Institute, Arizona State University for providing useful discussions in GUITOPE. This research was supported by XyZ.

CHAPTER 6 FACTORS AFFECTING IMMUNOSIGNATURE

6.1 ABSTRACT

Immunosignaturing is a technology that captures the humoral immune responses to be observed through binding of antibodies to random sequence peptides. Unlike genomics, the immune system is believed to be more homogenous across a population. Using DNA, RNA or other genetic biomarkers poses a challenge to behave consistently across population in terms of detecting lot of variation in the people and hence prohibit these biomarkers to use universally. We propose, using immunological biomarkers may reduce some factors that vary across population like gender, age, location etc. Immunosignaturing assay profiles antibody responses of an individual at any given time based on the IgG antibody binding to random sequence peptides. An important unanswered question relative to this technology is what are the factors like age, sex, and geographical location that change or affect immunosignatures across the population? This question is of the clinical relevance while sampling populations for biosignatures discovery studies for immunosignature. If the factors can be known, an ideal sampling methodology can be formulated to screen populations based on these factors. We tested the immunosignaturing platform for its ability to resolve differences in age, sex and geographical location. We found the age is a very prominent factor that affects the immunosignature, with children under age 12 showing relatively high binding on our arrays. In contrast to the genomics markers, we found immunosignature of healthy male/female adults are

similar and no significant patterns is observed. For geographical location, we found there is a very significant difference among the US adult and Sweden adults, but no significant difference is found among immunosignature of people from different US states. Given, the fact that immunosignature is not bias against sex or local geographical location, this platform can be used very efficiently for broader screening for disease diagnostics biomarkers.

6.2 INTRODUCTION

Ideally a biomarker molecule that could diagnose disease early would be the most effective way to improve patient health (Weston and Hood 2004). Genetics biomarkers heavily rely on characterization of an individual like DNA, RNA, SNP (Lucas et al. 2009; Le-Niculescu et al. 2009; Kurian et al. 2007; Tugwood, Hollins, and Cockerill 2003). These biomarkers have been used for screening various diseases (Okamoto 2009; Hennessy et al. 2008) and this has been successful in many causes especially for single base pair diseases (Saiki et al. 1985). The basic problem with relying on the genomics biomarkers is they are very heterogeneous across population (Duffy, Evoy, and McDermott 2010; Sawyers 2008). The factors like age, sex and geographical location highly affects the traditional biomarker and thus are very prone to individual to individual variation (Rotimi and Jorde 2010; Emilsson et al. 2008; Millen et al. 2009). Were there a candidate biomarker that was relatively abundant, unaffected by age, sex, race or genetic factors, and different between healthy and sick persons and physically stable, the problem of biomarkers for effective diagnostic would be much simpler. One such candidate, we propose are immunological molecules like antibodies might solve the heterogeneity problem. The immune system is known to behave more consistent and homogenous across a population compared to gene expression. Unlike basic DNA sequence which is static, the immune system is more complex and dynamic system which constantly monitors our health status. Antibodies are produced in both infection as well as chronic diseases like Alzheimer’s disease and cancer (Ada and Jones 1986; Brichory et al. 2001; Brydak and Machala 2000; Cox et al. 1994; DiFronzo et al. 2002; Hooks et al. 1979). The signals from the antibodies during the course of any disease are amplified to a huge extent due to their ~10^{10} concentrations. We have developed a machine readable platform known as immunosignaturing which monitors the antibody binding to unbiased antigen epitopes through random sequence peptides. While we have seen clear distinction between disease and healthy controls immunosignature which were matched for age, sex and geographical location (Restrepo et al. 2011; Chase, Johnston, and Legutki 2012; Legutki et al. 2010; Brown et al. 2011), we had not tested the idea that immunosignatures might be quite similar for healthy individuals differ in age, sex and geographical location. We examined healthy individuals from of different age groups, geographical locations and sex one at a time while controlling for the other two factors.

6.3 METHODS

6.3.1 MICROARRAY

The CIM 10k V.1 is a 1-up microarray containing 10,000 random sequence 20-mer peptides attached via a maleimide reaction to the NH_3 terminal sulfur of cysteine, creating a covalent attachment (Legutki et al. 2010).

6.3.2 SAMPLE PROCESSING

Sera samples from healthy controls were stored at -80 C after which samples were aliquoted and refrozen at -20 C. Samples were diluted at 1:500 in sample buffer containing 1XPBS, 0.5% Tween20, 0.5% Bovine Serum Albumin (Sigma, St. Louis, MO) and then applied to the array using the standard immunosignaturing protocol as mentioned (Legutki et al. 2010). Antibodies were detected with 5nm Alexafluor 555-labeled streptavidin (Invitrogen, Carlsbad, CA). This tertiary antibodies bind to 5nM biotinylated anti-human secondary antibody (Novus anti-human IgG (H+L), Littleton, CO). Microarrays were scanned and converted to tabular data after which median foreground signal was used as the value of antibody peptide binding.

6.3.3 SAMPLES

The Center for Innovations in Medicine, Biodesign Institute, Arizona State University has an existing IRB 0912004625, which allows analysis of blinded samples from collaborators.

For testing age as a factor: We collected 24 sera samples of healthy children from University of Arizona Asthma study were collected and 36 sera samples of local healthy adult donors from Biodesign Institute, Arizona State University.

For testing gender as factor: We collected sera samples from healthy 7 males US adults and 14 females US adults from the Alzheimer's disease study.

For testing geographical location as factor: We collected sera samples from healthy 29 US Adults from local donors from Arizona State University and 23 samples from Sweden Tularemia study.

6.3.4 DATA ANALYSIS

The raw tabular data were imported to GeneSpring 7.3.1 (Agilent, Santa Clara, CA). Data were median normalized per array and then transformed into $\log_{10}$ scale. To get an unbiased and unsupervised view of the immunosignature, first no feature selection was performed. If there is a bias in the unsupervised methods, we conclude that there is an effect of the factor, if there is no bias in the unsupervised methods, we performed Welsh t test after multiple testing corrections (FWER=5%) to select peptides that are differential among the classes and then the analysis is performed. Line graph, scatter plot, principal component analysis and classification have been performed for both unsupervised and supervised methods. For classification, we used Naïve Bayes with the leave one out cross validation method since this algorithm has been known to work best for immunosignaturing data (Kukreja, Johnston, and Stafford 2012)

6.4 RESULTS

6.4.1 EFFECT OF AGE ON IMMUNOSIGNATURE

36 US healthy adults and 24 US healthy children were run in duplicate on the CIM 10k V.2 peptide microarrays. Technical replicates with Pearson's correlation coefficient <0.90 were discarded. **Figure 6-1** (A) shows a line graph and (B) shows scatter plot of all 10,000 peptides intensity (normalized) for adults and children. The peptide intensity distribution for the children group was of significantly much broader range than that of adults. There were more than 2000 peptides which were 2 folds higher in children compared to adults while there were less than 100 peptides which were 2 folds higher in adults compared to children.

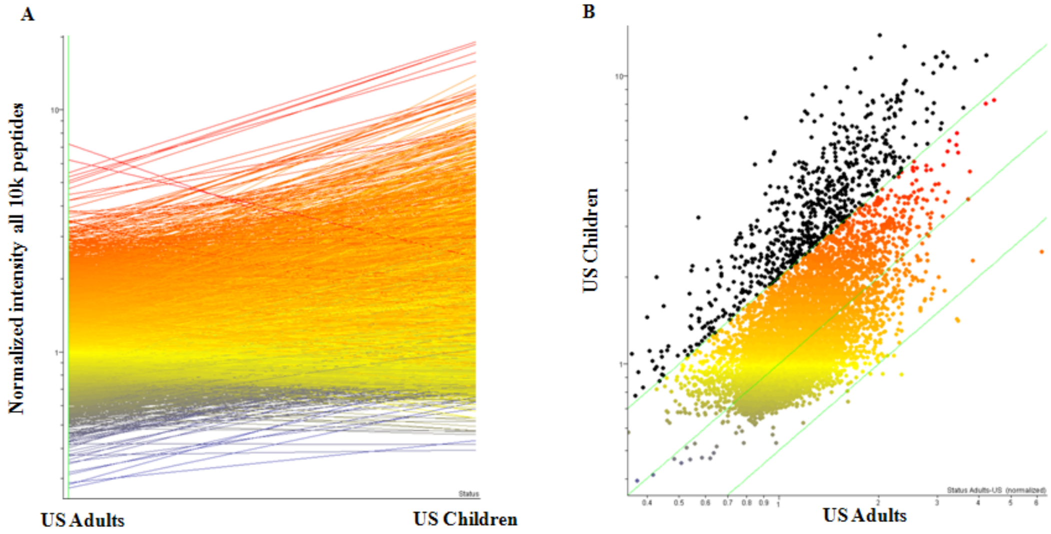
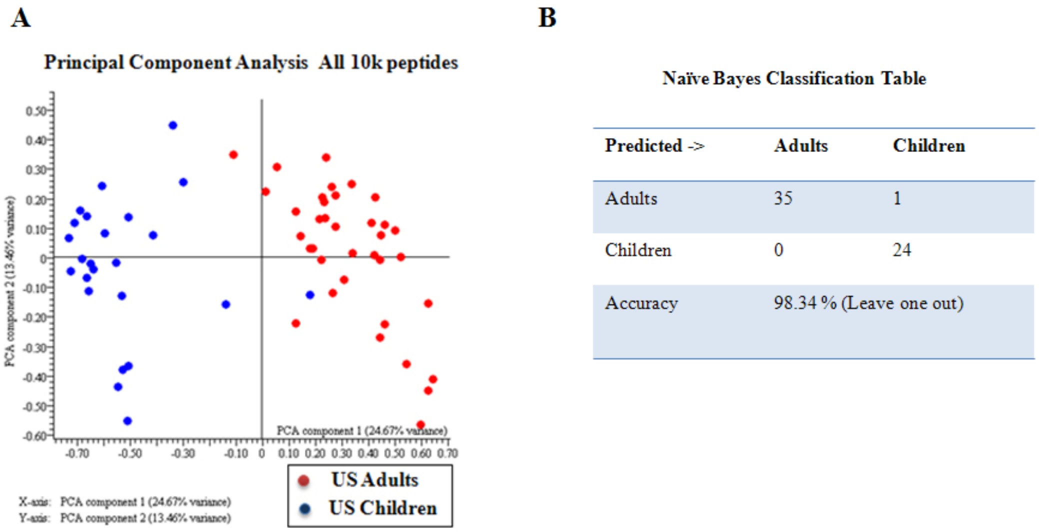


Figure 6-1 : Immunosignaturing of US Adults and Children

6.4.2 CLASSIFICATION PERFORMANCE ADULTS VS CHILDREN

When tested for separating the adults and children groups using unsupervised methods, the two classes were distinct. **Figure 6-2** (A) shows the principal component analysis for all 10,000 peptides explaining 38.13% of the variance in the population. The two classes can be very well separated with a linear line. (B) shows the classification table obtained from Naïve Bayes algorithm from all 10,000 peptides after leave one out cross validation. 98.34% accuracy was achieved with only 1 adult misclassified as children.



Naïve Bayes Classification Table

Predicted ->	Adults	Children
Adults	35	1
Children	0	24
Accuracy	98.34 % (Leave one out)	

Figure 6-2: PCA and classification table of adults vs children immunosignature

6.4.3 EFFECT OF GENDER ON IMMUNOSIGNATURE

7 US healthy male adults and 14 US healthy female adults were run in duplication on the CIM 10k V.2 peptide microarrays. Technical replicates with Pearson’s correlation coefficient <0.90 were discarded. From the unsupervised analysis, the two classes of males and females were not distinct and hence

supervised analysis was performed by selecting top 50 peptides from the welsh t-test with 5% FWER. **Figure 6-3** (A) shows line graph and (B) shows scatter plot of top 50 peptides intensity (normalized) for males and females. The peptide intensity distribution for both the groups were not significantly different and there was no specific pattern that can be observed regarding higher peptide intensities in any groups.

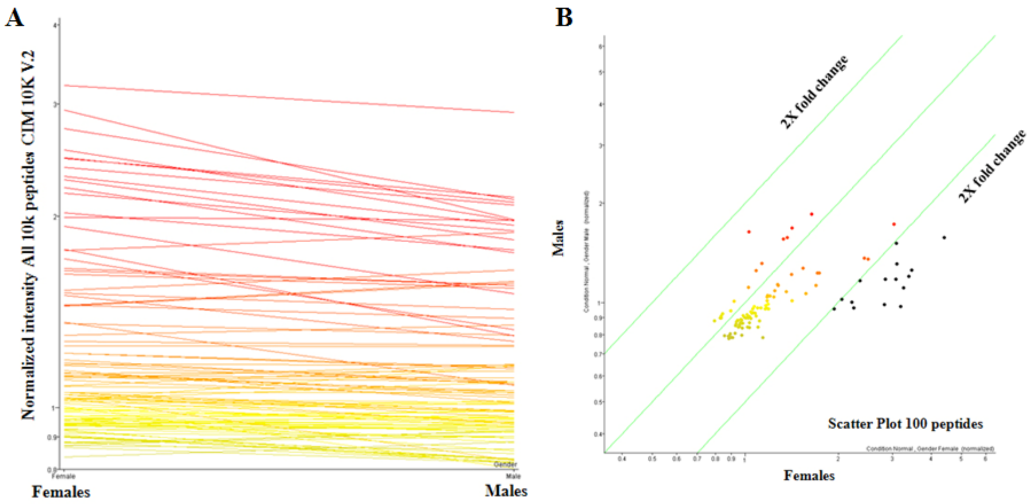
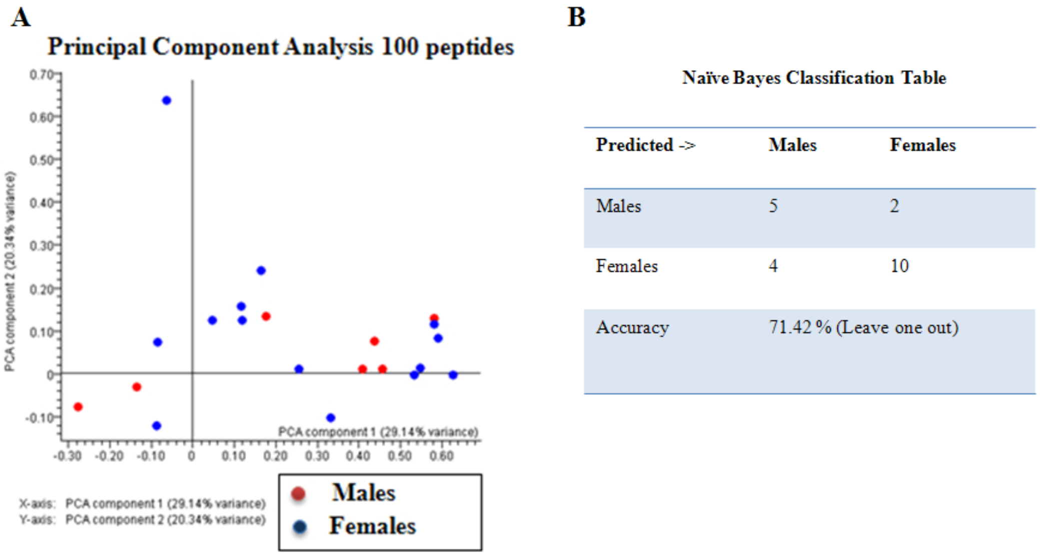


Figure 6-3: Immunosignature of males vs females

6.4.4 CLASSIFICATION PERFORMANCE OF MALES VS FEMALES

When tested for separating the males and females groups using supervised methods (top 50 peptides), the two classes were not distinct. **Figure 6-4** (A) shows the principal component analysis for top 50 peptides explaining 49.48 % of the variance in the population. The two classes cannot be separated with linear line and clustered together. (B) shows the classification table obtained from Naïve Bayes algorithm from the top peptides after leave one out cross validation. Only 71.42 % accuracy was achieved (6 misclassification among 21 samples). The difference between males and females immunosignature were not clear and subtle.



Naïve Bayes Classification Table

Predicted ->	Males	Females
Males	5	2
Females	4	10
Accuracy	71.42 % (Leave one out)	

Figure 6-4: PCA and classification table of males vs female immunosignature

6.4.5 EFFECT OF GEOGRAPHICAL LOCATION ON IMMUNOSIGNATURE

29 US healthy adults and 23 Sweden healthy adults were run in duplication on the CIM 10k V.2 peptide microarrays. Technical replicates with Pearson's correlation coefficient <0.90 were discarded. **Figure 6-5** (A) shows line graph and (B) shows scatter plot of all 10,000 peptides intensity (normalized) for US and Sweden adults. The peptide intensity distribution for Sweden adults group spanned significantly much broader range than that of US adults. There were more than 3000 peptides which were 2 folds higher in Sweden adults compared to US adults while there were less than 50 peptides which were 2 folds higher in US adults compared to Sweden.

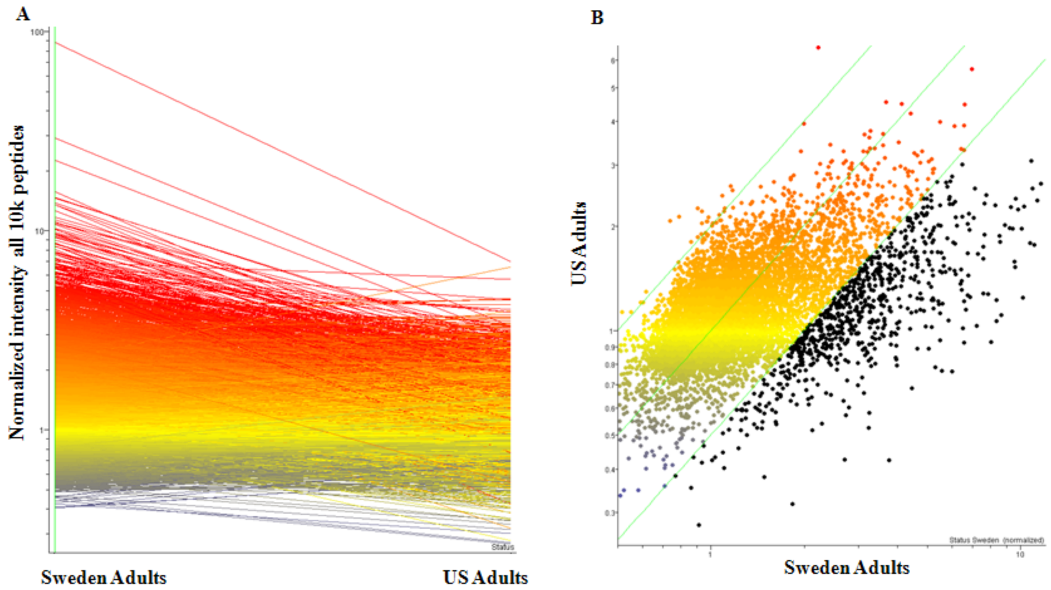
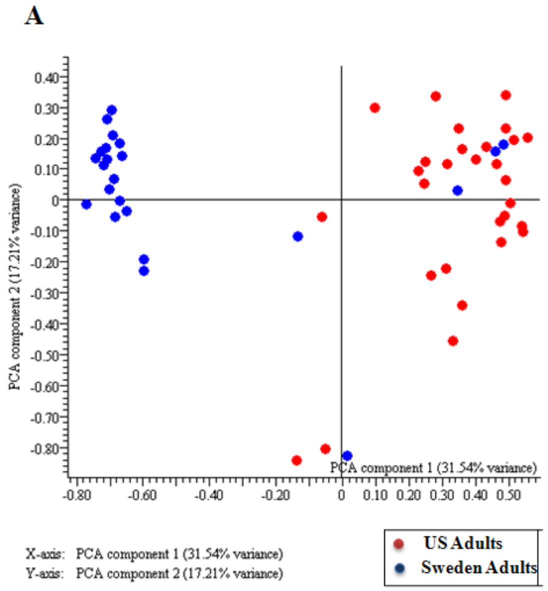
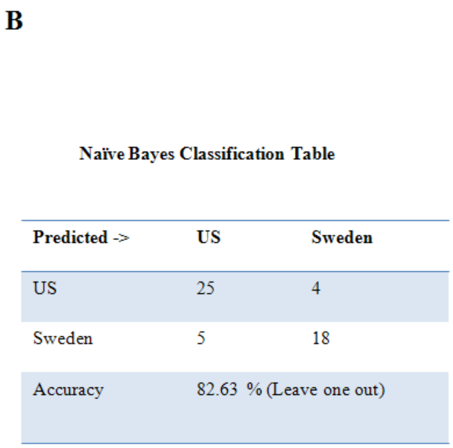


Figure 6-5: Immunosignature of US vs Sweden Adults

6.4.6 CLASSIFICATION PERFORMANCE OF US VS SWEDEN ADULTS

When tested for separating the US and Sweden adults groups using unsupervised methods, the two classes were distinct. **Figure 6-6** (A) shows the principal component analysis for all 10,000 peptides explaining 48.75% of the variance in the population. The two classes can be very well separated with linear line with few misclassifications. (B) shows the classification table obtained from Naïve Bayes algorithm from all 10,000 peptides after leave one out cross validation. 82.63% accuracy was achieved. Overall geographical location specific to country does make it an overall effect on the immunosignature.





Naïve Bayes Classification Table

Predicted ->	US	Sweden
US	25	4
Sweden	5	18
Accuracy	82.63 % (Leave one out)	

Figure 6-6: PCA and classification table for US and Sweden adults immunosignature

6.5 CONLCUSION AND DISCUSSION

We used different factors like age, gender and geographical location to examine if these factors have any effect on immunosignaturing technology. We tested if children have different immunosignature than adults and found that the overall binding of antibody to peptide is very high on our arrays for children sera. The range of the peptide intensity binding was also high compared to that of adults. There are a few plausible explanations for observing high binding for the children sera. One of the explanations might be due to vaccination of children due to which they might have high immune response against random peptides. Another explanation might be that the children immune system are much more vibrant than that of adults since they are being less exposed to infections at the early age. Since immunosignaturing technology measures the IgG antibody binding to the peptides, it might be possible the children have higher IgG antibodies in the first place.

Using unsupervised methods, the difference between the immunosignature of children and adults are clear. The can be separated well from each other with >95% accuracy. Hence age as a factor affects the immunosignature to a significant extent.

When tested for male and female immunosignature of adults, we did not find any clear indication of the difference using unsupervised methods. There was no significant difference between overall intensity of peptide binding on our arrays. When the top 50 peptides were selected that are differential between the males and females, the separation was not high and the two classes looked similar. In the principal component analysis, the males and females immunosignatures were clustered together and <75% accuracy was achieved in classify the two classes from each other. There have not been many reports indicating the difference in function of humoral immune response in males and females. This is one of the advantages of using proteomics biomarkers such as antibodies that are homogenous between genders. This would highly facilitate development of biomarkers that can be used for all the adult population irrespective of gender.

We then finally test the effect of geographical location by comparing the immunosignature of US adults and Sweden adults population. We have seen that there is no significant difference of local geographical location like people from

Arizona, California etc (data not shown). But when cross countries immunosignature are compared, we found a significant difference between the immunosignature of Sweden adults and US adults. Sweden adults spanned much higher range of peptide intensities compared to US adults and the overall peptide intensities of binding was higher than that of US adults. In the unsupervised analysis using all 10,000 peptides, principal component analysis yield a good separation between the two classes with few misclassifications. The unsupervised classification yielded >80% accuracy between the two classes. Overall, the difference between the immunosignature of US adults and Sweden adults are clear. This might be due to several reasons possibly due to climate and life style difference.

Finally, in order to establish the potential of this technology for creating a diagnostic, we tested factors like age, gender and geographical location for their affect on the immunosignaturing technology. Ideally in theory, this technology would eventually be used for health monitoring, a complimentary approach to diagnostics where every person would monitor their health status on regular basis. In that stage, the factors like age, gender and geographical location would no longer be of concern and all population variation would be normalized. But for more immediate use, it is extremely important and vital to understand the factors that would affect this upcoming immunosignaturing technology.

CHAPTER 7 TAKING IMMUNOSIGNATURING TEMPERATURE

7.1 ABSTRACT

The primary reason for the on growing health care crisis is due to post symptomatic medicine which relies on treating a disease after the person is sick or diseased. The major limitation of this approach is that the system in concern is already damaged and biologically, the concentration of antigens is very high. This in turn makes the system both challenging as well as expensive to repair. If this system is changed to pre symptomatic medicine, then the expense can be lowered to a significant extent as well the efficacy of the treatment methods could improve. In order to perform an early diagnosis rather than traditional diagnostics, it is imperative to establish a regular health monitoring system. One such upcoming technology that focuses on health monitoring is immunosignaturing. This technology profiles the humoral immune response through the binding of antibodies through random sequence peptides. Currently using 10,000 random peptides, this technology delivers an unbiased picture of the immune system creating a binding distribution of antibodies. Cells under the normal state are regulated but as soon as the body starts to enter a diseased state, the cells function starts to change, triggering the immune system. We hypothesize that during the abnormal state of immune system, the profile of antibodies would change leading the change in distribution pattern of their binding to the random sequence peptides. If we were to take a measurement like temperature of the immune system on a regular basis, we could potentially detect any subtle changes in the

immunosignature of a diseased individual even if the specific cause were not known. We tested our hypothesis by taking Coefficient of Variation (COV) of all 10k peptides of a healthy individual and compared it with diseased person on a population level. We found, that this measurement changed significantly in the disease group with sensitivity but not specificity. Also this measurement is stable over time, robust to protocol variation and even on the choice of 10k random peptides. Such unbiased mathematical and statistical measurements would help taking an unbiased view of immune systems and possibly alarm individual for any infection or disease.

7.2 INTRODUCTION

Healthcare expenditure is rising every year with total $2.5 trillion spent in 2010 alone in US (US Department of Health and Human Services 2010). About 90% of the expenditure goes in maintenance to take care of sick and diseased people (US Department of Health and Human Services 2010). This is primarily due to post symptomatic medicine methodology that we currently follow. Here, we wait for the individual to get sick and then start the treatment. The main limitation of this approach is that at a later stage of any disease, the concentration of antigens (defective molecules) is very high and probability of inexpensive and normal methodologies to normalize the situation decreases. A paradigm shift is needed to move the approach of post symptomatic to pre symptomatic diagnosis or a prognostic approach. Towards this approach, we need a novel technology for regular health monitoring to capture significant change in the health status of an individual. One such technology is immunosignaturing, which profiles the humoral immune responses through binding of antibodies to random sequence peptides. Antibodies can be used as biosignatures of an individual's health through observing their binding affinity on random sequence peptides. These random peptides have no homology with known epitopes, so they provide an unbiased profile of antibodies binding representing an individual immunological status.

The immune system in a normal state has a profile of antibodies that are produced during a healthy stage of an individual. When this person encounters a foreign antigen or self-abnormal activity, this profile would change indicating an alarm. Immunosignaturing technology can provide a health status monitoring tool in this case to measure the antibody profile at a normal stage and alarm when the profile of antibodies change during a disease state.Using a machine-readable platform, both healthy and diseased individual signatures can be obtained from the antibodies binding 10,000 random peptides. The challenge now is to obtain a single measurement out of 10,000 random peptides that is equivalent to taking a temperature of the immune system to monitor abnormal activity.

We tested the hypothesis of using a single metrics to measure the overall measure of antibody binding to 10k random peptides at a particular given time using immunosignaturing technology. Such metrics should incorporate several criteria. Firstly it should be independent on the choice of 10k random peptides for the unbiased measure. Secondly, it should be stable over time in an individual and also robust to protocol variations in the technology and thirdly it should change significantly in diseased group or when the person is having infection or chronic condition. Developing such metrics is both useful and challenging. Here, in this work we tested one such metrics Coefficient of Variation for its efficacy to achieve above milestones. COV by definition is just the inverse of signal to noise ratio. It is calculated by measuring the standard deviation over mean. The prime rational behind using this metrics is that when the immune system is in a normal state, the distribution of antibodies binding to random peptides will have some have some specific chaotic behavior. This pattern is likely to change; either becomes less chaotic, due to the antibodies generated towards a particular antigen or becomes more chaotic if several types of antibodies are raised against multiple complex sets of antigens. Measuring such distributional behavior in normal and disease conditions, antibody activity would eventually lead us to a measure that can diagnose disease conditions.

7.3 METHODS

7.3.1 MICROARRAY

The CIM 10k V.1, V.2 AND V.3 is a microarray containing 10,000 random sequence 20-mer peptides attached via a maleimide reaction to the NH_3 terminal sulfur of cysteine, creating a covalent attachment (Legutki et al. 2010).

7.3.2 SAMPLES PROCESSING

Sera samples from healthy controls were stored at -80 C after which samples were aliquoted and refrozen at -20 C. Samples were diluted at 1:500 in sample buffer containing 1XPBS, 0.5% Tween20, 0.5% Bovine Serum Albumin

(Sigma, St. Louis, MO) and then applied to the array using the standard immunosignaturing protocol as mentioned (Legutki et al. 2010). Antibodies were detected with 5nm Alexafluor 647-labeled streptavidin (Invitrogen, Carlsbad, CA). The tertiary antibody binds to 5nM biotinylated anti-human secondary antibody (Novus anti-human IgG (H+L), Littleton, CO). Microarrays were scanned and converted to tabular data after which median foreground signal was used as the value of antibody peptide binding.

7.3.3 SAMPLES

The center for Innovations in Medicine, Biodesign Institute, Arizona State University has an existing IRB 0912004625, which allows analysis of blinded samples from collaborators.

To test the effect of protocol variation on COV metric: We collected 10 sera samples each of healthy and T1D children from the diabetes study and ran them under 1nm secondary IgG antihuman antibody and AF-555 as tertiary antibody. For another protocol, we collected non overlapping 30 sera samples each of healthy and T1D children from the same study and ran them after 6 months using 5nm secondary IgG antihuman antibody and AF-647 as tertiary antibody.

To test the effect of peptide array variation on COV metric: We collected sera samples from 7 healthy US adults and ran them in both CIM 10k. V1 and V2 using different set of 10,000 random peptides.

To test the effect of time variation on COV metric: We collected 3 adults sera samples from Biodesign Institute, Arizona state university and followed their immunosignature on CIM 10k V.1 for 15 days with daily collection of blood.

To test the effect of early stage onset for COV metric: We collected 3 biological replicates each pre and post influenza challenge.

To test the consistency among normal group for COV metric: We collected 72 sera samples of healthy individual collected from 5 different sources.

To test measurement of Normal Vs Disease conditions on COV metric: We collected 14 samples of breast cancer, 23 samples of type 2 diabetes, 12 samples of Alzheimer, 5 samples of PanIN, 10 samples of pancreatitis, 19 samples of pancreatic cancer, 26 samples of esophageal cancer, 20 samples of GBM and finally 20 samples of Myeloma.

7.3.4 DATA ANALYSIS

The raw tabular data were imported to GeneSpring 7.3.1 (Agilent, Santa Clara, CA). Data were median normalized per array and then transformed into $\log_{(10)}$ scale. To get an unbiased and unsupervised view of the immunosignature, no feature selection has been performed. Coefficient of Variation for each sample is calculated by taking the mean of the 10,000 random peptides over standard deviation of all the 10,000 random peptides. For each class, normal or disease, 95% confidence interval has been constructed by taking 1.96 * standard error. Overlapping confidence interval indicates that the two groups mean difference is not significant.

7.4 RESULTS

7.4.1 ANALYSIS OF COV ON PROTOCOL VARIATION

We tested COV metric for its robustness for protocol variation on immunosignaturing technology. We tested subsets of disease and control samples collected from the same study, but processed under different time points and different secondary and tertiary antibody concentration. Immunosignature of 10 sera samples each of T1D and controls were collected using 1nm secondary antibody concentration and AF-555 tertiary antibody concentration in the Set 1. For Set 2, we collected 30 sera samples each of T1D and controls from the same study. COV was calculated for every sample; mean and 95% confidence interval was calculated for both diseased group and controls for both the sets. **Figure 7-1** shows the mean and confidence interval for both the sets. Mean of COV in the diseased group between the two sets were not significantly different, and so was the mean COV of control group in the two sets. Although run under different protocol conditions, the measurements were stable and consistent among the two sets.

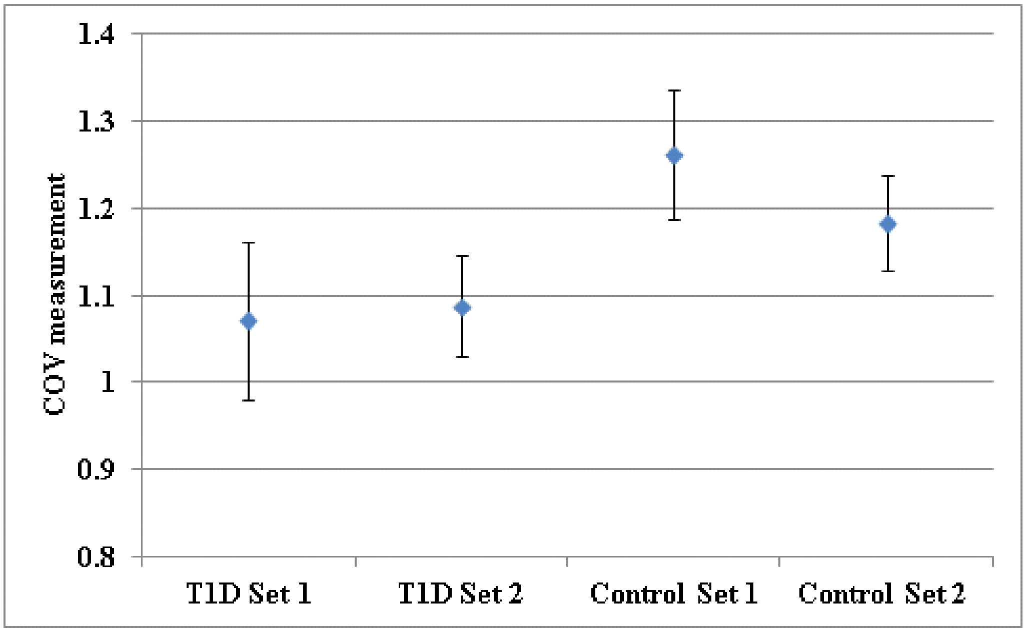


Figure 7-1: COV analysis for protocol variation

7.4.2 ANALYSIS OF COV ON PEPTIDE ARRAY VARIATION

We tested the COV metric for its robustness for the choice of 10,000 random peptides on immunosignaturing technology. We tested subsets of 7 healthy normal subjects on CIM 10k V.1 and V.2 consisting of non overlapping 10,000 random peptides. **Figure 7-2** shows the mean of 7 samples with 95% confidence interval for the two peptide arrays. The confidence interval for the two groups overlapped indicating there is no significant difference among the mean in the two groups between V.1 and V.2 CIM 10k arrays.

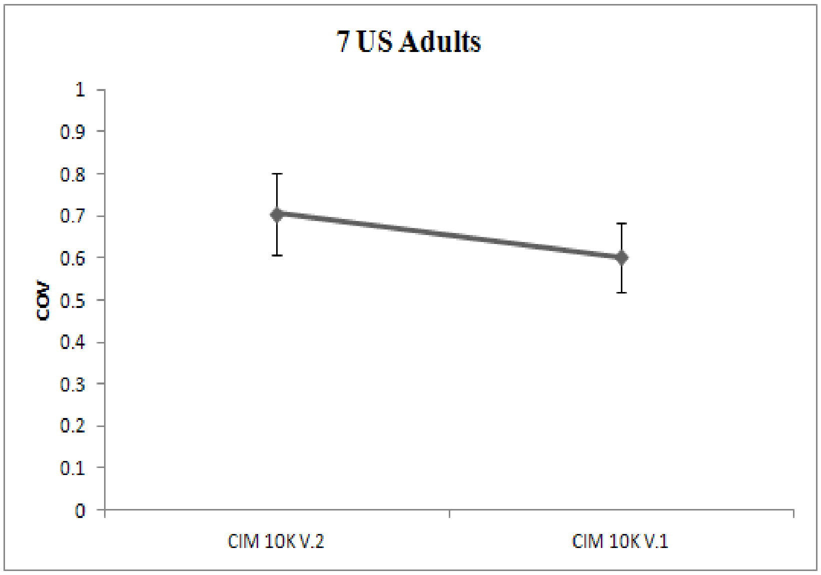


Figure 7-2: COV analysis on peptide array variation

7.4.3 ANALYSIS OF COV ON TIME COURSE

We tested COV metric for its consistency and stability over time on healthy individual subjects. Towards this goal, we tested 3 healthy individuals and followed their immunosignature for 14 days. **Figure 7-3** shows mean of COV for each individual over 14 days time course and 95% confidence interval. For 14 time points, COV was tightly regulated with <0.1 standard error. There was no significant difference among the three groups since the confidence interval overlapped indicating COV metrics is stable over time for healthy individuals.

We then followed a subject (ID 43) immunosignature for 30 days during which subject has been vaccinated against tetanus at day 17, we tested if we could see any differences in mean COV of all 10k peptides for that subject. We found a slight increase after day 17 to day 24 in the COV after which this measurement stabilizes. **Figure 7-4** shows COV measurement of subject 43 over 30 days.

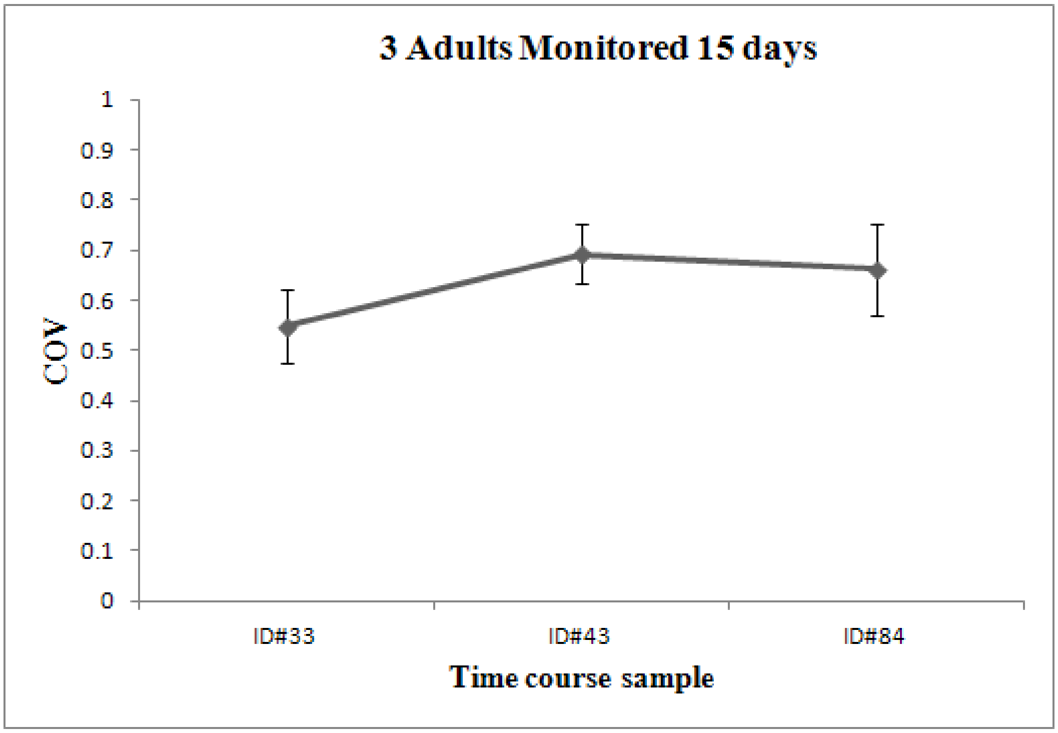


Figure 7-3: COV analysis of 3 healthy individuals on time course

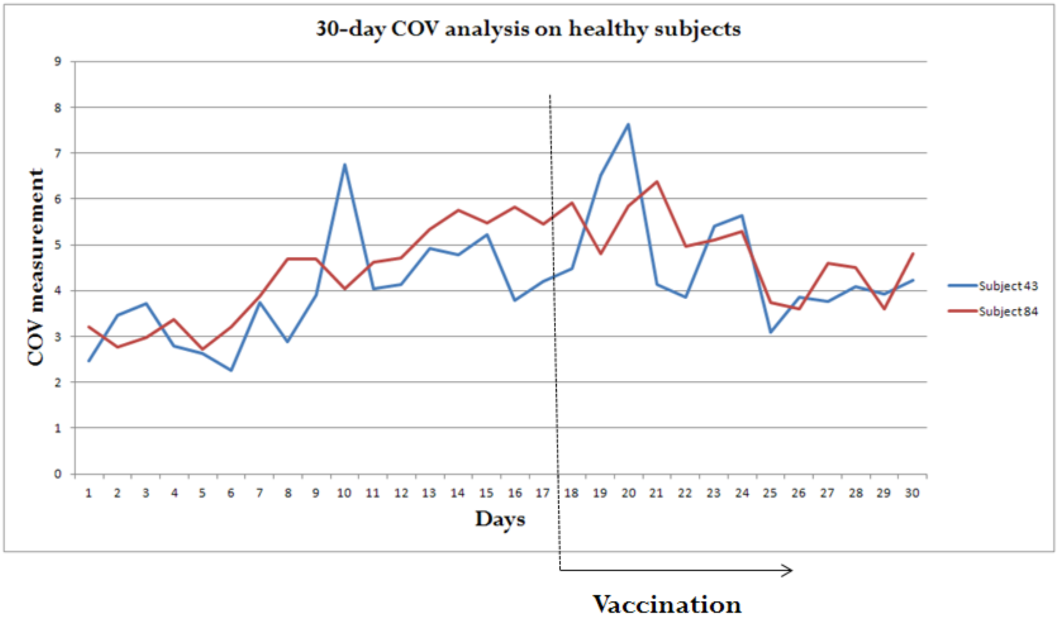


Figure 7-4: 30-day COV analysis on subject 43 and 84

7.4.4 ANALYSIS OF COV FOR EARLY STAGE DETECTION

We tested COV metric for its ability to detect disease stage at an early onset. For this, we collected 3 mice and followed their immunosignature pre (day 0) and post (14 days after the post influenza challenge). **Figure 7-5** shows the mean of COV of pre and post immunized group with 95% confidence interval. There was a statistical significant difference between the two groups (p value = 0.01) indicating the COV metrics changed at an early stage during the mouse immunization.

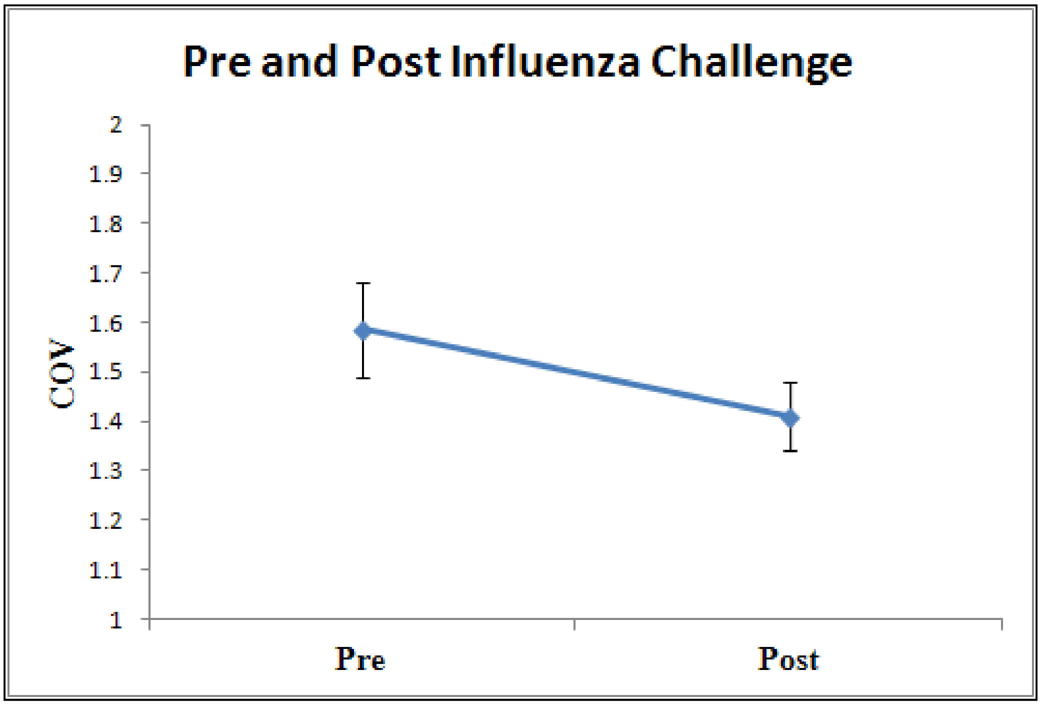


Figure 7-5: COV analysis for early stage detection

7.4.5 ANALYSIS OF COV ON HEALTHY POPULATION

We tested COV metric for its consistency among different sets of normal population. We tested 5 different sets of healthy individual sera collected from different studies across US. The group G1 consisted of 16 samples, G3 consisted of 11 samples, G4 consisted of 21 samples, G5 consisted of 10 samples and finally G6 consisted of 14 samples. **Figure 7-6** shows mean and 95% confidence interval of mean of COV for 5 different groups. G1, G3, G5 and G6 were quite consistent with similar mean G4 COV had lower mean but all the overlapping confidence intervals indicated no difference among mean (COV) for the 5 groups. This shows that COV metric is tightly regulated in the healthy population.

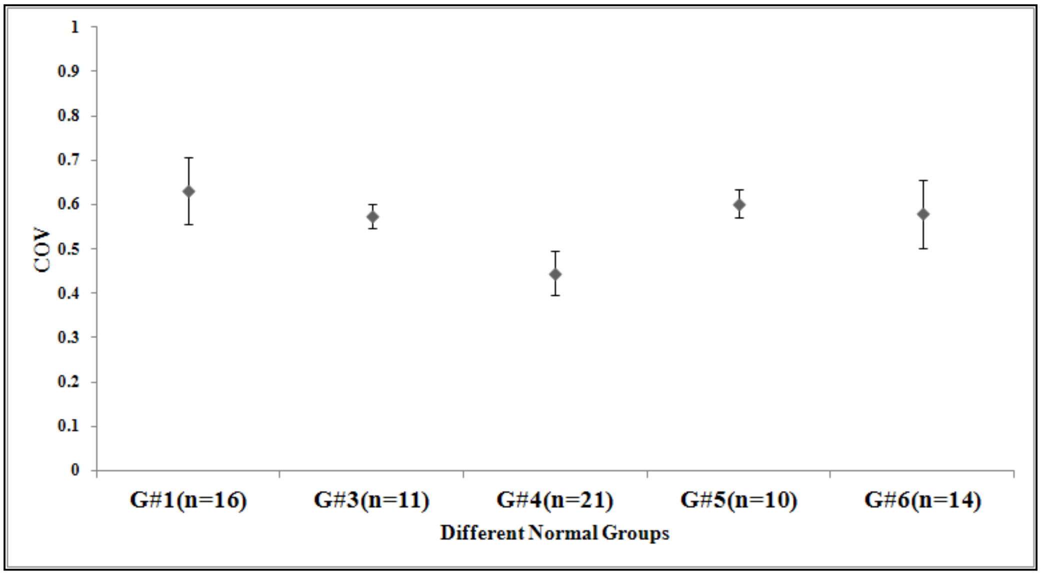


Figure 7-6: COV analysis on healthy population

7.4.6 ANALYSIS OF COV ON DISEASED AND HEALTHY POPULATION

We tested COV metrics for its differential ability in various disease populations compared to healthy individuals. Since COV metrics was consistent in different sets of healthy individual, it was combined into one set containing 72 samples. For disease status, we compared healthy individuals ‘mean COV’ to various infection and chronic diseases set. This includes various types of cancer like breast, pancreatic, myeloma, GBM and esophageal cancer. In other chronic diseases we compared the healthy individuals COV with type 2 diabetes, Alzheimer and panIN. For infectious disease, we compared it against pancreatitis. **Figure 7-7** shows mean COV of normal and other diseased group with 95% confidence interval. In breast cancer samples, the overall immunosignature was highly suppressed and there was no significant difference found in mean COV between the two groups. Also in type 2 diabetes group, the mean COV between the normal group and type 2 diabetes group was not significant. Apart from these two, for all the other disease states, there was a statistical significant difference among the mean COV of healthy individual and disease groups. This shows the during disease state, COV metrics on the immunosignature tend to change significantly compared to normal state.

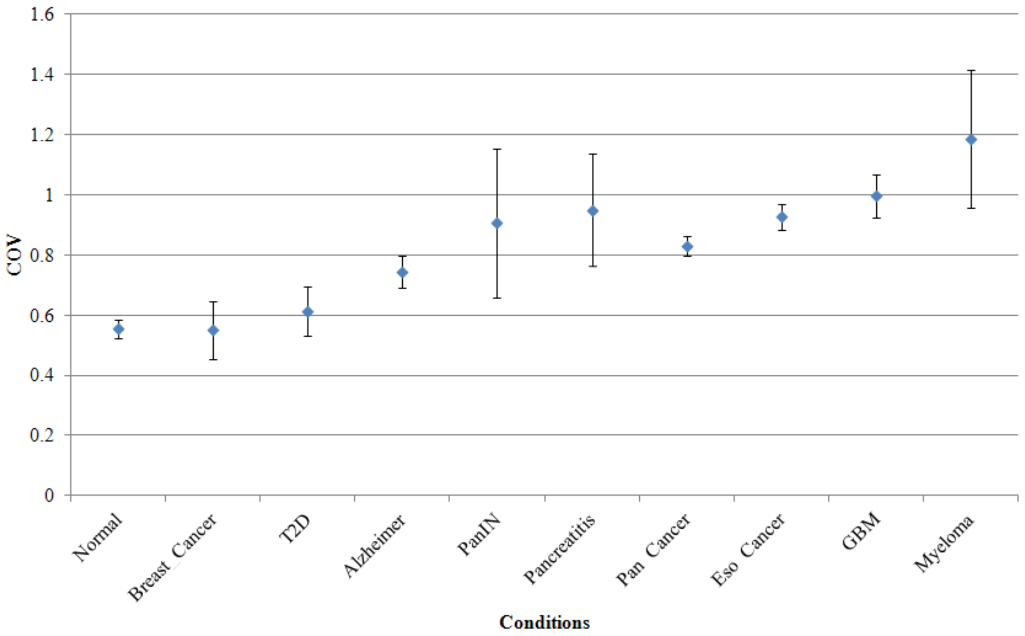


Figure 7-7: COV analysis on normal vs various disease immunosignature

We tested COV metric for its ability to be sensitive to general diseases and relax the criteria for specificity. Towards this goal, we merged several related diseases like into one group. Breast cancer, pancreatic cancer, myeloma, GBM and esophageal cancer were grouped into one cancer group. Pancreatitis, PanIN, type 2 diabetes and pancreatic cancer were grouped under pancreas diseases. And finally all diseases were clubbed into one group known as diseased group. **Figure 7-8** shows the mean COV for healthy individuals, pancreatic disease group, Cancer group and finally the diseased group with 95% confidence interval. Healthy subjects COV were statistically different from any of the disease related groups indicating the COV metrics changes when the system is diseased and this metric is sensitive but not specific enough to predict which exact type of disease is encountered.

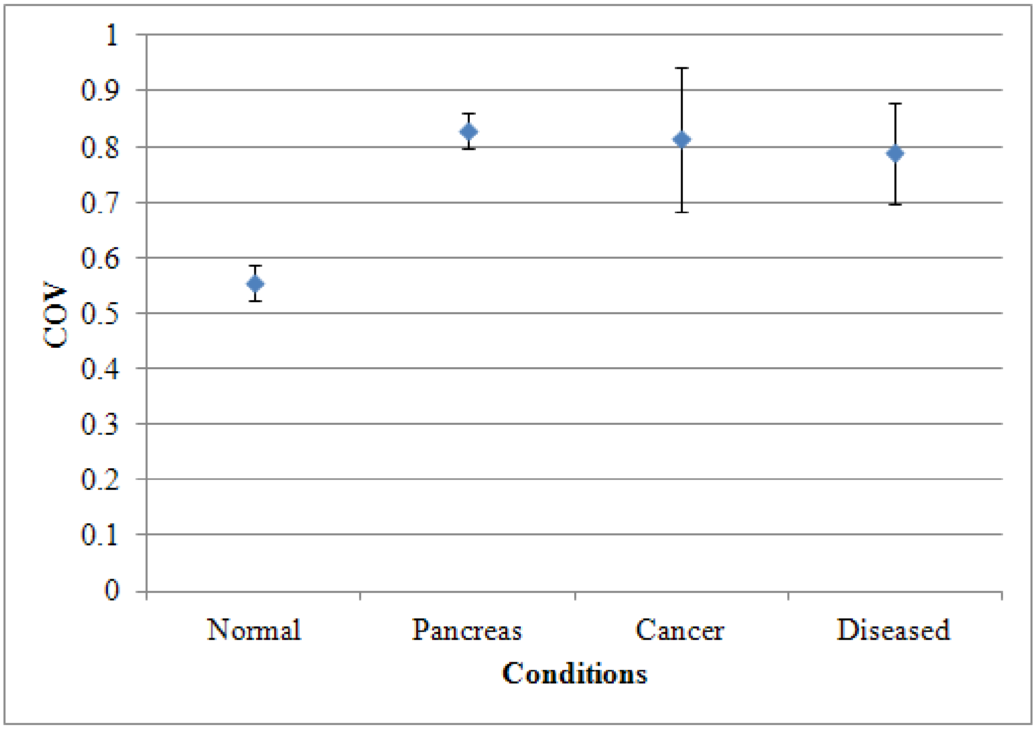


Figure 7-8: COV analysis of healthy vs general disease groups

7.5 CONCLUSION AND DISCUSSION

In this work, we proposed a diagnostic approach that can revolutionize healthcare using general monitoring of immune system by immunosignaturing technology. Immunosignaturing technology captures humoral immune response by binding of antibodies to random sequence peptides. When the immune system is in the steady or non-diseased state, we believe there is a general pattern of distribution of antibody binding in healthy individuals to random sequence peptides. This pattern is likely to change when the immune system encounters a self or non self attack due to the fact that antibodies will now be raised against a certain epitope of the antigens and certain set of random peptides will be mimotopes of these antigens. Overall distribution pattern of diseased individual antibody peptide binding compared to healthy individuals will change. Towards this goal, we proposed that using Coefficient of Variation as a single metrics can be used as a temperature like measurement of the immune system on the immunosignaturing technology; this metric would change when a health individual gets sick.

In gene expression studies, people hypothesized that cells under the normal state are tightly regulated, hence gene expression would be non chaotic but when a person enters a diseased state, the cells would enter into unsteady state, leading the gene expression to be more chaotic. But we observed an interesting phenomenon which may be different while working with immune system antibody profiles binding. When a person enters a disease state that have limited antigens like flu, or type 1 diabetes, we expect the antibody profile to be more tightly regulated since now immune system would produce antibodies specific to these limited antigens and hence the profile would be less chaotic. But when a person enters into a chronic condition or a disease responsible of several causes or antigens, then overall antibody profile would be diverge from normal behavior to a higher extent leading to chaotic behavior. We found COV of disease group in limited antigens diseases like T1D and flue as less compared to control group while the COV measurement of chronic disease like cancers were higher in disease group than that of control group.

For any single metrics per say, which would be an equivalent to take the temperature of immune system, there are several characteristics that this metric should achieve. We tested COV metric's potential for these characteristics; we first tested this measure against protocol variation in the technology. We found that the COV metrics is very robust against protocol variation such as different concentrations of secondary and different types of tertiary antibody in the immunosignaturing technology. One of the major challenges in working with the growing technology is that there persists a continuous process of optimization in protocols. Such optimizations we believe should not hinder the choice of a single metrics and it should be robust against change in protocol variations and calculates the overall average pattern of binding of antibodies to peptides. Secondly we tested whether choice of 10,000 random peptides depend on the calculation of COV metrics, we found that this metric does not change significantly for the same individuals when they are tested in two different versions of CIM 10k arrays. Since the peptides are random, it is extremely important that any metric is not dependent on the choice of the random peptides. Thirdly we tested COV metrics stability over time during a un-disease state. One of the challenges of the microarray data is the noise and random variation, which affects the stability of any metrics. COV when tested over 14 days on healthy individuals, showed consistency in measurements. This is an extremely important characteristic in any metrics; this allows low standard error hereby enabling small changes in measurements to be significant. Apart from being consistent in an individual, it is also important that the metric should be consistent in a population;

this lowers the standard error in population and enables to deliver promising measurement over population studies. To test this, we tested COV metric on 5 different non-overlapping healthy individual's sera sample sets collected from different sources. We found, no significant difference in the mean COV among the groups. This facilitates the use of COV metrics on population level studies. One of the major goals of the prognostic approach is to detect any infection/chronic disease at an early stage; hence we tested the efficiency of COV metrics for its ability to change significantly in a mouse model. There was a significant difference between COV of pre and post mice; after 14 days of influenza challenge. This shows this metric has the potential to be able to incorporate early disturbances in immune system during infections. Moreover, we tested the ability of COV metrics to be able to differentiate among healthy and chronic diseased conditions. We calculated COV of 72 healthy individuals and compared it against various diseased groups like different types of cancers, Alzheimer's etc and found a significant difference between mean COV of healthy individual's vs diseased groups. Overall, mean COV of healthy individuals was lower compared to any diseased groups. This indicates that when the immune system is diseased, it becomes more chaotic due to antibodies produced against various antigens during a disease state. In the normal state, antibodies binding to random peptides are spread uniformly and usually unbiased for any particular epitopes or mimotopes. When the system goes into the diseased state, antibodies are generated against a particular epitope of the antigens, and random peptides on our arrays may be mimotopes for some of the antigens, which lead to more diverse signals towards these mimotopes. Due to this, the overall chaotic behavior increases and mean COV rises in diseased individuals compared to healthy. Since COV only calculates the overall behavior of binding, it is not specific to disease conditions; hence we saw no significant difference between mean COV between disease groups alone. Hence, this metric is sensitive but not specific for predicting the type of diseased condition.

CHAPTER 8 CONCLUSIONS AND FUTURE WORK

Health related concerns, be it social, economical have been always an inseparable part of our society. With modern education and resources, people are becoming more aware of the importance of health in our lives. Improvements in technology have created methods, tools and devices that have significantly impacted both our life-style and age in a positive way. But at the same time, our efforts on improving health care have been depleting our finances to a significantly higher extent. If examined carefully, these expenditures are largely hospital bills of people who are already diseased. This is basically the cost of post symptomatic medicine methodology that we currently follow in our society. Due to lack to appropriate technology and ideas, there aren't many institution or research centers that are pursuing research towards pre symptomatic medicine. For most of the people, early diagnostics often means to detect the early recurrence of disease symptoms or cancer tumors so it is a bitter truth that early diagnostic has lost its name in reality. Another road block for early diagnostics research is there are not many technologies or research that are focusing truly on pre-symptomatic methods and thus this field, although very powerful, has been ignored by our society unconsciously.

There are two complementary approaches for adopting pre-symptomatic medicine methods; first is in the discovery of early diagnostic biomarkers and the second being able to monitor health monitoring on regular basis. The goal of discovering early biomarkers as been supported for the last 30 years and trillions of dollars has been spent on the approach of finding single, early biomarkers in the population. Mass spectrometry showed the most promise in delivering biomarkers but attempts were not significant since the population was heterogeneous and sample preparation was laborious and cumbersome. To date we only have 47 FDA approved biomarkers. The major problem of the approach to date is the basic process of discovery that is too conservative. The approach generally starts with 10:10 disease/control samples with a highly homogenous population and then researchers screen for significant differential biomarkers in the 10:10 sample study. The candidates are pushed in next stage of 100:100 disease control study and so on. But as the study progresses, the candidate biomarker molecules tend to fall out due to the problem of finding a single element in a heterogeneous population.

Lately health monitoring has been getting increased attention but there is a big technological limitation of doing a health monitoring. Most of the technologies, apart from Immunosignaturing, I am aware of are not well designed for health monitoring. The biggest reason why other technologies are not able to focus on health monitoring is due to their focus on assaying a single disease. For health monitoring, our focus should be unbiased against choice of biological elements but at the same time our choice of elements should be appropriate. So to achieve perfect balance in terms of getting unbiased as well as early obtaining individual normalized biosignatures is to use antibodies as biomarkers in the form of a random peptide microarray. Immunosignaturing has its root in traditional microarray with DNA fragments replaced with random peptides. This allows specific and sensitive measurement of immune system of an individual through binding of antibodies to random sequence peptides.

Immunosignaturing is different from traditional microarrays in terms of the complexities it captures; it is highly a complex multiplexed assay where a DNA microarray relies on one-to-one binding of molecule of interest to probes. The immunosignaturing microarray captures multi-level interactions. A single peptide on our array can bind to multiple antibodies and a certain antibody can bind to multiple peptides forming a many to many relationship. Hence, each peptide shows in the end, the resultant binding of multiple antibodies to it. This data is highly informative, but traditional methods, especially classifiers which worked for gene expression microarrays, may not be suitable for immunosignaturing data. So at first, I did a literature search for known and most commonly used classifiers and applied them on various immunosignaturing case studies. I tested 17 different classification algorithms on different datasets over multiple levels of stringency; the outperformance of Naïve Bayes was pretty clear. Some users of immunosignaturing data have found that the algorithm SVM tend to work slighter better than Naïve Bayes. But my studies have shown that overall Naïve Bayes outperforms in the long run and can be considered as gold standard if someone wants to stick to one classifier for the data analysis. The biggest variable on the choice of best classifier is number of features (peptides) that you want to select for your hypothesis. While some hypothesis are based on diagnostics test, while others are based on narrowing down significant peptides from 10,000 to top 200 in pilot studies and using small set of peptides for larger sample sets. Some classifiers like LDA (Linear Discriminant Analysis) performs well in the number of peptides selected is less than 25; in short this classifier is good for a small number of features. So if a person runs his/her immunosignaturing case study against the LDA classifier for the top 10 peptides and found the accuracy greater than the Naïve Bayes classifier it does not imply LDA outperforms Naïve Bayes over the long run. I have shown that overall a classifier performance decreases when the number of features increases. Naïve Bayes performance does not drop drastically in this case yielding a consistent measure irrespective of number of peptides. Also Naïve Bayes has never shown to underperform in my immunosignaturing case studies. It has ranked second on some datasets but I have not seen more than 10% difference between any classifier and Naïve Bayes performances. The main reason I support the Naïve Bayes performance is due to its simplicity and a unique property of independence which is satisfied by the immunosignaturing design. Naïve Bayes classifier is based on Bayes rule which assumes independence of features. In real world applications of classifiers even in DNA microarrays, there is always some sort of dependencies on features. In the case of gene expression microarrays, they are connected via regulation networks with master and slave genes; hence the probability of Naïve Bayes working is low due to its assumption of feature independence violation. But in immunosignaturing, random peptides are completely independent of each other and there is no significant sequence similarity we know between peptides. On this note, some people have argued about the independence of the peptide in immunosignaturing claiming it is not independent. When a monoclonal sera is run on our arrays there are multiple peptides that show signal in synchrony so they may be mimotopes of the same antigen so they are not independent. As simple as the argument sounds, it’s not completely true. As described the complexities of immunosignaturing design; each peptide signal is mixture of multiple antibodies binding, so although these peptides which are mimotopes might be correlated but they are often bound by other antibodies making the correlation less and more independent due to multiplicity. I think for a classifier the conclusion is pretty concrete that Naïve Bayes is very effective for immunosignaturing case studies. Overall performance of a classifier also depends on features selected, so more work can be done on feature selection methods apart from t-test to see if that increases Naïve Bayes performance. There are two parameters to consider. First are the feature selection methods, and second are classification algorithms. I used a single feature selection method and varied different algorithms to choose the best. Now when we know which classifier is best, future work can be extended to choose the best feature selection method for which Naïve Bayes performs the best. The reason for I do not feel very confident in using the t-test feature selection methods is due to the fact that biosignatures study like immunosignaturing are based on heterogeneous data and t-test look for consistency in terms of peptides intensity in a class. So even in the disease group; some samples might be of another type, while other will be of other. For example, I saw in one the immunosignaturing case studies was when I was analyzing type 1 diabetes data set. Although 40 samples were classified into one class T1D but based on their Radio Immuno Precipitation Assay titers, I found there was a sub grouping of subject's immunosignature having high and low IA-2 titer. So what I observed was peptides that were specific to IA-2 did not get selected in the t-test since there measurements weren't consistent across the disease group of T1D. Hence, a feature selection method should be adopted which also looks for sub grouping in the population and then look for consistency.

Finding a right classifier was the very first and important thing I did for my immunosignaturing case studies. Analysis of immunosignaturing data in any hypothesis often demands classification among different conditions of interests, so Naïve Bayes algorithm came in handy while I was researching other hypotheses relative to the technology. When my center first pursed immunosignaturing, the biggest critique that we got from immunologist was that this technology would not be specific for a particular condition. Their expectation was that we would only see general inflammations responses. So my first goal was to find the resolution ability of immunosignaturing technology. The aim was to test the sensitivity and specificity of the immunosignaturing technology. So I pursued a case study involving closely related diseases in the pancreas (type 2 diabetes, pancreatic cancer, pancreatitis and panIN). On immunosignaturing individual diseases and common a control group, it was clear that this technology was sensitive enough to detect the presence of disease with >90% accuracy when tested on independent test sets. Then for specificity, I compared each disease immunosignature among each other and found >90% accuracy on separating each disease. It was clear that this technology was not only specific but also sensitive enough to detect the presence of closely related inflammation and chronic diseases in the pancreas. While studying this hypothesis, I realized a unique mathematical property of immunosignaturing data. I postulated that in a population study immunosignaturing data can be considered as a vector which has three orthogonal components at a certain significance value. The primary component being a disease component which is set of the peptides that changed significantly from the control group. The secondary component, I refer to as the 'normal' or 'house keeping component' are the set of peptides that does not change significantly between the diseased and control group. The third component I termed as 'individual variation' are the set of peptides are unique for each individual and change according to personal variation. All the three components have their intrinsic value depending on the question in hand. The disease component is highly informative if we are studying diagnostics for disease, normal component play its role for quality control and providing a base-line for population studies. The third component is important when we are studying health monitoring of an individual over time. Although my research question in the beginning was to test if this technology is sensitive and specific for chronic diseases, I leveraged my disease component data for the pancreas related diseases to test if there are any similarities among specific disease immunosignatures. Epidemiology studies have reported a correlation and increased risk prediction of pancreas related disease if an individual already have another pancreatic disease (Noel et al. 2009). In specific there cases of individual having both pancreatic cancer and type 2 diabetes is pretty common. Although correlation between the two does not imply causation, researchers suspect that there might be some common biomarkers between the two diseases (Huxley et al. 2005). In my immunosignaturing data of pancreas related disease, I found consistency with the literature findings. When the top 200 peptides for each disease are compared, there was a significant overlap between signatures of type 2 diabetes (T2D) and pancreatic cancer. Also there was a significant overlap between signatures of T2D, pancreatic cancer and panIN which reveals that there is some common antigen driving the common signatures of these diseases. Harnessing this information, I proposed an idea of general monitoring of pancreas related diseases. Instead of taking each disease as individual disease, I proposed to take all diseases under one single disease and find the signatures that are consistent among all diseases. I found 673 peptides that had high specificity (>95%) on a test set which can be used for the general monitoring for pancreas related diseases. This study can be extended by considering peptides that are common in each disease component and decipher them for common antigen responsible for the cause.

Immunosignaturing when tested for various diseases and infections shows promise since the signature was reproducible, consistent and unique for every disease tested. It is a general lab practice to initially look for a small set of samples for disease and control. But in reality individuals often suffer from more than one complication, so we tested whether a disease immunosignature still remains consistent in the presence of another disease. When we tested subjects having type 2 diabetes and (CHF, MI) alone, their immunosignatures were unique, consistent and reproducible with 100% sensitivity obtained when comparing with healthy controls. But when I tested samples from people having both T2D and (CHF, MI), their immunosignatures had only a few features (peptides) in common with the individual complication. Apparently the presence of both diseases creates an interaction resulting a different immunosignature. The occurrence of these two complications are correlated and but some have even proposed a casual relationship (Haffner et al. 1998). When we tested the same concept on two diseases which do not interacted or correlated or in casual relationship like [(Cancer and flu), (KLH and PR8 immunization), we have found that even in the mixture of two diseases, individual disease immunosignature can be separated. Thus, if a subject is having multiple complications, immunosignaturing diagnostic for a single complication might still be reliable if multiple complications are not correlated or in casual relationship.

With successful attempts in classifying disease/conditions though immunosignaturing over the last few years, we developed confidence that this technology may have use as a diagnostic. But one thing which amazed almost everyone who has been encountered with this technology is the power of using random peptides. For some, it feels suspicious while for some it realizes that it makes sense to be able to deal with the stochastic behavior of immune system with the use of random peptides. But everyone at first seems a little dubious to trust the power of random peptides. Initially we attempted to trace back random peptides by applying monoclonal antibodies on our arrays but we found that although monoclonal antibodies were raised against a particular epitope, these antibodies recognize other epitopes on our array (mimotopes). Later our lab developed a tool GUITOPE which is an epitope mapping tool for random peptides onto given protein(s). On working with the immunosignature case study of type 1 diabetes, it occurred to me to work on tracing random peptides for T1D to its respective antigens. T1D has been studied well and 3 known autoantigens have been reported along with other candidate antigens. So I took T1D as a case study to trace random peptides to known T1D antigens. On a sample set of 40 disease and 40 controls, I found 679 peptides that were different in T1D compared to common controls. Out of 679 peptides, 210 were higher and 479 were lower significantly compared to controls. I observed that not only high binder random peptides were successfully mapped to 8 known antigens but also the low binder peptides were mapped to 3 known antigens in diabetes. We then hypothesized that there are already existing autoantibodies during the normal state against self proteins but modern technologies like ELISA are not that sensitive to detect at such low concentrations and when a person develops autoimmunity like T1D, the immune system may be trigger T-regulatory cells to shut down B-cells leading to lack of antibodies in the disease state which existed in normal state. Random peptides on our array mapped to certain parts of known antigens (predicted epitopes). In order to test our prediction we spotted the real peptides corresponding to our mapped random peptides and found our prediction to be true for both high and low binder peptides. One key thing I noticed in the mapping process is when random peptides are mapped using GUITOPE, there is lack of specificity since random peptides mapped not only to known antigens but also to some random antigens. I calibrated the score of GUITOPE (described in chapter 5) and found that now the peptides had a more specific behavior. One of the possible future works would be to test this calibration on other data sets and if successful incorporate that to GUITOPE. Another limitation I encountered during the mapping process is the required prior knowledge of antigens. Currently GUITOPE requires the input of a protein sequence, It would be best if a high throughput system can be designed which can automatically maps random peptides to the whole proteome for discovering unknown targets.

With the advent of immunosignaturing technology, the main idea our center had in mind was diagnostics but little different from what people having been doing so far in terms of looking at heterogeneous population rather than homogenous population. Although our aim is health monitoring of an individual, the technology also has data for population studies. One of the advantages of dealing with population data is the knowledge obtained from so many subjects immunosignature which helps in formulating the baseline. I collected all the data of healthy patients that were run in the past over our initially CIM 10 K V.1 arrays and analyzed the patterns of 10,000 peptides in a normal population I call the ‘Standard Normal Signature’. During the analysis, I encountered very interesting patterns in the 10k random peptides in normal population. While 82% of the peptides the reactivity was distributed as a normal distribution, there was a handful peptides that followed interesting distributions like uniform and bimodal.

Immunosignaturing relies on the immunological behavior patterns of an individual's antibody repertoire. Unlike genetics, I feel immune system biomarkers are more homogenous in terms of their basic variation in the population. Since we were doing lot of population studies on immunosignaturing, I studied factors affecting immunosignatures like age, sex and geographical location. Towards the end of my thesis work, I became interested in the health monitoring aspect of immunosignaturing which relies on taking signatures of individual at regular intervals to warn against any aberrations in ones immunosignature. For this, we required a measurement like an immune system temperature that we can calculate through our technology. Before moving to a mathematical choice of measurements, I realized a behavior of the immune system which might be related from traditional genomics. In gene expression studies, cells are known to be tightly regulated and controlled hence regulating the genes that are expressed during a normal state of an individual but when a person is sick, the cell enters into a chaotic state making gene expression more scattered. Looking into computational metrics like context mining, researchers have classified a person as disease or normal looking into the chaotic behavior of gene expression studies. I proposed that this might also be true when working the immune system. When we are in a normal state, the immune system produces a normal set of antibody repertoire hence their binding pattern to random sequence peptides would follow a certain distribution but when we suffer from disease this antibody repertoire distribution to random peptides would change. One of my lab members, Kurt proposed ‘Entropy’ as a single metric but I found it to be too averaged and did not significantly change when a person is exposed to a disease. I proposed ‘Coefficient of Variation’ as a simple metrics that might be useful to take a ‘temperature’ by immunosignaturing. This metric when I test on different case studies of immunosignaturing showed promise in changing significantly during a disease course, consistent in a normal population, and independent on the choice of random peptides. This is a very simple metric and in the future other computational and mathematical metrics should be tested as their potential to take temperature like measurements through our arrays.

While working on different case studies of immunosignaturing, I realized there is a great potential of this technology both as a population diagnostics, health monitoring as well as discovering drug targets. Since immunosignaturing has its roots in microarrays it is extremely useful to pay more attention to data analysis and developing mathematical and computational tools for efficient analysis of immunosignatures. In my Ph.D. work I laid out some ground work about unique potential of immunosignature and opened doors for other computational scientists to explore more interesting aspects of immunosignaturing.

REFERENCES

Ada, G. L., and P. D. Jones. 1986. The immune response to influenza infection. *Current topics in microbiology and immunology* 128:1-54.

Administration, Energy Information. 2009. Annual Energy Review, 2008.

Aha, David W., Dennis Kibler, and Marc K. Albert. 1991. Instance-based learning algorithms. *Machine Learning* 6 (1):37-66.

Amur, S., F. W. Frueh, L. J. Lesko, and S. M. Huang. 2008. Integration and use of biomarkers in drug development, regulation and clinical practice: a US regulatory perspective. *Biomark Med* 2 (3):305-11.

Anderson, Karen S., and Joshua LaBaer. 2005. The Sentinel Within:â€‰ Exploiting the Immune System for Cancer Biomarkersâ€ *Journal of Proteome Research* 4 (4):1123-1133.

Bacarese-Hamilton, T., J. Gray, and A. Crisanti. 2003. Protein microarray technology for unraveling the antibody specificity repertoire against microbial proteomes. *Curr Opin Mol Ther* 5 (3):278-84.

Barbosa, F., Jr., J. E. Tanus-Santos, R. F. Gerlach, and P. J. Parsons. 2005. A critical review of biomarkers used for monitoring human exposure to lead: advantages, limitations, and future needs. *Environ Health Perspect* 113 (12):1669-74.

Bell, A. J., and T. J. Sejnowski. 1995. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput* 7 (6):1129-59.

Bialek, K., A. Swistowski, and R. Frank. 2003. Epitope-targeted proteome analysis: towards a large-scale automated protein-protein-interaction mapping utilizing synthetic peptide arrays. *Anal Bioanal Chem* 376 (7):1006-13.

Boltz, K. W., M. J. Gonzalez-Moa, P. Stafford, S. A. Johnston, and S. A. Svarovsky. 2009. Peptide microarrays for carbohydrate recognition. *Analyst* 134 (4):650-2.

Bonifacio, E., V. Lampasona, L. Bernasconi, and A. G. Ziegler. 2000. Maturation of the humoral autoimmune response to epitopes of GAD in preclinical childhood type 1 diabetes. *Diabetes* 49 (2):202-8.

Braga-Neto, Ulisses, and Edward Dougherty. 2004. Bolstered error estimation. *Pattern Recognition* 37 (6):1267-1281.

Braga-Neto, Ulisses M., and Edward R. Dougherty. 2004. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* 20 (3):374-380.

Brazma, A. 2009. Minimum Information About a Microarray Experiment (MIAME)--successes, failures, challenges. *ScientificWorldJournal* 9:420-3.

Breiman, Leo. 2001. Random Forests. *Mach. Learn.* 45 (1):5-32.

Brichory, F., D. Beer, F. Le Naour, T. Giordano, and S. Hanash. 2001. Proteomics-based identification of protein gene product 9.5 as a tumor antigen that induces a humoral immune response in lung cancer. *Cancer Res* 61 (21):7908-12.

Brown, Justin, Phillip Stafford, Stephen Johnston, and Valentin Dinu. 2011. Statistical Methods for Analyzing Immunosignatures. *BMC Bioinformatics* 12 (1):349.

Brydak, L. B., and M. Machala. 2000. Humoral Immune Response to Influenza Vaccination in Patients from High Risk Groups. *Drugs* 60 (1):35-53.

Burbelo, P. D., K. E. Bren, K. H. Ching, E. S. Gogineni, S. Kottilil, J. I. Cohen, J. A. Kovacs, and M. J. Iadarola. 2007. LIPS arrays for simultaneous detection of antibodies against partial and whole proteomes of HCV, HIV and EBV. *Mol Biosyst* 7 (5):1453-62.

Burbelo, P. D., K. H. Ching, E. R. Bush, B. L. Han, and M. J. Iadarola. 2010. Antibody-profiling technologies for studying humoral responses to infectious agents. *Expert Rev Vaccines* 9 (6):567-78.

Burbelo, P. D., K. H. Ching, T. L. Mattson, J. S. Light, L. R. Bishop, and J. A. Kovacs. 2007. Rapid antibody quantification and generation of whole proteome antibody response profiles using LIPS (luciferase immunoprecipitation systems). *Biochemical & Biophysical Research Communications* 352 (4):889-95.

Burbelo, P. D., H. Hirai, A. T. Issa, A. Kingman, A. Lernmark, S. A. Ivarsson, A. L. Notkins, and M. J. Iadarola. 2010. Comparison of radioimmunoprecipitation with luciferase immunoprecipitation for autoantibodies to GAD65 and IA-2beta. *Diabetes Care* 33 (4):754-6.

Burbelo, P. D., Y. Hoshino, H. Leahy, T. Krogmann, R. L. Hornung, M. J. Iadarola, and J. I. Cohen. 2009. Serological diagnosis of human herpes simplex virus type 1 and 2 infections by luciferase immunoprecipitation system assay. *Clin Vaccine Immunol* 16 (3):366-71.

Burgoyne, F. H. 1961. The pathology of diabetes mellitus. *Can Med Assoc J* 84:1415-7.

Cenci, S., and R. Sitia. 2007. Managing and exploiting stress in the antibody factory. *FEBS Lett* 581 (19):3652-7.

Cessie, S. Le, and J. C. Van Houwelingen. 1992. Ridge Estimators in Logistic Regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41 (1):191-201.

Chapman, K. 2002. The ProteinChip Biomarker System from Ciphergen Biosystems: a novel proteomics platform for rapid biomarker discovery and validation. *Biochem Soc Trans* 30 (2):82-7.

Chase, B. A., S. A. Johnston, and J. B. Legutki. 2012. Evaluation of biological sample preparation for immunosignature-based diagnostics. *Clin Vaccine Immunol* 19 (3):352-8.

Chaudhuri, B. B., and U. Bhattacharya. 2000. Efficient training and improved performance of multilayer perceptron in pattern classification. *Neurocomputing* 34 (1,Äì4):11-27.

Cleary, John, and Leonard Trigg. 1995. K*: An Instance-based Learner Using an Entropic Distance Measure. Paper read at In Proceedings of the 12th International Conference on Machine Learning.

Colnaghi, M. I., S. Menard, and G. D. Porta. 1977. Natural anti-tumor serum reactivity in BALB/c mice. II. Control by regulator T-cells. *Int J Cancer* 19 (2):275-80.

Cooperman, J., R. Neely, D. T. Teachey, S. Grupp, and J. K. Choi. 2004. Cell division rates of primary human precursor B cells in culture reflect in vivo rates. *Stem Cells* 22 (6):1111-20.

Cox, Rebecca J., Karl A. Brokstad, Mark A. Zuckerman, John M. Wood, Lars R. Haaheim, and John S. Oxford. 1994. An early humoral immune response in peripheral blood following parenteral inactivated influenza vaccination. *Vaccine* 12 (11):993-999.

Cwirla, S. E., E. A. Peters, R. W. Barrett, and W. J. Dower. 1990. Peptides on phage: a vast library of peptides for identifying ligands. *Proc Natl Acad Sci U S A* 87 (16):6378-82.

Davies, H. A. 2000. The ProteinChip System from Ciphergen: a new technique for rapid, micro-scale protein biology. *J Mol Med (Berl)* 78 (7):B29.

de la Fuente, A., P. Brazhnik, and P. Mendes. 2002. Linking the genes: inferring quantitative gene networks from microarray data. *Trends Genet* 18 (8):395-8.

Derda, R., S. K. Tang, S. C. Li, S. Ng, W. Matochko, and M. R. Jafari. 2011. Diversity of phage-displayed libraries of peptides during panning and amplification. *Molecules* 16 (2):1776-803.

Deshpande, Vikram, Mari Mino-Kenudson, William Brugge, and Gregory Y. Lauwers. 2005. Autoimmune Pancreatitis: More Than Just a Pancreatic Disease?A Contemporary Review of Its Pathology. *Archives of Pathology & Laboratory Medicine* 129 (9):1148-1154.

Diamandis, E. P., and D. E. van der Merwe. 2005. Plasma protein profiling by mass spectrometry for cancer diagnosis: opportunities and limitations. *Clin Cancer Res* 11 (3):963-5.

DiFronzo, L. A., R. K. Gupta, R. Essner, L. J. Foshag, S. J. O'Day, L. A. Wanek, S. L. Stern, and D. L. Morton. 2002. Enhanced humoral immune response correlates with improved disease-free and overall survival in American Joint Committee on Cancer stage II melanoma patients receiving adjuvant polyvalent vaccine. *J Clin Oncol* 20 (15):3242-8.

Dotan, N., R. T. Altstock, M. Schwarz, and A. Dukler. 2006. Anti-glycan antibodies as biomarkers for diagnosis and prognosis. *Lupus* 15 (7):442-50.

Duffy, M. J., D. Evoy, and E. W. McDermott. 2010. CA 15-3: uses and limitation as a biomarker for breast cancer. *Clin Chim Acta* 411 (23-24):1869-74.

Dugernier, Thierry L., Pierre-Francois Laterre, Xavier Wittebole, Jean Roeseler, Dominique Latinne, Marc S. Reynaert, and Jerome Pugin. 2003. Compartmentalization of the Inflammatory Response during Acute Pancreatitis: Correlation with Local and Systemic Complications. *Am. J. Respir. Crit. Care Med.* 168 (2):148-157.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95 (25):14863-8.

Ekbom, Anders, Joseph K. McLaughlin, Britt-Marie Karlsson, Olof Nyr√©n, Gloria Gridley, Hans-Olov Adami, and Joseph F. Fraumeni. 1994. Pancreatitis and Pancreatic Cancer: a Population-Based Study. *Journal of the National Cancer Institute* 86 (8):625-627.

Elias, Joshua E., Wilhelm Haas, Brendan K. Faherty, and Steven P. Gygi. 2005. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Meth* 2 (9):667-675.

Emilsson, V., G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G. B. Walters, S. Gunnarsdottir, M. Mouy, V. Steinthorsdottir, G. H. Eiriksdottir, G. Bjornsdottir, I. Reynisdottir, D. Gudbjartsson, A. Helgadottir, A. Jonasdottir, U. Styrkarsdottir, S.

Gretarsdottir, K. P. Magnusson, H. Stefansson, R. Fossdal, K. Kristjansson, H. G. Gislason, T. Stefansson, B. G. Leifsson, U. Thorsteinsdottir, J. R. Lamb, J. R. Gulcher, M. L. Reitman, A. Kong, E. E. Schadt, and K. Stefansson. 2008. Genetics of gene expression and its effect on disease. *Nature* 452 (7186):423-8.

Fang, Y., A. G. Frutos, and J. Lahiri. 2002. Membrane protein microarrays. *Journal of the American Chemical Society* 124 (11):2394-5.

Fineberg, S. E., T. Kawabata, D. Finco-Kent, C. Liu, and A. Krasner. 2005. Antibody response to inhaled insulin in patients with type 1 or type 2 diabetes. An analysis of initial phase II and III inhaled insulin (Exubera) trials and a two-year extension trial. *J Clin Endocrinol Metab* 90 (6):3287-94.

Forster, I., and K. Rajewsky. 1990. The bulk of the peripheral B-cell pool in mice is stable and not rapidly renewed from the bone marrow. *Proc Natl Acad Sci U S A* 87 (12):4781-4.

Friedman, J, T Hastie, and R Tibshirani. 1998. Additive Logistic Regression: a Statistical View of Boosting. *Technical Report, Department of Statistics, Standford University*:1 - 45.

Friedman, Nir, Dan Geiger, and Moises Goldszmidt. 1997. Bayesian Network Classifiers. *Machine Learning* 29 (2):131-163.

Fujita, A., J. R. Sato, O. Rodrigues Lde, C. E. Ferreira, and M. C. Sogayar. 2006. Evaluating different methods of microarray data normalization. *BMC Bioinformatics* 7:469.

G¸venir, H. Altay, and Murat «akIr. 2010. Voting features based classifier with feature construction and its application to predicting financial distress. *Expert Systems with Applications* 37 (2):1713-1718.

Gardner, M. W., and S. R. Dorling. 1998. Artificial neural networks (the multilayer perceptron),Äîa review of applications in the atmospheric sciences. *Atmospheric Environment* 32 (14,Äì15):2627-2636.

Geijersstam, V., M. Kibur, Z. Wang, P. Koskela, E. Pukkala, J. Schiller, M. Lehtinen, and J. Dillner. 1998. Stability over time of serum antibody levels to human papillomavirus type 16. *J Infect Dis* 177 (6):1710-4.

Haab, Brian B. 2003. Methods and applications of antibody microarrays in cancer research. *Proteomics* 3 (11):2116-2122.

Haffner, S. M., S. Lehto, T. Ronnemaa, K. Pyorala, and M. Laakso. 1998. Mortality from coronary heart disease in subjects with type 2 diabetes and in nondiabetic subjects with and without prior myocardial infarction. *N Engl J Med* 339 (4):229-34.

Hall, D. A., J. Ptacek, and M. Snyder. 2007. Protein microarray technology. *Mech Ageing Dev* 128 (1):161-7.

Hall, M. A. 1998. Correlation-based Feature Subset Selection for Machine Learning, PhD Thesis, University of Waikato, Hamilton, New Zealand.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* 11 (1):10-18.

Halperin, R. F., P. Stafford, J. S. Emery, K. A. Navalkar, and S. A. Johnston. 2012. GuiTope: an application for mapping random-sequence peptides to protein sequences. *BMC Bioinformatics* 13:1.

Halperin, R. F., P. Stafford, and S. A. Johnston. 2011. Exploring antibody recognition of sequence space through random-sequence peptide microarrays. *Mol Cell Proteomics* 10 (3):M110 000786.

Hamilton, J. D. 1953. Pathology of diabetes mellitus. *Diabetes* 2 (3):180-3.

Hartigan, J. A. 1985. Statistical theory in clustering. *Journal of Classification* 2 (1):63-76.

Hastie, Trevor, and Robert Tibshirani. 1998. Classification by Pairwise Coupling. In *Advances in Neural Information Processing Systems*: MIT Press.

Hawa, M. I., D. Fava, F. Medici, Y. J. Deng, A. L. Notkins, G. De Mattia, and R. D. Leslie. 2000. Antibodies to IA-2 and GAD65 in type 1 and type 2 diabetes: isotype restriction and polyclonality. *Diabetes Care* 23 (2):228-33.

Hedenfalk, Ingrid, David Duggan, Yidong Chen, Michael Radmacher, Michael Bittner, Richard Simon, Paul Meltzer, Barry Gusterson, Manel Esteller, Mark Raffeld, Zohar Yakhini, Amir Ben-Dor, Edward Dougherty, Juha Kononen, Lukas Bubendorf, Wilfrid Fehrle, Stefania Pittaluga, Sofia Gruvberger, Niklas Loman, Oskar Johannsson, H√•kan Olsson, Benjamin Wilfond, Guido Sauter, Olli-P. Kallioniemi, √Öke Borg, and Jeffrey Trent. 2001. Gene-Expression Profiles in Hereditary Breast Cancer. *New England Journal of Medicine* 344 (8):539-548.

Hennessy, B. T., M. Murph, M. Nanjundan, M. Carey, N. Auersperg, J. Almeida, K. R. Coombes, J. Liu, Y. Lu, J. W. Gray, and G. B. Mills. 2008. Ovarian cancer: linking genomics to new target discovery and molecular markers--the way ahead. *Adv Exp Med Biol* 617:23-40.

Heyer, L. J., S. Kruglyak, and S. Yooseph. 1999. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* 9 (11):1106-15.

Hilsenbeck, S. G., W. E. Friedrichs, R. Schiff, P. O'Connell, R. K. Hansen, C. K. Osborne, and S. A. Fuqua. 1999. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J Natl Cancer Inst* 91 (5):453-9.

Hochberg, Y., and Y. Benjamini. 1990. More powerful procedures for multiple significance testing. *Stat Med* 9 (7):811-8.

Hooks, John J., Haralampos M. Moutsopoulos, Shirley A. Geis, Neil I. Stahl, John L. Decker, and Abner Louis Notkins. 1979. Immune Interferon in the Circulation of Patients with Autoimmune Disease. *New England Journal of Medicine* 301 (1):5-8.

Hruban, R. H., M. Goggins, J. Parsons, and S. E. Kern. 2000. Progression model for pancreatic cancer. *Clin Cancer Res* 6 (8):2969-72.

Hruban, R. H., C. Iacobuzio-Donahue, R. E. Wilentz, M. Goggins, and S. E. Kern. 2001. Molecular pathology of pancreatic cancer. *Cancer J* 7 (4):251-8.

Huxley, R., A. Ansary-Moghaddam, A. Berrington de Gonzalez, F. Barzi, and M. Woodward. 2005. Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies. *Br J Cancer* 92 (11):2076-2083.

Ian H. Witten, Eibe Frank, Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition).*

Imagawa, A., T. Hanafusa, J. Miyagawa, and Y. Matsuzawa. 2000. A novel subtype of type 1 diabetes mellitus characterized by a rapid onset and an absence of diabetes-related antibodies. Osaka IDDM Study Group. *N Engl J Med* 342 (5):301-7.

Inoue, Hiromu, Hiroyuki Miyatani, Yukihisa Sawada, and Yukio Yoshida. 2006. A Case Of Pancreas Cancer With Autoimmune Pancreatitis. *Pancreas* 33 (2):208-209 10.1097/01.mpa.0000232329.35822.3a.

John, George H., and Pat Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. Paper read at Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, at San Mateo.

Jupiter, D. C., and V. VanBuren. 2008. A visual data mining tool that facilitates reconstruction of transcription regulatory networks. *PLoS One* 3 (3):e1717.

Kari, L., and L. F. Landweber. 2000. Computing with DNA. *Methods Mol Biol* 132:413-30.

Katsilambros, N., E. Diakoumopoulou, I. Ioannidis, S. Liatis, K. Makrilakis, N. Tentolouris, and P. Tsapogas. 2006. Pathophysiology of Type 2 Diabetes. In *Diabetes in Clinical Practice*: John Wiley & Sons, Ltd.

Keerthi, S.S., S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy. 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation* 13 (3):637-649.

Kiehntopf, Michael, Robert Siegmund, and Thomas Deufel. 2007. Use of SELDI-TOF mass spectrometry for identification of new biomarkers: potential and limitations. *Clinical Chemistry and Laboratory Medicine* 45 (11):1435-1449.

Kijanka, G., S. Ipcho, S. Baars, H. Chen, K. Hadley, A. Beveridge, E. Gould, and D. Murphy. 2009. Rapid characterization of binding specificity and cross-reactivity of antibodies using recombinant human protein arrays. *J Immunol Methods* 340 (2):132-7.

Kukreja, M., S. A. Johnston, and P. Stafford. 2012. Comparative study of classification algorithms for immunosignaturing data. *BMC Bioinformatics* 13 (1):139.

Kurian, S., Y. Grigoryev, S. Head, D. Campbell, T. Mondala, and D. R. Salomon. 2007. Applying genomics to organ transplantation medicine in both discovery and validation of biomarkers. *Int Immunopharmacol* 7 (14):1948-60.

Kurian, S. M., R. Heilman, T. S. Mondala, A. Nakorchevsky, J. A. Hewel, D. Campbell, E. H. Robison, L. Wang, W. Lin, L. Gaber, K. Solez, H. Shidban, R. Mendez, R. L. Schaffer, J. S. Fisher, S. M. Flechner, S. R. Head, S. Horvath, J. R. Yates, C. L. Marsh, and D. R. Salomon. 2009. Biomarkers for early and late stage chronic allograft nephropathy by proteogenomic profiling of peripheral blood. *PLoS One* 4 (7):e6212.

L. Ackermann, Bradley, John E. Hale, and Kevin L. Duffin. 2006. The Role of Mass Spectrometry in Biomarker Discovery and Measurement. *Current Drug Metabolism* 7 (5):525-539.

Lamont, R. J., M. Meila, Q. Xia, and M. Hackett. 2006. Mass spectrometry-based proteomics and its application to studies of Porphyromonas gingivalis invasion and pathogenicity. *Infect Disord Drug Targets* 6 (3):311-25.

Lan, M. S., C. Wasserfall, N. K. Maclaren, and A. L. Notkins. 1996. IA-2, a transmembrane protein of the protein tyrosine phosphatase family, is a major autoantigen in insulin-dependent diabetes mellitus. *Proc Natl Acad Sci U S A* 93 (13):6367-70.

Landwehr, Niels, Mark Hall, and Eibe Frank. 2005. Logistic Model Trees. *Mach. Learn.* 59 (1-2):161-205.

Le-Niculescu, H., S. M. Kurian, N. Yehyawi, C. Dike, S. D. Patel, H. J. Edenberg, M. T. Tsuang, D. R. Salomon, J. I. Nurnberger, Jr., and A. B. Niculescu. 2009. Identifying blood biomarkers for mood disorders using convergent functional genomics. *Mol Psychiatry* 14 (2):156-74.

Legutki, Joseph Barten, D. Mitchell Magee, Phillip Stafford, and Stephen Albert Johnston. 2010. A general method for characterization of humoral immunity induced by a vaccine or infection. *Vaccine* 28 (28):4529-4537.

Lennon, Vanda A., Jon M. Lindstrom, and Marjorie E. Seybold. 1976. EXPERIMENTAL AUTOIMMUNE MYASTHENIA GRAVIS: CELLULAR AND HUMORAL IMMUNE RESPONSES*. *Annals of the New York Academy of Sciences* 274 (1):283-299.

Leslie, R. D. G., M. A. Atkinson, and Abner L. Notkins. 1999. Autoantigens IA-2 and GAD in Type I (Insulin-Dependent) Diabetes. *Diabetologia* 42:3-14.

Li, Donghui, Keping Xie, Robert Wolff, and James L. Abbruzzese. 2004. Pancreatic cancer. *The Lancet* 363 (9414):1049-1057.

Li, Jinong, Rosaria Orlandi, C. Nicole White, Jason Rosenzweig, Jing Zhao, Ettore Seregni, Daniele Morelli, Yinhua Yu, Xiao-Ying Meng, Zhen Zhang, Nancy E. Davidson, Eric T. Fung, and Daniel W. Chan. 2005. Independent Validation of Candidate Breast Cancer Serum Biomarkers Identified by Mass Spectrometry. *Clin Chem* 51 (12):2229-2235.

Li, Tao, Chengliang Zhang, and Mitsunori Ogihara. 2004. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* 20 (15):2429-2437.

Liu, H., J. Li, and L. Wong. 2002. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform* 13:51-60.

Long, A. D., H. J. Mangalam, B. Y. Chan, L. Tolleri, G. W. Hatfield, and P. Baldi. 2001. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *J Biol Chem* 276 (23):19937-44.

Lowenfels, Albert B., Patrick Maisonneuve, Giorgio Cavallini, Rudolf W. Ammann, Paul G. Lankisch, Jens R. Andersen, Eugene P. Dimagno, Ake Andren-Sandberg, and Lennart Domellof. 1993. Pancreatitis and the Risk of Pancreatic Cancer. *New England Journal of Medicine* 328 (20):1433-1437.

Lucas, J. E., C. M. Carvalho, J. L. Chen, J. T. Chi, and M. West. 2009. Cross-study projections of genomic biomarkers: an evaluation in cancer genomics. *PLoS One* 4 (2):e4523.

Mattoon, D., G. Michaud, J. Merkel, and B. Schweitzer. 2005. Biomarker discovery using protein microarray technology platforms: antibody-antigen complex profiling. *Expert Rev Proteomics* 2 (6):879-89.

Meloen, R. H., W. C. Puijk, and J. W. Slootstra. 2000. Mimotopes: realization of an unlikely concept. *J Mol Recognit* 13 (6):352-9.

Merbl, Y., R. Itzchak, T. Vider-Shalit, Y. Louzoun, F. J. Quintana, E. Vadai, L. Eisenbach, and I. R. Cohen. 2009. A systems immunology approach to the host-tumor interaction: large-scale patterns of natural autoantibodies distinguish healthy and tumor-bearing mice. *PLoS One* 4 (6):e6053.

Millen, A. E., M. Pettinger, J. L. Freudenheim, R. D. Langer, C. A. Rosenberg, Y. Mossavar-Rahmani, C. M. Duffy, D. S. Lane, A. McTiernan, L. H. Kuller, A. M. Lopez, and J. Wactawski-Wende. 2009. Incident invasive breast cancer, geographic location of residence, and reported average time spent outside. *Cancer Epidemiol Biomarkers Prev* 18 (2):495-507.

Montgomery, Douglas C. 2009. Design and Analysis of Experiments (7th Edition): John Wiley & Sons.

Nahtman, Tatjana, Alexander Jernberg, Shahnaz Mahdavifar, Johannes Zerweck, Mike Schutkowski, Markus Maeurer, and Marie Reilly. 2007. Validation of peptide epitope microarray experiments and extraction of quality data. *Journal of Immunological Methods* 328 (1,Äì2):1-13.

Noel, R. A., D. K. Braun, R. E. Patterson, and G. L. Bloomgren. 2009. Increased risk of acute pancreatitis and biliary disease observed in patients with type 2 diabetes: a retrospective cohort study. *Diabetes Care* 32 (5):834-8.

Notkins, A. L., and A. Lernmark. 2001. Autoimmune type 1 diabetes: resolved and unresolved issues. *J Clin Invest* 108 (9):1247-52.

Okamoto, O. K. 2009. Cancer stem cell genomics: the quest for early markers of malignant progression. *Expert Rev Mol Diagn* 9 (6):545-54.

Okazaki, K., and T. Chiba. 2002. Autoimmune related pancreatitis. *Gut* 51 (1):1-4.

Orchekowski, Randal, Darren Hamelinck, Lin Li, Ewa Gliwa, Matt VanBrocklin, Jorge A. Marrero, George F. Vande Woude, Ziding Feng, Randall Brand, and Brian B. Haab. 2005. Antibody Microarray Profiling Reveals Individual and Combined Serum Proteins Associated with Pancreatic Cancer. *Cancer Research* 65 (23):11193-11202.

Ozcan, U., Q. Cao, E. Yilmaz, A. H. Lee, N. N. Iwakoshi, E. Ozdelen, G. Tuncman, C. Gorgun, L. H. Glimcher, and G. S. Hotamisligil. 2004. Endoplasmic reticulum stress links obesity, insulin action, and type 2 diabetes. *Science* 306 (5695):457-61.

Platt, J. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. *Advances in Kernel Methods - Support Vector Learning*.

Quinlan, J.R. 1992. Learning with continuous classes. Paper read at Proceedings of the 5th Australian Joint Conference on Artificial Intelligence

Quinlan, JR. 1996. Bagging, Boosting and C4.5. *AAAI/IAAI* 1.

Ramachandran, N., J. V. Raphael, E. Hainsworth, G. Demirkan, M. G. Fuentes, A. Rolfs, Y. Hu, and J. LaBaer. 2008. Next-generation high-density self-assembling functional protein arrays. *Nature Methods* 5 (6):535-8.

Reimer, Ulf, Ulrich Reineke, and Jens Schneider-Mergener. 2002. Peptide arrays: from macro to micro. *Current Opinion in Biotechnology* 13 (4):315-320.

Reineke, U. 2009. Antibody epitope mapping using de novo generated synthetic peptide libraries. *Methods in Molecular Biology* 524:203-11.

Restrepo, Lucas, Phillip Stafford, D. Mitch Magee, and Stephen Albert Johnston. 2011. Application of immunosignatures to the assessment of Alzheimer's disease. *Annals of Neurology* 70 (2):286-295.

Robinson, W. H., C. DiGennaro, W. Hueber, B. B. Haab, M. Kamachi, E. J. Dean, S. Fournel, D. Fong, M. C. Genovese, H. E. de Vegvar, K. Skriner, D. L. Hirschberg, R. I. Morris, S. Muller, G. J. Pruijn, W. J. van Venrooij, J. S. Smolen, P. O. Brown, L. Steinman, and P. J. Utz. 2002. Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nat Med* 8 (3):295-301.

Rotimi, C. N., and L. B. Jorde. 2010. Ancestry and disease in the age of genomic medicine. *N Engl J Med* 363 (16):1551-8.

Saiki, R. K., S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230 (4732):1350-4.

Saluja, A. K., and M. L. Steer. 1999. Pathophysiology of Pancreatitis. *Digestion* 60 (Suppl. 1):27-33.

Salzberg, Steven L. 1994. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning* 16 (3):235-240.

Sawyers, C. L. 2008. The cancer biomarker problem. *Nature* 452 (7187):548-52.

Schellekens, G. A., H. Visser, B. A. de Jong, F. H. van den Hoogen, J. M. Hazes, F. C. Breedveld, and W. J. van Venrooij. 2000. The diagnostic properties of rheumatoid arthritis antibodies recognizing a cyclic citrullinated peptide. *Arthritis Rheum* 43 (1):155-63.

Schena, M., R. A. Heller, T. P. Theriault, K. Konrad, E. Lachenmeier, and R. W. Davis. 1998. Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol* 16 (7):301-6.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270 (5235):467-70.

Schuchhardt, J., D. Beule, A. Malik, E. Wolski, H. Eickhoff, H. Lehrach, and H. Herzel. 2000. Normalization strategies for cDNA microarrays. *Nucleic Acids Res* 28 (10):E47.

Sheehan, K. M., V. S. Calvert, E. W. Kay, Y. Lu, D. Fishman, V. Espina, J. Aquino, R. Speer, R. Araujo, G. B. Mills, L. A. Liotta, E. F. Petricoin, 3rd, and J. D. Wulfkuhle. 2005. Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. *Mol Cell Proteomics* 4 (4):346-55.

Sima, Chao, Sanju Attoor, Ulisses Brag-Neto, James Lowey, Edward Suh, and Edward R. Dougherty. 2005. Impact of error estimation on feature selection. *Pattern Recognition* 38 (12):2472-2482.

Spadone, S., F. de Pasquale, D. Mantini, and S. Della Penna. 2012. A K-means multivariate approach for clustering independent components from magnetoencephalographic data. *Neuroimage*.

Sreekumar, A., B. Laxman, D. R. Rhodes, S. Bhagavathula, J. Harwood, D. Giacherio, D. Ghosh, M. G. Sanda, M. A. Rubin, and A. M. Chinnaiyan. 2004. Humoral immune response to alpha-methylacyl-CoA racemase and prostate cancer. *J Natl Cancer Inst* 96 (11):834-43.

Stafford, P., and M. Brun. 2007. Three methods for optimization of cross-laboratory and cross-platform microarray expression data. *Nucleic Acids Res* 35 (10):e72.

Stafford, P., R. Halperin, J. B. Legutki, D. M. Magee, J. Galgiani, and S. A. Johnston. 2011. Physical characterization of the "immunosignaturing effect". *Mol Cell Proteomics* 11 (4):M111 011593.

———. 2012. Physical characterization of the "immunosignaturing effect". *Mol Cell Proteomics* 11 (4):M111 011593.

Stayoussef, M., J. Benmansour, F. A. Al-Jenaidi, H. B. Said, C. B. Rayana, T. Mahjoub, and W. Y. Almawi. 2011. Glutamic acid decarboxylase 65 and islet cell antigen 512/IA-2 autoantibodies in relation to human leukocyte antigen class II DR and DQ alleles and haplotypes in type 1 diabetes mellitus. *Clin Vaccine Immunol* 18 (6):990-3.

Stemke-Hale, Katherine, Bernhard Kaltenboeck, Fred J. DeGraves, Kathryn F. Sykes, Jin Huang, Chun-hui Bu, and Stephen Albert Johnston. 2005. Screening the whole genome of a pathogen in vivo for individual protective antigens. *Vaccine* 23 (23):3016-3025.

Stockert, E., E. Jager, Y. T. Chen, M. J. Scanlan, I. Gout, J. Karbach, M. Arand, A. Knuth, and L. J. Old. 1998. A survey of the humoral immune response of cancer patients to a panel of human tumor antigens. *J Exp Med* 187 (8):1349-54.

Thurlings, R. M., K. Vos, D. M. Gerlag, and P. P. Tak. 2006. [The humoral response in rheumatoid arthritis and the effect of B-cell depleting therapy]. *Ned Tijdschr Geneeskd* 150 (30):1657-61.

Tugwood, J. D., L. E. Hollins, and M. J. Cockerill. 2003. Genomics and the search for novel biomarkers in toxicology. *Biomarkers* 8 (2):79-92.

Urakami, T., A. Yoshida, J. Suzuki, H. Saito, M. Wada, S. Takahashi, and H. Mugishima. 2009. Differences in prevalence of antibodies to GAD and IA-2 and their titers at diagnosis in children with slowly and rapidly progressive forms of type 1 diabetes. *Diabetes Res Clin Pract* 83 (1):89-93.

US Department of Health and Human Services, Centers for Medicare Services. *National Health Expenditure Data* 2010 [cited. Available from http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-

Reports/NationalHealthExpendData/NationalHealthAccountsProjected.html.

Uttamchandani, M., and S. Q. Yao. 2008. Peptide microarrays: next generation biochips for detection, diagnostics and high-throughput screening. *Curr Pharm Des* 14 (24):2428-38.

Wang, D. G., J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, and E. S. Lander. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280 (5366):1077-82.

Weinberger, Kilian, John Blitzer, and Lawrence Saul. 2006. Distance metric learning for large margin nearest neighbor classification. Paper read at In NIPS.

Weston, Andrea D., and Leroy Hood. 2004. Systems Biology, Proteomics, and the Future of Health Care:‚Äâ Toward Predictive, Preventative, and Personalized Medicine. *Journal of Proteome Research* 3 (2):179-196.

Whiteaker, Jeffrey R., Lei Zhao, Heidi Y. Zhang, Li-Chia Feng, Brian D. Piening, Leigh Anderson, and Amanda G. Paulovich. 2007. Antibody-based enrichment of peptides on magnetic beads for mass-spectrometry-based quantification of serum biomarkers. *Analytical Biochemistry* 362 (1):44-54.

Wilder, R. L. 1995. Neuroendocrine-immune system interactions and autoimmunity. *Annu Rev Immunol* 13:307-38.

Yang, Y., P. Stafford, and Y. Kim. 2011. Segmentation and intensity estimation for microarray images with saturated pixels. *BMC Bioinformatics* 12:462.

Yang, Yee Hwa, Michael J. Buckley, Sandrine Dudoit, and Terence P. Speed. 2002. Comparison of Methods for Image Analysis on cDNA Microarray Data. *Journal of Computational and Graphical Statistics* 11 (1):108-136.

Yeh, H. Y., S. W. Cheng, Y. C. Lin, C. Y. Yeh, S. F. Lin, and V. W. Soo. 2009. Identifying significant genetic regulatory networks in the prostate cancer from microarray data based on transcription factor analysis and conditional independency. *BMC Med Genomics* 2:70.

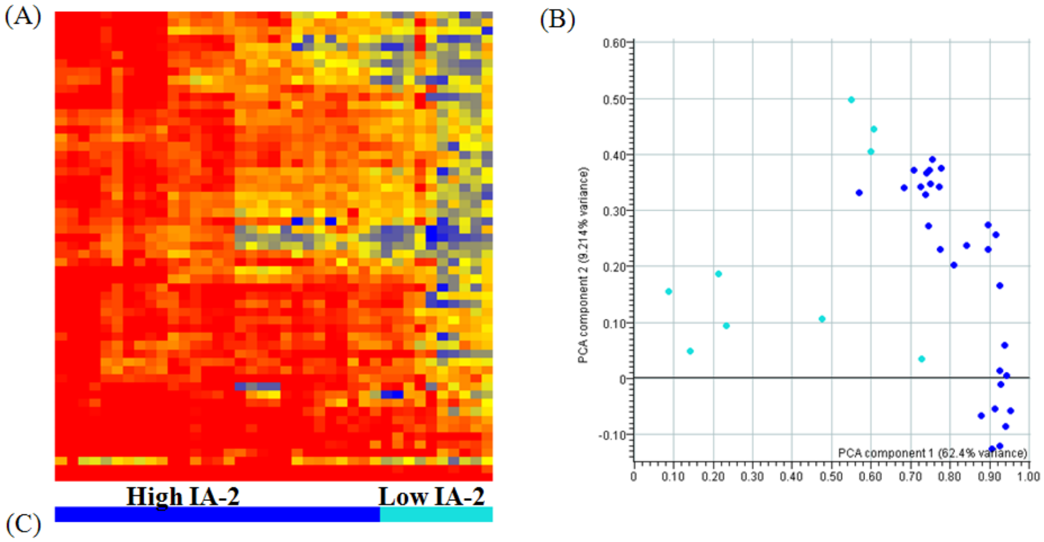
Yu, J, and XW Chen. 2005. Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data. *Bioinformatics* 21 Suppl 1:i487 - i494.

Yu, S., L. C. Tranchevent, X. Liu, W. Glanzel, J. A. Suykens, B. De Moor, and Y. Moreau. 2012. Optimized data fusion for kernel k-means clustering. *IEEE Trans Pattern Anal Mach Intell* 34 (5):1031-9.

APPENDIX

A.TITER ANALYSIS ON T1D SUBJECTS

We collected 29 samples of type 1 diabetes (T1D) having high IA-2 titer and 10 samples having low IA-2 titer. We selected 57 peptides that were 2 fold higher in higher titer samples compared to low titer samples. **Figure A.1** (A) shows heatmap of 57 peptides in high/low titer subjects. (B) shows principal component analysis where high and low titers subjects are separated well. (C) shows classification table using Naïve Bayes classification algorithm using leave one out methodology yielding 97% sensitivity.



(C)

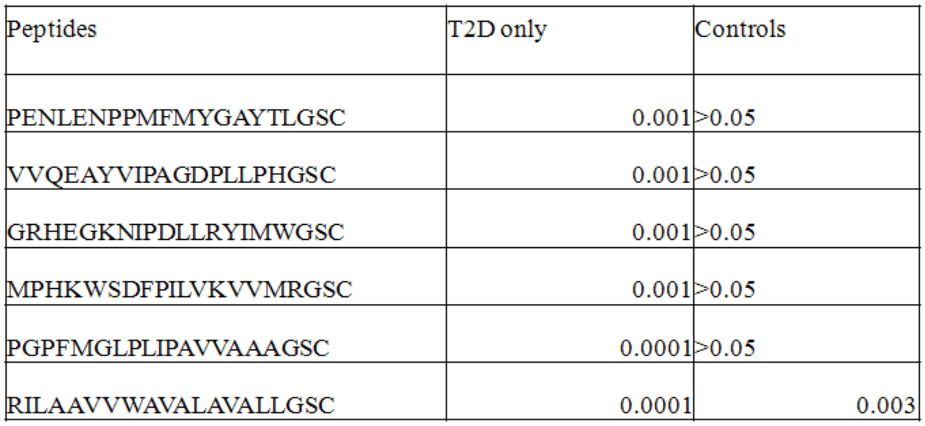
Predicted -> (89.7 %)	High IA-2	LOW
High	28	1
Low	3	7

Figure A.1: Immunosignaturing of high/low IA-2 titer of T1D subjects

B. MAPPING RANDOM PEPTIDES IN TYPE 2 DIABETES TO GADA PROTEIN

We selected 170 peptides that were significantly different in type 2 diabetes immunosignature compared to controls. We mapped these 170 peptides against a suspected autoantigen in T2D pathway and found 6 peptides that aligned. Figure B.1 (A) shows random peptide sequence along with false discovery rate corresponding to the mapping process. (B) shows the GUITOPE analysis for 1000 iterations which indicated 1 predicted epitope.

(A)



Peptides	T2D only	Controls
PENLENPPMFMYGAYTLGSC	0.001	>0.05
VVQEAYVIPAGDPLLPHGSC	0.001	>0.05
GRHEGKNIPDLLRYIMWGSC	0.001	>0.05
MPHKWSDFPILVKVVMRGSC	0.001	>0.05
PGPFMGLPLIPAVVAAAGSC	0.0001	>0.05
RILAAVVWAVALAVALLGSC	0.0001	0.003

(B)

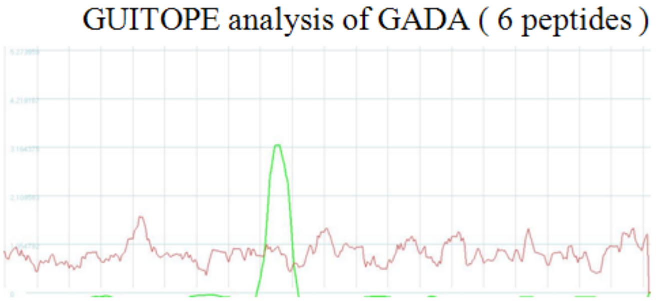


Figure B.1: GUITOPE analysis of T2D random peptides on GADA autoantigen

C. PEPTIDES LIST FOR VALIDATING PEPTIDES IN DECIPHER OF TYPE 1 DIABETES IMMUNOSIGNATURE

We ordered 20 peptides that were mapped on autoantigens of T1D. Table C.1 shows the peptide id. Ups indicate peptide that we predicted to be show higher binding. Down indicates peptides we predicted to be of less binding. Controls indicate peptides selected randomly on the part of the autoantigen.

Table C.1: Peptide list for decipher of T1D

Peptide Name	Ordered Peptides	Length
IA-2 Ups	LMSQGLSWHDDLTQYVIGSC	20
IA-2 Ups	DLTQYVISQEMERIPRLGSC	20
IA-2 Ups	RLPEPGGSSRAGDSSEGGSC	20
IA-2 Ups	EGTPASTRPLLDFRRKVGSC	20
IA-2 Ups	STRPLLDFRRKVNKCYRGSC	20
IA-2 Controls	MRLPGRPGGPGGSGGLRGSC	20
IA-2 Controls	VLLCLLLLGSRPGGCNAGSC	20
Insulin Up	YTPKTRREAEDLQVGQVGSC	20
Insulin Up	QCCTSICSLYQLENYCNGSC	20
Insulin Down	MALWMRLLPLLALLALWGSC	20
Insulin Down	SLQPLALEGSLQKRGIVEGSC	21
Insulin Control	DPAAAFVNQHLCGSHLVGSC	20
GAD Up	MWRAKGTTGFEAHVDKCGSC	20
GAD Up	EYLYNIIKNREGYEMVFGSC	20

GAD down	MSRKHKWKLSGVERANSGSC	20
GAD down	ANSVTWNPHKMMGVPLQGSC	20
GAD down	RHVDVFKLWLMWRAKGTGSC	20
GAD controls	ISNMYAMMIARFKMFPEGSC	20
GAD controls	VKEKGMAALPRLIAFTSGSC	20
GAD controls	IGTDSVILIKCDERGKMGSC	20

D. PATTERNS IN STANDARD NORMAL SIGNATURE

We ran more than 200 healthy samples on CIM 10k V.1 over the period of 5 years. We then analyzed all the normal individual immunosignature in terms of how each peptide reacts in the healthy normal population. I termed this as Standard Normal Signature (SNS). We then found some interesting patterns among 10k peptides among the healthy individuals. While more than 80% of the peptides followed the normal distribution, there were some peptides which followed bimodal and other interesting distributions. Figure D.1 shows the distribution of 10k peptides in Standard Normal Signature. We then asked how many peptides among the 10,000 are saturated in terms of their intensities. We observed that there were 12 peptides that had normalized intensity >1.75 in 95% of the healthy individual immunosignature (Standard Normal Signature). Table D.1 shows 12 peptides list from CIM 10k V.1 which were saturated in SNS.

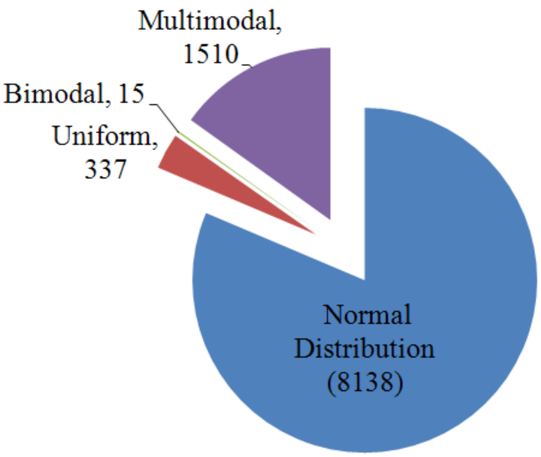


Figure D.1: Patterns in Standard Normal Signature

Table D.1: 12 peptides having intensities >1.75 in SNS

KRKFQRQHSPVRPEFFTGSC
AGAFRERRYKPMMWLHVGSC
PVKYWAKSRVHTRGSWFGSC
YMHRHFEGRGAPMNFRHGSC
RFLRRKPWSMEAHAAQPGSC
KEWQQRKARRYWHQWQDGSC
YRRGWIGMIQRHRIKYEGSC
VKGKLSNVPSWFNHFHSGSC
ARYWWANVDIIIKGGMRGSC
RWRSKYNPRPQYSNEYYGSC
TRMYILHKRWQEAHNVNGSC
VTGVKRPPLYNWTHGNVGSC