A Consolidated Approach to Analyzing Data from High Throughput Protein Microarrays with an Application to Immune Response Profiling in Humans

Mariusz Lubomirski[1*], Michael R. D'Andrea[1], Stanley M. Belkowski[1], Javier Cabrera[2], James M. Dixon[1], and Dhammika Amaratunga[1]

[1]*Johnson & Johnson Pharmaceutical Research & Development LLC* and [2]*Rutgers University*

Abstract

Motivation: DNA microarrays are a well known and established technology in biological and pharmaceutical research providing a wealth of information essential for understanding biological processes and aiding drug development. Protein microarrays are quickly emerging as a follow up technology, which will also begin to experience rapid growth as the challenges in protein to spot methodologies are overcome. Like DNA microarrays, their protein counterparts produce large amounts of data that must be suitably analyzed in order to yield meaningful information that should eventually lead to novel drug targets and biomarkers. Although the statistical management of DNA microarray data has been well described, there is no available report which offers a successful consolidated approach to the analysis of high throughput protein microarray data. We describe the novel application of a statistical methodology to analyze the data from an immune response profiling assay using human protein microarray with over 5000 proteins on each chip.

Introduction

Protein microarray technology is opening a new frontier in the profiling of protein expression. With this rapidly evolving technology, the expression patterns of thousands of proteins can be monitored in high throughput with the objective of selecting a small subset of proteins that are most relevant to the situation under study. This subset could be characterized further as potential biomarkers and targets. As such, it is expected that this technology will create a fascinating new horizon in diagnostic, prognostic and disease progression monitoring. In addition, it allows researchers to study protein functions at various levels and small molecule characterization in terms of efficacy, safety and selectivity, as well as antibody and immune system profiling.

Since the successful demonstration of the first proteome microarray based on yeast (Zhu et al. 2001), many difficulties related to arraying proteins have been recognized (Kusnezow and Hoheisel, 2002). Many of the difficulties in production, isolation and spotting were addressed through the use of bacterial and yeast expression vectors, mass spectrometry techniques and contact printing, respectively (reviewed by Bertone and Snyder, 2005). There has been a rapid evolution in application of these developments to human proteins resulting in arrays, which have been used both alone and in conjunction with other discovery techniques to develop biomarkers and novel targets (Ilyin et al., 2004; Lou et al., 2006).

As with DNA microarray technologies, the large amount of data generated by protein microarrays requires careful handling and appropriate analysis. The result of a data analysis is largely dependent on the quality of the data available but the analysis itself is instrumental in either correctly directing or misleading scientists in their research. Incorrect conclusions can extend the duration of projects, frustrate and further swell already stretched budgets. Every effort must be made to employ the best knowledge and experience to appropriately analyze and deliver correct conclusions. Although studies involving protein microarrays are new for expressional

experimentation, the microarray application of ligand binding is now reaching maturity with respect to the hardware advances and statistical methodologies employed in DNA applications (Amaratunga and Cabrera (2004) provide an extensive review of statistical analysis methodologies for DNA microarrays and also discusses early protein microarrays; Saundaresh et al (2006) discuss data analysis from small antigen protein microarray). Publications, studies and seminars on the subject provide an edge when approaching data generated by protein microarrays and the purpose of this paper is to describe a comprehensive set of statistical methods to apply to accurately analyze protein microarray data. To that end, we describe a successful adaptation of several DNA microarray statistical methodologies to this technology to select a small subset of possible protein targets, which have subsequently become the subject of further assay validation.

Methods

Protein Microarray Experiment

The data analysis methodology we outline in this paper was applied to a human protein microarray experiment. The purpose of the experiment was to identify proteins that are recognized by antibodies present in human serum with the intention to identify autoantibodies. An autoantibody is an antibody that reacts with proteins of the individual in which it was produced, which may lead to conditions called autoimmune diseases. The hallmarks of these autoimmune diseases are high levels of specific antibodies directed to a particular target protein, Nielen (2004). Reduction of the level of the target protein or disregulation of the target protein by the antibody results in pathogenesis of the disease as shown in diabetes where antibodies are formed to insulin and insulin producing cells inhibiting the ability to adequately monitor glucose levels in the system resulting in disease (Itoh 1989). It was discovered that measuring the levels of autoantibodies to insulin or insulin producing islet cells could be predictive of the onset of diabetes in children (Pietropaolo 2005).

Recently this type of serum biomarker has been extended to diseases not classically defined as autoimmune diseases. Detection of disease-related antibodies and autoantibodies may be used as biomarkers to predict disease and disease progression of other types of disease. Circulating IL-8 and anti-IL-8 antibody have been shown to be elevated in the sera of patients with ovarian cancer as compared with healthy controls and have therefore been proposed as potential biomarkers for ovarian cancer (Lokshin 2006). In addition, the presence of specific antibodies in the blood may be used as predictors of inflammatory bowel disease (Israeli 2005). Two of the most common antibodies found in the sera of patients with Crohn's disease and Ulcerative Colitis are anti-Saccharomyces cerevisiae mannan antibodies (ASCA) in Crohn's disease and perinuclear antineutrophil cytoplasm antibodies (pANCA) in ulcerative colitis. It is hoped that the detection of ASCA and pANCA may serve to predict the development of inflammatory bowel disease years in advance of clinical diagnosis of the disease (Israeli 2005). Interestingly, it was hypothesized that even Alzheimer's disease may be a form of an autoimmune disease, as the presence of Ig-positive neurons in brain tissues has been reported (D'Andrea, 2003; 2005a; 2005b).

In a related experiment, we examined a set of five normal serums and two sets of diseased serums, corresponding to two autoimmune diseases, to give a total of 20 serum samples. Using commercially available protein microarrays, serum antibodies from the normal and diseased patients were applied to interact with the proteins fixed on the array. Because the identity of each protein on the array is known, the autoantibodies present in the disease (or control) serum can be identified based on the proteins with which they interact. Our goal is to find autoantibodies that would form the basis for diagnostic, prognostic and disease progression biomarkers for the two human disease states. These potential biomarkers may be used separately to distinguish diseased from normal individuals or in a panel of several biomarkers to strengthen the decision making process.

Human protein microarray chip used in this experiment is Invitrogen high throughput device with 5056 proteins immobilized using hydrophobic surfaces. The proteins are expressed as N-terminal glutathione S-transferase (GST) fusion proteins, purified and double spotted within distinct subarrays with a number of assay specific positive controls and a set of negative controls included for quality assessment purposes. The proteins are mounted on nitrocellulose-coated glass slides, the serum and subsequent secondary fluorescent-labeled antibody were added, washed, dried and then scanned (Figure 1). The measurement of the fluorescence signal was corrected by the background signal, which is measured at a radius a small distance away from the circular feature.

INSERT FIGURE 1 HERE

Implementation

Data Processing

Since proteins do not express but either bind or not, the ideal measured signal should be in a form of binary sequence which then could be used for the separation of active entities. Separation of responders from non-responders in microarray experiments is a difficult task since microarrays are inherently noisy devices. In particular the immobilization techniques used here make the microarray prone to suffer from auto-fluorescence and unspecific binding The result is that the signal spans continuum rather than resembling dichotomes sequence. This is further compounded by the small number of samples available for investigation, which is usually the case in most early research and development situations. Based on lessons learned from the analysis of DNA microarrays, we take a consolidated approach that, for DNA microarrays, has performed well in selecting genes with high validation rates (Amaratunga and Cabrera, 2004). The essence of the analysis is to pull all the arrays together using normalization and variance stabilizing transformation and to thereby enable the application of a variety of statistical tests and data mining methodologies. The complete approach consists of four steps:

Step 1: A preprocessing step to suitably transform, normalize and quality check the data.
Step 2: A proof of concept step to verify that the purported differences among the distinct groups is indeed observable in the data.
Step 3: A feature detection step to identify the proteins and signatures associated with the differences among groups.
Step 4: A validation step to corroborate the features identified.

We will now describe this approach using the following notation: X_{ij} denotes the spot intensity corresponding to the *i*th protein on the *j*th array.

The preprocessing step

The first step is normalization, which is used to reduce disparities between arrays caused by technical effects such as scanner and operator effects. To apply normalization, a mock reference array $\{M_i\}$ is created by taking the median across arrays: $M_i = \text{median}_j(X_{ij})$. All arrays are normalized to this reference array using quantile normalization, whose objective is to make the distributions of the transformed spot intensities, $\{X_{ij}\}$, as similar as possible across the microarrays. To normalize the *j*th array to the reference array, sort the values of each of the arrays and use linear interpolation to predict the reference array value from the value on the array being normalized. The quantile normalized arrays should all have a distribution identical to the distribution of the mock reference array, unless there are ties that could cause small discrepancies.

Next the data are transformed to reduce the skewness in the data and the heterogeneity of variances across proteins. Although the log transformation is the most commonly used transformation for microarray data, we used a variant, a started log transformation: $Y_{ij} = \log(X_{ij}+c)$

as suggested by Rocke and Durbin (2002) as it is more effective at achieving our objectives and remains reasonably interpretable. The value c was chosen to optimize a criterion that is a composite of three measures: the average skewness across proteins, the correlation between protein mean and protein variance and the coefficient of variation across proteins. Scatterplots and boxplots of the data before transformation and normalization (Figure 2) and after transformation and normalization (Figure 3) show considerable improvement. A plot of protein means versus their standard deviations with lowess fit shows reasonably low correlation between the two (Figure 4). Thus, the normalized and log transformed data has satisfied distributional assumptions for hypothesis testing and has enabled direct comparison of protein profiles.

INSERT FIGURE 2 HERE

INSERT FIGURE 3 HERE

As a quality check, Spearman correlations were calculated between each pair of arrays. This, in combination with a boxplot of the negative controls (Figure 5), identified one outlier array, which was eliminated from further analysis. The remaining analyses are all carried out on this reduced normalized and transformed data matrix.

INSERT FIGURE 4 HERE

INSERT FIGURE 5 HERE

Proof of concept

The next step is to verify that the purported differences among the groups are indeed observable in the data. This is readily done via a spectral map. A spectral map, a variant of Gabriel's (1971) biplot, is a graph that displays markers for both proteins and biological samples, the markers being calculated from a weighted singular value decomposition of the data matrix $\{Y_{ij}\}$ as described by Lewi (1976) in a chemometrics setting and by Wouters et al (2003) for DNA microarrays. Figure 6 shows a spectral map of the protein array data, the circles are a sort of principal components display of the proteins, while the squares are a sort of principal components display of the samples. The separation of the diseased and normal samples is clearly evident. The normal samples are clustered around the center of the map, while the two sets of diseased samples appear at the opposite ends of the map. Thus this graph shows that the separation of these three groups is indeed the dominant signal in the data.

INSERT FIGURE 6 HERE

Identifying statistically significant proteins

For each protein, it is necessary to carry out a statistical test. The choice of test depends on the experimental design. In our case, since the objective is to examine any protein that has a significant pairwise difference between the three treatment groups (i.e., the control group and the two disease groups), we carried out a Tukey's studentized range test (Tukey, 1951, 1953) at α=0.01. This identified 277 proteins. The positive False Discovery Rate of this selection (Storey and Tibshirani, 2003) was estimated to be 19.6%.

Identifying significant protein combinations

A protein-by-protein analysis by definition precludes identifying groups of proteins that in combination may be more predictive than any individual protein. A number of multivariate

analysis approaches may be used to find such proteins. We used two methods that have a proven track record in the analysis of DNA microarray data: (1) random forest and (2) spectral map analysis.

(1) Application of random forest (Breiman 2001) has been found to give consistently good performance in classification and gene importance selection in the analysis of DNA microarray data (Lee et al 2005, Diaz-Uriarte and de Andres, 2006). In random forest classification, partitioning trees are built by successively splitting the samples according to a measure of impurity at a given node until terminal nodes are as homogenous as possible. The measure of impurity is usually determined by either entropy or the Gini index of diversity. The consequence of a small number of samples and a large number of expressions leads to the possibility of a non-unique solution due to many expressions leading to the same splits. Hence forests consisting of many trees, typically in excess of a thousand, are built. The 19 protein array samples were used in the supervised mode to build a classification model and the Gini index based importance measure was used for protein selection. Variable selection from random forests (Diaz-Uriarte and de Andres,2006) eliminated a large number of proteins to optimize the out-of-bag (OOB) error rates (Figure 7). In the end, 49 proteins were identified, 21 of which overlap with the proteins identified by Tukey Hypothesis testing.

(2) Spectral map analysis was already mentioned at the proof of concept stage described earlier. An additional advantage of a spectral map lies in its special ability to elicit correlations between the proteins and the separation of the biological samples. Thus the proteins located at the edges of the map and away from the center are noted as being the most highly associated with disease discrimination. In our study, 0.5% of most distal proteins were selected for further investigation. These proteins are shown in red in Figure 6. Only 10% of these were identified by Tukey Hypothesis testing.

The low degree of overlap is not particularly surprising as it is possible that a set of proteins may only be modestly differentially expressed across groups but may act in concert to separate them. Such sets of proteins would not be identified by an individual protein analysis but would be identified by a multi-protein analysis.

INSERT FIGURE 7 HERE

Validation of features identified

The collection of proteins identified by the above analyses were examined by the scientists for known biological relevance. Several of the proteins identified were known to be relevant for the diseases under investigation. A small subset was selected for further study. These proteins would be further examined by a second assay system. This would be best performed by an ELISA assay to determine levels of antibody to each of the specific antigens. An alternative approach would be analysis by western and immunohistochemistry. These protein analysis methods would provide valuable feedback as to the levels and distribution of the antigen proteins and would help us determine our interest in the associated specific antibody levels.

CONCLUSIONS

We have described a systematic approach for analyzing protein microarray data. The approach is consistent with the approach we use for analyzing data from routine DNA microarray experiments. One advantage of the consolidation is that software and data analysis systems developed for DNA microarray experiments can be gainfully employed for protein microarrays as well (see Prouty(2004) for a description of a microarray data analysis system that integrates the data analysis power of R with the data visualization power of Spotfire).

Our approach consists of a precise series of steps. In the preprocessing step, the use of normalization, data transformation and data quality assessment reduces disparities among arrays enabling subsequent application of analysis methods across all arrays simultaneously. In the

second step, spectral maps are used as an unsupervised classification tool to demonstrate the existence of a dominant signal that clearly separates the different groups. Any clusters of proteins associated with the separation are noted. In a supervised classification step, proteins associated with the separation of groups are identified. Finally, individual protein-by-protein analysis gives further insight into differentially expressed entities. Based on these findings, a small number of proteins can be identified for further study by specific secondary assays that are more accurate and precise.

Even though the individual statistical methodologies described here have been used to process and analyze data generated by DNA microarrays, they have not been reportedly applied to process proteomic data. These statistical methods are known to produce valuable clues in search for meaningful data features that could advance to novel targets and biomarkers. In fact, very little is known about a reliable statistical scheme to manage protein array data allowing the proteins that vary in response to the particular condition under study to be properly identified. While other methods based around non-consolidated approaches may leave protein array data unusable or misleading (as was our recent experience!), the approach we have demonstrated using DNA microarray analysis techniques applied to the analysis of protein microarray data, can yield much greater value. This multipronged approach also allows analysis of protein clusters rather than single proteins. This is advantageous in that panels of protein biomarkers have proved to be more useful in the diagnosis of disease states than single proteins, Xiao (2005). Abundant information is available about the individual proteins but this does not provide us with the integrated understanding of biological sytems. By studying many proteins simultaneously, as we make it possible by identifying clusters of proteins , we can study interactions to understand complex organisms better.

In general, the possibility of converging the data analysis tools for both, DNA and protein microarrays, offers an exciting prospect for biology and statistics research into future identification and development of disease biomarkers using protein microarrays. These methods provide added value to the analysis of the data in identifying important individual proteins and in the development of panels of biomarkers and provide enough information to continue to validate the data in other biological assays.

Software

All analyses were performed using the R software platform which can be freely downloaded from CRAN, the Comprehensive R Archive Network (http://cran.r-project.org/), with the use of the DNAMR library modules available at http://www.rci.rutgers.edu/~cabrera/DNAMR/. The spectral map software is available at http://alpha.luc.ac.be/~lucp1456/ while the random forest software is available as a CRAN library.

References

Amaratunga D. Cabrera J. 2004. *Exploration and Analysis of DNA Microarray and Protein Data*, John Wiley.

Bertone P. Snyder M. 2005. Advances in functional protein microarray technology, *FEBS J.* 272(21):5400-11.

Breiman L. 2001. Random forests. *Machine Learning*. 45:5–32.

D'Andrea MR. 2003. Evidence linking autoimmunity to neuronal cell death in Alzheimer's disease. *Brain Research* 982 (1): 19-30.

D'Andrea MR. 2005a. Evidence that the immunoglobulin-positive neurons in Alzheimer's disease are dying by the classical complement pathway. *Am J AD and Other Dementias* 20(3): 144-150.

D'Andrea MR. 2005b. Add Alzheimer's disease to the list of autoimmune diseases. *Medical Hypotheses* 64(3):458-463.

Diaz-Uriarte R. Alvarez de Andres S. 2006. Gene selection and classification of microarray data using random forest, *BMC Bioinformatics, **7**:3*

Gabriel KR. 1971. The biplot graphics display of matrices with application to principal component analysis. *Biometrika* 58:453-467.

Ilyin SE. Belkowski SM. Plata-Salaman CR. 2004. Biomarker discovery and validation: technologies and integrative approaches. *Trends Biotechnol.* 22(8):411-6.

Israeli E. Grotto I. Gilburd B. Balicer RD. Goldin E. Wiik A. Shoenfeld Y. (2005) Anti-Saccharomyces cerevisiae and antineutrophil cytoplasmic antibodies as predictors of inflammatory bowel disease, *Gut*. 54(9):1232-6.

Itoh M. 1989. Immunological aspects of diabetes mellitus: prospects for pharmacological modification, *Pharmacol Ther. 44(3):351-406.*

Kusnezow W. Hoheisel JD. 2002. Antibody microarrays: promises and problems. *Biotechniques.* Suppl:14-23.

Lee JW. Lee JB. Park M. Song SH. 2005. An extensive evaluation of recent classification tools applied to microarray data. *Computational Statistics and Data Analysis*. 48:869–885.

Lewi PJ. 1976. Spectral mapping, a technique for classifying biological activity profiles of chemical compounds, *Arzneimittel Forschung (Drug Research), 26, pp.1295-1300.*

Lokshin AE. Winans M. Landsittel D. Marrangoni AM. Velikokhatnaya L. Modugno F. Nolen BM. Gorelik E. 2006. Circulating IL-8 and anti-IL-8 autoantibody in patients with ovarian cancer, *Gynecol Oncol. 21.*

Lou XJ. Belkowski SM. Dixon JM. Hertzog B. Horwitz D. Ilyin SI. Lawrence D. Polkovitch D. Towers M. D'Andrea MR. 2006. Strategies of biomarker discovery for drug development. *Frontiers in Drug Design and Discovery*, eds. Caldwell, D'Andrea.

Nielen M. Schaardenburg D. Reesink H. Stadt R Bruinsma I. Koning M. Habibuw M. Vandenbroucke J. Dijkmans B. 2004. Specific Autoantibodies precede the symptoms of rheumatoid arthritis: A study of serial measurements in blood donors. *Arthritis and Rheumatism50 pp.380-386.*

Pietropaolo M. Yu S. Libman IM. Pietropaolo SL. Riley K. LaPorte RE. Drash AL. Mazumdar S. Trucco M. Becker DJ. 2005. Cytoplasmic islet cell antibodies remain valuable in defining risk of progression to type 1 diabetes in subjects with other islet autoantibodies, *Pediatr Diabetes, 6(4):184-92.*

Prouty S. Nathan D. Ledwith J. Salisbury M. Lyon G. Messer A. Amaratunga D. Go O. Wan J. Ilyin S. 2004. Integrative tools for data analysis in pharmaceutical R&D, *Pharmagenomics.*

Sundaresh S. Doolan DL. Hirst S. Mu X. Unal B. Davies DH. Felgner PL. Baldi P. 2006. Identification of humoral immune responses in protein microarrays using DNA microarray data analysis techniques, *Bioinformatics*, 22, pp.1760-1766.

Storey JD and Tibshirani R. 2003. Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, 100: 9440-9445.

Tukey JW. 1951. Reminder sheets for "Discussion of paper on multiple comparisons by Henry Scheffé." In The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948- 1983 469-475. Chapman and Hall, New York

Tukey JW. 1953. The problem of multiple comparisons. Unpublished manuscript. In The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948-1983 1-300. Chapman and Hall, New York.

Xiao Z. Prieto D. Conrads T. Veenstra T. Issaq H. 2005. Proteomic patterns: their potential for disease diagnosis, *Mol Cell Endocrinol*. 31 pp.95-106

Wouters L. Gohlmann HW. Bijnens L. Kass SU. Molenberghs G. Lewi PJ. 2003. Graphical Exploration of Gene Expression Data: A Comparative Study of Three Multivariate Methods , *Biometrics 59*, pp.1131-1139

Zhu H. Bilgin M. Bangham R. Hall D. Casamayor A. Bertone P. Lan N. Jansen R. Bidlingmaier S. Houfek T. Mitchell T. Miller P. Dean RA. Gerstien M. Snyder M. 2001. Global Analysis of Protein activities Using Proteome Chips, *Science (293).*

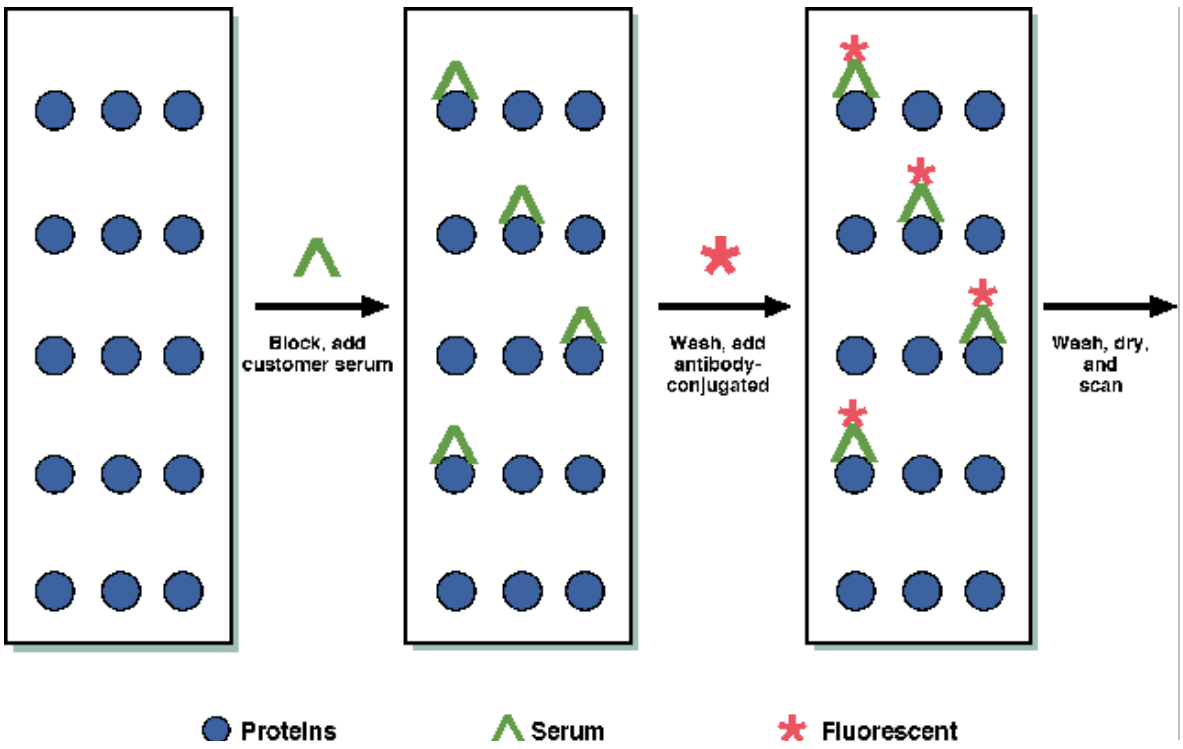


Figure 1. Experimental protocol involves application of serum sample to the surface of protein microarray with fixed proteins. In the next step fluorescence is added and after washing and drying, scanning takes place.

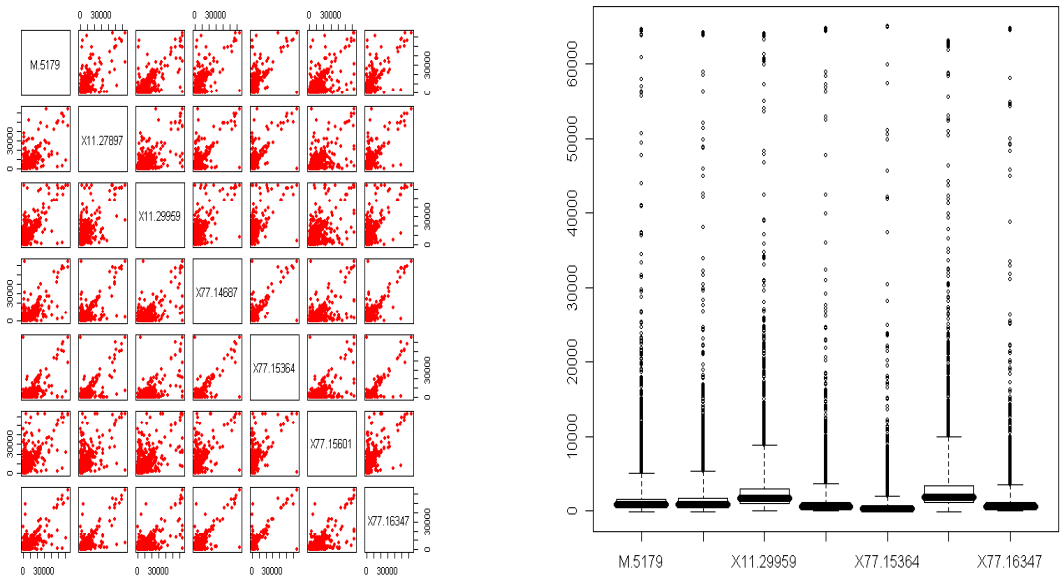


Figure 2. Pairs and Boxplot of raw data for a subset of samples

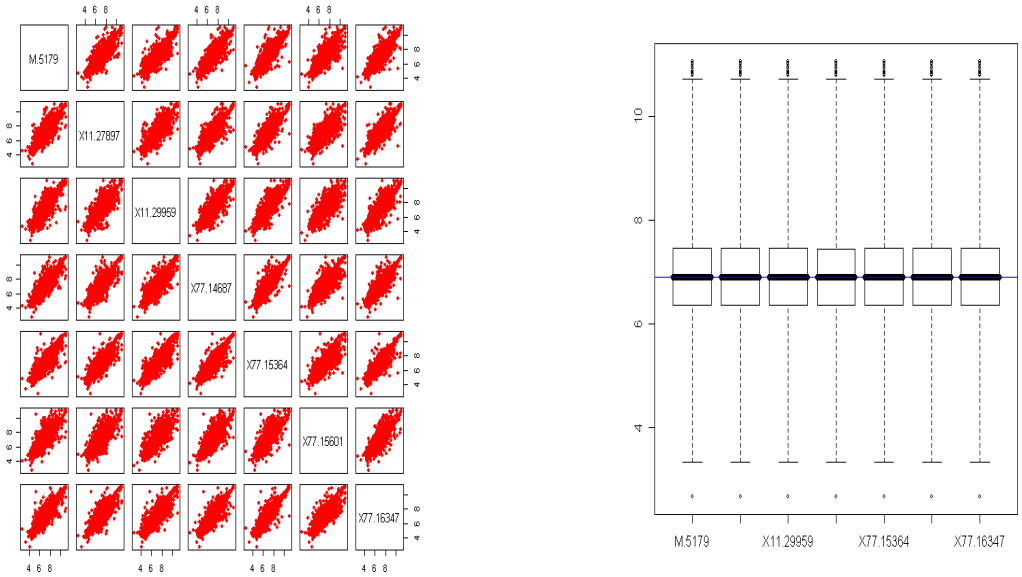


Figure 3. Pairs and Boxplot of normalized and started log transformed data

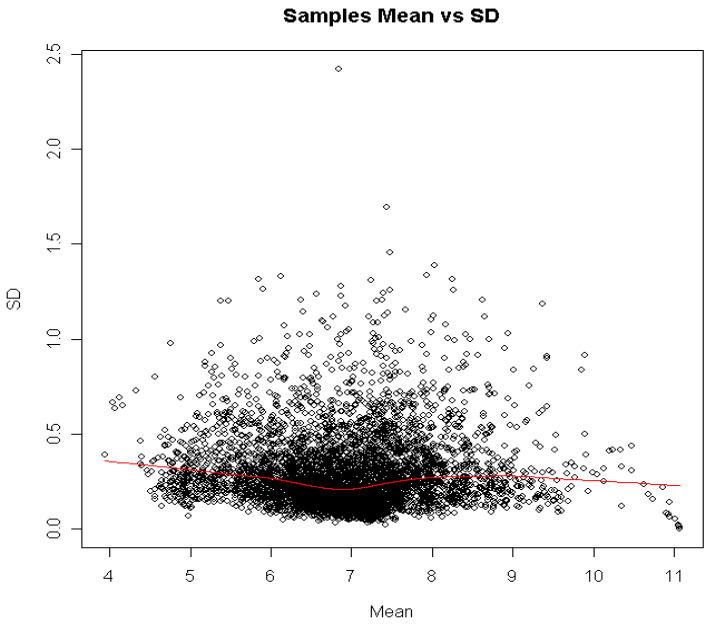


Figure 4. Standard Deviation versus Mean of normalized and log transformed samples. Lowess smoothing line shown in red, remains largely horizontal indicating low correlation.

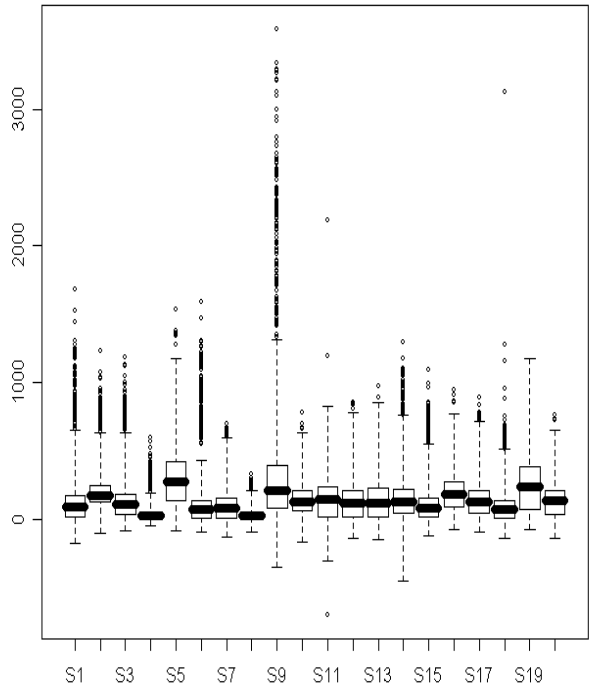


Figure 5. Negative control raw values of 20 chips in the experiment. The outlier chip S9 is clearly visible.

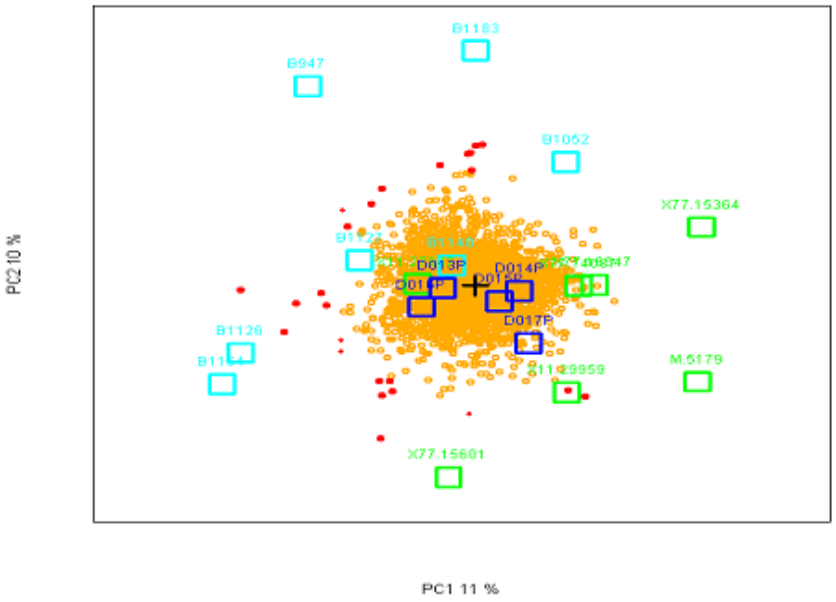
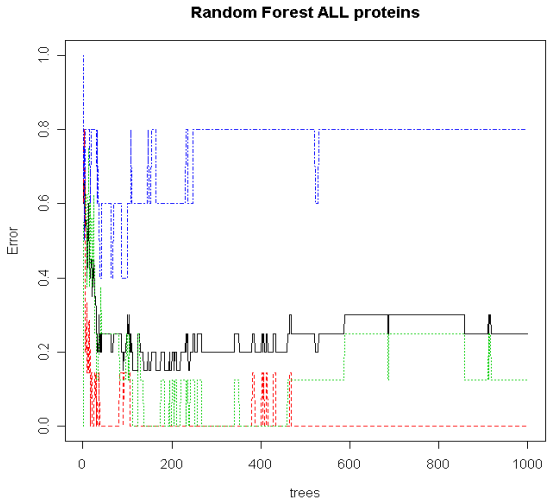
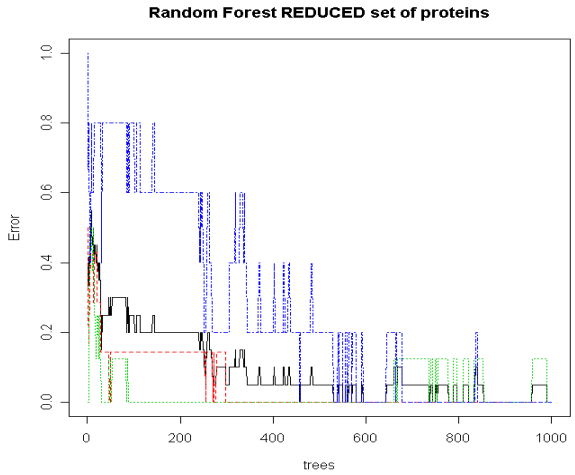


Figure 6. Spectral Map showing biological samples as squares and proteins as circles. Normal(blue), Disease1(green), Disease2(magenta), Proteins(orange). Proteins shown in red correspond to the most distal proteins.



a)



b)

Figure.7 Protein selection to minimize OOB error rates. The procedure calculates OOB error rates for consequtively shrinking set of proteins. The set of proteins corresponding to the lowest OOB is the optimum set. a) Error rates with total number of proteins. b) Error rates with reduced set of proteins.