Mining Whole-Sample Mass-Spectrometry Proteomics Data for Biomarkers - An Overview

Abstract

In this paper we aim to provide a concise overview of designing and conducting an MS proteomics experiment in such a way as to allow statistical analysis that may lead to the discovery of novel biomarkers. We provide a summary of the various stages that make up such an experiment, highlighting the need for experimental goals to be decided upon in advance. We discuss issues in experimental design at the sample collection stage, and good practise for standardising protocols within the proteomics laboratory. We then describe approaches to the data mining stage of the experiment, including the processing steps that transform a raw mass spectrum into a useable form. We propose a permutation-based procedure for determining the significance of reported error rates. Finally, because of its general advantages in speed and cost, we suggest that MS proteomics may be a good candidate for an early primary screening approach to disease diagnosis, identifying areas of risk and making referrals for more specific tests without necessarily making a diagnosis in its own right. Our discussion is illustrated with examples drawn from experiments on bovine blood serum conducted in the Centre for Proteomic Research (CPR) at Southampton University.

Key words:
mass spectrometry, proteomics, data mining, biomarkers

1 Introduction

Whole-sample proteomics is a rapidly developing technology that holds out the tantalising prospect of more reliable, cheaper and faster diagnostic tests across virtually the whole spectrum of diseases and disorders, as well as an opportunity to observe previously unknown immune mechanisms and biological processes in action. The method comprises the use of high-throughput mass spectrometry to capture a 'snapshot' of the proteomic makeup of a biological sample, followed by the application of statistical data mining and machine learning techniques to identify biomarkers. These are typically proteins, peptides or metabolites whose measured intensity varies in response to some known biological change, and which may be used singly or in combination for the purposes of classification. In many ways this problem resembles others that are well known the field of data mining - most notably genomics, but also pattern recognition and other forms of supervised learning.

In this article we aim to provide what we hope to be a timely overview and practical discussion of the experimental design and data mining aspects of such studies, based on our own experience in analysing data generated in the Centre for Proteomic Research at the University of Southampton. It is our intention to identify the strategies that we have found most useful and the pitfalls we have encountered in mining such data in the hope that this may lead to improved understanding and further development and standardisation of the experimental methodology and protocols employed in such studies.

In Section 2 of this paper we attempt to give an overview of a proteomics experiment for biomarker discovery, highlighting the various stages through from initial sampling through to final marker identification. In Sections 3 and 4 we highlight issues in experimental design and experimental protocol: Section 3 covers sampling and design issues at the sample extraction stage, while Section 4 explores lab design – replicates and randomisation. In Section 5 we outline the processing and data reduction steps needed to transform a raw mass spectrum into a form suitable for data mining. Section 6 covers unsupervised and supervised classification – feature selection and classification techniques. We conclude in Section 7 with a summary of our suggestions and some thoughts on the future potential of the technique.

2 Overview

The flow chart in Figure 1 illustrates a hypothetical layout for an experiment designed to uncover and validate genuine biomarkers. As with the design of any study, it is generally good practice to attempt to define a precise statement of goals in advance. Although on the face of it this may seem entirely self-evident, the collaborative nature of proteomics research often means that the biological samples analysed are a by-product of other kinds of studies and experiments, conducted under their own designs with their own goals, and in the rush to generate mass spectra it may not be until after the data capture step that proper consideration is given to the purpose of the proteomic study itself.

In studies of illness or disease, it is tempting to view the goal of an MS proteomics study as leading directly to a stand-alone diagnostic test, envisaging a situation in which samples would be collected and frozen under standardised protocols and sent to central laboratories where a full spectrum would be captured and an exact diagnosis would be made on the basis of the observed intensities of a small number of components. In this scenario, a good estimate of the generalisation performance of a classifier (ie. it's performance when presented with new, unseen data) will be the best criterion for measuring the success of any study. While such a technology seems entirely achievable, it will involve a more extensive study design exploring challenges in standardising the sample extraction, storage and preparation steps, inter-machine and inter-lab reproducibility and population sample bias. Such issues have only recently begun to be addressed in a serious fashion (see eg. [Hu et al., 2005], [Karp et al., 2005], [Wagner et al., 2004], [Vitzthum et al., 2005], [West-Nielsen et al., 2005]). It also seems in some sense wasteful to capture a complete proteomic spectrum containing potentially thousands of components when only a handful of known components will ultimately be used for diagnosis. In practise it would seem more logical to search for novel markers first, validate and identify them, and then begin to tailor the sample preparation, data capture and data processing stages to amplify the signal and to minimise the measurement error of these specific components to the greatest degree possible once they have been identified.

If our goal is simply to identify genuine novel markers for future analysis, a necessary intermediate step our emphasis shifts somewhat away from classification performance within a single data set and towards validation across multiple data sources. It is still unusual to find examples in the literature of putative markers validated under genuinely differing conditions - say, in a different lab, on a different machine, or on samples originating from a distinct source (though not entirely unheard of, see eg. [Li et al., 2005]). This is denoted in Figure 1 by the validation cycle - the more times we observe a change in a particular marker under changing conditions, the more confident we become that we are observing a genuine biological effect.

[Figure 1 about here]

Another consideration prevalent in such studies is their collaborative nature. Very often the samples used will have been collected and stored even before the proteomics study is conceived. This is represented by the double branch at the beginning of our flow chart. Although this may mean that the experimental design of the sampling experiment may be fixed in advance, in practise we have often found it possible to request further samples of a particular type on an ongoing basis - the validation cycle therefore becomes something analogous to a positive feedback loop.

3 Initial Sampling

A proteomics experiment for biomarker discovery begins with a set of clinical samples, though as we have already noted, collection of these samples may occur before the proteomics study itself is initiated. These samples often take the form of blood serum, but biological products such as sputum or urine are equally valid candidates for mass spectrometry. Each sample will be associated with at least one and possibly more class status labels, reflecting the biological difference or differences of interest. Frequently this will be an indicator of the presence or absence of a disease or infection (eg. much of the literature to date has focused on different types of cancer, see [Li et al., 2005] or [Conrads et al., 2004]). In principle, however, the difference might also be hereditary, dietary, or deliberately induced (an example of an application based on the latter might be a 'doping' test to identify athletes using artificial supplements).

There are several important considerations to note here, dependant upon the stated goals of our study. The first is the observation that even if we are able to identify clear, validated markers in our data, we cannot make any accurate statement of their true diagnostic potential unless we have at least one sample data set that reflects the population to which we hope to apply the test (note that this is not necessarily the same thing as a sample of the overall population, since we may envisage applying the test on a selective basis, identifying at-risk groups a-priori, for example, or in some cases the subjects may be self-selecting). Often biological samples employed in proteomic studies are collected as a by-product of randomised controlled clinical trials. Here the 'positive' samples may be drawn from control groups, or treatment groups prior to intervention, and the 'negative' samples to which they are compared are typically a similar sample of matched healthy volunteers.

Samples of this type are useful in the sense that they give us perhaps the strongest chance of finding genuine biomarkers clearly expressed. They may, however, be inadequate for determining the diagnostic potential of these markers. Such samples may be in some sense more 'pristine' than the reality, selected for inclusion in the study on the basis of the ease with which they may be assigned to clearly segregated groups. Another consideration that is invariably overlooked is that except in cases where the biological difference has been artificially induced (usually ruled out in humans in studies of disease on ethical grounds, but sometimes found in veterinary studies) patients must have been assigned to either of the two groups on the basis of an expert diagnosis or existing diagnostic test. It is important to recognise that when assessing the potential accuracy of our markers that in these cases we will always record an apparent error rate at least as high as that of the diagnostic method used to establish the disease status class labels of our samples.

This is also an important consideration when building a classification model, since it is well known that mislabelled training set observations can have a disproportionately detrimental impact on many model-building techniques.

We must also give careful consideration to the consistency of the protocols employed for collecting and storing samples. Since changes in the expressed intensity of biomarkers are usually determined on the basis of a two-way comparison between groups, and biological samples prove highly susceptible to small changes in environmental conditions (see eg. [West-Nielsen et al., 2005]), it is vitally important that these protocols remain constant across groups. For example, it is desirable that from the earliest possible stage samples from different groups should not be stored in separate freezers, or transported separately, and that tubes should not be contaminated. In reality it can be difficult to foresee all factors that might arise in an experiment in advance, and protocols frequently do change midway through an experiment for wholly practical reasons. In this case, a secondary guiding principle should be that no change should affect any one group more than any other.

Care should also be taken when combining samples from multiple sources. In one pilot-scale analysis, we analysed samples with two states (healthy and diseased) where samples were drawn in unequal within-group proportion from two identical clinical studies conducted in two geographically distinct locations. After capturing the data and identifying one apparently strong marker, we discovered that it was better correlated with the geographical source of the sample than with the disease status. This may have reflected some difference in plastic composition of the sample containers used at the two sites. In this instance, having excluded this marker, we were able to identify further markers for which the correlation was stronger with the disease state.

4 Experimental Design in the Lab

Once samples have been collected and transported to a proteomics lab, we are ready to consider our experimental design for capturing the mass spectra. Once again the guiding principle here must be consistency, primarily across class groups and secondarily through the course of the experiment. Samples must undergo a number of preparatory steps before being introduced into a mass spectrometer. In our studies, we typically analyse 20 samples per day alongside a quality control (one identical sample run on every day to assess reproducibility). Prior to the start of the experiment, all samples are thawed and randomised into separate well-plates before being refrozen. On the day that each plate is to be analysed, the samples are thawed again, and either bonded to a chip surface using a matrix compound in the case of the matrix assisted laser desorption/ionisation (MALDI) technique or introduced to an electrospray-ionisation (ESI) mass spectrometer through a Nanomate device. In addition to the quality control sample, we also frequently run a complete replicate of the entire first day at the end of the experiment for comparison.

Samples are known to degrade over time once thawed to room temperature (eg. blood samples clot, see [West-Nielsen et al., 2005]). Ideally the number of freeze-thaw cycles should be kept constant across the experiment, as should the length of time that samples are allowed to stand at room temperature once thawed.

Without doubt, the best means of protecting against unforeseen experimental drift is randomisation. The heatmap in panel A of Figure 2 illustrates a set of 233 MALDI mass spectra for bovine serum samples acquired as part of a study seeking biomarkers for bovine tuberculosis. Twenty mass spectra were acquired per day for the first eleven days, and thirteen on day 12. Only a section of the scanning range is shown, running from 750 - 2,500Da. Each of the 233 spectra (processed via the steps described in the following section) is represented as a horizontal line across the plot, running from top to bottom. The breaks across the plot represent the separation between separate days of the experiment. The samples were drawn from both healthy and diseased animals, but the run order was entirely randomised and disease state is not indicated here.

[**Figure 2 about here**]

Here we clearly expect to see variation between samples, since they originate from different animals. However one of the most striking aspects of the MALDI plot is the apparent variation from day to day in the course of the experiment, despite the efforts that were made to maintain a consistent protocol and calibration. This was also borne out by an analysis of the twelve quality control spectra This kind of daily drift has been observed before in [Conrads et al., 2004]. Evidently we risk introducing bias into our study if we run our samples in a non-random order (by group), which may in turn lead to spurious results. The best means to guard against this possibility is to fully randomise the run order of samples.

Panel B in the figure above shows the same samples with the same run order, this time analysed using an ESI mass spectrometer fitted with a Nanomate device, plotted in a range from 50-1200 m/z. This more automated approach yields data with a higher density at a higher resolution. It is, however, more likely than the MALDI approach to display heavy components as multiply-charged envelopes rather than single peaks, which means that some parts of the spectra must be subjected to a process of deconvolution to yield actual mass values. A peak-by-peak analysis of the quality control samples, however, suggested that this approach yielded significantly better between-day reproducibility than the MALDI, and deconvolution suggested that the number of actual components observed in a comparable range was also greater.

5 Data Processing

Before mass spectra may be analysed, they typically undergo a number of processing steps applied sequentially to render them in a standardised form. The order in which these steps occur is not mandated (for example, smoothing might occur before or after binning, but conversely it seems logical that normalisation should occur after baseline subtraction). Each of these steps presents us with a wealth of possible choices as to how to proceed, with the impact and suitability of each competing technique likely to be largely data-dependent.

Combination: Constructing a spectrum by aggregating raw data captured over several laser shots or over a continuous period of time is usually performed by the proprietary software that runs the mass spectrometer.

Calibration Correction: Small inadvertent errors in calibration typically arise when the mass spectrometer is regularly recalibrated in the course of an experiment (perhaps daily). This will not affect the shape of the spectra themselves, but may result in a constant shift along the mass / charge scale. This shift can be corrected by aligning a peak, either a reference component of known mass artificially introduced to all samples for this purpose, or any other peak found to appear consistently in a majority of samples.

Binning: If bin widths are chosen to be wider than the spacing of the original spectra, binning may have the effect of smoothing the data and reducing its size. It also ensures that the values recorded on the m/z scale are identical across all spectra. In our studies, we generally aim to preserve as much of the original data as possible, so we set the bin edges as equal to the midpoints between recorded values in a single reference spectrum. Because the scanning accuracy generally reduces with increasing mass, and this is reflected in the actual m/z values of the intensities that are recorded, our bin sizes increase with increasing m/z.

Smoothing: Any spectrum smoothing algorithm or technique may be used to smooth the data and remove the 'jaggedness' typical of mass spectrometry (noise). Usually these are parametric, and require some tuning according to the degree of smoothing desired. Among the many possible approaches are kernel smoothing, spline smoothing and smoothing by wavelet thresholding. Care must be taken at this stage not to distort or 'blunt' peaks of interest, and not to eliminate genuine peaks.

Baseline subtraction: A baseline shift occurs in a spectrum as the result of residual charge buildup in the ion channels during spectrum acquisition, and therefore tends to appear below components present at high concentration, raising them up above the m/z axis. Since the magnitude of this shift tends to vary from spectrum to spectrum, it is standard practise to subtract it in such a way that the bases of all peaks sit along the horizontal axis. In our studies we opt to model the baseline as a series of cubic splines passing through the lowest minima within windows a fixed number of bin centres apart in a manner similar to the *msbackadj* function from the Matlab bioinformatics toolbox. Another approach (found for example in the MassLynx software suite), is to model the baseline as a high order-polynomial. In our experience, baselines tend to be less pronounced and less variable in ESI data than in MALDI data.

Normalisation: Normalising the data is sometimes deemed necessary after baseline correction, depending on the degree of visible variation, and usually amounts to standardising the area under each spectrum by multiplying the intensity scale by an appropriate constant. One particular limitation of this approach, as discussed in ([Arneberg et al., 2007]), is its vulnerability to heteroscedastic noise in peak heights (ie. the tendency for measurement error to proportionately increase as peak intensity increases). They propose applying an n^{th} root or logarithmic transformation (see [Kvalheim et al., 1994]) to the full spectrum prior to normalisation, though this will adversely impact the correlation structure across spectra. Under some circumstances, we might consider normalising the height of a specific peak or a set of peaks. This may be most appropriate where the sample has been 'spiked' with one or more reference components in known proportion. It is also possible that normalisation could be applied after a peak extraction step (see below) to standardise the resulting peak list.

Alignment Unlike calibration correction, alignment typically involves the manipulation of the spectra themselves, usually with reference to a given set of reference peaks, to ensure that components that are thought to be the same appear consistently in the same place on the m/z scale. This is to correct for the measurement error of the mass spectrometer itself, which typically increases with increasing m/z. Many techniques have been proposed to accomplish this, with many attempting to preserve peak shape (cf. the function msalign in the Matlab Bioinformatics toolbox, which applies a linear transformation to the m/z scale). Other techniques have been proposed based on optimising the cross-correlation structure across multiple spectra (eg. [Wong et al., 2005]). The applicability of such manipulation will depend in large part on the number of peaks present, their density along the spectrum, and in particular their degree of overlap. The function of matching peaks across multiple spectra might also be addressed at the peak extraction stage.

Peak extraction: It is relatively common practise to search for biomarkers in the full data, with each peak represented by multiple adjacent m/z values. Often the data will have undergone some further transformation (such as the noise correction described above). Such approaches will, however, potentially be vulnerable to small misalignments that have not been corrected at the alignment stage. As we are generally interested in classifying based on the heights of peaks in the sample (which correspond to the intensities of specific components), we may alternatively seek to identify these, match them across multiple spectra, and extract a list of their heights. At this point the data size can be greatly reduced, since, for example, a raw ESI spectrum of approximately 200,000 recorded values may yield a peak list of typically 3,000-4,000 components, though some of these may exist in multiple charge states. Usually the peak matching and extraction steps are based on ad-hoc algorithms with parameters fine-tuned to suit the data under consideration, and will attempt to match a peak by searching for it within a defined window in all spectra.

Deconvolution: Because of the manner in which mass spectra are calculated, components that yield multiply charged ions may be recorded as a 'charge envelope' – a range of peaks appearing at whole fractions of their true mass value (singly charged ions should always appear at integer mass / charge values). Deconvolution methods such as the Maxent algorithm employed by the MassLynx software attempt to identify peaks with fractional charges and recombine them at their true integer masses. Multiply charged ions are comparatively rare when using laser-ionisation mass spectrometers, but are a common feature of electrospray ionisation. Deconvolution algorithms rely on their ability to isolate and recombine individual peaks given accurate measurements of their mass intensity. It is therefore possible that this step may inadvertently introduce noise into the deconvoluted spectrum in an unpredictable fashion if the original spectrum is itself noisy, and as a result it is perhaps likely to prove most useful in retrospectively identifying the true mass of components that emerge as candidate markers.

[**Figure 3 about here**]

6 Data Mining

Once we have acquired fully-processed peak lists for a set of positive and negative samples, how can we determine whether biomarkers are, in fact, present? A logical approach is to divide our samples into training and test sets (possibly with an additional validation set used to test choices of parameter values), and attempt to build a model on the training set that is able to predict observations in the test set with better accuracy than would be expected to be achieved by classifying at random. More generally, we can make use of techniques such as k-fold cross-validation and leave-one-out cross-validation to maximise our training set size and obtain a better unbiased estimate of our test performance. A useful metric here is the mean of the false positive rate (the proportion of negative samples mistakenly classified as positive) and the false negative rate (the proportion of positive samples mistakenly classified as negative). This quantity will have an expected value of exactly 50% in the absence of any true markers, and is related to the mean subjective utility (MSU) score [McDonald, 2006] under the assumption of equal misclassification costs. Typically, although we select training and test samples at random, we may wish to introduce the constraint that the prior proportions of positives and negative samples remain fixed.

Classifying our data will typically consist of a combination of feature selection (to reduce the data size and guard against overfitting) and a model building step. Both steps will be performed on the training data only, or possibly a combination of a training and validation sample (as distinct from the test sample). We shall consider these steps in turn.

Feature Selection: It is important to recognise that all feature selection techniques will introduce at least one tunable parameter, whether this denotes the number of features to be selected or some other threshold value. Determining the optimal value of this parameter, or justifying an ad-hoc choice of its value will generally amount to validation, either on a sample selected from the original data or on data generated separately as part of the validation cycle. Almost certainly the fastest methods of performing selection are based on feature ranking, computing a simple score for the discriminative power of each peak and using these to select the most likely candidate markers. Possible scoring methods include, among others, the p-values from the two sample t-test, the Mahalonobis distance, the Mann-Whitney-Wilcoxon ranked sum statistic, the Pearson correlation coefficient, and the performance of a stump classifier (a decision tree with only one split on one variable that optimises a criterion such as MSU).

Although feature ranking techniques are fast they are blind to interactions between components because they consider only one peak at a time. Other more computationally expensive techniques such as forward or backwards stepwise variable selection, or selection based on Artificial Neural Networks have been proposed (see eg. [Lancashire et al., 2005]), and these may capture such interactions.

Model building: Once a number of the most significant features have been selected, we are in a position to build a model. There are a great many standard supervised classification methodologies open to us here, including linear and logistic regression, artificial neural networks, decision trees, support vector machines (SVM). Some of these will introduce further tunable parameters - for instance the choice of kernel and slackness parameter in SVM, number of nodes in the intermediate layer of a neural network, or splitting criterion and node size for a decision tree. In general, the relative performance of these classifiers is determined by the data under consideration. Some meta-analysis of classification methods (such as [Jamain, 2004]) have found nonparametric linear classifiers to be among the most generally robust, and this appears to be borne out by our experience in this area. Cost-sensitivity, or the notion that in many 2-class classification scenarios false positive and false negative errors are often not considered equally serious, is also likely to be an important consideration at this stage, since many model-building methodologies (such as decision trees) are able to incorporate unequal cost weightings at the fitting stage.

Once we have an estimate of our classifier's performance, perhaps in the form of a mean of the false positive and false negative test error rates, we need to determine whether this performance is significant enough to imply the presence of biomarkers. Here we have found it useful to employ the following permutation method:

Step 1: Perform feature selection and classification to generate an estimate of the generalisation performance of our classifier, recorded as the mean of the false positive and false negative rates.

Step 2: Randomly permute the labeling of positive and negative observations in the training and test sets and repeat Step 1 N times with all other settings fixed, where N is some large integer. This yields a series of test errors $e_1, e_2, \ldots, e_N$ The mean of these errors should be close to 50%.

Step 3: Count the number of errors that exceed the error rate estimate calculated in Step 1. This value as a proportion of N gives us a rough estimate of the probability that our observed test error could have arisen by chance.

If our estimate of the generalisation performance of our classifier is very much smaller than could be expected to arise by chance, then we are in a position to say that biomarkers are likely to be among the components on our peak list. Potential markers can be identified by analysing the feature selection and classification steps to determine which peaks are selected most often and with highest significance, and which tend to be given greatest weight by our classification model.

As we have already noted, unless our data sample represents a genuine sample of the population to which a diagnostic test would be applied in practice, we cannot make any statement here about the true diagnostic potential of our markers. An iterative validation process is likely to be necessary in order to determine the validity of particular markers, and to determine how reproducible their discriminatory power can be shown to be from the initial sampling phase through the sample storage, preparation and mass spectrometry phases, as well as inter-machine and inter-lab. It also seems likely that in developing an actual diagnostic test, care would be taken both to optimise all these phases specifically to amplify the components that are known to be useful. For example, the ELISA interferon-gamma test for tuberculosis involves stimulating a blood sample with a non-virulent derivative of the bacterium to stimulate the production of interferon-gamma, a known marker.

7 Summary and Discussion

In this article we have attempted to draw attention to be some of the most important considerations in designing and performing an experiment to discover novel biomarkers in proteomics data, and we have also summarised the key processing stages that make up such an experiment. Evidently such research is usually broadly collaborative, involving requiring expertise in the design of clinical experiments, biological sciences, and statistical data mining / machine learning. We have found it helpful to draw a distinction between MS proteomics for the purposes of discovering novel biomarkers, and mining with a view to developing a novel diagnostic procedure, since the latter will require more advanced studies of reproducibility, standardisation and optimisation of protocols.

It is evident, however, that MS proteomics would have many advantages as a diagnostic tool - collection of samples is quick and non-invasive, throughput is high and results can be obtained relatively quickly. Furthermore, the method yields a complete snapshot of the proteome of a sample. For this reason, it is possible that the technique could be a useful tool in primary screening, able to identify risk factors for a wide range of potential conditions and referring patients for further testing, even where it is not possible to make an exact diagnosis. This could be analogous to the role of a general practitioner in a health service, who sits at the top of a hierarchy of successively more specialised (and also potentially more expensive or time consuming) diagnostic techniques. This idea, along with one possible means of constructing such a hierarchy, is discussed in greater detail in [McDonald, 2004].

References

[Aebersold and Mann, 2003] Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422:198 – 207.

[Arneberg et al., 2007] Arneberg, R., Rajalahti, T., Flikka, K., Berven, F. S., Kroksveen, A. C., Berie, M., Myhr, K.-M., Vedeler, C. A., Ulvik, R. J., and

Kvalheim, O. M. (2007). Pretreatment of mass spectral profiles: Application to proteomic data. *Analytical Chemistry*, 79:7014–7026.

[Baker, 2005] Baker, M. (2005). In biomarkers we trust? *Nature Biotechnology*, 23:297 – 304.

[Conrads et al., 2004] Conrads, T. P., Fusaro, V. A., Ross, S., Johann, D., Rajapakse, V., Hitt, B. A., Steinberg, S. M., Kohn, E. C., Fishman, D. A., Whitely, G., Barrett, J. C., Liotta, L. A., Petricoin, E. F., and Veenstra, T. D. (2004). High-resolution serum proteomic features for ovarian cancer detection. *Endocrine-Related Cancer*, 11:163 – 178.

[Coombes et al., 2005] Coombes, K. R., Morris, J. S., Hu, J., Edmonson, S. R., and Baggerly, K. (2005). Serum proteomics profiling - a young technology begins to mature. *Nature Biotechnology*, 23:291 – 292.

[Coombes et al., 2004] Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., Hung, M. C., and Kuerer, H. M. (2004). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Technical Report UTMDABTR-001-04, The University of Texas M. D. Anderson Cancer Center.*

[Hu et al., 2005] Hu, J., Coombes, K. R., Morris, J. S., and Baggerly, K. A. (2005). The importance of experimental design in protemic mass spectrometry experiments: Some cautionary tales. *Briefings in Functional Genomics and Proteomics*, 3:322 – 331.

[Jacobs et al., 2005] Jacobs, J. M., Adkins, J. N., Qian, W.-J., Liu, T., Shen, Y., Camp, D. G., and Smith, R. D. (2005). Utilizing human blood plasma for proteomic biomarker discovery. *Journal of Proteome Research*, 4:1073 – 1085.

[Jamain, 2004] Jamain, A. (2004). *Meta–Analysis of Classification Methods.* PhD thesis, Imperial College London, UK.

[Karp et al., 2005] Karp, N. A., Spencer, M., Lindsay, H., O'Dell, K., and Lilley, K. S. (2005). Impact of replicate types on proteomic expression analysis. *Journal of Proteome Research*, 4:1867 – 1871.

[Kvalheim et al., 1994] Kvalheim, O. M., Brakstad, F., and Liang, Y.-Z. (1994). Preprocessing of analytical profiles in the presence of homoscedastic or heroscrdastic noise. *Analytical Chemistry*, 66:43 – 51.

[LaBaer, 2005] LaBaer, J. (2005). So, you want to look for biomarkers (introduction to the special biomarkers issue). *Journal of Proteomic Research*, 4:1053 – 1059.

[Lancashire et al., 2005] Lancashire, L., Schmid, O., Shah, H., and Ball, G. (2005). Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis, and principal components analysis. *Bioinformatics*, 21:2191 – 2199.

[Li et al., 2005] Li, J., Orlandi, R., White, C. N., Rosenzweig, J., Zhao, J., Seregni, E., Morelli, D., Yu, Y., Meng, X.-Y., Zhang, Z., Davidson, N. E., Fung, E., and Chan, D. W. (2005). Independent validation of candidate breast cancer serum biomarkers identified by mass spectrometry. *Clinical Chemistry*, 51:2229 – 2235.

[Malyarenko et al., 2005] Malyarenko, D. I., Cooke, W. E., Adam, B.-L., Malik, G., Chen, H., Tracy, E., Trosset, M. W., Sasinowski, M., Semmes, O. J., and Manos, D. M. (2005). Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques. *Clinical Chemistry*, 51:65 – 74.

[McDonald, 2004] McDonald, R. A. (2004). *Combination in Supervised Classification Problems*. PhD thesis, Imperial College London, UK.

[McDonald, 2006] McDonald, R. A. (2006). The mean subjective utility score, a novel metric for cost-sensitive classifier evaluation. *Pattern Recognition Letters*, 27:1472 – 1477.

[Petricoin et al., 2002] Petricoin, E. F., Ardekani, A. M., Hitt, B., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359:572 – 577.

[Rosenblatt et al., 2004] Rosenblatt, K., Bryant-Greenwood, P., Killian, J. K., Mehta, A., Geho, D., Espina, V., Petricoin, E. F., and Liotta, L. A. (2004). Serum proteomics in cancer diagnosis and management. *Annual Review of Medicine*, 55:97 – 112.

[Sauve and Speed, 2004] Sauve, A. C. and Speed, T. P. (2004). Normalization, baseline correction and alignment of high-throughput mass spectrometry data. In *Proceedings of the Genomic Signal Processing and Statistics Workshop*.

[Tatay et al., 2003] Tatay, J. W., Feng, X., Sobczak, N., Jiang, H., Chen, C.-F., Roumyana, K., Struble, C., Wang, N. J., and Tonellato, P. J. (2003). Multiple approaches to data-mining of proteomic data based on statistical and pattern classification methods. *Proteomics*, 3:1704 – 1709.

[Villanueva et al., 2005] Villanueva, J., Philip, J., Chaparro, C. A., Li, Y., Toledo-Crow, R., DeNoyer, L., Fleisher, M., Robbins, R. J., and Tempst, P. (2005). Correcting common errors in identifying cancer-specific serum peptide signatures. *Journal of Proteome Research*, 4:1060 – 1072.

[Vitzthum et al., 2005] Vitzthum, F., Behrens, F., Anderson, N. L., and Shaw, J. H. (2005). Proteomics: From basic research to diagnostic application. a review of requirements and needs. *Journal of Proteome Research*, 5:1086 – 1097.

[Vivanco et al., 2005] Vivanco, F., Martin-Ventura, J. L., Duran, M. C., Barderas, M. G., Blanco-Colio, L., Darde, V. M., Mas, S., Meilhac, O., Michel, J. B., Tunon, J., and Egido, J. (2005). Quest for novel cardiovascular biomarkers by proteomic analysis. *Journal of Proteome Research*, 4:1181 – 1191.

[Wagner et al., 2004] Wagner, M., Naik, D., and Pothen, A. (2004). Protocols for disease classification from mass spectrometry data. *Proteomics*, 3:1692 – 1698.

[West-Nielsen et al., 2005] West-Nielsen, M., Hogdall, E. V., Marchiori, E., Hogdall, C. K., Schou, C., and Heegaard, N. H. (2005). Sample handling for mass spectrometric proteomic investigations of human sera. *Analytical Chemistry*, 77:5114 – 5123.

[Wong et al., 2005] Wong, J. W., Durante, C., and Cartwright, H. M. (2005). Application of fast fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Analytical Chemistry*, 77:5655 – 5661.