The importance of experimental design in proteomic mass spectrometry experiments: Some cautionary tales

Jianhua Hu, Kevin R. Coombes, Jeffrey S. Morris and Keith A. Baggerly

Date received (in revised form): 13th December 2004

Jianhua Hu
is Assistant Professor of Biostatistics at the University of Texas M. D. Anderson Cancer Center (UTMDACC). Her current research is focused on the development of biomarker panels from mass spectrometry data.

Kevin Coombes
is Associate Professor of Biostatistics and chief of the Bioinformatics section (including Hu and Baggerly) at UTMDACC. He has won multiple prizes for the analysis of microarray and proteomics data and has developed several methods for processing spectral data.

Jeffrey Morris
is Assistant Professor of Biostatistics at UTMDACC. His work on functional data analysis (ehixh includes proteomic specra) has been recognized with several prizes from the statistical community.

Keith Baggerly
is Associate Professor of Biostatistics at UTMDACC. He has won multiple prizes for the analysis of microarray and proteomics data and speaks frequently on experimental design.

Keywords: *calibration, data preprocessing, follow-up studies, proteomic mass spectrometry, randomised run order, validation*

Keith Baggerly,
Department of Biostatistics and Applied Mathematics,
University of Texas M. D. Anderson Cancer Center,
1515 Holcombe Blvd, Unit 447,
Houston, TX 77030-4009,
USA

Tel: +1 713 563 4290
Fax: +1 713 563 4243
E-mail: kabagg@mdanderson.org

Abstract

Proteomic expression patterns derived from mass spectrometry have been put forward as potential biomarkers for the early diagnosis of cancer and other diseases. This approach has generated much excitement and has led to a large number of new experiments and vast amounts of new data. The data, derived at great expense, can have very little value if careful attention is not paid to the experimental design and analysis. Using examples from surface-enhanced laser desorption/ionisation time-of-flight (SELDI-TOF) and matrix-assisted laser desorption–ionisation/time-of-flight (MALDI-TOF) experiments, we describe several experimental design issues that can corrupt a dataset. Fortunately, the problems we identify can be avoided if attention is paid to potential sources of bias before the experiment is run. With an appropriate experimental design, proteomics technology can be a useful tool for discovering important information relating protein expression to disease.

BACKGROUND

Proteomics is a promising field that has begun to bloom in recent years, following large-scale explorations of genomics. New techniques employed to understand protein expression at the cellular level have been widely applied to biomedical and clinical problems. One exciting technique is mass spectrometry. Typically, mass spectrometry has been used to derive proteomic expression patterns, which have been put forth as potential biomarkers for the early diagnosis of cancer and other diseases. There are many other exciting applications of proteomics, such as identifying proteins that are only expressed in infected cells to use in developing new antiviral drugs. New experimental opportunities in this field have attracted a lot of attention, and have led to a large number of experiments and vast amounts of new data.

The University of Texas M. D. Anderson Cancer Center (M. D. Anderson) is one of many institutions to have developed and implemented all kinds of proteomics techniques on different tissue samples. Some initial results from proteomics studies run at M. D. Anderson have included the clear separation of multiple known subtypes of one type of cancer; identification of a new disease subtype; differentiating between tissues from patients with and without disease; and extremely good separation of known stages of a disease where no non-invasive markers exist. Our initial analyses seemed encouraging but, in these cases, the results then proved to be wrong. In three of the studies described above, data were brought to us after the experiments had been run. In each study, problems with the experimental design had produced biases that distorted the results. We describe the problems we encountered in detail in case studies 1 to 3 below.

Better results have been obtained when we have had the opportunity to work with the investigators on the design and implementation of the experiments, allowing us to identify potential biases that could be introduced. We have been able to validate the results from these studies, showing them to be reproducible and more reflective of true biological conditions. Case study 4, by contrast, describes a successful experiment.

Experimental design is important

The problems we describe through these case studies are important because they are not unique to M. D. Anderson. Indeed, we believe that the results of some high profile studies show similar flaws.[1] We stress that such problems of bias do not arise from fundamental mechanisms underlying proteomic mass spectrometry, but rather from the complexity and sensitivity of the implementation of the technology. The errors we describe can be avoided with the use of careful experimental design, simple statistics and thorough validation of the results. Our descriptions are not comprehensive, and in particular we will not dwell on issues of multiple testing or complex steps dealing with the exceedingly high-dimensional nature of the data. Rather, we will focus on simple tests that can be undertaken to prevent or correct faulty results.

The importance of experimental design to proteomic mass spectrometry is being more widely recognised; several carefully designed studies are underway by various research groups[2] and some good reviews on the validation of findings have been published.[3] Our focus, however, is of a more visceral nature; we intend to alert experimenters to the need for caution in every step in an experiment, from sample collection to laboratory analysis to data analysis, by illustrating some ways that procedures can go wrong.

CASE STUDIES

Case study 1: Multiple subtypes

Researchers at M. D. Anderson conducted an experiment on serum samples from several patients having one of five types of cancer. A total of 247 patients were involved in the study: 40, 60, 65, 62 and 20 patients who had been diagnosed with cancer of the subtypes 1–5, respectively. Surface-enhanced laser desorption/ionisation time-of-flight mass spectrometry (SELDI-TOF-MS) was applied for protein profiling of the serum samples, producing spectra from four different fractions and three chip surfaces for each patient. Different chip surfaces make different classes of proteins available for analysis and fractionation subsets the groups still further. Such subsetting is often employed so that the signals from high-abundance proteins or peptides do not 'swamp' others nearby. The researchers sought to identify the protein peaks that uniquely defined a given subtype of cancer.

The mass spectrometry data for this study was brought to our group for analysis. We preprocessed the raw spectra using standard routines of simultaneous peak detection and baseline correction (SPDBC). We used an algorithm developed inhouse and implemented in The Math Works, Inc., Natick MA (MATLAB) to perform the SPDBC on each spectrum. Similar code is available as supplementary information to one of the papers by Coombes *et al.*[4] Our most recent codes, a package called Cromwell, is available from the same site.[5] The SPDBC algorithm produced a list of M/Z values corresponding to peaks and also a baseline-corrected spectrum. We then normalised each modified spectrum to the total ion current for the region of M/Z = 2,000 Th and above. Intensities at M/Z values below this showed frequent saturation. Based on the corrected spectra, we identified approximately 100 peaks per fraction/surface pairing shared by all the protein expression profiles from the patients.

We performed a hierarchical clustering analysis of all the samples to evaluate the ability of the peaks to discriminate between the five cancer subtypes. We used an agglomerative clustering algorithm with average linkage and a

distance metric based on the Pearson correlation coefficient. Surprisingly, we observed that simple clustering produced six groups instead of five (see the top panel of Figure 1). We investigated the clinical information and it turned out that the resulting six clusters matched the run dates of the samples, rather than the biologically different groups (see the bottom panel of Figure 1). We found that the serum samples from patients diagnosed with one cancer subtype had been run at least a month before all of the rest, and that the run date affected all of the sample spectra to some degree. We were able to verify this by examining the spectra from a material that is commonly used for quality control (QC), which the researchers had run concurrently. The spectra from the QC material showed the same clustering pattern as the biological samples. We attempted to apply simple additive shifts to align the QC samples to fix the problem, but failed.

Run date effects can be larger than biological effects

Comments

Proteomic profiles are not yet very reproducible over time, and the intensities are semiquantitative at best. To focus on the biological contrasts between groups of tissue samples, we recommend that investigators include some members from each contrasting sample in each laboratory-run group. If the run groups are large, simply randomising the run order will achieve this. Running all samples 'as they come in' is not yet a good way to operate experiments in proteomic mass spectrometry.

Case study 2: Collection protocols

Another group of researchers conducted an experiment at M. D. Anderson on tissue samples from 50 patients with cancer, which were believed to include two subtypes of the disease. The researchers applied three different fractionation protocols (identified as myo25, myo70 and bsa70) to produce three different spectra per sample. Splitting a sample into three fractions can better highlight different subsets of the proteins.

The disease subtype information was 'stripped out' and the resultant blinded

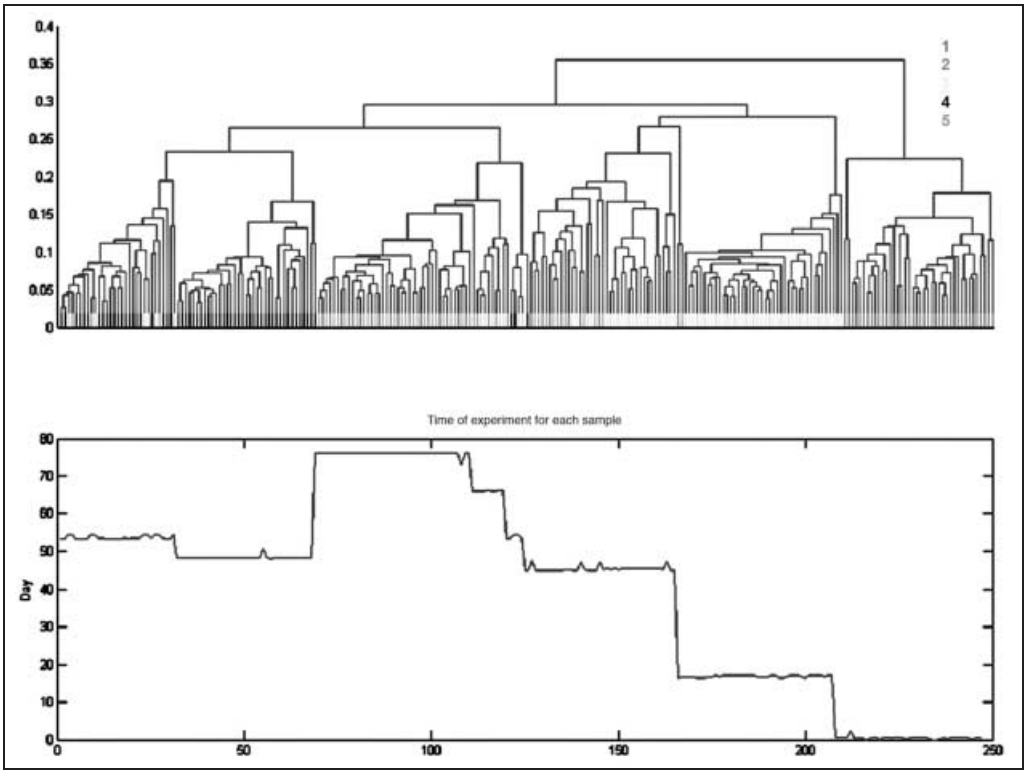

Figure 1: Detection of subtypes of cancer

dataset was brought to our group for analysis. The aim of the analysis was to perform unsupervised clustering of the data to see if the two subtypes could be identified correctly and blindly. We preprocessed the spectral data in a manner similar to that described in the first case study, including the methods of SPDBC and normalisation to the total ion current. We analysed the spectra within each of the three fractions separately. After aligning the peaks across the spectra within each fraction and filtering out the noise, we identified 172, 130 and 130 peaks, respectively, in the fractions from the myo25, myo70 and bsa70 protocols. We then performed hierarchical clustering analyses in each of the three fractions. The results seemed very exciting, with two distinct clusters clearly identified in each fraction. We also observed that the myo25 and myo70 fractions produced the same two clusters, and that clustering from the bsa70 fraction was identical to the others, except for the classification of a single sample. These results were communicated and the data were unblended; however, further exploration showed that the split that we had found did not match the subtypes assumed by the investigators. Rather, the split matched very closely with the day on which the sample collection protocol had been changed midway through the experiment. Figure 2 illustrates the clustering pattern within the fraction bsa70.

Changes in collection protocols can have large effects

Comments

Many features of an experiment affect protein expression profiles, and we have not yet been able to identify all of them. We recommend that investigators define a single protocol and follow it throughout the experiment. This will reduce the number of factors that are of concern during the data analysis. If a protocol must be altered, the investigator should make sure that samples representing both sides of the contrast of interest are present for each run batch that the laboratory processes, and should accordingly be prepared to analyse the data in batches.

Case study 3: Calibration and sample handling

A third group of researchers at M. D. Anderson collected urine samples from individuals for proteomic analysis in the study of cancer. The study focused on five categories of human subjects: disease-free individuals, patients presenting with low-

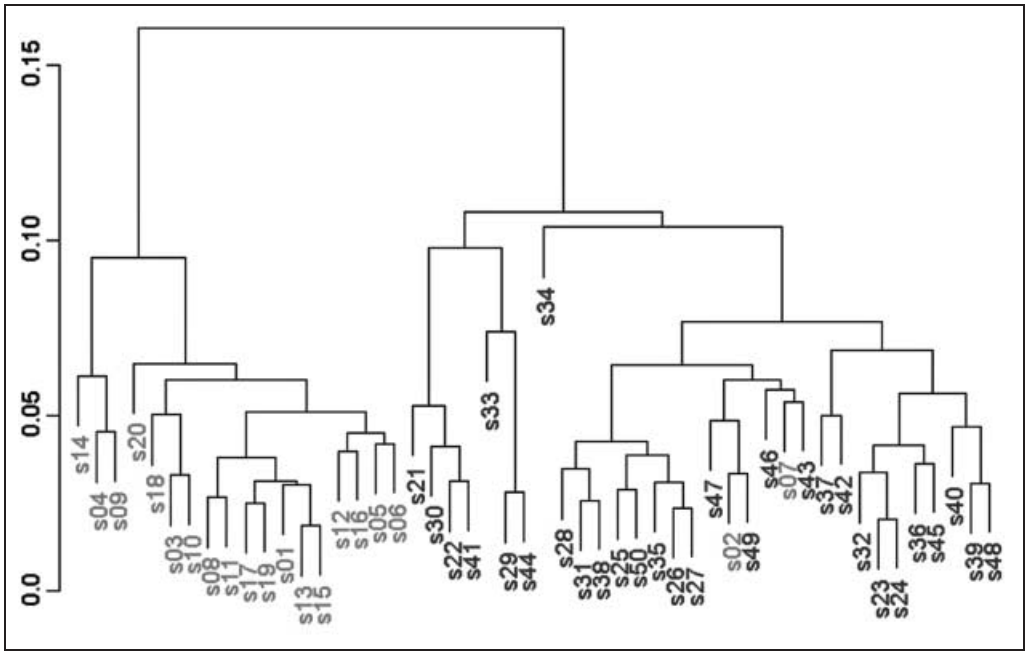

Figure 2: Discovery of clusters in data from bsa70 fraction of tumour samples

grade tumours, patients presenting with high-grade tumours and separate categories of patients who had histories of low-grade or high-grade tumours. A promising goal of the study was to identify important peaks in protein profiling that could differentiate between the disease-free patients and those with tumours of either low or high grades.

The spectral data that were brought to our researchers for analysis had been preprocessed carefully, including the use of baseline subtraction and normalisation methods. Our initial examination selected several peaks that could separate controls from cancers successfully, such as the three peaks shown in the first three rows of Figure 3. We noted that the triplet of peaks is in the same M/Z range as the triplet identified as defensin proteins in urine by Vlahou *et al.*[6] Upon closer examination, we discovered that the spectra from the disease-free individuals also produced the same three peaks, with only a *shift* in the location of the peaks when compared with those derived from the spectra of patients with cancer. We further confirmed the results to be due to an offset of the calibration of the spectra. Once the calibration had been corrected, the difference in protein expression was not present.

Misalignment or lack of calibration can give rise to misleading structures

Fortunately, we could still find some peaks that correlated with cancer based on the corrected spectra. To validate the results, we used an independent group of samples, a 'test set', which had been collected at a different clinic. The peaks that had been selected from the initial study at M. D. Anderson were used as the 'training set', to predict if the data from a subject represented in the test set indicated the presence of cancer. It turned out that the prediction algorithm did not work on the new test set. Careful comparison of the spectra from both sets showed that the spectral patterns were surprisingly *different* — even for subjects from the same category, such as the disease-free individuals. Figure 4 shows an example of a specific region of the spectra produced from the urine samples from

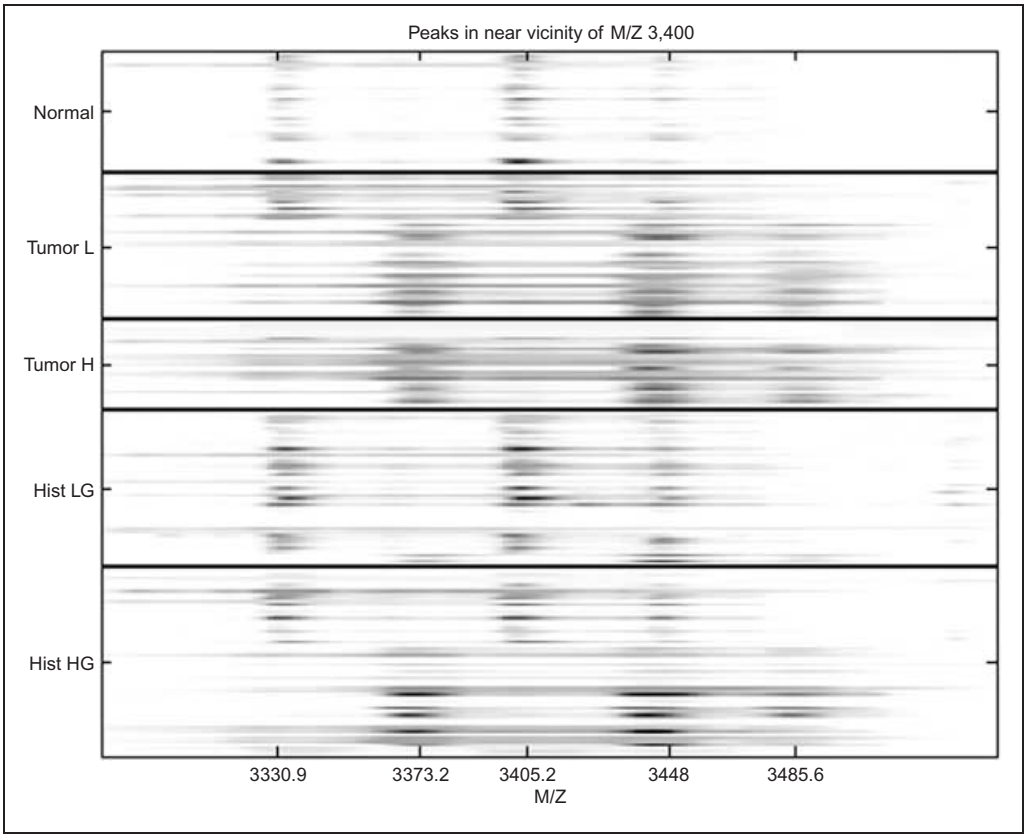

Figure 3: Comparisons of spectra in/near the vicinity of M/Z 3400 in a cancer study

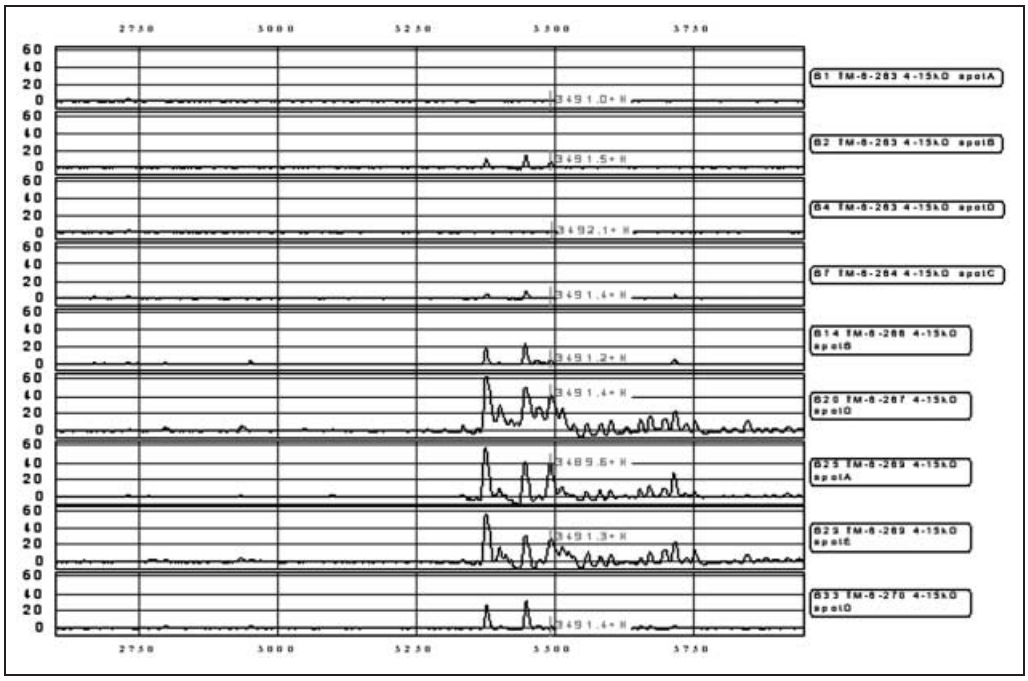

Figure 4: Comparisons of spectra of disease-free subjects from two independent sets (top four, bottom five) in a cancer study

four disease-free individuals in the training set and five in the test set, indicating that the spectra were strikingly different. Interestingly, the same three peaks identified earlier were present at high levels in the samples of the test set taken from disease-free individuals. We then learned that the urine samples in the test set from the disease-free individuals had *degraded*: the samples had not been processed immediately, and had been left sitting at room temperature for several hours prior to processing.

Randomization at several levels can balance effects so that they do not bias the results

Comments

The discovery of important protein biomarkers is very difficult because many factors in the experimental process can easily introduce bias. We recommend that researchers perform calibrations prior to every occasion on which they produce a total of proteomic spectra. It is very important to conduct calibrations before a whole batch of samples are run. When conducting a multicentre study, investigators must ensure that the experimental protocol of every centre is identical, in order to produce sensible results. We also emphasise the importance of pictures; the shift problem is evident when shown diagramatically.

Case study 4: Doing things right

A fourth group of researchers at M. D. Anderson conducted a study of breast cancer within the Nellie B. Connally Breast Center between 2001 and 2003.[7] The main aim was to examine proteomic changes in the plasma of patients with breast cancer in response to chemotherapy with paclitaxel or with a combination of s-fluorouracil, doxorubicin and cyclophosphamide (FAC). Protein biomarkers in plasma profiles that are associated with breast cancer, and which differentiate between women with and without the disease, were identified. The study data were derived from 69 patients with newly diagnosed stage I–III breast carcinoma and 15 healthy volunteers. Plasma samples were obtained on day 0 (before chemotherapy) and day 3 (post-treatment) for all patients, 29 of whom had received preoperative chemotherapy and 40 of whom had received postoperative chemotherapy. Plasma samples from the healthy volunteers (control subjects) were obtained at two time points no sooner than three days apart within a one-week period, for comparison. SELDI-TOF-MS profiling was used to examine protein activity.

The laboratory procedures included thoroughly cleaning the ProteinChips (Ciphergen Biosystems, Freemont, CA) beforehand, to eliminate noise in the spectra due to extraneous chemical materials attached to the chips. Samples were then randomly loaded to the spots on each chip; cases and controls were intermingled and duplicates of all samples were processed and run. In addition, a QC sample was prepared by pooling plasma from three randomly selected participants. Aliquots of the QC sample were spotted on each chip to determine the reproducibility of measurements and to serve as a control protein profile. The mass accuracy was calibrated on the day of measurement for all the spectra, additive white noise was removed, the baseline was subtracted and normalisation was carefully conducted.

We examined the stability of the protein profiles over time using the standard QC sample, as well as the spectra generated from the healthy women at two points in time. Analysis of the experimental data detected a single chemotherapy-inducible peak and five other peaks that distinguished cancer patients from healthy subjects. These results suggested that the peaks were candidates for disease-related biomarkers. Furthermore, a follow-up study was conducted: another set of samples from disease-free women and cancer patients were collected and analysed 3 months after the original study. It turned out that the same set of peaks could be found, indicating that the study was consistent and reproducible. Attempts were made to identify the proteins further; some identities are now known.

Validation can be achieved using different technologies, replication in different labs, or replication with new samples at different times

Comments

Case study 4 represents a 'successful' example of proteomics experiments: the experimental designs and implementations were consistent; identical protocols were strictly followed for the contrasting samples; samples were collected and run on the spot/chip in a random order; quality control materials were applied to achieve good calibration; control samples were prepared in duplicate to test the stability of the protein profile and validate the results; and the spectral data were preprocessed identically. Furthermore, follow-up studies were conducted later, and other technologies were used to identify the proteins. The data processing software Cromwell[8] and other tools are available at http://bioinformatics.mdanderson.org/software.html.

DISCUSSION

We have provided four examples illustrating the importance of careful experimental design for proteomics studies. Much of this involves intelligent application of three principles: randomisation, replication and blocking. These principles are elegantly summarised in classical statistics texts.[9] In order to apply these principles to the problems that researchers in the clinical and basic sciences wish to address, however, it is very useful to consult a statistician before collecting the data in order to determine a good experimental design. The collaborating or consulting statistician can design a trial ensuring that the data collected will be able to answer the question of interest and, equally importantly, can provide implementation guidelines that will prevent the results from being distorted by external biases. Figure 5 summarises the important points that we learned from the case studies.

Advice for the statistician

We recommend that the statistician obtain the clinical data prior to analysing the proteomic data. Clinical data include information on sample preparation and the run order. When provided with data from an experiment that has been designed and conducted without statistical consultation, the statistician may be able to identify obvious problems by skimming the clinical data. First, the statistician will want to check the mechanical contrasts, such as the effects caused by running samples on different days. Most of the problems we detected

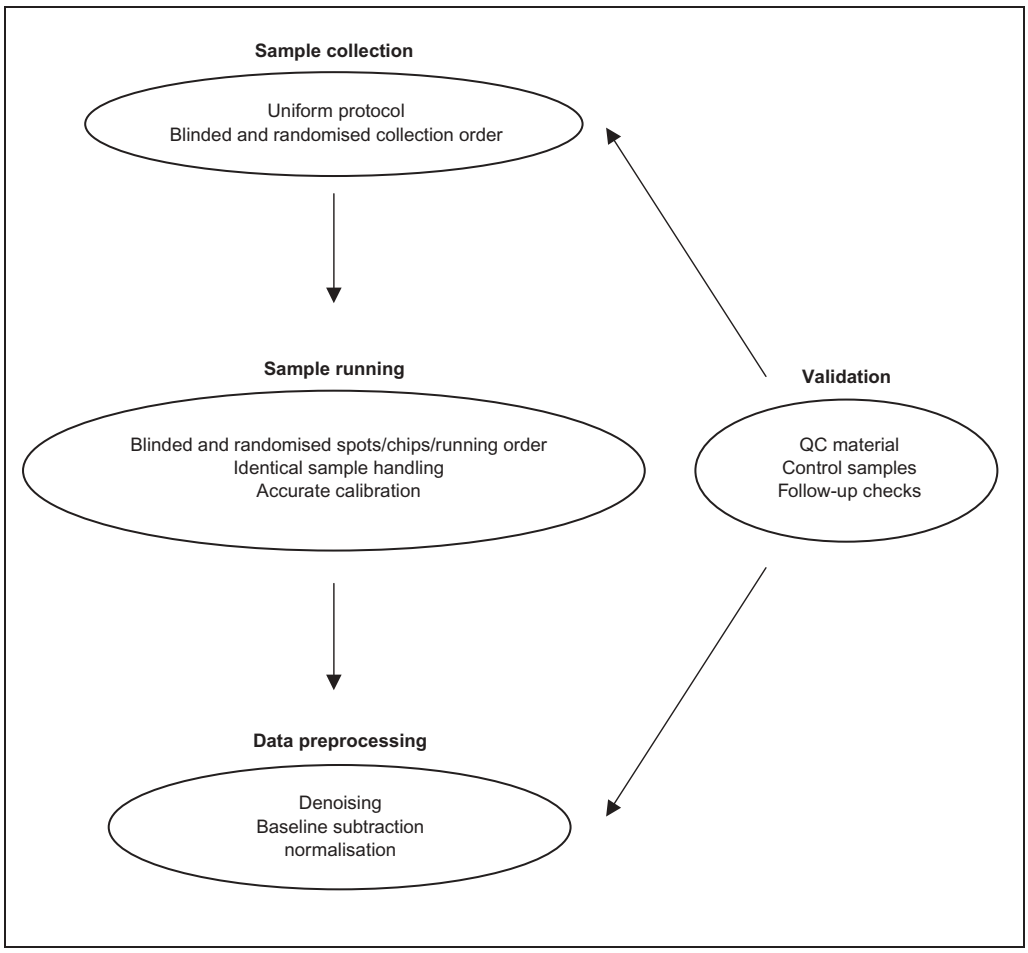

Figure 5: Flowchart of the important issues to know how to address before proteomic expression profiling

from the case studies and elsewhere were found through the use of simple pictures, such as the finding of sinusoidal noise with MALDI data[10] and the use of heat maps of the spectra sorted by run order.[1] Both examples were among the most illuminating findings in our experience with proteomics studies.

Secondly, the statistician should plan to validate the results from the beginning of the experiment. Ideally, this would include identifying the peptides involved and confirming the assay results using a technology that is different from that used in the original experiment (ie using an enzyme-linked immunosorbent assay on results obtained from protein chips and mass spectrometry). Failing that, we recommend that the statistician attempts to reproduce the results using samples produced by a different facility or using a new set of samples produced by the same facility after a gap in time of a few months.

The issues we have addressed are not unique to the field of proteomics; however, gaining specific training and knowledge in proteomics data and also in the physics of mass spectrometry will enable the statistician to spot obvious problems that will bias the results of a proteomic experiment. For example, most of the MALDI and SELDI data that we analyse cover the M/Z range from 0 to about 20,000 Th. In this range, most of the peptides will be singly charged, so that the biggest peak will occur where $Z = 1$; however, there will also be a smaller peak corresponding to the case where $Z = 2$. Superimposing plots of the spectra against M/Z and of the spectra against 2*M/Z provides a quick check of whether the data calibration is off, and the likely charge states of peaks of interest, as the singly and

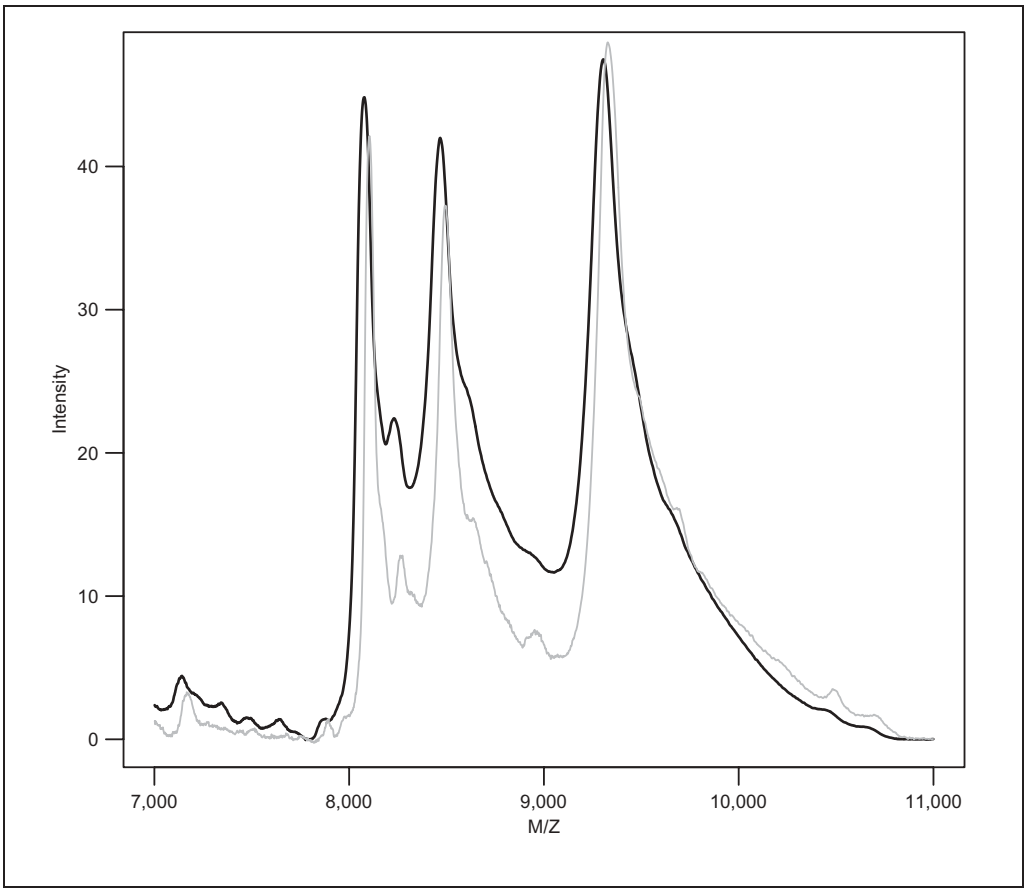

Figure 6: Superposition of singly and doubly charged peaks to check calibration and charge states

doubly charged peaks should line up. The data from a study on prostate cancer[11] are used as an example. Figure 6 shows a specific mass region, with a slight location shift of the set of singly and doubly charged peaks that are respectively illustrated by the black and the grey curves.

Current research in the field of proteomics and, more broadly, in the field of high-throughput biological assays and biomarker discovery is exciting. Using highly advanced technological tools, however, still requires careful adherence to rather simple guidelines and ensuring the reproducibility of experimental findings.

Acknowledgment

Jeffrey Morris' effort was partially supported by a grant from the NCI, R01 CA107304-01.

References

1. Baggerly, K. A., Morris, J. S. and Coombes, K. R. (2004), 'Reproducibility of SELDI-TOF protein patterns in serum: Comparing datasets from different experiments', *Bioinformatics,* Vol. 20, pp. 777–785.

2. Zhang, Z., Bast, R. C., Yu, Y. *et al.* (2004), 'Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer', *Cancer Res.,* Vol. 64, pp. 5882–5890.

3. Ransohoff, D. F. (2004), 'Rules of evidence for cancer molecular-marker discovery and validation', *Nat. Rev. Cancer,* Vol. 4, pp. 309–314.

4. Coombes, K. R., Fritsche, Jr., H. A., Clarke, C. *et al.* (2003), 'Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization', *Clin. Chem.,* Vol. 49, pp.1615–1623. Supplementary information is available at http://bioinformatics.mdanderson.org/supplements.html.

5. Coombes, K. R., Tsavachidis, S., Morris, J. S. *et al.* (2004), 'Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by deionising spectra with the undecimated discrete wavelet transform', *Biostatistics and Applied Mathematic Technical Report UTMDABTR-001-04*, http://bioinformatics.mdanderson.org/.

6. Vlahou, A., Schellhammer, P. F., Mendrinos, S. *et al.* (2001), 'Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine', *Am. J. Pathol.,* Vol. 158, pp. 1491–1502.

7. Pusztai, L., Gregory, B. W., Baggerly, K. A. *et al.* (2004), 'Pharmacoproteomic analysis of prechemotherapy and postchemotherapy plasma samples from patients receiving neoadjuvant or adjuvant chemotherapy for breast carcinoma', *Cancer,* Vol. 100, pp. 1814–1822.

8. Morris, J. S., Coombes, K. R., Koomen, J. M. *et al.* (2004), 'Feature extraction methodology for mass spectrometry data in biomedical applications using the mean spectrum', *Biostatistics and Applied Mathematic Technical Report UTMDABTR-010-04*, http://bioinformatics.mdanderson.org/.

9. Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978), 'Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building', Wiley, New York, NY.

10. Baggerly, K. A., Morris, J. S., Wang, J. *et al.* (2003), 'A comprehensive approach to the analysis of matrix-assisted laser desorption/ ionization time of flight proteomics spectra from serum samples', *Proteomics,* Vol. 3, pp.1667–1672.

11. Petricoin, III, E. F., Ornstein, D. K., Paweletz, C. P. *et al.* (2002), 'Serum proteomic patterns for detection of prostate cancer', *J. Natl. Cancer Inst.,* Vol. 94, pp. 1576–1578.