

Review

## Bioinformatics strategies for proteomic profiling

C. Nicole White,\* Daniel W. Chan, and Zhen Zhang

*Department of Pathology, Johns Hopkins Medical Institutions, Baltimore, MD, 21231, USA*

Accepted 4 May 2004

### Abstract

Clinical proteomics is an emerging field that involves the analysis of protein expression profiles of clinical samples for de novo discovery of disease-associated biomarkers and for gaining insight into the biology of disease processes. Mass spectrometry represents an important set of technologies for protein expression measurement. Among them, surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI TOF-MS), because of its high throughput and on-chip sample processing capability, has become a popular tool for clinical proteomics. Bioinformatics plays a critical role in the analysis of SELDI data, and therefore, it is important to understand the issues associated with the analysis of clinical proteomic data. In this review, we discuss such issues and the bioinformatics strategies used for proteomic profiling.

© 2004 The Canadian Society of Clinical Chemists. All rights reserved.

*Keywords:* Bioinformatics; Proteomic profiling; Mass spectrometry; SELDI; Quality control

### Introduction

Recently, advances in technologies for high-throughput genomic and proteomic expression analysis have introduced a new era of research. The simultaneous measurement of a large number of expressed proteins, known as proteomic profiling, has become an important screening tool for the discovery of new biomarkers. In addition to the direct clinical applications, such as early detection and diagnosis of disease, results of proteomic profiling research also facilitate the generation of hypotheses that may lead to new discoveries which may aid in the understanding of the disease process itself. Novel disease-associated biomarker patterns, identified through proteomic profiling, have recently been reported [1–11]. Yet, few of the published profiles have been verified by a second, independent study, an important requirement in validating the information for clinical use. Verification has lagged because validation studies take time to complete and because validation requirements are still being defined. This review examines the strategies available to validate proteomic profiling data.

Proteomic profiling is a high-throughput technology with a novel set of computational challenges. Some of these relate to identifying the major sources of variability that arise from the protein profiling techniques and experimental design. Non-disease-related sources of variation can be minimized through the selection of an optimal experimental design. By controlling non-disease-related sources of variability, the researcher can focus on evaluating disease-related variability.

This review also focuses on the bioinformatics approaches used to foster biomarker discovery by effectively mining the complex proteomic data streams, in particular, those produced by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI TOF-MS) profiling data.

### Proteomic profiling methods

A variety of analytic methods have been used for the screening, identification, and characterization of proteins. While each technology has its strengths, they also have their limitations. For instance, 2D gel electrophoresis [12] is very good at resolving proteins but is poor at equivalently visualizing all species over the broad concentration ranges spanning many orders of magnitude at which they may be present, especially those in low abundance [13]. Gel-to-gel reproduc-

\* Corresponding author. Center for Biomarker Discovery, Johns Hopkins Medical Institutions, Room 200, 419 North Caroline Street, Baltimore, MD 21231.

*E-mail address:* [cnwhite@jhu.edu](mailto:cnwhite@jhu.edu) (C.N. White).

ibility is another limiting factor. While the 2D gel technology has been available for over two decades and has been widely used, it is being replaced by new technologies optimized for proteomic profiling. Protein/antibody arrays hold considerable promise for functional proteomics and expression profiling, but problems limit their utility as well. Limitations include the lack of high-throughput technologies to express and purify proteins and to generate large sets of well-characterized antibodies [14]. Further research and development are required before the promise of this technique can be fully realized.

Mass spectrometry (MS) successfully addresses the throughput limitation of 2D gels and eliminates the need to purify, identify, and develop antibodies to proteins before proteomic profiling experimentation. SELDI is an affinity-based MS method in which proteins are selectively adsorbed to a chemically modified surface and impurities are removed by washing with buffer [15]. By combining different ProteinChip® array surfaces and wash conditions, SELDI allows on-chip protein capture and micropurification, thereby permitting high-throughput protein expression analysis of a large number of clinical samples [15]. After preprocessing steps involving mass calibration, baseline subtraction, and peak detection, the mass spectra from  $n$  individual samples are converted into peak intensity data typically organized as an  $m$  (peaks)  $\times$   $n$  (samples) matrix, where the intensity of a peak corresponds to the relative abundance of proteins at a molecular mass expressed as mass to charge ratio, or  $m/z$ . A proteomic profile constructed using the SELDI platform is typically a subset of these  $m$  peaks and is constructed using the intensity data of the peaks. SELDI has been used to evaluate specimens from nipple aspirates [16], serum [1–5], urine [7,9], and lysed tissue extracts [17] with little requirement for sample-specific preparation.

While SELDI has several strengths, including the types of biological fluids that may be evaluated, ease of sample preparation, and high-throughput capabilities, the technique also has weaknesses. Issues include the reproducibility of mass spectrometry, in general, and SELDI, in particular, [3,18,19] and the fact that peak amplitudes are measurements of relative protein abundance rather than absolute quantitative measures. Quality control (QC) measures that monitor peak amplitude can identify unacceptable deviations and are discussed later. Another limitation of the SELDI PBS-II mass reader is that protein identities are not returned as the protein profile is collected [18]. The proteins included in a biomarker pattern should be identified to further the understanding of the biology of disease. However, the SELDI profile per se may be useful in the clinical setting when the pattern itself has diagnostic or prognostic significance.

### Data variability

Among the issues associated with expression profiling using clinical samples, systematic biases arising from pre-

analytical variables can be among the most damaging. While careful statistical examination of results and their correlation with possible non-disease-related variables may reveal the existence of biases, no amount of statistical or computational processing can correct such problems within a single set of samples collected under the same conditions. Therefore, errors from systematic differences among samples should be minimized or eliminated whenever possible through good experimental design, careful analysis procedures, and quality control protocols.

Examples of non-disease-associated factors include (1) within-class biological variability which may include unknown subphenotypes among study populations; (2) pre-analytical variables such as systematic differences in study populations and/or in sample collection, handling, and preprocessing procedures; (3) analytical variables such as inconsistency in instrument conditions that result in poor reproducibility; and (4) measurement imprecision.

### Biological variability

When a protein profiling experiment is used for de novo discovery, an adequate sample size is of utmost importance. Defining “adequate” can be tenuous though, when no prior knowledge exists about the proteins the researcher is studying. While several different methods are available for defining sample size [20,21], none works particularly well in the context of proteomic profiling because so little is known about the complexity of the final model before an experiment is completed, or about the protein peaks included in that model. Because SELDI is a high-throughput platform, sample size is often limited by the availability of samples rather than the resources available to examine those samples.

Regardless of the number of tested samples, the consistency of protein peak amplitudes should be evaluated within the disease groups. One method to evaluate the consistency of a peak or peaks within a single data set is through bootstrap analysis. Bootstrap analysis involves resampling, with replacement, from the experimental data [22]. This can reveal, for instance, if a small percent of the cancer samples make the full cancer group appear statistically different from the noncancer group. This technique can be used in conjunction with almost any analysis/modeling technique.

### Site or center variability

Collection practices, sample handling, or storage conditions may be different from institution to institution, and such differences may influence the proteins present in biological fluids [9,18]. Since these biases are often specific to institutions (sites), the use of specimens from multiple institutions combined with sound study design is the preferred approach to discover biomarkers that are truly associated with the disease process. Generally, samples from multiple sites are randomly divided into a discovery, or training, data set and a

validation, or testing, data set. The advantage of such an approach is that the discovery set will be a better representative of the validation set. Statistically, the discovery set and the validation set are guaranteed, albeit artificially, to satisfy the independent and identically distributed (IID) condition, a prerequisite for most statistical inference and learning algorithms. Results are most meaningful when the number of sites is sufficiently large and diverse to form a truly representative sample of the target population. However, by pooling the sites' data, it is possible for a complex multivariate model to pick up the different types of systematic biases existing in the original data sets. An inherent weakness of such an approach, then, is that the artificially created IID condition ensures that the constructed model will perform well in the validation set, although the model's performance may rely on information unrelated to the disease. An alternative approach, which is more conservative, is to complete biomarker discovery and validation by using the sites' data sets separately. The top-ranking candidate biomarkers from the discovery phase are then cross-compared and validated using the other sites' data. This type of discovery model mimics the multicenter validation process that any clinical biomarker will eventually have to pass before clinical use. Considering the effort and cost required for postdiscovery validation, it is important to incorporate a sound and sometimes conservative study design into the discovery phase of biomarker research.

#### *Processing error*

Error is also introduced when processing specimens during preanalytical and analytical phases of the experiment. Before analysis, processing of specimens may include pipetting, diluting, and other manipulations to the sample. While each of these steps is, theoretically, applied to all the samples uniformly, differences can occur. One set of chips used in one bioprocessor (96 samples), for instance, may be processed differently than those in the second or third bioprocessor. Different batches of buffer may be used throughout an experiment. Samples spotted with an air bubble result in poor spectra results. Chip variability may be controlled (to some degree) by using chips from the same lot, and chemicals used during a single experiment should be from the same batch. Furthermore, an automated robot can remove variability during sample transfer and processing. Because an error introduced during processing can be difficult or impossible to trace once the experiment is completed, it is best to rigorously control the experimental procedure to minimize the introduction of variation in the first place.

Analyzing replicates of spotted samples is highly recommended. It is usually not feasible to run enough replicates to have the statistical power to remove any replicates that are "unusual." Instead, samples should be spotted in triplicate and the mean of the three observations used in lieu of a single observation. In this way, the effect of spotting errors is reduced.

#### *Instrument variability*

Use of a similar chip-reading protocol for all experiments reduces variation between experiments [9]. The standard operating procedures should specify procedures used to select the instrument parameter settings, the number of laser shots to be averaged, and reading frame from which to collect the spectra. None of these items should vary from run to run throughout the experiment.

Unfortunately, the same machine parameters will not continue to produce identical spectra over time. In SELDI, several pieces of the instrument, such as the laser and detector, have a limited life span. Consequently, the spectra from the same machine will look slightly different with ensuing use. Variation also arises from differences among batches or lots of chips; ideally, chips from the same batch should be used throughout the experiment. Spot-to-spot variation, although small, also occurs. Given these instrument and chip-related sources of variability, specimens should be spotted in random order on the chips. Randomization can limit the confounding effects of these variation sources. Published SELDI experiments do not uniformly report using randomization protocols. Ciphergen's new software system, Ciphergen Express Data Manager Version 2.1, includes a built-in randomization tool. This package may make it easier for research groups without statistical or bioinformatics support to carry out randomization protocols.

#### *Quality control (QC)*

Relatively little has been published on how to incorporate quality control procedures into the proteomic experimental protocol. Petricoin and Liotta [19] have discussed the need to evaluate between and within experimental changes. Coombes et al. [16] developed a protocol that assessed the reproducibility of SELDI spectra and, by including QC material, were able to identify and exclude chips that generated spectra significantly different than the norm. Sorace and Zhan [23] demonstrated that bias was evident within one SELDI experiment, leading them to conclude that further analysis should be completed within any SELDI experiment to discover if bias is present. However, in other publications, coefficient of variation (CV) of peak height within an experiment is the primary, if not the only, measure of reproducibility evaluated [2,3,5,8–10,17]. Any new technology, particularly one being presented as a potential clinically used diagnostic tool, requires stringent QC to evaluate analytical performance over time.

The variability arising from measurement processing error and instrument drift can be evaluated using QC procedures. To detect processing errors, pooled human serum, for instance, can be randomly spotted with the experimental serum. These samples are processed with the study samples and are compared to determine if any changes occurred during the course of the experiment. Instrument

performance, however, must be compared not only during one experiment, but also over the course of time. One large specimen pool with volume to last at least several months should be aliquot and frozen. An aliquot should be tested in each experimental run. By sampling from one uniform specimen and assuring that all samples go through the same number of freeze/thaw cycles, the (theoretically) same spectra may be compared for changes over time. QC procedures and rules well-established in clinical laboratory testing can also be applied in the research setting [24].

#### *Adverse effects caused by excessive variation and poor experimental design*

- (1) Bias can be introduced at any experimental step. It could include changes to medication or lifestyle in response to cancer diagnosis [23] or differing sample collection procedures for samples and controls. Systematic errors can artificially worsen or improve a profile's discriminatory accuracy. It is impossible to evaluate if the profile's accuracy is based on disease pathology or bias if systematic differences between samples are introduced before sample processing. However, QC procedures may differentiate between the two if they are introduced at or following SELDI chip preparation.
- (2) Excessive variability in spectral peak amplitude will substantially degrade the utility of experimental results. A case in point is the 50–60% imprecision reported by Yasui et al [8]. Such poor CVs forced the authors to forgo quantitative analysis and revert to the much less discriminating presence or absence of the peak as the outcome measure.
- (3) Mass shifts caused by machine variability and processing error lead to poor mass accuracy. Several groups report mass accuracy around 0.1% [2,8,17], but mass shifts can be greater, particularly in large experiments where chips are read over the course of a week rather than a day. With increased mass shifting, the researcher cannot rely on peak detection via available software packages. Ciphergen's automated peak detection software assumes each peak  $m/z$  varies less than 0.3%. When poor mass accuracy is observed, the researcher must select peaks manually. Manual peak selection can be more accurate than the software but is tedious and time consuming.

#### **Strategies for analysis**

One of the common characteristics of expression profile data is high dimensionality in comparison to a relatively small sample size. This characteristic was uncommon before the development of microarrays and necessitated the recent development of novel methods to analyze profiling data. Several of these methods are described below including those used to evaluate the stability of identified candidate

biomarkers through bootstrap analysis and/or validation data sets. While other analysis techniques exist, the methods selected here are some of the most popular and well-published methods available.

The most rudimentary statistical analysis involves univariate tests. Besides traditional statistics like the  $t$  test and its nonparametric equivalent, the Wilcoxon test, univariate methods have been developed specifically for the analysis of expression profiles. These determine the significance of observed changes while accounting for the large number of variables [6,25]. For example, the univariate methods developed for microarray analysis but applicable to proteomics assess the significance of discriminatory profiles by evaluating permutations of repeated measurements to estimate the percent of changes that would be identified by chance, known as the false discovery rate [25]. While this approach improves a researcher's ability to identify statistically significant changes in expression, it cannot account for the interdependence of the variables.

It is plausible to assume on biological grounds that the proteins present in the proteomic profile are not fully independent of each other *in vivo*. For this reason, a multivariate approach to analysis is preferred because it can address the correlations among variables. Unfortunately, one of the strengths of proteomic analysis, namely, the large number of variables that can be measured simultaneously, becomes a limitation for this type of analysis. The large number of variables compared with the (usually) small number of observations results in an unstable estimate of the covariance matrix. Simultaneous multivariate analysis requires a stable estimate of the covariance matrix.

To use a multivariate approach and circumvent the issue of covariance matrix estimation, a dimension reduction step is employed. Dimension reduction methods project a large number of genes or proteins onto a smaller and more manageable number of clusters [26,27], or some type of supervariable [28]. The conditional density function can be used to construct a decision rule. This decision rule combines peak intensities to cluster the samples into diseased or nondiseased clusters. In real world experiments, it is rare that all samples can be classified correctly, so the probability of incorrectly classified samples is calculated as the probability of error [29].

Some of the most commonly used dimension-reduction techniques employ clustering methods such as principle component analysis (PCA) [2,6,16,30]. PCA is an unsupervised analysis tool: samples are classified without including disease status in the training algorithm. In PCA, the training samples, regardless of their relative location to the underlying class boundaries in the variable space, contribute equally to the estimation of the data distributions and the classification function. On the other hand, in some supervised approaches to dimension reduction, the samples that are close to the boundaries are weighted much more heavily than the interior samples. As an extreme example, the support vector machine (SVM) [31,32] model solutions

are solely determined by the support vectors that consist only the boundary data points. The removal of interior samples does not affect the solution at all. Because each clinical sample represents a considerable amount of effort and cost, limiting analysis to the support vectors is not the most efficient use of information. In addition, analysis that relies solely on the support vectors could be very sensitive to labeling errors in the training samples of small sample studies. However, for the purpose of data classification rather than representation, inaccuracies may be introduced by treating all samples equally.

With the above shortcomings in mind, the unified maximum separability analysis (UMSA) algorithm was developed for genomic and proteomic expression data [1,5,33]. The conceptual framework of UMSA is very straightforward. In the original SVM learning algorithm [31], a constant,  $C$ , limits the maximum influence of any sample point on the final SVM model solution. In UMSA, this constant becomes an individualized parameter for each data point to incorporate additional statistical information about the data point's position relative to the distribution of all the classes of samples. The rationale behind UMSA is that information about the overall data distribution (although the estimation itself might not be perfect) can be used to prequalify the trustworthiness of any training sample to be a support vector. The final solution, therefore, will rely on the weighted contributions of the support vectors and be less sensitive to labeling errors of a small percentage of samples.

The construction of a linear UMSA classifier provides a supervised multivariate method to rank a large number of variables. Similar to unsupervised component analysis methods such as PCA, the UMSA-based procedure is also a linear projection of data. However, in PCA, the axes in the new space represent directions along which the data demonstrate maximum variations. In UMSA component analysis, the new axes represent directions along which the two classes of data are best separated by linear classification. When it is used for dimension reduction, the smaller number of axes may be viewed as composite features that retain most of the information relevant to the separation of data classes. The UMSA-based software system has been used for genomic expression data analysis [33] and more recently for biomarker discovery using clinical proteomic profiling [1,5,10,11] generated by SELDI.

After selecting a limited number of protein peaks as candidate biomarkers, more traditional linear modeling techniques may be used. Methods such as logistic regression [34,35] allow the user to define an equation describing the relationship between protein peaks as well as explicitly evaluating the significance of each peak's contribution toward the multivariate relationship.

Regardless of the modeling and/or analysis technique used, it is advisable to complete an additional step and evaluate the robustness of the final model. Procedures available to complete this step are determined by the sample size of the study. Many groups split their full data set using a

stratified random sampling procedure [2,4,6,11]. Two data sets are constructed from the original data set, and the decision rule derived using one data set is tested in the second data set. Testing the decision rule is necessary because of sample or biological variability.

If the number of samples collected for the study is too small for a stratified random sampling procedure, bootstrap analysis [5,23] may be performed. These procedures provide the advantage of using the entire data set during discovery and validation. However, because the two data sets are not independent, the results may be overly optimistic and difficult to verify in a second, independent, study. Overfitting can be a serious problem for complex multivariate models and may result in an amplification of non-disease-associated data variability on the analysis results.

Additionally, bootstrap analysis may be used during the discovery phase alone to evaluate peak consistency before validation is completed [11] or may be used to identify several different, but equally performing, sets of candidate biomarkers [6]. Regardless of what technique is employed, verification of analysis results should be completed before proteomic profiling analysis is published.

## Summary

Advances in high-throughput technologies, such as SELDI, have made it possible to obtain expression profiles of a large number of proteins using clinical samples. Recent reports have raised the expectation for the application of proteomic profiling to clinical diagnostics. In this paper, we have reviewed and discussed several critical and often overlooked issues in translating results from proteomic profiling to biomarker discovery and to eventual clinical applications.

The clinical evaluation of a diagnostic test relies mostly on its efficacy in terms of positive and negative predictive values in a targeted population. It has been proposed that if a particular protein expression pattern can be associated with a disease condition, the pattern itself could be used as a diagnostic test. However, the identification of the composing molecular entities, as part of the process of biomarker discovery, will not only facilitate assay development for large-scale validation and clinical use, but also help the understanding of the biology of the disease itself and lead to additional discoveries.

Improvement in profiling technologies, better quality control procedures, and the proper use of sophisticated bioinformatics and statistical tools are all important factors to ensure true discoveries in clinical proteomics. However, the single most significant factor that affects the discovery and verification of candidate protein expression patterns or biomarkers is the selection of clinical samples. False results can be obtained from studies using a sample set from a single institution that exhibits significant and systematic biases due to sample inclusion/exclusion criteria and/or specimen han-

dling, processing, and storage conditions. Recent reports have demonstrated the necessity and benefits of multicentered studies.

Proteomic profiling is a new approach to clinical diagnosis, and many computational challenges still exist. Not only are the platforms themselves still improving, but the methods used to interpret the high dimensional data are developing as well.

## Acknowledgments

This work was supported in part by a grant from CIPHERGEN Biosystems, Inc. (Fremont, CA), and by an NCI Grant 1P50 CA83639, UTMDACC Specialized Programs of Research Excellence (SPORE) in Ovarian Cancer. We would also like to thank Lori Sokoll for her assistance in reviewing the article.

## References

- [1] Rai AJ, Zhang Z, Rosenzweig J, et al. Proteomic approaches to tumor marker discovery. *Arch Pathol Lab Med* 2002;126:1518–26.
- [2] Petricoin III E, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002;359:572–7.
- [3] Qu Y, Adam BL, Yasui Y, et al. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin Chem* 2002;48:1835–43.
- [4] Petricoin III EF, Ornstein DK, Paweletz CP, et al. Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst* 2002;94:1576–8.
- [5] Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem* 2002;48:1296–304.
- [6] Zhu W, Wang X, Ma Y, Rao N, et al. Detection of cancer-specific markers amid massive mass spectral data. *Proc Natl Acad Sci U S A* 2003;100:14666–71.
- [7] Clarke W, Silverman BC, Zhang Z, Chan DW, Klein AS, Molmenti EP. Characterization of renal allograft rejection by urinary proteomic analysis. *Ann Surg* 2003;237:660–4.
- [8] Yasui Y, Pepe M, Thompson ML, et al. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* 2003;4:449–63.
- [9] Schaub S, Wilkins J, Weiler T, Sangster K, Rush D, Nickerson P. Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry. *Kidney Int* 2004;65:323–32.
- [10] Koopman J, Zhang Z, White N, et al. Serum diagnosis of pancreatic adenocarcinoma using surface-enhanced laser desorption and ionization mass spectrometry. *Clin Cancer Res* 2004;10:860–8.
- [11] Li J, White CN, Zhang Z, et al. Detection of prostate cancer using serum proteomic pattern. *J Urol* 2004;171:1782–7.
- [12] Anderson L, Anderson NG. High resolution two-dimensional electrophoresis of human plasma proteins. *Proc Natl Acad Sci U S A* 1977;74:5421–5.
- [13] Gygi SP, Rist B, Aebersold R. Measuring gene expression by quantitative proteome analysis. *Curr Opin Biotechnol* 2000;11:396–401.
- [14] Abbott A. A post-genomic challenge: learning to read patterns of protein synthesis. *Nature* 1999;402:715–20.
- [15] Fung ET, Thulasiraman V, Weinberger SR, Dalmasso EA. Protein biochips for differential profiling. *Curr Opin Biotechnol* 2001;12:65–9.
- [16] Coombes KR, Fritsche Jr HA, Clarke C, et al. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin Chem* 2003;49:1615–23.
- [17] Zheng Y, Xu Y, Ye B, et al. Prostate carcinoma tissue proteomics for biomarker discovery. *Cancer* 2003;98:2576–82.
- [18] Diamandis EP. Point: proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? *Clin Chem* 2003;49:1272–5.
- [19] Petricoin III E, Liotta LA. Counterpoint: the vision for a new diagnostic paradigm. *Clin Chem* 2003;49:1276–8 [22].
- [20] Munoz A, Rosner B. Power and sample size for a collection of  $2 \times 2$  tables. *Biometrics* 1984;40:995–1004.
- [21] Connor RJ. Sample size for testing differences in proportions for paired-sample design. *Biometrics* 1987;43:207–11.
- [22] Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat* 1979;7:1–26.
- [23] Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 2003;4:24.
- [24] Westgard JO. Internal quality control: planning and implementation strategies. *Ann Clin Biochem* 2003;40:593–611.
- [25] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 2001;98:5116–21.
- [26] Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 2002;18:39–50.
- [27] Tibshirani R, Hastie T, Narasimhan B, et al. Exploratory screening of genes and clusters from microarray experiments. *Stat Sin* 2002;12:47–59.
- [28] Huang E, Cheng SH, Dressman H, et al. Gene expression predictors of breast cancer outcomes. *Lancet* 2003;361:1590–6.
- [29] Fukunaga K. Introduction to statistical pattern recognition. San Diego: Academic Press; 1990. 52 pp.
- [30] Landgrebe J, Wurst W, Welzl G. Permutation-validated principle component analysis of microarray data. *Genome Biol* 2002;3:1–11.
- [31] Vapnik VN. Statistical learning theory. New York: Wiley-Interscience; 1998. 736 pp., 23.
- [32] Wagner M, Naik D, Pothan A. Protocols for disease classification from mass spectrometry data. *Proteomics* 2003;3:1692–8.
- [33] Zhang Z, Page G, Zhang H. Applying classification separability analysis to microarray data. In: Lin SM, Johnson KF, editors. *Methods of microarray data analysis*. Boston: Kluwer Academic Publishers; 2001. p. 125–36.
- [34] Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med* 1993;118:201–10.
- [35] Moss M, Wellman DA, Cotsonis GA. An appraisal of multivariable logistic models in the pulmonary and critical care literature. *Chest* 2003;123:923–8.