Michael Wagner[1]
Dayanand Naik[2]
Alex Pothen[3]

[1]Pediatric Informatics,
Cincinnati Children's Hospital
Medical Center, Cincinnati,
OH, USA
[2]Mathematics and Statistics
[3]Computer Science
Old Dominion University,
Norfolk, VA, USA

Protocols for disease classification from mass spectrometry data

We report our results in classifying protein matrix-assisted laser desorption/ionization-time of flight mass spectra obtained from serum samples into diseased and healthy groups. We discuss in detail five of the steps in preprocessing the mass spectral data for biomarker discovery, as well as our criterion for choosing a small set of peaks for classifying the samples. Cross-validation studies with four selected proteins yielded misclassification rates in the 10–15% range for all the classification methods. Three of these proteins or protein fragments are down-regulated and one up-regulated in lung cancer, the disease under consideration in this data set. When cross-validation studies are performed, care must be taken to ensure that the test set does not influence the choice of the peaks used in the classification. Misclassification rates are lower when both the training and test sets are used to select the peaks used in classification *versus* when only the training set is used. This expectation was validated for various statistical discrimination methods when thirteen peaks were used in cross-validation studies. One particular classification method, a linear support vector machine, exhibited especially robust performance when the number of peaks was varied from four to thirteen, and when the peaks were selected from the training set alone. Experiments with the samples randomly assigned to the two classes confirmed that misclassification rates were significantly higher in such cases than those observed with the true data. This indicates that our findings are indeed significant. We found closely matching masses in a database for protein expression in lung cancer for three of the four proteins we used to classify lung cancer. Data from additional samples, increased experience with the performance of various preprocessing techniques, and affirmation of the biological roles of the proteins that help in classification, will strengthen our conclusions in the future.

Keywords: Biomarker discovery / Discrimination methods / Matrix-assisted laser desorption/ionization-time of flight mass spectrometry / Support vector machines PRO 0519

1 Introduction

We report results on the second challenge problem provided for the First Annual Proteomics Datamining Conference, organized by the Departments of Radiology and Biostatistics at Duke University in September 2002. We discuss the preprocessing techniques employed on the mass spectra to obtain data amenable for classification, and we assess the significance of the classification results obtained using several algorithms. The data consist of protein mass spectra obtained from serum samples of 41 individuals, 24 of whom have been diagnosed with a disease (revealed at the conference to be lung cancer), and 17 healthy individuals. Each sample was further split into twenty fractions that were obtained by varying the pH during sample preparation. The challenge question was: Is it possible to find patterns among the protein mass spectra of these samples that characterize and distinguish healthy individuals from those with diesease? An affirmative answer could lead to a potential diagnosis tool; additionally, the identification of proteins with different expression levels in diseased and healthy samples provide valuable insight into the pathways that underlie the disease in question. We were provided with the raw data sets as well as a processed set which contained the locations and (raw) intensities of peaks as identified by the software that comes with the MALDI-TOF instrument. At the conference we presented classification results from the processed data; in this paper we present results on the raw data, since we wish to study the influence of preprocessing on the classification results.

Correspondence: Dr. Alex Pothen, Computer Science Department, Old Dominion University, Norfolk VA 23529-0162, USA
E-mail: pothen@cs.odu.edu
Fax: +1-757-683-4900

Abbreviations: B/W, Ratio between group sum of squares to within group sum of squares ratio; **kNN**, k nearest neighbor discrimination method

Preprocessing of the data is of crucial importance and significantly influences the quality of the classification results. Figure 1 illustrates the many steps involved in biomarker discovery beginning with the mass spectra. Each preprocessing step allows for a number of different options, and only with a thorough understanding of the experimental setup, the preprocessing methods and the significance of the output from the mass spectrometry instrument can one hope to generate meaningful classifiers. We will briefly mention the choices we made in the next few sections. There is no consensus in the literature about how the various preprocessing steps should be done. The hope remains that if a strong signal is truly present in the given data, then it will not be too sensitive to the details of the preprocessing, and information critical for building models with strong prediction capabilities will be retained. Our findings should be considered tentative for this reason. With more experience, the preprocessing methods will improve in sophistication and robustness. Biological insight (*i.e.*, identification of the proteins whose peaks we used in our classification) will help validate our choices; until then we want to explore and expose the options at hand, and provide a framework useful for diseases to be classified using protein profiles obtained through MS. A recent discussion of the role of MALDI-TOF MS in proteomics is included in [1].

2 Materials and methods

In order to make the data amenable to classification, we need to transform the mass spectra of each fraction (a total of 820) into vectors, preferably of low dimension, that characterize the samples (the peak profiles). Each raw spectrum consists of 60 831 intensity measurements at discrete mass/charge (*m/z*) values. Given the small sample size of 41, our first goal is to reduce this data to, say, less than twenty peaks that discriminate between healthy and diseased states.

2.1 Baseline identification and subtraction

Each mass spectrum exhitits a base intensity level (a baseline) which varies from fraction to fraction and consequently needs to be identified and subtracted. This noise varies across the *m/z* axis, and it generally varies across different fractions, so that a one-value-fits-all strategy cannot be applied. Figure 2 shows a sample spectrum from the dataset. We see a near-exponential decay in the noise at the beginning, after which the noise level appears to be a linear function of the *m/z*. The picture is very similar (but not identical) for other samples. Our goal was to get a rough approximation of the baseline, and we used local linear regression (as implemented in the SAS soft-

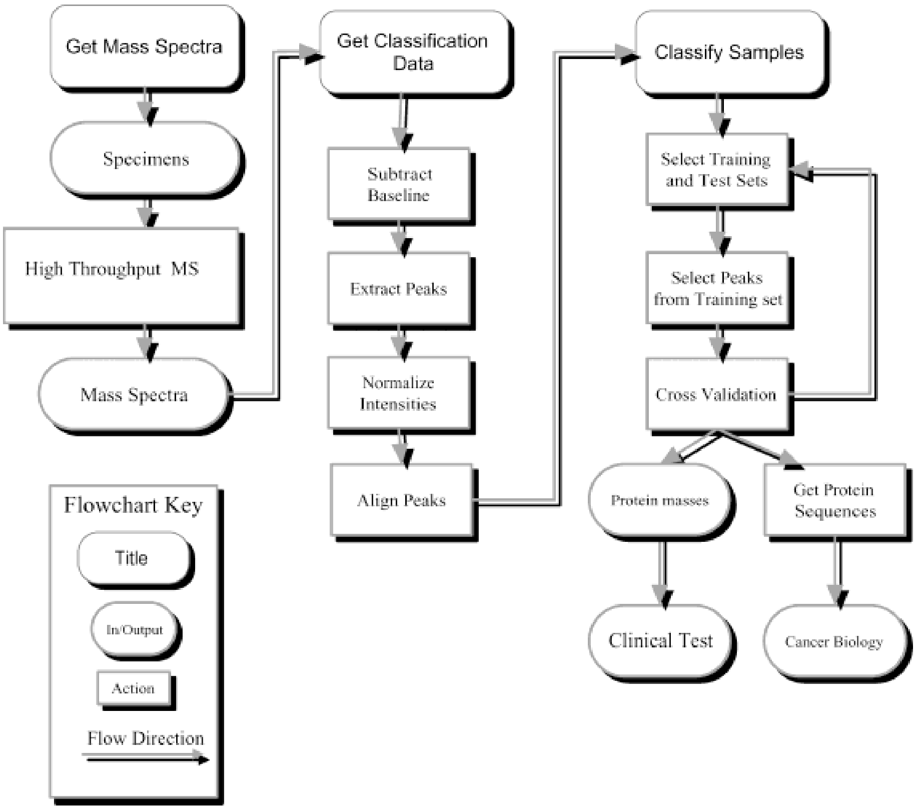

Figure 1. An overview of the various steps involved in biomarker discovery using mass spectrometry. For simplicity we have omitted the processing of fractions in the current data set.

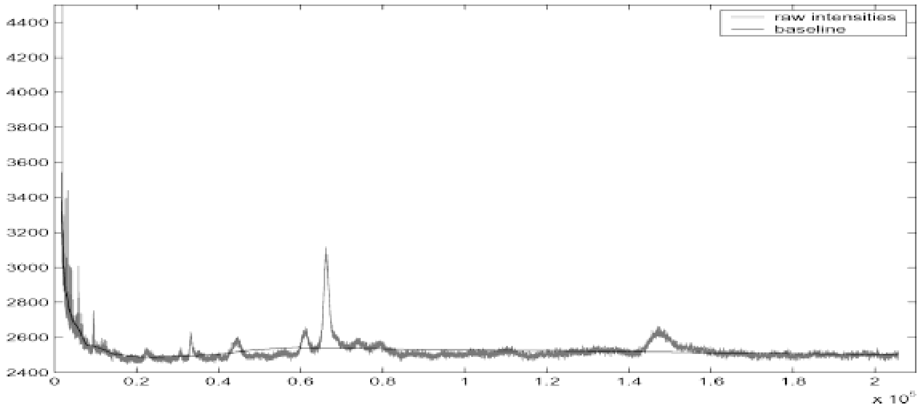

Figure 2. Sample protein mass spectrum (fraction 10 from sample A01) with baseline identified.

Table 1. Smoothing parameters used in local linear regression for baseline identification.

m/z (kDa/z)	1.7–2.3	2.3–3.7	3.7–10	10–15	15–40	40–60	60–100	100–205
smoothing parameter	1%	2%	5%	10%	20%	40%	50%	70%

ware package (SAS Institute, VA, USA) by the LOESS procedure) iteratively in order to smooth over the peaks. To deal with the exponential decay at the beginning we used varying degrees of smoothness, depending on the *m/z* interval being considered. A few attempts, validated by visual inspection of a few samples, yielded the choice of parameters tabulated in Table 1.

A look at the processed dataset (and insight gained from the discussion at the conference) revealed that peaks with *m/z* under 1.7 kDa were deemed to either stem from matrix molecules or contaminants, so we chose to ignore them. A second iteration of smoothing was applied by identifying intensity values which deviated from the baseline by more than one standard deviation. Those values were (temporarily) replaced by their corresponding baseline values, and the smoothing technique was re-applied. As can be seen in the example in Fig. 2, the resulting baseline appears to be satisfactory. Future work will explore alternative techniques for this step.

2.2 Peak identification and extraction

The problem of identifying peaks in a mass spectrum is a central one that deserves careful consideration. However, in order to be able to directly compare with the processed data given to us, we took the masses of the peaks from the processed data; these were identified by the software provided with the mass spectrometry instrument, coupled with some human processing. We noticed that the peak locations provided in the processed data did not always quite correspond to local maxima in our baseline-corrected mass spectrum, and so we took the mass location and intensity of the local maximum within thirty measurement points of the peak mass from the processed data. In rare cases this resulted in peak intensities that were negative (*i.e.*, the "peak" lies beneath the computed baseline). Visual inspection of several such examples revealed that the intensities at these points were not really distinguishable as peaks, and so we deemed it safe to ignore these rare cases. Future versions of our processing strategy will incorporate a custom peak identification procedure based on the distribution of intensities, taking advantage of the fact that peak intensities correspond to points in the tail of the overall distribution of intensities.

2.3 Intensity normalization

Details of the experimental setup are such that the absolute peak intensities are not comparable across different fractions, let alone samples. This motivates the need for a normalization scheme which ultimately enables the merging of the peak profile of fractions into a peak profile of a sample. One could think of a number of choices of how to normalize: with respect to the maximum intensity in a sample, using the sum of all peak intensities, or, possibly, using the total area under the peaks as reference value. None of these is an obvious choice, and all have severe defects in the presence of pathological examples. After some discussion we chose to normalize with respect to the sum of the intensities. And so each peak intensity was divided by the sum of all peak intensities in that fraction and multiplied by 1000, so that the processed intensities could be interpreted over a uniform range across fractions and samples.

2.4 Merger of fraction data into sample data

There are numerous cases where several fractions display peaks at very similar mass points, and so the question arises how to decide when two peaks in different fractions are to be considered to stem from the same protein and when they represent different proteins. In order to extract a single peak profile *per* sample one needs to merge the normalized peak profiles from the twenty fractions. The mass accuracy of the instrument was given to be approximately 0.1%. We chose the following heuristic when merging peaks: If the masses of two peaks are within 0.2%, we merge them and assign the new peak to have a mass of the average of the two and its intensity to be the maximum of the two peaks. The tolerance of 0.2% was intentionally chosen to be larger than the instrument accuracy to additionally smooth the data. This scheme was applied iteratively, with subsequent new peak masses to be chosen as the weighted average of the previous peaks.

2.5 Peak alignment across samples

Finally, in order to make the peak profiles comparable across different samples, we need to align them, *i.e.*, to find one common set of peak locations across all samples that will work as coordinates for the vectors we will use for each sample in the classification schemes to follow. The idea we used is identical to the one used when merging fractions into samples: if two peaks are within 0.2% of each other then they will be considered identical and their masses are reassigned.

2.6 Peak selection

The preceding steps result in vectors of length 603 for each sample which we now take to characterize the samples. Of the 603 peaks in this reduced dataset, over 60% appear in only very few samples and are thus not likely to be helpful in classifying the majority of the samples. Hence we chose to ignore any peaks that occurred in fewer than eight samples, a step which reduced the dimension of the identifying vectors down to 229. There is no hope of getting statistically significant results if the number of data points is less than the degrees of freedom, so we need to further reduce the number of peaks used in the classification. We ordered the peaks according to their information content as measured by the F-statistic. This is equivalent to computing the ratio of variances of peak intensities between and within the two groups (B/W ratio) and sorting in decreasing order. A similar technique has been used in classifying cancers using gene expression data, where it is called gene selection [2]. We subsequently experimented with the classification algorithms using between three and fifteen peaks as ordered by the B/W criterion. The questions we address now are: Can we, after these carefully chosen but admittedly still rather *ad hoc* preprocessing steps, identify peaks corresponding to biomarkers that are fundamentally affected by the disease? Furthermore, is it possible to assess the confidence of our prediction?

3 Results and discussion

We report classification results using established statistical and optimization-based tools: linear discrimination, quadratic discrimination, nonparametric discrimination using a kernel, nonparametric discrimination using *k*-nearest neighbor classification (kNN) using the Mahalanobis distance, and linear support vector machines (SVM). We do not give detailed descriptions of these classification methods here due to space considerations, but refer the interested reader to the excellent discussions of these methods in [3, 4].

We implemented the entire scheme from Fig. 1 using a combination of languages and tools such as Perl, SAS and Matlab (Mathworks, Natick, MA, USA). The SVM^{light} software [5] was used for the support vector machine. Some of these methods require an *a priori* choice of parameters, in particular we chose $k = 6$ for kNN, r = 0.5 for the kernel method and $C = 1$, where C is the tradeoff parameter between margin maximization and misclassification error in the SVM. We stress that we did not perform extensive parameter tuning experiments here. A common practice from both the statistics and machine learning literatures is to use cross-validation to test the power and quality of the methods. The data set is split into a training set, which is given to the classification method in order to build the model, and a test set, which is used to assess the quality of the model. The rather small size of the dataset (41 samples) constrains us to perform only a leave-one-out cross-validation test. Here 40 samples are used as training set and the remaining one is used as test sample. This is done until each of the 41 samples has been left out, and we report on the overall classification results.

We examined two cases, one where the peak selection (using B/W ratios) is performed on the entire dataset ("preselected peaks"), and another where peaks are reselected for each training set. The former biases the peak selection using information from the test set and, as such, is not quite a fair test of the generalization capabilities of the models. Consequently, one would expect the second validation test to be more stringent and to predict higher and more realistic error rates. In the case of preselected peaks, information such as covariance structure required by kNN and the nonparametric kernel method is also computed using the entire dataset, which further biases the classification. This is what is done by default in packages such as SAS, and as we will see it has dramatic consequences. Finally, we ran all the methods with various numbers of selected peaks, keeping in mind that results obtained with a smaller number of peaks are likely to be more robust. We present results for four and thirteen peaks selected using the B/W ratio.

Table 2 shows leave-one-out cross-validation results for the top four selected peaks. Here and in future tables, group A corresponds to lung cancer samples and group B to samples from healthy individuals. We see that, for example, the quadratic discrimination method misclassifies four samples from group A as belonging to group B, and one sample from B as belonging to A, the overall error rate is thus 12%. The results are (with one exception) identical for preselected and reselected peaks, since these four peaks consistently rank as the top four using the B/W criterion (although they do change order sometimes when reselecting during cross-validation). The only exception is kNN which classifies one additional sample from B correctly in the case of preselected peaks. The reason this occurs (despite using the same features) is that kNN uses covariance information in calculating the Mahalanobis distance, which in this case is based on the entire dataset, not just a subset of forty samples. Tables 3 and 4 show the behavior of the meth-

Table 2. Leave-one-out cross-calidation results using the top four peaks selected from a total of 229

Method	From	To		Error	Method	From	To		Error
		A	B	rate (%)			A	B	rate (%)
Lin. Discrim.	A	22	2	10	Quad. Discrim.	A	20	4	12
	B	2	15			B	1	16	
Nonpar. Kernel	A	20	4	12	kNN	A	21	3	15
	B	1	16			B	3	14	
Linear SVM	A	22	2	15					
	B	4	13						

Table 3. Leave-one-out cross-validation results using the top thirteen peaks, with reselection in every iteration

Method	From	To		Error	Method	From	To		Error
		A	B	rate (%)			A	B	rate (%)
Lin. Discrim.	A	18	6	27	Quad. Discrim.	A	20	4	34
	B	5	12			B	10	7	
Nonpar. Kernel	A	22	2	29	kNN	A	18	6	27
	B	10	7			B	5	12	
Linear SVM	A	23	1	2					
	B	0	17						

Table 4. Leave-one-out cross-validation results using the top thirteen peaks selected using the entire dataset

Method	From	To		Error	Method	From	To		Error
		A	B	rate (%)			A	B	rate (%)
Lin. Discrim.	A	22	2	12	Quad. Discrim.	A	22	2	24
	B	3	14			B	8	9	
Nonpar. Kernel	A	24	0	7	kNN	A	24	0	0
	B	3	14			B	0	17	
Linear SVM	A	23	1	2					
	B	0	17						

ods for thirteen peaks; in one case the peaks were selected using the entire dataset, and in the other we reselected them in every cross-validation step using only the training set.

Fisher's linear discrimination assumes homogeneous covariance matrices for the two groups, and uses a pooled sample covariance matrix for the construction of the discriminant functions. For the current data set we have observed that the covariance matrices are not homogeneous. This explains the poor performance of this method in general. kNN is a local method which can produce poor results for noisy data in general. However, both fared well when only four peaks were used. It is interesting to note in this case that the four selected peaks were the same whether the F-statistic (B/W ratio) is calculated using the pooled sample variance or using the weighted form of the sample variance (to account for the heterogeneity of the variances). Hence the performance of these methods is comparable to those that account for the difference in the covariance matrices.

The quadratic discrimination and nonparametric kernel methods account for the heterogeneity of the covariance matrices by using different covariance matrices for different groups in the construction of the discriminant functions.

However, the small sample sizes (only 17 samples for group B) in the current data set cause these covariance matrices to become nearly singular when a large number of peaks are used. This explains the poor performance of these methods when thirteen peaks are used for classification. It is interesting to note however that the SVM is apparently not affected by these drawbacks and fared well in all these situations. It seems to be significantly more robust in terms of performance with varying numbers of peaks. We also see that, as expected, the error rates are generally higher when the peak selection is based only on information in the training set. Again, the SVM is a notable exception.

To assess the significance of the results presented in the previous section we ran again our methods on the same data but with randomized group assignments. That is, while the peak profile vectors were kept the same, the assignment to groups A and B were randomized. The ratio of the numbers of samples in the two groups was kept at 24/17, as in the original data set. The purpose of this is to attempt to get insight into how significant the classification results are by comparing them to results on what essentially is random data of a similar nature. The corresponding results are presented in Table 5. We show the average leave-one-out error rate as well as the best and worst errors. These results provide a benchmark to the results in Table 2. They illustrate that the latter seem to be significant and not artifacts of the choices made in the preprocessing stages.

The masses of the four most significant peaks that we used to distinguish between cancerous samples and healthy samples (for the results reported in Table 2) are tabulated in Table 6. The first, third, and fourth of these peaks are down-regulated in lung cancer, while the second is up-regulated; indeed, the second peak appears

Table 5. Error rates for cross-calidation runs on ten datasets generated from original data by randomizing the group assignments. Four peaks with reselection were used

Method	Linear Discr.	Quadr. Discr.	Nonpar. Kernel	kNN	SVM
Average (min, max) error (%)	54 (37, 86)	48 (22, 80)	48 (27, 83)	52 (29, 73)	49 (24, 76)

Table 6. The masses of the four proteins used to classify lung cancer in this paper, and their matches with proteins from a database for lung cancer protein expression. Two close matches were found for the second protein, and no match was found for the fourth protein

Suggested peak	Change in cancer	Reported peak	Description of protein
28 088.9	down	27 774	Stratifin
		11 858	Cellular retinoic acid binding protein
11 695.2	up	11 521	Protein kinase C inhibitor
9 481.7	down	9 200	Cytochrome *c* oxidase
8 712.4	down	NA	NA

only in one of the healthy samples, at a low intensity. We compared these protein masses with a database of protein expression in lung cancer [6] to find matching proteins with close masses. Three of the four proteins we have employed for classification have closely matching proteins, as shown in Table 6. Of these, the stratifin and cellular retinoic acid binding protein are differentially expressed in small cell, adenocarcinoma, and squamous lung tumors. These matches should be considered tentative at this stage until the proteins in our study are sequenced and identified. Identifying these proteins or protein fragments and understanding their role in lung cancer would provide credence to the data processing techniques and classification algorithms that we have employed.

4 Concluding remarks

We have discussed in detail five of the steps in preprocessing the mass spectral data for biomarker discovery, as well as our criterion for choosing a small set of peaks for classifying the samples. Cross-validation studies with four proteins with the highest B/W ratio yielded misclassification rates in the 10–15% range for all the classification methods. Three of these proteins or protein fragments are down-regulated and one up-regulated in lung cancer. When cross-validation studies are performed, care must be taken to ensure that the test set does not influence the choice of the peaks used in the classification. Unfortunately, statistical packages do not guarantee this when default settings in their methods are employed. Misclassification rates are generally lower when both the training and test sets are used to select the peaks used in classification, than when only the training set is used. This expectation was dramatically borne out when thirteen peaks were used in cross-validation studies. However, when only four peaks were used to classify, the two approaches led to almost identical results; this was due to the fact that in these cases, the identity of the four peaks did not change when the training set changed. We take this as another strong indication that these four peaks are indeed good candidates for biomarkers.

Experiments with the samples randomly assigned to the two classes confirmed that misclassification rates were higher in such cases than those observed with the true data. We strongly believe that stringent validation experiments of this or similar nature should always be performed when dealing with high-dimensional data. The group covariance matrices were heterogeneous for the lung cancer and healthy groups, leading to poor expected performances for linear discrimination and nearest neighbor classification. Quadratic discrimination and nonparametric kernel discrimination methods account for the heterogeneity in the data, but suffered from the near singularity of the covariance matrices when the number of peaks used to classify was large relative to the number of samples in a group. The support vector machine exhibited robust performance when the number of peaks was varied from four to thirteen, and when the peaks were selected from the training set alone. Three of the four proteins we used in classifying lung cancer have closely matching proteins in a protein expression database for lung cancer.

We caution that the number of samples included in this study was small, and that our conclusions should be considered tentative for this reason. Data from additional samples, increased experience with the performance of various preprocessing techniques, and more insight into both the experimental setup and the underlying biology of the disease will strengthen our methodology in the future.

A part of the third author's work was done while he was visiting the Computer Science Research Institute at Sandia National Labs, Albuquerque NM.

Received October 16, 2002

5 References

[1] Patterson, S. D., Aebersold, R., Goodlett, D. R., in: Pennington, S., Dunn, M. J. (Eds.) *Mass Spectrometry based Methods for Protein Identification and Phosphorylation Site Analysis.* in: *Proteomics: From Protein Sequence to Function*, BIOS Scientific Publishers, Oxford 2001, pp. 87–130.

[2] Dudoit, S., Fridlyand, J., Speed, T. P., *J. Am. Stat. Assoc.* 2002, *97*, 77–87.

[3] Cristianini, N., Shawe-Taylor, J., *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.

[4] Hastie, T., Tbishirani, R., Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, New York 2001.

[5] Joachims, T., in: Schölkopf, B., Burges, C. T. C., Smola, A. J. (Eds.), in: *Advances in Kernel Methods – Support Vector Learning*, MIT Press, Cambridge, MA 1999, pp. 169–184.

[6] Oh, J. M. C., Brichory, F., Puravs, E., Kuick, R. *et al.*, *Proteomics* 2001, *1*, 1303–1319.