Author(s): Marc Zapatka[1]

[1] Division of Theoretical Bioinformatics, German Cancer Research Center, Germany

Mass spectrometry (MS) profiling of serum and other body fluids as well as tissue samples is utilized for high-throughput analysis of complex samples and allows discovery of candidate biomarkers for diseases such as cancer. Despite the enthusiastic reports of early studies, MS-based biomarker discovery is fraught with controversy (Diamandis 2003, Garber 2004, Diamandis 2006) and powerful diagnostic performance of proteomic profiling (Petricoin et al. 2002) has shown to be related to erroneous study design (Baggerly et al. 2004). Namely, preanalytical variabilities in sample handling and processing cause substantial changes of MS peptide profiles (Baumann et al. 2005, Banks et al. 2005, Findeisen et al. 2005). These in-vitro changes will markedly affect profiling experiments overriding disease-related peptide patterns, and thus may even completely abolish meaningful data interpretation (Karsan et al. 2005). Rigorous standardization from the first step of sample collection to subsequent sample processing is necessary (Baumann et al. 2005), but this can hardly be integrated into routine laboratory testing.

Datasets from proteomic MS are like microarrays typically characterized by a very high-dimensional input space (several thousand variables) but relatively small numbers of samples (a few hundred at best). This contradicts the use of classical statistical techniques such as linear discriminants or neural networks, unless the dimensionality of the problem is reduced using some intelligent feature reduction algorithm. Furthermore, some data pre-processing steps are necessary to utilize MS data reasonably.

For a study using time of flight mass spectrometry at least the important points proposed as minimal requirements should be considered (Diamandis 2006):

1. Examine the effect of sample collection and storage conditions on the data.
2. Examine other variables that could potentially bias the data, either biologically or bioinformatically (Ransohoff 2005).
3. Every derived diagnostic algorithm should be tested without retraining sets, on an equally large set of samples from different institutions and/or countries to verify its robustness.
4. It is essential to positively identify the discriminatory peaks and link them to disease pathobiology (do the identified discriminatory peaks make biological sense?). If a peak cannot be positively identified, it should not be considered as a useful marker. But it has to be considered that information can also be hidden in proteases which generate surrogate markers through processing of endogenious proteins (e.g. fibrinogen, complement system, …) (Villanueva et al. 2006).

5. Correlate peak intensities with tumor burden (e.g., stage).
6. In serum profiling, provide clues for peak intensity decreases in patients with cancer versus normal subjects.
7. Validation studies of previously reported data, even if negative, should be published, as done by Karsan et al. (2005), so that this apparently highly promising technology is put into perspective.
8. Good experimental design and laboratory practices should be exercised in executing the study. This should include (a) randomized experimental and control samples prior to analysis, it is essential that test sets are analyzed in a blinded fashion, (c) provide sufficient experimental details so that others could reproduce the study, (d) define how frequently the mass spectrometer has been calibrated, and (e) define how experimental consistency and reproducibility throughout the experiment has been monitored and controlled.

Depending on the evaluated sample a reduction of probe complexity (through chromatography, fractionation, ...) has the advantage to make low abundant proteins detectable otherwise not taken into account. For instance fractionation of serum proved to be helpful for detecting prostate-specific antigen (Solassol et al. 2005).

Having incorporated all minimal requirements for profiling using mass spectrometry data the bioinformatic preprocessing of the spectra is the most important step (Coombes et al. 2007). Due to the complex data acquisition procedure several steps should be used to compensate for some technical variations that might cover the interesting biological differences. Inadequate or incorrect preprocessing methods can result in data that exhibits substantial biases. Therefore a minimum sequence of preprocessing steps is recommended to remove at least some systematic bias and variability:

- Mass calibration transforms the data from the time of flight into the mass over charge (m/z) domain. The calibration function is a quadratic function constructed using the time measurements of a mixture of 4 to 6 peptides of known size.
- Baseline subtraction is a crucial step. Because of the complex underlying physical processes the displacement of the baseline can not easily be calculated. As a reasonable approximation often the local minimum can be used.

- Normalization has to be applied to remove variability due to differences in the total amount of protein detected. A widely accepted approach is the normalization on the total ion count. This is of course only applicable if in case of an ideal measurement similar intensities are expected for all samples.
- The realignment of the spectra is important to make the spectra comparable. For example the alignment can be performed using the algorithm of Jeffries based on corresponding peaks (Jeffries 2005)
- Peak detection to identify m/z values of peak positions can be performed on each spectrum. If applied to the mean spectrum a good alignment of the spectra is needed beforehand (Morris at al. 2005).
- Peak quantification can incorporate calculation of peak areas or heights and signal to noise ratio.
- Peak matching across the samples is required if the realignment of the spectra or the peak detections results in unsure correspondence peaks

A visual inspection of the data before and after preprocessing is essential and can help to reveal serious error in the data generation and data preprocessing. For the quality control an image plot of the mass spectra (gelview: spectra in rows, m/z in columns, intensity color coded) is a helpful visualization. Mass shifts, probe degradation, intensity variations, changes in the spectrometer settings (e.g. chamber cleaning, changes in laser intensity, day to day variations, ...) and many more can easily be identified and spectra of suspicious quality should be removed.

After preprocessing several machine learning algorithms can be applied to classify patients according to the spectra and/or identify biomarkers. For example support vector machines (SVM) and artificial neural networks (ANN) have shown good learning performance based on these high dimensional data sets.

Conclusion

General experience shows the importance of a proper study design and data pre-processing. Due to the complex structure of mass spectrometry data these are even more important. The machine learning algorithms applied afterwards are capable to reveal differences in the data. Whether these are biological meaningful or artifacts depends mainly upon the steps done before machine learning.

Acknowledgements

The author thanks Peter Findeisen and Franz Bosch for helpful discussions.

Recommended literature discussing these points:

1. Baggerly, K. A.; Morris, J. S.; Wang, J.; Gold, D.; Xiao, L. & Coombes, K. R. A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. Proteomics, 2003 , 3 , 1667-1672
2. Baggerly, K. A.; Morris, J. S. & Coombes, K. R. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics, 2004 , 20 , 777-85
3. Banks RE, Stanley AJ, Cairns DA, Barrett JH, Clarke P, Thompson D, Selby PJ. Influences of blood sample processing on low-molecular-weight proteome identified by surface-enhanced laser desorption/ionization mass spectrometry. Clin Chem 2005;51:1637-49.
4. Baumann S, Ceglarek U, Fiedler GM, Lembcke J, Leichtle A, Thiery J. Standardized approach to proteome profiling of human serum based on magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Clin Chem 2005;51:973-80.
5. Diamandis EP. Point: Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? Clin Chem 2003;49:1272-5.
6. Diamandis, E. P. Serum proteomic profiling by matrix-assisted laser desorption-ionization time-of-flight mass spectrometry for cancer diagnosis: next steps. Cancer Res, 2006 , 66 , 5540-5541
7. Findeisen P, Sismanidis D, Riedl M, Costina V, Neumaier M. Preanalytical impact of sample handling on proteome profiling experiments with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Clin Chem 2005;51:2409-11.
8. Garber K. Debate rages over proteomic patterns. J Natl Cancer Inst 2004;96:816-8.
9. Jeffries, N. Algorithms for alignment of mass spectrometry proteomic data. Bioinformatics, 2005 , 21 , 3066-3073
10. Karsan A, Eigl BJ, Flibotte S, Gelmon K, Switzer P, Hassell P et al. Analytical and preanalytical biases in serum proteomic pattern analysis for breast cancer diagnosis. Clin Chem 2005;51:1525-8.
11. Kevin R. Coombes, Keith A. Baggerly, and Jeffrey S. Morris. "Pre-Processing Mass Spectrometry Data" Fundamentals of Data Mining in Genomics and Proteomics. Ed. M

Author(s): Marc Zapatka[1]

[1] Division of Theoretical Bioinformatics, German Cancer Research Center, Germany

Dubitzky, M Granzow, and D Berrar. Boston: Kluwer, 2007. 79-99.

12. Morris, J. S.; Coombes, K. R.; Koomen, J.; Baggerly, K. A. & Kobayashi, R. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. Bioinformatics, 2005
13. Petricoin, E. F.; Ardekani, A. M.; Hitt, B. A.; Levine, P. J.; Fusaro, V. A.; Steinberg, S. M.; Mills, G. B.; Simone, C.; Fishman, D. A.; Kohn, E. C. & Liotta, L. A. Use of proteomic patterns in serum to identify ovarian cancer. Lancet, 2002 , 359 , 572-7
14. Ransohoff, D. F. Bias as a threat to the validity of cancer molecular-maker research Nature Reviews Cancer, 2005 , 5 , 142-149
15. Solassol, J.; Marin, P.; Demettre, E.; Rouanet, P.; Bockaert, J.; Maudelonde, T. & Mangé, A. Proteomic detection of prostate-specific antigen using a serum fractionation procedure: potential implication for new low-abundance cancer biomarkers detection. Anal Biochem, 2005 , 338 , 26-31
16. Sorace, J. M. & Zhan, M. A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics, 2003 , 4 , 24
17. Villanueva, J.; Shaffer, D. R.; Philip, J.; Chaparro, C. A.; Erdjument-Bromage, H.; Olshen, A. B.; Fleisher, M.; Lilja, H.; Brogi, E.; Boyd, J.; Sanchez-Carbayo, M.; Holland, E. C.; Cordon-Cardo, C.; Scher, H. I. & Tempst, P. Differential exoprotease activities confer tumor-specific serum peptidome patterns. J Clin Invest, 2006 , 116 , 271-284

Version	Tracking of changes	Name	Date