Beating the Noise: New Statistical Methods for Detecting Signals in MALDI-TOF Spectra Below Noise Level

Tim O.F. Conrad[1], Alexander Leichtle[2], Andre Hagehülsmann[3], Elmar Diederichs[1], Sven Baumann[2], Joachim Thiery[2], and Christof Schütte[1]

[1] Free University Berlin, Department of Mathematics, Germany
[2] Institute of Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics, University Hospital Leipzig, Germany
[3] Microsoft Research, Cambridge, UK

Abstract

Background: The computer-assisted detection of small molecules by mass spectrometry in biological samples provides a snapshot of thousands of peptides, protein fragments and proteins in biological samples. This new analytical technology has the potential to identify disease associated proteomic patterns in blood serum. However, the presently available bioinformatic tools are not sensitive enough to identify clinically important low abundant proteins as hormons or tumor markers with only low blood concentrations.

Aim: Find, analyze and compare serum proteom patterns in groups of human subjects having different properties such as disease status with a new workflow to enhance sensitivity and specificity.

Problems: Mass data acquired from high-throughput platforms frequently are blurred and noisy. This complicates the reliable identification of peaks in general and very small peaks even below noise level in particular.[1] However, this statement is only valid for single or few spectra. If the algorithm has access to a large number of spectra (e.g. $N > 1000$), new possibilities arise, one of such being a statistical approach.

Approach: Apply signal preprocessing steps followed by statistical analyses of the blurred data and the region below the typical noise threshold to identify signals usually hidden below this "barrier".

Results: A new analysis workflow has been developed that is able to accurately identify, analyze and determine peaks and their parameters even below noise level which other tools can not detect. A Comparison to commercial software[2] has clearly proven this gain in sensitivity. These additional peaks can be used in subsequent steps to build better peak patterns for proteomic pattern analysis. We belive that this new approach will foster identification of new biomarkers having not been detectable by most algorithms currently available.

[1] It is due to the fact that it is not possible to distinguish noise from signals if these two components fully overlay.

[2] ClinProTools 2.0, Bruker Daltronics: manufacturer of mass spectrometry instruments and accessories.

M.R. Berthold, R. Glen, and I. Fischer (Eds.): CompLife 2006, LNBI 4216, pp. 119–128, 2006.

1 Introduction

With the advantages of proteomic technologies it became possible to assess small and low abundant molecules in biological fluids. These peptide or protein patterns (often referred to as "fingerprints") can indicate status or progress of a specific disease. Several studies have shown the potential of such patterns for early detection of different types of cancer [1, 2].

After reducing sample complexity by biochemical separation techniques [3], the protein fingerprinting of biological samples consists of three main steps: first, generating of mass spectra, that is presenting the complete spectrum of peptides and proteins in the sample. Second, features (usually peaks) in the resulting spectra are detected that can be used to discriminate between groups of individuals with various phenotypes (eg. gender, age or disease). Third, the features have to be tested in independent studies to confirm and to detect the underlying molecules.

Mass spectrometry is a high-throughput profiling technique able to fulfill the first part of the task[3]. The second step is usually done by machine learning algorithms and statistical approaches which are used to analyze the data obtained from mass spectrometry and to detect phenotype specific patterns.

Todays mass spectrometry based protein fingerprinting techniques rely on the analysis of spectra from complex biological protein mixtures (e.g. serum) obtained from high-throughput platforms in clinical settings. An unsolved bioinformatic problem is the highly sensitive detection of peaks (potential features) within this "crossfire of influences". Embedded in systematic and random noise introduced during acquisition of data it is very difficult and questionable to detect peaks and correctly determine their parameters, such as location, height or width. Most methods simply ignore any signals below an estimated noise threshold and potentially lose many signals hidden in this region.

In this study, we propose a new statistical driven approach that allows to analyze noise and identify signals below the commonly used signal-to-noise threshold[4]. This is done by sophisticated preprocessing steps and statistical analysis of all potential signals in a large number of spectra by identifying even smallest features. Compared to commercial software[5] (see Tab.1) our approach - in the cases tested - is about 20000 times more sensitive without loss of specificity. Additionally peaks identified can be used in subsequent steps to build better patterns for proteomic fingerprinting analysis. We belive that this will foster identification of new biomarkers having not been detectable by most algorithms currently available.

[3] In this study we used the "Matrix-assisted Laser Desorption/Ionization - Time-Of-Flight - Mass Spectrometry" (MALDI-TOF MS). For a recent review and a good introdutcion to this topic see e.g.[4] and references therein.

[4] This method only regards peaks if their height is above a certain value determined by a noise-estimation step. A common setting for the minimum peak height is three times the estimated noise level.

[5] ClinProTools 2.0, Bruker Daltronics: manufacturer of mass spectrometry instruments and accessories.

2 Material and Methods

2.1 Spectra Preprocessing

The raw spectrum acquired by a TOF mass-spectrometer (see Fig.1 left) is a mixture of the real signal and noise. The noise itself consists of a low-frequency baseline and high-frequency chemical and random noise. Preprocessing of TOF spectra includes suppression of noise and enhancing of the real signal - it is therefore a crucial step prior to the actual signal extraction. The next sections describe steps performed to prepare the raw signal enabling subsequent reliable peak detection and analyses.

Baseline Correction. The baseline is an exponential like offset dependent on the m/z value (mass-to-charge; x-value). It is mainly caused by clusters of matrix components and small molecular fragments originating from degradation processes, desorption and collisions in the acceleration phase. A baseline correction is performed to remove this rather low-frequency noise from the spectrum.

Following [5–7] we use a morphological TopHat filter. Mathematical morphology is the analysis of spatial structures and is used here to eliminate certain spatial structures within the signal, in our case the baseline. Its simplicity and rapidity make it extremely handy for application to large amounts of data. Fig. 1 illustrates this method. Note that this technique does not produce negative intensity values as many other popular methods depending on polynomial fitting [8], piecewise linear regression [9] or convex hulls [10] do.

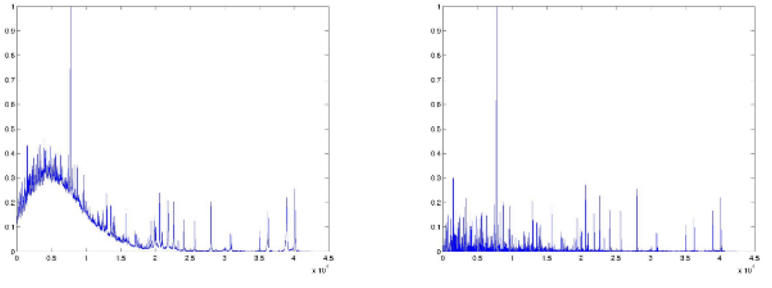

Fig. 1. Application of the TopHat Filter: the opening is substracted from the raw signal (left) yielding the filtered version (right)

Smoothing. Smoothing or denoising the raw signal X tries to separate the Gaussian contribution from the undisturbed signal S and generally yields better results in subsequent steps of the analysis workflow, since some general assumptions about smoothness can be taken. We define the Gaussian distributed component of the signal as noise. Consider the problem of denoising a raw signal $X \in \mathbb{R}$ having additive noise n with zero mean:

$$x_i = s_i + n_i \quad i = 1..N, \quad n \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

Since the noise described above occurs on much faster time scales as the signal, we use a multiresolution analysis on X [11] based on a time-invariant discrete orthogonal wavelet transformation [12].

Opposed to other denoising algorithms, such as moving average or low-pass filter (e.g. Savitzky-Golay), this approach utilizes the multi-scale nature of the signal and therefore has better energy conservation properties, that is, the amplitude of the signal decreases less through denoising.

Normalization. Inter-spectrum normalization is the process of removing systematic variations between spectra. Many different techniques exist such as "Inverse Normalization" [13] or "Logarithmic Normalization" [14]. Our implementation follows the idea of the most frequently used method which is global normalization with respect to the average *total ion current* (TIC[6])[7] [16, 17] with an important extension: from the set of spectra to be normalized all TIC values are computed, outliers removed and the remaining highest value (instead of the average) is used for the actual computation.

2.2 Peak Identification in Single Spectra

Most peak detection algorithms have in common that they use threshold driven detection techniques. That is, a peak will only be regarded if it is higher than a predetermined signal-to-noise threshold depending on the calculated noise level (see e.g. [18]).

As shown exemplarily in Fig. 2, by assuming a noise level of 50[8] and using a signal-to-noise ratio of 3[9] about 85% of the 1332 potential peaks in this particular spectrum would be discarded and their assigned information lost. Although most of these peaks essentially *are* noise, some might carry important information. This means, that these artificially introduced "barriers" would prevent detection of small signals in a very early pre-processing stage.

The subsequent sections describe our new approaches to overcome this signal-to-noise barrier, that means increasing sensitivity without decreasing specificity.

Detection of Candidate Peaks. The initial peak detection simply determines the location of potential peaks, a process often referred to as "seeding". Utilizing the properties of the TopHat filter (see Sec. 2.1) it is sufficient to detect interception points of the curve with the X-axis. These points define start- and end points of potential peaks and are stored in a database for further analyses.

The advantage of this approach is that even smallest peaks are considered for consecutive steps. However, a deliberate validation algorithm must be applied to this set of candidate peaks to distinguish real peaks from noise and detect and deconvolute overlapping peaks.

[6] The TIC is the sum of the area of all peaks in a spectra.

[7] The normalization by the ratio $R = \frac{\text{"TIC of spectrum"}}{\text{"Average TIC of all spectra"}}$ is reported to be superior to other methods tested [15, 6].

[8] Different noise-estimators compute values ranging from 50 to 150.

[9] A commonly used value to get reliable results.

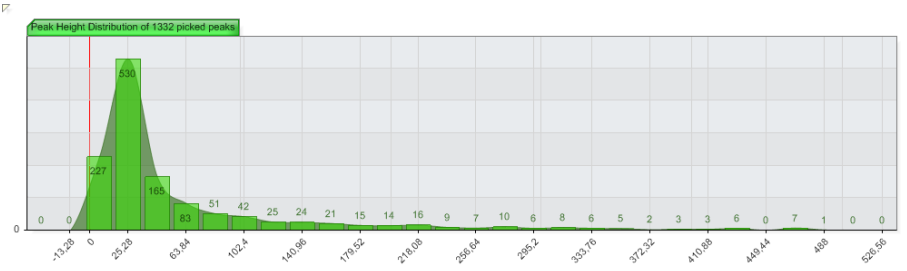

Fig. 2. Histogram of peak heights in a randomly chosen spectrum of the study described in [3]. Only peaks smaller than 500 are displayed.

Analyzing Candidate Peaks. In mass spectra of complex protein mixtures (such as serum) most of the peaks detected are broadened and/or highly convoluted. This results for example from molecular fragments having very similar masses and thus partly overlay, from poor machine resolution, or different isotopic forms of the same molecule. Therefore, a successful peak detection algorithm needs to deconvolute those peaks. A widely accepted method assumes a "blurred peak" to be a mixture of Gaussians and tries to resolve it back into its original components. A commonly used technique is "Maximum Entropy" that has been originally developed for clarifying blurred images (see [19]). Based on this idea we have developed an approach to separate and evaluate the assumed mixture of Gaussians. The key steps for each candidate peak found are as follows:

1. Determine number of Gaussian components by density estimation using the "Greedy Expectation Maximization" algorithm [20]. This algorithm has been shown to have a very good performance even on large mixtures often found in peaks at higher masses (> 3000Da). (For a comparative study see [21].)
2. To account for isotopic forms a de-isotoping step is carried out by fitting a mass dependent pre-calculated model (see [22] for details) if more than one Gaussian component is found. If successfull, the peaks involved are tagged as belonging to the same molecule(-fragment). If the quality of the fit is too poor the peak is split according to the number of Gaussians found and for each of the new parts step 1 is performed again.
3. Determine and store the parameters (height, width, center, area, shape quality[10]) of this peak.

2.3 Peak Assignment Across Spectra

In order to identify a particular peak across spectra a list of so called "masterpeaks" is maintained per spectra group (e.g. male or healthy). A masterpeak

[10] This is achieved by geometrical hashing [23]. The hashing algorithm returns a discrete value $c \in \{0, 1 \ldots 5\}$, indicating the class the hashing algorithm has assigned to this shape. $c = 0$ means "noise" and $c = 1..5$ peak, where $c = 5$ is assigned if the peak looks "perfect". The categories are trained a priori.

comprises peaks having similar properties (m/z value, height, shape, etc.) across spectra in this group. From the comprehensive distribution of property values the "real" values for a masterpeak will be derived in later stages.

Preprocessing: Finding Candidate Masterpeaks. To build a set of potential masterpeaks the following two steps are carried out:

1. Center and width of every peak identifyed in step 2.2 in a group of spectra under scrutiny (e.g. healthy or female) are stored in a temporary table, ordered by their center.
2. Candidate clusters of these peaks are built with respect to the centers and the width of these peaks. Peaks belong to the same cluster if they overlap in at least one point. Since we can simply "march" through this ordered set of peaks with a linear number of comparisons, the complexity is $\mathcal{O}(n)$. Alternatively, complete-linkage hierarchical clustering could be performed [24] to build the clusters which is computationally more expensive.

Masterpeak Property Determination. We now have a set of candidate clusters often containing more than one "real" group of similar peaks. This step is going to resolve these groups by a Bayesian Clustering approach. From the clusters found in this step all properties such as center, height or width are derived. From the law of large numbers we know that the average values will converge to the real values. The probabilistic object that underlies this approach is a distribution on partitions of integers[11] known as the (weighted) *Chinese restaurant process* (CRP) [27–29].

The CRP can be best described by a process where N customers sit down in a Chinese restaurant with an infinite number of tables $C_1, C_2, \ldots$ and each table has an infinite number of seats. Suppose customers arrive sequentially. Per definition the first guest sits down at the first table. The $n+1$th subsequent customer x_{n+1} sits at a table drawn from the following distribution:

$$\begin{aligned} p(\text{occupied table } i \mid \text{previous customers}) &= \tfrac{c_i}{\alpha+n} \cdot R \cdot s(x_{n+1}|x_j, j \in C_i) \\ p(\text{new table} \mid \text{previous customers}) &= \tfrac{\alpha}{\alpha+n} \end{aligned}$$

where c_i is the number of customers already sitting at table C_i, R is a rescaling factor and $\alpha > 0$ is a parameter defining the CRP. Obviously, the choice of the similarity function $s(\cdot)$ is crucial and is explained in the following paragraphs.

Let $C_i(A)$ be the average value of a property A of a set of peaks "sitting" at table i. (For example, $C_2(center)$ would be the average center of peaks at the second table.) Let $x_j(A)$ be the value of property A of peak j. $s(\cdot)$ has the following properties:

1. The distance of the center of a peak to the average center of an existing group of peaks can not be further away than 2 Da.
2. $s(\cdot)$ is the likelihood of x_j belonging to "table" C_i depending on how similar the properties of x_j are to the peaks already at "table" C_i.

[11] Interestingly, the partition after N steps has the same structure as draws from a Dirichlet process [25, 26].

This results in:

$$s(x) = \begin{cases} 0 & \text{if } |C_i(\text{center}) - x_j(\text{center})| > 2 \\ \sum_{A \in \mathcal{PP}} k_{i,A}(x(A)) & \text{otherwise} \end{cases}$$

$\mathcal{PP}$ being the set of peak properties and $k_{i,A}(\cdot)$ is constructed as follows:

1. Kernel Density Estimation (KDE)[12] [30] is performed on $x_j(A), j \in C_i$.
2. This resulting density is transformed through interpolation to the continuous function $k_{i,A}(\cdot)$.

The resulting sub-clusters (tables) are processed further in order to merge together similar groups and stored in the database.

3 Results and Discussion

Although having been designed for large amounts of spectra, we conducted first experiments with a small set of samples. To obtain a first "proof-of-principle" and to test the overall performance of our workflow we spiked a subset of human serum samples[13] with a peptide mix[14]. We split 16 different samples into five groups each. Before sample pretreatment and measurement each of the groups was spiked with one of the following concentrations: 121.21nMol/L, 0.76nMol/L, 0.30nMol/L, 3.03pMol/L, 0.075pMol/L, resulting in 320 spectra (64 for each concentration group due to 4-fold spotting).

We then processed each resulting raw spectrum as described above. For each of the five resulting concentration groups we evaluated the masterpeaks found. Subsequently we validated whether masterpeaks originating from the spiked peptides were identified by the algorithms and checked the deviation of the determined centers to the postulated ones. Note that at this stage no analysis of other detected peaks has been performed. Table 1 summarizes the findings:

1. The algorithms successfully detect peaks even for very small concentrations at pMol/L level. This is exemplarily shown for the hormones Angiotensin, Bombesin and ACTH clip 18-39 which can be detected in a very low and biologically relevant concentration range ($\sim$3 pMol/L). Peaks for Angiotensin and Bombesin are not detected by commercial software[15]. Therefore, in these examples, our algorithm is at least 20000 times more sensitive than a commercial algorithm using a signal-to-noise threshold.

[12] Following the Parzen Window approach with Gaussian Kernel.

[13] The protocol used for preprocessing and (magnetic bead) fractionation has been described in [3].

[14] Protein calibration standard mix (Part No.: 206355 & 206196 purchased from Bruker Daltronics, Leipzig, Germany).

[15] ClinProTools 2.0, Bruker Daltronics. Parameters used: Signal-to-Noise Level: 3, Peak Detection Algorithm: Centroid.

Table 1. Results of the spiking experiments (see text for explanation). Note that no calibration has been performed for the Proteomics.NET platform.

Substance	Platform[3]	Theor. Center	Center found (Concentration of spiked peptide mix as stated below)				
			121.21nMol/L	0.76nMol/L	0.30nMol/L	3.03pMol/L	0.075pMol/L
Angiotensin II	P.NET	1047.20	1047.1	-[1]	-[1]	-[1]	-[1]
Angiotensin II	CPT	1047.20	1046.9	-[2]	-[2]	-[2]	-[2]
Angiotensin I	P.NET	1297.51	1297.6	1298.0	1299.3	1299.2	-[1]
Angiotensin I	CPT	1297.51	1297.2	-[2]	-[2]	-[2]	-[2]
Bombesin	P.NET	1620.88	1620.9	1618.1	1617.2	1617.2	-[1]
Bombesin	CPT	1620.88	1620.6	-[2]	-[2]	-[2]	-[2]
ACTH clip 18-39	P.NET	2466.73	2466.8	2465.8	2465.8	2466.2	-[1]
ACTH clip 18-39	CPT	2466.73	2466.2	-[2]	2466.1	2465.9	-[2]
Somatostatin 28	P.NET	3149.61	3149.5	-[1]	-[1]	-[1]	-[1]
Somatostatin 28	CPT	3149.61	3149.0	-[2]	-[2]	-[2]	-[2]
Insulin	P.NET	5734.56	5734.3	-[1]	-[1]	-[1]	-[1]
Insulin	CPT	5734.56	5734.2	-[2]	-[2]	-[2]	-[2]

[1]: No significant masterpeak in this range at this concentration found, [2]: No significant peaks in this range at this concentration found, [3]: CPT: Bruker ClinprotTools, P.NET: Our Proposed Platform: Proteomics.NET.

2. The proposed methods are able to detect peak centers accurately since shifts and noise in the spectra largely are cancelled out after averaging by the cluster partition process described above[16]. The centers found are determined more precise than the commercial software does.

4 Conclusion

We have presented new methods for preprocessing MALDI-TOF MS spectra and detecting and evaluating peaks in these spectra. These steps lead to an enhanced sensitivity in the overall peak detection process. In a proof-of-concept setting, our results show that our algorithms with the statistical driven approach are able to detect spiked peptides in serum spectra in a concentration as low as 3.03pMol/L. We believe that this new approach will promote the sensitivity of proteome pattern diagnostics in laboratory medicine.

Acknowledgements

The study presented here is supported by a grant from Microsoft Research Ltd. to C. Schütte, Berlin and J. Thiery, Leipzig. We are grateful to Dr. Markus

[16] Note that the spectra have not been calibrated and therefore systematic shifts can occure. This is due to the fact that none of the calibration methods from the literature having been applied to this highly complex dataset has shown stable and reliable results. We belive that new ideas and approached have to be developed to successfully tackle this problem.

Kostrzewa (Bruker Daltonics, Leipzig, Germany) for technical support and helpful discussions. This work was also supported by a grant from the Sächsische Aufbaubank (SAB) to J. Thiery, Leipzig.

References

1. K. R. Kozak, F. Su, J. P. Whitelegge, K. Faull, S. Reddy, and R. Farias-Eisner. Characterization of serum biomarkers for detection of early stage ovarian cancer. *Proteomics*, 5(17):4589–4596, Nov 2005.
2. S. Becker, L. H. Cazares, P. Watson, H. Lynch, O. J. Semmes, R. R. Drake, and C. Laronga. Surfaced-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) differentiation of serum protein profiles of BRCA-1 and sporadic breast cancer. *Ann Surg Oncol*, 11(10):907–914, Oct 2004.
3. S. Baumann, U. Ceglarek, G. M. Fiedler, J. Lembcke, A. Leichtle, and J. Thiery. 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*, 51(6):973–980, Jun 2005.
4. Glen L Hortin. The MALDI TOF Mass Spectrometric View of the Plasma Proteome and Peptidome. *Clin Chem*, Apr 2006.
5. E. J. Breen, F. G. Hopwood, K. L. Williams, and M. R. Wilkins. Automatic poisson peak harvesting for high throughput protein identification. *Electrophoresis*, 21(11):2243 – 2251, 2000.
6. A. C. Sauve and T. P. Speed. Normalization, baseline correction and alignment of high-throughput mass spectrometry data. In *Procedings Gensips 2004*, 2004.
7. C. Gröpl, A. Hildebrandt, O. Kohlbacher, E. Lange, S. Lövenich, and M. Sturm. OpenMS - Software for Mass Spectrometry. Poster presented at the MBI Workshop on Computational Proteomics and Mass Spectrometry 2005, Ohio State University, 2005.
8. V. Mazet, D. Brie, and J. Idier. Baseline spectrum estimation using half-quadratic minimization. In *Proceedings of the European Signal Processing Conference*, Vienna, Autriche, September 2004.
9. M. Wagner, D. Naik, and A. Pothen. Protocols for disease classification from mass spectrometry data. *Proteomics*, 3(9):1692–1698, Sep 2003.
10. Q. Liu, B. Krishnapuram, P. Pratapa, X. Liao, A. Hartemink, and L. Carin. Identification of differentially expressed proteins using maldi-tof mass spectra. In *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1323– 1327, Nov 2003.
11. A.K. Louis, P. Maass, and A. Rieder. *Wavelets: Theorie und Anwendungen.* B.G. Teubner, Stuttgart, 2nd edition, 1998.
12. G. P. Nason and B. W. Silverman. The stationary wavelet transform and some statistical applications. In *Lecture Notes in Statistics*, volume 103, pages 281–300, 1995.
13. E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359(9306):572–577, Feb 2002.
14. L. Li, H. Tang, Z. Wu, J. Gong, M. Gruidl, J. Zou, M. Tockman, and R. A. Clark. Data mining techniques for cancer detection using serum proteomic profiling. *Artif Intell Med*, 32(2):71–83, Oct 2004.

15. J. L. Norris, D. S. Cornett, J. A. Mobley, S. A. Schwartz, H. Roder, and R. M. Caprioli. Preparing maldi mass spectra for statistical analysis: A practical approach. In *Proceedings of the 53rd ASMS Conference on Mass Spectrometry and Allied Topics*, San Antonio, TX, June 2005.
16. E. T. Fung and C. Enderwick. ProteinChip clinical proteomics: computational challenges and solutions. *Biotechniques*, Suppl:34–8, 40–1, Mar 2002.
17. K. A. Baggerly, J. S. Morris, J. Wang, D. Gold, L.-C. Xiao, and K. R. Coombes. A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics*, 3(9):1667–1672, Sep 2003.
18. R. N. McDonough and A. D. Whale. *Detection of Signals in Noise*. Academic Press, San Diego, 2nd edition, 1995.
19. S. Guiasu and A. Shenitzer. The principle of maximum entropy. *The Mathematical Intelligencer*, 7(1):42–48, 1985.
20. J. J. Verbeek, N. Vlassis, and B. Kröse. Efficient greedy learning of gaussian mixture models. *Neural Comput*, 15(2):469–485, Feb 2003.
21. P. Paalanen, J.-K. Kamarainen, J. Ilonen, and H. Kälviäinen. Representation and discrimination based on gaussian mixture model probability densities - practices and algorithms. Technical Report 95, Lappeenranta University of Technology, Department of Information Technology, 2005.
22. Jingfen Zhang, Wen Gao, Jinjin Cai, Simin He, Rong Zeng, and Runsheng Chen. Predicting molecular formulas of fragment ions with isotope patterns in tandem mass spectra. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 02(3):217–230, 2005.
23. H. J. Wolfson and I. Rigoutsos. Geometric hashing: An overview. *IEEE Computational Science & Engineering*, 4(4):10–21, /1997.
24. R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q.-T. Le. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics*, 20(17):3034–3044, Nov 2004.
25. T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.
26. D. Blackwell and J. MacQueen. Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.
27. D.J. Aldous. *Exchangeability and related topics*, volume 1117 of *Lecture Notes in Math - Ecole d'ete de probabilites de Saint-Flour*. Springer, Berlin, 1983.
28. H. Ishwaran and L. F. James. Generalized weighted chinese restaurant process for species sampling mixture models. *Statistica Sinica*, 3:1211–1235, 2003.
29. A. Y. Lo. Weighted chinese restaurant processes. *Cosmos*, 1(1):107–111, 2005. World Scientific Publishing Company.
30. S. J. Scheather. Density estimation. *Statistical Science*, 19(4):588–597, 2004.