Spectroscopy 17 (2003) 579–595
IOS Press

Automated Peak Harvesting of MALDI-MS spectra for high throughput proteomics

E.J. Breen *, W.L. Holstein, F.G. Hopwood, P.E. Smith, M.L. Thomas and M.R. Wilkins
Proteome Systems Ltd, Locked Bag 2073, North Ryde, NSW 1670, Australia

Abstract. High throughput proteomics is realized not only by the use of automated hardware but also by the application of efficient, automated software routines to complex data. In this paper, we present the recent developments of our software tool Peak Harvester for the automatic harvesting of monoisotopic peaks from MALDI-TOF mass spectra of peptides. Peak Harvester uses advanced mathematical morphology to convert mass spectra into stick representations. Poisson modeling of theoretical isotopic distributions is then applied to derive the monoisotopic peptide mass from an isotopically resolved group of peaks. The accuracy of Peak Harvester is demonstrated via the analysis of peptide spectra from low concentrations of bovine serum albumin blotted onto PVDF membranes and of tryptic digested platelet proteins derived from human blood following two-dimensional gel electrophoresis. The results demonstrate the power of this software as it can accurately assign monoisotopic masses, including those from overlapping isotopic distributions, and picks masses as accurately as an experienced human operator. We have further developed Peak Harvester to include peak harvesting from MALDI-TOF Post Source Decay (PSD) experiments. Here we demonstrate the versatility of the software by both the analysis of PSD data from 2DE and the analysis of peptide mass spectra collected directly from tryptic digests on a PVDF membrane.

Keywords: Mathematical morphology, proteomics, MALDI-TOF, isotopic distribution, Poisson, identification

Abreviations

MALD-TOF: matrix assisted laser desorption/ionization time of flight, MS: mass spectrometry, PMF: peptide mass fingerprinting, PSD: post source decay, PVDF: polyvinylidene difluoride, DE: dimensional electrophoresis, SDS-PAGE: sodium dodecyl sulfate polyacrylamide gel electrophoresis.

1. Introduction

Mass spectrometers are being increasingly applied to protein identification and characterization in proteomics. Many of these mass spectrometers are now offered with an advanced capacity for automatic analysis of samples, bringing true high throughput capacity to the field. For example, MALDI-TOF mass spectrometers from Kratos and Bruker are equipped with 384 sample targets. Typically, a plate containing this number of samples can be analysed automatically overnight, to yield a parent ion mass spectrum for each sample. Some MALDI-TOF instruments, namely those equipped with curved field reflectrons, are also capable of undertaking automatic post source decay (PSD) analysis of suitable peptides identified in the parent ion spectra. Together, these analyses can generate many hundreds of mass spectra per instrument per day, which forms a massive task to interpret manually. Unfortunately, the situation is even more challenging in LC-MS-MS instruments, which can generate thousands of spectra per day if long runs and high scanning frequencies are used. Clearly, there is a requirement for the automation of spectral analysis, as well as acquisition, to take advantage of the throughput that is currently available.

*Corresponding author. Fax: +61 2 98891805; E-mail: Ed.Breen@ProteomeSystems.com.

A number of groups, including ourselves, have recently described methods to automatically process peptide mass spectra to yield monoisotopic masses [1–5]. A variety of approaches have been used by the different groups, which generally involve the background removal of noise in spectra, approaches for finding the peaks (such as: peak picking, harvesting or deconvolution), a means of deriving mass dependent isotopic peptide distributions (such as, Poisson distributions, regression functions or template construction) and the application of this model to the processed spectrum to identify the monoisotopic peaks.

Peak picking approaches to monoisotopic peaks are essentially cross correlation approaches [2], where by monoisotopic peaks are selected from a correlation between a templating function and the mass spectra, rather than from the mass spectra itself. Similarly, deconvolution approaches select monoisotopic peaks from deconvolved spectra. In contrast, a harvesting approach, and as used here, extracts its peaks directly from the mass spectra, via an inplace sequential isotope distribution subtraction scheme [1]. Like deconvolution [3], but unlike peak picking [2], harvesting inherently detects overlapping distributions. Yet, it is considerably more efficient than deconvolution because it works directly with the modulation (isotope distributions) in the mass spectra, rather than attempting to remove it prior to peak detection.

The ability to produce mass defined isotopic patterns is one of the major problems to be faced when attempting to automate monoisotopic peak selection. Our solution [1] to this problem was to regress Poisson probability distributions against just 15 hypothetical peptides made up from a repeating unit of an average amino acid. This considerably simplified our approach as we didn't need to derive all possible isotopic patterns by averaging all peptides with a given mass and over all known masses [2]. This latter approach is further complicated by the mass alignment inaccurracies inherent within mass spectrometers and it also seems to be applicable to peptides with masses below 1500 Da, where the mass populations of peptides are clustered. Above 1500 Da the discrete clustering begins to disappear [3].

Here we present considerable refinements to our approach for the identification of monoisotopic peak masses in peptide mass spectra. We describe improved methods of resampling and pre-processing the spectra, a mathematical means to deal with a phenomenon we call "alignment error" during fitting of the identified peaks to a Poisson intensity model, and a new post-processing heuristic to better check resulting monoisotopic peak lists for error. We illustrate the versatility of the peak detection routine by application to two challenging MALDI-TOF mass spectra types, the first being peptide mass spectra generated from samples presented to the mass spectrometer on PVDF membranes; the second being postsource decay mass spectra of parent ions.

2. Materials and methods

2.1. One-dimensional gel electrophoresis (1DE) of bovine serum albumin

Titrated amounts of Bovine serum albumin (BSA) were prepared in SDS-PAGE sample buffer and reduced and alkylated (using dithiothreitol and acrylamide, respectively) for 1 hr at room temperature prior to electrophoresis. The sample was electrophoresed on 6–15% (w/v) polyacrylamide ProteoGel^{TM} (Sigma-Aldrich, St. Louis, MO) following the manufacturer's instructions.

2.2. Blotting of gels and on-membrane protein digestions

Proteins separated by 1D SDS-PAGE were electroblotted onto Immobilon-P PVDF membranes (Millipore, Bedford, MA), using a prototype ElectrophoretIQ^{TM} electroblotting apparatus (PSL, Sydney, Aus-

tralia). Electroblotting was carried out at 400 mA for 1.3 hr applying methods described by Khyse-Anderson [6] followed by protein staining using Direct Blue 71 (Sigma-Aldrich, St. Louis, MO). The PVDF membranes were then adhered to an Axima-CFR MALDI target plate (Kratos, Manchester, UK) and enzymatic digestions, were carried out directly on protein bands blotted onto PVDF membranes using Porcine trypsin (Sigma-Aldrich, St. Louis, MO) followed by matrix addition (α-cyano-4-hydroxycinnamic acid) to the resultant peptides as described in [7]. All dispensing of chemicals to the membrane was carried out using a an α-version Chemical Printer jointly developed by Proteome Systems Ltd. (Sydney, Australia) and Shimadzu-Biotech (Kyoto, Japan). Glass capillary piezoelectric devices (Microfab Technologies, Inc., Plano, TX) were used to micro-dispense all solutions which were pre-filtered through membrane filters (Millipore, Bedford, MA) prior to dispensing.

2.3. Two-dimensional gel electrophoresis of platelet proteins and in-gel digestion

Human platelets were sourced from the Red Cross Blood Bank (Sydney, Australia). Contaminating red blood cells were removed from the platelets by centrifugation at 200g for 10 min at 4°C. The platelet-rich plasma was then centrifuged at 1500g for 20 min at 4°C. The platelet component of the pellet was gently resuspended in 50 mM Tris-HCl, 90 mM NaCl 5 mM EDTA, pH 7.4 and washed twice more. The platelet pellet was freeze dried overnight. A crude platelet membrane preparation was prepared by suspending 200 mg of lyophilized platelets in 10 ml of 100 mM sodium carbonate and sonicated at 70% intensity in a Branson Digital sonicator Model450 four times for 15 seconds whilst keeping cool on ice. After sonication the sample was stirred for 1 hr at 4°C. The sample was centrifuged at 115,000g for 75 minutes at 4°C. The pellet was resuspended in 50 mM Tris pH 7.3 with the assistance of an ultrasonic bath. The centrifugation and resuspension were repeated another two times. The final pellet was resuspended in 2–5 ml of 7 M urea, 2 M thiourea 1% (w/v) C7, 40 mM Tris. Tributyl phosphine was added to a final concentration of 5 mM and incubated at room temperature for 1 hour. Acrylamide was added at a final concentration of 10 mM for 1 hour and a protein estimation performed, and the sample adjusted to obtain a final protein concentration of 4 mg/ml.

Two dimensional gel electrophoresis: (i) Before rehydration of IPG strips, sample was ultrasonicated for 2 min and then centrifuged at 21000g for 5 min. The supernatant was collected and 10 μl of Orange G finally added as an indicator dye. (ii) Dry 24 cm IPG strips (Amersham-Pharmacia Biotech., Uppsala, Sweden) were rehydrated for 6 hr with 400 μl of protein sample. Rehydrated strips were focused on a Protean IEF Cell (Bio-Rad, Hercules, CA) for 120 kV hr at a maximum of 10 kV. Focused IPG strips were equilibrated for 20 min in 6 M urea, 2% (w/v) SDS, 50 mM Tris-HCl, pH 7.0. (iii) Equilibrated strips were inserted into loading wells of 6–15% (w/v) tris-acetate SDS-PAGE pre-cast prototype 10 cm $\times$15 cm GelChips^{TM} (Proteome Systems, Sydney, Australia). Electrophoresis was performed at 50 mA for 1.5 hr. Proteins were stained overnight using Coomassie stain G-250 destained with 1% acetic acid and gels rehydrated in water prior to gel spot excision.

To digest proteins, to peptides gel pieces were excised using a prototype Xcise^{TM} system (Proteome Systems, Sydney, Australia and Shimadzu-Biotech, Kyoto, Japan) and then washed with 25 mM NH_4HCO_3, pH 8.5. Gel pieces were then dehydrated under vacuum for 15 min and digested with 10 μl of 20 μg/ml porcine trypsin in 25 mM NH_4HCO_3, pH 8.5, overnight at 30°C. Peptides were extracted from gel pieces with 10 μl of 0.5% (v/v) formic acid and by sonication for 10 min. Prior to MALDI-TOF MS analysis, peptides were concentrated and purified using a C_{18} ZipTip® (Millipore, Bedford, MA) eluted onto a target plate in 2 μl of matrix solution and allowed to dry.

2.4. Mass spectrometric analysis of peptides

Peptide and peptide post source decay spectra were collected using an Axima-CFR MALDI-TOF mass spectrometer (Kratos, Manchester, UK) using time delayed extraction in reflectron mode. All spectra were internally two-point calibrated on trypsin autodigestion peaks or a standard peptide (ACTH) added to the matrix.

2.5. Protein identification via database searching

To achieve database protein identification we used in house databases and tools (Proteome Systems Limited, Sydney, Australia) or PeptIdent from the ExPASy Molecular Biology Server (www.expasy.ch/tools/peptident.html). We searched the Swiss-Prot, TrEMBL and/or our in house databases. Mass error tolerances of 100 ppm and 1.0 Da were used for peptide mass fingerprinting and post source decay data, respectively.

3. Results and discussion

3.1. Pre-processing of spectra

Time-of-flight (TOF) mass spectrometers sample spectra, S, linearly in the time domain, t, and because TOF is proportional to the square root of the mass to charge (m/z) ratio of the ions upon conversion of the spectrum to the mass domain, the data is non-evenly spaced. This presents a challenge as it is easier to develop analysis procedures based on evenly sampled data. Yet in the case of TOF–MS, the mass (m/z) difference between sample points is not constant. Hence, it is necessary to resample the mass spectra at even mass intervals.

In a prior report on the Peak Harvester software [1], linear interpolation was applied to resample the original mass spectra data. However this routine was found to be unsatisfactory because in some situation this could lead to the underestimation of the peak heights, as shown in Fig. 1(a). Linear interpolation can also lead to a smoothing of the data and even to loss of peak information. Cubic interpolation is another interpolation method that is often used. With cubic interploation, as seen in Fig. 1(b), artificially high peaks can be introduced that lead to erroneous intensity values and peak positions.

In comparison, we have found that nearest-neighbour interpolation, as shown in Fig. 1(c), accurately preserves the peak intensity information. Although, as with all interpolation methods, shifts in positional information can and do occur. To date, we have found that nearest neighbour interpolation has proven to be the more reliable resampling routine, especially when considering the importance of peak height information when dealing with overlapping distributions, as will be discussed in more detail in Section 3.5.1. An example of the nearest-neighbour interpolation is applied to a peptide distribution in Fig. 2.

3.2. Mathematical morphology

Following resampling, we apply mathematical morphology [8–10], as it allows us to design filters that accurately focus in on peak widths and on the distance between peaks. Basically, a morphological filter is any filter which is idempotent $\Phi(\Phi(g)) = \Phi(g)$, increasing $g \geqslant f \Rightarrow \Phi(g) \geqslant \Phi(f)$ and is either antiextensive $\Phi(g) \leqslant g$ or extensive $\Phi(g) \geqslant g$. A mathematical morphology filter is any morphological filter that can be expressed in terms of erosions and dilations.

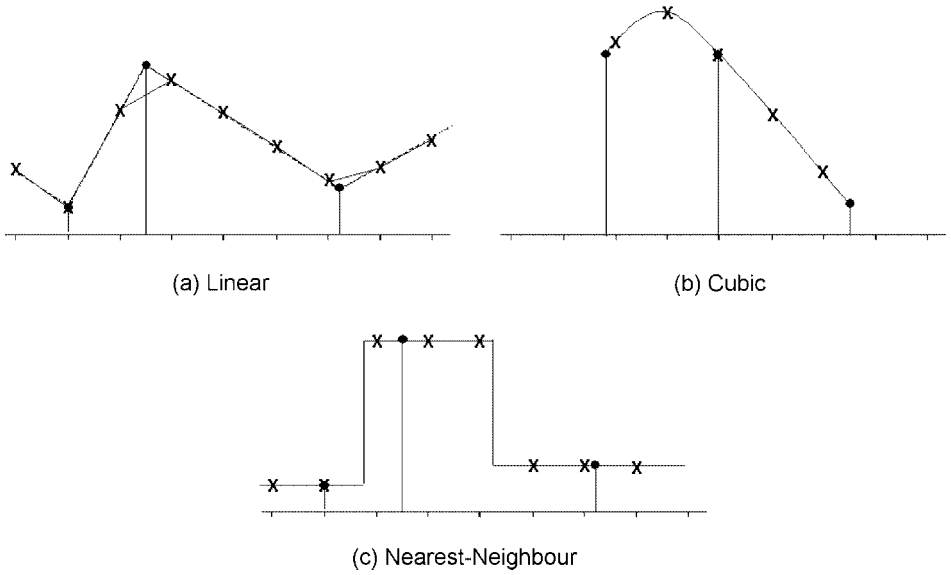


Fig. 1. Examples of interpolation schemes. Solid dots represent the original data points, while $\times$ represents the sampled points.

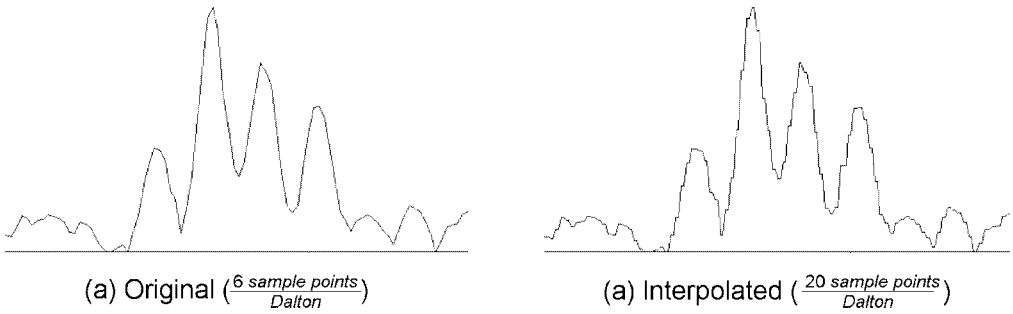


Fig. 2. Example of nearest neighbour interpolation on a 2800 Dalton peptide distribution.

The erosion (ε) of a spectrum S at mass m with a structuring element B is denoted by:

$$\varepsilon_B(S)(m) = \min_{b \in B} S(m + b), \tag{1}$$

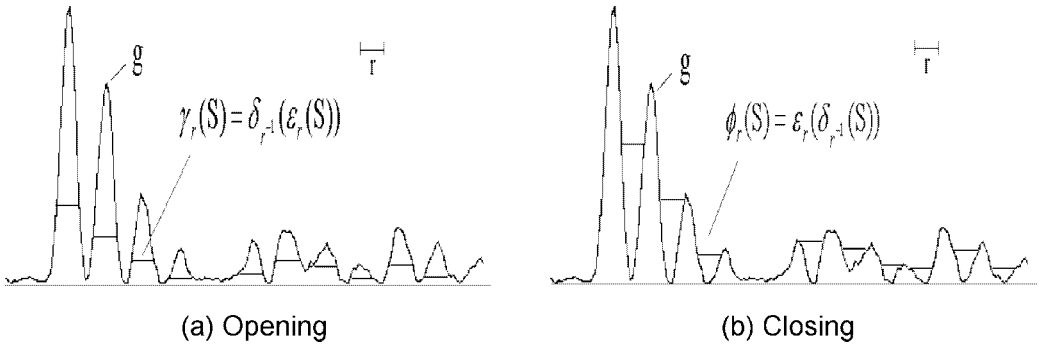
here the structuring element B is flat and is simply a line of some specified length, l, with the origin at its centre; that is, $B = \{-l/2, -1, 0, 1, l/2\}$.

The dilation of a spectrum with a structuring element is denoted by:

$$\delta_B(S)(m) = \max_{b \in B} S(m + b). \tag{2}$$

These two basic transformations are often combined to produce openings:

$$\gamma_B(S) = \delta_{\check{B}}(\varepsilon_B(S)) \tag{3}$$



(a) Opening (b) Closing

Fig. 3. Examples of openings and closings with a line or length r. Note that the symmetric set of r is represented by r^{-1}.

and closings:

$$\Phi_B(S) = \varepsilon_B\big(\delta_{\breve{B}}(S)\big), \tag{4}$$

where $\breve{B} = \{-b \mid b \in B\}$ is the symmetric set of B with respect to its origin. Any closing is the dual of a particular opening and vice-versa; that is:

$$\Phi_B(S) = \big(\gamma_B\big(S^c\big)\big)^c, \tag{5}$$

where f^c represents the complement of $f : f^c(x) = t_{\max} - f$, and $t_{\max}$ the maximum of f.

Figure 3 provides an example of an opening and a closing. Notice that an opening attenuates peaks, while a closing fills in troughs. The amount of attenuation or filling is determined by the length of the line r, as seen in Fig. 3.

3.3. Spectrum cleaning

A prominent feature in raw spectra is the common presence of a background trend, where the low mass range of the spectrum does not reach the baseline, refer to Fig. (4)a.

Typically large openings are found to be good for estimating the background level, however we have found better results by constructing filters that alternate between closings and openings. For example, an estimate of the lower envelop, L, of the spectrum can be obtained via:

$$L(S) = \varepsilon_x\big(\delta_{x+y}(\varepsilon_y)\big), \tag{6}$$

where the lengths of x, y are typically 100, 11 Da, respectively. While these values may seem large, these transformations can be computed very efficiently [11].

The background corrected spectrum, $C(S)$, shown in Fig. 4(b), was obtained via

$$C(S)(m) = \begin{cases} 0 & \text{if } S(m) - L(S)(m) < 0, \\ S(m) - L(S)(m) & \text{otherwise.} \end{cases} \tag{7}$$

To determine the noise level, N, in the spectrum, we first compute the upper envelop U of the spectrum by:

$$U(S) = \delta_x\big(\varepsilon_{x+y}(\partial_y)\big) \tag{8}$$

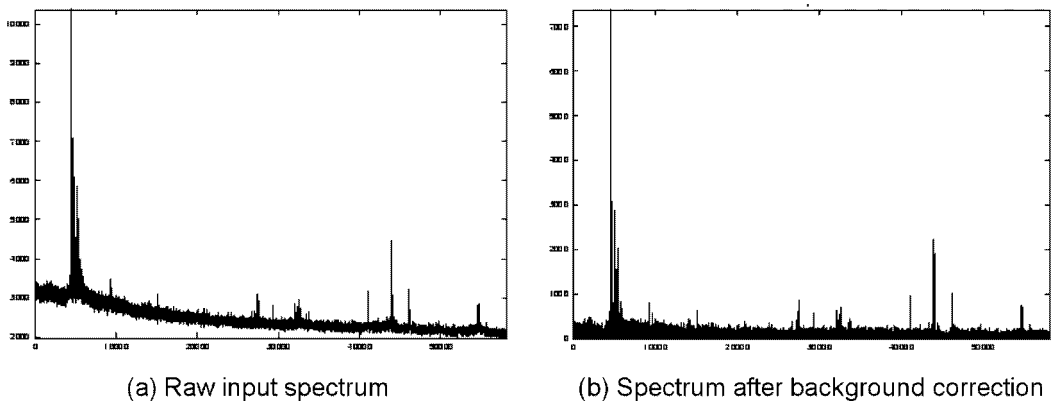


Fig. 4. Example of Background subtraction from MALDI-TOF spectra. Reproduced with permission from *Electrophoresis* **21** (2000), 2243–2251.

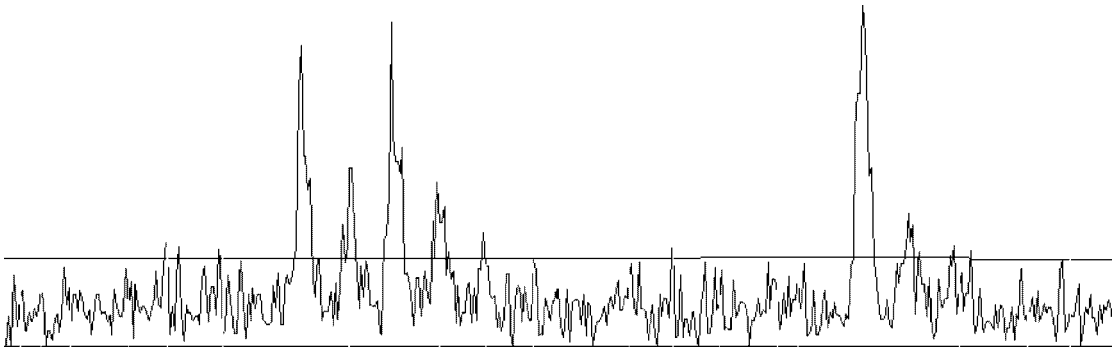


Fig. 5. Estimated noise level in a region of MALDI-TOF spectra. The upper line drawn across the signal represents the noise level, which can vary from point to point. See Eq. (9). Reproduced with permission from *Electrophoresis* **21** (2000), 2243–2251.

and then the noise level is computed via:

$$N(S)(m) = \begin{cases} 0 & \text{if } U(S)(m) - L(S)(m) < 0, \\ U(S)(m) - L(S)(m) & \text{otherwise.} \end{cases} \tag{9}$$

An example of the use of Eq. (9) is given in Fig. 5.

3.4. Spectrum segmentation

Following the above signal cleaning, small irrelevant maxima are then removed by applying a small opening. Next, regional maxima are extracted. Regional maxima are defined as points or groups of points that have intensity values that are strictly greater than their neighbouring points. The extracted maxima are then used as seeds for performing a one-dimensional watershed segmentation [10] on the spectrum, so as to obtain isolated peaks as shown in Fig. 6(b).

This information is then further reduced to a stick representation by replacing each isolated peak, Fig. 6(b), by a stick at its centroid position determined at 70% of the peaks maximum height. Refer to Fig. 7 for stick representation.

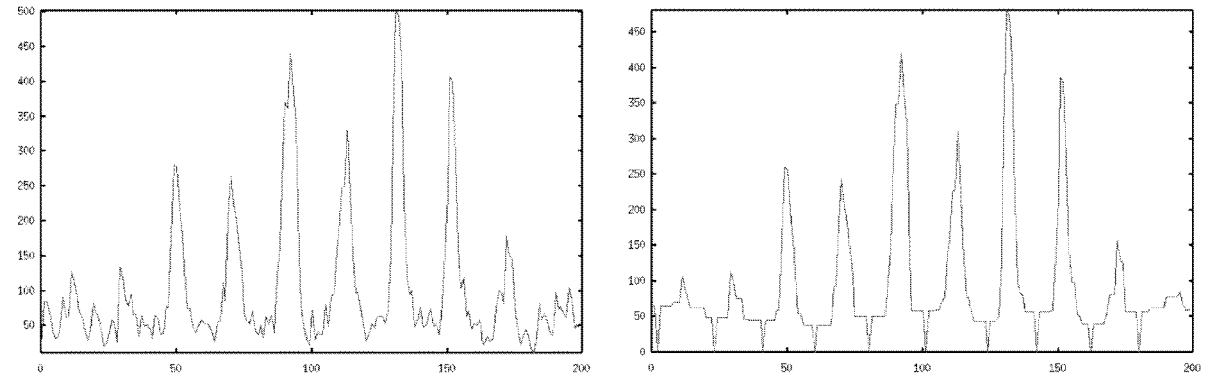


Fig. 6. Watershed segmentation of spectra to isolate major peaks. Reproduced with permission from *Electrophoresis* **21** (2000), 2243–2251.

3.5. Harvesting monoisotopic peaks

From each peptide mass spectrum we are interested in extracting the monoisotopic peaks from each isotopic distribution for each peptide mass. The isotopic distributions arrive due to the naturally occurring presence of H_2, C_{13}, N_{13}, O_{17} and O_{18} and sulfur elemental isotopes in the peptides analysed. To harvest these peaks, we use a Poisson model approach [1]. Basically, the Poisson model is a probability distribution that we use to relate the number of atoms, n, to proportion p of its isotopes:

$$P(x;M) = \begin{cases} \dfrac{\mathrm{e}^{-M}M^z}{x!} & \text{if } x = 0,1,2,\mathrm{K};\ M > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where M, the mean of the distribution, represents the product np.

Since there are many isotopic distributions and we will not in general know the appropriate values of n and p, we previously derived a mapping function $F : m \rightarrow M$, where m is mass. We did this by deriving a hypothetical average amino acid $u = C_{10}H_{16}N_3O_3$, scaled to whole numbers, forming peptides composed of repeating units u from 1 to 15 (corresponding to peptides masses between 245.1376 and 3410.8059 Da), and deriving the mapping function:

$$F(m) = 0.000594m - 0.03091. \tag{11}$$

Since we were able to capture our optimization steps [1] into a simple regression type equation (Eq. (11)) and the evaluation of Eq. (10) is fast, we are able to define an average peptide isotopic distribution for any value of m, and in real time. Other techniques are considerably more involved and do not readily lend themselves to analysis of higher mass peptides [3] or are slow for real time analysis [2], and hence require further sampling and estimation procedures.

Also, the peak harvester approach could be defined by using a different representation of u and therefore extending the application of this peak harvesting tool to a greater number of applications, including polymers and glycoproteins or any other system where a repeating unit (u) can be characterized.

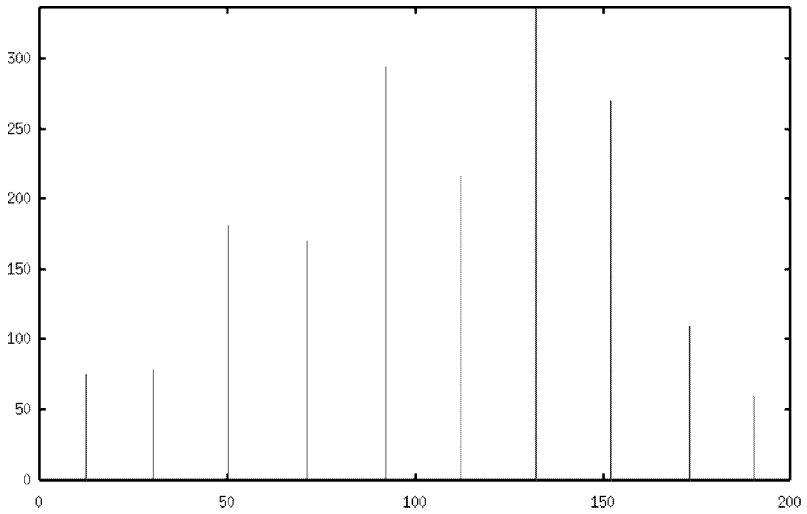


Fig. 7. Spectra processed into sticks. Reproduced with permission from *Electrophoresis* **21** (2000), 2243–2251.

3.5.1. Fitting the Poisson model

There are two predominant variables to consider when fitting the Poisson distribution to a stick representation of a mass spectrum, Fig. 7. These are the presence of overlapping isotopic distributions, illustrated in Fig. 8, and what we term the alignment error in the isotope separations, when the centroid peak values deviate from exactly 1.0 Da.

The alignment error is generally due to limitations of mass resolution, peak asymmetry and sample rate and is more often seen in weak signals than strong signals. To compensate for the alignment error (e), at each candidate monoisotopic position, $S(m)$. The signal upstream is inspected for potential isotopes. This is achieved by collecting at 1 Da intervals the largest stick value within a specified alignment error range, $\pm e$ (illustrated in Fig. 9), centered at each Da position with respect to the monoisotopic peak's mass m position:

$$S_*(i) = \max_{x=m-e}^{m+e} (\breve{S}(x)). \tag{12}$$

Here, $\breve{S}$ is the stick representation of the original spectrum, Fig. 7, and this is done $\forall i \in \{m + j | j \in [1, 2, 3, \infty], P(k, M) > 0.001\}$, where $M = F(m)$.

The next step is to fit the Poisson model (Eq. (10)) to each isotope k, where $P(k, M) > 0.001$, in the distribution for mass m:

$$\hat{P}(m; k) = \begin{cases} \min\big(p(m; k), S_*(m + k)\big) & \text{if } k > 0, \\ \breve{S}(m) & \text{otherwise,} \end{cases} \tag{13}$$

where $p(m; k) = \breve{S}(m) * A(k; M)$ and $A(k; M) = p(k; M)/p(0; M)$.

We now need to measure how well the extracted Poisson model $\hat{P}$ fits the theoretical Poisson model P at mass m:

$$H(m) = \sum_{\forall k \in \{P(k,M) > 0.001\}} P(k; M)\hat{P}(m; k). \tag{14}$$

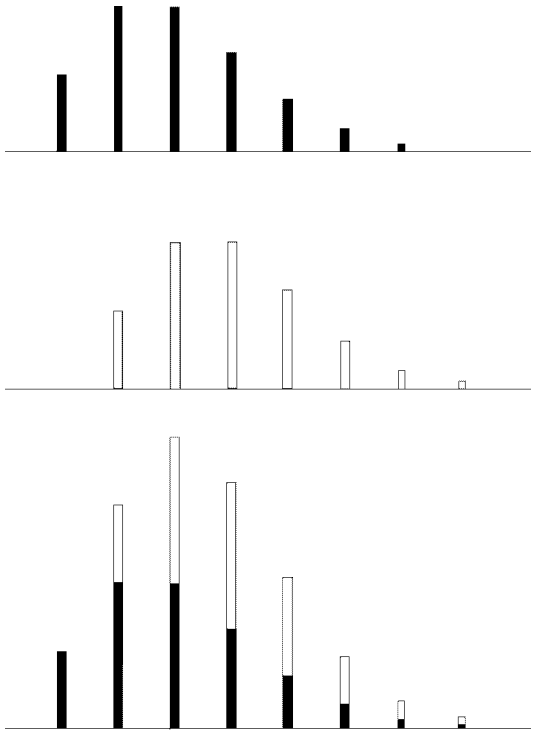


Fig. 8. Poisson model for deamidation for a peptide of mass 3410 Da. The top signal represents the monoisotopic peptide. The middle signal represents the deamidated version and the bottom signal represents the summation of the two signals above.

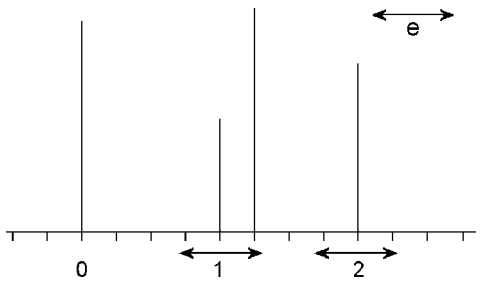


Fig. 9. Alignment error; the peak at position 0 represents a monoisotopic peak, and the location for isotopes at 1 and 2 Daltons are marked. Because there can be some error in correctly finding the centroid of a peak, we need to look for peaks that are slightly less or slightly more than 1 Da apart when we fit the poison model. The error range is given by e.

H can be viewed as representing the weighted average height of the peaks within the distribution $\hat{P}$. The height of the distribution can then be compared with the noise level at that location in the spectrum:

$$I(m) = \begin{cases} 1 & \text{if } H(m) > zN(m), \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

I is an indicator function for whether the average height of the distribution exceeds the level of the noise N by a specified amount z. All distributions that pass the threshold are considered to be valid distributions and their monoisotopic mass is recorded and the contribution of that distribution to the entire

signal is subtracted. The procedure then continues from the next position in the signal examining each further peak as a candidate monoisotopic peak. Note that this approach inherently handles overlapping distributions [1], as illustrated in Fig. 8.

3.6. Post processing: heuristics

The threshold approach described above (Eq. (15)) is effective but in some cases is too simple as it generally picks a superset of monoisotopic peaks, depending on the chosen value of z.

In order to ensure the peaks are real the list of monoisotopic peaks are further scanned to identify peaks that are within a few Daltons of each other. For all peaks that satisfy this condition the following criteria is used to eliminate the noise: if the distance between the monoisotopic peak sp at position (k) has a neighbour $(k+1)$ that is less than 3 Da away, $sp(k)$ is a valid peak if $sp(k)/sp(k+1) > 0.2$. Similarly, if the $sp(k)$ has neighbour $(k-1)$ that is less than 3 Da away, it is a valid peak if, $sp(k)/sp(k-1) > 0.6$.

This treatment helps ensure that any distribution that overlaps another is not simply due to model misfitting or low signal.

3.7. Application to data analysis

In our previous report on Peak Harvester [1], we demonstrated the successful application of this software to low concentration standard peptide analysis and in-gel digests of the protein bovine serum albumin, illustrating the fitting of the Poisson distribution to actual peptide distributions for a range of peptide masses.

Here, we further demonstrate the versatility of the peak harvesting routines by application to the detection of overlapping distributions from PMF data collected following 2DE of platelet proteins and to peptide mass fingerprinting (PMF) data generated directly from a PVDF surface (on which blotted proteins have been digested). We also demonstrate peak harvesting of post source decay (PSD) data collected following 2DE of platelet proteins.

3.8. Analysis of PMF spectra of platelet proteins

A MALDI-TOF mass spectrum was collected from a Coomassie stained protein spot isolated from a 2DE of human platelet proteins, shown in Fig. 10. A peak list from the resultant mass spectrum was processed with Peak Harvester using a maximum alignment error of 0.1 Da (the optimum error for PMF data). The peak list was searched against PMF databases using the ExPASy PeptIdent tool and the protein was identified as Chain 1: Integrin Beta-3 (Accession number P05106), with 26 peptides matched, covering 31% of the protein sequence.

The spectrum in Fig. 10 contains a large number of tryptic peptide peaks, resulting in a number of regions of overlapping distributions. An expanded view of two of these regions is illustrated in Fig. 11. In both cases Peak Harvester has identified multiple isotopic distributions and it was found that the monoisotopic masses derived and corresponding peptide sequences (identified by Peptident) are shown in Table 1. All five sequences were matched to tryptic peptides of Chain 1: Integrin Beta-3.

In the first example, Fig. 11(a), two overlapping distributions (1221.59, 1223.58 Da) can be easily seen due to their differences in intensity. Both of these distributions were detected by Peak Harvester and database searching identified the peptide masses as sequences of Chain 1: Integrin Beta-3 with no tryptic missed cleavages or other modifications. The second, and more interesting, example of overlapping distributions (Fig. 11(b)) is found at 1531.82, 1532.83 and 1533.81 Da. These masses were again

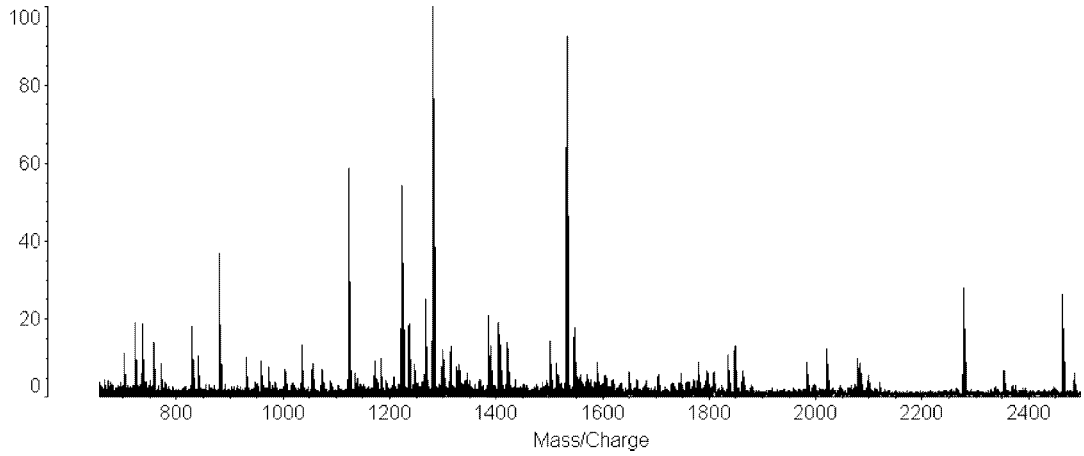


Fig. 10. Peptide mass spectrum of Chain 1: Integrin beta-3 identified from 2DE human platelet proteins.

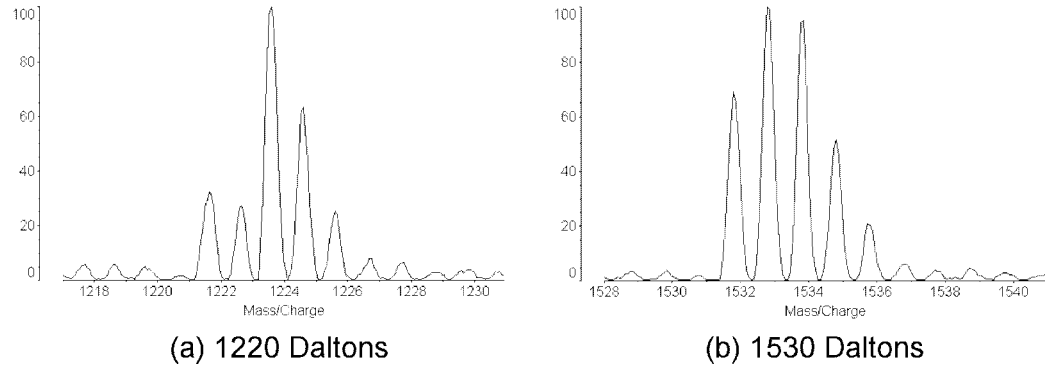


Fig. 11. Two sets of overlapping distributions from PMF mass spectrum of Chain 1: Integrin beta-3. Stars show peaks that were picked by the peak harvester software as monoisotopic masses.

found to be tryptic peptides from Chain 1: Integrin Beta-3 peptides where there are no tryptic missed cleavages. We believe the peptide at mass 1533.81 is due to the deamidation of one of the glutamine residues of the 1532.83 sequence. This preliminary assignment of the 1533.81 peptide requires further mass spectrometric confirmation to definitively assign the sequence (our standard practice is to use PSD analysis which cannot be applied for multiple peptide distributions within approximately 5–10 Da of each other). Nonetheless, the detection of these three overlapping distributions highlights the ability of Peak Harvester to extract complex distributions from MALDI PMF data.

3.9. Analysis of tryptic peptides of bovine serum albumin from a PVDF membrane

MALDI analysis of peptide digests directly from PVDF membranes is inherently challenging due to the non-flat nature of the membrane surface and lack of conductivity across the membrane. Both features can lead to broader peak resolution than that obtainable with standard MALDI data collection from a flat metal target. Here, we use data collected from the tryptic digestion of Bovine Serum Albumin (BSA) on PVDF membranes to demonstrate the rigor of the peak harvester tool in its ability to harvest such data despite variation from perfect behaviour.

Table 1
Monoisotopic masses harvested and corresponding peptide sequences matched for the five overlapping distributions shown in Fig. 11

Monoisotopic mass	Sequence matched
1221.59	WDTANNPLYK
1223.58	FQYYEDSSGK
1531.83	NDASHLLVFTTDAK
1532.82	GSGDSSQVTQVSPQR
1533.81	GSGDSSQVTQVSPQR*

*Where one of the Q residues is potentially deamidated.

To account for the reduced resolution typical of peptides collected directly from a PVDF surface, the mass spectra of BSA from PVDF were all harvested with an alignment error of 0.2 Da. BSA peptides were generated from tryptic digests on bands of BSA blotted from 1DE onto a PVDF membrane. The 1DE wells were loaded with BSA protein concentrations of 500, 250 and 100 fmol. Less than one eighth of the protein band was digested on the membrane, the equivalent of 62.5, 31.2 and 12.5 fmol of whole protein, respectively. Figure 12 shows the peaks of masses 1193.68, 1439.85 and 1479.89 Da after processing with peak harvester. The theoretical Poisson distribution is drawn inside the actual distribution of the mass spectrum. Note that the left most peak in each case is the monoisotopic mass.

For the 62.5 fmol sample (Fig. 12, upper row) all peaks have high signal to noise ratios with near baseline separation. The actual isotopic distributions map well to the Poisson distributions. At lower concentrations, for example the harvesting of the 1439.85 Da peak for the 12.5 fmol sample (Fig. 12, bottom row), the actual isotopic distribution starts to deviate from the Poisson distribution, although in this case, it is still an acceptable match for successful harvesting.

To assess the importance of the alignment error, the BSA data was reharvested with an alignment error of 0.1 Da. In this case, all peptides were correctly harvested with the exception of the 1193.68 Da peptide in the 12.5 fmol sample, as shown in Fig. 13. In this case, the distribution of the 1193.69 Da peptide did not fall within the 0.1 Da alignment error and this peptide was rejected as a potential monoisotopic candidate. The next mass at 1194.80 Da was then considered as a potential monoisotopic candidate and in this case, the isotopic distribution fitted the Poisson distribution, therefore incorrectly assigning the 1194.80 Da peptide as a monoisotopic mass. Fortunately, the peak harvesting tool described herein is adequately flexible to allow peak harvesting parameters to be optimized and preset for each type of system being analysed (i.e. our standard procedure is to harvest PMF data with an alignment error of 0.1 Da and PSD data with an alignment error of 0.2 Da), therefore removing the need for users to search with multiple parameter combinations for each sample.

3.10. Analysis of PSD spectra of platelet proteins

Our final demonstration of our peak harvesting tool is its ability to harvest peaks from PSD mass spectra. In some ways these spectra tend to be more complicated than PMF data because the spectra generally includes a lot of very small broad peaks and, due to a broader range of laser powers used, isotopic resolution varies from well resolved to poorly resolved (or completely unresolved) peptide fragments.

The processing of this data is also different than that derived for PMF analysis. Unlike the approach used with PMF data, where the Poisson model is fitted to a stick representation of the spectra as shown in Fig. 7, for PSD we do not reduce the spectra after segmentation (Section 3.4) to sticks but instead

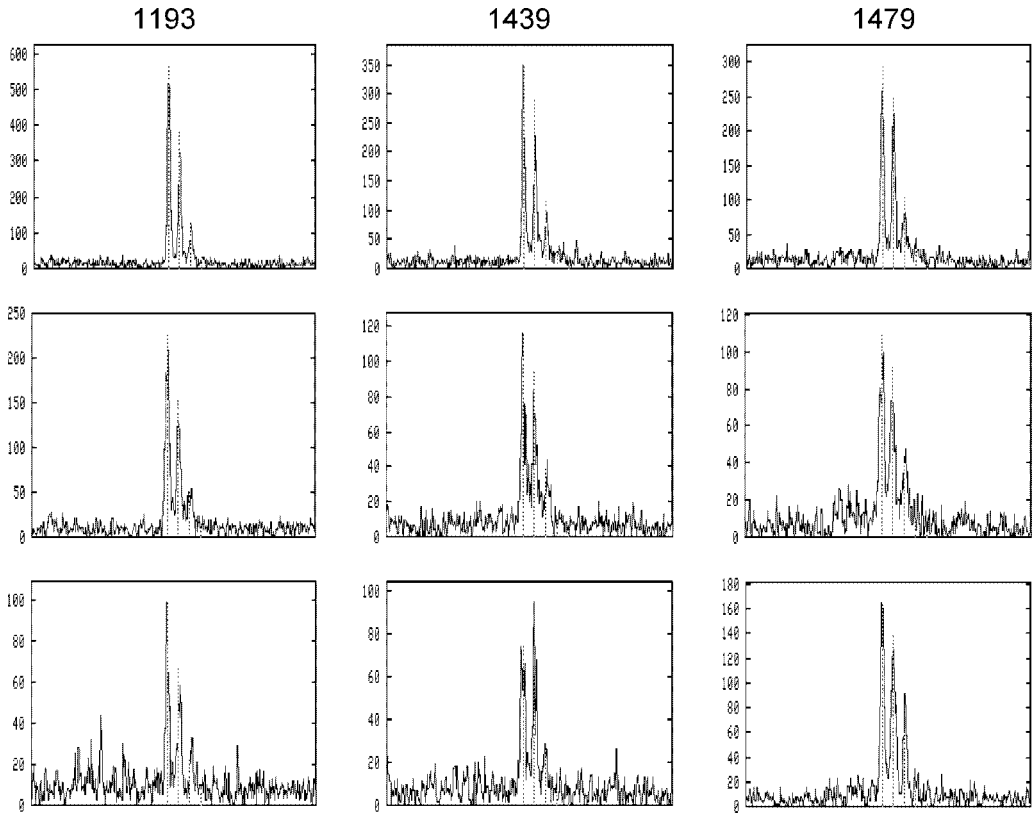


Fig. 12. Examples of peaks harvested from different concentrations of BSA electroblotted from a 1DE to a PVDF membrane. Upper row 62.5 fmol, middle 31.2 fmol and bottom 12.5 fmol.

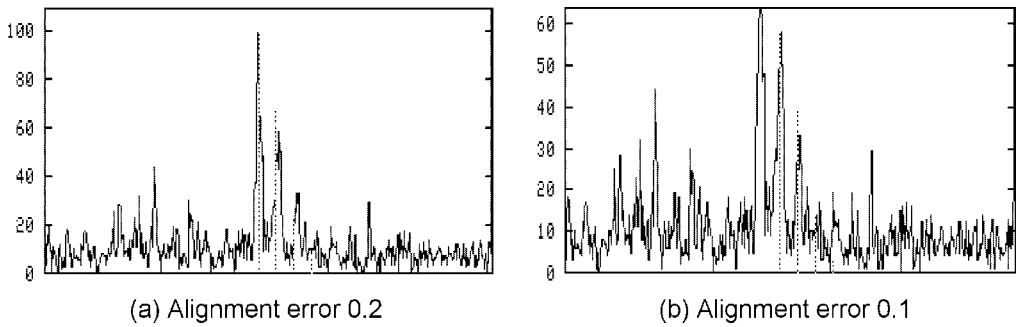


Fig. 13. Alignment error. The effect of the alignment error upon harvesting of spectra from PVDF membranes. In this case, an alignment error of 0.2 Da was required to ensure correct peak picking of the monoisotopic peak.

select all peaks which are wider than 1 dalton and above the noise level (Eq. (9)). From these peaks, we find the location of their maximum and, given their mass, we then fit a Poisson distribution such that the distribution's maximum intensity aligns with the maxima of the PSD peak (see Figs 15 and 17). Then the mass of the monoisotope is easily derived from this distribution.

PMF analysis was carried out on two more protein spots from 2DE of platelet proteins. The database searching results indicated that the two protein spots analysed contained Chain 1: Actin, cytoplasmic 1

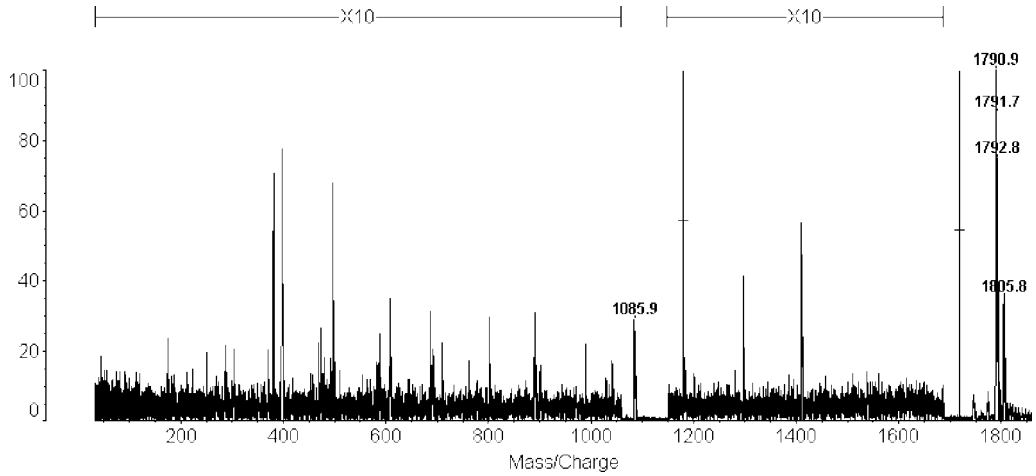


Fig. 14. PSD mass spectrum of 1791.05 Da peptide from a tryptic digest of Chain 1: Actin, cytoplasmic 1 or 2.

Table 2
Results of the PMF and corresponding PSD analysis of two proteins derived from 2DE of human platelets

Protein name	Accession No.	% Coverage	PSD mass	No. PSD hits
Chain 1: Actin, cytomplasmic 1 or 2	P02570 or P02571	50.0	1791.05	25
Beta tubulin, class VI, dJ543J19.4	Q9H4B7	53.7	1130.67	17

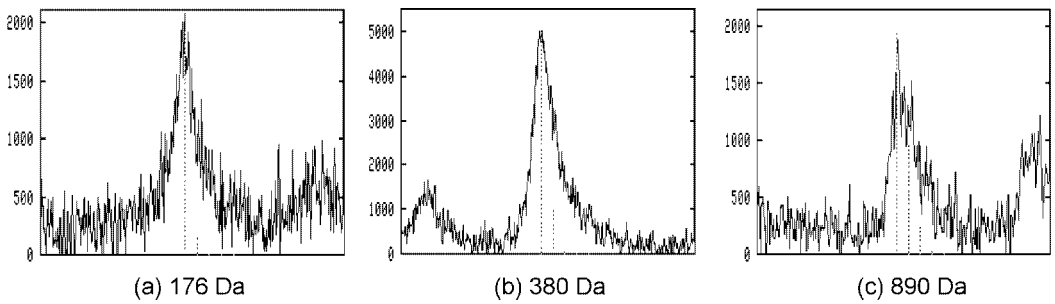


Fig. 15. PSD masses harvested from the PSD mass spectrum of the 1791.05 Da peptide of Chain 1: Actin, cytoplasmic 1 or 2 with Peak Harvester.

or 2 (accession number P02570 and P02571) and Beta tubulin 1, class VI, dJ543J19.4 (accession number Q9H4B7) as summarized in Table 2.

The two forms of the Chain 1: Actin identified, P02570 and P02571 differ only in the starting sequence of the first 9 residues. However as neither of these residues was identified in the PMF analysis, we can not distinguish between the presence of either (or both) proteins.

PSD was carried out on one peptide from each protein to confirm the identification (see Table 2). In the PSD analysis of Actin, fragmentation was carried out on the 1791.05 Da peptide, resulting in broad fragments with almost no isotopic resolution. These fragment peaks were detected using our peak harvester tool where the harvesting parameters have been previously optimized for PSD data. The PSD mass spectrum, presented in Fig. 14 shows a large number of PSD fragments. A few examples demonstrating the successful harvesting of these unresolved PSD fragments are shown in Fig. 15.

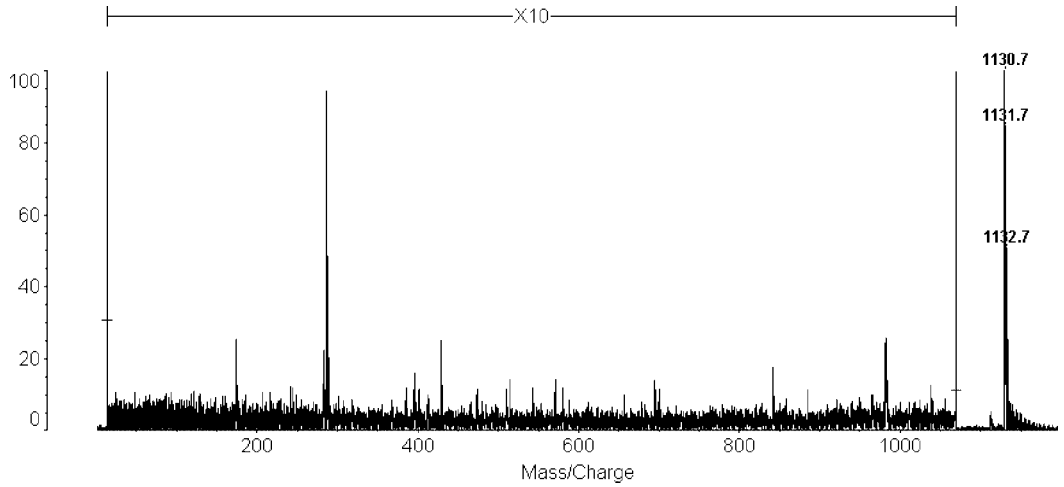


Fig. 16. PSD mass spectrum of 1130.67 Da peptide from a tryptic digest of Beta tubulin, class VI, dJ543J19.14.

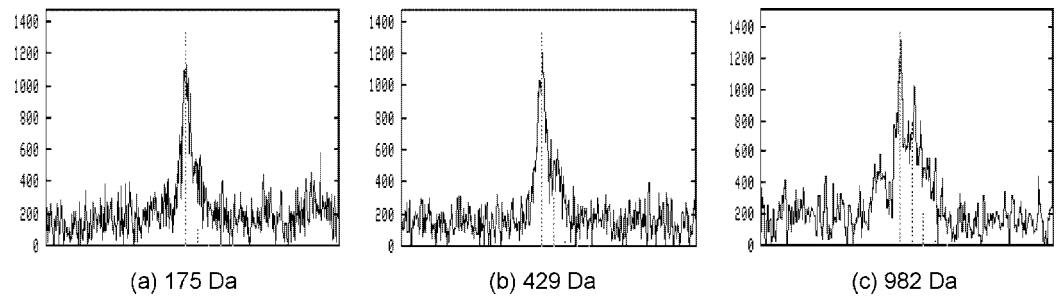


Fig. 17. PSD masses harvested from the PSD mass spectrum of the 1130.67 Da peptide from a tryptic digest of Beta tubulin, class VI, dJ543J19.14.

The resultant peak list from the peak harvester was analysed using an in house PSD database search engine. The results showed that the peptide fragments matched to 28 b and y fragments of the peptide sequence SYELPDGQVITIGNER of the protein Actin (cytoplasmic 1 or 2). This confirmed that the harvester picked peaks with the accuracy and frequency required to use PSD for confirmation of a protein indentification.

In the second example, Fig. 16, the PSD data is better resolved and therefore provides a good comparison of the ability of Peak Harvester to successfully harvest PSD spectra despite large variability in the nature of that data. PSD analysis was carried out on the 1130.67 Da peptide from Beta tubulin 1, class VI, dJ543J19.4 and the resultant spectrum was harvested with the same parameters as used for the Actin example above. Examples of the fragments harvested with Peak Harvester are shown in Fig. 17. Here, the peaks at masses 429 and 982 include some resolution of the second isotope. Peak Harvester has successfully derived the monoisotopic masses (within 1 Da of the database mass). The resultant peaks were searched with an in house PSD database search engine resulting in the identification of 17 b and y fragments from the protein sequence FPGQLNADLR of the protein Beta tubulin 1, class VI, dJ543J19.4.

4. Conclusion

In this manuscript we have further demonstrated the rigor of our Peak Harvester via its application to the analysis of complex data. The ability of the harvester to select monoisotopic peaks from data collected directly from PVDF membranes is a solid example of the flexibility of this tool and the ease at which the fundamental approach of the harvester can be easily manipulated to apply to a range of systems.

The incorporation of PSD harvesting into the tool, and along with the inherent speed at which the algorithms can be applied, has ensured that this tool is indispensable in the processing of all MALDI-TOF data generated in our laboratory. We routinely apply peak harvester to large data sets (typically 384 samples from automated MALDI-TOF analysis). We have also integrated this peak harvester into a platform which pipelines spectra though the harvester, databases the results, and then automatically sends the picked peaks through to a protein identification engine. This has given us very satisfying results, facilitating our high throughput proteomic studies.

References

[1] E.J. Breen, F.G. Hopwood, K.L. Williams and M.R. Wilkins, Automatic poisson peak harvesting for high throughput protein identification, *Electrophoresis* **21** (2000), 2243–2251.
[2] R. Gras, M. Muller, E. Gasteiger, S. Gay, P.-A. Binz, W. Bienvenut, C. Hoogland, J.-C. Sanchez, A. Bairoch, D.F. Hochstrasser and R.D. Appel, Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection, *Electrophoresis* **20** (1999), 3535–3550.
[3] M. Wehofsky, R. Hoffmann, M. Hubert and B. Spengler, Isotopic deconvolution of matrix-assisted laser desorption ionisation mass spectra for substance-class specific analysis of complex samples, *Eur. J. Mass Spectrom.* **7** (2001), 39–46.
[4] M. Wehofsky and R. Hoffman, Automated deconvolution and deisotoping of electrospray mass spectra, *J. Mass. Spectrom.* **37** (2002), 223–229.
[5] D.M. Horn, R.A. Zubarev and F.W. Mclafferty, *J. Am. Soc. Mass Spectrom.* **11** (2000), 320.
[6] J. Kyhse-Andersen, Electroblotting of multiple gels: A simple apparatus without buffer tank for rapid transfer of proteins from polyacrylamide to nitrocellulose, *J. Biochem. Biophys. Methods* **10** (1984), 203–209.
[7] A.J. Sloane, J.L. Duff, N.L. Wilson, P.S. Gandhi, C.J. Hill, F.G. Hopwood, P.E. Smith, M.L. Thomas, R.A. Cole, N.H. Packer, E.J. Breen, P.W. Cooley, D.B. Wallace, K.L. Williams and A.A. Gooley, High-throughput peptide mass fingerprinting and rotein macroarray analysis using chemical printing strategies, *Mol. Cell. Proteomics* (2002), (in press).
[8] G. Matheron, *Random Sets and Integral Geometry*, John Wiley, New York, NY, USA, 1975.
[9] J. Serra, *Image Analysis and Mathematical Morphology*, Academic Press, New York, NY, USA, 1982.
[10] P. Soille, *Morphological Image Analysis: Principles and Applications*, Springer-Verlag, Berlin, 1999.
[11] P. Soille, E. Breen and R. Jones, A fast algorithm for min/max filters along lines of arbitrary orientation, in: *IEEE Workshop on Nonlinear Signal and Image Processing*, Vol. II, I. Pitas, ed., Neos Marmaras, June 1995, pp. 987–990.