BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 7 2013, pages 925–932
doi:10.1093/bioinformatics/btt061

Data and text mining

Advance Access publication February 15, 2013

Chemical rule-based filtering of MS/MS spectra

Beáta Reiz[1,2,3], Attila Kertész-Farkas[1], Sándor Pongor[1,4] and Michael P. Myers[1,5,*]

[1]Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, 34149 Trieste, Italy, [2]Laboratory of Bioinformatics, Biological Research Centre, Hungarian Academy of Sciences, H-6726 Szeged, Hungary, [3]Institute of Informatics, University of Szeged, H-6720 Szeged, [4]Faculty of Information Technology, Pázmány Péter Catholic University, 1083 Budapest, Hungary and [5]Protein Networks Group, International Centre for Genetic Engineering and Biotechnology, 34149 Trieste, Italy

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Identification of proteins by mass spectrometry–based proteomics requires automated interpretation of peptide tandem mass spectrometry spectra. The effectiveness of peptide identification can be greatly improved by filtering out extraneous noise peaks before the subsequent database searching steps.

Results: Here we present a novel chemical rule-based filtering algorithm, termed CRF, which makes use of the predictable patterns (rules) of collision-induced peptide fragmentation. The algorithm selects peak pairs that obey the common fragmentation rules within plausible limits of mass tolerance as well as peak intensity and produces spectra that can be subsequently submitted to any search engine. CRF increases the positive predictive value and decreases the number of random matches and thus improves performance by 15–20% in terms of peptide annotation using search engines, such as X!Tandem. Importantly, the algorithm also achieves data compression rates of ~75%.

Availability: The MATLAB source code and a web server are available at http://hydrax.icgeb.trieste.it/CRFilter/

Contact: myers@icgeb.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 17, 2012; revised on December 23, 2012; accepted on February 3, 2013

1 INTRODUCTION

Identifying proteins from the mass spectra of their proteolytic peptides has become a standard method for analyzing complex biological samples (Aebersold and Mann, 2003). In a typical experiment, mass spectra obtained via liquid chromatography coupled to tandem mass spectrometry (MS/MS) are compared with theoretical spectra derived from a sequence database (Yates *et al.*, 1995) [for reviews, see Becker and Bern (2011), Deutsch *et al.* (2008), Jacob (2010), Johnson *et al.* (2005), MacCoss (2005), McDonald *et al.* (2004), Menschaert *et al.* (2010), Nesvizhskii (2010), Nesvizhskii and Aebersold (2004), Neumann and Bocker (2010), Noble and MacCoss (2012) and Webb-Robertson and Cannon (2007)]. The analysis of a single sample may require the evaluation of hundreds of thousands of experimental spectra. Commonly these are laden with extraneous peaks, which make peptide identification particularly difficult.

*To whom correspondence should be addressed.

In this work, we address the filtering of preprocessed spectra [for reviews, see Jacob (2010), Johnson *et al.* (2005), Mujezinovic *et al.* (2006), Reiz *et al.* (2012) and Salmi *et al.* (2009)], which consist of a series of monoisotopic peaks that are characterized by their fragmentation mass (m/z ratio) and their intensity. The goal of filtering is to increase the quality of the data while reducing the size of the dataset under analysis. Because the analysis capacity of instruments is constantly increasing, there is a growing need for fast and efficient filtering strategies. Briefly, filtering methods fall into two large categories: *spectrum filtering* approaches and *peak filtering* approaches.

Spectrum filtering approaches seek to identify low-quality spectra based on their peculiar intensity or m/z distributions, and subsequently exclude these spectra from further analysis. One general strategy is to use peptide fragmentation rules to identify high-quality spectra (Bern *et al.*, 2007; Flikka *et al.*, 2006; Geer *et al.*, 2004; Hoopmann *et al.*, 2007; Nesvizhskii *et al.*, 2006; Renard *et al.*, 2009). There are many ways to incorporate peptide fragmentation rules into spectrum annotation, and in a broad sense, all *de novo* algorithms and sequence tagging algorithms use fragmentation rules.

Peak filtering approaches, meanwhile, seek to discard unwanted peaks from the spectra of a particular experiment. Such unwanted peaks are identified by their low intensities and/or by the fact that they do not obey the fragmentation rules of a given experiment. Although technically incorrect, any peak that is deemed superfluous for a particular data processing workflow can be discarded from the spectrum as 'noise'. For instance, we can simply retain only the top 50 most intense peaks, or discard peaks with intensities <5% of the maximum intensity. A number of laboratories have reported on the use of such classical signal-to-noise filters that also reduce the number of peaks in a spectrum (Geer *et al.*, 2004; Renard *et al.*, 2009; Xu and Freitas, 2010). The advantage of noise filtering methods is that they can produce cleaner spectra. These can then be further analyzed with any search engine. However, the signal-to-noise filters are typically equally destructive to all low-intensity peaks, regardless of whether or not they follow the expected fragmentation rules.

To address this problem, we developed a set of spectral processing filters designed to reduce the number of uninformative peaks in peptide fragmentation spectra. The motivation is to use the chemical rules of gas-phase peptide fragmentation (Barlow and O'Hair, 2008; Biemann, 1990) to produce filtered spectra, which can then be further analyzed by any search engine of

choice. Importantly, the filters presented here remove peaks that do not form b–y (a–x, c–z) pairs or are not separated by the mass of an amino acid. In contrast to spectrum filtering methods that use related principles for controlling spectrum quality (Bern *et al.*, 2004; Flikka *et al.*, 2006; Nesvizhskii *et al.*, 2006), our strategy only discards peaks deemed to be uninformative, rather than discarding entire spectra. This peak-by-peak strategy has negligible computational costs, results in data compression rates of ~75% and a substantial improvement in the number of the annotated spectra at the same false discovery rate (FDR) level as compared with unfiltered data.

The article is structured as follows: Section 2 describes the data and the general computational methods, Section 3 describes the results (i.e. the principle, the algorithm and the tests conducted on the method) and Section 4 contains the discussion and conclusions.

2 METHODS

Datasets. The *Jurkat* dataset is an experimentally acquired set of tryptic peptide spectra obtained from a human cell extract, collected on a Thermo LTQ Orbitrap instrument (Kil *et al.*, 2011). The dataset contains 5853 MS/MS spectra (charges from +2 to +6) and were downloaded from Proteome Commons (https://proteomecommons.org/). UPS, a second experimental dataset, was created from a tryptic digest of the Universal Proteomics Standard I (Sigma) using an Applied Biosystems 4800 MALDI-TOF/TOF instrument as previously described (Bish *et al.*, 2008). The UPS I standard contains 50 human proteins, and our UPS dataset contains 3368 peptide spectra. HSPP2A dataset (Glatter *et al.*, 2009), the largest, contains 29 583 spectra (20 773 doubly, 8706 triply, 474 quadruply, 26 quintuply and 4 hexuply charged spectra) obtained with LTQ mass spectrometer on trypsin-digested human protein phosphatise 2A system and was acquired from www.peptideatlas.org/repository/publications/Glatter2008.

Top *N* intensity filter. We have implemented Top N filters for a comparison as well, which retains the N most intense peaks (Chalkley *et al.*, 2005; Hansen *et al.*, 2003). $N = 40$–100 gives good results for many datasets. In our experiments, N was set to 50.

Spectrum identification. The raw and the filtered experimental spectrum datasets have been annotated with X!tandem (version 2010.12.01.1), Mascot program (version 2.2), MassMatrix (Xu and Freitas, 2010) and InsPecT version 2012 (Tanner *et al.*, 2005). For all search engines, the parent and fragment ion mass tolerance were set to 0.1 and 0.3 Da, respectively, for the UPS dataset; 0.03 and 0.3 Da, respectively, for the Jurkat dataset and 0.1 and 0.3 Da, respectively, for the HSPP2A dataset. The IPI human sequence database (version 3.71) (Kersey *et al.*, 2004) was used as a reference with X!Tandem, InsPecT and MassMatrix. We included reversed sequences as the decoy dataset. The 'total peaks' parameter in the X!tandem program was set to 5000. Default values were used for the rest of the parameters. Mascot was used with above parameters using the SwissProt human database (version 2012.03).

Performance evaluation methods. The positive predictive value (PPV) is the ratio of the matching peaks in the experimental spectra over the total peaks [see Altman and Bland (1994)], formally

$$\mathrm{PPV} = \frac{\#\mathrm{matches}}{\#\mathrm{peaks}}.$$

Filtering performance was also evaluated by the receiver operator curve (ROC) technique as follows: a dataset (raw or filtered) was submitted to a search engine, and the identified peptides were listed in the order of decreasing significance (increasing E-value). In this list, peptides of the human proteome were considered positive hits and the peptides of the decoy dataset were considered negative hits. The number of positive hits was then plotted as a function of the number of negative (decoy) hits by varying the decision threshold (Sonego *et al.*, 2008). This plot yields a monotonously increasing curve, the ROC curve, and higher running ROC curves indicate better performance. The FDR (Storey and Tibshirani, 2003) was calculated as the ratio of the number of the decoy hits over the number of the positive hits at a certain threshold t by the following formula:

$$\mathrm{FDR}(t) = \frac{\#\mathrm{decoy}(t)}{\#\mathrm{target}(t)}.$$

The ROC plot was used to compare methods at the level of the same FDR value. For instance, FDR = 0.2% is a straight line in the ROC plot, and its intersections with the ROC curves indicate the number of target and decoy peaks found at the same level of FDR. FDR = 100% would coincide with the diagonal $x = y$ line.

All calculations and chemical rule-based filtering (CRF) algorithm were implemented in MATLAB (version R2010b). The source code and a web server are available at http://hydrax.icgeb.trieste.it/CRFilter/.

3 RESULTS

3.1 Principle of CRF

Given an experimental MS/MS spectrum, CRF seeks to retain (i) high-intensity peaks that are trusted without further conditions, and (ii) low-intensity peaks that are related to one of the high-intensity peaks according to any of the following three rules:

- Mass complementation rule: the sum of the masses of a pair of the high- and low-intensity peaks that add up to the precursor mass $MH^+ + 1$ (e.g. b–y pairs);
- Amino acid mass distance rule: the mass difference of two peaks equals to one of the known—native or modified—amino acid masses ('amino acid neighbors'); and
- Amino acid mass complementation rule: the sum of two peaks and the precursor mass MH^+ differs by one amino acid mass (e.g. a b-ion and the amino acid neighbor of its y-pair).

Certain types of instruments may not produce equally intense b–y (a–x or c–z) peak pairs, for instance, triple quadrupole or quadrupole-time-of-flight instruments tend to produce only either b- or y-ions, thus only one member of the ion pair is visible. In this case, CRF keeps such peaks if they are either intense enough to be considered high-intensity peaks or are separated by a mass of an amino acid from high-intensity peaks and hence pass the amino acid mass distance rule.

Furthermore, when amino acids lose NH_3 or H_2O on collision-induced dissociation (CID) fragmentation under certain circumstances, or when some amino acids carry post-translational modifications, CRF keeps these peaks if they are intense enough or they have a high-intensity amino acid neighbor and thus pass the first or the second rule. Peaks that do not have amino acid neighbors, such as internal fragment ions, might be retained if they are high-intensity peaks.

The high- and low-intensity peaks are defined via user-specified thresholds (an example is shown in Fig. 1). As an approximation, the high- and the low-intensity peaks are defined as the percentage of the estimated maximum peak number (EMP). Assuming that under the conditions of CID, seven

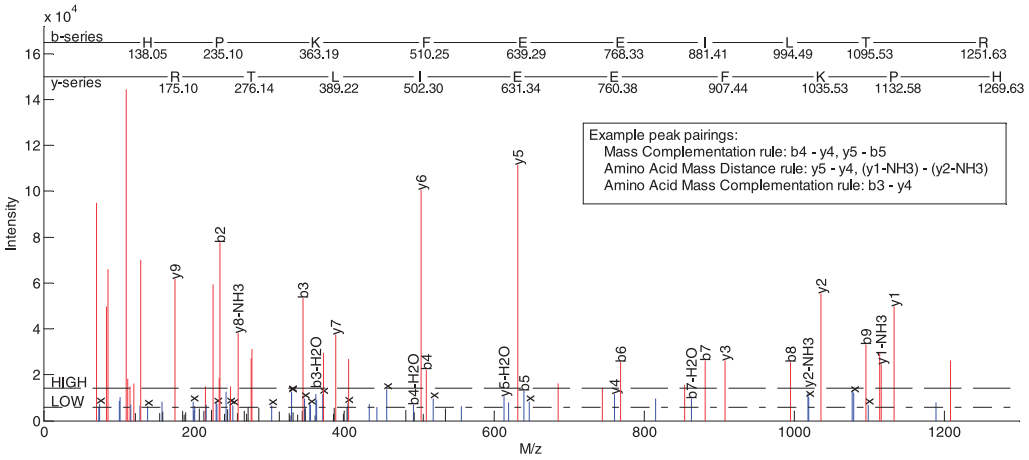

Fig. 1. An illustration of filtering rules [spectrum 953 from the UPS dataset (precursor mass = 1269.69), matching peptide from the human proteome: HPKFEEILTR]. The high-intensity peaks, marked by red, were obtained using $T_H = 50$. These peaks pass the filtering without further tests. The low-intensity peaks, marked by blue and obtained using $T_L = 110$. A blue peak is kept if it passes any of the chemical rules (i)–(iii) when paired with a red peak. For instance, blue peak $b_5 \in$ LOW (at 639.39 m/z) and read peak $y_5 \in$ HIGH (at 631.44 m/z) satisfy chemical rule (i) their masses add up to 1270.83, which equals the precursor mass +1 within 0.4 Da tolerance, so b_5 will pass. Similarly, blue peak $y_4 \in$ LOW (at 760.49 m/z) and red peak $y_5 \in$ HIGH (at 631.44 m/z) satisfy chemical rule (ii) because their masses differ by 129.06 Da, the mass of the Glutamic acid, within 0.4 Da tolerance, so y_4 will pass. Blue peak y_4 (at 760.49 m/z) and red peak $b_3 \in$ HIGH (at 363.34 m/z) satisfy chemical rule (iii) because their masses add up to 1123.83 Da, which is 146.86 Da apart from the precursor mass +1, and this difference is almost exactly the mass of phenylalanine (146.91 Da). The blue peaks that do not pass any filter (marked by a black x) are discarded. The very low-intensity peaks that are not selected to the LOW set are also discarded

ions can be produced for every peptide bond (Roepstorff and Fohlman, 1984). The number of peptide bonds can thus be roughly calculated by dividing the precursor mass by the average mass of amino acids, so EMP can be calculated as

$$\mathrm{EMP} = \frac{7 \cdot \mathrm{MH}^+}{\mathrm{AAM}},$$

where AAM denotes the average amino acid mass, which is calculated from a table of amino acid masses, plus the masses of any modifications that are specified. For instance, by specifying 30% as the upper threshold and 70% as the lower threshold, we select 30% of the EMP as 'high-intensity' peaks and the next 30–70% as 'low-intensity' ones. Peaks in the 71–100% interval are discarded. EMP is a rough estimate because it does not contain irregular fragmentation events. However, we found it useful because it does not penalize either small or large peptides.

3.2 Formal description of the CRF algorithm

Let $S = [(p_j, i_j)]_{j=1}^{n}$ denote a singly charged spectrum of n peaks, where the pair (p_j,i_j) denotes the mass (m/z) p_j and the intensity i_j of the jth peak, and without the loss of generality, let us assume the peaks are ordered by the intensity in descending order, i.e. $i_1 \geq i_2 \geq \ldots \geq i_n$. Let T_H and T_L ($T_H < T_L$) be two user-defined thresholds expressed in percentage of the EMP.

CRF uses T_H and T_L threshold parameters to calculate two sets, the HIGH(T_H) and LOW(T_H, T_L) that contain high- and low-intensity peaks, by

$$\mathrm{HIGH}(T_H) = \left\{ (p_j, i_j) \mid 1 \leq j \leq \frac{T_H}{100} \cdot \mathrm{EMP} \right\}$$

$$\mathrm{LOW}(T_H, T_L) = \mathrm{HIGH}(T_L) \backslash \mathrm{HIGH}(T_H) = \left\{ (p_j, i_j) \mid \frac{T_H}{100} \cdot \mathrm{EMP} < j \leq \frac{T_L}{100} \cdot \mathrm{EMP} \right\}.$$

CRF marks every peak in HIGH as accepted, and peaks $(p,i) \in$ LOW are marked only if there is a peak $(q,j) \in$ HIGH for which any of the following rules are satisfied:

(i) Mass complementation rule: $p + q = \mathrm{MH}^+ + 1$,
(ii) Amino acid mass distance rule: $|p - q| \in \mathrm{AM}$,
(iii) Amino acid mass complementation rule: $|\mathrm{MH} - |p - q| + 1| \in \mathrm{AM}$,

where AM denotes the set of the amino acid masses. Finally, CRF removes the unmarked peaks from the spectrum.

CRF is made tolerant to experimental error by considering the equations fulfilled within user-defined tolerance limits defined in Daltons. Two such tolerance values are used: Δ_F is the fragment ion mass (m/z) tolerance and Δ_P is the precursor ion mass (m/z) tolerance. The values of Δ_F and Δ_P depend on the accuracy of the mass spectrometer and should be set to approximately the same values as those used by the search engine. However, CRF is not indifferent to the tolerance parameters. In extreme cases of tolerance parameters, the effect of the three chemical rules becomes insignificant, and the filtering will be dominated by the intensity thresholds, which determine the sets HIGH and LOW. This happens because CRF lets all peaks pass that are in the sets HIGH or LOW if the tolerance parameters are too loose. Conversely, if tolerance parameters are too tight then only peaks in set HIGH will pass and peaks in LOW will be discarded. These examples illustrate that CRF will still keep the most intense peaks even under bad parameter settings.

When a spectrum has a multiply charged precursor ion, all possible charge states for fragmentation peaks are taken into consideration, and if such a peak obeys any of the three rules, the original peak is kept.

3.3 Testing the algorithm

To verify that the CRF improves the spectrum identification, we used the Jurkat and the UPS raw spectrum datasets. UPS is a standardized mixture of proteins, and the dataset was collected on a MALDI-TOF/TOF instrument. The Jurkat dataset was collected with an Orbitrap instrument, which is one of the most commonly used instruments for proteomics. We used X!Tandem or the MassMatrix search engines (using decoy sequences as described in Section 2). The filtering parameters were as follows: $T_H = 50\%$, $T_L = 110\%$ for the Jurkat dataset; $T_H = 70\%$, $T_L = 120\%$ for the UPS dataset and $T_H = 120\%$, $T_L = 160\%$ for the HSPP2A dataset (see Section 3.4).

The main hypothesis behind CRF is that selecting peak pairs based on appropriate chemical rules will decrease the number of random hits in a filtered spectrum as compared with a raw spectrum. For the first 'proof of principle' test, we produced a histogram of the number of matching peaks between the experimental and the theoretical spectra obtained from database searching using X!Tandem with the UPS dataset (Fig. 2). In general, it is convenient to divide such a histogram into a left and a right side roughly corresponding to high and low scores. In the low-scoring region, the distribution of experimental (target) and randomly simulated (decoy) datasets coincides. This is a consequence of the fact that a low number of matches frequently occurs at random when comparing a spectrum either with target (i.e. real) or with decoy spectra. On the other hand, in the high-scoring region, the frequency of the target hits is higher than those of decoy hits. Figure 2 shows that CRF filtering dramatically decreases random hits while leaving the true hits essentially intact. This is shown by the fact that the right-hand tail of the raw and filtered datasets are essentially identical (shown by solid red and green lines). It is also apparent that the histogram of the decoy dataset is shifted more by CRF filtering than the histogram of the target dataset (compare solid green with dashed green lines). Because significance (P-value) is usually expressed as a distance between target and decoy data, CRF filtering will naturally increase the significance of peptide identifications. This finding thus confirms that CRF removes the uninformative peaks in the spectrum dataset and decreases the influence of random matches.

Another way to show the efficiency of the filtering is to directly compare PPVs (Section 2, performance measures) obtained on filtered and raw spectra. We used the X!Tandem search engine for spectrum annotation and calculated the number of the theoretical peaks that match a peak in an experimental spectrum. The calculated PPVs were plotted as a function of the precursor mass (Fig. 3). This figure shows a large improvement in PPV for the CRF-filtered spectra. We think that this improvement is due to the elimination of the uninformative peaks, which improves the ratio of the matching (informative) peaks in a spectrum.

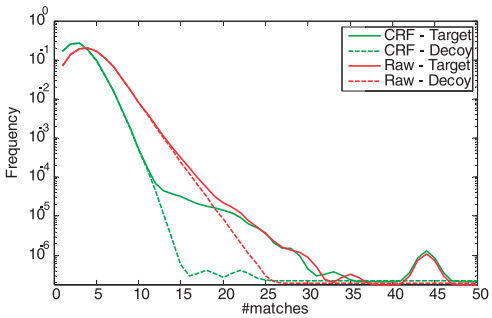

Fig. 2. Histogram of the number of matched peaks obtained during the database search on the UPS dataset. The data are normalized to frequency values (sum = 1.00)

The filtering also improves the E-values of the X!Tandem search results. Figure 4 compares X!Tandem E-values of raw and CRF-filtered data on the peptide as well as on the protein level. In this representation, each spectrum is represented by a dot of E-value pairs (y-axis corresponds to CRF-filtered E-value and x-axis to raw E-value). The oblique line corresponds to the diagonal, $y = x$, so dots above the line represent spectra where CRF filtering improves the E-values, while dots below the line are spectra where CRF filtering had a negative effect (Fig. 4 Left). It is apparent that the large majority of the annotated spectra are improved at the peptide level. Protein identification data show a similar trend (B), and here nearly all protein E-values are improved on CRF filtering. These data are consistent with the results shown in Figure 4 and indicate that CRF filtering improves the significance of peptide and protein identifications.

The efficiency of CRF and Top50 intensity filtering was compared using ROC analysis (Fig. 5). The raw and filtered datasets were submitted to the X!Tandem, MassMatrix, InsPect and Mascot search engines (using decoy sequences as described in Section 2), and the lists of identified peptides were analyzed by the ROC technique. The ROC curves in Figure 5 indicate that CRF generally identified more targets than the two other approaches.

The intersection points of the ROC curve and the straight line of FDR = 0.2% show, at the same FDR level, that CRF allows for the annotation of more spectra. Improvements in numbers are shown in Table 1, obtained with X!Tandem at the same FDR level. On the Jurkat dataset, X!Tandem with CRF filtering identifies 8.7 and 25.5% more spectra and 8.1 and 21.7% more unique peptide sequences compared with Top50 and raw, respectively. On the UPS dataset, the improvements achieved with CRF are 0.5 and 18% for spectra and 1 and 14% for unique peptide sequences compared with Top50 filtering and raw, respectively. When the Jurkat dataset is analyzed with X!Tandem, CRF identifies 8% more unique peptides than Top50, and 22% more than in the raw dataset. On the UPS dataset, the respective improvements are 1% as compared with Top50 and 14% as compared with the raw dataset. Table 1 shows the actual number of spectra and unique peptides identified with the two filters, respectively. Roughly speaking, CRF increased both the number of spectra and the number of the unique peptide sequences by 15–25% as compared with unfiltered data and by 1–5% compared with Top50 filtered data.

3.4 Parameter selection

CRF has two adjustable threshold parameters: T_H is used for defining the high-intensity peaks that are accepted without condition and T_L is used for defining low-intensity peaks that are accepted only if they form a pair with other high-intensity peaks. We examined how the thresholds T_H and T_L affect the spectrum

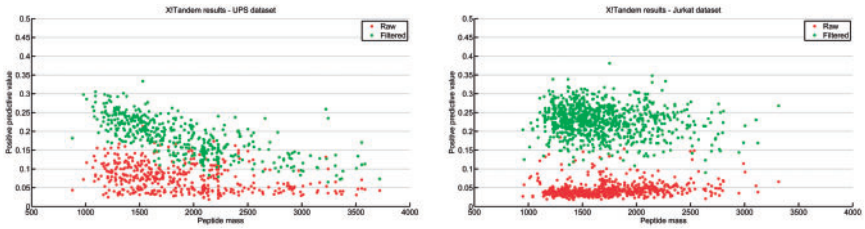

Fig. 3. PPV values versus precursor mass. CRF filtering improves generally the ratio of the informative peaks among all peaks

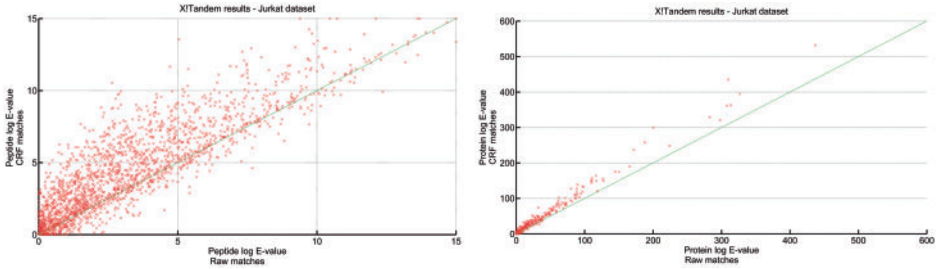

Fig. 4. Comparison of CRF-filtered and raw UPS datasets evaluated with X!Tandem search engine. The diagram on the left side represents peptide identifications, while the diagram on the right side shows protein identifications

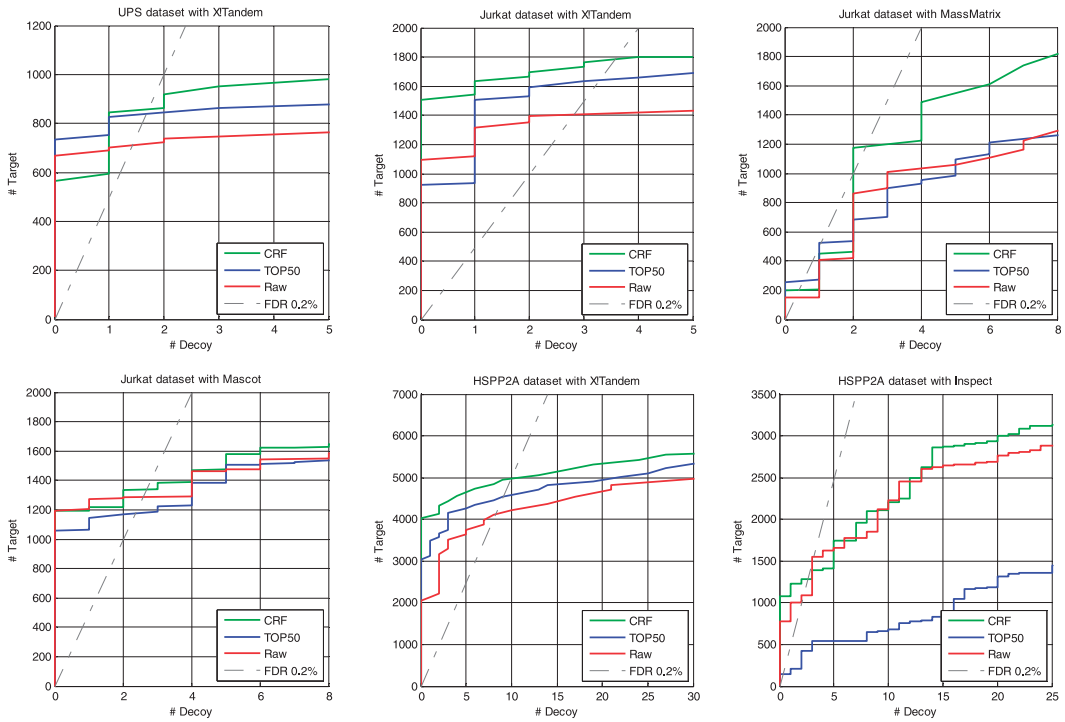

Fig. 5. ROC analysis of the spectrum annotation results of the raw data and the CRF and Top50 filtered Jurkat, HSPP2A and the UPS datasets, obtained with X!Tandem, Mascot, InsPecT and MassMatrix. These plots show independent improvements by the CRF filter on various search engines

identification results. The number of the identified peptides at FDR = 0.2% are plotted in Figure 6A as a function of the threshold parameters. We found a large plateau where the spectrum identification reaches its maximum. The greatest spectrum identification was seen when $T_H = 40–90\%$ and $T_L = 60–160\%$. The large plateau illustrates the robustness of CRF because even with a wide variety of settings, the filters are highly efficient at retaining the informative peaks. We also plotted the percentage of kept peaks (data compression) for various values of T_H and T_L (Fig. 6B). These results indicate that CRF is relatively refractory to the values for T_H and T_L, except for extremely low (too stringent) and high (too permissive) values. Therefore, we expect that only a minimal sample-to-sample tuning will be necessary for the use of CRF. Importantly, a substantial decrease in

Table 1. Search results using X!Tandem with CRF, Top50 filtered and raw datasets

Dataset	CRF versus Top50			CRF versus raw		
	CRF only	Common	Top50 only	CRF only	Common	Raw only
Spectra						
Jurkat	182	1604	38	380	1406	28
UPS	57	800	52	148	709	19
Unique peptides						
Jurkat	153	1316	43	306	1163	44
UPS	55	599	47	113	541	34

The numbers represent the number of the identified spectra or the number of the identified unique peptide sequences, as indicated at FDR = 0.2%.

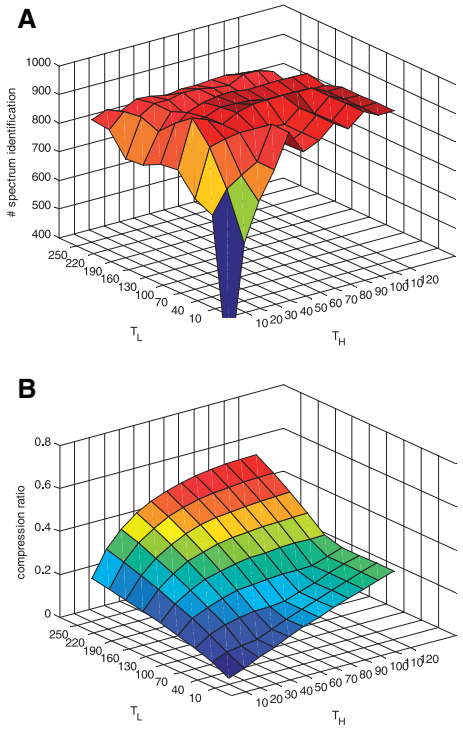

Fig. 6. CRF performance at various (T_H, T_L) parameter pairs on the UPS dataset using X!Tandem using parameters described in Section 2. (**A**) The number of annotated spectra, calculated at FDR = 0.2%. (**B**) Data compression ratio (% of peaks kept)

dataset size accompanies the improvement. Based on these results, we chose $T_H = 50$ and $T_L = 120\%$ as the default values.

CRF has two tolerance parameters: Δ_F is the fragment ion mass (m/z) tolerance and Δ_P is the precursor ion mass (m/z) tolerance; their role is briefly discussed above (end of Section 3.2). The values of Δ_F and Δ_P depend on the accuracy of the mass measurements and are to be set by the experimenter. In our case, we used $\Delta_F = 0.3$ Da and $\Delta_P = 0.3$ Da for the UPS dataset and $\Delta_F = 0.3$ Da and $\Delta_P = 0.03$ Da for the Jurkat dataset.

3.5 Post-translational and chemical modifications

Identification of post-translational and chemical modifications is becoming an increasingly important part of the analysis of proteomics data and accounting for modifications plays a fundamental role in MS data (Tanner *et al.*, 2005). To handle modified peaks, CRF uses an amino acid table that is adjustable for both complete and partial modifications. We studied this problem on the UPS dataset in which the modifications were accounted for during database searching: carbamidomethylation (57.02 Da on Cys) as a fixed modification, deamidation (0.98 Da on Asn and Gln) and oxidation (15.99 Da on Met) both as partial modifications. In the raw dataset, X!Tandem annotated 735 spectra at FDR = 0.2%. When CRF was used without specifying modifications, 864 spectra were identified; when the above modifications were applied during filtering, the number of annotated spectra increased to 866. In other words, the increase with respect to the raw data is substantial regardless of whether the modified amino acids are correctly specified during filtering. We checked the shifted peaks in the spectra (details not shown) and found that the marginal increase in the number of identified spectra (866 versus 864) was due only to a few shifted peaks. This may appear somewhat counterintuitive; however, we note that amino acid modifications do not change most of the inter-peak distances on which CRF is based. So CRF will detect spectra of peptides containing modified residues well even if the modifications are not, or incorrectly, specified (also see Supplementary Materials).

4 DISCUSSION

We have developed a spectrum filtering method based on the gas-phase chemistry of peptide fragmentation and we therefore refer to this strategy as CRF. The main goal of this filtering strategy is to enrich for peaks that most search engines recognize as coming from fragmented peptides. We have used a series of rules: one based on finding b–y complementary pairs, one based on finding peaks separated by the mass of an amino acid residue (amino acid neighbors) and a third one that is a combination of the first two. We found that CRF filtering improves peptide annotation by 15–25%, at the same FDR level, and provides an ~75% compression of the data.

We have tested the CRF on data collected with MALDI-TOF/TOF, ion trap and Orbitrap ion trap hybrid instruments. We have found that the CRF approach also improves the search engine performance with respect to these data. However, data compression and search engine performance gains with ion trap

data were not as great as those achieved with datasets coming from high-resolution instruments. The principle weaknesses with ion trap data are the high mass errors, which requires the mass tolerance to be >1 Da and ambiguities in the charge states. Both these issues cause filters to keep more extraneous peaks in the output. However, even with ion trap data, CRF yielded ~50% data compression and a 10% increase in terms of spectrum annotations.

CRF is most similar to sequence tagging algorithms. However, CRF produces peak list that can be analyzed using standard spectral matching algorithms. A number of groups have published similar algorithms designed to perform quality control checks on each spectra (Flikka *et al.*, 2006; Hoopmann *et al.*, 2007; Nesvizhskii *et al.*, 2006). Spectra that do not have enough amino acid neighbors, or b–y pairs, are typically excluded from the search, while those that pass the quality control check are searched in total. In essence, the rule-based filters are performing a similar quality control check. However, this control is performed at the level of individual peaks, rather than complete spectra. This removes ~75% of all the peaks in the dataset and results in 15–25% more annotated spectra at the same FDR level.

The BYONIC search engine uses a similar strategy, where lookup peaks are extracted from spectra and then searched using a custom algorithm (Bern *et al.*, 2007). Not only does our strategy differ in the rules that are used to select peaks, but we output the selected (and in a separate file, the discarded) peaks in the MASCOT Generic Format (mgf). Searching with the discarded peaks resulted in no legitimate peptide identifications, which indicates the rule-based filters are efficient at retaining the informative peaks and removing noise peaks.

Geer *et al.* (2004) and Renard *et al.* (2009) come close to using CRF, in that they apply their Top *N* filter within a set window surrounding the most intensive peaks. The rational for this filtering is that a true peak should have no more than one true neighbor within 57 Da (the mass of the smallest amino acid), which in the case of a b-ion would be a y-ion (Geer *et al.*, 2004). We have purposefully omitted this type of strategy because the small neutral losses of water and ammonia are often helpful in interpreting spectra and the a-type ion is particularly useful in identifying the b-ion series (Bern and Goldberg, 2006).

The MASCOT search engine uses a different approach to spectral filtering. MASCOT initially searches with a restricted set of high-intensity fragment ions and then iteratively repeats the search using the less intense peaks, until it is clear that the results are no longer being improved. This is incredibly effective at finding the best set of peaks in a spectrum, but it is not computationally efficient and does not allow for data compression.

Unfortunately, the current batch of *de novo* and sequence tagging search algorithms do not allow the output of the peaks they select to be analyzed. Many of these programs use sophisticated peak picking algorithms and they are likely to produce similar, or even greater, improvements than seen with the chemical rule-based filters (Bern *et al.*, 2007; Tanner *et al.*, 2005). However, one advantage of using a simpler algorithm is that it is relatively fast. For example, it takes only 2.8 s to filter 1000 spectra (347 821 peaks) on a PC [with a 3 GHz Intel Core2 processor (Q6850, 3.00 GHz, 8 GB RAM) running under Fedora release 8 Linux operating system]. The time estimate does not include input/output operations.

Because CRF efficiently extracts the peaks coming from peptide fragmentations, PPV and the expectation value of the spectra are increased (Figs 3 and 4, respectively), while the number of random matches is reduced (Fig. 2). Interestingly, the best scoring matches from the unfiltered data typically do not improve as much as more marginal spectra. We think this is largely owing to the fact that the best scoring matches may leave little room for further improvement. The use of the chemical rule-based filters greatly improved the performance of the tested search algorithms as has been shown by ROC analysis (Fig. 5).

CRF is easy to use and its performance compares favorably with the filters based solely on signal-to-noise or raw intensity measures. We hope this approach will be useful for the analysis of large-scale proteomics data.

ACKNOWLEDGEMENTS

The authors thank Dr Somdutta Dhir (ICGEB) for her help and advice.

Funding: The work at ICGEB, Trieste, was supported by the ICGEB PhD program in Molecular Biology (B.R.). The work at BRC, Szeged, was supported by the Informatics PhD Program of the University of Szeged, Hungary (B.R.), and by grant no. TAMOP-4.2.2-08/1/2008-008 from the Hungarian National Office for Research and Technology (NKTH). The work at Pázmány University was partially supported by NKTH grants TET_10-1-2011-0058, TAMOP-4.2.1.B_11/2/KMR-2011-0002 and TÁMOP-4.2.2/B-10/1-2010-0014.

Conflict of Interest: none declared.

REFERENCES

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Altman,D.G. and Bland,J.M. (1994) Diagnostic tests 2: Predictive values. *BMJ*, **309**, 102.

Barlow,C.K. and O'Hair,R.A. (2008) Gas-phase peptide fragmentation: how understanding the fundamentals provides a springboard to developing new chemistry and novel proteomic tools. *J. Mass Spectrom.*, **43**, 1301–1319.

Becker,C.H. and Bern,M. (2011) Recent developments in quantitative proteomics. *Mutat. Res.*, **722**, 171–182.

Bern,M. and Goldberg,D. (2006) De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *J. Comput. Biol.*, **13**, 364–378.

Bern,M. *et al.* (2004) Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics*, **20** (**Suppl. 1**), i49–i54.

Bern,M. *et al.* (2007) Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.*, **79**, 1393–1400.

Biemann,K. (1990) Sequencing of peptides by tandem mass spectrometry and high-energy collision-induced dissociation. *Methods Enzymol.*, **193**, 455–479.

Bish,R.A. *et al.* (2008) Conjugation of complex polyubiquitin chains to WRNIP1. *J. Proteome Res.*, **7**, 3481–3489.

Chalkley,R.J. *et al.* (2005) Bioinformatic methods to exploit mass spectrometric data for proteomic applications. *Methods Enzymol.*, **402**, 289–312.

Deutsch,E.W. *et al.* (2008) Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol. Genomics*, **33**, 18–25.

Flikka,K. *et al.* (2006) Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics*, **6**, 2086–2094.

Geer,L.Y. *et al.* (2004) Open mass spectrometry search algorithm. *J. Proteome Res.*, **3**, 958–964.

Glatter,T. *et al.* (2009) An integrated workflow for charting the human interaction proteome: insights into the PP2A system. *Mol. Syst. Biol.*, **5**, 237.
Hansen,K.C. *et al.* (2003) Mass spectrometric analysis of protein mixtures at low levels using cleavable 13C-isotope-coded affinity tag and multidimensional chromatography. *Mol. Cell. Proteomics*, **2**, 299–314.
Hoopmann,M.R. *et al.* (2007) High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal. Chem.*, **79**, 5620–5632.
Jacob,R.J. (2010) Bioinformatics for LC-MS/MS-based proteomics. *Methods Mol. Biol.*, **658**, 61–91.
Johnson,R.S. *et al.* (2005) Informatics for protein identification by mass spectrometry. *Methods*, **35**, 223–236.
Kersey,P.J. *et al.* (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
Kil,Y.J. *et al.* (2011) Preview: a program for surveying shotgun proteomics tandem mass spectrometry data. *Anal. Chem.*, **83**, 5259–5267.
MacCoss,M.J. (2005) Computational analysis of shotgun proteomics data. *Curr. Opin. Chem. Biol.*, **9**, 88–94.
McDonald,W.H. *et al.* (2004) MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom.*, **18**, 2162–2168.
Menschaert,G. *et al.* (2010) Peptidomics coming of age: a review of contributions from a bioinformatics angle. *J. Proteome Res.*, **9**, 2051–2061.
Mujezinovic,N. *et al.* (2006) Cleaning of raw peptide MS/MS spectra: improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise. *Proteomics*, **6**, 5117–5131.
Nesvizhskii,A.I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics*, **73**, 2092–2123.
Nesvizhskii,A.I. and Aebersold,R. (2004) Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov. Today*, **9**, 173–181.
Nesvizhskii,A.I. *et al.* (2006) Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol. Cell. Proteomics*, **5**, 652–670.
Neumann,S. and Bocker,S. (2010) Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Anal. Bioanal. Chem.*, **398**, 2779–2788.
Noble,W.S. and MacCoss,M.J. (2012) Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput. Biol.*, **8**, e1002296.
Reiz,B. *et al.* (2012) Data preprocessing and filtering in mass spectrometry based proteomics. *Curr. Bioinformatics*, **7**, 212–220.
Renard,B.Y. *et al.* (2009) When less can yield more—computational preprocessing of MS/MS spectra for peptide identification. *Proteomics*, **9**, 4978–4984.
Roepstorff,P. and Fohlman,J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.*, **11**, 601.
Salmi,J. *et al.* (2009) Filtering strategies for improving protein identification in high-throughput MS/MS studies. *Proteomics*, **9**, 848–860.
Sonego,P. *et al.* (2008) ROC analysis: applications to the classification of biological sequences and 3D structures. *Brief. Bioinformatics*, **9**, 198–209.
Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, **100**, 9440–9445.
Tanner,S. *et al.* (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.*, **77**, 4626–4639.
Webb-Robertson,B.J. and Cannon,W.R. (2007) Current trends in computational inference from mass spectrometry-based proteomics. *Brief. Bioinformatics*, **8**, 304–317.
Xu,H. and Freitas,M.A. (2010) A dynamic noise level algorithm for spectral screening of peptide MS/MS spectra. *BMC Bioinformatics*, **11**, 436.
Yates,J.R., 3rd *et al.* (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.*, **67**, 1426–1436.