BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i49–i54
DOI: 10.1093/bioinformatics/bth947

Automatic Quality Assessment of Peptide Tandem Mass Spectra

Marshall Bern[1], David Goldberg[1,], W. Hayes McDonald[2] and John R. Yates, III[2]*

[1]Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA and [2]The Scripps Research Institute, 10440 North Torrey Pines Road, La Jolla, CA 92037, USA

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: A powerful proteomics methodology couples high-performance liquid chromatography (HPLC) with tandem mass spectrometry and database-search software, such as SEQUEST. Such a set-up, however, produces a large number of spectra, many of which are of too poor quality to be useful. Hence a filter that eliminates poor spectra before the database search can significantly improve throughput and robustness. Moreover, spectra judged to be of high quality, but that cannot be identified by database search, are prime candidates for still more computationally intensive methods, such as *de novo* sequencing or wider database searches including post-translational modifications.

Results: We report on two different approaches to assessing spectral quality prior to identification: binary classification, which predicts whether or not SEQUEST will be able to make an identification, and statistical regression, which predicts a more universal quality metric involving the number of b- and y-ion peaks. The best of our binary classifiers can eliminate over 75% of the unidentifiable spectra while losing only 10% of the identifiable spectra. Statistical regression can pick out spectra of modified peptides that can be identified by a *de novo* program but not by SEQUEST. In a section of independent interest, we discuss intensity normalization of mass spectra.

Contact: goldberg@parc.com

1 INTRODUCTION

Proteomics studies the entire complement of proteins in a biological system, such as a cell or tissue, with the aim of understanding the workings of the system in various states. The techniques of proteomics (Liebler, 2001) involve a sequence of complex steps, such as protein separation, digestion and identification, which must be developed and optimized together in a 'systems approach' in order to extract the maximum amount of information from the entire pipeline. In this paper, we address the problem of improving the throughput of peptide identification by tandem mass spectrometry (Aebersold and Goodlett, 2001). We describe algorithms for assessing the quality of a tandem mass spectrum before attempting to identify the peptide. This algorithm can be used to prefilter spectra so that only reasonably good spectra are sent to time-consuming, database-search identification programs, such as SEQUEST (Eng *et al*., 1994) and Mascot (Perkins *et al*., 1999). The algorithm can also be used as a post-filter to identify high-quality spectra that warrant even more time-consuming analysis, such as SEQUEST with a database of post-translational modifications (MacCoss *et al*., 2002b), partial sequence identification using GutenTag (Tabb *et al*., 2003b), or fully *de novo* sequencing using programs, such as Lutefisk (Taylor and Johnson, 2001). We report below on successful *de novo* sequencing of spectra that could not be recognized by SEQUEST, a reversal of the usual situation in which database-search methods outperform *de novo* methods.

In a previous related work, Tabb *et al*. (2001) discuss spectral quality assessment and mention a number of simple rules for prefiltering, such as minimum and maximum thresholds on number of peaks and a minimum threshold on total peak intensity. They state that such rules can remove 40% or more of the bad spectra. The best algorithm described here can remove 75% of the bad spectra while losing only 10% of the high-quality (identifiable) spectra. Interestingly, the number of peaks and their intensities—often used by experts to 'eyeball' spectra—had little classification power relative to more detailed features such as the number of peak pairs differing by amino acid masses. Thus, we find that quality assessment is more easily done by a machine than by a human expert.

Finally, we note that a loss of 10% of the peptide identifications incurs a smaller loss in the number of protein identifications. In a large-scale study of the *Chlamydia* proteome, the filter of Section 2.2—applied in series after a simple rule-based filter—lost only 5% of the correct peptides and 3% of the correct protein identifications. It removed an additional 44% of the bad spectra beyond those removed by the simple filter, thus improving computer throughput by almost a factor of two, and—surprisingly—reduced the number of incorrect (non-*Chlamydia*) peptide and protein identifications (by 8 and 12%, respectively) when searching against a large, multispecies 'distractor' database.

*To whom correspondence should be addressed.

2 ALGORITHM DEVELOPMENT

We obtained 68 978 tandem mass spectra from a known mixture of five proteins (rabbit phosphorylase a, horse cytochrome c, horse apomyoglobin, bovine serum albumin and bovine β-casein), digested with four different proteases (trypsin, elastase, subtilisin and proteinase K), as described previously (MacCoss *et al*., 2002a). Of the 68 978 spectra, 5678 were labeled GOOD, meaning that they were matched by SEQUEST searching against the NCBI non-redundant protein database with 907 654 entries, to one of the five proteins in the mixture or to a likely contaminant such as keratin or one of the enzymes used for digestion. For the purposes of this study, the other 63 300 spectra were labeled BAD, although some of these are high-quality spectra of variant or modified peptides. Such a large proportion of BAD spectra is typical of Multidimensional Protein Identification Technology (Washburn *et al*., 2001), in which peptides eluted by two-dimensional liquid chromatography are electrosprayed continually into a mass spectrometer. The MS instrument used for these spectra (LCQ-Deca, ThermoFinnigan) is an ion-trap instrument with a lower m/z (mass over charge) cut-off ∼200–300 Da, and a resolution of ∼0.3 Da at $m/z \sim 1000$. (Here and elsewhere we informally write Da instead of Daltons per unit charge.)

Broadly speaking there are two competing approaches to developing an automatic classifier. The traditional approach devises a number of handcrafted features incorporating human knowledge; whereas, the more modern approach feeds less processed, high-dimensional data into a classifier algorithm, such as support vector machines (SVMs), that can in effect learn features from the data. We tried both approaches, reporting on handcrafted features in Section 2.2 and SVMs in Section 2.3. For the regression problem (Section 2.4) of predicting a continuous quality metric rather than simply GOOD or BAD, we reused the handcrafted features rather than attempting to learn features. Before describing these experiments, however, we delve into an issue common to all MS/MS analysis problems.

2.1 Intensity normalization

Intensity of peaks is widely recognized as highly variable from spectrum to spectrum (Havilio *et al*., 2003). Consequently there is no agreed-upon way to incorporate intensity information into algorithms. SEQUEST (Eng *et al*., 1994) uses only the largest 200 peaks and scores only the presence/absence of peaks, using two different constants for b- and y-ions. Havilio *et al*. (2003) develop an intensity-based scoring algorithm and claim significant improvement over SEQUEST. Intensity-based scoring, however, is not easy. Raw intensities are too variable to be used, with maximum and total intensities varying over two or three orders of magnitude within the GOOD data. Relative intensities (i.e. raw intensities divided by total intensity) as used by Havilio *et al*. are better, yet still too variable, because a single strong peak or a low background of noise peaks often shifts values by a factor of two or three.

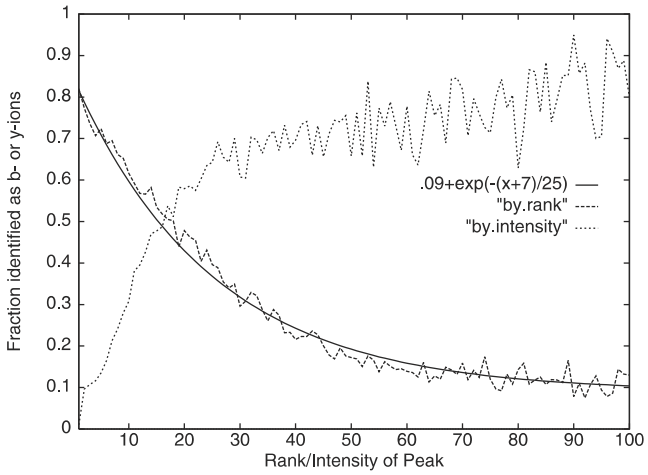

Fig. 1. The bumpy increasing line gives the probability that a peak of a given relative intensity turns out to be a b- or y-ion. For this line the x-axis is in hundredths of percentage, that is, 50 means 0.5% of the total ion intensity is in this peak. (The bin size was picked to give a curve that runs over roughly the same 0.1–0.8 range as the rank curve.) The y-axis shows $(\#b + \#y)/(\#b + \#y + \#?)$, where $\#b$ is the number of b-ion peaks of a given intensity (out of 1416 identified spectra), $\#y$ is the number of y-ion peaks and #? is the number of unidentified peaks. Other identified peaks (isotopes, a-ions, water or ammonia losses, internal fragments) were not counted in the probability. The less bumpy decreasing curve gives the probability that a peak of a given rank (rank 1 = most intense) turns out to be a b- or y-ion. The smooth curve is an exponential function shown for comparison. The fact that rank gives a less bumpy curve than relative intensity argues for improved (lower variance) probability estimation from rank.

For maximum robustness, we chose to use rank-based intensity normalization rather than relative intensities, where the most intense peak has rank 1, the second most intense has rank 2, and so forth. Figure 1 compares how well rank and relative intensities correlate with an a posteriori measure of peak quality, computed on the GOOD spectra in the training set: the probability that the peak is a b- or y-ion. Each spectrum has peaks of all ranks (at least up to rank 200 or so) but spectra differ considerably in relative intensities, and hence estimation of probability from rank has much lower variance than estimation from relative intensity. This advantage of rank over intensity extends to probability-based scores and features.

Moreover, Figure 1 justifies a particularly simple way to use ranks. The plot of rank versus probability fits a negative exponential function quite well. Thus the contribution of peak x to a probabilistic scoring function as advocated in the literature (Bafna and Edwards, 2001; Dančik *et al*., 1999; Havilio *et al*., 2003; Tabb *et al*., 2003a) should be proportional to a constant plus $1/\mathrm{Rank}(x)$, in order that a sum of contributions is equal to a constant plus the log-likelihood that the peaks in the sum are indeed b- and y-ions.

2.2 Classification using handcrafted features

Following the discussion above, our handcrafted features all use a normalized intensity of the form

$$\mathrm{Norm}I(x) = \max\{0, C_1 - (C_2/\mathrm{Max}mZ) \cdot \mathrm{Rank}(x)\},$$

where MaxmZ is the maximum significant m/z-value in the spectrum, and C_1 and C_2 are constants. The MaxmZ term means that generally more peaks are considered for longer peptides.

We learned values for C_1 and C_2 for each feature separately, by picking the C_1 and C_2 values that gave the best discrimination between GOOD and BAD in the training set. For example, $C_1 = 28$ and $C_2 = 400$ for the Good-Diff Fraction feature, meaning that Norm$I(x)$ is greater than zero if $\mathrm{Rank}(x) \leq 140$ when Max$mZ = 2000$, a typical value. Generally C_1 and C_2 were about the same for different features, with the exception of the Isotopes feature which used peaks of much lower rank. Evidently, the fact that a peak has appropriate m/z and intensity relative to another peak increases the likelihood that the peak is meaningful.

Each spectrum is mapped to a feature vector $(f_1, f_2, \ldots, f_7)$, a point in $\mathbb{R}^7$, where f_i is the value of the i-th feature below.

(1) *Npeaks.* The number of peaks in the spectrum. This feature is often recommended (Kinter and Sherman, 2000; Tabb *et al.*, 2001) for human assessment of spectrum quality.

(2) *Total Intensity.* The sum of the raw intensities of the peaks in the spectrum.

(3) *Good-Diff Fraction.* This feature measures how likely two peaks are to differ by the mass of an amino acid. Let

$$\mathrm{GoodDiffs} = \sum\big\{\mathrm{Norm}I(x) + \mathrm{Norm}I(y) | M(x) - M(y) \approx M_i \quad \text{for some } i = 1, 2, \ldots, 20\big\},$$

where $M(x)$ is the m/z-value of peak x and $M_1, M_2, \ldots, M_{20}$ are the amino acid masses (not all of which are unique). The comparison implied by $\approx$ uses a tolerance, which was set to 0.37 Da for our ion-trap spectra. Now let

$$\mathrm{TotalDiffs} = \sum\big\{\mathrm{Norm}I(x) + \mathrm{Norm}I(y) \mid 56 \leq M(x) - M(y) \leq 187\big\}.$$

Then $f_3 = \mathrm{GoodDiffs}/\mathrm{TotalDiffs}$.

(4) *Isotopes.* The total normalized intensity of peaks with associated isotope peaks. That is,

$$\sum\big\{\mathrm{Norm}I(x) \mid M(x) \approx M(y) - 1 \text{ and } I(x) \approx \text{Expected Intensity of } +1 \text{ Isotope}\big\}.$$

(5) *Complements.* The total normalized intensity of pairs of peaks with m/z-values summing to the mass of the parent ion. The feature is computed assuming both $+2$ and $+3$ charge states for the parent ion (i.e. two different M_{Parent} masses) and the larger feature value is used; the same technique is used in the program 2–3 to determine charge state (Sadygov *et al.*, 2002).

$$\sum\big\{\mathrm{Norm}I(x) + \mathrm{Norm}I(y) \mid M(x) + M(y) \approx M_{\mathrm{Parent}}\big\}.$$

(6) *Water Losses.* The total normalized intensity of pairs of peaks with m/z-values differing by 18 Da.

$$\sum\big\{\mathrm{Norm}I(x) + \mathrm{Norm}I(y) \mid M(x) - M(y) \approx 18\big\}.$$

(7) *Intensity Balance.* The m/z range is divided into 10 equal-width bands between 300 Da and the largest observed m/z. The feature is the total raw intensity in the two bands with greatest intensity minus the total raw intensity in the seven bands with lowest intensity.

For classification we used Quadratic Discriminant Analysis (QDA), a classical method that models feature vectors of each class by multivariate Gaussian distributions and thus determines quadratic decision boundaries between GOOD and BAD. This simple method tends to perform surprisingly well (Hastie *et al.*, 2001), especially with summation features such as ours that have approximate Gaussian distributions due to the central limit theorem.

We trained two separate classifiers, one for singly charged parent ions and one for multiply charged. Training a QDA classifier involves computing the means and covariance matrix for the features; we removed outlying feature vectors (if the value of any feature fell in the top or bottom 1% for that feature) in order to make the fitting more robust. For feature selection, we tested all subsets of the set of features, and chose the one that gave the best binary classification performance on the training set (one-fourth of GOOD and one-eighth of BAD). We imposed an Occam's razor: a subset of features was preferred if its percentage of correct classifications (both GOOD and BAD) was within 0.5% that of the superset. We adjusted the threshold on the decision surface (an isosurface for probability ratio) so that 90% of the GOOD spectra were classified good; users could of course adjust this threshold depending upon their requirements, e.g. using less aggressive filtering for one-dimensional high-performance liquid chromatography (HPLC), which does not produce as many spectra as two-dimensional HPLC. In developing the classifiers, we did not use the test set until reporting final results, a purity of approach afforded by the great amount of data.

The binary classifier for the singly charged spectra used four features: Good-Diff Fraction, Complements, Water Losses and Balance. The binary classifier for the multiply charged spectra used four slightly different features: Good-Diff Fraction, Isotopes, Water Losses and Balance. The results on the

	Called Good	Called Bad	% Correct
+1 GOOD	671	75	89.9%
+1 BAD	5585	11475	67.3%
+2/+3 GOOD	3166	348	90.1%
+2/+3 BAD	11611	26684	69.7%
All GOOD	3837	423	90.1%
All BAD	17196	38159	68.9%

Fig. 2. The results with handcrafted features. True classifications are on the left; e.g. 89.9% of the singly charged GOOD spectra were called Good by our binary classifier.

test set (3/4 of GOOD and 7/8 of BAD) are given in Figure 2. Error rates on the test set were essentially identical to those on the training set. The classification problem for spectra from singly charged parent ions is slightly more difficult than for multiply charged parent ions, due to the generally poor fragmentation of singly charged parent ions (Kinter and Sherman, 2000).

A binary classifier that uses only Npeaks and Total Intensity—the two features most often used by experts in quick manual assessment—gives much weaker results: only 54% rejection of BAD spectra when 90% of the GOOD spectra are classified good.

2.3 Classification with SVMs

Motivated by the success of features involving m/z differences between peaks (Good-Diff Fraction, Isotopes, etc.), we used a histogram of m/z differences as the input to an SVM classifer. For our first SVM experiment, we created from each spectrum a vector of length 187 (the maximum mass of an amino acid residue) with bins for m/z differences of [0.5, 1.5], [1.5, 2.5], and so forth up to [186.5, 187.5]. The entry in histogram bin i is a sum over all peak pairs in the spectrum:

$$\mathrm{Hist}(i) = \sum \big\{ \min\{1/\mathrm{Rank}(x), 1/\mathrm{Rank}(y)\} | M(x) - M(y) \in [i-0.5, i+0.5] \big\}.$$

This expression differs from Good-Diff Fraction in using $\min\{1/\mathrm{Rank}(x), 1/\mathrm{Rank}(y)\}$ rather than $\mathrm{Norm}I(x) + \mathrm{Norm}I(y)$. The difference between the expressions $1/\mathrm{Rank}(x)$ and $1/\mathrm{Norm}I(x)$ is inconsequential here, just shifting everything by a linear transformation. There is a difference between the sum and the minimum; we chose the minimum because it gave better SVM classification performance. We also tried using raw intensities instead of $1/\mathrm{Rank}(x)$ in order to test whether intensity normalization is even necessary for SVM input data; perhaps the SVM can learn an even better normalization. We found that $1/\mathrm{Rank}(x)$ normalization was helpful after all, improving classification performance by 2–3%.

For the SVM experiment, which takes significant training and testing time, we did not separate the spectra into singly and multiply charged data sets. We used SVM-Light (Joachims, 1999) and trained on 1/4 of the GOOD spectra and 1/32 of the BAD spectra; ∼30% of the training vectors ended up as support vectors. To speed up the experiments, we tested on three-fourth of the GOOD data and only one-fourth of the BAD. We used radial basis functions, and experimented to find a good value (500) for gamma, the width parameter of the basis functions. We used the default penalty value for training set errors, and we adjusted the relative costs of the two types of errors in order to obtain 90% correct classification of the GOOD spectra.

Figures 3 and 4 gives results for our SVM experiments. In addition to difference histograms with 1-Da bins from 1 to 187, we also tried some larger difference histograms: 1-Da bins from 1 to 384 and 0.5-Da bins from 1 to 187. The SVM approach gives appreciably better results than the handcrafted-feature approach, with performance improving

	Called Good	Called Bad	% Correct
1-Da bins, 1 to 187			
GOOD	3833	427	90.0%
BAD	4062	11738	74.3%
1-Da bins, 1 to 374			
GOOD	3835	425	90.0%
BAD	3894	11906	75.9%
0.5-Da bins, 1 to 187			
GOOD	3835	425	90.0%
BAD	3940	11860	75.1%

Fig. 3. The results with SVM classifiers.

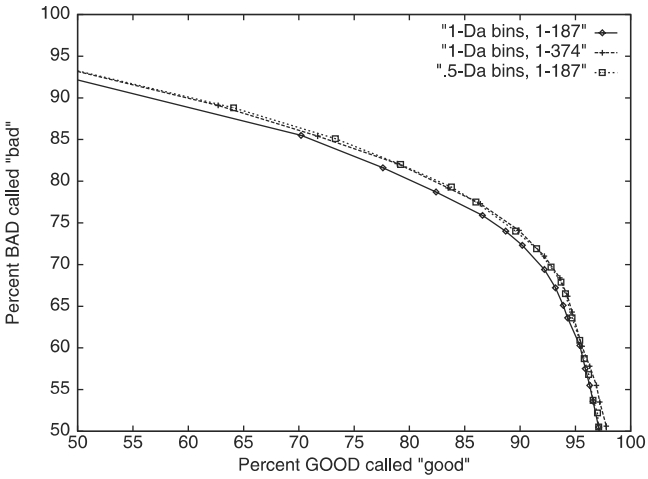

Fig. 4. Receiver operator characteristic (ROC) curves for the SVM classifiers show the trade off between false positives and false negatives. For example, if 15% loss of GOOD spectra is acceptable, then almost 80% of the BAD spectra can be removed, but if 5% loss of GOOD spectra is the maximum acceptable, then only ∼60% of the BAD spectra can be removed. (Numbers do not exactly match Fig. 3, because the width parameter gamma for the radial basis function kernel was changed in order to make more complete ROC curves.)

slightly with increasing size of input vectors. Of course the running time becomes quite a bit slower as the size increases. In general, the SVM classifiers are slower than the QDA classifiers, although not as slow as running SEQUEST itself. The fastest SVM classifier (1-Da bins from 1 to 187) takes 362 s to process 20 000 spectra, whereas the QDA classifier takes 114 s to process the same spectra. SEQUEST takes ∼1 s per spectrum using a small (1 MB) database and ∼15 s per spectrum on a large (100 MB) database.

2.4 Regression

A binary classifier is sufficient for filtering spectra in order to improve SEQUEST throughput, but we are also interested in the problem of assigning a numerical quality score to each spectrum, in order to prioritize the high-quality unidentified spectra for further processing. This is a regression problem, as it attempts to predict a continuous measure rather than a binary variable.

We defined the continuous measure of quality to be the fraction of b- and y-ions observed among the peaks of high intensity. More specifically, letting Length denote the number of amino acids in the peptide, we define

$$\text{Quality} = \tfrac{1}{2}(\#b + \#y)/(\text{Length} - 1),$$

where $\#b$ is the number of b-ion peaks with rank $< 6 \cdot \text{Length}$ and $\#y$ is the number of y-ion peaks with rank $< 6 \cdot \text{Length}$. We can compute this measure with an *a posteriori* analysis of the GOOD spectra. We experimented with other definitions of Quality, e.g. an analogous definition using normalized intensity rather than simply presence/absence of peaks, and another definition that penalized for unidentified peaks. The various definitions of Quality gave similar results. We settled on the definition above because it is most interpretable by humans; the feature runs from 0 to 1.0, from no b- and y-ions observed to all possible b- and y-ions observed. In addition, many peptide identification programs, both database-search and *de novo*, rely on presence/absence of b- and y-ions rather than some sort of normalized intensity.

We next ran a multivariate linear regression with the seven handcrafted classification features as explanatory variables and Quality as the response variable, in order to determine a linear combination of the features that is predictive of spectrum quality. [We used the handcrafted features rather than SVM regression (Vapnik, 1996), because our interest was in proof of concept rather than performance numbers, which we had no good means to assess.] The multivariate linear regression gave only two of the classification features (Good-Diff Fraction and Complements) highly significant non-zero coefficients as judged by P-values. The R^2 value for the regression was 0.537, which means that the linear combination has correlation coefficient $\sqrt{0.537} \approx 0.73$ with Quality, not overwhelming but certainly high enough to be useful.

Sequence	X-corr
[430.2]GSTWW[210.2]EMDKEACFA[154.1]AFR	.809
[430.2]GSTWW[210.2]EMDKEACFAVE[154.1]K	.789
[430.2]GSDGDW[211.1]KMDKEACFAVE[154.1]K	.781
[430.2]GSDGDW[211.1]KMDKEACAFVE[154.1]K	.756
[168.1][262.1]GSTWW[210.2]EMDKEACFAVE[154.1]K	.800

Fig. 5. Top five Lutefisk identifications for the best BAD spectrum.

The regression identified thousands of BAD spectra with predicted Quality scores better than the average Quality of GOOD spectra, which was ∼0.28, meaning that only 28% of all possible b- and y-ions appeared among the best-ranking peaks in the spectrum. We submitted the six best BAD spectra (all with predicted Quality over 0.44) to Lutefisk (Taylor and Johnson, 2001), a *de novo* peptide sequencer. On two of the six spectra, Lutefisk gave partial sequences that could be uniquely matched by BLAST to bovine serum albumin. Figure 5 illustrates one of these successes; a bracketed number indicates a 'mass gap', meaning unidentified residues, possibly with modifications, totaling that mass.

A BLAST search with MDKEACFAVE gives a match with bovine serum albumin, which has a subsequence of ENFVAFVDKCCAADDKEACFAVEGPK. The letters GP perfectly fill the mass gap of 154.1 Da, so we could be fairly confident of the identification even without knowing that bovine serum albumin was one of the proteins in the mixture. No suffix of the correct sequence ENFVAFVDKCCAAD, however, sums to the same mass as [430.2]GSTWW[210.2]EM, which means that all the peaks in the spectrum are shifted from where they should be in an unmodified peptide from bovine serum albumin. (Indeed Lutefisk recognized DKEACFAVE on the basis of a ladder of y-ion peaks, with no help from b-ions.) Thus this spectrum is likely to be from a modified or variant peptide.

3 CONCLUSION

The example of finding a modified peptide by using our quality regression program, along with SEQUEST and Lutefisk, illustrates how data analysis for proteomics depends upon a suite of tools. We believe that spectral quality assessment can play an important supporting role in a tool suite, maximizing throughput and applicability of the more central tools. Indeed the QDA filter of Section 2.2 is already in use in real-world proteomics at the Scripps Research Institute. Quality filters can help mine valuable information from very noisy data; e.g. in the proteomics set-up described here over 90% of the spectra are unidentifiable by SEQUEST. Finally, we believe that quality assessment can also play a significant role in tool development. Comparison of peptide identification approaches would be enabled by a standard way to measure the quality of spectra in different data sets.

REFERENCES

Aebersold,R. and Goodlett,D.R. (2001) Mass spectrometry in proteomics. *Chem. Rev.*, **101**, 269–296.

Bafna,V. and Edwards,N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, **17**, S13–S21.

Dančik,V., Addona,T.A., Caiser,L.R., Vath,J.E. and Pevzner,P.A. (1999) *De novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.*, **6**, 327–342.

Eng,J.K., McCormack,A.L. and Yates,J.R.,III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976–989.

Field,H.I., Fenyö,D. and Beavis,R.C. (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimizes protein identification, and archives data in a relational database. *Proteomics* **2**, 36–47.

Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Havilio,M., Haddad,Y. and Smilansky,Z. (2003) Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem.*, **75**, 435–444.

Joachims,T. (1999) Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge, MA.

Keller,A., Purvine,S., Nexvizhskii,A.I., Stolyar,S., Goodlett,D.R. and Kolker,E. (2002) Experimental protein mixture for validating tandem mass spectral analysis. *OMICS: J. Integr. Biol.*, **6**, 207–212.

Kinter,M. and Sherman,N.E. (2000) *Protein Sequencing and Identification using Tandem Mass Spectrometry.* John Wiley and Sons, New York.

Liebler,D.C. (2001) *Introduction to Proteomics: Tools for the New Biology*. Humana Press, Totowa, NJ.

MacCoss,M.J., Wu,C.C. and Yates,J.R.,III (2002a) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.*, **74**, 5593–5599.

MacCoss,M.J., McDonald,W.H., Saraf,A., Sadygov,R., Clark,J.M., Tasto,J.J., Gould,K.L., Wolters,D., Washburn,M., Weiss,A. Clark,J.I. and Yates,J.R.,III (2002b) Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc. Natl Acad. Sci. USA*, **99**, 7900–7905.

Perkins,D.N., Pappin,D.J.C., Creagy,D.M. and Cottrell,J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Sadygov,R.G., Eng,J., Durr,E., Saraf,A., McDonald,H., MacCoss,M.J. and Yates,J.R.,III (2002) Code developments to improve the efficiency of automated MS/MS spectra interpretation. *J. Proteome Res.*, **1**, 211–215.

Tabb,D.L., Eng,J.K. and Yates,J.R.,III. (2001) Protein identification by SEQUEST. In P. James, (ed.), *Proteome Research: Mass Spectrometry*. Springer, Berlin.

Tabb,D.L., Smith,L.L., Breci,L.A., Wysocki,V.H., Lin,D. and Yates,J.R.,III. (2003a) Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem.*, **75**, 1155–1163.

Tabb,D.L., Saraf,A. and Yates,J.R.,III (2003b) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.*, **75**, 6415–6421.

TaylorJ.A. and Johnson,R.S. (2001) Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Anal. Chem.*, **73**, 2594–2604.

Vapnik,V. (1996) *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Washburn,M.P., Wolters,D. and Yates,J.R.,III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.*, **19**, 242–247.