Bioinformatics Advance Access published December 13, 2005

BIOINFORMATICS

Quality classification of tandem mass spectrometry data

Jussi Salmi[1,*], Robert Moulder[2], Jan-Jonas Filén[2,3], Olli S. Nevalainen[1], Tuula A. Nyman[4], Riitta Lahesmaa[2] and Tero Aittokallio[2,5]

[1]Department of Information Technology and Turku Centre for Computer Science, University of Turku, Finland

[2]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Finland

[3]The National Graduate School in Informational and Structural Biology, Finland

[4]Finnish Institute of Occupational Health, Helsinki, Finland

[5]Department of Mathematics, University of Turku, Finland

ABSTRACT

Motivation: Peptide identification by tandem mass spectrometry is an important tool in proteomic research. Powerful identification programs exist, such as SEQUEST, ProICAT and Mascot, which can relate experimental spectra to the theoretical ones derived from protein databases, thus removing much of the manual input needed in the identification process. However, the time-consuming validation of the peptide identifications is still the bottleneck of many proteomic studies. One way to further streamline this process is to remove those spectra that are unlikely to provide a confident or valid peptide identification, and in this way to reduce the labour from the validation phase.

Results: We propose a prefiltering scheme for evaluating the quality of spectra before the database search. The spectra are classified into two classes: spectra which contain valuable information for peptide identification and spectra that are not derived from peptides or contain insufficient information for interpretation. The different spectral features developed for the classification are tested on a real-life material originating from human lymphoblast samples and on a standard mixture of 9 proteins, both labelled with the ICAT reagent. The results show that the prefiltering scheme efficiently separates the two spectra classes.

Availability: The software tools are available on request from the authors.

Contact: jussi.salmi@it.utu.fi

Supplementary information:The Mascot ion score distributions and the C4.5 classification rules can be found at address http://staff.cs.utu.fi/staff/jussi.salmi/Supplementary_material.pdf

1 INTRODUCTION

In most proteomic studies protein identification is achieved using mass spectrometry (MS) with database searches. Proteins are first digested into smaller fragments, peptides, which are then analysed with MS. Amongst the different types of MS-methods and instruments available for protein identification, liquid chromatography (LC) coupled with tandem mass spectrometry (MS/MS) is a standard method for identifying proteins from complex mixtures of proteins. In LC-MS/MS-analysis peptides are firstly separated using LC, followed with measurement of their mass-to-charge (*m/z*) ratios and fragmentation by the mass spectrometer. Protein identifications are made by interpreting specific pattern of ions in the resulting MS/MS spectrum. Identification is most often achieved using a database search program such as SEQUEST (Eng *et al.*, 1994) or Mascot (Perkins *et al.*, 1999). These programs compare the MS/MS spectra obtained from unknown peptides with theoretically predicted spectra derived from protein databases. As the technique is relatively reliable and suitable for automation, it allows the high throughput required in proteomics experiments. Alternatively, *de novo* sequencing may be used to reconstruct peptide sequences from the peaks in the spectrum, without direct reference to a database.

Even in the ideal case, identification of peptides by their MS/MS spectrum cannot be done with a 100% certainty, as fragments of different origin (e.g. b- and y-fragments) may overlap to any degree, so two different sequences may give indistinguishable (in terms of measured *m/z* of fragments) MS/MS spectra. In practice, however, even if the sequence of a fragmented peptide is present in the database its identification can be problematic, because the obtained peak patterns are not always equivalent to the theoretical ones. Besides missing peaks, there may be peaks in the spectrum which are not consistent with theoretical fragmentation. Such artefacts complicate the analysis process and compromise the reliability of automatic identifications. Database search programs need to rank the theoretical spectra according to the likelihood of correct match with the sample spectrum using complicated scoring schemes, see e.g. the comparison by Chamrad *et al.* (2004) and the references therein. Consequently, reliable protein identification with search algorithms is a time-consuming and program-dependent task (Moulder *et al.*, 2005) involving a considerable frequency of false positive identifications (Keller *et al.,* 2002b; Cargile *et al.,* 2004).

During the past few years, there have been a number of studies concerning the interpretation of the results of various search programs to distinguish correct peptide identifications from false positives. In particular these have addressed the SEQUEST-calculated features combined with additional parameters extracted from the spectra to predict correctly and incorrectly assigned peptides. Machine learning methods for such classifications

*To whom correspondence should be addressed.

include expectation maximization algorithm (Keller *et al.*, 2002a), support vector machines (Anderson *et al.*, 2003), and a combination of a neural network system and a statistical score (Razumovskaya *et al*., 2004). Motivated by manual interpretation rules for a peptide match, Sun *et al.* (2004) described two scores based on the rules that intensity of the fragment ions should be clearly above baseline noise, and that b- or y-ion series should be continuous in the spectrum. They complemented the scores with a filter based on the SEQUEST-outputs (Xcorr and DeltaCn scores), and tested the method by analyzing spectra produced from a known mixture of 18 proteins with an ion trap mass spectrometer. Compared to pure SEQUEST results, this method increased the number of positive identifications without increasing the number of negative identifications (Sun *et al.,* 2004).

Despite the recent advances in estimating the confidence of peptide identifications, the eventual decision whether the software-made identifications are correct and complete often relies on subjective, experience-dependent manual verification. One of the principal reasons is that there frequently exist poor-quality spectra that can give a nearly perfect peptide match by pure chance alone. As LC-MS/MS can produce hundreds of fragment ion spectra for a particular protein mixture in one hour, it is evident that the labour-intensive manual validation of the identifications is the bottleneck of the analysis pipeline. Therefore, there is a great need for automated MS/MS spectrum interpretation methods for objective determination of the quality of given spectra before the database search. Removal of bad quality spectra prior to the database search not only improves throughput, but also reduces the risk of false positive identifications. The challenge is to find the spectral features which could be used to classify spectra into the ones containing valid peptide information or noise only. As the classification system can not separate the two classes completely, there is inevitably a trade-off between incorporating as many good spectra and removing as many bad spectra as possible.

Previously, Bern *et al.* (2004) used a set of features in their spectra quality classification to filter MS/MS spectra before identifying them with SEQUEST. They reported achieving the specificity (true negative rate, TNR) of 67-70% at 90% sensitivity (true positive rate, TPR) in binary classification with quadratic discriminant analysis which predicts whether or not SEQUEST can make the peptide identification. Their test data consisted of a known mixture of 5 proteins processed with an ion-trap mass spectrometer, and they considered only those spectra good, which matched in the NCBI protein database with one of those proteins, with their known contaminants, or with enzymes used in sample preparation (Bern *et al.* 2004).

Purvine *et al.* (2004) used a prefilter with 3 features for MS/MS spectra classification. One feature addressed the uncertainty in charge state assignments, the second was based on total signal intensity, and the third on a signal-to-noise estimate. Excellent results were demonstrated with data produced with an ion trap mass spectrometer for the analysis of a standard protein/peptide test mixture. By adjusting the feature-specific thresholds on this material only, they obtained the specificity of 98% and sensitivity of 96.8%.

Recently, Savitski *et al.* (2005) complemented the Mascot-output (the M-score) with a database-independent scoring scheme (called S-score), in their test material, the prefiltering based on the maximum length of peptide sequence tag allows for removing 39% of the spectra before database identification phase. Their scoring scheme can also be used to classify Mascot-identified spectra to several reliability classes, and the combined use of the M- and S-scoring schemes provides a 40% improvement in peptide identification as compared to M-scoring alone. However, their approach and subsequent results were strongly dependent of the high mass accuracy provided by the Fourier transform ion trap mass spectrometer.

Data pre-processing filters have also been used to streamline data analysis in studies with automated *de novo* sequencing. Taylor and Johnson (2001) used the patterns in the predicted ion types, and the accountability of ions greater than the precursor *m/z*, to reduce the number of candidate sequences. Grossmann *et al.* (2005) have used pre-processing to distinguish between b- and y-ions and other peaks.

In the present study, we focus on revealing alternative features useful for spectra quality classification, especially in isotope-coded affinity tag (ICAT) experiments. The LC-MS/MS data was acquired using a quadrupole - time of flight mass spectrometer. We used decision tree supervised classification techniques to predict the quality of a given spectrum before the database search. The predictions were tested against manual validation performed by a human expert in terms of the receiver operating characteristic (ROC) and cross-validation. In addition to ROC-curves, we report also the results from specific (TPR, TNR)-pairs that allow comparison to the results of Bern *et al.* (2004).

2 METHODS

2.1 Test material

The mass spectrometric data used in this study were derived from a LC-MS/MS analysis of the isotope coded affinity tag (ICAT) labelled peptide fractions from previously described ICAT experiments (Moulder *et al*., 2005; Filén *et al.*, 2005). The LC-MS/MS data was produced using a *QSTAR*™ *Pulsar* quadrupole - time of flight mass spectrometer (Applied Biosystems) coupled with a nano-LC instrument (LC-Packings). The data included five repeated analyses from a 9-protein standard mixture, and the analysis of four separate cation-exchange fractions from an applied study of cytokine regulated protein expression in human lymphoblasts (Table 1).

For the purpose of assignment of "good" and "bad" MS/MS spectra, the data was analysed with MS/MS data analysis software ProICAT (version 1.1, Applied Biosystems) and SEQUEST (ThermoFinnigan). For the standard experiment, good spectra were quickly identified by association with the expected components (and contaminants), while performing searches against the full NCBI database with additional searches made against a custom database specifically containing the standard proteins and a reversed database (NCBI human). For the applied study, the data were searched against NCBI human specific database and the "good" spectra were identified by manual validation. The mass tolerances were 0.3 Da for MS spectra and 0.2 Da for MS/MS spectra in ProICAT database searches. For SEQUEST, the tolerances were 0.3 Da for MS spectra and the default value of 0 for MS/MS spectra.

The spectra considered for manual validation were those that were determined to provide peptide identifications with confidence scores greater than 50%, for the ProICAT interpretations, and with Xcorr and deltaCn values greater than or equal to 1.5 and 0.1, respectively, for the SEQUEST results. The rest of the spectra were marked "bad" without manual validation.

2.2 Spectral features

Bern *et al.* (2004) used features that measure the total intensity (B_1) and intensity balance (B_2) of the spectrum, the number of peaks (B_3), the total intensity of peaks with water losses (separated by 18 Da, B_4), the total intensity of peaks with isotopes (separated by 1 Da, B_5), the total intensity of peak pairs with mass gap similar to a known mass gap of an amino acid (B_6), and the total intensity of pairs of peaks which have the sum of masses similar to the parent ion mass (B_7). They also applied rank-based intensity normalization to the spectra to overcome the problem of varying mean intensity of peaks in different spectra, which would otherwise make it more difficult to compare the scores between different spectra.

The 9 additional features used in the present study can be divided into two groups. The first three of these measure the overall, global attributes of the spectrum, such as mean intensity or standard deviation of peak intensities. These features include F_1-F_3. Another type of features was based on finding specific, local attributes of the spectrum. Peaks with certain *m/z*-values should be present in a valid spectrum, even though their absence may not prohibit a successful identification of the spectrum in the database search phase. These features were developed according to the guidelines for manual interpretation (Kinter and Sherman, 2000), and they include F_4-F_9. All the features are computed for each spectrum, constituting a point in $\mathbb{R}^9$. Mass tolerance used for peak position accuracy was 0.2 Da.

F_1: The average intensity of the peaks in the spectrum:

$$F_1 = \frac{\sum_{x \in S} I(x)}{n} ,$$

where n is the number of peaks in the spectrum S and $I(x)$ is the intensity of a peak with *m/z*-value x. Features F3 to F9 were calculated after normalizing the peak intensities of the spectra relative to the highest F1-score of all the spectra in the data set.

F_2: The standard deviation of the peaks in the spectrum.

$$F_2 = \sqrt{\frac{1}{n-1} \sum_{x \in S} \left(I(x) - F_1 \right)^2} ,$$

The rationale behind this feature is that high variability in peak intensities increases the likelihood of peaks with larger intensities, corresponding to ions from amino acids present in the spectrum. A spectrum with little variability in peak intensities is presumed to contain mostly noise peaks.

F_3: The total intensity of exceptionally high peaks in the spectrum. The peak intensities tend to become smaller with increasing masses, and therefore a simple spectrum-wide threshold is not sufficient for deciding whether a peak is exceptionally large. Instead, the highest and lowest *m/z*-values of the spectra from the whole data set are first searched, and the maximal *m/z*-domain is divided into m regions of equal width (m=10 in our study). An upper 90% confidence interval level for peak intensity is calculated for each of these regions. As a result we get 10 (x,y)-pairs of the form (middle *m/z* point of the region, 90% confidence interval value in the region). A k-degree (k=3 in our study) polynomial g is fitted to these points using the least squares method (Cormen *et al.*, 1990); see Fig. 1. For a given spectrum, the total intensity of the peaks ranging above this curve constitutes the feature:

$$F_3 = \sum_{I(x) > g(x)} I(x)$$

F_4: The presence of immonium ions in the spectrum. Most of the immonium ions can be found from the data produced with the MS-instrument used in the production of our test material. The maximum number of immonium ions present in a spectrum depends on the amino acid composition and precursor mass of the peptide. If we denote by $M(x)$ the mass of ion x, then the lower bound of the maximum numbers of these peaks is calculated as:

$$lowerbound = \frac{M(precursor)}{\max(M(x))}$$

where $\max(M(x))$ is the mass of the immonium ion with the largest mass. The upper bound is calculated in a similar way using the amino acid with the lightest immonium ion. The final value of this feature is given by:

$$F_4 = \begin{cases} 1 & , \text{if } lowerbound \le n \le upperbound \\ \dfrac{upperbound}{n} & , \text{if } n > upperbound \\ \dfrac{n}{lowerbound} & , \text{if } n < lowerbound \end{cases} .$$

The feature will give a maximum score of 1, if there is a maximal number of immonium ions in the spectrum.

F_5: The total intensity of peaks resulting from the ICAT-reagent. When fragmenting ICAT-labelled peptides the ICAT-reagent fragments from the peptide producing characteristic ions.

F_6: The total intensity of peptide y_1-ion peak. This peak has either *m/z*-value 147 (lysine-containing peptides) or 175 (arginine-containing peptides), and one of them is usually present in the spectrum if it contains a tryptic peptide sequence. If both are present, the sum of these is used.

F_7: The total intensity of the precursor peak. The hypothesis here is that a spectrum, which contains a peptide sequence also contains a peak corresponding to the precursor ion.

F_8: The total intensity of ions y_{n-2}, b_2 and b_{n-1}. These are ions whose position can be easily computed, and whose presence may suggest, that a peptide sequence is present in the spectrum. The b_2-ion can be found by looking for a (b_2,a_2)-ion pair separated by 28 Da (Kinter and Sherman, 2000). Having found the b_2-ion, it is easy to calculate $M(y_{n-2})$ by deducting $M(b_2)$ from M(precursor). Finally,

$M(b_{n-1})$ is calculated by deducting $M(y_1)$ from M(precursor). If y_1-ion has both 147 and 175 *m/z* peaks, then the more intense of these is used.

F_9: A score based on the mass-ladder, a preliminary peptide sequence, built using rank-normalized peaks of the spectrum as suggested by Bern et al. (2004). Rank normalization of the peaks gives rank 1 to the peak with the largest intensity, rank 2 to the second largest etc. Hence, the number of peaks in the sequence can be controlled by setting the parameters of this normalisation method appropriately (150 largest peaks were used in this study). The starting point of the sequence can be calculated using either the y_{n-2}-ion with mass gap $M(y_{n-2})$-$M(y_1)$ or the b_2-ion with mass gap $M(b_{n-1})$-$M(b_2)$ (Kinter and Sherman, 2000). The sequence can then be constructed by trying to add new amino acids into it, and checking whether the corresponding peaks can be found in the rank-normalised spectrum. The isotopic structure of the peaks in the non-normalized spectrum is taken into account when deciding whether a peak is noise or produced by an amino acid ion. Normally, this kind of naive *de novo*-sequencing is too slow for practical use. However, as we are not trying to produce an exact and complete result, but rather to get a crude estimation whether a spectrum has a clear peptide sequence structure in it, this is sufficient. The feature equals the number of ions that could be added to the sequence plus 0.5 points for each ion in the sequence that could also be found in the immonium ions of the spectrum. If the sequence is completed (the mass of the sequence equals mass gap), then the score is increased by 10 points. It is also possible to use a cut off-value for the sequence length to avoid using too much time in the building of long sequences (cut-off value of 5 was used in this study).

2.3 Classification methods

Supervised classification methods were used to distinguish the two spectra classes "good" or "bad" as assessed by manual validation (see Section 2.1.). We used decision tree techniques, including C4.5 (Quinlan, 1993) and Random Forest (Breiman, 2001). Decision trees operate by selecting splits in the data at some feature values, which are thought to provide the best classification. The resulting groups are then split again recursively to provide smaller groups such that each group contains finally members of one class only. The final tree contains thus nodes which correspond to features and arcs which correspond to values of those features.

The C4.5 algorithm can handle continuous feature values and it can also prune subtrees to avoid over-fitting. A Random Forest is a set of trees, which have been trained with independent randomized subsets of training data set. The class membership of a new sample is determined by all these decision trees and the class obtaining the majority of votes is considered as the final classification. For all the classification and feature selection tasks with these two methods, we used the classification software package WEKA (Witten and Frank, 2000).

2.4 Testing procedure

Classification experiments were performed with our proposed features and with the features B_1 to B_7 of Bern *et al.* (2004), both normalized to [0,1] before classification. Errors were determined with 10-fold cross-validation, where the data set is first divided into 10 parts $D_1,..,D_{10}$, and each part D_i was in turn used as the test set while the other nine parts D_j, $j \neq i$ were used as the training set. Classification error was then estimated according to the 10 test sets.

The results of the classification experiment are displayed as ROC-curves which show the relationship between true positive rate (TPR, the proportion of "good" spectra judged by the algorithm to be good) and false positive rate (FPR, the proportion of "bad" spectra judged by the algorithm to be good), along with the area under curve (AUC), which gives the area under the ROC-curve. In the case of optimal performance TPR=1, FPR=0 and AUC=1.

Another way of reporting the classification results involves selecting the most useful features using Random Forest and RankSearch algorithms in WEKA. RankSearch is a forward attribute selection search method, which starts from an empty set of features and adds new ones if they benefit the classification model in terms of a smaller classification error.

3 RESULTS

3.1 Quality classification

The ROC-curves for the classification are shown in Figure 2 separately for each of the 10 different data sets, when using the Random Forest classifier. The closer the ROC-curve is to the upper-left corner the better the classifier. The corresponding AUC and the FPR at 0.9 TPR-level are shown in Table 2. A FPR of 0.25 means that 75% of the bad spectra are filtered out before identification phase while at the same time losing 10% of the good spectra.

For C4.5 the results were very much similar to those of Random Forest, but they are somewhat worse in all the tests for both the features of the present study and those of Bern *et al.* (2004). The rules used by the C4.5 algorithm in classifying the spectra are reported as supplementary information to provide information on the values of features used in separating the two classes (http://staff.cs.utu.fi/staff/jussi.salmi/C4_5_rules.pdf).

As can be seen from Fig. 2 and Table 2, the proposed features performed better in all the cases except with the SEQUEST-analysed fraction 17, but even there the AUC result was better than with features of Bern *et al.* (2004). As expected, the combined classification with all the features B_1 to B_7 and F_1 to F_9 was the best in most cases. In some cases, the features F_1 to F_9 alone performed marginally better, which is quite normal for decision trees, because the extra features can complicate the tree branching and provide contradictory information.

There was no clear difference between the results for the standard protein mixture and the lymphoblast study. From the analysed fractions, 19 was classified best by our system. Fraction 16 contained much less valid spectra than the other fractions. Differences between the SEQUEST and ProICAT -analysed test materials were only marginal, but consistently ProICAT-analysed results gave slightly better correspondence with our features, see Table 2.

3.2 Feature selection

In order to test the importance of the considered features, feature selection was performed for B_1 to B_7 and F_1 to F_9. The results are shown in Table 3, which contains the features used

by the Random Forest decision tree each time during the 10-fold cross-validation of the attribute selection.

Attribute F_3 was the most frequently used attribute; it was used every time the decision tree was constructed for each data set. Feature F_6 was among the best features with 7 times, B_7 5 times and F_1 and F_7 4 times. Feature F_9 is similar to feature B_7, as they both try to confirm the presence of a valid peptide mass ladder in the spectrum. However, B_7 was used more frequently, so it probably provides a better estimation.

All the data represent ICAT-labelled samples, and therefore feature F_5, which uses the known peaks resulting from the ICAT-reagent is feasible here. The compared method of Bern *et al.* (2004) did not presume the sample to be ICAT-labelled. However, the feature F_5 was only among the best features with just two data sets (1 and 10) and therefore it is not likely to be very important for the classification.

4 DISCUSSION

In the present study, we have considered the problem of filtering out poor quality MS/MS spectra to reduce the time needed for identifying the proteins from a complex biological sample. The spectral features and the classification methods were tested on real-life data and compared to the results of Bern *et al.*(2004). The results showed that the pre-filter can benefit the data analysis pipeline by reducing the number of spectra, which should otherwise be manually validated after the identification phase. The amount of time spent in database searches is reduced, since it is linearly proportional to the number of spectra. Also de novo-sequencing algorithms should perform faster and more reliably after the removal of bad spectra from the data.

The total intensity of high peaks in a spectrum (feature F_3) proved to be a useful measure of spectral quality, as well as the presence of a peptide sequence terminus peak (F_6), the precursor peak (F_7), and a pair of peaks together forming the mass of the precursor peak (B_7) as also suggested by Bern *et al.* (2004). Nevertheless, we recommend using all the features F_1 to F_9 when filtering spectra, because each of them were among the best features for some test material, see Table 3. Combining also the features of Bern et al. provided more reliable results in general.

There are some possible explanations for the observed discrepancy between our classification results and those of Bern *et al.* (2004). Firstly, differences between the types of mass spectrometers used may have contributed to differences in the MS/MS data. The data used by Bern *et al.* (2004) was acquired using an ion trap instrument, whereas our data was generated using a quadrupole time-of-flight instrument. Secondly, the methods for the data classification and validation are different. Bern *et al.* (2004) used the SEQUEST results directly as the basis for data classification, whereas we combined manual validation to the interpretation of our ProICAT and SEQUEST results.

Figure 3 shows SEQUEST Xcorr-score distributions for the manually validated "good" and "bad", as well as for the filtered and unfiltered spectra from material 5. It can be seen, that manual validation agrees well with the Xcorr-score, and that filtering keeps the spectra with the highest Xcorr-scores, although it also keeps many spectra with low Xcorr-scores. Similar distributions of the scores of "good" and "bad" spectra would be expected if the data analyses has been made with Mascot; Mascot analysis was made for this same test material, and the distribution of ion scores is shown in the supplementary figure.

In the work of Purvine *et al.* (2004), some excellent results were produced with ion trap data using filtering features based on charge state assignments and intensities. Whilst, ambiguity in charge assignment is a less common problem with data from quadrupole time-of-flight instrument instruments, inclusion of such a feature would be valuable addition for a generic data processing tool.

At a more general level, the scripts mzStar (Pedrioli *et al.*, 2004) and wiff2dta (Boehm *et al.*, 2004), which are used to convert QStar *.wiff data files to the generic mzXML format, and convert QStar data to a SEQUEST searchable format, respectively, include features that permit data refining by centroiding and the removal of peaks with intensity less than a chosen threshold. At the time of this study we found that these features did not provide satisfactory results with mzStar, and we did not pursue the use of wiff2dta due to compatibility problems with our data analysis pipeline. We have since learned that the recently released version (December 2005) of the latter script should solve these compatibility problems.

Although in recent years a number of algorithms have been developed for database searching and post-processing of the search results, only a limited effort has been devoted to the preprocessing of MS/MS spectra prior to searching. The presence of un-interpretable spectra and those not derived from peptides does, however, increase the burden of computer time and post processing validation. Accordingly, the understanding of the sources of noise and the development of filtering methods are useful in the streamlining of data analysis. The integration of the different, pre- and post- analysis phases of proteomic data analysis into a complete analysis framework will assist efficient large-scale proteomic studies.

5 ACKNOWLEDGEMENTS

The work was supported by the graduate school in Computational Biology, Bioinformatics and Biometry (ComBi), the national graduate school in Informational and Structural Biology (ISB), the National Technology Agency and the Academy of Finland (grants 203632, 53377 and 203654). We thank Petri Kouvonen from the Proteomics and Mass Spectrometry Unit of the Centre for Biotechnology, and Raija Andersen, Outi Melin and Marjo Linja for excellent technical assistance.

REFERENCES

Anderson, D. C., Li, W. Q., Payan, D. G. and Noble, W. S. (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.*, **2**, 137-146.

Bern, M., Goldberg, D., McDonald, W. H. and Yates, J. R., III (2004) Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics*, 20, i49-i54.

Boehm, A. M., Galvin, R. P. and Sickmann, A. (2004) Extractor for ESI quadrupole TOF tandem MS data enabled for high throughput batch processing. *BMC Bioinformatics,* **5,** 162.

Breiman, L. (2001) Random Forests. *J. Machine Learning*, **45**, 5-32.

Cargile, B. J., Bundy, J. L., Stephenson, J. L. Jr. (2004) Potential for false positive identifications from large databases through tandem mass spectrometry. *J. Proteome Res.*, **3**, 1082-5.

Chamrad, D. C., Körting, G., Stühler, K., Meyer, H. E., Klose, J., Blüggel, M. (2004) Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*, **4**, 619-28.

Cormen, T. H., Leiserson, C. E. and Rivest, R. L. (1990) *Introduction to Algorithms.* MIT Press, New York.

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**, 976-89.

Filén, J.-J., Nyman, T. A., Korhonen, J., Goodlett, D. R., Lahesmaa, R. (2005)Characterization of microsomal fraction proteome in human lymphoblasts reveals the down-regulation of galectin-1 by interleukin-12. Proteomics, Oct 24; [Epub ahead of print].

Grossmann, J., Roos, F. F., Cieliebak, M., Liptak, Z., Mathis, L. K., Muller, M., Gruissem, W., and Baginsky, S. (2005) AUDENS: A tool for automated peptide de novo sequencing. *J. Proteome Res.* **4**, 1768-1774

Fenyö, D., Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.*, **75**, 768-74.

Keller, A., Nesvizhskii, A., Kolker, E. and Aebersold, R. (2002a) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, **74**, 5383-92.

Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S., Goodlett, D. R., Kolker, E. (2002b) Experimental protein mixture for validating tandem mass spectral analysis. *OMICS*, **6**, 207-12.

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

MacCoss, M. J., Wu, C. C., Yates J. R., III (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.*, **74**, 5593-9.

Moulder, R., Filén, J.-J., Salmi, J., Katajamaa, M., Nevalainen, O. S., Oresic, M., Aittokallio, T., Lahesmaa, R. and Nyman, T. A. (2005) A comparative evaluation of software for the analysis of liquid chromatography-tandem mass spectrometry data from isotope-coded affinity tag experiments. Proteomics, **11**, 2748-60.

Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22**, 1459-1466

Perkins, D. N., Pappin, D. J. C., Creasy, D. M. And Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551-67.

Purvine, S., Kolker, N., Kolker E. (2004) Spectral quality assessment for high-throughput tandem mass spectrometry proteomics. *OMICS*, **8**, 255-65.

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, USA.

Razumovskaya, J., Olman, V., Xu, D., Uberbacher, E. C., VerBerkmoes, N. C., Hettich, R. L. and Xu, Y. (2004) A computational method for assessing peptide-identification reliability in tandem mass spectrometry analysis with SEQUEST. Proteomics, **4**, 961-969.

Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D., Jonscher, K. R., Pierce, K. G., Old, W. M., Cheung, H. T., Russell, S., Wattawa, J. L., Goehle, G. R., Knight, R. D., Ahn, N. G. (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.,* **76**, 3556-68.

Savitski, M. M., Nielsen, M. L. and Zubarev, R. A. (2005) New Database-independent, Sequence-tag-based Scoring of peptide MS/MS data validates mowse scores, recovers below-threshold data, singles out modified peptides and assesses the quality of MS/MS techniques. *Mol. Cell Proteomics*, **4**, 1180-8.

Sun, W., Li, F., Wang, J., Zheng, D. and Gao, Y. (2004) AMASS: software for automatically validating the quality of MS/MS spectrum from SEQUEST results. *Proteomics*, **3**, 1194-9.

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

Witten, I. H. and Frank, E. (2000) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Fransisco, USA

Fig. 1. A spectrum classified as good and manually validated. X-axis gives the *m/z*-values and y-axis the intensity. The features F_1 (solid horizontal line close to *m/z*-value of 0) and F_3 (solid downward-sloping curve) are shown. Feature values F_1 to F_9 are (before normalization to [0,1]): 4.6, 25.4, 188, 0.75, 175, 11, 18, 336, 0, respectively.

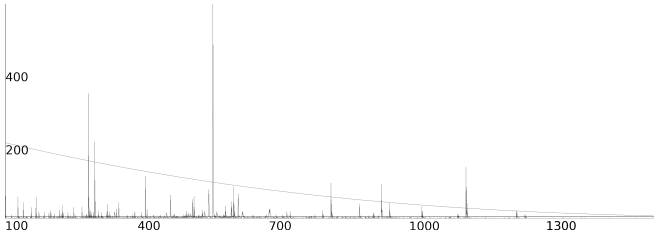

Table 1. The test material and the number of spectra in different data sets.

Material nr.	Data source	Analysis software	Good	Bad	Total
1	Applied fraction 16	SEQUEST	25	832	857
2	Applied fraction 16	Pro ICAT	45	812	857
3	Applied fraction 17	SEQUEST	89	374	463
4	Applied fraction 17	Pro ICAT	109	354	463
5	Applied fraction 18	SEQUEST	120	459	579
6	Applied fraction 18	Pro ICAT	156	423	579
7	Applied fraction 19	SEQUEST	67	592	659
8	Applied fraction 19	Pro ICAT	100	559	659
9	Standard Mixture	SEQUEST	336	1821	2157
10	Standard Mixture	Pro ICAT	397	1760	2157

Fig. 2. The ROC-curves for the Random Forest classifiers. Each of the panels shows the ROC-curve for a specific data set with both the proposed features (solid line) and the features of Bern *et al.* (dotted line). Also shown are the 90% true positive rate (TPR, horizontal straight line) and the corresponding false positive rate (FPR, vertical straight line). Test material used in each panel of the figure is marked left of the panel (see Table 1).

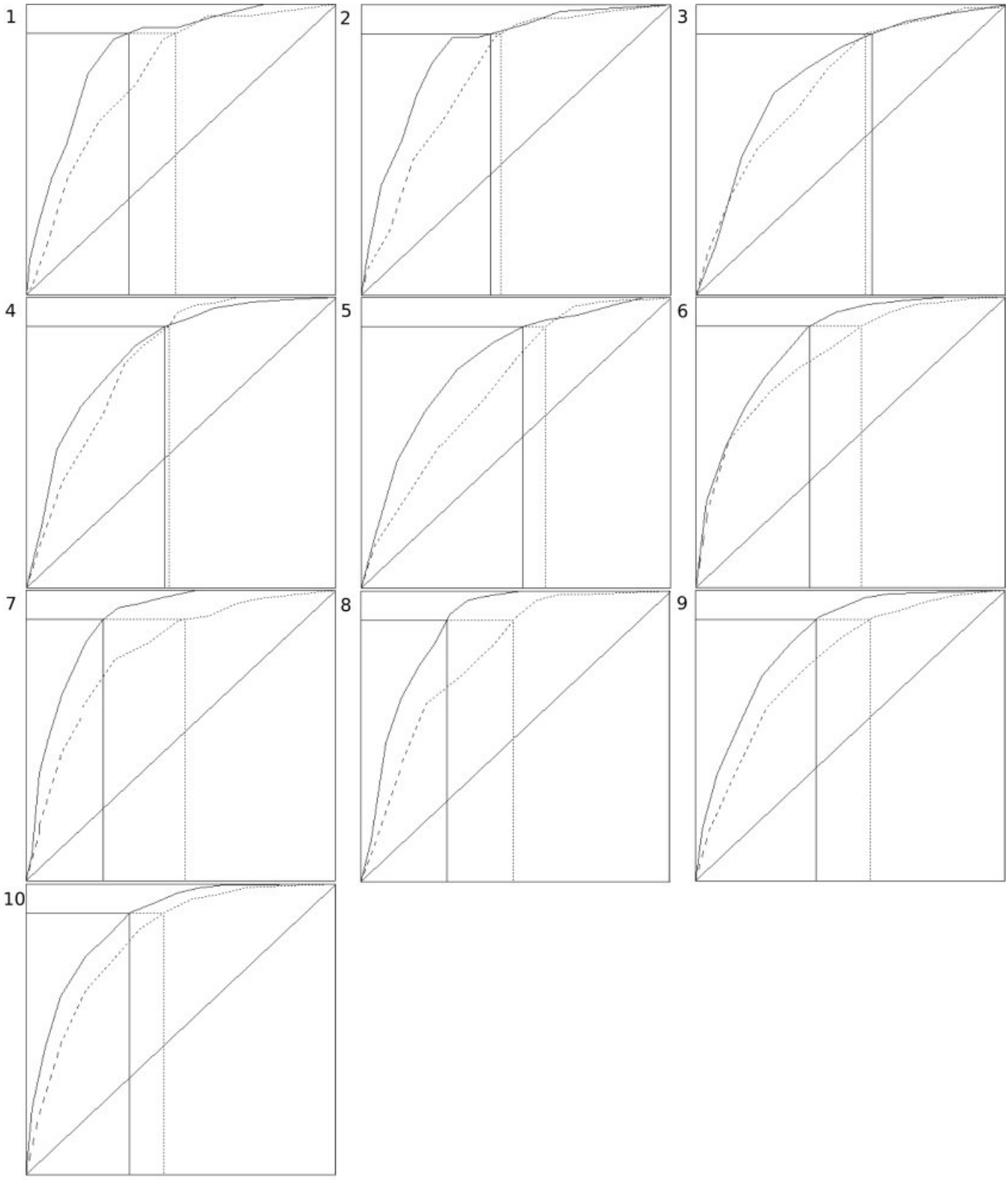
1
2
3
4
5
6
7
8
9
10

Table 2. A summary of test results with Random Forest classifier. "Our" column shows the results with the proposed features F_1 to F_9, "Bern" column with the features B_1 to B_7 by Bern *et al.* (2004) and the "Combined" column shows the results with the features combined. The FPR and AUC-values correspond to the panels 1-10 in Fig. 2. To provide easier comparison with the previous results of Bern *et al.*, we used the same level of good spectra by fixing TPR to 90%.

a) SEQUEST analysed test material

Test material	FPR (TPR≈0.90)			AUC		
	Our	Bern	Combined	Our	Bern	Combined
1	0.33	0.41	0.28	0.84	0.76	0.87
3	0.57	0.55	0.54	0.76	0.74	0.77
5	0.52	0.60	0.52	0.78	0.70	0.76
7	0.25	0.52	0.24	0.89	0.79	0.88
9	0.40	0.58	0.42	0.83	0.75	0.84

b) Pro ICAT analysed test material

Test material	FPR (TPR≈0.90)			AUC		
	Our	Bern	Combined	Our	Bern	Combined
2	0.30	0.43	0.32	0.84	0.76	0.85
4	0.45	0.46	0.44	0.81	0.79	0.82
6	0.42	0.53	0.39	0.84	0.80	0.84
8	0.29	0.51	0.25	0.87	0.77	0.86
10	0.34	0.45	0.36	0.87	0.81	0.87

Table 3. The features selected by the Random Forest decision tree in each 10-fold cross-validation step.

Test material	Selected features
1	F_2, F_3, F_5, F_6, F_7, B_7
2	F_3
3	F_1, F_3
4	F_3, B_3, B_7
5	F_1, F_3, F_7, F_6, B_7
6	F_1, F_3, F_6, F_9, B_1, B_3, B_7
7	F_3, F_6, F_7, B_2, B_4, B_7
8	F_3, F_6, F_7, B_5
9	F_1, F_3, F_6, F_8, B_1, B_2
10	F_3, F_4, F_5, F_6, F_8, B_1, B_2, B_3

Fig. 3. (a) SEQUEST Xcorr-score distribution in the test material 5 for the manually validated "bad" (dotted line) and "good" (continuous line) spectra at FPR=90% and (b) for the whole (dotted line) and filtered (continuous line) set of spectra.