Robust Prediction of the MASCOT Score for an Improved Quality Assessment in Mass Spectrometric Proteomics

Thomas Koenig[1,2], Bjoern H. Menze[1], Marc Kirchner[1,2,†], Flavio Monigatti[2,3,†], Kenneth C. Parker[4,†], Thomas Patterson[2], Judith Jebanathirajah Steen[5], Fred A. Hamprecht[1], Hanno Steen[2,3,*]

[1] Interdisciplinary Center for Scientific Computing, University of Heidelberg, Heidelberg, Germany

[2] Department of Pathology, Children's Hospital Boston, Boston, MA, USA

[3] Department of Pathology, Harvard Medical School, Boston, MA, USA

[4] Partners Healthcare Center for Genomics and Genetics, Harvard Medical School, Cambridge, MA, USA

[5] Department of Neurobiology, Harvard Medical School and Children's Hospital Boston, Boston, MA, USA

[*] Address correspondence to:
Hanno Steen, Ph.D.
Children's Hospital Boston
Department of Pathology
Enders 1130
320 Longwood Avenue
Boston, MA 02115
phone +1-617-919-2629
fax +1-617-730-0168
hanno.steen@childrens.harvard.edu

[†] Authors contributed equally to this work.

Keywords. classification, supervised learning, regression, random forest, peptide identification

Abstract

Protein identification by tandem mass spectrometry is based on the reliable processing of the acquired data. Unfortunately, the generation of a large number of poor quality spectra is commonly observed in LC-MS/MS, and the processing of these mostly non-informative spectra with its associated costs should be avoided. We present a continuous quality score that can be computed very quickly and that can be considered an approximation of the MASCOT score in case of a correct identification. This score can be used to reject low quality spectra prior to database identification, or to draw attention to those spectra that exhibit a (supposedly) high information content, but could not be identified. The proposed quality score can be calibrated automatically on site without the need for a manually generated training set.

Turning this score into a classifier and by using features independent of the instrument, the proposed approach performs equal to previously published classifiers and feature sets and also gives insights into the behavior of the MASCOT score.

Introduction

One of the hallmarks of proteomics is mass spectrometry-based protein identification which uses peptide fragment information and protein sequence databases[1,2]. To identify any substance with confidence, informative tandem mass spectra must be matched to peptides in protein sequence databases of interest. Current automated procedures (e.g. MASCOT[3], SEQUEST[4] or ProteinPilot[19]) assign peptide sequences contained in a database to mass spectra regardless of the actual quality of the spectrum under consideration. With "quality" we refer to the amount of useful information that can be extracted from a product ion spectrum. Unfortunately, the generation of a large number of bad quality spectra, which contain too little, irrelevant or ambiguous information, is commonly observed in LC-MS/MS, and is especially prevalent in cases of samples with limited amount of material (Fig. 1). Consequently, these spectra only slow down the identification process.

The majority of poor peptide hits results from the fragmentation of chemical background, or to a smaller extent from a low intensity of the signal when the selected precursor ion is low in abundance or does not fragment efficiently in the mass spectrometer[5]. As a database search is time-consuming, it is desirable to identify and filter out such spectra prior to the search process.

A second, more interesting class of unmatched spectra consists of high quality spectra that do not match to any sequence in the database according to the search parameters used. In most cases, these spectra relate to peptides with unexpected enzymatic cleavage sites or with unexpected chemical modifications that are not being considered during the search. These spectra might reveal important biological information not annotated in the database such as truncations, modifications, or splice variations[1,2,6]. In most cases, a manual inspection of the spectra would allow to identify that kind of data. However, as

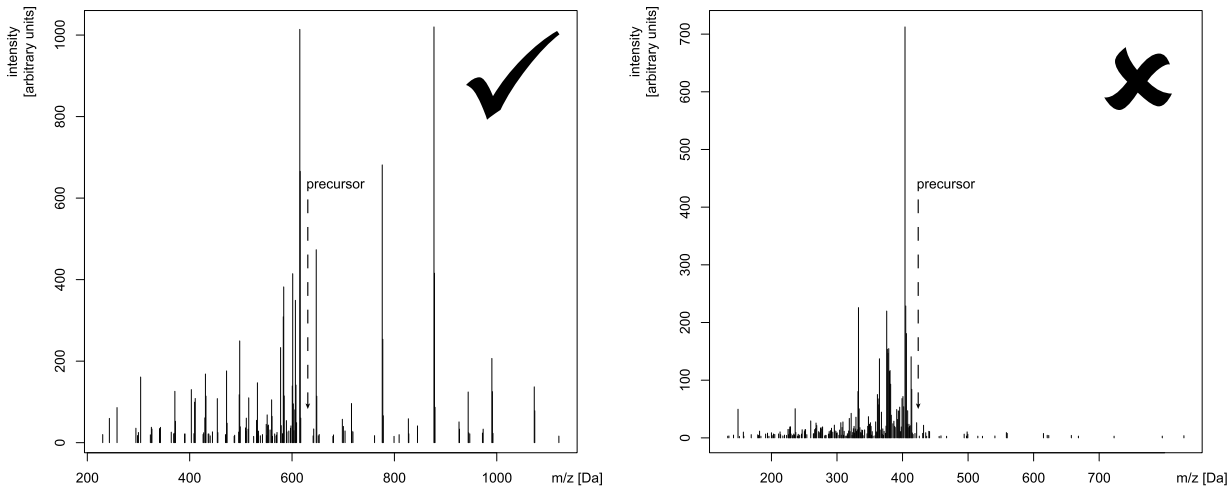

Figure 1: Examples for multiply charged precursors resulting in good (left) and bad (right) spectra. In general, the two groups comprise a variety of different types of spectra and thus cannot be distinguished only by the characteristics exhibited by the two examples shown here. This is why a machine learning approach is employed.

today's mass spectrometers acquire thousands of product ion spectra in a single LC-MS/MS experiment, manual inspection is not a viable option. In this paper, we describe a method to separate high quality peptide spectra from the initial set. This selection can be carried out

- prior to database searching to speed up the identification process, or
- on spectra which could not be identified with confidence during a database search, in order to find those spectra exhibiting a (supposedly) high information content, so that additional more time-consuming searches using modified search parameters can be performed.

Technically, the idea is to employ a machine learning approach that relates a class label to the quality of a product spectrum, which is described by a number of features tailored to this specific task[7−15]. In contrast to prior work in this field, we use a regression to predict the MASCOT score as a measure for spectral quality rather than applying a binary classifier which only decides whether a spectrum is of good quality or not. Since

this approach can rely on MASCOT scores in the training phase, it does not require the tedious manual generation of training data sets. This property of automatic *on-site calibration* is central for the robustness of the chosen strategy and is paired with a set of features independent of total ion currents, i.e. of instrument settings such as trapping times, automatic gain control target values and detector efficiencies.

Previous Work. Present approaches mainly focus on training binary classifiers distinguishing good and bad spectra. Training and test sets are usually compiled using the success of identifications on a mixture of known proteins, or by manually assessing spectral quality. The different strategies can be coarsely divided into two groups:

- *A posteriori* approaches, when the classification is performed after identification. In this case, the classifier's decision is based on the output of the individual search results[7−10]. By definition, they cannot serve as quality filters, but only give an estimate for the reliability of a specific protein identification instead.

- *A priori* approaches that aim at classifying spectra prior to applying an identification technique. Consequently, these procedures only use information directly obtained from the spectrum of interest and can therefore be used both as quality filters as well as a means for detecting supposedly false identifications.[11−15] The method presented in this paper belongs to the latter class.

To our knowledge, Moore *et al.*[11] were the first who tried to tackle the classification problem. Their approach was based on the creation of a binary dataset using thresholds on cross-correlation scores from SEQUEST searches. They then employed an evolutionary strategy to train the feature weights of a linear function. Bern *et al.*[12] classified spectra using a set of features human experts would consult when manually assessing quality, and trained a quadratic discriminant analysis (QDA) classifier on the outcome of those features. Additionally, they used a rank-normalized histogram of mass differences as the

input to a Support Vector Machine (SVM), which they found to perform better than their so-called handcrafted features. The integration of rank-normalized intensities makes their approach independent from absolute ion intensities, an important advantage in standard proteomics experiments. Bern *et al.*[12] also proposed a quality score based on a linear regression which was trained on the number of b- and y-ions matched during a database search. However, score and binary classifier were completely independent of each other, which amounted to two separate methods for rating spectral quality. Furthermore, the above number of correct matches is difficult to obtain.

Xu *et al.*[13] assembled a set of manually assessed spectra, trained a QDA classifier on general features associated with those spectra and then used the classifier's output as a score that, instead of monotonically increasing with spectral quality, exhibited an unintuitive behavior that assigned low and high scores to poor spectra, while tagging good ones with intermediate scores. They applied the classifier to a large number of yeast lysate spectra, but did not carry out a validation of its performance on ground truth.

Salmi *et al.*[14] used decision trees for rating the quality of ICAT-labeled data.

A recent study by Flikka *et al.*[15] deals with data acquired on Q-TOF, TOF-TOF and quadrupole iontrap instruments. It introduces a clustering technique prior to classifier training in order to avoid a selection bias when certain classes of spectra are much more abundant in the training set than others.

MASCOT Score Prediction as Quality Measure. All approaches mentioned above have in common that they make use of a binary classifier; as a consequence, their users cannot choose what they consider good and bad spectra without retraining. Once it has been trained, a classifier will be adapted to a certain problem that defines the characteristics of good and bad spectra. In order to obtain optimal results, retraining will be necessary if the requirements regarding its desired behavior change. Of course, common classifier implementations provide an implicit measure for the confidence of an

assignment, usually in terms of a class probability. Preferable, however, would be a continuous score that explicitly measures quality. Such a quality score would also allow for a comparison with results from database searches, such as MASCOT scores, and to find those spectra exhibiting a large deviation in the statements made by the two scores. It would also allow for filtering out spectra that exhibit scores below a *user*-definable threshold to both save processing time with regard to database searches, and to assist human experts in giving them a selection of the high quality spectra with low MASCOT scores, i.e.unassigned, but potentially interesting spectra.

In the following, we will describe a continuous *a priori* quality score which is based on regression rather than binary classification. We have chosen the MASCOT score as training label since it is considered a proxy for spectral quality by most human experts. The labels that are generated in this manner are thus very cheap, involving no time-consuming manual assignment of classes to a large number of spectra, making it more efficient to set up such a score. Based on that, we trained a Random Forest[17] (RF) to predict the MASCOT score in an *a priori* fashion. We hypothesized that this regression would generalize sufficiently well to assign high scores even to those spectra which are erroneously assigned a low MASCOT score (candidates for further validation by human experts).

Testing our hypothesis comprised the following steps:

- Spectra were tagged with a continuous, computer-generated label, the MASCOT score, which represented the ground truth for training.
- Starting from a large number of features, we trained the regression and successively reduced the initial feature set to optimize regression performance.

The following three steps involve different ground truths and are only necessary for validating our method. They are not required for establishing the regression.

- The *regression performance* of the resulting feature set is validated against a manually labeled, four-class ground truth.
- This feature set, along with five other sets (four of them published before), is then also compared in terms of its *classification performance* with a two-class gold standard created by using ProteinPilot as independent protein identification tool. Based on this, a feature set is assembled which ensures independence from absolute ion intensities and degrading detector efficiencies, making it well suited for the requirements arising in common mass spectrometric experiments.
- The practical usability is demonstrated for a dataset that was searched with an incorrect MASCOT setting for the static cysteine modification employed.

Materials & Methods

Automatic On-Site Calibration. The data used for this study stem from various tryptic in gel digests, for which we used generic protocols as previously described by Shevchenko *et al.*[18] In short, proteins were separated by SDS-PAGE prior to in-gel reduction and alkylation. Subsequently, the samples were digested overnight with sequencing grade trypsin (Promega, Madison, WI, USA). After digestion, the gel plugs were repeatedly extracted prior to vacuum centrifugation. The dried digests were reconstituted in loading buffer prior to LC-MS/MS analysis using a linear quadrupole ion trap. A LTQ (Thermo Scientific, San Jose, CA, USA) equipped with a microautosampler and a Surveyor HPLC pump (both: Thermo Scientific) was used for the analyses. Gradients from 5% to 40% buffer B (0.2% formic acid in acetonitrile) were used; the length of the gradient was in the range of 10 to 60 minutes depending on the expected sample complexity. An in-house written program was used to extract fragment ion information from the raw data and convert them into mgf-files for subsequent protein identifications using MASCOT. Thus, the foundation of our analysis was made up by a large database comprising all resulting spectra. We then focused on multiply charged precursor ions only. The reason for this was that we intended to use our method on spectra acquired on FT and Orbitrap instruments, which can reliably detect the charge states allowing the exclusion of singly charged ions from further analysis as they often correspond to irrelevant chemical background.

The database of spectra was characterized by a large preponderance of data sets with a low MASCOT score; in order to reach a more balanced representation in the training set, a stratified sampling strategy was applied, selecting equal numbers of spectra out of MASCOT score intervals of size one. Good coverage of the feature space is required in order to avoid the selection bias mentioned in the introduction, so only unique peptide

sequences were allowed in the chosen score intervals.

The group of spectra featuring very low MASCOT scores comprises heterogeneous product ion spectra of bad and good quality, the latter because they had erroneously been assigned a low score (false negatives) for various reasons discussed above. To avoid deterioration of the model by inclusion of those false negatives, only data with a MASCOT score of at least 10 was incorporated in the regression training and test sets. In contrast, spectra being awarded a very high score usually no longer exhibit differences in the quality as visually assessed by human experts. Instead, differences in their scores are mainly caused by the presence of additional single ions that could be matched by MASCOT. Given that behavior, the spectra with a MASCOT score higher than 80 were excluded from training and testing, i.e. regression output will be limited to the interval [10, 80]. We discuss and justify this procedure in the supplementary materials.

Taking into account all of the above considerations resulted in a total number of 70,000 spectra, grouped into score intervals containing 1,000 each. All spectra were randomly assigned to either a training set of size 14,000 or to a test set of size 56,000, respectively. Incidentally, to make the chosen approach as flexible as possible, no constraints on the MASCOT search parameters (e.g. mass tolerances, modifications etc.) were imposed when sampling the spectra from the database.

***A Priori* Score Prediction.** We used R[20] to train RFs as well as SVMs with linear and radial basis function kernels on a number of feature sets (see below). As it turned out, evaluation was considerably slower in case of the SVMs. In contrast, the simple tree structures of a RF can be evaluated very quickly, which makes it optimally suited for fast preprocessing. A RF is also easy to train (near optimal results with default parameters, no overfitting when the forest size is increased) and provides an intrinsic way of assessing feature importance. Additionally, it performed well in another similar study by our group in which the data quality in magnetic resonance spectroscopic images[21] was assessed. For

all these reasons, we chose RF for further regression and classification in all subsequent analyses. Its output will be referred to as the "quality score". The training procedure was invoked with standard parameters[22], i.e. 500 trees (*ntree*) and $p/3$ randomly chosen variables at each split (*mtry*), where p denotes the number of features.

As the regression input, a number of features comprising those proposed by Bern *et al.*[12], Moore *et al.*[11] and Xu *et al.*[13] were assembled (cf. table 1), and the obtained set was expanded by average intensity, the percentage of total ion current (TIC) above the precursor mass, the number of peaks greater than 1% TIC and overall peak density (average number of peaks per Da). Then a backward selection was applied based on RF's importance measure[17]: the features with lowest importance were successively removed from the set if their removal did not provoke a significant increase in out-of-bag mean squared error (MSE; i.e. an increase greater than 4). In case two features had a very similar importance, the one which was computationally more expensive was excluded from the set. The result is referred to as the "fast subset". It was combined with a histogram of mass differences of up to 186.5 Da with a binning of 1 Da as proposed by Bern *et al.*[12], i.e. using rank-normalization; this feature set is termed the "composite set". Rank-normalization was chosen in order to make the histogram independent from absolute ion intensities while maintaing interrelationships between peak heights, a behavior which simple presence/absence features would not have accounted for.

Performance Evaluation. In the regression case, 400 spectra were manually assessed, sampling from four characteristic regions of MASCOT scores vs. predicted quality scores (see below). The spectra were taken out of the smallest possible square necessary to obtain 100 spectra per location. They were independently rated by three experts according to the classes *excellent*, *good*, *intermediate* and *poor* without prior knowledge of the MASCOT or predicted quality score. The final label of a single spectrum was set to the median of the expert votes. The resulting labels were then used to assess the

compatibility of the regression results with what experts consider spectral quality.
In order to test classification behavior on a ground truth close to lab reality, an additional biological sample was analyzed. We considered this step necessary to ensure that our approach would be capable of handling common classification problems arising in standard proteomics experiments. For this purpose, the protein content of an aliquot of a human urine specimen was precipitated using a final concentration of 10% trichloroacetic acid. The precipitate was dissolved in SDS loading buffer and fractionated by SDS-PAGE. After Coomassie staining, the band corresponding to human serum albumin was excised and analyzed as detailed above. The data were searched using ProteinPilot 2.0. We decided in favor of ProteinPilot in order to obtain a ground truth independent from MASCOT searches, since we already trained the regression on them. This strategy thus avoids overly optimistic results and gives further confidence in the classification results. Finally, all spectra were assigned to the two classes "good" or "bad" according to their identification outcome: proteins were considered as being identified when two or more peptides could be assigned with a confidence of $> 99\%$. Consequently, the corresponding peptide spectra were considered "good" in this case, and bad otherwise. We opted for a binary classification task to represent the desired behavior of a quick prefilter prior to database searching.
Using these binary labels, the performances of (i) the composite set, (ii) the fast subset, (iii) the mass difference histogram and the features proposed by (iv) Bern *et al.*[12], (v) Moore *et al.*[11], and (vi) Xu *et al.*[13] were compared. In order to assess the quality of the feature sets, Receiver Operator Characteristics (ROCs) of the six feature sets were computed in two different ways:

- The regression previously trained on the database spectra was applied to the ones described above. Then the corresponding ROC was created by varying a threshold on the quality score for separating bad from good spectra.

- A binary RF classifier was trained directly on the binary labels created with ProteinPilot in order to investigate whether the proposed quality score can keep up with conventional techniques. The corresponding ROC was then generated by varying the probability threshold that had to be exceeded by an observation in order to be considered a good spectrum. This was done using five-fold cross-validation with ten iterations, and the median of the AUC was used to plot the corresponding ROC.

After rating using the ProteinPilot identification outcomes, the (randomly balanced) dataset comprised 1746 spectra, each of the two classes consisting of 873 spectra.

Ability to Uncover Good, but Unidentified Spectra. In many proteomics applications the run time of the database search is less important than the identification of as many spectra as possible. It can even be crucial to reveal spectra that are not correctly matched by standard search parameters. In order to simulate this case, we ran a MASCOT search on the sample described above with the static cysteine modification set to "acrylamide", and another search using the correct setting "iodoacetamide". The search parameters included deamidation, methionine oxidation, and pyroglutamic acid formation as dynamic modifications. As a consequence, MASCOT was unable to find the correct cysteine containg spectra1 during the first search. We then compared the results of both searches to pick out these spectra, and applied our regression to it. Thus, each of the spectra was assigned a quality score. We could then determine a sensitivity for this problem by choosing a threshold for the quality score, defining all the spectra with scores below it as bad and the other ones as good. By varying this threshold, we obtained the false negative rate as a function of it.

Implementation Issues. The methods presented above made use of the *R* packages *randomForest* 4.5-18, *e1071* 1.5-16 for training the SVMs and *ROCR* 1.0-2 to plot ROCs and calculate AUCs. We also integrated the regression into our in-house data conversion

software. A chart depicting the general workflow is shown in Fig. 2. It incorporates the call of an external R process, which takes over the score prediction. The interface between our preprocessing software and R is implemented by an exchange of binary files. The quality score is finally written into the mgf-header of the corresponding MASCOT search files.

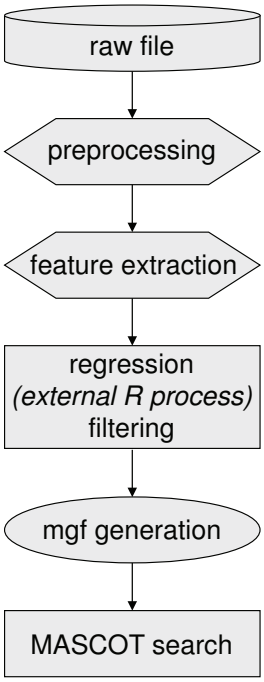

Figure 2: Preprocessing scheme

Results & Discussion

Selection of Features and their Relevance for Score Prediction. The feature selection strategy initially resulted in the following subset of seven features: TIC (1), the number of peaks (2), the percentage of TIC above the precursor mass (3), the average peak intensity (4), the standard deviation of the peak intensity (5), the number of peaks greater than 1% TIC (6) and overall peak density (7). Therefore, this set comprises information about the ion intensity present in a spectrum (1, 2 & 4), noise (5, 6 & 7) as well as the probability for a spectrum to originate from a multiply charged precursor ion (3). Combined with the mass difference histogram, this amounted to a total number of 193 features for the composite set, as shown in table 1.

The variable importance[17] of this composite feature set is depicted in Fig. 3. First of all,

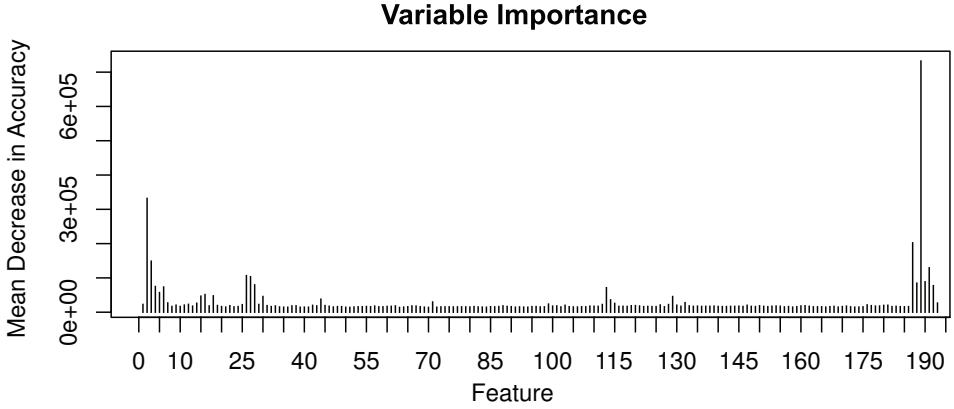

Figure 3: Variable importance[17] for the composite feature set. The first 186 channels describe the influence of the mass differences present in a spectrum up to 186.5 Da, whereas the last seven features represent general features described in the text.

it suggests that the features human experts take into account (187-193, order as described above) are more relevant than mass differences. However, it also provides evidence that the presence of isotope peaks (Δ of 2 and 3) is a good measure for spectral quality. An explanation for a Δ of 4 to 6 awaits further experiments.

To a lesser extent and in accordance with our expectations, relevance is also given to

certain amino acid residue masses (alanine at 71, valine at 99, (iso-)leucine at 113, asparagine at 114, aspartic acid at 115 and glutamic acid at 129). Interestingly, the presence of other amino acids than those stated apparently does not separate good from bad spectra using our strategy.

The small peak at a mass difference of 44 corresponds to the polymer polyethylene glycol, a frequently observed contaminant. It should be noted that a high importance does not imply a high quality score. In the above example, the regression obviously inferred that high intensities at this polymer-specific mass difference correlate with low MASCOT scores and thus represent spectra that are not desirable. The Δ of 28 Da can be explained by the mass difference between a- and b-ions, whereas a Δ of 18 Da most likely corresponds to water loss. The mass differences of 15, 16, 26, 27 and 30 Da that achieved a high importance are currently under investigation.

However, by considering the cases where a clear assignment was possible, it becomes evident that the importance measure was able to find biologically relevant information in the mass spectra the RF regression was trained on.

Results of the Score Prediction. Fig. 4 shows a comparison of the two scores for training and test set, calculated using the composite feature set. Interesting is a saddle point between the two density maxima at the lower left and the upper right corner, respectively, visible in both the training (Fig. 4a) and test set (Fig. 4b). For the test set, the marginal distribution of the two scores is depicted in Fig. 5. It shows that the regression has a slope which is too flat, i.e. it makes predictions that are biased towards the average score for most of the spectra located at the extremes of the chosen MASCOT score interval. Also striking is a non-constant bias with respect to the bisecting line (Fig. 4b). The results of the two Support Vector Machines also showed very similar behavior, and the same was also true for all feature sets examined (results not shown). This anomaly is investigated next and is shown to be a *fundamental relationship* between the

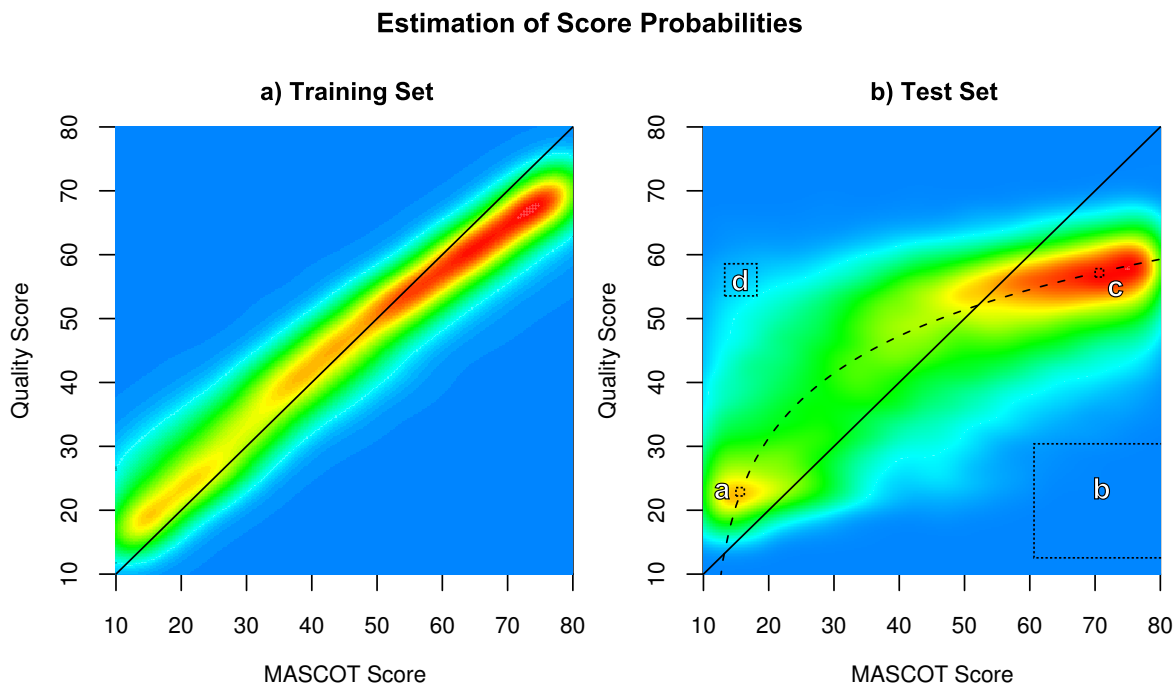

Figure 4: Two-dimensional kernel density estimation of predicted quality scores vs. MASCOT scores: a) training set (14,000 spectra), b) test set (56,000 spectra). Letters specify the positions at which manual assessment took place. The dashed line shows a logarithmic fit of quality as a function of the MASCOT score.

MASCOT score and the quality of a spectrum.

Validation of the Score Prediction. To further validate the test set results, the four regions marked with boxes in Fig. 4b were subject to manual scrutiny as described in the *Materials & Methods* section. The results for this procedure are displayed in Fig. 6. Groups with similar MASCOT and quality scores (*a* (true negatives), *c* (true positives)) were rated in very good agreement with their respective scores. Region *b* (false positives) represents the group containing spectra with high MASCOT scores, yet low predicted quality ones, and is difficult to analyze because the corresponding spot lies well outside the populated regions (cf. Fig. 4b). As a consequence, the number of spectra drawn from this region was reduced to 25. Even so, this procedure required including spectra up to a quality score of 30, which meant including samples with spectra better than desired.

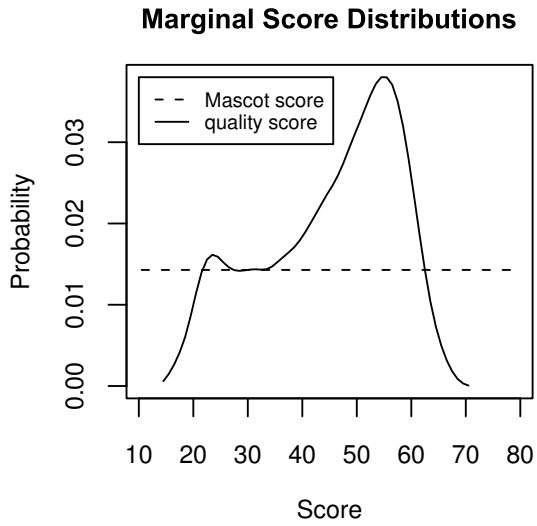

Figure 5: Marginal score probabilities for the test set in Fig. 4b. The curves were flattened using smoothing splines (for visualization only).

Because of that and due to the low number of spectra rated, this region eludes a valid quantitative analysis, and conclusions are somewhat ambiguous. Nevertheless, no excellent spectra were found at all, and the number of poor and intermediate ones clearly exceeds the frequency of good spectra.

The predictions of spectral quality in the regions featuring low MASCOT and high quality scores (d, false negatives) are corroborated by the independent human evaluation: not a single spectrum was classified as poor and only very few as intermediate. This gives compelling evidence that non-linear regression in conjunction with the composite feature set is able to generalize very well with regard to spectral quality and is therefore well suited to assist human observers in finding good spectra that are poorly interpreted by MASCOT.

Discussion. Since we were interested in obtaining a measure for spectral quality rather than exactly reproducing the MASCOT score, and because we designed our features to capture this aspect of a spectrum, we did not expect the same, constant distribution for the RF output as for the input. Considering the features to be a natural representation

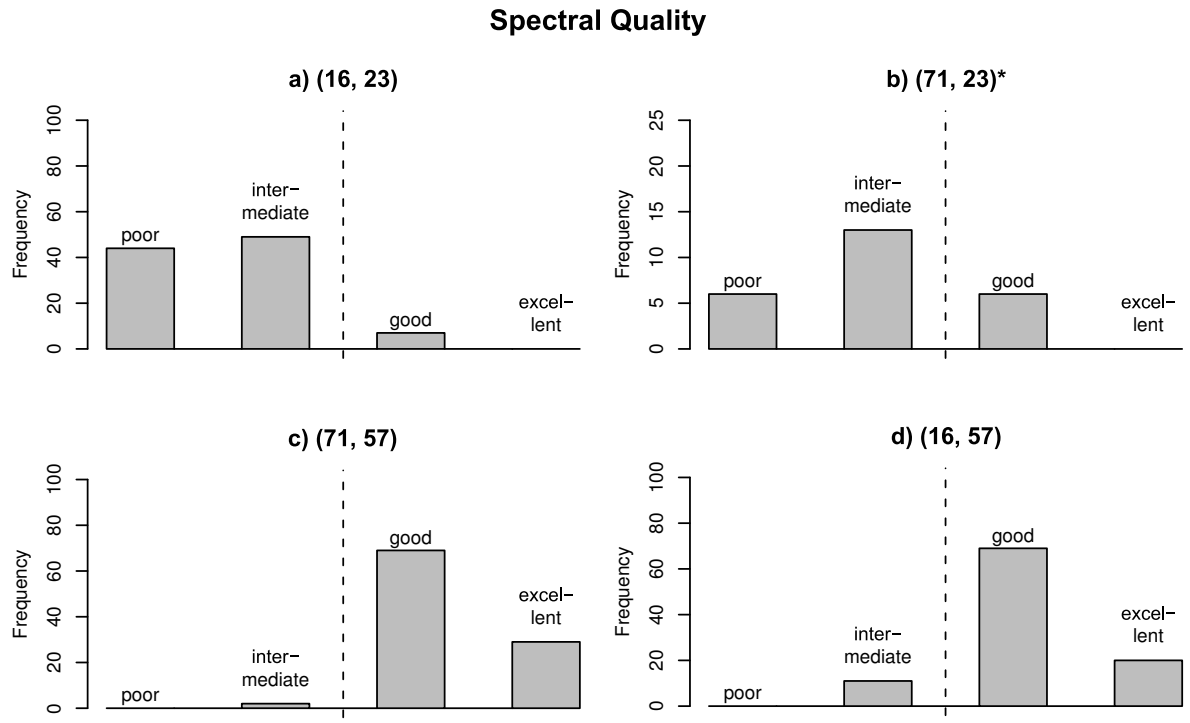

Figure 6: Results of the manual assessment that took place at the regions marked in Fig. 4b. The captions denote the coordinates in the density plot (MASCOT Score, quality score). *: see text.

of spectral quality, the prominent non-linearity must originate directly from the MASCOT scoring algorithm. The manual validation indicates that the resulting bimodal distribution is compatible with what experts consider spectral quality, so the regularization carried out by the regression is powerful enough to cope with imperfect training labels.

Consequences. This important property of *robustness against training label noise* allows for the training to be carried out on a large number of spectra in conjunction with their MASCOT scores. Given such a database, it enables users to easily train a regression according to the scheme presented, since the MASCOT search parameters used for training are not restricted. Therefore, such a quality score facilitates efficient *on-site calibration* and can be retrained every time it turns out to be necessary with the same low effort.

Interpretation. In addition to its suitability for rating spectral quality, regression seems to be an appropriate way to discover subtleties in the MASCOT score which otherwise are not observable: Fig. 4b suggests that spectral quality (q) as a function of the MASCOT score (m) behaves approximately logarithmically:
$q(m) = 14\ln(1.7m - 18) - 7.5$. In other words, the MASCOT score increases exponentially with spectral quality, meaning that a slight increase in quality, e.g. caused by an additional ion present, will cause a strong boost of the MASCOT score. As a consequence, the rather intuitive claim leading to the limitation of the score interval at the upper end is now justified from a data mining point of view.
To be precise, it is the nature of the chosen features which are responsible for the logarithmic flattening of the quality score. Since they were designed to capture spectral quality, they cannot reflect the results of the MASCOT scoring algorithm with its access to a sequence database to an arbitrary precision. Especially at high scores, the MASCOT score strongly depends on whether an additional ion present matches an entry in the underlying sequence database or not. The result is a behavior that is very similar to the intuitive perspective humans adopt for rating the quality of mass spectra, a property that nicely matches the aim of creating a tool that provides results which are easy to interpret. Additionally, in an approach similar to Moore *et al.*[23] and Peng *et al.*[24] using a reversed database, we have found (results not shown) a MASCOT score of about 33 to represent the optimal threshold for separating spectra corresponding to random matches from those which contain desired information. In this respect, the density minimum at a score of 30 revealed here (see Fig. 5) supports this result from a completely different viewpoint. It should be noted that a) this threshold score of 33 applies to searches against mammalian databases with 40.000 to 55.000 protein entries and b) the majority of the data used in this study were derived from searches against mammalian databases.
Suitability as Classifier and Feature Set Comparison. The results of our

regression-based classifier (which can be used to filter out bad spectra prior to the identification process) are shown in Fig. 7a. Our composite set performs best with an AUC of 0.91, incorporating both mass differences as well as features taken into account by human experts. It performs considerably better than the mass differences and the fast subset taken separately. The former was originally proposed in a work[12] that simultaneously assessed the performance of another set of features we are referring to as "Bern *et al.*".

According to Fig. 7a, accepting a false positive rate of 0.4 (i.e. 60% of the bad spectra will be filtered out) amounts to a true positive rate of 0.983, i.e. less than 2% of the good spectra get accidently removed. This case corresponds to a quality score threshold of about 38 and performs better than or equal to results reported in the literature (Flikka *et al.*[15]: 2%, Bern *et al.*[12]: 5%). We have therefore developed a method that can compete with existing techniques while providing both easy on-site calibration as well as a continous quality measure. We think it is important to remind the reader at this point that our regression was trained on multiply charged precursors only. As a consequence some of the misclassified spectra probably correspond to singly charged ions, which we did not filter out. Our results therefore give a pessimistic estimate for the classifier performance. The misclassification rate on a dataset comprising spectra arising from multiply charged precursors only should then be even lower, although we haven't formally checked for this. Fig. 7b shows the classification performance obtained by directly training a binary classifier instead of a regression. The results are a tad better, which might be due to the fact that the the regression was trained on a vast number of spectra in contrast to the classifier, which was trained *and* tested on spectra from a single sample only, therefore potentially "overfitting" to peptides predominantly contained in that sample.

Table 1 gives an overview of the feature sets used and their performances corresponding

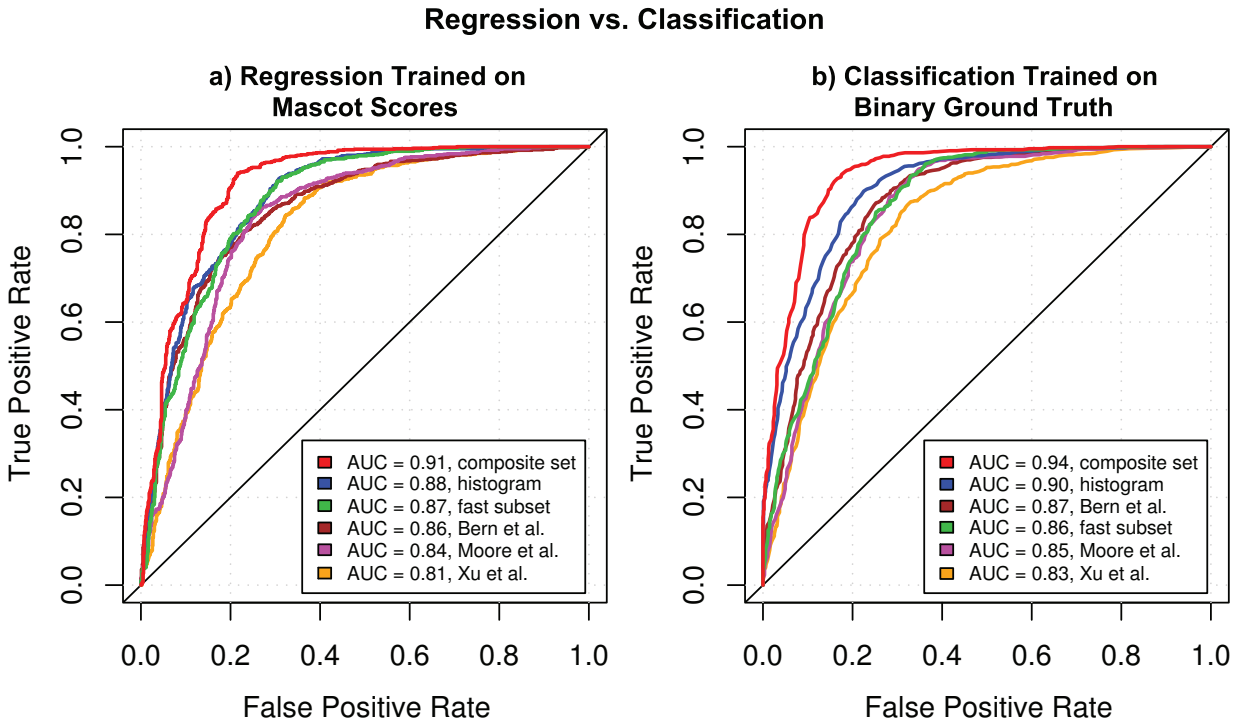

Figure 7: Classification performance for six feature sets: a) our regression turned into a classifier by varying the spectral quality score threshold; b) binary classifier (for comparison).

to the different experiments. In both classification cases, our composite set exhibits the highest AUC with values of 0.91 and 0.94, respectively.

Good, but Unidentified Spectra. As stated before, many proteomics applications aim at matching as many spectra as possible to peptides, many of which might represent novel forms of proteins that have never been annotated previously and therefore are not contained in the underlying sequence databases. These spectra would be mostly assigned low MASCOT scores despite having a high information content, that means we would expect them to be assigned quality scores in a regime ensuring them to be marked as high quality spectra. The simulation of this case by choosing false MASCOT settings for the modification yielded the results depicted in Fig. 8. It shows how many of the spectra searched with the wrong setting "acrylamide" are identified as "good": plotted is the

sensitivity as the function of the applied quality score threshold. This threshold defines the border between good and bad spectra according to the user who chooses its value. The higher the threshold, the stricter the classifier will select only the high quality spectra. Consequently, the number of spectra marked as "good" will decrease with increasing threshold. The question now is whether the performance of our method is good enough for this type of application.

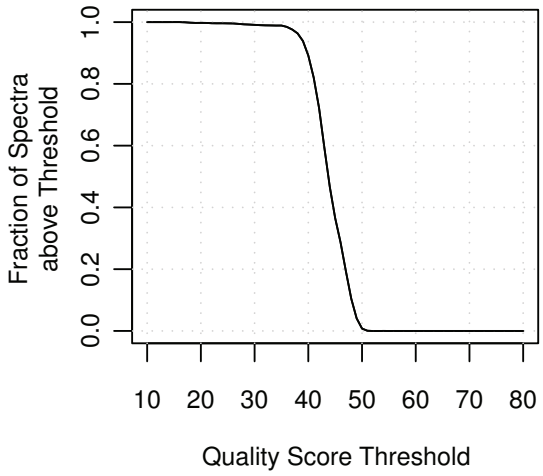

Figure 8: Fraction of modifications identified as good spectra when varying the quality score threshold. The higher the threshold, the lower the number of spectra rated "good" (see text).

Comparing Fig. 8 with our overall results stated in the previous paragraph, one finds that the score threshold of 38 roughly corresponds to 2% loss of good spectra during this application, too. This means, that this threshold will reject only about 2% of the meaningful spectra even if the MASCOT search parameters have been chosen in an inappropriate manner, while filtering out 60% of the low quality spectra. Good, but unmatched spectra can then be subject to manual inspection, searches with different

parameters or *de novo* sequencing.

Extension to Instrument Independent Features. To conclude this study, we would like to address an issue that usually is not paid attention to. In our composite set, we have incorporated three features which explicitly take into account peak intensities. However, the detector used in an LTQ mass spectrometer does not register single counts, but is based on an electron multiplier. The efficiency of this device depends on the voltage applied to the successive acceleration stages, which has to be adjusted over time to compensate for the deterioration of the used materials. Furthermore, increased trapping times and / or automatic gain control target values will result in higher product ion intensities. These facts can severely bias the results of a regression trained on different settings, and instead of attempting to normalize peak intensities to account for this behavior, we considered it a better strategy to construct a feature set not containing any explicit information about ion intensities at all. In fact, one can construct such a set by simply omitting the TIC, the average peak intensity and the standard deviation of the peak intensity from the composite feature set. The resulting features thus comprise the mass difference histogram, the number of peaks, the percentage of TIC above the precursor mass, the number of peaks greater than 1% TIC and the overall peak density. Although this subset was not a result of our backward feature selection strategy (probably due to correlations among the variables), the ROCs arising from the reduced set (not shown) are almost identical to the ones of the intensity dependent composite set (Fig. 7), and with a value of 0.906 the corresponding AUC turns out to be only slightly less than before (0.913) in case of the regression. It is therefore well suited for application in standard proteomics experiments, since it offers stability and portability between different instruments.

	composite set	detector independent set	fast subset	histogram	Bern *et al.*	Moore *et al.*	Xu *et al.*
TIC	•		•			log	
no. of peaks	•	•	•			log	
% of peaks above precursor mass	•	•	•				
mean intensity	•		•				
SD of intensity	•		•			•	
no. of peaks > 1% TIC	•	•	•				
peak density	•	•	•				
histogram features	•	•		•			
Good-Diff fraction					•		
isotopes					•		
water losses					•		
intensity balance					•		
BPI						log	
% of peaks > 1% BPI						•	
% of peaks > 20% BPI						•	
no. of peaks > 5% BPI							•
no. of peaks > 3% BPI							•
no. of peaks > 2% BPI							•
mean peak distance s.t. intensity > 2% TIC							•
mean peak distance s.t. 1.5% TIC < intensity < 2% TIC							•
MSE	**218**	227	246	241	269	321	292
AUC (regression)	**0.91**	0.91	0.87	0.88	0.86	0.84	0.81
AUC (classification)	**0.94**	0.93	0.86	0.90	0.87	0.85	0.83

Table 1: Shown are the assignments of the features to the seven sets examined, the final mean squared error (MSE) on the out-of bag sample after training the regression and the respective areas under curve (AUC) for regression and classification. Abbreviations used are TIC (total ion current), BPI (base peak intensity), SD (standard deviation) and log (logarithm of the respective quantity).

Conclusions

We have presented a score that assesses the quality of peptide product ion spectra in proteomics using a non-linear regression technique. We have been able to show that this score can be trained by using a database of spectra in conjunction with their respective MASCOT scores, i.e. without a manually labeled training set. It has turned out that the regression was not adversely affected by the false labels which are unavoidable when using an entire database of spectra as training set. This makes the proposed quality score well suited for *on-site calibration*, rendering it possible to easily adapt the quality score to different instrumentation according to the scheme presented.

Furthermore, it has been shown that this method performs better than or equal to classifiers presented in recent studies when it comes to the prefiltering of spectra prior to database searches, while at the same time equipping researchers with a continuous measure for spectral quality. By benchmarking several sets of features introduced in the past, we found that the proposed set performs best, and it has become apparent that only the combination of features capturing the visual appearance of a spectrum with those used by peptide matching algorithms (i.e. a mass difference histogram[12]) allows for a significantly increased classification performance.

Finally, a set of features has been introduced that does not require retraining the regression after the trapping time or the settings of the electron multiplier have been changed. It may even be possible to use a regression trained on the LTQ with a completely different mass spectrometer, provided that fragmentation mechanisms are comparable.

All this is done without incorporating any search results, so once trained, the regression might as well be used in conjunction with other protein identification techniques, such as SEQUEST or even a *de novo* sequencer, without any adjustments to the core of the

algorithm.

Altogether, a precise, robust and flexible scheme has been introduced for both the prefiltering of data as well as for assisting users in finding high-quality spectra that cannot be assigned a corresponding peptide by automated procedures.

Acknowledgements

The authors gratefully acknowledge financial support by the German Academic Exchange Service (T. K.), the Hans L. Merkle foundation (M. K.), the Robert Bosch GmbH (F. A. H.), the DFG under grant number HA-4364/2 (M. K., F. A. H.) and the Children's Hospital Trust (J. J. S. and H. S.).

References

[1] Choudhary J. S.; Blackstock W. P.; Creasy D. M.; Cottrell J. S. Interrogating the Human Genome Using Uninterpreted Mass Spectrometry Data. *Proteomics* **2001**, *1*, 651-667. Erratum in: *Proteomics* **2001**, *1*, 796.

[2] Mann M.; Pandey A. Use of Mass Spectrometry-Derived Data to Annotate Nucleotide and Protein Sequence Databases. *Trends Biochem. Sci.* **2001**, *26*, 54-61.

[3] Matrix Science, *http://www.matrixscience.com* (accessed 5/18/07).

[4] SEQUEST, *http://fields.scripps.edu/sequest* (accessed 5/18/07).

[5] Taylor J. A.; Johnson R.S. Implementation and Uses of Automated De Novo Peptide Sequencing by Tandem Mass Spectrometry. *Anal Chem.* **2001**, *73*, 2594-604.

[6] Wisniewski J.R.; Zougman A.; Kruger S.; Mann M. Mass Spectrometric Mapping of Linker Histone H1 Variants Reveals Multiple Acetylations, Methylations, and Phosphorylation as well as Differences between Cell Culture and Tissue. *Mol. Cell. Proteomics* **2007**, *6*, 72-87.

[7] Sun W.; Li F.; Wang J.; Zheng D.; Gao Y. AMASS: Software for Automatically Validating the Quality of MS/MS Spectrum from SEQUEST Results. *Mol. Cell. Proteomics* **2004**, *3*, 1194-1199.

[8] Razumovskaya J.; Olman V.; Xu D.; Uberbacher E. C.; VerBerkmoes N. C.; Hettich R. L.; Xu Y. A Computational Method for Assessing Peptide-Identification Reliability in Tandem Mass Spectrometry Analysis with SEQUEST. *Proteomics* **2004**, *4*, 961-969.

[9] Anderson D. C.; Qeiqun L.; Payan D. G.; Noble W. S. 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.* **2003**, *2*, 137-146.

[10] Keller A.; Nesvizhskii A. I.; Kolker E.; Aebersold R. Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **2002**, *74*, 5383-5392.

[11] Moore R.; Young M. K.; Lee T. D.; Method for Screening Peptide Fragment Ion Mass Spectra Prior to Database Searching. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 422-426.

[12] Bern M.; Goldberg D.; McDonald W. H.; Yates J. R. Automatic Quality Assessment of Peptide Tandem Mass Spectra. *Bioinformatics* **2004**, *20*, i49-i54.

[13] Xu M.; Geer L. Y.; Bryant S. H.; Roth J. S.; Kowalak J. A.; Maynard D. M.; Markey S. P. Assessing Data Quality of Peptide Mass Spectra Obtained by Quadrupole Ion Trap Mass Spectrometry, *J. Proteome Res.* **2005**, *4*, 300-305.

[14] Salmi J.; Moulder R.; Filn J.-J.; Nevalainen O. S.; Nyman T. A.; Lahesmaa R.; Aittokallio T. Quality Classification of Tandem Mass Spectromtry Data. *Bioinformatics* **2006**, *22*, 400-406.

[15] Flikka K.; Martens L.; Vandekerckhove J.; Gevaert K.; Eidhammer I. Improving the Reliability and Throughput of Mass Spectrometry-Based Proteomics by Spectrum Quality Filtering. *Proteomics* **2006**, *6*, 2086-2094.

[16] Hastie T.; Tibshirani R.; Friedman J.H. The Elements of Statistical Learning. *Springer*, **2003**

[17] Breiman L. Random Forests. *Machine Learning* **2001**, *45*, 5-32.

[18] Shevchenko A.; Wilm M.; Vorm O.; Mann M. Mass Spectrometric Sequencing of Proteins from Silver-Stained Polyacrylamide Gels. *Anal. Chem.* **1996**, *68*, 850-858.

[19] Shilov I.; Seymour S; Patel A. A., Loboda A.; Tang W. H.; Keating S. P.; Hunter C. L.; Nuwaysir L. M.; Schaeffer D. A. The Paragon Algorithm, a Next Generation Search Engine that Uses Sequence Temperature Values and Feature Probabilities to Identify Peptides from Tandem Mass Spectra. *Mol Cell Proteomics* **2007**, *6*, 1638-55.

[20] R Development Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing **2007**, http://www.R-project.org (accessed 2/2/07)

[21] Menze B. H.; Kelm B. M.; Weber M.-A.; Bachert P.; Hamprecht F. A. Mimicking the Human Expert: Pattern Recognition for an Automated Assessment of Data Quality in Magnetic Resonance Spectroscopic Images. *Magnetic Resonance in Medicine*, in press

[22] Liaw A.; Wiener M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18-22

[23] Moore R.; Young M. K.; Lee T. D. Qscore: An Algorithm for Evaluating SEQUEST Database Search Results. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 378-386.

[24] Peng J.; Elias J. E.; Thoreen C. C.; Licklider L. J.; Gygi S. P. Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome. *J. Proteome Res.* **2003**, *2*, 43-50.