A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores

D.C. Anderson*, Weiqun Li, Donald G. Payan,

and William Stafford Noble**

Rigel Incorporated, 240 East Grand Avenue, South San Francisco, California 94080

*** Department of Genome Sciences, University of Washington, Seattle, Washington 98195*

*To whom correspondence should be addressed: (650)624-1104, dca0210@earthlink.net

Running title: machine learning analysis of peptide shotgun sequencing data

Keywords: shotgun peptide sequencing, SEQUEST, support vector machine, machine learning, mass spectrometry, capillary LC/MS/MS, proteomics

Abstract

Shotgun tandem mass spectrometry-based peptide sequencing using programs such as SEQUEST allows high-throughput identification of peptides, which in turn allows identification of corresponding proteins. We have applied a machine learning algorithm, called the support vector machine, to discriminate between correctly and incorrectly identified peptides using SEQUEST output. Each peptide was characterized by SEQUEST-calculated features such as delta Cn and Xcorr, measurements such as precursor ion current and mass, and additional calculated parameters such as the fraction of matched MS/MS peaks. The trained SVM classifier performed significantly better than previous cutoff-based methods at separating positive from negative peptides. Positive and negative peptides were more readily distinguished in training set data acquired on a QTOF compared to an ion trap mass spectrometer. The use of 13 features, including four new parameters, significantly improved the separation between positive and negative peptides. Use of the support vector machine and these additional parameters resulted in a more accurate interpretation of peptide MS/MS spectra, and is an important step towards automated interpretation of peptide tandem mass spectrometry data in proteomics.

Introduction

The separation and sequencing by capillary hplc-tandem mass spectrometry of femtomole (or below) peptide levels is the basis for the high-throughput identification of proteins present in cell or tissue samples. The technique has broad applicability: applications include the identification of peptides binding individual MHC proteins of defined haplotype[1], identification of a peptide recognized by melanoma specific human CTL cell lines[2], the identification of individual protein complexes[3-4], large scale analysis of the yeast proteome[5], identification in yeast of interacting proteins for a large number of tagged protein baits[6-7], identification of proteins in urine[8], and definition of proteins of the nucleolus[9].

The analysis of peptide collision-induced dissociation spectra to give information on a peptide's sequence was developed by Hunt and coworkers[10-14] and Biemann[15]. To identify proteins from mass spectrometry data, protein database searches initially used peptide fragments[16] or sequence tags[17], and included sequenced genomes[18] and more sophisticated search techniques[19-22]. Yates and co-workers developed correlations of peptide tandem mass spectrometry data and sequences from protein databases[23-25], incorporated these in the program SEQUEST, and coupled this software with capillary LC/MS/MS data and database searches to identify proteins[26] and protein complexes[27]. Due to its early implementation, availability and the widespread use of ion trap, triple quadrupole, and quadrupole time-of-flight mass spectrometers that generate compatible data, SEQUEST is one of the most commonly used programs.

The use of database search programs introduces questions about how to interpret their output. SEQUEST outputs for each spectrum one or more peptides from the given database whose theoretical spectra closely match the given spectrum. Associated with each match is a collection of statistics. Initially, the difference between normalized cross-correlation functions (delta Cn) for the first and second ranked results from a search of a relatively small database was used to indicate a correctly selected peptide sequence[23, 25]. Additional criteria were subsequently added, including the cross-correlation score between the observed peptide fragment mass spectrum and the theoretically predicted one (Xcorr), followed by a manual examination of the MS/MS spectra[27]. More stringent criteria combined the use of Xcorr cutoffs, delta Cn, and the correspondence of peptide sequences with those expected for cleavage with the enzymes used for proteolysis[5, 28].

Recently, Moore et al. described a probabilistic algorithm called Qscore[29], for evaluating SEQUEST database search results. In contrast to previous heuristic techniques, Qscore is based upon a probability model which includes the expected number of matches from a given database, the effective database size, a correction for indistinguishable peptides, and a measurement of match quality. The algorithm performs well in distinguishing between true and false matches from SEQUEST outputs.

The approach described here addresses a similar problem using a different approach. Rather than building an algorithm by hand, we use a machine learning algorithm, called the support vector machine (SVM), to learn to distinguish between correctly and incorrectly identified peptides. The support vector machine (SVM) [30-32] is a supervised learning algorithm, useful for recognizing subtle patterns in complex data sets. The algorithm has been applied in domains as diverse as text categorization, image recognition, hand-written digit recognition[32] and in various bioinformatics domains, including protein remote homology detection[33], protein fold recognition[34], and microarray gene expression analysis[35-36]. The SVM is fundamentally a binary classifier: given two classes of data, the SVM learns to distinguish between them and to predict the classification of previously unseen examples. In the application described here, the algorithm is trained from a labeled collection of SEQUEST outputs, where the labels indicate whether the peptide represents a correct or incorrect identification. The SVM then learns to distinguish between peptides that were correctly and incorrectly identified by SEQUEST.

The SVM algorithm is surprisingly simple. It treats each training example as a point in a high-dimensional space and searches for a hyperplane that separates the positive from the negative examples. As such, the SVM is closely related to the perceptron algorithm[37], with three important differences. First, motivated by statistical learning theory[31], the SVM searches for a hyperplane that separates the two classes with the largest margin; i.e., the SVM finds a hyperplane which maximizes the minimum perpendicular distance to any training example. Choosing the maximum margin hyperplane reduces the chance that the SVM will overfit the training data. Second, for data sets that are not separable by a simple hyperplane, the SVM uses a mathematical trick, known as the kernel trick, to operate implicitly in a higher-dimensional space. By increasing the dimensionality of the space in which the points reside, the SVM can learn complex decision boundaries between the two given classes. Finally, for data sets that contain some mislabelled examples, the SVM incorporates a soft margin. The SVM may find a decision boundary that nearly, but not perfectly, separates the two given classes. A few outlier examples are allowed to fall on the wrong side of the decision boundary.

Here we use tryptic digests of mixtures of known protein standards, purified proteins, or of a variety of affinity extracts by specific antibodies or other binding proteins, to generate

LC/MS/MS data using ion trap or quadrupole time-of-flight (QTOF) mass spectrometers. Peptides are classified as positive examples (derived from proteins known or expected to be in the samples) or negative examples (peptides not expected to be in the samples). Each example in the training set is characterized by a vector of features, including observed data (peptide mass, precursor ion intensity) and SEQUEST-calculated statistics (such as the parameters Xcorr, delta Cn, Sp, and RSp). These labelled vectors are then used to train an SVM to distinguish between positive and negative examples.

Our experiments show that the trained SVM, when tested on its ability to classify previously unseen examples, exhibits high sensitivity and specificity. We illustrate the learning procedure using two differently-sized databases, as well as using data generated on ion trap and QTOF mass spectrometers. The SVM yields significantly fewer false positive and false negative peptides than any of the cutoffs previously proposed, and gives a cleaner separation of positive and negative peptides than Qscore-based single peptide analysis. The trained SVM is an accurate, high-throughput technique for the examination of SEQUEST results, which will enable the processing of large amounts of data generated from examinations of complex mixtures of proteins.

Experimental Section

Peptide Samples. Tryptic digest test mixtures containing 500 pmol of reduced, iodoacetic acid-alkylated hen egg white lysozyme, horse myoglobin and horse cytochrome c, bovine serum albumin and bovine carbonic anhydrase were purchased from Michrom Bioresources (Auburn CA). These standards were mixed so that individual test samples contained from 5 to 80 fmol of each of the five proteins, with 2-fold differences in each concentration. Affinity extracts of cultured human Jurkat cells were prepared using antibodies specific for individual antigens, and were carried out as described[38]. Individual baits or antigens and the source of the antibodies used for the affinity extractions included heat shock protein 90 (MA3-010 antibody, Affinity BioReagents, Golden CO), RbAp48 (13D10 antibody, Upstate Biotechnology, Lake Placid, NY), the synthetic C-terminal p21$^{cip1/waf1}$ peptide biotin-GSGSGSGSGSKRRQTSMTDFYHSKRRLIFS-acid, the fusion protein glutathione S-transferase-S5a (AFFINITI Research Products Ltd., Exeter, UK), and green fluorescent protein (rabbit polyclonal antibody, Molecular Probes, Eugene OR).

Mass Spectrometry and Database Searches. Ion trap mass spectrometry utilized a Finnigan LCQ (ThermoFinnigan, San Jose CA) and an LC Packings (San Francisco CA) Ultimate capillary hplc and custom packed 75 micron internal diameter capillary C18 reversed phase columns for sample injection and chromatography, as described[38]. Quadrupole time of flight mass spectrometry was carried out using a Micromass QTOF-1 mass spectrometer coupled to an LC Packings capillary hplc as above. Peptides were eluted using a 1% acetonitrile/min. gradient. Database searches utilized either the nonredundant human protein database of March 15, 2002, or the nonredundant protein database of March 6, 2002. Both were downloaded from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov). Proteins from the human immunodeficiency virus were removed from the nonredundant human protein database before use. The version of SEQUEST (ThermoFinnigan, San Jose, CA) used for database searches was SEQUEST 2.0 that was distributed with Sequest Browser. Tryptic cleavages at only lys or arg and up to two missed internal cleavage sites in a peptide were allowed. The maximal allowed uncertainty in the precursor ion mass was 1.5 m/z. SEQUEST searches allowed optional met oxidation and cys carboxamidomethylation since cysteines were derivatized in this fashion after protein thermal denaturation and reduction. Peptides with masses from 700-3500 m/z and precursor charge states of +1, +2 and +3 were allowed. A few peptides analyzed on the QTOF-1 were present as +4 ions and were left in the appropriate positive or negative category. For spectra collected on the LCQ, the minimum total ion current required for precursor ion fragmentation was 1.0 x 10^5, the minimum number of ions was 25, and IonQuest filtering was turned off. Single precursor ion scans from 350 - 1800 m/z were followed by 6 MS/MS scans from 50 to twice the precursor ion m/z, up to a limit of 1800 Da.. Dynamic exclusion was turned on for a duration of 1 min. A collision energy on the LCQ of 30 was used for all fragment ion spectra. For the QTOF-1, a precursor charge-dependent and peptide mass-dependent collision energy was used, ranging from 16-55 ev for +1 ions of 388-2000 Da, 22-65 ev for +2 ions of 400-2000 Da, 16-50 ev for +3 ions between 435-2000 Da and 19-36 ev for +4 ions between 547-2000Da. For database searches using non-human protein test samples, sequences for the non-human proteins were added to the nonredundant human protein database.

Positive and negative peptides. Positive peptides were selected by several criteria. One was tryptic peptides from five known proteins in tryptic digest standards. A second was peptides from proteins expected to be present in affinity extracts because they are derived from the antibody or affinity reagent used in the extraction, from the known antigen for the antibody, from known interacting partners of the antigen, or are autolytic fragments of trypsin. A third category includes peptides from extracted proteins thought to be present due to the identification of at least two peptides from that protein with SEQUEST scores that meet stringent criteria[5, 28]. This includes common contaminating proteins such as myosin, heat shock proteins, defined cytokeratins, and may include proteins not previously demonstrated to interact with a particular bait. Tryptic digests from isolated protein standards were injected at different levels between 5-1000 fmol to include SEQUEST scores from peptides with strong as well as weak signals.

Negative peptides were selected from tryptic digests of known protein standards, in which these peptides were assigned by SEQUEST to proteins other than the injected protein or its human homolog. A second category of negative peptides included peptides selected from lower scoring peptide matches (i.e. from incorrect proteins) to MS/MS data from peptides from a known standard protein.

Construction of training sets. Training sets were constructed using data collected and analyzed under three different conditions: data collected using an ion trap mass spectrometer and analyzed using the nonredundant human and full nonredundant databases, and data collected on a QTOF mass spectrometer analyzed using the nonredundant human database. All three sets included 9 experimentally measured and SEQUEST-calculated parameters[23-25]: MS/MS spectrum total ion current, peptide charge, peptide precursor ion mass, the difference in observed and theoretical precursor ion masses for the best-fit peptide, the SEQUEST variables Xcorr (cross-correlation score of the observed to the theoretical MS/MS spectrum for a peptide sequence), delta Cn (the magnitude of the difference in normalized cross-correlation parameter values for the first and second hits found by SEQUEST), Sp (the preliminary score for a peptide after correlation analysis to the predicted fragment ion values), RSp (the final correlation score rank), and the percent of predicted y and b ions matched in the MS/MS spectrum.

A training set representing ion trap data and a SEQUEST nonredundant human database search was constructed containing 696 positive peptides, including 338 unique peptides representing 47 different proteins. Multiple copies of some individual peptides were obtained from independent runs using from 5 fmol to 1 pmol of individual standard proteins, resulting in peptides with a large dynamic range in signal to noise. There were a total of 465 negative peptides, of which 435 were unique; 30 negatives were generated using peptides that were second or lower choices below a top ranked positive peptide. Initial support vector machine calculations incorrectly assigned negative labels to a number of positive peptides. Upon examination, in a number of cases the top ranked peptide from a SEQUEST database search, for a given precursor ion and MS/MS spectrum, was instead from a different protein. SEQUEST had selected a lower ranked peptide from the protein of interest and incorrectly listed it as being top ranked. As a result of this round of SVM calculations, all gi or accession numbers for positive peptides were verified as corresponding to the protein identified, and a number of positive peptides with relatively low scores were individually blasted against the nonredundant database to check the identity of their source protein.

A second training set representing ion trap data and a SEQUEST search using the full nonredundant database was constructed. It contained 497 positive peptides assigned to 280 unique sequences from 33 different proteins. It also contained 479 negative peptides assigned to 460 different peptide sequences; 67 negatives were generated using peptides that were second or lower choices below a top ranked postive peptide. This database had approximately 8 times as many sequences as the nonredundant human protein database, and may be useful for finding protein homologs from other organisms when the human protein sequence is not in a database (for analyses using human cells) or for analysis of non-human samples. The most significant outliers from initial SVM analyses were examined to uncover errors in SEQUEST-peptide sequence assignment or errors in data handling.

A third training set representing QTOF data and a SEQUEST search using the nonredundant human database was also constructed. It contained 1017 positive and 532 negative peptides analyzed on a quadrupole time-of-flight mass spectrometer. This training set was created for comparison with the two previous training sets since data for these peptides was collected on a different instrument. The positive peptides were derived from 45 different proteins, and represented 493 unique sequences. The negatives contained MS/MS spectra assigned to 335 different sequences. Seventy additional negative peptides were derived by selecting lower choices than the top ranked peptide, when the top ranked peptide was correctly assigned to a known protein from a standard peptide map. As before, initial support vector machine analysis was used to uncover mistakes in data entry or incorrect assignments of sequences to proteins by SEQUEST, by analysis of individual false positives and false negatives.

Four new parameters used to evaluate SEQUEST output. The basic parameters used to evaluate SEQUEST output included experimentally measured or calculated parameters such as precursor ion mass, precursor ion current, or peptide charge. They also included those calculated using SEQUEST: mass difference between observed and calculated precursor ions for the best fit sequence, Xcorr, delta Cn, Sp, RSp, and % y and b ions matched. Four additional parameters were measured or calculated. These included a count of the number of peaks in the MS/MS spectrum, and the fraction of these peaks matched by predicted peptide fragments. An MS/MS peak for ion trap data was defined as having over 10^3 counts, and for QTOF-1 data as having over 1 count. In a noisy MS/MS spectrum, the fraction of matched peaks should be low, for both positive and negative peptides. In other MS/MS spectra it should be lower for negative than for positive peptides. A third parameter is the fraction of the MS/MS spectrum total ion current that is in matched peptide fragments. For a good match, this fraction should be high, and for a poor match it should be low. A fourth parameter is the sequence similarity between the top peptide choice and second ranked choice. When delta Cn is low, this parameter is intended to mark these peptides for further examination. When the value of this parameter is close to 1 (high sequence similarity) and other scores are good, the individual peptides (and consequently proteins) identified are examined to see if they are similar. If so, the identification may be useful. If the sequences are different, a unique peptide/protein is not defined by the combined scores.

Support Vector Machine calculations. The SEQUEST output data is summarized in a (number of peptides) by (9 or 13 parameter) matrix, in which each row contains a vector consisting of the SEQUEST output parameters associated with a particular protein. This data is then normalized in two ways. First, in order to give equal importance to each of the features, the columns of the matrix are normalized by dividing each entry by the column sum. This operation ensures that the total for each column is 1.0. Second, each 9- or 13-element vector is converted to unit length by dividing each vector component by the Euclidean length of the vector. This operation projects the data onto a unit sphere in the 9- or 13-dimensional space defined by the data. Note that this latter normalization can be performed in the feature space, by defining a normalized kernel K' in terms of the original kernel K:

$$K'(X,Y) = \frac{K(X,Y)}{\mathrm{sqrt}(K(X,X)\ K(Y,Y))}$$

The kernelized normalization has the advantage of implicitly operating in the higher-dimensional kernel space.

Support vector machines are trained using a simple optimization algorithm[33]. A software implementation in ANSI C is freely available at http://microarray.cpmc.columbia.edu/gist. The output of the SVM optimization is a set of weights, one per peptide in the training set. The magnitude of each weight reflects the importance of that peptide in defining the separating hyperplane found by the optimization: peptides with zero weights are correctly classified and far from the hyperplane; peptides with small weights are correctly classified and close to the hyperplane; peptides with large weights are incorrectly classified by the hyperplane, as described next. The SVM weights, together with the original training set, can be used to predict the classification of a previously unseen peptide vector.

In most classification tasks, the positive and negative class labels assigned to the training set are not 100% correct. Therefore, the SVM employs a soft margin, which allows some of the training examples to fall on the "wrong" side of the separating hyperplane, as shown in Figure 1. An SVM soft margin may be implemented in several ways. We employ a 2-norm soft margin, which charges each misclassified example with a penalty term that increases quadratically according to the example's perpendicular distance from the hyperplane. In order to account for differences in the number of positive and negative examples, errors in the positive class (for which we have fewer examples) are charged more heavily than examples in the negative class. The asymmetric 2-norm soft margin is implemented by adding a constant to the diagonal entries in the kernel matrix[32]. The diagonal factor added to K(X,X) is 0.2 * (n_X / n), where n_X is the number of training examples in the same class as example X, and n is the total number of training examples[35].

To test the generalization performance of the algorithm, the SVM is trained and tested using leave-one-out cross-validation. In this paradigm, a single example is removed from the matrix, and the SVM is trained on the remaining examples. The resulting classifier is applied to the held-out example, and the predicted classification is compared to the true classification. The held-out example is counted as a true positive, false positive, true negative or false negative, depending upon the agreement between the true and predicted class. This leave-one-out procedure is repeated for every example in the data set.

Evaluation of results. A straightforward method for evaluating the quality of the predictions made by the SVM is to compare the classifications assigned by the SVM to the classifications assigned a priori. Disagreements between the two are counted either as false positives or false negatives. Prediction quality can be measured more precisely using the receiver operating characteristic (ROC) curve. Rather than depending upon a particular classification threshold, the ROC curve integrates information about the complete ranking of examples created by the SVM. The ROC curve plots, for varying classification thresholds, the number of true positives as a function of the number of false positives. The area under this curve, normalized to range from 0 to 1, is called the ROC score. A perfect classifier will rank all of the positive examples above negative examples, yielding an ROC score of 1. A random classifier will produce an approximately diagonal curve, yielding a score close to 0.5.

Results and Discussion.

The SVM provides good discrimination performance on three different data sets. Support vector machine calculations were run on all three datasets, and the results compared (Table 1). For the dataset derived from ion trap mass spectrometry and a SEQUEST search of the nonredundant human protein database, there were 48 false positives, 117 false negatives, 579 true positives and 417 true negative peptides and a ROC score of 0.929. Of the initial training set peptides, 14.2% were false positives or negatives. For the dataset derived from ion trap mass spectrometry and a SEQUEST search of the full nonredundant human protein database, there were 62 false positives, 81 false negatives, and a ROC score of 0.920. Of these peptides, 14.7% were false positives or negatives. For QTOF mass spectrometry data, searched using the nonredundant human database, calculations found 27 false positive and 81 false negative peptides. The ROC score for this analysis was 0.981; 7.0% of these peptides were false positives or negatives.

ROC plots for the 3 datasets examined with 9 parameters are shown in Figure 2A. Use of the full nonredundant protein database, containing approximately 8-fold more sequences, still allows a good separation between positive and negative peptides, but the ROC scores are slightly lower than for the smaller nonredundant human database. Using the same nonredundant human database for comparison, data collected on a quadrupole time-of-flight mass spectrometer is more readily separated by the SVM into positives and negatives than data collected on this ion trap.

In order to understand the errors made by the SVM, we looked in detail at each of the false positives and negatives. Many of the errors made by the SVM correspond to examples with noisy spectra or poor fragementation. For the ion trap-nonredundant human protein database training set, 7 of the 25 top false positive peptides had noisy MS/MS spectra, and another 5 had poor fragmentation, with much of the ion current in a few major peaks. Nine of the top 25 false negatives had noisy MS/MS spectra, while 13 had poor fragmentation. For the ion trap-full nonredundant protein database training set, 6 of the top 22 false positives had noisy MS/MS spectra and an additional spectrum had poor fragmentation of the precursor ion. Seven of the top 24 false negatives had noisy MS/MS spectra, and 7 had poor precursor ion fragmentation. For the QTOF data, 4 of the top 20 false positive peptides had low signal-to-noise MS/MS spectra and an additional 4 fragmented poorly. Eight of the top 23 false negatives fragmented poorly, and 4 had noisy MS/MS spectra. A lower information content could make it difficult to match the correct peptide sequence for peptides with poor MS/MS fragmentation or noisy MS/MS spectra.

For each of the three training sets, some of the false positives or false negatives that did not have noisy MS/MS spectra, or poorly fragmenting precursor ions, matched the predicted MS/MS spectrum from the best-fit peptide fairly well. It is possible that some of the false positives were contaminants of the known proteins used as standards, and thus were true positives. Some of the false negatives had poor SEQUEST scores and the SVM had trouble recognizing them as positive peptides. Overall the these peptides seem to represent a core of peptides that are currently difficult to correctly assign with the parameters used.

Using four additional parameters improves the SVM's performance. Based upon the initial analyses described above, we computed four additional parameters that we hypothesized would help the SVM recognize noisy or otherwise difficult examples. These parameters were tested in the analysis of the three training sets. The use of the number of peaks in an MS/MS spectrum, and the fraction of those peaks matched by fragments predicted from the best-fit database peptide sequence, was intended as an additional measure of the goodness-of-fit of a peptide sequence to the data. The fraction of the total ion current in the MS/MS spectrum matched by predicted peptide fragments was intended as an additional measure for the goodness-of-fit of a peptide sequence to the data, and to weight the fit by the intensity of the matched fragments. The sequence similarity between the top sequence and second choice sequence was intended to allow discrimination, for peptides with low delta Cn values, between dissimilar sequences almost equally well-matched to the data, and very similar sequences matched to the data. In the former case the top ranked sequence is not useful, while in the latter case it may be useful if the matched peptides are from similar proteins.

Training sets were constructed as above for positive and negative peptides associated now with 13 parameters, the original 9 and the four additional parameters described above (Table 1). For the ion trap-nonredundant human protein database training set with SVM calculations, there were 44 false positives, 100 false negatives, and the ROC score was improved to 0.950. This represents a loss of 4 false positives and 17 false negatives compared to the 9-parameter dataset. 12.4% of the peptides were false positives or negatives. Addition of these parameters thus improved the overall performance of the SVM calcuations.

For SVM calculations on the ion trap- full nonredundant protein database training set, use of the additional 4 parameters resulted in a reduction of false positives to 53 and the false negatives to 70. The ROC score was now 0.939; 12.6% of the peptides were false positives or negatives. Thus for this training set the use of the additional parameters also increased the separation between the positive and negative peptides.

For SVM calculations on the QTOF- full nonredundant protein database training set, the total false positive peptides decreased from 26 to 18, and the false negative peptides decreased from 81 to 51. The ROC score for this analysis was 0.988; now only 4.5% of the training set peptides were found to be false positives or negatives. Thus the best separation of positives and negatives for any training set was obtained using QTOF-collected data and 13 parameter analysis. The QTOF data was collected without internal calibration of each run, and SEQUEST searches utilized a 1.5 Da window. Thus the higher mass accuracy available with internal calibration or more advanced instruments may further improve the separation of these positive and negative peptides. The average mass deviation between observed and best-fit peptides for the positive peptides for QTOF data was 0.40 ± 0.25 Da, compared to an average mass deviation for ion trap positive peptides of 0.52 ± 0.38 Da. Thus the uncalibrated QTOF data as used here appears to have a slightly higher mass accuracy than ion trap data.

For all three datasets ROC scores increase with the use of the 4 additional parameters beyond the original 9 parameters. ROC curves for the 3 datasets examined with 13 parameters are shown in Figure 2B. For the full nonredundant protein database, containing ca. 8-fold more sequences, there is still a good separation between positive and negative peptides, but the ROC scores are slightly lower than for the smaller nonredundant human database. For the same database, data collected on the QTOF mass spectrometer is more readily separated by the SVM into positives and negatives than data collected on an ion trap. The QTOF ROC scores are noticeably higher for both 9 and 13-parameter training sets.

For the parameter representing the fraction of MS/MS peaks matched by predicted peptide fragments, this value was slightly higher in the ion trap training sets for positive peptides (0.499 ± 0.120 and 0.512 ± 0.113 for the NR human and full NR database sets) than for negative peptides (0.410 ± 0.098 and 0.438 ± 0.090 respectively). The difference was more pronounced for QTOF-1 training set data: 0.632 ± 0.120 for positive peptides, 0.352 ± 0.139 for negative peptides. For the parameter representing the average fraction of MS/MS total ion current matched by predicted peptide fragments, its value was slightly higher for ion trap positive peptides (0.646 ± 0.163 and 0.656 ± 0.156 for the NR human and full NR datasets) than for negative peptides (0.468 ± 0.153 and 0.520 ± 0.141 respectively). The difference was more pronounced for QTOF-1 data (0.750 ± 0.112 and 0.392 ± 0.182 for positive and negative peptides). This suggests that the QTOF-1 data may be less noisy than ion trap data, which is consistent with an examination of the MS/MS spectra.

Fisher scores can be used to understand what features are providing the most information. Although the support vector machine generally produces very accurate predictions, this accuracy comes at the price of reduced explanatory power. Unlike a decision tree classifier, the SVM does not explicitly select a few features that are most relevant to the classification task at hand. However, we can use a related technique to analyze the correlations between each feature and the classification labels associated with each peptide. The Fisher criterion score (FCS)[40] is a simple metric that is closely related to the Student's t-test. The score was developed in the context of linear discriminant analysis, which is closely related to the SVM methodology. The FCS has been used previously for feature selection in conjunction with the SVM classification of microarray data[36]. For a given pair of distributions A and B, with means A_m and B_m and standard deviations σ_A and σ_B, the FCS is defined as

$$\frac{(\{A_m\} - \{B_m\})^2}{\sigma_A + \sigma_B}$$

Here, A and B correspond to the distributions of a given feature (say, Xcorr) within the positive and negative training sets, respectively. A high FCS indicates that the distribution of Xcorr scores associated with positively labeled peptides is markedly different from the Xcorr scores associated with negatively labeled peptides. We can compute the FCS for each feature in our data set, and rank the features to determine which ones are providing the most information to the SVM.

Unfortunately, SVM results are particularly difficult to explain because the SVM can operate in a higher-order feature space defined by the kernel function. In general, it is not possible to compute Fisher criterion scores of the features in this high-dimensional space. Indeed, for some functions, such as the radial basis function, the feature space is of infinite dimension. However, for a relatively simple kernel function, such as the quadratic polynomial kernel used here, we can explicitly calculate the higher-order features and then compute FCS's for each one.

Based on FCS analysis, the most predictive single feature (Table 2) for all three 9 and 13 parameter training sets was delta Cn[23], the difference between the normalized cross-correlation parameters of the first and second ranked peptides. Xcorr, the raw correlation score of the top peptide sequence with the observed MS/MS spectrum, was the second most predictive single feature for all but two training sets. Threshold values of both of these parameters have been used previously to separate positive from negative peptides[5,25,27,28]. RSp, the ranking of the preliminary raw score, Sp, the preliminary score of the top peptide, and % ion match, the percent of predicted y and b ions for a given sequence that were matched in the experimental MS/MS spectrum, were also predictive. Two of the new parameters, fraction of matched MS/MS TIC and fraction of matched MS/MS peaks, were among the most highly predictive features, particularly for QTOF data. The least predictive features were delta mass, the difference between the observed and predicted masses for individual peptides, and the precursor ion current for individual peptides. The difference between observed and predicted precursor ion masses may not be predictive since this difference is already restricted when selecting peptides for SEQUEST analysis.

Some pairs of features are more informative than either feature alone. Combinations of individual features were also analyzed for their utility separating positive from negative peptides. Table 3 shows the results of a Fisher criterion score analysis of the different data sets using pairwise features. Only discriminant scores of 1.0 or above for at least one training set were included for illustration purposes. Compared to the analysis using single features, the analysis of pairs of features shows that correlations (or perhaps anti-correlations) among some pairs of features can be much more informative. The combination of fraction matched MS/MS TIC and delta Cn receives an FCS of 4.74, much higher than the scores assigned to either feature alone. The relatively high ranking of pairwise scores explains why the quadratic kernel function yields good SVM classification performance.

The most highly predictive combinations included the fraction of matched MS/MS ion current and the fraction of matched MS/MS peaks (7 combinations each). Other highly predictive combinations included delta Cn or Xcorr with other features. For each of these combinations the predictive value was higher with data acquired on the QTOF-1. This illustrates the ability of the SVM to learn the predictive value of combinations of features that might not be obvious a priori. The mass difference between the observed precursor ion mass and calculated mass of the best-fit peptide, which was poorly predictive when analyzed alone (Table 2), was also poorly predictive in combination with other parameters (data not shown). Thus not all parameters were highly predictive alone or in combination with other parameters. As a result of the utility of numerous pairwise feature combinations all combinations of features were included in the analysis. Individual variables that are highly predictive when analyzed in a pairwise fashion may be relatively independent variables.

The enhanced performance of the SVM with QTOF data compared to ion trap data appears to be due to better predictiveness of a number of parameters, including precursor ion charge measurement. This value was significantly more predictive for the separation of positives from negatives in QTOF data than for ion trap data (Table 2). Precursor charge was also highly predictive in pairwise feature analysis of QTOF data when combined with five other parameters (Table 3). One factor in this predictiveness might be the asymmetrical distribution of +1 charged precursors: 39 were included as part of the training set positive peptides, while 311 were included in the negative peptides. As discussed in the methods section, positive and negative peptides were not selected on this basis. Thus observation of a +1 precursor ion is significantly more likely for a negative than positive peptide. Other parameters, such as the MS/MS spectrum peak count, the fraction of matched MS/MS peaks, and fraction of matched MS/MS total ion current, were also significantly more predictive that for ion trap data, either alone (Table 2) or in combination with other parameters in pairwise scoring (Table 3). An enhanced signal-to-noise ratio for this data may also be valuable for the separation of positives and negatives.

One explanation for the difference in performance for the QTOF versus ion trap datasets might be the larger size of the QTOF training set. A subset of the QTOF data including 497 positive and 479 negative peptides, the same size as the smaller of the two ion trap datasets, was examined by the SVM using 13 parameters, and the ROC score computed. The test results contained 20 false positives and 23 false negatives, and a ROC score of 0.989. This compares well with the ROC score for analysis of the full sized QTOF dataset using 13 parameters (0.988). This suggests that the quality of the QTOF data, rather than the larger number of examples in the dataset, explains the improved performance compared to the ion trap-based results.

The SVM provides better performance than other techniques.

Comparison of SVM results with previous analyses of SEQUEST results based on thresholds. The results of the SVM analysis of the above training sets can be compared with approximations of previous methods, employing different cutoffs for delta Cn and/or Xcorr, used to evaluate SEQUEST-generated matches between peptide data and database sequences (Table 4). One simple method, used before protein sequence databases became large, involved selection of peptides as positives with delta Cn values larger than 0.1[23,25]. Using a criterion of minimizing false positives (defined here as negative peptides missed using the defined cutoffs) and false negatives (defined as positive peptides missed), this was the best performing cutoff of the 3 sets of cutoffs examined. A second method[27] included selection, as positives, of +1 peptides with Xcorr values larger than 1.5, selection of +2 and +3 peptides with Xcorr values larger than 2, and several other criteria including manual examination. Use of these cutoffs alone resulted in intermediate performance among the 3 sets of cutoffs. A more stringent method[5,28] included retention of tryptic peptides with Xcorr values above 1.9, 2.2 and 3.75 for +1, +2 and +3 peptides, a delta Cn of 0.1 or greater, and tryptic ends, followed by manual confirmation of the sequence match to the MS/MS spectrum under some circumstances. The cutoffs from this method resulted in the highest sum of false positives and false negatives for the 3 methods considered, although it gave lower levels of false positives than some of the other sets of cutoffs. The SVM results using both 9 and 13 parameters gave a significantly lower sum of false positives and false negatives than these sets of cutoffs.

Comparison of SVM results with Qscore results. Training set peptides analyzed using SVM calculations were also analyzed using the Qscore algorithm[29]. Qscore is a program that evaluates the quality of protein identifications from SEQUEST results using probabilistic scoring. The program requires at least two peptides for a protein identification, thus for comparison purposes with individual peptides contained in the nonredundant human database-ion trap training set, we modified the Qscore program to allow the display of calculated scores for single peptides. Qscore is not a binary classifier, thus true and false positives and negatives were not calculated. In Figure 3, the ROC curve for Qscore analysis of the ion trap-nonredundant human dataset is compared with ROC curves generated using SVM calculations. For both the 9 and 13 parameter SVM results, the ROC curves are shifted to the upper left, indicating that for a fixed percent of false positives, there are significantly more true positive peptides from the SVM analysis. While Qscore does not attempt to identify proteins with fewer than two peptides, these results suggest that a similar use of SVM peptides, combined with careful examination for mistakes of outliers from initial SVM analysis of SEQUEST data, might provide higher quality protein identifications.

Keller et al. [39] used an expectation maximization algorithm, incorporating for analysis four SEQUEST scores and the number of tryptic peptide termini present in the matched peptides. Plus 2 and 3 ions were analyzed separately for ion trap peptide data; Xcorr', delta Cn, and ln RSp contributed to most of the discrimination between positive and negative peptides. Our data includes more parameters, and +1, +2, and +3 ions are included in one analysis. For our training sets, we find that more parameters significantly contribute to the discrimination between positive and negative peptides, including delta Cn, Xcorr, Sp, % ion match, RSp, the fraction of matched MS/MS peaks and total ion current, which vary for different training sets.

Comments on results. The support vector machine is a binary classifier, and is thus useful for making decisions about membership of analyzed entities in either of two classes. Here we have defined the two classes as peptides correctly or incorrectly matching SEQUEST-assigned sequences. Additional applications using mass spectrometry data might include binary decisions between classes such as presence and absence of an early stage disease such as cancer[41]. Similar decision making could be applied to de novo sequenced peptides if there was sufficient information describing the fit of a de novo sequence to a peptide, and if the problem was constructed as to whether or not the de novo sequence was correct. This would likely involve other algorithms than SEQUEST, which relies mainly on pattern matching between predicted and observed MS/MS spectra.

Based on our experience and on the training set data examined, there are several categories of incorrectly predicted peptides. First, we initially encountered false positives based on the SEQUEST selection of peptides, matched to a given precursor ion and its MS/MS spectrum, that were not the top ranked peptides. These, and incorrectly labeled peptides, were removed after manual examination of results from initial rounds of SVM analysis. Second, analysis of some of the tryptic maps of individual “pure” proteins indicated that there were other proteins present with more than one high-scoring peptide. Examples of negative peptides were not taken from these samples. They were instead substituted with samples of at least 97% protein purity, which were limited to injections of no more than 100 fmol of peptides. The presumed levels of impurity should thus be below the routine limit of detection for our ion trap or QTOF mass spectrometers (ca. 10 fmol) as currently configured. Nonetheless it is possible that some of the proteins assigned as negatives might represent impurities present in the sample.

We have not been able to completely separate positives from negatives in any of the training sets examined, for data acquired on either mass spectrometer. Some of the reasons discussed below may help explain this observation. First, the training sets included the lowest scoring available positive peptides, which were often among multiple peptides correctly identifying a known protein. A number of false positive sequences with high SEQUEST scores, for example peptides selected as second choices for known positive peptides, were also included. Similar examples have been reported when using reversed-sequence databases as controls[29]. For the ion trap-non redundant human database-searched training set, there were 124 positive peptides with delta Cn values below 0.1. For +1, +2 and +3 ions there were 4, 33, and 75 additional peptides that did not meet the most stringent Xcorr cutoffs (method 3) in Table 4. There were 108 negative peptides with delta Cn values of 0.1 or above, and 14, 74, and 0 additional +1, +2 and +3 *negative* peptides with Xcorr values *above* those used for cutoffs in method 3 of Table 4. These were thus challenging training sets.

Second, a number of false positives and negatives were assigned to peptides with noisy MS/MS spectra, or with poor fragmentation in these spectra. In both cases the information content necessary for correct sequencing will be compromised, and it is expected that accurate sequence assignments will be difficult. Of the 22 poorly fragmenting positive peptides incorrectly assigned as negatives, all but one contained an internal residue (pro, his, arg) thought to cause uneven peptide fragmentation[42,43], and 14 contained more than one of these internal residues. It is not clear that even manual examination of these peptide MS/MS spectra will lead to a correct sequence match. A tentative identification of proteins based on these questionable peptides will require additional experiments, or additional matching peptides of higher quality, for verification. A computational indication of ambiguously identified peptides, indicated by the computed distance from the 9-parameter or 13-parameter hyperplane, should select any peptide so positioned for further scrutiny.

More generally, incorrect sequence assignments may also occur if the correct sequence is not in the database examined. For human protein sequences 80% of novel gene predictions from drafts of the Ensembl and Celera datasets occur in only one of these datasets[44], thus an accurate and complete human protein sequence database is not yet available. Other incorrect assignments may be due to modifications to individual amino acids not incorporated into the sequences searched, or to incorrect assignment of the precursor ion charge when a lower mass accuracy instrument is used and the ratio of MS to MS/MS scans is low. The best resulting sequence will then be incorrect.

Conclusions

Using appropriate training sets, our approach allows an automated computational first-pass analysis of SEQUEST data on individual peptides. This should allow a higher throughput analysis of shotgun peptide sequencing results. For tandem mass spectrometry data, SVM analysis of experimentally obtained parameters, SEQUEST-calculated statistics, and additional parameters allows a better match between this data and peptide sequences than previous methods, using our training sets. The use of four new parameters tested here contributed significantly to the separation of positive and negative peptides. A good but not complete separation between positive and negative peptides was obtained for ion trap data using two different databases. A significantly better separation was obtained for uncalibrated QTOF MS/MS data. Using SVM calculations, the contributions of the parameters to the separation were individually examined. The parameters delta Cn, Xcorr, Sp, the fraction of the MS/MS spectrum ion current matched by peptide fragments, and the fraction of the total number of MS/MS spectrum peaks matched by peptide fragments contributed significantly to the separation of positive and negative peptides. Each training set is customized to the mass spectrometer used to collect data and the database examined. Protein identifications from these peptides will then be based on the number of individual peptides identifying a particular protein, and the distance of each peptide from the hyperplane separating positives and negatives in the appropriate training set. The reproducibility and uniqueness of the identification will also be important[38] for correct protein identifications. Manual examination of spectra of peptides with poor or ambiguous SVM-calculated scores should identify noisy or poorly-fragmenting spectra that may compromise peptide identification.

Table 1. Analysis of training sets using different methods [a].

Method:		**SVM-9 analysis**		**SVM-13 analysis**	
Training set:	positive, negative peptides	false positives, negatives	ROC scores	false positives, negatives	ROC scores
ion trap, NR human	696, 465	48, 117	0.929	44, 100	0.950
ion trap, full NR	497, 479	62, 81	0.920	53, 70	0.939
QTOF, NR human	1017, 532	27, 81	0.981	18, 51	0.988

[a] The training sets used either the nonredundant human database (NR human) or the full nonredundant database.

Table 2. Contribution of single features to the separation of positive and negative peptides as reflected by their Fisher criterion scores

mass spectrometer:	ion trap	ion trap	QTOF-1
database:	NR human	NR full	NR human
features:	**9 or 13 parameters**		
delta Cn	1.401	1.018	2.861
Xcorr	0.935	0.477	2.444
Sp	0.714	0.604	1.158
MH	0.000	0.000	0.704
charge	0.118	0.102	0.488
RSp	0.273	0.447	0.313
% ion match	0.607	0.447	0.079
dM	0.000	0.014	0.024
TIC	0.016	0.011	0.008
	13 parameters		
fraction matched MSMS TIC	0.632	0.422	2.804
fraction matched MSMS peaks	0.335	0.260	2.314
peak count	0.062	0.018	0.209
seq similarity	0.060	0.130	0.115

Table 3. Pairwise contributions of individual feature Fisher scores to the separation of positive and negative peptides.

mass spectrometer:		ion trap	ion trap	QTOF-1
database:		NR human	NR full	NR human
feature 1	**feature 2**	**9 or 13 parameters**		
delta Cn	Xcorr	1.51	1.15	3.60
	charge	0.980	0.809	3.56
	MH	1.05	0.877	3.12
	% ion match	1.76	1.43	2.81
	SP	1.43	1.18	2.80
Xcorr	charge	0.208	0.068	1.94
	MH	0.366	0.162	1.89
	Sp	0.956	0.698	1.88
	%ion match	1.37	0.846	1.83
Sp	MH	0.743	0.593	1.92
	charge	0.502	0.402	1.77
%ion match	MH	1.53	1.13	2.09
	charge	0.402	0.322	1.25
	Sp	0.959	0.775	1.12
		13 parameters		
fraction matched MSMS peaks	delta Cn	1.48	1.21	4.23
	Xcorr	1.08	0.661	3.38
	Sp	0.998	0.811	2.38
	%ion match	0.998	0.811	2.38
	MH	0.114	0.087	1.67
	charge	0.014	0.007	1.53
	peak count	0.328	0.148	1.18
fraction matched MSMS TIC	delta Cn	1.68	1.34	4.74
	Xcorr	1.33	0.843	3.82
	fraction matched MSMS peaks	0.556	0.394	2.82
	Sp	1.17	0.931	2.58
	MH	0.327	0.229	2.10
	charge	0.108	0.057	1.90
	% ion match	1.08	0.700	1.47
peak count	delta Cn	1.01	0.861	1.42

Table 4. Analysis of training sets using different methods.

Method: [a]	**1**	**2**	**3**	**SVM-9**	**SVM-13**
Training Set:	--------------------**false positives, false negatives**-------------------				
ion trap, NR human	115, 142	133, 187	132, 369	48, 117	44, 100
ion trap, full NR	87, 142	305, 55	180, 251	62, 81	53, 70
QTOF, NR human	108, 81	126, 86	57, 285	27, 81	18, 51

[a] The cutoffs used for these comparative analyses are taken from Eng. et al[23] and Yates et al. [25] for method 1, from Link et al. [27] for method 2, and from Washburn et al.[5] and Gygi et al. [28] for method 3; the SVM analyses used both 9 and 13 parameters. False positives and negatives for methods 1-3 were calculated as the number of negative and positive peptides missed by the cutoffs, respectively.

Figure legends.

Figure 1. A support vector machine learns to recognize high-quality peptide matches. The figure illustrates how an SVM learns to discriminate between true and false peptide matches (listed as positives and negatives). **Peptide data is obtained from LC/MS/MS experiments analyzed by SEQUEST.** A training set consists of a collection of individual peptide matches, each characterized by a vector of statistics (as described in the text) and a binary classification (true or false match) provided by manual inspection of the training set. The SVM learning algorithm finds a decision boundary that separates the true matches from the false matches. This decision boundary can then be used by the SVM prediction algorithm to determine the classification of previously unseen peptides. The prediction produced by the SVM is a binary classification, along with a discriminant value that can be used to estimate the SVM's confidence in its prediction. **Analysis of training sets using single or pairwise feature analysis can indicate which individual or pairwise features contribute the most to separation of positive and negative peptides in 9- or 13-feature space. Comparison of training sets obtained using different mass spectrometers or databases estimates the contribution of these variables to the separation of positive and negative peptides, and thus to accurate peptide and protein identification.**

Figure 2. ROC plots of three different training sets used in SVM calculations. A. ROC plot of training sets containing 9 parameters. The normalized true positives are plotted against the normalized false positives for each training set. The QTOF-nonredundant human database set is represented in open black squares, the ion trap-nonredundant human database set in light gray, and the ion trap-nonredundant human database set in darker

gray. The QTOF training set has the fewest false positives relative to true positives of any set; the ion trap-full nonredundant database set, which has about 8 times as many entries as the ion trap-nonredundant human set, has the most false positives relative to true positives of any set. Thus the SVM has the most success separating true from false positives with the QTOF dataset, and less success with ion trap data using either the full nonredundant or nonredundant human databases. **B.** ROC plot of training sets containing 13 parameters. The QTOF-nonredundant human database set (open black squares) has the fewest false positives relative to true positives, the ion trap-nonredundant human database (light gray) is intermediate in this respect, and the ion trap-full nonredundant database (darker gray) has the most false positives relative to true positives of any set. Again the SVM has the most success separating true from false positives with the QTOF dataset.

Figure 3. Comparison of Qscore with SVM analyses of a peptide training set. A ROC plot was used to compare SVM and Qscore analysis of an ion-trap nonredundant human database training set using either 9 or 13 parameters. Qscore was modified to calculate values for single peptides rather than requiring two peptides for an analysis, and these scores were used for the comparison. Both SVM analyses gave a higher number of true positives for a fixed number of false positives than the modified Qscore analysis.

References.

1. Hunt ,D.; Michel, H.; Dickinson, T.; Shabanowitz, J.; Cox, A.; Sakaguchi, K.; Appella, E.; Grey, H.; Sette, A. *Science* **1992,** *256*,1817-20

2. Cox, A.; Skipper, J; Chen, Y.; Henderson, R.; Darrow, T.; Shabanowitz, J; Engelhard, V.; Hunt, D.; Slingluff, Jr. C. *Science,* **1994,** *264*, 716-719.

3. Neubauer, G.; Gottschalk, A.; Fabrizio, P.; Seraphin, B.; Luhrmann, R.; Mann M. *Proc Natl Acad Sci USA* **1997,** *94*, 385-90.

4. Rout, M.; Aitchison, J.; Suprapto, A.; Hjertaas, K.; Zhao, Y.; Chait B. *J Cell Biol* **2000,** *148*, 635-51.

5. Washburn, M.; Wolters, D.; Yates III, J. *Nature Biotechnology* **2002,** *19*, 242-247.

6. Gavin, A.; Bosche, M; Krause, R; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J.; Michon, A.; Cruciat, C.; Remor, M.; Hofert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M.; Copley, R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B.; Neubauer, G.; Superti-Furga, G. *Nature* ***2002,*** *415*, 141-7.

7. Ho, Y; Gruhler, A; Heilbut, A; Bader, G; Moore, L; Adams, S; Millar, A.; Taylor, P.; Bennett, K.; Boutilier, K.; Yang, L.; Wolting, C.; Donaldson, I.; Schandorff, S.; Shewnarane, J.; Vo, M.; Taggart, J.; Goudreault, M.; Muskat, B.; Alfarano, C.; Dewar, D.; Lin, Z.; Michalickova, K.; Willems, A.; Sassi, H.; Nielsen, P.; Rasmussen, K.; Andersen, J.; Johansen, L.; Hansen, L.; Jespersen, H.; Podtelejnikov, A.; Nielsen, E.; Crawford, J.; Poulsen, V.; Sorensen, B.; Matthiesen, J.; Hendrickson,R.; Gleeson, F.; Pawson, T.; Moran, M.; Durocher, D.; Mann, M.; Hogue, C.; Figeys, D.; Tyers, M. *Nature* **2002** *415*,180-3.

8. Spahr, S.; Davis, M.; McGinley, M.; Robinson, R.; Bures, E.; Beierle, J.; Mort, J.; Courchesne, P.; Chen, K.; Wahl, R.; Yu, W.,; Luethy, R.; Patterson, S. *Proteomics* **2001,** *1*, 93-107.

9. Andersen, J.; Lyon, C., Fox, A.; Leung, A.; Lam, Y.; Steen, H.; Mann, M.; Lamond, A. *Current Biol.* **2002,** *12*, 1-11.

10. Hunt, D.; Bone, W.; Shabanowitz, J.; Rhodes, J.; Ballard, J. *Anal. Chem.* **1981** *53*, 1704-1706.

11. Hunt D.; Buko A.; Ballard, J.; Shabanowitz, J.; Giordani A. *Biomed Mass Spectrom* **1981,** *8*, 397-408.

12. Hunt D.; Shabanowitz ,J.; Yates, J.; Melver, R.; Hunter, R.; Syka, J.; Amy, J. *Anal Chem* **1985,** *57*, 2728-33.

13. Hunt, D.; Yates, J.; Shabanowitz, J.; Winston, S.; Hauer, C. *Proc Natl Acad Sci U S A* **1986,** *83*, 6233-7.

14. Hunt, D.; Zhu, N.; Shabanowitz, J. *Rapid Commun Mass Spectrom* **1989,** *3,* 122-4.

15. Biemann, K. *Biomed. Environ. Mass Spectrom.* **1988,** *16*, 99.

16. Henzel, W.; Billeci, T.; Stults, J.; Wong, S.; Grimley, C.; Watanabe, C. *Proc Natl Acad Sci USA* **1993,** *90*, 5011-5.

17. Mann, M.; Wilm, M. *Anal Chem* **1994,** *66*, 4390-9

18. Shevchenko, A.; Jensen, O.; Podtelejnikov, A.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann M. *Proc Natl Acad Sci U S A* **1996,** *10*, 14440-5.

19. Qin, J.; Fenyo, D.; Zhao, Y.; Hall, W.; Chao, D.; Wilson, C.; Young, R.; Chait, B. *Anal Chem* **1997,***69,* 3995-4001.

20. Zhang, W.; Chait, B. *Anal Chem* **2000,** *72*, 2482-9.

21. Clauser, K.; Baker P.; Burlingame, A.. *Analytical Chemistry* **1999,** *71*, 2871-82.

22. Perkins, D.; Pappin, D.; Creasy, D.; Cottrell, J. *Electrophoresis* **1999,** *20,* 3551-67.

23. Eng, J.; McCormack, A.; Yates III, J. *J. Am. Soc. Mass Spectrom.* **1994,** *5*, 976-989.

24. Yates, J.; Eng, J.; McCormack, A.; Schieltz, D. *Anal Chem* **1995,** *15*, 1426-36.

25. Yates, J.; Eng J.; McCormack, A. *Anal Chem* **1995,** *15*, 3202-10.

26. McCormack, A.; Schieltz, D.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. *Anal Chem* **1997,** *69*, 767-76.

27. Link, A.; Eng, J.; Schieltz, D.; Carmack, E.; Mize, G.; Morris, D.; Garvik, B.; Yates, J. (1999). *Nat Biotechnol* **1999,** *17,* 676-82.

28. Gygi, S.; Rist, B.; Griffin, T.; Eng, J.; Aebersold, R. *J. Proteome Research* ***2002***, .

29. Moore, R.; Young, M.; Lee, T. *J. Am. Soc Mass Spectrom.* **2002,***13*, 378-386.

30. Boser, B.; Guyon, I.; Vapnik , V. In Haussler, D., *5th Annual ACM Workshop on COLT:* Pittsburgh, 1992; pp. 144-152.

31. Vapnik, V. *Statistical Learning Theory*. John Wiley and Sons: New York, 1998.

32. Cristianini, N.; Shawe-Taylor, J. *An introduction to support vector machines;* Cambridge University Press, Cambridge, 2000.

33. Jaakkola, T.; Diekhans, M.; Haussler, D. *Proc Int Conf Intell Syst Mol Bio*l, ***1999,*** 149-58.

34. Ding, C.; Dubchak, L. *Bioinformatics* **2001,***17*, 349-58.

35. Brown, M.; Grundy, W.; Lin, D.; Cristianini, N; Sugnet, C.; Furey, T.; Ares Jr., M.; Haussler, D. *Proc. Nat. Acad. Sci.* **2000,** *97*, 262-267.

36. Furey T.; Cristianini, N; Duffy, N.; Bednarski D.; Schummer, M; Haussler, D. *Bioinformatics 2000 16*, 906-14.

37. Rosenblatt, F. *Psychol. Rev.* **1959,** *65*, 386-408.

38. Gururaja, T.; Li, W.; Bernstein, J.; Payan, D.; Anderson, D. *Journal of Proteome Research* **2002,** *1*, 253-261.

39. Keller, A.; Nesvizhskii, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002,** 74, 5383-5392.

40. Duda, R.; Hart, P. *Pattern Classification and Scene Analysis.* New York: John Wiley and Sons, 1973.

41. Petricoin III , E.; Ardekani, A.; Hitt, B. et al. *Lancet* **2002,** 359, 572-577.

42. Pappayanopoulos, I. *Mass Spectrom. Rev.* **1995,** *14*, 49.

43. Willard, B.; Kinter, M. *J Am Soc Mass Spectrom* **2001,** *12*, 1262-1271.

44. Hogenesch, J.; Ching, K.; Batalov, S.; Su, A.; Walker, J.; Zhou, Y.; Kay, S.; Schultz, P.; Cooke, M. *Cell* **2001,***106*, 413-415.

Fig. 1

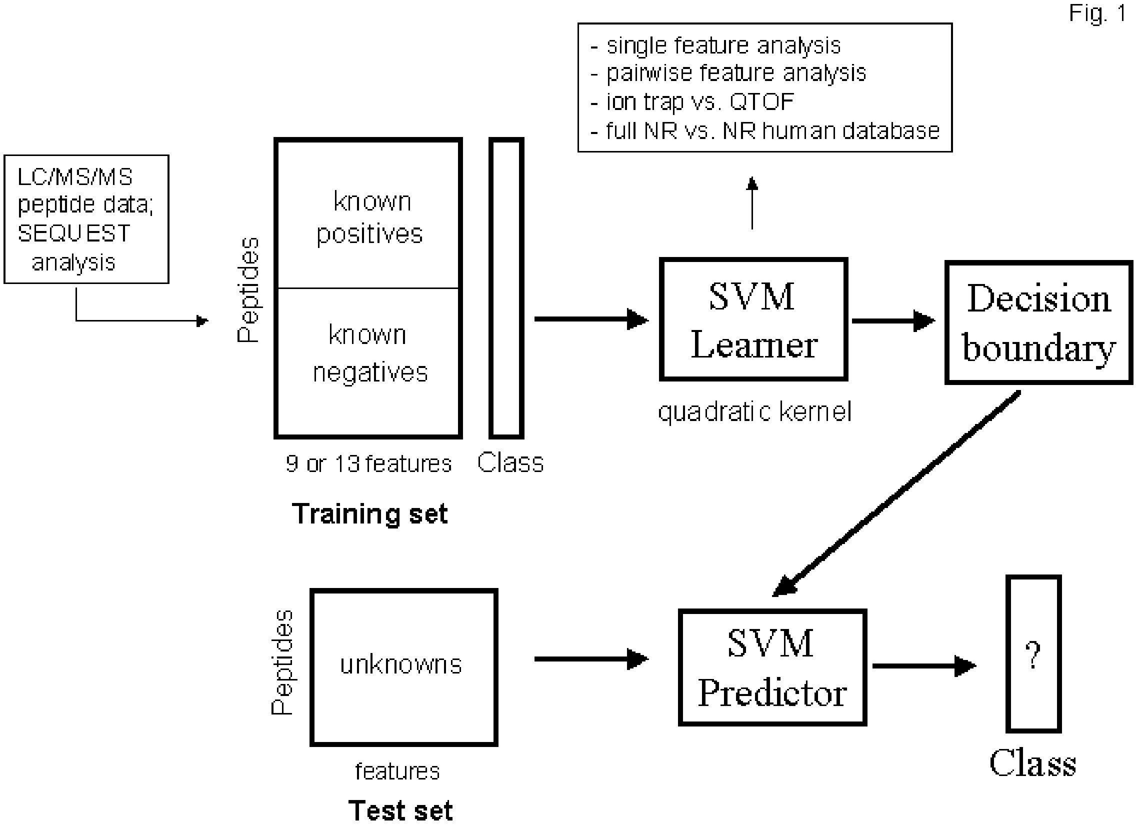

- single feature analysis
- pairwise feature analysis
- ion trap vs. QTOF
- full NR vs. NR human database
LC/MS/MS peptide data; SEQUEST analysis
Peptides
known positives
known negatives
9 or 13 features
Class
Training set
SVM Learner
quadratic kernel
Decision boundary
Peptides
unknowns
features
Test set
SVM Predictor
?
Class

Fig. 2A

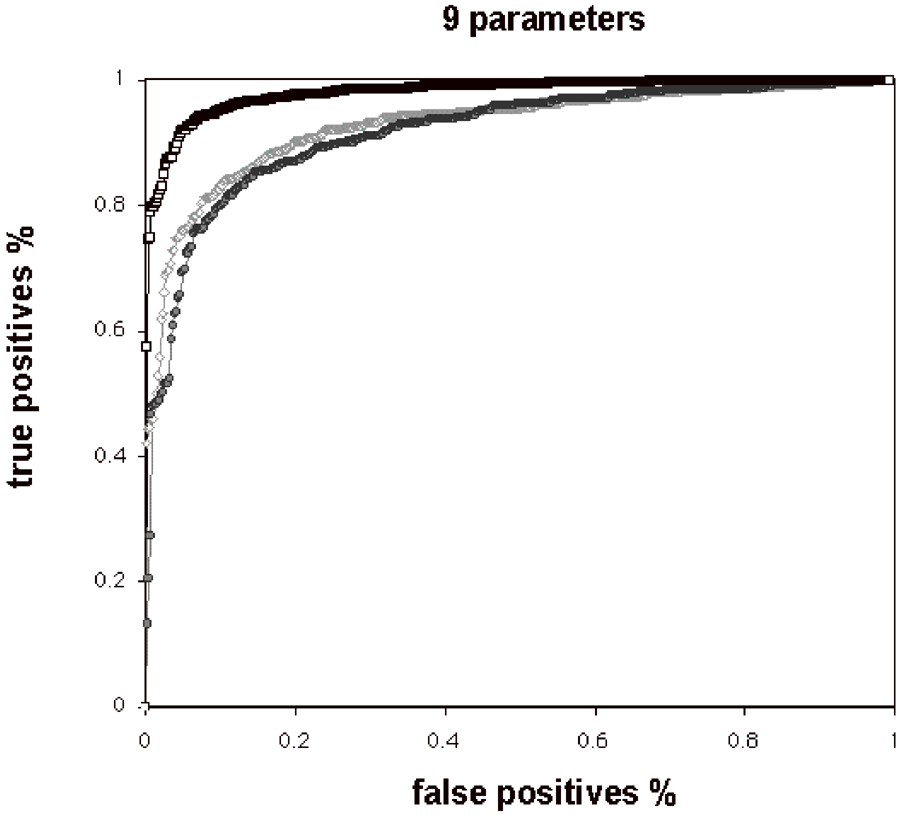

9 parameters
true positives %
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
false positives %

Fig. 2B

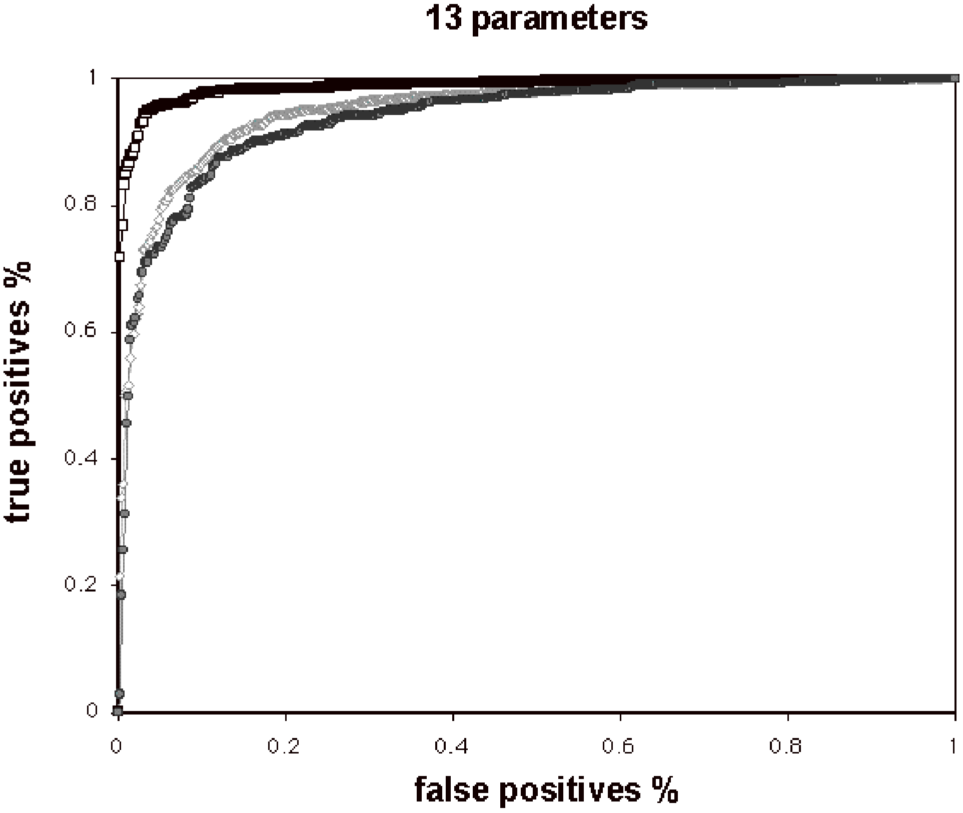
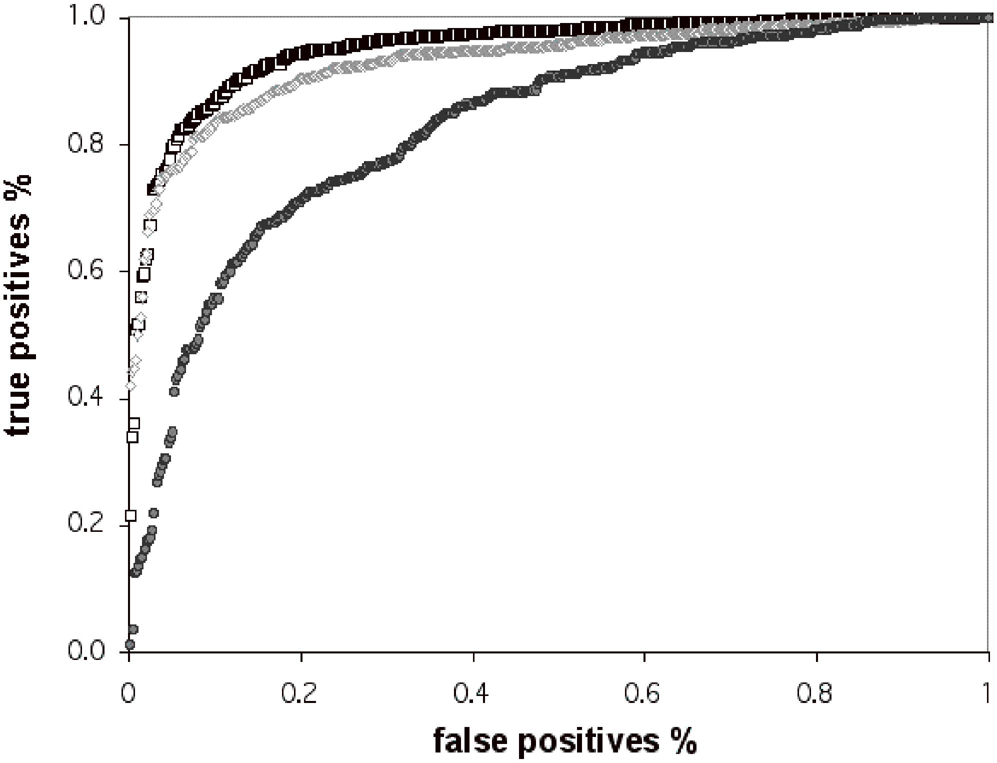

13 parameters
true positives %
false positives %
0
0.2
0.4
0.6
0.8
1
0
0.2
0.4
0.6
0.8
1

Fig. 3


true positives %
false positives %