A Systematic Statistical Analysis of Ion Trap Tandem Mass Spectra in View of Peptide Scoring

Jacques Colinge, Alexandre Masselot, and Jérôme Magnin

GeneProt Inc., Pré de la Fontaine 2, CH-1219 Meyrin, Switzerland
Jacques.Colinge@geneprot.com

Abstract. Tandem mass spectrometry has become central in proteomics projects. In particular, it is of prime importance to design sensitive and selective score functions to reliably identify peptides in databases. By using a huge collection of 140 000+ peptide MS/MS spectra, we systematically study the importance of many characteristics of a match (peptide sequence/spectrum) to include in a score function. Besides classical match characteristics, we investigate the value of new characteristics such as amino acid dependence and consecutive fragment matches. We finally select a combination of promising characteristics and show that the corresponding score function achieves very low false positive rates while being very sensitive, thereby enabling highly automated peptide identification in large proteomics projects. We compare our results to widely used protein identification systems and show a significant reduction in false positives.

1 Introduction

Tandem mass spectrometry (MS/MS) combined with database searching has become central in proteomics projects. Such projects aim at discovering all or part of the proteins present in a certain biological tissue, e.g. tears or plasma. Before MS/MS can be applied, the complexity of the initial sample is reduced by protein separation techniques like 2D-page or liquid chromatography (LC). The proteins of the resulting simpler samples are digested by an enzyme that cleaves the proteins at specific locations. Trypsin is frequently used for this purpose. MS/MS analysis is performed on the digestion products, which are named peptides. Alternatively, early digestion can be applied and peptide separation techniques used. In both cases, the peptides are positively ionized and fragmented individually [18] and, finally, their masses as well as the masses of their fragments are measured. Such masses constitute a data set, the experimental MS/MS spectrum, that is specific to each peptide. The MS/MS spectra can be used to identify the peptides by searching into a database of peptide sequences. By extension, this procedure allows to identify the proteins [9,16].

MS/MS database searching usually involves the comparison of the experimental MS/MS spectrum with theoretical MS/MS spectra, computed from the

G. Benson and R. Page (Eds.): WABI 2003, LNBI 2812, pp. 25–38, 2003.

peptide sequences found in the database. A (peptide) score function or scoring scheme is used to rate the matching between theoretical and experimental spectra. The database peptide with the highest score is usually considered as the correct identification, provided the score is high enough and/or significant enough.

Clearly, the availability of sensitive and selective score functions is essential to implement reliable and automatic MS/MS protein identification systems. In [3] we proposed a generic approach (OLAV) to design such score functions. This approach is based on standard signal detection techniques [22]. In this paper we apply it to LC electrospray ionization ion trap (LC-ESI-IT) mass spectra and we study the relative interest of various quantities we can compute when we compare theoretical and experimental spectra. We finally select a combination of such quantities and establish the performance of the corresponding score function by performing large-scale computations. For reference purposes, we give results obtained with Mascot [20], a widely used commercial protein identification program (available from Matrix Sciences Ltd), and we report the performance we obtain on a generally available data set [13] for which Sequest [6,28] (available from ThermoFinnigan) results have been published [13,12].

Currently available protein identification systems can be classified into three categories: heuristic systems, systems based on a mathematical model and hybrid systems. In the heuristic category there are well known commercial programs: Mascot, Sequest and SONAR MS/MS [7]. Sequest and SONAR correlate theoretical and experimental spectra directly, without involving a model. Mascot includes a limited model [19] as well as several heuristics intended to capture some properties related to signal intensity and consecutive fragment matches. Model-based systems use stochastic models to assess the reliability of matches. In this category we find: MassLynx (available from Micromass Limited [25]), SCOPE [2], ProbId [29] and SHERENGA [4]. SCOPE considers fragment matches as independent events and estimates a likelihood by assuming a Gaussian distribution of mass errors. MassLynx uses a Markov chain to estimate the correct match likelihood and to model consecutive fragment matches. ProbId uses Bayesian techniques to estimate the probability a match is correct. It integrates several elementary observations like peak intensities and simultaneous detection of fragments in several series. SHERENGA estimates a likelihood ratio by considering every fragment match as an independent Bernoulli random variable. [8] improves over SHERENGA by considering signal intensity and neutral losses. The hybrid category generally uses multivariate analysis techniques to filter the results returned by heuristic systems [1,12,14,17,23].

The knowledge of which are the essential quantities to include in a score function is certainly beneficial to most of the approaches above.

According to the relative performance of the various score functions we tested, the most important quantity to include in a score function is the probability to detect each ion type. The next quantity is the intensity of detected fragment: intense fragment must match with probabilities depending on the ion type. Then, different extra quantities improve performance: probability to detect a fragment depending on its amino acid composition, probability to observe consecutive fragment matches. By combining these quantities in a naive Bayesian classifier, we design a score function that has a false positive rate as low as 3% is the true positive rate is fixed at 95%. On data set [13], the false positive rate is inferior to 0.5%.

It is difficult to compare peptide score functions without testing them on the same data set. As a matter of fact, MS/MS data are noisy and of variable precision. Hence, the absolute performance of a given score function may change depending on data set quality. According to our experience, the relative advantage of a score function compared to another one is generally stable from data sets to data sets. To allow readers to compare our results with their own experience, we report them by using an available data set or with a standard algorithm tested on the same set. We observe a strong advantage in favor of the approach we propose.

2 Mass Spectrometry Concepts

2.1 ESI Ion Trap Instruments

Current peptide ionization methods that are common in proteomics include electrospray ionization (ESI) and matrix assisted laser desorption ionization (MALDI) [10]. Several technologies are also available for selecting and fragmenting the accelerated peptides, one of which is quadrupole ion trap (IT) [11,26]. IT represents a significant and growing portion of the mass spectrometers used in proteomics. The approach presented in [3] is not specific to ESI-IT mass spectra.

ESI produces positively charged ions, whose charge states are mainly two or three. An IT instrument breaks peptides by low-energy collision-induced dissociation (CID) [26]. The fragmentation process yields several ion types (a, b, y) [18], depending on the exact cleavage location. The proportion of each ion type produced changes with the MS/MS technology. Additionally, certain amino acids can loose one water (H_2O) or ammonia (NH_3) molecule. Consequently, fragment ion masses can be shifted by -18 Da and/or -17 Da.

Every mass spectrometry instrument produces a signal that can be assumed to be continuous. Peak detection software is used for extracting peptide or fragment masses from this signal. The list of extracted masses is named a peak list. When we refer to a spectrum, we always refer to the corresponding peak list, which is the primary data for identification.

2.2 Matching Theoretical and Experimental Spectra

We do not describe here how to compute theoretical spectra [26]. It is sufficient to know that, given an amino acid sequence, there exit precise rules to compute the mass of every possible fragment of each ion type. The theoretical spectrum consists of the masses of fragments for a selected set S of ion types. S is instrument technology dependent. S also depends on the peptide charge state.

We name the comparison of an experimental spectrum with a theoretical spectrum a *match*. A match can be either correct or random. From the match we compute several quantities that are then used by the score function. These quantities are modeled as random variables. It is convenient to represent them by a random vector E.

The score function is intended to distinguish between correct and random matches. This problem can be viewed as an hypothesis testing problem. In [3] we propose to build score functions as log-likelihood ratios as this approach yields optimal decision rules [22], provided E probability distributions are known in the correct (H_1) and random (H_0) cases. In practice, we have to approximate these two distributions. Nevertheless, we believe that log-likelihood ratios are very effective for peptide scoring, which is confirmed by the high performance we achieve, see Section 4. In [3] we give other arguments to justify this point of view.

3 Statistical Modeling

3.1 Data Set

We analyzed by proteomics two pools of 2.5 liters of plasma. One control pool and one diseased pool (coronary artery disease), each containing roughly 50 selected patients. Multidimensional liquid chromatography was applied, yielding roughly 13 000 fractions per pool, which were digested by trypsin and analyzed by mass spectrometry (LC-ESI-IT) using 40 Bruker Esquire 3000 instruments.

The set of ion trap mass spectra we use is made of 146 808 correct matches, 33 000 of which have been manually validated. The other matches have been automatically validated by a procedure, which, in addition to fixed thresholds, includes biological knowledge and statistics about the peptides that were validated manually. There are 3329 singly charged peptides (436 distinct), 82 415 doubly charged peptides (3039 distinct) and 61 064 triply charged peptides (2920 distinct).

Every performance reported in this paper is obtained by randomly selecting independent training and test sets, which sizes are 3000 and 5000 matches respectively (1000/2329 for charge 1). This procedure is repeated 5 times and the results averaged. We also checked that both model parameters and performance barely change from set to set.

In order to validate one important hypothesis at Section 3.6, we use another set of 1874 doubly charged peptides (73 distinct) and 4107 triply charged peptides (90 distinct). The spectra were acquired on 4 Bruker Esquire 3000+ instruments. We refer to this data set as data set B. Data set [13] has been generated by a ThermoFinnigan LCQ ion trap instrument.

3.2 Theoretical Spectrum and Neutral Losses

As we mentioned in Section 2.1, certain amino acids may lose water or ammonia (a so-called neutral loss). By considering the chemical structure of amino

Table 1. Neutral loss statistics. Relative abundance in Cys_CAM, Asn, Gln, Arg, Ser and Thr between b-17 and b, b-18 and b, etc. Other amino acids are not enriched significantly. Mean and standard deviation are computed from all amino acid enrichments. *Cys_CAM.

Ions	CysC*	Asn	Gln	Arg	Mean	std dev	Ions	Ser	Thr	Mean	std dev
Singly charged peptides											
a-17	1.2	1.7	1.3	0.7	1.00	0.25	a-18	1.2	1.4	0.98	0.24
b-17	1.2	2.1	1.5	1.9	1.08	0.37	b-18	1.2	1.3	0.94	0.19
y-17	1.0	1.9	1.2	2.3	1.13	0.42	y-18	1.2	1.0	1.04	0.14
Doubly charged peptides											
a-17	1.0	1.9	0.9	0.4	1.02	0.28	a-18	1.2	1.2	0.99	0.21
b-17	1.2	1.6	1.2	0.9	1.00	0.18	b-18	1.2	1.2	0.93	0.19
y-17	1.2	1.5	1.5	0.9	1.04	0.19	y-18	1.1	1.2	1.00	0.13
Triply charged peptides											
b-17	1.0	1.4	0.9	0.9	1.00	0.20	b-18	1.4	1.1	0.95	0.20
y-17	0.8	1.5	1.3	0.8	0.99	0.23	y-18	1.1	1.6	0.94	0.28
b^{++}-17	0.9	1.0	1.0	1.0	0.99	0.05	b^{++}-18	1.1	1.1	0.98	0.06
y^{++}-17	1.0	1.0	1.0	0.8	0.99	0.11	y^{++}-18	1.0	1.2	0.99	0.13

acids [26], we observe that Arg (R), Asn (N) and Gln (Q) may loose ammonia, and Ser (S) and Thr (T) may loose water. In order to break disulfur bonds, Cys (C) are modified. A common modification is S-carboxamidomethyl cysteines (Cys_CAM, +57 Da) whose chemical structure [26] suggests a potential loss of ammonia. In the data sets we use (except [13]), Cys are modified as Cys_CAMs.

To be able to compute realistic theoretical spectra, we check which of the amino acids above loose water or ammonia significantly. This point has been already considered by [27] for doubly charged peptides and based on a much smaller data set. We follow a similar approach, i.e. we assume that every amino acid may loose water or ammonia and we compute the amino acid composition of matched ions a, b, y and a,b,y-17,18. Finally, the amino acid compositions are compared to find significant enrichments. The results are shown in Table 1 and we decide to exclude Cys_CAM. Asn is kept although it is only significant for singly charged peptides. We checked that including it provides a small benefit for singly charged peptides without penalizing higher charge states (data not shown).

3.3 Comparing Peptide Score Functions

To supplement the peptide scores with p-values, we have introduced in [3] a method for generating random matches and thus random scores. The general principle is the following: given a match, we generate a fixed number of random peptide sequences, with appropriate masses and possible PTMs, by using a Markov chain. The random peptides are matched with the original spectrum to estimate the random score distribution. Now, to compare the relative performance of various score functions, we compute, for each correct match, the ratio of the correct match score with the best of 10 000 random match scores.

3.4 A Basic Reference Score Function

We define a first score function L_1, which we will refer to as the minimal score function. Every potentially improved score function will be compared to this one. L_1 is a slight extension (charge state dependence) of a score function introduced in [4] and it can be derived as follows. We assume that fragment matches (tolerance given) constitute independent events and the probability of these events depends on the ion type $\theta \in S$ and the peptide charge state z. We denote this probability $p_\theta(z)$. Now, let $s = a_1 \cdots a_n$ be a peptide sequence and a_i its amino acids. The probability of a correct match between s and an experimental spectrum is estimated by taking the product of $p_\theta(z)$ for every matched fragment and of $1 - p_\theta(z)$ for every unmatched fragment. The null-model is identical with random fragment match probabilities $r_\theta(z)$. We find

$$L_1 = \log \left(\prod_{i=1}^{n} \left[\prod_{\theta \in M(s,i)} \frac{p_\theta(z)}{r_\theta(z)} \prod_{\theta \in S(s,i)-M(s,i)} \frac{1 - p_\theta(z)}{1 - r_\theta(z)} \right] \right) .$$

$S(s,i) \subset S$ is the set of ion types ending at amino acid a_i, $M(s,i) \subset S(s,i)$ is the set of ion types matching experimental fragment mass. $S(s,i)$ may be a proper subset of S because certain ions are not always possible depending on the fragment last amino acid (neutral loss). $p_\theta(z)$, $\theta \in S$, are learnt from a set of correct matches. The probabilities of random fragment matches $r_\theta(z)$ are learnt from random peptides.

We use relative entropy in bit $H_\theta(z) = p_\theta(z) \log_2(p_\theta(z)/r_\theta(z))$ to measure the importance of each ion type. For $z = 1, 2, 3$, we empirically determined a threshold that is a constant divided by the average peptide length given z. The performance of L_1 and the list of ion types selected are robust with respect to threshold variations. In decreasing order of $H_\theta(z)$, we use: ($z = 1$) y, b-18, b, y-17, b-17, ($z = 2$) y, b, b-18, b-17, y-18, y-17 and ($z = 3$) y, b^{++}, b, b^{++}-18, y^{++}, b^{++}-17, y^{++}-18, y-18, y^{++}-17, b-18. The performance results are shown in Table 2.

3.5 Consecutive Fragment Matches

The score function L_1 is based on a very strong simplifying assumption: fragment matches are independent events. In theory (very simplified), if an amino acid in the peptide has been protonated, then successive fragments should also contain this protonation site and hence be detected. In reality, the protonation sites are not the same for every copy of a peptide and a probabilistic approach should be followed.

A natural improvement of L_1 would be a model able to “reward” consecutive matches, while tolerating occasional gaps. This can be achieved by using a

Table 2. Performance comparison. Percentage of correct match scores that are equal (ratio=1.0), 20% superior (1.2) or 40% superior (1.4) to the best of 10 000 random match scores, which are generated for each correct match. As the test sets we use comprise numerous good matches, which are treated easily, we also report performance on matches having a L_1 score between 0 and 10. Such lines are marked by an asterisk*. Lines marked with ^{B} refer to the complementary data set B.

	Charge 1			Charge 2			Charge 3		
Function	1.0	1.2	1.4	1.0	1.2	1.4	1.0	1.2	1.4
L_1	75.0	56.2	42.1	97.4	94.7	90.4	96.4	94.4	91.9
L_{consec}	73.3	54.8	40.8	97.8	95.0	91.1	97.8	96.2	93.9
L_{intens}	79.5	57.5	40.9	97.8	95.3	91.2	97.7	95.8	93.5
$L_{1,\text{class}}$	73.2	57.0	46.5	97.6	95.2	91.9	96.4	94.7	91.8
L_{iClass}	76.6	57.5	42.7	97.8	94.9	90.8	97.5	96.0	93.1
L_1^*	66.5	47.7	36.7	83.9	79.3	75.0	82.7	77.8	74.0
L_{consec}^*	73.8	55.6	42.5	85.5	79.3	75.7	89.0	84.8	82.4
L_{intens}^*	71.1	48.5	35.2	83.9	78.9	76.3	87.0	83.7	79.4
$L_{1,\text{class}}^*$	65.5	50.5	41.9	87.8	84.2	81.2	84.2	80.7	76.6
L_{iClass}^*	67.7	48.5	36.8	84.9	79.6	75.7	86.3	83.0	77.6
L_1^B				98.4	96.5	93.7	98.7	94.7	85.9
L_{intens}^B				99.9	99.8	99.4	99.9	99.9	99.3
L_1^{B*}				88.8	83.1	79.8	95.4	85.1	82.8
L_{intens}^{B*}				99.2	98.1	96.5	98.9	98.9	97.7

Markov chain (MC) or a hidden Markov model (HMM) [5]. In this perspective, as observed in [3], it may be advantageous to unify several ion types in one generalized ion type to better capture the consecutive fragment match pattern. For instance, one may want to consider ion types b, b-17, b-18, b^{++} as one general ion type B.

We consider one MC and two HMMs, see Figure 1, and we denote by L_2 the log-likelihood ratio of models for consecutive matches. Given a choice of model (MC or HMM), we define a new score function $L_{\text{consec}} = L_1 L_2$, $L_2 = \prod_{\theta \in S'} L_{2,\theta}$, where S', the set of generalized ion types, and $L_{2,\theta}$, the corresponding log-likelihood ratios.

For each peptide charge state, we test every combination of the models hmmA, hmmJ and mcA both (Fig. 1) for the alternative (H_1) and the null hypotheses (H_0), with orders $n = 2, 3, 4$, order(H_0 model) $\leq$ order(H_1 model). That is 84 L_2 models in total. We empirically set S' to ($z = 1$) Y={y}, ($z = 2$) B={b, b-18, b-17}, Y={y, y-18, y-17} and ($z = 3$) B={b, b^{++}, b^{++}-18, b^{++}-17}, Y={y, y^{++}, y^{++}-18}. The parameters are learnt by expectation maximization (Baum-Welch Algorithm, [5]).

We found no unique combination that would dominate the other ones. Several combinations perform well and, as a general tendency, we have observed that HMMs have a slight advantage for the H_1-model, whereas MCs are sufficient for the H_0-model. The performance of the various combinations is simi-

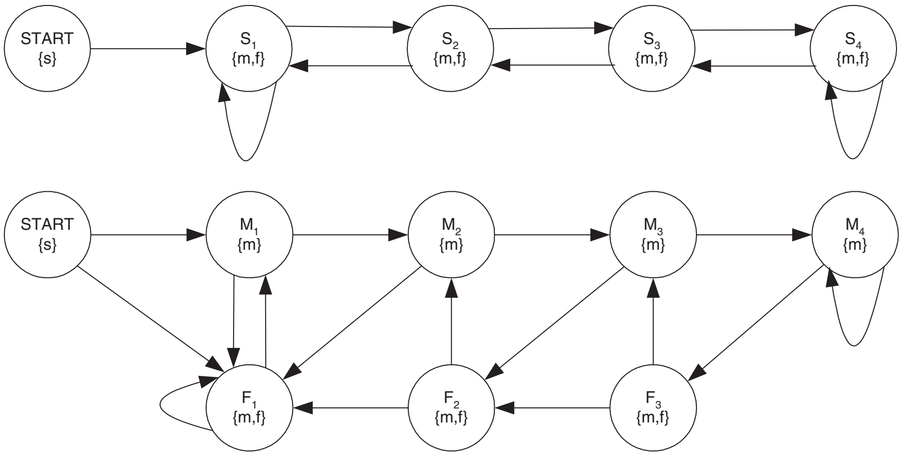


Fig. 1. Consecutive fragment matches models. Two models aimed at capturing successive fragment match patterns: hmmJ (top) and hmmA (bottom). Circles represent states and the letters between curly brackets are the emitted symbols. The structures shown correspond to what we name order 4. The symbol 'm' emitted by a state M_i represents a correct fragment match, while the symbol 'm' emitted by a state F_i represents a random fragment match (this is even possible in a correct peptide match). 'f' represents a theoretical fragment mass not matched in the experimental data.The model mcA is identical to hmmA except for the states F_i, which only emit the symbol 'f'. The structure of these models is designed to allow for higher match probability after a first match has been found. It is also designed for accepting a few missing matches in a longer match sequence.

lar, therefore indicating the intrinsic importance of considering consecutive "no matter" the exact method. The performance of the best combinations ($z = 1$: hmmA(3)/mcA(2), $z = 2$: hmmA(4)/mcA(3), $z = 3$: hmmA(4)/hmmJ(2)) is shown in Table 2. In every case, the transition and emission probabilities nicely fit the model structures. For hmmJ(2), Y ions and charge 2, we find the transitions START to S_1 (probability 1), S_1 to S_1 (0.59), S_1 to S_2 (0.41), S_2 to S_2 (0.88), S_2 to S_1 (0.12) and emissions S_1 ('m' with probability 0.05, 'f' 0.95), S_2 ('m' 0.97, 'f' 0.03).

3.6 Signal Intensity

It is well known among the mass spectrometry community that different ion types have different typical signal intensity. In the case of tryptic peptides, C-terminal ion types (x, y, z) naturally produce more intense peaks. This is due to the basic tryptic cleavage sites (Lys, Arg), which facilitate protonation. [27] and [8] even report fragment relative length intensity dependence.

Here we use a simple model that orders the experimental peaks by intensity and then split them into 5 bins. We obtain $L_{\text{intens}} = L_1 L_3$, $L_3 = \prod_{\theta \in S''} L_{3,\theta}$, S'', a set of ion types, and $L_{3,\theta}$, the corresponding log-likelihood ratios. By selecting ion types for their significance (relative entropy), we set S'' to ($z = 1$) b, b-17, b-18, y, y-17, y-18, ($z = 2$) b, y and ($z = 3$) b-17, b^{++}, y, y-17, y^{++}. The performance of L_{intens} is shown in Table 2.

Although signal intensity improves performance, we expected a more spectacular change. By further investigating, we found a direct explanation for this disappointing result. Bruker peak detection software allows for exporting the n most intense peaks above noise level into the peak list. At the time we generated our main data set, n was set to 100. Ion types like b-18, b-18, y-17, y-18, b^{++}, y^{++} are generally important for scoring, although their signal is much less intense than b or y signal [27]. Now, given that longer peptides statistically have more protonation sites, it is clear that singly charged peptides are shorter. Therefore, $n = 100$ is sufficient to cover most of the fragment masses. On the other hand, it turns out that it is not sufficient to include enough fragment masses in the model L_{intens} for $z = 2, 3$. We used the complementary data set B, which was generated with $n = 200$. We denote by L^B_{intens} the model L_{intens} trained and tested on this set. The results shown in Table 2 nicely confirm the explanation. Since the spectra in data set B were acquired on a different (better) instrument and the samples were made of purified proteins (not a biological sample), we repeat the performance of L_1 (renamed L^B_1) for reference purpose.

In theory, L_3 includes L_1 and one could expect that L_3 performance is similar to L_{intens}. In practice, L_3 performance is very inferior to L_{intens} (less than 66% for a ratio of 1 in Table 2), even on data set B. The reason is that the pattern captured by L_1 is more or less always available, whereas the intensity pattern is more variable.

3.7 Amino Acid Dependence

Depending on their amino acid sequence, fragments may be more or less easily detected. The actual dependence involves several phenomena (dissociation, ionization). A model making use of the whole fragment sequence would contain too many parameters. It is commonly accepted that the last amino acid of a fragment (cleavage site) plays a significant role in the above mentioned phenomena. We limit the number of model parameters by only considering the last amino acid of a fragment and by grouping them in classes.

Basic score revisited. We designed an improved version of L_1, which we name $L_{1,\text{class}}$, that uses parameters $p_\theta(z, c)$, $r_\theta(z, c)$, c a set of amino acids, and whose performance is shown in Table 2. We empirically determined the following amino acid classes (amino acids with similar probabilities): (N-tern ions) 'AFHILMVWY', 'CDEGNQST', 'KPR' and (C-term ions) 'HP', 'ACFIMDEGLNQSTVWY', 'KR'.

Consecutive fragment matches revisited. It is possible to extend hmmA, hmmJ and mcA by replacing their states by one state per amino acid class to separate amino acids that inhibit fragmentation from amino acids that favor fragmentation. As we prefer to stay with simple and robust models, we have not implemented the amino acid dependent versions of hmmA, hmmJ and mcA.

Signal intensity revisited. The model we introduced in Section 3.6 can be extended as we did for $L_{1,\text{class}}$, thus obtaining a model $L_{\text{iClass}} = L_1 L_{3,\text{class}}$. The performance of the latter is reported in Table 2. L_{iClass} performs much worse than L_{intens}, what we explain by the fact that, although, the last amino acid plays a major role in the fragment dissociation phenomena, it is no strong relation with signal intensity. Signal intensity is more a consequence of the ion type and the entire peptide sequence.

4 An Efficient Score

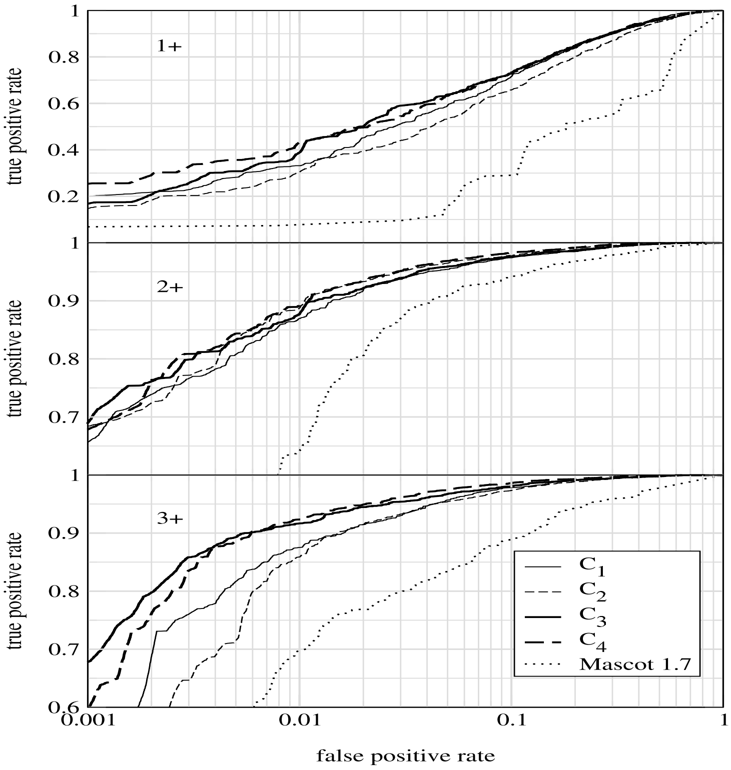
L_1 is significantly improved by considering signal intensity. Consecutive fragment matches as well as the amino acid dependent version of L_1 also improve the performance. Accordingly, we tested 4 combinations, which are $C_1 = L_{\text{intens}}$, $C_2 = L_{1,\text{class}} L_2$, $C_3 = L_{\text{intens}} L_3$ and $C_4 = L_{1,\text{class}} L_2 L_3$, see Figure 2.

The receiver operating characteristics (ROC) curves of Figure 2 (top) are obtained by setting a threshold on match p-values. The correct match p-values are computed by searching the peptides of our data set against a database of 15 000 human proteins with variable Cys_CAM and oxidation (Met, His, Trp) modifications. The random match p-values are computed by searching against a database of 15 000 random proteins with the same variable modifications and by taking the best match. The random protein sequences were obtained by training an order 3 MC on the human protein database. From Figure 2 we observe that C_4 is the best combination. C_4 performance on data set and database [13] is shown in Figure 2 (bottom).

5 Discussion

In this work we designed score functions by assuming the independence of many statistical quantities. This choice allowed us to rapidly design efficient score functions as naive Bayesian classifier. This can be seen as preparative work to identify key contributors to successful peptide score functions in order to design future models, comprising fewer simplifying assumptions. We found that ion type probabilities, signal intensity and consecutive fragment matches are essential to a peptide score function. We omitted one aspect of a match that may be pertinent: the simultaneous detection of several ion types [29] (neutral loss, complementary fragments). Peptide elution times may also provide additional information to select correct matches [21]. New approaches based on the mobile proton model are under development, see for instance [24], and it should be possible to combine them with what we presented here.

The general applicability of modern stochastic score functions has not been addressed here. In particular, two central questions might limit their value: the



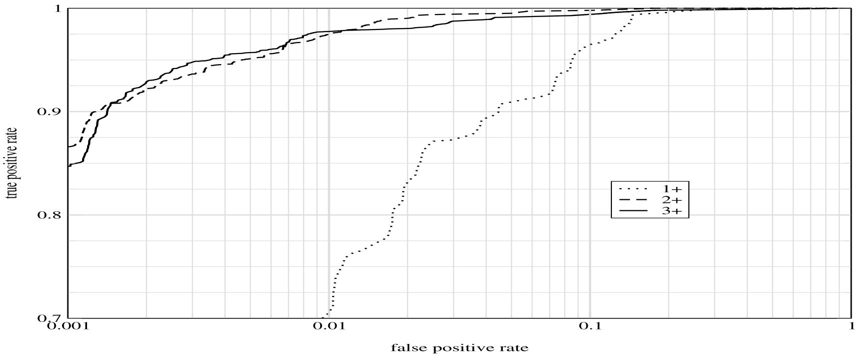


Fig. 2. ROC curves. **Top**. Score functions $C_{1,2,3,4}$ on Bruker Esquire 3000 data. $C_4 = L_{1,\text{class}} L_2 L_3$ is the best score function at every charge state. At charge states 2 and 3, if we fix a true positive rate of 95%, the improvement is 3.5-fold (charge 2) or 5.8-fold (charge 3) over Mascot. At charge state 1, if we accept a false positive rate of 5%, 4 times more peptides are identified. **Bottom**. ThermoFinnigan LCQ data [13]. The improvement is 14-fold compared to [12].

minimal size of the training set and performance robustness. These two points are addressed in [15] and the results are positive.

We described a generic method for designing peptide score functions. We applied it systematically to a large and diverse MS data set (140 000+ peptides) to design a series of models of increasing complexity. By selecting an appropriate combination of models, we obtained a very efficient score function, which, given a true positive rate of 95%, has a false positive rate as low as 3% (doubly charged peptides) or 2% (triply charged peptides). This is 3.5 to 5.8 times less than Mascot 1.7. On data set [13] our false positive rate is less than 0.5%, which is a 14-fold improvement compared to [12] (Figure 5). In [13] (Table 3), Sequest performance is reported when used "traditionally", i.e. by setting thresholds on Xcorr and other quantities exported by Sequest. The smallest Sequest false positive rate reported (threshold set 4) is 2% with 59% true positive rate. We obtain a corresponding false positive rate of 0.004%, which is 50 times less. The highest Sequest true positive rate reported (threshold set 2) is 78% with 9% false positive rate. We obtain again a corresponding false positive rate of 0.004%, which is 187 times less. Similar performance has been obtained on real samples by using Bruker Esquire 3000+ instruments. Such low false positive rates, the lowest ones ever reported to our best knowledge, combined with the competitive price and robustness of ion trap instruments, provide a highly appropriate technology platform in the perspective of the many academic and private large-scale proteomics projects currently emerging.

6 Acknowledgements

We thank the protein separation, mass spectrometry, LIMS and database teams for providing us with so much interesting data. We are grateful to Paul Seed who read an early version of the manuscript and to Eugene Kolker for data set [13]. We also would like to especially thank the annotation group and Lydie Bougueleret.

References

1. D. C. Anderson, W. Li, D. G. Payan, and W. S. Noble. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.*, 2:137–146, 2003.
2. V. Bafna and N. Edwards. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, 17:S13–S21, 2001.
3. J. Colinge, A. Masselot, M. Giron, T. Dessingy, and J. Magnin. OLAV: Towards high-throughput MS/MS data identification. *Proteomics*, to appear in August, 2003.
4. V. Dancik, T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner. De novo peptide sequencing via tandem mass spectrometry: a graph-theoretical approach. *J. Comp. Biol.*, 6:327–342, 1999.

5. R. Durbin et al. *Biological sequence analysis.* Cambridge University Press, Cambridge, 1998.
6. J. K. Eng, A. J. McCormack, and J. R. III Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, 5:976–989, 1994.
7. H. L. Field, D. Fenyö, and R. C. Beavis. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identifications, and archives data in a relational database. *Proteomics*, 2:36–47, 2002.
8. M. Havilio, Y. Haddad, and Z. Smilansky. Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem.*, 75:435–444, 2003.
9. W. J. Henzel et al. Identifying protein from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA*, 90:5011–5015, 1993.
10. P. James. *Mass Spectrometry.* Proteome Research. Springer, Berlin, 2000.
11. R. S. Johnson et al. Collision-induced fragmentation of $(m+h)^+$ ions of peptides. Side chain specific sequence ions. *Intl. J. Mass Spectrom. and Ion Processes*, 86:137–154, 1988.
12. A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search. *Anal. Chem.*, 74:5383–5392, 2002.
13. A. Keller, S. Purvine, A. I. Nesvizhskii, S. Stolyar, D. R. Goodlett, and E. Kolker. Experimental protein mixture for validating tandem mass spectral analysis. *OMICS*, 6:207–212, 2002.
14. D. C. Liebler, B. T. Hansen, S. W. Davey, L. Tiscareno, and D. E. Mason. Peptide sequence motif analysis of tandem MS data with the SALSA algorithm. *Anal. Chem.*, 74:203–210, 2002.
15. A. Masselot, J. Magnin, M. Giron, T. Dessingy, D. Ferrer, and J. Colinge. OLAV: General applicability of model-based MS/MS peptide score functions. In *Proc. 51st Am. Soc. Mass Spectrom.*, Montreal, 2003.
16. A. L. McCormack et al. Direct analysis and identification of proteins in mixture by LC/MS/MS and database searching at the low-femtomole level. *Anal. Chem.*, 69:767–776, 1997.
17. R. E. Moore, M. K. Young, and T. D. Lee. Qscore: An algorithm for evaluating sequest database search results. *J. Am. Soc. Mass Spectrom.*, 13:378–386, 2002.
18. I. A. Papayannopoulos. The interpretation of collision-induced dissociation mass spectra of peptides. *Mass Spectrometry Review*, 14:49–73, 1995.
19. D. J. Papin, P. Hojrup, and A. J. Bleasby. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.*, 3:327–332, 1993.
20. D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551–3567, 1999.
21. K. Petritis, L. J. Kangas, P. L. Fergusson, G. A. Anderson, L. Paša-Tolić, M. S. Lipton, K. J. Auberry, E. F. Strittmatter, Y. Shen, R. Zhao, and R. D. Smith. Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analysis. *Anal. Chem.*, 75:1039–1048, 2003.
22. H. V. Poor. *An Introduction to Signal Detection and Estimation.* Springer, New York, 1994.
23. R. G. Sadygov, J. Eng, E. Durr, A. Saraf, H. McDonald, M. J. MacCoss, and J. Yates. Code development to improve the efficiency of automated MS/MS spectra interpretation. *J. Proteome Res.*, 1:211–215, 2002.

24. F. Schütz, E. A. Kapp, J. E. Eddes, R. J. Simpson, T. P. Speed, and T. P. Speed. Deriving statistical models for predicting fragment ion intensities. In *Proc. 51st Am. Soc. Mass Spectrom.*, Montreal, 2003.
25. J. K. Skilling. Improved methods of identifying peptides and protein by mass spectrometry. European Patent Application EP 1,047,107,A2., 1999.
26. A. P. Snyder. *Interpreting Protein Mass Spectra.* Oxford University Press, Washington DC, 2000.
27. D. L. Tabb, L. L. Smith, L. A. Breci, V. H. Wysocki, D. Lin, and J. Yates. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem.*, 75:1155–1163, 2003.
28. J. Yates, J. K., and Eng. Identification of nucleotides, amino acids, or carbohydrates by mass spectrometry. United States Patent 6,017,693, 1994.
29. N. Zhang, R. Aebersold, and B. Schwikowski. ProbId: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, 2:1406–1412, 2002.