David N. Perkins[1]
Darryl J. C. Pappin[1]
David M. Creasy[2]
John S. Cottrell[2]

[1]Imperial Cancer Research Fund, London, UK
[2]Matrix Science Ltd., London, UK

Probability-based protein identification by searching sequence databases using mass spectrometry data

Several algorithms have been described in the literature for protein identification by searching a sequence database using mass spectrometry data. In some approaches, the experimental data are peptide molecular weights from the digestion of a protein by an enzyme. Other approaches use tandem mass spectrometry (MS/MS) data from one or more peptides. Still others combine mass data with amino acid sequence data. We present results from a new computer program, Mascot, which integrates all three types of search. The scoring algorithm is probability based, which has a number of advantages: (i) A simple rule can be used to judge whether a result is significant or not. This is particularly useful in guarding against false positives. (ii) Scores can be compared with those from other types of search, such as sequence homology. (iii) Search parameters can be readily optimised by iteration. The strengths and limitations of probability-based scoring are discussed, particularly in the context of high throughput, fully automated protein identification.

Keywords: Protein identification / Mass spectrometry / Bioinformatics EL 3725

1 Introduction

Mass spectrometry (MS) has become the method of choice for the rapid identification of proteins and the characterisation of post-translational modifications [1]. Several algorithms and computer programs have been described in the literature for protein identification by searching a sequence database using mass spectrometry data. Since the first publications on this topic appeared in 1993, there have also been a number of reviews. A recent article by Yates [2] provides a concise overview of the subject and comprehensive references to the literature. In some approaches, the experimental data are peptide molecular weights from the digestion of a protein by an enzyme (a peptide mass fingerprint) [3–7]. Other approaches use MS/MS data from one or more peptides (an MS/MS ions search) [8]. Still others combine mass data with explicit amino acid sequence data or physicochemical data which infer sequence or composition (a sequence query) [9].

The general approach in all cases is similar. The experimental data are compared with calculated peptide mass or fragment ion mass values, obtained by applying appropriate cleavage rules to the entries in a sequence database. Corresponding mass values are counted or scored in a way that allows the peptide or protein which best matches the data to be identified. If the "unknown" protein is present in the sequence database, then the aim is to pull out the correct entry. If the sequence database does not contain the unknown protein, then the aim is to identify those entries which exhibit the closest homology, often equivalent proteins from related species. While several algorithms assign scores to matches, we are not aware of any systematic attempts to report scores which accurately reflect true probabilities. The advantages of probability-based scoring include: (i) A simple rule can be used to judge whether a result is significant or not. This is particularly useful in guarding against false positives. (ii) Scores can be compared with those from other types of search, such as sequence homology. (iii) Search parameters can be readily optimised by iteration.

We present results from a new search engine, Mascot, which incorporates probability-based scoring. All three types of search are supported: peptide mass fingerprint, sequence query, and MS/MS ions search. Any FASTA format sequence database can be searched, nucleic acid databases being translated in all six reading frames on the fly. The program, which is threaded for parallel execution on multiprocessor machines and clusters, has been ported to Microsoft Windows NT, SGI Irix, Sun Solaris, and DEC Unix, and can be freely accessed across the World Wide Web at Uniform Resource Locator (URL) http://www.matrixscience.com.

Correspondence: Dr. D. J. C. Pappin, Imperial Cancer Research Fund, Protein Sequencing Laboratory, 44 Lincoln's Inn Fields, London WC2A 3PX, UK
E-mail: d.pappin@icrf.icnet.uk
Fax: +44-171-269-3093

Abbreviations: PSD, post-source decay; **SMA,** *N*-succinimidyl-2-morpholine acetate; **URL**, uniform resource locator

0173-0835/99/1818-3551 $17.50+.50/0

2 Materials and methods

2.1 Sample preparation and mass spectrometry

Protein bands were stained with silver or Coomassie Brilliant Blue, excised from an SDS-PAGE gel, and digested overnight with trypsin [10]. Aliquots of 0.5–1 µL were generally sampled directly from the digest supernatant for MS fingerprint analysis using a TofSpec 2E MALDI time-of-flight (TOF) instrument (Micromass, Manchester, UK). The remaining digested peptides (>90% of total digest) were then reacted with *N*-succinimidyl-2-morpholine acetate (SMA) in order to enhance b ion abundance and facilitate sequence analysis by MS/MS [11, 12]. Derivatised peptides were eluted with a single step gradient to 75% v/v methanol/0.1% v/v formic acid and fragmented by low-energy collision-activated dissociation using an LCQ ion-trap MS (ThermoQuest, San Jose, CA, USA) fitted with a nano-electrospray source [13, 14].

2.2 Database search engine

The search engine used in this work, Mascot, is a development of the MOWSE computer program [6, 15]. Significant differences between MOWSE and Mascot are the addition of probability-based scoring, support for matching MS/MS data, and the removal of prebuilt indexes. Mascot works directly from the FASTA format sequence databases which, for maximum search speed, may be compressed and mapped into memory. For interactive searching, the user interface to Mascot is a web browser, and searches are defined using hypertext mark-up language (HTML) forms. A form may be used to enter search parameters and data and may also specify a local text file to be uploaded to the server. This uploaded file can contain both experimental data and search parameters. The Mascot search engine, written in ANSI C, is executed as a common gateway interface (CGI) program, (Fig. 1). On completion of a search, it calls a Perl CGI script which reads the results file and returns an HTML report to the client browser. Links to additional CGI scripts provide more detailed views of the results.

MS data are submitted to Mascot in the form of peak lists. That is, lists of centroided mass values, optionally with associated intensity values. In the case of MS/MS data, peak detection is also required in the chromatographic dimension, so that multiple spectra from a single peptide are summed together and spectra from the chromatogram baseline are discarded. Accurate and efficient data reduction is a critical factor in getting the best out of the search engine. Figure 2 illustrates the search form for a peptide mass fingerprint. Although Mascot accepts all three types of searches, putting the parameters for all search types into a single form was found to make it too complex. The fields are mostly self-explanatory, and further details can be found in the web site help text.

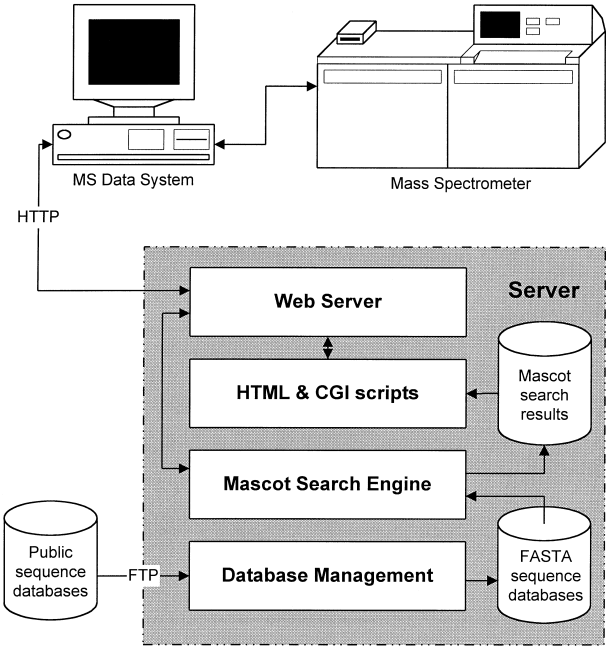
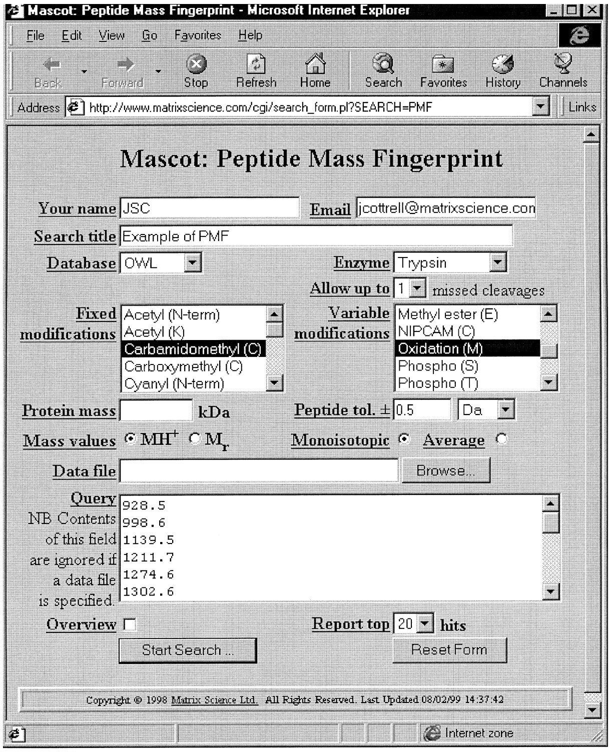


Figure 1. Functional block diagram of web-based interactive searching


Mascot: Peptide Mass Fingerprint - Microsoft Internet Explorer
File Edit View Go Favorites Help
Back Forward Stop Refresh Home Search Favorites History Channels
Links
Browse...
Start Search ...
Reset Form
Internet zone

Address http://www.matrixscience.com/cgi/search_form.pl?SEARCH=PMF

Mascot: Peptide Mass Fingerprint

Your name JSC Email jcottrell@matrixscience.con

Search title Example of PMF

Database OWL Enzyme Trypsin

Allow up to 1 missed cleavages

Fixed modifications: Acetyl (N-term), Acetyl (K), Carbamidomethyl (C), Carboxymethyl (C), Cyanyl (N-term)

Variable modifications: Methyl ester (E), NIPCAM (C), Oxidation (M), Phospho (S), Phospho (T)

Protein mass kDa Peptide tol. ± 0.5 Da

Mass values ● MH^+ ○ M_r Monoisotopic ● Average ○

Data file

Query
NB Contents of this field are ignored if a data file is specified.

```
928.5
998.6
1139.5
1211.7
1274.6
1302.6
```

Overview □ Report top 20 hits

Copyright © 1998 Matrix Science Ltd. All Rights Reserved. Last Updated 08/02/99 14:37:42

Figure 2. Mascot peptide mass fingerprint search form

2.3 Probability-based scoring

The fundamental approach is to calculate the probability that the observed match between the experimental data set and each sequence database entry is a chance event. The match with the lowest probability is reported as the best match. Whether the best match is also a significant match depends on the size of the database. To take a simple example, the calculated probability of matching six out of ten peptide masses to a particular sequence might be 10^{-5}. This may sound like a promising result but, if the real database contains 10^6 sequences, several scores of this magnitude may be expected by chance. A widely used significance threshold is that the probability of the observed event occurring by chance is less than one in twenty ($p < 0.05$). For a database of 10^6 entries, this would mean that significant matches were those with probabilities of less than 5×10^{-8}. The probability for a good match is usually a very small number, which must be expressed in scientific notation. This can be inconvenient, so we have adopted a convention often used in sequence similarity searches, and report a score which is $-10\text{Log}_{10}(P)$, where P is the probability. This means that the best match is the one with the highest score, and a significant match is typically a score of the order of 70.

2.4 Testing

Pearson [16] has described how the performance of biological sequence comparison algorithms should be judged on two criteria: (i) sensitivity, the ability to calculate high-ranking scores for distantly related sequences; and (ii) selectivity, the ability to calculate low-ranking scores for unrelated sequences. The performance of algorithms for protein identification based on MS data can be judged on a similar basis: (i) sensitivity, the ability to make a correct identification using weak or noisy data; and (ii) selectivity, the ability to calculate low-ranking scores for spurious, random matches. Judging the sensitivity and selectivity of the algorithms in Mascot can only be done with knowledge of the "correct" answer. While this could be approached by using artificial data sets, all the examples given here use real experimental data. We do not believe that calculated data can provide a valid basis for evaluating sensitivity and selectivity. Factors such as systematic calibration errors, nonspecific enzyme behaviour, gas-phase ion fragmentation kinetics, contributions from contaminating proteins, instrument artefacts, unsuspected modifications, *etc.*, are extraordinarily difficult to simulate with any realism. It is also important to test the algorithms against the widest possible variety of data sets.

As far as statistical significance is concerned, the validity of the probabilities calculated by Mascot can be tested by repeating a search against a randomised sequence database. In this work, we use a database of representative sequences [17]. That is, a database in which the overall amino acid composition, the number of entries, and the distribution of entry lengths are identical to a real database, but with random sequences. No attempt has been made to preserve nearest-neighbour frequencies. Another valuable check is to submit the same search to multiple search engines and compare the results. Details of other search engines can be found at the following URLs: MassSearch [4] http://vinci.inf.ethz.ch/ServerBooklet/MassSearchEx.html; MOWSE [6] http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse; Expasy tools [18] http://www.expasy.ch/tools/; PeptideSearch [9] http://www.mann.embl-heidelberg.de/Services/PeptideSearch/PeptideSearchIntro.html; Protein Prospector [19] http://prospector.ucsf.edu/; Prowl [20] http://prowl.rockefeller.edu/PROWL/prowl.html; and Sequest [8] http://thompson.mbt.washington.edu/sequest/.

2.5 The model

A critical step in any statistical analysis is the definition of an appropriate model. An ideal model would faithfully represent the underlying physical system. Unfortunately, the physical processes which determine the observed data in a protein identification experiment are of great complexity, and only the most important factors can be included in the model. In addition, there are some physical factors which can be modelled, but which result in overly complex expressions, or mathematical series without closed forms. Even with powerful computer hardware, simple and efficient code is essential in order to complete a search of a large database in a reasonable amount of time. This means that it is sometimes necessary to ignore a physical factor in the interests of throughput even though, in principal, it could be included in the model.

2.5.1 Proteolysis

MOWSE [6] was the first protein identification program to recognise that the relative abundance of peptides of a given length in a proteolytic digest depends on the lengths of both peptide and protein. For trypsin, cleaving after arginine and lysine unless followed by proline, approximately 10% bonds are cleavage sites. In a protein of infinite length, the fractional abundance of ideal trypsin limit peptides of length N residues is simply $A(1\text{-}A)^{N\text{-}1}$, where A is the fractional abundance of bonds which are cleavage sites. This distribution is shown in Fig. 3 for three different cleavage agents.

Of course, real proteins are not of infinite length. Finite proteins have an "end effect" which increases the abun-

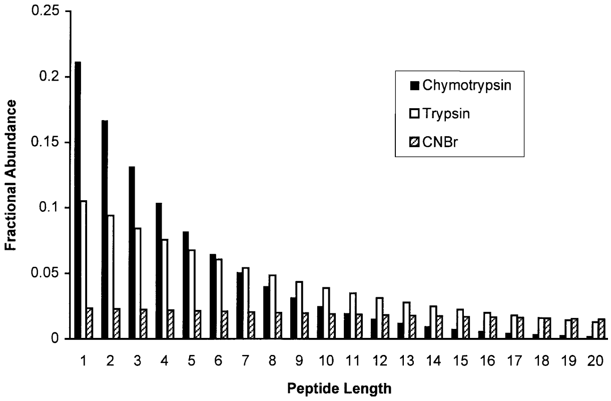


Figure 3. Calculated peptide length distributions for three cleavage agents of differing specificity acting on a protein of infinite length: chymotrypsin, trypsin, and cyanogen bromide

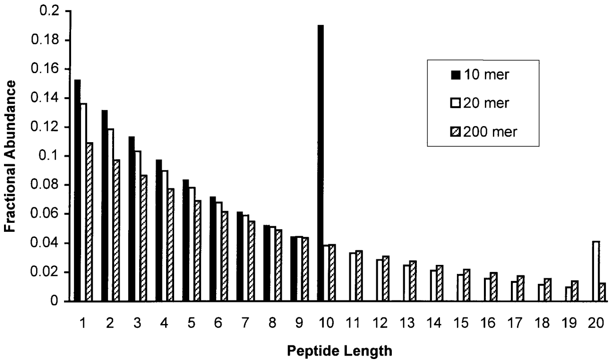


Figure 4. Calculated peptide length distributions for tryptic limit peptides from proteins of length 10, 20 and 200 residues

dance of short peptides and dramatically increases the probability of finding the peptide equal in length to the protein (*i.e.*, no cleavage). Figure 4 shows the fractional abundance of peptides as a function of their length for trypsin acting on proteins of length 10, 20, and 200 residues.

The next level of complexity in modelling proteolysis is to allow for missed cleavage sites. Missed cleavages occur for a number of reasons. One mechanism, which we are unable to include in the model, is steric hindrance, making a cleavage site inaccessible to the enzyme. Another factor, which can significantly influence cleavage probability, is the identity of the residue adjacent to the cleavage site. For example, trypsin is less likely to cleave a substrate when there is a basic residue (arginine, lysine) adjacent to the cleavage site [21, 22]. Although this effect was included in the original MOWSE model, it has been dropped from Mascot in the interests of simplicity and execution speed. The final cause of missed cleavages is simple kinetics. Either the enzyme-to-substrate ratio is too low or the time allowed is insufficient for digestion to proceed to completion. This factor is included in the Mascot model by allowing the user to specify that a peptide may include missed cleavage sites up to an arbitrary maximum number.

2.5.2 Modifications

Post-translational modifications, and modifications due to chemical derivatisation, contribute greatly to the complexity of mass-based searching. Often, there is uncertainty as to whether a particular modification is present or not. Even if present, a modification may not be quantitative. For example, a peptide may contain some oxidised and some nonoxidised methionine residues. Three classes of modification can be identified: (i) Modifications which affect a specific residue, only when that residue is at a peptide terminus (*e.g.*, conversion of *N*-terminal glutamine to pyro-glutamic acid); (ii) modifications which affect a peptide terminus, independent of the identity of the residue (*e.g.*, esterification of the *C*-terminus); (iii) modifications which affect a residue independent of its position in the peptide (*e.g.*, oxidation of methionine). Mascot supports all three classes of modification, which may be specified as being quantitative or nonquantitative. However, the number of nonquantitative modifications is limited to a maximum of four. This is because nonquantitative modifications substantially increase the number of calculated mass values, and so raise the level of random matches. This makes it inadvisable to specify a large number of nonquantitative modifications in a search; better to risk missing one or two peptides than compromise specificity on the remainder.

Matching MS/MS data from a peptide which contains nonquantitative modifications raises an interesting issue. Consider a peptide which contains three methionine residues, one of which is oxidised. Assuming that all three methionines are equally susceptible to oxidation, the experimental MS/MS spectrum will contain contributions from three different permutations of oxidised and nonoxidised methionines. All three permutations have the same molecular weight, but give rise to differing MS/MS spectra. Thus, the Mascot model attempts to match the experimental MS/MS data to the sum of the contributions from all possible permutations of nonquantitative modifications which fall within the mass error window specified for the peptide. Some nonquantitative modifications, such as enzymatic phosphorylation, are likely to be site-specific. In such cases, with good data, a more thorough matching procedure which included individual permutations and combinations of permutations might be expected to reveal the location of the modified residues. However, this has not yet been incorporated into the Mascot code.

Mascot does not attempt to make use of the information concerning known post-translational modifications and processing present in database annotations. The feasibility of reading SWISS-PROT annotations has been demonstrated by the MultiIdent program [18]. This facility, though undoubtedly useful when searching SWISS-PROT and other well-annotated protein databases, does not eliminate the need to search for nonquantitative modifications. Also, database annotations cannot help with modifications due to sample handling, such as oxidation of methionine, or acrylamide adduction to cysteine. In any case, the bulk of database entries are translated from nucleic acid sequences, and so cannot include information on experimentally observed modifications.

2.5.3 Mass accuracy

Mascot, in common with most other search engines, requires the user to provide an error window on the measured mass values. This is a particularly important parameter. Specifying a window which is too large will increase the level of random matches and so reduce discrimination. However, specifying too narrow a window is much worse, because valid matches will be missed. The Mascot model assumes that mass measurement errors should be treated as being uniformly distributed across the specified error window. Although the random component of the error might be expected to follow some kind of quasi-normal distribution about zero, there is also a systematic component, due to calibration error, which will result in values being high or low as a function of mass. Thus, if the estimated error window is ± 0.25 Da, then a match with an error of 0.2 Da is assumed to be "as good as" one with an error of 0.02 Da. If this was not the case, then mass error could be treated as a variable in the probability calculation, and used to select the set of matches with the lowest probability [4].

The Mascot model further assumes that mass values are smoothly distributed, which is not actually the case. As described by Mann [23], the limited elemental composition of proteins means that peptide mass values are clustered around discrete values, separated by intervals of 1.00048 Da (monoisotopic). In consequence, for accurate data, the number of random matches is not proportional to the width of the error window once the error window becomes comparable in size, or smaller than one Dalton. To obtain well-behaved scores from accurate data, the width of error window is treated as asymptotically approaching ± 0.25 Da. Thus, for perfectly accurate data, the score of the best match would tend to a maximum as the width of the error window was reduced to zero. Note that this is distinct from the accuracy with which mass values are calculated to determine if there is a match, currently set to 1/65536 Da.

2.5.4 Average amino acid composition

Mascot calculations are based on the average amino acid composition of the Owl database [24]. For example, the length of a peptide for scoring purposes is estimated by dividing its molecular mass by 111. Although small differences in average amino acid composition are found between the major databases, the consequences for the scoring scheme are negligible.

2.5.5 MS/MS fragment ion series

MS/MS fragment ion data are matched to calculated values for user-selected ion series [25, 26]. The choice of ion series is important. Failure to select a series which is well represented in the experimental data will mean that potential matches are missed. Conversely, selection of a series which is not well represented in the data simply contributes to the tally of random matches. The ion series supported by Mascot are listed in Table 1. There are three sets of series for common experimental conditions, while any selection of the nine supported series can be saved and used as a custom set. Several common types of instrument have lower mass accuracy for MS/MS fragments than for intact peptides. A typical mass error window for MS/MS fragments might be ± 0.5 Da. Since each ion series contributes one calculated mass value per residue, the probability of finding a random match between a calculated and experimental value for a ± 0.5 Da error window is approximately 1% per ion series. Unless the MS/MS data are exceptionally clean, selecting more than four ion series can only bring diminishing returns. High charge state precursors pose a problem, because there is the potential for multiple charge states for each ion series. Matching fragment ions with charge states greater than 2^+ should probably be limited to data from instruments capable of determining and specifying the charge states of the fragment ions. Otherwise, the calculated values will tend to swamp the mass scale, and discrimination will be lost.

2.5.6 Protein molecular mass

Peptide mass fingerprint algorithms which simply count matches rely on the user to specify a molecular mass for the protein. Otherwise, the best match will always be to the most massive proteins, such as titin (3 MDa). Specifying the molecular mass of the intact protein in Mascot is not normally necessary, because the score is a true probability that the match is random,

Table 1. The MS/MS fragment ion series supported by Mascot

Ion type	Ion mass[a)]	Low energy CID	High energy CID	PSD	Custom weighting factor
a	[N]+[M]-CO		1	1	□
a^*	a-NH_3			1	□
a^o	a-H_2O				
a^{++}	(a+H)/2		1		□
b	[N]+[M]	1	1	1	☑
b^*	b-NH_3			1	□
b^o	b-H_2O				
b^{++}	(b+H)/2	1	1		□
c	[N]+[M]+NH_3				
d	a-partial side chain				
v	y-complete side chain				
w	z-partial side chain				
x	[C]+[M]+CO				
y	[C]+[M]+H_2	1	1	1	☑
y^*	y-NH_3				□
y^o	y-H_2O				
y^{++}	(y+H)/2	1	1		□
z	[C]+[M]-NH				

a) [N], mass of *N*-term group
[C], mass of *C*-term group
[M], mass of the sum of the neutral amino acid residue masses

which takes protein length into account. If there are valid reasons to specify the protein molecular mass, simply restricting matches to database entries based on the calculated mass of the entire sequence is highly inadvisable, because many of the sequence database entries are for the least processed form of a protein. For example, the SWISS-PROT entry for bovine insulin, INS_BOVIN, is actually the sequence of the precursor protein including signal and connecting peptides. This adds up to a molecular mass of 11 394 Da, so that a search based too tightly around an experimental measurement of the molecular mass of this protein (5734 Da) would fail to find a correct match.

In Mascot, if a protein molecular mass is specified, this is applied as a sliding window on the database sequences, as first suggested by Yates [7]. For example, if the protein molecular mass was specified as 20 kDa then, in any database entry which exceeds this mass, the code looks for the highest scoring set of matches which occur within a 20 kDa window. In this way, a protein can be correctly scored even though it is substantially shorter than the database entry, for example a proteolytic fragment of a larger protein.

2.5.7 Making use of peak intensity values

Intensity information is ignored in a peptide mass fingerprint. The dominant ionisation techniques, MALDI and ESI, are far from quantitative. Peak intensities depend strongly on the physical and chemical properties of the analytes, so that it would be rash to assume that the more intense peaks were more “valid” than the weaker ones. While it is true that peaks below a certain intensity are more likely to be random noise, it has been our experience that this is not a serious problem in data sets submitted for peptide mass fingerprint searches. Large peaks are as likely to remain unassigned as small ones. In other words, the “noise” is mainly chemical (peptides from other proteins, nonspecific enzyme cleavage, unsuspected modifications, *etc.*) rather than random (shot noise, electrical and electronic artefacts, *etc.*).

In the case of MS/MS spectra, relative peak intensities within a fragment ion series are a function of several complex processes, including composition-based fragmentation kinetics, parent ion activation parameters, and mass analyser artefacts [26]. Because MS/MS spectra tend to exhibit much higher levels of apparently random noise, often a peak at every mass, it becomes essential for peaks to be selected on the basis of intensity. The Mascot code iteratively searches for the set of the most intense peaks which yields the highest score. At least, in the case of an MS/MS spectrum, we know what an ideal spectrum should look like: a uniform ladder of peaks for each fragment ion series. This suggests the possibility of correcting for mass analyser artefacts by normalising peak intensities so as to approach an ideal ladder spectrum prior to intensity-based peak selection. This is standard practice in library search algorithms for electron impact mass spectra, where a typical approach is to select the most intense peak in each 14 Da mass interval. Intensity normalisation is a direction that will be pursued in future work.

2.5.8 Nucleic acid translation

Nucleic acid databases are translated on the fly in all six reading frames. In most cases, the databases of interest contain expressed sequence tags (ESTs) [27]. For EST searches, the code does not look for a start codon, but begins translation at the start of the entry. If it finds a stop codon, this is treated as a gap, and translation is restarted at the next codon. Codons containing base ambiguities sometimes translate to nonambiguous amino acid residues. For example, ATH, where H is A or C or T, translates to isoleucine. The current version of Mascot does not attempt to identify such cases; all codons which include ambiguities are translated to the unknown amino acid residue, X.

Table 2. Syntax for specifying amino acid sequence information in a Mascot search

Prefix	Meaning	Example
b-	*N*- > *C*-sequence	seq(b-DEFG)
y-	*C*- > *N*-sequence	seq(y-GFED)
*-	Orientation unknown	seq(*-DEFG)
n-	*N*-terminal sequence	seq(n-ACDE)
c-	*C*-terminal sequence	seq(c-FGHI)

If the sequence orientation is unknown, Mascot searches for both senses. If no prefix is specified, the default is b-.

2.5.9 Sequence query

In a sequence query, amino acid sequence or composition data may be associated with one or more peptide masses [9]. If such information is present, it is treated as a rigorous filter on the candidate sequences. Ambiguous sequence or composition data can be used (in a manner similar to a regular expression search in computing) but it still functions as a filter, not a probabilistic match of the type found in a BLAST or FASTA homology search. The sequence information is specified in standard one-letter code, preceded by a prefix as outlined in Table 2, to indicate in which direction the sequence should be read.

All examples in Table 2 would match a peptide with the sequence ACDEFGHI. Note also that y-GFED is written *C*-term to *N*-term, whereas c-FGHI is written *N*-term to *C*-term. An unknown amino acid may be indicated by an 'X'. More than one amino acid may be specified for a position by putting them between square brackets. A line may contain several sequence information qualifiers.

Amino acid composition data may be specified by a number, followed by one or more amino acid codes in square brackets. An asterisk means at least one. For example 1234 comp(2[H]0[M]3[DE]*[K]) indicates a peptide which contains two histidines, no methionines, a total of three acidic residues (glutamic or aspartic acid) and at least one lysine. Note that 'X' is not meaningful in a composition query and is not allowed.

The code does not make exhaustive checks on the validity of combinations of multiple sequence and composition qualifiers. For example, the following would all be accepted, even though they are not reasonable: (i) specifying (c-ACD) for a tryptic digest, even though the *C*-termini of all but one peptide per protein can only be K or R; (ii) conflicts between sequence and composition qualifiers, *e.g.*, seq(*-ACD) comp(0[C]); and (iii) duplicate sequence qualifiers, *e.g.*, seq(c-ACD) seq(*-ACD).

3 Results and discussion

Figures 5–7 illustrate typical result reports. (The experimental details of this search are discussed in Section 3.3). At the top of the page are a few lines to identify the search uniquely: title, date, user name, *etc.* The database is identified with either a release number or a date stamp. Following the header is a histogram of the score distribution for the 50 best-matching proteins. In this particular example, scores greater than 68 were reported to be significant. That is, the chance of a random match getting a score of 68 is 1 in 20, (p <0.05). An (optional) overview table provides an animated summary of the results, and is the starting point for "drilling down" into more detailed views. Search results are multidimensional, and cannot be adequately represented by a single, flat table.

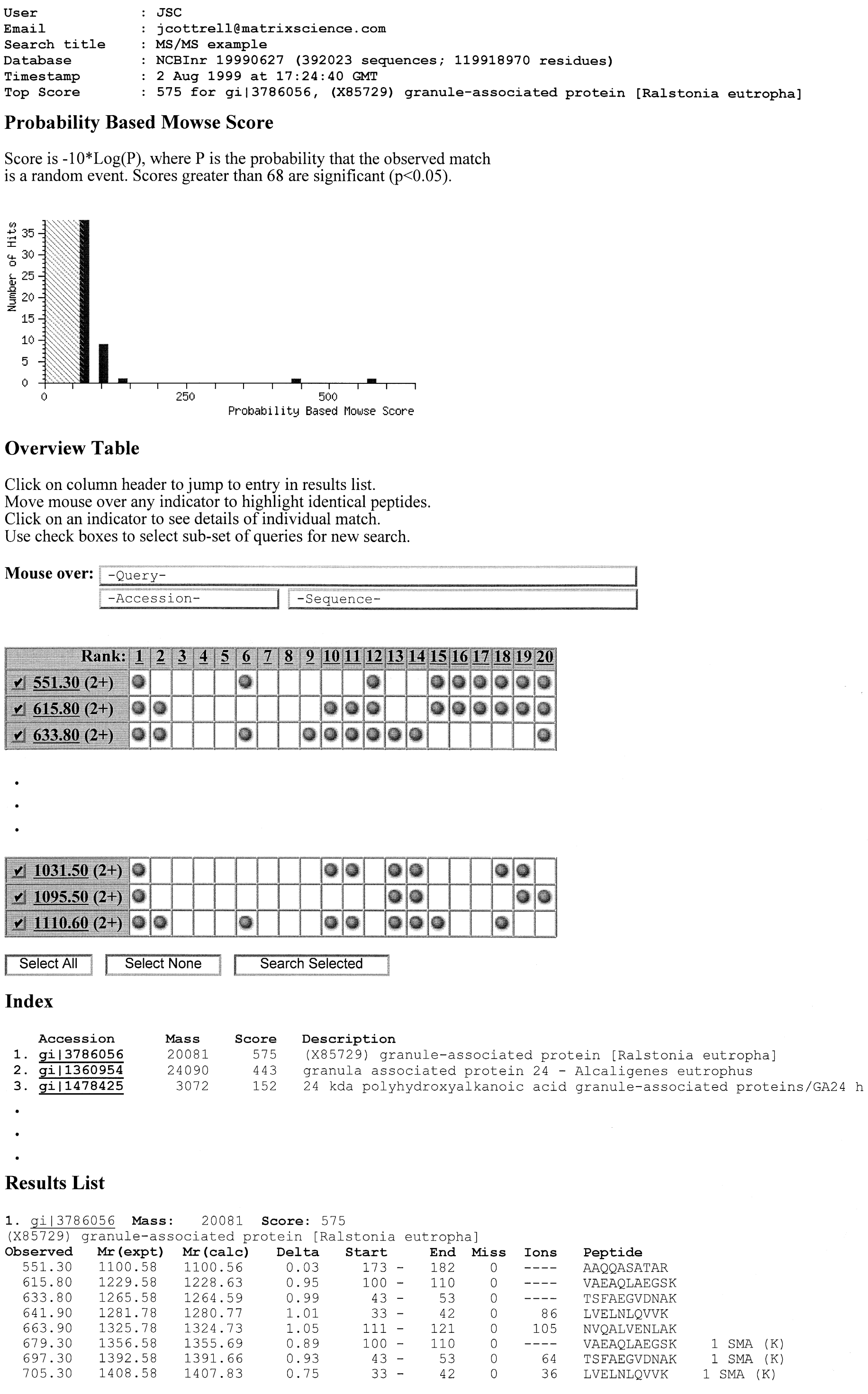
The next section of the result report contains a tabular summary of the matching proteins. For each protein, the first line contains the accession number (linked to the corresponding protein view), the protein molecular mass, and the overall score. This is followed by the FASTA title line, then a table summarising the matched peptides. If a search includes variable modifications, any found in a peptide are listed after the sequence string. At the end of the report, the search parameters are summarised.

By following hyperlinks from the main report page, more detailed reports are available. The Protein View report (Fig. 6) includes the formatted sequence of the protein in 1-letter code with matched peptides highlighted in bold, red type.

If an enzyme has been specified, this is followed by a table of the peptides expected from the digest, including all partials up to the limit specified by the missed cleavages parameter. The matched peptides are shown in bold, red type, together with a link to the corresponding peptide view. If no enzyme was specified, this table contains only the matched peptides. At the bottom of the page, the annotation text from the full database entry is reproduced. The Peptide View report (Fig. 7) for a matched peptide displays a mass spectrum in which all of the matched fragment ions are labelled. The matched fragment ions are also shown in tabular format below the spectrum. As in other views, the sequence is shown in 1-letter code and matched values are highlighted in bold, red type. Only the ion series which were included in the search (*i.e.*, had a non-zero weighting) are included.

3.1 Peptide mass fingerprint

An in-gel tryptic digest of a protein purified from a thermophilic bacterium was analysed by MALDI-TOF-MS. The



```
User           : JSC
Email          : jcottrell@matrixscience.com
Search title   : MS/MS example
Database       : NCBInr 19990627 (392023 sequences; 119918970 residues)
Timestamp      : 2 Aug 1999 at 17:24:40 GMT
Top Score      : 575 for gi|3786056, (X85729) granule-associated protein [Ralstonia eutropha]
```

Probability Based Mowse Score

Score is -10*Log(P), where P is the probability that the observed match is a random event. Scores greater than 68 are significant (p<0.05).

Overview Table

Click on column header to jump to entry in results list.
Move mouse over any indicator to highlight identical peptides.
Click on an indicator to see details of individual match.
Use check boxes to select sub-set of queries for new search.

Mouse over: -Query-
-Accession- -Sequence-

Rank:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
☑ 551.30 (2+)	●					●						●			●	●	●	●	●	●
☑ 615.80 (2+)	●	●								●	●	●			●	●	●	●	●	●
☑ 633.80 (2+)	●	●				●			●	●	●	●	●	●						●

•
•
•

☑ 1031.50 (2+)	●									●	●		●	●				●	●	
☑ 1095.50 (2+)	●												●	●					●	●
☑ 1110.60 (2+)	●	●				●				●	●		●	●	●			●		

Index

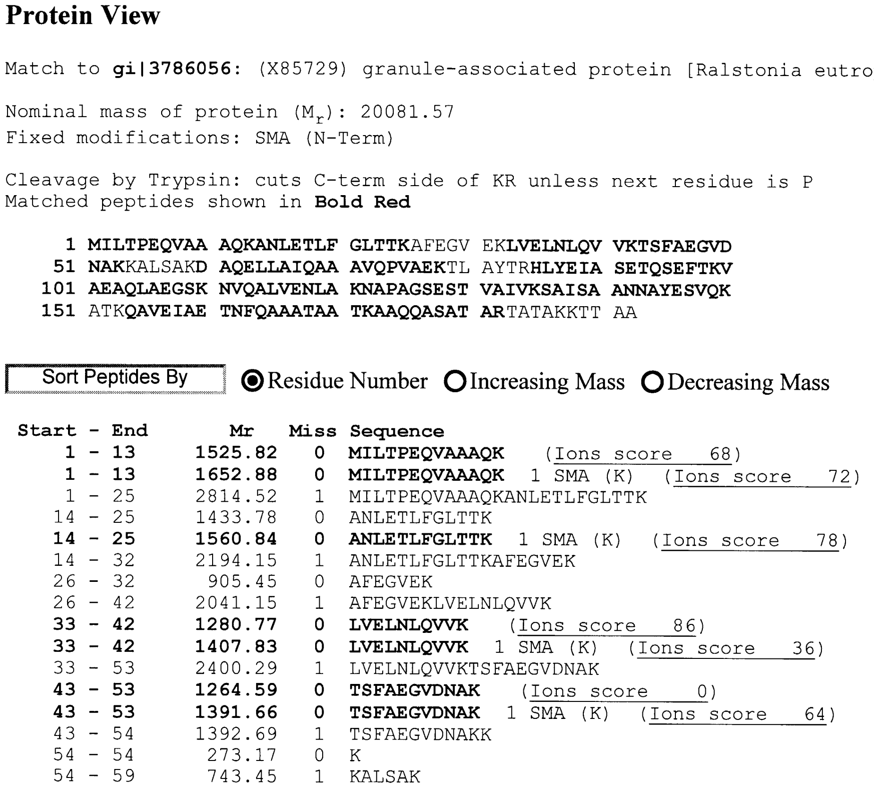
```
    Accession      Mass    Score   Description
 1. gi|3786056     20081     575   (X85729) granule-associated protein [Ralstonia eutropha]
 2. gi|1360954     24090     443   granula associated protein 24 - Alcaligenes eutrophus
 3. gi|1478425      3072     152   24 kda polyhydroxyalkanoic acid granule-associated proteins/GA24 h
 .
 .
 .
```

Results List

```
1. gi|3786056  Mass:   20081  Score: 575
(X85729) granule-associated protein [Ralstonia eutropha]
Observed   Mr(expt)  Mr(calc)    Delta    Start      End  Miss   Ions    Peptide
  551.30   1100.58   1100.56     0.03     173 -     182     0   ----    AAQQASATAR
  615.80   1229.58   1228.63     0.95     100 -     110     0   ----    VAEAQLAEGSK
  633.80   1265.58   1264.59     0.99      43 -      53     0   ----    TSFAEGVDNAK
  641.90   1281.78   1280.77     1.01      33 -      42     0     86    LVELNLQVVK
  663.90   1325.78   1324.73     1.05     111 -     121     0    105    NVQALVENLAK
  679.30   1356.58   1355.69     0.89     100 -     110     0   ----    VAEAQLAEGSK    1 SMA (K)
  697.30   1392.58   1391.66     0.93      43 -      53     0     64    TSFAEGVDNAK    1 SMA (K)
  705.30   1408.58   1407.83     0.75      33 -      42     0     36    LVELNLQVVK    1 SMA (K)
```

Figure 5. Example of the main result report from an MS/MS ions search of 20 peptides from an in-gel tryptic digest of a protein from *Ralstonia eutropha* against the NCBInr database. Peptide mass tolerance ± 2 Da, fragment ion tolerance ± 1 Da, no restriction on protein mass, one missed cleavage allowed. The peptides had been derivatised with SMA.

spectrum, internally calibrated on two trypsin autolysis peaks, is shown in Fig. 8. Peak detection yielded 72 mass values. Searching all 72 values against Owl release 31.3 (290 043 entries), without any restriction on the protein mass, a mass tolerance of ± 200 ppm, and no modifications, matched 36 masses to UVRB_THETH, an excinuclease subunit from *Thermus aquaticus*. Although only half the values have been matched, the score of 359 gives an extremely high level of confidence that this identification is correct. The second best matching protein received a score of 51, well below the significance threshold of 67 for a database of this size. When the identical seach was repeated against a random database of the same size, the highest score was 63. Two subsets of



Protein View

```
Match to gi|3786056: (X85729) granule-associated protein [Ralstonia eutropha]

Nominal mass of protein (Mr): 20081.57
Fixed modifications: SMA (N-Term)

Cleavage by Trypsin: cuts C-term side of KR unless next residue is P
Matched peptides shown in Bold Red

     1 MILTPEQVAA AQKANLETLF GLTTKAFEGV EKLVELNLQV VKTSFAEGVD
    51 NAKKALSAKD AQELLAIQAA AVQPVAEKTL AYTRHLYEIA SETQSEFTKV
   101 AEAQLAEGSK NVQALVENLA KNAPAGSEST VAIVKSAISA ANNAYESVQK
   151 ATKQAVEIAE TNFQAAATAA TKAAQQASAT ARTATAKKTT AA
```

Sort Peptides By ● Residue Number ○ Increasing Mass ○ Decreasing Mass

```
 Start - End       Mr   Miss Sequence
     1 - 13    1525.82   0  MILTPEQVAAAQK    (Ions score   68)
     1 - 13    1652.88   0  MILTPEQVAAAQK  1 SMA (K)   (Ions score   72)
     1 - 25    2814.52   1  MILTPEQVAAAQKANLETLFGLTTK
    14 - 25    1433.78   0  ANLETLFGLTTK
    14 - 25    1560.84   0  ANLETLFGLTTK  1 SMA (K)   (Ions score   78)
    14 - 32    2194.15   1  ANLETLFGLTTKAFEGVEK
    26 - 32     905.45   0  AFEGVEK
    26 - 42    2041.15   1  AFEGVEKLVELNLQVVK
    33 - 42    1280.77   0  LVELNLQVVK    (Ions score   86)
    33 - 42    1407.83   0  LVELNLQVVK  1 SMA (K)   (Ions score   36)
    33 - 53    2400.29   1  LVELNLQVVKTSFAEGVDNAK
    43 - 53    1264.59   0  TSFAEGVDNAK    (Ions score    0)
    43 - 53    1391.66   0  TSFAEGVDNAK  1 SMA (K)   (Ions score   64)
    43 - 54    1392.69   1  TSFAEGVDNAKK
    54 - 54     273.17   0  K
    54 - 59     743.45   1  KALSAK
```

Figure 6. Example of a protein view report (hit 1 from the search shown in Fig. 5)

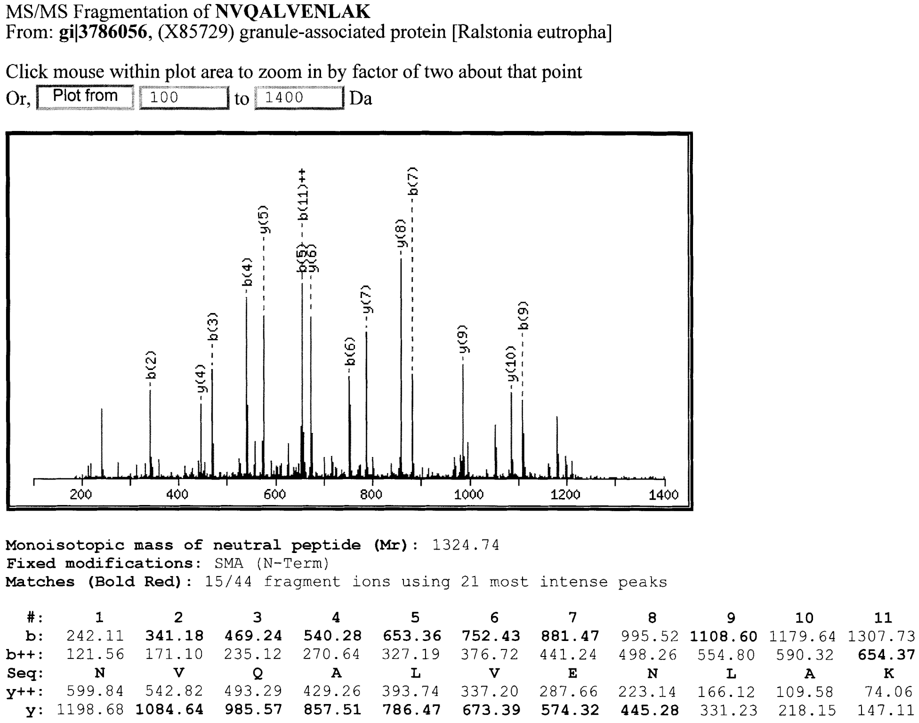
these mass values were selected using the RAND() function in a Microsoft Excel spreadsheet and all three sets (72, 24, and 8 values) searched against Owl using identical conditions. Each search was then repeated several times with progressively wider mass tolerances until the top scoring match was no longer correct.

Figure 9 shows how the score for the UVRB_THETH falls steadily as the full width of the mass tolerance window is increased. For the full set of 72 values, of which approximately half match, the mass tolerance can be opened out to an astonishing ± 0.5% before the score drops into the region of insignificance. The set of 24 values, of which 13 match, maintains a significant score down to ± 0.1%. The set of eight values, of which five match, never achieves a significant score, even at the tightest mass tolerance of ± 0.02%. Note that UVRB_THETH was the highest scoring protein for all the data points plotted in Fig. 9, even though some scores are below the significance threshold. A score of less than the significance threshold does not mean that an answer is wrong, only that one cannot be confident in the identification without additional data or other supporting evidence.

This example illustrates that mass accuracy in a peptide mass fingerprint is not as critical as might be expected, provided that a reasonable number of mass values can be matched. In the context of high throughput proteomics, one should aim to obtain a representative set of mass values at a mass accuracy which can be achieved under routine conditions using external calibration. This will allow the majority of samples to be identified rapidly and inexpensively, conserving resources for more intractable samples. There is no "correct" choice for the significance threshold. In a fully automated context, where the consequences of a false positive are more of a concern than routing a sample through an additional stage of analysis, the threshold should be set high. For Owl 31.3, a one-in-a-million event corresponds to a score of 115. On the other hand, if the protein was known with certainty to be a yeast protein, this would reduce the 5% significance threshold to 54, because there are only some 11 000 yeast entries in Owl 31.3.

3.2 Sequence queries

The ready availability, through the World Wide Web, of the EMBL PeptideSearch package has contributed to the popularity of the Sequence Tag search [9]. This approach relies on being able to interpret a few residues of amino acid sequence from an MS/MS spectrum. In many cases, the combination of peptide mass, interpreted sequence, and the fragment ion masses which enclose the sequence, are sufficient to identify the peptide unambiguously. However, it is difficult to see any fundamental



MS/MS Fragmentation of **NVQALVENLAK**
From: **gi|3786056**, (X85729) granule-associated protein [Ralstonia eutropha]

Click mouse within plot area to zoom in by factor of two about that point
Or, Plot from 100 to 1400 Da

```
Monoisotopic mass of neutral peptide (Mr): 1324.74
Fixed modifications: SMA (N-Term)
Matches (Bold Red): 15/44 fragment ions using 21 most intense peaks

   #:       1       2       3       4       5       6       7       8       9      10      11
   b:  242.11  341.18  469.24  540.28  653.36  752.43  881.47  995.52 1108.60 1179.64 1307.73
 b++:  121.56  171.10  235.12  270.64  327.19  376.72  441.24  498.26  554.80  590.32  654.37
 Seq:       N       V       Q       A       L       V       E       N       L       A       K
 y++:  599.84  542.82  493.29  429.26  393.74  337.20  287.66  223.14  166.12  109.58   74.06
   y: 1198.68 1084.64  985.57  857.51  786.47  673.39  574.32  445.28  331.23  218.15  147.11
```

Figure 7. Example of a peptide view report (*m/z* 663.9 from hit 1 of the search shown in Fig. 5)

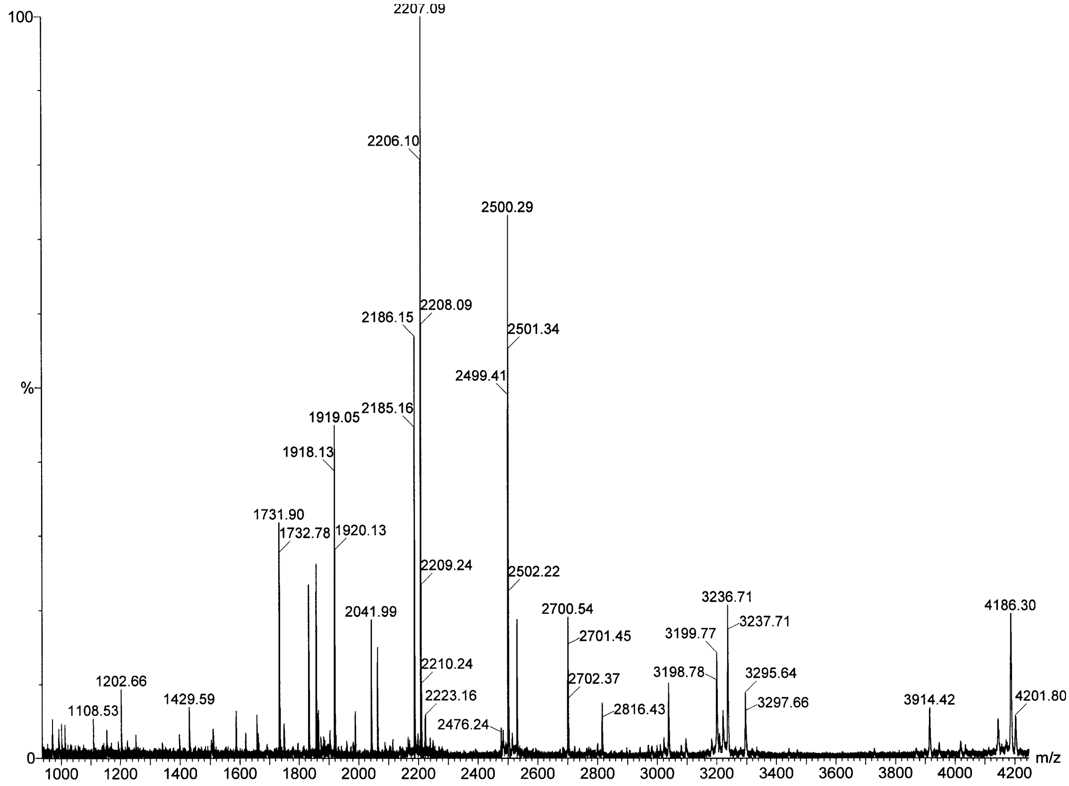


Figure 8. MALDI-TOF spectrum of an in-gel tryptic digest of a protein isolated from a thermophilic bacterium. The spectrum was internally calibrated using two trypsin autolysis peaks at *m/z* 805.42 and *m/z* 2163.06.

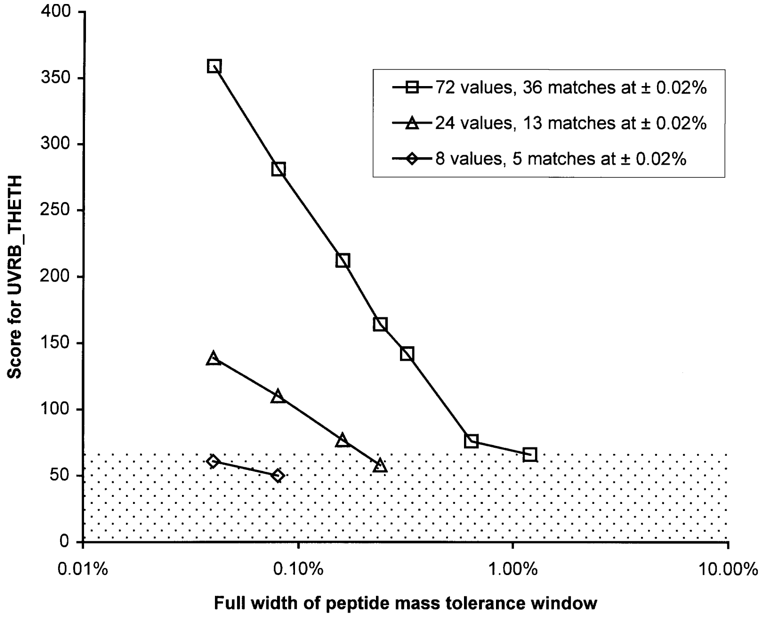


Figure 9. Score behaviour in a series of peptide mass fingerprint searches of data from the spectrum in Fig. 8. Refer to Section 3.1 for full details. Scores falling in the shaded area are not significant.

advantage to performing manual sequence interpretation of this type when it is possible to perform a search using the raw MS/MS data. Although the presence of the sequence data should make the search less computationally intensive, hence faster, the overall throughput is severely constrained by the interpretation step. The other difficulty with the sequence tag approach is that the sequence may be called incorrectly, particularly when data quality is poor.

Perhaps the enduring value of a sequence query is that it enables data from orthogonal techniques to be combined. For example, simply knowing the *N*-terminal residue for each peptide in a peptide mass fingerprint provides a substantial increase in specificity. Such information can be easily obtained by performing a single cycle of manual Edman degradation followed by reanalysis [28]. As an example, a search of four mass values with an error of ± 1 Da from the analysis of a tryptic digest gave a match to an actin, but with a nonsignificant score of 52. A single cycle of manual Edman degradation was performed in an attempt to identify the *N*-terminal residues but, because of the limited mass accuracy, there was ambiguity in two cases: 795.6 seq(n-[LIN]); 976.5 seq(n-A); 1516.9 seq(n-[QEK]); and 1791.2 seq(n-S). (Mass values are average values for singly protonated ions). Nevertheless, this combination of mass and sequence data was sufficient to increase the score to a highly significant 90, confirming that the source protein was indeed actin. This score increase corresponds to a difference in probability of 10^4. To a first approximation, this is because the population of potential matches has been reduced by an order of magnitude for each of the four peptides. Similarly, the ragged end information generated by carboxypeptidase digestion can provide positive identification, even though the short runs of sequence data may be ambiguous or contain gaps [29].

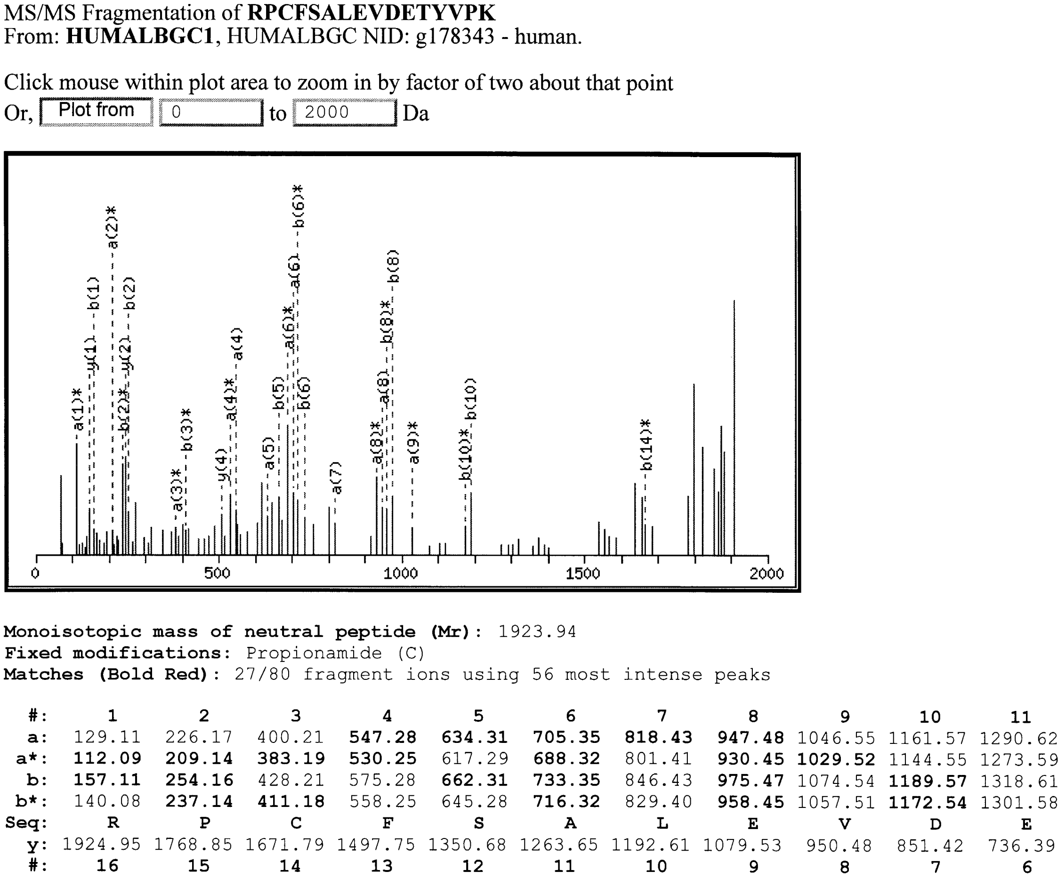
The additional specificity provided by amino acid sequence qualifiers allows the peptide mass tolerance to be relaxed. Searching the average mass value 1854.1 seq(n-TCP) seq(*-DST) seq(c-R) against Owl 31.3 with a ± 25% mass tolerance gives several matches to proteins containing the peptide TCPVQLWVDSTPPPGTR (1854.1 Da), and also three matches to a peptide which differ by a single substitution: TCPVQLWVDSTPPPGSR (1840.1). This pseudohomology search mode can be useful when the analyte protein itself is not present in the database. Amino acid composition can be deduced from the mass shifts following microscale chemical derivatisation [15], or from physical properties such as UV absorbance. The knowledge that a particular amino acid is present or absent significantly increases the specificity of a peptide mass fingerprint [20].

3.3 MS/MS ions search

Although the overwhelming majority of MS/MS fragment ion searches are conducted on electrospray data from triple quadrupole or ion trap instruments [8, 30], the technique is equally applicable to post-source decay (PSD) data from MALDI-TOF instruments [31]. While PSD spectra can sometimes appear “scratchier” than ESI data, excellent matches can still be obtained. Figure 10 is an example of a good match to a peptide from human serum albumin. A score of 99 was obtained by finding 27 matches from 80 calculated fragment ions using the 56 most intense peaks. Note that this does not mean that the remaining peaks are noise. In fact, examination of the data shows that several of the lower intensity peaks correspond to fragment ions. It simply means that matching the 56 most intense peaks gave the lowest probability. In addition to providing estimates for the expected errors on the precursor and fragment ion masses, the user is required to specify which fragment ion series can be expected in the data. As mentioned earlier, searching for too many ion series causes a rapid loss of specificity. In the previous example, the PSD data were searched using the default ion series set (a, a-17, b, b-17, and y). Chemical derivatisation of peptides has long been used to assist in *de novo* sequencing of peptides by directing fragmentation along particular pathways. Derivatisation, which brings a dual benefit of improved S/N and sparser MS/MS spectra, is equally beneficial to database searching. An example of this is illustrated in Figs. 5–7.

An in-gel tryptic digest of a protein purified from intracellular polyhydroxyalkanoic acid granules from *Ralstonia eutropha* was analysed by ESI-MS on a Finnigan MAT LCQ ion trap [32]. Initial PCR and DNA studies had shown that the DNA sequence coding for this protein, phasin, differed significantly from that deposited in Genbank. Analysis of the expressed protein was thus undertaken to confirm these results. One of the DNA sequence differences was a single base insertion, giving rise to a frame shift, resulting in significant alteration of the *C*-terminal region. To facilitate *de novo* sequencing of the affected peptides, SMA derivatisation was used to direct fragmentation into the b and y series.

SMA derivatisation modifies lysine residues and the *N*-terminus. In this particular experiment, derivatisation of *C*-terminal lysine residues with SMA did not go to completion. This may have been due to the influence of buffer salts eluted from the bulk polyacrylamide matrix. Figure 5 shows the result of searching with SMA modification of lysine specified as nonquantitative, and serves to illus-



MS/MS Fragmentation of **RPCFSALEVDETYVPK**
From: **HUMALBGC1**, HUMALBGC NID: g178343 - human.

Click mouse within plot area to zoom in by factor of two about that point
Or, Plot from 0 to 2000 Da

Monoisotopic mass of neutral peptide (Mr): 1923.94
Fixed modifications: Propionamide (C)
Matches (Bold Red): 27/80 fragment ions using 56 most intense peaks

#:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
a:	129.11	226.17	400.21	**547.28**	**634.31**	**705.35**	**818.43**	**947.48**	1046.55	1161.57	1290.62
a*:	**112.09**	**209.14**	**383.19**	**530.25**	617.29	**688.32**	801.41	**930.45**	**1029.52**	1144.55	1273.59
b:	**157.11**	**254.16**	428.21	575.28	**662.31**	**733.35**	846.43	**975.47**	1074.54	**1189.57**	1318.61
b*:	140.08	**237.14**	**411.18**	558.25	645.28	**716.32**	829.40	**958.45**	1057.51	**1172.54**	1301.58
Seq:	**R**	**P**	**C**	**F**	**S**	**A**	**L**	**E**	**V**	**D**	**E**
y:	1924.95	1768.85	1671.79	1497.75	1350.68	1263.65	1192.61	1079.53	950.48	851.42	736.39
#:	**16**	**15**	**14**	**13**	**12**	**11**	**10**	**9**	**8**	**7**	**6**

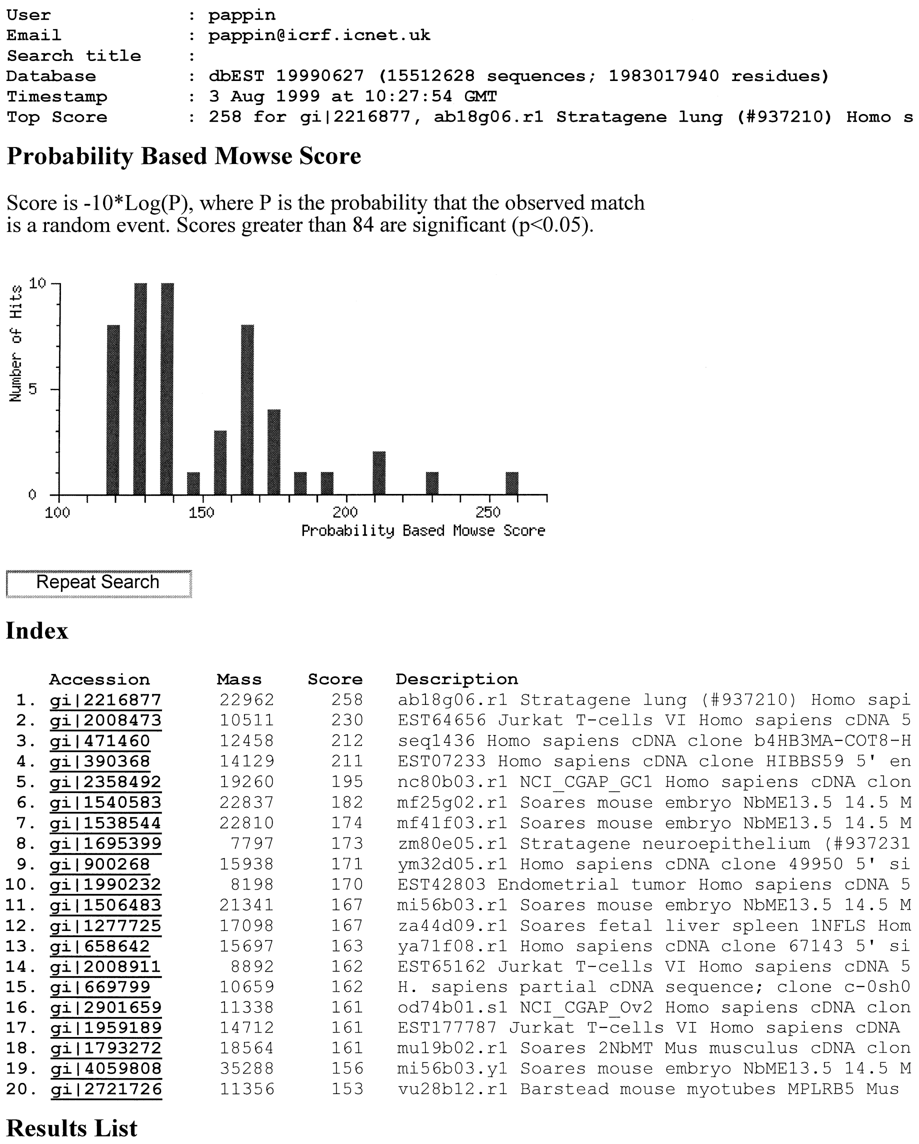
Figure 10. Peptide view of a PSD spectrum which matches to the sequence RPCFSALEVDETYVPK from human serum albumin. Peptide mass tolerance ± 0.15 Da, fragment ion tolerance ± 1 Da, no restriction on protein mass, one missed cleavage allowed. Cysteine was assumed to be derivatised by acrylamide adduction.

trate the ability of the software to handle incomplete chemical modifications in a transparent fashion. Of 20 peptides analysed, 11 contained fully modified lysine groups. Sequence coverage was 83%, as shown in Fig. 6. The effect of the SMA group can be seen in Fig. 7. Complete b and y ion series can be identified, even though the lowest probability match misses a few. In this particular spectrum, doubly charged fragment ions are not abundant; thus a search restricted to just the b and y series would give an increased score.

In general, the sequences in the dbEST database do not correspond to intact proteins. When a number of peptides from a tryptic digest of a protein are searched collectively against dbEST, it is not unusual to see matches to several overlapping clones that, taken together, span the complete protein (see Fig. 11).

Peptides from an in-gel tryptic digest of a human protein were fully derivatised with SMA and analysed by ESI-MS on a Finnigan MAT LCQ ion trap. Parallel searching of 16 peptide MS/MS spectra (submitted as a single search task) against dbEST produced a ranked hit-list containing highly significant matches to several human clones. Using the highest ranked clone as input to the nucleic acid sequence alignment program ESTBlast (http://www.hgmp.mrc.ac.uk/ESTBlast/; R. Gill, personal communication) produced the alignments shown in Fig. 12.

Comparison of the Mascot and ESTBlast results reveals extensive similarity. Because the two reports use different accession numbers, the Mascot hit numbers have been added to Fig. 12 as an additional column. All but one of the clones listed in the ESTBlast report are present in the top 13 Mascot matches. Several additional matches



```
User            : pappin
Email           : pappin@icrf.icnet.uk
Search title    :
Database        : dbEST 19990627 (15512628 sequences; 1983017940 residues)
Timestamp       : 3 Aug 1999 at 10:27:54 GMT
Top Score       : 258 for gi|2216877, ab18g06.r1 Stratagene lung (#937210) Homo s
```

Probability Based Mowse Score

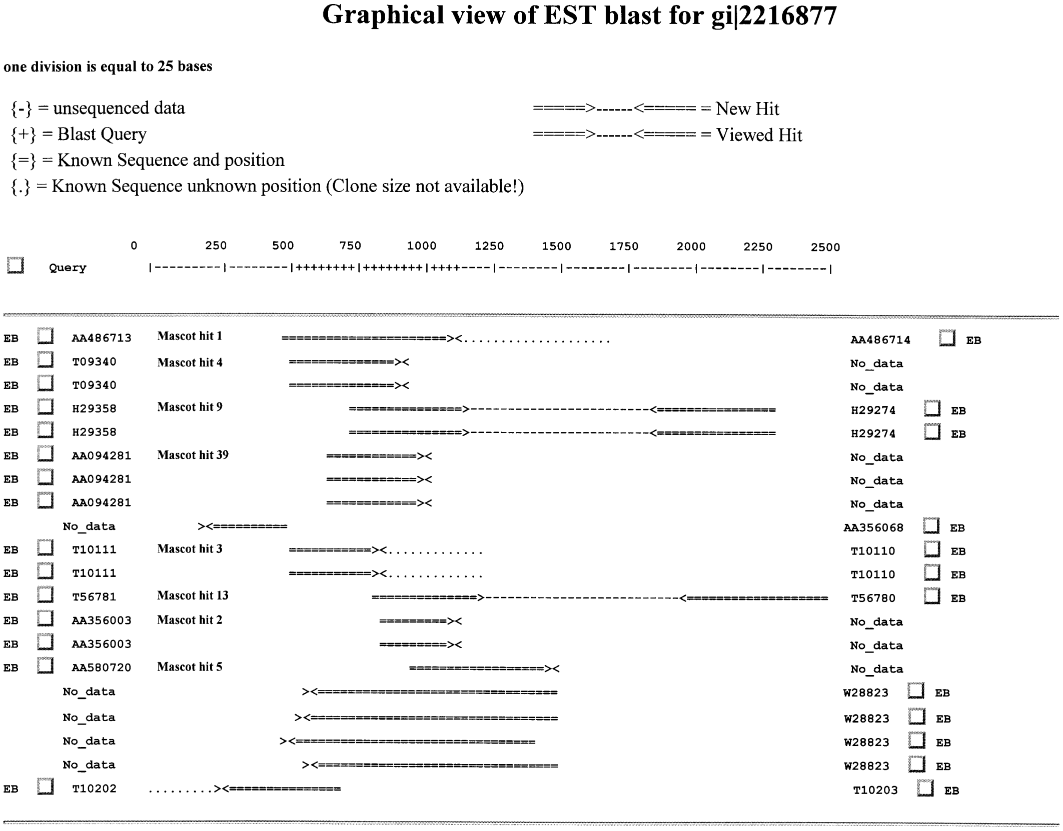
Score is -10*Log(P), where P is the probability that the observed match is a random event. Scores greater than 84 are significant (p<0.05).

Index

```
    Accession      Mass    Score   Description
 1. gi|2216877     22962     258   ab18g06.r1 Stratagene lung (#937210) Homo sapi
 2. gi|2008473     10511     230   EST64656 Jurkat T-cells VI Homo sapiens cDNA 5
 3. gi|471460      12458     212   seq1436 Homo sapiens cDNA clone b4HB3MA-COT8-H
 4. gi|390368      14129     211   EST07233 Homo sapiens cDNA clone HIBBS59 5' en
 5. gi|2358492     19260     195   nc80b03.r1 NCI_CGAP_GC1 Homo sapiens cDNA clon
 6. gi|1540583     22837     182   mf25g02.r1 Soares mouse embryo NbME13.5 14.5 M
 7. gi|1538544     22810     174   mf41f03.r1 Soares mouse embryo NbME13.5 14.5 M
 8. gi|1695399      7797     173   zm80e05.r1 Stratagene neuroepithelium (#937231
 9. gi|900268      15938     171   ym32d05.r1 Homo sapiens cDNA clone 49950 5' si
10. gi|1990232      8198     170   EST42803 Endometrial tumor Homo sapiens cDNA 5
11. gi|1506483     21341     167   mi56b03.r1 Soares mouse embryo NbME13.5 14.5 M
12. gi|1277725     17098     167   za44d09.r1 Soares fetal liver spleen 1NFLS Hom
13. gi|658642      15697     163   ya71f08.r1 Homo sapiens cDNA clone 67143 5' si
14. gi|2008911      8892     162   EST65162 Jurkat T-cells VI Homo sapiens cDNA 5
15. gi|669799      10659     162   H. sapiens partial cDNA sequence; clone c-0sh0
16. gi|2901659     11338     161   od74b01.s1 NCI_CGAP_Ov2 Homo sapiens cDNA clon
17. gi|1959189     14712     161   EST177787 Jurkat T-cells VI Homo sapiens cDNA
18. gi|1793272     18564     161   mu19b02.r1 Soares 2NbMT Mus musculus cDNA clon
19. gi|4059808     35288     156   mi56b03.y1 Soares mouse embryo NbME13.5 14.5 M
20. gi|2721726     11356     153   vu28b12.r1 Barstead mouse myotubes MPLRB5 Mus
```

Results List

Figure 11. The result of searching 16 MS/MS spectra from an in-gel tryptic digest of a human protein against a 6-frame translation of dbEST showing matches across multiple entries. Peptide mass tolerance ± 0.1%, fragment ion tolerance ± 0.5 Da, no restriction on protein mass, one missed cleavage allowed. The peptides had been derivatised with SMA. Oxidation of methionine and acrylamide adduction to cysteine were specified as nonquantitative modifications.



```
                    Graphical view of EST blast for gi|2216877

one division is equal to 25 bases

 {-} = unsequenced data                                   =====>------<===== = New Hit
 {+} = Blast Query                                        =====>------<===== = Viewed Hit
 {=} = Known Sequence and position
 {.} = Known Sequence unknown position (Clone size not available!)

               0        250       500       750      1000      1250      1500      1750      2000      2250      2500
□   Query       |---------|--------|++++++++|++++++++|++++----|--------|--------|--------|--------|--------|

EB □ AA486713  Mascot hit 1       ======================><...................                 AA486714  □ EB
EB □ T09340    Mascot hit 4        ==============><                                           No_data
EB □ T09340                        ==============><                                           No_data
EB □ H29358    Mascot hit 9               ===============>------------------------<===============  H29274  □ EB
EB □ H29358                               ===============>------------------------<===============  H29274  □ EB
EB □ AA094281  Mascot hit 39           ============><                                         No_data
EB □ AA094281                          ============><                                         No_data
EB □ AA094281                          ============><                                         No_data
      No_data              ><==========                                                      AA356068  □ EB
EB □ T10111    Mascot hit 3        ===========><.............                                 T10110  □ EB
EB □ T10111                        ===========><.............                                 T10110  □ EB
EB □ T56781    Mascot hit 13                ==============>-------------------------<==================  T56780  □ EB
EB □ AA356003  Mascot hit 2                  =========><                                      No_data
EB □ AA356003                                =========><                                      No_data
EB □ AA580720  Mascot hit 5                      ==================><                         No_data
      No_data                          ><==================================                  W28823  □ EB
      No_data                         ><===================================                  W28823  □ EB
      No_data                       ><==================================                     W28823  □ EB
      No_data                          ><==================================                  W28823  □ EB
EB □ T10202     .........><===============                                                     T10203  □ EB
```

Figure 12. A sequence alignment report from the program ESTBlast for the highest scoring clone in Fig. 11

appear in the Mascot list because the search was not restricted to human clones. In this example, therefore, the MS-based search was able to identify the majority of overlapping clones required for the subsequent assembly of the entire protein sequence.

3.4 Limitations of the method

3.4.1 Nonindependent experimental data

The bane of peptide mass fingerprint searching has always been false positives. Although the probability-based scoring scheme described here provides a quantitative measure of the significance of a match, it is based on certain assumptions. One of these assumptions is that the experimental data are independent measurements sampled from the population of all possible measurements. If the data are not independent, then the absolute score becomes an unreliable guide to significance, (although the relative scores within a search can still be useful). In fact, it is perfectly possible for a nonindependent data set to get a highly significant score from a search against the randomised database.

Most commonly, the problem is duplicate mass values. In peptide mass fingerprint data, this may be because the mass error window is too large, or the peaks are split by noise or faulty peak picking. In an MS/MS data set, it may be that scans belonging to a single chromatographic peak have been submitted as independent spectra rather than averaged together. Whatever the reason, if there are pairs of duplicate mass values, and one matches, then so does the other. The score, calculated on the basis that these are independent matches, is then too high. Less obviously, data sets without duplicate values can gain significant scores in searches against the random database when nonquantitative modifications are included. If a modification is specified, it is quite likely that the data contain pairs of values which are separated by the mass difference of that modification. If one value of a pair matches a sequence which contains the modifiable residue, then the probability of the other matching the same sequence is much greater than random. Just as in the case of duplicate masses, the experimental data are not independent and the reported score may be too high.

3.4.2 Atypical sequence entries

Some of the entries in the sequence databases exhibit extended repeats, such as AF005273, porcine submaxillary apomucin. Although the molecular mass of this protein is 1.2 MDa, over 80% of the sequence is composed of an identical 7 kDa repeat. It is difficult to know how to treat such cases. If a single experimental peptide mass is allowed to match to multiple calculated masses, then a single experimental mass which matches within a repeat will give a large and meaningless score. But, if duplicate matches are not permitted, it will be virtually impossible to get a match to such a protein because the number of measurable mass values is too small to give a statistically significant score.

3.5 High throughput protein identification

3.5.1 Closed loop automation

Maximum throughput at minimum cost will generally be achieved by "sieving" samples through a series of analytical screening stages of increasing complexity. The prerequisites for this are a scoring scheme which allows rule-based software to decide whether identification has been achieved, and rapid searching to enable real-time decisions to be made. For example, mass analysis of a tryptic digest by MALDI-TOF is rapid, sensitive and inexpensive. In many cases, a peptide fingerprint of the intact digest mixture will be all that is needed to identify a protein. A peptide mass fingerprint takes only a few seconds; it can thus be done in real-time, while the sample is still in the instrument. If the result is not conclusive, there is the possibility to return to the sample and select peptide signals for MS/MS analysis by PSD. While the quality of a PSD spectrum may not be as high as could be achieved by other means, it will often be sufficient to confirm an ambiguous identification.

Proteins which cannot be identified by these means might proceed to be analysed by LC-MS/MS on an electrospray instrument. The achievable throughput of LC-MS/MS is limited by the chromatography stage, so it is not usually cost-effective to use this technique for primary screening. (Faster separations technologies, such as CE or step-gradient elution or short columns will certainly lead to increased throughput. But, continuing advances in MALDI technology seem likely to maintain the differential in throughput). There are two options for handling the flow of data from an LC-MS/MS run. One is to submit the completed data set for searching by Mascot at the end of the run. The other is to submit the accumulating data in real-time, ideally after each new peptide elutes as indicated by the reconstructed ion chromatogram. Although the latter option involves repetitive searching, it actually gives higher throughput because the analysis can be terminated as soon as there are data from sufficient peptides to identify the protein. In practice, this can only be implemented as a closed loop, in which software performs instrument control according to decisions based on search results. Such an arrangement is illustrated schematically in Fig. 13.

3.5.2 Search speed

The preceding paragraphs illustrate the importance of rapid searching. As discussed earlier, there are trade-offs to be made between search speed and the sophistication of the underlying model. Improved search speed could be achieved by precalculating and indexing quantities such as peptide molecular masses. Unfortunately, this would restrict the range of cleavage agents and modifications which could be supported, because a new index would be required for each combination of these parameters. Benchmarks for search speed can be misleading, because small changes in certain parameters can make large differences in the search time. In Mascot, search time is roughly proportional to the number of calculated peptides. Thus, search time increases *pro rata* with both database size and the width of the peptide mass tolerance window. A lower specificity enzyme, such as chymotrypsin, will take longer than a higher one, with a "no-enzyme" search as the worst case.

Quantitative modifications cause no increase in search time because they are simply a shift in the effective residue mass. In contrast, each additional nonquantitative modification causes a geometrical increase in the search time, dependent on the abundance of the affected resi-

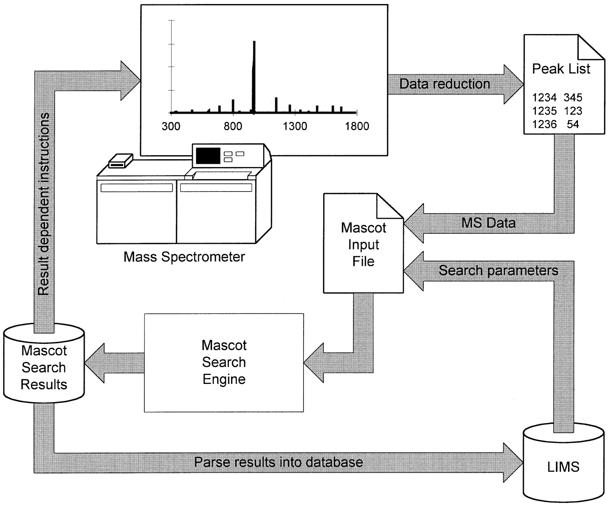


Figure 13. Functional block diagram illustrating closed loop automation for high throughput protein identification

due. For an MS/MS ions search, the search time increases only slowly with the number of peptides. This is because the experimental data can be indexed at the beginning of the search for efficient reference during a single pass through the sequence database. Table 3 contains two search speed benchmarks. One represents a simple peptide mass fingerprint. The other represents a more computationally intensive search of the full dbEST database as of June 27, 1999, including two variable modifications and 16 MS/MS spectra. Provided that the FASTA sequence databases are memory-mapped, searches are processor-bound, and throughput can scale linearly with the number of available processors where the system architecture allows.

4 Concluding remarks

Since the first papers on peptide mass fingerprinting appeared, many authors have observed that the availability of each completed genome would turn protein identification within that species into a bounded problem. Today, the sequencing of the human genome is near completion. Searching genomic data introduces some additional complexity [33]. In the case of a protein database, an algorithm simply searches for the best match between the experimental data and each discrete entry. In the case of genomic sequence data, an algorithm must seek to localise individual peptide matches to a defined region which may correspond to an open reading frame. Since there is a high probability of frame shifts across any extended stretch of genomic sequence data, it becomes essential to correlate matches to the translation in all three frames. We conclude that the use of MS data for protein identification is still at an early stage of development. The introduction of probability-based scoring is a significant step towards the integration of this methodology with purely sequence-based bioinformatics tools, but much remains to be done.

Table 3. Two search speed benchmarks

Example 1	
Platform	1 × 400 MHz Pentium II, Windows NT 4
Search type	Peptide mass fingerprint
Database	NCBInr
Number of entries	~380 000
Input data	8 peptide masses
Peptide mass tolerance	± 0.1%
Enzyme	Trypsin, 0 missed cleavage
Variable modifications	None
Execution time	14 s
Example 2	
Platform	6 × 350 MHz Pentium II, Windows NT 4
Search type	MS/MS ions search
Database	DbEST
Number of entries	~15 500 000 (6 frame translation)
Input data	16 MS/MS spectra
Peptide mass tolerance	± 0.1%
Fragment ion mass tolerance	± 0.5 Da
Enzyme	Trypsin, 1 missed cleavage
Variable modifications	Met oxidation, Cys propionamide
Execution time	4 min 44 s

DNP and DJCP were supported by the Imperial Cancer Research Fund.

Received August 27, 1999

5 References

[1] Blackstock, W. P., Weir, M. P., *Trends Biotechnol.* 1999, *17*, 121–127.

[2] Yates III, J. R., *Electrophoresis* 1998, *19*, 893–900.

[3] Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., Watanabe, C., *Proc. Natl. Acad. Sci. USA* 1993, *90*, 5011–5015.

[4] James, P., Quadroni, M., Carafoli, E., Gonnet, G., *Biochem. Biophys. Res. Commun.* 1993, *195*, 58–64.

[5] Mann, M., Hojrup, P., Roepstorff, P., *Biol. Mass Spectrom.* 1993, *22*, 338–345.

[6] Pappin, D. J. C., Hojrup, P., Bleasby, A. J., *Curr. Biol.* 1993, *3*, 327–332.

[7] Yates III, J. R., Speicher, S., Griffin, P. R., Hunkapiller, T., *Anal. Biochem.* 1993, *214*, 397–408.

[8] Yates III, J. R., Eng, J. K., McCormack, A. L., Schieltz, D., *Anal. Chem.* 1995, *67*, 1426–1436.

[9] Mann, M., Wilm, M., *Anal. Chem.* 1994, *66*, 4390–4399.

[10] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal. Chem.* 1996, *68*, 850–858.

[11] Sherman, N. E., Yates, N. A., Shabanowitz, J., Hunt, D. F., Jeffery, W., Bartlet-Jones, M., Pappin, D. J. C., *Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Allied Topics*, Atlanta, GA, May 21–26, 1995, p. 626.

[12] Hoss, M., Robins, P., Naven, T. J. P., Pappin, D. J. C., Sgouros, J., Lindahl, T., *EMBO J.* 1999, *18*, 3868–3875.

[13] Hunt, D. F., Yates III, J. R., Shabanowitz, J., Winston, S., Hauer, C. R., *Proc. Natl. Acad. Sci. USA* 1986, *83*, 6233–6237.

[14] Wilm, M., Mann, M., *Anal. Chem.* 1996, *68*, 1–8.

[15] Pappin, D. J. C., Rahman, D., Hansen, H. F., Bartlet-Jones, M., Jeffery, W., Bleasby, A. J., in: Burlingame, A. L., Carr, S. A. (Eds.), *Chemistry Mass Spectrometry and Peptide-Mass Databases: Evolution of Methods for the Rapid Identification and Mapping of Cellular Proteins*, Humana, Totowa, NJ 1996, pp. 135–150.

[16] Pearson, W. R., *Protein Sci.* 1995, *4*, 1145–1160.

[17] Fitch, W. M., *J. Mol. Biol.* 1983, *163*, 171–176.

[18] Wilkins, M. R., Gasteiger, E., Wheeler, C. H., Lindskog, I., Sanchez, J.-C., Bairoch, A., Appel, R. D., Dunn, M. J., Hochstrasser, D. F., *Electrophoresis* 1998, *19*, 3199–3206.

[19] Clauser, K. R., Hall, S. C., Smith, D. M., Webb, J. W., Andrews, L. E., Tran, H. M., Epstein, L. B., Burlingame, A. L., *Proc. Natl. Acad. Sci. USA* 1995, *92*, 5072–5076.

[20] Fenyo, D., Qin, J., Chait, B. T., *Electrophoresis* 1998, *19*, 998–1005.

[21] Keil, B., *Specificity of Proteolysis*, Springer Verlag, Berlin 1992.

[22] Pappin, D. J. C., Rahman, D., Hansen, H. F., Jeffery, W., Sutton, C. W., in: Atassi, M. Z., Appella, E. (Eds.), *Peptide-Mass Fingerprinting as a Tool for the Rapid Identification and Mapping of Cellular Proteins*, Plenum, New York, NY 1995, pp. 161–173.

[23] Mann, M, *Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Allied Topics*, Atlanta, GA, May 21–26, 1995 p. 639.

[24] Bleasby, A. J., Wootton, J. C., *Protein Engineer.* 1990, *3*, 153–159.

[25] Roepstorff, P., Fohlman, J., *Biomed. Mass Spectrom.* 1984, *11*, 601.

[26] Papayannopoulos, I. A., *Mass Spectrom. Rev.* 1995, *14*, 49–73.

[27] Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R., Ventner, J. C., *Science* 1991, *252*, 1651–1656.

[28] Bradley, C. V., Williams, D. H., Hanley, M. R., *Biochem. Biophys. Res. Commun.* 1982, *104*, 1223–1230.

[29] Korostensky, C., Staudenmann, W., Dainese, P., Hoving, S., Gonnet, G., James, P., *Electrophoresis* 1998, *19*, 1933–1940.

[30] Yates III, J. R., Eng, J. K., McCormack, A. L., *Anal. Chem.* 1995, *67*, 3202–3210.

[31] Griffin, P. R., MacCoss, M. J., Eng, J, K., Blevins, R. A., Aaronson, J. S., Yates III, J. R., *Rapid Commun. Mass Spectrom.* 1995, *9*, 1546–1551.

[32] Hanley, S. Z., Pappin, D. J., Rahman, D., White, A. J., Elborough, K. M., Slabas, A. R., *FEBS Lett.* 1999, *447*, 99–105.

[33] Küster, B., Mortensen, P., Mann, M., *Proceedings of the 47th ASMS Conference on Mass Spectrometry and Allied Topics*, Dallas, TX, June 13–17, 1999, pp. 1897–1898.