JOURNAL OF PROTEOMICS 74 (2011) 1842–1851

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Tutorial

Protein identification using MS/MS data☆

*John S. Cottrell**

Matrix Science Ltd., London, UK

ARTICLE INFO

Article history:
Received 2 February 2011
Accepted 9 May 2011
Available online 15 May 2011

Keywords:
Protein identification
Database search
Protein inference
Tutorial

ABSTRACT

The subject of this tutorial is protein identification and characterisation by database searching of MS/MS Data. Peptide Mass Fingerprinting is excluded because it is covered in a separate tutorial. Practical aspects of database searching are emphasised, such as choice of sequence database, effect of mass tolerance, and how to identify post-translational modifications. The relationship between sensitivity and specificity is discussed, as is the challenge of using peptide match information to infer which proteins were present in the sample.
Since these tutorials are introductory in nature, most references are to reviews, rather than primary research papers. Some familiarity with mass spectrometry and protein chemistry is assumed. There is an accompanying slide presentation, including speaker notes, and a collection of web-based, practical exercises, designed to reinforce key points. This Tutorial is part of the *International Proteomics Tutorial Programme* (IPTP 6).

© 2011 Elsevier B.V. Open access under CC BY-NC-ND license.

1. Overview

Fig. 1 illustrates a typical experimental workflow for protein identification and characterisation using MS/MS data. The starting point is a protein sample, which may be a single protein or a complex mixture of proteins. An enzyme, often trypsin, digests the proteins to peptides. In most cases, one or more stages of chromatography are used to regulate the flow of peptides into the mass spectrometer. Peptides are selected one at a time using the first stage of mass analysis. Each isolated peptide is then induced to fragment, possibly by collision, and the second stage of mass analysis used to capture an MS/MS spectrum.

For each MS/MS spectrum, software is used to determine which peptide sequence in a database of protein or nucleic acid sequences gives the best match. Each entry in the database is digested, in silico, using the known specificity of the enzyme, and the masses of the intact peptides calculated. If the calculated mass of a peptide matches that of an observed peptide, the masses of the expected fragment ions are calculated and compared with the experimental values. Some search engines also predict and compare the relative intensities of the fragment ions [1]. Many scoring algorithms have been devised to decide which peptide sequence best matches a given spectrum.

Because the data in each MS/MS spectrum correspond to an isolated peptide, it makes no difference whether the original sample was a single protein or a mixture. Individual peptide sequences are identified, then the set of peptide sequences is used to infer which proteins may have been present. Unless a peptide is unique to one particular protein, there may be some ambiguity as to which protein it originated from.

There are many variations of this workflow. 1D or 2D gel electrophoresis may be used for separation followed by a single stage of chromatography. In Top-down proteomics (matching MS/MS spectra of intact proteins [2]), or if the sample was of endogenous peptides (sometimes called peptidomics [3]), there would be no enzyme digest step.

☆ This Tutorial is part of the *International Proteomics Tutorial Programme* (IPTP 6). Details can be found at: http://www.proteomicstutorials.org/.

* *Matrix Science Ltd.*, 64 Baker Street, London W1U 7GB, UK. Tel.: +44 20 7486 1050; fax: +44 20 7224 1344.
E-mail address: jcottrell@matrixscience.com.

1874-3919 © 2011 Elsevier B.V. Open access under CC BY-NC-ND license.
doi:10.1016/j.jprot.2011.05.014

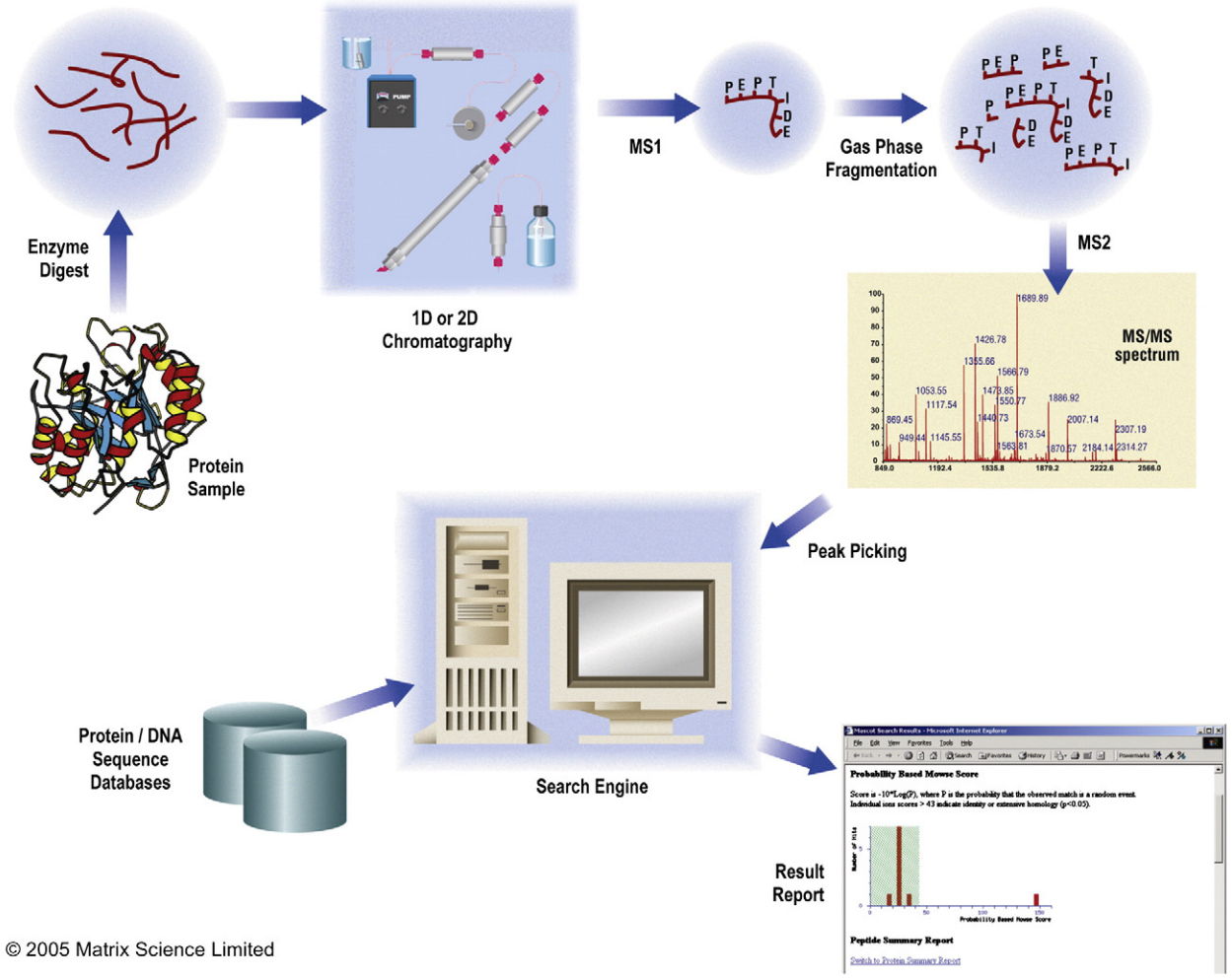

Fig. 1 – A typical experimental workflow for protein identification and characterisation using MS/MS data.

Database matching of MS/MS data is possible because peptide molecular ions fragment preferentially at certain points along the backbone [4,5]. For many instruments, the major peaks in an MS/MS spectrum are b ions, where the charge is retained on the N-terminus, or y ions, where the charge is retained on the C-terminus [6]. The dominant fragmentation pathways depend on the ionisation technique, the mass analyser, and the peptide structure. Electron transfer dissociation, for example, produces predominantly c and z ions [7].

Peptide fragmentation is rarely a clean process, and the spectrum will often show significant peaks from side chain cleavages and internal fragments, where the backbone has been cleaved twice. This creates ambiguity and makes it difficult to 'read off' the sequence. Worse, there may be no recognisable fragments at all to define some parts of the sequence. So, although de novo sequence interpretation can be used in favourable cases [8], the only option for many MS/MS spectra is database searching, where we settle for the best match in a limited pool of sequences. There may be no sequence ions from one half of the peptide, but if there is a sequence in the database that gives a good match to the other half, and the molecular mass fits, we accept this as the correct sequence.

2. Sequence tag searches

The first approach to database searching with MS/MS data was the sequence tag, originally described by Mann and Wilm [9]. Whilst the quality of a typical MS/MS spectrum is not good enough for 'end to end' de novo interpretation, it is often possible to pick out a short sequence ladder, and read off three or four residues of sequence. In a sequence homology search, this would be worth almost nothing, since any given triplet can be expected to occur by chance many times in even a small database. What Mann and Wilm recognised was that a short stretch of amino acid sequence might provide sufficient specificity for unambiguous identification if it was combined with the fragment ion mass values that enclose it, the peptide mass, and the enzyme specificity.

Picking out an accurate tag is not trivial, and requires both skill and experience (or sophisticated software). Fig. 2 shows a promising four residue tag. Each search engine has its own syntax for the mass and sequence information. The example shown here is for Mascot, where the masses are entered as observed m/z values. Searching this tag with trypsin as the enzyme and a mass tolerance of 0.5 Da gives a match to the

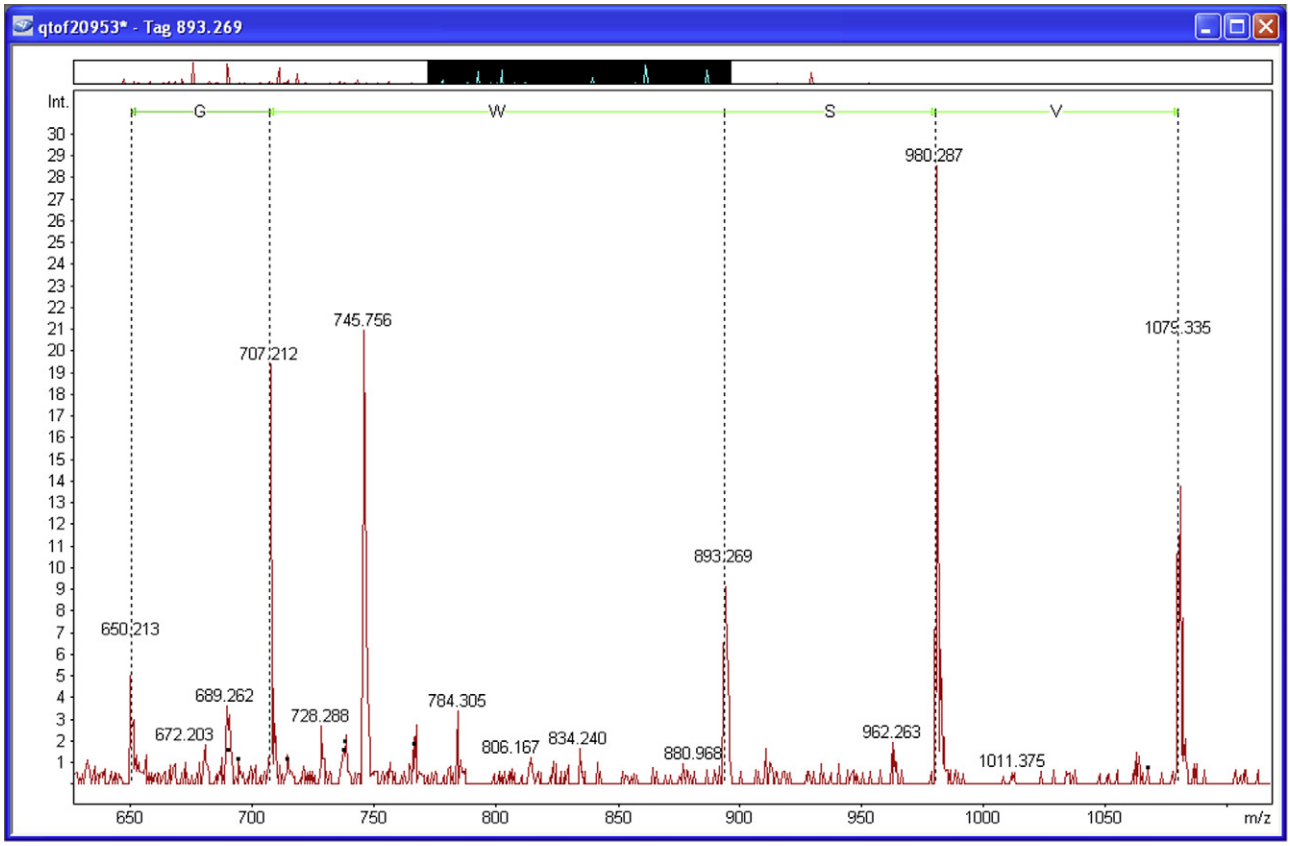

Fig. 2 – An example of a sequence tag. Each search engine has its own syntax for the mass and sequence information. The example shown here is for Mascot, where the masses are entered as observed m/z values.

trypsin autolysis peptide LQGIVSWGSGCAQK. The peaks defining the tag turn out to be y ions, which is why the sequence reads from right to left.

If the tag is not called correctly, then no match will be found. With some search engines, ambiguity is OK, as long as it is recognised and the query is formulated correctly. Obviously, I=L and, unless the mass accuracy is high, Q=K and F=MetOx. Software or a table of mass values can help identify the more common ambiguities.

These days, the standard sequence tag search is essentially obsolete. It is much easier to skip the interpretation stage and search the MS/MS peak list directly, as will be described below. The reason the sequence tag is still important is its so-called 'error tolerant' mode. This consists of relaxing the specificity, usually by removing the peptide molecular mass constraint. When this is done, the tag becomes disconnected from one terminus, so that a match is possible even if there is a mass difference to one side or the other of the tag. This is one of the few ways of getting a database match to a peptide when there is a truly unknown modification or a variation in the primary amino acid sequence.

Other algorithms that combine mass and sequence information in novel ways include OpenSea [10], MS-BLAST [11], GutenTag [12], and MultiTag [13].

3. Searching uninterpreted MS/MS data

The more widespread approach to database searching of MS/MS data is to skip the interpretation step and let the search engine try to match calculated mass values directly. This method was pioneered by John Yates and Jimmy Eng at the University of Washington, Seattle, who used a cross correlation algorithm to compare an experimental MS/MS spectrum against spectra predicted from candidate peptide sequences. Their ideas were implemented as the Sequest program [14].

There are now many search engines on the web for performing searches of uninterpreted MS/MS data. Fig. 3 lists those available in late 2010 together with some of the free and commercial packages that can be downloaded or purchased to run locally.

Searching of uninterpreted MS/MS data is readily automated for high throughput work, and most 'proteomics pipelines' use this approach. A sample may generate tens or even hundreds of thousands of MS/MS spectra, which can be searched as a single data set. It also offers the possibility of getting useful matches from spectra of marginal quality, from which it would be difficult to call a reliable sequence tag. In

InsPect	http://proteomics.ucsd.edu/LiveSearch/
Mascot	http://www.matrixscience.com/search_form_select.html
MS-Tag (Protein Prospector)	http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=mstagstandard
OMSSA	http://pubchem.ncbi.nlm.nih.gov/omssa/index.htm
PepProbe	http://bart.scripps.edu/public/search/pep_probe/search.jsp
Phenyx	http://phenyx.vital-it.ch/pwi/login/login.jsp
Popitam	http://www.expasy.org/tools/popitam/
RAId_DbS	http://www.ncbi.nlm.nih.gov/CBBResearch/qmbp/RAId_DbS/index.html
Sonar	http://hs2.proteome.ca/prowl/knexus.html
X!Tandem (The GPM)	http://thegpm.org/TANDEM/index.html
Not on-line	ByOnic, Crux, MassMatrix, MyriMatch, Paragon, PepSplice, ProLuCID, ProbID, ProteinLynx GS, SIMS, Sequest, SpectrumMill, greylag, pFind, etc.

Fig. 3 – Search engines for uninterpreted MS/MS data.

isolation, a weak match to a single spectrum might not be worth much, but if several spectra match peptides from the same protein, this may give a higher degree of confidence that the protein is truly present. This is usually only practical for small data sets, where the probability of getting multiple peptide matches to the same protein by chance is very low.

Searches of large data sets can be slow, particularly if searched without enzyme specificity or with many variable modifications. It may be necessary to spread the workload across multiple processors in a computer cluster or grid.

The remainder of this tutorial focuses on key aspects of searching uninterpreted MS/MS data: search parameters, site localisation of modifications, multi-pass searches to locate unsuspected modifications and non-specific cleavage, scoring, the trade-off between sensitivity and specificity, and protein inference.

4. Search parameters

Besides the mass spectrometry data, all search engines require additional information in the form of 'search parameters'. Some try to keep the user interface simple, and ask for few parameters, others aim for greater flexibility and accept some additional complexity. This section looks at the more important search parameters.

4.1. *Sequence database*

There are two major repositories for sequence databases, one hosted by the National Center for Biotechnology Information in the USA (http://www.ncbi.nlm.nih.gov/) and the other hosted by the European Bioinformatics Institute (http://www.ebi.ac.uk/). All search engines use the ubiquitous Fasta format and all support searching of databases of protein sequences. Some also support searching DNA sequences, which may be sequences corresponding to proteins (mRNA), expressed sequence tags (EST), or the genomic DNA sequence for a particular organism. Nucleic Acids Research produces an annual databases issue, containing articles describing the major databases (http://dx.doi.org/10.1093/nar/gkp1077).

SwissProt is a high quality, curated, non-redundant protein database. That is, it contains a consensus sequence for each distinct protein, and known variants have been collapsed into a single entry. This makes it relatively small, so searches are fast and reports are concise. However, if a protein of interest is present in the sample at a very low level, and only represented in the MS/MS data by one or two spectra, there is a risk that these critical sequences could be missing from a non-redundant database. The alternative is a comprehensive, non-identical database, where every known protein sequence is explicitly represented, such as NCBInr or UniRef100. The penalty is that the database becomes larger by a factor of approximately 20. A search will take 20 times as long and protein inference becomes more difficult.

EST databases can be very large and very redundant. They are worth trying with high quality MS/MS data if a good match could not be found in a protein database or if studying an organism that is not well represented in the protein databases.

The genomes of an increasing number of organisms have been sequenced. But, in most cases, as soon as a genome sequence is published, the proteins corresponding to the coding sequences are added to NCBInr and UniRef100. So, the main motivation for searching a genomic DNA sequence would be to find matches to sequences from coding sequences missed by the gene finding software. Unfortunately, the genomes of higher organisms have an exon/intron structure, which causes many potential matches to be lost because the peptide sequence is broken across two exons [15].

To illustrate, Fig. 4 summarises the search results for a public domain dataset (ABRF Proteome Informatics Research Group Study iPRG2010). The data, from human proteins, were searched using Mascot against a curated protein database (IPI human), the human sequences of the GenBank EST division, and the human genome assembly. Search conditions were as shown. The table shows the number of peptide matches obtained at a 1% false discovery rate (to be explained in greater detail below) and the average score threshold used to get 1% FDR.

EST_human gets far fewer matches than IPI_human. The main reason is that EST_human is a much larger database, by more than a factor of 100. This means that score thresholds are higher, so we lose all the weaker matches that had scores between 36 and 60. There may be additional matches in EST_human, but the net change is to lose many matches.

The human genome results are even worse. This is not because of a higher threshold; the databases are very similar in size. One reason is that a proportion of potential matches are missed because they are split across exon–intron boundaries. Based on average tryptic peptide length, approximately 20% of matches are lost for this reason. In this particular example, the difference is much larger than 20%. The other factor is that the human genome is only 1.5% coding sequence, and represents a single consensus genome. EST_human is 100% coding sequence and represents a wide range of SNPs and variants.

To search a nucleic acid database, a utility can be used to translate the sequences to amino acids prior to the search or the search engine can translate automatically, in the course of the search. Usually, this is a 6 frame translation, 3 reading frames from the forward strand and 3 reading frames from the

Type of search	: MS/MS Ion Search
Enzyme	: Trypsin/P
Fixed modifications	: Carbamidomethyl (C)
Variable modifications	: Acetyl (N-term), Phospho (Y), Phospho (ST), Oxidation (M), Gln->pyro-Glu (N-term Q)
Mass values	: Monoisotopic
Protein mass	: Unrestricted
Peptide mass tolerance	: ± 10 ppm (# ^{13}C = 1)
Fragment mass tolerance	: ± 0.6 Da
Max missed cleavages	: 1
Instrument type	: ESI-TRAP
Number of queries	: 8,797

Database	Size	Avg. 1% threshold	# matches @ 1% FDR
IPI_human 3.66	3.5 x 10^7 residues	36	2961
EST_human 20100415	4.2 x 10^9 bases	60	1899
Human Genome 20060306	3.1 x 10^9 bases	60	1241

Fig. 4 – Results for a public domain dataset searched using Mascot against a curated protein database (IPI human), the human sequences of the GenBank EST division, and the human genome assembly. Search conditions were as shown.

complementary strand. This is partly because the strand and frame are often uncertain and partly to catch frame shifts. All organisms do not share the same genetic code. The differences may not be great, but it is worth using the correct code if the taxonomy is known.

4.2. *Taxonomy*

If a database contains taxonomy information, most search engines can use this to restrict the search to entries for a particular organism or taxonomic rank. This speeds up the search because, in effect, it makes the database smaller. Limiting the taxonomy also simplifies the result report, because it removes homologous proteins from other species.

It isn't a good idea to specify a very narrow taxonomy in a search. If the correct protein from the correct species is not in the database, it can be very helpful to see a match to a protein from a similar species. This is especially important for poorly represented species. For example, in Swiss-Prot 2010_08, there are 25,000 entries for rodents, all but 1500 of which are either mouse or rat. So, even if you are studying hamster or porcupine, choose 'Rodentia' or something broader, not 'Other rodentia'.

If using a narrow taxonomy or a single organism database, remember to include sequences from common contaminants. Otherwise, you may get misleading matches such as human serum albumin when it should be BSA or mouse keratin when it should be human.

4.3. *Mass tolerance*

Most search engines support separate mass tolerances for precursors and fragments. Hybrid instruments, in particular, may have very different accuracies for MS and MS/MS.

Precursor m/z comes from peak picking the (MS) survey scan. Sometimes, the ^{13}C peak may be selected rather than the ^{12}C. In extreme cases, the $^{13}C_2$ peak may be taken. This can happen even when the underlying accuracy and resolution are very high. Some search engines allow for this and will look at the correct mass, plus and minus the tolerance, and also in narrow windows 1 Da and 2 Da higher. This is preferable to opening out the precursor mass tolerance to 1 or 2 Da.

Specifying too tight a mass tolerance is a very common reason for failing to get a match. Unless the search engine performs some type of re-calibration [16,17], it is mass accuracy that matters, not precision. Making an estimate of the mass accuracy doesn't have to be a guessing game. Search result reports usually include graphs showing the mass errors for matched peptides as a function of mass. Search a standard that gives many strong matches, such as a BSA digest, and look at these error graphs. The extent of random scatter and any systematic trend will become clear. Add on a safety margin and this is your error estimate.

If you have very high precursor accuracy, beware of searching a small database. The combination of low ppm mass tolerance, tryptic cleavage specificity, and a database limited to proteins from a single taxon can mean that there is only a single candidate peptide sequence for many of the MS/MS spectra. It is difficult to judge the reliability of such matches, when there are few or no alternatives to compare against. Better to search a larger database or open out the mass tolerance so as to ensure each spectrum is tested against a diverse range of candidate sequences.

4.4. *Enzyme*

If the peptides result from an enzyme digest, you need to know what the enzyme was and select it in the search form.

Setting the number of allowed missed cleavage sites to zero simulates a limit digest. If you are confident that the digest was perfect, with no partial fragments present, this will give maximum discrimination.

If experience shows that the digest mixtures usually include some partials (peptides with missed cleavage sites) you should choose a setting of 1, or maybe 2 missed cleavage sites. Don't specify a higher number without good reason, because each additional level of missed cleavages increases the number of calculated peptide masses to be matched against the experimental data. Just like mass tolerances, the missed cleavage parameter is best set by looking at some successful search results to see how complete the digests actually are.

Some people like to perform searches without enzyme specificity, then gain confidence that a match is correct if the matched peptides are tryptic. The downside is that this makes the search space 100 to 1000 times larger, so that many weak matches will be lost. If there is evidence for a lot of non-specific cleavage, which may actually be in-source fragmentation, then a semi-specific enzyme allows one end of the peptide to be non-specific, but not both. Only abandon enzyme specificity completely if you have no other choice, such as when searching endogenous peptides.

4.5. *Peak list file format*

There are a number of different file formats for peak lists. DTA and PKL were developed for Sequest and Waters Masslynx respectively, and are relatively simple, containing little more than mass and intensity pairs. MGF is the Mascot Generic Format, which also supports embedded search parameters and meta data. mzML is the standard interchange format controlled by the Proteomics Standards Initiative [18]. It can be used for either raw data or peak lists and replaces both mzData, which was mainly for peak lists, and mzXML, which was mainly for raw data.

4.6. *Modifications*

Proteins and peptides can be modified in hundreds of ways [19]. Some modifications relate to biological function, others are artefacts of sample handling. The most comprehensive databases of protein modifications are Unimod (http://www.unimod.org), which focuses on modifications relevant to mass spectrometry, and RESID (http://www.ebi.ac.uk/RESID/), which concentrates on natural modifications, mostly post-translational.

In database searching, modifications are handled in two ways. First, there are the quantitative modifications, usually called fixed or static. An example would be the efficient alkylation of cysteine. Since all cysteines are modified, this is effectively just a change in the mass of cysteine. It carries no penalty in terms of search speed or specificity.

In contrast, most post-translational modifications do not apply to all instances of a residue. For example, phosphorylation might affect just one serine in a peptide containing many serines. Non-quantitative modifications, usually called variable or differential, are expensive in the sense that they increase the time taken for a search and reduce its specificity. This is because the software has to permute out all the possible arrangements of modified and unmodified residues that fit to the peptide molecular mass. As more and more modifications are considered, the number of combinations and permutations increases geometrically; a so-called combinatorial explosion. This is why it is very important to be as sparing as possible with variable modifications.

A third class of modifications is sometimes used for stable isotope labels in a quantitation experiment. Some peptides will carry a light label (or no label) and other peptides will carry a heavy label, but no peptide ever carries a mixture of light and heavy. To keep the search space small, this can be implemented as two separate fixed modification searches, one each for the light and heavy labels.

5. Site analysis

For modifications that relate to biological function, the site of a modification can be just as important as the nature of the modification. If a peptide contains both modified and unmodified sites, identifying the presence of a modification is not the same as localising it to a particular residue. A standard search result report may list only the highest scoring arrangement, or it may list several arrangements with differing scores, but leaving interpretation to the user. Many software tools have been developed to try and quantify site localisation, often with a particular focus on phosphorylation: Ascore [20], MaxQuant [21], InsPect [22], MS-Alignment [23], PTMFinder [24], PhosphoScore [25], Debunker [26], SloMo — ETD/ECD [27], ModifiComb [28], and Delta Score [29].

6. Multi-pass searches

Multi-pass searching is the efficient way to find less common modifications, including point mutations in the primary sequence, and non-specific peptides. The first pass search is a simple search of the entire database with minimal modifications. The protein hits found in the first pass search are then selected for an exhaustive second pass search. During this second pass search, a long list of potential modifications is tested serially. That is, only peptides containing a single unsuspected modification will be matched, but this will cover most post-translational modifications. Single residue substitutions can be handled in exactly the same way as modifications, and the enzyme specificity will often be relaxed so as to find matches to non-specific peptides.

Because only a handful of entries are being searched, search time is not an issue. It is difficult to apply any kind of statistical treatment to the results, because the entries being searched have been pre-selected. Best to think of the matches from the first pass search as the evidence for the presence of the proteins. The matches from the second pass search give increased coverage and may give clues as to unsuspected modifications or SNPs that merit closer investigation.

Often, there are many possible explanations for an observed mass difference. For example, a delta of 28 Da could be formyl or dimethyl or ethyl or one of several amino acid substitutions, such as Lys->Arg. The search report can give a list of possibilities, but the researcher must use their knowledge and

experience to decide on the best explanation for the observed differences.

The main limitation of multi-pass searching is that it can only find matches to proteins that have at least one match to an unmodified peptide. Not so useful when studying endogenous peptides or very heavily modified proteins, such as histones. Similar limitations apply to 'unrestricted' modification searches, where peptide mass spectra are compared with one another, rather than matched to database sequences [30].

7. Scoring

Many different ways of scoring peptide matches have been developed. A review by Sadygov [31] classified scoring algorithms as Descriptive (e.g. Sequest [14], Sonar [32]), Interpretative (e.g. PeptideSearch [9], MS-Seq [33]), Stochastic (e.g. Scope [34], Olav [35]), and Probability-based (e.g. Mascot [36], OMSSA [37]). It is beyond the scope of this tutorial to go into detail about individual algorithms.

For search engines that have non-statistical scoring algorithms, there is the possibility to process the results with something like PeptideProphet [38] or Percolator [39], which converts the scores into probabilities. This makes it easier to apply a threshold to remove unreliable matches.

Whether or not the scoring algorithm is probability-based, database searching is a statistical process. Most MS/MS spectra do not encode the complete peptide sequence; there are gaps and ambiguities. Hopefully, most of the time, we are able to report the correct match, a 'true positive', but not always. If the sequence of the peptide is not in the database, and we obtain a match below our score or significance threshold, that is also OK, and we have a 'true negative'. The other two quadrants represent failure. A 'false positive' is when we report a significant match to the wrong sequence. A 'false negative' is when we fail to report a match even though the correct sequence is in the database. For real-life datasets, when we cannot be certain that all the correct sequences are present in the database, we don't know whether a failure to get a match to a spectrum is a true negative or a false negative. The usual way to measure the quality of a set of search results is with a false discovery rate, which doesn't require estimates of true negatives or false negatives [40].

Many search engines report expect (or expectation) values, either as scores or in addition to scores. An expect value is the number of times you would expect to get a score at least as high by chance. Small expect values are good, and a match with an expect value of 1 or more indicates a random match. You can calculate an expect value for a non-statistical score by fitting the tail of the score distribution to a straight line in log space [41]. Some software reports a PEP value (Posterior Error Probability), which is similar to an expect value for values much less than 1.

Even the best scoring scheme cannot fully separate the correct and incorrect matches. This is often illustrated in the form of a Receiver Operating Characteristic or ROC plot, which shows the relationship between the true positive and false positive rates as the threshold is varied (Fig. 5). The origin is a very high threshold, which lets nothing through. At the top right, we have a very low threshold, that allows everything through. Neither extreme is a useful place to be. The diagonal represents a useless scoring algorithm, that is equally likely to let through a false match as a true one. The solid curve shows a useful scoring algorithm, and the more it pushes up towards the top left corner, the better. Setting a threshold towards this top left corner gives a high ratio of correct matches to false matches.

8. Sensitivity and specificity

It could be argued that, a few years ago, there was too much focus on sensitivity and not enough consideration given to specificity, so that some of the published lists of proteins from database searching were not as accurate as the authors might have hoped. A growing awareness of this issue led to initiatives from various quarters. Most notably, the Editors of Molecular and Cellular Proteomics held a workshop in 2005 to define a set of guidelines [42]. Similar guidelines have been adopted by other journals and by the Proteomics Standards Initiative [43].

For large scale studies, there is a requirement to estimate the false discovery rate. One of the most reliable ways to do this is with a so-called target–decoy search. This is a very simple but powerful way of validating search results. The search is repeated, using identical search parameters, against a database in which the sequences have been reversed or shuffled. The number of matches from the decoy database is an excellent estimate of the number of false positives in the results from the target database [44].

There is a good deal of discussion in the literature about whether the decoy sequences should be reversed or randomised; whether to search a single database containing both target and decoy sequences or separate databases. These considerations may change the numbers in a small way, but the most important thing is to do some type of decoy search and estimate whether the level of false positives is 0.1% or 1% or 10%.

Although a decoy search is an excellent validation method for large data sets, it isn't useful when there are only a small

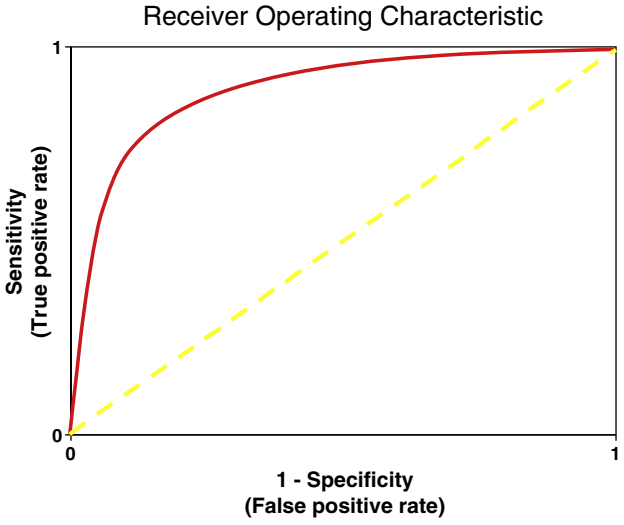

Fig. 5 – Receiver Operating Characteristic or ROC plot, which shows the relationship between the true positive and false positive rates as the threshold is varied.

number of spectra, because the numbers are too small to give an accurate estimate. Hence, this is not a substitute for a stable scoring scheme. You might think of it as a form of external calibration for the scoring algorithm.

In addition to transforming non-statistical scores into probabilities, PeptideProphet also seeks to improve sensitivity through better discrimination. It takes information about the matches, besides the score, and uses a machine learning algorithm called linear discriminant analysis to learn what distinguishes correct from incorrect matches. Examples of additional information would be precursor mass error, number of missed cleavages, or the number of tryptic termini [38].

A more recent development has been to use the matches from a decoy database as negative examples for a classifier. Percolator trains a machine learning algorithm called a support vector machine to discriminate between a sub-set of the high-scoring matches from the target database, assumed correct, and the matches from the decoy database, assumed incorrect [39].

9. Protein inference

It is essential to remember that database matching of MS/MS spectra identifies peptides, not proteins. Using the peptide sequences to deduce which proteins were present in the original sample is surprisingly difficult because many of the peptide sequences in a typical search result can be assigned to more than one protein. The guiding principal of protein inference is to create a minimal list of proteins. That is, the minimum number of proteins that can account for the observed peptides. Some people call this approach the principal of parsimony, others call it Occam's razor [45].

Imagine the very simple case where we have three peptide matches, which can be assigned to three proteins, as illustrated in Fig. 6. Do we have evidence for all three proteins, or just one? By the principal of parsimony, we will report that the sample contained protein A. Proteins B and C are classified as sub-set proteins, and given an inferior status. This is certainly a reasonable decision, but there is no guarantee that it is correct. It is possible that the sample actually did contain a mixture of proteins B and C, but not protein A. Another thing to watch for is the possibility that peptide 2 is a very weak match, maybe spurious. If so, then there is nothing to choose between proteins A and B.

This ambiguity is made worse in a shotgun proteomics or MudPIT experiment, where the proteins from a cell lysate are digested to peptides without any prior fractionation or separation. In general, no matter how good the data, there will be some ambiguity concerning which proteins were present in the sample. This can be a serious problem if someone who is not familiar with these issues ends up with the impression that the search gives evidence for the presence of all of the proteins in a family, rather than just one or two.

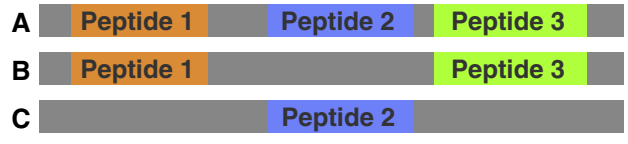

Fig. 6 – Three peptide matches, which can be assigned to three proteins. Do we have evidence for all three proteins, or just one?

Protein false discovery rate is not the same as peptide false discovery rate. It may be higher or lower, depending on the rules for accepting a protein or protein family. It is usually advisable to require that a protein has significant matches to more than one distinct peptide sequence. A protein with matches to just a single peptide sequence is commonly referred to as a 'one-hit wonder' and is often treated as suspect. This is actually a slight over-simplification. In a search with a large number of spectra and a small database, even though the peptide false discovery rate is low, a protein can pick up multiple false matches by chance. One way to guard against this is to look at the decoy search results and ensure that your rules for accepting a protein in the target database give no false positive proteins from the decoy.

10. Further reading

Other tutorials in this series cover related topics, in particular:

- De novo and sequence homology searching
- Protein ID by MALDI (Peptide Mass Fingerprinting)
- ID verification principles. PeptideProphet, etc.
- Proteomics databases (Data repositories).

Recent review articles that cover database searching of MS/MS data in greater depth or give a different perspective include those from Duncan et al. [46], Ma [47], Kumar and Mann [48], McHugh and Arthur [49], Matthiesen [50], and Hernandez et al. [51]. From the many text books that are available, *Computational Methods for Mass Spectrometry Proteomics* covers the area systematically, and each chapter contains a good selection of literature citations [52].

11. Practical exercises

A web-based collection of practical exercises has been compiled to accompany this tutorial. The starting page is http://www.ms-ms.com/exercises/exercises.html.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.jprot.2011.05.014.

REFERENCES

[1] Barton SJ, Whittaker JC. Review of factors that influence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors. Mass Spectrom Rev 2009;28(1):177–87.
[2] Kelleher NL. Top-down proteomics. Anal Chem 2004;76(11): 196A–203A.

[3] Fricker LD, Lim JY, Pan H, Che FY. Peptidomics: identification and quantification of endogenous peptides in neuroendocrine tissues. Mass Spectrom Rev 2006;25(2): 327–44.
[4] Papayannopoulos IA. The interpretation of collision-induced dissociation tandem mass spectra of peptides. Mass Spectrom Rev 1995;14(1):49–73.
[5] Paizs B, Suhai S. Fragmentation pathways of protonated peptides. Mass Spectrom Rev 2005;24(4):508–48.
[6] Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed Mass Spectrom 1984;11(11):601.
[7] McAlister GC, Berggren WT, Griep-Raming J, Horning S, Makarov A, Phanstiel D, et al. A proteomics grade electron transfer dissociation-enabled hybrid linear ion trap–orbitrap mass spectrometer. J Proteome Res 2008;7(8):3127–36.
[8] Seidler J, Zinn N, Boehm ME, Lehmann WD. De novo sequencing of peptides by MS/MS. Proteomics 2010;10(4): 634–49.
[9] Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem 1994;66(24):4390–9.
[10] Searle BC, Dasari S, Turner M, Reddy AP, Choi D, Wilmarth PA, et al. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. Anal Chem 2004;76(8):2220–30.
[11] Shevchenko A, Sunyaev S, Loboda A, Shevehenko A, Bork P, Ens W, et al. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time of flight mass spectrometry and BLAST homology searching. Anal Chem 2001;73(9):1917–26.
[12] Tabb DL, Saraf A, Yates JR. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem 2003;75(23):6415–21.
[13] Sunyaev S, Liska AJ, Golod A, Shevchenko A, Shevchenko A. MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. Anal Chem 2003;75(6):1307–15.
[14] Eng JK, McCormack AL, Yates III JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 1994;5(11):976–89.
[15] Choudhary JS, Blackstock WP, Creasy DM, Cottrell JS. Matching peptide mass spectra to EST and genomic DNA databases. Trends Biotechnol 2001;19(10):S17–22.
[16] Luethy R, Kessner DE, Katz JE, McLean B, Grothe R, Kani K, et al. Precursor-ion mass re-estimation improves peptide identification on hybrid instruments. J Proteome Res 2008;7(9): 4031–9.
[17] Cox J, Mann M. Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap. J Am Soc Mass Spectrom 2009;20(8):1477–85.
[18] Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, et al. mzML — a community standard for mass spectrometry data. Mol Cell Proteomics 2010. http://www.mcponline.org/content/10/1/R110.000133.
[19] Jensen ON. Interpreting the protein language using proteomics. Nat Rev Mol Cell Biol 2006;7(6):391–403.
[20] Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 2006;24(10):1285–92.
[21] Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 2006;127(3):635–48.
[22] Tanner S, Shu HJ, Frank A, Wang LC, Zandi E, Mumby M, et al. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem 2005;77(14): 4626–39.
[23] Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. Identification of post-translational modifications by blind search of mass spectra. Nat Biotechnol 2005;23(12):1562–7.
[24] Tanner S, Payne SH, Dasari S, Shen Z, Wilmarth PA, David LL, et al. Accurate annotation of peptide modifications through unrestrictive database search. J Proteome Res 2008;7(1): 170–81.
[25] Ruttenberg BE, Pisitkun T, Knepper MA, Hoffert JD. PhosphoScore: an open-source phosphorylation site assignment tool for MSn data. J Proteome Res 2008;7(7): 3054–9.
[26] Lu BW, Ruse C, Xu T, Park SK, Yates J. Automatic validation of phosphopeptide identifications from tandem mass spectra. Anal Chem 2007;79(4):1301–10.
[27] Bailey CM, Sweet SMM, Cunningham DL, Zeller M, Heath JK, Cooper HJ. SLoMo: automated site localization of modifications from ETD/ECD mass spectra. J Proteome Res 2009;8(4):1965–71.
[28] Savitski MM, Nielsen ML, Zubarev RA. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. Mol Cell Proteomics 2006;5(5):935–48.
[29] Savitski MM, Lemeer S, Boesche M, Lang M, Mathieson T, Bantscheff M, et al. Confident phosphorylation site localization using the Mascot Delta Score. Mol Cell Proteomics 2010. http://www.mcponline.org/content/10/2/M110.003830.
[30] Ahrne E, Muller M, Lisacek F. Unrestricted identification of modified proteins using MS/MS. Proteomics 2010;10(4): 671–86.
[31] Sadygov RG, Cociorva D, Yates JR. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. Nat Methods 2004;1(3): 195–202.
[32] Field HI, Fenyo D, Beavis RC. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. Proteomics 2002;2(1):36–47.
[33] Clauser KR, Baker P, Burlingame AL. Role of accurate mass measurement (+/–10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal Chem 1999;71(14):2871–82.
[34] Bafna V, Edwards N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics 2001;17(suppl 1):S13–21.
[35] Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 2003;3(8):1454–63.
[36] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999;20(18):3551–67.
[37] Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, et al. Open mass spectrometry search algorithm. J Proteome Res 2004;3(5):958–64.
[38] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74(20):5383–92.
[39] Spivak M, Weston J, Bottou L, Kall L, Noble WS. Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. J Proteome Res 2009;8(7): 3737–45.
[40] Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 2007;4(10):787–97.
[41] Eng JK, Fischer B, Grossmann J, MacCoss MJ. A fast SEQUEST cross correlation algorithm. J Proteome Res 2008;7(10):4598–602.

[42] Bradshaw RA, Burlingame AL, Carr S, Aebersold R. Reporting protein identification data: the next generation of guidelines. Mol Cell Proteomics 2006;5(5):787–8.
[43] Binz P-A, Barkovich R, Beavis RC, Creasy D, Horn DM, Julian RK, et al. Guidelines for reporting the use of mass spectrometry informatics in proteomics. Nat Biotechnol 2008;26(8):862.
[44] Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 2007;4(3):207–14.
[45] Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data — the protein inference problem. Mol Cell Proteomics 2005;4(10):1419–40.
[46] Duncan MW, Aebersold R, Caprioli RM. The pros and cons of peptide-centric proteomics. Nat Biotechnol 2010;28(7):659–64.
[47] Ma B. Challenges in computational analysis of mass spectrometry data for proteomics. J Comput Sci Technol 2010;25(1):107–23.
[48] Kumar C, Mann M. Bioinformatics analysis of mass spectrometry-based proteomics data sets. FEBS Lett 2009;583 (11):1703–12.
[49] McHugh L, Arthur JW. Computational methods for protein identification from mass spectrometry data. PLoS Comput Biol 2008;4(2).
[50] Matthiesen R. Methods, algorithms and tools in computational proteomics: a practical point of view. Proteomics 2007;7(16):2815–32.
[51] Hernandez P, Muller M, Appel RD. Automated protein identification by tandem mass spectrometry: issues and strategies. Mass Spectrom Rev 2006;25(2):235–54.
[52] Eidhammer I, Flikka K, Martens L, Mikalsen S-O. Computational methods for mass spectrometry proteomics. Chichester, UK: Wiley; 2007.