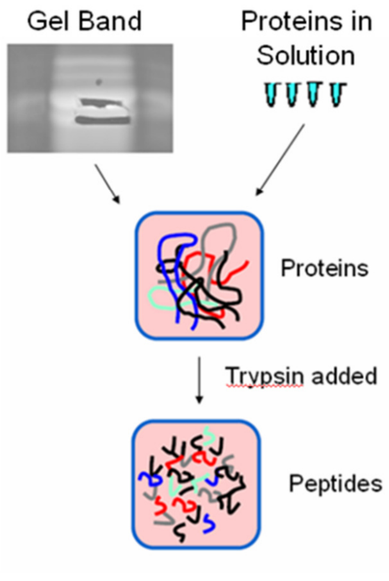
Identifying Proteins: from Gel Bands to Useable Data.

Protein identification is a core task in nearly every proteomics experiment. Whether these studies are looking for potential binding sites for a viral protein on the cell surface, searching for new biomarkers in human serum, correct and confident identification is essential for success, and to avoid wasting time and money chasing bad leads.

Protein identification is a complex process with many steps and we have a multitude of different tools and techniques available to help us with protein identification. What follows is a summary of how we identify proteins using the mass spectrometer.

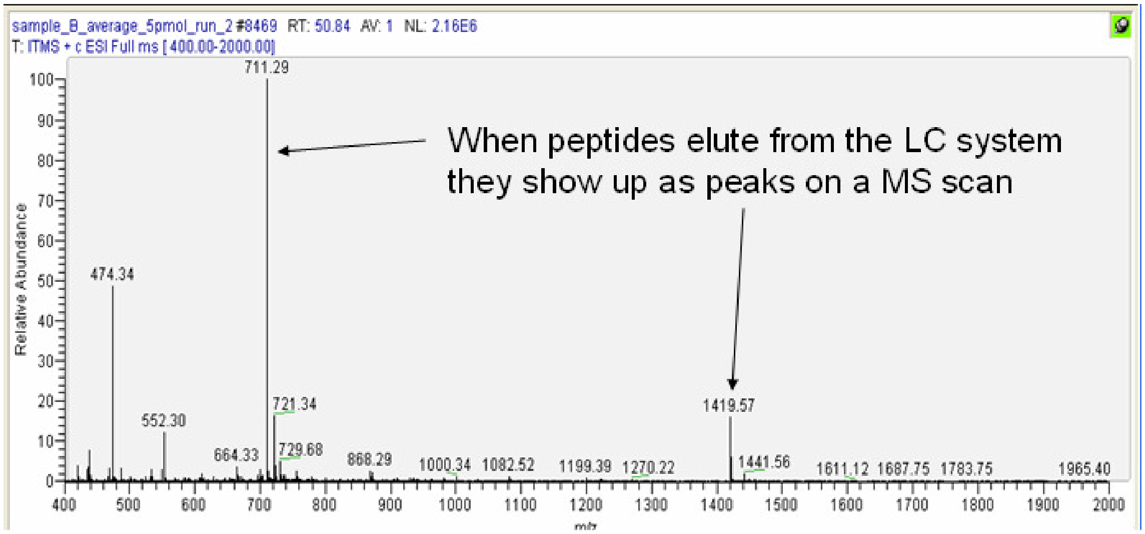
Sample preparation

Protein samples are generally submitted to PSR either as a gel band or in a liquid solution, such as a Co-IP elution buffer. PSR employs a bottom-up approach to protein identification, which involves breaking down the protein into smaller peptides for analysis. Before this happens the cystine residues on the protein are usually reduced and alkylated to prevent them from forming di-sulfide bonds. The protein is then broken down through enzymatic digestion. The most common enzyme used in the field is trypsin, but there are dozens of other enzymes on the market which cut the protein at different places along the amino acid chain, and different enzymes can be used upon request.

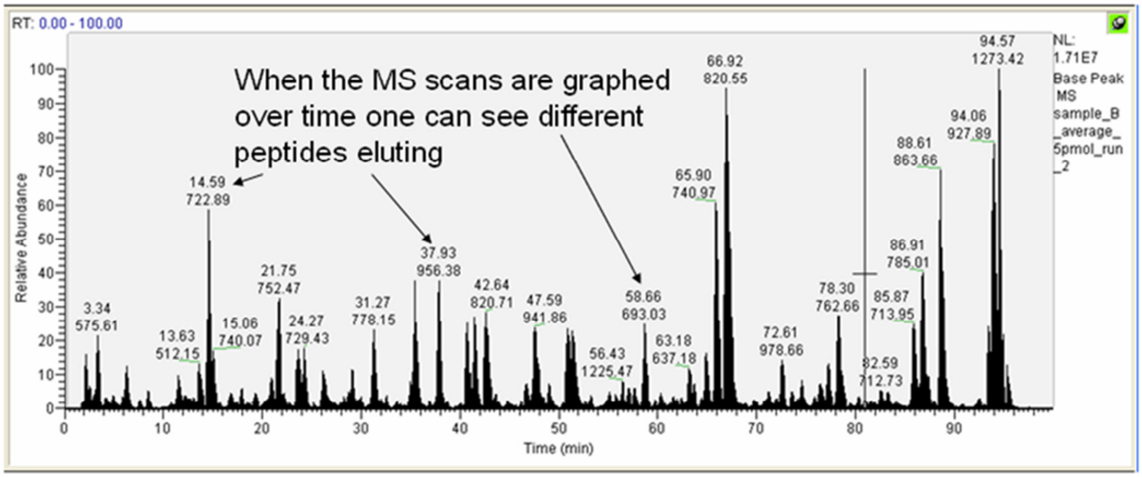


Once the protein has been digested it is ready for introduction into a liquid chromatography (LC) system. Here proteins are adhered to a C18 column in an acetic aqueous solution. Next a liquid pump slowly increases the amount of organic solvent (usually Acetonitrile) flowing over the sample. In the presence of the acetonitrile peptides lose their grip on the column and are eluted. Once eluted from the column they are introduced into the mass spectrometer via electrospray. When they enter the mass spec a peak can be seen in a MS Spectra at a point corresponding to their m/z value.

Peptides that are more hydrophobic tend to elute first and more hydrophilic peptides elute later in the gradient. When looking at the height of the peaks in the MS scans over time you can often see individual peptides eluting. Very complex mixtures are often subjected to more than one phase of fractionation. Salt gradients can be added into a series of organic phase gradients to enhance separation of the peptides. These multi-dimensional separations are commonly done by PSR for TMT projects, or other experiments that require gathering data on thousands of proteins.



When peptides elute from the LC system they show up as peaks on a MS scan

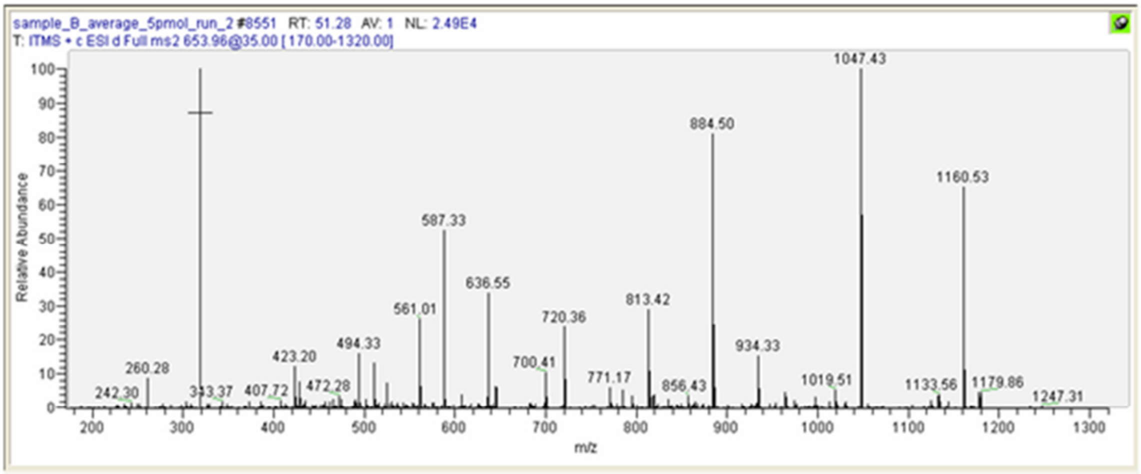


When the MS scans are graphed over time one can see different peptides eluting

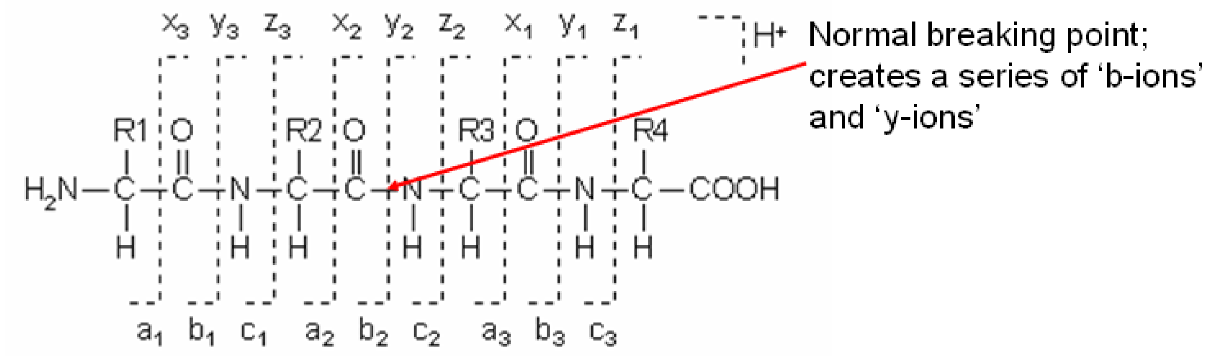
After peaks are identified in the MS spectra they are then selected for 2nd order or MS/MS scans which are used for identification. In this process first the m/z of the peptide is isolated and then the peptide is then broken apart. At PSR this process is accomplished most frequently via a process called collision-induced dissociation or (CID) or Higher-energy collisional dissociation (HCD). In these methods peptide ions are forced to collide with an inert gas which breaks down the peptide down into its component parts. These parts are then scanned into the detector of the mass spectrometer creating an MS/MS Spectra. Once a peptide has been fragmented it is then put onto an exclusion list to allow the mass spectrometer to choose other less intense peaks for analysis.

How Peptides Fragment

Below is an MS/MS Spectra created from the fragmentation of a peptide in one of our ion trap mass spectrometers. The different fragments each produce a different m/z peak creating a fragmentation pattern unique to that peptide.



After collision with the inert gas the peptide can break apart at any point along its amino acid backbone, or on its side chains. Mass spectrometers are usually calibrated to deliver a specific amount energy, in the form of a voltage, which excites the peptides to the point where the breakage occurs mostly along their backbone.


x3 y3 z3 x2 y2 z2 x1 y1 z1
H+
Normal breaking point; creates a series of 'b-ions' and 'y-ions'
R1 O R2 O R3 O R4
H2N C C N C C N C C N C COOH
H H H H H H H
a1 b1 c1 a2 b2 c2 a3 b3 c3

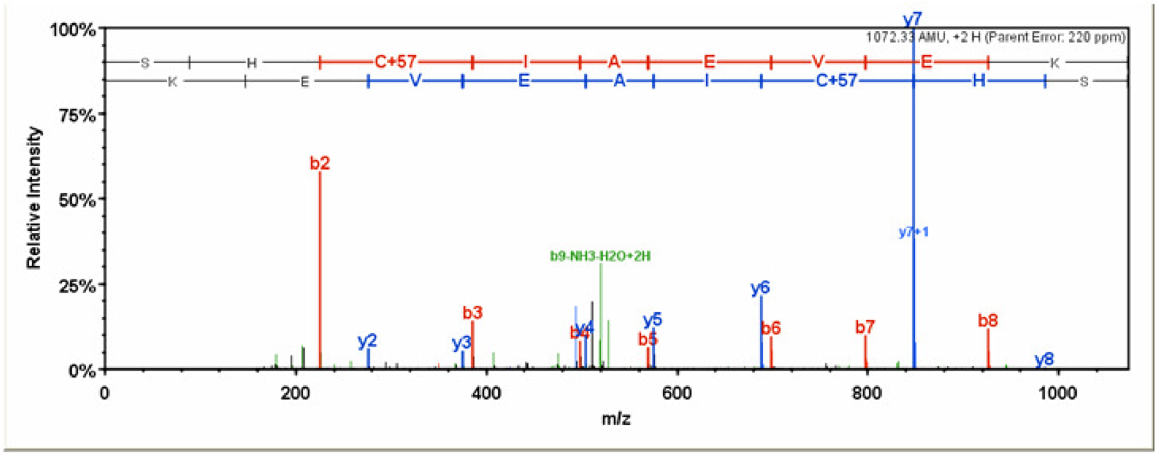
Because the point the peptide breaks is mostly random (some locations are more common than others), and varies from molecule to molecule, a series of peaks are generated from the collision of a single peptide. Each collision generates a pair of masses each representing one part of the whole peptide. Below is an example of the series of masses created from the fragmentation of a single peptide.

Initial Peptide	Mass + 2H+	m/z peak
L-M-S-T-A-A-R	732.88	366.44

Mass	B-ion series	Y-ion series	Mass
114.16	L	M-S-T-A-A-R	618.72
245.35	L-M	S-T-A-A-R	487.53
332.43	L-M-S	T-A-A-R	400.45
433.53	L-M-S-T	A-A-R	299.35
504.61	L-M-S-T-A	A-R	228.27
575.69	L-M-S-T-A-A	R	157.19

Fragmentation Pattern Analysis

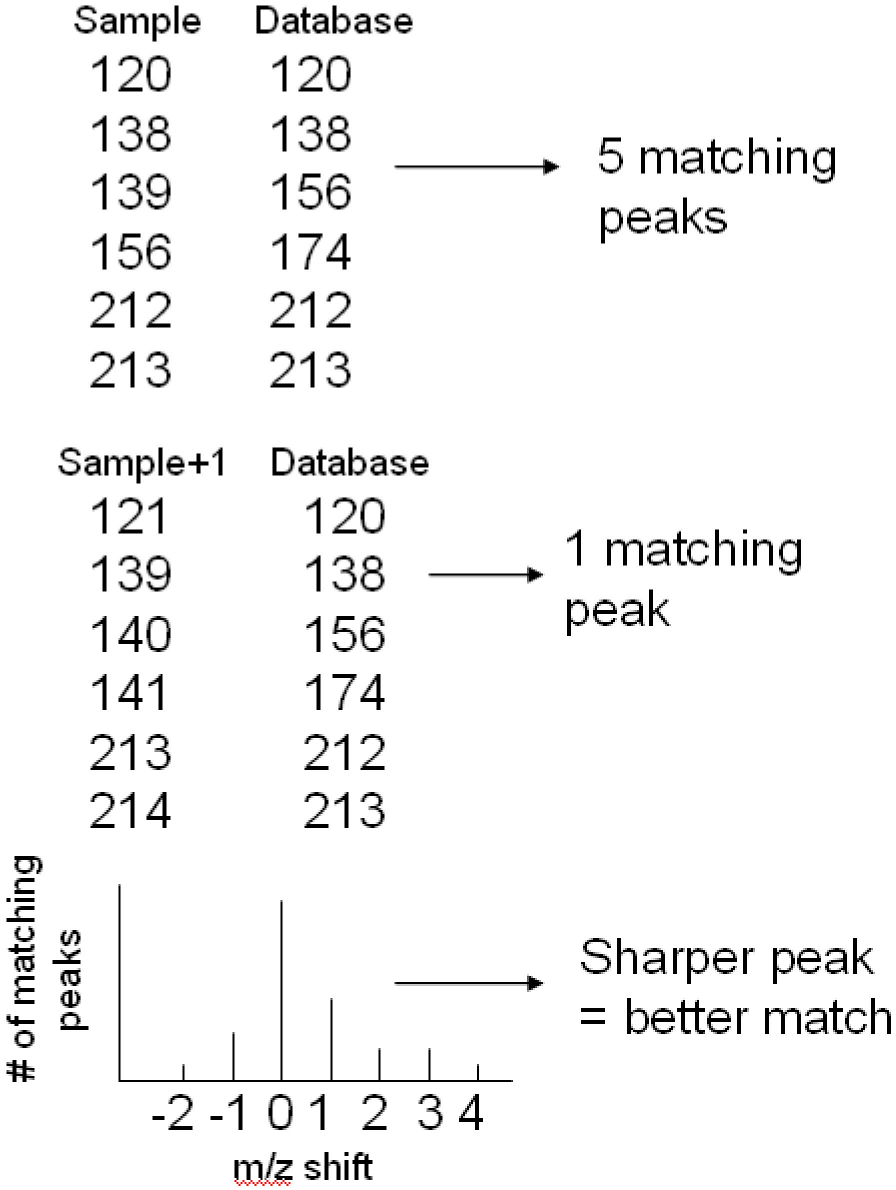
Once a MS/MS spectra has been created the next step is trying to determine which peptide the fragmentation pattern represents. This process used to be done by hand, and often involved many hours spent looking at the spectra, calculating the spacing between the peaks, weeding out 'bad' spectra, and coming up with a possible identification



compares the fragmentation pattern to a database which can contain the sequences of 100,000+ proteins, and makes a determination of the best possible match. At PSR we have several different computer programs which can do this for us. Comet is our most commonly used software, it is the opensource version of what has grown from the original Sequest algorithm. We also have Sequest HT which is part of Proteome Discoverer and mostly used for TMT analysis, Andromeda which is integrated with MassQC and used for SILAC samples, and Byonic which allows for unrestricted PTM searches. Each program has its own pros and cons, but they all use similar logic to identify peptides.

A Simple Program (Sequest)

Each of the spectral matching algorithms works in a slightly different manner, but they make a lot of similar assumptions. What follows is an overview of how the classic Sequest program matches spectra to peptides. Sequest was the first program that was written to do this, and is a good model for understanding how these programs work.



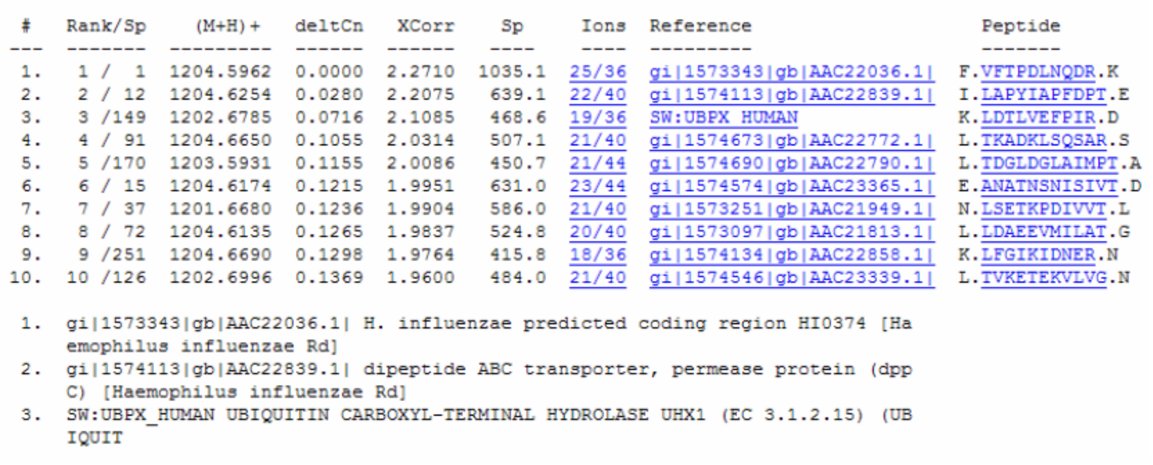
Sample	Database
120	120
138	138
139	156
156	174
212	212
213	213

Sample+1	Database
121	120
139	138
140	156
141	174
213	212
214	213

To start a 'peak list' is generated from the raw mass spectrometer data. This consists of a list of masses and intensities from MS/MS spectra. The sequest algorithm only uses the list of masses in matching the spectra to potential spectra. The first step is to put the data into 1 Da 'bins.' To do this the algorithm looks to see if there is a peak present for each whole number m/z value. If there is then it will list that value, if not then it will not. Next the algorithm does the same thing for each potential sequence in the database it is searching against. Then Sequest compares the two lists and determines the number of matching values.

Following this an additional mass unit is added to one of the datasets and the process is repeated. This happens for multiple times for multiple different m/z shifts. The result of this comparison can be seen above, with a different number of matching m/z values for various m/z shifts. In a correct match there should be a high number of matches with no shift, and many less matches when a mass shift is applied to one of the datasets. Incorrect matches should see little to no change in the number of matching peaks when shifted. The sharpness of the drop off in the number of matching peaks is interpreted into a correlation score; which is the basis for determining how well a MS/MS spectra matches a potential sequence. This is referred to as the XCorr value.

The other factor which is given a lot of weight is the size of the gap in XCorr scores between the best and 2nd best matching sequences. This is reflected in the deltCn score. A correct pairing of spectra and hypothetical sequence should result in a larger gap in correlation scores than an incorrect pairing. An example of Sequest output can be seen below.



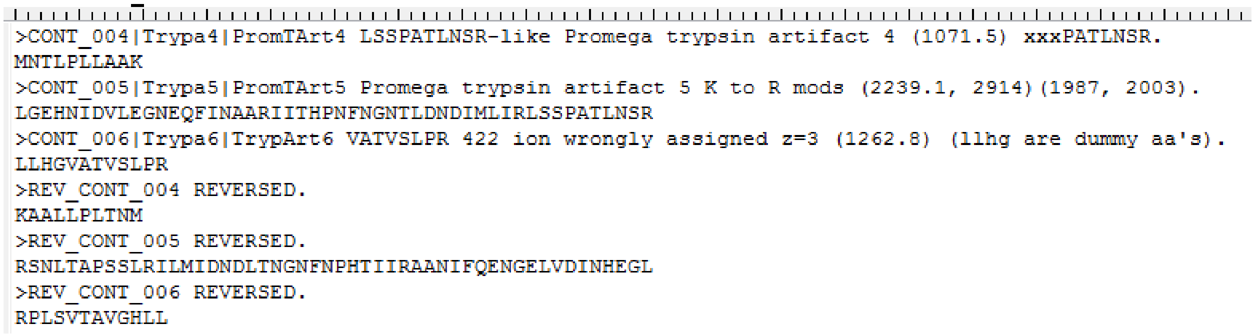
```
 #   Rank/Sp    (M+H)+    deltCn   XCorr    Sp     Ions   Reference                    Peptide
---  -------  ---------- ------  ------   ----    ----   ---------                    -------
 1.   1 /  1  1204.5962 0.0000  2.2710  1035.1  25/36  gi|1573343|gb|AAC22036.1|  F.VFTPDLNQDR.K
 2.   2 / 12  1204.6254 0.0280  2.2075   639.1  22/40  gi|1574113|gb|AAC22839.1|  I.LAPYIAPFDPT.E
 3.   3 /149  1202.6785 0.0716  2.1085   468.6  19/36  SW:UBPX_HUMAN              K.LDTLVEFPIR.D
 4.   4 / 91  1204.6650 0.1055  2.0314   507.1  21/40  gi|1574673|gb|AAC22772.1|  L.TKADKLSQSAR.S
 5.   5 /170  1203.5931 0.1155  2.0086   450.7  21/44  gi|1574690|gb|AAC22790.1|  L.TDGLDGLAIMPT.A
 6.   6 / 15  1204.6174 0.1215  1.9951   631.0  23/44  gi|1574574|gb|AAC23365.1|  E.ANATNSNISIVT.D
 7.   7 / 37  1201.6680 0.1236  1.9904   586.0  21/40  gi|1573251|gb|AAC21949.1|  N.LSETKPDIVVT.L
 8.   8 / 72  1204.6135 0.1265  1.9837   524.8  20/40  gi|1573097|gb|AAC21813.1|  L.LDAEEVMILAT.G
 9.   9 /251  1204.6690 0.1298  1.9764   415.8  18/36  gi|1574134|gb|AAC22858.1|  K.LFGIKIDNER.N
10.  10 /126  1202.6996 0.1369  1.9600   484.0  21/40  gi|1574546|gb|AAC23339.1|  L.TVKETEKVLVG.N

 1.  gi|1573343|gb|AAC22036.1| H. influenzae predicted coding region HI0374 [Ha
     emophilus influenzae Rd]
 2.  gi|1574113|gb|AAC22839.1| dipeptide ABC transporter, permease protein (dpp
     C) [Haemophilus influenzae Rd]
 3.  SW:UBPX_HUMAN UBIQUITIN CARBOXYL-TERMINAL HYDROLASE UHX1 (EC 3.1.2.15) (UB
     IQUIT
```

Probability Software

Once a potential matching spectra and peptide have been paired together another computer program is run to assess the probability of that pairing being correct. There are two major ways that this gets done, one is by estimating the error rate by using reversed or similar nonsense protein databases, the other is by modeling the distribution of the scores and fitting multiple curves under what is often a bi-modal distribution. Below we'll take a closer look at both these approaches.

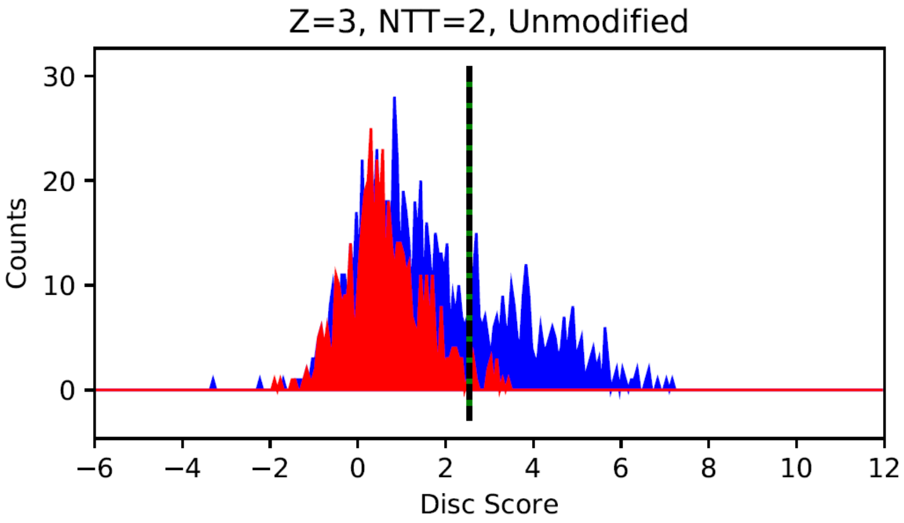
Nonsense Databases

The most common method today for estimating error in a bottom-up proteomics experiment is by using a reversed database. A reversed database is just as simple as it sounds; the amino acid sequences in the protein entries are reversed to run C-terminal to N-terminal. Most commonly these sequences are then appended to the initial protein database as seen below.



```
>CONT_004|Trypa4|PromTArt4 LSSPATLNSR-like Promega trypsin artifact 4 (1071.5) xxxPATLNSR.
MNTLPLLAAK
>CONT_005|Trypa5|PromTArt5 Promega trypsin artifact 5 K to R mods (2239.1, 2914)(1987, 2003).
LGEHNIDVLEGNEQFINAARIITHPNFNGNTLDNDIMLIRLSSPATLNSR
>CONT_006|Trypa6|TrypArt6 VATVSLPR 422 ion wrongly assigned z=3 (1262.8) (llhg are dummy aa's).
LLHGVATVSLPR
>REV_CONT_004 REVERSED.
KAALLPLTNM
>REV_CONT_005 REVERSED.
RSNLTAPSSLRILMIDNDLTNGNFNPHTIIRAANIFQENGELVDINHEGL
>REV_CONT_006 REVERSED.
RPLSVTAVGHLL
```

The idea is that MS/MS spectra correctly matching to proteins that are present will cluster in the forward database, while incorrectly assigned spectra will match equally often to both the original protein sequences and the additional reversed entries. Because of this the matches to the reversed entries can be used to estimate the number of matches to the forward entries, and thus the false-positive rate of different score thresholds in the experiment. So if you have 4 proteins 'identified' in the reversed database, and 350 matched in the forward database you could assume approximately 4 of those 350 proteins have been incorrectly identified. This yields a false discovery rate estimate of 1.1% (4/350).



There are different variations of the reversed-database strategy that all work similarly, with various pros and cons. Efforts can be made to keep the enzymatic cut sites in the same location which may better mimic the peptide size distribution of the original database. Randomized protein sequences can also be used instead of straight reversals, which can remove problems with repeating motifs, and allow for multiple randomized copies of a database to be appended in cases where a smaller database is a potential problem. Reversed databases can also be searched without having the original protein sequences present, and score distributions from one search can be used to infer distribution in the other. While some of these additional methods likely do a better job of estimating the amount of truly random errors they may also mask more non-random errors, which can lead to higher estimates confidence levels than they should. The appended reversed database is usually seen as the more conservative method for estimating error.

Distribution Matching

Once a potential matching spectra and peptide have been paired together another computer program is run to assess the probability of that pairing being correct. This computer program can take score input from a spectral matching program and use this data to give the match a probability. To do this software will start by calculating a discriminant score for the spectra.

The discriminant score is based on a number of different values including:

-the correlation score (XCorr) from Sequest (or equivalent from other programs)

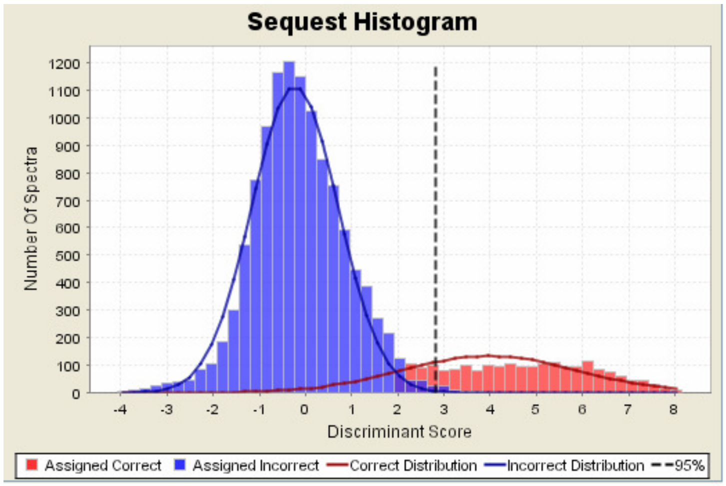
-the deltCn from Sequest

-the number of peptide termini that match enzyme cleavage sites

-the number of missed cleavages

-the charge state

-and other values



The software then assumes that there are two distributions in the discriminate scores: correct and incorrect, and attempts to fit curves over the distributions to determine at a set value how many of the spectra are correctly and incorrectly identified. The relative height of these two curves at any point is used to determine the probability that an identification is correct.

From Peptides back to Proteins

Once a list of identified peptides have been assembled the next step is to work backwards from the peptide level and compile a list of proteins. There are several factors that complicate the translation of peptides back into proteins. Some of these are as follows:

-Same peptide found in many proteins

-False Positive rate

-Database size

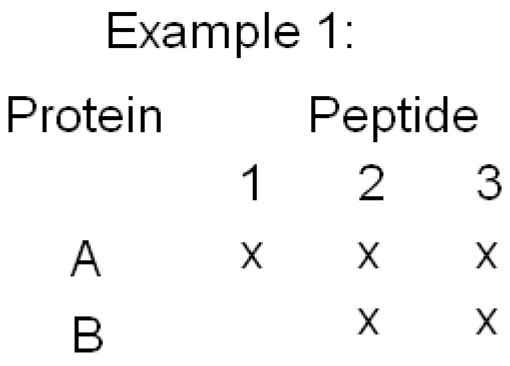
-Search criteria

-And others…

Some of these problems are easier to handle than others. For example many simpler organisms have fewer proteins, and thus a smaller protein database. The Swissprot human database contains about 20,000 entries, but the E.coli database only has about 4,000 proteins. This is a problem as false-positive peptide identifications tend to be randomly distributed throughout the database, while true-positives tend to cluster. With 5x the space to spread out in the human database we can tolerate far more noise at the peptide level; whereas with the E.coli database peptides will appear to cluster sooner simply by chance. What this means in practice is that we have to set stricter score thresholds at the peptide level in smaller organisms to keep the false-positive rate at the protein level manageable.

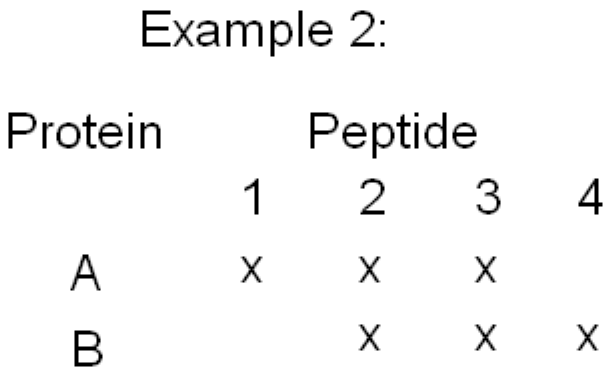
A different problem occurs when the same peptide is found in multiple proteins. This happens most commonly among members of the same protein family, proteins with alternative splicing forms, proteins with precursor forms, and among similar species when using a multi-species database. If there are no identified peptides which can be used to distinguish the proteins from each other one of them (generally the first one listed in the database by default) will be reported and the other proteins will be listed as redundant sequences, or listed elsewhere depending on the software used.

Things get even more complicated when two or more proteins partly overlap and the program must decide which one or ones are present. In these cases the software will usually apply an Occam's razor approach: it tries to create the shortest list of proteins that accounts for all the peptides present in the sample. Some examples of this decision making process are found below:



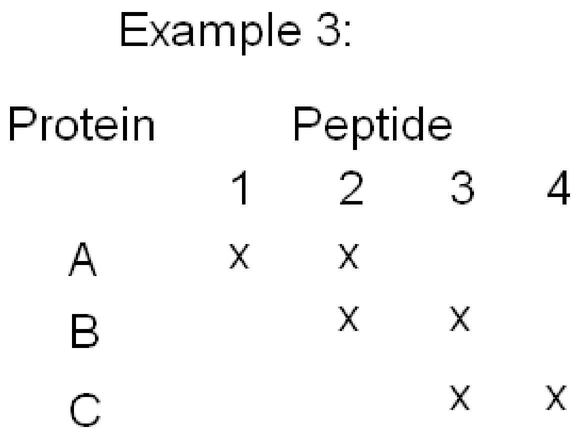
Example 1:

Protein	Peptide		
	1	2	3
A	x	x	x
B		x	x



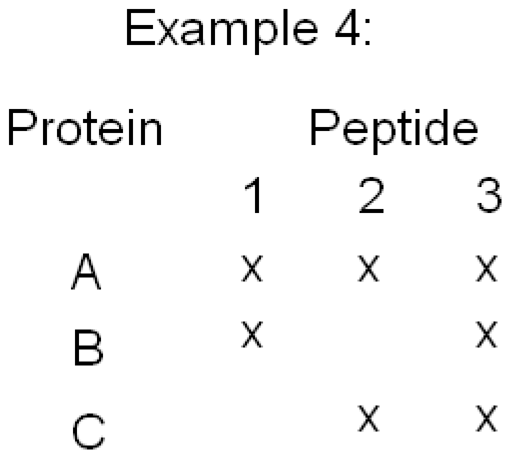
Example 2:

Protein	Peptide			
	1	2	3	4
A	x	x	x	
B		x	x	x



Example 3:

Protein	Peptide			
	1	2	3	4
A	x	x		
B		x	x	
C			x	x



Example 4:

Protein	Peptide		
	1	2	3
A	x	x	x
B	x		x
C		x	x

In example 1 only Protein A would be assumed to be present because it is the only protein with independent evidence, and Protein B would not show up in the list of identified proteins with most software. In example 2 both proteins assumed to be present because they both have independent evidence. In example 3 only Proteins A and C will be identified because there are no peptides assigned to Protein B which cannot be explained by other proteins. This is a clear example of the Occam's razor approach, as is example 4. In this case only Protein A will be listed because it can account for all three peptides by itself.

So Where is my Protein?

If you were expecting to find a particular protein in an experiment, and it didn't appear on your list of identified proteins there are several possible reasons.

- The protein isn't in the database (relatively easy to check/fix)
- That there is sequence overlap with other protein ID's (Blast search ID'ed peptides or compare sequences). For example, in the situation above, if Peptide 1 is a false match, and you were looking for Protein B, then Scaffold may not have given you the correct result. Also, Protein A could be a precursor or closely related protein.
- That there was sufficient protein to ID the sample

Some Final Thoughts on Protein ID Work

It is important to remember that the while the identification of proteins using mass spectrometry has started to mature as a science there are many aspects about it that are still difficult, and still actively being researched. For example, it can be next to impossible to identify which form of a protein is present in a sample if there are many forms present in the cell. This is because of the large amount of sequence overlap. There are also some peptides which you will never see in a mass spectrometer. Peptides from membrane proteins are particularly hard to see as many of them are highly hydrophobic and don't stick well to our column material or ionize well in the mass spectrometer. The process only becomes more complicated as the searches expand into looking for post-translational modifications, and/or attempting to do quantitation.

Well I hope you know a little more about Proteomics now. You can always contact us here at PSR if you have any questions. We're always happy to help out!