De novo sequencing of mass spectrometry data

You have just performed a large mass spectrometry experiment looking at the cell cycle (see associated R notebook for details, but the details do not matter so much for this part of the practical). We would like you to interpret some raw mass spectrometry data so that you can get an idea of how you can deduce peptide sequence from this data and how you get from MS/MS spectrum to sequence to protein identification.

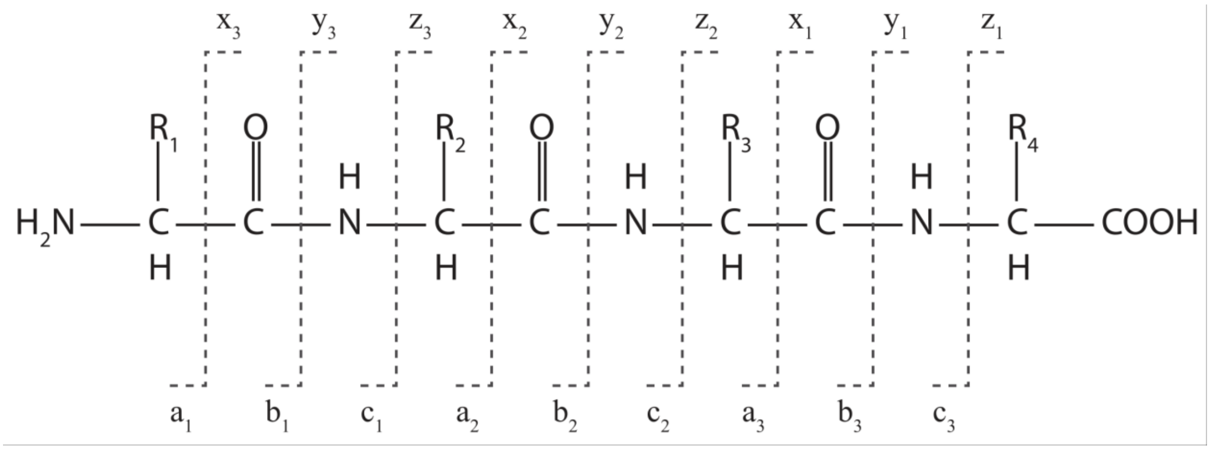

Figure 1. Possible fragmentations of peptide ions in tandem MS.

Table 1. Amino acid residue masses

Amino Acid	3 Letter Code	Single Letter Code	Residue Mass	
			Monoisotopic	Average
Glycine	Gly	G	57.02147	57.052
Alanine	Ala	A	71.03712	71.079
Serine	Ser	S	87.03203	87.078
Proline	Pro	P	97.05277	97.117
Valine	Val	V	99.06842	99.133
Threonine	Thr	T	101.04768	101.105
Cysteine	Cys	C	103.00919	103.144
Isoleucine	Ile	I	113.08407	113.160
Leucine	Leu	L	113.08407	113.160
Asparagine	Asn	N	114.04293	114.104
Aspartic Acid	Asp	D	115.02695	115.089
Glutamine	Gln	Q	128.05858	128.131
Lysine	Lys	K	128.09497	128.174
Glutamic Acid	Glu	E	129.04260	129.116

Methionine	Met	M		131.04049	131.198
Histidine	His	H		137.05891	137.142
Phenylalanine	Phe	F		147.06842	147.177
Arginine	Arg	R		156.10112	156.188
Tyrosine	Tyr	Y		163.06333	163.170
Tryptophan	Trp	W		186.07932	186.213

Q1. Why are there two masses given per amino acid residue? What is the difference between monoisotopic and average mass?

Exercise 1.

First of all, we would like you to look at a simpler example of *de novo* sequencing from a spectrum that is unrelated to this practical but illustrates how one performs sequencing (Figure 2). The spectrum is an MS/MS spectrum of a peptide from a tryptic digest that has been acquired on a time-of-flight instrument and the fragmentation used is collision-induced dissociation (CID). This form of fragmentation produces predominantly b- and y-ions. From consulting the associated survey (MS) scan we know that the precursor is doubly charged (z = 2).

The easiest way to sequence is to assign the y-ion series from this spectrum. Tryptic digest produces peptides that terminate in K or R so the y_1 ion should be one of these amino acids. In Figure 2, all of the m/z values you will need to sequence this peptide have been labelled (but there are some others which do not correspond to y-ions).

Q2. What is the C-terminal residue of this peptide?

The y-ion series is always most prevalent at higher m/z values. The y-ion series in this MS/MS spectrum is complete and the difference in m/z between y-ions is equal to an amino acid residue mass. We can sequence the peptide based on the delta m/z of the y-ion series, as it is complete.

Deduce the sequence of the peptide. Start at m/z 1464.65 and work in both directions.

Q3. What is the sequence of this peptide?

Perform a BLAST search on the sequence you have deduced (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Select “protein BLAST” and enter the sequence into the text box. Leave all parameters as they are and click on “BLAST”.

Q4. From which protein is the peptide likely to have been derived?

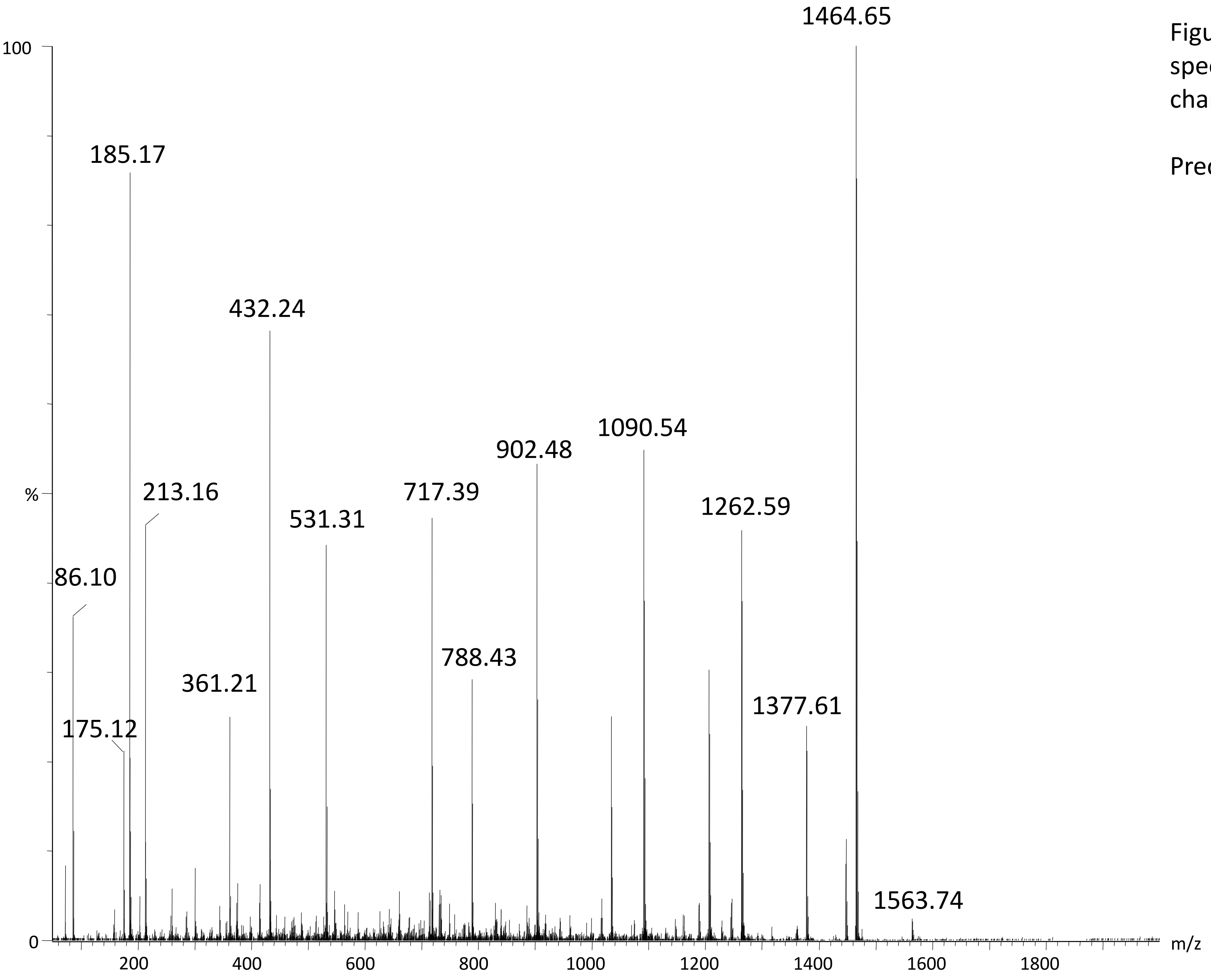

Figure 2. MS/MS spectrum of doubly charged precursor.

Precursor m/z = 838.9

Exercise 2.

Hopefully you will have worked out that this peptide is most definitely not derived from a human protein! Now we would like you to try deducing the sequence from a more complex MS/MS spectrum (Figure 3), which is a peptide from a tryptic digest of a human sample. This spectrum has been acquired on an Orbitrap instrument and the fragmentation used is higher-energy collisional dissociation (HCD). Like CID, this form of fragmentation produces predominantly b- and y-ions. From consulting the associated survey (MS) scan we know that the precursor is doubly charged (z = 2). This peptide has a full y-ion series. Using a similar approach to that used in Exercise 1, deduce the sequence of the peptide. Not all of the m/z values you will need have been labelled.

Q5. What is the sequence of the peptide?

Perform a BLAST search on the sequence you have deduced (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Select "protein BLAST" and enter the sequence into the text box. This time under database, specify "UniProtKB/Swiss-Prot(swissprot)" and under "Organism" type "Homo sapiens (taxid:9606)".

Q6. From which protein is the peptide likely to have been derived? Can you tell from this sequence?

Q7. What do the resulting "hits" represent? How might you go about determining which protein the peptide has been derived from?

Q8. To what do the ions at 86.10 in Figure 2 and 120.08 in Figure 3 correspond? How are they formed?

We have illustrated here how to sequence a peptide *de novo* from an MS/MS spectrum. Mass spectrometry experiments are large and the raw data are very complex, typically containing tens or even hundreds of thousands of spectra, so you can appreciate that sequencing all of the peptides in the entire experiment using this approach is neither practical nor sensible. This is compounded by the fact that often it is not possible to assign all y-ions in a given MS/MS spectrum as the y-ion series is not complete.

Instead this process is generally automated, using an algorithm such as Mascot which perform an *in silico* digest of the entire organism-specific database and match observed m/z values from the mass spectrometry experiment with theoretical m/z values derived from the *in silico* digest, given tolerances for the specific mass analyser used in the experiment.

Now you have finished this section of the practical, move on to the associated R Notebook exercise.

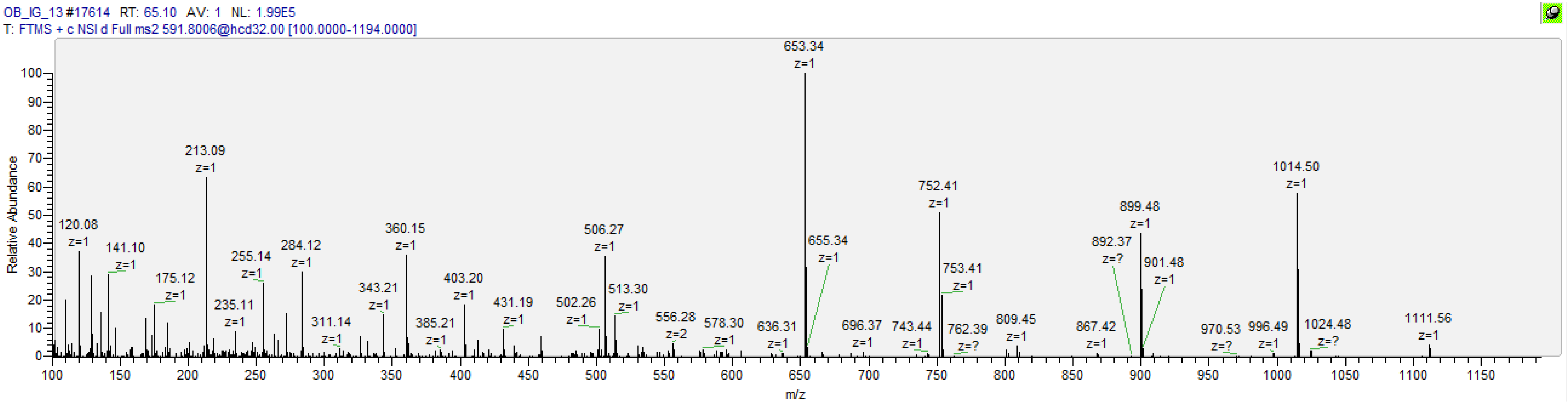

Figure 3. MS/MS spectrum of doubly charged precursor. Precursor m/z = 591.80.