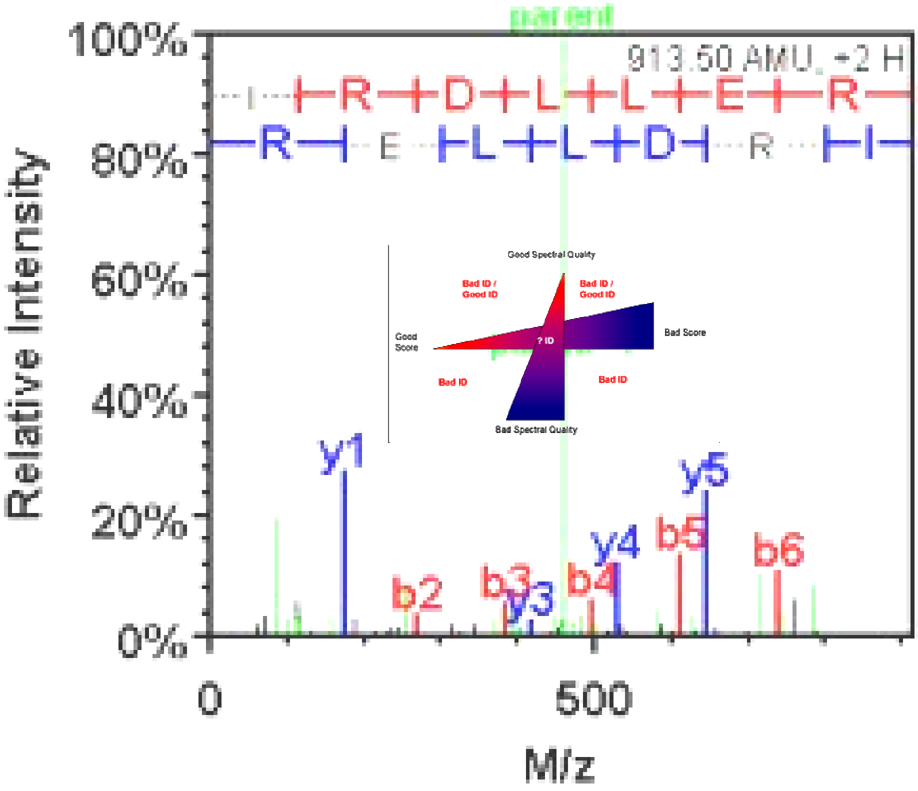


VALIDATING SEQUENCE ASSIGNMENTS FOR PEPTIDE FRAGMENTATION PATTERNS:

A PRIMER IN MS/MS SEQUENCE IDENTIFICATION

by Karen R. Jonscher, PhD

PROTEOME SOFTWARE

June, 2005

Copyright

Trademarks and Permissions

The Author

Karen R. Jonscher obtained her Ph.D. in Applied Physics from the California Institute of Technology, working with Drs. Leroy E. Hood and John R. Yates, III. Dr. Jonscher is a widely published author of research on protein identification with mass spectrometry.

Table of Contents

Validating Sequence Assignments for Peptide Fragmentation Patterns

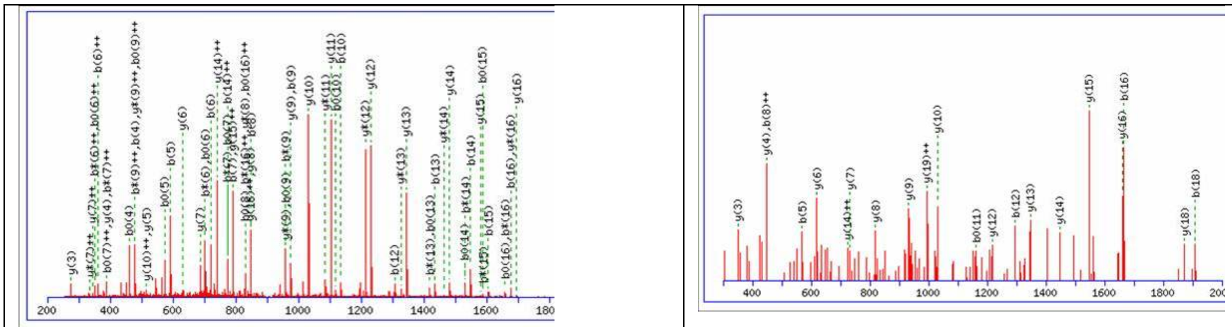
by Karen Jonscher, PhD

With the advent of high throughput proteomics, data is being generated at an astonishing rate. Validating peptide sequence assignments from database search engines is an increasingly important, but often overlooked, aspect of protein identification using tandem mass spectrometry.

Why Is Validating Data Important?

Every database search generates some false positive and false negative assignments, for a wide variety of reasons. We would like eliminate, or at least materially reduce, these incorrect hits. Decisions made downstream of peptide identification often involve a great deal of money for bioassays. In this era of tight funding, we must provide highly accurate data as the basis for these crucial decisions.

The simple example in Figure 1 demonstrates the potential problem. Two fragment ion spectra, from a multi-dimensional LC proteomics experiment using a quadrupole ion trap, were searched with Mascot. The two spectra had similar scores. But were they equivalent?



Mascot Score = 101 / Good ID
High quality spectrum with good signal-to-noise and all of the prominent ions assigned**.**

Mascot Score = 101 / Bad ID
Poor quality spectrum less prominent signal-to-noise. Fragment ions are most likely randomly assigned to noise peaks.

Figure 1: Same score, different credibility

In this tutorial you will learn about some of the factors important in low energy peptide fragmentation and how to use this information to accept or reject database search engine peptide sequence assignments.

How Do Peptides Break Apart?

In order to assess whether a sequence assignment is correct or not, it is important to understand how and why peptides break apart. Under low energy dissociation conditions, peptides primarily fragment at the C – N bond between amino acids, while higher energy instruments may also generate breaks at an internal C – C bond. Standard types of resulting fragment ions are characterized by the relative position of the break and by which end of the chain retains a positive charge. (See Appendix A: Peptide Chain Definitions for a review of terminology used in this section.)

- **b-type ions**: If the parent's peptide charge is retained on the N-terminal end of the peptide, the ion is known as a **b**-type ion.
- **y-type ions**: If the charge is retained on the C-terminal end, the ion is termed a **y**-type ion.

a-type ions: If the fragmentation energy in the instrument, especially triple quadrupoles or quadrupole-time-of-flight hybrids (Q-TOFs), is sufficient to generate C-C bond cleavage, the **b** ion loses CO. The resulting ions are known as **a**-type ions.

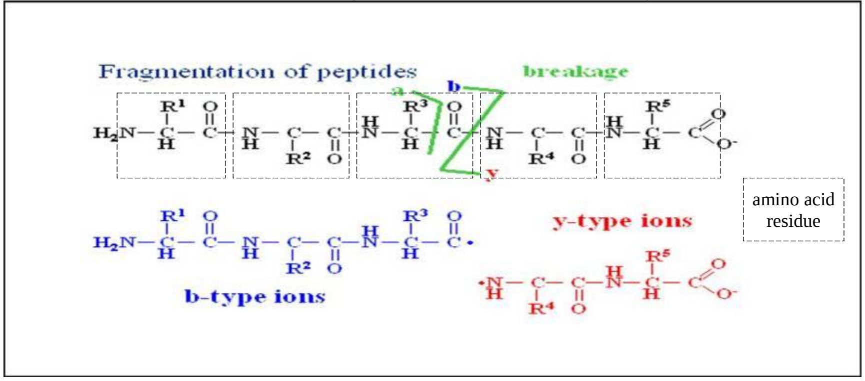


Figure 2: Peptide fragmentation

This paper examines only the more easily seen and interpreted **b** and **y** ions.

Peptides dissociate into nested sets of fragments

If A, B, C, D, E, F and G represent a peptide's full chain of amino acid residues, the peptide can dissociate to form any of the following **b** and **y** fragments. Note that b_1 is a complement to y_6. In other words, a break between the A and B amino acids creates both $\mathbf{b_1}$ and $\mathbf{y_6}$ fragments, and the sum of their two masses adds up to the total peptide mass.

N-terminal	C-terminal
H-A+ = b_1	y_1 = +G-OH
H-AB+ = b_2	y_2 = +FG-OH
H-ABC+ = b_3	y_3 = +EFG-OH
H-ABCD+ = b_4	y_4 = +DEFG-OH
H-ABCDE+ = b_5	y_5 = +CDEFG-OH
H-ABCDEF+ = b_6	y_6 = +BCDEFG-OH

Figure 3 shows a typical format for summarizing the full list of potential **b** and **y** ions that may result from a single peptide.

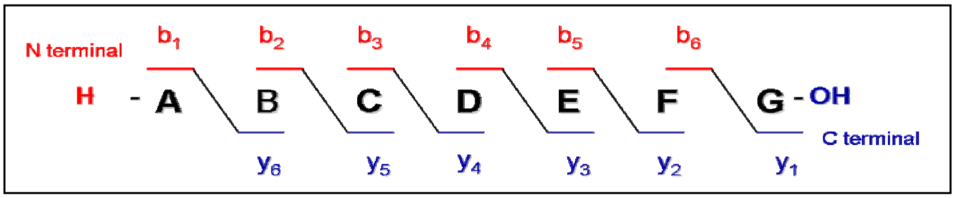


Figure 3: b and y ion diagram

Mass differences give clues to peptide sequence

Because successive **b** or **y** ions differ by one amino acid, we can deduce a peptide's sequence by calculating the difference in mass between spectrum peaks. If the mass difference corresponds to an amino acid residue, then that amino acid is assigned to the peak representing the difference.

Several general guidelines help in the identification process. (The general rules presented throughout this document are summarized and numbered sequentially in Appendix B: Sequence Identification Rules for ease of reference when working through the subsequent examples.)

- The **largest y ion** will appear anywhere between 57 to 186 amu below the total mass of the precursor ion. (largest **y** ion = precursor ion – $\mathbf{b_1}$.)
- The **smallest y ion, $\mathbf{y_1}$,** will appear at its single amino acid residue mass plus 19 amu.
  ($\mathbf{y_1}$ = one residue + the C-terminal OH + charge)
- The **largest b ion** will be below the precursor by {18 amu + its single residue mass}, i.e. in a range of 75 to 204 below the precursor mass. (largest **b** ion = precursor ion – $\mathbf{y_1}$.)
- The **smallest b ion, $\mathbf{b_1}$,** will be at the residue mass + 1.
  ($\mathbf{b_1}$ = one residue + the N-terminal H+)
- **Ion trap** results typically do not reveal $\mathbf{b_1}$, $\mathbf{y_1}$, or immonium ions, because of the low mass cut-off of the equipment.

With a good quality spectrum, it's possible to work successively through the peaks shown to determine the full peptide sequence. The ideal result would be a clearly labeled sequence of complete **b** and **y** ion fragmentation as shown in Figure 4.

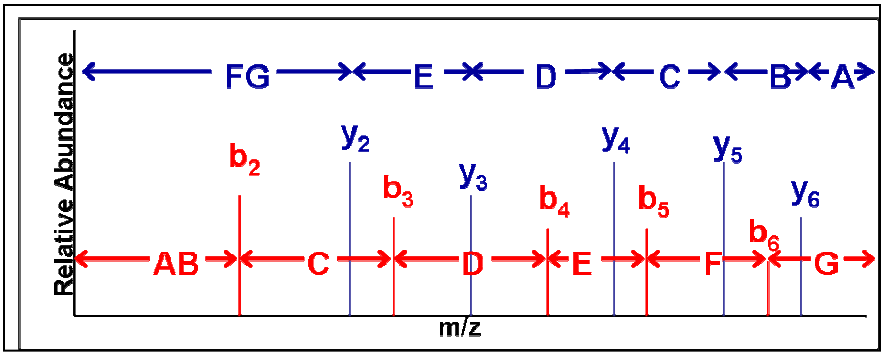


Figure 4: Fully identified peptide sequence

Fragmentation chemistry

Table 1 lists the amino acids with the molecular weights for their residues, structures of the side chains, and various chemical characteristics. Several additional chemical properties need to be considered when validating sequence assignments.

- **Basic** amino acids (R, H, L) must be present for a doubly-charged ion.
- Ion signal can be intense (high peaks) for cleavages C-terminal to **acidic** amino acids (D, E). These residues also tend to lose water and cyclize to randomly eject portions of the sequence.

- **Isobaric** amino acids (I vs L, or K vs Q) cannot be differentiated using low energy fragmentation instruments such as Q-TOF.
- **Serine** (**S**) and **threonine** (**T**) can lose water. Loss of water from **T** is particularly intense if the amino acid is near a terminal end of the peptide.
- If a peptide is tryptic, **y_1** will either be **lysine (K)** at 147 or **arginine (R)** at 175.
  (**y_1** mass = 19 + residue mass from Table 1.)
  Because trypsin causes cleavages at K and R, these amino acids are unlikely to be found at the N-terminal or in the interior of a tryptic precursor.

Table 1: Amino acid fragmentation characteristics and residue masses

Name	Characteristics	Symbol	Mass (-H_2O)	Side Chain
Alanine		**A**, Ala	71.08	CH_3-
Arginine	lose ammonia (-17)	**R**, Arg	156.19	$HN=C(NH_2)-NH-(CH_2)_3-$
Asparagine	lose ammonia (-17)	**N**, Asn	114.10	$H_2N-CO-CH_2-$
Aspartic acid		**D**, Asp	115.09	$HOOC-CH_2-$
Cysteine	lose H_2S = 34	**C**, Cys	103.15	$HS-CH_2-$
Glutamine	lose ammonia (-17)	**Q**, Gln	128.13	$H_2N-CO-(CH_2)_2-$
Glutamic acid		**E**, Glu	129.12	$HOOC-(CH_2)_2-$
Glycine	suppress b ions	**G**, Gly	57.05	H-
Histidine		**H**, His	137.14	$N=CH-NH-CH=C-CH_2-$ |__________|
Isoleucine		**I**, Ile	113.16	$CH_3-CH_2-CH(CH_3)-$
Leucine		**L**, Leu	113.16	$(CH_3)_2-CH-CH_2-$
Lysine		**K**, Lys	128.17	$H_2N-(CH_2)_4-$
Methionine	lose CH_3SH (-48)	**M**, Met	131.20	$CH_3-S-(CH_2)_2-$
Phenylalanine		**F**, Phe	147.18	Phenyl-CH_2-
Proline	suppress b ions; typically dominant	**P**, Pro	97.12	$-N-(CH_2)_3-CH-$ |_______|
Serine	lose water (-18)	**S**, Ser	87.08	$HO-CH_2-$
Threonine	lose water (-18), especially near end of peptide	**T**, Thr	101.11	$CH_3-CH(OH)-$
Tryptophan	abundant y ions	**W**, Trp	186.21	Phenyl-$NH-CH=C-CH_2-$ |___________|
Tyrosine		**Y**, Tyr	163.18	4-OH-Phenyl-CH_2-
Valine		**V**, Val	99.13	$CH_3-CH(CH_2)-$

blue rows → basic **pink rows → acidic**

Some pairs of amino acids add up to the mass of a different amino acid, as shown in Table 2. The same can happen with acetylated amino acids, a common modification.

Table 2: Ambiguous masses

Amino Acid Combination	Mass (amu)	Single Amino Acid	Acetylated Acid	Mass (amu)	Unmodified Amino Acid
G-G	114	N	Ac-G	99	V

G-A	128	K/Q	Ac-A	113	L/I
V-G	156	R	Ac-S	129	E
G-E	186	W	Ac-N	156	R
A-D	186	W			
S-V	186	W			

Manually Deducing a Sequence

To clarify how the identification process works, we will manually interpret an MS/MS spectrum generated by a quadrupole ion trap. The process we'll follow is termed ***de novo*** sequencing. In our example, the deconvoluted mass of the precursor was 1449.38, and the observed ion was doubly charged.

1. High-mass dominant peaks

We'll start by looking at the dominant peaks just below the mass of the precursor ion. We'll look for possible **y** ions between 57 and 186 amu below 1450, where the exact amount below 1450 is determined by the mass of the single residue *excluded* from this largest **y** ion (rule 2) and possible **b** ions in a range of 75 to 204 below 1450 (rule 1[1]). Figure 5 shows the peaks being identified at this step. Peaks marked with blue are **y** ions. Peaks marked with red are **b** ions, with their corresponding amino acids in the upper right. As ions are identified, we show their defining amino acids in the ladder across the top, above the spectrum ranges on which they are based.

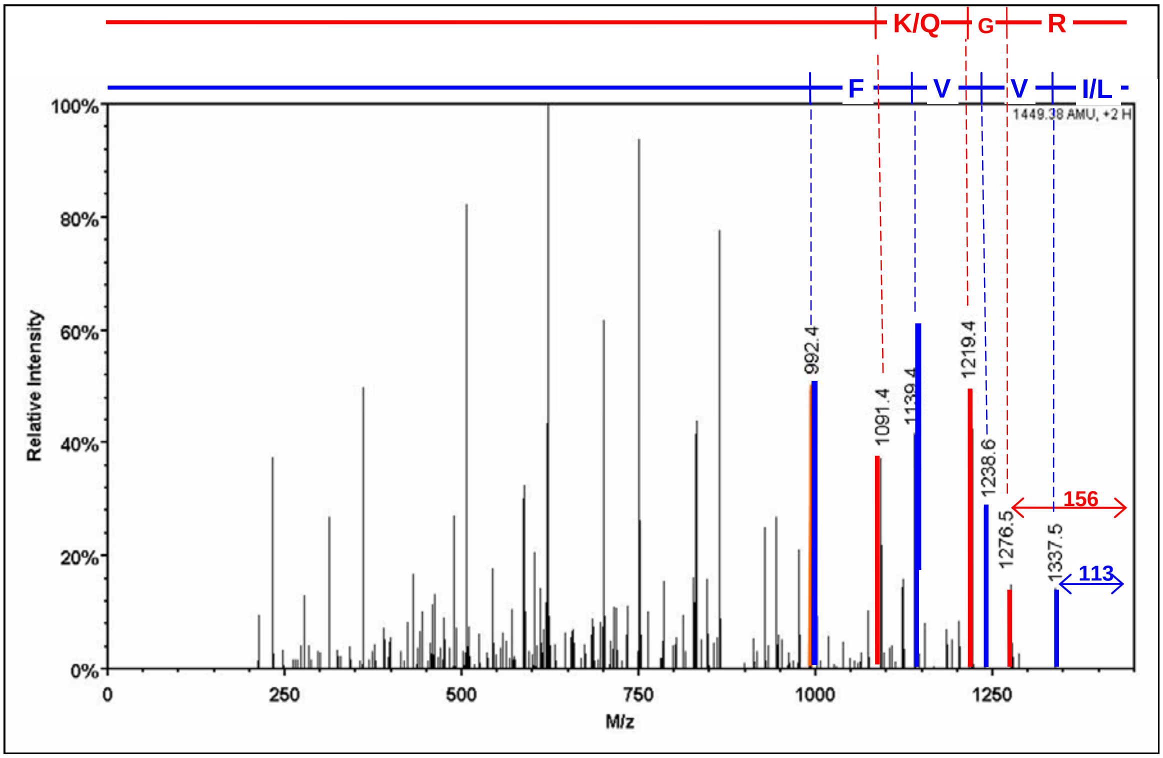


Figure 5: Spectrum ID: largest ions

[1] See Appendix B: Sequence Identification Rules for the numbered list of rules referenced throughout these examples.

- The first ion is at **1337.5**. 1450-1337=**113=I/L,** so we assign the largest **y** ion as either **leucine** or **isoleucine**. Note: our shorthand terminology here "assigns" the largest ions according to the amino acid where the cleavage occurred. Thus, the largest **y** ion ends with the residue (so far unidentified) just *before* the identified I/L in the sequence. Also, because we don/t know yet how many amino acids are in the precursor, we can't yet assign a numeric subscript to the largest fragment ions.
- The next ion is at **1276**. 1450-18-1276=**156=R**, so we assign the largest **b** ion as **arginine**. (rule 1) Because the sample was digested with trypsin, we would expect lysine or arginine just after the cleavage at the end of the largest **b** ion. (rule 5)
- Next we look at the ion at **1238**. We assume it is a **y** ion, because it is less than one full amino acid residue in mass from our last **b** ion, 1276. Noting that 1337-1238=**99=V**, we assign the next **y** ion as **valine.**
- Continuing in this manner, we find b ions at:
    - 1276-**1219=57=G**, glycine and
    - 1219-**1091=128= K/Q**, either lysine or glutamine. This technique can't distinguish between those two amino acids with nearly the same mass (rule 4).
- And y ions at:
    - 1238-**1139=99=V**, **valine**
    - 1139-**992=147=F, phenylalanine**.

2. Complementary low-mass b and y ions

To verify the high mass **y** ion assignments, we look for the complimentary low mass **b** ions. Since the data is from an ion trap, we will probably not see **b$_1$** (rule 6). Therefore, we start by looking for **b$_2$**.

- Since the largest y ion was either leucine or isoleucine and the next y ion in the series was valine, we look for the complementary ion at 113+1+99=**213=b$_2$** (rule 1 for the total mass of a **b** ion). Thus our first visible **b**-ion peak represents two residues, I/L + V.
- The next **b** ion should result from addition of another valine (paralleling the reverse C-terminal series of **y**-ion components), therefore we'd expect a signal at 213+99=**312=b$_3$**, and we do find a peak there.
- When phenylalanine is added, we find the b$_4$ ion at 312+147=**459=b$_4$**.

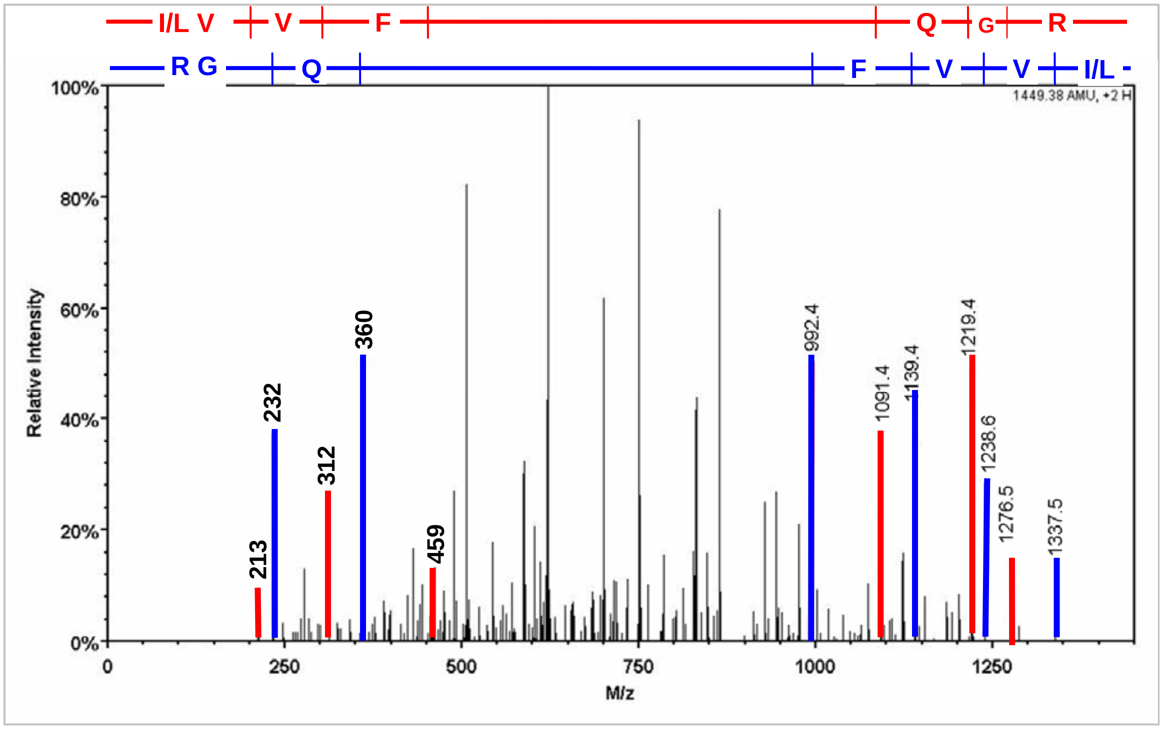


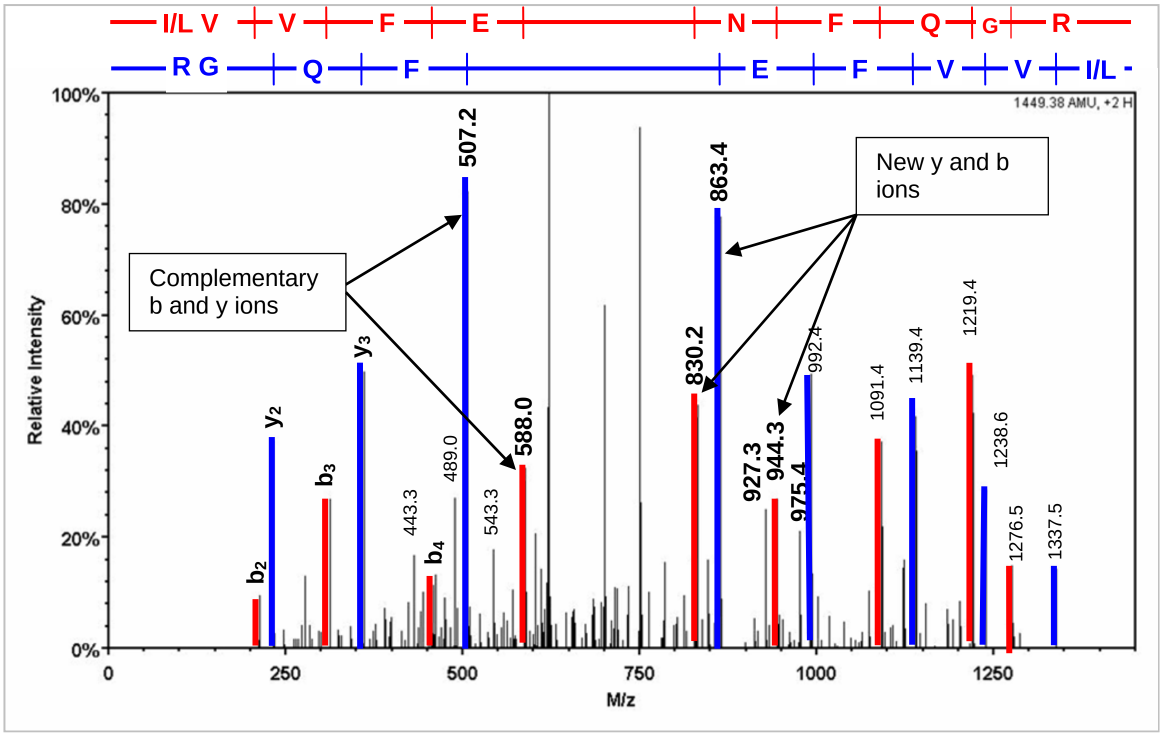
Figure 6: Spectrum ID: complementary low-mass b and y ions

Similarly, to verify the high mass **b** ion assignments, we look for the complimentary low mass **y** ions. Since the data is from an ion trap, we will probably not see **y$_1$**. However the assignment of arginine makes sense given that the sample was digested using trypsin (rule 5). Therefore, we start by looking for **y**$_2$.

- Assuming R is **y$_1$**, we look for a signal resulting from the addition of Glycine, G, at 156+19+57=**232=y$_2$**. We add the 19 amu to account for the carboxyl group on the end of the **y** ions (see Figure 2). Thus, our first visible **y** peak represents R + G.
- The next complementary **y** ion should be at 232+128=**360=y$_3$** resulting from the addition of either lysine, K, or glutamine, Q. Q is the most likely of the two. That's because the sample was digested using trypsin, and we would expect a lysine only at the C-terminal end (rule 5).

3. Finding the next largest ions

We'll now proceed to work further down from the top, identifying mass differences for more of the high-mass dominant spectra.



Complementary b and y ions

New y and b ions

Figure 7: Spectrum ID: finding the next largest ions

We look for the next largest **b** ion below 1091. There are three choices:

1091-975=116
1091-944=**147=F, phenylalanine**
1091-927=164

Only the signal at **944** corresponds to an amino acid residue mass, so the next **b** ion is **F.**

- The next largest **y** ion will appear below 992 Since **992-863=129=E**, we assign the next **y** ion as **glutamic acid**.
- The next prominent peak should correspond to a **b** ion so we look below 944 and note that 944-**830=114=N, asparagine.**
- Validating our high mass ion assignments, we expect to find **b_5** resulting from addition of glutamic acid at 459+129**=588=b_5** and **y_4** resulting from addition of phenylalanine**,** 360+147**=507=y_4**.

4. Completing the sequence

As we can see from Figure 7, our identification of high-mass ions and their complements has almost filled in all the prominent spectra. The next largest **y** ion will appear below 863.

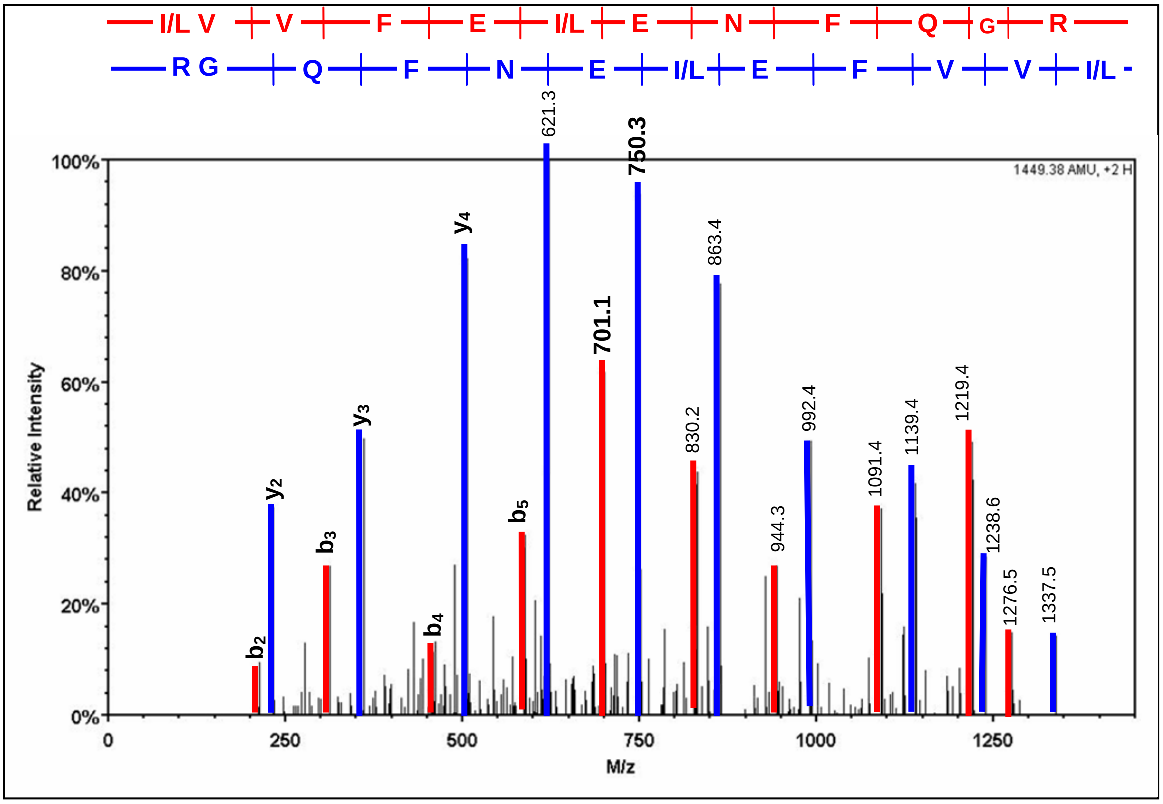


Figure 8: Spectrum ID; completing the sequence

- Since 863-**750=113=I/L**, we assign the next **y** ion as leucine/isoleucine.
- The next prominent peak should correspond to a **b** ion so we look below 830 and note that 830-**701=129=E, glutamic acid.**
- The next **y** ion will appear below 750. Since 750-**621=129=E**, we assign the next **y** ion as glutamic acid. Because this is complementary to the **b** ion we just assigned (701 + 750 **.** 1450, and the most recent residue to be dropped from a **y** ion, E = the most recent residue added to a **b** ion). So our sequence is complete!

5. Final sequence

With the full sequence completed, we can now fill in all the **y** and **b** ion numbers.

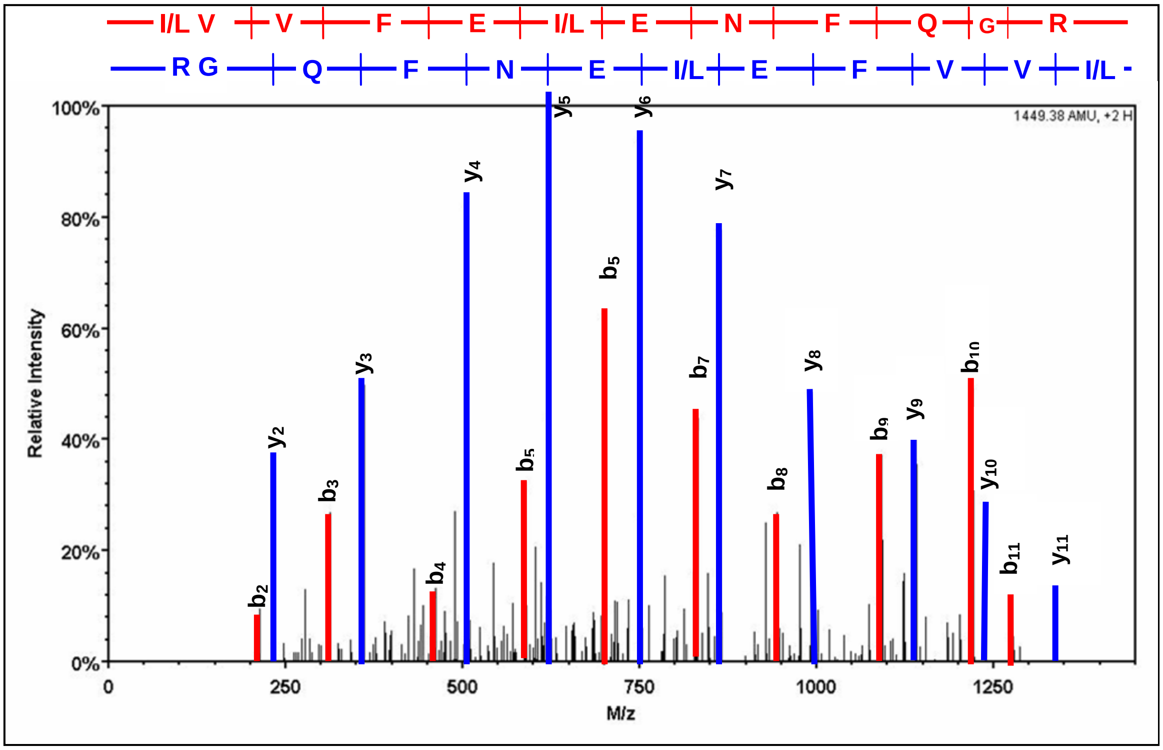


Figure 9: The final identified spectrum

In this example we have observed a complete series of complementary **b** and **y** ions and identified all the prominent fragments. Now we'll look at some alternate identifications that might have arisen for this spectrum.

Alternate Identifications: Sequence Database Searching

Because of ambiguities in identifying peaks, several possible sequences may be identified for a single spectrum. Obtaining clear results from *de novo* sequencing is particularly difficult if the spectrum does not display clear peaks for all the ions in the **y** and **b** ladders. An alternate technique, used by search engines such as Mascot and SEQUEST, is to search the spectrum peaks against databases of known peptide sequences. These search engines typically identify several possible matching sequences; a calculated score for each reflects the likelihood of the match being correct. Here we will look at another possible interpretation of the spectrum just completed.

Figure 10 shows this spectrum, but with one of the lower-ranked sequence identifications from SEQUEST. Although most of the prominent ions were assigned, they were not assigned to **b** and **y** ions, but rather to water loss and other ions that are generally less abundant. Thus, this tentative identification is unlikely to be correct

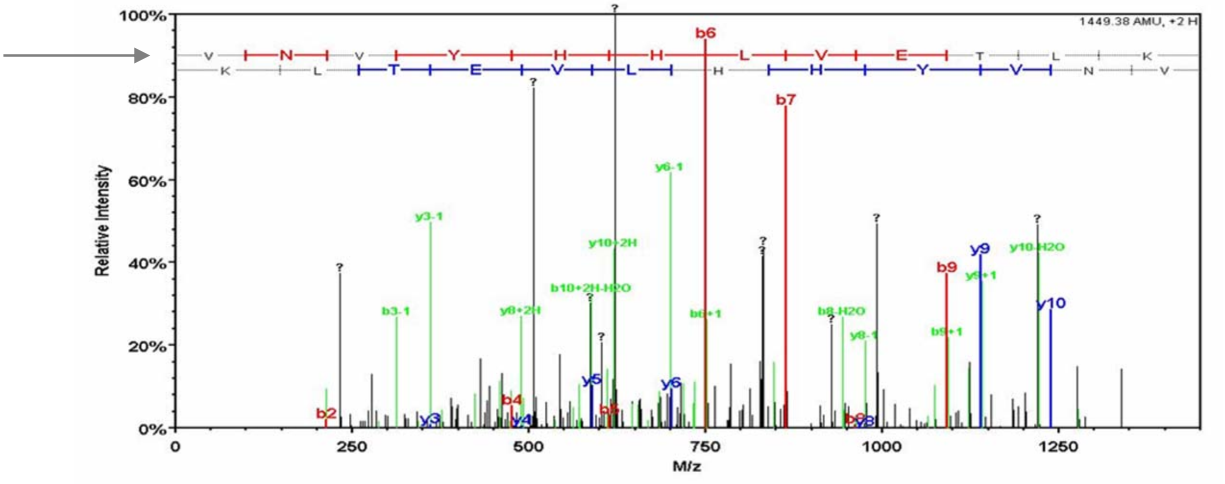


Figure 10: A low confidence identification

SEQUEST's output file, in Table 3, shows twelve possible peptides for this spectrum, ranked from most likely to least likely. The program calculates XCorr scores to reflect identification quality, the higher the score the better the ID. Experience has shown that XCorr scores above 1.9, 2.2, or 3.7 (for peptides with 1, 2, or 3 charges respectively) often correspond to correct peptide identification. The first row of the table, with an XCorr of 4.7 matches the identification we manually determined in the previous section. The score here is well above the 2.2 threshold for a doubly charged peptide.

Table 3: SEQUEST output file

Red circles show sequence identified in Figure 10. Green circles show sequence identified in Figure 11.

```
      Rank
 #    / Sp   Id#   (M+H)+    deltCn   XCorr    Sp     Ions    Reference          Peptide
  1   1 / 1   0   1450.769   0.000   4.757  2559.1   20/22   CRB1_HUMAN     R.LVVFELENFQGR.R
  2   2 / 2   0   1451.753   0.025   4.636  2541.5   20/22   CRB1_HUMAN     R.LVVFELEN*FQGR.R
  3   3 / 2   0   1451.753   0.057   4.485  2541.5   20/22   CRB1_HUMAN     R.LVVFELENFQ*GR.R
  4   4 / 3   0   1452.737   0.280   3.423  2036.0   18/22   CRB1_HUMAN     R.LVVFELEN*FQ*GR.R
  5   5 / 6   0   1451.757   0.404   2.836  1057.7   15/22   TP3B_HUMAN     K.LN*M#VKFLQ*VEGR.G
  6   6 / 5   0   1450.773   0.462   2.560  1426.3   17/22   TP3B_HUMAN     K.LN*M#VKFLQVEGR.G
  7   7 /13   0   1450.655   0.465   2.543   840.0   13/22   NEUM_HUMAN     R.TKQ*VEKN*DDDQ*K.I
  8   8 /13   0   1449.671   0.475   2.496   840.0   13/22   NEUM_HUMAN     R.TKQ*VEKNDDDQ*K.I
  9   9 / 5   0   1449.671   0.476   2.494   817.4   13/22   NEUM_HUMAN     R.TKQ*VEKN*DDDQK.I
 10  10/11    0   1451.706   0.504   2.359   843.6   13/22   ING_HUMAN      K.S]VETIKEDM#NVK.F
 11  11/16    0   1451.801   0.515   2.309   817.3   14/22   GGT5_HUMAN     R.VNVYHHLVETLK.F
 12  12/ 4    0   1451.749   0.518   2.293  1458.5   16/20   DESP_HUMAN     R.LTYEIEDEKRR.R
```

When we compare the low confidence identification in Figure 10 with SEQUEST's top-ranked identification, shown in Figure 11, we see that the better ID assigns **b** and **y** ions to all of the prominent peaks, with water losses now accounting for many of the lower abundance peaks (the green peaks in the diagram). Both those characteristics are signs of a correct identification.

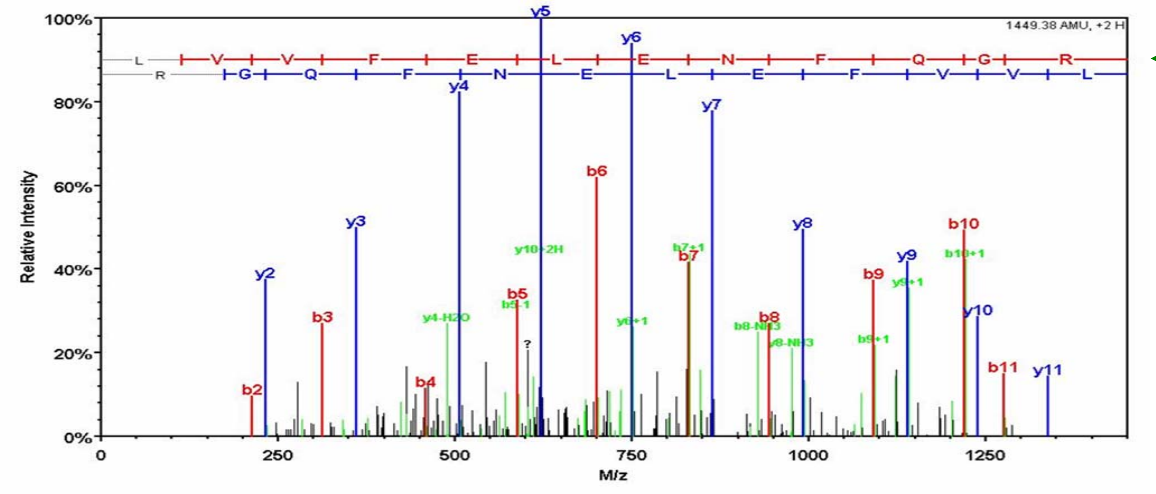
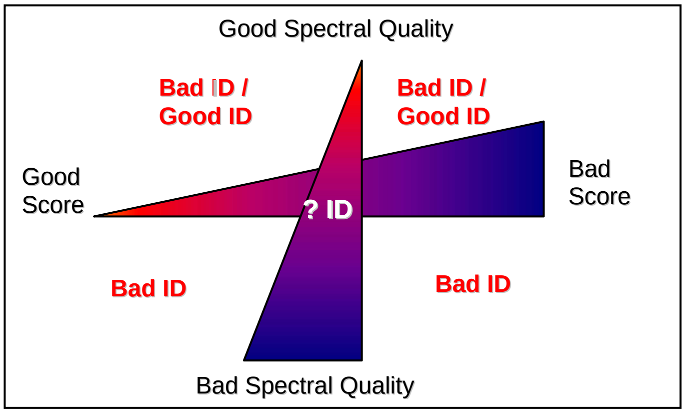


Figure 11: Recap of correct identification

Different proteomics programs use different conventions in displaying identified spectra. This figure is from Scaffold, which shows **y** ions in blue, **b** ions in red, and modifications such as loss of water in green. The summary sequence ladder across the top is another characteristic of Scaffold output.

Score vs. Spectral Quality



As we have seen, a high score may be an indicator that an identification is correct. However, this does not hold true in all cases. In the next few examples, we will see instances where good scores actually corresponded to bad identifications, and bad scores corresponded to good identifications. The examples demonstrate the importance of spectral quality and review cases of questionable identification. Each looks at proteomics data from a Mascot search.

Example 1: Great Score / Nice Spectrum / Bad ID

This mass spectrum has a **good distribution of fragment ions** and has **good signal to noise** ratio. Oftentimes, good quality spectra like this provide good search results.

1. IPI00001661 **Mass:** 45425 **Total score:** 111 **Peptides matched:** 1 Tax_Id=9606

Regulator of chromosome condensation

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Rank	Peptide
1	950.77	1899.53	1899.00	0.53	0	111	1	VVQVSAGDSHTAALTDDGR

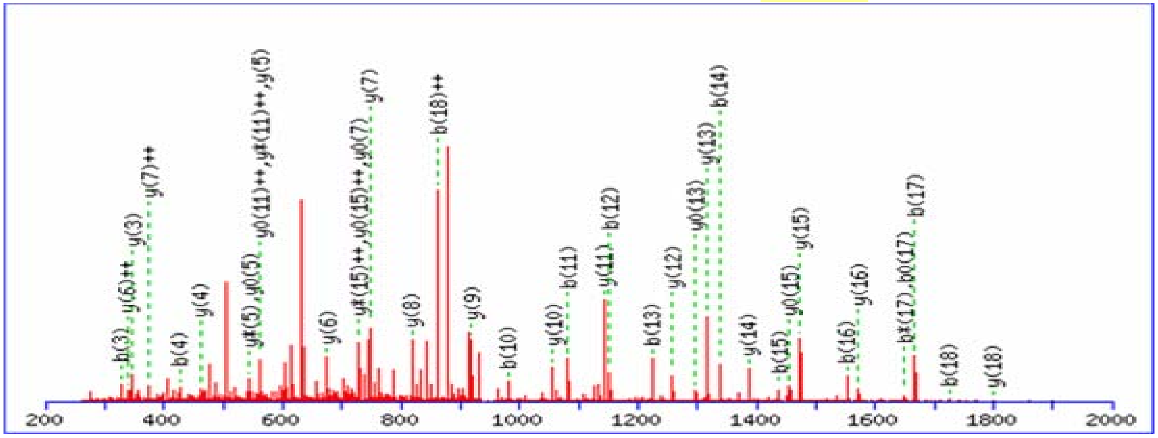


Figure 12: Mascot high score search results

The **Mascot** score is 111. Scores over 40 or 50 typically generate correct identifications. The **score** is well beyond the 95% confidence level and is well separated from the other possibilities, generally a positive indicator of a correct ID.

The masses of the possible sequence ions are summarized in the data table below. Masses labeled in red were observed in the experiment. We see a fairly long contiguous run of for the **y** ions and a shorter run for the **b** ions. There is some overlap of the **b** and **y** ions, however, so we have a complementary set.

Table 4: Mascot data table 1 (m/Z values)

#	b	b++	b*	b*++	b0	b0++	Seq.	y	y++	y*	y*++	y0	y0++	#
1	100.1	50.6					V							19
2	199.3	100.1					V	1801	901	1784	892	1783	892	18
3	327.4	164.2	310.4	155.7			Q	1702	851	1685	843	1684	842	17
4	426.5	213.8	409.5	205.3			V	1574	787	1557	779	1556	778	16
5	513.6	257.3	496.6	248.8	495.6	248.3	S	1474	738	1457	729	1456	729	15
6	584.7	292.9	567.7	284.3	566.7	283.8	A	1387	694	1370	686	1369	685	14
7	641.8	321.4	624.7	312.9	623.7	312.4	G	1316	659	1299	650	1298	650	13
8	756.8	378.9	739.8	370.4	738.8	369.9	D	1259	630	1242	622	1241	621	12
9	843.9	422.5	826.9	413.9	825.9	413.5	S	1144	573	1127	564	1126	564	11
10	981.1	491.0	964.0	482.5	963.0	482.0	H	1057	529	1040	521	1039	520	10
11	1082.2	541.6	1065.1	533.1	1064.1	532.6	T	920	460	903	452	902	451	9
12	1153.2	577.1	1136.2	568.6	1135.2	568.1	A	819	410	802	401	801	401	8
13	1224.3	612.7	1207.3	604.2	1206.3	603.7	A	748	374	731	366	730	365	7
14	1337.5	669.2	1320.5	660.7	1319.5	660.2	L	677	339	660	330	659	330	6
15	1438.6	719.8	1421.6	711.3	1420.6	710.8	T	564	282	547	274	546	273	5
16	1553.7	777.3	1536.6	768.8	1535.7	768.3	D	462	232	445	223	444	223	4
17	1668.8	834.9	1651.7	826.4	1650.7	825.9	D	347	174	330	166	329	165	3
18	1725.8	863.4	1708.8	854.9	1707.8	854.4	G	232	117	215	108			2
19							R	175	88	158	80			1

Column headings as in standard Mascot:

b or y: unmodified b or y ions, with single charge
++: doubly charged ion (thus shown at half the m/Z of the unmodified ion)
* : lost ammonia (unmodified ion – 17, for NH_3)
0 : lost water (unmodified ion – 18, for H_20)

Despite the many of the factors which seemed to lead to a correct identification, it is **wrong**. It fails to account for three of the dominant ions.

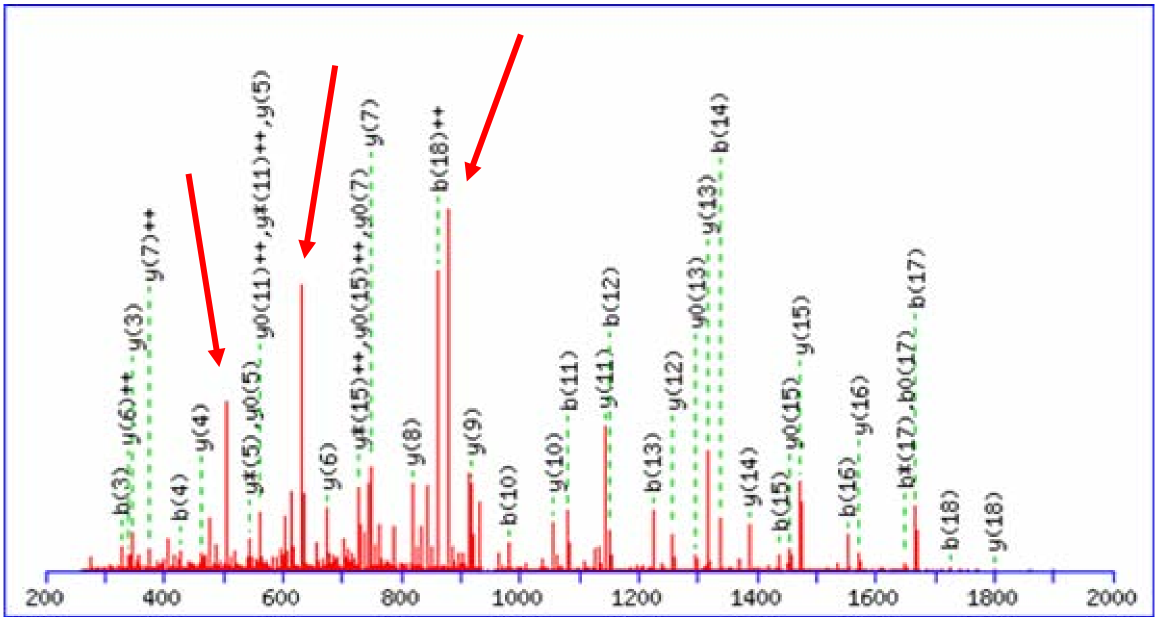


Figure 13: Unidentified dominant ions

Since this is a **good quality** spectrum, it would be worth pursuing other interpretation options. For one, the peptide may be modified and it would be worth re-searching the data using a database including modifications such as **phosphorylation** or **glycosylation**, among others. Several searches may be required. ***De novo* searching** would also be a possible approach if modification searches do not provide acceptable results.

Example 2: Good Spectrum / Bad Score / Bad ID

In this case, the spectrum again has good signal to noise and well-separated fragment ions. However, most of the dominant ions are unidentified, and the Mascot score of 29 is below generally accepted thresholds.

No significant hits to report **Unassigned queries:** (no details means no match)

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Rank	Sequence
1	868.07	1734.13	1730.98	3.15	2	29	1	KGVASTDNTLIARSLGK

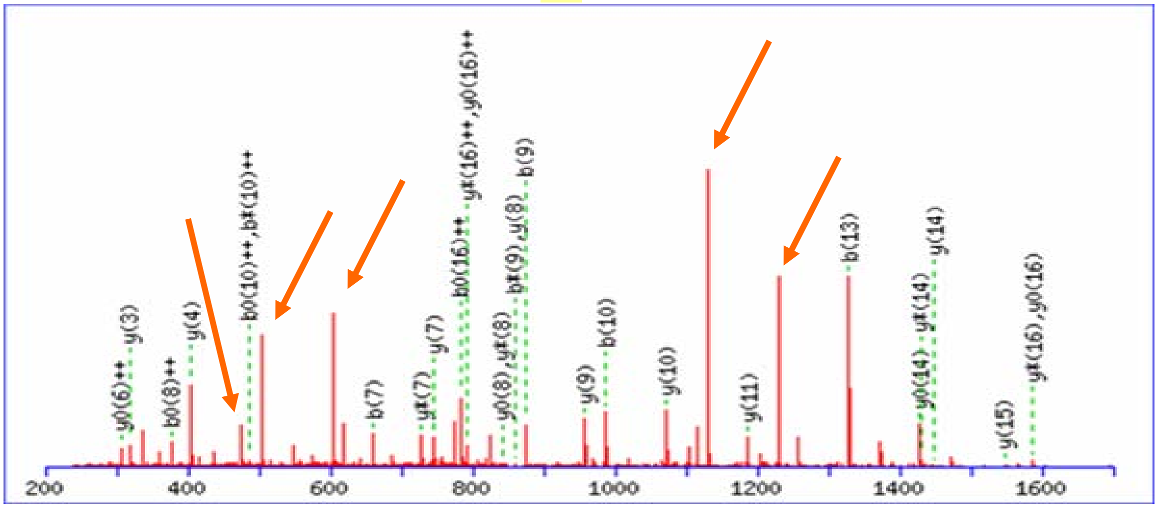


Figure 14: Lack of dominant ion identification

Table 5 shows **only short runs** of contiguous sequence. There is **little complementarity** between the **b** and **y** ions.

Table 5: Mascot data table 2 (m/Z values)

#	b	b++	b*	b*++	b0	b0++	Seq.	y	y++	y*	y*++	y0	y0++	#
1	129.2	65.1	112.2	56.6			K							17
2	186.2	93.6	169.2	85.1			G	1603.8	802.4	**1586.8**	**793.9**	**1585.8**	**793.4**	16
3	285.4	143.2	268.3	134.7			V	**1546.8**	773.9	1529.7	765.4	1528.8	764.9	15
4	356.5	178.7	339.4	170.2			A	**1447.6**	724.3	**1430.6**	715.8	**1429.6**	715.3	14
5	443.5	222.3	426.5	213.8	425.5	213.3	S	1376.6	688.8	1359.5	680.3	1358.5	679.8	13
6	544.6	272.8	527.6	264.3	526.6	263.8	T	1289.5	645.2	1272.4	636.7	1271.5	636.2	12
7	**659.7**	330.4	642.7	321.9	641.7	321.4	D	**1188.4**	594.7	1171.3	586.2	1170.4	585.7	11
8	773.8	387.4	756.8	378.9	755.8	**378.4**	N	**1073.3**	537.1	1056.3	528.6	1055.3	528.1	10
9	**874.9**	438.0	**857.9**	429.5	856.9	429.0	T	**959.2**	480.1	942.2	471.6	941.2	471.1	9
10	**988.1**	494.6	971.1	**486.0**	970.1	**485.5**	L	**858.1**	429.5	**841.0**	421.0	**840.1**	420.5	8
11	1101.3	551.1	1084.2	542.6	1083.2	542.1	I	**744.9**	373	**727.9**	364.5	726.9	364.0	7
12	1172.3	586.7	1155.3	578.2	1154.3	577.7	A	631.8	316.4	614.7	307.9	613.7	**307.4**	6
13	**1328.5**	664.8	1311.5	656.2	1310.5	655.8	R	560.7	280.8	543.6	272.3	542.7	271.8	5
14	1415.6	708.3	1398.6	699.8	1397.6	699.3	S	**404.5**	202.8	387.5	194.2	386.5	193.7	4
15	1528.8	764.9	1511.7	756.4	1510.7	755.9	L	**317.4**	159.2	300.4	150.7			3
16	**1585.8**	**793.4**	1568.8	784.9	1567.8	**784.4**	G	204.3	102.6	187.2	94.1			2
17							K	147.2	74.1	130.2	65.6			1

With **dominant ions unidentified** the sequence assignment is obviously **incorrect** (rule 20). Since the spectrum is of good quality, the next step should be to consider modifications, which is beyond the scope of this paper.

Example 3: Good Spectrum / Bad Score / Good ID

The spectrum in Figure 15 shows good signal to noise ratio (rule 17) and a clear spread of peaks (rule 18). The Mascot score, however, was only 31 (rule 21).

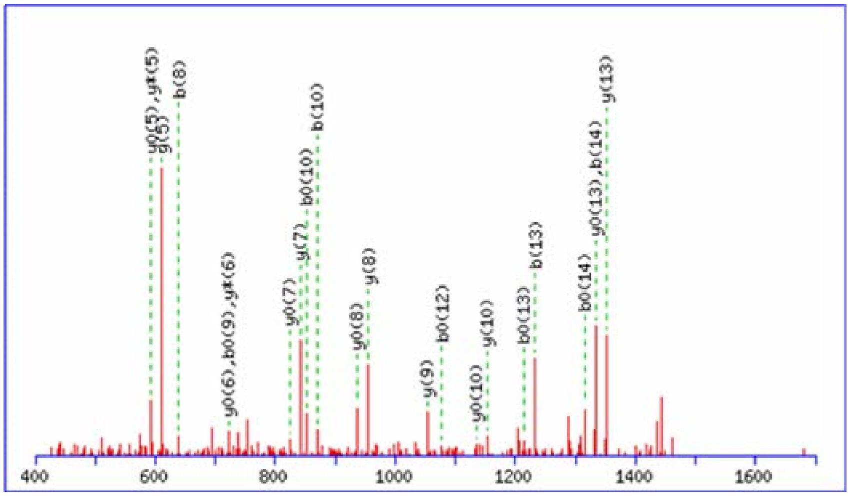


Figure 15: All dominant peaks accounted for

Despite the low score, the identification does account for all prominent ions as **y** and **b** ions (rule 20). Looking at the Mascot data table, we see that the assignments are also consistent with plausible chemistry. The largest peak corresponds to $\mathbf{y_5}$ (mass 611) cleavage at proline (rule 8). There is no water loss ("0" postscript) for **b** ions until $\mathbf{b_9}$, when threonine appears, and the water loss for **y** ions also occurs in many of the fragments including T (rule 11). It is possible that the presence of basic histidine (H) and acidic glutamic acid (E) inhibit water loss at $\mathbf{y_3}$ and $\mathbf{y_4}$.

Table 6: Mascot data table 3 (m/Z values)

#	b	b0	Seq	y	y*	y0	#
1	72.1		A				15
2	129.1		G	1409.6	1392.6	1391.6	14
3	186.2		G	1352.5	1335.5	1334.5	13
4	257.3		A	1295.5	1278.5	1277.5	12
5	328.4		A	1224.4	1207.4	1206.4	11
6	427.5		V	1153.3	1136.3	1135.3	10
7	526.6		V	1054.2	1037.2	1036.2	9
8	639.8		I	955.1	938.0	937.0	8
9	740.9	722.9	T	841.9	824.9	823.9	7
10	870.0	852.0	E	740.8	723.8	722.8	6
11	967.1	949.1	P	611.7	594.7	593.7	5
12	1096.2	1078.2	E	514.6	497.5	496.5	4
13	1233.4	1215.4	H	385.4	368.4	367.4	3
14	1334.5	1316.5	T	248.3	231.3	230.3	2
15			K	147.2	130.2		1

A SEQUEST search provided the same identification as the Mascot search (rule 24). Hence, despite the low score, this is likely a correct assignment.

Example 4: Bad Spectrum / Bad Score / Bad ID

Figure 16 presents a very weak spectrum. It has virtually no baseline (rule 17), and we must assume that we are simply seeing noise. The ions are likely assigned by random chance.

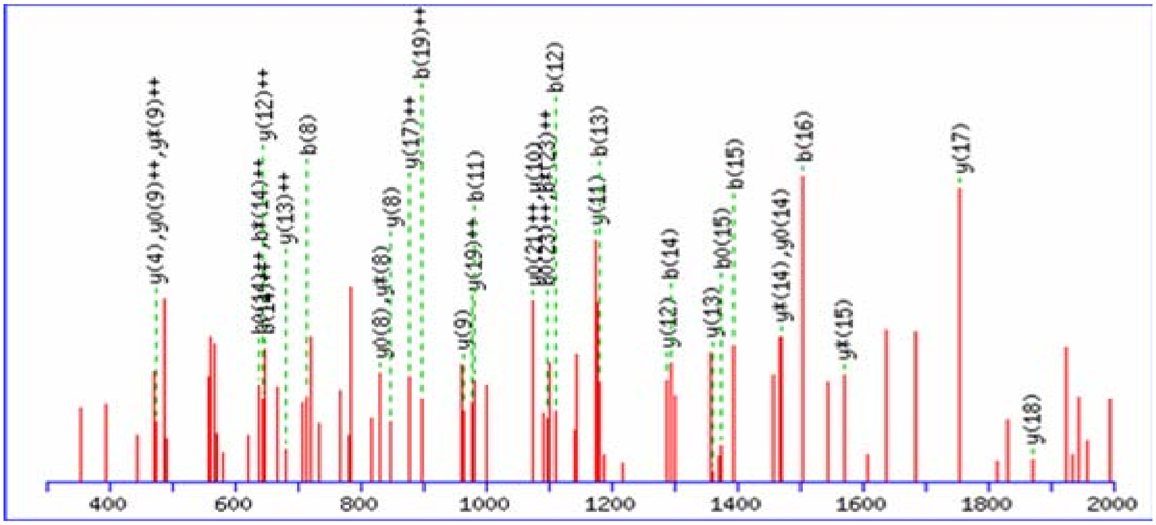


Figure 16: Noisy spectrum

The Mascot score was only 18 (rule 21), and the data table below raises a number of questions.

Table 7: Mascot data table 4 (m/Z values)

#	b	b++	b*	b*++	b0	b0++	Seq	y	y++	y*	y*++	y0	y0++	#
1	58	30					G							25
2	157	79					V	2414	1207	2397	1199	2396	1198	24
3	228	115					A	2315	1158	2298	1149	2297	1149	23
4	299	150					A	2244	1122	2226	1114	2226	1113	22
5	413	207					L	2172	1087	2155	1078	2154	**1078**	21
6	514	257			496	248	T	2059	1030	2042	1022	2041	1021	20
7	601	301			583	292	S	1958	**980**	1941	971	1940	971	19
8	**716**	358			698	349	D	**1871**	936	1854	928	1853	927	18
9	813	407			795	398	P	**1756**	**879**	1739	870	1738	870	17
10	884	442			866	433	A	1659	830	1642	821	1641	821	16
11	**983**	492			965	483	V	1588	794	**1571**	786	1570	785	15
12	**1111**	556	1094	548	1093	547	Q	1489	745	**1472**	736	**1471**	736	14
13	**1182**	592	1165	583	1164	583	A	**1361**	**681**	1344	672	1343	672	13
14	**1295**	**648**	1278	**640**	1277	**639**	I	**1289**	**645**	1272	637	1271	636	12
15	**1395**	698	1378	689	**1377**	689	V	**1176**	589	1159	580	1158	580	11
16	**1508**	754	1491	746	1490	745	L	**1077**	539	1060	531	1059	530	10
17	1623	812	1606	803	1605	803	D	**964**	483	947	**474**	946	**474**	9
18	1724	862	1707	854	1706	853	T	**849**	425	**832**	416	**831**	416	8
19	1795	**898**	1778	890	1777	889	A	748	374	731	366	730	365	7
20	1882	942	1865	933	1864	933	S	677	339	660	330	659	330	6
21	1997	999	1980	991	1979	990	D	590	295	573	287	572	286	5
22	2096	1049	2079	1040	2078	1040	V	**475**	238	458	229	457	229	4
23	2210	1105	2192	**1097**	2191	**1096**	L	375	188	358	180	357	179	3
24	2325	1163	2308	1154	2307	1154	D	262	132	245	123	244	123	2
25							K	147	74	130	66			1

Although there are some **complementary runs** of contiguous sequence ions (rule 19), it is **unlikely** they are **significant**, given the quality of the mass spectrum.

- We do see a dominant fragment ion at the $\mathbf{y_{17}}$ proline (P), identified in the box in the above table (rule 8). However **other large signals do not correspond to expected cleavages** C-terminal to the acidic amino acids (rule 9). The arrows show these **b** ions that we would expect to generate large signals.

- The assignment of **doubly-charged ions** without the presence of a basic ion is also highly unlikely (rule 10). This identification is incorrect and this spectrum could likely be discarded.

Example 5: Marginal Spectrum / Marginal Score / Questionable ID

The spectrum in this case has a few clear peaks and little noise, but fragmentation is limited. The most abundant peak results from loss of water from serine, S, not proline cleavage as expected (rule 8). The doubly-charged $\mathbf{y_8}$ ion is reasonable, given the presence of basic arginine, R (rule 10). However, $\mathbf{b_2}$ is shown doubly charged with no basic amino acids, which is highly unlikely.

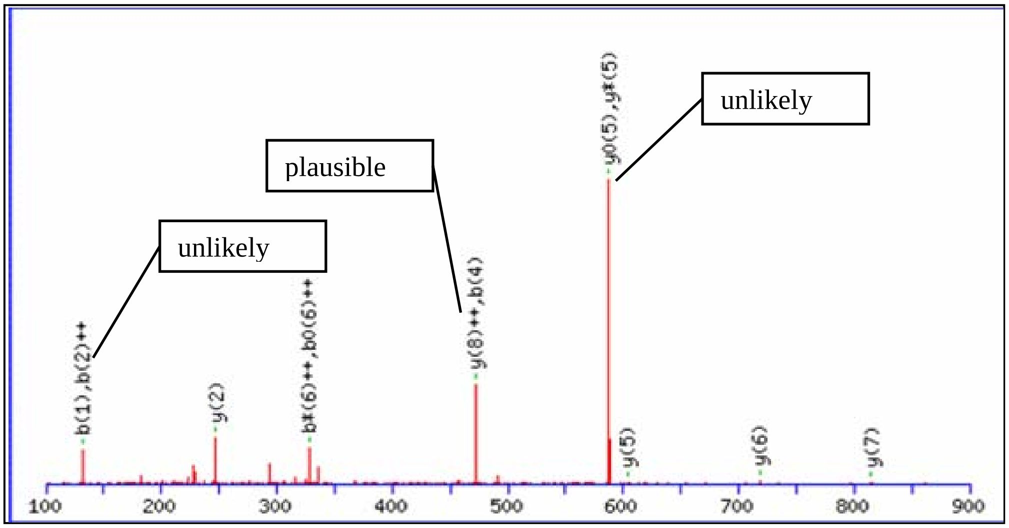


Figure 17: Limited fragmentation

The Mascot score is 38, just below the typical cut-off for a probable match (rule 21). This particular peptide identification could possibly be disregarded if the protein assignment was confirmed by another peptide.

Table 8: Mascot data table 5 (m/Z values)

#	b	b++	b*	b*++	b0	b0++	Seq	y	y++	y*	y*++	y0	y0++	#
1	**132.2**	66.6					M							9
2	261.3	**131.2**			243.3	122.2	E	945.1	**473.0**	928.0	464.5	927.1	464.03	8
3	358.4	179.7			340.4	170.7	P	**816.0**	408.5	798.9	400.0	797.9	399.47	7
4	**472.5**	236.8	455.5	228.3	454.5	227.8	N	718.8	359.9	701.8	351.4	700.8	350.91	6
5	559.6	280.3	542.6	271.8	541.6	271.3	S	**604.7**	302.9	**587.7**	294.4	**586.7**	293.86	5
6	672.8	336.9	655.7	**328.4**	654.8	**327.9**	L	517.7	259.3	500.6	250.8	499.6	250.32	4
7	829.0	415.0	811.9	406.5	811.0	406.0	R	404.5	202.8	387.5	194.2	386.5	193.74	3
8	930.1	465.5	913.0	457.0	912.1	456.5	T	**248.3**	124.7	231.3	116.1	230.3	115.65	2
9							K	147.2	74.1	130.2	65.6			1

Example 6: Good Spectrum / Good Score / Good ID

This case presents a very clean spectrum. It displays a complete b and y series with very little noise. The ion chemistry is plausible, with abundant ions at the aspartic acid, D, (**b$_9$**) and tryptophan, W, (**y$_4$**) cleavages, as would be expected (rules 9 and 13).

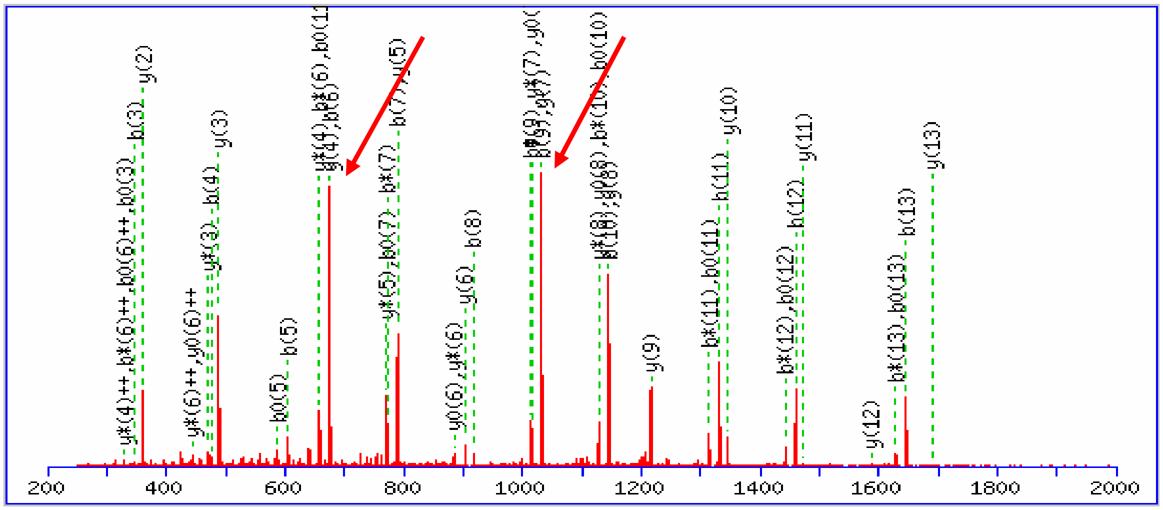


Figure 18: Clean, complete spectrum

The Mascot score is a respectable 79 (rule 21), and the identification seems very likely correct.

Table 9: Mascot data table 6 (m/Z values)

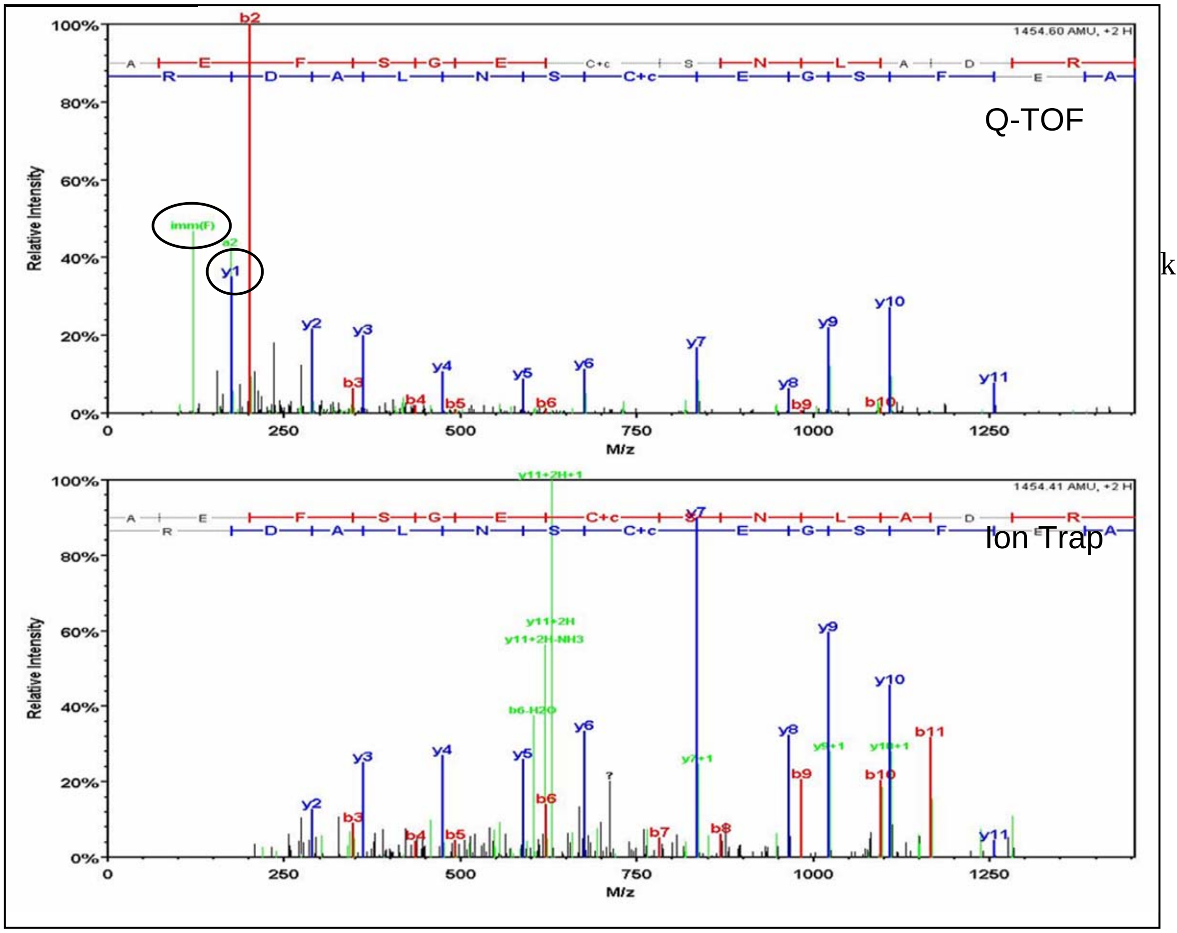
#	b	b++	b*	b*++	b0	b0++	Seq	y	y++	y*	y*++	y0	y0++	#
1	132.2	66.6					M							14
2	233.3	117.2			215.3	108.2	T	**1689.8**	845.4	1672.8	836.9	1671.8	836.4	13
3	**348.4**	174.7			**330.4**	165.7	D	**1588.7**	794.9	1571.7	786.4	1570.7	785.9	12
4	**476.5**	238.8	459.5	230.3	458.5	229.8	Q	**1473.6**	737.3	1456.6	728.8	1455.6	728.3	11
5	**605.6**	303.3	588.6	294.8	**587.6**	294.3	E	**1345.5**	673.3	1328.5	664.7	1327.5	664.3	10
6	**676.7**	338.9	**659.7**	**330.4**	**658.7**	**329.9**	A	**1216.4**	608.7	1199.4	600.2	1198.4	599.7	9
7	**789.9**	395.4	**772.9**	386.9	**771.9**	386.4	I	**1145.3**	573.2	**1128.3**	564.6	**1127.3**	564.2	8
8	**918.0**	459.5	901.0	451.0	900.0	450.5	Q	**1032.2**	516.6	**1015.1**	508.1	**1014.1**	507.6	7
9	**1033.1**	517.1	**1016.1**	508.5	**1015.1**	508.1	D	**904.0**	452.5	**887.0**	**444.0**	**886.0**	**443.5**	6
10	**1146.3**	573.6	**1129.2**	565.1	**1128.2**	564.6	L	**788.9**	395.0	**771.9**	386.5			5
11	**1332.5**	666.7	**1315.4**	**658.2**	**1314.5**	**657.7**	W	**675.8**	338.4	**658.7**	**329.9**			4
12	**1460.6**	730.8	**1443.6**	722.3	**1442.6**	721.8	Q	**489.6**	245.3	**472.5**	236.8			3
13	**1646.8**	823.9	**1629.8**	815.4	**1628.8**	814.9	W	**361.4**	181.2	344.4	172.7			2
14							R	175.2	88.1	158.2	79.6			1

Comparing Q-TOF and Ion Trap Spectra

The following examples compare spectra that were acquired on a **Q-TOF** and an **ion trap**, showing cases of both correct and incorrect identifications. Q-TOF results display a different pattern for the fragmentation spectra, as well as the presence of immonium ions and low mass fragment ions. Therefore, different considerations arise, depending on the instrument used.

Comparison with correct identification

Note the presence of **y_1** and the **immonium ion** for **F** in the Q-TOF spectrum. The QTOF spectrum shows **b** ion suppression except for **b_2** (rule 7)**.**



kely to

Comparison with incorrect Identification

In this case, each spectrum has left many prominent ions unaccounted for. The Q-TOF spectrum has assigned several b-type ions, which is unlikely with this equipment (rule 7). Thus, these identifications are likely to be incorrect.

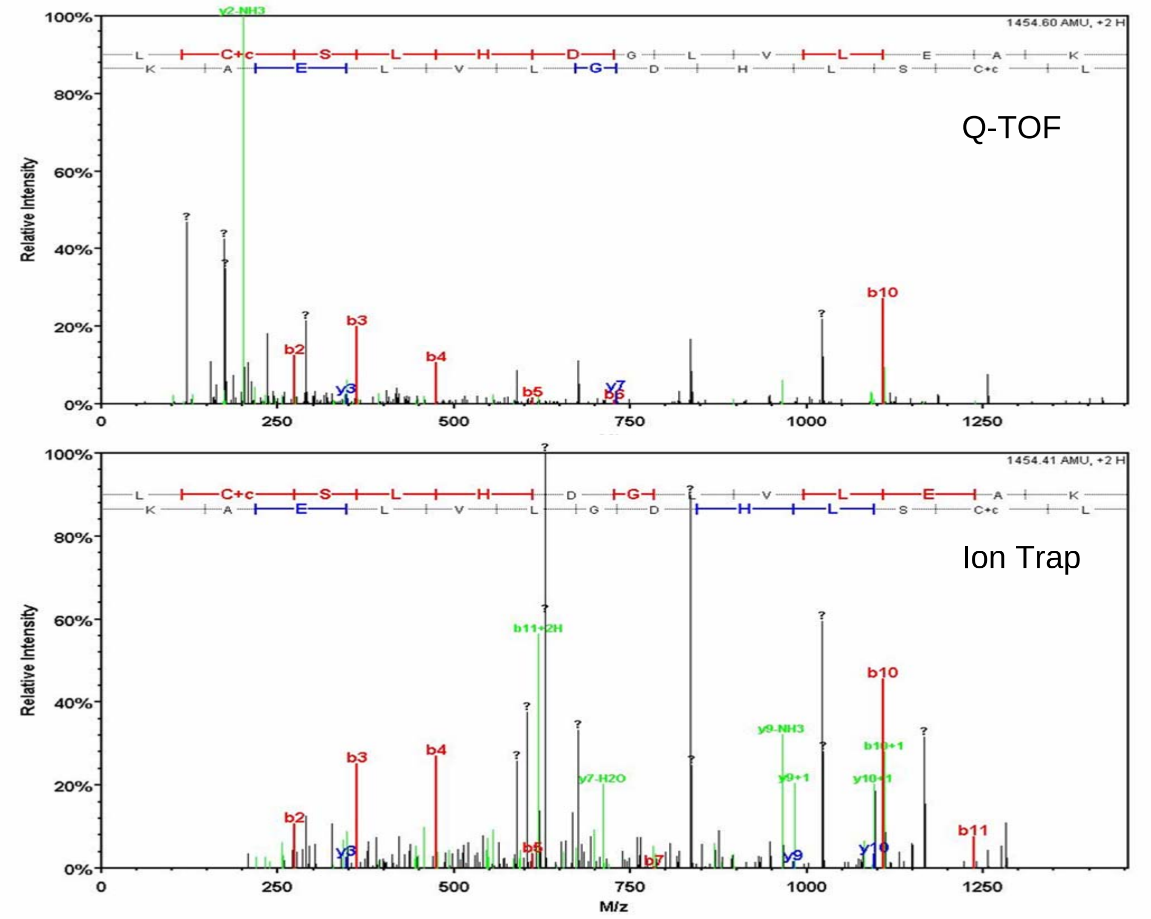


Figure 20: Q-TOF and Ion Trap spectra for an incorrect identification

Reader exercises

The following two figures present two more sets of spectra. In each case, identify the characteristics that support or raise questions about the validity of the identification.

Exercise 1

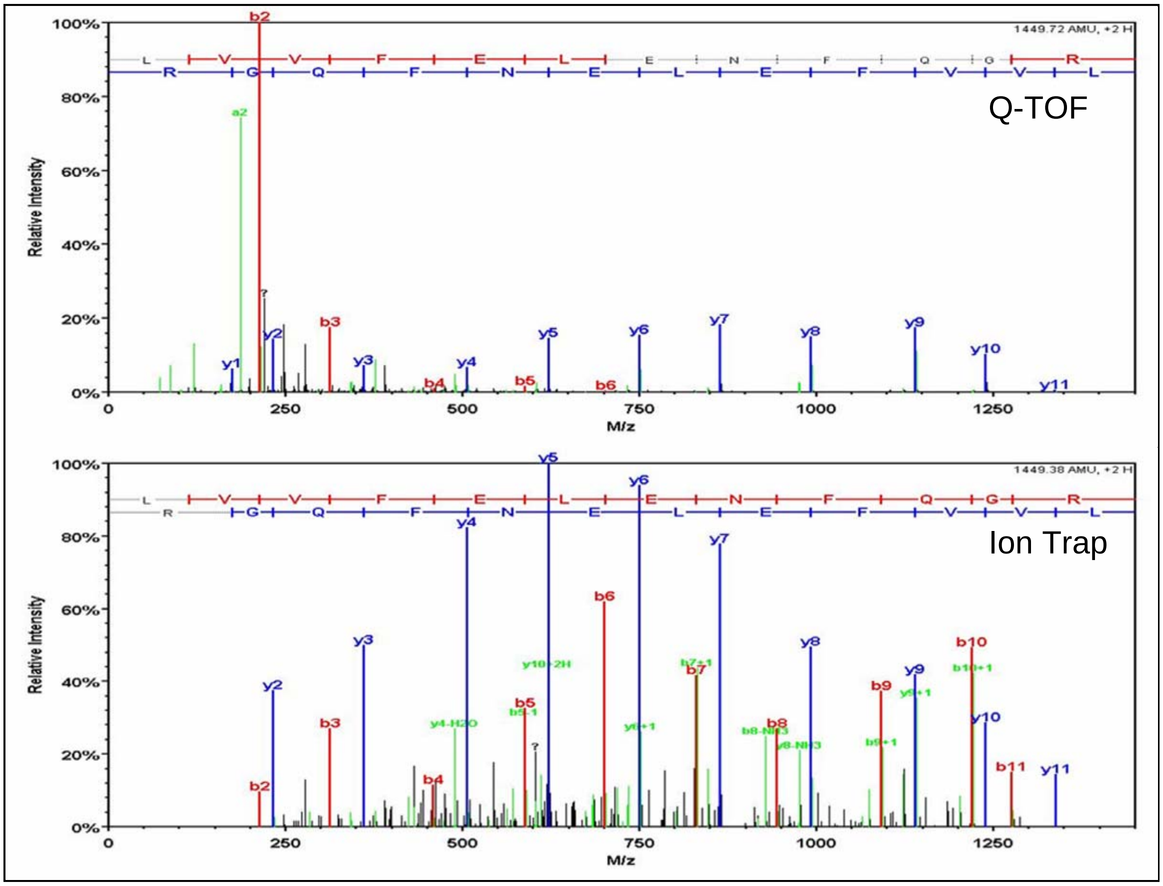


Figure 21: Test Case 1

Exercise 2

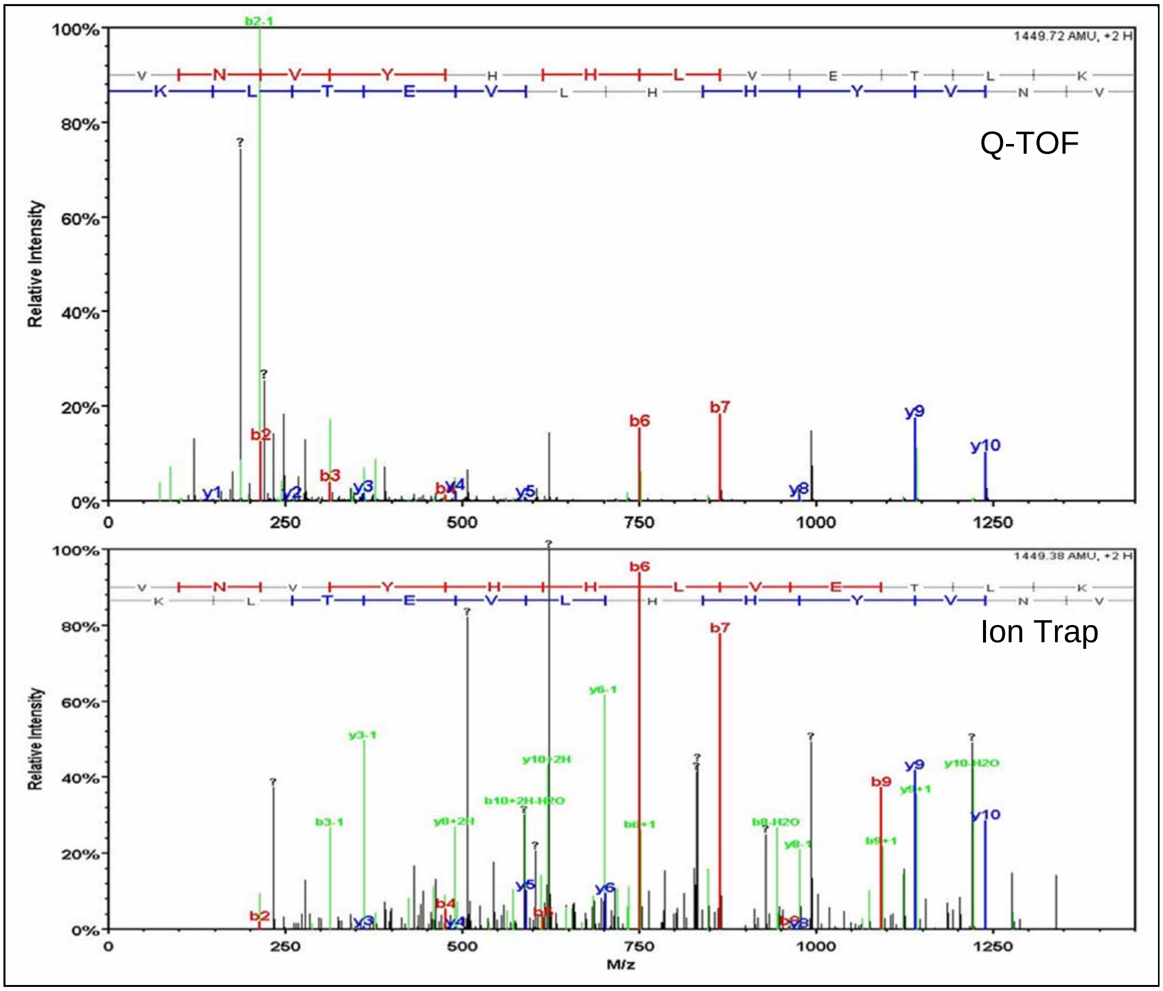


Figure 22: Test Case 2

See Appendix C: Notes for User Exercises for answers.

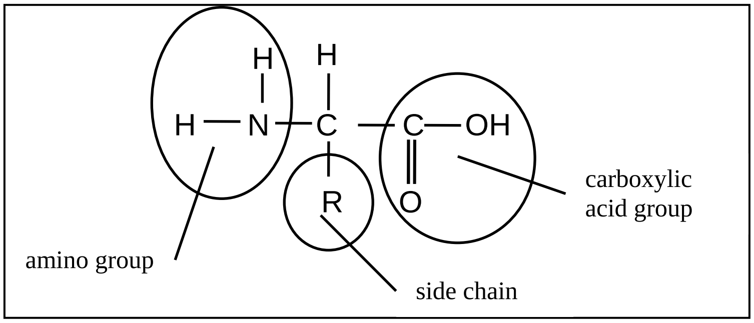
Summary

When interpreting the results of search engines it is important to consider all of the following factors.

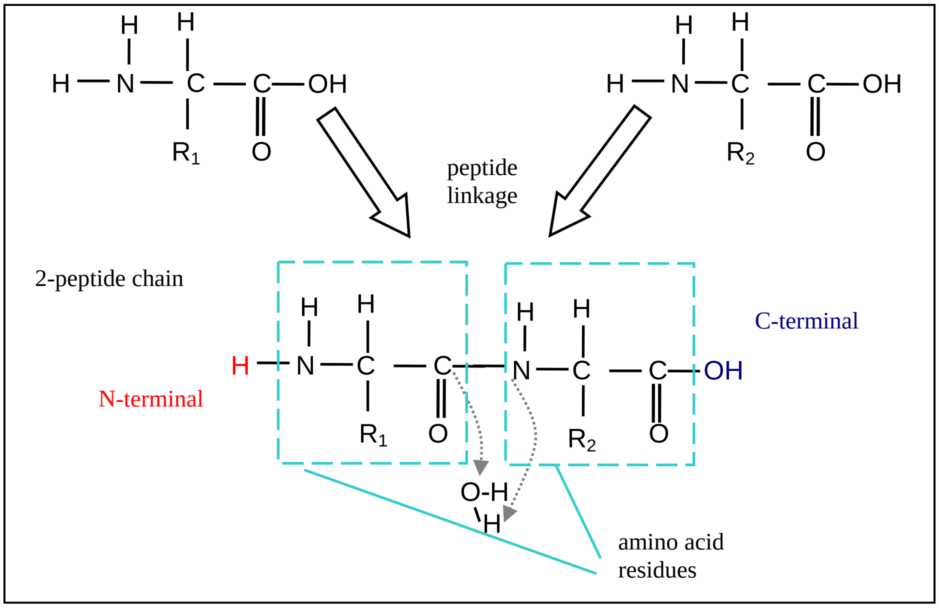
- **Look at the score**. Mascot or SEQUEST scores give one good indicator of confidence in peptide identification. Confirmation, or disagreement, from two search engines adds further information to confirm marginal scores or screen out false positives.
- **Look at the sequence runs**. A good spectrum should have clearly prominent **b** and **y** peaks above a low baseline of background noise. It should also display good fragmentation with a wide variety of peaks, covering most of the potential ions.
- **Consider the ion fragmentation chemistry**. Identifications based on mass differences should also be consistent with known fragmentation characteristics of the peptides identified and the digestion enzyme used.
- **Consider the instrument**. Q-TOF equipment typically displays more low mass and immonium ions, and suppresses most **b**-type ions.
- **Put it all together**. As seen in this paper’s examples, it is rare to have incontrovertible evidence for all of the above indicators in any one spectrum. A combination good professional judgment and good proteomics analysis software is necessary to make determinations in the context of your own experiments.

Appendix A: Peptide Chain Definitions

Amino Acid: A hydrocarbon chain of the general form shown below. Variations in the composition of the **side chain**, R, create the 20 different amino acids making up all human proteins.



Peptide: A peptide is formed when at least two amino acids join by an amide linkage (releasing H_2O), as shown below. An amide linkage between two amino acids is often called a **peptide linkage**. The portion of each amino acid that is retained in a peptide chain is referred to as a **residue**. Peptide chains are typically drawn with the amino end on the left, referred to as the **N-terminal**, and the carboxyl end on the right, referred to as the **C-terminal**.



Immonium Ion: a small fragment consisting of a single amino acid without the carboxyl group.

Isobaric: having the same mass, e.g. isoleucine and leucine.

Parent Ion, Precursor Ion: the peptide being fragmented in an MS experiment.

Appendix B: Sequence Identification Rules

This list summarizes the general rules and guidelines presented throughout the document.

De novo sequencing

1. The mass of a **b ion** equals the mass of its amino acid residues + 1 amu (from the N-terminal H^+).

   $\mathbf{b_1}$ will appear at one residue mass + 1, but $\mathbf{b_1}$ is rarely seen because of its low mass.

   The **heaviest b ion** will be below the precursor by {18 amu + one single residue mass}, i.e. in a range of 75 to 204 below the precursor mass. (heaviest **b** ion = precursor ion – $\mathbf{y_1}$.)

2. The mass of a **y ion** equals the mass of its amino acid residues + 19 amu (from the C-terminal OH plus 2).

   $\mathbf{y_1}$**,** will appear at its single amino acid residue mass plus 19 amu.

   The **heaviest y ion** will usually be the largest of all identified ions, appearing anywhere between 57 to 186 amu below the total mass of the precursor ion. (largest **y** ion = precursor ion – $\mathbf{b_1}$.)

3. If either the N- or C-**terminal** is **modified**, the masses of the **b** and **y** ions will be modified accordingly.

4. **Isobaric** amino acids (isoleucine, **I** and leucine, **L**; or lysine, **K** and glutamine, **Q**) cannot be differentiated using low energy fragmentation instruments.

5. If a peptide is **tryptic**, $\mathbf{y_1}$ will either be **lysine (K)** at 147 or **arginine (R)** at 175.
   ($\mathbf{y_1}$ mass = 19 + residue mass from Table 1. from rule 2)

   K and R are *not* likely to be found at the N-terminal or in the *interior* of a tryptic precursor ion.

Instrument specific

6. **Ion trap** results typically do not reveal $\mathbf{b_1}$, $\mathbf{y_1}$, or immonium ions, because of the low mass cut-off of the equipment.

7. **Q-TOF** results tend to display **y** ions and suppress **b** ions other than $\mathbf{b_2}$.

Fragmentation chemistry

8. **Proline** (**P**) cleaves easily on its N-terminal side, resulting in **dominant y-ion peaks** and **suppressed b ions**.

9. **Acidic** amino acids (**D** and **E**) can generate **high b ion peaks** for cleavages C-terminal to the residue. These residues also tend to **lose water** and **cyclize** to randomly eject portions of the sequence.

10. **Basic** amino acids (**K**, **R**, and **H**) must be present for a doubly-charged ion.

11. **Serine (S)** and **threonine (T)** can **lose water** (-18 amu). Water loss is especially intense when **T** is near the end of a peptide.

12. **Glycine** (**G**) tends to **suppress b** ions.
13. **Trytophan** (**W**) tends to create abundant y ions.
14. **Arginine (R), asparagine (N),** and **glutamine (Q)**, can **lose ammonia** (-17 amu).
15. **Cysteine (C),** can **lose H_2S** (-34), if not alkylated.
16. **Methionine** (**M**) can **lose CH_3SH** (-48 amu), and **oxidized M** can lose **CH_3SOH** (-64 amu).

Good spectrum characteristics

17. Spectrum should display **good signal-to-noise**, with clearly prominent peaks above a much lower baseline .
18. Peaks should be **well-separated** over a wide range of masses.

Good ID characteristics

19. **Long, contiguous ladders are good.** For ion trap equipment**,** which reveals both **b** and **y** ions, a complementary **b** and **y** set including some overlap is best.
20. **Prominent peaks** should be identified as **b** and **y** ions.
21. **Mascot** scores are often considered reliable when at least 40.
22. **SEQUEST XCorr** scores are often considered reliable above 1.9 or 2.2, or 3.7, for peptides with 1, 2, or 3 charges respectively.
23. A Mascot or SEQUEST **score** that is **well-separated** from the next highest score (deltaCn > 0.1 for SEQUEST) means less chance of an alternate identification that would be just as good.
24. **Agreement** between two different search engines increases identification confidence.

Appendix C: Notes for User Exercises

The identification in exercise 1 is likely to be correct . The ion trap generates long contiguous ladders, with overlapping **b** and **y** ion sets (Rule 19). Prominent peaks are identified as **b** and **y** ions (Rule 20).   No results appear materially inconsistent with the other instrument-specific and fragmentation chemistry rules presented in this paper.

The identification in exercise 2 is probably not correct. Although significant portions of the **b** and **y** ladders are identified (Rule 19), several prominent peaks are *not* identified to **y** and **b** ions. In addition, the Q-TOF results prominently display several b ions, which is *not* consistent with Rule 7.