A Novel Forwards-working Approach to Peptide Sequencing

Anthony Ferrige[1], Stuart Ray[1], Rob Alecio[1], Keith Waddell[2] and Kate Zhang[3]

[1]PPL Ltd, Isleham, Cambs., UK, CB7 5RX; [2]Applied Biosystems, Framingham, MA 01701, USA; [3]Genzyme, Framingham, MA 01701, USA

Overview

There are two commonly used methods aimed at determining the sequence of a peptide from its fragmentation pattern. In database search methods, the deisotoped m/z values are compared with those of known peptides. The other searches for expected mass differences and compares all possible combinations to see which sequences account for most masses above an arbitrarily user-selected threshold. Neither method uses the concept of working forwards to obtain the most probable sequences along with an estimate of how likely each is. The methodology described in this work uses predictive data reconstruction techniques on all the information in the data. The methodology is currently at an early stage of its development but is already showing promise and potential.

Introduction

Regardless of the many options available for acquiring fragmentation data, all rely on the degradation of the peptide into a number of fragments. Some fragmentation pathways are common to most methods and ‘y’ and ‘b’ ions are generally present, even if incomplete and weak. By the nature of the fragmentation process, a range of charges is usually present. Typically, Z=1 to Z=3 will be present but Z=5 and Z=6 are not uncommon for larger peptides. The modern generation of instruments almost invariably resolve the ions into their respective isotope clusters and most programs make use of this charge information to obtain either a zero-charge or zero-charge+1 spectrum as the starting point for peptide sequencing. However, it is crucial that any centroiding or spectrum deconvolution should provide reliable peak positions and that the multi-charge deisotoping should not produce artefacts.

For the data presented here the resolution was sufficient to centroid the data rather than perform a rather more time-consuming spectrum deconvolution. However, because traditional centroiding methods are very prone to noise, we developed a very fast data reconstruction method that substantially reduces noise without broadening peaks. The reconstruction may then be efficiently centroided and the error bars computed. Irrelevant features may be removed from peak tables using significance filters – arbitrary thresholds do not apply. The program is comparable in speed to other commonly used centroiding methods. The certainty of peak positions and intensities is naturally high for intense, isolated peaks and decreases with both decreasing S/N and for severely overlapped peaks. This is illustrated in Figure 1.

In addition, the intensity errors are available for the multi-charge deisotoping. By its very nature, algebraic deisotoping assumes that there is no intensity error for each isotope peak. This places an extremely severe constraint on the fitting process and frequently generates numerous artefact peaks. We have therefore developed a ReSpect™-based deisotoping program that performs its fitting within both the noise level and the intensity errors. This freedom – absent for algebraic methods – ensures that there is positive evidence in the data for any reconstructed deisotoped peak and that the results are free of artefacts, other than those that arise from the applied empirical formula being an average and therefore a compromise for any particular peptide or fragment. The new ReSpect™-based deisotoping method therefore provides very clean and reliable zero-charge results that are well suited to peptide sequencing studies. An example is shown in Figure 2. The top part of the diagram shows the reconstructed intensity patterns that fit the data within the noise level. The centre trace (black) shows the summed reconstructed centroids for comparison with the data (bottom). Note that the peak at 384.2 in the data is broad because there is an imperfect coincidence of a Z=2 and Z=3 cluster.

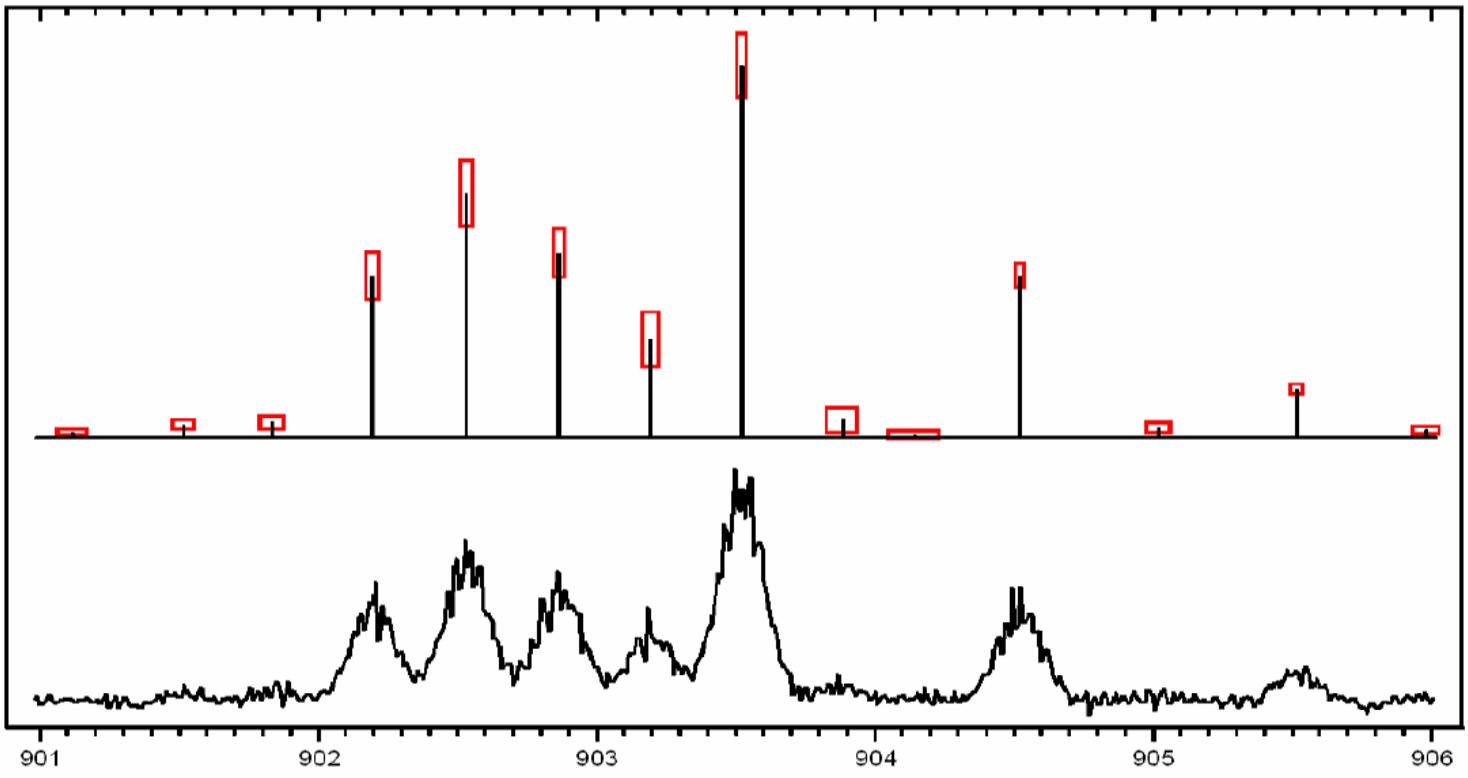
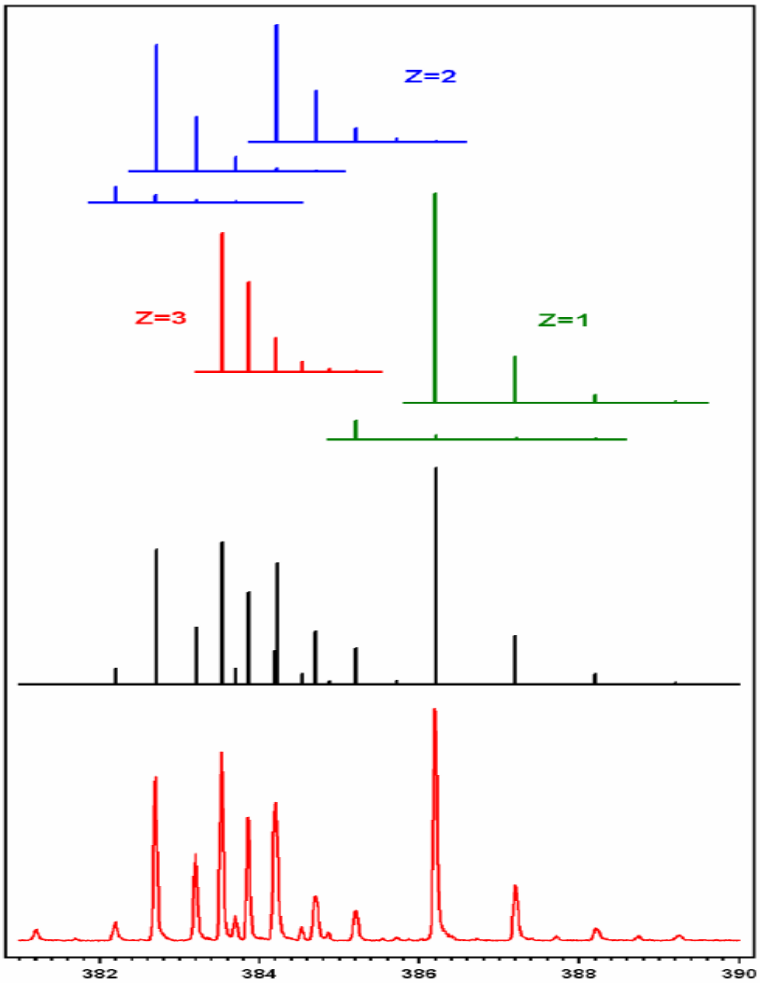


Fig. 1. How errors change with S/N and peak overlap



Problem Complexity

De novo peptide sequencing is a particularly complex problem because of the extremely large number of possibilities that need to be considered. Given fast enough computers and a vast amount of memory the problem could be solved without resorting to ways in which the number of permutations may be reduced. Indeed, it is the large number of variables involved that cause the currently available methods to fail frequently.

Another common difficulty is that the mass of the peptide may be unknown at the outset and it is not possible to apply constraints to the computation. An example of this complexity is illustrated in Figure 3. This shows the total number of combinations that need to be considered for peptides up to a mass of 1000 as being almost 4 million. However, this is only part of the story because for each amino acid combination at a given mass, numerous different sequences are possible.

Fig. 2. Top, Reconstructed isotope patterns; Centre, Reconstructed centroids from patterns Bottom – Data

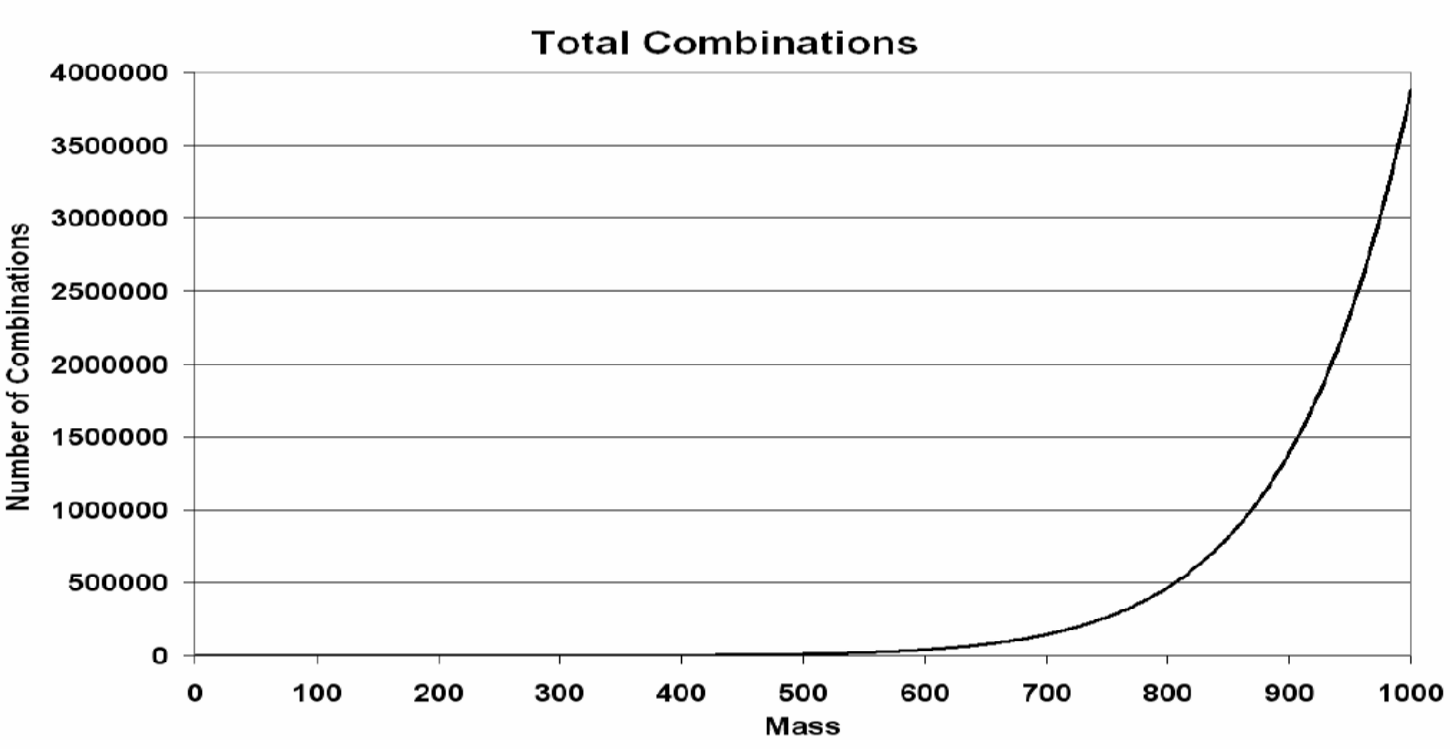


Fig. 3 The total number of combinations for peptides with a mass up to 1000

A further illustration of this is shown in Table 1. Here, all 81 amino acid combinations are listed within 100 ppm of the reference mass of 509.198. For this purpose, **I** and **L** are treated as being the same amino acid.

Table 1 – Amino Acid Combinations for a Specific Mass

Mass	Amino acids	Ref M	ppm
509.158	S1C1D1W1	509.198	82.0
509.161	S3D2	509.198	76.9
509.161	S1P1T1C2	509.198	75.2
509.161	C1Q1E1M1	509.198	75.1
509.161	A1V1C2D1	509.198	75.1
509.161	A2C1D1M1	509.198	75.1
509.161	G1C2L1D1	509.198	75.1
509.161	G1A1C1E1M1	509.198	75.1
509.161	G2D1M2	509.198	75.1
509.161	N1D1M2	509.198	75.1
509.161	G1V1C2E1	509.198	75.1
509.173	C2E1R1	509.198	52.3
509.176	N1D2F1	509.198	45.9
509.176	G2D2F1	509.198	45.9
509.177	V1C2W1	509.198	44.1
509.177	A1C1M1W1	509.198	44.1
509.179	A1S1T1C1E1	509.198	39.0
509.179	G1T2C1E1	509.198	39.0
509.179	S2V1C1D1	509.198	39.0
509.179	A1T2C1D1	509.198	39.0
509.179	A1S2D1M1	509.198	39.0
509.179	G1S1T1D1M1	509.198	39.0
509.179	G1S2E1M1	509.198	39.0
509.181	C1N1H2	509.198	36.2
509.181	G2C1H2	509.198	36.2
509.194	G1S1P1C1F1	509.198	8.0
509.194	G1A1P1C1Y1	509.198	8.0
509.194	T2C1W1	509.198	7.9
509.194	P1C1Q1Y1	509.198	7.9
509.194	S2M1W1	509.198	7.9
509.197	S3T1E1	509.198	2.9
509.197	S2T2D1	509.198	2.8
509.198	C1K1E1M1	509.198	1.1
509.198	A1S1V1C1M1	509.198	1.1
509.198	G1S1C1L1M1	509.198	1.1
509.198	T1Q1M2	509.198	1.0
509.198	S1V2C2	509.198	1.0
509.198	A1T1C2L1	509.198	1.0
509.198	A2S1M2	509.198	1.0
509.198	G1V1T1C1M1	509.198	1.0
509.198	G1A1T1M2	509.198	1.0
509.206	G1P2C1H1	509.198	-14.9
509.209	T1C1M1R1	509.198	-21.8
509.212	G1A2E1Y1	509.198	-28.1
509.212	T1D1Q1F1	509.198	-28.2
509.212	T1N1E1F1	509.198	-28.2
509.212	V1N1D1Y1	509.198	-28.2
509.212	S1Q1E1F1	509.198	-28.2
509.212	A1Q1E1Y1	509.198	-28.2
509.212	A2S1D1F1	509.198	-28.2
509.212	A3D1Y1	509.198	-28.2
509.212	G1S2P1Y1	509.198	-28.2
509.212	G1A1T1D1F1	509.198	-28.2
509.212	G1A1S1E1F1	509.198	-28.2
509.212	G2T1E1F1	509.198	-28.2
509.212	G2V1D1Y1	509.198	-28.2
509.216	S1V1T2C1	509.198	-35.1
509.216	S3V1M1	509.198	-35.1
509.216	A1S1T2M1	509.198	-35.1
509.216	S2T1C1L1	509.198	-35.1
509.216	G1T3M1	509.198	-35.1
509.223	P1Q1E1H1	509.198	-51.1
509.223	A2P1D1H1	509.198	-51.1
509.223	G1D1R1Y1	509.198	-51.1
509.223	G1A1P1E1H1	509.198	-51.1
509.227	A1S1F1W1	509.198	-59.3
509.227	A2Y1W1	509.198	-59.3
509.227	G1T1F1W1	509.198	-59.3
509.231	P4C1	509.198	-66.1
509.231	A2V1C1F1	509.198	-66.1
509.231	A3M1F1	509.198	-66.1
509.231	G1A1C1L1F1	509.198	-66.1
509.231	G2V1M1F1	509.198	-66.1
509.231	P1C1K1Y1	509.198	-66.1
509.231	C1L1Q1F1	509.198	-66.2
509.231	V1N1M1F1	509.198	-66.2
509.233	S1T4	509.198	-71.3
509.234	T1K1M2	509.198	-73.0
509.239	A1P1H1W1	509.198	-82.2
509.242	G1M1F1R1	509.198	-89.0

However, this is only part of the story because for each specific peptide combination there will be numerous different sequences. The average number of amino acids in Table 1 is about 5. Therefore, on average there are about 120 sequences per entry. This means that there are almost 10000 sequences to be considered within a 100 ppm error.

Methodology

Any peptide sequencing program relies heavily on the quality of the data and full advantage is taken of the fast reconstruction centroiding and the ReSpect™-based artefact-free multi-charge deisotoping programs since there is evidence for any reconstructed zero-charge mass. The ReSpect™ deisotoping method provides a zero-charge peak table of masses and their intensities along with their errors. Confidence filters may therefore be applied to remove peaks of low significance, making arbitrary thresholding unnecessary. The resulting peak table is used purely as a reference and the new predictive program constructs both masses and intensities so that low intensity masses are generally less favourable than high intensity masses.

Just like other de novo methods, the method described here is capable of generating millions of possible sequences unless it is guided towards only those solutions with a high probability. Before making selections based on consistency with the data, it is necessary to establish a reasonable starting point. Because instruments are frequently calibrated using rather low mass calibrants, it is common to find unacceptably high mass errors towards the high m/z limit of the data. Absolute mass measurements may therefore be unreliable and any attempt to identify a peptide from its mass is fraught with danger.

Instead, we look for common mass differences as this reduces the impact of calibration errors and additionally makes the method more robust against any substitution at the peptide terminals. By searching for common mass differences, a table of possible partial peptides is generated, together with confidence levels based on the quality of the fit to the mass and intensity data. The list of possible small trial peptides that are considered will fall within an allowed error of typically 100 ppm. Using a database of amino acids, the program works forwards from each trial peptide to predict the most likely amino acids to extend the sequence. The predictive process is applied towards both the C- and N-terminals with the aim of accounting for the greatest number of peaks in the data and the greatest intensity. However, it is also important to take into account the quality of the fit of each new candidate sequence to the data with respect to the mass errors. Because noise will be present, sequences that are weak are penalised even if they account for a large number of peaks.

In addition, a major constraint that is applied in the fitting process is that there must be evidence for both forward and reverse fragmentation so that the sequence may be read in both directions. This constraint dramatically reduces the number of potential candidates.

Early experiments have shown that it is usually possible to generate reliable candidate tri- and tetra-peptides and each is considered as a suitable starting point for reconstructing the data. During the sequence expansion phase the number of possible peptides increases rapidly so that there may be many candidates that have a similar probability. However, as the expansion continues towards a conclusion, more and more of the data are accounted for and the number of plausible candidates falls. This is because it becomes impossible for the vast majority of the candidates to fit the remaining peaks. Typically, we find that at least 90% of the intensity in the data is fitted to high precision. The way the number of possible sequences changes during the fitting is shown in Figure 4.

In all our tests to date, the maximum number of potential sequences has never exceeded reasonable limits and the final hit list has always reduced to a small number of sequences that fit the data to varying degrees.

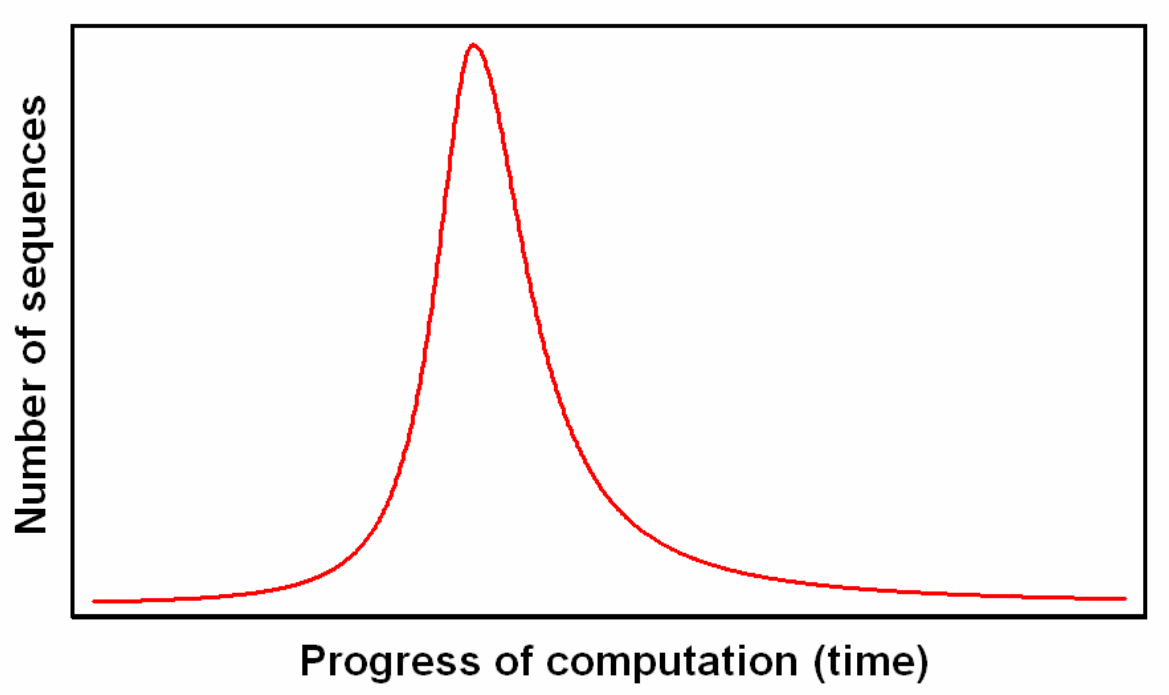


Fig. 4 The way the number of potential sequences changes during the course of the calculation

The program terminates when the greatest intensity and largest number of peaks are accounted for.

Experimental

Two peptides were used in this study with the sequences:

Example 1: **FLFHTEYVV**

Example 2: **TGPNLHGLFGR**

Both peptides (less than 0.1 mg each) were dissolved in 600µl DMSO:ACN:0.1% formic acid solution (1:1:1 v/v). 4µl peptide solution was loaded to the nanospray tip. The nano electrospray MS was operated under the positive mode with 900-1300 v. The mass scan range was m/z 100-1500. The instrument used was an ABI Q-Star.

Results

Example 1

The zero-charge spectrum obtained by centroiding and multi-charge deisotoping is shown in Figure 5 as a spike plot. The peak table was used as the input for the peptide sequencing program using an error of 100 ppm.

This particular peptide showed considerable internal cleavage and many of the y ions were weak. Some of the y and b ion assignments obtained from the program are shown in the figure.

In this example the target peptide was the top hit and accounted for more peaks and more intensity than any other peptide. At this early stage of the methodological development the scoring was based simply on how much of the data are accounted for in terms of the number of peaks and intensity that is reconstructed. Following the top hit were sequences that were missing one or other or both of the terminal amino acids.

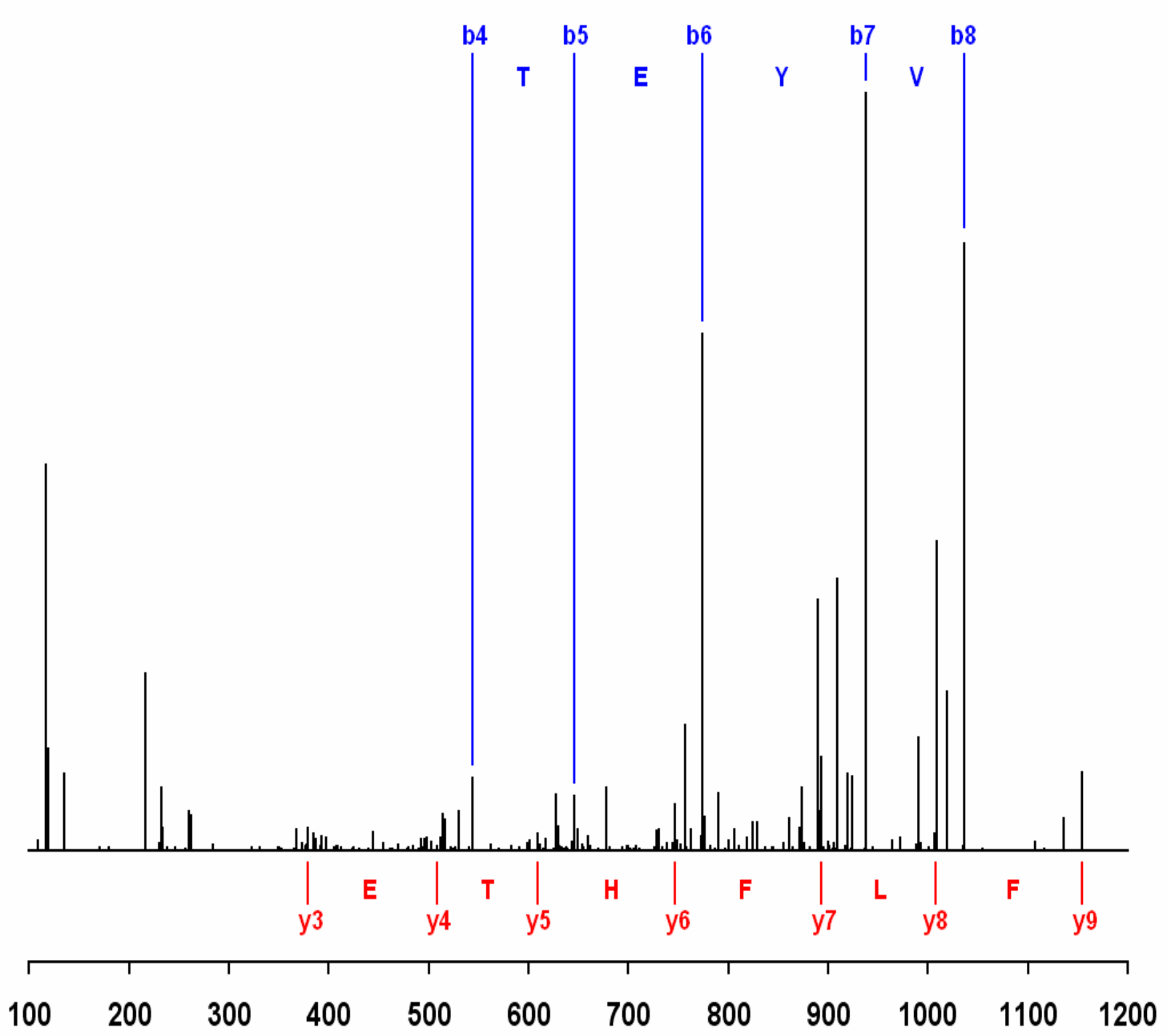


Fig. 5. ZC spectrum of FLFHTEYVV with y and b ion assignments

The complete list of 167 assigned ions is shown in Table 2. Of course, some of the very weak ions may be spurious but they all fit within the calibration error.

The total intensity in the data was 1419741 and the total reconstructed intensity for the assigned ions was 1344145. The new methodology has therefore assigned **94.7%** of the intensity in the data.

Ignoring the very few serious outliers, a calibration error is evident and is shown in Figure 6. Although the calibration error is almost 0.1 Da over ~1000 mass units, the new methodology is almost immune to this because it is predicting single amino acids. The error over the mass range for tyrosine (Y) reduces to only 0.015 Da and is well within the 100 ppm allowed in the data reconstruction calculation.

Table 2 – Identified Ions

Ion	MassFnd	Int	ppm
a*3	362.215	462	44.1
a*4	499.292	275	68.2
a*5	600.349	1176	71.8
a*6	729.406	249	78.2
a*7	892.494	5356	92.0
a*8	991.569	1349	89.5
a03	361.220	368	13.2
a04	498.293	3071	38.3
a05	599.358	1757	59.6
a06	728.413	4593	66.1
a07	891.481	9076	60.1
a08	990.564	25970	68.1
a1	119.077	23354	25.3
a2	232.170	14436	55.5
a3	379.233	5241	19.7
a4	516.322	7117	72.0
a5	617.374	2756	67.5
a6	746.430	10790	73.0
a7	909.513	61763	81.8
a8	1008.590	70415	82.1
a9	1107.658	2250	74.4
b03	389.219	204	21.0
b04	526.297	852	52.1
b05	627.354	13051	58.8
b06	756.418	28833	77.2
b07	919.498	17520	81.6
b08	1018.581	36365	88.4
b2	260.168	9179	60.5
b3	407.242	1394	53.1
b4	544.320	16763	73.0
b5	645.377	12620	77.2
b6	774.429	117370	76.6
b7	937.508	171879	79.7
b8	1036.586	137895	81.1
c05	644.323	23	-31.7
c07	936.506	486	60.7
c08	1035.583	1324	63.5
c1	164.097	31	9.5
c3	424.226	339	-49.6
c4	561.339	220	58.3
c5	662.373	1242	28.9
c7	954.555	38	99.9
iax2-4	395.199	73	9.0
iax2-5	496.260	2773	33.2
iax2-7	788.423	55	93.2
iax3-6	512.245	3093	83.6
iay02-4	351.195	176	-31.5
iay02-6	581.333	391	63.0
iay02-7	744.429	1808	94.0
iay02-8	843.490	854	73.7
iay03-4	238.121	926	-3.0
iay03-5	339.180	307	29.9
iay03-7	631.321	1404	72.9
iay03-8	730.400	4978	76.3
iay04-5	192.108	57	35.7
iay04-6	321.160	290	49.7
iay04-7	484.241	1306	70.1
iay04-8	583.318	1253	73.8
iay05-6	184.092	31	40.4
iay05-7	347.167	38	53.5
iay05-8	446.236	277	44.5
iay06-8	345.191	219	63.7
iay2-3	232.170	14436	55.5
iay2-6	599.358	1757	85.0
iay2-7	762.430	4894	78.9
iay2-8	861.501	7420	72.9
iay3-4	256.148	700	60.0
iay3-5	357.193	249	35.3
iay3-6	486.260	293	76.0
iay3-7	649.342	4887	86.0
iay3-8	748.415	2595	81.3
iay4-5	210.122	6	51.4
iay4-6	339.180	307	75.0
iay4-7	502.257	2085	78.7
iay4-8	601.340	2466	89.4
iay5-6	202.106	403	51.9
iay5-7	365.191	484	88.7
iay5-8	464.259	730	69.2
iay6-8	363.206	7	73.9
iay7-8	234.150	5287	56.3
iaz2-4	352.213	653	64.5
iaz2-5	453.255	202	37.9
iaz2-6	582.326	359	78.6
iaz2-7	745.416	1390	97.5
iaz2-8	844.486	894	87.8
iaz3-5	340.174	16	61.5
iaz3-7	632.321	160	97.4
iaz4-5	193.092	1	35.7
iaz5-7	348.157	39	72.6
ibx2-4	423.231	509	95.3
ibx2-5	524.265	525	50.3
iby02-4	379.233	5241	85.9
iby02-5	480.279	1028	64.1
iby02-6	609.340	3995	79.8
iby02-7	772.422	3429	87.2
iby02-8	871.498	5330	85.9
iby03-4	266.118	13	5.8
iby03-5	367.180	5083	43.6
iby03-7	659.326	3453	84.2
iby03-8	758.407	1027	89.7
iby04-6	349.171	1033	92.6
iby04-7	512.245	3093	83.6
iby04-8	611.323	1570	85.3
iby05-6	212.092	7	59.4
iby05-7	375.168	465	66.6
iby05-8	474.247	287	73.9
iby06-7	274.111	107	56.9
iby06-8	373.201	1926	99.5
iby2-3	260.168	9179	60.4
iby2-4	397.232	3051	53.1
iby2-5	498.293	3071	69.0
iby2-6	627.354	13051	83.1
iby2-7	790.432	13119	84.2
iby2-8	889.506	56945	81.6
iby3-4	284.146	1558	64.8
iby3-5	385.200	4121	65.0
iby3-6	514.261	8456	84.3
iby3-7	677.339	14364	85.3
iby3-8	776.415	7824	84.4
iby4-5	238.121	926	61.2
iby4-6	367.180	5083	85.3
iby4-7	530.257	9255	84.5
iby4-8	629.329	5580	77.0
iby5-6	230.103	1856	55.3
iby5-7	393.181	3511	69.0
iby5-8	492.260	2856	76.3
iby6-7	292.124	296	63.0
iby6-8	391.208	1157	85.0
iby7-8	262.147	8148	60.0
ibz2-6	610.268	229	-11.0
icx7-8	305.134	50	-10.1
icy2-6	644.323	23	-8.1
icy3-4	301.159	19	15.4
icy3-5	402.211	270	22.9
icy3-7	694.367	1035	86.5
icy4-6	384.200	816	64.0
icy4-7	547.281	217	76.6
icy5-7	410.213	164	80.9
icy6-7	309.155	56	72.8
x03	387.207	2801	71.2
x07	901.486	1406	98.2
x3	405.228	848	94.6
x7	919.498	17520	98.2
y03	361.220	368	55.4
y04	490.276	628	68.1
y05	591.330	939	67.3
y06	728.413	4593	87.1
y07	875.488	2007	80.6
y08	988.584	1691	83.3
y09	1135.654	7523	73.7
y1	117.082	87496	23.8
y2	216.158	40196	50.9
y3	379.233	5241	60.0
y4	508.285	1113	62.1
y5	609.340	3995	63.7
y6	746.430	10790	93.5
y7	893.501	21417	81.8
y8	1006.600	4071	87.2
y9	1153.666	17847	74.1
z*3	345.191	219	96.5
z*4	474.247	287	97.8
z*8	972.549	3230	92.6
z03	344.197	46	68.0
z04	473.258	215	89.1
z3	362.215	462	86.3
z5	592.322	59	80.4
z6	729.406	249	99.2

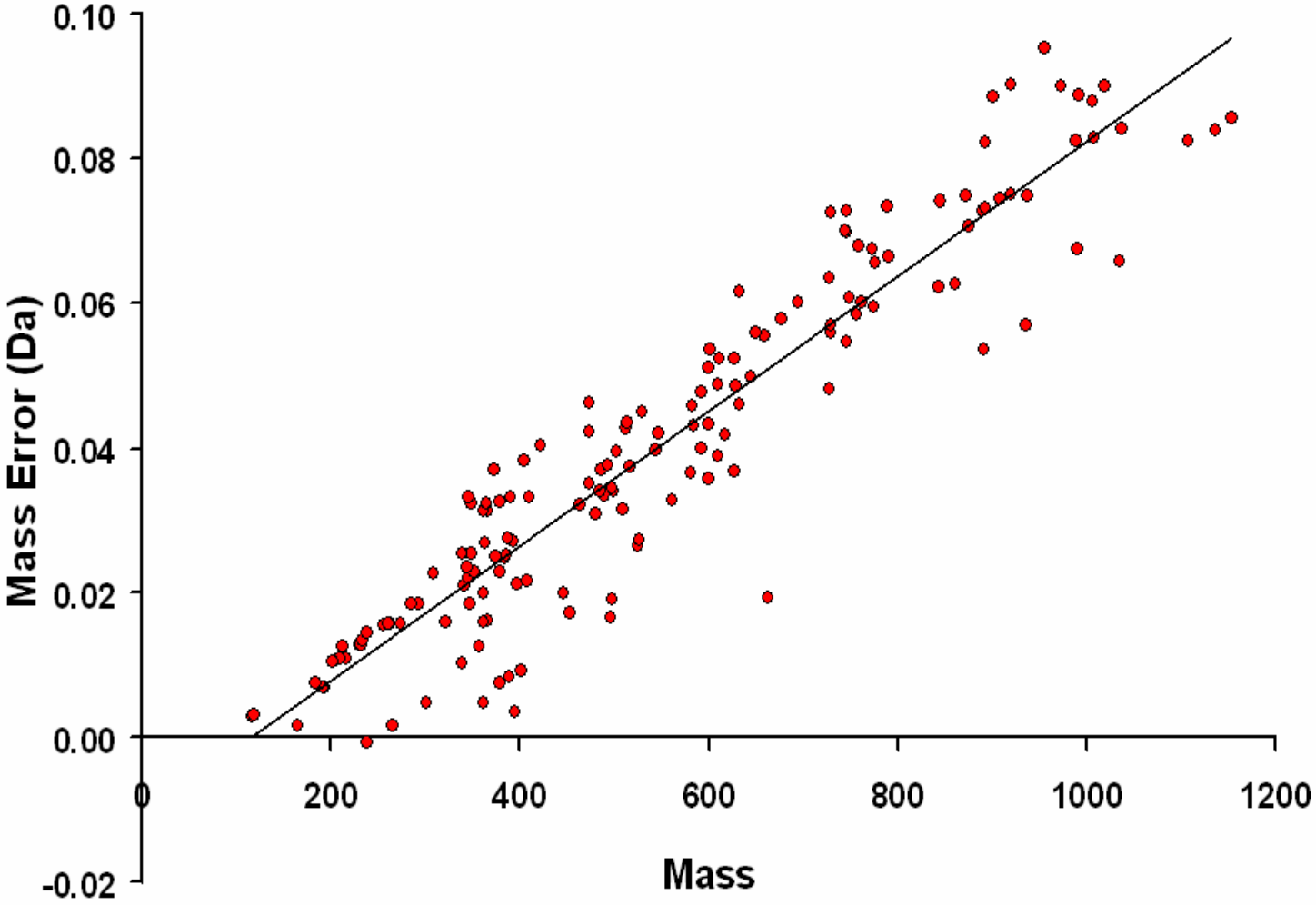


Fig. 6. Calibration error for FLFHTEYVV

Example 2

The zero-charge spectrum obtained by centroiding and multi-charge deisotoping is shown in Figure 7 as a spike plot. The peak table was used as the input for the peptide sequencing program using an error of 100 ppm.

Again, considerable internal cleavage was observed and the b ions were weak. Most of the y ions were strong and are shown on the figure.

In this example the target peptide was again the top hit and accounted for more of the data than any other peptide. As with Example 1, the following hits were for sequences that were missing one or other or both of the terminal amino acids.

The total intensity in the data was 30360 and the total reconstructed intensity for the assigned ions was 27707. Therefore **91.3%** of the intensity in the data has been accounted for.

The calibration for this example had no trend and all errors were well within the 100 ppm allowed in the data reconstruction calculation.

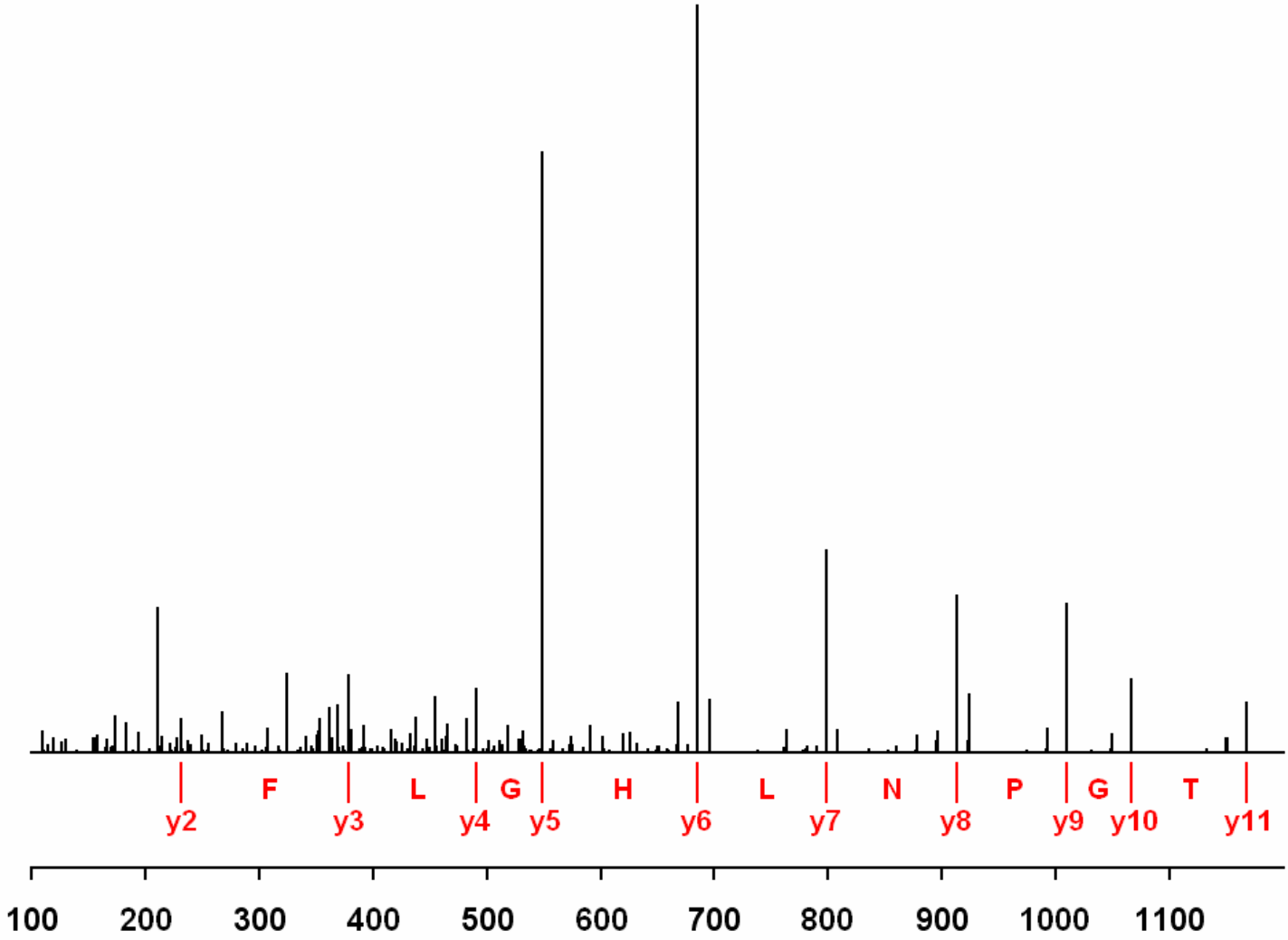


Fig. 7. ZC spectrum of TGPNLHGLFGR with y ion assignments

Discussion

It is notable that the sequencing technique described here relies only on general peptide fragmentation reactions. The method does not make use of empirical rules that attempt to predict how the behaviour of a peptide depends on the particular amino acids that are present. It is likely that the performance of the method would be improved by incorporating such rules.

The types of peptide fragment that are observed depend on the details of the measurement technique, which in practice are not always closely controlled, and also on the nature of the peptide itself. Thus it is unlikely that any single sequencing method will be applicable in all cases. In this regard, it may be significant that the new technique does not depend on the presence of complete fragmentation series, and is therefore complementary to conventional ladder sequencing methods.

At present the ppm error that is allowed is quite critical, particularly when the data are not particularly well calibrated as in Example 1. Setting this too high allows numerous additional sequences to fit the data and unrelated sequences can displace the target sequence. Setting the error too low may prevent the reconstruction of the target sequence and hits high on the list are more likely to be partial rather than full sequences. As the methodology is developed to incorporate confidence levels it is anticipated that the program will become much more robust.

Conclusions

We have described a peptide sequencing technique that complements conventional methods. The technique shows promise, is capable of further improvement and could form part of a more widely applicable combined approach.

Future Work

Work continues on determining the confidence level for each plausible sequence so that they may be ranked according to their probability. It is hoped to have a fully functional and robust program operating during the coming months.

Acknowledgements

We are grateful to Genzyme (Framingham) and ABI (Framingham) for providing the data for analysis.