pNovo+: *De novo* Peptide Sequencing using Complementary HCD and ETD Tandem Mass Spectra

November 9, 2011

Abstract

(1)Complementarity HCD and ETD, including hydrogen rearrange ions in ETD data;

(2)Relaxation of the anti-symmetry constraint;

(3)Efficient algorithm for finding the *k* longest paths in a directed acyclic graph.

Abstract

(1)pNovo+ identified 76.2% of full sequence which were yielded by database search , while PEAKS HCD was 52.6%, PEAKS ETD was 33.6%, PEAKS (HCD∪ETD) was 63.6%.

(2)pNovo+ takes on average 0.07 second per spectrum.

Outline

(1)Methods

(1.1)Peak Selection

(1.2)Ion Type Determination

(1.3)Directed Acyclic Graph Construction

(1.3.1)Generating graph vertices

(1.3.2)Generating graph edges

(1.3.3)Finding the k longest paths

(1.3.4)Ranking all candidate peptides

(2)Materials and Results

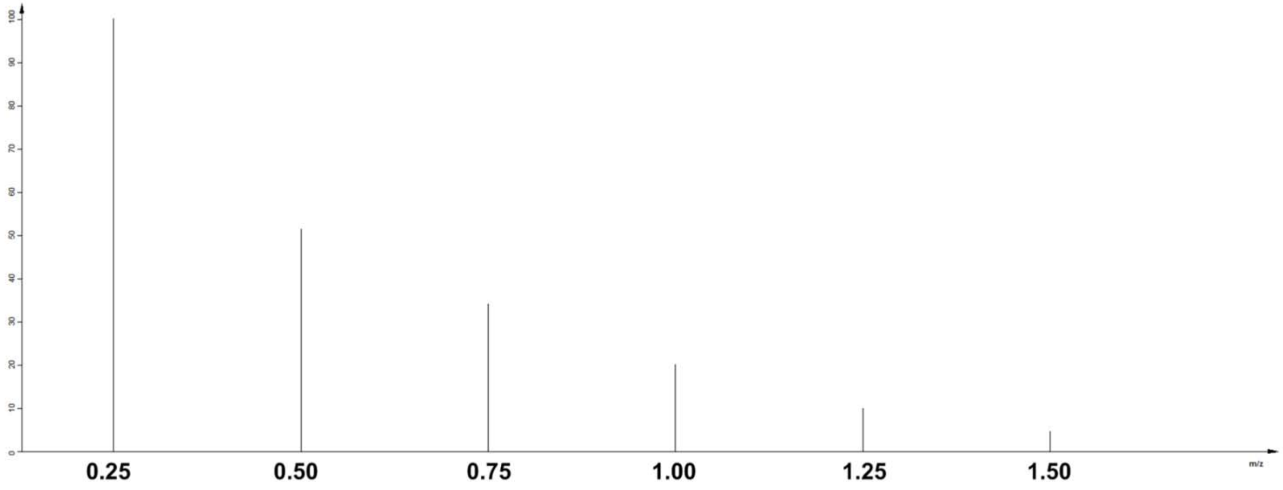
Comparison between pNovo+ and PEAKS

Peak Selection

(1)The weight of each peak is set equal to the natural logarithm of its intensity.

(2)For each peak, charge state by finding the best fitting one from 1+ to 2+, 3+…

If one peak is possibly *C*+ charged, we do not check its charge as the divisor of *C*.



Peak Selection

(3)For an isotopic cluster, we selected the peak *p* with the lowest m/z.

(4)In ETD spectra, we also selected the peaks whose intensities were greater than the intensity of *p*.

The peaks with intensity greater than *p* in ETD were selected in order to inclusion of [*c*−H] and [*z*+H] ions that are very abundant in ETD spectra and [*c*−H] and [*z*+H] ions differ from their cognate *c* and *z* ions by ~1 Dalton, a high degree are superimposed onto isotopic clusters.

Peak Selection

(5)Peaks that are not associated with any isotopic clusters are treated as both singly charged and doubly charged ions.

(6)All peaks were transformed to their singly charged m/z value, and merge peaks of equal mass within a given tolerance range.

(7)The precursor ion peaks and those peaks with neutral losses such as the loss of water or ammonia were detected and deleted.

Ion Type Determination

$$\mathbf{b - c = -17.0265} \qquad (1)$$

$$\mathbf{y - z = +16.0188} \qquad (2)$$

$$\mathbf{b - [c - H] = -16.0192} \qquad (3)$$

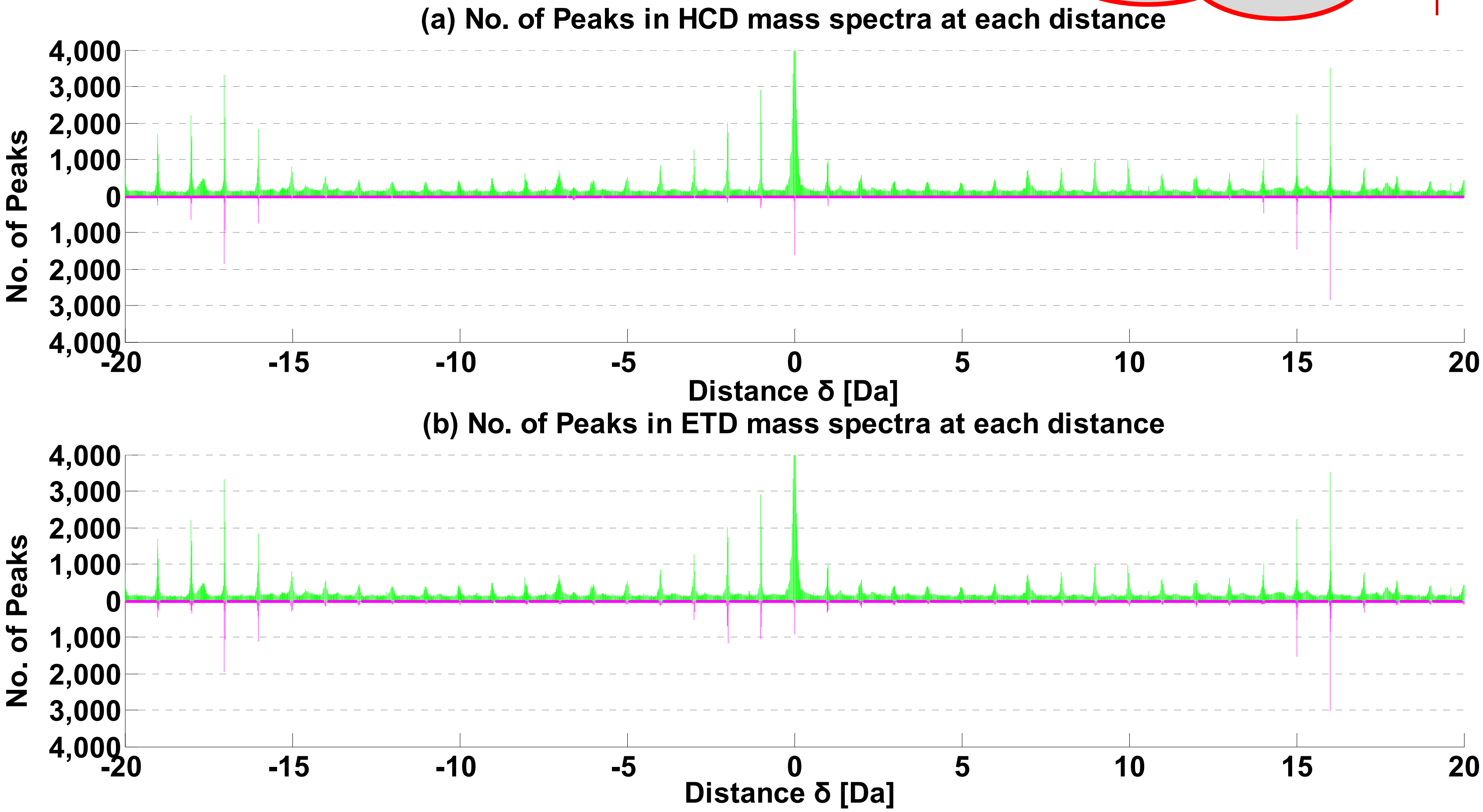
$$\mathbf{y - [z + H] = +15.0115} \qquad (4)$$

Horn, D.M., R.A. Zubarev, and F.W. McLafferty, *Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. Proc Natl Acad Sci U S A, 2000.* ***97(19): p. 10313-7.***

Sun, R.X., et al., *Improved peptide identification for proteomic analysis based on comprehensive characterization of electron transfer dissociation spectra. J Proteome Res, 2010.* ***9(12): p. 6354-67.***

No. of Peaks at each distance

$$\mathbf{b-c=-17.0265}$$
$$\mathbf{y-z=+16.0188}$$
$$\mathbf{b-[c-H]=-16.0192}$$
$$\mathbf{y-[z+H]=+15.0115}$$


(a) No. of Peaks in HCD mass spectra at each distance

(b) No. of Peaks in ETD mass spectra at each distance

Figure 1. (a) The area above the x-axis represents the number of HCD peaks at each distance δ, and the area below the x-axis represents the number of HCD peaks that were matched to the correct peptide at each distance δ. (b) The area above the x-axis represents the number of ETD peaks at each distance δ, and the area below the x-axis represents the number of ETD peaks that were matched to the correct peptide at each distance δ.

Match accuracies

$$\mathbf{b-c=-17.0265}$$
$$\mathbf{y-z=+16.0188}$$
$$\mathbf{b-[c-H]=-16.0192}$$
$$\mathbf{y-[z+H]=+15.0115}$$

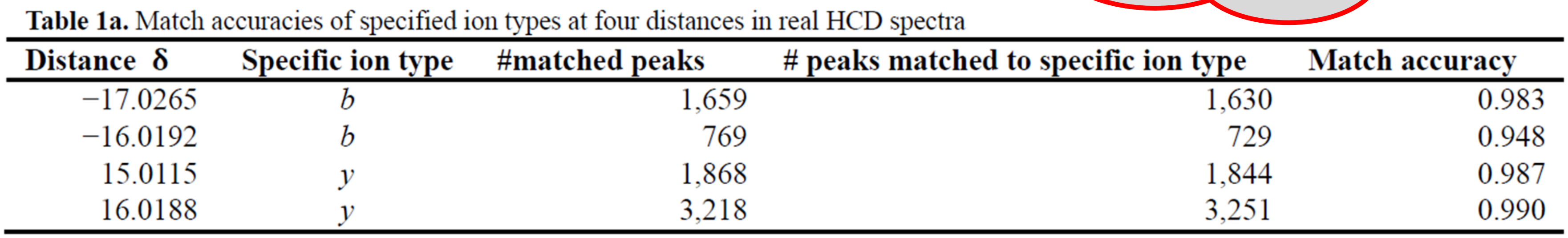


Table 1a. Match accuracies of specified ion types at four distances in real HCD spectra

Distance δ	Specific ion type	#matched peaks	# peaks matched to specific ion type	Match accuracy
−17.0265	*b*	1,659	1,630	0.983
−16.0192	*b*	769	729	0.948
15.0115	*y*	1,868	1,844	0.987
16.0188	*y*	3,218	3,251	0.990

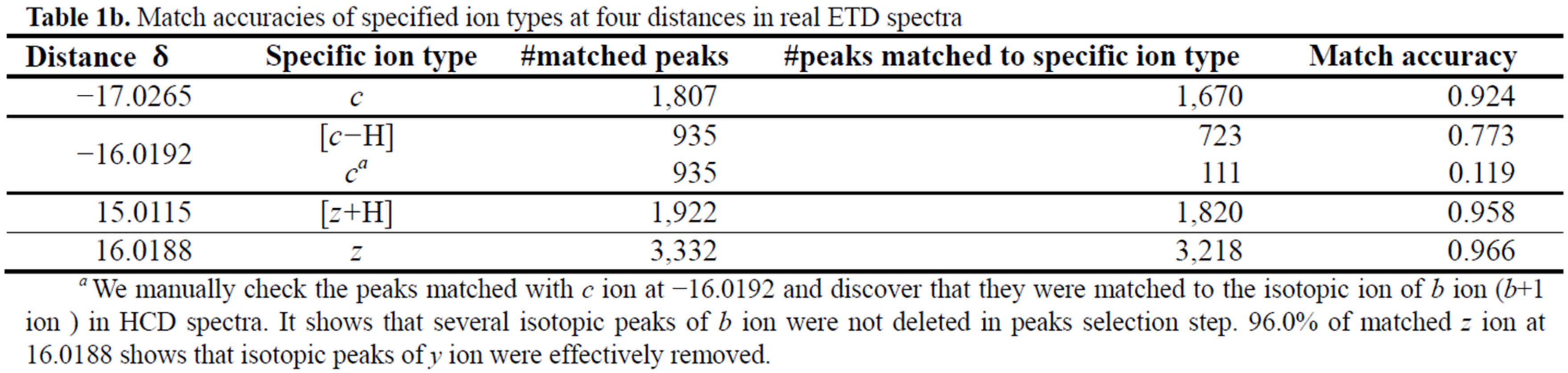


Table 1b. Match accuracies of specified ion types at four distances in real ETD spectra

Distance δ	Specific ion type	#matched peaks	#peaks matched to specific ion type	Match accuracy
−17.0265	*c*	1,807	1,670	0.924
−16.0192	[*c*−H]	935	723	0.773
	c[a]	935	111	0.119
15.0115	[*z*+H]	1,922	1,820	0.958
16.0188	*z*	3,332	3,218	0.966

[a] We manually check the peaks matched with *c* ion at −16.0192 and discover that they were matched to the isotopic ion of *b* ion (*b*+1 ion) in HCD spectra. It shows that several isotopic peaks of *b* ion were not deleted in peaks selection step. 96.0% of matched *z* ion at 16.0188 shows that isotopic peaks of *y* ion were effectively removed.

Match accuracies at different mass deviation

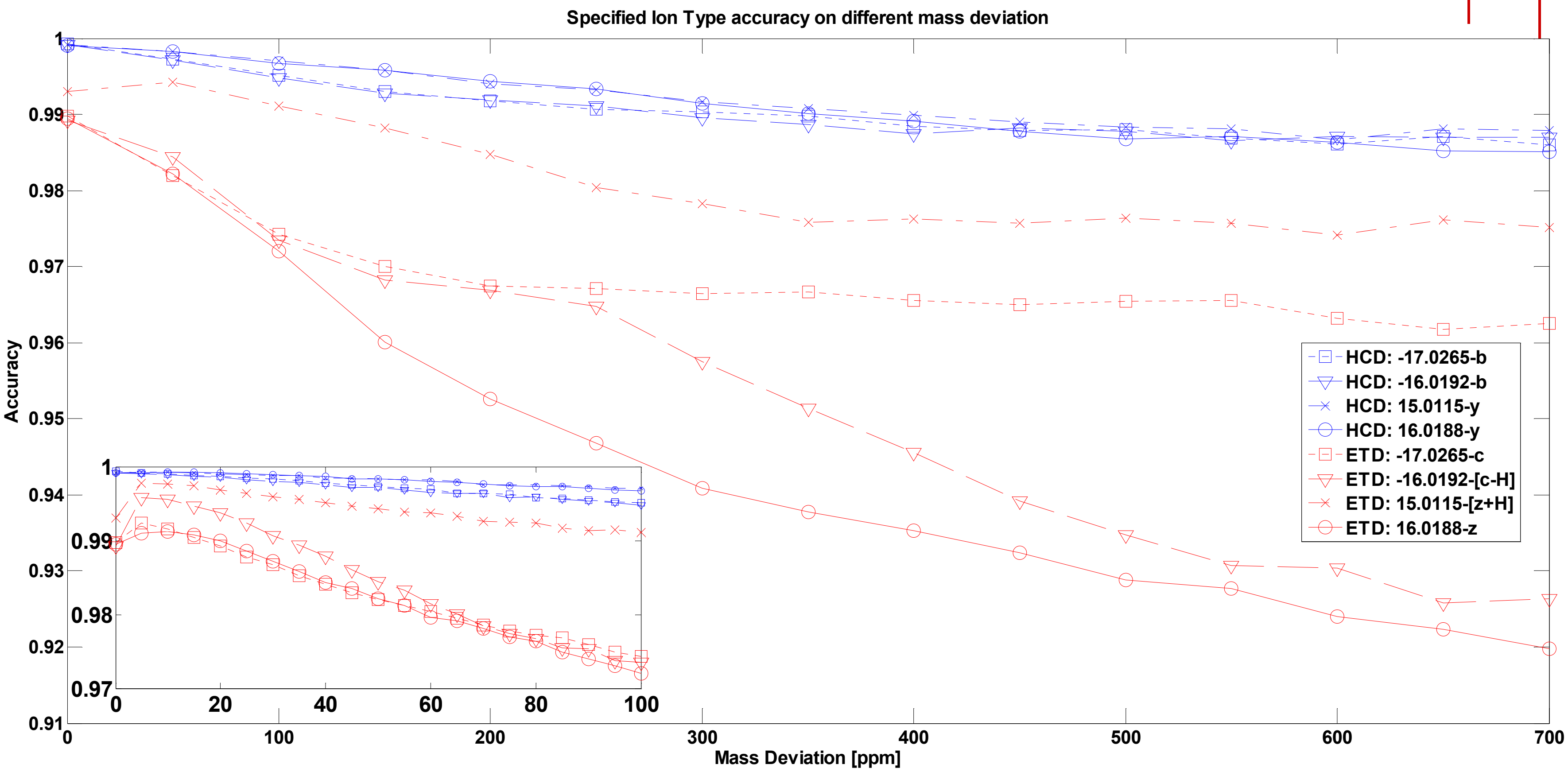


Figure S1. Accuracies of matched specified ion types at four distances in theoretical HCD and ETD spectra. We discover that the accuracy decreases at all distances. In addition, the accuracies of the ETD spectra are lower than the HCD spectra at all four distances. In the ETD spectra, we consider *c*, *z*, [*c*−H], and [*z*+H] ions, while we only consider *b* and *y* ions in the HCD spectra. So ETD spectra have a higher probability of being randomly matched to incorrect peaks. Furthermore, when the mass deviation is greater than ~ 250 ppm, the ETD data curves at −16.0192-[*c*−H] and 16.0188-*z* drop dramatically. The inset shows the details for mass deviations from 0 to 100 ppm.

Why not divide-and-conquer approach

(1)Some peak ion types would be determined incorrectly.

(2)If these peaks were used as pivot ions in the divide-and-conquer approach, then the true peptides would not be obtained.

Zhang, Z., *De novo peptide sequencing based on a divide-and-conquer algorithm and peptide tandem spectrum simulation. Anal. Chem., 2004.* ***76: p. 6374-6383.***
Bertsch, A., et al., *De novo peptide sequencing by tandem MS using complementary CID and electron transfer dissociation. Electrophoresis, 2009.* ***30(21): p. 3736-3747.***

Outline

(1)Methods

(1.1)Peak Selection

(1.2)Ion Type Determination

(1.3)Directed Acyclic Graph Construction

(1.3.1)Generating graph vertices

(1.3.2)Generating graph edges

(1.3.3)Finding the k longest paths

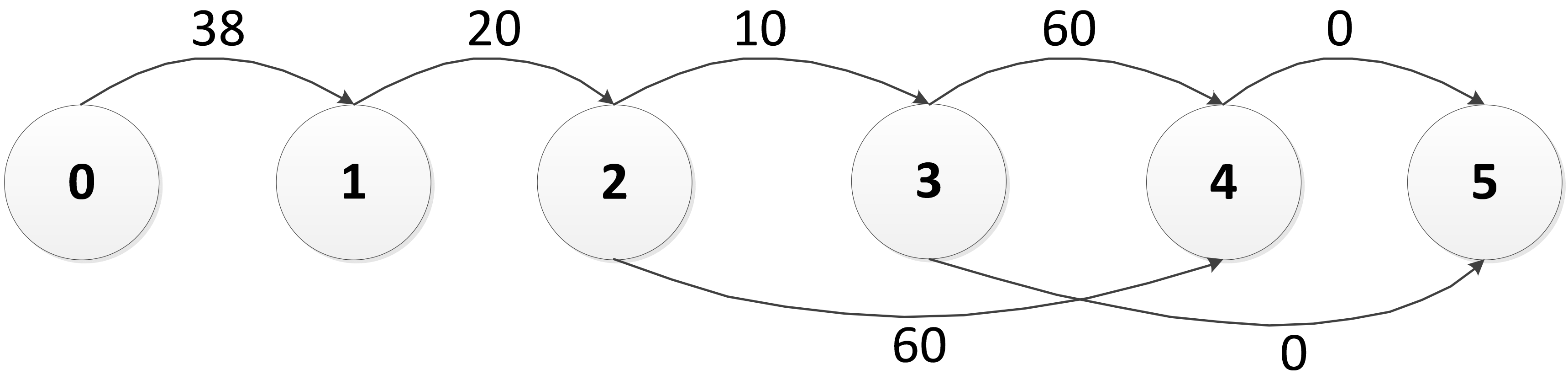
(1.3.4)Ranking all candidate peptides

(2)Materials and Results

Comparison between pNovo+ and PEAKS

Directed Acyclic Graph Construction

Directed Acyclic Graph



Unlike previously reported approaches

(1)The restriction of anti-symmetry is not considered;

(2)An efficient algorithm to find the *k* longest paths;

(3)We used a simple but effective approach to rank all of the candidate peptides.

Generating graph vertices

(1)*b* and *y* ions for HCD data and *c*, *z*, [*c*−H] and [z+H] ions for ETD data;

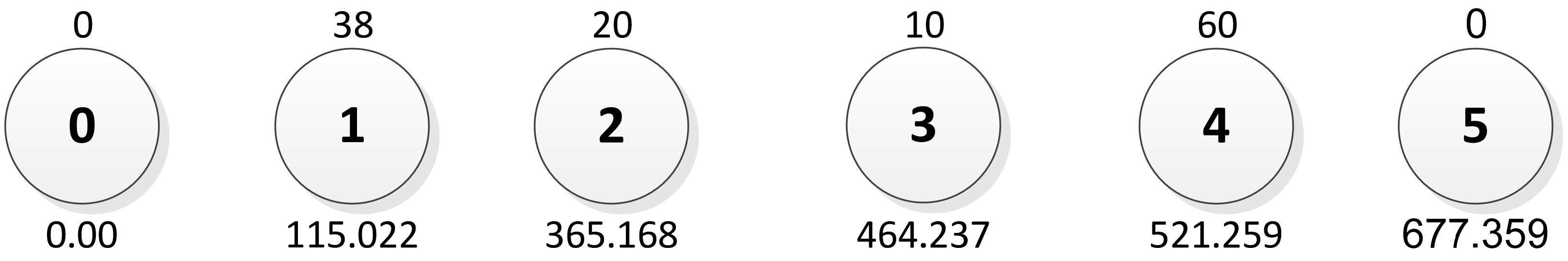
(2)All peaks were transformed to singly charged *b* ions;

(3)All peaks from HCD and ETD spectra were integrated into a new spectrum;

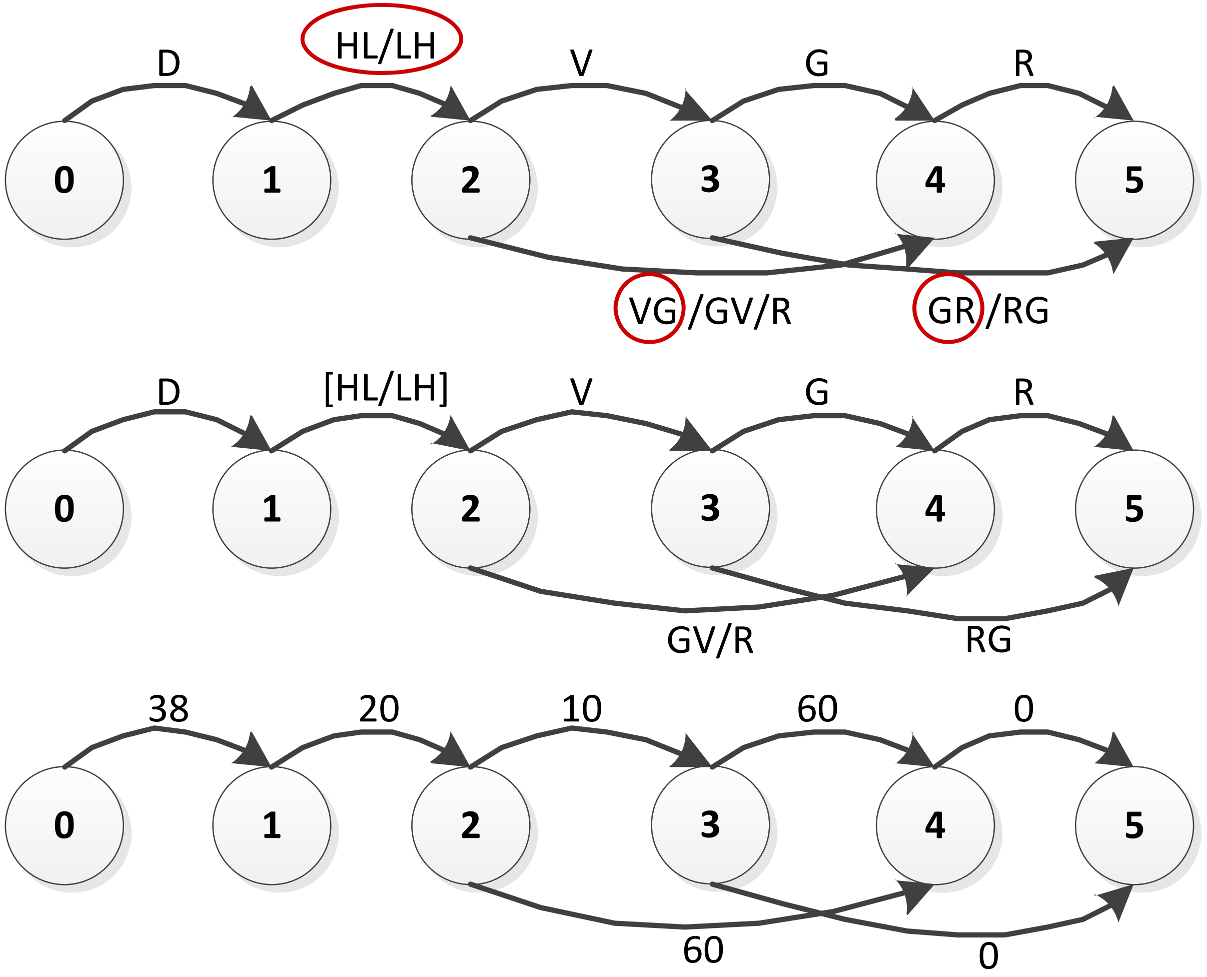
(4)If two or more vertices are of equal mass within a given tolerance range, they are merged as a new peak.

Generating graph vertices

(5)We add a source vertex and a destination vertex to each spectrum graph.



Generating graph edges



Why not anti-symmetry paths

(1)Every peak whose ion type is unknown is converted to more than one vertex; therefore, each such vertex has at least one “fake” vertex.

(2)However, the anti-symmetric longest path-finding problem is NP-hard.

(3)In addition, the longest path does not usually contain the correct peptide sequence.

Dancik, V., et al., *De novo peptide sequencing via tandem mass spectrometry. J Comput Biol, 1999.* ***6(3-4): p. 327-42.***

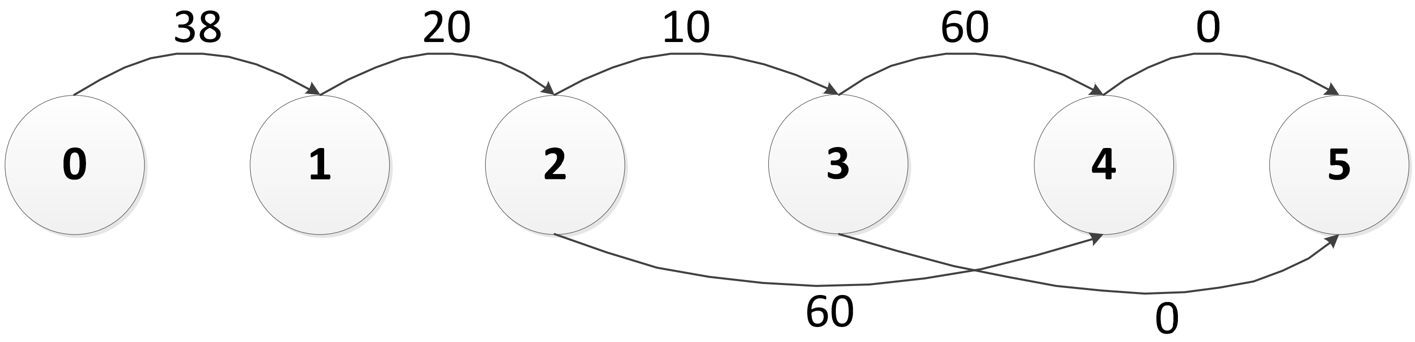
Why not anti-symmetry paths

Thus, we converted the path-finding problem to two sub-problems.

(1)The first one deals with how to find the *k* longest paths in a directed acyclic graph <u>without considering the restriction of anti-symmetry</u>,

(2)and the second one deals with how to generate candidate peptides from the *k* longest paths and rank them.

Finding the *k* longest paths



The time complexity of our algorithm is $O(|E|*\log d + k|V|*\log \bar{d})$.

The proof of the time complexity and pseudo code of the FKLP algorithm are shown in the Appendix.

Yen, S.H., D.H. Du, and S. Ghanta, *Efficient algorithms for extracting the K most critical paths in timing analysis, in Proceedings of the 26th ACM/IEEE Design Automation Conference. 1989, ACM: Las Vegas, Nevada, United States.* ***p. 649-654***.

Ju, Y.-C. and R.A. Saleh, *Incremental techniques for the identification of statically sensitizable critical paths, in Proceedings of the 28th ACM/IEEE Design Automation Conference. 1991, ACM: San Francisco, California, United States.* ***p. 541-546.***

Kundu, S., *An incremental algorithm for identification of longest (shortest) paths. Integr. VLSI J., 1994.* ***17(1): p. 25-31.***

Time complexity of FKLP

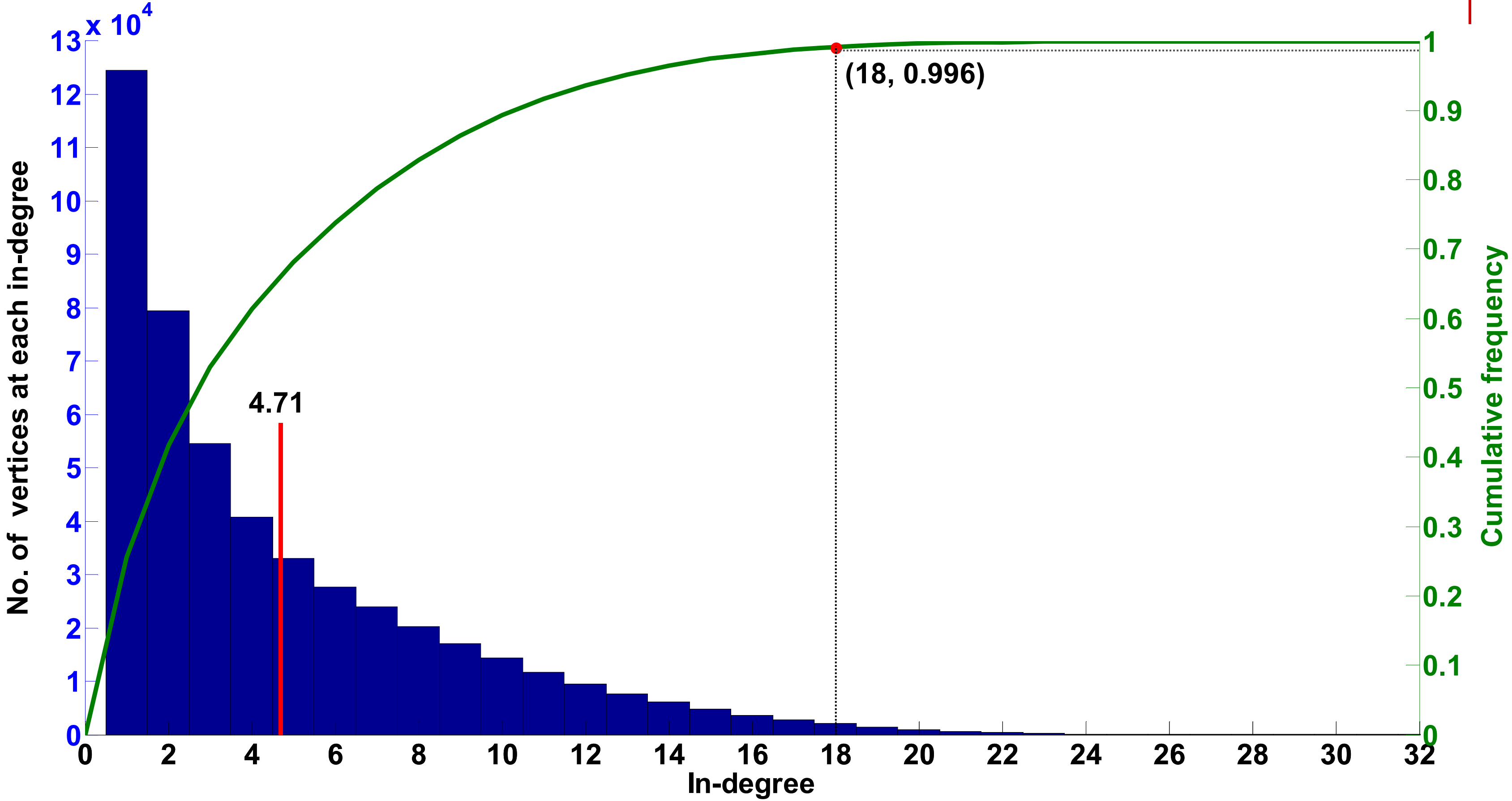
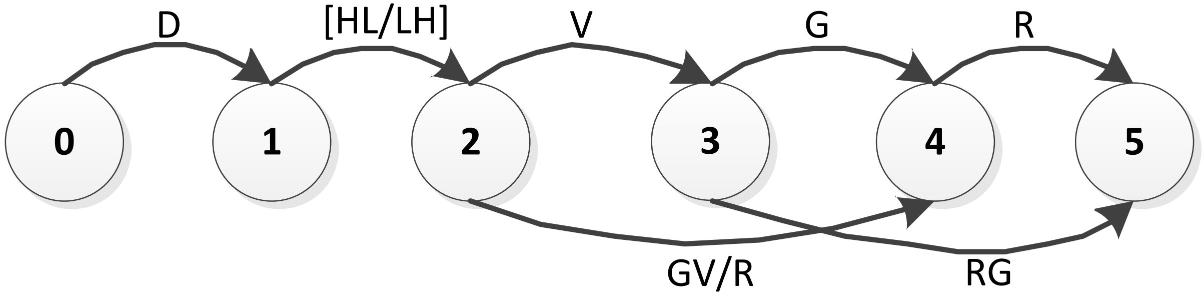


Figure S2. Number of vertices at each in-degree. The green curve is cumulative frequency of the histogram. The figure shows that 99.6% of vertices have an in-degree of 18 or less. The average in-degree of these vertices is 4.71 while most of the spectrum graphs in our dataset have more than 150 vertices, thus the spectrum graph is a sparse graph. The time complex of our algorithm is $O(|E|*\log d + k|V|*\log\overline{d})$, thus, the algorithm requires about linear time to the number of vertices.

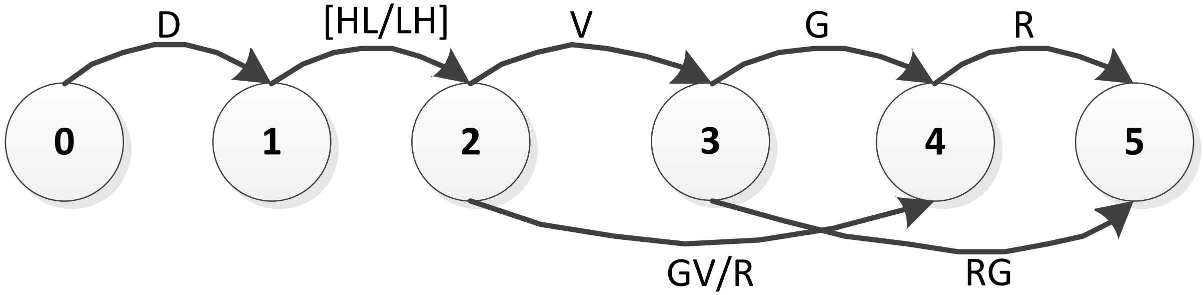
Ranking candidate peptides



(1)We use the breadth first search method to generate all candidate peptides from the *k* longest paths.

(2)The main problem in this step is how to correctly rank the candidate peptides.

Why not weights of the paths

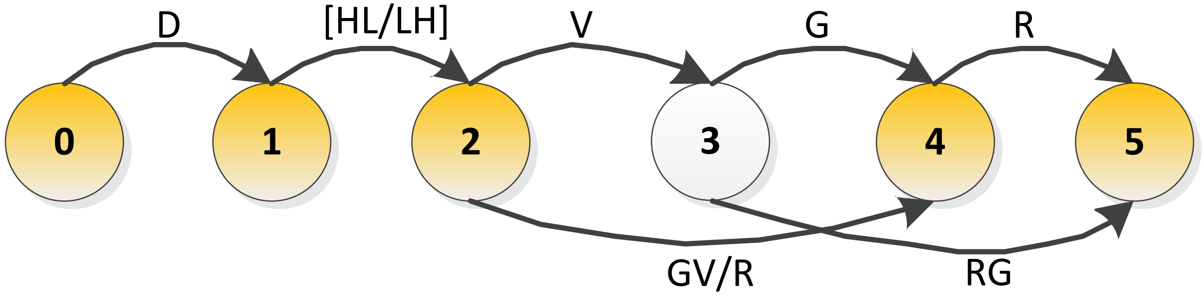


As a first intuition, we use the weights of paths to directly rank the peptides.

(1)However, there could more than one candidate peptide from a given path with the same path weight.

(2)The weights of the paths is not always perfect.

How to distinguish peptides with the same path weights



0→1→2→4→5

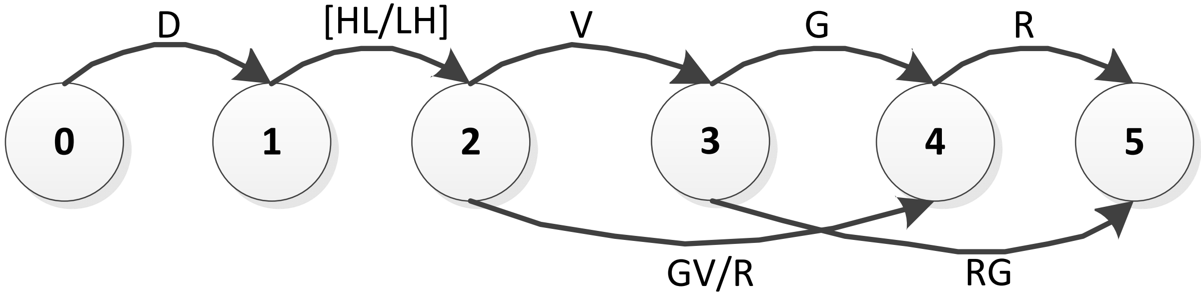
D[HL/LH]GVR
D[HL/LH]RR

We define GAP_{pep} for each candidate peptide:

$$GAP_{pep} = L_{pep} - L_{path}$$

L_{pep} is the number of amino acids in the peptide
L_{path} is the number of edges in the path

How to distinguish peptides with the same *GAP*



(a)

0→1→2→③→5	D[HL/LH](RG)
0→1→2→④→5	D[HL/LH](GV)R

(b)

0→1→2→4→5	D[HL/LH]RR
0→1→2→③→4→5	D[HL/LH]VGR

We rank each pair of peptides that has the same *GAP* by the average weight of the unique vertices.

Ranking candidate peptides

(1)In summary, we first rank all candidate peptide by *GAP*.

(2)Then we rank each pair of peptides that has the same *GAP* by the average weight of the unique vertices.

(3)If both the *GAP* and the average weight of the unique vertices are equal, we rank them by the sum of the mass deviations of each edge in ascending order.

Outline

(1)Methods

(1.1)Peak Selection

(1.2)Ion Type Determination

(1.3)Directed Acyclic Graph Construction

(1.3.1)Generating graph vertices

(1.3.2)Generating graph edges

(1.3.3)Finding the k longest paths

(1.3.4)Ranking all candidate peptides

(2)Materials and Results

Comparison between pNovo+ and PEAKS

MS/MS Data

(1)8-protein STD, HCD and ETD spectrum pairs generated by the same precursor ion

(2)Enzyme: Asp-N, Elastease, Glu-C, Lys-C and Trypsin

(3)pFind 2.6 database search

(1)The mass tolerances of both precursor and fragment ions was set to ±20 ppm.
(2)Carbamidomethyl on cysteines and oxidation on methionines were set as variable modifications.
(3)pBuild is used to filter out the results under the 1% FDR control at the spectrum level.
(4)Then we select the spectrum pairs whose identified peptides are the same and lengths are between 6 and 19.

Search parameter

(1)A set of 1,144 spectrum pairs remain as a test set, including 170 from Asp-N, 388 from Elastease, 149 from Glu-C, 231 from Lys-C and 206 from Trypsin.

(2)pNovo+ and PEAKS 5.3

The parameters kept similar as pFind 2.6.

Additionally, for all test sets, we had also run PEAKS by setting the error tolerance of the fragment ions to ±0.01 Da and ±0.02 Da . At last, we chose the result at ±0.01 Da because PEAKS at ±0.01 Da was slightly better than the results at ±0.02 Da.

Comparison between pNovo+ and PEAKS

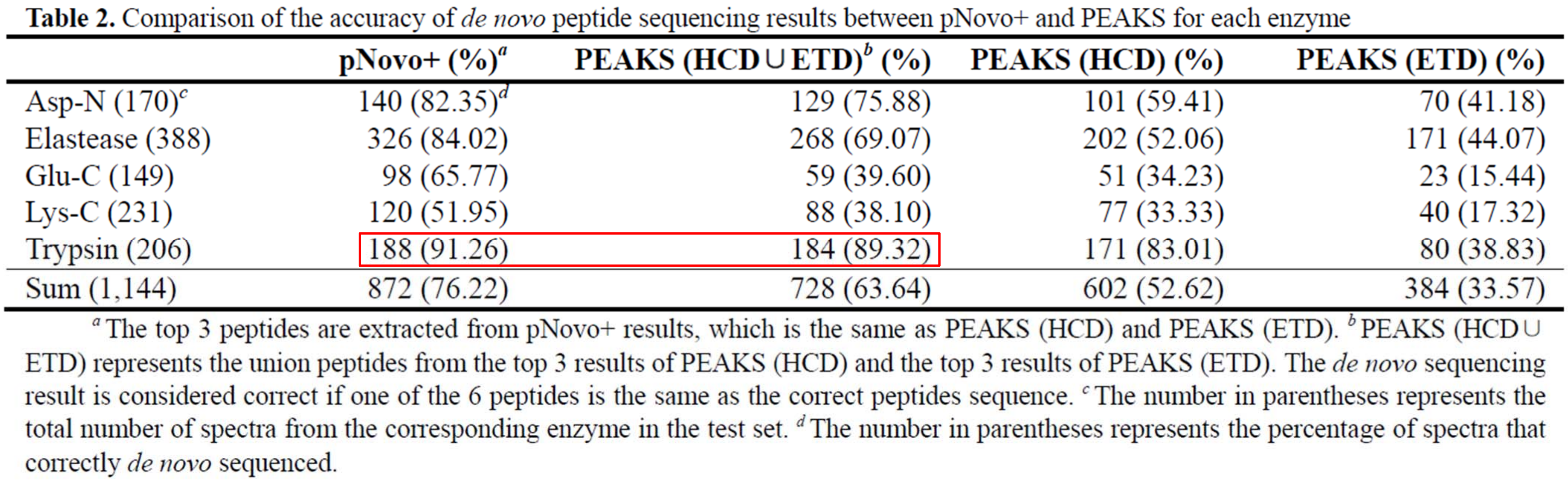


Table 2. Comparison of the accuracy of *de novo* peptide sequencing results between pNovo+ and PEAKS for each enzyme

	pNovo+ (%)[a]	PEAKS (HCD∪ETD)[b] (%)	PEAKS (HCD) (%)	PEAKS (ETD) (%)
Asp-N (170)[c]	140 (82.35)[d]	129 (75.88)	101 (59.41)	70 (41.18)
Elastease (388)	326 (84.02)	268 (69.07)	202 (52.06)	171 (44.07)
Glu-C (149)	98 (65.77)	59 (39.60)	51 (34.23)	23 (15.44)
Lys-C (231)	120 (51.95)	88 (38.10)	77 (33.33)	40 (17.32)
Trypsin (206)	188 (91.26)	184 (89.32)	171 (83.01)	80 (38.83)
Sum (1,144)	872 (76.22)	728 (63.64)	602 (52.62)	384 (33.57)

[a] The top 3 peptides are extracted from pNovo+ results, which is the same as PEAKS (HCD) and PEAKS (ETD). [b] PEAKS (HCD∪ETD) represents the union peptides from the top 3 results of PEAKS (HCD) and the top 3 results of PEAKS (ETD). The *de novo* sequencing result is considered correct if one of the 6 peptides is the same as the correct peptides sequence. [c] The number in parentheses represents the total number of spectra from the corresponding enzyme in the test set. [d] The number in parentheses represents the percentage of spectra that correctly *de novo* sequenced.

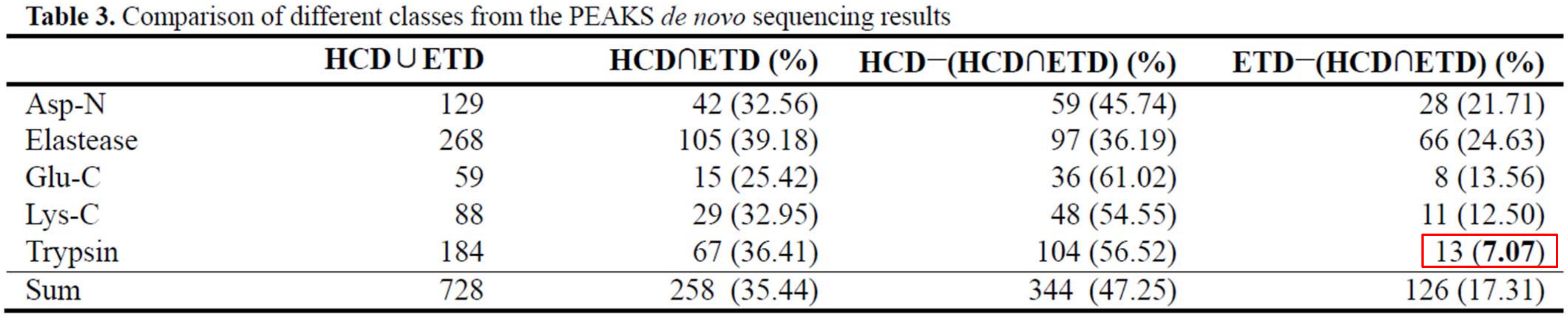


Table 3. Comparison of different classes from the PEAKS *de novo* sequencing results

	HCD∪ETD	HCD∩ETD (%)	HCD−(HCD∩ETD) (%)	ETD−(HCD∩ETD) (%)
Asp-N	129	42 (32.56)	59 (45.74)	28 (21.71)
Elastease	268	105 (39.18)	97 (36.19)	66 (24.63)
Glu-C	59	15 (25.42)	36 (61.02)	8 (13.56)
Lys-C	88	29 (32.95)	48 (54.55)	11 (12.50)
Trypsin	184	67 (36.41)	104 (56.52)	13 (**7.07**)
Sum	728	258 (35.44)	344 (47.25)	126 (17.31)

Comparison between pNovo+ and PEAKS

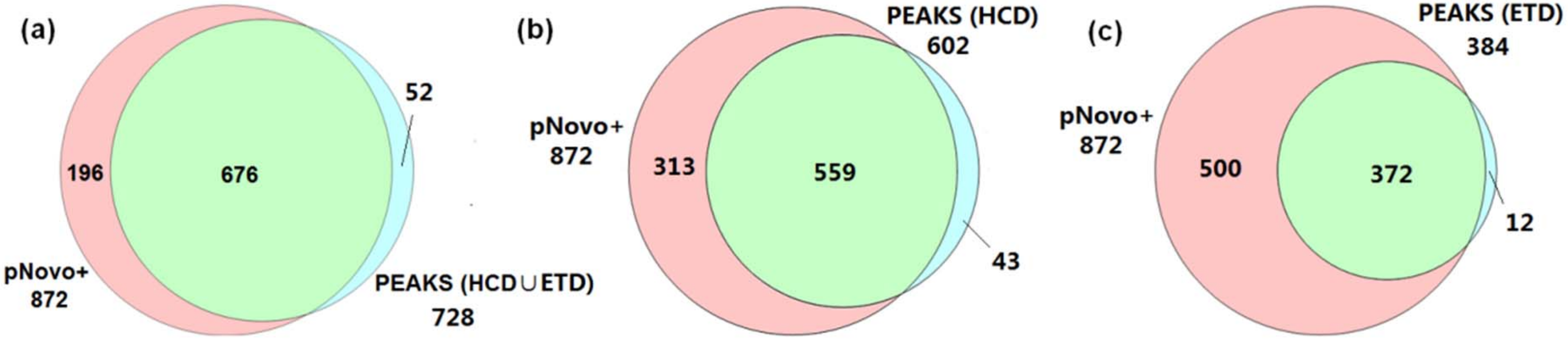


Figure 4. Comparisons of the consistency between (a) pNovo+ and PEAKS (HCD∪ETD), (b) pNovo+ and PEAKS (HCD) and (c) pNovo+ and PEAKS (ETD). The numbers of correctly identified spectra are shown.

In the 52 spectra that were uniquely identified by PEAKS, 40% of them were identified by pNovo+ in top 10 but not in top 3. While in the 196 spectra that were uniquely identified by pNovo+, only 15% of them were identified by PEAKS in top 10 but not in top 3.

Comparison between pNovo+ and PEAKS

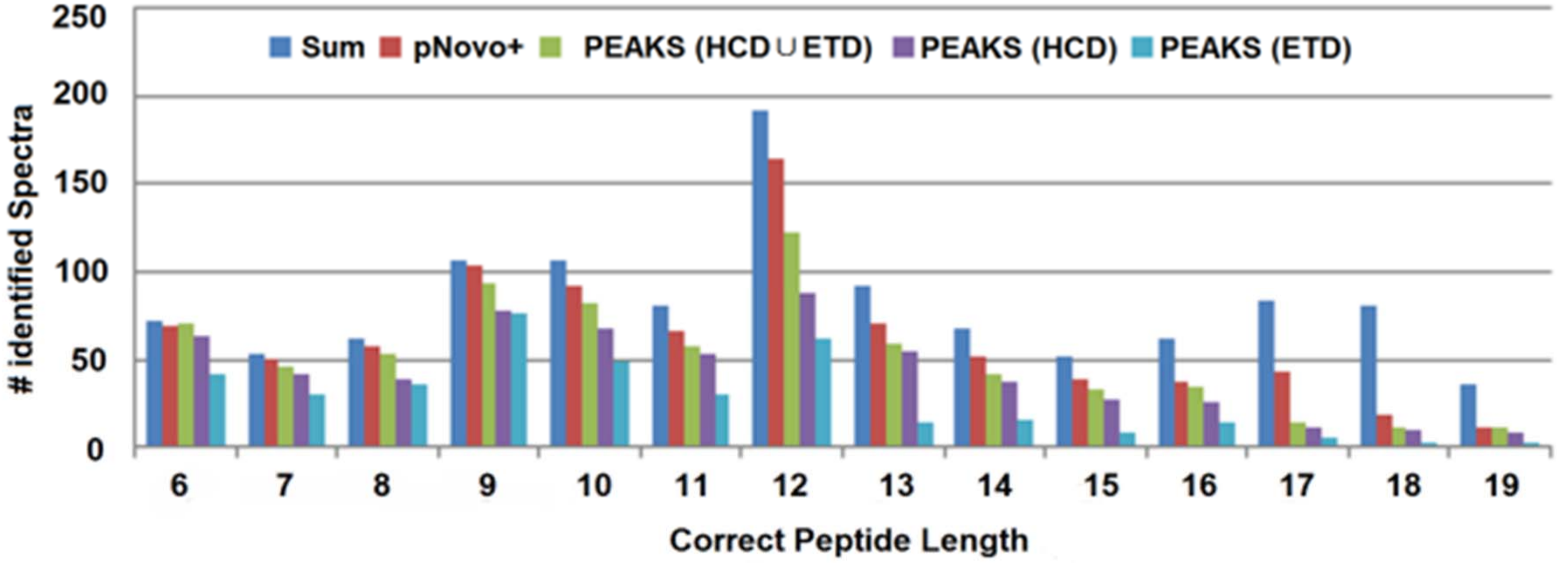


Figure 5. Comparison between pNovo+ and PEAKS with regard to different peptide lengths. The percentage of identified peptides clearly decreased with increasing peptide length. However, the percentage decreased more significantly for PEAKS (HCD∪ETD) compared with pNovo+.

As the peptide length increases, the percentages of mass spectra that are correctly identified using pNovo+ and PEAKS (HCD∪ETD) both decreased. However, at almost every peptide length, pNovo+ identified more spectra, especially when the peptide length is greater than 8.

Comparison between pNovo+ and PEAKS

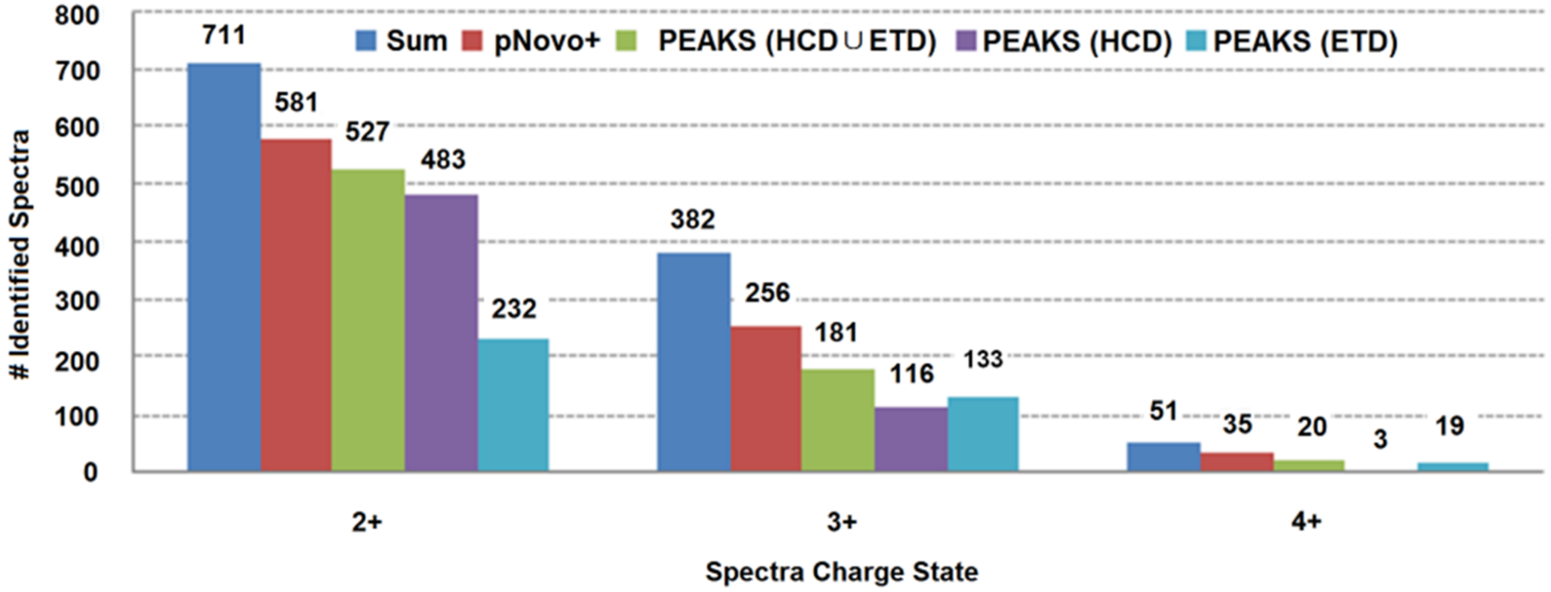


Figure 6. Comparison between pNovo+ and PEAKS with regard to different charge states. The number above each bar indicates how many spectra correctly identified.

We observed that unlike charge state 2+, in charge state 3+ and 4+ PEAKS (ETD) identified more spectra than PEAKS (HCD). This indicates that ETD is more suitable for higher charge states than HCD. Additionally, pNovo+ improves on PEAKS (HCD∪ETD) by 45.4% for charge state 3+ and 75.0% for charge state 4+, much higher than the 10.2% improvement for charge state 2+.

Comparison between pNovo+ and PEAKS

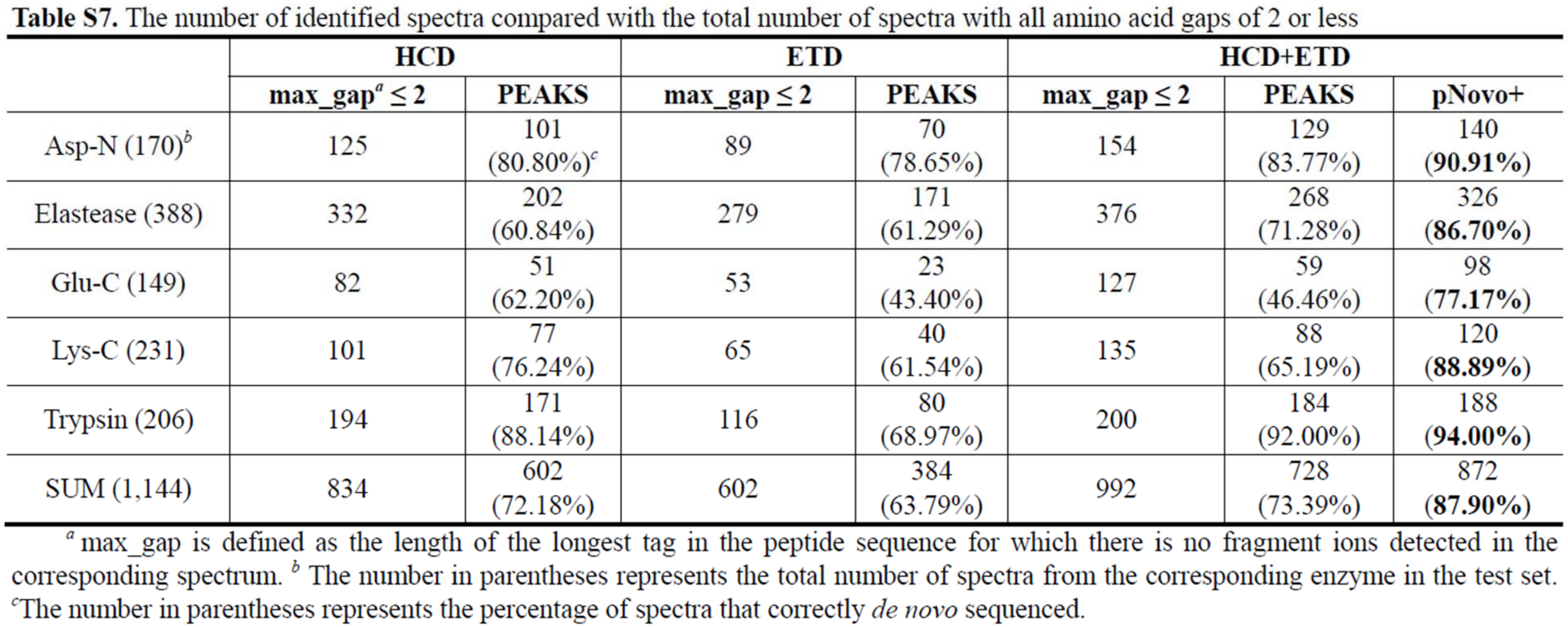


Table S7. The number of identified spectra compared with the total number of spectra with all amino acid gaps of 2 or less

	HCD		**ETD**		**HCD+ETD**		
	max_gap[a] ≤ 2	**PEAKS**	**max_gap ≤ 2**	**PEAKS**	**max_gap ≤ 2**	**PEAKS**	**pNovo+**
Asp-N (170)[b]	125	101 (80.80%)[c]	89	70 (78.65%)	154	129 (83.77%)	140 **(90.91%)**
Elastease (388)	332	202 (60.84%)	279	171 (61.29%)	376	268 (71.28%)	326 **(86.70%)**
Glu-C (149)	82	51 (62.20%)	53	23 (43.40%)	127	59 (46.46%)	98 **(77.17%)**
Lys-C (231)	101	77 (76.24%)	65	40 (61.54%)	135	88 (65.19%)	120 **(88.89%)**
Trypsin (206)	194	171 (88.14%)	116	80 (68.97%)	200	184 (92.00%)	188 **(94.00%)**
SUM (1,144)	834	602 (72.18%)	602	384 (63.79%)	992	728 (73.39%)	872 **(87.90%)**

[a] max_gap is defined as the length of the longest tag in the peptide sequence for which there is no fragment ions detected in the corresponding spectrum. [b] The number in parentheses represents the total number of spectra from the corresponding enzyme in the test set. [c] The number in parentheses represents the percentage of spectra that correctly *de novo* sequenced.

For spectra with maximum gap lengths of 2 or less, pNovo+ yielded nearly 90% of correct identifications, which indicates that pNovo+ takes full advantage of the complementary HCD and ETD data.

Running time comparison

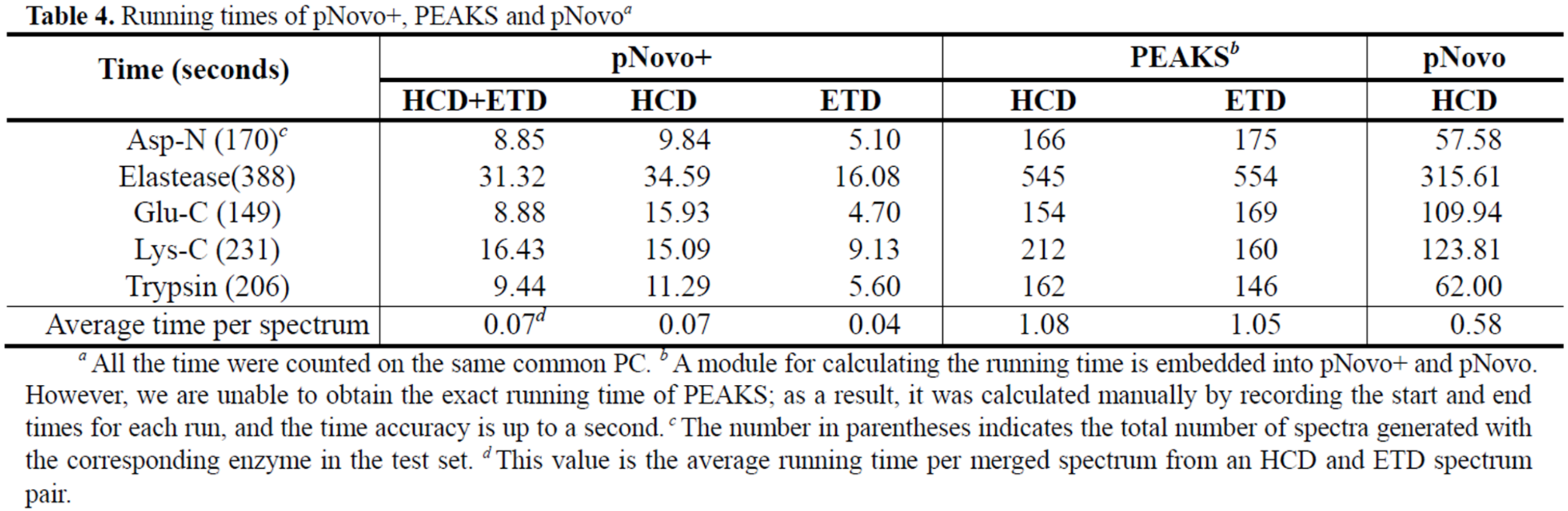


Table 4. Running times of pNovo+, PEAKS and pNovo[a]

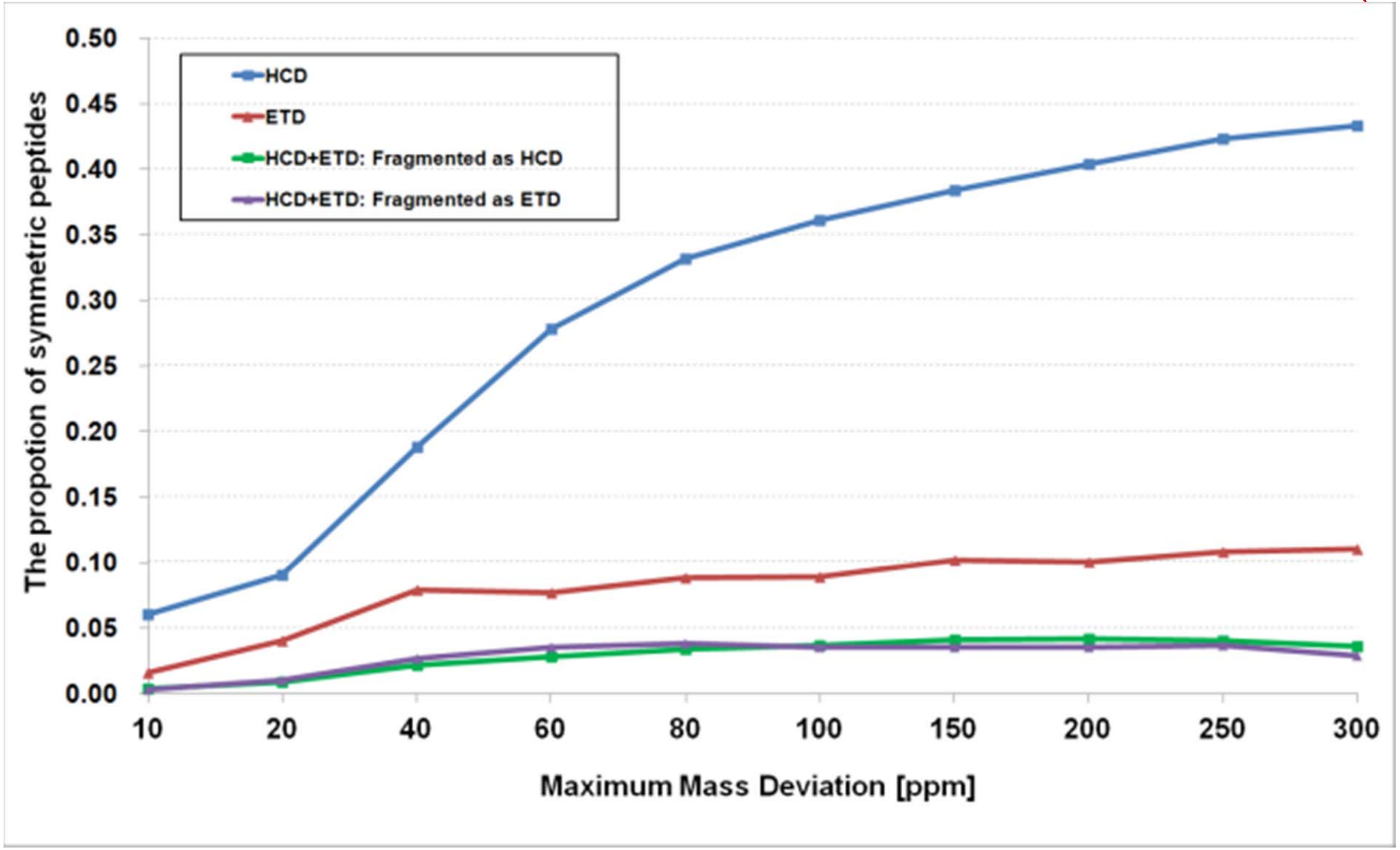
Time (seconds)	pNovo+			PEAKS[b]		pNovo
	HCD+ETD	HCD	ETD	HCD	ETD	HCD
Asp-N (170)[c]	8.85	9.84	5.10	166	175	57.58
Elastease(388)	31.32	34.59	16.08	545	554	315.61
Glu-C (149)	8.88	15.93	4.70	154	169	109.94
Lys-C (231)	16.43	15.09	9.13	212	160	123.81
Trypsin (206)	9.44	11.29	5.60	162	146	62.00
Average time per spectrum	0.07[d]	0.07	0.04	1.08	1.05	0.58

[a] All the time were counted on the same common PC. [b] A module for calculating the running time is embedded into pNovo+ and pNovo. However, we are unable to obtain the exact running time of PEAKS; as a result, it was calculated manually by recording the start and end times for each run, and the time accuracy is up to a second. [c] The number in parentheses indicates the total number of spectra generated with the corresponding enzyme in the test set. [d] This value is the average running time per merged spectrum from an HCD and ETD spectrum pair.

The generation and scoring of candidate peptides is the most time-consuming part of pNovo. However, with the fast FKLP algorithm, pNovo+ is very efficient at finding the *k* longest paths. Compared with the running time of the de novo sequencing algorithms presented in the previous study, the speed of pNovo+ is 10 to 100 times faster than other reported tools.

Andreotti, S., G.W. Klau, and K. Reinert, *Antilope-A Lagrangian Relaxation Approach to the de novo Peptide Sequencing Problem. IEEE/ACM Trans Comput Biol Bioinform, 2011.* ***PP Issue:99 p. 1-1.***

Why the anti-symmetry restriction is not necessary



Thank you for your attention!