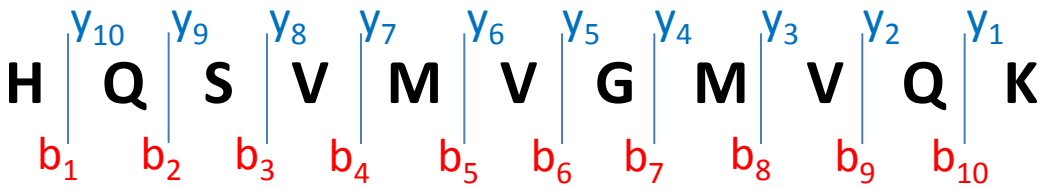
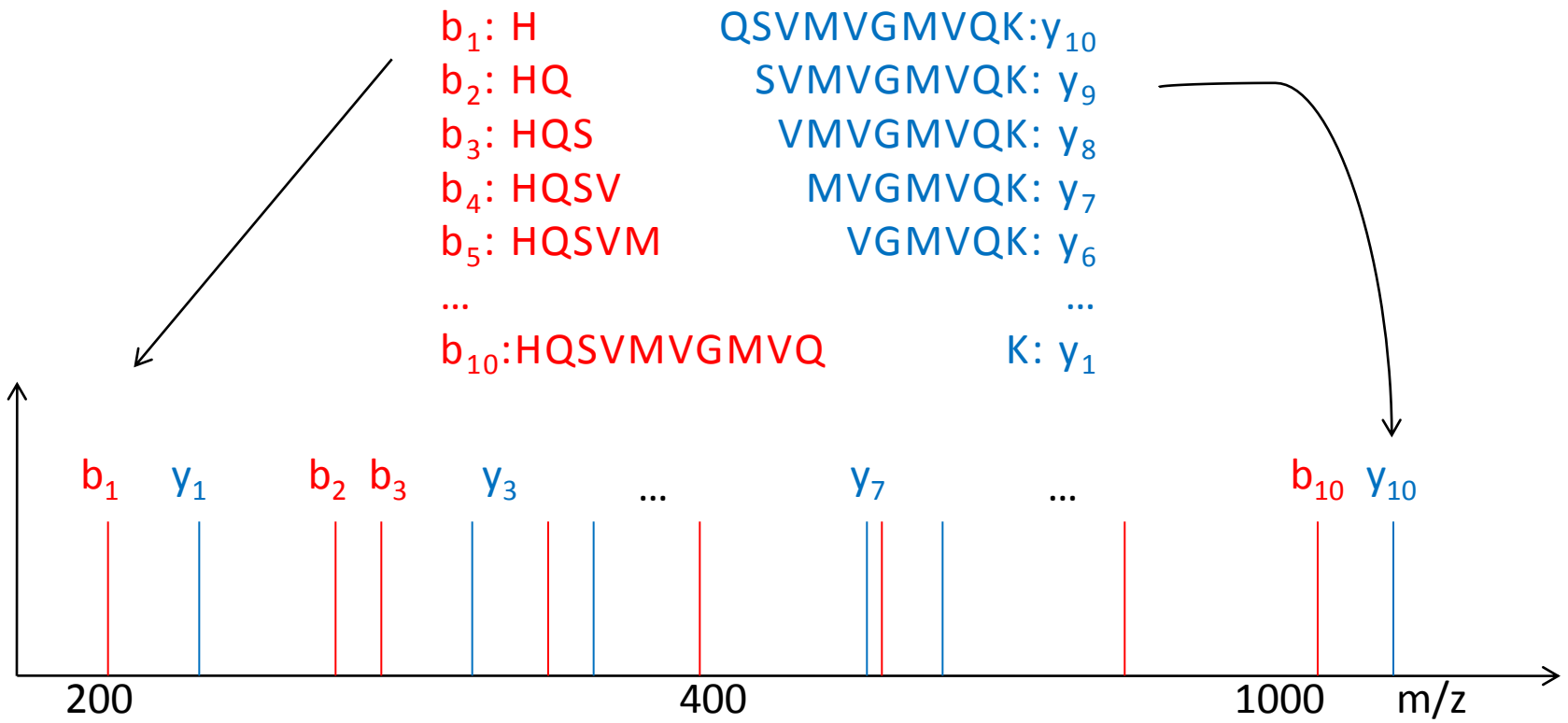
PTMSearch: a Greedy Tree Traversal Algorithm for Finding Protein Post-Translational Modifications in Tandem Mass Spectra

Attila Kertesz-Farkas, Beata Reiz, Michael P. Myers, and Sandor Pongor

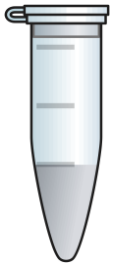
ICGEB, Trieste, Italy

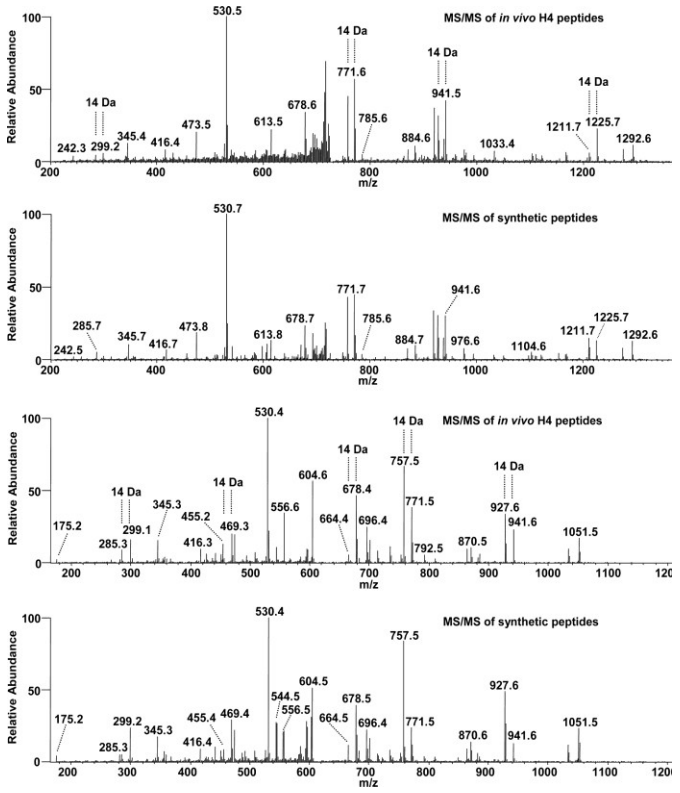
Mass spectrum





Mass Spectrometry





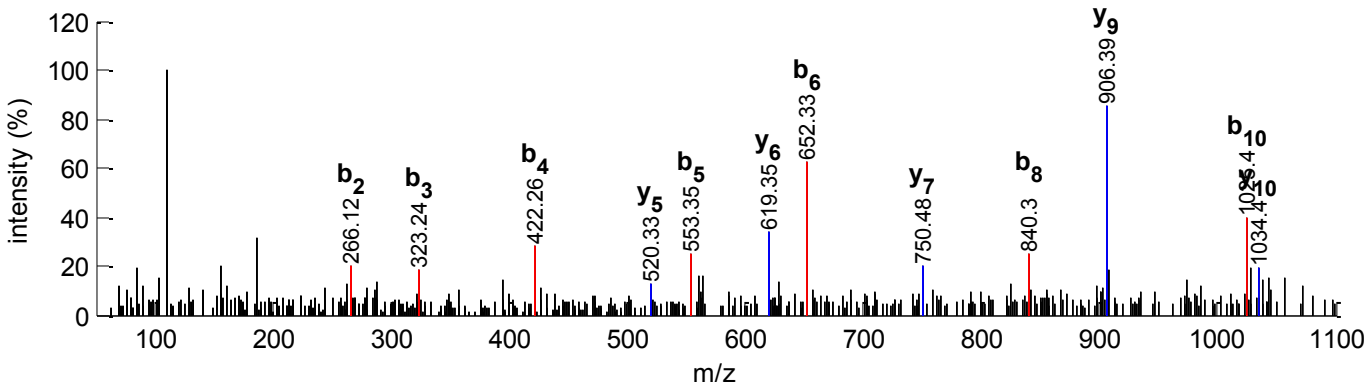
Protein mixture

Mass Spectrometer

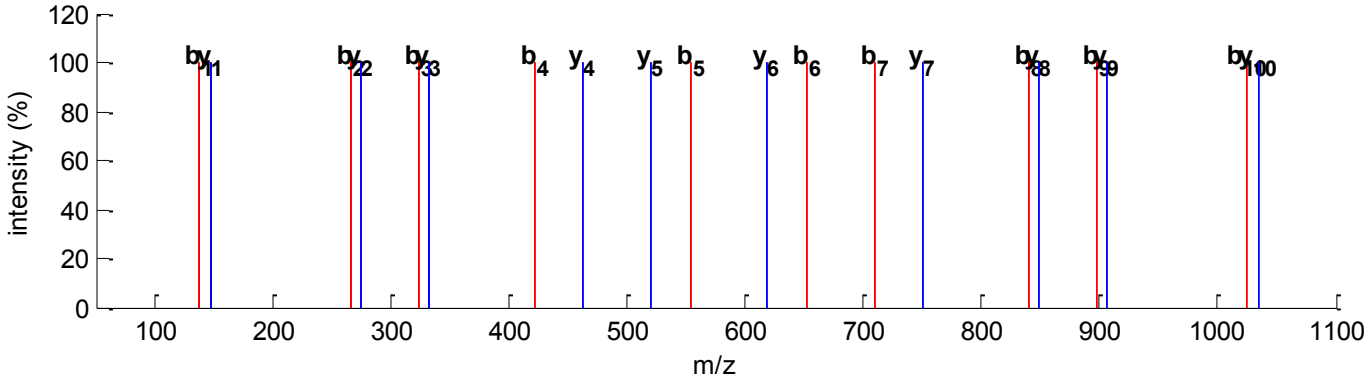
Thousands of spectra to analyze

Experimental Spectra Annotation

Experimental spectrum:



Shared Peak Count: 12



Theoretical spectrum of the peptide: HQSVMVGMVQK

DB of peptides:

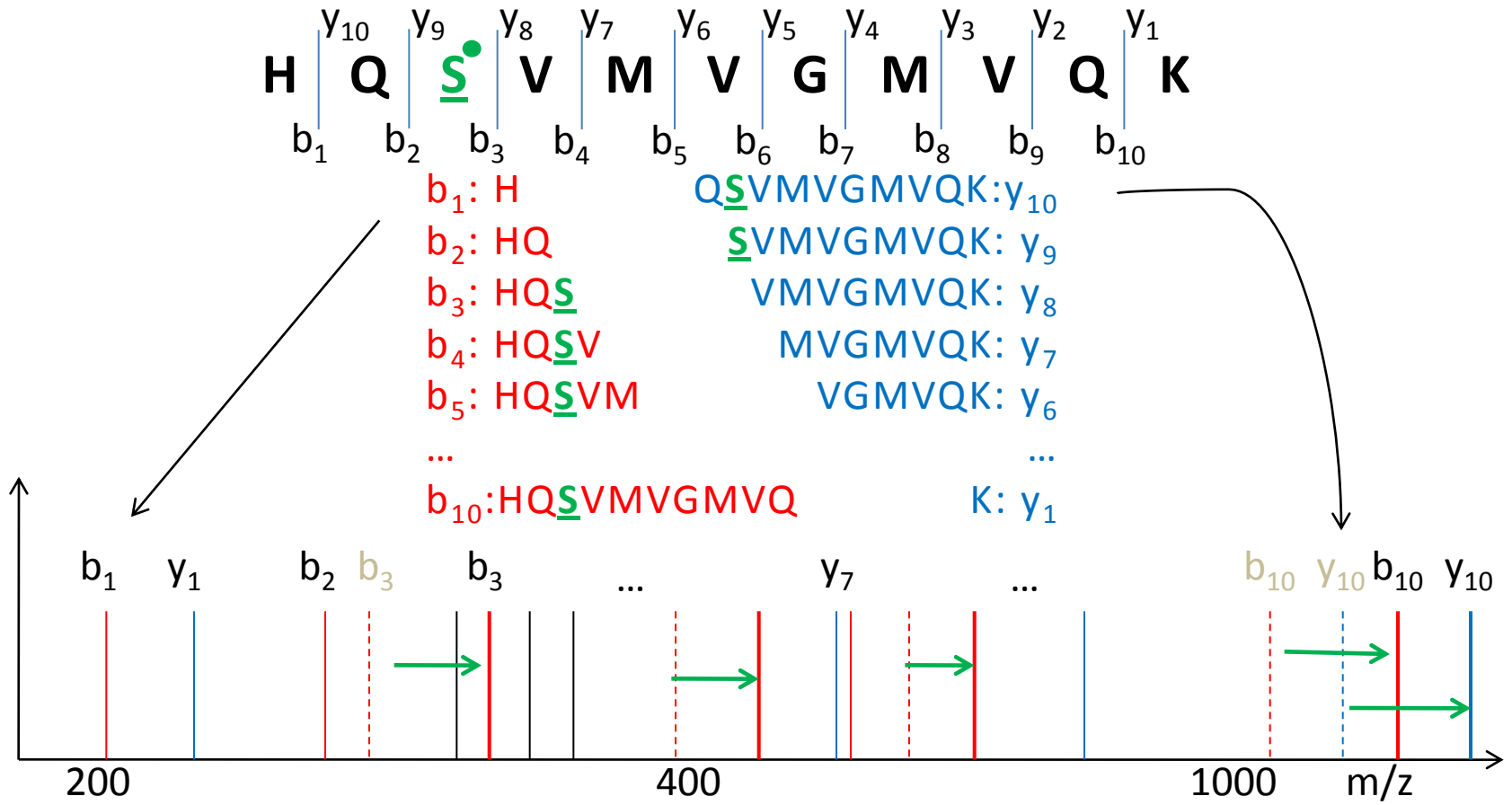
AQDWSK
QSDSSWRQK
QWEEFGEETK
QWESVRGTK
ERTGFDWK
ERTWERTGR
EWRTGHJKK
YJNFGTR
WERFFGDFHR
…
…
HQSVMVGMVQK
…
…

Post-Translational Modifications (PTMs)

- In nature, proteins undergo extensive chemical modifications that fine-tune their functions and structures.
- Modified proteins cannot be recognized using simple theoretical spectra.

Post-Translational Modifications (PTMs)

- PTM alters the weight of amino acids and the peptide that results peak shifts in the spectrum:



Obstacles

1. Complexity
   – longer execution time

# modifications	# modified theoretical spectra
0 modification:	1
#modifications: K Peptide length: L PTMs per amino acids: M	$\binom{L}{K} M^K$

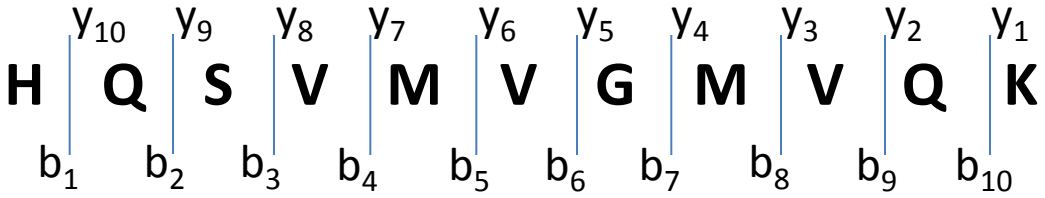
2. Significance of the hit
   – Inserting many PTMs make the theoretical spectra too flexible and in the end all experimental spectra can be aligned to a theoretical spectrum.

Our method: PTMSearch

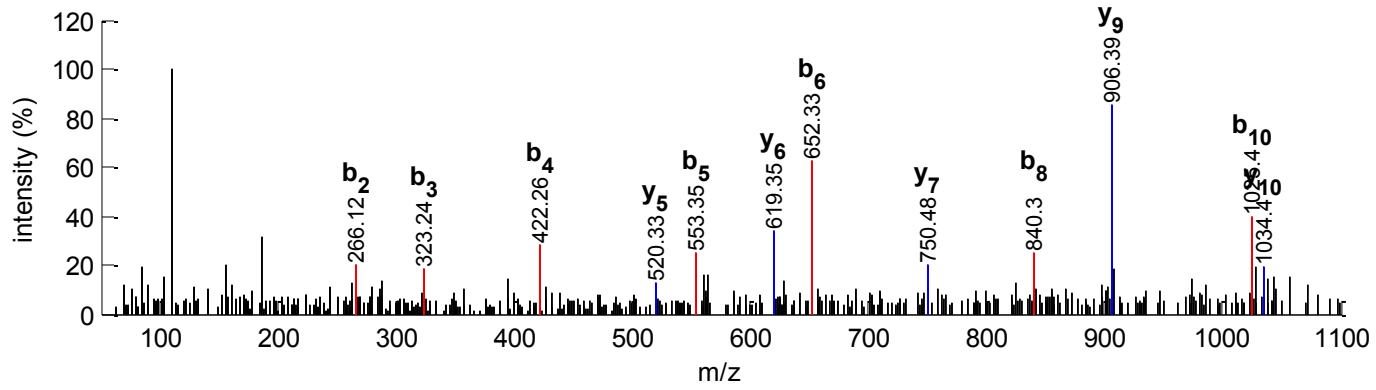
- A new, algorithm for identifying PTMs in the mass spectrum.
- It uses a database of potential PTMs.
- It represents the search space as a tree, and uses a tree traversal to find PTMs.

PTMSearch

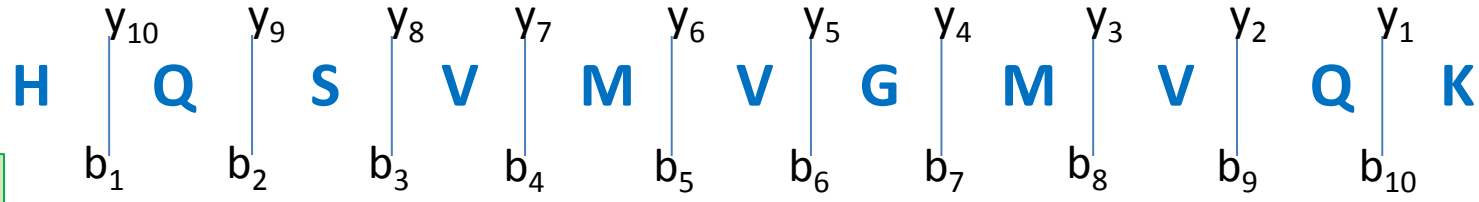
- Theoretical peptide (molecular mass:1245.01Da):



- And an experimental spectrum (mass: 1324.98 Da)



- Mass difference: Δ=79.97 Dalton

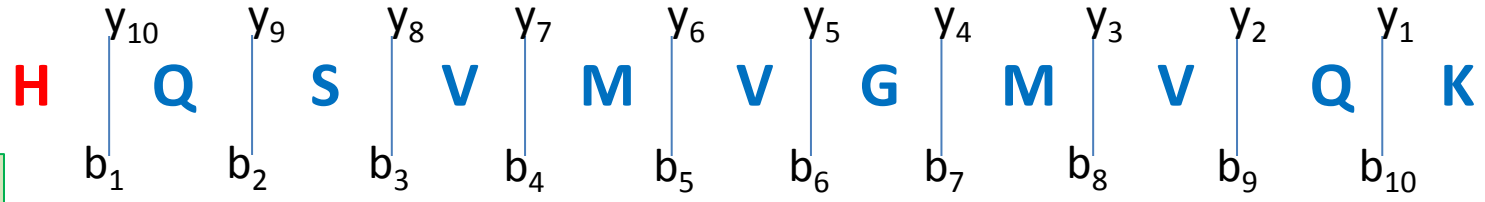


DB of PTMs

H :	**14.01, 57.02**
Q :	**--**
S :	**79.97**
V:	**-**
K:	**238.40**
…	

b_0:0
y_{11}:1324.98
S: 0
Δ: 79.97
Mod: --

Level: 0



DB of PTMs

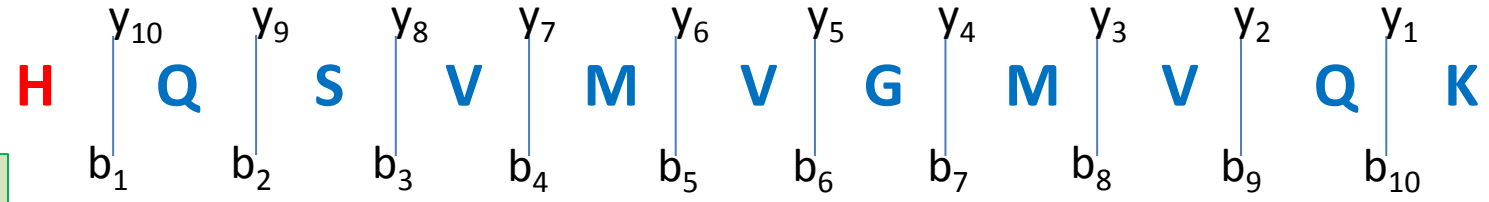
H :	**14.01, 57.02**
Q :	**--**
S :	**79.97**
V:	**-**
K:	**238.40**
…	

b_0:0
y_{11}:1324.98
S: 0
Δ: 79.97
Mod: --

Level: 0

b_1:137.04
y_{10}:1187.94
S: 2
Δ: 79.97
Mod: --

Level: 1



DB of PTMs

H : 14.01, 57.02
Q : --
S : 79.97
V: -
K: 238.40
…

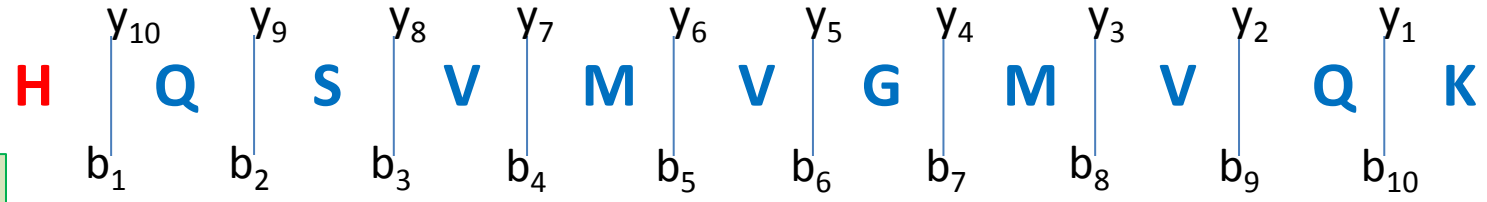
b_0:0
y_{11}:1324.98
S: 0
Δ: 79.97
Mod: --

Level: 0

b_1:137.04
y_{10}:1187.94
S: 2
Δ: 79.97
Mod: --

b_1:151.05
y_{10}:1173.93
S: 1
Δ: 65.96
Mod: 14.01

Level: 1



DB of PTMs

H : 14.01, 57.02
Q : --
S : 79.97
V: -
K: 238.40
…

b_0:0
y_{11}:1324.98
S: 0
Δ: 79.97
Mod: --

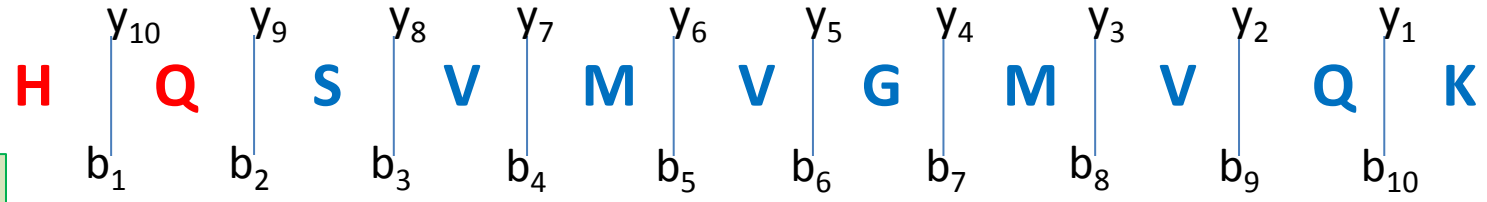
Level: 0

b_1:137.04
y_{10}:1187.94
S: 2
Δ: 79.97
Mod: --

b_1:151.05
y_{10}:1173.93
S: 1
Δ: 65.96
Mod: 14.01

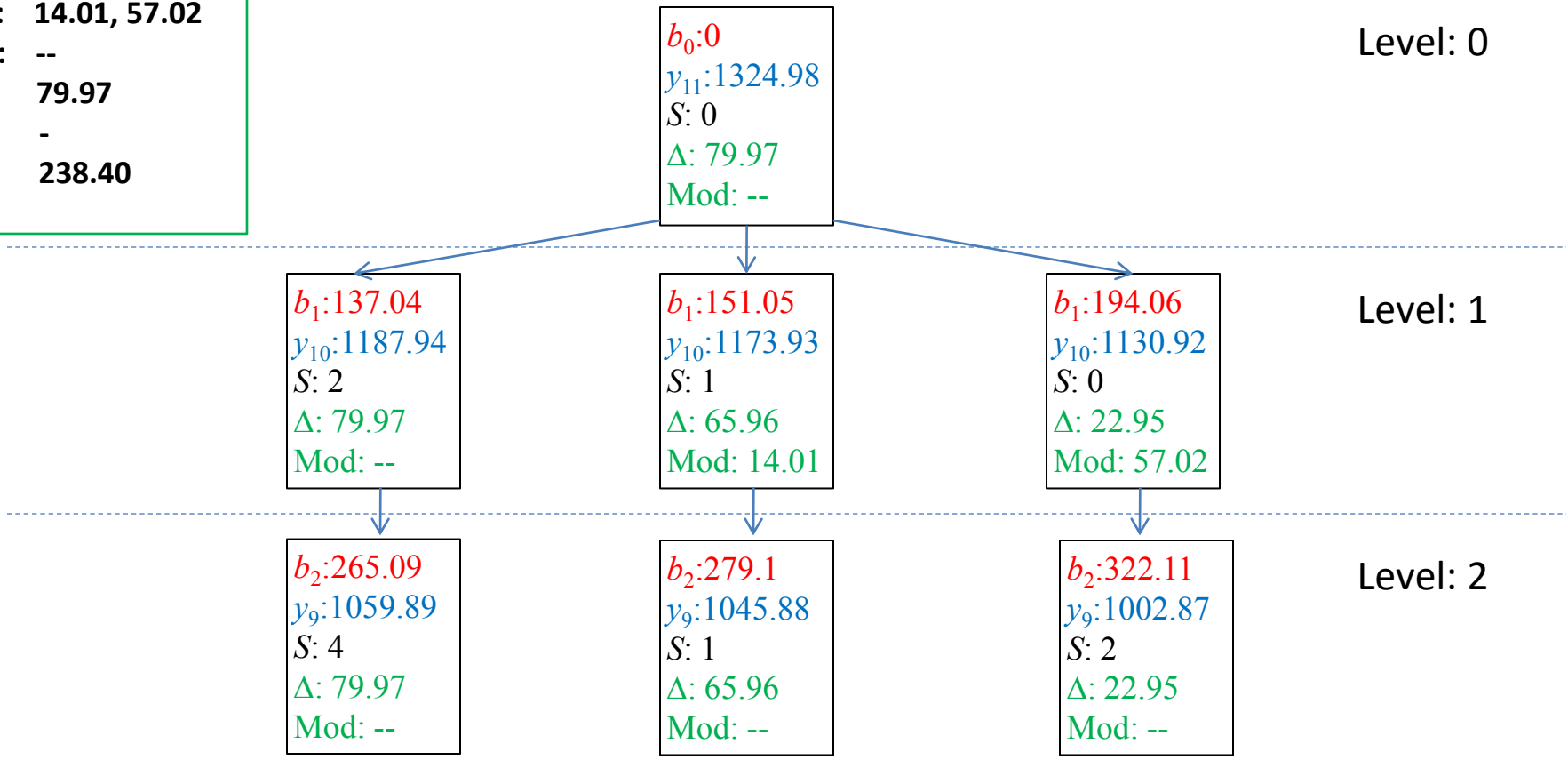
b_1:194.06
y_{10}:1130.92
S: 0
Δ: 22.95
Mod: 57.02

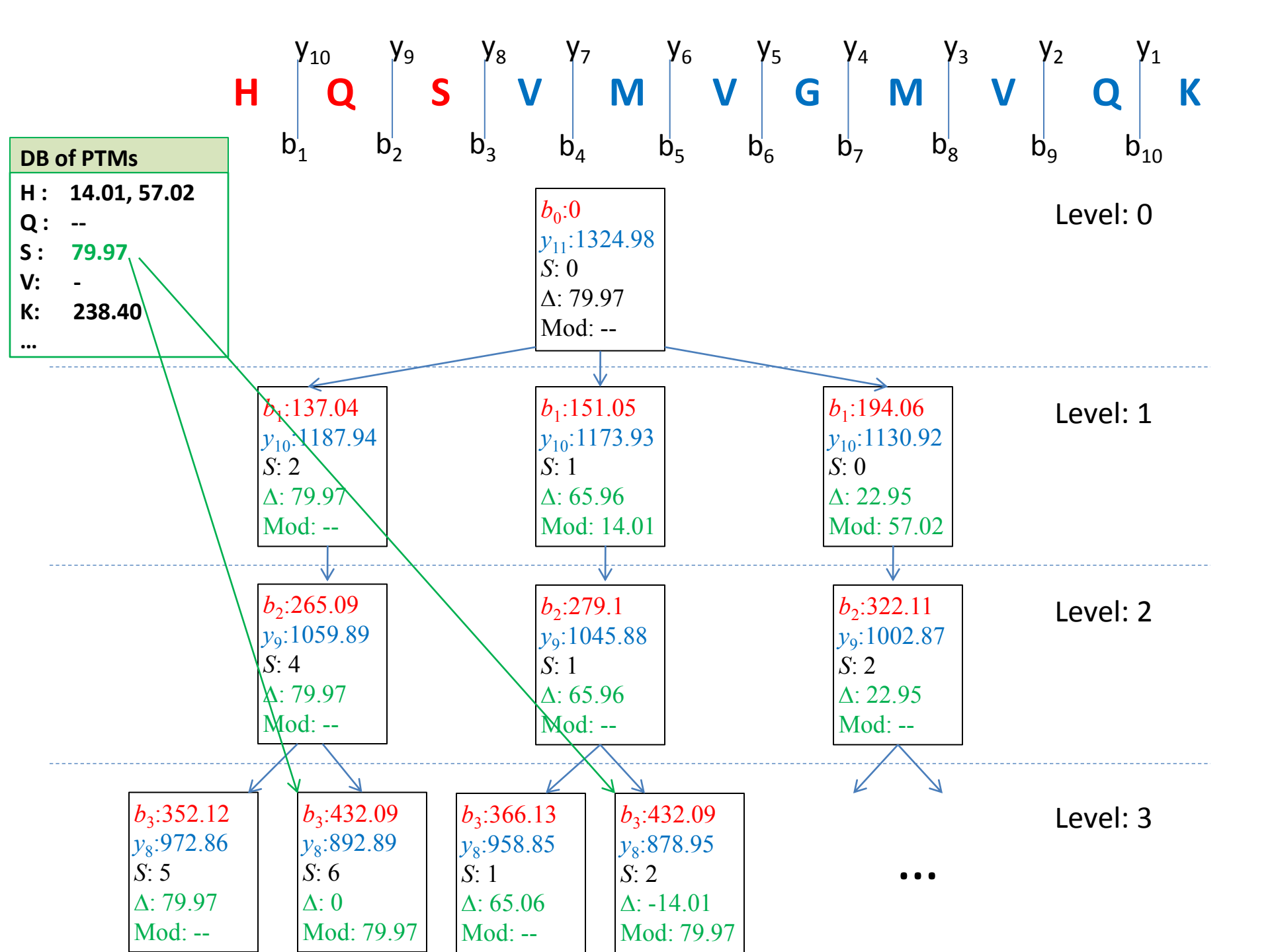
Level: 1



DB of PTMs

H :	**14.01, 57.02**
Q :	**--**
S :	**79.97**
V:	**-**
K:	**238.40**
…	




H Q S V M V G M V Q K
y10 y9 y8 y7 y6 y5 y4 y3 y2 y1
b1 b2 b3 b4 b5 b6 b7 b8 b9 b10
DB of PTMs
H : 14.01, 57.02
Q : --
S : 79.97
V: -
K: 238.40
...
b0:0
y11:1324.98
S: 0
Δ: 79.97
Mod: --
Level: 0
b1:137.04
y10:1187.94
S: 2
Δ: 79.97
Mod: --
b1:151.05
y10:1173.93
S: 1
Δ: 65.96
Mod: 14.01
b1:194.06
y10:1130.92
S: 0
Δ: 22.95
Mod: 57.02
Level: 1
b2:265.09
y9:1059.89
S: 4
Δ: 79.97
Mod: --
b2:279.1
y9:1045.88
S: 1
Δ: 65.96
Mod: --
b2:322.11
y9:1002.87
S: 2
Δ: 22.95
Mod: --
Level: 2
b3:352.12
y8:972.86
S: 5
Δ: 79.97
Mod: --
b3:432.09
y8:892.89
S: 6
Δ: 0
Mod: 79.97
b3:366.13
y8:958.85
S: 1
Δ: 65.06
Mod: --
b3:432.09
y8:878.95
S: 2
Δ: -14.01
Mod: 79.97
...
Level: 3

Properties

- All nodes at level i^{th} correspond to the i^{th} amino acid in the peptide
- The number of the nodes in the computational tree T is:

$$|T| = \sum_{j=1}^{n} \prod_{i=1}^{j} (n_{a_i} + 1)$$

where n_{a_i} is the number of the possible modifications of the amino acid a_i, n is the length of the peptide.

Speedup techniques

- Limit the number of the PTMs K to be included to the peptide at the same time.
- This keeps the size of the search space polynomial wrt the length of the input:

$$|T| = O(n^K)$$

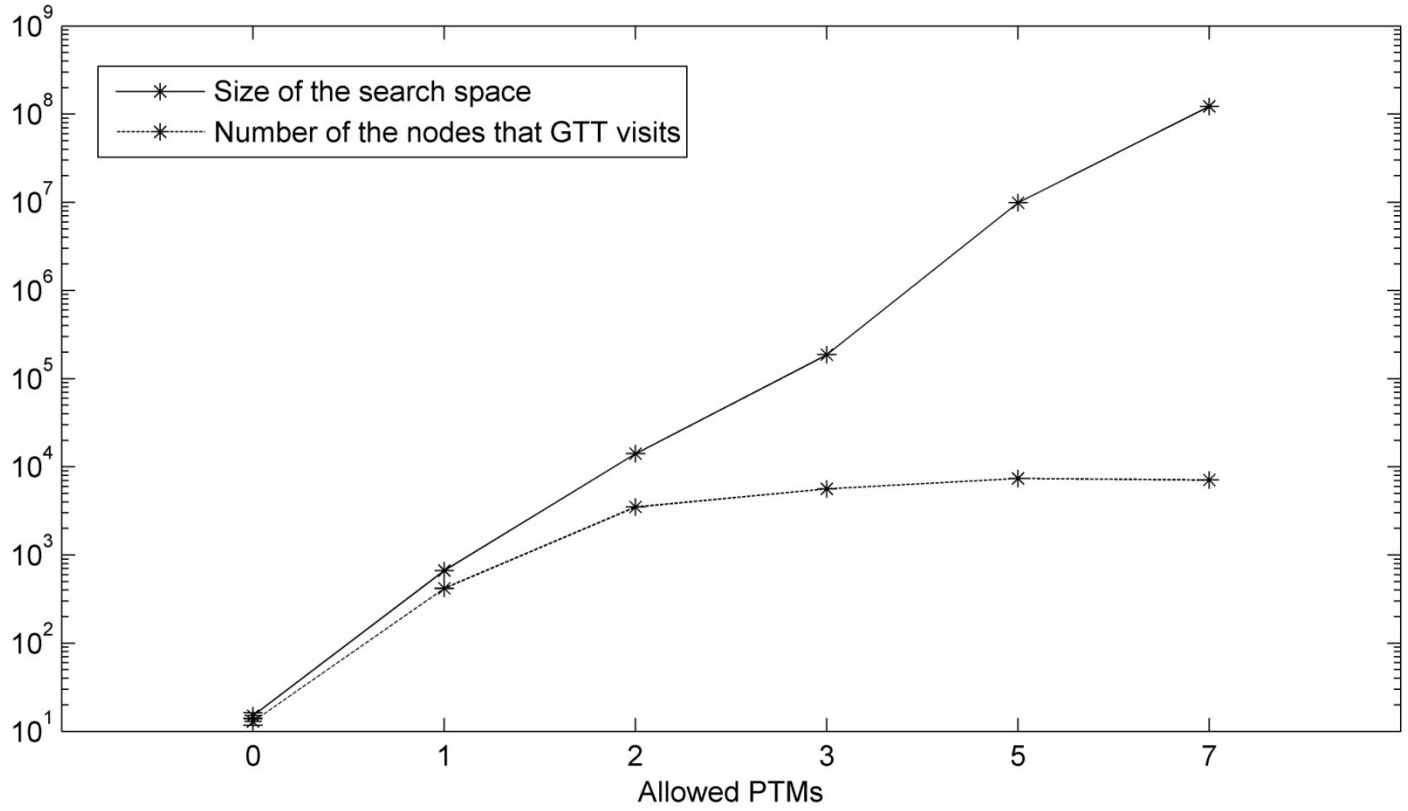
Speedup techniques

- Greedy tree traversal with backtracking.
  - Extend the node with the highest score.
  - queue Q stores the best $\#Q_T$ nodes for backtracking.
    - Q can be a priority queue with bounded size.
  - We can assess the probability $P(\varepsilon)$ of eliminating the correct solution from the search space:

$$P(\varepsilon) = N_L \cdot p\left(K + \sum_{j=1}^{H} p_e^j \right),$$

where N_L is the average PTMs per amino acids, p is the probability of a random match, p_e is the probability of a peak missing, and $H = \left| Q_T /(N_L)^K \right|$

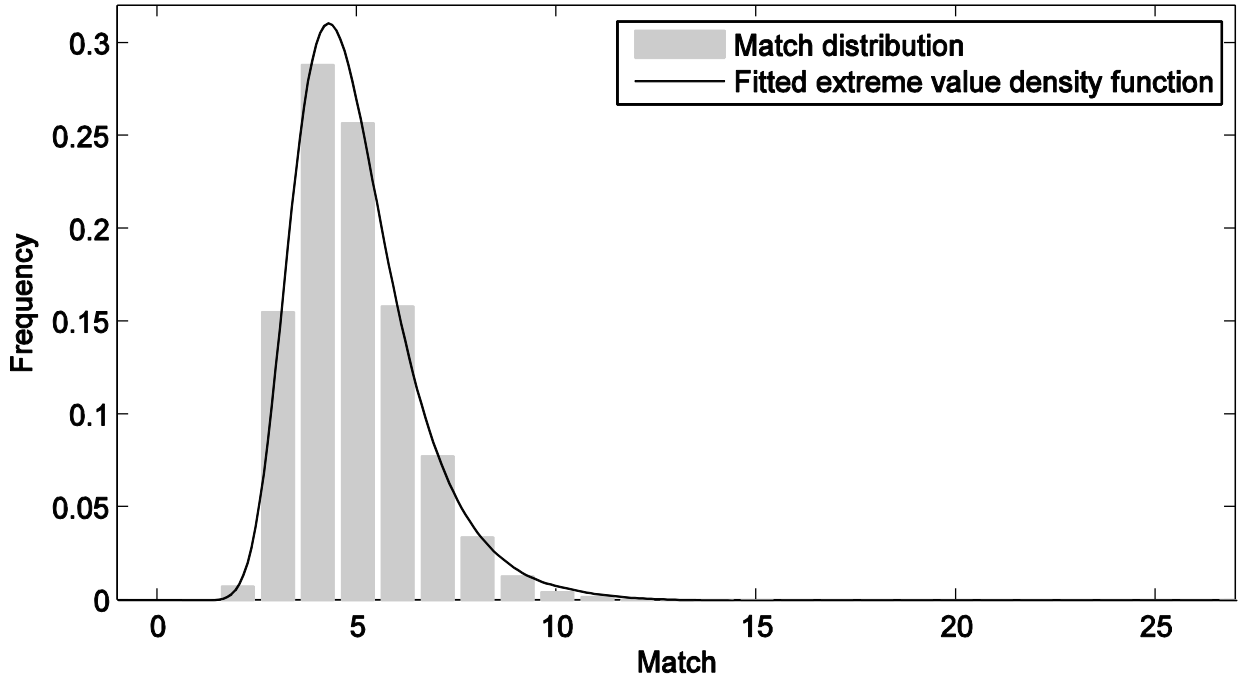
Size of the search space



Size of the search space as a function of the PTM limit

Significance calculation

- The distribution of the random matches follows an Extreme Value Distribution



Experimental tests

- Aurum dataset: 1834 high quality spectra from 246 protein samples

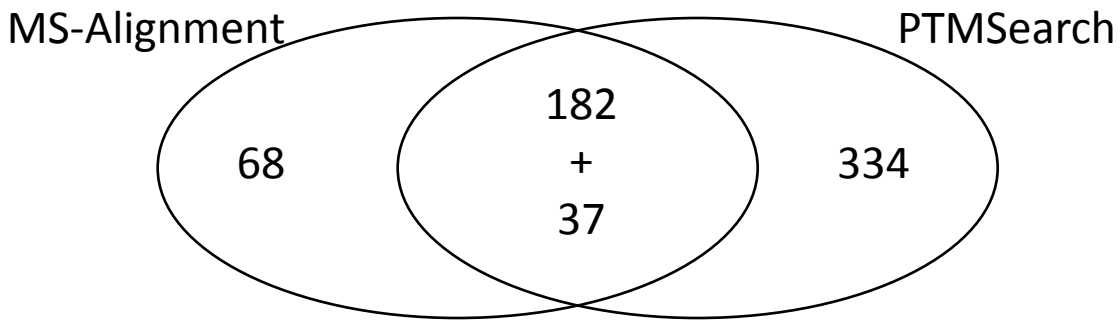
PTM Limit	**0**	**1**	**2**	**3**
0	410	295	243	194
1		258	180	105
2			91	42
3				75
Σ	**410**	**553**	**514**	**416**
Time (min):	11	120	754	1281

- FPR=5%

Experimental tests

- MS-Alignment: standard method for PTM identification

	0 PTM	1 PTM	**Sum**	Time
MS-Alignment	196	91	**287**	5.5 days
PTMSearch	295	258	**553**	2 hours

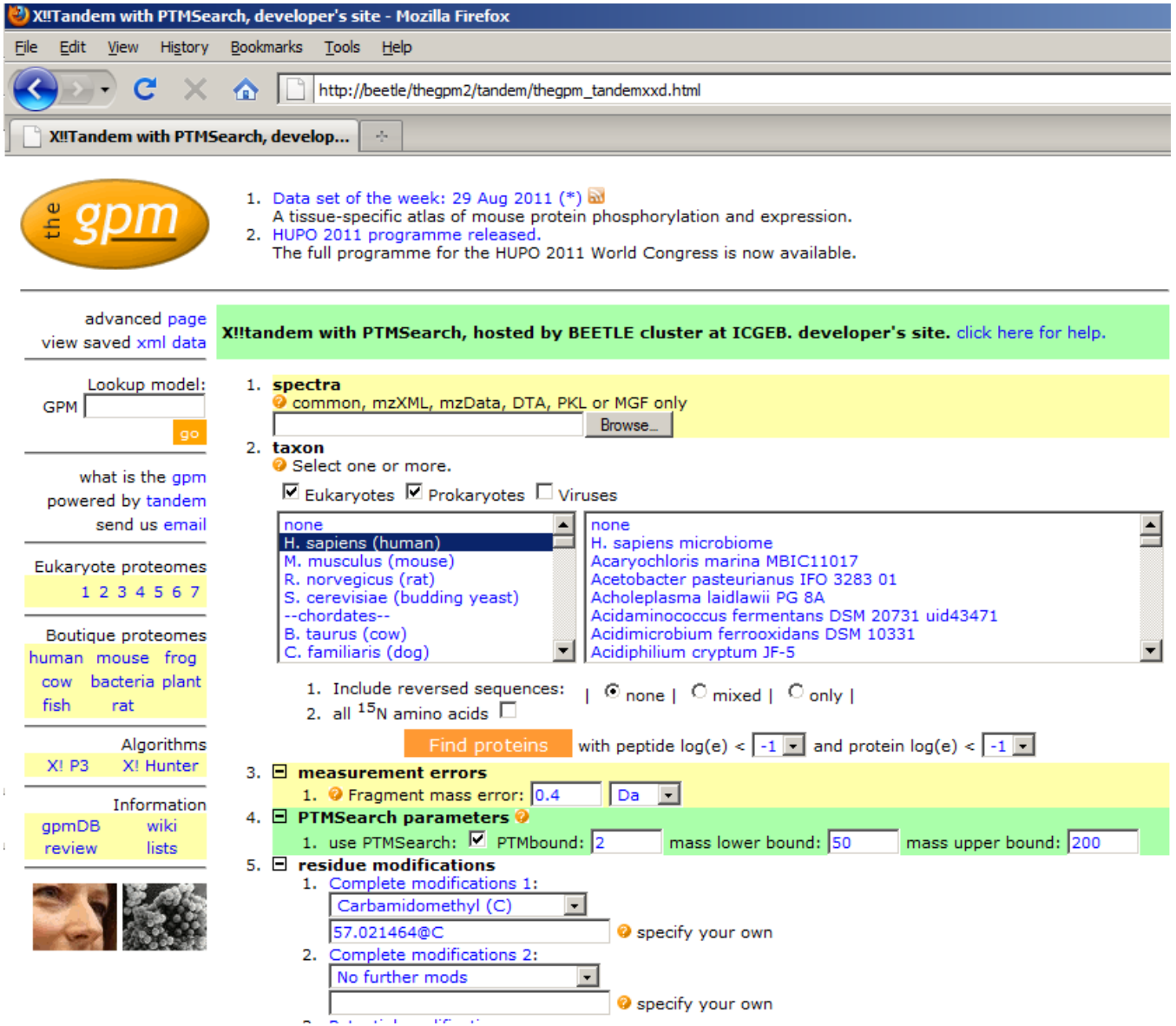


- FPR=5%

Summary and conclusions

- PTM identification computationally extensive
- We have proposed PTMSearch:
  - Novel search space representation for PTM identification
  - Greedy tree traversal method with backtracking
  - Estimation on error of the greedy approach
  - Significance calculation
- PTMSearch identifies more spectra with PTMs at the same FPR.

- PTMSearch with X!tandem software will be provided as free server service.



Thank you for your attention