

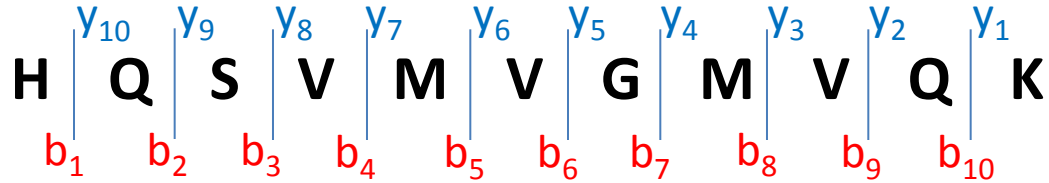
# PTMSearch: a Greedy Tree Traversal Algorithm for Finding Protein Post- Translational Modifications in Tandem Mass Spectra

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

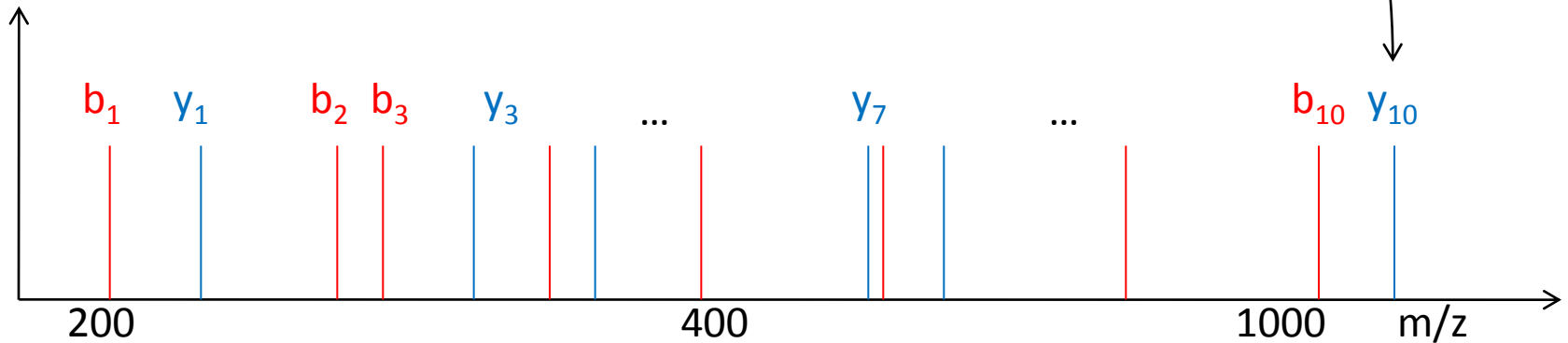
ICGEB, Trieste, Italy



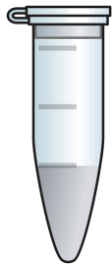
# Mass spectrum



b <sub>1</sub> : H	QSVGMVQK: y <sub>10</sub>
b <sub>2</sub> : HQ	SVMVGMVQK: y <sub>9</sub>
b <sub>3</sub> : HQS	VMVGMVQK: y <sub>8</sub>
b <sub>4</sub> : HQSV	MVG MVQK: y <sub>7</sub>
b <sub>5</sub> : HQSVM	VGMVQK: y <sub>6</sub>
...	...
b <sub>10</sub> : HQSVMVGMVQ	K: y <sub>1</sub>

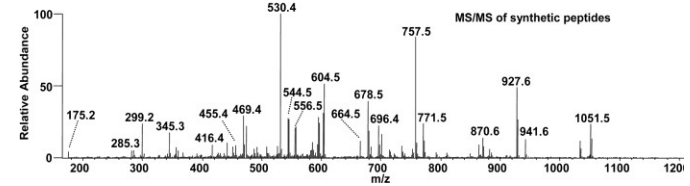
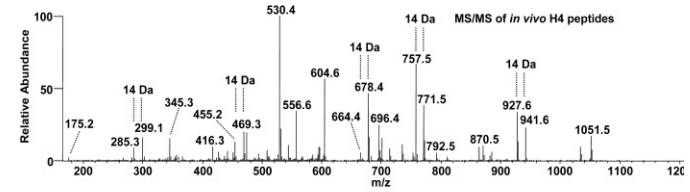
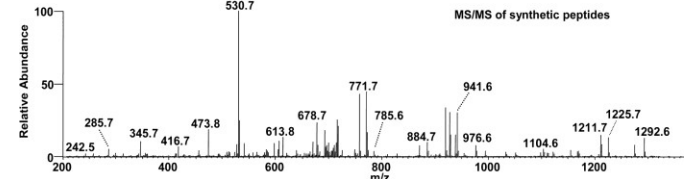
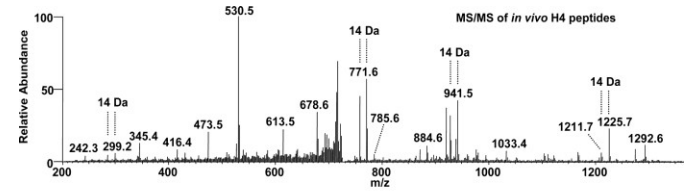


# Mass Spectrometry



Protein mixture

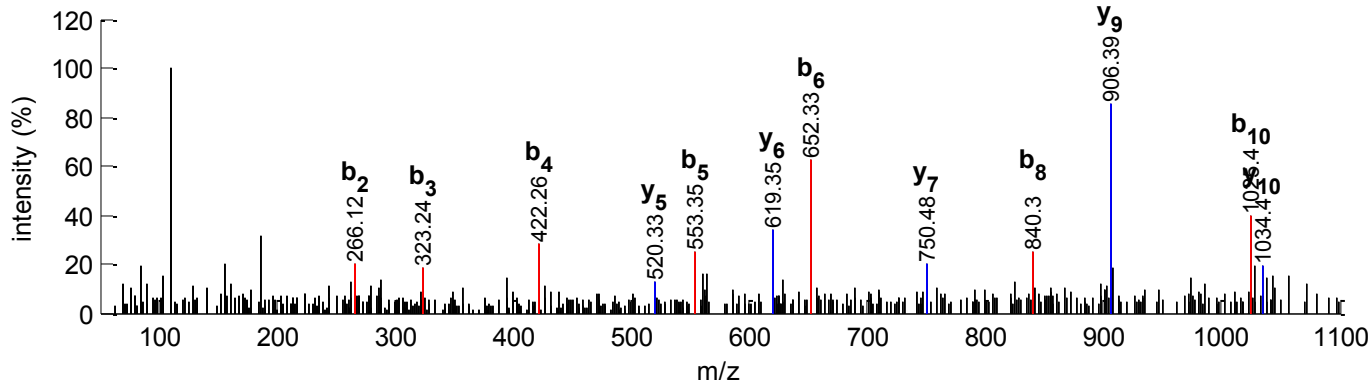
Mass Spectrometer



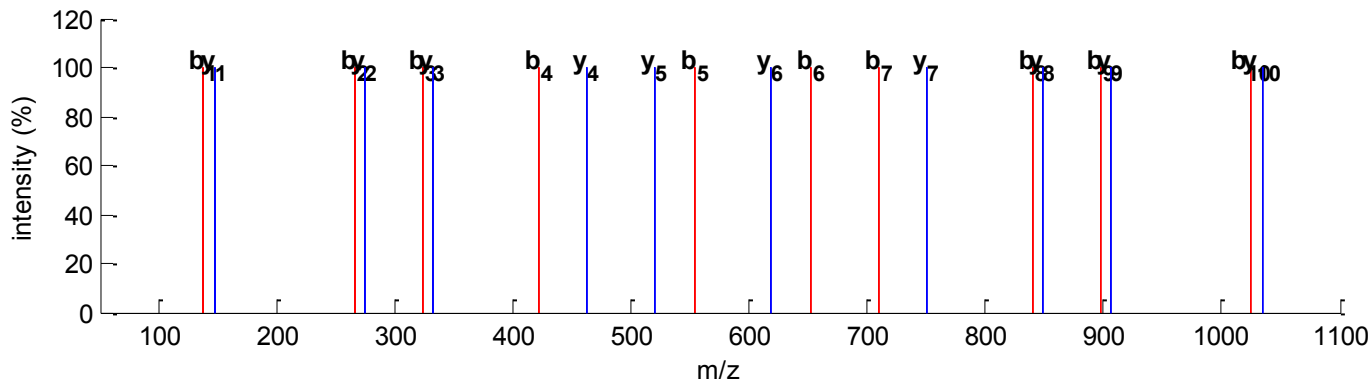
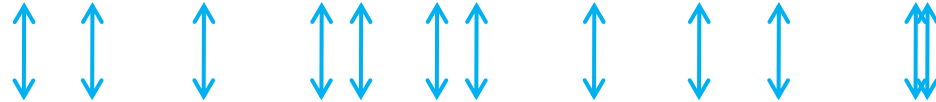
Thousands of spectra to analyze

# Experimental Spectra Annotation

Experimental spectrum:



Shared Peak  
Count: 12



Theoretical spectrum of the peptide: HQSVMVGMVQK

DB of peptides:

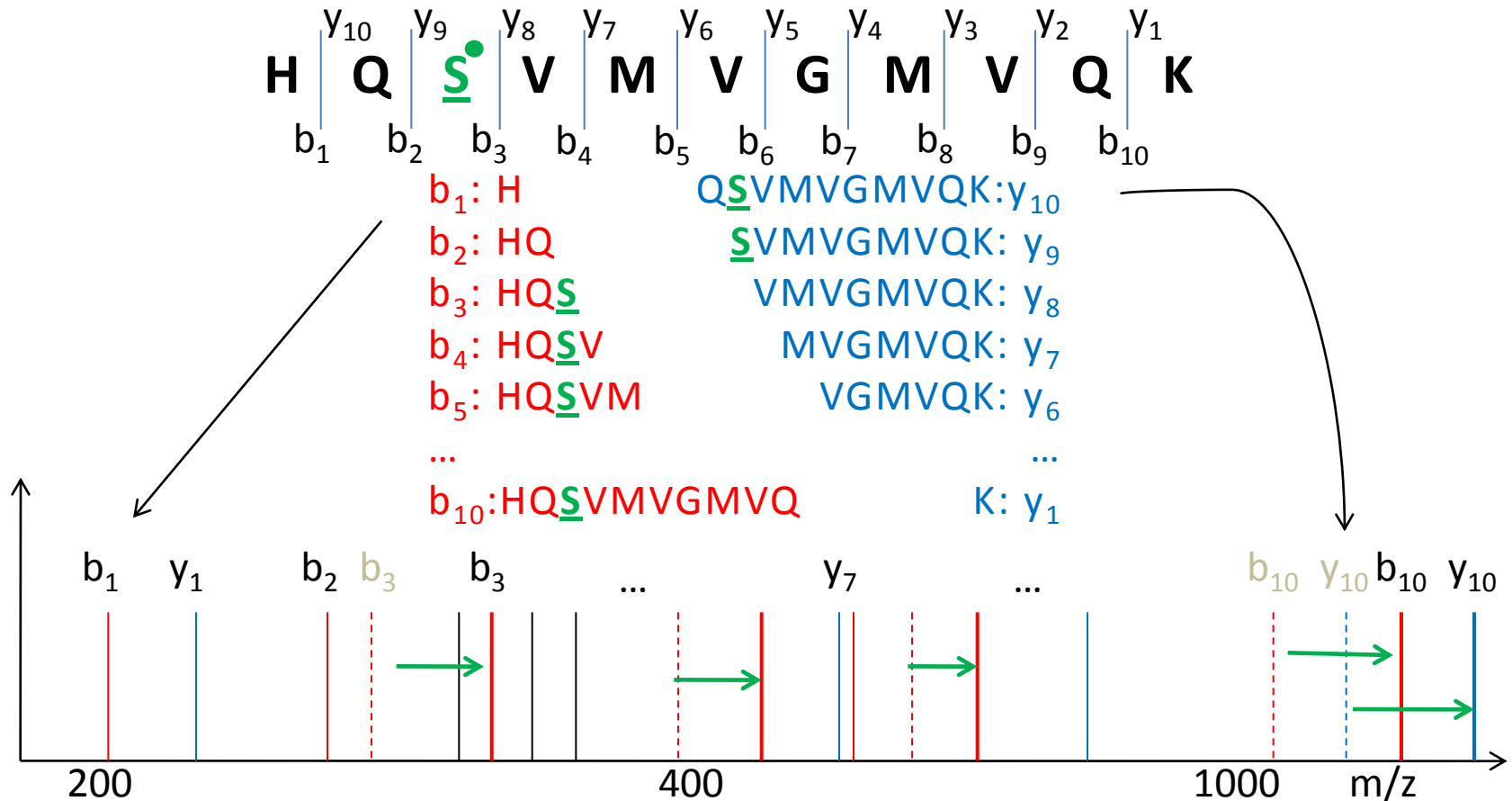
- AQDWSK
- QSDSSWRQK
- QWEEFGEEK
- QWESVRGTK
- ERTGFDWK
- ERTWERTGR
- EWRTGHJJK
- YJNFGTR
- WERFFGDFHR
- ...
- ...
- HQSVMVGMVQK
- ...
- ...

# 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.

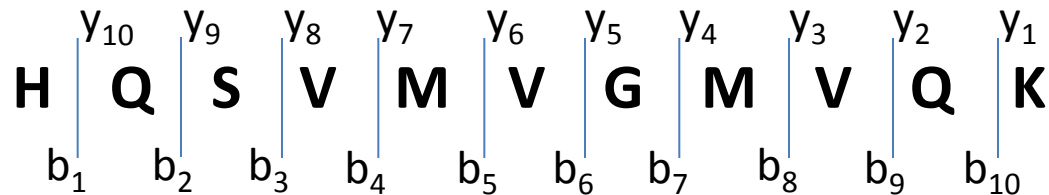
# Our method: PTMSearch

- 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.

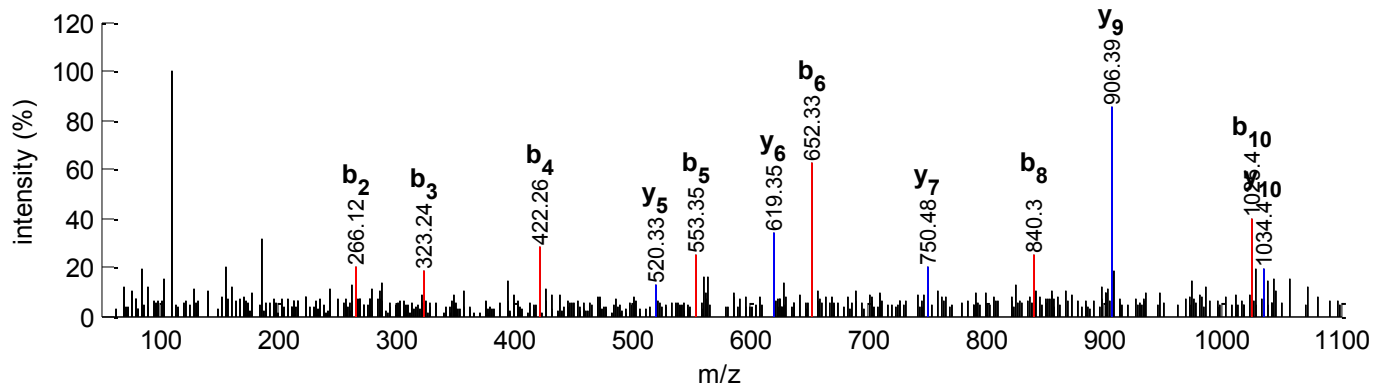


# PTMSearch

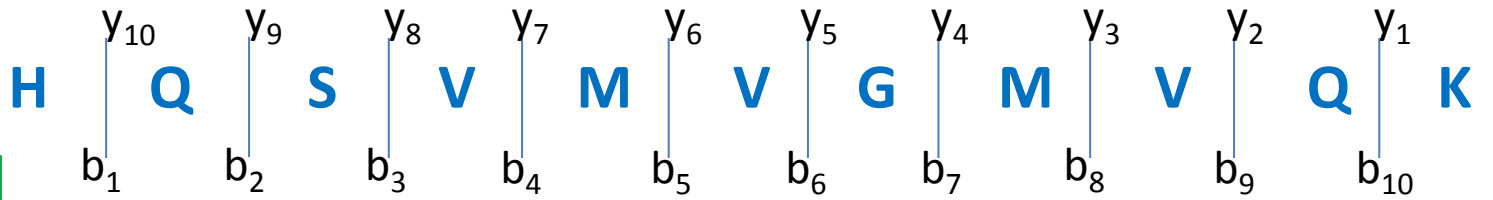
- Theoretical peptide (molecular mass:1245.01Da):



- And an experimental spectrum (mass: 1324.98 Da)



- Mass difference:  $\Delta=79.97$  Dalton

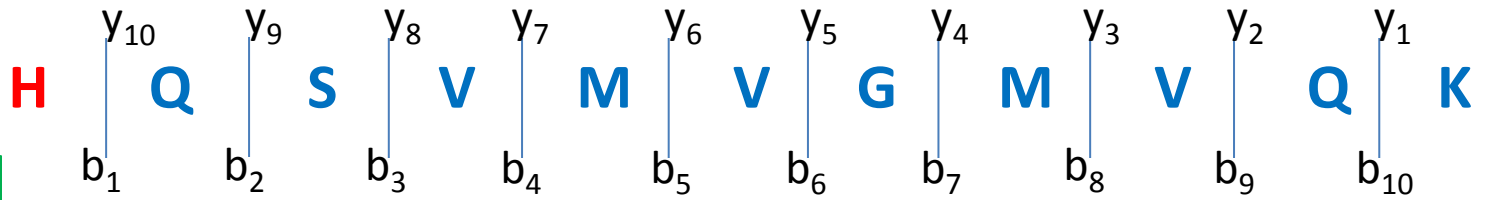


**DB of PTMs**

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

*b*<sub>0</sub>: 0  
*y*<sub>11</sub>: 1324.98  
*S*: 0  
*Δ*: 79.97  
 Mod: --

Level: 0



Level: 0

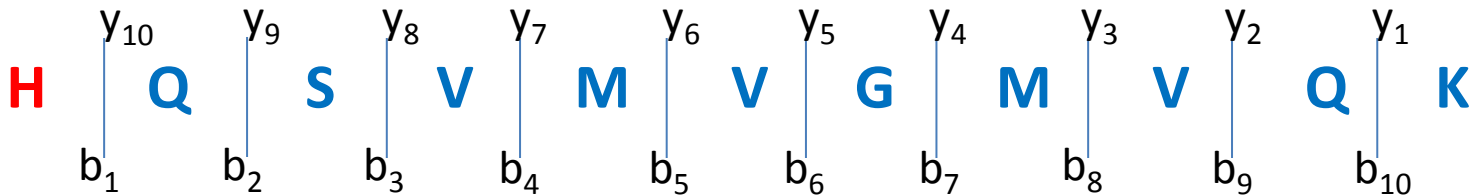
$b_0:0$   
 $y_{11}:1324.98$   
 $S:0$   
 $\Delta:79.97$   
 $Mod:--$

Level: 1

$b_1:137.04$   
 $y_{10}:1187.94$   
 $S:2$   
 $\Delta:79.97$   
 $Mod:--$

**DB of PTMs**

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



Level: 0

Level: 1

**DB of PTMs**

H: 14.01, 57.02

Q: --

S: 79.97

V: -

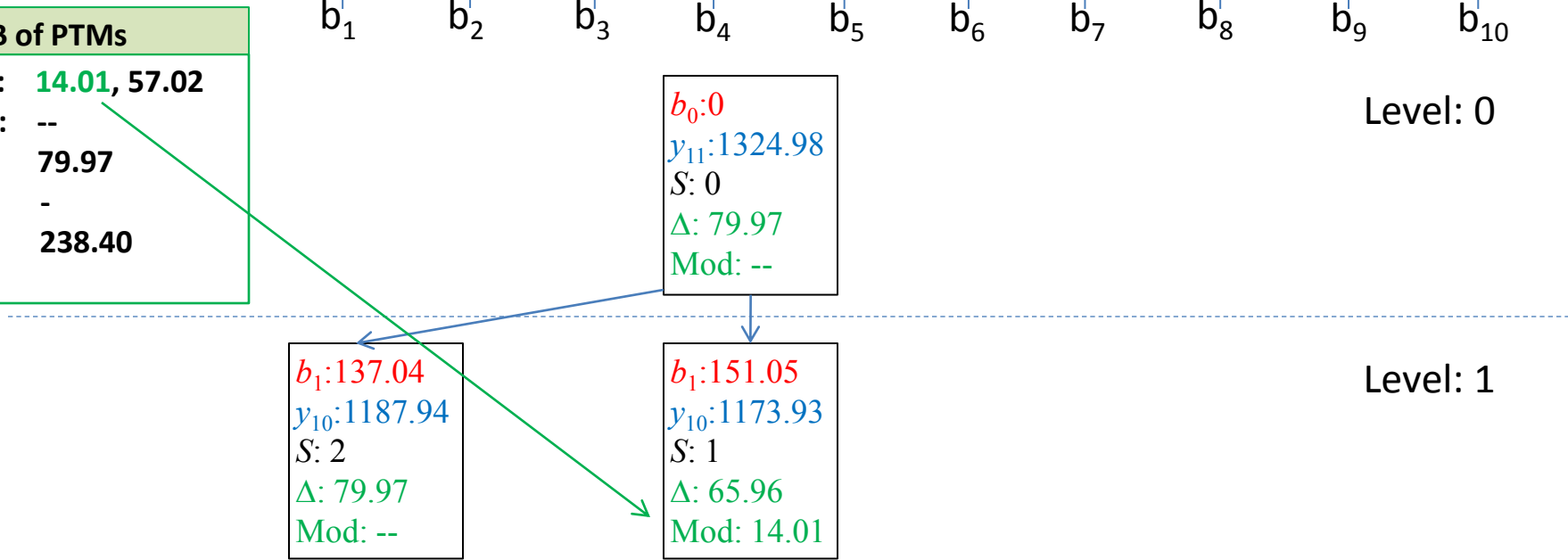
K: 238.40

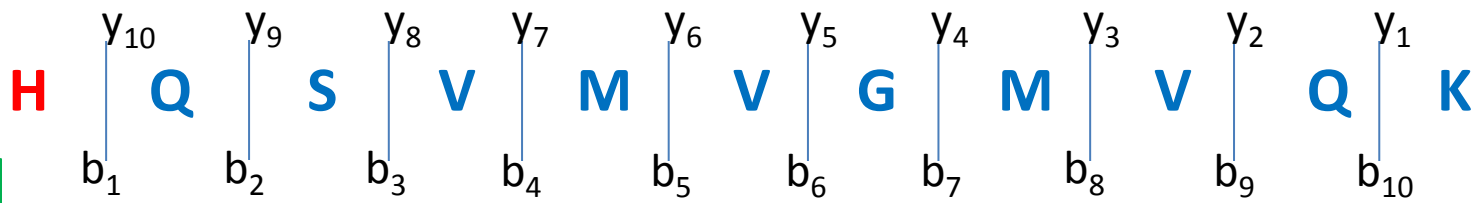
...

$b_0:0$   
 $y_{11}:1324.98$   
 $S:0$   
 $\Delta:79.97$   
 $Mod:--$

$b_1:137.04$   
 $y_{10}:1187.94$   
 $S:2$   
 $\Delta:79.97$   
 $Mod:--$

$b_1:151.05$   
 $y_{10}:1173.93$   
 $S:1$   
 $\Delta:65.96$   
 $Mod:14.01$





Level: 0

Level: 1

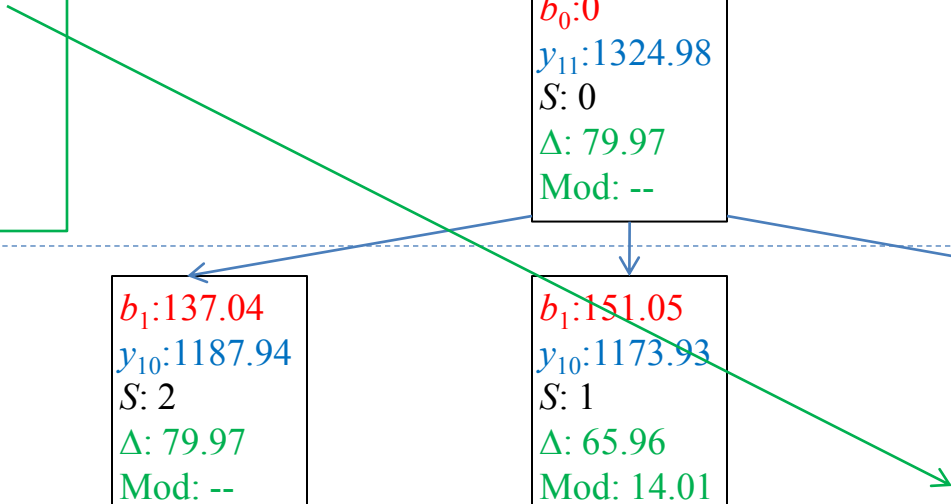
DB of PTMs	
H:	14.01, 57.02
Q:	--
S:	79.97
V:	-
K:	238.40
...	

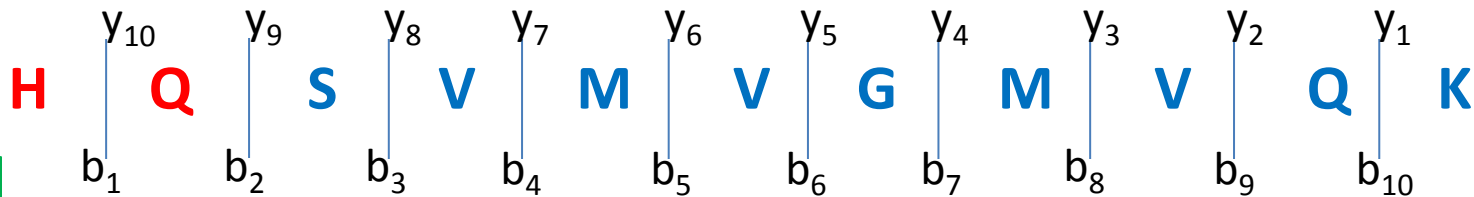
$b_0:0$   
 $y_{11}:1324.98$   
 $S:0$   
 $\Delta:79.97$   
 $Mod:--$

$b_1:137.04$   
 $y_{10}:1187.94$   
 $S:2$   
 $\Delta:79.97$   
 $Mod:--$

$b_1:151.05$   
 $y_{10}:1173.93$   
 $S:1$   
 $\Delta:65.96$   
 $Mod:14.01$

$b_1:194.06$   
 $y_{10}:1130.92$   
 $S:0$   
 $\Delta:22.95$   
 $Mod:57.02$





Level: 0

**DB of PTMs**

H : 14.01, 57.02

Q : --

S : 79.97

V : -

K : 238.40

...

$b_0:0$   
 $y_{11}:1324.98$   
 $S:0$   
 $\Delta:79.97$   
 $Mod:--$

Level: 1

$b_1:137.04$   
 $y_{10}:1187.94$   
 $S:2$   
 $\Delta:79.97$   
 $Mod:--$

$b_1:151.05$   
 $y_{10}:1173.93$   
 $S:1$   
 $\Delta:65.96$   
 $Mod:14.01$

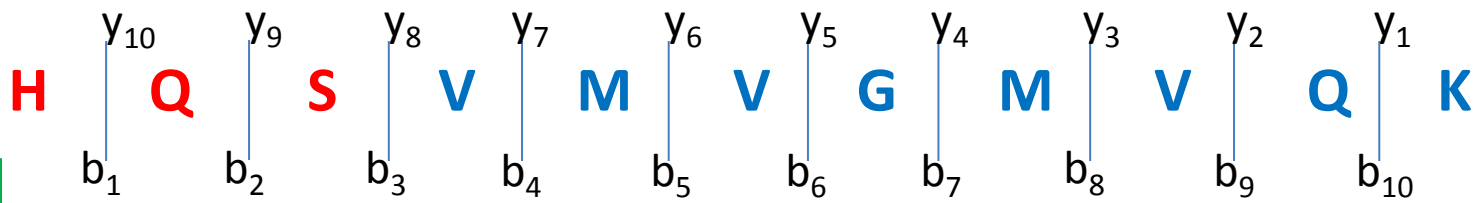
$b_1:194.06$   
 $y_{10}:1130.92$   
 $S:0$   
 $\Delta:22.95$   
 $Mod:57.02$

Level: 2

$b_2:265.09$   
 $y_9:1059.89$   
 $S:4$   
 $\Delta:79.97$   
 $Mod:--$

$b_2:279.1$   
 $y_9:1045.88$   
 $S:1$   
 $\Delta:65.96$   
 $Mod:--$

$b_2:322.11$   
 $y_9:1002.87$   
 $S:2$   
 $\Delta:22.95$   
 $Mod:--$



DB of PTMs	
H :	14.01, 57.02
Q :	--
S :	79.97
V :	-
K :	238.40
...	

$b_0:0$   
 $y_{11}:1324.98$   
 $S:0$   
 $\Delta:79.97$   
 $Mod:--$

Level: 0

$b_1:137.04$   
 $y_{10}:1187.94$   
 $S:2$   
 $\Delta:79.97$   
 $Mod:--$

$b_1:151.05$   
 $y_{10}:1173.93$   
 $S:1$   
 $\Delta:65.96$   
 $Mod:14.01$

$b_1:194.06$   
 $y_{10}:1130.92$   
 $S:0$   
 $\Delta:22.95$   
 $Mod:57.02$

Level: 1

$b_2:265.09$   
 $y_9:1059.89$   
 $S:4$   
 $\Delta:79.97$   
 $Mod:--$

$b_2:279.1$   
 $y_9:1045.88$   
 $S:1$   
 $\Delta:65.96$   
 $Mod:--$

$b_2:322.11$   
 $y_9:1002.87$   
 $S:2$   
 $\Delta:22.95$   
 $Mod:--$

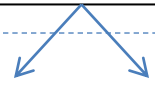
Level: 2

$b_3:352.12$   
 $y_8:972.86$   
 $S:5$   
 $\Delta:79.97$   
 $Mod:--$

$b_3:432.09$   
 $y_8:892.89$   
 $S:6$   
 $\Delta:0$   
 $Mod:79.97$

$b_3:366.13$   
 $y_8:958.85$   
 $S:1$   
 $\Delta:65.06$   
 $Mod:--$

$b_3:432.09$   
 $y_8:878.95$   
 $S:2$   
 $\Delta:-14.01$   
 $Mod:79.97$



...

Level: 3

# Properties

- All nodes at level  $i^{\text{th}}$  correspond to the  $i^{\text{th}}$  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:

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

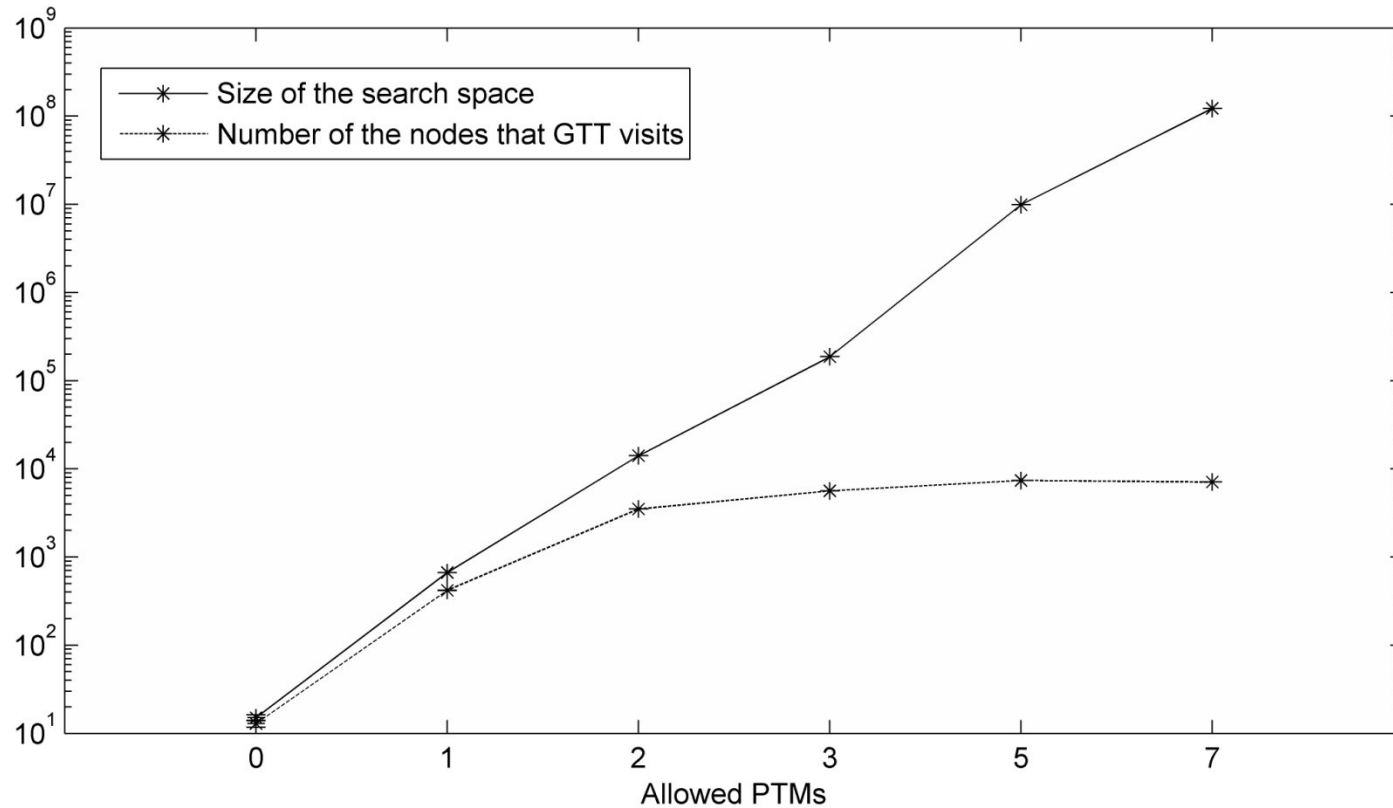
# Speedup techniques

- 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  $H = \left\lfloor \frac{Q_T}{(N_L)^K} \right\rfloor$

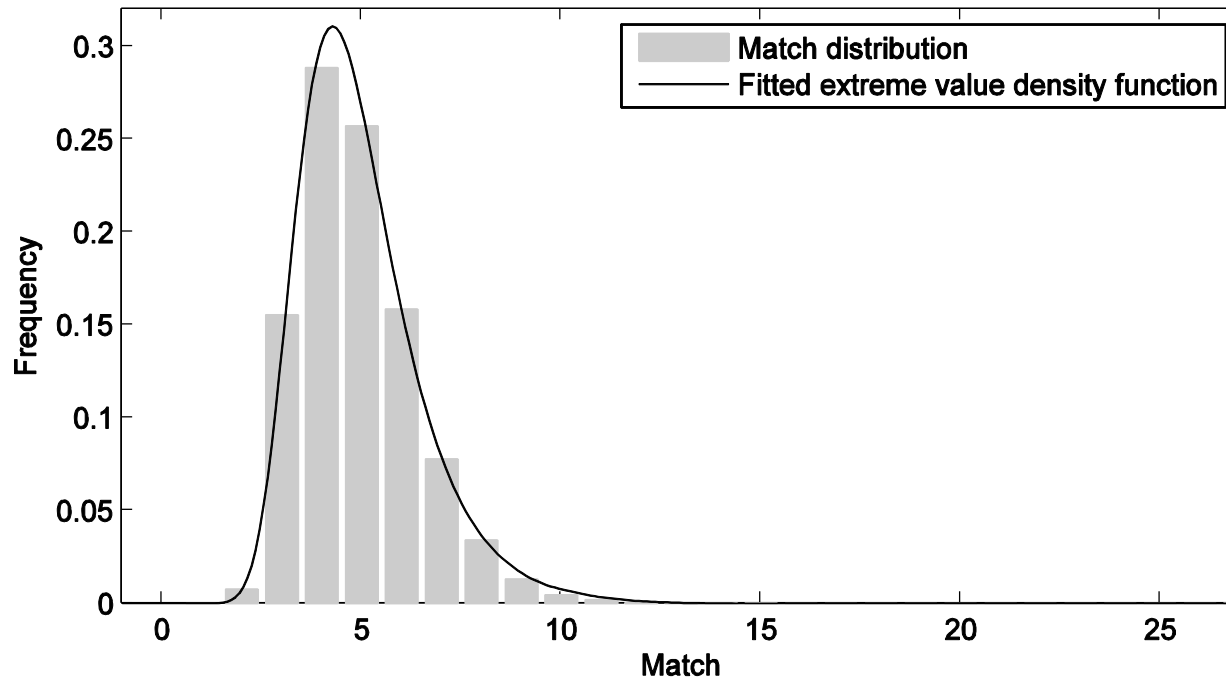
# Size of the search space



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

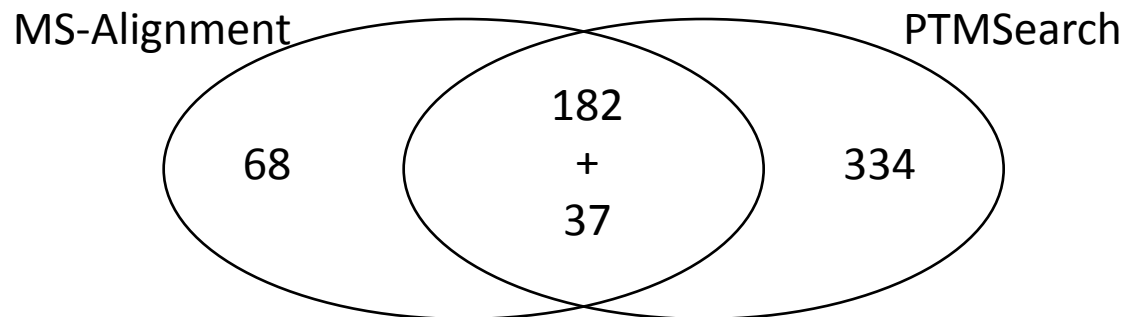
PTM Limit	0	1	2	3
0	410	295	243	194
1		258	180	105
2			91	42
3				75
$\Sigma$	<b>410</b>	<b>553</b>	<b>514</b>	<b>416</b>
Time (min):	11	120	754	1281

- FPR=5%

# Experimental tests

- MS-Alignment: standard method for PTM identification

	0 PTM	1 PTM	Sum	Time
MS-Alignment	196	91	<b>287</b>	5.5 days
PTMSearch	295	258	<b>553</b>	2 hours

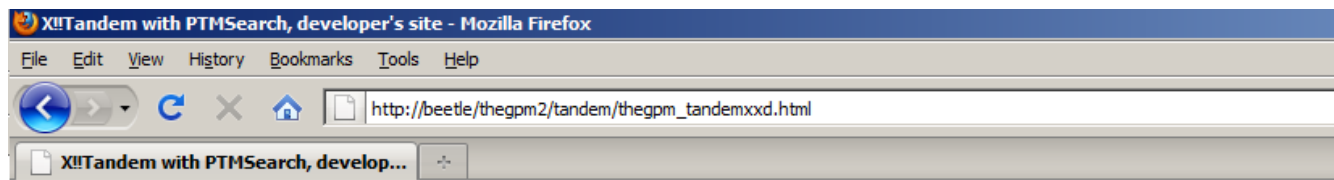


- FPR=5%

# 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.



- Data set of the week: 29 Aug 2011 (\*)**  
A tissue-specific atlas of mouse protein phosphorylation and expression.
- HUPO 2011 programme released.**  
The full programme for the HUPO 2011 World Congress is now available.

advanced [page](#)  
view saved [xml](#) [data](#)

**X!Tandem with PTMSearch, hosted by BEETLE cluster at ICGEB. developer's site. [click here for help.](#)**

Lookup model:  
GPM

- spectra**  
common, mzXML, mzData, DTA, PKL or MGF only

what is the [gpm](#)  
powered by [tandem](#)  
[send us email](#)

- taxon**  
Select one or more.  
 Eukaryotes  Prokaryotes  Viruses

Eukaryote proteomes  
[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#)

none	none
H. sapiens (human)	H. sapiens microbiome
M. musculus (mouse)	Acaryochloris marina MBIC11017
R. norvegicus (rat)	Acetobacter pasteurianus IFO 3283 01
S. cerevisiae (budding yeast)	Acholeplasma laidlawii PG 8A
--chordates--	Acidaminococcus fermentans DSM 20731 uid43471
B. taurus (cow)	Acidimicrobium ferrooxidans DSM 10331
C. familiaris (dog)	Acidiphilium cryptum JF-5

Boutique proteomes  
[human](#) [mouse](#) [frog](#)  
[cow](#) [bacteria](#) [plant](#)  
[fish](#) [rat](#)

1. Include reversed sequences:  none |  mixed |  only |  
2. all <sup>15</sup>N amino acids

with peptide log(e) <  and protein log(e) <

Algorithms  
[X! P3](#) [X! Hunter](#)

- measurement errors**  
1. Fragment mass error:

Information  
[gpmDB](#) [wiki](#)  
[review](#) [lists](#)

- PTMSearch parameters**  
1. use PTMSearch:  PTMbound:  mass lower bound:  mass upper bound:



- residue modifications**  
1. Complete modifications 1:  
  
 specify your own  
2. Complete modifications 2:  
  
specify your own



Thank you for your attention