Towards solving the peptidomics problem with ProSpect

Emile Benoist[1], Géraldine Jean[1], Hélène Rogniaux[2], Guillaume Fertin[1], Dominique Tessier[2]

1.Nantes Université, LS2N (UMR 6004), team ComBi, Nantes, France
2.Inrae, BIA (UR 1268), Nantes, France
{emile.benoist,geraldine.jean,guillaume.fertin}@univ-nantes.fr,
{dominique.tessier,helene.rogniaux}@inrae.fr

Monday November 20th 2023

Intro

Peptidomics:

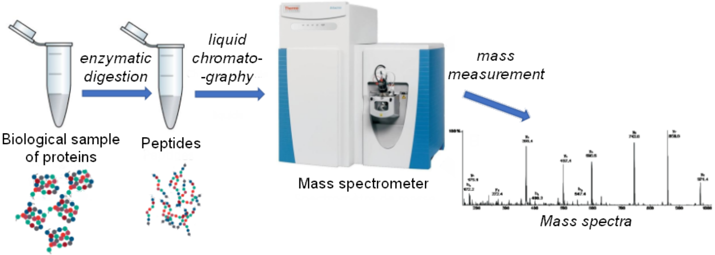
The branch of proteomics dedicated to identifying peptides (i.e., small protein fragments), in a given organism.

(proteomics → the discipline aiming at identifying proteins in a given organism)

protein: long amino acid sequence (more than 50)
MLPPA......LPPQETPK.......KRNIL

peptide: small amino acid sequence (not more than a few tens)
LPPQETPK

How identify the proteins of a sample ? (proteomics)



Biological sample of proteins

enzymatic digestion

Peptides

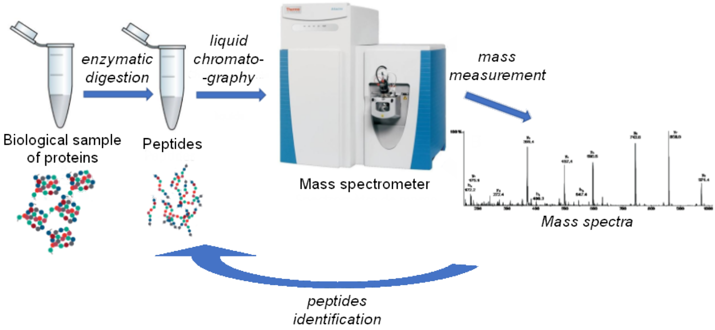
liquid chromato- -graphy

Mass spectrometer

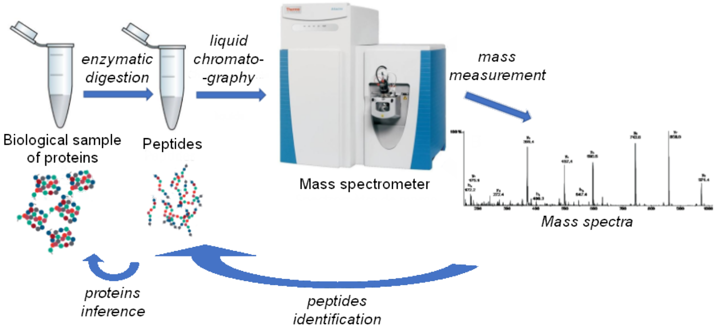
mass measurement

Mass spectra

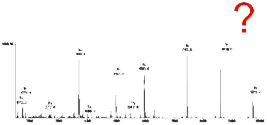
How identify the proteins of a sample ? (proteomics)

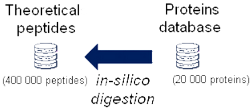


How identify the proteins of a sample ? (proteomics)



Identify a spectrum using a proteins database (proteomics)




Theoretical peptides

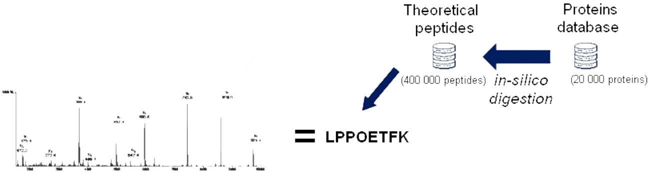
(400 000 peptides)

in-silico digestion

Proteins database

(20 000 proteins)

Identify a spectrum using a proteins database (proteomics)



Theoretical peptides

(400 000 peptides)

Proteins database

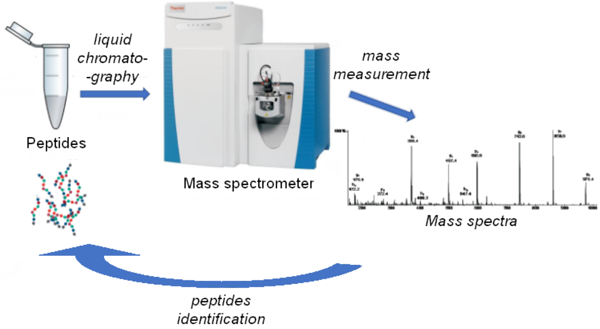
(20 000 proteins)

in-silico digestion

= **LPPOETFK**

PSM (Peptide-Spectrum-Match)

How identify the peptides of a sample ? (peptidomics)



liquid chromato--graphy

Peptides

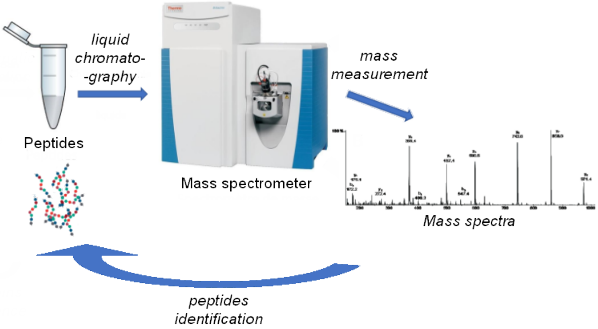
Mass spectrometer

mass measurement

Mass spectra

peptides identification

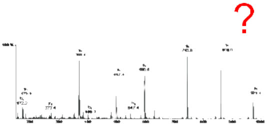
How identify the peptides of a sample ? (peptidomics)

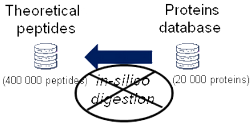


→ Our objective: implement an algorithm (ProSpect) for the identification of peptides in peptidomics

Encountered difficulties

→ First problem: *in-silico* digestion is infeasible





Theoretical peptides

Proteins database

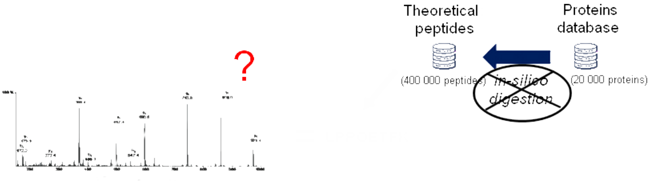
(400 000 peptides)

~~*in-silico* digestion~~

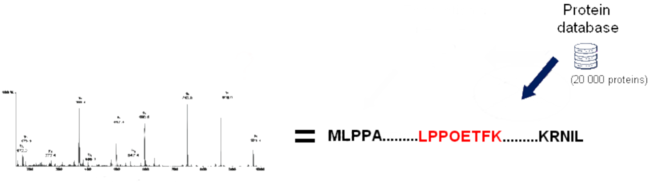
(20 000 proteins)

Encountered difficulties

→ First problem: *in-silico* digestion is infeasible



→ Our solution: compare the spectra directly with the proteins



Encountered difficulties

→ Second problem: peptides in a biological sample of peptidomics are more likely to carry modifications (alteration of their structure or sequence)

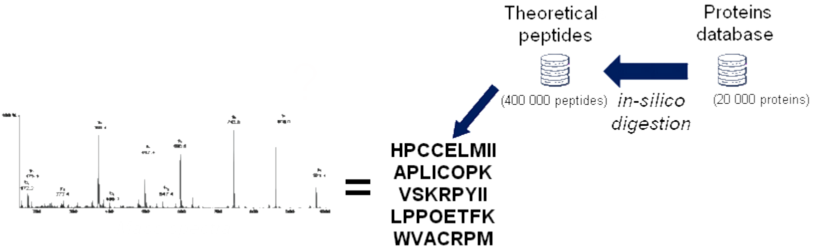
Encountered difficulties

→ Second problem: peptides in a biological sample of peptidomics are more likely to carry modifications (alteration of their structure or sequence)

→ Our solution: be based on SpecGlobX *, an algorithm used in proteomics for a better identification of modified peptides

* G. Prunier, M. Cherkaoui, A. Lysiak, O. Langella, M. Blein-Nicolas, V. Lollier, E. Benoist, G. Jean, G. Fertin, H. Rogniaux, and D. Tessier, "Fast alignment of mass spectra in large proteomics datasets, capturing dissimilarities arising from multiple complex modifications of peptides," bioRxiv, 2023. [Online]. Available: https://www.biorxiv.org/content/earl/2023/03/12/2023.03.09.531667

How SpecGlobX can be used ?



Theoretical peptides

Proteins database

(400 000 peptides)

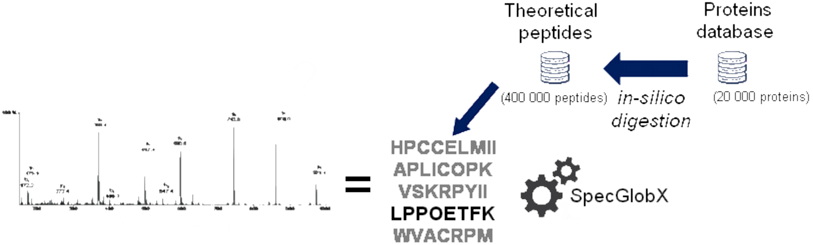
in-silico digestion

(20 000 proteins)

=

HPCCELMII
APLICOPK
VSKRPYII
LPPOETFK
WVACRPM

How SpecGlobX can be used ?



Theoretical peptides

(400 000 peptides)

in-silico digestion

Proteins database

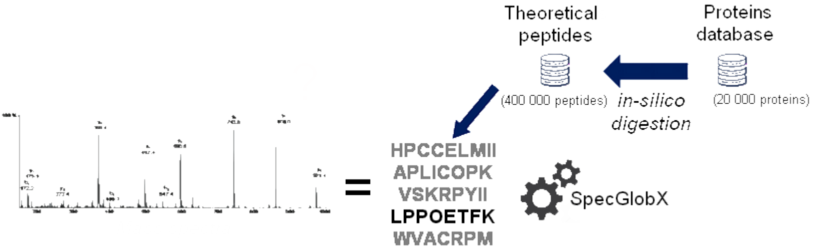
(20 000 proteins)

=

HPCCELMII
APLICOPK
VSKRPYII
LPPOETFK
WVACRPM

SpecGlobX

How SpecGlobX can be used ?



Theoretical peptides

Proteins database

(400 000 peptides)

in-silico digestion

(20 000 proteins)

=

HPCCELMII

APLICOPK

VSKRPYII

LPPOETFK

WVACRPM

SpecGlobX

× SpecGlobX directly apply for each couple spectrum-peptide would give too high execution times

A problem of scale

→ Without any improvement:

- normal size of a peptidomics dataset: at least 1 billion of PSMs (50 000 spectra, 20 000 proteins)
- SpecGlobX processes 1 million of PSMs in 10 minutes (1 thread)
- 1 protein ≈ 40 peptides

↪ more than 6000 hours for a classic peptidomics dataset

The work carried out

→ Implement ProSpect based on SpecGlobX

- adapt SpecGlobX to the new context of peptidomics
- significantly improve SpecGlobX performance

How SpecGlobX performs a spectrum-peptide alignment ?

→ A dynamic algorithm that fill a score matrix

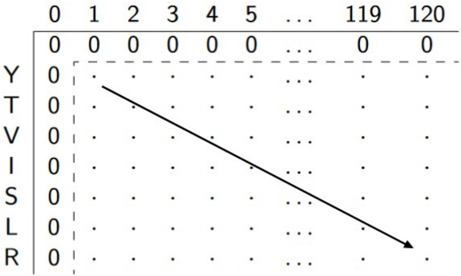
$$\begin{array}{c|ccccccccc}
 & 0 & 1 & 2 & 3 & 4 & 5 & \dots & 119 & 120 \\
\hline
 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\
Y & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\
T & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\
V & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\
I & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\
S & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\
L & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\
R & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot
\end{array}$$

- each column correspond to one peak of the spectrum (120 in average)
- each row correspond to an amino acid of the peptide (between 7 and 25)

How SpecGlobX performs a spectrum-peptide alignment ?

→ How the matrix is filled ?

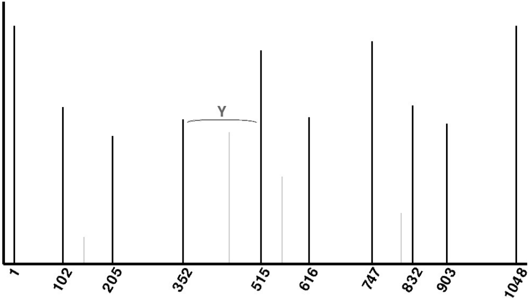
- the matrix is filled by score from left to right and from top to bottom
- the score of each cell is based on the score of a cell on its left and in the previous row.



	0	1	2	3	4	5	...	119	120
	0	0	0	0	0	0	...	0	0
Y	0	.	.	.	.	.	...	.	.
T	0	.	.	.	.	.	...	.	.
V	0	.	.	.	.	.	...	.	.
I	0	.	.	.	.	.	...	.	.
S	0	.	.	.	.	.	...	.	.
L	0	.	.	.	.	.	...	.	.
R	0	.	.	.	.	.	...	.	.

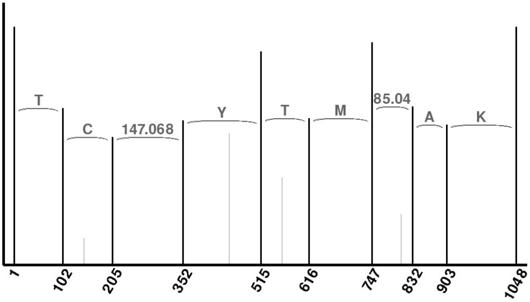
Amino acids in a spectrum

→ a mass delta between 2 peaks and equal to the mass of an amino acid indicates the presence of this amino acid in the peptide



How to align peaks and amino acids ?

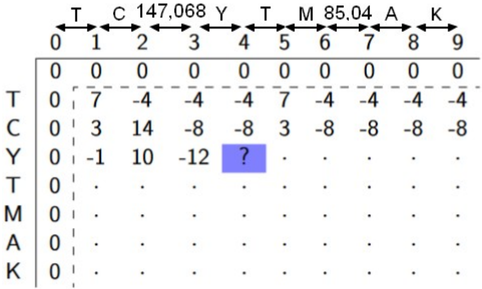
→ this spectrum can be interpret as the peptide TC[147,068]YHM[85,04]AK (with 2 unknown masses)



How SpecGlobX performs a spectrum-peptide alignment ?

→ There is 3 possibilities to fill a cell :

- case *found*
- case *found with shift*
- case *not found*



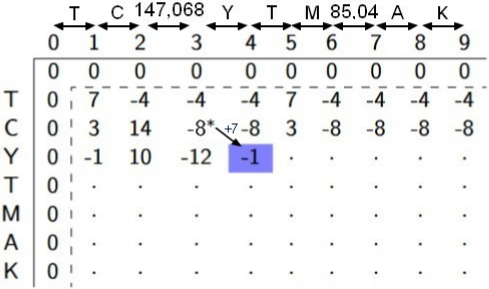
T C 147,068 Y T M 85,04 A K

	0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0	0
T	0	7	-4	-4	-4	7	-4	-4	-4	-4
C	0	3	14	-8	-8	3	-8	-8	-8	-8
Y	0	-1	10	-12	?	.	.	.	.	.
T	0	.	.	.	.	.	.	.	.	.
M	0	.	.	.	.	.	.	.	.	.
A	0	.	.	.	.	.	.	.	.	.
K	0	.	.	.	.	.	.	.	.	.

How SpecGlobX performs a spectrum-peptide alignment ?

→ There is 3 possibilities to fill a cell :

- case *found*
- case *found with shift*
- case *not found*


T
C
147,068
Y
T
M
85.04
A
K
+7

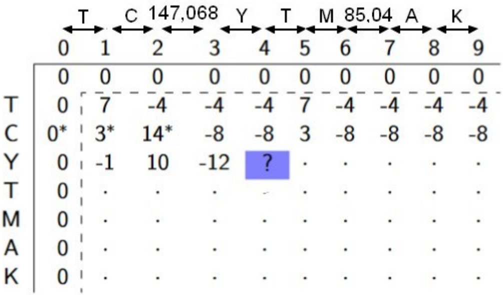
	0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0	0
T	0	7	-4	-4	-4	7	-4	-4	-4	-4
C	0	3	14	-8*	-8	3	-8	-8	-8	-8
Y	0	-1	10	-12	-1	.	.	.	.	.
T	0	.	.	.	.	.	.	.	.	.
M	0	.	.	.	.	.	.	.	.	.
A	0	.	.	.	.	.	.	.	.	.
K	0	.	.	.	.	.	.	.	.	.

→ a bonus of 7 is added from the score

How SpecGlobX performs a spectrum-peptide alignment ?

→ There is 3 possibilities to fill a cell :

- case *found*
- case *found with shift*
- case *not found*

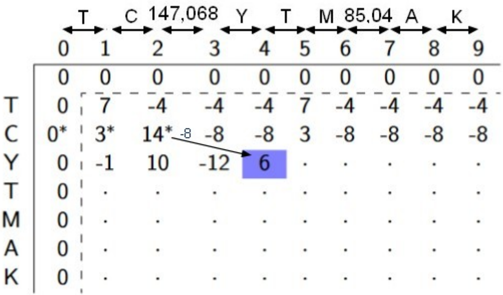

T
C
147,068
Y
T
M
85,04
A
K

	0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0	0
T	0	7	-4	-4	-4	7	-4	-4	-4	-4
C	0*	3*	14*	-8	-8	3	-8	-8	-8	-8
Y	0	-1	10	-12	?	.	.	.	.	.
T	0	.	.	.	.	.	.	.	.	.
M	0	.	.	.	.	.	.	.	.	.
A	0	.	.	.	.	.	.	.	.	.
K	0	.	.	.	.	.	.	.	.	.

How SpecGlobX performs a spectrum-peptide alignment ?

→ There is 3 possibilities to fill a cell :

- case *found*
- case *found with shift*
- case *not found*



T C 147,068 Y T M 85,04 A K

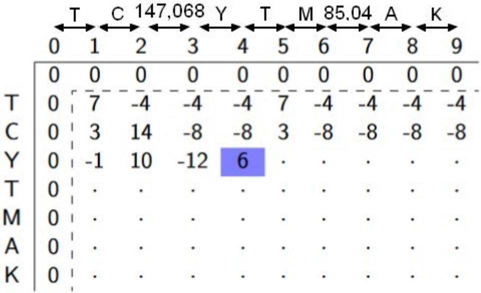
	0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0	0
T	0	7	-4	-4	-4	7	-4	-4	-4	-4
C	0*	3*	14*	-8	-8	3	-8	-8	-8	-8
Y	0	-1	10	-12	6	.	.	.	.	.
T	0	.	.	.	.	.	.	.	.	.
M	0	.	.	.	.	.	.	.	.	.
A	0	.	.	.	.	.	.	.	.	.
K	0	.	.	.	.	.	.	.	.	.

→ a penalty of 8 is subtracted from the score

How SpecGlobX performs a spectrum-peptide alignment ?

→ There is 3 possibilities to fill a cell :

- case *found*
- case *found with shift*
- case *not found*



T C 147,068 Y T M 85,04 A K

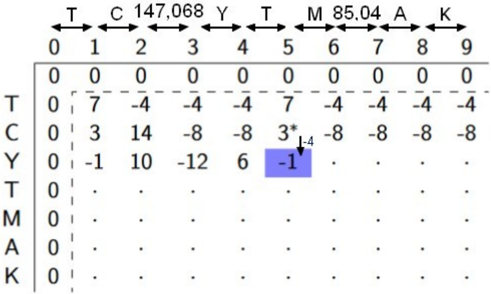
	0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0	0
T	0	7	-4	-4	-4	7	-4	-4	-4	-4
C	0	3	14	-8	-8	3	-8	-8	-8	-8
Y	0	-1	10	-12	6	.	.	.	.	.
T	0	.	.	.	.	.	.	.	.	.
M	0	.	.	.	.	.	.	.	.	.
A	0	.	.	.	.	.	.	.	.	.
K	0	.	.	.	.	.	.	.	.	.

→ the choice which generate the best score for the cell is selected

How SpecGlobX performs a spectrum-peptide alignment ?

→ There is 3 possibilities to fill a cell :

- case *found*
- case *found with shift*
- case *not found*

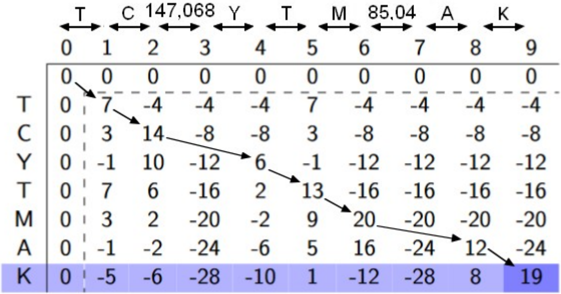

T
C
147,068
Y
T
M
85,04
A
K
-4

	0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0	0
T	0	7	-4	-4	-4	7	-4	-4	-4	-4
C	0	3	14	-8	-8	3*	-8	-8	-8	-8
Y	0	-1	10	-12	6	-1	·	·	·	·
T	0	·	·	·	·	·	·	·	·	·
M	0	·	·	·	·	·	·	·	·	·
A	0	·	·	·	·	·	·	·	·	·
K	0	·	·	·	·	·	·	·	·	·

→ a penalty of 4 is subtracted from the score

How SpecGlobX performs a spectrum-peptide alignment ?

→ when the matrix is completely filled, the cell of the last row with the best score is considered as the best alignment



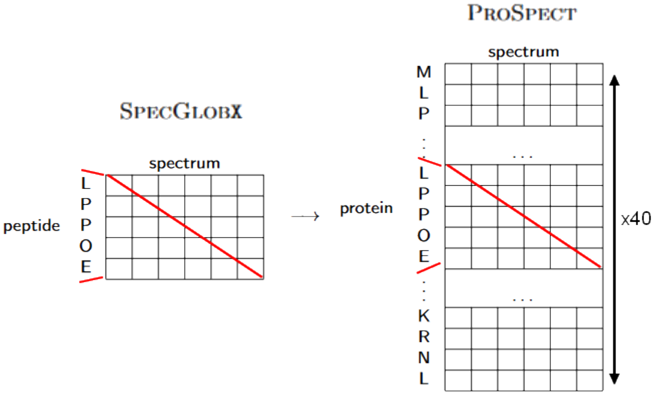
T
C
147,068
Y
T
M
85,04
A
K

	0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0	0
T	0	7	-4	-4	-4	7	-4	-4	-4	-4
C	0	3	14	-8	-8	3	-8	-8	-8	-8
Y	0	-1	10	-12	6	-1	-12	-12	-12	-12
T	0	7	6	-16	2	13	-16	-16	-16	-16
M	0	3	2	-20	-2	9	20	-20	-20	-20
A	0	-1	-2	-24	-6	5	16	-24	12	-24
K	0	-5	-6	-28	-10	1	-12	-28	8	19

→ the spectrum is interpreted as the peptide TC[147,068]YTM[85,04]AK

How to adapt SpecGlobX to peptidomics ?

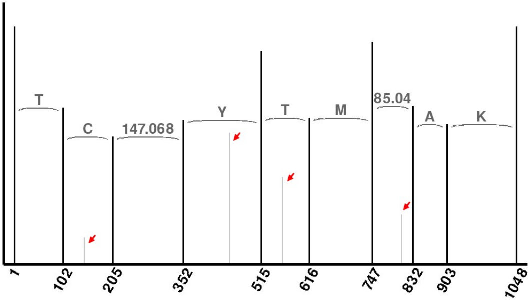
→ Move from global to semi-global alignment:



→ not only the best score of the last row is considered, but any score at any position can be

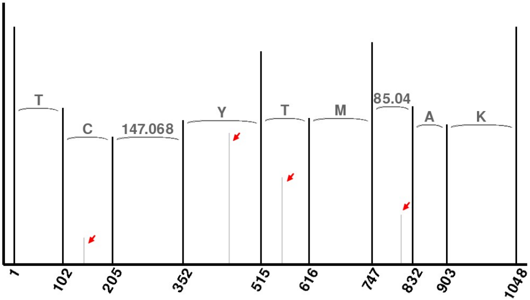
An example of improvement

→ a large proportion of peaks in a spectrum are useless, there is no need to generate a column for them



An example of improvement

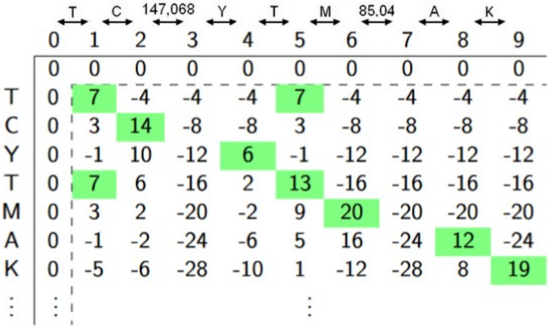
→ a large proportion of peaks in a spectrum are useless, there is no need to generate a column for them



→ reduction of the number of columns in the matrix: by 120 columns to 50 columns on average

Another example of improvement

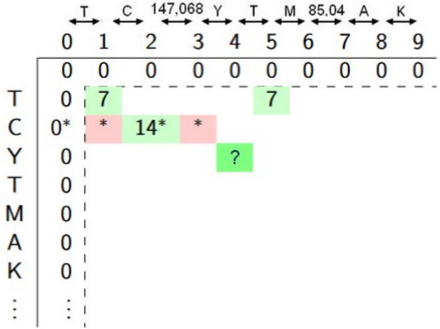
→ we only compute the score of the *cells of interest* (the cells where the amino acid is found in the spectrum)



		T	C	147,068	Y	T	M	85.04	A	K
	0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0	0
T	0	7	-4	-4	-4	7	-4	-4	-4	-4
C	0	3	14	-8	-8	3	-8	-8	-8	-8
Y	0	-1	10	-12	6	-1	-12	-12	-12	-12
T	0	7	6	-16	2	13	-16	-16	-16	-16
M	0	3	2	-20	-2	9	20	-20	-20	-20
A	0	-1	-2	-24	-6	5	16	-24	12	-24
K	0	-5	-6	-28	-10	1	-12	-28	8	19
⋮	⋮					⋮				

Another example of improvement

→ we only compute the score of the *cells of interest* (the cells where the amino acid is found in the spectrum)

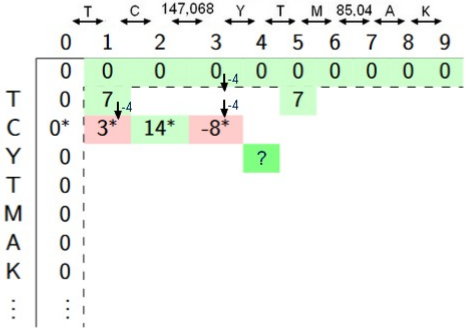


	0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0	0
T	0	7				7				
C	0*	*	14*	*						
Y	0				?					
T	0									
M	0									
A	0									
K	0									
⋮	⋮									

Problem: the score of a cell of interest is calculated from the score of cells which are not always of interest

Another example of improvement

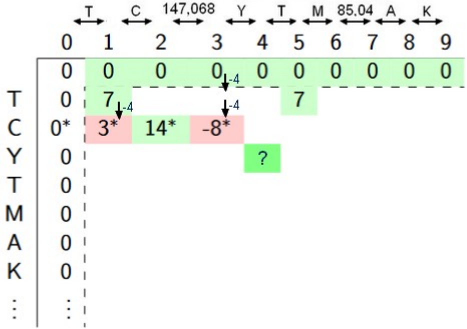
→ the score of any cell can be computed relying on the score and the row number of the cell of interest above it.



	0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0	0
T	0	7				7				
C	0*	3*	14*	-8*						
Y	0				?					
T	0									
M	0									
A	0									
K	0									
⋮	⋮									

Another example of improvement

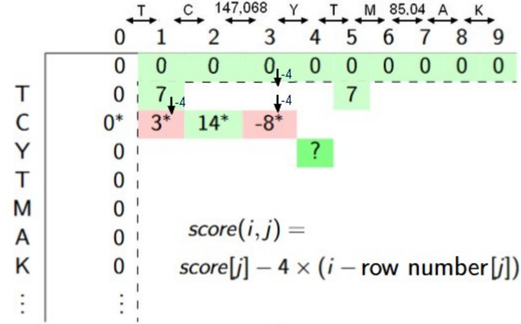
→ the score of any cell can be computed relying on the score and the row number of the cell of interest above it.



	0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0	0
T	0	7				7				
C	0*	3*	14*	-8*						
Y	0				?					
T	0									
M	0									
A	0									
K	0									
⋮	⋮									

→ Our solution: save for each column the score and the row number of the last cell of interest computed until now

Another example of improvement



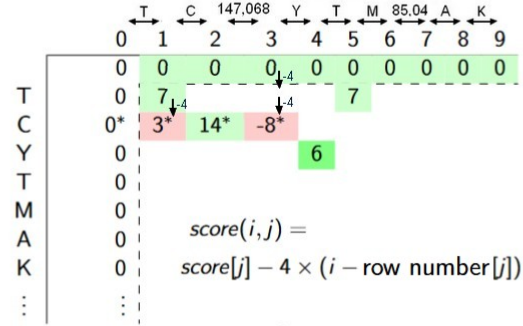
	0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0	0
T	0	7				7				
C	0*	3*	14*	-8*						
Y	0				?					
T	0									
M	0									
A	0									
K	0									
⋮	⋮									

$$score(i,j) =$$
$$score[j] - 4 \times (i - \text{row number}[j])$$

↓

row number	1	2	0	0	1	0	0	0	0
score	7	14	0	0	7	0	0	0	0

Another example of improvement



	0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0	0
T	0	7				7				
C	0*	3*	14*	-8*						
Y	0				6					
T	0									
M	0									
A	0									
K	0									
⋮	⋮									

$$score(i,j) =$$
$$score[j] - 4 \times (i - \text{row number}[j])$$

↓

row number	1	2	0	3	1	0	0	0	0
score	7	14	0	6	7	0	0	0	0

A good memory management

→ A huge effort was made on memory management:

- limit memory allocation and deallocation to limit the use of the garbage collector
- most structures are allocated once and for all at the start of execution
- these structures having to be reset between each spectrum-protein alignment, they must be designed to minimize this reset time

The impact of these improvements

Without any improvement:

↪ more than 6000 hours for a classic peptidomics dataset

The impact of these improvements

Without any improvement:

↪ more than 6000 hours for a classic peptidomics dataset

Now:

↪ approximately 9 hours for the same dataset (47 000 spectra, 20 000 proteins)

Experimental results

→ Comparison of ProSpect with 2 other algorithms, including SpecGlobX, on a proteomics dataset:

- 694 spectra → the analysis by mass spectrometry of a sample containing only the protein cytochrome C
- a protein dataset → the protein cytochrome C and 4403 other proteins

each spectrum identified by a peptide coming from the protein cytochrome C is considered rightly identified

SpecOMS/ SpecGlobX	MS-GF+	ProSpect
228	192	**273**

the number of correct identifications

Thank you for your attention