Evaluation of open search methods based on theoretical mass spectra comparison

Albane LYSIAK

PhD student 2019-2022, Université de Nantes

Supervisors : Guillaume FERTIN, Géraldine JEAN (Laboratoire des Sciences du Numérique de Nantes), Dominique TESSIER (INRAE Nantes)

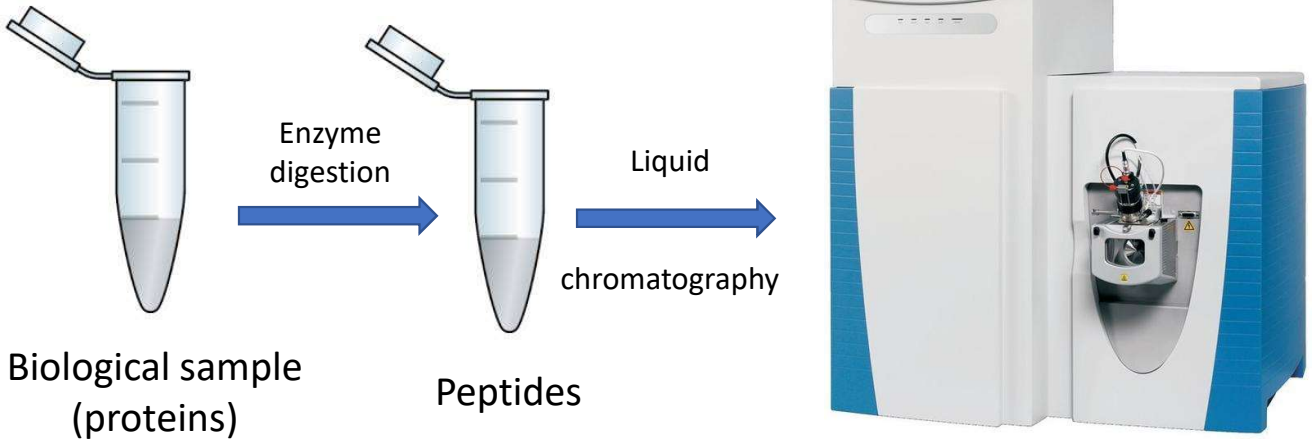
SeqBIM, November 23, 2020

UNIVERSITÉ DE NANTES

INRAE

Context : Mass spectrometry

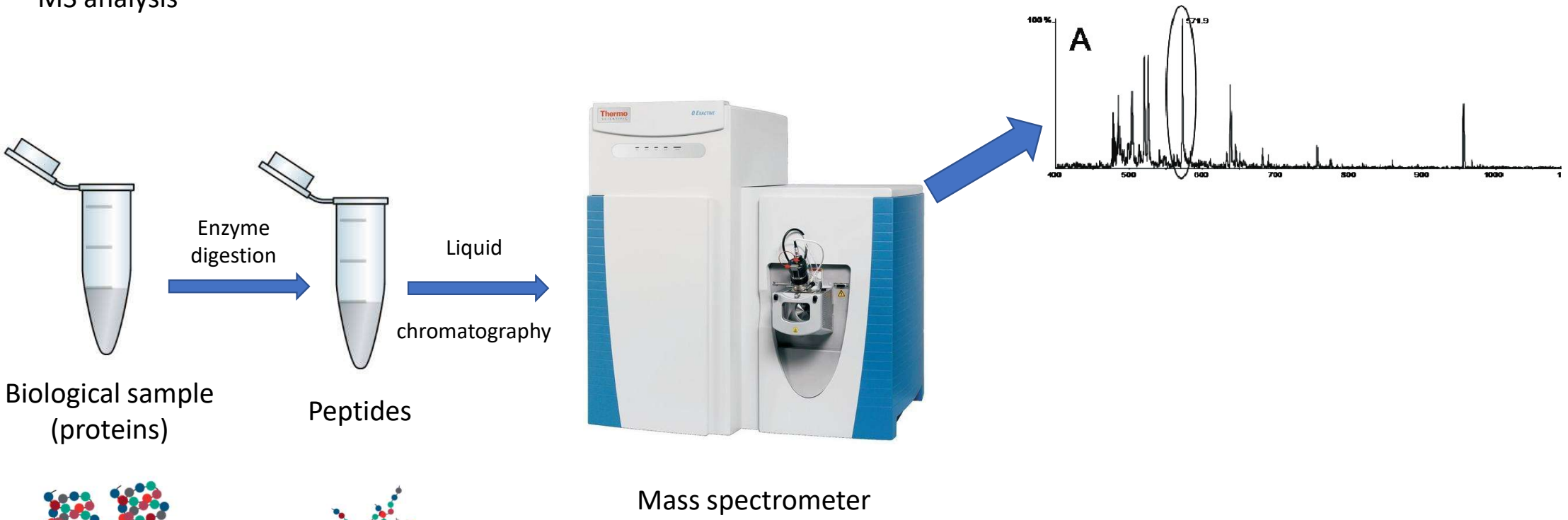
MS analysis



Context : Mass spectrometry

MS analysis

MS spectrum



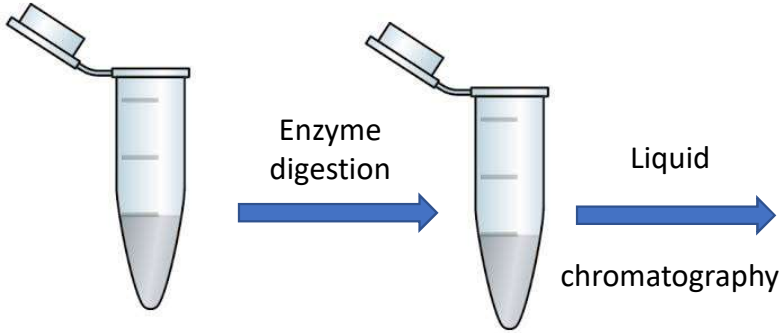
Biological sample
(proteins)

Peptides

Mass spectrometer

Context : Mass spectrometry

MS analysis



Enzyme digestion

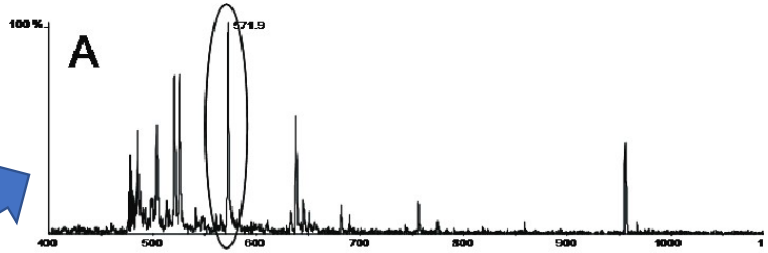
Liquid chromatography

Biological sample (proteins)

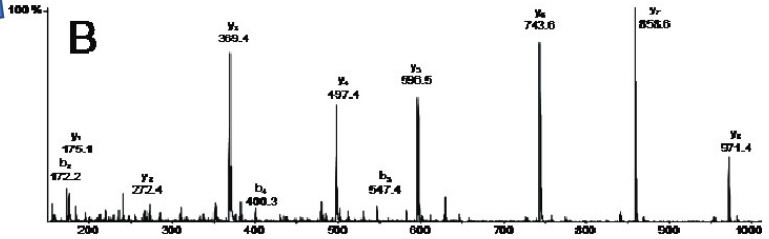
Peptides

Mass spectrometer

MS spectrum



MS/MS spectrum

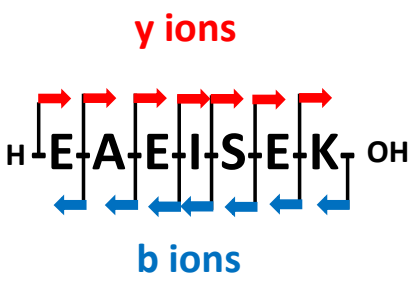


Context : Mass spectrometry

Why does a MS/MS spectrum contain sequence information ?

Sequences and masses of ions of the peptide EAEISEK (Dalton)

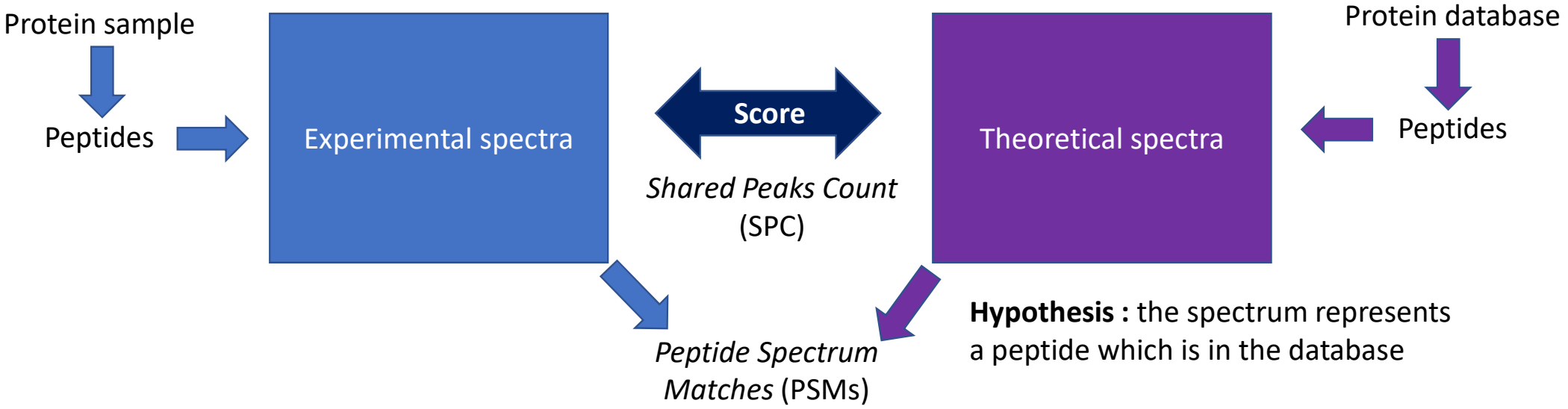
b1	E-H^+	130.0499
b2	EA-H^+	201.087
b3	EAE-H^+	330.1296
b4	EAEI-H^+	443.2136
b5	EAEIS-H^+	530.2457
b6	EAEISE-H^+	659.2883
b7	EAEISEK-H^+	787.3832



y1	H_2^+-K-OH	147.1128
y2	H_2^+-KE-OH	276.1554
y3	H_2^+-KES-OH	363.1874
y4	H_2^+-KESI-OH	476.2715
y5	H_2^+-KESIE-OH	605.3141
y6	H_2^+-KESIEA-OH	676.3512
y7	H_2^+-KESIEAE-OH	805.3938

Context : Mass spectrometry

- Mass spectrometry : reference method to identify proteins



➢ Modified peptides ?

Context : modifications and OMS methods

- OMS (*Open Modification Search*) methods
  - Accept a high mass difference (Δm) between spectra when they are compared
  - Hypothesis :
    - Conventional method : the spectrum represents a peptide which is in the database
    - OMS method : the spectrum represents a peptide which a variant is in the database

- SpecOMS (Matthieu David *et al*, 2017) :
  - Fast calculation of a simple similarity score between all couples of spectra
  - Determination of the optimal place of the supposed modification (mass shift algorithm)

Context : the mass shift algorithm

Δm=10

Masses of EAEISEK

b1	E-H^+	130.0499
b2	EA-H^+	201.087
b3	EAE-H^+	330.1296
b4	EAEI-H^+	443.2136
b5	EAEIS-H^+	530.2457
b6	EAEISE-H^+	669.2883
b7	EAEISEK-H^+	797.3832

« Raw » spc = 6 shared peaks

Masses of EAEISEK

130.0499	E-H^+	b1
201.087	EA-H^+	b2
330.1296	EAE-H^+	b3
443.2136	EAEI-H^+	b4
530.2457	EAEIS-H^+	b5
659.2883	EAEISE-H^+	b6
787.3832	EAEISEK-H^+	b7

y1	H_2^+-K-OH	147.1128
y2	H_2^+-KE-OH	286.1554
y3	H_2^+-KES-OH	373.1874
y4	H_2^+-KESI-OH	486.2715
y5	H_2^+-KESIE-OH	615.3141
y6	H_2^+-KESIEA-OH	686.3512
y7	H_2^+-KESIEAE-OH	815.3938

147.1128	H_2^+-K-OH	y1
276.1554	H_2^+-KE-OH	y2
363.1874	H_2^+-KES-OH	y3
476.2715	H_2^+-KESI-OH	y4
605.3141	H_2^+-KESIE-OH	y5
676.3512	H_2^+-KESIEA-OH	y6
805.3938	H_2^+-KESIEAE-OH	y7

Context : the mass shift algorithm

Δm=10

Masses of EAEISEK

b1	E-H^+	130.0499
b2	EA-H^+	201.087
b3	EAE-H^+	330.1296
b4	EAEI-H^+	443.2136
b5	EAEIS-H^+	530.2457
b6	EAEISE-H^+	669.2883
b7	EAEISEK-H^+	797.3832

y1	H_2^+-K-OH	147.1128
y2	H_2^+-KE-OH	286.1554
y3	H_2^+-KES-OH	373.1874
y4	H_2^+-KESI-OH	486.2715
y5	H_2^+-KESIE-OH	615.3141
y6	H_2^+-KESIEA-OH	686.3512
y7	H_2^+-KESIEAE-OH	815.3938

Δm simulated at position 6

Spc = 14

Loc	Spc
1	4
2	4
3	8
4	10
5	12
6	14
7	12

Shift spc = 14

Δm=10

Masses of EAEISEK

130.0499	E-H^+	b1
201.087	EA-H^+	b2
330.1296	EAE-H^+	b3
443.2136	EAEI-H^+	b4
530.2457	EAEIS-H^+	b5
669.2883	EAEISE-H^+	b6
797.3832	EAEISEK-H^+	b7

147.1128	H_2^+-K-OH	y1
286.1554	H_2^+-KE-OH	y2
373.1874	H_2^+-KES-OH	y3
486.2715	H_2^+-KESI-OH	y4
615.3141	H_2^+-KESIE-OH	y5
686.3512	H_2^+-KESIEA-OH	y6
815.3938	H_2^+-KESIEAE-OH	y7

Problematic

- Remains complicated to perform and evaluate (mass shift, spectra quality, unknown sequences, …)
- Understand concepts underlying algorithms

Methods

- Identifications bewteen **theoretical** spectra (human proteome)
- Each peptide is compared to all the others ; « at the limits » ; but same DB, same organism
- Modifications = sequence modifications ; insertions/deletions/substitutions

- ➢ High quality spectra
- ➢ All sequences are known : quality of identifications easier to assess => new indicators

Methods

OMS research tools : varied principles

Compare to…? Theoretical database ? Spectral librairies ?

How to filter ? D*e novo* tags ?

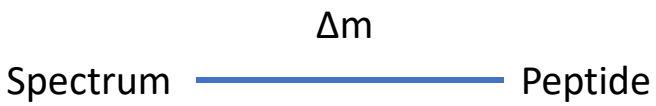
How to compare ? Large variety of scores

Use the mass difference to select the best identification ?

Selection which does not takes the Δm into account (possibly used after for a realignment)

Selection which takes the Δm into account

- Both strategies are available in SpecOMS



Methods

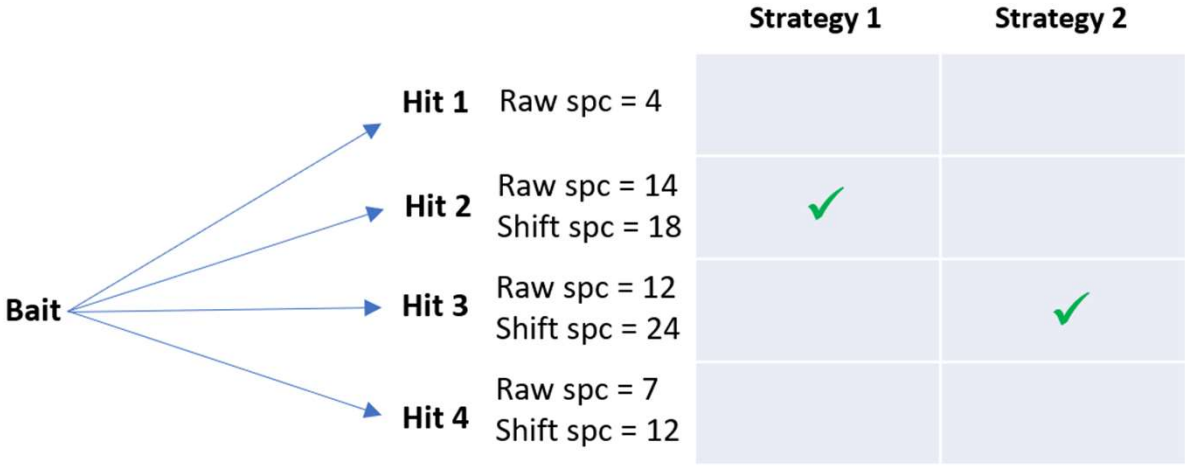
Evaluation of two OMS strategies

Strategy 1 :

The best solution for a bait is chosen according to the best *raw spc* among candidate hits

Strategy 2 :

The best solution for a bait is chosen according to the best *shift spc* among candidate hits



Methods

Evaluation of the quality of the results with two new criteria :

- **Color classification (green / orange / red)** : difficulty to get the bait sequence with available information
- ***Low Information Peaks Rate*** **(LIPR)** : common masses in a PSM which do not give any sequence information

Methods

Green / Orange / Red color classification

Bait	*Hit*	*Δm* (Da)	Shift location	Interpretation : from the hit to the bait
TSSDSISR	TSDSISR	87.0320	3	TSDSISR + S = TSSDSISR
YQEFQNR	EPPNPEYQEFQNR	-663.2864	6	EPPNPEYQEFQNR – EPPNPE = YQEFQNR
ETVHIPGAR	ETIPGAR	236.1273	2	ETIPGAR + Δm = ET-Δm-IPGAR
VISPEDGK	VIESPDGK	0	/	/
VCASIAQK	VTIQCQK	0	/	/
WFSIYDQR	FWSIQDYFR	-147.0684	/	/

Methods

Low Information Peaks Rate (LIPR)

SISSINR / SISQIESINR

Masses	Sequences
88.0393	S <---> S ✓
201.1234	SI <---> SI ✓
288.1554	SIS <---> SIS ✓
175.119	R <---> R ✓
289.1619	RN <---> RN ✓
402.2459	RNI <---> RNI ✓
489.278	RNIS <---> RNIS ✓

(a)

LIPR = 0%

NEDIIQSIQR / NDEIMVSIQR

Masses	Sequences
115.0502	N <---> N ✓
359.1197	NED <---> NDE ✗
472.2038	NEDI <---> NDEI ✗
175.119	R <---> R ✓
303.1775	RQ <---> RQ ✓
416.2616	RQI <---> RQI ✓
503.2936	RQIS <---> RQIS

(b)

LIPR = 2/7 = 28.57%

TEGIPIGIIAHGK / EASIPIGIIVVR

Masses	Sequences
288.119	TEG <---> EAS ✗
401.2031	TEGI <---> EASI ✗
498.2558	TEGIP <---> EASIP ✗
611.3399	TEGIPI <---> EASIPI ✗
668.3614	TEGIPIG <---> EASIPIG ✗
781.4454	TEGIPIGI <---> EASIPIGI ✗
894.5295	TEGIPIGII <---> EASIPIGII ✗

(c)

LIPR = 100%

Results

Strategy 1

Min raw SPC	#Green PSMs	#Orange PSMs	#Red PSMs	Total	LIPR (avg %)
7	51562	28438	375404	455404	38.5
10	29697	10160	89964	129821	35.79
15	12852	4438	8415	25705	5.05
17	**9153**	**3071**	**5006**	**17230**	**3.93**
20	5460	1676	2382	9518	4.73

Strategy 2

Min shift SPC	#Green PSMs	#Orange PSMs	#Red PSMs	Total	LIPR (avg %)
7	110279	28724	316401	455404	22.97
10	110279	28724	294605	433608	20.72
15	75330	22763	249798	347891	18.29
20	32866	9999	34028	76893	4.74
21	**27211**	**8334**	**22773**	**58318**	**2.53**
25	19176	6022	13955	39153	1.79
30	13734	4375	8960	27069	1.62
35	8108	2496	4348	14952	1.36
40	5684	1670	2505	9859	1.28

- More green PSMs for S2
- % of green PSMs increases with the score
- LIPR is smaller for S2 and decreases with the score

Conclusion

- Taking the Δm into account for the choice of the best candidate gives results easier to interprete
- Use of theoretical spectra bring interesting response elements
- Perspectives : take a closer look at red PSMs

Thank you !