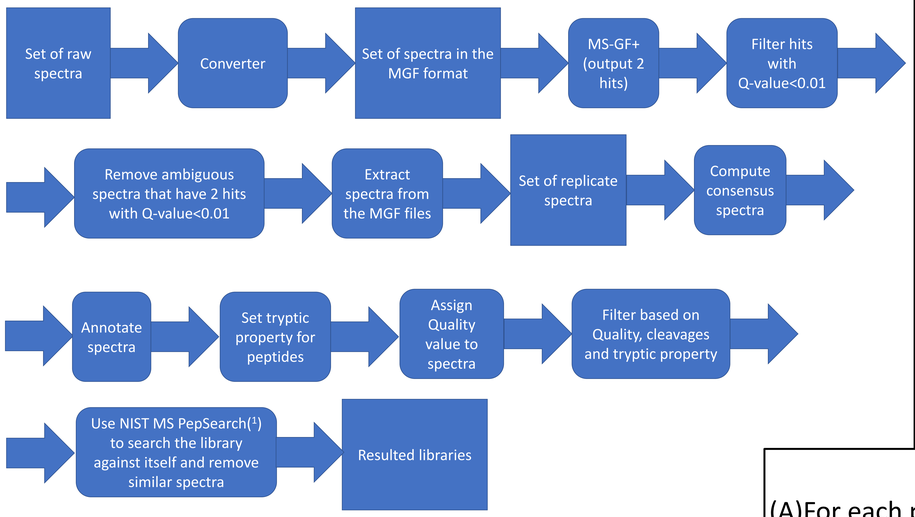
Filtering and optimization of peptide tandem mass spectral libraries

Sergey L. Sheetlin, Guanghui Wang, Dmitrii V. Tchekhovskoi , Zheng Zhang, Stephen E. Stein
Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

NIST National Institute of Standards and Technology U.S. Department of Commerce | Mass Spectrometry Data Center

General scheme of building library of consensus spectra



Filters

- F_1: number of replicate spectra.
- F_2: total annotated abundance above precursor divided by total abundance above precursor.
- F_3: absolute difference between experimental and theoretical precursors.
- F_4: penalizes large unannotated peaks depending on their abundance and m/z.
- F_5: ratio of annotated abundance above precursor and the total annotated abundance.
- F_6: $N_{a,b,y}$ divided by its average value (for a given value of peptide length) where $N_{a,b,y}$ is the maximum number of consecutive a, b or y fragment ions.
- F_7: the rank of the first unannotated peak above precursor.

Thresholds

The following inequalities are checked:

$F_1 \leq 2, F_2 \leq 0.5, F_3 \geq 5ppm, F_4 \geq 1.5, F_5 \leq 0.03, F_6 \leq 0.4,\ F_7 \leq 1$

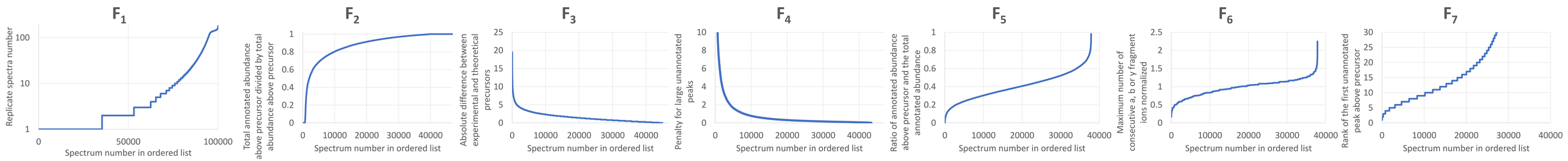
Quality is the number of conditions satisfied

Pearson correlation coefficients between filters

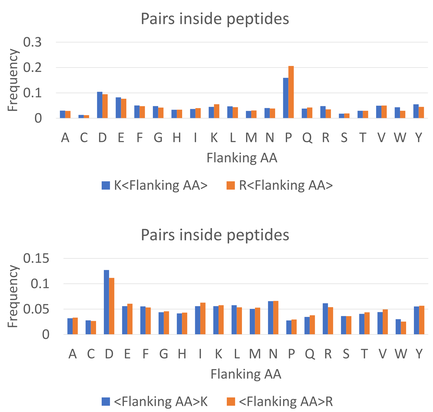
	F_1	F_2	F_3	F_4	F_5	F_6	F_7
F_1		0.15	0.06	0.07	0.44	0.14	0.12
F_2	0.15		0.35	0.27	0.45	0.46	0.67
F_3	0.06	0.35		0.10	0.21	0.40	0.35
F_4	0.07	0.27	0.10		0.46	0.13	0.24
F_5	0.44	0.45	0.21	0.46		0.25	0.33
F_6	0.14	0.46	0.40	0.13	0.25		0.47
F_7	0.12	0.67	0.35	0.24	0.33	0.47	

Applying filters for 100,000 test set of spectra

(A)For each plot, the spectra are sorted according to the filter, (B)Then the spectra with the values below the threshold are removed, (C)The remaining part is transferred to the next filter



Missed cleavages accepted into the library



Accepted:

KD or RD
KE or RE
KP or RP
DK or DR

Peptides with missed cleavages in the search results

- A sample is searched in the library using NIST MS PepSearch
- Hits with common substrings are grouped together. An example:

Fullname	Mods	Charge	Count	Average score	
R.ADDRKDDRVEERDPPR.R		4	1	750	The tryptic subsequences are too short to be included in the library
K.DDRVEERDPPR.R		3	1	401	
R.KDDRVEERDPPR.R		4	2	396	
R.AEGSDVANAVLDGADCIMLSGETAKGDYPLEAVR.M	2(15,C,CAM)(17,M,Oxidation)	3	2	581	The tryptic subsequence belongs to the library but with different charge
R.AEGSDVANAVLDGADCIMLSGETAKGDYPLEAVR.M	1(15,C,CAM)	3	2	473	
A.EGSDVANAVLDGADCIMLSGETAKGDYPLEAVR.M	3(14,C,CAM)(14,C,CAM)(16,M,Oxidation)	3	3	367	
K.GDYPLEAVR.M		2	1	626	
K.AENGKLVINGNPITIFQERDPSKIKWGDAGAEYVVESTGVFTTMEK.A		4	1	307	Multiple subsequences but charge does not match for the tryptic peptides without missed cleavages
K.IKWGDAGAEYVVESTGVFTTMEK.A	1(20,M,Oxidation)	3	1	245	
K.IKWGDAGAEYVVESTGVFTTMEK.A		3	4	690	
V.INGNPITIFQERDPSK.I		2	1	834	
K.LVINGNPITIFQER.D		2	1	759	
K.LVINGNPITIFQERD.P		2	1	636	
K.LVINGNPITIFQERDPSK.I		3	2	654	
I.NGNPITIFQERDPSK.I		2	1	870	
F.QERDPSKIKWGDAGAEYVVESTGVFTTMEK.A		4	1	485	
K.WGDAGAEYVVESTGVFTTMEK.A	1(18,M,Oxidation)	2	2	565	
K.WGDAGAEYVVESTGVFTTMEK.A		3	2	641	
K.WGDAGAEYVVESTGVFTTMEK.A		2	8	512	
					No tryptic subsequences
R.AEPEAGEEQAGDDRDSGGPVLQFDYEAVANR.L		3	1	505	
					Variations in modifications
R.AFADALEVIPMALSENSGMNPIQTMTEVR.A	2(10,M,Oxidation)(18,M,Oxidation)	3	2	533	
R.AFADALEVIPMALSENSGMNPIQTMTEVR.A	1(24,M,Oxidation)	3	1	470	
R.AFADALEVIPMALSENSGMNPIQTMTEVR.A	1(10,M,Oxidation)	3	1	715	
R.AFADALEVIPMALSENSGMNPIQTMTEVR.A		3	1	830	

Peptides with missed cleavages included into the libraries

- Peptides with missed cleavages are not very frequent but sometimes uniquely represent some region of a protein
- The number of peptides exponentially decreases with the number of missed cleavages
- Peptides with the number of missed cleavages greater than 2 are placed into a separate library

H. sapiens Orbitrap HCD libraries of consensus spectra are available (2)

Library	Spectra	Peptide ions	Peptides	Peptide ions missed cleavages
Tryptic of **Quality** 7 without missed cleavages	398,373	398,373	257,122	0
Tryptic of **Quality** 4,5,6,7 with missed cleavages	200,477	200,477	142,824	122,535
Semi-Tryptic of **Quality** 4,5,6,7 with missed cleavages	312,933	312,933	252,194	48,502
Total	911,783	911,783	652,140	171,037
The old NIST library	1,127,970	489,921	324,877	70,944

References

(1)NIST MS PepSearch:
https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:mspepsearch
(2)NIST peptide libraries:
https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:cdownload