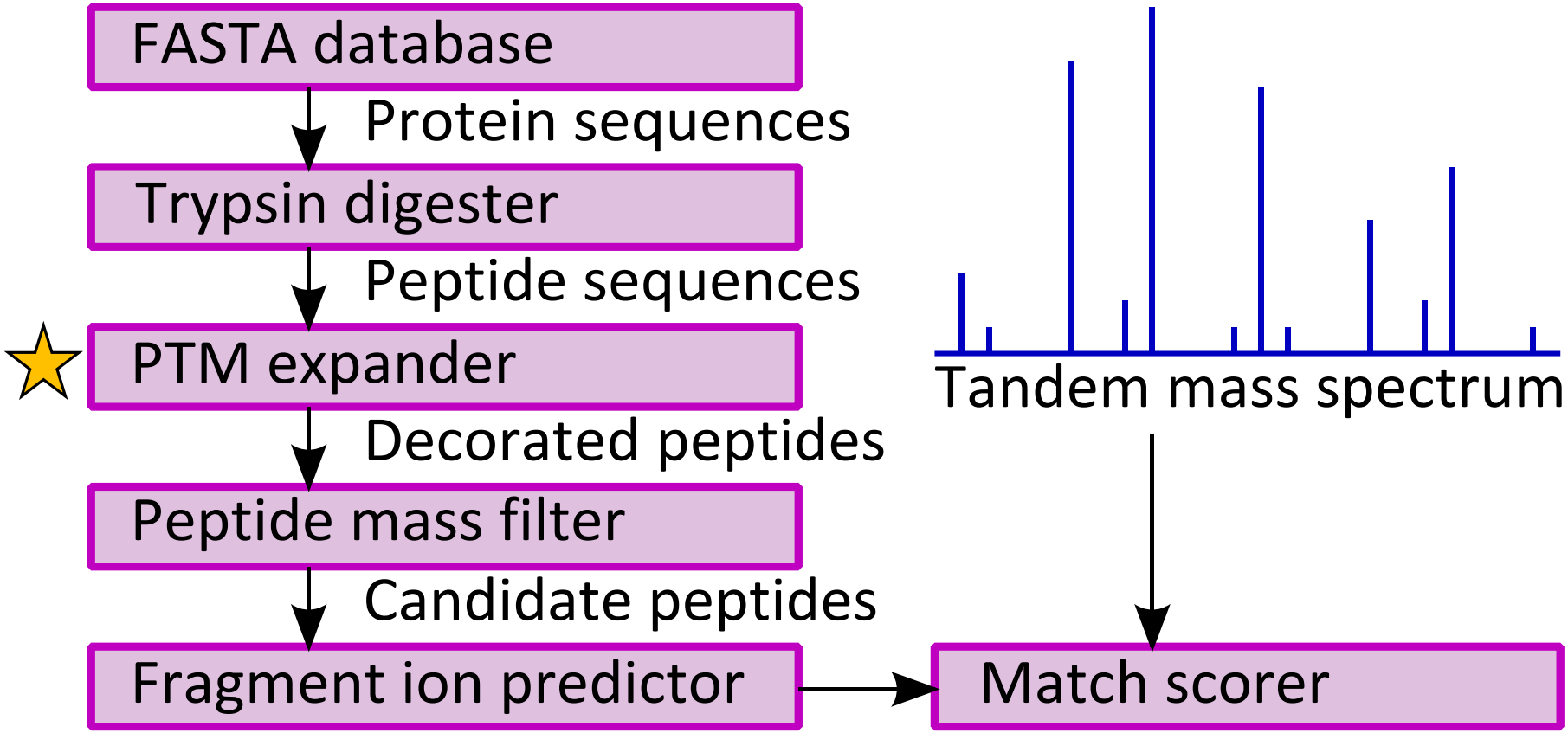
Bioinformatics for Post-translational modification ID

DAVID L. TABB, PH.D.

Overview

- Adapting DB search with “dynamic mods”
- Understanding common modifications
- Opening DB search to ignore precursor mass
- Controlling FDR in PTM-containing PSMs
- Localizing post-translational modifications

Database search overview



Eng et al (1994) *J. Amer. Soc. Mass Spectrom*. 5: 976-989.

Eng et al (1995) *Anal. Chem.* 67: 1426-36.

Each dynamic PTM branches search space

Because multiple PTMs may be in each peptide, adding PTMs to a search creates an *exponential* cost.

Three sites in two states give eight decorations

$$2^3 = 8$$

D I G S E S T E K

D I G S*E S T E K

D I G S E S*T E K

D I G S E S T*E K

D I G S*E S*T E K

D I G S*E S T*E K

D I G S E S*T*E K

D I G S*E S*T*E K

PTM Catalog: unimod.org

UNIMOD protein modifications for mass spectrometry

Help

Unimod

Logged as **Guest** Log out Change password Advanced search

Search for: Any field | Contains | acetylation | Search | Show all

Details found: 6 Page 1 of 1

Records Per Page:: 20

	Accession #	PSI-MS Name	Interim name	Description	Monoisotopic mass	Average mass	Composition
View	766		Met-loss+Acetyl	Removal of initiator methionine from protein N-terminus, then acetylation of the new N-terminus	-89.029920	-89.1594	H(-7) C(-3) N(-1) S(-1)
View	1	Acetyl	Acetyl	Acetylation	42.010565	42.0367	H(2) C(2) O
View	37	Trimethyl	tri-Methylation	tri-Methylation	42.046950	42.0797	H(6) C(3)
View	1372		Acetyl:13C(2)	heavy acetylation	44.017274	44.0220	H(2) 13C(2) O
View	1042		Acetyldeoxyhypusine	Acetyldeoxyhypusine	97.089149	97.1582	H(11) C(6) N
View	1043		Acetylhypusine	Acetylhypusine	113.084064	113.1576	H(11) C(6) N O

DM Creasy and JS Cottrell. *Proteomics* (2004) 4: 1534-1536

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Chemical modifications: the usual suspects

Sometimes on N-termini, too

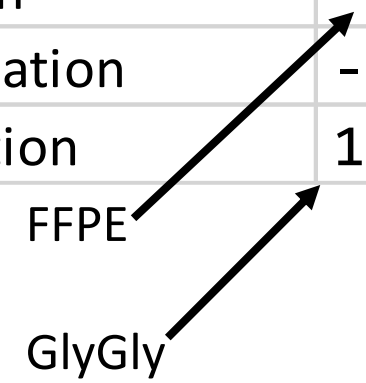
Site	Cause	Chemical identity	Shift in Da
Cys	Iodoacetamide	Carbamidomethylation	57.021464
Cys	MMTS	Beta-methythiolation	45.987721
Met	Exposure to air	Oxidation	15.994915
N-term Gln	Side chain attack	Pyro-Glutamate	-17.026549
N-term and Lys	iTRAQ-4	Isobaric label	144.102063
N-term and Lys	TMT-6	Isobaric label	229.162932
Lys	Old urea solution	Carbamylation	43.005814
Asn	Deamidation of NG	Aspartate	0.984016

See also PNGase F

Also consider adducts from acrylamide gels (+71.037114 Da)

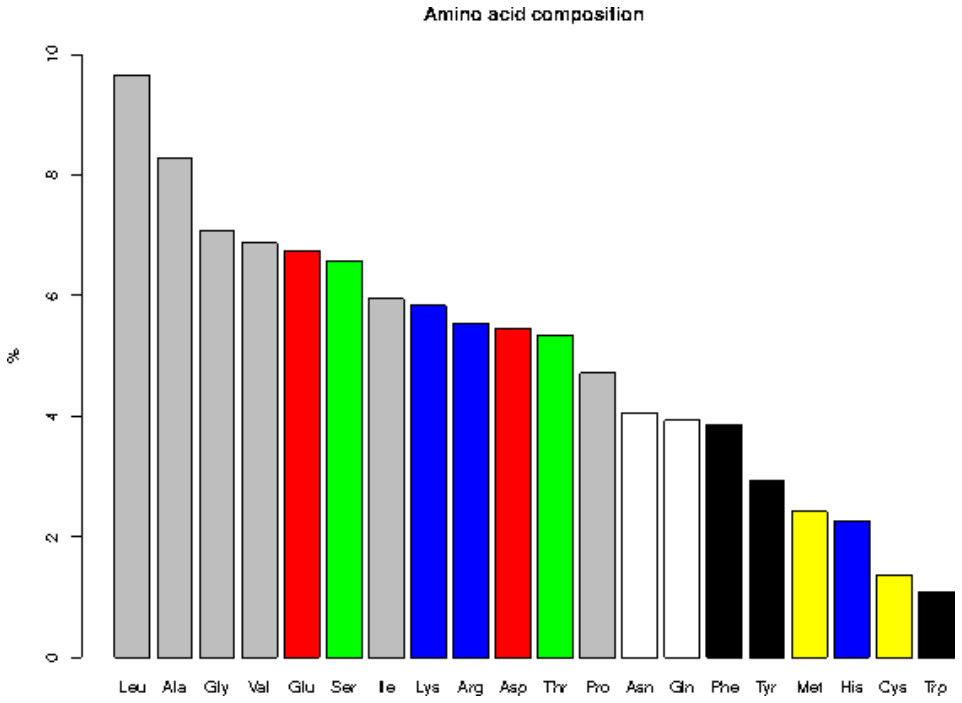
In vivo modifications: the usual suspects

Site	Cause	Chemical identity	Shift in Da
N-term and Lys	Acetyltransferases	Acetylation	42.010565
Ser, Thr, Tyr	Kinases	Phosphorylation	79.966331
Pro in tissues	Prolyl hydroxylase	Hydroxyproline	15.994915
Arg, Lys	Methyltransferases	Methylation	14.015650
C-term	PA monooxygenases	Alpha amidation	-58.005479
Lys	E3 ubiquitin ligases	Ubiquitination	114.042927



PTMs on common AAs are costly

Ala	8.26	Gly	7.08	Pro	4.71
Arg	5.53	His	2.27	Ser	6.58
Asn	4.05	Ile	5.94	Thr	5.34
Asp	5.46	Leu	9.66	Trp	1.09
Cys	1.37	Lys	5.83	Tyr	2.92
Gln	3.93	Met	2.41	Val	6.87
Glu	6.74	Phe	3.86		



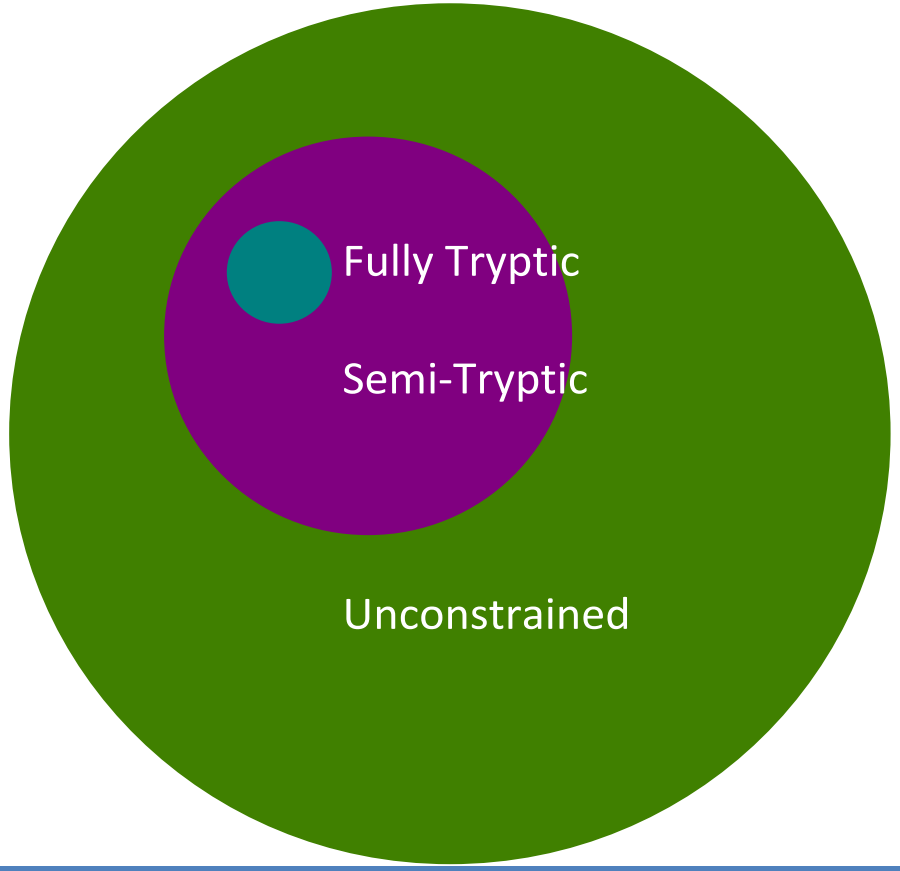
Adding a mass shift for a common amino acid slows performance far more than for a rare amino acid.

Swiss-Prot release notes, Jan 9, 2015

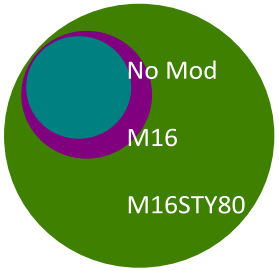
UNIVERSITEIT STELLENBOSCH UNIVERSITY

How does search time scale?

Trypsin Specificity



PTMs, fully tryptic

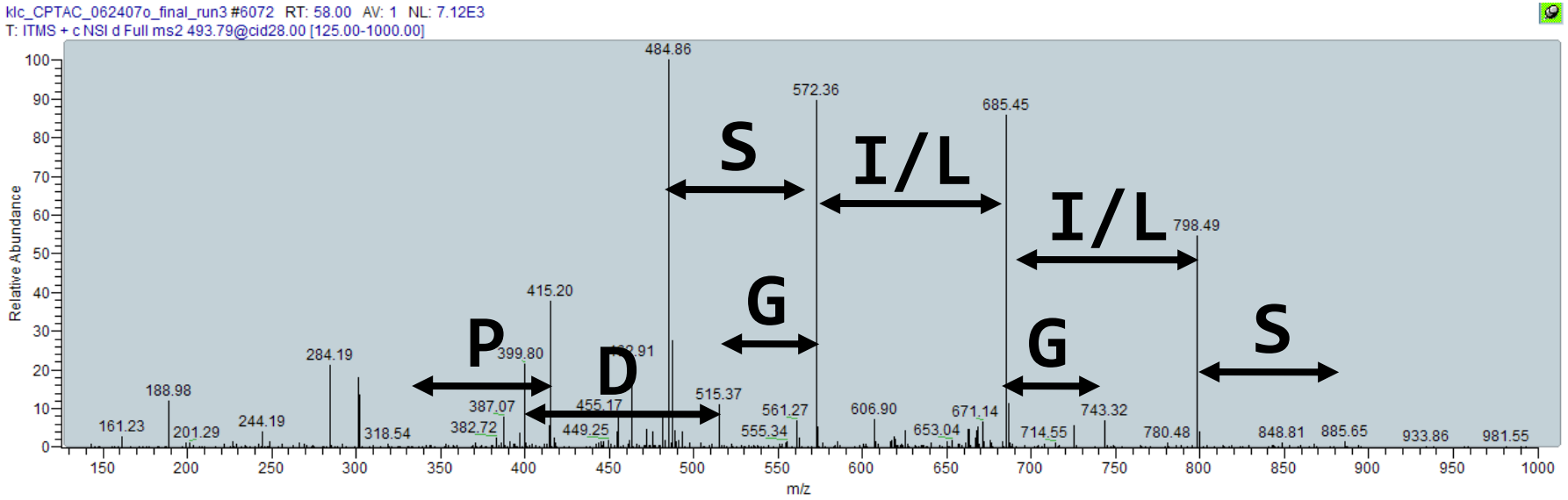


Area of circle represents the number of comparisons between a decorated peptide and an MS/MS.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Sequence tagging route to PTMs hindered by inference and reconciliation

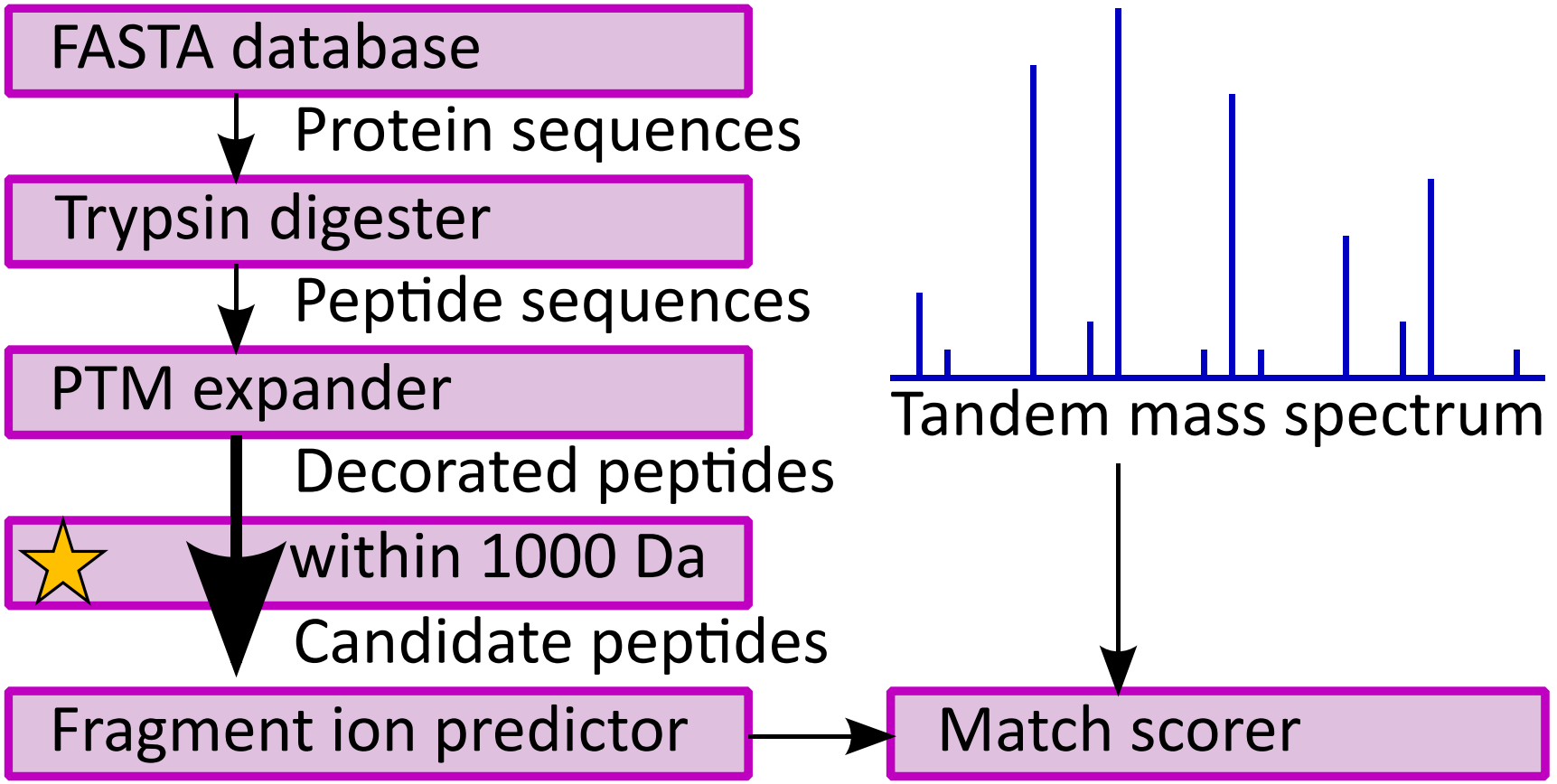
Which gaps are best to chain together? Given a good "tag" for three AAs, how do we interpret that against DB sequences?



S. Dasari et al. *Chem. Res. Toxicol*. (2011) 24: 204-216

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Open search matches scans to DB peptides with very different masses



The promise of Open Search

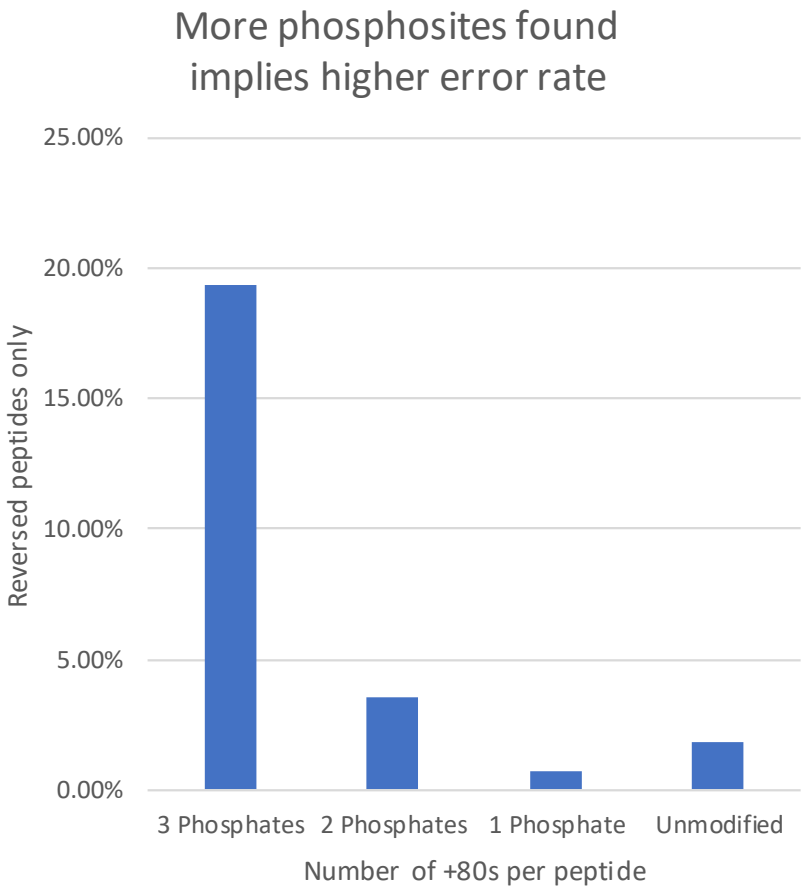
- Wide precursor tolerance allows matching despite large precursor mass difference.
- Open Search claims many more spectra identified and far more PTMs recognized.
- Making DB search fast enough required fragment index-based matching:
  - MSFragger: Kong et al. *Nat. Methods* (2017) 14: 513.
  - Open pFind: Chi et al. *Nat. Biotech.* (2018) 36: 1059.

The cautions for Open Search

1. Enlarged search space requires greater conservatism in PSM filtering.
2. The Open Search is hypothesis-forming; *don't use it as your final search*!
3. Distilling a PTM configuration for DB search from open search takes thought.

Global PSM FDR control may not control PTM FDR!

- TiO_2 peptide enrichment
- 24 RAWs PXD006230
- 2% PSM FDR control
  - 160 3Phos peptides
  - 1237 2Phos peptides
  - 7549 1Phos peptides
  - 3259 0Phos peptides



MSFragger followed by IDPicker 3.1

“Warp, spindle, fold, and mutilate”

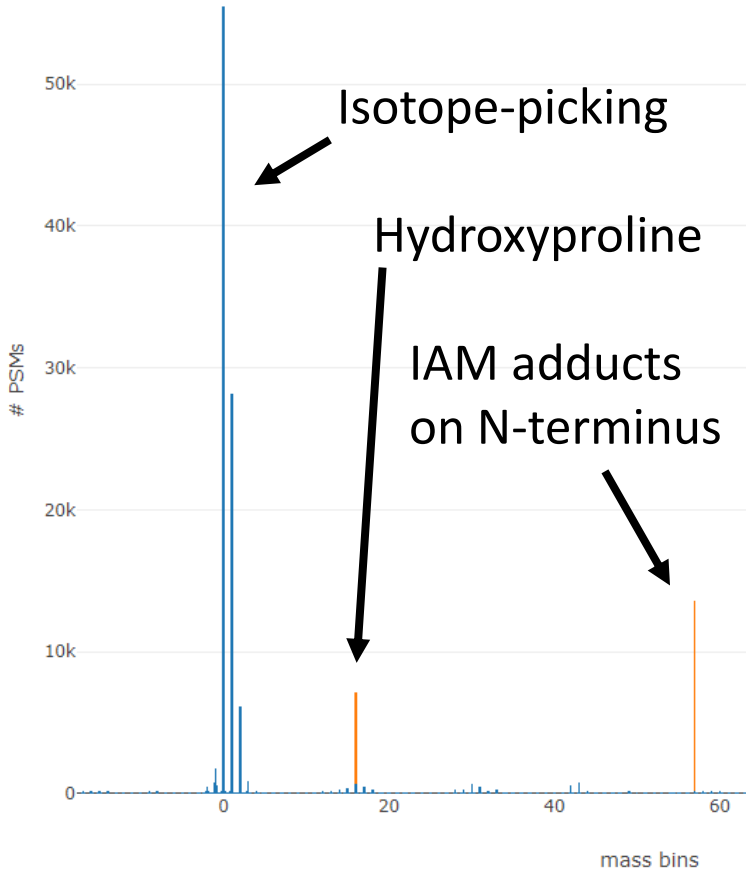
- Allowing your algorithm to alter peptide structures significantly will produce false matches by distorting sequences with PTMs.
- The more degrees of freedom given a PSM, the greater the proportion of false discoveries at a given score.
- The PSMs with the most PTMs are the class with the highest false discovery rate.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Blind search: one “delta” of any mass in any one location

Sequence tagging often specifies amino acid bearing the PTM; open search may specify a mass shift for the PSM without location.

At right, PSM mass deltas from MSFragger on “Nair” HCD FFPE glioblastoma set



Mass shift tables highlight dominant mass shift types

UNIVERSITEIT STELLENBOSCH UNIVERSITY

DeltaMass	A	C	D	E	F	G	H	I	K	L	M	N	P
6	1			1		8			2	1	2		
7				1	1	1							
8	1		5	1				13		2			
9			9			1		6		1			1
10			5										
11	13	2	2	9	4		2	3	5				5
12	30	23	22	44	19	20	133	36	56	30	41	21	8
13	26		7	11	7	14	33	10	65	12	2	16	9
14	55		18	38	7	127	147	73	866	21	2	24	45
15	16	12	15	33	16	70	61	17	267	20	4	28	114
16	64	1	23	25	41	237	8	17	65	16	3339	31	1289
17	16		5	14	13	81	1	8	16	3	69	7	265
18	5		4	2	1	27	2	6	1	6	13		48
19			1			4		2		1	4		13

Savitski, *Mol. Cell. Proteomics* (2006) 5:935-948.

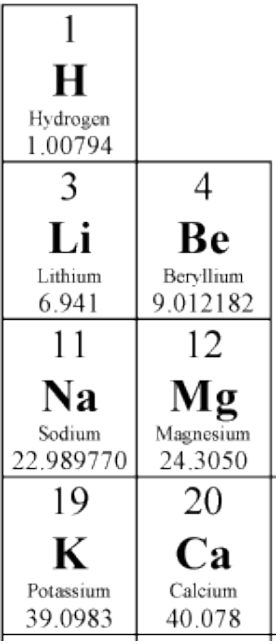
Dasari, *Chem. Res. Tox.* (2011) 24: 204-216.

Gaining expertise in blind interpretation

- Boring is more often correct than brilliant.

  +22 Da is Sodium, not Asp → His

- Blind PTMs are useful for finding *patterns* of mass shifts; do *not* put faith in individual peptide-spectrum matches.
- Unusual cleavages can appear as peptide-terminal mass shifts in blind searches.



Site localization of PTMs

- Multiple PTM-decorations of a sequence may tie in score for a spectrum, or nearly tie.
- One can say this peptide and PTM explain the spectrum, but is position of the PTM correct?
- Ascore defines a new score to estimate probability from differentiating fragments.
- Delta score techniques compare original DB search scores of variants to assess site error.

Beausoleil. *Nature Biotechnology* (2006) 24: 1285-1292.
Bailey. *J. Proteome Res*. (2009) 8: 1965-1971.
Savitski. *Mol. Cell. Proteomics* (2011) 10: M110.003830.
Taus. *J. Proteome Res.* (2011) 10: 5354-62.

Takeaway messages

- Many techniques support the identification and verification of PTMs.
- Stepping beyond database search is useful when multiple PTMs can be found.
- Resisting false discoveries is essential because PTM search spaces are much larger.
- Biological databases contribute perspective to evaluate the reasonableness of PTMs.