Identifying Post-translational Modifications

Nuno Bandeira

Executive Director

NIH/NCRR Center for Computational Mass Spectrometry

Skaggs School of Pharmacy and Pharmaceutical Sciences

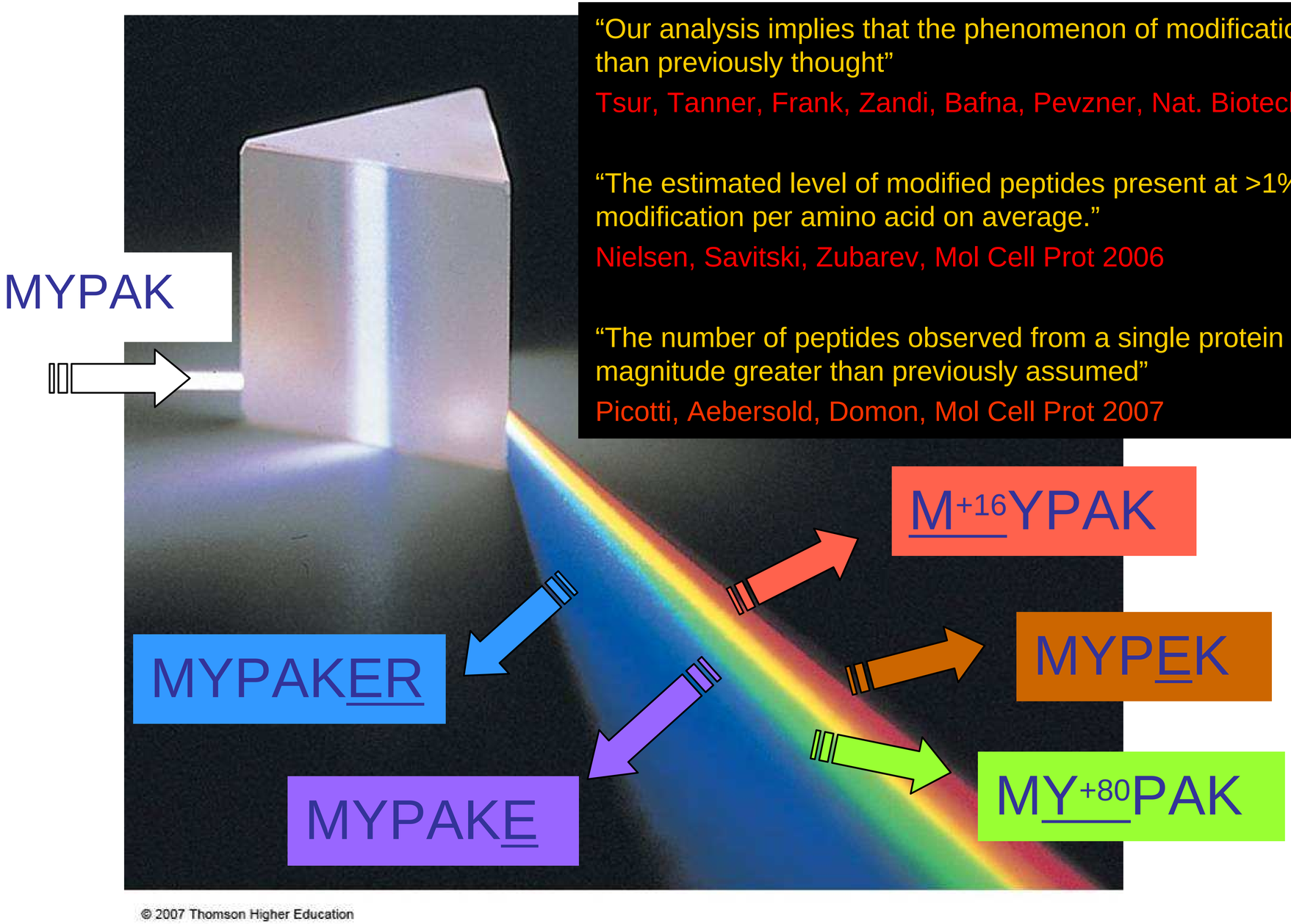
Dept. Computer Science and Engineering

University of California, San Diego

Center for Computational Mass Spectrometry

National Center for Research Resources

The dynamic proteome



"Our analysis implies that the phenomenon of modification is much more widespread than previously thought"

Tsur, Tanner, Frank, Zandi, Bafna, Pevzner, Nat. Biotech 2005

"The estimated level of modified peptides present at >1% level is approaching one modification per amino acid on average."

Nielsen, Savitski, Zubarev, Mol Cell Prot 2006

"The number of peptides observed from a single protein is at least one order of magnitude greater than previously assumed"

Picotti, Aebersold, Domon, Mol Cell Prot 2007

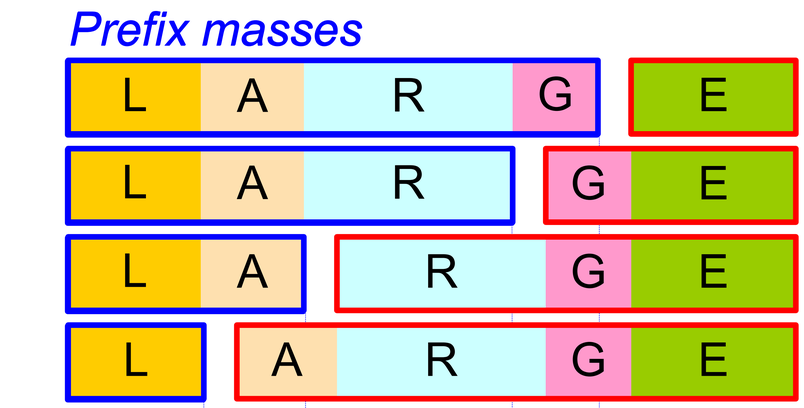
http://apollo.lsc.vsc.edu/classes/met130/notes/chapter19/graphics/prism.jpg

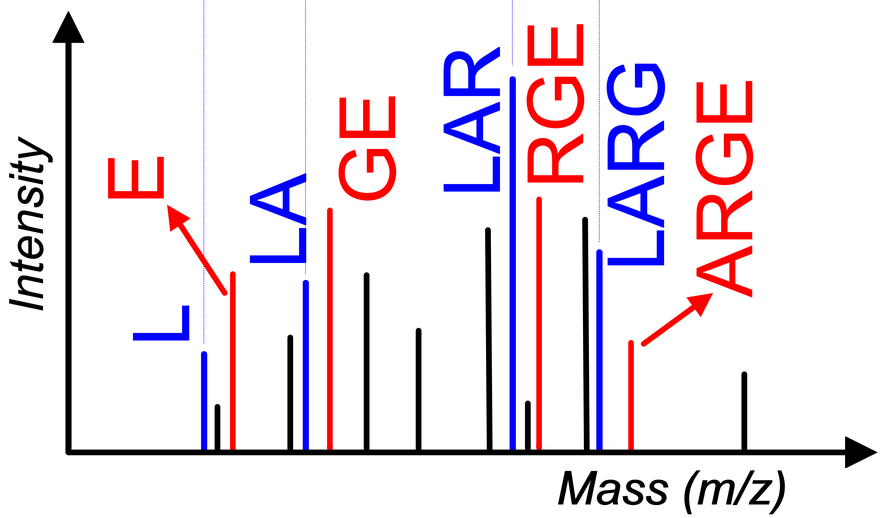
http://proteomics.ucsd.edu

Center for Computational Mass Spectrometry

Tandem Mass Spectrometry (MS/MS)

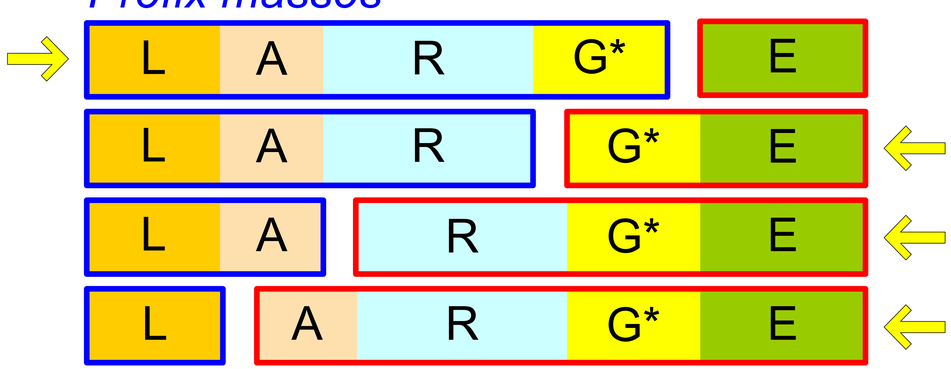
Peptide LARGE

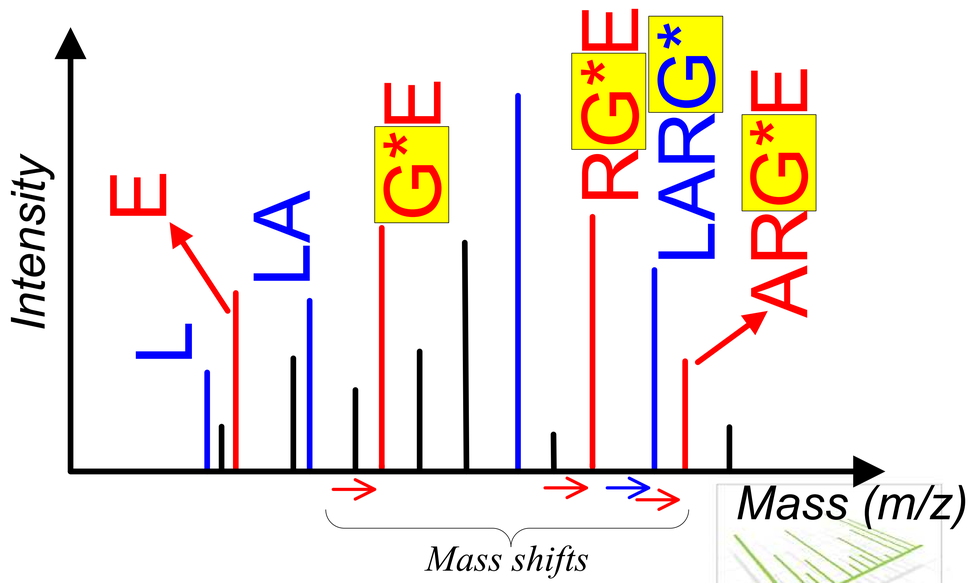




MS/MS spectrum

Modified peptide LARG*E



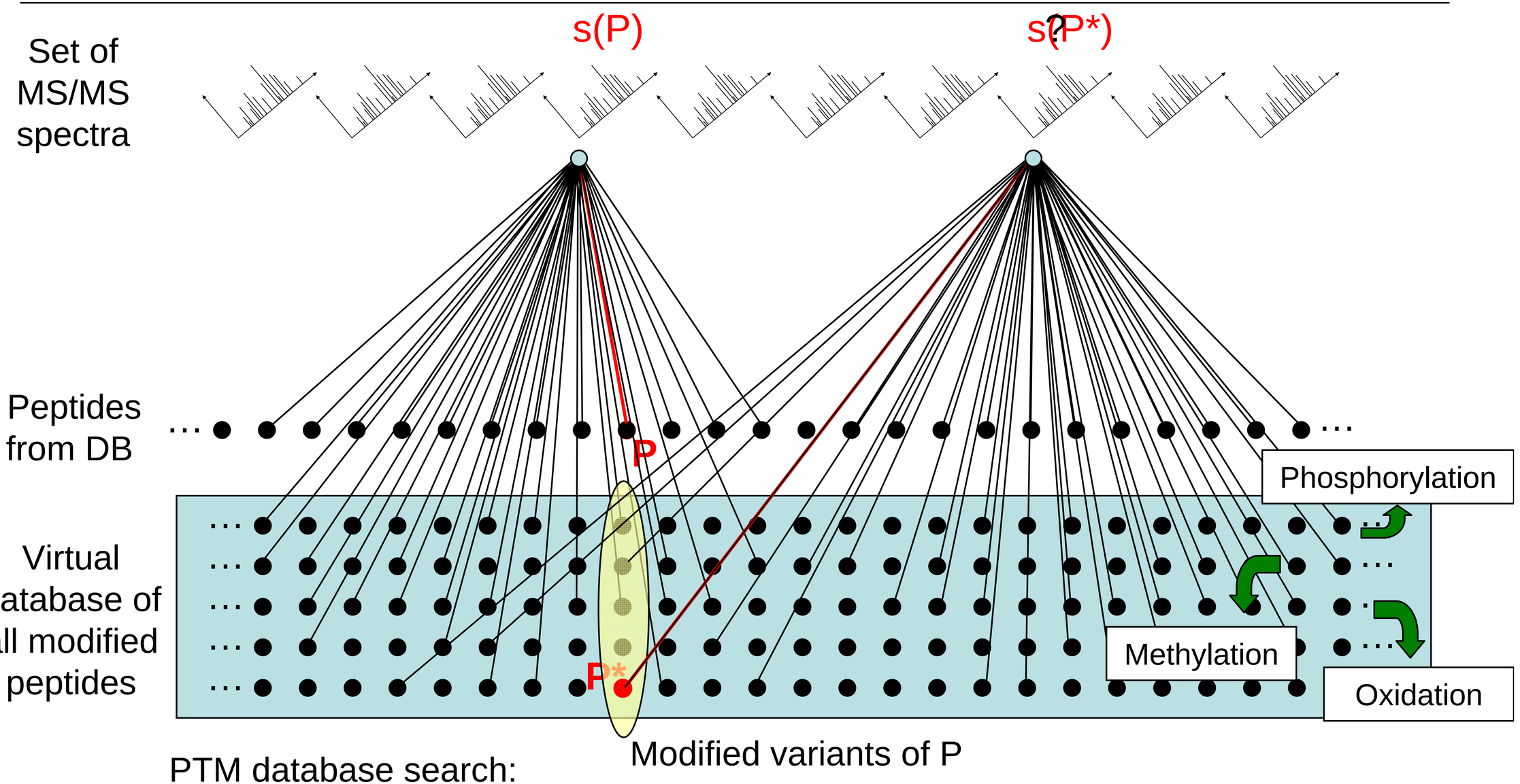


Modification: any event that changes the mass at a specific site.

http://proteomics.ucsd.edu

Center for Computational Mass Spectrometry

MS/MS spectrum identification



PTM database search:

⇒ Virtual database size restricts the allowed number of modifications

⇒ Becomes computationally heavy (i.e., slow)

⇒ Stricter thresholds for same False Discovery Rate

http://proteomics.ucsd.edu

Center for Computational Mass Spectrometry

Computational strategies

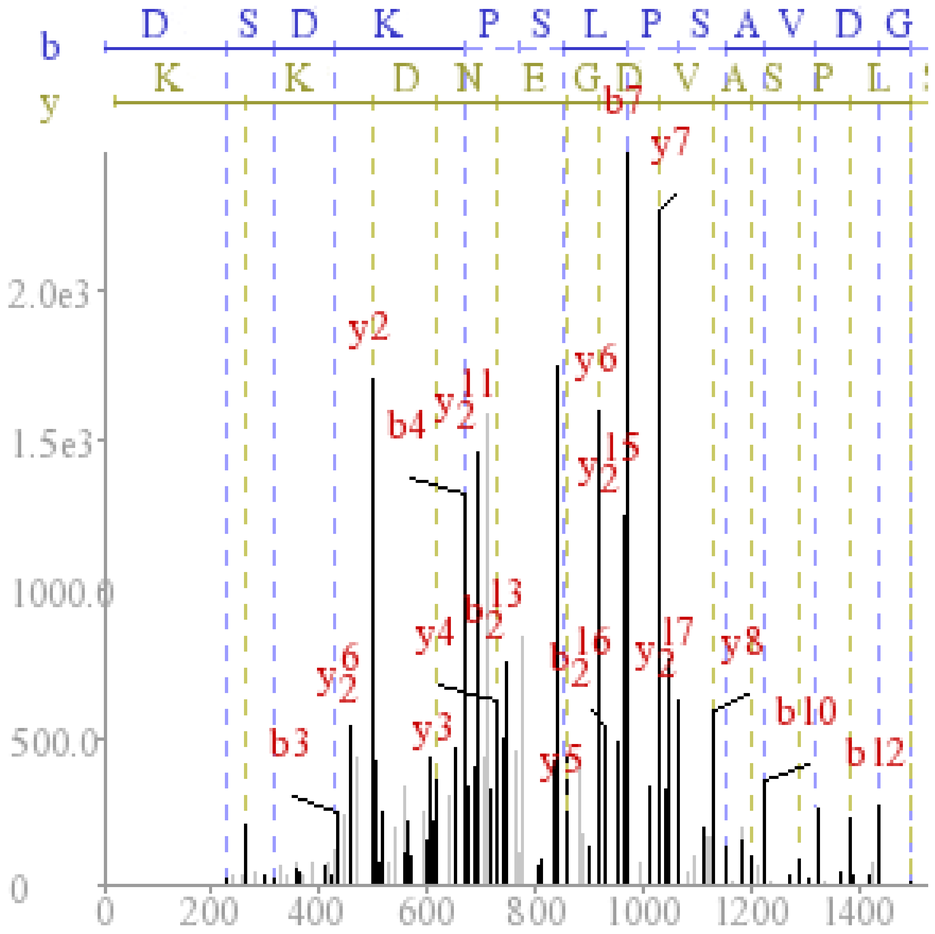
- InsPecT: tag-based search
  - Derives amino acid sequence tags from each modified spectrum and only considers DB peptides containing one of the reconstructed tags
  - Pro: filtered virtual database reduces FDR; much faster than standard approaches
  - Con: misses identifications if spectrum has no correct sequence tag (typically 95%+ sensitivity for top 50 tags)

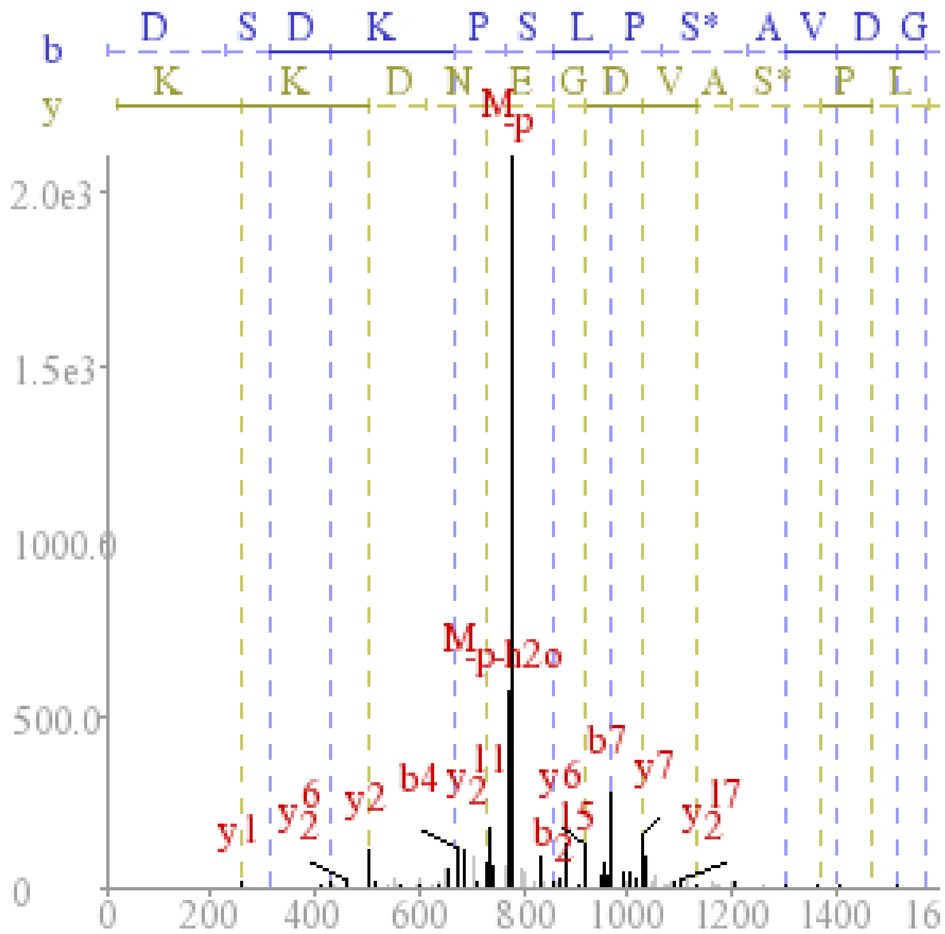
- Alternative approach: two-pass search
  - First identify proteins using spectra from unmodified peptides then search for modifications only on proteins from the first pass
  - Pro: speedup inversely proportional to complexity of the sample
  - Con: misses modified proteins with no unmodified peptides, difficulties estimating FDRs (small Decoy databases, should not re-search spectra identified in first pass)

Tanner et al, Anal Chem 2005

PTMs may change fragmentation

Phosphorylation: weak signal in b and y ions due to phosphate loss





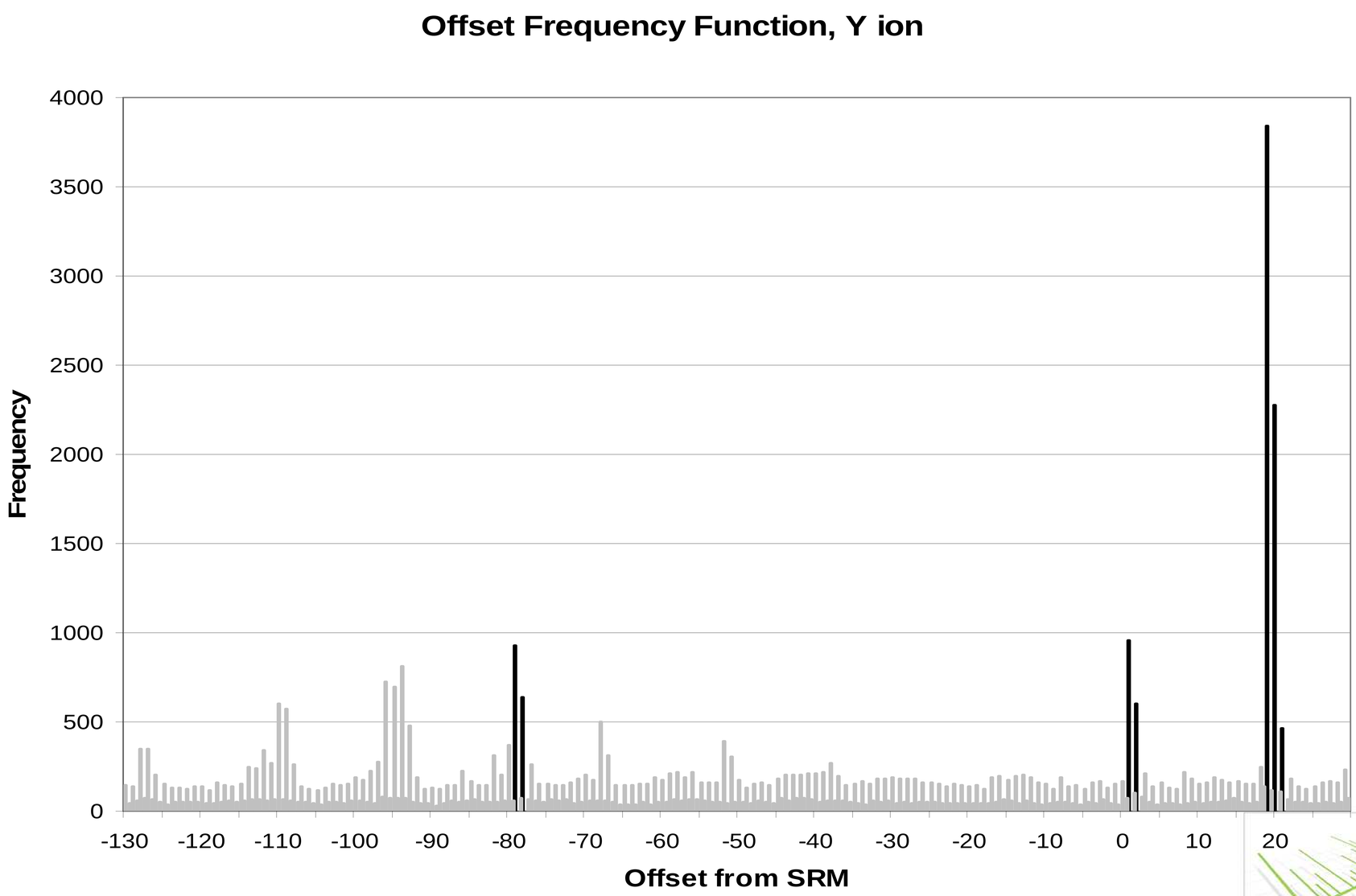
Payne et al, J Prot Res 2008

http://proteomics.ucsd.edu

Center for Computational Mass Spectrometry

Modification Changes Fragmentation

- New ion observed, fragment neutral loss



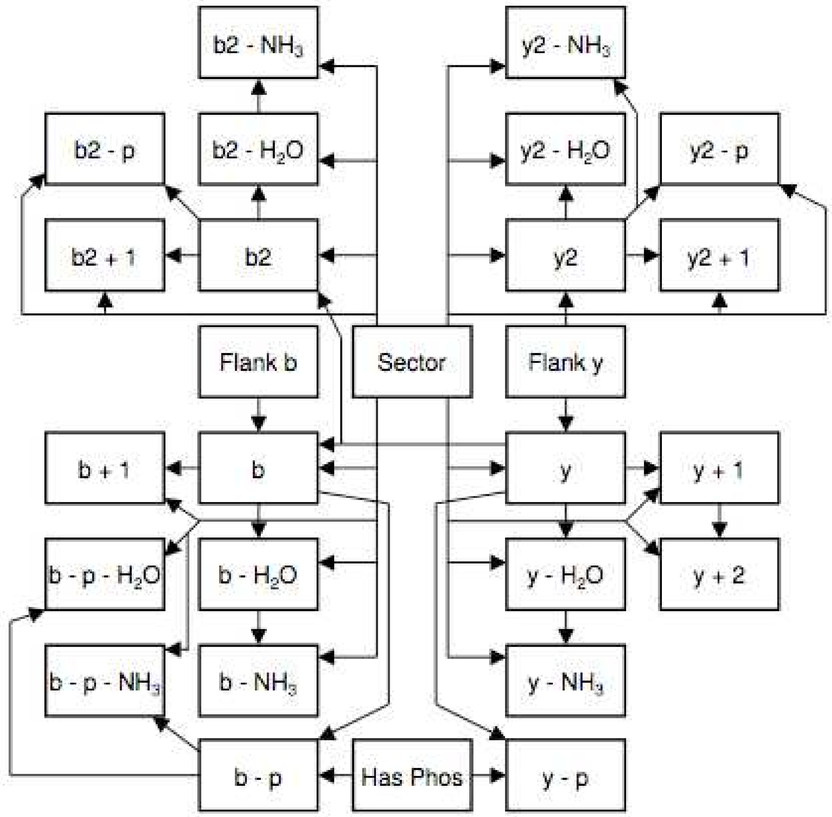
Payne et al, J Prot Res 2008

http://proteomics.ucsd.edu

Center for Computational Mass Spectrometry

InsPecT Scoring Paradigm

- Ions generated by fragmentation are not independent
- Peak intensities taken into account
- Model the probability of observing in CID with a Bayesian network.



$$P_{CID}(\vec{I} = [I_0, I_1, I_2, \ldots] \mid P_j, S) \approx \prod_i P_{CID}(I_i \mid P_j, I_{\pi(i)}, S)$$

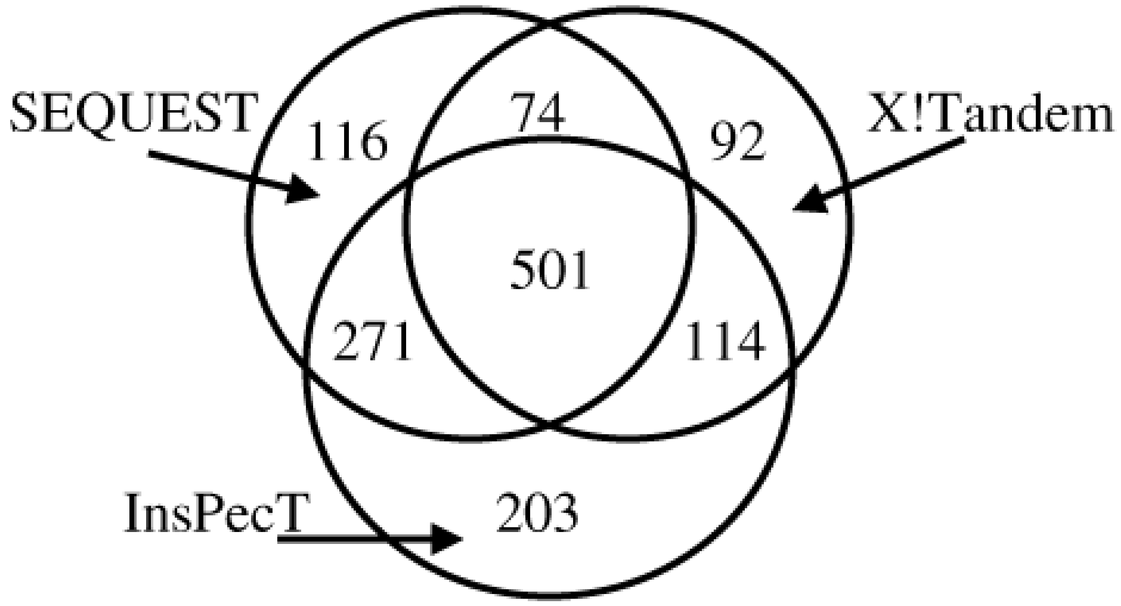
Payne et al, J Prot Res 2008

http://proteomics.ucsd.edu

Center for Computational Mass Spectrometry

InsPecT results

- Benchmark with SEQUEST and X!Tandem
  - 6410 LTQ MS/MS, IMAC, *S. cerevisiae*
  - Up to 2 phosphorylations (+80 on S,T,Y) per peptide
  - 1% FDR



Run Time
Inspect: 30 min
X!Tandem: 6 hours
SEQUEST: 36 hours

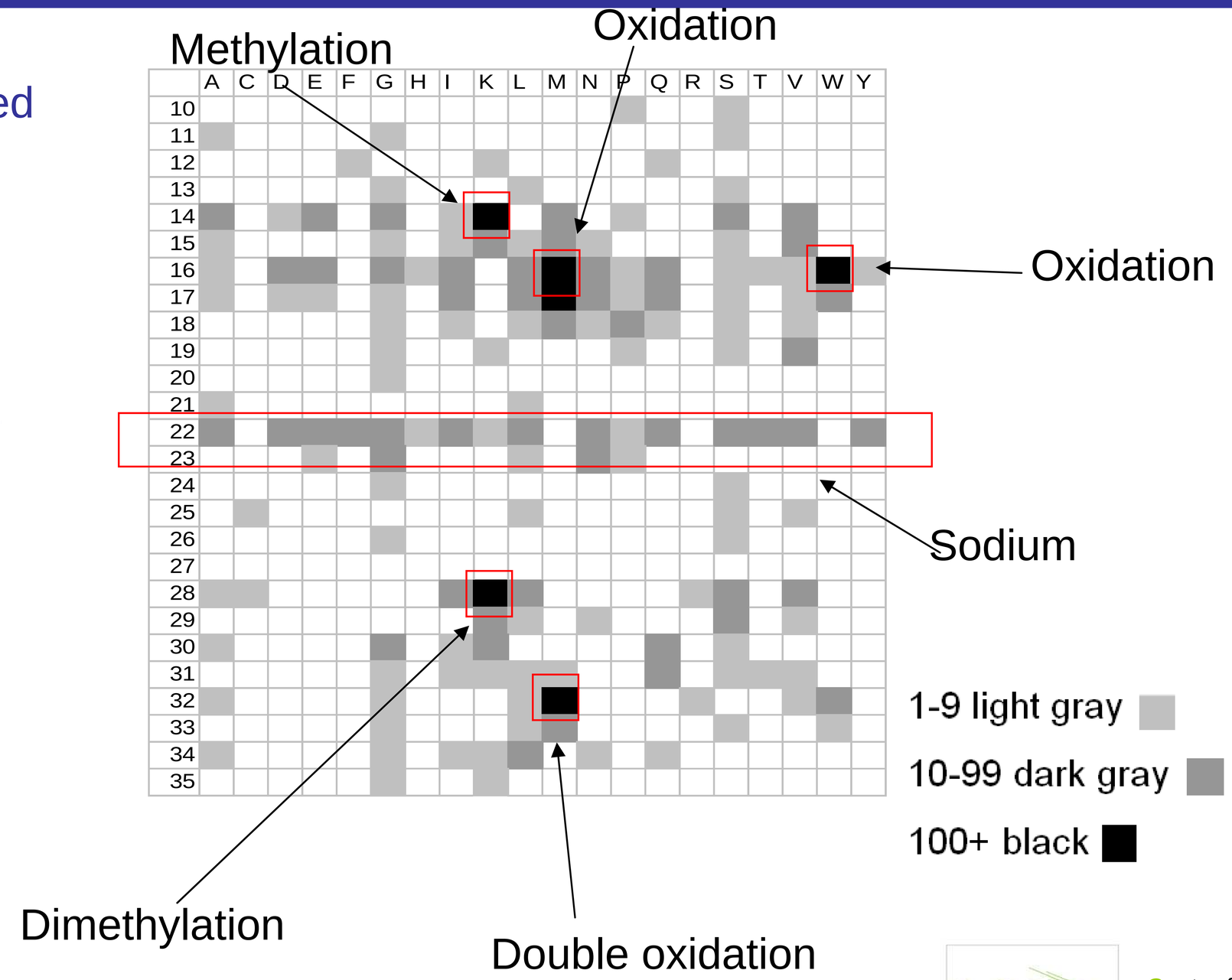
Payne et al, J Prot Res 2008

http://proteomics.ucsd.edu

Center for Computational Mass Spectrometry

PTM Frequency Matrix: strength in numbers

- Over-represented mass-shifts represent the ubiquitous modifications.
- Can we reliably detect the lower abundance modifications?



Tsur et al, Nat Biotech 2005

http://proteomics.ucsd.edu

Center for Computational Mass Spectrometry

PTMFinder

14 on K (methylation)		
K*LSSPATL	9	0
K*LSSPATLN	1	0
K*LSSPATLNS	36	0
K*LSSPATLNSR	8	0
IMLIK*LSSPATLNSR	1	0
TLDNDIM+16LIK*	4	11
IITHPNFNGNTLDNDIMLIK*	4	6
IITHPNFN+1GNTLDNDIMLIK*	2	2
IITHPNFNGNTLDNDIM+16LIK*	4	24

28 on S (mutation to D)		
GPGTS*ILSTWIGGSTR	3	0
FGPGTS*ILSTWIGGSTR	1	0
DIFGPGTS*ILSTWIGGSTR	21	0
DIFGPGTS*ILSTWIGGSTRSISGT	2	0
DIFGPGTS*ILSTWIGGSTRSISGTSMATPHVAGLA	3	0

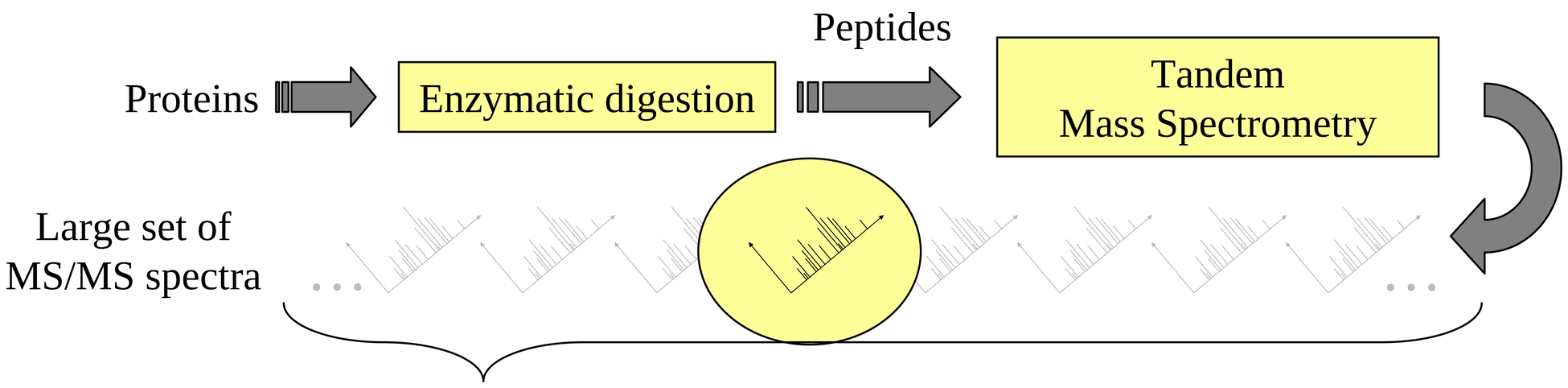
Overlapping peptides help confirm modifications

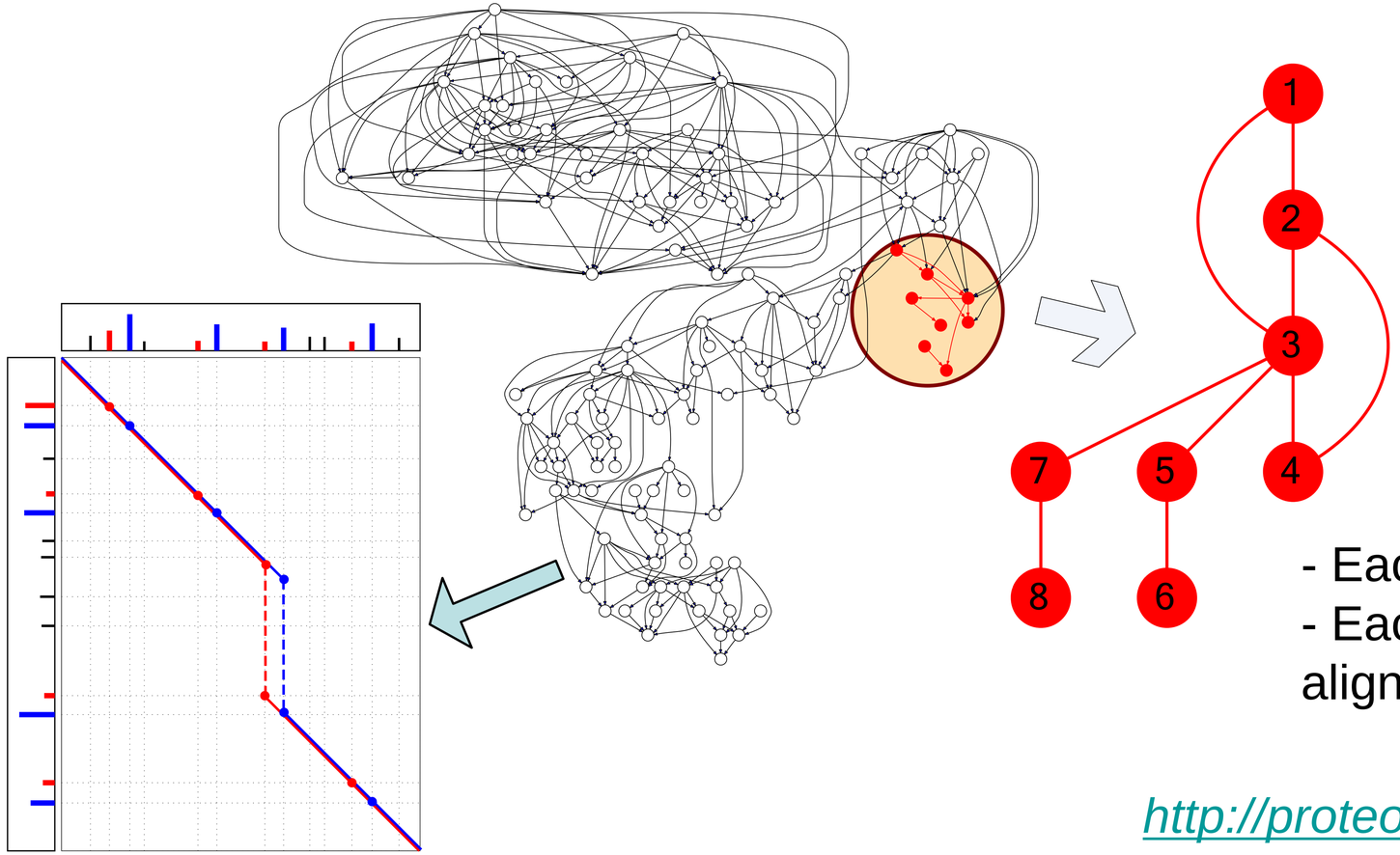
Tsur et al, Nat Biotech 2005
Tanner et al, J Prot Res 2007

http://proteomics.ucsd.edu

Center for Computational Mass Spectrometry

Spectral Networks






```
1   KQGGTLDD   LEE   QAREL
2   KQGGTLDD   LEE   QARE
3   KQGGTLDD   LEE   QAR
4   KQGGTLDD   LEE   QA
5   KQGGTLDD   LEE-18QAR
6   KQGGTLDD   LEE-18Q
7    QGGTLDD   LEE   QAR
8    QGGTLDD-53LEE   QAR
```

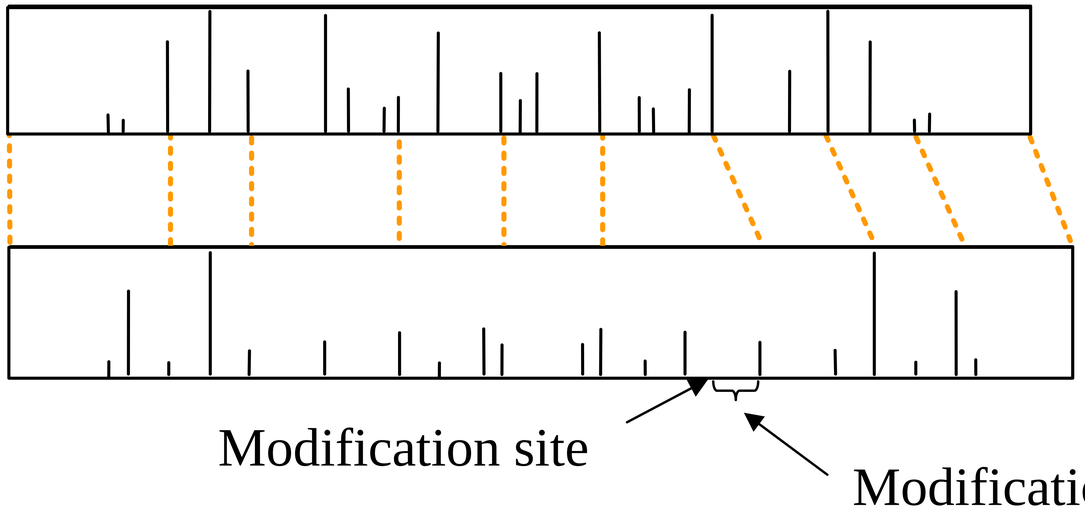
- Each node is a spectrum
- Each edge is a spectrum/spectrum alignment

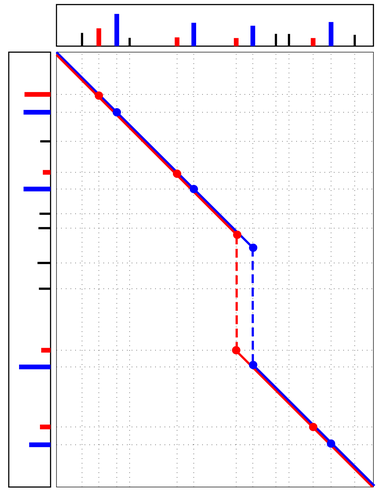
http://proteomics.ucsd.edu

Computational Mass Spectrometry

Spectral Alignment

Spectral alignment reveals the mass and location of post-translational modifications.





Sample of cataractous lens from a 93-year old patient

- Collaboration with Larry David @ Oregon Health and Science University
- Lens proteins do not turnover and accumulate modifications over time
- Intensively studied in Searle et al.'04, Tsur et al.'05 and Wilmarth et al.'06
- Detected over 70,000 spectral alignments

OREGON HEALTH&SCIENCE UNIVERSITY

Bandeira et al, PNAS 2007

http://proteomics.ucsd.edu

Center for Computational Mass Spectrometry

Modifications on cataractous lens

Location	Modification Mass	Putative annotation
S,T	-18	dehydration
Q	-17	deamidation
W	-2	cross-linking
H	14	methylation
M,W	16	oxidation
S,H	28	double methylation
N-term	42	acetylation
N-term	43	carbamylation
K,non-terminal	43	carbamylation
W	44	carboxylation
R	55	unknown
K	58	carboxymethylation
K	72	carboxyethylation

Table 1: Rediscovered all modifications previously identified by blind database search.

Table 2: Identified 6 new modification events

Location	Modification mass	Type	Putative annotation	Comment
M	-48	Chem. artifact	loss of methane sulfenic acid	*reported on same site*
W	4	PTM	kynurenine	*reported in cataractous lenses*
S	30/73	unknown	unknown	
W	32	PTM	formylkynurenine	*reported in cataractous lenses*
N-term	57	unknown	carboxyamidomethylation	*In-vivo N-term modification?*
N-term	271	unknown	unknown	

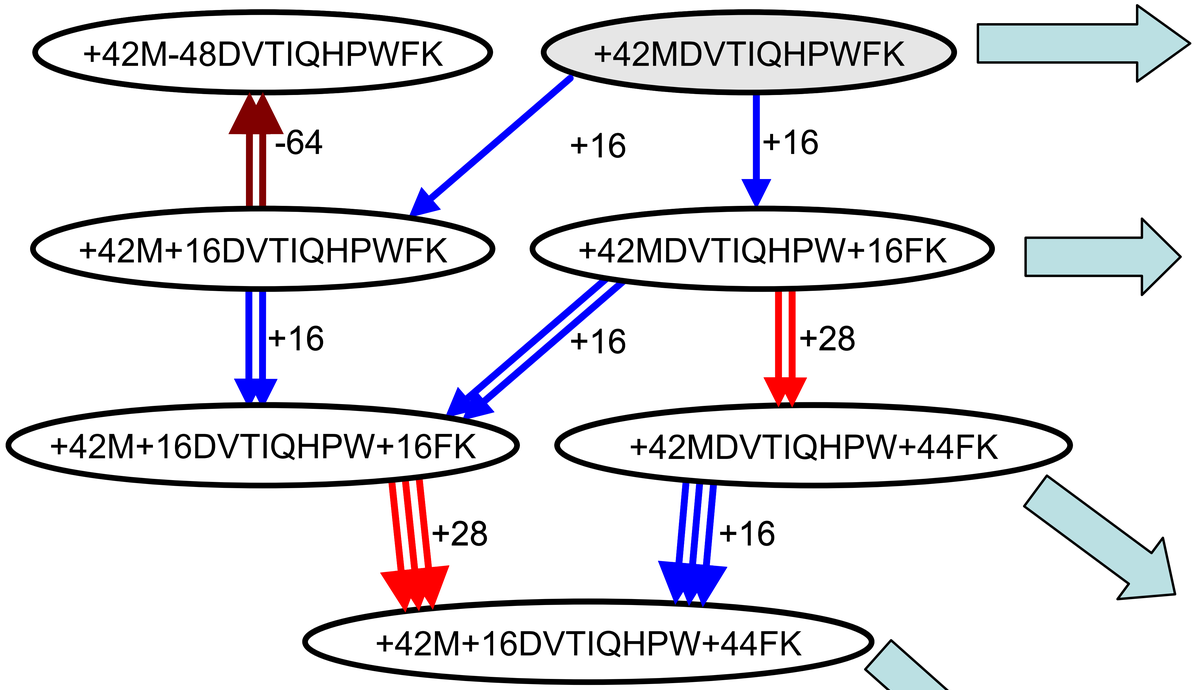
Bandeira et al, PNAS 2007

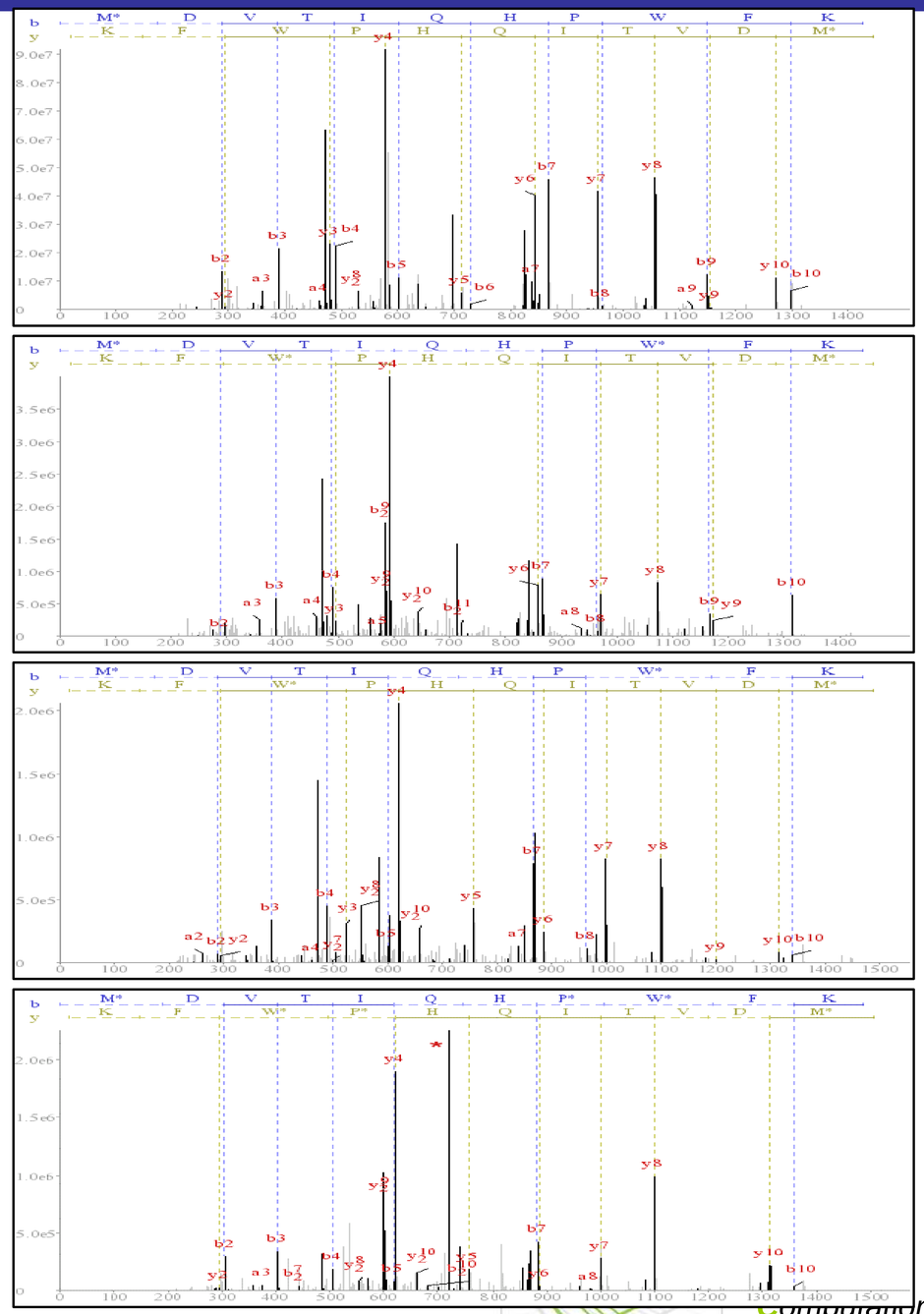
http://proteomics.ucsd.edu

Center for Computational Mass Spectrometry

Spectral networks of modified variants

Focus on a single peptide: MDVTIQHPWFK





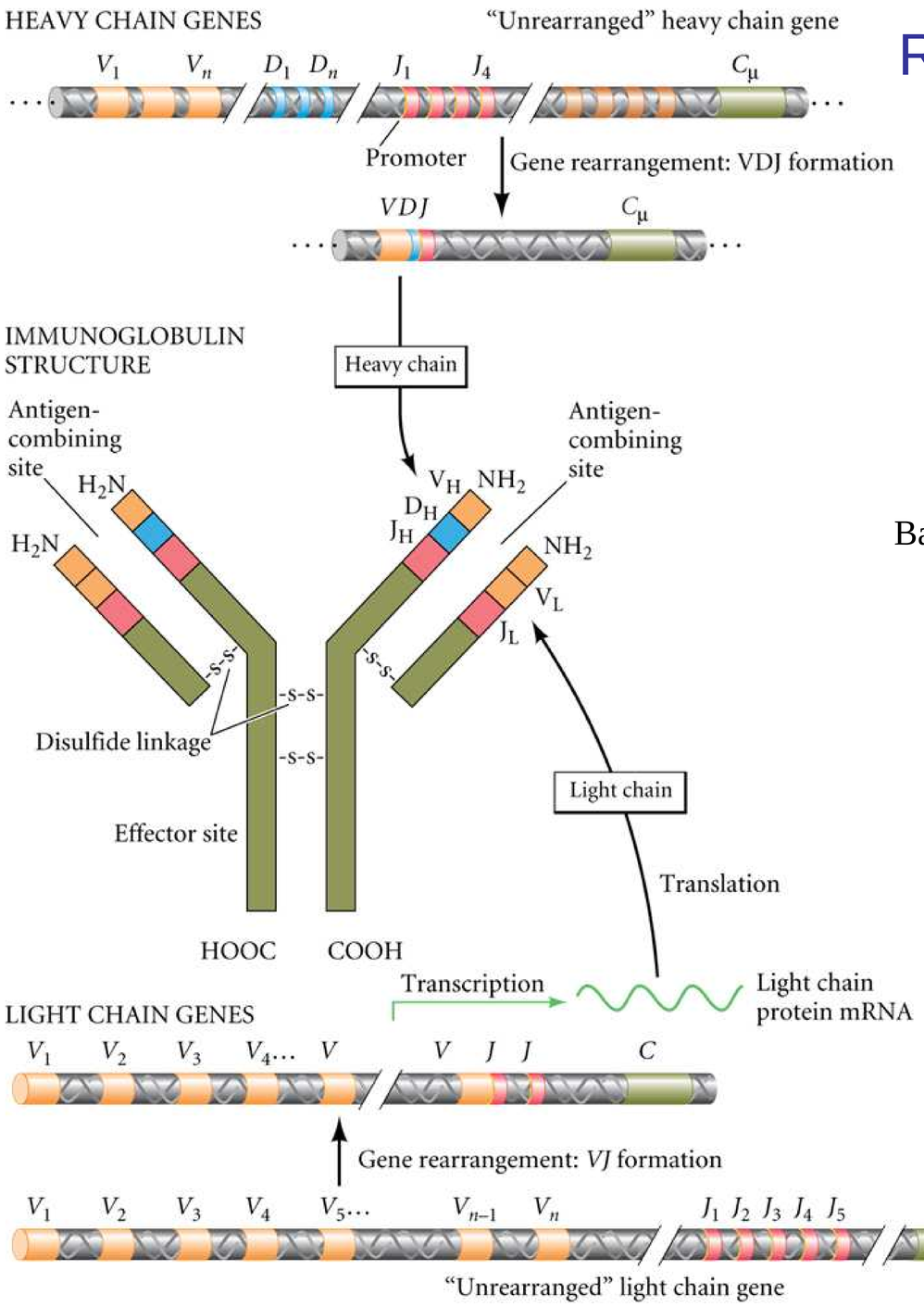
All modifications discovered on cataractous lens were supported by spectral networks.
-PTMs without a database
-Triply-modified peptide
-All unrestricted modifications

Bandeira et al., PNAS 2007

http://proteomics.ucsd.edu

Computational Mass Spectrometry

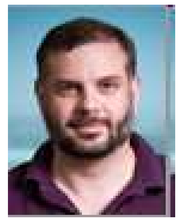
Characterizing monoclonal antibodies



Recombinant sequences create a challenge for de novo sequencing

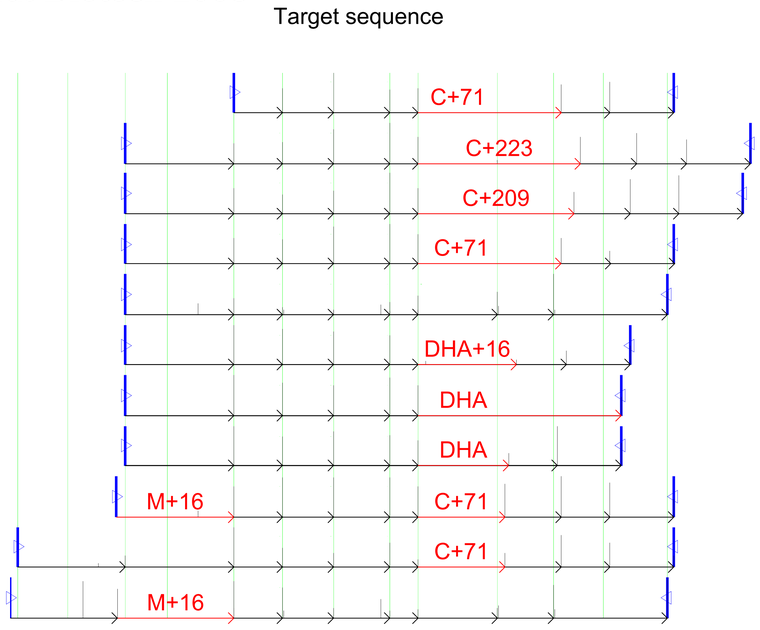
- Full characterization includes PTMs
- Mutations are a *"special kind of modification"*

(see PSGR, Tue 10:45am)



Genentech
IN BUSINESS FOR LIFE

Bandeira et al, Nat Biotech 2008



er for putational ; Spectrometry

http://proteomics.ucsd.edu

CCMS

Conclusions

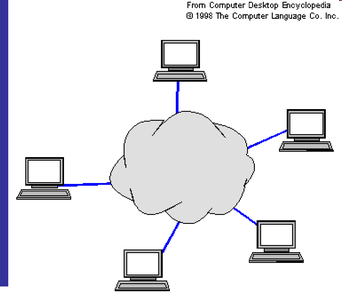
Possible strategies

- Known modifications: filtration, PTM-specific scoring
- Blind search: search any mass offsets, singly-modified peptides
- Spectral Networks: search spectra against spectra, consensus interpretation, highly modified peptides

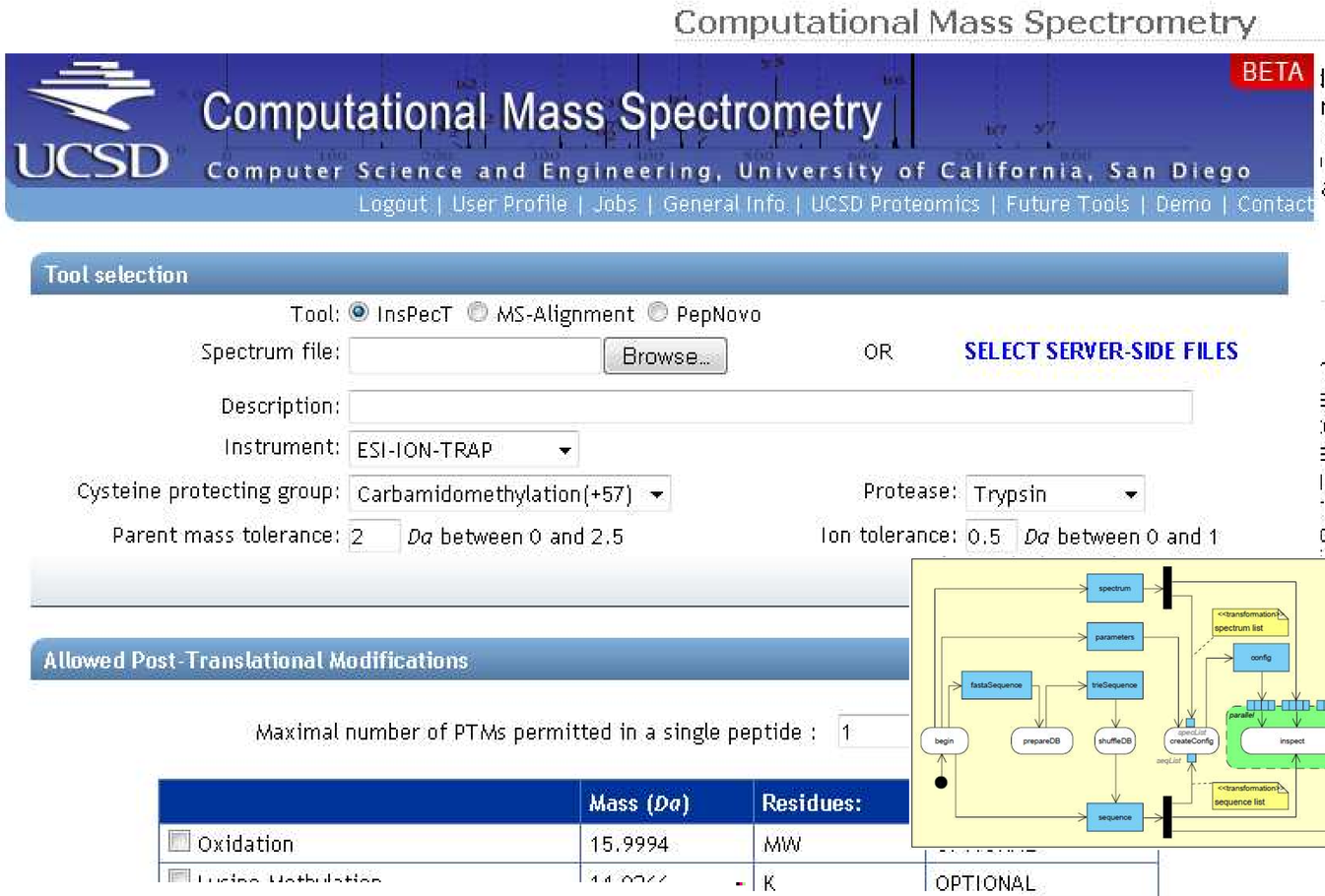
Main considerations

- False Discovery Rate stringency depends on size of virtual database, strategies may not be Target/Decoy compliant
- PTM site assignments are often ambiguous
  - AScore, Phospho-Loc. Score (PLS)
- Charged PTMs are typically not considered (e.g., phosphopantetheinyl)
- Glycosylation, SUMOylation and Ubiquitination (chains) require special approaches

Center for Computational Mass Spectrometry



Compute-intensive discovery proteomics at the click of a button



Latest Releases

MS-GeneratingFunction
2008.09.04

PepNovo
2008.07.08

MS-Clustering
2008.06.09

Inspect, MS-Alignment
2008.04.04

MS-Dictionary
2007.11.30

Spectral Networks
Sept 2007

Media Coverage

A powerful tool for PTM discovery (Jan 2008, Journal of Proteome research, Vol 7, Issue 1)

National Center for Research Resources

http://proteomics.ucsd.edu

Center for Computational Mass Spectrometry

Acknowledgements

CCMS, UCSD

Pavel Pevzner
Vineet Bafna
Dumitru Brinza
Yuval Inbar
Ari Frank
Stephen Tanner
Banu Dost
Nitin Gupta
Sangtae Kim
Roy Liu
Julio Ng

UCSD

Steven Briggs
Pieter Dorrestein
Vivian Hook
Susan Taylor
Virgil Woods
Huilin Zhou
Steven Bark
Zhouxin Shen
Wei-Ting Liu
Dario Meluzzi
Josué Pérez
Jian Wang

Broad Institute MIT/Harvard

Karl Clauser

Max Planck Institute, Germany

Matthias Mann
Jesper Olsen

OHSU

Larry David
Phil Wilmarth

Pacific Northwest National Lab

Richard Smith
Joshua Adlkins
Gordon Anderson
Kim Hixson
Mary Lipton

Indiana University

Haixu Tang

Genentech

Jennie Lill
David Arnott
Victoria Pham
Wendy Sandoval

UCLA

Parag Mallick
Roland Luethy

NIH NIAID/NIMH

Jon Yewdell
Sanford Markey

Utrecht University, Netherlands

Albert Heck

U Seoul, Korea

Eunok Paek
Seungjin Na

U Victoria, Canada

Cristoph Borchers

Inst. Ricardo Jorge, Portugal

Deborah Penque
Nuno Charro

Merck
Arana Therapeutics
Agilent

National Center for Research Resources

Calit2

//proteomics.ucsd.edu

Center for Computational Mass Spectrometry