# Bioinformatics for Top-Down Proteomics

DAVID L. TABB, PH.D.



with help from Richard LeDuc, Northwestern Proteomics

# Overview

- Motivation: the *proteoform*
- Separations and fragmentation
- Identification in ProSight PTM
- Identification by spectral alignment
- Consortium and standardization

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Humans have how many genes, transcripts, and proteins?

GENCODE 34, released April 2020:

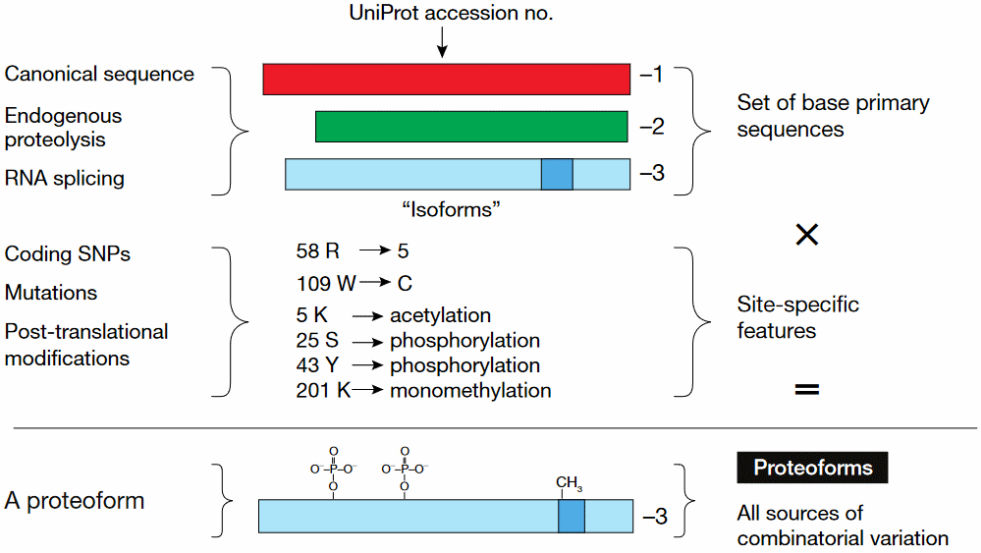
- Total number of genes: 60,669
- Protein-coding genes: 19,959
  *(See also pseudogenes and lncRNAs)*
- Protein-coding transcripts: 84,068
- Total # of distinct translations: 62,269
- Genes with multiple distinct translations: 13,717
  *(69% of all protein-coding genes)*

J. Harrow et al. *Genome Res*. (2012) 22: 1760-1774.

# A gene yields mRNA isoforms that yield multiple proteoforms

***Proteoforms*** differ through three primary factors:

- Genetic variation and RNA splicing give different mRNA sequences.
- Proteolysis and nonsense mutations truncate mature proteins.
- Post-translational modification may dramatically alter activity.




*Phenotype may be specific to a particular proteoform.*

*DNA sequencing is blind to post-translational modifications.*

LM Smith, NL Kelleher. *Nature Methods* (2013) 10: 186-187.

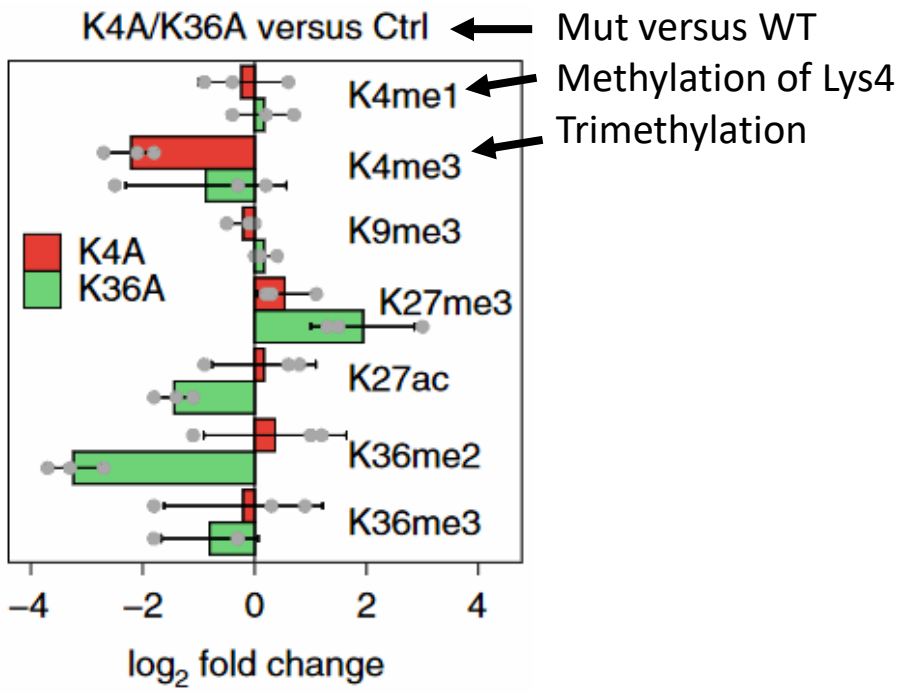
# Known blind spots for shotgun proteomics

## WHICH VEGF ISOFORM RISES IN CANCER COHORT?

>sp|P15692-2|VEGFA_HUMAN Isoform VEGF189
MNFLLSWVHWSLALLLYLHHAKWSQAAPMAEGGGQNHHE
VVKFMDVYQRSYCHPIETLVDIFQEYPDEIEYIFKPSCVP
LMRCGGCCNDEGLECVPTEESNITMQIMRIKPHQGQHIGE
MSFLQHNKCECRPKKDRARQEKKSVRGKGKGQKRKRKKSR
YKSWSVPCGPCSERRKHLFVQDPQTCKCSCKNTDSRCKAR
QLELNERTCRCDKPRR

>sp|P15692|VEGFA_HUMAN Isoform VEGF206
AMNFLLSWVHWSLALLLYLHHAKWSQAAPMAEGGGQNHHE
VVKFMDVYQRSYCHPIETLVDIFQEYPDEIEYIFKPSCVP
LMRCGGCCNDEGLECVPTEESNITMQIMRIKPHQGQHIGE
MSFLQHNKCECRPKKDRARQEKKSVRGKGKGQKRKRKKSR
YKSWSVYVGARCCLMPWSLPGPHPCGPCSERRKHLFVQDP
QTCKCSCKNTDSRCKARQLELNERTCRCDKPRR

## WHICH OF THESE PTMS CO-OCCUR ON HISTONE H3.3?






M. Gehre et al. *Nature Genet*. (2020) 52: 273-282.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Analytical chemistry for intact proteins is more challenging.

"Numerically, the intact proteome appears to be a much simpler mixture than its corresponding peptide digests. In practice, however, protein-level fractionation and separation are daunting tasks due to the diverse physicochemical properties (e.g., size, charge, and hydrophobicity) and the wide dynamic range of the proteome."



B. Chen et al. *Anal. Chem*. (2018) 90: 110-127.

>sp|P01308|INS_HUMAN Insulin (11,981 Da)
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha (15,258 Da)
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR

>sp|P07477|TRY1_HUMAN Trypsin-1 (26,558 Da)
MNPLLILTFVAAALAAPFDDDDKIVGGYNCEENSVPYQVSLNSGYHFCGGSLINEQWVVS
AGHCYKSRIQVRLGEHNIEVLEGNEQFINAAKIIRHPQYDRKTLNNDIMLIKLSSRAVIN
ARVSTISLPTAPPATGTKCLISGWGNTASSGADYPDELQCLDAPVLSQAKCEASYPGKIT
SNMFCVGFLEGGKDSCQGDSGGPVVCNGQLQGVVSWGDGCAQKNKPGVYTKVYNYVKWIK
NTIAANS

>sp|P60709|ACTB_HUMAN Actin, cytoplasmic 1 (41,737 Da)
MDDDIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQS
KRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKMT
QIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDL
AGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALDFEQEMATAASSSSLEKSY
ELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVLS
GGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQ
EYDESGPSIVHRKCF

>sp|Q14533|KRT81_HUMAN Keratin, type II cuticular Hb1 (54,928 Da)
MTCGSGFGGRAFSCISACGPRPGRCCITAAPYRGISCYRGLTGGFGSHSVCGGFRAGSCG
RSFGYRSGGVCGPSPPCITTVSVNESLLTPLNLEIDPNAQCVKQEEKEQIKSLNSRFAAF
IDKVRFLEQQNKLLETKLQFYQNRECCQSNLEPLFEGYIETLRREAECVEADSGRLASEL
NHVQEVLEGYKKKYEEEVSLRATAENEFVALKKDVDCAYLRKSDLEANVEALIQEIDFLR
RLYEEEILILQSHISDTSVVVKLDNSRDLNMDCIIAEIKAQYDDIVTRSRAEAESWYRSK
CEEMKATVIRHGETLRRTKEEINELNRMIQRLTAEVENAKCQNSKLEAAVAQSEQQGEAA
LSDARCKLAELEGALQKAKQDMACLIREYQEVMNSKLGLDIEIATYRRLLEGEEQRLCEG
IGAVNVCVSSSRGGVVCGDLCVSGSRPVTGSVCSAPCNGNVAVSTGLCAPCGQLNTTCGG
GSCGVGSCGISSLGVGSCGSSCRKC

>sp|P02768|ALBU_HUMAN Serum albumin (69,367)
MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPF
EDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEP
ERNECFLQHKDDNPNLPRLVRPEVDVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLF
FAKRYKAAFTECCQAADKAACLLPKLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAV
ARLSQRFPKAEFAEVSKLVTDLTKVHTECCHGDLLECADDRADLAKYICENQDSISSKLK
ECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYAR
RHPDYSVVLLLRLAKTYETTLEKCCAAADPHECYAKVFDEFKPLVEEPQNLIKQNCELFE
QLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVV
LNQLCVLHEKTPVSDRVTKCCTESLVNRRPCFSALEVDETYVPKEFNAETFTFHADICTL
SEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLV
AASQAALGL

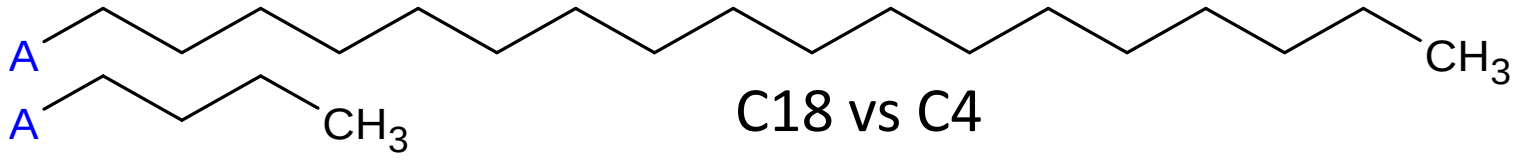
>sp|P02452|CO1A1_HUMAN Collagen alpha-1(I) chain (138,941 Da)
MFSFVDLRLLLLLAATALLTHGQEEGQVEGQDEDIPPITCVQNGLRYHDRDVWKPEPCRI
CVCDNGKVLCDDVICDETKNCPGAEVPEGECCPVCPDGSESPTDQETTGVEGPKGDTGPR
GPRGPAGPPGRDGIPGQPGLPGPPGPPGPPGPPGLGGNFAPQLSYGYDEKSTGGISVPGP
MGPSGPRGLPGPPGAPGPQGFQGPPGEPGEPGASGPMGPRGPPGPPGKNGDDGEAGKPGR
PGERGPPGPQGARGLPGTAGLPGMKGHRGFSGLDGAKGDAGPAGPKGEPGSPGENGAPGQ
MGPRGLPGERGRPGAPGPAGARGNDGATGAAGPPGPTGPAGPPGFPGAVGAKGEAGPQGP
RGSEGPQGVRGEPGPPGPAGAAGPAGNPGADGQPGAKGANGAPGIAGAPGFPGARGPSGP
QGPGGPPGPKGNSGEPGAPGSKGDTGAKGEPGPVGVQGPPGPAGEEGKRGARGEPGPTGL
PGPPGERGGPGSRGFPGADGVAGPKGPAGERGSPGPAGPKGSPGEAGRPGEAGLPGAKGL
TGSPGSPGPDGKTGPPGPAGQDGRPGPPGPPGARGQAGVMGFPGPKGAAGEPGKAGERGV
PGPPGAVGPAGKDGEAGAQGPPGPAGPAGERGEQGPAGSPGFQGLPGPAGPPGEAGKPGE
QGVPGDLGAPGPSGARGERGFPGERGVQGPPGPAGPRGANGAPGNDGAKGDAGAPGAPGS
QGAPGLQGMPGERGAAGLPGPKGDRGDAGPKGADGSPGKDGVRGLTGPIGPPGPAGAPGD
KGESGPSGPAGPTGARGAPGDRGEPGPPGPAGFAGPPGADGQPGAKGEPGDAGAKGDAGP
PGPAGPAGPPGPIGNVGAPGAKGARGSAGPPGATGFPGAAGRVGPPGPSGNAGPPGPPGP
AGKEGGKGPRGETGPAGRPGEVGPPGPPGPAGEKGSPGADGPAGAPGTPGPQGIAGQRGV
VGLPGQRGERGFPGLPGPSGEPGKQGPSGASGERGPPGPMGPPGLAGPPGESGREGAPGA
EGSPGRDGSPGAKGDRGETGPAGPPGAPGAPGAPGPVGPAGKSGDRGETGPAGPTGPVGP
VGARGPAGPQGPRGDKGETGEQGDRGIKGHRGFSGLQGPPGPPGSPGEQGPSGASGPAGP
RGPPGSAGAPGKDGLNGLPGPIGPPGPRGRTGDAGPVGPPGPPGPPGPPGPPSAGFDFSF
LPQPPQEKAHDGGRYYRADDANVVRDRDLEVDTTLKSLSQQIENIRSPEGSRKNPARTCR
DLKMCHSDWKSGEYWIDPNQGCNLDAIKVFCNMETGETCVYPTQPSVAQKNWYISKNPKD
KRHVWFGESMTDGFQFEYGGQGSDPADVAIQLTFLRLMSTEASQNITYHCKNSVAYMDQQ
TGNLKKALLLQGSNEIEIRAEGNSRFTYSVTVDGCTSHTGAWGKTVIEYKTTKTSRLPII
DVAPLDVGAPDQEFGFDVGPVCFL

DOI /10.1186/1756-0500-5-85



# Proteins cover a wide range of lengths and sizes.

# Protein recovery from LC columns requires shorter chains.




C18 vs C4

- Peptides / proteins are lured away from beads through increasing hydrophobicity.
- Column length, pump pressure, and pore size are substantial factors for separation.

pressure drop → $P = \frac{\eta v L}{d_p^2}$ ← viscosity, velocity, and length

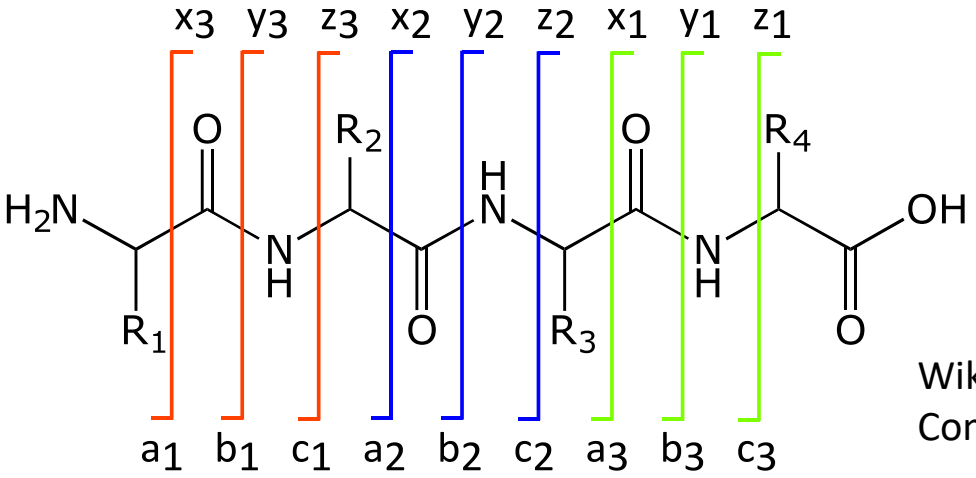
particle diameter →



Y. Shen et al. *J of Chromatography A* (2017) 1498: 99-110.

# Fragmenting peptides and proteins

- Collision-induced dissociation
  - Standard quadrupole technique
- Electron transfer dissociation
  - Ion-ion reaction for gentle bond cleavage



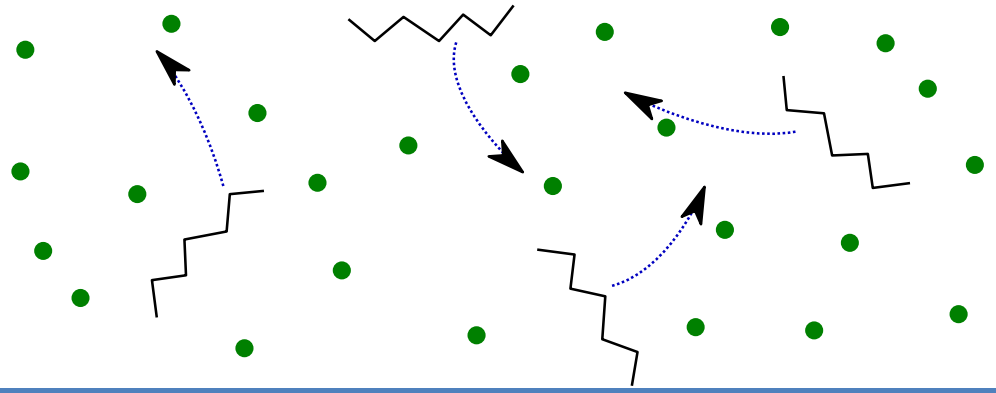

Wikimedia Commons

Roepstorff and Fohlman *Biomed. Mass Spectrom.* (1984) 11: 601.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Collision-Induced Dissociation (CID)

- When the quadrupole adds energy to ions, they collide more frequently with gas molecules, gaining energy.
- Protons become mobile, destabilizing peptide bonds (creating *b*-*y* fragments).

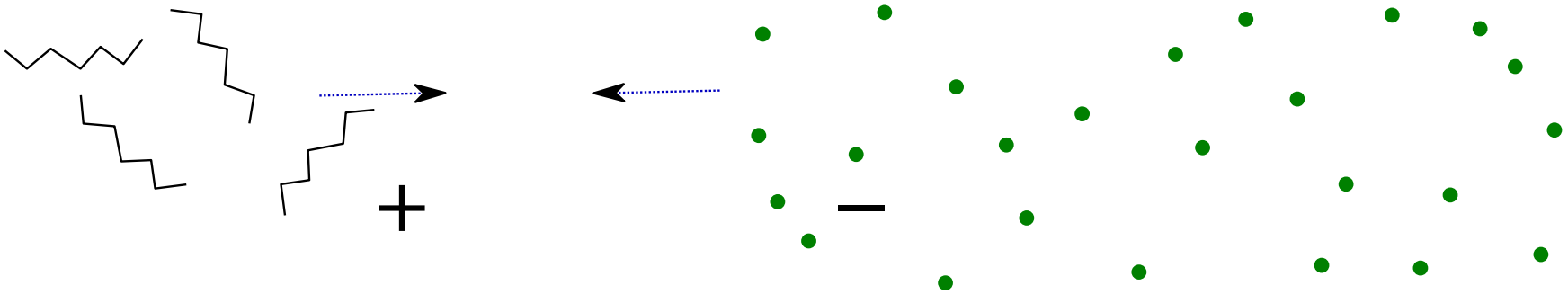


B. Paizs and S. Suhai. *Mass Spectrom. Rev*. (2005) 24: 508-548.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Electron Transfer Dissociation (ETD)

- Charge draws positively-charged proteins to accept electrons from radical anions.
- An amino acid backbone cleaves between nitrogen and alpha carbon (*c*-*z* fragments).




JJ Coon et al. *PNAS*. (2005) 102: 9463-9468.

# Radical chemistry in peptides

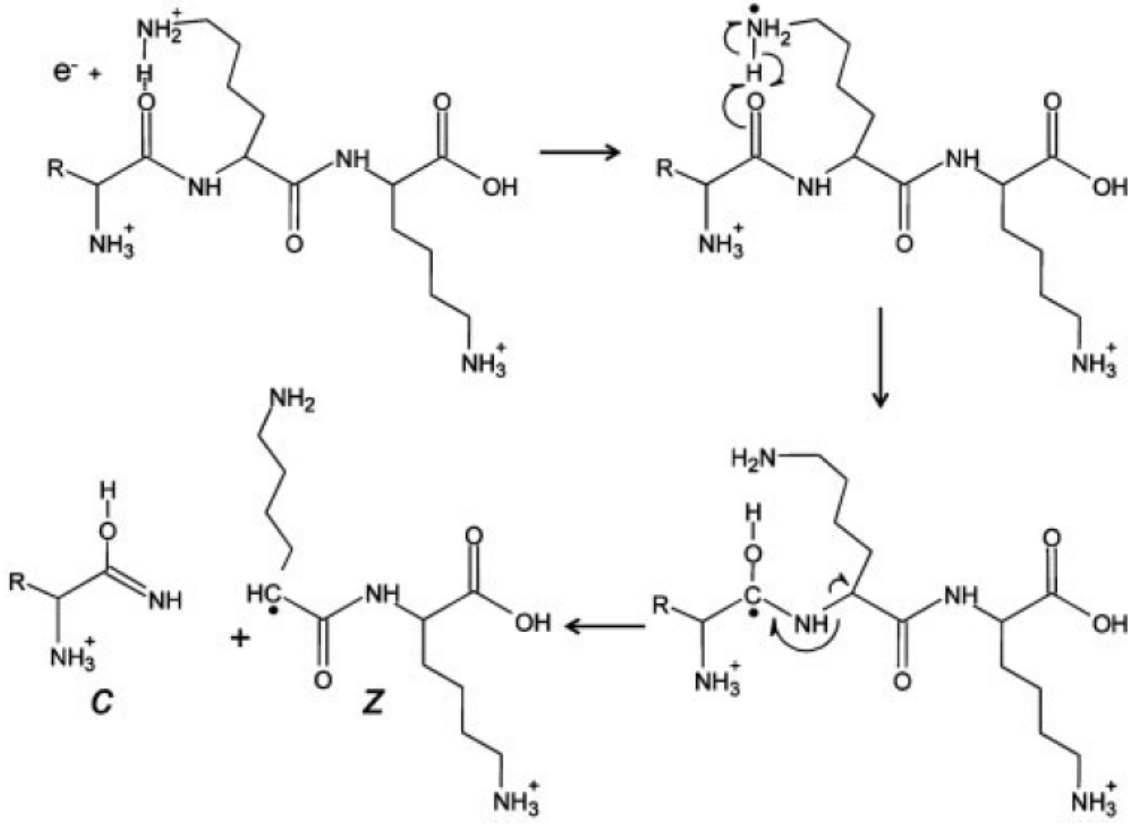


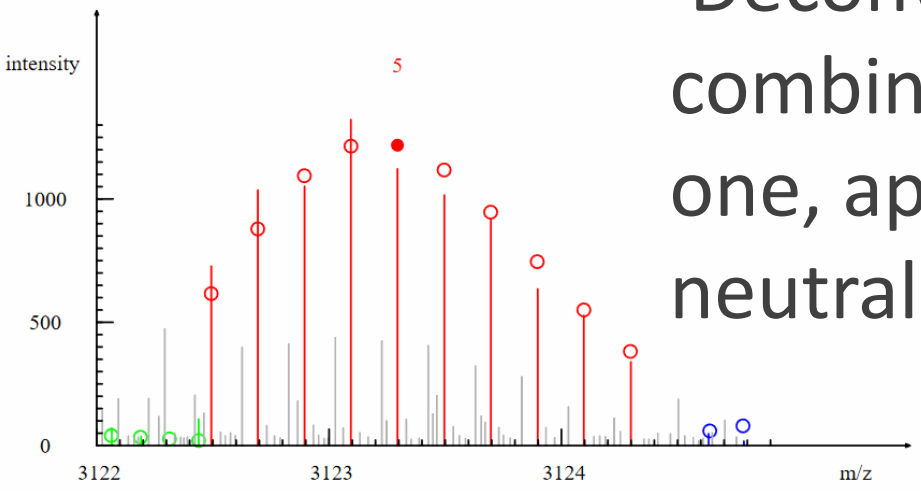

**Fig. 1.** Fragmentation scheme for production of c- and z-type ions after reaction of a low-energy electron with a multiply protonated peptide.

Syka et al (2004) *PNAS* 101: 9528-9533.

# Intermission

# Deconvolution goals

- A molecule appears in many isotopes and at many charges to produce isotopic *envelopes* in both MS and MS/MS scans.




- Deconvolution attempts to combine these peaks to only one, appearing at +1 or neutral monoisotopic mass.



X. Liu et al. *Mol. Cell. Proteomics*. (2010) 9: 2772-2782.

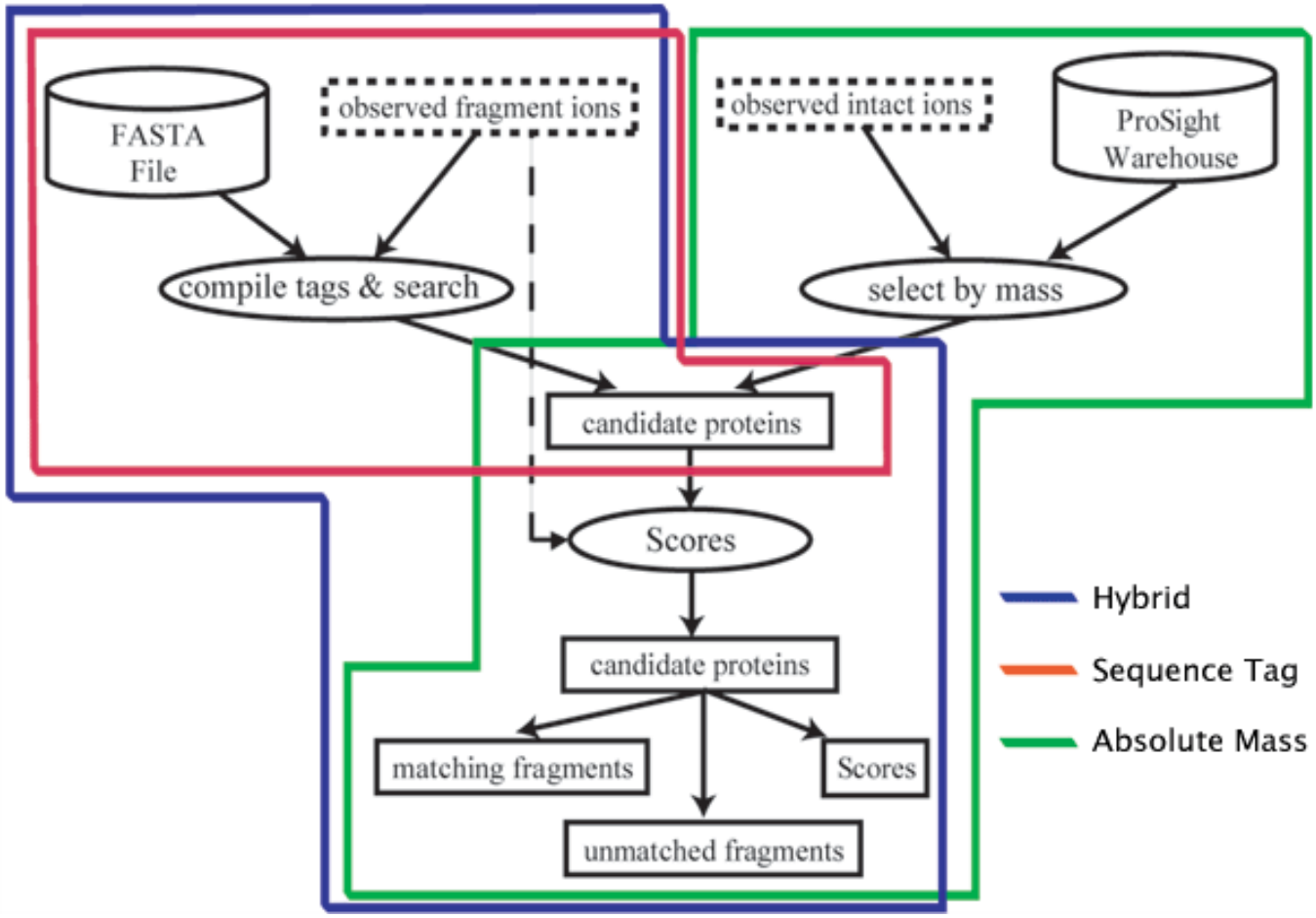
UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Limiting the sequence expansion for *PrSMs*

← Proteoform Spectrum Matches

- In shotgun PTM ID, we decorate peptide sequences with allowable mass shifts. Top-down benefits from known PTM annotation.
- Partial inferred sequence tags can narrow protein candidate list considerably.
- Signal peptides and other backbone cleavages take on special importance.

# ProSight PTM schematic






GK Taylor et al. *Anal. Chem*. (2003) 75: 4081-4086.

# Poisson scoring model

- $P = \frac{(xf)^n * e^{-xf}}{n!}$

Probability of random match

Number of matched ions

Number of predicted ions

See also OMSSA: Geer *J. Proteome Res*. (2004) 3: 958-964.

- $x = \frac{1}{111.1} * 2 * (M_a * 2)$

Mean prob. of random match

Freq-weighted avg AA mass

Fragments per bond

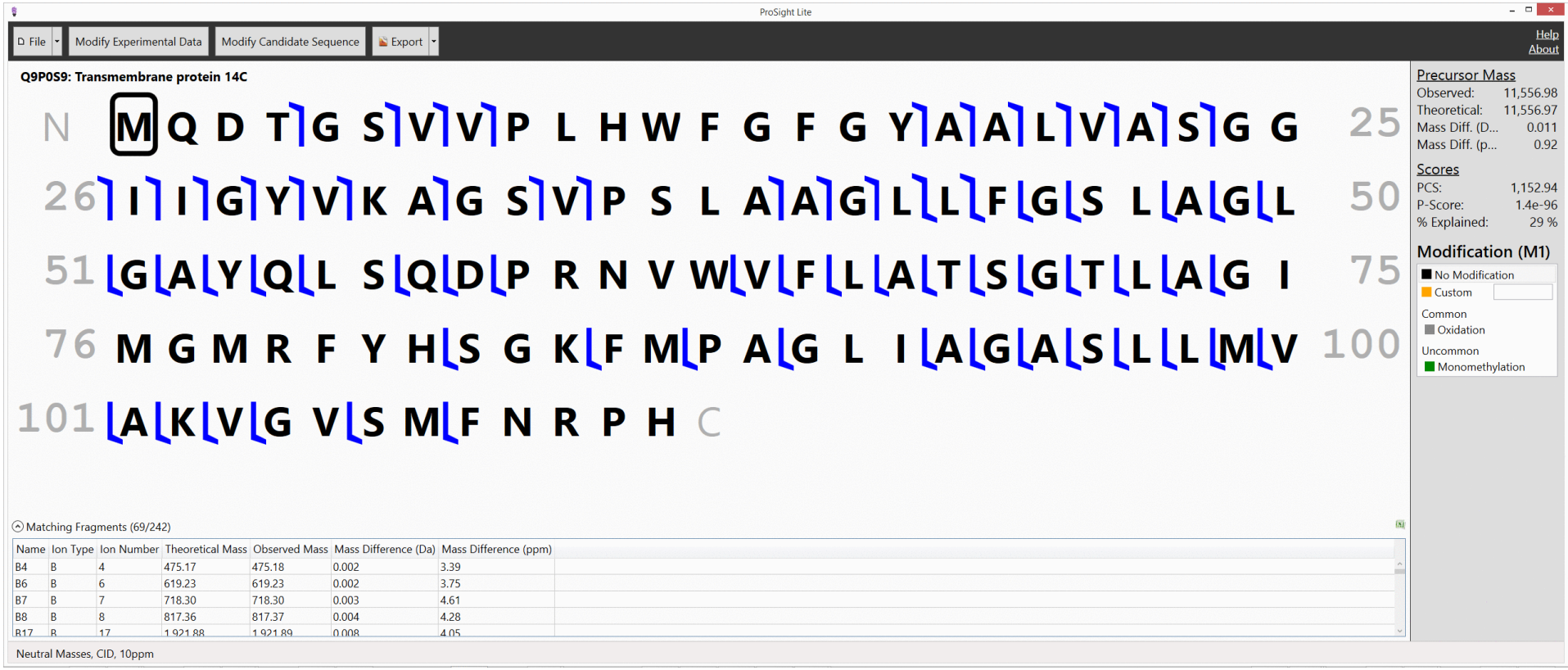
Mass tolerance



F Meng et al. *Nature Biotech*. (2001) 19:952-957.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Visualizing supporting fragment ions

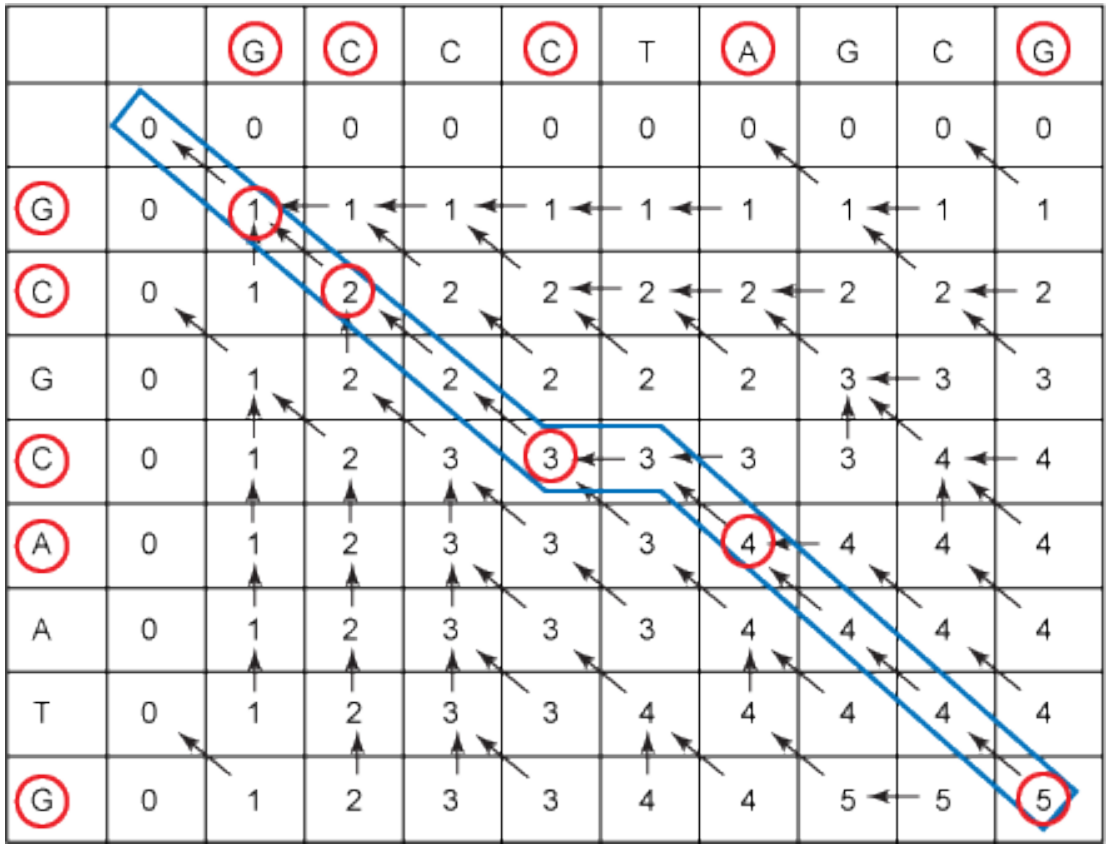





RT Fellers et al. *Proteomics* (2015) 15: 1235-1238

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Dynamic programming is for more than sequence alignment.



| | | G | C | C | C | T | A | G | C | G |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| C | 0 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| G | 0 | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 |
| C | 0 | 1 | 2 | 3 | 3 | 3 | 3 | 3 | 4 | 4 |
| A | 0 | 1 | 2 | 3 | 3 | 3 | 4 | 4 | 4 | 4 |
| A | 0 | 1 | 2 | 3 | 3 | 3 | 4 | 4 | 4 | 4 |
| T | 0 | 1 | 2 | 3 | 3 | 4 | 4 | 4 | 4 | 4 |
| G | 0 | 1 | 2 | 3 | 3 | 4 | 4 | 5 | 5 | 5 |

- In Smith-Waterman, we use gaps to represent INDEL differences between sequences.
- Approach can be adapted to many additive optimization problems in proteomics!

https://www.ibm.com/developerworks/library/j-seqalign/

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Dynamic programming in shotgun proteomics

- Infer sequences from MS/MS *de novo*
  - V Dancik et al. J. Computat. Bio. (1999) 6: 327.
- Align LC retention times of features
  - M Ono et al. *Mol. Cell. Proteomics* (2006) 5: 1338.
- Localize phosphorylations within peptide
  - F Saeed et al. *IEEE* (2012) 10.1109/BIBMW.2012.6470210
- Compute exact p-values for XCorrs of PSMs
  - JJ Howbert et al. *Mol. Cell. Proteomics* (2014) 13: 2467.

 Not comprehensive-- this list is intended to show diversity.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

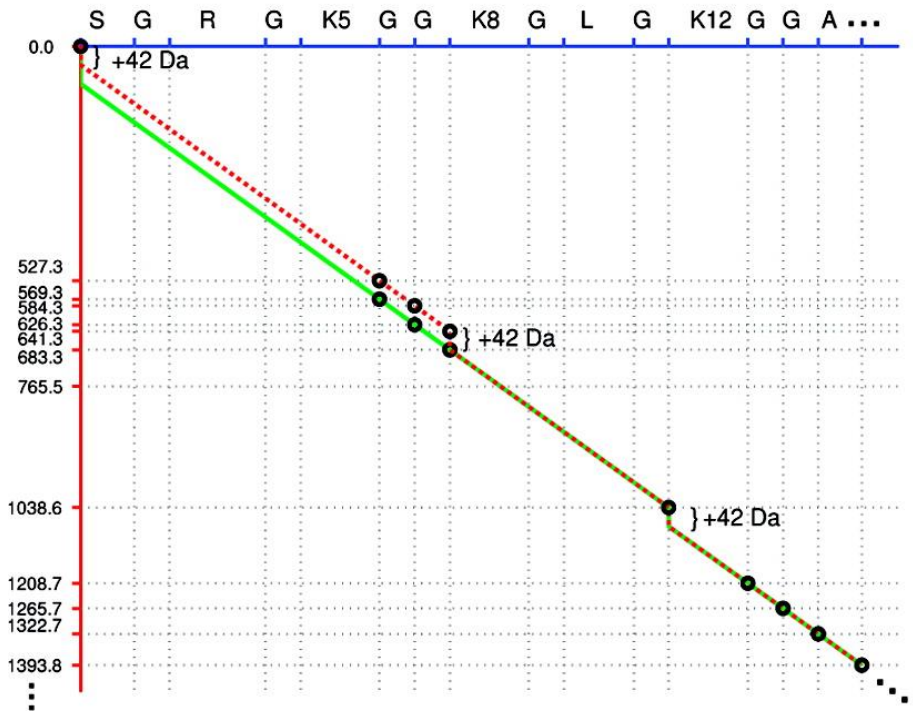
# Problems leveraging sequence and PTM composition

- "The candidate expansion method… leads to an exponential growth in the number of candidate protein forms that need to be considered."
- "top-down spectral alignment may deal with as many as 10-20 PTMs to a protein"
- "one often deals with multiple isobaric protein forms in the same spectrum"



A. Frank et al. *Anal. Chem*. (2008) 80: 2499-2505.

# MS-TopDown and histone H4




- Acetylation adds 42 Da to N-terminus and two other sites in first 15 amino acids.
- Lys5 and Lys8 ambiguity may result from co-fragmenting variants in one MS/MS.



A. Frank et al. *Anal. Chem*. (2008) 80: 2499-2505.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Inheritors of MS-Align+

MS-TopDown (UCSD 2008)

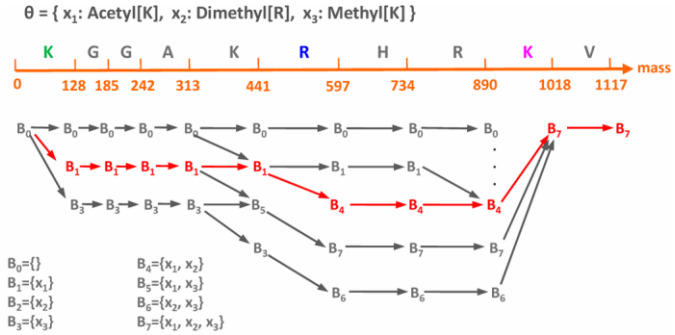
MS-Align+ (UCSD 2012)

TopPIC (IUPUI 2016)

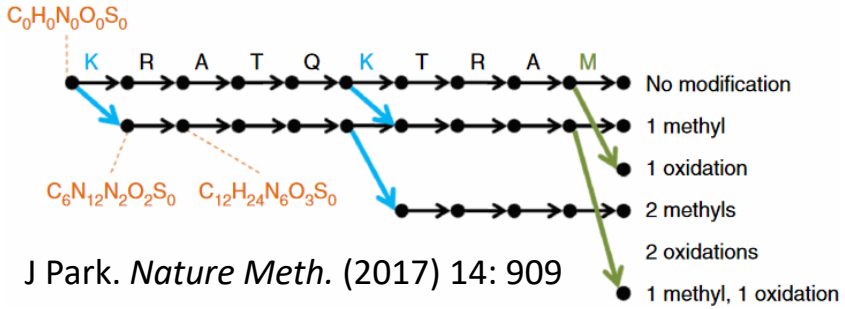
TopMG (IUPUI 2017)

MSPathFinder (PNNL 2017)

pTop (Beijing ICT 2016)




RX Sun. *Anal. Chem*. (2016) 88: 3082




J Park. *Nature Meth.* (2017) 14: 909

Solid lines reflect participation of former UCSD personnel.

# How do we communicate proteoforms?

## ProForma Proteoform Notation Rules

### The Basics

**1.** The amino acid sequence is written. Ambiguous amino acids can be specified.

```
SEQVENCE SEQXXNCE
```

**2.** Modifications and are written inside square brackets.

```
SEQVK[Unimod:Label:13C(3)][Acetyl]ENCE
```

**3.** Tags contain descriptors in key : value pairs.

```
SEQVEN[mass:+14.02]CE
```

**4.** Multiple descriptors are separated by pipes.

```
SEQVEN[mod:Methyl|mass:+14.02]CE
```

### Advanced Usage

**6.** Prefix tags define the key for all subsequent tags.

```
[RESID]+S[AA0037]EQVE[AA0234]NCE
[mass]+S[80]EQVE[14]NCE
[formula]+S[HPO(3)]EQVE[CH(2)]NCE
```

**7.** Terminal modifications are separated from the sequence by a dash.

```
[mass:-17.027]-QVENCE-[Amidation]
```

### The Specifics

**5a. Modification Name**

```
PRT[Phospho]EFRM
PRT[Phosphothreonine(UniProt)]EFRM
PRT[O-phospho-L-threonine(RESID)]EFRM
PRT[O-phospho-L-threonine(PSIMod)]EFRM
```

**5b. Database Accession**

```
PRT[Unimod:21]EFRM
PRT[UniProt:PTM-0254]EFRM
PRT[RESID:AA0038]EFRM
PRT[PSI-MOD:MOD:00047]EFRM
```

**5c. Mass**

```
SEQ[mass:+15.995]VENCE
SEQ[mass:+16]VENCE
SEQ[mass:16]VENCE
```

**5d. Chemical Formula**

```
SEQVEN[Methyl|formula:H(2)C]CE
```

**5e. Additional Information**

```
SEQ[info: unstructured text]VENCE
```



RD LeDuc et al. *J. Proteome Res*. (2018) 17: 1321-1325.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# ProForma in practice

## Examples of Best Practices

**i.** Histone H4 with several modifications. This example is human-readable and conforms to best practices.

```
[Acetyl]-S[Phospho|mass:79.966331]GRGK[Acetyl|Unimod:1|mass:42.010565]QGGKA
RAKAKTRSSRAGLQFPVGRVHRLLRKGNYAERVGAGAPVYLAAVLEYLTAEILELAGNAARDNKKTRIIPRHLQL
AIRNDEELNKLLGKVTIAQGGVLPNIQAVLLPKKT[Unimod:21]ESHHKAKGK
```

**ii.** This is a valid and compact way of specifying Unimod accessions in multiple locations in the sequence.

```
[Unimod]+[1]-S[21]GRGK[1]QGGKARAKAKTRSSRAGKVTIAQGGVLPNIQAVLLPKKT[21]ESHHKAKGK
```

**iii.** Extensive description of a modification using descriptors and IDs from different databases.

```
MTLFQLREHWFVYKDDEKLTAFRNK[p-adenosine| N6-(phospho-5'-adenosine)-L-lysine
(RESID)| RESID:AA0227| PSI-MOD:00232|N6AMPLys(PSI-MOD)]SMLFQRELRPNEEVTWK
```

**iv.** Unknown modifications are best described by their mass shift and marked as unknown.

```
MTLFQLDEKLTA[mass:-37.995001|info:unknown modification]FRNKSMLFQRELRPNEEVTWK
```

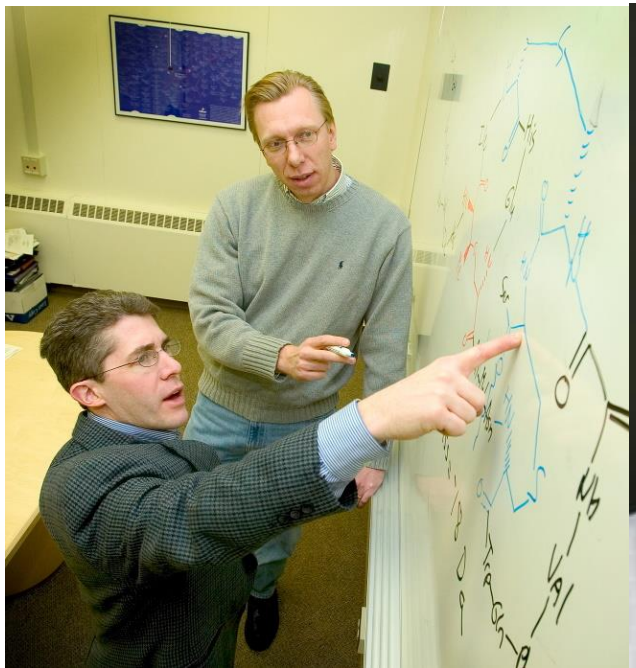
RD LeDuc et al. *J. Proteome Res*. (2018) 17: 1321-1325.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# People of interest

- Neil Kelleher
- Lloyd Smith
- Ying Ge
- Jeff Agar
- Ljiljana Pasa-Tolic
- Pavel Pevzner
- Xiaowen Liu
- Rui-Xiang Sun

CTDP
Consortium for Top-Down Proteomics
Bringing Proteoforms to Life℠



# Takeaway messages

- If we digest proteins to measure them, we lose peptide relationships.
- Top-down proteomics relies upon high-resolution MS/MS and good separations.
- The dominant ProSight PTM framework has competition from alignment-based software.
- The CTDP seeks to broaden the use of top-down tech throughout biomedical research.