Post-translational Modifications and Proteomics

Gary Van Domselaar

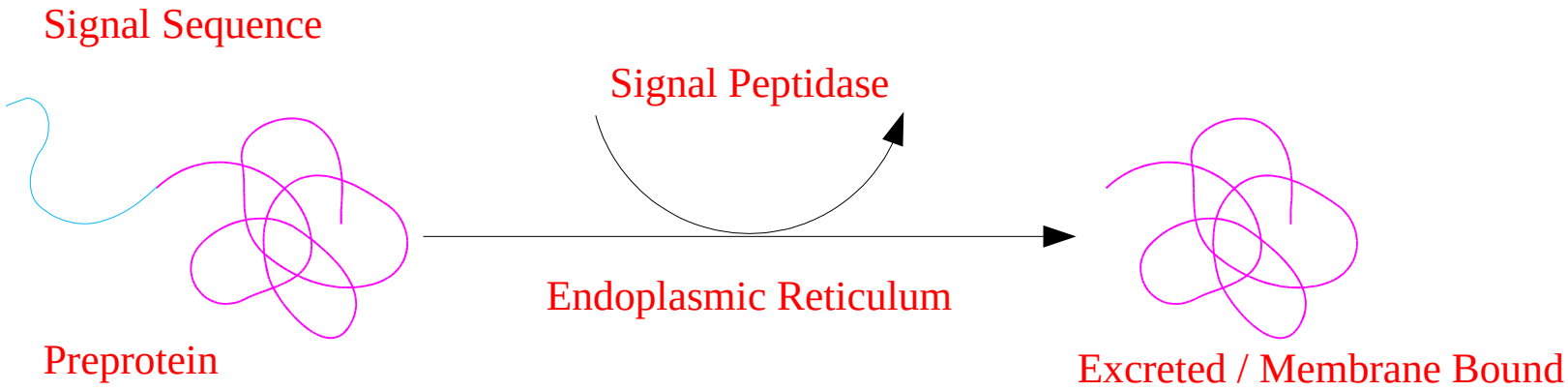
University of Alberta

Post-translational Modifications

- **Definition**
  - **Covalent modification of side chains in proteins after they are translated and exit from the ribosomes.**
- **General Purpose:**
  - **To alter the function, destination, or fate of the translated protein**
  - **To expand the functional group repertoire of proteins beyond the 20 'natural' amino acids.**

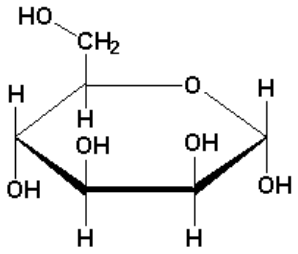
Post-translational Modifications

- **Proteolytic Cleavage**
  - **Conversion of an inactive protein precursor (*proprotein*) to the active form**
  - **Removal of a signal peptide from a *preprotein*.**
  - **Conversion of *zymogens* to active enzymes**
  - **Removal of initiating methionine**

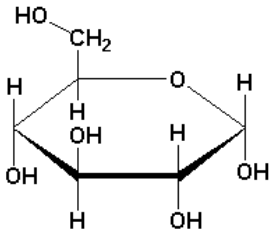


Post-translational Modifications

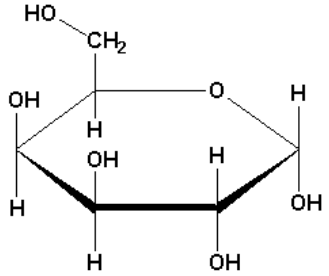
- **Glycoproteins**
  - **Proteins covalently linked to carbohydrate (Glucose, galactose, mannose, fucose, GalNAc, GlcNAc, NANA).**



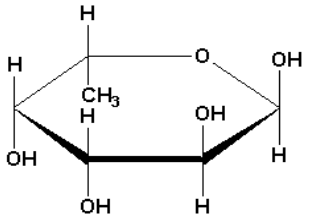
Mannose



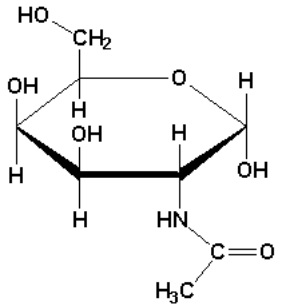
Glucose



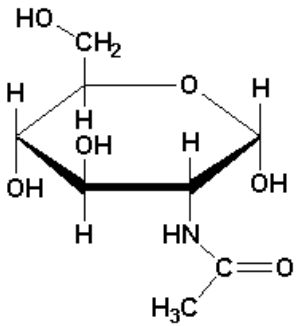
Galactose



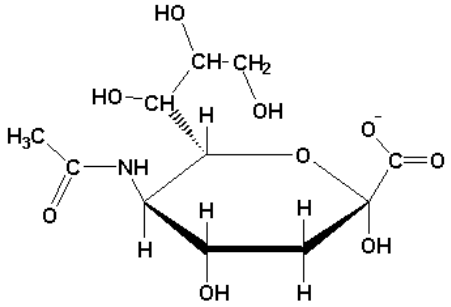
L-Fucose



N-Acetylgalactosamine



N-acetylglucosamine

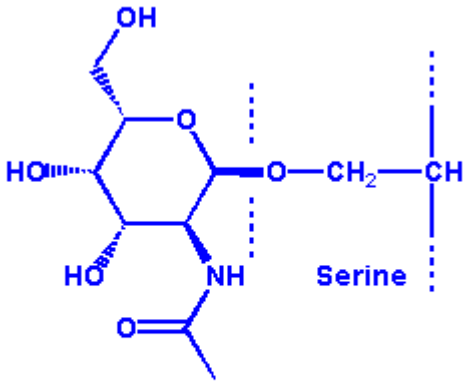


N-Acetyl-Neuraminic Acid

Post-translational Modifications

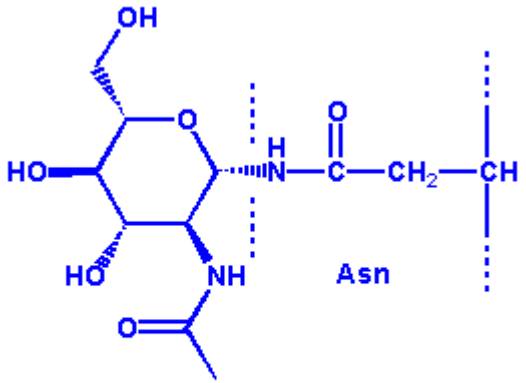
- **Glycosidic bonds**
  - **O-linked (Ser, Thr, Hydroxy-Lys),**
  - **N-linked (predominant).**

GalNAc linkage to Ser, Thr



O-linkage to GalNAc

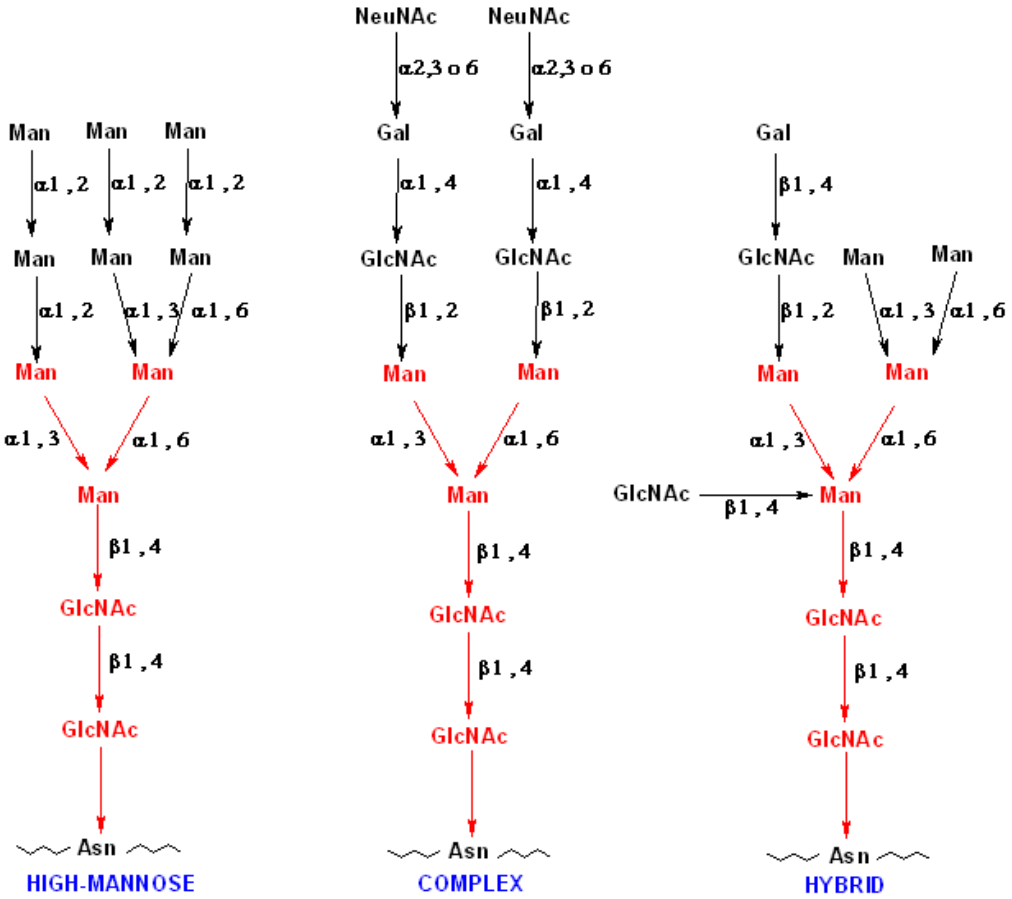
GlcNAc linkage to Asn



N-linkage to GlcNAc

Post-translational Modifications

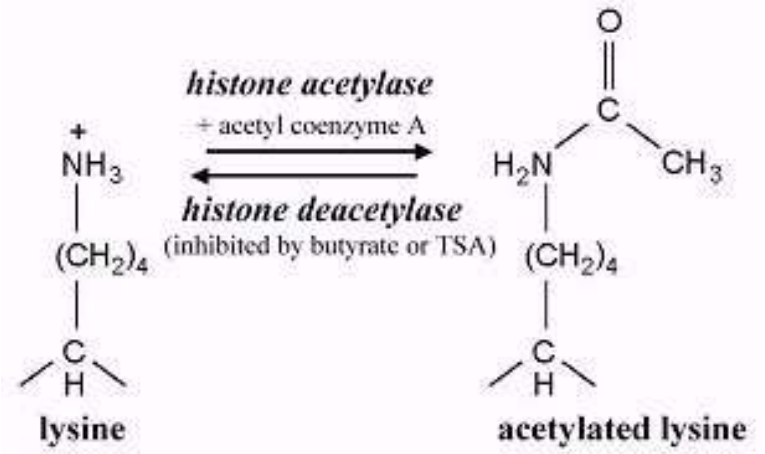
- **Oligosaccharides**

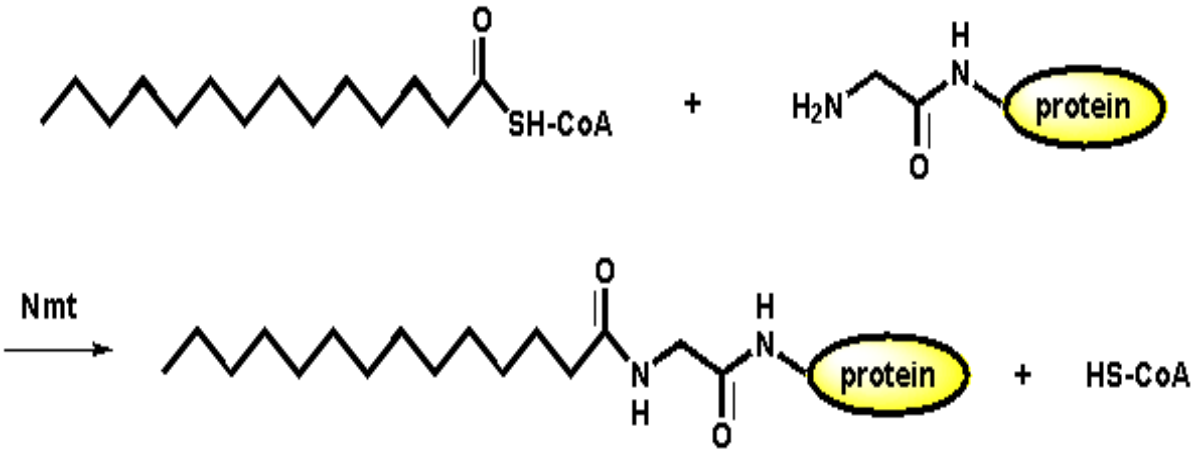


http://www.med.unibs.it/~marchesi/protmod.html

Post-translational Modifications

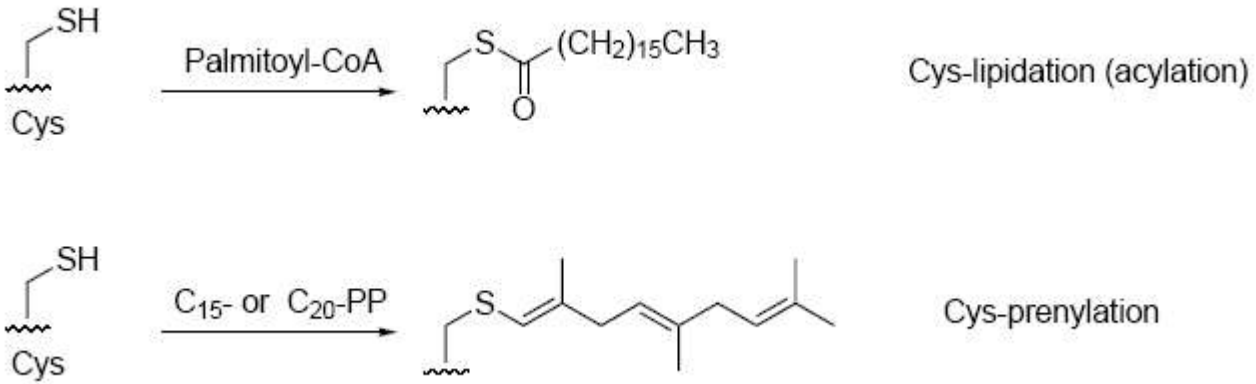
- **N-terminal and Lysine side-Chain Acylation**
  - **acetylation**
  - **myristoylation**
  - **palmitoylation**





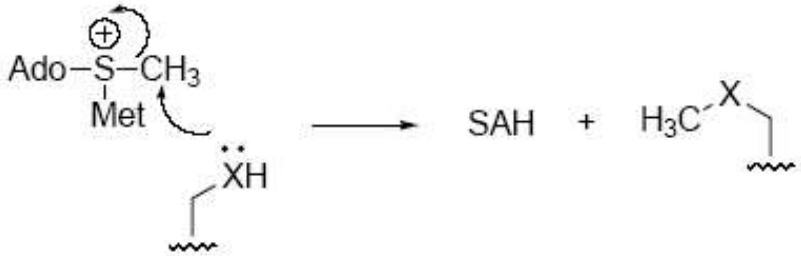
Post-translational Modifications

- **Cys Lipidation**
  - **Prenylation**
    - **Farnesylation**
    - **Geranylgeranylation**
  - **Palmitoylation**

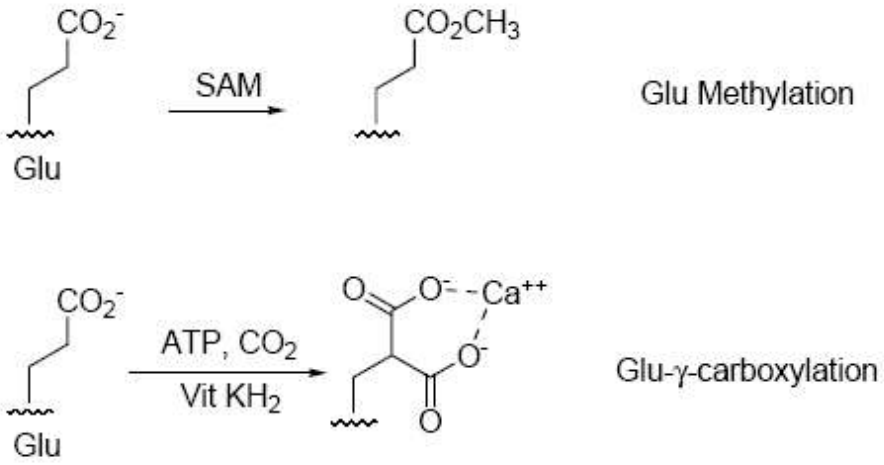


Post-translational Modifications

- **Methylation**
  - **Lysine**
  - **Arginine**
  - **C- terminal**

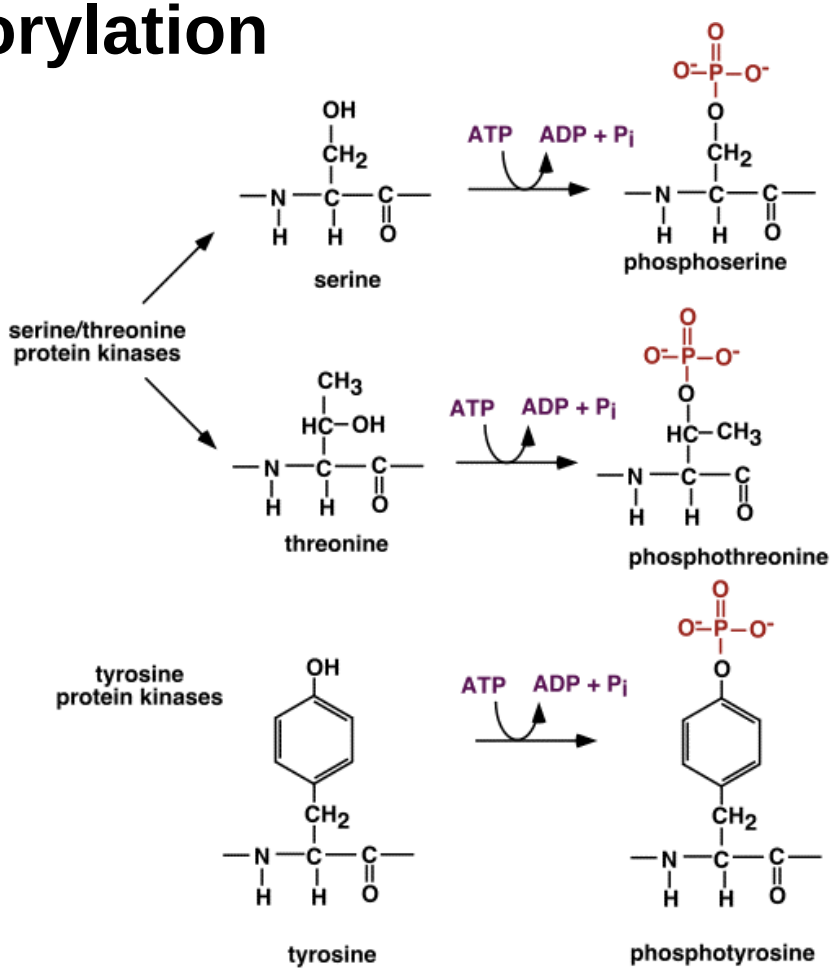


- **Glu Methylation and Carboxylation**



Post-translational Modifications

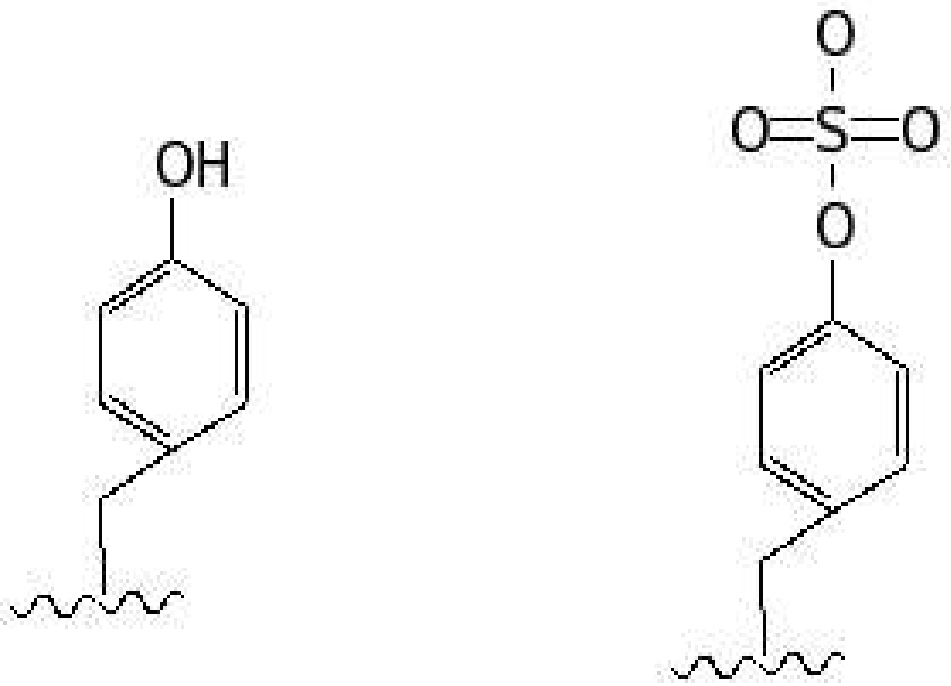
- **Phosphorylation**



http://anx12.bio.uci.edu/~hudel/bs99a/lecture26/lecture7_3.html

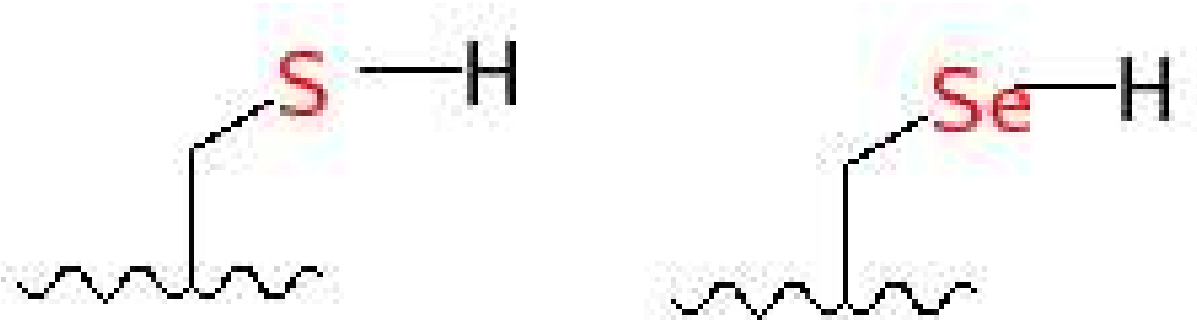
Post-translational Modifications

- **Sulfation**



Post-translational Modifications

- **Selenocysteine**



Post-translational Modifications

Carbohydrates
N-linked oligosaccharides
O-linked oligosaccharides
Glycophosphoinositol-lipid anchors
C-Mannosylation

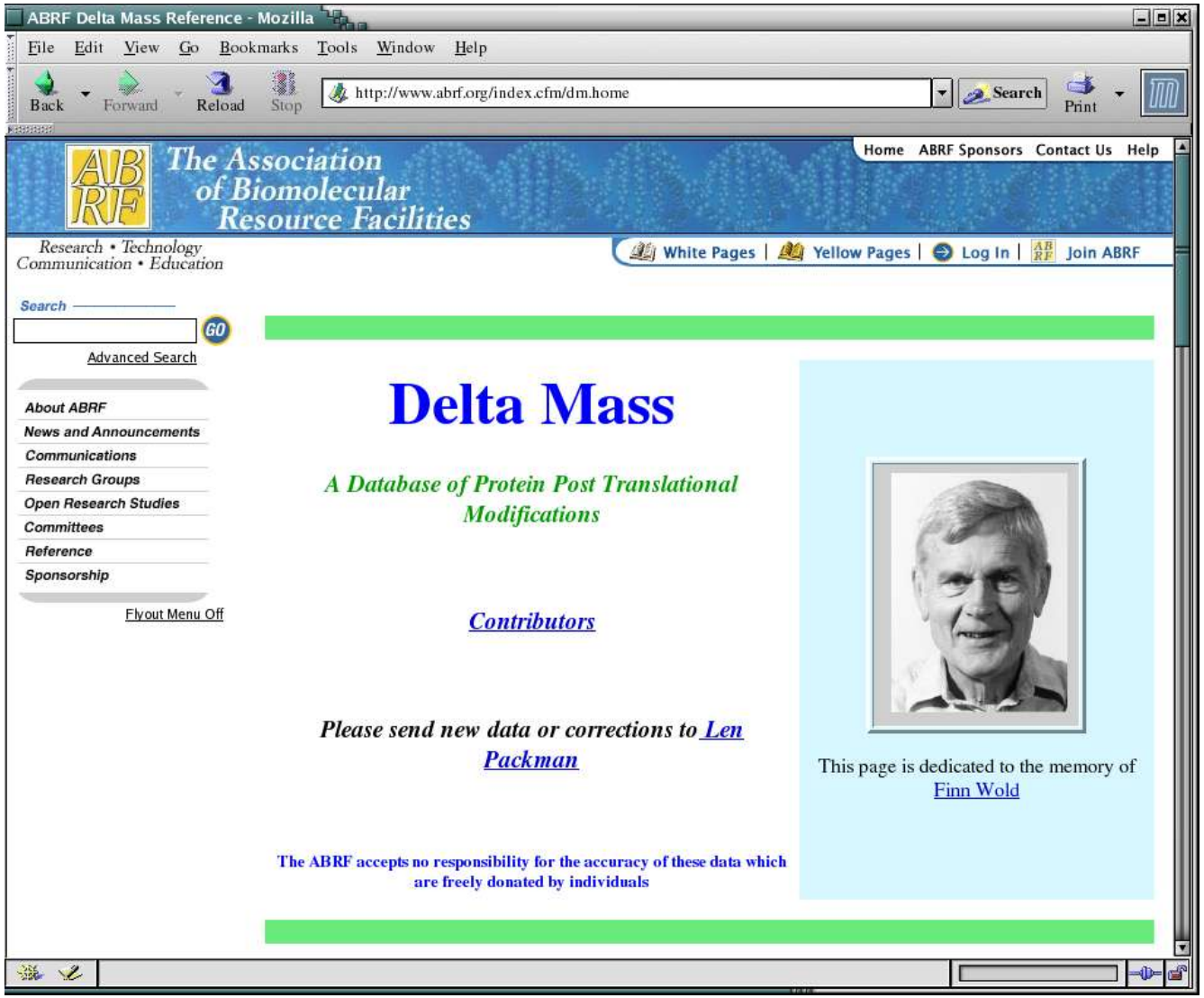
Lipids
Acylation
Palmitylation
Myristylation
Farnesylation
Prenylation
Geranylgeranylation

Cross-linking
Disulfide formation
Thiolester formation
Transglutamylation

Amino acid modification
Amidation–deamidation
Cysteinylation
Formylation
Citrullination
Isoaspartylation
Detyrosination
Thioether addition
Gamma-carboxylation
Deimidation

Small adducts
Phosphorylation
Sulfation
Nitration
Acetylation
Formylation
Methylation
Oxidation
Hydroxylation
Iodination
Biotinylation

Post-translational Modifications

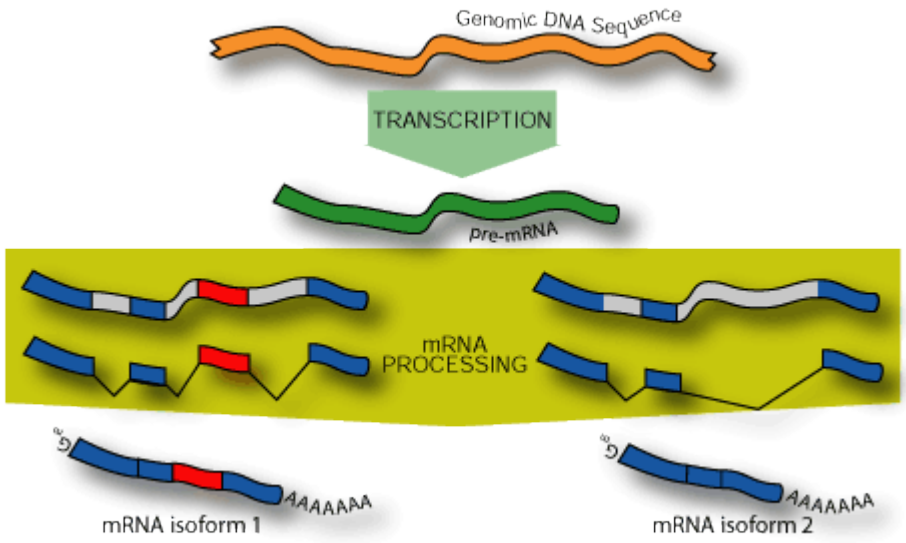


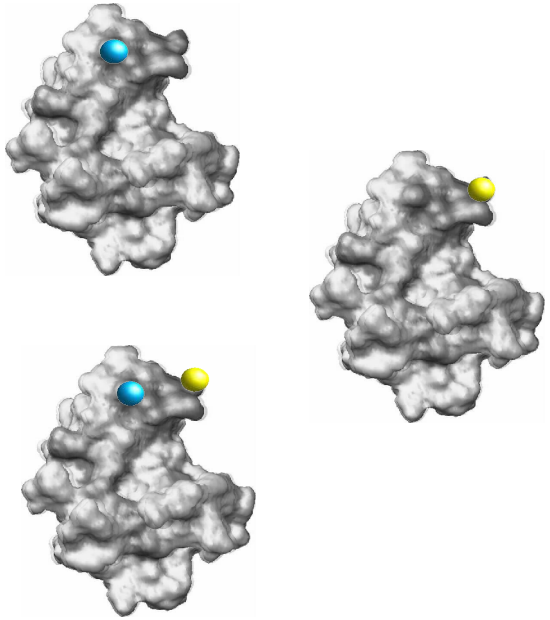
Post-translational Modifications

- **Many PTMs have been discovered serendipitously during studies of individual proteins and biological processes.**
- **Direct PTM analysis provides a complete analysis, but is slow and sometimes not even possible for certain proteins because of problems of isolation and obtaining sufficient amounts.**
- **Proteomics analysis is high throughput, but typically incomplete in its ability to completely characterize the protein primary sequence**
- **How can proteomics be applied to the analysis of PTMS?**

Isolation of Modified Proteins

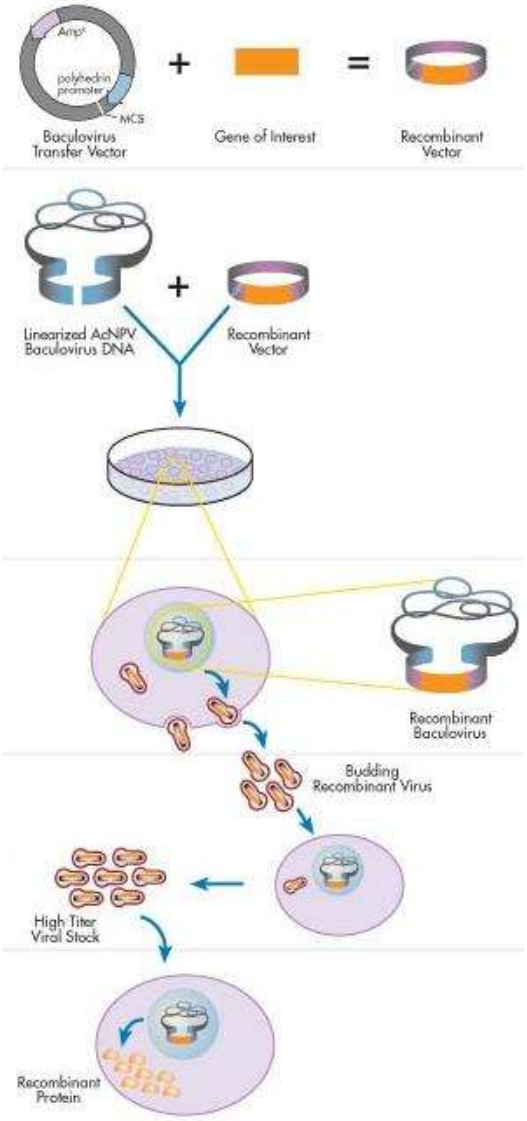
- **Consideration: A single gene can have many gene products due to alternative splicing and multiple modification combinations.**
- **A lot of protein is necessary (~1ug).**





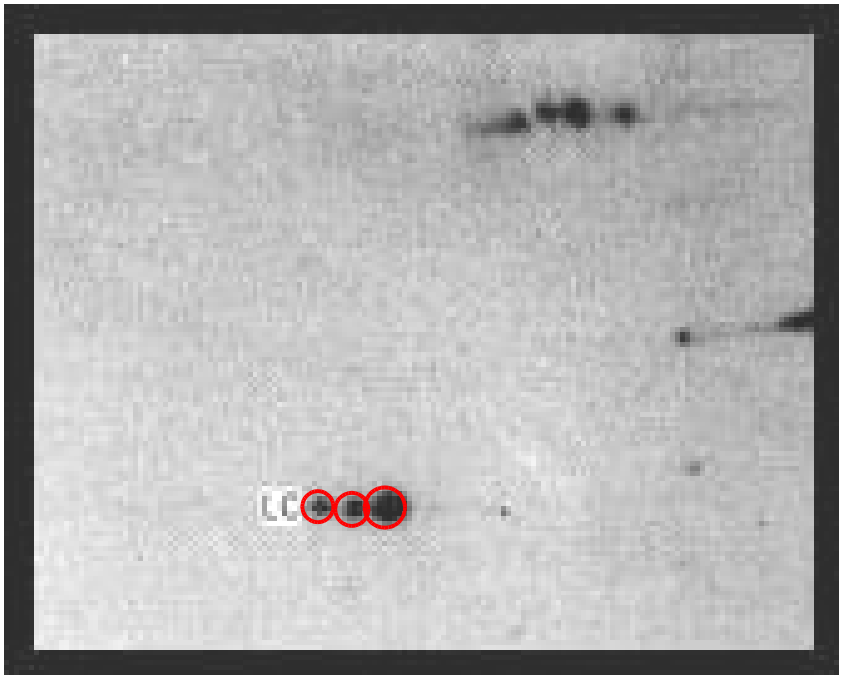
Isolation of Modified Proteins

- **A suitable expression (eg baculovirus) can sometimes be employed to recombinantly express the protein(s) to be studied.**
- **Caveat: recombinant systems do not always produce identical PTMs to the those of mammalian cells.**

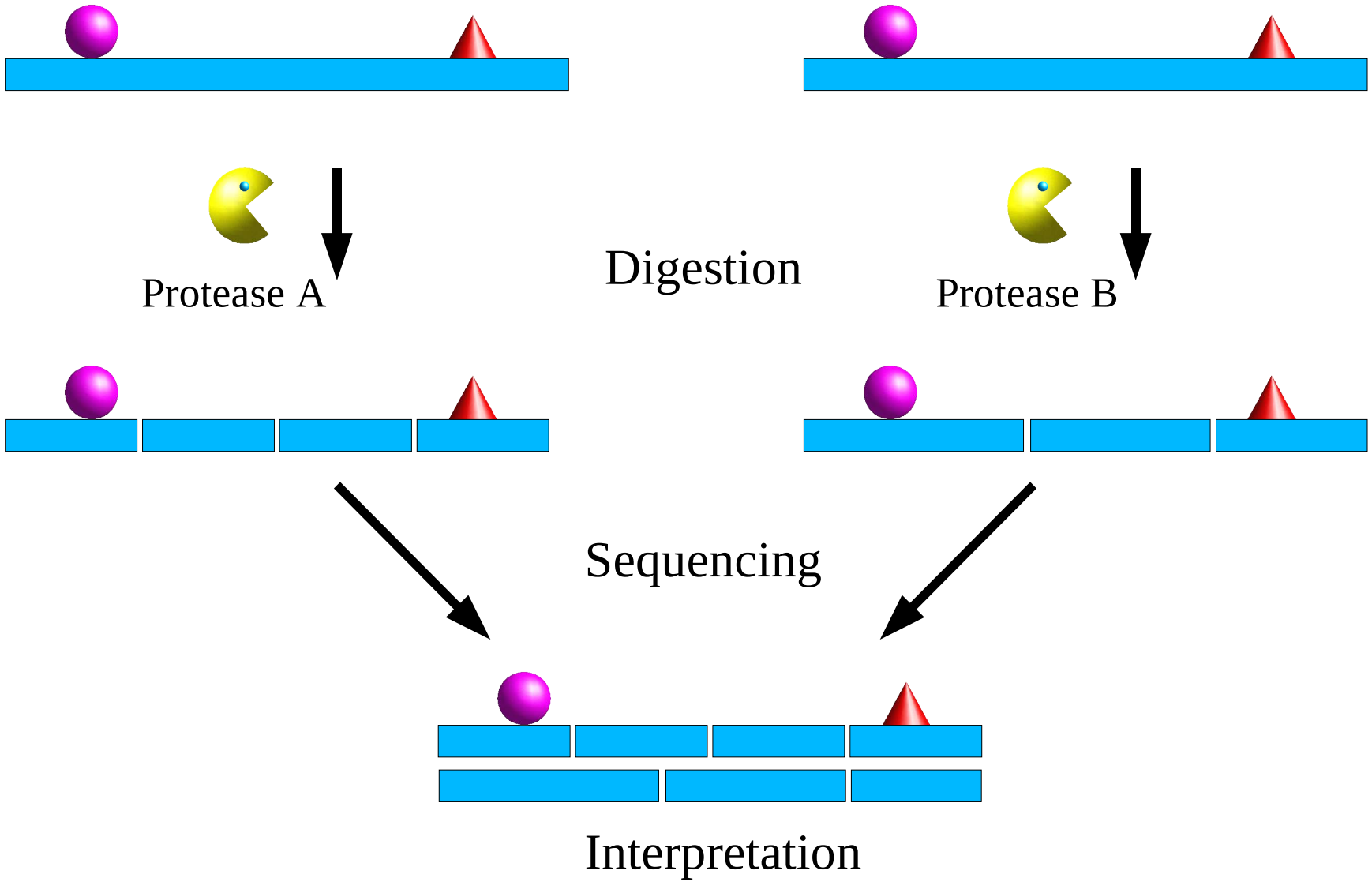


Isolation of Modified Proteins

- **2D Gel electrophoresis typically can be used to separate modified proteins.**

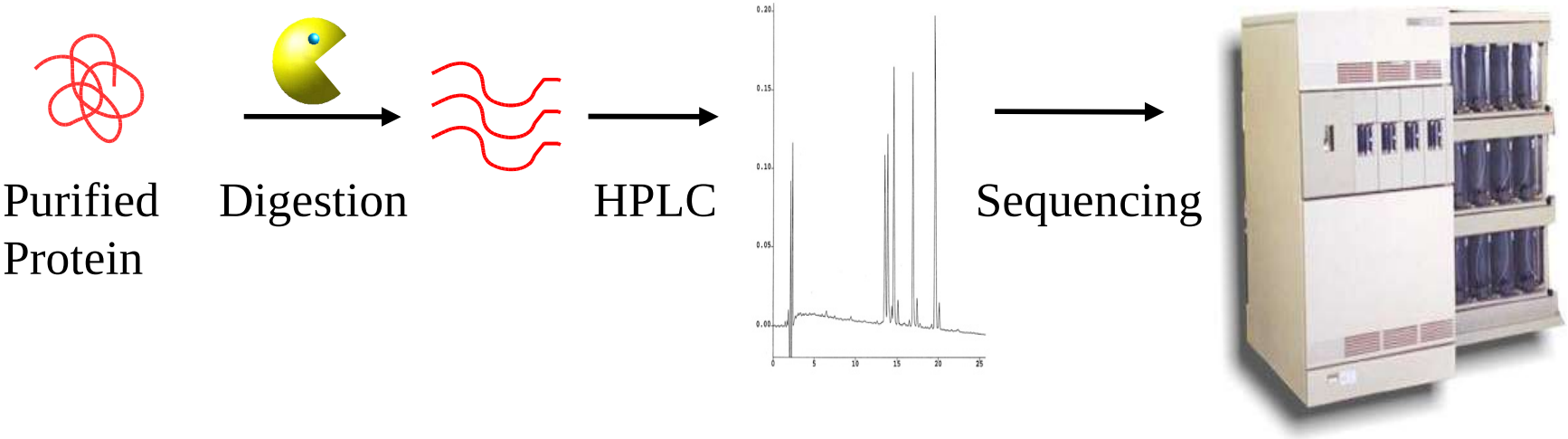


PTM Mapping a Purified Protein



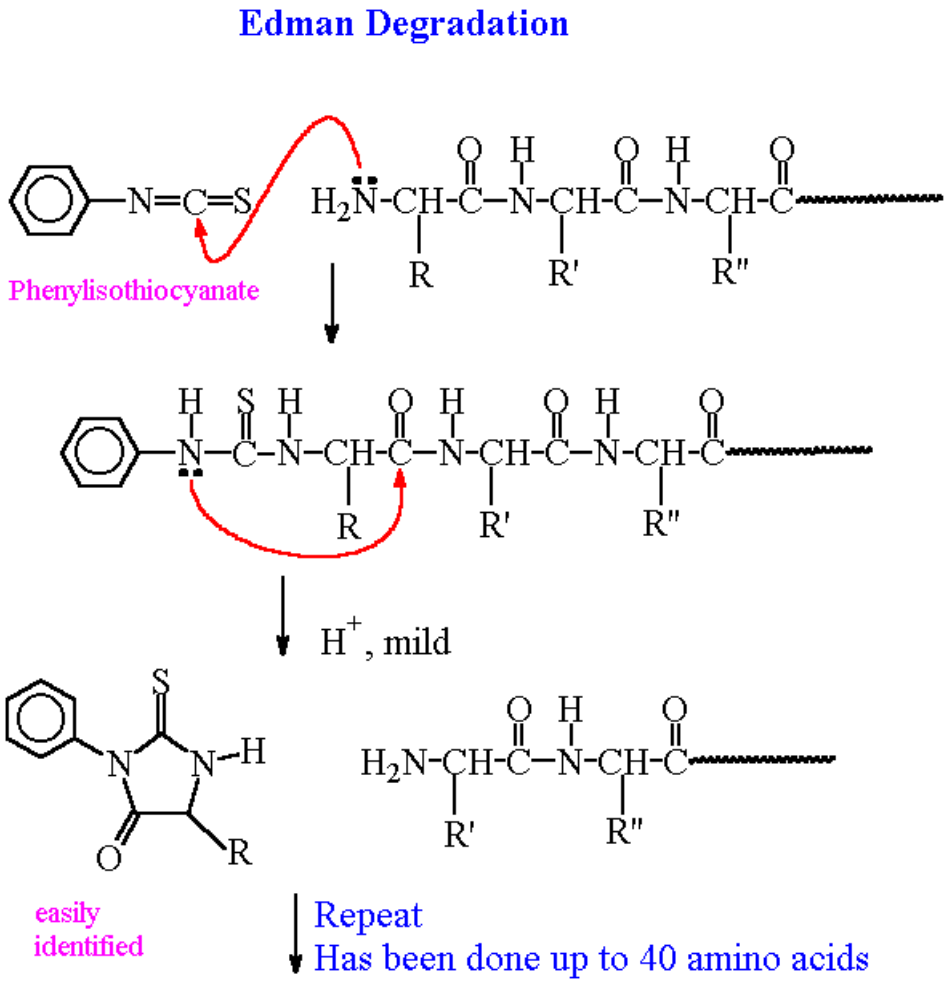
PTM Mapping a Purified Protein

- **Sequencing: Edman degradation**



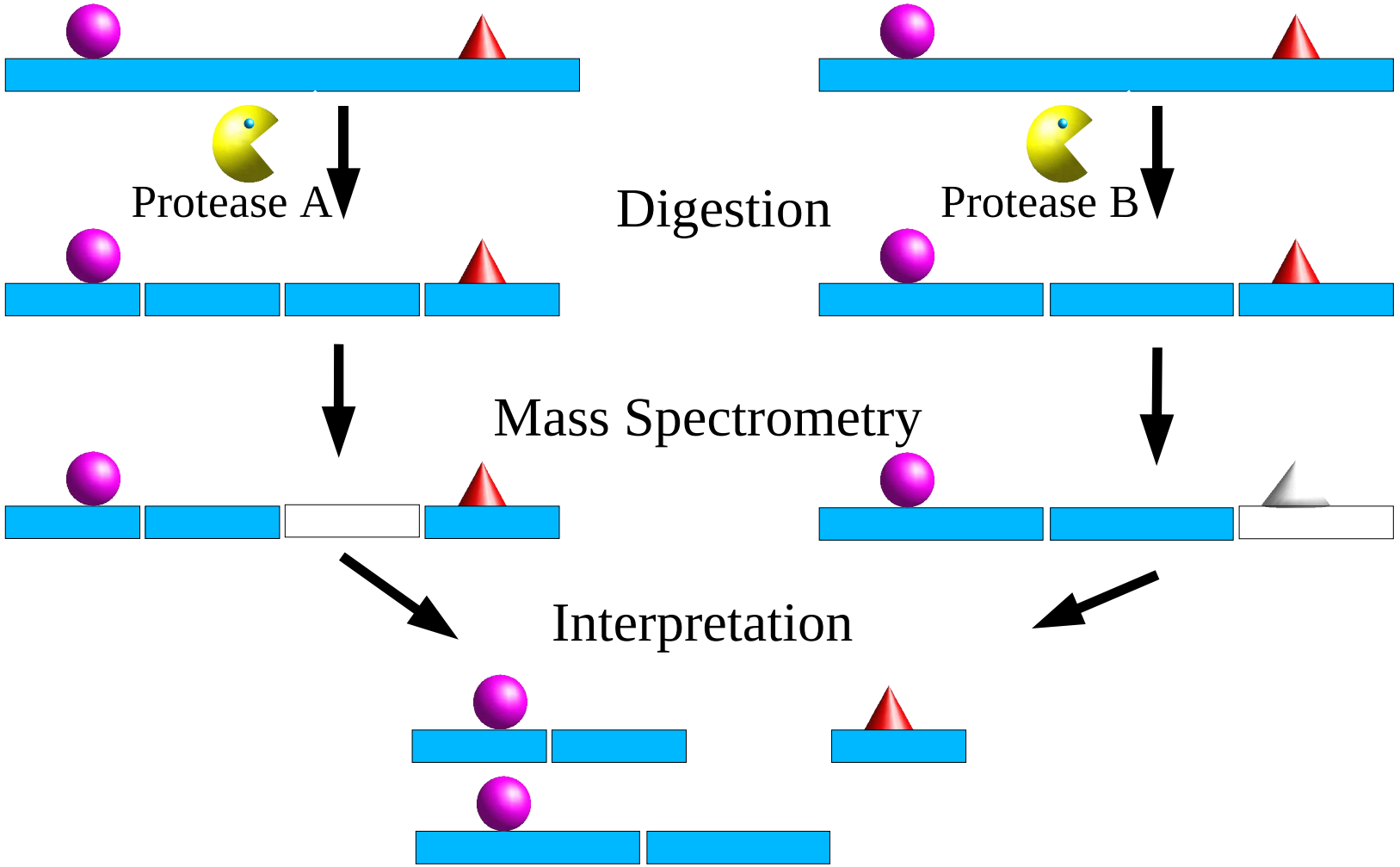
PTM Mapping a Purified Protein

- **Sequencing: Edman degradation**



PTM Mapping a Purified Protein

- **Sequencing: Mass Spectrometry**

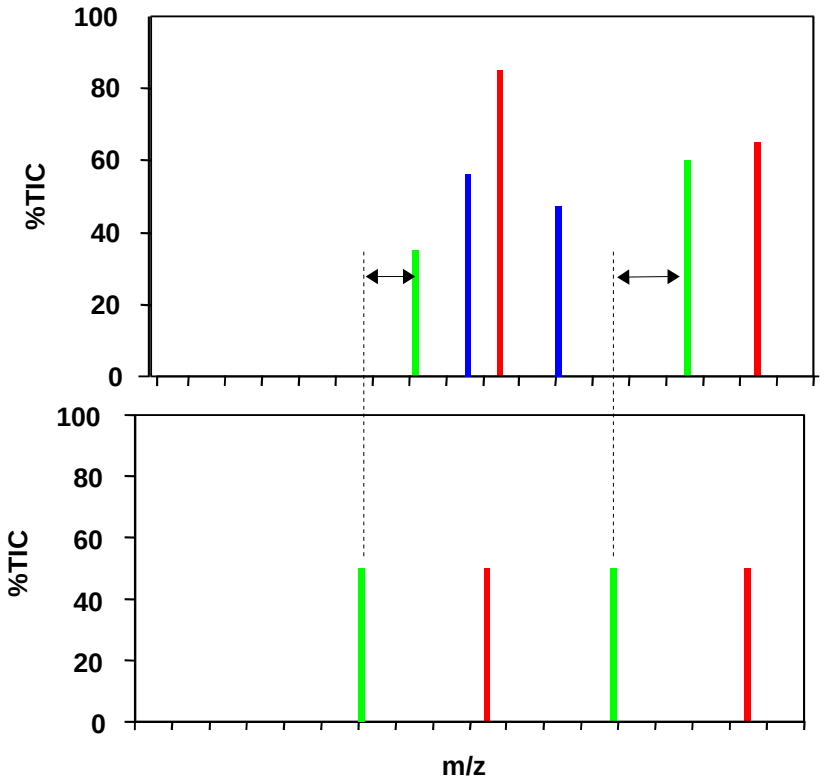


PTM Mapping a Purified Protein

Peptide Mass Fingerprinting
(MALDI, LC/ESI-MS)

Experimental

Theoretical

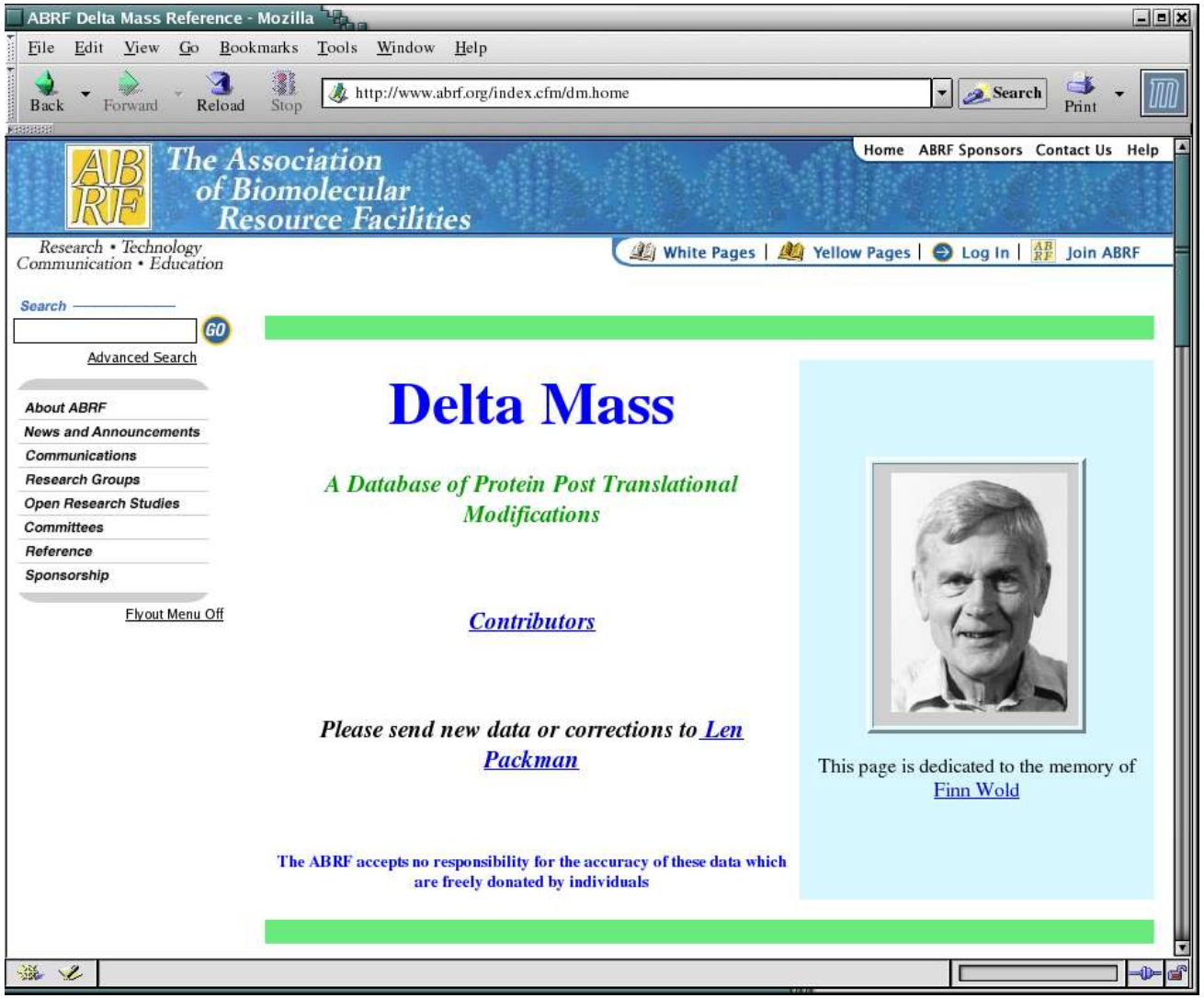


Hits

Misses

Other
Unmatched
Masses

Post-translational Modifications



http://www.abrf.org/index.cfm/dm.home

PTM Mapping a Purified Protein

Table 1. Some common and important post-translational modifications

PTM type	ΔMass[a] (Da)	Stability[b]	Function and notes
Phosphorylation			Reversible, activation/inactivation of enzyme activity, modulation of molecular interactions, signaling
pTyr	+80	+++	
pSer, pThr	+80	+/++	
Acetylation	+42	+++	Protein stability, protection of N terminus. Regulation of protein–DNA interactions (histones)
Methylation	+14	+++	Regulation of gene expression
Acylation, fatty acid modification			Cellular localization and targeting signals, membrane tethering, mediator of protein–protein interactions
Farnesyl	+204	+++	
Myristoyl	+210	+++	
Palmitoyl	+238	+/++	
etc.			
Glycosylation			
N-linked	>800	+/++	Excreted proteins, cell–cell recognition/signaling
O-linked	203, >800	+/++	*O*-GlcNAc, reversible, regulatory functions
GPI anchor	>1,000	++	Glycosylphosphatidylinositol (GPI) anchor. Membrane tethering of enzymes and receptors, mainly to outer leaflet of plasma membrane
Hydroxyproline	+16	+++	Protein stability and protein–ligand interactions
Sulfation (sTyr)	+80	+	Modulator of protein–protein and receptor–ligand interactions
Disulfide bond formation	–2	++	Intra- and intermolecular crosslink, protein stability
Deamidation	+1	+++	Possible regulator of protein–ligand and protein–protein interactions, also a common chemical artifact
Pyroglutamic acid	–17	+++	Protein stability, blocked N terminus
Ubiquitination	>1,000	+/++	Destruction signal. After tryptic digestion, ubiquitination site is modified with the Gly-Gly dipeptide
Nitration of tyrosine	+45	+/++	Oxidative damage during inflammation

[a]A more comprehensive list of PTM Δmass values can be found at: http://www.abrf.org/index.cfm/dm.home
[b]Stability: + labile in tandem mass spectrometry, ++ moderately stable; +++ stable.

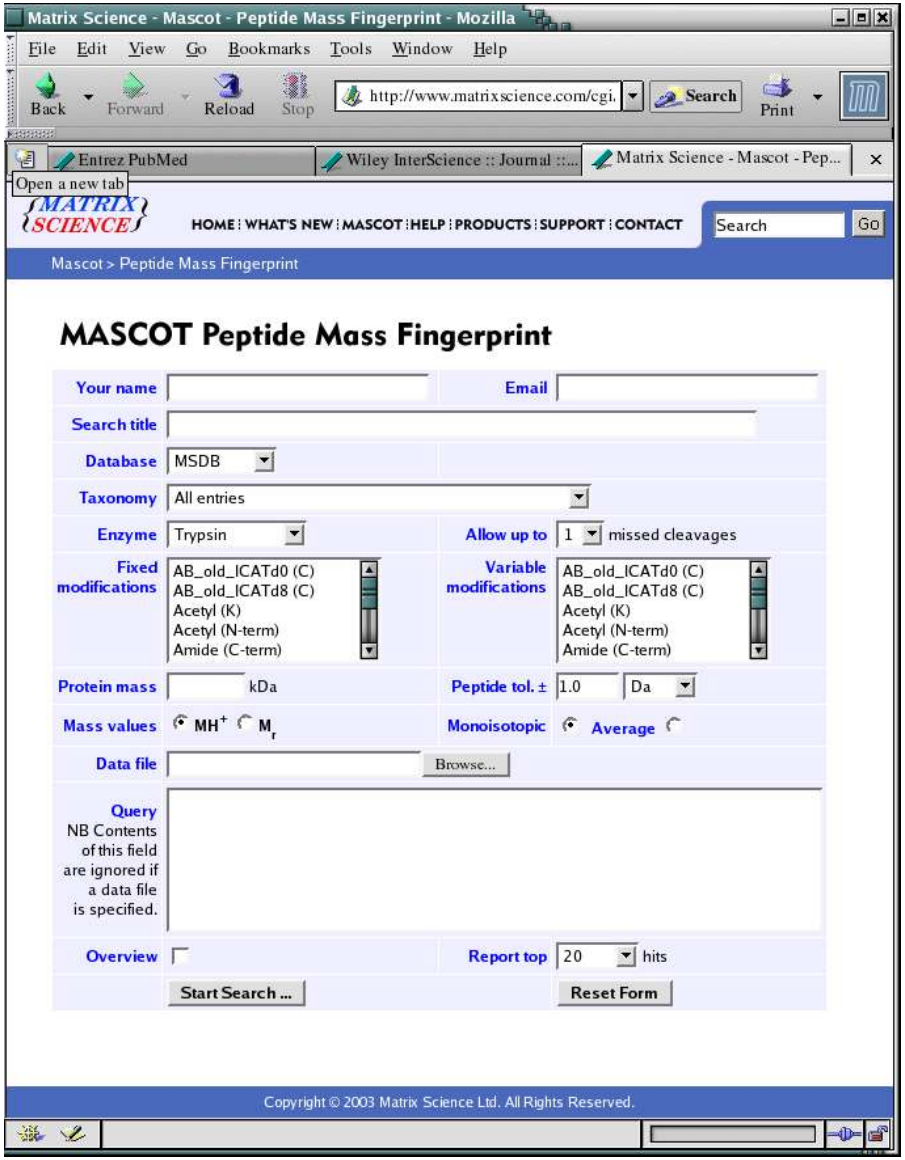
Mann M, Jensen ON. (2003) *Nat Biotechnol.* **21**:255-61.

PTM Mapping a Purified Protein

- **Unmatched masses can arise from several sources:**
  - **Contaminating proteins**
  - **Chemical modifications (eg MetO)**
  - **Unexpected cleavage**
- **The remaining unmatched matches can be considered PTMs.**

PTM Mapping a Purified Protein

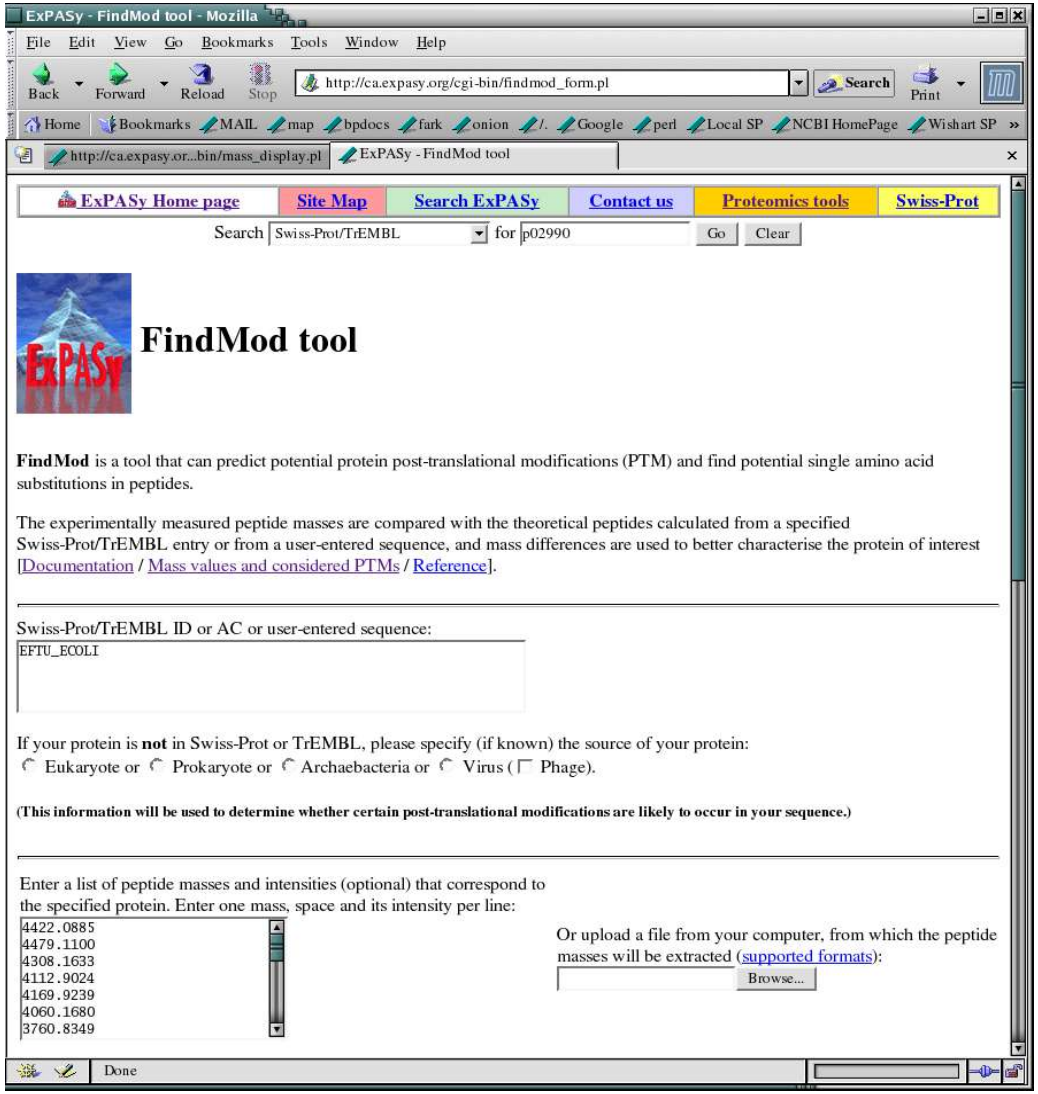
- **MS-based protein sequence prediction software can look for a limited set of modifications if they are known or suspected in advance.**
- **The combinatorial explosion involved in matching modified masses between thousands of experimental spectra and millions of database peptides is intractable.**



PTM Mapping a Purified Protein

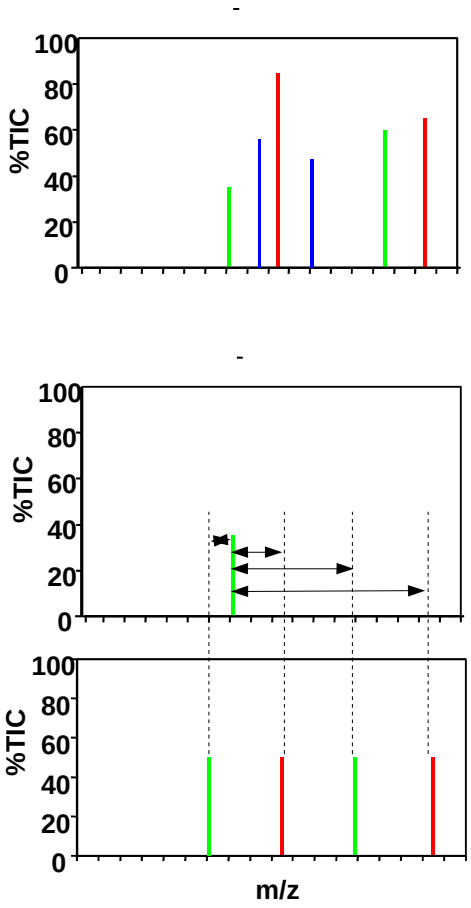
- **“Second pass” software has been written to help identify PTMs *after protein identification*.**
  - **FindMod**
  - **GlycoMod**
  - **FindPept**
  - **Salsa**

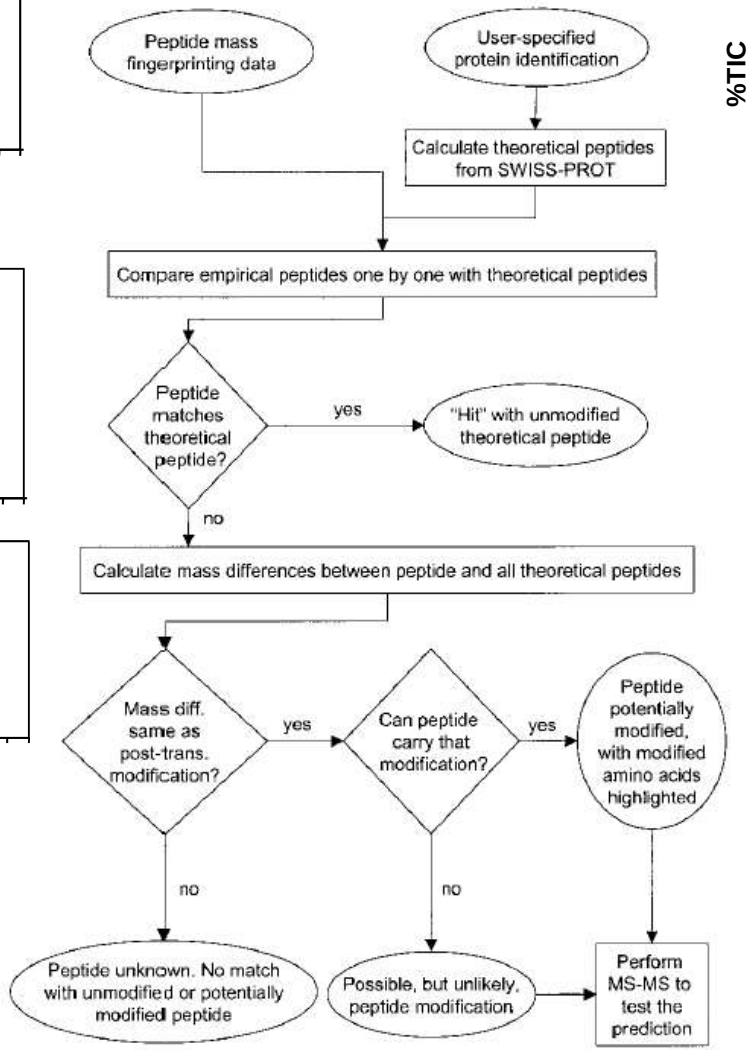
FindMod

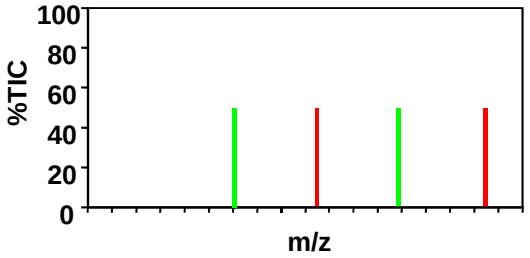


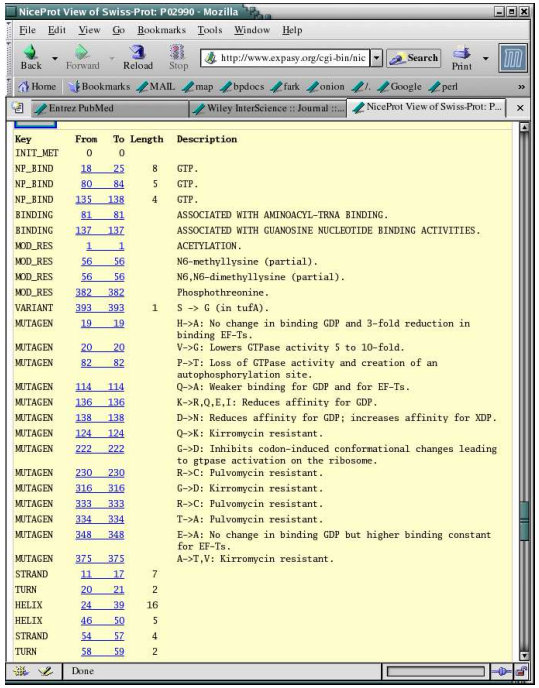
http://ca.expasy.org/tools/findmod/

FindMod


%TIC
100
80
60
40
20
0
m/z

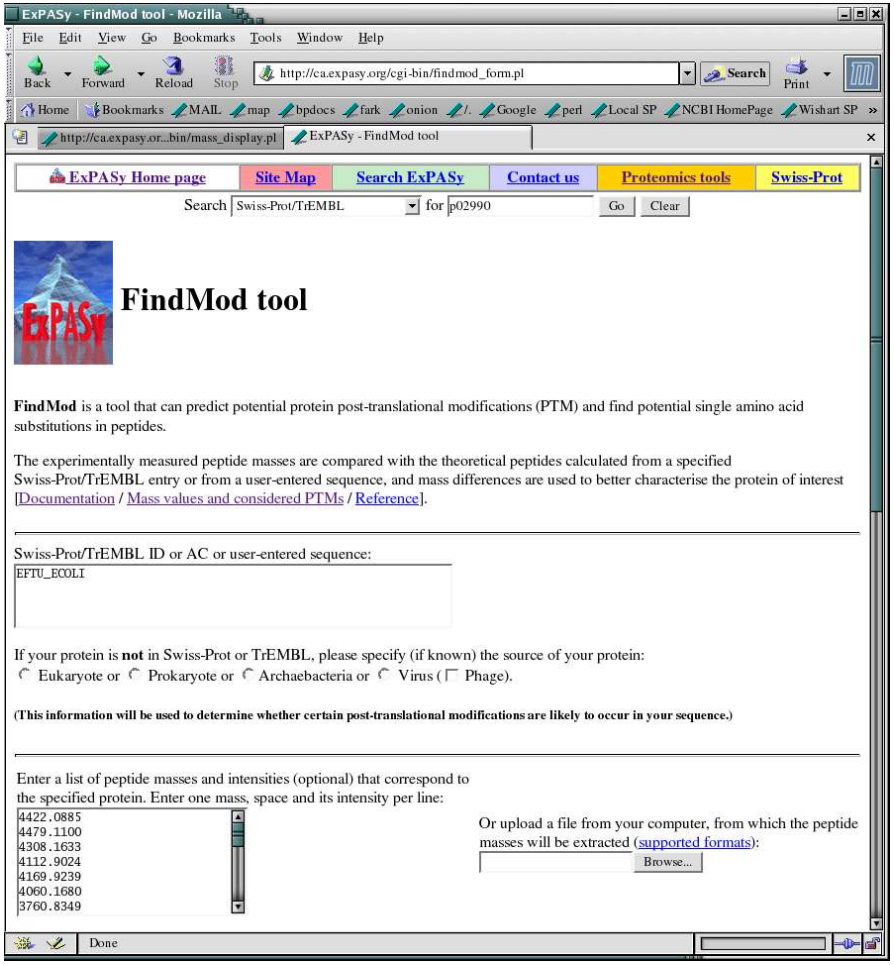

Peptide mass fingerprinting data
User-specified protein identification
Calculate theoretical peptides from SWISS-PROT
Compare empirical peptides one by one with theoretical peptides
Peptide matches theoretical peptide?
yes
"Hit" with unmodified theoretical peptide
no
Calculate mass differences between peptide and all theoretical peptides
Mass diff. same as post-trans. modification?
yes
Can peptide carry that modification?
yes
Peptide potentially modified, with modified amino acids highlighted
no
no
Peptide unknown. No match with unmodified or potentially modified peptide
Possible, but unlikely, peptide modification
Perform MS-MS to test the prediction


%TIC
100
80
60
40
20
0
m/z


NiceProt View of Swiss-Prot: P02990 - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.expasy.org/cgi-bin/nic
Search Print
Home Bookmarks MAIL map bpdocs fark onion Google perl
Entrez PubMed
Wiley InterScience :: Journal :::
NiceProt View of Swiss-Prot: P...
Key From To Length Description
INIT_MET 0 0
NP_BIND 18 25 8 GTP.
NP_BIND 80 84 5 GTP.
NP_BIND 135 138 4 GTP.
BINDING 81 81 ASSOCIATED WITH AMINOACYL-TRNA BINDING.
BINDING 137 137 ASSOCIATED WITH GUANOSINE NUCLEOTIDE BINDING ACTIVITIES.
MOD_RES 1 1 ACETYLATION.
MOD_RES 56 56 N6-methyllysine (partial).
MOD_RES 56 56 N6,N6-dimethyllysine (partial).
MOD_RES 382 382 Phosphothreonine.
VARIANT 393 393 1 S -> G (in tufA).
MUTAGEN 19 19 H->A: No change in binding GDP and 3-fold reduction in binding EF-Ts.
MUTAGEN 20 20 V->G: Lowers GTPase activity 5 to 10-fold.
MUTAGEN 82 82 P->T: Loss of GTPase activity and creation of an autophosphorylation site.
MUTAGEN 114 114 Q->A: Weaker binding for GDP and for EF-Ts.
MUTAGEN 136 136 K->R,Q,E,I: Reduces affinity for GDP.
MUTAGEN 138 138 D->N: Reduces affinity for GDP; increases affinity for XDP.
MUTAGEN 124 124 Q->K: Kirromycin resistant.
MUTAGEN 222 222 G->D: Inhibits codon-induced conformational changes leading to gtpase activation on the ribosome.
MUTAGEN 230 230 R->C: Pulvomycin resistant.
MUTAGEN 316 316 G->D: Kirromycin resistant.
MUTAGEN 333 333 R->C: Pulvomycin resistant.
MUTAGEN 334 334 T->A: Pulvomycin resistant.
MUTAGEN 348 348 E->A: No change in binding GDP but higher binding constant for EF-Ts.
A->T,V: Kirromycin resistant.
MUTAGEN 375 375
STRAND 11 17 7
TURN 20 21 2
HELIX 24 39 16
HELIX 46 50 5
STRAND 54 57 4
TURN 58 59 2
Done

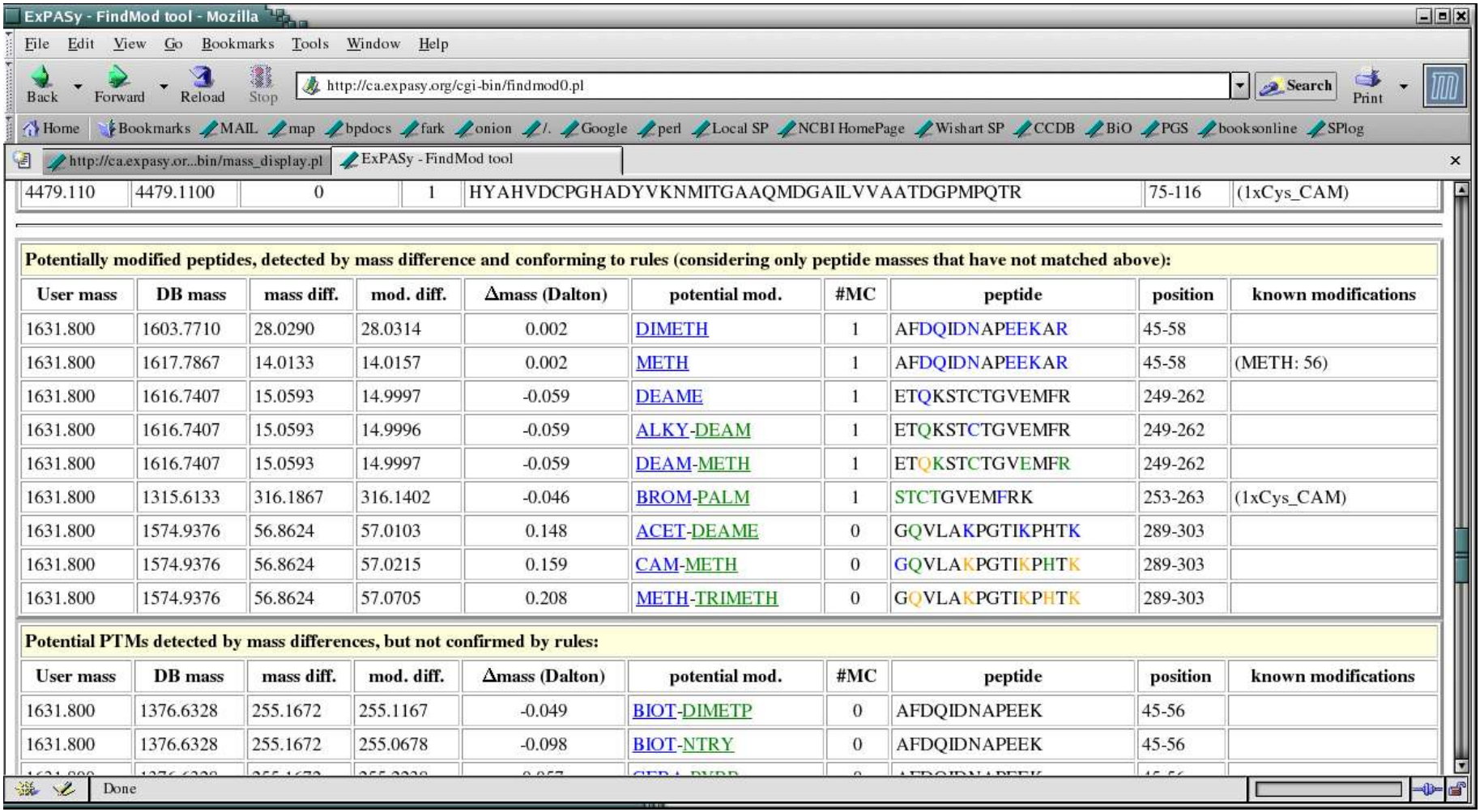
FindMod

- **Example: Lysine Dimethylation in *E. coli* elongation factor TU**



FindMod

- **Example: Lysine Dimethylation in *E. coli* elongation factor TU**



4479.110	4479.1100	0	1	HYAHVDCPGHADYVKNMITGAAQMDGAILVVAATDGPMPQTR	75-116	(1xCys_CAM)

Potentially modified peptides, detected by mass difference and conforming to rules (considering only peptide masses that have not matched above):

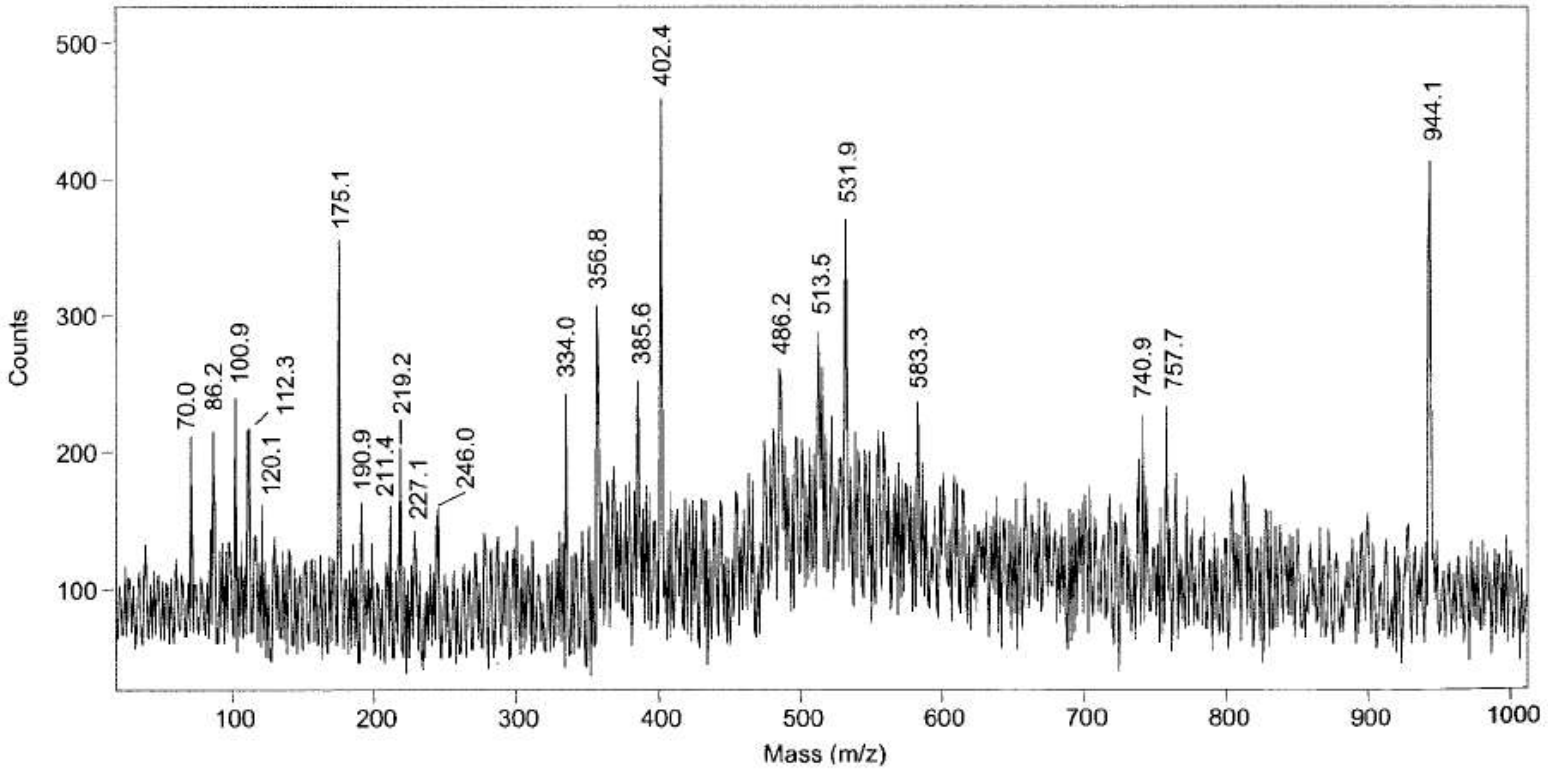
User mass	DB mass	mass diff.	mod. diff.	Δmass (Dalton)	potential mod.	#MC	peptide	position	known modifications
1631.800	1603.7710	28.0290	28.0314	0.002	DIMETH	1	AFDQIDNAPEEKAR	45-58	
1631.800	1617.7867	14.0133	14.0157	0.002	METH	1	AFDQIDNAPEEKAR	45-58	(METH: 56)
1631.800	1616.7407	15.0593	14.9997	-0.059	DEAME	1	ETQKSTCTGVEMFR	249-262	
1631.800	1616.7407	15.0593	14.9996	-0.059	ALKY-DEAM	1	ETQKSTCTGVEMFR	249-262	
1631.800	1616.7407	15.0593	14.9997	-0.059	DEAM-METH	1	ETQKSTCTGVEMFR	249-262	
1631.800	1315.6133	316.1867	316.1402	-0.046	BROM-PALM	1	STCTGVEMFRK	253-263	(1xCys_CAM)
1631.800	1574.9376	56.8624	57.0103	0.148	ACET-DEAME	0	GQVLAKPGTIKPHTK	289-303	
1631.800	1574.9376	56.8624	57.0215	0.159	CAM-METH	0	GQVLAKPGTIKPHTK	289-303	
1631.800	1574.9376	56.8624	57.0705	0.208	METH-TRIMETH	0	GQVLAKPGTIKPHTK	289-303	

Potential PTMs detected by mass differences, but not confirmed by rules:

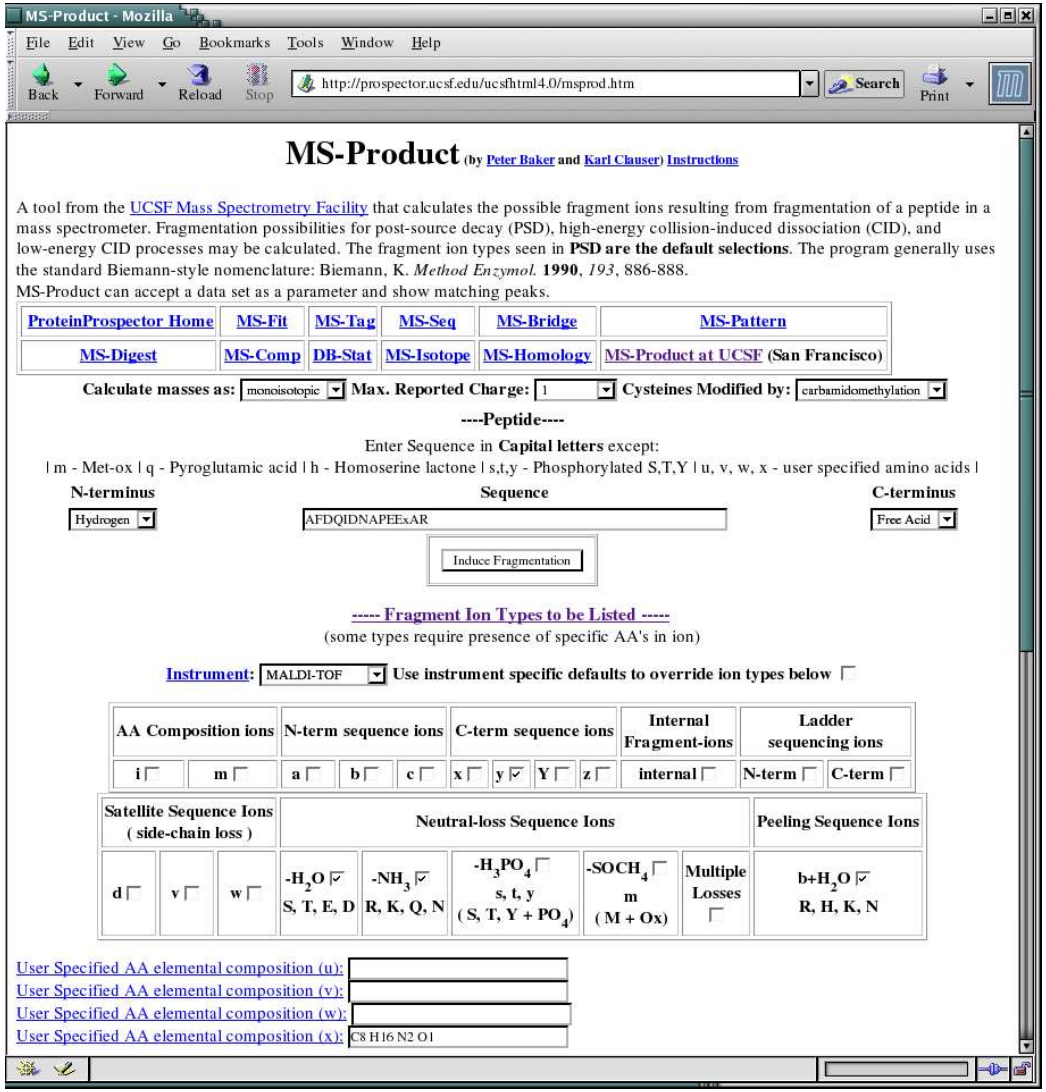
User mass	DB mass	mass diff.	mod. diff.	Δmass (Dalton)	potential mod.	#MC	peptide	position	known modifications
1631.800	1376.6328	255.1672	255.1167	-0.049	BIOT-DIMETP	0	AFDQIDNAPEEK	45-56	
1631.800	1376.6328	255.1672	255.0678	-0.098	BIOT-NTRY	0	AFDQIDNAPEEK	45-56	

FindMod

- **Post Source Decay Spectrum of AFDQIDNAPII-K(dimethyl)-AR**

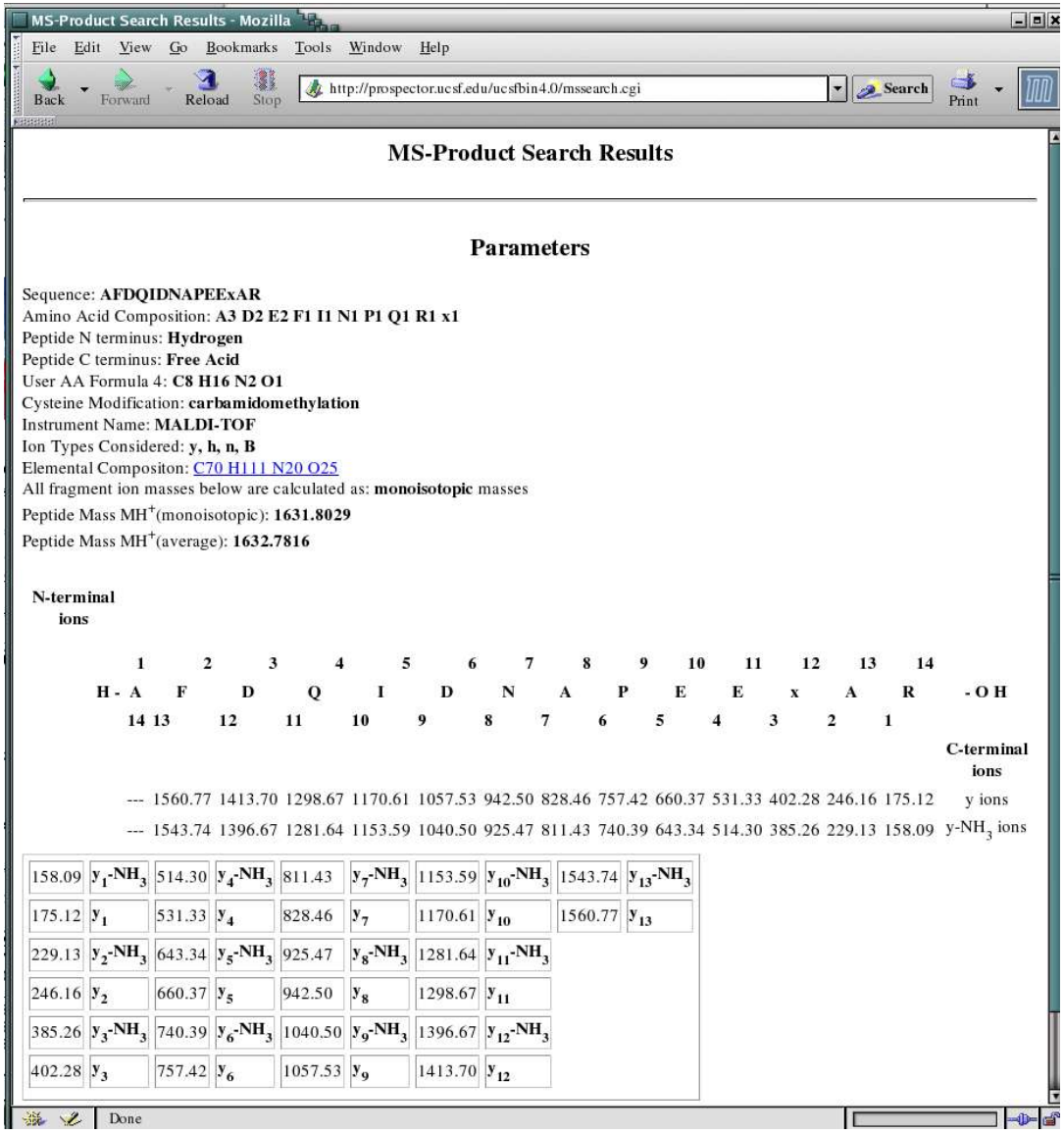


FindMod



http://prospector.ucsf.edu/ucsfhtml4.0/msprod.htm

FindMod



MS-Product Search Results

Parameters

Sequence: **AFDQIDNAPEExAR**
Amino Acid Composition: **A3 D2 E2 F1 I1 N1 P1 Q1 R1 x1**
Peptide N terminus: **Hydrogen**
Peptide C terminus: **Free Acid**
User AA Formula 4: **C8 H16 N2 O1**
Cysteine Modification: **carbamidomethylation**
Instrument Name: **MALDI-TOF**
Ion Types Considered: **y, h, n, B**
Elemental Compositon: C70 H111 N20 O25
All fragment ion masses below are calculated as: **monoisotopic** masses
Peptide Mass MH^+(monoisotopic): **1631.8029**
Peptide Mass MH^+(average): **1632.7816**

N-terminal ions

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
H -	A	F	D	Q	I	D	N	A	P	E	E	x	A	R	- O H
	14	13	12	11	10	9	8	7	6	5	4	3	2	1	

															C-terminal ions
---	1560.77	1413.70	1298.67	1170.61	1057.53	942.50	828.46	757.42	660.37	531.33	402.28	246.16	175.12		y ions
---	1543.74	1396.67	1281.64	1153.59	1040.50	925.47	811.43	740.39	643.34	514.30	385.26	229.13	158.09		$y\text{-}NH_3$ ions

158.09	$y_1\text{-}NH_3$	514.30	$y_4\text{-}NH_3$	811.43	$y_7\text{-}NH_3$	1153.59	$y_{10}\text{-}NH_3$	1543.74	$y_{13}\text{-}NH_3$
175.12	y_1	531.33	y_4	828.46	y_7	1170.61	y_{10}	1560.77	y_{13}
229.13	$y_2\text{-}NH_3$	643.34	$y_5\text{-}NH_3$	925.47	$y_8\text{-}NH_3$	1281.64	$y_{11}\text{-}NH_3$		
246.16	y_2	660.37	y_5	942.50	y_8	1298.67	y_{11}		
385.26	$y_3\text{-}NH_3$	740.39	$y_6\text{-}NH_3$	1040.50	$y_9\text{-}NH_3$	1396.67	$y_{12}\text{-}NH_3$		
402.28	y_3	757.42	y_6	1057.53	y_9	1413.70	y_{12}		

FindMod

Ion type	Amino acid or sequence	Measured mass	Theoretical mass
Immonium	P	70.0	70
	I/L	86.2	86
	Q	100.9	101
	R	112.3	112
	F	120.1	120
a 2	AF-	190.0	191.3
b 2	AF-	219.2	219.3
b 3	AFD-	334.0	334.4
y 1	-R	175.1	175.2
y 2	-AR	246.0	246.3
y 3-NH_3	-K(dimeth)AR	385.6	385.5
y 3	-K(dimeth)AR	402.4	402.5
y 4-NH_3	-EK(dimeth)AR	513.5	514.6
y 4	-EK(dimeth)AR	531.9	531.6
y 6-NH_3	-PEEK(dimeth)AR	740.9	740.8
y 6	-PEEK(dimeth)AR	757.7	757.9
Internal Fragments[a]	-PE-	227.1	227.2
	-PEE-	356.8	356.4
	-EEK(dimeth)A-	486.2	486.5
	-PEEK(dimeth)A-	583.3	583.7
	-DQIDNAPEEK(dimeth)A-	1222.2	1222.3

Wilkins et al. (1999) J. Mol. Biol. 289:645-657

FindMod

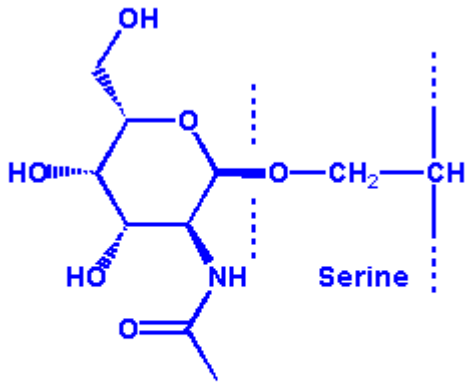
- **Considers about 30 modifications**
- **Cannot handle complex modifications such as glycosylation (except O-GlcNac).**

GlycoMod

- **Glycosylation is one of the most common PTMs (estimated >50% of all proteins are glycosylated).**
- **Glycosylations can be quite complex**
- **GlycoMod finds all possible compositions of a glycan structure from its experimentally determined mass.**

Glycosylation

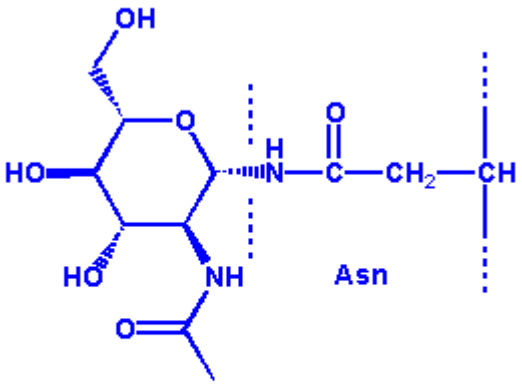
- **O-linked**
  - **Ser, Thr, Hydroxy-Lys, or hydroxy-Pro**
  - **Glycan can be linked to Ser or Thr via a phosphate group (a phosphodiester linkage).**



O-linkage to GalNAc

Glycosylation

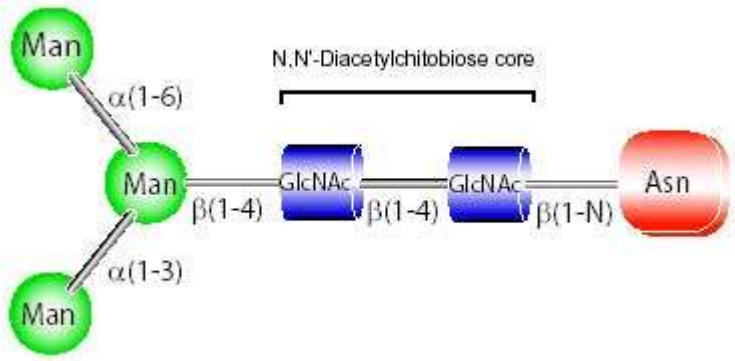
- **N-linked**
  - **Linkage is through amide nitrogen of Asparagine**
  - **Almost always have GlcNAc at the reducing terminus**
  - **Consensus sequence: -Asn-[!Pro]-[Ser|Thr|(Cys)]**



N-linkage to GlcNAc

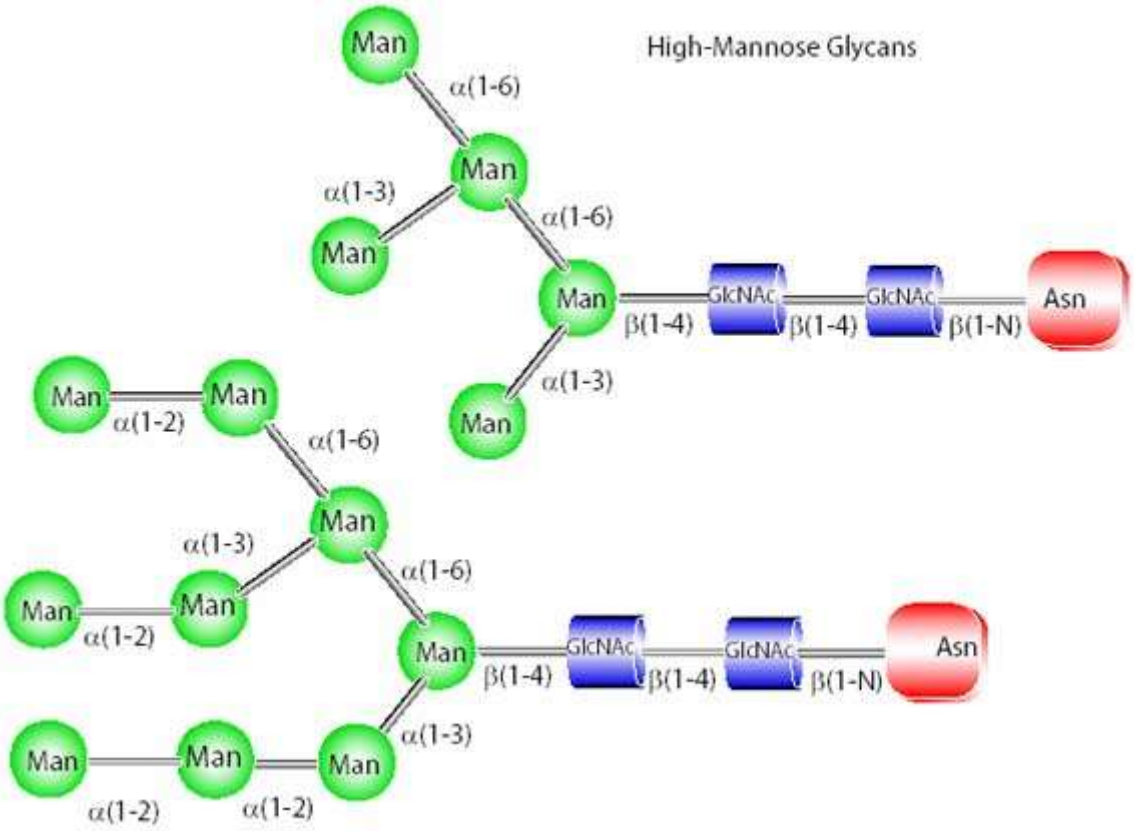
Glycosylation

- **N-linked**
  - **Almost always contains the core pentasaccharide**
    Man(α1-6)[Man(α1-3)]Man(β1-4)GlcNAc(β1-4)GlcNAc(β1-4)



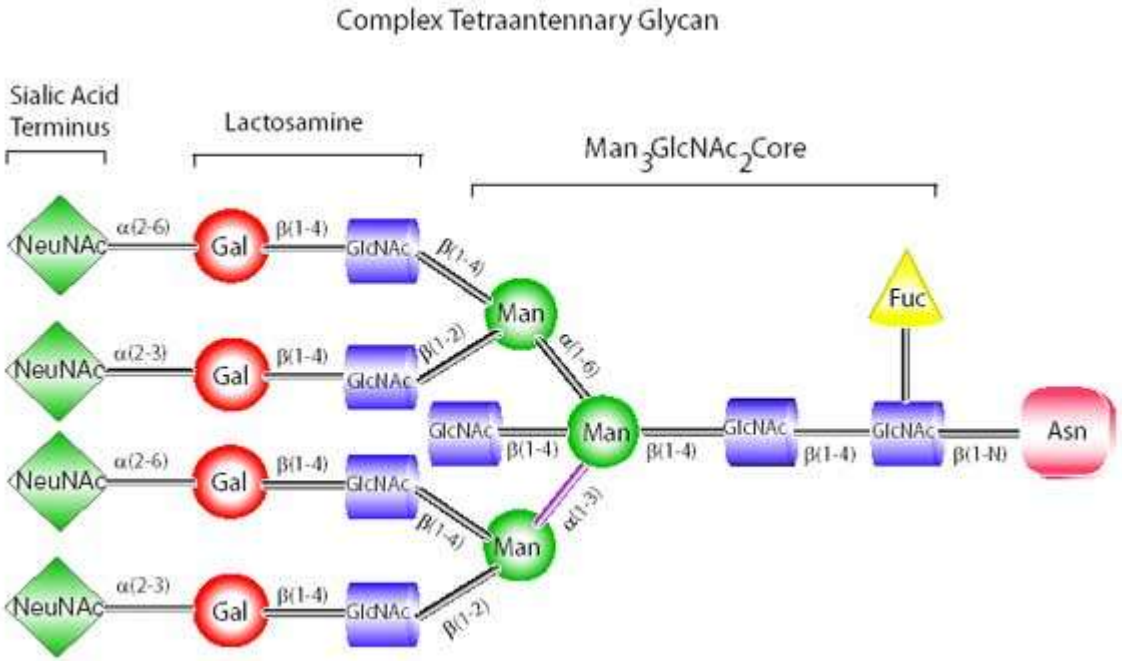
Glycosylation

- **N-linked**
  - **High mannose type**



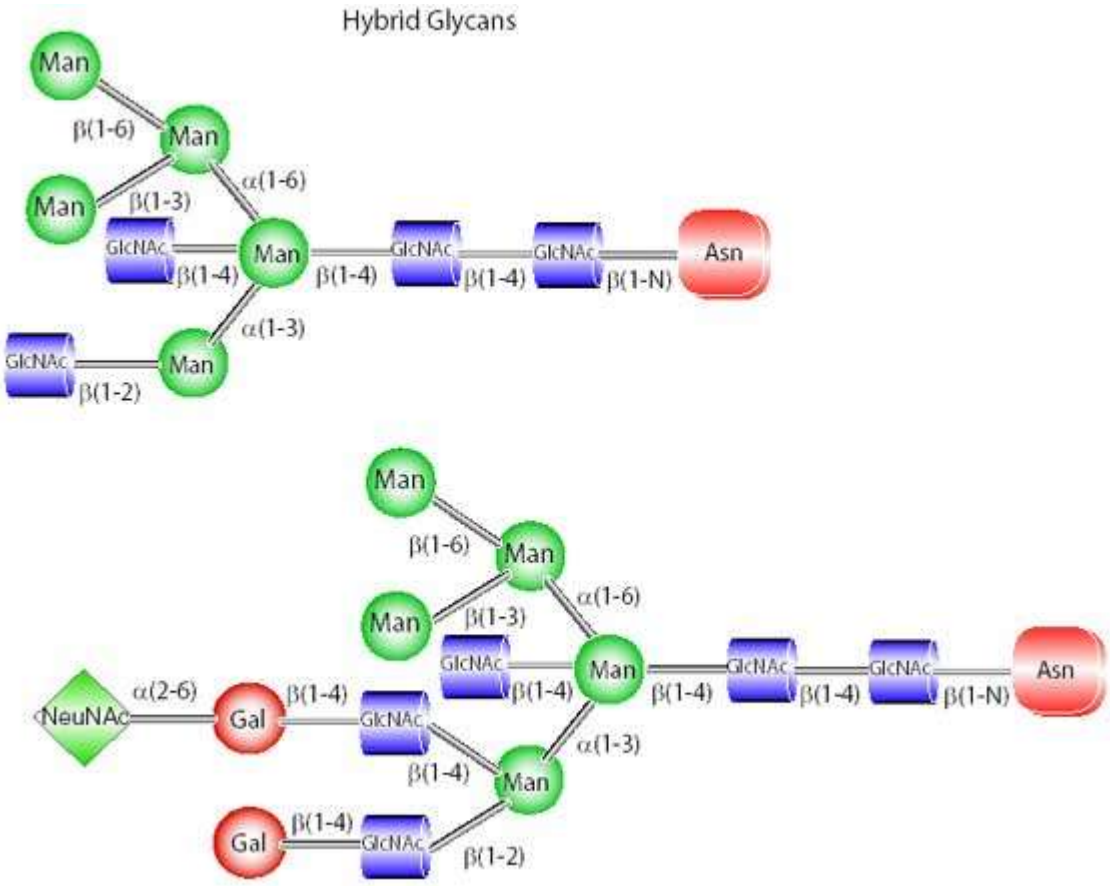
Glycosylation

- **N-linked**
  - **Complex type**



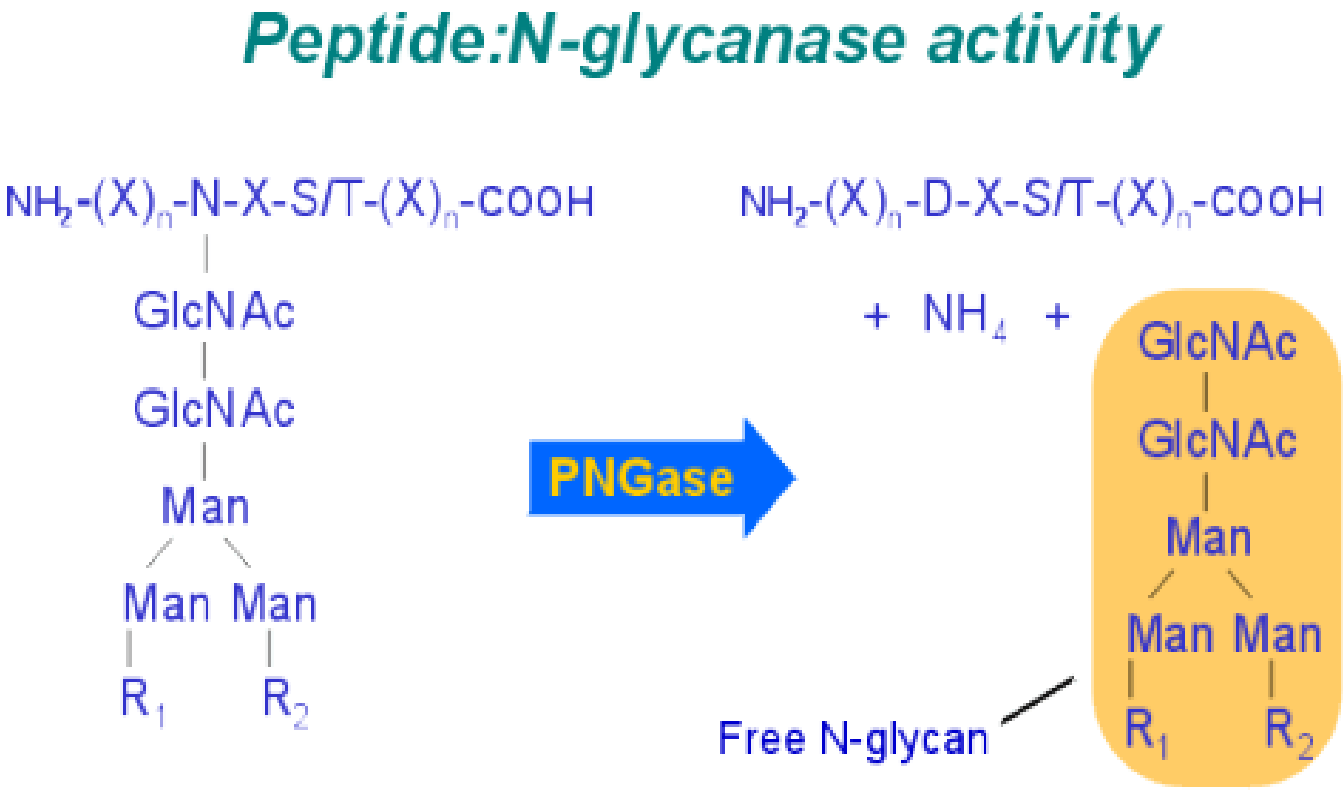
Glycosylation

- **N-linked**
  - **Hybrid type**



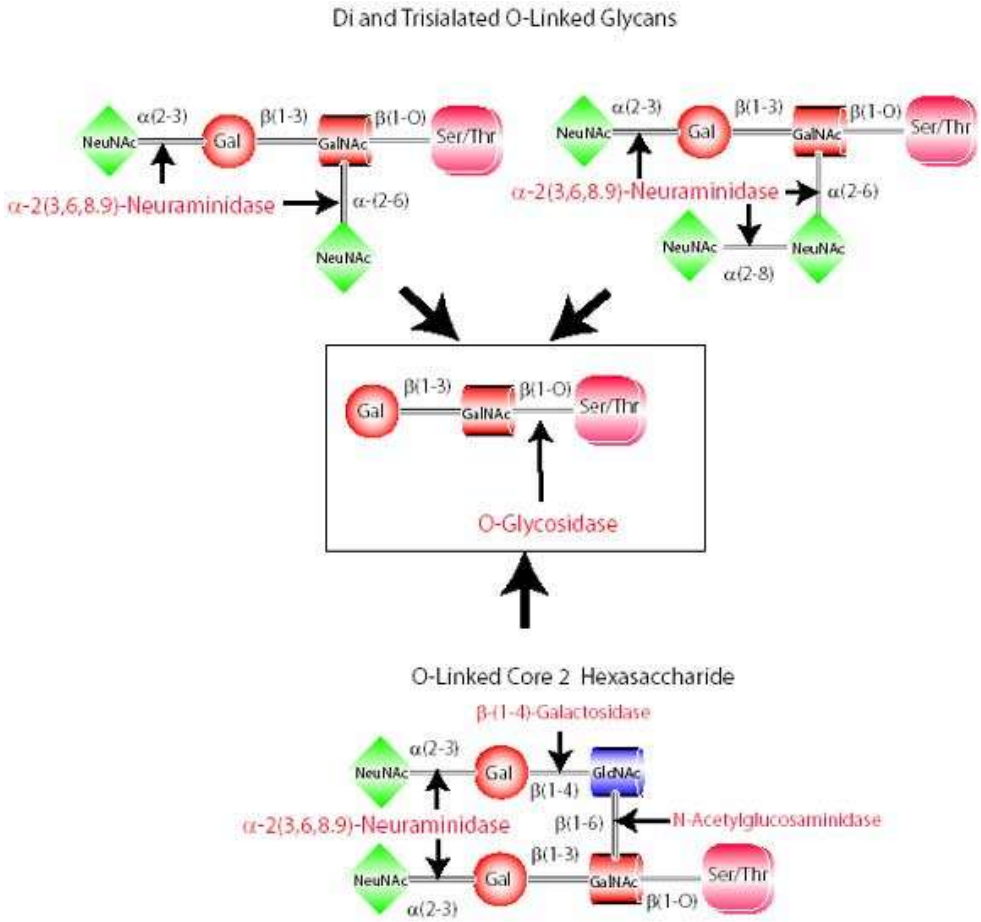
Glycosylation

- **PNGase**



Glycosylation

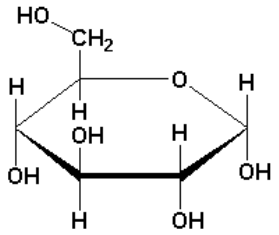
- **O-Glycosidase**



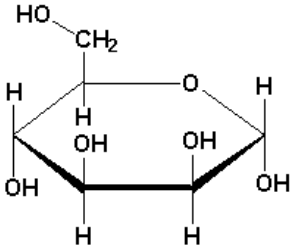
Glycosylation

- **MS cannot distinguish between monosaccharides with the same mass:**

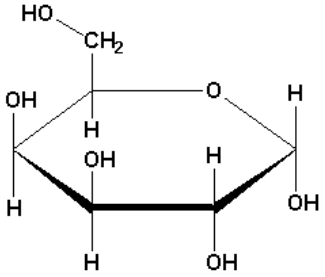
"Hexoses"



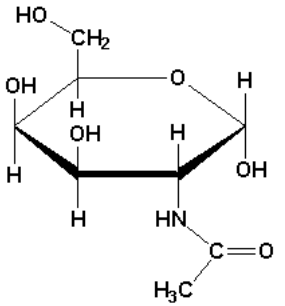
Glucose



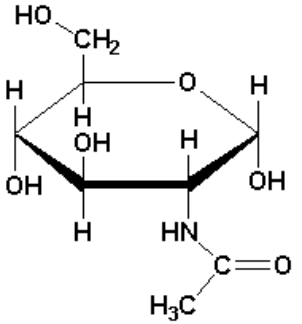
Mannose



Galactose



N-Acetylgalactosamine



N-acetylglucosamine

Glycosylation

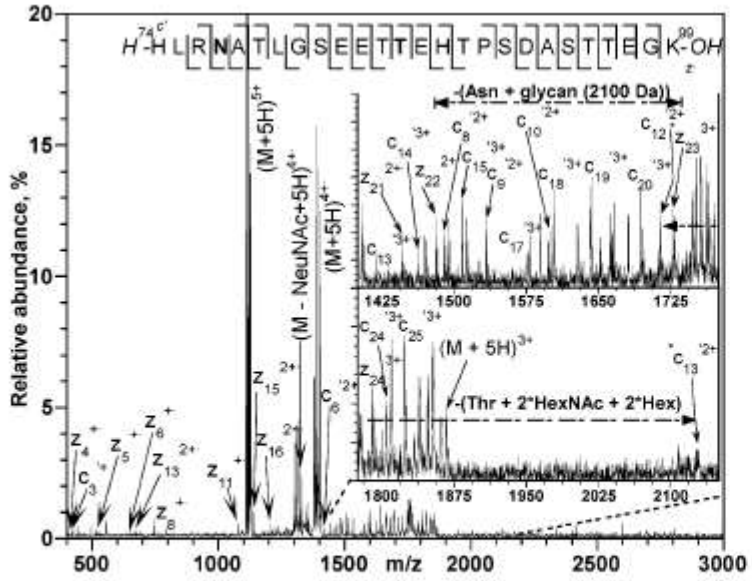


Figure 4. ECD FT mass spectrum of 5+ of peptide 74–99. Two out of 10 potential sites were glycosylated, by N-linkage at Asn77 (add-on mass 2100.7765 Da) and O-linkage at Thr86 (add-on mass 730.2535 Da).

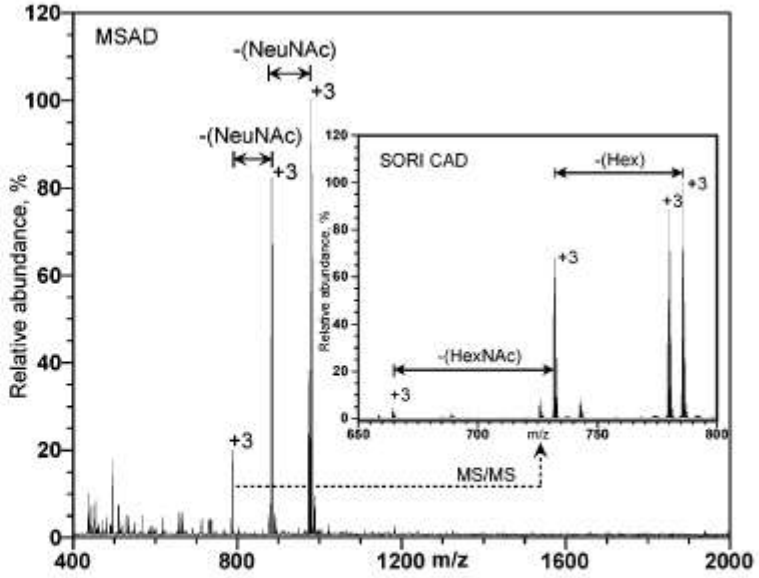
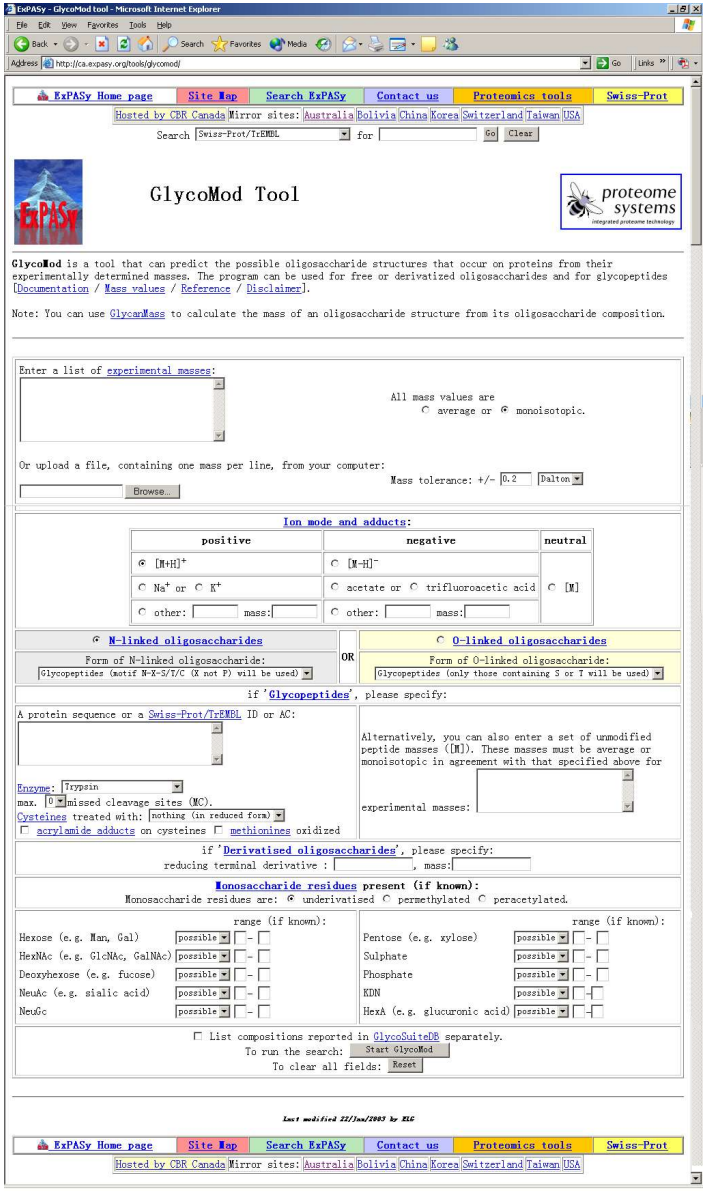


Figure 5. MSAD and SORI CAD (inset) FT mass spectra of 3+ of the peptide from fraction Lys-C 18 showing consecutive losses of two units of NeuNAc (291 Da), one unit of Hex (162 Da), and one unit of HexNAc (203 Da).

Kjeldsen et al. Anal. Chem.2003, 75,2355-2361

GlycoMod

- **Finds all possible compositions of a glycan structure from its experimentally determined mass by comparing the the mass against a list of precomputed glycan masses.**
- **http://ca.expasy.org/tools/glycomod/**



GlycoMod

Table 1. Upper limits imposed on the number of residues of a particular monosaccharide (based on a survey of the literature) found in *O*-linked and *N*-linked oligosaccharides using GlycoMod

Monosaccharide residue	*O*-linked oligosaccharides	*N*-linked oligosaccharides
Hexose	0–14	0–20
HexNAc	0–14	0–20
Deoxyhexose	0–6	0–6
NeuAc	0–7	0–5
NeuGc	0–7	0–5
Pentose	0–3	0–3
Sulphate	0–6	0–3
Phosphate	0–6	0–2
KDN	0–2	0–0
HexA	0–2	0–0

Cooper et al. (2001) *Proteomics* **1**:340-349

GlycoMod

Example: 2-aminopyridine derivatised N-linked oligosaccharide
From horseradish peroxidase

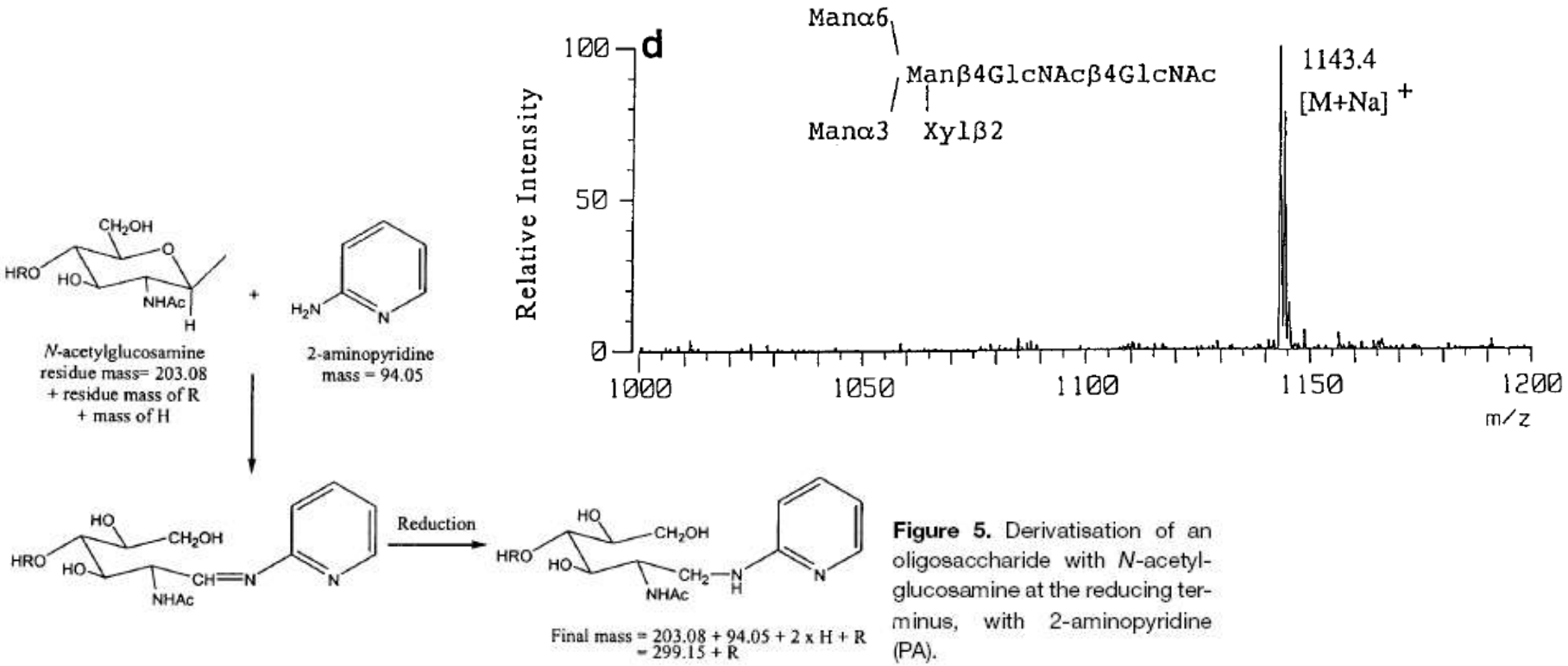


Figure 5. Derivatisation of an oligosaccharide with *N*-acetylglucosamine at the reducing terminus, with 2-aminopyridine (PA).

GlycoMod

Example: 2-aminopyridine derivatised N-linked oligosaccharide From horseradish peroxidase

- Mass Value: 1143.4
- Mass Type: monoisotopic
- IonMode/Adducts: [M+Na]+
- N-linked oligo form: Derivatised oligosaccharide
- Reducing terminal derivative id: 'PA'
- Reducing terminal derivative monoisotopic mass: 94.053

GlycoMod

Example: 2-aminopyridine derivatised N-linked oligosaccharide
From horseradish peroxidase

User mass: 1143.4

Adduct (Na^+): 22.989768
Derivative mass (PA): 96.06874

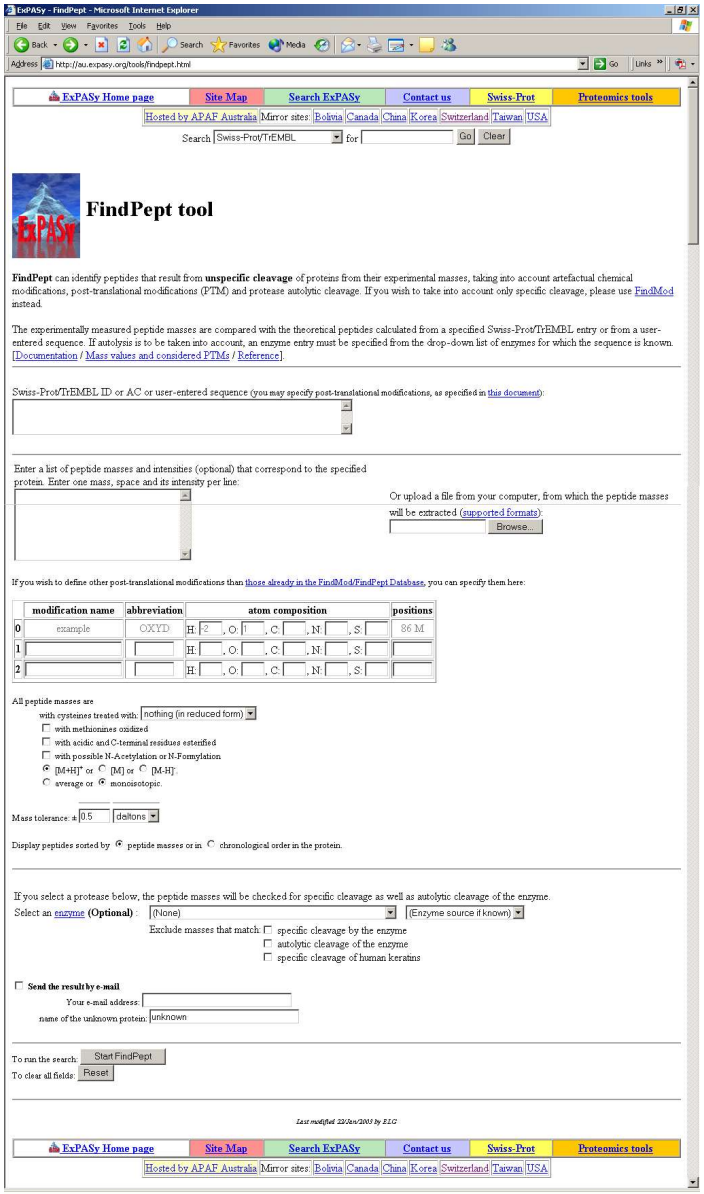
glycoform mass	Δmass (Dalton)	structure	type	Links
1024.323	0.018	$(Hex)_1$ $(NeuAc)_1$ $(NeuGc)_1$ $(Pent)_2$	-	
1024.36	-0.018	$(Hex)_3$ $(HexNAc)_2$ $(Pent)_1$	-	GlycoSuiteDB

2 structures found.

FindPept

- **Identifies the origin of "missing" peptide masses resulting from unspecific proteolytic cleavage, missed cleavage, protease autolysis or keratin contaminants.**
- **Takes into account PTMs and protease autolysis peaks (and keratin peaks).**

http://ca.expasy.org/tools/findpept.html



FindPept

Table 1. List of post-translational and artifactual modifications handled by FindPept. Modifications are applied optionally, except those marked with [*] which are quantitatively applied. For more details on PTMs, see http://www.expasy.org/tools/findmod/findmod_masses.html.

PTMs read in the feature table (FT lines) of a SWISS-PROT entry:	
FT key	*description*
INIT_MET	Initiator methionine cleavage
MOD_RES and BINDING	Acetylation; amidation; beta-methylthiolation; biotin; carbamylation; citrullination; deamidation; dimethylation; FAD; formylation; gamma-carboxyglutamic acid; hydroxylation; methylation; phosphorylation; pyridoxal phosphate; phosphopantetheine; pyrrolidone carboxylic acid; sulfation; trimethylation
CARBOHYD	N-linked glycation, O-linked N-acetylglucosamination, C-linked mannosylation
LIPID	N-acyl diglyceride, farnesyl, geranylgeranyl, lipoyl, myristyl, palmityl, n-octanoyl
CHAIN and PEPTIDE	Cleavage of chains

Artificial modifications specified by the user:

- Treatment of cysteine by iodoacetamide, iodoacetic acid or 4-vinylpyridene [*]
- Acrylamide adducts on cysteine
- Modification of methionine to homoserine lactone by CNBr cleavage
- Oxidation of methionine
- Methyl esterification of carboxylate side chains and terminus [*]

PTMs occurring frequently naturally which are considered by the program:

- Possible N-terminal acetylation
- Possible N-terminal formylation

FindPept

- **Example: Atypical cleavage of chicken lysozyme**
- **Scenario: Chicken lysozyme C (P00698) digestion with lysyl endopeptidase (Lys C) gives a mass at 1262.591 that cannot be explained by the cleavage rule for this enzyme (carboxy side of Lys, including Lys Pro).**

You have selected LYC_CHICK (P00698) from Swiss-Prot:

Lysozyme C precursor (EC 3.2.1.17) (1,4-beta-N-acetylmuramidase C) (Allergen Gal d 4) (Gal d IV).
Signal in positions 1-18 has been removed.

- Chain Lysozyme C at positions 19 - 147 [Theoretical pI: 9.32 / Mw (average mass): 14313.14 / Mw (monoisotopic mass): 14303.88]

mass	position	#MC	artif. modification(s)		peptide sequence
6897.1896	52-114	0	Cys_CAM: 82, 94, 98, 112	7125.2755	FESNFNTQATNRNTDGSTDY GILQINSRWWCNDGRTPGSR NLCNIPCSALLSSDITASVN CAK
2280.1090	32-51	0	Cys_CAM: 48	2337.1305	RHGLDNYRGYSLGNWVCAAK
2177.0491	116-134	0	Cys_CAM: 133	2234.0705	IVSDGNGMNAWVAWRNRCK
1474.7583	135-147	0	Cys_CAM: 145	1531.7798	GTDVQAWIRGCRL
1295.6598	20-31	0	Cys_CAM: 24	1352.6813	VFGRCELAAAMK

98.4% of sequence covered (you may modify the input parameters to display also peptides < 500 Da):

```
      1          11         21         31         41         51
      |          |          |          |          |          |
    1                    kV FGRCELAAAM KRHGLDNYRG YSLGNWVCAA KFESNFNTQA    60
   61 TNRNTDGSTD YGILQINSRW WCNDGRTPGS RNLCNIPCSA LLSSDITASV NCAKkIVSDG   120
  121 NGMNAWVAWR NRCKGTDVQA WIRGCRL
```

FindPept

- **Example: Atypical cleavage of chicken lysozyme**
- **MS-MS analysis confirms (K)GTDVQAWIRGC(R).**
- ***Myxobacter* Lys C has been shown to cleave C-term of Carboxymethylcysteine.**

Matching peptides for specific cleavage:

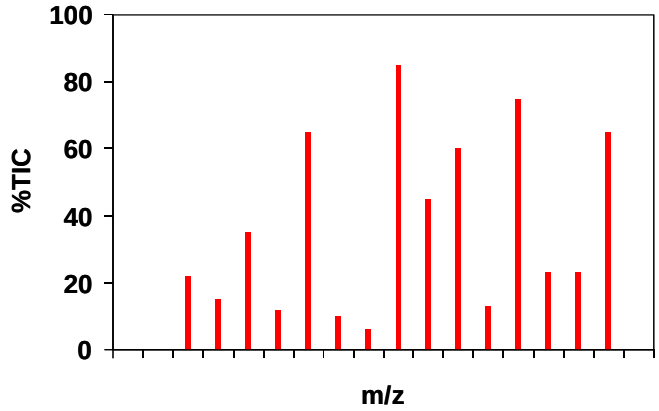
User mass	DB mass	Δmass (ppm)	peptide	position	modifications	missed cleavages
1352.681	1352.681	0.200	(K)/VFGRCELAAAMK/(R)	20-31	CYS_CAM	0
1531.780	1531.780	-0.100	(K)/GTDVQAWIRGCRL	135-147	CYS_CAM	0
2234.070	2234.071	0.200	(K)/IVSDGNGMNAWVAWRNRCK/(G)	116-134	CYS_CAM	0
2337.131	2337.130	-0.200	(K)/RHGLDNYRGYSLGNWVCAAK/(F)	32-51	CYS_CAM	0
7125.275	7125.275	0.000	(K)/FESNFNTQATNRNTDGSTDY GILQINSRWWCNDGRTPGSR NLCNIPCSALLSSDITASVN CAK/(K)	52-114	CYS_CAM CYS_CAM CYS_CAM CYS_CAM	0

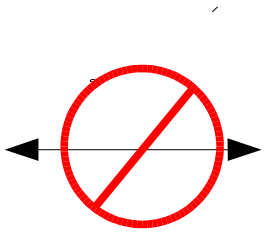
Matching peptides for unspecific cleavage:

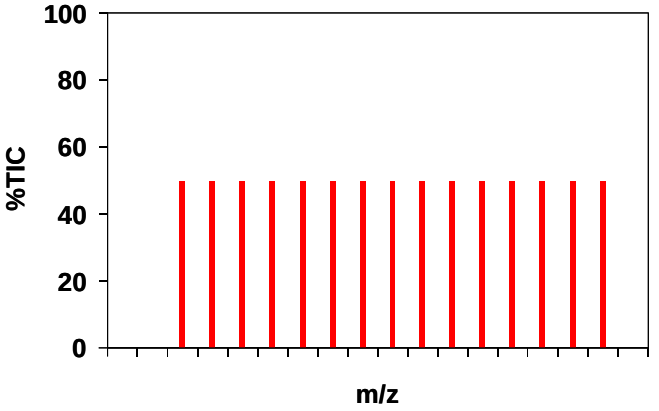
User mass	DB mass	Δmass (ppm)	peptide	position	modifications	missed cleavages
1262.591	1262.583	-6.000	(K)/IVSDGNGMNAWV(A)	116-127		0
1262.591	1262.595	2.800	(K)/GTDVQAWIRGC(R)	135-145	CYS_CAM	0
1262.591	1262.610	14.900	(I)LQINSRWWC(N)	74-82	CYS_CAM	0
1262.591	1262.616	19.500	(V)NCAKKIVSDGNG(M)	111-122	CYS_CAM	1

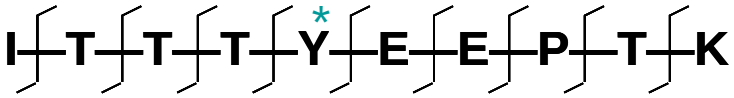
SALSA

- **MS/MS-based sequence prediction packages cannot identify 'unanticipated' modifications.**







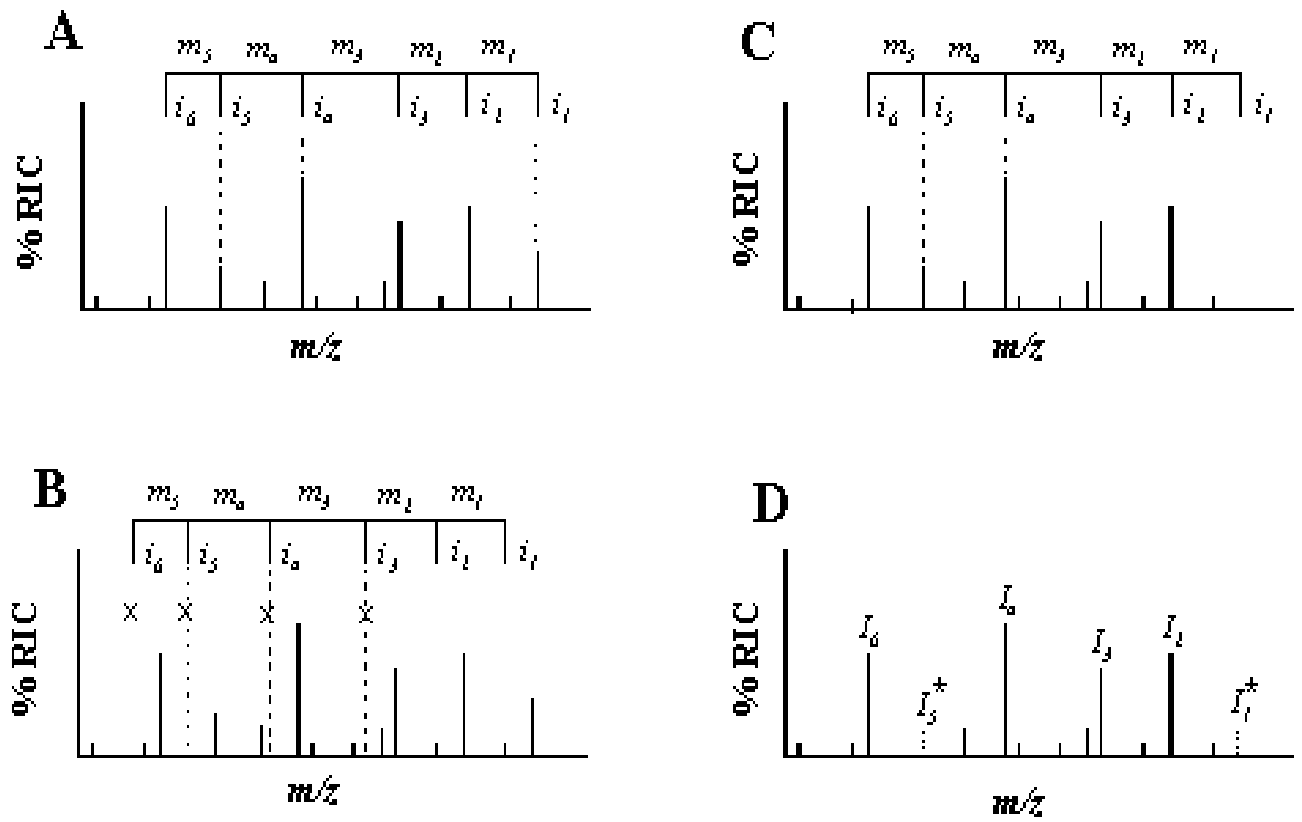


(M+H)+ = X+m

(M+H)+ = X

- ***SALSA (S*coring *AL*gorithm for *S*pectral *A*nalysis) Uses pattern recognition identify motifs and PTMs in MS/MS data.**

SALSA



SALSA

- **SALSA uses Sequences to find patterns in MS/MS spectra; this is the opposite of sequence prediction programs like Mascot and SEQUEST.**
- **Can find modified peptides by specifying the peptide sequences manually**
- **SALSA score is not statistically formulated.**

PTMS and The Hypergeometric Distribution

- The hypergeometric distribution models the total number of “successes” in a fixed size sample drawn without replacement from a finite population.
  - Eg. A deck of 52 cards has 16 face cards. Say you were to draw randomly 10 cards from the deck. What is the probability of getting 8 face cards?

The Hypergeometric Distribution

$$y = f(x|M,K,n) = \frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}} \qquad \binom{n}{x} = \frac{(n)!}{(n)!(n-x)!}$$

Where:

M is the size of the population (52 cards).

K is the number of items with the desired characteristic in the population (16 face cards).

n is the number of samples drawn (10).

x is the desired outcome (8 face cards).

$$y = \frac{\binom{16}{8}\binom{36}{2}}{\binom{52}{10}} = \frac{\frac{(16)!}{(8)!(8)!}\frac{(36)!}{(2)!(34)!}}{\frac{(52)!}{(10)!(42)!}} = 5.1\times10^{-4}$$

The Hypergeometric Distribution

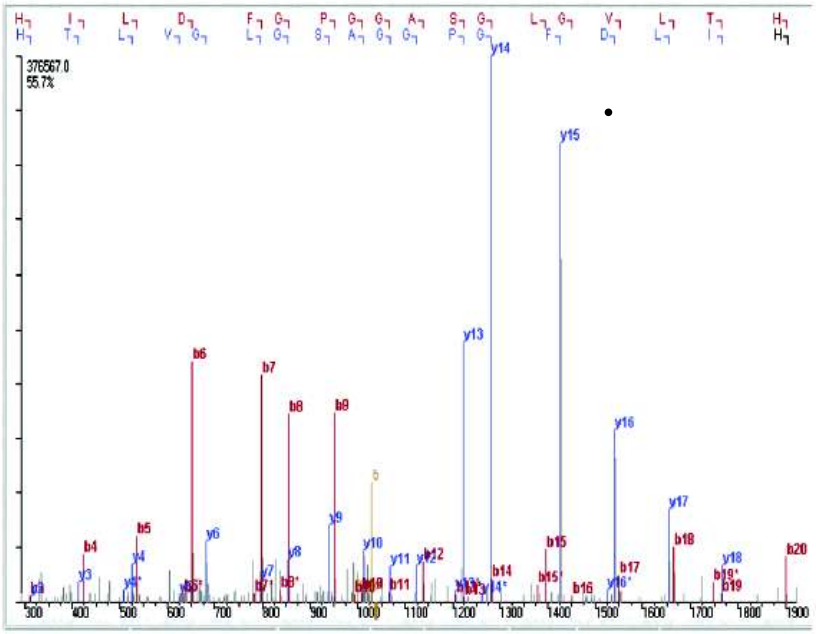
1. Perform an *in silico* digestion of the protein database and identify the (M+H)+ mass for each peptide.
2. Perform an *in silico* fragmentation of each peptide and associate them with the parent ion.
3. Extract all the peptides from the database with a parent (M+H)+ mass matching the experimental parent mass and count the number of fragments for each. The total number all these fragment masses is the population size, M .

The Hypergeometric Distribution

4. Identify the number of these fragment ion masses from peptide sequences that have a match to a peak in the experimental spectrum, K.
5. For each peptide sequence, the total number of theoretical fragment ion masses is the sample size, n.
6. The number of matches between the theoretical and experimental spectrum is x.

The Hyergeometric Distribution

- Example: ATHILDFGPGGASGLGVLTHR Against the NCBI Non-redundant Protein Database



M is the total number of theoretical fragment ions from the database that have (M+H)+ masses matching the experimental spectrum (5 284 150).

K is the subset of these fragment ions that match a fragment ion peak in the experimental spectrum (569 160).

n is the number of fragment ions for the peptide sequence under consideration (40).

x is the number of fragment ion matches between the experimental and theoretical fragmentation spectra (34).

The Hypergeometric Distribution

$$y = f(x|M,K,n) = \frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}} \qquad \binom{n}{x} = \frac{(n)!}{(n)!(n-x)!}$$

$$y = \frac{\binom{569160}{34}\binom{34714990}{6}}{\binom{5284150}{40}} = \frac{\frac{(569160)!}{(34)!(569126)!}\frac{(4714990)!}{(6)!(4714984)!}}{\frac{(5284150)!}{(40)!(5284110)!}} = 1.0\times10^{-26.62}$$

M is the total number of theoretical fragment ions from the database that have (M+H)+ masses matching the experimental spectrum (5 284 150).

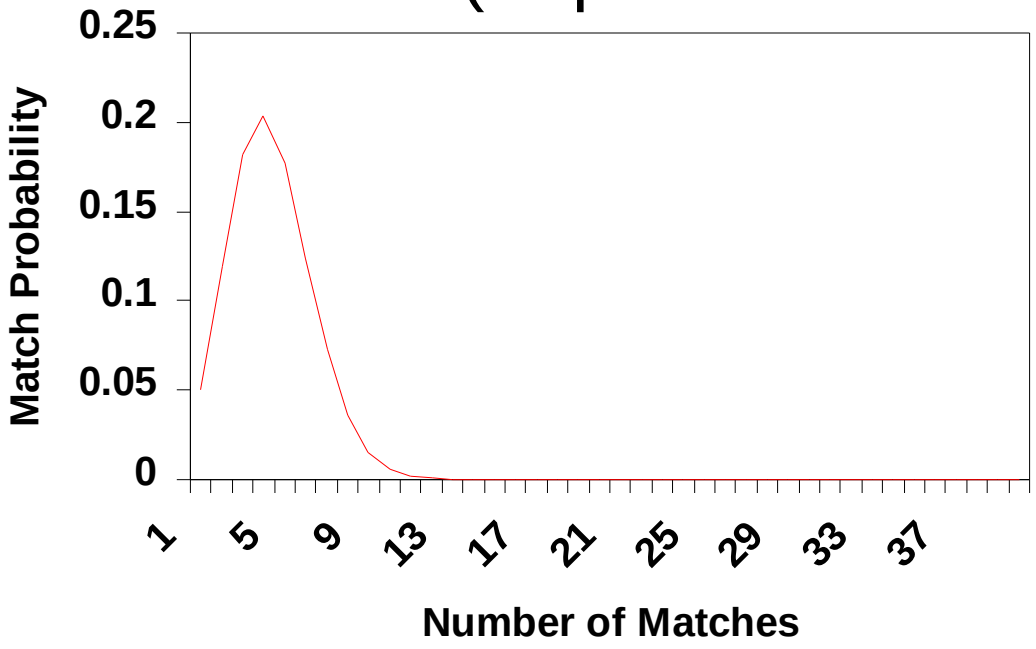
K is the subset of these fragment ions that match a fragment ion peak in the experimental spectrum (569 160).

n is the number of fragment ions for the peptide sequence under consideration (40).

x is the number of fragment ion matches between the experimental and theoretical fragmentation spectra (34).

The Hypergeometric Distribution

- Significance of a match: How many peptides in a database are expected to have the same or better matches to the experimental spectrum?
- P-Value (Expectation Value)



$$E = M \sum_{P<P_0}^{P_0} P$$

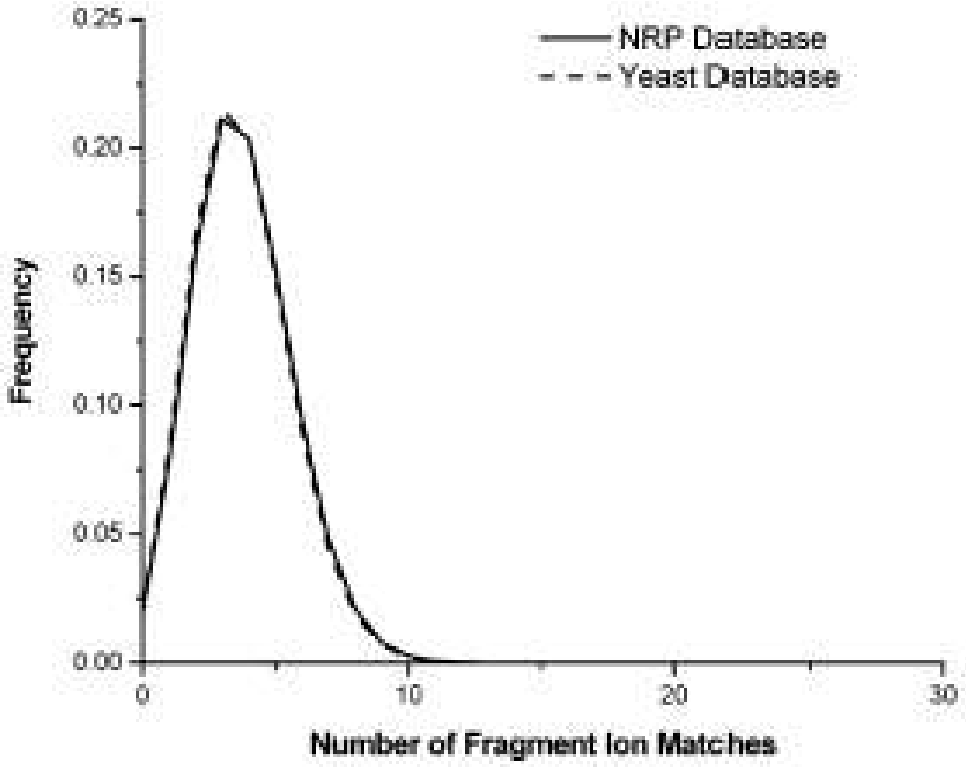
$$= 5284150 \times 10^{-26.62}$$

$$= 1 \times 10^{-19.90}$$

$E < 0.05$ considered significant

The Hypergeometric Distribution

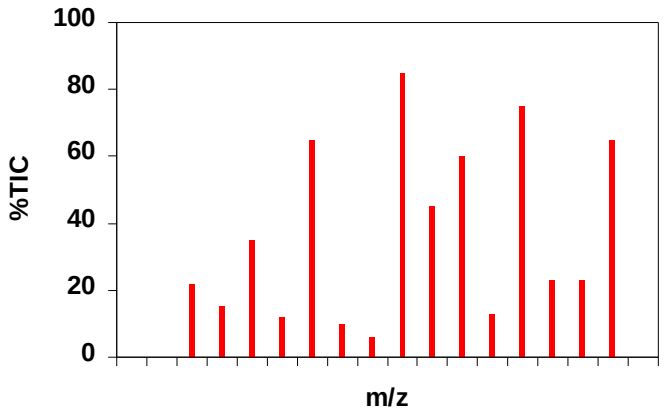
- The hypergeometric model is database independent:



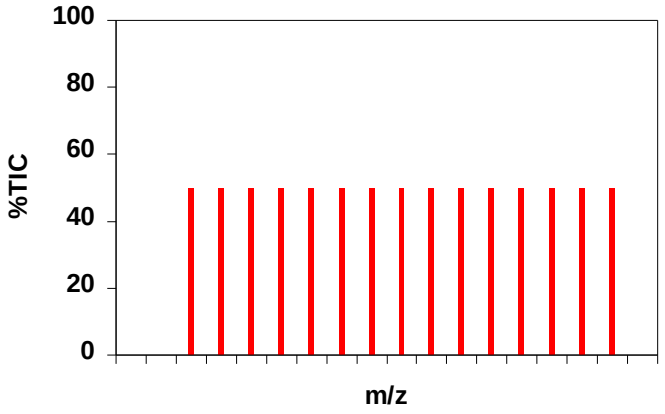
The same experiment conducted against the S. cerevisiae databases (6200 sequences) gives a match probability of $10^{-26.51}$ (vs $10^{-26.62}$ for NCBI NR).

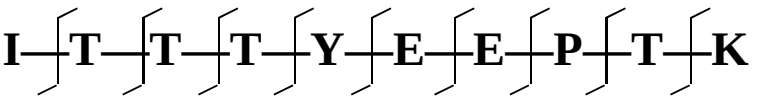
Identifying Posttranslational Modifications

- Current protein sequence prediction packages have limited ability to detect posttranslational modifications.



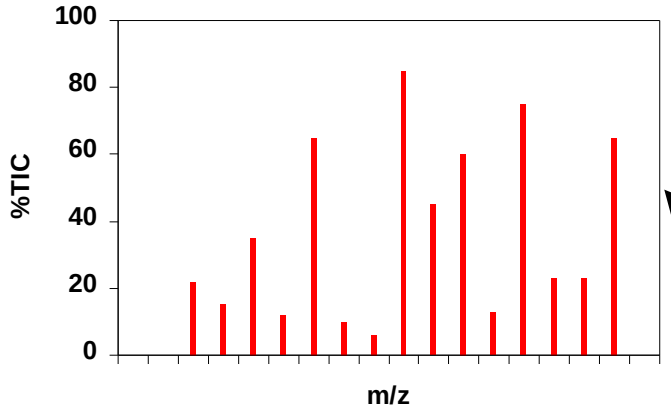
$(M+H)+ = X+m$



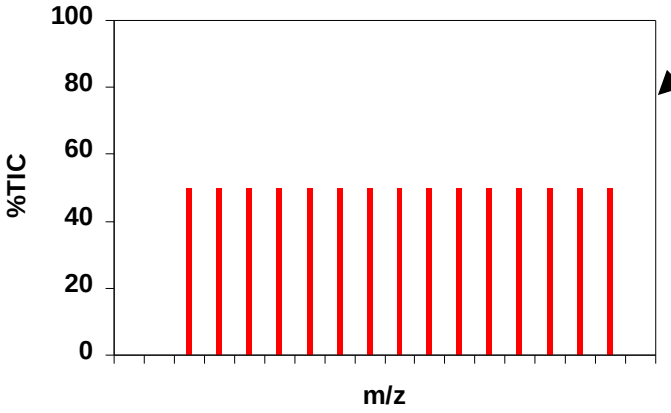


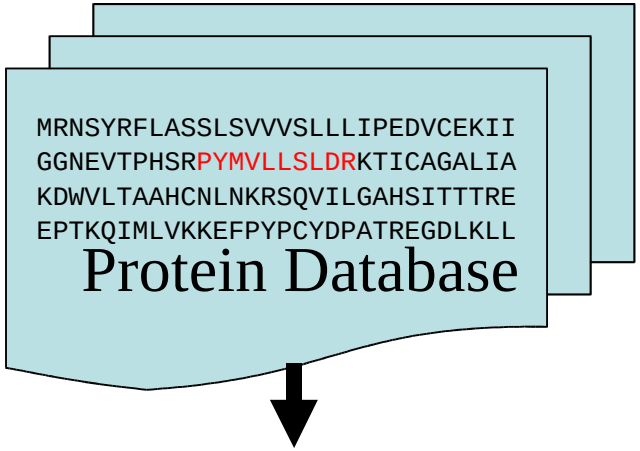
$(M+H)+ = X$

Identifying Posttranslational Modifications



P—Y—M—V—L—L—S—L—D—R





In Silico Digestion

LASSLSVVVSLLLIPEDVCEK
IIGGNEVTPHSR
PYMVLLSLDR
TICAGALIAK
DWVLTAAHCNLNKR
ITTTYEEPTK
QIMLVK
EFPYPCYDPATR
EGDLKLL

In Silico Fragmentation

Identifying Posttranslational Modifications

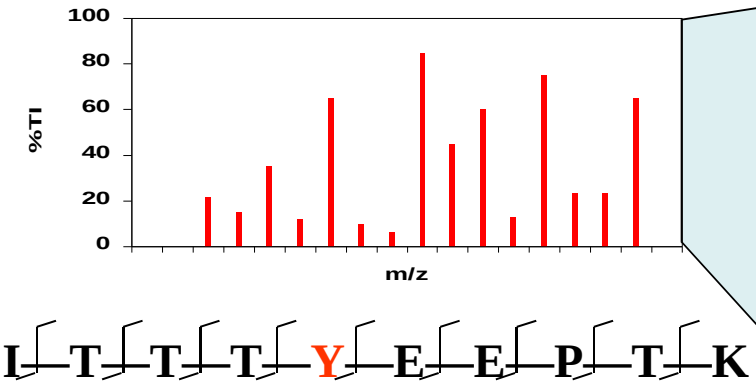
1. Assume each of the spectra corresponds to a candidate peptide.
2. Calculate the (M+H)+ difference between the spectrum and the candidate peptide and determine *m*.
3. Assign m to each residue in the sequence, in turn.
4. Calculate the hypergeometric probability of the spectrum and this candidate peptide.
5. If the match score improves over the original score, then it is more probable that the spectrum corresponds to a PTM peptide.
6. A look-up table can be used to identify the probable PTM and determine if is compatible with the modified residue.

Identifying Posttranslational Modifications

LASSLSVVVSLLLIPEDVCEK
IIGGNEVTPHSR
PYMVLLSLDR
TICAGALIAK
DWVLTAAHCNLNKR
ITTTYEEPTK
QIMLVK
EFPYPCYDPATR
EGDLKLL

PTM Candidate Peptides

b Ions	Fragment	y Ions
----------	I TTTYEEPTK	1062.54
215.14	IT TTYEEPTK	961.50
316.19	ITT TYEEPTK	860.45
417.23	ITTT YEEPTK	759.40
573.34	ITTTY EEPTK	603.30
702.38	ITTTYE EPTK	474.26
831.42	ITTTYEE PTK	345.21
928.47	ITTTYEEP TK	248.16
1029.52	ITTTYEEPT K	147.11
1175.62	ITTTYEEPTK	----------



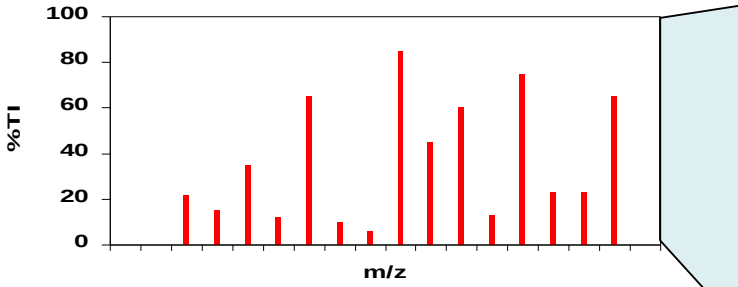
b Ions	Fragment	y Ions
----------	I TTTYEEPTK	1062.54+m
215.14	IT TTYEEPTK	961.50+m
316.19	ITT TYEEPTK	860.45+m
417.23	ITTT YEEPTK	759.40+m
573.34+m	ITTTY EEPTK	603.30
702.38+m	ITTTYE EPTK	474.26
831.42+m	ITTTYEE PTK	345.21
928.47+m	ITTTYEEP TK	248.16
1029.52+m	ITTTYEEPT K	147.11
1175.62+m	ITTTYEEPTK	----------

Identifying Posttranslational Modifications

LASSLSVVVSLLLIPEDVCEK
IIGGNEVTPHSR
PYMVLLSLDR
TICAGALIAK
DWVLTAAHCNLNKR
ITTTYEEPTK
QIMLVK
EFPYPCYDPATR
EGDLKLL

PTM Candidate Peptides

b Ions	Fragment	y Ions
----------	I TTTYEEPTK	1062.54
215.14	IT TTYEEPTK	961.50
316.19	ITT TYEEPTK	860.45
417.23	ITTT YEEPTK	759.40
573.34	ITTTY EEPTK	603.30
702.38	ITTTYE EPTK	474.26
831.42	ITTTYEE PTK	345.21
928.47	ITTTYEEP TK	248.16
1029.52	ITTTYEEPT K	147.11
1175.62	ITTTYEEPTK	----------



I—T—T—T—Y—E—E—P—T—K

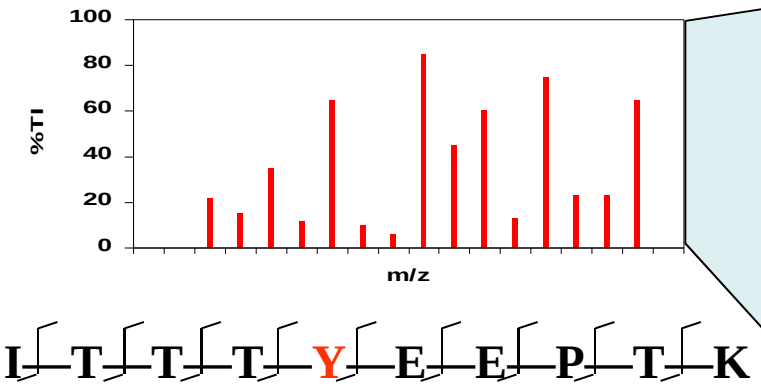
b Ions	Fragment	y Ions
----------	I TTTYEEPTK	1062.54+m
215.14	IT TTYEEPTK	961.50+m
316.19	ITT TYEEPTK	860.45+m
417.23	ITTT YEEPTK	759.40+m
573.34+m	ITTTY EEPTK	603.30
702.38+m	ITTTYE EPTK	474.26
831.42+m	ITTTYEE PTK	345.21
928.47+m	ITTTYEEP TK	248.16
1029.52+m	ITTTYEEPT K	147.11
1175.62+m	ITTTYEEPTK	----------

Identifying Posttranslational Modifications

LASSLSVVVSLLLIPEDVCEK
IIGGNEVTPHSR
PYMVLLSLDR
TICAGALIAK
DWVLTAAHCNLNKR
ITTTYEEPTK
QIMLVK
EFPYPCYDPATR
EGDLKLL

PTM Candidate Peptides

b Ions	Fragment	y Ions
----------	I TTTYEEPTK	1062.54
215.14	IT TTYEEPTK	961.50
316.19	ITT TYEEPTK	860.45
417.23	ITTT YEEPTK	759.40
573.34	ITTTY EEPTK	603.30
702.38	ITTTYE EPTK	474.26
831.42	ITTTYEE PTK	345.21
928.47	ITTTYEEP TK	248.16
1029.52	ITTTYEEPT K	147.11
1175.62	ITTTYEEPTK	----------



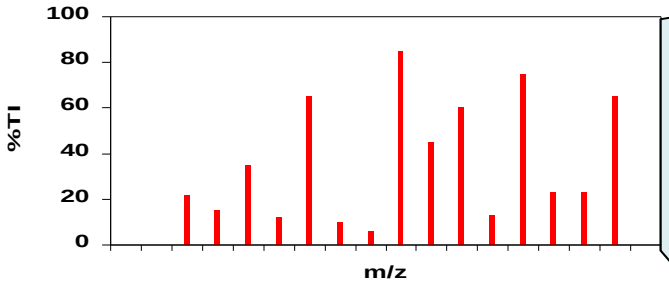
b Ions	Fragment	y Ions
----------	I TTTYEEPTK	1062.54+m
215.14	IT TTYEEPTK	961.50+m
316.19	ITT TYEEPTK	860.45+m
417.23	ITTT YEEPTK	759.40+m
573.34+m	ITTTY EEPTK	603.30
702.38+m	ITTTYE EPTK	474.26
831.42+m	ITTTYEE PTK	345.21
928.47+m	ITTTYEEP TK	248.16
1029.52+m	ITTTYEEPT K	147.11
1175.62+m	ITTTYEEPTK	----------

Identifying Posttranslational Modifications

LASSLSVVVSLLLIPEDVCEK
IIGGNEVTPHSR
PYMVLLSLDR
TICAGALIAK
DWVLTAAHCNLNKR
ITTTYEEPTK
QIMLVK
EFPYPCYDPATR
EGDLKLL

PTM Candidate Peptides

b Ions	Fragment	y Ions
----------	I TTTYEEPTK	1062.54
215.14	IT TTYEEPTK	961.50
316.19	ITT TYEEPTK	860.45
417.23	ITTT YEEPTK	759.40
573.34	ITTTY EEPTK	603.30
702.38	ITTTYE EPTK	474.26
831.42	ITTTYEE PTK	345.21
928.47	ITTTYEEP TK	248.16
1029.52	ITTTYEEPT K	147.11
1175.62	ITTTYEEPTK	----------



I–T–T–T–Y–E–E–P–T–K

(M+H)+ = 1175.62+m

b Ions	Fragment	y Ions
----------	I TTTYEEPTK	1062.54+m
215.14	IT TTYEEPTK	961.50+m
316.19	ITT TYEEPTK	860.45+m
417.23	ITTT YEEPTK	759.40+m
573.34+m	ITTTY EEPTK	603.30
702.38+m	ITTTYE EPTK	474.26
831.42+m	ITTTYEE PTK	345.21
928.47+m	ITTTYEEP TK	248.16
1029.52+m	ITTTYEEPT K	147.11
1175.62+m	ITTTYEEPTK	----------