A Software Application for MS Data Processing and Post Translational Modification Database Generation

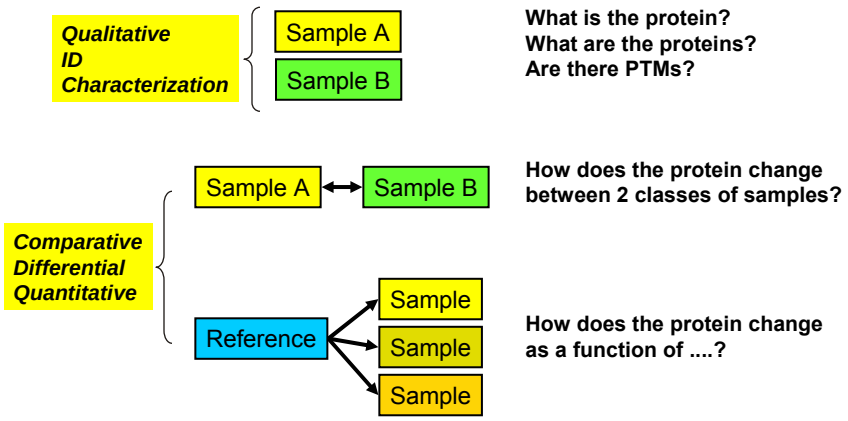
James West, Hua Huang, Weiwei Tong, Yang Su, David H. Perlman, Catherine E. Costello, Mark E. McComb

Cardiovascular Proteomics Center
Boston University School of Medicine
Boston, MA

MS and Proteomics

- Mass spectrometry is widely used for characterization of protein primary structure and structural changes, such as sequence mutations and post-translational modifications.
- It is now possible to apply MS to protein based diseases
- Mass spectrometry (protein based)
  - Speed
  - Sensitivity
  - Direct protein characterization
  - Post-translational modifications
  - Unambiguous sequence determination
  - A one size fits all methodology
- Advent of new instrumentation
  - high resolution, high mass accuracy, robustness, cheaper
  - advent of friendly instrumentation
- Advent of powerful data analysis tools
  - able to rapidly process large and complex data sets correctly
  - advent of friendly data analysis tools

Protein Identification & Protein Quantitation



What do we want to know? What are we looking for?

MS and MS/MS Information

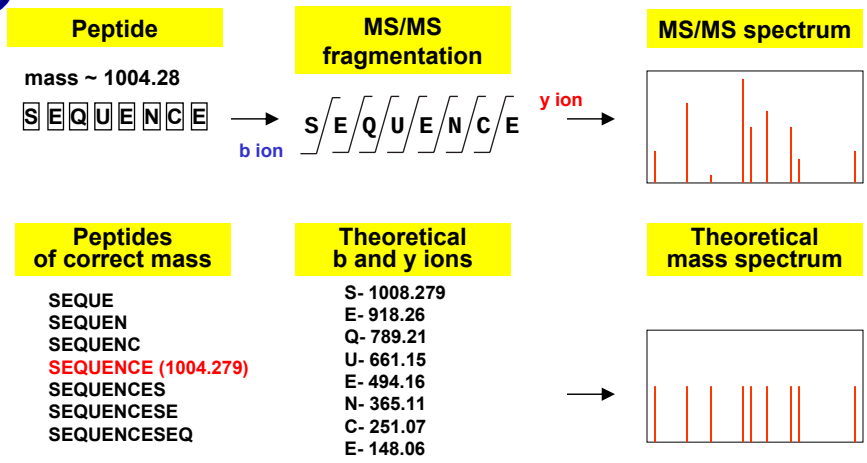
- Qualitative information → *m/z*: *tentative*: ID, map, % coverage, PTMs
- Quantitative information → %I
- *m/z*: Instrument dependent: Trap vs QTOF vs FT
  - Accuracy: Resolution: Dynamic range
- %I: MALDI vs ESI
  - MS %I → chemical constitution of peptide
  - PTM may be lost in MALDI
  - LC > TOF, FT for quantitation
  - ICAT → quantitation via relative abundance
- Normal vs Control
- ID via map: yes/no expression
- %I: differential expression
- Variant: differential expression
- PTM: yes/no, ID, location
- %PTM: differential expression

	TOF	QTOF	FT
R	7k	10k	50k
ppm	100	10	5
S/N	10^3	10^5	10^3
pmol	.001	.200	1

Protein and Gene Databases

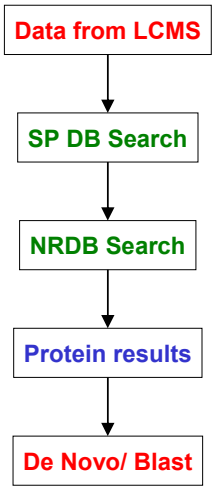
- SWISSPROT and TREMBL
  - Small and highly curated protein knowledge and sequence database
  - Manually confirmed sequences.
- NCBI non-redundant (NCBInr)
  - Non-redundant database from the National Center for Biotechnology Information
  - Used for BLAST and Entrez
  - Comprised of translated sequences from the Genbank, SwissProt, Protein Information Resource (PIR), and Brookhaven Protein Data Bank (PDB)
- EST Clusters (dBEST)
  - "single-pass" cDNA sequences, or Expressed Sequence Tags (EST's)
  - Short, usually 3' end sequences from isolated mRNA.
  - expressed sequences (no introns) including splice variants
  - Short lengths, redundancy, and low quality affect results
  - DB search: nucleic acid sequences are translated in all 6 reading frames
  - dbEST is a very large database: EST_human, mouse, others

Protein Identification by MS/MS Probability Match Database Search



- a DB search will search for precursor peptide masses
- then align theoretical b/y ion masses with observed masses
- a probability algorithm will determine if the match is good
- %I for b/y ions NOT calculated!

Iterative MS/MS Data Analysis Approach



MS/MS spectra: many experiments

Standard search: trypsin, 1 missed site
- first pass to id components remove "other" components

Standard search: trypsin, 1 missed site
- first pass to id components remove "other" components

Protein results database
- re-search other enzymes, non-specific cleavage
- increased sequence coverage

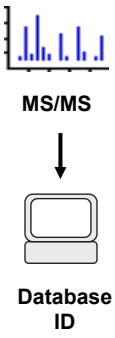
Second pass to ID novel proteins

Protein identification based on "probabilities"
Typical DB search yields 30% return on a data set
Ambiguous results ⇒ repeat to reduce complexity of data

Proteomics: Intelligent data analysis

- Typical LC-MS/MS experiment
  - several hundred MS and MSMS peaks
  - data analysis possible but time consuming
- *smart* MS/MS
  - complete sequence coverage
  - conclusive location and ID of variant
  - location and identification of PTMs
- Proteomics and automated data interpretation
  - Protein identification
  - Protein sequencing
  - denovo sequencing
  - Blast homology
  - PTM characterization

increasing difficulty



Proteomics

Protein identification ⟷ *Protein characterization*

Post-translational Modifications (PTMs)

- PTMs change the way proteins behave and interact with other proteins in the cell.
- While some PTMs have been identified and are well know to occur, many PTMs remain unknown and so far undetected. These PTMs may be involved in many disease states and therefore their identification is important.
- Identification of PTMs poses a complex problem for researchers principally due to a lack of methodology and therefore many unknown PTMs are only discovered by chance.
- MS based proteomics is only just beginning to be applied to the characterization of post-translational modifications of proteins.
- With accurate mass tandem mass spectrometry (MS/MS) data obtained from LC-MS/MS experiments of proteolytic digests of proteins, it is possible to characterize a limited number of known PTMs. However, for the discovery of unknown PTMs or the identification of unknown localization sites, the use of a standard PTM database fails.

Database Design: for targeting PTMs

- Large non-redundant protein database:
  - >1,000,000 proteins
  - many false positives for single protein analysis
  - assignments forced to tryptic peptides
  - false negative results for PTMs
- Single protein database:
  - single amino acid "*corrected sequence*"
  - search the digital dataset against the experimental data
  - eliminate false negatives
  - improved positive results for PTMs
- Single peptide database
  - ie: >gi|TC1-15| Mass 1533 → GPTGTGESB**C**PLMVB
  - with and without PTMs preprogrammed
  - *maximum results*

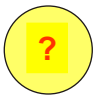
Construct custom databases to solve custom datasets
Single protein/peptide databases target PTMs

Characterization of PTMs Via MS and MS/MS

- **Post-translational-modifications**
  - 100s known and unknown PTMs are possible
  - multitude of locations feasible
  - dynamic range
  - temporal probability ⇒ biochemical processes

[Protein] x 1 to $1E^{-2}$

- **Characterization**
  - simple amino acid substitution matrices are insufficient
  - PTM databases ⇒ knowledge of the sample is required
  - PTM follow biochemical rules



- **Catalog of known PTMs**
  - RESID Database >800 known PTMs !
- **Create a PTM Database**
  - Nominal mass
  - Exact mass

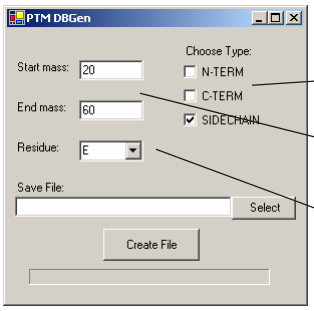
PTM	change	ΔMass
S-Cysteinyl	C3H5NO2S	119.0041
unknown	CnHxNyOzSm	n,x,y,z,m>1

Validation: Are the results real? ⇒ *in vitro* modification
Fortunately PTMs follow some rules ⇒ analysis possible

Nominal Mass and Exact Mass Databases

- We have constructed a database of nominal mass and exact mass PTMs and have applied a series of successive searches to determine the feasibility of using this approach for PTM characterization.
- The PTM Database Generator (PTM-DBGen) software application was developed using Microsoft Visual C# .NET utilizing the Microsoft .NET Framework 1.1 (version 1.1.4322 SP1) in the Microsoft Visual Studio .NET 2003 Development Environment and was used to generate these custom nominal mass PTM databases in an XML format that ProteinLynx Global Server 2.2 (Waters Corporation) can use.
- LC-MS and LC-MS/MS data sets that were obtained in house and derived from a large number of ongoing projects were processed using ProteinLynx Global Server 2.2 and the results were matched against the SwissProt database and the custom generated protein databases from PTM-DBGen.

PTM-DBGen



Interface of the PTM-DBGen application.

Check boxes allows the user to indicate the location of the PTM on the peptide.

Fields allow the user to designate starting and ending nominal masses.

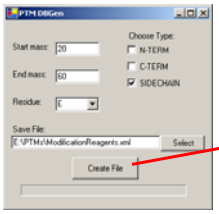
Dropdown box prevents the user from mistakenly generating a database with multiple residues.

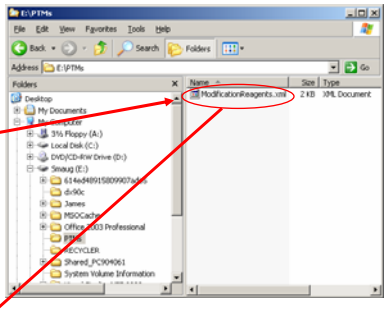
Example Single Amino Acid Database generated by PTM-DBGen

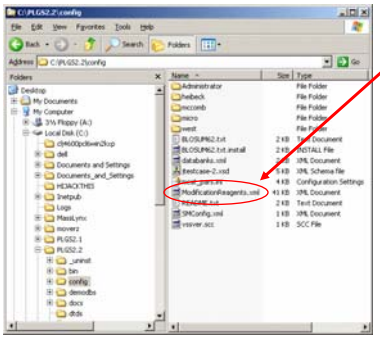
```
<MODIFICATION_REAGENTS>
        <MODIFIER NAME="SIDECHAIN NOMINAL MASS = 20 RESIDUE = E" MCAT_REAGENT="FALSE">
                <MODIFIES APPLIES_TO="E" DELTA_MASS="20" TYPE="SIDECHAIN">
                </MODIFIES>
        </MODIFIER>
        <MODIFIER NAME="SIDECHAIN NOMINAL MASS = 21 RESIDUE = E" MCAT_REAGENT="FALSE">
                <MODIFIES APPLIES_TO="E" DELTA_MASS="21" TYPE="SIDECHAIN">
                </MODIFIES>
        </MODIFIER>
        <MODIFIER NAME="SIDECHAIN NOMINAL MASS = 22 RESIDUE = E" MCAT_REAGENT="FALSE">
                <MODIFIES APPLIES_TO="E" DELTA_MASS="22" TYPE="SIDECHAIN">
                </MODIFIES>
        </MODIFIER>
        <MODIFIER NAME="SIDECHAIN NOMINAL MASS = 23 RESIDUE = E" MCAT_REAGENT="FALSE">
                <MODIFIES APPLIES_TO="E" DELTA_MASS="23" TYPE="SIDECHAIN">
                </MODIFIES>
        </MODIFIER>
...
```

Replacing the ProteinLynx Global Server 2.2 PTM Database

1. Generate the custom PTM database.



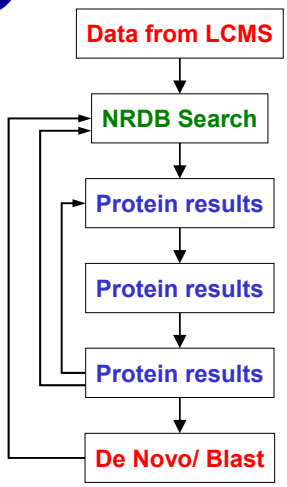




2. Replace the existing ModificationReagents.xml file for ProteinLynx Global Server 2.2 with the new custom PTM database.

3. Do another search in ProteinLynx Global Server 2.2 looking only for PTMs.

Iterative MS Data Analysis Approach



MS/MS spectra: many experiments

Standard search: trypsin, 1 missed site
first pass to id components remove "other" components

Other cleavages and PAM AA matrix
increased sequence coverage, amino acid substitutions

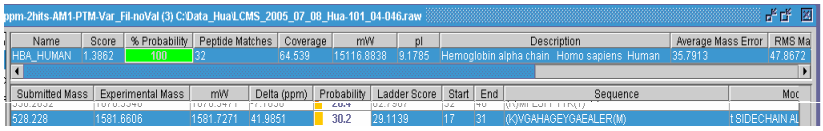
First pass PTM analysis
known PTMs and known artifacts

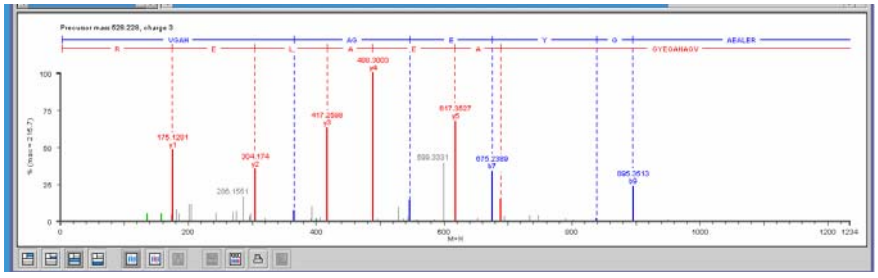
Second pass PTM analysis
new post translational modifications

new results

Correlate results to sequence sequentially
Ambiguous results ⇒ repeat to reduce complexity of data

Example of LC MS/MS of $[M+2H]^{2+}$ = 791.845 *m/z*





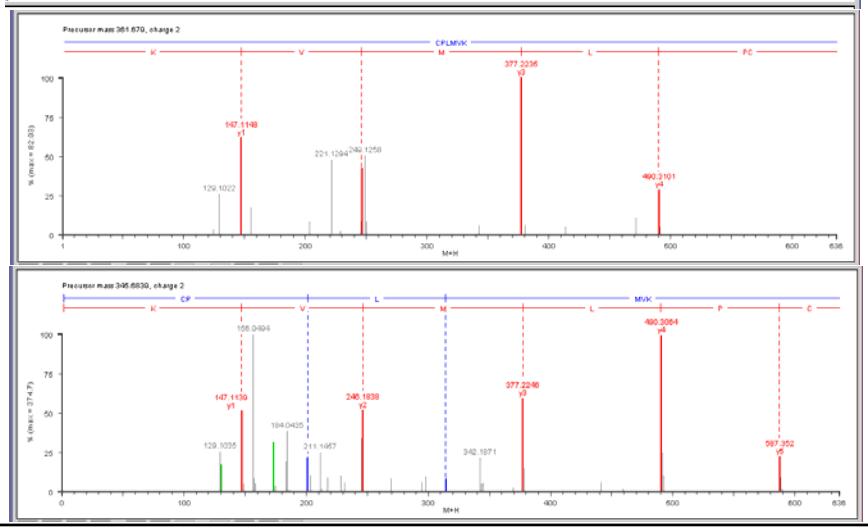
Example of the database search results obtained from the LC MS/MS analysis of a tryptic digest of human hemoglobin. Data were search against the hemoglobin database with successive addition of nominal mass PTMs. The result here shows the MS/MS confirmation of the +53 mass PTM on the beta chain tryptic peptide 17-31.

PTM Database Search Results from a Hemoglobin Sample

- 403 MS/MS returned results from the database search
- 201 results returned with a probability > 0
  - 0 is the cutoff for probability in PLGS 2.2
  - ~ 50% false positive
- 111 results returned with a ladder score > 14
  - 14 is the cutoff we use for the ladder score in PLGS 2.2
  - ~ 75% false positive
- Manual confirmation identified over 100 assignable MS/SM spectra
- 27 PTMs were observed including dehydration, deamidation, oxidation and 2 novel PTMs which were previously not seen
- The example here shows a PTM of Δmass = 53 which we would not normally expect: this mass difference is due to iron adduction on the hemoglobin molecule

LC MS/MS: oxidized and nominal peptide of TTR

Submitted Mass	mW	Delta (ppm)	Probability	Ladder Score	Start	End	Sequence	Modifications
390.6735	779.3604	37.2791	21.4	48	10	15	(K)CPLMVK(V)	SIDECHAIN NOMINAL MASS = 90 RESIDUE = C
361.679	721.3604	25.0455	21	56	10	15	(K)CPLMVK(V)	SIDECHAIN NOMINAL MASS = 32 RESIDUE = C
345.6839	689.3604	11.9529	6.8	92	10	15	(K)CPLMVK(V)	



Typical Results Table for TTR Protein

m/z	Exp. mW	mW	Δ(ppm)	Prob.	L-Score	Start	End	Sequence	Modifications	Subs.	OK
345.684	689.352	689.360	12	67	92	10	15	(K)CPLMVK(V)			y
361.679	721.342	721.360	25	21	56	10	15	(K)CPLMVK(V)	SC Nom. MASS = 32 AA = (1)		y
390.674	779.331	779.360	37	21	48	10	15	(K)CPLMVK(V)	SC Nom. MASS = 90 AA = (1)		y
526.274	1050.532	1050.540	8	7	2.0408	16	25	(K)VLDAVRGSPA(I)	SC Nom. MASS = 67 AA = (9)		y
482.757	963.498	963.514	16	44	46.9388	22	31	(R)GSPAINVAVH(V)			y
590.264	1178.512	1178.524	10	34	36.7347	35	44	(R)KAADDTWEPF(A)			y
526.220	1050.424	1050.429	5	52	39.5349	36	44	(K)AADDTWEPF(A)			y
537.206	1072.396	1072.429	31	10	11.6279	36	44	(K)AADDTWEPF(A)	SC Nom. MASS = 22 AA = (7)		y
697.811	1393.606	1393.615	6	90	70.1493	36	48	(K)AADDTWEPFASGK(T)			y
713.802	1425.589	1425.615	18	18	28.3582	36	48	(K)AADDTWEPFASGK(T)	SC Nom. Mass = 32 AA = (8)		y
713.804	1425.592	1425.656	45	11	26.8657	36	48	(K)AAFDTWEPFASGK(T)		F for D (3)	y
515.240	1028.465	1028.477	12	63	65.3061	49	58	(K)TSESGELHGL(T)			y
616.287	1230.559	1230.573	12	57	50.8197	49	60	(K)TSESGELHGLTT(E)			y
680.808	1359.601	1359.615	10	52	49.2537	49	61	(K)TSESGELHGLTTE(E)			y
482.223	1443.645	1443.677	22	50	36.0656	59	70	(L)TTEEEFVEGIYK(V)			y
722.840	1443.664	1443.677	9	74	59.0164	59	70	(L)TTEEEFVEGIYK(V)			y
557.748	1113.480	1113.487	6	25	11.6279	61	69	(T)EEEFVEGIY(K)			y
621.791	1241.566	1241.582	13	63	71.4286	61	70	(T)EEEFVEGIYK(V)			y
557.273	1112.530	1112.539	8	66	58.1395	62	70	(E)EEFVEGIYK(V)			y
352.701	703.387	703.390	5	55	61.2903	81	87	(K)ALGISPF(H)			y
485.746	969.476	969.492	16	33	58.1395	81	89	(K)ALGISPFHE(H)			y
484.232	966.449	966.456	7	62	37.8378	88	95	(F)HEHAEVVF(T)			y
634.326	1266.636	1266.646	8	55	62.2951	115	126	(Y)SYSTTAVVTNPK(E)			y
509.276	1016.537	1016.550	13	17	34.6939	117	126	(Y)STTAVVTNPK(E)			y

Results Table Discussion

- The table is an example of typical results from a series of database searches using multiple PTM databases.
- Each PTM database allowed for 1 PTM on a single amino acid. PTM databases of this sort were used for all 20 amino acids, A thru Y.
- The table is truncated: replicate positive results and all negative results (score < correct score) were removed for clarity.
- Oxidation as well as other unknown PTMs are seen in this table.

The TTR database that we used

>sp(P02766) TTHY_HUMAN Transthyretin precursor Prealbumin corrected
GPTGTGESKCPLMVKVLDAVRGSPAINVAVHVFRKAADDTWEPFASGKTSESGELHGLTT
EEEFVEGIYKVEIDTKSYWKALGISPFHEHAEVVFTANDSGPRRYTIAALLSPYSYSTTA
VVTNPKE

>sp(P02753) Retinol binding protein holoform corrected
ERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGR
VRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSC
RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSER
NL

>sp(P00761) TRYP_PIG Trypsin precursor
FPTDDDDKIVGGYTCAANSIPYQVSLNSGSHFCGGSLINSQWVVSAAHCYKSRIQVRLGE
HNIDVLEGNEQFINAAKIITHPNFNGNTLDNDIMLIKLSSPATLNSRVATVSLPRSCAAA
GTECLISGWGNTKSSGSSYPSLLQCLKAPVLSDSSCKSSYPGQITGNMICVGFLEGGKDS
CQGDSGGPVVCNGQLQGIVSWGYGCAQKNKPGVYTKVCNYVNWIQQTIAAN

Discussion

- **Results included a significant number of false positives**
  - **Scores generated from the probability and ladder scores**
  - **Scores assigned by ProteinLynx Global Server 2.2**
  - **Scores for all false positives were below a determined acceptable range**
  - **All false positives could thus be easily filtered out**
- **The result set was manually inspected**
  - **Contained several peaks that had previously been missed by ProteinLynx Global Server.**
- **This test resulted in a manageable number of false positives**
  - **Allowed for the identification of several PTMs that had previously been missed as false negatives.**

Summary

- Amino acid variants always observed
  - Real and DB sequence errors
- PTMs always observed
  - Redox induced, enzymatic, signaling, etc.
- Targeted database search improves results
  - search faster
  - better orientation forces positive match
  - easier to handle the results
- Accurate mass and high resolution
  - Increases likelihood of accurate results
- Dynamic range and sensitivity
  - Better MS/MS of low abundance PTMs

Acknowledgements

- We are thankful to Waters Corporation for support with PLGS
- This project was funded by NIH grants P41-RR10888 and NHLBI contract N01-HV-28178.
- Thanks to all the members of the Mass Spectrometry Resource and the Cardiovascular Proteomics Center.