William R. Wiley EMSL
Environmental Molecular Sciences Laboratory

Proteomic Analysis using Accurate Mass Tags

Gordon Anderson
PNNL
January 4-5, 2005

Outline

- **Accurate Mass and Time Tag (AMT) based proteomics**
- Instrumentation
- Data analysis
- Data management
- Challenges

Approach for Proteome Analysis

Cell lysis

↓

Protein extraction/fractionation

↓

Optional protein stable-isotope labeling

↓

Proteolytic digestion

↓

Capillary separation on-line with ESI-FTICR

↓

Data processing and display

PNNL Accurate Mass and Time Tag Approach

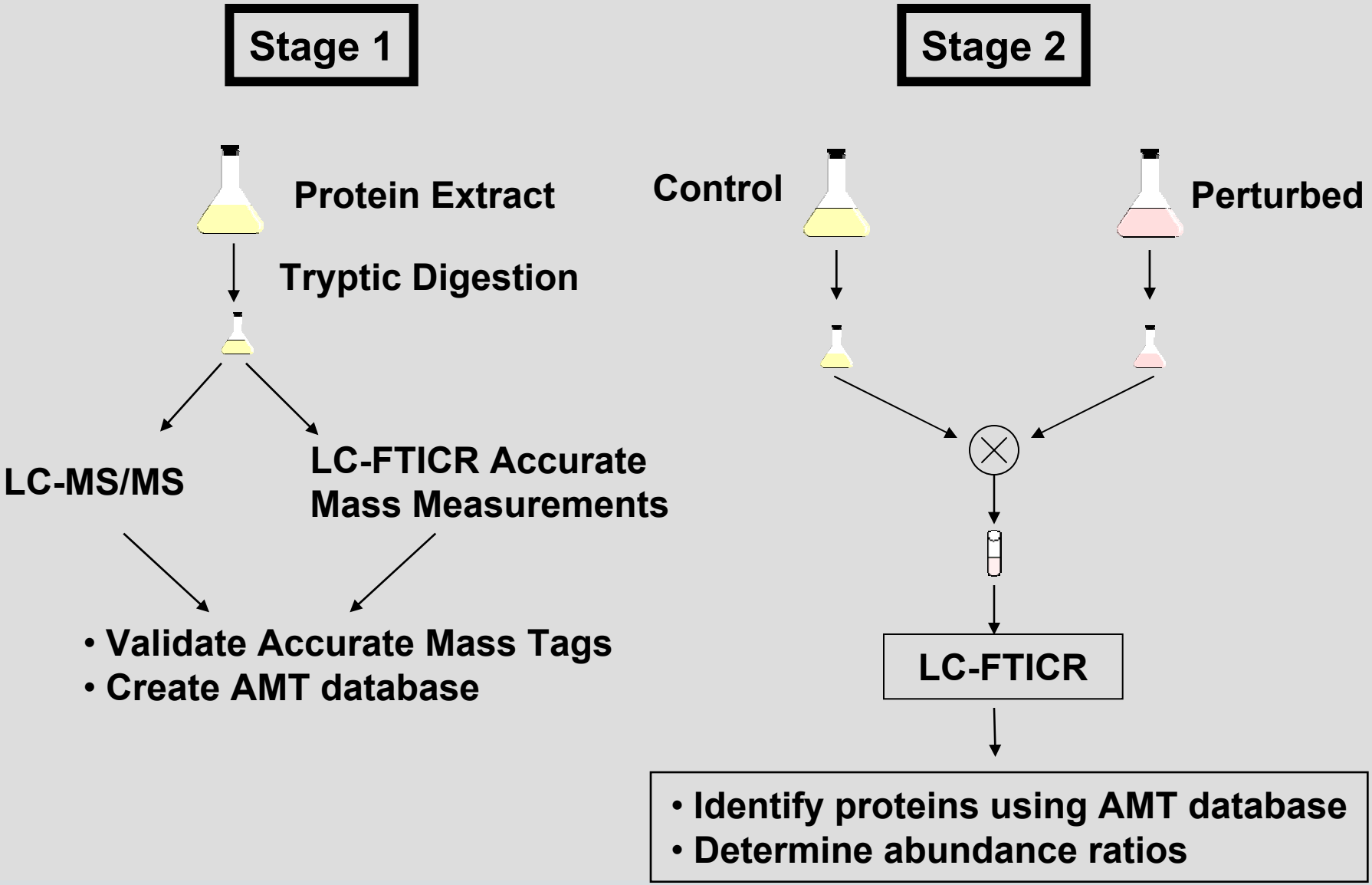
Given the constraint of a sequenced genome, high accuracy mass measurements by FTICR provide unique "biomarkers" for nearly all proteins :

Accurate Mass and Time Tags (AMTs)

Use of AMTs for protein identification

- Subsequent MS/MS not required
- Enables high throughput proteomics

Overall Methodology



...APEEKAR**GITINTAHVEYQTETR**HYSHVDCPGHADYVK...

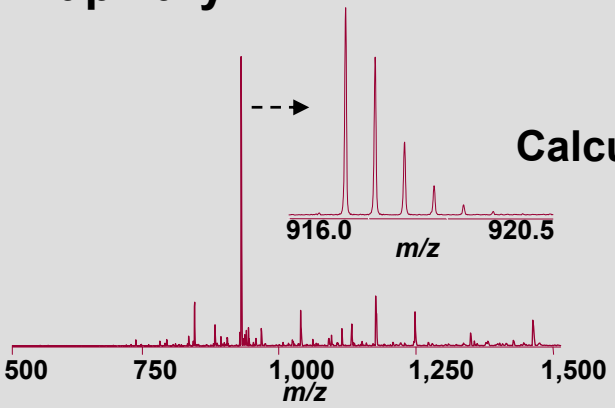
Fragment from Elongation factor Tu ↓ **Tryptic digestion**

APEEKAR **GITINTAHVEYQTETR** HYSHVDCPGHADYVK

↓

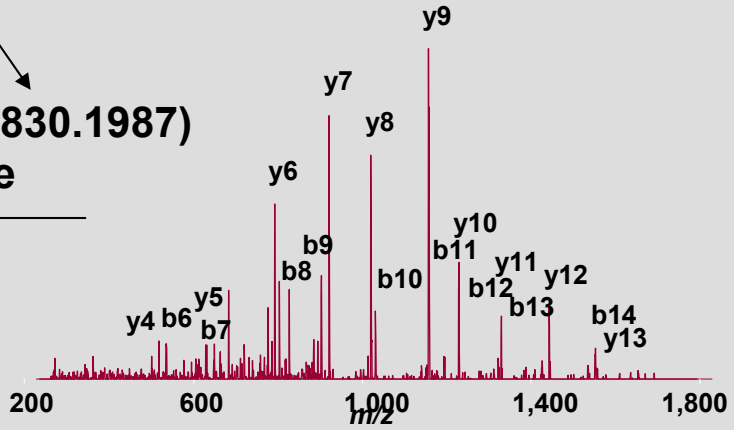
GITINTAHVEYQTETR

Capillary LC- FTICR



Calculated PMT mass (1830.1987)
***and* elution time**

Capillary LC-MS/MS (e.g. with ion trap)



Accurate mass observed: 1830.1987

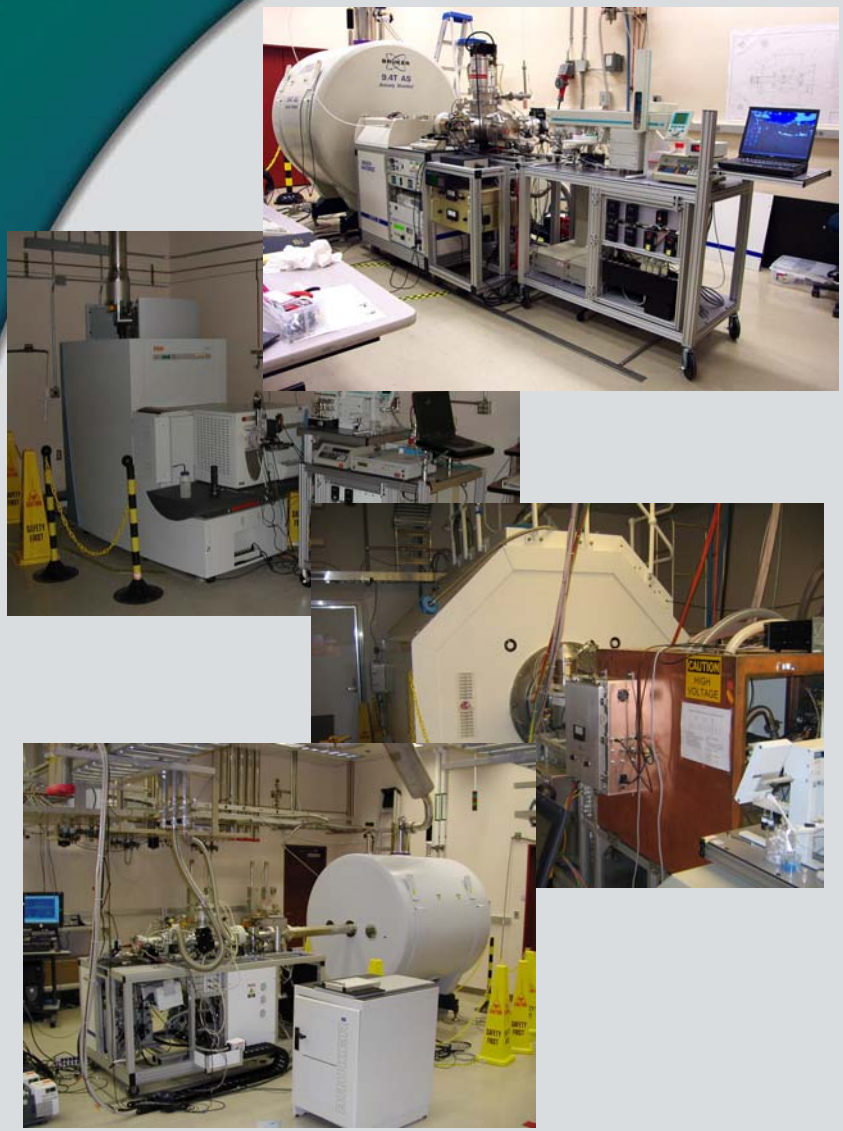
PMT identified from *D. radiodurans*:
GITINTAHVEYQTETR

Validated AMT (GITINTAHVEYQTETR)

Outline

- Accurate Mass and Time Tag (AMT) based proteomics
- **Instrumentation**
- Data analysis
- Data management
- Challenges

FTICR/TOF Mass Spectrometers



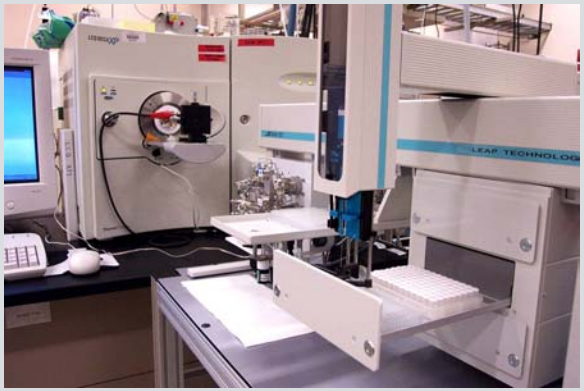
Raw data formats
- FTICR
  - Bruker
  - Finnigan odyssey
  - LTQ-FT
- TOF
  - Agilent
  - Waters

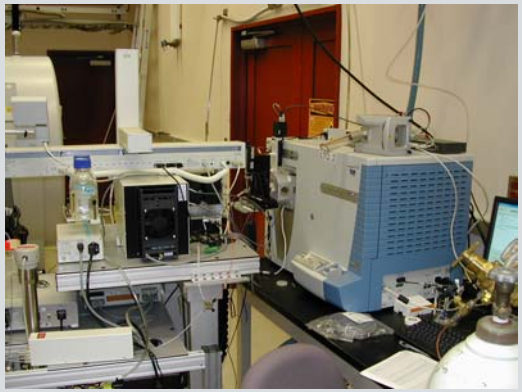
Data analysis
- De-isotoping
- In house developed software

Large data volumes

Standards needs
- Raw spectrum
  - Instrument settings
- Analysis results
- Analysis parameters

ION Trap Mass Spectrometer



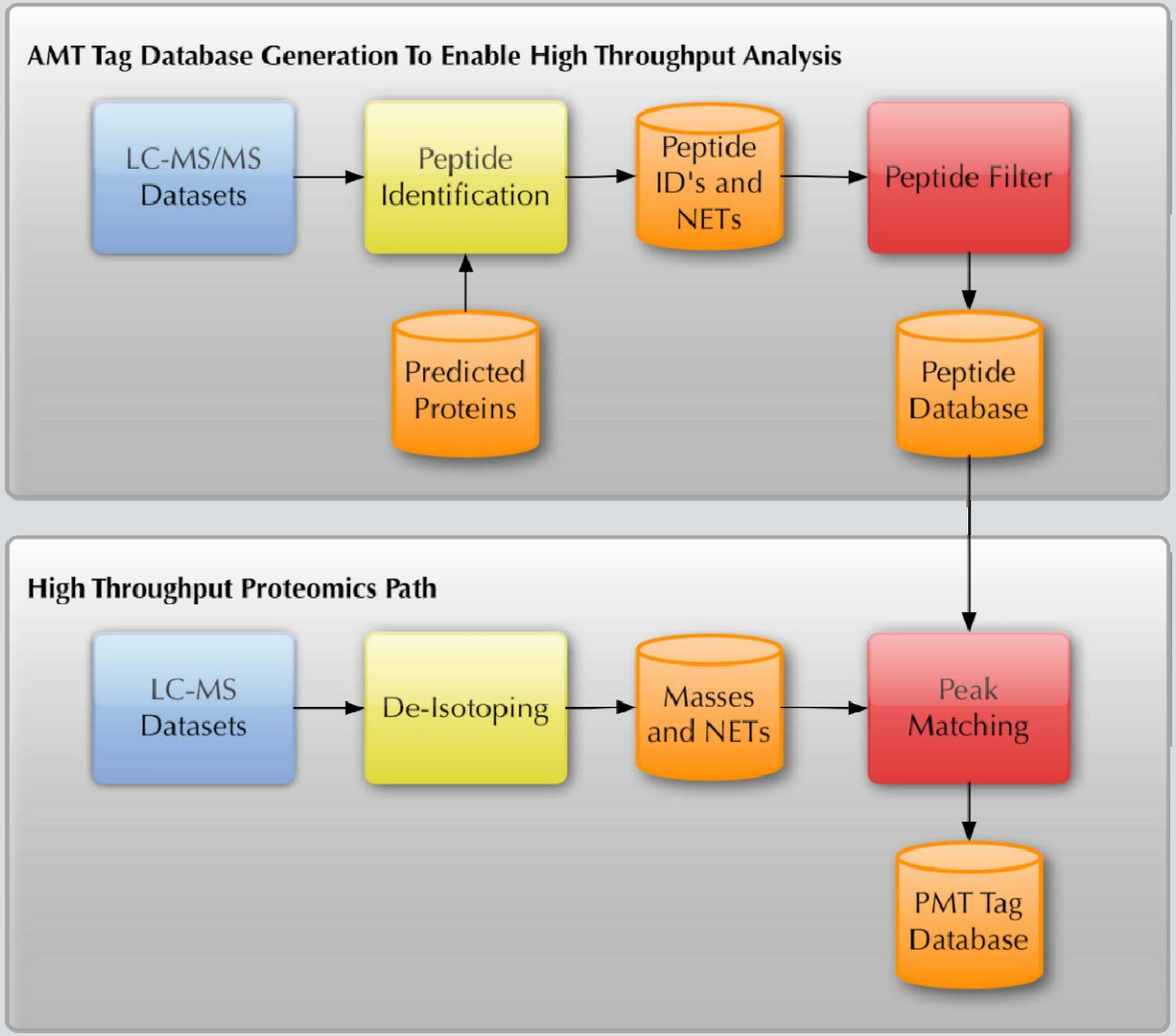


- Raw data formats
  - Thermo LCQ
  - Thermo LTQ
  - Agilent ION trap
- Data analysis
  - Peptide ID
    - SEQUEST
    - MASCOT
    - In house developed tools
- Standards needs
  - Raw spectrum
    - Instrument settings
  - Analysis results
    - Score normalization
  - Analysis parameters

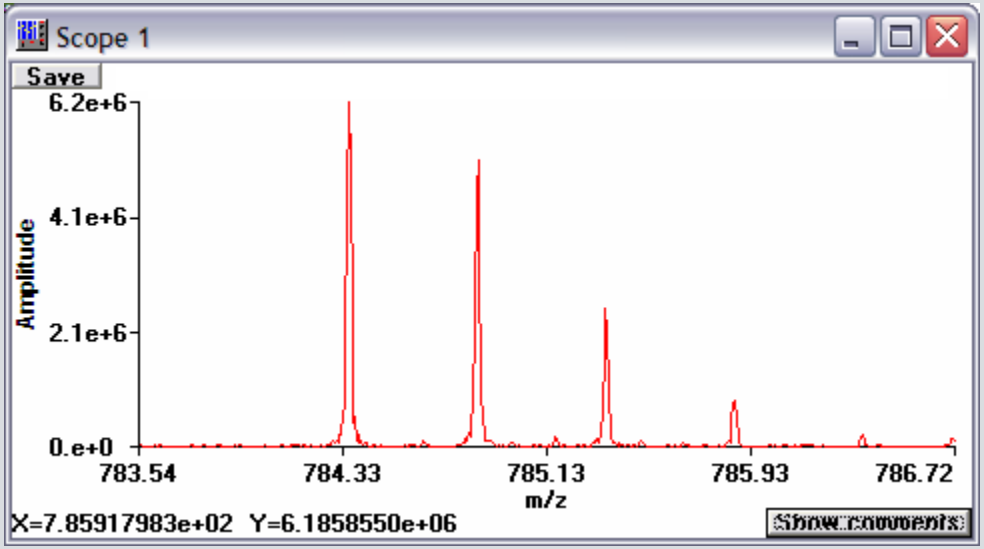
Outline

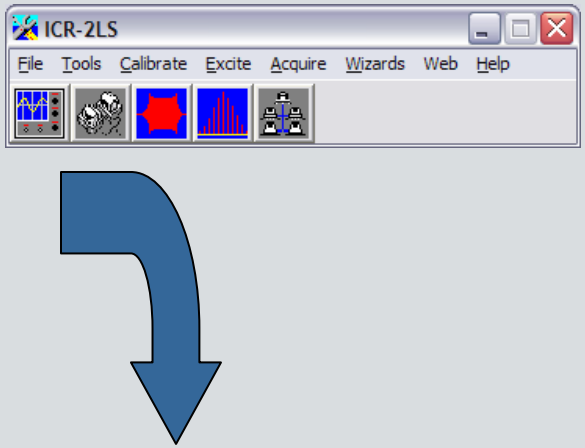
- Accurate Mass and Time Tag (AMT) based proteomics
- Instrumentation
- **Data analysis**
- Data management
- Challenges

Proteomics Analysis Pipeline



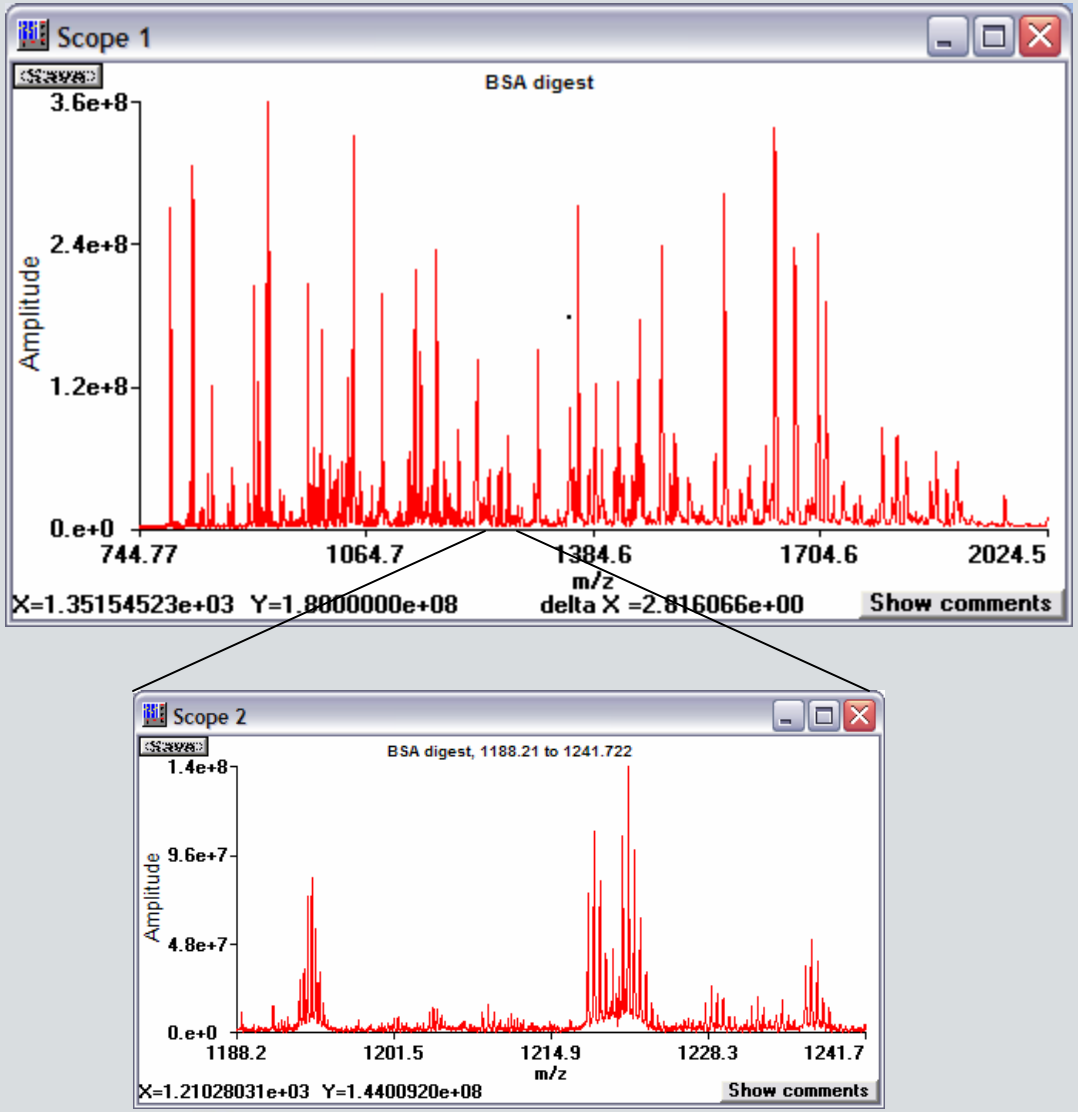
High resolution mass spectrum analysis



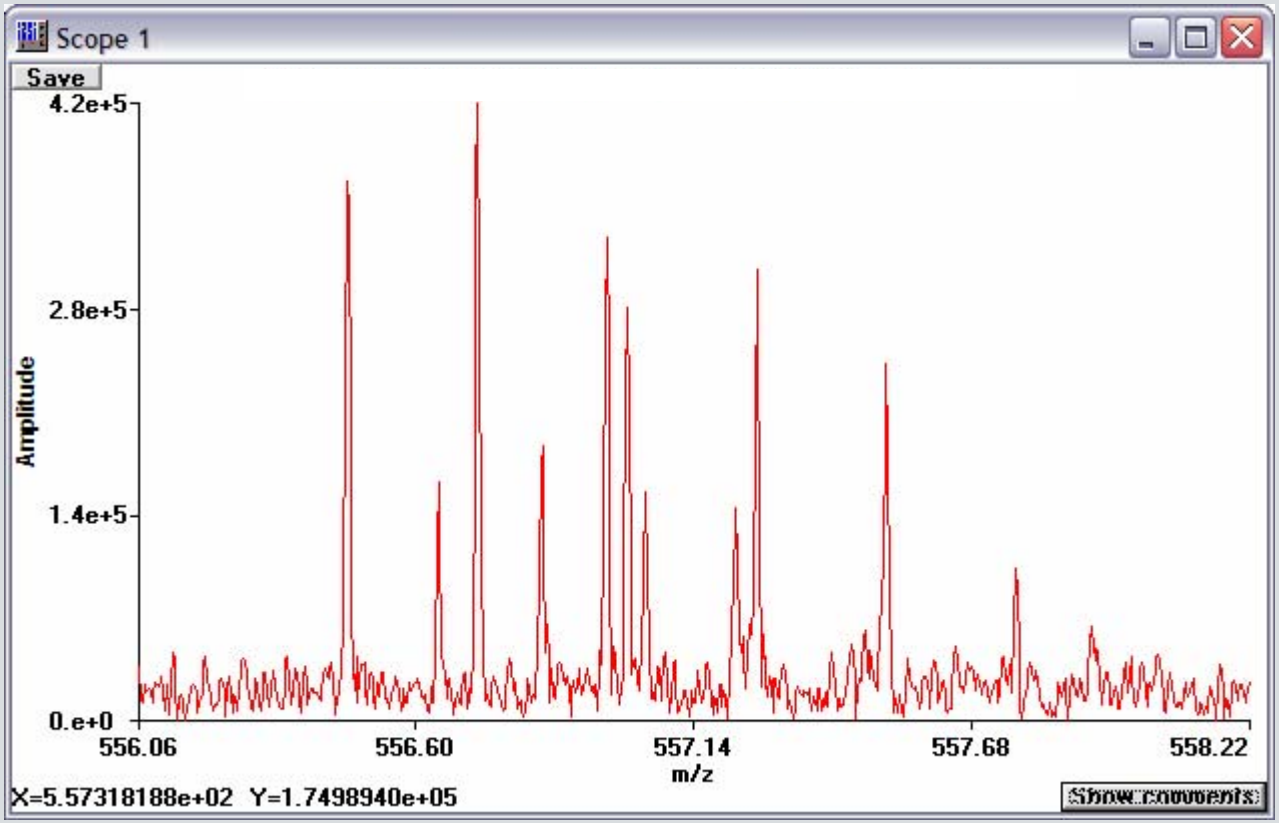


CS	Abundance	m/z	Fit	Average MW	Monoisotopic MW	Most abundant MW
2	6.19E+06	784.366	0.006	1567.764	1566.7175	1566.7175

High Resolution MS Data

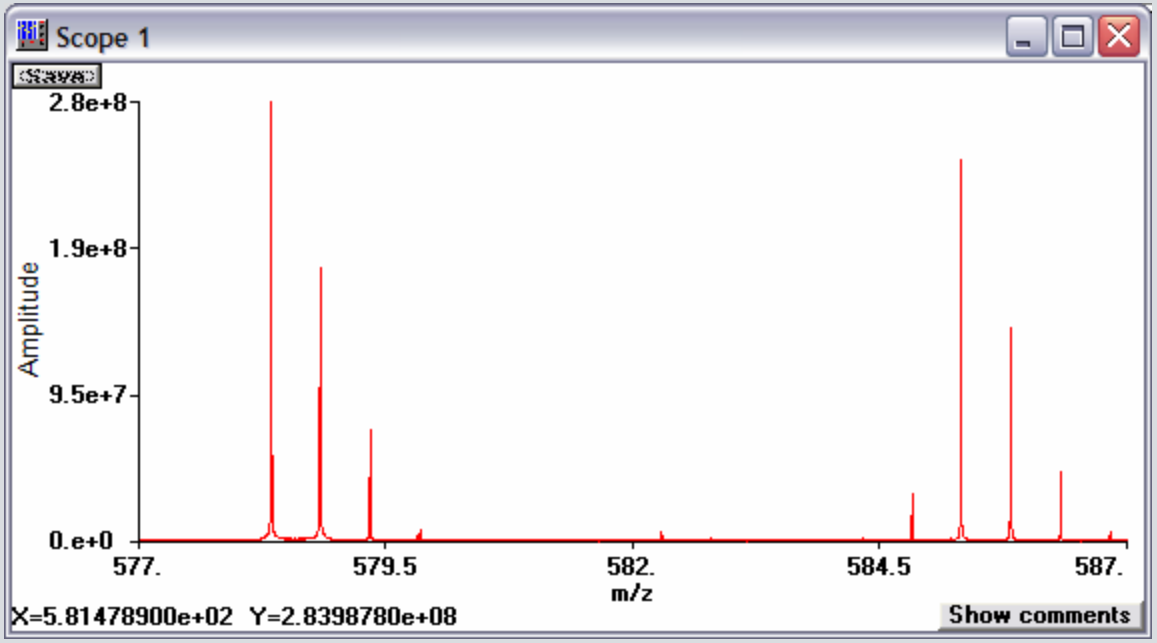


Overlapping Isotopic Distributions



CS	Abundance	m/z	Fit	Average MW	Monoisotopic MW	Most abundant MW
4	3.70E+05	556.4735	0.021	2223.3104	2221.8649	2222.8679
4	3.10E+05	557.2672	0.026	2225.4828	2224.0367	2225.0397
5	1.89E+05	556.8497	0.036	2779.9962	2778.2092	2779.2121

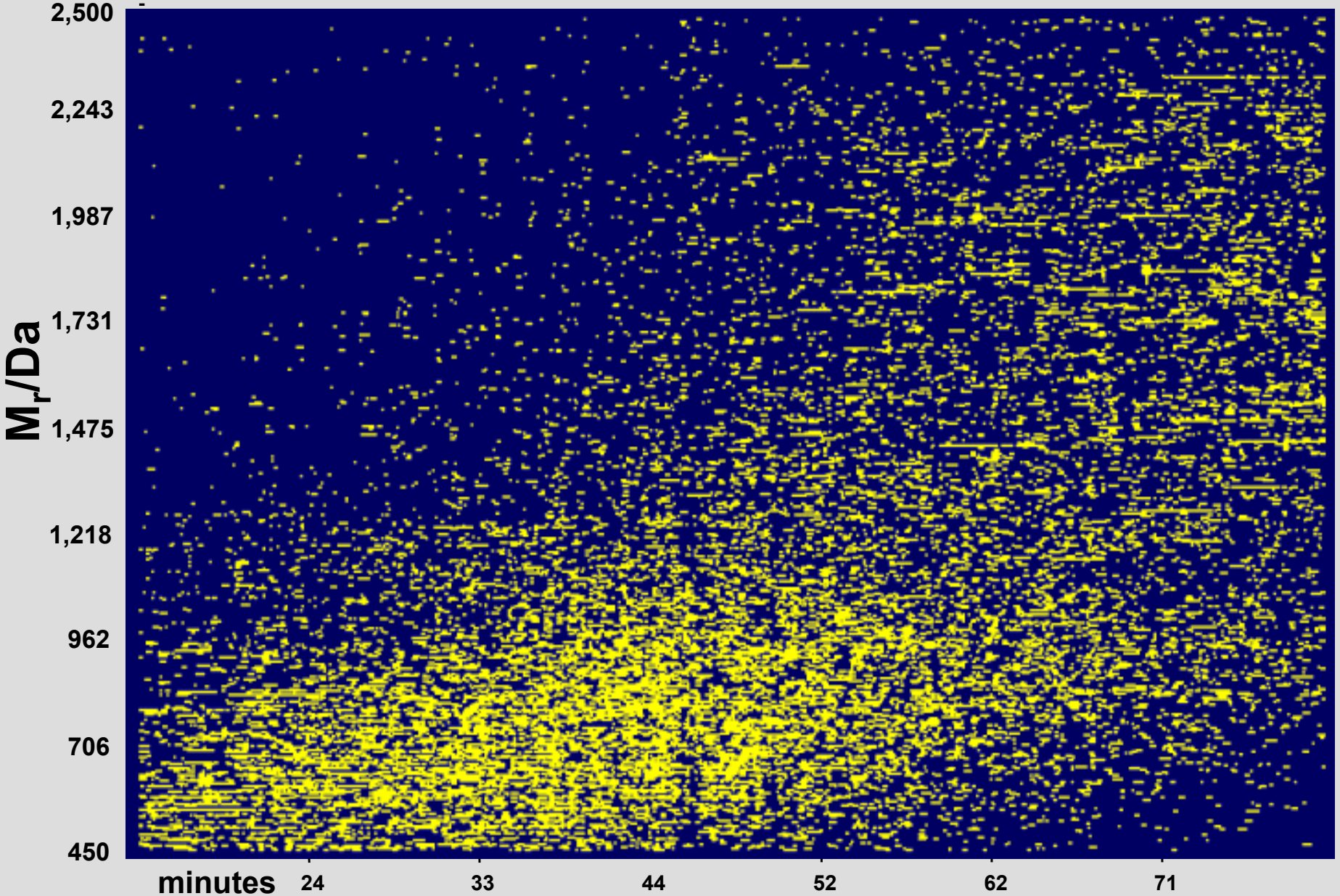
Isotopic signature Identification, N14/N15



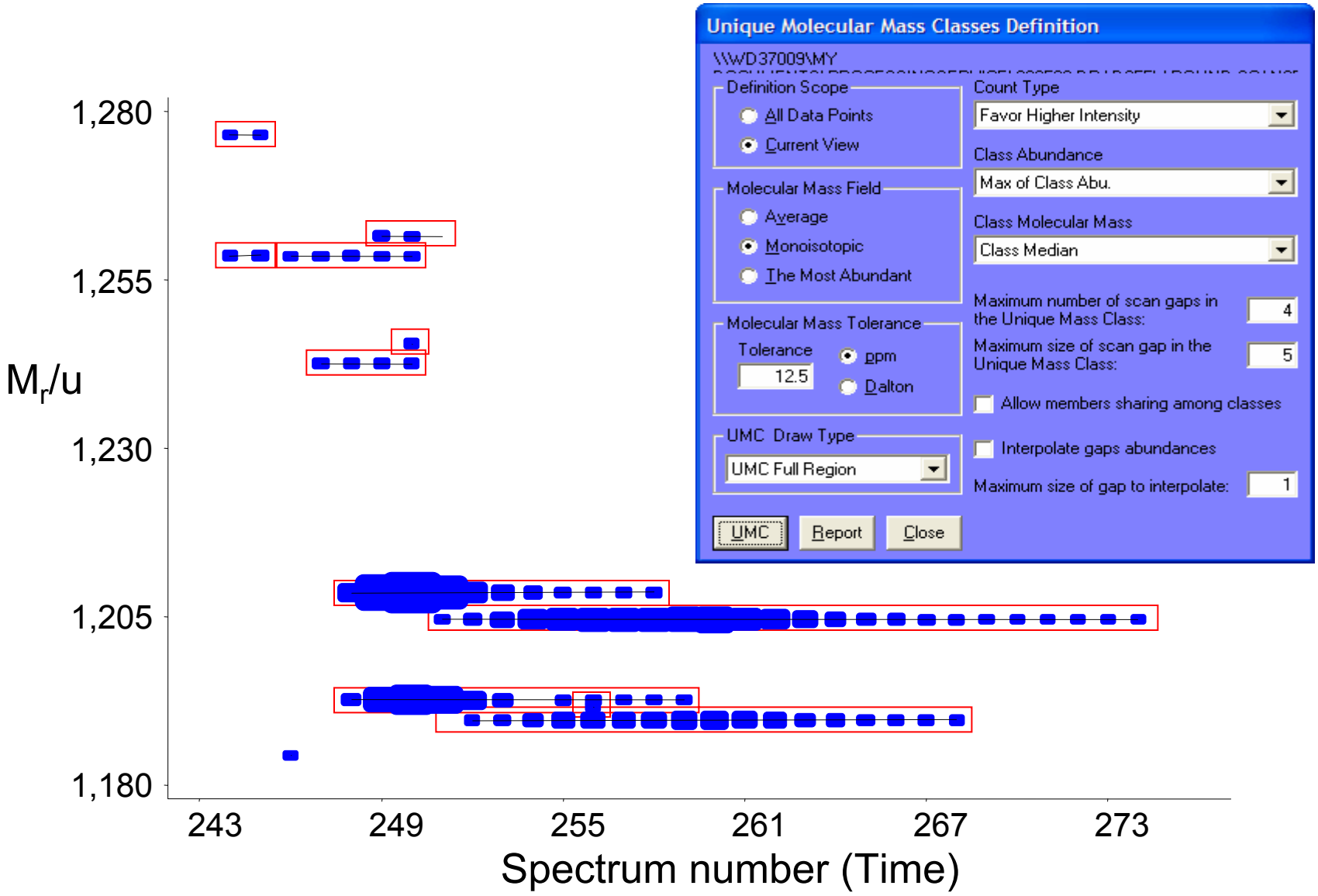
CS	Abundance	m/z	Fit	Average MW	Monoisotopic MW	Most abundant MW	Isotope
2	2.84E+08	578.3451	0.097	1155.3760	1154.6757	1154.6757	N14
2	2.47E+08	585.3236	0.05	1169.1556	1154.6741	1168.6326	N15

Capillary LC-FTICR 2-D Display of Peptides from a Yeast Soluble Protein Digest

>160,000 isotopic distributions corresponding to thousands of peptides



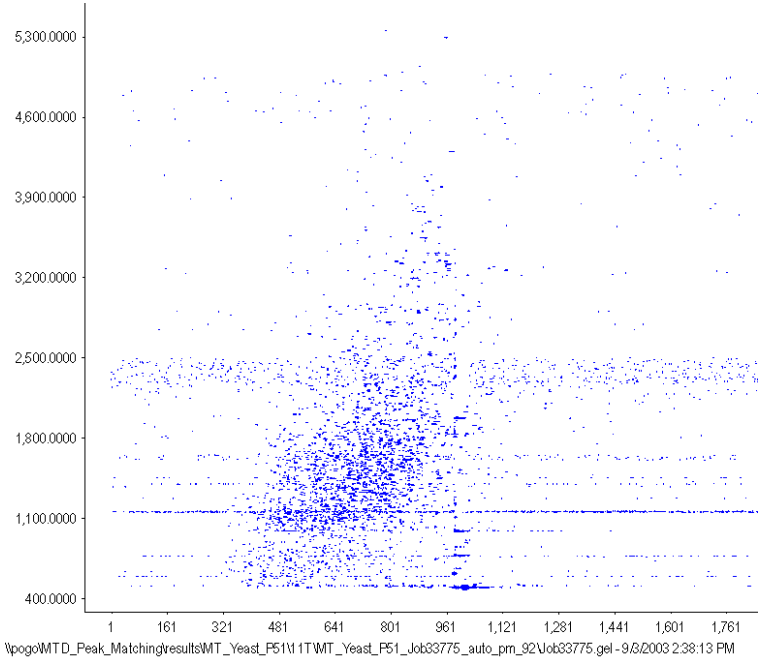
Unique Mass Classes Generation

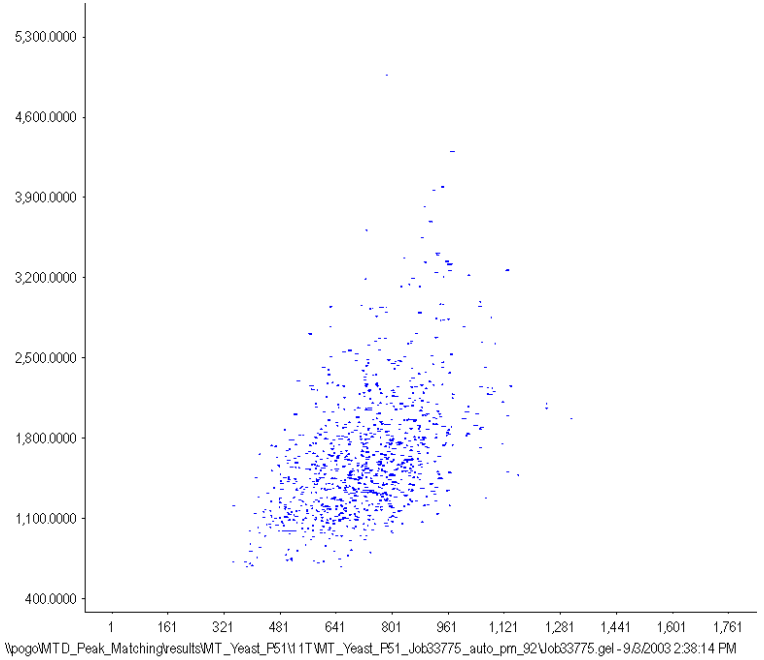


Q Rollups – Summary Sheet

Quantitation ID	SampleName	Comment	Experiments	Jobs	Threshold for Inclusion (%Max)
153	MGYMO_007a, 4 replicates	Jobs: 33775, 33777, 33779, 33781	MGYMO_007a	33775, 33777, 33779, 33781	33%

Unique Mass Tag Count	UMC Count	NetAdj NET Min	NetAdj NET Max	MMA Tolerance PPM	NET Tolerance	Refine Mass Cal PPM Shift	Results Folder Path
1184	1848	-0.04058	0.90892	3	0.05	2.8714	\\pogo\MTD_Peak_Matching\results\MT_Yeast_P51\11T\MT_Yeast_P51_Job33775_auto_pm_92





Q Rollups – Peptide Sheet

Reference	Abundance Average	Mass Tag ID	Mass Tag Abundance	Mass Tag Abundance StDev	Peptide	Monoisotopic Mass	UMC Match Count Avg	UMC Mass Tag Hit Count Avg
Q0045	24.33	17578	24.33	4.8	APDFVESNTIFNLNTVK	1907.9628	1	1
Q0045	24.33	358707	4.47	1.6	SSSIEFLLTSPPAVHSFNTPAVQS	2515.2594	1	1
Q0085	0.93	42610	0.93	0.2	WLISQEAIYDTIMNMTK	2056.0008	1	1
Q0105	2.75	355044	2.75	0.9	DVHNGYILR	1085.5618	1	1
Q0110	2.75	355044	2.75	0.9	DVHNGYILR	1085.5618	1	1
Q0115	2.75	355044	2.75	0.9	DVHNGYILR	1085.5618	1	1
Q0120	2.75	355044	2.75	0.9	DVHNGYILR	1085.5618	1	1
Q0250	24.36	4831	2.52	0.1	FVVTAADVIHDFAIPSLGIK	2112.1619	1	1
Q0250	24.36	17624	27.52	4.0	LLDTDTSMVVPVDTHIR	1910.9771	1	2
Q0250	24.36	19869	21.20	5.2	LNQVSALIQR	1140.6615	1	1
Q0250	24.36	350982	4.95	0.7	AIGYQWYWK	1213.5920	1	1
Q0250	24.36	353150	2.85	0.3	IEAVSLPK	855.5065	1	1

Reference	Mass Tag ID	Scan Minimum	Scan Maximum	Mass Error PPM Avg	Replicate Count Avg	Replicate Count Min	Replicate Count Max	Used For Abundance Computation
Q0045	17578	864	1102	-0.324	4	4	4	1
Q0045	358707	902	1147	-0.920	4	4	4	0
Q0085	42610	1226	1328	-1.540	3	3	3	1
Q0105	355044	580	734	0.331	4	4	4	1
Q0110	355044	580	734	0.331	4	4	4	1
Q0115	355044	580	734	0.331	4	4	4	1
Q0120	355044	580	734	0.331	4	4	4	1
Q0250	4831	1103	1199	-1.060	3	3	3	0
Q0250	17624	743	947	-0.250	4	4	4	1
Q0250	19869	653	831	0.286	4	4	4	1
Q0250	350982	842	1068	0.193	4	4	4	0
Q0250	353150	606	771	0.096	4	4	4	0

Q Rollups – Protein Sheet

Reference	Abundance Average	Abundance StDev	Mass Tag Count Unique Observed	Mass Tag Count Used For Abundance Avg	Mass Tag Matching Ion Count	Fraction Scans Matching Single Mass Tag	Mass Error PPM Avg
YNL055C	88.66	24.8	21	10	253	96%	0.167
YBL030C	51.67	19.4	18	7	151	60%	0.040
YBR085W	48.85	21.6	7	3	49	61%	0.049
YJR077C	42.15	16.4	12	4	103	100%	-0.051
YGL008C	39.64	23.0	27	2	195	77%	-0.285
YPL036W	39.64	23.0	18	2	132	77%	-0.111
YMR056C	36.71	17.9	6	2	36	44%	0.119
YBL104C	27.52	0	1	1	12	0%	1.514
Q0250	24.36	4.5	5	2	38	68%	-0.147
Q0045	24.33	0	2	1	19	100%	-0.622
YNL309W	19.87	0	1	1	6	100%	-2.661
YOR065W	16.28	4.9	6	5	63	100%	-0.283
YBR054W	15.36	0	4	1	29	55%	-0.261

Reference	Replicate Count Avg	Replicate Count StDev	Replicate Count Max
YNL055C	3.7	0.9	4
YBL030C	3.4	1.0	4
YBR085W	3.6	0.8	4
YJR077C	4.0	0.0	4
YGL008C	3.8	0.6	4
YPL036W	3.9	0.2	4
YMR056C	3.5	0.8	4
YBL104C	4.0	0.0	4
Q0250	3.8	0.4	4
Q0045	4.0	0.0	4
YNL309W	4.0	0.0	4
YOR065W	4.0	0.0	4
YBR054W	3.8	0.5	4

Q Rollups – Crosstab Sheets

Protein Crosstab

Reference	Job 33775 (QID147)	Job 33777 (QID148)	Job 33779 (QID149)	Job 33781 (QID150)	Avg Abu
YNL055C	87.14	100.73	98.28	69.94	89.02
YBL030C	55.72	56.68	60.00	39.07	52.87
YBR085W	52.41	54.94	51.68	33.62	48.16
YJR077C	45.13	46.60	45.75	36.02	43.38
YGL008C	34.66	42.73	45.94	27.81	37.79
YPL036W	34.66	42.73	45.94	27.81	37.79
YMR056C	41.45	41.62	30.66	26.06	34.95
YBL104C	27.41	30.64	30.18	21.84	27.52
Q0250	25.85	26.92	27.06	17.61	24.36
Q0045	27.99	26.38	25.70	17.24	24.33
YNL309W	18.37	19.85	26.58	14.68	19.87
YOR065W	16.31	18.36	18.18	12.29	16.28
YBR054W	15.35	16.09	19.17	10.85	15.36

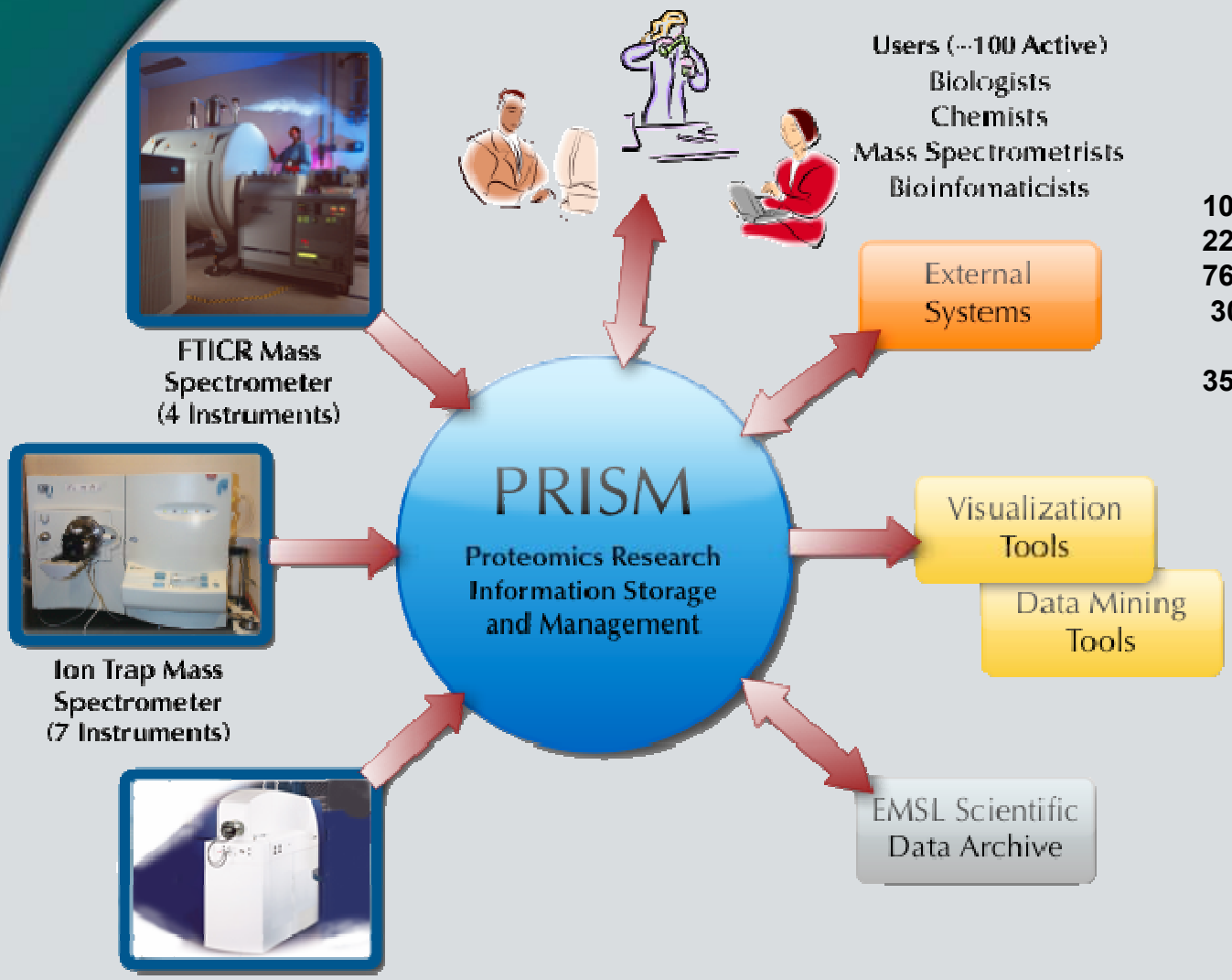
Proteins with Peptides Crosstab

Reference	Mass Tag ID	Job 33775 (QID147)	Job 33777 (QID148)	Job 33779 (QID149)	Job 33781 (QID150)	Avg Abu
Q0045	17578	27.99	26.38	25.70	17.24	24.33
Q0045	358707	4.10	6.05	5.42	2.32	4.47
Q0085	42610	1.12	0.88	0.78		0.93
Q0105	355044	3.18	3.27	3.23	1.33	2.75
Q0110	355044	3.18	3.27	3.23	1.33	2.75
Q0115	355044	3.18	3.27	3.23	1.33	2.75
Q0120	355044	3.18	3.27	3.23	1.33	2.75
Q0250	4831	2.50	2.40	2.67		2.52
Q0250	17624	27.41	30.64	30.18	21.84	27.52
Q0250	19869	24.28	23.19	23.94	13.37	21.20
Q0250	350982	5.38	5.36	5.19	3.87	4.95
Q0250	353150	2.50	3.19	3.07	2.65	2.85

Outline

- Accurate Mass and Time Tag (AMT) based proteomics
- Instrumentation
- Data analysis
- **Data management**
- Challenges

Create information about the proteins in biological samples using mass spectrometer data



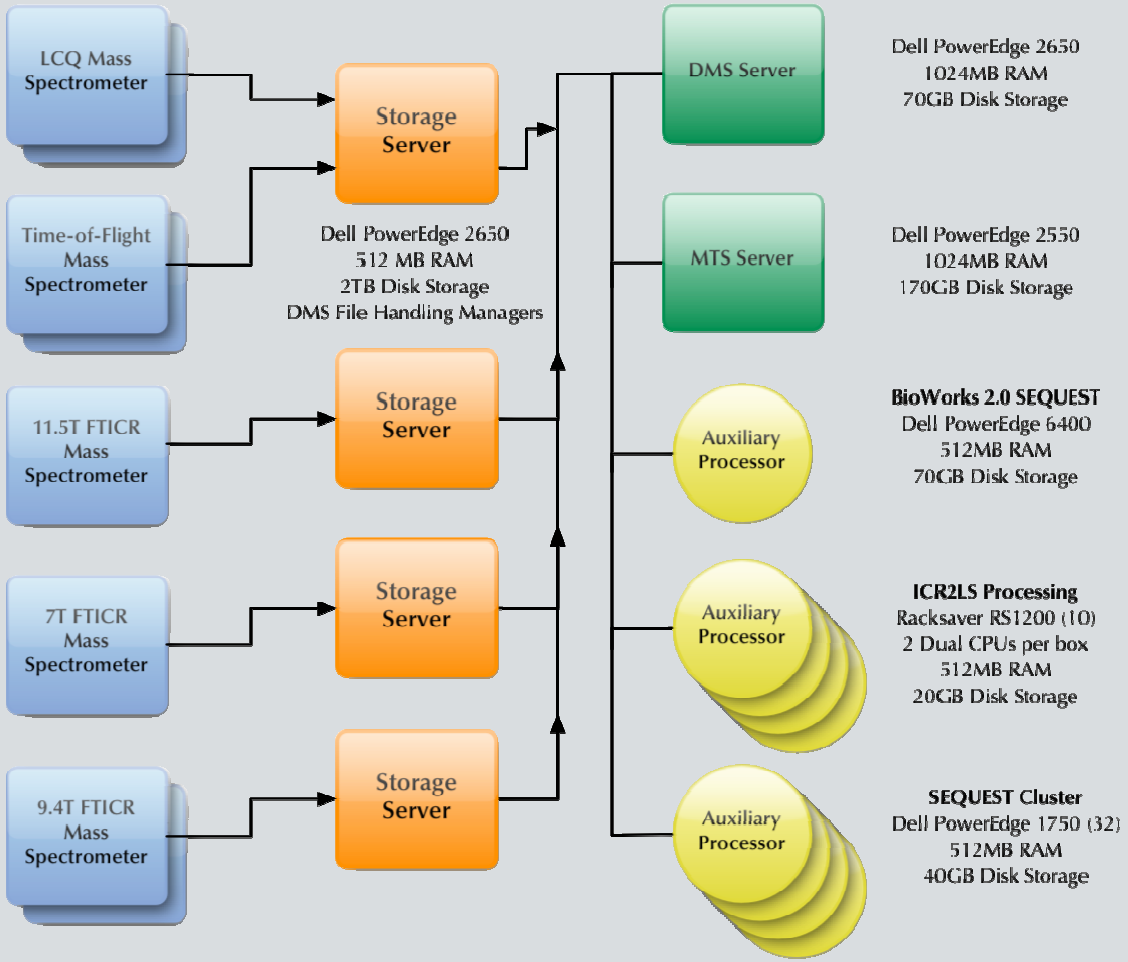
36	Organisms
129	Research Campaigns
10,950	Prepared Samples
22,341	MS Instrument Runs
76,543	Automated Analyses
30 TB	Data / results files
210	Databases
350 GB	Data in databases

PRISM Functions

- Collect Information
  - Capture spectra files from MS instruments
  - Gather metadata (biomaterial, sample prep., etc.)
- Store Information
  - Raw spectra files
  - Analysis results files
  - Lab information
- Process Information
  - Analyze raw spectra with multiple tools
  - Identify mass tags, peptides and proteins
  - Formulate reports
- Present Information
  - Web interface for general user access
  - Database and file access for external tools and systems
  - Utility programs for special purpose data access

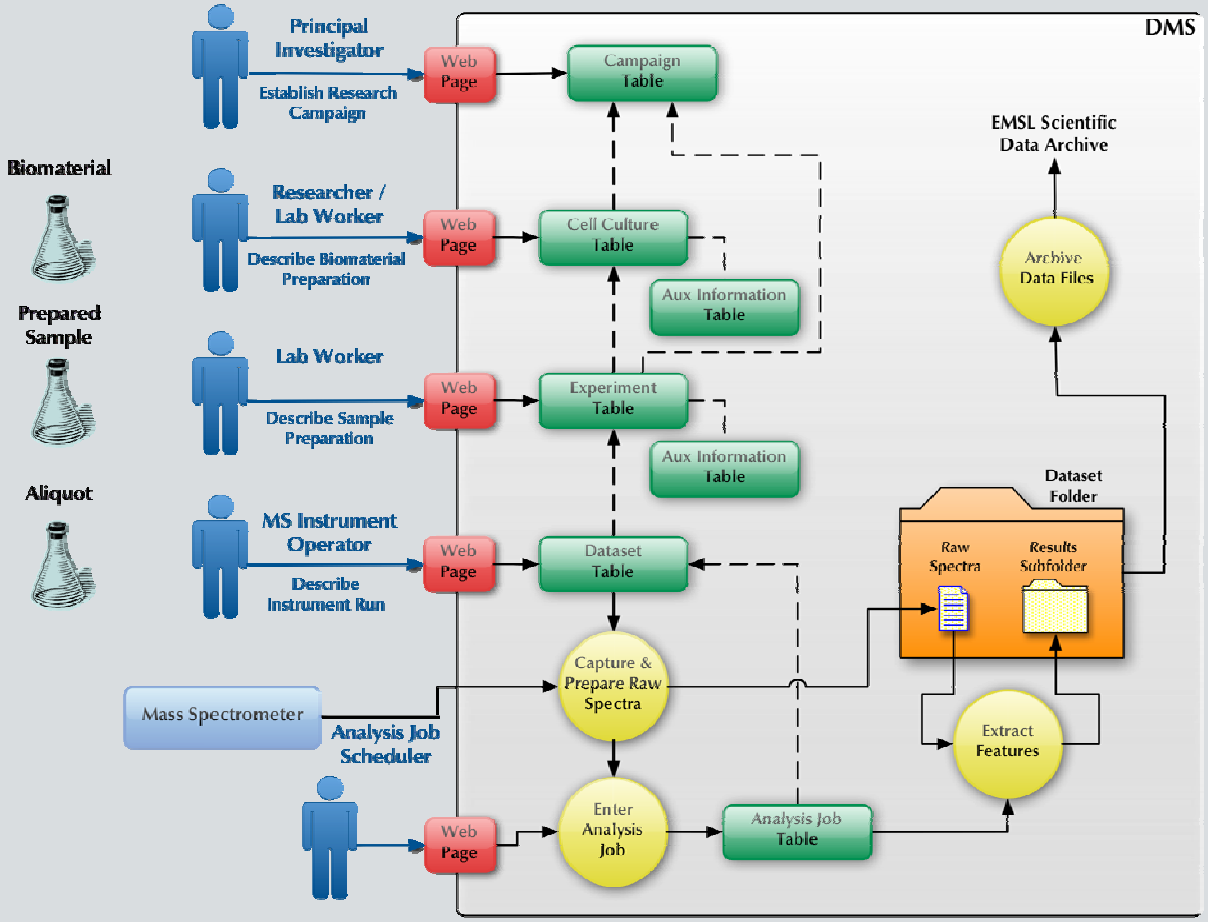
PRISM Hardware

- Primary Servers
  - Databases
  - Web User Interfaces
  - Data Management System (DMS)
  - Mass Tag System (MTS)
- Storage Servers
  - Bulk file storage and automated file handling
  - Add more to handle more MS instruments
- Auxiliary Processors
  - Automated data analysis
  - Add more to address increasing sample volume

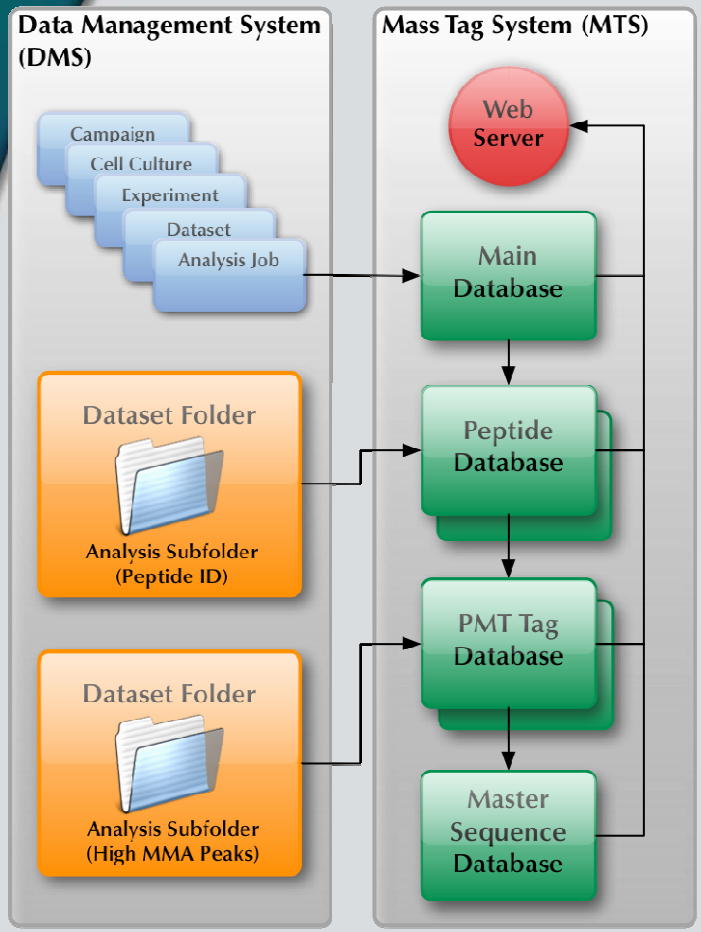


Lab Information / Data Analysis

- Campaign
  - Line of research
- Cell Culture
  - Biomaterial source
- Experiment
  - Prepared sample
- Dataset
  - MS Instrument run
- Analysis Job
  - Peptide Identification
  - Peak Identification
  - Total Ion Current (TIC)



Mass Tag Creation



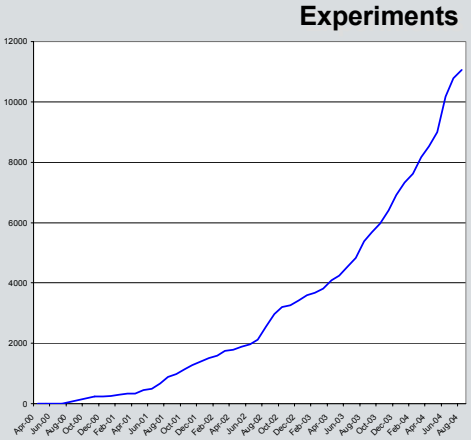
Mass Tag System (MTS)

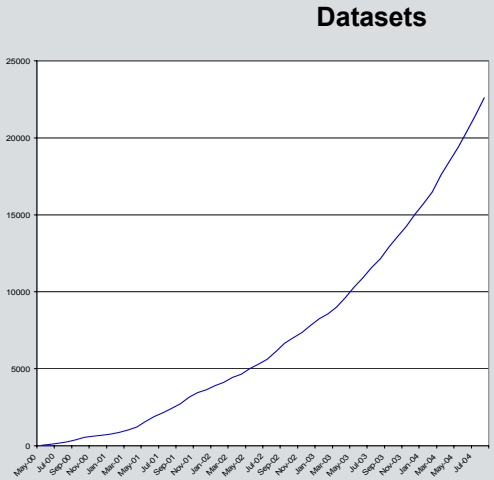
- Main Database
  - Centralize references to DMS metadata
  - Manage automatic update cycle
- Peptide Database
  - Select Analysis Jobs from DMS
    - Based on metadata in DMS tracking DB
  - Extract peptide identifications
    - From DMS analysis results files
    - Based on selection criteria
- Mass Tag Database
  - Select analysis jobs from peptide database
  - Import selected peptides and make PMTs
  - Extract and match monoisotopic peaks
    - From DMS analysis results files
    - Match to PMTs by mass and elution time

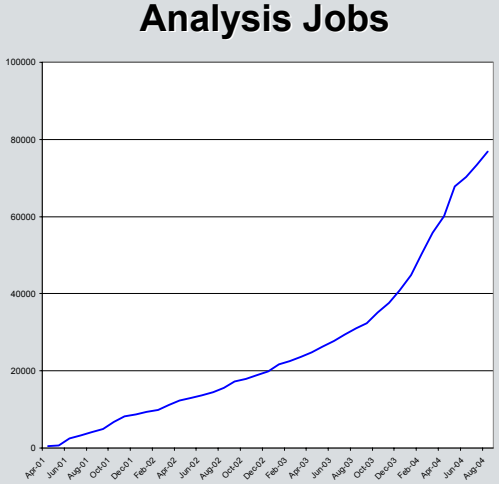
PRISM Operation

- Steady operation for several years
- ~99% Uptime

- Phased Improvements
  - Phase 1 (March 2000) Basic Dataset Capture
  - Phase 2 (March 2001) Automated Analyses
  - Phase 3 (March 2002) Mass Tag System
  - Phase 4 (April 2003) High-capacity storage servers
  - Phase 5 (June 2004) Quantitation Automation, More tools, Enhanced Filtering UMC Automation







Outline

- Accurate Mass and Time Tag (AMT) based proteomics
- Instrumentation
- Data analysis
- Data management
- **Challenges**

Challenges

- Multiple data standards are required
  - Experiment description
  - Raw mass spec data
    - Capture the instrument configuration
  - Spectral analysis results
    - De-isotoping
    - Peptide ID
  - Protein expression results
- Common dictionary is needed to support interoperability
- Proteomics method description (analysis workflow)
- Analyses tool meta data capture
  - Tools used, executables available
  - Versions
  - Parameter files
- Current efforts
  - HUPO PSI
  - mzXML intergraded into one PNNL development

Enabling Actions

- Small working groups
  - Develop a roadmap
  - Dictionary
- Multi-institution programs targeted at interoperability
  - Funded to develop and implement standards
  - Samples analyzed at multiple sites
  - Data exchanged and analyzed at multiple sites
  - Development of a unified web based dissemination of data and methods
- Efforts integrated with HUPO PSI
- Publication requirements for data access

Proteomics Software Team

- Team Members
  - Gordon Anderson (Team leader)
  - Ken Auberry (Scientific liaison)
  - Dave Clark (Process automation SW development)
  - Lars Kangas (Artificial intelligence SW development)
  - Gary Kiebel (System architecture, database, web interface)
  - Matt Monroe (Analytic tool development)
  - Nikola Tolic (Peak matching, visualization SW development)
  - Eric Strittmatter (Analytic tool development)
- Contributors
  - Edward Ellis
  - Nathan Trimble
  - Kyle Littlefield
  - John Sandoval
- Funding
  - DOE, OBER
- For more details
  - http://ncrr.pnl.gov

Dick Smith and his research group