Identifying and locating PTMs in complex peptides or proteins based on acquired high resolution Full-Scan and MS/MS data

Ray Fyhr – Merck IT, July 15th 2013

Proteomics and Bioinformatics 2013, Philadelphia, PA

MERCK
Be well

The SME and ME

- The talents necessary to throw the football and catch the touchdown pass do not exist in the same person
- Requires two EXTREME experts.
- Must overcome inherent technical language barrier
- They think differently
- They work differently
- Each has to be able to learn from and teach the other
- Each has to be able to minimize the details for the other

Bottom-Up versus Top-Down

Stick to what we know and continue to use Bottom-Up even though we are aware of it's restrictive limitations ?

versus

Adopt something new(er) which has shown interesting results but involves a learning curve and lacks the robust software tools needed to fully exploit ?

Advantages of Top-Down

Decoding protein modifications using top-down mass spectrometry

Nertila Siuti and Neil L Kelleher
Nature Methods. 2007 October; 4(10): 817–821

- Top-down mass spectrometry ... strives to preserve the post-translationally modified (PTM) forms of proteins present *in vivo* by measuring them intact …

- … PTMs are a key driving force behind cellular signaling. … intact proteins are less susceptible to instrumental biases than are their small peptide counterparts

It’s been over 10 years

What’s up with TD?

- Mass Spectrometers are better
- More PhD scientists in the field
- More Bioinformatics IT people
- Computers are much faster
- Clusters are much bigger
- Amazon has got the CLOUD (cheaper)

Top-Down analysis needs new software

Proteomics is pretty complex and Programming is pretty complex

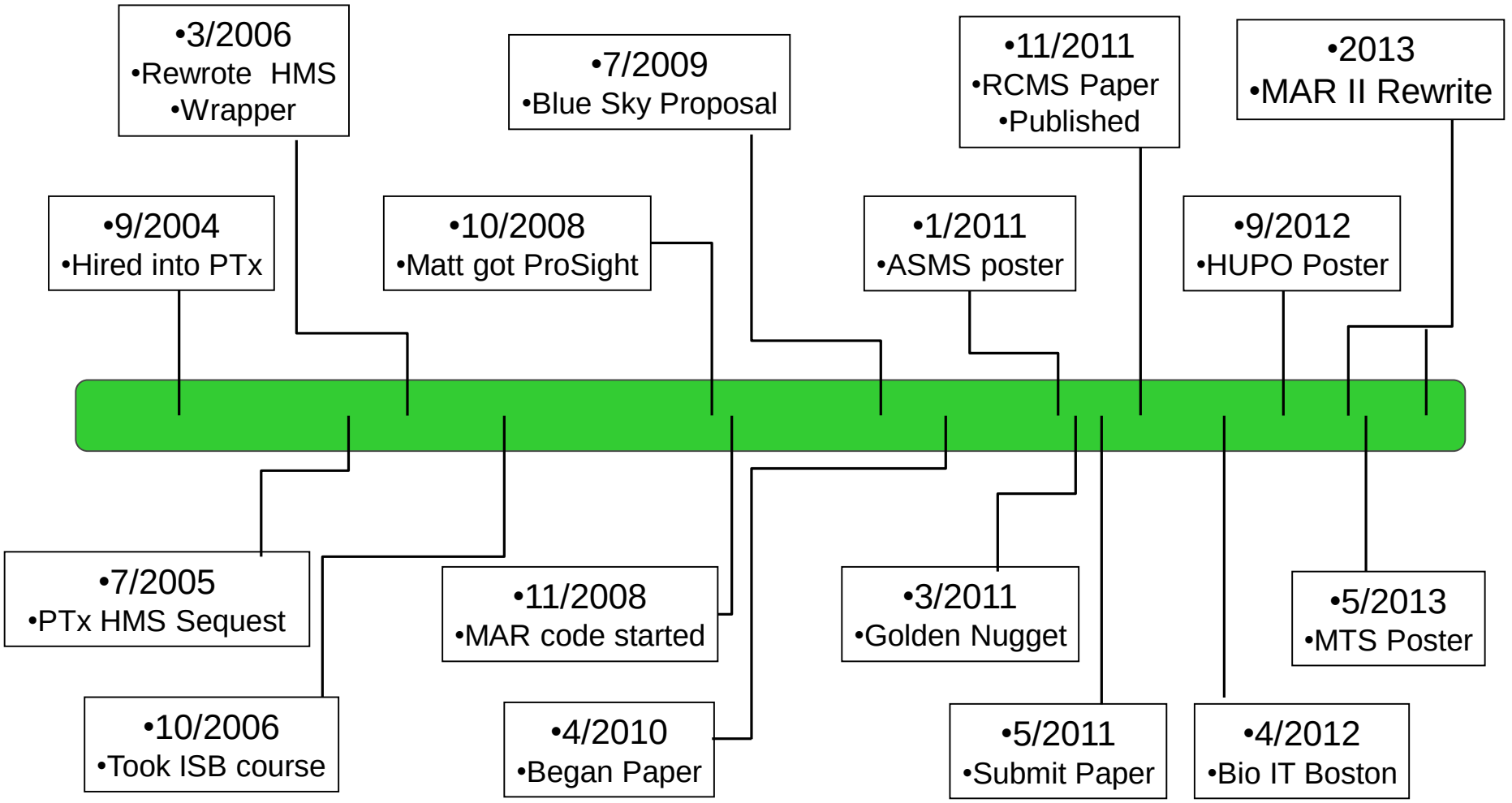
Scientist - “I have all these features of interest, but I can’t get IDs for them. This is really hindering my progress in understanding cellular signaling and biomarker discovery.”

Programmer - “Seems like the scientists are doing all the interesting work. Is Google still recruiting? Wish there were some good video game companies around here. I feel my skills are deteriorating”

The Software Developer needs to

- Ask the Right Questions
- Learn how to find their own answers
- Read the Relevant Literature
- Survey the Current Software Tools
- Understand the Data from the Instruments
- Necessity is the Mother of Invention
- Dream It Then Build It
- “10% inspiration and 90% perspiration”

Time Line of the MAR algorithm



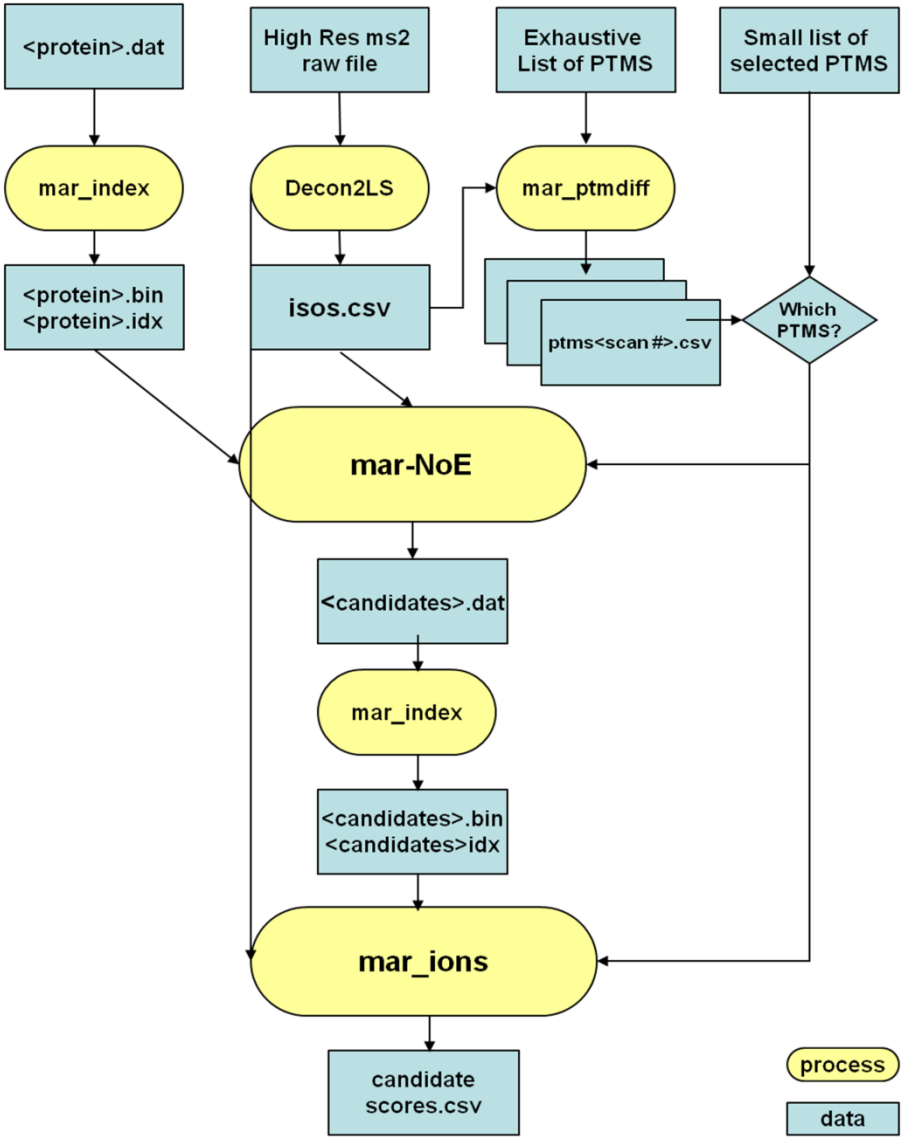
MERCK
Be well

Original design goals for MAR

Create a software tool to convert high resolution MS/MS data into peptide or protein identifications

- Employ a simple fasta-formatted protein database
- Allow pre-defined “differential” modifications for searching
- Eliminate intact protein or enzymatic restrictions
- Consider high mass accuracy data for scoring
- Perform full-scan surveying to determine high probability PTMs
- Make it parallelizable for high performance
- Develop new functionality to locate residue within peptide

MAR – process flow diagram



MERCK
Be well

mar_index – creates binary data and index

Input File: FASTA-formatted protein database (.dat)

```
ID       APOC3_HUMAN
SQ
     MQPRVLLVVA LLALLASARA SEAEDASLLS FMQGYMKHAT
     KTAKDALSSV QESQVAQQAR GWVTDGFSSL KDYWSTVKDK
     FSEFWDLDPE VRPTSAVAA
//
```

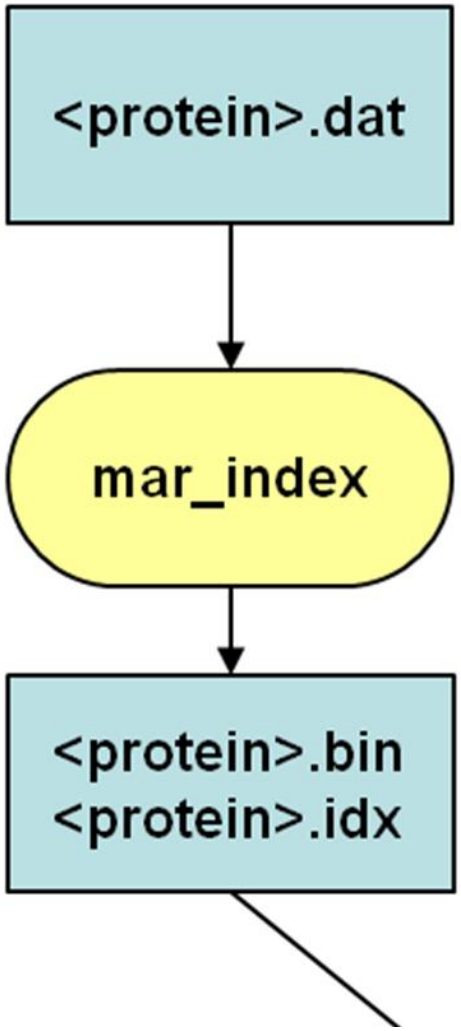
Output file: index (.idx)

10827.488350, APOC3_HUMAN, 12897907, 606069

1.Theoretical molecular weight
2.The Protein Name
3.Pointer to the binary file of it's amino acid molecular weights
4.Pointer to the ascii .dat file of this protein

Output file: binary array of molecular weights (.bin)

1.3104049e+02, 1.2805858e+02, 9.7052764e+01
1.5610111e+02, 9.9068414e+01, 1.1308406e+02
1.1308406e+02, 9.9068414e+01, 9.9068414e+01
7.1037114e+01, 1.1308406e+02, 1.1308406e+02
7.1037114e+01, 1.1308406e+02, 1.1308406e+02, ……..



MERCK
Be well

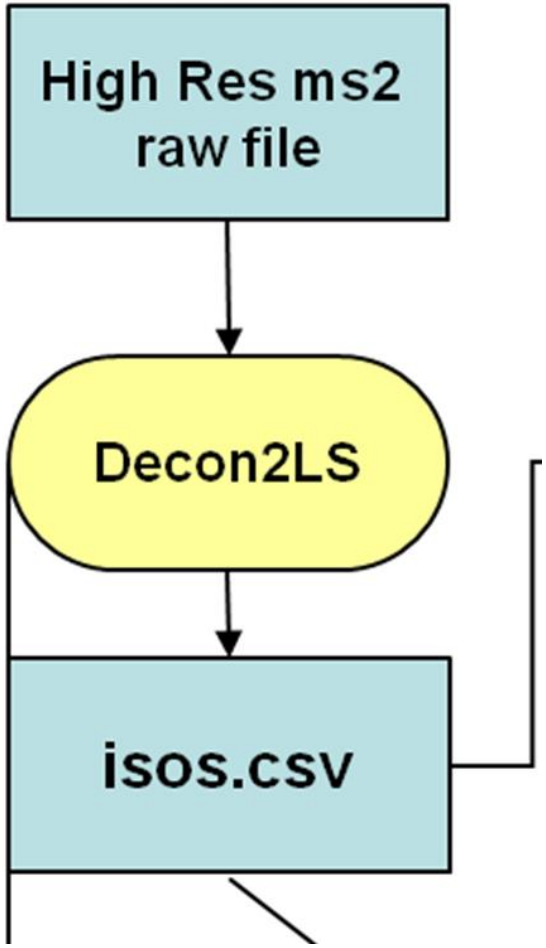
MAR uses “Decon2LS” to THRASH raw data

Thermo High Res ms2 .raw files converted into lists (.csv format) of scan number and monoisotopic neutral masses using the Horn transformation function of the publically available program Decon2LS (PNNL)

N. Jaitly, A. Mayampurath, K. Littlefield, J. N. Adkins, G. A. Anderson, R. D. Smith. Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. BMC Bioinforma.2009, 10, 87

Typical Raw File – 1 GByte
<filename>_isos.csv – 10 Mbytes (* use ~2.5Mbytes)

scan_num *	**monoisotopic_mw ***
charge	**mostabundant_mw**
abundance	**fwhm**
mz *	**signal_noise**
fit	**mono_abundance**
average_mw	**mono_plus2_abundance**



MERCK Be well

mar_ptmdiff – chooses best PTMs using MS scans

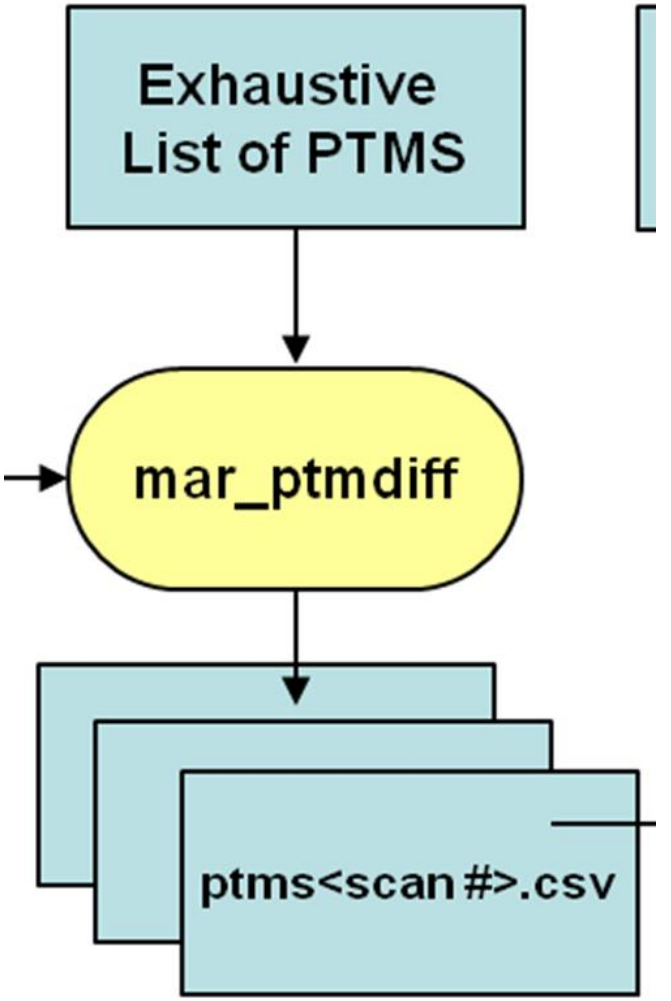
List of ’differential’ modifications in a simple csv

use,aa,delta,uid,max,name
Y,T,656.203700,10025,1,O-glycosylation

mar_ptmdiff reduces computation time by discovering matches between the experimental data from the full-scans and the theoretical list of ’differential’ modifications. The matches within a user specified delta are then ordered and a unique list is produced for each MS/MS scan.

Reduced the exhaustive list of 351 choices to 4

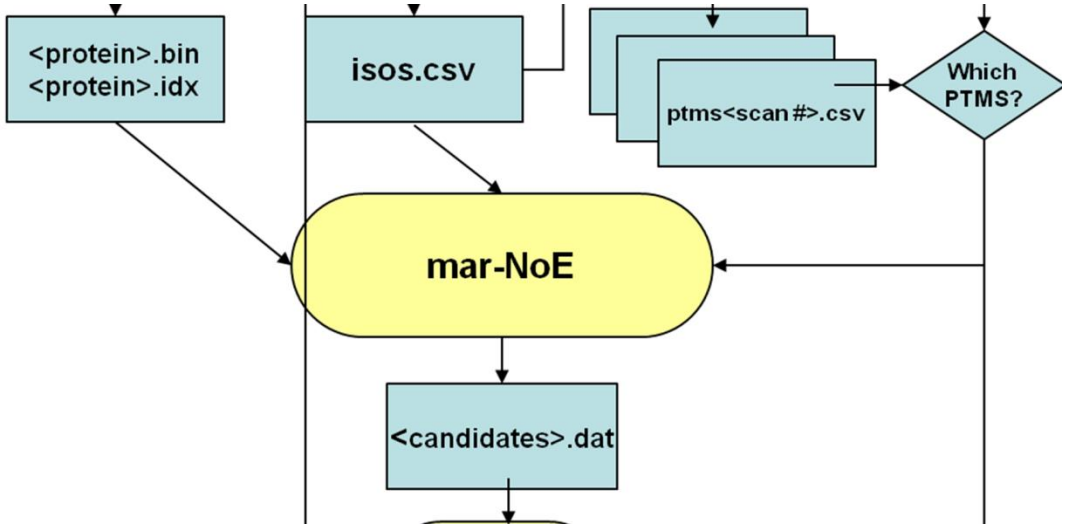
use,aa,delta,uid,max,name
Y,N,656.227600,149,1,Hex1HexNAc1NeuAc1
Y,S,656.227600,10015,1,Hex1HexNAc1NeuAc1
Y,T,656.227600,10016,1,Hex1HexNAc1NeuAc1
Y,T,656.203700,10025,1,O-glycosylation



MERCK
Be well

mar_NoE – selects candidates using PTMs

Executes a "no enzyme" inch-worm type search that greatly increases the candidate search space.



Uniprot_Human database (release 05/03/2011) comprised of 20,238 proteins with an average protein length of 558 residues

The number of unmodified peptide candidates 5 amino acids or longer in length is **6.8 billion**

Imposing the trypsin restriction requiring cleavage at Lys and Arg residues, that number is reduced to approximately **3.3 million.**

That's a 2050-fold difference

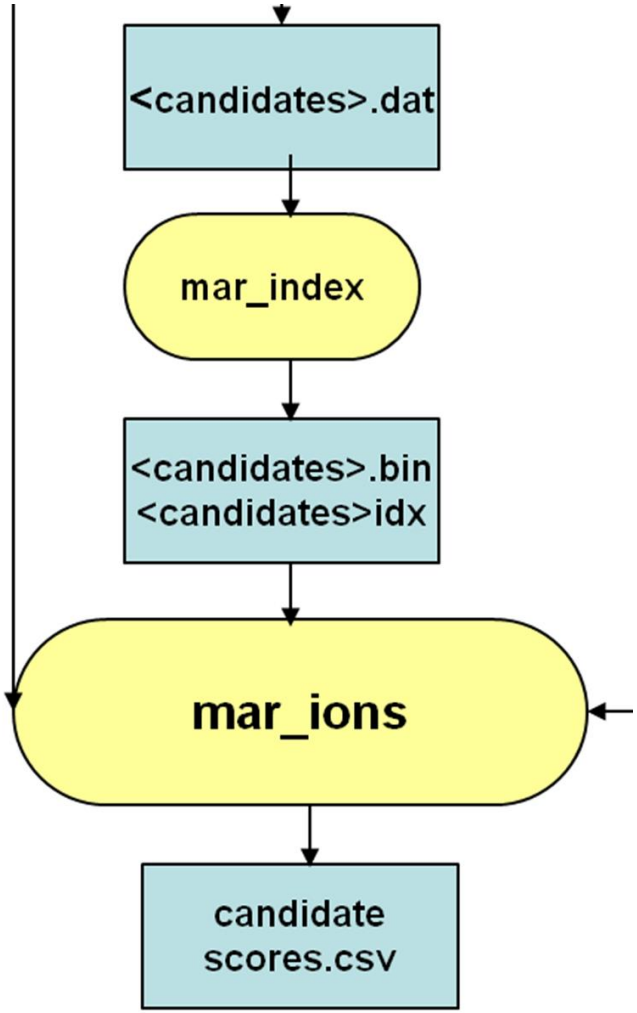
MERCK
Be well

mar_ions – scores the candidate sequences

Insilico generated fragments of these candidate sequences are created by the *mar_ions* module and compared against experimental fragment ions from the deconvoluted “isos” file data.

Every candidate fragment is first assigned an internal score based on the sum of the number of ions matched between in silico and experimental molecular weights within the tolerance the user requested (30 ppm).

This sum is then divided by the variance of the ion matches giving an internal scoring mechanism that can be calculated quickly. Upon completion of the *mar_ions* module, the highest 200 candidate scores are then assigned P-scores.



MERCK
Be well

2011 – "Golden Nugget" from RCMS paper

mw - 9415.4568

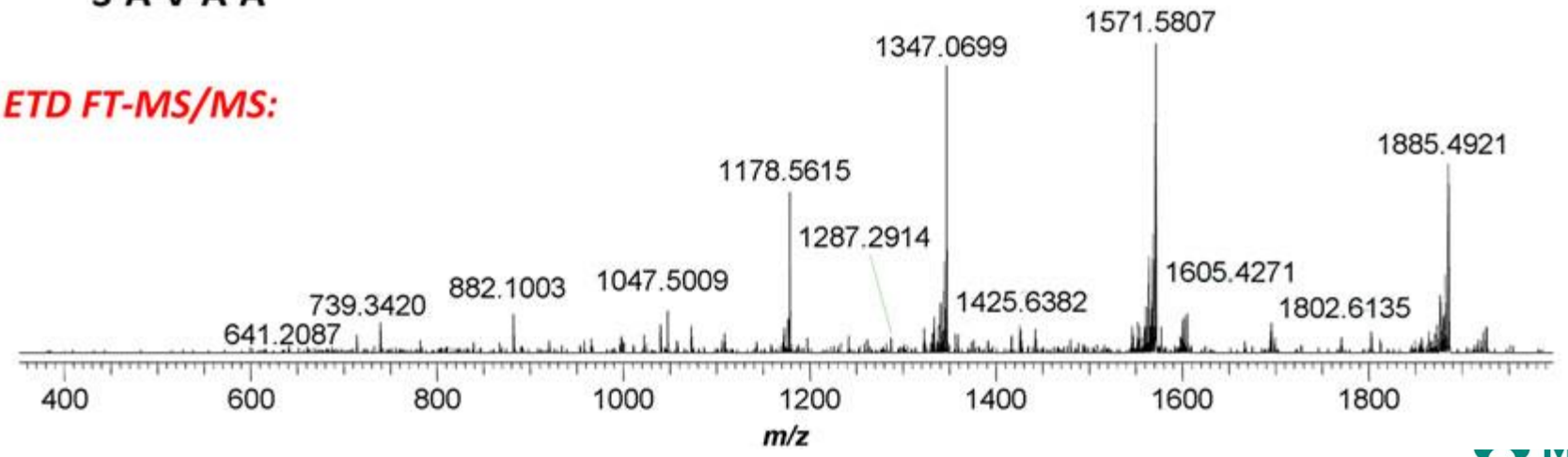
Search Results:

P-score	c ions	z ions	Scan Number	Protein Name	PTM id	PTM mass	Modification Name
2.0E-13	23	18	1979	APOC3_HUMAN_NoE_21_99	10025	656.2393	T O-glycosylation
2.0E-13	23	18	1979	APOC3_HUMAN_NoE_21_99	10015	656.2393	S Hex1HexNAc1NeuAc1
2.0E-13	23	18	1979	APOC3_HUMAN_NoE_21_99	10016	656.2393	T Hex1HexNAc1NeuAc1
0.6	10	17	1979	ACRBP_HUMAN_NoE_241_322			
0.6	12	15	1979	FOXP4_HUMAN_NoE_29_120			

Fragmentation Map:

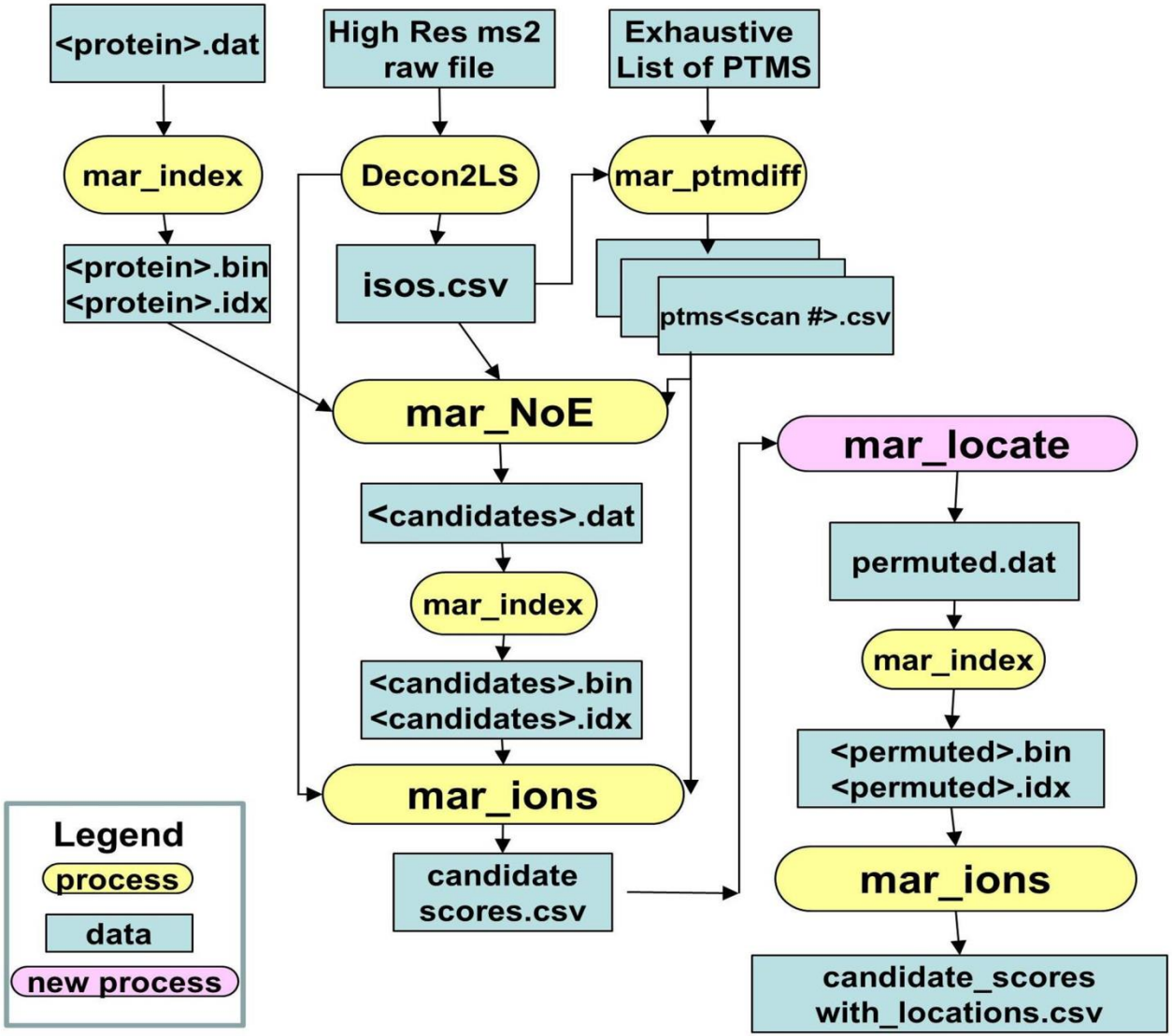
S-E-A-E-D-A-S-L-L-S-F-M-Q-G-Y-M-K-H-A-T-K-T-A-K-D-A-L-S-S-V-Q-E-S-Q-V-A-Q-
Q-A-R-G-W-V-T-D-G-F-S-S-L-K-D-Y-W-S-T-V-K-D-K-F-S-E-F-W-D-L-D-P-E-V-R-P-T-
S-A-V-A-A

ETD FT-MS/MS:



MERCK Be well

2012 – added PTM residue location



MERCK
Be well

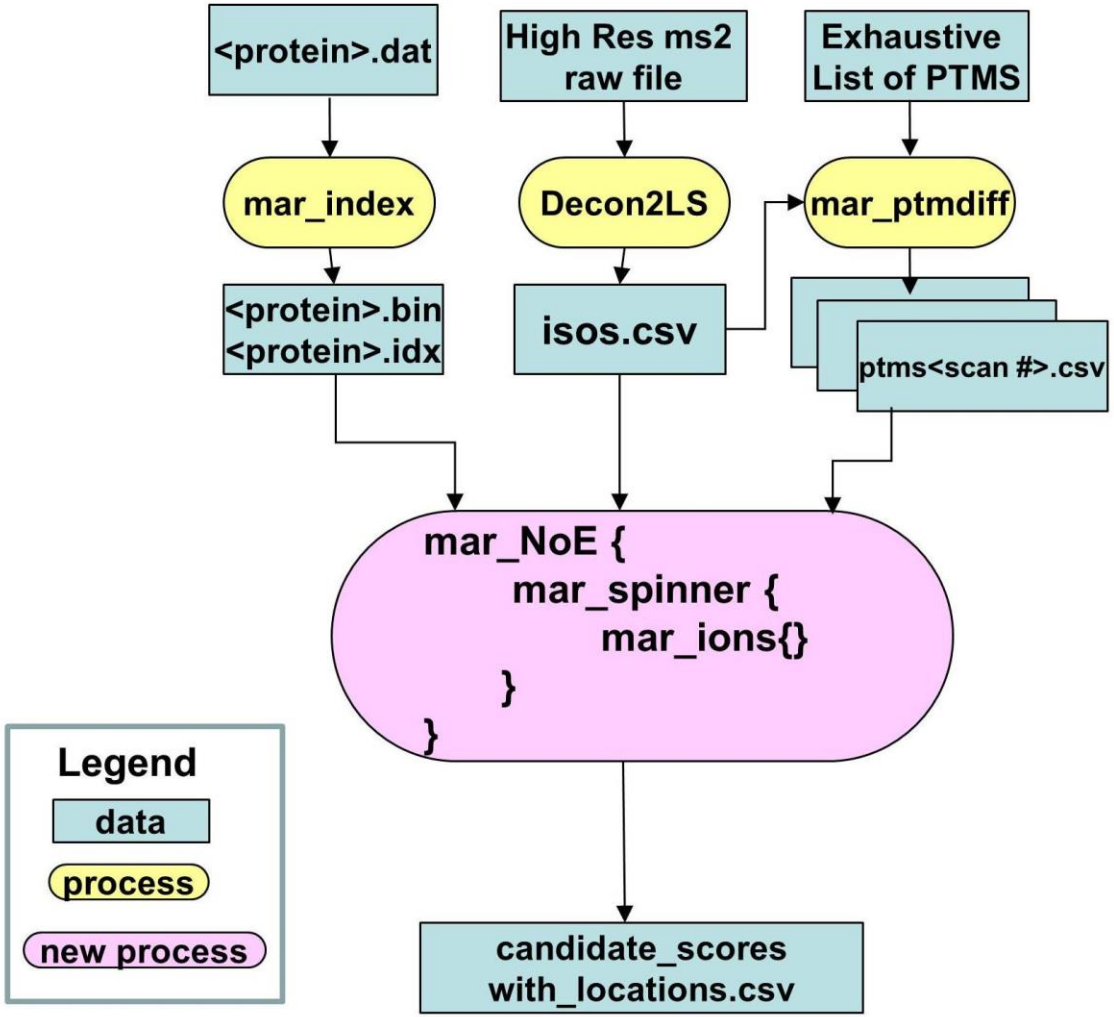
2012 - PTM residue location results

- Developed a first version of the PTM locator extension to the MAR algorithm
- Results:
  - 9 raw files with total of 628 ms2 scans
  - 123 Identifications with P-Scores < $1e^{-5}$
  - 54 of those had PTMs which were algorithmically located

2012 - Abbreviated list of some top polypeptide identification results

P-score	B/C ion	Y/Z ion	polypeptide	mw	mw diff	PTM mw	PTM description	aa RES
3.87E-59	29	50	TRX2_YEAST_NoE_2_104	11063.5735	0.0273	-2.0146	C 1 disulfide bridge	30
6.06E-57	38	45	G3P3_YEAST_NoE_240_332	10149.2956	-0.9718	0.9840	N Deamidated	9
4.03E-41	22	43	RS28B_YEAST_NoE_1_67	7602.1650	-0.0191	42.0106	K Acetyl	11
7.49E-33	26	31	G3P3_YEAST_NoE_240_332	10150.2809	0.0135	0.9840	N Deamidated	9
8.90E-28	30	23	HSP12_YEAST_NoE_2_109	11596.6444	0.0154	42.0106	K Acetyl	11
1.82E-20	17	34	G3P3_YEAST_NoE_208_332	13372.0599	-0.0033	14.0157	L methyl	101
6.95E-16	19	16	SDO1L_YEAST_NoE_2_111	11913.1289	0.0169	42.0106	K Acetyl	5
1.57E-14	4	40	MAL12_YEAST_NoE_496_584	10427.5542	-0.6747	-17.0265	Q Q pyroglutamic ac	18
4.52E-10	7	23	ENO1_YEAST_NoE_320_437	12646.6507	-0.9759	0.9840	Q Gln->Glu	42
1.34E-08	16	16	G3P3_YEAST_NoE_283_332	5573.7878	0.0095	14.0157	L methyl	43

2013 – Mar II is a major rewrite
many simultaneous PTMs with location



MERCK
Be well

MAR's Advantages

1. Considers high mass accuracy data for scoring
2. Does complete "No Enzyme" based search, it is not dependent on predicting the cleavages in advance
3. Allows for many simultaneous differential modifications
4. Indexing of protein database files is very fast allowing for any number of custom protein databases
5. Provides full-scan surveying for potential ion mass differences to allow "on-the-fly" generation of differential modification list used for searching
6. Parallelized and performance scales linearly

Scientific Software Collaboration

Expert to Expert

Can’t be Scheduled

Science is the Driver

Minimize Interference

Maximize Synchronicity

Acknowledgements

We would like to thank Ekaterina Deyanova and Helene Cardasis for the collection of high resolution MS/MS data. The authors thank Fanyu Meng, Laurence Lee, and Cloud Paweletz for their helpful discussions and careful review of the original manuscript. The authors also thank Ronald Hendrickson, Nathan Yates, Neil Kelleher, John "Jack" Kellie, and Martin Leach for their support. Decon2LS was graciously provided free of charge by the Pacific Northwest National Laboratory. Merck IT management has been very helpful throughout the years and they include Gary Mallow, Alec Anuka, Allan Ferguson, Dermot BarryWalsh, Mark Hill, and Clark Golestani.

Backup Slides Begin

Short Abstract

Up until recently, most mass spectrometry biomarker discovery strategy focused on small peptide fragments ignoring the post translational landscape of larger peptides and intact proteins. Top down proteomics analyzes the intact protein and all its post translational modification in one single run. Here we describe an extension to a new top-down proteomics algorithm developed at Merck called "**MAR".** The application of these new developments for protein id may be very useful in areas such as neuroproteomics and neurology.

Short Biography

- **Ray Fyhr –** transitioning Sr. Specialist, Business Analysis, Merck
- **Biography -** Ray Fyhr has an AB degree from Colgate University and an MS degree in Computer Science from New York Institute of Technology. With over 30 years of professional experience as a software developer and business analyst, his career has traversed many industries including pharmaceutical, communications, financial, transportation, manufacturing, and operations research. In 2004, Mr. Fyhr joined Merck supporting the information technology needs of the Proteomics department. This included a customized LIMS deployment and other lab automation endeavors involving Elucidator and HPC. Of particular interest to Ray is developing software which solves complex research problems. Now he is involved with implementing new algorithms for Top-Down Proteomics and other types of biomarker analyses in Big Data Platforms.

Two modes of PTM searching

All at a Time:

The key computational advantage of the differential modification consideration is the multiplicative (not factorial) expansion of possible candidate peptides:

Max multiplying factor = (Mod1+1) x (Mod2+1) x (Mod3+1)…x…(Modn+1)

Where n is the number of different modifications (Mod) considered.

One at a Time:

The number of modifications is restricted to 1 for any candidate polypeptide

Maximum multiplying factor = N (Where N is the number of PTMs).

P-value - candidate scoring and database size estimation

MS/MS search results using the MAR algorithm are scored according to a 'random-match' probability to all matching candidate peptides. Using the equation defined by Meng et al, [1]

$$P = ((xf)nxe-xf)/n!$$

a P-value is calculated for each candidate polypeptide.

Calculation of the x-term given in the probability equation (above) is adapted from the original equation [1] as such:

$$x = (1/111.1)*(2^(n+1))*(Ma*2)$$

where the mass accuracy (Ma) is 0.5 Da, and the number of fragments 2^(n+1) term indicates that the location of the modification is not considered for scoring. The location of the modification is determined later using the MAR_ions algorithm, which compares the 'differential' modification residue specificity with the number of matching fragment ions for the matching candidate protein. Protein forms with the highest number of fragment ions are considered correct; however, manual inspection of the PTM location is required.

Protein database size is determined by the number of possible modifications considered. The number of forms for each protein of AA amino acid length (limited in size to >5 amino acids) is

$$\#\ \text{candidates per protein} = (AA+1-5)*[(AA+1-5+1)/2]*(1+1*350)$$

This is calculated for each of the 20238 proteins of the Uniprot_Human database (release 05/03/2011) and sum totaled, equaling a value of 2,385,665,247,328.

1 Meng, F.; Cargile, B.J.; Miller, L.M.; Forbes, A.J.; Johnson, J.R.; Kelleher, N.L. Nat.Biotechnol. 2001, 19 (10), 952-957

Lines of Code – all ‘C’ with no packages

MAR

Module name	LOC
mar_NoE.c	350
mar_find.c	320
mar_index.c	252
mar_ions.c	696
mar_ptmdiff.c	311
mar_common.c	781
mar_master.c	502
mar_client.c	277
TOTAL	3489

MAR II

Module name	LOC
mar_NoE.c	1794
mar_find.c	320
mar_index.c	252
mar_ions.c	-
mar_ptmdiff.c	311
mar_common.c	-
mar_master.c	502
mar_client.c	277
TOTAL	3456

MAR II has a lot more functionality with LESS code