Center for
Computational
Mass
Spectrometry

Approach for large-scale identification of linked peptides from tandem mass spectra

Jian Wang[1], Veronica Anania[2], Jeff Knott[3], John Rush[3], Jennie R Lill[2], Philip E Bourne[4] and Nuno Bandeira[4,5,6]

1. Bioinformatics Program, UCSD, La Jolla, CA
2. Protein Chemistry Department, Genentech Inc., 1 DNA Way, South San Francisco, California
3. Cell Signaling Technologies, Danvers, MA
4. Skaggs School of Pharmacy and Pharmaceutical Science, UCSD, La Jolla, CA
5. Center for Computational Mass Spectrometry, UCSD, La Jolla, CA
6. Computer Science and Engineering, UCSD, La Jolla, CA

Contact: jiw006@ucsd.edu   bandeira@ucsd.edu

Introduction:

Chemical cross-linking and mass spectrometry have been shown to constitute a powerful tool to study protein-protein interactions and to help elucidate the structure of large protein complexes. However computational methods to interpret the convoluted MS/MS spectra from linked peptides are still in their infancy, thus making the high-throughput application of this approach largely impractical. Here we use disulfide-linked peptides as an example to describe a generic procedure to a) efficiently generate large mass spectral reference data for linked peptides and b) use this data to automatically train an algorithm that can efficiently and accurately identify linked peptides from MS/MS spectra.

Challenges in identification of linked peptides:

1) Linked Peptides has substantially different fragmentation pattern than unlinked, linear petpides

2) MS/MS spectra from linked peptides contain a mixture of fragments from more than one peptides

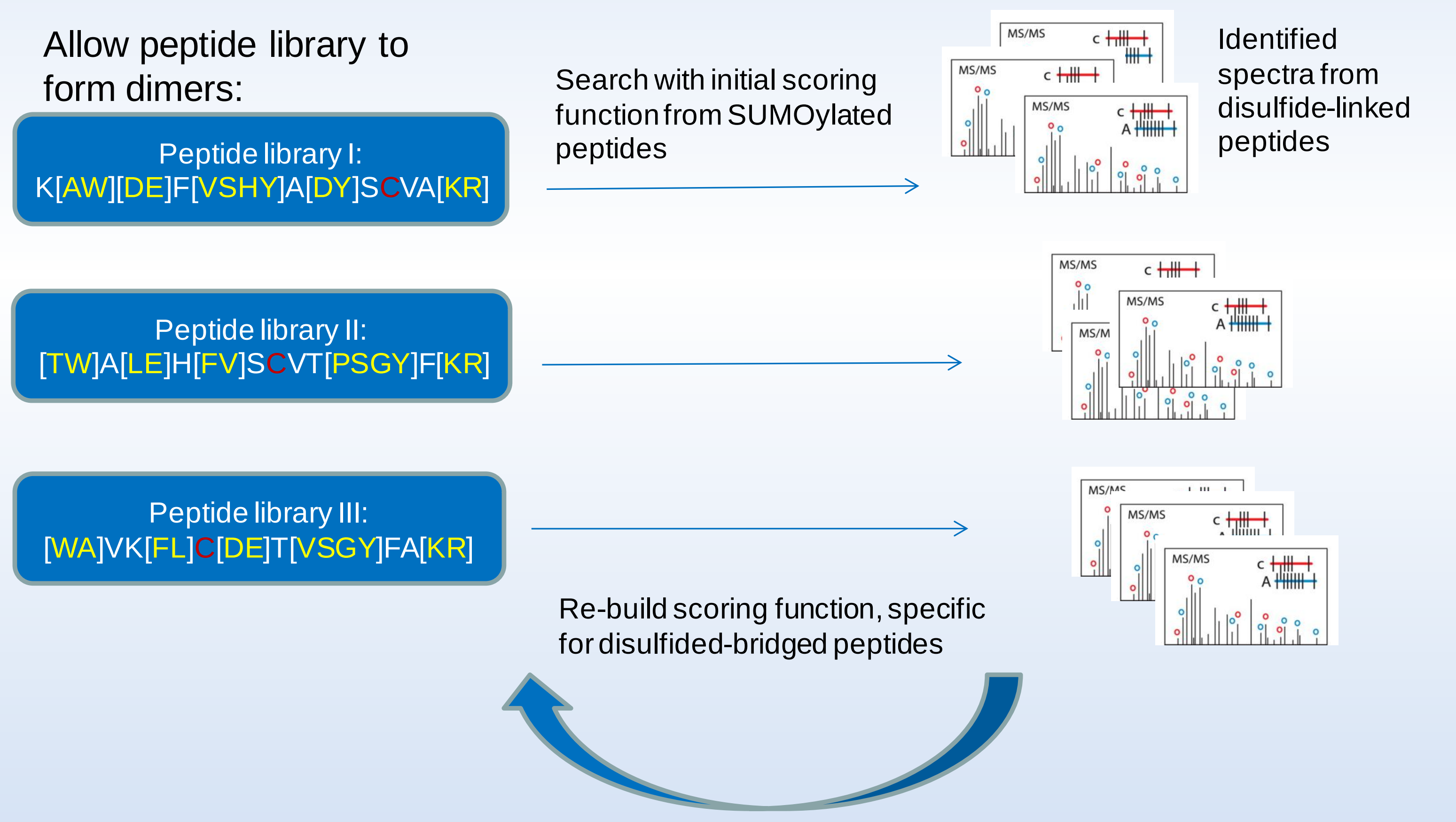
3) Lack of large and reliable annotated dataset to learn fragmentation pattern of linked peptides

Most current approaches/methods either use fragmentation models from unlinked, linear peptides or learn the model from data with limited size, which may not generalized well.

Method:

I. Generating large training datasets using combinatorial synthetic peptide libraries:

Building efficient and accurate scoring models for peptide identification usually requires a large set of reliably identified spectra. However, such datasets are usually hard/impossible to obtain without first having a computational method to identify those spectra in the first place: a “chicken and egg” problem. Here we break this cycle using combinatorial synthetic peptide libraries:



Allow peptide library to form dimers:

Peptide library I: K[AW][DE]F[VSHY]A[DY]SCVA[KR]

Peptide library II: [TW]A[LE]H[FV]SCVT[PSGY]F[KR]

Peptide library III: [WA]VK[FL]C[DE]T[VSGY]FA[KR]

Search with initial scoring function from SUMOylated peptides

Identified spectra from disulfide-linked peptides

Re-build scoring function, specific for disulfided-bridged peptides

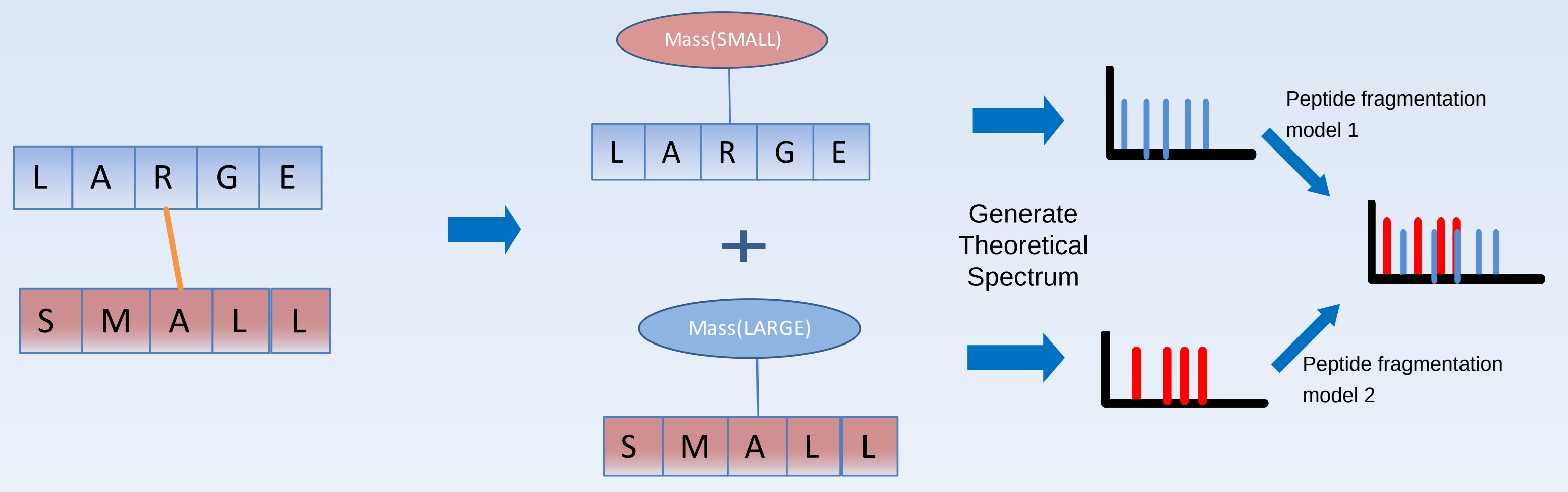
Reference:

[1] J.Wang, PE. Bourne, N Bandeira MCP(10) 2011
[2] Leither, et. al. MCP 2012

II. *Developing scoring function for linked peptides:*

1) Linked peptides are modeled as a mixture of two peptides with PTM

We model fragments of linked peptides as a mixture of fragment ions from two peptides, each carrying a PTM with mass of the other peptide at the linking site. Each peptide will be scored with a different scoring models accounting for the fact that two peptides are presented in the same MS/MS spectrum.



Generate Theoretical Spectrum

Peptide fragmentation model 1

Peptide fragmentation model 2

2) Probabilistic scoring model for a peptide pair

Our scoring function is base upon a probabilistic model that describes how a pair of co-eluting peptides fragments in a mixture MS/MS spectra [3]. We obtain model parameters for each peptide, accounting for their difference in fragmentation patterns.

Spectrum: represented as vector of peak rank (rank by intensity)

S = [0, 10, 0 , 0, 40, 0, 80, 0,10, 100, 50, 0, 5, 90, 0 ......] 0: no peak presented

Peptide: represented as vector of ion-types

P = [0, b, 0 , y, 0, 0, b-H20, 0, y, 0, 0, 0, b, 0 ......] 0: noise peak

$$Score = \log\frac{\Pr(s1|p1)}{\Pr(s1|0)} + \log\frac{\Pr(s2|p2)}{\Pr(s2|0)} + \ldots\ldots + \log\frac{\Pr(sn|pn)}{\Pr(sn|0)}$$

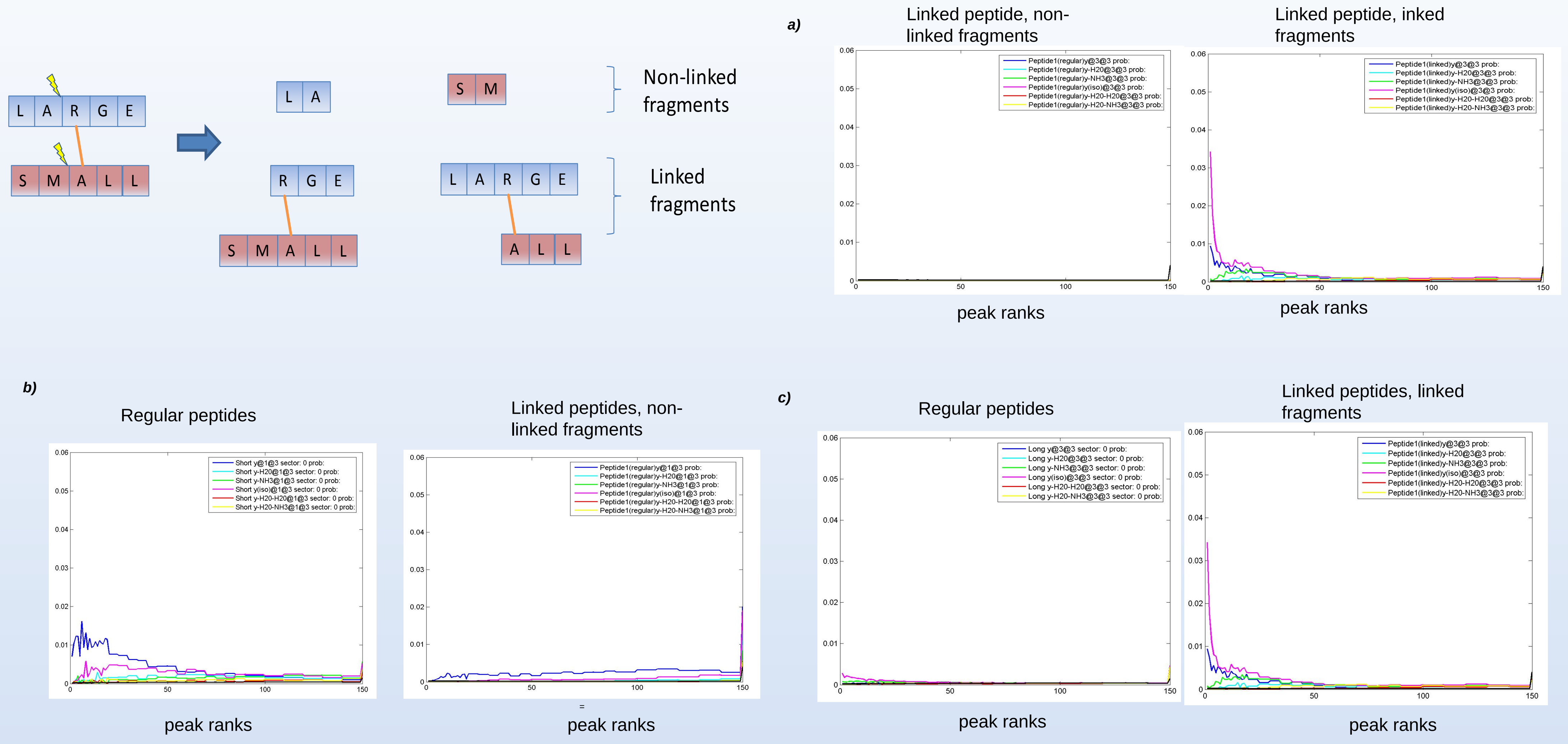
Lined peptides are represent as a mixture of two peptides with PTM:
e.g.LAR[+900]GE & SMA[+1500]LL assume Mass(LARGE)=1500, Mass(SMALL) = 900
-represent each peptide in vector format, then combine to represent a pair

```
P1 =      [ y,   b,  0,  y,   0,  0,   b-H20,   0,   y,   0,   0,       0,    b,   0   ...... ]
P2 =      [ y,   0,  0,  b,   b,  0,            0,   0,   0,   b,   0,  y-NH3,   0,   y   ...... ]
P1+P2 =  [y2,  b1,  0,  y1,  b2,  0,  b1-H20,   0,  y1,  b2,   0, y2-NH3,  b1,  y2   ...... ]
```

Learn parameters *Pr*(*si*|*pi*) from synthetic peptide library, separate scoring model for each peptide to account for their difference in fragmentation

3) Capturing linked-peptide specific fragmentation patterns

We separate fragment ions from linked peptides into two type of ions: linked and non-linked fragments. We notice three general characteristics about the fragmentation pattern of linked peptides: i) linked-fragments and non-linked fragments has VERY different fragmentation pattern (panel a); ii) Non-linked fragments of linked peptide, although share some similar characteristics with fragments from unlinked peptides, their intensity are generally suppressed (panel b); iii) Linked-fragments also has very different fragmentation compare to those of unlinked pepitdes, particularly high-charge fragments are very prominent (panel c).



Non-linked fragments

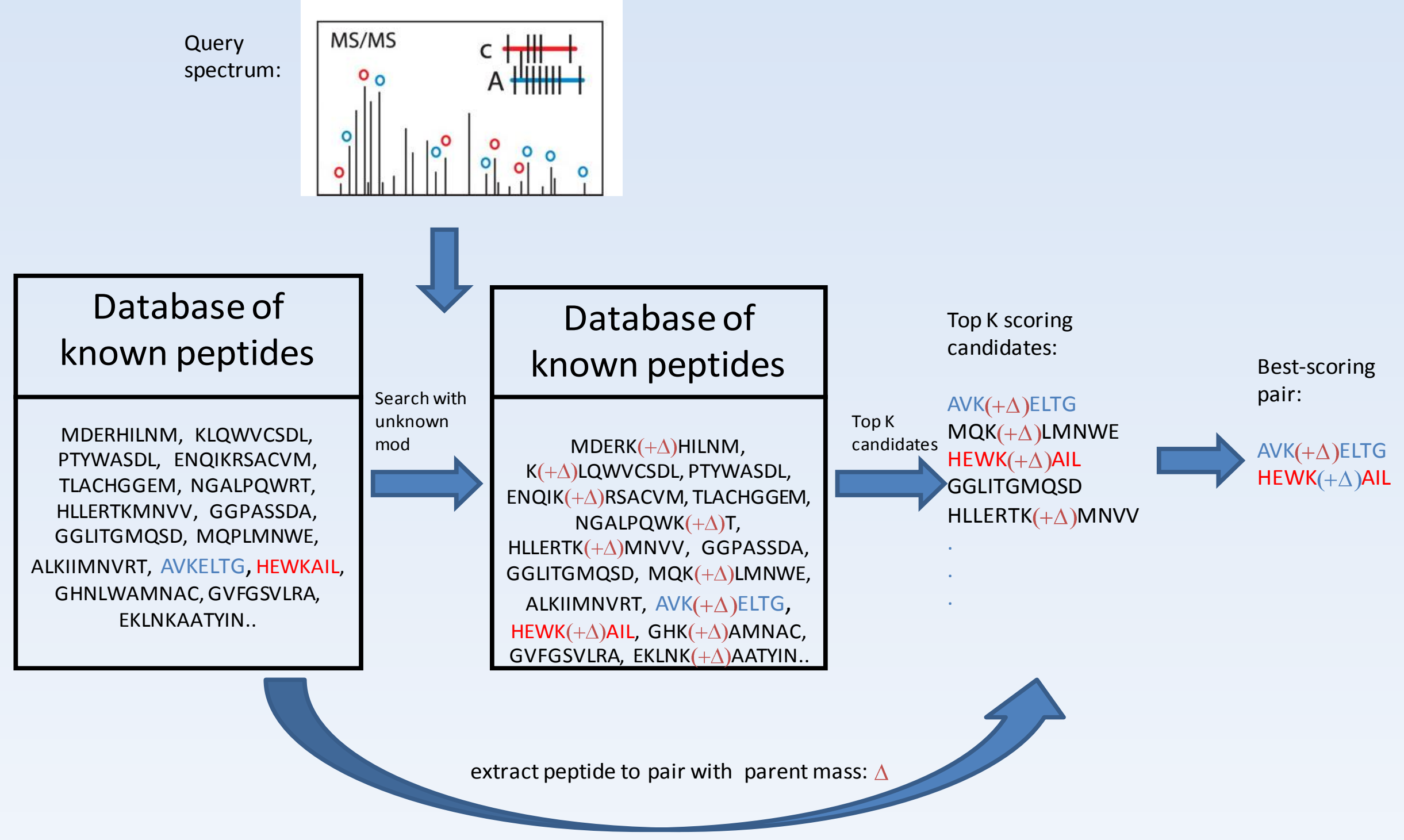
Linked fragments

a) Linked peptide, non-linked fragments — Linked peptide, inked fragments

b) Regular peptides — Linked peptides, non-linked fragments

c) Regular peptides — Linked peptides, linked fragments

III. *Workflow of database search method for linked peptides*



Query spectrum:

Database of known peptides

MDERHILNM, KLQWVCSDL, PTYWASDL, ENQIKRSACVM, TLACHGGEM, NGALPQWRT, HLLERTKMNVV, GGPASSDA, GGLITGMQSD, MQPLMNWE, ALKIIMNVRT, AVKELTG, HEWKAIL, GHNLWAMNAC, GVFGSVLRA, EKLNKAATYIN..

Search with unknown mod

Database of known peptides

MDERK(+Δ)HILNM, K(+Δ)LQWVCSDL, PTYWASDL, ENQIK(+Δ)RSACVM, TLACHGGEM, NGALPQWK(+Δ)T, HLLERTK(+Δ)MNVV, GGPASSDA, GGLITGMQSD, MQK(+Δ)LMNWE, ALKIIMNVRT, AVK(+Δ)ELTG, HEWK(+Δ)AIL, GHK(+Δ)AMNAC, GVFGSVLRA, EKLNK(+Δ)AATYIN..

Top K candidates

Top K scoring candidates:

AVK(+Δ)ELTG
MQK(+Δ)LMNWE
HEWK(+Δ)AIL
GGLITGMQSD
HLLERTK(+Δ)MNVV
.
.
.

Best-scoring pair:

AVK(+Δ)ELTG
HEWK(+Δ)AIL

extract peptide to pair with parent mass: Δ

Results:

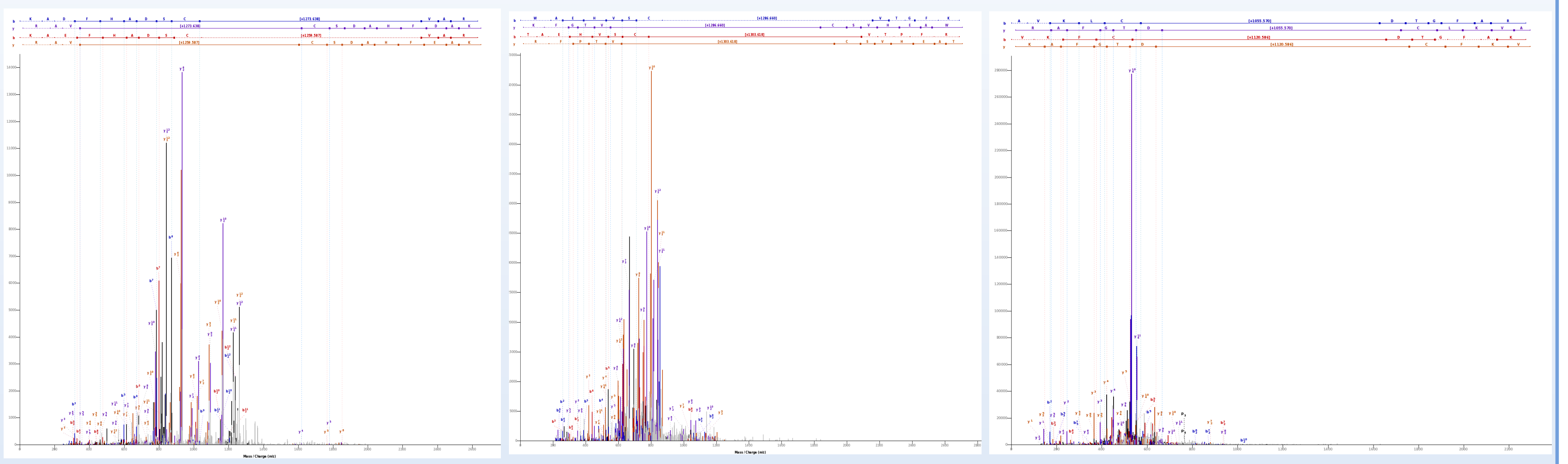
Dataset	# of MS/MS spectra identified (5% FDR)	Unique peptide pairs
Library 1 K[AW][DE]F[VSHY]A[DY]SCVA[KR]	2239	1190
Library 2 [TW]A[LE]H[FV]SCVT[PSGY]F[KR]	2636	995
Library 3 [WA]VK[FL]C[DE]T[VSGY]FA[KR]	1077	791

	# of MS/MS spectra identified from disulfides		
False discovery rate (FDR)	Library peptides only (~36,000 peptide pairs)	Library peptides + Ecoli decoy DB (~800X library)	Library peptides + Yeast decoy DB (~1900X library)
2%	1837	1796	1220
3%	1997	1853	1360
4%	2161	1919	1483
5%	2239	1971	1553

Sample	Cross-linker used	# dead-end-link IDs	# loop-links IDs	# cross-link IDs
Bovine Serum Albumin (69.3KDa)	BS3	151 (44*)	50 (10)	34 (23)
Rabbit Aldolase (157KDa)	BS3	958 (95)	534 (64)	106 (15)
Yeast 20S Proteasome (700KDa)	DSS	690 (255)	82 (50)	90 (23), 90% ⬆
Rabbit 20S Proteasome (1.4MDa)	DSS	500 (179)	258 (23)	284(64), 52% ⬆

*Proteasome data are from ref [2]

Examples of identified MS/MS spectra from linked peptides



Conclusion:

- Current database search methods do not try to capture the specific fragmentation of linked peptides because there are limited number of annotated data to learn fragmentation statistics of linked peptides
- Using disulfide-bridged peptides as an example, we demonstrate that the use of combinatorial synthetic peptide libraries is an efficient way to generate a *large* and *reliable* reference MS/MS dataset for linked peptides.
- We developed a rigorous probabilistic models that capture the specific fragmentation patterns of linked peptides.
- We show that this new approach can identify thousands of MS/MS spectra from disulfide-bridged peptides even against whole-proteome scale sequence database
- Our approach can be generalized to identify peptides with other linked peptides