Center for
Computational
Mass
Spectrometry

UCSD

# A turn-key approach for large-scale identification of SUMOylated peptides from tandem mass spectra

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

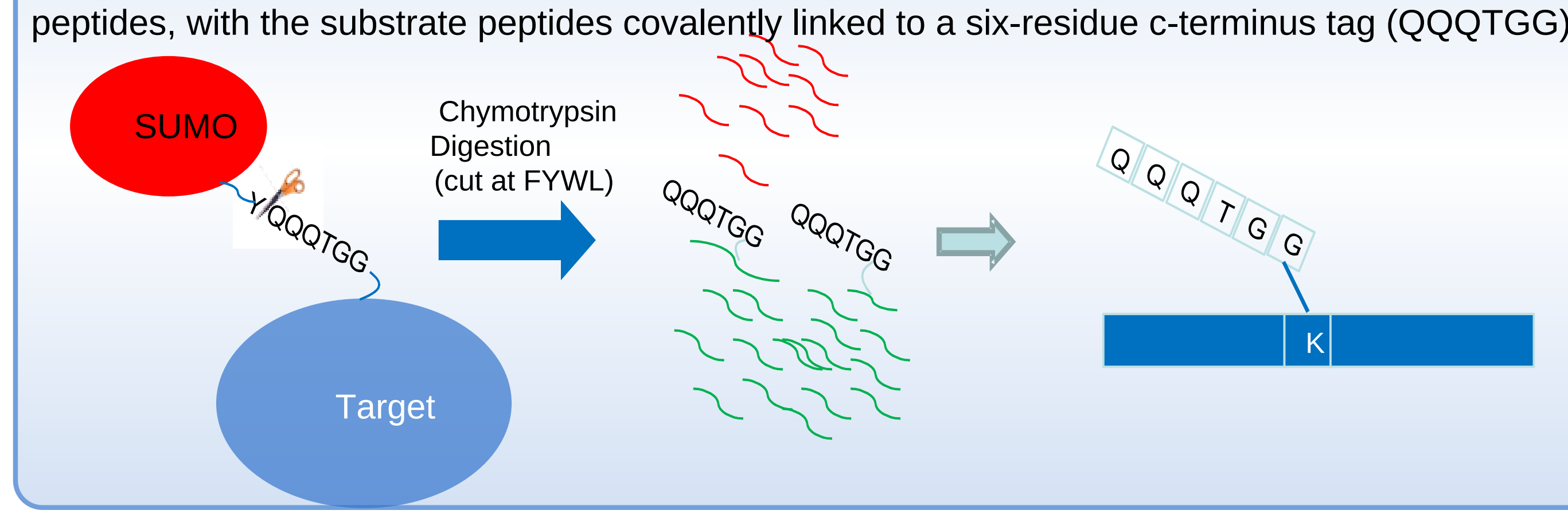
1. Bioinformatics Program, UCSD, La Jolla, CA
2. Protein Chemistry Department, Genentech Inc., 1 DNA Way, South San Francisco, California
3. Cell Signaling Technologies, Danvers, MA
4. Skaggs School of Pharmacy and Pharmaceutical Science, UCSD, La Jolla, CA
5. Center for Computational Mass Spectrometry,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

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

## Introduction:

Detection of complex post-translational modifications (PTMs) remains a challenging open problem in proteomics. Simple PTMs (e.g. methylation, deamidation etc.) can be readily identified by tandem mass (MS/MS) spectrometry by considering characteristic mass shifts in peptide fragment masses. However, more complex PTMs (e.g. glycosylation, small ubiquitin-like modification (SUMOylation) etc.) present a more difficult problem because 1) the PTM itself can generate multiple fragment ions and 2) it can cause significant changes in peptide fragmentation. Here we describe a procedure for generating statistical scoring models for identification of complex PTMs from MS/MS spectra and demonstrate how PTM- and site-specific scoring can dramatically improve the identification of SUMOylated peptides to up to seven times more identifications than conventional database search

SUMOylation: small ubiquitin-like modifiers (SUMO) are small proteins (~100 residues) that attach to target proteins to regulate their function. SUMO is known to be involved in regulation of intra-nuclear trafficking, cell cycle, DNA repair, replication and stress responses[1] . It is also linked to neurodegenerative diseases such as Huntington and Alzheimer's disease [2]. After chymotrypsin digestion, SUMOlyated proteins generate "y-shaped" peptides, with the substrate peptides covalently linked to a six-residue c-terminus tag (QQQTGG) from SUMO.

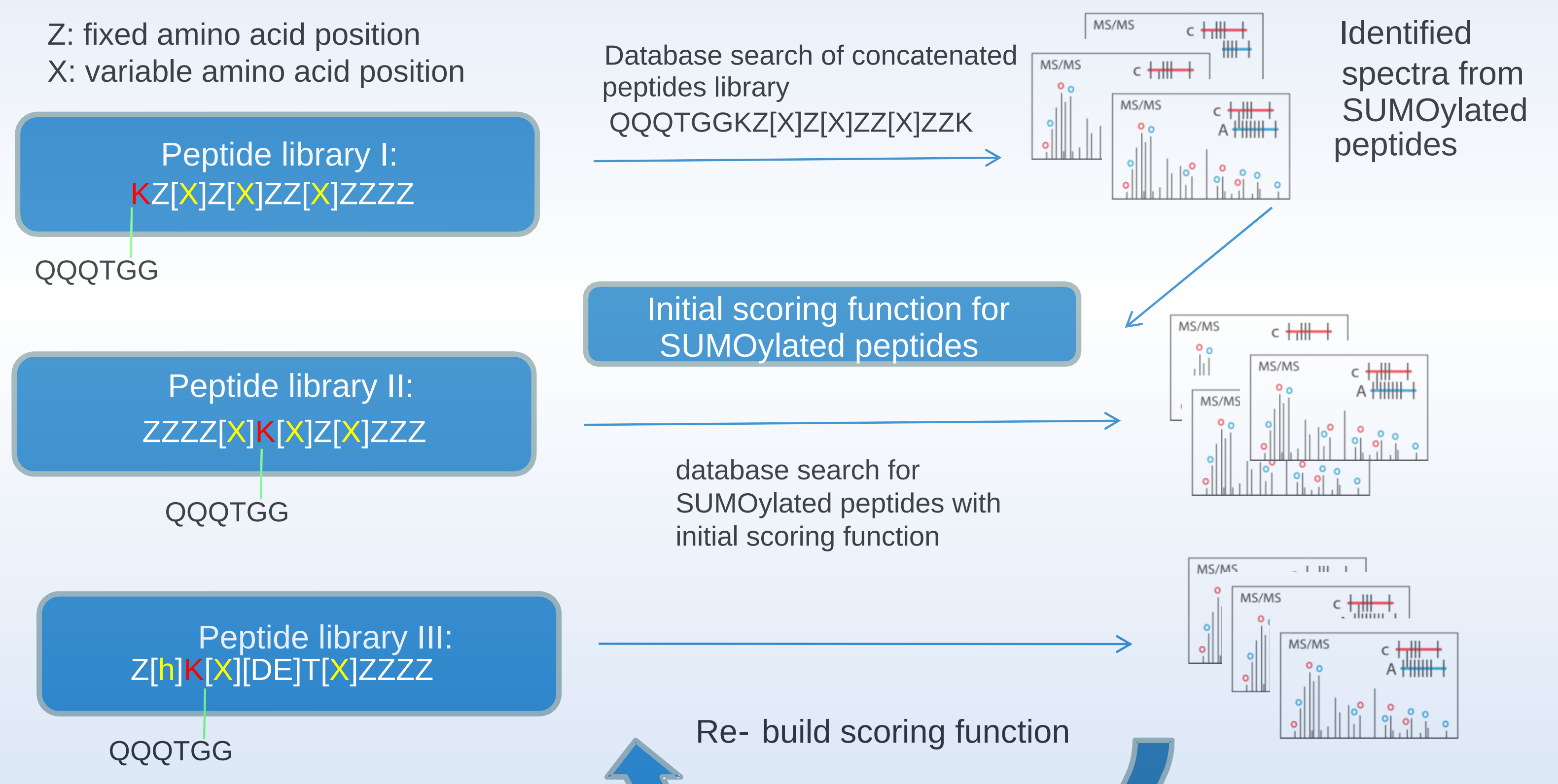


## Method:

*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:

Z: fixed amino acid position
X: variable amino acid position


Peptide library I:
KZ[X]Z[X]ZZ[X]ZZZZ
QQQTGG
Database search of concatenated peptides library
QQQTGGKZ[X]Z[X]ZZ[X]ZZK
MS/MS
c
A
Identified spectra from SUMOylated peptides
Initial scoring function for SUMOylated peptides
Peptide library II:
ZZZZ[X]K[X]Z[X]ZZZ
QQQTGG
database search for SUMOylated peptides with initial scoring function
Peptide library III:
Z[h]K[X][DE]T[X]ZZZZ
QQQTGG
Re- build scoring function

Library 3 is base on the canonical SUMOlyation motif:
h-K-X-[D/E] where h is a aliphatic residue, X is any amino acid, D or E is an acidic residue

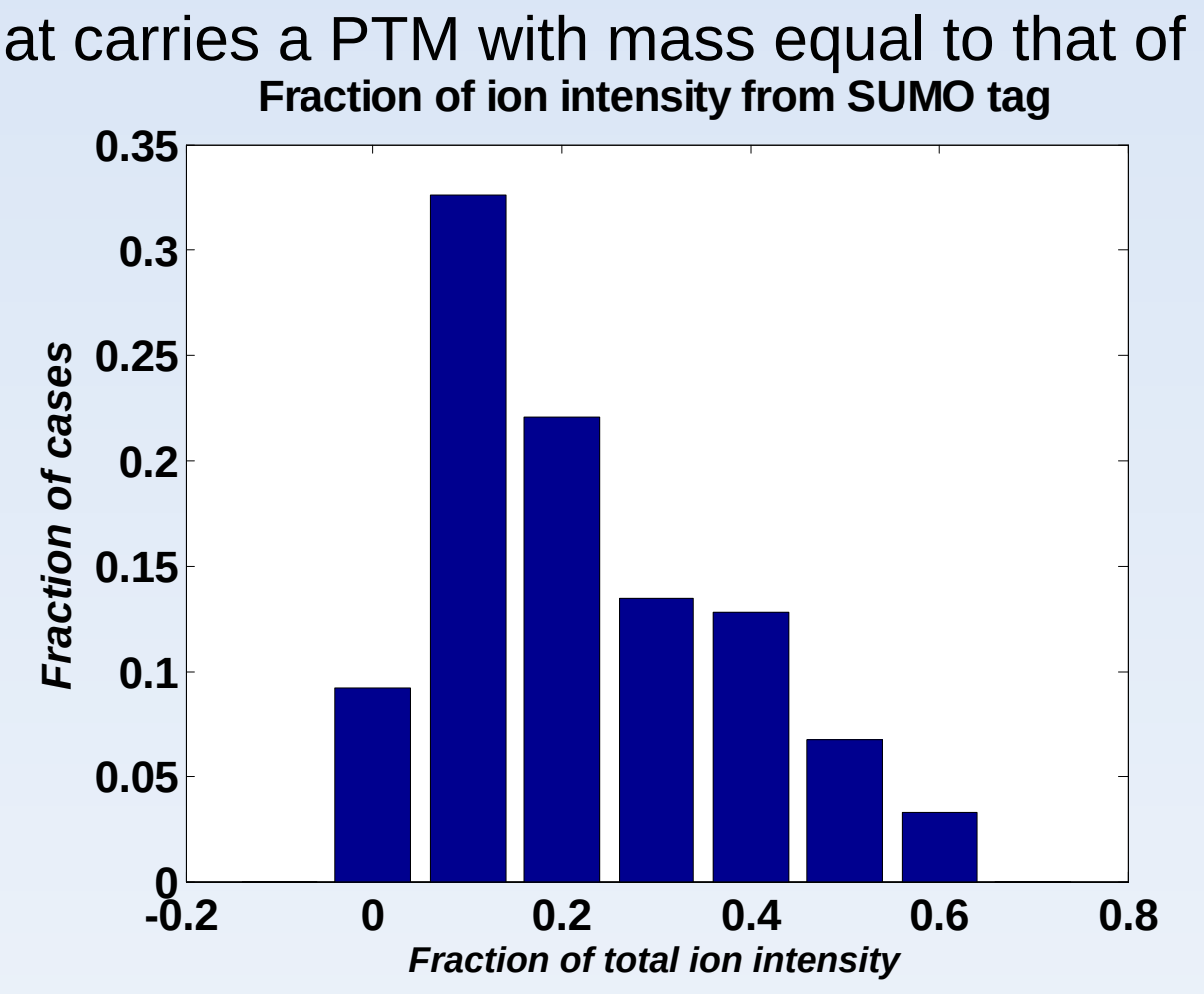
## Reference:

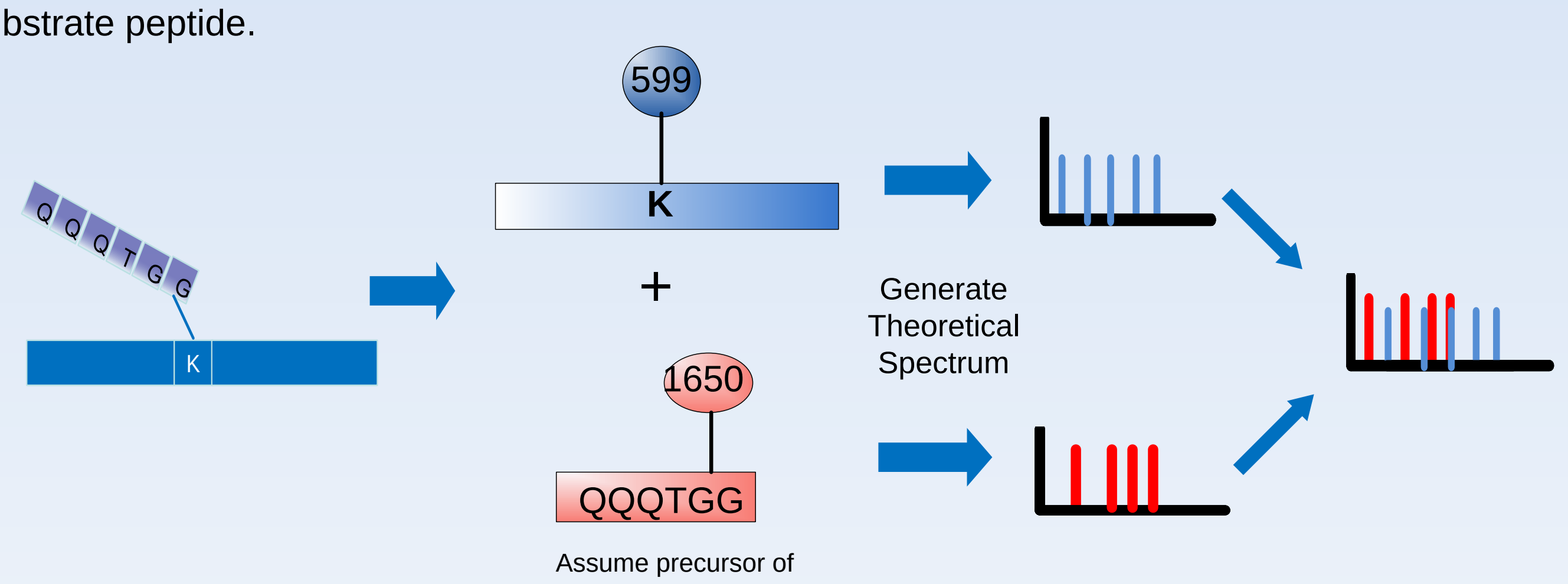
[1] F. Galission et. al. MCP 2010 (10):1535-9476
[2] E. Meulmeeter & F. Melchior Nature 2008 (452):709-711
[3] J.Wang, PE. Bourne, N Bandeira Database search method for mixture tandem mass spectra ASMS 2010 poster

II. *Developing a PTM specific scoring function:*

1) Accounting for fragment ions from PTM

PTM fragment ions from SUMO tag (QQQTGG) contribute 10-60% to the total ion intensity in a given MS/MS spectra. We therefore model fragments of a SUMOylated peptides as a mixture of fragment ions from two peptides, one is the substrate peptide carry a +599 PTM and the other is the tag (QQQTGG) that carries a PTM with mass equal to that of the substrate peptide.





Assume precursor of substrate peptide is 1650

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 the substrate peptides and the SUMO tag separately, accounting for their difference in fragmentation patterns.

Spectrum: represented as vector of peak rank (rank by intensity)[1]

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

SUMOylated peptides are represent as a mixture of two peptides with PTM:
e.g. FVIGGK[+599]PQR & QQQTGG[+1000]
-represent each peptide in vector format, then combine to represent a pair

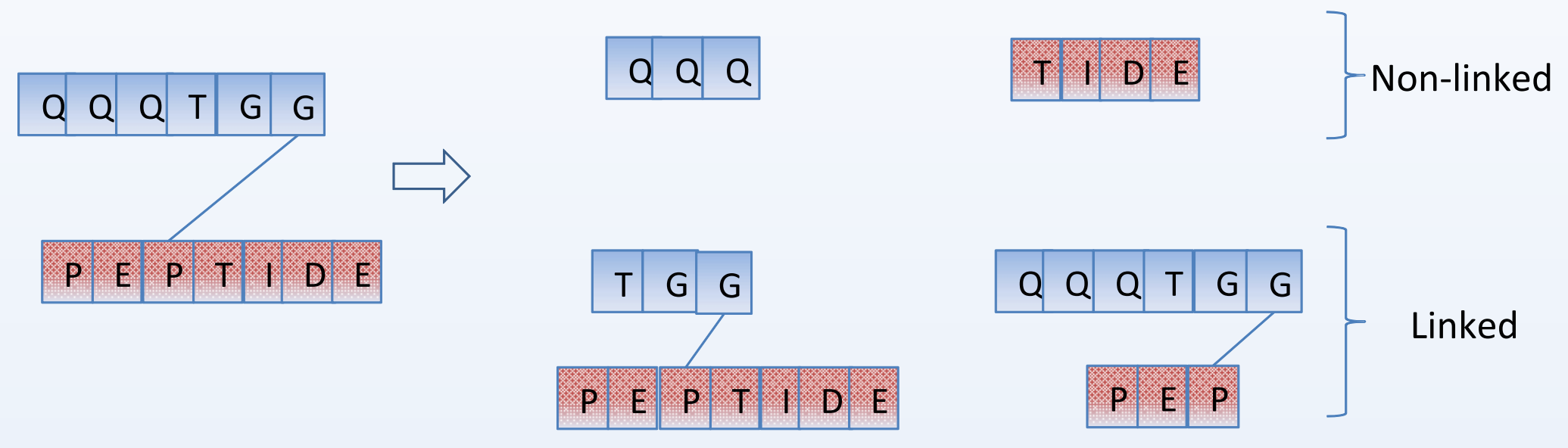
```
P1 =      [ y,   b,  0,  y,   0,  0,   b-H20,   0,   y,    0,   0,        0,    b,   0  …… ]
P2 =      [ y,   0,  0,  b,   b,  0,            0,   0,    0,   b,   0,  y-NH3,    0,   y  …… ]
P1+P2 =  [y2,  b1,  0,  y1,  b2,  0,  b1-H20,   0,  y1,  b2,   0, y2-NH3,  b1,  y2  …… ]
```

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

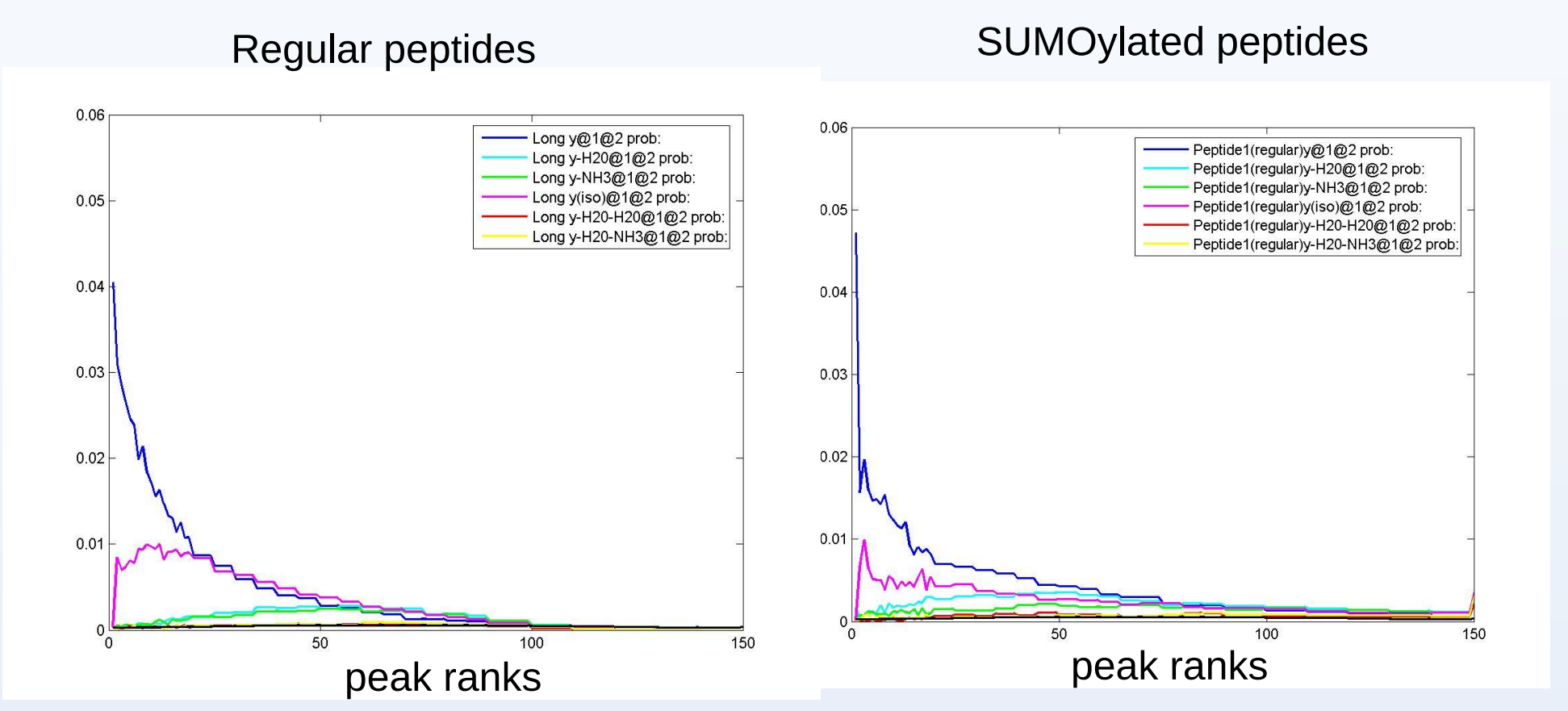
Learn parameters *Pr*(*si*|*pi*) from synthetic peptide library, separate scoring model for substrate peptide and SUMO tag to account for their difference in fragmentation

3) Different scoring model for linked and non-linked (non-linked) fragments

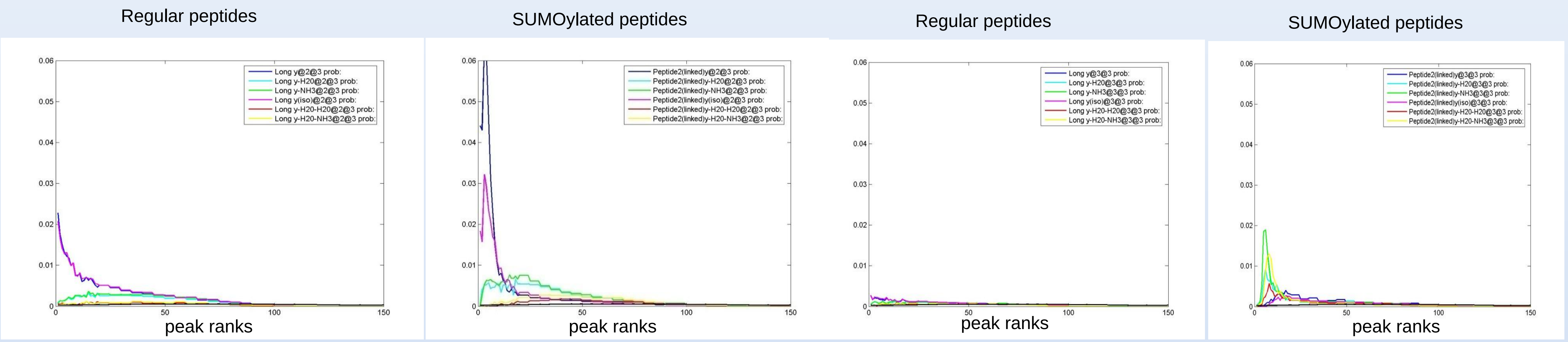
Fragmentation of SUMOylated peptides generate two type of fragments ions: linked and non-linked fragments. As illustrated below, normal fragments consists of fragmentation pattern similar to those seen in linear peptides, while linked fragments poses alternative fragmentation patterns. Particularly, due to their linkage to a second peptide, highly-charged fragment ions are more typically prevalent.



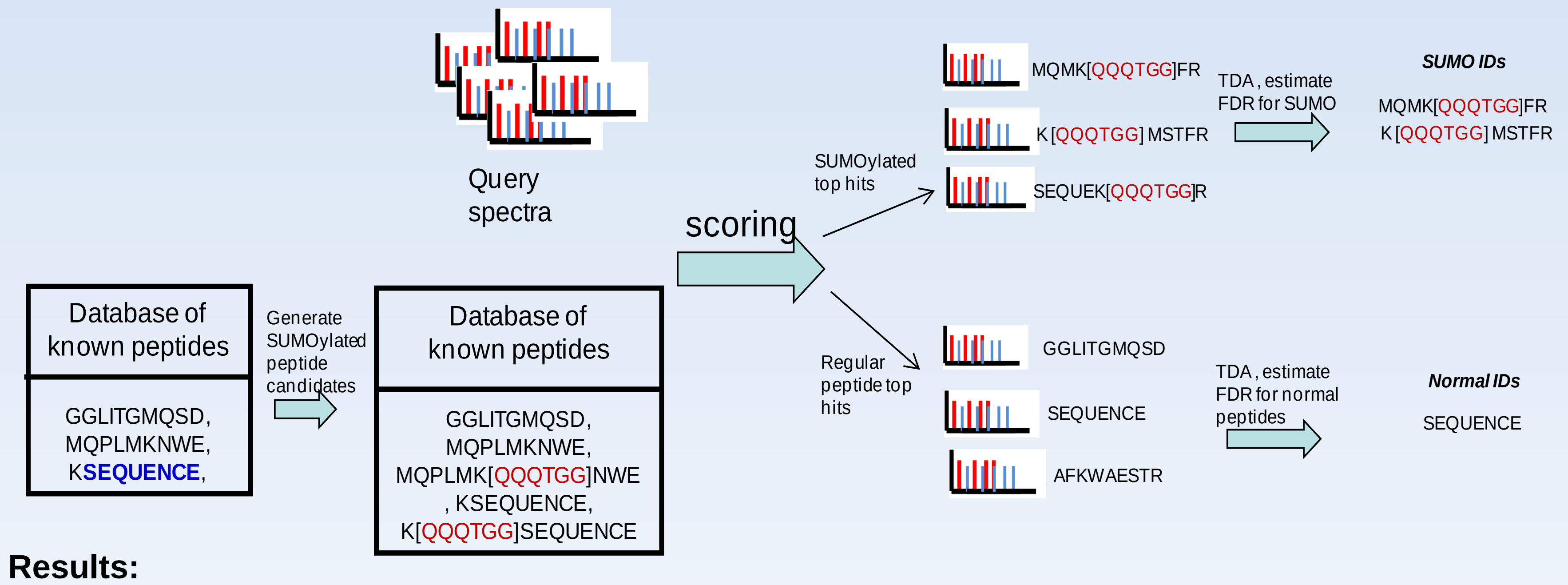
Non-linked fragments:



Linked fragments:



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


Query spectra
Database of known peptides
GGLITGMQSD, MQPLMKNWE, KSEQUENCE,
Generate SUMOylated peptide candidates
Database of known peptides
GGLITGMQSD, MQPLMKNWE, MQPLMK[QQQTGG]NWE , KSEQUENCE, K[QQQTGG]SEQUENCE
scoring
SUMOylated top hits
MQMK[QQQTGG]FR
K[QQQTGG]MSTFR
SEQUEK[QQQTGG]R
TDA, estimate FDR for SUMO
SUMO IDs
MQMK[QQQTGG]FR
K[QQQTGG]MSTFR
Regular peptide top hits
GGLITGMQSD
SEQUENCE
AFKWAESTR
TDA, estimate FDR for normal peptides
Normal IDs
SEQUENCE

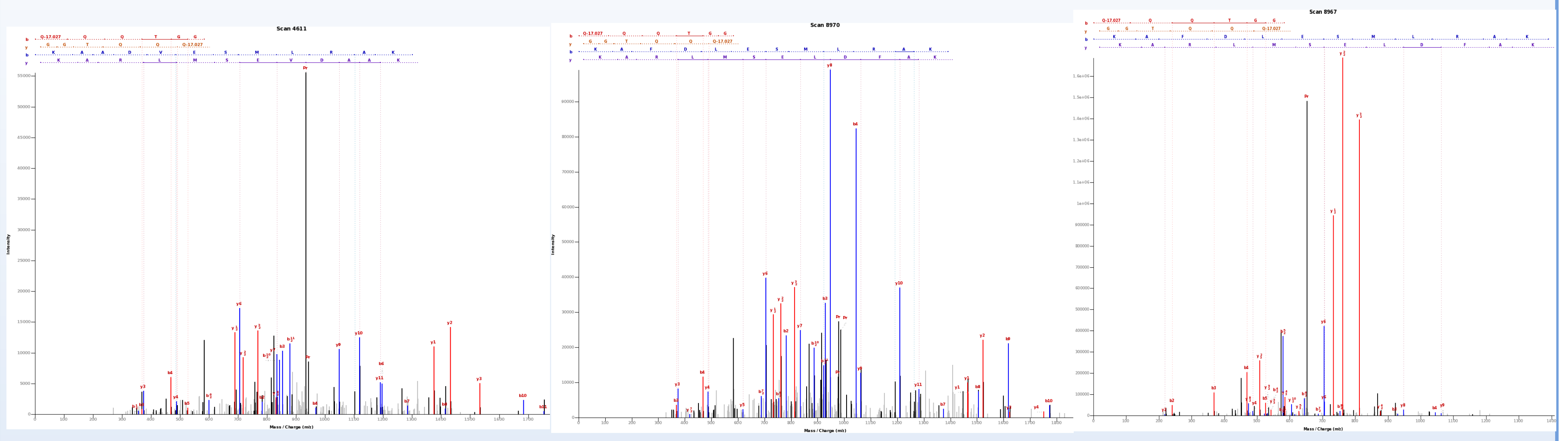
## Results:

*I. Combinatorial synthetic peptide libraries:*

MS/MS spectra from the SUMOylated synthetic peptide library were analyzed using: 1) InsPecT allowing +599 (SUMOylation) and +582 (SUMOylation with pyro-glutamic acid on the tag) modification and 2) MXDB, our new database search tool with scoring model specific for SUMOylated peptides using a 3.0 Da precursor tolerance and 0.5 Da fragment mass tolerance against a database consist of all E.Coli proteins sequences and the library peptide sequences. Results for MXDB are obtained using two fold cross-validation. As shown below at 1% false discovery rate, the identification rate of InsPecT for SUMOylated peptides decrease significantly as compare to its identification rate on ordinary tryptic peptides while the identification rate of MXDB for SUMOylated peptides is comparable to InsPecT's identification rate on normal peptides.

| | # of identified spectra from SUMOylated peptides (identification rate) | | | # of identified spectra |
|---|---|---|---|---|
| Dataset | Lib_sumo1 | Lib_sumo2 | Lib_sumo3 | Yeast tryptic digest of cell lystate |
| InsPecT | 743 (6.1%) | 1826 (16.4%) | 392 (3.2%) | 22658 (29.7%) |
| MXDB | 2320 (19.0%) | 4967 (44.7%) | 2929 (24.2%) | n/a |
| # of MS/MS spectra | 12202 | 11113 | 12177 | 76177 |

Examples of identified MS/MS spectra from SUMOylated peptides



Conclusion:

- Current database methods that identify PTMs by considering characteristic mass shift in peptide fragment masses do not handle complex PTM well for two reasons: 1) the PTM itself generate fragment ions that contribute significantly to the total ion intensity in the MS/MS spectra and 2) the PTM causes significant changes to peptide fragmentation pattern of the peptides
- 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 SUMOylated peptides.
- We developed a rigorous probabilistic models that capture the specific fragmentation patterns of SUMOylated peptides.
- We show that this new approach can identify three to seven times more spectra from SUMOlyated peptides compare to current database search tools and that its overall identification rate for SUMOylated peptides is comparable to current database search tool's identification rate for ordinary (linear, unmodified) tryptic peptides.
- Our approach can be generalized to identify peptides with other complex PTMs.