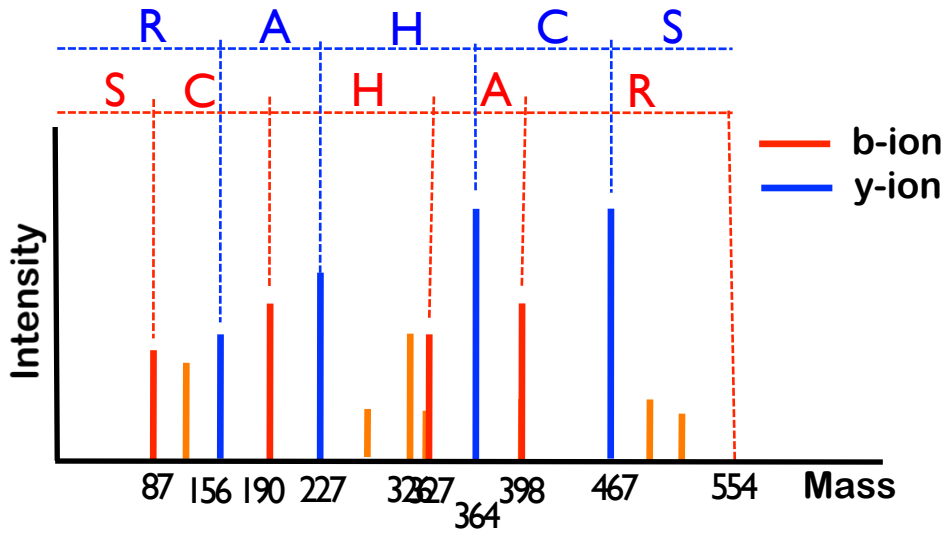
Time to say hello to MS-GFDB and say goodbye to InsPecT

Sangtae Kim
University of California San Diego

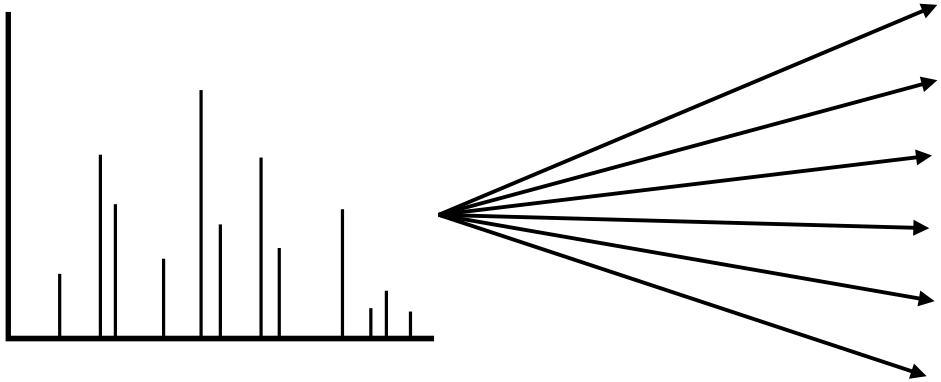
Center for
Computational
Mass
Spectrometry

Mass Spectrometry Based Proteomics



Identifying a MS/MS spectrum to a peptide.

MS/MS Database Search



Protein Sequence Database

```
>Seq1
ARNDDQGGHILKMFKKLILLK
>Seq2
MFPVYWTSPNRAARNDCEHLL
>Seq3
KMMMYVVVYWPSSFMKILLHG
>Seq4
QEQGGHILLKMMFPSDDQQGH
>Seq5
HKLMFPSTWYVVDRNNASSCE
>Seq6
FFPPFSTWWYVEQGHHDDCNE
……
```

InsPecT

- First released in 2005
- Served us well for 6 years

Time to Retire InsPecT

Mr. InsPecT

Ms. MS-GFDB

Which is better?

	InsPecT	MS-GFDB
Identifying more peptides at 1% FDR		✓
Applicability to various spectral types		✓
Easiness to use		✓
Search speed		✓

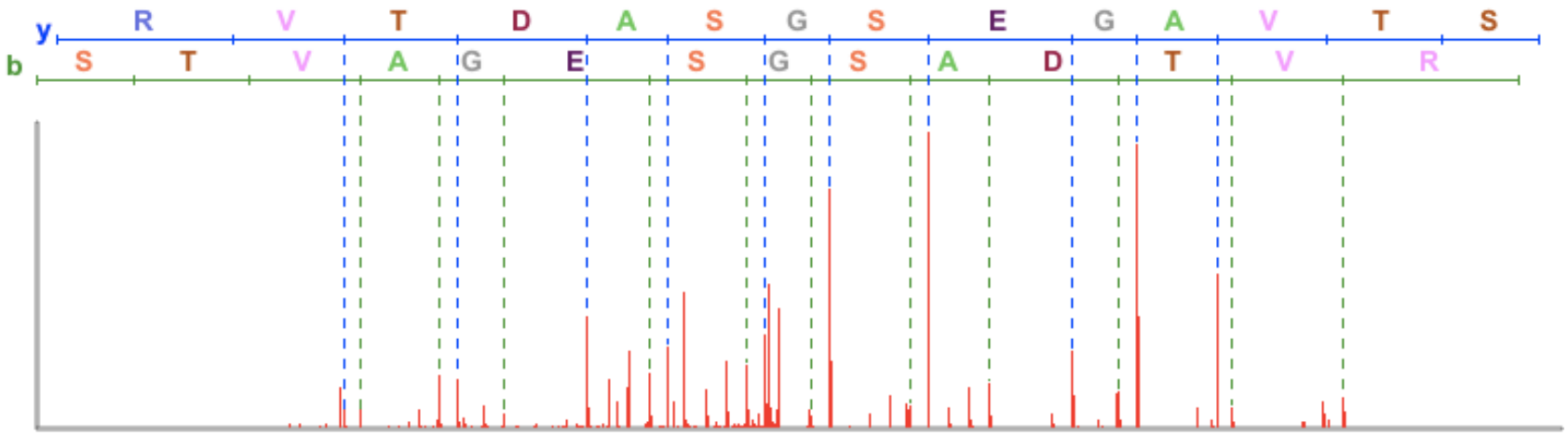
Outline I – Something Old

- The generating function approach (MS-GF)
- MS-GF scoring
- Previous results

Outline II – Something Fresh

- How to extend the generating function approach to modified peptides?
- How to benefit from high-precision MS/MS spectra?
- How to efficiently search the database?
- New results
  - [CID|ETD]-[HighRes|LowRes], HCD
  - α LP: new enzyme
  - Phosphorylation

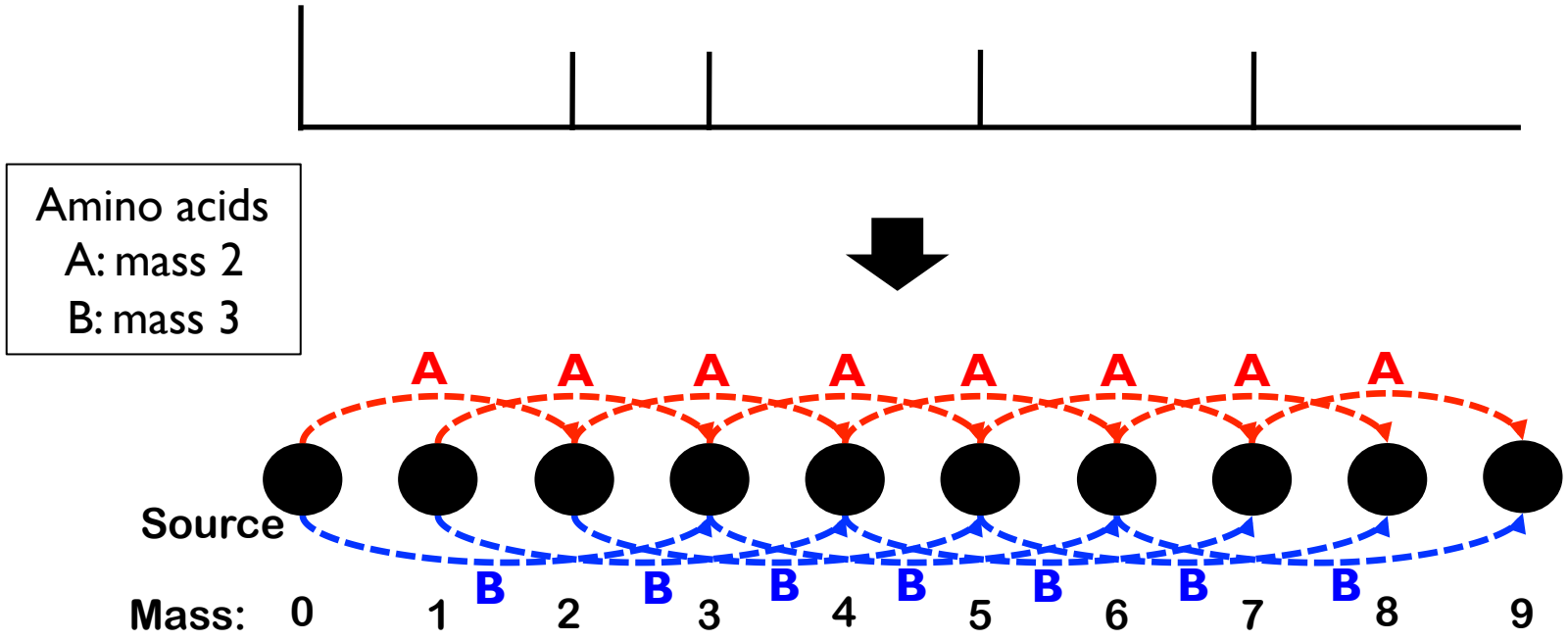
Generating Function Approach (MS-GF)



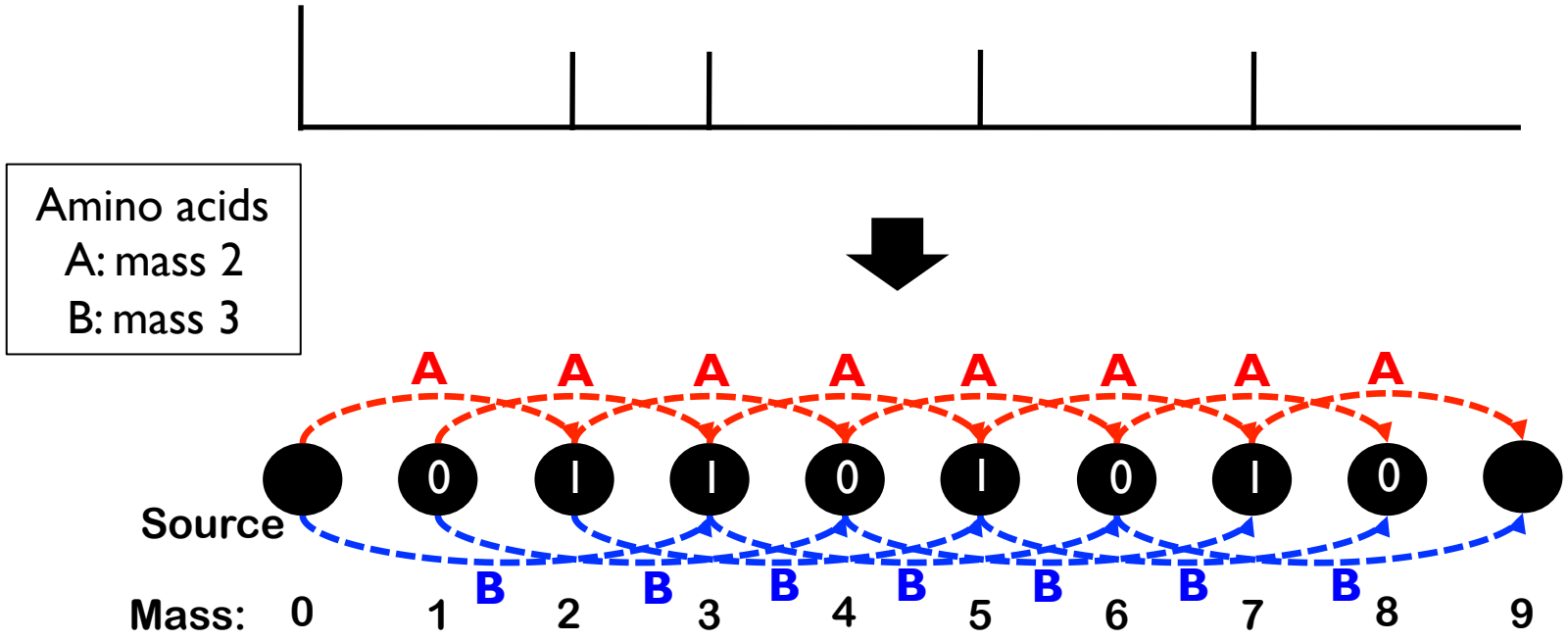
Given a Peptide-Spectrum Match (PSM),
what is the statistical significance (P-value or E-value)
of the PSM?

Kim et al., JPR 2008

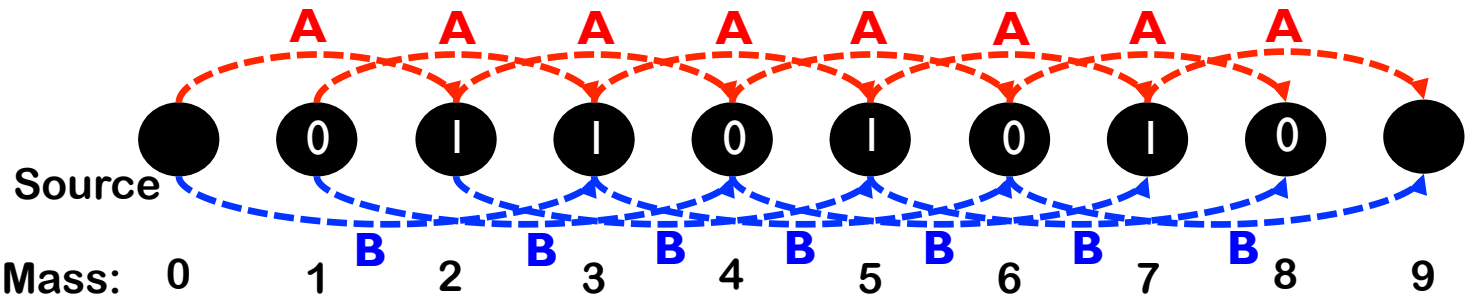
Converting Spectrum into Graph



Assign Scores to Nodes



Compute Score Histogram of All Peptides



Node Score	0	0	I	I	0	I	0	I	0	0	
Score=0	I	0	0	0	0	0	0	0	0	0	
Score=I	0	0	I	I	I	0	2	0	2	2	
Score=2	0	0	0	0	0	2	0	I	2	I	AABA
Score=3	0	0	0	0	0	0	0	2	0	2	ABAA BAAA

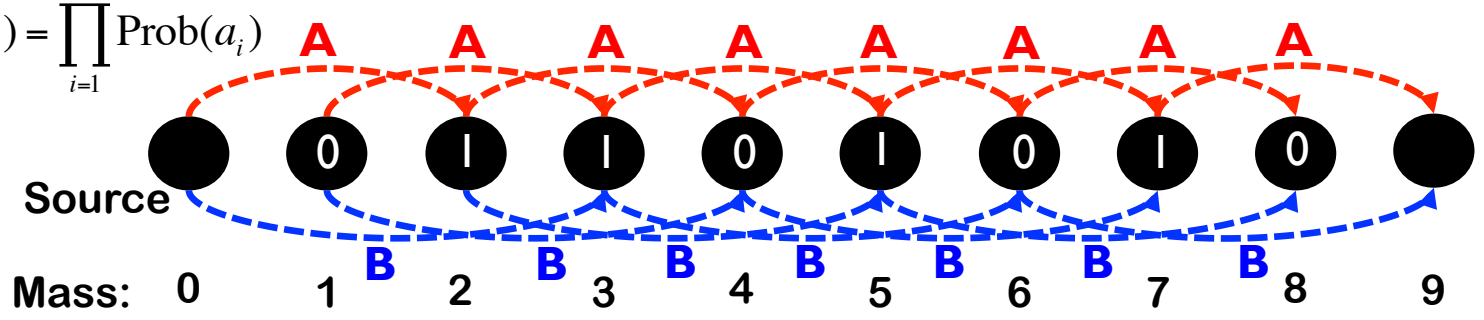
Score Histogram of All Peptides

Compute Score Histogram of All Peptides (Weighted)

$\text{Prob}(A) = 0.4$

$\text{Prob}(B) = 0.6$

$\text{Prob}(a_1 \ldots a_k) = \prod_{i=1}^{k} \text{Prob}(a_i)$

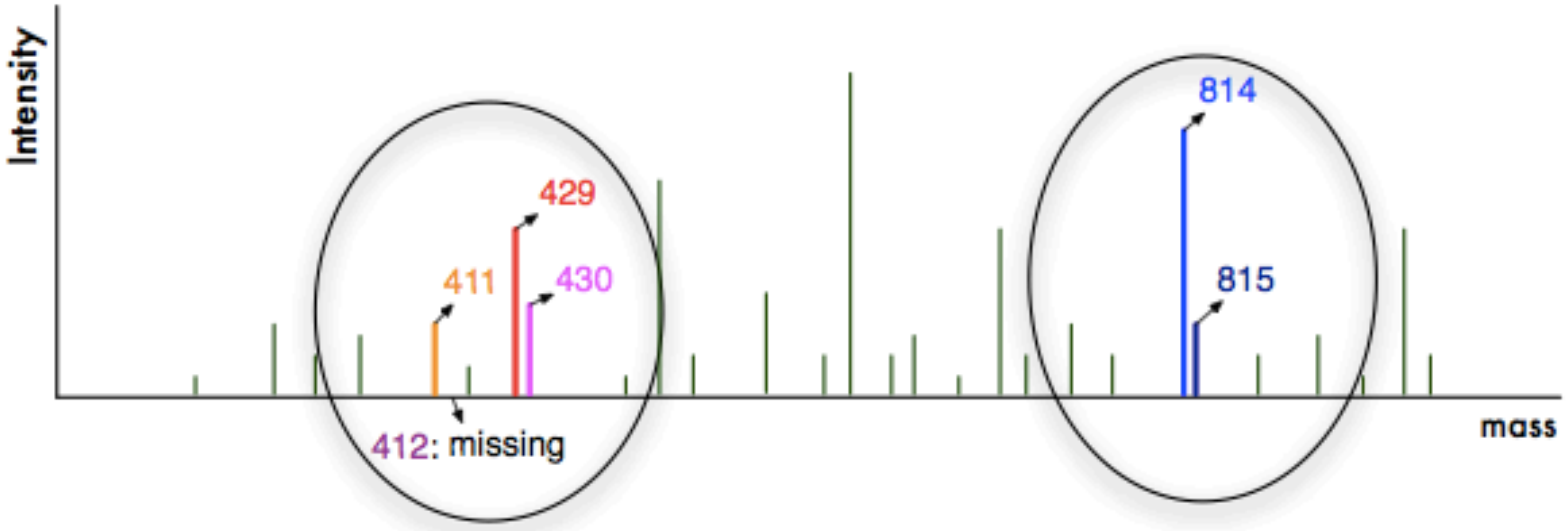


Node Score	0	0	1	1	0	1	0	1	0	0	
Score=0	1	0	0	0	0	0	0	0	0	0	
Score=1	0	0	.4	.6	.16	0	.424	0	.026	.254	
Score=2	0	0	0	0	0	.48	0	.096	.288	.038	AABA
Score=3	0	0	0	0	0	0	0	.192	0	.077	ABAA BAAA

SpecProb(Score=2)=0.038+0.077=0.115
=Prob(ABAA)+Prob(BAAA)+Prob(AABA)

SpecProb*DBSize
≈DB P-value

How to Compute Node Scores?



Prefix Residue Mass (PRM) : 428

	Ion types	Mass	Rank	Prob (Ion)	Prob (Noise)	Score
	b	429	5	0.15	0.001	5.01
	b+H	430	56	0.06	0.005	2.48
Rank score	b-H2O	411	69	0.05	0.006	2.12
	b-NH3	412	none	0.003	0.007	-0.85
	y	814	2	0.58	8.9E-4	6.47
	y+H	815	31	0.09	0.003	3.40

Rank score at 428:

5.01+2.48+2.12-0.85+6.47+3.40
= 18.6

MS-GF Scoring

Input:
Set of PSMs

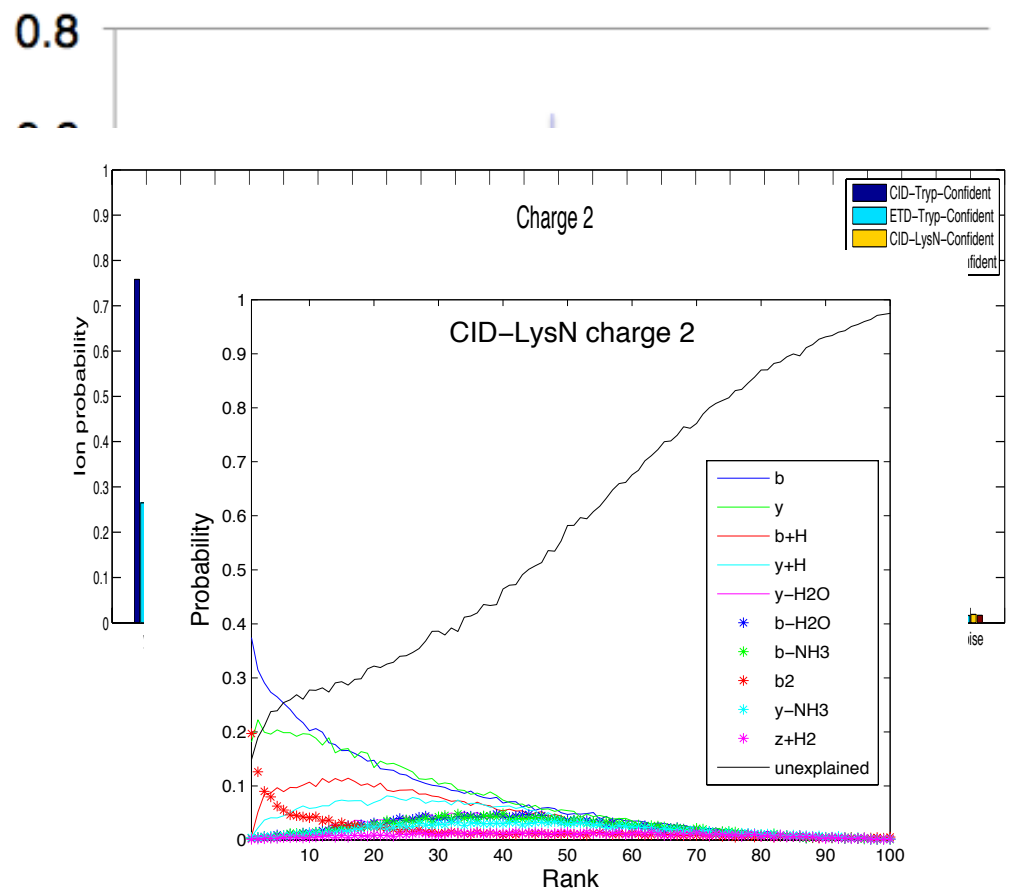
Partition PSMs
by Charges and Parent Masses

Remove Precursor Peaks
Offset Frequency Function of Precursors

Select Ion Types
OFF of Prefix/Suffix Masses

Compute Rank Probabilities
Compute Prob(ion/rank)

Output:
Scoring Parameter File



Previous Results

40% more spectra than OMSSA

20% more peptides than OMSSA

180% more peptides than Mascot (ETD)

300% more peptides than SEQUEST (ETD)

40% more peptides than PeptideProphet

32% more spectra than SEQUEST+iProphet

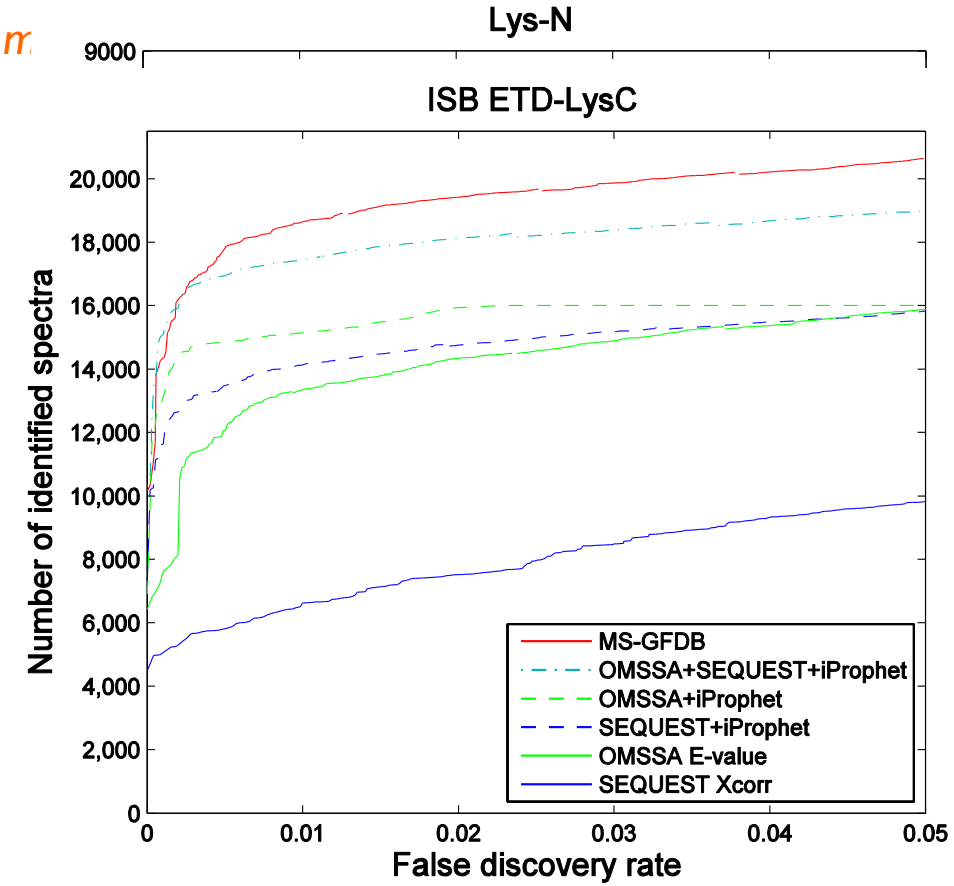
7% m

Tool

- MS-GF
- SEQUEST
- Mascot
- X!Tandem
- OMSSA
- InsPecT
- SEQUEST+ PeptideProphet
- OMSSA+ PeptideProphet

Spectrum Type

- CID
- ETD
- Trypsin
- Lys-N
- Lys-C



Kim et al., JPR 2008

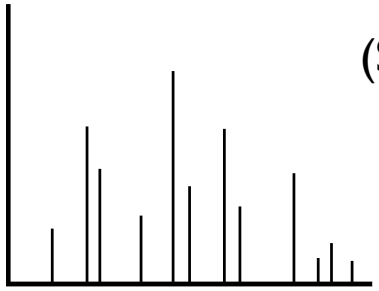
What happened after Chloe was born?

I did not mention about

- Modified peptides
- High-precision MS/MS spectra
- Search speed

Stat. Significance of Modified Peptides

Spectral Probability: the probability of finding an equal or better scoring peptide
(at a single position of the i.i.d. random database)



GAGM+16MSQR

Spectral Dictionary
(Set of equal or better scoring peptides)

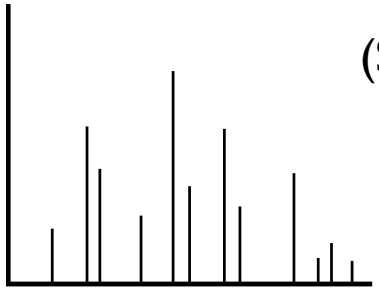
GAGM+16MSQR
GAGMM+16SQR
GQMM+16SGAR

} Same Peptide
(different mod variants)

SpecProb = Prob(GAGMMSQR)+Prob(GQMMSGAR)
not
2*Prob(GAGMMSQR)+Prob(GQMMSGAR)

From P-value to E-value

Spectral E-Value: the expected number of modification variants
with equal or better scores (at a single position of the i.i.d. random database)



GAGM+16MSQR

Spectral Dictionary
(Set of equal or better scoring peptides)

GAGM+16MSQR
GAGMM+16SQR
GQMM+16SGAR

} Same Peptide
(different mod variants)

SpecProb = Prob(GAGMMSQR)+Prob(GQMMSGAR)

SpecEValue = 2*Prob(GAGMMSQR)+Prob(GQMMSGAR)

Spectral Probability and E-Value

S : Spectrum

P : Peptide

$\text{Score}(S,P)$: score of P against S

- Spectral Probability (SpecProb)

$$\text{SpecProb}(S,P^*) = \sum_{\forall P \text{ where } \text{Score}(P,S) \geq \text{Score}(P^*,S)} \text{Prob}(P)$$

- Spectral E-Value (SpecVal)

$\text{Variants}(P)$: set of all modification variants of P

$$\text{Mult}(S,P,t) = \sum_{\forall V \in \text{Variants}(P) \text{ where } \text{Score}(S,V) \geq t} 1$$

$$\text{SpecVal}(S,V^*) = \sum_{P} \text{Prob}(P)\,\text{Mult}(S,P,\text{Score}(S,V^*))$$

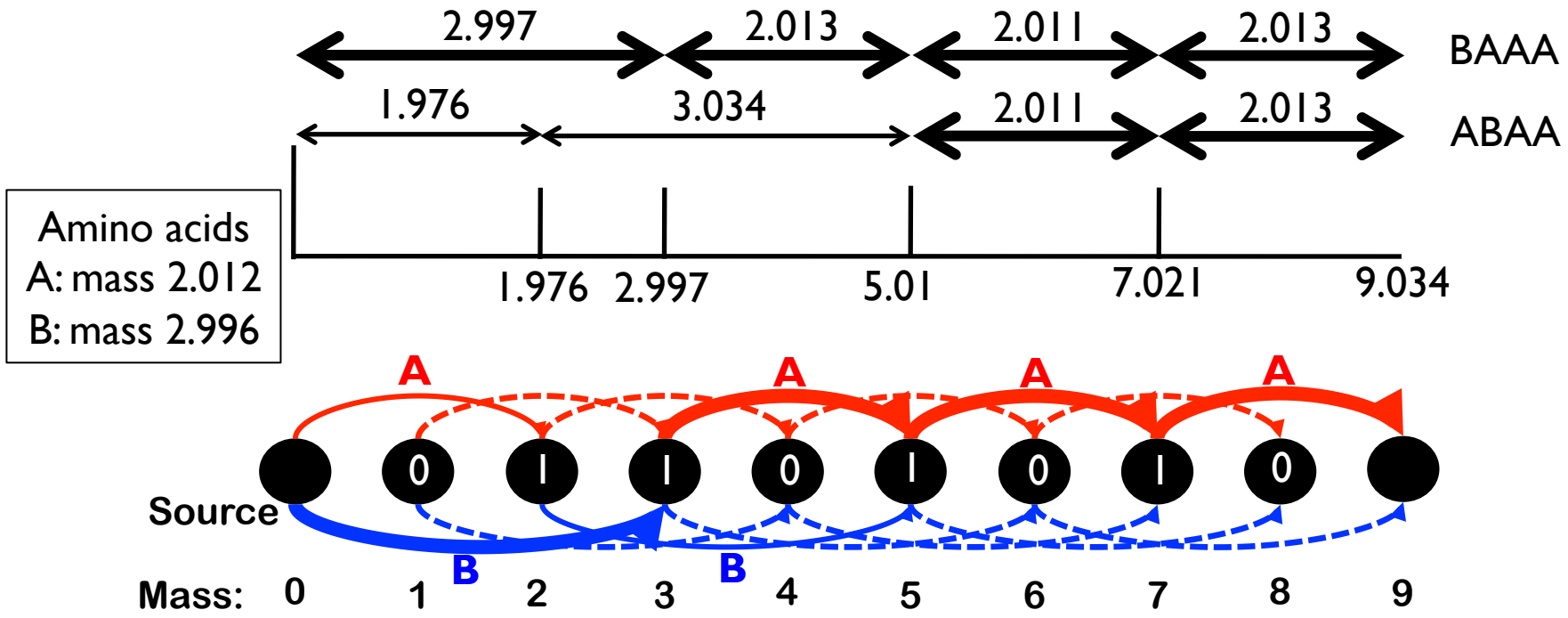
MS-GFDB now supports:

- Finding modified peptides
- Reporting SpecEValue

High-precision MS/MS Spectra

- How to benefit from high-precision MS/MS spectra?

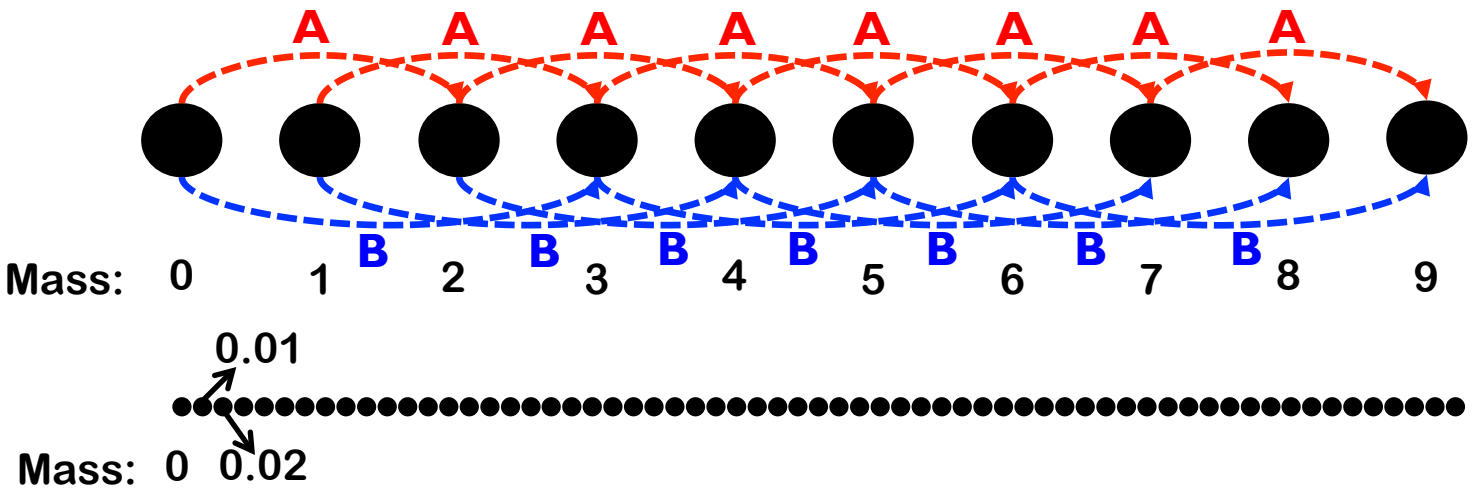
High-precision MS/MS Spectra



Score(BAAA) = Score(ABAA) = 3

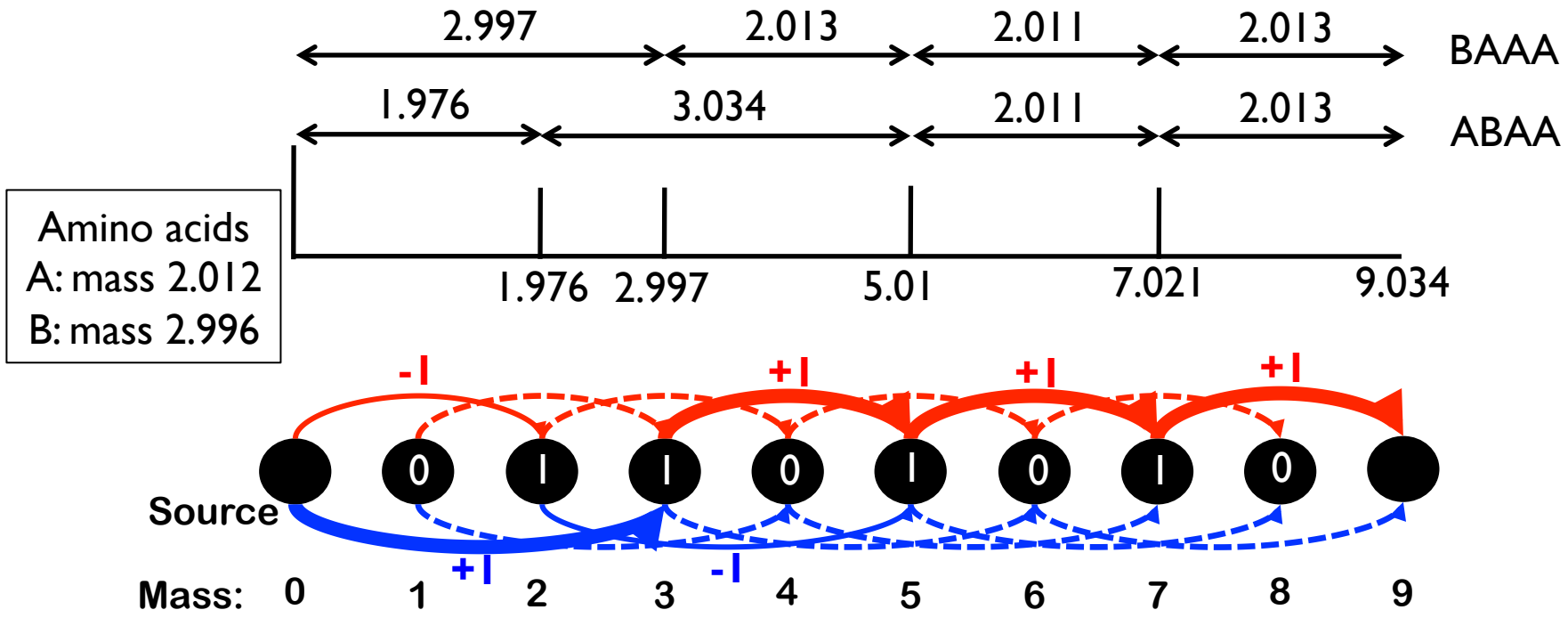
Which is more probable?

Using Smaller Bins? – Too slow



- Currently, spectrum peak masses are rescaled to the closest integer (nominal mass) after multiplying 0.999497
- Change the constant to 274.335215 (max error < 2ppm)
- MS-GF becomes 40X slower
- MS-GFDB identifies only 6% more PSMs at 1% FDR

Assigning Scores to Edges



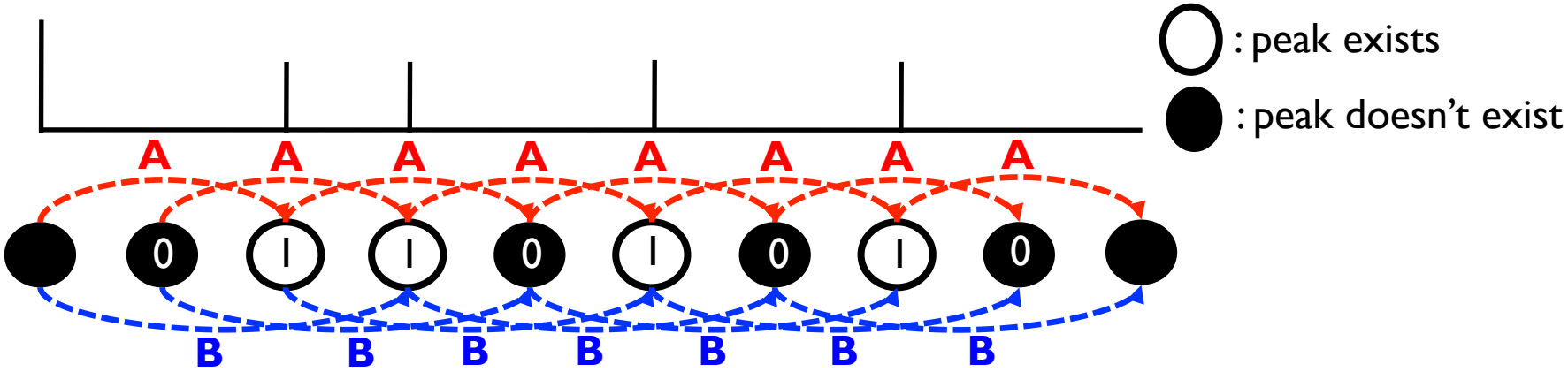
NodeScore(BAAA) = NodeScore(ABAA) = 3

EdgeScore(BAAA) = 4, EdgeScore(ABAA) = 0

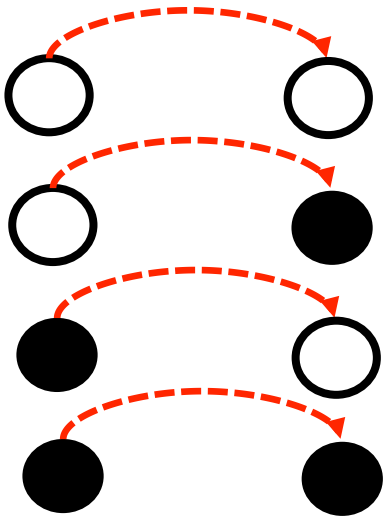
Score(BAAA) = 7 > Score(ABAA) = 3

How to Compute Edge Scores?

Recruit peaks (e.g. y ion peak) corresponding to a nominal mass bin



Prob(Peak) ≈ #Peaks / #Bins = 0.2



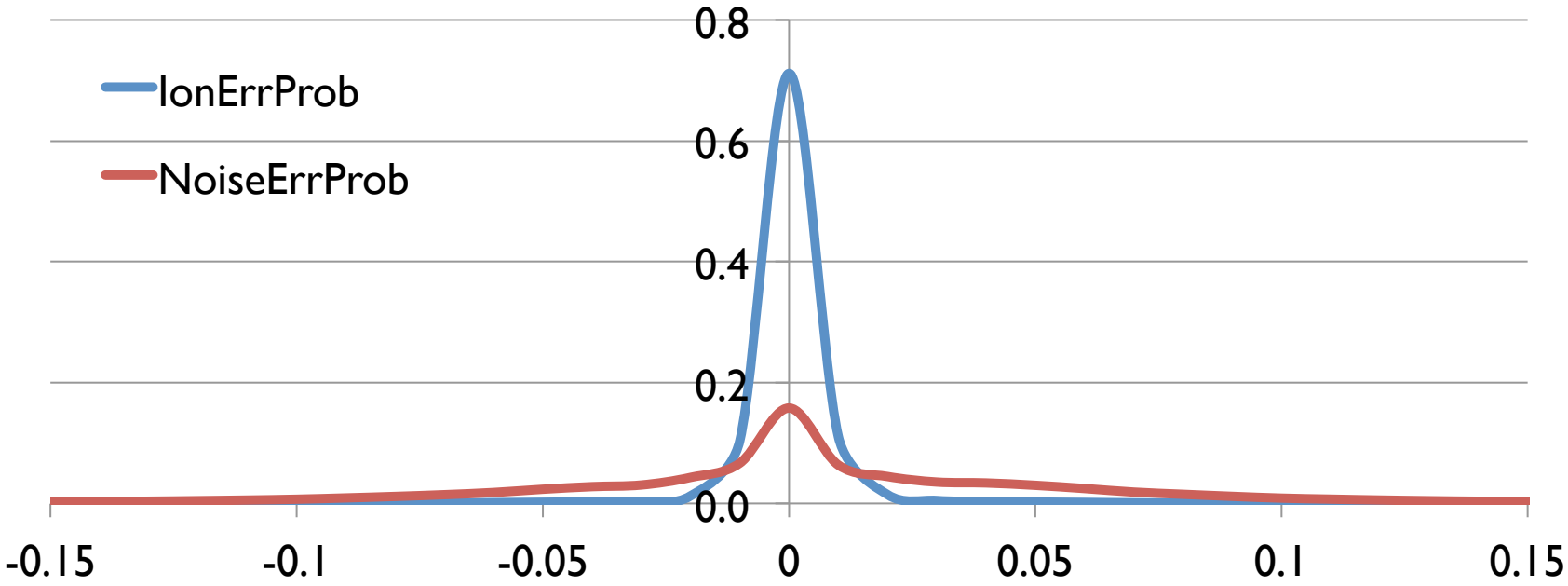
Prob(YY)=0.45, IonExistenceScore(YY)=log(0.45/(0.2*0.2))

Prob(YN)=0.19, IonExistenceScore(YN)=log(0.19/(0.2*0.8))

Prob(NY)=0.19, IonExistenceScore(NY)=log(0.19/(0.2*0.8))

Prob(NN)=0.17, IonExistenceScore(NN)=log(0.17/(0.8*0.8))

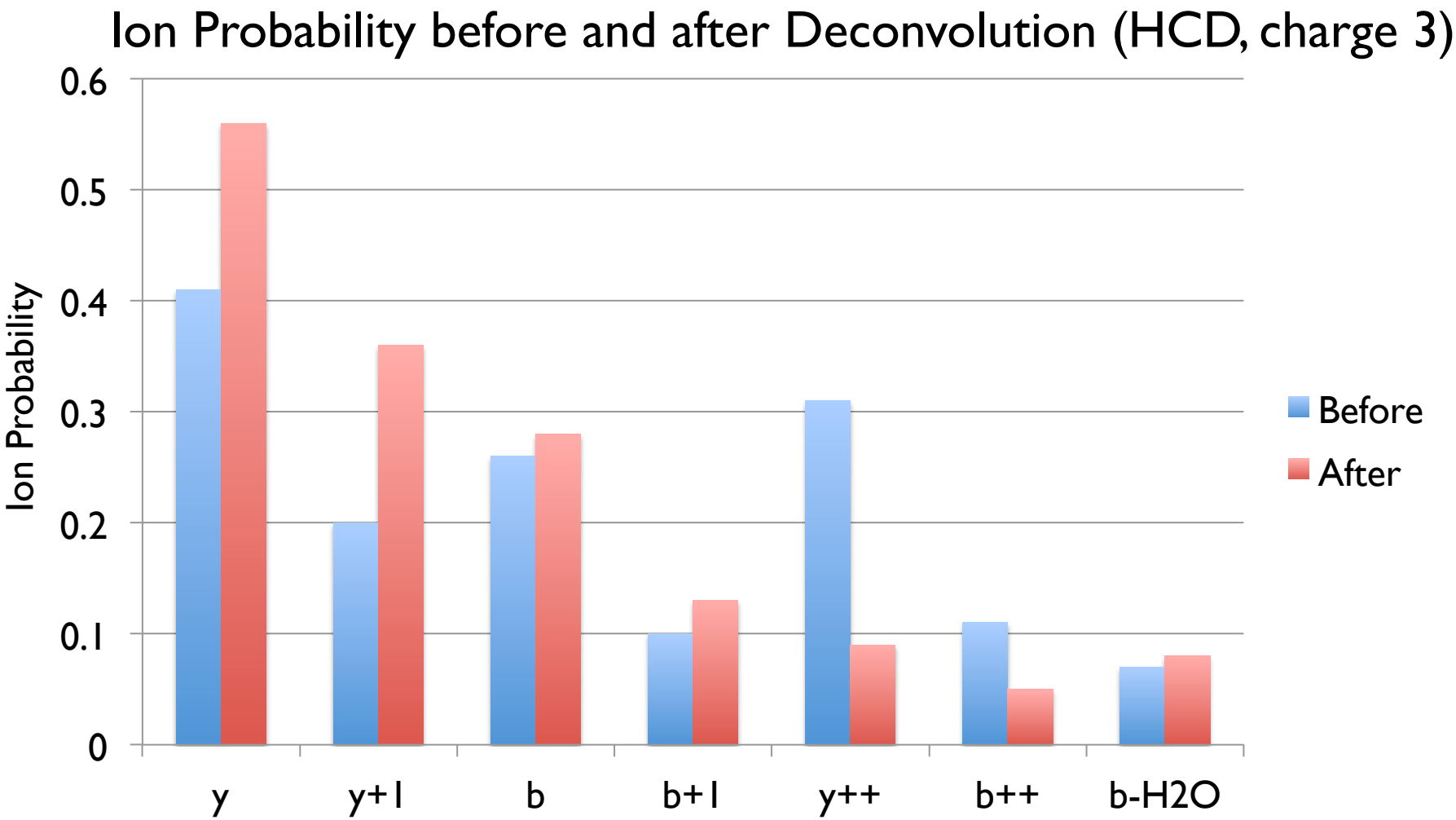
Error Score



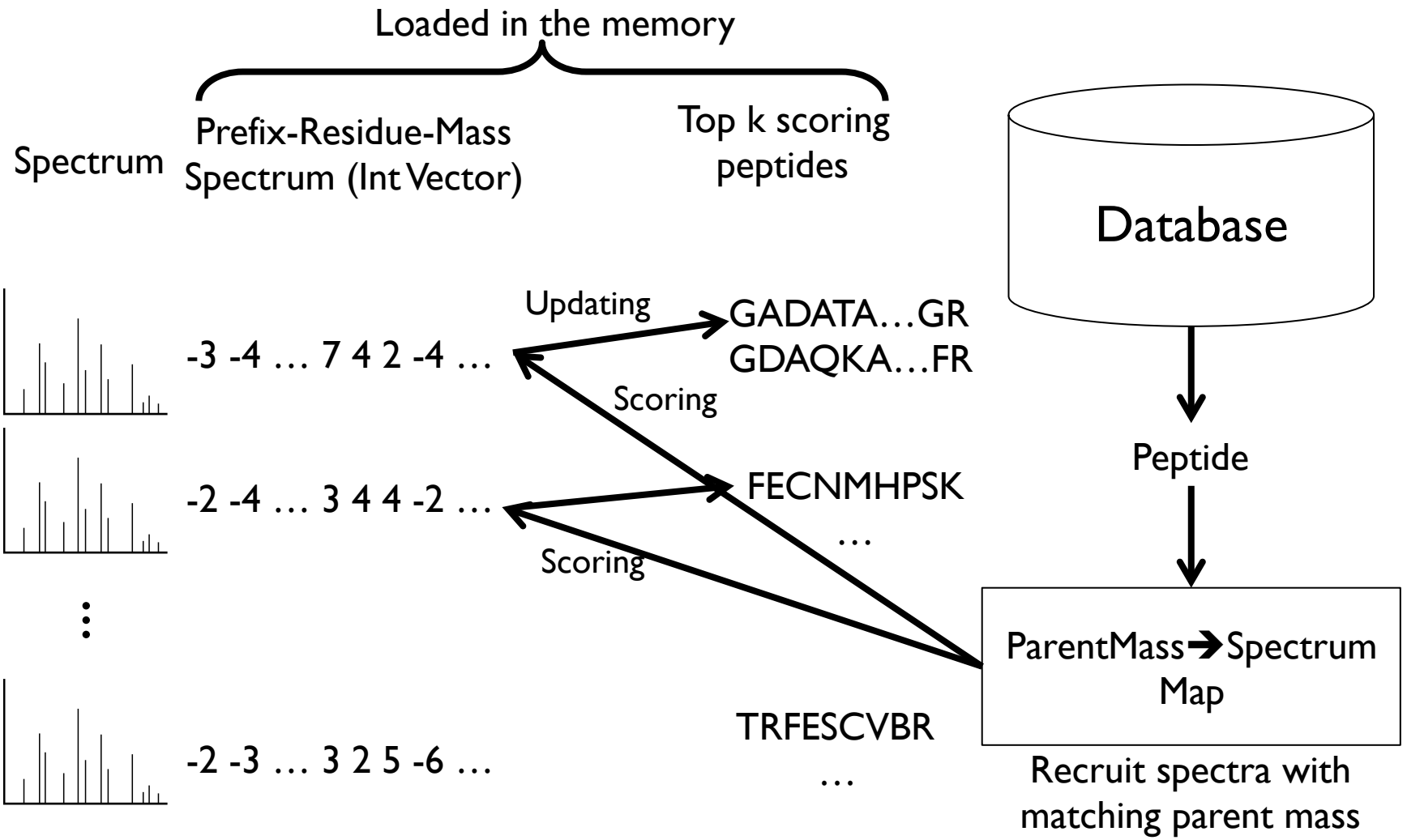
$$\text{ErrorScore}(e) = \log \frac{\text{IonErrProb}(e)}{\text{NoiseErrProb}(e)}$$

EdgeScore = IonExistenceScore + ErrorScore

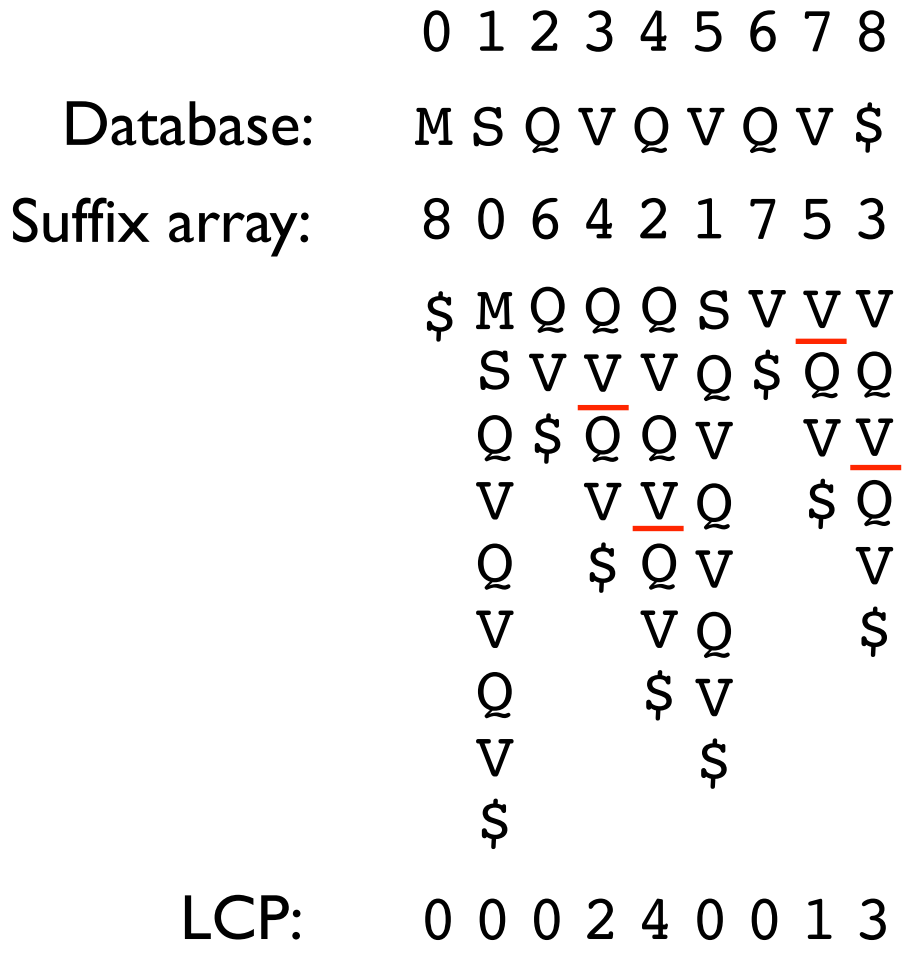
Deconvolution



Database Search



Enumerating Distinct Peptides using Suffix Array



Dinstinct Peptides (length 3-4)

MSQ,MSQV

QVQ

SQV,SQVQ

VQV

VQVQ

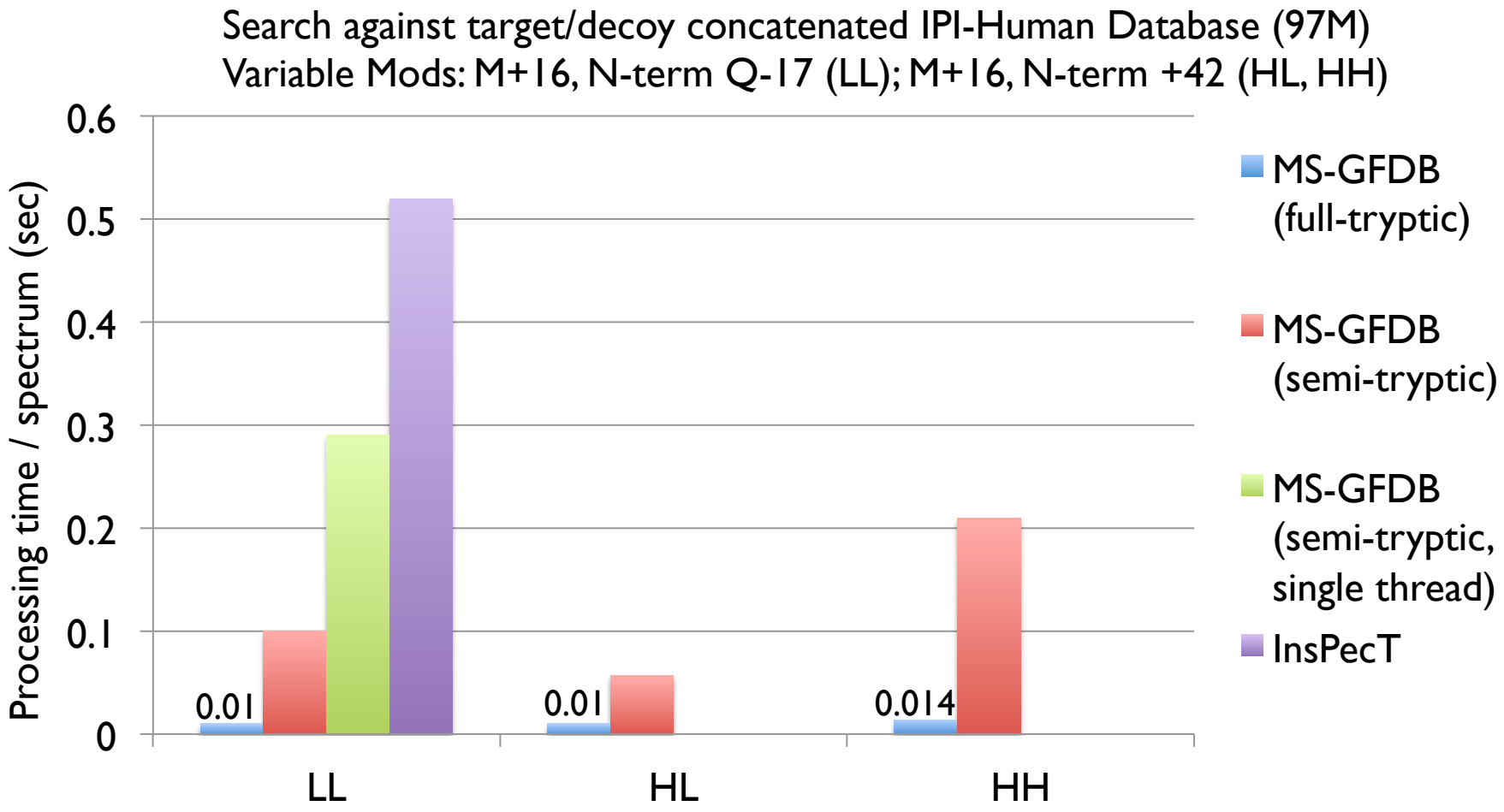
Zhou et al., BMC Bioinfo. 2010

Other Changes

- Multi-threading support: 3X speed-up on a quad-core machine
- Amino acid probabilities are derived from the database (previously, 0.05 for all amino acids)
- Reporting database-level E-value
- Reporting Expected FDR (EFDR) or Target-decoy approach based FDR (FDR)
  - EFDR is computed theoretically without using the decoy database.
  - EFDR is accurate when the parent mass tolerance is larger than 0.5Da
  - Otherwise EFDR is slightly larger than what it should be (conservative estimation)

Results

Running Time



LL: Low-precision MS1, Low-precision MS2
HL: High-precision MS1, Low-precision MS2
HL: High-precision MS1, High-precision MS2

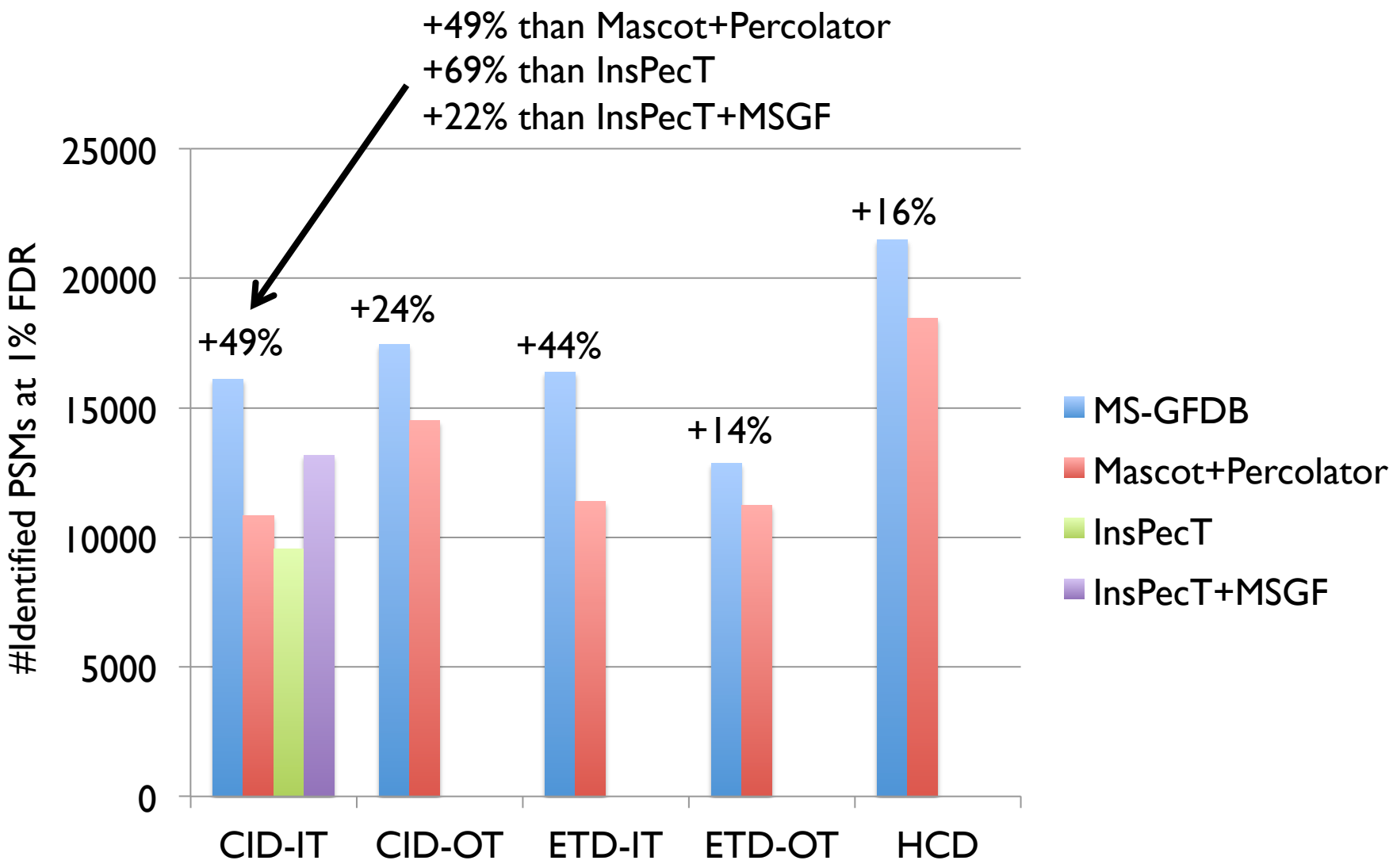
Tested with Core i7 920 (2.67Ghz, quad-core)

Application to Various Fragmentation Methods

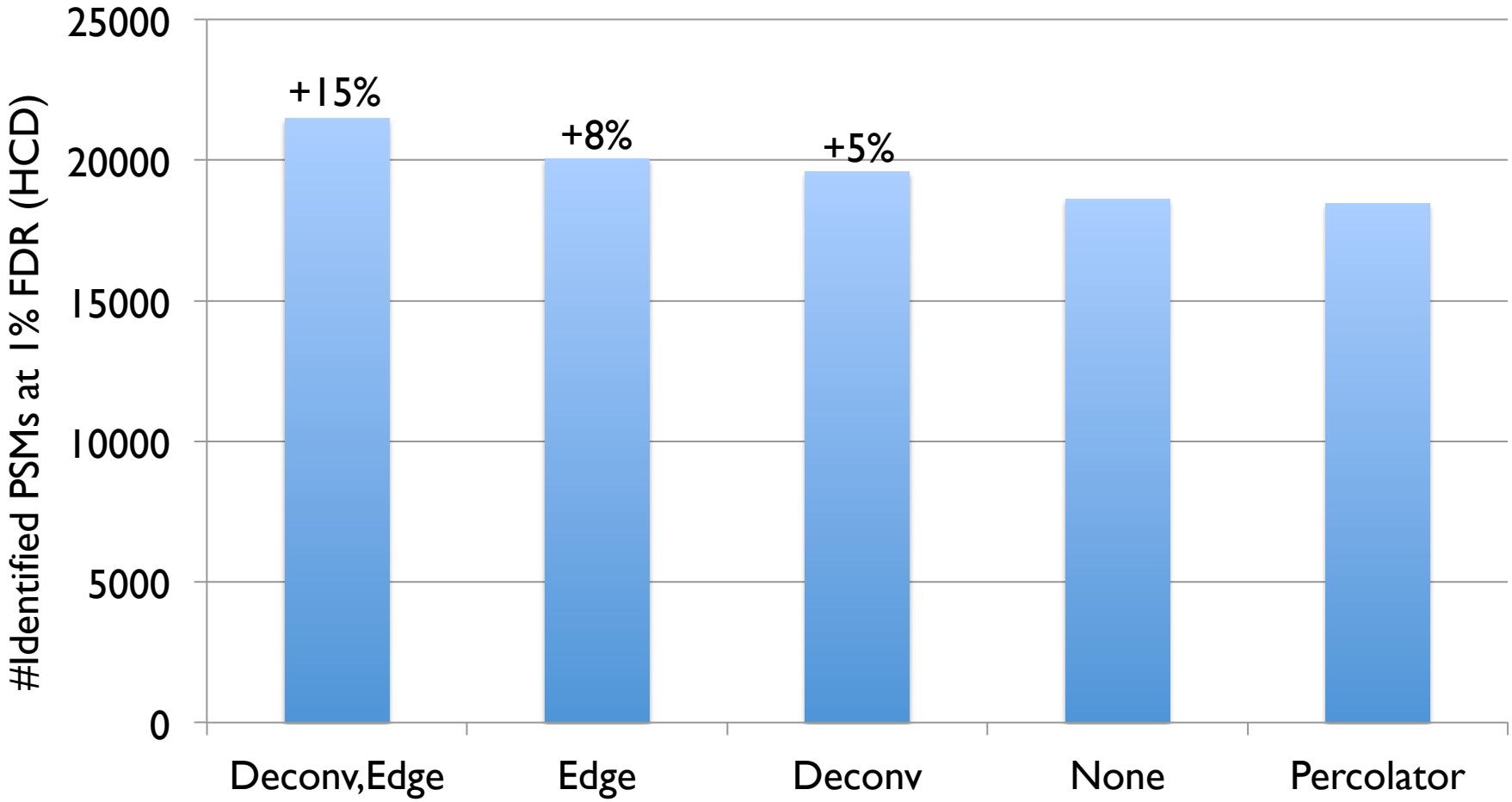
- HEK293 whole cell lysate, digested by Trypsin
  - Analyzed the same sample using 5 different set-ups
  - CID-Iontrap: 38,401 spectra (CID-IT)
  - CID-Orbitrap: 33,586 spectra (CID-OT)
  - ETD-IonTrap: 30,451spectra (ETD-IT)
  - ETD-Orbitrap: 25,734 spectra (ETD-OT)
  - HCD: 37,810 spectra (HCD)

Frese et al., JPR 2011

Search Results



Is Edge Scoring / Deconvolution Helpful?



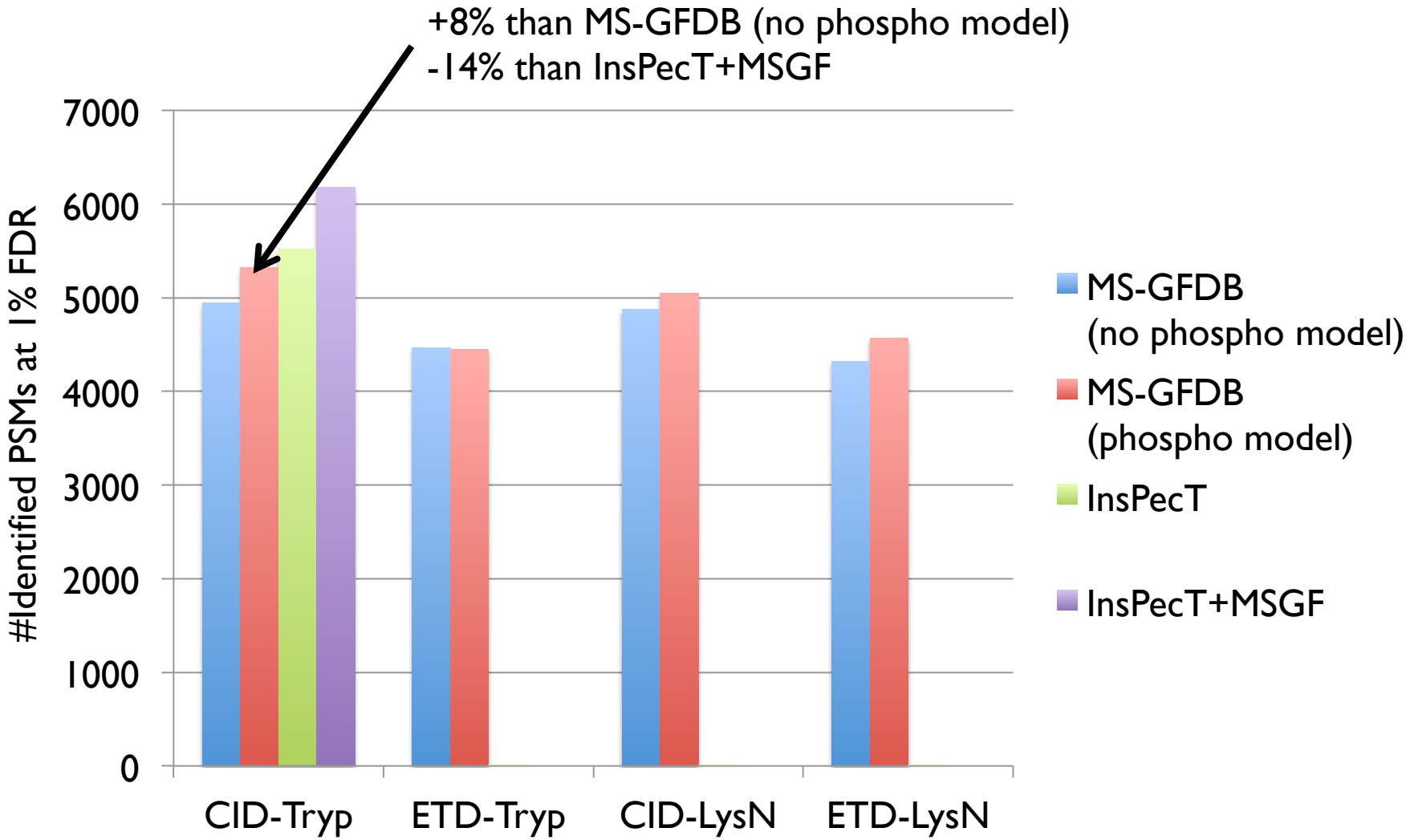
Application to Novel Enzyme

- MS-GFDB greatly outperformed ProteinProspector on a dataset of novel enzyme digests.
- After re-training, MS-GFDB identified 30% more peptides.
- The results will be available after the paper is submitted.

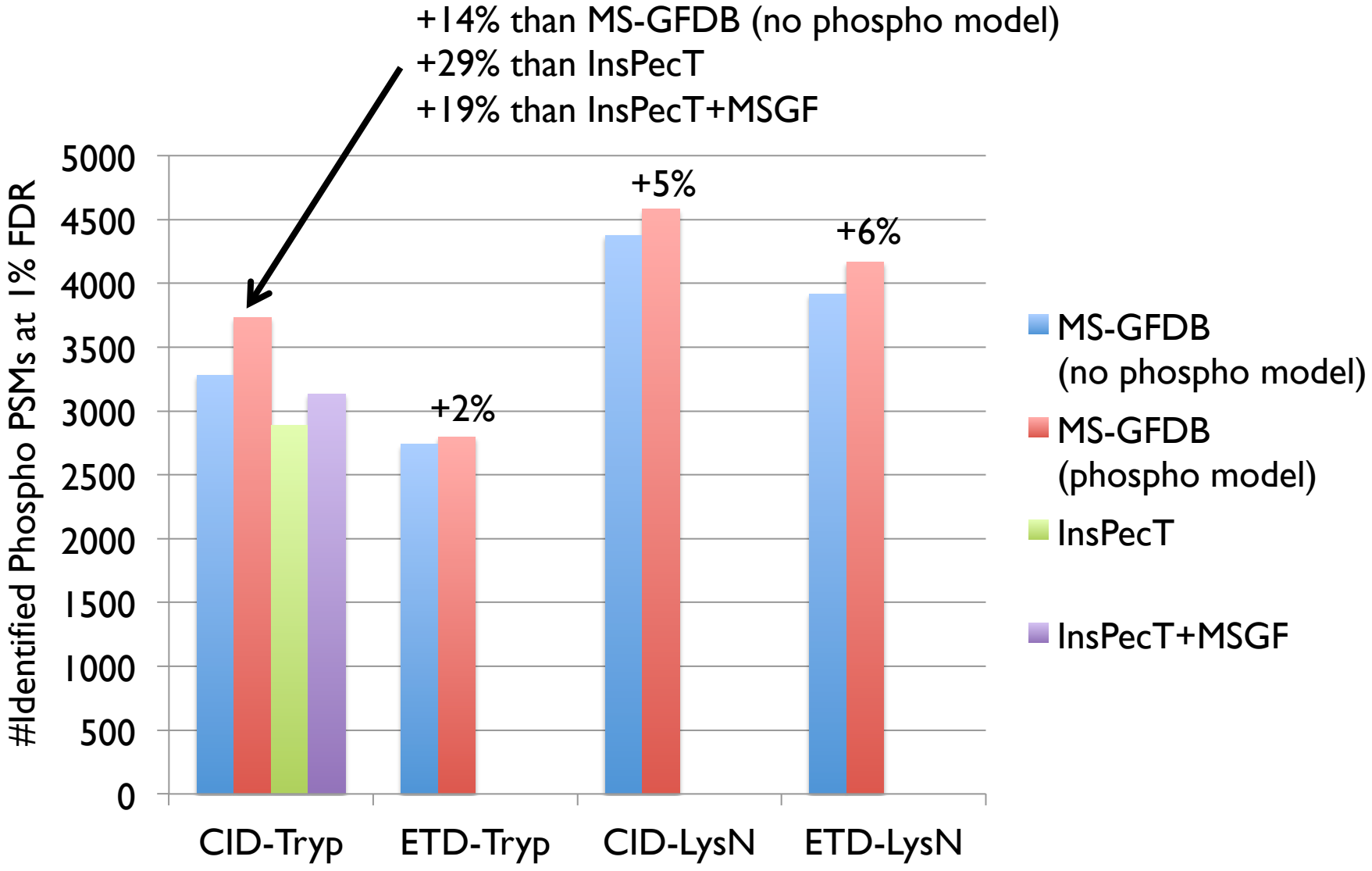
Phosphorylation-enriched Dataset

- HEK293 cell line, phosphopeptides are enriched using SCX
- 68,670 CID/ETD pairs
  - 33,463 from Trypsin digests (CID-Tryp / ETD-Tryp)
  - 35,207 from Lys-N digests (CID-LysN / ETD-LysN)
- Scoring parameters are trained for phosphorylated spectra – no modification of the algorithm

Search Results (#PSMs)



Search Results (#Phospho PSMs)



Phosphorylation-Specific MS/MS Scoring for Rapid and Accurate Phosphoproteome Analysis

Samuel H. Payne,*,† Margaret Yau,‡ Marcus B. Smolka,§ Stephen Tanner,† Huilin Zhou,§,|| and Vineet Bafna‡

Bioinformatics Program, University of California San Diego, Department of Computer Science and Engineering, University of California San Diego, Ludwig Institute for Cancer Research, and Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, California 92093

Received February 16, 2008

The promise of mass spectrometry as a tool for probing signal-transduction is predicated on reliable identification of post-translational modifications. Phosphorylations are key mediators of cellular signaling, yet are hard to detect, partly because of unusual fragmentation patterns of phosphopeptides. In addition to being accurate, MS/MS identification software must be robust and efficient to deal with increasingly large spectral data sets. Here, we present a new scoring function for the Inspect software for phosphorylated peptide tandem mass spectra for ion-trap instruments, without the need for manual validation. The scoring function was modeled by learning fragmentation patterns from 7677 validated phosphopeptide spectra. We compare our algorithm against SEQUEST and X!Tandem on testing and training data sets. At a 1% false positive rate, Inspect identified the greatest total number of phosphorylated spectra, 13% more than SEQUEST and 39% more than X!Tandem. Spectra identified by Inspect tended to score better in several spectral quality measures. Furthermore, Inspect runs much faster than either SEQUEST or X!Tandem, making desktop phosphoproteomics feasible. Finally, we used our new models to reanalyze a corpus of 423 000 LTQ spectra acquired for a phosphoproteome analysis of *Saccharomyces cerevisiae* DNA damage and repair pathways and discovered 43% more phosphopeptides than the previous study.

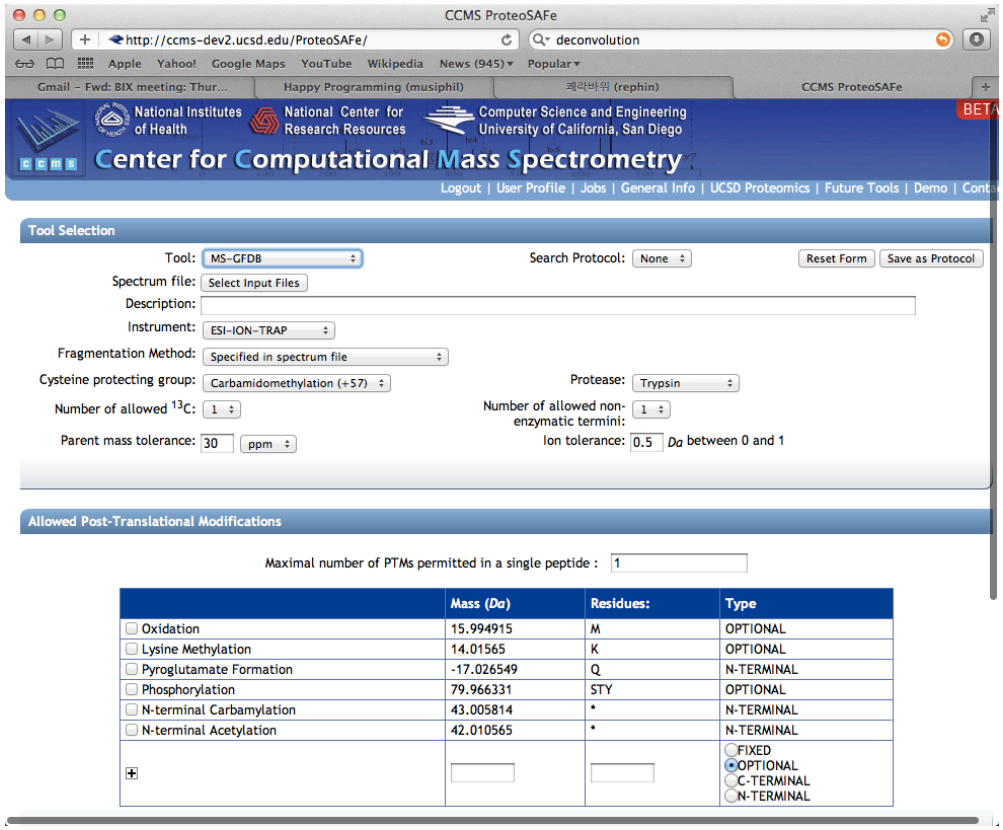
- InsPecT scoring is trained for phosphorylated spectra.
- InsPecT identified 13% more phosphorylated spectra than SEQUEST and 39% more than X!Tandem.
- MS-GFDB identified 29% more phosphorylated spectra than InsPecT.

Is MS-GFDB Easy to Use?

- Command-line interface (MS-GFDB is available at http://proteomics.ucsd.edu/Software/MSGFDB.html)
  - Only 3 required parameters: spectrum file path, database file path, parent mass tolerance
  - Optional parameters: output file path, number of threads to run, whether to use the target-decoy approach for FDR calculation, fragmentation method, instrument type, enzyme, number of allowed isotope errors, number of allowed non-enzymatic termini, modifications, min/max peptide length, min/max charge, number of matches per spectrum to report, whether to use uniform aa probability

Is MS-GFDB Easy to Use?

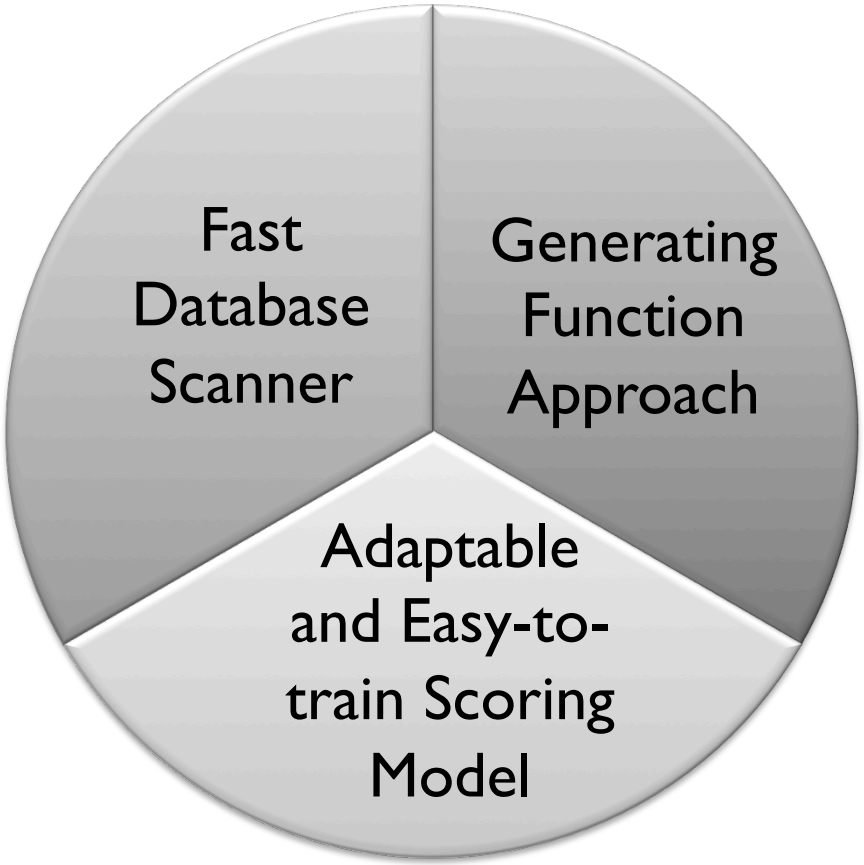
- Web-based Interface: ProteoSAFe



Then, Why Isn't MS-GFDB Popular?

- At the time of publication (MCP 2010), it was inconvenient to use.
  - Preprocessing database took long
  - No modification support
  - Slow
- It is getting popular
  - Incorporated into the pipeline of Pacific Northwest National Laboratory

Summary – What is MS-GFDB?



Summary – How Good is MS-GFDB?

Which is better?

	Any DB Search Tool Or Post-processing Tool	MS-GFDB
Identifying more peptides at 1% FDR		✓
Applicability to various spectral types		✓
Easiness to use	✓	
Search speed	✓	✓

To Do – One-click Search

- Minimizing Search Parameters
  - Don't let users worry about search parameters
  - Automatically decides the followings:
    - Parent mass tolerance
    - # allowed isotope errors
    - # allowed non-enzymatic termini
    - Modifications

To Do – Scalability

- For a database containing n amino acids, currently MS-GFDB requires:
  - 4n+ α bytes of memory for the suffix array construction (α = 200-300MB)
  - n+ α bytes of memory for database searching
- Reduce the requirement to n+ α and α

To Do – User Interface

- Easy-to-use UI for Scoring Parameter Training
  - Run MS-GFDB with training mode and output a parameter file
  - Register the parameter file for later uses

To Do – Naming

- Change the name
- Any Idea?
- MS-GF+ (?)

Acknowledgments

- Pavel Pevzner
- Nuno Bandeira, Jeremy Carver, To-ju Huang
- Kyowon Jeong, Natalie Castellana, Xiaowen Liu and all the lab members
- Julio Ng at Microsoft
- Taejoon Kwon at UT Austin
- Matthew Monroe at PNNL
- Magnus Palmblad at Leiden University, Netherlands
- Jesse Meyer at Komives lab, UCSD
- Christian Frese and Albert Heck at Utricht University, Netherlands

National Institutes of Health
The Nation's Medical Research Agency

NIH P-41-RR024851