Center for Computational Mass Spectrometry

UCSD

MixGF: spectral probability for mixture spectra of more than one peptides

Jian Wang[1], Philip E Bourne[2] and Nuno Bandeira[3,4,5,]

1. Bioinformatics Program, UCSD, La Jolla, CA
2. 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:

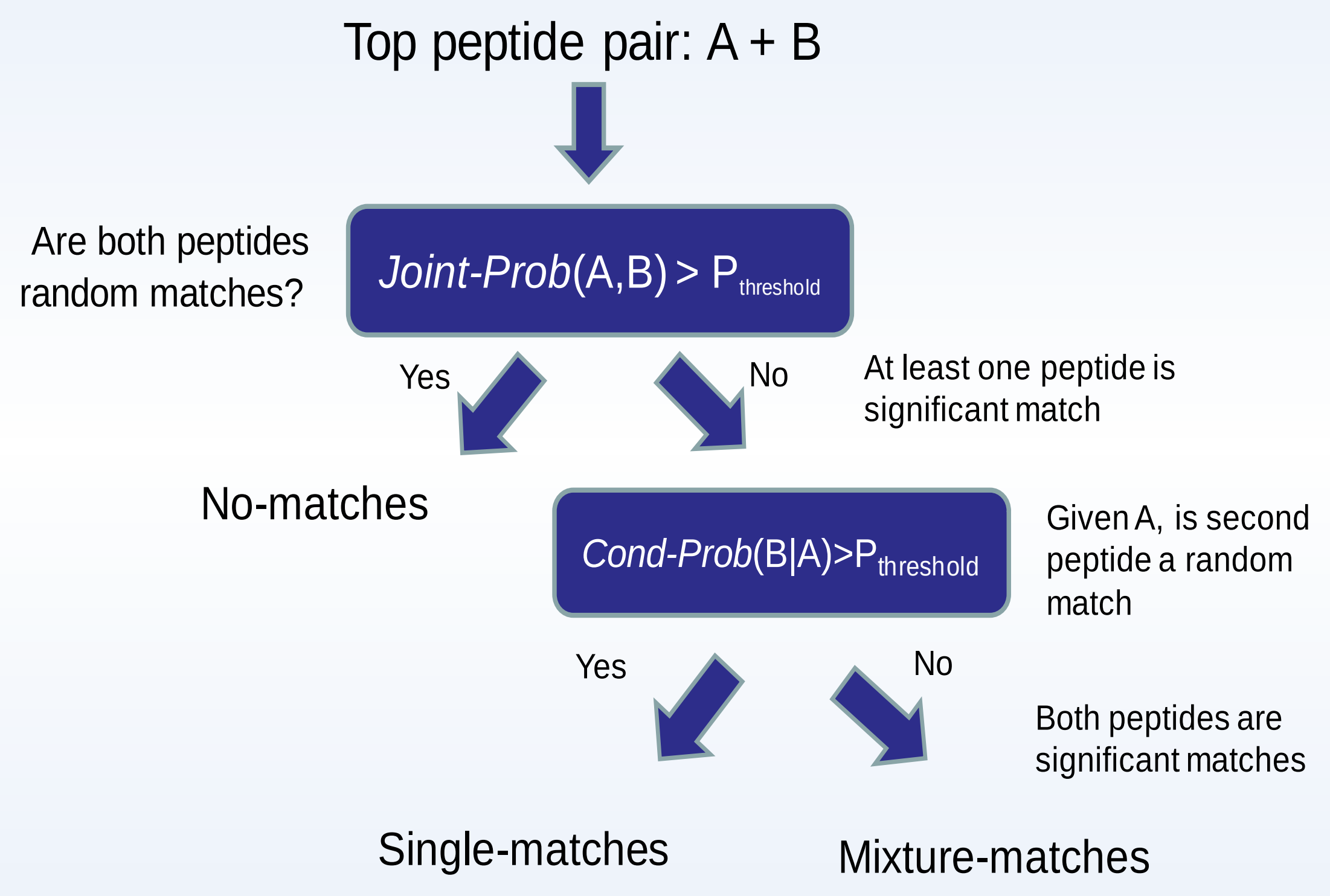
Recent advances in data acquisition protocols such as MS^E , Q-Exactive, SWATHMS where multiple peptides are fragmented simultaneously in one MS/MS *mixture* spectrum have the potential to greatly increase the throughput of peptide identification in proteomics. However the successful application of these protocols partly depends on computational methods that can sequence more than one peptide per MS/MS spectrum. In previous work we showed that current tools for identifying mixture spectra suffers from relative low sensitivity because of their limited ability to separate true matches from false positives. Here we describe how to rigorously compute the statistical significance of peptide identifications for mixture spectra and show that this approach substantially improves the sensitivity of state-of-the-art database search tools for identifying mixture spectra.

Overview:

We model a mixture spectrum as a linear combination of two single-peptide spectra and want to calculate the statistical significance for a given pair of peptides (A,B) matched to a spectrum (M). We formulate this problem into two questions:

1)*Joint-probability*: what is the probability that a random pair of peptides (out of all possible *peptide pairs*) match M with score greater than score(M,A,B)?

2) *Conditional-probability*: what is the probability that a random peptide (out of all possible peptides) that pair with the first-matched peptide A will have a score greater than score(M, A, B)?



Top peptide pair: A + B

Are both peptides random matches?

Joint-Prob(A,B) > $P_{threshold}$

Yes

No

At least one peptide is significant match

No-matches

Cond-Prob(B|A)>$P_{threshold}$

Given A, is second peptide a random match

Yes

No

Both peptides are significant matches

Single-matches

Mixture-matches

Challenges:

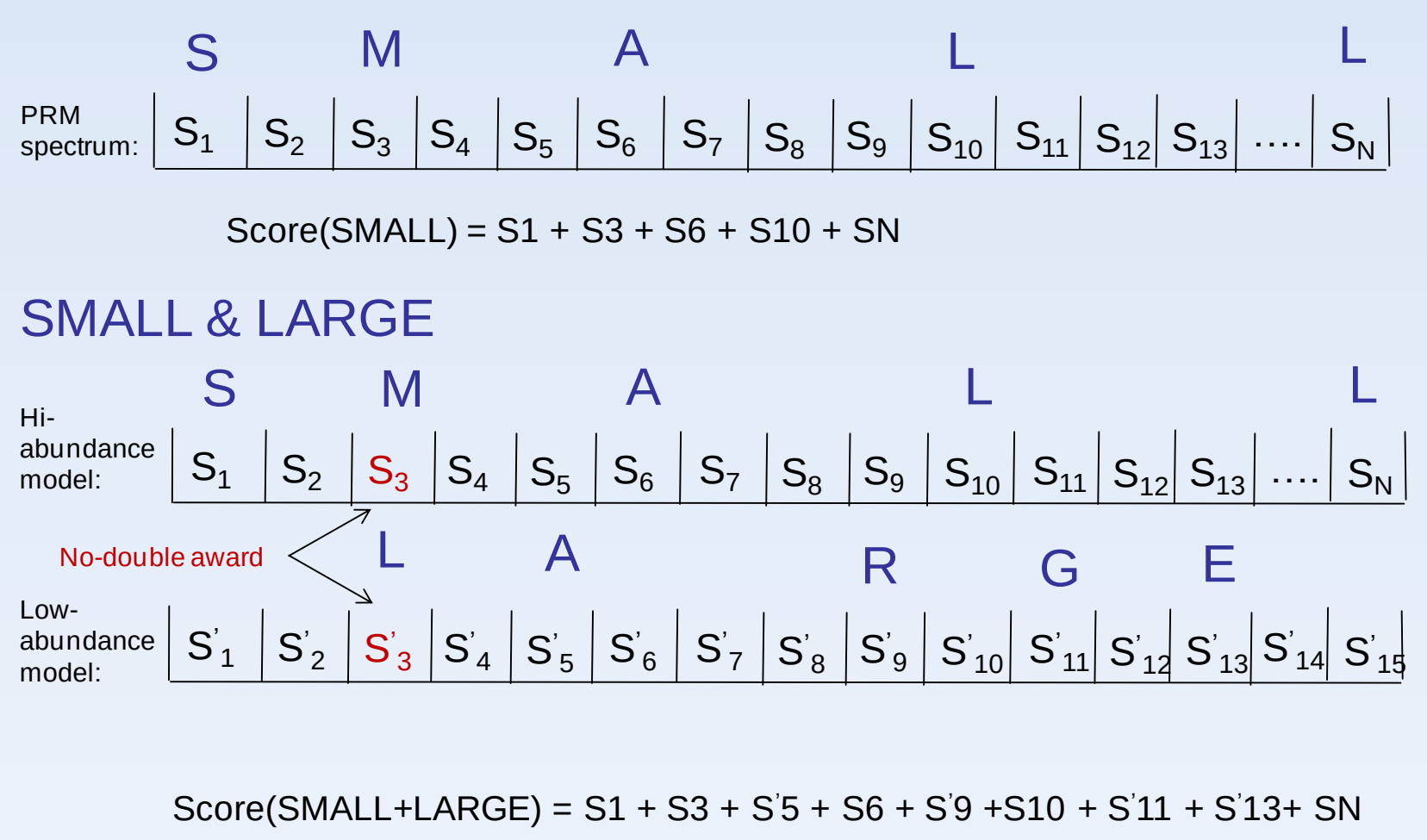
The statistical questions are straight-forward to formulate, but to compute *joint-* and *conditional-probability* we need to generate the score distribution of all peptides and peptide pairs which is computationally expensive. The challenge is to compute the probability efficiently without explicitly consider scores of all peptide and peptide pairs.

Reference:

[1] J.Wang, PE. Bourne, N Bandeira MCP(10) 2011
[2] J. Wang, J. Perez-Santiago, J.E. Katz, P. Mallick, and N. Bandeira. MCP 9(7):1476–85, 2010.
[3] Kim, S., Gupta, N., and Pevzner, P. A. (2008), J. Proteome Res. 7, 3354 –3363

Prefix-residue mass (PRM) spectrum for mixture spectrum:

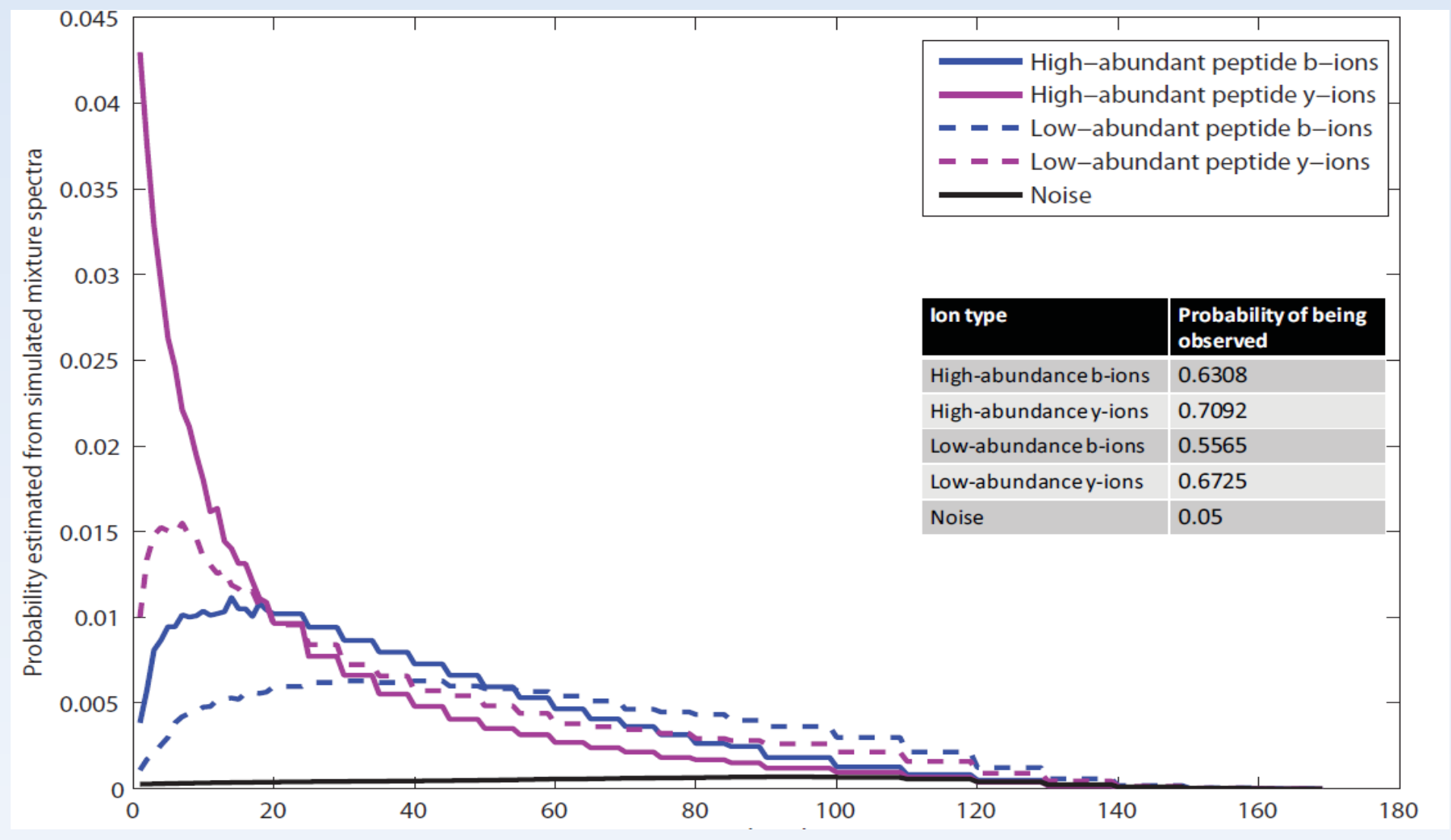
A prefix-residue spectrum is a scored version of the MS/MS spectrum that has a score at each mass position from 0 to precursor mass M. The score at position i represents the log-likelihood that a peptide with prefix mass i generate the observed MS/MS spectrum.


S
M
A
L
L
PRM spectrum:
S_1 S_2 S_3 S_4 S_5 S_6 S_7 S_8 S_9 S_{10} S_{11} S_{12} S_{13} …. S_N
SMALL & LARGE
Hi-abundance model:
No-double award
L
A
R
G
E
Low-abundance model:
S'_1 S'_2 S'_3 S'_4 S'_5 S'_6 S'_7 S'_8 S'_9 S'_{10} S'_{11} S'_{12} S'_{13} S'_{14} S'_{15}

Score(SMALL) = S1 + S3 + S6 + S10 + SN

Score(SMALL+LARGE) = S1 + S3 + S'5 + S6 + S'9 +S10 + S'11 + S'13+ SN

Need different scoring models to model high and low abundance peptide in mixture spectra because they have very different fragmentation patterns



Ion type	Probability of being observed
High-abundance b-ions	0.6308
High-abundance y-ions	0.7092
Low-abundance b-ions	0.5565
Low-abundance y-ions	0.6725
Noise	0.05

Dynamic programming to compute joint-probability and conditional probability:

Let *Joint*(*m*1,*m*2, *S, T*) be the probability that a pair of peptide with parent mass m1 and m2 when match to S with score higher than T. Also define S^H represents the scoring model for high-abundance peptide and S^L represents the scoring model for low-abundance peptide. Then we can define the following recurrence relationship for Joint-probability:

$$
Joint(m1,m2,S,T) = \left\{ \begin{array}{l}
if\ m1 < m2: \\
\sum_{all\ amino\ acid\ a} Joint(m1, m2-|a|, T - S^L(m2)) \times prob(a) \\
if\ m1 > m2: \\
\sum_{all\ amino\ acid\ a} Joint(m1-|a|, m2, T - S^H(m1)) \times prob(a) \\
if\ m1 == m2: \\
\sum_{a1}\sum_{a2} Joint(m1-|a1|, m2-|a2|, s - max\left\{ \begin{array}{l} S^H(m1) \\ S^L(m2) \end{array} \right\}) \\
\times prob(a1) \times prob(a2)
\end{array} \right.
$$

Let Cond(*m2*, *S*, *T*| *A*) be the conditional probability that a peptide with parent mass m2 pair with A when match to S with score higher than T. Since we are conditioned on the first peptide being valid match, to avoid double counting, we give a score of zero at mass position corresponds to the prefix masses of A:

$$For: P = p_1, p_2 ... p_n \quad S^L(p_i) = 0$$

then we can compute the conditional probabitly with the following recurrence:

$$Conditional(m2, S, T|P) = \sum_{all\ amino\ acid\ a} Conditional(m2 - |a|, T - S^L(m2)|P) \times prob(a)$$

Approximating Joint-probability by product of conditional-probability:

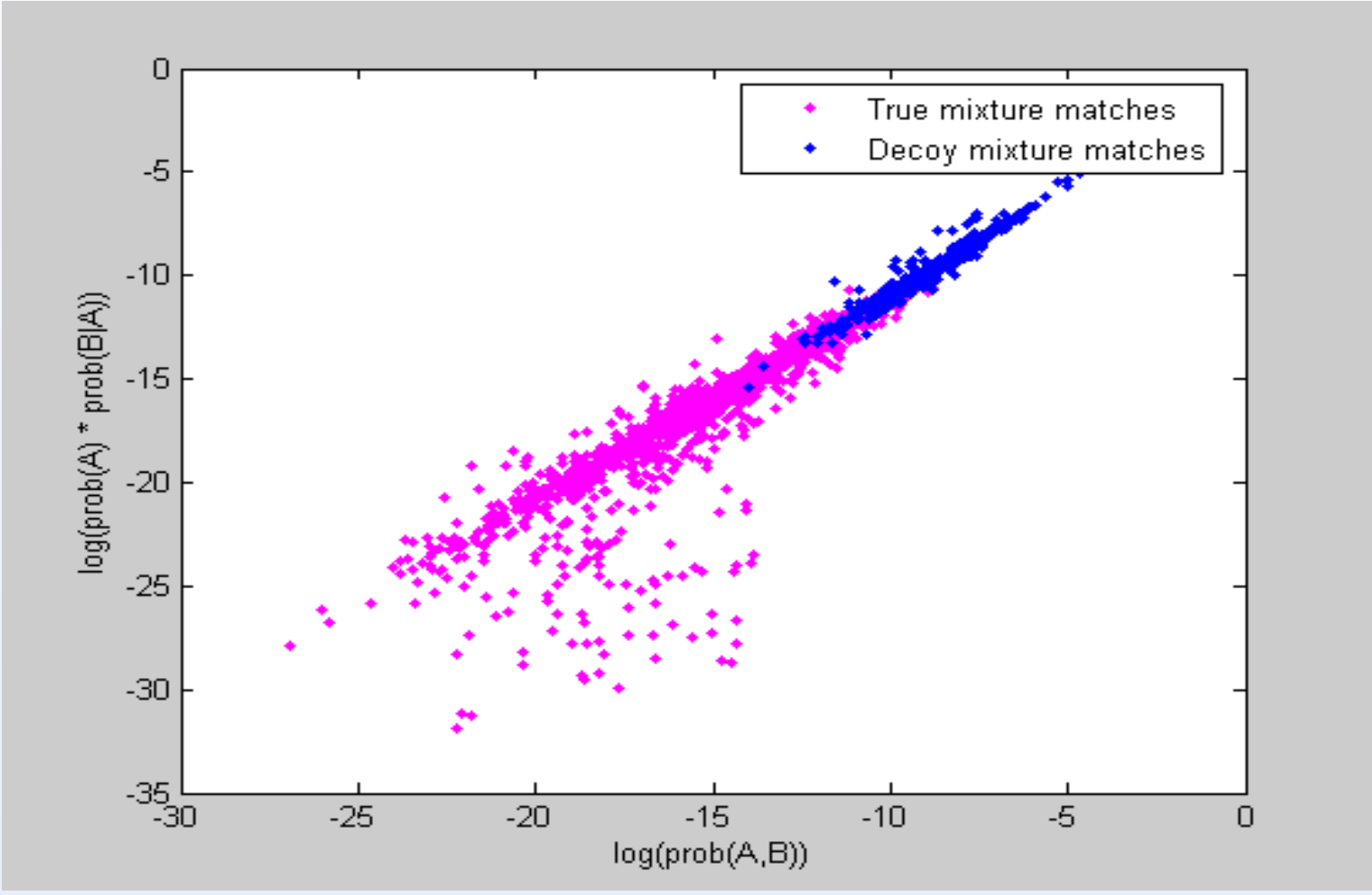
Joint-probability can be computed rigorously, but still scale exponentially to the number of peptides. We want to approximate it with conditional probability that we know how to compute efficiently

Definition of conditional probability:

$$\Pr(B \mid A) = \frac{\Pr(A \wedge B)}{\Pr(A)}$$

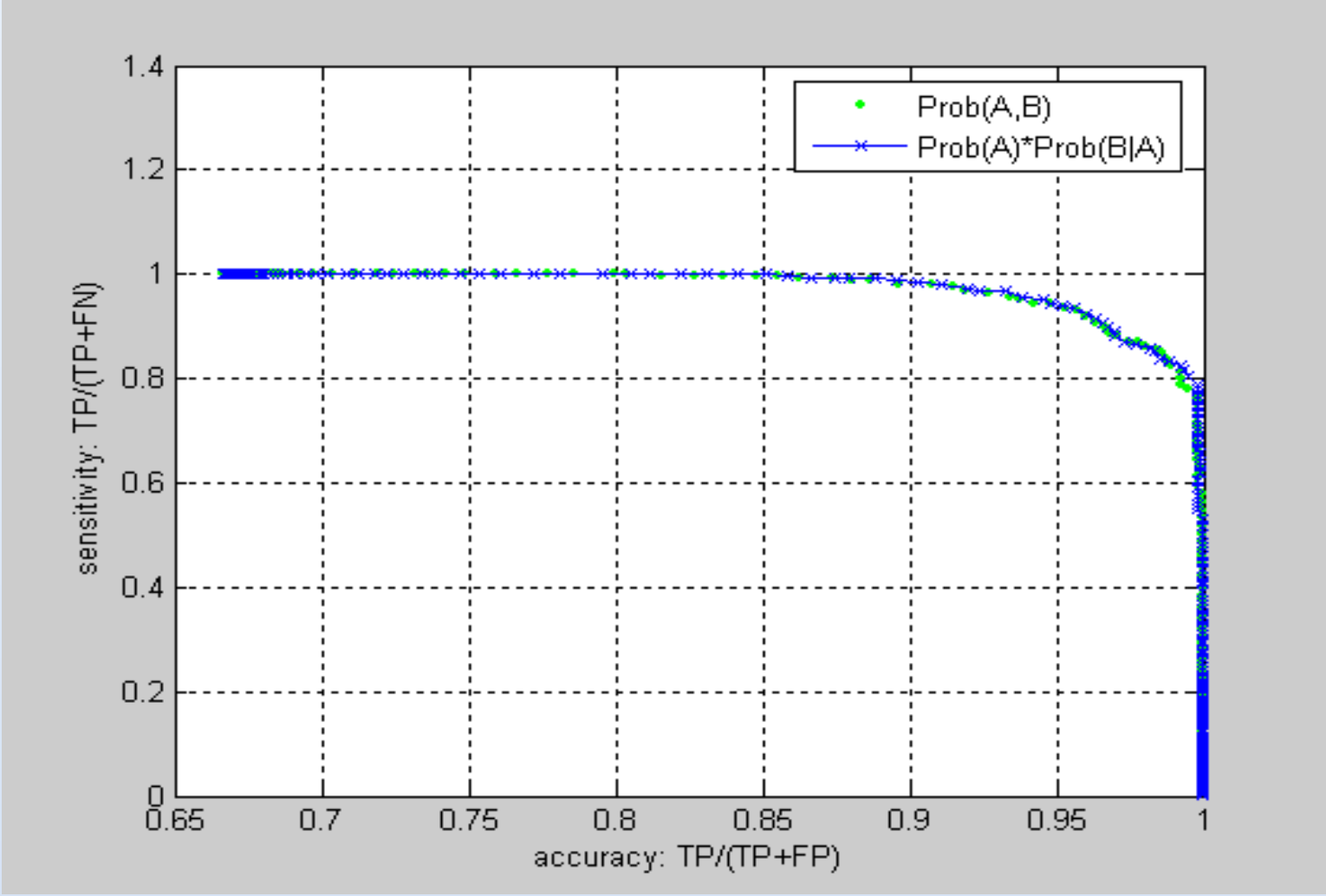
$$\Pr(A \wedge B) = \Pr(A) \times \Pr(B \mid A)$$

$$\mathbf{Joint}(A, B) \approx \Pr(A) \times Cond(B \mid A)$$



Target-decoy approach for mixture spectrum:

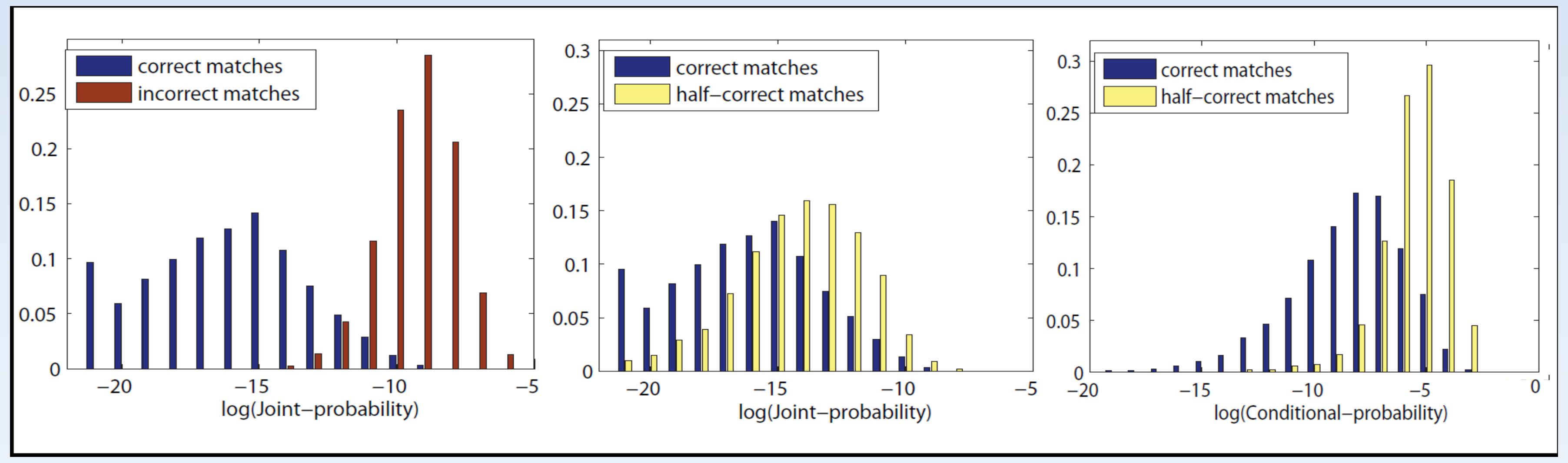
For mixture-spectrumIDs
Matches are in four categories: TT, TD, DT, DD
A target hit can be correct (C) or incorrect (I)
A decoy hit is by definition incorrect
TT = CC + CI + IC + II
TD = CI + II
DT = IC + II
DD = II
Target-Decoy assumption: CI, IC and II are equal
Want to compute
FDR(mixture) = (IC+IC+II)/CC
= (TD + DT – DD) / TT
FDR(IDs) = (½CI +½IC + II)/CC
= ½(TD + DT)/ TT

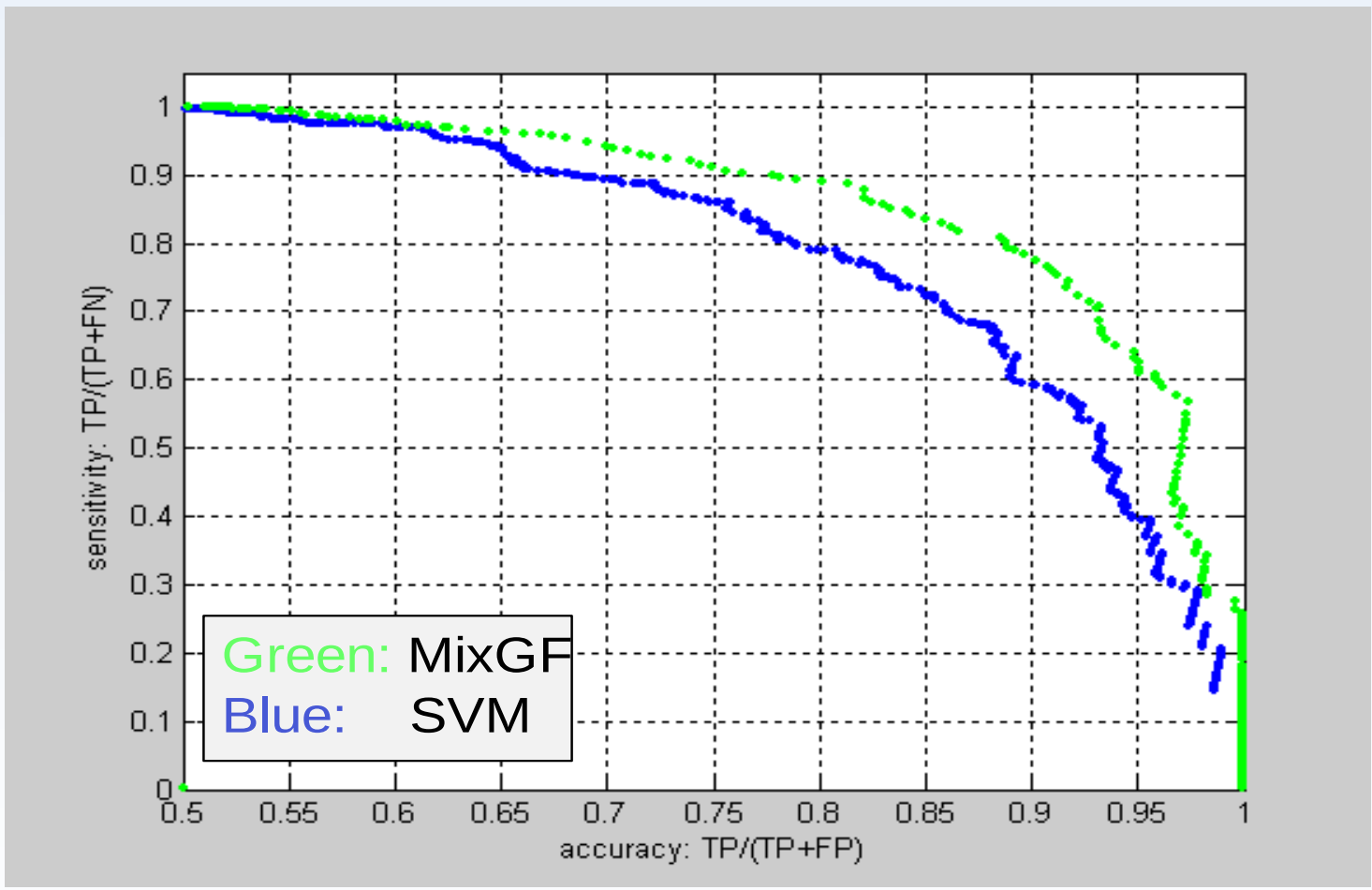


Results:

Separating true matches from false matches in simulated dataset:

--Half-correct matches are matches where one peptide is correct and another peptide is incorrect



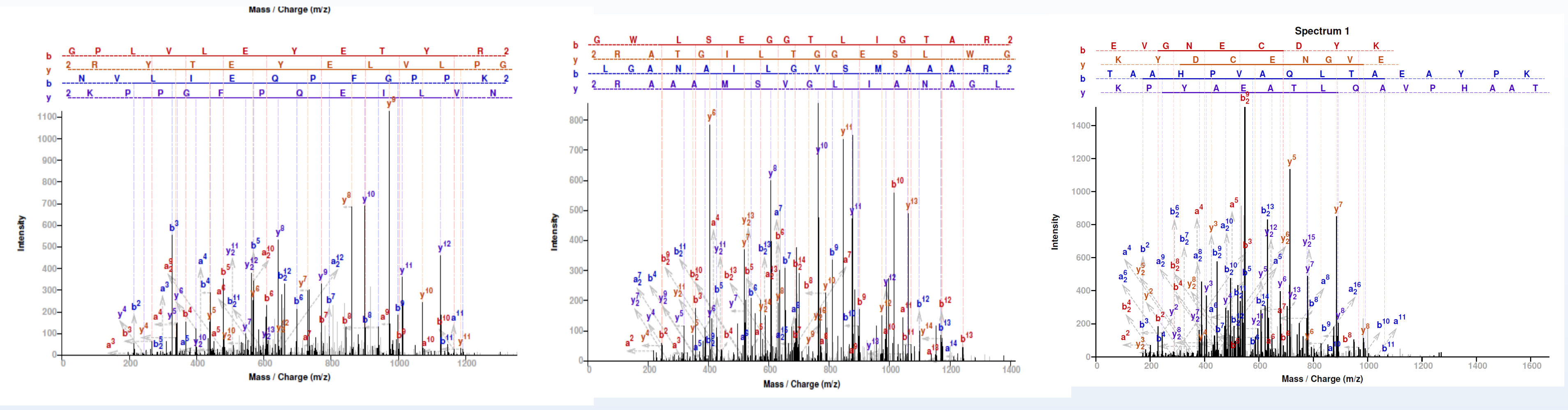


Benchmarking on Yeast whole-cell lysate:

–Public available in Tranche/Proteome Commons (from Univ of Vanderbilt)
–Analyzed on LTQ Orbitrap XL mass spectrometer
–Total of 76177 spectra

Method	1% FDR	2% FDR	3% FDR	4% FDR	5% FDR
svm	748	1214	1620	1905	2124
Joint-prob, cond-prob	1320	1580	1972	2268	2676
product-prob, cond-prob	1011	1646	2038	2356	2688
Single-prob, cond-prob	1310	1664	2091	2452	2760

Examples of mixture spectra:



Conclusion:

- Statistical significance of a peptide-peptide-spectrum matches (PPSM) can be formulated as two questions: 1) Joint-probability and 2) Conditional-probability.
- These two probability can be computed analytically and efficiently using a dynamic programming approach.
- Joint-probability is a good metric to separate true mixture-spectrum matches from false matches where both peptides are incorrect and conditional-probability is a good metric to separate true matches from false matches where one peptide is correct and the other peptide is incorrect.
- Joint-probability can be efficiently approximated by a product of conditional-probability, enabling MixGF's applicability to mixture-spectra with more than two peptides.
- MixGF approach increase the sensitivity of current database search methods at identifying mixture spectra from more than one peptides