Center for
Computational
Mass
Spectrometry

UCSD

A Database Search Method for Identifying Mixture Tandem Mass Spectra

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

1. Bioinformatics Program, University of California, San Diego, La Jolla, CA, USA 2. Skaggs School of Pharmacy and Pharmaceutical Science, UCSD, La Jolla, CA, USA
3. San Diego Supercomputer Center, UCSD, La Jolla, CA, USA 4. Center for Computational Mass Spectrometry, UCSD, La Jolla, CA, USA 5. Department of Computer Science and Engineering, UCSD, La Jolla, CA, USA

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

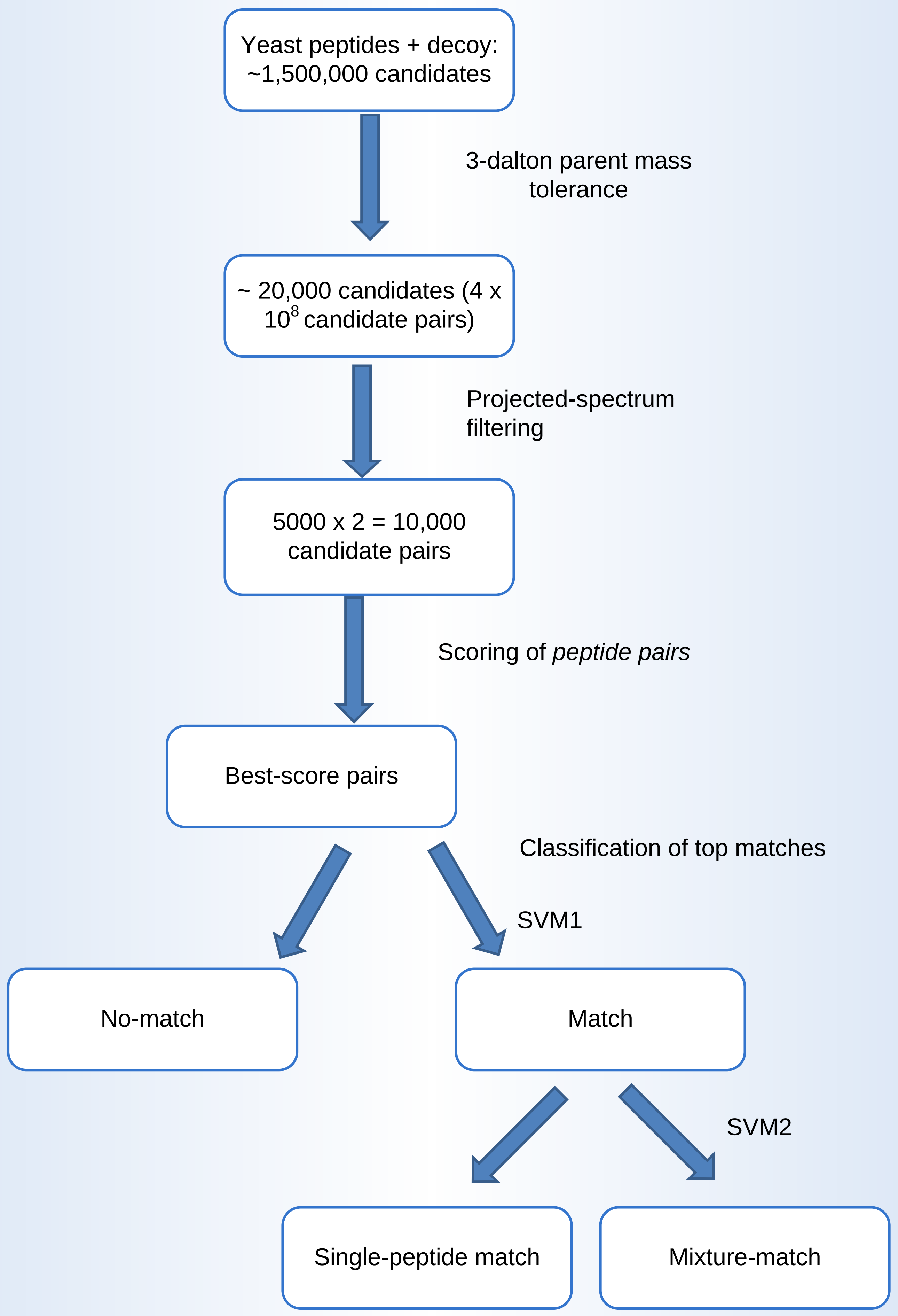
Overview:

A database search algorithm for supporting identification of tandem mass spectra from more than one peptide – *mixture spectra.*

Introduction:

The success of high-throughput proteomics hinges on the ability of computational methods to identify peptides from tandem mass spectra. However, a common limitation of most peptide identification approaches assumes each MS/MS spectrum is generated from a single peptide. We propose a new database search tool and demonstrate that peptides can be reliably identified from mixture spectra while considering only a fraction of possible peptide pairs.

Method Overview:



Yeast peptides + decoy: ~1,500,000 candidates

3-dalton parent mass tolerance

~ 20,000 candidates (4 x 10^8 candidate pairs)

Projected-spectrum filtering

5000 x 2 = 10,000 candidate pairs

Scoring of *peptide pairs*

Best-score pairs

Classification of top matches

SVM1

No-match

Match

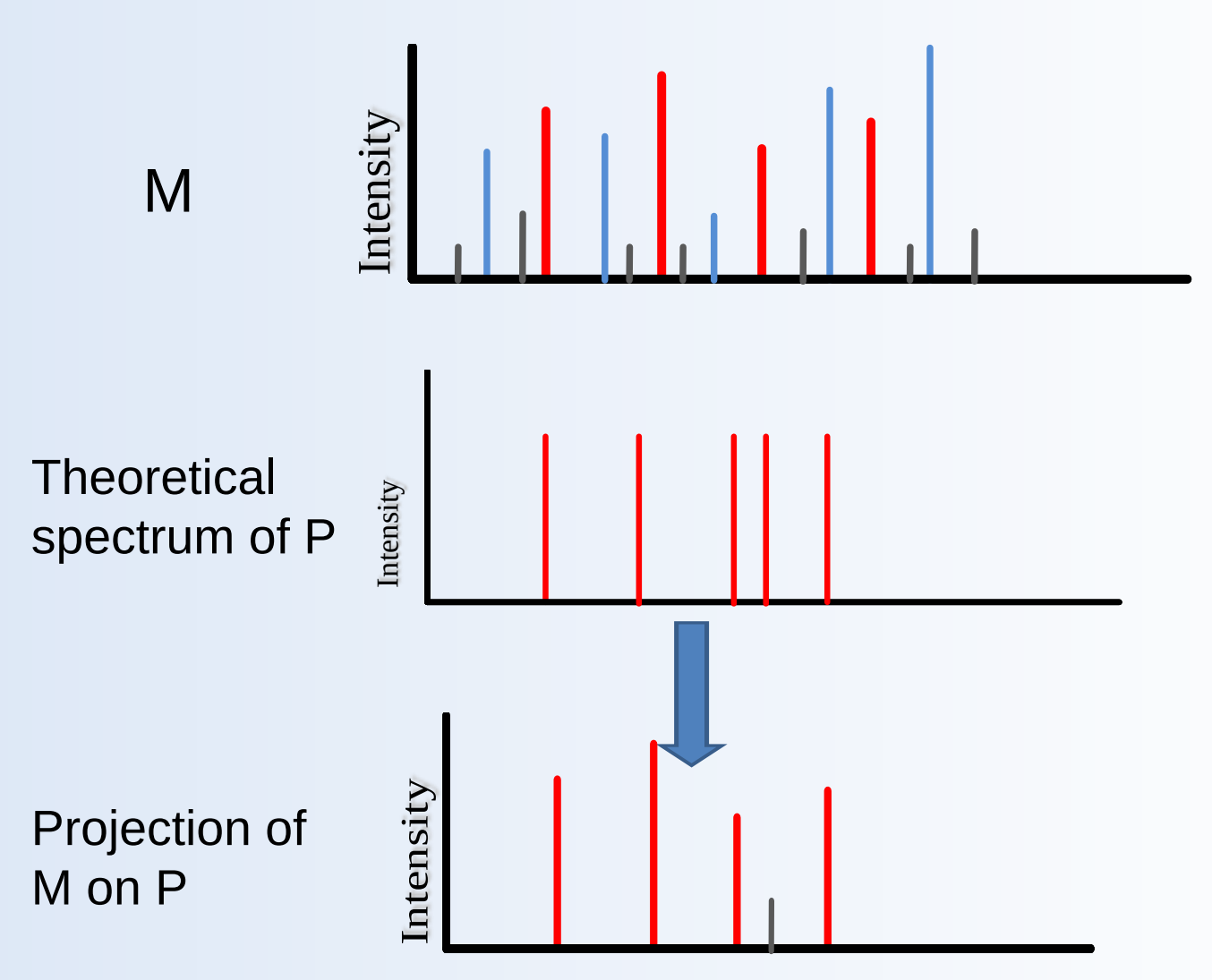
SVM2

Single-peptide match

Mixture-match

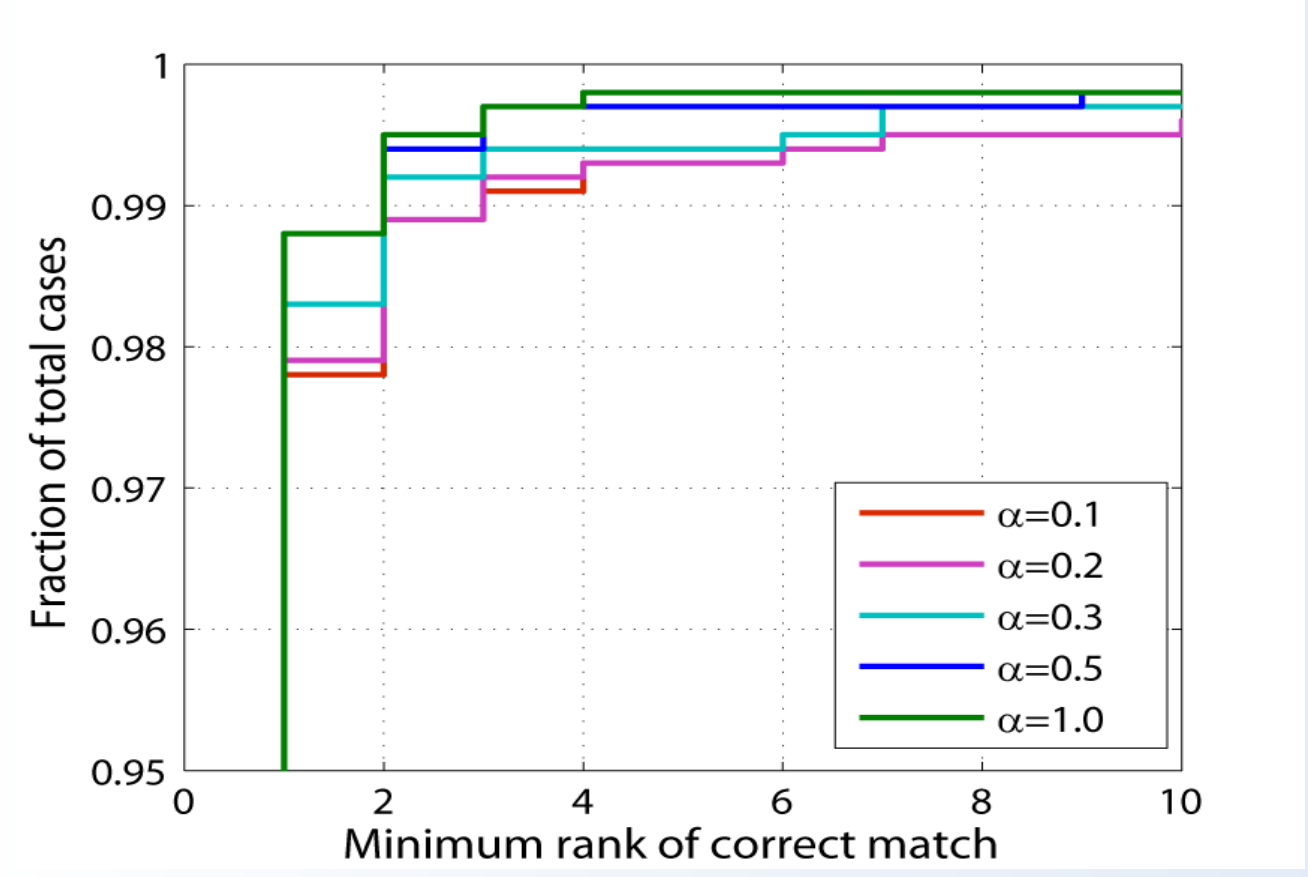
Projected-spectrum Filtering:

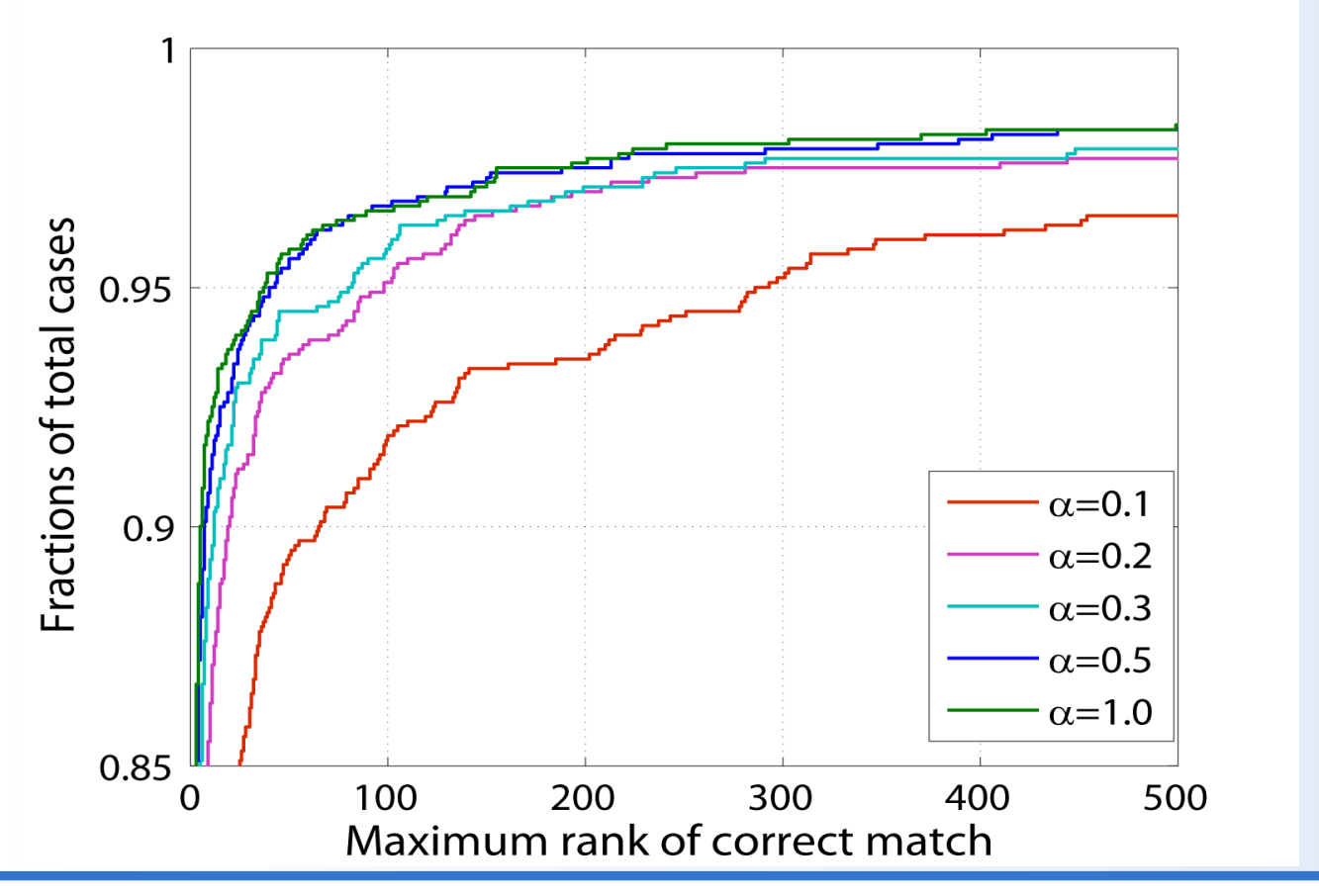
-given query spectrum *M* and candidate peptide *P* only consider peak in *M* if it also present in *P*



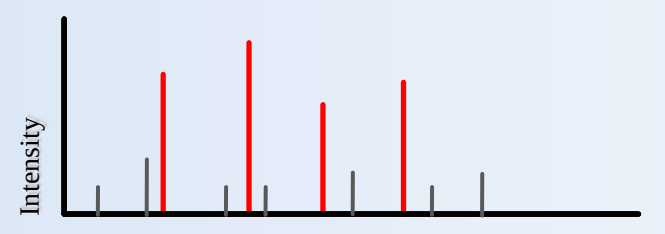
-score every candidate Pi in database against projection of M on Pi and keep only top N highest scoring candidates

Efficiency of filter, measured by the ranks of correct peptide match in a candidate list sorted by score





Scoring model for peptide pair:



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

FVGGPQR

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

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

Peptide pair: FVIGGPQR & AHSSMVG
-represent each peptide in vector format, then combine to represent a pair

```
P1 =      [ 0,   b,  0,  y,   0,  0,  b-H20,   0,  y,   0,  0,         0,   b,  0  …… ]
P2 =      [ y,   0,  0,  0,   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 simulated mixture spectra
Different models for spectra with different charge state and different peptide length

Percentage of cases with correct top peptide pairs:

Mixture coefficient (a)	M-SPLIT (Spectral library search)	MDB Search (all yeast peptides)	MDB Search (only spectral library peptides)	Iterative approach
1:1	97	87(97)	95(98)	81
1:0.5	92	79(92)	90(98)	74
1:0.3	80	66(86)	79(92)	57
1:0.2	63	50(77)	69(87)	30
1:0.1	34	19(43)	34(70)	6

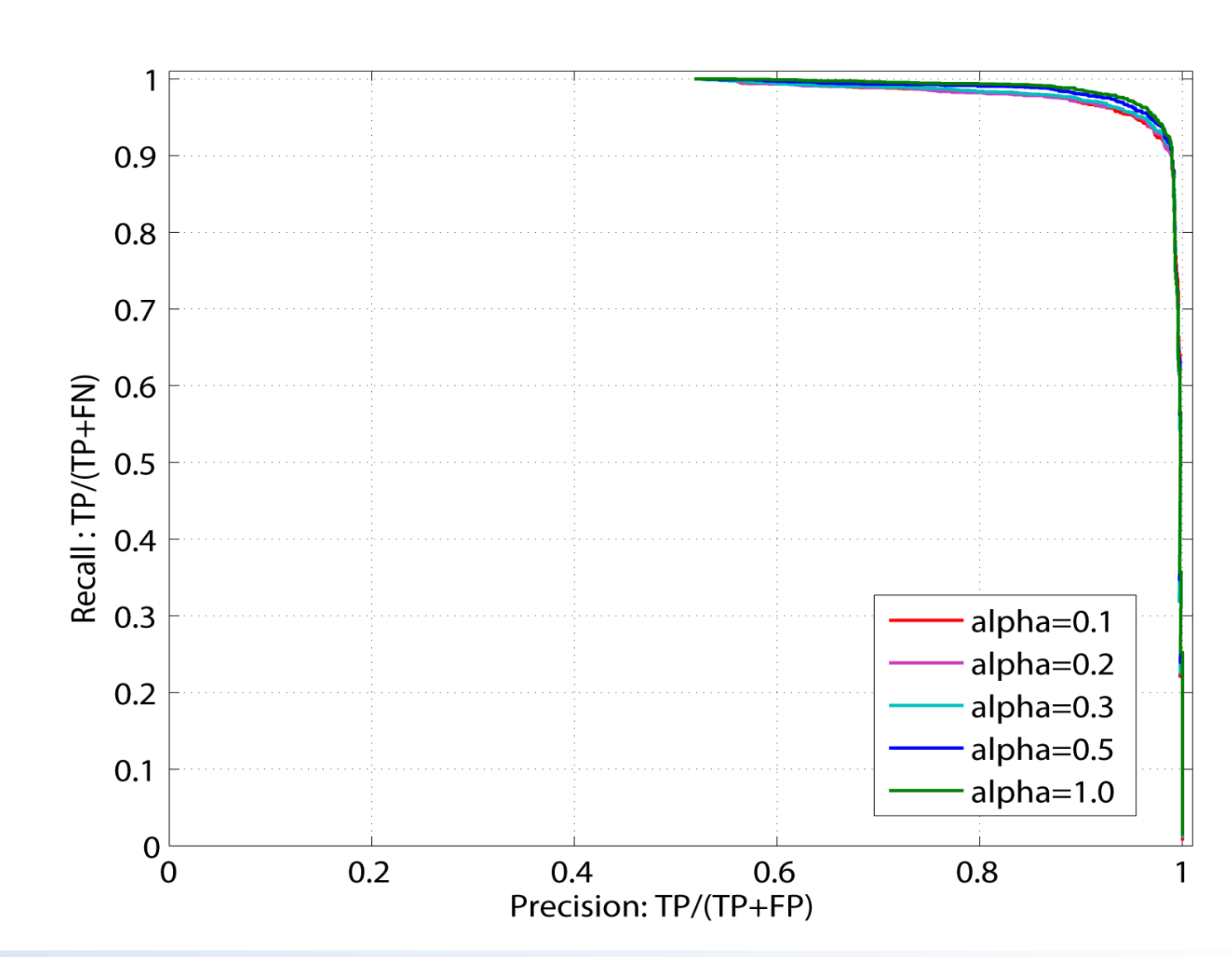
*M-SPLIT[2]: spectral library search method using NIST spectral library[3]

*number in parenthesis represents percentage of cases with correct pairs in top ten candidates

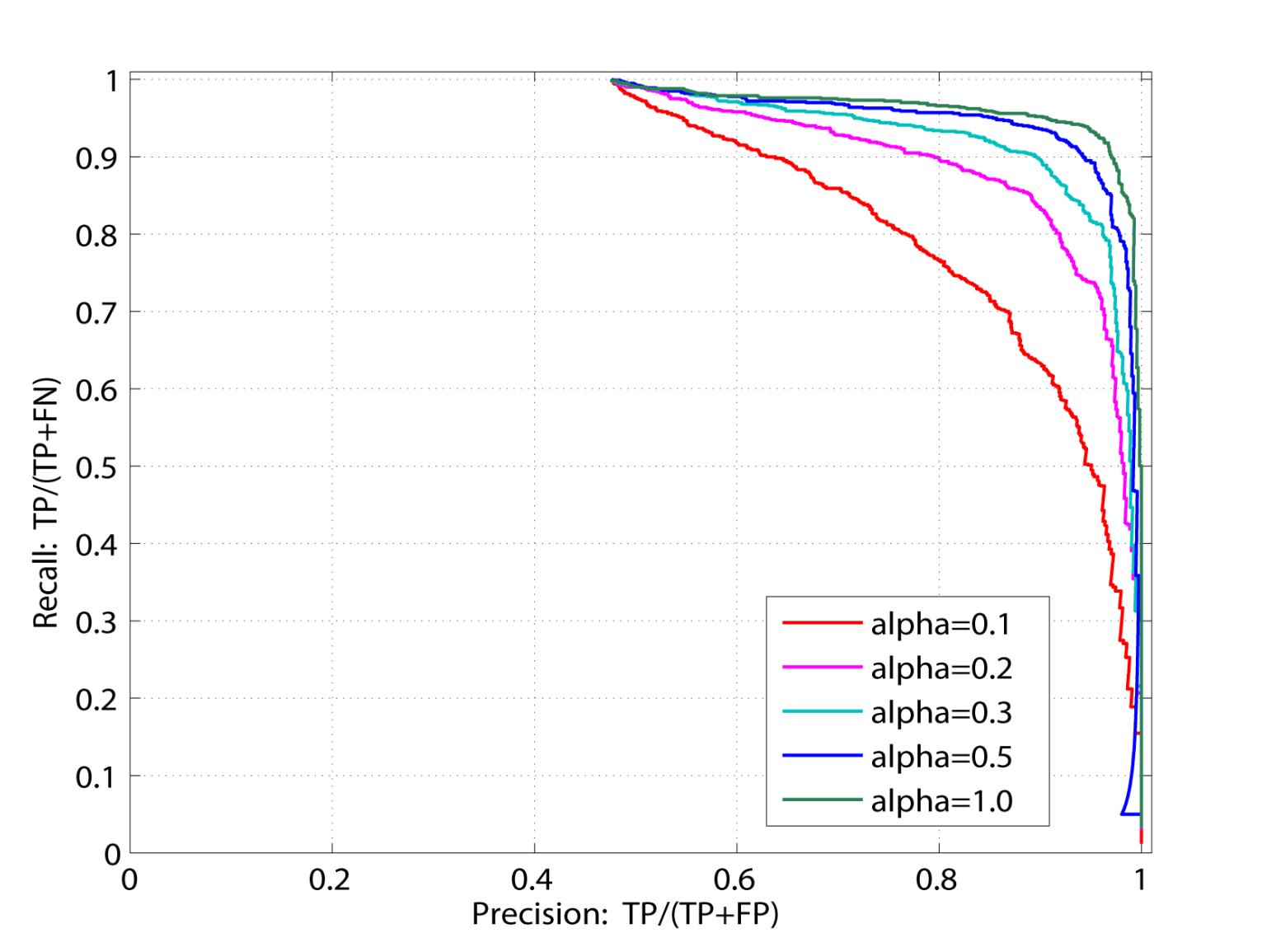
Classification of top matches:

Support Vector Machine (SVM) were used to learn discriminative models from following features:
1) Likelihood score of candidate peptide pair, 2) Likelihood score for each peptide alone, 3) explained intensity 4) % b/y presented, 5) longest contiguous stretches of b/y ions, 6) average mass errors between observed and theoretical peaks

No-matches vs. Matches (single + mixture)



Single-peptide matches vs. Mixture-matches



Results:

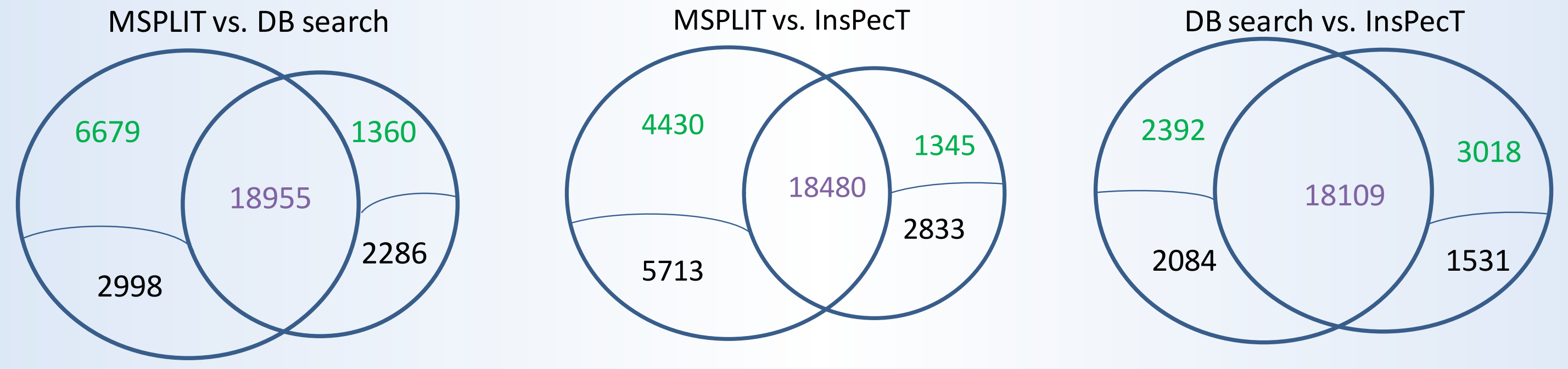
Dataset summary:
1) NIST yeast spectral library (*ver.* 6/06) [3].
2) Yeast cell lysate dataset: downloaded from Tranche/ProteomeCommons[4], made available by Vanderbilt [5]
~70,000 MS/MS spectra collected on yeast tryptic digest
Instrument: LTQ Orbitrap XL (Thermo Fisher Scientific)
One full MS Scan (m/z 300—2000) at resolution 60,000
Followed by 8 MS/MS scans on the LTQ

Comparison of M-SPLIT, InsPecT, and Database Search

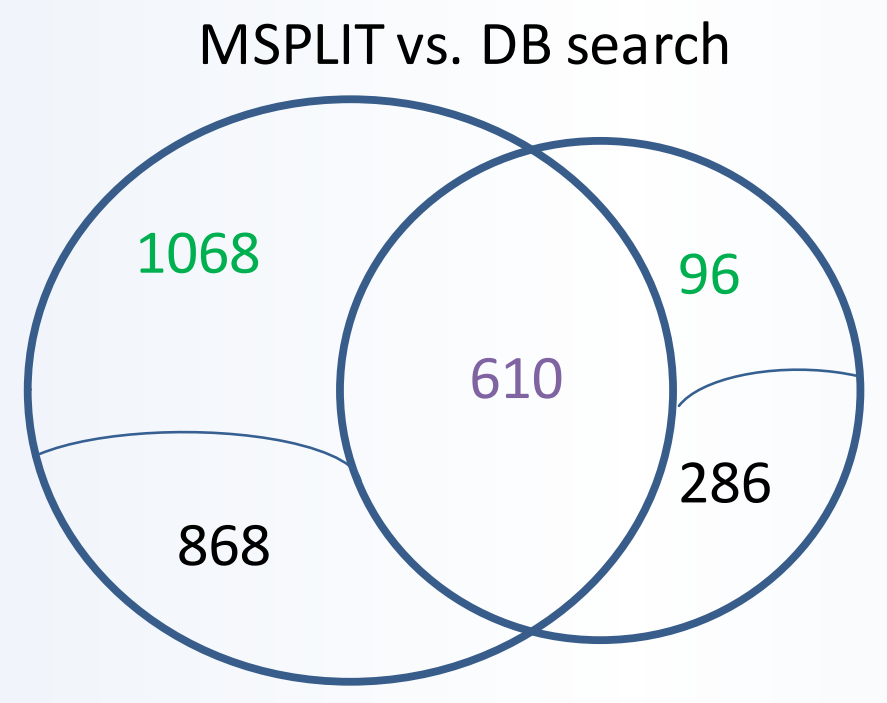
- All searches used 3 dalton parentmass tolerance and 0.5 dalton fragment mass tolerance
- FDR was estimated using target/decoy strategy @ 1%

	Spectra Identified			Unique peptides		
	Single	Mixture	Total	Single	Mixture	Total
M-SPLIT	26083	2549	28632	5833	2351	6304
MDBSearch	21611	974	22585	5092	1121	5304
InsPecT	22658	n/a	22658	5272	n/a	5272

Single-peptide matches:



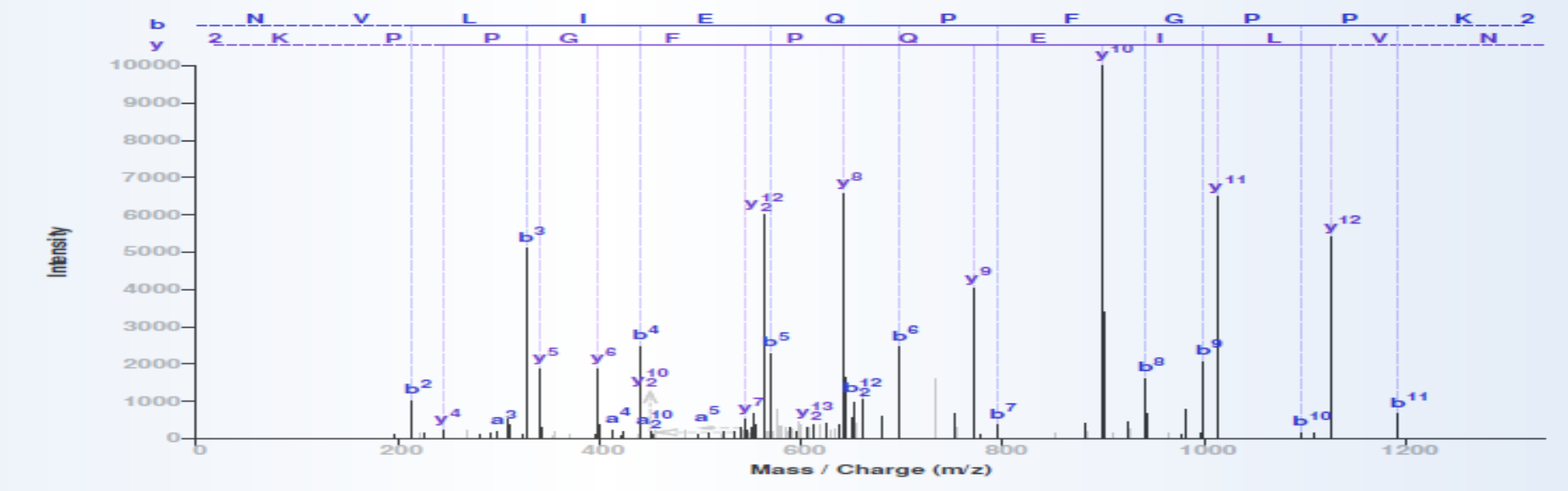
Mixture matches:



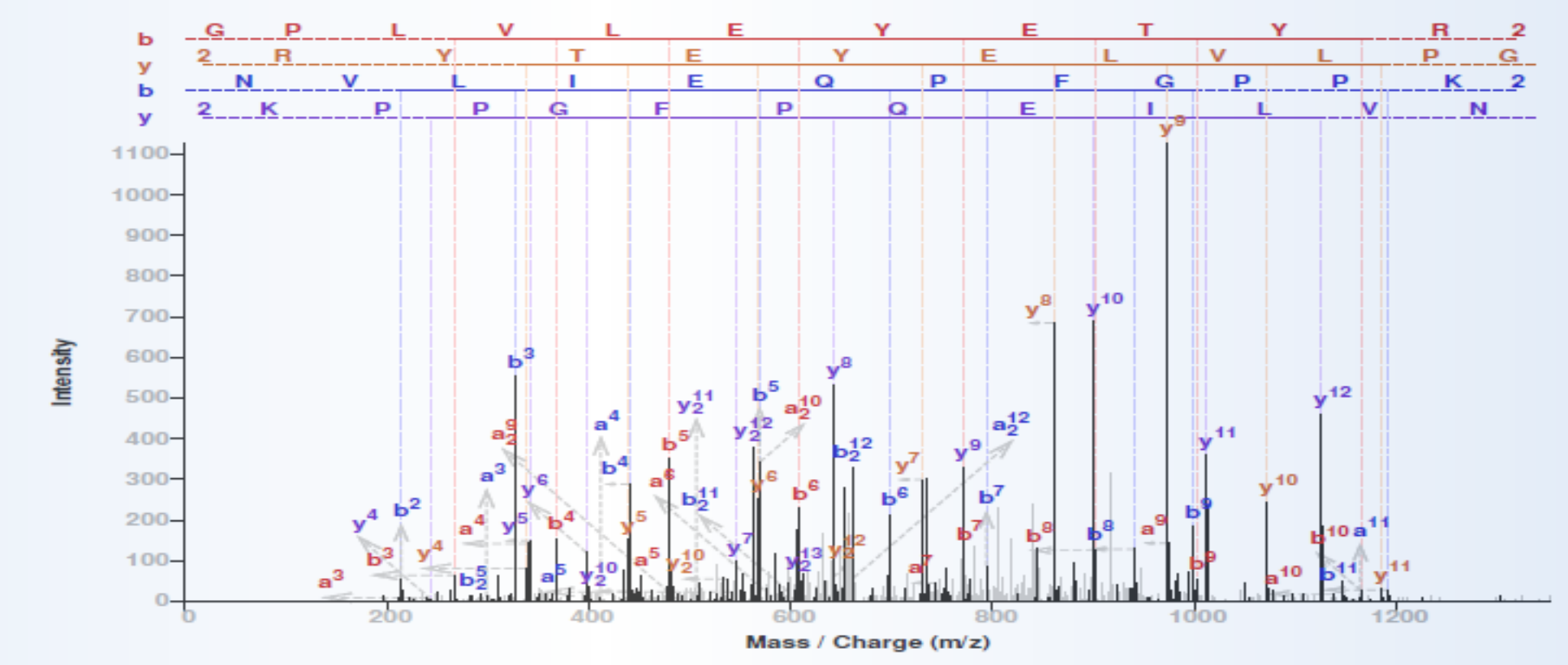
Purple: same top hit, classify as match
Green: same top hit, but below SVM threshold
Black: different top hit

Example:

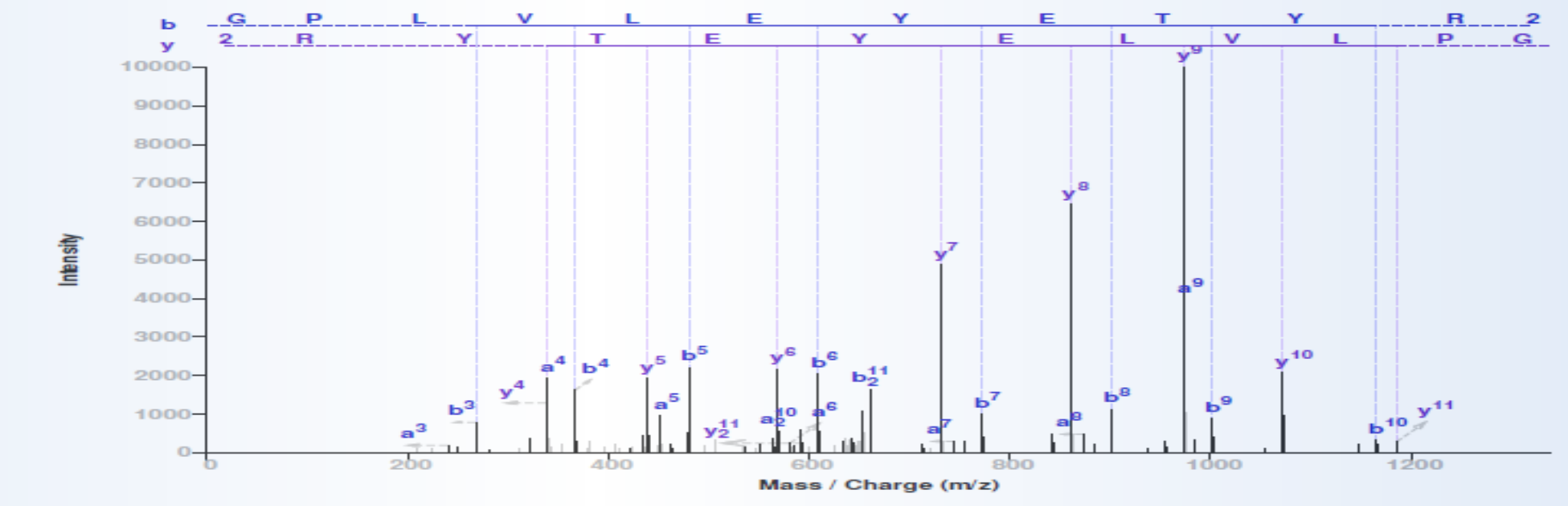
NIST library spectrum for peptide: NVLIEQPFQPPK



Mixture spectra from yeast dataset



NIST library spectrum for peptide: GPLVLEYETYR



Conclusion:

- Mixture spectra can be reliably identified using a database search method.
- Mixture spectra represented 5-10% of all identifiable spectra in a typical high-throughput experiment.
- Mixture spectra have a higher information content than single-peptide spectra, since each spectrum contains two peptides.
- Roughly 5-10% of unique peptides identified are present only in mixture spectra, thus contributing 5-10% gain in peptide identification.

Acknowledgment: This work was supported by the National Institutes of Health grant 1-P41-RR024851 from the National Center for Research Resources

References:

1. Kim et. al. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra *MCP*, 5(1), 2009.
2. Wang et. al. Peptide identification from mixture tandem mass spectra *MCP*, 2010
3. Deutsch et. al. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows EMBO reports 9, 5, 429-434 (2008)
4. Falkner et. al. ProteomeCommons.org IO Framework: reading and writing multiple proteomics data formats *Bioinformatics*, 23(2):262, 2007
5. Li et. al. Network-assisted protein identification and data interpretation in shotgun proteomics. *Mol Sys. Bio.,* 5(1), 2009