基于相关谱图对的非限制性修饰检测
Unrestrictive Modification Detection Based on Related Spectral Pairs

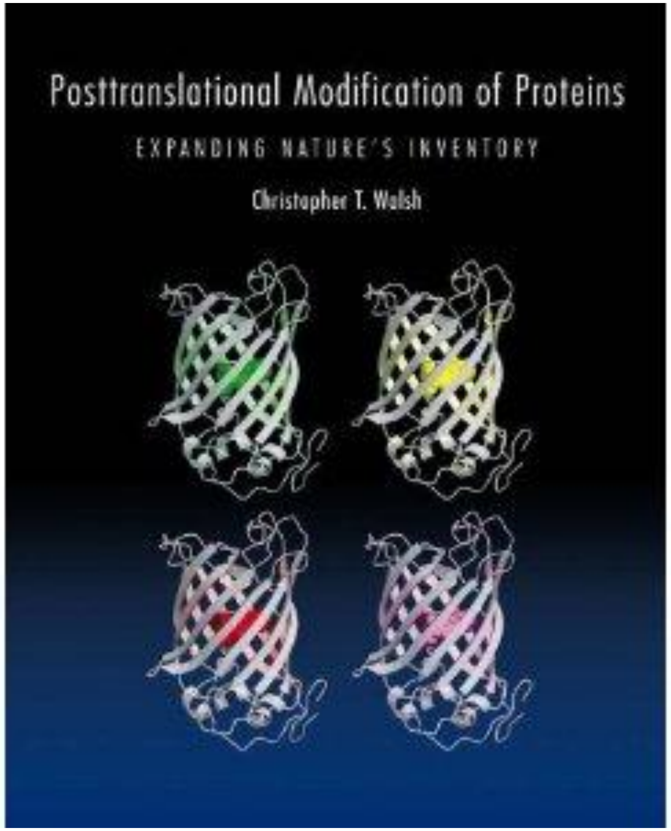
付岩
中国科学院计算技术研究所

2010.11.11

Joint work with Ding Ye, Liyun Xiu, and other pFind group members

Complexity of life

- # protein isoforms >> # genes
- This is in large part due to posttranslational modifications of proteins that provide covalent alterations to protein backbones and side chains that increase proteome complexities.


Posttranslational Modification of Proteins
EXPANDING NATURE'S INVENTORY
Christopher T. Walsh

MS-based proteomic analysis of PTM

- Wilkins MR, Gasteiger E, Gooley AA, Herbert BR, Molloy MP, Binz PA, Ou K, Sanchez JC, Bairoch A, Williams KL *et al:* ***High-throughput mass spectrometric discovery of protein post-translational*** **modifications**. *J Mol Biol 1999, 289(3):645-657.*

- Mann M, Jensen ON: **Proteomic analysis of post-translational modifications**. *Nat Biotechnol 2003,* 21(3):255-261.

- Jensen ON (2004) **Modification-specific proteomics: Characterization of posttranslational modifications by mass spectrometry**. *Curr Opin Chem Biol 8:33–41.*

- Witze ES, Old WM, Resing KA, Ahn NG: **Mapping protein post-translational modifications with mass spectrometry**. *Nat Methods 2007, 4(10):798-806.*

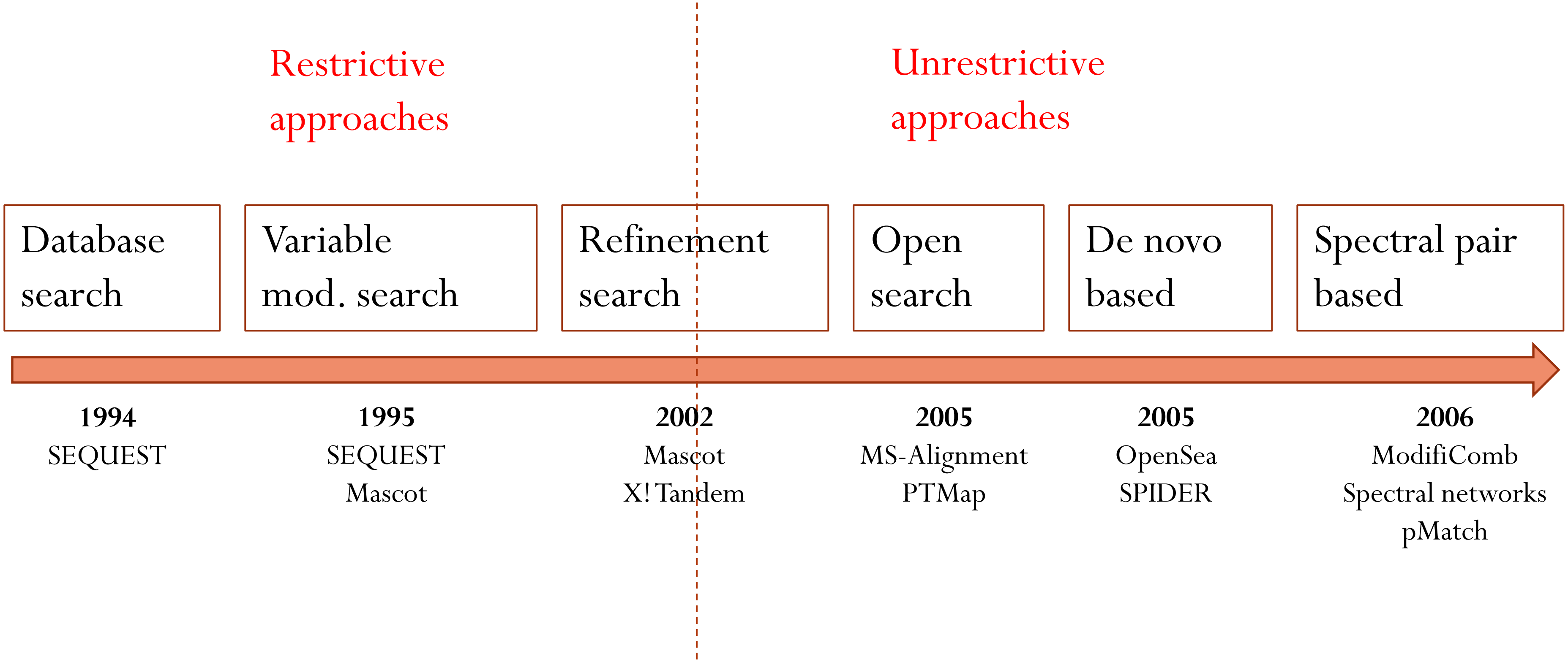
(Open) problems of modification identification via mass spectrometry

- Site identification of interested modifications (e.g., phosp. acyl. methy. ect)
- Effects of modifications on peptide fragmentation (e.g., neutral loss, suppression)
- Unambiguous site assignment (e.g., phosp.)
- Search speed(e.g, when many variable modifications are specified)
- False discovery rate control of modification identifications

(Open) problems of modification identification via mass spectrometry (cont.)

- Recognition of signals of interested modifications (e.g., phosphorylation spectra)
- Discrimination between in vivo and in vitro modifications (e.g., deamidation)
- Identification of complex modification structures (e.g., glycan, ubiquitin)
- Combinatorics of modifications (e.g., histones)
- **Discovery of unanticipated/novel modifications**

History of modification detection algorithms via MS/MS



Restrictive approaches

Unrestrictive approaches

Database search

Variable mod. search

Refinement search

Open search

De novo based

Spectral pair based

1994
SEQUEST

1995
SEQUEST
Mascot

2002
Mascot
X! Tandem

2005
MS-Alignment
PTMap

2005
OpenSea
SPIDER

2006
ModifiComb
Spectral networks
pMatch

Computational difficulties of modification detection

- Restrictive approaches
  - Combinatory explosion (hundreds known, a few allowed)
  - Guessing of modification types
  - Inability to find novel modifications

- Unrestrictive approaches
  - Low search speed
  - Increasing random matches
  - Sensitive to spectrum quality

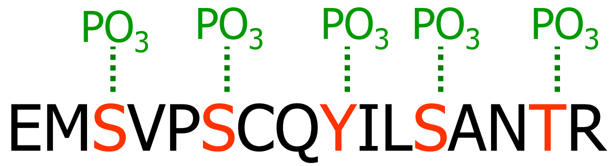
- Spectral identification rate is 10-30% in current proteomics

Restrictive approach
pFind: sequence DB search engine

Fu Y, et al. Bioinformatics 2004; Wang L, et al. RCM 2007

- Scoring function: kernel spectral dot product
- Significance evaluation: E-value, FDR

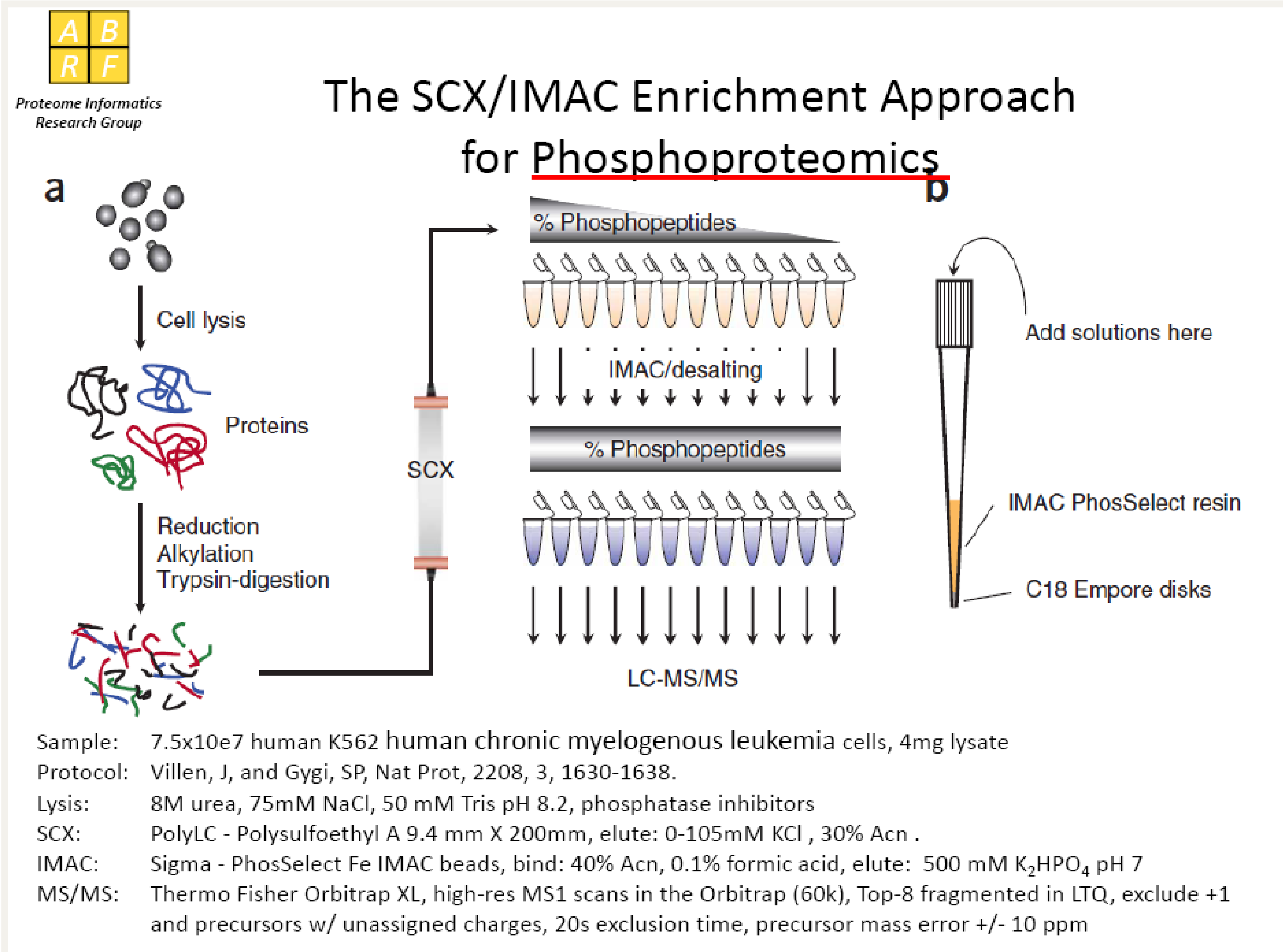
- Variable modification search
  - Enumerate all possible modification forms



Five potential modification sites, 32 candidate peptides

  - Database indexing and optimized search flow

ABRF-iPRG2010



Proteome Informatics Research Group

The SCX/IMAC Enrichment Approach for Phosphoproteomics

Sample: 7.5x10e7 human K562 human chronic myelogenous leukemia cells, 4mg lysate

Protocol: Villen, J, and Gygi, SP, Nat Prot, 2208, 3, 1630-1638.

Lysis: 8M urea, 75mM NaCl, 50 mM Tris pH 8.2, phosphatase inhibitors

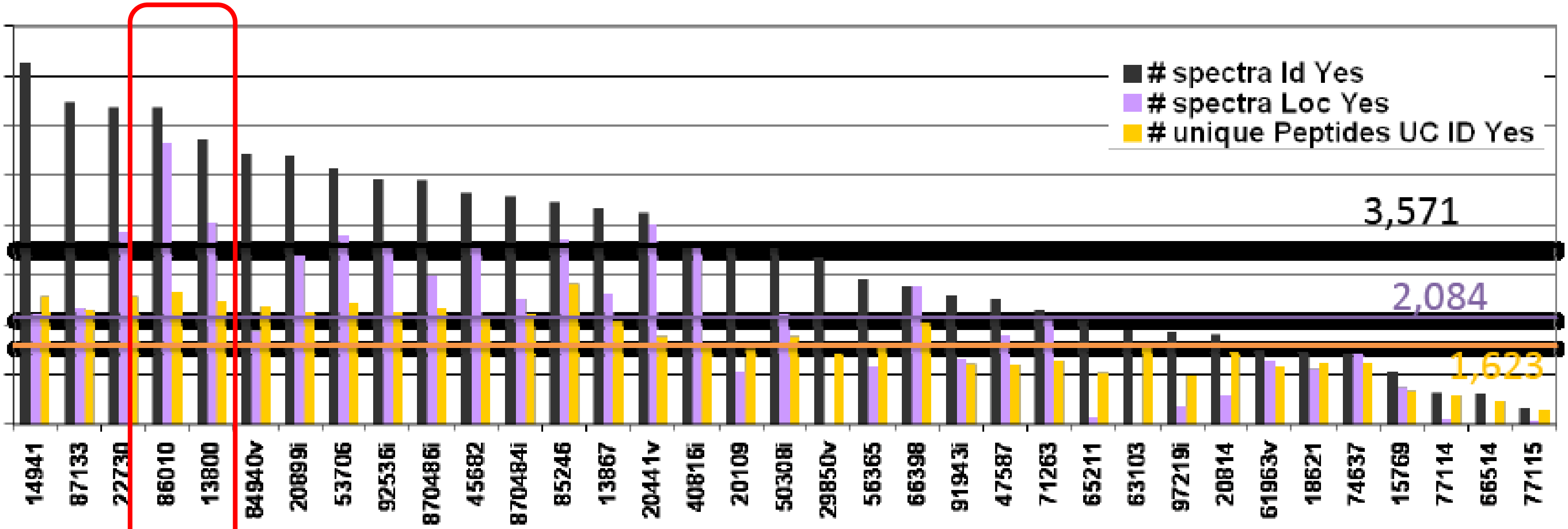
SCX: PolyLC - Polysulfoethyl A 9.4 mm X 200mm, elute: 0-105mM KCl , 30% Acn .

IMAC: Sigma - PhosSelect Fe IMAC beads, bind: 40% Acn, 0.1% formic acid, elute: 500 mM K_2HPO_4 pH 7

MS/MS: Thermo Fisher Orbitrap XL, high-res MS1 scans in the Orbitrap (60k), Top-8 fragmented in LTQ, exclude +1 and precursors w/ unassigned charges, 20s exclusion time, precursor mass error +/- 10 ppm

A B
R F
Proteome Informatics Research Group

Results



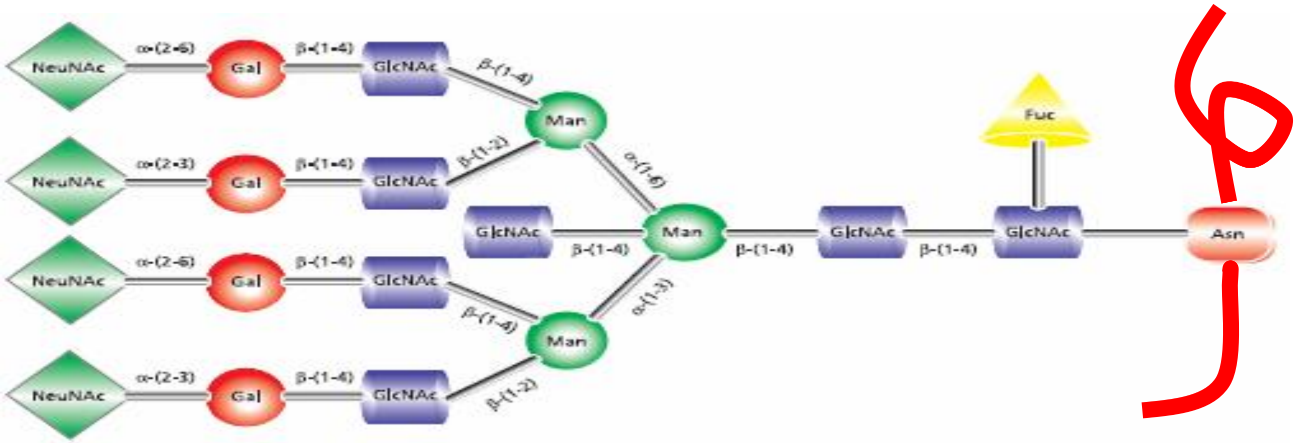
	14941	87133	22730	86010	13800	84940v	20899i	53706	92536i	870486i	45682	870484i	85246	13867	20441v	40816i	20109
Pre masses adjusted	n	n	y	y	y	y	y	y	y	y	n	y	n	n	n	n	y
N term acetyl	y			y	y		y		y	y					y		
Used > 1 search engine	y	y	n	n	n	n	y	n	y	n	y	n	y	y	n	n	y
Pre mass filtering	n	y	y	n	y	y	y	y	n	y	n	y	n	n	n	n	n
Solution charge filtering	n	n	y	n	n	n	n	n	n	y	n	y	n	n	n	n	n
Localization software	y	y	y	n	y	y	n	y	y	y	y	y	y	y	n	n	n
Peptide ID Software	Mu*	Mu*	Se	Pf	Pf	Se	Mu*	Om	Ma	Sm	Mu	Sm	Mu	Mu	Xt	Ma	Mu
Localization software	Ih	Ih	As		Ih	As		Ph	Mq	Sm	Ih	Sm	Ih	As			

	50308i	29850v	56365	66398	91943i	47587	71263	65211	63103	97219i	20814	61963v	18621	74637	15769	77114	66514	77115
Pre masses adjusted	n	n	n	n	y	n	n	n	n		n	n	n	y	n		y	n
N term acetyl				y	y		y											
Used > 1 search engine	n	n	n	n	y	n	y	n	y		n	y	y	n	n		n	y
Pre mass filtering	n	n	n	n	n	n	n	n	n		y	n	n	n	n		y	n
Solution charge filtering	n	n	n	n	n	n	n	n	n		n	n	n	n	n		n	n
Localization software	y	n	y	y	y	y	y	y	n		y	n	n	n	y		y	n
Peptide ID Software	Ma	Ma	In	Ma	Mu	Se	Mu	Se	Mu		Ma	Xt	Mu	Pv	Ma		Se	Sc
Localization software	As		Pl	Mq		Ps	Pl	Nn			As	Ih			Ih		As	

Ma=Mascot, Ms=MsInspect, Mu=Multiple, Mu*=Multiple + Spec Lib, Om=OMSSA, Pf=pFind, Pv=Pview, Sc=Scaffold, Se=SEQUEST, Sm=Spectrum Mill, Xt=X!Tandem
As=Ascore, Ih=In-house, In=InsPecT, Mq=MaxQuant, Nn=NNScore, Ph=Phosphinator, Pl=Phosphate Localization Score, Ps=PhosphoScore, Sm=Spectrum Mill

Identification of Core Fucosylated Glycoproteins using pFind search engine

Jia W, Lu Z, Fu Y, et al. Molecular & Cellular Proteomics 2009

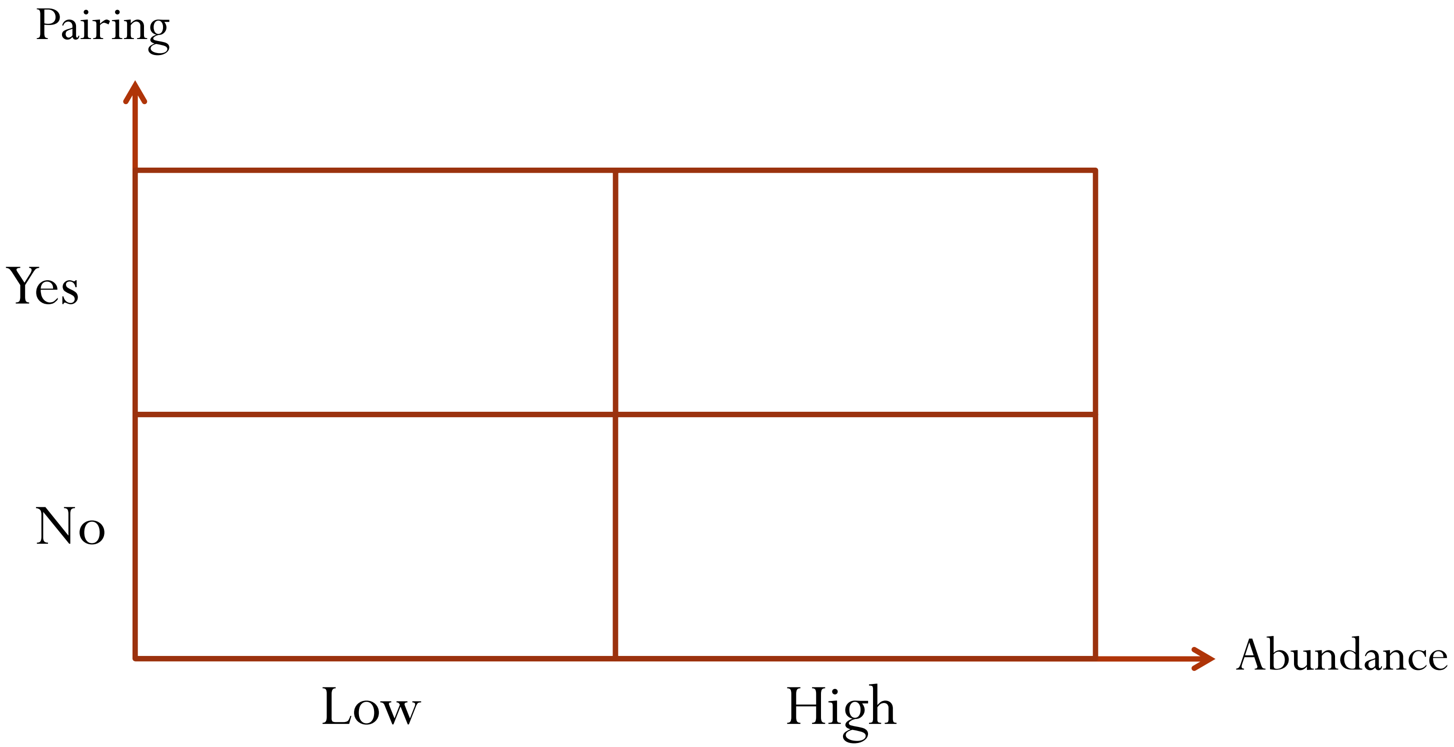


6134 spectra	Identified peptide	Identified spectra	Identification rate
pFind	115	1973	32.16%
SEQUEST	82	1311	21.37%
Mascot	98	1813	29.56%

Unrestrictive approaches

- Divide and conquer
  - Divide modifications in MS/MS into different categories
  - Use different strategies for different categories of modifications
  - Goal: maximize modification detections with minimal effort

Categorization of modifications in MS/MS



Pairing

Yes

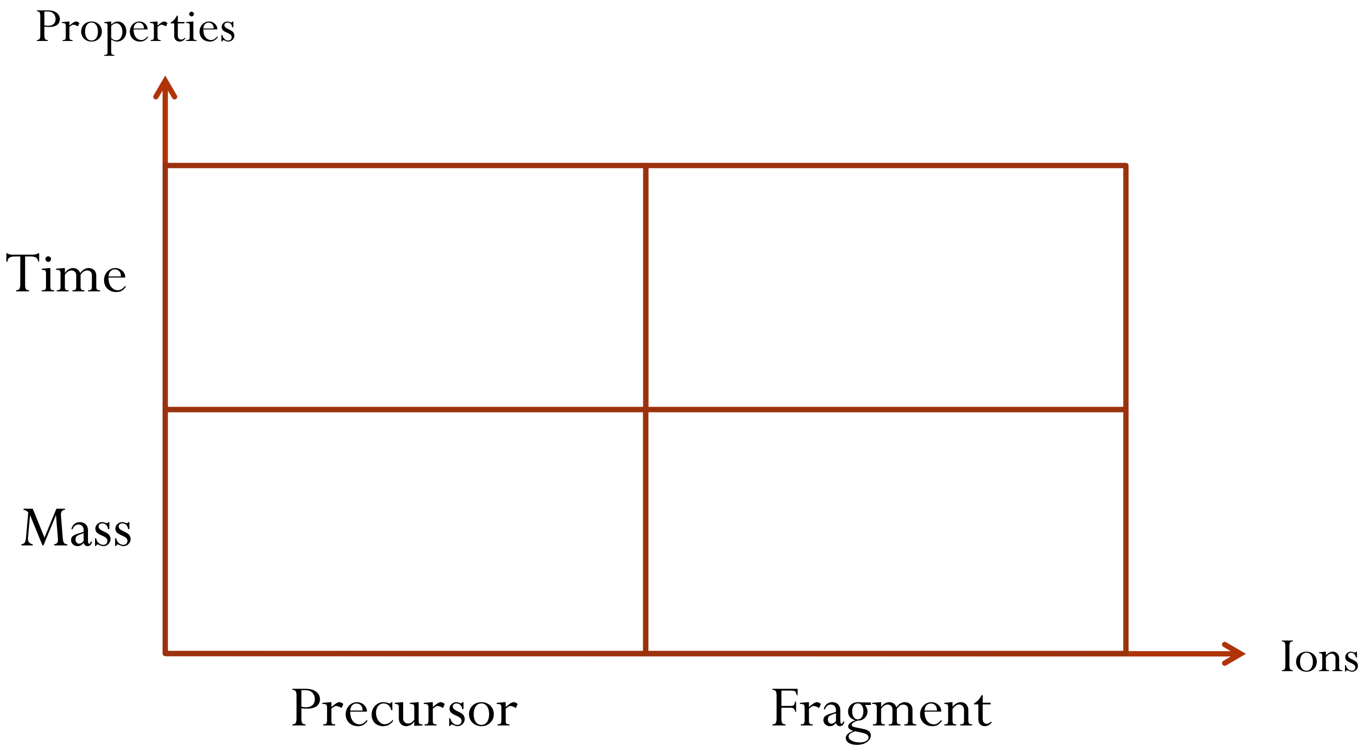
No

Abundance

Low

High

Information from LC-MS/MS



Properties

Time

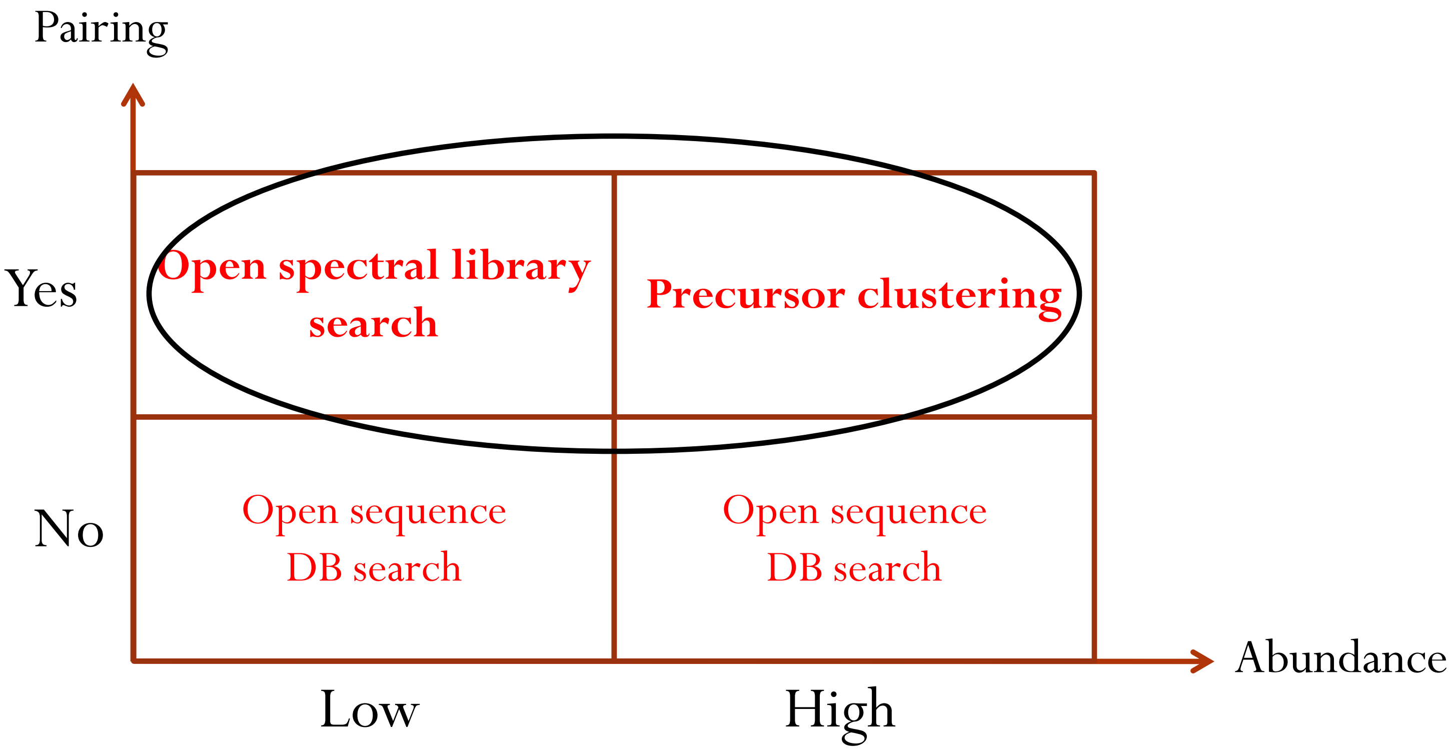
Mass

Ions

Precursor

Fragment

Unrestrictive approaches



	Low	High
Yes	**Open spectral library search**	**Precursor clustering**
No	Open sequence DB search	Open sequence DB search

pMatch tool

Open spectral library search

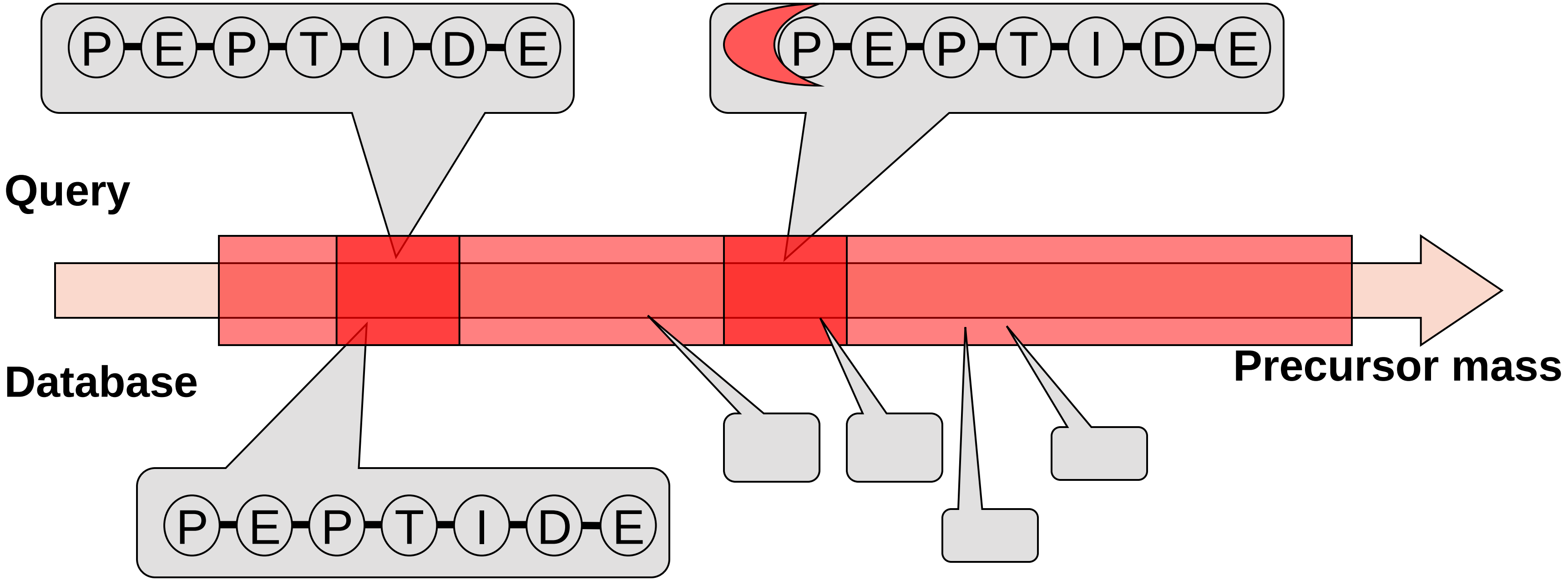
pMatch: Open spectral library search

Ye D, Fu Y, et al. Bioinformatics 2010

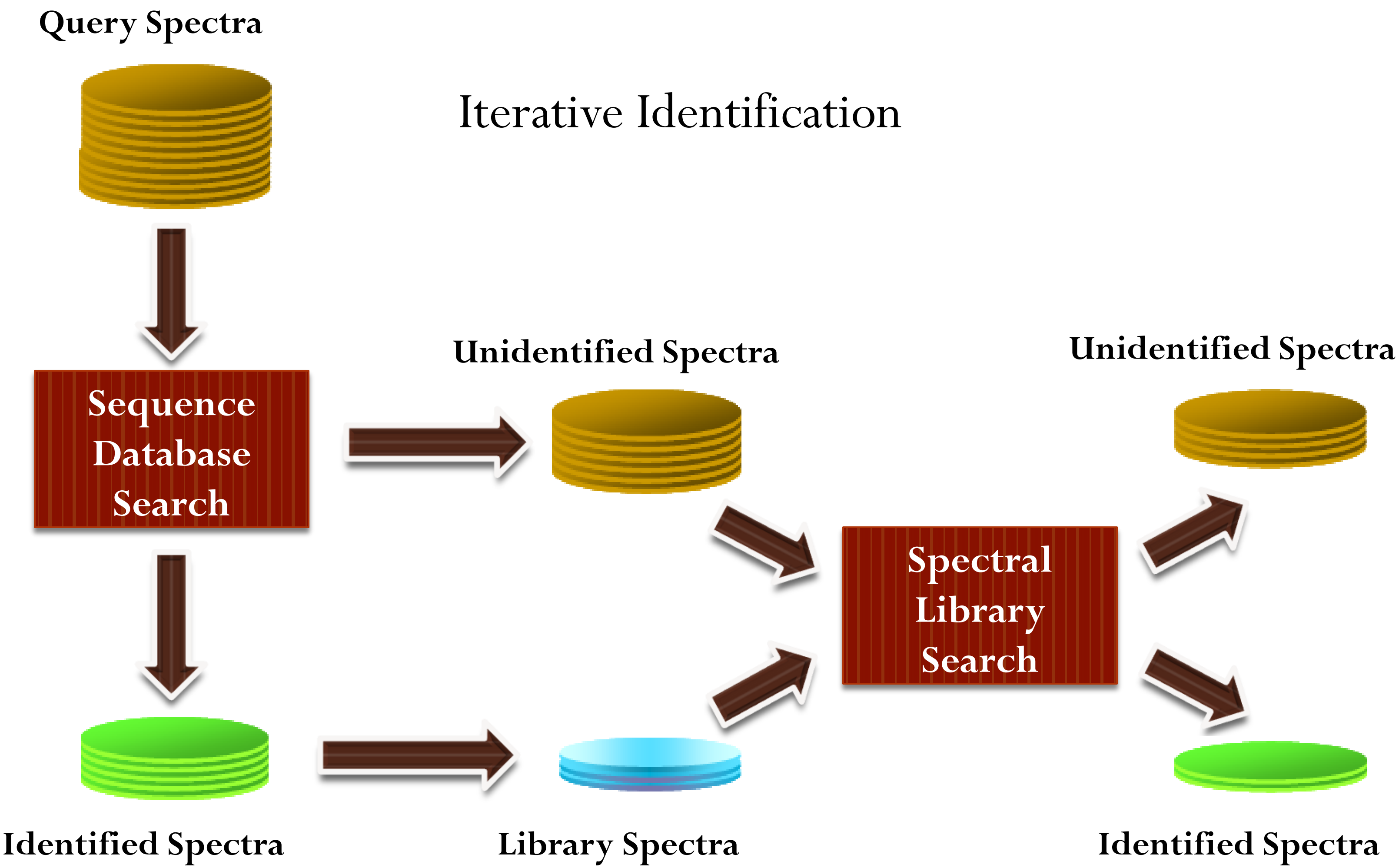
- Spectral library search
  - Search against identified experimental spectra rather than peptide sequences
- Open search mode
  - Match unmodified spectra to modified spectra
- Assumption
  - Modified/unmodified peptide pairs

Open search

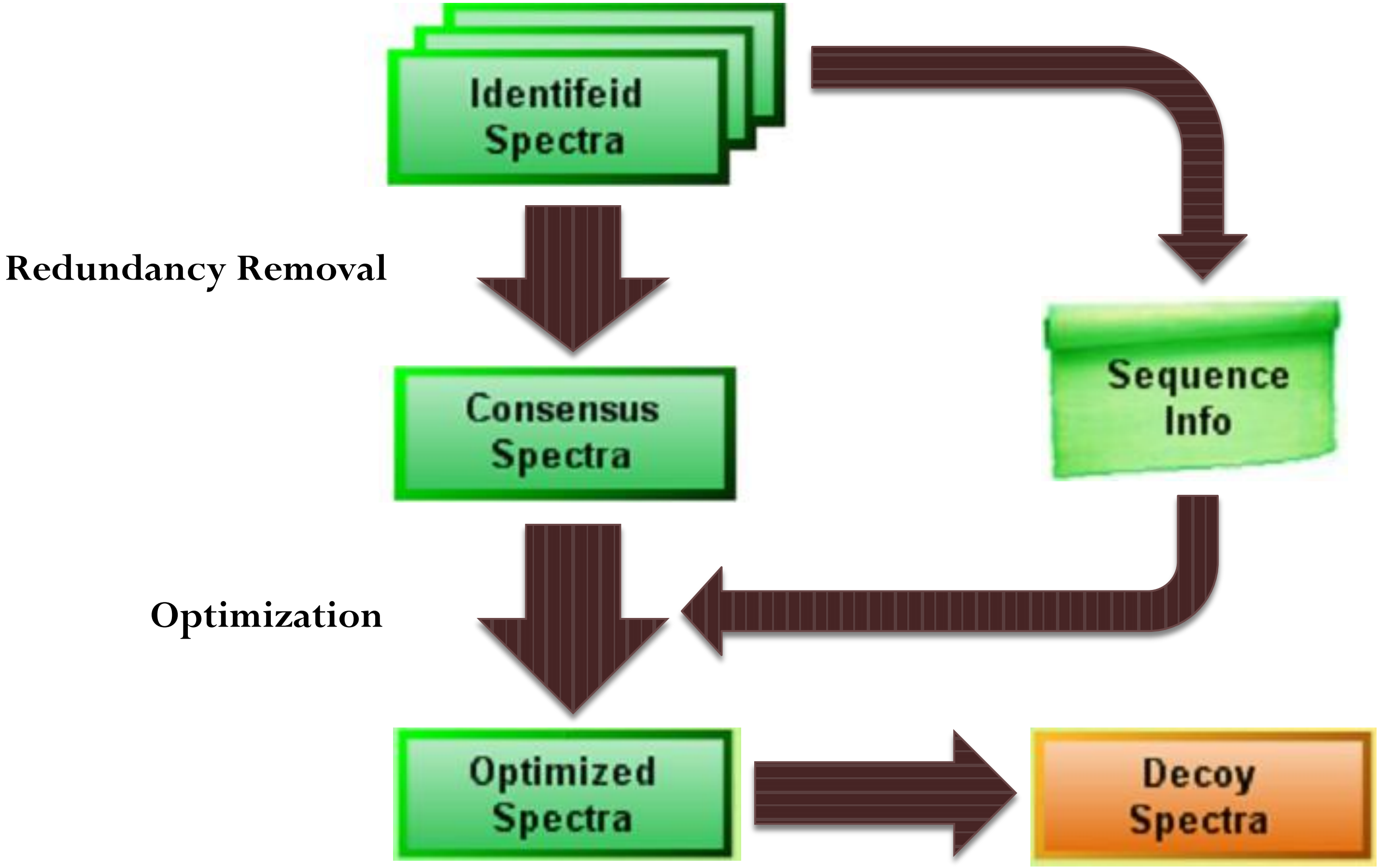
- Precursor mass tolerance:
  - Conventional search mode: ±3 Da
  - Open search mode: ±300 Da or more



Experimental Workflow



Library Construction



Identifeid Spectra

Redundancy Removal

Consensus Spectra

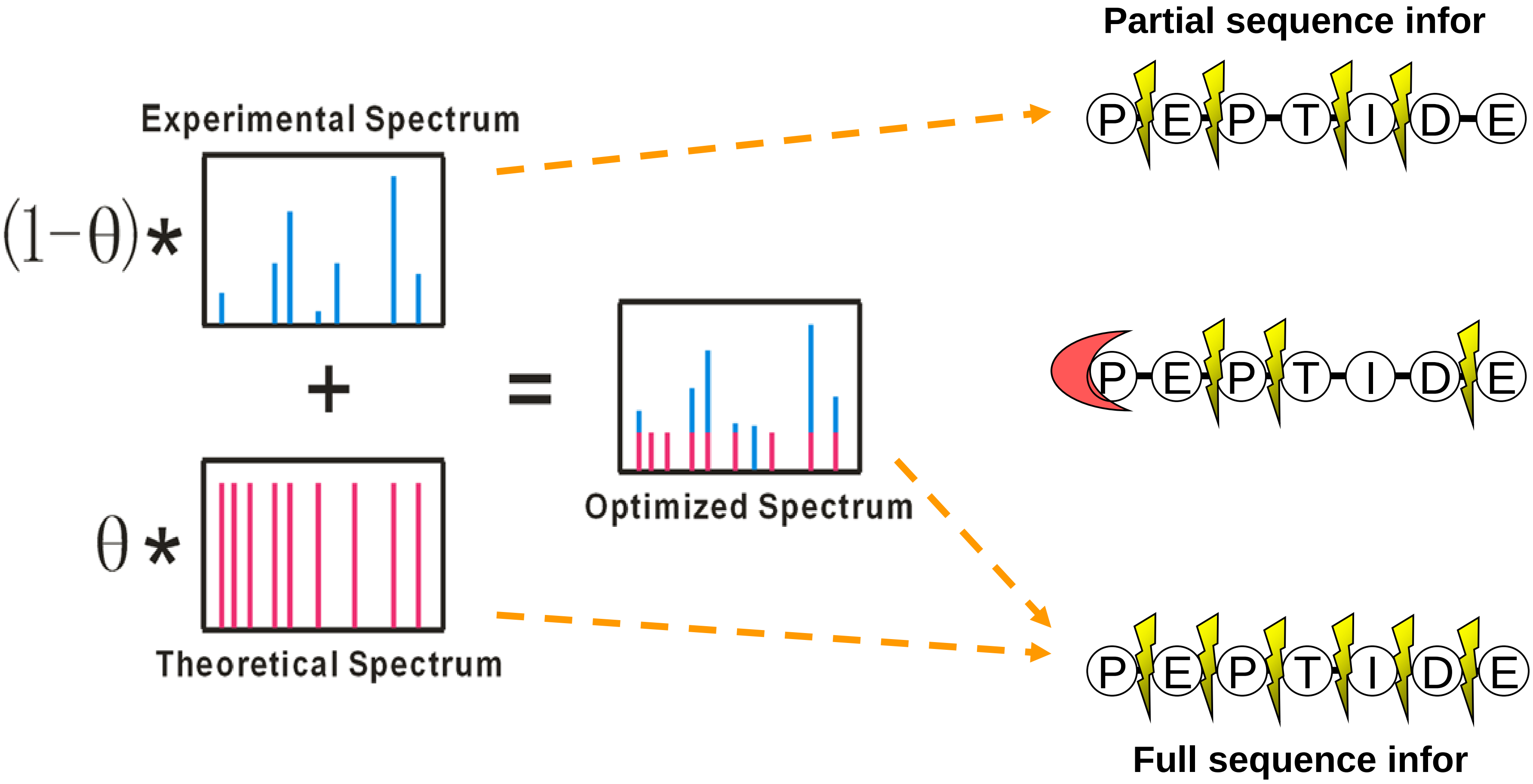
Sequence Info

Optimization

Optimized Spectra

Decoy Spectra

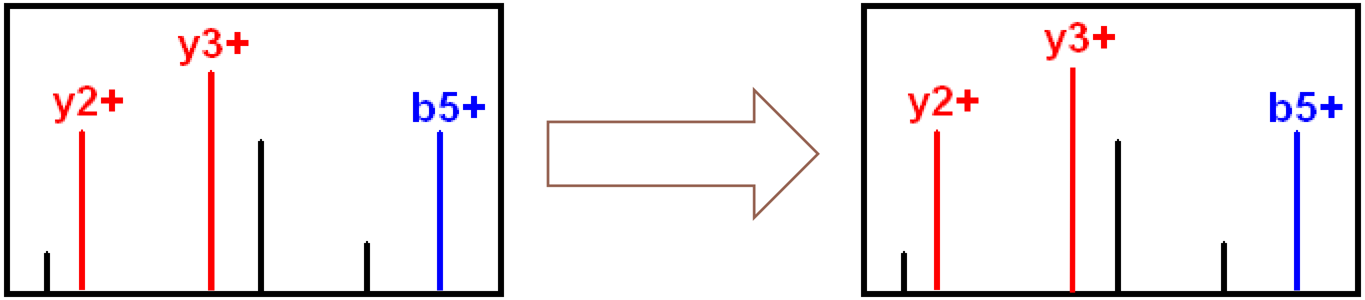
Spectrum optimization



Decoy Spectra for FDR control

- “pseudo-reversed” peptide sequence
  - **ABCDEFK -> FEDCBAK**

- Peak m/z values are shifted



Real Spectrum

Decoy Spectrum

Scoring

- Peak hits are determined considering **mass shifts** caused by unanticipated modifications

- Match score is calculated from peak hits
- Two sub-scores:
  - How similar are the two spectra?
    - SDP_Score : Spectral Dot-Product Score

  - How does one match stand out from the rest candidates?
    - P_Score : Probability-based Score

SDP_Score

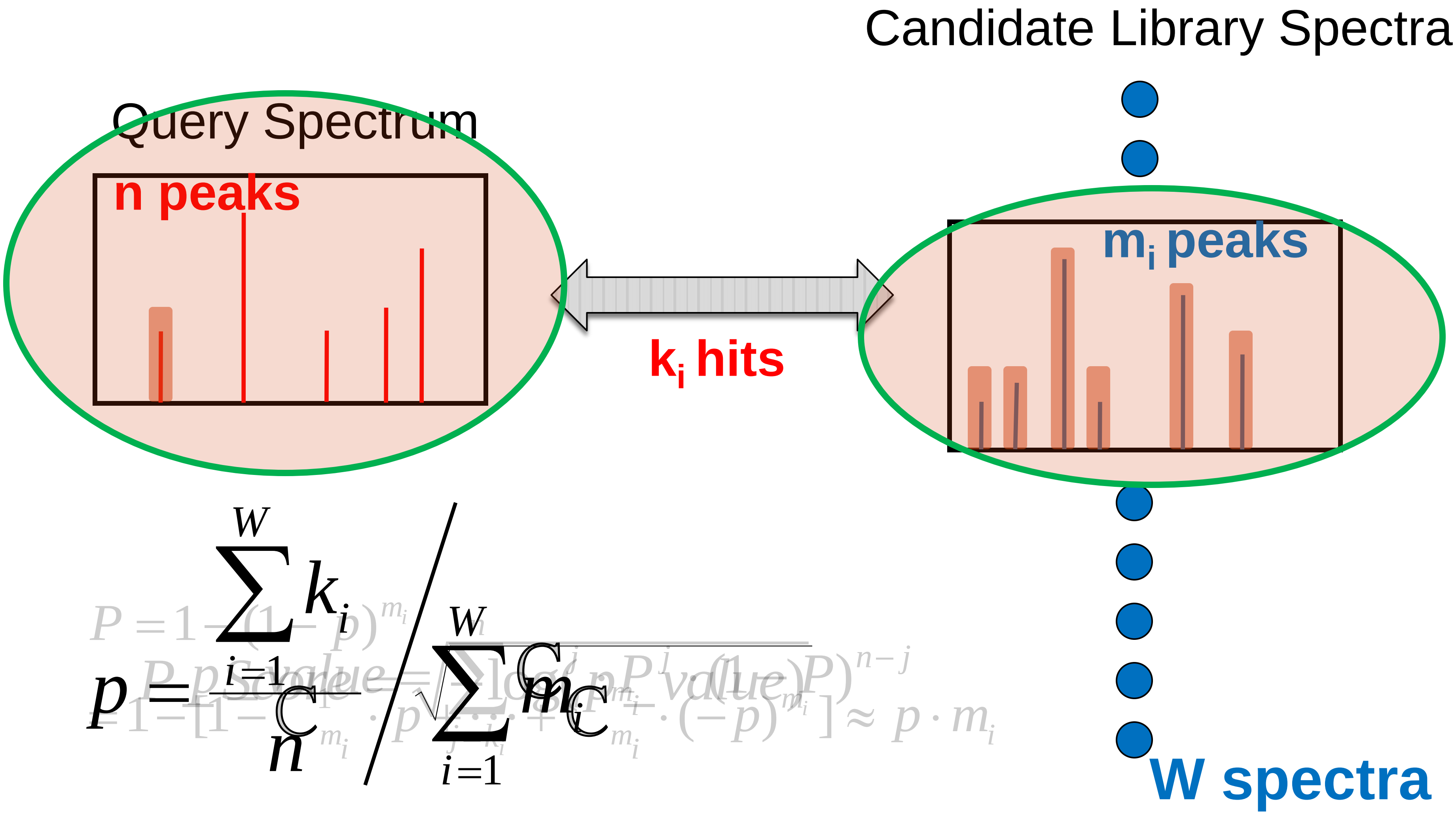
- SDP_Score is the cosine of the angle between the two spectral vectors

$$SDP\_Score = \frac{\sum\limits_{peak\_hits} I_Q \cdot I_L}{\sqrt{\sum\limits_{query\_peaks} I_Q{}^2} \cdot \sqrt{\sum\limits_{library\_peaks} I_Q{}^2}}$$

I_Q : Intensity of the Query peak

I_L : Intensity of the Library peak

P_Score



$$p = \frac{\sum_{i=1}^{W} k_i}{n} \Bigg/ \sum_{i=1}^{W} m_i$$

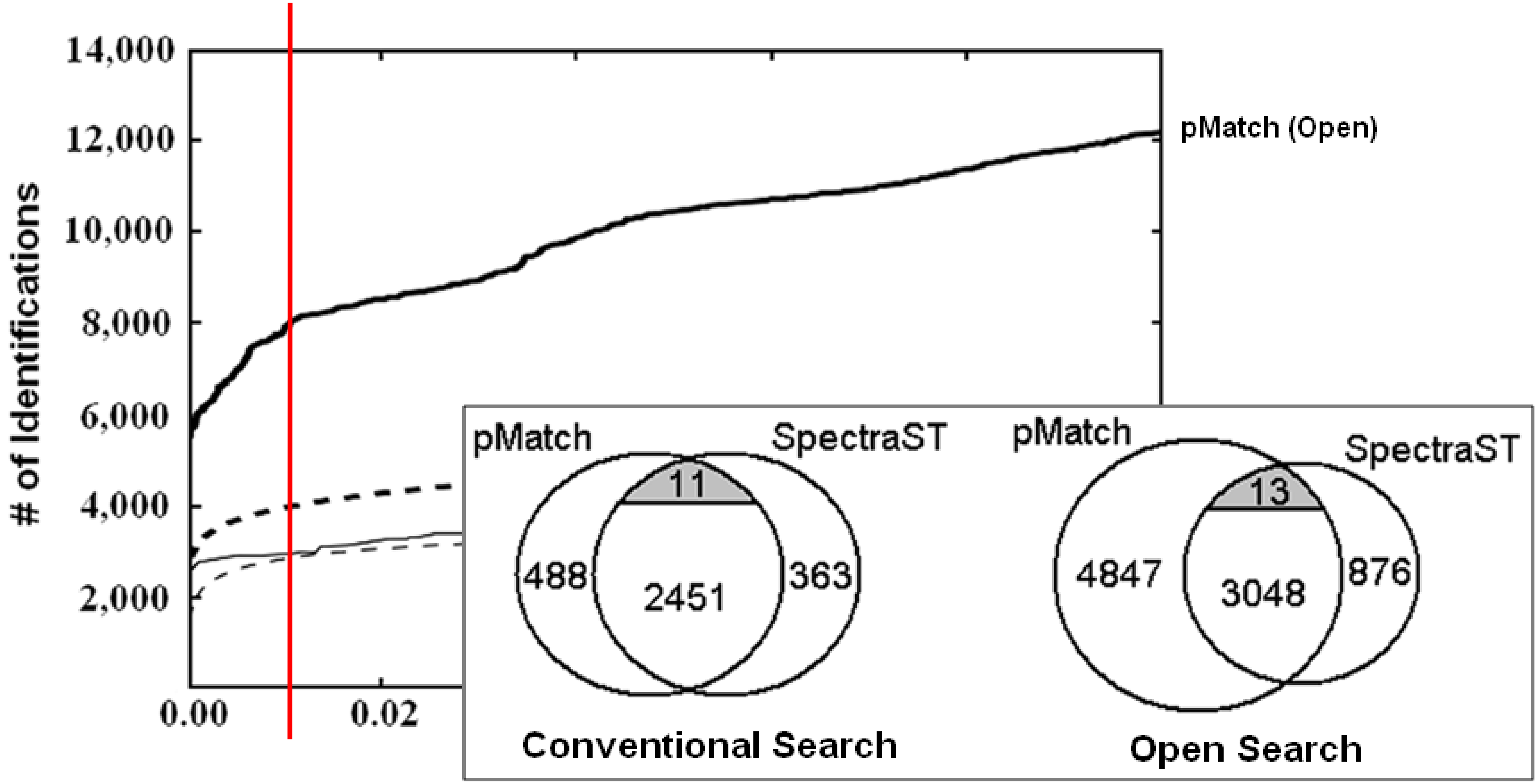
$$P = 1-(1-p)^{m_i}$$

$$P\_Score = -\log(P\_value)$$

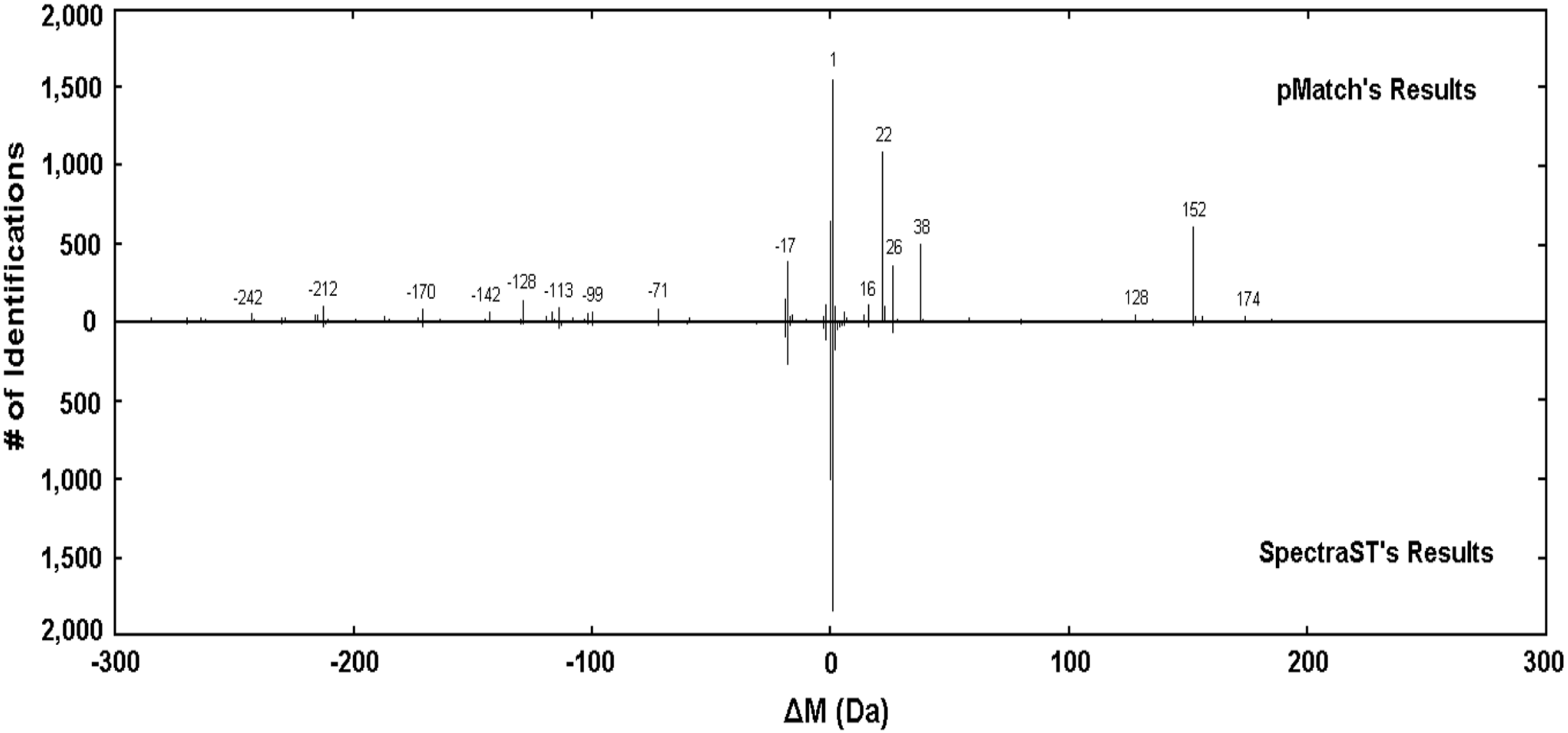
$$P\_value = \sum C_n^j P^j \cdot (1-P)^{n-j}$$

$$P = 1-[1-C_{m_i}^1 \cdot p + \cdots + C_{m_i}^{m_i} \cdot (-p)^{m_i}] \approx p \cdot m_i$$

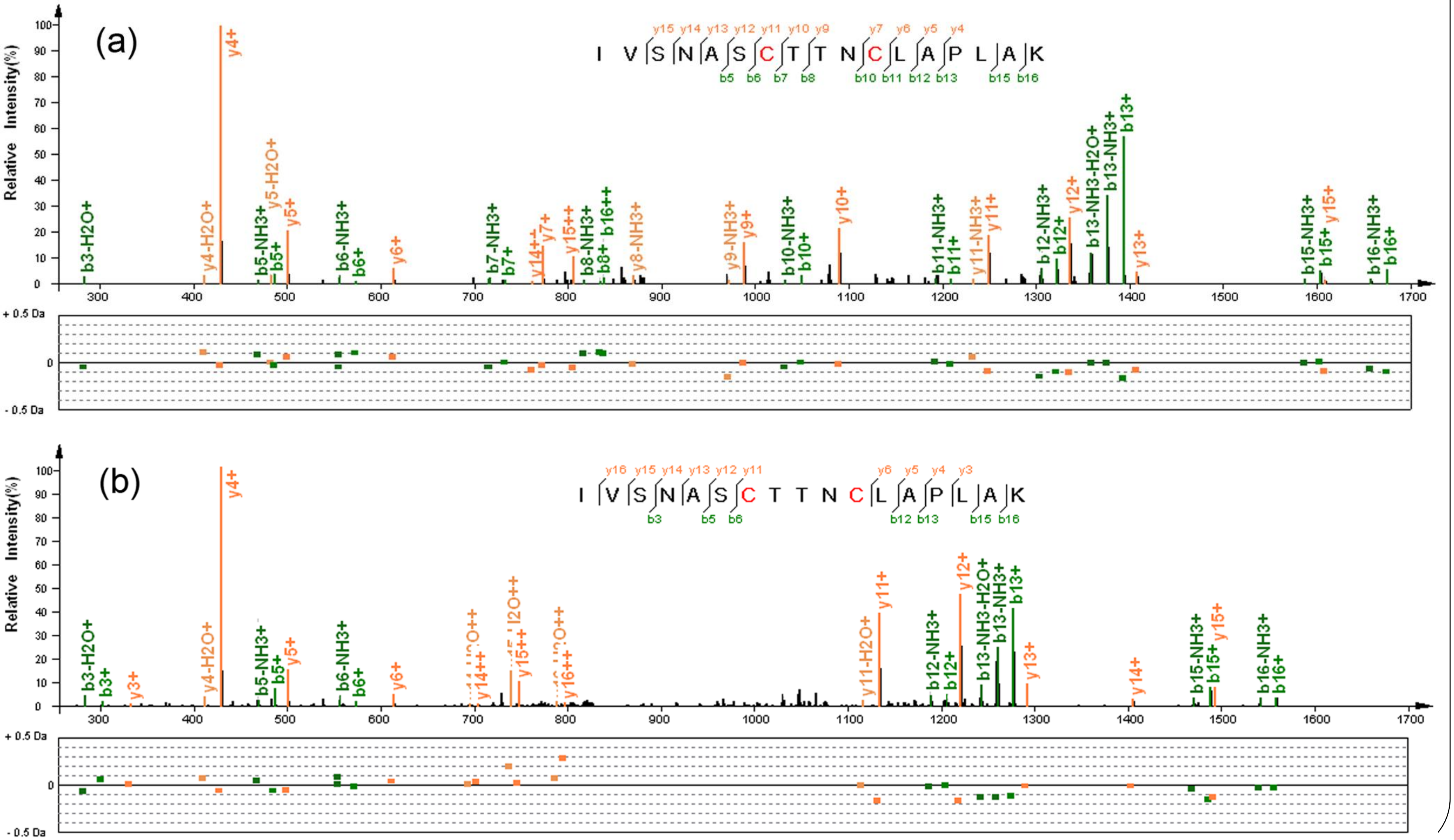
ISB-18mix dataset (Klimek, J, et al. JPR 2008)



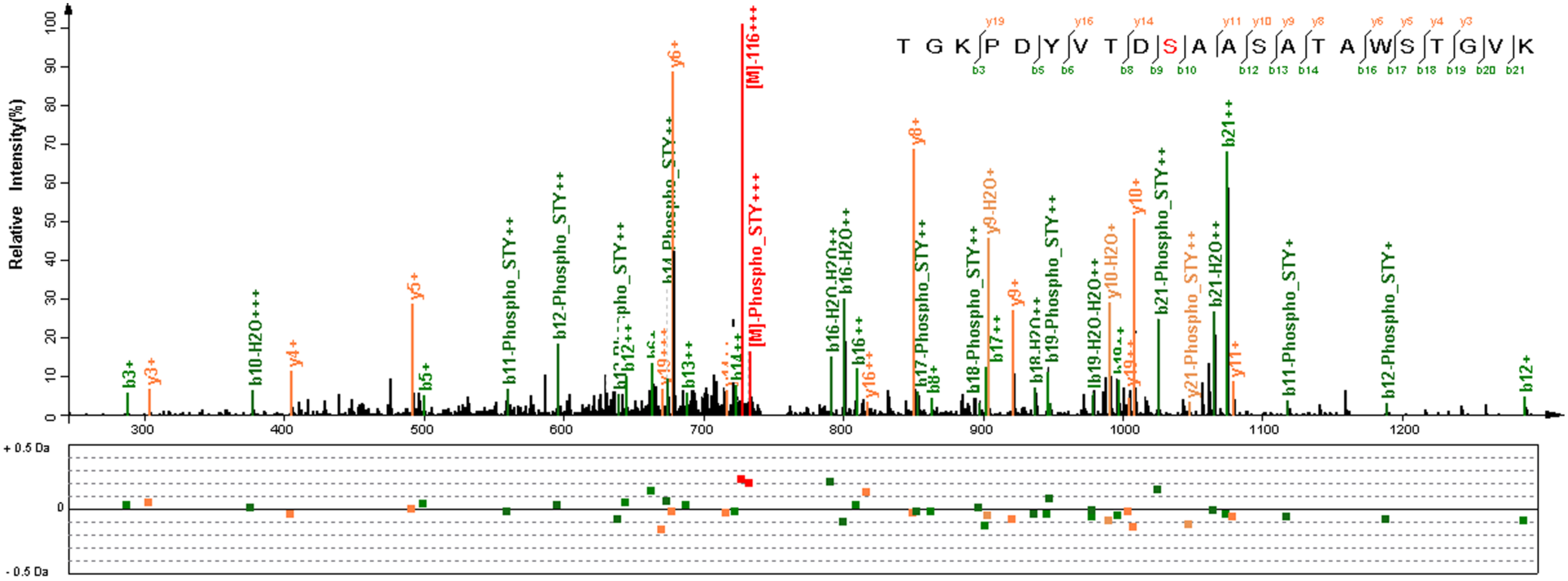
Detected mass shifts



A Disulfide Bridge



Phosphorylation



Results on five Datasets

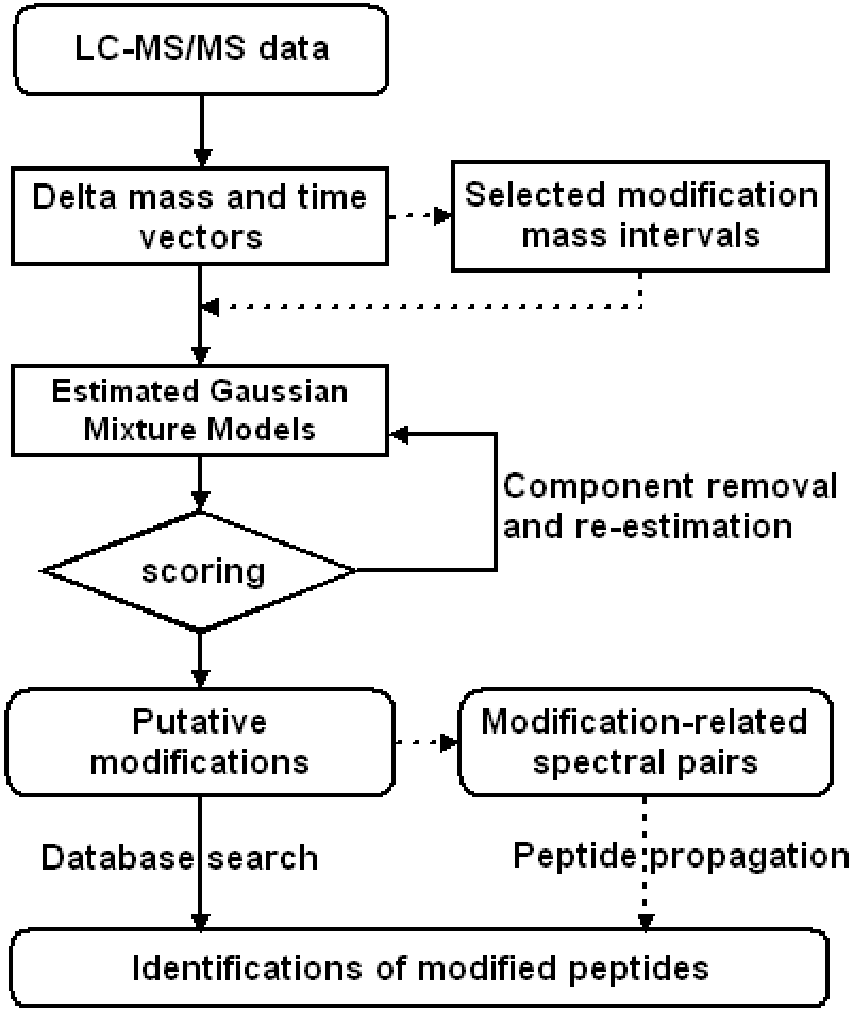
Dataset	Total MS/MS	Identified Spectra		Identification Rate Raised	Abundant Modifications (Da)
		pFind	pMatch		
ISB-18mix	40,376	12,032	+8,025	29.80% → **49.68%**	-116 (A disulfide bridge); -18 (Dehydration); -17 (Ammonia loss); -16 (Ammonia loss & Deamidation); 1 (Deamidation); 2 (Two deamidations); 16 (Oxidation); 22 (Sodium); 23 (Sodium & Deamidation); 26 (Acetaldehyde +26); 38 (Calcium); 39 (Calcium & Deamidation); 152 (CarbamidomethylDDT); 153 (CarbamidomethylDDT & Deamidatoin); 174 (CarbamidomethylDDT & Sodium)
TAP-PSD95	36,387	3,575	+1,882	9.82% → **15.00%**	-18 (Dehydration); -17 (Ammonia loss); 1 (Deamidation); 14 (Methylation); 16 (Oxidation); 22 (Sodium); 26 (Acetaldehyde +26); 28 (Formylation); 32 (Dioxidation); 42 (Acetylation); 54 (Acetaldehyde +26 & Formylation); 70 (Formylation & Acetylation); 80 (Phosphorylation)
HUPO-14	15,221	7,281	+2,418	47.84% → **63.72%**	-17 (Ammonia loss); 1 (Deamidation); 12 (Formaldehyde induced modification); 14 (Methylation); 26 (Acetaldehyde +26); 42 (Acetylation)
Haas-Data	56,599	9,172	+2,558	16.21% → **20.74%**	-17 (Ammonia loss); 1 (Deamidation); 43 (Carbamylation); 171 (Carbamylation & Lysine added)
Gygi-Qstar	46,195	9,255	+4,357	20.03% → **29.40%**	1 (Deamidation); 12 (Formaldehyde induced modification); 22 (Sodium); 28 (Formylation)

pCluster tool

Fast detection of abundant modifications by precursor clustering

pCluster: fast detection of abundant modifications by precursor clustering

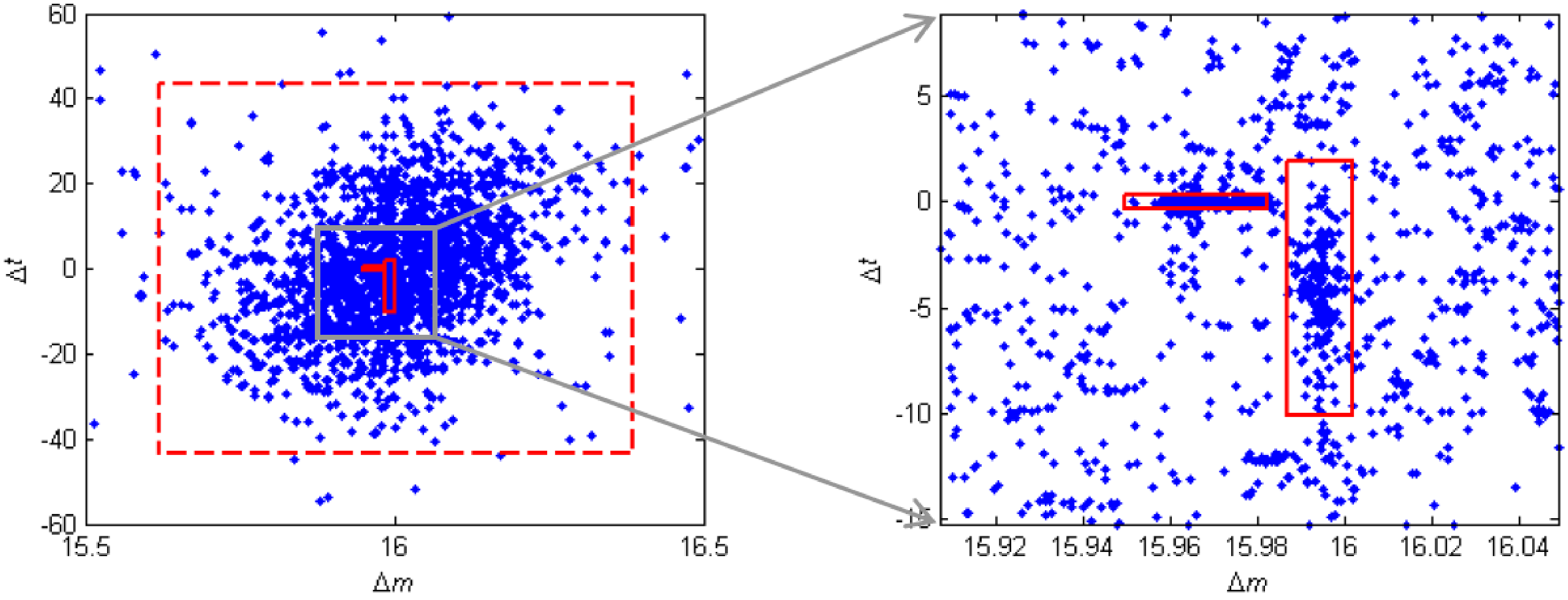
Fu Y, et al. BMC Bioinformatics 2009; Fu Y,et al. Submitted 2010


LC-MS/MS data
Delta mass and time vectors
Selected modification mass intervals
Estimated Gaussian Mixture Models
Component removal and re-estimation
scoring
Putative modifications
Modification-related spectral pairs
Database search
Peptide propagation
Identifications of modified peptides

Every pair of spectra is represented by a 2-d vector, called a *delta mass and time vector*

$$\Delta = \left\langle \Delta m, \Delta t \right\rangle$$

Modification-induced delta vector distributions



Two-dimensional Gaussian mixture distribution model

$$f\left(\Delta\right)=\alpha_{Rand}\,f_{Rand}\left(\Delta\right)+\sum_{k=1}^{n}\alpha_{Mod,k}\,f_{Mod,k}\left(\Delta\right)$$

$$\alpha_{Rand}+\sum_{k=1}^{n}\alpha_{Mod,k}=1$$

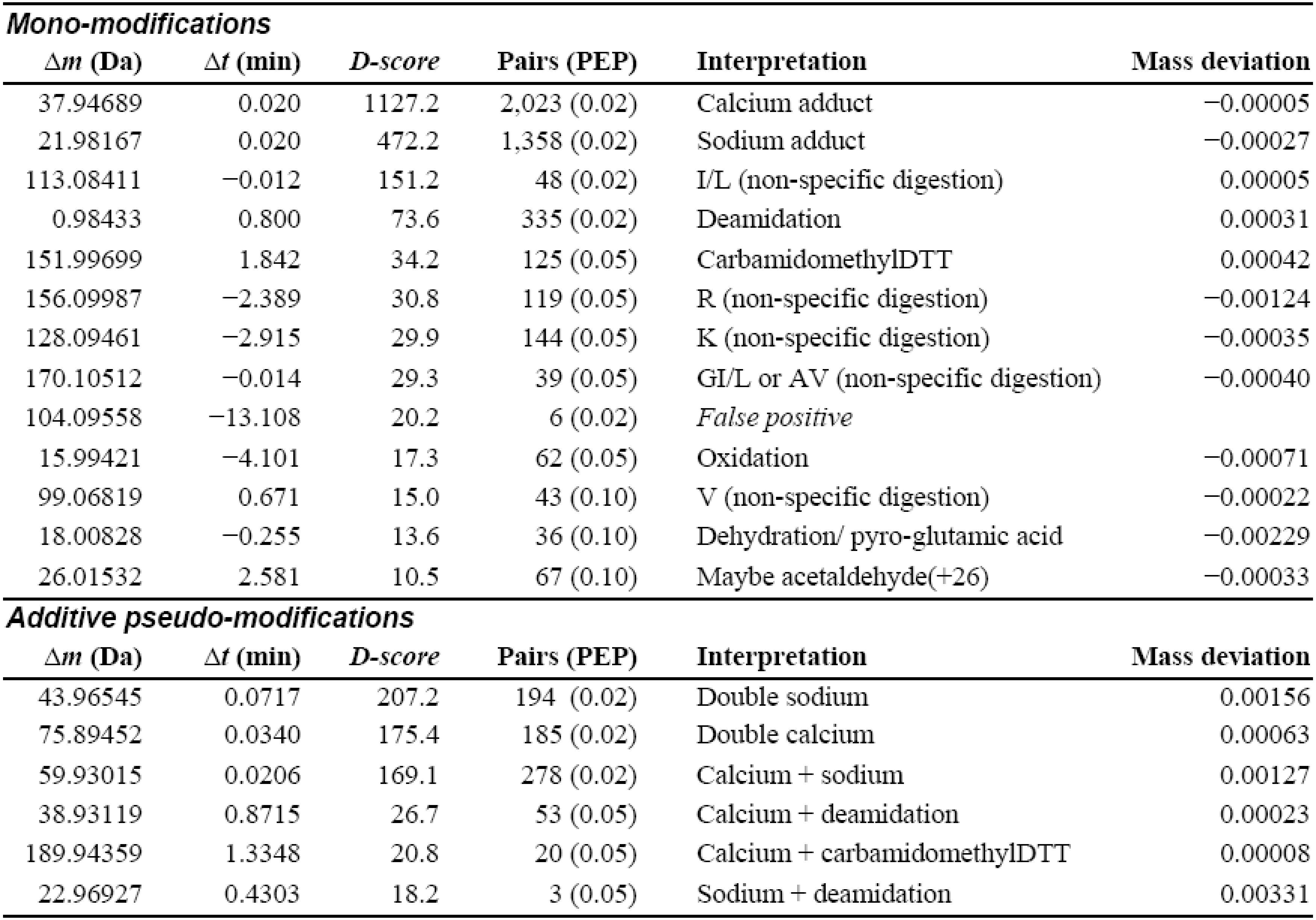
where $f_{Rand}(\Delta)$ represents the *pdf* of the random distribution in a mass interval, $f_{Mod,k}(\Delta)$ represents the *pdf* of the *k*-th modification-induced distribution, *n* is the total number of modifications in this mass interval, and α_{Rand} and $\alpha_{Mod,k}$ are mixing coefficients.

Results

- ISB standard protein mix data
  - Eighteen purified proteins
  - Digestion by trypsin
  - Mixture 3 on an LTQ-FT mass spectrometer
  - The third run selected
  - 4,085 MS/MS scans

Klimek, J., et al. (2008)The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. *J Proteome Res* 7, 96-103

Detected putative modifications



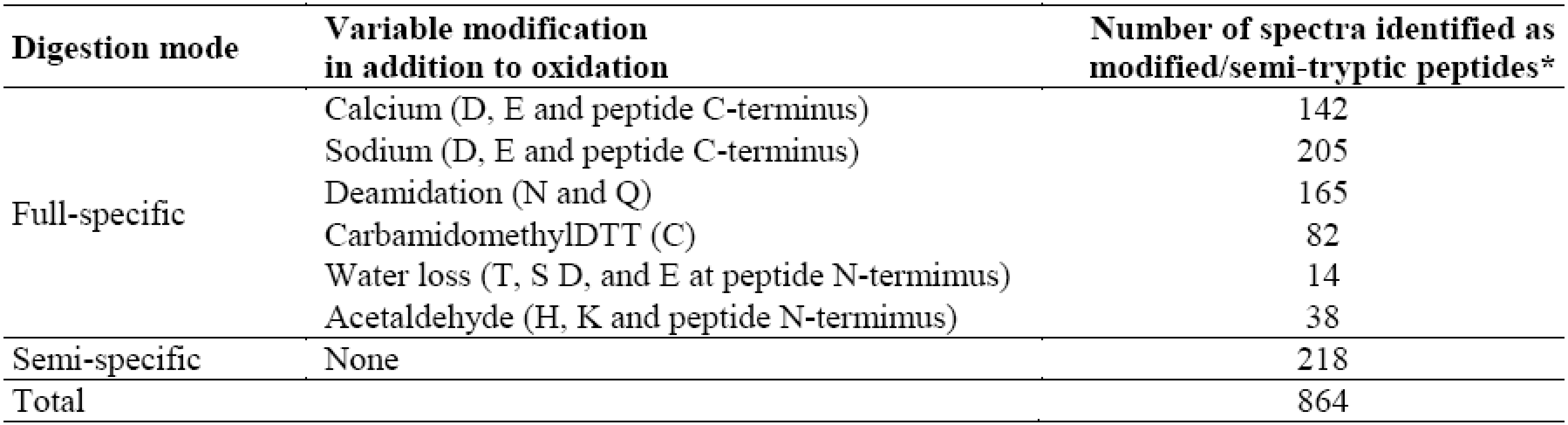
Mono-modifications

Δ*m* (Da)	Δ*t* (min)	*D-score*	Pairs (PEP)	Interpretation	Mass deviation
37.94689	0.020	1127.2	2,023 (0.02)	Calcium adduct	−0.00005
21.98167	0.020	472.2	1,358 (0.02)	Sodium adduct	−0.00027
113.08411	−0.012	151.2	48 (0.02)	I/L (non-specific digestion)	0.00005
0.98433	0.800	73.6	335 (0.02)	Deamidation	0.00031
151.99699	1.842	34.2	125 (0.05)	CarbamidomethylDTT	0.00042
156.09987	−2.389	30.8	119 (0.05)	R (non-specific digestion)	−0.00124
128.09461	−2.915	29.9	144 (0.05)	K (non-specific digestion)	−0.00035
170.10512	−0.014	29.3	39 (0.05)	GI/L or AV (non-specific digestion)	−0.00040
104.09558	−13.108	20.2	6 (0.02)	*False positive*	
15.99421	−4.101	17.3	62 (0.05)	Oxidation	−0.00071
99.06819	0.671	15.0	43 (0.10)	V (non-specific digestion)	−0.00022
18.00828	−0.255	13.6	36 (0.10)	Dehydration/ pyro-glutamic acid	−0.00229
26.01532	2.581	10.5	67 (0.10)	Maybe acetaldehyde(+26)	−0.00033

Additive pseudo-modifications

Δ*m* (Da)	Δ*t* (min)	*D-score*	Pairs (PEP)	Interpretation	Mass deviation
43.96545	0.0717	207.2	194 (0.02)	Double sodium	0.00156
75.89452	0.0340	175.4	185 (0.02)	Double calcium	0.00063
59.93015	0.0206	169.1	278 (0.02)	Calcium + sodium	0.00127
38.93119	0.8715	26.7	53 (0.05)	Calcium + deamidation	0.00023
189.94359	1.3348	20.8	20 (0.05)	Calcium + carbamidomethylDTT	0.00008
22.96927	0.4303	18.2	3 (0.05)	Sodium + deamidation	0.00331

Identified peptides with detected modifications

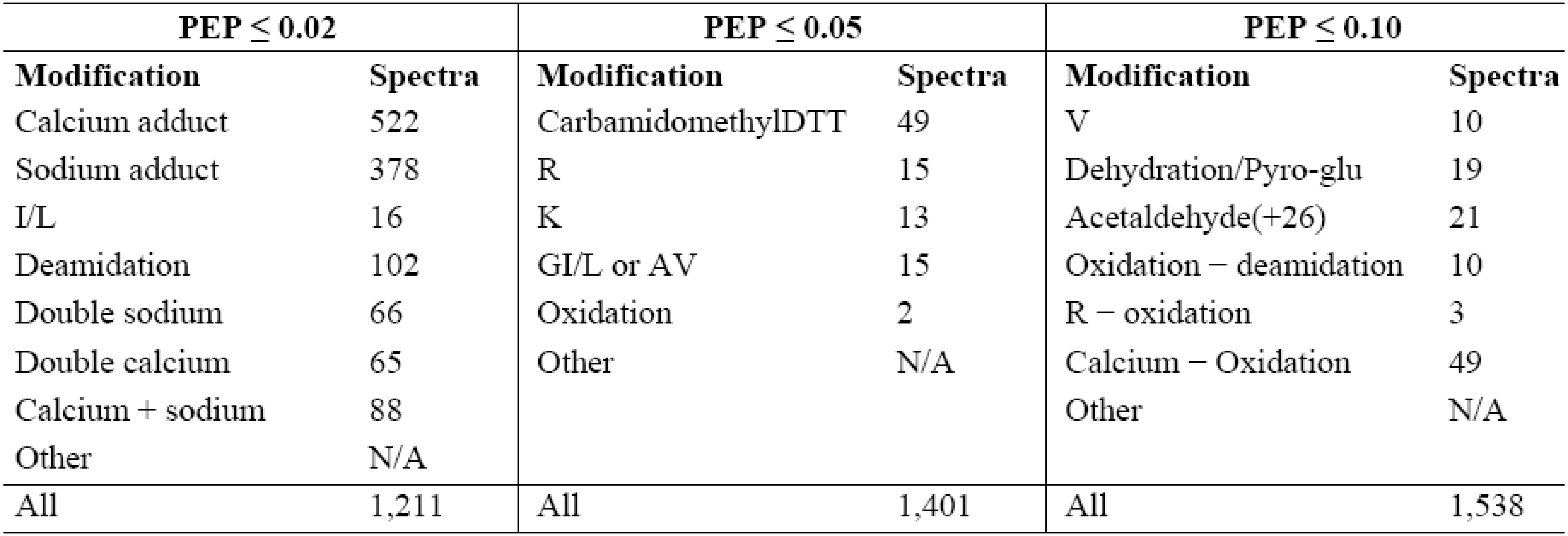


Digestion mode	Variable modification in addition to oxidation	Number of spectra identified as modified/semi-tryptic peptides*
Full-specific	Calcium (D, E and peptide C-terminus)	142
	Sodium (D, E and peptide C-terminus)	205
	Deamidation (N and Q)	165
	CarbamidomethylDTT (C)	82
	Water loss (T, S D, and E at peptide N-termimus)	14
	Acetaldehyde (H, K and peptide N-termimus)	38
Semi-specific	None	218
Total		864

Total spectra: 4085

Identified by initial search: 1032

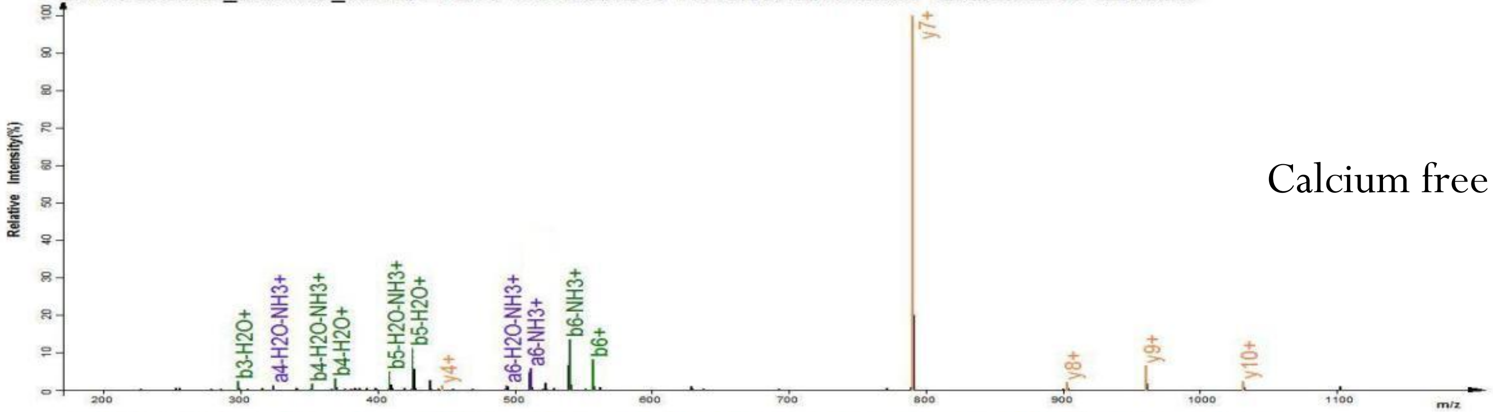
Peptide propagation among related spectral pairs



PEP ≤ 0.02		PEP ≤ 0.05		PEP ≤ 0.10	
Modification	**Spectra**	**Modification**	**Spectra**	**Modification**	**Spectra**
Calcium adduct	522	CarbamidomethylDTT	49	V	10
Sodium adduct	378	R	15	Dehydration/Pyro-glu	19
I/L	16	K	13	Acetaldehyde(+26)	21
Deamidation	102	GI/L or AV	15	Oxidation − deamidation	10
Double sodium	66	Oxidation	2	R − oxidation	3
Double calcium	65	Other	N/A	Calcium − Oxidation	49
Calcium + sodium	88			Other	N/A
Other	N/A				
All	1,211	All	1,401	All	1,538

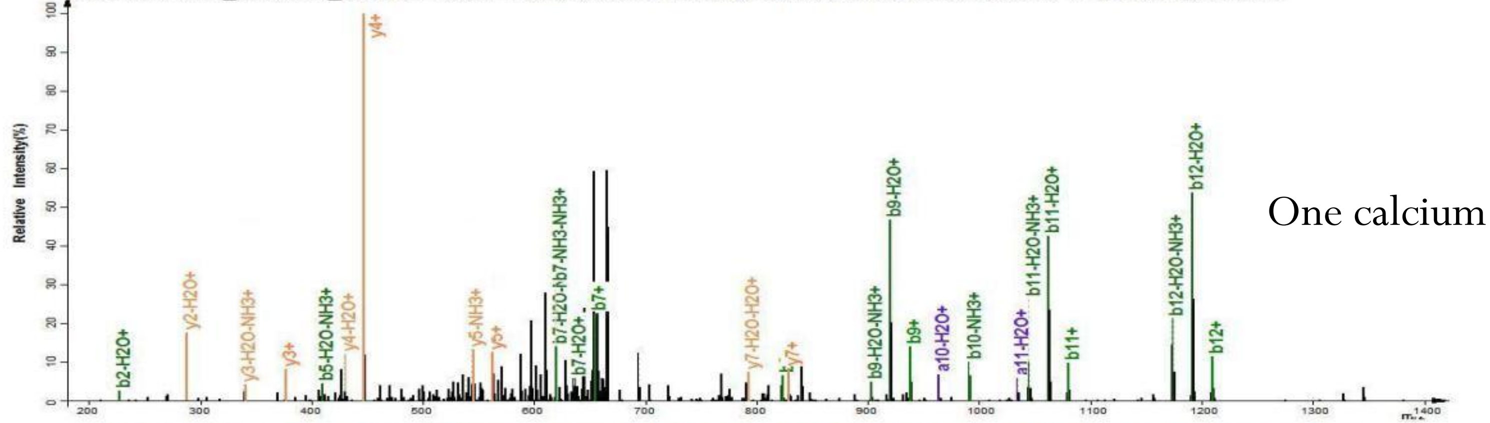
Examples of spectra of calcium adducts

E:\MSdata\ISB_data\Mix_3\LTQ-FT\B06-11073.dta\B06-11073.2634.2634.2.dta SEQ:ENAAGIPMDAAER



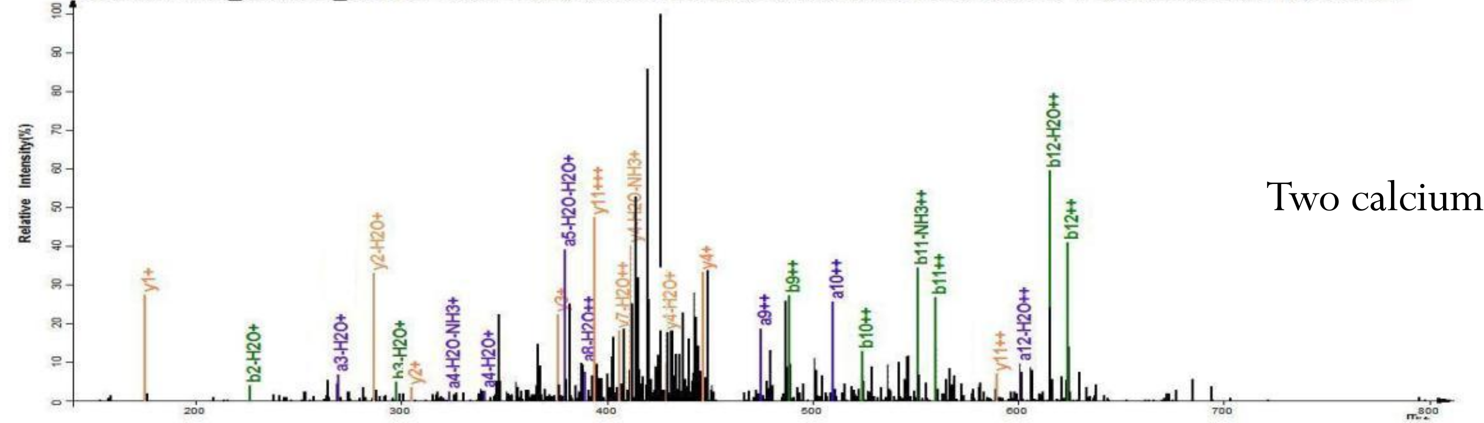
Calcium free

E:\MSdata\ISB_data\Mix_3\LTQ-FT\B06-11073.dta\B06-11073.2639.2639.2.dta SEQ:ENAAGIPMDAAER 8,Calcium



One calcium

E:\MSdata\ISB_data\Mix_3\LTQ-FT\B06-11073.dta\B06-11073.2650.2650.3.dta SEQ:ENAAGIPMDAAER 6,Calcium 9,Calcium



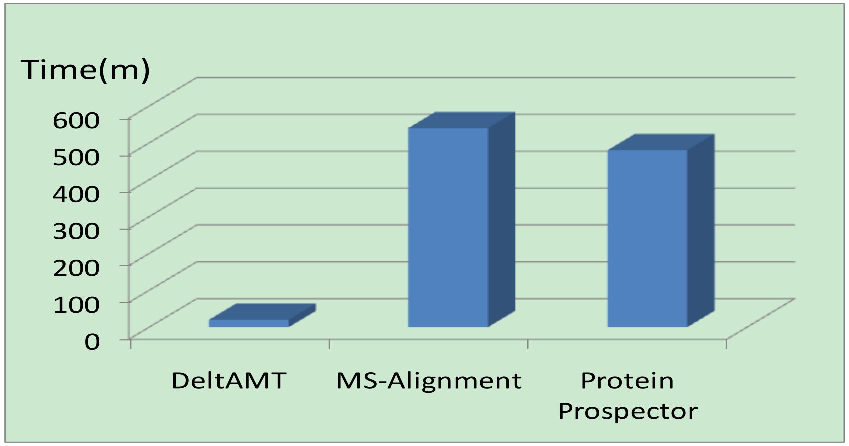
Two calcium

Running time: From reading the raw mass spectra to reporting potential modifications (mass and time shifts), the procedure took less than 4 min. The whole process of data analysis, including the restrictive database search and the peptide propagation, was completed within 20 min.

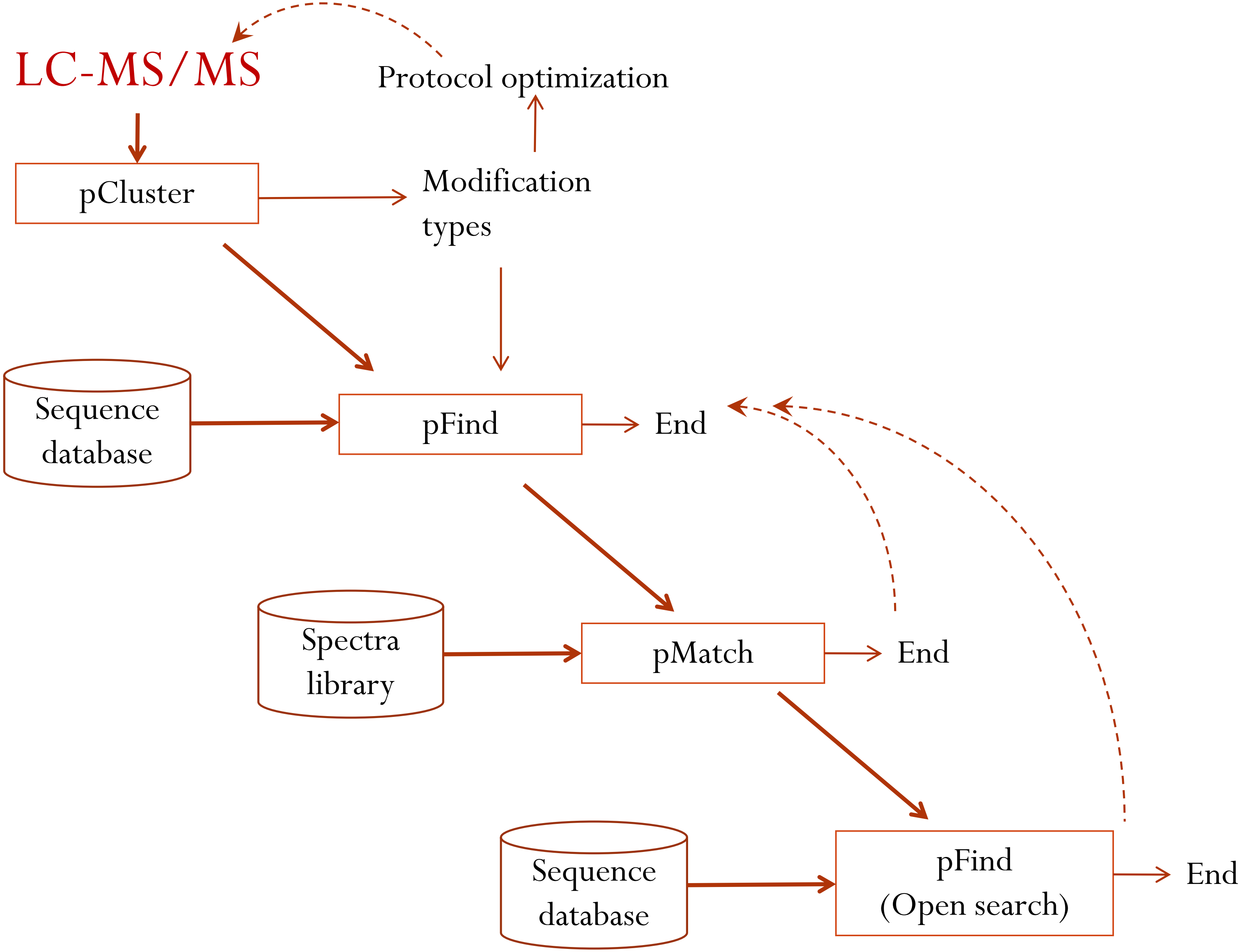
Modification detection				Basic Database search	Total
Reading raw data	Reporting modifications	Reporting spectral pairs	Propagation & location		
0.3 min	3.4 min	1.4 min	5 min	9.7 min	19.8 min

MS-Alignment, Protein Prospector and other fragment information based methods can detect low-abundance modifications, but they are generally very slow.

The precursor information based DeltAMT is a powerful complement to them.



For MS-Alignment, the same database (less than 6 M amino acids) was searched as was used for the basic search by pFind, and one unanticipated modification was allowed per peptide. The result for Protein prospector is from (Chalkley et al., 2008, MCP, pp2386-1398), in which a database containing 46 proteins was searched.



LC-MS/MS

Protocol optimization

pCluster

Modification types

Sequence database

pFind

End

Spectra library

pMatch

End

Sequence database

pFind (Open search)

End

Summary

- Deeper insights into mass spectral data
- Increasing spectral identification rates
- Experimental protocol optimization
- Discovery of biologically interesting or novel modifications
- Potentially useful for quantitation analysis

- ???

Software available

- pFind: http://pfind.ict.ac.cn
- pMatch: http://pfind.ict.ac.cn/pmatch
- pCluster: http://pfind.ict.ac.cn/pcluster

Acknowledgement

- Group members and collaborators

BPRC
SIBS
NIBS
……

- Support funds
  - NSFC(30900262),CAS(KGGX1-YW-13)
  - 973(2010CB912701),863(2007AA02Z315)

Thank you for your attention!

yfu@ict.ac.cn

http://pfind.ict.ac.cn