Assigning Confidence to Peptide Id

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Alexey Nesvizhskii

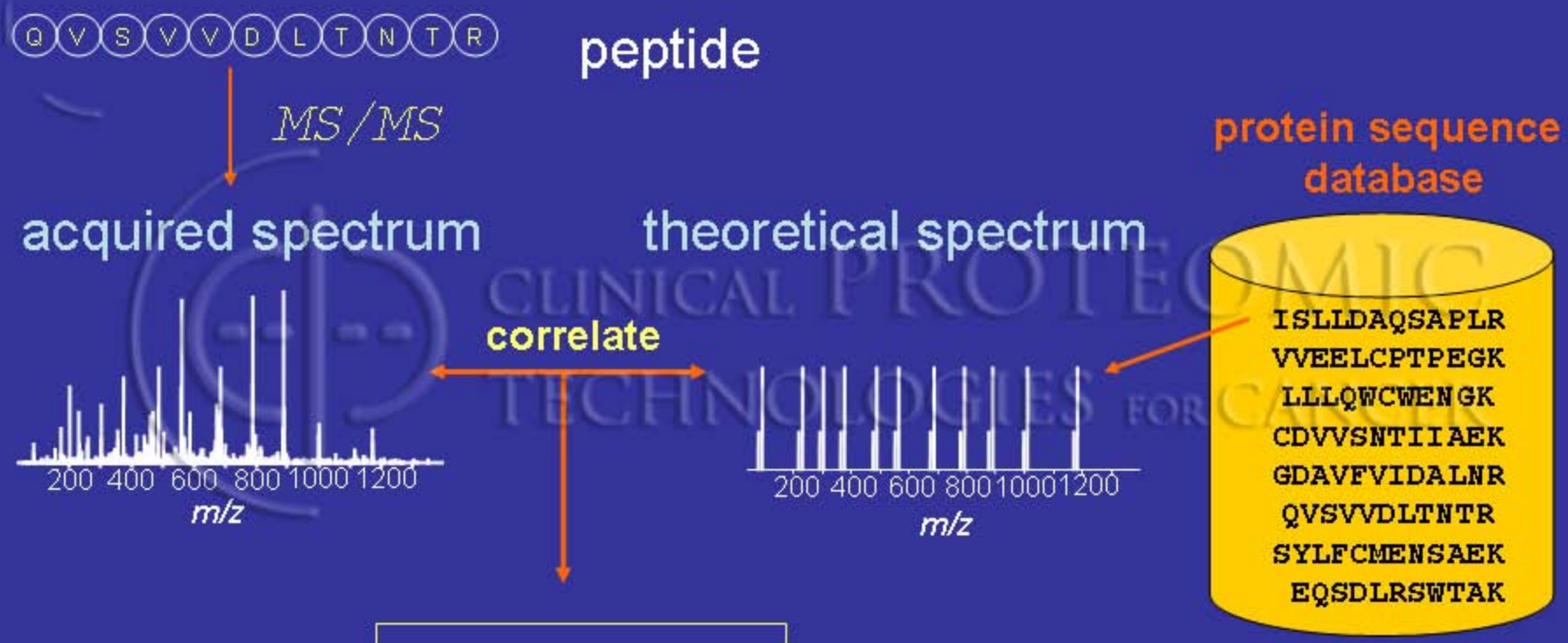
University of Michigan

Department of Pathology and
Center for Computational Medicine and Biology

nesvi@med.umich.edu

NIST
November 15 2007

Database Search-based Identification



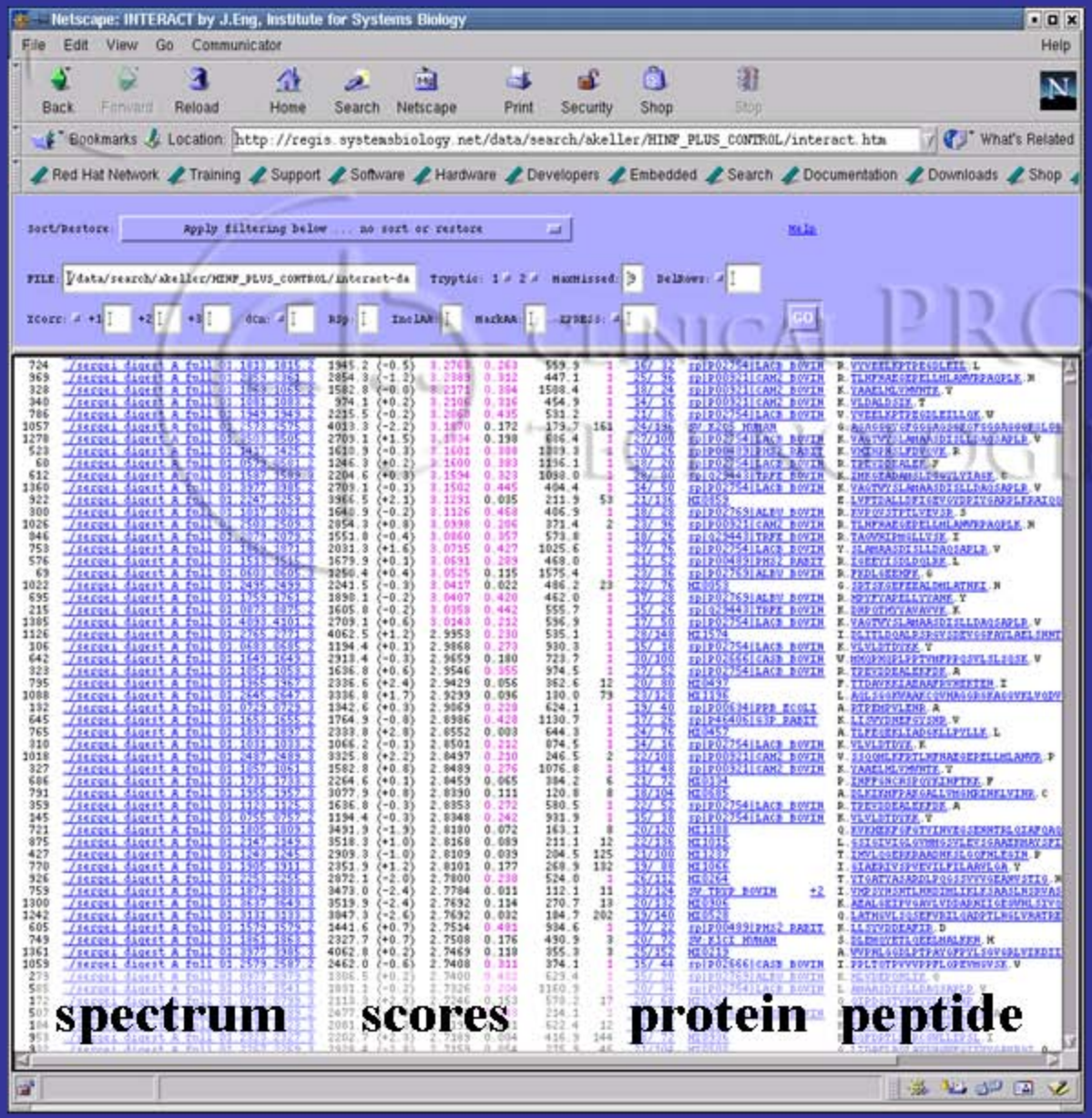
output:
ranked list of peptide matches

peptide	score
QVSVVDLTNTR	5.6
ISLLDAQSAPLR	1.3
CDVVSNTIIAEK	1.1
DLLLQWCWENGK	0.9
SYLFCMENSAEK	0.8
…	

best match
QVSVVDLTNTR

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Validation of Peptide Identifications

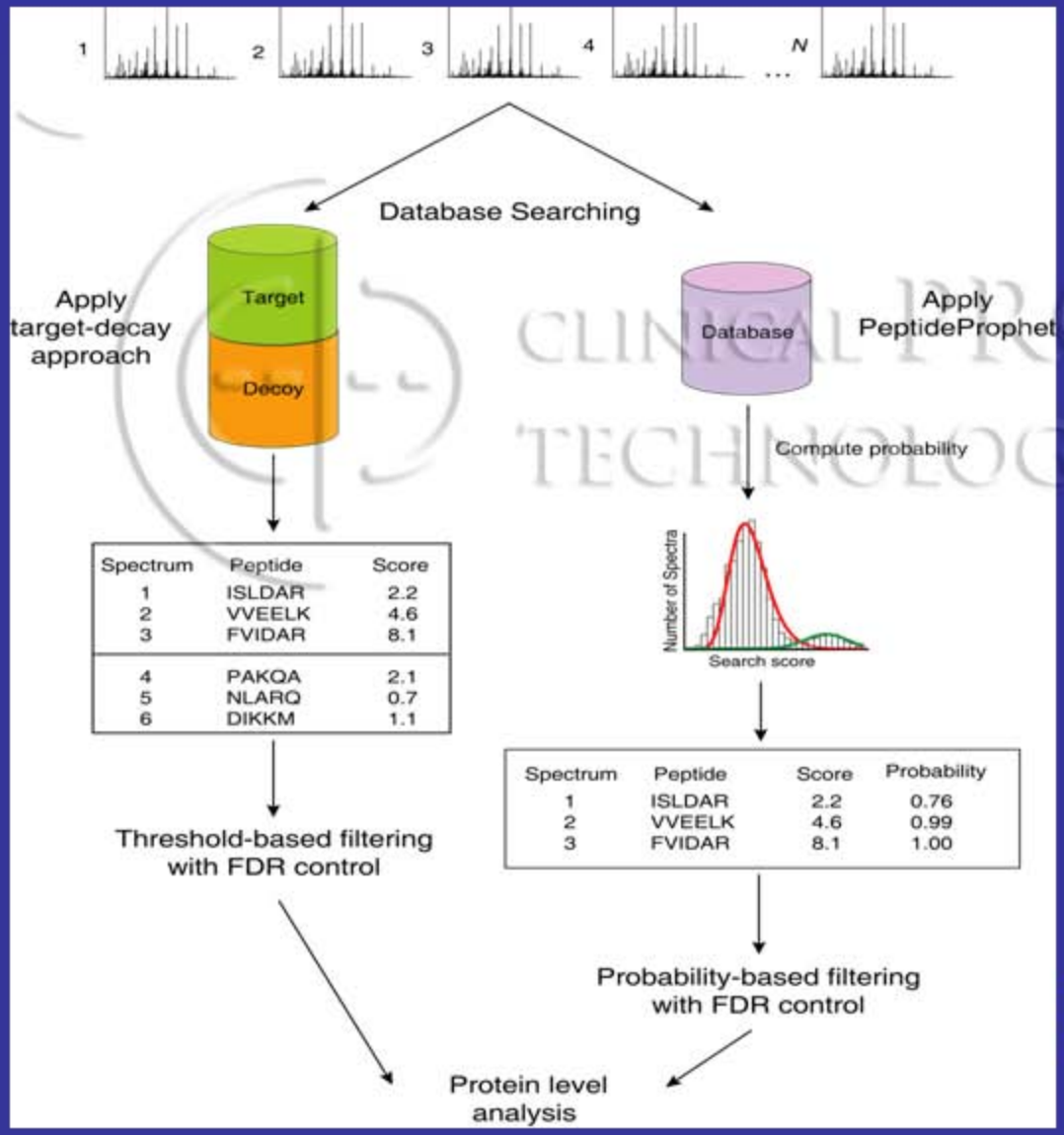


output from MS/MS database search algorithms

Instruments can generate thousands of spectra per hour!

- which peptide identification are correct?
- how to filter data?

Assessing Confidence in Peptide Identifications



Commonly used methods:

1. Expectation values
2. Target-decoy database searching
3. PeptideProphet

A.I. Nesvizhskii et al. Nature Methods 4, 787 (2007)

Statistical Validation of Peptide Identifications

- Single spectrum measures: p-values or expectation values

null hypothesis: random match

compute p-value, reject null if p-value is small

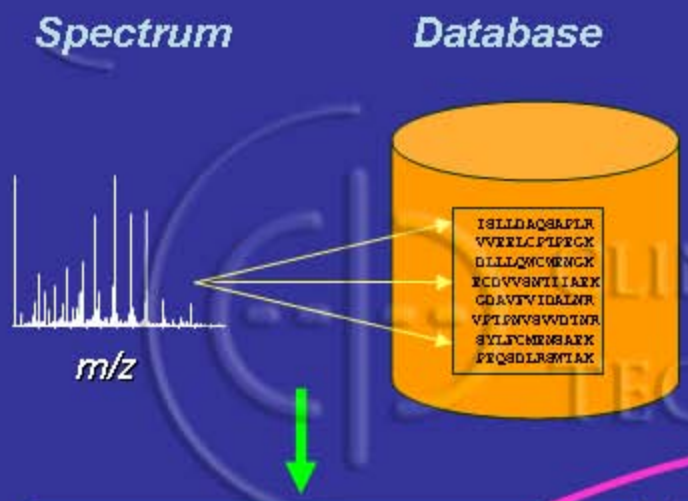
example: sequence similarity searching (BLAST)

- Global Methods

false discovery rates (FDR) as well as individual posterior peptide probabilities (complement of local fdr)

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Expectation Values (empirical model)



$$p(s_m) = \sum_{s \geq s_m} P(s)$$

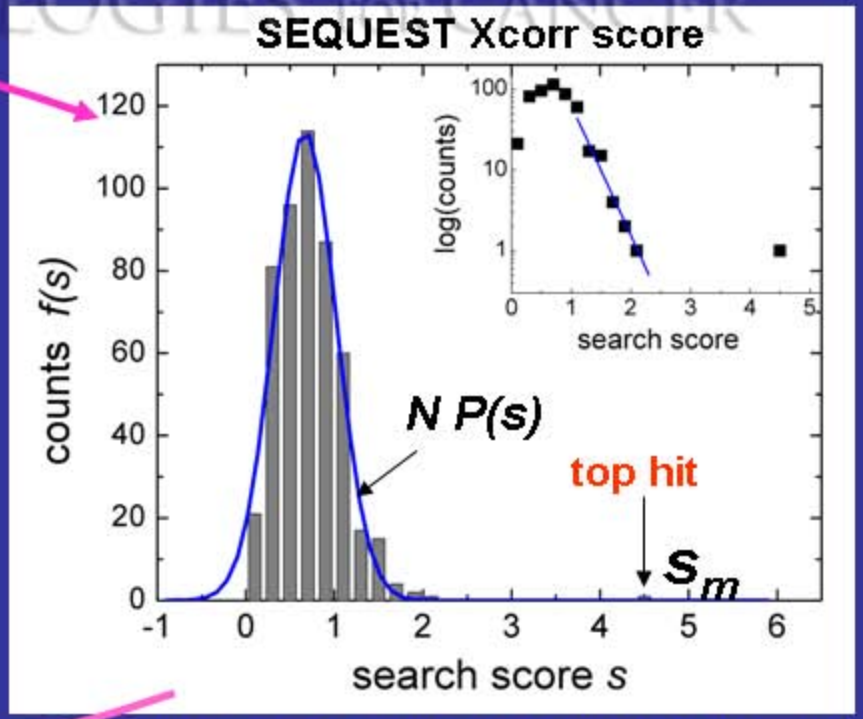
probability to get score $s \geq s_m$ by chance

$$E(s_m) = N \sum_{s \geq s_m} P(s)$$

expected number of random matches with $s \geq s_m$

Rank	Peptide	Score	p
1	ISLLDAQSAPLR	4.5	10^{-5}
2	VVEELCTPEGK	2.1	0.11
3	DLLLQWCWENGK	2.0	0.23
4	ECDVVSNTIIAEK	1.9	0.52
5	GDAVFVIDALNR	1.7	0.72
6	VPTPNVSVVTNR	1.6	0.86
7	SYLFCMEAEK	1.6	0.86
8	PEQSDLRSWTAK	1.5	0.94
	...		...

N peptides



Fenyo & Beavis *Anal. Chem.* (2003)

Expectation Values (or p-values): Limitations

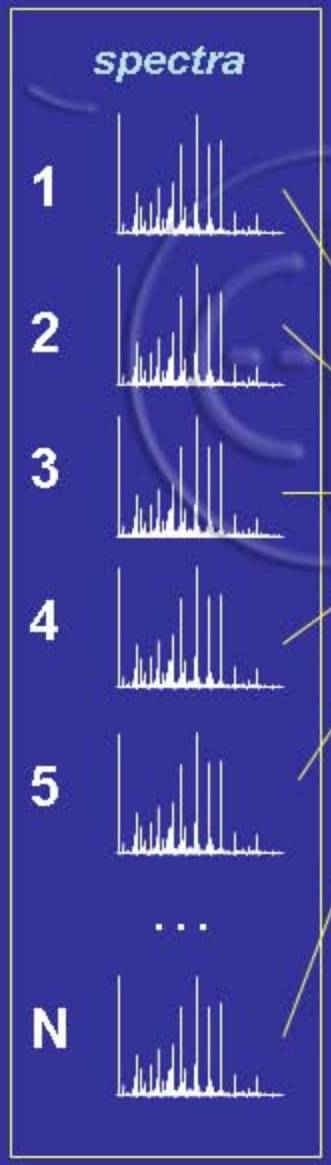
1. *p*-values or *E*-values are not well suited for the analysis of large-scale datasets (do not allow estimation of error rates as a function of filtering threshold)

2. Do not reflect additional useful information (e.g., number of missed cleavages, peptide retention time, mass accuracy)

3. Need to compute protein probabilities by combining probabilities of peptides corresponding to the same protein. Whether peptide expectation values can be used for that purpose is not clear

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Target-Decoy Database Approach

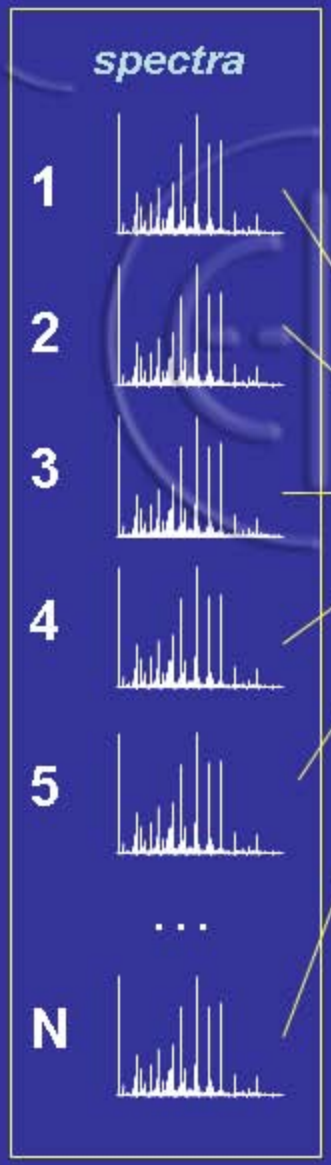


entire dataset, M spectra
(1 or more LC/MS/MS runs)

Spectrum	Peptide	Protein	Score
1	ISLLDAQSAPLR	N	4.5
2	VVEELCTPEGK	N	3.9
3	DLLLQWCWENGK	Rev	1.2
4	ECDVVSNTIIAEK	N	0.9
5	GDAVFVIDALNR	Rev	0.8
	...		
N	SYLFCMEAEK	Rev	0.5

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Target-Decoy Database Approach



entire dataset, M spectra
(1 or more LC/MS/MS runs)

Spectrum	Peptide	Protein	Score
1	ISLLDAQSAPLR	N	4.5
2	VVEELCTPEGK	N	3.9
3	DLLLQWCWENGK	Rev	1.2
4	ECDVVSNTIIAEK	N	0.9
5	GDAVFVIDALNR	Rev	0.8
	...		
N	SYLFCMEAEK	Rev	0.5

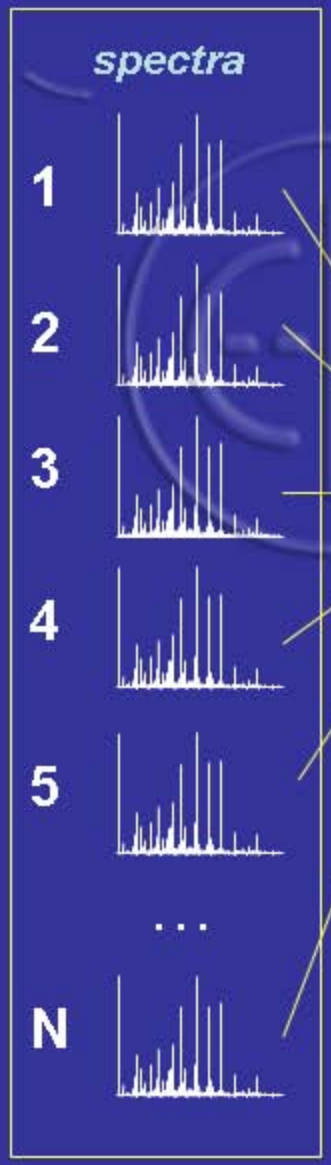
select score threshold s_T
calculate $N(s_T)$ and $N_{rev}(s_T)$

estimated false discovery rate

$$Err(s_T) = \frac{2 \cdot N_{rev}(s_T)}{N(s_T)}$$

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Mixture Modeling Approach



entire dataset, M spectra
(1 or more LC/MS/MS runs)

Spectrum	Peptide	Score
1	ISLLDAQSAPLR	4.5
2	VVEELCTPEGK	3.9
3	DLLLQWCWENGK	1.2
4	ECDVVSNTIIAEK	0.9
5	GDAVFVIDALNR	3.6
	...	
N	SYLFCMEAEK	1.1

best match
to each spectrum

raw score
E value, p-value

A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold
Anal. Chem. 74, 5383 (2002)

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Mixture Modeling Approach

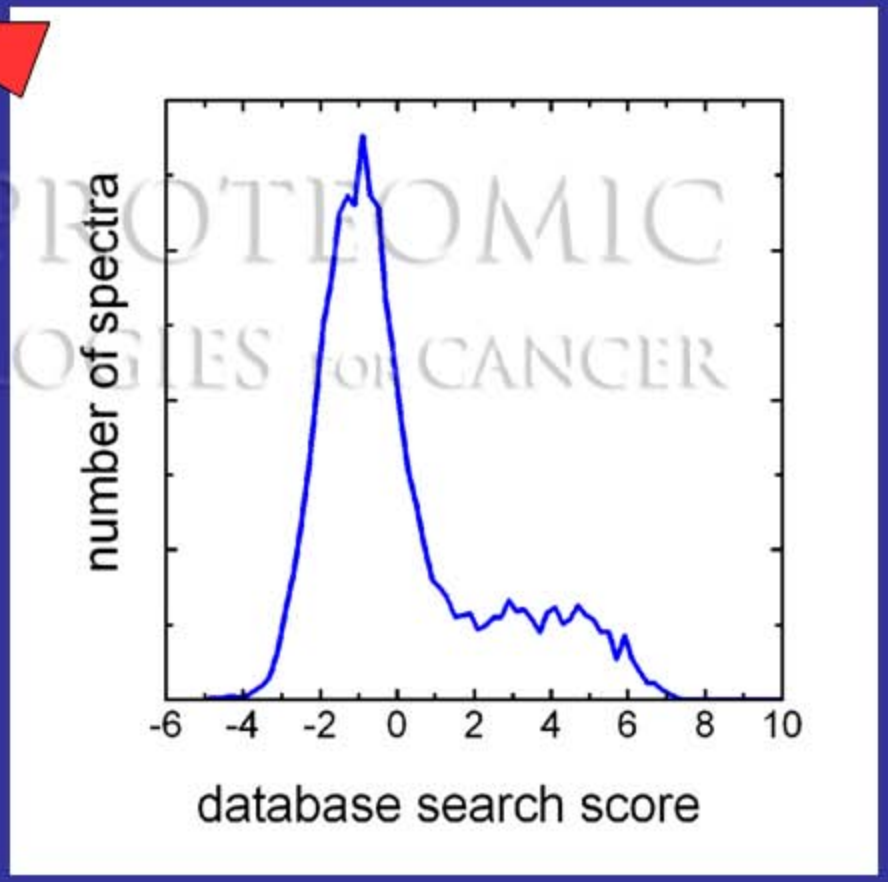
entire dataset:

Spectrum	Peptide	Score
1	ISLLDAQSAPLR	4.5
2	VVEELCTPEGK	3.9
3	DLLLQWCWENGK	1.2
4	ECDVVSNTIIAEK	0.9
5	GDAVFVIDALNR	3.6
	…	
M	SYLFCMEAEK	1.1

spectrum

best match

score

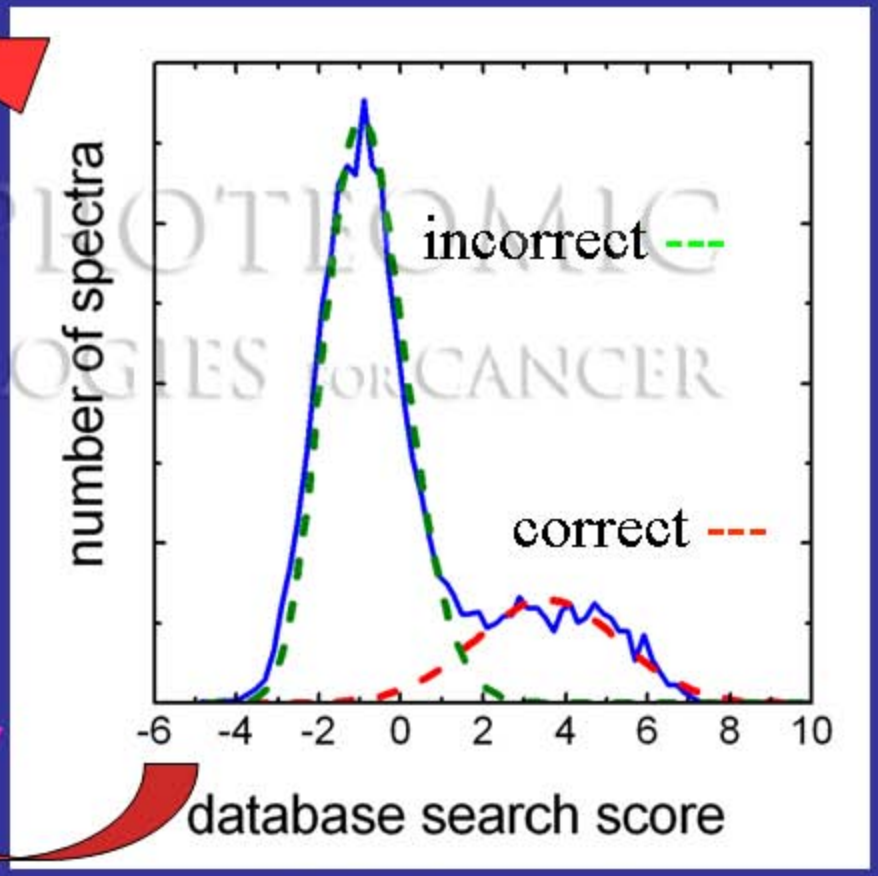


CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Computing Peptide Probabilities

entire dataset:

Spectrum	Peptide	Score	probability
1	ISLLDAQSAPLR	4.5	1.00
2	VVEELCTPEGK	3.9	0.99
3	DLLLQWCWENGK	1.2	0.11
4	ECDVVSNTIIAEK	0.9	0.00
5	GDAVFVIDALNR	3.6	0.87
	…		…
M	SYLFCMEAEK	1.1	0.02

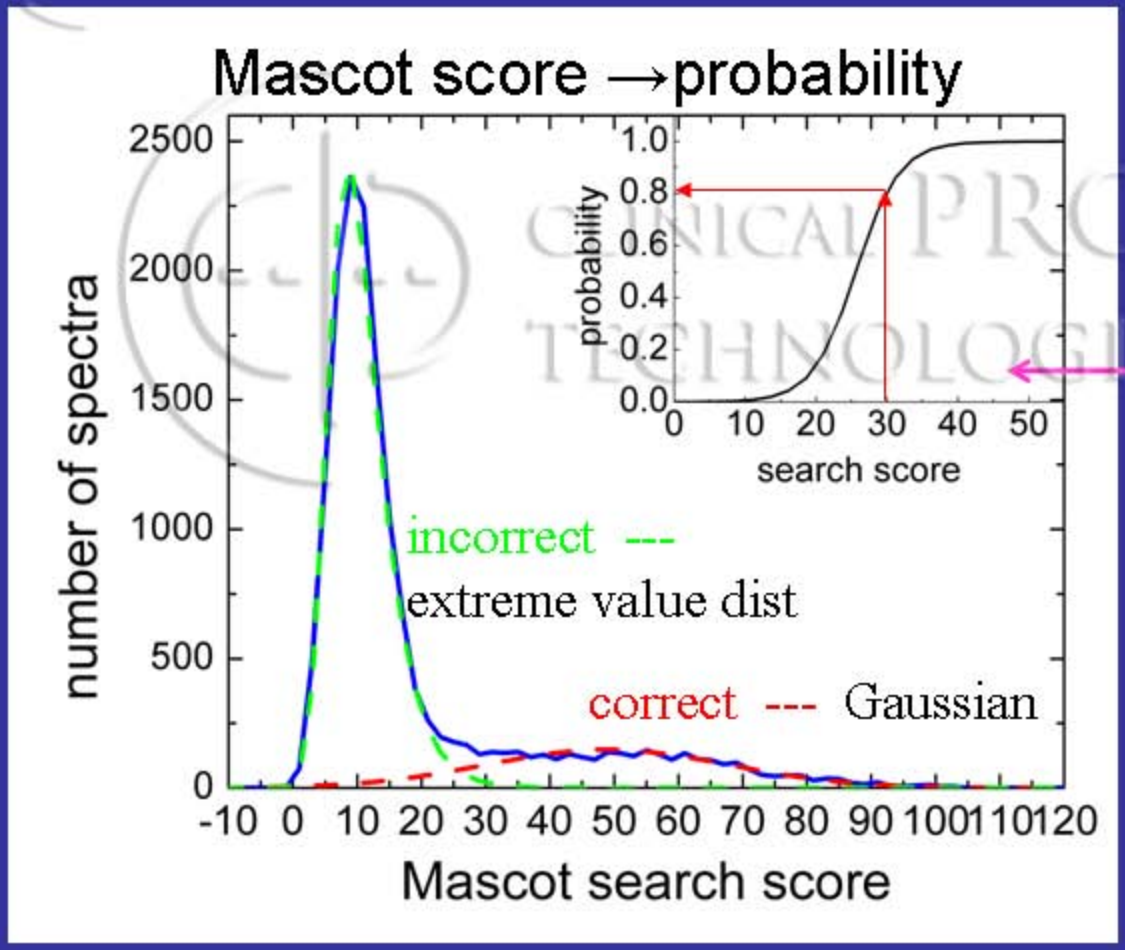


'unsupervised clustering'

EM mixture model algorithm learns the most likely distributions among correct and incorrect peptide assignments given the observed data

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Illustration: Assigning Probabilities to Mascot Search Results

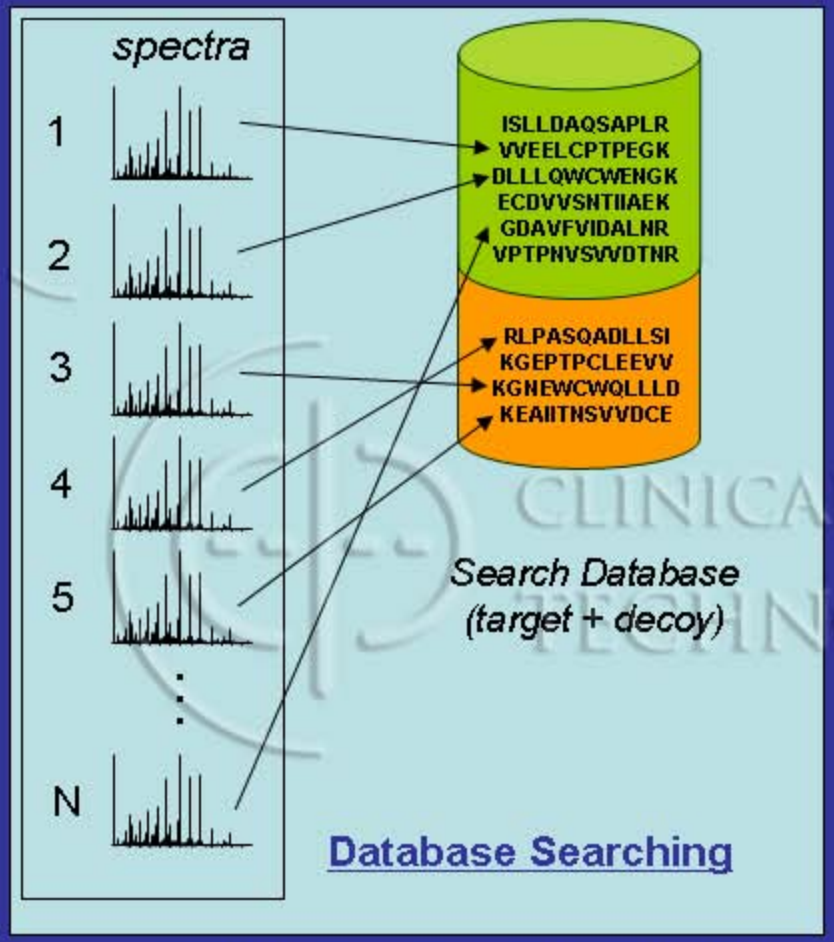


distributions are learned from the data

conversion of database Search scores into probabilities

Distribution parameters are learned from the data for each analyzed dataset anew using the EM mixture model algorithm

H. Influenzae, membrane fraction, 15 LC/MS/MS runs (~30,000 spectra)



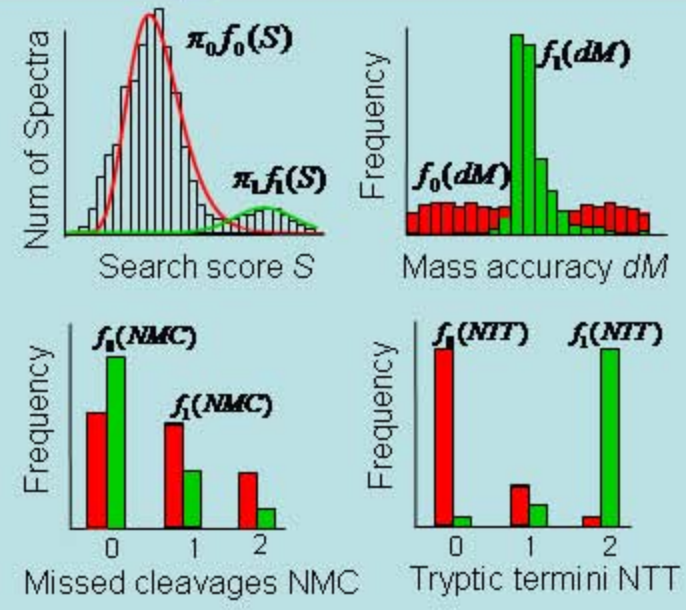
	Spec	S	NTT	NMC	dM	p_+
target	1	2.2	2	0	0.1	0.76
	2	4.6	2	1	0.2	0.94
	4	8.1	2	0	-0.1	0.99
	5	4.3	1	0	0.1	0.87
	7	3.7	2	0	-1.1	0.83
			...			...
decoys	3	1.2	0	2	0.7	0(0.51)
	6	1.1	1	1	-2.1	0(0.43)
	5	4.3	0	0	1.6	0(0.13)
	8	3.7	0	2	1.2	0(0.03)
			...			...

Probability filtering with FDR control

Protein level analysis

$$p_+ = p(+ \mid S, dM, NTT, NMC) = \frac{\pi_1 f_1(S, dM, NTT, NMC)}{\pi_0 f_0(S, dM, NTT, NMC) + \pi_1 f_1(S, dM, NTT, NMC)}$$

Posterior Probability



Mixture Fitting

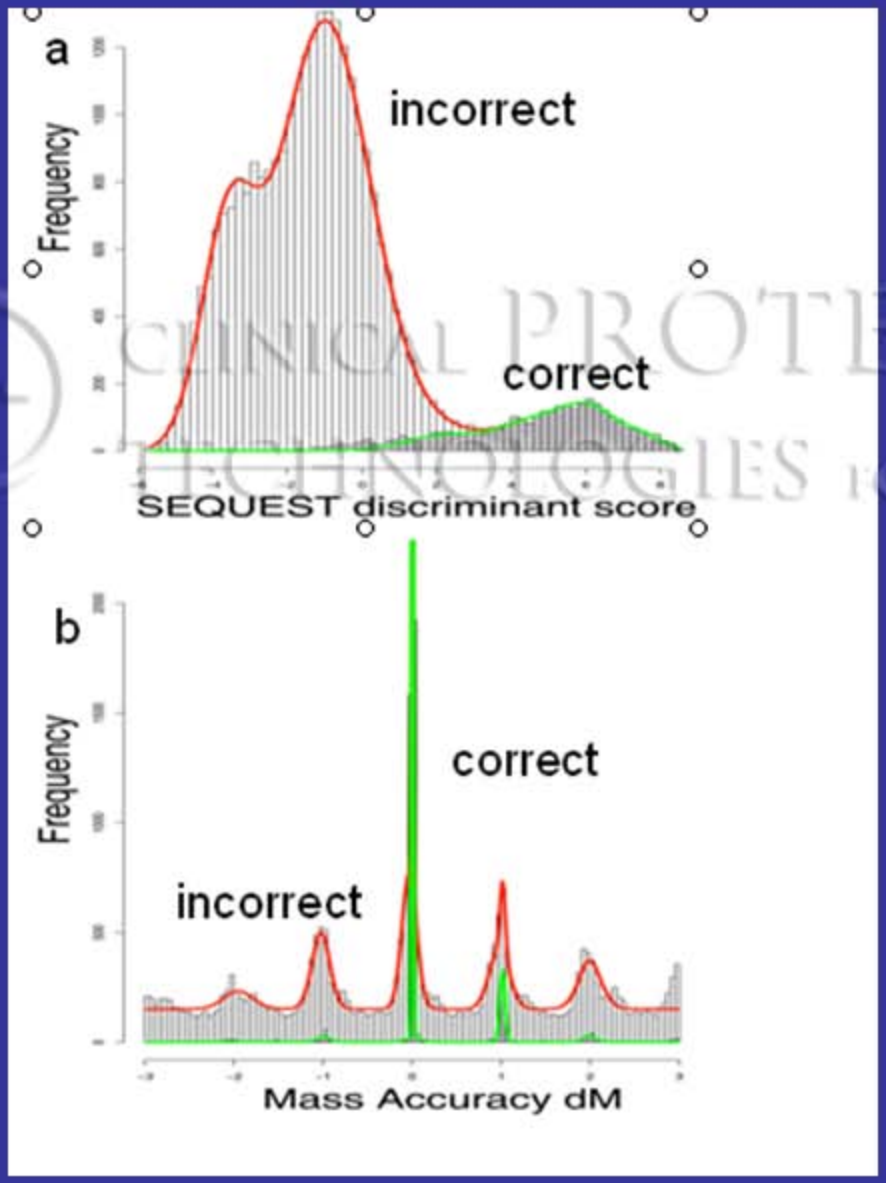
$$\sum_{i \in D} \log f_0(S, dM, NTT, NMC) + \sum_{i \in T} \log \{\pi_0 f_0(S, dM, NTT, NMC) + \pi_1 f_1(S, dM, NTT, NMC)\}$$

where $f_k(S, dM, NTT, NMC) = f_k(S) f_k(dM) f_k(NTT) f_k(NMC)$

Likelihood Decomposition

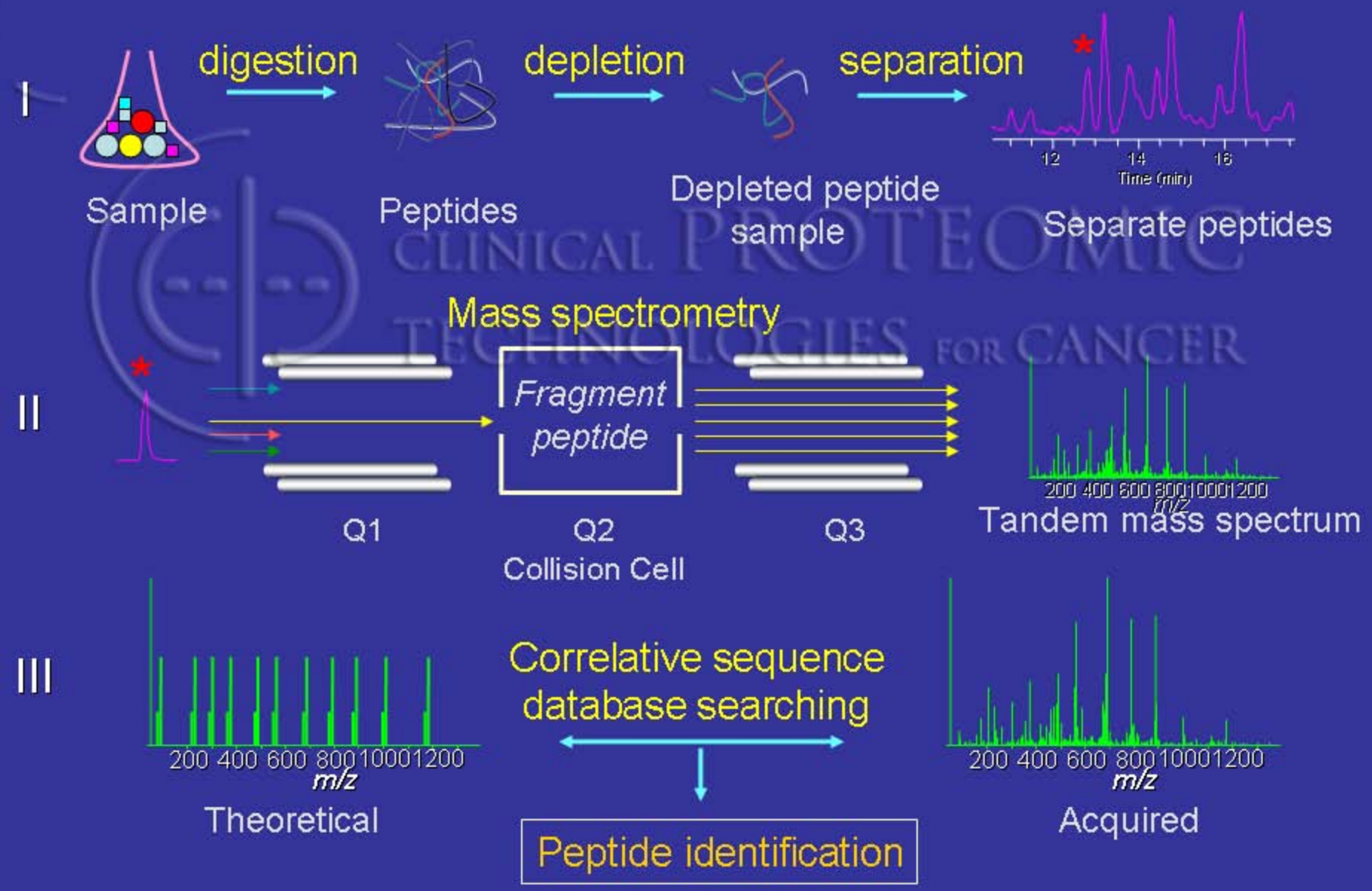
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Flexible Mixture Models



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Protein Identification by MS/MS



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Protein Identification by MS/MS

digestion

num of tryptic peptides
NTT (0,1,2)
num of missed cleavages
NMC (0,1,2,…)

depletion

glycopeptide?
glyco (0,1)
cys-containing?
cys (0,1)

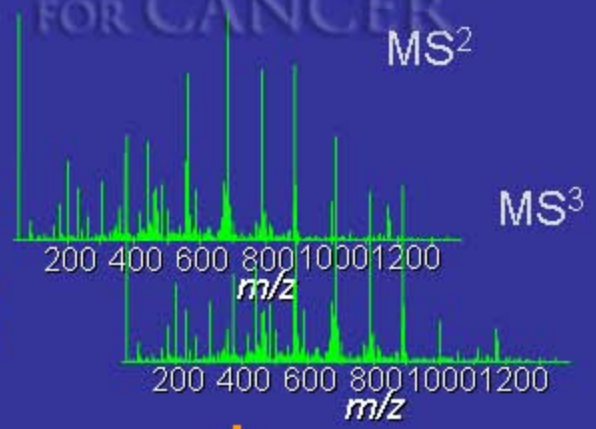
separation

retention time *r*
pI value (FFE)
SCX fraction number

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Mass spectrometry
(MS, MS^2, MS^3)

precursor *m/z*
charge state *z*



probability assignment is correct

Database searching

search scores:
Ion score (MASCOT)
Xcorr (SEQUEST)
Expect (TANDEM)

Spectrum Library searching

K.TMSKLQR.Y

peptide sequence

Auxiliary Discriminant Information

database search scores

+

peptide sequence properties:

- number of tryptic termini
- mass accuracy
- number of missed cleavages
- presence of a specific a.a. (C if ICAT)
- presence of a specific sequence motif (NxS/T)

peptide separation coordinates:

- pI value (isoelectric focusing)
- elution time (reversed-phase LC)

multiple search tools

multiple stages of mass spectrometry (MS3+MS2)

CENTER FOR PROTEOMIC TECHNOLOGIES FOR CANCER

Using Peptide Separation Coordinates

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

database search scores

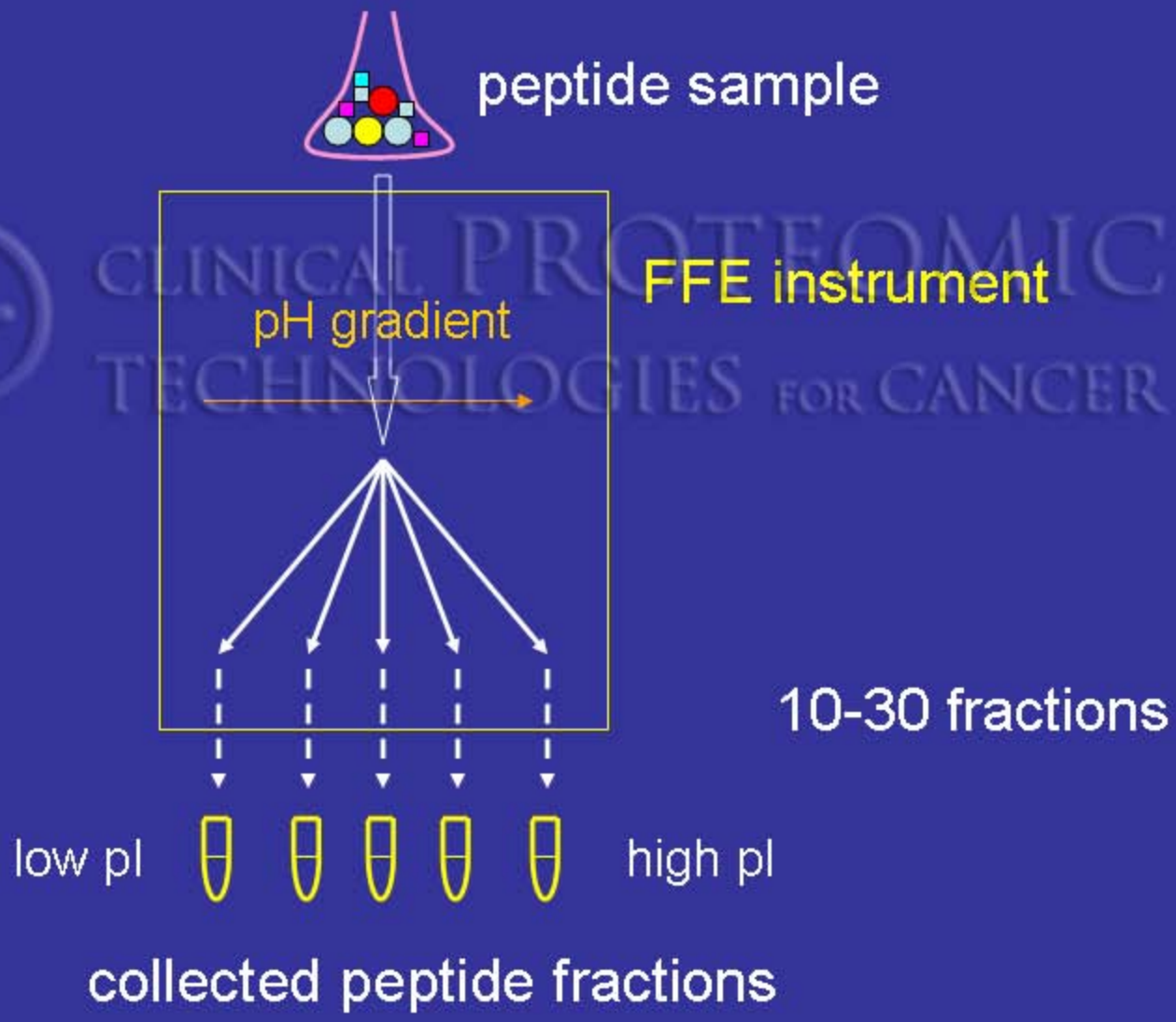
+

peptide separation coordinates:

- pI value (isoelectric focusing)
- elution time (reversed-phase LC)

J. Malmstroem et al. *J. Proteome Research* (2006)

Peptide Isoelectric Focusing

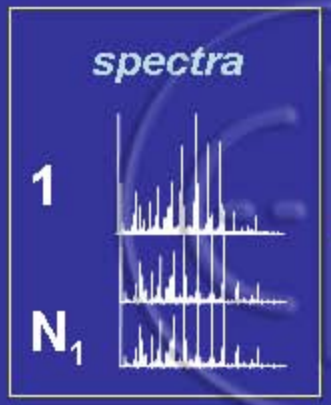


CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

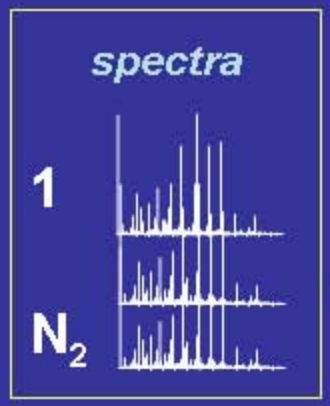
Peptide Identification

2 fractions -> 2 LC-MS/MS runs

pI fraction 1



Spectrum	Peptide	Score	calc pI
1	ISLLDAQSAPLR	4.4	4.5
2	VVEELCTPEGK	3.9	4.7
3	DLLLQWCWENGK	1.2	9.2
4	ECDVVSNTIIAEK	0.9	3.5
…			
N_1	FMMWEDCMEAEK	0.7	7.3

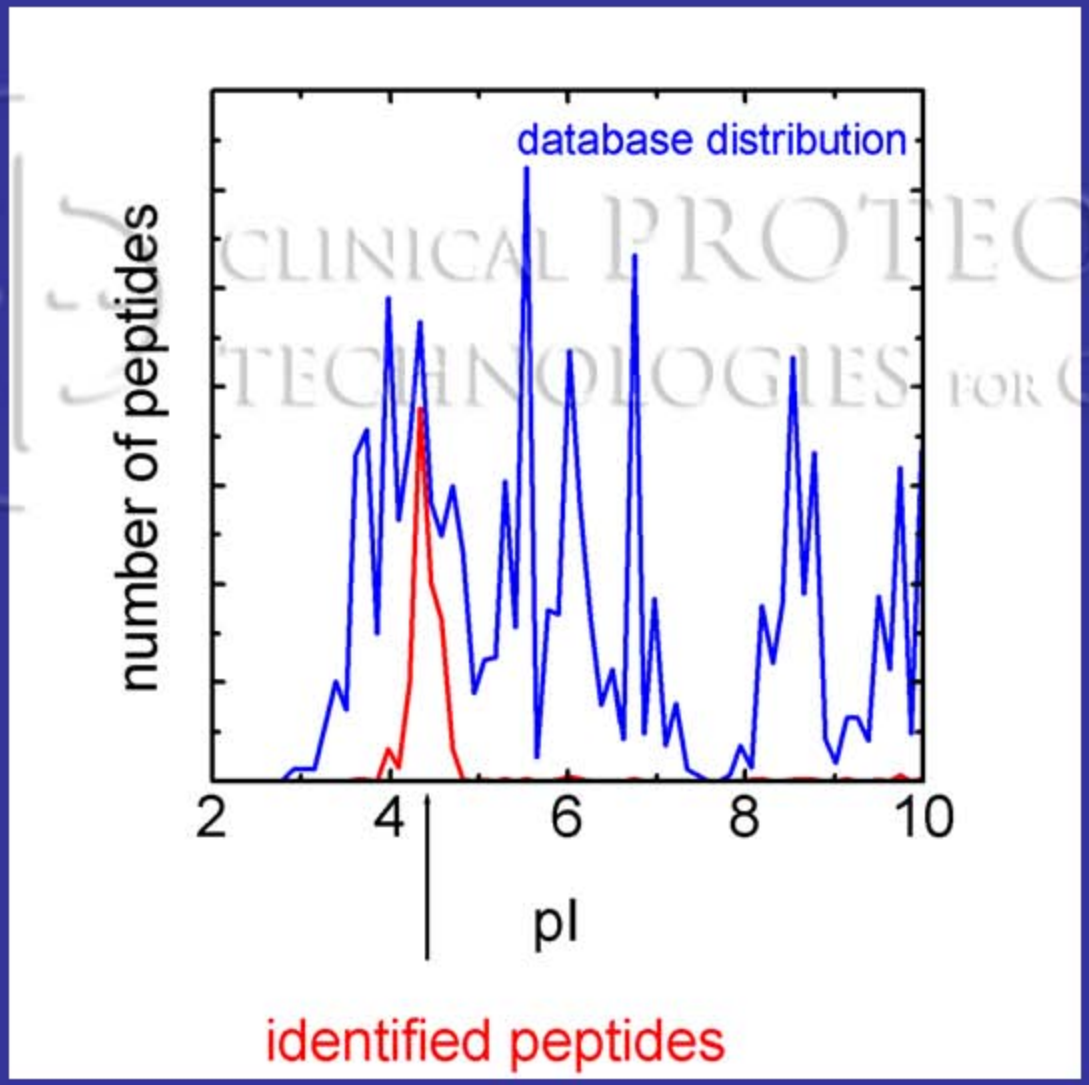


pI fraction 2

Spectrum	Peptide	Score	calc pI
1	LDAQSPWRAPTLR	4.2	6.5
2	CTPEVVLLGYK	2.9	6.7
3	GKMMDLLCWENGK	1.1	5.2
4	PLYSSECDVVSNTR	0.9	3.6
…			
N_2	SYLFCMEAEK	0.6	8.3

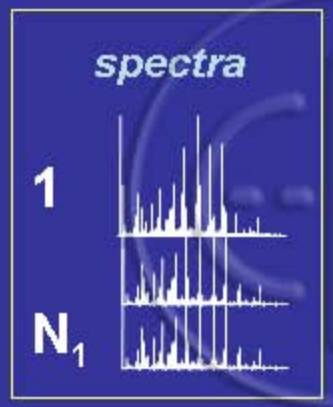
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Distribution of pI values among correct and incorrect peptides



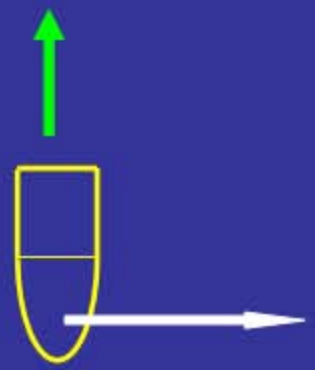
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

How to add pI Information?



Spectrum	Peptide	Score	calc pI
1	ISLLDAQSAPLR	4.4	4.7
2	VVEELCTPEGK	3.9	4.1
3	DLLLQWCWENGK	1.2	9.2
4	ECDVVSNTIIAEK	0.9	3.5
N_1	FMMWEDCMEAEK	0.7	7.3

measured pH=4.5



commonly used approach:

experimentally measure pH

calculate:

delta pI= pI – expected pI (based on pH)

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Data-dependent statistical model for incorporating pI Information

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

No need to measure pH

We can learn from the data what pI values we should expect for each fraction

Joint probability of search score F and pI

database search scores (search score F) only:

$$p_+ = p(+ \mid F) = \frac{p(F \mid +)\,p(+)}{p(F \mid +)\,p(+) + p(F \mid -)\,p(-)}$$

Add peptide separation information:

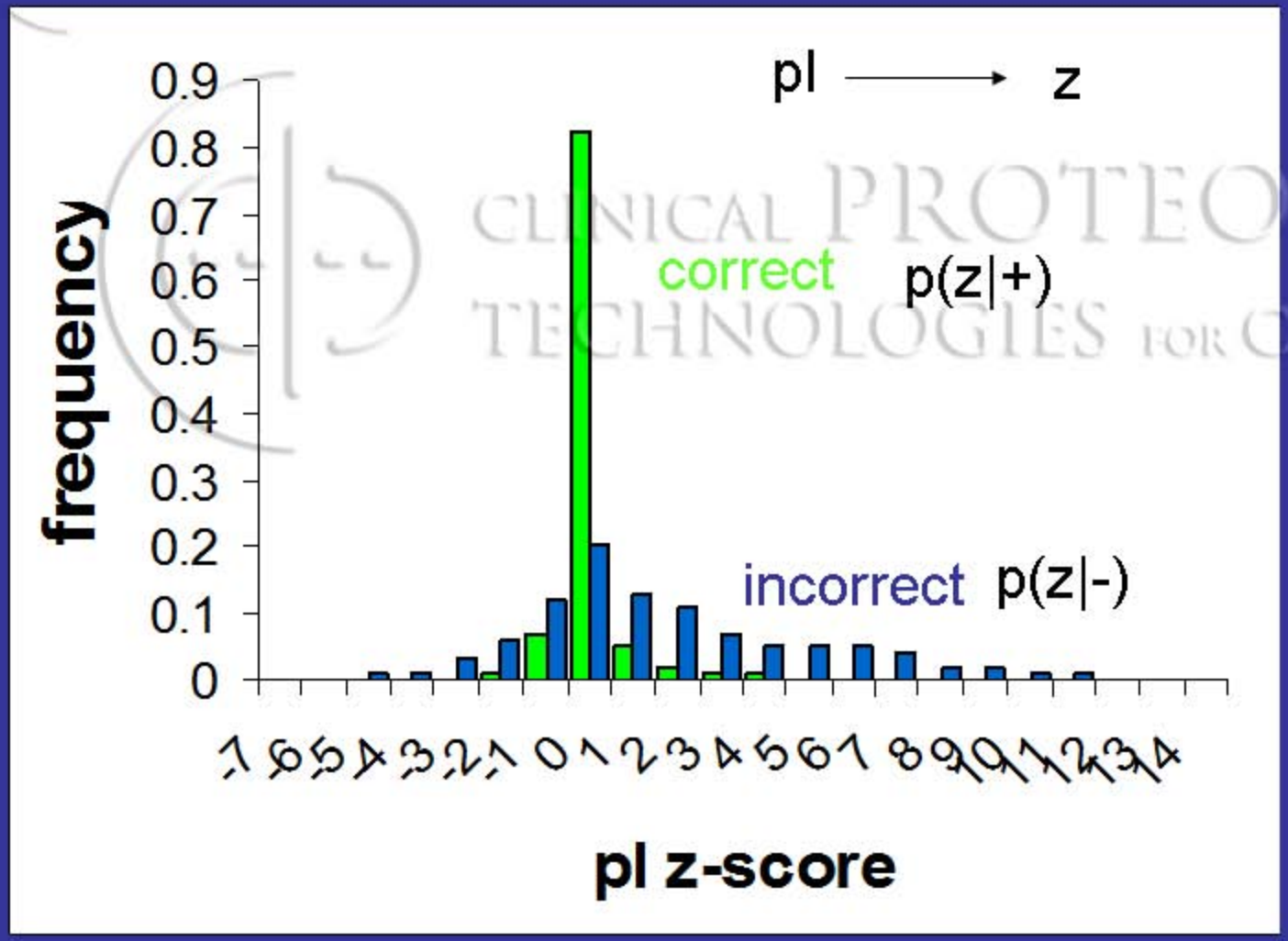
- pI value $p(pI \mid +), p(pI \mid -)$

$$p(+ \mid F, pI) = \frac{p(F \mid +)\,p(pI \mid +)\,p(+)}{p(F \mid +)\,p(pI \mid +)\,p(+) + p(F \mid -)\,p(pI \mid -)\,p(-)}$$

$$p(+ \mid F, pI) = \frac{p_+}{p_+ + (1 - p_+)\alpha^{-1}}$$ where $$\alpha = \frac{p(pI \mid +)}{p(pI \mid -)}$$

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

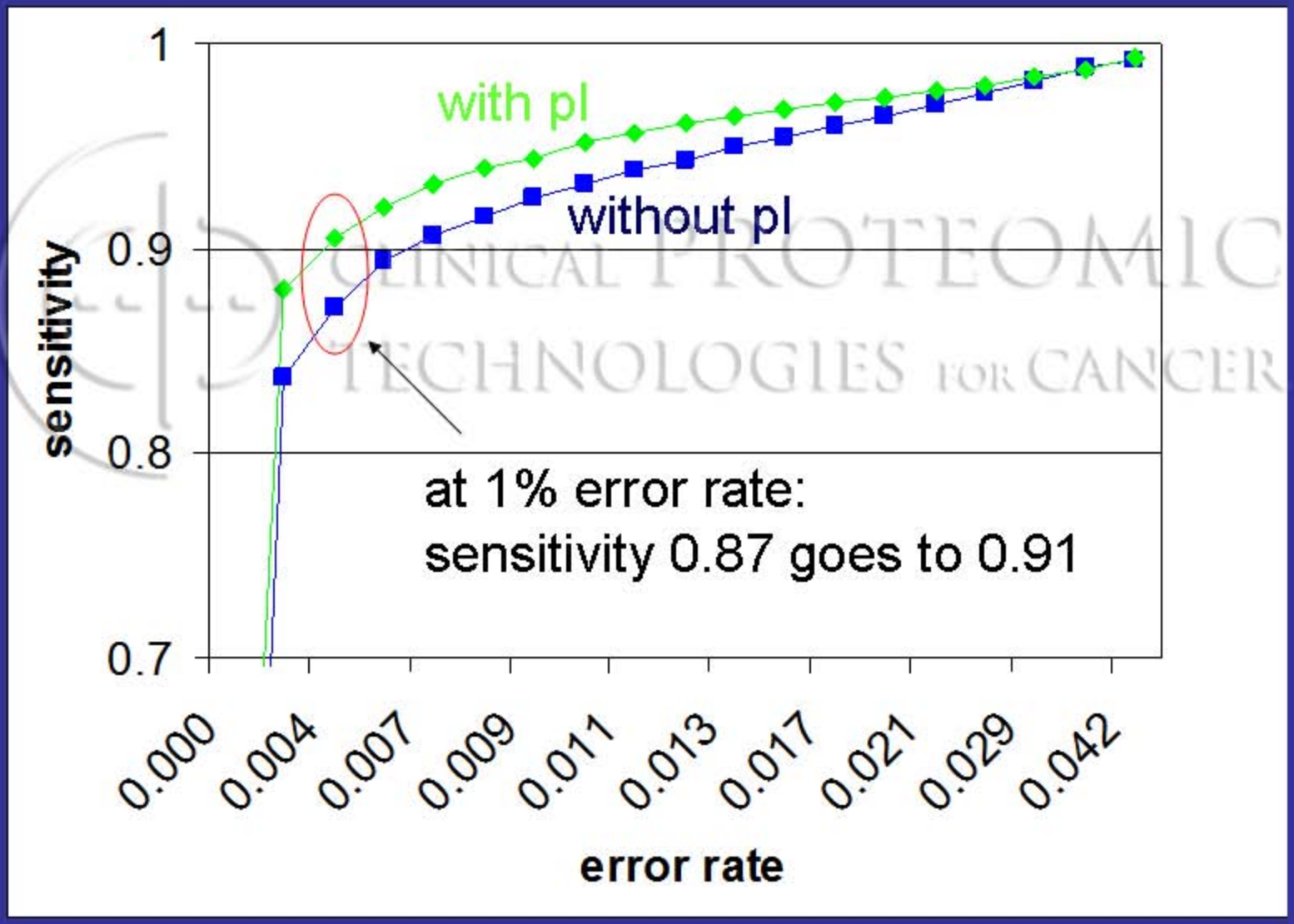
Distributions of pI z scores



data from all fractions

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Improve discrimination



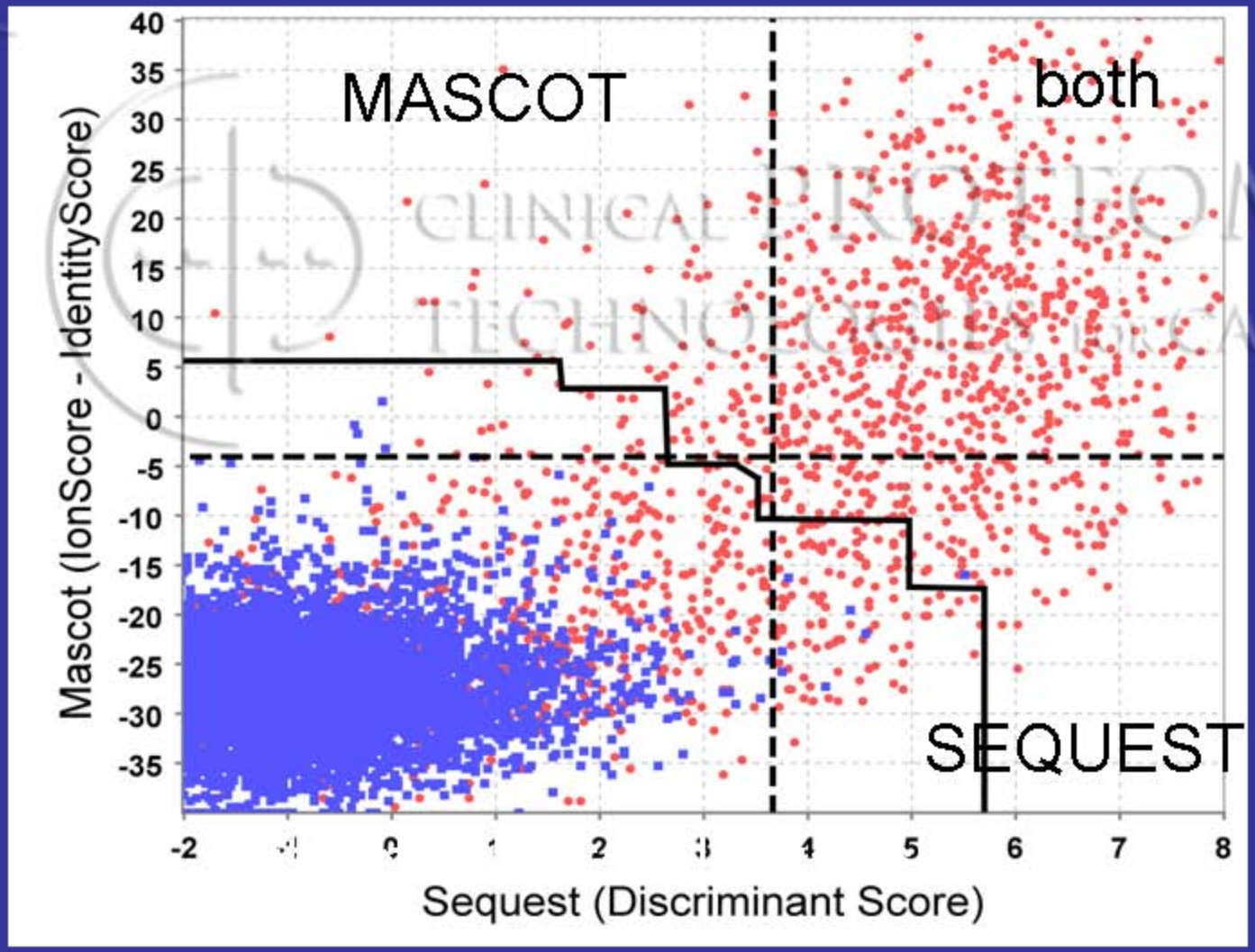
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Using Multiple Search Tools

coupling output from multiple search engines for increased accuracy and sensitivity of peptide identification

B. Searle, M. Turner, and A.I. Nesvizhskii, Submitted to *J. Prot. Research*

Improved Sensitivity

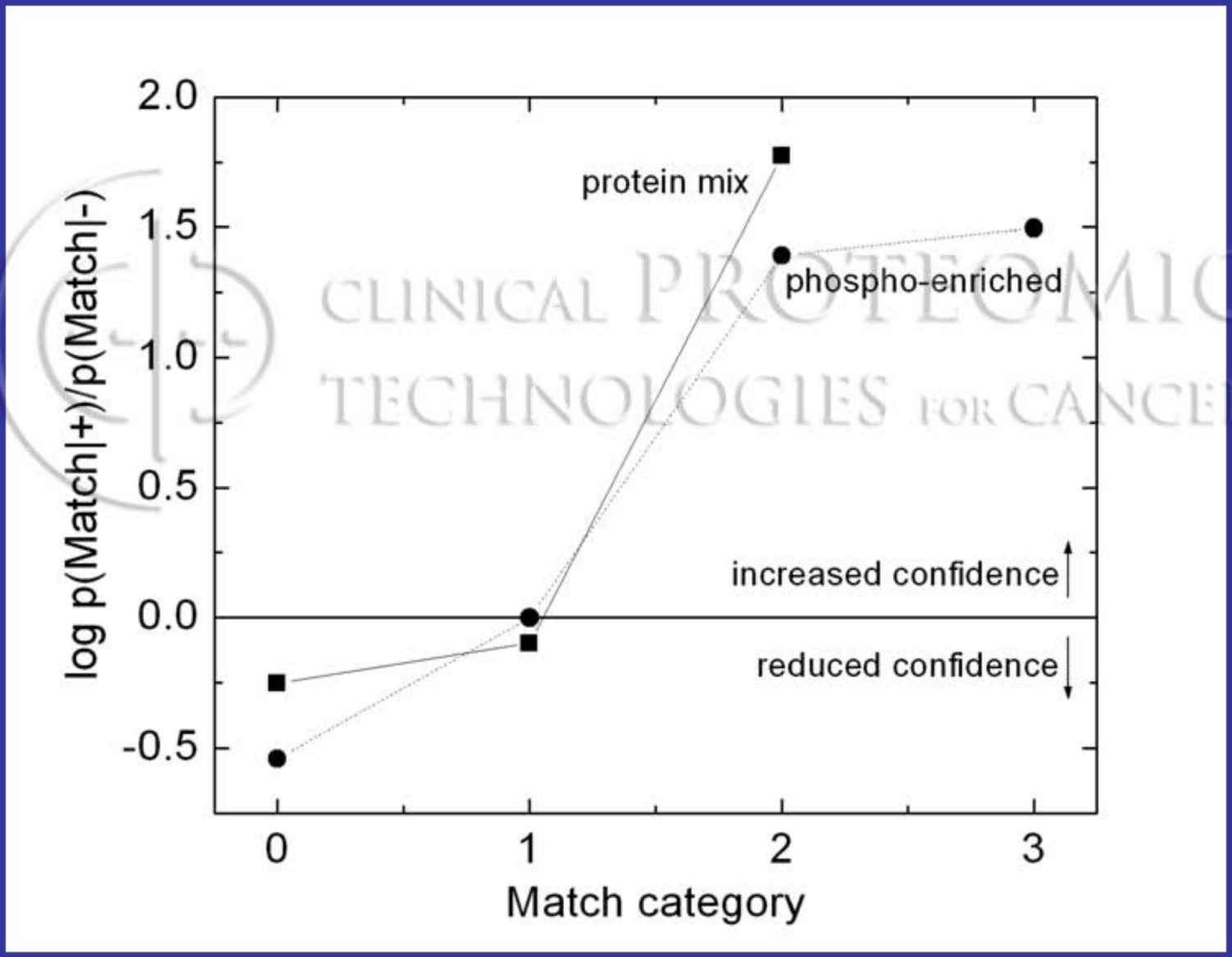


Investigating MS^2-MS^3 matching statistics:

coupling consecutive stage mass spectrometry data for increased peptide identification confidence

P. Ulintz, B. Bodenmiller, P. Andrews, R. Aebersold, A.I. Nesvizhskii,
Mol. Cell. Proteomics, in press

Match parameter distributions



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Acknowledgements

Hyunwon Choi (Biostat student)
Xia Cao (post-doc, Computer Science)
Yind Ding (Biostat student)
Damian Fermin (Bioinformatics programmer)

David Shteynberg and Seattle Proteome Center (ISB)
Brian Searle, Proteome Software (Portland, Oregon)
Johan Malmstroem, Bernd Bodenmiller, Ruedi Aebersold (ETH Zurich)

UM Collaborations:
Pete Ulintz, Phil Andrews (NRPP)
Arul Chinnaiyan, Kojo Elenitoba-Johnson (UM Pathology)
Gil Omenn and Proteomics Alliance for Cancer Research

Funding from NIH (NCI R01 CA 126239)
Subcontract from Burnham Institute (Center for Proteolytic Pathways)