Statistical Validation of Database Search Results

Alexey Nesvizhskii

nesvi@systemsbiology.org

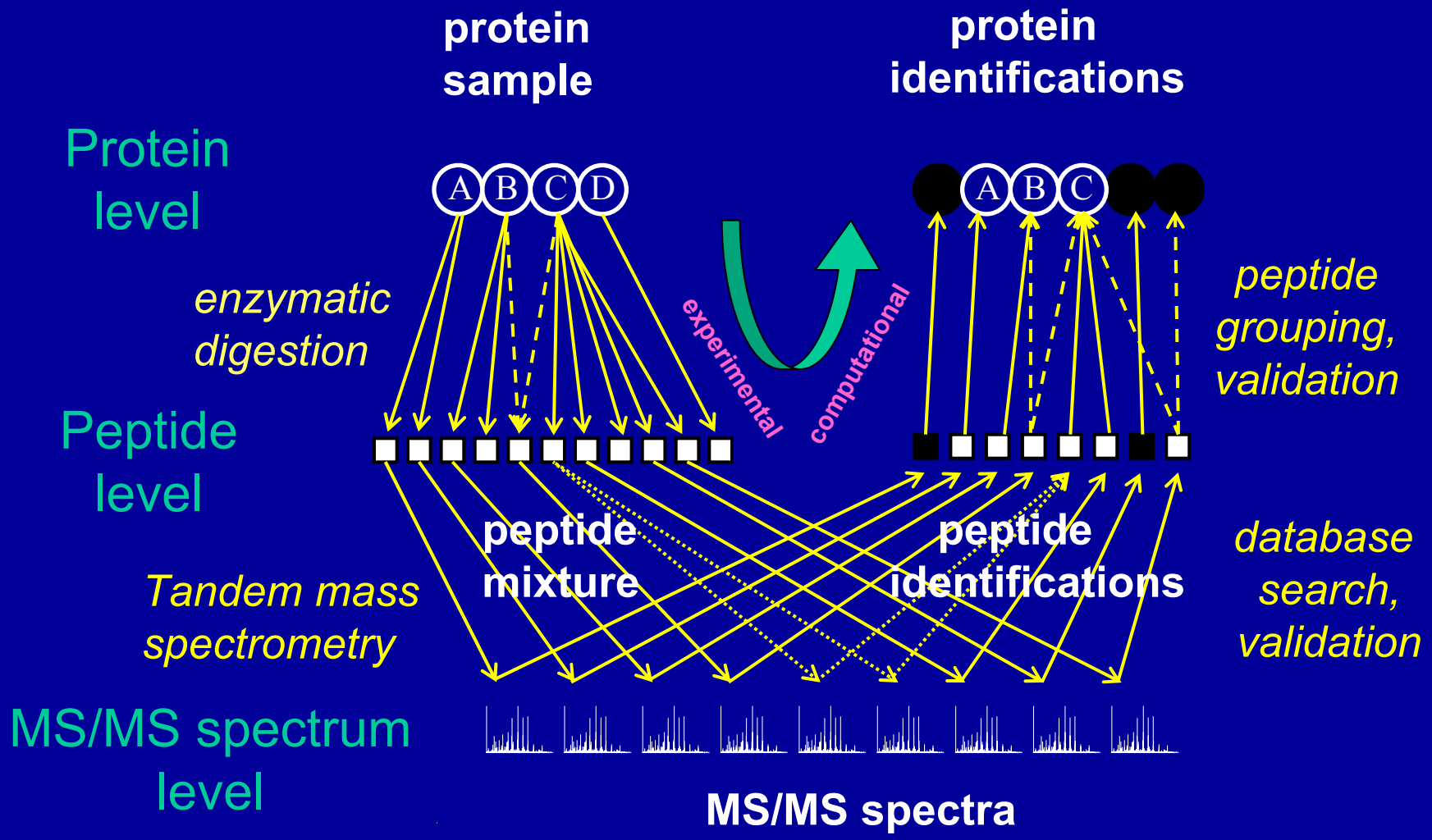
Institute for Systems Biology
Seattle, WA
U.S.A.

Standards in Proteomics
January 4, 2005

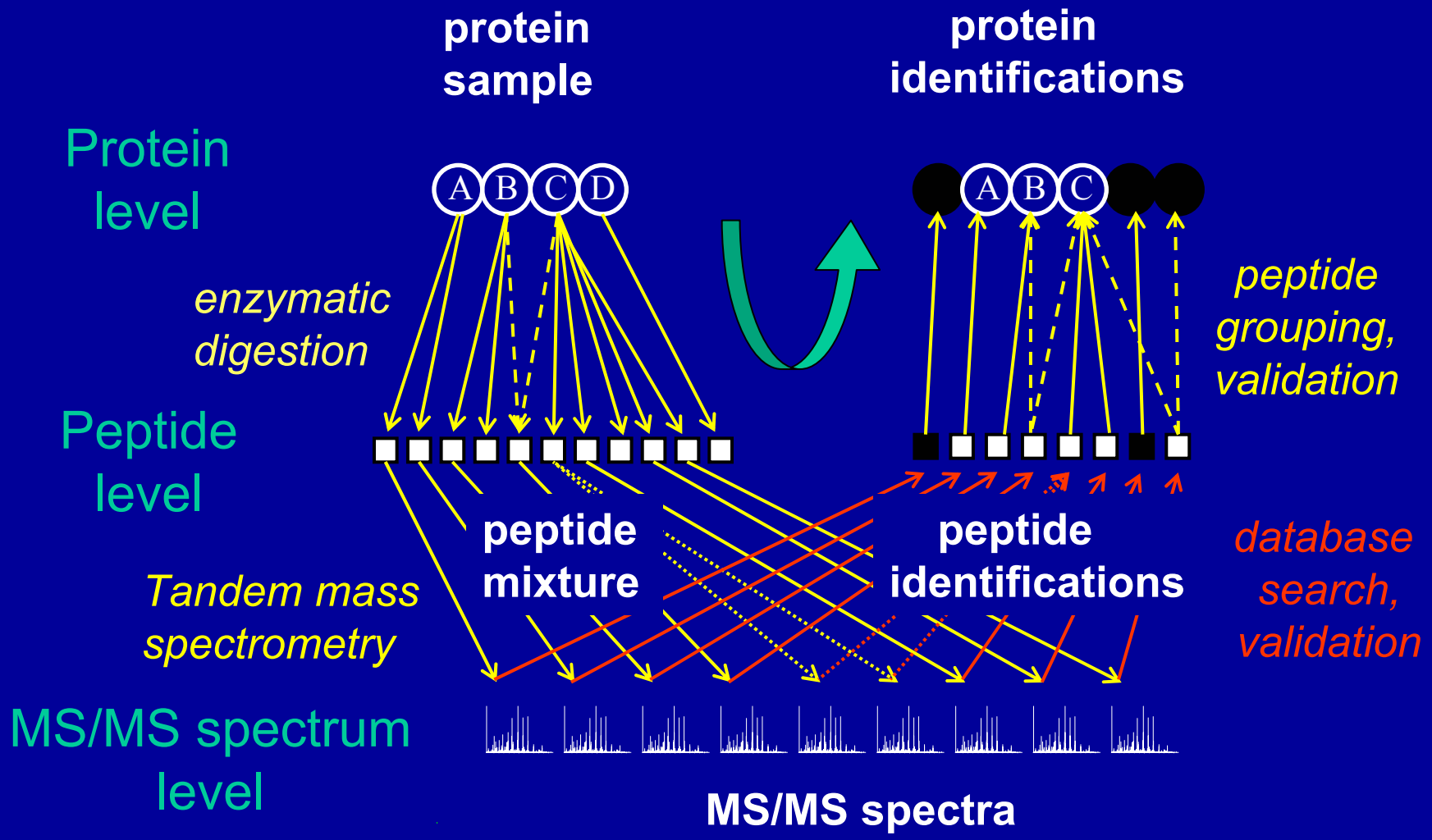
Outline

- High throughput MS/MS-based proteomics
- Peptide identification and validation
  - p-value, expectation value
  - Probabilities, error rates
- Protein Identification
  - error rate inflation (due to non-random peptide grouping)
  - protein inference problem (degenerate peptides)

Shotgun Protein Identification

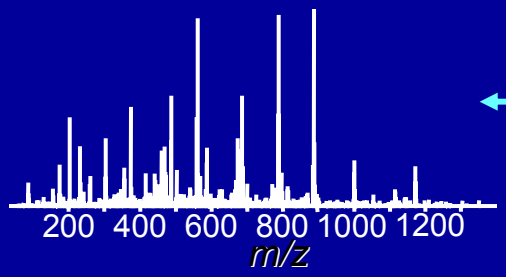


Shotgun Protein Identification



MS/MS Database Search

Acquired MS/MS spectrum

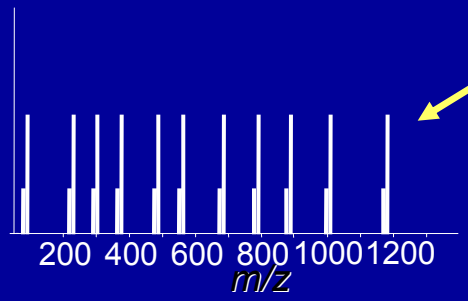


correlate

similarity score

Sequence Database

Theoretical spectrum



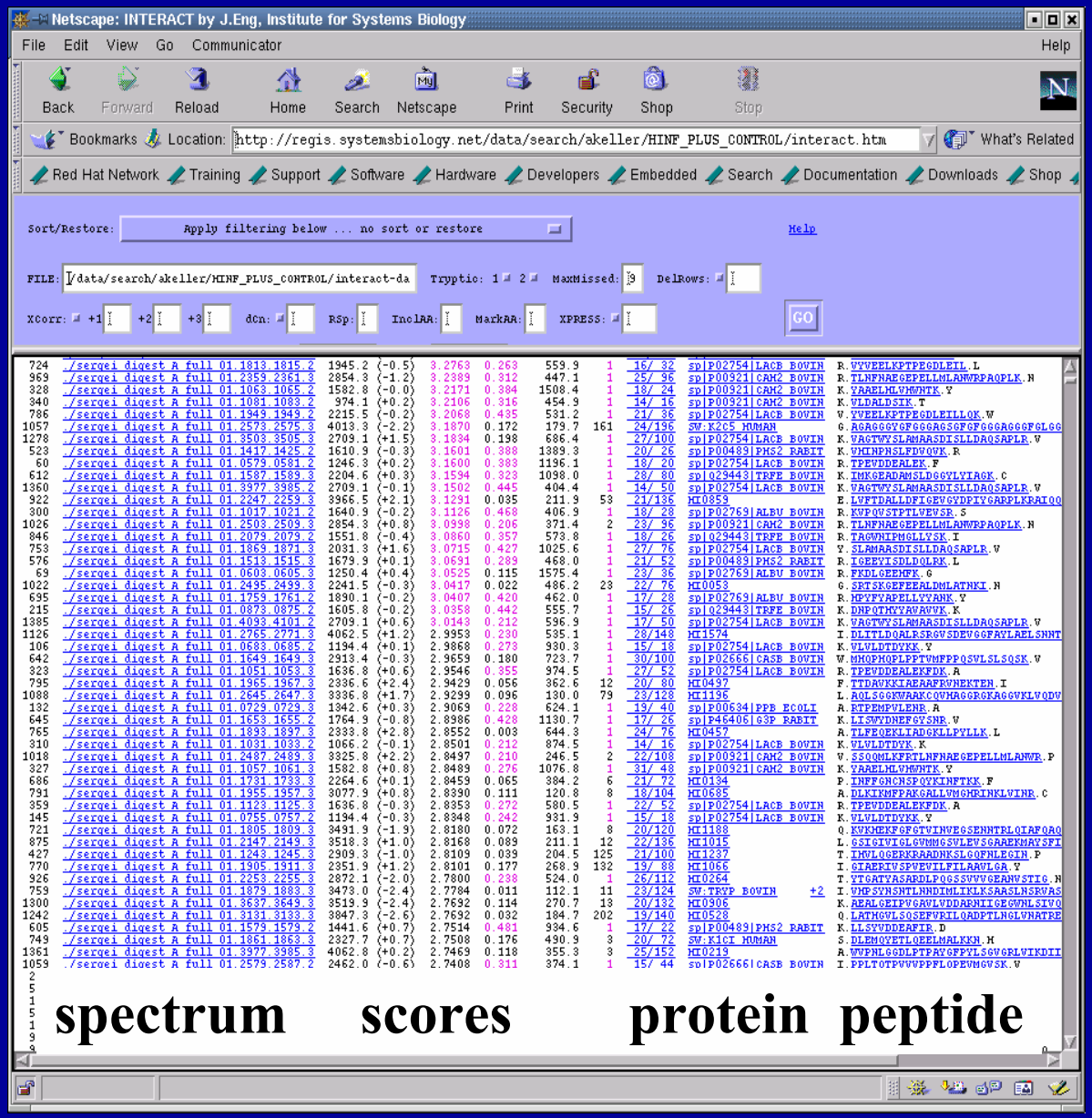
ISLLDAQSAPLR
VVEELCPTPEGK
DLLLQWCWENGK
ECDVVSNTIIAEK
GDAVFVIDALNR
VPTPNVSVVDLTNR
SYLFCMENSAEK
PEQSDLRSWTAK

best matching database peptide

Algorithms: SEQUEST, Mascot, Sonar, SpectrumMill, …

best matching database peptide can be correct or incorrect identification

Validation of Peptide Identifications

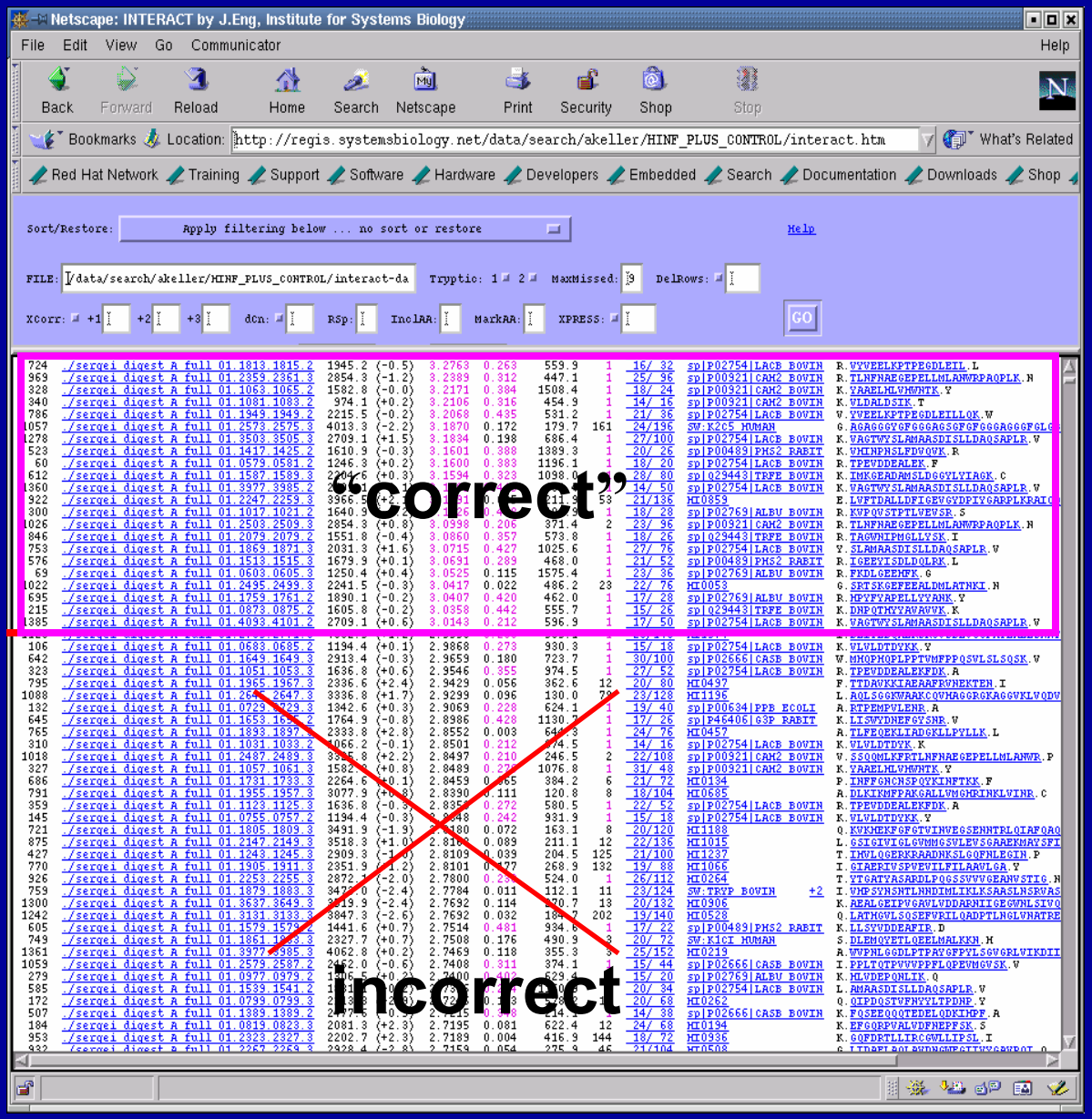


output from MS/MS database search algorithms

thousands of spectra

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

Threshold Model



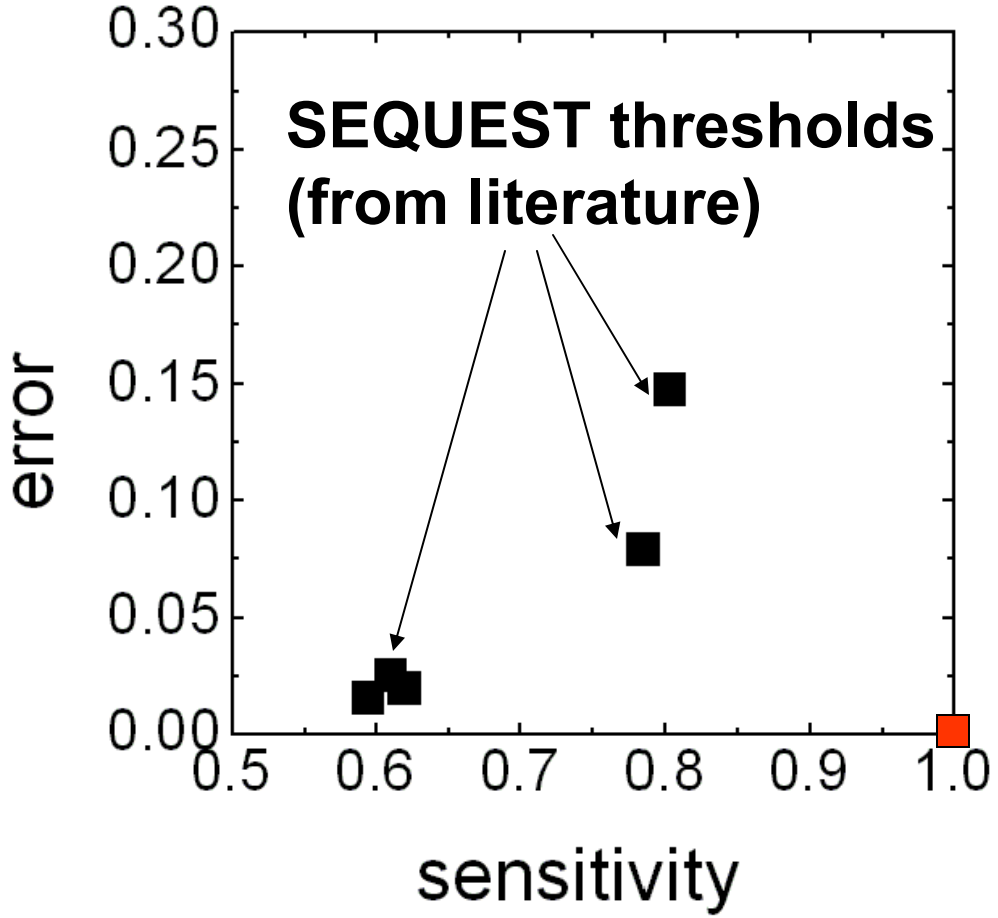
sort by search score

threshold

SEQUEST:
Xcorr > 2.0
ΔC_n > 0.1

MASCOT:
Ion Score > 30

Threshold Model: Bad Discrimination and Inconsistency



test data (18 proteins): OMICS 6(2), 207 (2002)

Sensitivity: fraction of all correct results passing filter

Error Rate: fraction of all results passing filter that are incorrect

Ideal Spot

Difficulties in Interpreting Peptide Identifications based on MS/MS

Applies to both SEQUEST and Mascot (as it is used in practice) and, to large degree, to more recent tools

- No ‘useful’ measures of confidence

  (Mascot: ‘identity threshold’ guideline is not practical and rarely used)

- Different criteria used to filter data
- Unknown and variable false positive error rates

Just as assignment of quality scores to each base in DNA sequencing was essential for the genome sequencing programs, statistical models for estimating the accuracy of peptide and protein identifications are crucial for the success of high throughput proteomics

Statistical Validation

- p-values or expectation values

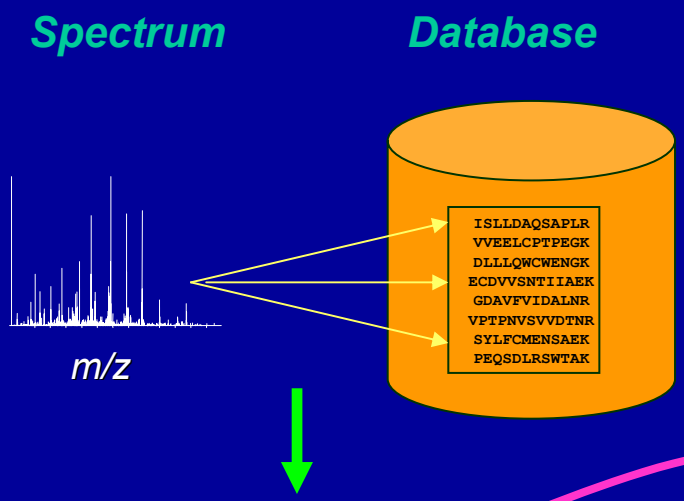
  used, e.g., in sequence similarity searching

- Probabilities (Bayes)

  based on the ratio of two distributions (correct and incorrect) derived from the data (entire dataset)

  used, e.g., in information retrieval (relevant vs. non-relevant documents)

Expectation Values (empirical model)



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

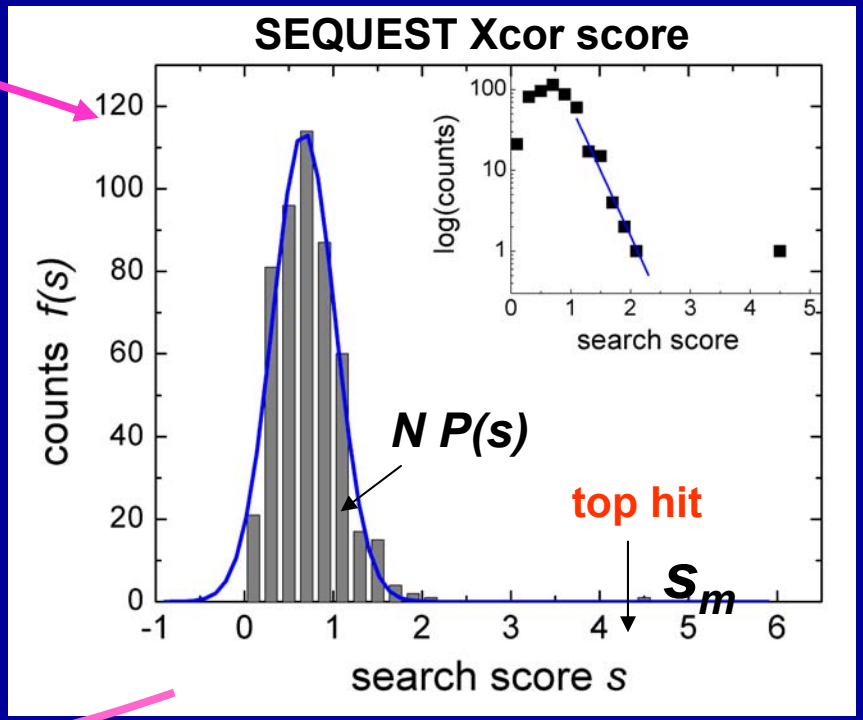
probability to get score $s \geq s_m$ by chance

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

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

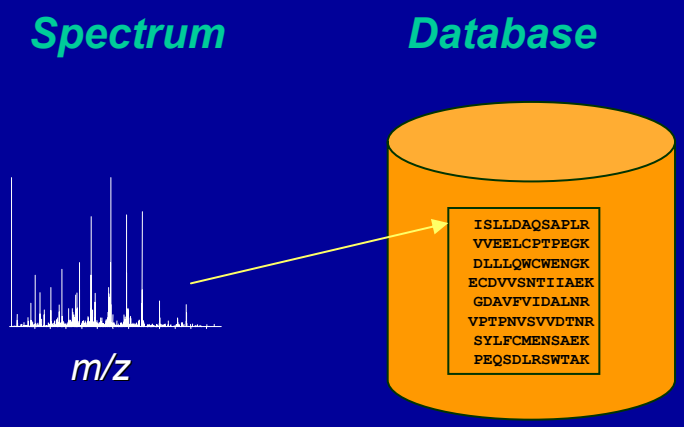
Rank	Peptide	Score	p
1	ISLLDAQSAPLR	4.5	10^{-5}
2	VVEELCTPEGK	2.1	0.11
3	DLLLQWCWENGK	2.0	0.23
4	ECDVVSNTIIAEK	1.9	0.52
5	GDAVFVIDALNR	1.7	0.72
6	VPTPNVSVVTNR	1.6	0.86
7	SYLFCMEAEK	1.6	0.86
8	PEQSDLRSWTAK	1.5	0.94
	...		...

N peptides



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

Expectation Values (explicit model)



Sadygov & Yates *Anal. Chem.* (2003)
Geer et al. *J. Proteome Res.* (2004)

$$P(s) = \frac{\mu^s}{s!}\exp(-\mu)$$

Poisson distribution

μ: function of mass tolerance, number experimental peaks number of calculated ions mass, charge

s: number of matched peaks

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

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

Geer et al. (upper bound)

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

$$E(s_m) = N(1-(1-\sum_{s \geq s_m} P(s))^N) \approx N^2 \sum_{s \geq s_m} P(s)$$

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

Sadygov et al. (lower bound)

Expectation Values (or p-values): Limitations

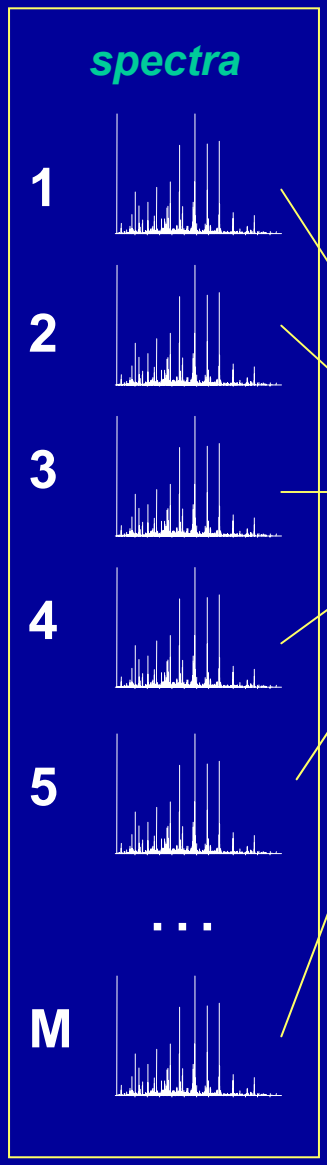
1. P-values or E-values are not well suited for the analysis of large-scale datasets (do not allow estimation of error rates as a function of filtering threshold)

   see, e.g., recent papers by Tsibshirani and others on the subject of p -values vs. False Discovery Rate (FDR) approach

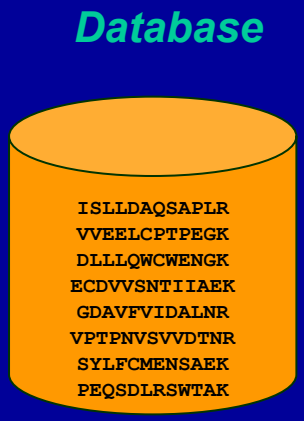
2. Difficult to take advantage of other useful information (e.g., number of missed cleavages, peptide retention time)

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

Modeling Large-Scale Datasets



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
…		
M	SYLFCMEAEK	1.1

best match
to each spectrum

raw score
E value
p-value

Statistical Model for Computing Peptide Probabilities (PeptideProphet)

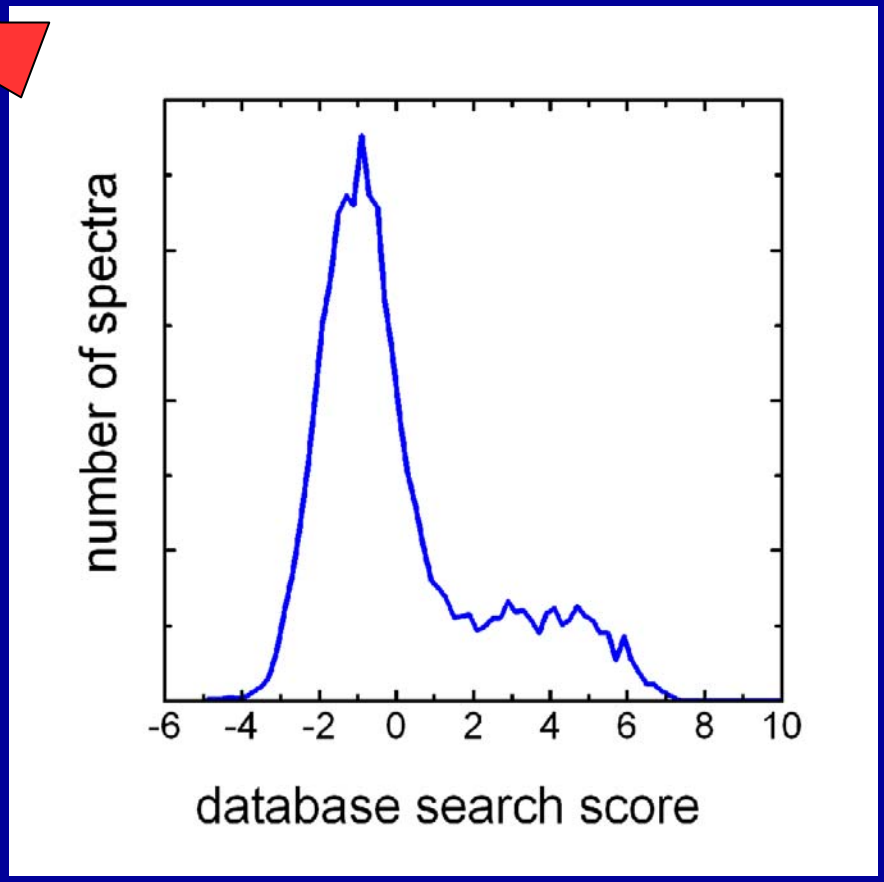
entire dataset:

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



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

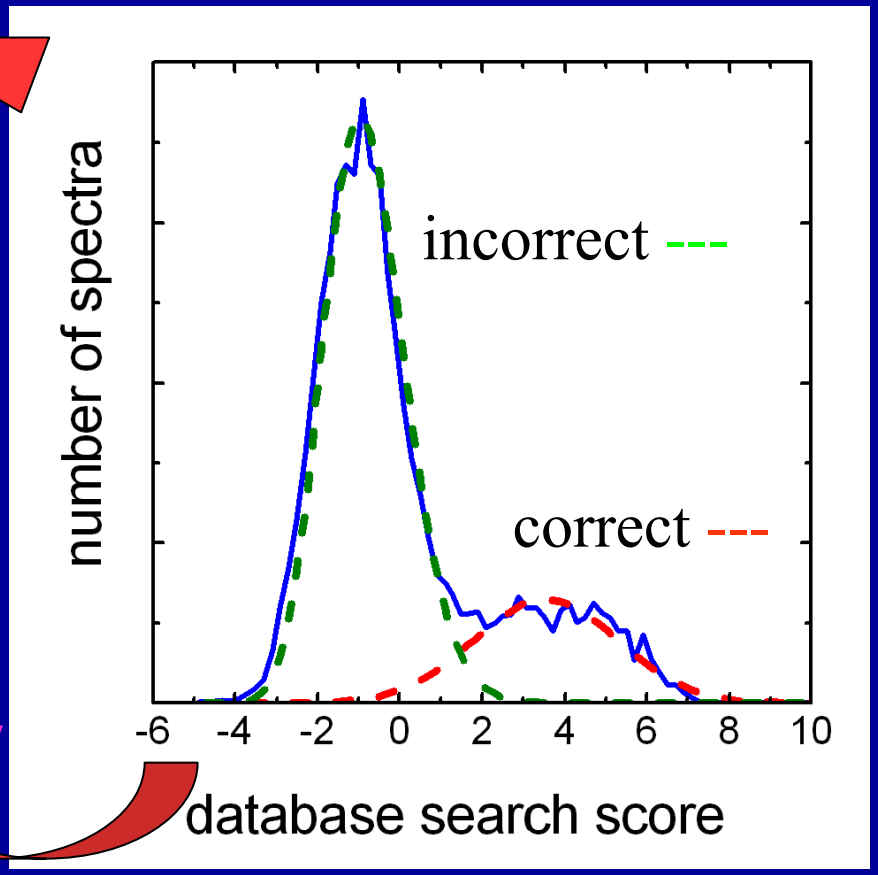
Statistical Model for Computing Peptide Probabilities (PeptideProphet)

entire dataset:

	Spectrum	Peptide	Score	
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

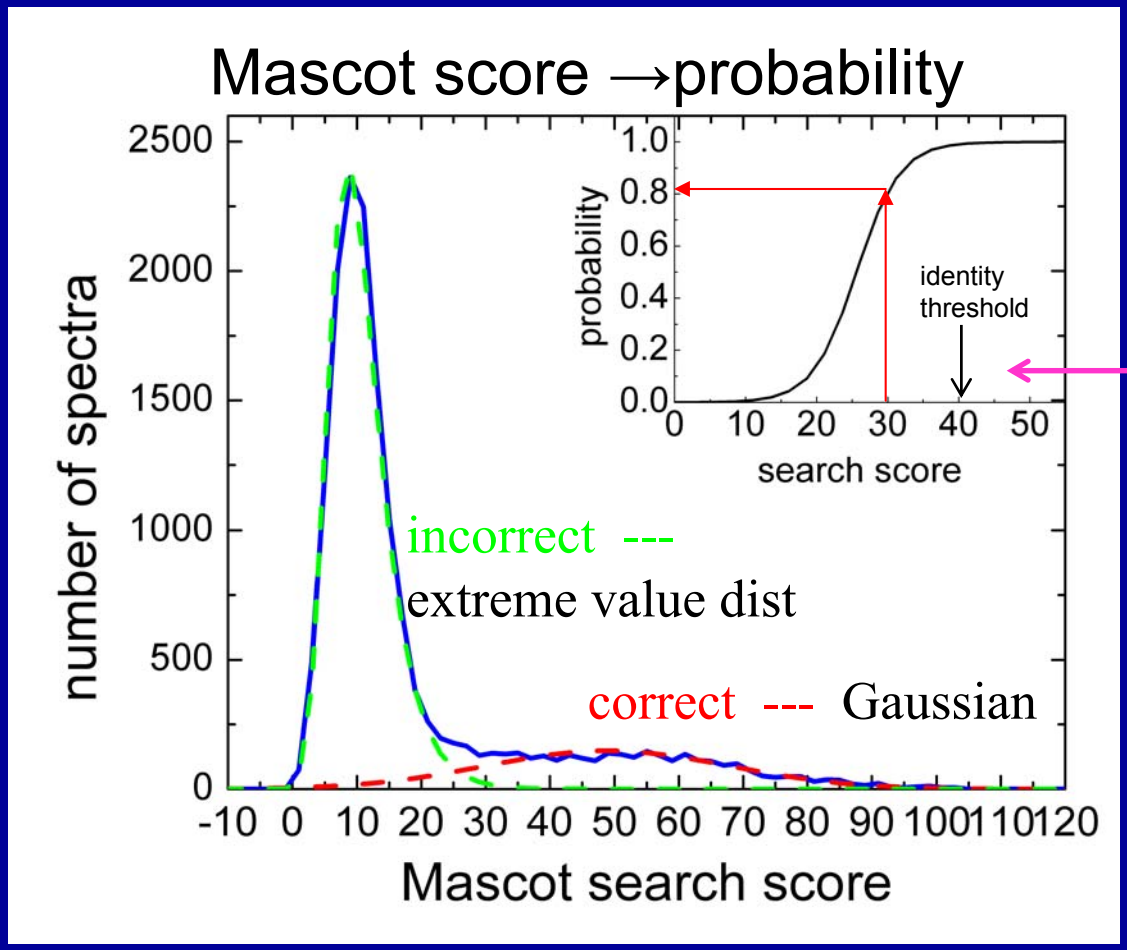
probability

‘unsupervised clustering’



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

Illustration: Assigning Probabilities to Mascot Search Results



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

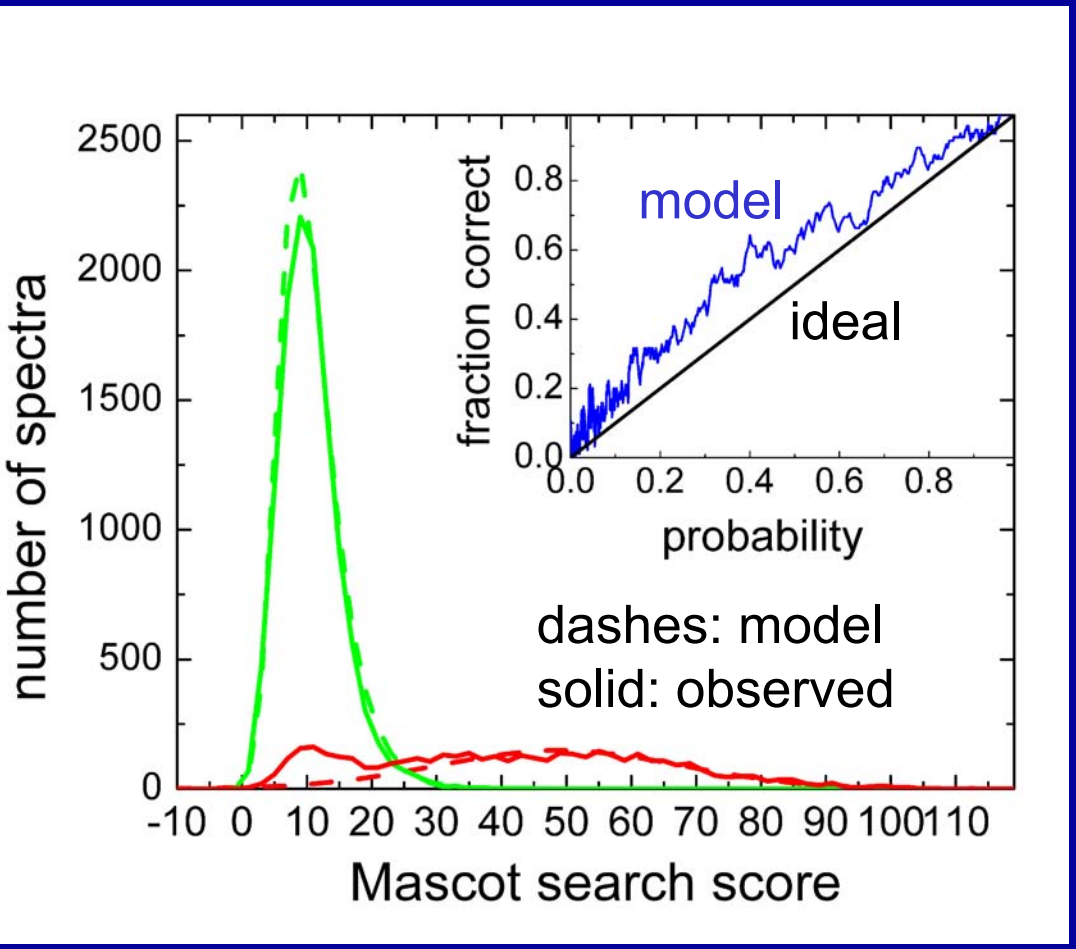
distributions are learned from the data

conversion of Mascot search scores into probabilities

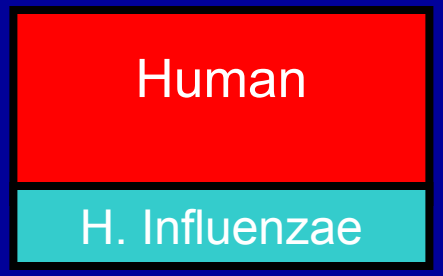
To address a common misunderstanding:

distribution parameters ARE NOT determined using a control dataset of 18 proteins, or any other training dataset for that matter. They are learned from each analyzed dataset anew using the EM mixture model algorithm

Accuracy of Learned Distributions and Computed Probabilities



database searched:



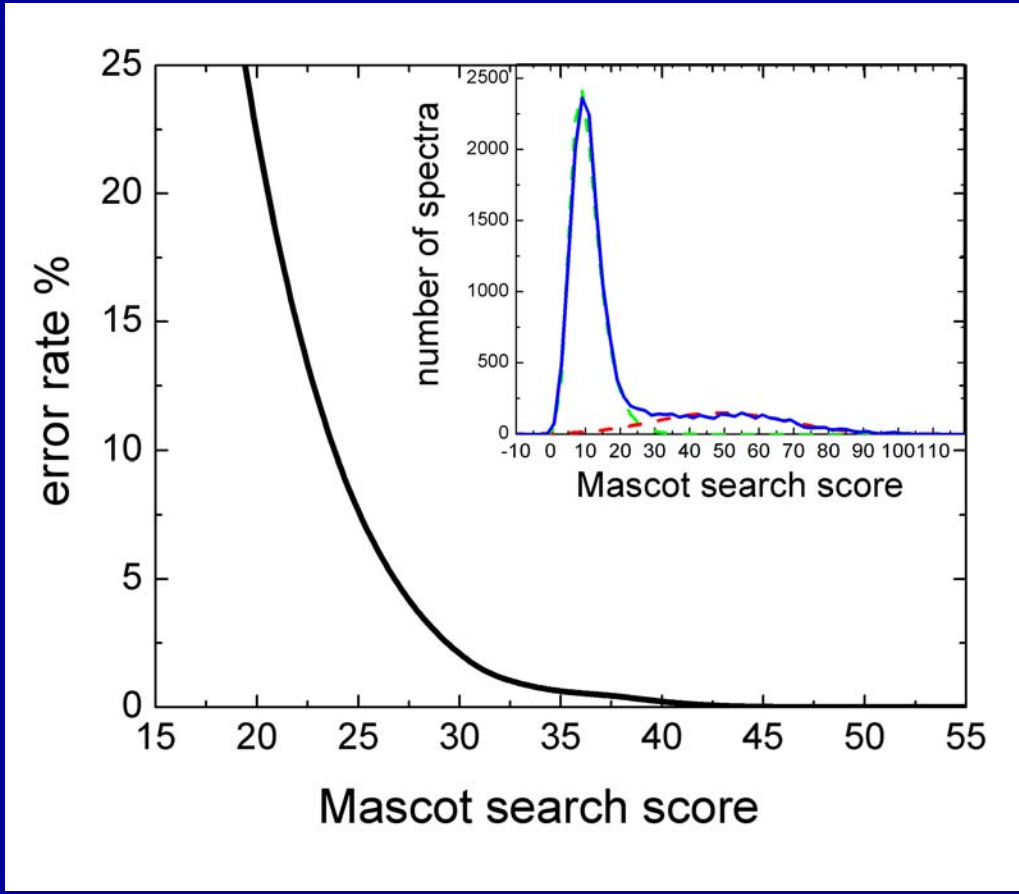
size ratio: ~ 20:1

For those familiar with "reverse database search" approach:
This is an equivalent of appending 20 randomized databases of equal size.

Method is accurate

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

Estimation of False Positive Error Rates



Computed probabilities allow to estimate sensitivity and false positive error rate, e.g.

$$< Err >= \frac{\sum_{\{i|p_i \geq p_T\}} (1 - p_i)}{\sum_{\{i|p_i \geq p_T\}} 1}$$

Use model predictions to achieve desired balance between sensitivity and the error rates

no need to do “reverse database searching” to estimate the error rates

Information Useful for Validation

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:

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

PeptideProphet statistical model can use all this information → method is more discriminative

PeptideProphet Software

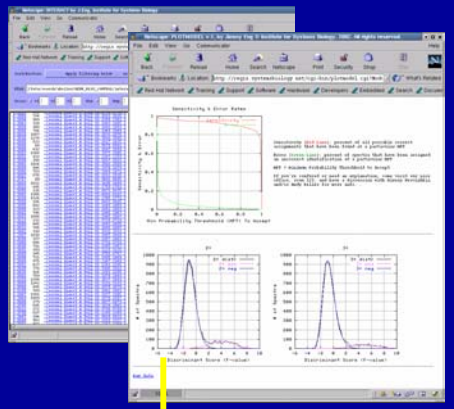


LCQ TOF-TOF Q-TOF

SEQUEST MASCOT

hard to compare datasets, unknown error rates

Also: different sample preparation sample complexity quality of spectra



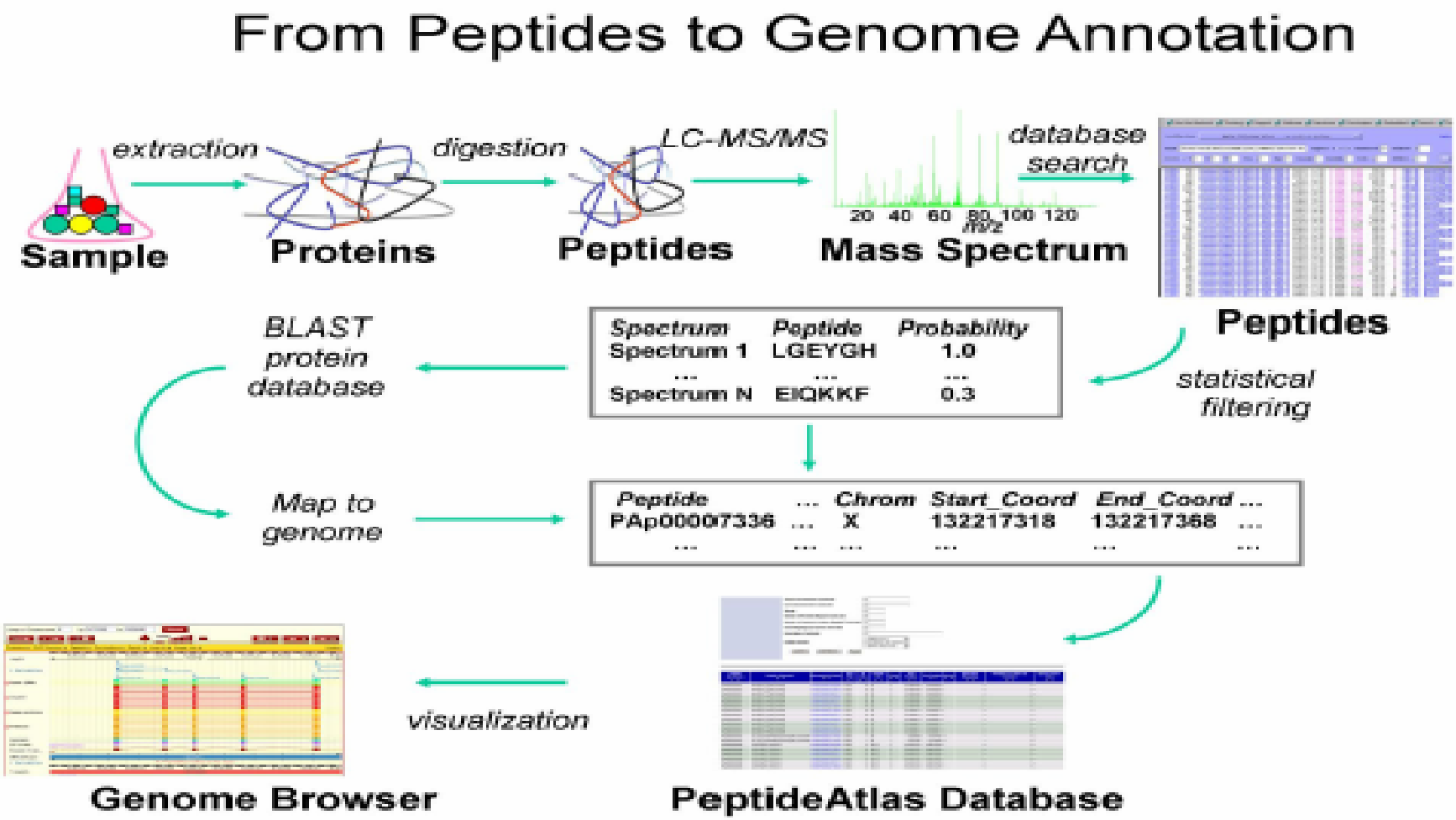
PeptideProphet

easy comparison quality control, known error rates

peptide identifications with accurate probabilities

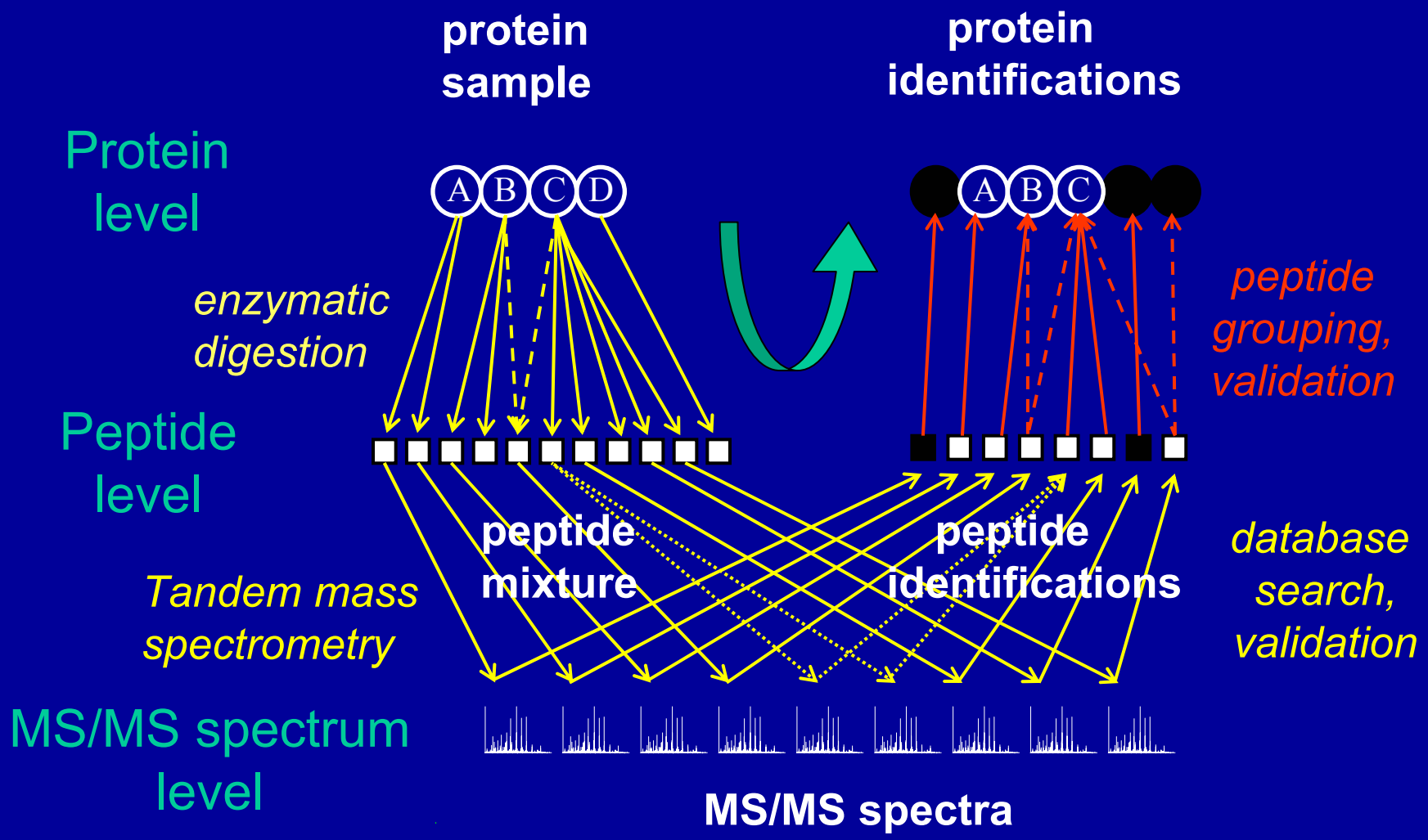
download from *www.proteomecenter.org*

PeptideAtlas: High Throughput Pipeline



Protein Identification

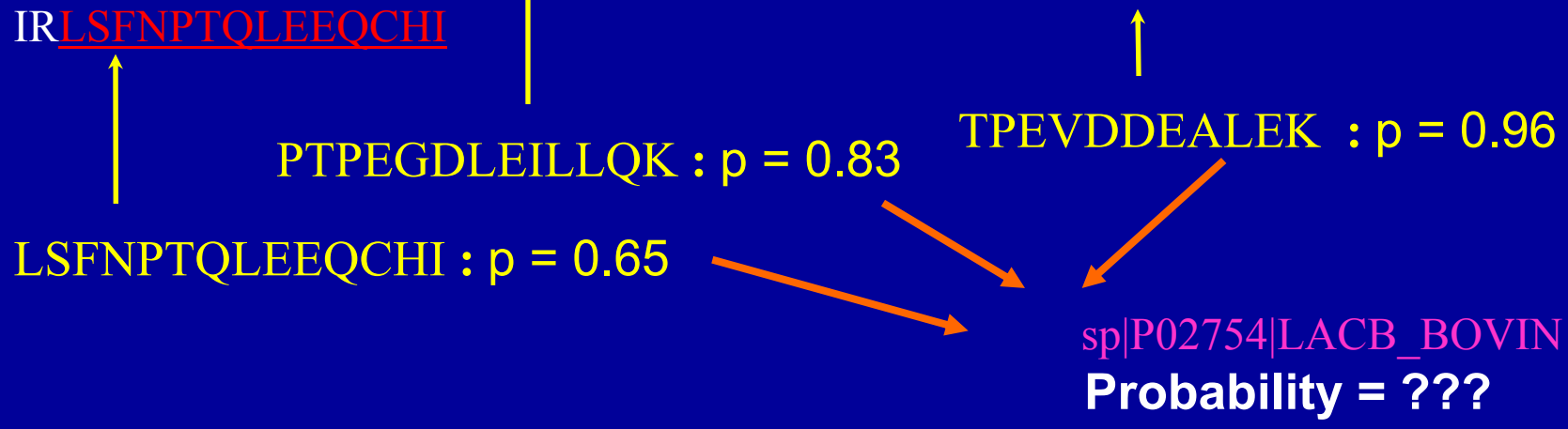
Shotgun Protein Identification



Protein Identification

>sp|P02754|LACB_BOVIN BETA-LACTOGLOBULIN PRECURSOR (BETA-LG) (ALLERGEN BOS D 5) - Bos taurus (Bovine).

MKCLLLALALTCGAQALIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSA
PLRVYVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKL
VLDTDYKKYLLFCMENSAEPEQSLACQCLVRTPEVDDEALEKFDKALKALPMH
IRLSFNPTQLEEQCHI



ProteinProphet combines probabilities of peptides assigned to MS/MS spectra to compute *accurate* probabilities that corresponding proteins are present?

Issues for Protein Identification

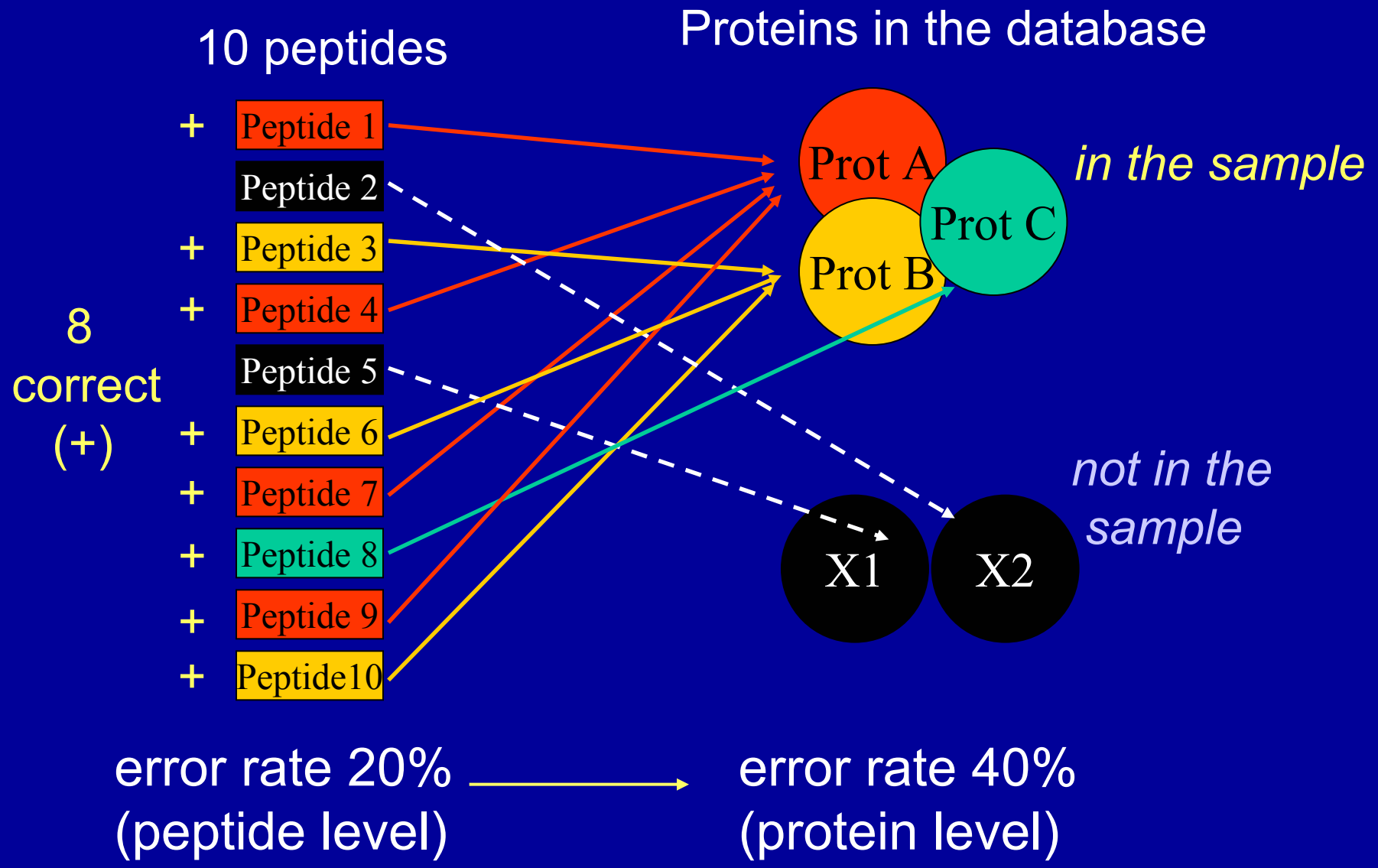
- Peptides corresponding to 'single-hit' proteins are less likely to be correct than those corresponding to 'multi-hit' proteins

- Many peptides are present in more than a single protein (isoforms, homologous proteins etc.)

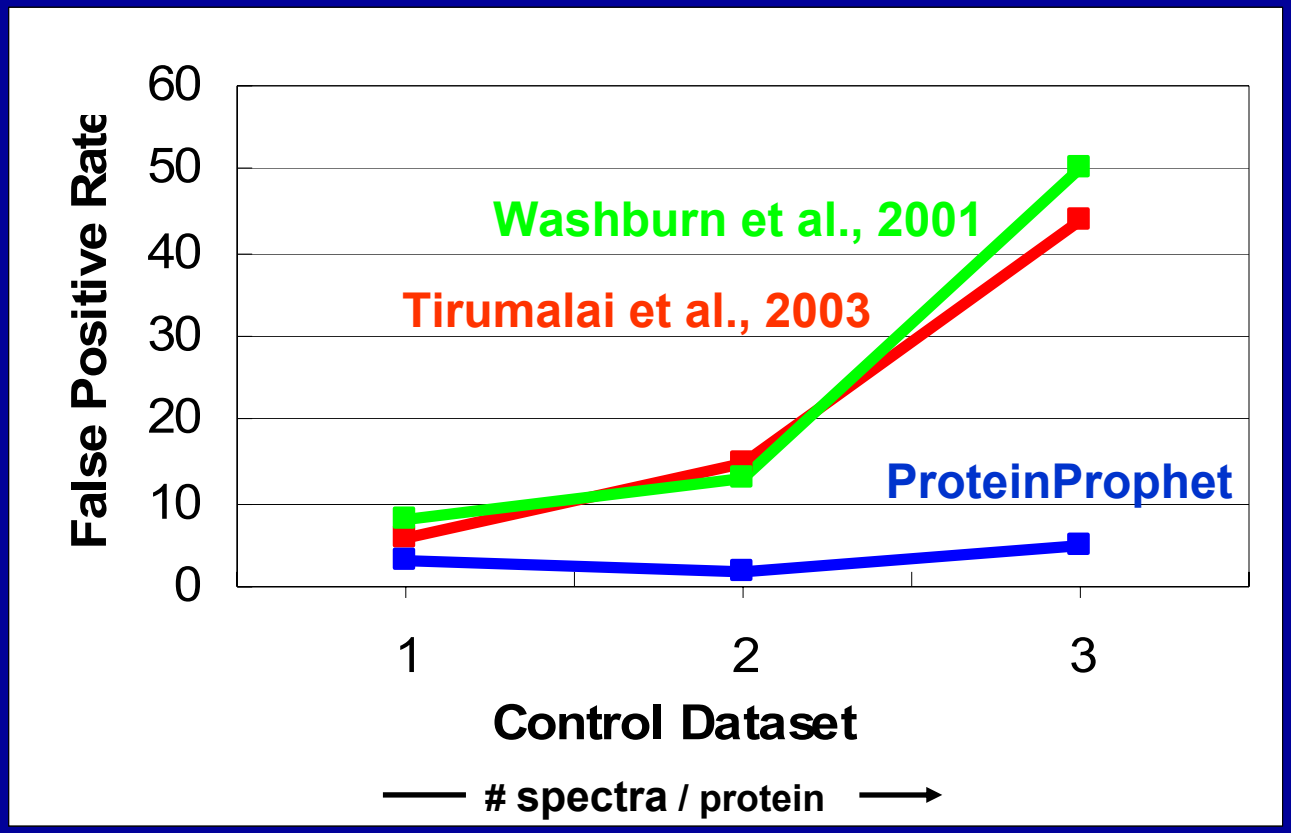
A. I. Nesvizhskii *et al.*, *Anal. Chem.* **75**, 4646-4658 (2003)

A. I. Nesvizhskii & R. Aebersold *Drug Discov Today* **9**, 173-81 (2004)

Amplification of False Positive Error Rates from Peptide to Protein Level



Protein ID False Positive Rates: ProteinProphet vs. Score Thresholds



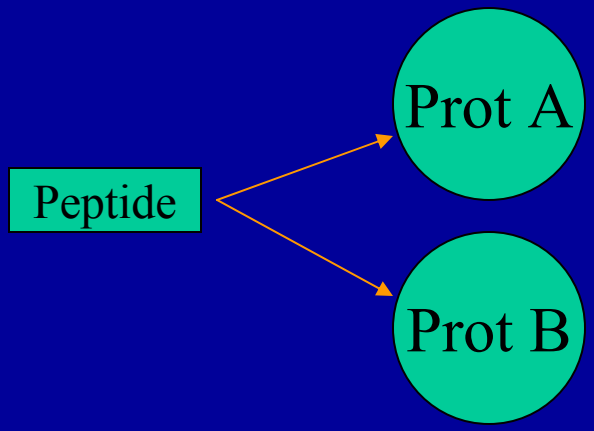
Keep acquiring data and you will "identify" everything in the database!

Control Datasets:

1 Halobacterium vs. Halo+Human (4 runs)
2 Halobacterium vs. Halo+Human (45 runs)
3 18 purified proteins vs. 18+Human (22 runs)

Protein Inference Problem

Degenerate peptides: correspond to more than a single entry in protein database



protein A or protein B ??

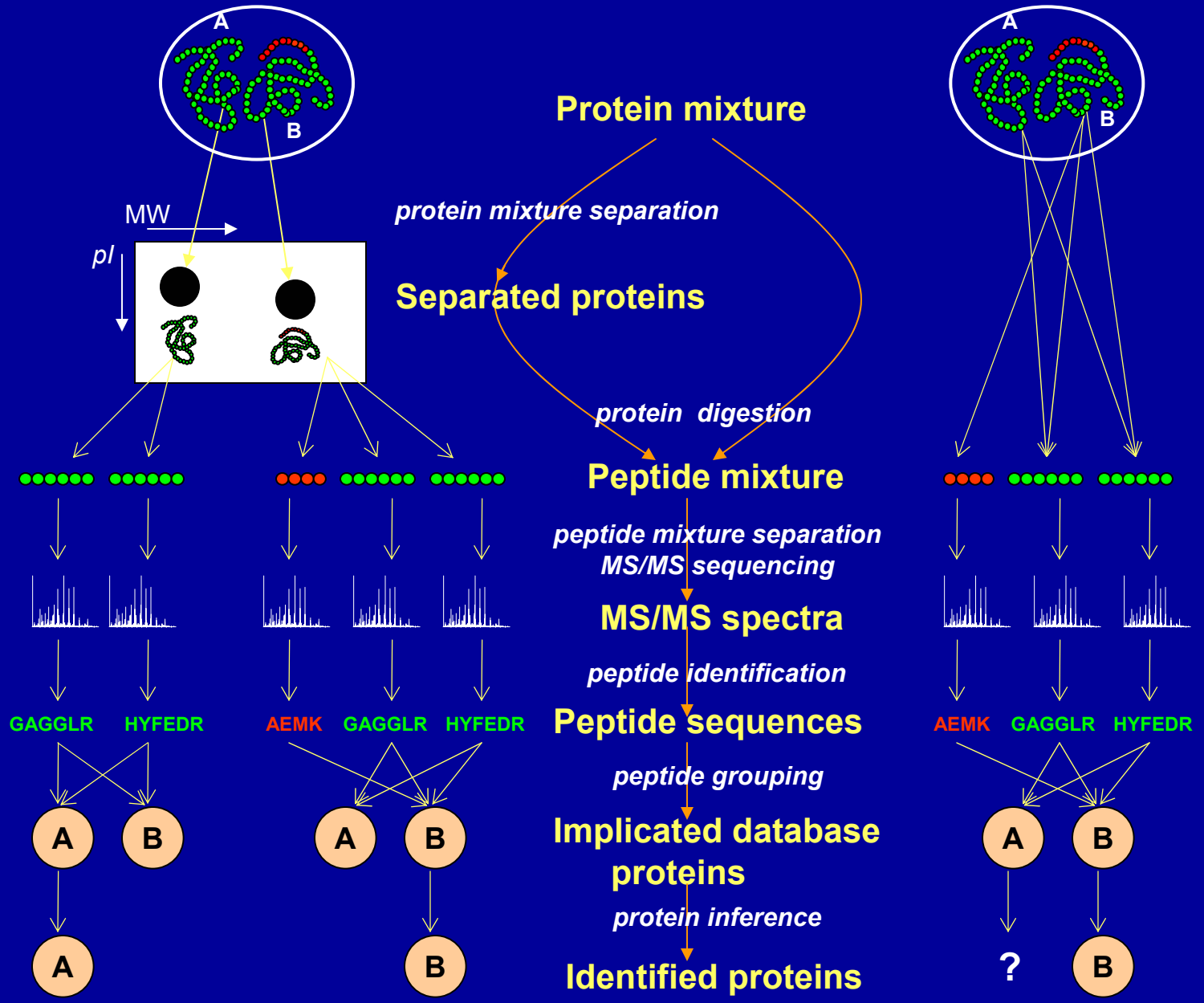
Or both?

In shotgun proteomics the connectivity between peptides and proteins is lost

Degenerate peptides are more prevalent with databases of higher eukaryotes due to the presence of:

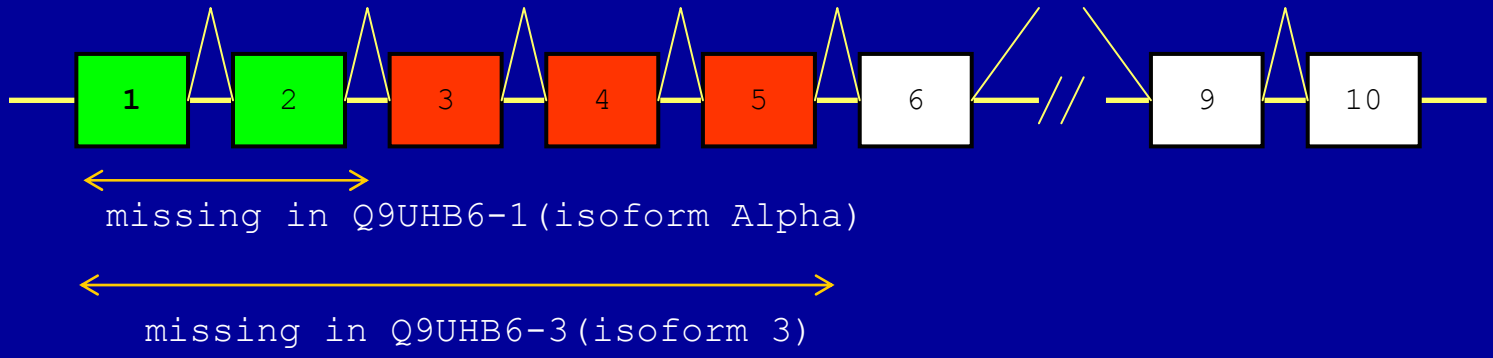
- related protein family members
- alternative splice forms
- partial sequences

2D Gel-based Approach vs. Shotgun Approach



Isoforms

Gene: EPLIN location: 12q13 Q9UHB6-1 (isoform Beta)



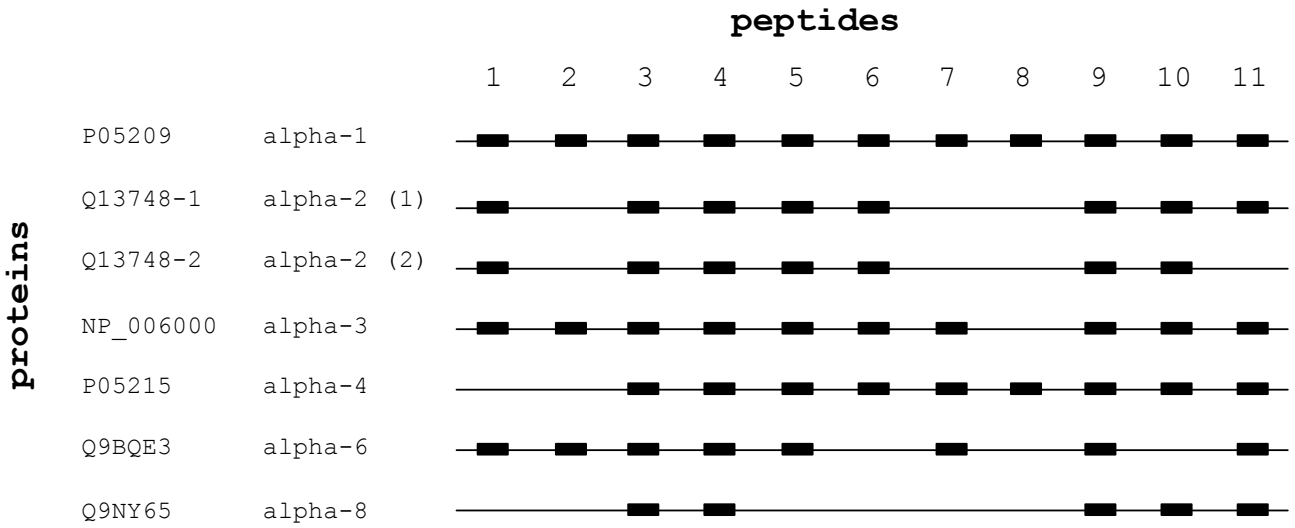
>IPI:IPI00220465 Splice isoform Alpha of Q9UHB6 Epithelial protein lost in neoplasm

MESSPFNRRQWTSLSLRVTAKELSLVNKNKSSAIVEIFSKYQKAAEETNMEKKRSNTENLSQHFRKGTLTVLKKKWENP
GLGAESHTDSLRNSSTEIRHRADHPPAEVTSHAASGAKADQEEQIHPRSRLRSPPEALVQGRYPHIKDGEDLKDHSTES
KKMENCLGESRHEVEK**SEISENTDASGKIEK**YNVPLNRLKMMFEKGEPTQTKILRAQSRSASGRKISENSYSLDDLEIG
PGQLSSSTFDSEKNESRRNLELPRLSETSIKDRMAKYQAAVSKQSSSTNYTNELKASGGEIKIHKMEQKENVPPGPEVC
ITHQEGEKISANENSLAVRSTPAEDDSRDSQVKSEVQQPVHPKPLSPDSRASSLSESSPPKAMKKFQAPARETCVECQK
TVYPMERLLANQQVFHISCFRCSYCNNKLSLGTYASLHGRIYCKPHFNQLFKSKGNYDEGFGHRPHK**DLWASKNENEEI**
LERPAQLANARETPHSPGVEDAPIAKVGV**LAASMEAKASSQQEK**EDKPAETKKLRIAWPPPTELGSSGSALEEGIKMSK
PKWPPEDEISKPEVPEDVDLDLKKLRRSSSLKER**SRPFTVAASFQSTSVK**SPKTVSPPIRKGWSMSEQSEESVGGRVAE
RKQVENAKASKKNGNVGKTTWQNKESKGETGKRSKEGHSLEMENENLVENGADSDEDDNSFLKQQSPQEPKSLNWSSFVD
NTFAEEFTTQNQKSQDVELWEGEVVKELSVEEQIKR**NRYYDEDEDE**E

Gene families (paralogues)

C **Tubulin-alpha protein family**

1	TIGGGDDSFNTFFSETGAGK	5	IHFPLATYAPVISAEK	9	VGINYQPPTVVPGGDLAK
2	AVFVDLEPTVIDEVR	6	AYHEQLSVAEITNACFEPANQMVK	10	AVCMLSNTTAIAEAWAR
3	QLFHPEQLITGKEDAANNYAR	7	YMACCLLYR	11	LDHKFDLMYAK
4	NLDIERPTYTNLNR	8	SIQFVDWCPTGFK		



>IPI00387144 IPI:IPI00387144.2|SWISS-PROT:P05209 Tubulin alpha-1 chain

MRECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSDK**TIGGGDDSFNTFFSETGAGK**HVPR**AVFVDLEPTVIDEVR**TGTYR**QLFHPEQ**
LITGKEDAANNYARGHYTIGKEIIDLVLDRIRKLADQCTGLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPAPQVSTAVV
EPYNSILTTHTTLEHSDCAFMVDNEAIYDICRR**NLDIERPTYTNLNR**LISQIVSSITASLRFDGALNVDLTEFQTNLVPYPR**IHFPLATYA**
PVISAEKAYHEQLSVAEITNACFEPANQMVKCDPRHGK**YMACCLLYR**GDVVPKDVNAAIATIKTKR**SIQFVDWCPTGFKVGINYQPPTVVP**
GGDLAKVQR**AVCMLSNTTAIAEAWARLDHKFDLMYAK**RAFVHWYVGEGMEEGEFSEAREDMAALEKDYEEVGVDSVEGEGEEEGEEY

Mature Protein Forms

A

>IPI00019906.1|UniProt/Swiss-Prot:P35613 Basigin precursor

signal peptide

MAAALFVLLGFALLGTHGASG**AAGTVFTTVEDLGSK**ILLTCSLNDSATEVTGHR**WLKGGVVLKEDALPGQK**TEFKVDSDDQWGEYSCVFL
PEPMGTANIQLHGPPRVKAVK**SSEHINEGETAMLVCK**SESVPPVTDWAWYKITDSEDKALMNGSESRFFVSSSQGR**SELHIENLNMEADP**
GQYRCNGTSSKGSDQAIITLRVRSHLAALWPFLGIVAEVLVLVTIIFIYEKR**RKPEDVLDDDDAGSAPLK**SSGQHQNDKGKNVRQRNSS

signal peptide removed

Protein Inference using ProteinProphet

- Peptides corresponding to 'single-hit' proteins are less likely to be correct than those corresponding to 'multi-hit' proteins

  ProteinProphet learn from the data itself by how much peptide probabilities should be adjusted to reflect this protein grouping information

- Many peptides are present in more than a single protein (isoforms, homologous proteins etc.)

  ProteinProphet resolves ambiguous cases (when possible) and presents the list of protein identifications in a convenient "biologist-friendly" format

A. I. Nesvizhskii *et al.*, *Anal. Chem.* **75**, 4646-4658 (2003)

Publishing Large-Scale Datasets: Example

The Application of New Software Tools to Quantitative Protein Profiling Via Isotope-coded Affinity Tag (ICAT) and Tandem Mass Spectrometry

II. EVALUATION OF TANDEM MASS SPECTROMETRY METHODOLOGIES FOR LARGE-SCALE PROTEIN ANALYSIS, AND THE APPLICATION OF STATISTICAL TOOLS FOR DATA ANALYSIS AND INTERPRETATION*

Priska D. von Haller‡, Eugene Yi, Samuel Donohoe, Kelly Vaughn, Andrew Keller, Alexey I. Nesvizhskii, Jimmy Eng, Xiao-jun Li, David R. Goodlett, Ruedi Aebersold, and Julian D. Watts§

Proteomic approaches to biological research that will prove the most useful and productive require robust, sensitive, and reproducible technologies for both the qualitative and quantitative analysis of complex protein mixtures. Here we applied the isotope-coded affinity tag (ICAT) approach to quantitative protein profiling, in this case proteins that copurified with lipid raft plasma membrane domains isolated from control and stimulated Jurkat human T cells. With the ICAT approach, cysteine residues of the two related protein isolates were covalently labeled with isotopically normal and heavy versions of the same reagent, respectively. Following proteolytic cleavage of combined labeled proteins, peptides were fractionated by multidimensional chromatography and subsequently analyzed via automated tandem mass spectrometry. Individual tandem mass spectrometry spectra were searched

ing the rapid evaluation of large proteomic datasets possible. Finally, by repeating the experiment, information relating to the general reproducibility and validity of this approach to large-scale proteomic analyses was also obtained. *Molecular & Cellular Proteomics* 2:428–442, 2003.

A main objective of proteomics research is the systematic identification and quantification of the proteins expressed in a cell, or contained within a cell compartment or other protein complex. The common approach to quantitative protein analysis to date has been the combination of protein separation, most commonly high-resolution two-dimensional polyacrylamide gel electrophoresis (2DE)[1] and tandem mass spectrometry (MS/MS). For this approach, protein identification is

P. Von Haller *et al.* Molecular & Cellular Proteomics (2003)

Conclusions

- consistent and transparent analysis;
  reduce if not eliminate need for manual validation

need to make tools available (and published)!

- known error rates → comparison between different experiments becomes possible

- facilitates publishing large-scale datasets

  - publish entire (minimally filtered) datasets of protein and peptide identifications together with statistically computed confidence measures

  - provide access to raw data (e.g., MS/MS spectra)

Publishing Large-scale Datasets

The Need for Guidelines in Publication of Peptide and Protein Identification Data

WORKING GROUP ON PUBLICATION GUIDELINES FOR PEPTIDE AND PROTEIN IDENTIFICATION DATA*

Steven Carr‡§, Ruedi Aebersold¶, Michael Baldwin‖, Al Burlingame‖, Karl Clauser, and Alexey Nesvizhskii¶**

Over the past few years, the number and size of proteomic datasets composed of mass spectrometry-derived protein identifications reported in the literature have grown dramatically. This is a direct result of the widespread availability of instruments, methods, and easy-to-use software for collecting large amounts of data and for converting the observed peptide and fragment-ion masses to peptide and then protein identities. In particular, the analysis of samples containing large numbers of proteins by multidimensional liquid chromatography (LC/LC)[1] coupled on-line with tandem mass spectrometry (MS/MS) is now a common component of many biological projects. Clearly it is in the interest of the scientific

the most probable peptide assignment to the top of the "hit" list. In addition, new filtering criteria are being developed that, when layered onto the results from the above algorithms, help to eliminate a certain additional percentage of false positives (3, 4). It is very important that the users of these tools, our authors, have at least a working understanding of how the algorithm they use works. However, even the judicious use of scoring, threshold parameters, and additional filtering criteria for search engines, while serving the very important purpose of reducing the number of misassigned peptides and proteins, does not eliminate the problem. It is almost always possible to match a MS/MS spectrum to a peptide in

Molecular & Cellular Proteomics **3**, 532-534 (2004)

Acknowledgements

Ruedi Aebersold
Andrew Keller Jimmy Eng
David Goodlett Eugene Yi Sam Donahue
Julian Watts Jeff Ranish
Dan Martin Victor Ng
Serg Stolyar E. Kolker

Funding: NHLBI Proteomics Center at ISB

www.proteomecenter.org