Omics (2024) – Proteomics data: identification and quantification

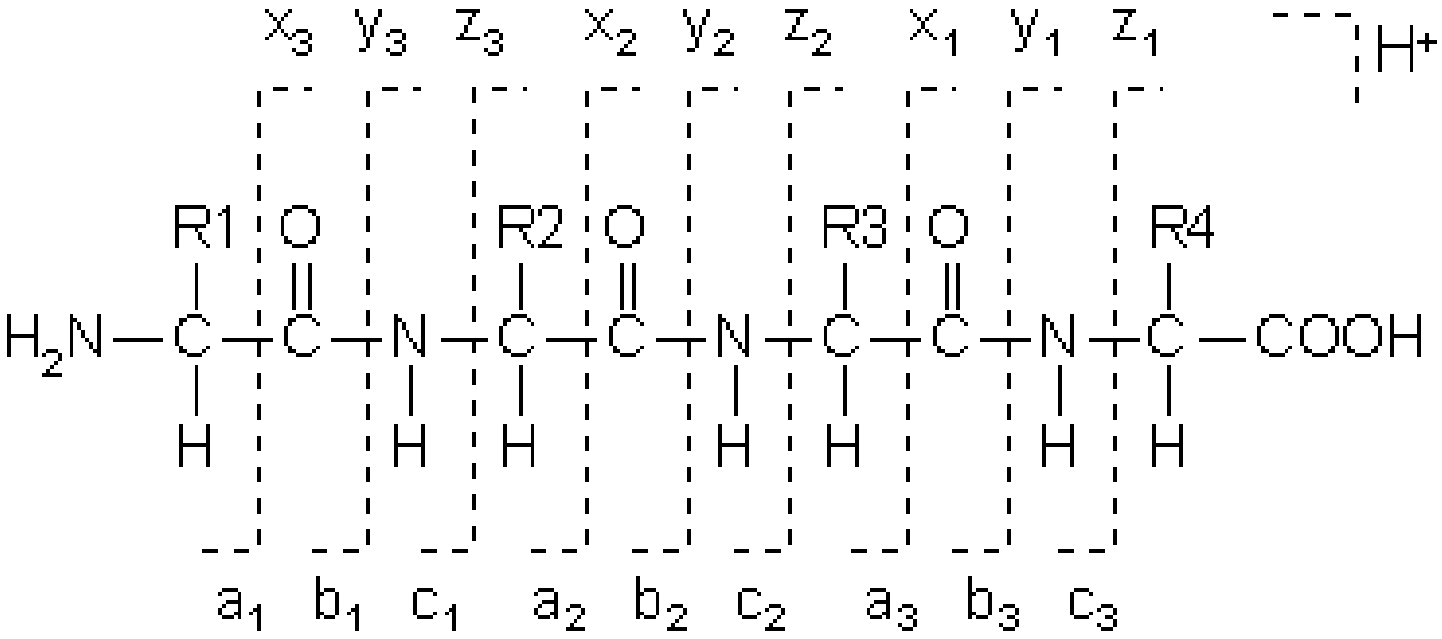
FREDRIK.LEVANDER@IMMUN.LTH.SE

Learning goals

Be able to describe schematically:

- Principles of MS/MS peptide identification
- Important parameters for MS/MS identification
- How to control identification error rates
- Obstacles of protein identification and quantification using peptide data
- Principles of quantification using MS1 data
- Spectral libraries – pros and cons
- DIA vs DDA data processing

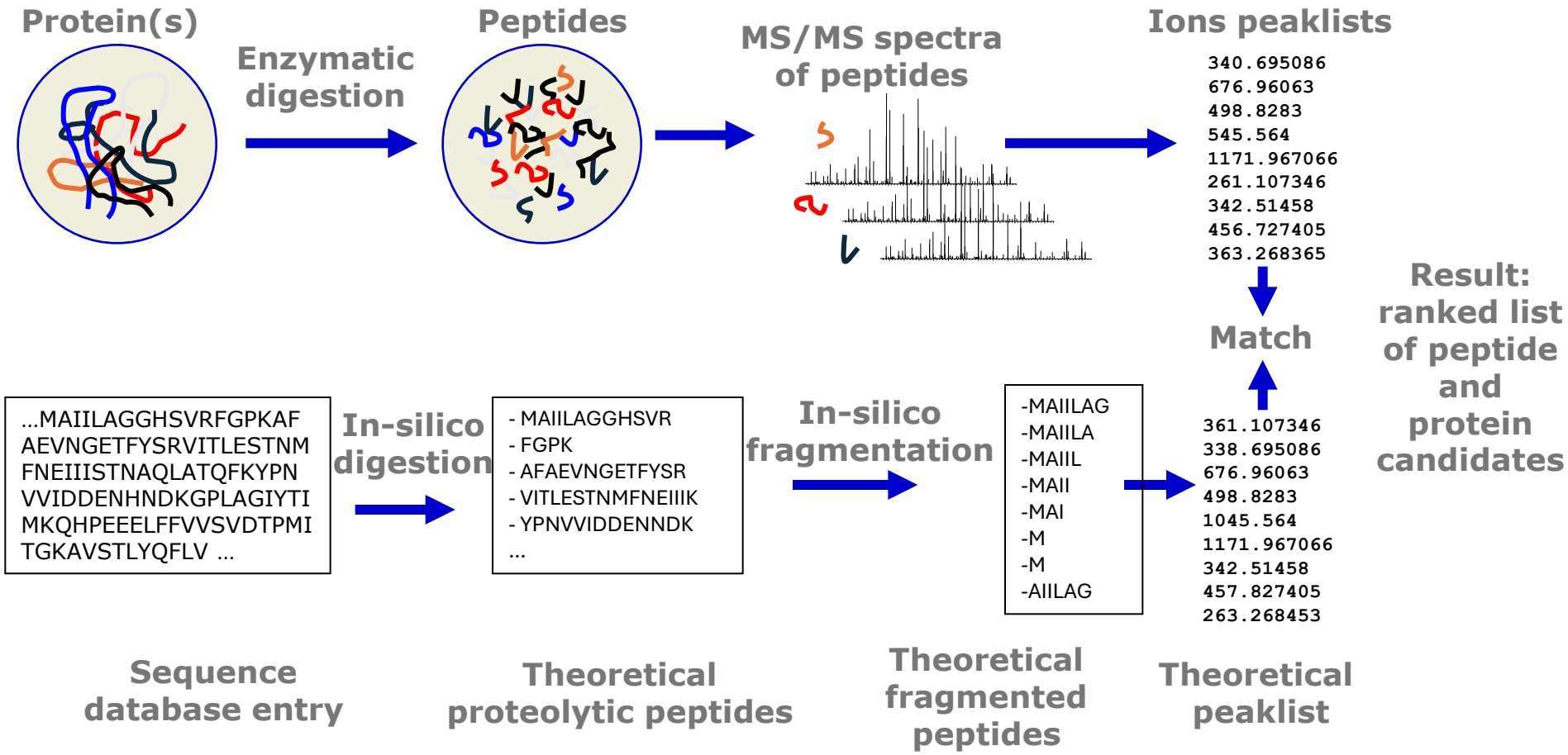
MS/MS peptide fragmentation in mass spectrometer collision cell



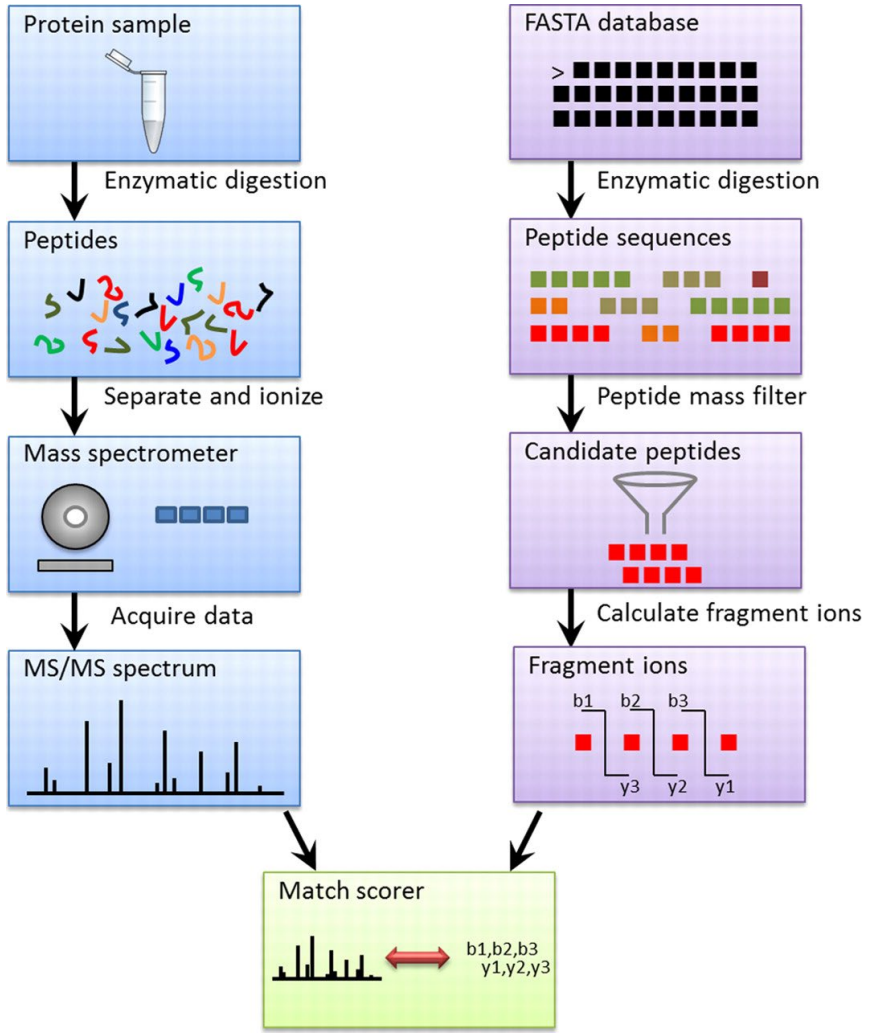
(Plus internal ions, immonium ions, loss of water etc.)

How to identify peptides in MS/MS spectra?

Peptide fragmentation fingerprinting (PFF)
= ion search = MS/MS database matching



Data acquisition and database search.



Eng J K et al. Mol Cell Proteomics 2011;10:R111.009522

Tutorial

Protein identification using MS/MS data ☆

John S. Cottrell · ✉

Matrix Science Ltd., London, UK

Received 2 February 2011, Accepted 9 May 2011, Available online 15 May 2011

Show less

doi:10.1016/j.jprot.2011.05.014

Some PFF tools

Many tools developed over the years

Software	Source website
InsPecT	peptide.ucsd.edu/inspect.py
Mascot	www.matrixscience.com/search_form_select.html
MS-Tag and MS-Seq	prospector.ucsf.edu
PepFrag	prowl.rockefeller.edu/prowl/pepfragch.html
Phenyx	phenyx.vital-it.ch
Popitam	www.expasy.org/tools/popitam
ProID (download)	sashimi.sourceforge.net/software_mi.html
Sequest*	fields.scripps.edu/sequest/index.html
Sonar	65.219.84.5/service/prowl/sonar.html
SpectrumMill*	www.home.agilent.com
VEMS	www.bio.aau.dk/en/biotechnology/vems.htm
X!Tandem (download)	www.thegpm.org/TANDEM

New additions: MSGF+ , Andromeda, Myrimatch , Crux , Tide , MS Amanda, OMSSA

*Commercial

Non exhaustive list!

Important parameters for sucessful peptide identification

- Search engines match theoretical spectra with experimental spectra.
- Main differences between engines are in how matching peaks are scored.
- Search parameters are essential.

Protein needs to be in the search database to be found

- The search database (typically a FASTA file with protein sequences needs to be provided)
- UniProt is the most used resource for protein sequences and FASTA files can be downloaded for many organisms. It consists of:
  - SwissProt (reviewed sequences) with the prefix sp
  - TrEMBL (Automatically Translated EMBL sequences) with the prefix tr

Extract from UniProt human proteome

```
>sp|O60888|CUTA_HUMAN Protein CutA OS=Homo sapiens OX=9606 GN=CUTA PE=1 SV=2
MSGGRAPAVLLGGVASLLLSFVWMPALLPVASRLLLLPRVLLTMASGSPPTQPSPASDSG
SGYVPGSVSAAFVTCPNEKVAKEIARAVVEKRLAACVNLIPQITSIYEWKGKIEEDSEVL
MMIKTQSSLVPALTDFVRSVHPYEVAEVIALPVEQGNFPYLQWVRQVTESVSDSITVLP
>sp|P24310|CX7A1_HUMAN Cytochrome c oxidase subunit 7A1, mitochondrial OS=Homo sapiens OX=9606 GN=COX7A1 PE=1 SV=2
MQALRVSQALIRSFSSTARNRFQNRVREKQKLFQEDNDIPLYLKGGIVDNILYRVTMTLC
LGGTVYSLYSLGWASFPRN
>sp|P14406|CX7A2_HUMAN Cytochrome c oxidase subunit 7A2, mitochondrial OS=Homo sapiens OX=9606 GN=COX7A2 PE=1 SV=1
MLRNLLALRQIGQRTISTASRRHFKNKVPEKQKLFQEDDEIPLYLKGGVADALLYRATMI
LTVGGTAYAIYELAVASFPKKQE
>sp|Q9NTM9|CUTC_HUMAN Copper homeostasis protein cutC homolog OS=Homo sapiens OX=9606 GN=CUTC PE=1 SV=1
MKRQGASSERKRARIPSGKAGAANGFLMEVCVDSVESAVNAERGGADRIELCSGLSEGGT
TPSMGVLQVVKQSVQIPVFVMIRPRGGDFLYSDREIEVMKADIRLAKLYGADGLVFGALT
EDGHIDKELCMSLMAICRPLPVTFHRAFDMVHDPMAALETLLTLGFERVLTSGCDSSALE
GLPLIKRLIEQAKGRIVVMPGGGITDRNLQRILEGSGATEFHCSARSTRDSGMKFRNSSV
AMGASLSCSEYSLKVTDVTKVRTLNAIAKNILV
```

Limitations of Fasta databases

- Sequence variants not included by default -> will miss peptides with sequence deviations at the individual level.
- Signal peptides included in the database sequence, but normally not in the cell…
- Variable splicing can generate different proteins.
- Such info and information about post-translational modifications may be present in UniProt but not be used by search engines as plain FASTA used for the search.

Protein modifications needs to be considered

- *In vivo* modifications. Phosphorylation etc. Mod line in Swissprot
- *In vitro* modifications. Cysteins reduced and alkylated in sample preparation giving rise to carbamidomethylation.
- Needs to be considered as they change the mass of the peptide and fragments
- For example cystein carbamidomethylation gives mass shift of 57.02146 Da.

Missed cleavages

- Trypsin (and other endoproteoases) do not cleave completely. Need to consider peptides with missing cleavages as well.
- Search setting: usually one or two missed cleavages considered

Example (Trypsin cleaves after R and K):

>sp|P24310|CX7A1_HUMAN Cytochrome c oxidase subunit 7A1, mitochondrial OS=Homo sapiens OX=9606 GN=COX7A1 PE=1 SV=2

MQALRVSQALIRSFSSTARNRFQNRVREKQKLFQEDNDIPLYLKGGIVDNILYRVTMTLCLGGTVYSLYSL GWASFPRN

MQALR no missed cleavage

MQALRVSQALIR one missed cleavage

MQALRVSQALIRSFSSTAR two missed cleavages.

Mass tolerances

- As m/z values of peaks measured by a mass spectrometer are not exact, a certain tolerance is needed for matchng
- A larger mass tolerance increases the search space as more possible matches.
- Tolerance at precursor level -> affecting how many peptide candidates that need to be scored
- Tolerance at fragment level -> affecting how many fragments match
- Well calibrated instruments with high mass accuracy allows for low mass tolerance settings.
- Tolerances given as Da or ppm. A mass tolerance of 10 ppm equals 0.01 Da at m/z 1000. And at 400 m/z 10ppm is +/- 0.004
  - ppm mass tolerances often better corresponding to instrument characteristics

What do the search engines do?

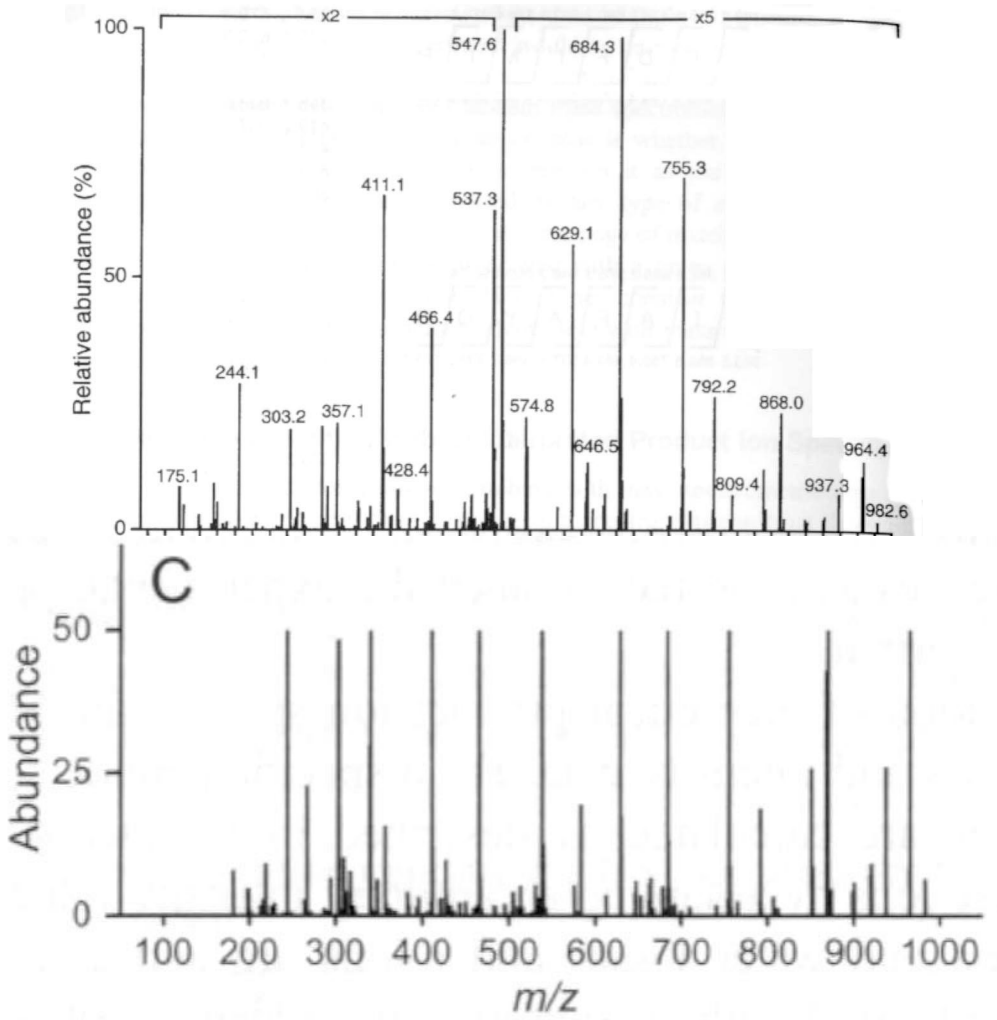
- Sequest example (first developed MS/MS search engine)
- No need to learn the specifics of the algorithm.

Sequest / Turbo Sequest

Algorithm (peptide matching)

1. Experimental spectrum: data reduction
   - Convert m/z ratios to their nominal values (nearest integer)
   - Remove all but the 200 most abundant ions
   - Divide spectrum into 10 evenly spaced sub-sections
   - In each sub-section: normalize ions to the most abundant ion in the sub-section, which is given value of 50

Data reduction



Experimental spectrum

Processed spectrum

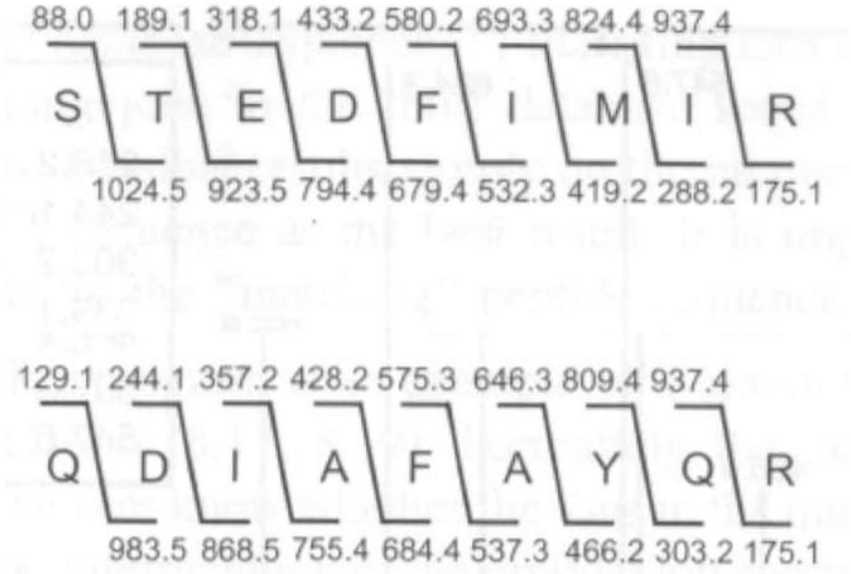
Sequest – Algorithm (II)

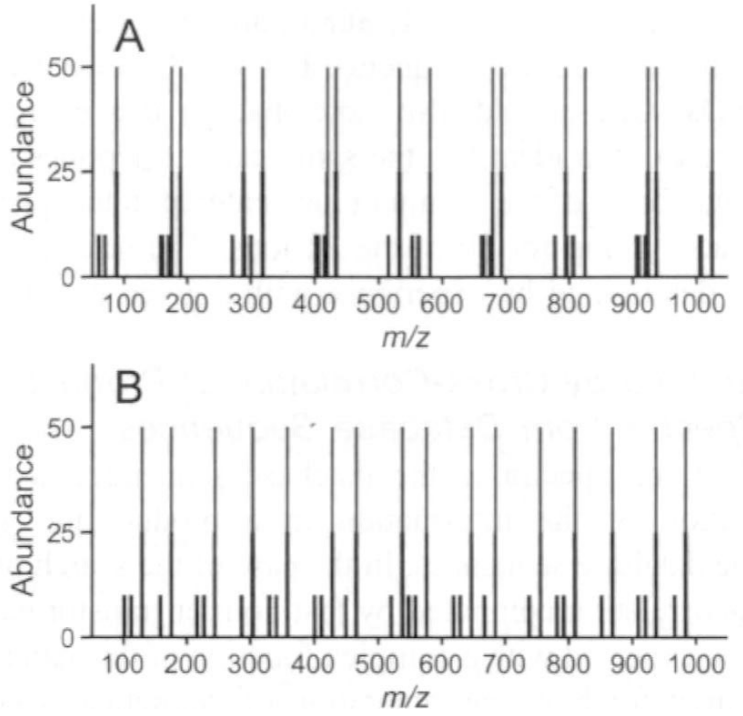
2. Build theoretical spectra

   For each peptide sequence:

   - Calculate b- and y-ions from sequence; assign intensity value of 50
   - Calculate neutral loss ions for ammonia, water losses and a-ions; assign values of 10
   - Add width to b- and y-ions: assign value of 25 to *m/z* $\pm$ 1 from calculated b- and y-ions
   - Same for neutral loss and a-ions; assign value of 10

Build theoretical spectra – examples





Sequest – Algorithm (III)

3. Candidates selection

   For each theoretical spectrum:

   - Compute simple correlation with processed experimental spectrum (*Sp* score ≅ Shared Peak Counts [with intensities])
   - Keep 500 top ranked spectra (peptides)

Sequest – Algorithm (IV)

4. Compute cross-correlation (XCorr)

   For each of 500 theoretical spectra:

   - Compute XCorr correlation
   - Sort peptides by XCorr result
   - Top one is best peptide match

Sp	XCorr	deltCn	Peptide	Reference
795.4	2.2530	0.0000	R.QDIAFAYQR.R	162779
416.5	1.9844	0.1192	R.ENITLIDHR.N	2650045
472.4	1.8290	0.1882	K.ELLAAFYRK.H	3874358
407.3	1.7757	0.2118	K.EIDSQKTYK.T	4704782
317.0	1.6542	0.2658	K.GWALFRSFK.A	15769

   - deltCr

Sp score

$$S_p = \left(\sum_k I_k\right) m(1+\beta)(1+\rho)/L$$

- where I_k = intensity of matched peak k,
- m = number of matches,
- β = reward for consecutive match of an ion series,
- ρ = reward for presence of immonium ion
- L = number of theoretical ions

Xcorr score

$$Corr(E,T) = \sum_{I=0}^{N-1} x_i y_{i+\tau}$$

- where E = experimental spectrum
- T = theoretical spectrum
- x_i and y_i are intensity values of E and T
- τ = displacement value
- N = number of data points in spectrum
- Cross-correlation is computed using a Fast Fourier Transform (FFT)

Sequest/TurboSequest

- Transformation of experimental masses: intensities are normalised, low-intensity peaks are removed, m/z values are round to next integer.
- Virtual digestion of the database which considers also combinations of modifications.
- Match between the theoretical and experimental masses and computation of the score:

$$S_p = \left(\sum_k I_k \right) m(1+\beta)(1+\rho)/L$$

- Reconstruction of a virtual spectrum of the 500 best peptide sequences with three different intensity values.
- Selection of 500 best peptides.
- Cross-correlation: Fourier transform of both signals (experimental and virtual) and multiplication of results.

Sequest output

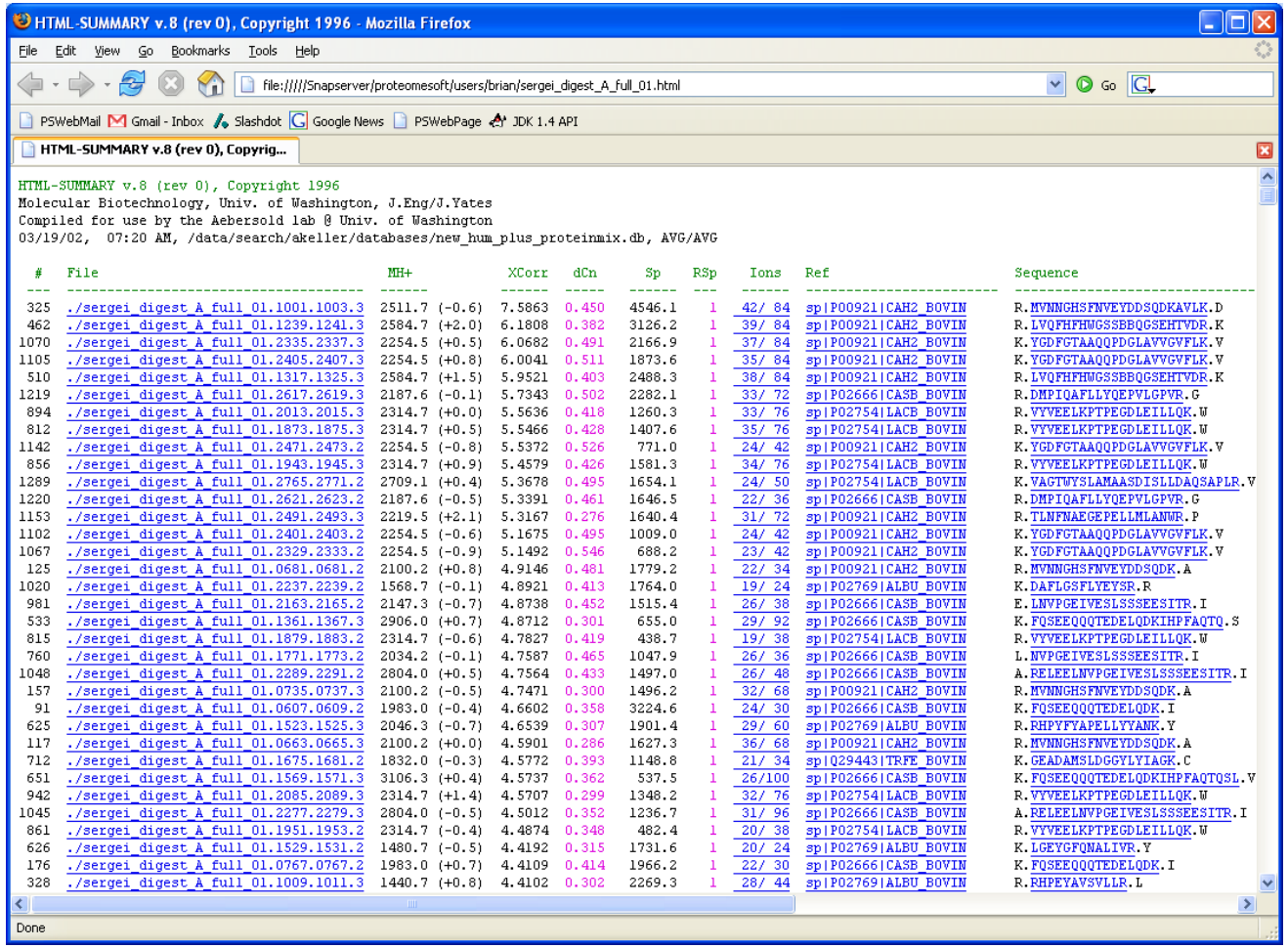
```
al941203#2.0402.0406.2.out
SEQUEST v.B22, Copyright 1993-95
Molecular Biotechnology, Univ. of Washington, J.Eng/J.Yates
Licensed to John Yates' Lab @ Univ. of Washington
11/15/95, 08:53 AM, 2 min. 32 sec. on thompson
mass=1472.0(+2), fragment tol.=0.00, mass tol.=3.00, AVG
# amino acids = 2904160, # proteins = 6254, # matched peptides = 151688
immonium (HFYWM) = (00000), total inten. = 6927.9, lowest Sp = 170.3
ion series nA nB nY ABCDVWXYZ: 0 1 1 0.5 1.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0
rho=0.200, beta=0.075, top 10, /wfs/dbase/OWL/yeast

 #    Rank/Sp   (M+H)+     Cn     delCn   C*10^4    Sp      Ions   Reference         Peptide
---   -------   ------   ------   ------   ------   ----    ----   ---------         -------
 1.    1 /   1  1471.7   1.0000   0.0000   3.8603   851.3  22/39   G3P1_YEAS+4    (R)VPTVDVSVVDLTVK
 2.    2 /   8  1469.7   0.6042   0.3958   2.3323   381.5  16/39   S52527         (L)QAPPPPPSSTKSKF
 3.    3 /   2  1472.9   0.5877   0.4123   2.2688   448.7  17/39   KEX1_YEAS      (A)VVVTIVTFLIVVLG
 4.    4 /   9  1469.6   0.5573   0.4427   2.1515   378.5  17/39   CB31_YEAS      (R)VPMTGDLSTGNTFE
 5.    5 /  12  1471.8   0.5356   0.4644   2.0677   368.2  17/39   ODPA_YEAS      (S)VKAVLAELMGRRAG

 1.  G3P1_YEAST GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE 1 (EC 1.2.1.12). - SACCHAROMYCES
     CEREVISIAE (BAKER'S YEAST).
 2.  S52527 hypothetical protein - yeast (Saccharomyces cerevisiae)
 3.  KEX1_YEAST CARBOXYPEPTIDASE KEX1 PRECURSOR (EC 3.4.16.6) (CARBOXYPEPTIDASE D). - SAC
     CHAROMYCES CEREVISIAE (BAKER'S YEAST).
```

Problem: How to know if a hit is correct?

Threshold model for filtering identifications – the traditional way

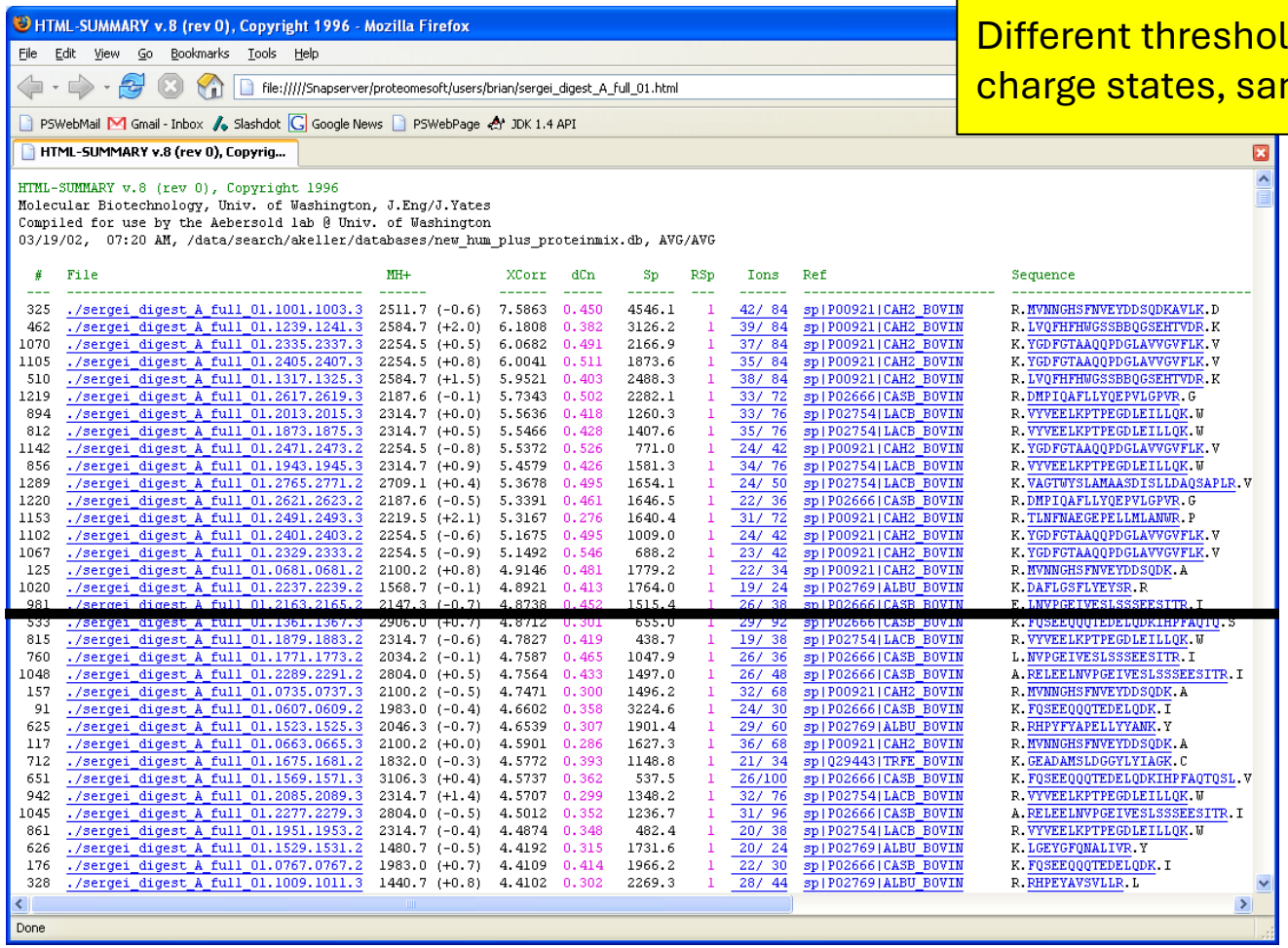


sort by match score

spectrum **scores** **protein** **peptide**

Set a threshold

Next, a threshold value was set. Different programs have different scoring schemes, so SEQUEST, Mascot, and X!Tandem use different thresholds.
Different thresholds may also be needed for different charge states, sample complexity, and database size.



sort by match score

SEQUEST
XCorr > 2.5
dCn > 0.1

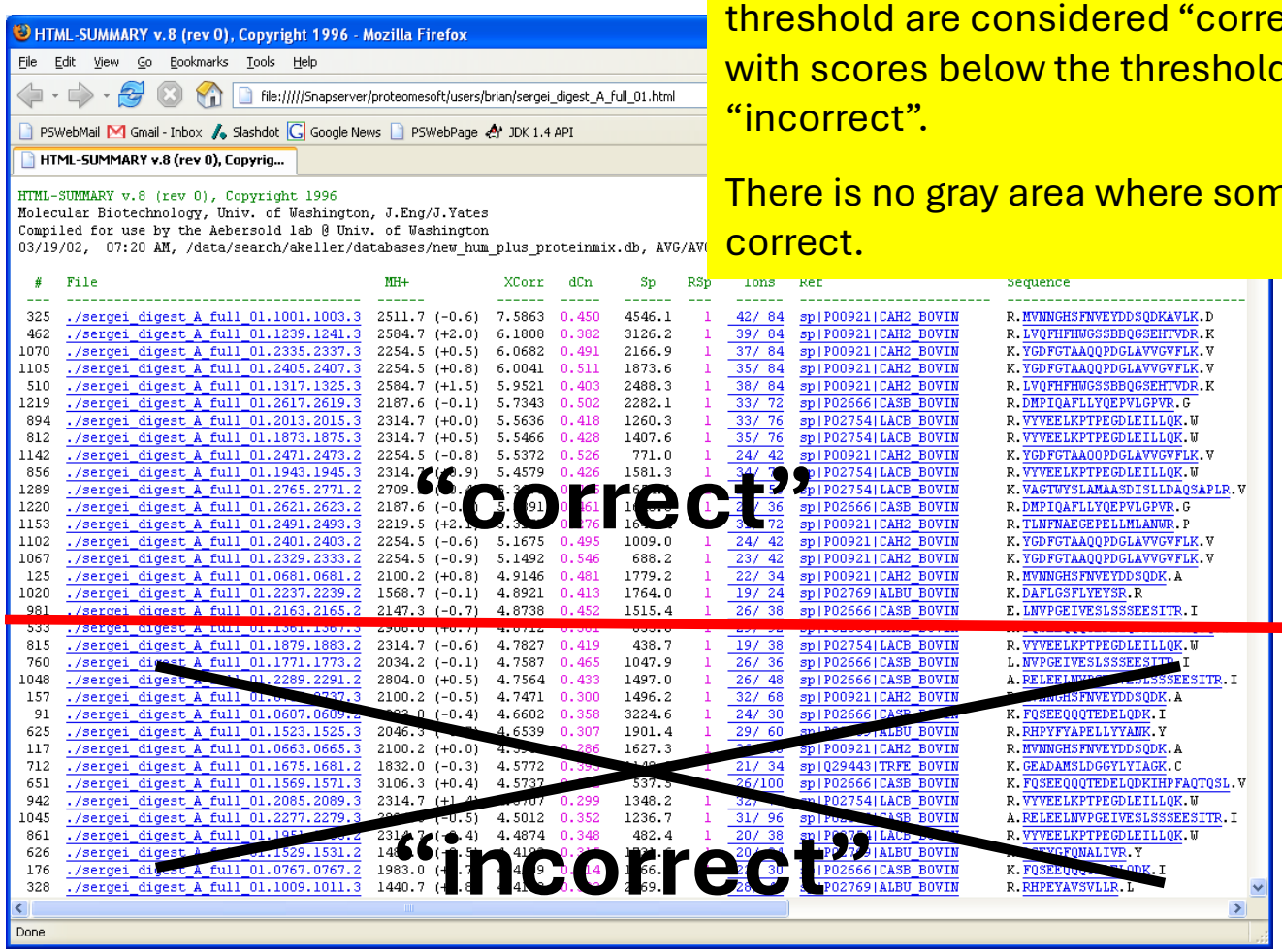
Mascot
Score > 45

peptide

Matches below threshold are dropped

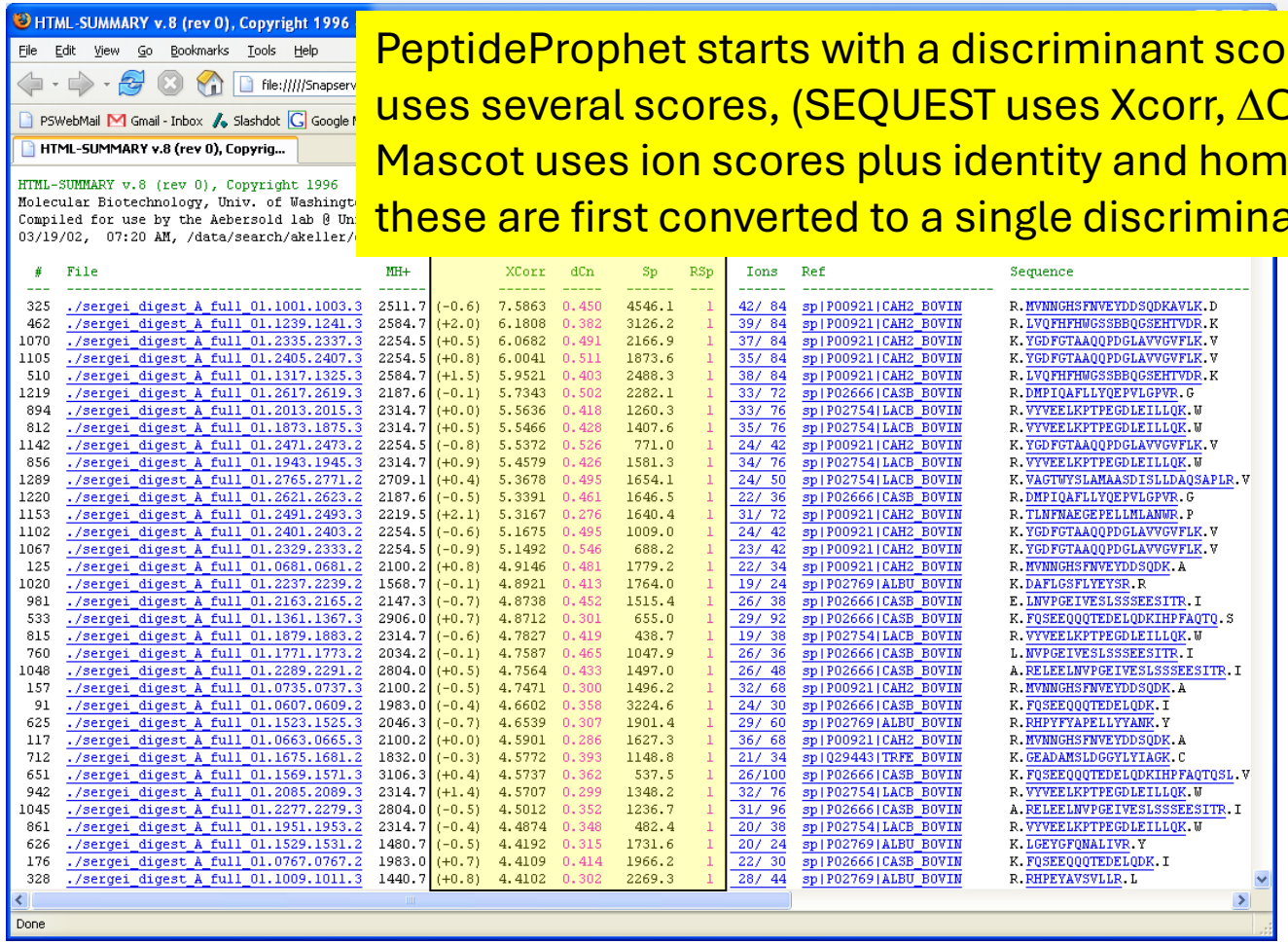
Peptides that are identified with scores above the threshold are considered “correct” matches. Those with scores below the threshold are considered “incorrect”.

There is no gray area where something is possibly correct.



peptide

Creating a discriminant score – the PeptideProphet approach



PeptideProphet starts with a discriminant score. If an application uses several scores, (SEQUEST uses Xcorr, ΔCn, and Sp scores; Mascot uses ion scores plus identity and homology thresholds), these are first converted to a single discriminant score.

sort by match score

peptide

https://doi.org/10.1021/ac025747h
Keller et al, 2002

Discriminant score for SEQUEST – the PeptideProphet approach

$$D = \begin{pmatrix} +8.4* \dfrac{\ln(XCorr)}{\ln(\#AAs)} \\ +7.4* \Delta Cn \\ -0.2*\ln(rankSp) \\ -0.3* \Delta Mass \\ -0.96 \end{pmatrix}$$

For example, here's the formula to combine SEQUEST's scores into a discriminant score:

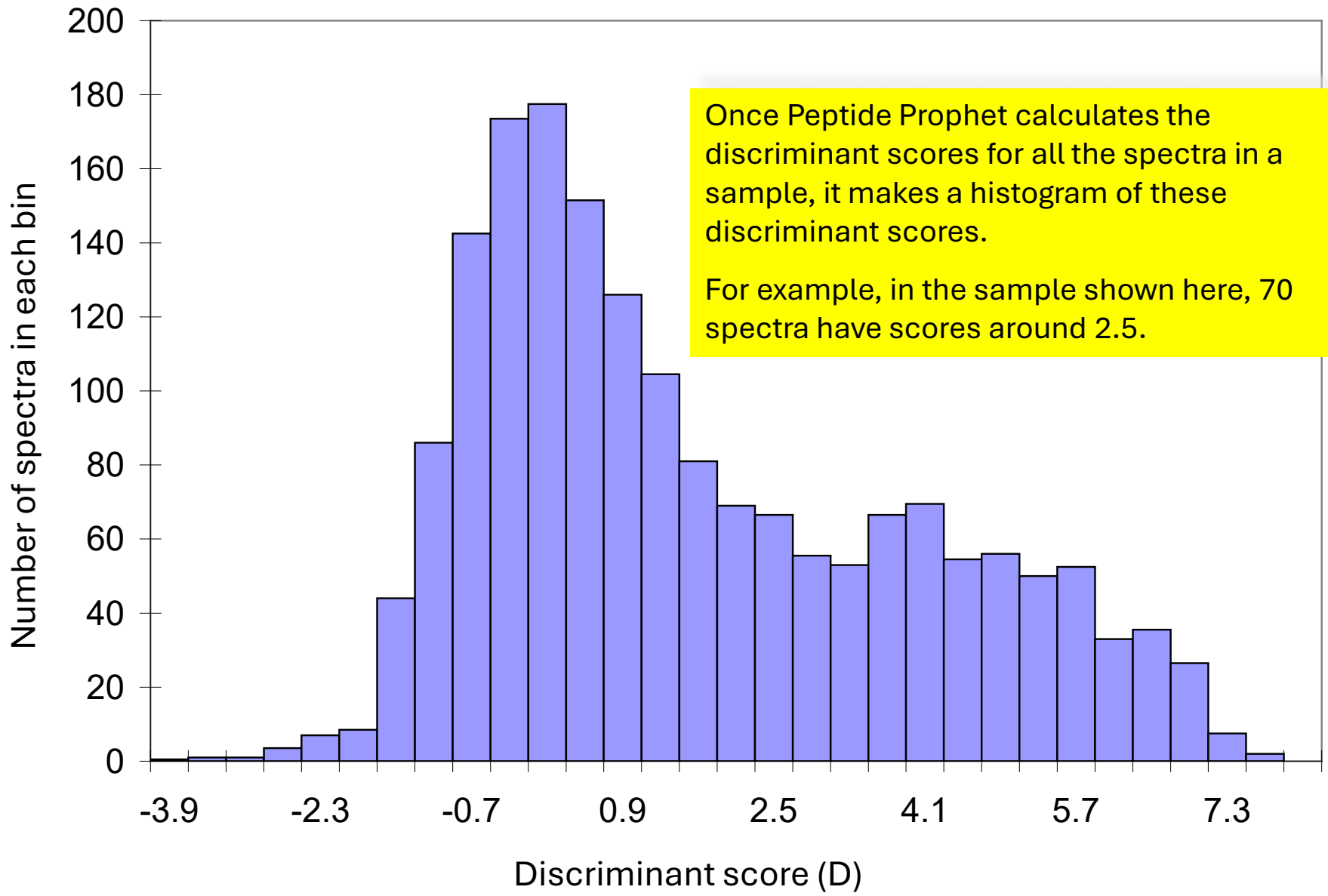
SEQUEST's **XCorr** (correlation score) is corrected for length of the peptide. High correlation is rewarded.

SEQUEST's **ΔCn** tells how far the top score is from the rest. Being far ahead of others is rewarded.

The top ranked by SEQUEST's **Sp** score has **ln(rankSp)**=0. Lower ranked scores are penalized.

Poor mass accuracy (big **ΔMass**) is also penalized.

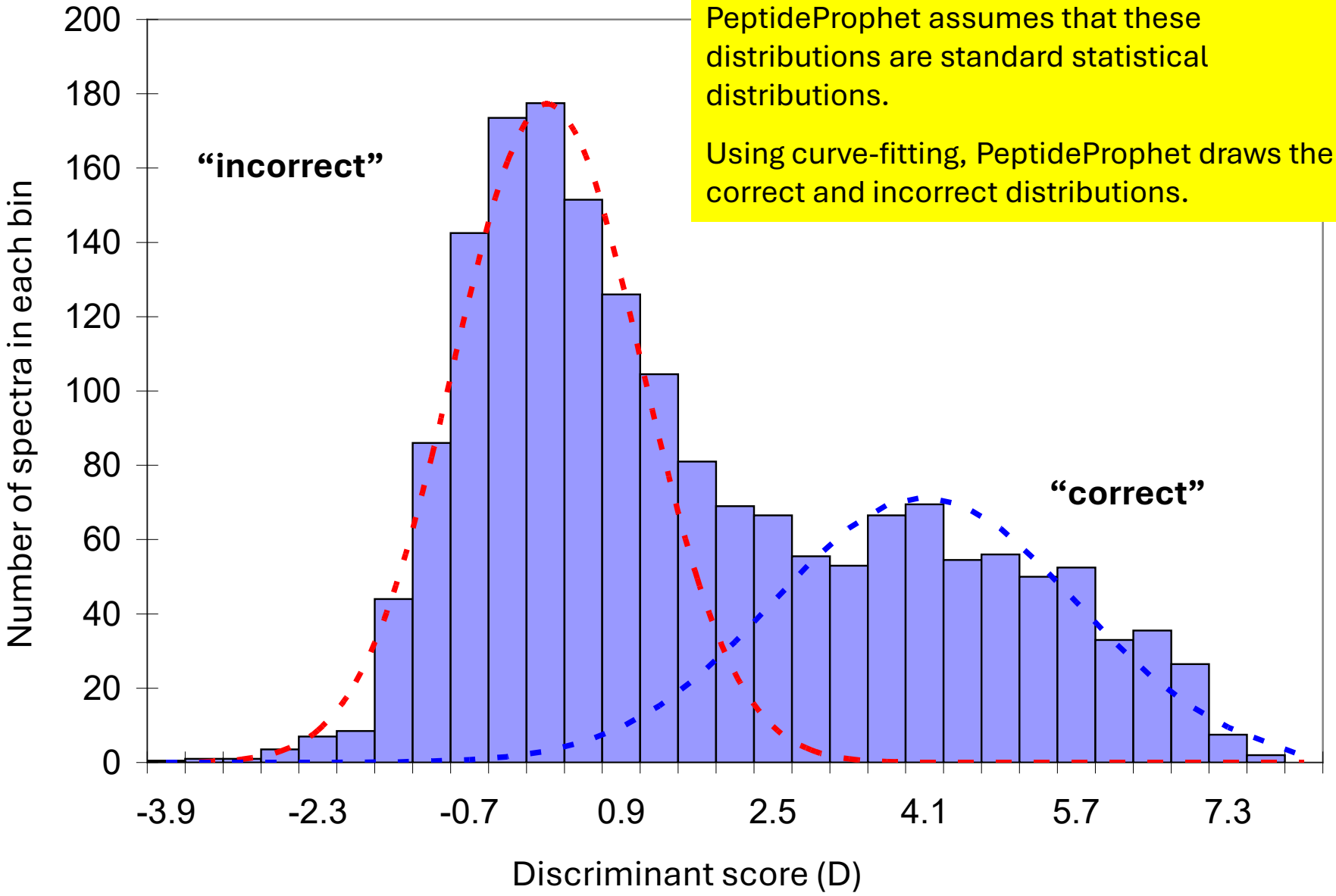
Histogram of scores



Once Peptide Prophet calculates the discriminant scores for all the spectra in a sample, it makes a histogram of these discriminant scores.

For example, in the sample shown here, 70 spectra have scores around 2.5.

Mixture of distributions

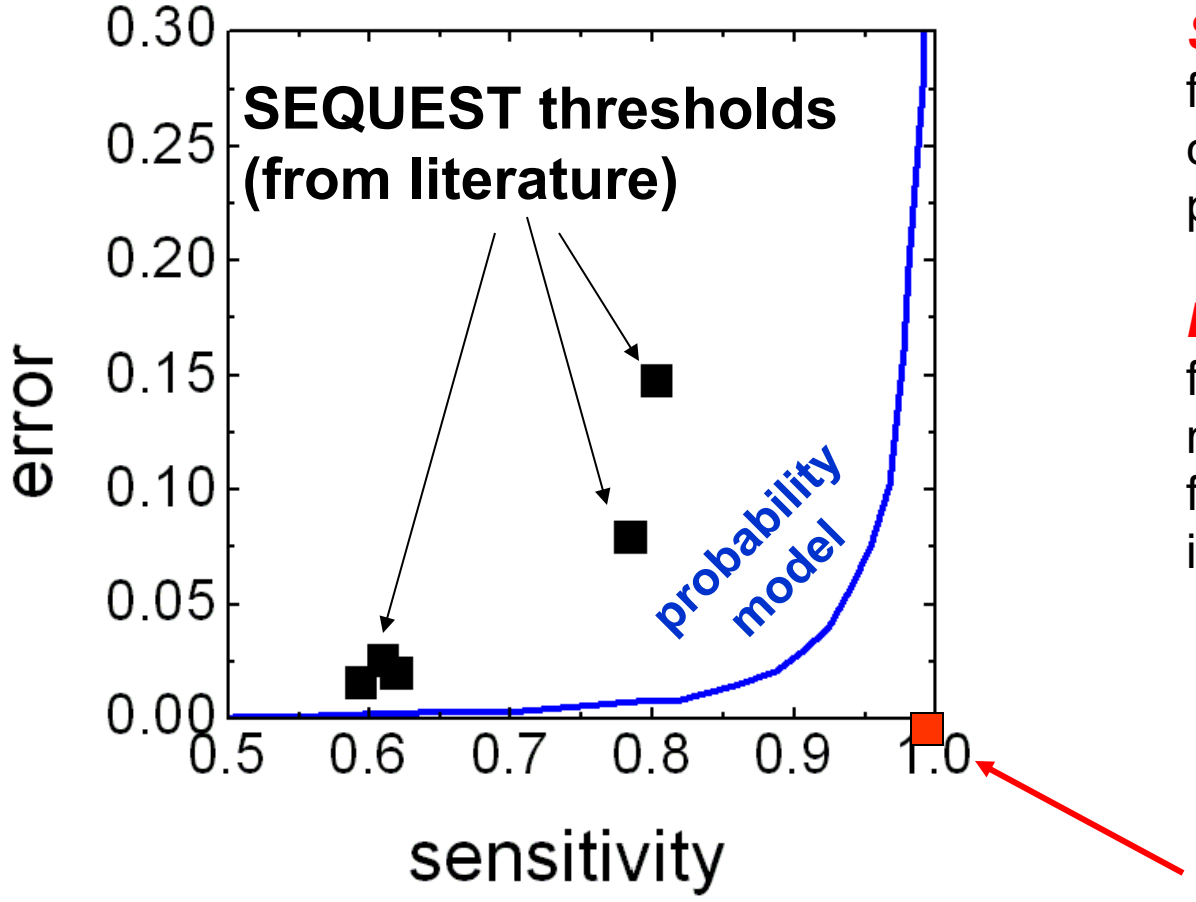


This histogram shows the distributions of correct and incorrect matches.

PeptideProphet assumes that these distributions are standard statistical distributions.

Using curve-fitting, PeptideProphet draws the correct and incorrect distributions.

Discriminating power of Peptide Prophet



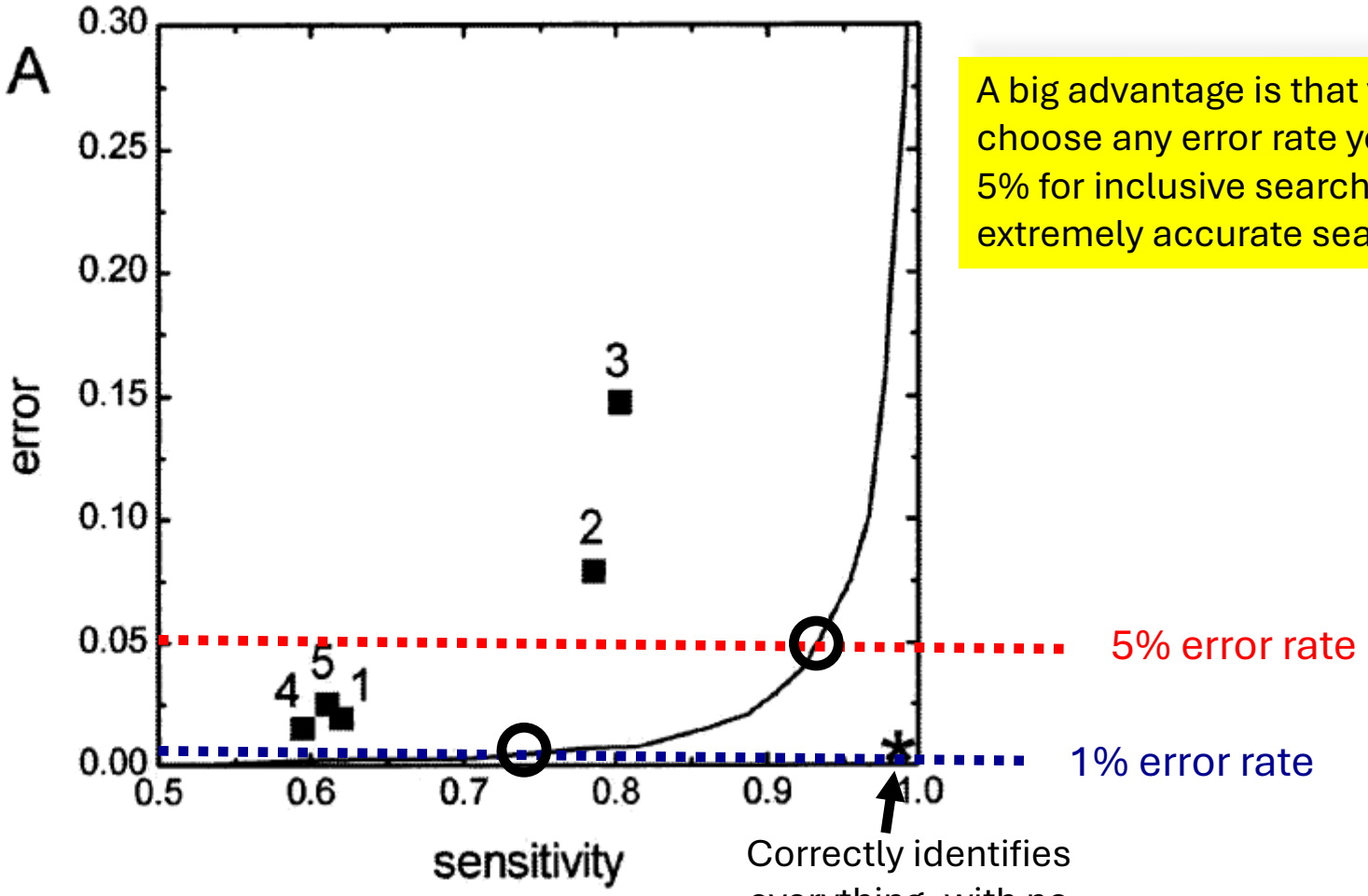
Sensitivity:
fraction of all correct results passing filter

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

Ideal Spot

Improved discrimination: more identifications (at the same error rate)

Choose an error rate with PeptideProphet



A big advantage is that you can choose any error rate you like, such as 5% for inclusive searches, or 1% for extremely accurate searches.

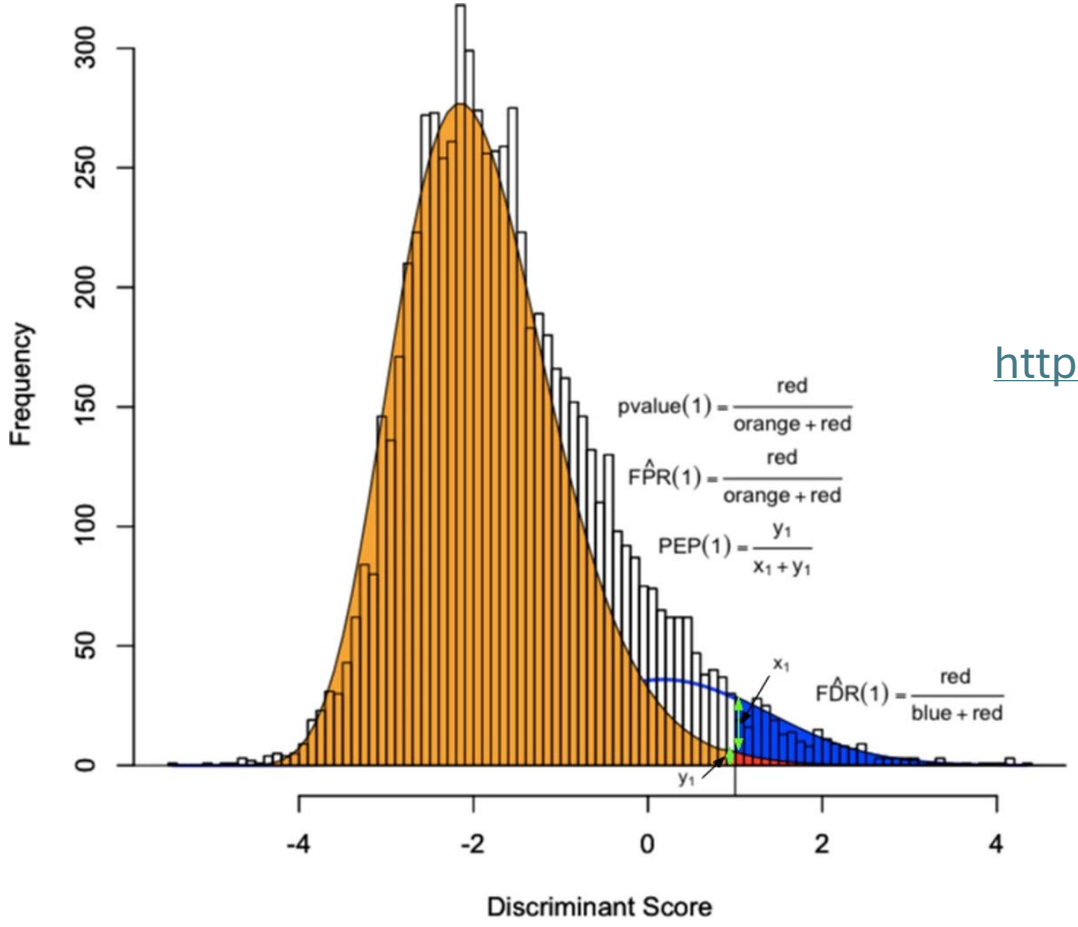
5% error rate

1% error rate

Correctly identifies everything, with no error

Keller et al., Anal Chem 2002

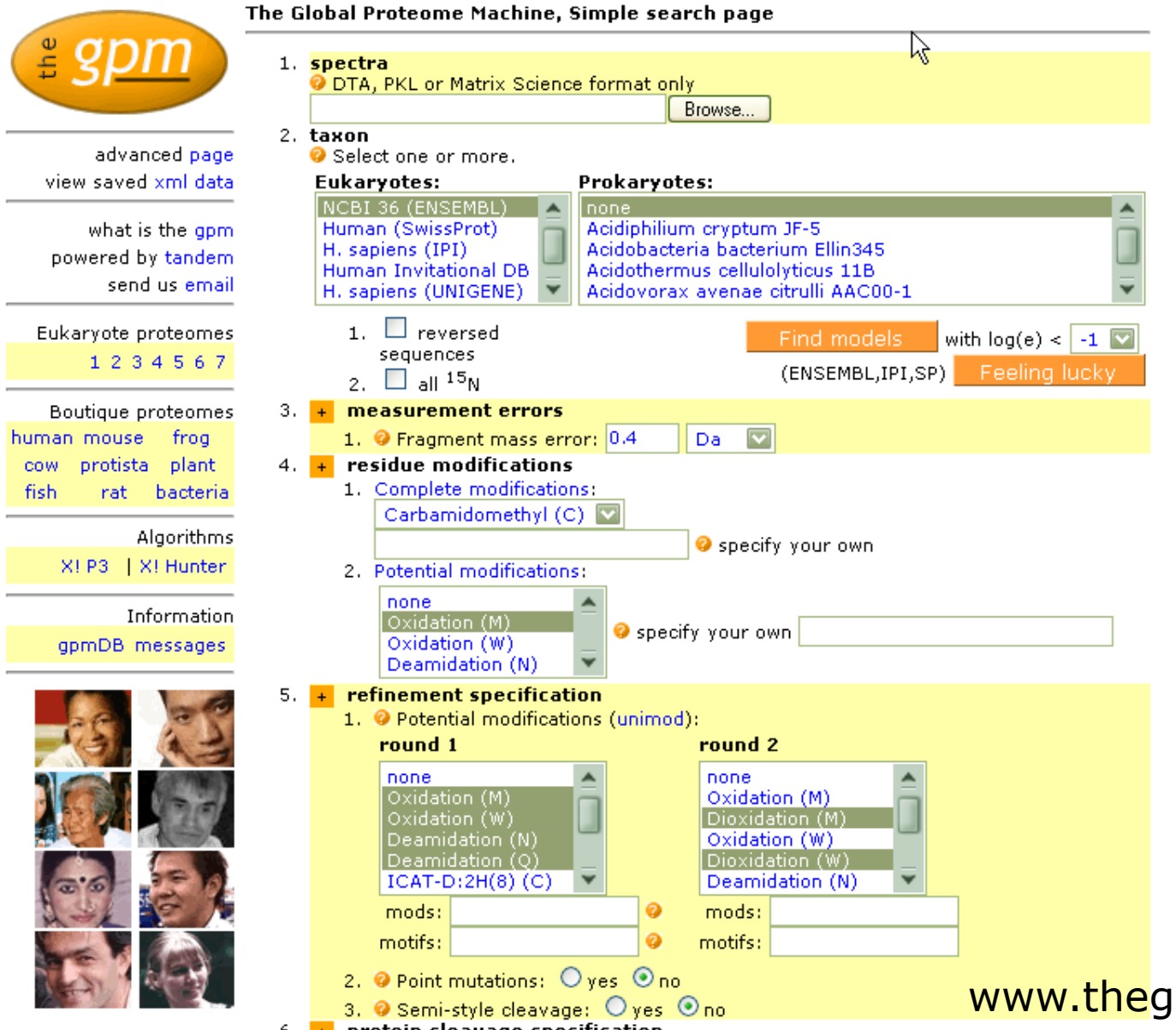
PeptideProphet developments



https://doi.org/10.1186/1471-2105-13-S16-S1

Peptide Prophet still used after further developments,
but FDR-based approaches with decoy databases has mostly taken over. We'll get back to that!

The X!Tandem search engine



www.thegpm.org

X!Tandem

- Calculates statistical confidence (e-values) for all of the individual spectrum-to-sequence assignments

- Reassembles all of the peptide assignments in a data set onto the known protein sequences and assign the statistical confidence that this assembly and alignment is non-random.

- E- values are then transformed to log values to remove the powers of 10

E-values

- For a given score S, it indicates the number of matches that are expected to occur by chance in a database with a score at least equal to S.
- The e-value takes into account the size of the database that was searched. As a consequence it has a maximum of the number of sequences in the database.
- The lower the e-value, the more significant the score is.
- An e-value depends on the calculation of the p-value.

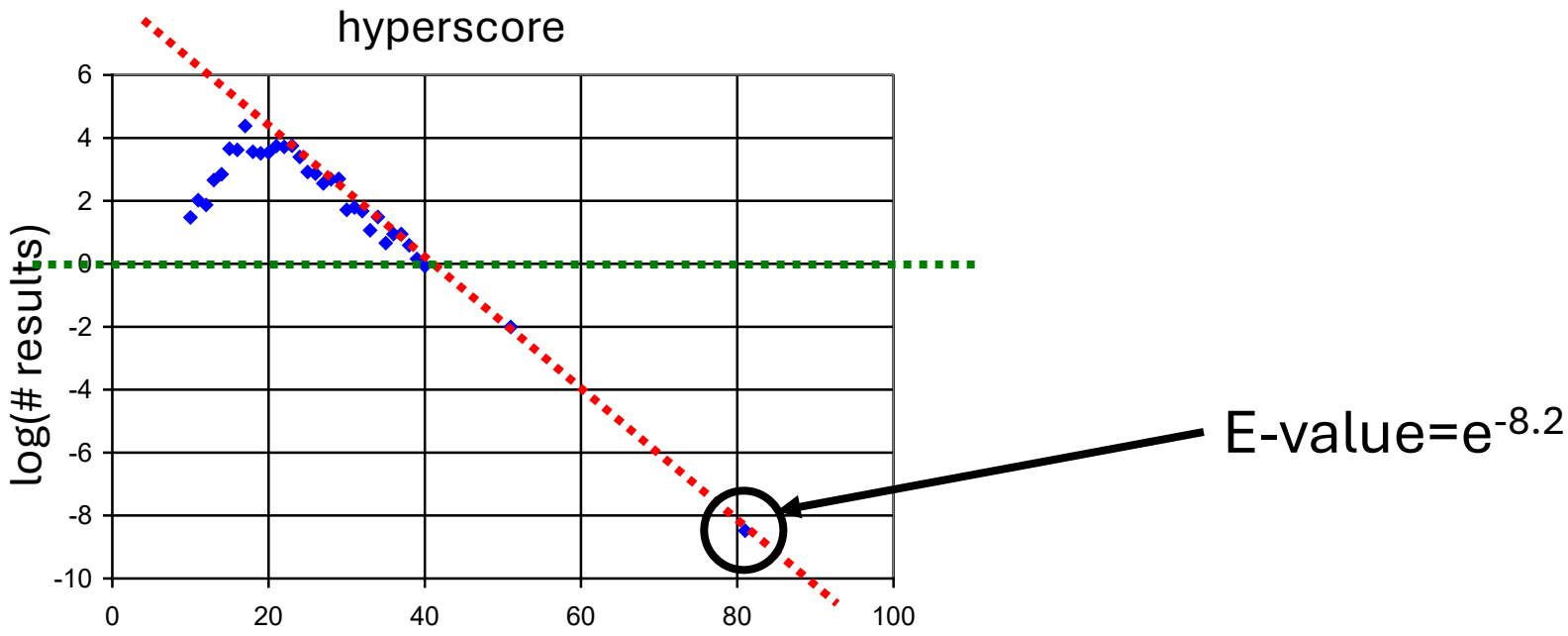
X!Tandem

- Match experimental versus theoretical spectra
- Preliminary score = dot product of experimental versus theoretical spectra (because only similar peaks are considered, this is the sum of the intensities of the matched y and b ions)
- Hyperscore = the preliminary score by multiplying by N factorial for the number of b and y ions assigned
- It makes a histogram of all the hyperscores for all the peptides in the database that might match this particular spectrum
- Log transformation of these values, a line interpolates them
- A match is significant if is greater than the point at which the straight line through the log data intersects the log(#results)=0 line.

Source: Brain Searle (Proteome Software) – XTandem to be explained

Log transformation and e-value

- X!Tandem calculates the E-value by extrapolating the red line of the log histogram.
- For the example shown, a hyperscore of 83 would occur by chance where the red line crosses 83. The log of this value — the E-value — is -8.2, as shown.



Source: Brain Searle (Proteome Software) – XTandem to be explained

X!Tandem - output

Models from 'test_Phenyx01.dta'
Main model display

model | intersect | gel | chip | peptide | table | details | XML |

GPM32100018784 | performance | parameters | FASTA | PRIDE | get annotation

Display: log(e) < -1 and # > 0 , show as | table | html | valid log(e) < -0.9, p-score = 79 (out of 100)

rank	log(e)+	log(I)	%	#	total	Mr	accession
1	-90.2	4.88	40	10	28	32.6	sp|NPM_HUMAN| gpmDB homologues protein Nucleophosmin
2	-2.3	3.13	1.3	1	1	90.1	sp|DACT1_HUMAN| gpmDB homologues protein Dapper homolog 1
3	-1.3	2.57	21	1	1	10.7	sp|KV3C_HUMAN| gpmDB homologues protein

plist.pl, v. 2007.01.31

1

sp|NPM_HUMAN|
protein model: sp|NPM_HUMAN|

model | homologues | details | XML |

| swissprot | gpmDB |

Peptide clustering: **unique** | modified | none *Display:* table | **html** |

sp|NPM_HUMAN|: Nucleophosmin
log(e) = -90.2

(validate)

```
  1 MEDSMDMDMSPLRPQNYLFGCELKADKDYHFKVDNDENEHQLSLRTVSLGAGAKDELHIV  60
 61 EAEAMNYEGSPIKVTLATLKMSVQPTVSLGGFEITPPVVLRLKCGSGPVHISGQHLVAVE 120
121 EDAESEDEEEEDVKLLSISGKRSAPGGGSKVPQKKVKLAADEDDDDDEEDDDEDDDDDD 180
181 FDDEEAEEKAPVKKSIRDTPAKNAQKSNQNGKDSKPSSTPRSKGQESFKKQEKTPKTPKG 240
241 PSSVEDIKAKMQASIEKGGSLPKVEAKFINYVKNCFRMTDQEAIQDLWQWRKSL        294
```

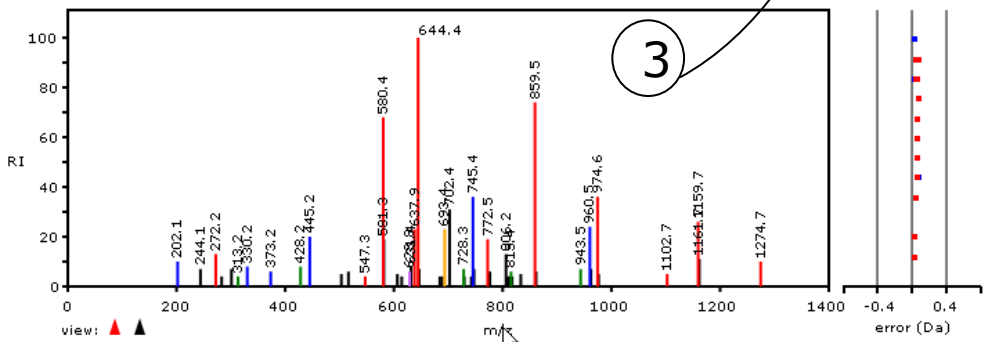
2

#	log(e)	log(I)	m+h	delta	z	sequence | validate | studio
1	-6.8	3.20	1603.767	0.153	2	207 SNQNGKDSKPSSTPR 221

mods: 210N+0.984016

no description

S N Q N G K D S K P S S T P R



3

bond	+1 y	+1 y-17	+1 y-18	+1 b	+1 b-17	+1 b-18
S1	1516.735	1499.708	1498.724	88.039	71.013	70.029
N2	1402.692	1385.666	1384.682	202.082	185.056	184.072
Q3	+1,2 1274.634	1257.607	1256.623	330.141	313.114	312.130
N4	+1,2 1159.607	1142.580	1141.596	445.168	428.141	427.157
G5	1102.585	1085.559	1084.575	502.189	485.163	484.179
K6	974.490	957.464	956.480	630.284	613.258	612.274
D7	859.463	842.437	841.453	+1,2 745.311	728.285	727.301
S8	772.431	755.405	754.421	832.343	815.317	814.333
K9	644.336	627.310	626.326	960.438	943.412	942.428
P10	547.283	530.257	529.273	1057.491	1040.464	1039.480
S11	460.251	443.225	442.241	1144.523	1127.496	1126.512

(e)	log(I)	m+h	delta	z	sequence
.5	2.53	2981.249	0.273	3	[1 MEDSMDMDMS PLRPQNYLFG CELK 24adkd
.1	3.34	2165.813	0.170	2	[1 MEDSMDMDMS PLRPQNY 17lfgc
.7	4.33	1568.730	0.143	2	yhfk33 VDNDENEHQL SLR 45tvsl
.2	3.44	1212.513	0.114	2	yhfk33 VDNDENEHQL 42slrt
.9	2.74	2929.461	0.197	3	lslr46 TVSLGAGAKD ELHIVEAEAM NYEGSPIK 73vtla
2.4	2.57	2161.012	0.149	2	agak55 DELHIVEAEA MNYEGSPIK 73vtla
3.8	3.71	2227.216	0.166	2	atlk81 MSVQPTVSLG GFEITPPVVL R 101lkcg
.8	3.20	1603.767	0.153	2	naqk207 SNQNGKDSKP SSTPR 221skgq
.6	3.83	931.473	0.092	2	ktpk240 GPSSVEDIK 248akmq
.4	2.98	1835.838	0.138	2	ncfr278 MTDQEAIQDL WQWR 291ksl]

s if the information below appears incomplete

'y

Niceprot View | SRS View

Entry: P06748

MATION	
	NPM_HUMAN
	P06748; P08693; O12826; O13440; O13441; O14115; O5EU94; O5EU95; O5EU96;

Source of errors in assigning peptides

- Scores not adapted
- Parameters are too stringent or too loose
- Low MS/MS spectrum quality (many noise peaks, low signal to noise ratio, missing fragment ions, contaminants)
- Homologous peptides
- Incorrectly assigned charge state
- Pre-selection of the 2nd isotope (the parent mass is shifted of 1 Da. A solution is to take the parent mass tol. larger, but may draw the good peptide too)…
- Novel peptide or variant

Hints to know when an identification is correct

- Scores: the higher, the better.
- Better when high intensity peaks are matched and ion series are extended, without too many and too big holes.
- Better when the mass errors are more or less constant among all ions or follow a trend.
- If you have time, try many tools and compare the results

But what to do with large scale data?

Decoy database

- Is used to repeat the search, using identical search parameters, against a database in which the sequences have been reversed or randomised.
- Do not expect to get any real matches from the "decoy" database
- Helps estimate the number of false positives that are present in the results from the real database.
- It is a good validation method for MS/MS searches of large data sets, it is not as useful for a search of a small number of spectra, because the number of matches is too small to give an accurate estimate.
- In Mascot, a random sequence of the same length is automatically generated with the same average amino acid composition if the ‘decoy’ option is selected

Possible Decoy databases

Original sequence

- Reverse: each sequence of the real database is reversed (back to forth)

- Shuffle: each sequence of the real database is shuffled with the same average AA composition

- Random: each sequence of the real database is a new randomised sequence based on the AA composition of the database

False Discovery Rate (FDR) target-decoy strategy

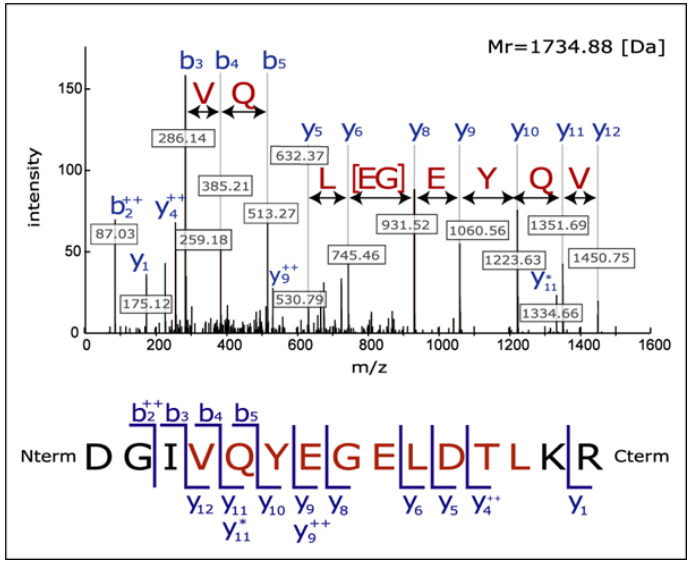
- The expected number of false positive identifications can be estimated by use of decoy databases.
- If searching in a database which is an equal size combination of decoy and true peptides the FDR can be estimated:
- FDR = decoy hits / forward hits above a certain score cutoff
- For example 10 hits with random (decoy) accession numbers and 100 hits with normal (target) accession numbers would indicate an FDR of 0.1 (10% of target hits can be estimated to be falsed discoveries).

Concluding remarks

- A higher score cut off will lead to a lower number of true false identifications, but also to a lower number of true identifications
- For a specific experiment the acceptable rate of false identifications should be determined.
  - Filtering at 1% FDR is frequent
- Small datasets harder to validate

Alternative approaches to MS/MS identification

- De Novo Sequencing:
  - Sequencing = « read » the full peptide sequence out of the spectrum (from scratch)
  - Then, eventually search database for sequence (not necessarily)



Why de novo sequencing is difficult

1. Leucine and isoleucine have the same mass
2. Glutamine and lysine differ in mass by 0.036Da
3. Phenylalanine and oxidized methionine differ in mass by 0.033Da
4. Cleavages do not occur at every peptide bond (or cannot be observed on the MS-MS
    - Poor quality spectrum (some fragment ions are below noise level)
    - The C-terminal side of proline is often resistant to cleavage
    - Absence of mobile protons
    - Peptides with free N-termini often lack fragmentation between the first and second amino acids

Why de novo sequencing is difficult (II)

5. Certain amino acids have the same mass as pairs of other amino acids
   - Gly +Gly (114.0429) Asn (114.0429)
   - Ala +Gly (128.0586) Gln (128.0586)
   - Ala +Gly (128.0586) Lys (128.0950)
   - Gly + Val (156.0742) Arg (156.1011)
   - Ala + Asp (186.0641) Trp (186.0793)
   - Ser + Val (186.1005) Trp (186.0793)
6. Directionality of an ion series is not always known (are they b- or y-ions?)

New approaches

- Open searches. Search for the unknown! These algorithms use clever strategies to allow matching unknown modifications with reasonable speed.

Published: 08 October 2018

Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine

Hao Chi, Chao Liu, Hao Yang, Wen-Feng Zeng, Long Wu, Wen-Jing Zhou, Rui-Min Wang, Xiu-Nan Niu, Yue-He Ding, Yao Zhang, Zhao-Wei Wang, Zhen-Lin Chen, Rui-Xiang Sun, Tao Liu, Guang-Ming Tan, Meng-Qiu Dong, Ping Xu, Pei-Heng Zhang & Si-Min He

Nature Biotechnology **36**, 1059–1061(2018) | Cite this article

Published: 10 April 2017

MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics

Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu & Alexey I Nesvizhskii

Nature Methods **14**, 513–520(2017) | Cite this article

Spectral library searching

- Generate library of spectra with condfidently identified spectra
- Match MS/MS with library instead of predicted spectra.
  - Make use of real world peak intensities allowing for high confidence matching.
  - Drawback is that intensities can vary beween MS platforms and fragmentation methods
  - Only spectra in the library will be identified!

Large scale effort

Building ProteomeTools based on a complete synthetic human proteome

Daniel P Zolg, Mathias Wilhelm, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Bernard Delanghe, Derek J Bailey, Siegfried Gessulat, Hans-Christian Ehrlich, Maximilian Weininger, Peng Yu, Judith Schlegl, Karl Kramer, Tobias Schmidt, Ulrike Kusebauch, Eric W Deutsch, Ruedi Aebersold, Robert L Moritz, Holger Wenschuh, Thomas Moehring, Stephan Aiche, Andreas Huhmer, Ulf Reimer & Bernhard Kuster

Nature Methods **14**, 259–262 (2017) | Cite this article

Abstract

We describe ProteomeTools, a project building molecular and digital tools from the human proteome to facilitate biomedical research. Here we report the generation and multimodal liquid chromatography–tandem mass spectrometry analysis of >330,000 synthetic tryptic peptides representing essentially all canonical human gene products, and we exemplify the utility of these data in several applications. The resource (available at http://www.proteometools.org) will be extended to >1 million peptides, and all data will be shared with the community via ProteomicsDB and ProteomeXchange.

Predicting spectra using AI

Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning

Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, Ulf Reimer, Hans-Christian Ehrlich, Stephan Aiche, Bernhard Kuster ✉ & Mathias Wilhelm ✉

Nature Methods **16**, 509–518 (2019) | Cite this article

AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics

Wen-Feng Zeng, Xie-Xuan Zhou, Sander Willems, Constantin Ammar, Maria Wahle, Isabell Bludau, Eugenia Voytik, Maximillian T. Strauss & Matthias Mann ✉

Nature Communications **13**, Article number: 7238 (2022) | Cite this article

Sequence-to-sequence translation from mass spectra to peptides with a transformer model

Melih Yilmaz, William E. Fondrie, Wout Bittremieux, Carlo F. Melendez, Rowan Nelson, Varun Ananth, Sewoong Oh & William Stafford Noble ✉

Nature Communications **15**, Article number: 6427 (2024) | Cite this article

Prediction of peptide spectra and their LC retention times is boosting identification rates

- Predicted spectra sometimes even better than spectral libraries
- Can adapt between experimental setups using transfer learning

AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics

Wen-Feng Zeng, Xie-Xuan Zhou, Sander Willems, Constantin Ammar, Maria Wahle, Isabell Bludau, Eugenia Voytik, Maximillian T. Strauss & Matthias Mann

Nature Communications **13**, Article number: 7238 (2022) | Cite this article

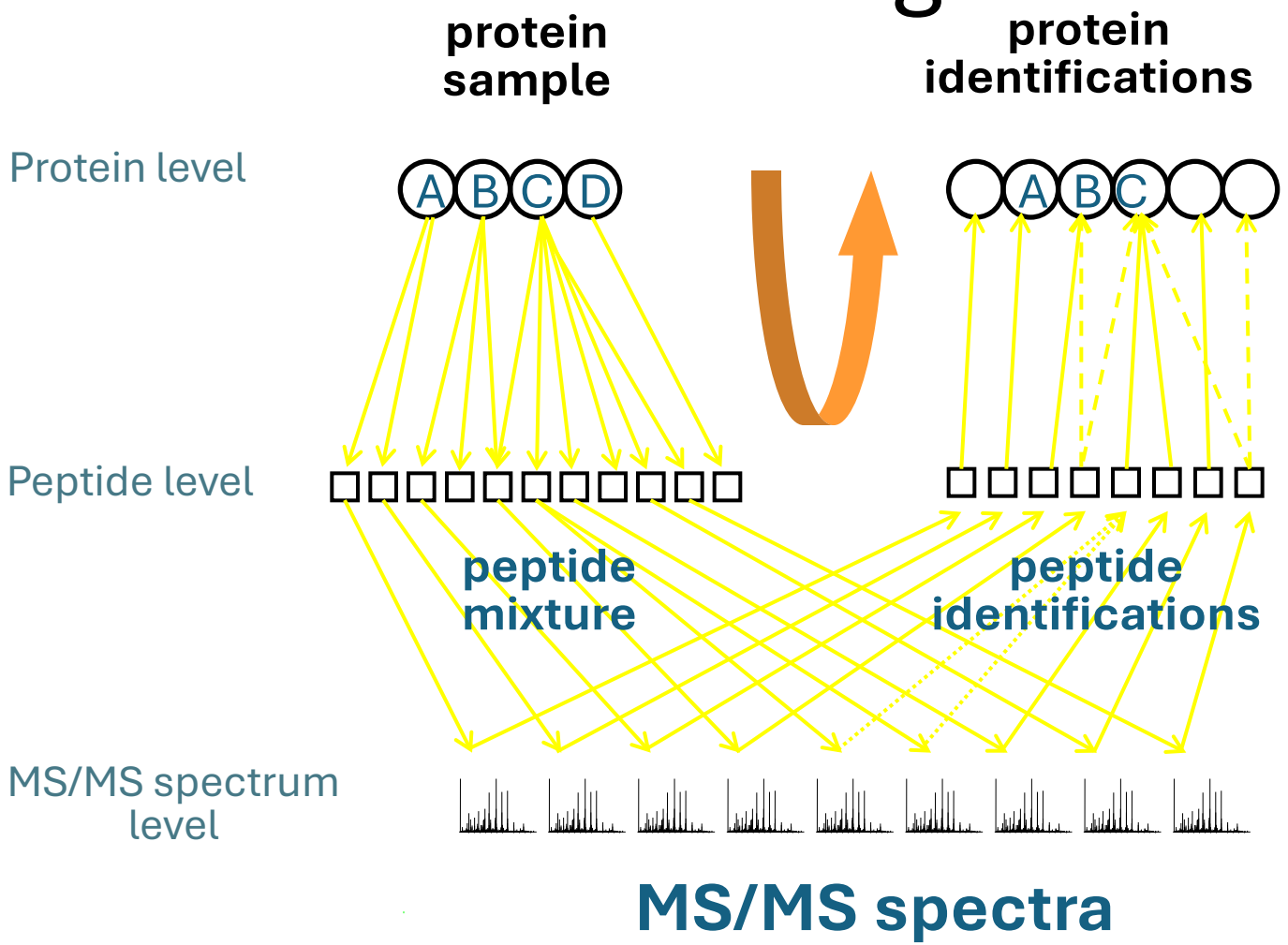
Abstract

Machine learning and in particular deep learning (DL) are increasingly important in mass spectrometry (MS)-based proteomics. Recent DL models can predict the retention time, ion mobility and fragment intensities of a peptide just from the amino acid sequence with good accuracy. However, DL is a very rapidly developing field with new neural network architectures frequently appearing, which are challenging to incorporate for proteomics researchers. Here we introduce AlphaPeptDeep, a modular Python framework built on the PyTorch DL library that learns and predicts the properties of peptides (https://github.com/MannLabs/alphapeptdeep). It features a model shop that enables non-specialists to create models in just a few lines of code. AlphaPeptDeep represents post-translational modifications in a generic manner, even if only the chemical composition is known. Extensive use of transfer learning obviates the need for large data sets to refine models for particular experimental conditions. The AlphaPeptDeep models for predicting retention time, collisional cross sections and fragment intensities are at least on par with existing tools. Additional sequence-based properties can also be predicted by AlphaPeptDeep, as demonstrated with a HLA peptide prediction model to improve HLA peptide identification for data-independent acquisition (https://github.com/MannLabs/PeptDeep-HLA).

Protein inference

- Peptides are measured as proxies for proteins
- Which proteins do the identified peptides represent?

Protein inference challenge



Protein inference

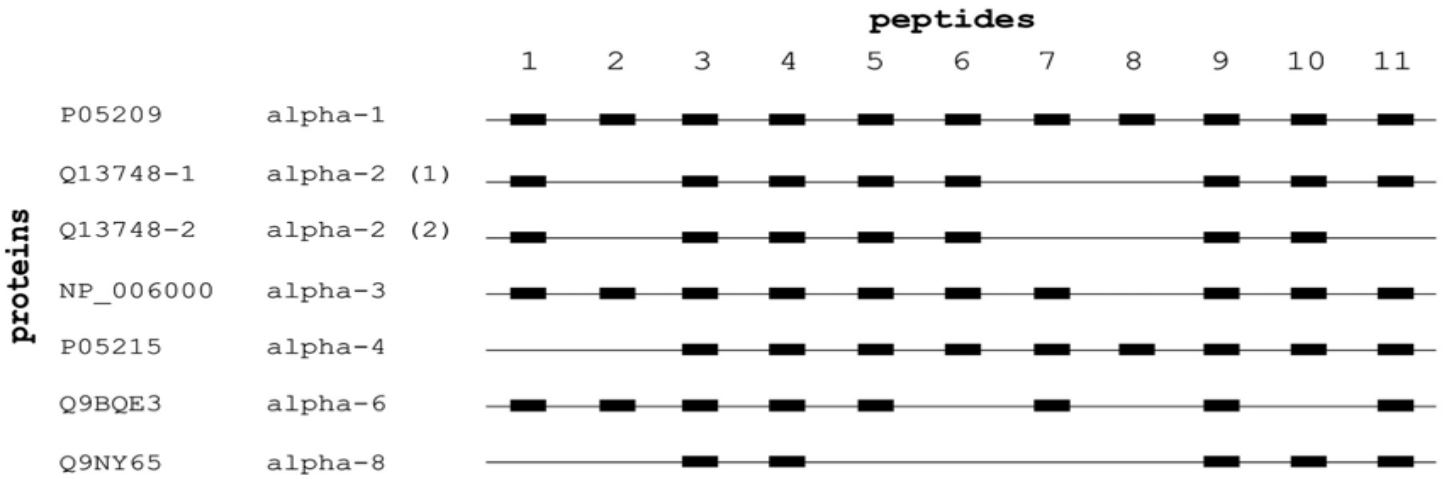
- Similar to RNA seq, but:
  - Shorter sequences in proteomics
  - More bases (~20 amino acids) in protein data compared to 4 in RNA.
  - Protein sequences cut at defined positions using specific enzyme. As a consequence: little sequence overlap.

How to group peptides to proteins?

Peptides identified:

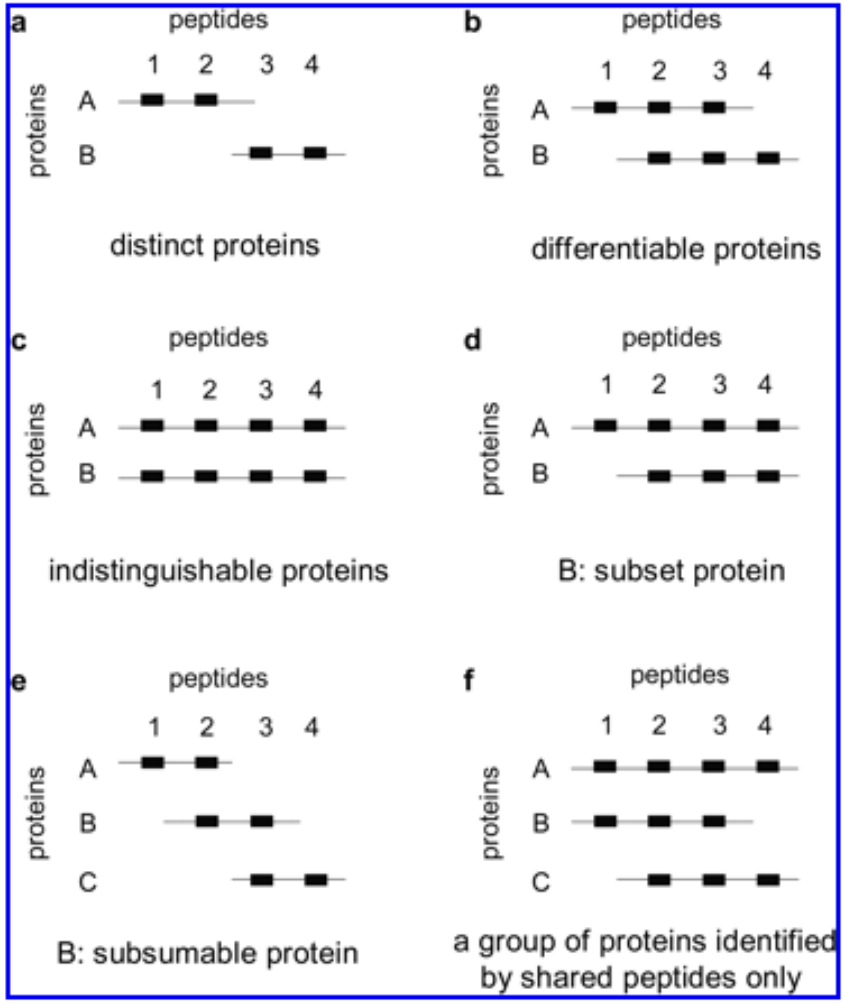
1	TIGGGDDSFNTFFSETGAGK	5	IHFPLATYAPVISAEK	9	VGINYQPPTVVPGGDLAK
2	AVFVDLEPTVIDEVR	6	AYHEQLSVAEITNACFEPANQMVK	10	AVCMLSNTTAIAEAWAR
3	QLFHPEQLITGKEDAANNYAR	7	YMACCLLYR	11	LDHKFDLMYAK
4	NLDIERPTYTNLNR	8	SIQFVDWCPTGFK		

Assignment of peptides to proteins:



https://doi.org/10.1074/mcp.r500012-mcp200

Occam’s razor approach



“Occam’s razor constraint, would provide a minimal list of proteins sufficient to explain all observed peptides. Such a minimal list would contain all distinct and differentiable proteins, e.g. proteins A and B in Fig. 5, a and b, and proteins A and C in Fig. 5e but no subsumable or subset proteins, e.g. only protein A would be included in the list in the cases shown in Fig. 5, d and f. In the case of indistinguishable protein identifications, Fig. 5c, it would be most accurate to collapse all such identifications into a single entry in the protein summary report as there is often no basis to eliminate any of them”

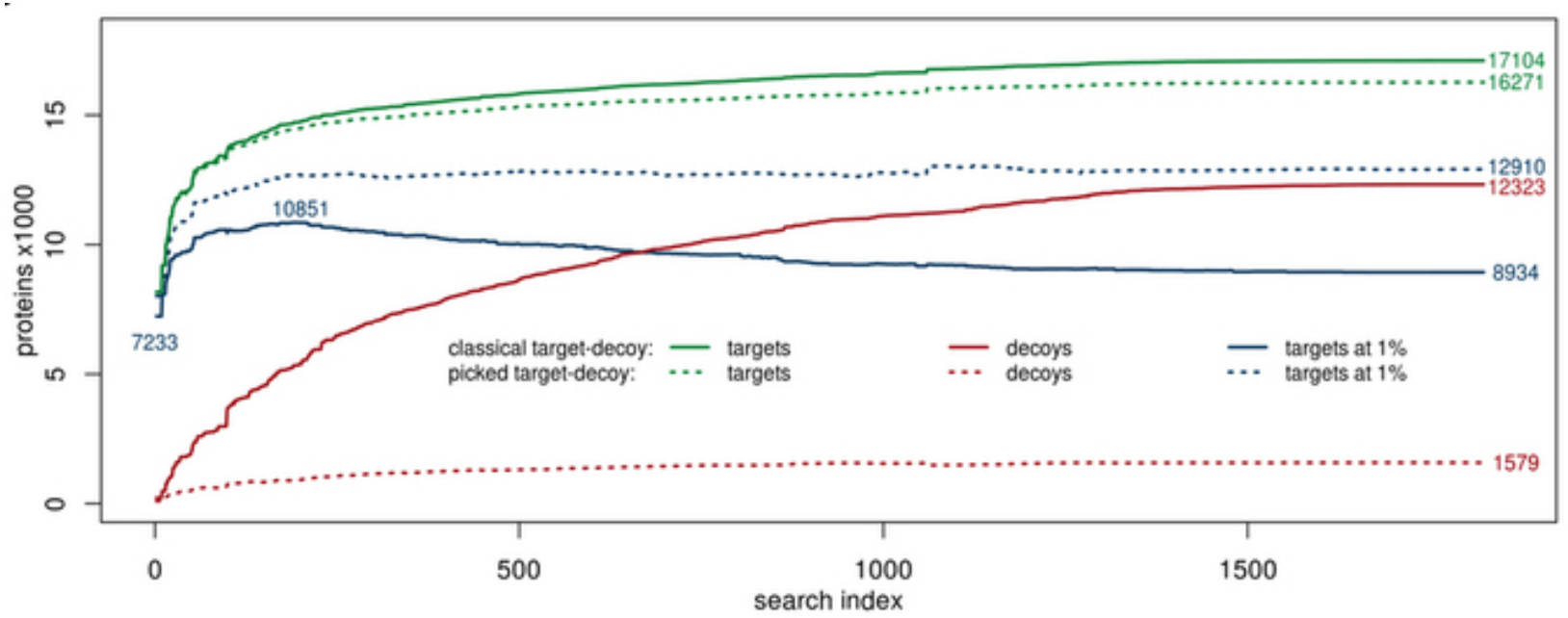
From Nesvizhskii, and Aebersold

https://doi.org/10.1074/mcp.r500012-mcp200

Target-decoy approach for FDR estimation

- Can be used at protein level as well as at PSM or peptide level
- How many target and decoy proteins are reported after filtering, including criteria such as peptide score cutoff and minimum number of peptides?

Scaling up problem



From

Mass-spectrometry-based draft of the human proteome

Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M. Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, Toby Mathieson, Simone Lemeer, Karsten Schnatbaum, Ulf Reimer, Holger Wenschuh, Martin Mollenhauer, Julia Slotta-Huspenina, Joos-Hendrik Boese, Marcus Bantscheff, Anja Gerstmair, Franz Faerber & Bernhard Kuster

Nature **509**, 582–587 (29 May 2014) | doi:10.1038/nature13319

Overcoming the problem

“Picked” model. Select best scores for any peptide in a protein. Calculate the same for the decoy version of the protein. If better than decoy= target, otherwise = decoy. Calculate FDR based on remaining.

Mol Cell Proteomics. 2015 Sep; 14(9): 2394–2404. PMCID: PMC4563723
Published online 2015 May 17. doi: 10.1074/mcp.M114.046995 PMID: 25987413

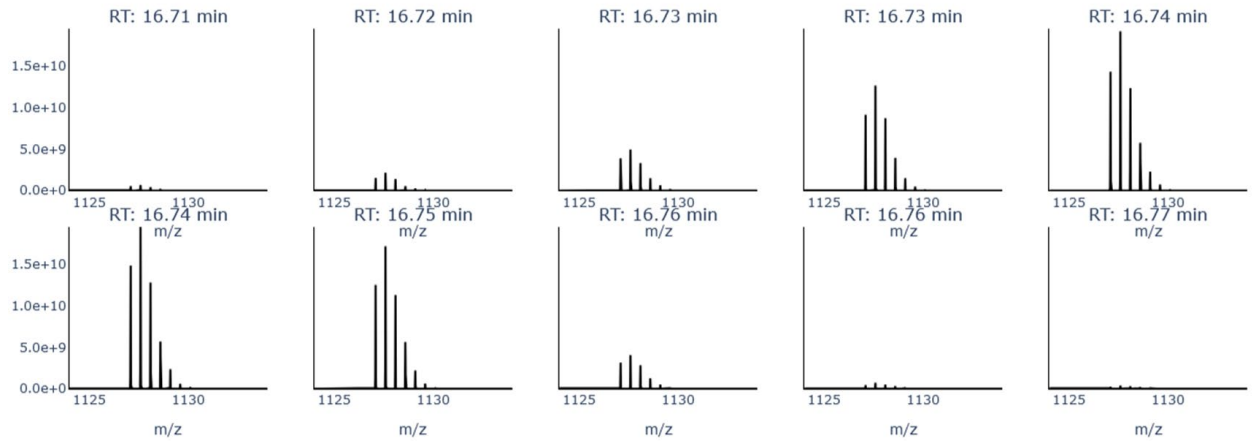
A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets[S]

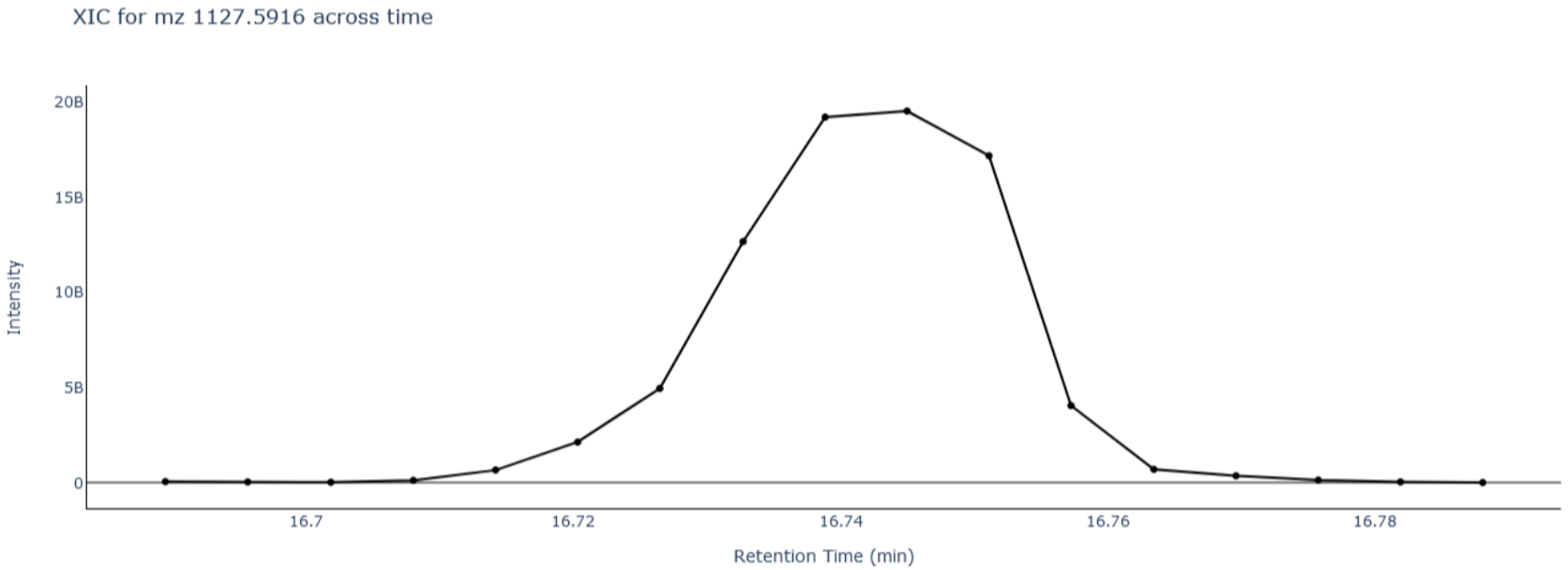
Mikhail M. Savitski,[‡‡‡] Mathias Wilhelm,[§¶‡‡‡] Hannes Hahne,[§] Bernhard Kuster,[§‖**] and Marcus Bantscheff[‡**]

Implementation can be found in Percolator:

https://github.com/percolator/percolator

Quantification

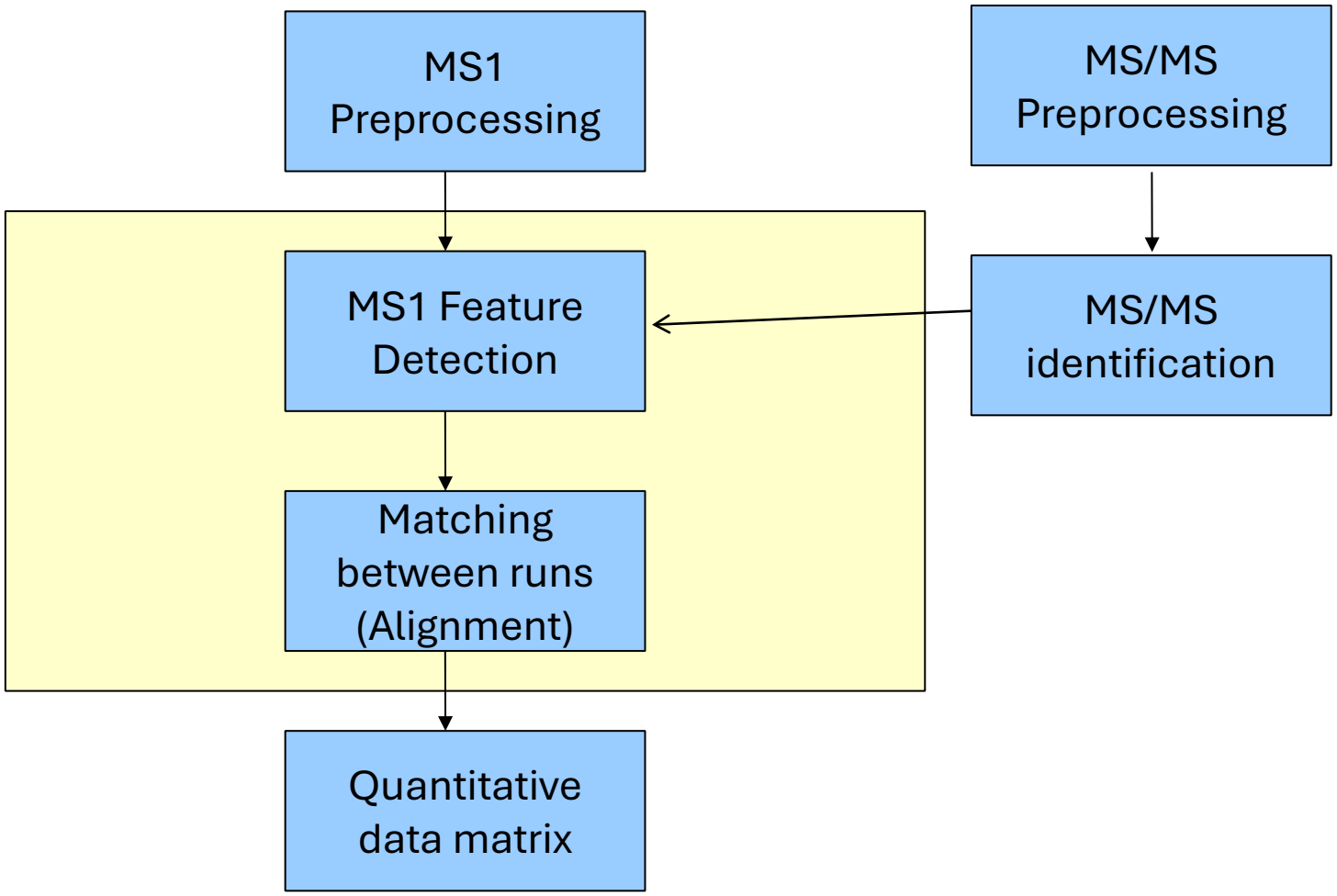




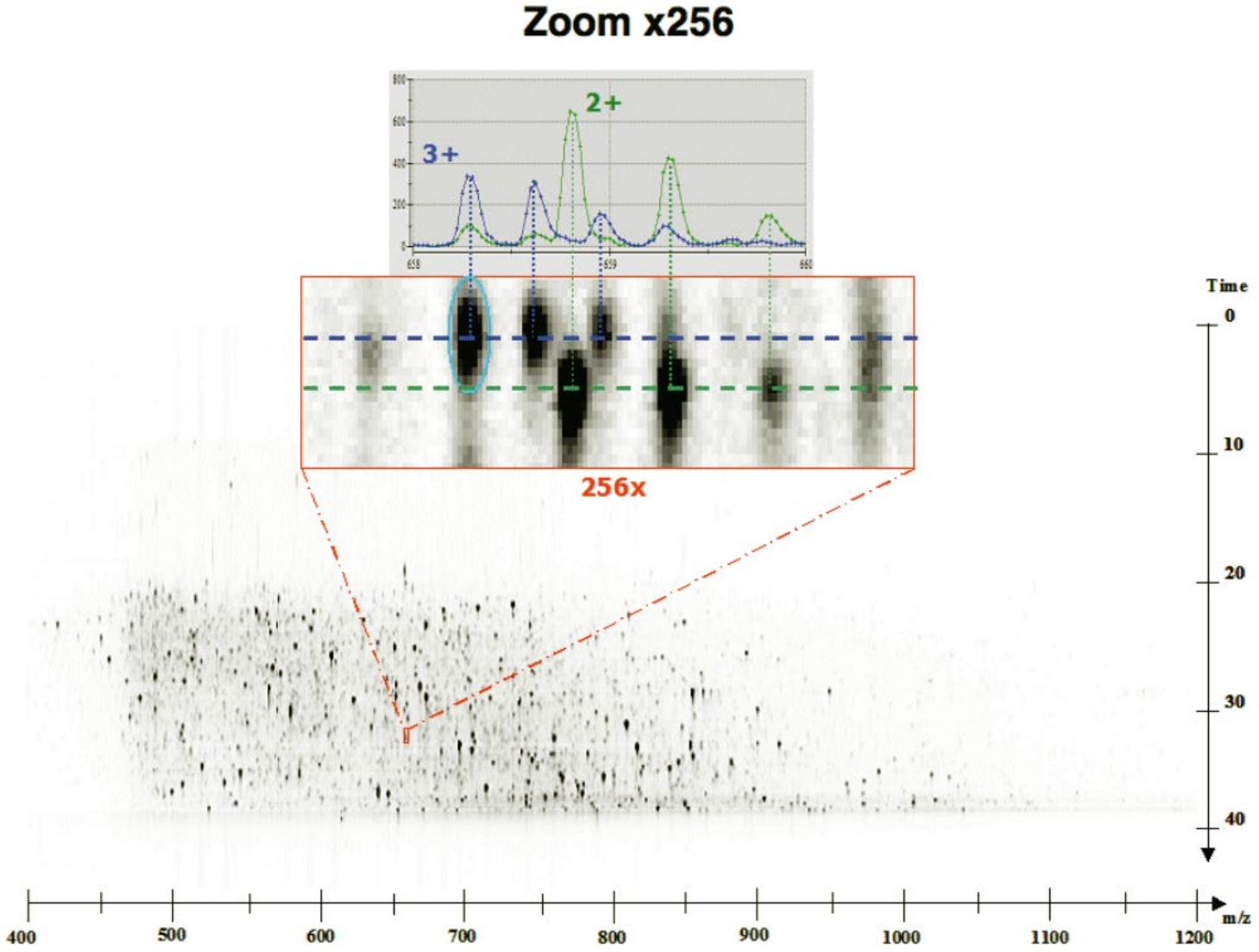
Quantification

- Peak intensities can be used for quantification, but:
  - Peptides have different ionisation properties and only relative quantities (between sample comparisons) can be derived.
  - To obtain absolute quantities we need parallel analyses of reference peptides of known concentration
  - Peaks can have MS/MS identity info in one run, but not in others (missing values)

Label free quantification workflow DDA data

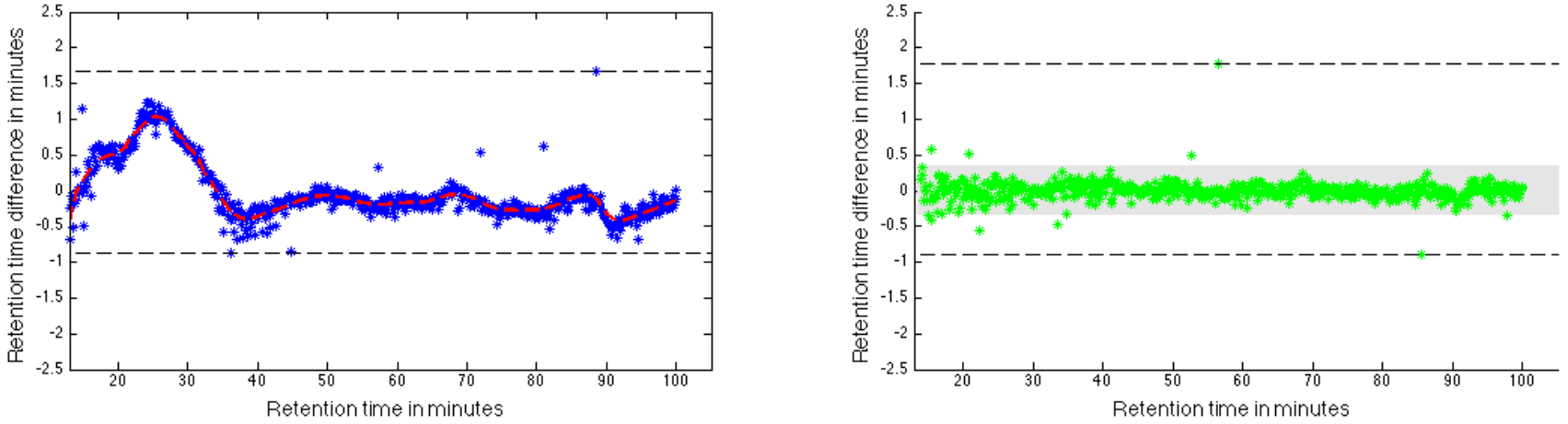


Quantification using MS1 peaks “peptide features”



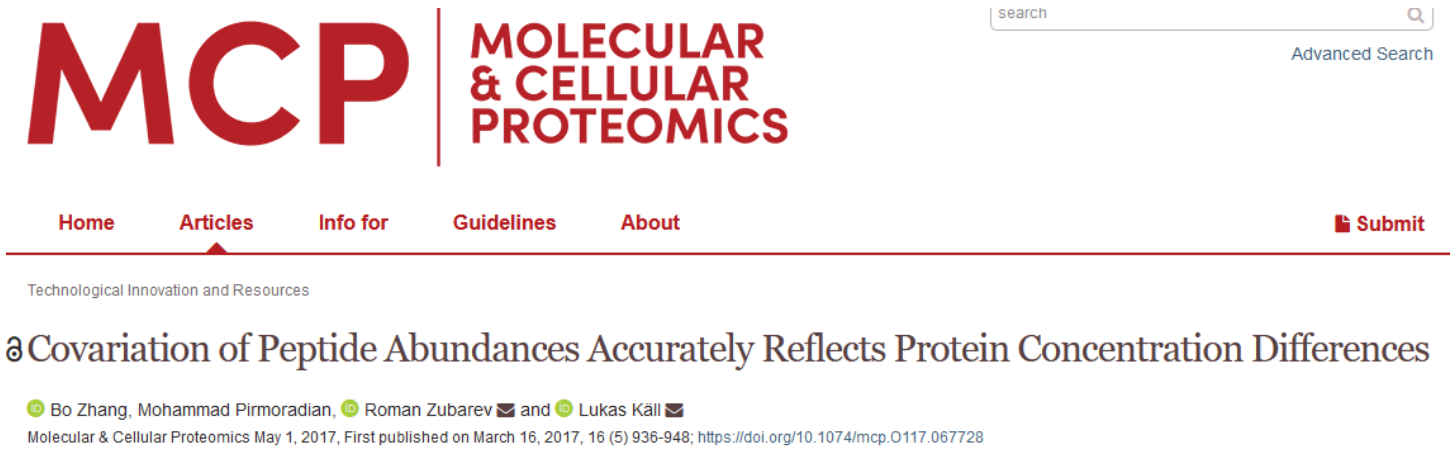
Matching of features in different LC-MS files

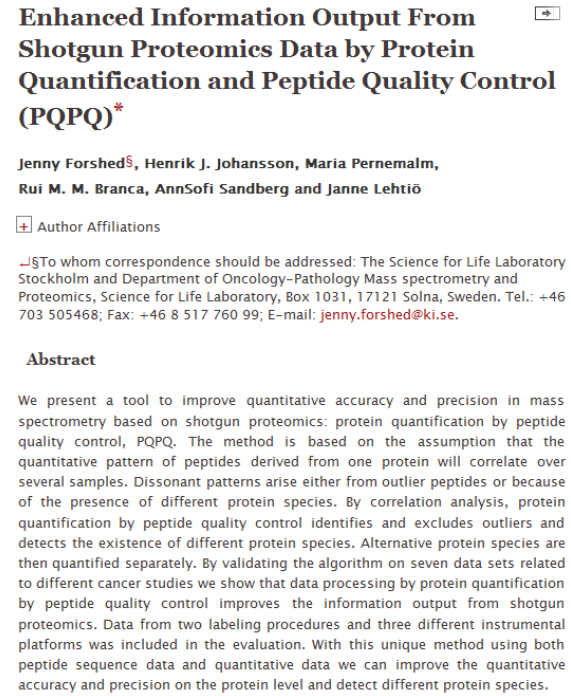
- critical for comprehensive and reproducible quantification



First: correction function for retention time drifts. *Alignment*
Then: match everything within m/z and retention time tolerance.

But what about protein quantities?





Enhanced Information Output From Shotgun Proteomics Data by Protein Quantification and Peptide Quality Control (PQPQ)*

Jenny Forshed§, Henrik J. Johansson, Maria Pernemalm, Rui M. M. Branca, AnnSofi Sandberg and Janne Lehtiö

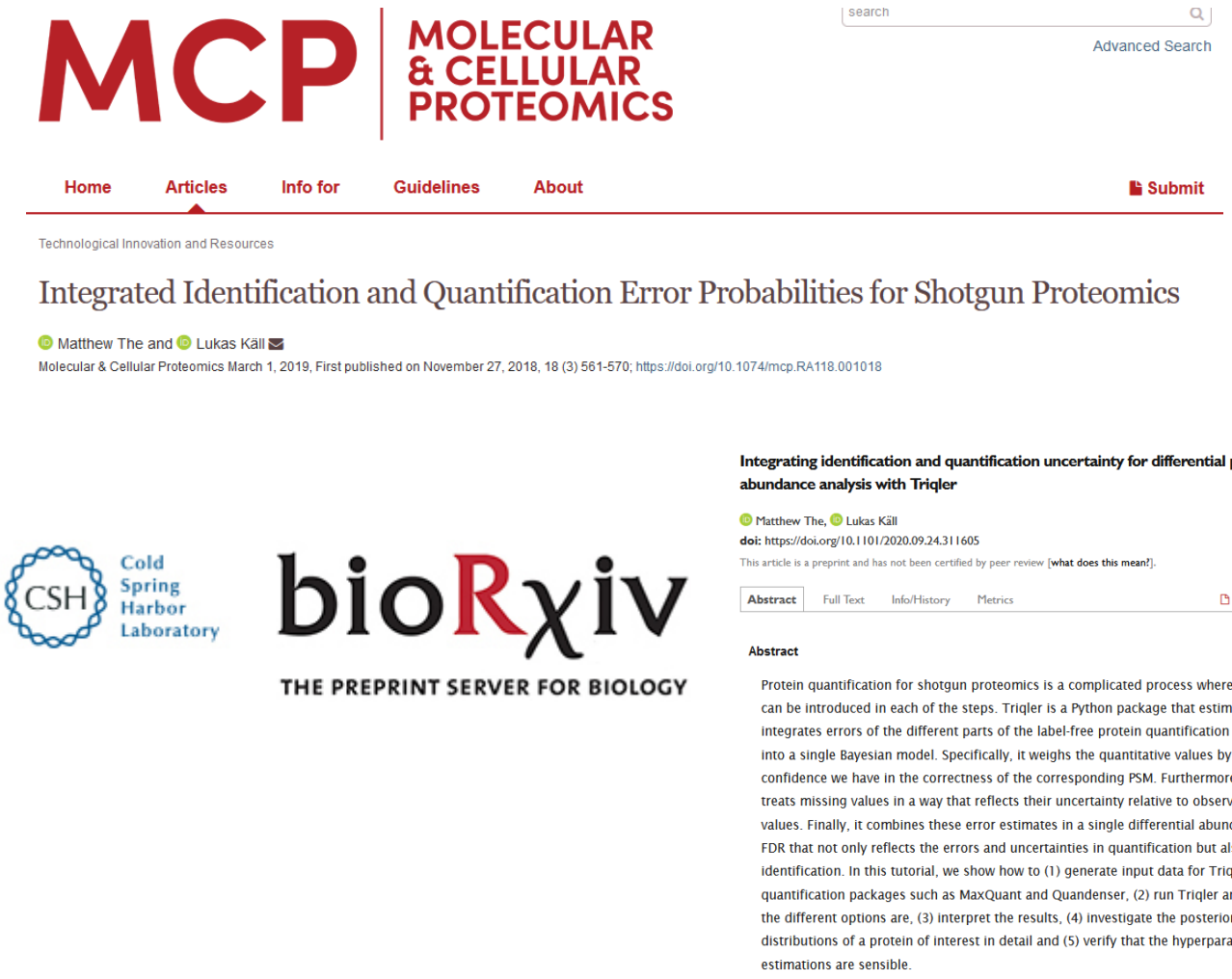
Author Affiliations

§To whom correspondence should be addressed: The Science for Life Laboratory Stockholm and Department of Oncology–Pathology Mass spectrometry and Proteomics, Science for Life Laboratory, Box 1031, 17121 Solna, Sweden. Tel.: +46 703 505468; Fax: +46 8 517 760 99; E-mail: jenny.forshed@ki.se.

Abstract

We present a tool to improve quantitative accuracy and precision in mass spectrometry based on shotgun proteomics: protein quantification by peptide quality control, PQPQ. The method is based on the assumption that the quantitative pattern of peptides derived from one protein will correlate over several samples. Dissonant patterns arise either from outlier peptides or because of the presence of different protein species. By correlation analysis, protein quantification by peptide quality control identifies and excludes outliers and detects the existence of different protein species. Alternative protein species are then quantified separately. By validating the algorithm on seven data sets related to different cancer studies we show that data processing by protein quantification by peptide quality control improves the information output from shotgun proteomics. Data from two labeling procedures and three different instrumental platforms was included in the evaluation. With this unique method using both peptide sequence data and quantitative data we can improve the quantitative accuracy and precision on the protein level and detect different protein species.

Combining error estimates?



MCP MOLECULAR & CELLULAR PROTEOMICS

search

Advanced Search

Home Articles Info for Guidelines About Submit

Technological Innovation and Resources

Integrated Identification and Quantification Error Probabilities for Shotgun Proteomics

Matthew The and Lukas Käll

Molecular & Cellular Proteomics March 1, 2019, First published on November 27, 2018, 18 (3) 561-570; https://doi.org/10.1074/mcp.RA118.001018

CSH Cold Spring Harbor Laboratory

bioRxiv THE PREPRINT SERVER FOR BIOLOGY

Integrating identification and quantification uncertainty for differential protein abundance analysis with Triqler

Matthew The, Lukas Käll

doi: https://doi.org/10.1101/2020.09.24.311605

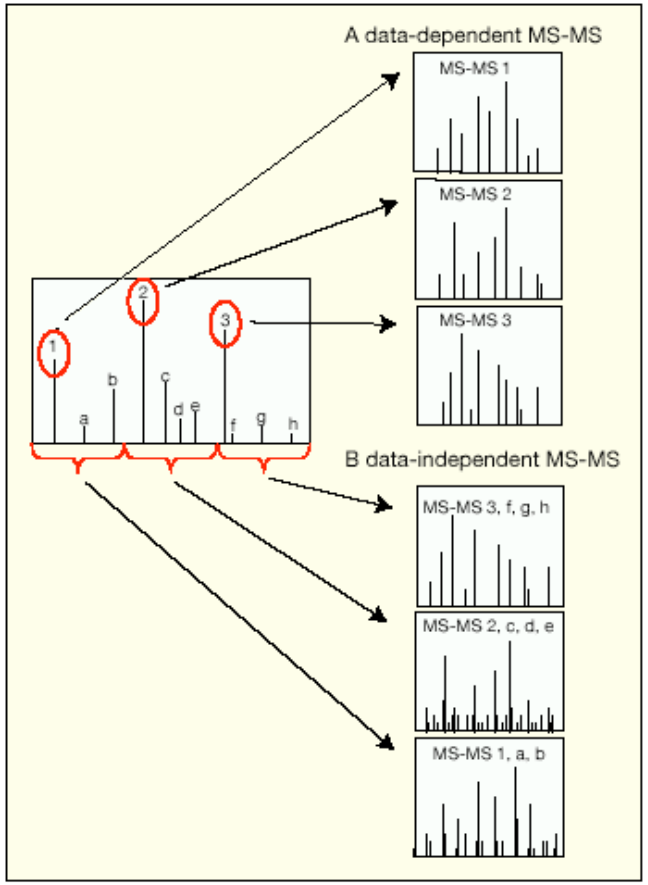
This article is a preprint and has not been certified by peer review [**what does this mean?**].

Abstract Full Text Info/History Metrics Preview PDF

Abstract

Protein quantification for shotgun proteomics is a complicated process where errors can be introduced in each of the steps. Triqler is a Python package that estimates and integrates errors of the different parts of the label-free protein quantification pipeline into a single Bayesian model. Specifically, it weighs the quantitative values by the confidence we have in the correctness of the corresponding PSM. Furthermore, it treats missing values in a way that reflects their uncertainty relative to observed values. Finally, it combines these error estimates in a single differential abundance FDR that not only reflects the errors and uncertainties in quantification but also in identification. In this tutorial, we show how to (1) generate input data for Triqler from quantification packages such as MaxQuant and Quandenser, (2) run Triqler and what the different options are, (3) interpret the results, (4) investigate the posterior distributions of a protein of interest in detail and (5) verify that the hyperparameter estimations are sensible.

DIA data processing



DDA

DIA

Nature Methods - 1, 16 - 17 (2004)
https://doi.org/10.1038/nmeth1004-16

DIA data processing

- DIA typically isolates ions in wider m/z windows than DDA
  - Complex spectra with multiple peptides fragmented
- Two main approaches for identification and quantification:
  - Spectrum-centric: try to match spectra with a sequence database as in standard MS/MS searches
  - Peptide-centric: look for peptides in a spectral library in the spectra

Mostly peptide-centric approaches used. A library with peptide spectra and their LC retention times needed as input.

➔Need to acquire spectral library using DDA. Instrument-specific fragmentation

➔Or use predicted spectral library?

➔Quantification using MS2 fragment traces and / or MS1 precursor traces.

Advances in DIA data processing using AI

DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput

Vadim Demichev, Christoph B. Messner, Spyros I. Vernardis, Kathryn S. Lilley & Markus Ralser

Nature Methods **17**, 41–44 (2020) | Cite this article

49k Accesses | **879** Citations | **201** Altmetric | Metrics

Abstract

We present an easy-to-use integrated software suite, DIA-NN, that exploits deep neural networks and new quantification and signal correction strategies for the processing of data-independent acquisition (DIA) proteomics experiments. DIA-NN improves the identification and quantification performance in conventional DIA proteomic applications, and is particularly beneficial for high-throughput applications, as it is fast and enables deep and confident proteome coverage when used in combination with fast chromatographic methods.

AlphaDIA enables End-to-End Transfer Learning for Feature-Free Proteomics

Georg Wallmann, Patricia Skowronek, Vincenth Brennsteiner, Mikhail Lebedev, Marvin Thielert, Sophia Steigerwald, Mohamed Kotb, Tim Heymann, Xie-Xuan Zhou, Magnus Schwörer, Maximilian T. Strauss, Constantin Ammar, Sander Willems, Wen-Feng Zeng, Matthias Mann

doi: https://doi.org/10.1101/2024.05.28.596182

This article is a preprint and has not been certified by peer review [**what does this mean?**].

Abstract

Mass spectrometry (MS)-based proteomics continues to evolve rapidly, opening more and more application areas. The scale of data generated on novel instrumentation and acquisition strategies pose a challenge to bioinformatic analysis. Search engines need to make optimal use of the data for biological discoveries while remaining statistically rigorous, transparent and performant. Here we present alphaDIA, a modular open-source search framework for data independent acquisition (DIA) proteomics. We developed a feature-free identification algorithm particularly suited for detecting patterns in data produced by sensitive time-of-flight instruments. It naturally adapts to novel, more eTicient scan modes that are not yet accessible to previous algorithms. Rigorous benchmarking demonstrates competitive identification and quantification performance. While supporting empirical spectral libraries, we propose a new search strategy named end-to-end transfer learning using fully predicted libraries. This entails continuously optimizing a deep neural network for predicting machine and experiment specific properties, enabling the generic DIA analysis of any post-translational modification (PTM). AlphaDIA provides a high performance and accessible framework running locally or in the cloud, opening DIA analysis to the community.