COMPUTATIONAL PROTEOMICS AND METABOLOMICS

Oliver Kohlbacher, Sven Nahnsen, Knut Reinert

9. Protein Inference

Overview

- The protein inference problem
  - Isoforms and protein groups
  - Problem definition
- Protein inference algorithms
  - ProteinProphet
- Protein false discovery rates
  - Difference between PSM FDR and protein FDR
  - Computing protein FDRs
  - MAYU

LEARNING UNIT 9A
PROTEIN INFERENCE PROBLEM

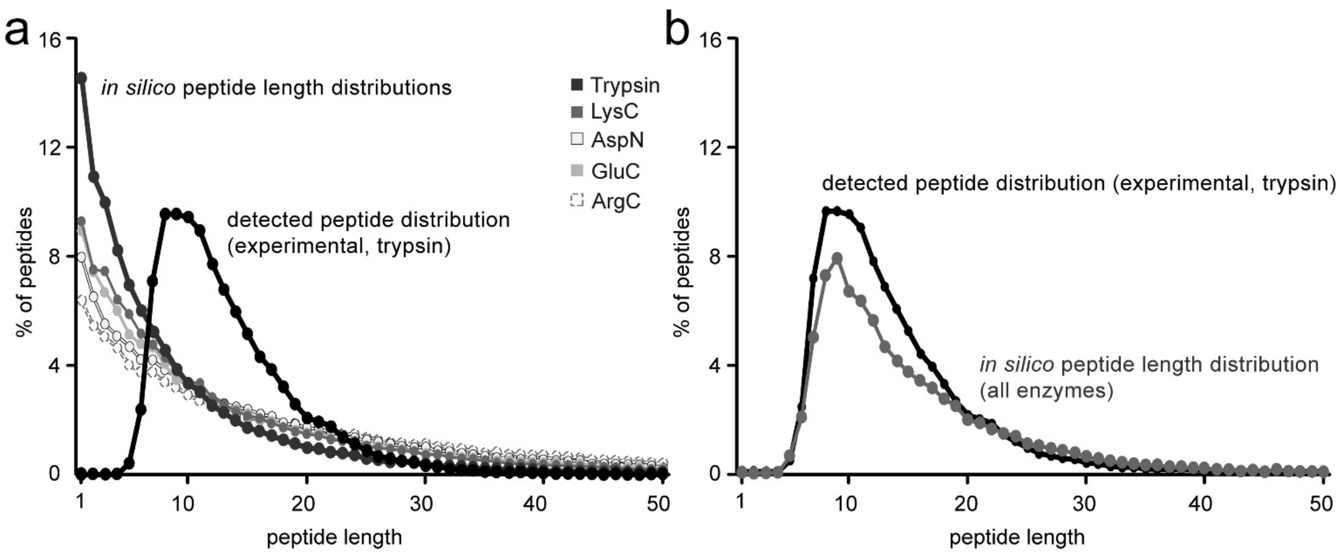
- Problem definition
- Protein families
- Protein ambiguity groups
- Inference through quantification
- Significance of inferred hits
- One hit wonders

Identifying Proteins

- Identification methods so far only identify peptide-spectrum matches (PSMs)
  - Search a database
  - Return a ranked list of PSMs with associates scores
- PSM false discovery rates (FDRs) can be computed through a target-decoy approach
- An FDR of 1% would mean that 1% of the PSMs with a score above the threshold are expected to be incorrect
- Note that this is a statement on the individual PSM, not per peptide or protein!

Identifying Proteins

- Each PSM above the threshold contributes
  - a match of a spectrum to a peptide
  - a match of a peptide to a protein
- Peptides are not necessarily unique!
- Length distribution of observed peptides deviates from theoretical distribution: short peptides (length 6 and shorter) are usually not observed

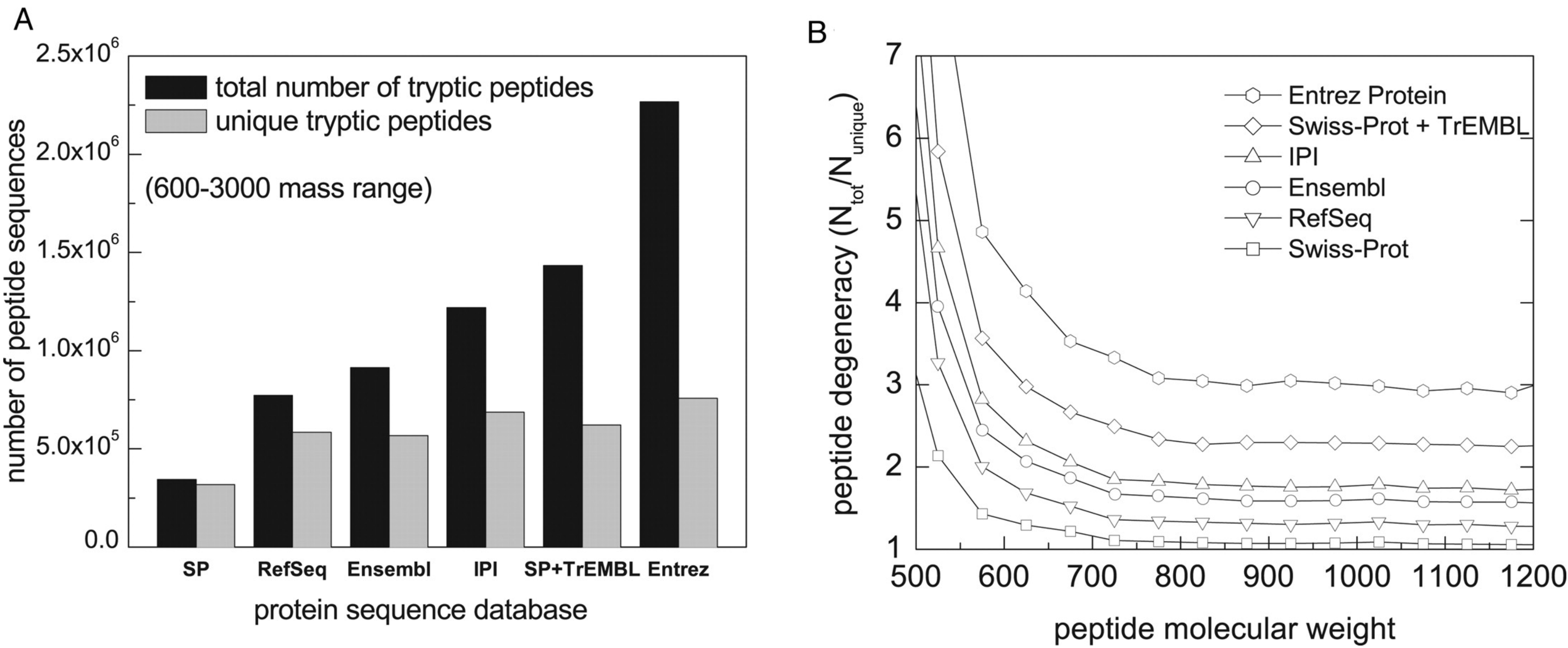


Danielle L. Swaney; Craig D. Wenger; Joshua J. Coon; *J. Proteome Res.* 2010, 9, 1323-1329.

Uniqueness

- If we are interested in proteomics (in contrast to peptide identification in metabolomics, MHC ligandomics etc.), we want to quantify proteins
- Non-unique peptide sequences can stem from different proteins
- Obviously, uniqueness depends on the chosen database
- Uniqueness becomes more likely for longer peptide sequences
- Reasons for non-uniqueness
    - Chance hits
    - Different isoforms
    - Conserved regions shared within a protein family

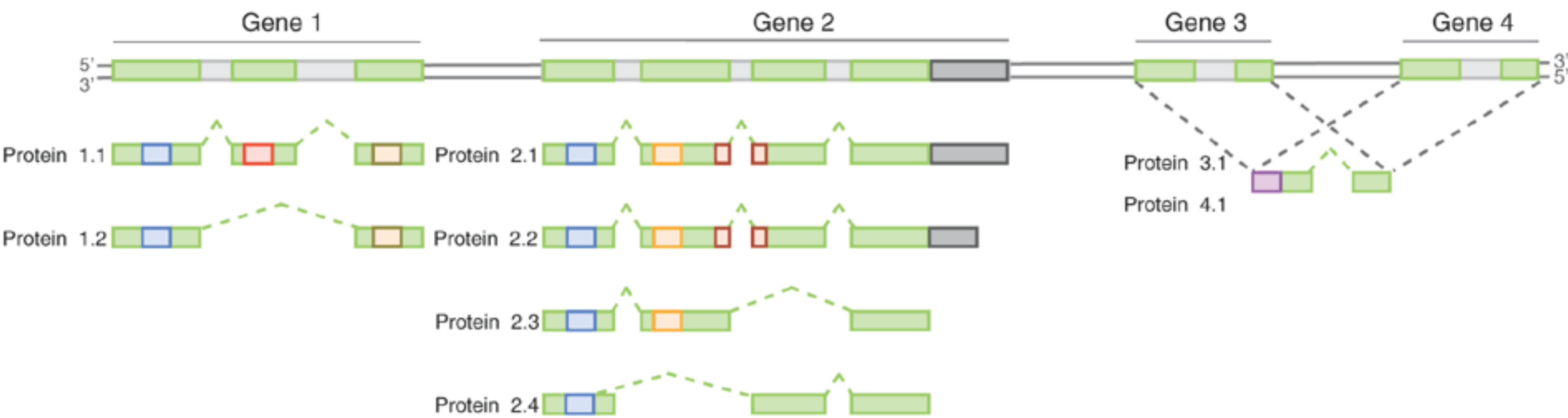
Uniqueness



- Uniqueness depends on the size of the database
- Searching an appropriate (non-redundant) database is thus preferable
- Reference databases (SwissProt) usually contain few degenerate (non-unique) tryptic peptides above a mass of 750 Da
- Problem: isoforms of proteins/splice variants!

Nesvizhskii A I , Aebersold R Mol Cell Proteomics 2005;4:1419-1440.

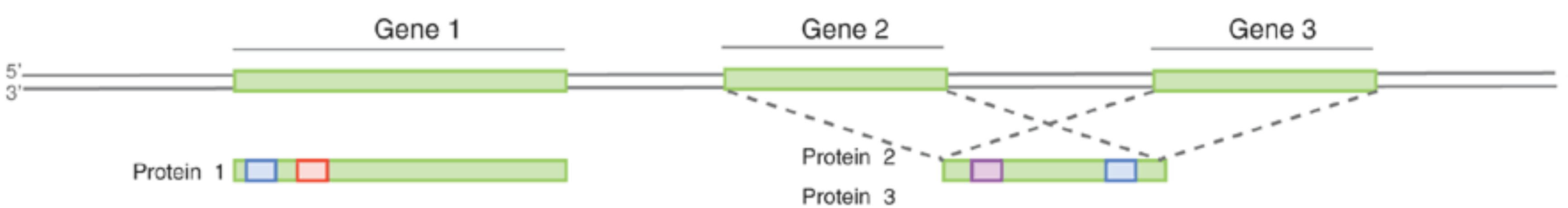
Uniqueness



Class	Protein sequence(s)	Protein isoform(s)	Gene(s)
1a	Unambiguous	Unambiguous	Unambiguous
1b	Unambiguous	Ambiguous	Unambiguous
2a	Ambiguous	Ambiguous	Unambiguous
2b	Ambiguous	Ambiguous	Unambiguous
3a	Unambiguous	Ambiguous	Ambiguous
3b	Ambiguous	Ambiguous	Ambiguous

Eukaryotes

Prokaryotes



Qeli & Ahrens, Nature Biotechnology 28, 647–650 (2010)

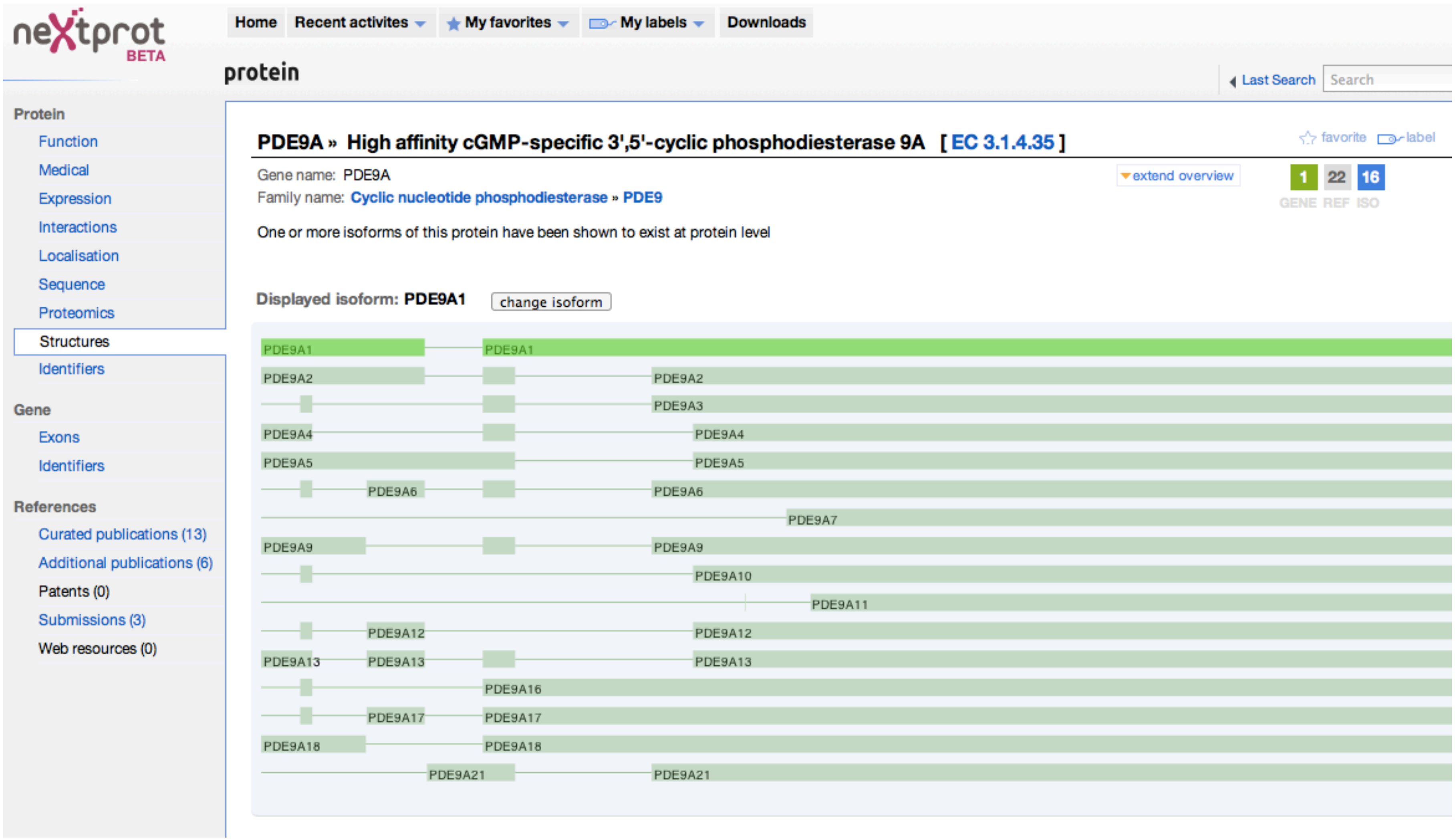
Protein Isoforms

www.nextprot.org/db/statistics/release?viewas=numbers

- NextProt Release 3.0.20
  - 20,140 human proteins
  - 39,565 sequences resulting from alternative isoforms
- On average 2.96 different splice variants for each protein sequence
- Some proteins have a much larger number of variants
- Resolving the different isoforms is only possible, if peptides crossing the right exon boundaries are observed

NextProt Release 3.0.20, 2013-11-01, http://www.nextprot.org/db/statistics/release?viewas=numbers

Protein Isoforms



- phosphodiesterase 9A has 16 documented isoforms
- Peptides stemming from the second half of the sequence are entirely indistinguishable between isoforms

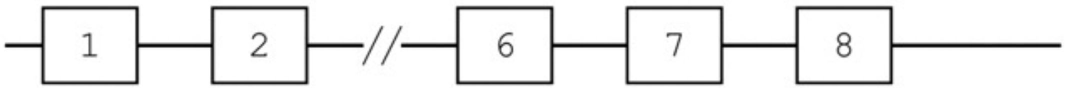
http://www.nextprot.org/db/entry/NX_O76083/structures

Protein Isoforms

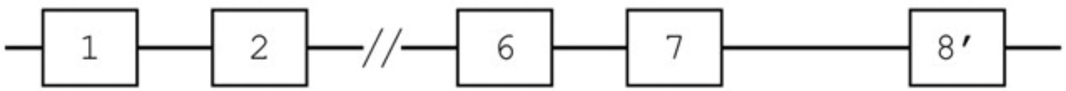
A

Gene CAPZB

>IPI00026185 IPI:IPI00026185.4|Swiss-Prot:P47756-1|ENSEMBL:ENSP00000264202
Tax_Id=9606 Splice isoform 1 of P47756 F-actin capping protein beta subunit



>IPI00218782 IPI:IPI00218782.1|Swiss-Prot:P47756-2|ENSEMBL:ENSP00000264203
Tax_Id=9606 Splice isoform 2 of F-actin capping protein beta subunit



P47756-1: MSDQQLDCALDLMRRLPPQQIEKNLSDLIDLVPSLCEDLLSSVDQPLKIARDKVVGKDYL 60
MSDQQLDCALDLMR**RLPPQQIEK**NLSDLIDLV**PSLCEDLLSSVDQPLK**IARDKVVGK**DYL**
P47756-2: MSDQQLDCALDLMRRLPPQQIEKNLSDLIDLVPSLCEDLLSSVDQPLKIARDKVVGKDYL 60

P47756-1: LCDYNRDGDSYRSPWSNKYDPPLEDGAMPSARLRKLEVEANNAFDQYRDLYFEGGVSSVY 120
LCDYNRDGDSYRSPWSNKYDPPLEDGAMPSARLR**KLEVEANNAFDQYR**DLYFEGGVSSVY
P47756-2: LCDYNRDGDSYRSPWSNKYDPPLEDGAMPSARLRKLEVEANNAFDQYRDLYFEGGVSSVY 120

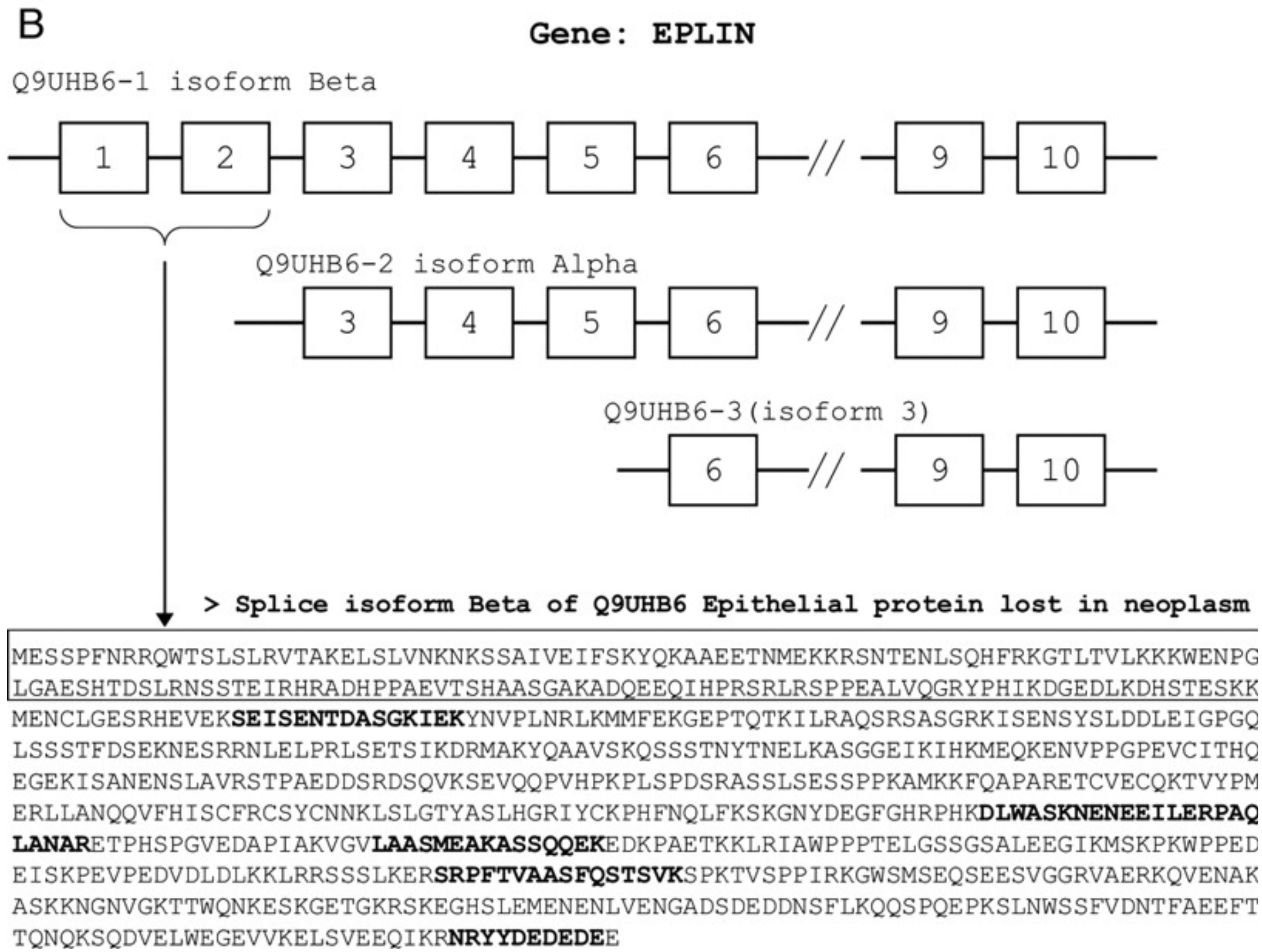
P47756-1: LWDLDHGFAGVILIKKAGDGSKKIKGCWDSIHVVEVQEKSSGRTAHYKLTSTVMLWLQTN 180
LWDLDHGFAGVILIKKAGDGSKKIK**GCWDSIHVVEVQEK**SSGRTAHYKLTSTVMLWLQTN
P47756-2: LWDLDHGFAGVILIKKAGDGSKKIKGCWDSIHVVEVQEKSSGRTAHYKLTSTVMLWLQTN 180

P47756-1: KSGSGTMNLGGSLTRQMEKDETVSDCSPHIANIGRLVEDMENKIRSTLNEIYFGKTKDIV 240
KSGSGTMNLGGSLTR**QMEKDETVSDCSPHIANIGRLVEDMENKIRSTLNEIYFGK**TKDIV
P47756-2: KSGSGTMNLGGSLTRQMEKDETVSDCSPHIANIGRLVEDMENKIRSTLNEIYFGKTKDIV 240

P47756-1: NGLRSIDAIPDNQKFKQLQRELSQVLTQRQ 270
NGLRS+   D K + L+ +L + L ++Q
P47756-2: NGLRSVQTFADKSKQEALKNDLVEALKRKQ 270

Nesvizhskii A I , Aebersold R Mol Cell Proteomics 2005;4:1419-1440.

Protein Isoforms



Nesvizhskii A I , Aebersold R Mol Cell Proteomics 2005;4:1419-1440.

Protein Families

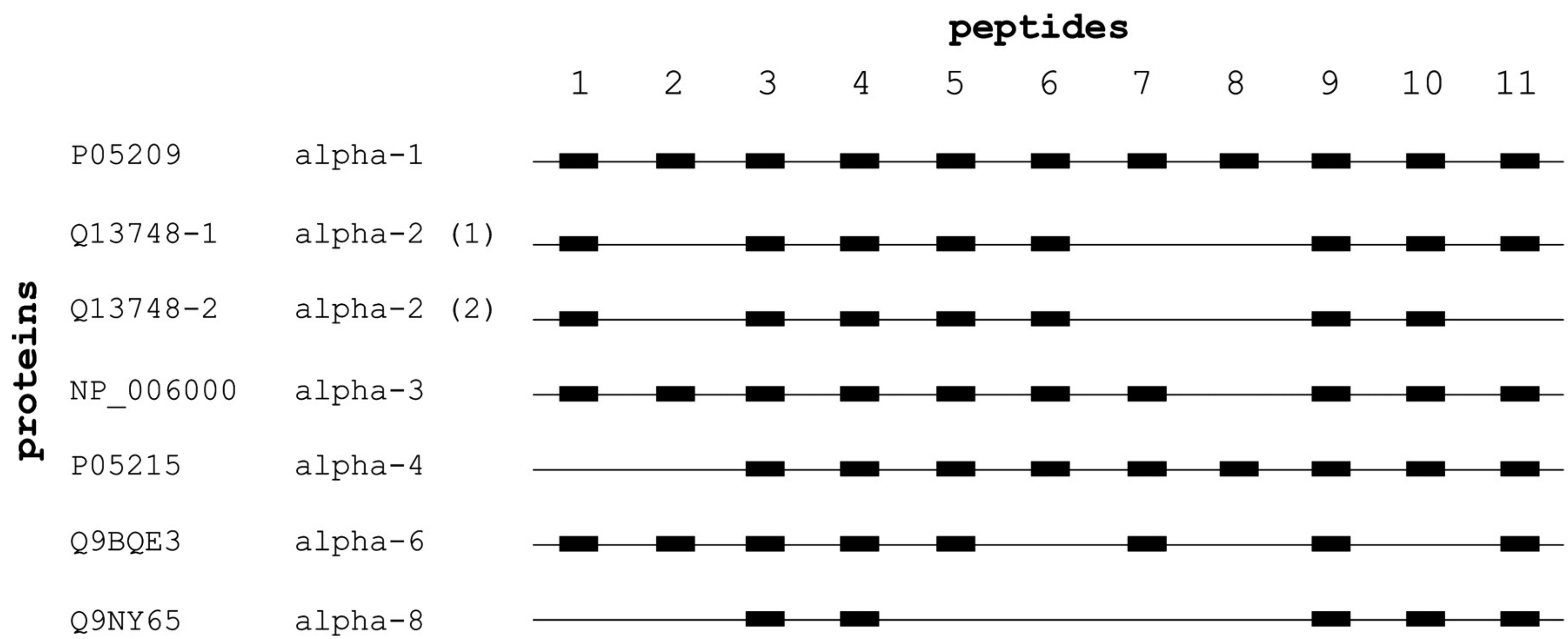
- Sequence coverage is often poor in large scale studies: many proteins are identified through very few peptides only
- In prokaryotes, typically over 90% of the identified peptides are unique in the whole proteome
- In particular in eukaryotes the large number of orthologs leads to significant sequence identity between different proteins that are not isoforms
- In eukaryotes, the number of unique identified peptides can thus easily drop below 50% (Gupta & Pevzner, 2009)

Protein Families

Peptides identified:

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

Assignment of peptides to proteins:



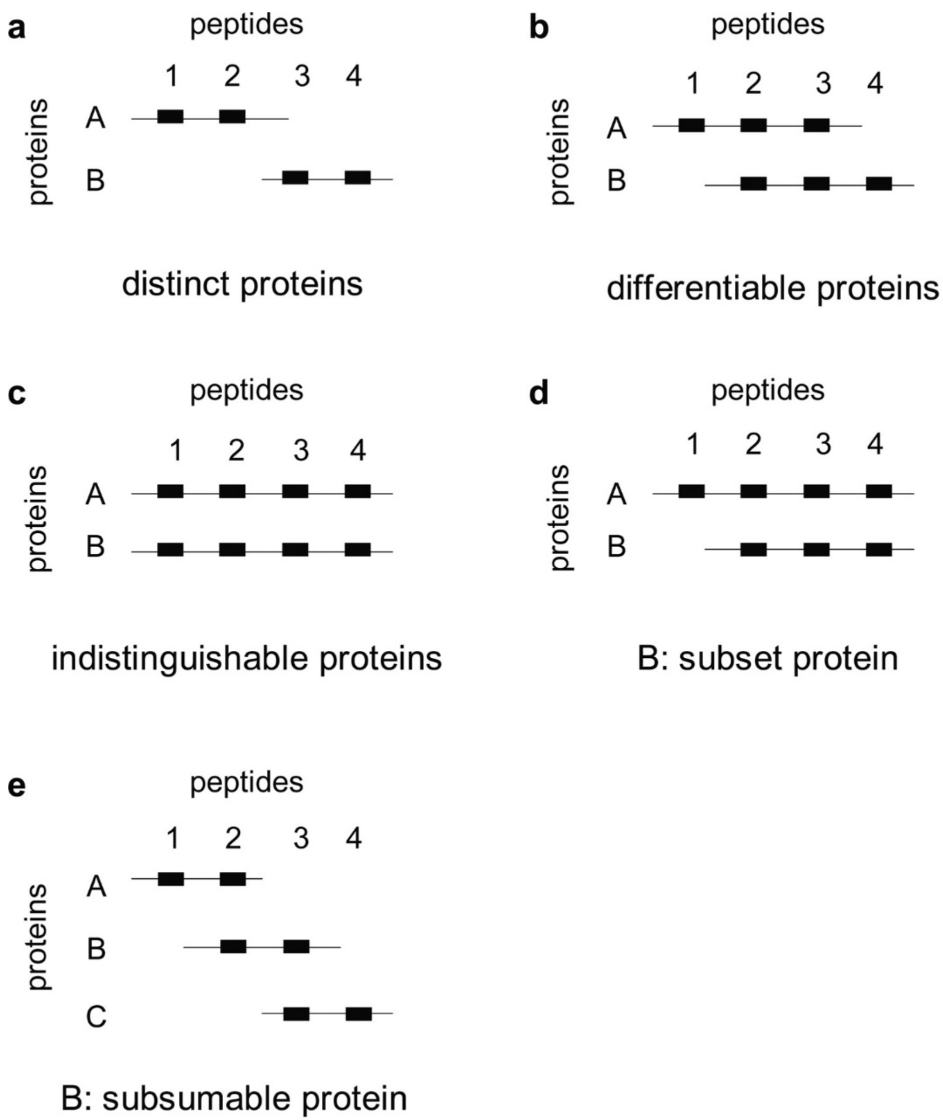
Parsimony-Based Inference

- **Idea**

  Find the smallest set of proteins explaining all observed peptides
- If all peptides mapping to one protein family can be explained by a single protein, then it is quite likely, that only this protein is present (but this must not necessarily be the case)
- Basically: applying **Occam's razor** to the dataset – find the simplest explanation possible (**maximum parsimony**)

Parsimony-Based Inference

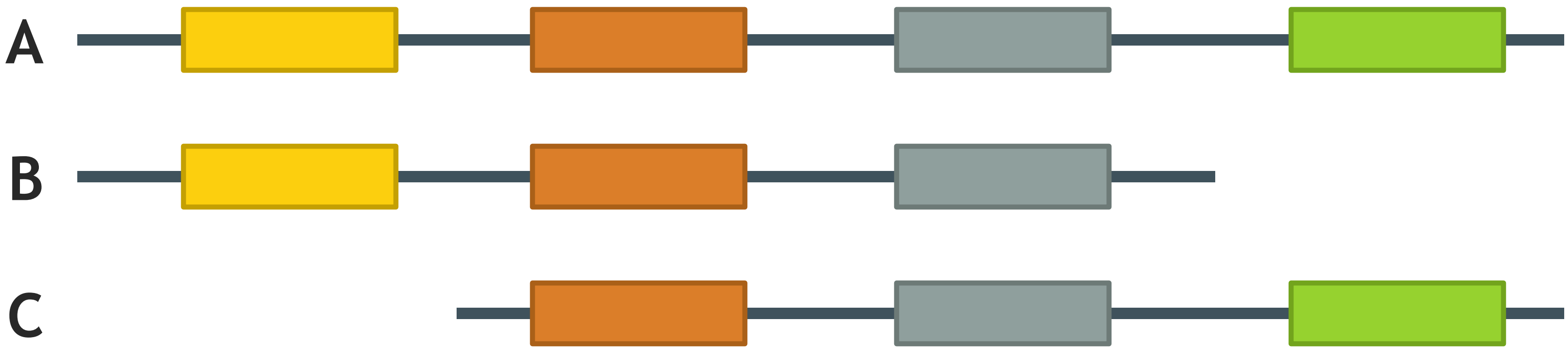
- Scenarios for different proteins given a set of observed peptides
  - **Distinct** proteins do not share peptides
  - **Differentiable** proteins can be distinguished by at least one distinct peptide
  - **Indistinguishable** proteins share all peptides
  - **Subset** proteins contain only peptides also contained in another protein
  - **Subsumable** proteins contain only peptides that are also contained in other proteins



Nesvizhskii A I , Aebersold R Mol Cell Proteomics 2005;4:1419-1440.

Protein Ambiguity Groups

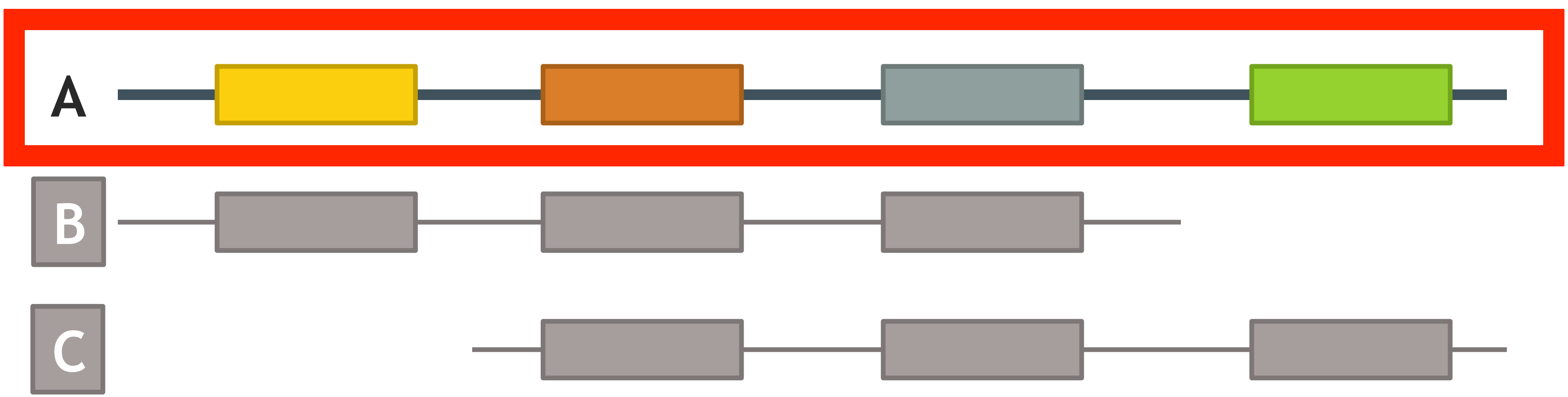
Example:



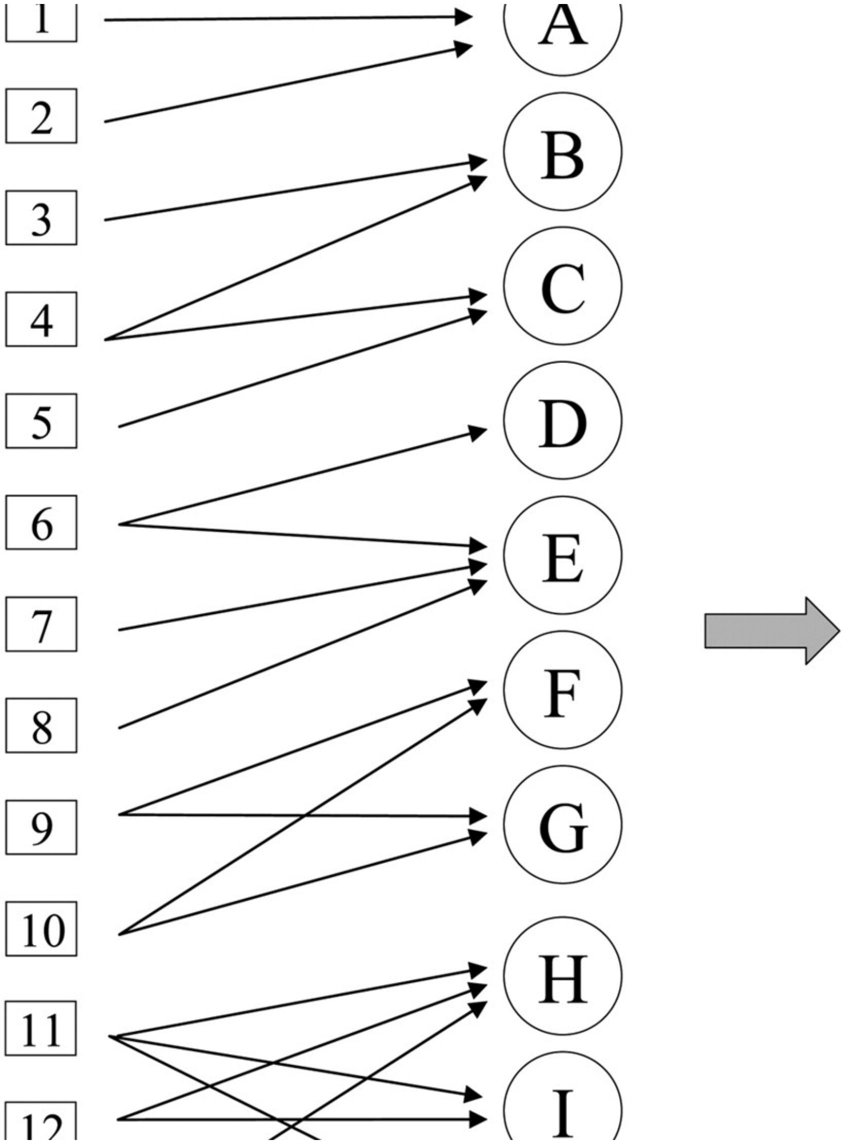
- Note that even though the presence of A is sufficient to explain all observed peptides, this does not automatically imply the absence of B and C
- The data is explained equally well by the presence of A, the presence of A + B, A + C, B + C, or A + B + C
- The set of proteins sharing one or multiple peptides is often referred to as a **protein ambiguity group**

Parsimony-Based Inference

- Maximum parsimony inference results in a **minimal list of proteins**
- It thus removes all distinct and differentiable proteins of a protein ambiguity group
- It does not contain any subsumable or subset proteins
- In the previous example, A would be sufficient to explain the observed peptides, B and C would not be reported



Reporting of PAGs



1. Protein A
   peptides 1, 2
2. Protein B
   peptides 3, 4*
3. Protein C
   peptides 4*, 5
4. Protein E
   peptides 6*, 7, 8
5. Protein F, Protein G
   peptides 9*, 10*
6. Protein group:

(1) Protein H
   peptides 11*, 12*, 13*

(2) Protein I
   peptides 11*, 12*

(3) Protein J
   peptides 11*, 13*

"protein" count: 6

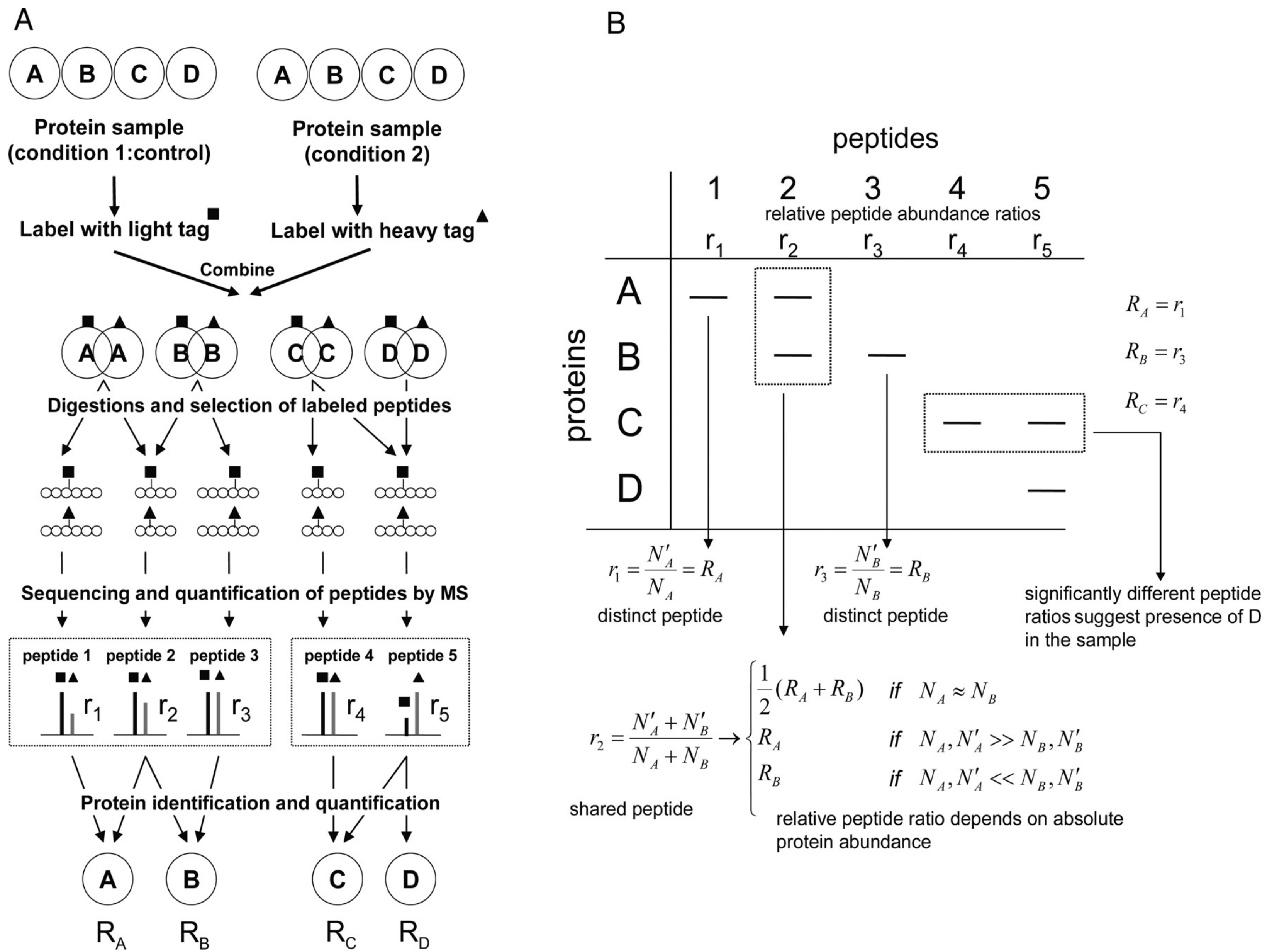
no conclusive evidence:

Nesvizhskii A I , Aebersold R Mol Cell Proteomics 2005;4:1419-1440.

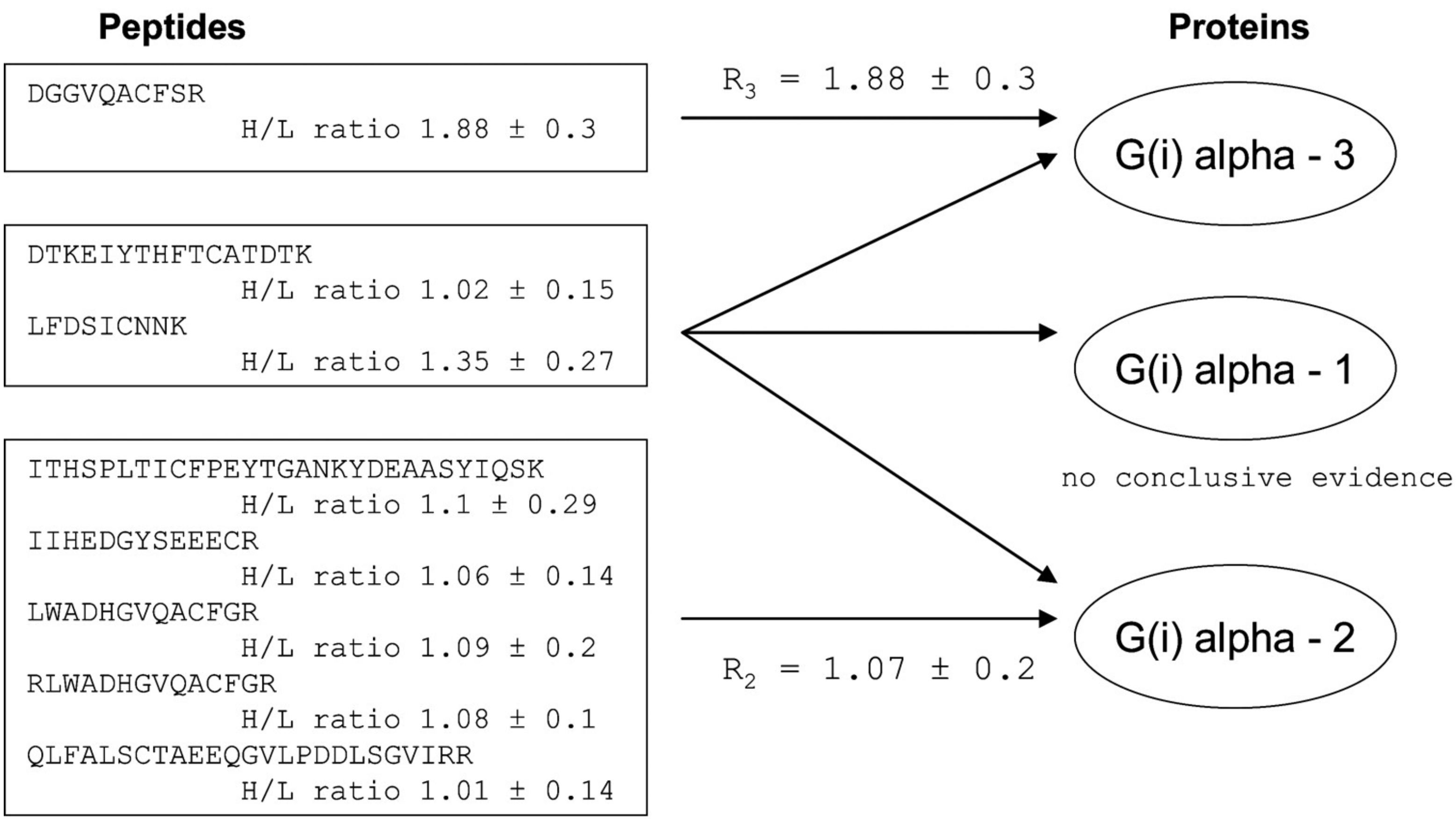
Inference through Quantification

- Quantitative data can be used for inference as well (similar to transcript data)
- This is, however, non-trivial and usually done manually and on a case-by-case basis
- Distinct peptides can be used to quantify their source proteins
- Shared peptides result in an averaging of the quantitative information
- This results in (often underdetermined) systems that can be used to quantify isoforms
- Quantitative information can also be used to prove the presence of a specific isoform (through deviating ratios of shared peptides)

Inference through Quantification



Inference through Quantification



- Based on six unique and two shared peptides from a protein ambiguity group (three G proteins) one cannot decide whether G(i) alpha 1 is actually present in the sample
- Often the quantification accuracy is not sufficient to provide a conclusive result

Nesvizhskii A I , Aebersold R Mol Cell Proteomics 2005;4:1419-1440.

Significance of Inferred Hits

- What is the meaning of a PSM for a protein identification?
  - FDR is calculated on the PSM level
  - 1% FDR means that one in 100 identifications yields a an incorrect protein identification
- This does not mean that there is also an FDR rate of 1% on the protein level!
- In particular in large-scale studies (tens of thousands of spectra), protein FDRs are much higher than peptide FDRs
- PSMs for a large number of (mostly) identical samples
  - Number of correctly identified proteins does not increase significantly with the number of spectra (it is always the same proteins being identified, additional (correct) PSMs do not increase the number of proteins)
  - Number of false positives increases with the number of PSMs (yields hits to random proteins, so initially mostly novel false positives!)

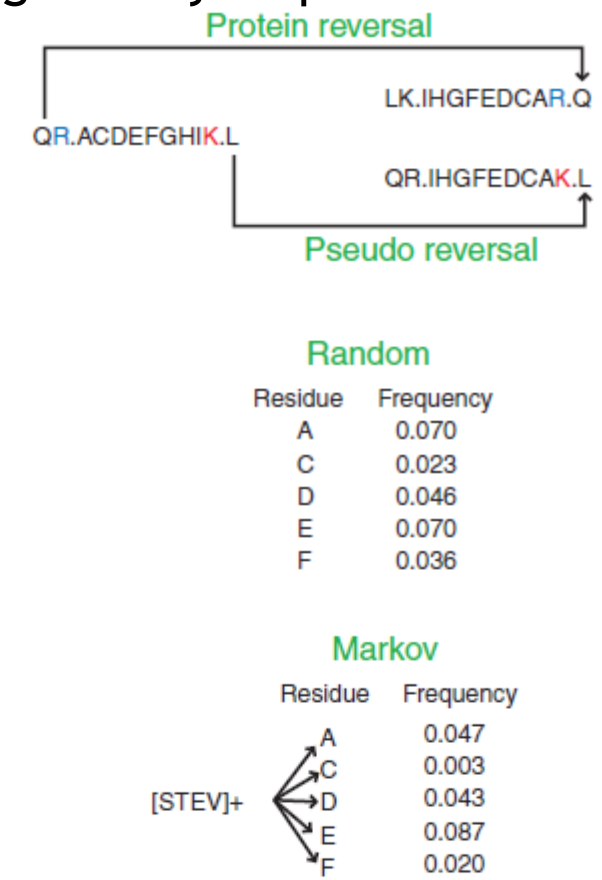
One Hit Wonders

- In many cases, proteins are identified through a single PSM only
- These ‘single hit wonders’ have long been considered problematic: a single false PSM can lead to a wrongly identified protein
- In fact, the so-called ‘Paris guidelines’ for data deposition in proteomics recommend only reporting identifications for which at least two peptides have been identified
- This also became known as the ‘two peptide rule’
- Obviously, just dropping a large part of PSMs is inadequate to address this problem

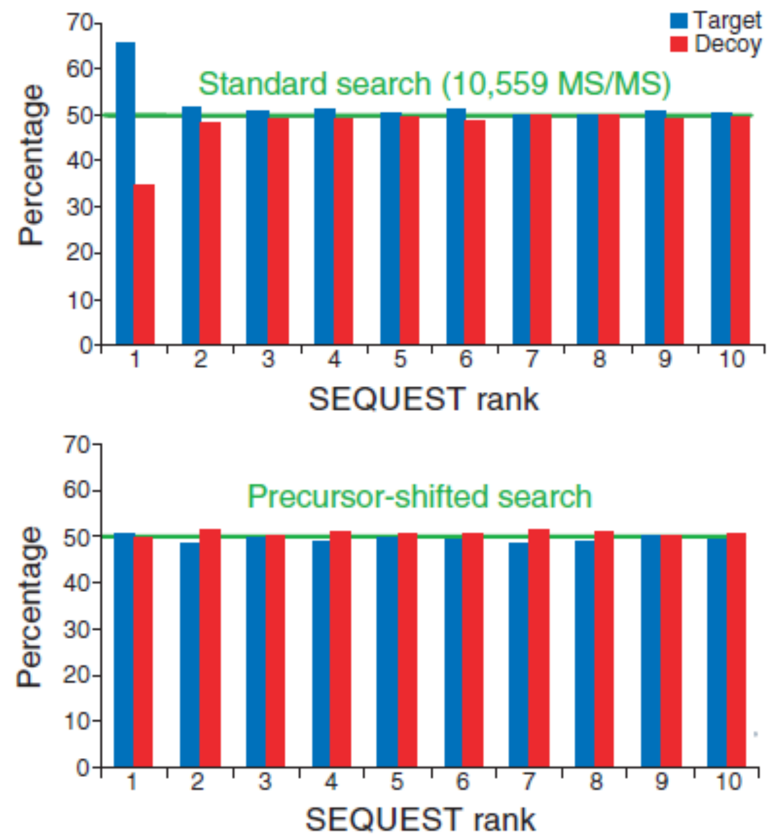
- Bradshaw RA, Burlingame AL, Carr S, Aebersold R. Mol Cell Prot 2006, 5:787-8
  - http://www.mcponline.org/site/misc/ParisReport_Final.xhtml

Recap: Target-decoy databases

Design decoy sequences



Separation of target and decoy results



Recap: FDR Calculation

- General equation for FDR calculation (see statistics lecture)

$$FDR = \frac{FP}{FP+TP}$$

There are two ways how FDRs are calculated based on target-decoy search results:

- Käll et al. suggest (*Käll et al., Proteome Res.* **2008**, *7*, 29- 34)

$$FDR = \frac{\#decoy}{\#target}$$

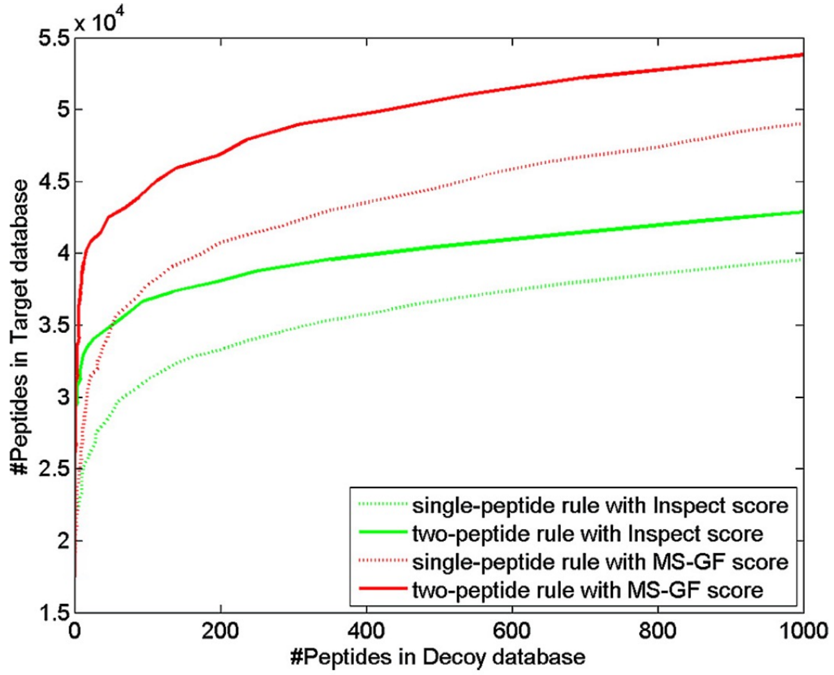
- Zhang et al. suggest (*Zhang et al.*, J Proteome Res 2007;6(9):3549-3557)

$$FDR = \frac{2\#decoy}{\#target+\#decoy}$$

- OpenMS::TOPP::FalseDiscoveryRate uses the *Käll* metrics

One Hit Wonders

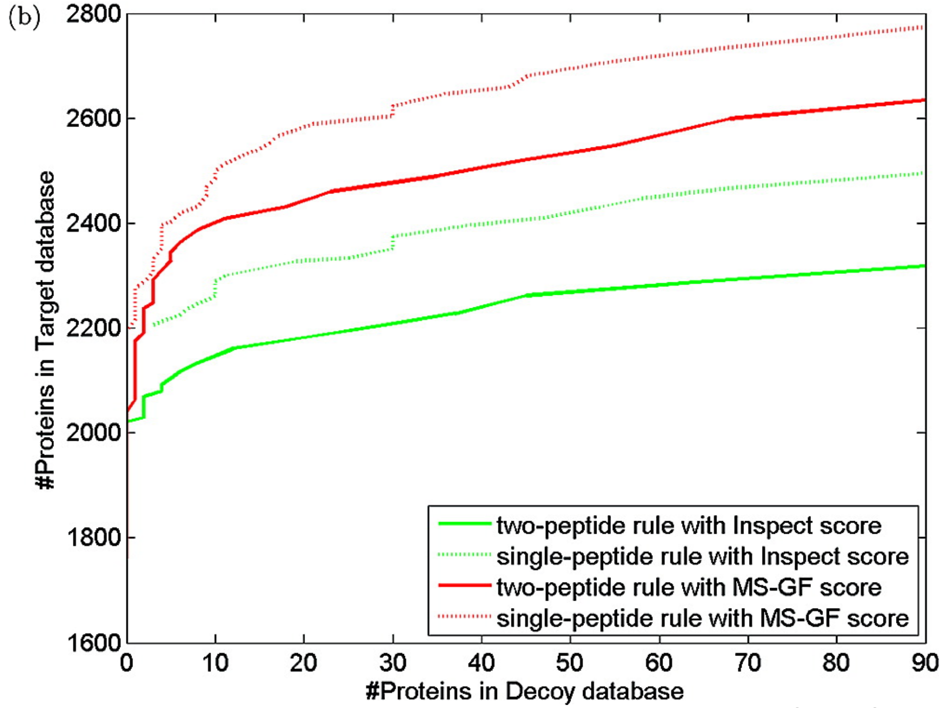
- Gupta & Pevzner argued in 2009 that the application of the two peptide rule actually results in increased false discovery rates
- Removing one-hit wonders should improve the FDR of peptide identifications – this is indeed the case
- For a given number of decoy hits, the number of target peptides increases compared to keeping all PSMs ('single peptide rule')



Gupta & Pevzner, J. Proteome Res. 2009, 8, 4173-4181.

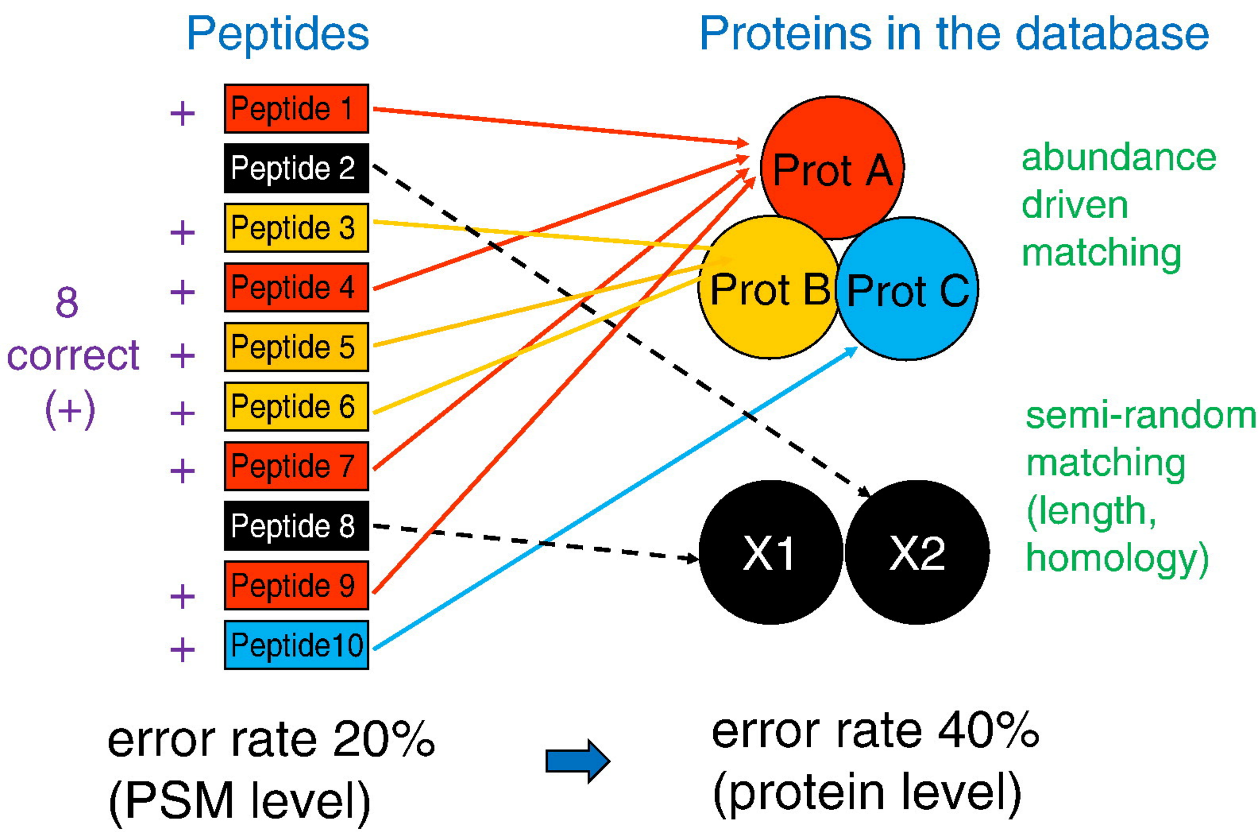
One Hit Wonders

- On the ***protein level*** things are different, however
- For the same dataset, the number of identified proteins is ***higher*** using the single peptide rule than using the two peptide rule at the same FDR!
- More peptide identifications thus do not necessarily imply a higher protein discovery rate



Gupta & Pevzner, J. Proteome Res. 2009, 8, 4173-4181.

Protein FDRs



- Error rates increase when going from peptides to proteins
  - Correct peptide IDs tend to group into a small set of correct proteins
  - Incorrect IDs are semi-random and scatter over the whole protein database

A. Nesvizhskii, J. Proteomics (2010), 73:2092-2123

LEARNING UNIT 9B
PROTEIN PROPHET

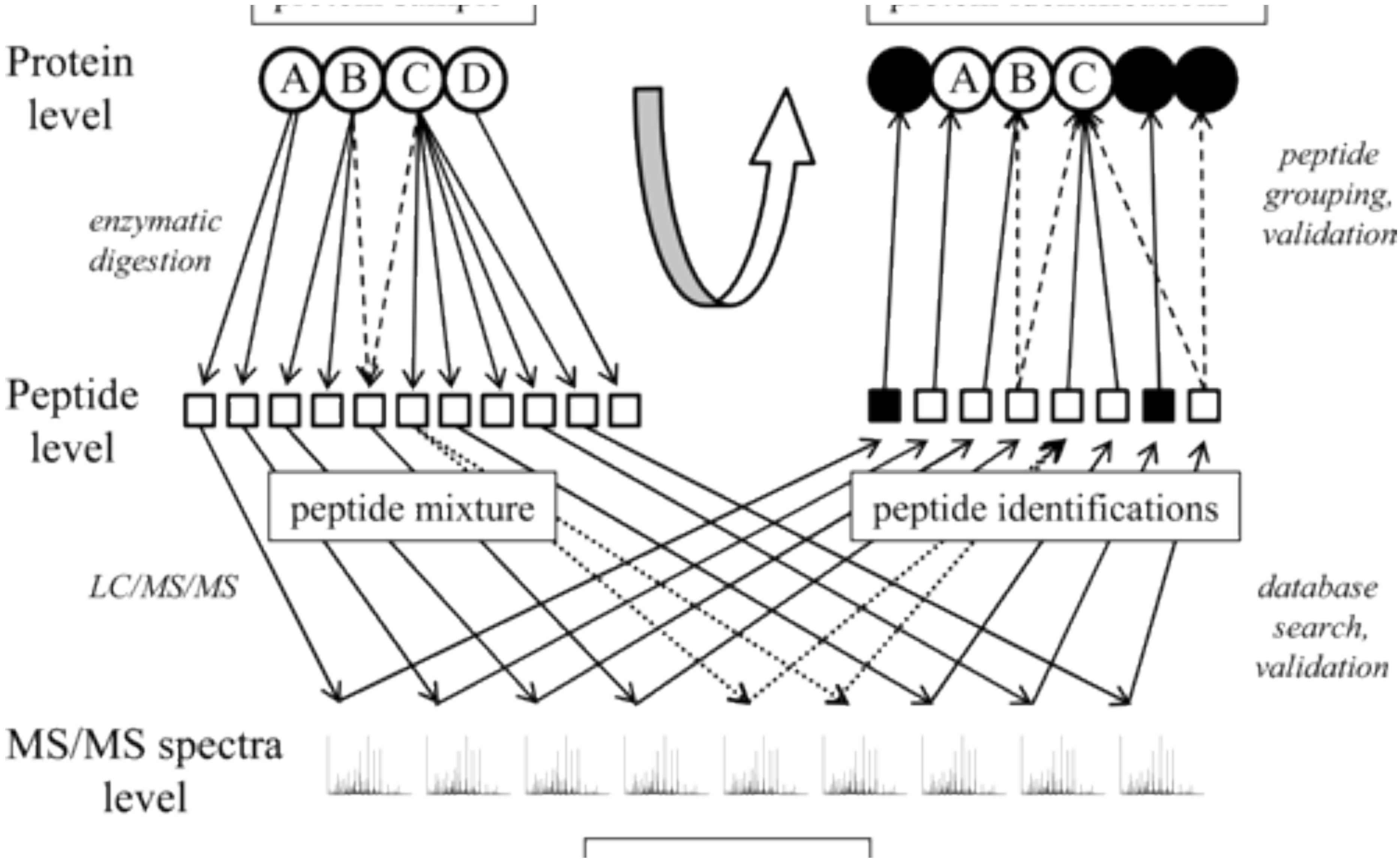
- Peptide probability estimates
- Protein probability estimates
- Sibling peptides correction
- Degenerate peptides

ProteinProphet

- ProteinProphet is an open-source software tool for protein inference and currently one of the standard tools in the area
- **Key ideas**
  - Maximum parsimony approaches to compile protein lists
  - Reporting of protein ambiguity groups
  - **Protein probability estimation**: estimate the probability that a given protein is correctly identified given all evidence for it

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

ProteinProphet - Overview



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

PeptideProphet

- Peptide Probability Estimates (PPE)
  - Computed by **PeptideProphet**
  - Converts search engine scores into a *probabilities*
  - Similar ideas have been discussed in the context of consensus identification
  - PeptideProphet uses **expectation maximization** to compute a **mixture model** of the score distributions of correct and incorrect PSMs
  - Given a PSM and a search engine score, we can thus compute a probability that the PSM is correct
- In contrast to a (raw) score, PPEs are a simple way to determine the trust in each individual PSM

Nesvizhskii, et al., Anal. Chem. (2002), 74, 5383-5392

Protein Probability Estimates

- Given the PPEs, we can easily compute the probability for each of the induced protein IDs
- Assuming all peptides are unique, we can compute the probability P for an protein identification as 1 minus the probability of all peptide identifications inducing this peptide being wrong
- We could do this on the peptide level quite simply as follows:

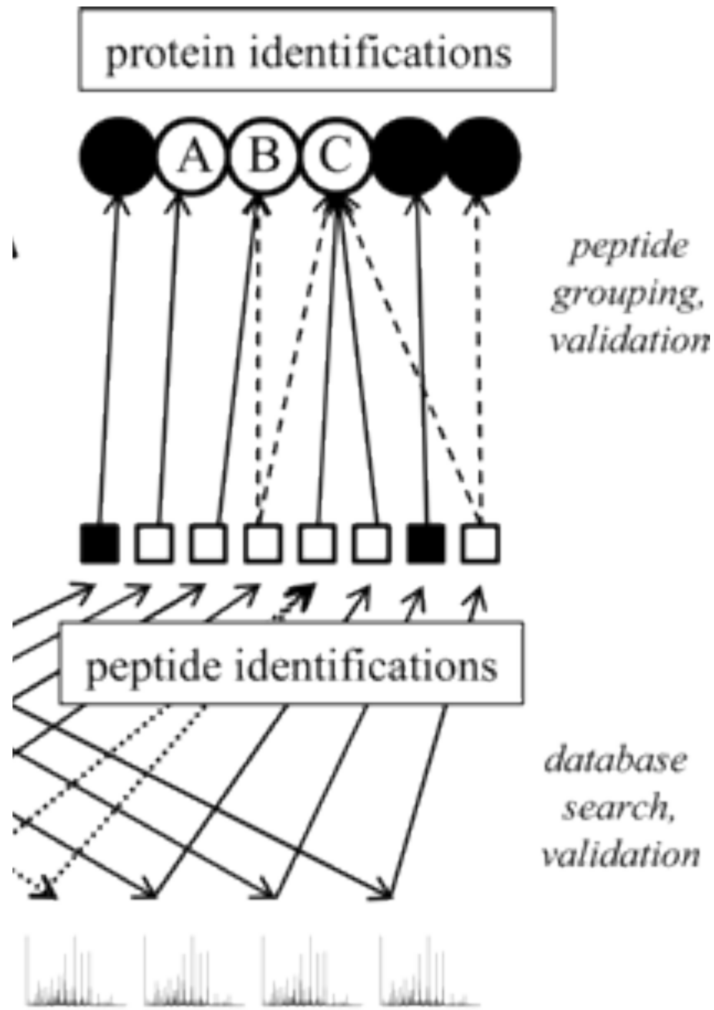
$$P = 1 - \prod_i (1 - p_i)$$

  with probabilities p_i for the peptide identification of peptide *I* being correct
- However, we also need to consider multiple evidence for different spectra giving evidence for the same peptide

Protein Probability Estimates

- We thus need to consider probabilities for each PSM independently
- Each PSM is assigned a PPE by PeptideProphet
- Probability that a protein is **not** present in a sample despite its PSMs depends on the probabilities $p(+|D_i^j)$ for the peptide ID of peptide *i* based on the observed data (spectrum) *j* being correct
- We can thus compute *P* based on PPEs of all PSMs:

$$P = 1 - \prod_i \prod_j (1 - p(+|D_i^j))$$



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

Protein Probability Estimates

- There are a few problems with this:
  - **PSMs are not independent**

    There is a high probability for multiple spectra of the same peptide to hit the same incorrect ID if the spectra are of high quality, but do not match the database (e.g., due to post-translational modification)
  - **Ambiguous peptide-protein matches**

    If a peptide matches multiple proteins, its evidence cannot simply be shared across these proteins

Protein Probability Estimates

- A simple way to deal with multiple PSMs is to
  - Include each peptide just once
  - Consider only the PSM with the best PPE of all PSMs to the same peptide:

$$p_i = \max_j p(+|D_i^j)$$

- P would then be computed as follows:

$$P = 1 - \prod_i (1 - \max_j p(+|D_i^j)) = 1 - \prod_i (1 - p_i)$$

- This procedure yields a more conservative estimate of protein probabilities

ProteinProphet

Example:

>gi|125910|sp|P02754.3|**LACB_BOVIN**
MKCLLLALALTCGAQALIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLR**VYVEELKPTPEGDL**
EILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKVLVLDTDYKKYLLFCMENSAEPEQSLACQCLVR
TPEVDDEALEKFDKALKALPMHIR**LSFNPTQLEEQCHI**

LSFNPTQLEEQCHI : $p = 0.48$
LSFNPTQLEEQCHI : $p = 0.65$
max = 0.65

TPEVDDEALEK : $p = 0.91$

VYVEELKPTPEGDLEILLQK : $p = 0.81$

$$P(LACB\_BOVIN) = 1 - (1 - 0.81)\,(1 - 0.91)\,(1 - 0.65) = 0.99$$

After: Nesvizhskii, et al., Anal. Chem. (2003), 75, 4646-4658

Sibling Peptides

- Correct assignments tend to cluster to the same proteins
- Incorrect assignments tend to be hits to proteins with no other assigned peptides
- As a result, the computed PPEs, while correct in the context of the ***whole*** dataset, need to be corrected for an accurate estimate in the context of their source protein
- ProteinProphet introduces the notion of **sibling peptides**
- Sibling peptides are peptides hitting the same protein
- Rather than counting them, ProteinProphet defines the number of sibling peptides NSP_i for a peptide *i* as the sum of the PPEs:

$$\mathrm{NSP}_i = \sum_{\{m \mid m \neq i\}} p(+ \mid D_m)$$

where the sum runs over all peptides *m* hitting the same protein as *i* and PPEs p_i are the maximum values for a given peptide reached in the dataset

Sibling Peptides

Example:

>gi|125910|sp|P02754.3|**LACB_BOVIN**
MKCLLLALALTCGAQALIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLR**VYVEELKPTPEGDL**
EILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKVLVLDTDYKKYLLFCMENSAEPEQSLACQCLVR
TPEVDDEALEKFDKALKALPMHIR**LSFNPTQLEEQCHI**

LSFNPTQLEEQCHI : $p = 0.48$
LSFNPTQLEEQCHI : $p = 0.65$
max = 0.65

TPEVDDEALEK : $p = 0.91$

VYVEELKPTPEGDLEILLQK : $p = 0.81$

NSP(VYV…) = 0.91 + 0.65 = 1.56
NSP(TPE…) = 0.65 + 0.81 = 1.46
NSP(LSF…) = 0.91 + 0.81 = 1.72

After: Nesvizhskii, et al., Anal. Chem. (2003), 75, 4646-4658

Sibling Peptides

- Intuitively, one would trust identifications with a high NSP more than those with a low NSP (more evidence per protein)
- We can thus refine PPEs in the context of the source protein as follows:

$$p(+|D, NSP) = \frac{p(+|D)p(NSP|+)}{p(+|D)p(NSP|+) + p(-|D)p(NSP|-)}$$

with

- *p(NSP|+)* and *p(NSP|-)* being the probabilities of having a particular NSP value for correct/incorrect assignments
- *p(+|D)* and *p(-|D)* are the uncorrected probabilities for the peptide assignment being correct/incorrect

Sibling Peptides

- Values for *p(NSP|+)* and *p(NSP|-)* can be computed for the whole dataset
- NSP values are binned and counted for correct and incorrect assignments

$$p(NSP|+) = \frac{1}{N\,p(+)} = \sum_{\{i|NSP_i \in k\}} p(+|D_i, NSP_i)$$

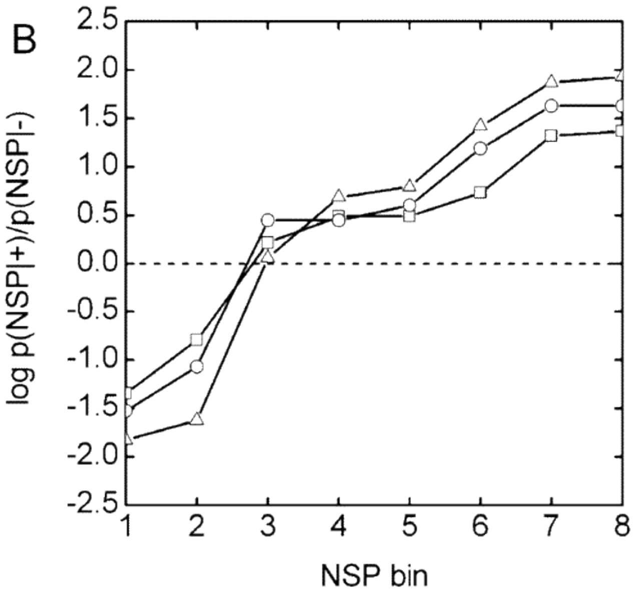
where *N* is the total number of peptides assignments and *p(+)* is the prior probability of a peptide identification being correct

- *p(+)* can be computed by summation over all peptide identifications of the dataset:

$$p(+) = \frac{1}{N} \sum_i p(+|D_i, NSP_i)$$

NSP Distributions

- NSP distributions can be determined using expectation maximization
- As a first guess, unadjusted p(+|D) values are used to compute an estimated NSP value for each assignment
- Applying EM then yields adjusted probabilities, this is repeated until convergence has been reached
- NSP distributions depend on the dataset and the dataset size



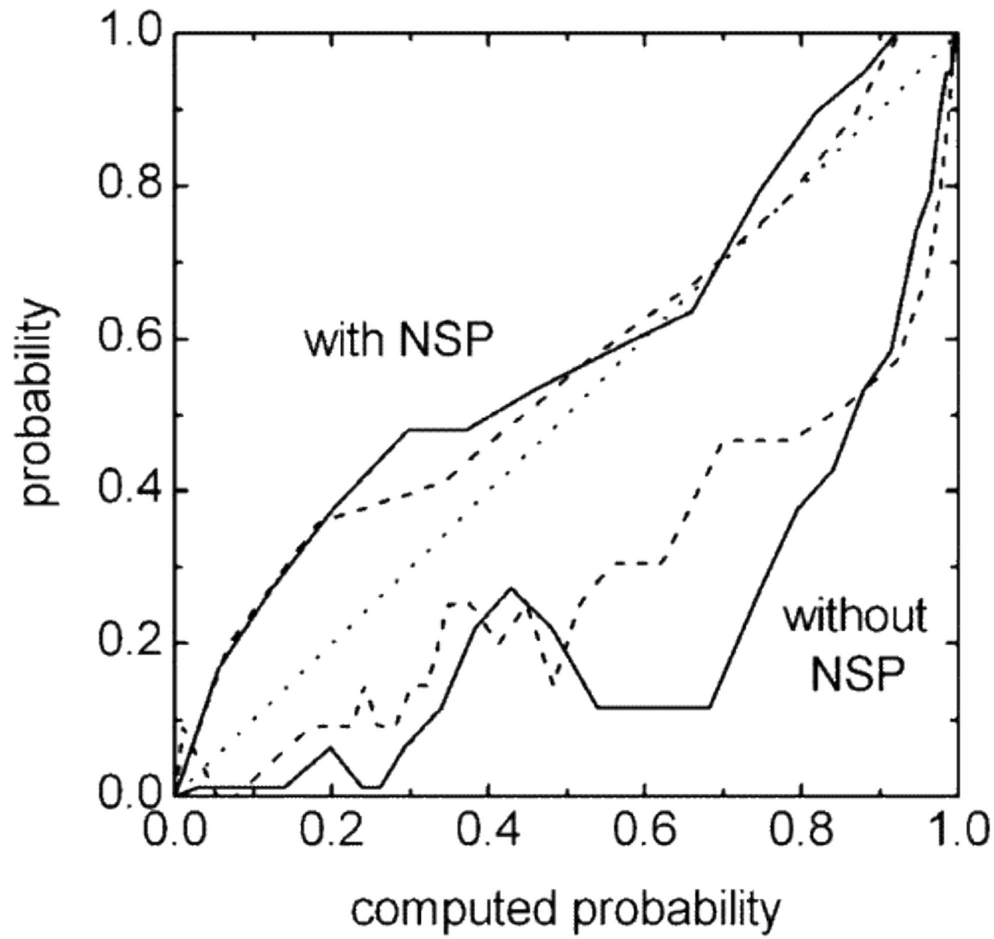
NSP distribution for datasets of varying size:

- squares: single run of a low-complexity sample
- circles: four runs of the same sample
- triangles: 22 runs

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

Influence of NSP Correction

- NSP correction yields better predictions of protein probabilities
- Figure on the right shows the predicted vs. true protein probabilities with and without NSP
- Different lines correspond to different datasets
- Dotted line: perfect prediction



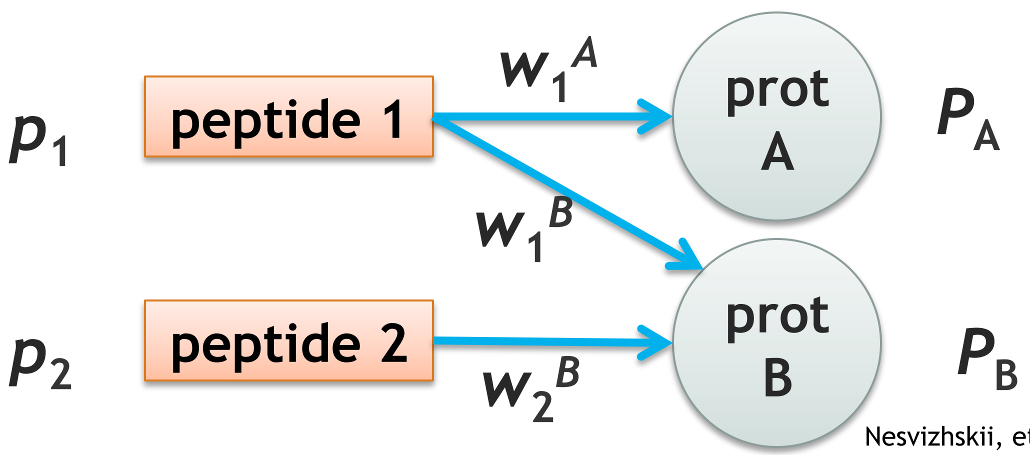
Nesvizhskii, et al., Anal. Chem. (2003), 75, 4646-4658

Protein Ambiguity

- Shared peptides within a PAG cause issues as well
- Their probabilities can be distributed over their potential source proteins through a weighting scheme based on the protein probabilities:

$$P_n = 1 - \prod_i \left(1 - w_i^n \, p(+|D_i)\right) \qquad w_i^n = \frac{P_n}{\sum_{s=1 \ldots N_s} P_s}$$

- Weights w_i^n are again estimated iteratively using an EM-like algorithm



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

Protein Ambiguity Group

PROTEIN GROUP 1: "flagellin_precursor"

1 **FLA4_HALN1 1.00**

>FLA4_HALN1 (P13077) Flagellin B2 precursor

1.00	1	INTAGY	1.00 / 1.00	4
1.00	1	STIQWIGPDTATTL	1.00 / 1.00	4
1.00	2	GSATGEEASAQVSNR	1.00 / 1.00	4
1.00	2	ANVPESLK	0.92 / 0.90	4
1.00	1	INIVSAY	0.86 / 0.83	4

2 **FLA1_HALN1 0.00**

>FLA1_HALN1 (P13074) Flagellin A1 precursor

0.00	2	GSATGEEASAQVSNR	1.00 / 1.00	4
0.00	1	STIQWIGPDTATTL	1.00 / 1.00	3
0.00	2	ANVPESLK	0.92 / 0.90	4
0.00	1	INIVSAY	0.86 / 0.83	4

3 **Q9HQT8 0.00**

>Q9HQT8 Flagellin A2 precursor

0.00	2	GSATGEEASAQVSNR	1.00 / 1.00	3
0.00	1	INIVSAY	0.83 / 0.83	3
0.00	2	ANVPESLK	0.78 / 0.90	2

4 **Q9HQX4 FLA3_HALN1 0.00**

>Q9HQX4 Flagellin B3 precursor

>FLA3_HALN1 (P13076) Flagellin B1 precursor

0.00	2	GSATGEEASAQVSNR	1.00 / 1.00	2
0.00	1	INTAGY	1.00 / 1.00	2
0.00	1	INIVSAY	0.83 / 0.83	2

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

LEARNING UNIT 9C
PROTEIN FDR CALCULATION

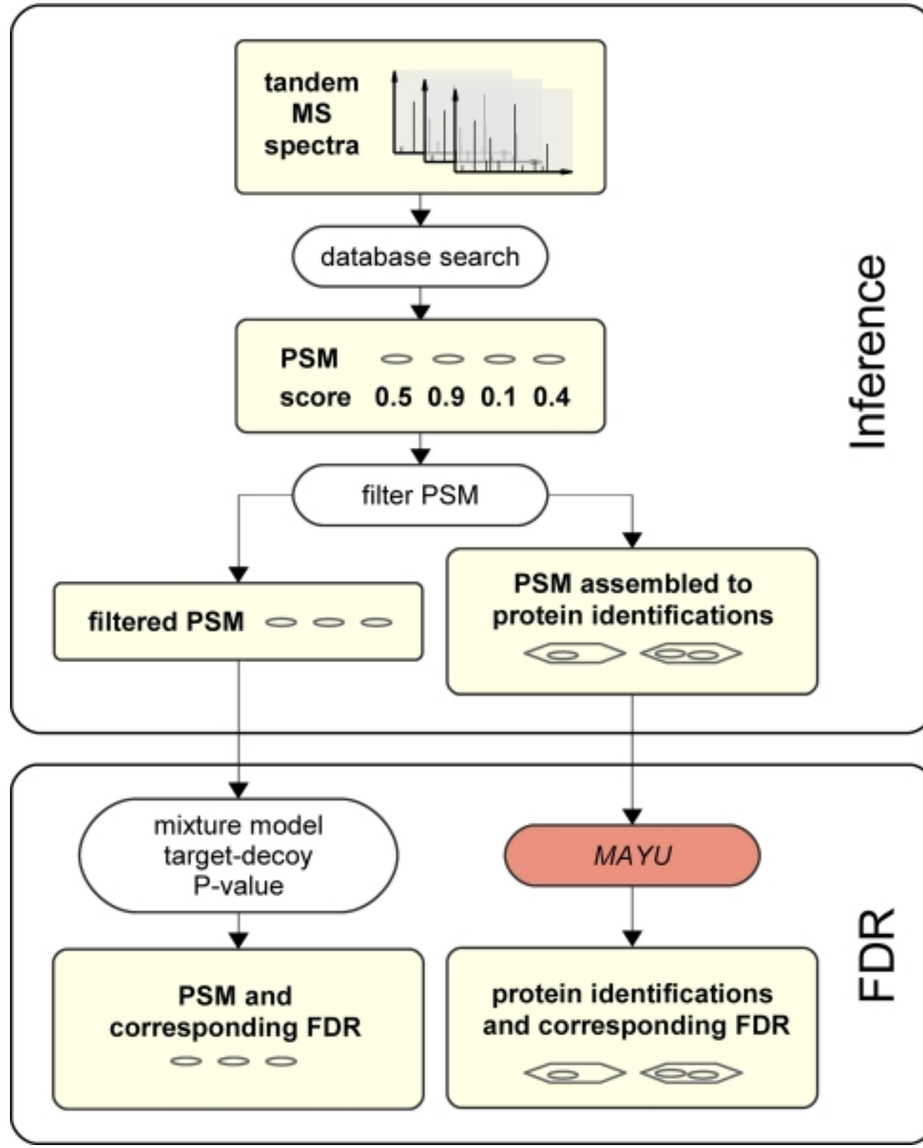
- Protein FDR calculation
- MAYU

Estimating Protein FDRs

- Peptides FDRs do not correspond to protein FDRs
- Currently, large-scale studies often have dozens or hundreds of LC-MS runs that are being accumulated
- Repeated measurements lead to an accumulation of false positive identifications
- As a rule of thumb, protein FDR increases linearly with the number of repeat measurements
- FDRs can be estimated in the same fashion as PSM FDRs through a naïve target-decoy approach

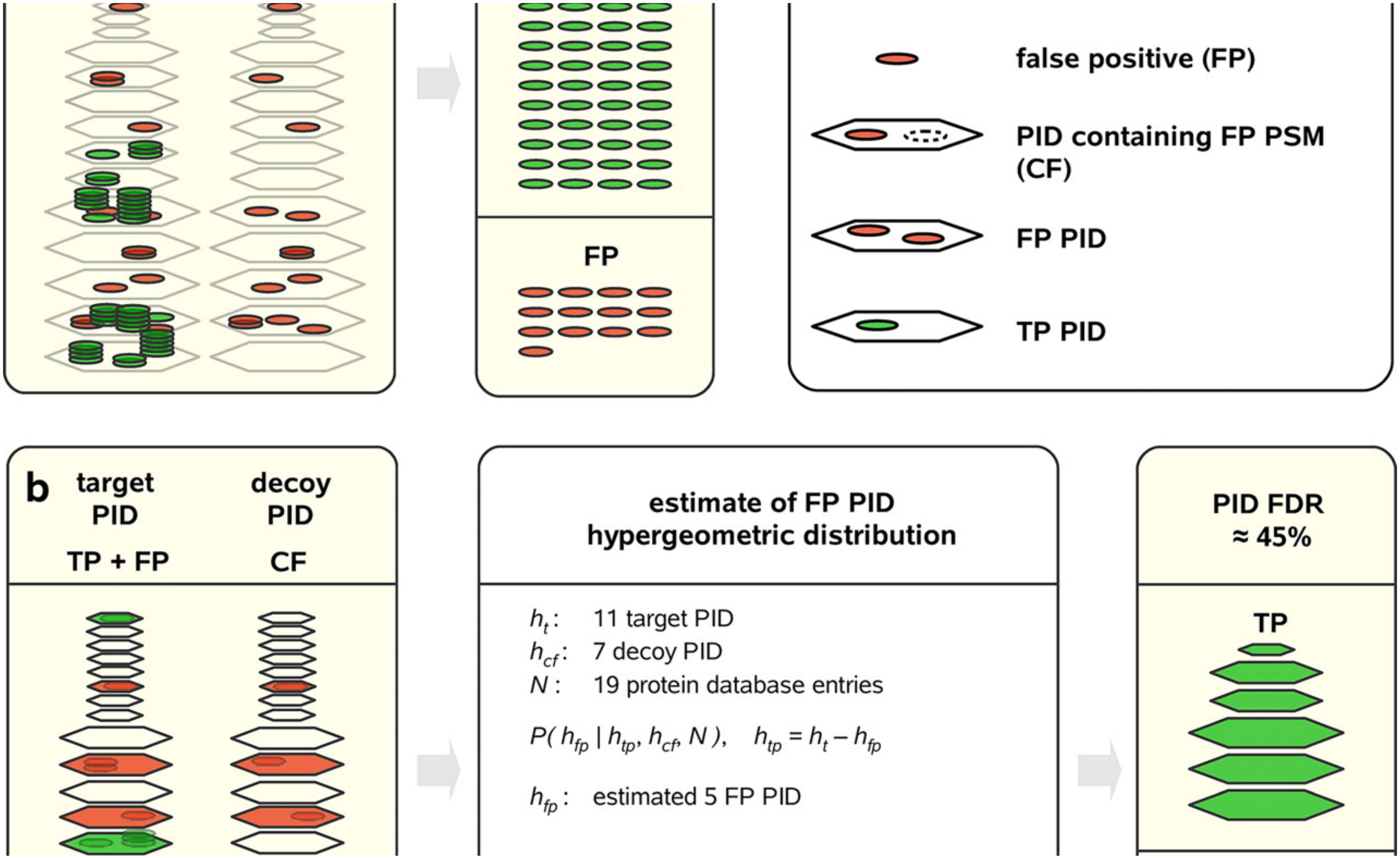
MAYU

- MAYU estimates protein FDRs for large-scale datasets
- The approach is similar to the PSM FDR determination done in PeptideProphet, but on the level of proteins
- MAYU fits a hypergeometric distribution to determine the expected number of false positives



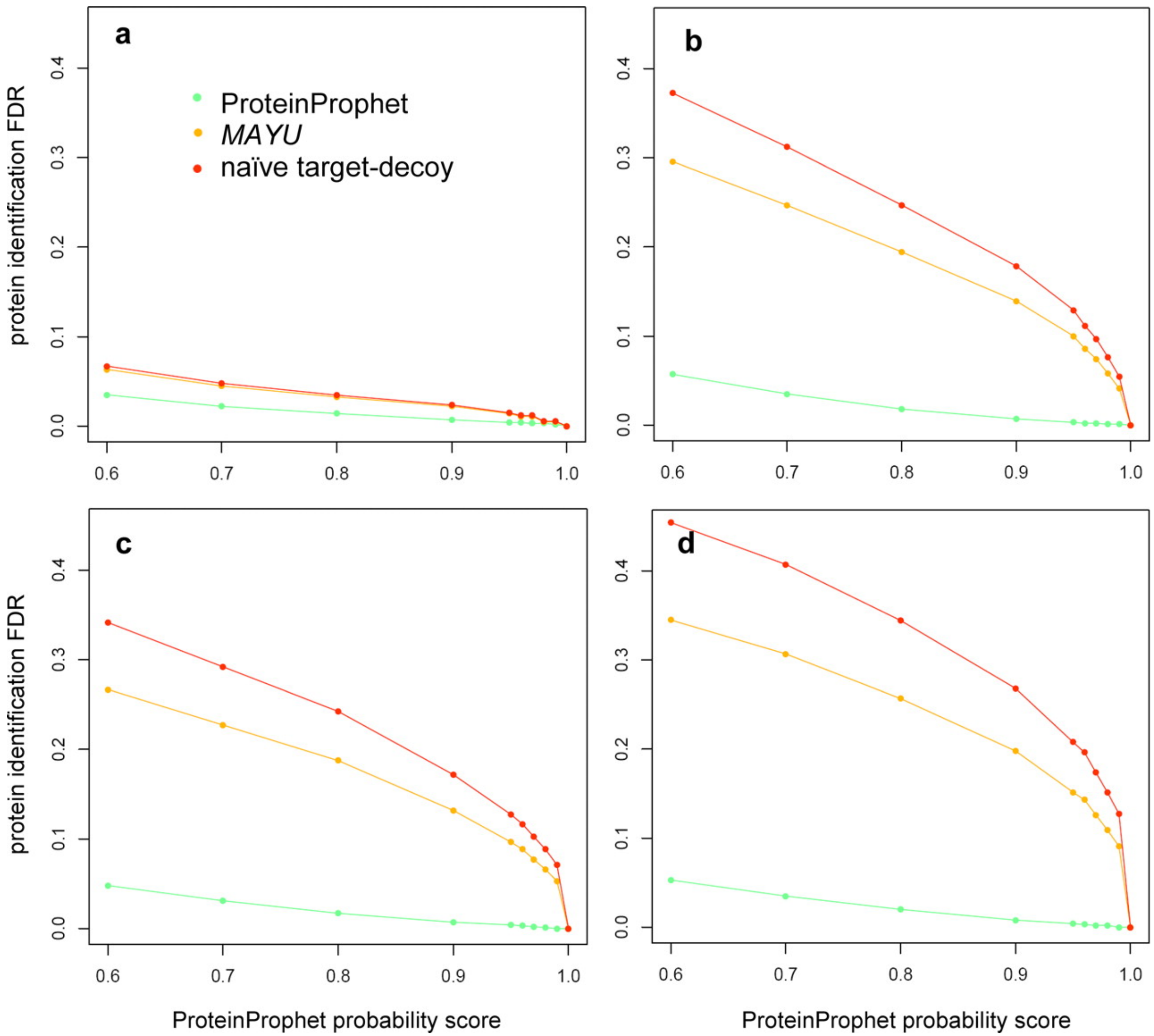
Reiter et al., Mol. Cell. Proteomics, 2009, 8, 2405-2417

MAYU



Reiter et al., Mol. Cell. Proteomics, 2009, 8, 2405-2417

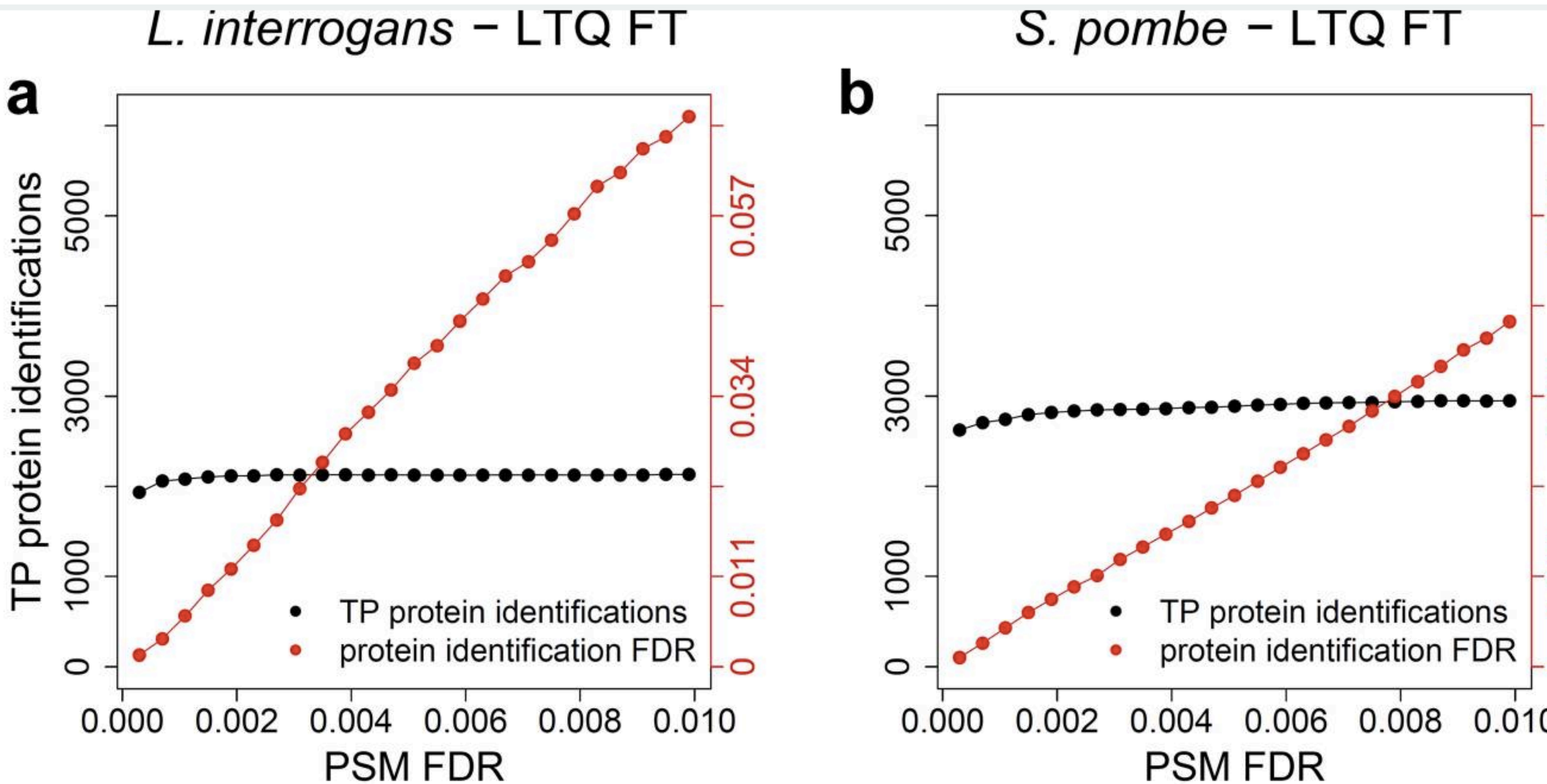
MAYU vs. ProteinProphet



a. 1 run
b. 5 runs
c. 10 runs
d. 20 runs

Reiter et al., Mol. Cell. Proteomics, 2009, 8, 2405-2417

MAYU

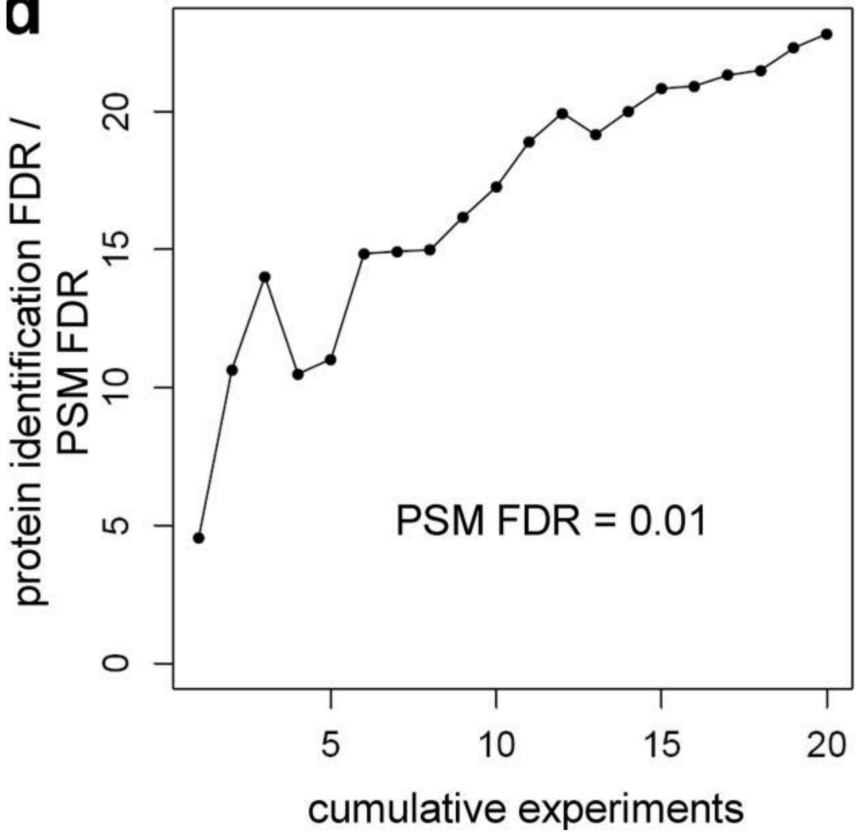


- Interestingly, increasing the PSM FDR does not yield an increased rate of true protein identification
- Currently popular values of 1-5% PSM FDR seem to be much to high and yield very large protein FDRs (>10%)

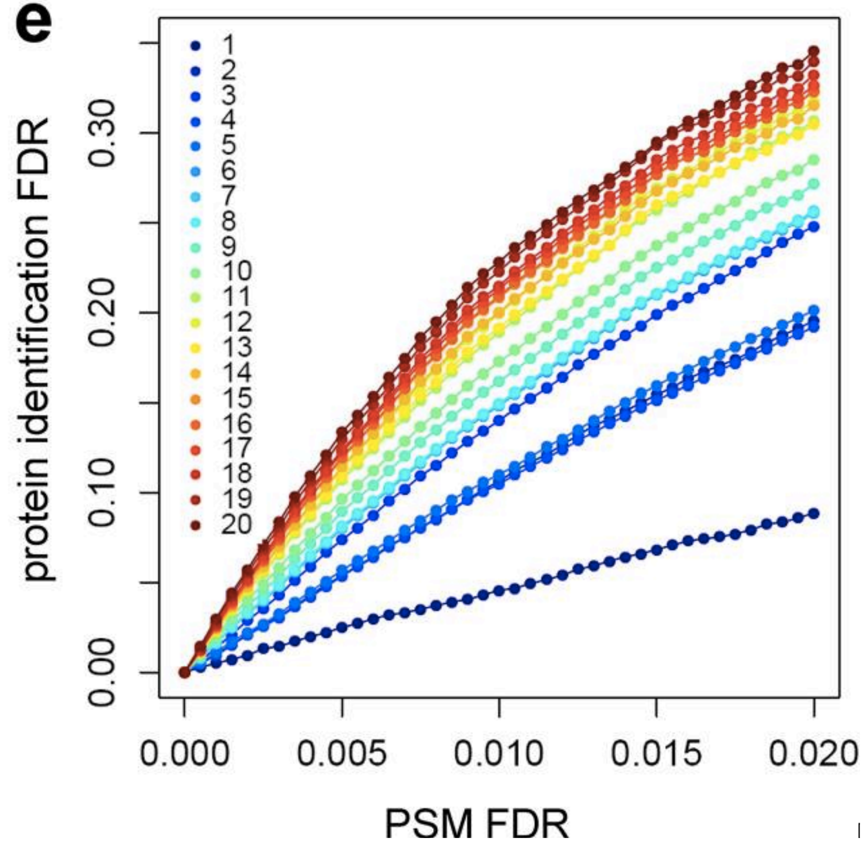
Reiter et al., Mol. Cell. Proteomics, 2009, 8, 2405-2417

MAYU

d



e

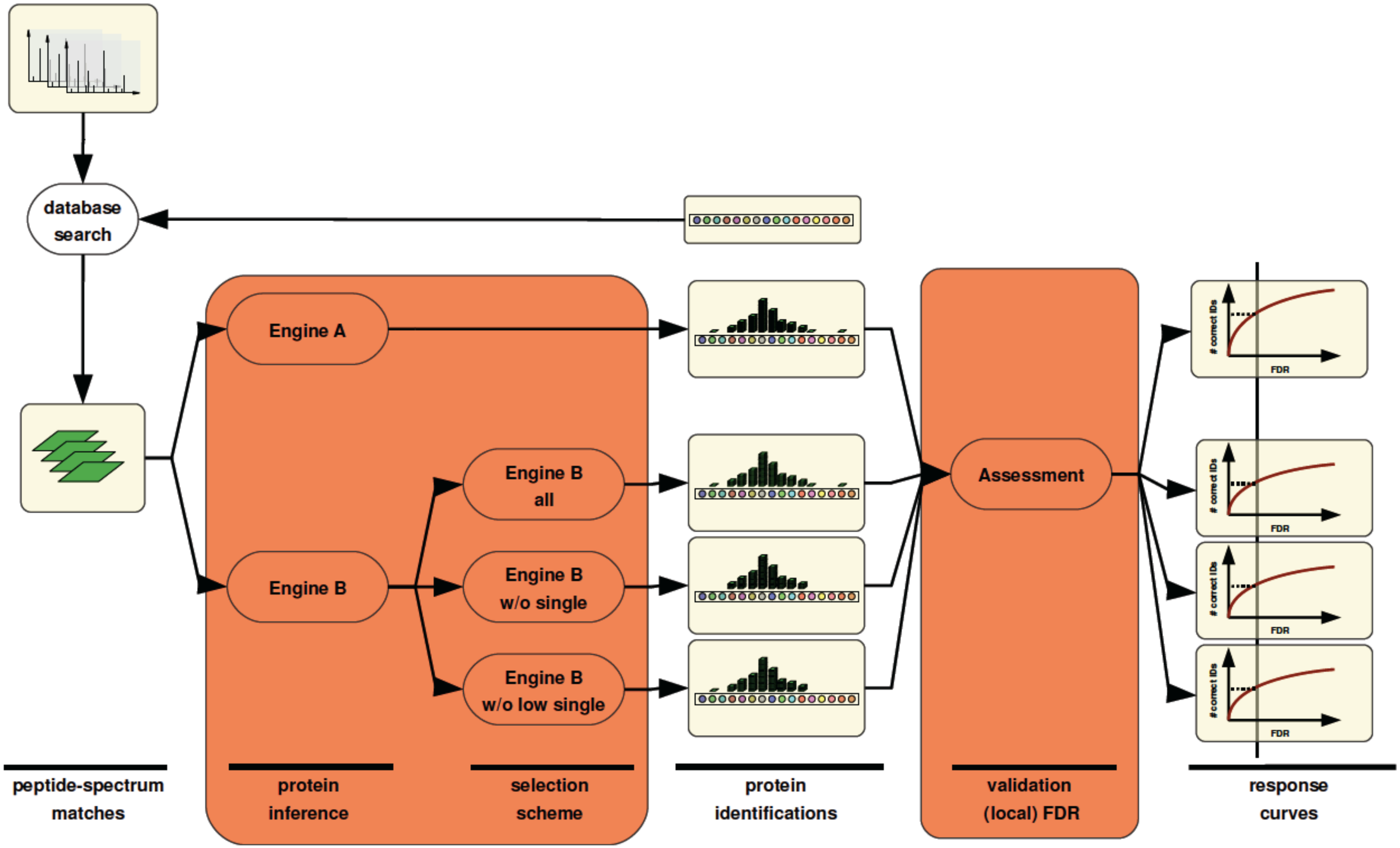


- These figures show the increase of protein FDR with the number of repeat measurements (right: color = number of runs)
- As can be seen from these plots, large-scale studies are particularly prone to FP accumulation
- Protein FDRs can thus easily reach values of over 50%, i.e. half of reported protein identifications can be incorrect!

Reiter et al., Mol. Cell. Proteomics, 2009, 8, 2405-2417

Benchmarking Inference Engines

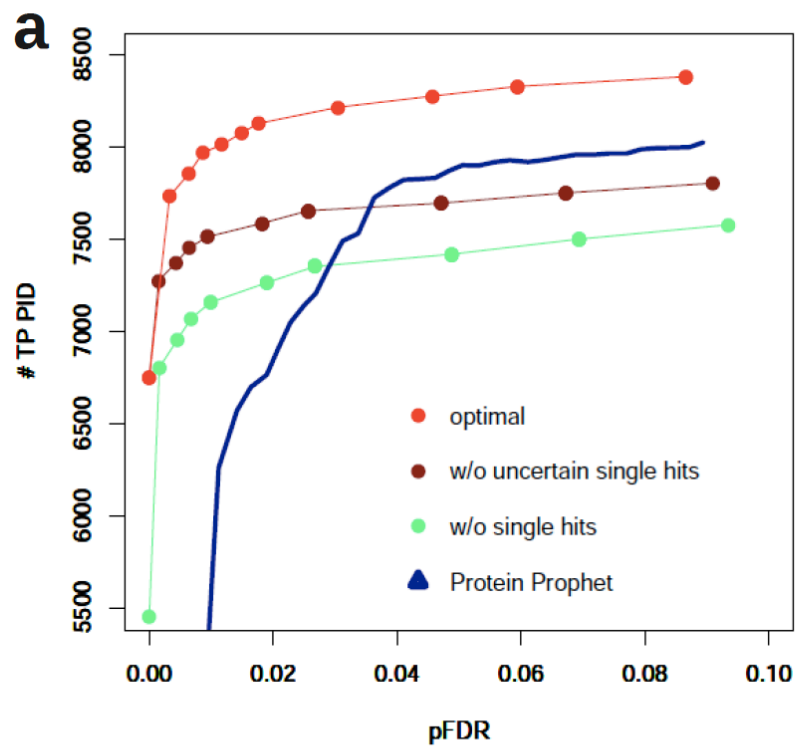
- With MAYU it is possible to benchmark different protein inference engines and PSM selection strategies (e.g., two-peptide vs. single-peptide rule)



Claassen et al., Mol Cell Proteomics (in press)

Benchmarking Inference Engines

- **Conclusions**
  - Keep all high quality hits, independent of whether they are single-hit wonders or not
  - Stringent FDR filtering on the PSM level is required to get a good protein FDR
  - Optimal strategy might depend on the dataset and on the organism (database size!)



Claassen et al., Mol Cell Proteomics (in press)

References

- **One-hit wonders, two peptide rule**
  - http://www.mcponline.org/site/misc/ParisReport_Final.xhtml
  - Gupta, Pevzner, False Discover Rates of Protein Identifications: A Strike against the Two-Peptide Rule, J. Proteome Res. 2009, 8, 4173-4181.
- **Protein inference methods**
  - Nesvizhskii A I , Aebersold R, Interpretation of Shotgun Proteomics Data, Mol Cell Proteomics 2005;4:1419-1440
  - Nesvizhskii, Keller, Kolker, Aebersold, A Statistical Model for Identifying Protein by Tandem Mass Spectrometry, Anal. Chem. 2003, 75, 4646-4658.
  - Keller, Nesvizhskii, Kolker, Aebersold, Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search, Anal. Chem. 2002, 74, 5383-5392
  - ProteinProphet and PeptideProphet:

    http://proteinprophet.sourceforge.net
- **Protein FDR Estimation (MAYU) and inference engine benchmarking**
  - Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, Hengartner MO, Aebersold R, Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry, Mol Cell Proteomics. 2009, 8:2405-17
  - Claassen, Reiter, Hengartner, Buhmann, Aebersold, Generic Comparison of Protein Inference Engines, Mol. Cell. Proteomics (in press, PMID: 22057310)