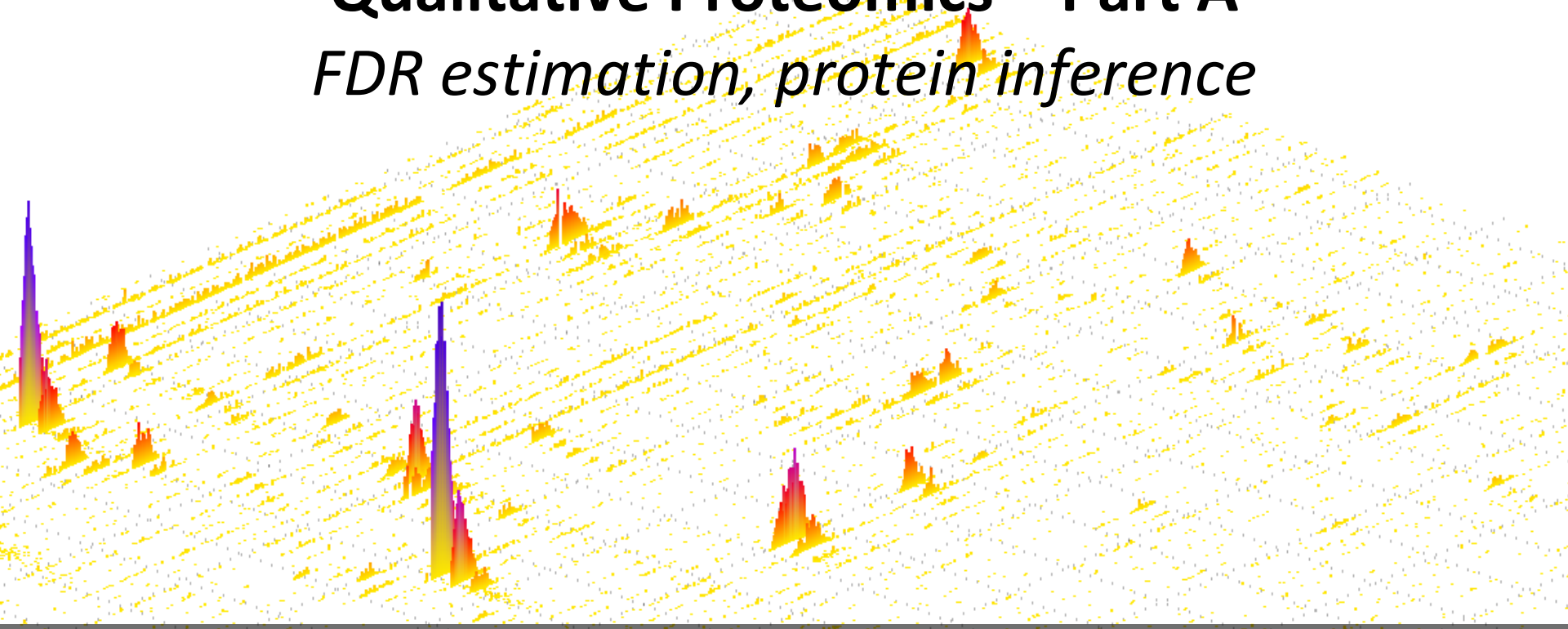
May Institute 2017

Computation and statistics for mass spectrometry and proteomics

Qualitative Proteomics – Part A

FDR estimation, protein inference



Oliver Kohlbacher

University of Tübingen and

MPI for Developmental Biology

KohlbacherLab.org | @okohlbacher

MAX-PLANCK-GESELLSCHAFT

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Today's Schedule

Monday 5/1/2017	Proteomics and metabolomics with OpenMS
8:00 AM	Registration
9:00 AM	Fundamentals of non-targeted proteomics and metabolomics
10:30 AM	**Refreshments**
11:00 AM	Hands-on: Tutorial: Introduction to OpenMS and KNIME
12:30 PM	**Lunch Break**
1:30 PM	Lecture: Database search, peptide-spectrum matches.
2:00 PM	Hands-on: Peptide and protein identification by database search.
3:00 PM	**Refreshments**
3:30 PM	Lecture: FDR estimation, protein inference, quality control.
4:00 PM	Hand-on: Peptide and protein identification by database search.
5:00 PM	Improvised poster session
6:00 PM	**Adjourn**

Overview

- **Concepts of Database Search**
  - Fundamentals of peptide fragmentation, ion series
  - Database search: key ideas
- **Database Search Engines**
  - X!Tandem
  - Scoring function and underlying statistics
  - SEQUEST

FDR ESTIMATION

Database Settings

- The database should contain all protein sequences that are potentially in the sample (e.g., all human proteins of your looking at proteomics data from human cell lines)
- From the database and the enzyme's 'cutting rule' settings, the peptide candidates are calculated
- Apart from the expected proteins, the database should also contain common contaminants, such as trypsin (or other enzymes), keratins or BSA (bovine serum albumin, often used for instrument calibration)
- Databases can also be designed in a way to give an intuitive idea of false discovery rates -> **target/decoy databases**

Target-Decoy Databases

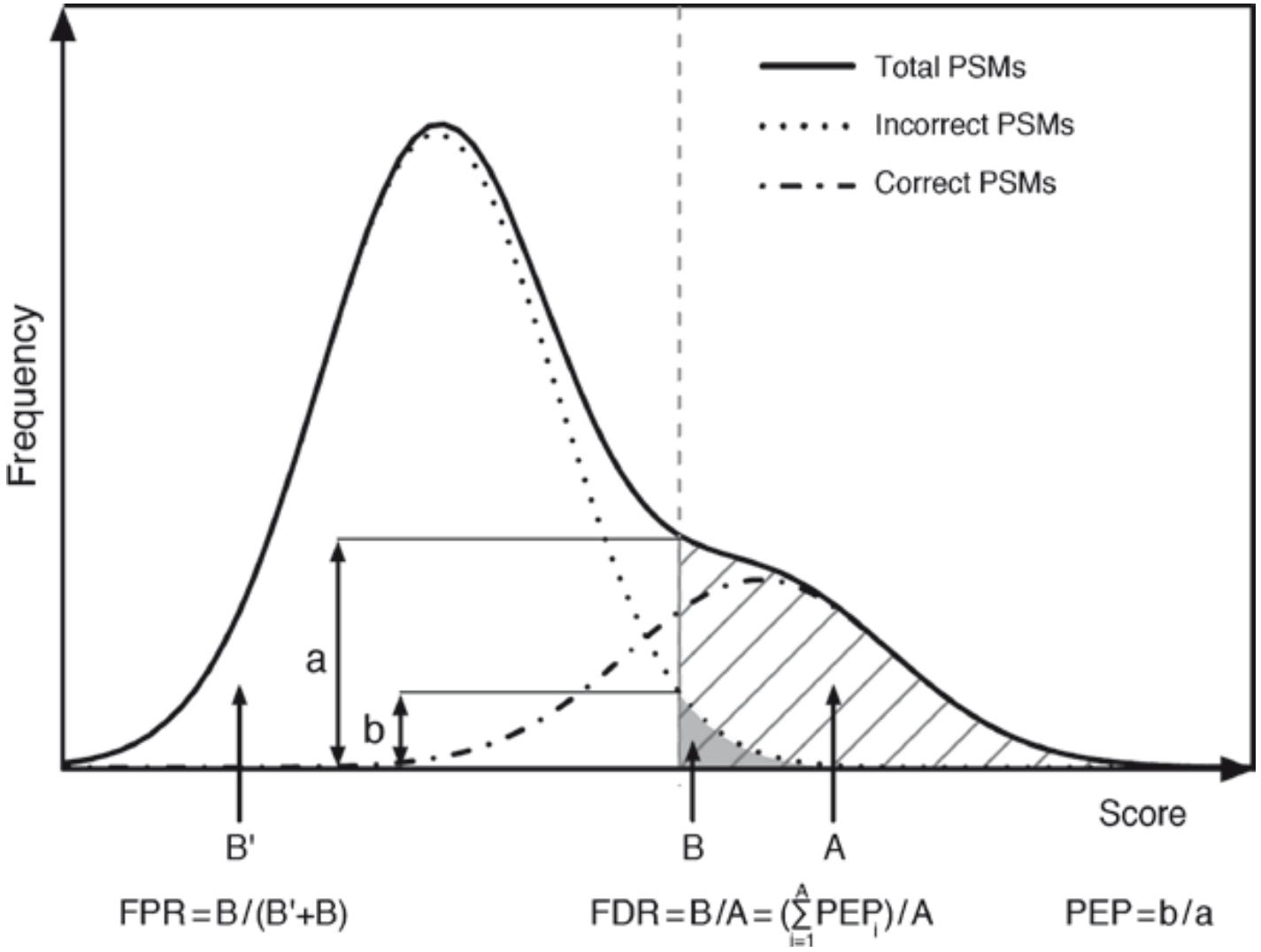
- Take the original protein sequences (target sequences) and reverse, pseudo-reverse, randomize or shuffle these sequences to create **decoy sequences**
- Spectra are either searched twice (first against target, then decoy database) or against the concatenated database (target + decoy)
- Decoy sequences are random sequences and should not be present in the sample
- PSMs against decoy proteins have to be **false positive identifications**
- Note
  - The decoy database design should provide equal numbers of decoy peptides as there are target peptides per search space (with randomized sequences this is hard to control)
  - Ideally one should avoid large overlap between target and decoy peptides

Target-Decoy Approach

Peptide identification	Search engine score	TARGET/DECOY
LCEVEEGDKEDVDK	s_1	TARGET
YTAQVDAEEKEDVK	s_2	TARGET
IVADKDYSVTANSK	s_3	TARGET
TGIEIIKK	s_4	TARGET
DLGEEHFK	s_5	TARGET
TASSDTSEELNSQDSPK	s_6	**DECOY**
GAGGENEPPAAAPEPR	s_7	TARGET
IKDPDAAKPEDWDDR	s_8	TARGET
VDEVGGEALGR	s_9	TARGET
SEEQLKEEGIEYK	s_{10}	**DECOY**
LHVDPENFK	s_{11}	TARGET
FSTVAGESGSADTVRDPR	s_{12}	TARGET
AEEDEILNR	s_{13}	**DECOY**

- PSMs are sorted by (deteriorating) score
- As the score gets worse, the likelihood of finding a decoy hit increases, likelihood for target hit decreases
- By choosing an appropriate score threshold, one can ensure a given false-discovery rate (FDR)

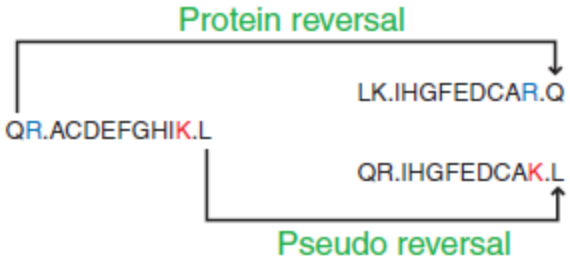
PSM Score Distribution



Brosch and Choudhary, Methods in Molecular Biology, 2010, Volume 604, 43-5

Target-Decoy Approach

Design decoy sequences



Random

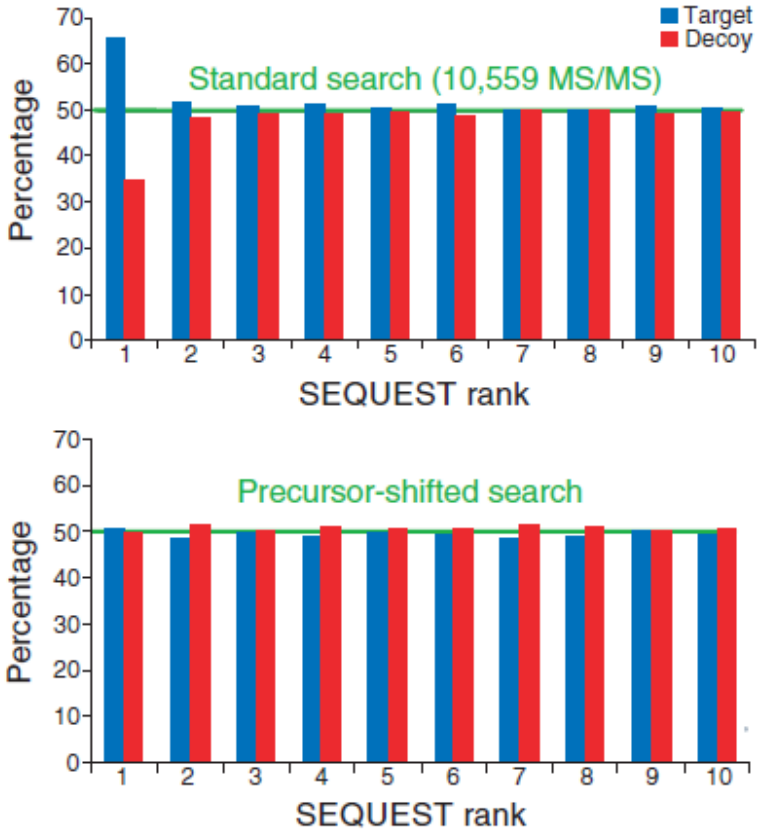
Residue	Frequency
A	0.070
C	0.023
D	0.046
E	0.070
F	0.036

Markov

Residue	Frequency
A	0.047
C	0.003
D	0.043
E	0.087
F	0.020

[STEV]+

Separation of target and decoy results



Although different decoy database designs produce very similar results, the most frequently used approaches are the reversed and pseudo-reversed decoy databases

Elias and Gygi, Nature Methods. Vol. 4, No. 3, March 2007

Calculation of FDRs

- 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 tool **FalseDiscoveryRate** uses the *Käll* metrics

THE PROTEIN INFERENCE PROBLEM

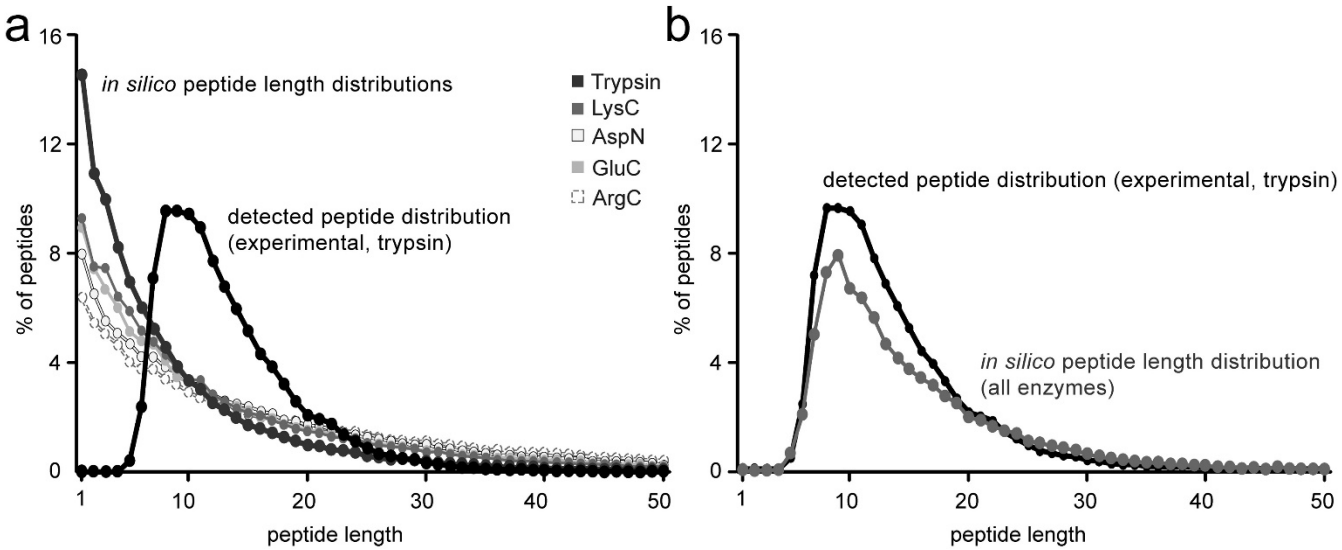
- Problem definition
- Protein families
- Protein ambiguity groups
- Inference through quantification
- Significance of inferred hits
- One hit wonders

Identifying Proteins

- Identification methods so far only identify peptide-spectrum matches (PSMs)
  - Search a database
  - Return a ranked list of PSMs with associated 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

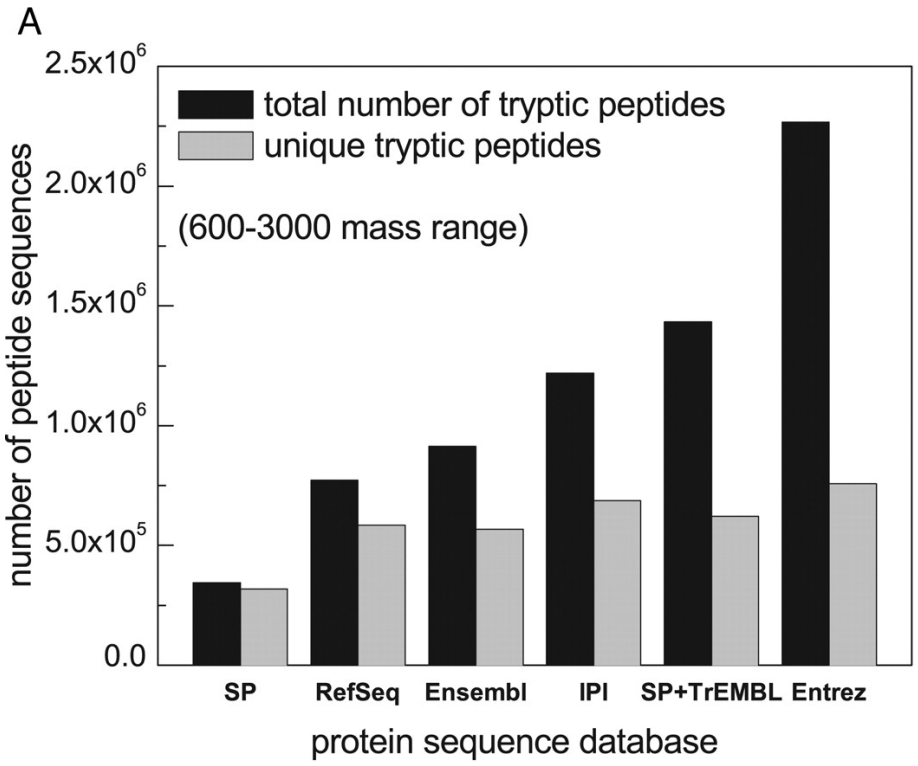


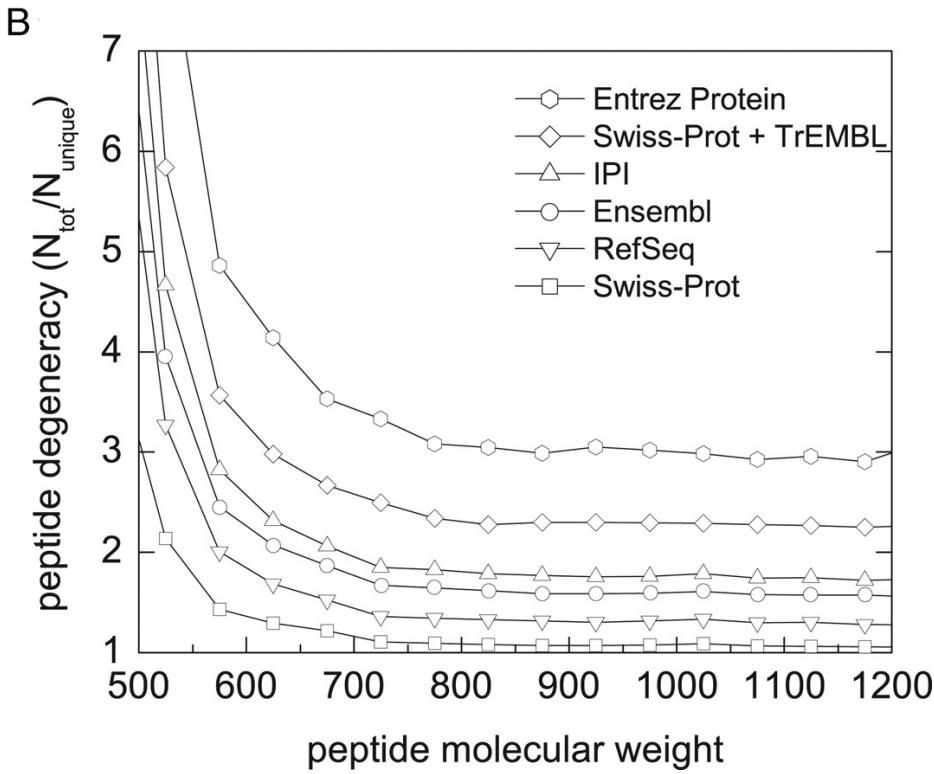
Danielle L. Swaney; Craig D. Wenger; Joshua J. Coon; *J. Proteome Res.* 2010, 9, 1323-1329.

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
- **Uniqueness**
  - **depends on the chosen database**
  - **becomes more likely for longer peptide sequences**
- Reasons for non-uniqueness
  - Chance hits
  - Different isoforms
  - Conserved regions shared within a protein family

Uniqueness

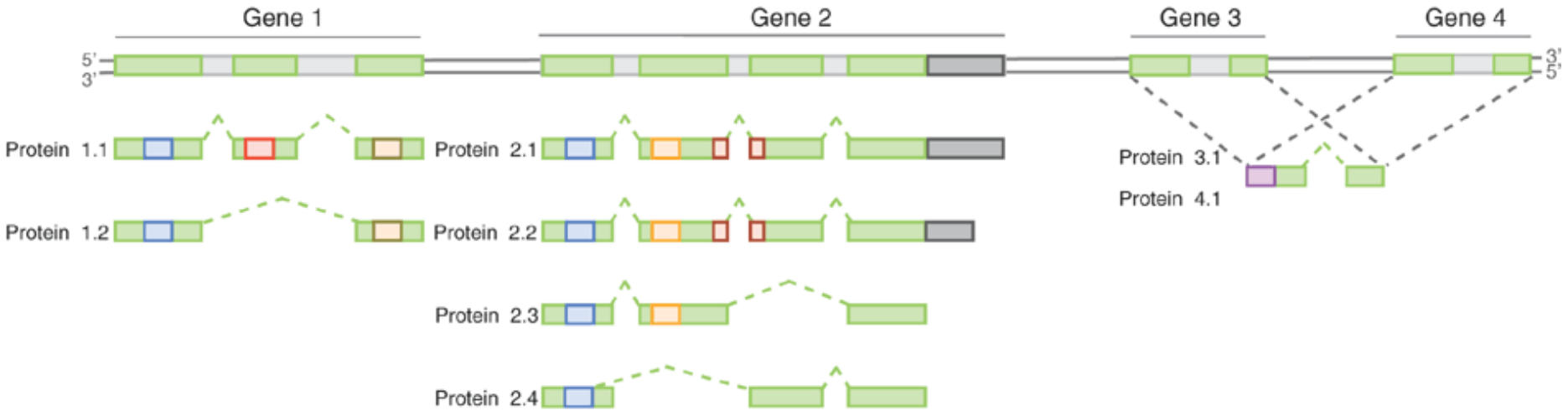




- Uniqueness depends on the size of the database
- Searching an appropriate (non-redundant) database is thus preferable
- Reference databases (SwissProt) usually contain few degenerate (non-unique) tryptic peptides above a mass of 750 Da
- **Problem: isoforms of proteins/splice variants!**

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

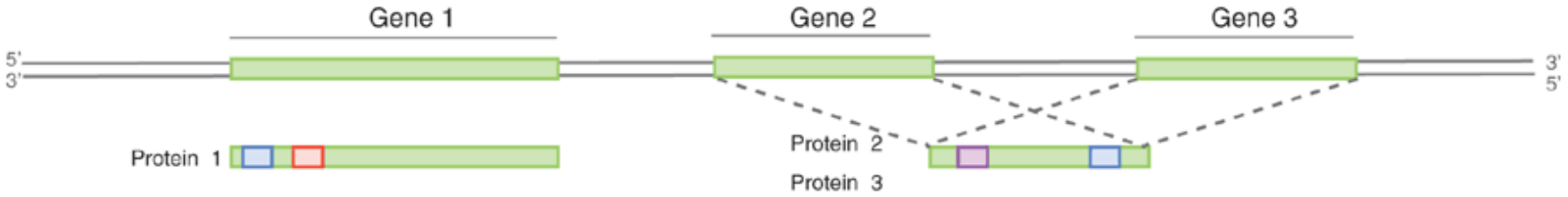
Uniqueness



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



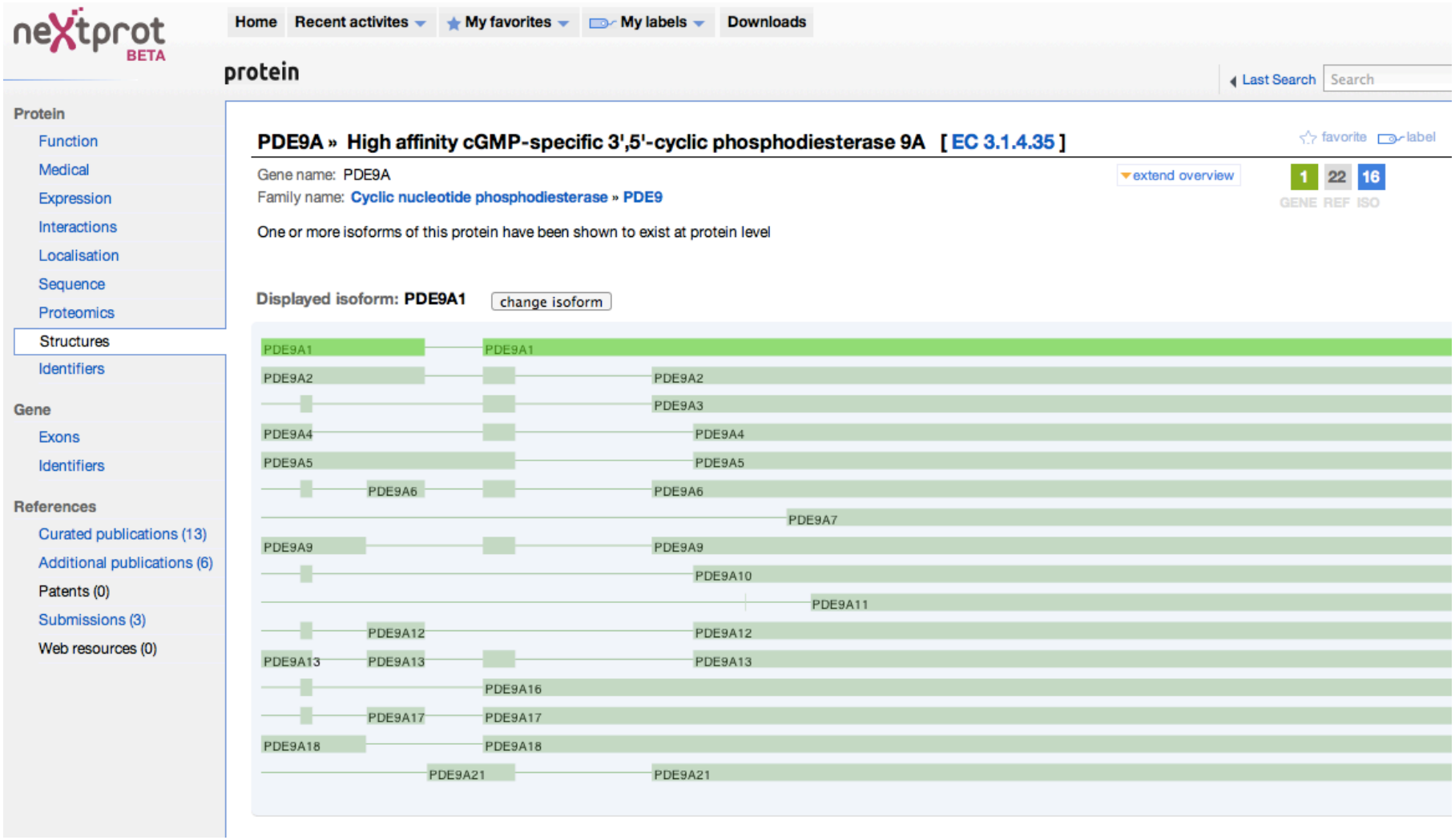
Qeli & Ahrens, Nature Biotechnology 28, 647–650 (2010)

Protein Isoforms

- 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

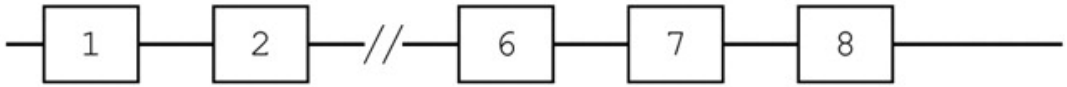
http://www.nextprot.org/db/entry/NX_O76083/structures

Protein Isoforms

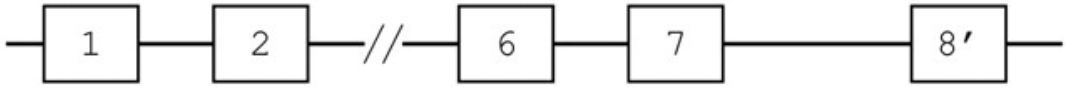
A

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

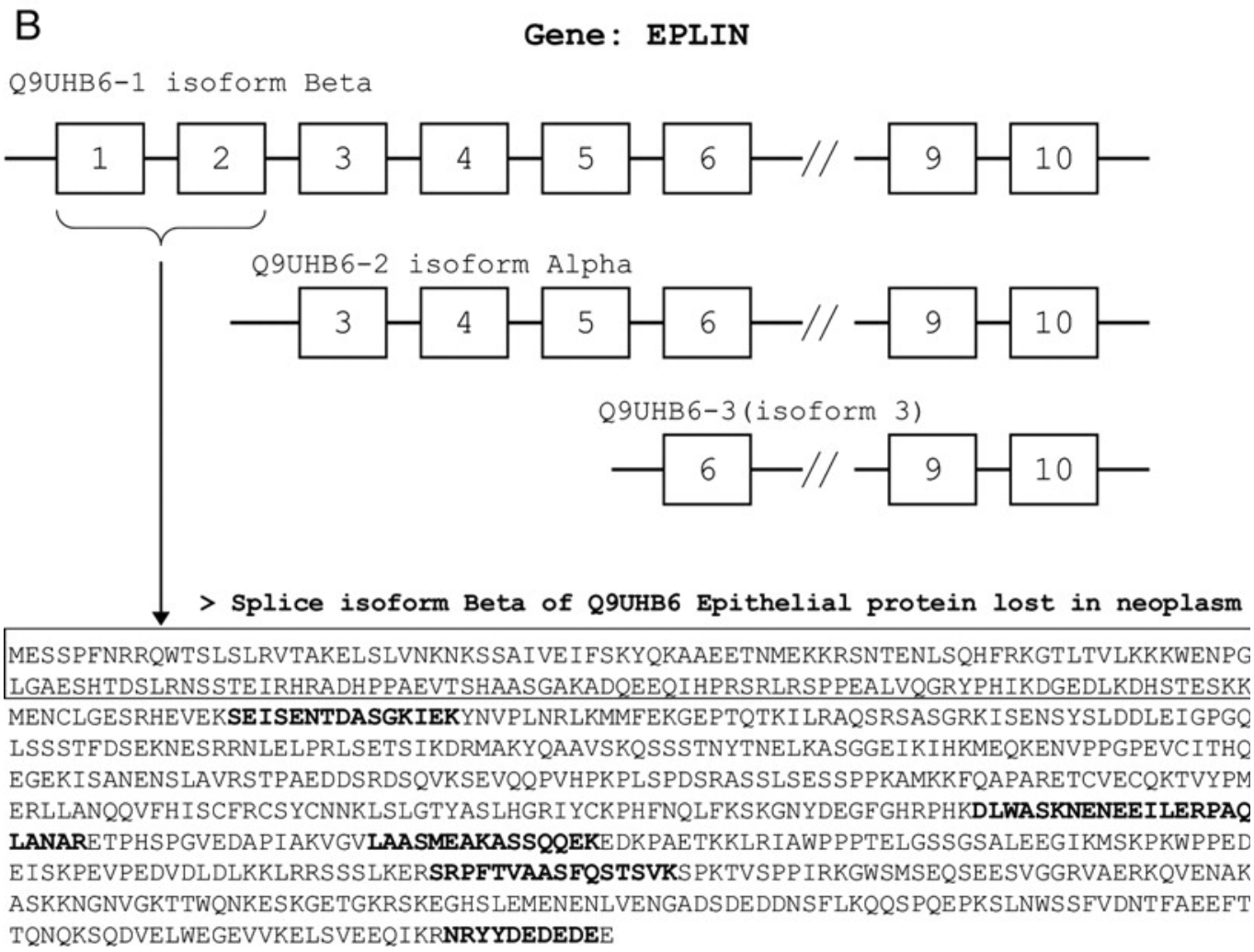


>IPI00218782 IPI:IPI00218782.1|Swiss-Prot:P47756-2|ENSEMBL:ENSP00000264203
Tax_Id=9606 Splice isoform 2 of F-actin capping protein beta subunit




```
P47756-1: MSDQQLDCALDLMRRLPPQQIEKNLSDLIDLVPSLCEDLLSSVDQPLKIARDKVVGKDYL 60
          MSDQQLDCALDLMRRLPPQQIEKNLSDLIDLVPSLCEDLLSSVDQPLKIARDKVVGKDYL
P47756-2: MSDQQLDCALDLMRRLPPQQIEKNLSDLIDLVPSLCEDLLSSVDQPLKIARDKVVGKDYL 60

P47756-1: LCDYNRDGDSYRSPWSNKYDPPLEDGAMPSARLRKLEVEANNAFDQYRDLYFEGGVSSVY 120
          LCDYNRDGDSYRSPWSNKYDPPLEDGAMPSARLRKLEVEANNAFDQYRDLYFEGGVSSVY
P47756-2: LCDYNRDGDSYRSPWSNKYDPPLEDGAMPSARLRKLEVEANNAFDQYRDLYFEGGVSSVY 120

P47756-1: LWDLDHGFAGVILIKKAGDGSKKIKGCWDSIHVVEVQEKSSGRTAHYKLTSTVMLWLQTN 180
          LWDLDHGFAGVILIKKAGDGSKKIKGCWDSIHVVEVQEKSSGRTAHYKLTSTVMLWLQTN
P47756-2: LWDLDHGFAGVILIKKAGDGSKKIKGCWDSIHVVEVQEKSSGRTAHYKLTSTVMLWLQTN 180

P47756-1: KSGSGTMNLGGSLTRQMEKDETVSDCSPHIANIGRLVEDMENKIRSTLNEIYFGKTKDIV 240
          KSGSGTMNLGGSLTRQMEKDETVSDCSPHIANIGRLVEDMENKIRSTLNEIYFGKTKDIV
P47756-2: KSGSGTMNLGGSLTRQMEKDETVSDCSPHIANIGRLVEDMENKIRSTLNEIYFGKTKDIV 240

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

Protein Isoforms



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

Protein Families

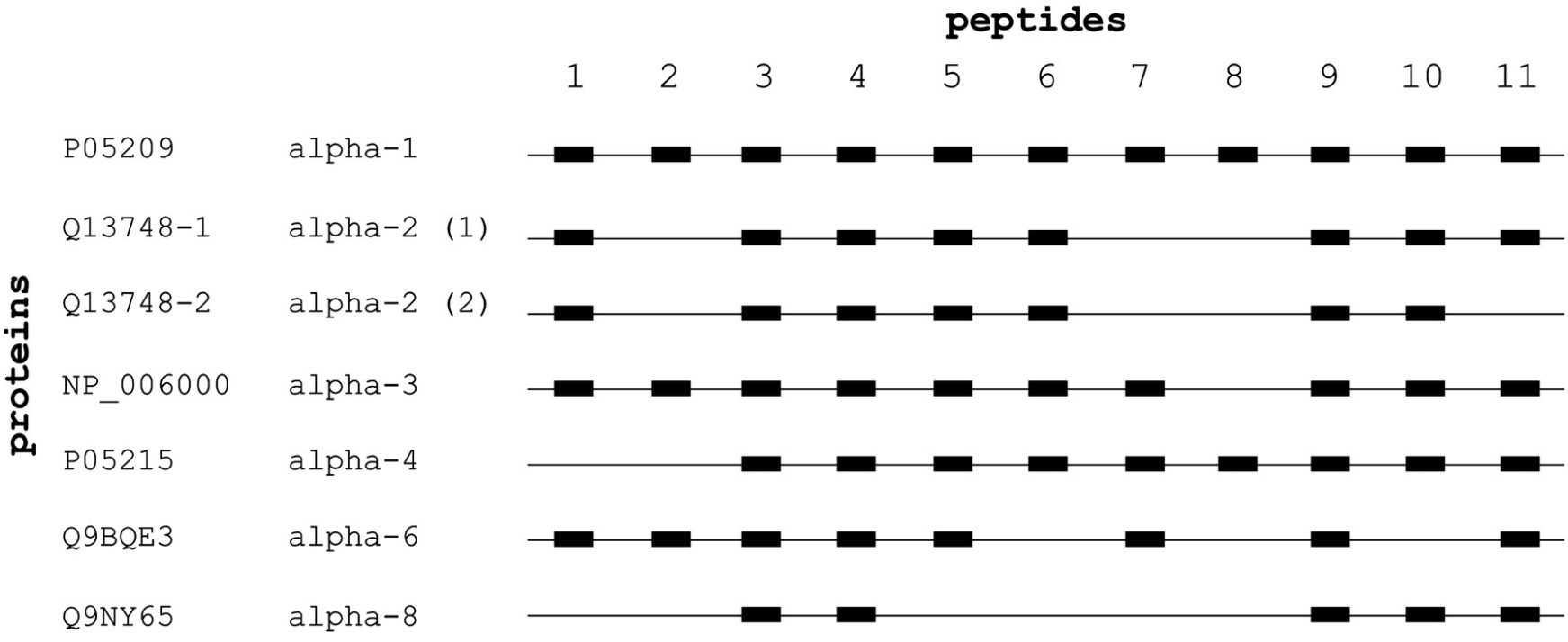
- Sequence coverage is often poor in large scale studies: many proteins are identified through very few peptides only
- In prokaryotes, typically over 90% of the identified peptides are unique in the whole proteome
- In particular in eukaryotes the large number of orthologs leads to significant sequence 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	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:



Parsimony-Based Inference

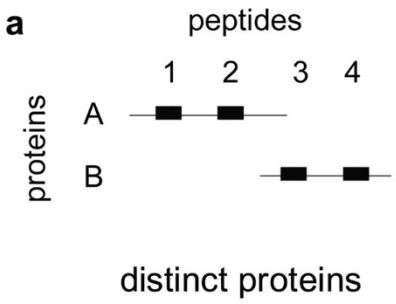
- **Idea**

  Find the smallest set of proteins explaining all observed peptides

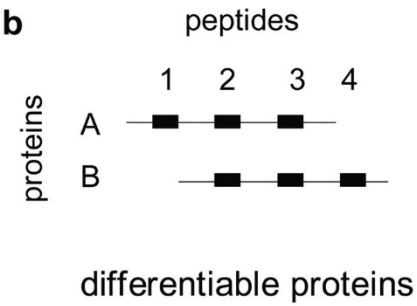
- If all peptides mapping to one protein family can be explained by a single protein, then it is quite likely, that only this protein is present (but this must not necessarily be the case)
- Basically: applying **Occam's razor** to the dataset – find the simplest explanation possible (**maximum parsimony**)

Parsimony-Based Inference

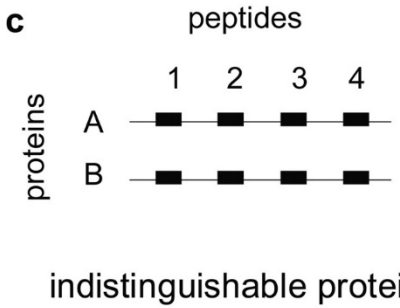
- Scenarios for different proteins given a set of observed peptides
  - **Distinct** proteins do not share peptides
  - **Differentiable** proteins can be distinguished by at least one distinct peptide
  - **Indistinguishable** proteins share all peptides
  - **Subset** proteins contain only peptides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



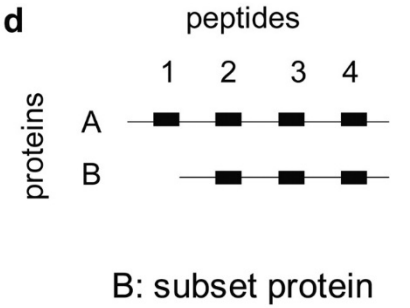
distinct proteins



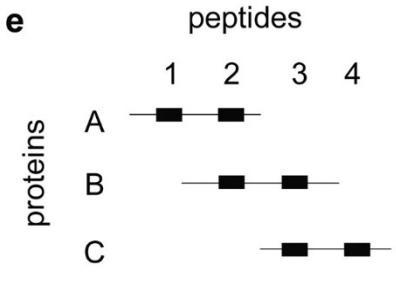
differentiable proteins



indistinguishable proteins



B: subset protein

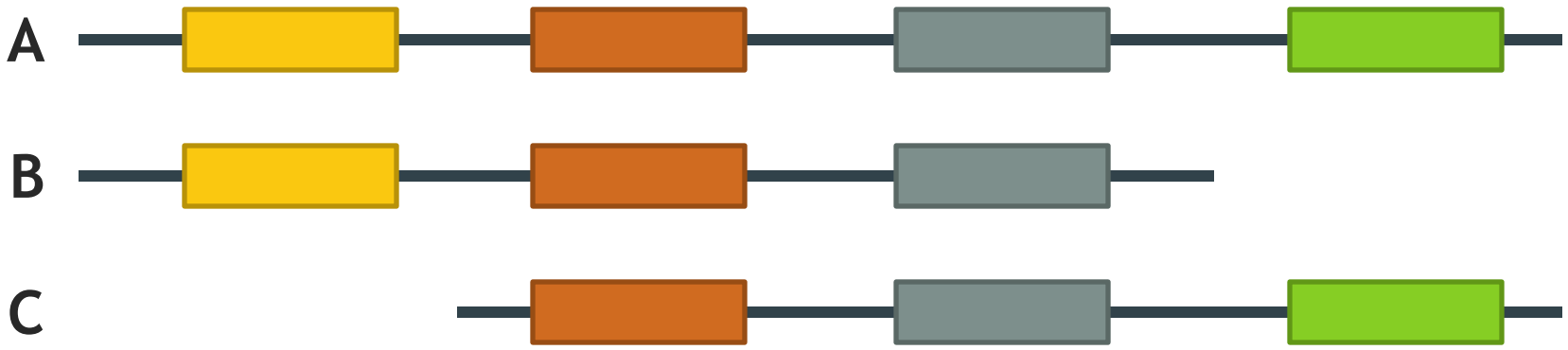


B: subsumable protein

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

Protein Ambiguity Groups

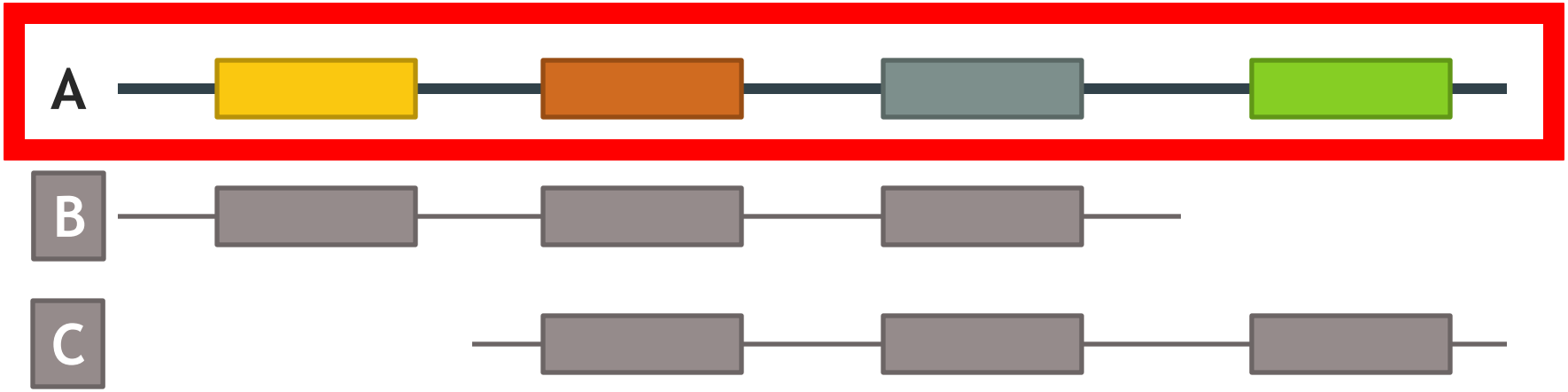
Example



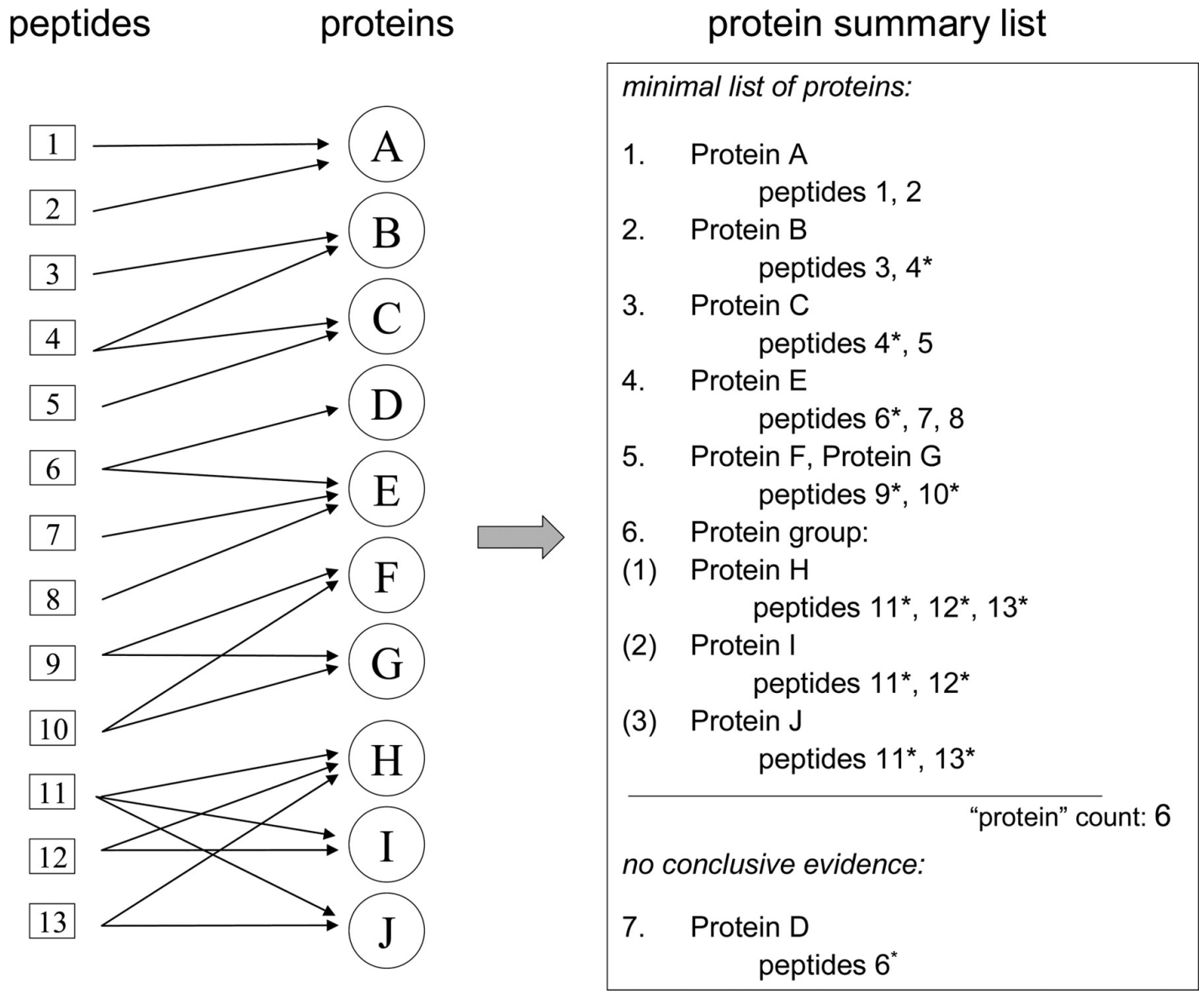
- Note that even though the presence of A is sufficient to explain all observed peptides, this does not automatically imply the absence of B and C
- The data is explained equally well by the presence of A, the presence of A + B, A + C, B + C, or A + B + C
- The set of proteins sharing one or multiple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



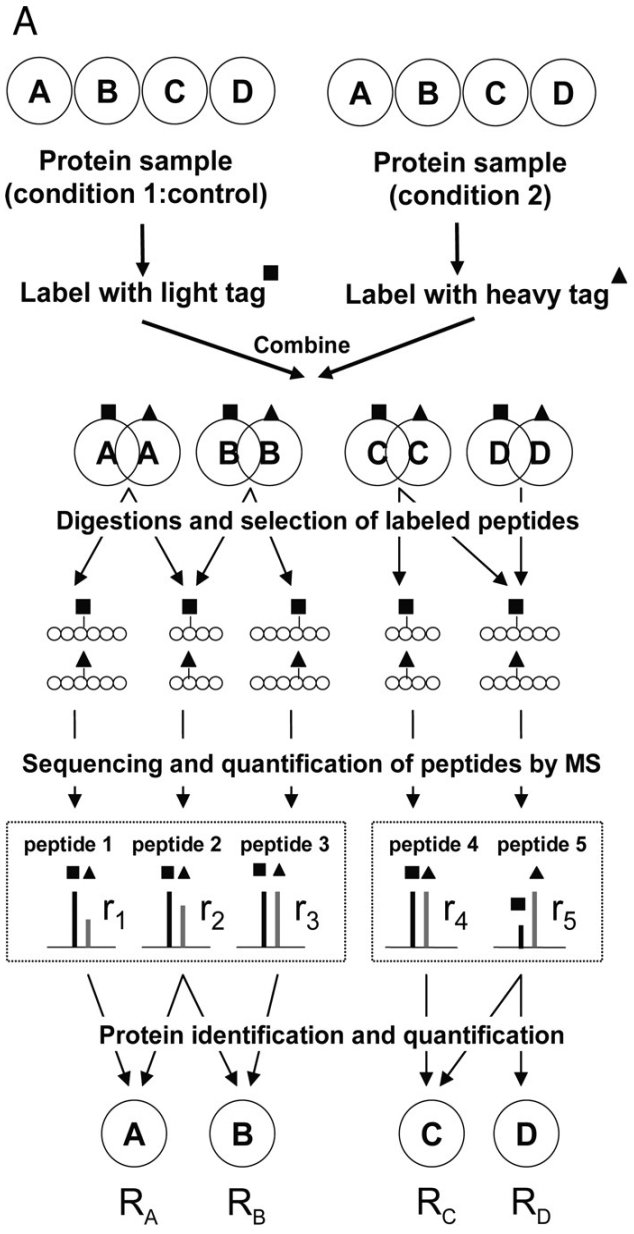
Reporting of PAGs

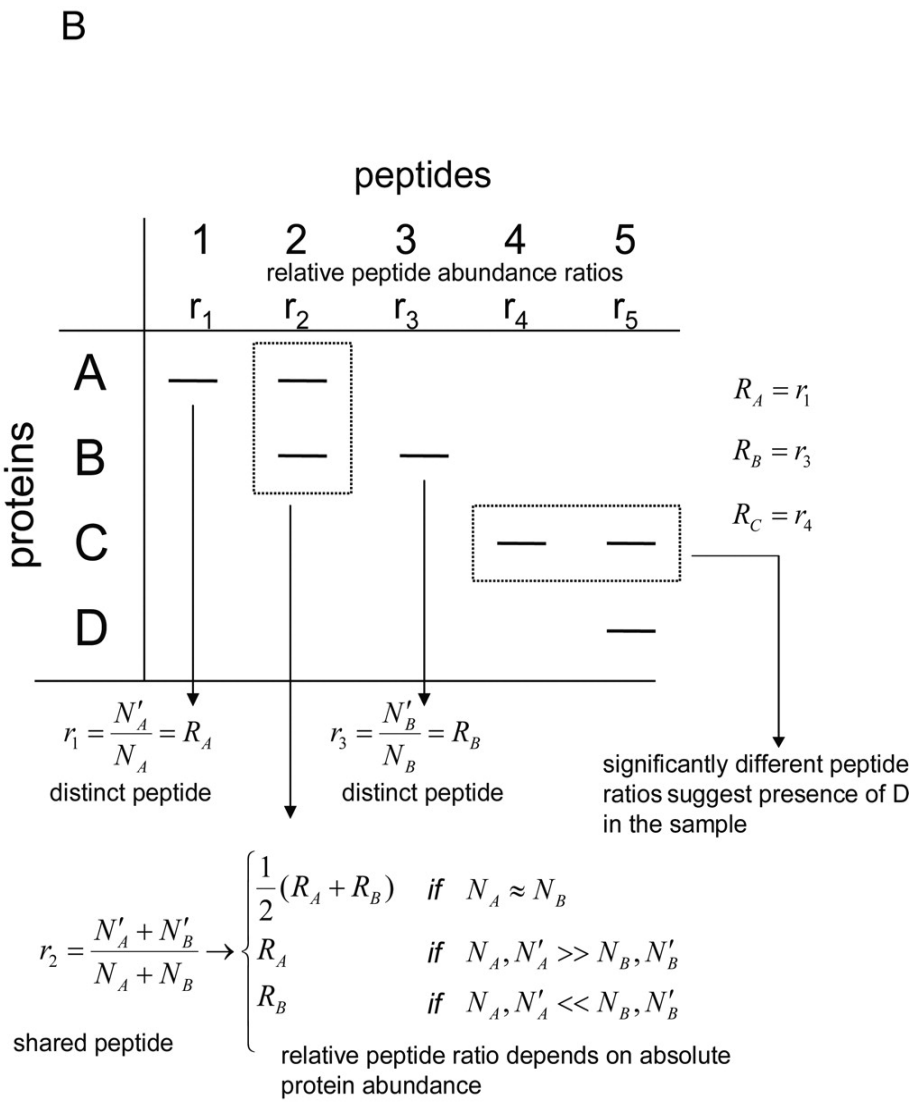


Inference Through Quantification

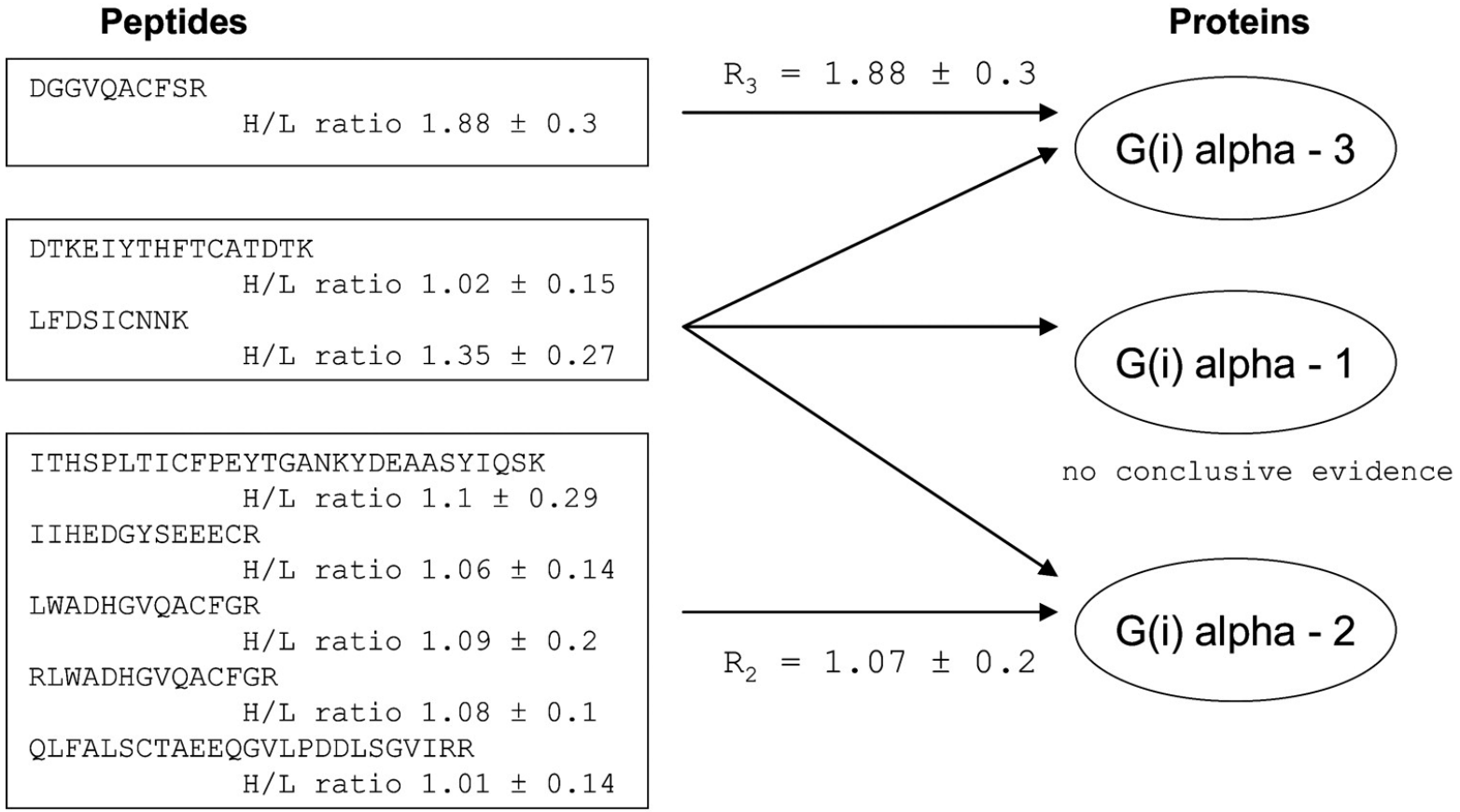
- Quantitative data can be used for inference as well (similar to transcript data)
- This is, however, non-trivial and usually done manually and on a case-by-case basis
- Distinct peptides can be used to quantify their source proteins
- Shared peptides result in an averaging of the quantitative information
- 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 peptide 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!)

Single Hit Wonders

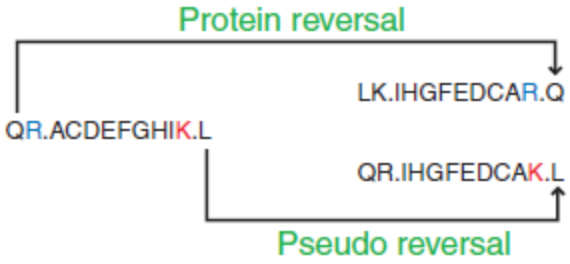
- In many cases, proteins are identified through a single PSM only
- These '**single hit wonders**' have long been considered problematic: a single false PSM can lead to a wrongly identified protein
- In fact, the so-called '**Paris guidelines**' for data deposition in proteomics recommend only reporting 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

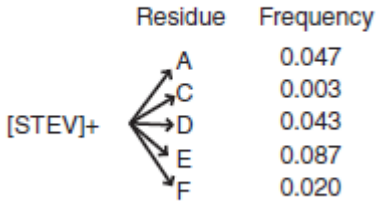
Design decoy sequences



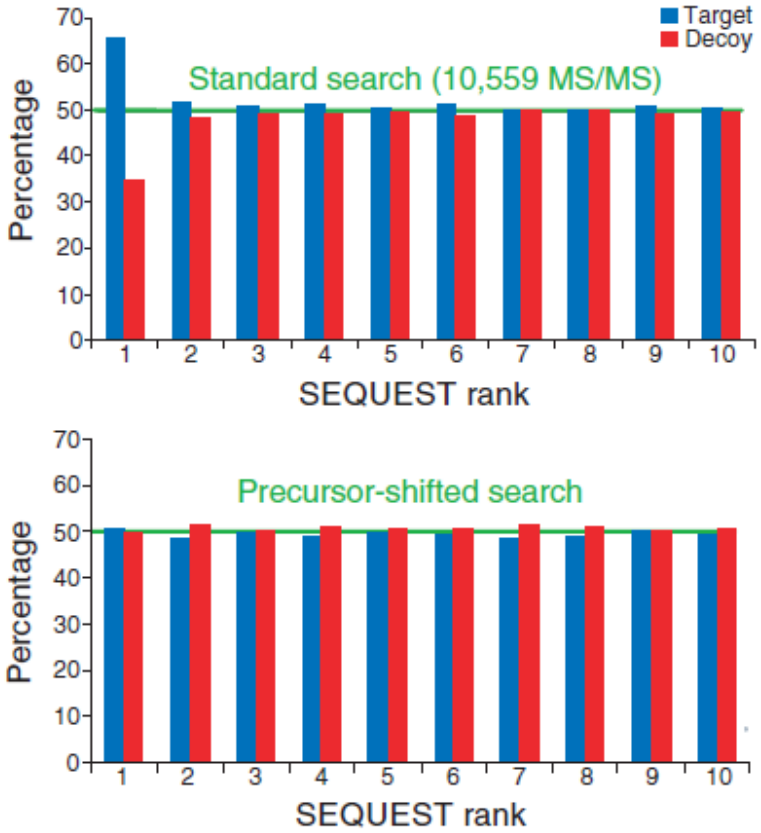
Random

Residue	Frequency
A	0.070
C	0.023
D	0.046
E	0.070
F	0.036

Markov



Separation of target and decoy



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)

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

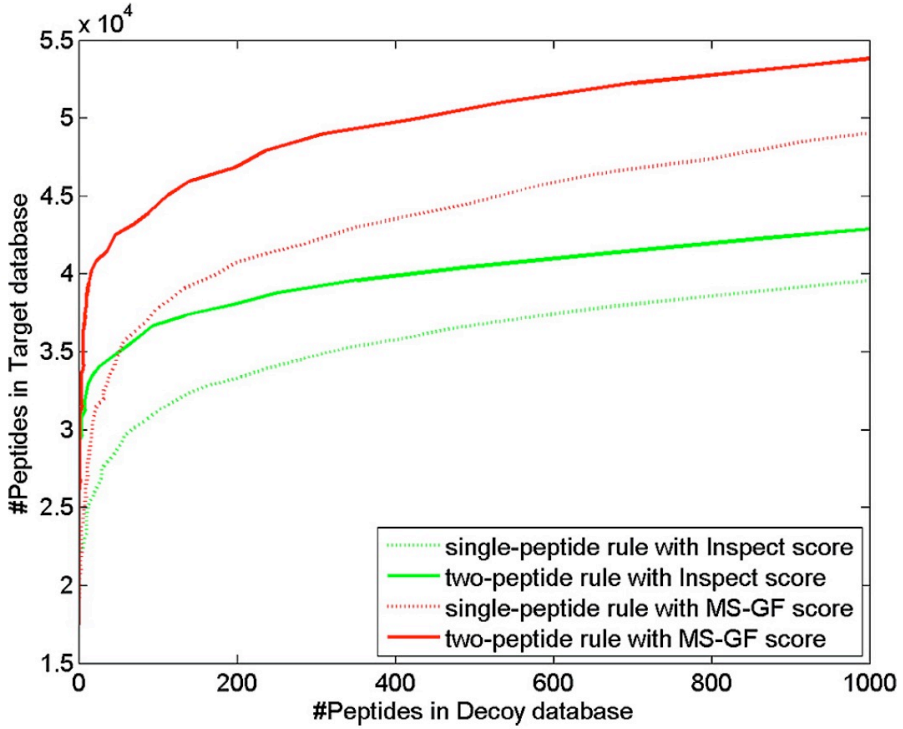
- Zhang et al. suggest *(Zhang et al.,* J Proteome Res 2007;6(9):3549-3557)

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

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

Single Hit Wonders

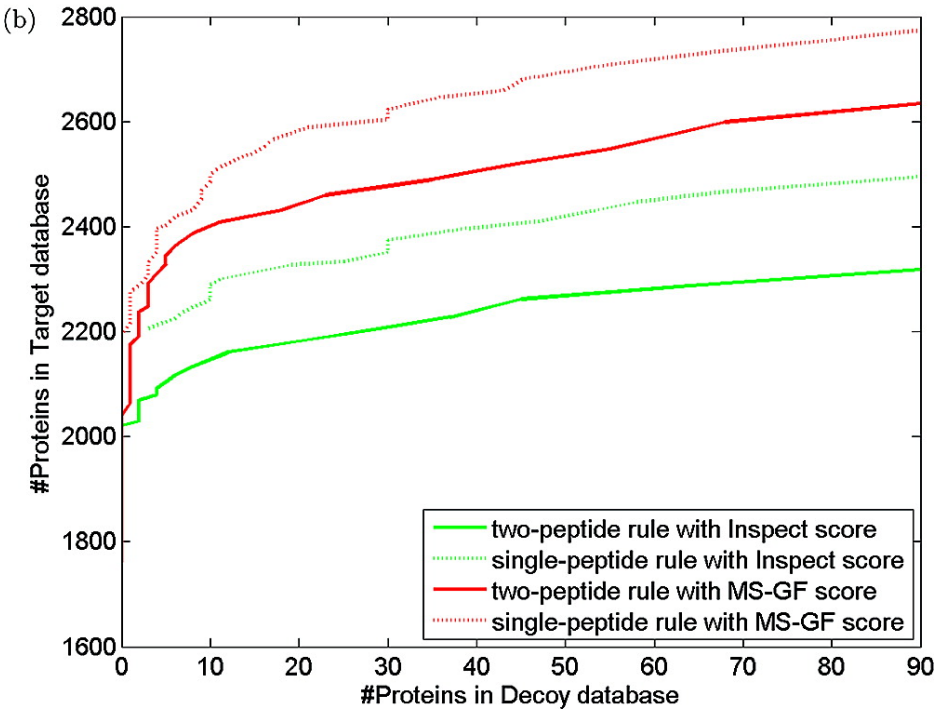
- Gupta & Pevzner argued in 2009 that the application of the two peptide rule actually results in increased false discovery rates for proteins
- Removing one-hit wonders should improve the FDR of peptide identifications – this is indeed the case
- For a given number of decoy hits, the number of target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



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

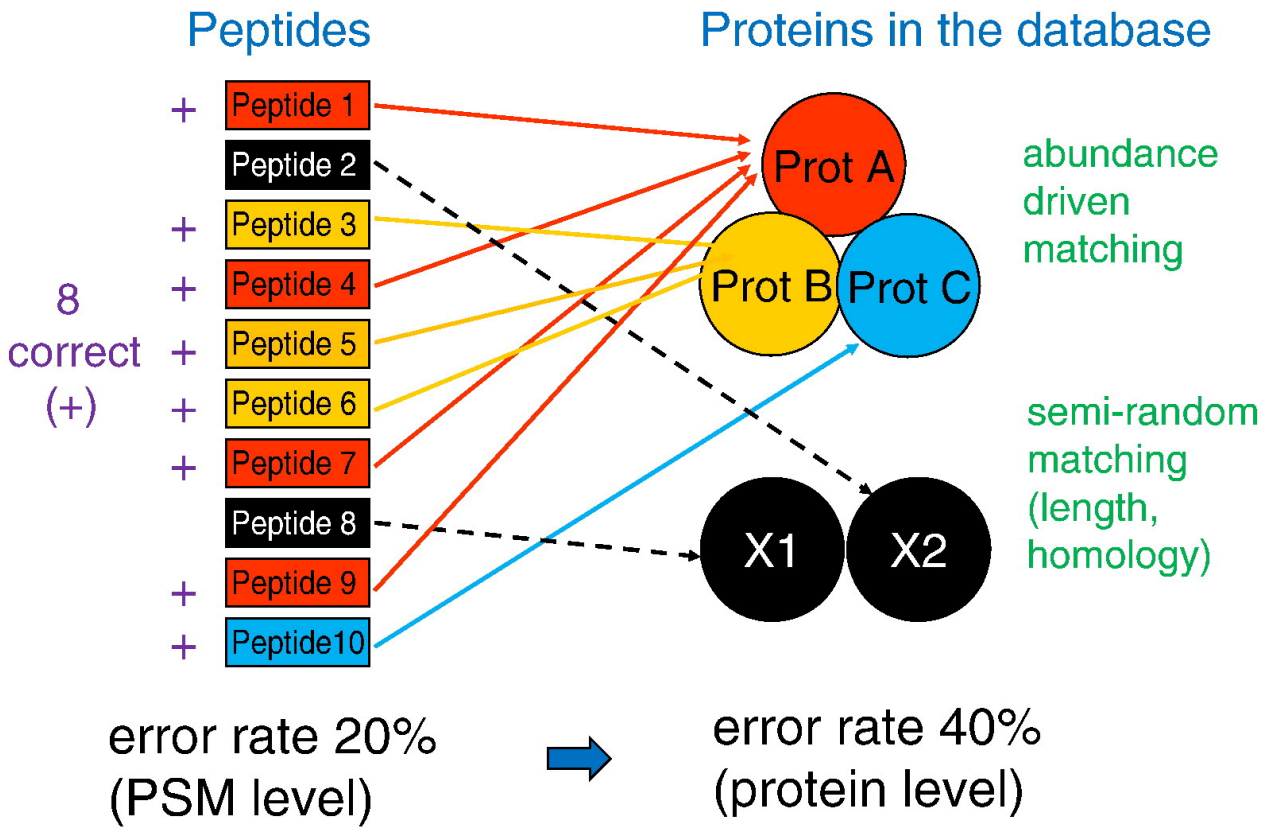
One Hit Wonders

- 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

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)