Bioinformatics 2

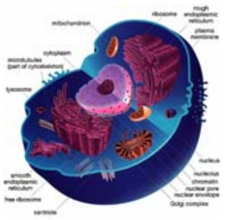
Lecture 2

Proteomics

Juri Rappsilber
Wellcome Trust Centre for Cell Biology, UoE
http://rappsilber.bio.ed.ac.uk/

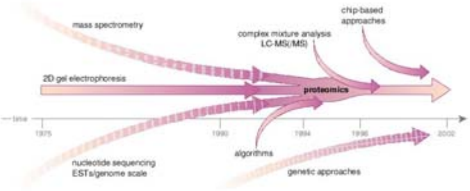
Key questions of proteomics

- **What proteins are there**?
- **How much** is there of each of the proteins?
  - Absolute quantitation
  - Stoichiometry
- What (modification/splice) **state** are the proteins in?
- Which proteins **interact** with each other or with other molecules (DNA, RNA)?
- How does all of the above **change** with time/stimulation/mutation of a key protein/… ?



Foundation of proteomics

- **M**ass spectrometry
- **A**lgorithms
- **D**NA sequencing

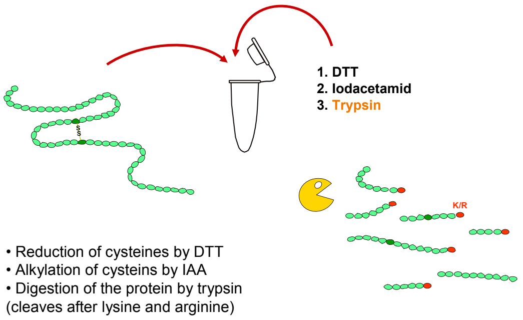


What proteins are there?

Protein identification is achieved by

- Proteolysis of the proteins into peptides
- Mass spectrometric detection of the peptides
(shortcut to protein identification: peptide mass fingerprinting)
- Mass spectrometric fragmentation of the peptides
- Database search to identify the peptides

Protein digestion



1. DTT
2. Iodacetamid
3. Trypsin

- Reduction of cysteines by DTT
- Alkylation of cysteins by IAA
- Digestion of the protein by trypsin
(cleaves after lysine and arginine)

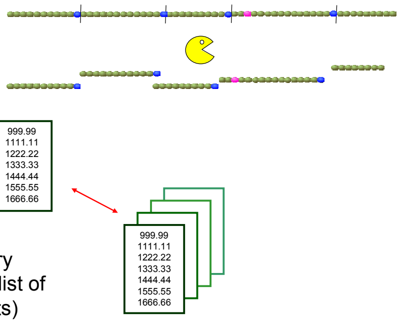
Peptide mass fingerprinting

Isolated protein

Digestion

MALDI MS

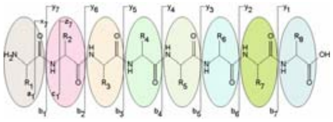
Database Query
(compare with list of
'in-silico' digests)



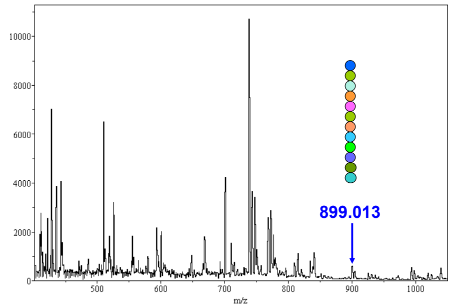
Peptide Fragmentation

(Low-Energy Collision induced fragmentation)

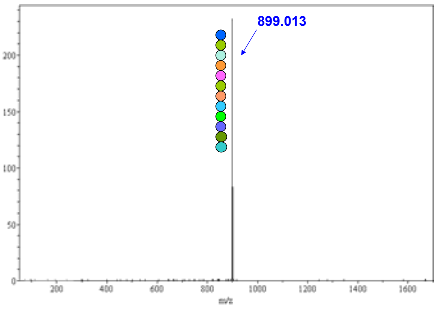
- Peptides fragment preferentially between amino acids
- The chemical bond that cleaves depends on the fragmentation method.
- Low-Energy Collision Induced Dissociation (CID) is most common. Leads to b and y ions
- Electron Transfer Dissociation (ETD) is up and coming. Leads to c and z ions.



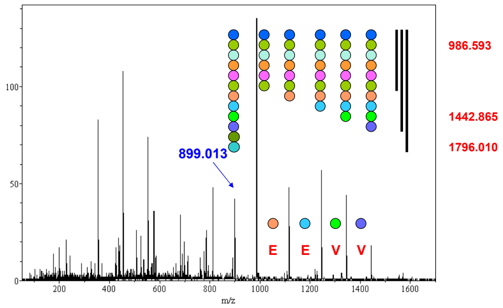
MS of a Peptide Mixture



MS/MS of a Peptide (low collision energy)

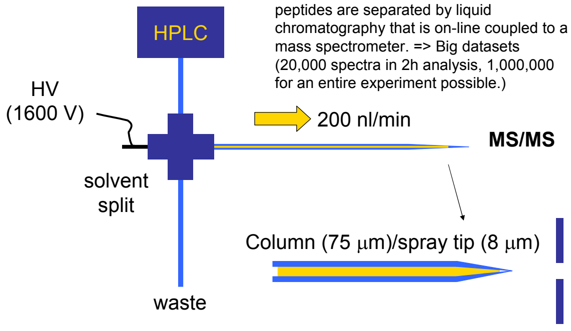


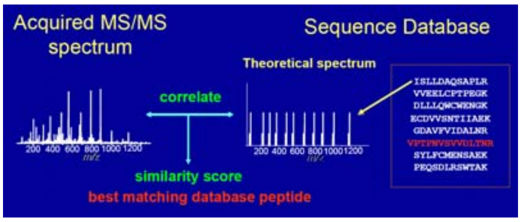
MS/MS of a Peptide (high collision energy)



LC-MS interface

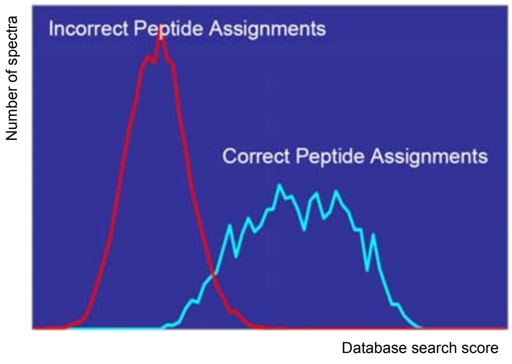
For the analysis of complex mixtures peptides are separated by liquid chromatography that is on-line coupled to a mass spectrometer. => Big datasets (20,000 spectra in 2h analysis, 1,000,000 for an entire experiment possible.)



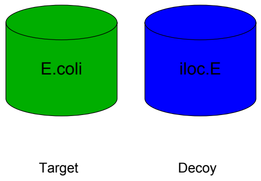


- Many programs available for this matching of fragmentation spectra with peptide sequences from databases (Mascot, Sequest, OMSSA, XTandem!)
- Each program has its own score.
- None of the scores is truly statistical.
- Results for the same dataset vary (overlap between any two ca. 50-60%).

How to find the rate of incorrect assignments => confidence?



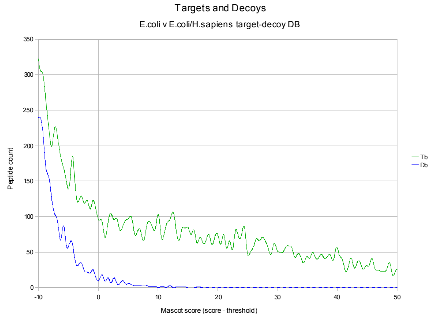
Decoy Database



Targets and decoys v score

🟩 *Target*

🟦 Decoy



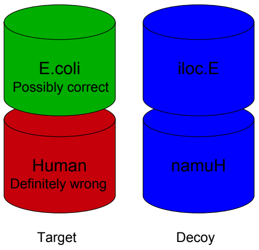
FPR calculation methods

$$\text{False positive rate} = \frac{\text{Decoy count}}{\text{Target count}}$$

Locally (within a window around a given score)
or
cumulative (everything above a given score)

Test the impact of FPR calculation

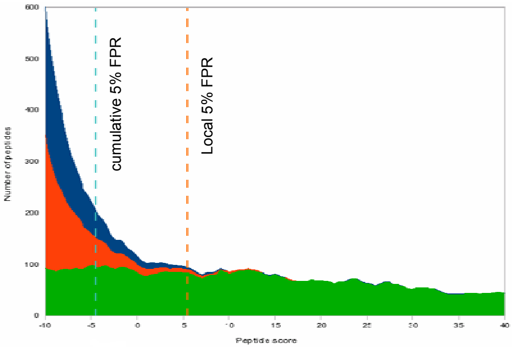
🟩 *E.coli*
🟥 Human
🟦 Decoy



The addition of the human sequences allows us to check if our decoy based approach correctly models our incorrectly identified target peptides.

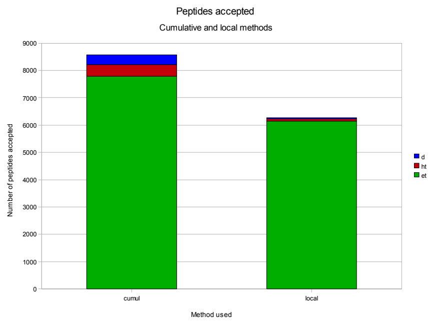
Two methods for counting the false positives

🟩 *E.coli*
🟥 Human
🟦 Decoy



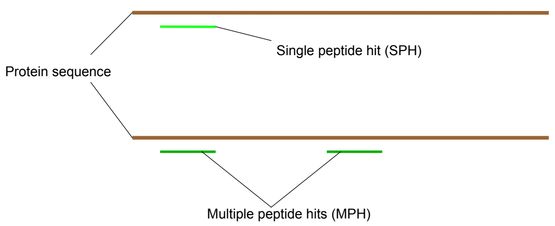
Peptide counts for cumulative and local methods

🟩 *E.coli*
🟥 Human
🟦 Decoy



More peptides identified by cumulative method, but also more false peptides included.

Single and multiple peptide hits

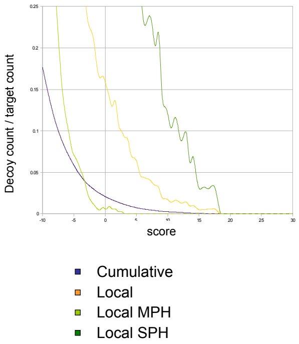


Significance of SPH and MPH

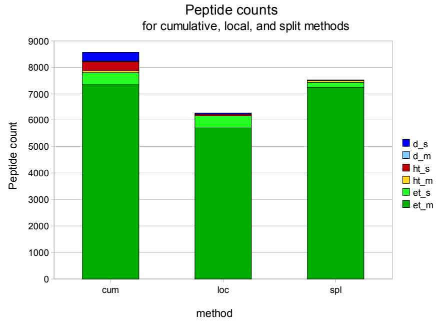
- MPHs confer additional corroboration to each-other
- SPH are often disregarded in practice
- What if we treat MPH and SPH separately?

Improved confidence by SPH/MPH

- MPH and SPH show very different curves.
- Local MPH cutoff is similar to cumulative cutoff.
- Local SPH cutoff is much higher than cumulative cutoff => cumulative method overestimates confidence in SPH leading to high false discovery rates for SPH proteins and their rejection.

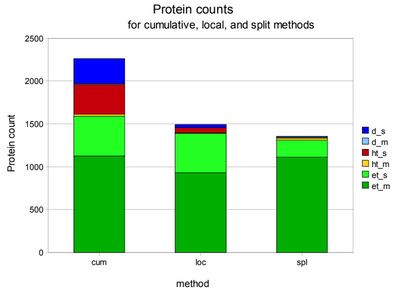


Final comparison: peptide counts



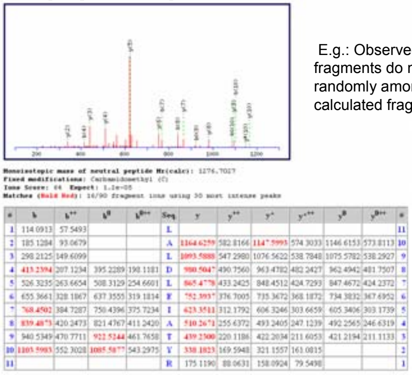
Split method optimizes the number of correct peptides while minimizing the number of incorrect identifications.

Final comparison: protein counts



The impact of the FPR method is MUCH more severe on protein level as most peptides match to proteins together with other peptides. Few peptides match to a protein alone. However, essentially all incorrectly identified peptides match alone to a protein. The protein list grows hence by a protein per false peptide. Therfore, with current approach (cum) proteins cannot be identified reliably with a single peptide. This is possible using local split method (spl).

Non-statistical component of peptide-spectra matches



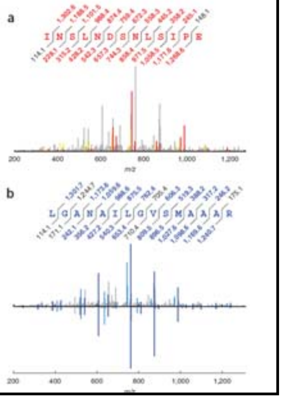
Monoisotopic mass of neutral peptide Mr(calc): 1276.7027
Fixed modifications: Carbamidomethyl (C)
Ions Score: 64 Expect: 1.1e-05
Matches (Bold Red): 16/90 fragment ions using 30 most intense peaks

#	b	b++	b°	b°++	Seq.	y	y++	y*	y*++	y°	y°++	#
1	114.0913	57.5493			L							11
2	185.1284	93.0679			A	**1164.6259**	582.8166	**1147.5993**	574.3033	1146.6153	573.8113	10
3	298.2125	149.6099			L	**1093.5888**	547.2980	1076.5622	538.7848	1075.5782	538.2927	9
4	**413.2394**	207.1234	395.2289	198.1181	D	**980.5047**	490.7560	963.4782	482.2427	962.4942	481.7507	8
5	526.3235	263.6654	508.3129	254.6601	L	**865.4778**	433.2425	848.4512	424.7293	847.4672	424.2372	7
6	655.3661	328.1867	637.3555	319.1814	E	**752.3937**	376.7005	735.3672	368.1872	734.3832	367.6952	6
7	**768.4502**	384.7287	750.4396	375.7234	I	**623.3511**	312.1792	606.3246	303.6659	605.3406	303.1739	5
8	**839.4873**	420.2473	821.4767	411.2420	A	**510.2671**	255.6372	493.2405	247.1239	492.2565	246.6319	4
9	940.5349	470.7711	**922.5244**	461.7658	T	**439.2300**	220.1186	422.2034	211.6053	421.2194	211.1133	3
10	**1103.5983**	552.3028	**1085.5877**	543.2975	Y	**338.1823**	169.5948	321.1557	161.0815			2
11					R	175.1190	88.0631	158.0924	79.5498			1

E.g.: Observed fragments do not scatter randomly among the calculated fragments.

SVM approach

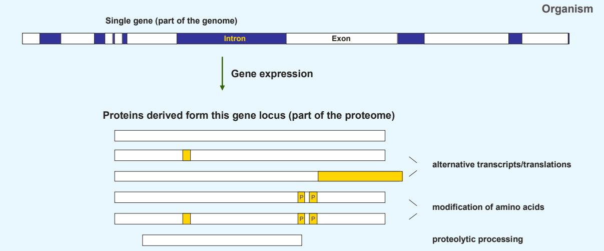
- Collect long list of features characterizing the peptide-spectrum match (this includes the score but also other parameters)
- Use decoy matches as false positives
- Train the SVM with each dataset new
- Gives significant improvement (20-400%) over search program alone or alternative procedures.



Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ.
Semi-supervised learning for peptide identification from shotgun proteomics datasets.
Nat Methods. 2007 Nov;4(11):923-5. Epub 2007 Oct 21.

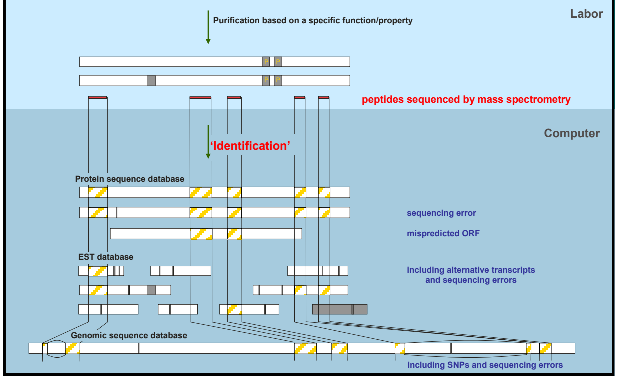
What does the peptide based analysis mean for identifying proteins?

Sequence space in the cell


Organism
Single gene (part of the genome)
Intron
Exon
Gene expression
Proteins derived form this gene locus (part of the proteome)
alternative transcripts/translations
P
P
modification of amino acids
proteolytic processing

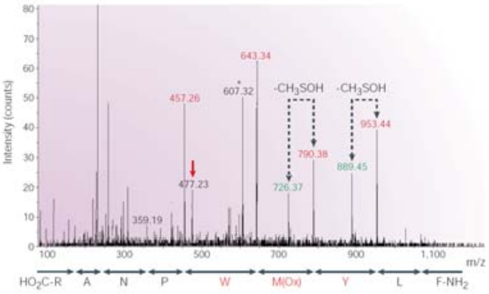
What does it mean to identify a protein in proteomics?
Rappsilber J, Mann M.
Trends Biochem Sci. 2002 Feb;27(2):74-8.

Scientific approach


Labor
Purification based on a specific function/property
peptides sequenced by mass spectrometry
Computer
'Identification'
Protein sequence database
sequencing error
mispredicted ORF
EST database
including alternative transcripts and sequencing errors
Genomic sequence database
including SNPs and sequencing errors

Modified peptides

- Include modification as possibility in the database search
- For informatics the same problem as peptide identification

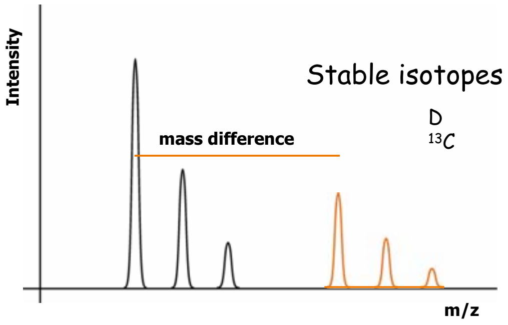


Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol. 2004 Sep;5(9):699-711. Review.

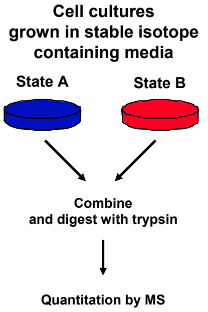
Quantitation in MS

- Absolute quantitation possible by using a labelled peptide as reference standard.
- Differential analysis possible by labelling on sample and not labelling the other. Both can then be mixed and analyzed together.

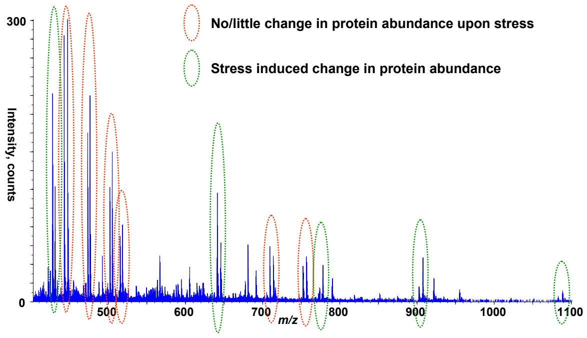
Quantitation in MS



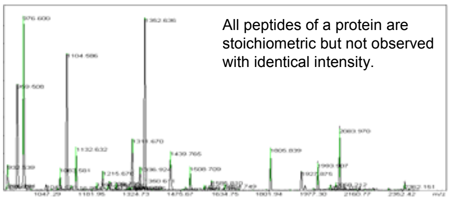
In vivo labeling with *SILAC*



Analysis of proteins from stressed cells



Stoichiometry



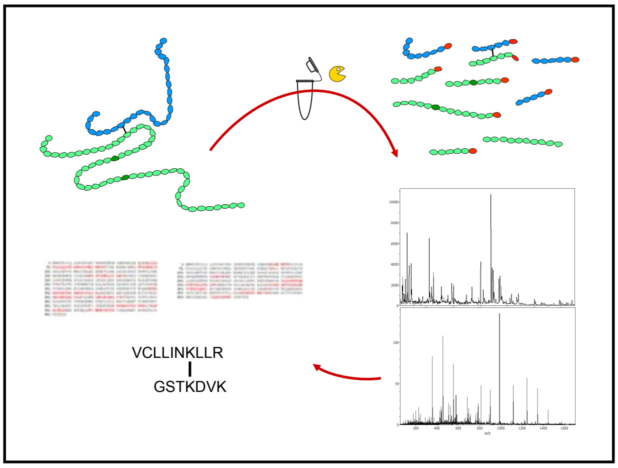
All peptides of a protein are stoichiometric but not observed with identical intensity.

Intensity in mass spectrum not direct consequence of abundance but influenced by many molecule-specific factors
=> Apple-orange problem

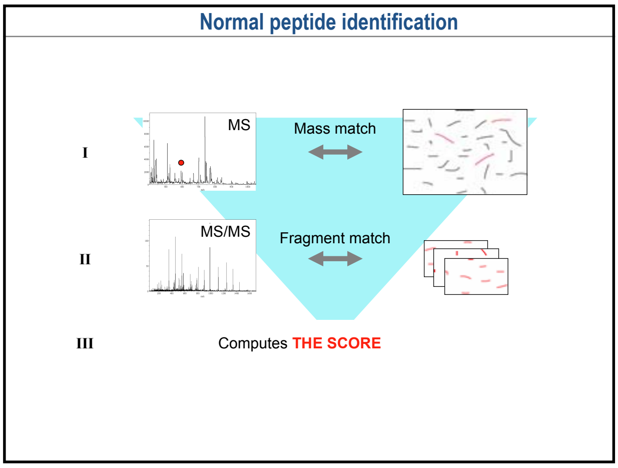
Approximation possible by summing up the mass spectrometric evidence gathered for a protein and normalizing this by the expected volume of evidence
Example: number of observed peptides / number of observable peptides

Protein-protein interactions

- Can be analyzed using same tools as for protein identification (mass spectrometry and database searching).
- Need to cross-link proteins to maintain their proximity also after proteolysis.



VCLLINKLLR
|
GSTKDVK

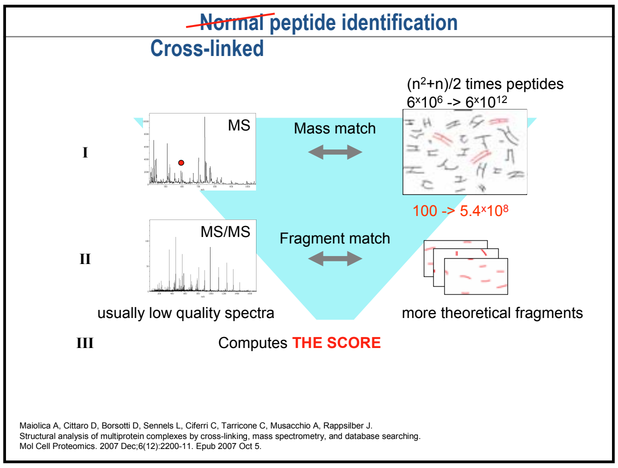


Normal peptide identification

I — MS — Mass match

II — MS/MS — Fragment match

III — Computes **THE SCORE**



~~Normal~~ peptide identification

Cross-linked

(n²+n)/2 times peptides
$6{\times}10^{6}$ -> $6{\times}10^{12}$

I — MS — Mass match

100 -> $5.4{\times}10^{8}$

II — MS/MS — Fragment match

usually low quality spectra

more theoretical fragments

III — Computes **THE SCORE**

Maiolica A, Cittaro D, Borsotti D, Sennels L, Ciferri C, Tarricone C, Musacchio A, Rappsilber J.
Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching.
Mol Cell Proteomics. 2007 Dec;6(12):2200-11. Epub 2007 Oct 5.