Bioinformatics 2

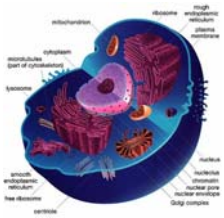
Lecture X

Proteomics

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

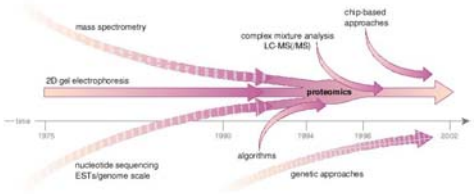
Key questions of proteomics

- **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

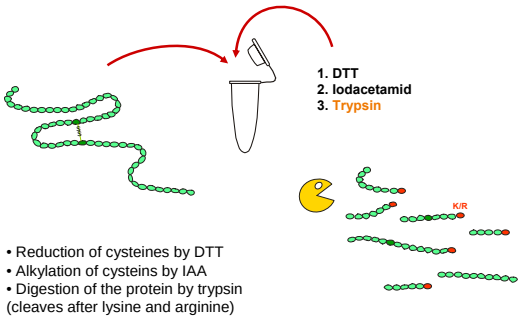


What proteins are there?

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

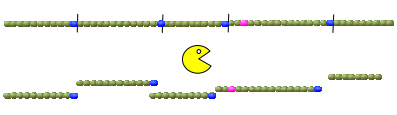
Protein digestion



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

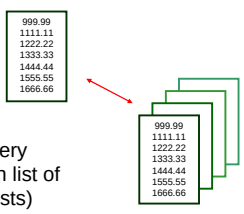
Peptide mass fingerprinting

Isolated protein



Digestion

MALDI MS

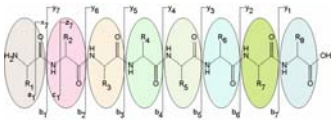


Database Query
(compare with list of
'in-silico' digests)

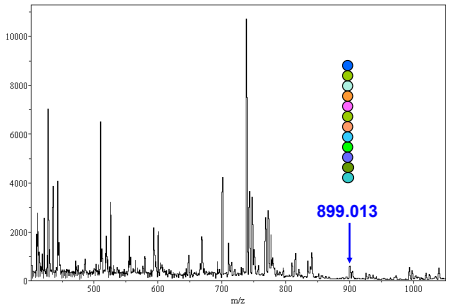
Peptide Fragmentation

(Low-Energy Collision induced fragmentation)

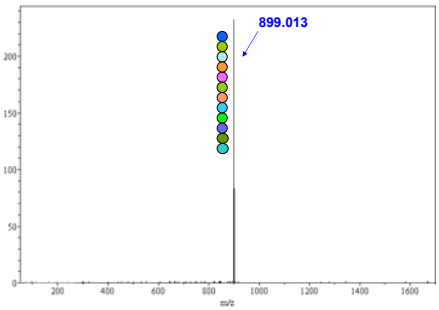
- Peptides fragment preferentially between amino acids
- The chemical bond that cleaves depends on the fragmentation method.
- Low-Energy Collision Induced Dissociation (CID) is most common. Leads to b and y ions
- Electron Transfer Dissociation (ETD) is up and coming. Leads to c and z ions.



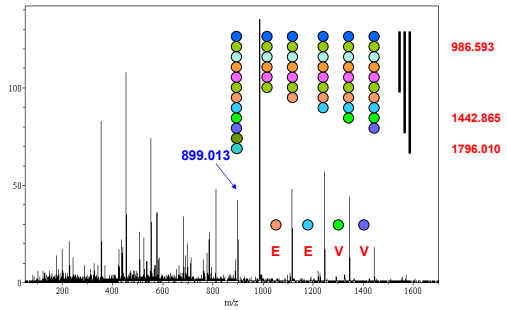
MS of a Peptide Mixture



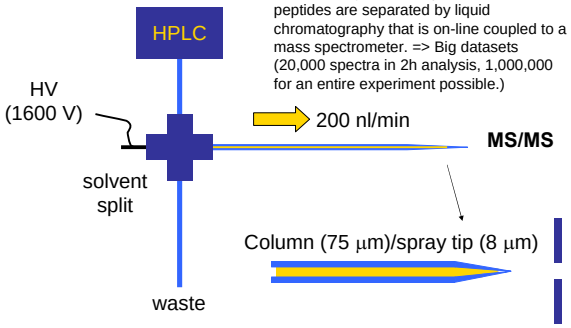
MS/MS of a Peptide (low collision energy)



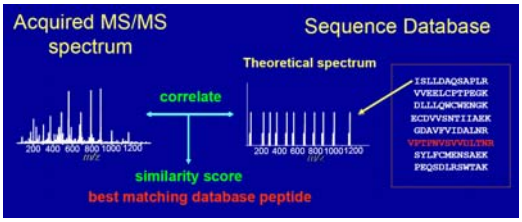
MS/MS of a Peptide (high collision energy)



LC-MS interface

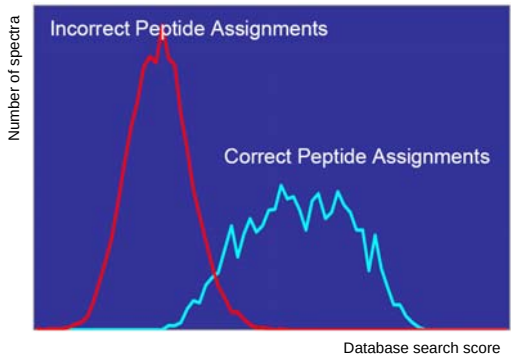


For the analysis of complex mixtures peptides are separated by liquid chromatography that is on-line coupled to a mass spectrometer. => Big datasets (20,000 spectra in 2h analysis, 1,000,000 for an entire experiment possible.)

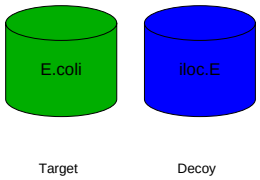


- Many programs available for this matching of fragmentation spectra with peptide sequences from databases (Mascot, Sequest, OMSSA, XTandem!)
- Each program has its own score.
- None of the scores is truly statistical.
- Results for the same dataset vary (overlap between any two ca. 50-60%).

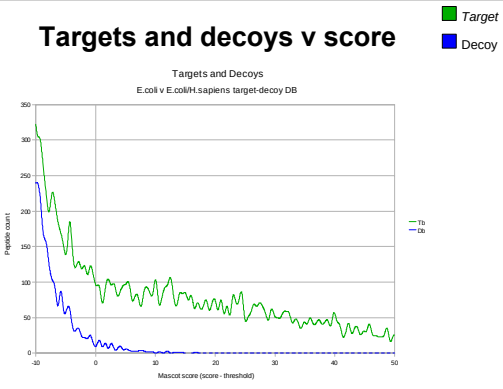
How to find the rate of incorrect assignments => confidence?



Decoy Database



Targets and decoys v score

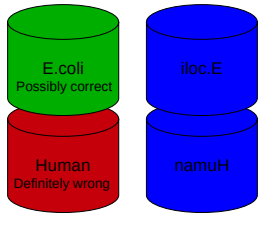


FPR calculation methods

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

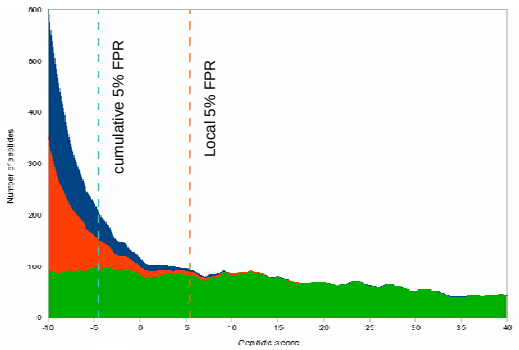
Locally (within a window around a given score)
or
cumulative (everything above a given score)

Test the impact of FPR calculation

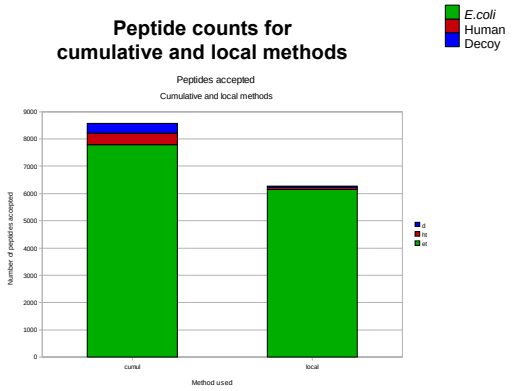


The addition of the human sequences allows us to check if our decoy based approach correctly models our incorrectly identified target peptides.

Two methods for counting the false positives

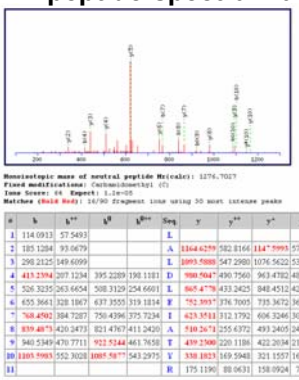


Peptide counts for cumulative and local methods



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

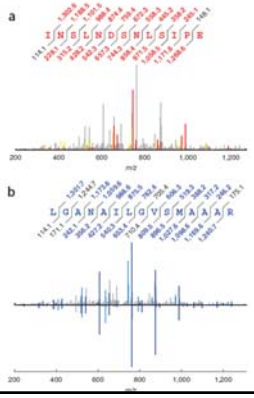
Non-statistical component of peptide-spectra matches



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

SVM approach

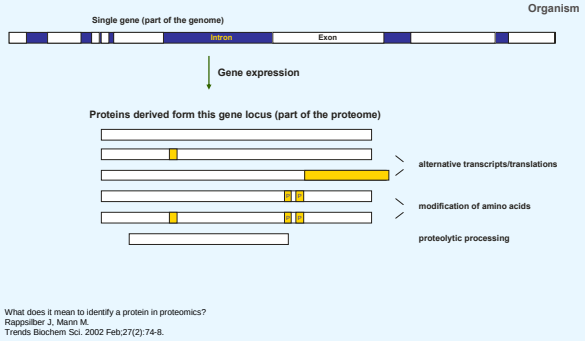
- Collect long list of features characterizing the peptide-spectrum match (this includes the score but also other parameters)
- Use decoy matches as false positives
- Train the SVM with each dataset new
- Gives significant improvement (20-400%) over search program alone or alternative procedures.



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.

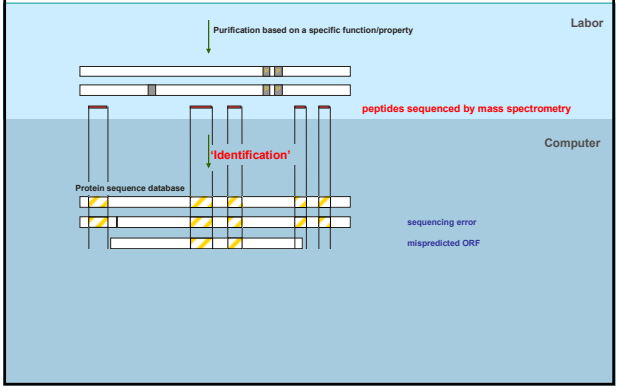
What does the peptide based analysis mean for identifying proteins?

Sequence space in the cell

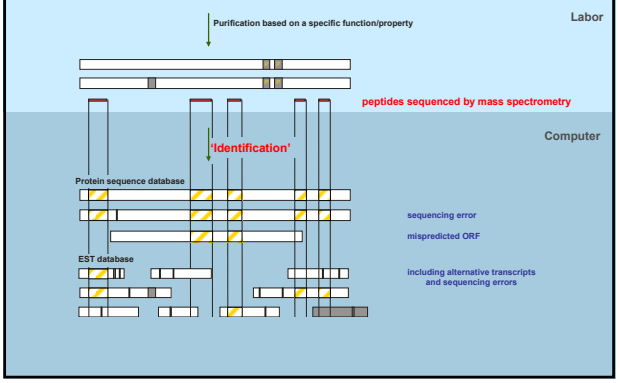


What does it mean to identify a protein in proteomics?
Rappsilber J, Mann M.
Trends Biochem Sci. 2002 Feb;27(2):74-8.

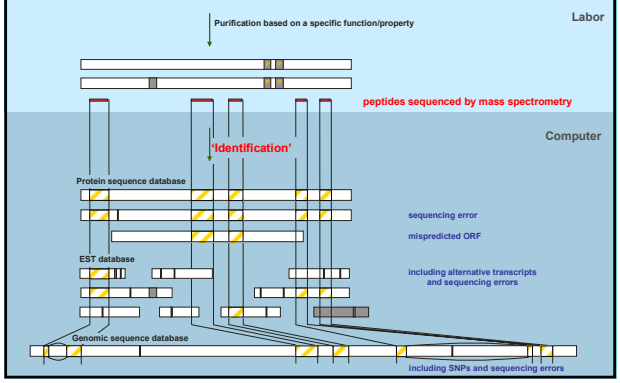
Scientific approach



Scientific approach

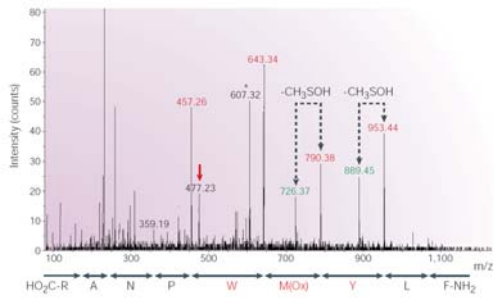


Scientific approach



Modified peptides

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

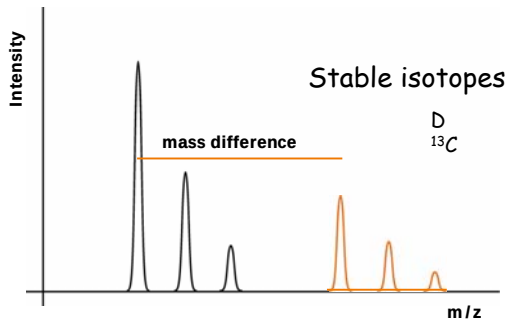


Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol. 2004 Sep;5(9):699-711. Review.

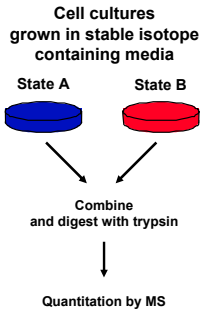
Quantitation in MS

- 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.

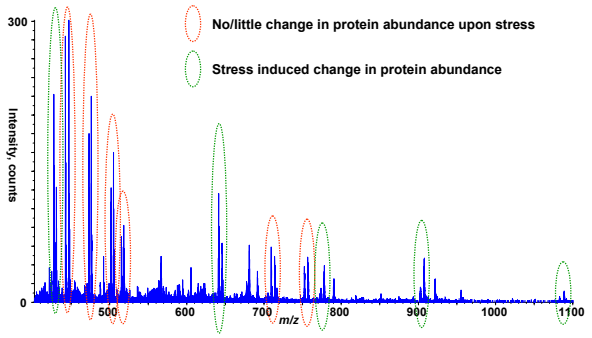
Quantitation in MS



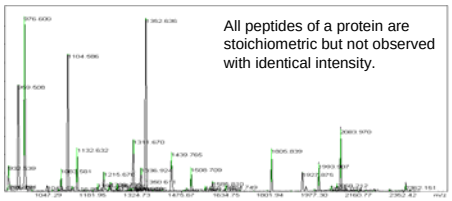
In vivo labeling with *SILAC*



Analysis of proteins from stressed cells



Stoichiometry



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

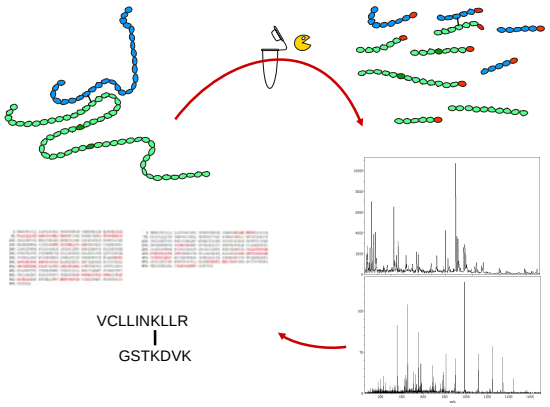
Intensity in mass spectrum not direct consequence of abundance but influenced by many molecule-specific factors
=> Apple-orange problem

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

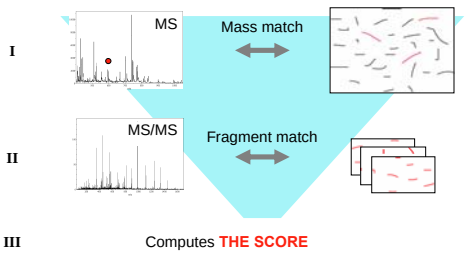
Protein-protein interactions

• Can be analyzed using same tools as for protein identification (mass spectrometry and database searching).

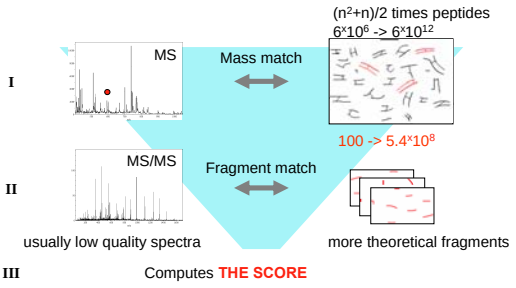
• Need to cross-link proteins to maintain their proximity also after proteolysis.



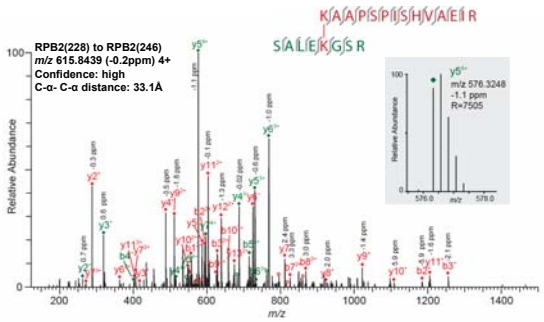
Normal peptide identification



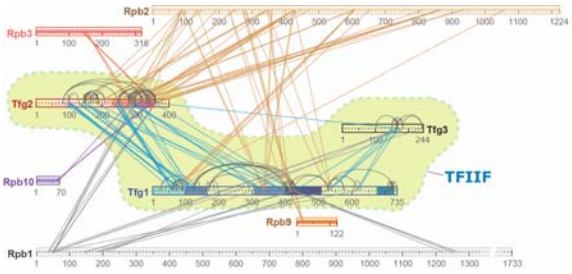
~~Normal~~ peptide identification
Cross-linked



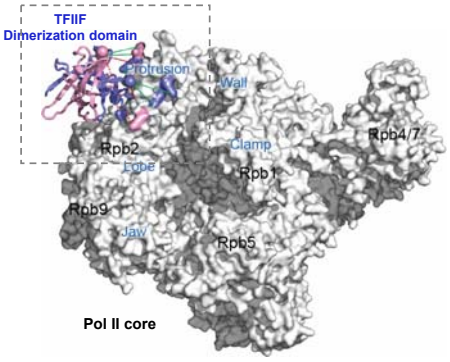
Maiolica A, Cittaro D, Borsotti D, Sennels L, Ciferri C, Tarricone C, Musacchio A, Rappsilber J.
Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching.
Mol Cell Proteomics. 2007 Dec;6(12):2200-11. Epub 2007 Oct 5.



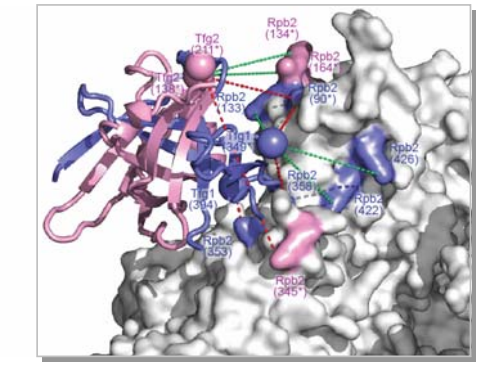
Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry.
Chen ZA, Jawhari A, Fischer L, Buchen C, Tahir S, Kamenski T, Rasmussen M, Lariviere L, Bukowski-Wills JC, Nilges M, Cramer P, Rappsilber J.
EMBO J. 2010 Feb 17;29(4):717-26.



Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry.
Chen ZA, Jawhari A, Fischer L, Buchen C, Tahir S, Kamenski T, Rasmussen M, Lariviere L, Bukowski-Wills JC, Nilges M, Cramer P, Rappsilber J.
EMBO J. 2010 Feb 17;29(4):717-26.



Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry.
Chen ZA, Jawhari A, Fischer L, Buchen C, Tahir S, Kamenski T, Rasmussen M, Lariviere L, Bukowski-Wills JC, Nilges M, Cramer P, Rappsilber J.
EMBO J. 2010 Feb 17;29(4):717-26.



Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry.
Chen ZA, Jawhari A, Fischer L, Buchen C, Tahir S, Kamenski T, Rasmussen M, Lariviere L, Bukowski-Wills JC, Nilges M, Cramer P, Rappsilber J.
EMBO J. 2010 Feb 17;29(4):717-26.