Introduction to Proteomics Analysis and Databases

Nicky Mulder: nicola.mulder@uct.ac.za

CBIO Computational Biology @ UCT

UNIVERSITY OF CAPE TOWN

What is Proteomics?

- Large-scale study of proteins to determine their function
- Proteome is protein complement of the genome
- Includes the study of:
  - Protein structure and function
  - Protein-protein interactions
  - Protein expression
  - Protein localization
  - Protein modifications
  - Etc.

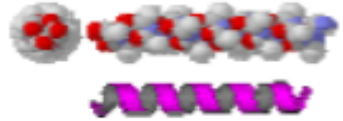
Proteomics studies

- Primary structure (*sequence*)

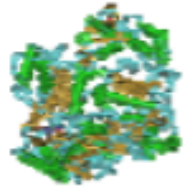
  …YSFVATAER…

Mass spectrometry

- Secondary structure (*structural elements*)



- Tertiairy structure (*3D shape*)



Xray, NMR

- Modifications (*dynamic, function*)
  phosphorylation

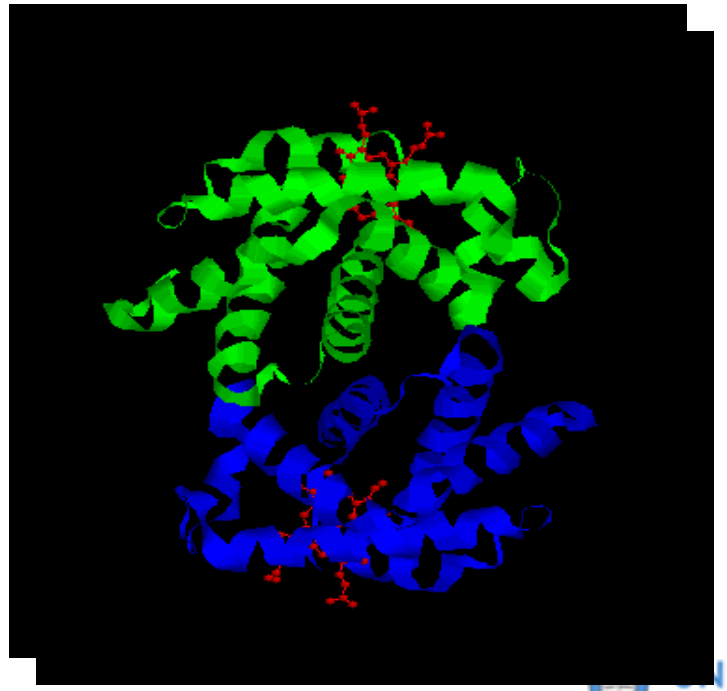
Mass spectrometry

- Processing (*targetting, activation*)
  trypsin
  platelet activity

Localization studies

Protein structure

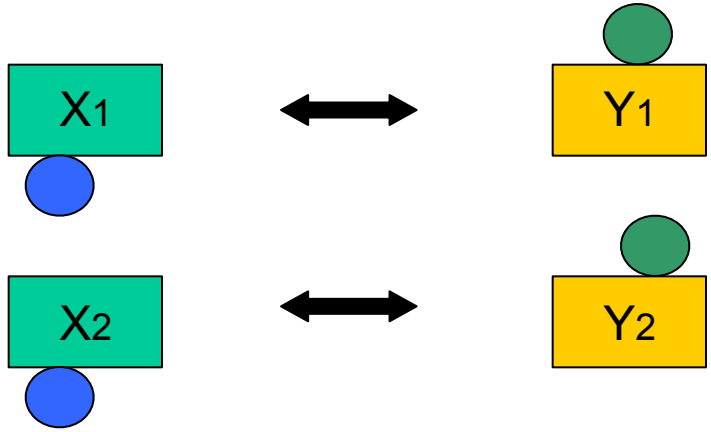
- Determined by x-ray crystallography or NMR
- Provides clues about protein function
- Very time-consuming and not always possible to crystallize a protein!



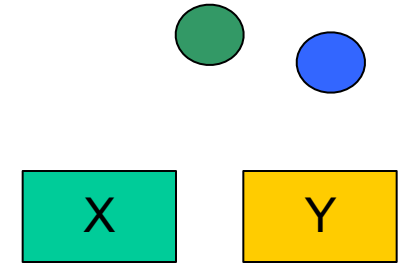
Physical interactions

- Experiments to identify physical interactions between DNA and proteins (for example TFs) or between two proteins:
  - Yeast two hybrid
  - Protein arrays

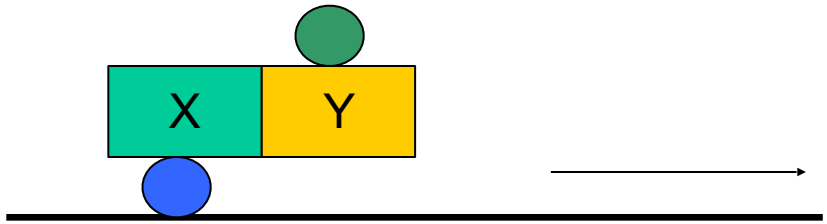
Yeast two hybrid



Two proteins X and Y fused to DNA-binding and activation domains



Interaction of X and Y = expression of reporter gene

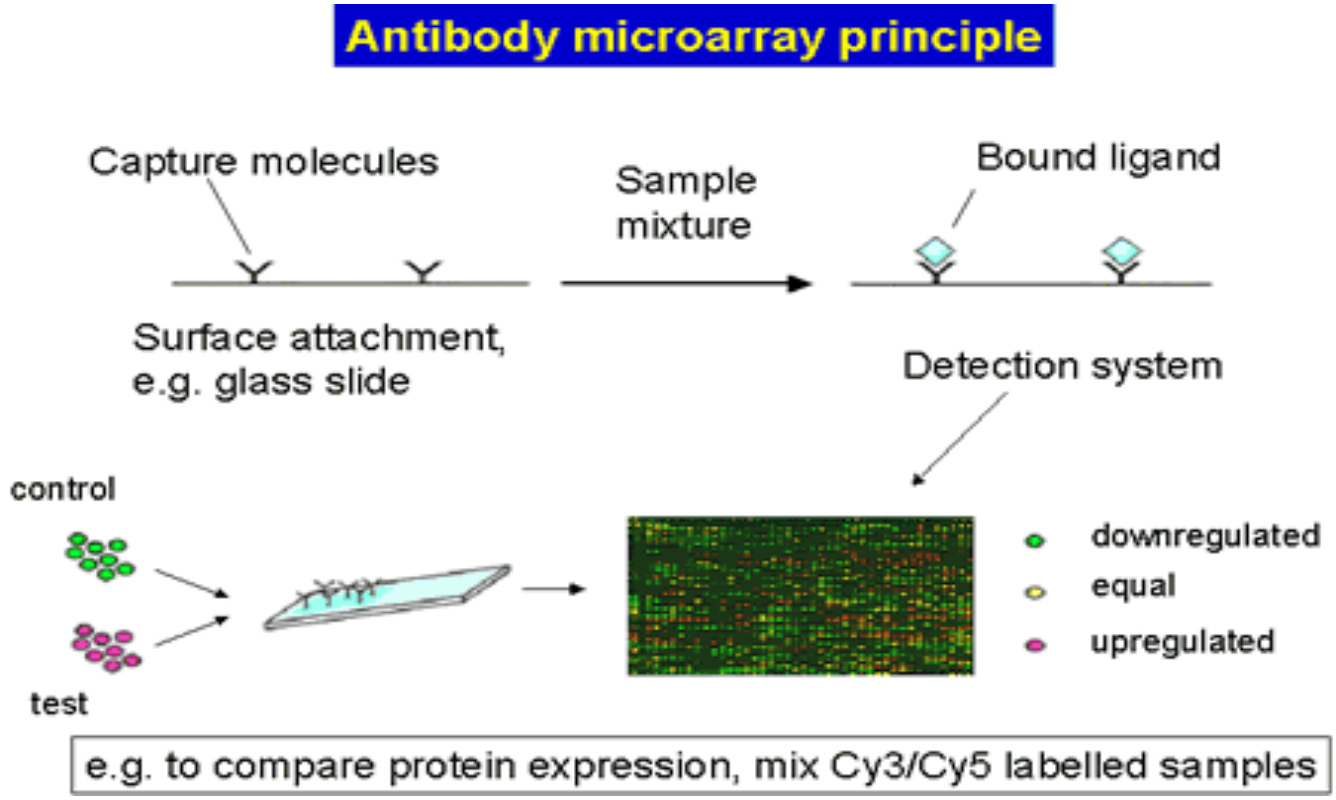


Protein arrays

Most common are antibody arrays

Good for well studied organisms

Limited for new ones



http://www.functionalgenomics.org.uk/sections/resources/protein_arrays.htm

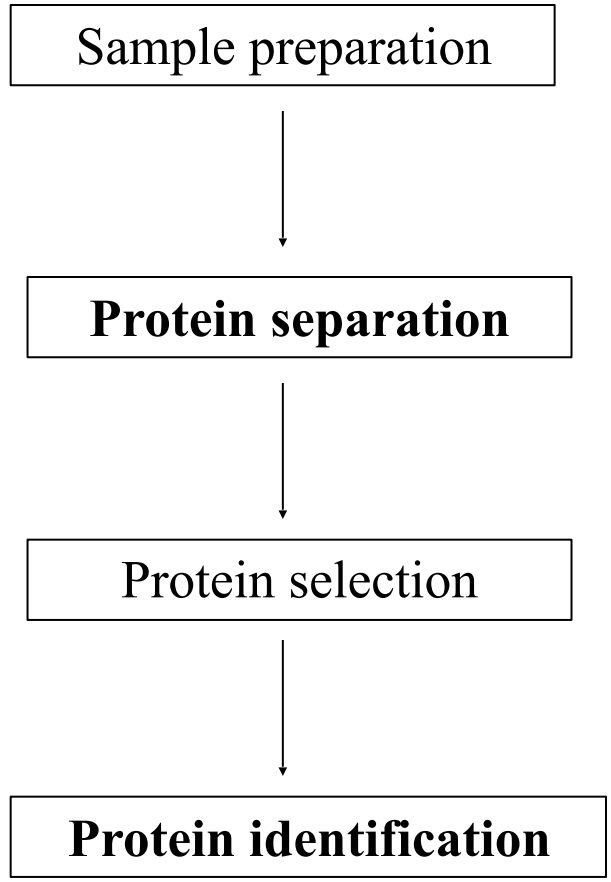
Protein-protein interaction databases

- Protein-protein interaction databases store pairwise interactions or complexes
- IntAct
- DIP (Database of Interacting Proteins)
- BIND (Biomolecular Interaction Network Database)

Other experimental data

- **Protein localization:**
  - Co-localized proteins may share functional relationships
  - Not always case, e.g. in cytoplasm
  - Localization can change with environment
- **Protein expression:**
  - Measuring quantity of protein in more than 1 sample
  - Identification of relevant proteins

Workflow of a proteomics experiment



Sample can be from patient cohort, cell selection, fraction, etc.

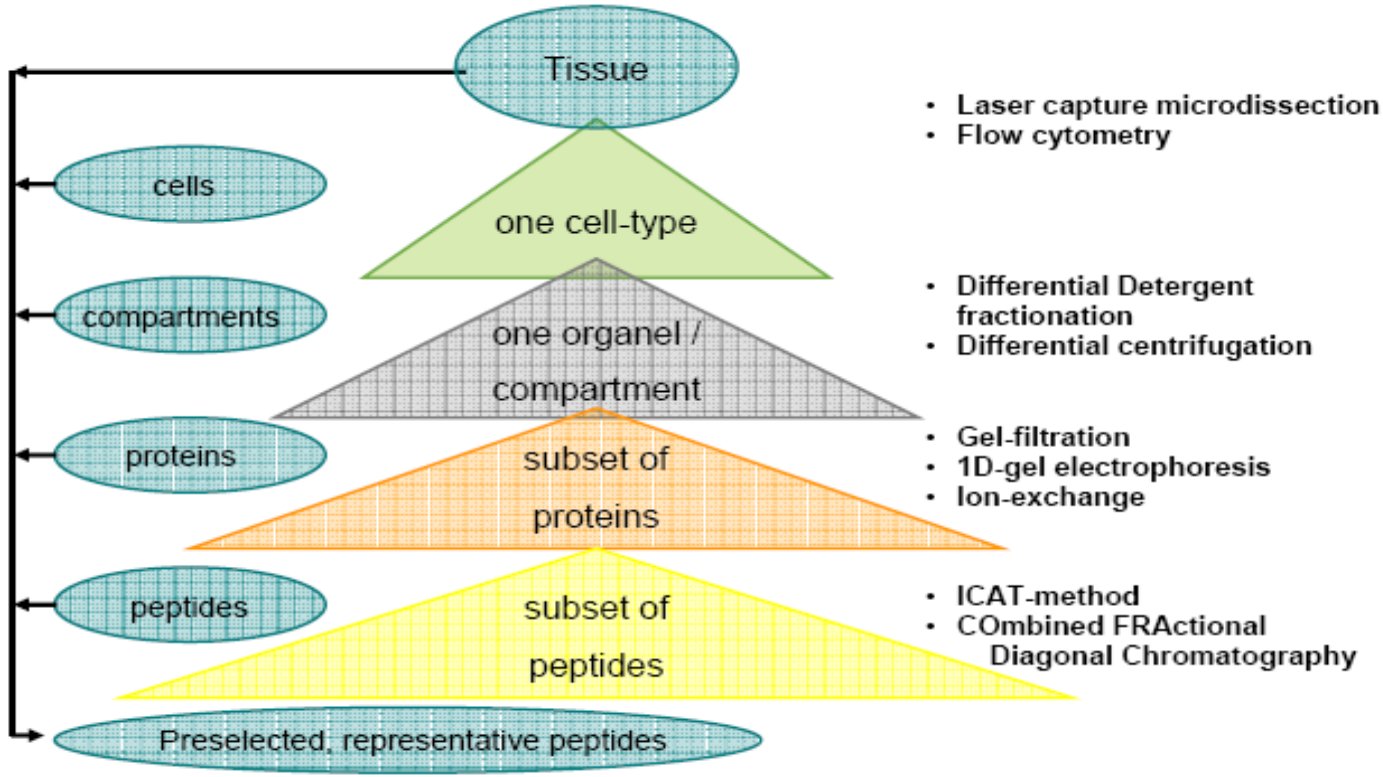
Different separation techniques, e.g. 2-D PAGE, HPLC, ICAT, etc.

Depends on separation method

Usually mass spectrometry

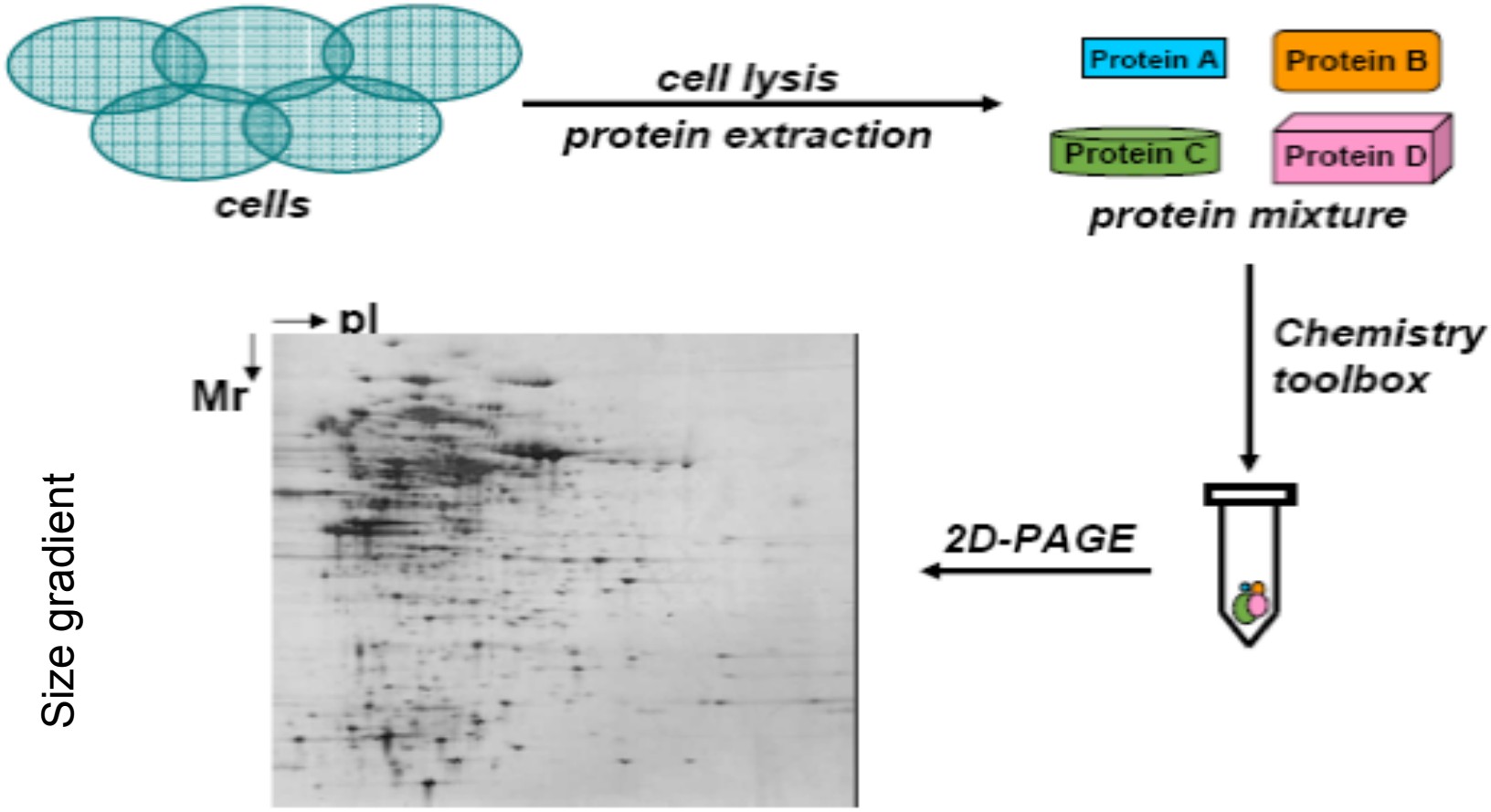
Protein separation

- 2D PAGE
- Gel-free systems:
  - ICAT
  - HPLC

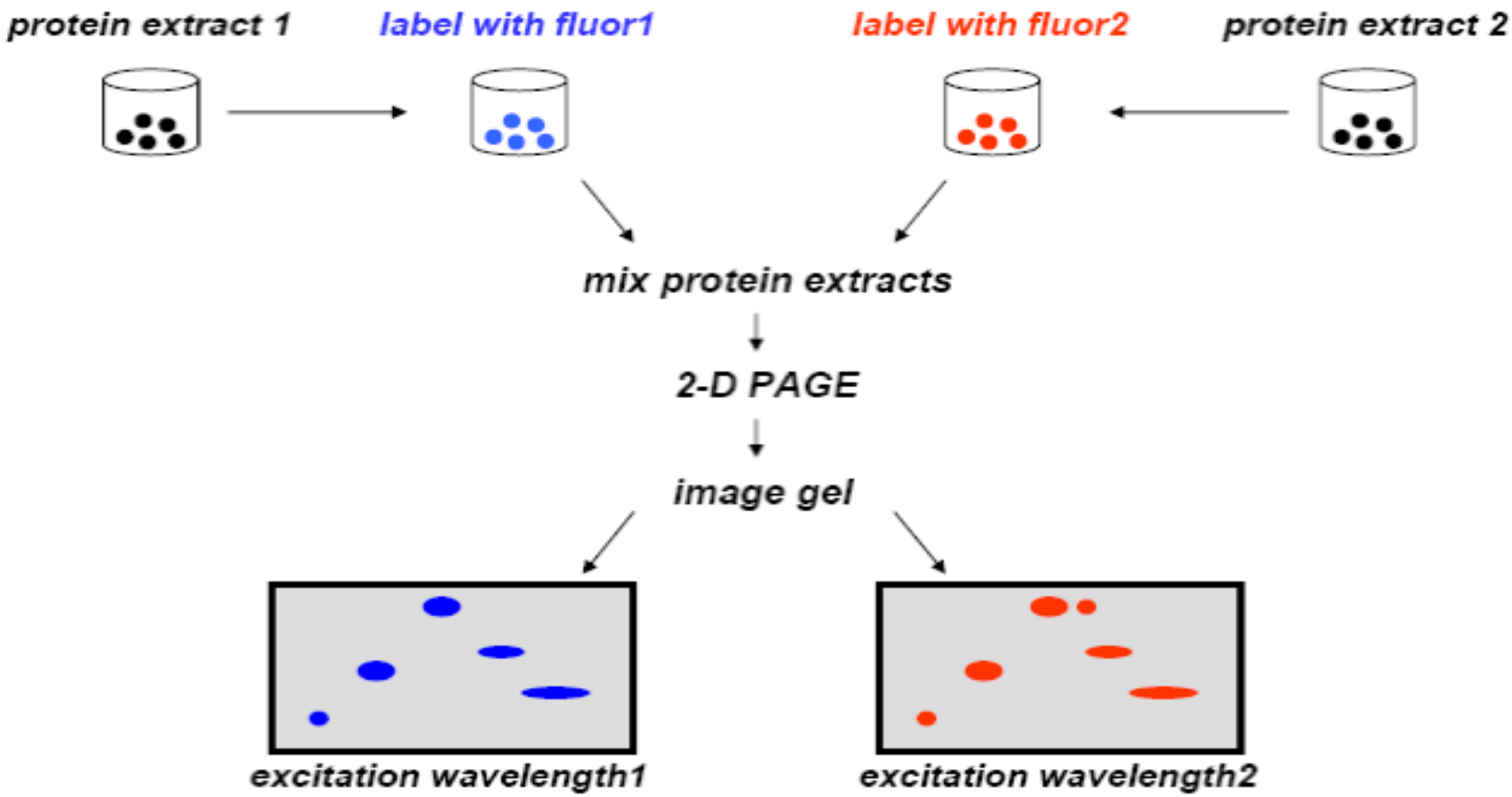


Mass spec –digest proteins further

Protein separation -2D PAGE



Comparing 2 samples



Bioinformatics component

- Sample tracking
- Image capture
- Image analysis and comparison:
  - Convert to matrix for example
  - Measuring intensities
  - Removing background noise
  - Finding difference between gels

Problems with 2D PAGE

- Some proteins can't be detected
  - Low abundance
  - Highly charged (run off gel)
- Reproducibility -can't easily compare across different gels
- Imperfect separation –multiple proteins in a spot
- Experiment takes time and special skills

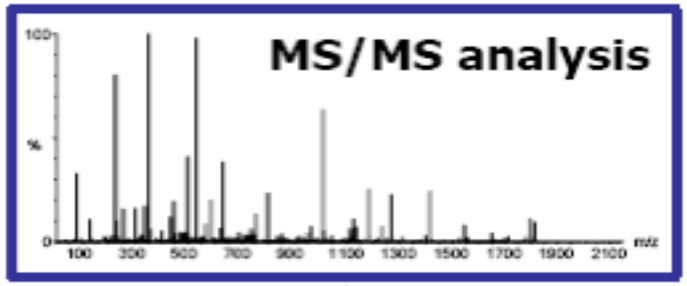
After 2D PAGE

http://www.akh-wien.ac.at/biomed-research/hrz/platweb1.htm

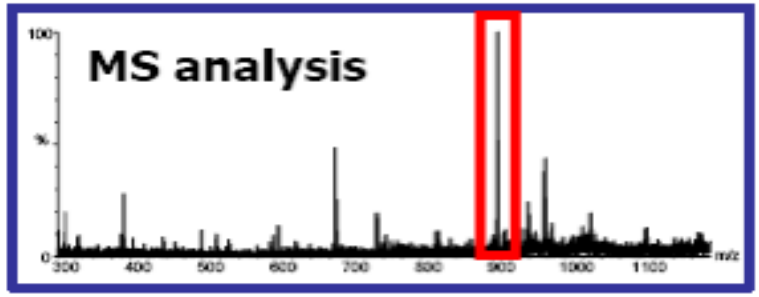
protein extraction

complex protein mixture

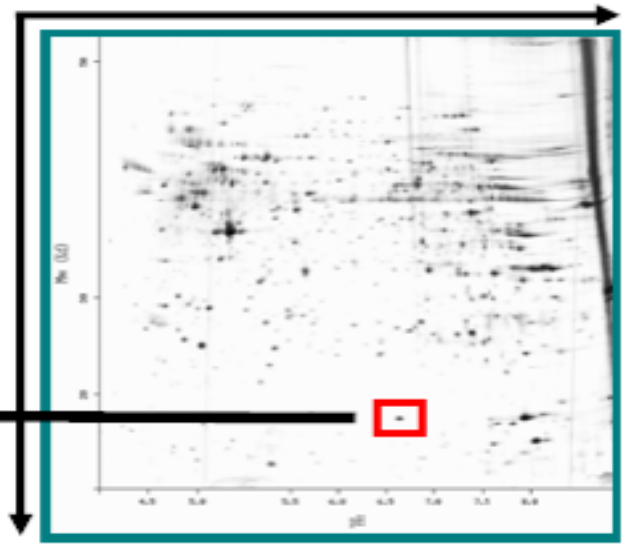
2D-PAGE separation



fragmentation

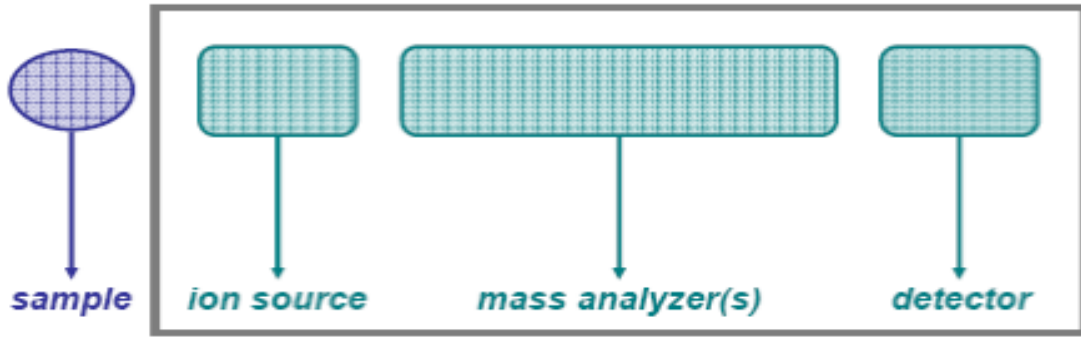


tryptic digest



Mass spec

- Digest proteins with e.g. trypsin (lysine or arginine)
- Proteins are ionized and brought into gas phase
- Move through mass analyzer which separates them based on mass
- Detector records presence of ions

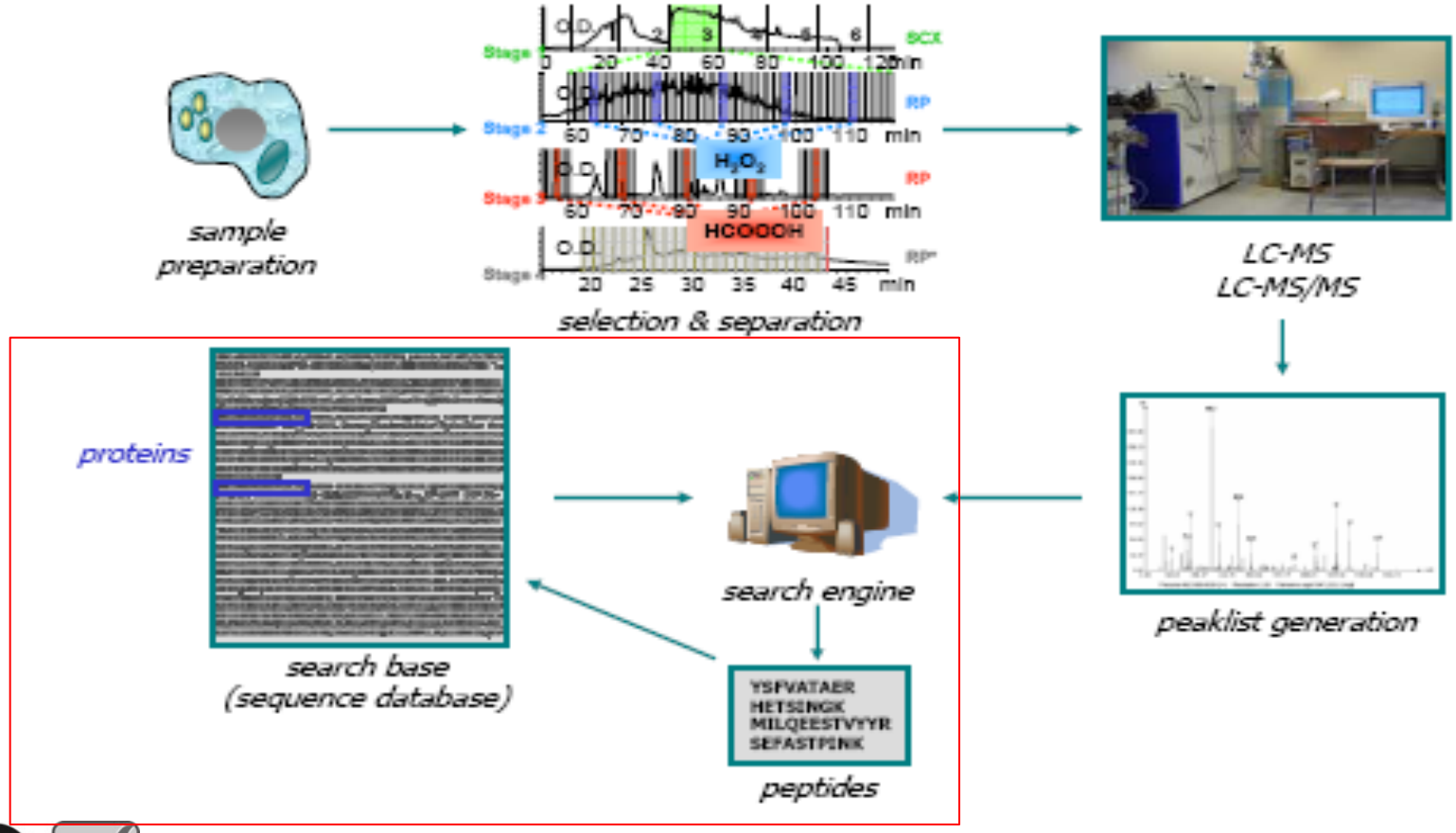


Generalized mass spectrometer

F=ma

F related to charge, electric field, velocity

Protein identification with Mass Spectrometry



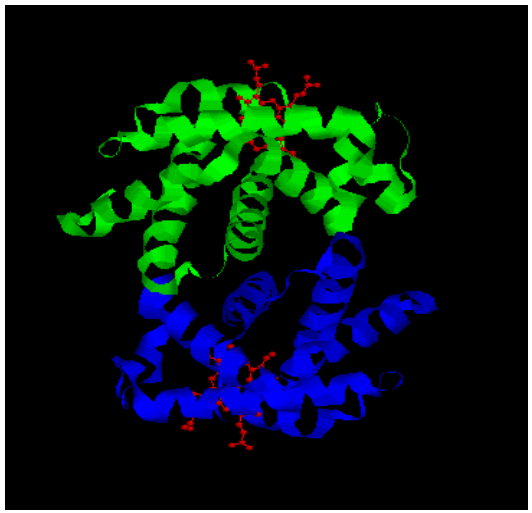
Protein identification

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFA
TLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHK

denature

digest with trypsin

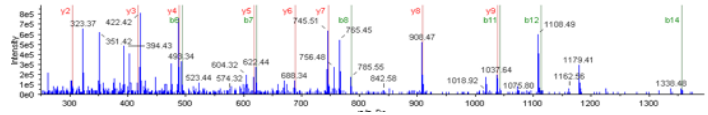
Recognises lysine (K) & arginine (R)



Mass spec

VHLTPEEK
SAVTALWGK
VNVDEVGGEALGR
LLVVYPWTQR
FFESFGDLSTPDAVMGNPK
VK
AHGK
K
VLGAFSDGLAHLDNLK
GTFATLSELHCDK
LHVDPENFR
LLGNVLVCVLAHHFGK
EFTPPVQAAYQK
VVAGVANALAHK

mass spectrum



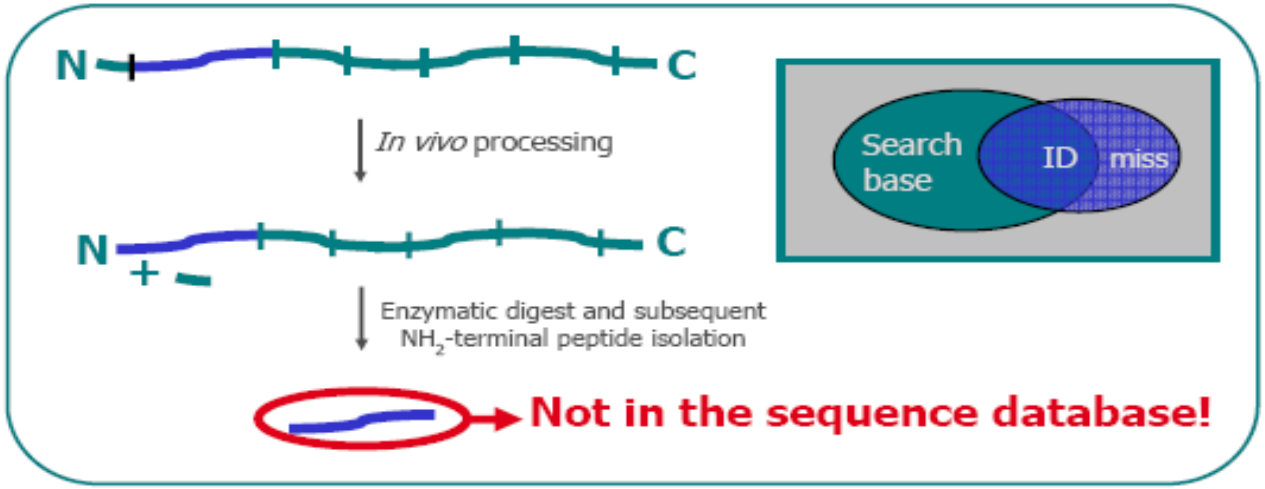
compare with theoretical peptide spectra; ID = best similarity

Problems with mass spec ID

- Protein samples often contain a mixture of proteins
- Digestion/fragmentation isn't always complete
- Not all proteins get ionized
- Background noise in spectra
- Proteins can contain modifications, which will change mass
- Differences with sequence databases

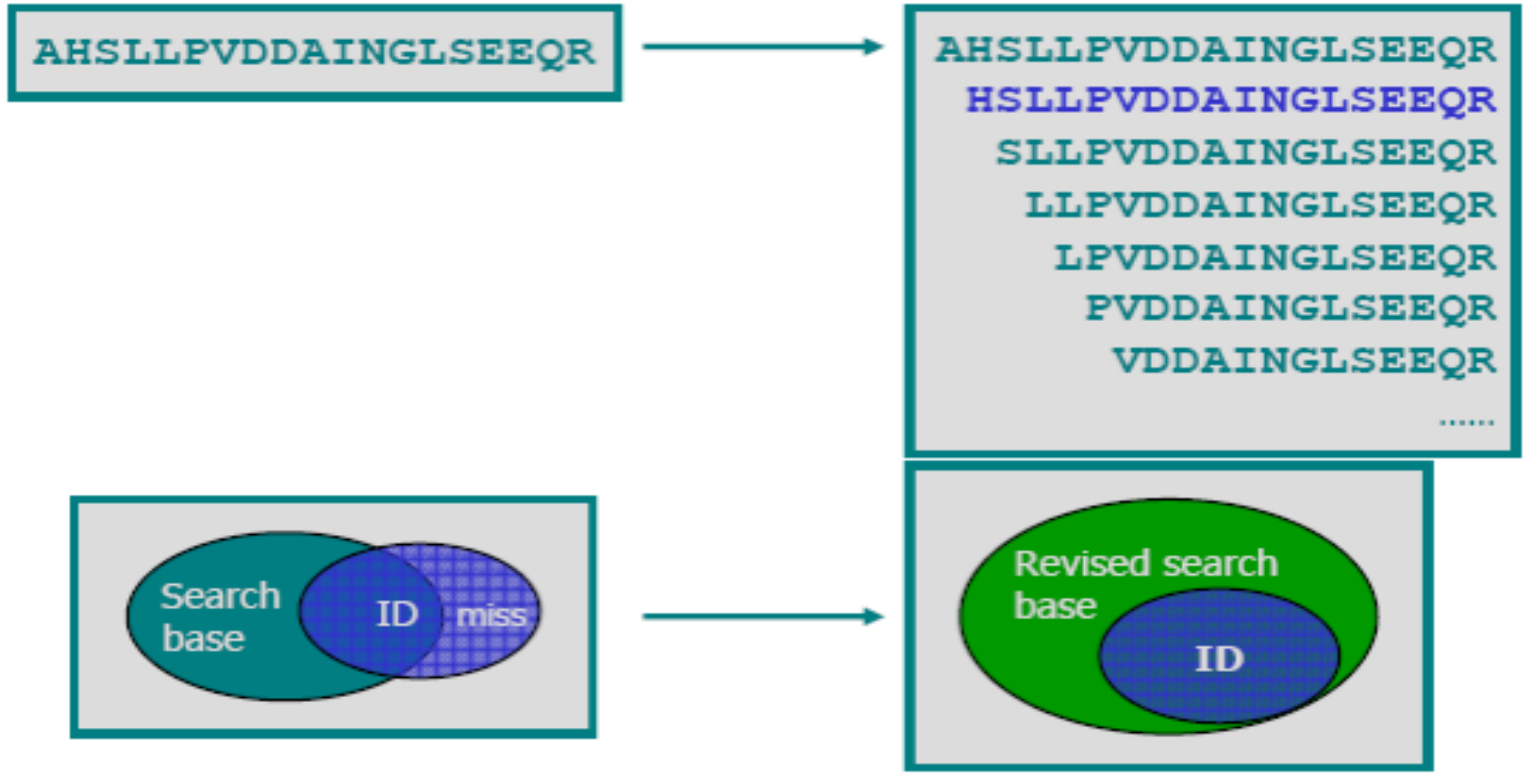
Issue with protein sequence databases

- Protein sequence database –UniProt -redundant
- NCBI non-redundant database (contains fragments) – redundant at peptide level
- Problem:



Changes theoretical spectrum

Extending database content



But: don't want to end up with too many redundant (identical) peptides

Solving problems –Tandem MS

- Two rounds of mass spec
- Fragment peptides and obtain spectrum
- Select peak you want then fragment this again
- Able to better separate peptides/proteins

MS/MS or Tandem MS

http://www.akh-wien.ac.at/biomed-research/hitz/platweb1.htm

protein extraction

complex protein mixture

2D-PAGE separation

pI

MW

tryptic digest

MS analysis

fragmentation

MS/MS analysis

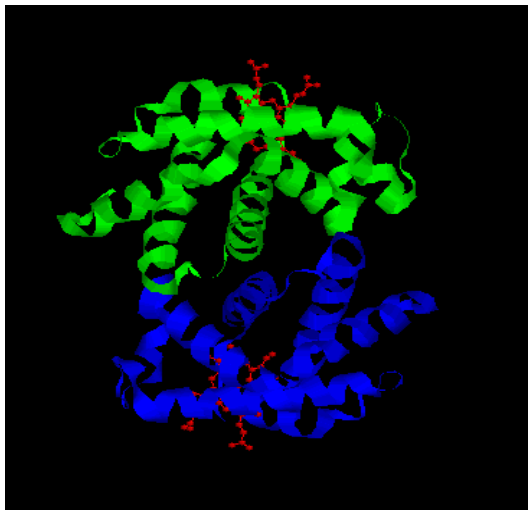
CBIO Computational Biology @ UCT

UNIVERSITY OF CAPE TOWN

Protein identification

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFA
TLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHK

denature



digest with trypsin

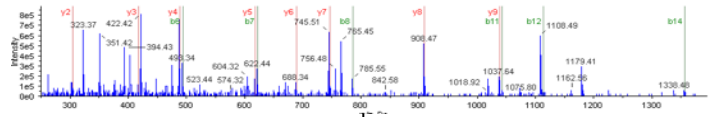
VHLTPEEK
SAVTALWGK
VNVDEVGGEALGR
LLVVYPWTQR
FFESFGDLSTPDAVMGNPK
VK
AHGK
K
VLGAFSDGLAHLDNLK
GTFATLSELHCDK
LHVDPENFR
LLGNVLVCVLAHHFGK
EFTPPVQAAYQK
VVAGVANALAHK

Recognises lysine (K) & arginine (R)

Mass spec

V	HLTPEEK
VH	LTPEEK
VHL	TPEEK
VHLT	PEEK
VHLTP	EEK
VHLTPE	EK
VHLTPEE	K

mass spectrum



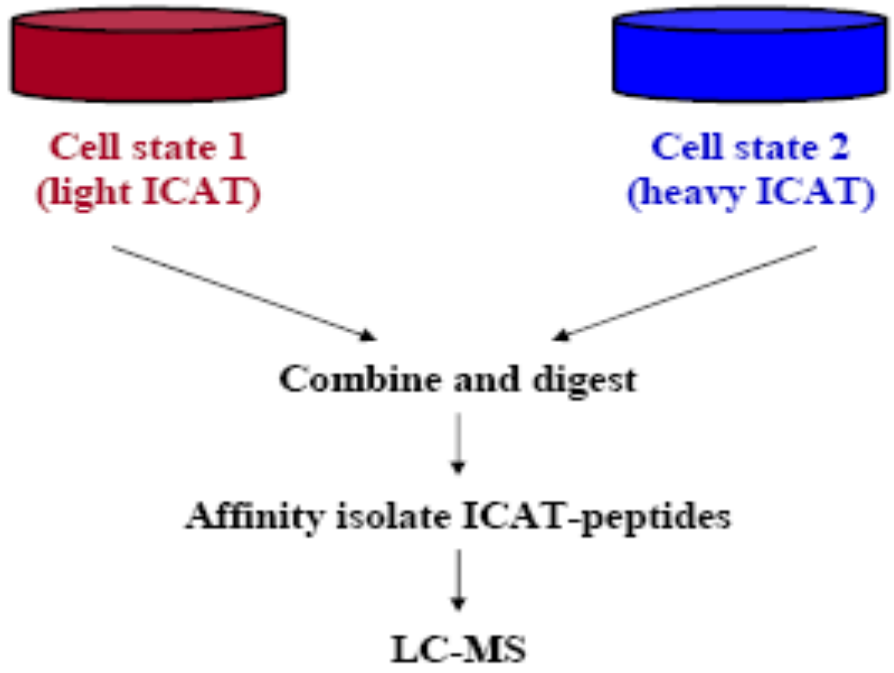
compare with theoretical peptide spectra; ID = best similarity

CBIO Computational Biology @ UCT

MS for comparative proteomics

- Isotope Coded Affinity Tag (ICAT)
- To identify and quantify proteins in 2 populations – extract proteins from 2 samples
- Modify cysteine residues using the ICAT molecule (biotin + linker (heavy -deuterium or light –hydrogen) + thiol-reactive group)
- Extract proteins with avidin affinity column
- Analyze with mass spec
- Measures every protein containing a cysteine

ICAT



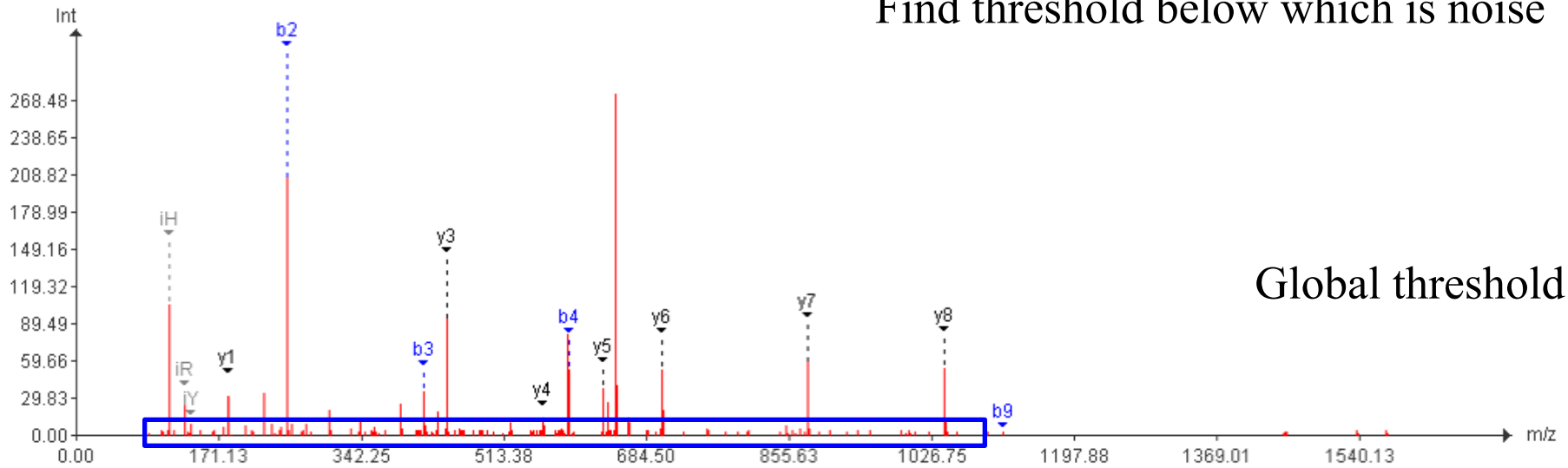
Provides way of introducing mass difference between 2 samples

Outline of data analysis

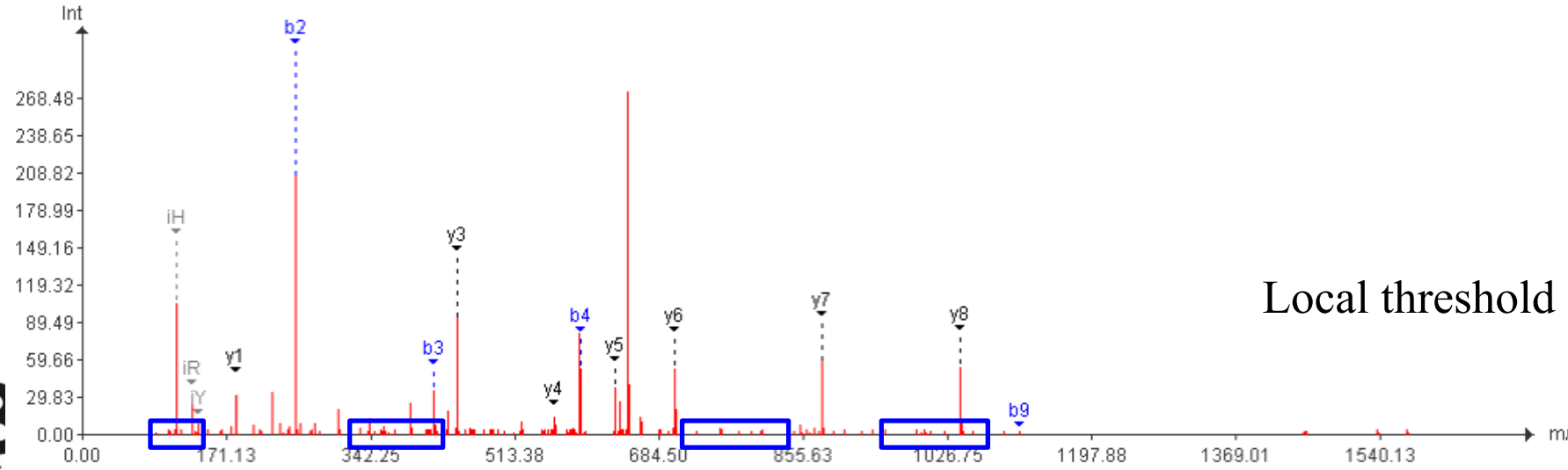
- Proteomics data analysis (MS)
  - Pre-processing
  - Protein/peptide identification
  - Proteomics-related databases
- Data mining:
  - Pathway analysis
  - Enrichment analysis

Pre-processing 1: noise reduction

Find threshold below which is noise

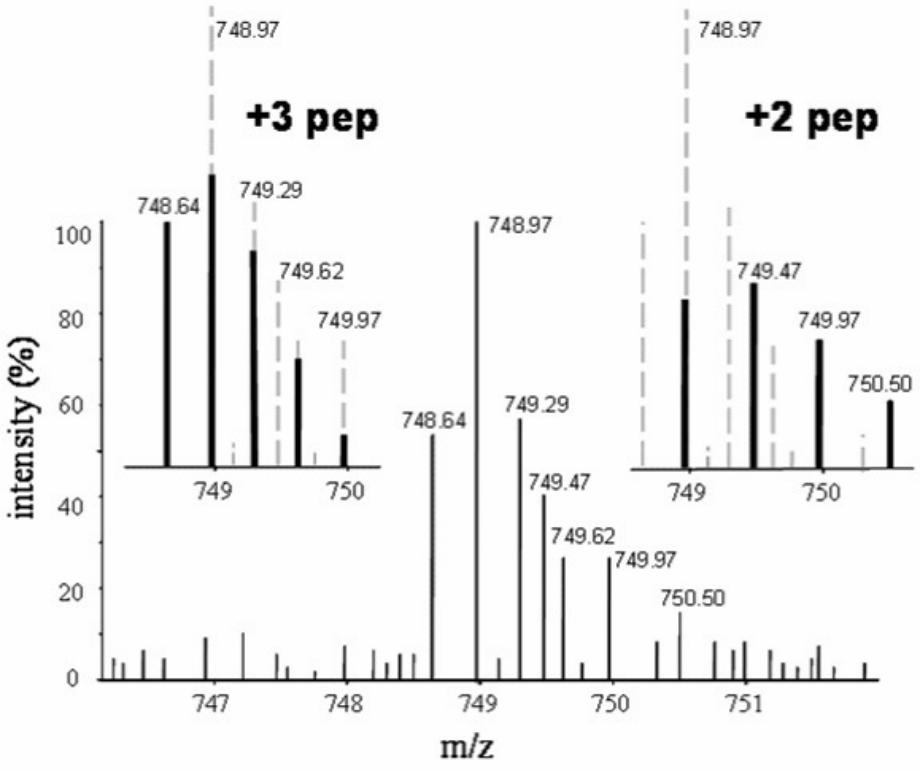


Global threshold



Local threshold

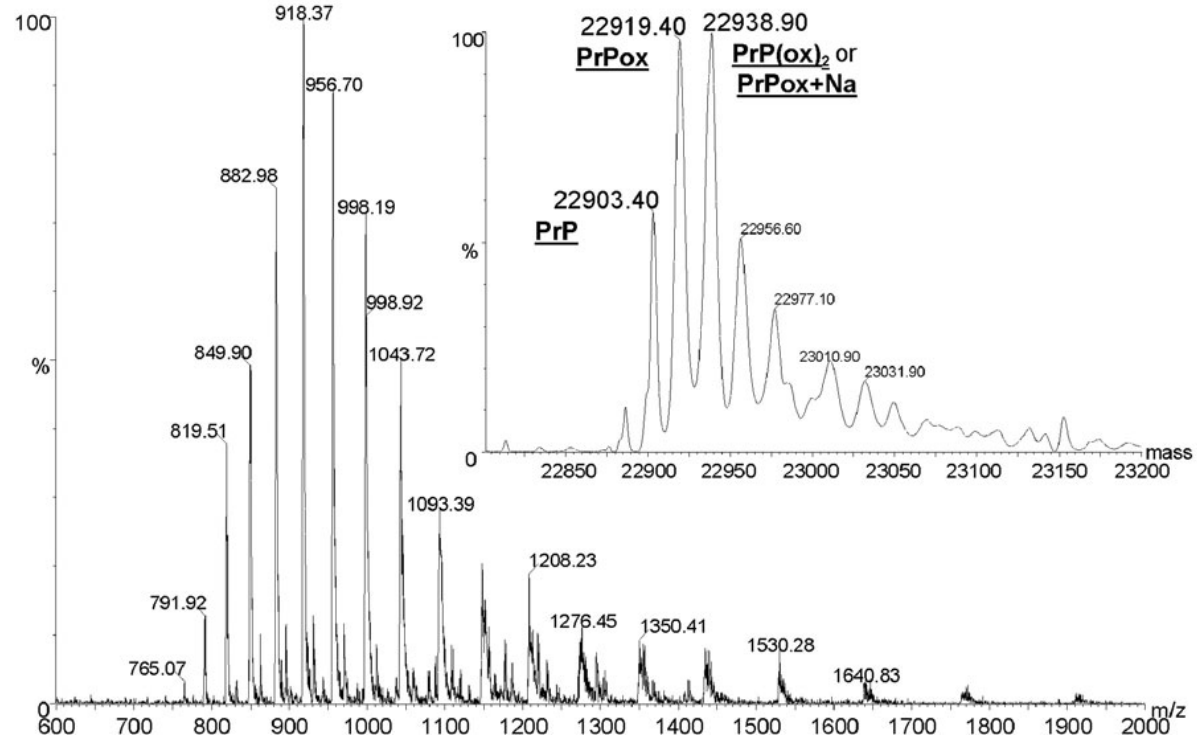
Pre-processing 2: charge deconvolution



peptides

www.purdue.edu/dp/bioscience/image/spectrum.jpg

proteins

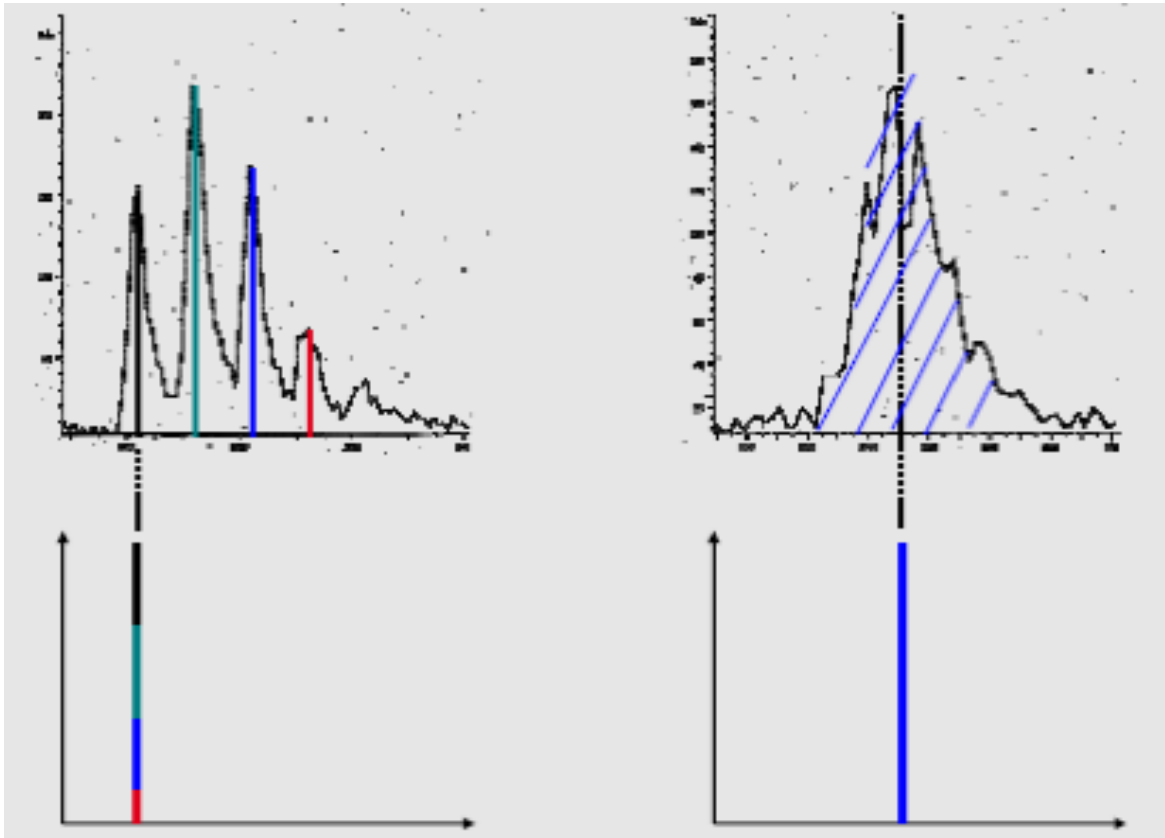


Gill et al. EMBO journal, 2000

CBIO Computational Biology @ UCT

UNIVERSITY OF CAPE TOWN

Pre-processing 3: peak picking



Spectrum filtering and clustering

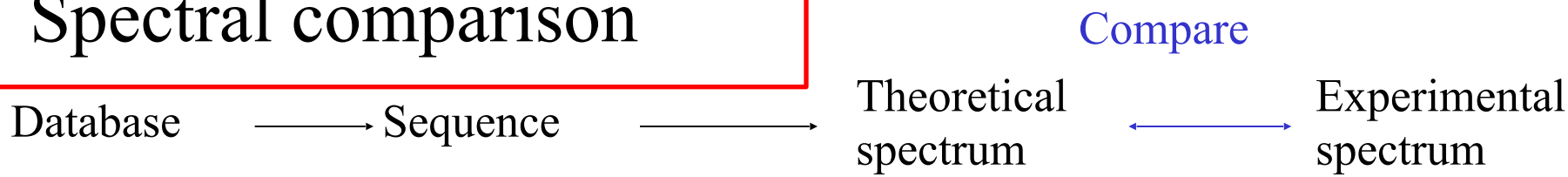
- Filter out low quality spectra
- Cluster spectra - same peptide could be fragmented several times
- To resolve redundancy –group and merge spectra
- Now ready for identification!

Assumptions made in ID

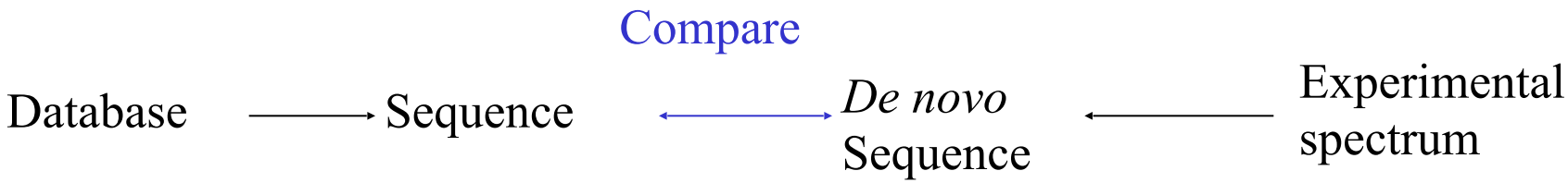
- All peaks in spectrum are from the same protein
- The protein is in the same form as it is in the database
- Protein is completely digested
- All pieces produce a signal

Types of PFF

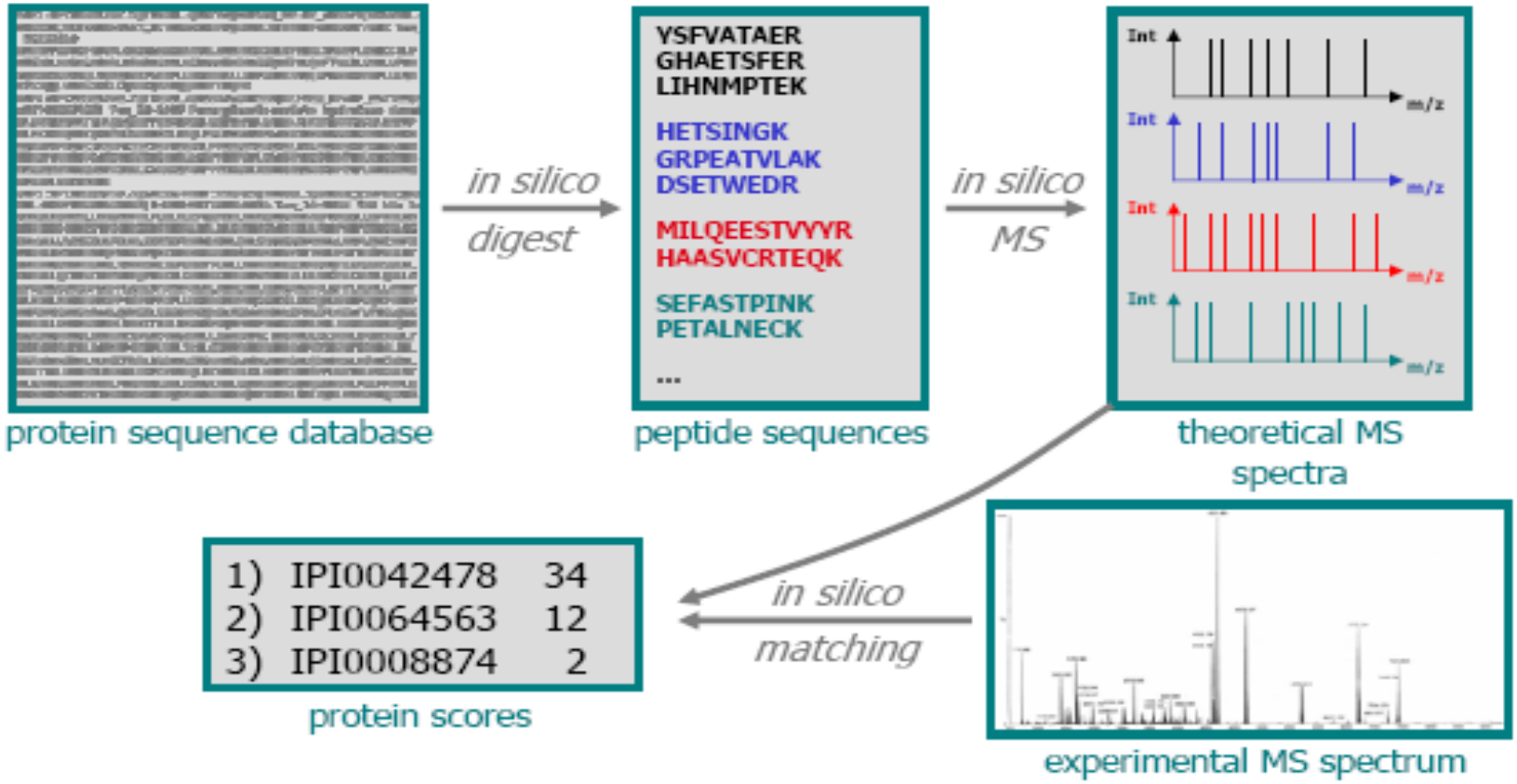
- Spectral comparison



- Sequence comparison



Peptide Fragment Fingerprinting (PFF)



Software to do this

- MASCOT (http://www.matrixscience.com)
  - Predicts threshold score that needs to be passed
  - Provides rank, score and threshold
- SEQUEST (http://fields.scripps.edu/sequest)
  - User decides on threshold
  - Provides rank and score
- XTandem (http://www.thegpm.org/TANDEM)

Example results

List of proteins and the peptides they matched

IPI00025363
Protein Detail
(top ↑)

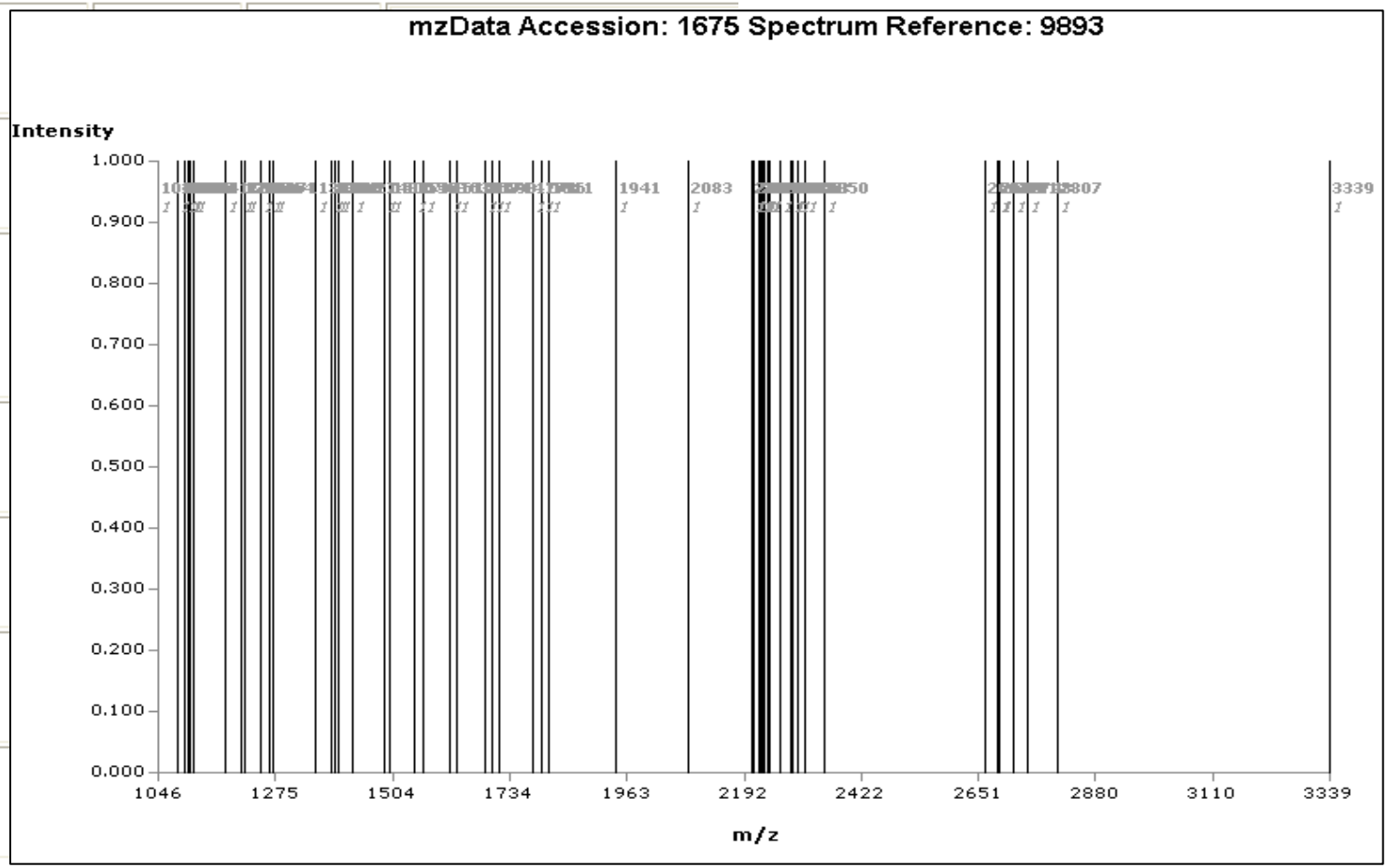
VDFSLAGALNAGFKETR	Location: 50 - 66	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
ALAAELNQLR	Location: 96 - 105	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
LADVYQAELR	Location: 112 - 121	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
DNLAQDLATVR	Location: 142 - 152	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
LEAENNLAAYR	Location: 163 - 173	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
KIESLEEEIR	Location: 189 - 198	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
QLQSLTCDLESLR	Location: 288 - 300	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
	Modification MOD:00397 from database PSI-MOD [1 .0] at position		
EQEERHVR	Location: 312 - 319	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
EAASYQEALAR	Location: 320 - 330	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
EAASYQEALARLEEEGQSLK	Location: 320 - 339	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
HLQEYQDLLNVK	Location: 345 - 356	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint

Example results

VDFSLAGALNAGFKETR	Location: 50 - 66	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
ALAAELNQLR	Location: 96 - 105	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
LADVYQAELR	Location: 112 - 121	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
DNLAQDLATVR	Location: 142 - 152	Spectrum: 9893	Source: PRIDE; Name: Identified by peptide mass fingerprint
LEAENNLAAYR			
KIESLEEEIR			
QLQSLTCDLESLR			
EQEERHVR			
EAASYQEALAR			
EAASYQEALARLEEEGQSLK			
HLQEYQDLLNVK			

IPI00025363 Protein Detail

(top ↑)



Results continued

- Usually get multiple proteins resulting
- Peptides can match different proteins

  Protein inference:

Minimal set

Peptide	a	b	c	d
Proteins				
proteinX	x		x	
proteinY	x			
proteinZ		x	x	x

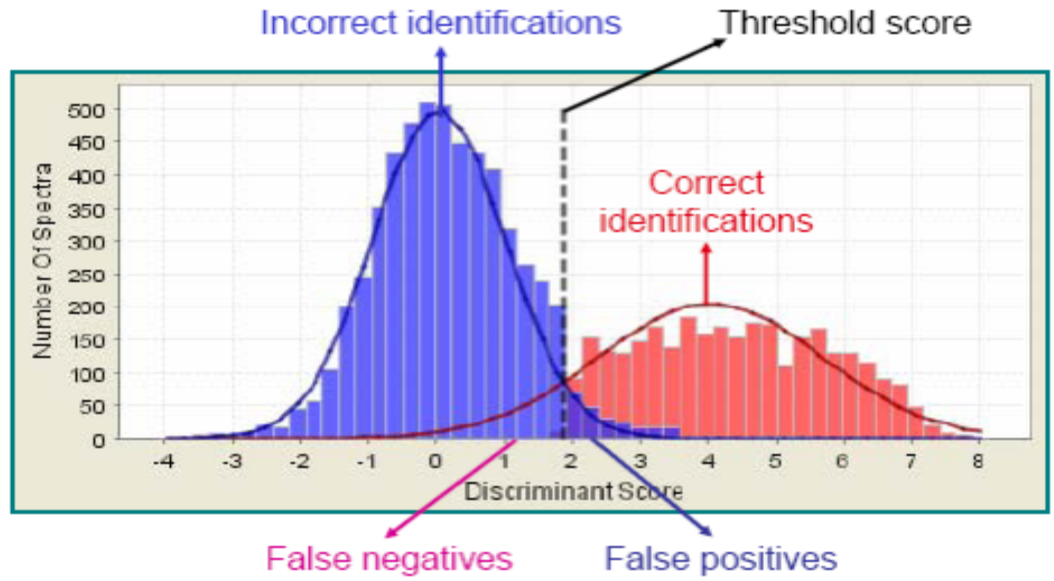
Maximal set

Peptide	a	b	c	d
Proteins				
proteinX	x		x	
proteinY	x			
proteinZ		x	x	x

Truth is in between –or go for best annotated proteins

Problems with peptide ID

- Don’t end up with actual sequence
- Working on unsequenced genome
- Ambiguity with protein families
- False positive and false negative matches



Adapted from: www.proteomesoftware.com – Wiki pages

Combining search algorithms

- Diff programs have different strengths
- All give some Fs and Ns
- Run a combination of search engines then:
  - Union of results –extends identifications –fewer Ns
  - Intersection of results –stricter set of results –fewer Fs

  - What is your research question?

Validation: Peptide- and ProteinProphet

- PeptideProphet:
  - Post-processes peptide identification data to assess the probability that an ID is correct
- ProteinProphet:
  - Tries to produce minimum protein set given list of proteins and probabilities
  - Probability based on peptide score
  - Degenerate peptides (mapping to >1 protein) get lower score

Decoy databases

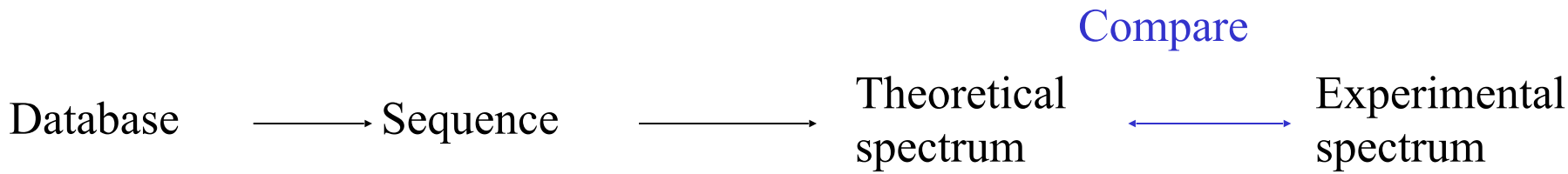
- Use to calculate probability of identifications and FP rate
- Three main types:
  - Reversed databases –reverse all sequences, e.g. RKLYWSML -> LMSWYLKR
  - Shuffled databases –shuffle all sequences, e.g. RKLYWSML -> YKRWLMSL
  - Randomized databases –all sequences are randomly generated, e.g RKLYWSML -> GKYSQTTDV

Decoy databases continued

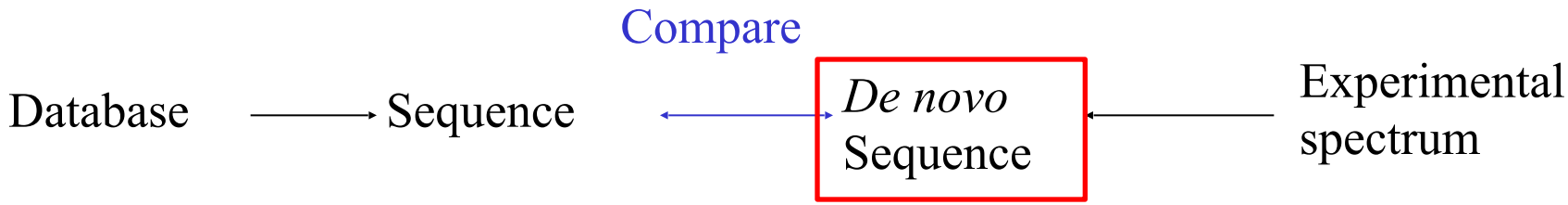
- Run ID on decoy database to see how many matches you get
- Some overlaps/identifications do occur –measure false positive rate
- Repeat decoy searches several times to estimate standard deviation

Types of identification

- Spectral comparison

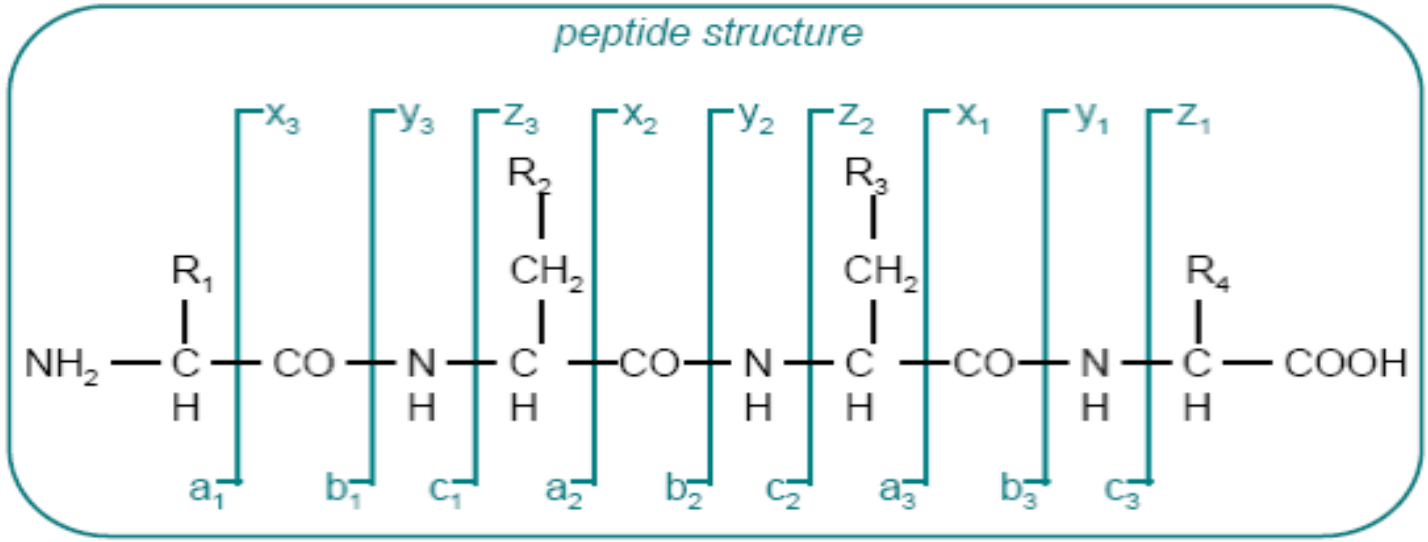


- Sequence comparison



Peptide fragmentation (MS/MS)

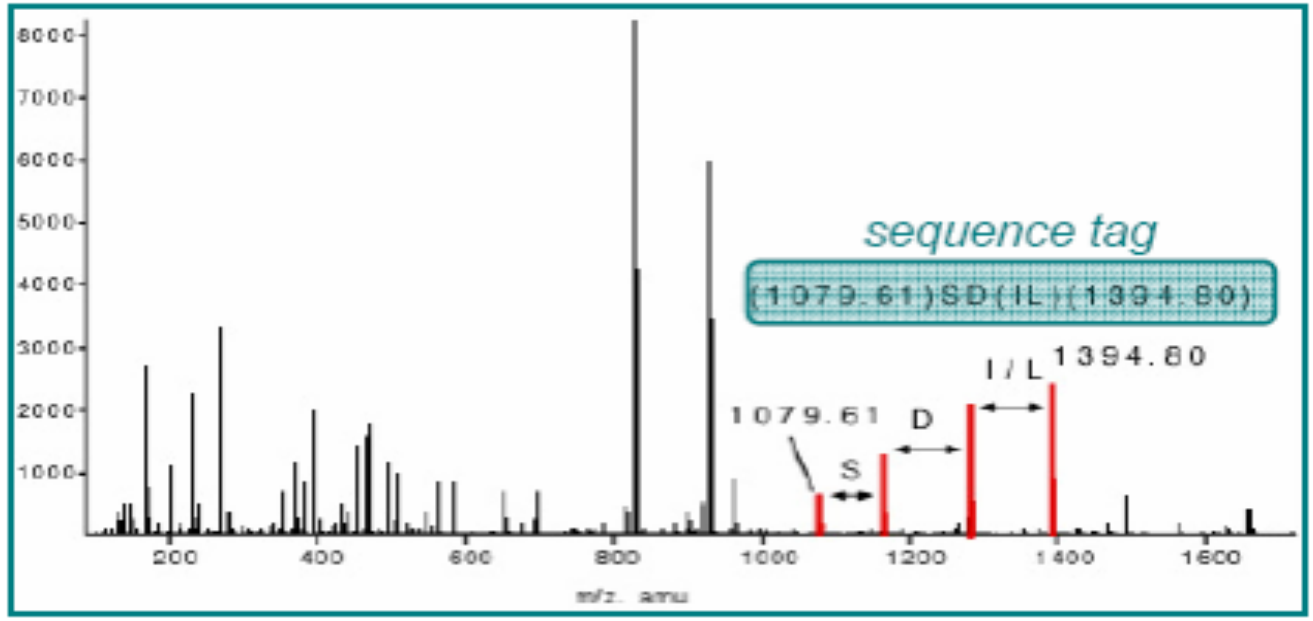
- Protein gets fragmented into different components



Sequencing with mass spec

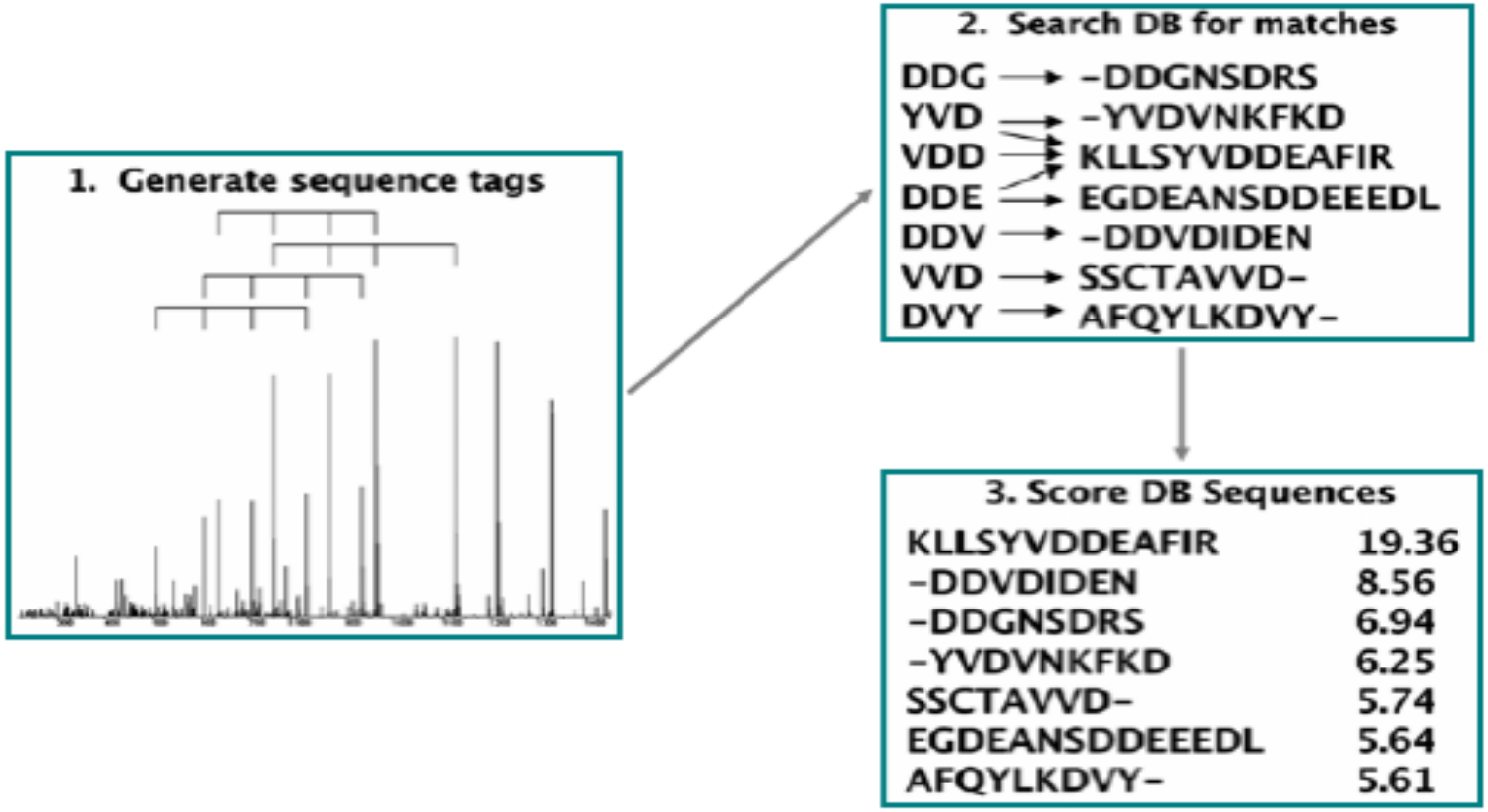
Compare

Database → Sequence ↔ *De novo* Sequence ← Experimental spectrum



Use known weights of amino acids

Sequencing continued



From: Tabb et al., Anal. Chem., 2003

Proteomics data repositories

- Why are they needed?
  - Data usually provided in PDF tables
  - No raw data provided in publications
  - Can't compare data
  - Links to other data sources
- Some examples:
  - PeptideAtlas
  - PRIDE

PRIDE example 1

PRIDE Experiment Collection Version 2.1

Experiment 1: HUPO Brain Proteome Project: BPP_PilotProject_Lab_10: human, post mortem (11h), brain number RZ104: Peptide Mass Fingerprint Identifications 1149-96-7-RZ104 pH 4-7 Overlay

Experiment: (top ↑)

Description: HUPO Brain Proteome Project: BPP_PilotProject_Lab_10: human, post mortem (11h), brain number RZ104: Peptide Mass Fingerprint Identifications 1149-96-7-RZ104 pH 4-7 Overlay

Short Title: HUPO BPP: BPP_PilotProject_Lab_10: human, post mortem (11h), brain number RZ104: PMF Identifications 1149-96-7-RZ104 pH 4-7 Overlay

Accession: 1689

Links:

mzData Section (Tissue / sample details and Contact details)
MS Instrument details
Data processing details
Spectrum List

Meyer HE, Hamacher M. Quintessence from proteomics networks - the HUPO Brain Proteome Project Pilot Studies. Proteomics. 2006 Sep;6(18):4887-9

Source	Name	Value
PubMed	16967475	16967475
DOI	10.1002/pmic.200690105	
PRIDE	Reference reporting this experiment	

Hamacher M, Apweiler R, Arnold G, Becker A, Bluggel M, Carrette O, Colvis C, Dunn MJ, Frohlich T, Fountoulakis M, van Hall A, Herberg F, Ji J, Kretzschmar H, Lewczuk P, Lubec G, Marcus K, Martens L, Palacios Bustamante N, Park YM, Pennington SR, Robben J, Stuhler K, Reidegeld KA, Riederer P, Rossier J, Sanchez JC, Schrader M, Stephan C, Tagle D, Thiele H, Wang J, Wiltfang J, Yoo JS, Zhang C, Klose J, Meyer HE. HUPO Brain Proteome Project: summary of the pilot phase and introduction of a comprehensive data reprocessing strategy. Proteomics. 2006 Sep;6(18):4890-8

Source	Name	Value
PubMed	16927433	16927433
DOI	10.1002/pmic.200600295	
PRIDE	Reference reporting this experiment	
PRIDE	Reference describing data analysis	
PRIDE	Reference describing sample preparation	

Stephan C, Reidegeld KA, Hamacher M, van Hall A, Marcus K, Taylor C, Jones P, Muller M, Apweiler R, Martens L, Korting G, Chamrad DC, Thiele H, Bluggel M, Parkinson D, Binz PA, Lyall A, Meyer HE. Automated reprocessing pipeline for searching heterogeneous mass spectrometric data of the HUPO Brain Proteome Project pilot phase. Proteomics. 2006 Sep;6(18):5015-29

Source	Name	Value
PubMed	16927432	16927432
DOI	10.1002/pmic.200600294	
PRIDE	Reference reporting this experiment	
PRIDE	Reference describing data analysis	

Dowsey AW, English J, Pennington K, Cotter D, Stuehler K, Marcus K, Meyer HE, Dunn MJ, Yang GZ. Examination of 2-DE in the Human Proteome Organisation Brain Proteome Project pilot studies with the new RAIN gel matching technique. Proteomics. 2006 Sep;6(18):5030-47

Source	Name	Value
PubMed	16927431	16927431
DOI	10.1002/pmic.200600152	
PRIDE	Reference reporting this experiment	

PRIDE example 2

Protocol:				
	Name:	2DGE-IPG		
	Step #1	Source	Name	Value
		PRIDE	Separation	2DGE-IPG
	Step #2	Source	Name	Value
		PRIDE	Peptide cleavage	Trypsin

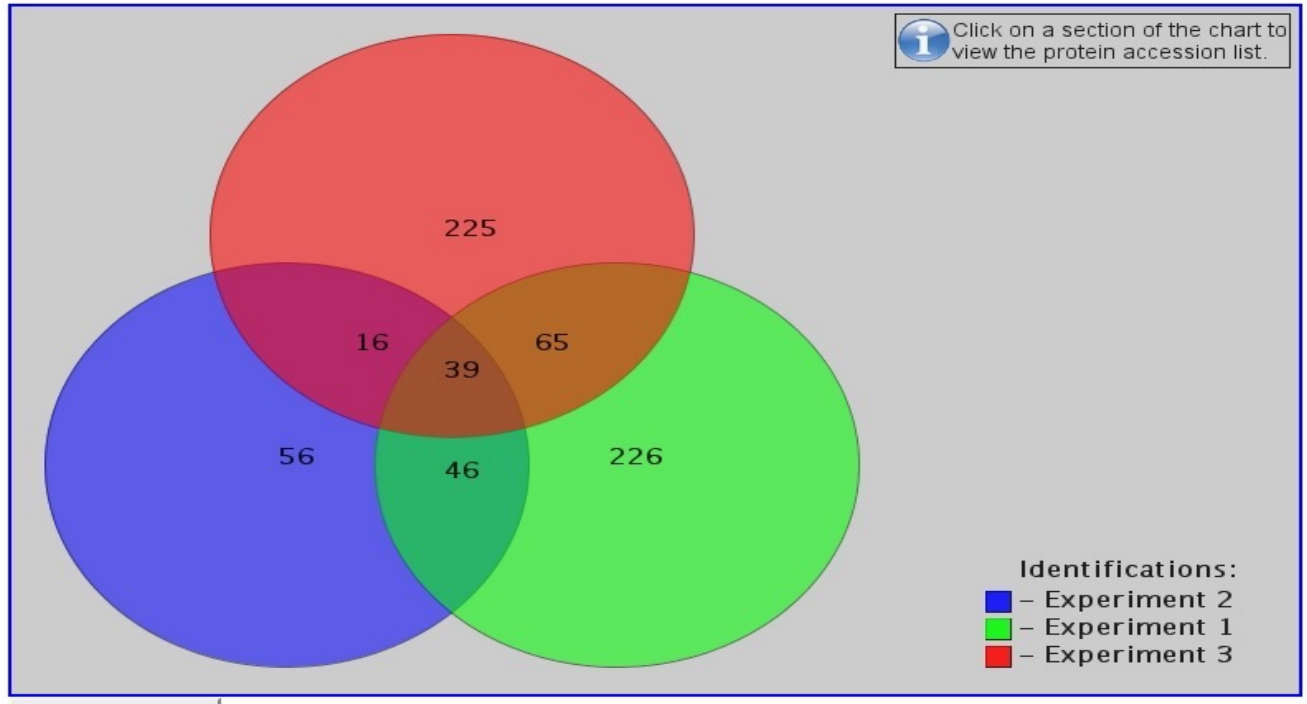
Additional:	Source	Name	Value
	PRIDE	Project	HUPO Brain Proteome Project

Identifications

Accession	Splice Isoform	Database	Score	Threshold	Search Engine	Sequence Coverage	Molecular Weight	pI	Additional Information		
									Source	Name	Value
IPI00022434 Peptide List (top ↑)		IPI			HUPO BPP Protein Merger		0.0 kDa	0.0	PRIDE	Identified by peptide mass fingerprint	
									PRIDE	Search database protein sequence	MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGE ENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVAD ESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEP ERNECFLQHKDDNPNLPRLVRPEVDVMCTAFHDNEETFLK KYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAA CLLPKLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAV ARLSQRFPKAEFAEVSKLVTDLTKVHTECCHGDLLECADD RADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVEND EMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYAR RHPDYSVVLLLRLAKTYETTLEKCCAAADPHECYAKVFDE FKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVP QVSTPTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVV LNQLCVLHEKTPVSDRVTKCCTESLVNRRPCFSALEVDET YVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHK PKATKEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLV AASQAALGL
IPI00025363 Peptide List (top ↑)		IPI			HUPO BPP Protein Merger		0.0 kDa	0.0	PRIDE	Identified by peptide mass fingerprint	
									PRIDE	Search database protein sequence	MERRRITSAARRSYVSSGEMMVGGLAPGRRLGPGTRLSLA RMPPPLPTRVDFSLAGALNAGFKETRASERAEMMELNDRF ASYIEKVRFLEQQNKALAAELNQLRAKEPTKLADVYQAEL RELRLRLDQLTANSARLEVERDNLAQDLATVRQKLQDETN LRLEAENNLAAYRQEADEATLARLDLERKIESLEEEIRFL RKIHEEEVRELQEQLARQQVHVELDVAKPDLTAALKEIRT QYEAMASSNMHEAEEWYRSKFADLTDAAARNAELLRQAKH EANDYRRQLQSLTCDLESLRGTNESLERQMREQEERHVRE AASYQEALARLEEEGQSLKDEMARHLQEYQDLLNVKLALD IEIATYRKLLEGEENRITIPVQTFSNLQIRETSLDTKSVS EGHLKRNIVVKTVEMRDGEVIKESKQEHKDVM
IPI00025363 Peptide List (top ↑)		IPI			HUPO BPP Protein Merger		0.0 kDa	0.0	PRIDE	Identified by peptide mass fingerprint	
									PRIDE	Search database protein sequence	MERRRITSAARRSYVSSGEMMVGGLAPGRRLGPGTRLSLA RMPPPLPTRVDFSLAGALNAGFKETRASERAEMMELNDRF ASYIEKVRFLEQQNKALAAELNQLRAKEPTKLADVYQAEL RELRLRLDQLTANSARLEVERDNLAQDLATVRQKLQDETN LRLEAENNLAAYRQEADEATLARLDLERKIESLEEEIRFL RKIHEEEVRELQEQLARQQVHVELDVAKPDLTAALKEIRT QYEAMASSNMHEAEEWYRSKFADLTDAAARNAELLRQAKH EANDYRRQLQSLTCDLESLRGTNESLERQMREQEERHVRE AASYQEALARLEEEGQSLKDEMARHLQEYQDLLNVKLALD IEIATYRKLLEGEENRITIPVQTFSNLQIRETSLDTKSVS EGHLKRNIVVKTVEMRDGEVIKESKQEHKDVM

Links to spectra and peptides

PRIDE experiment comparison tool



Save Diagram

Intersectionset ABC:		
Identifications common to all three Experiments:		39
Similarity Score {A∩B∩C}/{A∪B∪C}:		0.05794948
Protein accession:	Protein version:	Database:
IPI00298497	3	IPI human
IPI00025252	1	IPI human
IPI00291175	2	IPI human

MS data analysis summary

- Spectra need processing: -background noise, peak finding, clustering etc.
- Peptide or protein ID: via spectra or sequence
  - Get list of proteins with scores
  - Validate
  - Need to take into account known issues with ID and DBs
- Issues related to complexity in samples
  - Solution: tandem MS
- Next - interpretation of results
  - What is the protein of interest
  - Does it have a structure, known domains
  - What about the collection of proteins?

Applications of proteomics

- Analysing protein expression
- Determining which proteins are present under different conditions or in different samples – comparative proteomics
- Biomarker discovery:
  - MS profiling