

PROTEOMICS

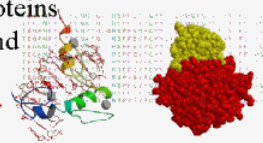
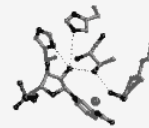
Proteome —
proteins
expressed by
a genome



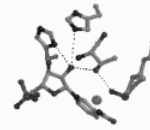
Presented by: Jan-Olov Höög and Jan Carlstedt-Duke,
Karolinska Institutet, Stockholm, Sweden

The tools of life

- All cellular processes in biological organisms involve proteins, the tools of life.
- Chemical reactions are catalyzed and controlled by enzymes, molecules are transported, signals are transduced, gene expression is controlled, and energy is harvested through the multifaceted activities of proteins.
- Information underlying all these activities is coded by the genomes.
- A cornerstone for functional genomics/proteomics is therefore the expression of the encoded proteins and the characterization of their functional and structural properties.



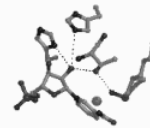
Some genomes completed



Organism	Genome size (Mb)	Number of genes	Year
<i>Mycoplasma genitalium</i>	0.58	470	1995
<i>Haemophilus influenzae</i>	1.8	1,743	1991
<i>Escherichia coli</i>	4.6	4,285	1996
<i>Saccharomyces cerevisiae</i>	12.1	6,266	1996
<i>Caenorhabditis elegans</i>	97	19,000	1998
<i>Drosophila melanogaster</i>	180	13,600	2000
<i>Homo sapiens</i>	3,000	40,000	2001



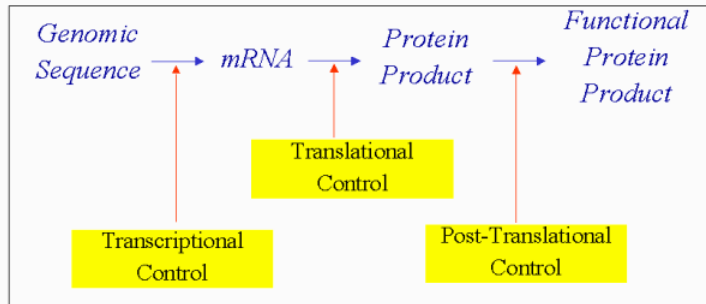
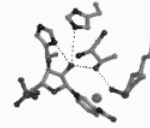
Proteomics – usually Analysis of a large number of proteins with a combination of 2D-gels and masspectrometry



Today it includes both the indentification of a large number of expressed proteins by a genome as well as the characterization of functional and structural relationship



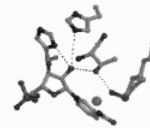
Protein production pathway



mRNA level \neq expressed protein level nor does it indicate the nature of the functional protein product



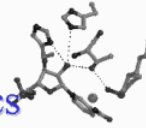
Large scale protein analysis



- The **Proteome** is the protein complement of the genome
- **Proteomics** is the use of qualitative and quantitative protein level measurements to characterize biological processes (e.g. diseases)
- **Diseases** are treated at the protein level
- Proteins are **drug** targets



Challenges: Genomics vs Proteomics



DNA

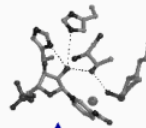
- Static
- Can be amplified
- Little complexity: Single component
- Good solubility characteristics

Protein

- Very dynamic
- Can not be amplified
- Post-translationally modified (very complex)
- Variable solubility



Key technologies in proteomics



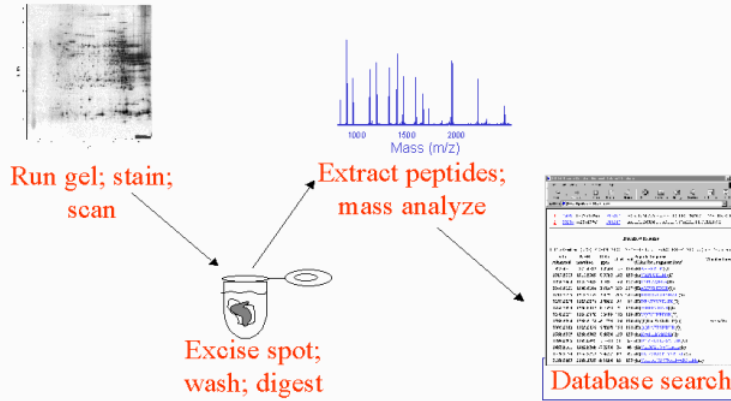
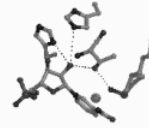
- Reproducible 2D Gel Technology
- Staining and Scanning Technology
- **Mass Spectrometry for Identification**
- Databases (protein and genome)
- Database Searching Algorithms

Expanding
Automation

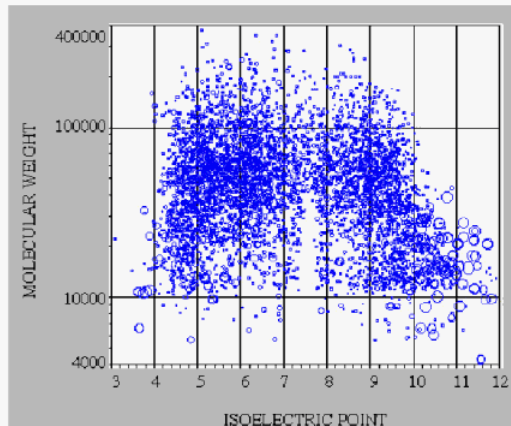
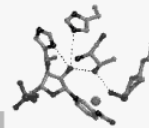
The key challenge in Proteome research is the automation and integration of these technologies



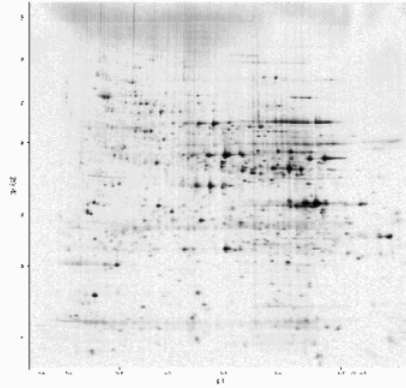
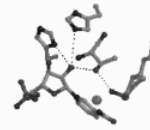
Experimental approach



Theoretical 2D gel of yeasts



Experimental 2D gel of yeasts

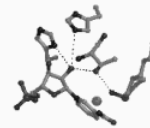


From EXPASY (<http://www.expasy.ch>)

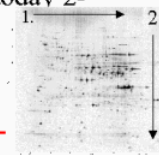


Key technologies in proteomics

Reproducible 2D gels (1)

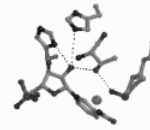


- Electrophoresis is today the central technique for the analysis of a large number of proteins at the same time.
- The methodology as such was invented by Tiselius almost half a century ago. The first "2-D gel electrophoresis" was published in 1956 by Smithies and Poulik describing the use of paper and starch gel electrophoresis. 1975 O'Farrell optimised the 2-D separation procedure that is the basis for today 2-D PAGE.

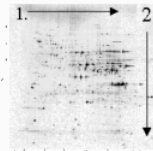


Key technologies in proteomics

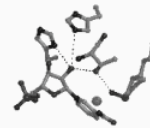
Reproducible 2D gels (2)



- First gel dimension – Isoelectric focusing (IEF) with immobilised pH gradients (IPG)
- Second gel dimension – SDS polyacrylamide gel electrophoresis (PAGE)
- In both cases precast gels of high quality is a basic requirement.

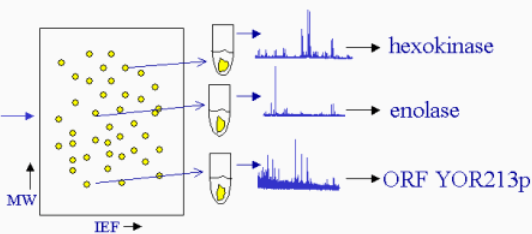


Need for scalable proteomics



Samples → **Separation** → **Digestion** → **Identification**

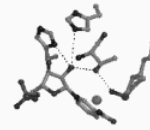
- normal samples
- diseased samples
- serum, CSF...
- ± drug
- time-course



1000s of samples = 1000s of gels → 1000s of proteins/gel → 10^6 - 10^{12} proteins



Sample preparation



Pretreatment of samples before running gel

Solubilisation \Rightarrow **Denaturation** \Rightarrow **Reduction**

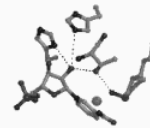
- Non-protein components have to be removed, e.g. nucleic acids.
- Typical solubilisation buffer: 8 M urea, 4% CHAPS, 50 mM DTT and 40 mM Tris.



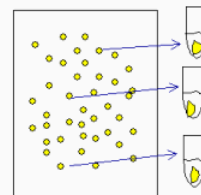
This step is very often the most important for a successful result.



In-gel digestion



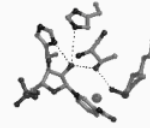
- Dice each gel slice into small pieces and add 25mM NH_4HCO_3 /50% ACN
- Extract the supernatant and transfer to a separate tube (to be discarded).
- Repeat.
- Speed Vac the gel pieces to complete dryness.
- Add trypsin solution (trypsin in 25mM NH_4HCO_3) and incubate at 4°C for 30 min.
- Spin briefly and incubate at 37°C overnight.



<http://donatello.ucsf.edu/ingel.html>



Identification of peptide fragments with masspectrometry



- Direct MALDI-TOF MS¹ of protein immobilised on membrane or in gel matrix (low yield from directly eluted proteins)
- MALDI or ESI-MS² analysis of peptides following digestion of proteins from gels or immobilised on membranes

¹Matrix assisted laser desorption/ionisation - time of flight masspectrometry

²Electrospray ionisation-masspectrometry



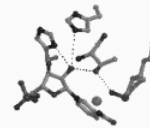
Proteome diversity

gene

```
ATGGGCACCCAGGAAAAATTATTAATGCAAGGCAGCCATTGCCTGGAAAACTGGCAGT
CCOCTTTGCATTSAGAAAAATTAAGGTGTCCOCTCTAAGGCTTGTAAAGGTGGGATTCAG
GTAAATCGCCACGTGTGTGTGCCCTACTGACATCAATGCCACCGATCTTAAGAAAGAAAGCT
CTCTTCOCGGTAGTCTTGGTCCATGAGGTGTGCAAGGAATTGTAGAAAAGTGTGGGCCCGGGA
GTGACCAACTTCAAAACAGGTGACAAAGTAATCCATTCTTTGACCCACAGTGCAAAAGG
```

predicted protein

```
MetGlyThrGlnGlyLysValIleLysCysLysAlaAlaIleAlaTrpLysThrGlySer
ProLeuCysIleGluLysIleLysValSerProProLysAlaCysLysValArgIleGln
ValIleAlaThrCysValCysProThrAspIleAsnAlaThrAspProLysLysLysAla
LeuPheProValIleLysGlyHisGluCysAlaGlyIleValGluSerValGlyProGly
ValThrAsnPheLysProGlyAspLysValIleProPhePheAlaProGlnCysLysArg
```



N-terminal truncation



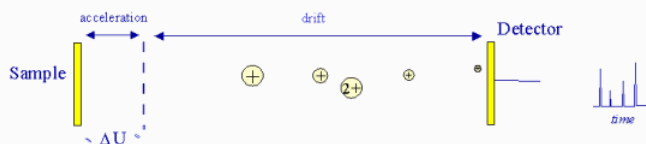
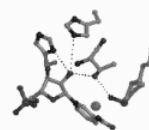
Modified gene product

co- and post-translational modifications of the proteins

From one gene sequence to one predicted protein sequence to many protein products, the latter to be identified with masspectrometry



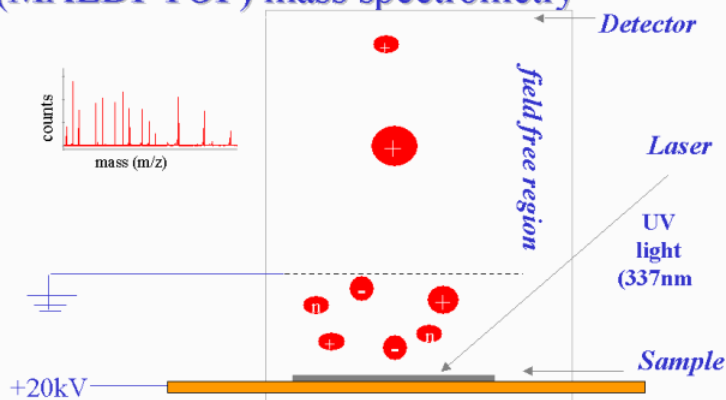
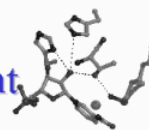
The Time-of-Flight (TOF) method



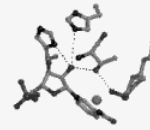
$$TOF = A (m/z)^{1/2} + B$$



Matrix Assisted Laser Desorption/Ionization Time-of Flight (MALDI-TOF) mass spectrometry



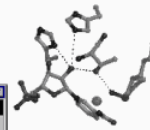
Database searching



A peptide-mass fingerprinting tool from the [UCSC Mass Spectrometry Facility](http://donatello.ucsf.edu/) that tries to fit a user's mass spectrometry data to a protein sequence in an existing database and thus suggest the identity of the user's protein.

<http://donatello.ucsf.edu/>

Database searching: hit list



MS-Index	# (%)	Database	SwissProt Accession #	Species	MW	Protein Name
1	11300 (93%)		R09914	YEAST	46671.1	ENOLASE 1 (EC 4.2.1.11) (2-FRUCTOSE-BISPHOSPHATE DEHYDROGENASE) (G-FCB3750-D-GLYCERATE HYDROLYTASE) (c567.1 Da)
2	11302 (93%)		R09912	YEAST	46783.2	ENOLASE 2 (EC 4.2.1.11) (2-FRUCTOSE-BISPHOSPHATE DEHYDROGENASE) (G-FCB3750-D-GLYCERATE HYDROLYTASE) (c567.1 Da)

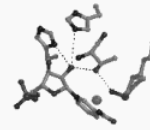
Detailed Results

1: 107.6 matches (93%). YEAST ENOLASE 1 (EC 4.2.1.11) (2-FRUCTOSE-BISPHOSPHATE DEHYDROGENASE) (G-FCB3750-D-GLYCERATE HYDROLYTASE) (c567.1 Da)

Index	MW	Delta	start end	Peptide Sequence	Modifications
726.3620	726.3622	17.2003	9 14	(R)SYVDSR(G)	
756.4800	756.4752	8.9919	409 414	(K)EYQLLEQ	
807.4560	807.4365	14.1777	178 184	(K)LFPAALQ	
1159.6110	1159.6111	-6.1276	185 194	(R)GSEVYHDLQ	
1286.7040	1286.7109	-6.1102	67 78	(G)NVDYVAPFVQ(A)	
1412.8380	1412.8225	16.9442	105 119	(K)LGANALGVSLAASR(A)	
1579.7990	1578.8015	-1.6910	89 102	(K)AVDDFLSLDUTANQ(S)	
1627.9630	1627.9495	8.2746	103 119	(K)KLGANALGVSLAASR(A)	
1755.8630	1755.8190	25.0774	255 269	(K)DGKYLDFKPNRDK(S)	
1755.8630	1755.9493	49.1214	312 328	(K)TGRVIVADLLVYNEK(R)	

<http://falcon.ludwig.ucl.ac.uk/ucshtml3.2/msfit.htm>

Mass accuracy in database searching

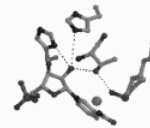


<i>Molecular ion</i>	<i>Allowed error (ppm)</i>	<i># Matched Proteins</i>	
		<i>All Species</i>	<i>Yeast</i>
1975.954	1000	754	173
	100	200	76
	10	44	6
1526.884	1000	982	120
	100	197	27
	10	19	2
1055.541	1000	1300	173
	100	514	76
	10	87	6

* SWISSPROT r.33, total proteins all species= 52205,
yeast = 3653



Mass accuracy in database searching



<i># of peptides</i>	<i>error (ppm)</i>	<i># Matched Proteins</i>	
		<i>All Species</i>	<i>Yeast</i>
2 of 3	1000	149	20
3 of 3		2	1
2 of 3	100	10	1
3 of 3		1	1
2 of 3	10	2	1
3 of 3		1	1

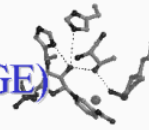
Masses searched = 1288.7113, 1420.7549, 2447.0978

* SWISSPROT r.33, total proteins = 52205



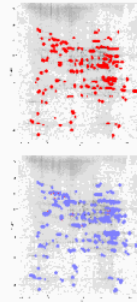
Developments

Difference gel electrophoresis (DIGE)



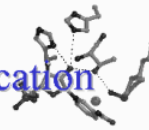
Dye-labelled proteins co-separated on the same gel

- First protein extract is labelled with e.g. Cy2 and second protein extract is labelled with Cy3
- Co-separation by 2-D PAGE
- Image gel at two different wavelengths
- Differential analysis
- Inducible proteins can be detected with only one gel run



Developments

Multidimensional protein identification technology (MudPIT)



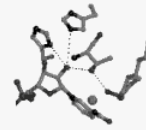
- Multidimensional liquid chromatography
- Tandem mass spectrometry
- Database searching

MudPIT has been applied to the proteome of the *Saccharomyces cerevisiae* and yielded the largest proteome analysis to date. A total of 1,484 proteins were detected and identified. Categorization of these hits demonstrated the ability to detect and identify proteins rarely seen in proteome analysis, including low-abundance proteins like transcription factors and protein kinases.







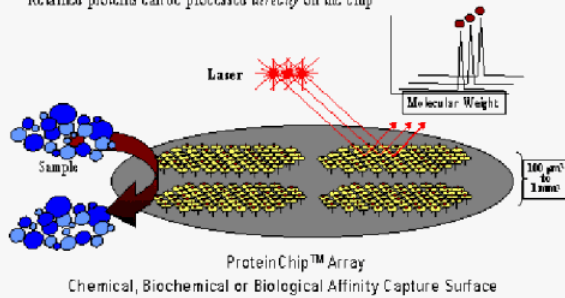
Developments

Protein-array chips (1)



The SELDI Process and ProteinChip™ Arrays

- Sample  goes *directly* onto the ProteinChip™ Array 
- Proteins  are captured, retained and purified directly on the chip (affinity capture) 
- Retentate Mag™ is "read" by Surface-Enhanced Laser Desorption/Ionization (SELDI)
- Retained proteins can be processed *directly* on the chip

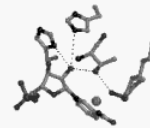


• <http://www.ciphergen.com/> -

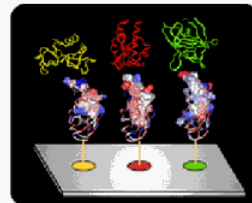


Developments

Protein-array chips (2)



The HIP™ chip, the first high-throughput addressable array of capture proteins, which will revolutionize drug discovery and functional proteomics. With this chip it is possible to rapidly assess the protein expression profile of a biological sample so as to measure the relative level of hundreds and ultimately thousands of different proteins simultaneously. Differential analysis of these profiles can then be used to identify the proteins in human tissues that are affected as a result of a particular disease or in response to a specific therapeutic drug.

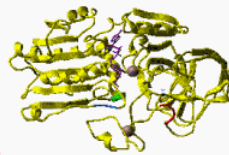
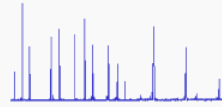


• <http://www.phylos.com/> -



Useful links for proteomics research

- <http://base-peak.wiley.co.uk/links/>
- <http://donatello.ucsf.edu/>
- <http://prospector.ucsf.edu/>
- <http://ai.sri.com/ecocyc/ecocyc.html>
- <http://www.ebi.ac.uk/>
- <http://www.embnet.org/>
- <http://www.expasy.ch/>
- <http://www.hon.ch/>
- <http://www.pdb.bnl.gov/>
- <http://www.protocol-online.net/index.htm>
- <http://www.nature.com/genomics/>



To obtain the complete picture of Proteomics/Functional genomics

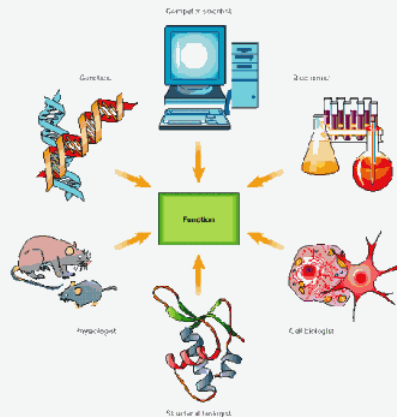


Figure 2.10. The function of genes is determined by the products they encode. The function of genes is determined by the products they encode. The function of genes is determined by the products they encode.