

# Proteomes and Proteomics: "P" words in the world of Functional Genomics

Dr Marc Wilkins



## Genes, Genomes and Gene Products

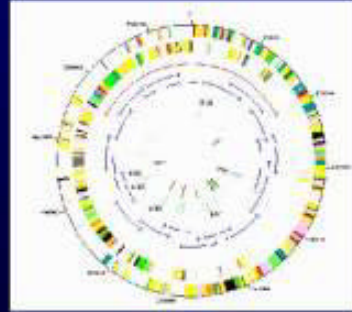
- availability of genomic sequences has changed biology forever

- genomes now available for many dozens of organisms

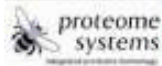
e.g. *Escherichia coli*  
*Bacillus subtilis*  
*Saccharomyces cerevisiae*  
*Drosophila melanogaster*  
*Homo sapiens*

- genomes define the informational content of an organism

- however, a genome sequence does not tell you how an organism works



*Mycoplasma genitalium* genome map  
Fraser et al, 1995



- study of proteins, the functional molecules, is essential

## What is a Proteome?

- proteome = PROTEin complement expressed by a genOME or tissue  
(Wilkins *et al.* 1995 Biotechnology and Genetic Engineering Reviews 13, 19-50)
- proteomes are dynamic
- proteomes change as a function of:
  - time
  - development
  - extracellular conditions
  - intracellular conditions



*Dictyostelium discoideum* development

## What is Proteomics?

- proteomics is the study of proteomes
- proteomics aims to:
  - separate, identify and characterise proteins on a large scale
  - define levels of proteins in cells / tissues and how these change
  - investigate protein complexes
  - elucidate protein functions, pathways, and interrelationships

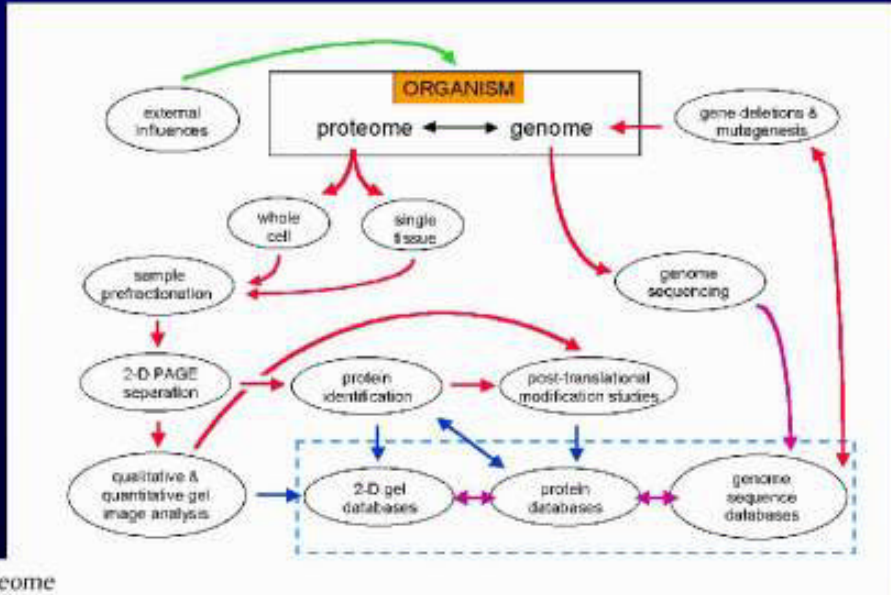


*Dictyostelium discoideum* development



*D. discoideum* slugs after 2-D PAGE

# Proteomics, Databases and Things In-Between



## *Major Issues in Proteomics*

- how can you separate and visualise the proteins in a proteome?
- how can this be used to study protein complexes and pathways?
- how can you quickly identify separated proteins?
- how can you characterise proteins in detail?
  
- why do companies want to spend millions on proteomics?

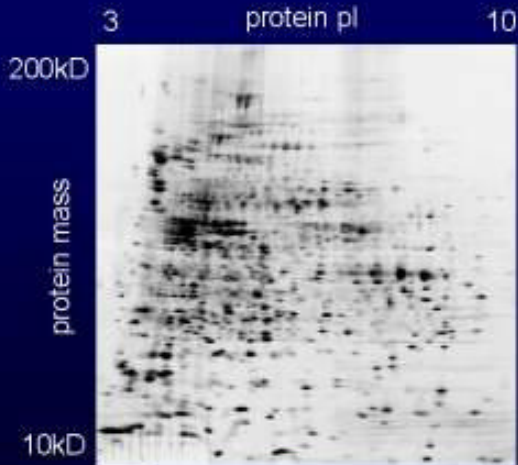
## *Traditional Protein Chemistry*



- column chromatography  
(e.g. ion exchange or reversed phase)
- one protein at a time
- purification may take months



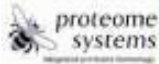
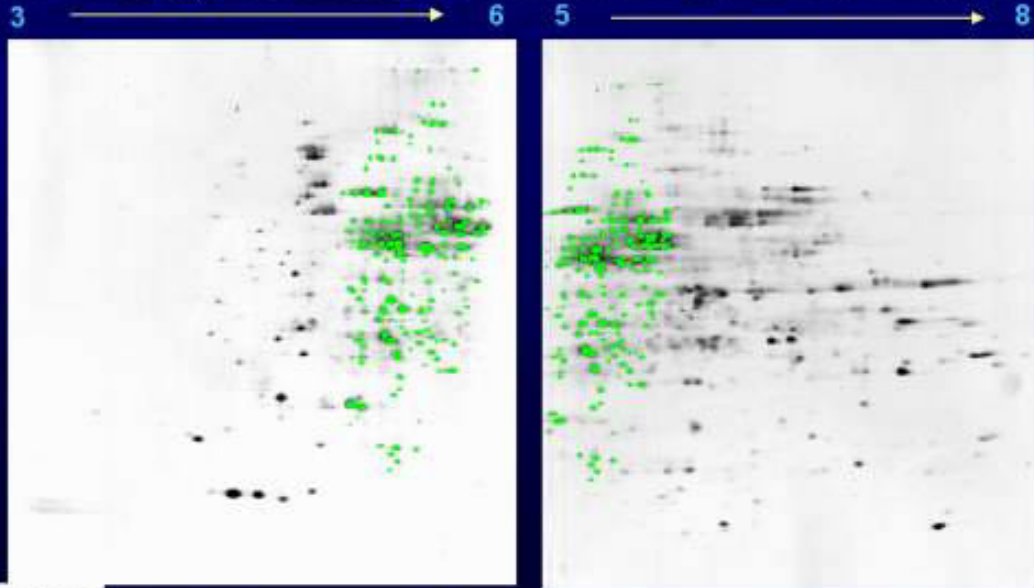
## *Proteomics Array: two dimensional gel electrophoresis*



- first dimension charge-based separation
- second dimension mass-based separation
- up to thousands of proteins purified at once (e.g. 6000)
- proteins purified in parallel in 1-3 days
- image is reference map for cell, tissue or protein complex

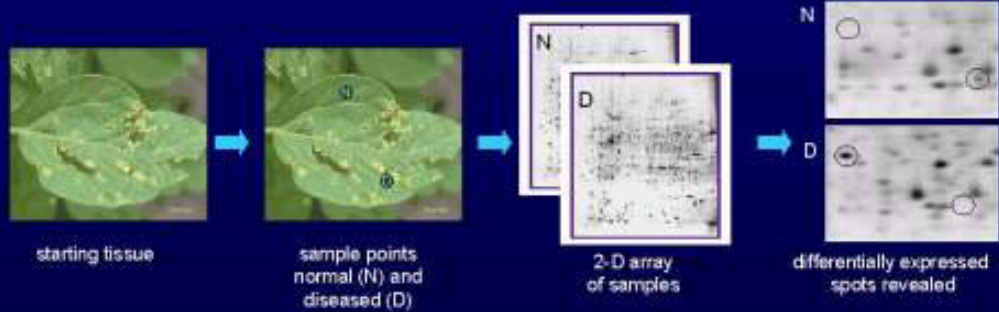


## Narrow Range Gels Increase Resolution and can be Matched



*E. coli* - silver stained, matched by Melanie II software

## Discovery of Differentially Expressed Proteins



Differential display can reveal:

- changes in protein expression
- altered processing or modification of proteins

These changes may be responsible for a phenotype.

Co-regulation or co-modification of proteins can provide clues to pathways and function

## *Protein Identification & Characterisation*

Proteomics requires:

- high throughput protein identification
- characterisation of modifications (e.g. phosphorylation)

Previously, proteins have been identified with:

Edman sequencing, antibodies  
amino acid composition, co-migration

Now, proteins are identified using mass spectrometry and:

peptide mass fingerprinting  
peptide fragmentation

## Mass Spectrometry of Proteins

Mass spectrometers precisely measure the mass of molecules

Mass spectrometers have two parts:

- 1) ion source
- 2) measuring apparatus

Common ion sources for protein analysis are:

- electrospray ionisation (ESI)
- matrix assisted laser desorption/ionisation (MALDI)

Measuring apparatus these are teamed with are:

- quadrupoles
- time of flight (TOF) detectors
- ion traps
- combinations of the above

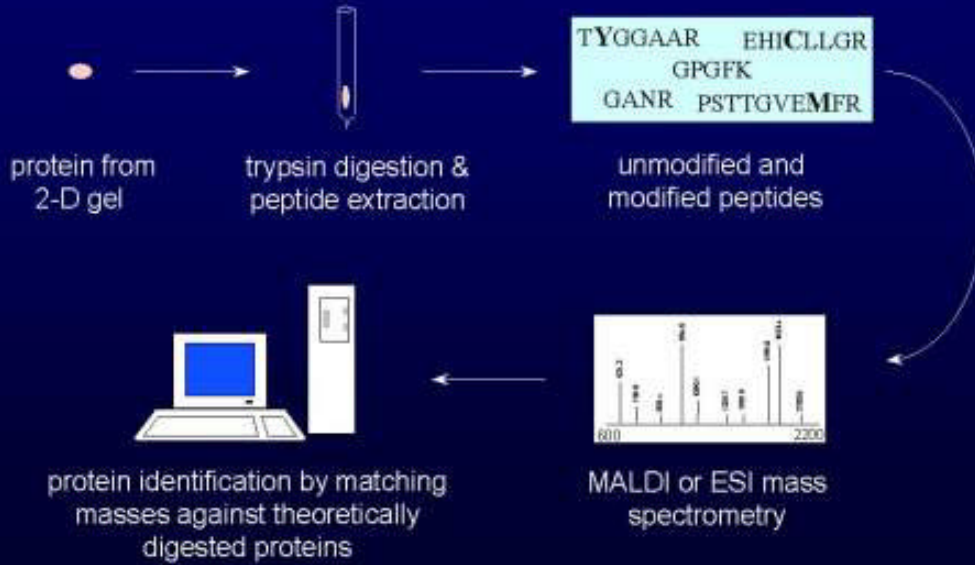
Common platforms:

- MALDI-TOF, ESI-TOF, ESI / ion trap, (triple) quadrupole TOF

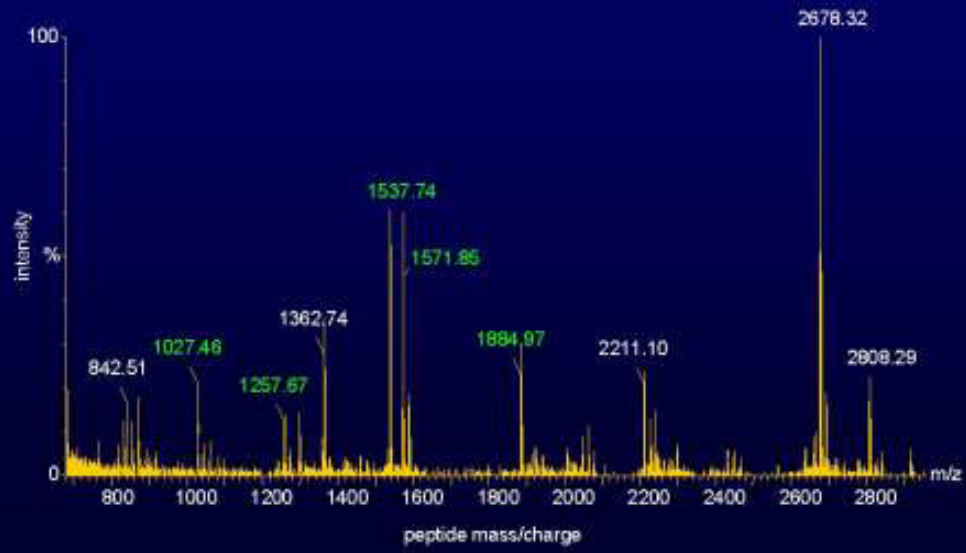


Bruker Biflex III

## Protein Identification by Peptide Mass Fingerprinting



## MALDI-TOF Spectra for *E. coli* Protein from 2-D gel



## E. coli Protein is TRP Repressor Binding Protein

Score: 0.71, 5 matching peptides: [P30849](#) (WRBA\_ECOLI) pI: 5.60, Mw: 20714.36  
TRP REPRESSOR BINDING PROTEIN - Escherichia coli

user mass	matching mass	$\Delta$ mass (Dalton)	#MC	modification	position	peptide	links
1884.94	1884.949	0.009	0		88-105	TFLDQTGGLWASGAL YGE	<a href="#">FindMod</a>
1537.72	1537.724	0.0041	0		155-171	GGIFYGATTIAGGDG SR	<a href="#">PeptideMass</a>
1571.85	1571.825	-0.0249	1		37-49	RVPETMPPQLFEK	<a href="#">BioGraph</a>
1257.64	1257.643	0.0033	0		172-182	QPSQEELSIAR	
1027.45	1027.445	-0.0051	0		79-87	EGNMSGQMB	

34.5% of sequence covered.

```

      1          11          21          31          41          51
      |          |          |          |          |
1  akviviyyw yghietvra vaegaslvdg aevvleRVE TFPPQLFEK ggigtapva 60
61 tpqeladya iifgptrFG NMSGQRTFL DQTGGLWANG ALVSKlavf astgtggqe 120
121 qtitstottl ahbgeivps gyaaelidv sqvrGGTPYD ATTIAGGDG QPSQEELSI 180
181 ARyqgeyva lsvkng
    
```

Matched against ~90,000 proteins in SWISS-PROT with PeptIdent  
<http://expasy.proteome.org.au/tools/peptident.html>



## *Automation of Protein Identification*

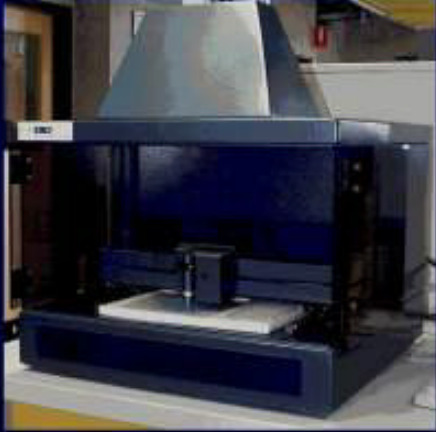


384 well plates

High-throughput labs now use automated:

- spot excision
- trypsin digestion
- MALDI target loading
- spectra acquisition
- calibration
- peak picking
- database matching

## *ARRM 214 Protein Excision Robot - Prototype*



- excision from gel or PVDF



- transfer to 96 well plates

## *Protein Excision Robot - Production Model*



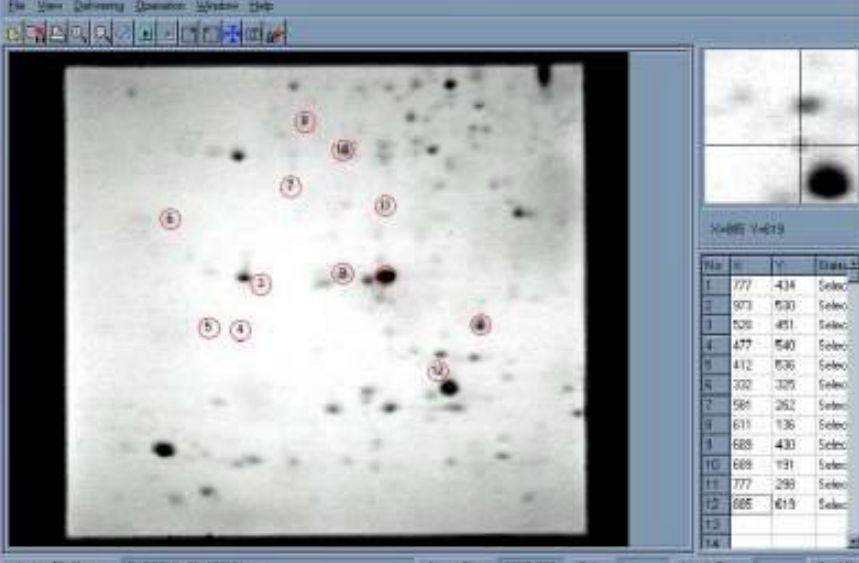
- developed by Proteome Systems, APAF and ARRM Biotech
- marketed internationally by Bio-Rad

## Excision Robot - Gel Imaging



- whole gel images are shown, mouse guides the zoom window

## Excision Robot - Spot Selection

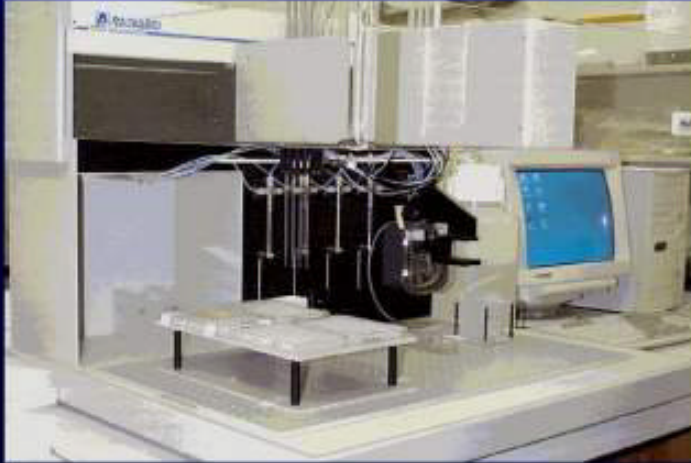


The screenshot shows the software interface for the Excision Robot. The main window displays a grayscale image of a spot array with 12 spots circled in red and numbered 1 through 12. A smaller inset window shows a magnified view of the selected spot. A table on the right lists the coordinates and status of the selected spots.

File	X	Y	Status
1	777	434	Selec
2	973	630	Selec
3	508	451	Selec
4	477	540	Selec
5	412	636	Selec
6	332	325	Selec
7	586	262	Selec
8	611	136	Selec
9	688	430	Selec
10	688	131	Selec
11	777	298	Selec
12	685	619	Selec
13			
14			

- spots selected by point & click, co-ordinates are logged and exported

## Automated Liquid Handling & Delivery



Canberra Packard MultiProbe 104



↓  
trypsin digestion



## Automated Mass Spectrometry

### *Proteomics requires automated*

- spectra acquisition
- calibration
- (deconvolution)
- peak picking
- data export



Voyager DE-STR

### *MALDI-TOF MS offers*

- high throughput (500+ per day)
- high degree of automation

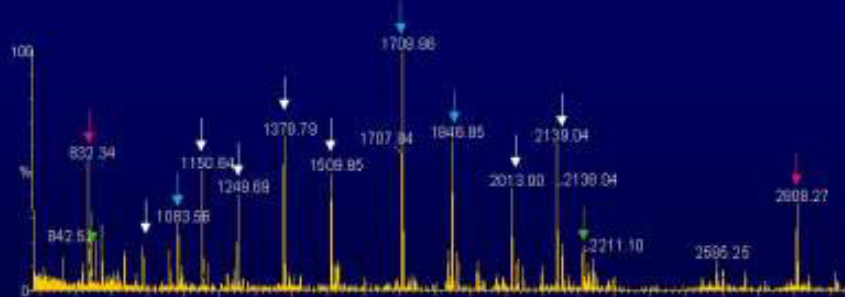


Micromass LCT

*ESI-TOF* machines currently slower & less automated



## Mass Spectra Require Interpretation

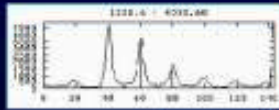
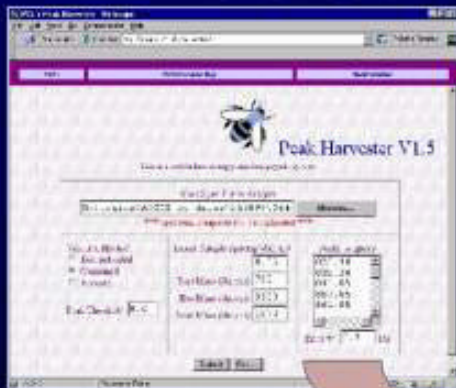


2 calibration peaks, > 2 junk peaks  
7 hits with P02339, 3 large unmatched peaks - modifications  
lots of noise

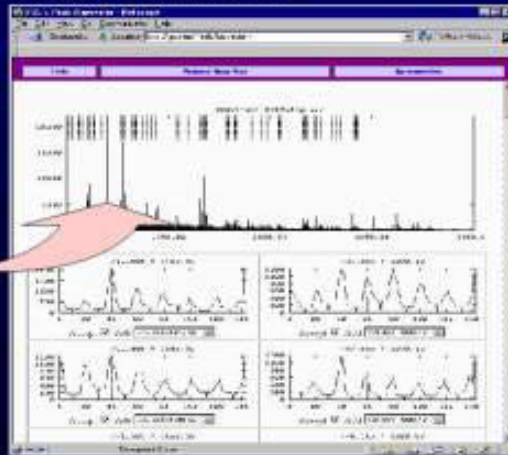
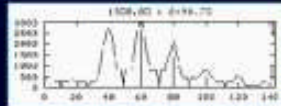
Can interpretation and database matching be done automatically?

## Advanced Peak-Picking Module

- "best spectrum" selection
- image analysis filters, poisson isotopic modelling
- fuzzy logic deisotoping and deamidation detection
- interlaced peak detection
- cross-platform applicability



Poisson isotopic modelling



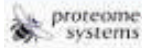
# Integrated Protein Identification Tools

- peptide mass fingerprinting engine
- queries proprietary sequence databases
- integrated with other in-house tools

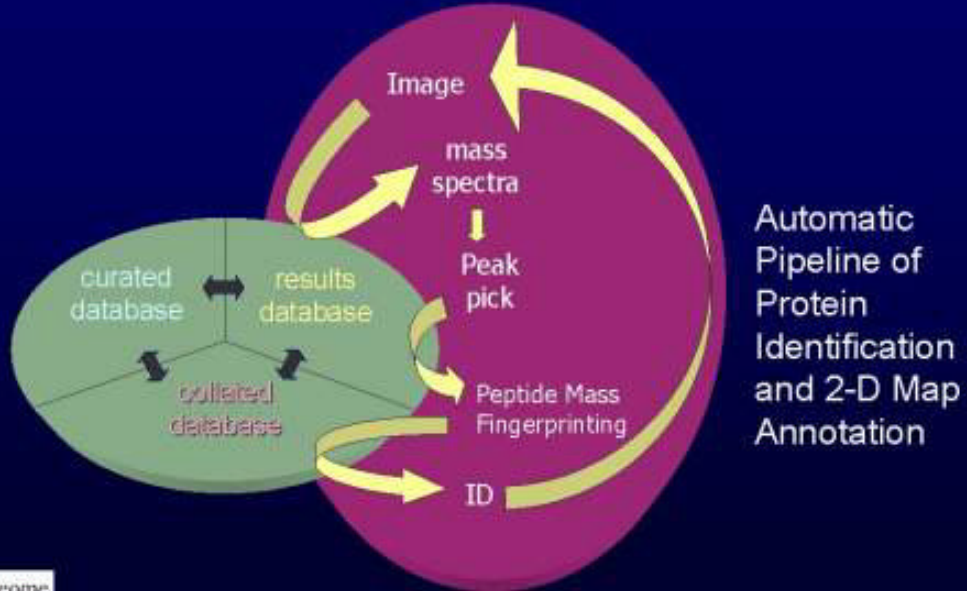
The screenshot displays the Proteome Systems software interface, which is used for protein identification. It features several windows and panels:

- Left Panel:** A list of mass values (e.g., 100.07, 116.08, 129.09) and a search criteria section.
- Search Results Window:** Displays the results of a search, including a list of proteins with their accession numbers and scores. The top result is **ADENOSINE DEAMINASE** (P00931) with a score of 7.9.
- Right Panel:** Shows a table of protein identification results, including the protein name, accession number, and score. The top result is **ADENOSINE DEAMINASE** (P00931) with a score of 7.9.

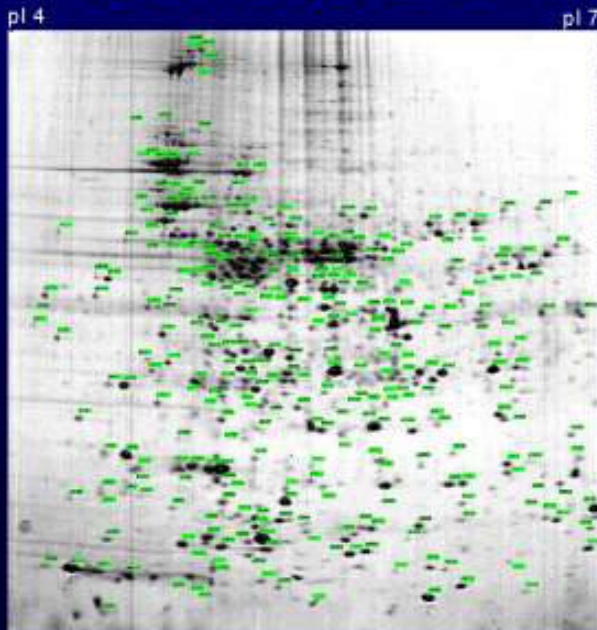
The Proteome Systems logo is visible in the bottom left corner of the interface.



# Pipelining Tools Automate Analyses

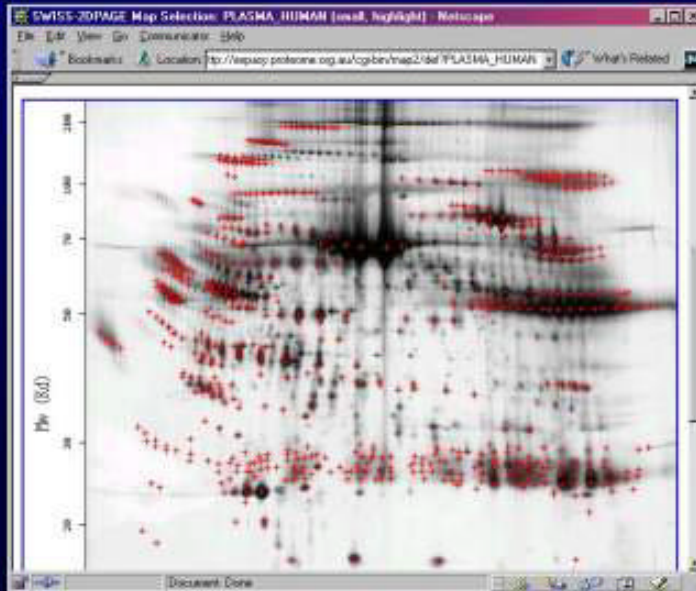


## 2-D Array of *E. coli* Whole Cell Lysate



- 1069 analyses from 5 gels
- 587 (55%) spots identified
- 55 (5%) putatively identified
- 324 spots ID on 3mg master
- 192 gene products
- All steps automated but for reference image annotation
- This project completed in 8 weeks with 1.5 people (including development time)

## Annotated Reference Maps on the Web



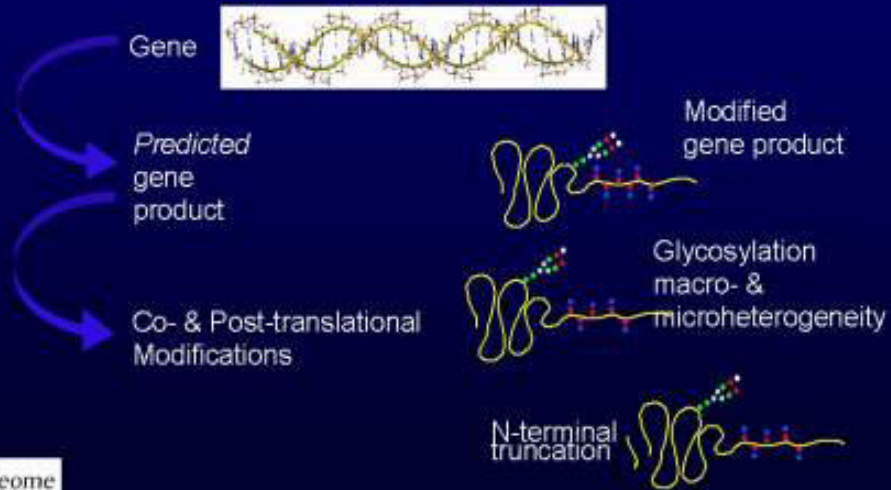
E.g. human plasma  
from SWISS-2DPAGE

[www.expasy.ch](http://www.expasy.ch)



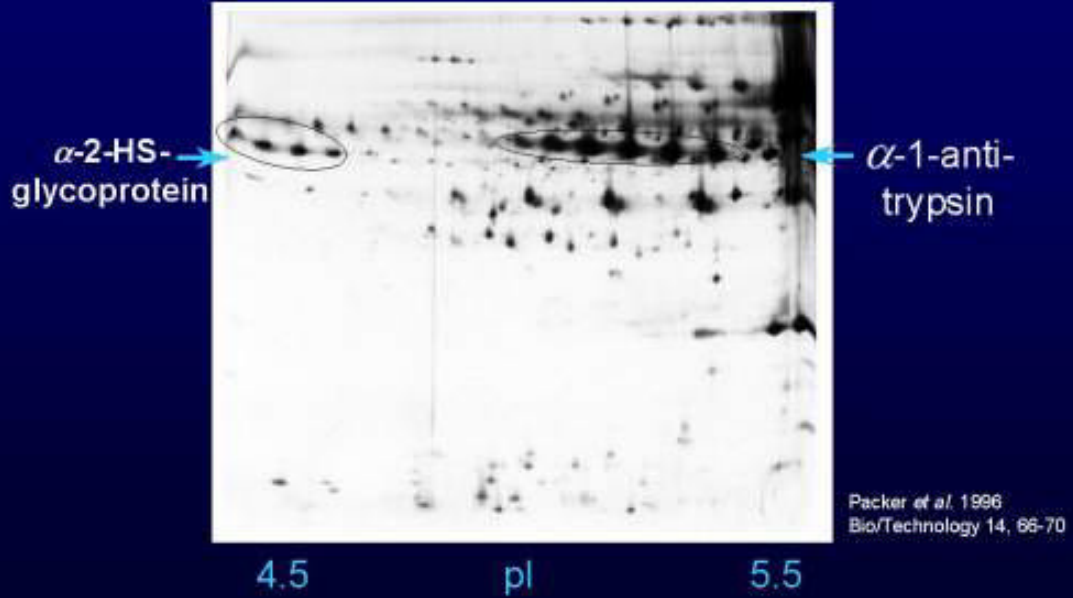
# One Gene can give Many Proteins

This is NOT predictable from genomic sequences!

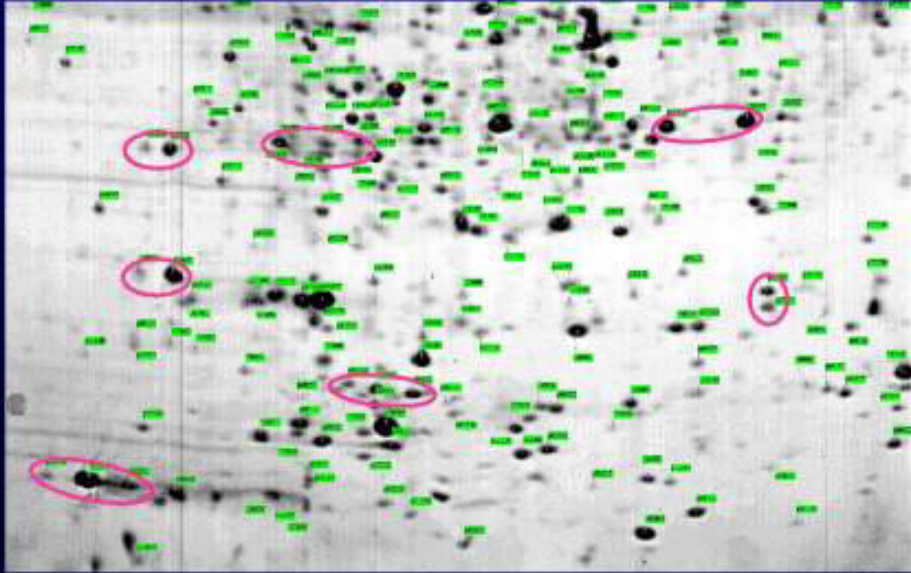




## Isoforms in 2-D PAGE of Human Plasma



## Isoforms in *E. Coli* 2-D Reference Map



3mg of *E. coli* lysate, coomassie stained, proteins identified by MS

## *Studying Protein Modifications by Mass Spectrometry*

Isoforms on gels are due to protein processing or modifications

Most modifications change protein or peptide mass  
e.g. methylation changes mass by 14 Da

Some modifications change protein charge as well as mass  
e.g. phosphorylation makes protein more acidic

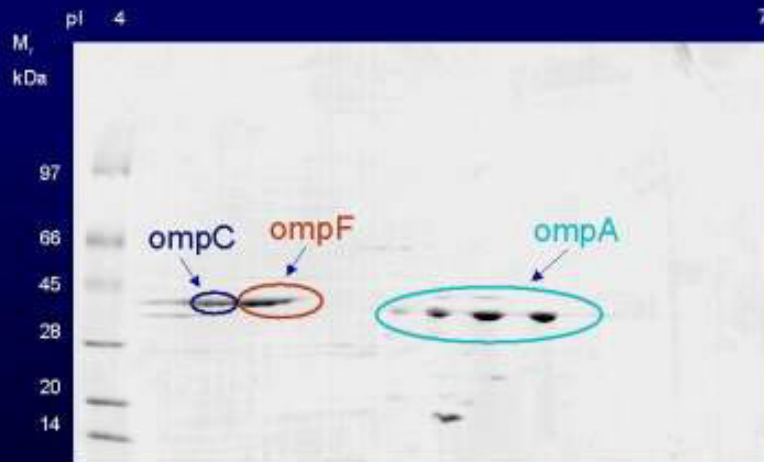
Many modifications can be studied with mass spectrometry.  
Sometimes, this can be done with peptide mass fingerprinting.  
But usually this requires fragmenting peptides using MS-MS.

## *Masses of Some Protein Modifications*

Modification Type	Mass Change
Acetylation	42
Amidation	-1
Deamidation	1
Formylation	28
Methylation	14
Phosphorylation	79
Sulfation	80
Glycosylation	162-3000

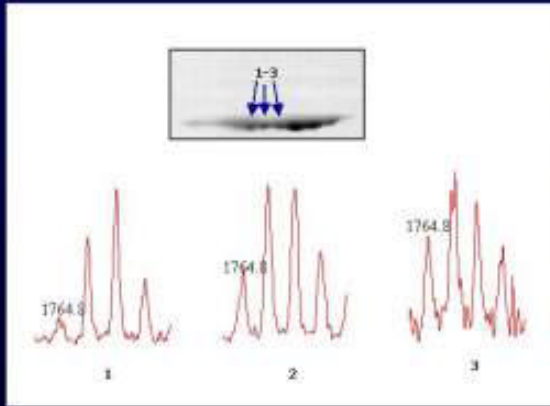
## *E. coli* Membrane Proteins on Mini 2-D PAGE

Why are there Isoforms?



solubilisation solution = 7M urea, 2M thiourea, 1% ASB 14, 2mM TBP, 40mM Tris base, 0.5% CA  
coomassie stained gel

## Deamidation of 1764.8 in ompC Causes Isoforms



peptide sequence    protein pI

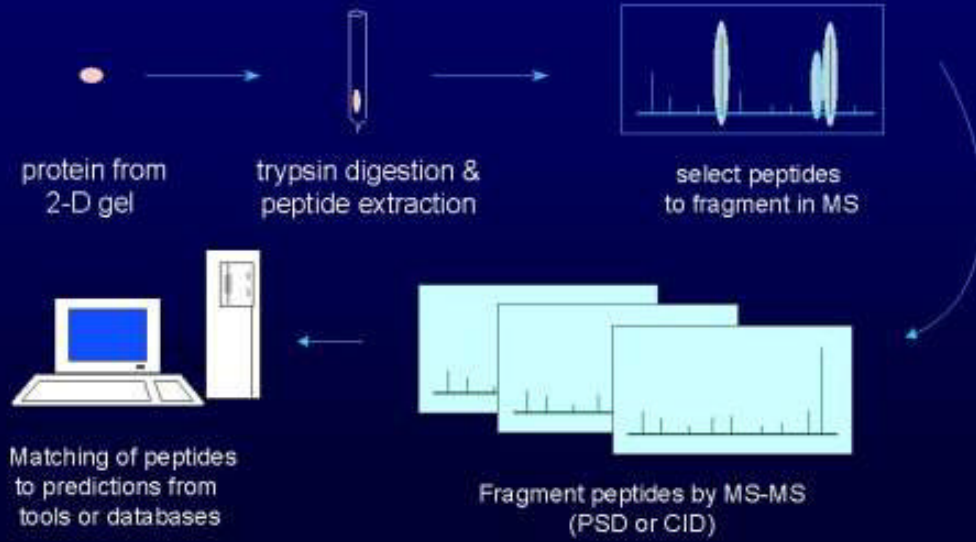
NGNPSGEGFTSGVTNGR    4.48

DGNPSGEGFTSGVTNGR  
or  
NGNPSGEGFTSGVTDGR } 4.45 (-0.03)

DGNPSGEGFTSGVTDGR    4.4 (-0.08)

**N** to **D** deamidation  
causes 1 Da increase  
in peptide mass

## Peptide Fragmentation & Characterisation by MS-MS





## *Peptide Fragmentation by MS-MS: peptides fragment in predictable ways*

Peptides usually fragment at the amide bonds between amino acids

e.g. peptide of sequence DGHYSR gives fragments:

	GHYSR	HYSR	YSR	SR	S	(y ions)
as well as						
	DGHYS	DGHY	DGH	DG	D	(b ions)

Because masses of each peptide fragment can be calculated, peptides can be "sequenced"

Modifications can also be localised to a single amino acid.

## FindMod Output - Application of Rules

- potentially modified peptides that agree with rules are listed
- amino acids that potentially carry modifications are shown

Potentially modified peptides, detected by mass difference and conforming to rules (considering only peptide masses that have not matched above):									
User mass	DB mass	mass diff.	mod. diff.	$\Delta$ mass	potential mod.	#MC	peptide	position	known modifications
1631.81	1603.771	28.039	28.031	-0.007	<a href="#">DIMETH</a>	1	AFDQIDNAPEEKAR	45-58	
1631.81	1617.787	14.023	14.016	-0.006	<a href="#">METH</a>	1	AFDQIDNAPEEKAR	45-58	(METH: 56)

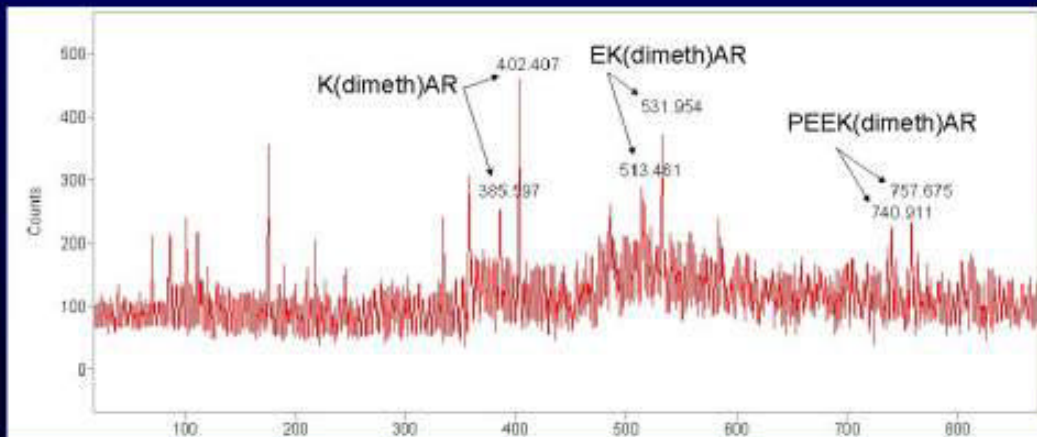
- peptides potentially modified only by mass difference

Potential PTMs detected by mass differences, but not confirmed by rules:									
User mass	DB mass	mass diff.	mod. diff.	$\Delta$ mass	potential mod.	#MC	peptide	position	known modifications
1631.81	1603.771	28.039	27.995	-0.043	<a href="#">FORM</a>	1	AFDQIDNAPEEKAR	45-58	
1631.81	1632.736	-0.925	-0.983	-0.057	<a href="#">AMID</a>	1	ETQKSTCTGVEMFR	249-262	(1xMSO)

- predictions can be tested by MS-MS peptide fragmentation

## Testing FindMod Predictions by MS-MS

- PSD fragmentation of 1631.8 verified dimethylated lysine in peptide AFDQIDNAPEEK(dimeth)AR



## *Why Spend Millions on Proteomics?*

### Proteomics will:

- define the proteome of a cell or tissue
- provide means of comparing proteomes to explain phenotypes (e.g. disease vs normal states)
- provide clues to protein function by defining co-stimulated and co-regulated proteins
- be powerful in combination with other technologies such as two-hybrid functional assays and gene knockout

### Proteomics will not:

- replace genome sequencing
- be as easy as genome sequencing

## References and Further Reading



### General references on proteomics:

Proteome Research: new frontiers in functional genomics, 1997. Eds. Wilkins et al. Springer.

Blackstock WP, Weir MP.  
Proteomics: quantitative and physical mapping of cellular proteins.  
Trends Biotechnol. 1999, 17:121-7.

Anderson NL, Anderson NG.  
Proteome and proteomics: new technologies, new concepts, and new words.  
Electrophoresis. 1998 19:1853-61.