

Proteomics and some of its Mass Spectrometric Applications

Research Seminar on Data Analysis for Bioinformatics

Reetta Nylund
April 1st, 2005

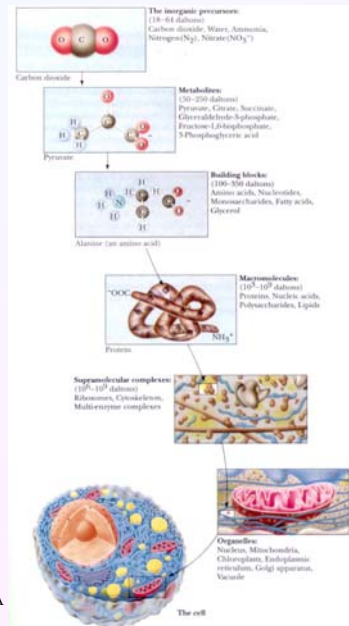
What?

Large scale screening of proteins, their
expression, modifications and interactions
by using high-throughput approaches

Why?

The number of found genes (ORFs) in human genome is much less than the number of expressed proteins, and therefore, more information is needed about the “working units” of the cells

From Garrett & Grisham, Biochemistry, 1999, Saunders College Publishing, USA



April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

3

Reetta Nylund

How?

- Proteomics “branches”:
 - Proteomic analysis (analytical protein chemistry)
 - Characterization of proteins and their post-translational modifications
 - Expression proteomics (differential display proteomics)
 - Profiling of expressed proteins using quantitative methods
 - Cell-mapping proteomics (cataloging of protein-protein interactions)
 - Identification of protein complexes
- Used approaches:
 - Gel-based proteomics
 - Mass spectrometry driven proteomics
 - Protein arrays, Yeast two-hybrid arrays,... etc.

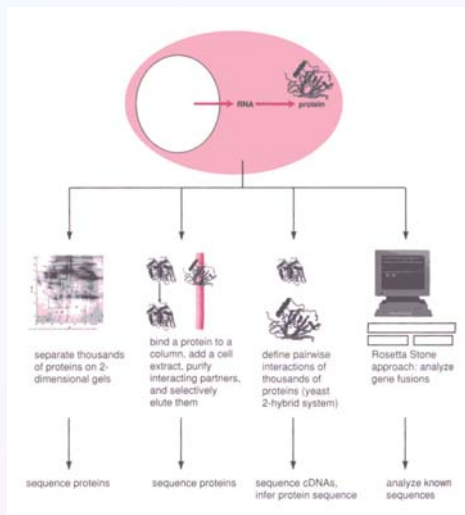
April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

4

Reetta Nylund

Proteomics approaches



From Pevsner, 2003, p.248
Approaches to high-throughput
protein analysis

April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

5

Reetta Nylund

Gel-based proteomics

- “Older” approach to screen the protein expression at the large scale
- The typical flow of gel-based proteomics (2DE & MS)
 - Sample preparation
 - First-dimension isoelectric focusing (IEF)
 - Second-dimension SDS-PAGE
 - Visualization & evaluation
 - Expression analysis
 - Protein identification by MS

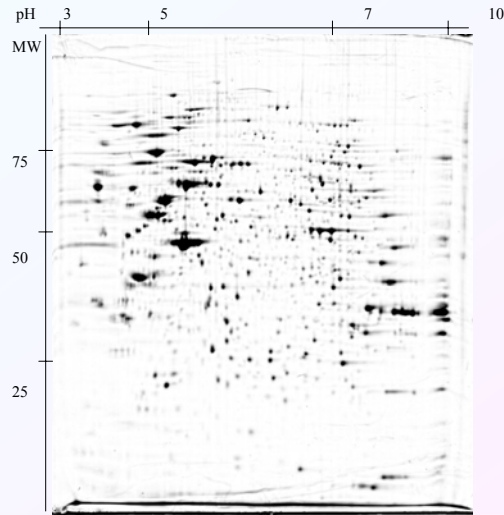
April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

6

Reetta Nylund

Two-dimensional gel electrophoresis (2DE)



- 1st dimension separation based on the pI of proteins
- 2nd dimension separation based on the molecular weight of proteins
- Several visualization/detection possibilities

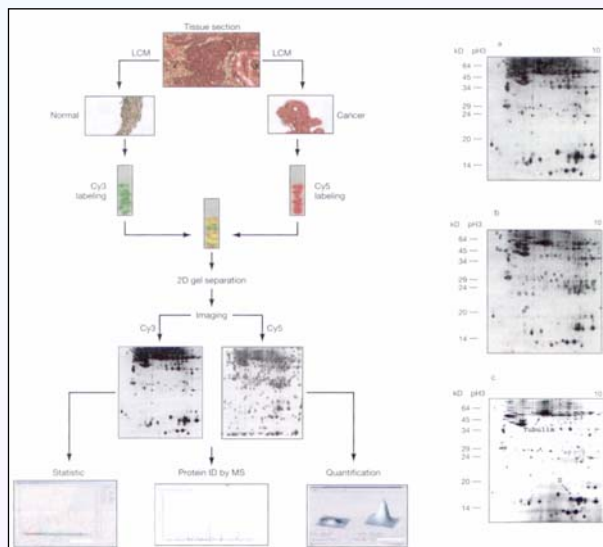
April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

7

Reetta Nylund

Quantitative proteomics - 2DE & MS



From Simpson, 2002, p.4
General outflow of
DIGE procedure
(a control, b cancer)

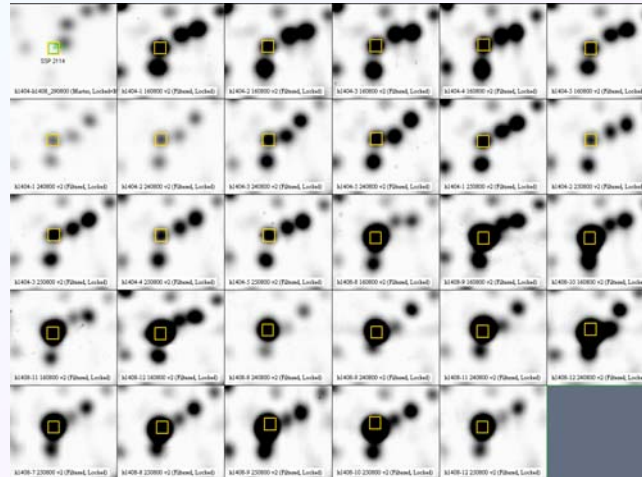
April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

8

Reetta Nylund

An example of a quantitative expression analysis (2DE)



Master gel,
14 controls,
14 cases,
protein shown
6-fold upregulated
t-test $p=2.5 \cdot 10^{-8}$

April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

9

Reetta Nylund

2DE - Highlights and pitfalls

- Highlights
 - Resolving capacity of hundreds of proteins at the same time (Sample fractionation needed for analyzing proteome more completely)
 - Possibility to identify post-translational modifications (based on pI and MW)
- Pitfalls
 - Laborious (time-consuming) and technically challenging
 - Technical limitations e.g. staining method

April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

10

Reetta Nylund

Protein arrays

- “Basic idea similar as in cDNA arrays”;
 - Substrate (protein, antibody etc) is bound on the surface of the array
 - The sample is introduced to array → binding
 - Detection and analysis
- Used for several purposes
 - Screening/profiling
 - Protein - protein interactions
 - Protein - small molecule interactions
 - Kinase - substrate interactions
 - etc.

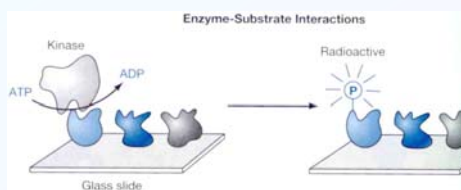
April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

11

Reetta Nylund

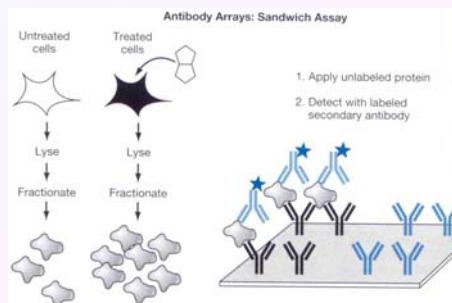
Protein arrays cont.



Both figures from Simpson, 2002

The use of protein arrays to identify enzyme - substrate interactions; a kinase modifies a substrate by transferring of a phosphate group to the substrate

Antibody arrays:
proteins are bound to antibodies
and binding is detected



April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

12

Reetta Nylund

Protein arrays cont.

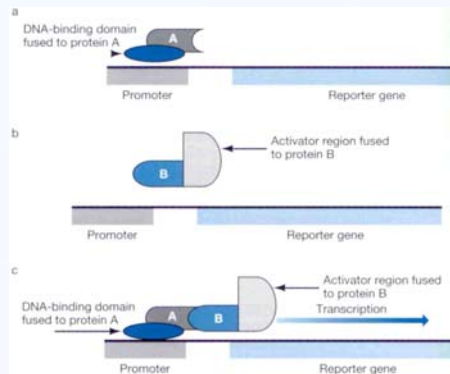
- New screening technique - not much knowledge yet
- “Similarity” to cDNA arrays
 - “Technique exists”
 - Problems e.g.
 - Specificity of substrates, also some proteins are more “sticky” than others
 - Background variations
 - Normalization among experiments

Yeast two-hybrid system

Used to identify protein-protein interactions

- Systematic testing of binding candidates
- Used firstly with yeast (*S. cerevisiae* the whole genome know), but nowadays applied also to other organisms

Yeast two-hybrid cont.



From Simpson 2002

- “Bait” (i.e. a protein whose interacting proteins are searched) is bound to the DNA-binding domain of a reporter gene. Bait alone isn’t capable to turn on the transcription of a reporter gene.
- “Prey” (candidate interacting protein) is bound to the activation region of reporter gene. Also prey alone isn’t capable to turn on the transcription of a reporter gene.
- If bait and prey interact the transcription of a reporter gene is turned on → measurement of reporter gene activity

April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

15

Reetta Nylund

Yeast two-hybrid cont.

The method has been highly successful in detecting many potential protein-protein interactions, but it has several limitations

- A bait capable to self-activation of the reporter gene is unsuitable
- If post-translational modifications (e.g. phosphorylation) are required for binding activity interaction is less likely to be detected
- Also if additional non-peptide factors (e.g. DNA sequence) are needed, the interaction is less likely to be detected
- Some interactions may lack biological content
- Discovery of novel interacting components not possible

April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

16

Reetta Nylund

MS driven techniques

- MS identification of proteins after quantitative analysis by 2DE
 - Peptide mass fingerprinting (Maldi MS)
 - Sequence based identification (MS/MS)
- Identification and quantitation using MS
 - Labeling samples (e.g. ICAT, iTRAQ) for quantitative analysis
 - Identification of the post-translational modifications

April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

17

Reetta Nylund

Generation of peptides for MS analysis

- Proteins digestion
 - Typically by trypsin; cleaves on the C-side of Arg and Lys i.e. generates peptides having R or K at the C-terminus
 - Also other cleaving enzymes e.g. clostripain, endopeptidase Lys-C
- Peptide analysis
 - Total mass of a peptide
 - Peptide fragmentation (mass of single amino acids)

April 1st, 2005

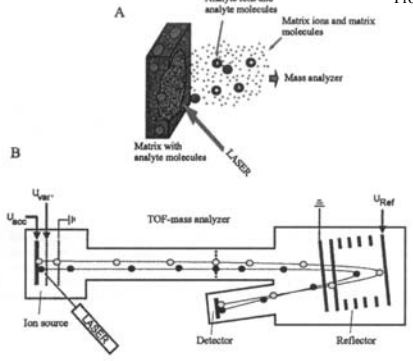
Research Seminar on Data Analysis for Bioinformatics

18

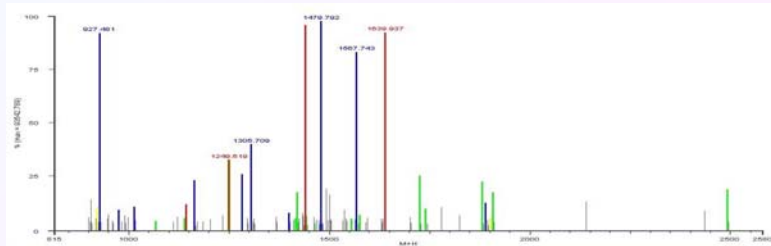
Reetta Nylund

From Mann et al., 2001

Maldi-ToF MS



Sample is cocrystallized with matrix and irradiated with laser, which leads to ionization of the peptides.
The time-of-flight of the peptides is measured, which allows the determination of the peptide masses.



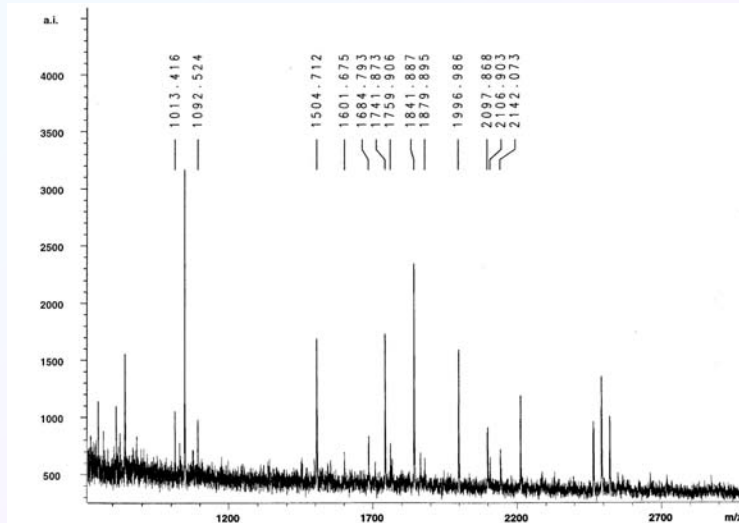
April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

19

Reetta Nylund

Peptide mass fingerprinting analysis



April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

20

Reetta Nylund

Peptide mass fingerprinting analysis cont.

Database search for peptide masses

- Select:
 - Database
 - Taxonomy
 - Enzyme
 - Modifications
 - Peptide tolerance
- Perform the search based on peptide masses

April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

21

Reetta Nylund

Peptide mass fingerprinting analysis cont.

Probability Based Mowse Score

Score is $-10 \cdot \log(P)$, where P is the probability that the observed match is a random event. Protein scores greater than 76 are significant ($p < 0.05$).

Concise Protein Summary Report

Format: [Format As](#) | [Concise Protein Summary](#) | [Help](#)

Significance threshold: Max number of hits:

Hit-Search All Search Unmatched

1. [gi115815573](#) Mass: 49394 Score: 370 Expect: 2.4e-11 Queries matched: 11
 HNF1B [Homo sapiens]

2. [gi112655001](#) Mass: 49404 Score: 369 Expect: 3e-11 Queries matched: 11
 HNF1B protein [Homo sapiens]

3. [gi115815572](#) Mass: 49404 Score: 347 Expect: 4.7e-09 Queries matched: 10
 Heterogeneous nucleic ribonucleoprotein H1 [Homo sapiens]

4. [gi115561538](#) Mass: 49377 Score: 342 Expect: 1.5e-08 Queries matched: 10
 FREDICTED: hypothetical protein_2F_519151 [Pan troglodytes]

5. [gi115815560](#) Mass: 49470 Score: 341 Expect: 1.0e-08 Queries matched: 10
 Heterogeneous nucleic ribonucleoprotein H1 [Homo sapiens]

Mascot Search Results

Protein View

Batch to: [gi115815573](#) Score: 178 Expect: 2.4e-11
 HNF1B [Homo sapiens]

Nominal mass (kDa): 49384 Calculated pI value: 5.79
 NCBI BLAST search of [gi115815573](#) against nr
 Unformatted sequence listing for pasting into other applications

Taxonomy: [Homo sapiens](#)

Fixed modifications: Carbamidomethyl (C)
 Variable modifications: Oxidation (M)
 Cleavage by Trypsin: cuts C-term side of K unless next residue is P
 Number of mass values searched: 13
 Number of mass values matched: 11
 Sequence Coverage: 35%

Matched peptides shown in Bold Red

```

1 VQLVYRGGK PVOVQGLFW SCARVDFP FDCCKIQRGQ QIIFITTR
31 GDFSRAPVE LKSRHVELL LKSRITRGS RYVYVFRSH VERDYLDT
101 GDFSRVFRM GDFVRLDLP GDFEYVDFP FDFLELVNG IFLVDFQR
131 STGSRVDFR SGRSRKALK EDRERLGGY IIFKSSGAK VDTYDFPR
201 LRKRGKDFP IPRKAGDTH IIRGKGFES RPRGATGGY GDTLDTQTH
231 SDFPQDFP GDFKRVETK RQVDFVDFK VDFYDFRQK VDFHGLFTR
301 RFDITDFYS FDFVDFVDFE IPRDFVDFE IDFVFATRE IVALRDKKA
351 NDFSRVDFP LKSRHVELL LKSRITRGS RYVYVFRSH VERDYLDT
481 GLRDKSTGQ PARQQLDQV GDFVDFVDFE IIFKSSGAK VDTYDFPR
    
```

Show predicted peptides also

Sort Peptides By	# Residue Number	Increasing Mass	Decreasing Mass		
Start - End	Observed	M (exp)	M (calc)	Delta	Miss Sequence
1 - 16	1741.97	1740.97	1740.95	0.01	VQLVYRGGKPPVQVDFE I Oxidation (M)
15 - 29	1739.91	1738.98	1738.85	0.03	VQLVYRGGKPPVQVDFE
17 - 29	1584.78	1583.78	1583.68	0.03	GLDFVDFRDFVDFE
50 - 68	2164.98	2163.98	2163.98	-0.00	SDSRVDFRDFVDFE
99 - 114	1684.79	1683.79	1683.76	0.03	STGSRVDFRDFVDFE
131 - 167	1841.89	1840.88	1840.88	-0.00	STGSRVDFRDFVDFE
263 - 275	1681.67	1680.67	1680.64	0.03	RLVDFVDFRDFVDFE
276 - 294	2097.87	2096.86	2096.86	-0.02	VDFVDFVDFRDFVDFE
300 - 316	1896.99	1895.98	1895.97	0.01	ATSDITDFVDFVDFE
317 - 326	1897.92	1891.92	1891.92	-0.06	VDFVDFVDFE
336 - 375	2162.67	2161.67	2161.62	0.03	VQLVYRGGKPPVQVDFE

April 1st, 2005

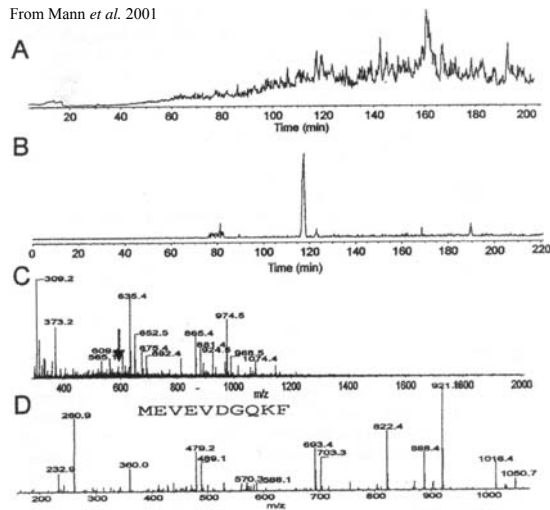
Research Seminar on Data Analysis for Bioinformatics

22

Reetta Nylund

LC-MS analysis

From Mann *et al.* 2001



LC (Liquid Chromatogram) separation of complex mixtures of peptides

- A) Total ion current chromatogram
- B) Ion current for $m/z = 591.5$, selected ion current chromatogram
- C) Mass spectra at 117 min
- D) MS/MS spectrum at $m/z 591.5$ and the interpreted amino acid sequence

April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

23

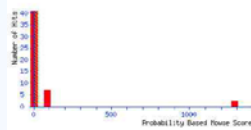
Reetta Nylund

Peptide fragmentation - ESI MS

MS data file: DATA.TXT
 Database: M08 20041018 (154924): sequences: 82244(492 residues)
 Taxonomy: Homo sapiens (human) (26129 sequences)
 Timestamp: 28 Oct 2004 at 15:11:42 GMT
 Significant hits: 12287 - TRANSFERRIN precursor (val101-02) - human
 4052141 - TRANSFERRIN precursor (val101-02) - human
 4052141 - TRANSFERRIN precursor (val101-02) - human
 4052141 - TRANSFERRIN precursor (val101-02) - human
 4052141 - TRANSFERRIN precursor (val101-02) - human
 4052141 - TRANSFERRIN precursor (val101-02) - human
 4052141 - TRANSFERRIN precursor (val101-02) - human
 4052141 - TRANSFERRIN precursor (val101-02) - human
 4052141 - TRANSFERRIN precursor (val101-02) - human
 4052141 - TRANSFERRIN precursor (val101-02) - human

Probability Based Mowse Score

Low score is $-10^{\log(P)}$, where P is the probability that the observed match is a random event.
 Individual low scores < 30 indicate identity or extreme homology (p < 0.05).
 Protein scores are derived from low scores as a non-probabilistic basis for ranking protein hits.



Peptide Summary Report

[Click to Print Summary Report](#)

To create a bookmark for this report, right click this link: [Peptide Summary Report \(print.asp? 307\)](#)

Select All | Select None | Search Selected | Error tolerance | Archive Report

1. TRANSFERRIN precursor (val101-02) - human
 Mass: 79280 Score: 1283 Peptide matched: 134

transferrin precursor (val101-02) - human

Check to include this hit in your tolerance search or archive report

query	observed	MS (ppm)	MS (calcd)	Delta	Miss	Score	Expect	Rank	Peptide
1	635.03	634.02	634.37	-0.35	0	23	3.1	1	SLLEP
2	635.03	634.04	634.37	-0.33	0	23	3.1	1	SLLEP
3	635.06	634.04	634.37	-0.31	0	23	3.1	1	SLLEP
4	635.06	634.04	634.37	-0.31	0	23	3.1	1	SLLEP
5	635.19	642.18	642.38	-0.19	0	12	8.1	1	SLLEP
6	734.03	734.02	734.40	-0.37	0	13	2.6	2	QVQAPVF
7	734.13	734.12	734.40	-0.27	0	12	1.2	1	QVQAPVF
8	734.14	734.13	734.40	-0.26	0	12	1.2	1	QVQAPVF
9	734.22	734.21	734.40	-0.18	0	18	1	1	QVQAPVF

MASCOT Search Results

Peptide View

MS/MS Fragmentation of **SASDFEWDQK**

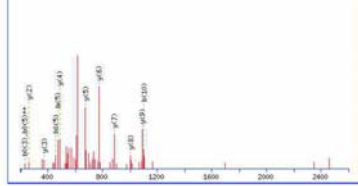
Found in **TRANSFERRIN precursor (val101-02) - human**

Match to Query 2: 1248 96748 (mass: 82244(492))

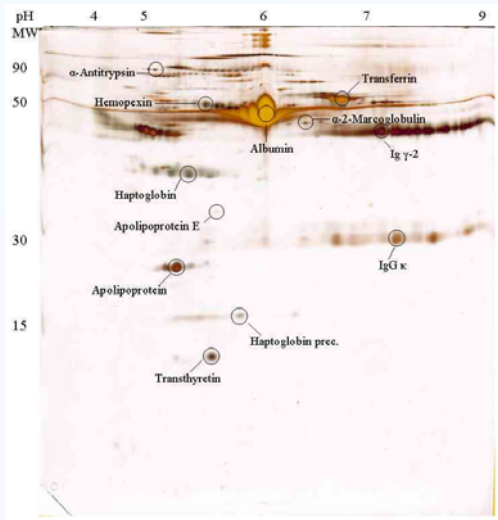
From data file DATA.TXT

Click mouse within plot area to zoom in. In Factor of two about that point

Plot from 400 to 2000 Da



Evaluation of the results - an example



Human serum sample (w/o albumin depletion), 2DE and protein analysis using MS

- Proteins analyzed with Maldi and ESI Ion Trap
- α -Antitrypsin
 - No identification by Maldi
 - Identification by ESI with high score (sample containing also human keratin...)
 - Databases MW ~ 47kDa, pI ~ 5.4
 - Correct?

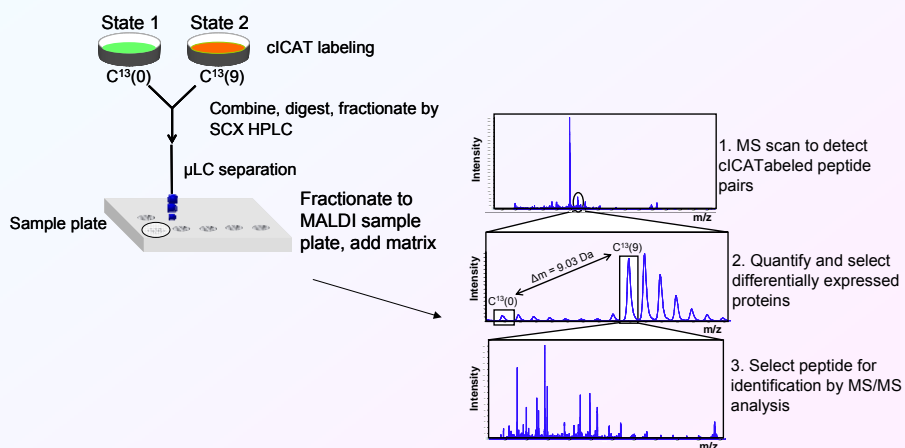
April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

25

Reetta Nylund

ICAT



Courtesy of Timothy Griffin, Univ. Minnesota, USA

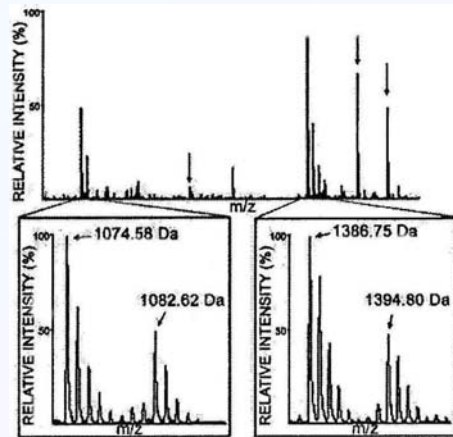
April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

26

Reetta Nylund

ICAT analysis



From Smolka *et al.* 2002

- Peptide identification by MS/MS or protein analysis by Maldi
- Quantitative analysis by ICAT peaks

April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

27

Reetta Nylund

Analysis of post-translational modifications



From Mann *et al.* 2001

- Several strategies to analyze post-translational modifications
- Phosphorylated peak in MS; additional mass of 79.966 Da

April 1st, 2005

Research Seminar on Data Analysis for Bioinformatics

28

Reetta Nylund

Proteomics summary

- There are several different approaches used in proteome analysis; most of them are labourous at least at some point of analysis
- The methods are generally efficient for profiling the proteome and analyze expression differences in case-control studies, but to obtain complete proteome usually one method is not enough...

References

J Pevsner, *Bioinformatics and Functional Genomics*, John Wiley & Sons, NJ USA, 2003

RJ Simpson, *Proteins and Proteomics*, Cold Spring Harbor Laboratory Press, NY USA, 2002

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R, Quantitative analysis of complex protein mixtures using isotope-coded affinity tags, *Nat Biotechnol* 1999, 17:994-999

Mann M, Hendrickson RC, Pandey A, Analysis of Proteins and Proteomes by Mass Spectrometry, *Annu. Rev. Biochem.* 2001, 70:437-473

Smolka M, Zhou H, Aebersold R, Quantitative Protein Profiling Using Two-dimensional Gel Electrophoresis, Isotope-coded Affinity Tag Labeling, and Mass Spectrometry, *Molecular and Cellular Proteomics* 2002, 1.1:19-29