CNR - ITB
Institute for Biomedical Technologies
National Research Council

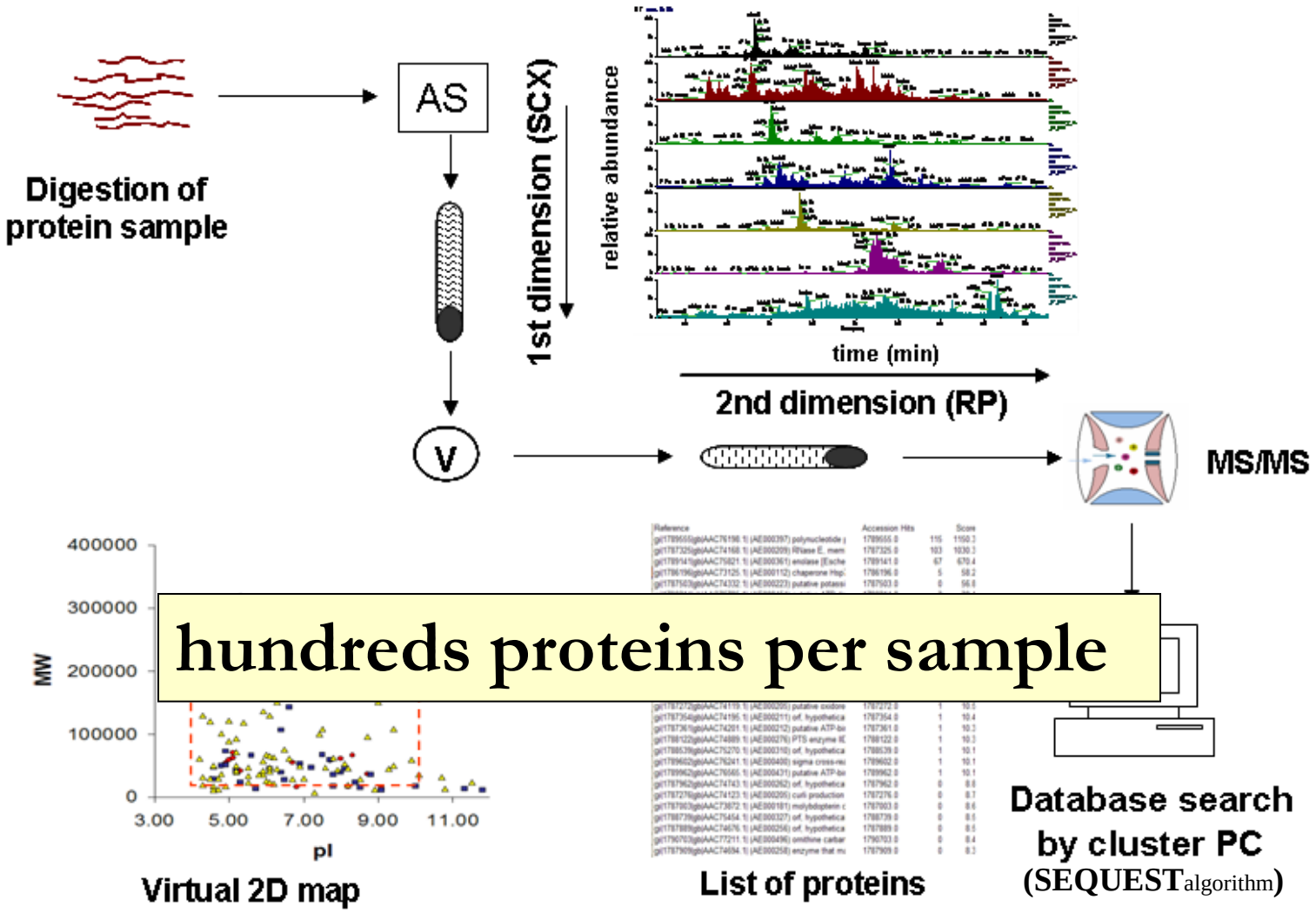
User-friendly bioinformatics tools for handling proteomics data in biomarker discovery and cluster analysis

Dipartimento
PROGETTAZIONE
MOLECOLARE

Dario Di Silvestre

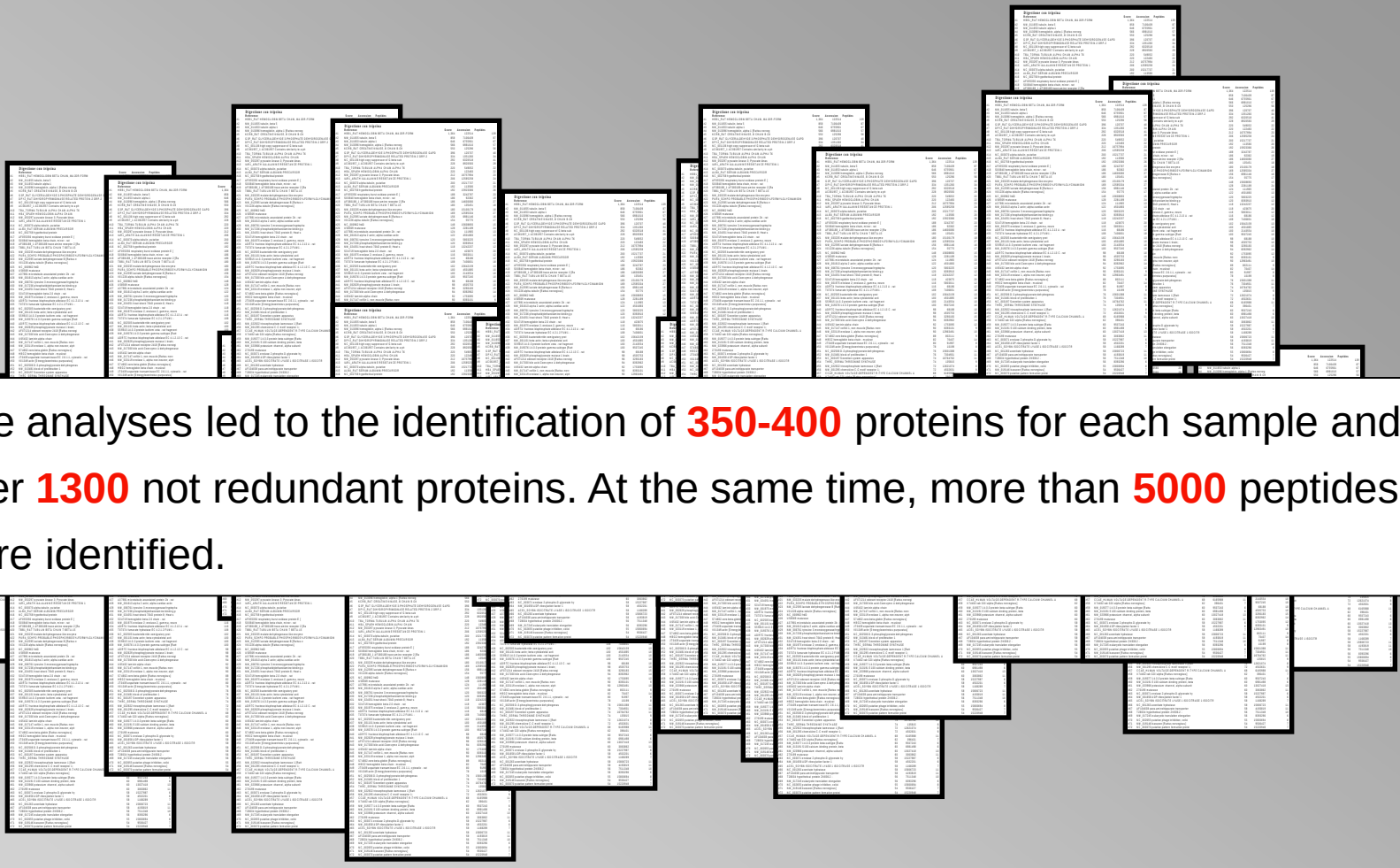
Proteomics and Metabolomics unit

MudPIT - Multidimensional Protein Identification Technology



Mauri & Dehò, *Methods in Enzymology*, 2008

2DC-MS/MS analysis of *Sus scrofa* cardiac tissues

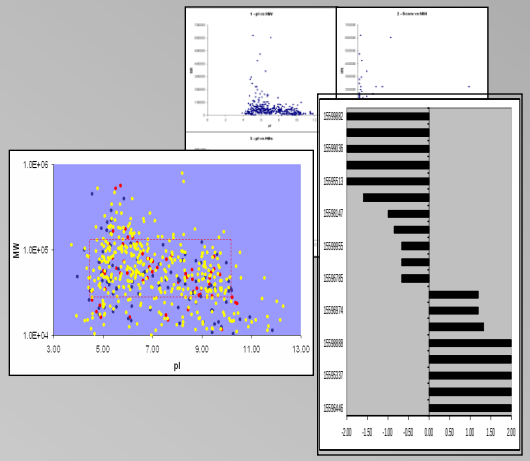


The analyses led to the identification of **350-400** proteins for each sample and over **1300** not redundant proteins. At the same time, more than **5000** peptides were identified.

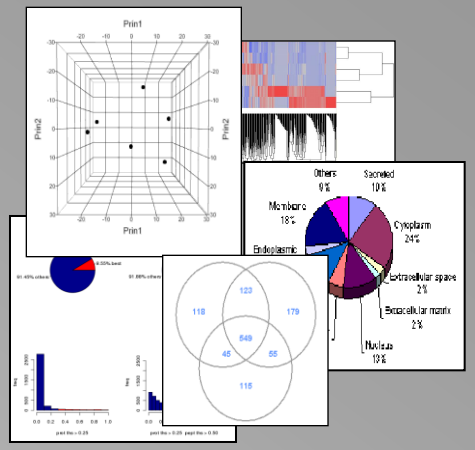
in collaboration with Scuola Superiore Sant'Anna, Sector of Medicine, Pisa, Italy

Data processing workflow

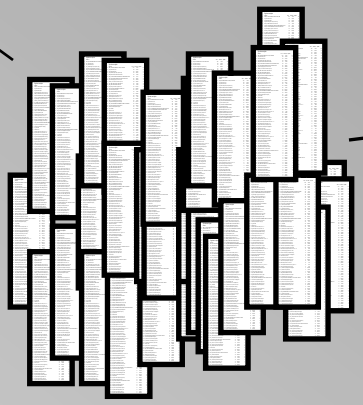
Biomarker discovery



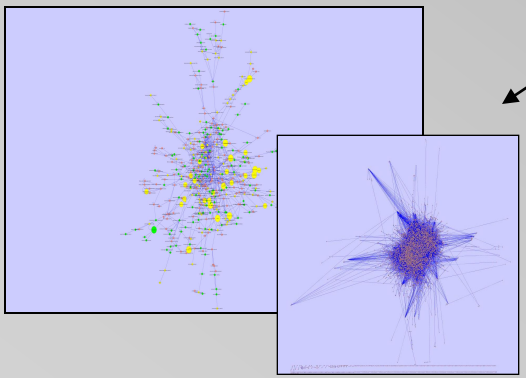
Statistical evaluations



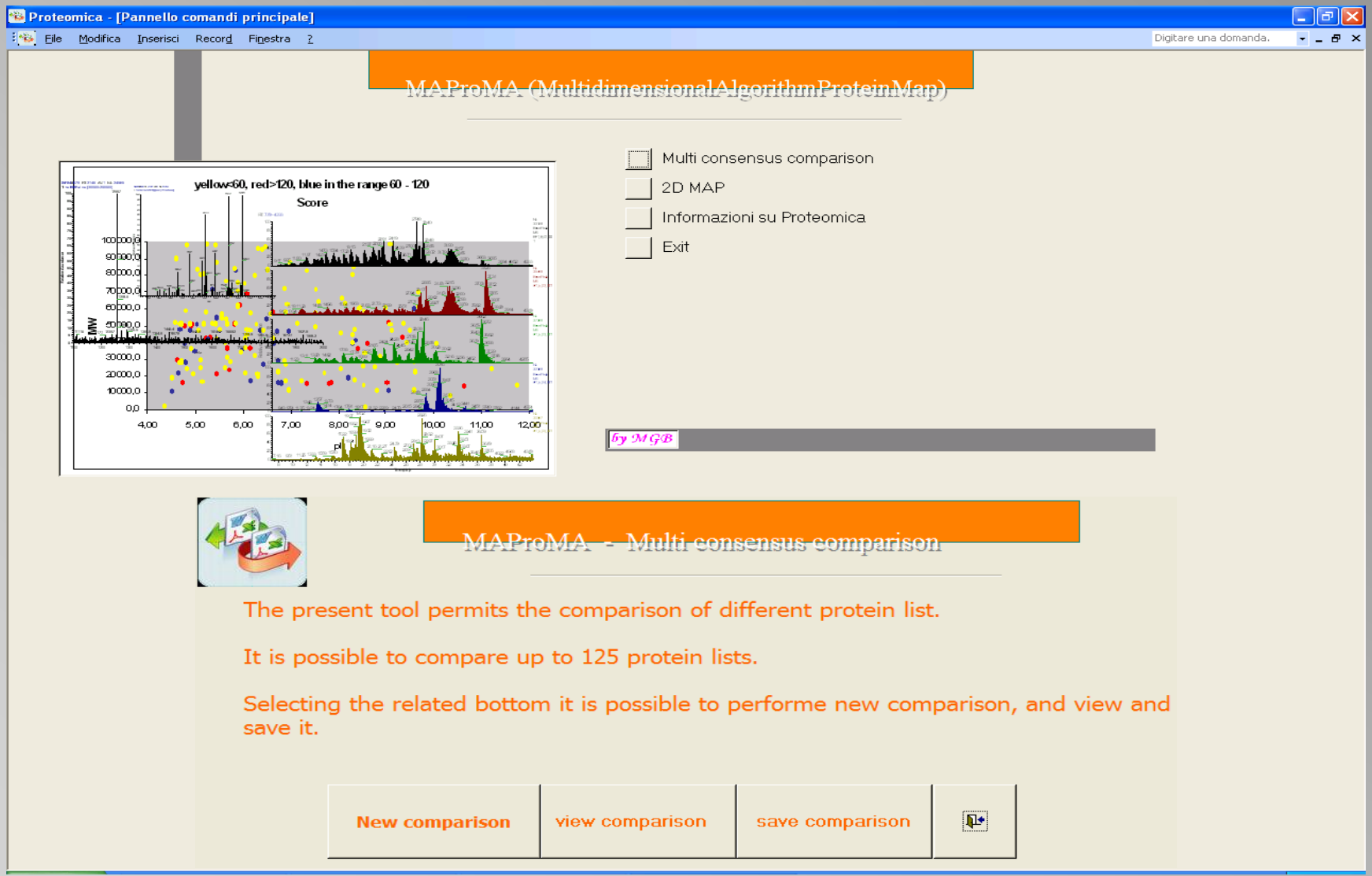
Protein and peptide lists



Protein-protein interaction network



MAProMa (MultidimensionalAlgorithmProteinMap)



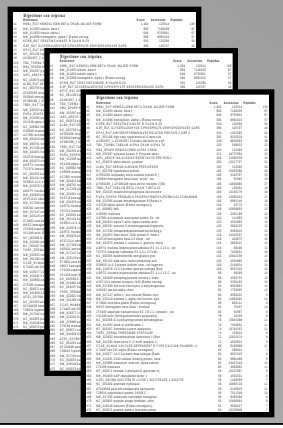
Proteomic data analysis

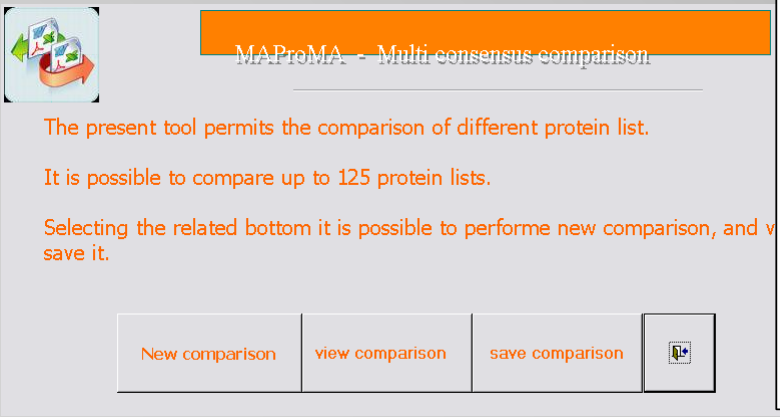
MAProMa

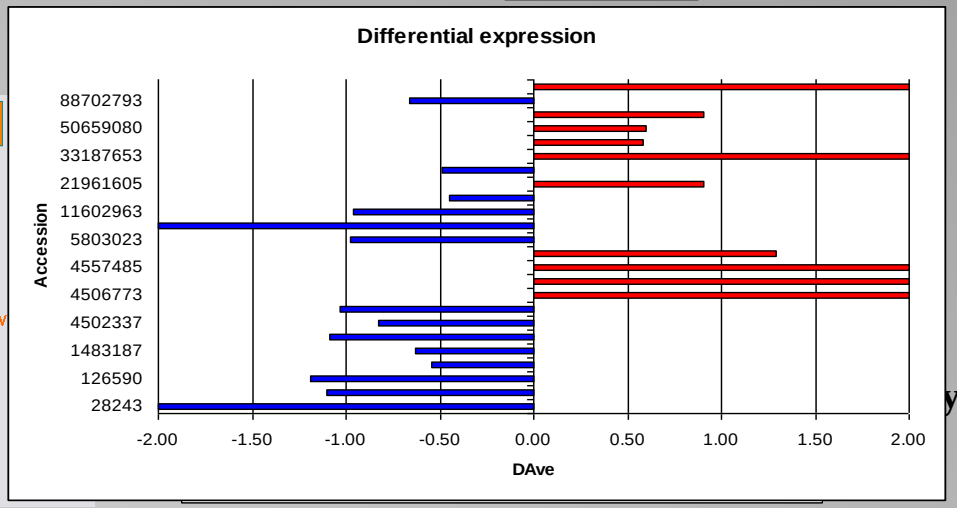
Multidimensional **A**lgorithm **Pro**tein **M**ap

- *2D-Maps*
  - pI-MW plot of proteins from a list.
- *Differential expression*
  - Using DAve and DCI formulas (Mauri et al. FASEB J., 2005)
- *Comparison*
  - To compare up to 125 protein lists.

Proteins/Peptides lists



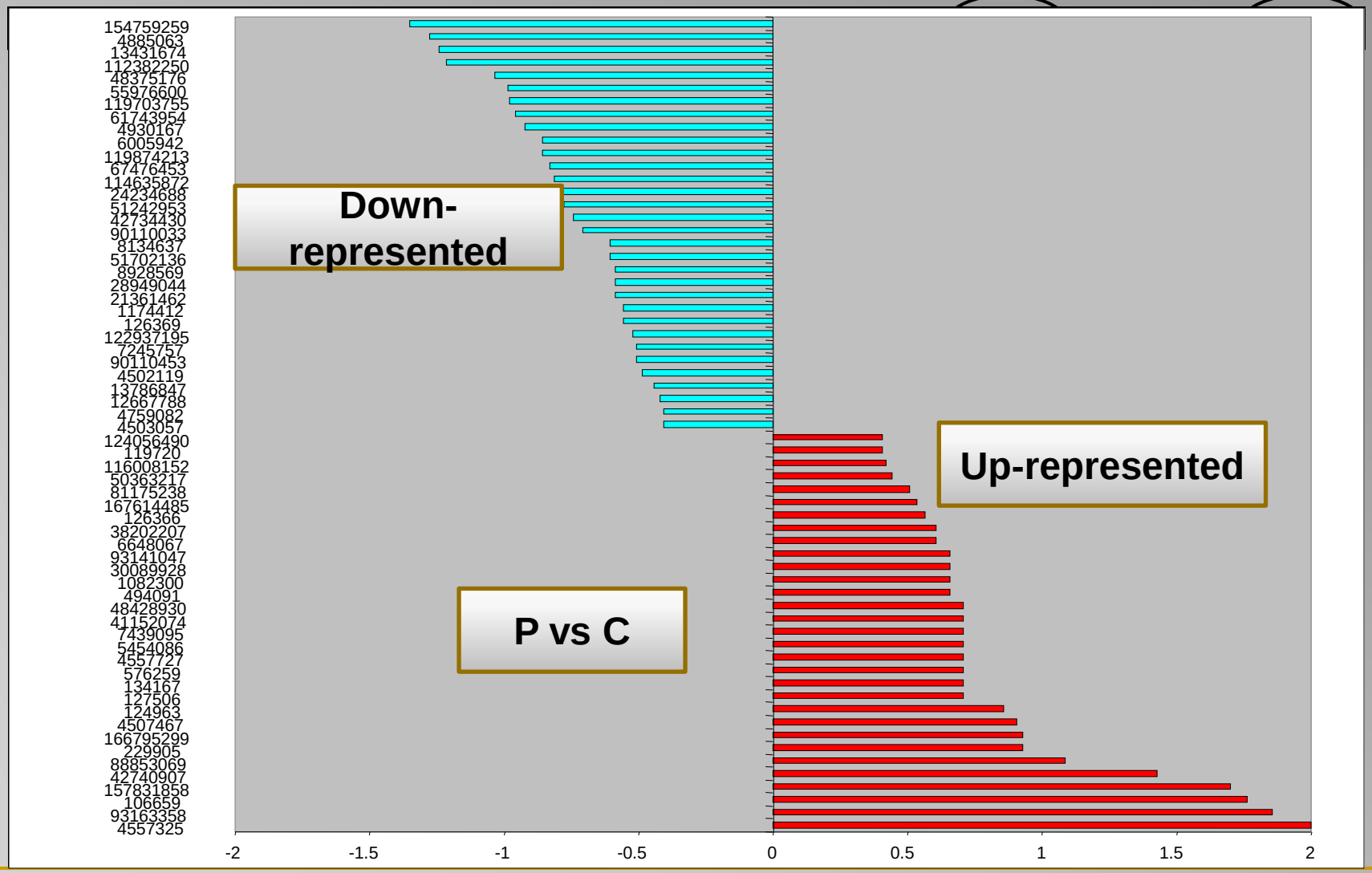




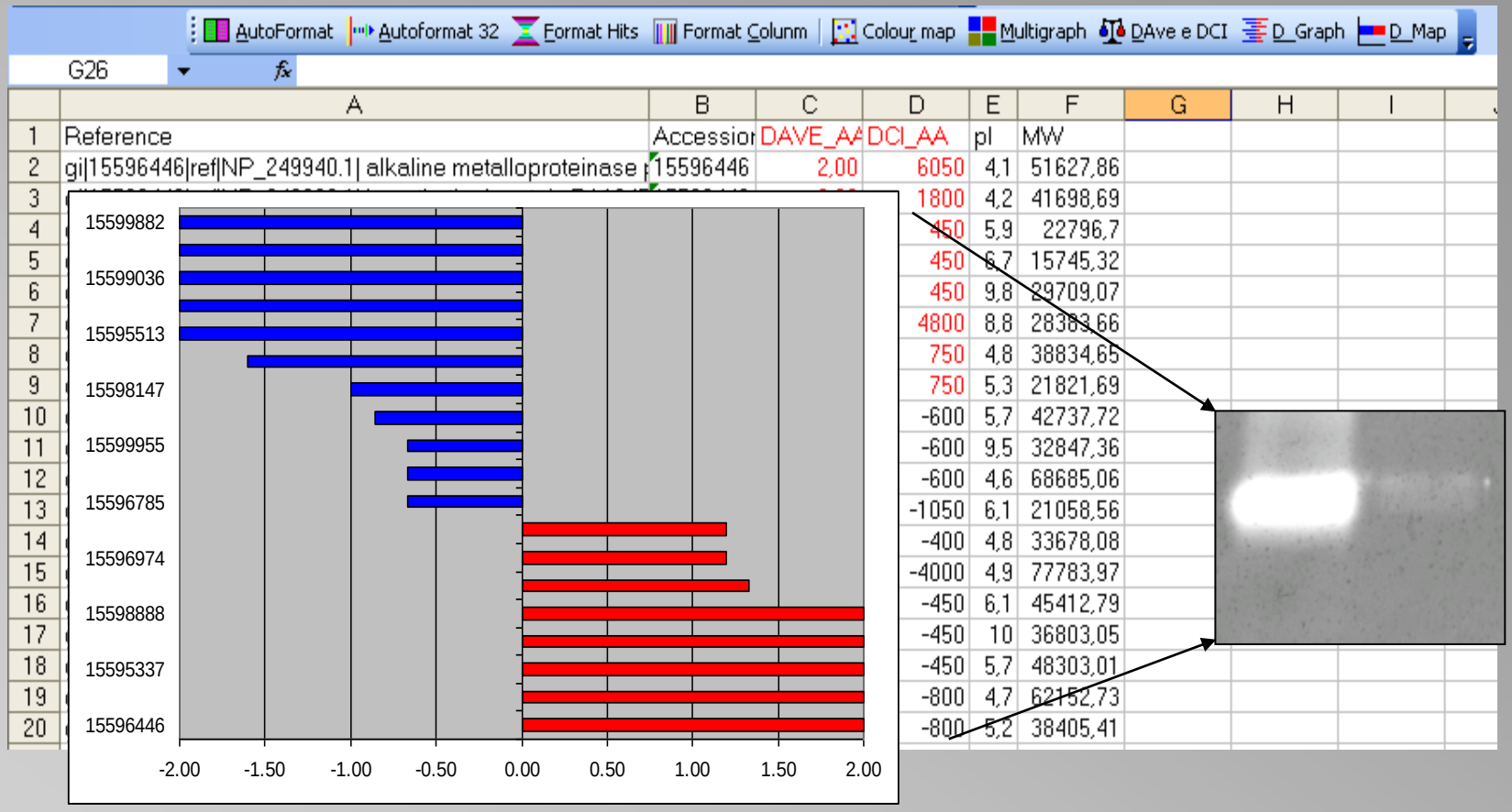
ysis

Patients *vs* Volunteers

Comparison → Differential analysis → Frequency → Differential proteins



MAProMa tool bar



	A	B	C	D	E	F	G	H	I
1	Reference	Accessior	DAVE_AA	DCI_AA	pI	MW			
2	gi|15596446|ref|NP_249940.1| alkaline metalloproteinase p	15596446	2,00	6050	4,1	51627,86			
3				1800	4,2	41698,69			
4				450	5,9	22796,7			
5				450	6,7	15745,32			
6				450	9,8	29709,07			
7				4800	8,8	28383,66			
8				750	4,8	38834,65			
9				750	5,3	21821,69			
10				-600	5,7	42737,72			
11				-600	9,5	32847,36			
12				-600	4,6	68685,06			
13				-1050	6,1	21058,56			
14				-400	4,8	33678,08			
15				-4000	4,9	77783,97			
16				-450	6,1	45412,79			
17				-450	10	36803,05			
18				-450	5,7	48303,01			
19				-800	4,7	62152,73			
20				-800	5,2	38405,41			

in collaboration with Cystic Fibrosis Center of Verona
Bergamini et al, 2010 submitted

Hierarchical Clustering

and

Principal Component Analysis

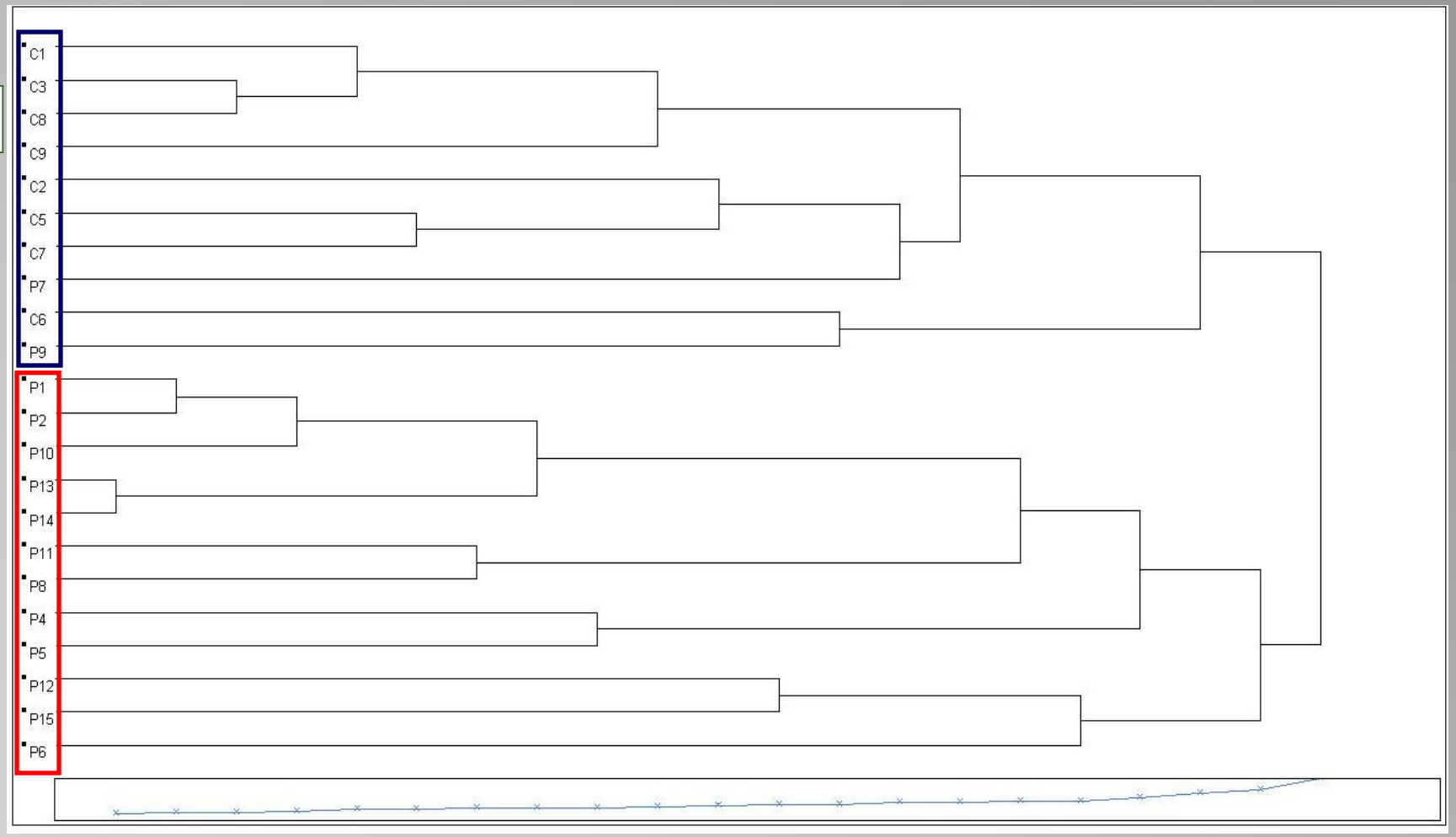
To describe in concise form the nature and structure of data

Task: to assign objects (samples) to classes on the basis of measurements (proteins) made on the objects

Unsupervised learning: the classes are unknow a priori and need to be "discovered" from the data

Automatic script for hierarchical clustering and PCA

Based on MAProMa "Comparison"

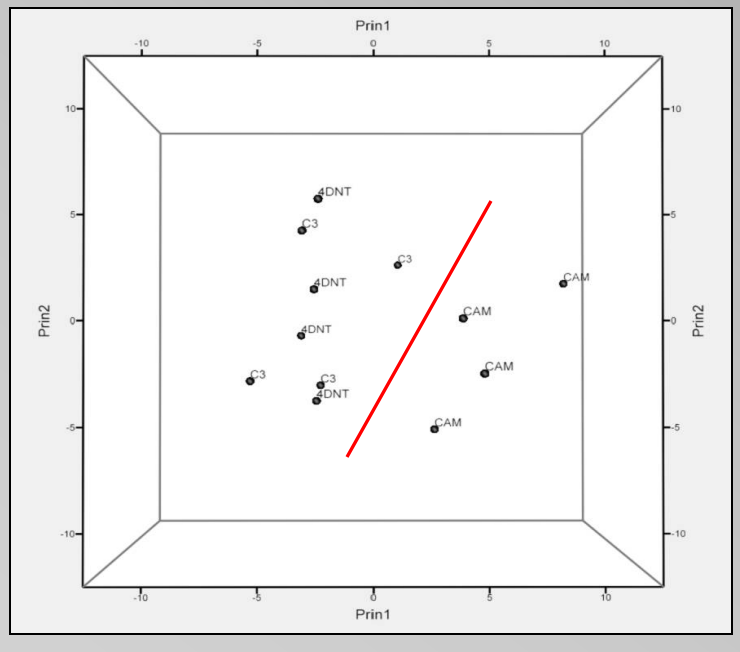


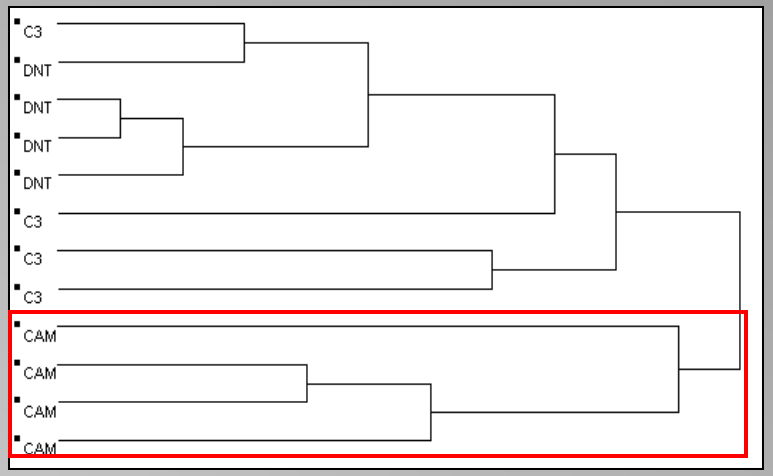
Proteomic profiles of abdominal subcutaneous fat tissue from patients with systemic amyloidoses and healthy controls

Mesembryanthemum crystallinum

C3 photosynthesis **CAM photosynthesis**

C3 metabolic state	**4_DNT intermediate phase**	**CAM metabolic state**
NaCl stress T0	After 4 days	After 12 days



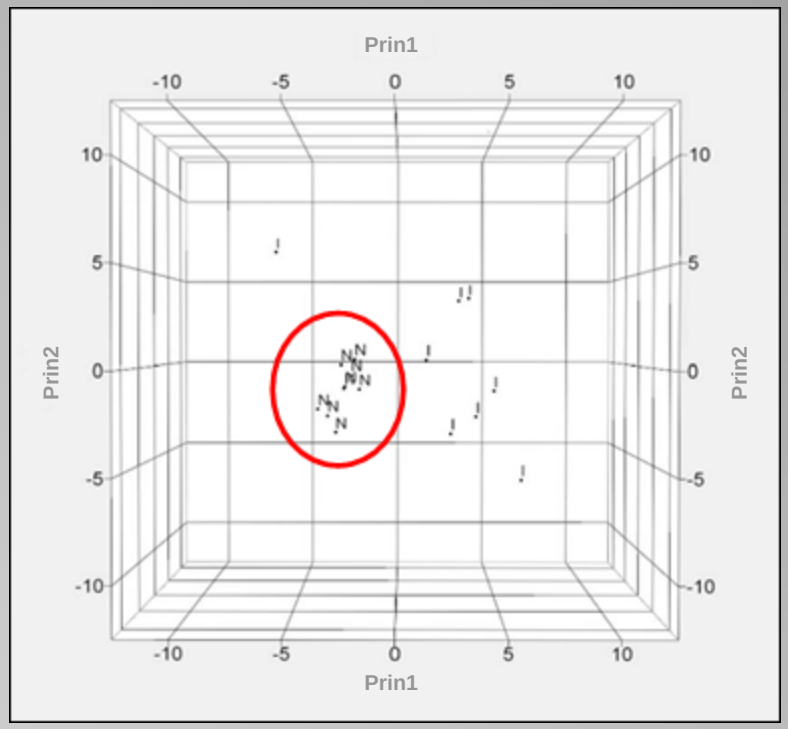


in collaboration with Institut für Botanik, Technische Universität Darmstadt, Germany
Cosentino & Di Silvestre, 2010 submitted

Sus scrofa cardiac tissues

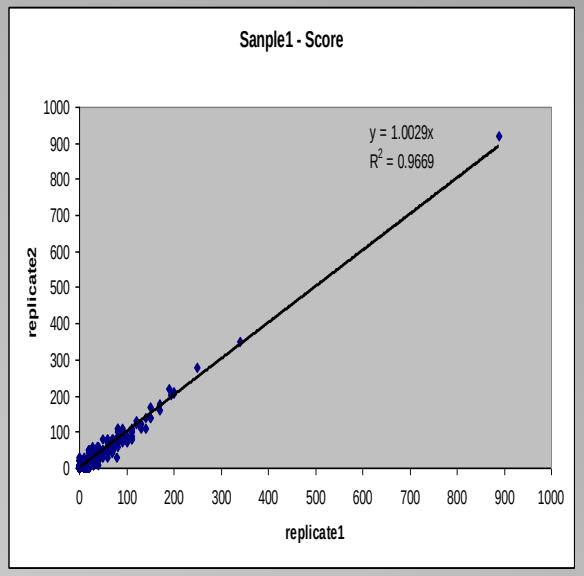
N→ Normal Tissues

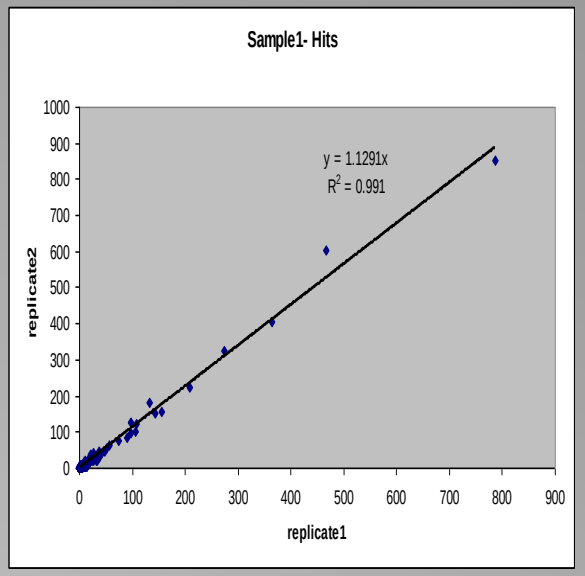
I→ Myocardial infarction



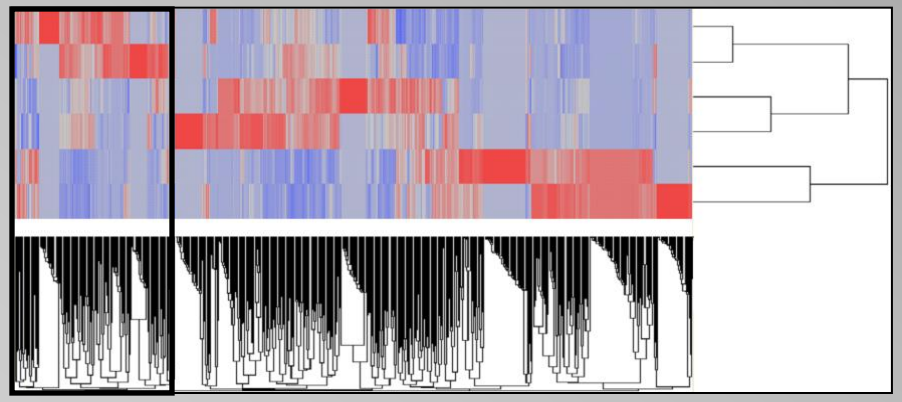
in collaboration with Scuola Superiore Sant'Anna, Sector of Medicine, Pisa, Italy

MudPIT reproducibility





Sample1 – Replicate1
Sample1 – Replicate2
Sample2 – Replicate1
Sample2 – Replicate2
Sample3 – Replicate1
Sample3 – Replicate2



Advantages:

Hierarchical clustering and PCA

are applicable to MudPIT proteomic data.

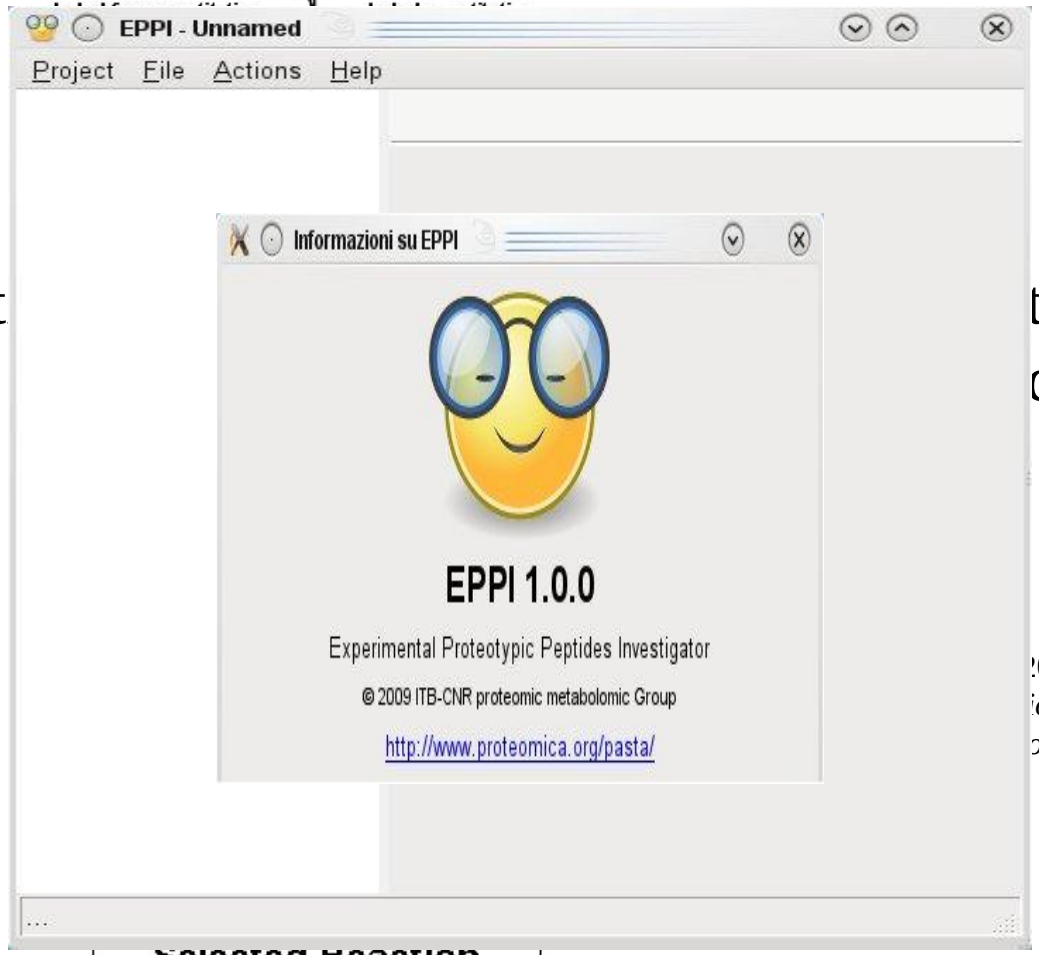
- **Samples discrimination based only on experimental data.**
- **Rapid identification of potential biomarkers from MudPIT proteins lists.**
- **Rapid evaluation of MudPIT analysis performance.**

EPPI - Experimental Proteotypic Peptides Investigator

Shotgun proteomics

Profiling

Mutant



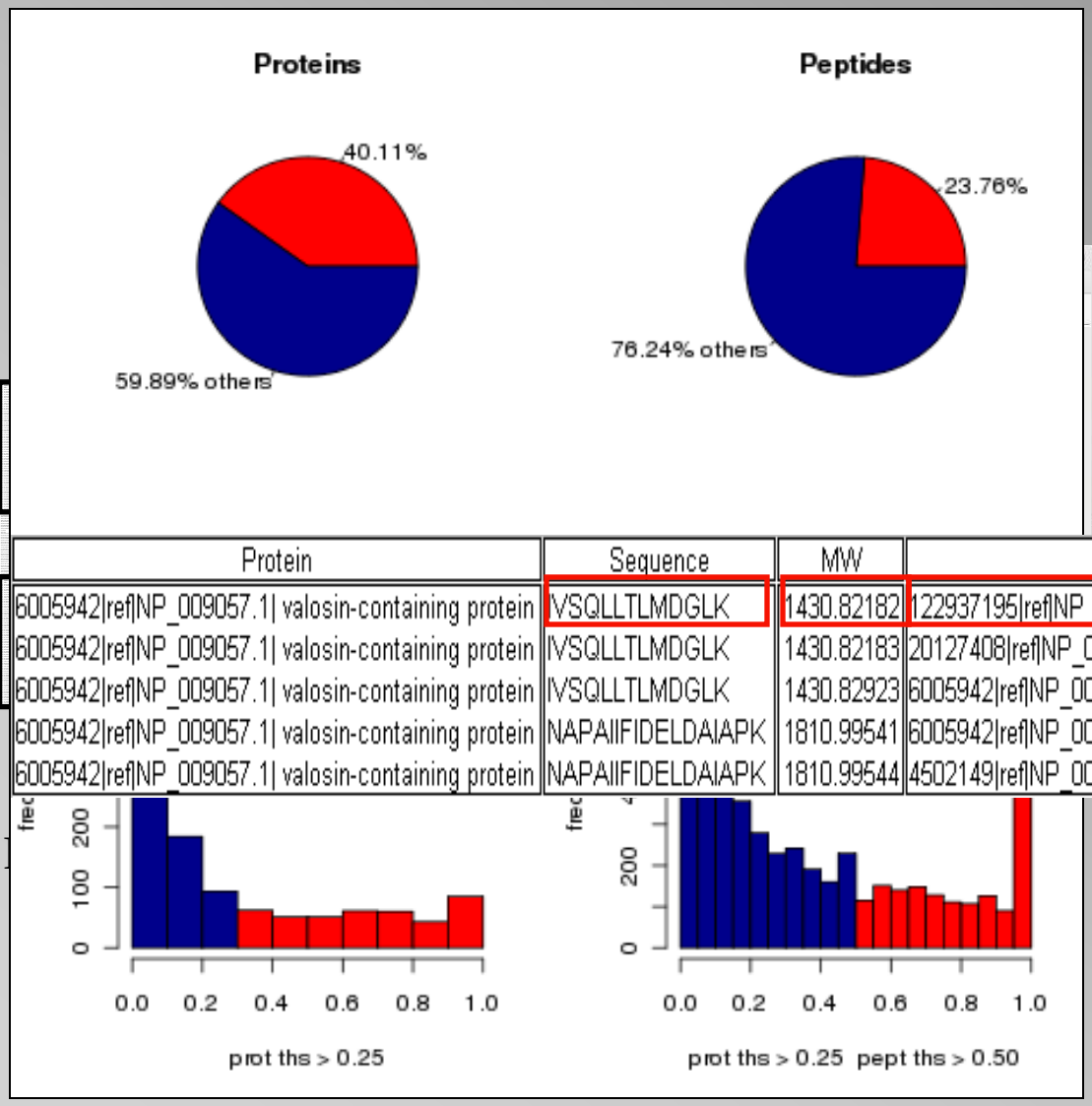
Those pept t likely to be observed ds and that

2005, 19(13), 1844-1850.
iol., 2005, 6(7), 577-583.
ol., 2007, 25(1), 125-131

Selected Reaction Monitoring

Di Silvestre et al. Current Pharmaceutical Analysis, 2010

EPPI - Experimental Proteotypic Peptides Investigator



Protein	Sequence	MW	Protein match
6005942|ref|NP_009057.1| valosin-containing protein	IVSQLLTLMDGLK	1430.82182	122937195|ref|NP_001073869.1| plasma membrane associated protein, S3-12
6005942|ref|NP_009057.1| valosin-containing protein	IVSQLLTLMDGLK	1430.82183	20127408|ref|NP_000173.2| mitochondrial trifunctional protein
6005942|ref|NP_009057.1| valosin-containing protein	IVSQLLTLMDGLK	1430.82923	6005942|ref|NP_009057.1| valosin-containing protein
6005942|ref|NP_009057.1| valosin-containing protein	NAPAIIFIDELDAIAPK	1810.99541	6005942|ref|NP_009057.1| valosin-containing protein
6005942|ref|NP_009057.1| valosin-containing protein	NAPAIIFIDELDAIAPK	1810.99544	4502149|ref|NP_001634.1| apolipoprotein A-II preproprotein

- N° of samples: 40
- N° of entries in the fasta database:12915
- N° of proteins: 1142
- Number of peptides: 4894
- Thresholds (Proteins/Peptides): 0.25/0.5
- N° of host proteins: 458
- N° of unique peptide sequences: 220
- N° of unique peptide mass: 28

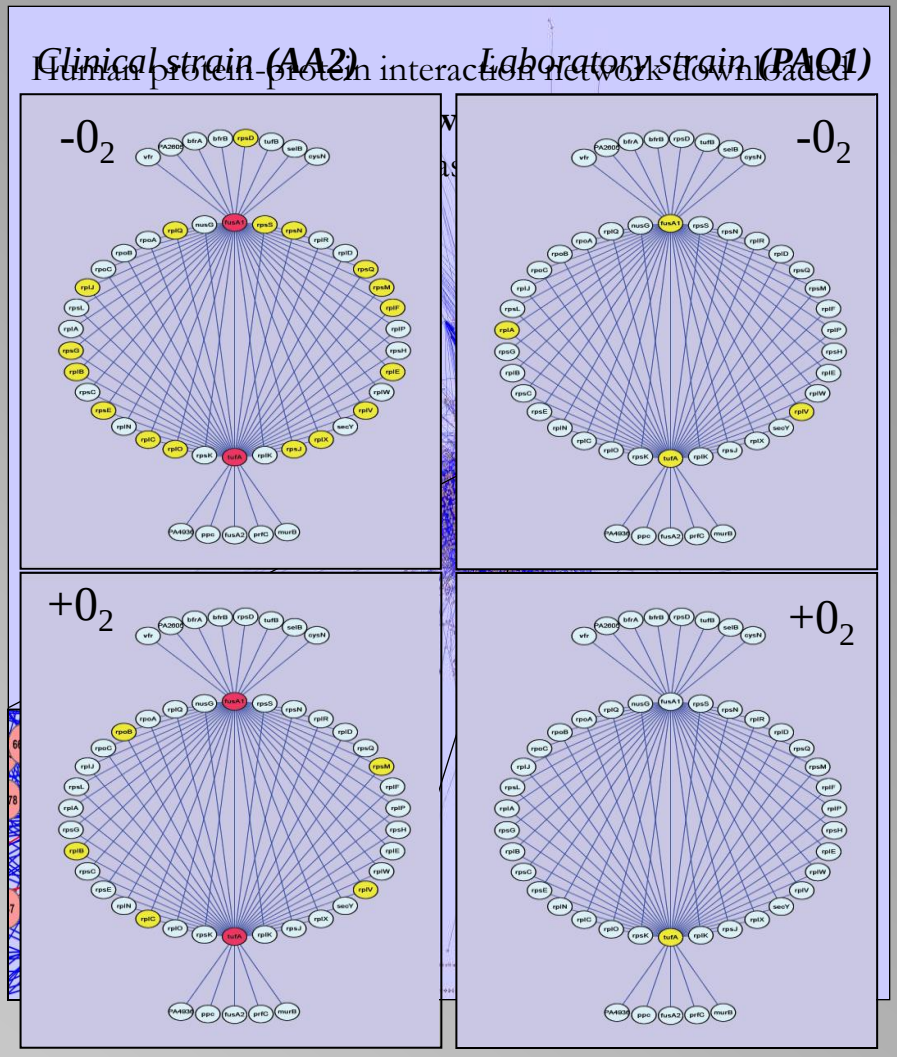
Exp.

Proteotypic peptides

Network based approach

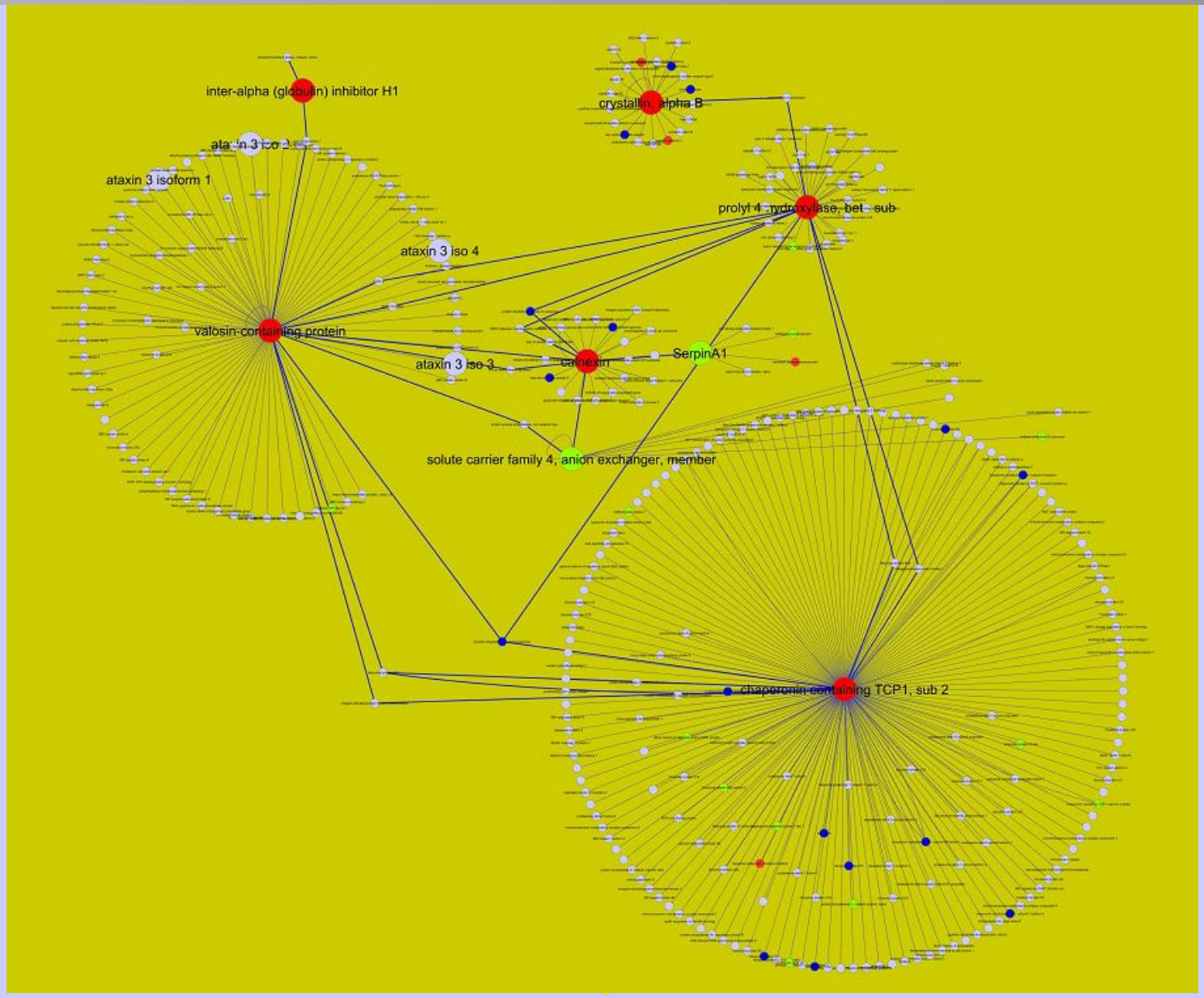
P. Aeruginosa

- **Given high-throughput measurements (LC/LC-MS/MS) comparing two conditions, identify groups of over-, under- and normal-expression.**
- **Non-quantified proteins are NOT (necessarily) unchanged in expression**
- **Assume all pre-processing done and that we have relative protein concentrations between two conditions**
- Problem: poor coverage: 10-20%



Red – Differentially represented proteins (MAProMA)
Yellow – Proteins identified by MudPIT analysis
White - Other proteins interacting with red proteins

Bergamini et al, 2010 submitted



in collaboration with Amyloidosis Research and Treatment Center, IRCCS Policlinico San Matteo, University of Pavia, Italy

Conclusion

The very large amount of data produced by MS-based proteomics approaches, such as MudPIT, requires powerful informatics supports, **at various levels of sophistication.**

Using simple software solutions coupled to software, such as Cytoscape, proved to be very useful for the organization of data and for their interpretation in an unbiased manner.

The MudPIT approach has proved to be a powerful technology suitable to investigate complex biological samples. In this area of research its importance increases in parallel with the rapid evolution of bioinformatics science.

Acknowledgements

Scientific Team

Pierluigi Mauri (PI)
L. Benazzi
F. Basilico
A. De Palma
D. Di Silvestre
F. Brambilla
E. Rosti
R. Rossi
E. Franceschi
C. Comunian
P. Brunetti
S. Motta
B. Carulli
V. Bellettato

Gone
but
not forgotten

M. Fer
S. Daminelli
A. Sharma

Thank you for attention!

dario.disilvestre@itb.cnr.it