POLITECNICO
DI TORINO

COMPUTATIONAL TOOLS FOR THE INTERACTIVE EXPLORATION OF PROTEOMIC DATA AND AUTOMATIC BIO-NETWORKS RECONSTRUCTION

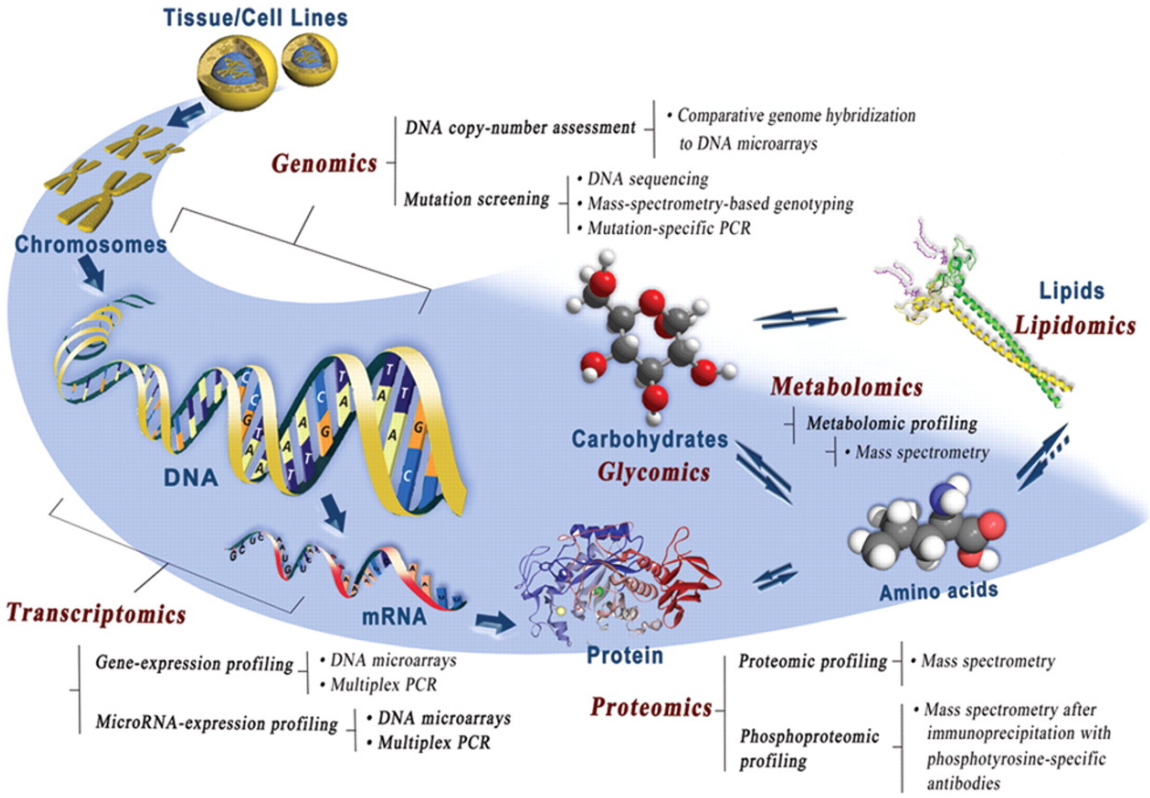
Massimo Natale

Tutor:
Enrico Macii
Elisa Ficarra

Agenda

- ✓ State of the art
- ✓ Aims of the work
- ✓ Methods
- ✓ Projects
- ✓ Research Activities and Results
- ✓ Publications and impact of resarch

Omics studies are substantially **changing the face** of biomedical and life science research. The generation of a **huge amount of data** is exceeding researchers' ability to effectively manage them.



Moreover, most of the knowledge, produced by 'omics' sciences, is collected within **biomedical publications** (hundreds of thousound per year), in **heterogeneous databases**.

Aims

Goal of my PhD was carried out the **development** of a new bioinformatics **approaches**, **software**, and **tools**, able to perform the analysis of experimental and published 'omics' data, using big data and systems biology strategies

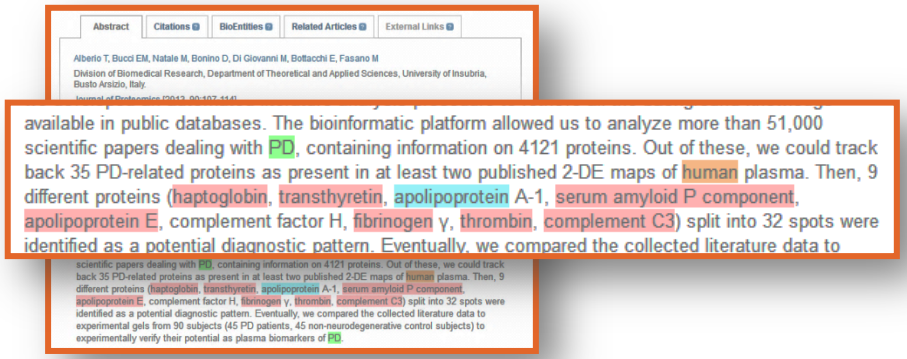
big data

The 'omics' experiments can collect data without an existing hypothesis, paving the way for the arrival of big biology and systems biology approach to scientific practice. It is fundamentally a science- and data-driven approach to bioprocessing. The data driven biology encourages the development of Big Data technologies.

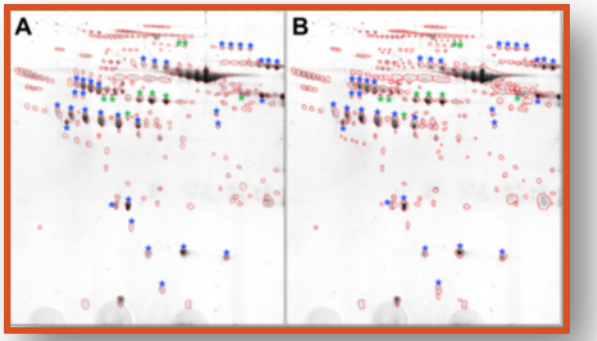
systems biology

Systems biology is the computational and mathematical modeling of complex biological systems. An **emerging engineering approach** applied to biomedical and biological scientific research. Systems biology is a biology-based inter-disciplinary field of study that focuses on **complex interactions** within biological systems, using a holistic approach.

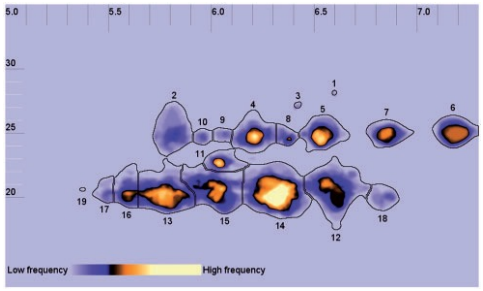
In order to improve analysis, management, and access to 'omics' data, we use different ICT methods for combining information coming from:



1. PubMed literature



2. experimental results



3. image meta-analysis

ExPASy Bioinformatics Resource Portal

STRING 9.1

BioGRID 3.2

4. public available database

NUTRALP VDA

September 2013 – October 2015

This research project aims to evaluate the nutraceutical properties and antioxidant activity of apple and grapevine.

Open Source Drug Discovery Platform (OSDD)

March 2013 – February 2015

A research project which aims to set up a first open-access informatics platform dedicated to the Open Source Drug Development, allowing software and data sharing in a protected, cloud-friendly environment;

VDNA Barcoding

March 2013 – February 2015

The Research Unit (UR) VDNA Barcoding buds from the idea of creating an advanced biotechnology center in the Aosta Valley that studies the alpine ecosystems using genomic analysis based on DNA sequencing and highly polymorphic molecular markers.

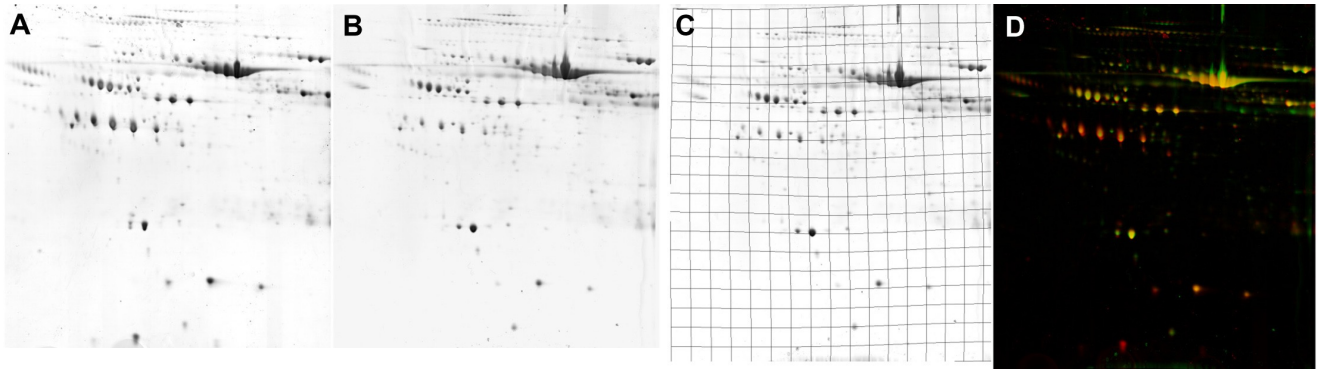
“Meglio a Casa”
September 2013 – September 2014
The goal of this project is to create an telemedicine and domiciliary care service for Parkinson’s Disease patients.

ParIS - Parkinson Information System
July 2010 – Genuary 2012
A research project devoted to the testing of a new procedure for finding perypheral biomarkers in Parkinson Disease.

IMAGE - Image Meta Analysis Generation and Exploitation
April 2009 – April 2010
A project dedicated to the meta-analysis of proteomic two dimensional gel electrophoresis (2D-GE) images in the literature.

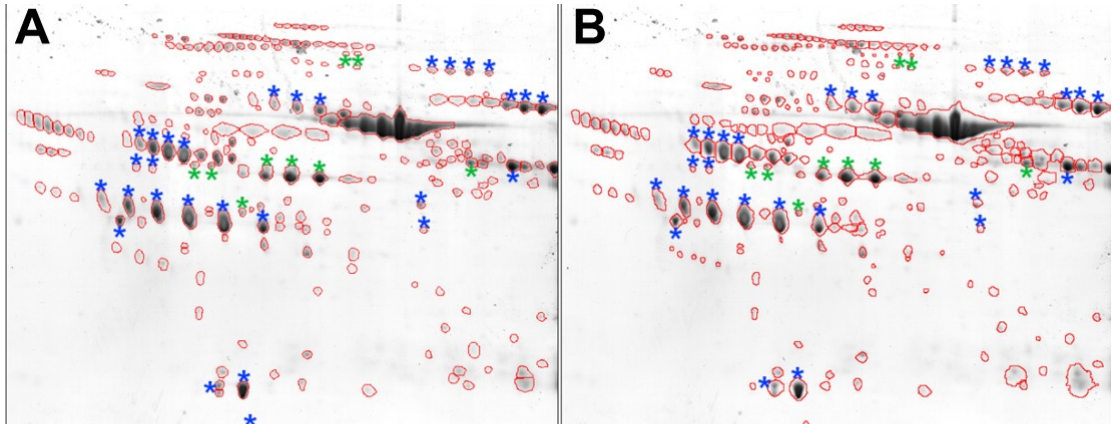
Open source 2D-GE analysis workflow, and automatic registration (IMAGE)

We implemented an new image analysis workflow using an open source Java library (ImageJ), in order to manage all the steps of a two dimensional gel electrophoresis (2D-GE) analysis.

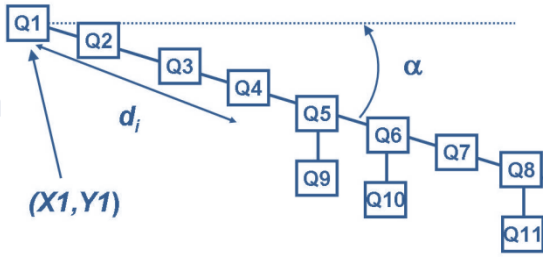


To test it, we performed a set of 2D-GE experiments on plasma samples from patients of acute myocardial infarction

We compared the results obtained by our procedure to those obtained using a widely diffuse commercial package, finding similar performances.



We have developed a 2D-GE method for matching annotated 2D-GE images extracted from publicly available papers. The method allow to localize a protein of interest by means of a deformable template.



Denoising: median filter to eliminate the salt and pepper noise; a cubic smoothing spline procedure and rolling ball algorithm for background subtraction.
Warping: a Levenberg-Marquardt minimization enhanced by a Broyden-Fletcher-Goldfarb-Shanno (BFGS)
Feature extraction (spot detection): Watershed algorithm
Automatic registration: a deformable template, composed of eleven squares (Qi) and described by ten parameters, is deformed by minimizing the energy function

Natale M, Maresca B, Abrescia P, Bucci Em (2011) Image Analysis Workflow for 2-D Electrophoresis Gels Based on ImageJ. In: PROTEOMICS INSIGHTS, vol. 4, pp. 37-49. - ISSN 1178-6418

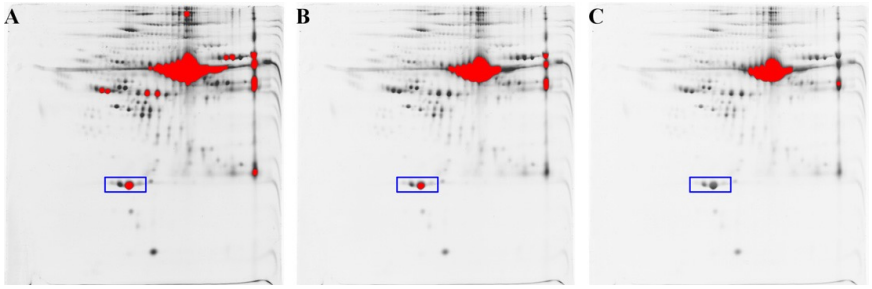
Rozza A., Arca S., Casiraghi E., Campadelli P., Natale M., Bucci E., Consoli P. (2011) Automatic Alignment of Gel 2D Images. In: Neural Nets WIRN11, pp. 3-10

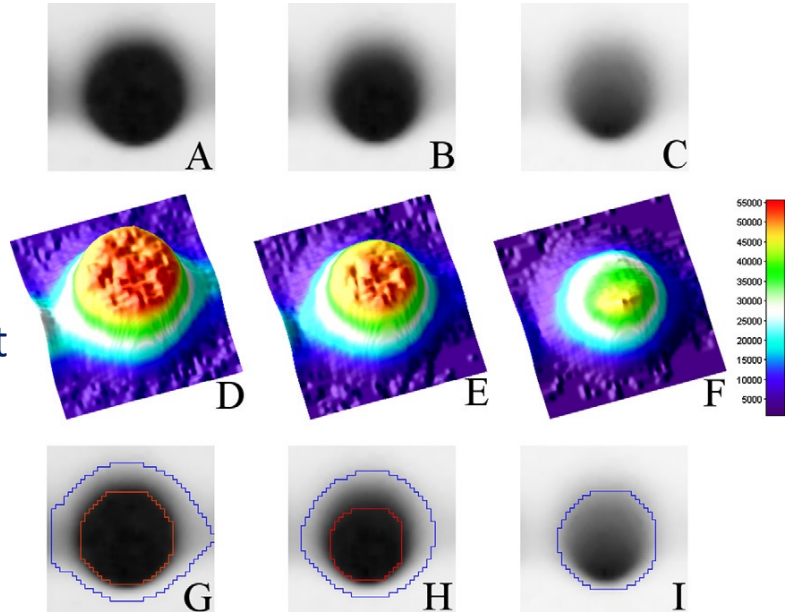
Novel Gaussian Extrapolation Approach for 2D-GE saturated spots (ParIS)

In order to enable the analysis of 2D-GE image extracted from biomedical literature and public repositories we developed an effective technique for the detection and the reconstruction of over-saturated protein spots. Pixel correction in saturated and smeared spots allows more accurate feature extraction (spot detection) allowing more reliable meta-analysis of 2D-GE images.

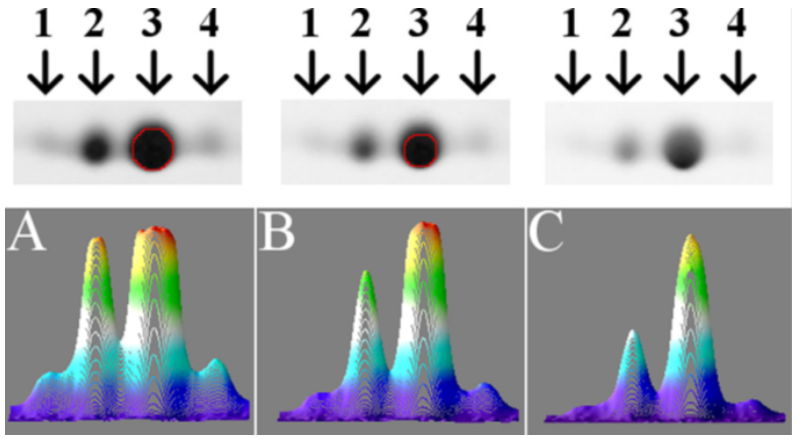
Firstly, the algorithm reveals overexposed areas, where spots may be truncated, and plateau regions caused by smeared and overlapping spots.

We implemented a morphological filter, inspired by the rolling-ball algorithm, where a standard element is moved along each scan line of the image.

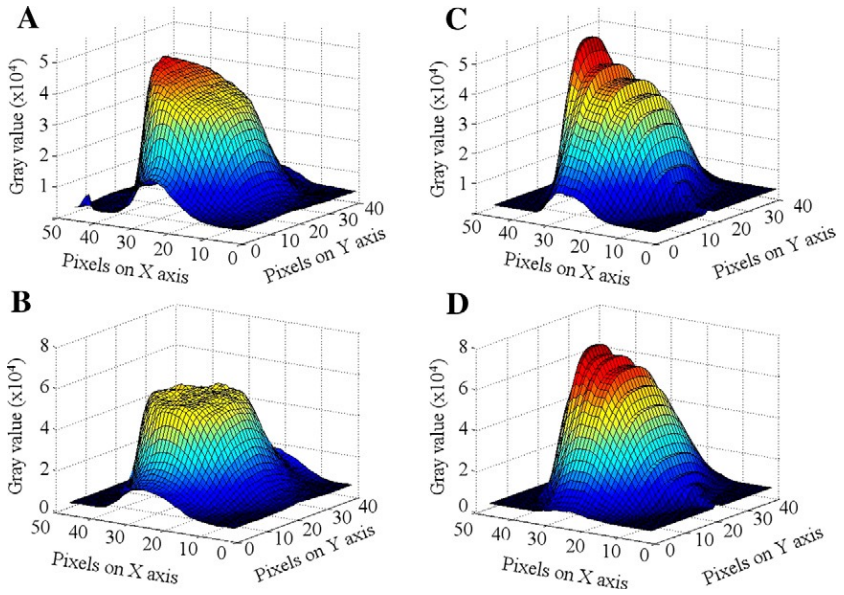




Next, the algorithm reconstructs the correct distribution of pixel values in these overexposed areas and plateau regions, using a two-dimensional least-squares fitting based on a generalized Gaussian distribution.



$$f(x, M(y), \sigma(y), x_0, b) = \frac{M(y)}{\sigma(y)} \exp\left(-\frac{|x - x_0(y)|^b}{b\sigma(y)^b}\right)$$

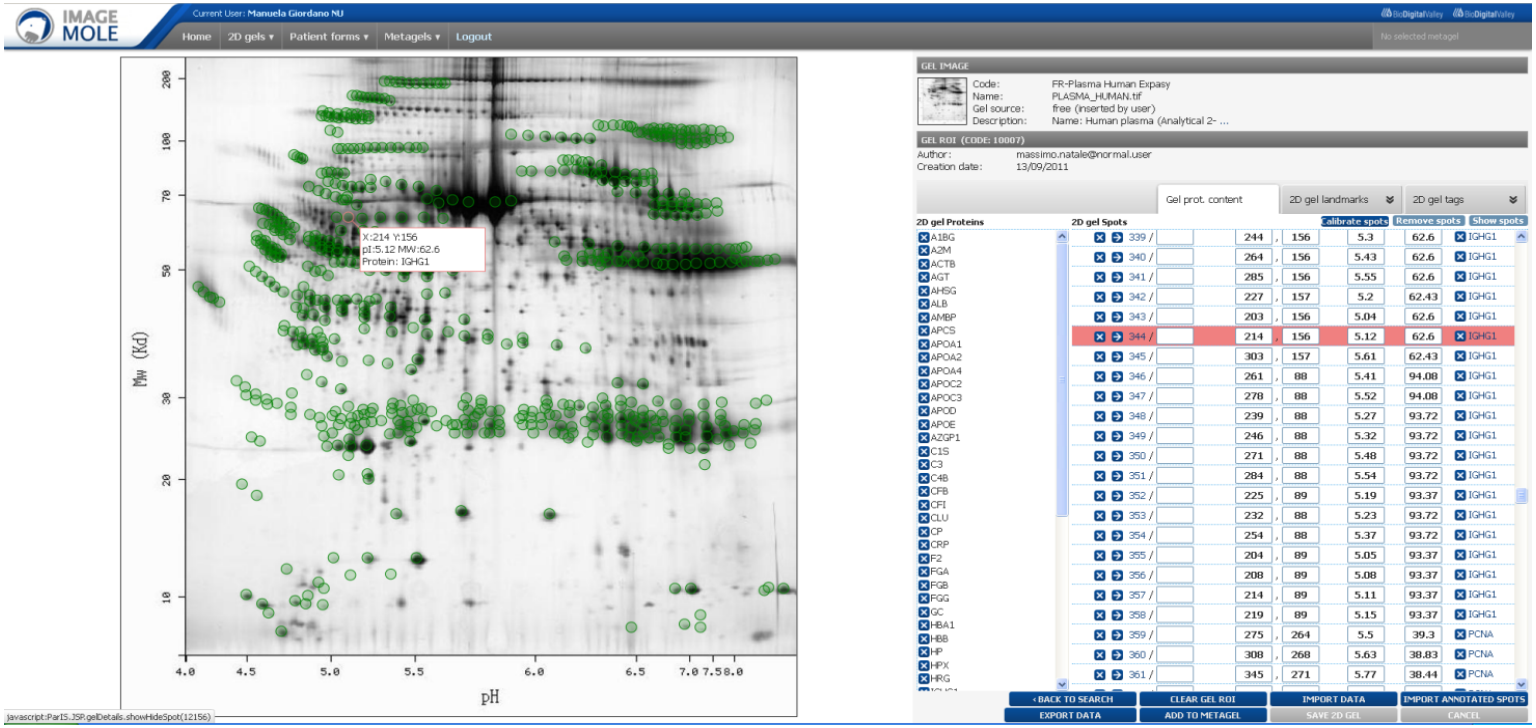


Natale M., Caiazzo A., Bucci E.M., Ficarra E. (2012) A Novel Gaussian Extrapolation Approach for 2D Gel Electrophoresis Saturated Protein Spots. In: GENOMICS, PROTEOMICS & BIOINFORMATICS, pp. 336-344. - ISSN 1672-0229

Natale M., Caiazzo A., Bucci E.M., Ficarra E. (2012) A novel Gaussian fitting approach for 2D gel electrophoresis saturated protein spots. International Conference on Bioinformatics Models, Methods and Algorithms, pp. 335-338

iMole Platform (ParIS)

We developed iMole a platform that automatically extracts images and captions from biomedical literature. Images are indexed using biomedical terms identified in figure captions. We used iMole to build a 2D-GE database, which yielded more than 16,500 images.



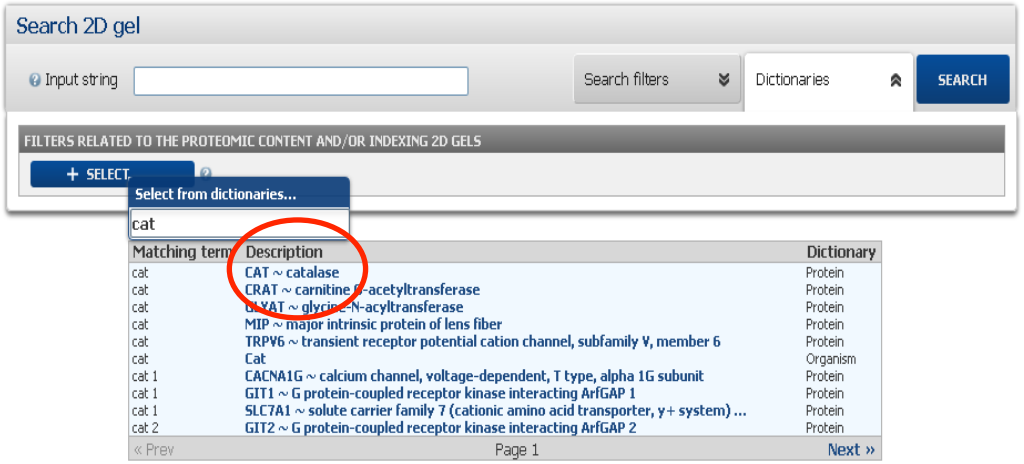
Technologies

Semantic: ontologies imported from open sources (Gene Ontology), thesaurus in house developed from Expasy and Entrez (PubMed) dictionary.

Text mining: name entity recognition, ambiguities in the terminology are resolved using multiple searches for more than one alias, as well as the co-occurrence of specific words that can either deny or force the tagging; supervised learning neural network (NN) developed using Weka. Tested OpenNLP, Stanford NLP, GATE (for NER).

Implementation: Java Server Pages (on Apache Tomcat) and MySQL database.

Giordano M., Natale M., Cornaz M., Ruffino A., Bonino D., Bucci E.M. (2013) iMole, a web based image retrieval system from biomedical literature. ELECTROPHORESIS, vol. 34, pp. 1965-1968. - ISSN 0173-0835



Proteomics biomarker validation (ParIS)

We developed an automated literature analysis procedure to retrieve all the background knowledge available in public databases. This work was devoted to the testing of a new procedure for validating peripheral biomarkers of Parkinson Disease, previously described in literature.

Core technologies

Text mining: as in iMole

Image analysis: as in open source 2D-GE analysis workflow

Data analysis: in R, we performed, non-parametric Wilcoxon test, Pearson linear correlation, linear discriminant analysis (LDA), receiver operating characteristic (ROC) curve.

Alberio T., Bucci E.M., Natale M., Bonino D., Di Giovanni M., Bottacchi E., Fasano M.(2013) Parkinson's disease plasma biomarkers: An automated literature analysis followed by experimental validation. JOURNAL OF PROTEOMICS, vol. 90, pp. 107-114. - ISSN 1874-3919

Bucci E.M., Natale M., Bonino D., Cornaz M., Gullusci M., Montagnoli L., Poli A., Ruffino A., Alberio T., Fanali G., Fasano M., Bottacchi E., Di Giovanni M., Gagliardi S., Cereda C. (2011) Parkinson Informative System (ParIS): a pipeline fo the evaluation and clinical validation of Parkinson' disease proteomic biomarkers. In: XLII Congress of the Italian Neurological Society. pp. 447-448

- Using text mining technologies we extract from PubMed database papers more than 51,000 scientific papers dealing with Parkinson Disease.
- We identify 4121 proteins cited in these papers.
- Out of these, we could track back 35 proteins describe as biomarker of Parkinson and present in at least two published 2-DE maps of human plasma.
- Among those biomarkers we identified 9 different proteins (haptoglobin, transthyretin, apolipoprotein A-1, serum amyloid P component, apolipoprotein E, complement factor H, fibrinogen γ, thrombin, complement C3) split into 32 spots as a potential diagnostic pattern.
- Some of the biomarker candidates, that arose from the literature analysis, were confirmed from a proteomics study on 90 subjects (45 Parkinson Disease patients, 45 non-neurodegenerative control subjects.

FunMod a Cystoscape plugin (OSDD)

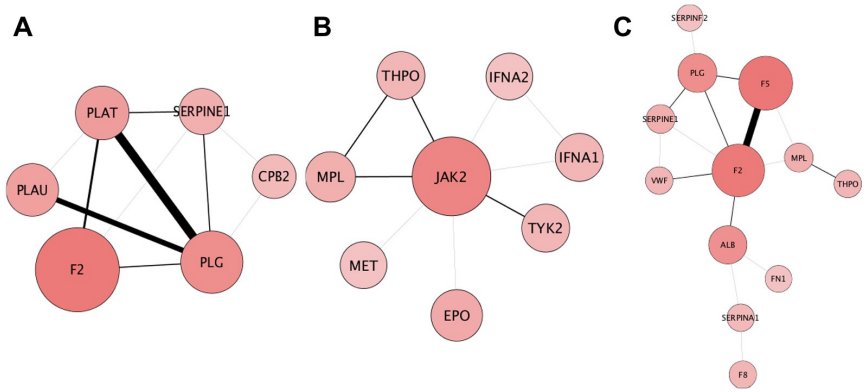
We developed FunMod an innovative Cytoscape plugin able to identify biologically significant sub-networks within informative protein networks, enabling new opportunities for elucidating pathways involved in diseases. FunMod has the ability of combining pathways and topological analysis allowing the identification of the key proteins within sub-network functional modules.

FunMod identifies within protein-protein network, subnetworks belonging to the same biological pathways.

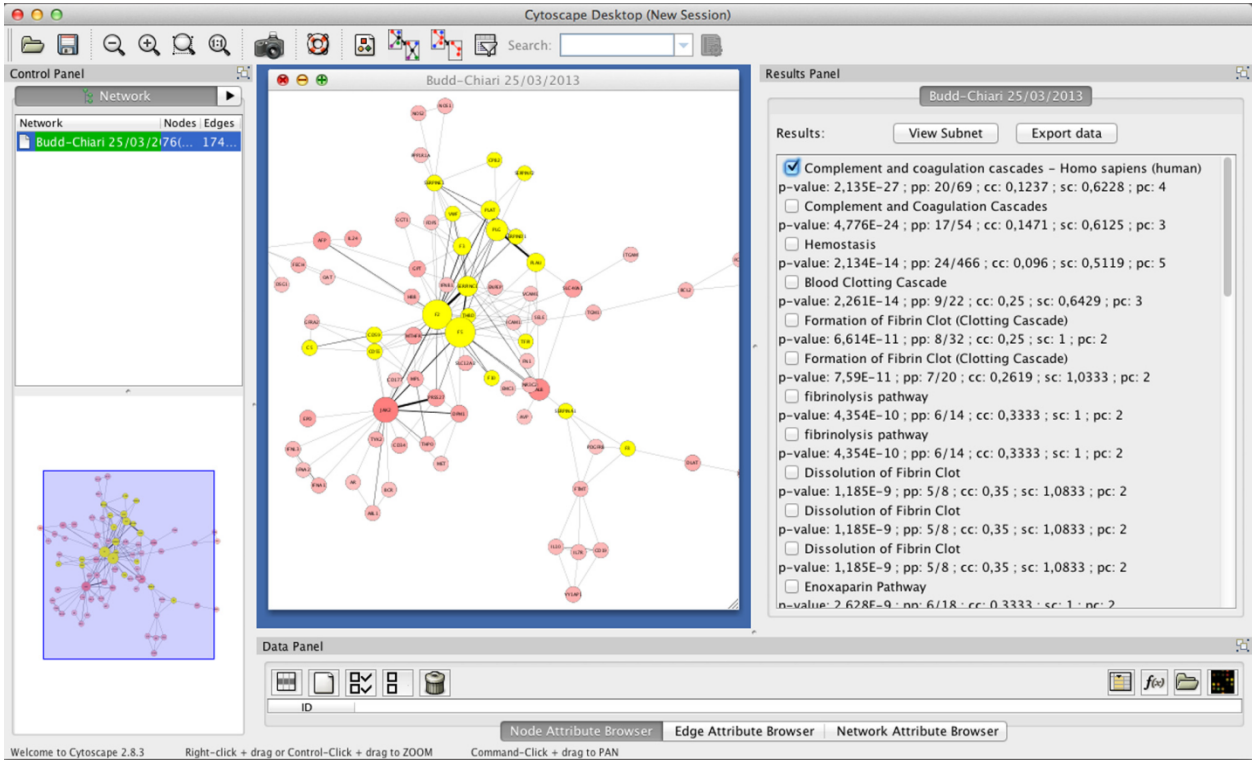
$$h\,(x; X, n, N) = \frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}}$$

The statistical significance of the sub-network is determined by performing a hypergeometric test,

It then analyzes the topology of the identified sub-network to infer the topological relations (motifs) of its nodes.



FunMod was developed as a Cytoscape plugin.
Cystoscape is the most widely used open source platform for network data integration, analysis, and visualization.



Natale M., Benso A., Di Carlo S., Ficarra E. (2014) FunMod: A Cytoscape Plugin for Identifying Functional Modules in Undirected Protein-Protein Networks. GENOMICS, PROTEOMICS & BIOINFORMATICS, vol. 12 n. 4, pp. 178-186. - ISSN 1672-0229

Bucci EM, Natale M, Poli A (2011) Protein Networks: Generation, Structural Analysis and Exploitation. In: Systems and Computational Biology - Molecular and Cellular Experimental Systems. INTECH, pp. 125-146. ISBN 9789533072807

PepSirio and PeptideHunter

We developed a new bioinformatics software which is able to analysed mass spectrometry data and identify the protein peptides from the mass spectrum which arise from. This software enables the identification and characterization of the oligopeptide fraction extracted from a cheese at different ages of ripening and subsequently identified by an in-source fragmentation detectable with a single-quadrupole mass analyzer.

Core technologies

The software performed a the brute force pattern matching algorithm, and for the front end we developed a Java Server Pages (JSP) web application and a Java Swing descktop application.

Valentini S., Natale M., Ficarra E., Barmaz A. (2012) New Software for the Identification and Characterization of Peptides Generated during Fontina Cheese Ripening Using Mass Spectrometry Data. JOURNAL OF CHEMISTRY AND CHEMICAL ENGINEERING, vol. 6 n. 4, pp. 323-326. - ISSN 1934-7375

Bioinformatics data analysis

Generation of literature-based networks was performed using ProteinQuest (PQ). Using PQ we found co-occurring proteins into both article abstracts and image captions. Connections mediated by a relevant biological concept (enhance/repress expression or activity) were used to create and extend a protein-rotein network. All such connections were controlled in order to verify the consistency between the retrieved literature data and the experimental results.

De Paula R.G., De Magalhães Ornelas A.M., Rezende Morais E., De Castro Borges W., Natale M., Guidi Magalhães L., Rodrigues V.(2014) Biochemical characterization and role of the proteasome in the oxidative stress response of adult Schistosoma mansoni worms. PARASITOLOGY RESEARCH, vol. 113, pp. 2887-2897. - ISSN 0932-0113

Gatti S., Leo C., Gallo S., Sala V., Bucci E.M., Natale M., Cantarella D., Medico E., Crepaldi T. (2013) Gene expression profiling of HGF/Met activation in neonatal mouse heart. TRANSGENIC RESEARCH, vol. 22, pp. 579-593. - ISSN 0962-8819

Publications

8 journal papers
3 conference proceeding
1 book chapter
1 patent* (Application 2009)

Goolge scholar metrics

Indici citazioni	All	from 2010
Citazioni	189	131
Indice H	6	6
i10-index	3	3

Impact factor

14,34	from 2011
34,01	all