POLITECNICO DI TORINO

Ph.D. School
Ph.D. in Computer and Control Engineering – XXVI cycle

Ph.D. Thesis

Computational tools for the interactive exploration of proteomics data and automatic bio-networks reconstruction

Massimo Natale

Tutor
Enrico Macii
Elisa Ficarra

Febbraio 2015

A Ilaria

I dwell in Possibility
A fairer House than Prose
More numerous of Windows
Superior for Doors

Of Chambers as the Cedars
Impregnable of Eyes
And for an Everlasting Roof
The Gambrels of the Sky

Of Visitors the fairest
For Occupation This
The spreading wide my narrow Hands
To gather Paradise

I dwell in Possibility
(Emily Dickinson)

Acknowledgements

Desidero innanzitutto ringraziare il Prof. Enrico Macii, tutore del mio percorso di Dottorato. La mia gratitudine va inoltre al Prof.ssa Elisa Ficarra con cui ho svolto buona parte del mio lavoro di ricerca e che mi ha spinto ad approfondire le tematiche relative all'analisi di immagine. Voglio anche esprimere la tutta mia gratitudine nei confronti del Prof. Alfredo Benso, per la collaborazione sul progetto di ricerca Open Source Drug DIscovery e per il costante e approfondito confronto avuto sulle tematiche della network analisi. Ringrazio anche l'Ing. Stefano Di Carlo, l'Ing. Alessandro Savino e l'Ing. Santa Di Cataldo per la collaborazione e l'aiuto che mi hanno dato durante il mio percorso di Dottorato.
Un grande ringraziamento all'Ing. Patrizia Iacovone, all'Ing. Marcello Bianchetti e a tutto il team di ITC Engineering Coordination di Unicredit per il supporto e la disponibilità dimostrata in questi mesi.
Ringrazio sinceramente i colleghi della Biodigitalvalley, Luca, Andrea, Moreno, Joey, Manuela, Cristina e Fausto per aver creduto e collaborato con passione ai progetti di ricerca condotti insieme.

Ringrazio con affetto i miei genitori Dino e Francesca per avermi insegnato che la dedizione e l'onestà intellettuale sono principi fondamentali nel lavoro come nella vita.

Come alla fine di ogni bella storia il lieto fine è d'obbligo. Il mio più grande abbraccio a Giorgia e Ilaria, sono la mia casa della possibilità.

Abstract

Revolutionary improvements in high-throughput technologies, also called 'omics' technologies, enable qualitative and quantitative monitoring of various biomolecule classes, providing broader insights into fundamental biological processes of living systems. In order to investigate, at the same time, a whole genome, transcriptome or proteome of a cell, a tissue or an organism, modern high-throughput instruments lead to the generation of a vast amounts of raw experimental data. For example, in genomics applications, next generation sequencing instruments can produce nearly 1 terabyte of data from each sample run.

Data creation in today's research is exponentially more rapid than anything we anticipated even a decade ago, and biomedical data generation is exceeding researchers' ability to capitalize on the data. Omics studies generating large amounts of data continue to proliferate, producing billions of data points and providing opportunities for the original researchers and other investigators to use these results in their own work to advance our knowledge of biology and biomedicine. Moreover, much of this information was collected within biomedical publications and in heterogeneous databases. The discovery and extraction of useful information from unstructured sources, as biomedical literature and public available database, are a trivial tasks, but necessary to enable a deep knowledge and understanding of the state of the art in a specific field of interest.

Based on these premises, the increasing availability of 'omics' data represents an unprecedented opportunity for bioinformatics researchers, but also a major challenge behind the need for novel systems biology approaches. The development of new approaches, software, and tools are requested to improve access to these data and store data, for its annotation and integration and stimulate the ability to make new discoveries using them.

Moreover, the value of 'omics' data is greatly enhanced when bioinformatics and systems biology strategies allow the integration of several data sources. In particular, a systems biology approach facilitates a multi-targeted approach, allows the integration of experimental and literature data, leading to a deeper understanding of physiologically complex processes and cellular functions.

The first objective of this dissertation is discuss the problems related to the management of high volume of experimental data, and how extract meaningful informations form biomedical literature and other open source of biomedical data.

Later this dissertation describe the bioinformatics tool and software that can store interactively and neatly proteomics data, perform analysis and meta-analyses for obtaining new insights and understanding from the large amount of data generated in high-throughput screening.

Finally, this dissertation aims to provide evidence of the effectiveness of systems biology approaches to integrate experimental 'omics' data and informations form biomedical literature.

Summary

The information included in this dissertation is fully self-contained and does not assume any prior knowledge of specific aspects of bioinformatics on the part of the reader.

A reader with minimum understanding of data and image analysis be able to read through the whole dissertation with ease, and gradually build up the information necessary to understand the described concepts.

The thesis is organized as follows:

Chapter 1 briefly describes high-throughput technologies, a set of technologies widely used for understanding the behavior of cells, tissues, organs, and the whole organism at the molecular level using methods such as genomics, proteomics. This chapter also introduces the systems biology and bioinformatics tools needed to analyze and make sense of the 'omics' data. These technologies have the potential to extract new results from the large amount of 'omics' data. This chapter discusses why the 'omics' data should be considered as Big Data, and bioinformatics as the data science applied into biomedical scenario.

Chapter 2 describes the problem of the management of Big Data in the healthcare and biomedical sector. Large volume of data, produced with high velocity from a high number of various types of sources, is now relevant for leaders across every sector of healthcare services. Improve data collection and analysis IT infrastructure has the potential to facilitate the efficiency and effectiveness of the whole health care sector, and facilitate the transfer of research results into clinical applications.

Chapter 3 presents an overview of the emerging solutions to deal with for Big Data analysis in biomedical field. New IT infrastructures are needed to store and access data, and new paradigms of semantic representation are requested for its annotation and integration. Provide the proper tools and resources to manage Big Data is one of the biggest challenge that IT researchers face today.

Chapter 4 discusses the reproducibility problems in biomedical research. Recent studies shown that experimental findings from several scientific papers cannot be reliably reproduced by other researchers. At the same time some proteomics authors have reported that the differentially expressed proteins, commonly observed in 2-DE publications, represent common cellular stress responses and are a reflection of the technical limitations of 2-DE. This chapter discusses the various factors that

contribute to the problem, as statistical mistakes or bias in the data sources, and how bioinformatics and systems biology enable a deep knowledge and understanding of the state of the art in a specific field of interest. A methodological analysis and comprehension of the whole biomedical data is crucial to provide solid theoretical basis for proposing studies, interpret the outcome of experiments.

Chapter 5 presents a tool for extracting images and text from biomedical literature. This method will be based on the simultaneous analysis of scientific papers, biomedical thesaurus and ontologies. This step reaches the dual objective of making a massive search of all information related to specific proteins through both text and images in literature. Since, in the biomedical research community, much attention is drawn by figures because often summarize the findings of the research work, provide a computational tool able to mine image data is a central task during the implementation of a software suite for the interactive exploration of proteomics data.

Chapter 6 introduces the concept of meta-analysis and its application as a new tool for assessing 2D-GE images extracted from proteomics papers and publicly available databases. Most proteomics studies move to identify specific two-dimensional electrophoresis (2-DE) pattern of proteins specifically related to a physiological or pathological condition. However, the information arising from these investigations is often incomplete due to inherent limitations of the technique, to extensive protein post-translational modifications and sometimes to the paucity of available samples. The meta-analysis of proteomic data can provide valuable information pertinent to various biological processes that otherwise remains hidden. This chapter shows a meta-analysis of the Parkinson Disease protein DJ-1 in heterogeneous 2-DE experiments. The protein was shown to segregate into specific clusters associated with defined conditions. Interestingly, these results were experimentally validated on human brain specimens from control subjects and Parkinson Disease patients.

Chapter 7 presents image strategies able to analyze two dimensional gel electrophoresis (2D-GE) image. The first part of the chapter describes an ImageJ-based procedure able to manage all the steps of a 2D-GE gel analysis. ImageJ is a free available image analysis application, developed by National Institutes of Health (NIH) and provides an open source alternative to commercial software allowing all researchers to developed meta-analyses of 2D-GE images. The second part of the chapter describes an image analysis processing procedure for detection and reconstruction of over-saturated protein spots, a common problems of image downloaded from web repositories. Aims of this chapter is provide an example of how open source tools might support the image analysis of proteomics studies.

Chapter 8 discusses the biological networks analysis topic. In recent years biological networks have attracted a lot of interest within the scientific community and many methods have been introduced for their inference. This Chapter discusses the perspective of biological network analysis and the way in which the results could be ported to a clinical context. The procedure, presented in this chapter, extracts information from unstructured text data and analyse network combining biological ontologies and network topology data.

Contents

Chapter 1

The 'omics' technologies and the big data problem

The modern high-throughput technologies allow a large or even exhaustive number of measurements that can be taken in a fairly short time period. These technologies are leading to the generation of a copious amounts of data at multiple levels of biology from gene sequence and expression to protein and metabolite patterns, and are substantially changing the face of biomedical and life science research. This signals a new era in how we approach to scientific inquiries.
Nowadays, wide data can be collected in an 'omics' experiment 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 the bioinformatics tools and computational approaches that are required to extract value and generate new biological understanding from the huge volume and diversity of bioscience data now available and so underpin and enable biological research as it continues to evolve as a data intensive discipline.

1.1 High-throughput technologies

High-throughput technologies allow to measure within a single analysis, several features of a large family of cellular molecules, such as genes, proteins, or small

metabolites (Idris et al., 2013). High-throughput technologies were introduced from the Human 'Genome' Project since 1990s (Naido et al., 2011), and have been named by appending the suffix 'omics'. The suffix 'ome', etymologically derived from the Sanskrit OM, describe the ability of the high-throughput technologies to perform complete and comprehensive analysis(Özdemir, 2013). By combining 'gene' and 'ome' Hans Winkler (1920) created the term genome, and Victor McKusick and Frank Ruddle added 'genomics' to the scientific lexicon as the title for the new journal they co-founded in 1987, with emphasis on linear gene mapping, DNA sequencing and comparison of genomes from different species (McKusick and Ruddle, 1987).

Nowadays 'omics' refers to the collective technologies used to explore the roles, relationships, and actions of the various types of molecules that make up the cells of an organism. These technologies include:

- Genomics: the study of genes and their function;
- Proteomics; the study of proteins;
- Metabonomics: the study of molecules involved in cellular metabolism;
- Transcriptomics: the study of the mRNA;
- Glycomics: the study of cellular carbohydrates;
- Lipomics: the study of cellular lipids;

'Omics' technologies and various neologisms that define their application contexts, however, are more than a simple play on words. They substantially transformed both the throughput and the design of scientific experiments. 'Omics' technologies provide the tools needed to look at the differences in DNA, RNA, proteins, and other cellular molecules between species and among individuals of a species (see, Figure 1).

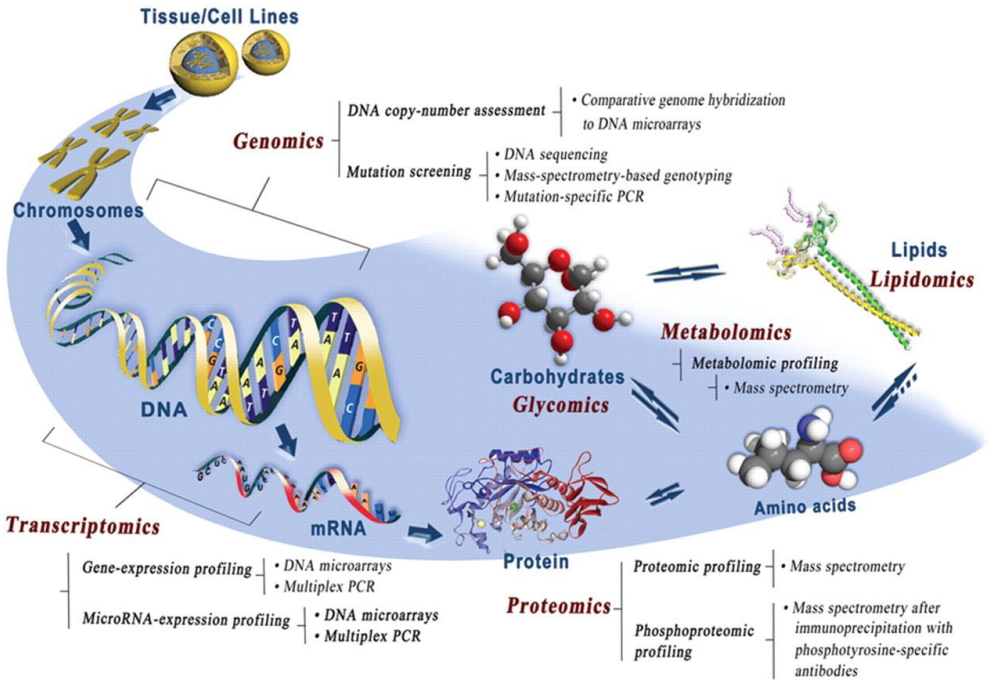

Figure 1. Schematic of the 'omic hierarchy: genomics, transcriptomics, proteomics, metabolomics, glycomics and lipodomics.

These types of molecular profiles can vary with cell or tissue exposure to chemicals or drugs and thus have potential use in toxicological assessments. These new methods have already facilitated significant advances in our understanding of the molecular responses to cell and tissue damage, and of perturbations in functional cellular systems (Aardema and MacGregor, 2002).

1.2 'omics' science and technologies

The 'omics' science and technologies improve the simplistic and reductionist experimental models that offer merely a temporal snap shot of the much more complex, longitudinal and dynamic nature of molecule interactions (and their fluctuations in response to social/environmental exposures) that fundamentally govern human health and disease. The process of research is fundamentally altered in 'omics' science. Ordinarily, scientists have accustomed to hypothesis-driven research wherein a clearly articulated scientific question/hypothesis would be posed (Ozdemir et al. 2009). Subsequently experiments would be carried out to obtain data in order to test the study hypothesis. With the 'omics' approach, asking an initial research question is not always necessary or a pre-requisite. Genome or proteome wide data can be collected in an 'omics' experiment without an existing hypothesis, followed by generation and testing of biological hypotheses. This reversal from the 'first hypothesize-then-experiment' tradition to 'first experiment-then-hypothesize' mode of operation offers the promise to discover unprecedented pathophysiological mechanisms of disease as well as response and toxicity to drugs and nutrition.

'Omics' experiments are conducted thanks to high-throughput measurement technologies, in which a large or even exhaustive number of measurements can be taken in a fairly short time period, leading to the generation of a copious amounts of data at multiple levels of biology from gene sequence and expression to protein and metabolite patterns underlying variability in cellular networks and function of whole organ systems. In fact this led to overabundance of data in biomedical experiments recently. This signals a new era in how we approach to scientific inquiries. That is, the arrival of 'big biology' and a systems (integrative) approach to scientific practice with global measurements of molecular pathways in health and disease.

Nowadays, 'omics' experiments provide a huge amount of molecular measurements for each single experiment so now on of the main challenge is to develop a bioinformatics strategy that uses the 'omics' measurements to predict a clinical outcome of interest, such as disease status, survival time, or response to therapy (Kitano, 2002). Bioinformatics is used to abstract knowledge and principles from large-scale data, to present a complete representation of the cell and the organism, and to predict computationally systems of higher complexity, such as the interaction networks in cellular processes and the phenotypes of whole organisms. Bioinformatics tools include several computational tools able to mine information from large databases of biological data (Huang et al., 2012). These tools are most

commonly used to analyze large sets of omics data. Bioinformatics and databases of biological information can be used to generate biological networks of cellular and physiological pathways and responses. This integrative approach is called systems biology (O'Brien EJ and Palsson, 2015). Systems Biology is an integration of data from all levels of complexity genomics, proteomics, metabolomics, and other molecular mechanisms using advanced computational methods to study how networks of interacting biological components determine the properties and activities of living systems (de Vargas and Claassen, 2014). The goal is to create overall computational models of the functioning of the cell, multicellular systems, and ultimately the organism. These in silico models will provide virtual test systems for evaluating the responses of cells, tissues, and organisms to diseases, therapies or other pathological conditions.

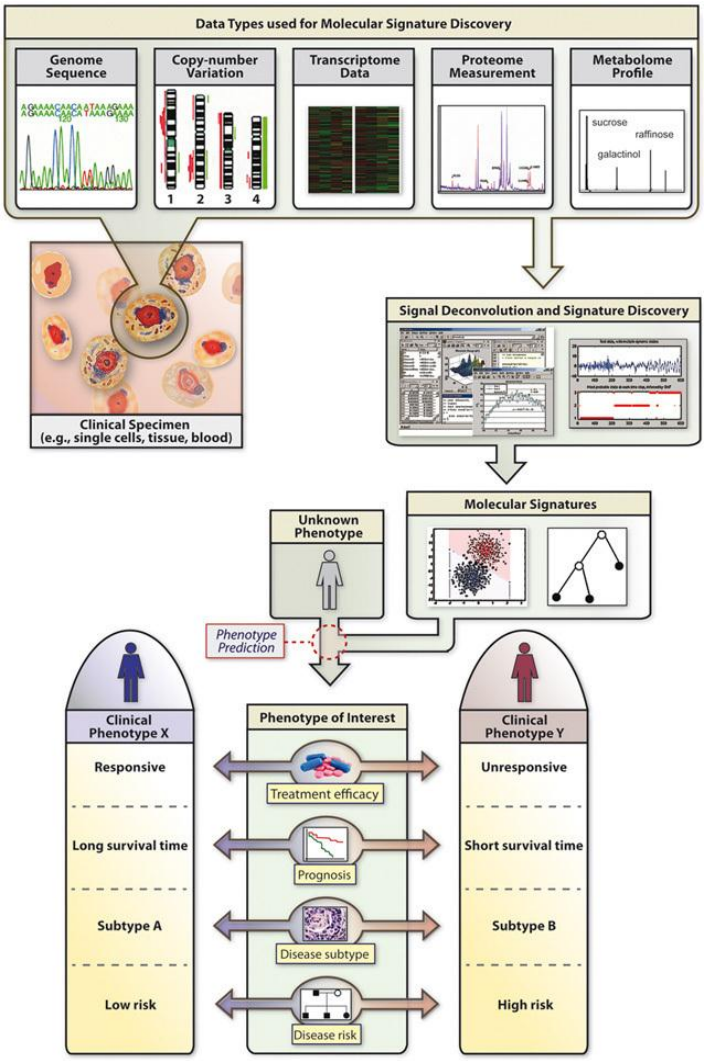

Figure 2. Overview of the discovery and application of molecular signatures from omics data. Molecular signatures can be derived from a broad range of omics data types (e.g. DNA sequence, mRNA, and protein expression) and can be used to predict various clinical phenotypes (e.g. response to therapy, prognosis) for previously unseen patient specimens

Bioinformatics data mining is a fundamental step within the process of using 'omics' data to discover a molecular signature. A molecular signature is as a set of

biomolecular features (e.g. DNA sequence, DNA copy number, RNA, protein, and metabolite expression) together with a predefined computational procedure that applies those features to predict a phenotype of clinical interest on a previously unseen patient sample (Sung et al., 2012). A signature can be based on a single data type or on multiple data types. The overall process of identifying molecular signatures from various omics data types for a number of clinical applications is summarized in Figure 2.

These increasing availability of OMICS data represents an unprecedented opportunity for bioinformatics researchers. A similar scenario arises for the healthcare systems, where the digitalization of all clinical exams and medical records is becoming a standard in hospitals. Such huge and heterogeneous amount of digital information, nowadays called Big Data, is the basis for uncovering hidden patterns in data, since it allows the creation of predictive models for real-life biomedical applications. But the main issue is the need of improved technological solutions to deal with them.

Chapter 2

Big data in healthcare

Data have become a torrent flowing into every area of the modern life. Companies churn out a burgeoning volume of transactional data, capturing trillions of bytes of information about their customers, suppliers, and operations. Millions of networked sensors are being embedded in the physical world in devices such as mobile phones, smart energy meters, automobiles, and industrial machines that sense, create, and communicate data in the age of the Internet of Things. Indeed, as companies and organizations go about their business and interact with individuals, they are generating a tremendous amount of digital "exhaust data," data that are created as a by-product of other activities.

Social media sites, smartphones, and other consumer devices including PCs and laptops have allowed billions of individuals around the world to contribute to the amount of big data available. Like other essential factors of production such as hard assets and human capital, it is increasingly the case that much of modern economic activity, innovation, and growth simply couldn't take place without data.

Many citizens around the world regard this collection of information with deep suspicion, seeing the data flood as nothing more than an intrusion of their privacy. But there is strong evidence that big data can play a significant role to the benefit not only of private pharmaceutical firms but also of national healthcare systems and their citizens.

The chapter is structured as follows. This chapter seeks to understand the state of digital data, how different domains can use large datasets to create value, the potential value across healthcare stakeholders, and the implications for the biomedical research and can create significant value for the healthcare.

Section 1 discusses the use of simulations and modelling on preclinical or early clinical datasets, and in Section 2 the statistical tools and algorithms. Section 3 presents how to identify additional indications and discover adverse effects analysing

clinical trials data and patient records, Section 4 presents the promising application on personalized medicine.

Digital data is now everywhere and big data is now relevant for leaders across every sector, and consumers of products and services stand to benefit from its application (Tene and Polonetsky, 2012). Big data refers to huge data sets that are orders of magnitude larger (volume), more diverse, including structured, semistructured, and unstructured data (variety), and arriving faster (velocity)(Howe, 2008)(Schilling PL and Bozic, 2014) that exceed an organization's storage or compute capacity for accurate and timely decision making (Mayer-Schönberger and Cukier, 2013). However, big data is defined less by volume, which is a constantly moving target, than by its ever-increasing variety, velocity, variability and complexity (Boyd and Crawford, 2012). These four factors help define the major issues that IT[1] needs to address:

Variety. Up to 80 percent of biomedical data is unstructured, not numeric, and embedded in plain text but it still must be folded into quantitative analysis and decision making (Tao et al., 2013). Text unstructured data require different architecture and technologies for analysis.

Velocity. Thornton May says, "Initiatives such as the use of RFID[2] tags and smart metering are driving an ever greater need to deal with the torrent of data in near- real time. This, coupled with the need and drive to be more agile and deliver insight quicker, is putting tremendous pressure on organizations to build the necessary infrastructure and skill base to react quickly enough."

Variability. In addition to the speed at which data comes your way, the data flows can be highly variable – with daily, seasonal and event-triggered peak loads that can be challenging to manage.

Complexity. Difficulties dealing with data increase with the expanding universe of data sources and are compounded by the need to link, match and transform data across business entities and systems. Organizations need to understand relationships, such as complex hierarchies and data linkages, among all data.

For instance, if United States healthcare could use big data creatively and effectively to drive efficiency and quality, we estimate that the potential value from data in the sector could be more than $300 billion in value every year, two-thirds of which would be in the form of reducing national health care expenditures by about 8 percent (Bromer et al., 2011).

The US health care system has four major pools of data within health care, each

[1] IT: Information technology (IT) is the application of computers and telecommunications equipment to store, retrieve, transmit and manipulate data, often in the context of a business or other enterprise.
[2] RFID: Radio-frequency identification (RFID) is the wireless use of electromagnetic fields to transfer data, for the purposes of automatically identifying and tracking tags attached to objects.

primarily held by a different constituency (Jog, 2012). Data are highly fragmented in this domain. The four pools are provider clinical data, payor activity (claims) and cost data, pharmaceutical and medical products R&D[3] data, and patient behavior and sentiment data. The amount of data that is available, collected, and analyzed varies widely within the sector. For instance, health providers usually have extensively digitized financial and administrative data, including accounting and basic patient information. In general, however, providers are still at an early stage in digitizing and aggregating clinical data covering such areas as the progress and outcomes of treatments. Depending on the care setting, we estimate that as much as 30 percent of clinical text/numerical data in the United States, including medical records, bills, and laboratory and surgery reports, is still not generated electronically. Even when clinical data are in digital form, they are usually held by an individual provider and rarely shared. Indeed, the majority of clinical data actually generated are in the form of video and monitor feeds, which are used in real time and not stored.

The pharmaceutical and medical products (PMP) subsector is arguably the most advanced in the digitization and use of data in the health care sector. PMP captures R&D data digitally and already analyzes them extensively. Additional opportunities could come from combining PMP data with other datasets such as genomics or proteomics data for personal medicine, or clinical datasets from providers to identify expanded applications and adverse effects. In addition to clinical, activity (claims) cost data, and pharmaceutical R&D datasets, there is an emerging pool of data related to patient behaviour (e.g., propensity to change lifestyle behaviour) and sentiment (e.g., from social media) that is potentially valuable but is not held by the health care sector. Patient behaviour and sentiment data could be used to influence adherence to treatment regimes, affect lifestyle factors, and influence a broad range of wellness activities.

It will be imperative for organizations, and possibly policy makers, to figure out how to align economic incentives and overcome technology barriers to enable the sharing of data. The rearcher identified a set of levers that have the potential to improve the efficiency and effectiveness of the health care sector by exploiting the tremendous amount of electronic information that is, and could become, available throughout the US health care sector.

2.1 Predictive modelling

The first lever is the aggregation of research data so that PMP companies can perform predictive modelling for new drugs and determine the most efficient and cost-effective allocation of R&D resources. This “rational drug design” means using simulations and modelling based on preclinical or early clinical datasets along the

[3] R&D: Research and development (R&D) is a general term for an activities related to the enterprise of corporate or governmental innovation.

R&D value chain to predict clinical outcomes as promptly as possible. The evaluation factors can include product safety, efficacy, potential side effects, and overall trial outcomes. This predictive modeling can reduce costs by suspending research and expensive clinical trials on suboptimal compounds earlier in the research cycle.

The benefits of this lever for the PMP sector include lower R&D costs and earlier revenue from a leaner, faster, and more targeted R&D pipeline. The lever helps to bring drugs to market faster and produce more targeted compounds with a higher potential market and therapeutic success rate. Predictive modelling can shave 3 to 5 years off the approximately 13 years it can take to bring a new compound to market.

2.2 Statistical tools and algorithms to improve clinical trial design

Another lever is using statistical tools and algorithms to improve the design of clinical trials and the targeting of patient recruitment in the clinical phases of the R&D process (Raghupathi W and Raghupathi, 2014). This lever includes mining patient data to expedite clinical trials by assessing patient recruitment feasibility, recommending more effective protocol designs, and suggesting trial sites with large numbers of potentially eligible patients and strong track records. The techniques that can be employed include performing scenario simulations and modelling to optimize label size (the range of indications applicable to a given drug) to increase the probability of trial success rates. Algorithms can combine R&D and trial data with commercial modelling and historic regulatory data to find the optimal trade-off between the size and characteristics of a targeted patient population for trials and the chances of regulatory approval of the new compound. Analyses can also improve the process of selecting investigators by targeting those with proven performance records.

2.3 Analyzing clinical trials data

A third R&D-related lever is analysing clinical trials data and patient records to identify additional indications and discover adverse effects. Drug repositioning, or marketing for additional indications, may be possible after the statistical analysis of large outcome datasets to detect signals of additional benefits. Analysing the (near) real-time collection of adverse case reports enables pharmacovigilance, surfacing safety signals too rare to appear in a typical clinical trial or, in some cases, identifying events that were hinted at in the clinical trials but that did not have sufficient statistical power.

These analytical programs can be particularly important in the current context in

which annual drug withdrawals hit an all-time high in 2008 and the overall number of new drug approvals has been declining. Drug withdrawals are often very publicly damaging to a company. The 2004 removal of the painkiller Vioxx[4] from the market resulted in around $7 billion in legal and claims costs for Merck and a 33 percent drop in shareholder value within just a few days (Fielder, 2005).

2.4 Personalized medicine

Another promising big data innovation that could produce value in the R&D arena is the analysis of emerging large datasets (e.g., genome data) to improve R&D productivity and develop personalized medicine (Murdoch and Detsky, 2013). The objective of this lever is to examine the relationships among genetic variation, predisposition for specific diseases, and specific drug responses and then to account for the genetic variability of individuals in the drug development process (Sarachan et al., 2003).

Personalized medicine holds the promise of improving health care in three main ways: offering early detection and diagnosis before a patient develops disease symptoms; more effective therapies because patients with the same diagnosis can be segmented according to molecular signature matching (i.e., patients with the same disease often don't respond in the same way to the same therapy, partly because of genetic variation); and the adjustment of drug dosages according to a patient's molecular profile to minimize side effects and maximize response(Garrison and Austin, 2006).

Personalized medicine is in the early stages of development. Impressive initial successes have been reported, particularly in the early detection of breast cancer, in prenatal gene testing, and with dosage testing in the treatment of leukaemia and colorectal cancers. Experts estimated that the potential for cost savings by reducing the prescription of drugs to which individual patients do not respond could be 30 to 70 percent of total cost in some cases. Likewise, earlier detection and treatment could significantly lower the burden of lung cancer on health systems, given that early-stage surgery costs are approximately half those of late-stage treatment (Swan, 2012).

[4] Vioxx: Rofecoxib is a nonsteroidal anti-inflammatory drug that has now been withdrawn over safety concerns. It was marketed by Merck & Co. to treat osteoarthritis, acute pain conditions, and dysmenorrhoea. Rofecoxib was approved by the Food and Drug Administration in 1999, and was marketed under the brand names Vioxx, Ceoxx, and Ceeoxx. Worldwide, over 80 million people were prescribed rofecoxib at some time. On September 30, 2004, Merck withdrew rofecoxib from the market because of concerns about increased risk of heart attack and stroke associated with long-term, high-dosage use. Merck withdrew the drug after disclosures that it withheld information about rofecoxib's risks from doctors and patients for over five years, resulting in between 88,000 and 140,000 cases of serious heart disease. Rofecoxib was one of the most widely used drugs ever to be withdrawn from the market (source at http://en.wikipedia.org/wiki/Rofecoxib).

Chapter 3

Big data strategy

Simple definition of Big Data, introduced in the previous chapter, is based on the concept of data sets whose size is beyond the management capabilities of typical relational database software. This definition of Big Data is based on the three 'V' paradigm[1]: volume, variety, and velocity. The volume recalls for novel storage scalability techniques and distributed approaches for information query and retrieval. The second V, the variety of the data source, prevents the straightforward use of neat relational structures. Finally, the increasing rate at which data is generated, the velocity, has followed a similar pattern as the volume. This "need for speed," particularly for web-related applications, has driven the development of techniques based on key-value stores and columnar databases behind portals and user interfaces, because they can be optimized for the fast retrieval of precomputed information. Thus, smart integration technologies are required for merging heterogeneous resources: promising approaches are the use of technologies relying on lighter placement with respect to relational databases (i.e., NoSQL databases[2]) and the exploitation of semantic and ontological annotations.

The chapter is structured as follows. In Section 1 architectural solutions for Big Data are described, paying particular attention to the needs of the bioinformatics community. Section 2 presents parallel platforms for Big Data elaboration, while Section 3 is concerned with the approaches for data annotation, specifically

[1] V paradigm: in a 2001 research report and related lectures the analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.

[2] NoSQL database: A NoSQL database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.

considering the methods employed in the computational biology field. Section 4 introduces data access measures and security for biomedical data. Finally, Section 5 presents some conclusions and future perspective.

Although the Big Data ecosystem can still be considered quite nebulous, it does not represent just a keyword for researchers or an abstract problem: the United States Administration launched a 200 million dollar "Big Data Research and Development Initiative" in March 2012[3], with the aim to improve tools and techniques for the proper organization, efficient access, and smart analysis of the huge volume of digital data. Such a high amount of investments is justified by the benefit that is expected from processing the data, and this is particularly true for omics science.

A meaningful example is represented by the projects for population sequencing. The first one is the 1000 genomes[4] (Haraksingh and Snyder, 2013), which provides researchers with an incredible amount of raw data. Then, the ENCODE project[5] (Graur et al., 2015), a follow-up to the Human Genome Project[6] (Genomic Research), is having the aim of identifying all functional elements in the human genome (Duncan et al., 2014). Presently, this research is moving at a larger scale, as clearly appears considering the Genome 10K project[7] and the more recent 100K Genomes Project[8]. Just to provide an order of magnitude, the amount of data produced in the context of the 1000 Genomes Project is estimated in 100 Terabytes (TB), and the 100K Genomes Project is likely to produce 100 times such data. The targeting cost for sequencing a single individual will reach soon $1000, which is affordable not only for large research projects but also for individuals. The explosion of data is leading into the paradox that the cheapest solution to cope with these data will be to resequence genomes when analyses are needed instead of storing them for future reuse (Merelli et

[3] Big Data Research and Development Initiative: the Obama Administration launched the initiative in order to improve the researchers ability to extract knowledge and insights from large and complex collections of digital data. Additional information at http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal.

[4] 1000 Genomes: a deep catalogue of human genetic variation (http://www.1000genomes.org/).

[5] ENCODE Project: the Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active. (https://www.encodeproject.org/)

[6] Human Genome Project: The Human Genome Project (HGP) was one of the great feats of exploration in history, an international research effort to sequence and map all of the genes, together known as the genome, of members of our species, Homo sapiens. Completed in April 2003, the HGP gave us the ability, for the first time, to read nature's complete genetic blueprint for building a human being (http://www.genome.gov/10001772).

[7] Genome 10K Project: https://genome10k.soe.ucsc.edu/.

[8] To bring the predicted benefits of genomics to NHS patients is why the England Prime Minister launched the 100,000 Genomes Project in late 2012. Genomics England, a company wholly owned and funded by the Department of Health, was set up to deliver this flagship project which will sequence 100,000 whole genomes from NHS patients by 2017 (http://www.genomicsengland.co.uk/the-100000-genomes-project/).

al., 2014).

Storage represents only one side of the medal. The final goal of research activities in omics sciences is to turn such amount of data into usable information and real knowledge. Biological systems are very complex, and consequently the algorithms involved in analysing them are very complex as well. They still require a lot of effort in order to improve their predictive and analytical capabilities (Kirschner et al., 2014). The real challenge is represented by the automatic annotation and/or integration of biological data in real-time, since the objective to reach is to understand them and to achieve the most important goal in bioinformatics: mining information.

3.1 Big Data Architectures

Domains concerned with data-intensive applications have in common the above mentioned three 'V', even though the actual way by which this information is acquired, stored, and analysed can vary a lot from field to field. The main common aspect is represented by the requirements for the underlying IT architecture (Villars et al., 2011). The mere availability of disk arrays of several hundreds of Terabytes is in fact not sufficient, because the access to the data will have, statistically, some fails. Thus, reliable storage infrastructures have to be robust with respect to these problems(Chen and Zhang, 2014).

Moreover, the analysis of Big Data needs frequent data access for the analysis and integration of information, resulting in considerable data transfer operations. Even though we can assume the presence of a sufficient amount of bandwidth inside a cluster, the use of distributed computing infrastructure requires adopting effective solutions. Other aspects have also to be addressed, as secure access policies to both the raw data and the derived results. Choosing a specific architecture and building an appropriate Big Data system are challenging because of diverse heterogeneous factors. In this scenario the leading solutions that are adopted form bioinfomatics institution are the open source distributions (Jee and Kim, 2013).

3.2 Managing and Accessing Big Data

The first and obvious concern with Big Data is the volume of information that researchers have to face, especially in bioinformatics. At lower level this is an IT problem of file systems and storage reliability, whose solution is not obvious and not unique (Bohlouli et al., 2013). Open questions are what file system to choose and will the network be fast enough to access this data. The issue arising in data access and retrieval can be highlighted with a simple consideration: scanning data on a modern physical hard disk can be done with a throughput of about 100 Megabytes/s.

Therefore, scanning 1 Terabyte takes 5 hours and 1 Petabyte takes 5000 hours (Raicu et al., 2011). The Big Data problem does not only rely in archiving and conserving huge quantity of data. The real challenge is to access such data in an efficient way, applying massive parallelism not only for the computation, but also for the storage.

Moreover, the transfer rate is inversely proportional to the distance to cover. HPC clusters[9] are typically equipped with high-level interconnections as InfiniBand[10], which have a latency of 2000 ns (only 20 times slower than RAM and 200 times faster than a Solid State Disk) and data rates ranging from 2.5 Gigabit per second (Gbps) with a single data rate link (SDR), up to 300 Gbps with a 12-link enhanced data rate connection (Cai et al. 2015). But this performance can only be achieved on a LAN[11], and the real problems arise when it is necessary to transfer data between geographically distributed sites, because the Internet connection might not suitable to transfer Big Data. Although several projects, such GÉANT[12], the pan-European research and education network, and its US counterpart Internet2[13], have been funded to improve the network interconnections among states; the achievable bandwidth is insufficient. For example, BGI[14], the world largest genomics service provider, uses FedEx for delivering results.

If a local infrastructure is exploited for the analysis of Big Data, one effective solution is represented by the use of client/server architectures where the data storage is spread among several devices and made accessible through a network (Bakshi, 2012). The available tools can be mainly subdivided in distributed file systems, cluster file systems, and parallel file systems. Distributed file systems consist of disk arrays physically attached to a few I/O servers through fast networks and then shared to the other nodes of a cluster. In contrast, cluster file systems provide direct disk access from multiple computers at the block level (access control must take place on the client node). Parallel file systems are like cluster file systems, where multiple processes are able to concurrently read and write a single file, but exploit a client-server approach, without direct access to the block level.

The Hadoop Distributed File System (HDFS) is a Java-based file system that

[9] HPC clusters: High-Performance Computing Cluster is an open source, data-intensive computing system platform developed by LexisNexis Risk Solutions. The HPCC platform incorporates a software architecture implemented on commodity computing clusters to provide high-performance, data-parallel processing for applications utilizing big data.

[10] InfiniBand: InfiniBand (abbreviated IB) is a computer network communications link used in high-performance computing featuring very high throughput and very low latency. It is used for data interconnect both among and within computers. As of 2014 it is the most commonly used interconnect in supercomputers. InfiniBand host bus adapters and network switches are manufactured by Mellanox and Intel.

[11] LAN: a local area network (LAN) is a computer network that interconnects computers within a limited area such as a home, school, computer laboratory, or office building, using network media.

[12] GÉANT is the pan-European data network for the research and education community. It interconnects national research and education networks across Europe, enabling collaboration on projects ranging from biological science to earth observation and arts and culture (http://www.geant.net).

[13] Internet2: Internet2 is an exceptional community of U.S. and international leaders in research, academia, industry and government who create and collaborate via innovative technologies (http://www.internet2.edu).

[14] BGI Americas: more information at http://bgiamericas.com/

provides scalable and reliable data storage that is designed to span large clusters of commodity servers (Shvachko et al., 2012). It is an open source version of the GoogleFS introduced in 2003 (Ghemawat et al., 2003). The design principles were derived from Google's needs, as the fact that most files only grow because new data have to be appended, rather than overwriting the whole file, and that high sustained bandwidth is more important than low latency. As regards I/O operations, the key aspects are the efficient support for large streaming or small random reads, besides large and sequential writes to append data to files. The other operations are supported as well, but they can be implemented in a less efficient way.

At the architectural level, HDFS requires two processes: a NameNode service, running on one node in the cluster and a DataNode process running on each node that will process data. HDFS is designed to be fault-tolerant due to replication and distribution of data, since every loaded file is replicated (3 times is the default value) and split into blocks that are distributed across the nodes. The NameNode is responsible for the storage and management of metadata, so that when an application requires a file, the NameNode informs about the location of the needed data. Whenever a data node is down, the NameNode can redirect the request to one of the replicas until the data node is back online. Since the cluster size can be very large (it was demonstrated with clusters up to 4,000 nodes), the single NameNode per cluster forms a possible bottleneck and single point of failure. This can be mitigated by the fact that metadata can be stored in the main memory and the recent HDFS High Availability feature provides the user with the option of running two redundant NameNodes in the same cluster, one of them in standby, but ready to intervene in case of failure of the other.

3.3 Middleware for Big Data

The file system is the first level of the architecture. The second level corresponds to the framework/middleware supporting the development of user-specific solutions, while applications represent the third level. Besides the general-purpose solutions for parallel computing like the Message Passing Interface, hybrid solutions based on it, or extension to specific frameworks like the R software environment for statistical computing (Dudoit et al., 2003), there are several tools specifically developed for Big Data analysis.

The first and more famous example is the above mentioned Apache Hadoop, an open source software framework for large-scale storing and processing of data sets on distributed commodity hardware. Hadoop is composed of two main components, HDFS and MapReduce (Webster et la., 2004). The latter is a simple framework for distributed processing based on the Map and Reduce functions, commonly used in functional programming. In the Map step the input is partitioned by a master process into smaller subproblems and then distributed to worker processes. In the Reduce step the master process collects the results and combines them in some way to provide the

answer to the problem it was originally trying to solve.

Hadoop was designed as a platform for an entire ecosystem of capabilities, used in particular by a large number of companies, vendors, and institutions. To provide an example in the bioinformatics field, in The Cancer Genome Atlas project[15] researchers implemented the process of "sharding," splitting genome data into smaller more manageable chunks for cluster-based processing, utilising the Hadoop framework and the Genome Analysis Toolkit (Van der Auwera et al., 2013). Many other works based on Hadoop are present in literature, for example, some specific libraries were developed as Hadoop-BAM (Niemenmaa et al., 2012), a library for distributed processing of genetic data from next generation sequencer machines, was also developed relying on this framework.

Hadoop was the basis for other higher-level solutions as Apache Hive16, a data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. At the same time this language also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL. Apache Hive supports analysis of large datasets stored in HDFS and compatible file systems such as the Amazon S3 file system. It provides an SQL-like language called HiveQL while maintaining full support for map/reduce. To accelerate queries, it provides indexes, including bitmap indexes, and it is worth exploiting in several bioinformatics applications. Apache Pig[17] is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets. Apache Pig has the similar aim of Apache Hive to allow domain experts, who are not necessarily IT specialists, to write complex MapReduce transformations using a simpler scripting language. It was used for example for sequence analysis.

Hadoop is considered almost as synonymous for Big Data. However, there are some alternatives based on the same MapReduce paradigm like Disco[18], a lightweight, open-source framework for distributed computing based on the MapReduce paradigm. Disco is powerful and easy to use, thanks to Python. Disco distributes and replicates your data, and schedules your jobs efficiently. Disco even includes the tools you need to index billions of data points and query them in real-time. Disco was born in Nokia Research Center in 2008 to solve real challenges in handling massive amounts of data. Disco has been actively developed since then by Nokia and many other companies who use it for a variety of purposes, such as log

[15] The Cancer Genome Atlas project: The Cancer Genome Atlas will identify the genomic changes in more than 20 different types of human cancer. By comparing the DNA in samples of normal tissue and cancer tissue taken from the same patient, researchers can identify changes specific to that particular cancer.
[16] Apache Hive: https://hive.apache.org/
[17] Apache Pig: http://pig.apache.org/
[18] Disco: http://discoproject.org/

analysis, probabilistic modelling, data mining, and full-text indexing.

Nowadays, the attention of Big Data community is focused on Apache Spark[19], an open-source cluster computing framework originally developed in the AMPLab at UC Berkeley. In contrast to Hadoop's two-stage disk-based MapReduce paradigm, Spark's in-memory primitives provide performance up to 100 times faster for certain applications (Xin et al. 2013). By allowing user programs to load data into a cluster's memory and query it repeatedly, Spark is well suited to machine learning algorithms (Zaharia, 2011.). Spark requires a cluster manager and a distributed storage system. For cluster manager, Spark supports standalone (native Spark cluster), Hadoop YARN, or Apache Mesos[20]. For distributed storage, Spark can interface with a wide variety, including Hadoop Distributed File System (HDFS), Cassandra[21], OpenStack Swift, and Amazon S3. Spark also supports a pseudo-distributed mode, usually used only for development or testing purposes, where distributed storage is not required and the local file system can be used instead; in the scenario, Spark is running on a single machine with one worker per CPU core.

Spark has over 465 contributors in 2014, making it the most active project in the Apache Software Foundation and among Big Data open source projects.

3.4 Semantics, ontologies and open format for data integration

The previous sections focused mainly on how to analyse Big Data for inferring correlations, but the extraction of actual new knowledge requires something more. The key challenges in making use of Big Data lie, in fact, in finding ways of dealing with heterogeneity, diversity, and complexity of the information, while its volume and velocity hamper solutions are available for smaller datasets such as manual curation or data warehousing.

Semantic web technologies are meant to deal with these issues. The development of metadata for biological information on the basis of semantic web standards can be seen as a promising approach for a semantic-based integration of biological information (Dräger and Palsson, 2014). On the other hand, ontologies, as formal models for representing information with explicitly defined concepts and relationships between them, can be exploited to address the issue of heterogeneity in data sources.

[19] Apache Spark: https://spark.apache.org/

[20] Apache Mesos: Mesos is built using the same principles as the Linux kernel, only at a different level of abstraction. The Mesos kernel runs on every machine and provides applications (e.g., Hadoop, Spark, Kafka, Elastic Search) with API's for resource management and scheduling across entire datacenter and cloud environments.

[21] Cassandra: The Apache Cassandra database is the right choice when you need scalability and high availability without compromising performance. Linear scalability and proven fault-tolerance on commodity hardware or cloud infrastructure make it the perfect platform for mission-critical data.

In domains like bioinformatics and biomedicine, the rapid development and adoption of ontologies prompted the research community to leverage them for the integration of data and information (Hoehndorf et al., 2014). Finally, since the advent of linked data a few years ago, it has become an important technology for semantic and ontologies research and development. We can easily understand linked data as being a part of the greater Big Data landscape, as many of the challenges are the same. The linking component of linked data, however, puts an additional focus on the integration and conflation of data across multiple sources (Mayer et al., 2014).

3.4.1 Semantics

The semantic web is a collaborative movement (W3C)[22], which promoted standard for the annotation and integration of data. By encouraging the inclusion of semantic content in data accessible through the Internet, the aim is to convert the current web, dominated by unstructured and semistructured documents, into a web of data (Bonino et al. 2004). It involves publishing information in languages specifically designed for data: Resource Description Framework (RDF), Web Ontology Language (OWL), and Protocol and RDF Query Lenaguage (SPARQL) which is a query language for semantic web data sources.

RDF represents data using subject-predicate-object triples, also known as "statements." This triple representation connects data in a flexible piece-by-piece and link-by-link fashion that forms a directed labelled graph (Ruttenberg et al., 2007). The components of each RDF statement can be identified using Uniform Resource Identifiers (URIs). Alternatively, they can be referenced via links to RDF Schemas (RDFS), Web Ontology Language (OWL), or to other (nonschema) RDF documents. In particular, OWL is a family of knowledge representation languages for authoring ontologies or knowledge bases. The languages are characterised by formal semantics and RDF/XML-based serializations for the semantic web. In the field of biomedicine, a notable example is the Open Biomedical Ontologies (OBO) project[23], which is an effort to create controlled vocabularies for shared use across different biological and medical domains (Aranguren et al., 2007). OBO belongs to the resources of the US National Center for Biomedical Ontology (NCBO), where it will form a central element of the NCBO's BioPortal. The interrogation of these resources can be

[22] W3C: The Semantic Web Coordination Group is tasked to provide a forum for managing the interrelationships and interdependencies among groups focusing on standards and technologies that relate to this goals of the Semantic Web Activity. This group is designed to coordinate, facilitate and (where possible) help shape the efforts of other related groups to avoid duplication of effort and fragmentation of the Semantic Web by way of incompatible standards and technologies.

[23] Open Biomedical Ontologies : the OBO Foundry is a collaborative experiment involving developers of science-based ontologies who are establishing a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain. The groups developing ontologies who have expressed an interest in this goal are listed below, followed by other relevant efforts in this domain (http://www.obofoundry.org/).

performed using SPARQL, which is an RDF query language similar to SQL, for the interrogations of databases, able to retrieve and manipulate data stored in RDF format. For example, BioGateway organizes the SwissProt database[24], along with Gene Ontology Annotations (GOA), into an integrated RDF database that can be accessed through a SPARQL query endpoint, allowing searches for proteins based on a combination of Gene Ontology and SwissProt data (Huntley et al., 2015).

The support for RDF in high-throughput bioinformatics applications is still small, although researchers can already download the UniProtKB and its taxonomy information using this format or they can get ontologies in OWL format, such as Gene Ontology. The biggest impact RDF and OWL can have in bioinformatics, though, is to help integrate all data formats and standardise existing ontologies (González-Beltrán et al., 2014). If unique identifiers are converted to URI references, ontologies can be expressed in OWL, and data can be annotated via these RDF-based resources. The integration between them is a matter of merging and aligning the ontologies (in case of OWL using the "rdf:sameAs" statement). After the data has been integrated we can use the plus that comes with RDF for reasoning: context embeddedness. Organizations in the life sciences are currently using RDF for drug target assessment (Williams et al., 2012), and for the aggregation of omics data (Smith et al., 2007).

In addition, semantic web technologies are being used to develop well-defined and rich biomedical ontologies for assisting data annotation and search, the integration of rules to specify and implement bioinformatics workflows, and the automation for discovering and composing bioinformatics web services.

3.4.2 Ontologies

An ontology layer is often an invaluable solution to support data integration, particularly because it enables the mapping of relations among data stored in a database (Osier et al., 2004). Belonging to the field of knowledge representation, an ontology is a collection of terms within a particular domain organised in a hierarchical structure that allows searching at various levels of specificity. Ontologies provide a formal representation of a set of concepts through the description of individuals, which are the basic objects, classes, that are the categories which objects belong to, attributes, which are the features the objects can present, and relations, that are the ways objects can be related to one another. Due to this "tree" (or, in some cases, "graph") representation, ontologies allow the link of terms from the same domain, even if they belong to different sources in data integration contexts, and the efficient matching of apparently diverse and distant entities. The latter aspect can not only

[24] SwissProt: UniProtKB/Swiss-Prot is the manually annotated and reviewed section of the UniProt Knowledgebase (UniProtKB). It is a high quality annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions. Since 2002, it is maintained by the UniProt consortium and is accessible via the UniProt website.

improve data integration, but even simplify the information searching.

In the biomedical context a common problem concerns, for example, the generality of the term cancer. A direct query on that term will retrieve just the specific word in all the occurrences found into the screened resource. Employing a specialised ontology (i.e., the human disease ontology—DOID) the output will be richer, including terms such as sarcoma and carcinoma that will not be retrieved otherwise. Ontology-based data integration involves the use of ontologies to effectively combine data or information from multiple heterogeneous sources. The effectiveness of ontology-based data integration is closely tied to the consistency and expressivity of the ontology used in the integration process. Many resources exist that have ontology support: SNPranker, G2SBC, NSD, TMARepDB, Surface, and Cell cycle DB.

A useful instrument for ontology exploration has been developed by the European Bioinformatics Institute[25] (EBI), which allows easily visualising and browsing ontologies in the OBO format: the open source Ontology Lookup Service (Mayer et al., 2014). The system provides a user-friendly web-based single point to look into the ontologies for a single specific term that can be queried using a useful autocompletion search engine. Otherwise it is possible to browse the complete ontology tree using an Ajax library, querying the system through a standard SOAP web service described by a WSDL descriptor. The following ontologies are commonly used for annotation and integration of data in the biomedical and bioinformatics (Schulz et al., 2006):

- Gene ontology is the most exploited multilevel ontology in the biomolecular domain. It collects genome and proteome related information in a graph-based hierarchical structure suitable for annotating and characterising genes and proteins with respect to the molecular function and biological process they are involved in, and the spatial localisation they present within a cell (Aranguren et al., 2007).
- KEGG ontology (KOnt), which provides a pathway based annotation of the genes in all organisms. No OBO version of this ontology was found, since it has been generated directly starting from data available in the related resource.
- Brenda Tissue Ontology (BTO), to support the description of human tissues.
- Cell Ontology (CL), to provide an exhaustive organisation about cell types.
- Disease ontology (DOID), which focus on the classification of breast cancer pathology compared to the other human diseases.
- Protein Ontology (PRO), which describes the protein evolutionary classes to delineate the multiple protein forms of a gene locus.
- Medical Subject Headings thesaurus (MESH), which is a hierarchical controlled vocabulary able to index biomedical and health-related information.
- Protein structure classification (CATH), which is a structured vocabulary used for the classification of protein structures.

[25] EBI: EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry. (http://www.ebi.ac.uk/)

3.4.3 Linked Data

Linked data describes a method for publishing structured data so that they can be interlinked, making clearer the possible interdependencies. This technology is built upon the semantic web technologies previously described (in particular it uses HTTP, RDF, and URIs), but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by IT systems.

The linked data paradigm is one approach to cope with Big Data as it advances the hypertext principle from a web of documents to a web of rich data. The idea is that after making data available on the web (in an open format and with an open license) and structuring them in a machine-readable fashion (e.g., Excel instead of image scan of a table), researchers must work to annotate this information with open standards from W3C (RDF and SPARQL), so that people can link their own data to other people's data to provide context information.

In the field of bioinformatics, a first attempt to publish linked data has been performed by the Bio2RDF project (Tilahun et al., 2014). The project's goal is to create a network of coherently linked data across the biological databases. As part of the Bio2RDF project, an integrated bioinformatics warehouse on the semantic web has been built. Bio2RDF has created a RDF warehouse that serves over 70 million triples describing the human and mouse genomes.

A very important step towards the use of linked data in the computational biology field has been done by the above mentioned EBI, which developed an infrastructure to access its information by exploiting this paradigm. In detail, the EBI RDF[26] platform allows explicit links to be made between datasets using shared semantics from standard ontologies and vocabularies, facilitating a greater degree of data integration. SPARQL provides a standard query language for querying RDF data. Data that have been annotated using ontologies, such as DOID and the Gene Ontology, can be integrated with other community datasets, providing the semantics support to perform rich queries. Publishing such datasets as RDF, along with their ontologies, provides both the syntactic and semantic integration of data, long promised by semantic web technologies. As the trend toward publishing life science data in RDF increases, we anticipate a rise in the number of applications consuming such data. This is evident in efforts such as the Open PHACTS platform[27] and the

[26] EBI RDF paltform: The EBI RDF Platform aims to bring together the efforts of a number of EMBL-EBI resources that provide access to their data using Semantic Web technologies. It provides a unified way to query across resources using the W3C SPARQL query language (https://www.ebi.ac.uk/rdf/platform).

[27] Open PHACTS platform: the Open PHACTS Explorer provides a simple web-based user interface for querying and viewing the integrated data within the Open PHACTS Discovery Platform (http://www.openphacts.org/open-phacts-discovery-platform).

AtlasRDF-R package[28]. The final aim is that EBI RDF platform can enable such applications to be built by releasing production quality services with semantically described RDF, enabling real biomedical use cases to be addressed.

3.5 Perspectives and Open Problems

Data is considered the fourth paradigm in science, besides experimental, theoretical, and computational sciences (Darema, 2004). This is becoming particularly true in computational biology, where, for example, the approach "sequence first, think later" is rapidly overcoming the hypothesis-driven approach. In this context, Big Data integration is really critical for bioinformatics that is "the glue that holds biomedical research together."

There are many open issues for Big Data management and analysis, in particular in the computational biology and healthcare fields. It is critical to face the variety of the information that should be managed by such infrastructures, which should be organized in scheme-less contexts, combining both relaxed consistency and a huge capacity to digest data. Therefore, a critical point is that relational databases are not suitable for Big Data problems. They lack horizontal scalability, need hard consistency, and become very complex when there is the need to represent structured relationships. Nonrelational databases (NoSQL) are the interesting alternative to data storage because they combine the scalability and flexibility.

From the computational point of view, the novel idea is that jobs are directly responsible of managing input data, through suitable procedures for partitioning, organizing, and merging intermediate results. Novel algorithm will contain large parts of not functional code, but essential for exploiting housekeeping tasks. Due to the practical impossibility of moving all the data across geographical dispersed sites, there is the need of computational infrastructure able to combine large storage facilities and HPC. Virtualization can be the key in this challenge, since it can be exploited to achieve storage facilities able to leverage in-memory key/value databases to accelerate data-intensive tasks.

[28] AtlasRDF-R package: query the Gene Expression Atlas RDF data at the European Bioinformatics Institute using genes, experimental factors (such as disease, cell type, compound treatments), pathways and proteins.

Chapter 4

Biomedical research reproducibility

Two-dimensional gel electrophoresis (2D-GE) is a powerful technology to compare complex protein mixtures. 2D-GE has been applied to many fields of biomedical research and is widely used in biomarker discovery.

In recent years proteomics authors collected and analysed many 2D-GE based articles featuring lists of the differentially expressed proteins, denoting a disturbing déjà vu. The same proteins seem to predominate regardless of the experiment, tissue or species. The occurrence of individual differentially expressed proteins in 2D-GE published papers has been quantified, and the appearance of protein biomarkers identified in human, mouse, and rat tissues were calculated, showing groups of frequently identified proteins. These reports suggest that the commonly observed changes represent common cellular stress responses or are a reflection of the technical limitations of 2D-GE, and the frequent identification of these proteins must be considered in the interpretation of any 2D-GE studies.

Nowadays, serious concerns regarding the trustworthiness of research findings have been voiced in major journals and at professional meetings of fields as diverse as medicine, physics, cell biology, economics, and social psychology. There are many possible reasons, in a few cases researchers committed fraud and literally made up their data, but such fraud is actually rare, and the typical reasons for failures to replicate are different. To list just a few: the replication study might not follow exactly the same methods as the original study, or the new investigators may not have the skills necessary to successfully repeat a complex experimental procedure; the finding in question might have undiscovered "moderator variables," factors that cause the finding to get stronger, or go away. The mechanisms of nature are complicated, sometimes even almost chaotic. Scientists work hard to find signal amidst all that noise, and when they think they find something, are eager to report it to their colleagues and to the world.

The recommendations encourage different experts are numerous, and some are rather technical, but the recommendation that might be the most important is also the simplest: do more research. This means simply to improve the ability of manage more data.
This capter discusses in Section 1 the problem of reproducibility in research, in Section 2 try to draw some recommendations. Section 3 discusses how the meta-analyses of proteomic data can provide invaluable information pertinent to various biological processes provide invaluable information for biomedical research.

4.1 Research reproducibility a problem of modern science

A simple idea underpins science: "trust, but verify". Results should always be subject to challenge from experiment. That simple but powerful idea has generated a vast body of knowledge (Alvarenga, 2013). Since its birth in the 17th century, modern science has changed the world beyond recognition, and overwhelmingly for the better. But success can breed complacency. Modern scientists are doing too much trusting and not enough verifying, to the detriment of the whole of science, and of humanity. Too many of the findings that fill the academic ether are the result of shoddy experiments or poor analysis (Begley and Ioannidis, 2015).

Over the past few years various researchers have made systematic attempts to replicate some of the more widely cited priming experiments. Many of these replications have failed. For instance, a paper in PLoS ONE reported that nine separate experiments had not managed to reproduce the results of a famous study from 1998 purporting to show that thinking about a professor before taking an intelligence test leads to a higher score than imagining a football hooligan (Moran et la., 2014).

The idea that the same experiments always get the same results, no matter who performs them, is one of the cornerstones of science's claim to objective truth (Gómez-Pérez and Mazurek, 2014). If a systematic campaign of replication does not lead to the same results, then either the original research is flawed (as the replicators claim) or the replications are (as many of the original researchers on priming contend). Either way, something is awry.

Academic scientists readily acknowledge that they often get things wrong. But they also hold fast to the idea that these errors get corrected over time as other scientists try to take the work further. Evidence that many more dodgy results are published than are subsequently corrected or withdrawn calls that much-vaunted capacity for self-correction into question. There are errors in a lot more of the scientific papers being published, written about and acted on than anyone would normally suppose, or like to think.

4.2 The statistics mistakes in biomedical research

Various factors contribute to the problem. Statistical mistakes are widespread. The peer reviewers who evaluate papers before journals commit to publishing them are much worse at spotting mistakes than they or others appreciate.

First, the statistics, which if perhaps off-putting are quite crucial. Scientists divide errors into two classes. A type I error is the mistake of thinking something is true when it is not (also known as a "false positive"). A type II error is thinking something is not true when in fact it is (a "false negative"). When testing a specific hypothesis, scientists run statistical checks to work out how likely it would be for data which seem to support the idea to have come about simply by chance. If the likelihood of such a false-positive conclusion is less than 5%, they deem the evidence that the hypothesis is true "statistically significant". They are thus accepting that one result in 20 will be falsely positive but one in 20 seems a satisfactorily low rate.

Many researchers agree with the idea that the customary approach to statistical significance ignores three things:

- the "statistical power" of the study (a measure of its ability to avoid type II errors, false negatives in which a real signal is missed in the noise);
- the unlikeliness of the hypothesis being tested;
- the pervasive bias favouring the publication of claims to have found something new.

A statistically powerful study is one able to pick things up even when their effects on the data are small. In general bigger studies, those which run the experiment more times, recruit more patients for the trial, or whatever are more powerful. A power of 0.8 means that of ten true hypotheses tested, only two will be ruled out because their effects are not picked up in the data; this is widely accepted as powerful enough for most purposes. But this benchmark is not always met, not least because big studies are more expensive. A study by Dr Ioannidis and colleagues found that in neuroscience the typical statistical power is a dismal 0.21; writing in Perspectives on Psychological Science, Marjan Bakker of the University of Amsterdam and colleagues reckon that in that field the average power is 0.35.

Moreover, it remains to consider that the negative results are much more trustworthy than positive ones. But researchers and the journals in which they publish are not very interested in negative results. They prefer to accentuate the positive, and thus the error-prone. Negative results account for just 10-30% of published scientific literature, depending on the discipline. This bias may be growing. A study of 4,600 papers from across the sciences conducted by Daniele Fanelli of the University of Edinburgh found that the proportion of negative results dropped from 30% to 14% between 1990 and 2007.

Some scientists use inappropriate techniques because those are the ones they feel comfortable with; others latch on to new ones without understanding their

subtleties. Some just rely on the methods built into their software, even if they don't understand them.

Software can also be a problem for would-be replicators. Some code used to analyse data or run models may be the result of years of work and thus precious intellectual property that gives its possessors an edge in future research. Although most scientists agree in principle that data should be openly available, there is genuine disagreement on software. Journals which insist on data-sharing tend not to do the same for programs.

4.3 Frequently identified proteins in proteomics literature

The problem of the reproducibility of results affects also the omics sciences. Although conventional 2D-GE remains a fundamental tool in expression proteomics, this technique presents some limitations. Among its most criticized weaknesses are low dynamic range and relatively low resolution. Usually only a few hundred proteins can be detected on one gel representing the most abundant soluble cytosolic proteins. A typical published 2D-GE-based expression proteomics experiment features 400-1500 spots and reports between 10 and 40 identified up- or down- regulated proteins. After reading many 2D-GE-based articles presenting lists of the differentially expressed proteins, one starts experiencing a disturbing sense of déjà vu. The same proteins seem to predominate regardless of the experiment, tissue, and species. To explore this observation and to quantify the occurrence of individual differentially expressed proteins in 2D-GE experiment reports, some research team performed a proteomic meta-analysis.

These studies showed that a typical published 2-DE experiment demonstrates differential expression of several cytoskeletal and stress proteins, proteasome subunits, glycolytic enzymes, elongation factors, and heterogeneous ribonucleoprote in particle subunits or glutathione transferases. These proteins are, in some cases, accompanied by less abundant regulatory proteins. This result suggests that we should use extreme caution in the interpretation of differential expression of the most frequently identified proteins. Table 1 shows each identified protein as an individual polypeptide. However, many proteins belong to protein families of structurally very closely related molecules, often of similar, or overlapping functions such as annexins, tubulins, or peroxiredoxins. To take this fact into account Table 2 shows the frequently identified protein families as individual units.

These studies demonstrated that meta-analyses of proteomic data can provide invaluable information pertinent to various biological processes or methods involved. Targeted, statistically robust proteomic meta-analyses could provide invaluable information for biomedical research. For instance a global analysis of disease specific expression patterns based on a large dataset that also consider whether the protein was up- or down-regulated, could provide a brand new tool for dissecting the complex

processes of pathological bioprosses. Several authors agree that such cross heuristic views of proteomic data have the potential to discover limitations or weaknesses of 2D-GE as a method and consequently help to improve and further develop this technique.

Table 1. Top 15 most often identified differentially expressed proteins

Individual proteins							
	Humans				**Rodents**		
	Protein name	N. of identi-fications	Identified in percents of experiments		Protein name	N. of identi-fications	Identified in percents of experiments
1	HSP27	34	31 (%)	**1**	Enolase 1	25	32 (%)
2	Enolase 1	31	29	**2**	HSP60	16	21
3	Triosephosphate isomerase	22	20	**3**	ATP synthase beta subunit	14	18
4-6	Pyruvate kinase M1/M2	21	19	**4-8**	Vimentin	13	17
4-6	Peroxiredoxin1	21	19	**4-8**	Grp75	13	17
4-6	Peroxiredoxin2	21	19	**4-8**	Apolipoprotein A1	13	17
7	Vimentin	20	19	**4-8**	Dihydropyrimidinase like 2 protein	13	17
8	Annexin A4	19	18		Peroxiredoxin 6	13	17
9	HSC71	18	17	**9-10**	Phosphoglycerate mutase1	12	15
10-11	Peptidyl-prolyl isomerase A	17	16		HSC71	12	15
	Cytokeratin 8	17	16	**11-12**	Triosephosphate isomerase	10	13
12	Cathepsin D	16	15	**11-12**	Calreticulin	10	13
13	ATP synthase beta subunit	15	14	**13-15**	RhoGDI 1	9	12
14-15	Grp78/Bip	14	13	**13-15**	Grp78/Bip	9	12
14-15	RhoGDI 1	14	13	**13-15**	GAPDH	9	12

Table 2. Frequently identified protein families as individual units

Individual proteins							
	Humans				**Rodents**		
	Protein name	N. of identi-fications	At least one member identi-fied in percent of experiments (%)		Protein name	N. of identi-fications	At least one member identi-fied in percent of experiments (%)
1	Keratins	70	41	**1**	Peroxiredoxins	34	38
2	Annexins	67	40	**2**	Enolases	33	42
3	Peroxiredoxins	61	46	**3–4**	Tubulins	24	20
4	Actins	36	30	**3–4**	PDIs	24	26
5–6	HSP27	34	31	5	Annexins	22	26
5–6	Tropomyosins	34	23	6–7	Actins	21	23
7	GSTs	33	29		GSTs	21	19
8–10	Enolases	32	30	8–9	Tropomyosins	17	16
	PDIs	32	26		Dihydropyrimidinase-like proteins	17	19
	Tubulins	32	21		HSP60	16	21
	Cathepsins	26	22		Carbonic anhydrases	17	18
	TCP-1	12	24		Apolipoproteins	15	17
	Triosephosphate isomerase	22	20		ATP synthase beta subunit	14	18
	Pyruvate kinases	22	20		Malate dehydrogenases	14	18
	Vimentin	20	20		14-3-3 proteins	14	14

Chapter 5

Two dimensional gel electrophoresis image database

Authors of biomedical articles do often use images to present experimental results, or to communicate important concepts by means of appropriate charts. In the biomedical research community, much attention is drawn by figures, since figures often summarize the findings of the research work under consideration. Mining this source of information would allow a better evaluation of experimental results and allows a greater comprehension of the biological insight that can be collected from this data.
Different bioinformatics research groups develop biomedical image database systems, in which the researchers can store the images extracted form paper of interest, image found on the web repositories and their own experimental images.
This chapter present in Section 1 the state of art of the database for biomedical images. Section 2 present a new platform that automatically extract images and captions from biomedical literature. Section 3 presents a database comparison. Finally Section 4 describe how the user can manually tag images by using curated biomedical ontologies.

5.1 State of art of 2D-GE image repositories

Authors of biomedical articles do often use images to present experimental results, or to communicate important concepts by means of appropriate charts (Kim and Yu, 2011). In the biomedical research community, much attention is drawn by

figures, since figures often summarize the findings of the research work under consideration. There is growing evidence of the need for automated systems that would help biologists in finding the information contained in images and captions quickly and satisfactorily.

Different figure mining systems have indeed been proposed, including the Subcellular Location Image Finder (SLIF) system (Ahmed et al., 2010), BioText (Hearst et al., 2007) and Yale Image Finder (YIF)(Xu et al., 2008). All these systems store the processed information in a web- accessible, searchable database. Interestingly, these instruments allow researchers to structurally browse, in a precise and rapid way, otherwise unstructured knowledge. In particular, the development of proteomic image databases (2D-GE gel DB), with tools for the comparison of proteomic experimental maps, is widely encouraged (Drews and Görg, 2005). The possibility of using text mining and image processing technologies to create an interactive 2D-GE gel repository could therefore be of interest to the field. Furthermore, the 2D-GE gel DB are often used to support experimental identification of spots, as obtained by mass spectrometry analysis.

Since it was launched in 1993, the ExPASy website (http://www.expasy.org/) has been a reference in the proteomics community. Through the World-2D PAGE Portal, users can access more than 60 federated databases containing nearly 420 2D-GE gel images. The Proteome Experimental Data Repository (PEDRo), in addition, provides access to a collection of descriptions of proteomic protocols (Garwood et al., 2004). The Integrated Proteomics Exploring Database (IPED), based on the schema of PEDRo, was introduced in 2009 with the aim of standardize, store and visualize data obtained, not only from proteomics experiments but also from mass spectrometry (MS) assays (Zheng et al, 2009). However, current 2D-GE gel databases do not provide tools for high-throughput data uploading, and they lack convenient tools for data access.

5.2 A tool for image extraction and annotation

This section present iMole[1] (Giordano et al., 2003), an innovative system for automated image extraction, annotation and storage into a database, which is able to parse biological journal articles. Using BFO[2] Java library, iMole is capable of parsing biomedical articles and to automatically extract images and captions from Portable Document Format (PDF) documents. iMole allows the user to upload experimental gel images adding them to literature-retrieved images.

Furthermore, the user of iMole can add and modify various information to each 2D-GE image as detailed spots annotations, isoelectric point (pI) and protein

[1] iMole is available with a preloaded set of 2DE gel data at http://imole.biodigitalvalley.com.
[2] BFO: the BFO Java PDF Library is a smart, fast, flexible way to create, edit and display PDF documents (http://bfo.com),

molecular weight (MW) landmarks, free text description and controlled vocabulary terms. Finally, the user can create a personal set of tagged 2DE images.

Interestingly, this instrument allows researchers to browse the vast amount of unstructured data in a precise and rapid way. In order to retrieve images, users can query the database using terms belonging to different ontologies.

iMole has been built according to the principles of Expasy federated database, and it provides:

- gel images accessible directly on the iMole website;
- annotated gels with clickable protein spots, that if selected, display basic protein information;
- spot linked to other databases (i.e. Uniprot, Gene);
- keyword search query.

iMole stored a 2DE gel electrophoresis database. The database was created by using the software ProteinQuest[3] (PQ) to search all free full text articles with available captions stored in Medline. For all these target scientific articles we retrieved the PDF files and we extracted the images from these files by BFO library (Gatti et al., 2013).

In order to identify the captions referring to 2D-GE gel experiments, PQ performed two different classifiers: a supervised in house developed indexing method (iMole text mining tool) and a supervised learning neural network (NN) developed using Weka data mining software (Hall et al., 2009). Two independent sets of 2000 and 981 captions were been selected for training and testing steps, respectively. To train the multilayer perceptron of Weka we used the training set of 2000 captions (100 concerning 2D-GE gel and 1900 not concerning 2D-GE gel). The multilayer perceptron of Weka was implemented with three hidden levels and ten nodes in each level.

iMole text mining tool parsed all captions searching for terms related to the "2-Dimensional Gel Electrophoresis" technique and its alias (e.g. 2D-GE, DIGE, bi-dimensional electrophoresis, etc.); these terms belong to the methods dictionary. Ambiguities in the terminology are resolved using multiple searches for more than one alias, as well as the co- occurrence of specific words which can either deny or force the tagging.

To check the efficiency of the two classifiers, the control set of 981 captions was analysed (391 concerning 2D-GE gels and 590 not concerning 2D-GE gel). iMole text mining tool tool wrongly identifies as positive 97 captions (false positives, fp), and it correctly classifies, as negative, all captions that do not concern 2D-GE gels. The NN wrongly identifies 104 captions as positives (fp), and 91 as negative (false negative, fn). The comparison between the two systems reveals that iMole text mining tool gives more reliable results than the NN system.

This data is represented in Table 1.

[3] ProteinQuest: ProteinQuest is a complete web suite for the retrieval and automated analysis of biomedical information (www.proteinquest.com).

Using iMole the fraction of retrieved captions that are correct are 80.12 % (Precision), instead of 74.26 % obtained with the NN classification. Furthermore the fraction of correct captions retrieved by iMole tool, are 100.00 % (Recall) instead of 76.73 % performed by the NN system.

This data is represented in Table 2.

Other existing classifiers which can be used for retrieval of 2D-GE gel images include BioText and YIF[4], which are based on full-featured text search engine, such as Apache Lucene. With respect to the implementation of these software, iMole text mining tool indexed paper images with the help of semantic dictionaries describing biological entities, thus resolving ambiguities in a better way.

Table 1. Results of the text mining engine: the Weka supervised learning neural network (NN), and iMole text matching method.

	NN text mining tool	**iMole text mining tool**
true positive	300	391
false positive	104	97
true negative	486	493
false negative	91	0

Table 2. Results of the analysis described in the text:

- precision (true positive / (true positive + false positive))
- recall (true positive / (true positive + false negative))
- F1 score (2 x (precision x recall/precision + recall)

	precision	**recall**	**F1 score**
NN	74.26%	76.73%	0.754717
iMole	80.12%	100.00%	0.889647

5.3 Biomedical image database

In order to compare our classifier to existing tools, iMole was tested side by side with YIF. YIF was queried using "2-Dimensional Gel Electrophoresis" and all the aliases obtained from our dictionary, finding 20 196 images vs 16 752 obtained by the same query in iMole.

To allow for a direct, manual comparison of the results, we restricted the search, filtering for the protein ApoE and analyzing all the images returned by iMole and YIF.

[4] YIF: http://krauthammerlab.med.yale.edu/imagefinder/

iMole retrieves a higher number of correct images than the YIF platform. The higher accuracy of the process (no false positive or false negative images retrieved by iMole) is demonstrated by the precision index calculated on the iMole results. In this particular case, iMole reached a precision index of 100.00%, whereas the precision index calculated on YIF results, did not reach 10%. By repeating the experiment with other proteins were been retrieved similar results, demonstrating a better retrieval accuracy for iMole.

Images positively identified, by iMole, are tagged as 2D-GE gels and stored into the database. In addition, the software associates other tags to the extracted images using several other ontologies, referring to diseases, proteins, tissue, cell type and organisms. Extra tags can be manually added using an input box with an auto-complete function, so to be compliant with a precompiled dictionary derived from Entrez MeSH[5] with some manually editing. Images of 2DE gels maintain a link to the original publication through a PubMed identifier (ID) tag, that allows researchers to retrieve the original article containing them.

In order to fine tune the evaluation of the specificity of the images collected in the 2D-GE gels database was been verified how many captions, that contain the term "2D-GE gel" (or its alias), are really associated to 2D-GE gel images. This analysis revealed that 22 366 images that iMole identify as 2D-GE gels and we found that 17 238 are 2DE gels images and 5128 images are other than 2D-GE gels. The precision calculated on these data is 77.07 %. Hence, out of a hundred captions that contain the term "2DE gel", 77 are actually associated with a 2D-GE gel images. Taking into consideration these results as a whole, the iMole image selection system returns much more relevant images than irrelevant ones.

Furthermore, iMole can also support the manual images uploading performed by users. These images can be manually tagged and named by the user, and are associated to a free text description instead of a caption.

Table 3. Results of the query of 2DE gel AND "apoe" using Yale Image Finder and iMole.

	YIF ("apoe" filter)	**iMole ("apoe" filter)**
true positive	60	84
false positive	606	0
true negative	0	622
false negative	24	0

[5] MeSH: Medical Subject Headings (MeSH) is the National Library of Medicine controlled vocabulary thesaurus used for indexing articles for PubMed.

Table 4. Results of the comparison between Yale Image Finder and iMole:

- precision (true positive / (true positive + false positive))
- recall (true positive / (true positive + false negative))
- F1 score (2 x (precision x recall/precision + recall)

	precision	**recall**	**F1 score**
NN	9.01%	71.43%	0.16
iMole	100.00%	100.00%	1

5.4 Image annotation and spot tagging

Supplementary information, such as isoelectric point (pI) and molecular weight (MW) data, can be added to all images stored in the database. Furthermore, the users can add protein names to spots not yet annotated. Possible ambiguities are avoided at this stage by means of the auto-complete text function, which will force the user to select among official Entrez protein symbols. All protein annotations are automatically linked to Entrez, so to have a rapid access to protein information.

Additionally, users can associate further information with the gel, like 2D-GE landmarks or any terms of provided dictionaries. All image information that results from paper processing and from user uploading, is available through an interactive web-interface. Users can query the corresponding database by searching for the image names or for terms contained into the preloaded dictionaries to retrieve a specific image.

iMole database was developed in a MySQL environment on a Red-Hat server. The iMole web interface is implemented with servlets and Java Server Pages (JSP) technologies, utilizing Apache Tomcat as the JSP engine.

Chapter 6

2D-GE Image mata-analysis

Most proteomics studies move to identify specific two-dimensional electrophoresis (2D-GE) pattern of proteins specifically related to a physiological or pathological condition. However, the information arising from these investigations is often incomplete due to inherent limitations of the technique, to extensive protein post-translational modifications and sometimes to the paucity of available samples.

Several proteomics 2D-GE studies have reported partially redundant lists of differently expressed proteins. To be able to further extract valuable information from existing 2D-GE data, the power of a meta-analysis will be evaluated in this chapter.

A bioinformatics and statistical approach allows studying protein patterns rather than one protein at a time. A meta-analysis combines the data from several studies and can for example be used to study more general protein patterns over several different diseases. Merging data from several pathological conditions enables the investigation of several clinical issues not possible to answer based on a single data set. Specific biomarkers thus have the potential to help provide a more certain and rapid disease diagnosis.

This chapter shows how a meta-analysis of proteomic data can provide valuable information pertinent to various biological processes that otherwise remains hidden.

Section 1 discusses the pool of different DJ-1 modified forms. Section 2 presents a robust procedure able to re-evaluate 2D-GE experiments available in the literature. Section 3 and 4 presents how to perform a 2D-GE meta-analysis and gain correlations between specific spot patterns respectively. Section 5 describes the statistical analysis performed on meta-analysis data.

6.1 Protein post-translational modifications

DJ-1/PARK7 is an ubiquitous, highly conserved protein that was originally identified because of its ability to transform NIH3T3 mouse cells in cooperation with Ras (Nagakubo et al., 1997). Starting from the association between loss-of-function mutations in the DJ-1 gene and PARK7, a monogenic, autosomal-recessive form of Parkinson's disease (PD) (Bonifati et al., 2012), an accumulating body of evidence pinpointed the important role of DJ-1 in this neurodegenerative condition. Very recently, it has been shown that oxidized dopamine can covalently modify DJ-1 (Van Laar et al., 2009); however, whether this can affect dopamine cell degeneration is unknown. Some hints may come from the involvement of DJ-1 into many cellular functions, including evidence linking this protein to oxidative stress response, a fact well known even before the association of DJ-1 with PD (Mitsumoto et al., 2001), mitochondrial function (Zhang et al., 2005) and transcription (Zhong and Xu, 2008), nuclear transcription (Xu et al., 2005), mRNA binding and protein interaction (Hod et al., 1999; van der Brug et al., 2008) and protein degradation (Xiong et al., 2009). Mirroring the involvement of DJ-1 in multiple cellular activities, this protein has been found in complex with multiple molecular partners, including DJ-1 itself (Tao and Tong, 2003), PINK-1 and Parkin (Xiong et al., 2009), alpha-synuclein (Meulener et al., 2005).

DJ-1 has been shown to modulate dopamine toxicity in cellular models of oxidative stress with reference to PD (Fasano et al., 2008). Dopamine exposure leads to upregulation of DJ-1 that in turn increases cell resistance to dopamine itself and reduces intracellular oxidants (Lev et al., 2008). On the other hand, α-synuclein overexpression leads to upregulation of DJ-1, and DJ-1 overexpression protects cells from α-synuclein toxicity (Batelli et al., 2009). Besides being in complex with a number of different partners, DJ-1 is often post-translationally modified. DJ-1 modifications mainly include oxidations at different sites, which is related to its antioxidant role (Choi et al., 2006), but there are also evidences of ubiquitination (Olzmann et al., 2007) and SUMOylation (small ubiquitin-like modifier) (Shinbo et al., 2006). Not unexpectedly, DJ-1 binding to its molecular counterparts, and thus its pleiotropic effects, are affected by DJ-1 post-translational modification. For example, oxidation regulates homodimerization (Ito et al., 2007) and affects DJ-1 binding to mRNA (van der Brug et al., 2008).

Multiple DJ-1 modified forms are simultaneously present, so that DJ-1 can be better considered as a pool of different forms, with different modifications and in different amounts. It is very likely that, instead of the total amount of DJ-1 or of a defined DJ-1 form, the composition of this pool and the precise balance between different forms is the main factor determining DJ-1 global activity.

In particular, alterations of this pool, instead of DJ-1 mutations, are expected to play a role in those non-genetic conditions correlated to DJ-1 activity, including idiopathic PD. Indeed, since many different DJ-1 forms can be separated on the basis of their isoelectric point (pI), it is a common finding that DJ-1 oxidation, correlated to

aging, Parkinson's and Alzheimer's diseases, produces an increase in DJ-1 species of acidic pH, and a decrease in basic species, so that these conditions are characterized by a pool of DJ-1 forms different from that observed in controls (Choi et al., 2006). Ubiquitination and SUMOylation of DJ-1, on the other side, affects also the molecular weight (MW) of the modified species, and are thus separated by mono-dimensional electrophoresis accordingly (Olzmann et al., 2007).

6.2 Principles of 2D-GE meta-analysis

While two-dimensional electrophoresis (2D-GE) is able in principle to separate DJ-1 modified forms on the basis of both pI and MW changes, to the best of our knowledge there is no experimental evaluation of DJ-1 modifications, which takes into account both dimensions simultaneously, so as to completely describe the pool of DJ-1 possible forms and to evaluate the pool composition in a given condition.

To this purpose, a robust procedure was developed to re-evaluate 2-DE experiments available in the literature by automatic alignment. Since detailed information on experimental conditions, protocols and samples was available, it was possible to segregate different DJ-1 forms into pools associated with a defined condition independently by the particular experimental protocols used (Natale et al., 2010).

Interestingly, the DJ-1 pool from neural tissues, in particular from brain, displayed a specific and characteristic MW and pI pattern. Moreover, changes in this pattern might reflect neurodegenerative processes and aging. These results were experimentally validated by 2-DE western blotting on human post-mortem brain specimens from control subjects and PD patients.

The meta-analysis of proteomic data provides valuable information pertinent to various biological processes (or methods involved) that otherwise remains hidden if a single paper is considered. The occurrence of frequently identified proteins, for instance, may represent common cellular stress responses or simply reflect technical limitations of 2-DE, as previously described in the Chapter 4.

This chapter presents an innovative procedure to examine the status of DJ-1, a protein extensively and heterogeneously modified, by comparing its 2-DE pattern documented in 160 experiments from 37 different papers. Despite the presence of many different forms of the protein, with different pI and MW, it was found that DJ-1 main forms consistently associated with two different pools, with a different MW and average pI. These two pools are mutually exclusive, and the pool with a MW $\approx$ 20 kDa and a more acid average pI exhibits strong tissue specificity for brain and neural models. Moreover, we could show that this pool is indeed found experimentally in the SNpc of both PD patients and control subjects and, in agreement with the previous literature, there is a tendency at the level of population toward more acid forms in PD patients.

Altogether, these findings support the value of a proteomic meta-analysis for the unraveling of complex proteomic signatures and biomarker analysis.

6.3 Performing a 2D-GE meta-analysis

A set of relevant data to be included into the analysis, i.e. 2D-gel experiments where the protein DJ-1 was clearly identified, was obtained by extracting the relevant 2D-gel images from Pubmed-indexed Portable Document Format (PDF) documents. The document figures containing 2D-gel representations were identified by a text-matching procedure, and able to evaluate each caption for the occurrence of at least one term identified as connected to a 2D-gel experiment. This step led to the identification of 53 figures from 37 papers, which were automatically segmented to isolate 160 2D-gel images. The 2D-gel images were processed using ImageJ (National Institutes of Health) (Girish and Vijayalakshmi, 2004), including DJ-1 spots that could not be directly aligned, given the differences in resolution of the original figures and the different pI/MW range represented. Therefore, gel images were rescaled by calibrating the pixel-to-pI ratio using the following factor:

$$F = \mathrm{pix}_{pI} \times \frac{\mathrm{pI}_B - \mathrm{pI}_A}{\mathrm{pix}_B - \mathrm{pix}_{pA}}$$

where pix_{pI} was the desired resolution in the final image (i.e. 400 pixels per pI unit), pI_B and pI_A were the pI values of the center of mass of the most basic and of the most acid spot, respectively, and pix_B and pix_A the pixel number of the center of mass of the most basic and of the most acid spot, respectively. For MW calibration, the average position of the mass center of the four most intense DJ-1 spots was assumed to correspond to the mass reported for DJ-1. Once rescaled and registered on the reference pI/MW grid, gel images were background subtracted, segmented for the spots and binarized. To this aim, a rolling-ball procedure was used first (rolling ball radius = 50 pixel, performing a sliding paraboloid), then a watershed segmentation procedure was applied (smooth radius 5.0, dark objects on bright background, neighbourhood of 8 pixels, minimum level ‘0’, maximum level ‘120’, display outputs ‘binary image’) and finally an automatic thresholding to get the binarized image (Pleissner et al., 1999).

The 160 binarized images were used to get a ‘metagel’, i.e. an average intensity projection image, wherein each pixel stores the average intensity over all images in the 160 original images at corresponding pixel location (Figure 1). The color scale of this metagel accounts for the occurrence frequency of each spot in the considered experiments, with less frequent spot positions in blue and more frequent spot positions in white. The same approach was used to build two additional ‘metagels’

which represent all the experiments of Alzheimer's/PD patient brain (n=11) and all the experiments of control human brain origin (n=12).

Figure 1 reports the distribution of all the DJ-1 major forms represented in the scientific literature considered in the present work. From this Figure, it becomes immediately evident that the protein was reported with different MWs (20 and 25 kDa).

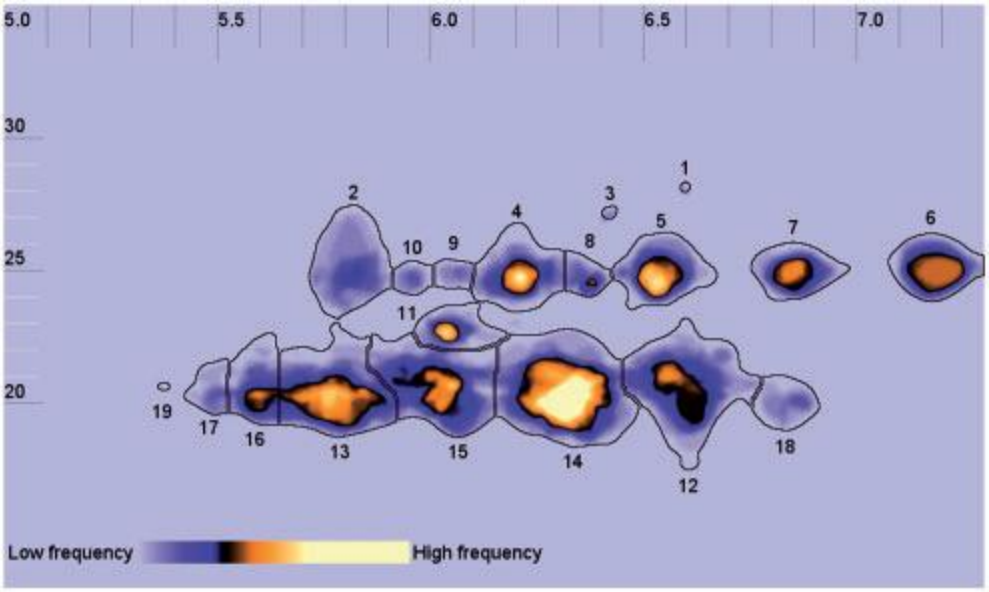

Figure 1. DJ-1 metagel, derived from 160 different bidimensional experiments, representing the overall pool of DJ-1 possible forms. Conserved spots are numbered and bordered in black. The color scale range from blue (spots poorly represented in the considered experiments) to white (spots very represented in the considered experiments).

The spots in the two groups are separated on the basis of the apparent MW, because the first group includes only DJ-1 forms with MW≈25 kDa, while in the second group only forms with MW≈20 kDa are present. Therefore, DJ-1 appears either at the expected MW (MW≈20 kDa) or as an heavier, modified form. The latter cannot be explained by a different sequence, since all the species considered in the present study (human, mouse, ovine and bovine) express a protein of predicted MW≈20 kDa. Moreover, DJ-1 forms with MW≈25 kDa are widely represented in the considered papers, and therefore they are not correlated with a particular protocol or a particular laboratory.

The existence of a 25 kDa DJ-1 form is documented in the literature also by a number of one-dimensional western blot experiments. For example, human DJ-1 of a similar MWappears in the works of Shinbo (Shinbo et al., 2006) and Bieler (Bieler et al., 2009), and in rat a similarly heavier DJ-1 was observed in the work ofYanagida (Yanagida et al., 2006). Further support to the existence of a heavier form of DJ-1 comes from the one-dimensional western blot reported as quality control by DJ-1 commercial antibody manufacturers (Supplementary Table 3). Again, it appears that, independently from the particular antibody used, the protein appears alternatively as a

25 kDa or 20 kDa band. Like the MW, the pI of the two identified pools is also different. The average pI of the 25 kDa pool tends to be more basic, if compared with the average pI of the 20 kDa pool, so that the increase in MW appears to be correlated to a corresponding increase in pI.

6.4 'Spot matrix' generation

In order to gain insight on possible correlations between specific spot patterns on 2D-gel experiments and biologically relevant information, an objective matrix was generated to correlate each spot in a given 2D-gel to one of the reference spot of the metagel, i.e. to one of the DJ-1 forms occurring in at least 5% of the considered experiments. Briefly, the following, fully automated procedure was performed:

(i) the 'metagel' was segmented as described above for the experimental gels;
(ii) the Euclidean distance of the center of mass of each experimental spot from the center of mass of each spot in the 'metagel' was calculated
(iii) a 'spot matrix' was build, where rows correspond to two dimensional gel (2D-gel) experiments, and columns to the spots in the 'metagel'.

A value of 1 was assigned if the Euclidean distance between any spot of the considered experiment and the 'metagel' spot defining the matrix column was minor or equal to the Feret diameter of the 'metagel' spot (i.e. the considered experiment contains a spot that is covered at least partially by a specific spot of the 'metagel').

The 'spot matrix' so obtained is thus a synthetic representation of the occurrence of each spot, identified in the 'metagel' and corresponding to a form of DJ-1 in a given experiment, allowing the description of each experiment in terms of which DJ-1 forms were present.

Correlated DJ-1 forms (i.e. DJ-1 forms that appear together in different experiments) were identified by building the correlation matrix reported in Figure 2A, where the cells of the matrix are colored according to the Pearson's correlation coefficient of the corresponding spot pair (see Section 3). Two groups of spots are positively correlated: a first group, including spots 4, 5, 6, 7, 8, and a second group, including spots 12, 13, 14, 15, 16, 17. Out of 160 experiments considered in the present work, 68 contained at least one spot of the correlation group 1, and 54 contained at least one spot from correlation group 2.

Interestingly, spots that are positively correlated in one group tend to be negatively correlated with the spots of the other group, so that the two identified pools are mutually exclusive. This fact could be due to at least one kind of post-translational modification of DJ-1, changing the mass of the protein, which acts like an alternative switch. The functional meaning of this switch will be investigated in future studies.

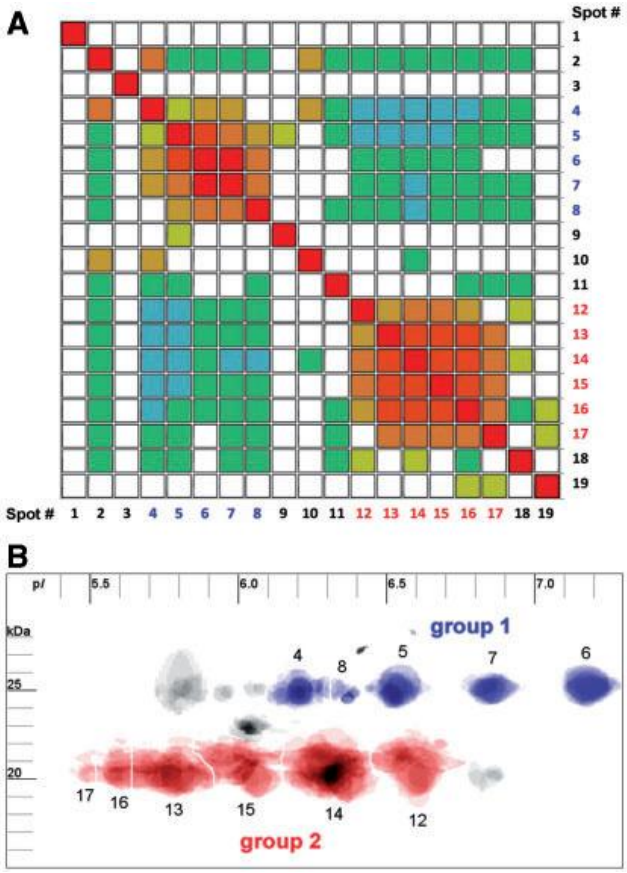

Figure 2. (A) Spot correlation matrix. Cells of the matrix are colored according to the Pearson's coefficient of correlation, from blue (negatively correlated spots) to red (positively correlated spots). White cells represent absence of correlation between the corresponding spots. (B) The correlation analysis identified two groups of positively correlated spots, colored in red and blue, respectively, which are mutually exclusive. The average MW and pI of the two groups is different (heavier and more basic in group 1, lighter and more acidic in group 2).

6.5 Statistical analysis of meta-analysis data

To ascertain whether the identified pools have any biological relevance, all the experiments (i.e. all the 2-DE patterns) were grouped according to their similarity in terms of Euclidean distances between the corresponding rows in the 'spot matrix'.

The spot matrix was analyzed using the software XLSTAT (Addinsoft, Paris, France) as described hereafter. The co-occurrence of different spots was measured to group those spots which most often are present simultaneously (i.e. to get pools of DJ-1 forms) according to Pearson's correlation method applied row by row in the 'Spot matrix': the linear correlation was measured between the presence of a given spot in a given experiment (expressed by a value of 1 in the corresponding cell of the spot matrix) and the presence of the other spots in the same experiment (expressed by the value, 0 or 1, of the other spots in the same row). The correlation alpha significance

threshold to include a spot in a pool was set to 0.01.

Grouping of the considered experiments was achieved by agglomerative hierarchical clustering in order to put together different experiments on the basis of the similarity of their vectorial representation in the 'spot matrix', i.e. on the basis of the occurrence of similar spots. The similarity between experiments was computed as Euclidean distances for each pair of rows in the 'spot matrix'. The Euclidean distances (d) are given for each pair of rows, A=(a_1, a_2, ...,a_p) and B=(b_1, b_2, ...,b_p), by the equation:

$$d(A,B) = \sqrt{\sum_{j=1}^{p} (a_j - b_j)^2}$$

where p is the number of columns/variables (19 in our case). The resulting distance matrix was used to cluster the experiments according to the Ward's method (Ward, 1963). Here, the clustering criterion is based on the error sum of squares, E, which is defined as the sum of the squared distances of individuals from the center of gravity of the cluster to which they have been assigned. Initially, E is 0, since every individual is in a cluster of its own. At each stage, the link created is the one that makes the least increase to E. Aggregation was stopped by automatic truncation, based on the entropy level, so to get groups as homogeneous as possible.

The resulting dendrogram is reported in Figure 3. Most of the experiments segregates into two groups: one contains mainly 2-DE gels from 'human, brain' samples (cluster 2, including neural cell lines), while the other contains experiments of heterogeneous, mostly non-neural and/or non-human origin (cluster 1).

Figure 3. Cluster analysis of the considered 2D experiments, based on the presence/absence of spots shown in Figure 1. Three clusters are formed according to the method explained in the text; cluster 2 contains only experiments on DJ-1 of neural origin.

Moreover, there is a separate 'non-neural' cluster (cluster 3), which contains experiments coming from a single paper, with an overall DJ-1 pattern not supported by any other paper and consequently suspicious of experimental biases. For this

reason, we did not consider this cluster further. Interestingly, the 'human, neural' cluster 2 contains only three samples originating from cellular lines (dopaminergic neuroblastoma M17 line), while all the other 23 experiments in this cluster are 2-DE conducted on proteins from human brain.

Once the experiments were grouped, it was possible to verify whether the clusters formed on the basis of the similarity between the images could be supported also by some non-biological bias (such as the paper of origin of the images, the authors of the experiments and the particular protocol used). The latter hypothesis would imply either that the results are strongly dependent on the particular author presenting them (i.e. they are not reproduced by others) or they are extremely sensitive to the experimental conditions used (i.e. they are not reproducible). Beside the aforementioned bias in cluster 3, both clusters 1 and 2 contain experiments from different papers, with different staining procedures, so that neither of these bias factors could explain the different DJ-1 patterns responsible for the observed clustering.

Therefore, it is possible to distinguish, at least in humans, the brain origin of DJ-1 on the basis of its 2-DE pattern, i.e. there is a specific experimental pattern associated with DJ-1 of human brain origin.

Chapter 7

Open source software for 2D-GE image analysis

A number of commercial software packages are currently available to perform digital two-dimensional electrophoresis (2D-GE) gel analysis. However, both the high cost of the commercial packages and the unavailability of a standard data analysis workflow, have prompted several groups to develop freeware systems to perform certain steps of gel analysis. Unfortunately, none of them offer a package that performs all the steps envisaged in a 2D-GE gel analysis. This Chapter describes an ImageJ-based procedure, able to manage all the steps of a 2D-GE gel analysis. ImageJ is a free available image processing and analysis application developed by National Institutes of Health (NIH) and widely used in different life sciences fields as medical imaging, microscopy, western blotting and PAGE. Nevertheless no one has yet developed a procedure enabled to compare spots on 2D-GE gels. The workflow described allows us to perform the whole 2D-GE analysis.

Moreover, this Chapter presents an effective technique for the detection and the reconstruction of over-saturated protein spots. Firstly, the algorithm reveals overexposed areas, where spots may be truncated, and plateau regions caused by smeared and overlapping spots. Next, it 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. Pixel correction in saturated and smeared spots allows more accurate quantification, providing more reliable image analysis results, particularly important when a meta-analysis of 2D-GE image is performed on images downloaded from web repositories.

Section 1 presents a rapid overview of the available 2D-GE image analysis software, Section 2 describes a 2D-GE image analysis performed by using an open source stack of software, and in Section 3 discusses the results obtained from this analysis. Section 4 discusses the prerequisites for implementing an effective detection the

reconstruction of over-saturated protein spots. Section 5 describe the algorithm for the detection of over-saturated protein spots in 2D-GE image. Finally, Section 6 presents the algorithm for the gaussian extrapolation of 2D-GE saturated protein spots.

Since the pioneer work of O'Farrel (1975), two-dimensional gel electrophoresis (2D-GE) has been demonstrated to be the most comprehensive technique for the analysis of proteome, allowing the simultaneous analysis of very large sets of gene products. In the post-genomic era, 2D-GE becomes a powerful tool that is widely used for the analysis of complex protein mixtures extracted from cells, tissues, or other biological samples (Gorg et al., 2004).

The main goal of 2D-GE is to match protein spots between gels and define differences in the expression level of proteins in different biological states (e.g., healthy versus diseased, or control versus treated).

In a 2D-GE gel thousands of proteins are separated in well defined spots; these protein spots can be revealed via a variety of staining techniques (Coomassie, Silver Stain, Sypro), and captured by one or more digitized computer images per gel(CCD camera, laser scanner, and optical scanner). The image capturing phase transforms the biological information of the 2D-GE gel into a quantitative computer-readable data set (Miller et al., 2006). Once all the studied gels have been collected and digitized the software-based image analysis can be started. Image analysis is crucial in extracting biologically relevant information from a two-dimensional gel electrophosesis experiment. Despite the availability of several software applications to analyze 2D-GE images, there is no general consensus on 2D-GE data analysis protocol (Berth et al., 2007).

7.1 Available 2D-GE image analysis software

Moreover several authors reported that the commercial packages are time consuming, can often miss values or give false positives, and induce variance in quantitative measures (Millioni et al., 2010). The commercially available software perform the analysis workflow in two different ways. The classical package condensed the information onto spots. The spot detection is performed prior to matching and expression profile extraction. The second image analysis software group is based on the whole image information. These packages apply a warping procedure to remove running differences between gels, and the spot detection and protein expression profiles extraction occurred in a separated and independent step.

The emphasis in this analysis software has been on reducing the subjectivity of the image analysis.

The fact that the alignment step is performed prior to the spot detection facilitates simultaneous spot detection on all gel images in an experiment and the resulting spot boundaries are identical on all gel images (Dowsey et al. 2010). In Table 1 are collected the most popular commercial software for 2D-GE gel analysis.

Table 1. The most popular commercial software for 2D-GE gel analysis.

Software package	**Company**	**Type**	**Web link**
PDQuest	BioRad, Hercules, CA	Spot based	www.bio-rad.com
ImageMaster 2D and DeCyder	GE Healthcare	Spot based	www.gehealthcare.com
Dymension	Syngene, Cambridge, UK	Spot based	www.syngene.com
Melanie	GeneBio, Geneva, Switzerland	Spot based	www.genebio.com
Delta2D	Decodon, Greifswald, Germany	Warping	www.decodon.com
Progenesis SameSpots	Nonlinear Dynamics, Newcastle, UK	Warping	www.nonlinear.com

Several research groups have developed freeware systems to handle certain key aspects of gel analysis, including archiving (SwissProt 2D)(Appel et al., 1999), comparison (Flicker)(Lemkin, 1997), interactive exploration (WebGel)(Lemkin et al., 1999), registration (bUnwarpJ and Sili2DGel)(Sorzano et al., 2008)(Pérès et al., 2008), spot detection (dos Anjos et al., 2011), spot quantification precision and differential expression (Pinnacle)(Morris et al., 2010). However nobody has developed a complete package freely available and platform independent able to perform all the steps of a 2D-GE gel analysis experiment (Marengo et al. 2005).

7.2 Open source workflow for performing 2D-GE image analysis

Leveraging also on these experiences this Chapter presents an image analysis workflow based on the popular public domain image analysis software package

ImageJ[1]. ImageJ and its plug-in is easy-to-use software and can be used in routine applications. The workflow has been developed according to the whole image information procedure (Bandow et al., 2008). It is based on six steps:

- aligning all the images;
- computing image fusion;
- creating a consensus spot pattern;
- propagating the consensus spot pattern to all gel images for quantification;
- statistical analysis.

In order to test the procedure, in this Chapter provide an example of 2D-GE image analysis performed on plasma from patients immediately after an acute myocardial event, comparing the results obtained using a widely diffused commercial package (Melanie; GeneBio, Geneva)(Natale et al., 2011). We looked for biomarkers of pathology and/or treatment in acute myocardial infarction (AMI) patients treated with common anticoagulant protocols.

With this aim, we enrolled 9 patients admitted within 6 hours after the onset of chest pain symptoms, with myocardial infarction defined according to ESC/ ACC criteria. All subjects signed informed consent forms prior to standard sample collection. 2D-GE was performed according to Maresca (2010) and each sample was run in duplicate.

Gels of plasma samples were visibly stained with Coomassie Blue, scanned using transmission mode to avoid saturation effects and saved in 16-bit TIFF format. Once all gels in the study had been collected and digitalized, they were analyzed using the ImageJ and some of its plugins or the commercial software Melanie. In Table 2 is collected the list of steps that describes how to perform the analysis using ImageJ and its plug-in.

Table 2. List of steps that describes how to perform the analysis.

Step	Description	Web link
1	Download and install ImageJ on your computer following the installation instructions specific to your platform (Windows, Mac OS, Linux, etc.);	http://rsbweb.nih.gov/ij/download.html
2	Align all images in pairs using bUnwarpJ plugin and taking always the same image as reference, and save the warped images;	http://biocomp.cnb.uam.es/~iarganda/bUnwarpJ/

[1] ImageJ: ImageJ is a public domain, Java-based image processing program developed at the National Institutes of Health. ImageJ was designed with an open architecture that provides extensibility via Java plugins and recordable macros. (http://rsb.info.nih.gov/ij/)

Step	Description	Web link
3	Open all the warped images and save these in a stack as a sequence using the "Image > Stacks > Images To Stack command";	
4	Sum image using "Image.Stacks.ZProject>Sum Slices";	
5	Perform spot detection on the fused image by the Watershed plug-in. Selected the binary output;	http://bigwww.epfl.ch/sage/soft/watershed
6	Apply the blob analyzer of ImageJ using "Analyze > Analyze Particles…", to measure the catchment basins and save the blots as a list of ROI;	
7	Open the stack image(saved in point 4) and propagated to all gel images the list of ROI obtained in by the spot detection procedure "ROI Manager > Show All";	
8	Measure the spots volume values using "ROI Manager > Measure" and save the Results as OpenOffice compatible (.ods) file;	http://www.openoffice.org
9	For quantitative comparison of spot intensities choose Integrated Density measure. This value is the integral of all pixel intensities within the spot boundary;	
10	Normalize the volume of each spot on a given gel image versus the total volume of all spots on that image, perform the ANOVA Test on normalized data.	

First, all images were warped by bUnwarpJ (Sorvano et al., 2008), an algorithm for elastic and consistent image registration developed as an ImageJ plug-in. It performs a simultaneous registration of two images, allowing us to solve the problem of spatial distortions due to run-time differences and dye-front deformations. The warping step is showed in Figure 1.

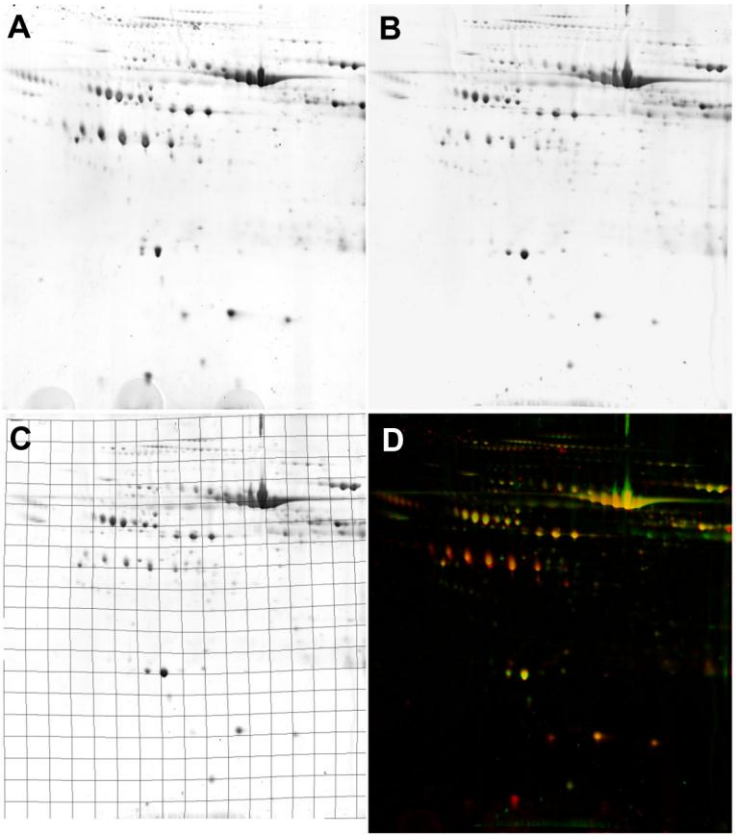

Figure 1. (A and B) shown two different 2D-GE images of two control subjects. In (C) is shown the elastic registration obtained during using bUnwarpJ plug-in. In Figure 1D is shown the overlap of the two gels after the warping step, in the red channel is shown the reference gel (Fig. 1A) and in the green channel is shown a warped gel.

bUnwarpJ aligns all images in pairs, taking always the same image as reference and producing the corresponding warped images of the others. The reference image and warped images were subsequently displayed in a single stack image and summed to generate a fused image. We followed an image sum approach to retain as much information as possible from the original images.

Spot detection was performed on the fused image by the watershed plug-in written by Daniel Sage (Tsukahara et al., 2008). This plug-in is able to segment an image using the watershed algorithm by flooding directly on graylevel image. Of the several kind of outputs provided, we selected the binary output that allows us to apply the blob analyzer of ImageJ, so as to measure the catchment basins and save the blots, one for each protein spot, as a list of regions of interest (ROI). Each ROI corresponds exactly to a spot in the fused image. The list of ROI obtained by the spot detection procedure was our consensus spot pattern that is valid for the whole gel set of the experiment.

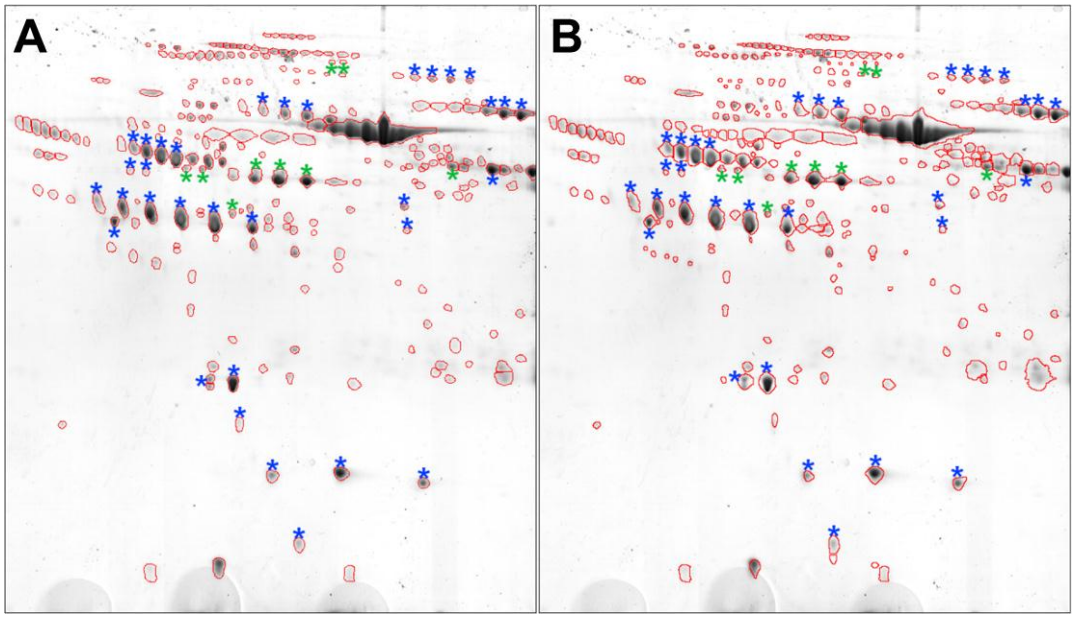

Figure 2. 2D-GE from the plasma of a control individual. (A) Spots detected and matched using Melanie. (B) Spots detected using the ImageJ procedure. **Note**: The stars show the spots used for comparison of quantification methods (unchanged spots in blue, differential spots in green).

In other words, the list of ROI obtained is equivalent to the grid used in gene chip analysis, this grid was imposed on each of the aligned gel images so that a defined number of areas were quantified on every gel image of the experiment.

ImageJ procedure was capable of detect 232 conserved spots, while with Melanie identified a pattern of 205 matched spots. The spot detection and the matching were manually checked in both the procedures; Figure 2 shows the spots detected by the two procedure on one of the control subject gels.

The spot volume values extracted from each image were listed in a ImageJ "Results table". The resulting table of "the whole image information procedure" did not have empty cells, while some commercial software, such as Melanie, are not able to eliminate all bias due to missing spot values (Chang et al., 2004). All the data were analyzed by Calc (OpenOffice)[2]. For the normalization the volume of each spot on a given gel image was diveded by the total volume of all spots on that image (Nishihara and Champion, 2002). The resulting table of our method did not have empty cells, while some commercial software, such as Melanie, are not able to eliminate all bias due to missing spot values (Chang et al., 2004).

The scatter plot in Figure 3 shows that there is a linear relationship between the spot volumes evaluated by Melanie and the corresponding values obtained by the ImageJ procedure. In particular 42 spots, 33 more abundant spots (blue stars in Fig. 2) and the 9 differentially expressed spots (green stars in Fig. 2, see the next paragraph for the identification procedure) were considered for the comparison; the fact that the straight line in Figure 3 has a slope 1 means that volume values calculated for the

[2]OpenOffice: OpenOffice is an open source suite of programs downloadable from the web site http:// www.openoffice.org/.

same spot are on average larger by using the ImageJ procedure, which can be related to the slightly larger area segmented for each spot by ImageJ due to the fact that spots were segmented on the fused image (and not on every single gel, as Melanie does).

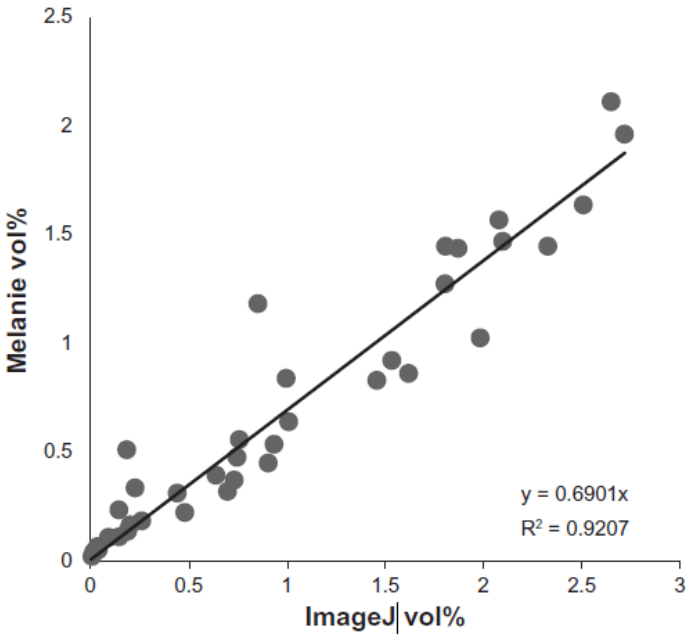

Figure 3. Scatter plot of spot mean volumes as evaluated by Melanie and ImageJ. **Notes**: The ImageJ values are plotted on the X-axis, the Malenie values on the Y-axis. Spots were normalized based on total spot volume.

7.3 2D-GE image analysis results

The 9 differential spots (shown in Figure 4), whose mean normalized volume was significantly decreased in the myocardial infarction versus the control group, were all identified by t test (P-value ,0.001); the test was run independently on the list of the spots, quantified by ImageJ-based procedure or by Melanie. 7 out of the 9 spots were well above the selected *P*-value threshold for both Melanie and the ImageJ-based procedure, while each of the 2 methods identified an additional spot which was missed by the other (with reference to Figure 4, spots 133 and 173 where identified only by Melanie and ImageJ respectively).

These 2 spots have a *P*-value slightly higher than the threshold and anyway with a significance under 0.05. All data of the spots are shown in table in supplementary material. By using the procedure described by Lemkin (Lemkin and Thornwall, 1999), by matching the spots of a gel with those of a reference map of human

plasma[3], we were able to tentatively identify 5 of the 7 significantly different spots as fibrinogen gamma chain fragments (with reference to Figure 4, spots 142, 143, 146, 147, and 148).

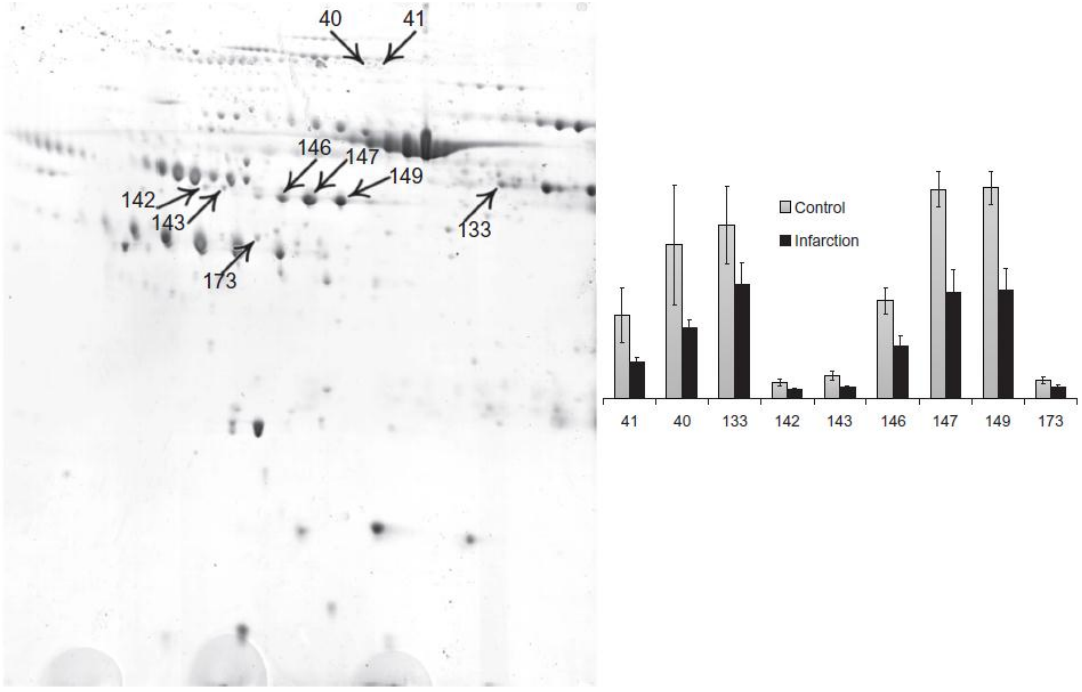

Figure 4. Profile of proteins differentially expressed in the plasma of myocardial infarction (MI) patients vs. control subjects. 7 spots were found significantly decreased in the plasma of the myocardial infarction patients by both methods (*P*-value , 0.001 , ANOVA test). Spots 133 and 173 were found to be differentially expressed by one method only (see text for explanation). Spots 142, 143, 146, 147 and 149 were identified as fibrinogen gamma chain fragments.

In conclusion, the ImageJ-based workflow is a free and easy alternative to a common commercial package for the segmentation and quantification of 2D gel spots; the procedure proved to be so effective, to confirm the results obtained by an established commercial solution.

This solution can help proteomic laboratories to quickly and inexpensively evaluate 2D-gel experimental results, without losing the required accuracy and providing a common reference for future analyses.

[3] Reference map oh Human plasma at http://expasy. org/swiss-2dpage/viewer.

7.4 Reliable reconstruction of protein spots in saturated 2D-GE image

An important prerequisite to guarantee optimal performance of automated image analysis packages is good image capture, which should allow for detection of various protein amounts ranging from very low to high abundance. In particular, when a gel image is acquired, it is important to avoid saturation effects, which occur when grey levels exceed the maximum representable level on the used scale (e.g., in an 8-bit image an overexposed or saturated area contains a large number of pixels with gray-level values close to 256).

In practice, to avoid saturation during the scanning procedure, the operators should assure that the more abundant protein spots are represented by pixels slightly below the maximum intensity available, while keeping the dynamic range as wide as possible and acquiring images at 16-bit rather than 8-bit, because a 16-bit image has 65,536 gray levels while a 8-bit has only 256 gray levels.

The use of fluorescent protein stains can extend the linear dynamic range by 3-5 orders of magnitude (Miller et al., 2013). However, the amplitude of the dynamic range is still a relevant issue in biological applications, since protein concentrations in biological systems may vary by seven or more orders of magnitude. Currently, there is no acquisition system that can cover this wide range. In practice, the differences in protein concentration in biological systems compromise the weak and low abundant protein spots, which are often more biologically relevant, but generally too faint to be accurately quantified (Hortin and Sviridov, 2010).

The intensity of a pixel is expressed within a given range, whose depth indicates how accurately the grey value can be represented. 2D-GE images are usually acquired at 16 bit resolution in the grey scale, resulting in maximum grey values of 2^{16} . The bit depth together with the type of staining influences the linear dynamic range. Some of the most widely-used staining methods are summarised in Table 3. The linear dynamic range refers to the number of orders of magnitude where the grey scale values increase proportionally with protein amount. In other words, the orders of magnitude define the volume difference between smallest and largest spots that be reliability quantified. For example, 1 order of magnitude means a difference from 1 to 10, while 5 orders from 1 to 100,000. The linear dynamic range is necessary to examine these 2D-GE spots containing varied concentrations of the target proteins, so that the correlation between protein concentration and spot volume can be generated and evaluated.

Table 3. Some of the most widely used staining methods

Stain/Dye	Residues associated	Sensitivity (ng)	Linear dynamic range (order)	MS compatibility
Coomassie Brilliant Blue	Arginine, lysine	8–10	1	Good
SYPRO Ruby	Primary amines	1	3-4	Good
Deep Purple	Primary amines	>1	4	Good
Minimal Cy Dyes	Lysine	0.1–0.2	4-5	Challenging

Moreover, due to the highly complex nature of biological samples, in most cases several proteins have similar pI and MW values, and the corresponding spots migrate to the same portion of the gel image, resulting in complex regions with multiple overlapping spots. In this context, commercial software available currently often requires manual revision of spot detection and refinement of computer generated matches (Clark and Gutstein, 2008).

The aim of the spot detection step is to define the location, the true boundary and the intensity (usually the total pixel volume within the boundary) of each spot. In particular, in order to achieve fast and reliable image analysis, the algorithm used for image segmentation must provide well-defined and reproducible spot detection, quantification and normalization (dos Anjos et al., 2011). However, despite the availability of several spot detection algorithms, the accurate definition of protein spot might be critical in certain regions, where defining the boundaries is challenging due to variable background, spots overlapping and spot saturation (Daszykowski et al., 2010).

In a saturated spot, intensity values are truncated, preventing the resolution of high intensity pixel, since the missing area of the spot will not be measured. In other words, a saturated spot does not provide any reliable quantitative data, and it might also compromise the overall normalization. In particular, when comparing different experimental states for expression changes representative of a particular condition, the inclusion of highly to medium abundant and saturated spots might bias normalization, especially if their variance is a significant portion of the total spot volume (Miller et al., 2006). For these reasons, several authors recommend manually deleting the saturated spots, before analysing the gels (Berth et al., 2007). In fact, currently available commercial software (as Delta2D, ImageMaster, Melanie, PDQuest, Progenesis and REDFIN) or academic packages (as Pinnacle and Flicker) are not able to deal with specific protein spot distortions found in the gel images (Maurer, 2006).

In a recently-proposed approach to improve spot detection, Kim and Yoon (Kim and Yoon, 2009) noticed that the gradients of the valley under a plateau spot orient the peak. They therefore suggested a spot separation method that computes the accumulated gradient of each point in the potential complex regions. The accumulated gradient image is then segmented with a watershed segmentation algorithm. Using

this method, they can detect 75%-96% over-saturated protein spots. However, this method is not able to provide quantitative information.

Since automatic detection of oversaturated protein spots is, to date, infeasible, software developers suggest recovering missing data by rescanning the gel at a lower exposure. Although theoretically possible, it must be noted that this is often prohibitive, since gel staining is photosensitive and colours fade away in a few minutes upon light exposure, or because image analysis is performed a long time after the acquisition of the gels so that the expensive 2DE procedure has to be repeated for another scanning (Kim and Yoon, 2009). Moreover, acquiring images with long exposure might provide more information, since the acquired images contain a larger number of spots and thus also the lower-abundance protein spots.

7.5 Detection of over-saturated protein spots in 2D-GE images

This section presents a novel two-step algorithm for detection and reconstruction of over-saturated protein spots (Natale et al., 2012). Firstly, it reveals overexposed areas, where spots may be truncated, and plateau regions caused by smeared and overlapping spots. Secondly, it processes the greyscale distribution of saturated spots, reconstructing the saturated region by a generalized Gaussian approximation.

The method yields a highly-refined and fully-automatic spot detection that does not need further manual corrections. Furthermore, the pixel correction in saturated and smeared spots according to the generalized Gaussian curve allows more accurate quantification, providing more reliable results from 2D-GE analysis.

To validate the proposed method, this section presents the steps of an image analysis in which highly- exposed 2D-GE image containing saturated spots were processed and compared the reconstructed spots to the corresponding spots of the unsaturated image. Results of this analysis proved that the method proposed can effectively reconstruct the saturated spots and improve spot quantification.

In order to test the algorithm, for localizing plateau areas, were examined 12 2D-GE gels with images acquired at 3 different exposures for each gel (36 images in total). After SDS–PAGE, the gel images were acquired by Gel-Doc-it 310 Imaging system (UVP, Upland, CA), with the following setting: exposure 7 s, gain 3; and three different apertures: 6, 7 and 8. Aperture is expressed as F-stop (e.g., F2.8 or f/2.8): the smaller is the F-stop number (or f/value), the larger is the aperture. Using a larger exposure yields a larger number of spots and, at the same time, more saturated areas. In our case, only the image acquired at the lower exposure (Figure 5C) could be properly analysed by commercial and academic software currently available, while

the saturation in some areas in the other two images (Figure 5A and B) prevented accurate evaluation of protein expression.

In order to identify the plateau regions, we implemented a morphological filter, inspired by the rolling-ball algorithm described by Sternberg (1983), which allows segmentation of the plateau zones. This method is based on a structural element (*SE*) defined by a circle of given radius (*RD*) and a grey-scale value tolerance (*GVT*). In particular, for each pixel (x, y). of the image *I*, the *SE* is defined as a circular neighbourhood of *RD*

$$SE = \left\{(s,t) \in I / \sqrt{(s-x)^2 + (t-y)^2}\right\} < RD$$

denotes the spatial domain of the image. The *RD* and the *GVT* are defined by a single parameter. For instance, setting the parameter to 10, the *RD* is 10 pixels and the *GVT* in the 10% of grey values of the pixel at position (x, y).

The centre of *SE* is moved along each pixel of the image and the maximum and minimum grey values of the pixels for each point (x, y) within the given *RD* are calculated. When the difference between maximum and minimum is less than *GVT*, the area defined by the local operator is considered as a plateau area.

An example of the plateau areas detected using our algorithm is shown in Figure 5. In this case, only the image in Figure 5C (acquired at the lowest exposure) could be properly analysed by commercial software available, while the other two images would be discarded because of the large saturated areas. However, in several cases researchers would be interested in processing also the image acquired with the highest exposure (Figure 5A), since it contains the highest number of spots (and thus also the lower-abundance protein spots). It must be said that, in practice, operators only acquire a single image per gel, choosing then the image that can give the largest amount of useful information. In the example considered here, image shown in Figure 5A provide more information due to more spots revealed. However, without an extrapolation approach, this image would be discarded due to the higher number of saturated spots. Therefore, Melanie (GeneBio - Geneva) was used to perform the spot detection on these three gel images and detected 254, 213 and 172 spots in images shown in Figure 5A, B and C, respectively. Hence, being able to analyse the higher exposure images would yield 30% more spots, which represent low abundant proteins and relatively faint spots.

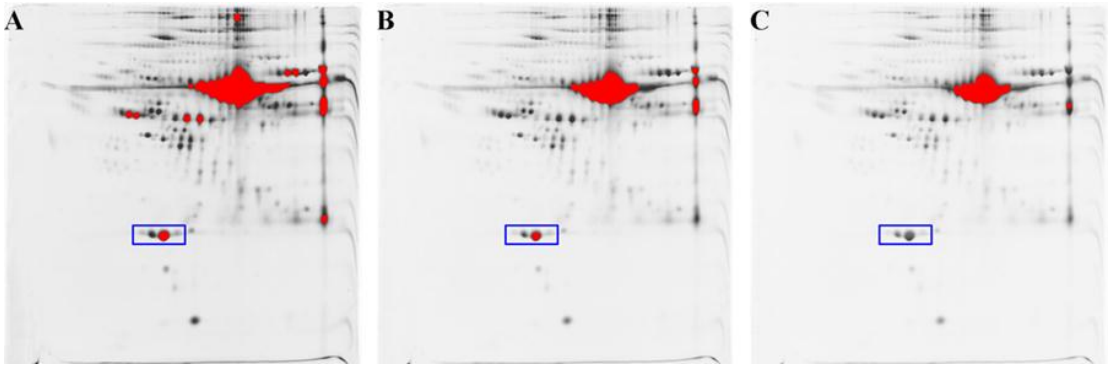

Figure 5. Saturation zones in 2D-GE images acquired at different exposures. The figure shows images of the same gel acquired at three different exposures as high (A), medium (B) and low (C). The saturation zones detected from our algorithm are indicated in red. Spots boxed in blue were further analyzed in figure 5B.

As mentioned earlier, the commercial software is not able to detect or report saturated areas and, furthermore, does not warn the operator of the presence of saturated regions. In fact, saturated spots can only be viewed using the 3D layout, while it would be desirable that the 2D-GE could be detected automatically by the software and the saturated spot were reported before performing an analysis. The proposed method is valid both for over-stained spots and for highly-abundant protein spots. An example of a highly-abundant protein spot is provided in Figure 5, where albumin is the largest spot in the middle of the images. The albumin spots are saturated in all the three images (A, B and C), and our algorithm is able to detect these areas correctly. Figure 5 shows that the same gel can be analyzed with or without saturated spots, by acquiring the images at different exposures. These images provide the opportunity to see how the grey values are distributed in the same spot.

The effect of overexposure on a single spot is shown in detail in Figure 6, which refers to the saturated spot contained in the blue box in Figure 5. In particular, Figure 6A-C show the spot acquired at three different exposure conditions from a greater to lower opening. From the 3D plot of the spot (Figure 6D-F), one can identify a plateau zone for the higher exposures (Figure 6D and 6E), where the red colour denotes the maximum possible value of the grey scale.

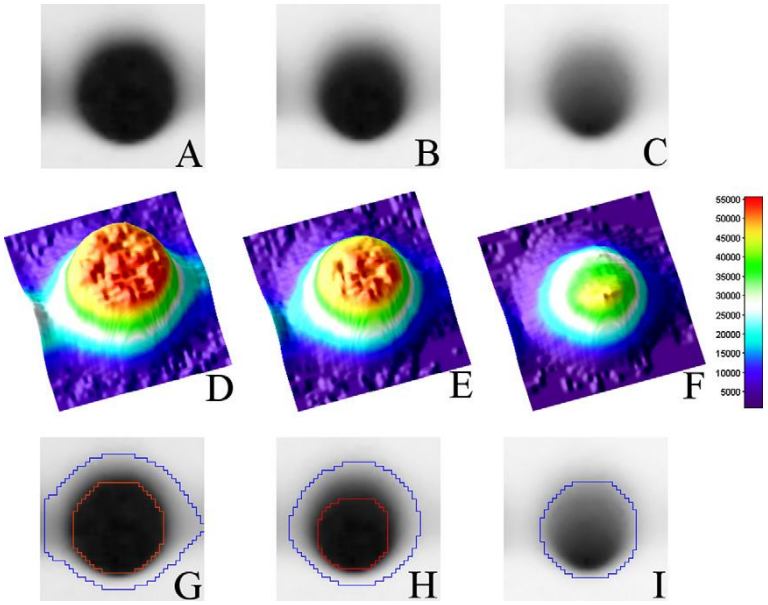

Figure 6. Visualization of plateau areas in 3D view and by our algorithm
The spot shown in panels A–C was the 3rd spot in the blue box in Figure 5 with high, medium and low exposure, respectively. D–F. The 3D visualization of the spot in panels A–C was shown in D–F, respectively, according to the color scale on the right. Red areas in D and E indicated saturated areas containing pixels with similar intensity (the saturated areas contains pixels with gray values close to 65,536 in 16-bit gray scale). G–I. The spots were detected using watershed segmentation, with the spot boundary in blue and the plateau boundary in red as detected by our algorithm. The red areas in D and E coincide with those in panels G and H, respectively.

Finally, Figure 6G-I show the result of the plateau-area finder (red contour) within the spot detection result using a watershed algorithm (blue contour). In particular, we observe that the spot in Figure 6G has a larger area, but also a wide plateau area. The effect of overexposure on spot volume quantification is shown in Figure 7, where the spot intensity distributions are truncated (Figure 7A and 7B).

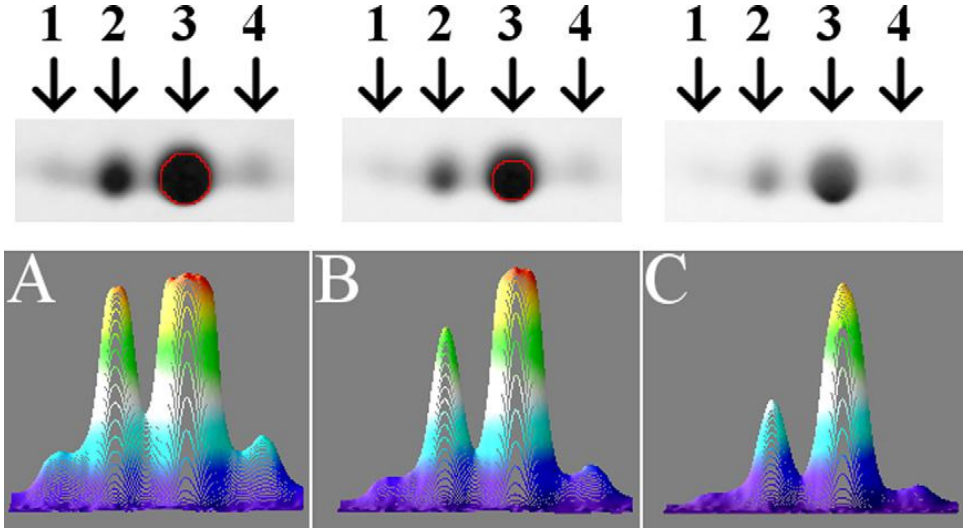

Figure 7. Visualization of unsaturated and saturated spots in 2D view. Upper panel indicated the regions in the blue box as shown in Figure 5A–C with high, medium and low exposure. These regions were visualized in 2D view in the lower panels accordingly. Arrows indicated the spots analyzed in Table 4. The profiles in this figure clearly show the Gaussian distribution of unsaturated spots.

7.6 Gaussian extrapolation approach for 2D-GE saturated protein spots

Before adopting the Gaussian extrapolation to reconstruct the saturated spots, we validated the fitting model on unsaturated spots. As described by Matuzevičius et al. (2007) , the algorithm found a significant correlation between this mathematical model and value distribution for unsaturated spots (Figure 8A and 8C), with an error in volume estimation below 5%. Two reconstructed spots were shown to validate the reconstruction properties of the Gaussian extrapolation method (Figure 8B and 8D).

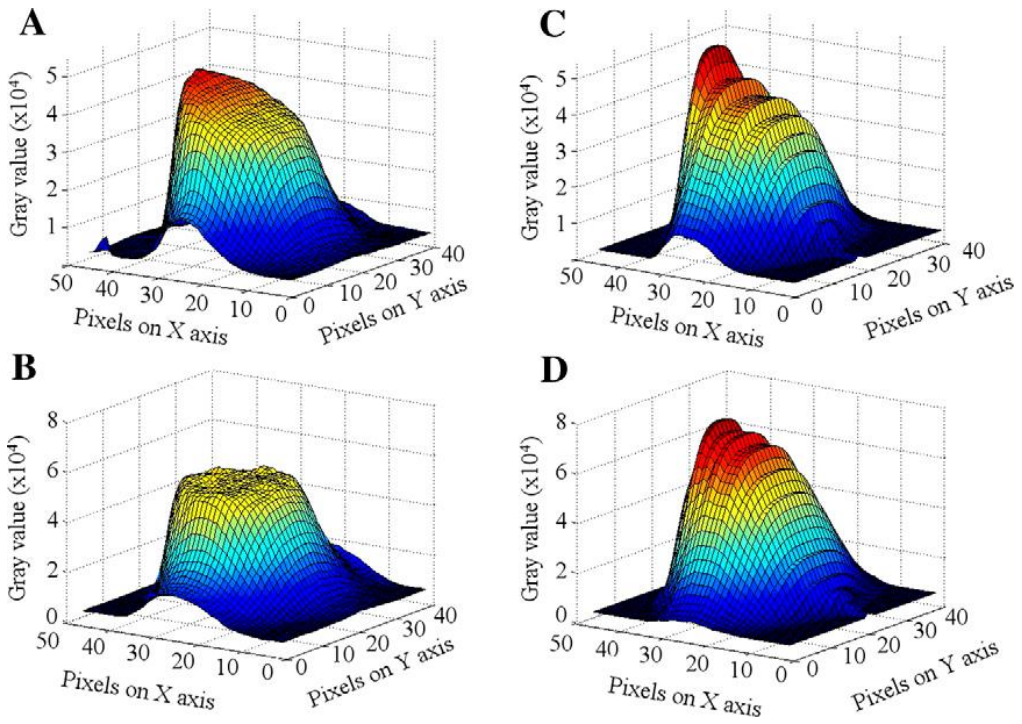

Figure 8. Visualization of unsaturated and saturated spots in 3D view before and after reconstruction. The 3D visualization of the unsaturated spot (Figure 6C) and saturated spot (Figure 6A) was shown in panels A and B, respectively. Pixels are shown on X and Y axes and the gray values are shown on Z axis. Shape reconstruction of the unsaturated spot shown in panel A and saturated spot shown in B was shown in panels C and D, respectively.

The reconstruction of the saturated spots has been done considering the unsaturated spot to be described by an analytical function, depending on a restricted set of parameters. In particular, we assumed each cross section of the spot intensity along both vertical and horizontal axes to be approximated by a generalized Gaussian distribution.

Namely, for each value of the Y-coordinates we considered a function of the form:

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

For b = 2, this equation defines the kernel of a standard Gaussian distribution centred in $x_0(y)$, where $\sigma(y)$ and $M(y)$ is the standard deviation and the maximum of intensity values, respectively. Note that, unlike the other parameters, b does not depend on y, assuming that the approximating Gaussian can have different maximum, center and variance in different sections.

The reconstruction problem can be formulated as follows. Given

$$\{(x_i, y_i), i = 1, \dots, N_x, j = 1, \dots, N_y\}$$

and the corresponding intensities $\{I_{ij}\}$, we determine the set of parameters (M, σ, x_0, b) for which the function defined in this equation fits at best the values of the intensity in the unsaturated region. In practice, we have to minimize an error function that defines how good a particular parameter set is.

For example, a standard least-squared criterion can be used

$$E\big(M(y_i), \sigma(y_j), x_0(y_j), b\big) = \sum_{i=1}^{N_x} \big\| f\big(x_i, M(y_j), \sigma(y_j), x_0(y_j), b - I_{ij}\big)\big\|^2$$

However, the error measure can be defined in different ways, *e.g.*, according to different norms or including different weighs for the different parameters and/or the different pixels.

In particular, we modified the equation in 5 in order to control the variation of the parameters for different sections. In our case the error function in 6 can be formulated as follows

$$\begin{aligned}\hat{E}\big(M(y_j), \sigma(y_j), x_0(y_j), b\big) \\ &= E\big(M(y_j), \sigma(y_j), x_0(y_j), b\big) + \Big(M(y_j) - M(y_{j-1})\Big)^2 \\ &+ \delta_\sigma \Big(\sigma(y_j) - \sigma(y_{j-1})\Big)^2 + \delta_{x_0} \Big(x_0(y_j) - x_0(y_{j-1})\Big)^2\end{aligned}$$

For positive values of the parameters $(\delta_M, \delta_\sigma, \delta_{x_0})$, the problem is then reduced to finding the parameters yielding the minimum of the selected error function. One possibility is to perform an exhaustive search on all the values of a pre-defined parameters space. However, if the size of theparameter space is large, a more effective Newton–Raphson algorithm (Ben-Israel, 1996) to find the zero of the gradient could be employed.

Next, to evaluate the effect of saturation on a quantitative proteomic analysis, was performed a 2D-GE analysis on 12 gels with images acquired at three different exposures (36 images in total). The image analysis was performed using the method describe by Natale et al. (2010). In most cases saturated spots are found with the maximum or medium exposure. The goal of this analysis was to quantify how the spot volume increases with the three different exposures (low, medium and high). The hypothesis was that the volume of the same spot grows linearly with increasing exposure, but with a different proportion depending on whether the spot was saturated or not. Spot detection was performed on the 12 images at higher exposure and then exported all the spots acquired from the images at high exposure to those from the medium and low exposures. the high exposure was chosen because of the highest number of spots detected. From all images, once matched, was extracted the raw volume of the four spots shown in Figure 7 (two small, one medium and one big spot). In order to compare the spots among the different gels, we normalized raw volumes by dividing each spot volume in the medium and higher exposure by the spot volume detected in the gel at the lower exposure.

In Table 4 reported the variation in the volume increase of the saturated and unsaturated spots at different exposures. At the low exposure, all spots have been normalized with respect to themselves (100%). At the maximum exposure, volume of unsaturated spots was increased by more than two times (239%), whereas the volume of saturated spot was increased only 1.5 times (163%), thus losing a considerable amount of information. This problem, if not reported, could lead to errors in statistical analysis, producing inconsistent results.

Table 4. Variation in the volume increase of the saturated and unsaturated spots at different exposures

	Low exposure	**Medium exposure**	**High exposure**
Unsaturated spots	100%	150% ± 21%	239% ± 40%
Saturated spots	100%	129% ± 12%	163% ± 14%

Note: Data was shown as average ± standard deviation.

The method developed in this work allows us to automatically identify the saturated pixels, and recalculate the correct distribution of grey values. In Figure 8B we showed a saturated spot, while in Figure 8D the same spot was shown after applying the algorithm of Gaussian extrapolation.

The method introduced in this chaper allows us to extend the linear dynamic range acting on the bit depth. When the largest spots are saturated because they reach the maximum grey value permitted by the bit resolution, the algorithm is able to reconstruct the saturated spots, recalculating a grey value of individual pixels well above 65,536. In particular, we demonstrated that our algorithm can successfully reconstruct the spot volume up to 2-3 times the maximum grey value. Therefore the values of the pixels of the reconstructed images should be stored as a values array. Of course, the linear dynamic range is strongly influenced by the type of staining used

and by the acquisition tool.

Saturation of abundant spots is an important issue in 2-DE evaluation, in particular when dealing with complex samples like serum or plasma, where protein distribution might be highly heterogeneous. Since commercial software currently available is not yet able to perform 2D-GE analysis in the presence of saturated spots, the only alternatives are based on gel image acquisition or gel rescanning with different parameters.

The proposed method has been validated by comparing the reconstructed spots in the same gel acquired at different exposures, showing that the method allows us to automatically extract the pixel data within the spot boundaries, and in the presence of a plateau area, recalculate the correct distribution of grey values.

Chapter 8

Network Analysis in Systems Biology

In recent years, extensive and multi-dimensional data sets generated from recent omics technologies have presented new challenges and opportunities for unravelling the complexity of biological systems and role of cells, tissues and organs, under physiological and pathological conditions.
In particular, distinguishing the interactions of genes, proteins and metabolites that drive pathological bioprocesses is necessary for translating the full benefit of omics experimental results into the clinic. This chapter address this need by presenting a focus on the latest cutting-edge biological network analysis strategies and the way in which the results could be ported to a clinical context. Integrated network analysis that combines omics data with systems biology models enables the characterization of the interacting behaviours of complex biological systems under study.
Moreover, this Chapter presents an novel instrument that is able to mine undirected protein–protein networks and to infer sub-networks of interacting proteins intimately correlated with relevant biological pathways. This software may enable the discovery of new pathways involved in diseases. In order to describe the role of each protein within the relevant biological pathways, this software computes and scores three topological features of the identified sub-networks. By integrating the results from biological pathway clustering and topological network analysis, this software proved to be useful for the data interpretation and the generation of new hypotheses in two case studies

8.1 Network analysis in systems biology

Systems biology broadly uses networks to model and discover emerging properties among genes, proteins and other relevant biomolecules. Theoretical studies have indicated that biological networks share many features with other types of networks, as computer or social networks (Barabasi and Oltvai, 2004). Therefore, biological network analyses allow the application of mathematical and computational methods of the graph theory to biological studies (Huber et al., 2007). The computational analysis of biological networks has therefore become increasingly useful to mine the complex cellular processes and signalling pathways (Spirin and Mirny, 2003). Many types of biological networks exist, depending on the information associated with their nodes and edges. In general, biological networks can be classified as directed and undirected networks (Pieroni et al., 2007). In directed networks, the nodes are molecules and edges represent causal biological interactions, such as the transcription and translation regulations (Li wt al., 2012). In contrast, in undirected networks, an edge indicates a shared property, such as the sequence similarity (Kuchaiev and Przulj, 2011), gene co-expression (Prifti et al., 2010), protein–protein interaction (Chen et al., 2013), or the term co-occurrence in the scientific literature (Gatti et al., 2013)(Gabow et al., 2008).

In order to extract relevant biological implication from undirected networks, which are also called informative network (Lysenko et al., 2011), it is useful to complement the topological information with the independent biological information retrieved from Gene Ontology (GO) and pathway databases. Often the goal is to identify densely-interconnected areas and correlate them with a specific biological function (Bu et al., 2003)(Weatheritt et al., 2012). Several algorithms and bioinformatic tools have been proposed for partitioning the network into structural modules, or for clustering sub-network modules within an informative network(Thomas and Bonchev, 2010)(Shen et al., 2012).

Researchers in Bioinformatics developed Cytoscape plugins (Smoot et al., 2011) to mine functional modules in a varieties network types, such as Clust&See (Spinelli et al., 2013), clusterMaker (Morris et al., 2011), CyClus3D (Audenaert et al., 2011), GLay (Su et al., 2010) and Enrichment Map (Merico et al., 2010). These plugins mainly function on the basis of topological properties. The groups of highly-interconnected nodes may form clusters based on a ‘‘first topological clustering’’ strategy, which aimed at partitioning complex networks into modules. The biological functions are then assigned assuming that members within each sub-network shared a similar biological function (Dong and Horvath, 2007). However, clusters are identified solely on the basis of the topology. Therefore, the possibility of co-occurrence events cannot be ruled out using these methods. Moreover, these strategies are heavily influenced by the topological structure of the network itself, and the way that the network is constructed (Aittokallio and Schwikowski, 2006).

In practice, the connectivity of the informative networks is established by

experimental methods, which can lead to sampling of a subnet of the real biological network(Heatha and Kavrakia, 2009). Often, biases in the sampling strategies lead to apparent scale-free topologies, which do not reflect the actual complete network topology. As an alternate to the ‘‘first topological clustering’’ methods, some authors used a ‘‘first pathway enrichment’’ strategy (Griswold et al., 2012), which enables analyzing gene networks and extracting functional modules starting from the biological enrichment analysis. Several Cytoscape plugins thus have been implemented owning to this strategy, such as BiNGO (Maere et al, 2005), ClueGO (Bindea et al., 2009), ClusterViz (Cai et al., 2010), JEPETTO (Glaab et al., 2012) and Reactome (Jupe et al., 2012).

BiNGO and ClueGO are widely used tools to determine which biological functions are statistically over-represented in a list of genes or a network. These plugins offer the possibility to calculate enrichment by using different statistical algorithms. However, they do not evaluate the connectivity between genes, and do not take into account the possibility that nodes spread in the network could still represent a significant biological function (Mitra et al., 2013).

A recent focus of bioinformatics has been to develop the computational tools that are able to mine the connectivity of gene networks and uncover the sets of molecules that participate in a common biological function. Reactome is established based on an un-weighted human protein functional interaction network and the functional interaction score was calculated with Pearson correlation coefficients among all gene pairs in the biological database (Wu and Stein, 2012). The weighted network was clustered into a series of gene interaction modules using the Markov clustering algorithm. Each module of the Reactome consists of a set of genes that are both connected in the protein functional interaction network and highly-correlated in biological databases. This approach, however, does not consider the topology or the connectivity of gene interaction modules. JEPETTO identifies functional associations between genes and pathways using protein interaction networks and topological analyses. Although JEPPETTO combines network analysis with functional enrichment, this tool requires a list of genes in input and the selection of a database of interest from which the reference gene sets will be extracted.

Reactome and JEPETTO are based on an internal gene interaction database, in which the users are not able to filter this pre-defined database or to analyze their own protei informative networks. Nowadays there are a great number of tools able to produce undirected protein networks from a user-defined query, such as protein–protein interaction network (STRING, BioGRID, etc.), co-expression network (COXPRESdb)(Obayashi et al., 2013) and co-occurrence network (ProteinQuest) (Benso et al., 2013).

8.2 Mining undirected protein–protein networks

This section presents a novel software able to identify sub-networks of genes that are highly connected and belong to the same biological pathways. Moreover, the aforementioned plugins do not allow the analysis of user-defined undirected networks, such as protein–protein interaction, functional association, gene co-expression and literature co-occurrence.

This section presents a new method using a ‘‘first biological assignment’’ strategy. This method is implemented as a Cytoscape plugin, called FunMod (Natale et al., 2014). According to the principle that interacting proteins drive common biological processes, FunMod analyzes informative networks combing topological and biological information to identify and extract sub-network modules in proteins that are involved in the same biological pathway. Moreover, in order to describe the shape of the modules and discriminate the proteins’ topological proprieties within the single sub-network, FunMod analyzes subnetwork features by using three topological scores. The subnetworks that are statistically overrepresented can act as building blocks of complex informative networks and carry out a specific biological function. Assessment of the sub-network topological proprieties and shapes can consequently be used for the gene ranking in the context of a specific research domain, such as a disease.

FunMod proves to be a useful method for identifying functional sub-networks in an informative protein network, exploring biomedical information and inferring automated functional hypotheses between an user defined protein–protein network. FunMod is unique at its capability of analyzing user-defined undirected networks, in order to provide more realistic models that incorporate information from certain cellular types, developmental states and/or disease conditions.

FunMod analyses the protein informative network displayed in the Cytoscape Main Network View window. The plugin supports many standard protein annotation formats and the protein nodes can be identified (node ID) by six different dictionaries: Entrez Gene ID, Ensembl, Official Symbol (HGNC symbol), NCBI Reference Sequence, UniGene, Uniprot Entry Name.

FunMod iterativaly selects all edges of the network and assigns a functional annotation to an edge when the two linked nodes are annotated in the same biological group or pathway in the ConsensusPathDB (DB) database (Liekens et al., 2011). In other words, FunMod considers a network $G = (V, E)$ with n vertices V joined with edges E, and collects, for each ConsensusPathDB pathway, pairs of linked nodes modelling a functional sub-network $G_p = (V_p, E_p)$, where $V_p \subseteq$ pathway and $E_p \in E$.

FunMod performs a global enrichment analysis screening the pathways whose proteins are co-annotated with their neighbors. Accordingly, all the pathways identified by FunMod are also significant in a global enrichment analysis, because the connected nodes are a fraction of nodes in the network. So all pathways enriched in a sub-network are enriched also in the global network, but only few pathways enriched

in the global network are also enriched in a cluster. The aim of our algorithm is to find the pathways enriched in the global network and whose proteins are densely connected in a sub-network.

Afterwards FunMod extracts all pairs of nodes annotated for the same pathway in a new sub-network, it tests the statistical significance of the sub-network, and calculates the topological properties of the sub-network. In this way FunMod is able to identify sub-networks that are statistically enriched in biological functions and that show interesting topological features.

The statistical significance of the sub-network is determined by calculating the p-value performing a hypergeometric test, a well-established method used in gene enrichment analysis (Griswold et al., 2012). The hypergeometric probability is based on the following formula:

$$h(x;X,n,N) = [XCx][N\text{-}XCn\text{-}\ x]/[NCn]$$

where x is the number of nodes of the sub-network (the items in the sample that are classified as successes), n is the number of genes in the network (items in the sample); X is the number of genes annotated in the DB with that pathway (items in the population that are classified as successes); and N the number of all genes annotated in DB (items in the population). FunMod preserves the sub-networks with a p-value < 0.05.

For a better understanding of the systemic functions and the cooperative interactions between genes within the functional modules, FunMod checks whether the sub-network topology fits into a specific module. Network modules represent patterns in complex networks occurring significantly more often than in randomized networks. They consist of sub-graphs of local interconnections between network elements. FunMod calculates a fitting score of each sub-network for three modules: clique, star, and path; the most common motifs that are found in various networks (Dolinski et al., 2013).

A clique is a sub-network in which all genes are connected to each other. Cliques are the most widely used modules for assigning a biological function to a topological sub-network. FunMod calculates the tendency to be a clique by Graph Density (GD), a score that can also be defined as the local clustering coefficient, using the formula:

$$\mathrm{GD} = \frac{2\mathrm{E}}{\mathrm{n} \times (\mathrm{n} - 1)}$$

where E is the number of edges in the sub-network and n the number of genes in the sub-network.

The star module, particularly interesting for identifying drug targets, is characterized by a central gene with a high degree (the hub) connected to a set of first-

degree neighbours loosely connected among each other. In a star sub-network the hub gene has influence on its neighbourhood genes and possibly on the whole network. In order to identify a star module, FunMod calculates the sub-network centralization (CE) using the formula:

$$CE = \frac{n}{(n - 2)} \times \left[\frac{\max(k)}{n - 1} - GD\right] \quad (3)$$

where max(k) is the highest degree in the sub-network.

The path module corresponds to a real pathway where the genes contribute to a signal transduction. The path score is calculated as the sub-network diameter (D), the maximum length of all shortest paths between any two connected nodes, using the formula:

$$D = \max_{i,j} \delta_{min}(i, j) \quad (4)$$

where δ_{min} is the minimum path between two nodes i and j of the network.

FunMod displays into the Cytoscape Results Panel the identified pathways, ranking them, according to their p-value. For each pathway FunMod displays its clique, star and path coefficient.

By clicking the pathway button FunMod selects the corresponding nodes in the network, and by clicking the "View subnet" button, it creates a new network containing only those genes and edges annotated with that pathway. Finally FunMod enables to save the results into a tab-delimited file. FunMod plugin, user guide, screenshot and demo networks can be freely downloaded from the SourceForge project page at: http://sourceforge.net/projects/funmodnetwork/. FunMod is platform independent, developed in Java as Cytoscape 2.8.4 plugin. FunMod is freely available for non-commercial purposes.

8.3 Subnetwork functional modules

Gene Ontology (GO) provides information on the location of genes and gene products in a cell or the extracellular environment and also on the molecular function they carry out. However, GO does not provide information about the interaction of proteins in the same biological context. For example, GO does not allow us to describe genes in terms of which cells or tissues they are expressed in, which developmental stages they are expressed at, or their involvement in disease (http:// www.geneontology.org/GO.doc.shtml). We thus chose ConsensusPathDB to identify proteins that are strictly involved in the same pathways. We also assess the

topological shape of each sub-network in order to reveal evidence of its biological function and the function of its components. Three topological scores are calculated to describe the global features of the sub-network: graph density, network centralization and shortest path. Other topological scores, such as centrality and degree, describe the relative importance of a single node within the sub-network and thus left out in FunMod.

To demonstrate the usage and performance of FunMod, we showed here case studies to analyze two different informative networks: a bibliometric network of proteins related to the Budd–Chiari syndrome and a co-expression network of proteins related to spinal muscular atrophy. Identification of the sub-network functional modules in the undirected protein networks allows the reduction in network complexity, clustering of proteins on the basis of common biological functions and discovery of the mechanisms underlying a disease.

Using a ‘‘first topological assignment’’ strategy to identify subnetwork functional modules, such as stars and cliques, can be tricky because informative networks are known to have a huge number of edges that are not always pertinent to biological functions. In this work we presented FunMod, a new Cytoscape 2.8 plugin, which can analyze undirected protein networks, such as co-occurrence and co-expression networks, and guide the discovery of sub-network functional modules.

A functional module can be considered as a distinct group of interacting proteins within a pathway relevant to a condition of interest. FunMod identifies within an informative network, pairs of nodes belonging to the same biological pathways and assesses their statistical significance. It then analyzes the topology of the identified sub-network to infer the topological relations (motifs) of its nodes. In this work, the network topology is influenced by the biomedical knowledge since the link between two proteins was established when two gene symbols appear in the same MEDLINE record. The study of the connection between biomedical concepts by using co-occurrence network extracted from MEDLINE proved capable of guiding the discovery of novel knowledge from scientific literature. This sub-network profiling combined with information from the biological database will help us to better understand the biological significance of the protein– protein network.

FunMod was tested using the co-occurrence network of proteins cited in Budd–Chiari syndrome papers, identifying 33 different biological pathways that are significantly enriched; and using the co-expression network of proteins discussed in publications on SMA. FunMod proves to be a useful tool for a better understanding of the cooperative interactions between proteins and discriminating the biological role played by each protein within a functional module.

Bibliography

Aardema MJ1 and MacGregor JT. Toxicology and genetic toxicology in the new era of "toxicogenomics": impact of "-omics" technologies. Mutat Res. 2002. 499(1):13-25

Ahmed A, Arnold A, Coelho LP, Kangas J, Sheikh AS, Xing E, Cohen W, and Murphy RF. Structured Literature Image Finder: Parsing Text and Figures in Biomedical Literature. Web Semant. 2010. 8(2-3):151-154

Aittokallio T, Schwikowski B. Graph-based methods for analysing networks in cell biology. Brief Bioinform 2006;7:243-55

Alvarenga CA. How science goes wrong, The Economist, 2013

Appel RD, Hoogland C, Bairoch A, Hochstrasser DF. Constructing a 2-D database for the World Wide Web. Methods Mol Biol. 1999;112:411–6

Aranguren ME, Bechhofer S, Lord P, Sattler U, and Stevens R. Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. BMC Bioinformatics. 2007. 8:57

Aranguren ME, Bechhofer S, Lord P, Sattler U, and Stevens R. Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. BMC Bioinformatics. 2007. 8:57

Audenaert P, Van Parys T, Brondel F, Pickavet M,Demeester P, Van de Peer Y, et al.CyClus3D: aCytoscape plugin for clustering network motifs in integrated networks. Bioinformatics 2011;27:1587–8

Bakshi K. Considerations for big data: Architecture and approach. In Aerospace Conference, 2012 IEEE. 2012. 1-7. IEEE

Bandow JE, Baker JD, Berth M, Painter C. Improved image analysis workflow for 2-D gels enables large-scale 2-D gel-based proteomics studies-COPD biomarker discovery study. Proteomics. 2008;8:3030–41

Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5:101–13

Batelli S, Albani D, Rametta R, Polito L, Prato F, Pesaresi M, Negro A, and Forloni G. DJ-1 modulates alpha-synuclein aggregation state in a cellular model of oxidative stress: relevance for Parkinson's disease and involvement of HSP70. PLoS One. 2008 Apr 2;3(4):e1884

Begley CG, and Ioannidis J P. Reproducibility in Science. 2015

Ben-Israel A. A Newton-Raphson method for the solution of system of equations. Journal of Mathematical analysis and applications 1996;15:243-252

Benso A, Cornale P, Di Carlo S, Politano G, Savino A. Reducing the complexity of complex gene coexpression networks by coupling multiweighted labeling with topological analysis. Biomed Res Int 2013;2013:676328

Berth M, Moser FM, Kolbe M, and Bernhardt J. The state of the art in the analysis of two-dimensional gel electrophoresis images. Appl Microbiol Biotechnol. 2007; 76:1223–43

Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 2009;25:1091-3

Bohlouli M, Schulz F, Angelis L, Pahor D, Brandic I, Atlan D, and Tate R. Towards an integrated platform for big data analysis. In Integration of practice-oriented knowledge technology: Trends and prospectives. 2013. 47-56. Springer Berlin Heidelberg

Bonifati V. Autosomal recessive parkinsonism. Parkinsonism Relat Disord. 2012.18 Suppl 1:S4-6

Bonino D. Corno F. Farinetti L. and Bosca A. Ontology driven semantic search. WSEAS Transaction on Information Science and Application. 2004. 1(6), 1597-1605

Boyd D, and Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society. 2012. 15(5): 662-679

Bromer JG, Ata B, Seli M, Lockwood CJ, and Seli E. Preterm deliveries that result from multiple pregnancies associated with assisted reproductive technologies in the USA: a cost analysis. Current Opinion in Obstetrics and Gynecology, 2011. 23(3), 168-173

Bu D, Zhao Y, Cai L, Xue H, Zhu X, Lu H, et al. Topological structure analysis of the protein–protein interaction network in budding yeast. Nucleic Acids Res 2003;31:2443–50

Cai J, Chen G, Wang J. ClusterViz: a Cytoscape plugin for graph clustering and visualization. School of Information Science and Engineering 2010;1

Cai Q, Arumugam RV, Xu Q, and He B. Understanding the Behavior of Solid State Disk. In Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems. 2015. 1:341-355. Springer International Publishing

Chang J, Van Remmen H, Ward WF, Regnier FE, Richardson A, and Cornell J. Processing of data generated by 2-dimensional gel electrophoresis for statistical analysis: missing data, normalization, and statistics. J Proteome Res. 2004;3:1210–8

Chen CP, and Zhang CY. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences. 2014. 275, 314-347

Chen L, Xuan J, Riggins RB, Wang Y, Clarke R. Identifying protein interaction sub-networks by a bagging Markov random field-based method. Nucleic Acids Res 2013;41:e42

Choi J, Sullards MC, Olzmann JA, Rees HD, Weintraub ST, Bostwick DE, Gearing M, Levey AI, Chin LS, and Li L. Oxidative damage of DJ-1 is linked to sporadic Parkinson and Alzheimer diseases. J Biol Chem. 2006 Apr 21;281(16):10816-24

Clark BN, and Gutstein HB. The myth of automated, high-throughput two-dimensional gel analysis. Proteomics 2008;8:1197–203

Darema F. Dynamic data driven applications systems: A new paradigm for application simulations and measurements. In Computational Science-ICCS 2004. 2004. 662-669. Springer Berlin Heidelberg

Daszykowski M, Bierczynska-Krzysik A, Silberring J, Walczak B. Avoiding spots detection in analysis of electrophoretic gel images. Chemometr Intell Lab Syst 2010;104:2–7

de Vargas Roditi L, Claassen M. Computational and experimental single cell biology techniques for the definition of cell type heterogeneity, interplay and intracellular dynamics. Curr Opin Biotechnol. 2014. Nov 9;34C:9-15

Dolinski K, Chatr-Aryamontri A, Tyers M. Systematic curation of protein and genetic interaction data for computable biology. BMC Biol 2013;11:43

Dong J, Horvath S. Understanding network concepts in modules. BMC Syst Biol 2007;4:1-24

dos Anjos A, Møller AL, Ersbøll BK, Finnie C, Shahbazkia HR. New approach for segmentation and quantification of two-dimensional gel electrophoresis images. Bioinformatics. February 1, 2011;27(3):368–75

Dowsey AW, Morris JS, Gutstein HB, and Yang GZ. Informatics and statistics for analyzing 2-d gel electrophoresis images. Methods Mol Biol. 2010;604:239–55

Dräger A, and Palsson BØ. Improving collaboration by standardization efforts in systems biology. Front Bioeng Biotechnol. 2014. 2:61

Drews O, and Görg A. DynaProt 2D: an advanced proteomic database for dynamic online access to proteomes and two-dimensional electrophoresis gels. Nucleic Acids Res. 2005. 33(Database issue):D583-7

Dudoit S, Gentleman RC, and Quackenbush J. Open source software for the analysis of microarray data. Biotechniques. 2003. 34(S45-S51), 13

Duncan E, Brown M, Shore EM. The revolution in human monogenic disease mapping. Genes (Basel). 2014. 5(3):792-803

Fasano M, Alberio T, Colapinto M, Mila S, and Lopiano L. Proteomics as a tool to investigate cell models for dopamine toxicity. Parkinsonism Relat Disord. 2008;14 Suppl 2:S135-8

Fielder JH. the Vioxx debacle. Engineering in Medicine and Biology Magazine, IEEE, 2005. 24(2), 106-109

Gabow AP, Leach SM, Baumgartner WA, Hunter LE, Goldberg DS. Improving protein function prediction methods with integrated literature data. BMC Bioinformatics 2008;9:198

Garrison LP, and Austin MF. Linking pharmacogenetics-based diagnostics and drugs for personalized medicine. Health Affairs. 2006. 25(5), 1281-1290

Garwood K, McLaughlin T, Garwood C, Joens S, Morrison N, Taylor CF, Carroll K, Evans C, Whetton AD, Hart S, Stead D, Yin Z, Brown AJ, Hesketh A, Chater K, Hansson L, Mewissen M, Ghazal P, Howard J, Lilley KS, Gaskell SJ, Brass A, Hubbard SJ, Oliver SG, andPaton NW. PEDRo: a database for storing, searching and disseminating experimental proteomics data. BMC Genomics. 2004. 5:68

Gatti S, Leo C, Gallo S, Sala V, Bucci E, Natale M, Cantarella D, Medico E, and Crepaldi T. Gene expression profiling of HGF/Met activation in neonatal mouse heart. Transgenic Res. 2013. 22(3):579-93

Ghemawat S, Gobioff H, Leung ST. The Google file system. In ACM SIGOPS operating systems review. 2003. 37(5):29-43

Giordano M, Natale M, Cornaz M, Ruffino A, Bonino D, and Bucci EM. iMole, a web based image retrieval system from biomedical literature. Electrophoresis. 2013. 34(13):1965-8

Girish V, and Vijayalakshmi A. Affordable image analysis using NIH Image/ImageJ. Indian J Cancer. 2004 Jan-Mar;41(1):47

Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A, EnrichNet: network-based gene set enrichment analysis. Bioinformatics 2012;28:i451-57

Gómez-Pérez J M, and Mazurek C. ROHub—A Digital Library of Research Objects Supporting Scientists Towards Reproducible Science. Semantic Web Evaluation Challenge: SemWebEval 2014 at ESWC 2014, Anissaras, Crete, Greece, 2014, 475, 77

González-Beltrán A, Maguire E, Sansone SA, and Rocca-Serra P. linkedISA: semantic representation of ISA-Tab experimental metadata. BMC Bioinformatics. 2014.15 Suppl 14:S4. doi: 10.1186/1471-2105-15-S14-S4

Gorg A, Weiss W, and Dunn MJ. Current two-dimensional electrophoresis technology for proteomics. Proteomics 2004;4:3665–85

Graur D, Zheng Y, and Azevedo RB. An evolutionary classification of genomic function. Genome Biol Evol. 2015

Griswold AJ, Ma D, Cukier HN, Nations LD, Schmidt MA, Chung RH, Evaluation of copy number variations reveals novel candidate genes in autism spectrum disorder-associated pathways. Hum Mol Genet 2012;21:3513-23

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, and Witten IH. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter. 2009. 11(1), 10-18

Haraksingh RR, and Snyder MP. Impacts of variation in the human genome on gene regulation. J Mol Biol. 2013. 425(21):3970-7

Hearst MA, Divoli A, Guturu H, Ksikes A, Nakov P, Wooldridge MA, andYe J. BioText Search Engine: beyond abstract search. Bioinformatics. 2007. 23(16):2196-7

Heatha AP, Kavrakia LE. Computational challenges in systems biology. Computer Science Review 2009;3:117

Hiroaki Kitano, Systems Biology: A Brief Overview. Science. 2002. 295(5560):1662-1664

Hod Y, Pentyala SN, Whyard TC, and El-Maghrabi MR. Identification and characterization of a novel protein that regulates RNA-protein interaction. J Cell Biochem. 1999 Mar 1;72(3):435-44

Hoehndorf R, Haendel M, Stevens R, and Rebholz-Schuhmann D. Thematic series on biomedical ontologies in JBMS: challenges and new directions. J Biomed Semantics. 2014. 5:15

Hortin GL, and Sviridov D. The dynamic range problem in the analysis of the plasma proteome. J Proteomics 2010;73:629–36

Howe D, Costanzo M, Fey P, Gojobori T, Hannick, L, Hide W, and Rhee SY. Big data: The future of biocuration. Nature. 2008. 455(7209), 47-50

Huang da W, Sherman BT, Lempicki RA Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009. 37(1):1-13

Huber W, Carey VJ, Long L, Falcon S, Gentleman R. Graphs in molecular biology. BMC Bioinformatics 2007;8:S8

Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, and O'Donovan C. The GOA database: Gene Ontology annotation updates for 2015. Nucleic Acids Res. 2015. 43(Database issue):D1057-63

Idris SF, Ahmad SS, Scott MA, Vassiliou GS, Hadfield J. The role of high-throughput technologies in clinical cancer genomics. Expert Rev Mol Diagn. 2013. 13(2):167-8

Ito G, Ariga H, Nakagawa Y, and Iwatsubo T. Roles of distinct cysteine residues in S-nitrosylation and dimerization of DJ-1. Biochem Biophys Res Commun. 2006 Jan 13;339(2):667-72

Jee K, and Kim GH. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Healthcare informatics research. 2013. 19(2), 79-85

Jog CR. Healthcare technology, patient engagement and adherence: systems and business opportunity analysis. 2012. Doctoral dissertation. Massachusetts Institute of Technology

Jupe S, Akkerman JW, Soranzo N, Ouwehand WH, Reactome - a curated knowledgebase of biological pathways: megakaryocytes and platelets. J Thromb Haemost 2012;doi:10.1111/j

Kim D, and Yu H. Figure text extraction in biomedical literature. PLoS One. 2011. 6(1):e15338

Kim MA, and Yoon YW. Spot detection of complex regions by accumulated gradient in two dimensional electrophoresis images. Proceedings of the 11th International Conference on Advanced Communication Technology (ICACT'09) 2009;3:1982-5. IEEE Press, Piscataway, NJ, USA

Kirschner DE, Hunt CA, Marino S, Fallahi-Sichani M, and Linderman JJ. Tuneable resolution as a systems biology approach for multi-scale, multi-compartment computational models. Wiley Interdiscip Rev Syst Biol Med. 2014. 6(4):289-309

Kuchaiev O, Przulj N. Integrative network alignment reveals large regions of global network similarity in yeast and human. Bioinformatics 2011;27:1390–6

Lemkin PF, and Thornwall G. Flicker image comparison of 2-D gel images for putative protein identification using the 2DWG meta-database. Mol Biotechnol. Sep 1999;12(2):159–72

Lemkin PF, Myrick JM, Lakshmanan Y, et al. Exploratory data analysis groupware for qualitative and quantitative electrophoretic gel analysis over the Internet-WebGel. Electrophoresis. Dec 1999;20(18):3492–507

Lemkin PF. Comparing two-dimensional electrophoretic gel images across the Internet. Electrophoresis. Mar-Apr 1997;18(3–4):461–70

Lev N, Ickowicz D, Melamed E, and Offen D. Oxidative insults induce DJ-1 upregulation and redistribution: implications for neuroprotection. Neurotoxicology. 2008 May;29(3):397-405

Li J, Hua X, Haubrock M, Wang J, Wingender E. The architecture of the gene regulatory networks of different tissues. Bioinformatics 2012;28:i509–14

Liekens AM, De Knijf J, Daelemans W, Goethals B, De Rijk P, Del-Favero J. BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. Genome Biol 2011;12:R57

Lysenko A, Defoin-Platel M, Hassani-Pak K, Taubert J, Hodgman C, Rawlings CJ, et al. Assessing the functional coherence of modules found in multiple-evidence networks from Arabidopsis. BMC Bioinformatics 2011;12:203

Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 2005;21:3448-9

Marengo E, Robotti E, Antonucci F, Cecconi D, et al, Numerical approaches for quantitative analysis of two-dimensional maps: a review of commercial software and home-made systems. Proteomics. 2005;5:654–66

Maresca B, Cigliano L, Corsaro MM, Pieretti G, Natale M, Bucci EM, Dal Piaz F, Balato N, Nino M, Ayala F, and Abrescia P. Quantitative determination of haptoglobin glycoform variants in psoriasis. Biol Chem. 2010;391(12):1429-39

Matuzevicius D, Serackis A, Navakauskas D. Mathematical models of oversaturated protein spots. Electron Electr Eng 2007;1:63–8

Maurer MH. Software analysis of two-dimensional electrophoretic gels in proteomic experiments. Curr Bioinform. 2006;1:255-62

Mayer G, Jones AR, Binz PA, Deutsch EW, Orchard S, Montecchi-Palazzi L, Vizcaíno JA, Hermjakob H, Oveillero D, Julian R, Stephan C, Meyer HE, and Eisenacher M. Controlled vocabularies and ontologies in proteomics: overview, principles and practice. Biochim Biophys Acta. 2014. 1844(1 Pt A):98-107

Mayer G, Jones AR, Binz PA, Deutsch EW, Orchard S, Montecchi-Palazzi L, Vizcaíno JA, Hermjakob H, Oveillero D, Julian R, Stephan C, Meyer HE, and Eisenacher M. Controlled vocabularies and ontologies in proteomics: overview, principles and practice. Biochim Biophys Acta. 2014. 1844(1 Pt A):98-107

Mayer-Schönberger V, and Cukier K. Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, 2013

McKusick VA, and Ruddle FH.. Toward a complete map of the human genome. Genomics. 1987. 1, 103–106

Merelli I, Pérez-Sánchez H, Gesing S, D'Agostino D. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. Biomed Res Int. 2014. 2014:134023

Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment Map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One 2010;5:e13984

Meulener MC, Graves CL, Sampathu DM, Armstrong-Gold CE, Bonini NM, and Giasson BI. DJ-1 is present in a large molecular complex in human brain tissue and interacts with alpha-synuclein. J Neurochem. 2005 Jun;93(6):1524-32

Miller I, Crawford J, and Gianazza E. Protein stains for proteomic applications: which, when, why? Proteomics 2006;6:5385–408

Millioni R, Sbrignadello S, Tura A, Iori E, Murphy E, and Tessari P. The inter- and intra-operator variability in manual spot segmentation and its effect on spot quantitation in two-dimensional electrophoresis analysis. Electrophoresis. 2010; 10:1739–42

Mitra K, Carvunis AR, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet 2013;14:719-32

Mitsumoto A, Nakagawa Y, Takeuchi A, Okawa K, Iwamatsu A, and Takanezawa Y. Oxidized forms of peroxiredoxins and DJ-1 on two-dimensional gels increased in response to sublethal levels of paraquat. Free Radic Res. 2001. 35(3):301-10

Moran JK, Weierstall R, and Elbert T. Differences in brain circuitry for appetitive and reactive aggression as revealed by realistic auditory scripts. Frontiers in behavioral neuroscience, 2014. 8

Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, et al. ClusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics 2011;12:436

Morris JS, Clark BN, Wei W, Gutstein HB. Evaluating the performance of new approaches to spot quantification and differential expression in 2-dimensional gel electrophoresis studies. J Proteome Res. 2010;9:595–604

Murdoch TB, and Detsky AS. The inevitable application of big data to health care. Jama. 2013. 309(13), 1351-1352

Nagakubo D, Taira T, Kitaura H, Ikeda M, Tamai K, Iguchi-Ariga SM, and Ariga H. DJ-1, a novel oncogene which transforms mouse NIH3T3 cells in cooperation with ras. vBiochem Biophys Res Commun. 1997. 231(2):509-13

Naidoo N, Pawitan Y, Soong R, Cooper DN and Ku CS. Human genetics and genomics a decade after the release of the draft sequence of the human genome. Hum Genomics. 2011, 5(6):577-622

Natale M, Benso A, Di Carlo S, Ficarra E. FunMod: a Cytoscape plugin for identifying functional modules in undirected protein-protein networks. Genomics Proteomics Bioinformatics. 2014 Aug;12(4):178-86

Natale M, Bonino D, Consoli P, Alberio T, Ravid RG, Fasano M, and Bucci EM. A meta-analysis of two-dimensional electrophoresis pattern of the Parkinson's disease-related protein DJ-1. Bioinformatics. 2010;26(7):946-52

Natale M, Caiazzo A, Bucci EM, and Ficarra E. A novel Gaussian extrapolation approach for 2D gel electrophoresis saturated protein spots. Genomics Proteomics Bioinformatics. 2012;10(6):336-44

Natale M, Maresca B, Abrescia P, and Bucci EM. Image analysis workflow for 2-D electrophoresis gels based on ImageJ. Proteomics Insights, 2011.4 37-49

Niemenmaa M, Kallio A, Schumacher A, Klemelä P, Korpelainen E, and Heljanko K. Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. Bioinformatics. 2012 Mar 15;28(6):876-7

Nishihara JC, and Champion KM. Quantitative evaluation of proteins in one- and two-dimensional polyacrylamide gels using a fluorescent stain. Electrophoresis. July 23, 2002;(14):2203–15

O'Farrel PH. High resolution two-dimensional electrophoresis of proteins. J Biol Chem 1975;250:4007–21

Obayashi T, Okamura Y, Ito S, Tadaka S, Motoike IN, Kinoshita K. COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. Nucleic Acids Res 2013;41:D1014–20

O'Brien EJ, Palsson BO. Computing the functional proteome: recent progress and future prospects for genome-scale models. Curr Opin Biotechnol. 2015. Jan 7;34C:125-134

Olzmann JA, Li L, Chudaev MV, Chen J, Perez FA, Palmiter RD, and Chin LS. Parkin-mediated K63-linked polyubiquitination targets misfolded DJ-1 to aggresomes via binding to HDAC6. J Cell Biol. 2007 Sep 10;178(6):1025-38

Osier MV, Zhao H, and Cheung KH. Handling multiple testing while interpreting microarrays with the Gene Ontology Database. BMC bioinformatics. 2004. 5(1), 124

Ozdemir V, Suarez-Kurtz G, Stenne R, Somogyi AA, Someya T, Kayaalp SO, and Kolker E. Risk assessment and communication tools for genotype associations with multifactorial phenotypes: the concept of "edge effect" and cultivating an ethical bridge between omics innovations and society. OMICS. 2009. Feb;13(1):43-61

Pérès S, Molina L, Salvetat N, Granier C, Molina F. A new method for 2D gel spot alignment: application to the analysis of large sample sets in clinical proteomics. BMC Bioinformatics. October 28, 2008;9:460

Pieroni E, de la Fuente van Bentem S, Mancosu G, Capobianco E, Hirt H, de la Fuente A. Protein networking: insights into global functional organization of proteomes. Proteomics 2008;8:799–816

Pleissner KP, Hoffmann F, Kriegel K, Wenk C, Wegner S, Sahlström A, Oswald H, Alt H, Fleck E. New algorithmic approaches to protein spot detection and pattern

matching in two-dimensional electrophoresis gel databases. Electrophoresis. 1999; 20(4-5):755-65

Prifti E, Zucker JD, Cle´ment K, Henegar C. Interactional and functional centrality in transcriptional coexpression networks. Bioinformatics 2010;26:3083–9

Raghupathi W, and Raghupathi V. Big data analytics in healthcare: promise and potential. Health Information Science and Systems. 2014. 2(1), 3

Raicu I, Foster IT, and Beckman P. Making a case for distributed file systems at exascale. In Proceedings of the third international workshop on Large-scale system and application performance. 2011. 11-18. ACM

Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V, Kinoshita J, Luciano J, Marshall MS, Ogbuji C, Rees J, Stephens S, Wong GT, Wu E, Zaccagnini D, Hongsermeier T, Neumann E, Herman I, and Cheung KH. Advancing translational research with the Semantic Web. BMC Bioinformatics. 2007. 8 Suppl 3:S2

Sarachan BD, Simmons MK, Subramanian P, and Temkin JM. Combining medical informatics and bioinformatics toward tools for personalized medicine. Methods Inf Med. 2003. 42(2), 111-115

Schilling PL, and Bozic KJ. The Big To Do About "Big Data". Clinical Orthopaedics and Related Research, 2014. 472(11), 3270-3272

Schulz M, Uhlendorf J, Klipp E, and Liebermeister W. SBMLmerge, a system for combining biochemical network models. Genome Inform. 2006.17(1):62-71

Shen R, Goonesekere NC, Guda C. Mining functional subgraphs from cancer protein–protein interaction networks. BMC Syst Biol 2012;6:S2

Shinbo Y, Niki T, Taira T, Ooe H, Takahashi-Niki K, Maita C, Seino C, Iguchi-Ariga SM, Ariga H. Proper SUMO-1 conjugation is essential to DJ-1 to exert its full activities. Cell Death Differ. 2006 Jan;13(1):96-108

Shvachko K, Kuang H, Radia S, and Chansler R. The hadoop distributed file system. In Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on. 2010. 1-10. IEEE

Smith AK, Cheung KH, Yip KY, Schultz M, and Gerstein MK. LinkHub: a Semantic Web system that facilitates cross-database queries and information retrieval in proteomics. BMC Bioinformatics. 2007. 8 Suppl 3:S5

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 2011;27:431–2

Sorzano CO, Arganda-Carreras I, Thévenaz P. Elastic image registration of 2-D gels for differential and repeatability studies. Proteomics. 2008; 8:62–5

Spinelli L, Gambette P, Chapple CE, Robisson B, Baudot A, Garreta H, et al. Clust&See: a Cytoscape plugin for the identification, visualization and manipulation of network clusters. Biosystems 2013;113:91–5

Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci U S A 2003;100:12123–8

Sternberg S. Biomedical image processing. IEEE Computer. 1983;16:22–34

Su G, Kuchinsky A, Morris JH, States DJ, Meng F. GLay: community structure analysis of biological networks. Bioinformatics 2010;26:3135–7

Sung J, Wang Y, Chandrasekaran S, Witten DM, Price ND. Molecular signatures from omics data: from chaos to consensus. Biotechnol J. 2012. 7(8):946-57

Swan M. Health 2050: the realization of personalized medicine through crowdsourcing, the Quantified Self, and the participatory biocitizen. Journal of personalized medicine 2.3. 2012, 93-118

Tao C, Song D, Sharma D, and Chute CG. Semantator: semantic annotator for converting biomedical text to linked data. J Biomed Inform. 2013. 46(5):882-93

Tao X, and Tong L. Crystal structure of human DJ-1, a protein associated with early onset Parkinson's disease. J Biol Chem. 2003 Aug 15;278(33):31372-9

Tene O, and Polonetsky J. "Big data for all: Privacy and user control in the age of analytics." Nw. J. Tech. & Intell. Prop. 2012. 11: xxvii

Thomas S, Bonchev D. A survey of current software for network analysis in molecular biology. Hum Genomics 2010;4:353–60

Tilahun B, Kauppinen T, Keßler C, and Fritz F. Design and development of a linked open data-based health information representation and visualization system: potentials and preliminary evaluation. JMIR Med Inform. 2014. 2(2):e31

Tsukahara M, Mitrović S, Gajdosik V, Margaritondo G, Pournin L, Ramaioli M, Sage D, Hwu Y, Unser M, and Liebling TM. Coupled Tomography and Distinct-Element-Method Approach to Exploring the Granular Media Microstructure in a Jamming Hourglass. Physical Review E. 2008. 77(6)

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. From FastQ data to high confidence variant calls: the

Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013 Oct 15;11(1110):11.10.1-11.10.33

van der Brug MP, Blackinton J, Chandran J, Hao LY, Lal A, Mazan-Mamczarz K, Martindale J, Xie C, Ahmad R, Thomas KJ, Beilina A, Gibbs JR, Ding J, Myers AJ, Zhan M, Cai H, Bonini NM, Gorospe M, and Cookson MR. RNA binding activity of the recessive parkinsonism protein DJ-1 supports involvement in multiple cellular pathways. Proc Natl Acad Sci USA. 2008 Jul 22;105(29):10244-9

Van Laar VS, Mishizen AJ, Cascio M, and Hastings TG. Proteomic identification of dopamine-conjugated proteins from isolated rat brain mitochondria and SH-SY5Y cells. Neurobiol Dis. 2009. 34(3):487-500

Villars RL, Olofson C.W, and Eastwood, M Big data: What it is and why you should care. White Paper, IDC. 2011

Vural Özdemir. OMICS 2.0: A Practice Turn for 21st Century Science and Society. OMICS: A Journal of Integrative Biology. January 2013, 17(1): 1-4

Ward JH. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc., 1963. 58, 236–244

Weatheritt RJ, Jehl P, Dinkel H, Gibson TJ. IELM-a web server to explore short linear motif-mediated interactions. Nucleic Acids Res 2012;40:W364–9

Webster J. MapReduce: Simplified Data Processing on Large Clusters. *Search Storage*. 2004

Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G, Goble C, and Mons B. Open PHACTS: semantic interoperability for drug discovery. Drug Discov Today. 2012. 17(21-22):1188-98

Winkler H. (1920). Verbreitung und Ursache der Parthenogenesis im Pflanzen-und Tierreiche. Verlag von Gustav Fischer; Jena

Wu G, Stein L. A network module-based method for identifying cancer prognostic signatures. Genome Biol 2012;13:R112

Xin R, Rosen J, Zaharia M, Franklin M, Shenker S, and Stoica I. "Shark: SQL and Rich Analytics at Scale". SIGMOD 2013

Xiong H, Wang D, Chen L, Choo YS, Ma H, Tang C, Xia K, Jiang W, Ronai Z, Zhuang X, and Zhang Z. Parkin, PINK1, and DJ-1 form a ubiquitin E3 ligase complex promoting unfolded protein degradation. J Clin Invest. 2009 Mar;119(3):650-60

Xu J, Zhong N, Wang H, Elias JE, Kim CY, Woldman I, Pifl C, Gygi SP, Geula C, and Yankner BA. The Parkinson's disease-associated DJ-1 protein is a transcriptional co-activator that protects against neuronal apoptosis. Hum Mol Genet. 2005 May 1;14(9):1231-41

Xu S, McCusker J, and Krauthammer M. Yale Image Finder (YIF): a new search engine for retrieving biomedical images. Bioinformatics. 2008. 24(17):1968-70

Zaharia M. Spark: In-Memory Cluster Computing for Iterative and Interactive Applications. Invited Talk at NIPS 2011 Big Learning Workshop: Algorithms, Systems, and Tools for Learning at Scale

Zhang L, Shimoji M, Thomas B, Moore DJ, Yu SW, Marupudi NI, Torp R, Torgner IA, Ottersen OP, Dawson TM, and Dawson VL. Mitochondrial localization of the Parkinson's disease related protein DJ-1: implications for pathogenesis. Hum Mol Genet. 2005. 14(14):2063-73

Zheng G, Li H, Wang C, Sheng Q, Fan H, Yang S, Liu B, Dai J, Zeng R, Xie L. A platform to standardize, store, and visualize proteomics experimental data. Acta Biochim Biophys Sin (Shanghai). 2009. 41(4):273-9

Zhong N, and Xu J. Synergistic activation of the human MnSOD promoter by DJ-1 and PGC-1alpha: regulation by SUMOylation and oxidation. Hum Mol Genet. 2008. 17(21):3357-67

Research Activities

The main contribution of this thesis lies in the introduction of a new bioinformatics technologies.

I implemented an image analysis workflow, using ImageJ an open source library, able to manage all the steps of a 2D-GE analysis. In this workflow, I collected and optimized different ImageJ tools in a plug-in, making available an unique open source software solution for analysing 2D-GE images. To test it, I performed a set of 2D-GE experiments on plasma samples from patients of acute myocardial infarction. I compared the results obtained by this procedure to those obtained using a widely diffuse commercial package, finding similar performances.

In order to enable the analysis of 2D-GE image extracted for biomedical literature and public repository I developed an effective technique for the detection and the reconstruction of over-saturated protein spots. Firstly, the algorithm reveals overexposed areas, where spots may be truncated, and plateau regions caused by smeared and overlapping spots. 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. The method is validated processing highly exposed 2D-GE images, comparing reconstructed spots with the corresponding non-saturated image, confirming that the algorithm enable correct spots quantification. Pixel correction in saturated and smeared spots allows more accurate feature extraction (called spot detection) allowing more reliable meta-analysis of 2D-GE images.

In the constest of the project Parkinson Information System (ParIS), I developed an automated literature analysis procedure to retrieve all the background knowledge available in public databases. This project was devoted to the testing of a new procedure for validating perypheral biomarkers of Parkinson Disease, previously described in literature. In this work I used ProteinQuest (PQ), a software developed by Biodigitalvalley. PQ is a web based platform for biomedical literature retrieval and analysis. PQ retrieves PubMed abstracts and extracts the text of the image captions from free full text articles. The image captions are extracted using the BFO Java

library (http://bfo.com) on the PDF version of the scientific papers (Natale et al., 2010). Briefly, PQ text-mining tool parses target documents searching for terms related to cured ontologies (proteins, diseases, bioprocess, body parts, etc.). Ambiguities in the terminology are resolved using a multiple search for more than one alias, as well as the co-occurrence of specific words which can deny or force the tagging process. PQ retrieved more than 51,000 scientific papers dealing with Parkinson Disease, and identify 4121 proteins cited in these papers. Out of these, I 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, I 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. Eventually, I compared the collected literature data to experimental gels from 90 subjects (45 Parkinson Disease patients, 45 non-neurodegenerative control subjects) to experimentally verify their potential as plasma biomarkers of PD. Some of the candidates, that arose from the literature analysis, were confirmed.

I contributed to the development of iMole a platform that automatically extracts images and captions from biomedical literature. iMole is a platform that automatically extracts images and captions from biomedical literature. Images are tagged with terms contained in figure captions by ProteinQuest iMole allows the user to upload directly their own images within the database and manually tag images by curated dictionary. I used iMole to build a 2DE database. Images were obtained by automatically parsing 16,608 proteomic publications, which yielded more than 16,500 images. Briefly, tagged 2DE gel images were collected and stored in a searchable 2DE gel database, available to users through an interactive web interface. The database can be further expanded by users with images of interest trough a manual uploading process.

I developed FunMod an innovative Cytoscape version 2.8 plugin able to identify biologically significant sub-networks within informative protein networks, enabling new opportunities for elucidating pathways involved in diseases. FunMod is the first Cytoscape plugin with the ability of combining pathways and topological analysis allowing the identification of the key proteins within sub-network functional modules. FunMod identifies within an informative network, pairs of nodes belonging to the same biological pathways and assesses their statistical significance. It then analyzes the topology of the identified sub-network to infer the topological relations (motifs) of its nodes.

I contributed in the development of a new software solution that could isolate images of gel electrophoresis, and check them for simple features such as possibly duplicated portions. This software runs an automatic check of ten of thousands of papers. This gel-checking software has proved to be able to identify various features, such as reused gel images or markings that suggested gels with potential irregularities.

I 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. I developed this software in order to analyse the oligopeptide fraction extracted from Fontina cheese at different ages of ripening and subsequently identified by an in-source fragmentation detectable with a single-quadrupole mass analyzer. This software performs an in-silico digestion of the major milk proteins, it calculates all the possible peptide fragments which are generated by the loss of the first N- or C-terminal amino acids, and finally, it matches the experimental ion chromatogram with the in-silico which generated theoretical spectrum to identify the exact amino-acid protein sequence of the unknown oligopeptide. With this tool, the useful insights into the proteolytic processes which occur during Fontina cheese aging are obtained, which leads to a better knowledge about the functional features of the proteolysis end product.

As bioinformatics researcher I supported different laboratory groups to perform experimental data validation by using literature-based network 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- protein network. All such connections were controlled in order to verify the consistency between the retrieved literature data and the experimental results. The protein-protein network obtained from PQ was exported for visualization to Cytoscape, a popular software platform for network analysis.

Curriculum Vitae

Massimo is an ICT project manager and Data Scientist at UniCredit SpA.
Massimo was graduated with a master degree in Food Science and Technology from the University of Parma in 2003, and is now completing his PhD in Information and Control Engineering at Politecnico di Torino.
Massimo has gained valuable experiences in design and management of innovative food, biomedical and ICT research projects, raising funds from public and private stakeholders.
Having worked in large companies and start-ups, Massimo brings a broad set of experiences in data analysis and interpretation, software development for image analysis, complex systems analysis, social media analysis, business and competitive intelligence.
Massimo is author or co-author of several scientific papers published on international peer reviewed journals, patents and conference presentations.

Experience

UniCredit SpA
ICT Project Manager
July 2014 – Present
Carry out feasibility studies and engineering design for innovative ICT solutions for big data analysis, machine learning, text mining, natural language processing.

Politecnico di Torino
PhD Student
January 2011 – Present
Develop and implement algorithms and software solutions for networks and complex systems analysis.
Author and co-author of several publications.

BioDigitalValley Srl (RGI Group)
Bioinformatics Scientist

January 2009 – May 2014
Collect, organize and analyse data to finalize scientific project results.
Design and develop innovative projects for food, biomedical and ICT research.
Study and implement innovative algorithms and software solutions for image analysis, data mining, social media analytic, web analytic, sentiment analysis, and natural language processing.
3 projects

Bioindustry Park del Canavese SpA
Senior researcher
November 2005 – December 2008
Manage the team of 2D-electophoresis and proteomics lab at L.I.M.A.
Carry out research projects, data collection and analysis, transfer technology activities.

Sodexo SpA
Quality Control Manager
January 2005 – November 2006
Develop and audit of quality and safety management systems according to International Standards ISO 9000, HACCP and 81/08 (Ex 626).

Consiglio Nazionale delle Ricerche
Scientist
2002 – December 2004
Validate and apply proteomics technologies in the study of food allergy.
Analysis and management of mass spectrometry data

Technical skills

Massimo is a Scala, Java, JavaScript, HTML5 programmer, and is familiar with Hadoop ecosystem managing both Map Reduce and Apache Spark frameworks.

Massimo has a strong background in text mining and natural language processing, and he knows the main open source library as Apache Open NLP, Stanford NLP, MALLET, UIMA, GATE, LingPipe.

Massimo works on image analysis using open java libraries as ImageJ. and network analysis, and and has a deep understanding of network analysis using different frameworks: Cystoscape, Gephy, JUNG, Tinkerpop Blueprints.

Massimo has a good experience in machine learning and manages some tools as Weka, and RapidMiner. In the Hadoop ecosystem he uses Apache Mahout and MLLib on Apache Spark.

Projects

As project manager in Biodigitalvalley Srl Massimo managed different resaerch projects:

VDNA Barcoding
March 2013 – Present
The Research Unit (UR) VDNA Barcoding buds from the idea of creating an advanced biotechnology center in the Aosta Valley that undertakes multidisciplinary research projects aimed at the study and the protection of the alpine biodiversity , and which provides services for genomic analysis. The UR is located in the Center of scientific and wildlife research of the Marais, headquarters of the Regional Museum of Natural Sciences. More specifically, the UR aims at investigating biological, ecological, genetic and taxonomic aspects related to the flora, fauna and micro-flora of the alpine ecosystems using genomic analysis based on DNA sequencing and highly polymorphic molecular markers.
The research team included: Museo Regionale di Scienze Naturali della Valle d'Aosta, Biodigitalvalley Srl, Parco Nazionale del Gran Paradiso, 3Bite, Parco Naturale del Monte Avic.

Open Source Drug Discovery Platform
March 2013 – Present
A research project which aims to:
1. Build a so-called regional "Unità di Ricerca" (Research Unit) dedicated to exploring the paradigm of Open Source Drug Development;
2. Set up a first open-access informatic platform dedicated to the Open Source Drug Development, allowing software and data sharing in a protected, cloud-friendly environment;
3. Attract other research partners, for future funding applications and industrial service set-up.
The research team included: Biodigitalvalley Srl, Research Area Advanced Computing and Electromagnetics (ACE) at ISMB, and Consorzio Interuniversitario Nazionale per l'Informatica.

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. This project involved the collaboration of 100 patients (50 PD and 50 neurological controls) who donated their blood for protein and DNA analysis.
ParIS was partially funded by the Valle d'Aosta regional government.
The research team included: Biodigitalvalley Srl, Università degli Studi dell'Insubria,

Istituto Neurologico Casimiro Mondino Pavia, Struttura Complessa di Neurologia e Neurofisiopatologia e Stroke Unit - Ospedale Regionale - Azienda Unità Sanitaria Locale della Valle d'Aosta.

IMAGE - Image Meta Analysis Generation and Exploitation
April 2009 – April 2010
A project dedicated to the meta-analysis of proteomic 2D-gel images in the literature. Total cost: about 830.000 euro, partially funded under the frame of the "legge regionale n. 84 del 7 dicembre 1993" (VAlle d'Aosta region)
The research team included: Biodigitalvalley Srl, Dipartimento di Informatica – Università di Milano.

List of Publications

Journal articles

- 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

- 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

- 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

- 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

- 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

- 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

- Articolo di rivista 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

- Articolo di rivista 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

Proceedings

- Articolo in atti di convegno Natale M., Caiazzo A., Bucci E.M., Ficarra E. (2012) A novel Gaussian fitting approach for 2D gel electrophoresis saturated protein spots. In: Proceedings of the International Conference on Bioinformatics Models, Methods andAlgorithms,, Vilamoura, Algarve, Portugal, 1-4 Feb. 2012. pp. 335-338

- Articolo in atti di convegno 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, Vietri sul Mare, Salerno, Italy, June 3-5. pp. 3-10

- Articolo in atti di convegno 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

Book Chapters

- 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