STATISTICAL TOOLS FOR CLASSIFICATION, INTERPRETATION AND PREDICTION OF BIOLOGICAL DATA

Oriol Senan Campos

DOCTORAL THESIS

Statistical tools for classification, interpretation and prediction of biological data

Author
Oriol Senan Campos

Supervisors
Marta Sales Pardo
Roger Guimerà Manrique

Tarragona, 4th August 2017

UNIVERSITAT ROVIRA i VIRGILI

Acknowlegdements

First of all I really would like to thank all the support and guidance from my thesis supervisors, Roger and Marta, from which I have tried to learn as much as I have been able. Now after four years of PhD I feel lucky for being part of the Sees Lab. Here I have found the perfect cocktail between an open friendly environment and a fine selection of outstanding science. This is also thanks to Núria, Manu, Francesco, Toñi, Toni V, Toni A, Sergio, Marc and Ignasi. I also would like to thank URV for my year of Martí Franquès fellowship, and the support from the personnel, specially from our department DEQ, secretary, professors and other PhD students. Finally, thanks to the workers of Catalonia for the funding destinated to my fellowhip FI-2014 during the last three years of my PhD.

If I have done a PhD is also thanks to people that trusted in me before. In my little scientific career I had the help of Dr Rui Alves during my BsC project, Dr Jordi García Ojalvo during my MsC thesis and Dr David Rosell in my first scientific job. My motivation to do research comes as well from all the nice discussions with my friends from Biotechnology Lleida, and my curiosity and love for nature since my early childhood, thanks to my family education, specially due to my mother Carmen and my father Pepe.

UNIVERSITAT ROVIRA I VIRGILI

WE STATE that the present study, entitled "Statistical tools for classification, interpretation and prediction of biological data", presented by Oriol Senan Campos for the award of the degree of Doctor, has been carried out under my supervision at the Department of Chemical Engineering of this university.

Doctoral Thesis Supervisors

Dr Roger Guimerà Manrique

Dra Marta Sales Pardo

Tarragona, 5th June 2017

Summary

Mathematical modeling has been used for more than 100 years in theoretical biology, but now has become a fundamental part of biological research. The improvement of omics technologies is making it possible to systematically profile DNA, RNA, proteins and metabolites in living organisms. This data opens the door for a paradigm shift towards a systemic biology approach.

Despite the sequentiation of the genome and a wide coverage of the proteome, one fundamental question remains open: How many different metabolites are in a given organisms or in a biological sample? Metabolomics could offer us an answer, but with the current best technique for detecting the maximum number of metabolites, a liquid chromatography coupled to mass spectrometry (LC/MS), oftenly only 20-30 metabolites are annotated among the thousands of signals in the data. One of the main causes of this partial annotation of metabolomics experiments is the lack of a proper method to correctly group the multiple signals produced per metabolite. To fill this gap our first goal is the following:

- To develop a new method to group and annotate the multiple adducts and isotopes produced by metabolites in LC/MS experiments.

For this purpose we have developed CliqueMS, a novel method which groups signals belonging to the same metabolite, based on similarity property that can be applied when we transform our metabolomics experiment into a newtork. Then we annotate metabolites within each group. Our method outperforms current annotating methods and may contribute to overcome one of the main bottlenecks for a better annotation of metabolomic experiments.

While part of the research is focused on developing new algorithms and devices to overcome the current limitations of omics technologies, there is already a massive use of omics devices which is producing a steady growth in the recorded biological data. The way we obtain and analyze this data is one of the main challenges of biology, sometimes refered as “big data to knowledge”. The use of large amounts of data is a necessary but not sufficient condition to a paradigm shift in biology. We need a combination of the multiple sources of biological data to achieve a better comprehension of systems as a whole. This combination won’t be straightforward, as we do not see simple associations, for example there is no general correlation between mRNA and proteins abundance. Integration of data demands new mathematical models, to unveil the complex relations between the different biomolecules.

Regarding the combination of multiple omics data, we study the effects of Hibbiscus sabdariffa extracts in humans by analyzing the metabolomic and transcriptomic response after its ingestion. Our goal in this investigation is the following:

- To elucidate the role of the polyphenols present on Hibbiscus extracts, associated to a positive impact in human metabolism.

From the metabolomic profile we report for the first time the molecular composition of Hibbiscus sabdariffa extracts. By combining transcriptomic and metabolomic patterns we observe an alteration of the immune response, the mitochondrial function and the energy homeostasis. These results show an example of data integration. However, there is another important problem of omics data: how to cope with the different interpretations provided by different matemathical models?

Omics data is complex, oftenly highly variable and findings might be hard to reproduce, or even contradictory. Results are obtained through mathematical models, and they are validated by its capacity for prediction. Nevertheless, are good predicting models also good for interpretation? This issue is not only restricted to omics data, but is a general problem for biological data. We analyze the interplay between prediction and interpretability by evaluating the role of different computational models for predicting platelet deposition. In this research our goal is:

- To develop three complementary approaches to predict platelet deposition, as a first step towards a multiscale model for thrombosis.

Platelet deposition is the trigger of thrombus formation, a very important pathology leading to hearth stroke and embolia. In this study, we demonstrate that by measuring platelet concentration, vessel tissue and other variables our models can predict the platelet deposition in a new sample.

Resum (Català)

Els models matemàtics porten usant-se en biologia desde fa més de 100 anys, a l'àmbit de la biologia teòrica. Avui en dia, però, són ja una part fonamental del conjunt de la recerca biològica.

Les tecnologies òmiques han millorat tant que ja és possible la caracterització massiva molecular de ADN, ARN, proteïnes i metabòlits. Aquestes noves dades obren la possibilitat d'un canvi de paradigma, cap a una biologia més sistémica. S'ha aconseguit la sequenciació del genoma, una bona estimació del nombre total de proteïnes en molts éssers vius però una pregunta resta a l'aire: Quants metabòlits trobem en un organisme? I en una mostra biológica?

La millor técnica per a detetectar el màxim nombre de metabòlits, una cromatografia líquida acoplada a espectroscopia de masses, (LC/MS) malauradament només permet anotar entre 20 i 30 metabòlits, d'un total de senyals que habitualment ronden els milers.

Per a millorar l'anotació d'aquests experiments, hem desenvolupat el CliqueMS, un nou algorisme per a agrupar i anotar les múltiples senyals que un mateix metabolit produeix en els experiments de LC/MS. Aquestes senyals són variants isotòpiques, ionitzacions amb diferents ions, anomenades aductes, i fragments. El nostre métode millora els actuals mètodes d'anotació. Amb aquest métode volem contribuir a superar un dels principals colls d'ampolla per a l'anotació completa dels experiments en metabolòmica.

Una aplicació directa que esperem per al nostre nou métode es calcular la distribució d'aductes als experiments de metabolòmica no dirigida. L'estimació d'aquesta distribució serà una informació molt útil per als algorismes d'anotació, i en general per a millorar la metabolómica no dirigida.

És necessari, per tant, seguir millorant les tècniques i els mètodes per a una major qualitat i precisió de les dades òmiques. Com hem vist en el cas de la metabolòmica, totes les tecnologies òmiques tenen alguns inconvenients o mancances, però el seu ús no para d'augmentar i hi ha una creixement en el total de dades biologiques emmagatzemades. Aquestes dades són necessàries, però no suficients, per al canvi de paradigma en la

biologia. Necessitem combinar les diverses fonts d'informació per a comprendre els sistemes desde la seva totalitat. Sabem que aquesta combinació no serà directa, ja que no es veuen simples associacions matemàtiques entre les diferents biomolècules, per exemple no hi ha una correlació general entre els nivells de proteïna i ARN. La integració d'aquestes dades requereix nous models matemàtiques, per a descobrir les complicades relacions i dependències entre les biomolècules.

Com a exemple de combinar diverses dades òmiques, estudiem els effectes terapèutics d'una infusió de Hibisscus sabdariffa. Per coneixer aquests efectes analitzem la resposta metabolòmica i transcriptòmica després de la ingestió d'una infusió a partir d'un extracte de la planta. Observem una alteració a la funció mitocondrial, i al metabolisme energètic encarregat de l'homeostasi. Aquestes alteracions es veuen tant a la resposta metabòlica com als canvis als patrons d'expressió génica. A més del repte de l'integració de dades metabolòmiques, hi ha un altre problema fonamental de les dades òmiques: com avaluar les diferents interpretacions provinents de diferents models matemàtics?

Al món de les dades òmiques, trobem que aquestes són complexes, que poden ser altament variables i que els resultats són difícils de reproduïr, quan no contradictoris. Els resultats s'obtenen mitjançant model matemàtics, i la validesa d'aquests es determina per la seva capacitat de predicció. Aixi doncs, és incompatible la predicció amb la interpretació? Aquest problema no es restringeix només a les dades òmiques, sinó que és un problema general de les dades biològiques. En aquesta tesi incloem un estudi on avaluem com diferents models computacionals prediuen l'acumulació de plaquetes, el procés que desencadena la trombosis. Analitzem com la capacitat de predicció i la interpretació no són incompatibles, mitjançant tres models molt diferents, un basat en equacions mecanistiques, un model d'aprenentatge de màquina i un model fenomenològic derivat de la informació del model d'aprenentatge de màquina.

En quan al procés d'acumulació de plaquetes, podem predir aquesta acumulació en una mostra desconeguda a partir de les variables del nostre model, que són molt més fàcils de mesurar que la mateixa acumulació. Veiem que la influència del teixit de la vena on hi ha la lesió és molt important per a desencadenar la resposta de les plaquetes. Esperem en una futura aproximació a la trombosi usar també informació espacial, per a tenir en compte la localització de les plaquetes acumulades i incloure també l'efecte del fibrinògen.

Contents

Chapter 1

Introduction

1.1 Biology: Once upon an experimental science

Biology is the science that studies living organisms. Our notion of biology has changed with time together with the still much debated definition of life. These changes have occurred in parallel to the establishment of new forms of acquiring knowledge on biological systems, including mathematical modeling, a fundamental tool in many areas of biology.

Contemporary biological theories are based on three complementary paradigms: experimentation, observation and inference. A brief walk through the history of biology is useful to illustrate how the use of mathematical models has become a cornerstone of many areas in biology.

The study of living organisms and our environment has been a constant necessary factor for the development of human civilization. To obtain food, to find shelter, to have clothes or to avoid danger, humans always needed to comprehend natural phenomena. This has generated many forms of practical knowledge, particularly regarding living organisms. This sort of practical knowledge is still necessary in many present day human activities, like fishing, plant and animal breeding, forest exploitation, etc ... Over time, and for certain civilizations, there was a privileged group of people that could dedicate their life to observe and study nature. Liberated from the burden of “practicality”, they could make much deeper analysis of observed natural phenomena and extract generalities from the particular cases; a paradigmatic example being the works of Aristotle, who set the basis of biology for a very long period.

The start of the Renaissance triggered a scientific-technical revolution, which launched biology far beyond the knowledge gained during Middle Ages. The establishment of the scientific method, together with new technological innovations, like the microscope, brought great advances, including the study of microorganisms by Antonie van Leeuwenhoek, or the first observation of the cell by Robert Hook.

More importantly, there was a paradigm shift; biology turned into an experimental science, and hypotheses had to be proven or rejected in controlled experiments. This change in paradigm lead to the rejection of old principles that were not based on empirical evidence, such as spontaneous generation.

Despite being an empirical science, biology also relies on direct observation of nature. There are a series of branches of biology (and other related sciences) which require field work, since a lab experiment cannot always replicate the conditions we find in nature. In this regard, what we can do is to systematically classify and to study the habits and the distribution of species of plants, animals and microorganisms on the biosphere.

During the XVIII and XIX centuries, the growing number of expeditions obtained new records of plants, animals and fossils. The organization of this data by Carl von Linné created the modern taxonomy system. All this ordered data provided a more complete picture of nature, and was a source for many new discoveries, among of them the most important was Darwin’s theory of evolution.

So far, we have seen that the progress of biology has come from experimental evidence, and also from direct observation and systematic description of nature. What if, there is a biological process that we want to study, but it is impossible to reproduce it in the lab neither observe it in nature?

The alternative is to measure an indirect variable, either in a controlled experiment or in field work, that is related with the unobservable one. In addition, we might use probability theory to confirm the relation between the observed and the unobserved variable, which is known as inference.

This was the rather the case in another of the findings that changed biology forever: The discovery of Mendelian inheritance and the beginning of genetics.

1.2 More than 100 years of mathematical biology

Ernest Rutherford, Nobel Price in chemistry in 1908, famously and provocatively stated: “All science is either physics or stamp collecting”, (although the attribution to him is disputed). Biology has been criticized for being too descriptive, “stamp collecting”, rather than formulating general laws that can be expressed in mathematical terms. It is true that there is a certain resistance in biology to the incorporation of mathematics, and to focus more on exceptions than on generalities.

That being said, despite the fact that biology is not as quantitative as chemistry or physics, the use of mathematical modeling in biology is older than it might seem. A great example of this is the first modern evolutionary theory, which with the help of mathematical models combined two of the most fundamental theories in biology: genetics and Darwin’s evolution theory.

Going back to the origin of genetics, Gregor Mendel studied how an observable trait, the colour of pea seeds, was inherited. He observed that when crossing a pure green line with a pure yellow line, the resulting offspring was only yellow. Nevertheless, if these offspring were crossed between them, the green seeds were observed again in the second generation. The yellow was “dominant” approximately at a frequency of 3:1 versus the green, which was “recessive”.

With this and other experiments Mendel formulated general inheritance laws, later called laws of Mendelian inheritance. This laws were “rediscovered” by new experimental evidence almost 30 years after Mendel’s publication in 1865. It is worth noting that no molecular details of DNA were known at the time, actually it would take fifty years more for Watson, Crick and Franklin to elucidate the structure of DNA.

In the early twentieth century, more and more experimental evidence for Mendelian inheritance and for its corresponding cellular mechanisms was found. A group of geneticists, spearheaded by William Bateson, were trying to match the results with the theory of evolution of Darwin. They were known as the Mendelians. Mendelian inheritance stated that an observed trait depends on the genetic dotation, which consists on a maternal and a paternal allele (genotype). Depending on the alleles and their hierarchy (dominant, recessive ...) we will observe a certain phenotype. As a result Mendelians expected evolution to be discontinuous. To ilustrate this concept, let's us assume one pea seed has this genotype for the color: green(paternal)/
green(maternal). The pea is green. Then we mate this plant with another green/green plant. In sexual reproduction only one allele per individual is transmitted. In the absence of mutation for pea_1 = green and pea_2 will be green and the offspring will have a green/green phenotype. However, if the allele from pea_2 has a mutation such that pea_2 = yellow, the genotype of the offspring will be green/yellow and we will observe a sudden change of green into yellow, because if we remember Mendel's example, yellow is the dominant allele.

Darwin proposed that natural selection would lead to a gradual change of traits, instead of the discontinuous one proposed by the Mendelians. The scientists who supported Darwin's continuous evolution of traits were called the Biometricians. Biometricians were working with the growing amount of biological data to obtain useful knowledge. The work of Francis Galton, Karl Pearson and others developed a core of ideas and tools that transformed probability theory and statistics. Some of these are the standard deviation, the variance, the Pearson Correlation, the Chi-Square test or the p-value, which are still used today.

For instance, these scientist observed that certain variables, such as human height, follow the pattern of gradual or continuous evolution. In their data, taller parents typically have taller offspring, so Biometricians observed correlation between father and son height, and a continuous distribution of heights.

The agreement between the statements of the Biometricians and those of the Mendelian geneticists seemed impossible. However, the first modern evolutionary synthesis was able to reconcile both arguments.

Modern evolutionary synthesis, credited mainly to R. A. Fisher, J.B.S Haldane, and S. Wright could unite ideas from different disciplines and created a new paradigm: Evolution through genetic changes. Remarkably, this theory used mathematical models to provide evidences and to formalize concepts.

For example, R. Fisher demonstrated that, for a Mendelian trait controlling height, father and son would have deviations from perfect correlation, due to changes from dominant/recessive to recessive/recessive [1]. Then correlation would be different depending on familiar relationship (sibling, cousin, father ...) and generation. Therefore Mendelian genetics could give (imperfect) correlation as reported previously by the Biometricians.

Broadly, the modern evolutionary synthesis opened the new field of population genetics, and through mathematical demonstrations created many concepts that nowadays are a very important body of genetics.

This example illustrates the importance of mathematical modeling for the advance of biology, and the feasibility to formulate general laws of biology. Getting closer to the investigations carried out in this thesis, today the use of modeling and statistics is not only limited to theoretical research, but is equally important for the interpretation of experiments and for the description of nature.

If exploration journeys during the eighteenth and the nineteenth centuries were characterized by the discovery new species on remote lands, today we are witnessing to a great effort for the characterization of living organisms at the molecular scale.

The rise of omics has allowed a very fine profiling of biomolecules: DNA, RNA, proteins and metabolites. Many other types of biological data are being recorded as well. Thus we see a fast growth in the accumulation and in the complexity of stored data. We also refer to this as “big data”.

What are the main features of each of this omics technologies? Which are the limitations, and the challenges that we face by using this kind of data? In the following section we will try to answer to these and other questions related to the rise of omics in biology.

1.3 Omics: Big data in biology

Omics technologies are a collection of devices and tools that can perform high-throughput, automated, analysis of DNA, RNA, proteins and metabolites. They are an interesting source of data for understanding global and emerging properties of living organisms, and to elucidate the multiple interdependencies between the different biomolecules.

In addition, great promises have been put in omics technologies for applied biomedical research. One of the goals of omics technologies is to find biomarkers: patterns of genes, transcripts, proteins or metabolites which can be used for disease subtype classification, disease progression, selection of treatment, early diagnosis and many other applications.

Now let's see the particularities each type of omics technology.

Genomics and transcriptomics

Sequentiation of DNA is the most developed omics technology. DNA is a polymer whose monomers are four different nucleotides, which can be easily identified. Genome databases have the whole sequence of many species [2], and equipment is constantly improving. The goal of genomics is to study the structure, function and evolution of genomes.

In the first sequenciation of the human genome ([3]) the reported number of genes was approximately 20.000 genes, much fewer than expected. Therefore, and given the complexity of human body, the role of genes was not completely described using the traditional axiom: "one gene, one protein, one function".

Much of the applied research in genomics consists on finding disease-causing genetic variants. Genetic linkage analysis was developed for this specific purpose. This methodology has been successful for the identification of genes responsible of Mendelian diseases (single gene disease), but not for complex multifactorial diseases [4]. Another statistical method to identify disease-causing genes are genetic wide association studies (GWAS). This method has found many variants associated to disease, but it has a limited reproducibility and biological relevance [5].

DNA's relative simplicity has allowed the complete sequentiation of the genome in many organisms and a straightforward application of mathematical formalism. However, and because of its function, it does not provide enough information for the study of many processes. Therefore, the product of DNA, RNA, could be a much better alternative.

Contrary to DNA, RNA changes with time and location. The complete profiling of mRNA and other types of RNAs to understand biological processes and to find biomarkers is called transcriptomics.

The transcriptome is more complex than the genome because one gen may have multiple transcripts, a process known as alternative splicing. Microarrays [6] were the first high-throughput devices to monitor RNA. They can measure thousands of transcripts, but have a limited threshold regarding RNA abundance. The newer RNA-Seq technology [7] can detect, in principle, many more transcripts including alternative splicing events, and has a bigger range of detection.

Given this inner complexity of the transcriptome, there is a wide variety of experimental designs, depending on what we need to study, it might be the change in expression during time (time series), comparing conditions (treatment vs control, ...), different tissues or cell types, etc... Generally we want to report differentially expressed genes (DEG) in the different groups or conditions. Another commonly reported result is sets of genes, sometimes associated to a function or process, which have a different pattern of expression in one of the particular groups. Both DEG or sets of genes can potentially be a biomarker, for example a sign of an early diagnosis.

In this regard the massive use of transcriptomics has described RNA expression for many biological processes, like cancer, and has been used to discover many RNA biomarkers, [8]. Unfortunately, many of this findings are hard to reproduce [9]. There are multiple causes for this, firstly the difficulty to control all the variables affecting RNA expression, which can be solved by better experimental designs. Spurious results tend to appear when statistical methods to find out DEG or sets of genes are not used properly [9]. Finally, it is challenging to associate RNA expression to a particular function or biological process, given the layers of regulation between the gene expression and the final process, and that is why it is very important to measure other biomolecules, like proteins.

Proteomics: Uncovering the proteome

Proteins are the main product of the translation of mRNA. They perform a wide variety of cellular functions: catalyzers of biochemical reactions, main elements of cell structures like the cell wall, cellular machines like ion transport channels, electron transport chain, microtubules... Thus, proteins perform more functions and play a direct role at processes compared to RNA, which is more in the regulatory domain.

Due to its perceived importance, the proteomics domains quickly received a lot of attention. In relation to proteomic relevance, one of the early proteomics paper [10] stated: “By the turn of the millennium if not sooner, we will see a dramatic shift of emphasis from DNA sequencing and mRNA profiling to proteomics”. Is this the current situation?

The fact is, although proteomics has advanced a lot in the recent years, the specific nature of proteins makes its measurement much more challenging than profiling DNA or RNA. Firstly, extraction of all proteins from a sample is very difficult. The proteome is dynamic in space and time, and is more complex than the transciptome. This is because proteins might be in different states, which really affects their function. They can be altered by post-translational modifications (PTMs), by conformational changes or by interaction with other proteins.

There is a rich variety of analytical techniques to profile the proteome, or the different subproteomes coming from different tissues, body fluids, etc... There are qualitative approaches, more focused on identifying the maximum number of putative proteins. Others are more quantitative, they seek for precise measurements of the levels of protein abundance, to detect differences between groups of samples. (See [11] for a review on the multiple proteomics strategies).

Identification of proteins in proteomics is more complicated than in the case of RNA or DNA. Sequence information of gene or transcripts is a prior knowledge used this identification of proteins. The functional role of proteins is difficult to infer from the RNA or DNA data. For instance, mRNA abundance might be useful to predict protein concentration, but there is not a general correlation between the level of mRNA and protein [12]. So, the transcriptome and the genome are not enough to comprehend the proteome. We cannot understand many biological processes without a good quantification of the proteins, because there are many processes that are direct interactions between proteins, such as protein complexes, PTMs or signalling. On the other hand, proteins are very large molecules, susceptible to many changes and modifications. That complicates the interpretation of proteomics data, and might increase the factors that we have to take into account for a protein or a group of proteins to become a biomarker.

Name	**Measured Molecules**	**Identification**	**Coverage**
Genomics	DNA Sequence	4 Nucleotides	Complete for many organisms
Transcriptomics	RNA sequence, RNA abundance, variable in location and time	4 Nucleotides	Still ongoing, very complete for mRNA in some species
Proteomics	Protein identification, protein abundance, variable in location and time	20 Aminoacids, modifications	Not complete, great advances in some organisms
Metabolomics	Metabolite identification, metabolite abundance, variable in location and time	Different molecules	Incomplete

Table 1.1: Summary of different omics techniques

Metabolomics: An ensemble of small molecules

Metabolomics is the last of the omics to become a high throughput technology. We call metabolites to all the small molecules transformed in biochemical reactions. The goal of metabolomics is to measure, as precisely as possible, the maximum number of metabolites from a biological sample. Unfortunately, the metabolome is such and ensemble of different chemical structures. Their chemical properties are so different that purification and extraction processes are very challenging. In addition, the large heterogeneity of chemical structures, even more than in proteins, makes it very hard to annotate all metabolites from a sample.

Metabolites are the reactives and products of many cellular reactions. That makes them as the best signature for measuring biochemical activity. [13]. Moreover, data might be easier to interpret because, unlike proteins, metabolites do not have modifications such as denaturalizations, different structural conformations, PTMs, etc...

The two main technologies for metabolomics are based in Nuclear Magnetic Resonance and Mass Spectometry (MS). As in proteomics, the metabolome is variable in location and time. Therefore, there are many strategies for measuring the metabolome. Some are centered in high-precise measuring of a selected group of metabolites (targeted metabolomics). Other strategies are suited for a particular subtype of metabolites (like lipidomics for lipids) and some are more general, called untargeted metabolomics, which try to identify the maximum number of metabolites from a sample.

The total putative number of proteins has been estimated, in several organisms, due to bioinformatic tools and proteomic experiments [14]. That is not the case for the metabolome. First, we cannot use the genomic and transcriptomic information to make predictions of putative metabolites, or not as equally as in proteins. Second, even in the most broad technique for detecting the major number of metabolites: a liquid chromatography coupled to mass spectrometry (LC/MS), we do not know exactly how many metabolites are present in the samples.

A recent review [15] noted that the thousands of signals detected in an LC/MS experiment (those signals are called features and have time, mass to charge (m/z) and abundance values) may belong to a reduced number of metabolites. The reasons are mainly two.

Firstly each metabolite produces multiple features. Natural isotopic variants are detected as different features. Also, metabolites can ionize with different ion species, producing multiple features, called adducts of the same metabolite. Finally, metabolites can be fragmented and have covalent interactions with other metabolites, detected as well as different features.

Secondly, there are many other features which appear due to contamination, chemical noise and errors in the signal processing. Given the complexity of the spectral data resultant from untargeted metabolomics experiments, it is necessary to correctly group the multiple features belonging to the same metabolite.

The next step after reducing the thousands of observed features to hundreds of putative metabolites is its annotation. The number of features whose m/z match with an entry on a metabolomics database is small. This happens because the putative number of metabolites is much larger (millions) than the number of entries in the spectral databases (thousands) [16]. To solve this gap many algorithms have been developed to annotate metabolites not present in spectral databases. See [17] for the state of the art of current methods. The whole process of metabolite annotation is not fully automated, still depending in manual work. So often the total number of annotated metabolites is roughly 20-30 [15]. This is the main bottleneck for reporting new metabolites and for the goal of a more complete picture of the metabolome.

Multiomics and other biodata, communication breakdown?

Monitoring biomolecules is giving us new perspectives about the molecular composition and organization of living organisms. Massive omics data has to combine with other sources of data: databases for disease, drugs or toxics, medical records, molecular biology studies, etc ...

How can we connect information from multiple sources, involving different data types and often corresponding to different scales? This is one of the main challenges in providing comprehensive systems descriptions of biological processes.

Some attempts [18] have tried to reorganize data, to find hidden relations between phenomena at different levels of description (microscopic and macroscopic). This is the case of the diseasome and the toxome, two initiatives that connect diseases or toxics, by their similarities at the molecular level (genes, proteins, etc ...) and also at the macroscopic level (treatment, symptoms, etc...). This and other projects exhaustively use bibliographical resources, like the genomic-scale metabolic models. They will help to cope with distants sources of information of things that we already known and they will be a better data source for new investigations.

One example that systemic desciption is challenging is mRNA and protein. Even though they are two related biomolecules they are not generally correlated. While simple association between some data may be straightforward [12], others will require the development of new mathematical models that can offer insights to the intrincated dependencies between biomolecules and from single cells to tissues, organs and organisms.

1.4 Modeling perspectives

The more we know about the fundamentals of the cell, the more complicated it seems. We are discovering new elements that influence cell function, at all levels. The world of epigenomics is changing our vision of DNA, we see new functions for different types of RNA that makes gene expression more complex, promiscuous enzymes that perform side reactions... It seems difficult to manage all these levels of detail. We need to focus on the key elements that provide more information about the process we want to study.

A very good approach to find out the most relevant variables of a biological process is the use of mathematical models. We can use all this new biological data to build models, and then use this models to understand the data. Finally, the validity of a model is tested for its capacity to make predictions.

Each modeling approach entails a certain degree of simplification and some assumptions, necessary to fit the model to the data. A very usual process is to evaluate the output produced by the model, and correct the model depending on the outcome, or correct the model based on an accuracy measure.

Using this procedure, we may obtain some results that may generate hypothesis, such as a new interpretation of a phenomena, a pattern which is a signature for a disease... It is important, then, to modify the approach.

When we are in a confirmatory study, we should, if possible, establish our model, our accuracy measure and the design of the experiment previously to the data. This may prevent the so called human in the loop overfitting [19], which is the selection and reselection of data, model and accuracy measure until the reported model performance shows some spurious accuracy.

If we readapt the model on and on to obtain the maximum performance, the final model might not provide an explanatory answer to the hypothesis we wanted to test, and the accuracy measure may not be the most appropiate for the data that we are reporting. The perils of this practique are erroneous interpretations of the results and low reproducibility.

Models for classification

Classificators use categorical, ordinal and quantitative variables to predict a qualitative outcome. Classificator models are very integrated in the generation of omics data. In genomics and transcriptomics identification of nucleotides is fully automated and even the reconstruction of the genome, while in proteomics and metabolomics we still need better models to identify proteins and specially metabolites.

We called supervised learning when we know the categories of the classification, and we train our model to predict these categories with the input variables. This is the already mentioned case of protein or metabolite identification, but also one of the most important applications of omics data: the discovery of biomarkers to diagnose a disease, select a treatment...

Another approach for classification is the unsupervised learning. In this methods we group our samples, by clustering or other means, based on a distance measure computed by our classificator. The difference with supervised learning is that the we do not set previously the groups or categories of our classificator. Unsupervised learning my be used to discover some hiden groups in our samples, unobserved patterns, unknown effects of some variables...

When we build a model, we want our model to be very predictive and accurate. That is a clear case, for example, in the design of a classificator for diagnosing cancer. Nevertheless, we also want our model to be interpretable. In the same example, we would like to understand why a sample is classified as cancer, to elucidate a possible mechanisms of the disease. Given that, it is compatible prediction with interpretation?

Is there a tradeoff between prediction and interpretation?

Prediction and interpretation may indeed be seen as two opposite poles, but the truth is that they are two complementary aspects of modeling. It is more likely for a highly predictive variable to be as well the causal explanation of a phenomena.

We wanted the best predictive models, but we also want to understand biological processes. Certainly a more complex model, which can be better at predicting, is also harder to interpret. A solution for this is to derive simplified models from the complex models that give the best performance.

A method for simplification is the selection of the most important variables of a model, or in machine learning terminology, feature selection. The ongoing principle is that once we fit a model that has the desired degree of accuracy, there might be some input variables that are irrelevant or redundant. We can imagine this situation, for example, in transcriptomics data when thousands of transcipts can be used as a signature for a disease, or in a complex model with many quantitative and categorical variables. Selection of the most important variables can also avoid overfitting, because fitting the model with non relevant variables can adjust to some values which are specific of the training sample and not of the general phenomena that we want to predict.

Simplification, however should not compromise prediction. An apparently simple system can produce complex outcomes depending on the relation between its components. From an experimental and applied point of view, there are some variables which are more difficult or costly to measure, so reducing to the most important variables can, apart from aiding in the interpretation of a model, guide upcoming experimentations.

Works presented in this thesis

Senan O, Sales-Pardo M, Guimerà R *et al.* **CliqueMS: A tool for adduct annotation for LC/MS spectral data**. *In preparation.* 2017

Senan O, Pallarés J, Cito S *et al.* **A comprehensive study on different modeling approaches to predict platelet deposition rates in a perfusion chamber**. *Sci Rep.* 5 (13606). 2015

Beltran-Debon R, Senan O, Joven J *et al.* **The acute impact of polyphenols from Hibiscus sabdariffa in metabolic homeostasis: an approach combining metabolomics and gene-expression analyses.**. *Food Funct* 6, 2957 – 66. 2015

Article	**Classification**	**Interpretation**	**Prediction**
CliqueMS	X		
Platelets		X	X
Hibiscus		X	

Table 1.2: Scope of of the models developed in the articles

Bibliography

[1] Fisher R. A. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[2] Cunningham F., Ridwan Amode M., and Flicek P. *et al.* Ensemble 2016. *Nuc Acid Res*, 44(Database Issue):D710–6, 2016.

[3] Venter J. C, Adams M. D., and Zhu X. *et al.* The sequence of human genome. *Science*, 291(5507):1304–51, 2001.

[4] Ott J., Wang J, and Leal S. M. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genetics*, 16(5):275–84, 2015.

[5] Visscher P, Brown A. M., and Yang J. Five years of gwas discovery. *A J Hum Genetics*, 90(1):7–24, 2012.

[6] Schena M., Shalon D., and Brown PO. *et al.* Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–70, 1995.

[7] Wang Z., Gerstein M., and Snyder M. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genetics*, 10(1):57–63, 2009.

[8] Su Z., Fang H., and Tong W. *et al.* An investigation of biomarkers derived from legacy microarray data for their utility in the rna-seq era. *Genome Biol*, 15(12), 2014.

[9] Shi L, Jones W. D., and Tong W. *et al.* The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinformatics*, 9(S9), 2008.

[10] Anderson N. L. and Anderson N. G. Proteome and proteomics: New technologies, new concepts and new words. *Electrophoresis*, 19:1853–61, 1998.

[11] Angel T. E., Aryal U. K., and Smith R. D. Mass spectrometry based proteomics: existing capabilities and future directions. *Chem Soc Rev*, 41(10):3912–28, 2012.

[12] Ebrahim A., Brunk E., and Palsson B. O. Multi-omic data integration enables discovery of hidden biological regularities. *Nat Communications*, 7, 2016.

[13] Patti G., Yanes O., and Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol*, 13:263–69, 2012.

[14] Kühner S, van Noor V., and Gavin A. *et al.* Proteome organization in a genome-reduced bacterium. *Science*, 326(5957):1235–40, 2009.

[15] Zamboni N, Zahatelian A., and Patti G. Defining the metabolome: Size, flux, and regulation. *Mol Cell*, 58(4):699–706, 2015.

[16] Aguilar-Mogas A., Sales-Pardo M, and Yanes O. *et al.* imet: A network-based computational tool to assist in the annotation of metabolites from tandem mass spectra. *Anal Chem*, 89(6):3474–82, 2017.

[17] Kind T, Tsugawa H., and Fiehn O. *et al.* Identification of small molecules using accurate mass ms/ms search. *Mass Spectrom Rev*, 2017.

[18] Baker M. Big biology: The 'omes puzzle. *Nature*, 494(7438):416–9, 2013.

[19] Hofman M. J., Sharma A., and Watts D. Prediction and explanation in social systems. *Science*, 355(6324):486–88, 2017.

Chapter 2

CliqueMS: A tool for adduct annotation for LC/MS spectral data

ARTICLE *in preparation*

Authors
Oriol Senan Campos
Antoni Aguilar-Mogas
Miriam Navarro
Oscar Yanes
Marta Sales Pardo
Roger Guimerà Manrique

2.1 Introduction

A powerful technique for untargeted metabolomic studies is the use of liquid chromatography tandem mass spectroscopy (LC/MS). However, the need to substantially process raw data from LC/MS samples in order to obtain reliable annotations that later can be used for metabolite identification using tandem MS, poses a serious obstacle for the real throughput analysis of complex samples.

There has been a huge progress for a complete automatization of metabolite identification, see [1] for a review on Computational Metabolomics. Despite all this new software and tools, there are two limiting steps that still make the whole process low throughput: the multiple signals produced per metabolite and the small number of metabolites in spectral databases compared with the large number of putative metabolites.

Each metabolite produces multiple signals: natural isotopic variants, ionizations with different ion species (called adducts), fragmentations and covalent interactions with other metabolites. Grouping correctly the signals of the same metabolite is crucial for a posterior identification.

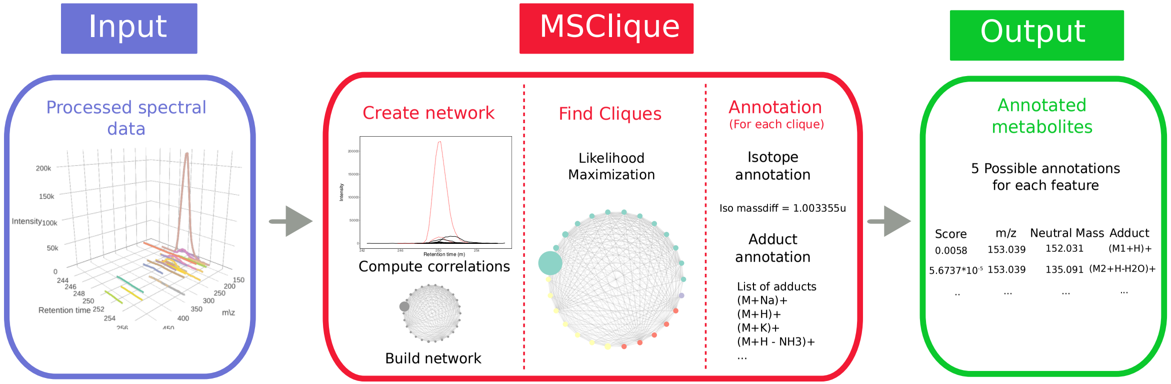
However, despite the availability of computational solutions, the truth is that many steps of the annotation process are still done manually.

In here, we will focus in the annotation of molecule adducts from LC/MS data. The lack of reliability of available methodologies for this task lies on the complexity of the spectral data from complex samples in which we have millions of signals for different m/z values localized in a continuum of retention times.

To aid in the automatization of this process we have developed CliqueMS, a computational tool that produces reliable annotations for molecule adducts from signal picking XCMS data (see Fig. 2.1).

CliqueMS annotates adducts in complex LC/MS samples based on the following assumptions: 1) Adducts of the same metabolite exit the column at the same retention time; 2) The plausibility of a specific adduct annotation is proportional to the frequency with which such adducts are observed in real samples.

Based on these assumptions, we have developed a tool that unlike other



Score	m/z	Neutral Mass	Adduct
0.0058	153.039	152.031	(M1+H)+
$5.6737*10^{-5}$	153.039	135.091	(M2+H-H2O)+
..	...	...	...

Figure 2.1: Schematic representation of the CliqueMS algorithm. Given processed spectral data of a complex metabolite sample, the algorithm identifies the signals belonging to the same metabolite. After establishing correlations between all signals, CliqueMS looks for cliques in the correlation network. Each clique is composed of the signals corresponding to adducts of the same metabolite. Then, for each clique, the algorithm proceeds to annotate each signal by establishing the parental ion neutral mass. The final output is a list containing all annotated adducts of all the metabolites.

approaches uses a mathematically principled approach to obtain the most plausible groupings of signals according to the similarities between them. Then we use existing data on the observed occurrence of potential adducts in real samples in order to annotate the signals in each one of the groups.

We find that CliqueMS is to consistently correctly annotate more metabolites and identifies a larger number of signals than existing widely-used approaches such as CAMERA [2], for both pure and complex samples.

2.2 Materials and Methods

2.2.1 Description of CliqueMS

Formally the problem that CliqueMS faces is the following. Our spectral data is comprised of a set of signals characterized by an m/z value and intensity vector $\{[(m/z)_i, \boldsymbol{f}_i]\}$. For each signal i, we obtain the intensity vector discretizing the signal into K equal bins so that $\boldsymbol{f}_i = (f_i(t_k); k = 1, \ldots, K)$, where $f_i(t_k)$ is the measured intensity at retention time tk, where $tk = t_{k-1} + \mathit{\Delta} t$ and $t_0 = 0$ (in our analysis, $\mathit{\Delta} t$ depends on the the mass detector operational parameters and the spectral data processing program). Given this data CliqueMS aims at providing a set of plausible annotations for complex samples that reflect the two previously mentioned assumptions.

To achieve this goal, we have identified three main steps (see Fig. 2.1): 1) the construction of a similarity network, where each node represents a signal and edges are weighted according to the similarity between signals; 2) the identification of the most plausible division of the similarity network into cliques (fully connected groups); 3) the annotation of the adducts corresponding to the same neutral mass for the signals within each clique.

Step 1: Construction of a similarity network between signals In order to provide meaningful annotations, first we need to group signals that are similar, so that signals corresponding to adducts of the same metabolite belong to the same group.

As previously mentioned, CliqueMS is based on the expectation that all adducts of the same metabolite have a similar retention pattern. Or in other words, we expect that signals corresponding to adducts of the same metabolite have non-zero intensities for the same retention time values. A critical step is thus select an appropriate measure of similarity between signals that reflects our expectations and allows the construction of a similarity network to obtain reliable groups of signals.

A possible choice of similarity function is the Pearson correlation between intensity vectors as considered in [2] within the context of spectral signal similarity. The caveat of the Pearson correlation coefficient is that it is suited to detect similarity of signals that are monotonously growing/decreasing and therefore it is *a priori* not an optimal option when signals are non-monotonous such as the spectral data we consider.

To overcome this caveat, we propose to use the cosine similarity, a simple measure that assesses the proportionality between intensity vectors:

$$\cos_{ij} = \frac{\sum_k f_i(t_k) f_j(t_k)}{\|\boldsymbol{f}_i\| \|\boldsymbol{f}_j\|} \tag{2.1}$$

where $\|\boldsymbol{f}_i\| = \sqrt{\sum_k f_i(t_k)^2}$.

To assess the ability of Pearson and cosine similarities to discriminate between adducts of the same metabolite from adducts of different metabolites that are coeluting, we performed the following validation experiment.

In our lab (see Sec. 2.2.2) we obtained spectra for a mix of standards. We then selected 43 signals belonging to 9 different molecules which were easy to manually identify due to the differences in retention times and the m/z values. To simulate coelution we manually aligned signals corresponding to adducts of different metabolites, using as a reference the retention time at the maximum intensity of the signal. (See Fig. 2.2). We then computed Pearson and cosine similarities between all pairs of signals and obtained the Receiver Operating Characteristic (ROC) curve [3] and its area under the curve (AUC) for both methods. The AUC value is the probability that a pair of signals corresponding to adducts of the same metabolite has a larger similarity than a pair of signals corresponding to adducts from different metabolites. Therefore, the larger the AUC value the larger the discriminatory power. We obtained AUC values equal to 0.887 for the cosine similarity and to 0.760 for the Pearson correlation, which show that the cosine similarity has a superior discriminatory power than the Pearson correlation. This high discriminatory power of the cosine similarity comes from the fact that the proportion of signal pairs corresponding to adducts of the same molecule decays rapidly as the value of the similarity decreases.

Therefore, as a first step we construct a weighted undirected similarity network $C^{\mathcal{O}}$ in which each node corresponds to a signal and the weight of each edge between nodes (i, j) corresponds to $c_{ij} = \cos_{ij}$. Note that this is not a fully connected network, because signals that have non-overlapping intensity vectors are not connected.

Step 2: Principled identification of groups of signals (cliques) in the similarity network Our next step is to identify groups of signals that are similar. Specifically, since our hypothesis is that signals with $c_{ij} = 0$ are for sure not adducts of the same metabolite, we aim at identifying cliques of signals in the network, that is groups of signals that are fully connected so that $c_{ij} \neq 0$ for any pair of signals within a clique.

Formally, the task of finding these groups is equivalent to a label assignment problem in which we want to assign a label to each signal σ_i so that signals corresponding to adducts of the same metabolite have the same label.

In order to produce a generative model for node label assignments we note that the cosine similarity between two signals is a good proxy for how likely two signals are to be adducts of the same metabolite. Therefore, a plausible assumption is that the probability of two signals (i, j) having the same label (i.e. belonging to the same clique) given a certain similarity c_{ij} between intensity vectors is precisely a function of that similarity:

$$p(\sigma_i = \sigma_j | c_{ij}) = g(c_{ij}) \tag{2.2}$$

Conversely, the probability that two nodes (i, k) have different labels given their similarity c_{ik} is $p(\sigma_i \neq \sigma_k | c_{ik}) = 1 - p(\sigma_i = \sigma_k | c_{ik})$.

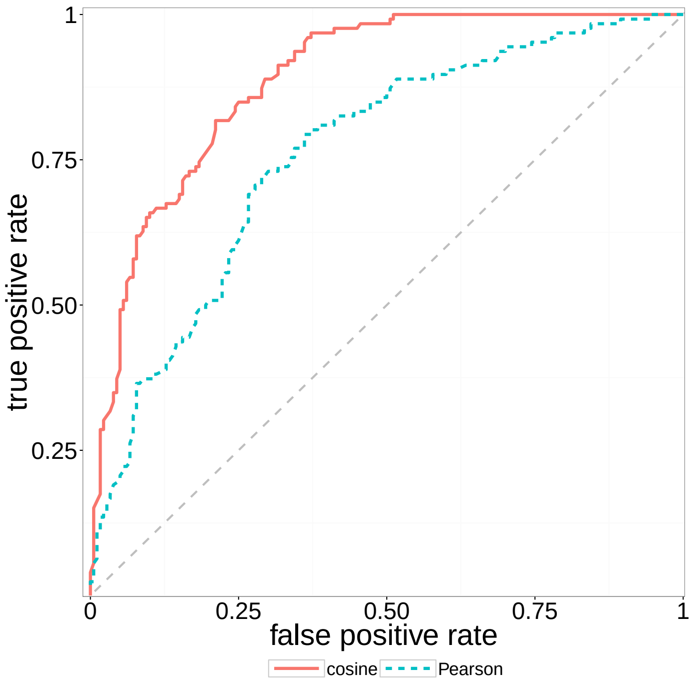


Figure 2.2: We compare the power of Pearson correlation and cosine similarity to distinguish pairs of peaks that are adducts of the same metabolite (true positives) from pairs of peaks belonging to different metabolites (false positives) in simulated coelution. For this purpose we use receiving operating characteristic curves (ROC), where we classify pairs of peaks as "same metabolite" or "different metabolite" if their correlation value is higher than a certain threshold. The threshold spans from 0 (cosine similarity) and -1 (Pearson corrrelation) to 1. Cosine similarity (black, area under the curve (AUC) = 0.887) is a better classifier than Pearson correlation (grey, AUC = 0.760). Total number of random correlations: 180 Total number of real correlations: 126

To specify the precise dependency of $p(\sigma_i = \sigma_j | c_{ij})$ on c_{ij}, we note that $p(\sigma_i = \sigma_j | c_{ij})$ needs to fulfill the following conditions: i) it has to be equal to zero if $c_{ij} = 0$ (that is two nodes whose intensity vectors do not overlap cannot belong to the same group) ; and ii) it has to be equal to one if $c_{ij} = 1$ (that is, signals with proportional intensity vectors have to belong to the same group). Because in our sample $\cos_{ij} \in [0, 1]$, any power of the similarity will satisfy these two conditions. Hence, we assume that

$$p(\sigma_i = \sigma_j | c_{ij}) = c_{ij}^{\alpha} \tag{2.3}$$

Under these assumptions, we can express the probability of an assignment of labels $\boldsymbol{\sigma}$ conditioned on the observed network of similarities $C^{\mathcal{O}}$ as

$$P(\boldsymbol{\sigma} | C^{\mathcal{O}}) = \prod_{\sigma \in \boldsymbol{\sigma}} \prod_{i<j} p(\sigma_i = \sigma_j | c_{ij})^{\delta_{\sigma\sigma_i} \delta_{\sigma\sigma_j}} \\ \times [1 - p(\sigma_i = \sigma_j | c_{ij})]^{(1 - \delta_{\sigma\sigma_i} \delta_{\sigma\sigma_j})} \tag{2.4}$$

where $\delta_{\sigma\sigma_j}$ is the Kronecker delta function. This probability is the so-called likelihood of the model given the data.

Within this probabilistic framework, the most plausible label assignment $\boldsymbol{\sigma}^{\star}$ is the one that maximizes Equation (2.4). In practice, instead of maximizing the likelihood in Equation (2.4), we find the $\boldsymbol{\sigma}^{\star}$ that maximizes the log-likelihood $\mathcal{L} = \log P(\boldsymbol{\sigma} | C^{\mathcal{O}})$. To do so, we use the following algorithm:

1. Start from a configuration in which each node has a different label.
2. Propose a new label assignment.
3. Accept the new label assignment if $\mathcal{L}$ increases.
4. Return to step 2 and iterate until no more changes are accepted.

In step 2, in order to propose a new label assignment we use a combined strategy that alternates between: 1) merging existing cliques; 2) moving nodes from one clique to another clique. To merge existing cliques, we follow the heuristic approach in [4] which is computationally fast. Specifically, we compute the mean-similarity between nodes within each pair of cliques. We then propose to merge the pair of cliques with the largest mean similarity. To move a node (i.e. to change the label of that node to that of a different clique), we select the label assignment that produces the largest increase in $\mathcal{L}$. In our implementation, we propose a node move after ten consecutive attempts at merging pairs of cliques. When $\mathcal{L}$ cannot be increased by merging any pair of cliques in the network, we try to move all nodes of the network from its clique to a different one. At the point where we cannot further increase the log-likelihood with single-node changes, the algorithm stops. For time performance reasons, specially in large samples, we include a parameter l_{min}, that impose a minimum relative change in $\mathcal{L}$ to consider an increase in the log-likelihood.

In order to estimate the best value for the parameter α in Eq. (2.3), we measure the accuracy of our algorithm to correctly assign the same label to signals that have similar retention patterns. Specifically, starting from the spectral data for the mixture of standards (see Sec. 2.2.2), we simulated differences in the coelution of metabolites by manually displacing all the signals of the same metabolite along the retention time axis. We quantify the overlap between signals belonging to different metabolites displaced in this way in terms of the distance between their peaks with the highest intensity, which we call time shifts. We used time shifts raging from 0s (complete coelution) to 4s (little overlap between intensity vectors). As in the previous validation, we consider the spectra of 9 pure compounds within the mixture.

We then simulate the coelution of 2, 3 and 4 compounds at different time shifts, and evaluate the accuracy of our algorithm at correctly labeling signals using the adjusted mutual information [5]. This measures the accuracy of the labeling by comparing the real and the proposed assignment while taking into account the number of signals associated to each metabolite. This value is scaled, so the adjusted mutual information is 0 for any value below the mutual information of a random assignation, which in this case is grouping each signal as a different group.

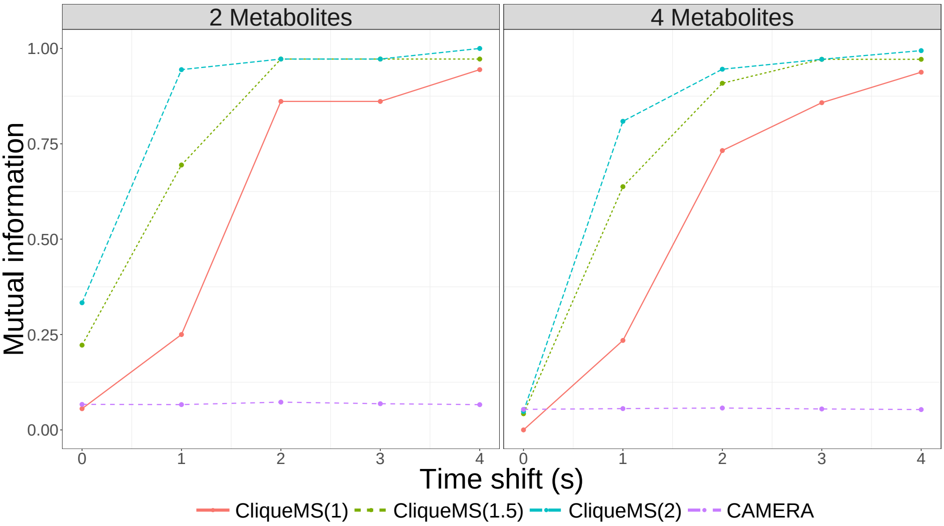


Figure 2.3: Identification of groups of signals with similar coelution patterns. Grouping with CliqueMS is more similar to real assignment than CAMERA grouping algorithm, and it is better when molecules are more separated. CliqueMS was tested with different cosine similarity exponents in Equation (2.3), and overall the best assignment corresponds to $\alpha = 2$.

In Fig. 2.3 we show the accuracy of our algorithm for different values of α : 1, 1.5 and 2. For reference, we also show the results obtained with the signal grouping algorithm in CAMERA. We find that for any choice of α our algorithm outperforms the signal grouping algorithm in CAMERA, the main reason being that the algorithm in CAMERA tends to produce too many groups of signals. We also find that higher values of the exponent lead to more groups of adducts, which improves performance when the time shift decreases. Conversely, when coelution is not as accentuated larger values of α result into too many cliques, slightly decreasing the algorithm's accuracy. Therefore, we use a value of $\alpha = 2$ in our analysis.

Step 3: Annotation of adducts by isotope and neutral mass identification After obtaining the maximum likelihood configuration $\boldsymbol{\sigma}^{\star}$, we use the differences in (m/z) values for all the signals within a clique to identify isotopes and putative adducts associated to the neutral mass of the metabolite.

An exception to this are signals corresponding to isotopic variations of the same metabolite, as they can be determined by the exact mass difference between signals and their relative intensities. Whenever this mass difference between two signals corresponds to 1.003355 $\pm\epsilon_I$, the relative error of the isotope search, the two signals are candidates for being isotopes. If their intensity ratios also correspond to the relative abundance of such isotopes, then these two signals are considered to belong to two isotopic variants of the same metabolite.

Specifically, consider we have a clique γ comprising $\varGamma$ signals $\gamma = \{[(m/z)_i, \boldsymbol{f_i}]; i = 1, \ldots, \varGamma\}$. Once all possible isotopes have been identified (N_I) we are left to provide possible adduct annotations for the remaining $\varGamma' = \varGamma - N_I$ signals within clique γ. The annotation of the isotopes will then follow from the annotation of the precursor ion. In order to do that, CliqueMS considers a list of possible adducts $\{A_i\}^{\circ}$ and their associated mass difference $\{\varDelta M_i\}^{\circ}$ taken from the NIST database [6] for samples with positive and negative annotation (see Supplementary Tables S1 and S2).

First, we determine the possible adduct annotations for each signal that are compatible with the observed mass differences. Specifically, for signal $(m/z)_i$, we obtain all the possible neutral masses M_k that are compatible with signal i being adduct A_k $(A_k \in \{A_i\}^{\circ})$, that is, those that fulfill:

$$\frac{m_i - (M_k + \varDelta M_k)}{M_k} \leq \text{tol}. \tag{2.5}$$

In our analysis we set tol = 10ppm, but this parameter can be tuned by the user. For the remaining signals $(m/z)_j \in \gamma; j \neq i$, we establish that $(m/z)_j$ is compatible with being adduct A_l with neutral mass M_k if:

$$\frac{m_j - (M_k + \varDelta M_l)}{M_k} \leq \text{tol}. \tag{2.6}$$

Following this procedure for all the signals $i \in \gamma$, we obtain for clique γ all possible neutral masses $\{M_k\}^\gamma$ that are compatible with at least two signals being adducts. For each such neutral mass M_k, we construct an adduct vector $\boldsymbol{a^k}$ in which each component a_i^k corresponds to the adduct annotation of signal i compatible with neutral mass M_k. If there is no compatible adduct for signal i then $a_i^k = \texttt{NULL}$.

The second step is to assess the plausibility of each one of these annotations. In order to do this, we note two facts. First, we note that in manual annotation the observation of some adducts such as [M+H]$^+$ or [M+Na]$^+$ is typically considered as a more reliable neutral mass identification than finding adducts [M-H+2Na]$^+$ and [2M+Na]$^+$. The reason for this is that the former couple of adducts are more common than the latter couple of adducts. To formalize this intuition and quantify the plausibility of a specific annotation, CliqueMS uses observed frequencies of adducts in available LC/MS spectra for pure components available in the NIST database (see Supplementary Tables S1 and S2). Specifically, for each M_k the plausibility s_k of annotation $\boldsymbol{a^k}$ is then:

$$s_k = \prod_{i=1}^{\Gamma'} p(a_i^k) \tag{2.7}$$

where $p(x)$ is the frequency of observation of adduct x and $p(\texttt{NULL}) = \epsilon$. In our analysis, we set $\epsilon = 10^{-6}$, so that the frequency of a non-annotated adduct is lower than that of the least common adduct in our database. Note that since available LC/MS spectra are likely to increase in the future, these parameters can be changed by the user as needed.

Second, we note that CliqueMS is based on the expectation that adducts of the same metabolite have similar retention patterns. However, we cannot avoid the fact that in the clique identification procedure we can be grouping together adducts of different metabolites that coelute. Taking this into consideration, CliqueMS allows for the annotation of adducts compatible with more than one neutral mass for signals in the same clique. Therefore, given the set of neutral masses $\{M_k\}^\gamma$ and their associated annotations $\{\boldsymbol{a^k}\}^\gamma$ we can in principle obtain complex annotations $\{\varphi\}^\gamma$ with multiple compatible neutral masses, so that $\varphi_i = a_i^k$ and $\varphi_j = a_i^{k'}$ with k not necessarily equal to k'. These annotations are also subject to the constraint that we have at least two adducts for each neutral mass. Nonetheless, because we expect the number of metabolites in coelution to be low, we assume the plausibility of annotations with a large number of neutral masses N_M to be low. To formalize this idea, the plausibility of such complex annotations s_c is then:

$$s_c = \prod_{i=1}^{\Gamma'} p(\varphi_i) \times \exp\left[-a(N_M - 1)\right] \tag{2.8}$$

where we have introduced an exponential penalty if the number of neutral masses is larger than one and $a = 10$ in our analysis. While this may seem a rather large penalty, we note that the most common adducts have $p(x) \sim \mathcal{O}(10^{-3})$ and rarest adducts have $p(x) \sim \mathcal{O}(10^{-5})$. Therefore, in order to prioritize annotations of a large amount of adducts associated to the same neutral mass over splitting of the annotation into that of two molecules with more common adducts, one needs to introduce exponentially large penalties. On the other hand, the penalty has to be low enough to enable the use of more than one neutral mass when no other annotations are possible. Using a value of $a = 10$ strikes the balance between both undesirable situations.

Unfortunately, the number of potential annotations can grow very fast and it is unfeasible to produce and score all possible annotations. Since we are actually interested in producing a few annotations with the largest plausibilities, we follow a greedy procedure to produce complex annotations. Specifically, we limit the list of neutral masses $\{M_k\}^\gamma$ to include: i) those masses that have the largest overall plausibilities s_k; and ii) consider the top scoring masses for annotating each signal $i \in \gamma$. In our analysis we use M_ks with the 15 top overall s_ks and the most plausible M_k for each signal; these parameter choices show a good compromise between speed of the calculations for large cliques and the retrieval of the most plausible annotations obtained from exhaustive annotation searches.

Finally, we rank annotations $\{\varphi\}^\gamma$ according to their plausibility $\{s_c\}^\gamma$ and produce for each clique the five most plausible annotations.

2.2.2 Spectral data acquisition

We have tested our algorithm with two sets of complex biological samples and a mixture of pure standards. The first set of the complex biological samples come from an immortalized human cell line of a retinal pigment epithelial cell called ARPE-19. These samples were cultured at normoxic and hippoxic conditions and so we called this set “NormHippo”. After removing the cell medium, metabolites were extracted into a extraction solvent by adding 2 mL of a cold mixture of chloroform/methanol (2:1 v/v). The resulting suspension was bath-sonicated for 3 minutes, and 2 mL of cold water was added. Then, 1 mL of chloroform/methanol (2:1 v/v) was added to the samples and bath-sonicated for 3 minutes. Cell lysates were centrifuged (5000g, 15 min at 4 C) and the aqueous phase was carefully transferred into a new tube. The sample was frozen, lyophilized and stored at -80 °C until further analysis. LC/MS analyses were performed using an UHPLC system (1290 series, Agilent Technologies) coupled to a 6550 ESI-QTOF MS (Agilent Technologies) operated in positive (ESI+) electrospray ionization mode. Vials containing extracted metabolites were kept at -20 °C prior to LC/MS analysis. Metabolites were separated using an Acquity UPLC (HSS T3) C18 reverse phase (RP) column (2.1 x 150mm, 1.8 μ) and the solvent system was A1 = 0.1% formic acid in water and B1 = 0.1% formic acid in acetonitrile. The linear gradient elution started at 100% A (time 0–2 min) and finished at 100% B (10-15 min). The injection volume was 5 μL. ESI conditions: gas temperature, 150 °C; drying gas, 13 L min-1; nebulizer, 35 psig; fragmentor, 400 V; and skimmer, 65 V. The instrument was set to acquire over the m/z range 100-1500 in full-scan mode with an acquisition rate of 4 spectra/sec.

The other complex set comes aswell from retina cells, but in this case from transgenic mice retina cells, and has been called “Retina”. The extraction of metabolites begun first with the lyophilization of mouse’s retinas. Metabolites were extracted adding 190 μL of MeOH and 120 muL of H2O, then vortex during 30 seconds. Afterwards, samples were frozen during 1 min in N2 liq. and thawed by cold sonication during 30 seconds. This step was applied three times. Then 380 μL of chloroform were added and vortexed during 30 seconds. Finally, samples were centrifuged (15000 rpm, 15 min a 4°C). The supernatant was extracted and dried. The sample was suspended in 100 μL of H2O:MeOH (1:1) and stored at -80 °C until further analysis. LC/MS analyses were performed in the same equipment than NormHippo samples, but in positive and negative ionization mode (ESI+ and ESI-). Metabolites were separated using the same column and conditions in the positive mode, when the instrument was operated in negative ionization mode, metabolites were separated using an Acquity UPLC (BEH) C18 RP column (2.1 x 150 mm, 1.7 μm) and the solvent system was A2 = 1 mM ammonium fluoride in water and B2 = acetonitrile. ESI conditions and acquisition of the spectra was the same than in NormHippo samples.

In the mix of standards samples, all standards were pulled to a final concentration of 1ppm in H2O:ACN (5:95) with 0.1% formic acid. LC/MS analysis was performed using the same equipment than in the complex biological samples. Metabolites were separated using an Acquity UPLC BEH HILIC column (2.1 x 150 mm, 1.8 μm) and the solvent system was A1 = 20mM ammonium acetate and 15 mM NH4OH in water and B1 = 95% ACN and 5% H2O. Samples were operated in positive electrosprai (ESI+) ionization mode. The linear gradient elution started at 100% B (time 0–2 min) and finished at 75% A (10-15 min). Electrosprai conditions and acquisition of spectra was similar to the complex biological samples.

2.3 Results and discussion

2.3.1 Mixture of standards

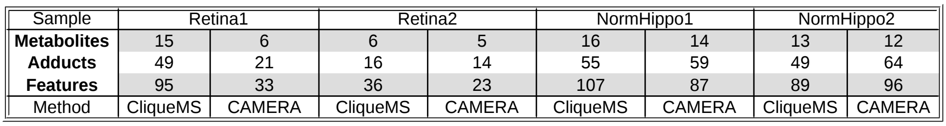
To validate CliqueMS we performed the algorithm in two sets. First we use a mixture of 9 pure known standard metabolites, which allows us to easily analyze the results. We compared the results with CAMERA.

In Fig. 2.5 we can see in colours the peaks belonging to the 9 different metabolites. We correctly annotate all 9 metabolites with CliqueMS. The total number of annotated peaks is 42. This results are better than with CAMERA, that annotates correctly 5 molecules and a total of 30 peaks.

CliqueMS correctly groups peaks belonging to the same metabolite, while CAMERA separates peaks that belong to the same metabolite, like in the case of Uracil or Fructose. Isotope annotation function does not set as isotopes the “wrong” peaks, while CAMERA does. Altough both algorithms are set wit the same error, CliqueMS computes it differently than CAMERA.

2.3.2 Biological samples

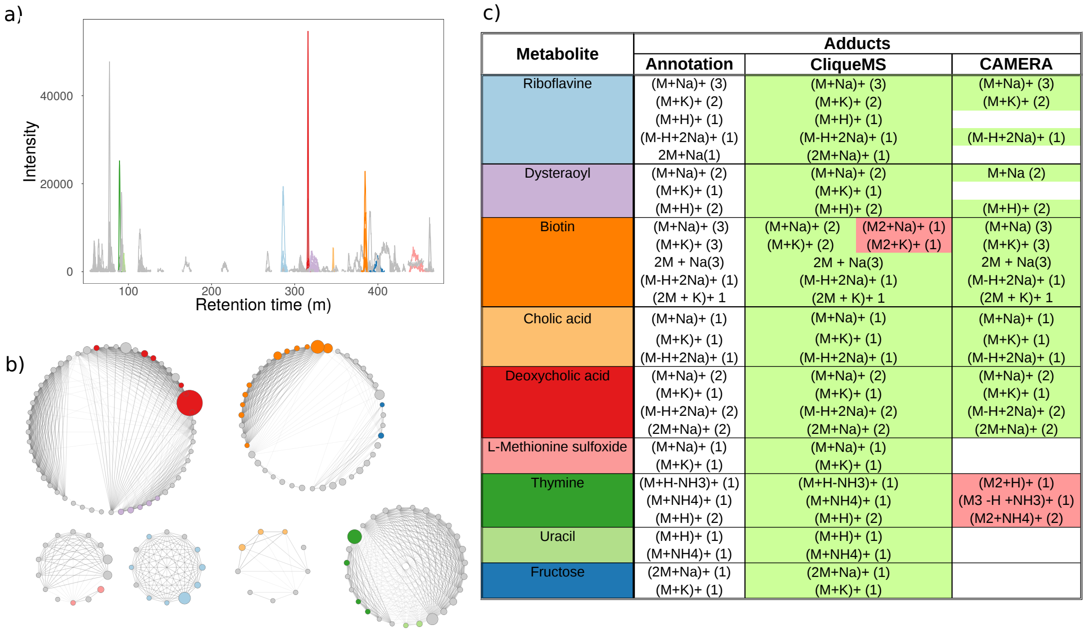
We shrunked the thousands signals of the spectra to hundreds of cliques. Some correctly annotated metabolites appear inside the same clique. This shows both the analytical limit, because many signals from different origin coelute in spite of the chromatography and the clique grouping, which cannot completely separate when many compounds appear together, as we saw in Fig. 2.3 for small differences in retention times. Samples in Retina set have significantly less signals than in NormHippo set, for example Retina1 has 8489 signals and NormHippo1 has 22367, but the number of cliques is not that much different, having Retina1 606 and NormHippo1 707. Both sets of experiments have a similar duration of the cromatography, but for a selected time interval NormHippo samples have, generally, a larger number of signals. As a result cliques in NormHippo have a larger number of signals than in Retina.



Sample	Retina1		Retina2		NormHippo1		NormHippo2	
Metabolites	15	6	6	5	16	14	13	12
Adducts	49	21	16	14	55	59	49	64
Features	95	33	36	23	107	87	89	96
Method	CliqueMS	CAMERA	CliqueMS	CAMERA	CliqueMS	CAMERA	CliqueMS	CAMERA

Table 2.1: Table summarizing results of CliqueMS and CAMERA with complex biological samples.

We have compared highly confident manual annotations with CliqueMS and CAMERA methods, a summary is in Table 2.1. We see that with Clique MS we annotate more metabolites in all samples and sets, compared with CAMERA. The number of total annotated signals is also bigger in



Metabolite	Adducts		
	Annotation	CliqueMS	CAMERA
Riboflavine	(M+Na)+ (3) (M+K)+ (2) (M+H)+ (1) (M-H+2Na)+ (1) 2M+Na(1)	(M+Na)+ (3) (M+K)+ (2) (M+H)+ (1) (M-H+2Na)+ (1) (2M+Na)+ (1)	(M+Na)+ (3) (M+K)+ (2) (M-H+2Na)+ (1)
Dysteraoyl	(M+Na)+ (2) (M+K)+ (1) (M+H)+ (2)	(M+Na)+ (2) (M+K)+ (1) (M+H)+ (2)	M+Na (2) (M+H)+ (2)
Biotin	(M+Na)+ (3) (M+K)+ (3) 2M + Na(3) (M-H+2Na)+ (1) (2M + K)+ 1	(M+Na)+ (2) (M2+Na)+ (1) (M+K)+ (2) (M2+K)+ (1) 2M + Na(3) (M-H+2Na)+ (1) (2M + K)+ 1	(M+Na) (3) (M+K)+ (3) 2M + Na(3) (M-H+2Na)+ (1) (2M + K)+ 1
Cholic acid	(M+Na)+ (1) (M+K)+ (1) (M-H+2Na)+ (1)	(M+Na)+ (1) (M+K)+ (1) (M-H+2Na)+ (1)	(M+Na)+ (1) (M+K)+ (1) (M-H+2Na)+ (1)
Deoxycholic acid	(M+Na)+ (2) (M+K)+ (1) (M-H+2Na)+ (2) (2M+Na)+ (2)	(M+Na)+ (2) (M+K)+ (1) (M-H+2Na)+ (2) (2M+Na)+ (2)	(M+Na)+ (2) (M+K)+ (1) (M-H+2Na)+ (2) (2M+Na)+ (2)
L-Methionine sulfoxide	(M+Na)+ (1) (M+K)+ (1)	(M+Na)+ (1) (M+K)+ (1)	
Thymine	(M+H-NH3)+ (1) (M+NH4)+ (1) (M+H)+ (2)	(M+H-NH3)+ (1) (M+NH4)+ (1) (M+H)+ (2)	(M2+H)+ (1) (M3 -H +NH3)+ (1) (M2+NH4)+ (2)
Uracil	(M+H)+ (1) (M+NH4)+ (1)	(M+H)+ (1) (M+NH4)+ (1)	
Fructose	(2M+Na)+ (1) (M+K)+ (1)	(2M+Na)+ (1) (M+K)+ (1)	

Figure 2.4: a) Extracted Ion Chromatogram (EIC) of standards experiment. The nine ionized molecules were annotated with CliqueMS, in colors we show signals that are adducts of that molecule. b) Network of the same experiment after computing cosine correlation. The intensity of the link increases with the correlation, the size of the nodes increases according to signal intensity. The corresponding colors are the same than the molecules and their respective adducts in the EIC. c) Results of CliqueMS and CAMERA. For each molecule different adducts are annotated, in parenthesis is the total number of isotopic variants of this particular adduct. Correctly annotated adducts are in green, non-annotated signals are in white and wrong annotation in red.

CliqueMS than in CAMERA in all samples except in one NormHippo2. There are some metabolites that were correctly annotated but the correct annotation is not among the top-five scores. This is more likely to happen when cliques are larger in the number of signals, which are the cases that many metabolites are coeluting.

2.4 Conclusions

We have showed that CliqueMS is capable of grouping the multiple signals of a metabolite and then annotate its neutral mass. We have seen that grouping based on a network principle gives better results than other methods, but also it has a limit when metabolites are strongly coeluting, and if that case it tends to group together more than one metabolite.

In all our data, simple and complex experiments, CliqueMS is generally annotating more molecules and more adducts than the most used current method, CAMERA.

Grouping based on cliques reduces the complex spectra to hundreds of groups whose peaks are of great similarity. If more than one metabolite is among those peaks CliqueMS annotation algorithm can annotate both metabolites.

Tolerance is an important parameter for the reported adduct list. More restrictive values will not annotate some correct adducts, but can also improve overall annotation, because wrong annotated adducts appear less, so when scoring correct annotations are more likely to be placed first.

Final anotation outcome depends also on the list of possible adducts. We have used NIST annotated adduct to build our list. We think that altough this list is good for an starting point, it should be combined whenever possible with the observed annotated adducts of previous experiments, to include adducts not observed in NIST, or to change the frequencies of adducts more ocurrent in some equipments.

We think that the increase on the use of CliqueMS and other annotating methods will provide more data on adduct frequencies, which in turn will be a source for improve CliqueMS.

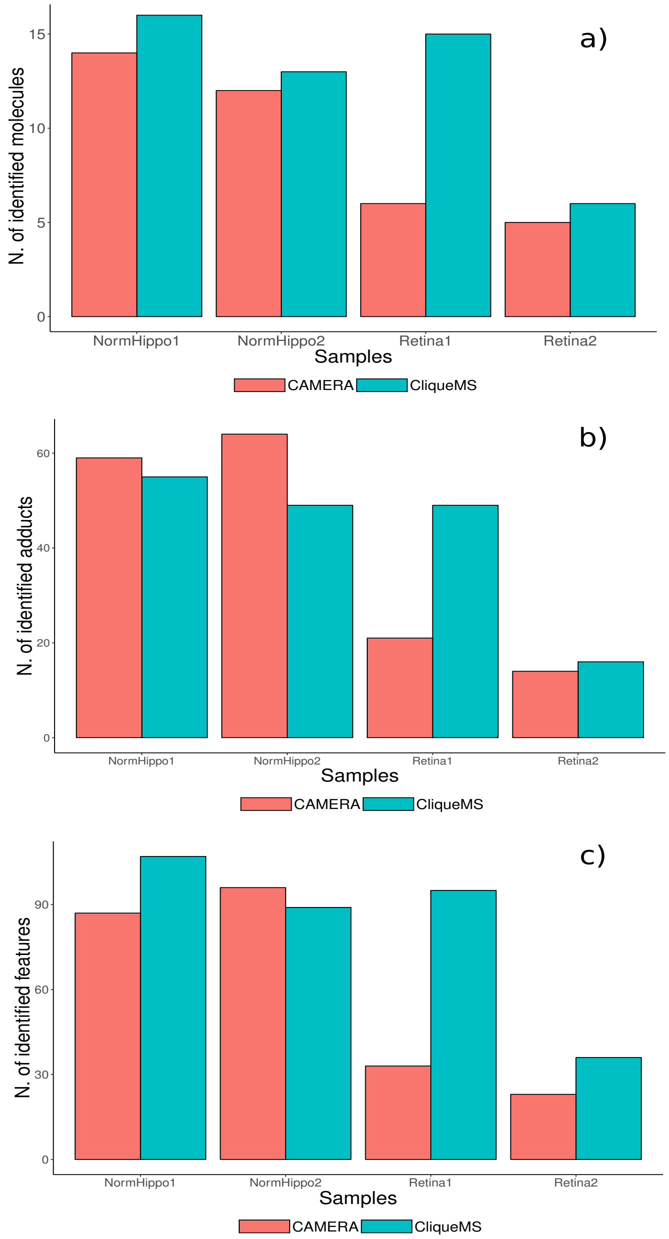


Figure 2.5: Results comparing CliqueMS and CAMERA methods. a) Number of correctly annotated metabolites. b) Number of correctly annotated adducts. c) Number of correctly annotated features (isotopes and adducts).

Bibliography

[1] Karan Uppal, Douglas I. Walker, Ken Liu, Shuzhao Li, Young-Mi Go, and Dean P. Jones. Computational metabolomics: A framework for the million metabolome. *Chem. Res. Toxicol.*, 29(12):1956–1975, 2016.

[2] Carsten Kuhl, Ralf Tautenhahn, Christoph Böttcher, Tony R. Larson, and Steffen Neumann. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.*, 84(1):283–289, 2012.

[3] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

[4] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech. Theor. Exp.*, 2008(10):P10008, 2008.

[5] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, 11:2837–2854, 2010.

[6] National Institute of Standards and Technology. *NIST/EPA/NIH Mass Spectral Library v2014*. US Secretary of Commerce, Gaithersburg, Maryland, USA, 2014.

Chapter 3

The acute impact of polyphenols from Hibiscus sabdariffa in metabolic homeostasis: an approach combining metabolomics and gene-expression analyses

ARTICLE *Food and Function* 6, pp2957-2966, **2015**

Authors
Raúl Beltrán-Debón, Esther Rodríguez-Gallego, Salvador Fernández-Arroyo, Oriol Senan-Campos, Francesco A. Massucci, Anna Hernández-Aguilera, Marta Sales Pardo, Roger Guimerà Manriquem, Jordi Camps, Javier A. Menendez and Jorge Joven

3.1 Introduction

The notion that nutrition is associated with the preservation of health supports the challenging search for bioactive food components [1, 2, 3]. Epidemiological and intervention studies suggest that plant-derived polyphenols are correlated with beneficial health outcomes, which is probably due to their potential action as regulators of the expression of metabolically important genes and/or their intrinsic antioxidant and anti-inflammatory activities [4, 5, 6, 7, 8]. The emergence of unmet global clinical needs (e.g., obesity and the associated conditions) may present an opportunity for the designers of functional foods to provide beneficial products, but some bioactive compounds may be incompatible with consumer acceptance (e.g., due to bitterness or astringency) [9, 10]. The lack of a clear theoretical basis or accepted mechanisms of action also complicates the acceptance of the therapeutic potential of polyphenols. This study is based on the perspective that data obtained from different types of "omics" may be instrumental in tackling the complexity of the mechanisms of action of polyphenols from Hibiscus sabdariffa by integrating the outcomes of multiple effects that occur simultaneously.

It has been argued that polyphenols may act as moderate toxins (i.e., hormesis), which is counterintuitive and contradictory with the fact that polyphenols are apparently nontoxic. The idea of xenohormesis was conceived to indicate that mammals are beneficiaries of phytochemicals because they may respond to "the same chemical cues" developed in plants. [11] To prove or discard this hypothesis is arduous. In addition, because the bioavailability of polyphenols is low, it is difficult to understand how foodstuffs provided in normal amounts could elicit significant effects [12, 13] . We assume that polyphenols are extremely bioactive in humans and/or that the observed effects are the result of multiple beneficial and synergistic interactions. Two strategies can be used to study this aspect and both are under debate: (1) to provide polyphenol-rich extracts, usually in higher doses than those provided in current diets, as products that influence multiple molecular targets, [14, 15] or (2) to manipulate the endogenous antioxidant levels by supplying weak pro-oxidants.[16]. The present study was performed under the rationale that the chemical composition and conformational changes of dietary polyphenols are responsible for binding to different metabolically active enzymes and/or receptors, and consequently may have the inherent potential to exert multiple effects. [17, 18, 19]. This may sound "heretical" to the pharmaceutical industries, which ignore the fact that supposedly selective drugs that are already in the market simultaneously modulate dozens of proteins and receptors [20]. To reduce the complexity found in the composition of plant foods in common diets, which include hundreds of polyphenols, [21] we studied the acute effects of a polyphenol-rich, fully characterized aqueous extract from the calyces of Hibiscus sabdariffa Linnaeus (Malvaceae) (HS) [15]. Our aim was to assess the influence of these polyphenols on the overall metabolic host response using a combination of metabolomics and gene-expression analyses. The result is a useful tool with potential application for monitoring phytochemical exposure in humans, which may be complementary to previous efforts in the quest for nutrition biomarkers, and it may provide support to the more comprehensive concept of foodomics [22, 23, 24].

3.2 Materials and methods

3.2.1 Experimental design

All the experimental procedures were performed in accordance with protocols approved by our Ethics Committee and Institutional Review Board (EPINOLS, 12-03-29/3proj6 and OBESPAD 14-07-31/7proj3). Written informed consent was obtained from the participants prior to their entry into the study. Based on previous results [25, 26], we calculated the sample size using formulas for the 1-sample Z test with a default power of 0.90. Accordingly, to avoid possible gender and age biases, the participants comprised ten healthy male non-obese individuals, non-smokers, free of medication and any metabolic derangements, and with ages ranging from 23 to 35 years. While designing the study, we found that a case-control or case-referent design for liquid ingestion (i.e., using water as a comparator) is unnecessary and is likely to provide confounders. Moreover, the basal or normal metabolic response in a control group could not be matched with the participants, and the previous exposure to different nutrients (diets) was difficult to assess. The use of the same participants on different days might also be a source of confounders, assuming that marked metabolic and hormonal differential changes can occur during the wash-up period. We thus adopted a repeated-group design (before–after study) with data collected in a short (3-hour) period, assuming that a moderate amount of water is metabolically inert and that the fasting state was clearly described and established [27]. The unit of analysis was the pair; another advantage of this design is that each pair serves as its own control, thus reducing the error and increasing the statistical power. The time of the study was limited to a few hours because the assessment of long-term effects would require chronic ingestion and a control group. Time-points for the measurements were also inferred as described in previous studies [13, 14, 15]. The importance of adhering to the fasting recommendations was repeatedly reinforced since the recruitment stage. Participants were asked to avoid strenuous physical activity the day before the experiment and were instructed to avoid ingestion of alcohol and polyphenol-rich foods or beverages (i.e., coffee, tea, juice, oil, chocolate, fruits, and vegetables) during the previous 7 days. Participants remained in the fasting state during the experiment, water was not allowed, and their activity was supervised and restricted. Strict fasting was indicated 12 hours prior to experiment following standardized policies. Clinical measurements or manipulations were avoided to prevent a placebo-like effect [15]. The extract from HS calyces containing 560 mg of polyphenols per 5 g of dried material was prepared by Monteloeder S.L (Elche, Spain) dissolved in water (200 mL) and was immediately ingested to ensure that each participant received 8 mg kg_1 of organic acids and phenolic compounds. Details on the composition of these compounds are provided in Table S1.† Participants were asked to remain recumbent for 10 min prior to the drawing of blood samples, which took place immediately before (08:00 AM) and 3 hours after the ingestion. Serum and plasma were obtained, frozen in aliquots at -80°C within 2 h of collection, and were then stored until batch analysis.

3.2.2 Laboratory measurements

In vitro antioxidant activity of HS extract was measured as previously described [12, 13, 14, 15]. The analytical methods for the separation and identification of phenolic fraction and other soluble compounds of the HS extract have been already described [28, 29]. Total and HDL-cholesterol, glucose, uric acid, bilirubin, triglycerides, cortisol and insulin were measured with standard methods (Boehringer, Mannheim, Germany). Serum aldosterone and renin activity were measured as described [30] . The homeostatic model assessment index (HOMA-IR) was calculated as an estimate of insulin resistance [31]. The ferric reducing ability of serum (FRAP) [32] and measurements of the concentrations of total polyphenols and malondialdehyde (MDA) [33] from the serum were performed essentially as described [34]. Chemokine (C–C motif) ligand 2 (CCL2), interleukin-6 (IL6), interleukin-8 (IL8) and tumor necrosis factor-alpha (TNFα) were measured with ELISA (Invitrogen, Carlsbad, USA). High sensitivity C-reactive protein (CRP) was measured using reagents from Biokit (Barcelona, Spain). None of these measured molecules was found in the extract.

3.2.3 Metabolomic platform

Selected samples were outsourced to Metabolon (Research Triangle Park, Durham, NC, USA), extracted upon arrival and divided into fractions for analysis. The instrument and overall process variability were 5% and 10%, respectively. The chromatographic conditions have been previously described [35, 36]. In brief, the liquid chromatography-mass spectrometry (LC-MS, LC-MS2; separately under positive mode and negative mode) platform was based on a Waters ACQUITY UPLC and a Thermo-Finnigan LTQ mass spectrometer, which consisted of an electrospray ionization source and a linear ion-trap mass analyzer. The samples for the gas chromatography/mass spectrometry (GC/MS) analysis were derivatized and analyzed on a Thermo-Finnigan Trace DSQ fast-scanning single-quadrupole mass spectrometer using electron impact ionization. Metabolites were identified by comparing the ion data, retention time, mass (m/z), and MS or MS/MS spectra with a reference library of chemical standards.

3.2.4 Transcriptomic profiling

Peripheral blood mononuclear cells (PBMCs) were isolated at two time points using BD Vacutainer Cell Preparation Tubes as tentative surrogate cells for the markers of gene expression in other tissues [37]. Cells were lysed and stored at $-80°C$ until RNA isolation using a QIAamp RNA Blood mini kit (QIAgen, Izasa, Barcelona, Spain). The quality was checked by capillary electrophoresis and further purified using sequential DNase digestion and QIAgen RNeasy microcolumns prior to the microarray analysis. RNA samples were sent to the Center of Excellence for Fluorescent Bioanalytics (KFB, Regensburg, Germany). The RNA expression profile was analyzed using a GeneChip® High-Throughput HG-U133, which measured the gene expression of 47[thin space (1/6-em)]000 transcripts and variants, combined with the Perfect Match array to remove possible mismatches (HT HG-U133+ PM 24-array plate, Affymetrix, Santa Clara, CA, USA). Numerical data were obtained using Affymetrix Expression Console 1.1.1 software. Gene expression was first measured using the robust multi-array average methodology, followed by quintile normalization. The quality of the data and sources of the batch effect were assessed using the affyPLM package version 1.34.0 version 2.11 and principal component analysis (PCA). Probe set annotation was downloaded from Affymetrix's website and mapped to 20[thin space (1/6-em)]741 genes. We also measured the expression of selected genes using real-time PCR amplifications with TaqMan primers and probes obtained from validated Assays-on-Demand products (Applied Biosystems, Foster City, CA) on the 7900HT Fast Real-Time PCR system (Applied Biosystems) [13]. When there were redundancies, the greatest average expression across all the samples was chosen to represent each gene [38].

3.2.5 Statistical and functional association analyses

We used the sample size and power calculator from Statistical Solutions (Clearwater, FL, USA) using known μ and θ values for control variables in the population and study sample. The power was set at 0.90 to minimize Type II errors. The before-after design required the analysis of aggregated data and use of the Tukey test to decrease the probability of Type I errors. We subsequently assumed that the variation among experimental data may not be fully captured in pre-treatment predictors but would manifest itself in the outcomes [26].

For the metabolomic analysis, we performed comparisons in the metabolomic profile with Welch's t-tests and/or Wilcoxon's rank sum tests as well as ANOVA for repeated measures. To correct for multiple testing, we used the False Discovery Rate estimated using the q-value as described earlier [39, 40, 41]. To obtain a full reconstruction of human metabolism, we considered all the biochemical reactions in the KEGG database in which humans are known to synthesize the required enzymes (or that happen spontaneously), but only main reactant pairs were considered for reconstruction [42, 43, 44, 45]. Then, we mapped onto the reconstruction of all metabolites in the metabolomic essays with a known KEGG identifier, and then we analyzed the microarray data using different "R" packages (http://cran.r-project.org) from Bioconductor (http://www.bioconductor.org/). After controlling the quality and batch effect of the samples, we assessed the differentially expressed genes using the Limma package 3.14.4 with a linear model to test the effect of HS ingestion on gene expression. Gene-Set Enrichment Analysis (GSEA) was performed using GSEA software version 2.0.10, employing Gene Ontology (GO) as a gene-set database and gene annotation for the GO terms based on the Bioconductor package version 2.8. For GSEA calculation, we used the p-value as the statistic for ranking the gene list or the median p-value for genes with multiple probe sets. We used the GSEA Pre-Ranked tool with the classic scoring scheme, a minimum gene-set size of 15, a maximum gene-set size of 1000 and 1000 permutations. Our significant gene-set list had an FDR of 30%. We further validated the metabolomic and transcriptomic analyses using Ingenuity

Pathway Analysis (IPA; Ingenuity Systems Inc., Redwood City, CA, USA; http://www.Ingenuity.com) to interpret the network functions, canonical signaling pathways and toxicity functions.

3.3 Results and discussion

3.3.1 The composition of the polyphenolic extract

The extract was a complex mixture of bioactive compounds prepared from the calyces of HS harvested in Senegal; the resulting beverage was acidic (pH = 2.8), sweet and resembled the cranberry in flavor. Small amounts of similar extracts are currently used in Western countries to market highly consumed herbal teas ("red or sour teas") [46]. The beverage prepared as described, however, was considered to be of low acceptability by 40 of the participants with bitterness as a common concern. Despite polyphenols being generally accepted as the relevant molecules in the quest for pharmacological action (Table S1†), we also measured the contribution of other compounds, including unknown proteins and/or peptides (2 mg kg_1), soluble fiber (5 mg kg_1), and minute quantities (¡100 μg) of citric, malic, ascorbic and protocatechuic acids. We also found mucilage (not measured) and carbohydrates (3.9 mg kg_1), including arabinose, galactose and glucose. According to the in vitro antioxidant activity of the extract, each participant received the equivalent in Trolox of 140 mg kg_1, measured as FRAP [47]. Utilizing comparisons and values for bioavailability in a rat model, [13] the highest concentration of an individual polyphenol, using our design, should be 0.2 μg mL_1. Employing the above-mentioned methods, polyphenols and/or their metabolites were not detectable. Nevertheless, several compounds (namely, hibiscus acid, quercetin-glucuronide and quercetin-diglucuronide) were detected using a triple quadrupole mass spectrometer, but the values remained under the limit of quantification. Bioavailability or pharmacokinetic experiments were out of the scope of this report, but these preliminary results suggest that a simple optimization of sample concentration and extraction should be used in the design of further studies.

3.3.2 Effects in selected laboratory variables

At the 3-hour time-point, we did not observe any changes in glucose metabolism, but lipid metabolism was affected. We also observed a significant decrease in serum cortisol and aldosterone concentration as well as a trend towards higher values in serum renin activity (Table 3.1) .The antioxidant activity of serum measured as FRAP and MDA concentration remained unchanged. This finding probably reflects the fact that the contribution of polyphenols to serum antioxidants is relatively low ($< 2\%$). The serum concentration of other contributors to the antioxidant activity, such as proteins, ascorbate and tocopherols, did not change significantly (data not shown), which is in concordance with the lack of variation in the major contributor serum uric acid ($> 40\%$) [48]. In contrast, we observed a significant increase in the serum concentration of bilirubin, another relatively minor contributor (3%–4%) of serum antioxidant activity under normal circumstances. Chromatographic values confirmed these data and provided results, which indicated the activation of the heme oxygenase-biliverdin reductase axis (i.e., a simultaneous increase in the concentration of heme and bilirubin and a decrease in biliverdin concentration; Fig. 3.1). We also measured selected variables to check the anti-inflammatory activity, but these variables, with the exception of serum CCL2 concentration, remained unchanged.

3.3.3 Metabolomic changes

It is important to note that metabolic changes were qualitatively similar in all participants (i.e., they followed the same trend—either a decrease or increase), suggesting that the observed results refer to the actions of the compounds in the HS extract. We detected 471 metabolites in untargeted metabolomic analyses, but we found uncertainties in the interpretation of 176 metabolites. The remaining 295 metabolites were positively identified, and a significant number ($n = 107$; 36%) were significantly different between groups. The final assessment was limited to 77 metabolites (25 metabolites increased and 52 metabolites decreased after the ingestion of the HS extract) after discarding marginally significant changes, which were scattered

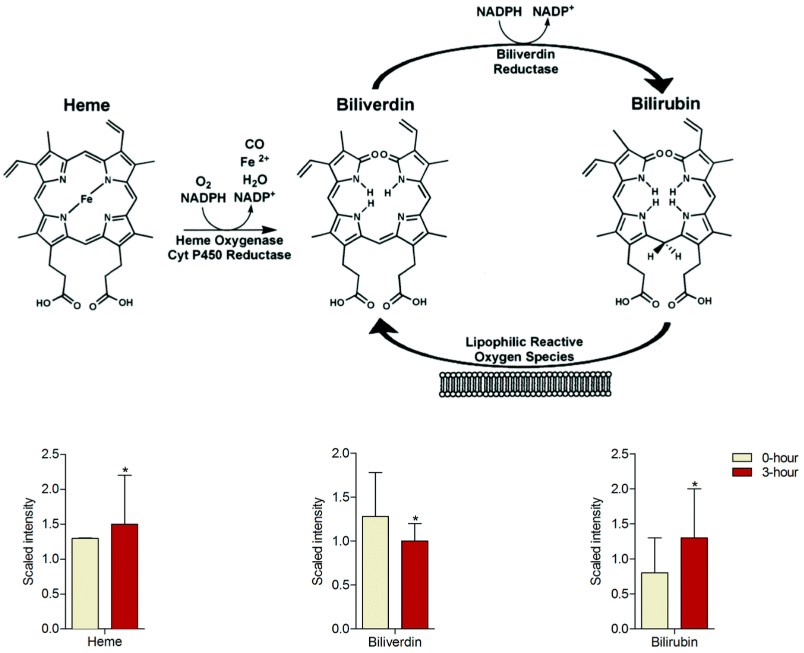
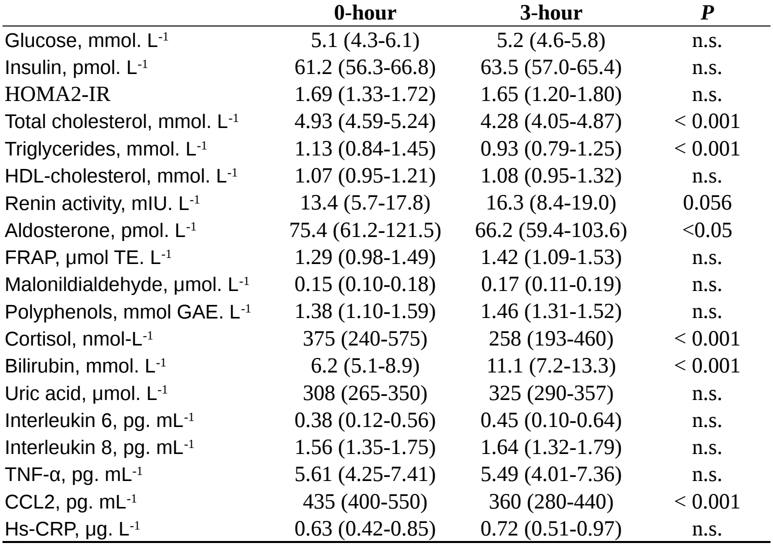


Figure 3.1: The effect of polyphenols in bilirubin metabolism. The ingestion of the HS extract increased the plasma concentration of heme and bilirubin with a significant decrease in biliverdin levels, suggesting activation of the heme-oxygenase-biliverdin reductase axis; *P ¡ 0.05 with respect to the 0-hour time-point.



	0-hour	**3-hour**	***P***
Glucose, mmol. L^{-1}	5.1 (4.3-6.1)	5.2 (4.6-5.8)	n.s.
Insulin, pmol. L^{-1}	61.2 (56.3-66.8)	63.5 (57.0-65.4)	n.s.
HOMA2-IR	1.69 (1.33-1.72)	1.65 (1.20-1.80)	n.s.
Total cholesterol, mmol. L^{-1}	4.93 (4.59-5.24)	4.28 (4.05-4.87)	< 0.001
Triglycerides, mmol. L^{-1}	1.13 (0.84-1.45)	0.93 (0.79-1.25)	< 0.001
HDL-cholesterol, mmol. L^{-1}	1.07 (0.95-1.21)	1.08 (0.95-1.32)	n.s.
Renin activity, mIU. L^{-1}	13.4 (5.7-17.8)	16.3 (8.4-19.0)	0.056
Aldosterone, pmol. L^{-1}	75.4 (61.2-121.5)	66.2 (59.4-103.6)	<0.05
FRAP, µmol TE. L^{-1}	1.29 (0.98-1.49)	1.42 (1.09-1.53)	n.s.
Malonildialdehyde, µmol. L^{-1}	0.15 (0.10-0.18)	0.17 (0.11-0.19)	n.s.
Polyphenols, mmol GAE. L^{-1}	1.38 (1.10-1.59)	1.46 (1.31-1.52)	n.s.
Cortisol, nmol-L^{-1}	375 (240-575)	258 (193-460)	< 0.001
Bilirubin, mmol. L^{-1}	6.2 (5.1-8.9)	11.1 (7.2-13.3)	< 0.001
Uric acid, µmol. L^{-1}	308 (265-350)	325 (290-357)	n.s.
Interleukin 6, pg. mL^{-1}	0.38 (0.12-0.56)	0.45 (0.10-0.64)	n.s.
Interleukin 8, pg. mL^{-1}	1.56 (1.35-1.75)	1.64 (1.32-1.79)	n.s.
TNF-α, pg. mL^{-1}	5.61 (4.25-7.41)	5.49 (4.01-7.36)	n.s.
CCL2, pg. mL^{-1}	435 (400-550)	360 (280-440)	< 0.001
Hs-CRP, µg. L^{-1}	0.63 (0.42-0.85)	0.72 (0.51-0.97)	n.s.

Table 3.1: Selected laboratory variables in plasma from fasting participants used to explore metabolic changes and anti-oxidative or anti-inflammatory effects prior to (0-hour) and 3 hours (3-hour) following consumption of the *Hibiscus sabdariffa extract*

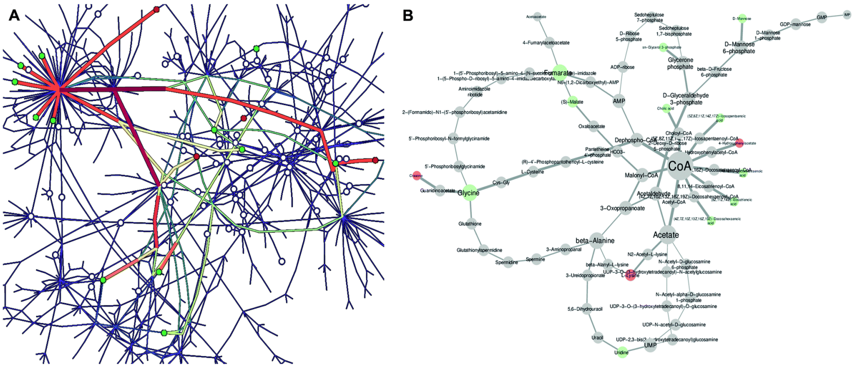

A
B
Acetoacetate
4-Fumarylacetoacetate
Fumarate
1-(5'-Phosphoribosyl)-5-amino-4-(N-succino...)-imidazole
1-(5-Phospho-D-ribosyl)-5-amino-4-imidazolecarboxyla...
N6-(1,2-Dicarboxyethyl)-AMP
Aminoimidazole ribotide
2-(Formamido)-N1-(5'-phosphoribosyl)acetamidine
5'-Phosphoribosyl-N-formylglycinamide
5'-Phosphoribosylglycinamide
Creatine
Guanidinoacetate
Glycine
Cys-Gly
L-Cysteine
(R)-4'-Phosphopantothenoyl-L-cysteine
Pantetheine 4'-phosphate
HCO3-
(S)-Malate
Oxaloacetate
AMP
ADP-ribose
D-Ribose 5-phosphate
Sedoheptulose 7-phosphate
Sedoheptulose 1,7-bisphosphate
sn-Glycerol 3-phosphate
Glycerone phosphate
D-Glyceraldehyde 3-phosphate
beta-D-Fructose 6-phosphate
D-Mannose 6-phosphate
D-Mannose
D-Mannose 1-phosphate
GDP-mannose
GMP
IMP
Cholic acid
Choloyl-CoA
(5Z,8Z,11Z,14Z,17Z)-Icosapentaenoic acid
(5Z,8Z,11Z,14Z,17Z)-Icosapentaenoyl-CoA
4-Hydroxyphenylacetate
Hydroxyphenylacetyl-CoA
Dephospho-CoA
2-Deoxy-D-ribose 5-phosphate
CoA
Malonyl-CoA
Docosatetraenoyl-CoA
8,11,14-Eicosatrienoyl-CoA
(8Z,11Z,14Z)-Icosatrienoic acid
Acetaldehyde
(4Z,7Z,10Z,13Z,16Z,19Z)-Docosahexaenoyl-CoA
Acetyl-CoA
(4Z,7Z,10Z,13Z,16Z,19Z)-Docosahexaenoic acid
3-Oxopropanoate
Acetate
beta-Alanine
3-Aminopropanal
Spermine
Spermidine
Glutathionylspermidine
Glutathione
N2-Acetyl-L-lysine
beta-Alanyl-L-lysine
L-Lysine
3-Ureidopropionate
N-Acetyl-D-glucosamine 6-phosphate
UDP-3-O-(3-hydroxytetradecanoyl)-N-acetylglucosamine
N-Acetyl-alpha-D-glucosamine 1-phosphate
UDP-3-O-(3-hydroxytetradecanoyl)-D-glucosamine
5,6-Dihydrouracil
UDP-N-acetyl-D-glucosamine
Uracil
UDP-2,3-bis(3-hydroxytetradecanoyl)glucosamine
UMP
Uridine

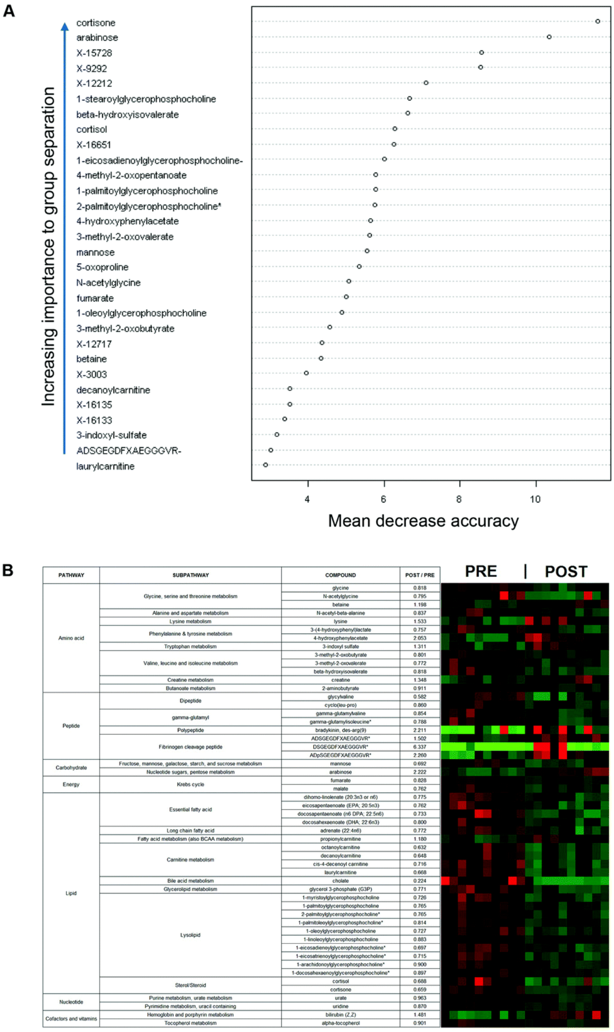
Figure 3.2: Overall representation of metabolic disturbances. The affected pathways are highlighted in the metabolic network connecting altered metabolites (red nodes increased, green nodes decreased, and black nodes remain unchanged) through the shortest possible metabolic routes. The color and width of each reaction (link) represent the number of shortest paths connecting the altered metabolites (A). We inferred centrality in CoA and acetate, most likely acetyl-CoA, but this metabolite was not identified experimentally (B).

across the human metabolic network (Fig. S1†). The perturbed metabolic routes were inferred using the "network parsimony principle", [49], and acetyl-CoA was the most central metabolite in the propagation of the perturbation (Fig. 3.2). The HS extract significantly decreased the concentration of branched-chain amino acids (i.e., isoleucine, leucine and valine) and long-chain fatty acids. The combined effect was a differential production of circulating carnitine conjugates, which suggests that cells take up these compounds to provide energy. Further analyses confirmed significant effects in the canonical pathways of amino acid metabolism (P = 0.00002) and the citric acid cycle (P = 0.000001). In addition, we found a decreased capacity to form triglycerides and an increased capacity for mitochondrial oxidation, indicating an improvement in metabolism and mitochondrial function (Fig. S2 and S3†).

The most representative metabolites that differentiate samples after the ingestion of HS extract were obtained by a random forest analysis (predictive accuracy > 80%) and ranked in order of their importance in the classification scheme (Fig. 3.3 A). The application of LDA, PCA and heat-map graphic representations yielded similar results for group clustering, pattern recognition and the most perturbed pathways and sub-pathways (Fig. 3.3 B and S4†). The HS extract acutely decreased the serum cortisone/cortisol levels (Fig. S5†) associated with changes in the expression of the SGK1 (serum/glucocorticoid regulated kinase 1) gene (Table 3.2). Seemingly, these changes might have multiple beneficial effects on metabolism. The increase in serum arabinose also appears as an important differentiator (Fig. 3.3 A). This is an intriguing finding that illustrates the possible influence of other soluble compounds present in plant-derived extracts. We also found that the HS extract significantly increased the serum concentrations of known products of gut microbiome metabolism such as catechol sulfate, 3-indoxyl sulfates, 3-phenylpropionate and 4-hydroxyphenylacetate. We also noticed a uniform and significant increase in serum concentrations of des-Arg(9)-bradykinin—the active metabolite of bradykinin. This metabolite causes blood vessels to dilate and is one of the substrates of angiotensin I-converting enzyme (ACE). Thus, this observation strongly suggests that HS extracts may act as an ACE inhibitor [15, 50, 51, 52].

3.3.4 Transcriptomic changes

The primary changes in the differentially expressed genes that may demonstrate an overall effect of the HS extract (some are depicted in Table 3.2) illustrate that PBMCs are a source of biological samples that could detect global changes with metabolic, oxidative and inflammatory implications. The GSEA p-value based on a ranked list of genes revealed the relative importance of the biological processes associated with the cellular response to organic substances, the immune system process, the maintenance of protein localization in organelles and the biological regulation of lipid and glucose metabolism (Fig. S6†). Similar studies on molecular functions showed an over-representation of genes related to the activity of cytokine



PATHWAY	SUBPATHWAY	COMPOUND	POST / PRE
Amino acid	Glycine, serine and threonine metabolism	glycine	0.818
		N-acetylglycine	0.795
		betaine	1.198
	Alanine and aspartate metabolism	N-acetyl-beta-alanine	0.837
	Lysine metabolism	lysine	1.533
	Phenylalanine & tyrosine metabolism	3-(4-hydroxyphenyl)lactate	0.757
		4-hydroxyphenylacetate	2.053
	Tryptophan metabolism	3-indoxyl sulfate	1.311
	Valine, leucine and isoleucine metabolism	3-methyl-2-oxobutyrate	0.801
		3-methyl-2-oxovalerate	0.772
		beta-hydroxyisovalerate	0.818
	Creatine metabolism	creatine	1.348
	Butanoate metabolism	2-aminobutyrate	0.911
Peptide	Dipeptide	glycylvaline	0.582
		cyclo(leu-pro)	0.860
	gamma-glutamyl	gamma-glutamylvaline	0.854
		gamma-glutamylisoleucine*	0.788
	Polypeptide	bradykinin, des-arg(9)	2.211
	Fibrinogen cleavage peptide	ADSGEGDFXAEGGGVR*	1.502
		DSGEGDFXAEGGGVR*	6.337
		ADpSGEGDFXAEGGGVR*	2.260
Carbohydrate	Fructose, mannose, galactose, starch, and sucrose metabolism	mannose	0.692
	Nucleotide sugars, pentose metabolism	arabinose	2.222
Energy	Krebs cycle	fumarate	0.828
		malate	0.762
Lipid	Essential fatty acid	dihomo-linolenate (20:3n3 or n6)	0.775
		eicosapentaenoate (EPA; 20:5n3)	0.762
		docosapentaenoate (n6 DPA; 22:5n6)	0.733
		docosahexaenoate (DHA; 22:6n3)	0.800
	Long chain fatty acid	adrenate (22:4n6)	0.772
	Fatty acid metabolism (also BCAA metabolism)	propionylcarnitine	1.180
	Carnitine metabolism	octanoylcarnitine	0.632
		decanoylcarnitine	0.648
		cis-4-decenoyl carnitine	0.716
		laurylcarnitine	0.668
	Bile acid metabolism	cholate	0.224
	Glycerolipid metabolism	glycerol 3-phosphate (G3P)	0.771
	Lysolipid	1-myristoylglycerophosphocholine	0.726
		1-palmitoylglycerophosphocholine	0.765
		2-palmitoylglycerophosphocholine*	0.765
		1-palmitoleoylglycerophosphocholine*	0.814
		1-oleoylglycerophosphocholine	0.727
		1-linoleoylglycerophosphocholine	0.883
		1-eicosadienoylglycerophosphocholine*	0.697
		1-eicosatrienoylglycerophosphocholine*	0.715
		1-arachidonoylglycerophosphocholine*	0.900
		1-docosahexaenoylglycerophosphocholine*	0.897
	Sterol/Steroid	cortisol	0.688
		cortisone	0.659
Nucleotide	Purine metabolism, urate metabolism	urate	0.963
	Pyrimidine metabolism, uracil containing	uridine	0.870
Cofactors and vitamins	Hemoglobin and porphyrin metabolism	bilirubin (Z,Z)	1.481
	Tocopherol metabolism	alpha-tocopherol	0.901

Figure 3.3: Altered metabolites differ in their relative importance. Random forest analysis (A), a supervised classification technique, distinguishes between groups based on their metabolic profiles with a predictive accuracy of > 80% and produces a list of primary differentiators. Heat map as a graphical representation of data (B), where the individual values contained in the metabolic profiles matrix are represented as colors. The red or green colors indicate increased or decreased plasma concentration, respectively. The represented metabolites were selected according to their relative importance to depict the fact that the actions of the HS extract are scattered across a significant number of metabolic pathways.

Symbol	***p*** **value**	**FDR**
CCL3L3	0.001	0.24
CEP152	0.016	0.14
CX3CR1	1.76x10^{-4}	0.2
CXCL10	2.33x10^{-4}	0.21
CXCL8	3.73x10^{-4}	0.22
CYP2R1	4.67x10^{-4}	0.23
EIF1	4.52x10^{-4}	0.23
EIF5	1.28x10^{-4}	0.2
ERN1	1.22x10^{-4}	0.2
FKBP5	2.42x10^{-4}	0.21
HNRNPDL	0.001	0.06
IFRD1	1.16x10^{-4}	0.2
MGAT4A	0.064	0.19
MIB2	2.64x10^{-6}	0.07
NID1	6.82x10^{-5}	0.19
PCMTD1	0.001	0.24
PMAIP1	1.06x10^{-4}	0.2
PPP1R15A	0.001	0.25
PTGS2	1.67x10^{-4}	0.2
SCAF4	1.10x10^{-4}	0.2
SGK1	1.91x10^{-4}	0.2
TAGAP	0.005	0.08
WDR20	0.065	0.19
ZBTB16	2.58x10^{-4}	0.21
ZBTB24	1.89x10^{-4}	0.2

Table 3.2: List in alphabetical order of the top 25 differentially expressed genes that best describe transcriptomic changes after the ingestion of *Hibiscus sabdariffa* extract

and chemokine receptors and ligands. There was a clear and significant association between functions related to the ligand binding to vitamin D and G-protein coupled receptors (Fig. 3.4). The gene ontology numbers and term names, gene size and false discovery rates, as well as the list of common genes significantly involved, may be found in Table S2.† Curiously, the response to biotic stimulus (GO# 0009607) is clearly overexpressed (n = 558 genes) and the major contribution was provided by CXCL8 (interleukin 8), CCL3, CCL2, IL-6 and TNF-α, indicating the anti-inflammatory component of HS.

3.3.5 Inferring the routes of interacting biological macromolecules

PA analysis further confirmed these findings, and the top associated network functions (score ¿40) were gene expression, post-translational modification, cell cycle, molecular transport, RNA trafficking and cellular function and maintenance. The top canonical pathway was glucocorticoid receptor signaling (P ¡ 0.000001). The examination of gene expression, summarizing the differences between the 0-hour and 3-hour time-points, indicated a down-regulation in the genes involved in cholesterol and triglyceride synthesis, lipid transport, gluconeogenesis and glycolysis. Notably, the differential changes in the expression of several genes suggest a possible effect of the HS extract in energy homeostasis via regulatory pathways involving the mechanistic targeting of rapamycin (MTOR) and/or the AMP-activated protein kinase (AMPK) (i.e., the regulation of nutrients and energy sensors [53]). Further confirmation was obtained by analyses of the metabolites and the genetic expression of acetyl-CoA carboxylases (ACC1 and ACC2), CERB-regulated transcriptional coactivator-2 (CRTC2), PPARγ coactivator-1α (Ppargc1α), ribosomal protein S6 kinase (S6K), and eukaryotic initiation factor 4E binding protein 1 (4EBP1). The genes and metabolites with known gene symbols were combined and the results on the main associated network functions and top canonical pathways did not change, except for a higher representation of the incorporation of bile acid-related functions. In this regard, the top up-regulated molecules were des-Arg-(9) bradykinin,

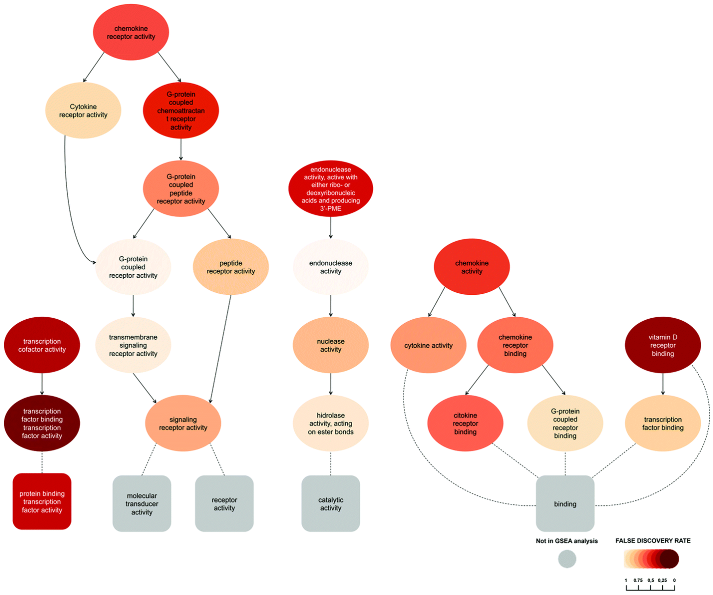

chemokine receptor activity
Cytokine receptor activity
G-protein coupled chemoattractan t receptor activity
G-protein coupled peptide receptor activity
endonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 3'-PME
G-protein coupled receptor activity
peptide receptor activity
endonuclease activity
chemokine activity
transcription cofactor activity
transmembrane signaling receptor activity
nuclease activity
cytokine activity
chemokine receptor binding
vitamin D receptor binding
transcription factor binding transcription factor activity
signaling receptor activity
hidrolase activity, acting on ester bonds
citokine receptor binding
G-protein coupled receptor binding
transcription factor binding
protein binding transcription factor activity
molecular transducer activity
receptor activity
catalytic activity
binding
Not in GSEA analysis
FALSE DISCOVERY RATE
1
0.75
0,5
0,25
0

Figure 3.4: Gene-Set Enrichment Analysis (GSEA) performed using Gene Ontology (GO) as the gene-set database for gene annotation. The figure depicts the significant overrepresentation of molecular function GO terms. Each node corresponds to a distinct molecular function, including gene sets with a low false discovery rate. The color scheme is at the bottom of the figure and the grey nodes correspond to terms without gene representation in the array. Dashed lines indicate missing intermediate terms between the nodes. The expression of genes with functional chemokine activity was significantly associated with the expression of genes, which indicate binding to both G-protein coupled receptors and vitamin D receptors.

bilirubin and CX3CR1, and the top down-regulated molecules were cholic acid, cortisone/cortisol and EGR3. Finally, we found an association between the ingestion of HS extract and a decrease in the depolarization of the mitochondrial membrane (P = 0.0015), mechanisms of gene regulation by peroxisome proliferators (P = 0.009) and p53 signaling (P = 0.002).

3.3.6 Overall discussion

The combination of metabolomic and transcriptomic analyses uncovers complex and multiple metabolic transformations following the ingestion of polyphenols and may help to predict the multiple interactions of food components on metabolic health. Inferring the routes of interacting biological macromolecules may be considered as a promising and complementary tool for capturing the metabolic complexity of phytochemical exposure[7, 15, 22, 23, 24, 54, 55, 56, 57]. Tissue-specific transcriptomic information requires invasive procedures, and whether changes in PBMCs are indicative of the metabolism in other tissues, although suggestive, needs confirmation. We found high serum arabinose concentrations, which may represent a cautionary note because the obvious source was the HS extract. Although human arabinose metabolism is unknown, our data may explain the effects observed in lipid metabolism because, at least in rats, arabinose reduces hepatic lipogenesis and the serum concentration of both cholesterol and triglycerides [50].

Polyphenols are potential antioxidants in vitro, but it has not been unequivocally established that the consumption of polyphenols in humans evokes in vivo antioxidant effects [48]. Although unexpected in a short-term experiment, our data confirm that the antioxidant activity of polyphenols may be partially derived from actions in the digestive tract via the upregulation of the heme-oxygenase (HO)-biliverdin reductase axis. The induction of HO expression explains the antioxidant action of serum bilirubin, contributes to the synergism with PPAR-agonists, and improves insulin sensitivity. Moreover, HO expression may suppress key steps associated with the activation of inflammatory and oxidative pathways [58, 59, 60, 61, 62, 63].

The metabolic effects of the HS extract converge on acetyl-CoA and may improve mitochondrial function via the transport of carbon atoms to the citric acid cycle (i.e., to be oxidized for energy production). In addition, these polyphenols regulate energy sensors (the AMPK/MTOR pathway) and increase the capacity for the oxidation of conjugates derived from branched-chain amino acids and long-chain fatty acids. The metabolism of protein, fat and carbohydrates may be also affected by the HS extract as it reduces the serum concentration of cortisol. This may be a significant finding because we have previously found that the HS extract lowers blood pressure and improves endothelial function in humans, [15] which is in line with the increasingly recognized association between excess cortisol and metabolic syndrome. In particular, the link between HS polyphenols and decreased cortisol may sustain findings that indicate that the combination of oxidation, inflammation and endothelial dysfunction are interrelated mechanisms with a role in the pathogenesis of hypertension [64, 65]. Excess cortisol induces hypertension, [64] and the rapid modifications induced by HS in blood pressure and serum cortisol levels confirm that the expression of the SGK1 gene may be crucial in the transport of sodium [65]. Similarly, we found significant associations between the HS extract and the expression of cytochrome P450, family 2, subfamily R, polypeptide 1 (CYP2R1), which were connected with the expression of genes affecting vitamin D receptor binding. Clinically, high serum levels of vitamin D seem to accompany a reduced risk of high blood pressure but the causality of the association remains to be ascertained [66]. Moreover, evidence presented here are concordant with our previous findings, indicating that the HS extract decreases the activity of the renin-angiotensin (RAS) system in patients with metabolic syndrome and hypertension [15]. The possible action of the HS extract as an ACE inhibitor in vivo might be sustained by the finding of elevated serum concentrations of the vasodilator, des-Arg(9)-bradykinin, and the interrelated effects that result in a decrease of the RAS activity may help to understand the beneficial actions of polyphenols and/or associated compounds from the HS extract. Hypertension, diabetes, obesity and cortisol stimulate RAS activity, and activated RAS is closely related to metabolic syndrome [67, 68, 69, 70, 71]. Conversely, the inhibition of RAS activity improves these disturbances [72, 73]. We also describe that genes acting on the molecular action of G-protein-coupled receptors are differentially expressed by the HS extract. This is important because cytokines, hormones and other active components that cause the deleterious metabolic effects induced by high tissue RAS activity act through G-protein-coupled receptors [74, 75].

3.4 Conclusion

Herein, we propose that polyphenols from HS are a potential source of bioactive compounds that may provide protection for the cardiovascular system. The effects described and those provided by other authors might be used for modeling combinations that are capable of optimizing the view that polyphenols play a pivotal regulatory role in metabolic reprogramming [76, 77, 78, 79, 80]. In addition, investigating multiple metabolic effects and affected pathways should be considered in the assessment of therapeutic strategies.

3.5 Acknowledgements

This study was supported by grants from the Instituto de Salud Carlos III (PI081175, PI081381, PI1102817, and PI1100130), Madrid, Spain; and James S. McDonnell Foundation Research Award, Spanish Ministerio de Economía y Comptetitividad (MINECO) Grant FIS2010-18639 (RG and MSP), European Union Grant PIRG-GA-2010-277166 (RG), European Union Grant PIRG-GA-2010-268342 (to MSP), and European Union FET Grant 317532 (MULTIPLEX).

Bibliography

[1] SM Solon-Biet, McMahon AC, and Le Couter D. G *et al.* The ratio of macronutrients, not caloric intake, dictates cardiometabolic health, aging, and longevity in ad libitum-fed mice. *Cell Metab*, 19:418–30, 2014.

[2] H Ley, Hamdy O., and Hu F.B. Prevention and management of type 2 diabetes: dietary components and nutritional strategies. *The Lancet*, 383(9933):1999–2007, 2014.

[3] Savica V, Bellinghieri G, and Kopple J. The effect of nutrition on blood pressure. *Annual Review of Nutrition*, 30:365–401, 2010.

[4] Corella D and Orovás JM. How does the mediterranean diet promote cardiovascular health? current progress toward molecular mechanisms. *Bioessays*, 36(5):526–37, 2014.

[5] Del Rio D, Rodriguez-Mateos Anna, and Crozier A *et al.* Dietary (poly)phenolics in human health: Structures, bioavailability, and evidence of protective effects against chronic diseases. *Antioxidants and Redox Signaling*, 18(14):1818–92, 2013.

[6] Hopkins A., Lamm M, and Ritenbaugh C. *et al.* Hibiscus sabdariffa l. in the treatment of hypertension and hyperlipidemia: A comprehensive review of animal and human studies. *Antioxidants and Redox Signaling*, 85:84–94, 2013.

[7] Joven J., Micol V., and Menéndez J. *et al.* Polyphenols and the modulation of gene expression pathways: Can we eat our way out of the danger of chronic disease? *Critical Reviews in Food Science and Nutrition*, 54(8), 2014.

[8] Treserra Rimbau A, Rimm E., and Lamuela-Raventós R. *et al.* Polyphenol intake and mortality risk: a re-analysis of the predimed trial. *BMC Medicine*, 12(77), 2014.

[9] Drewnowski A and Gomez-Carneros C. Bitter taste, phytonutrients, and the consumer: a review. *American Journal of Clinical Nutrition*, 72(6):1424–35, 2000.

[10] Lesschaeve I. and Noble A. Polyphenols: Factors influencing their sensory properties and their effects on food and beverage preferences. *American Journal of Clinical Nutrition*, 81(1S):330S–335S, 2000.

[11] Howitz T and Sinclair D. Xenohormesis: Sensing the chemical cues of other species. *Cell*, 133(3):387–91, 2008.

[12] Christelle M., Schlafteier R, and Larondelle Y *et al.* Gene expression changes related to the production of phenolic compounds in potato tubers grown under drought stress. *Phytochemistry*, 70(9):1007–116, 2009.

[13] Fernández-Arroyo S, Herranz-Lopez M, and Micol V *et al.* Bioavailability study of a polyphenol-enriched extract from hibiscus sabdariffa in rats and associated antioxidant status. *Mol Nutr Food Re*, 56(10):1590–5, 2012.

[14] Joven J., Espinell E., and Bertran-Debón R *et al.* Plant-derived polyphenols regulate expression of mirna paralogs mir-103/107 and mir-122 and prevent diet-induced fatty liver disease in hyperlipidemic mice. *Biochim Biophys Acta*, 1820(7):894–9, 2012.

[15] Joven J., March I., and Camps J *et al.* Hibiscus sabdariffa extract lowers blood pressure and improves endothelial function. *Molecular Nutrition and Food Research*, 58(6):1374–8, 2014.

[16] Halliwell B. The antioxidant paradox: less paradoxical now? *British Journal of Clinical Pharmacology*, 75(3):637–644, 2013.

[17] Herranz-Lopez M., Fernández-Arroyo S, and Alonso-Villaverde C *et al.* Synergism of plant-derived polyphenols in adipogenesis: Perspectives and implications. *Phytomedicine*, 19(3-4):253–61, 2012.

[18] Khan N, Farruch A, and Muhtar H *et al.* Synergism of plant-derived polyphenols in adipogenesis: Perspectives and implications. *Cancer Research*, 66(5), 2006.

[19] Hong Fang J, Xue Juan L, and Zhang H*et al.* Natural products and drug discovery. *Embo Reports*, 2009(10):194–200, 2009.

[20] Rull A, Geraeert B, and Camps J*et al.* Rosiglitazone and fenofibrate exacerbate liver steatosis in a mouse model of obesity and hyperlipidemia. a transcriptomic and metabolomic study. *Journal of Proteome*, 13(3):1731–43, 2014.

[21] Bravo L. Polyphenols: Chemistry, dietary sources, metabolism, and nutritional significance. *Nutrition Reviews*, 56(11):317–33, 1998.

[22] Pujos-Guillot E., Hubert J, and Manach C*et al.* Mass spectrometry-based metabolomics for the discovery of biomarkers of fruit and vegetable intake: Citrus fruit as a case study. *Journal of Proteome*, 12(4):1645–59, 2013.

[23] Khymenets O., Lacueva A, and Llorach R *et al.* Metabolic fingerprint after acute and under sustained consumption of a functional beverage based on grape skin extract in healthy human subjects. *Food & Function*, 6(4), 2015.

[24] Valdés A, García-Cañas V, and Cifuentes A *et al.* Comprehensive foodomics study on the mechanisms operating at various molecular levels in cancer cells in response to individual rosemary polyphenols. *Phytomedicine*, 86:9807–9815, 2014.

[25] Bertran-Debon R., Alonso-Villaverde C, and Joven J *et al.* The aqueous extract of hibiscus sabdariffa calices modulates the production of monocyte chemoattractant protein-1 in humans. *Phytomedicine*, 17:186–91, 2010.

[26] Jonhson V. Revised standards for statistical evidence. *PNAS*, 110(48), 2013.

[27] Simunic A., Cornes M, and Nybo M. Standardization of collection requirements for fasting samples: For the working group on preanalytical phase (wg-pa) of the european federation of clinical chemistry and laboratory medicine (eflm). *Clinica Chimica Acta*, 432:33–37, 2014.

[28] Rodriguez-Medina I., Beltrán-Debón R, and Fernández-Gutiérrez A *et al.* Direct characterization of aqueous extract of hibiscus sabdariffa using hplc with diode array detection coupled to esi and ion trap ms. *J of Separation Science*, 32(20):3441–48, 2009.

[29] Singha S and Kumar V. Fabrication and study of lignocellulosic hibiscus sabdariffa fiber reincorced polymer composites. *Bioresources*, 3:1173–1186, 2008.

[30] Espinel E, Joven J, and Serón D *et al.* Risk of hyperkalemia in patients with moderate chronic kidney disease initiating angiotensin converting enzyme inhibitors or angiotensin receptor blockers: a randomized study. *BMC Res Notes*, 6(306), 2013.

[31] Matthews D, Hosker J, and Turner C *et al.* Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia*, 28(7):412–419, 1985.

[32] Cao G and Prior R. Comparison of different analytical methods for assessing total antioxidant capacity of human serum. *Clin Chem*, 44(6 Pt1):1309–1315, 1998.

[33] Esterbauer H, Lang J, and Slater F. Detection of malonaldehyde by high-performance liquid chromatography. *Methods Enzym*, 105:319–28, 1984.

[34] Funes L, Fernández Arroyo S, and Micol V *et al.* Correlation between plasma antioxidant capacity and verbascoside levels in rats after oral administration of lemon verbena extract. *Food Chem*, 117(4):589–598, 2009.

[35] Evans AM., DeHaven CD., and Milgram E. *et al.* Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Anal Chem*, 81(16):6656–67, 2009.

[36] Gall VE. Beebe K and Ferranini E. *et al.* α-hydroxybutyrate is an early biomarker of insulin resistance and glucose intolerance in a nondiabetic population. *Plos One*, 5(5), 2010.

[37] De Mello V., Kolehmanien M., and Uusitupa M. *et al.* Gene expression of peripheral blood mononuclear cells as a tool in dietary intervention studies: What do we know so far? *Mol. Nutr. Food Res*, 56(7):1160–72, 2012.

[38] Irrizarry RA., Hobbs B., and Speed T *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, 2003.

[39] Eisen MB., SpeelMan P., and Botstein D. *et al.* Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–68, 1998.

[40] Saldanha A. Java treeview–extensible visualization of microarray data. *Bioinformatics*, 20(17):3246–48, 2004.

[41] Storey JD. and Tibshirani R. Statistical methods for identifying differentially expressed genes in dna microarrays. *Methods Mol. Biol.*, 224:149–57, 2003.

[42] Kanehisa M. and Goto S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Ac Res.*, 28(1):27–30, 2000.

[43] Kanehisa M., Goto S., and Tanabe M *et al.* Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic Ac Res.*, 42(Database Issue):199–205, 2014.

[44] Guimerà R, , Sales-Pardo M, and Amaral L. A network-based method for target selection in metabolic networks. *Bioinformatics*, 23(13):1616–1622, 2007.

[45] Guimerà R and Amaral L. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[46] Bechof M., Cissé G, and Tomlins I *et al.* Relationships between anthocyanins and other compounds and sensory acceptability of hibiscus drinks. *Food Chem*, 178:112–19, 2014.

[47] Fernández Arroyo S, Rodríguez Medina I, and Fernández-Gutiérrez A. *et al.* Permeability study of polyphenols derived from a phenolic-enriched hibiscus sabdariffa extract by uhplc-esi-uhr-qq-tof-ms. *Food Res Int*, 44(5):1490–5, 2011.

[48] Holman PC., Cassidy B., and Vidry J. *et al.* The biological relevance of direct antioxidant effects of polyphenols for cardiovascular health in humans is not established. *J Nutr*, 141:989–1009, 2011.

[49] Barabási L., Gulnhance N., and Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Gen*, 12(1):56–98, 2011.

[50] Osaki S., Kimura T., and Ritani N. *J Nutr*, 131:796–99, 2001.

[51] Cyr M., Roleau J. L., and Adam A. *et al.* Bradykinin and des-arg9-bradykinin metabolic pathways and kinetics of activation of human plasma. *Am. J. Physiol. Heart Circ. Physiol*, 281(1):275–83, 2001.

[52] Sharma J. N. Hypertension and the bradykinin system. *Curr. Hypertens. Rep.*, 11(3):178–181, 2009.

[53] Inoki K., Kim J., and Guan K. L. Ampk and mtor in cellular energy homeostasis and drug targets. *Annu. Rev. Pharmacol. Toxicol.*, 52:381–400, 2012.

[54] Altmaier E., Suhre K., and Kastenmüller G *et al.* Metabolomics approach reveals effects of antihypertensives and lipid-lowering drugs on the human metabolism. *Eur. J. Epidemiol.*, 29(5):325–36, 2014.

[55] Gottlieb A and Altman R. B. Integrating systems biology sources illuminates drug action. *Clin. Pharmacol. Ther*, 95:663–669, 2014.

[56] O'Grada C. M, Gibney M. J., and Roche H. M. Nutritional aspects of metabolic inflammation in relation to health—insights from transcriptomic biomarkers in pbmc of fatty acids and polyphenols. *Mol. Nutr. Food Res*, 58(8):808–20, 2014.

[57] Rodríguez Gallego E., Arola L., and Joven J. *et al.* Mapping of the circulating metabolome reveals [alpha]-ketoglutarate as a predictor of morbid obesity-associated non-alcoholic fatty liver disease. *Int. J. Obes*, 39:1–9, 2014.

[58] González J., Brito R., and Rodrigo R. Essential hypertension and oxidative stress: New insights. *World J. Cardiol.*, 6(6):353–66, 2014.

[59] Barbagallo I., Gazzolo D., and Li Volti G. *et al.* Potential therapeutic effects of natural heme oxygenase-1 inducers in cardiovascular diseases. *Antioxid. Redox Signaling*, 18:507–21, 2013.

[60] Chang E. C., Liu G. S., and Jiang F. *J. Hypertens*, 32:1379–86, 2014.

[61] Mancuso C., Santagelo R., and Calabrese V. The heme oxygenase/biliverdin reductase system: a potential drug target in alzheimers disease. *J. Biol. Regul. Homeost. Agents*, 27(2 Suppl):75–87, 2013.

[62] Barrajón-Catalán E., Menéndez J. A, and V. Micol *et al.* Molecular promiscuity of plant polyphenols in the management of age-related diseases: far beyond their antioxidant properties. *Adv. Exp. Med. Biol*, 824:141–159, 2014.

[63] Ndisang J. F. Heme oxygenase in cardiac repair and regeneration. *Front. Biosci*, 19:916–35, 2014.

[64] Reinehr T., Welzel B., and Holterhus P. M. *et al.* Relationships between 24-hour urinary free cortisol concentrations and metabolic syndrome in obese children. *J. Clin. Endocrinol. Metab*, 99(7):2391–2399, 2014.

[65] Rao D., Adler K. G., and Williams J. S. *J. Clin. Endocrinol. Metab*, 27:176–80, 2013.

[66] Kunutsor S. K., Munroe P. B, and Khan H. Vitamin d and high blood pressure: causal association or epiphenomenon? *Eur. J. Epidemiol.*, 29(1):1–14, 2014.

[67] Aubert J., Darimont C., and Negrel R. Regulation by glucocorticoids of angiotensinogen gene expression and secretion in adipose cells. *J Biochem*, 328:701–6, 1997.

[68] Kalupahana N. S. and Moustaid-Moussa N. The renin-angiotensin system: a link between obesity, inflammation and insulin resistance. *Obes Rev*, 13(2):136–49, 2012.

[69] Putnam K., Shoemaker R., and Cassis LA. The renin-angiotensin system: a target of and contributor to dyslipidemias, altered glucose homeostasis, and hypertension of the metabolic syndrome. *Am J Physiol Heart Circ Physiol*, 302(6):H1219–30, 2012.

[70] Rull A., Alonso-Villaverde C., and Joven J. Insulin resistance, inflammation, and obesity: Role of monocyte chemoattractant protein-1 (orccl2) in the regulation of metabolism. *Mediators Inflamm*, page pii326580, 2010.

[71] Singh VP., Baker KM., and Kumar R. High-glucose-induced regulation of intracellular ang ii synthesis and nuclear redistribution in cardiac myocytes. *Am J Physiol Heart Circ Physiol*, 293(2):H939–48, 2007.

[72] Argwal D., Keller JN., and Francis J. Chronic exercise modulates ras components and improves balance between pro- and anti-inflammatory cytokines in the brain of shr. *Basic Res Cardiol*, 106(6):1069–85, 2011.

[73] Vaidya A. and Williams J. The relationship between vitamin d and the renin-angiotensin system in the pathophysiology of hypertension, kidney disease, and diabetes. *Metabolism*, 61(4):450–58, 2012.

[74] Kloet A., Krause E. G., and Woods S. C. The renin angiotensin system and the metabolic syndrome. *Physiology & behavior*, 100(5):525–534, 2010.

[75] Skov J., Persson F., and Christiansen JS. Tissue renin-angiotensin systems: a unifying hypothesis of metabolic disease. *Front Endocrinol (Lausanne)*, 28:5–23, 2014.

[76] Patarroyo M and Francis G. Effect of angiotensin-converting enzyme inhibitors and angiotensin receptor antagonists in atherosclerosis prevention. *Curr Car Rep*, 14(4):433–42, 2012.

[77] Csermely P., London G., and Nussinov R. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol Ther*, 138(3):333–408, 2013.

[78] Kitano H. A robustness-based approach to systems-oriented drug design. *Nat Rev Drug Discov*, 6(3):202–10, 2007.

[79] Masana L., Joven J., and Turner PR. The mediterranean-type diet: is there a need for further modification? *Am J Clin Nutr*, 53(4):886–9, 1991.

[80] Menéndez J., Alarcón T, and Joven J. *Cell Cycle*, 13:699–709, 2014.

Chapter 4

A comprehensive study on different modelling approaches to predict platelet deposition rates in a perfusion chamber

ARTICLE *Scientific Reports* 5,13606, **2015**

Authors
Oriol Senan Campos
Jordi Pallarés
Roger Guimerà Manrique
Anton Vernet
Antoni Aguilar-Mogas
Gemma Villahur
Lina Badimon
Marta Sales Pardo
Salvatore Cito

4.1 Introduction

Thrombosis is the main responsible for the leading causes of mortality and morbidity worldwide: heart attack and ischemic stroke ([1]). Thrombus formation is an extremely complex pathological process that starts upon platelet interaction with the exposed vascular thrombogenic surface upon atherosclerotic plaque rupture. Concomitantly, tissue factor exposure triggers the activation of the coagulation cascade and thrombin formation further promoting platelet activation and aggregation. Thrombin, in turn, also leads to fibrin formation and thrombus stabilization

Experimental evidence shows that platelet activation and deposition depends on hemodynamic and rheological variables such as shear rate, shear stress ([2]), red blood cell margination ([3, 4]), exposed substrate (subendothelium, collagen, tendon, etc.) and local concentration of activated platelets and pro-thrombotic factors ([5, 6]). Despite the development of several theoretical models that describe the many contributors to thrombus formation and growth ([7]), with special emphasis on the platelet aggregation process ([3, 8, 9, 10, 11, 12]) as well as the spatial and temporal aspects of early stage thrombus dynamics ([13]), the role of each of the aforementioned variables on thrombus formation is still not clear thus hindering the development of comprehensive and computationally fast multiscale models ([14, 15, 16]).

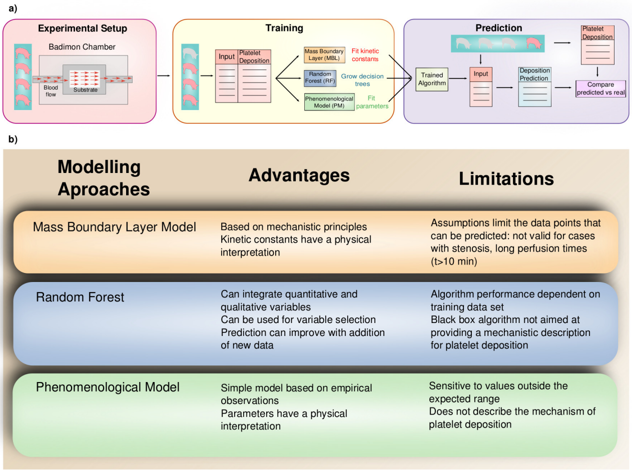
In view of this challenge, and as a first step towards the understanding of the role and limitations of different modelling approaches for thrombus formation, our goal is to compare distinct computationally fast approaches to predict platelet deposition levels. While platelet deposition has been extensively studied, especially within the hemodynamics literature ([17, 18, 19, 20]), very little emphasis has been placed on the assessment of the predictive power of such models. Specifically on the evaluation of whether models adjusted to a set of empirical data (training data set) provide a good description of a different empirical data set (test data set). To a large extent, this is due to the lack of extensive, systematic empirical data on platelet deposition for a wide range of experimental conditions.

To cover this gap, we analyze the ability of different computational approaches to predict platelet deposition values for a large variety of empirical conditions. Note that as a first step, we focus on total platelet deposition counts and do not take into account the spatial dimension of thrombus formation [13]. Specifically, we consider the following approaches: a) a mechanistic modeling approach, b) a machine learning approach; and c) a phenomenological approach. We find that a phenomenological approach built upon empirical facts of the platelet deposition process has the largest predictive power thus offering novel insights into what are the effective roles of different blood factors in platelet deposition.

4.1.1 Approach and rationale

Figure 4.1 illustrates the approach we followed in our study. Specifically, we first collected the platelet deposition data. Then, in order to asses the predictive power of the different computational approaches, we performed a cross-validation analysis. In this type of analysis, we divide the collected data into a training dataset and a test dataset. We use the training dataset to train our model or algorithm (that is to obtain model parameters) so that we obtain a good agreement between model/algorithm outputs and the known empirical platelet deposition value. Then, for each experimental condition in the test dataset, we use the trained model/algorithm to make a prediction of the platelet deposition value. We compare the predicted value with the real value obtained from the experiments to assess the error of the prediction of each approach.

Experimental data collection In our analysis, we consider platelet deposition data of pig blood obtained using a validated ex vivo perfusion chamber (Badimon chamber, [5, 21]), (see Methods). The Badimon chamber provides an excellent proxy for the patho-physiological environment that affects platelet deposition because: i) it is a bio-reactor that retains the cylindrical shape of vascular conduits in which one can simulate a broad range of flow conditions [22, 21]; ii) it is flexible enough to test the thrombogenicity associated with different vascular surfaces or atherosclerotic lesions [23]; and, iii) it allows to analyze different blood conditions and blood



Modelling Aproaches	Advantages	Limitations
Mass Boundary Layer Model	Based on mechanistic principles Kinetic constants have a physical interpretation	Assumptions limit the data points that can be predicted: not valid for cases with stenosis, long perfusion times (t>10 min)
Random Forest	Can integrate quantitative and qualitative variables Can be used for variable selection Prediction can improve with addition of new data	Algorithm performance dependent on training data set Black box algorithm not aimed at providing a mechanistic description for platelet deposition
Phenomenological Model	Simple model based on empirical observations Parameters have a physical interpretation	Sensitive to values outside the expected range Does not describe the mechanism of platelet deposition

Figure 4.1: Flowchart and summary of our approach. a) Flowchart of the analysis. Our study is divided in three steps: i) experimental setup and data collection; ii) training of models/algorithms; iii) prediction. *Experimental setup and data collection:* In the experiments, pig blood circulates from the animal to a perfusion chamber (Badimon Chamber) containing one of the three different vascular tissues considered triggering thrombi (tunica media, pig tendon, subendothelium). We collected platelet deposition counts for different experimental conditions such as perfusion time or shear rate (see Table 4.1 and Methods). We performed experiments with four different animals. *Training:* We consider all the collected input (experimental conditions) and corresponding platelet deposition data for three pigs. With this information we train the models/algorithms to get a good agreement between model/algorithm outputs and known platelet deposition values. *Prediction:* We now consider the data collected for the remaining pig. We use the experimental conditions in that dataset as inputs to the trained model/algorithm to obtain predictions of platelet deposition values for each set of conditions. We test the prediction power of each model/algorithm by comparing predicted platelet deposition values to measured platelet deposition values. We carry out steps ii) and iii) for the four different combinations of training (3 pigs) and test (1 pig) datasets. b) Advantages and limitations of each of the computational approaches for platelet deposition prediction that we consider in our study: a mass-transfer boundary layer model, the Random Forest algorithm and a phenomenological model (see text).

treatments [24, 25]. Specifically, we obtain platelet deposition data for four different pigs under a number of different experimental conditions including variation in shear rate, perfusion time, vascular tissue, hematocrit and platelet concentration levels (see Table 4.1 for a summary of the collected data).

Table 4.1: Experimental data.

Variable	Values (Mean, range)
Shear rate (s^{-1})	212, 1390 and 1690
Perfusion time (min)	3, 5, 10, 20 and 30
Hematocrit (%)	mean: 26.46 (PCV), [22.0, 31.30]
Platelet concentration (platelets/μl) × 10^{-3}	mean: 341.096, [182.0 , 449.0]
Blood	native blood and heparinized blood
Vascular tissue	PT - pig tendon; TM - tunica media; SE - subendothelium
Platelet deposition (platelets/cm^2× 10^{-6})	**Mean: 130.68 [0.63, 2013.5]**

Computational approaches We consider three complementary computationally fast approaches to model platelet deposition (see Figure 4.1 for a summary of the main advantages and limitations of each approach):

(a) A novel mechanistic model based on the mass-transfer boundary layer theory (MBL) ([27]). This is an approach that has been extensively used to investigate hemodynamics and platelet deposition in particular ([8, 10, 28, 29, 30, 31, 32]). This type of models assume that the platelet deposition rate is proportional to a reaction kinetics constant and to the platelet concentration at the wall ([8, 10]). We consider a generalization of a simple model of platelet deposition that includes implicitly the effect of the convective force using boundary-layer theory and as a novelty differentiates between the first monolayer of platelet deposition [platelets in contact with the substrate (e.g. endothelial layer)] and the following multi-layer platelet aggregates [platelet-platelet interaction and thrombus growth] (see Methods and Supplementary Material). As a result, the number of deposited platelets depends on the platelet and hematocrit levels in blood, the vascular lesion dimensions and two kinetic reaction constants that need to be determined: k_1 for the formation of the first monolayer and k_2 for the formation of subsequent layers (see Methods). Note that within our approach deposited platelets cannot detach.

The MBL approach has the advantage that it provides a mechanistic description of the platelet deposition process in which parameters have a clear physical meaning. However, due to MBL assumptions its application is limited to experiments with no stenosis (since the flat plate boundary layer assumptions would be violated) and for short perfusion times (see Methods).

(b) A machine-learning approach using the Random Forest algorithm (RF) ([33]). Methods such as the RF ([33]) are especially suited to predict the outcome (for instance, number of deposited platelets) of an event given the observation of certain features (such as the hematocrit level, shear rate and platelet concentration), without a priori knowledge of the mechanisms governing the specific phenomenon. Indeed, the RF has been successfully applied in a variety of biological contexts such as protein interaction prediction ([34]), gene classification ([35]) and feature selection in biological models ([36]).

Importantly the RF can process both qualitative and quantitative variables, which make it suitable for our analysis in which we have both types of variables (e.g. vascular tissue and blood type are qualitative, while the remaining variables are quantitative — see Table 4.1). However, the predictive power of the RF is severely affected by the range of the training dataset, and will produce very bad predictions for any new input data that falls out of that range.

(c) A phenomenological model (PM) constructed from empirical evidences collected in platelet deposition experiments. We consider a model that takes into account the a priori most relevant features, based on the following observations from the empirical data and from the literature, and further refined with the analysis of variable importance using the RF (see Supporting Figure S1-3):

- Platelet deposition counts increase, in general, with perfusion time and show no apparent signs of saturation in the measured times (see Supporting Figure S1-4);
- Platelets cannot deposit on a surface if there are no platelets circulating in blood;
- Tissue type affects the rate of platelet deposition ([2, 32, 37]);
- The shear rate affects the rate at which platelets deposit on a surface ([8, 21, 38]).

Taking into account these simple facts, we propose the following phenomenological model for the logarithm of the total platelet deposition P under certain experimental conditions:

$$\log_{10} P = \beta_C \log_{10} C + \beta_t \log_{10} t + \beta_\gamma \log_{10} \gamma + \beta(T) \tag{4.1}$$

where P is the platelet accumulation, C is the platelet concentration in blood, t is the perfusion time, γ is the shear rate, $\{\beta_C, \beta_t, \beta_\gamma\}$ are constants, and $\beta(T)$ is a constant that depends on the vascular tissue type (therefore it takes 3 different values).

Our cross validation analysis reveals that the PM has a larger predictive power than MBL and RF approaches: average median errors of 21% (MBL), 20.7% (RF) and 14.2% (PM).

4.2 Results

4.2.1 Model validation

We first assess the validity of the three approaches we consider by fitting the models to all available data points. Figure 4.2 shows that the three approaches we propose - (a) MBL, (b) RF, (c) PM - are, in principle, suited to obtain accurate platelet deposition values under different empirical conditions. The fitting parameters for the MBL model are the kinetic constants of the platelet adhesion process on the substrate (k_1) and on a layer of a previously deposited platelets (k_2). The PM has four fitting parameters: β_C, β_t, β_γ and $\beta(T)$, associated, respectively, to the platelet concentration in blood, the perfusion time, the shear rate and the substrate. The top rows in tables 4.2 and 4.3 show the model parameters estimated for MBL and PM approaches, respectively.

In the MBL approach, we find that platelet deposition counts on tunica media corresponding to a severely damaged vessel wall in which deeper vascular layers are exposed (i.e., vascular smooth muscle cell), does not depend on the values of k_1 and k_2. This suggests that for the experimental conditions under consideration, the deposition on this substrate was limited by the advective and diffusive transport of platelets towards the wall. For the other two substrates (pig tendon and subendothelium), we find that k_1 and k_2 are roughly independent of the tissue and that the values of are k_2 about one order of magnitude larger than k_1. This is consistent with the fact that in the PM (Table 4.3) we obtain the same value for the tissue parameters corresponding to subendothelium and pig tendon and a different value for tunica media.

This observation agrees with the expectation that platelet deposition occurs in a similar manner on both substrates because of their similar constituents. Pig tendons are a rich source of collagen fibers which are precisely one of the main constituents of the basal membrane, the layer that is exposed (but not damaged) in a subendothelial exposure. On the other hand, tunica media encompasses endothelial denudation with damage to both intima and the vascular media exposing to the blood flow not only collagen proteins but vascular smooth muscle cells and their constitutive proteins. Such proteins are highly thrombogenic ([5]) and therefore affect differently the platelet deposition process.

Table 4.2: MBL model parameters. The top row shows the values for k_1 and k_2 obtained considering all the available data for which the model can produce a prediction (no stenosis). The remaining rows show the values obtained for the cross-validation analysis. PT- pig tendon; SE - subendothelium

Test	k_1 (PT)	k_2 (PT)	k_1 (SE)	k_2 (SE)
(1 pig)	(m/s)$\times 10^7$	(m/s)$\times 10^5$	(m/s)$\times 10^7$	(m/s)$\times 10^5$
-	9.5 (0.3)	5.4 (0.4)	8.7 (0.3)	20.0 (0.4)
CP89	10.0 (0.3)	18.0 (0.4)	12.0 (0.3)	13.0 (0.4)
CP90	6.6 (0.3)	5.9 (0.4)	13.0 (0.3)	1.0 (0.4)
CP92	10.0 (0.3)	16.0 (0.4)	12.0 (0.3)	7.5 (0.4)
CP98	6.6 (0.3)	7.2 (0.4)	7.1 (0.3)	9.8 (0.4)

4.2.2 Predictive power assessment

In order to assess the predictive power of each one of the approaches, we performed four cross-validation experiments (Figure 4.1). In each one of these experiments, we consider three pigs as the 'training' data set, and the remaining pig as our 'test' data set. Therefore, we use data from three pigs to estimate the kinetic constants in the MBL approach (see Table 4.1), to train the RF and to estimate the parameters in the PM (see Table 4.2). We then evaluate the error of each of these three approaches in predicting platelet deposition values for the remaining pig. Figure 4.3 shows, as an example, the cross-validation plot for pig CP89.

Our analysis shows that the three approaches we propose produce reasonable predictions of the amount of deposited platelets (Figure 4.4). Note that we can build further confidence in the MBL and PM because model

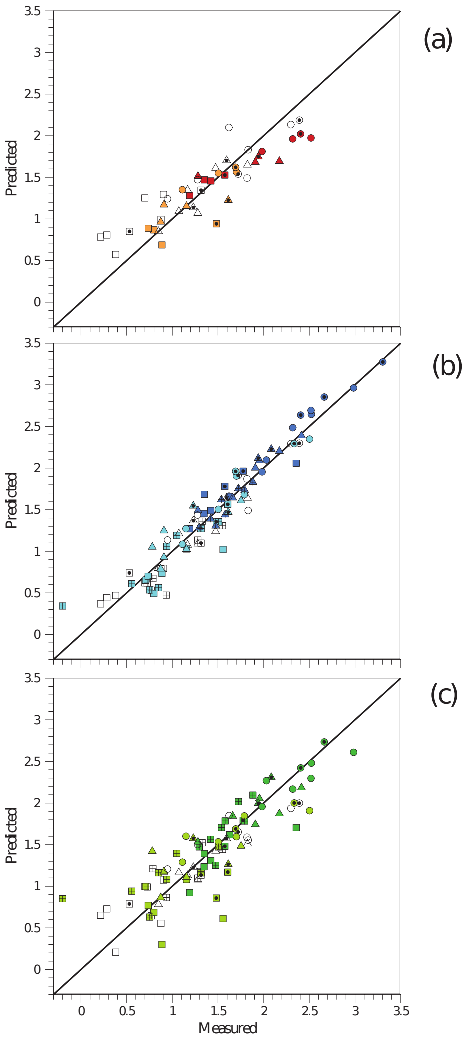


Figure 4.2: Platelet deposition predicted by (a) the mass-boundary layer model (MBL) (b) Random Forest (RF) and (c) the phenomenological model (PM). We show the predictions as $\log_{10}$(number of platelets/cm^2 $\times 10^{-6}$) versus the corresponding experimental values. Open symbols correspond to a perfusion time of 3 minutes, light color symbols to 5 minutes and dark color symbols to 10 minutes. Symbols with a cross represent data of native blood, symbols with dots and without dots correspond to different concentration of heparin (35+35U/K/H and 120+100U/K/H, respectively).

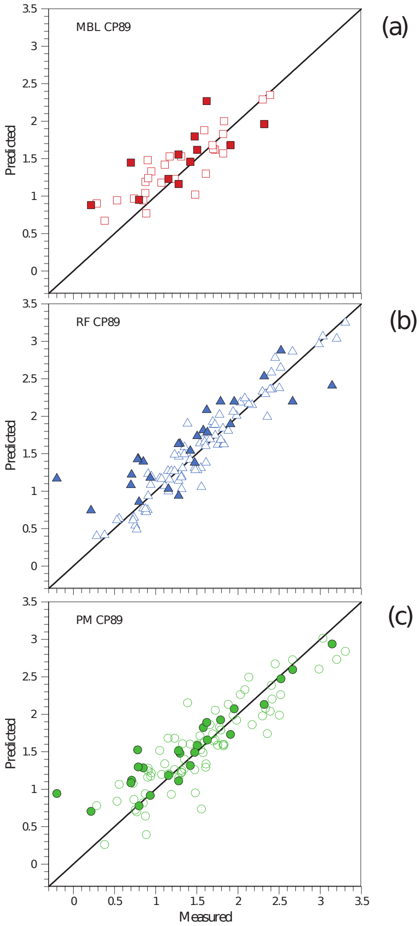


Figure 4.3: Cross-validation plot for pig CP89 showing platelet deposition predicted by (a) the mass-boundary layer model (MBL, red squares), (b) Random forest (RF, blue triangles) and (c) the phenomenological model (PM, green circles). We show model predictions as $\log_{10}$(number of platelets/cm^2 $\times 10^{-6}$) versus the corresponding experimental values for which MBL can produce a prediction (no stenosis). Open symbols correspond to the training set and filled symbols correspond to the test set. Parameters for PM: $\beta(T)$ = -6.3 (PT), -6.3 (SE), -5.8 (TM), β_C = 2.2, β_t = 1.33, β_γ = 0.402.

parameters show little variation (that is, are always in the same orders of magnitude) across the set of cross-validations. We note that in the PM all parameters in Eq (4.1) are significantly different from zero. In addition, in the case of pig tendon and subendothelium, the tissue parameters ($\beta(T)$ in Eq (4.1)) are very similar, confirming that there is little difference in platelet deposition on these two substrates as expected.

Table 4.3: PM parameters. The top row shows the values [value (error))] for β_C (platelet concentration), β_t (perfusion time), β_γ (shear rate) and $\beta(T)$ (tissue) obtained considering all the available data. The remaining rows show the values obtained for the cross-validation analysis considering data for the specified pig as the test set and data for the remaining pigs as the training set. PT- pig tendon; SE - subendothelium, TM - tunica media.

Test	β_C	β_t	β_γ	$\beta(T)$
-	2.2(0.3)	1.4(0.1)	0.38(0.07)	$-6.4(0.8)$ (PT) $-6.7(0.8)$ (SE) $-5.3(0.8)$ (TM)
CP89	2.2(0.3)	1.3(0.1)	0.42(0.08)	$-6.4(0.8)$ (PT) $-6.3(0.8)$ (SE) $-5.8(0.8)$ (TM)
CP90	2.1(0.3)	1.3(0.1)	0.30(0.09)	$-5.7(0.9)$ (PT) $-5.8(0.9)$ (SE) $-5.2(0.9)$ (TM)
CP92	2.6(0.5)	1.7(0.1)	0.40(0.08)	$-8.0(1.0)$ (PT) $-8.0(1.0)$ (SE) $-7.0(1.0)$ (TM)
CP98	2.0(0.4)	1.3(0.1)	0.40(0.09)	$-6.0(1.0)$ (PT) $-6.0(1.0)$ (SE) $-5.0(1.0)$ (TM)

In order to quantify the predictive power of each one of the approaches, we compute the relative error for each one of the cross-validations performed with the three approaches (Figure 4.5 and Supporting Figure S1-5). We note that the median error is typically low, and that the PM is the model that performs best. On average the PM shows relative errors typically about 14.2%, while MBL and RF approaches have median errors of 21% and 20.7%, respectively. This is also the case if we only consider data points for which MBL can produce predictions (that is, experiments with no stenosis), for which the PM has an average median error of 12.9%, while MBL and RF approaches on average have median errors of 22% and 17.2%, respectively.

We also note that in one of the cases (when predicting platelet deposition for pig CP92) we find that the RF and PM approaches have a much lower predictive power. An inspection of the data reveals that this dataset has a narrow range of platelet deposition values – CP92 platelet deposition: (platelets/cm$^2 \times 10^{-6}$) [2.4, 135.3] –, while the rest of data has a wider range – [0.62, 2013.74] (platelets/cm$^2 \times 10^{-6}$) – and that values are lower for CP92 ([182.0, 287.07] (platelets/μl $\times 10^{-3}$) than for the other three pigs (platelet concentration [289.07, 498.89] (platelets/μl $\times 10^{-3}$). Therefore, the loss of predictive power is probably due to the fact that the training data set has 'less' information in the region where CP92 points lie since the training set covers a broader range. This issue highlights the importance of the training set in order to obtain accurate predictions.

4.3 Discussion

Our study showcases the validity of computational approaches to predict platelet deposition in vascular tissues in a number of different conditions. First, we empirically assessed platelet deposition exposing animal blood to a thrombus triggering substrate during different time periods and at different shear rates. Then, we tested the predictive power of three complementary approaches: i) a principle based approach using a mass-transfer model; ii) a machine learning approach that has no information about the physico-chemistry behind the biological process (Random Forest); iii) a phenomenological model constructed from empirical evidence.

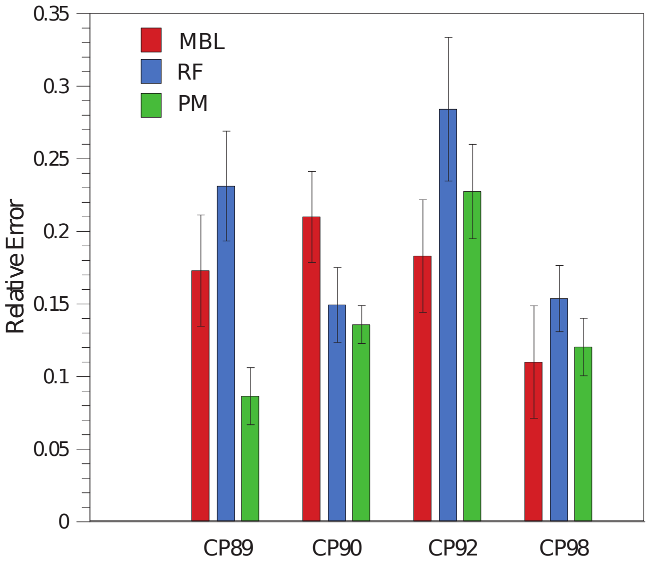


Figure 4.4: Median relative error in the test sets. For each one of the cross-validation analysis we show the median relative error: difference between the predicted and the measured value, relative to the measured value. Error bars correspond to median absolute deviation divided by the square root of observations. For each one of the approaches: MBL– Mass Boundary Layer Model, RF– Random Forest and PM– phenomenological model.

Our study shows that the three approaches have a consistent predictive power, the phenomenological model having an overall better performance. Furthermore, our analysis highlights the main advantages and disadvantages of the different approaches (see Figure. 4.1).

Our analysis also shows that RF and PM approaches would significantly benefit from the availability of platelet deposition data for a larger variety of empirical conditions (for instance, different shear rates and perfusion times). However, this is not necessarily the case for the MBL model. The assumptions made in such model impose certain limitations on the range of applicability of the model. In particular, our MBL approach is not applicable to cases with stenosis or for long times of perfusion when platelet detachment may occur (see for example Supporting Figure S1-4c, where a decrease of deposited platelets is observed for perfusion times between 10 and 30 minutes). The extension of the range of applicability of the MBL model to these cases would require to take into account and parametrize a) the variation of the wall shear rate along the substrate with stenosis and b) the mechanisms responsible for the platelet detachment, thus entailing an increase in the number of fitting parameters.

The availability of a larger variety of empirical conditions would help improve the prediction power of the PM in two aspects. One the one hand, it would yield a more robust set of model parameter values that would give good predictions for a larger range of empirical conditions. On the other hand, new experimental data could help uncover new empirical facts that could be used to refine our model.

Finally, our study shows that the parameter based approaches we propose are biologically sound. Remarkably, our mass-transfer model is a novel model that built upon common approaches in literature that explicitly differentiates between the formation of the first monolayer and that of the subsequent layers. The fact that the kinetic constants associated to each of these mechanisms are different by an order of magnitude indicates that this is an important aspect of the platelet deposition process. In the PM, the fact that all the model parameters are different from zero all the variables we selected have a distinct impact in the platelet deposition process. Additionally, for both approaches we obtain parameter values that are consistent with our expectation of the differences of deposition on different substrates. In particular, in the PM approach tissue dependency is well captured by a single parameter that is similar for pig tendon and subendothelial tissues and different for the tunica media. In contrast, the parameters associated to shear rate, platelet concentration in blood, and perfusion time remain the same throughout the analysis. In fact, according to Table 4.3 the largest contribution is that of platelet concentration in blood and perfusion time, which is also consistent with the assumptions in the MBL model.

All in all, our study opens the door toward further studies that aim to integrate macroscopic description of the models we propose by coupling it to more refined models of the microscopic processes behind platelet deposition.

4.4 Methods

4.4.1 Data description and prediction experiments

Experimental animal model Experiments were performed in Large White x Landrace commercial pigs (n=4, m≈36 kg), individually caged in a light-, temperature-, and humidity-regulated environment with controlled feeding and free access to water. The investigation conforms to the Guide for the Care and Use of Laboratory Animals published by the US National Institute of Health (NIH Publication No. 85-23, revised 1996).

Radioactive labeling of platelets We performed radioactive labeling of platelets to monitor their deposition (monolayer and multilayer). To that purpose, after overnight fasting, 43 ml of pig blood was drawn in 7 ml of anticoagulant citrate dextrose solution by femoral venipuncture. Platelets were isolated and labeled with ^{111}In (Amersham Biosciences, UK) as described in [25] suspended in a final volume of 4 ml of autologous plasma, and reinjected into the pig (ear vein) within 2h. Labeling efficiency was around 90% and the injected activity was around 250mCi. Post-mortem ^{111}In biodistribution indicated a correct platelet distribution with maximal accumulation in blood.

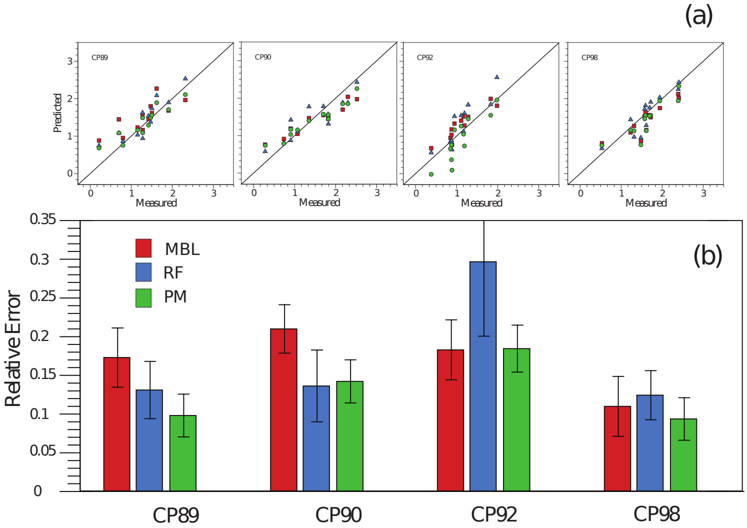


Figure 4.5: (a) Prediction and true value of platelet deposition of test sets. We use the same data points in each test set to directly compare the three modelling methods. (b) Median error of previous cross-validation, relative error: difference between the predicted and the measured value, relative to the measured value. Error bars correspond to median absolute deviation (MAD) divided by the square root of observations. MBL: Mass Boundary Layer (squares) Model, RF = Random Forest (Triangles), PM = Phenomenological Model (circles).

Extracorporeal perfusion system in the Badimon chamber The study protocol was approved by the institutional ethics committee (CSIC-ICCC) and all animal procedures were performed conform the guidelines from Directive 2010/63/EU of the European Parliament on the protection of animals used for scientific purposes or the NIH guidelines. In addition, we have followed the ARRIVE guidelines ([26]). We assessed platelet behavior by exposing the animal blood to a thrombus triggering substrate during different time periods and at different shear rates in the previously validated and standardized Badimon perfusion chamber ([21]). To that end, after overnight fasting, animals were tranquilized (8 mg kg^{-1} Stressnil®, Esteve), anesthetized (10 mg kg^{-1}, B. Braum, Spain), and a carotid artery-jugular vein shunt was established to place the Badimon perfusion chamber as described in ([25]). All of the animals received low-dose anticoagulation with heparin (50 IU kg^{-1}) as a continuous infusion to avoid clotting inside the tubing system. This heparin regime does not affect platelet deposition ([21]).

Blood was perfused through the chamber for different time periods (3, 5, 10, 20 and 30 minutes) at shear rates of 212s^{-1}, 1690s^{-1} and at an experimental stenosis of 80%, that corresponds to a shear value of 1390s^{-1}, in order to mimic the rheological conditions within blood vessels (see the following section for details on the calculation of these values). The thrombogenic substrates (platelet-triggering surfaces) included homologous porcine vessel walls with 2 types of damage [mild (denuded vessel wall or subendothelium SE) and severe (disrupted vessel wall or tunica media TM)] and pig tendon (PT). Several perfusions with varying time of perfusion, hemodynamic conditions and triggering substrate were performed in each animal. After the perfusion, vessels were fixed in 4% paraformaldehyde to count labelled platelets using a gamma counter (Wizard, Wallac, USA). Values were normalized by blood ^{111}In activity (counts), platelet counts in blood, and area exposed surface ([25]). At the end of the experiment, animal's heart was arrested with a 10ml potassium chloride 2M intravenous injection.

Hematological and hemodynamic parameters We determined hematocrit and platelet count throughout the experimental period with as System 9000 Serono cell analyzer.

Overview of the data Table 4.1 provides an overview of the type and range of data collected from the experiments.

For the perfusions performed with 80% of stenosis, we computed the shear rate solving numerically the Navier-Stokes equations in the three dimensional domain that emulate the perfusion chamber with and without the stenosis (see S3 for details).

An analysis of the empirically measured platelet deposition counts reveals that the distribution of the logarithm of the number of deposited platelets has no gaps and is smoother than the distribution of the number of deposited platelets (see Figure. S1-1). For this reason, we focus on predicting the $\log_{10}$ of the number of deposited platelets.

4.4.2 Computational approaches to platelet deposition

Mass-transfer boundary-layer model (MBL) Convection-diffusion-reaction models assume that the platelet deposition rate is proportional to a reaction kinetics constant and to the platelet concentration at the wall ([8, 10, 28, 29, 30, 31, 32, 39, 40, 41, 42, 43]). In here, we consider a generalization of a simple model of platelet deposition that includes implicitly the effect of the convective force using boundary-layer theory and differentiates between the first monolayer of platelet deposition [platelet in contact with the substrate (e.g. endothelial layer)] and the following multi-layer platelet aggregates (platelet-platelet interaction and thrombus growth).

Specifically, in our model we assume two different kinetic reaction constants: k_1 for the formation of the first monolayer and k_2 for the formation of subsequent layers. Therefore, we consider that as the first layer is being covered, with a maximum number of platelets $P_\infty = \frac{4A}{\pi d_p^2}$ where A is the area of the substrate and $d_p = 2 \cdot 10^{-6}$m is the diameter of an adhered platelet ([10]), the second layer starts to form. We model the two adhesion processes with first order kinetics.

In our model, for each one of the layers i we consider, the platelet deposition rate $N_i^{''}$ given certain wall flux of platelets depends on the available deposition area WL_i,

$$\frac{dP_i}{dt} = N_i^{''} W L_i \quad i = 1, 2 \tag{4.2}$$

with $L_1 = \left(1 - \frac{P_1}{P_\infty}\right)$ and $L_2 = \frac{P_1}{P_\infty}$

We assume that the diffusion, advection and reaction processes occur within a two-dimensional mass transfer boundary layer much thinner than the diameter of the perfusion chamber; and that there is a defect of concentration of platelets in comparison with the bulk concentration in the blood (see Supporting Material S2 for a full derivation and for a discussion about the physical interpretation of the equations), the platelet flux on a substrate of length L can be written as ([27]) (see Supporting Material S2),

$$N_i^{''} = \frac{C_0}{\frac{1}{k_i} + 1.238\left(\frac{L_i}{\delta\gamma D^2}\right)^{1/3}} \quad i = 1, 2 \tag{4.3}$$

where C_0 is the bulk concentration of platelets in the blood flow, γ is the shear rate, which is assumed to be constant within the mass transfer boundary layer thickness and D is the diffusion coefficient that depends on the hematocrit concentration ([44]) (see Supporting Material S2).

To numerically determine the kinetic constants using the MBL model, we assume that k_1 depends only on the type of substrate used in the experiments. For each set of experiments with a given substrate, we then compute the time evolution of P_1 and P_2 (see Eqs. S2-10 and S2-11). We then perform the calculations for several values of k_1 and k_2 in the ranges $10^{-3} \leq k_1 \leq 10^{-8}$ m/s and $10^{-3} \leq k_2 \leq 10^{-8}$ m/s. For each pair of values (k_1 , k_2), we then compute the absolute difference between the predicted value of the total number of platelets deposited and the corresponding experimental value at a given time. For each different substrate, we select the pair of values (k_1 , k_2) that minimizes the absolute difference between the measured and predicted values.

Random Forest (RF) We use Random Forest to predict the $\log_{10}$ of the platelet deposition count using four quantitative features and two qualitative features (see Table 4.1). In our analysis, we used the Random Forest Package version 4.6-7 ([45]) within R version 3.0.2 ([46]). We set the algorithm to the following parameters (mtry = $\sqrt{6}$, ntree = 1000). In order to control for the slight variation of each forest due to the bagging process, we performed 100 times each RF. For the estimation of the feature importance, we leaved one feature out of the Random Forest and computed the error rate. Additionally, we applied a linear correction to initial RF predictions to improve the error rate (see Supporting Figure S1-2).

Phenomenological model for platelet deposition (PM) We estimate the parameters by performing a least-squares fit of the data using the R software ([46]).

Bibliography

[1] Nichols, M. *et al.* European cardiovascular disease statistics 2012. http://www.escardio.org/about/documents/eucardiovascular-disease-statistics-2012.pdf (2012). European Heart Network, Brussels;European Society of Cardiology, Sophia Antipolis. Date accesed: 12/10/2014.

[2] Vandrangi, P., Sosa, M., Shyy, J. & Rodgers, V. Flow-dependent mass transfer may trigger endothelial signaling cascades. *PLoS ONE* **7**, e35260 (2012).

[3] Jordan, D., Homer-Vanniasinkam, S., Graham, A. & Walke, R. P. The effects of margination and red cell augmented platelet diffusivity on platelet adhesion in complex flow. *Biorheology* **41**, 641–53 (2004).

[4] Skorczewski, T., Erickson, L. & Fogelson, A. Platelet motion near a vessel wall or thrombus surface in two-dimensional whole blood simulations. *Biophys J* **104**, 1764–72 (2013).

[5] Badimon, L., Padro, T. & Vilahur, G. Extracorporeal assays of thrombosis. *Methods Mol Biol* **788**, 43–57 (2012).

[6] Badimon, L. & Vilahur, G. Thrombosis formation on atherosclerotic lesions and plaque rupture. *J Intern Med* **276**, 618–32 (2014).

[7] Cito, S., Mazzeo, M. & Badimon, L. A review of macroscopic thrombus modeling methods. *Thromb Res* **131**, 116–24 (2013).

[8] Affeld, K., Goubergrits, L., Watanabe, N. & Kertzscher, U. Numerical and experimental evaluation of platelet deposition to collagen coated surface at low shear rates. *J Biomech* **46**, 430–36 (2013).

[9] Kulkarni, S. *et al.* A revised model of platelet aggregation. *J Clin Invest* **105**, 783–91 (2000).

[10] Tokarev, A., Butylin, A. & Ataullakhanov, F. Platelet adhesion from shear blood flow is controlled by near-wall rebounding collisions with erythrocytes. *Biophys J* **100**, 799–808 (2011).

[11] Weller, F. Platelet deposition in non-parallel flow: influence of shear stress and changes in surface reactivity. *J Math Biol* **57**, 333–59 (2008).

[12] Wootton, D., Markou, C., Hanso, N. S. & Ku, D. A mechanistic model of acute platelet accumulation in thrombogenic stenoses. *Ann Biomed Eng* **29**, 321–29 (2001).

[13] Wang, W., Lindsey, J. P., Chen, J., Diacovo, T. G. & King, M. R. Analysis of early thrombus dynamics in a humanized mouse laser injury model. *Biorheology* **51,** 3–14 (2014).

[14] Flamm, M. & Diamond, S. Multiscale systems biology and physics of thrombosis under flow. *Ann Biomed Eng* **40**, 2355–64 (2001).

[15] Tahir, H., Bona-Casas, C. & Hoekstra, A. Modelling the effect of a functional endothelium on the development of in-stent restenosis. *PLoS ONE* **8**, e66138 (2013).

[16] Wang, W. & King, M. Multiscale modeling of platelet adhesion and thrombus growth. *Ann Biomed Eng* **40**, 2345–2354 (2012).

[17] Allender, S. *et al.* European cardiovascular disease statistics 2008. http://hdl.handle.net/10536/DRO/DU:30020501 (2008). European Heart Network, Brussels, England. Date accesed: 12/10/2014.

[18] Eckstein, E. On the simultaneous motions of many blood cells. *Biopl-hys J* **104**, 1839 (2013).

[19] Zhang, G., Zang, S. & B, D. Multiscale particle-based modeling of flowing platelets in blood plasma using dissipative particle dynamics and coarse grained molecular dynamics. *Cell Mol Bioeng* **7**, 552–574 (2014).

[20] Zhang, G., Zang, S. & B, D. A multiple time stepping algorithm for efficient multiscale modeling of platelets flowing in blood plasma. *J Comput Phys* **284**, 668–686 (2015).

[21] Badimon, L. & Badimon, J. Mechanisms of arterial thrombosis in non parallel streamlines: platelet thrombi grow on the apex of stenotic severely injured vessel wall. experimental study in the pig model. *J CLin Invest* **84**, 1134–44 (1989).

[22] Badimon, L., Badimon, J. J., Turitto, V. T. & Fuster, V. Role of von willebrand factor in platelet interaction with an expanded ptfe surface. *ASAIO Trans* **33**, 621–625 (1987).

[23] Fernandez-Ortiz, A. *et al.* Characterization of the relative thrombogenicity of atherosclerotic plaque components: Implications for consequences of plaque rupture. *J Am Coll Cardiol* **23**, 1562–1569 (1994).

[24] Badimon, J. J., Weng, D., Chesebro, J. H., Fuster, V. & Badimon, L. Platelet deposition induced by severely damaged vessel wall is inhibited by a boroarginine synthetic peptide with antithrombin activity. *Thromb Haemost* **71**, 511–516 (1994).

[25] Vilahur, G., Segalés, E., Salas, E. & Badimon, L. Effects of a novel platelet NO-donor (LA816), aspirin, clopidogrel and combined therapy in inhibiting flow and lesion-dependent thrombosis in the porcine ex vivo model. *Circulation* **110**, 1686–93 (2004).

[26] Kilkenny C, Browne WJ, Cuthill IC, Emerson M & Altman DG. Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. *PLoS Biol* **8,** e00412 (2010).

[27] Pallares, J. & Grau, F. Mass transfer rate of a first-order chemical reaction on a wall at high schmidt numbers. *Int J Heat Mass Transfer* **69**, 438–42 (2014).

[28] Tovar-Lopez, F. *et al.* An investigation on platelet transport during thrombus formation at micro-scale stenosis. *PLoS ONE* **8**, e74123 (2013).

[29] Colace, T., Tormoen, G., McCarty, O. & Diamond, S. Microfluidics and coagulation biology. *Ann Rev Biomed Eng* **15**, 283–303 (2013).

[30] Li, M., Hotaling, N., Ku, D. & Forest, C. Microfluidic thrombosis under multiple shear rates and antiplatelet therapy doses. *PLoS ONE* **9**, e82493 (2014).

[31] Badimon, L., Turitto, V., Rosemark, J., Badimon, J. & Fuster, V. Characterization of a tubular flow chamber for studying platelet interaction with biologic and prosthetic materials: Deposition of indium-111 labeled platelets on collagen, subendothelium and expanded polytetrafluoroethylene. *J Lab CLin Med* **110**, 706–18 (1987).

[32] Yamaguchi, T. *et al.* Particle-based methods for multiscale modeling of blood flow in the circulation and in devices: challenges and future directions. *Ann Biomed Eng* **38**, 1225–35 (2010).

[33] Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).

[34] Qi, Y., Bar-Yoseph, P. & Klein-Seetharaman, J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: Structure, Function and Bioinformatics* **63**, 490–500 (2006).

[35] Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using Random Forest. *BMC Bioinformatics* **7**, 3 (2006).

[36] Strobl, C., Boulesteix, A., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25 (2007).

[37] Weiss, H., Turitto, V. & Baumgartner, H. Platelet adhesion and thrombus formation on sub- endothelium in platelets deficient in glycoproteins IIb-IIIa, I b, and storage granules. *Blood* **67**, 322–30 (1986).

[38] Markou, C., Hanson, S. & Ku, D. The role of high wall shear rate on thrombus formation in stenoses. *Adv Biomed Eng* **26**, 555–58 (1993).

[39] Patel, D. *et al.* Dynamics of GPIIb/IIIa-mediated platelet-platelet interactions in platelet adhesion/thrombus formation on collagen in vitro as revealed by videomicroscopy. *Blood* **101**, 929 (2003).

[40] David, T., de Groot, P. & Walker, P. Boundary-Layer type solutions for initial platelet activation and deposition. *J Theor Med* **4**, 95–108 (2002).

[41] Bark, D. & Ku, D. Platelet transport rates and binding kinetics at high shear over a thrombus. *Biophys J* **105**, 502–11 (2013).

[42] Moiseyev, G. & Bar-Yoseph, P. Computational modeling of thrombosis as a tool in the design and optimization of vascular implants. *J Biomech* **46**, 248–52 (2013).

[43] Stubley, G., Strong, A., Hale, W. & Absolom, D. A review of mathematical models for the prediction of blood cell adhesion. *Physicochem Hydrodyn* **8**, 221–35 (1987).

[44] Zydney, A. & Colton, C. Augmented solute transport in the shear flow of a concentrated suspension. *Physicochem Hydrodyn* **10**, 77–96 (1988).

[45] Liaw, A. & Wiener, M. Classification and regression by RandomForest. *R news* **2**, 18–22 (2002).

[46] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013). http://www.R-project.org/. Date accesed: 07/01/2013.

4.5 Author contributions statement

R.G., M.S.P. and S.C. designed the research. G.V. and L.B. performed the experiments. J. P., O. S. and S. C. performed the research. A.A.M. assisted in the research. J. P., O. S., A.V., R. G., A.A.M., M.S.P. and S. C. discussed the results. J. P., O. S., R. G., A.V., A.A.M., G. V., L. B., M.S.P. and S. C. wrote the paper.

4.6 Additional information

The authors declare no competing financial interests.

Chapter 5

Conclusions

Metabolomics is a great technology to study biology and biomedicine. We have seen that it is a very interesting source of data, first because it is the most direct readout of cell activity, and secondly for its relative simplicity compared with proteomics, because metabolites don't have alterations like post-translational modifications, denaturalization, etc... that complicate the interpretation of data. The bottleneck for metabolomics is the lack of automatization in the annotation process. To achieve its full potential we need developments in equipment and in statistical methodology. With CliqueMS, we contribute towards a more automated high-throughput metabolomics, improving the annotation of adducts and isotopic variants. We expect that a better annotation can help us to answer a very fundamental question: How many metabolites are in our samples?.

We observe that our method is able to consistently provide better annotations than existing methods, both in the number of annotated metabolites and in the number of annotated adducts. CliqueMS works sequentially, first creating a network of similarity between features, then grouping those features that belong to the same metabolite, and finally annotating isotopes and adducts, so we can estimate the neutral mass of many metabolites. Nevertheless, annotation step also depends on the list of adducts, which can be provided by the user.

A first application of CliqueMS should be annotating a large group of samples, and study the distribution of adducts across untargeted metabolomics experiments. A better estimation of this distribution will be very useful for annotation, so we expect an improvement in the performance of CliqueMS along its use, as it will have more and more data for the distribution of adducts. Next steps for CliqueMS would be adapting its algorithm to gas chromatography metabolomics (GC/MS), and also to annotate fragmentation adducts, which apart from generation of adducts and isotopes is the other source of multiplicity of signals per metabolite.

Complex metabolomic samples, have thousands of features. With algorithms like CliqueMS we can reduce this number to hundreds of groups, but still many metabolites are coeluting. The limit of statistical methods is the limit of experimental devices. The new single-cell metabolomics, will simplify the complexity of omics data. Firstly, annotation will be easier because we will observe less metabolites in the samples. It also opens new ways to understand the data, as population of cells show an inner variability, that may confuse the interpretation of metabolomics and other omics data.

Reproducibility in reported results using omics data is an important issue. Better experimental designs, more control of variability, like in the case of single cell omics, can provide more consistent findings. Another crucial aspect to get consistent results and to have deeper analysis is the integration of data. We have studied the effects of Hibiscus sabdariffa by combining metabolomic and transcriptomic data. We have reported the first characterization of Hibiscus sabdariffa extracts, that show a therapeutical activity due its polyphenol content. We have seen that it has promising effects for energy metabolism and immune activation. To explore more this findings we would need to investigate the role of polyphenols in metabolic networks. It is of a great importance, for this an other data-driven hypothesis, to preregister confirmatory studies, to avoid spurious interpretations of the data. To think more about general theories and mathematical models that capture this general rules. In this way we will be able to connect distant sources of data and to think in novel experiments.

Finally, what we want is to build theories that connect massive molecular data, like omics data, with phenomena at different scales, including macroscopic phenomena. Thrombus formation is an example of a complex phenomena where we need models to integrate multiple source of data and processes at different scales. Thrombus formation is mainly triggered by platelet deposition. We demonstrate that it is possible to predict platelet deposition from some easily measurable variables, like platelet concentration and the vessel tissue. We expect that a better approximation to thrombus formation will be to integrate the spatial information of platelet deposition, and additionally to include the effect of fibrinogen in our model. Moreover, we have seen that different models, based on equations, machine learning or derived from feature selection of machine learning, can be combined to achieve complementary interpretations and predictions.