Endogenous Metabolites in Drug Discovery: from Plants to Humans

Joaquim Olivés Farrés

TESI DOCTORAL UPF / ANY 2016

DIRECTOR DE LA TESI:

Dr. Jordi Mestres

CEXS Department

The research in this Thesis has been carried out at the Systems Pharmacology Group, within the Research Programme on Biomedical Informatics (GRIB) at the Parc de Recerca Biomèdica de Barcelona (PRBB).

The research presented in this Thesis has been supported by Ministerio de Ciencia e Innovación project BIO2014-54404-R and BIO2011-26669.

Printing funded by the Fundació IMIM's program "Convocatòria d'ajuts 2016 per a la finalització de tesis doctorals de la Fundació IMIM."

Agraïments

Voldria donar les gràcies a tanta gent que em fa por deixar-me ningú. Però per començar haig agrair en especial al meu director la tesi, Jordi Mestres, per donar-me la oportunitat de formar part del seu laboratori i poder desenvolupar aquí el treball que aquí es presenta. A més d'oferir l'ajuda necessària sempre que ha calgut.

També haig de donar les gràcies a tots els companys del grup de Farmacologia de Sistemes que he anat coneguent durants tots aquests anys en què he estat aquí, en especial en Xavi, a qui li he preguntat mil coses, en Nikita, pels sdfs que m'ha anat llençant a CTLink, i la Irene i la Cristina, que els seus treballs també m'ajuden a completar la tesis. I cal agrair també a la resta de companys del laboratori, l'Albert, la Viktoria, la Mari Carmen, l'Andreas, en George, l'Eric i l'Andreu; de Chemotargets, en Ricard i en David; i altres membres del GRIB, com són l'Alfons, en Miguel, en Pau, l'Oriol i la Carina.

També voldria fer una menció al grup de Química Computacional de la UdG, especialment en Miquel Solà i en Jordi Poater, que em van introduir en la bioinformatica, i amb la seva ajuda vaig poder cursar el màster que m'ha portat aquí.

Finalment, i no menys important, és necessari també agrair a la meva familia, especialment als meus pares, per el suport durant tots aquests anys de doctorat, i de vida en general.

I per acabar, ara sí., vuoldria agrair també als companys de carrera, Josep, Pedro, Sergi i Jaco, bé, tu no estaves a la carrera però sempre has estat aquí... I altres que poc tenen a veure amb la biologia, però si han ajudat a fer més amenes les etapes més complicades, la gent del grup de jocs de taula i de Blood Bowl, del Beach Volley Olimpic Team, Late Volley, els Peixets i companys de pis.

Table of contents

Table of contents vii
Abstract ix
Preface xi
Part I: Introduction 1
I.1 History of Pharmacology 5
I.2 History of Ethnopharmacology 15
I.3 Medicinal plants 19
I.4 Parallelism between plants extracts and polypharmacy 23
I.5 Target deconvolution 25
I.6 Polypharmacology 31
I.7 Side effects/Toxicity 33
I.8 Exogenous compounds on human metabolism 38
I.9 Metabolomics 39
Part II: Objectives 45
Part III: Result 49
III.1: Closing the gap between therapeutic use and mode of action in remedial herbs 51
III.2: Unveiling the mode of action of traditional medicines 77
III.3: Exploring polypharmacology of bioactive compounds 115
III.4: Completeness of metabolic databases 127
III.5: Construction of genome-scale metabolic networks of Mycoplasma pneumoniae 141
III.5: Insight on systemic metabolic prediction 169
III.6: Human Metabolome Interference to Drug Polypharmacology 187
III.7: Food Interference to Human Endogenous Metabolome 217
Part IV: Discussion 237

IV.1:Ethnopharmacology 239
IV.2: Polypharmacology in bioactive chemical groups 241
IV.3: Human metabolic networks, current state 243
IV.4: Metabolome reconstruction 244
IV.5: Inerference between human endogenous metabolome and exogenous compounds 245
IV.6 Future directions of research 247
Part V: Conclusions 251
References 257

Abstract

The ability of small molecules to bind to multiple proteins is not exclusive of drugs but general to most chemicals, including endogenous metabolites of living organisms, from plants to humans. In this respect, herbal medicines have been used since the dawn of time for treating discomforts and maladies, but their exact mode of action remains, still nowadays, unknown for most of them. Remedial herbs are composed of hundreds of active compounds interacting between them and with many proteins, forming what we can refer to as a therapeutic cocktail. Some of these interactions are essential for the therapeutic effect of the plant, but some others may be detrimental to their pharmacological action or even cause undesired side effects. Gaining a deeper understanding of the mechanism of action of remedial herbs is one of the main objectives of this Thesis. In addition, recent findings indicate a key role of the endogenous metabolome in drug discovery, a dimension seldom being considered so far. In particular, we are interested in comparing the set of metabolomes currently available as a means to assess the degree of variability among species. Based on this, a second main objective of this Thesis is to develop a computational approach to generate the metabolome from its genome. From plants to humans, accounting for the endogenous metabolome of biological systems emerges as a new paradigm in drug discovery.

Resum

L'habilitat de petites molècules per interaccionar amb múltiples proteines no és exclusiva dels fàrmacs, sino de la majoria de compostos, incloent els metabòlits endogens dels organismes, desde les plants fins els humans. Respecte això, les plantes medicinals han estat utilitzades desde el principi dels temps per tractar malestars i malaties, no obstant el seu mode d'acció roman, encara actualment,

desconegut per la majoria d'ells. Les herbes remeieres estan compostos per centenars de compostos actius interactuant entre ells i amb diverses proteines, creant el que anomenem un cocktail terapèutic. Algunes d'aquestes interaccions son necessàries per produir l'efecte terapèutic, pero d'altres poden ser perjudicials per l'acció farmacològica or fins i tot produir efectes secundaris no desitjats. Aprofundir en el conèixement del mode d'acció de les plantes medicinals és una dels principals objectius d'aquesta Tèsis. A més, estudis recents indiquen un paper clau dels metabolits endogens en la recerca de fàrmacs, una dimensió rarament considerada. En particular, estem interessats en comparar els sets de metabolomes actualment disponible per tal d'assessurar el grau de variabilitat entre espècies. Basant-nos en això, el segon objectiu principal d'aquesta Tèsis és desenvolupar un mètode computacional per generar el metabolome a partir del propi genoma. Desde plantes fins a humans, tenir en compte el metabolome endogen dels sistemes biològics surgeix com un nou paradigme en el desenvolupament de fàrmacs.

Preface

In the last two decades, drug discovery has gradually moved from the traditional "one drug – one target" paradigm to a polypharmacological perception of "one drug – multiple targets". This new view of drug design includes both single drugs acting on multiple targets within a disease pathway and single drugs acting on multiple targets interfering with multiple disease pathways. At the same time, it is observed that the number of approved drugs keep on decreasing in recent years, whereas the interest for natural products in drug discovery increases significantly. One of the advocated advantages of natural products over synthetic molecules is that they often resemble human endogenous metabolites. In addition, natural products tend to offer high structural diversity within the biologically-relevant chemical space.

Many natural products are obtained from traditional medicines as part of complicated mixtures. The study of their effects on complex biological systems has been facilitated by metabolomics. Research on cellular chemical processes and the emergence of systems biology allow now a detailed study of organisms as complex biological system. Also, the development of new computational methods to predict the polypharmacology of small molecules offers a new perspective to predictive toxicology Altogether, we have never been in a better position to investigate at a molecular level the effects of natural products and traditional medicines in the organism, their therapeutics benefits and their potential adverse reactions.

The ultimate aim of this thesis is to investigate the applications and implications of the endogenous metabolome of organisms in drug discovery. It has been divided in five sections. It starts with an historical overview of pharmacology, traditional medicines, and metabolomics. Following this introductory section, the main objectives of this thesis are introduced. It is then followed by a description and discussion of the main results obtained. Finally,

the main conclusions derived from this thesis are outlined, and the document is concluded with a bibliographic section containing a list of cited references

Part I: Introduction

Since early human history, plants have been used in the form of remedies, potions and oils for medicinal purposes (Ji et al., 2009). Normally, people who use traditional medicine don't understand the scientific rationale behind these remedies; but they know from personal experience that some medicinal plants can be effective to treat some diseases. Most of this knowledge has been developed through trial and error over many centuries and has been transmitted through generations. Many of these plants were not selected based on a track record of remedial properties since people didn't have a scientific insight to explain and predict the therapeutic action of plants. In fact, in many cases their use was associated to rather irrational concepts, such as witchcraft and superstition. One example is the assumption that the appearance of plants may give clues to their medicinal properties, for example, yellow flowers are associated with bile and jaundice (Sibley, 2015). Sometimes this assumption worked, like in the case of *Chelidonium majus*, that contains yellow flowers and a yellow alkaloid containing latex has been used successfully to treat jaundice (Gurib-Fakim, 2006).

Figure 1 (left):*Chelidonium majus* (http://www.floracatalana.net)
Figure 2 (right): *Codex vindobonensis* in De Materia Medica, written by Dioscorides.

As mentioned, most of this traditional knowledge passed verbally from one generation to another; however, some written evidences have been identified. The earliest records about the use of traditional medicines are from paintings in clay tablets in cuneiform from Mesotopamia, dating from about 2600 BC. Among the substances documented there are oils of *Cedrus* species (cedar), *Cupressus sempervirens* (cypress), and *Glycyrrhiza glabra* (licorice) among others, some of which are still used today to treat many ailments, from cough and colds to inflammation and parasitic infection. But perhaps the earliest comprehensive document about these practices is the *Ebers Papyrus*, an ancient Egyptian book written in 1500 BC that contains medicinal knowledge from before 3000 BC for about 700 drugs, mostly plants, covering all sorts of illnesses. On the other hand, in China, thousands of herbal formulae have been documented over centuries. The first record dates back from 1100 BC and contains 52 prescriptions. Altogether, more than 100.000 formulae are documented on traditional Chinese medicine. Likewise, documentation of the Indian Ayurvedic medicine system dates back from about 1000 BC. It is also a very ancient medicinal culture, perhaps even older than Chinese medicine (Enrique Raviña Rubira, 2011). Both Ayurvedic and Chinese Medicines are sharing some axioms, for example, that illness is the loss of harmony (Gausachs, 2008).

In contrast, in western civilizations, the Greeks contributed substantially to the rational use of herbal drugs. In the first century AD, the physician Dioscorides wrote the *Materia Medica*, perhaps the most famous book in western medicine (Enrique Raviña Rubira, 2011). During the Dark and Middle Ages, this Western knowledge was preserved in monasteries of England, Ireland, France and Germany, but the Arabs have to be recognized as the culture who preserved much of the Greco-Roman knowledge and expanded it throughout Europe, including their own resources together with chinese and indian herbs unknown

to the Greco-Roman world. The Arabs were also the first to establish drug stores in the VIIIth century (Cragg and Newman, 2005).

It wasn't until XIXth century that scientists isolated various active components from medicinal plants (Goerig and Schulte am Esch, 1991). With the isolation of pure alkaloids and other active principles, as well as the preparation of synthetic pure organic chemicals, it was possible to study and examine their effects with accurately measured quantities. This process led to the establishment of pharmacology as a science (Raviña Rubira, 2011).

I.1 History of Pharmacology

Etymologically, pharmacology is the study of drugs (Greek *pharmakon,* medicine or drug; and *logos*, study). From its original purpose limited to the study of drug action, nowadays pharmacology can be regarded as a broad scientific discipline aiming at studying the changes produced by chemically active substances in living organisms (Brenner and Stevens, 2010).

Birth of pharmacology

The infancy of pharmacology can be traced back to France in the early XIXth century, with the works of François Magendie (1783-1855) and his pupil Claude Bernard (1813-1878). They were both convinced of the importance of using experimental methods and, influenced by the development of organic chemistry, extended their work to studying the physiological effect of certain alkaloids. Magendie was the first physiologist to use alkaloids for the treatment of diseases. Thus, he studied the action of nux vomica (a strychnine-containing plant drug) on dogs and he was able to show that the spinal cord is where its convulsant action occurs. He wrote a *Formulaire* that was extensively used by

doctors in those times (Tubbs et al., 2008). On the other hand, Bernard was interested on the study of curare, an arrow poison from various plants originated in central and south America. He discovered that curare acts at neuromuscular junctions to interrupt the stimulation of the muscle by nerve impulses. Bernard is known as the father of modern experimental medicine (van Bronswijk and Cohen, 2008). Their teachings gave a strong impetus to pharmacology and new scientists were immediately attracted by this discipline, like Rudolph Buchheim and Oswald Schmiedeberg, who are generally recognized as the founders of modern pharmacology (Raviña Rubira, 2011).

Rudolph Buchheim organized the first laboratory on this discipline and published a textbook on pharmacology (*Lerhuch der Arzneimittelehre*). He turned the purely descriptive and empirical study of medicines into an experimental science; however, his reputation is overshadowed by that of his student, Oswald Schmiedeberg (Habermann, 1974).

Schmiedeberg is generally considered as the father of pharmacology. He worked in Dorpat under Buschheim until 1872, when he became professor of pharmacology at University of Strasbourg and created the magnificent Institute of Pharmacology. He studied the pharmacology of chloroform and chloralhydrate. In 1883 he published the first edition of *Grundis der Pharmacheologie* (*Textbook of Pharmacology*), a textbook that soon became a classic. Schmiedeberg had a great influence on the development of pharmacology as a science. In fact, he is considered the most prominent pharmacologist of his time. In 46 years at his Institute, Schmiedeberg trained numerous disciples from about 20 different countries, many of which became then professors of pharmacology in their own countries (Muscholl, 2001).

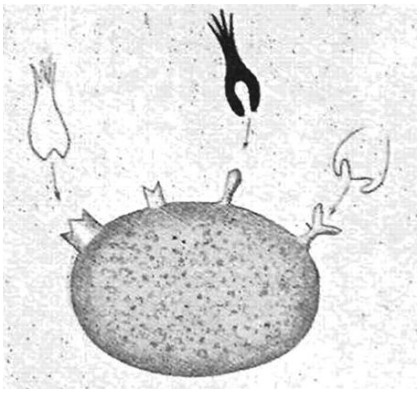


Figure 3: Ehrlich side-chain theory – diagram used to illustrate his lecture to the Royal Society of London in 1900

Drug receptor theory formulation

During the XIXth century, a key theory for pharmacology was proposed by Paul Ehrlich: the side chain theory. He spent many years experimenting with dyes to stain cells. Fascinated by the ability of some different cell types to accept or exclude dyes, Ehrlich developed the idea of the selective interactions of chemical substances with receptors expressed in cells. He proposed for the first time the idea of a receptor as a selective binding for chemotherapeutic agents (Figure 3) (Bosch and Rosich, 2008).

Ehrlich, together with Paul Guttman, pioneered the use of fully synthetic drugs in medicine. In the 1880s, after discovering that methylene blue could stain *Plasmodiidae,* which includes the malaria pathogen, Ehrlich administered this substance to two patients suffering this disease. The publication of the results is the first report of a synthetic drug being used successfully to treat a specific disease (Ehrlich and Guttmann, 1891). Ehrlich made further advances in chemotherapy when he introduced arsenicals for the treatment of syphilis. He took the premise that an infection caused by a micro-organism could be cured if the drug of choice was selectively taken by the invading microbes. Looking for a cure or treatment for syphilis, he began an exhaustive search for an arsenic

compound that would be a 'magic bullet', able to kill the microbe but not the person with the disease. After screening more than 600 arsenicals, compound 606 was found to be active in rabbits against the bacteria responsible of syphilis, *Treponema pallidum*. This compound was marketed as 'Salvarsan'. Ehrlich was the first investigator to develop an agent with a specific therapeutic effect on the basis of theoretical considerations. The systemic approach introduced by Paul Ehrlich became the cornerstone of drug search strategies in the pharmaceutical industry and resulted in thousands of drugs identified and tested clinically (Gelpi et al., 2015).

Drug receptor theory acceptance

Ehrlich receptor theory was not readily adopted by pharmacologists. It was in the early 1930s when significant support for the concept of drug receptors emerged from quantitative analysis of drug action receptor on cells by Alfred Joseph Clark. He analyzed mathematically a large amount of pharmacological data, showing that for many drugs the relationship between the drug concentration and the biological effect corresponded to a hyperbolic curve. Clark argued that the hyperbolic curve of drug action expressed the equilibrium between a drug present in excess that reacts with a finite number of cell receptors. He concluded that the pharmacological action that was produced was "directly proportional to the number of receptors occupied" (Maehle et al., 2002).

The last breakthroughs that strengthened the confidence in receptor theory were the direct measurement of drug binding and the development of a selective inhibitor of beta-adrenoreceptor, propranolol.

Early drug discovery

Together with pharmacology, drug discovery process has been also evolving during the last two centuries. This process could be divided in two main periods; the first one originated at the beginning of XIXth century, whereas the second one would start in the 1930s.

During the first period, the main aim was to isolate and purify natural products, obtaining new compounds by chemical synthesis and study their physiological properties. In this period, two generation of drugs were introduced. The first one is composed by alkaloids and organic products, while the second generation includes analgesics, hypnotics and antipyretics. The complete catalogue of effective drugs included morphine for pain, salicylates for fever, quinine for malaria, phenorbital for seizures, ether chloroform for analgesia, and not much else until 1935, when Gerhard Domagk discovered prontosil, which is considered the inflexion point of the second generation of drugs (Raviña Rubira, 2011).

The discovery of Prontosil: the first antibiotic

Domagk was a german physician and scientist of the Institute of Experimental Pathology, where he created a group focused on identifying antibacterial activity in dyes. Since one of the difficulties encountered by early researchers working in this area was the lack of reliable test for antibacterial activity, Domagk developed a method for screening the survival of mice that had been inoculated with *Streptococcus pyogenes*. Domagk obtained favorable results with derivatives of sulfonamide, with the most effective compound in protecting mice from lethal dose of *H. streptococcus* being Prontosil. After successfully using Prontosil to treat rabbits infected with *H. Streptococcus*, this drug began to be supplied to physicians to treat patients with life-threatening streptococcal

infections. Prontosil can thus be considered the first effective synthesized chemotherapeutic agent for treating bacterial infections (Bentley, 2009).

Prior to Prontosil, researchers and clinicians had very little options to alter the course of infectious diseases; it was considered untenable and ludicrous the idea that bacterial infection could be cured through the systematic administration of chemical substances. However, Domagk and his colleagues provided the impetus for changing these perceptions. An insightful remark made by Alexander Fleming, discoverer of penicillin, gave an interesting perspective of the importance of Domagk's discovery: "Without Domagk, there would be no sulfonamide! Without sulfonamide, there would be no penicillin! And without penicillin, there would be no antibiotics".

The capacity to treat a bacterial infection with an effective anti-microbial single pure drug provided clinicians with novel and substantial opportunities in therapeutics during an era where infectious deaths were common. In spite of the fact that Domagk's discovery was less celebrated than many others, it was the first step that allowed an increase in new drugs and treatments. His contributions represent an inflexion point of this new age, the Golden Age of drug discovery (Grundmann, 2004).

The discovery of Penicillin: the era of antibiotics

The second period begins right before the start of World War II, in 1935. It started with the introduction of the 3rd generation of drugs, which includes vitamins, hormones, sulfonamides, antibiotics, and their derivatives (Raviña Rubira, 2011). One of the most relevant discoveries of this period was penicillin, by Alexander Fleming, that replaced sulfonamides in phage therapy, since it showed better effects and fewer side effects.

Fleming was a microbiologist at the University of London. He spent a major portion of his research career studying a variety of diverse substances that

interfered with bacterial growth. In 1922, he isolated lysozyme from the nasal passages of a patient suffering acute rhinitis, and found that it was an antibacterial substance that could protect against certain non-pathogenic micro-organisms from becoming virulent. Fleming also found lysozyme in lacrimal fluid and saliva; this motivated him to seek other natural substances with antibacterial activity. This interest in natural substances promoted the accidental discovery of penicillin.

In St. Mary's Hospital, Fleming had a disorganized and untidy laboratory where there were contaminated Petri dishes and other detritus for extended periods of time. The disorganized state of his laboratory would have and important bearing on the discovery of penillicin. One day during the summer of 1928, a spore from a mold produced in the laboratory, floated into the laboratory and settled in a Petri dish impregnated with staphylococci. After several weeks, when Fleming was back to the laboratory after holidays, the contaminant mold had grown. Fleming inspected the Petri dishes and right in the corner of one of them in which he had grown a strain of staphylococci, he observed a small mold. He found that around the mold, the colonies had almost completely disappeared. He was intrigued by this observation, because staphylococci were known to be notoriously resistant to lysis. Since the mold could attack pathogen micro-organisms, he considered that the contaminant in the dish could have clinical utility. So he studied the properties of the unknown substance. With the aid of C. J. Latouche, a mycologist, he identified the mold as *Penicillium rubrum*. Since mold belonged to the genus *Penicillium*, the antibacterial substance was named *penicillin*. He found that penicillin killed streptococci, pneumococci, gonococci, meningococci, and diphteria bacilli. He also observed that penicillin was nontoxic to animals, and it was more effective against gram-positive than gram-negative. Fleming also identified some microorganisms insensitive to penicillin, like enterocoli, tubercle bacillus influenza and typhoid bacilli. Penicillin opened the door to a new era in the treatment of bacterial infections.

After the discovery of penicillin, and subsequently of other antibiotics, many drug companies established departments of microbiology and fermentation units (Aminov, 2010; Ligon, 2004).

The discovery of safe and effective antibiotic therapy was followed by the development of drugs effective in psychic, neurological, and cardiovascular disorders.

Target-based drug discovery: the paradigm change

After the identification of alpha- and beta-adrenoceptors by Raymond P. Alquist, the first selective beta-blocker was introduced. In 1965, propranolol became the first drug with high affinity for beta-receptor and low affinity for alpha-receptor. Propranolol was a best-selling drug still used today that boosted the acceptance of 'magic bullet' concept and promoted the rational drug design. Beta-blockers are included in the fourth drug generation (1960-1980) together with semisynthetic antibiotics, psychopharmacological agents, and cardiovascular agents (Quirke, 2006).

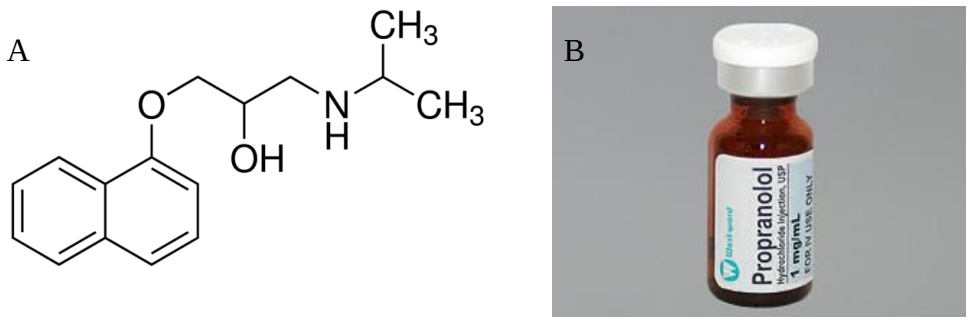


Figure 4: A) Propranolol structure ("Propranolol hydrochloride | Sigma – Aldrich,", n.d.).
B) Propanolol marketed. ("propranolol intravenous," n.d.)

Finally, the 5th generation includes enzyme inhibitors and biotechnologically derived drugs. In the mid-1980s, technological advances and the acceptance of receptor theory enabled a paradigm change in drug discovery. The advent of genomic sciences, DNA sequencing, combinatorial chemistry, and high throughput screening led ultimately to reverse pharmacology. So, drug discovery became a process based on the hypothesis that the modulation of the activity of a specific protein target will have beneficial therapeutic effects (Drews, 2000; Rubin, 2007).

Early animal experimentation

Pharmacology has depended largely on experiments in animals (Hajar, 2011). In fact, in early times, our ancestors used self-experimentally trial-and-error search for finding plants with therapeutic value (Scheindlin, n.d.). These human experiments allowed ancient civilizations to learn about hundreds of substances. During lots of centuries, it was not possible to isolate the active ingredients of the plants and botanical substances. Doctors and pharmacists did not have the necessary knowledge or technology. So, the cause-and-effect relationships of the administration of a drug were not measured scientifically. Each time that a herbal drug was prescribed was equivalent to an experiment. The strength of prescribed remedies depended on how they were prepared and administered, and thus the effectiveness was very variable. Due to it, there were discrepancies between doctors about the benefits of a drug (Altman, 1987; Weisse, 2012).

Pharmacologists soon learned that if they really wanted to know the effect of a herbal preparation on patient, it had to be tried first on people. So, many of them tested their potions on themselves. They could learn that the action of a particular drug could vary depending on how it was administered. For example, some drugs are ineffective when they are swallowed because stomach acids

inactivate them. Self-experimentation provided crucial insights into the way drugs can be used in medicine (Altman, 1987; Weisse, 2012).

In 1803 Friedrich Serturner, the German pharmacist who isolated the first alkaloid from opium, experimented on himself to produce the drug that even nowadays is the principal painkiller, morphine. Few scientists took more risks than Serturner. He observed that some samples of the drug could ease pain, while others did not. He hypothesized that opium contained an active principle that was responsible for the effect, but only when the concentration was high enough. He isolated the narcotic substance using basic techniques of chemical analysis to pharmacology, and named it 'morphine', for Morpheus, the god of sleep. Once he had isolated morphine, he began testing it in animals. He put crystals of pure morphine in food for mice and food for some dogs. In both cases, it put animals to sleep and ultimately killed them. Undaunted by the effects of the drug, Serturner decided to test a greatly reduced dosage on him and 3 friends because, in his own words, *experiments on animals do not give exact results*. Serturned and his friends administered themselves 100mg of morphine in 3 dosages. They began taking the first dosage of morphine and their faces became flushed and felt feverish. After half hour, they took another dosage, feverish increased and they began to feel nauseous and dizzy. Fifteen minutes later they took another dosage, and they experienced a sharp pain in stomach and felt they were about to faint. They layd down and fell into a dreaming state. Serturner and his friends experienced the symptoms of severe opium poisoning for several days (Altman, 1987; Weisse, 2012).

Toxicology and safety pharmacology in XXth century

With the appearance of pharmaceutical industry in the late XIXth century, the number of radical scientist using self-experimentation in their quest for new drugs was highly reduced. The use of humans for testing also decreased, while

the use of animals became more important, especially after some tragic incidents caused for not testing previously the safety of drugs. In 1937, a pharmaceutical company created a preparation of sulfanilamide, using diethylene glycol (DEG) as a solvent, called 'Elixir Sulfanilamide'. The company's chief pharmacist was not aware that DEG was poisonous to humans, and the product was marketed. It caused the death of more than a hundred people. No animal testing was done. After this incident, in 1938 safety testing of drugs on animals became compulsory before they could be marketed. The most frequently animals used in this safety testing and other pharmacological studies are mammals. Mice, rats, guinea pigs, rabbits and dogs are used depending on the type of test, since each one has special characteristics that make them more or less optimal. But mice are usually preferred because of their small size, ease of breeding, and short generation time (Hajar, 2011).

Although humans are no longer used as laboratory animals, they are still essential in clinical pharmacology. After a new drug compound has gone through sufficient preclinical testing to show therapeutic action and safety on short-term administration, and has passed the strict review of the Food and Drug Administration (FDA), the compound is administered to a small number of healthy human volunteers under closely controlled and monitored conditions in the Phase I of clinical trials. This phase provides information about the dosage and the common side effects likely to be expected (Scheindlin, n.d.).

I.2 History of Ethnopharmacology

Ethnopharmacology can be defined as a multidisciplinary area of research focused on the observation, identification, description, and experimental

investigation of ingredients of indigenous medicines of past and present cultures. It studies beneficial, toxic or other direct pharmacological effects. Ethnopharmacology is an excellent tool for drug discovery and it has contributed to the finding of many important plant-derived drugs (Heinrich et al., 2009; Soejarto et al., 2005).

The term 'Ethnopharmacology' was first used in 1967 as the title of a book on hallucinogens by Efron *et al.*, namely "Ethnopharmacologic Search for Psychoactive Drugs". It was however proposed significantly later than the term "ethnobotany", which appeared in a 1896 study of human's plant use by William Harshberger, an american botanist. Both ethnopharmacology and ethnobotany investigate the relationship between humans and plants and all its complexity (Heinrich, 2014).

Despite the fact that the term ethnopharmacology has a rather short history, for centuries researchers have been interested in natural medicines. Many studies involving the documentation and systemic study of local and traditional uses of plants have been performed. Explorers, missionaries, merchants, and experts in the respective healing traditions describe the uses of such medicinal plants, which form the basis of ethnopharmacology-based drug development. One of the first ethnopharmacolgical studies is attributed to one of the founding fathers of pharmacology and physiology, Claude Bernard (1813-1878). He was interested on the study of curare and the reasons behind why it was non-toxic when it was applied orally. In his own words, Bernard highlighted that *one of the facts noted by all those who reported on curare is the lack of toxicity of the poison in the gastrointestinal tract. The Indians indeed used curare as a poison and as a remedy for the stomach*. Bernard described also the different pharmacological effects depending on the administration: *if curare is applied into a living tissue via an arrow or a poisoned instrument, it results in death more quickly if it gets into the blood vessels more rapidly. Therefore death occurs more rapidly if one uses dissolved curare instead*

of the dried toxin (Black, 1999; Paton, 1976). He also demonstrated that animals didn't show any nervousness and no sign of pain. If the blood flow in the hind leg of a frog is interrupted using a ligature without interrupting the innervations and it is poisoned via an injury of the hind leg, it retains its mobility and the animal does not die from curare poisoning. These and subsequent studies provided a better understanding of the pharmacological effects of curare. The main toxin of curare was isolated from *Chondrodendron tomentosum* and it was identified as D-tubocurarine. In 1947, the structure of this alkaloid was determined and, in 1970, the tubocurarine structure was resolved using Nuclear Magnetic Ressonance (NMR). Nowadays, tubocurarine is used sporadically in some European countries, for example in France, as a muscle relaxant during surgery (Heinrich, 2014).

We can consider an ethnopharmacological approach any empirical use and medical testing of a plant for novel uses. Another example of a systemic study of the medical properties of a herbal medicine, is the study performed by William Withering (1741-1799) on foxglove, *Digitalis purpurea*. Foxglove was reportedly used by an English housewife to treat dropsy. Withering used the orally transmitted knowledge of British herbalism to develop a medicine that could be used by medical doctors (Heinrich, 2014). After identifying foxglove side-effects, he described the best and safest way of using it.

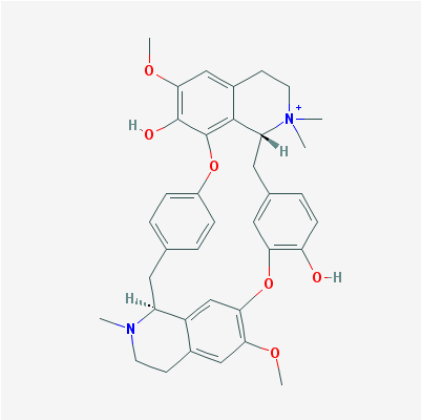


Figure 5:
D-tubocurarine structure (Pubchem, n.d.)

Figure 6:
Digitalis purpurea
(http://www.floracatalana.net)

The first natural products were identified during the early years of the XIXth century, and they were characterized in subsequent research. Some examples of compounds isolated during the early XIXth century are listed below (Heinrich, 2010):

- 1804 – Morphine, from opium poppy (*Papaver somniferum* L., Papaveraceae) was identified by F. W. Sertürner
- 1817 – Emetine, from ipecacuanha (Cephaelis ipecacuanba (Brot.) A. Rich., Rubiaceae)
- 1817 – Strychinine, from *Strychnos* spp., Loganiaceae.
- 1820 – Quinine, from *Cinchona* spp. (Rubiaceae)
- 1821 – Caffeine, from coffe tree (*Coffea arabica* L., and *C. canephora* Pierre ex. Froehn, Rubiaceae)
- 1826 – Coniine, from hemlock (*Conium maculatum* L., Apiaceae)
- 1833 – Atropine, from belladonna (*Atropa belladonna* L. Solanaceae)
- 1838 – Salicin, from willow bark (*Salix* spp., Salicaceae)

The above-mentioned studies and discoveries helped to coin the term ethnopharmacology, which offered focus and clear concept of the field of

research interested in the interface of traditional and local medical use of plants and their biological characteristics (Heinrich, 2010).

After the first use of the term 'Ethnopharmacology' in the context of hallucinogenic plants, it was occasionally used until 1979 when Laurent Rivier and Jan Bruhn founded the Journal of Ethnopharmacology. This term replaced many other terms used previously, like pharmacoethnologia, aboriginal botany or *Pharmakoëthnologie* used already, which was used by Tschirch (1910) in his '*Handbuch der Pharmakognosie*' (Heinrich, 2014).

With the appearance of the Journal of Ethnopharmacology, the scope was broadened to *a multidisciplinary area of research concerned with the observation, description, and experimental research of indigenous drugs and their biological activity*. Currently, many journals are publishing ethnopharmacological research and demonstrate the research interest in how humans use plants as medicine. The most relevant articles in ethnopharmacology are those that study the biological and pharmacological activity of locally and traditionally used medicinal plants and their historical uses (Heinrich, 2014).

I.3 Medicinal plants

Herbs have been throughout the history of mankind. They have not only provided an important source of food, but also served to cure different ailments. Animals have also taken advantage of herbs in the co-evolution of plant-animal relationships. In fact, several animal species began to utilize plants rich in bioactive compounds for protection against predators and parasites. So, it is arguable that the actual origins of herbal medicine can be traced back to the animal kingdom (Etkin and Elisabetsky, 2009).

Herbal medicine, also called traditional or natural medicine, has existed in one way or another in different cultures and civilizations, such as Egyptians, Western, Chinese, Kampo (Japan), Greco-Arab or Unani/Tibb (south Asia) (Dias et al., 2012). The knowledge on traditional medicines has been usually acquired through trial and error over the centuries. The reasons why people selected certain plants over others are complex, and a variety of factors, both ideational (based on ideas or beliefs) and tangible. The organoleptic properties of some plants were important elements to infer their therapeutic properties, information that was usually conveyed by the smell, texture, appearance, and even sound of plants. The meanings of organoleptics may vary between cultures (Leonti, 2002). Thus, red colour in plants may be a symbol for native medicines to identify plants that can be used to treat wounds, since red is the colour of blood. However, the red colour of some of these plants can also mean that their red quinones are hemostatic and antimicrobial, properties that have been identified by the users of these plants through their own experiences (Etkin and Elisabetsky, 2009).

The ideational component of plants valuation includes the characterization via binary opposition, namely, cold-hot (ying-yang) and wet-dry. Cold-hot, for example, is a binary characterization which has been used for many different culture medicines, from Chinese, to Galenic or Aztecs. In traditional chinese medicine, this characterization is very important. It doesn't refer to the temperature of the food but to the effect of the food on the body (Heinrich et al., 2006). Cold foods provide low-energy and help balance hot foods, whereas hot foods provide greater energy for activity, are higher in calories, and are used to treat pallor and weakness. Those who consume too many hot foods may feel overly warm, anxious or constipated, while those who consume too many cold foods experience diarrhoea, weakness, and depression. These principles are part of models that emphasize balance and proportion of medicine intake (Jiang et al., 2011).

Phytotherapeuticals have also had a close relationship to traditions. In this respect, different ethnopharmacological uses of some traditional medicines can be found in different regions. One example is rosemary (*Rosmarinus officinalis* L. Lamiaceae), a native plant of of the Mediterranean region used for a variety of maladies depending on the culture. In Spain, it is used to treat several forms of pain, including rheumatic and traumatic muscular pain and pain in the bones, but it is also widely used for gastrointestinal and respiratory disorders. These ethnopharmaceutical uses are significantly different from those in Mexico and Guatemala, where rosemary was adopted by native Mesoamericans due to medical syncretism. Still today in the medical systems of Mexico and Guatemala rosemary is used as a postpartum remedy, to treat respiratory problems, and against skin infections (Heinrich et al., 2006).

Figure 7: *Rosmarinus officinalis* (http://www.floracatalana.net)

Traditional Catalan Medicine

Catalonia is a country located in the Mediterranean area, in the north-eastern part of the Iberian Peninsula, conveying an area of 32,108 km² with a population around 7.5 million people. In this region, climate ranges from

mediterranean to nival, and the mean temperature varies from 17º C to 0º C due to the altitude range, which goes up to 3,000 metres. One the other hand, the altitude ranges and the presence of an orography compartmentalized in mountains and depressions produces an irregular rainfall. It varies from 1,200 mm/year in some points of the Pyrenees, to less than 400 mm/year in the west of the central depression. But in general, winter is cool or slightly cold depending on the location, spring and autumn are typically the rainiest seasons, and summers are hot and dry, except for the Pyrenean valleys, where summer is typically stormy (de Bolòs et al., 1997).

The evolution of the use of medicinal plants, like in every place on Earth, started with the adaptation of human to the medium. Catalonia has a set of geographic characteristics which favour the presence of a high climatic diversity and, as a consequence, there is also a wide range of ecosystems and flora, which has been the factor that gave rise to a strong ethnobotanical culture.

Throughout the centuries, the use of herbal medicines to treat ailments was closely accompanied by superstitious, magic and religious practices. One example of these beliefs is the elixir 'cordial', a medicinal drink used to treat cough and flu. It was composed of borage flower, mallow, poppy, violet, rose, gentian and holy water. Water had to be holy, otherwise the elixir would not work (Gausachs, 2008).

The first important person in Catalan ethnobotany history is documented back in the XIIIth century. His name was Arnau de Vilanova, an alchemist, writer, philosopher, and theologist. He mixed theology with astrology in his medical practices and wrote one of the first recipe books of the traditional Catalan medicine history, the *Antidotarium*. Later on in the XVIth century we find Francesc Micó, the most important catalan botanist. He created one of the first and most important herbariums and he was the first to work with a scientific methodology. Finally, in the XXth century, modern pharmaceutical

ethnobotany was established by Dr. Pius Font i Quer, who picked up a great deal of information on popular uses of medicinal plants in his floristic expeditions in the Iberian Peninsule, Balearic Islands and Pyrenees. He published *Plantas Medicinales*, a book to renew and update *Materia Medica* of Dioscorides. Although he only included 678 hispano-lusitans and western Mediterraneam herbs, this book became a reference of medicinal botany and ethnobotany (Gausachs, 2008).

I.4 Parallelism between plants extracts and polypharmacy

Plants contain hundreds of different chemical compounds, so the ingestion of herbal medicines is equivalent to the ingestion of dozens of chemicals including fatty acids, sterols, alkaloids, flavonoids, glycosides, saponins and others. A plant extract prepared to be used as a herbal medicine is thus, literally, a chemical cocktail prepared to interact with almost everything. Moreover, the number of compounds can be even increased when traditional medicines combine multiple herbs, some formulations containing more than 10 different plants. In spite of it, herbal medicines are in general associated to less side effects than those of modern drugs. It is accepted that, even if a specific compound is identified and its action understood, it is the effect of the other supporting or modifying compounds present in the whole plant which complement the therapeutic action and minimize side effects and adverse reactions in herbal reactions (Firenzuoli and Gori, 2007; Vickers et al., 2001).

In this respect, the use of herbal medicines can be compared with the common practice of polypharmacy, that is, the use of more than one drug in medicine. Polypharmacy is quite common in the elderly population, with many patients

taking four or more medications. It is estimated that affects about 40% of adults over the age of 65. Some of the concerns of polypharmacy are the possibility of drug-drug interactions and thus, the increase in adverse reactions. It is often associated with a decreased quality of life, decreased mobility, and cognition (Fulton and Riley Allen, 2005).

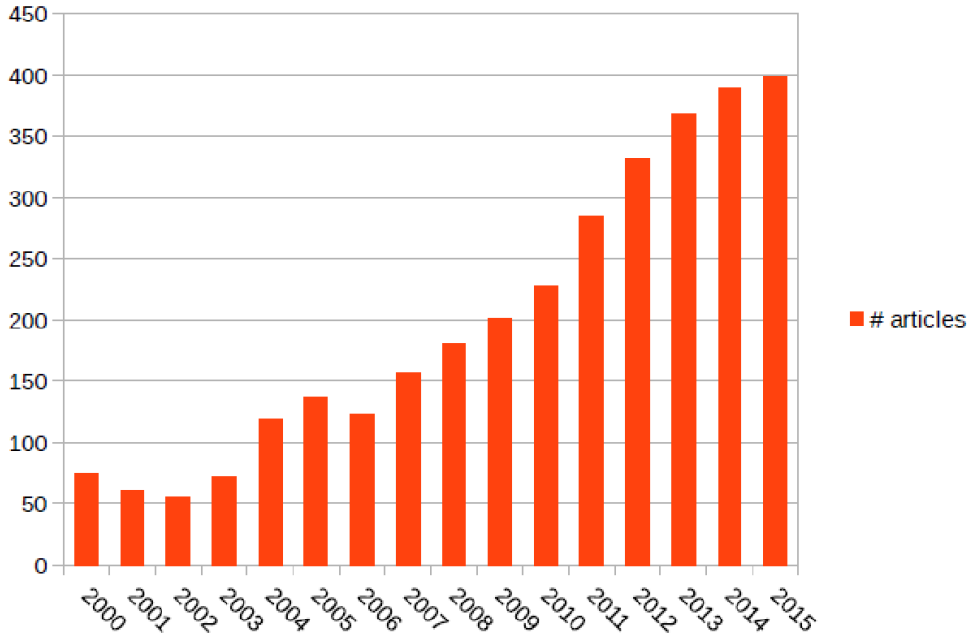
Herbalists maintain that side effects and adverse reactions common in polypharmacy are due to the single element philosophy of modern medicine. Oppositely to herbal medicine, most of drugs used in modern medicine contain a single active ingredient. It often implies that higher concentrations of isolated chemical compounds may be ingested, which may increase the therapeutic action, and subsequently side effects, that can be even increased in combinations with other drugs (Wachtel-Galor and Benzie, 2011).

It is frequently believed that herbal medicines are free of toxins or adverse effects, since they are originated from nature. But, like it may also happen in polypharmacy, the high number of active ingredients present in herbal medicines will increase the possibilities of interactions between them, which may result in unexpected and undesired side effects. To ensure the safe use of herbal medicine products, they should be managed as drugs. It should be considered that, even though they can treat/cure a disease or maintain health, they may also cause some adverse reactions. And, as drugs, to avoid these adverse effects, it is important to understand how herbal medicines act and what is the most appropriate dosage and formulation for treating the disease (Wachtel-Galor and Benzie, 2011; Zhang et al., 2015).

I.5 Target deconvolution

When chemical analysis became available in the early XIXth century, chemists began making their own version of plant compounds and the use of herbal medicines declined in favour of drugs. However, over the last 40 years the interest in medicinal herbs resurged. A rapid and continued growth of herbal market stimulated the interest in the scientific understanding of how herbs work and their efficacy. As a consequence, more and better analytical techniques have been developed and used to infer the mechanism of action of herbal medicines (Wills et al., 2000).

Until the last quarter of the XXth century, basically only experimental methods were used to study the therapeutic effects of medicinal plants and natural



Graphic 1: Pubmed publications showing the increased scientific interest in discovery of the mode of action of herbal medicines. Each column is the quantity of articles with 'herbal medicine mechanism' in title or abstract

products and to validate the hypotheses on their mechanism of action.

For instance, in 1988, King *et al.* investigated the effects of tetrandine, a bis-benzylisoquinoline alkaloid from *Radix stephania tetrandrae* and present also in other Chinese and Japanese herbs, which has been traditionally used to treat hypertension. Tetrandrine was a putative L-type Ca^{2+} entry blocker whose mechanism of action was unknown (King et al., 1988). L-type Ca^{+} channels include three separate binding sites, those for dihydropyridines, benzothiazepines and phenyl alkylamines (Catterall et al., n.d.). To investigate the tetrandrine mechanism, King *et al.* characterized the effects of tetrandrine on binding to the three chemical classes of L-type Ca^{2+} entry blockers. Their results suggested that tetrandrine interacts directly at the benzothiazepine-binding site of the Ca^{+} entry blocker receptor complex and allosterically modulates ligand binding at other receptors in this complex. In a study performed later by Liu *et al* on the same alkaloid, it was shown that tetrandrine blocked also T-type Ca^{+} channel. All these findings explained the therapeutic effectiveness of this alkaloid as vasodilatory agent, thus confirming its appropriate use to reduce blood pressure (Liu et al., 1992).

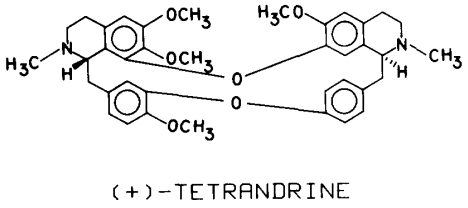


Figure 8 (left): *Radix stephania tetrandrae* (macognosy, n.d.)
Figure 9 (right): tetrandrine structure (King et al., 1988)

Traditional medicine current state

According to the World Health Organization (WHO), about three quarters of the world population use traditional remedies (mainly herbs) for the health care (Wang et al., 2012). In the same way, research on the identification of effective ingredients of medicinal plants and functioning targets was intensified during the last years, and it keeps on increasing at present. A great amount of new data about the composition of herbal medicines appeared, and several databases about medicinal plants and their ingredients have been established, particularly databases about traditional Chinese medicine, such as the Traditional Chinese Medicine Database (TCMID).

Since botanical mixtures contain hundreds of potentially bioactive natural products, in many cases it is not feasible to elucidate their mode of action using conventional biochemical technologies. The recent construction of databases with information on herbs, their chemical ingredients, and their therapeutic use opens an avenue for the application of computational methodologies in herbal medicine research. In this respect, computational methods offer a cost-effective and efficient approach to predict the mode of action of traditional medicines (Gertsch, 2011).

Rollinger *et al.* published one of the firsts ground-breaking examples of computational methods applied to investigating the potential mode of action of a plant. They used a computational approach to *in silico* target fishing to suggest the potential targets of 16 secondary metabolites identified from the aerial parts of *Ruta graveolens* (rue). Rue is an important medicinal plant native to the Mediterranean region and the Balkans. It has been used since ancient times to prevent contagion, to repel insects, and to heal their bites. Based on the fact that previous *in vitro* screening studies confirmed a significant acetylcholinesterase inhibiting activity for some extracts of Rutae herba, Rollinger *et al.* focused their study on three biological proteins,

Acetylcholinesterase, Human rhinovirus coat protein, and Cannabinoid receptor type-2. To predict ligand-target interactions, they generated 3-D pharmacophore models and performed a parallel screening against the putative biological targets. The major virtual hits that were obtained in that way showed a relatively good predictability and were corroborated in biological assays. In their results, arborinine was the compound with higher inhibitory activity for Acetylcholinesterase and Human rhinovirus coat protein, whereas rutamarin was confirmed as and active ligand of the Cannabinoid receptor type-2 (Rollinger et al., 2009).

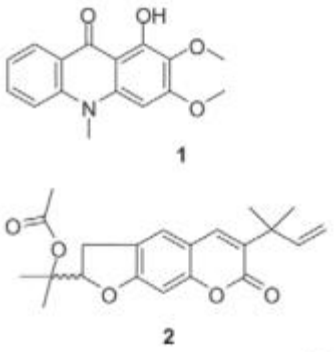


Figure 10 (left):
Ruta graveolens (Freckman, W. F. – https://gobotany.newenglandwild.org/species/ruta/graveolents/)
Figure 11 (up):
1-Rutamarin structure. 2-Arborinine structure. (Rollinger et al., 2009).

As shown above, traditional medicines may not only work through the inhibition of a particular protein target, but through an optimal balance of interactions with multiple protein targets within the same network. Thus, the ultimate pharmacological effect could be the result of is now referred to as network pharmacology or biochemical synergism. The term network

pharmacology was coined by Andrew L. Hopkins (2007) to emphasize that there are drug-target networks rather than single drug targets. On the other hand, Wagner stated in 2011 that in botanical drugs, multitarget effects predominate over other synergistic mechanisms (Wagner, 2011). Synergism can be produced by different molecular machinery, ligand interactions in one single protein or at the level of downstream effects. The fact that natural products may interact with multiple targets (polypharmacology), implies that Paul Ehrlich's concept of 'magic bullet' (single drug target) applied in drug discovery and pharmacology may be changed to 'magic shotgun' (multiple drug targets) (Strebhardt and Ullrich, 2008).

Polypharmacology is one of the major challenges in drug discovery, its rational design is complex, and new methods are needed to validate target combinations and optimize multiple structure-activity relationships while maintaining drug-like properties. The development of drug combinations is difficult, however herbal medicines can offer good starting points. Taking apart and reassembling all their bioactive compounds, it would be possible to find which natural products contribute to the pharmacodynamics of a given pharmacological effect, directly or indirectly (Gertsch, 2011).

During the past decade, there has been an important and fast development of 'omics' technologies and systems biology. Systems biology depicts the complex interactions at different levels as various networks and elucidates the underlying mechanisms of biological systems by studying those networks. This has facilitated a systems-level understanding of biological process concerning the interactions of genes, proteins and environmental factors (Ma'ayan, 2011). Thus, the application of network-based system biology to study the pharmacology of traditional medicines opens the possibility to understand the targets of bioactive ingredients and their interactions in the molecular network (Pelkonen et al., 2012).

The creation of drug-target networks that reflect the pharmacology of traditional medicines requires the previous identification of active compounds from botanical medicines and the targets of each compound. Some of the most conventional methods to identify bioactive compounds are the direct extraction of components from the plant, their pharmacological evaluation, and biochromatography, were active components are identified through screening (Wang et al., 2000). On the other hand, for the identification of targeted proteins of these compounds, with the birth of databases providing information about herbal compounds, it is now possible to apply data mining of bioactive compounds and *in silico* screening approaches to perform rapid and low-cost predictions of their targets (Zhao et al., 2010).

One of the first studies where systems pharmacology was applied to elucidate the mechanism of action of a traditional herb was performed by Zhao *et al*. In this investigation, they tried to shed light on the antidepressant mechanism of action of *Hypericum perforatum*. First, they collected the main active compounds of rue, hyperforin, hypericin, pseudohypericin, amentoflavone and several flavonoids. Through comprehensive literature search, they collected the neurotransmitter receptors, transporter proteins, and ion channels on which the St. John's wort active compounds show effects. Mapping these proteins onto KEGG pathways, they observed that active compounds acted on the system of neuroactive ligand-receptor interaction. So, the actions of multiple compounds of *Hypericum perforatum* resulted in an additive or synergistic antidepressant efficacy, producing the same antidepressant action as normal monotherapy but at much lower doses for separate compounds (Zhao et al., 2010).

I.6 Polypharmacology

Not so long ago, the paradigm in drug discovery was to pursue "one drug for one target". That implied that the ultimate objective was to obtain potent and selective small-molecule candidates against individual protein targets. Unfortunately, most of those alleged "selective candidates" were not thoroughly screened against a large panel of proteins. When several public and private initiatives collected and stored drug-target interaction data, it was observed that many of these selective compounds were in fact interacting with multiple targets and the percepcion of drug selectivity changed (Mestres et al., 2008; Nat Biotechnol). The ability of small molecules to interact with multiple targets was defined as "polypharmacology".

The term polypharmacoly was used for first time in 1971 by Domino as a synonym of polypharmacy. However, in 1997, Kenny *et al* used the term polypharmacology to refer to the lack of a_1-subtype adrenoceptor selectivity of the drug indoramin, that was known at the time to interact with other receptors, such as serotonin and histamine, leading to sedation as a side effect (Jalencas and Mestres, 2013). The number of articles and reviews published in the past 10 years containing the term polypharmacology in title, abstract or keywords (according to Scopus search results) has constantly increased. Out of 488 papers, 265 have been published in the last 2 years.

Now, it is widely recognised that selective drugs are more the exception rather than the rule and that most of therapeutically effective molecules tend interact with multiple proteins. Despite they were not designed on purpose, we can find numerous drugs that are known to have multi-targeting activities. For example, Aspirin, which is often used as an analgesic to relieve minor pains or as an antipyretic to reduce fever, also acts as an anti-inflammatory medication to treat

rheumatoid arthritis, pericarditis and kawasaki diseases (Reddy and Zhang, 2013).

On the other hand, there are some complex pathologies that have evidenced to be often polygenic in nature and tend to involve the deregulation of a complex and extended network of proteins. For their treatment, pharmacological interventions based on a single target are unlikely to be successful. The inherent biological networks might impede the efforts of shutting down a cellular pathway if a single switch is turned off, so the modulation of an optimal array of targets may provide a more efficient strategy (Anighoro et al., 2014). Some of these polygenic disorders beneficed by polypharmacology are psychiatric diseases and cancer (Peters, 2013; Roth et al., 2004). For example, all marketed antipsychotics are polypharmacological by nature. Chlorpromazine, the oldest known typical antipsychotic, which was originally developed more than 50 years as a sedative histamine receptor antagonist, was subsequently discovered to be a high-affinity dopamine antagonist. The examination of various public databases, including Psychoactive Drug Screening Program (PDSP) database, GLIDA, DrugBank, and PubChem showed that chlorpromazine interacts with numerous molecular targets. The list of targeted protein in PDSP database includes 29 different receptors: 9 different 5-HT receptors, 6 alpha-adrenergic receptors, 5 muscarinic receptors, 5 dopamine receptors, 3 histamine receptors and the imidazoline l1 receptor. In the case of PubChem, 20 different screening assays were found where chlorpromazine was considered to be an 'active' compound. The activity as D2 antagonism is considered to be the core of chlorpromazine effectiveness as antipsychotic, but some of these other mentioned activities may contribute to its effectiveness. On the other hand, some of these interactions could also be responsible for some of the side effects linked to chlorpromazine, such as hypotension, drowsiness or weight gain (Peters, 2012). When we are talking about a multitarget (promiscuous) drug with a wide and usually unpredictable spectrum of biological activities it could

eventually lead to adverse reactions. The interaction with non-therapeutic targets can cause severe side effects (Anighoro et al., 2014).

I.7 Side effects/Toxicity

Toxicity can be defined as the degree to which a substance can damage the organism. In pharmacology, it used to be consequence of adverse drug reactions, defined as the response to a drug that is unintentionally noxious. It may occur at therapeutical doses normally used for the treatment, prophylaxis, or diagnoses of disease, but it is more often encountered at higher doses due to abuse, medicational error and unintended overdose. On the other hand, when adverse reactions are 'minor' and predictable, normally the term side effect is usually used as synonym for toxicity. Side effects are any effect caused by a drug other than the intended therapeutic effect, whether beneficial, neutral or harmful. It may occur as part of a pharmacological action of the drug or may unpredicable in its occurrence (Flora et al., 2012).

In spite of the fact that adverse reactions are usually associated to drugs, traditional medicines can also be responsible for them. There are many side effects that have been reported upon herb ingestion. One example is a case that describes 105 patients in Belgium who had been taking a chinese herbal product for weight and developed nephropathy, *Aristolochia fangchi*. Among them, 18 patients were found to have urothelial carcinoma, which was shown to be related to the formation of DNA adducts from the aristolochic acid in this herb. On the other hand, some other herb side effects may be caused by the excessive biological effects of their active compounds. For instance, ephedra, a herb widely used in traditional medicine for treating asthma, bronchitis and fever. Ephedra was subjected to an analysis and it was found that it was 40 times more likely to lead to a report of a side-effect, compared to other

commonly used herbal products. It had a great risk to provoke nausea, vomiting, psychiatric symptoms and palpitations. Ephedra was banned from FDA in 90s (Bent, 2008).

Many other factors may actually be affecting the occurrence of adverse drug reactions. We could divide them in three types related to patient, social, and drug factors.

Patient- and social-related toxicity factors

Some of the principal patient-related factors are associated with age, gender, body weight and fat distribution.

Age is a very important factor since elderly people are at high risk of developing them for several reasons. They are likely to have many health problems and thus take several prescriptions and over-the-counter drugs. Moreover, as people get older the liver loses the ability to metabolize drugs and kidneys are less able to excrete drugs into the urine. On the other hand, in aged people the amount of water decreases and the amount of fat tissue relative to water increase. Thus, drugs that dissolve in water reach higher concentrations, while drugs that dissolve in fat are more accumulated in fat because there is more fat tissue to store them. As a consequence many drugs tend to stay in an older person's body much longer than they would in a younger person's body, prolonging the drug effects and increasing the risk of side effects (Malinovska et al., 2015).

The biological differences between males and females may also affect the action of drugs. In comparison to men, women have in general lower body weight and organ size, more body fat, different gastric motility and lower glomelurar filtration rate. All these differences may alter the pharmacokinetics and pharmacodynamics of drugs including absorption, distribution, metabolism and elimination. For example, chlorpromazine and fluspirilene seem to be more

effective in women than in men for the same dosage in plasma concentration (Anderson, 2008).

Body weight and fat distribution can also be a factor responsible of adverse drug reactions. Once absorbed, some water-soluble drugs tend to stay within the blood and the fluid that surrounds cells, while some fat-soluble drugs tend to concentrate in fatty tissues; other drugs concentrate mainly in one small part of the body. These accumulated drugs are slowly released into the bloodstream, keeping blood levels of the drug from decreasing rapidly and thereby prolonging the effect. The distribution of a given drug may vary from person to person, and for example obese people may store large amounts of fat-soluble drugs, whereas very thin people may store relatively little. In the case of older people, even when thin, they may store large amounts of fat-soluble drugs because the proportion of body fat increases with ageing (Alomar, 2014).

On the other hand, among the social factors we can include alcoholism, drinking, smoking, race and ethnicity factors. One of the effects of alcohol consumption is the activation of enzymes that transform some drugs into toxic chemicals that can damage the liver and other body organs. Alcohol can also magnify the inhibitory effects of sedative and narcotics in the brain (Lewis et al., 2015). Smoking affects liver enzymes, acting as a potent inducer of hepatic cytochrome P450 (CYP) 1A1, 1A2 and 2E1. The hepatic CYP1A2 enzyme metabolizes many drugs, resulting in the decrease of their pharmacological effects (Faber and Fuhr, 2004). Finally, ethnic factors are linked to genetic factors, which in turn account for some of the inter-individual differences due to polymorphisms in genes encoding drug metabolising enzymes, drug transporters, and receptors. For instance, Morimoto *et al* observed that African americans were found to be more susceptible to developing angiotensin-converting enzyme (ACE)-related angioedema than other ethnic groups (Morimoto et al., 2004).

Diseases may also influence susceptibility to adverse drug reactions. Multiple co-occurring diseases make even cause that drugs that are helpful in one disease can be harmful in another. For example, some beta-blockers taken for heart disease or high blood pressure can worsen asthma and provoke problems in people with diabetes since these drugs raise blood sugar levels (Alomar, 2014).

Drug-related factors: Polypharmacy and polypharmacology

Finally, adverse drug reactions can be induced also by drug-related factors, which can be divided into drug dosage and frequency, polypharmacy and polypharmacology. These reactions can also be classified according to the required doses for obtaining the therapeutic effect, doses above the maximum dose required for a therapeutic effect (toxic effects), doses within the therapeutic range (collateral effects), and effects that occur at doses below the therapeutic range in susceptible patients (hypersusceptibility) (Ferner and Butt, 2012).

Polypharmacy and polypharmacology are also risk factors for adverse effects. In the first case, they may occur due to drug interaction, synergism, duplication, additive effect, discontinuation of therapy, changing the dose to save money, skipping some medication and physiological antagonism. In the second, they are related to the capacity of a drug to modulate multiple targets. The interactions of drugs with unintended targets can be the cause of several adverse effects.

Some adverse effects can be extremely severe, such as the secondary effects caused by antipsychotics. Off-target activity against histamine H_1, serotonin 5-HT_{2C}, and the muscarinic receptors, are assumed to be responsible for metabolic adverse effects, such as weight gain, hyperprolactinaemia, and diabetes (Roth et al., 2004).

Prevention and prediction

Adverse reactions are one of the leading causes of morbidity and mortality in healthcare. The American Society of Health System Pharmacists found that 85% of patients consulted in a survey expressed concerns about at least one drug-related issue. Adverse drug reactions are a significant public health problem in the world and therefore, it is important to prevent them. Prevention strategies should target the prescribing and monitoring stages of pharmaceutical care. Moreover, understanding the activity of drugs and adverse reactions may also help preventing them. In order to understand adverse drug reactions and being able to predict them, systems biology approaches can become very useful. The ever increasing amount of data resources in genomics, transcriptomics, metabolomics, proteomics and their relationship to human physiology are paramount to create a new generation of predictive systems.

Nowadays some approaches to predict toxicity are already being used in early drug discovery to select compounds to move forward in development and reduce pharmaceutical research and development costs. These methods use vast sets of experimental data from databases that cover information such as metabolic pathways, protein interactions, metabolomics, signal transduction and transcriptional regulation. Network visualization methods can also be applied in drug mapping to explore associative networks of drugs, pathways and diseases (Shoshi et al., 2015).

For instance, Xie et al. (2009) introduced a computational strategy for the systematic identification of protein-drug interactions networks in order to elucidate the molecular mechanisms associated with the adverse effects of Cholesteryl Ester Transfer Protein (CETP) inhibitors. These inhibitors are used for the treatment of cardiovascular diseases. One of these inhibitors is trocetrapib, which has deadly off-target effects and was withdrawn from phase III clinical trials. They predicted the off-targets of Torcetrapid and other CETP

inhibitors and mapped them to biological pathways. The predicted protein-ligand network obtained was consistent with experimental results, revealing that side effects are modulated through the combinatorial control of multiple interconnected pathways (Xie et al., 2009).

I.8 Exogenous compounds on human metabolism

Biological networks are emerging as a analyses and interpretative tools for a better understanding of both drugs therapeutic activity and associated side effects. They may include all types of omics data. For example, metabolic networks include reactions occuring in an organism and their metabolites. They help in the identification of essential proteins and have been applied in pharmacology and toxicology (Csermely et al., 2013).

But drugs (including traditional medicines) are not the unique chemicals targeting proteins in the biological system. Organisms need food to keep on producing energy for survival and, through nutrition, many other exogenous compounds are introduced in our body. Food supply includes many different molecules that can be classified as nutrients and non-nutrients. Non-nutrients are outweighting the number of nutrients. Some of the non-essential nutrients found in vegetables and traditional herbal medicines are secondary metabolites accumulated by plants for defense, reproduction, and so for. among them we find, for example, flavones, with metabolic effects on heart disease, or stanins, that are affecting cholesterol metabolism. In addition to those non-nutrients found in food supplies, some others are man made, intentionally or accidentally, and all of them are factored into the metabolome. To understand the effects of nutrition on human metabolic regulation, metabolomics is currently being applied also to nutritional research.

Most of these exogenous compounds (drugs, natural products, nutrients and non-nutrients) are transiently coexisting with endogenous metabolites, at least in biofluids, so they may be critically important in metabolic studies. In the same way that exogenous metabolites may have some metabolic effect, endogenous metabolites may also exert some effect on these exogenous metabolites activity if they are, for example, interacting with the same target.

I.9 Metabolomics

Understanding cell metabolism is not only important for drug discovery and clinical treatment of metabolic disorders, it is also essential in other fields, such as metabolic engineering and synthetic biology. As a consequence, in recent years the importance of metabolomics research has been growing fast (Kotera et al., 2014).

Metabolomics is the scientific study of chemical processes involving metabolites, the intermediates and products of metabolism. Usually, the term "metabolite" is restricted to small molecules. They are involved in several functions, including structure, signalling, stimulatory and inhibitory effects on enzymes, catalytic activity, defence, and interactions with other organisms. Metabolomics allows the construction of metabolic networks and the conversion of biological knowledge into mathematical format and the subsequent computation of physiological states to address a variety of scientific applied questions.

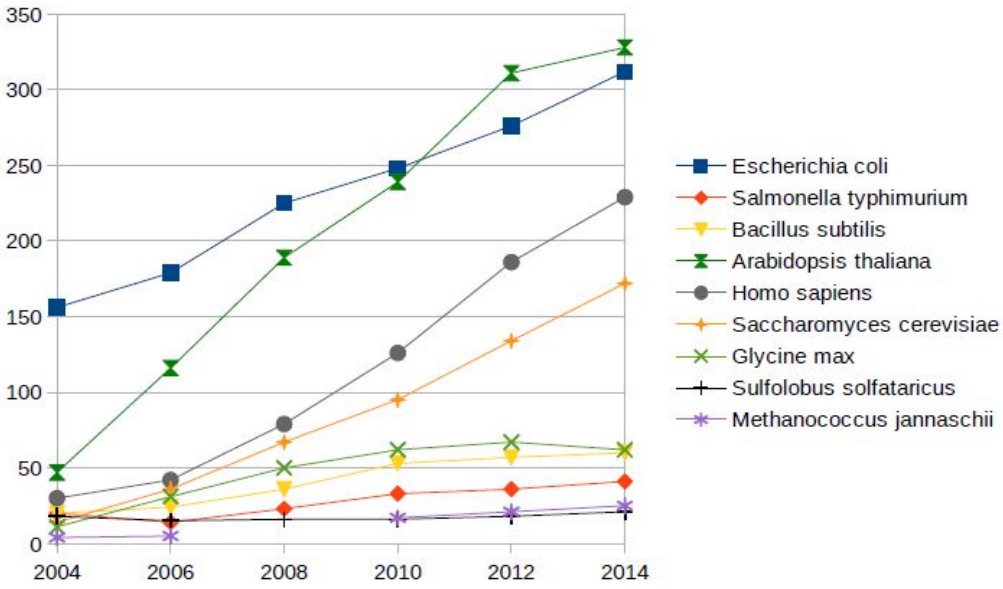
The knowledge about metabolism has proven useful to investigate biomarkers of diseases such as cancer (Koulman et al., 2009), quality of foods (Fitzgerald et al., 2009) and assessment of environmental pollution (Krauss et al., 2010)

Metabolomics databases

Currently, we can find several sources for metabolic data, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) and MetaCyc. KEGG is a knowledge base for systematic analysis of gene functions in terms of networks of genes and molecules. Its major component is the PATHWAY database, that consists of graphical diagrams of biochemical pathways including most of known metabolic pathways and some of the known regulatory pathways. It also contains information about orthologous and paralog gene groups among different organisms. It has been developed since 1995 and, in recent years, their efforts have been focused on capturing and representing knowledge on diseases as a perturbed state of the molecular network and drugs as perturbants to the molecular network (Kanehisa and Goto, 2000; Ogata et al., 1999). On the other hand, another important database is MetaCyc that contains a collection of metabolic pathways experimentally determined. With more than 2,100 pathways derived from >37,000 publications, it is the largest collection of metabolic pathways available. Pathway reactions are linked to one or more well-characterized enzymes, and both pathways and enzymes are annotated with reviews, evidences, codes, and literature citations. It contains the metabolic pathways for 2,205 organisms. We can find species belonging to the six kingdoms. The species with the highest number of experimentally elucidated pathways in MetaCyc are *Arabidopsis thaliana* (328 pathways), *Escherichia coli* (312), *Homo sapiens* (229), and *Saccharomyces cerevisiae* (172).

But still a large number of metabolites and metabolic pathways remain unknown, and many reaction steps are still missing, even in well-known pathways. For example, it is estimated that at least 1,060,000 metabolites are produced within plants, for which most chemical transformations remain to be identified. The elucidation of potential metabolite pathways in plants would provide a significant benefit for environmental, agricultural, pharmaceutical and

public health matters. But not only in plants most of their metabolomes are unknown. Those for many other organisms are completely unknown as well. Since experimental determination of metabolic pathways is difficult, expensive, and time consuming, even well-investigated species, like *Homo sapiens*, still have unknown or poorly known metabolic pathways.



Graphic 2: Evolution of the number of metabolites available for each organism in Metacyc Database.

Metabolome prediction

For metabolite detection, Mass Spectrometry (MS) is one of the most popular methods for its advantageous sensitivity and throughput factors. It generates a vast amount of data. However, the identification of the metabolites is a bottleneck, causing the finalization of the data analysis to take quite a substantial amount of time. As result of this long analysis period, some experimental data may be lost or become difficult to trace (Ara et al., 2015). Thus, there is a strong need to develop *in silico* methods to infer unknown but

possible metabolites and metabolic pathways. Automatic pathway reconstruction on a metabolome scale is a challenging issue in current computational biology (Kotera et al., 2014, 2013).

A variety of computational methods for reconstructing metabolic pathways have been developed. We can divide them in two categories: the traditional *in silico* methods for metabolic pathway reconstruction from a reference pathway, and the '*de novo*' reconstruction methods which have been developed to elucidate novel reactions based on metabolite chemical structures, known enzymatic reactions, and possible chemical transformations.

The most common *in silico* methods are those developed from a "*reference-based framework"*. In this framework, many known pathways are collected from literature to construct a combined pathway, named the '*reference pathway*', that considers only chemical transformation without distinguishing the difference between organisms. For an organism of interest, enzyme genes are assigned to appropriate positions in predefined reference pathways based on orthologous information about genes across different species. This methodology is limited for the available genome information of organisms and the set of genes, enzymes, and metabolites associated to pathway information. Another common approach is to consider chemical structures to identify pathways that conserve atoms from the original to the target compound in predefined pathway diagrams. Since these approaches reconstruct metabolic pathways from predefined pathways, they are unable to identify previously unknown pathways.

Conversely, *de novo* pathway metabolome reconstruction approaches have been developed to elucidate novel reactions based on metabolite chemical structures, known enzymatic reactions, and possible chemical transformations. They can be categorized into "*compound-filling framework"* and "*reaction-filling framework"*. The former predicts pathways by hypothesizing intermediate compounds necessary between the original and target compounds, whereas the latter

predicts pathways by filling in reactions among many existing compounds at a time. These methods are computationally very expensive, so large-scale prediction is not yet computationally feasible because of prohibitive computational burden (Kotera et al., 2014, 2013).

Many groups have developed reference-based techniques for predicting metabolic pathways of an organism from its genome and producing integrated pathway-genome databases that model the resulting predictions. Some of them are, for example, the previously mentioned KEGG project and BioCyc. KEGG project used this methodology to reconstruct portions of the metabolic pathways by using the reference biochemical knowledge (Bono et al., 1998). On the other hand, BioCyc offers a component called PathoLogic* that uses the MetaCyc database to predict the metabolic-pathway complement of an organisms from its genome (Karp et al., 1999). It contains a collection of more than 3,000 organism specific Pathway/Genome Databases (PGDB), each containing the genome and predicted metabolic network including metabolites, enzymes, reactions, metabolic pathways, predicted operons, transport systems, and pathway-hole fillers (Caspi et al., 2014).

Part II: Objectives

This PhD thesis started with the aim of studying the activity of exogenous (especially from herbal medicines) and endogenous chemicals on biological systems and to understand their effects. We started studying the mechanisms of action of medicinal plants from traditional Catalan medicine and followed with metabolites. The main objectives of this Thesis can be summarized as follows:

i) To elucidate the mode of action of traditional medicines

ii) To compare the polypharmacology between different bioactive compounds, including drugs, synthetic chemical libraries and natural products.

iii) To study the completeness of metabolic databases

iv) To develop a reconstruction protocol of organism metabolomes from their genome and validate it on *Mycoplasma pneumoniae*

v) To explore the interference between endogenous human metabolites and exgoneous compounds (drugs and food ingredients)

The first objective was accomplished on Chapter III.2, where virtual profiling methodology allowed studying the activity of herbal compounds and the identification of those responsible of therapeutic effects. Documentation accumulated during investigation on herbal medicines allowed the elaboration of a review on the topic in Chapter III.1. In Chapter III.3 we compared the polypharmacology of several bioactive compounds, which was followed by an analysis of the data included in metabolic databases, HMDB, KEGG and BioCyc. The data comparison between these databases let to achieve the third objective. Following with metabolic data completeness, it was elaborated a genome-scale metabolome reconstruction framework able to predict organism metabolomes from genomic data. It was applied in Chapter III.4 on *Mycoplasma pneumoniae* and other common species from several kingdoms in Chapter III.5,

including *Escherichia coli, Saccharomyces cerevisiae, Homo sapiens* and *Arabidopsis thaliana*. Finally, in Chapter III.6 and III.7, with the knowledge about metabolome, we studied its interference on exogenous compounds, including drugs and food intake compounds. We infer on metabolites relation with side-effects and drugs efficacy, and the impact of diet on human metabolome.

Part III: Result

III.1: Closing the gap between therapeutic use and mode of action in remedial herbs

Joaquim Olivés, and Jordi Mestres*

Systems Pharmacology, Research Programme on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute and University Pompeu Fabra, Parc de Recerca Biomèdica, 08003 Barcelona, Catalonia, Spain

Abstract

The ancient tradition of taking parts of a plant or preparing plant extracts for treating certain discomforts and maladies has long been lacking a scientific rationale to support its preparation and still widespread use in several parts of the world. This work presents a systems approach to generate mechanistic hypotheses for some of the therapeutic uses of remedial herbs. Both retrospective confirmation and prospective validation of some of the mechanistic hypotheses generated provide proof-of-principle for the validity of the approach.

* Corresponding autor. *E-mail address:* jmestres@imim.es

1. Introduction

Plant leaves, roots, barks, and extracts have been used since the dawn of human history to treat various discomforts and maladies. The healing properties of remedial herbs were most likely identified through a long and serendipitous learning process that once acquired was carefully passed through generations. Still today, traditional medicines represent a well-established therapeutic alternative to synthetic drugs in vast parts of the world (Tao et al., 2014). However, there is still a profound lack of understanding about the concrete chemical ingredient(s) and the exact mechanism(s) of action by which medicinal plants exert their therapeutic effect.

In recent years, global efforts to generate, collect, store, and make publicly available data connecting plants with their chemical ingredients, interacting proteins, and disease indications have set the ground to develop novel systems approaches to unveiling the mode of action of remedial herbs (Liu et al., 2013). This is schematically illustrated in **Fig. 1**. Most current ethnomedicinal studies focus on which parts of the plant are used to treat common ailments (Chassagne et al., 2016). Initiatives to identify and isolate some of the chemical structures present in those parts of therapeutic interest are still expensive and inefficient. This notwithstanding, it is estimated that approximately 50,000 endogenous plant metabolites have been already identified (Hounsome et al, 2008).

Affinity data between plant metabolites and human proteins are scarce to find in public repositories (Bolton et al., 2008; Gaulton et al., 2012). Therefore, most efforts to close the gap between therapeutic use and mode action in remedial herbs are needed in this direction (represented as a dotted line in **Fig. 1**). One option is to process large libraries of isolated small molecules from plants through *in vitro* high-throughput screening assays to identify affinities for therapeutically relevant proteins. This is a highly tedious and expensive

endeavour if one wants to be comprehensive. Alternatively, modern state-of-the-art computational methods to predict the affinity of small molecules across thousands of proteins can be applied to prioritize any further *in vitro* testing of selected small molecules on particular proteins (Vidal et al., 2010; Garcia-Serna et al., 2015).

The last step involves connecting those confirmed interacting proteins with the actual disease for which the plant is prescribed. This task is now facilitated by the recent construction of databases connecting human genes with diseases (Piñero et al., 2015). The aim of this work is to collect and integrate all pieces of data and processes that allow for automatically generate mechanistic hypotheses for the known therapeutic uses of plants. Details on how data was collected and stored into an integrated database are given next followed by some examples of retrospective confirmation, as well as prospective validation, that serve as proof of concept for the entire approach.

2. Materials and Methods

2.1. Linking plants to diseases

A very first version of the database was created containing the therapeutic uses of plants used mainly in traditional Catalan medicine (Gausachs, 2007). Plants were stored using their scientific name in Latin, whereas therapeutic uses were mapped to their corresponding disease identifier in ICD-10 (International Classification of Diseases Version 10). This initial database was complemented with additional therapeutic uses found for those plants in other public sources (Bonet et al., 1999; Raja et al., 1997; Rigat et al., 2007). Data from the different sources was integrated using Latin names for plants and ICD-10 identifiers for diseases from 18 categories. In total, 372 medicinal plants associated with 187

therapeutic uses (diseases) were collected at this stage.

2.2. Linking plants to molecules

The chemical composition of every single plant in the database was extracted from three different sources, namely, Dr Duke Phytochemical and Ethnobotanical Database (Duke, 2016), the KNApSAcK database (Nakamura et al., 2014) and Gausachs' work on Catalan remedial herbs (Gausachs, 2007). Among those, only KNApSacK contains chemical structures linked to chemical names. Structures for the chemical names available only in the other two sources were extracted from PubChem (Bolton et al., 2008) The final set of chemical structures from all sources was unified and stored using InChI Keys. In the end, a total of 7,443 unique chemical structures present in those 372 medicinal plants could be gathered and added to the database.

2.3. Linking proteins to diseases

Next, a list of both known and explored human proteins associated with diseases was extracted from the Therapeutic Target Database (Zhu et al., 2012). These data was complemented with curated protein-disease links, with focus on cardiovascular diseases, available in the literature (Cases and Mestres, 2009). The final list of proteins was stored and unified using their UniProt identifies (The Uniprot Consortium, 2015). A final number of 724 unique proteins known to be relevant for 166 out of the initial 187 diseases were ultimately entered into the database.

2.4. Linking molecules to proteins

Finally, affinity data (pK_i, pK_d, pIC_{50}, pEC_{50}) between the chemical structures

and proteins was extracted from various public sources (Gaulton et al., 2012; Bolton et al., 2008; Gilson et al., 2016; Southan et al., 2016; Roth et al., 2004). Up to 9,342 known interactions between 282 molecules, present in 193 plants, and 170 proteins were identified and collected into the database at this stage. In addition, since affinity data are well recognised to be suffering from completeness issues (Mestres et al., 2008; 2009), known interactions between molecules and proteins were complemented with high-confidence predictions obtained using ligand-based computational models implemented in the CT-link software (Garcia-Serna et al., 2015). Accordingly, 12,000 additional predicted interactions between 1,353 molecules, present in 223 plants, and 246 proteins were generated. In the end, affinity data for a total of 21,305 interactions between 1,353 endogenous molecules, present in 223 plants, and 246 therapeutically-relevant proteins were assembled and stored in the database.

2.5. Experimental in vitro assays

For the prospective validation, two molecules and one herbal extract were selected for testing with *in vitro* assays at Cerep (Cerep Inc.) . Ribosylzeatin was tested in binding assays to confirm the predicted interactions with adenosine A1 and A3 receptors. Cellular assays were used to confirm the predicted interactions between isorhamnetin and the dopamine D4 receptor, as well as between a bilberry extract and 5-lipoxygenase.

For the binding assay, ribosylzeatin was tested twice at a test concentration of 10 μM. The reference agonist ligands used to calculate the compound activity were CPA for the adenosine A1 receptor and IB-MECA for the adenosine A3 receptor, which have IC_{50} values of 0.75 nM and 0.31 nM, respectively. The adenosine A1 receptor assay was performed in the presence of 1 nM of [3H]CCPA. After 60 min. of incubation with shaking, bound radioactivity was separated from free by vacuum filtration and determined scintillation counting.

A similar procedure was followed for the adenosine A3 receptor assay. In this case, it was performed in the presence of 0.15 nM of [125I]AB-MECA. It was incubated with shaking during 120 min. After that, bound radioactivity was filtered and measured with scintillation counting. For these binding assays the results are expressed as a percent of measured specific binding relative to control specific binding.

Fot the dopamine D4 receptor assay, isorhamnetin was tested at a concentration of 10 μM. The reference agonist ligand was dopamine, with an EC_{50} value is 28 nM. D4.4 was incubated for 10 min. at 37ºC and, after that, cAMP was detected and measured with HTRF. Results are expressed as a percent of measured response relative to control response.

Finally, for the 5-lipoxigenase enzyme assay, the bilberry extract was tested at a concentration of 10 μM. The reference compound used was NDGA, which have an IC_{50} of 910 nM. 5-lipoxigenase was incubated 20 min. with shaking and 25 μM arachidonic acid as substrate. After it, rhodamine 123 was measured using fluorimetry. Results are expressed as a percent of measured specific binding relative to control specific binding.

Compounds showing an inhibition or stimulation higher than 50% were considered to be active for the proteins tested, whereas interactions showing activity values between 25% and 50% were considered to be indicative of at least weak to moderate effects.

3. Results and Discussion

Among the 372 medicinal plants present in our database, *Sambucus nigra* (black elder) is the plant associated with the highest number of therapeutic uses, 31. It is used for the treatment of many different illnessess, such as bronchitis, migraine, diarrhoea, nausea, hyperuricemia and influenza. Genus *Sambucus*

belongs to Caprifoliaceae and includes eighteen species all over the world, six of them distributed in subtropical areas of America, Eurasia and Africa. In Catalonia, we can find *Sambucus nigra* and *Sambucus ebulus,* whose leaves, flowers and berries are traditionally used for several medicinal applications in many countries of the world (Dulf et al., 2013; Mahmoudi et al., 2014) Some other plants with a high number of therapeutic uses are *Allium sativum* (24)*, Rosmarinus officinalis* (22), *Mentha spicata* (21), *Urtica dioica* (21), *Salvia officinalis* (21) and *Thymus vulgaris* (21), all of them found easily in the Catalan countryside, and used as food and/or spice in many other countries.

On the other hand, if we focus on cardiovascular diseases, a set of 169 plants are associated with 48 different therapeutic uses. For illustrative purpose, the network of plants linked to cardiovascular diseases is shown in **Fig. 2**. Among plants, *Ginkgo biloba* is the plant with the most cardiovascular uses, 15, by *Crataegus monogyna, Aesculus hippocastanum*, and *Vitis vinifera* with 7, and C*amellia sinensis* and *Allium cepa* with 6. On the other hand, among diseases, hypertension, hypercholesterolemia, hyperglycemia, and haemorrhoids are clearly the cardiovascular aspects being most addressed by remedial herbs.

Ginkgo biloba, Camellia sinensis, and *Aesculus hippocastanum* are not native Catalan plants per se. However, as in many other parts of the world, they are cultivated and used also often as ornamental plants. *Ginkgo biloba* and *Camellia sinensis* are indigenous plants from Asia (Cybulska-Heinrich et al., 2012; Moore et al., 2009). The extracts of the leaves and nuts from *Ginkgo biloba* have been used for hundreds of years to treat a wide variety of disorders, such as asthma, vertigo, tinnitus, as well as general circulatory problems (Cybulska-Heinrich et al., 2012). *Camellia sinensis* is a plant from which green tea can be produced. This beverage has a long traditional use as social drink but also as medicine in the treatment and prevention of disorders, dysfunctions, or diseases in humans and other animals (Moore et al., 2009; Batista et al., 2009). *Aesculus hippocastanum*, horse chestnut, is native to the countries of the Balkan Peninsula, but it is

cultivated worldwide for its beauty. Historically, seed extracts from this plant have been used as a treatment for many ailments ("Aesculus hippocastanum (Horse chestnut). Monograph," 2009).

On the other hand, *Crataegus monogyna* (hawthorn), *Vitis vinifera* (grapevine) and *Allium cepa* (onion) can be found in Catalonia in the wild. The first one is also known as a traditional medicinal plant in many countries, growing in shrub communities and decidious thin forests (Öztürk and Tunçel, 2011). About grapevine, it is an indigenous plant from southern and Western Asia, but it is cultivated today in all temperature regions of the world (Nassiri-Asl and Hosseinzadeh, 2009). Finally, *Allium cepa* (onion) is one of the most important vegetables worldwide and is extensively cultivated. It is a herbaceous bulbous plant that has a long tradition of being beneficial against inflammation, general cardiovascular diseases, and cancer (Slimestad et al., 2007).

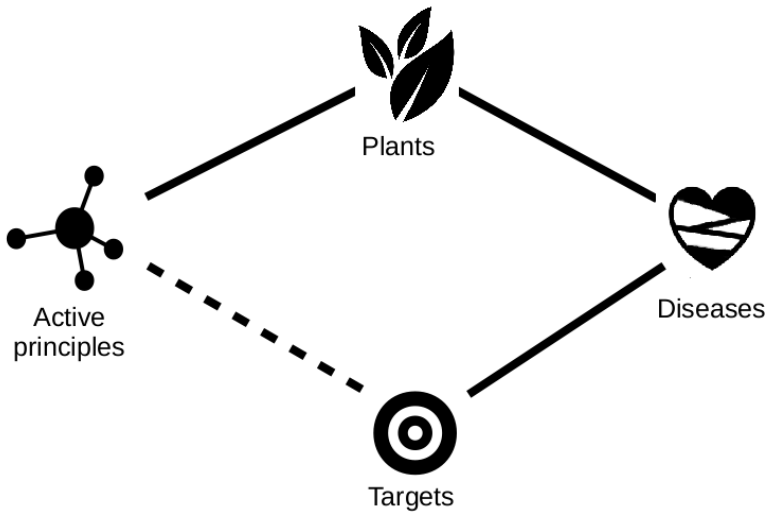


Figure 1. Scheme showing the process of closing the gap between the therapeutic use of plants and the protein targets predicted from the active principles extracted

About completeness of knowledge on chemical composition, *Camellia sinensis* (green tea) is the plant with the highest number of chemical structures identified, 710, followed by *Zea mays* (maize) and *Panax ginseng* (ginseng), with 677 and 601 chemical structures, respectively. None of them is naturally found in Catalonia, but they are certainly cultivated. Among the autochtonous plants with the highest number chemical structures identified we found *Citrus sinensis* (orange tree), *Apium graveolens* (celery), and *Daucus carota* (carrot), 589, 533, and 507 molecules, respectively. In contrast, many plants in the dabase have only one or very few endogenous metabolites identified, such as *Rhamnus alaternus* (Mediterranean Blackthorn), *Lonicera etrusca* (honeysuckle), or *Hernaria glabra* (herniaria).

A detailed analysis of the inter-links between plants, molecules, proteins, and therapeutic uses in our database (**Fig. 1**) identified a total of 21,305 mechanistic hypotheses. A mechanistic hypotheses is generated if a given plant known to have some therapeutic use contains a chemical structures that is either known or predicted to interact with a human protein associated with its original therapeutic use. The plant with the highest number of mechanistic hypotheses for their therapeutic uses is *Glycine max* (soybean). It is followed by*, Vitis vinifera* (grapevine)*, Ginkgo biloba, Citrus limon*, and *Camellia sinensis,* with 713, 712, 693 and 584, respectively. Out of the 21,305 mechanistic hypotheses, 9,342 involve known molecule-protein interactions, wherease 13,963 of them involve predicted interactions. Thus, while the former will be used to perform retrospective confirmations, the latter will be used to pursue some prospective validations.

3.1. Retrospective confirmations

The plants with the highest number of mechanistic hypotheses generated are *Sambucus nigra* and *Allium sativum*, with 25 out of 31, followed by *Ginkgo biloba,*

Thymus vulgaris, and *Allium cepa*. Among the therapeutic uses for which mechanistic hypotheses were generated on the basis of experimentally confirmed data and associations, we identified some well known active

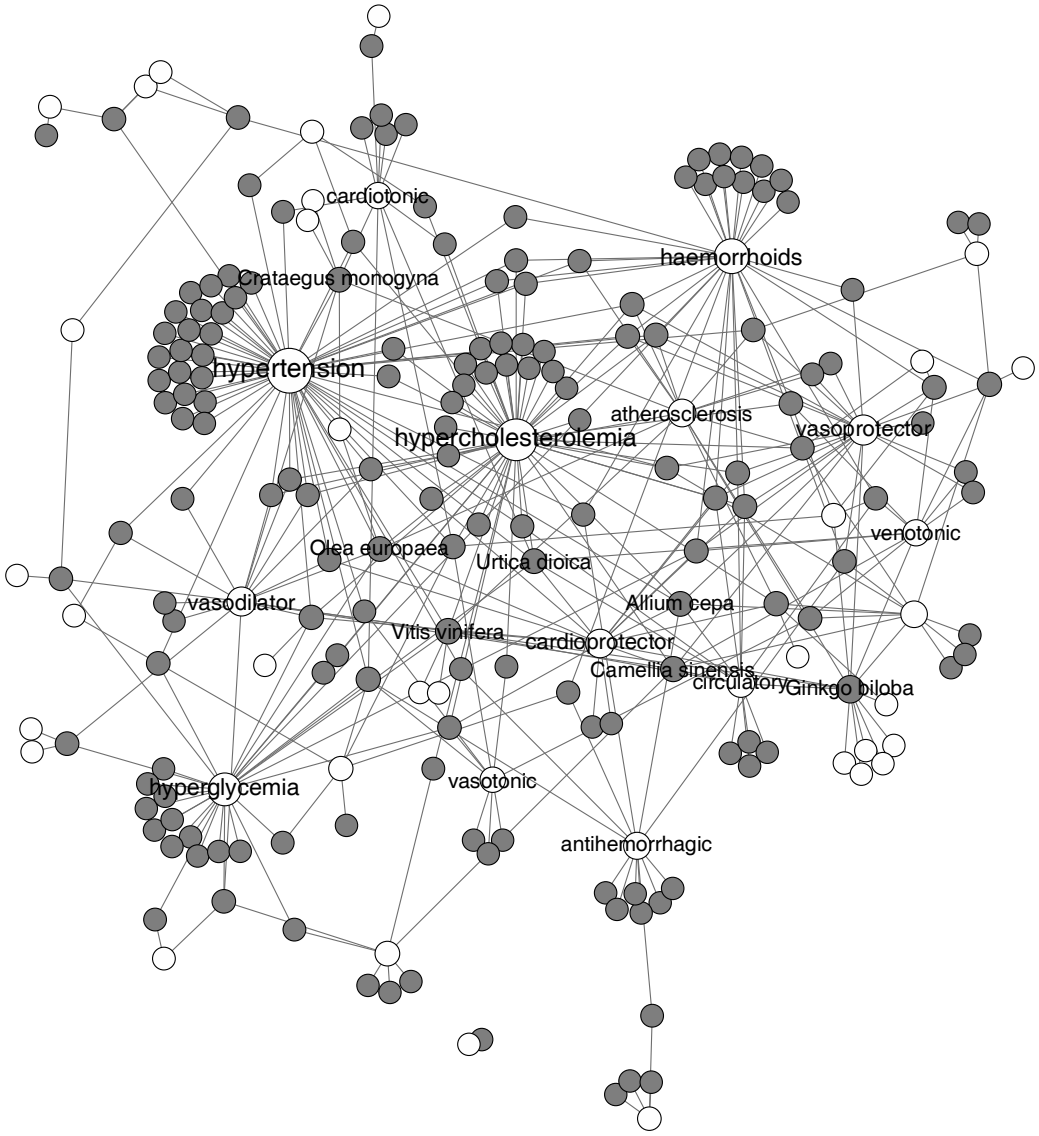


Figure 2. Network of remedial herbs (green circles) linked to therapeutical applications in cardiovascular diseases (white circles).

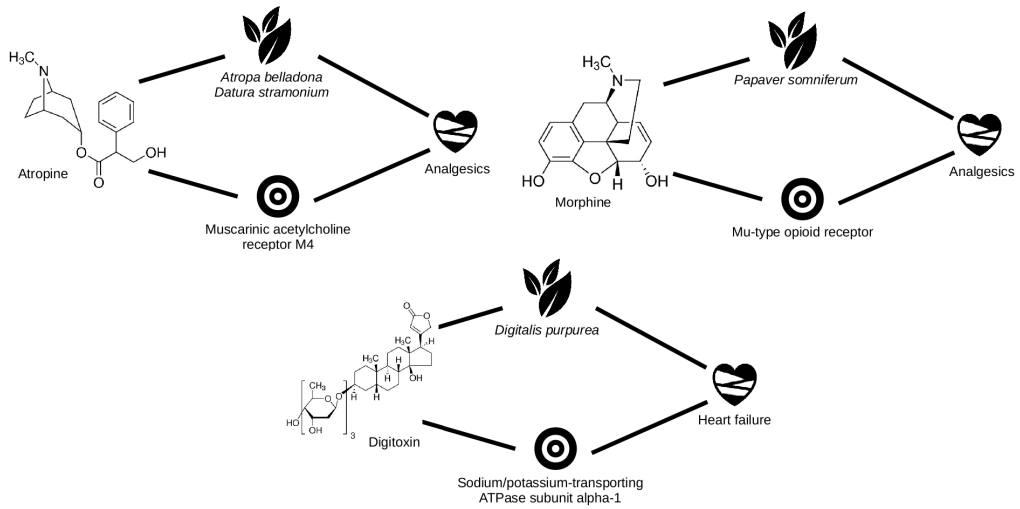


Figure 3. Scheme showing some of the closed circles confirmed retrospectively for three plants used as cardiovascular remedies.

principles such as atropine, morphine, and digitoxin (**Fig. 3**).

One example is atropine, found mainly in *Atropa belladonna* and *Datura stramonium* (Caksen et al., 2003; Kurzbaum et al., 2001), both used commonly for their analgesic action (Duttaroy et al., 2002; Gausachs, 2008a; Overington et al., 2006). This molecule is known to be active against the muscarinic acetylcholine receptor M4, a therapeutic target associated with some analgesics. Therefore, we have all links confirmed and forming a mechanistic hypothesis for the analgesic action of these plants (Owais et al., 2014; Soni et al., 2012.

A second example is morphine, a compound with reported analgesic activity. It is found in *Papaver somniferum* (opium poppy) and it was the first active alkaloid extracted from this plant (Jurna, 2003). Opium has been used in traditional medicinal as sedative and analgesic (Calixto et al., 2001, Gausachs, 2008b). According to all links established in our database, morphine emerges as a candidate for the analgesic action of opium through its interaction with Mu-type opioid receptor (Choi et al., 2006; Yamada et al., 2006), a receptor well

recognized to be associated with analgesia (Inturrisi, 2002).

And a third selected example is digitoxin, a glycoside with activity against the sodium/potassium-transporting ATPase subunit alpha-1, a protein associated with heart failure (J. J. Chen et al., 2001; Hauck et al., 2009; Müller-Ehmsen et al., 2002). Digitoxin is found is *Digitalis purpurea* (J.-J. Chen et al., 2001), a plant used in traditional medicine for treating precisely this disease (Gausachs, 2008a). Digitoxin has not only been proven to indeed interact with the sodium/potassium-transporting ATPase subunit alpha-1, but it also has been show to be effective in heart failure (Belz et al., 2001).

Some other, a bit more speculative, examples of mechanistic hypotheses generated directly from connecting known data for other herbal therapeutic uses could be the therapeutic effect of *Salvia officinalis* for the treatment of Alzheimer's disease and the use of *Achillea millefolium* for treating depression. *Salvia officinalis* has been shown to have some beneficial effect in Alzheimer's disease (Obulesu and Rao, 2011). One of the compounds present in this plant ellagic acid (Gašić et al., 2015), that has been shown to be active against Casein kinase II subunit alpha, a protein associated to Alzheimer's disease (Perez et al., 2011; Rosenberger et al., 2016). In addition, *Achillea millefolium* is traditionally used for treating depression. Multiple proteins have been associated to this disease, Monoamine oxidase A being one of them (Thase et al., 1995). Our analyses reveal that three compounds from yarrow, namely, quercetin, luteolin, and apigenin have indeed biologically relevant affinities for this enzyme (Benetis et al., 2008; Lemmens-Gruber et al., 2006; Bandaruk et al., 2014; Han et al., 2007).

Among the disease categories, the circulatory system is the one for which the highest number of mechanistic hypotheses were generated, 231, followed by the respiratory and musculoeskeletal system. Within these categories, a number of diseases were selected for close inspection, namely, analgesia, cough,

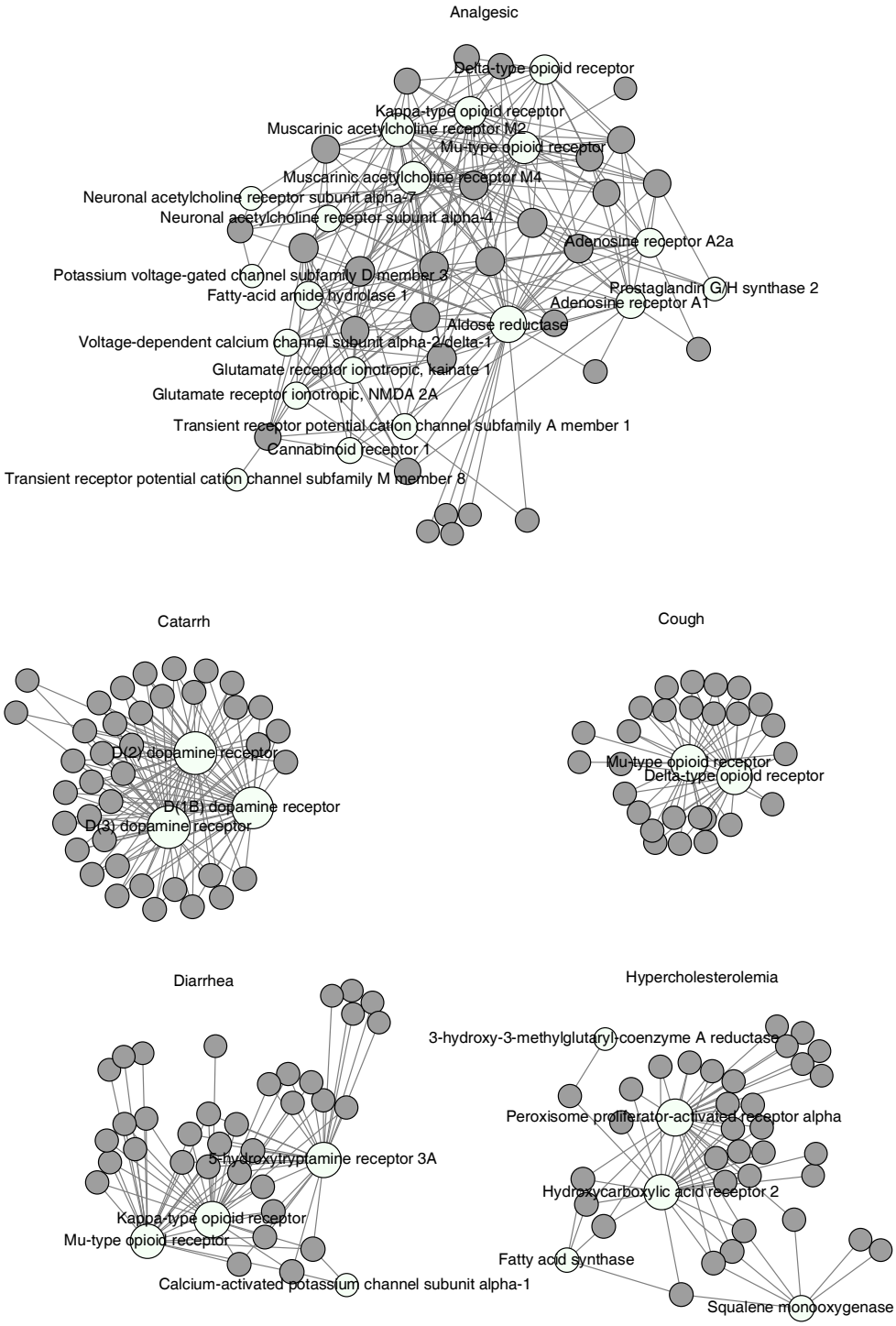


Figure 4. Network of remedial herbs (green circles) linked to proteins (orange circles) associated with various maladies, namely, analgesia, cough, hypercholesterolemia, catarrh, and diarrhea.

hypercholesterolemia, diarrhea, and catarrh. One ought to bear in mind that, even though plants may be composed of chemicals targeting common proteins, the ultimate mechanistic hypotheses of each plant for addressing each disease may be essentially different.

The plant-protein network emerging for the selection of five therapeutic uses of plants is provided in **Fig. 4**. For analgesia, we can see that many different proteins associated with this disease are being targeted by at least one of the endogenous metabolites found in those plants, the most among them being aldose reductase (Young et al., 1983, p. 198), muscarinic acetylcholine receptors (Duttaroy et al., 2002), and opioid receptors (Inturrisi, 2002). About plants being used for treating catarrh and cough, it can be observed that almost all of them are targeting the same targets. For catarrh, all plants contain some chemical that is active on the dopamine D1B. D2, and D3 dopamine receptors, all of them associated with respiratory diseases (Birrell et al., 2002). For cough, all plants contain at least one chemical with affinity for the μ and δ opioid receptors (Kotzer et al., 2000). On the other hand, for diarrhea and hypercholesterolemia, different mechanistic hypotheses are retrieved for different plants linked to these therapeutic uses. For Diarrhea, the list of proteins comprises the μ and κ opioid receptors (Callahan, 2002), the 5-hydroxytryptamine 3A receptor channel (Sikander et al., 2009), and the calcium-activated potassium channel subunit alpha-1 (Deng et al., 2015). The first three proteins are the most targeted proteins. Even though many plants are targeting all them, others seem to target only one or two. Similarly, most of the plants used for treating hypercholesterolemia contain at least one active ingredient on the peroxisome proliferator-activated receptor α (Rimando et al., 2005) and the hydroxycarboxylic receptor 2 (Karpe and Frayn, 2004). However, other plants may be exerting their therapeutic effect through interactions with other proteins associated to this disease, such as fatty acid synthase (Marseille-Tremblay et al., 2007) and squalene monooxygenase (Belter et al., 2011).

It is important to highlight at this stage that the use of known data only is prone to the effects of completeness and thus, several links may actually be missing in the networks discussed. In fact, looking at the distribution of the affinity values for those known interactions, we observe that in many cases these interacting chemicals are found also in plants that are not used for treating the disease associated with the interacting protein. Some of the reasons why these plants have not been used for these illnessess could be, for example, that the compound concentration is not enough in the plant or the plant really has this therapeutic action but it simply is not used for it. Another possible reason could be that we have focused on the therapeutic use of these plants in Catalonia, while the traditional uses of those plants can be different in other regions. Last but not least, it could well be that the action of some of these compounds require the presence of some bioenhancer in the plant as well (Dudhatra et al., 2012), being the therapeutic action a result of multiple compounds acting synergistically. Overall, from the initial number of 372 plants associated with at least one therapeutic use, only 193 contain known data for all necessary links to derive a mechanistic hypothesis. Therefore, adding predictions and thus, improving the lack of completeness, could generate mechanistic hypotheses for the remaining plants.

3.2. Prospective validations

Before embarking into the analysis of the mechanistic hypotheses emerging from predicted interactions, we validated the accuracy of those predictions for which known data was available. Overall, a good correlation was found between known and predicted affinity values for the same molecule-protein interactions. As can be observed in **Fig. 5**, the median of the difference in affinities was 0.332, with 25% and 75% quartiles being at -0.1 and 0.7 with respect to the median, respectively, with a standard deviation of 0.794.

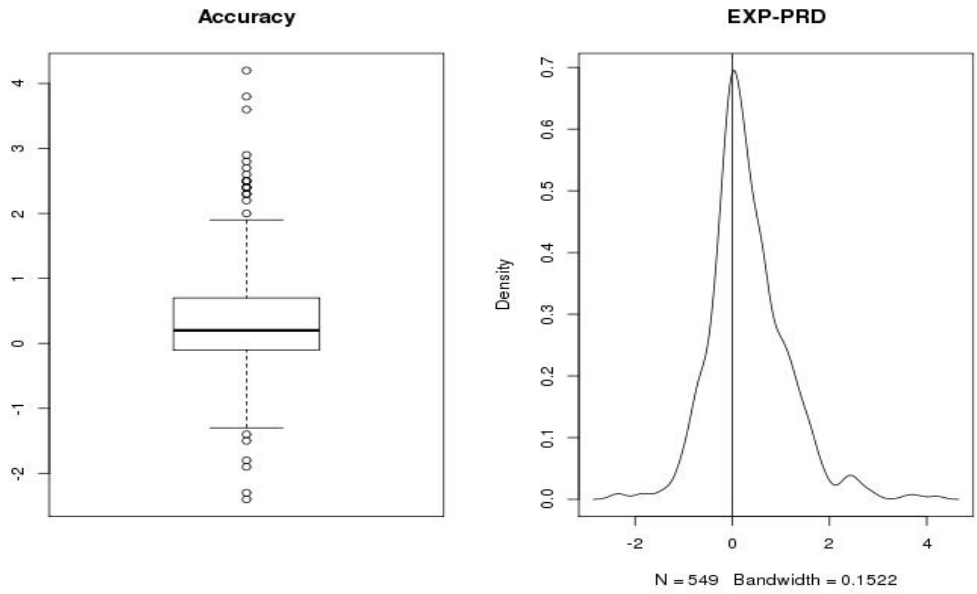


Figure 5. Boxplot (left) and density (right) of the accuracy between known and predicted affinities.

From all mechanistic hypotheses generated on the basis of predicted molecule-protein interactions, a focused set of interactions to be confirmed *in vitro* was selected based on a balance between potency of the predicted affinity and novelty of the prediction, as regarded by the similarity to the closest molecule for which the affinity for the same protein is known already (**Fig. 6**). Among those, we prioritsed the confirmation of the proposed mechanistic hypotheses for two single compounds, namely, isorhamnetin and rybosylzeatin, and one compound mixture, composed of cyaniding, delphinidin, and malvidin (**Fig. 7**).

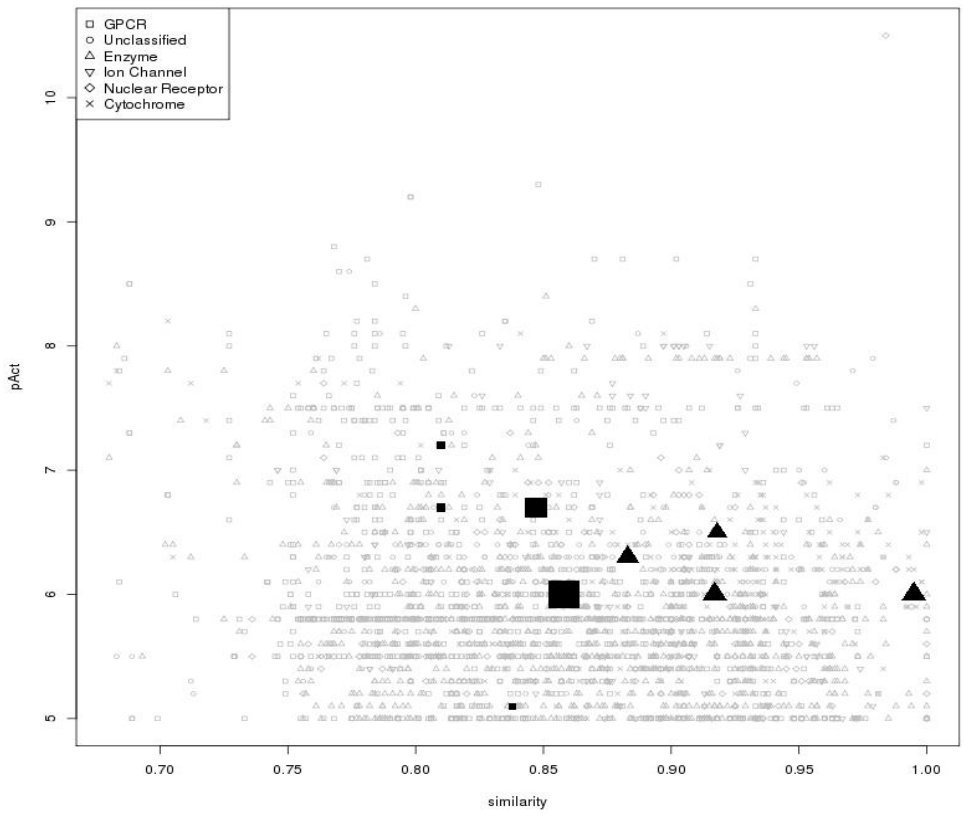


Figure 6. Relationship between the similarity to the closest bioactive neighbor and the predicted affinity (pAct) for all mechanistic hypotheses linked to cardiovascular uses of plants. A total of 2,860 molecule-protein interactions are depicted. The nine interactions selected for *in vitro* testing are highlighted in black.

For ribosylzeatin, we could confirm experimentally the two interactions predicted for the adenosine A1 and A3 receptors, for which 57% and 65% binding, respectively, was obtained at 10μM concentration. Ribosylzeatin is participating in the therapeutic action of *Ginkgo biloba*, *Glycine max*, and *Vitis vinifera*. Accordingly, the mechanistic hypothesis would suggest that, for these plants, the interaction of one of their chemical ingredients (ribosylzeatin) with the adenosine A1 and A3 receptors may be contributing to their beneficial effect in the treatment of a number of cardiovascular diseases,

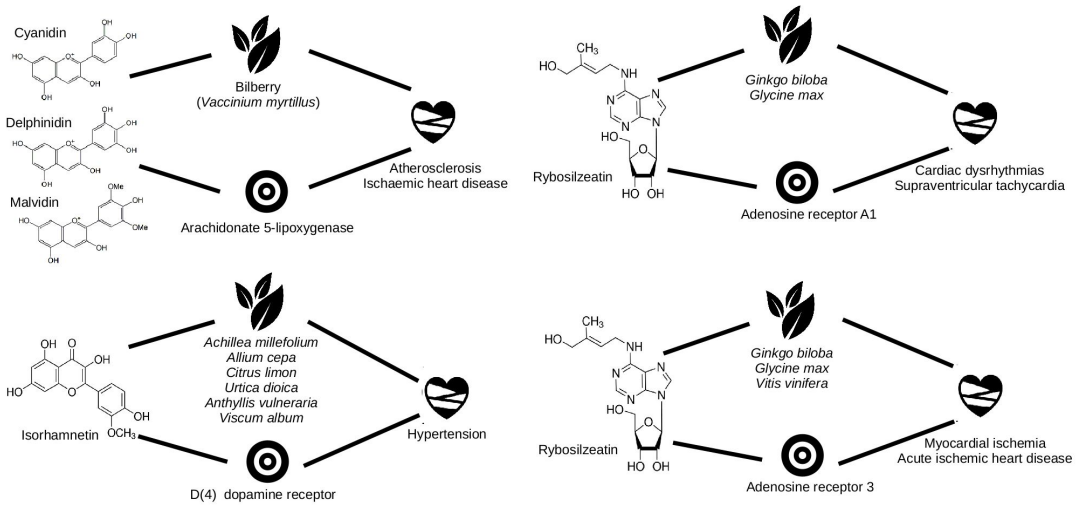


Figure 7. Scheme showing some of the closed circles confirmed prospectively for several plants used as cardiovascular remedies.

namely, cardiac dysrhythmias, supraventricular tachycardia, acute ischaemic heart disease, and mycocardial ischemia by targeting adenosine receptor A3.

It is interesting to note that the therapeutic action of grapevine against ventricular tachycardia has been already demonstrated in rats (Zhao et al., 2010). However, in this case, they only tested a proanthocyanidin grape seed extract. Therefore, on the basis of the hypothesis generated here, we would suggest that ribosylzeatin is participating also on this therapeutic effect in synergy with proanthocyanidins. On the other hand, *Ginkgo biloba* is the sole surviving species of the division Ginkgophyta, that seems to have existed for over 250 million years. It is indigenous to Korea, Japan and China, but nowadays can be found worlwide (Mahadevan and Park, 2008; van Beek and Montoro, 2009). Ginkgo was introduced in Traditional Chinese Medicine on 2,800 BC. For over 5,000 years, the seeds and leaves have been used to treat various diseases, like pulmonary disorders, heart and lung dysfunction and skin diseases. More recenlty, its use has been suggested to address cognitive

deficiencies and dementia. On the other hand, grapevine (*Vitis vinifera*) is a decidious woody climber plant indigenous of Southern Europe and Western Asia, that is cultivated today in all temperature regions of the world (Nassiri-Asl and Hosseinzadeh, 2009). The grape has been used in folk medicine since ancient times, it has diuretic and cardioprotector properties useful for many cardiovascular diseases. The leaves are also used to treat diarrhoea and heal wounds and the sap is used as an antiseptic for eye wash (Delíorman Orhan et al., 2009; Gausachs, 2008b). Finally, soybean is an annual legume of the Fabaceae family. It is indigenous to East Asia and China but now is extensively cultivated in many temperate regions of the world (Munro et al., 2003). Soybean is known as an important source of proteins in diet, but it is also used widely as herbal medicine for the treatment of many diseases such as atherosclerosis and other cardiovascular diseases, depression, obesity or osteoporosis (Gausachs, 2008b).

The predicted activity of the compound mixture composed of cyaniding, delphinidin, and malvidin against arachidonate 5-lipoxygenase was also confirmed experimentally, with 41% inhibition. This results provides a mechanistic hypothesis for the therapeutic use of *Vaccinium myrtillus* in atherosclerosis and ischemic heart disease. It has been already suggested that quercetin is partially responsible for the therapeutic action of this plant due to its affinity for the arachidonate 5-lipoxygenase, but we could add now that delphinidin, cyanidin and malvidin, all of them anthocyanidins, may be also contribute to the action of the plant. Anthocyanidins are compounds present in high concentrations in bilberry fruit (Chu et al., 2011; Ciro Cassinese, 2007). On the other hand, other anthocyanidins present in bilberry, such as peonidin and petunidin, were also predicted to be active against this protein. So all these compounds may actually contribute synergistically to the therapeutic effect attributed to bilberry for the treatment of atherosclerosis and ischemic heart disease.

Finally, the affinity value obtained for the predicted interaction between isorhamnetin and the dopamind D_4 receptor was only 28%. Despite this rather low affinity value, we could suggest that this phytochemical may well be contributing to some extent to the effect of the plants in which it is present for the treatment of hypertension. A chemical present in most of the plants listed and suggested to be responsible for this therapeutic action is quercetin, reported to have also a low experimental affinity value of 5μM against the dopamind D_4 receptor.

4. Conclusions

An effort to integrate data linking plants, molecules, proteins, and diseases has demonstrated to be useful to generate mechanistic hypotheses that provide a scientific basis for some of the therapeutic uses of remedial herbs. In this respect, the use of predicted interactions largely increases our ability to generate mechanistic hypotheses for plants for which known data is scarce. The selected examples that provide retrospective confirmation, as well as those that offer prospective validation, anticipate very good perspectives for the applicability of this type of system approaches for finding a scientific rationale for many traditional medicines. There is much more to learn about nature and how to use it. More research in this direction is underway in our laboratory.

Acknowledgments

This research was supported by the Spanish Ministerio de Economía y Competitividad (project BIO2014-54404-R).

References

Batista, G.A., Cunha, C.L., Scartezini, M., von der Heyde, R., Bitencourt, M.G., Melo, S.F., 2009. Prospective double-blind crossover and study of Camellia sinensis (Green Tea) in dyslipidemias. Arquivos Brasileiros de Cardiologia 93, 128-188.

Bleicher, K.H., Böhm, H.J., Müller, K., Alanine, A.I., 2003. Hit and lead generation: beyond high-throughput screening. Nature Reviews Drug Discovery 2, 269-278.

Bolton, E.E., Wang, Y., Thiessen, P.A., Bryant, S.H., 2008. PubChem: Integrated Platform of Small Molecules and Biological Activities, in: Annual Reports in Computational Chemistry. Washinton, DC.

Bonet, M.A., Parada, M., Selga, A., Vallès, J., 1999. Studies on pharmaceutical ethnobotany in the regions of L'Alt Empordà and Les Guilleries (Catalonia, Iberian Peninsula). Journal of Ethnopharmacology 68, 145-168.

Cases, M., Mestres, J., 2009. A chemogenomic approach to drug discovery: focus on cardiovascular diseases. Drug Discovery Today 14, 479-485.

Cerep Inc. (http://www.cerep.fr).

Chan, E., Tan, M., Xin, J., Sudarsanam, S., Johnson, D.E., 2010. Interactions between traditional Chinese medicines and Western therapeutic. Current Opinion in Drug Discovery & Development 13, 50-65.

Chassagne, F., Hul, S., Deharo, E., Bourdy, G., 2016. Natural remedies used by Bunong people in Mondulkiri province (Northeast Cambodia) with special reference to the treatment of most common ailments. Journal of Ethnopharmacology 191, 41-70.

Cybulska-Heinrich, A.K., Mozaffarieh, M., Flammer, J., 2012. Ginkgo biloba and adjuvant therapy for progressive normal and high tension glaucoma. Molecular Vision 18, 390-402.

De Bolòs, O., Vigo, J., Masalles, R.M., Ninot, J.M., 1993. Flora manual dels Països Catalans. Editorial Portic, Barcelona.

Deliorman Orhan, D., Orhan, N., Özçelik, B., Ergun, F., 2008. Biological activities of Vitis vinifera L. leaves. Turkish Journal of Biology 33, 341-348.

Dickson, M., Gagnon, J.P., 2004. Key factors in the rising cost of new drug discovery and development. Nature Reviews Drug Discovery 3, 417-429.

Duke, J.A. 2016. Dr Duke Phytochemical and Ethnobotanical Database (http://www.ars-grin.gov/duke/).

Garcia-Serna, R., Mestres, J., 2010. Anticipating drug side effects by comparative pharmacology. Expert Opinion on Drug Metabolism & Toxicology 6, 1253-1263.

Garcia-Serna, R., Vidal, D., Remez, N., Mestres, J., 2015. Large-scale predictive drug safety: from structural alerts to biological mechanisms. Chem. Res. Tox. 28, 1875-1887.

Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J.P., 2012. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40, D1100–D1107.

Gausachs, R., 2007. Les herbes hemeieres: de la cultura al fàrmac, una aproximació etnobotànica. 2 volumes. Rafael Dalmau Editor, Barcelona.

Gilson, M.K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., Chong, J., 2016. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 44, D1045-1053.

González-Tejero, M.R., Casares-Porcel, M., Sánchez-Rojas C.P., Ramiro-Gutiérrez, J.M., Molero-Mesa, J., Pieroni A., Giusti, M.E., Censorii, E., de Pasquale, C., Della, A., Paraskeva-Hadijchambi, D., Hadjichambis, A., Houmani, Z., El-Demenashar, M., El-Zayat, M., Hmamouchi, M., Eljohrig, S., 2008. Medicinal plants in the mediterranian: synthesis of the project Rubia. Journal of Ethnopharmacology 116, 341-357

Gregori-Puigjané, M., Mestres, J., 2006. SHED: Shanon entropy descriptors from topological feature distributions. Journal of Chemical Information and Modeling 46, 1615-1622.

Holland, B.K., 1994. Prospecting for drugs in ancient texts. Nature 369, 702

Harvey, A.L., 2008. Natural products in drug discovery. Drug Discovery Today 13, 894-901.

Hounsome, N., Hounsome, B., Tomos, D., Edwards-Jones, G. 2008. Plant metabolites and nutritional quality of vegetables. Journal of Food Science 73, 48-65.

Kong, D.X., Li, X.J., Zhang, H.Y., 2009. Where is the hope for drug discovery? Let history tell future. Drug Discovery Today 14, 115-124.

Krejsa, C.M., Horvath, D., Rogalski, S.L., Penzotti, J.E., Mao, B., Barbosa, F., Migeon, J.C., 2003. Predicting ADME properties and side effects: the

BioPrint approach. Current Opinion in Drug Discovery & Development 6, 470-80.

Mahadevan, S., Park, Y., 2008. Multifaceted therapeutic benefits of Ginkgo biloba L.: chemistry, efficacy, safety, and uses. Journal of Food Science 73, R14-R19.

Melino, V.J., Hayes, M.A., Kathleen, S.L., Ford, C.M., 2011. The role of light in the regulation of ascorbate metabolism during development in the cultivated Vitis vinifera. Journal of the Science of Food and Agriculture 91, 1712-1721.

Mestres, J., Gregori-Puigjané, E., Valverde, S., Solé, R. V., 2008. Data completeness: the Achilles of drug-target networks. Nature Biotechnology 9, 983-986.

Mestres, J., Gregori-Puigjané, E., Valverde, S., Solé, R. V., 2009. The topology of drug-target interaction networks: implicit dependence on drug properties and target families. Molecular BioSystems 9, 1051-1057.

Mestres, J., Seifer, S.A., Oprea, T.I., 2011. Linking pharmacology to clinical reports: cyclobenzaprine and its possible association with serotonin syndrome. Clinical Pharmacology & Therapeutics 90, 662-666.

Moore, R.J., Jackson, K.G., Minihane, A.M., 2009. Green tea (Camellia sinensis) catechins and vascular functions. British Journal of Nutrition 102, 17980-1802.

Mullard, A., 2012. 2011 FDA drug approvals. Nature Reviews Drug Discovery 11, 91-94.

Nakamura, Y., Afendi, F.M., Parvin, A.K., Ono, N., Tanaka, K., Hirai Morita, A., Sato, T., Sugiura, T., Altaf-Ul-Amin, M., Kanaya, S., 2014. KNApSAcK Metabolite Activity Database for retrieving the relationships between metabolites and biological activities. Plant Cell Physiol. 55, e7

Nassiri-Asl, M., Hosseinzadeh, H., 2009. Review of the pharmacological effects of Vitis vinifera (grape) and its bioactive compounds. Phytotherapy Research 23, 1197-1204.

Newman, D.J., 2008. Natural products as leads to potential drugs: an old process or the new hope for drug discovery? Journal of Medicinal Chemistry 51, 2589-2599.

Liu, H., Wang, J., Zhou, W., Wang, Y., Yang, L., 2013. Systems approaches and polypharmacology for drug discovery from herbal medicines: an example using licorice. Journal of Ethnopharmacology 146, 773-793.

Öztürk, N., Tunçcel, M., 2011. Assessment of phenolic acid content and in vitro antiradical characteristics of hawthorn. Journal of Medicinal Food 14, 664-669.

Pan, S.Y., Pan, S., Yu, Z.L., Ma, D.L., Chen, S.B., Fong, W.F., Han, Y.F., Ko, K.M., 2010. New perspectives on innovate drug discovery: an overview. Journal of Pharmacy & Pharmaceutical Sciences 13, 450-471.

Patwardhan, B., Mashelkar, R.A., 2009. Traditional medicine-inspired approaches to drug discovery: can Ayurveda show way forward? Drug Discovery Today 14, 804-811.

Paul, S.M., Mytelka, D.S., Dunwiddie, C.T., Persinger, C.C., Munos, B.H., Lindborg, S.R., Schacht, A.L., 2010. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery 9, 203-214.

Piñero, J., Queralt-Rosiñach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., Furlong, L.I., 2015. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database bav028.

Raja, D., Blanché, C., Vallés Xirau, J., 1997. Contribution to the knowledge of the pharmaceutical ethnobotany of La Segarra region (Catalonia, Iberian Peninsula). Journal of Ethnopharmacology 57, 149-160.

Rigat, M., Bonet, M.A., Garcia, S., Garnatje, T., Vallès, J., 2007. Studies on pharmaceutical ethnobotany in the high river Ter valley (Pyrenees, Catalonia, Iberian Peninsula). Journal of Ethnopharmacology 113, 267-277.

Rollinger, J.M., Shuster, D., Danzl, B., Schwaiger, S., Markt, P., Schmidtke, M., Gertsch, J., Wolber, G., Langer, T., Stuppner, H., 2009. In silico target fishing ligand discovery exemplified on constituents of Ruta graveolens. Planta Medica 75, 195-204.

Roth, B.L., Lopez, E., Beischel, S., Westkaemper, R.B., Evans, J.M., 2004. Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. Pharmacol. Ther. 102, 99–110.

Southan, C., Sharman, J.L., Benson, H.E., Faccenda, E., Pawson, A.J., Alexander, S.P.H., Buneman, O.P., Davenport, A.P., McGrath, J.C., Peters, J.A., Spedding, M., Catterall, W.A., Fabbro, D., Davies, J.A., NC-IUPHAR, 2016. The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. Nucleic Acids Res. 44, D1054-1068

Tao, L., Zhu, F., Qin, C., Zhang, C., Xu, F., Tan, C.Y., Jiang, Y.Y., Chen, Y.Z., 2014. Nature's contribution to today's pharmacopeia. Nature Biotechnology 32, 979-980.

The UniProt Consortium, 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–D212. doi:10.1093/nar/gku989

Schmid, E.F., Smith, D.A., 2004. Is pharmaceutical R&D just a game of chance or can strategy make the difference? Drug Discovery Today 9, 18-26.

Swinney, D.C., Anthony, J., 2011. How were new medicines discovered? Nature Reviews Drug Discovery 10, 507-519.

Verpoorte, R., Crommelin, D., Danhof, M., Gilissen, L.J., Schuitmaker, H., van der Greed, J., Witkamp, R.F., 2009. A systems view on the future of medicine: inspiration from Chinese medicine? Journal of Ethnopharmacology 121, 479-481.

Vidal, D., Garcia-Serna, R., Mestres, J., 2011. Ligand-based approaches to in silico pharmacology. Methods in Molecular Biology 672, 489-502.

Vidal, D., Mestres, J., 2010. In silico receptorome screening of antipsychotic drugs. Molecular Informatics 29, 543-551.

Zhao, G., Gao, H., Qiu, J., Lu, W., Wei, Z., 2010. The molecular mechanism of protective effects of grape seed proanthocyanidin extract on reperfusion arrhythmias in rats in vivo. Biological & Pharmaceutical Bulletin 33, 759-767.

van Beek, TA., Montoro, P., 2009. Chemical analysis and quality control of Ginkgo biloba leaves, extracts, and phytopharmaceuticals. Journal of Chromatography A 1216, 2002-2032.

III.2: Unveiling the mode of action of traditional medicines

Joaquim Olivés, and Jordi Mestres*

Systems Pharmacology, Research Programme on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute and University Pompeu Fabra, Parc de Recerca Biomèdica, 08003 Barcelona, Catalonia, Spain

Abstract

Medicinal plants have been very important along history for the treatment of many diseases. For centuries researchers have been interested in the observation, description, and experimental investigation of these medicines and their biological activites. Today, a huge amount of articles about ethnopharmacology have been published, and many of the data have been collected in several databases. At the same time, many other databases have been developed containin data from different but related fields. This has allowed the development of computational methodologies that take advantge of the available data to obtain information of interest, for example, the elucidation of herbal medicines therapeutic mode of action. In this review we have investigated on the diversity of computational approaches used on ethnopharmacological studies, providing information about the current state of the art.

* Corresponding autor. *E-mail address:* jmestres@imim.es

1 Introduction

Since early human history, natural products has been utilized to treat and prevent diseases, they have formed the basis of most of the modern medicines. In XIX century, with the chemopharmacological revolution, many pharmacological laboratories started to isolate the compounds of many plants in order to extract their active principles, which were responsible of the medicinal or toxicological properties of the plants. The first time that a natural compounds(Jones, 2011) were isolated was on 1804, when F.W.A. Sertürner isolated for first time morphine from opium (*Papaver somniferum*) and in the years following he investigated the effects of this compound. Some years later it started to be produced derived compounds from natural products, the most famous example probably is acetylsalicylic acid, which was produced for the first time in 1853 by the french chemist Charles Frederic Gerhardt and patented in 1899 by Bayer, it was derived from salicin, a natural product isolated from the bark of willow tree (*Salix alba*) (Jones, 2011) .

Over the last century natural products have been the major source of chemical diversity for starting materials while driving pharmaceutical discovery (Mishra and Tiwari, 2011). The investigation of natural compounds reached his peak in the period 1970-1980, after this period, with the advent of combinatorial chemistry technology and other synthetic chemistry techniques, the work with natural compounds went in wane (Newman, 2008). From late 1980s to late 1990s, combinatorial chemistry let to create many libraries containing hundreds of thousands to millions of new compounds. However, these new techniques didn't increase drug productivity and natural products remained as an important source of new drugs, drug leads and new chemical entities (NCE). From 1981 to 2002, natural products and natural product derived and synthetic derived natural product accumulated approximately a 48% of the NCE reported (Balunas and Kinghorn, 2005). They have had a dominant role in anticancer

compounds and drugs for infectious diseases, where 60% and 70%, respectively (McChesney et al., 2007).

Natural product structures have a greater complexity, a high chemical diversity, biochemical specificity, and other molecular properties that make the favourable for drug discovery (Koehn and Carter, 2005). Moreover, as the herbal medicine has been routinely used through the ages and their beneficial and adverse effects should be known, natural products may be safer than synthetic drugs (Cohen & Ernst, 2010).

Nowadays, according to the World Health Organization (WHO), 80% of the world's population still relies on plant-based medicines for primary health ("Traditional Medicine Growing Needs and Potential - WHO Policy Perspectives on Medicines, No. 002, May 2002," 2002). But most of these herbal medicines are based on traditional knowledge rather than evaluation and laboratory, so their safety and efficacy is relatively unknown and herbs with pharmacological activity may also be toxic, especially if they are used incorrectly. For example, recently it was reported a poisoning of *Mandragora officinarum,* a plant used as sedative and to treat some cardiovascular and respiratory system diseases, it is also known as magic, aphrodisiac and hallucinogenic. In this poisoning case report, a man ingested five 'aphrodisiac' berries of mandragora and after 1h of the ingestion he experienced nausea, vomiting, abdominal pain, agitation, aggression, hallucinations, mydriasis, dry mouth and skin, hyperthermia, tachycardia, and increased blood pressure, so, an unappropiated use of a medicinal plant can drive to unwanted consequences (Nikolaou et al., 2012;Gausachs, 2008). Therefore, it is important to identify the mode of action of these medicines in order to develop better medicines from traditional knowledge, potentiating the desired effects and removing the adverse ones.

The little knowledge of efficacy and the unwanted side effects are because the

mode of action of these herbal extracts remains unknown for the vast majority of plants, despite that the chemical composition is well known for a great number of plant species. Moreover, quite often there is not sufficient clinical data about herbal pharmacological effects and molecules responsible of it (Dunnick and Nyska, 2013).

In the recent years, the interest in the identification of effective ingredients of medicinal plants and functioning targets has increased. As result, several databases about medicinal plants and their ingredients have been established, specially, databases centered principally in Traditional Chinese Medicine, such as Traditional Chinese Medicine Database (TCMID) (Xue et al., 2013).

Traditional medicines contains hundreds of compounds and only a few bioactive compounds contribute to the therapeutic effect, one of the strategies to identify the compounds that are responsible of the plants benefit is by using high-throughput screening methods, however they are still very limiting and vastly expensive (Naoghare and Song, 2010). Moreover, the complex composition and polypharmacology of traditional medicines make it even harder to use experimental methods to elucidate multitarget mode of action from a holistic point of view (Zhao et al., 2013).

On the other hand, computer-assisted tools offer cheaper and better methods to predict the mode of action of traditional medicines. They can predict a large number of new drug-target interactions, and allow constructing drug-target networks. These strategies try to predict new drug targets, and to achieve their endeavour they can use several different techniques (C. Huang et al., 2014) .

In this paper, we will investigate the different computational strategies that have been used on these recent years to enquire in the mode of action of traditional medicines.

2 Data sources

To perform any *in silico* study of traditional medicine it is necessary to collect data. Literature is always an available source of data; however, there are many database and sources where it is much easier to retrieve data. These databases can be divided in in four categories according to their content:

2.1 Diseases – Plants

- **International Ethnobotany Database (ebDB)**: it provides a wide variety information about plants, including medicinal uses, plants parts used and locations where the plants growth (Skozen and Bussman, 2006).
- **Natural Products Alert (NAPRALERT)**: It contains ethnomedical information of organisms derived from abstracts and original articles ("NAPRALERT," 2012).
- **Dr Duke Phytochemical and Ethnobotanical Database**: There information about the therapeutic uses of more than 1000 plants ("Dr Duke Phytochemical and Ethnobotanical Database," n.d.).
- **the Traditional Chinese Medicine Database (TCMD)**: It contains 1102 unique herbs and information about their therapeutic uses (He et al., 2001).
- **Herbal Ingredients' Targets (HIT)**: There are more than 1300 entries for reputable Chinese herbs where we can find its functions (Ye et al., 2011).
- **TCMDatabase@Taiwan**: It contains 21 differents classes of traditional medicines therapeutic uses and the medicines found in each functional class. The number of plants in each functional class varies from 3 to 62 (Chen, 2011).
- **Traditional Chinese Medicine Integrated Database (TCMID)**: It contains 8159 herb entries which may provide information about their therapeutic uses (Xue et al., 2013).

- **KNApSAcK:** There are 1432 species entries with information about their biological activities. It containts almost 2000 different activites (Afendi et al., 2012).

2.2 Plant - Ingredients

- **Natural Products Alert (NAPRALERT)**: There is pharmacological/biochemical information of extracts of organisms ("NAPRALERT," 2012).
- **Dr. Duke Phytochemical and Ethnobotanical Databases:** There is the chemical composition of more than 1000 plants. Chemical structures are not available ("Dr Duke Phytochemical and Ethnobotanical Database," n.d.).
- **the Traditional Chinese Medicine Database (TCMD)**: There are 12120 unique compounds for 1102 herbs (He et al., 2001).
- **Herbal Ingredients' Targets (HIT)**: It contains about 586 active compounds from more than 1300 chinese herbs (Ye et al., 2011).
- **TCMDatabase@Taiwan:** It contains 32.364 constituents from 352 different herbs, animal products and minerals (Chen, 2011).
- **Traditional Chinese Medicine Integrated Database (TCMID)**: It contains 8159 chinese herbs with 25210 ingredients, the structure of the compounds may or not be available (Xue et al., 2013).
- **Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform (TCMSP):** This database contains 499 herbs registered in Chinese Pharmacopeia, with a total of 12144 chemicals (Ru et al., 2014).
- **KNApSAcK**: There are 22399 species and 50897 metabolites entries, many of them with structure information. It also contains the biological activity of species (Afendi et al., 2012).

- **Dictionary of Natural Products**: There is information of 170.000 natural products which includes their source data, however, all these data is not available to be downloaded ("Dictionary of Natural Products," 2014).

But electronic databases are not the unique source of information about traditional medicine plants compositions; it can also be collected from books, like *Les Herbes Remeires* of Gausachs, R. (2007), and articles of studies about plant composition.

2.3 Ingredients - Proteins

Herbal ingredients database doesn't use to include the activity of the compounds, and it necessary to search this information in other databases. There are many existing databases with a great amount of information about chemical activities:

- **Herbal Ingredients' Targets (HIT):** There is information derived from more than 3250 literatures, it contains about 1301 known protein targets (221 of them described as direct targets) affected by 586 herbal compounds (Ye et al., 2011).
- **Traditional Chinese Medicine Integrated Database (TCMID)**: This database comprises information about targets of natural ingredients which is collected from different resources, STITCH, HIT and published articles (Xue et al., 2013).
- **Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform (TCMSP):** It contains experimental and predicted drug-target information. Experimental data is retreived from HIT, while predicted data is obtained with SysDT model (Ru et al., 2014).

- **ChEMBL**: It comprises > 1 million compounds and > 12 million of activites (Gaulton et al., 2012
- **PubChem:** It contains the activity and the structure for > 19 million unique chemical structures (Bolton et al., 2008).
- **DrugBank:** It contains 7680 drug entries, which includes some natural products (Law et al., 2014).
- **STITCH:** There is interaction information for over 68.000 different chemicals (Kuhn et al., 2014).
- **The Comparative Toxicogenomics Database (CTD):** This database includes 983049 chemical-gene interactions in 513 organisms, between 105590 chemicals and 35095 genes (Davis et al., 2014)
- **BindingDB:** It contains about 620.000 binding data for 5.500 proteins and over 270.000 drug-like molecules (Liu et al., 2007).
- **PDB:** This database doesn't contain the reported activity of compounds, however, it contains over 100.000 proteins structures, which are necessary for docking studies (Berman, 2000).

2.4 Proteins - Diseases

- **Therapeutic Target Database (TTD):** It provides information about explored therapeutic proteins and the targeted disease. It contains 1535 targets (Zhu et al., 2012).
- **PharmGKB:** It provides pharmacogenomic knowledge, with associations between > 20.000 genes, and 53 pathways and > 3000 diseases (Whirl-Carrillo et al., 2012).
- **Potential Drug Target Database (PDTD):** It contains 1207 entries covering 841 known and potential drug targets with structures from

PDB categorized into 15 types according to their therapeutic areas (Gao et al., 2008).

- **The Comparative Toxicogenomics Database (CTD):** CTD provides almost 30000 gene-disease associations, including 7455 genes and 4912 diseases (Davis et al., 2014).
- **KEGG:** One of the most widely used databases; it contains 372 pathways linked to protein/enzyme (Kanehisa et al., 2014).
- **OMIM:** This database contains information on all known mendelian disorders and ober 12.000 genes. It is focused in relations between phenotype and genotype ("Online Mendelian Inheritance in Man, OMIM," n.d.).
- **Genetic Association Database (GAD):** Archive of human genetic association studies of complex diseases and disorders. Data extracted from published articles and GWAS studies (Becker et al., 2004).
- **Gene Expression Omnibus (GEO):** public repository that archives microarray, next-generation sequencing and other forms of high-throughput functional genomics data submitted by research community (Barrett et al., 2013).
- **ConsesusPathDB-human (CPDB):** It integrates 32 public resources; some of the information information available comprises biochemical pathways linked to proteins (Kamburov et al., 2011).
- **Traditional Chinese Medicine Integrated Database (TCMID)**: Data retrieved from DrugBank and OMIM (Xue et al., 2013).
- **Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform (TCMSP):** It contains disease information retrieved from TTD and PharmaGKB (Ru et al., 2014).

2.5 Side effects

Despite the common use of herbal medicines, it is not always ensured if they are safely. Herbal products contains a huge number of compounds, and it isn't known the activity of all them, this can cause that an unsuitable use of the herbal product may produce undesired effects

Moreover, several drug-drug interactions have interactions have been reported. These interactions can alter the way of one or both of the drugs in the body, or cause unexpected side effects. For example, taking together selective serotonin reuptake inhibitors (SSRI) with non-steroidal anti-inflammatory drugs (NSAID), may increase the risk of bleeding, and the simultaneous administration of ibuprofen and acetylsalicylic acid produce antagonistic interactions, reducing the effects (Cascorbi, 2012).

It is well known that plants contains a great amount of ingredients, with many different activities, some of them may be contributing to the therapeutic use of the plant, but some other could cause undesirable effects. And the number of compounds is even higher when several plants are merged to create a herbal medicine. So, in the same way this plants may treat diseases through a synergistic interactions of their ingredients, they could also produce undesired effects, this is why they are also important those databases about side effects.

- **SIDER**: It contains information on marketed medicines and their recorded adverse drug reactions. There are 4192 side effects and 996 drugs (Kuhn et al., 2010).
- **AERS**: It contains over seven million reports of adverse effects ("AERS," 2014).

2.6 Database Selection Criteria

To select which databases will be used for the studies, researchers must consider some database appreciations. Here are listed some of the most important:

- Availability of the database: Usually researches will try to use those sources that are publicly available.
- How easy is to download data: It is important to use database that can be entirely downloaded in an easy way. For example in chemical databases it is a limitation to have to download structures individually.
- The data included in the database: It must be considered that a database contains all the data we need; it is useless to download a great amount of data with missing necessary information.
- Supported database: supported databases that are being actualized periodically are better.
-

3 Computational methodologies in herbal medicine research

3.1 Computational methodologies to obtain herbal ingredients activity

In herbal medicine research, many different methodologies have been used to collect and predict ligand-protein interactions. These strategies can be divided in methods that collect information from existing databases; and methods that try to predict new targets for the compounds, chemogenomic approaches.

3.1.1 Data mining

Data mining is defined generally as the process of extracting meaningful information from large datasets through the use of any relevant data analysis techniques. These techniques can be applied to extract information from large volumes of text data, called text mining, or from structured databases, which is called data mining (Yang et al., 2013).

3.1.1.1 Data mining

Data mining is a computational method that has been traditionally used to search patterns in structured databases.

3.1.1.2 Text mining

Text mining is a process which comprises the discovery of knowledge from text through the extraction of meaningful patterns and trends. It is intended to explore relationship among the objects stored in unstructured database (Yang et al., 2013).

This method is applied in drug discovery and biomedicine to the identification of entities such as genes or diseases, as well as the identification of relationships between those entities, including protein-protein interactions, disease-associated entities (genes/proteins), diseases-related networks and the interaction of herbal active ingredients and the targets.

Text mining allows finding useful information from an extremely large number of articles, however, despite having access to a large number of papers, often can only be accessed to limited information because the full text is restricted.

3.1.2 Entity grammar systems

Entity grammar systems are a formal grammar tool proposed for study the complex hierarchies in biological systems. Using it as a framework, the formal model of complex biological systems can be easily constructed. It has been successfully used in the modelling, learning and simulation of biological cells (Wang, 2004).

3.1.3 Chemogenomic methods

Chemogenomic approaches for drug discovery consist in ligand-based, target-based and target-ligand methods, which are used to reveal novel relationships between compounds and targets (C. Huang et al., 2014).

These *In silico* methods have their origins in Quantitative Structure-Activity Relationships (QSAR), which consists in the construction of a mathematical model relating molecular structure to a chemical property or biological effect by means of statistical techniques.

QSAR use molecular descriptors as numerical representations of chemical structures. There are a large number of different molecular descriptors that are classified according to the dimensionality of the chemical representation from which they are computed. One-dimensional descriptors encode properties such as molecular weight, molar refractivity and octanol/water partition, offering a fair reflection of the size, shape and lipophilicity of molecules. On the other hand, 2D-descriptors are computed from topological representation of molecules. Finally, 3D-descriptors are obtained directly from 3D structure of molecules (Ekins et al., 2007).

For computing these molecular descriptors, the structure of the chemicals is required. Usually, these chemical structures can be obtained from some herbal

medicines-ingredients databases or chemical databases like TCMSP, PubChem, etc., however, some chemical structures are not available and they have to be constructed in chemical structure drawing program such as MDL ISIS/Draw.

To perform these *in silico* methodologies, one of the most used programs is Discovery Studio, from Accelrys (Accelrys Software Inc., n.d.), it allows performing docking studies, molecular dynamics simulations and generate pharmacophores and molecular descriptors

A. Ligand-Based Approach

The ligand-based approach is also known as the chemical approach, the basic assumption sustaining this method is that similar compounds are expected to have similar affinities for a given target, so, similar compounds should display a similar pharmacological profile (Vidal et al., 2011) **.** The general practices of this approach are describing compounds with chemical descriptors and calculate a similarity coefficient between ligands (Zhao et al., 2013)**.**

There is a diverse range of ligand-based methods with different computational costs depending on the type of structural information that they use. The most commonly used approaches are those using topological fingerprints encoding the presence of substructural fragments.

Similarity Ensemble Approach (SEA) (Keiser et al., 2007) is one of the most used methods to predict human targets; its similarity criterion is the widely used Tanimoto coefficient (Tc). Tc has been also applied in many studies to evaluate drug-likeness of herbal ingredient. Another important approach used to calculate similarity between two molecular compounds is feature-pair distribution (FDP) (Vidal et al., 2011).

On the other hand, regarding methods that require 3D structure representations, the most widely used is pharmacophore model, which have been applied to screen potentian ligangs for many targets. These models can be generated by many programs, like GALAHAD, a Tripos Ltd pharmacophore module, from sets of compounds ("GALAHAD Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144-2319. www.tripos.com. Contact company for pricing information.," 2007). A pharmacophore is defined to be the 3D arrangement of molecular features necessary for bioactivity, and it can be generated by many docking programs such as Glide, from Schrödinger (Friesner et al., 2006), and Discovery Studio (Accelrys Software Inc., n.d.). There is also some available web servers able to identify potential targets using pharmacophore models, one example is PharmMapper, a freely accessed web-server designed to identify potential target for the given molecules using pharmacophore mapping approach.

But pharmacophore screening only considers compounds who are mimics of the ligand from which the pharmacophore was generated, so it may neglect other positive binding modes. This limitation can be avoided by constructing multiple pharmacophore models with different modes of interaction, which is called virtual parallel screening.

B. Target-Based Approach

Target-based methods predict ligand-target interactions through the structural information of proteins and ligands. Their aim is to predict the conformation and orientation of the ligand within the protein cavity (docking), as well as the binding affinity of the ligand and the protein (scoring).

There are two target-based approaches, docking and inverse docking. The first one predicts the orientation of a compound in the cavity of a given target,

forming and stable complex. While inverse docking fishes targets from known ligands. Both of them play an important role in virtual screening.

Some of the most important docking programs used are AutoDock, Discovery Studio and INVDOCK. AutoDock is free suite of automated docking tools, it consist of Autodock 4 and AutoDock Vina. Both AutoDock programs are often used in ethnopharmacology studies to perform virtual screening and chemical mechanism studies (Morris et al., 2009).

On the other hand, INVDOCK is a inverse docking method used in target fishing studies where the authors tries to identify protein and nucleic acids targets of a small number of phytochemicals (Chen and Zhi, 2001).

The performance of these methods is highly dependent on the targets, since it is necessary the structural information of the protein. Protein structures are not always available, alternatively, often homology modelling is used to build protein structure. Swiss-Model Automated Protein Modelling Server, for example, is one of the tools available for protein modelling (Schwede et al., 2003, p.).

The structure of the compounds is also necessary in these studies. It implies that the structure have to be built and optimized when it is not available. Between the programs used for the optimization and minimization we can find Maestro, from Schrodinger, and Sybyl, from Tripos.

On the other hand, docking implementation depend highly on the nature of target, to alleviate this situation has been suggested the use of multiple active site corrections to compensate the ligand-dependent biases.

C. Machine Learning

Machine learning is a high-throughput method of artificial intelligence that enables computers to learn from known data, like ligand chemistry, structural information, and ligand-proteins networks, and predict unknowns, such as new drugs, targets, and drug-target interactions.

Machine learning can be supervised or unsupervised. In the first one, the objective is to build a mathematical model from input variables and predict unknown interactions involving new compounds and proteins, while the objective of unsupervised machine learning is to extract patterns and interactions between a series of input variables (Yamanishi, 2013). Usually the data is divided into a training dataset, which is used to create the model, and a validation dataset that check the robustness of the model (Zhao et al., 2013).

Machine learning can use many different techniques. In unsupervised learning the most common approaches are principal-component-based methods. While in supervised learning the most used techniques are Support vector machines (SVM) and Random forests (RF), these techniques can be used individually or in combination with another technique (Jensen and Bateman, 2011).

Support vector machine is a powerful tool to classify objects into two classes; it uses a training set of objects and their known classes to create a model to predict the classes for new objects [biological applications of SVM, rong yang]. SVM has two main categories, support vector classification (SVC) and support vector regression (SVR), which is the most common form of SVM (Basak et al., 2007). On the other hand, random forest is a technique that includes an ensemble of decision trees and incorporates feature selection (Qi, 2012).

Some other techniques that are used in some studies are Naïve Bayes classifier and linear regression models. The first one is a probabilistic classifier based on

Bayes' Theorem; it is usually used in early prediction, it can process large amount of data, learn fast, and its tolerant to random noise (Tian et al., 2013). Otherwise, linear regression models try to stablish a relationship between two or more variables.

3.2 Network construction

Last years, one of the most used methods to elucidate the mode of action of medicinal plants has been the exploration of ligand-protein networks. Vast majority of of studies in traditional medicine network pharmacology are about Traditional Chinese Medicine, where synergism is the principle core and plays an essential role improving clinical. Chinese Medicine is considered the pioneer of multicomponent-multitarget pharmacology (Zhang et al., 2013). The number of publications in this area has been increasing exponentially the last 5 years (Zhang et al., 2013).

The construction of the network allows identifying active ingredients and synergistic combinations, so, it helps to understand the therapeutic mechanism of traditional medicine.

Networks use to be built in programs that allow visualizing complex networks and integrate attribute data. In this case, in most of studies, the authors use Cytoscape which is an open source software platform for biological network visualization, data integration and statistical modelling of molecular networks (Cline et al., 2007).

3.3 Research studies

In the last years, it has been performed many different studies focused on traditional medicine using the mentioned methods. A part from a great variety of methodologies, in these papers we can also find many different aims. The

authors not only try to elucidate the therapeutic plants mode of action, but also try to identify natural compounds that can potentially be new drugs. In medicinal plants research we can find principally these type studies:

3.3.1 Drug discovery

It have been published many studies where the authors use virtual screening on traditional medicine to try to identify potential ligands for protein targets of therapeutic interest. They predict interactions and affinities of compounds against proteins that must be subsequently confirmed experimentally. In these studies, usually the authors screen a large number of natural compounds, and some of the predicted interactions are tested experimentally, in order to validate their prediction. There is also target fishing studies, where a small number of phytochemicals are screened against a target database to identify protein targets *ab initio*.

3.3.2 Multi-target studies

Multi-target studies, unlike Western medicine studies, which are based on 'reductionism' philosophy and follow the paradigm of 'one gene - one drug - one disease', are based on 'holism' philosophy and replace the drug design of 'magic bullets' by the search of multitarget drugs that act on biological networks (Wang et al., 2011; Zhang et al., 2013). The authors of these studies assume that drugs commonly act on multiple targets, and that the mode of action of medicinal plants is due to the synergistic action of many constituents of the plant. So they search for a group of ligands that interact with one or more targets from a set of functionally/pathologically related targets, and single phytochemicals that potentially can target a variety of targets from the set of related proteins.

4 Aplications

Recently Li Y. *et al.* combined methods of drug-likeness evaluation, oral bioavailability prediction, drug targets prediction and network pharmacology techniques to investigate the mechanism of *Eucommia ulmoides Oliv.*, an herb widely used to help regulate hypertension and the immune system. The herbal composition was retreived from TCMSP, and using a systematic approach based on RF and SVM models the candidate targets were predicted. This model was built using a dataset of 6511 drugs and 3987 targets from DrugBank. Finally, compound-potential target and target-diseases networks were built in Cytoscape. From the screening of 41 ingredients they found 39 potential targets hits, which are associated to many diseases, such as neoplasms, cardiovascular diseases, immune diseases, etc (Li et al., 2014). This methodology have been applied previously in many studies to elucidate the pharmacological properties and mode of actions many other traditional medicines; such as licorice (*Glycyrrhiza glabra*), a widely used herb with many medicinal properties (H. Liu et al., 2013); Radix Curcumae formula, which consist of four herbs and it is applied to prevent CCVD (Tao et al., 2013); Ma-huang Decoction, and herbal formula composed of four herbs which is used to treat several diseases such as cough, asthma, headache, arthalgia, etc. (Yao et al., 2013); seven herbs clinically used for the treatment of cardiovascular disease, *Radix astragali Monogolici, Radix puerariae Lobatae, Radix ophiopogonis japonici,* and *Radix salviae miltiorrhiza (Wang et al., 2012)*, and *Ligusticum chuanxiong, Dalbergia odorifera and Corydalis yanhusuo (B. Li et al., 2012)*; and two sets of tonic herbs, called Qi-enriching herbs and Blood-tonifying herbs (J. Liu et al., 2013).

Li X *et al* studied the underlying therapeutic mechanism of Compound Danshen Formula (CDF) with a systems-pharmacological model. CDF is a traditional Chinese medicine applied in the treatment of cardiovascular diseases; it is composed of 3 herbs. The chemical composition of each herb was

retrieved from Chemistry Database (www.organchem.csdb.cn), Chinese Herbal Drug Database (Qiao et al., 2002) and literature, and their structures were downloaded from LookChem (http://www.lookchem.com) or produced **in** ISIS Draw 2.5 and optimized in Sybyl 6.9. Potential targets of these compounds were searched in PharmMapper and the information of predicted target candidates related with CVDs was collected from TTD, PharmGKB and DrugBank. To validate compound-target associations related with CVD, they performed a molecular docking simulation on each bioactive compound using AutoDock software, and molecular dynamics simulation in the Amber 10 suite of programs (Case et al., 2005). Finally, they used Cytoscape to generate Compound-Target, Compound-Pathways and Target-Disease networks. Compound-target Network, these networks illustrated the interactions of 85 compounds with 41 targets which revealed the mechanism of CDF on CVD (X. Li et al., 2012).

Shujing S. *et al.* used a docking-approach in an integrated model of systems pharmacology to study the mechanism of actions of Fufang Xueshuatong (FXST) Capsule. FXST is a traditional Chinese remedy used for the treatment of cardiovascular diseases, it is composed of four-herb formula of *Panax notoginseng, Radix astragali, Salvia miltiorrhize* and *Radix scrophulariaceae.* In this study the chemical composition of the herb was not retreived from any existing database but from chromatographic studies they performed previously. Otherwise, 115 candidate proteins related to blood coagulation and thrombotic diseases were collected using data-mining on literature and public database sources, including PubMed, PubChem, Drugbank, PDTD, TTD and PharmGKB. Protein structures were downloaded from PDB. Surflex-Dock (Jain, 2007) was used to perform the docking studies, where the top 10 targets for each compound where selected as potential targets. The predicted interactions were represented in a compound-target network. Their results showed that 22 ingredients of FXST interact with 41 targets related to some

cardiovascular diseases, elucidating the synergistic mechanisms of this medicine (Sheng et al., 2014).

Duhuo Jisheng Decoction (DHJSD), a traditional Chinese medicine used to treat osteoarthritis (OA), has been also studied to elucidate his therapeutic mechanism of action. Zheng *et al,* performed a study of this medicine using LigandFit, a docking program within DS 2.0, to perform a virtual screening of phytochemicals. DHJSD is composed of 15 medicinal herbs whose chemical compounds were retrieved from Chinese Herbal Drug Database and Handbook of the Constituents in Chinese Herbal Drug, obtaining 496 compounds. The chemical structures were docked against 20 proteins associated to OA, and compounds were sorted according to their DockScore. The top 3% compounds of the DockScore sorting were linked to their corresponding proteins to construct drug-targets networks in Cytoscape, in order to search for multi-target compounds of DHJSD. A drug-drug association network was also built to classify compounds into clusters. Their results suggested that DHJSD had compounds with potential synergy and polypharmacology against OA (Zheng et al., 2013b). Using a similar methodology, the same research group also investigated the molecular mechanism of Taohong Siwu decoction (THSWD), another formulation prescribed in traditional Chinese medicine used in the treatment of OA (Zheng et al., 2013a), and two herbs used for the therapy of CVDs, *Salvia miltiorrhiza* and *Panax ginseng*, both investigated in many other TCM studies (Zheng et al., 2013b).

To elucidate the mechanism of action of a group of 32 herbs found in Chinese medicines used for the treatment of type II diabetes mellitus (T2DM), in 2013 Tian S *et al* applied a protocol that combines molecular docking and pharmacophore mapping to discriminate potential inhibitors from non-inhibitors for the selected proteins. 2479 phytochemicals were retrieved from

TCMCD and TCM-Datatabase@Taiwan and 42 T2DM-related targets and their structure were collected from TTD, Drugbank, KEGG and literature. Molecular docking and pharmacophore-based ligand profiling were performed and their accuracy and reliability was examined using a validation dataset composed of inhibitors and non-inhibitors. Since the accuracies of these two methods were variable, Bayesian classifiers were also used to identify potential inhibitors. Summarizing, 1590 drug-like compounds were determined to be potential inhibitors for 14 proteins, but to achieve more reliable predictions, only the top 693 compounds ranked by decreasing the Bayesian scores for each target were chosen. The analysis of the compound-target networks demonstrated that a small portion of inhibitors can interact with multi-targets (Tian et al., 2013).

A mechanistic study of an anti-cancer Chinese medicine, Yadanzi (*Brucea javanica)*, has been recently performed by Zhang *et al* applying a reverse docking-based approach. 13 major ingredients of Yadanzi were collected from TCMD and their putative targets were identified using INVDOCK. 902 proteins (including 113 known therapeutic targets of marketed drugs) and 7119 ingredient-target interactions were predicted, 2100 of them with a better binding affinity than their corresponding drug-target interactions according to a comparative docking analysis. 17 of the 902 targeted proteins were mapped in KEGG pathway of non-small lung cancer (NSCLC), and the network analysis suggested that anti-cancer activity of Yadanzi is result of the manipulation of MAPK signalling and the phosphorilation process of anti-apoptosis (Zhang et al., 2014).

Shi *et al* has applied a network pharmacology approach to understand the mechanisms of action of Bu-shen-Huo-xue formula (BSHX) against chronic kidney disease (CKD). This medicine is composed of five herbs, for which are retrieved 774 compounds. 478 genes associated with CKD were collected from

OMIM, Genetic Association Database (GAD) and microarrays of Gene Expression Omnibus (GEO), and 31 proteins associated to CKD from TTD and Drugbank. For the construction of a natural-product target network a molecular docking was performed to predict phytochemical putative targets. Finally, it was built a network including PPi extracted from human protein-protein databases. Networks analysis revealed that BSHX exerts his therapeutic effect through multi-channel network regulation. Tanshinone IIA, rhein, curcumin, calycosin, and quercetin were identified as the potential effective ingredients of this medicine (Shi et al., 2014).

Network construction and analysis has also been used by Song J *et al.* to elucidate the molecular mechanism of the tradicional medicine Chinese formula Shu-feng-jie-du, which consist of 8 herbs and is used to treat influenza infection. They developed a module analysis approach to investigate complex networks, and used it identify pharmacological units, which are connected subneworks where a set of compounds with similar physicochemical properties modulate the activities of a group of function-similar gene-products. Their approach consisted of three steps, network construction, module detection and pathway analysis. To construct the compound-target network, they collected all the necessary data from existing databases, the herb composition was extracted from the Chemistry Database (http://www.organchem.csdb.cn), and the 2D structure from PubChem, obtaining 243 chemicals, on the other hand, the potential targets are retrieved from Drugbank, CTD and STITCH. In the network it was also integrated protein-protein interactions (PPi) data from databases such as Human Protein Reference Database (Peri et al., 2004) and BioGrid (Stark et al., 2006) (protein-protein databases). They identified four pharmacological units, and 24 out of 40 enriched pathways that were ranked in the top 10 corresponding to each pharmacological unit were relevant for the process of influenza infection (Song et al., 2013).

The mechanisms of action of other 18 plants used for treating influenza have also been studied by Gu S *et al* using molecular computational methods. They collected the structures of plant ingredients in TCMD and the available structures of influenza viral proteins from PDB in order to perform docking simulations in AutoDockTools. The predicted interactions suggested that these herbs can inhibit influenza via the targeting of various viral proteins, and are effective against different influenza subtypes (Gu et al., 2013).

A similar approach has been applied by Ma S *et al* to investigate the mechanism at the molecular level of TCM for the treatment of sepsis. They identified 16 targets involved in sepsis diseased and 343 compounds from 5 herbs to perform a virtual docking using Schrodinger Glide. To validate the predicted interactions, compounds that inhibited thrombin protein in computational studies were tested in vitro. Docking results showed that multiple bioactive compounds targeted multiple proteins and the first 10 compounds were characterized. On the other hand, the *in vitro* assays suggested a good correlation with the virtual screening (Ma et al., 2013).

Otherwise, Wang X *et al* applied molecular docking and virtual screening to identify potential natural ingredients able to inhibit inducible nitric oxide synthase (iNOS). They generated a pharmacophore model with GALAHAD from a set of iNOS inhibitors selected from the literature; and after testing it, the model was used to screen TCMD, which contains 23033 compounds. From screening it was obtained a hit list of 498 ingredients to be used in the molecular docking, performed in Surflex-Dock. For the top 20 compounds with a higher docking score, they searched in related literature for experimental evidence of their capability decreasing the activity or production of NO (Wang et al., 2014).

Arya H and Coumar MS searched fragment-like lead molecules in TCMD for filariasis target asparaginyl-tRNA synthase. For this study, 95 chemicals reported from eight plants used for the treatment of worm infection were retrieved from TCMD@Taiwan. Two different virtual screening approaches were used to identify the hits, docking-based virtual screening and E-pharmacophore virtual screening. The best hits from both screening were two aglycones of *Agrimonia*, that were later used to perform a molecular dynamics simulation study, which revealed that both compounds are forming stable interactions with the target protein (Arya and Coumar, 2014).

In a similar way, it has been performed some studies to identify candidate compounds that inhibit Human immunodeficiency virus type-1 (HIV). Huang HJ *et al* screened TCM compounds to identify new candidates that inhibit HIV integrase (IN), a required factor for the infection of HIV. They used a docking approach to identify compounds with a higher dock score than D77, a known drug with demonstrated inhibition against HIV by binding IN. Subsequently, multiple linear regression and support vector machine were used to predict the potential bioactivity of TCM candidates. 9-hydroxy-(10E)-octadecenoic acid and beauveriolide I were identified as potential inhibitor and a molecular dynamics simulation confirmed that both compounds were capable of forming stable complexes with IN (Huang et al., 2014). On the other hand, Yanuar *et al* performed a virtual screening using AutoDock to identify potential inhibitors of HIV-1 protease. They downloaded the phytochemicals from HerbalDB (Yanuar et al., 2011) and chemical structures were docked against HIV-1 protease using AutoDock in PyRx. Top 10 ranked compounds were list as hits from screening, 8-Hydroxyapigenin 8-(2",4"-disulfatoglucuronide), isoscutellarin 4'-methyl ether, amaranthin, torvanol A, ursonic acid, 5-Carboxypyranocyanidin 3-O-(6"-O-malonyl-beta-glucopyranoside), Oleoside, jacoumaric acid, platanic acid and 5-carboxypyranocyanidin 3-O-beta-glucopyranoside(Yanuar et al., 2014).

Docking simulations have also been applied by Sathishkumar N *et al* to identify ginsenosides from *Panax ginseng* with binding affinity to 3 anti-apoptotic proteins, BCL-2, BCL-XL and MLC-1. Ginseng is one of the most valuable medicinal plants in eastern Asia and many studies have revealed that his derivatives reduce tumor growth. 12 ginsenosides downloaded from PubChem were docked against the anti-apoptotic proteins in AutoDock. Rg1, Rg3, Rf and Rh2 were found to have binding affinity with BCL-2, BCL-XL and MLC-1, therefore they were identified as potent cancer inhibitors that could be used in chemotherapy (Sathishkumar et al., 2012).

Src kinase is also an attractive target associated to cancer for which have been searched potential ligands in traditional medicine databases. In 2012 Tou WI and Chen CY screened TCM Database@Taiwan against Src kinase using LigandFit program within DS 2.5 (Venkatachalam et al., 2003). DS 2.5 was used to calculate individual molecular property descriptors of 53 Src inhibitors with known pIC50 values, which were used to construct 4 models to predicting the bioactivity of TCM candidates, MLR and SVM models for QSAR, and CoMFA and CoMSIA models for 3D-QSAR. DS 2.5 was also used to perform molecular dynamics to evaluate the stability of candidates with Src kinase. Isopraeroside IV, 9alpha-hydroxyfraxinellone-9-O-beta-D-glucoside and aurantiamide were the top three TCM candidates identified from docking, and based on their high stability and predicted bioactivities, they may be directly used as candidate lead compounds in biological studies (Tou et al., 2013). This methodology was also applied by the same authors to discover potential FAAH-like anandamide transporter (FLAT) antagonist, a drug target for pain regulation. This study suggested Guineensine as an antagonist of FLAT, with a potential application in relieving neuropathic pain (Tou et al., 2013).

Paulke *et al* performed a study to compare the activity of lysergic acid diethylamide (LSD) and lysergic acid amide (LSA), a psychoactive ergotalkaloid

found in seeds of *Argyreia nervosa*, a medicinal plant in Ayurvedic medicine. This herb is used in number diseases, such as nervousness, bronchitis, tuberculosis, arthritis and diabetes, and LSA is considered as a natural substitute of LSD, due to their similar chemical structure. To predict the pharmacological profiles of both compounds and other *Argyreia nervosa* ergotalkaloids, they performed an in silico prediction model based on the similarity of the compounds against a dataset of ligands retrieved from ChemblDB with reported Ki values for serotonin, norepinephrine, dopamine, muscarine, and histamine receptor subtypes. The comparison of LSA and LSD pharmacological profiles exhibited that LSA has a weaker psychedelic activity than LSD, and should not be regarded as LSD-like psychedelic drug. On the other hand, a broad spectrum of possible targets was predicted for *Argyreia nervosa* ergotalkaloids that need to be more investigated (Paulke et al., 2013).

Luo J. *et al* proposed an approach called directed TCM grammar systems (dTGS), based on EGS, to identify effective components from TCM formula. In order to testing their approach they studied the component-disease relationship of the TCM formula Bai-Hu decoction plus Wasting-Thirsting (BHDWT) and type 2 diabetes mellitus (T2D). The components of the formula were extracted from TCMD and the compound-target interactions were derived from STITCH, on the other hand the signal pathways of T2D were retrieved from KEGG and the chemical components used to treat T2D from TTD. The collected data was used to create a compound-target network of the formula and a biological network of T2D, and performing their EGS approach it could be identified the effective components groups. They found 19 compounds acting on 20 proteins in T2D (Luo et al., 2013).

5 Conclusions

In the recent years, many different *in silico* methodologies has been applied in the traditional medicine research. These tools facilitate the identification of therapeutic targets of medicinal plants and allow elucidating the mode of action of these plants, identifying which constituents are responsible of the therapeutic action of medicinal plants and putative new leads for drugs.

The used methodologies include several different approaches, such as data mining, QSAR approaches and statistical machine learning methods. The *in silico* tool used in each study depends on the purpose of the research and it's a decision of the researchers. Most of these tools predict new chemicals activities which previously haven't been reported experimentally, so it is important to know the reliability of the used approach and to combine *in silico* methodologies with experimental tests in order to validate the predictions.

Some of medicinal plants researches use existing data to seek out the therapeutic mechanism of these traditional medicines. They use to apply data mining or entity grammar systems methods to collect and select data. Despite there is enough existing data to achieving this purpose, there is still many data that remain unknown or is private, so the use of *in silico* tools would allow complementing their methodology.

Frequently, the therapeutic action of traditional medicines use to be considered as the result of the synergistic effect of their active ingredients, so, the study of therapeutic mechanism can result in a complex network. Given the complexity of traditional medicines, these studies requires the performance of many experimental work, despite of *in silico* tools offer and economical and efficient way to explore the chemical composition activity, it suggest hypothesis that has to be tested experimentally subsequently, in vitro or in vivo.

Network pharmacology can be useful to elucidate the mode of action of medicinal plants and confirm their effective ingredients. However, despite of that there are many studies about the chemical composition of many plants with information about the concentration of each compound; the studies about traditional medicine therapeutic action don't use to take into account this information. So, it is not possible to it is not possible to know which is the contribution of each compound to the plant therapeutic effect.

Nevertheless, network pharmacology can be helpful to identify synergistic combinations and optimize the formula of traditional medicine and find new leads for drugs.

References

Accelrys Software Inc., n.d. Discovery Studio Modeling Environment, Release 4.0.

AERS [WWW Document], 2014. . FDA Adverse Event Report. Syst. FAERS. URL http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ (accessed 11.13.14).

Afendi, F.M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K., Ikeda, S., Takahashi, H., Altaf-Ul-Amin, M., Darusman, L.K., Saito, K., Kanaya, S., 2012. KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. Plant Cell Physiol. 53, e1. doi:10.1093/pcp/pcr165

Arya, H., Coumar, M.S., 2014. Virtual screening of traditional Chinese medicine (TCM) database: identification of fragment-like lead molecules for filariasis target asparaginyl-tRNA synthetase. J. Mol. Model. 20, 2266. doi:10.1007/s00894-014-2266-9

Balunas, M.J., Kinghorn, A.D., 2005. Drug discovery from medicinal plants. Life Sci. 78, 431–441. doi:10.1016/j.lfs.2005.09.012

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., Soboleva,

A., 2013. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 41, D991–D995. doi:10.1093/nar/gks1193

Basak, D., Pal, S., Patranabis, D.C., 2007. . Neural Inf. Process. – Lett. Rev. 11, 22.

Becker, K.G., Barnes, K.C., Bright, T.J., Wang, S.A., 2004. The Genetic Association Database. Nat. Genet. 36, 431–432. doi:10.1038/ng0504-431

Berman, H.M., 2000. The Protein Data Bank. Nucleic Acids Res. 28, 235–242. doi:10.1093/nar/28.1.235

Bolton, E.E., Wang, Y., Thiessen, P.A., Bryant, S.H., 2008. PubChem: Integrated Platform of Small Molecules and Biological Activities, in: Annual Reports in Computational Chemistry. Washinton, DC.

Cascorbi, I., 2012. Drug interactions--principles, examples and clinical consequences. Dtsch. Ärztebl. Int. 109, 546–555; quiz 556. doi:10.3238/arztebl.2012.0546

Case, D.A., Cheatham, T.E., Darden, T., Gohlke, H., Luo, R., Merz, K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.J., 2005. The Amber biomolecular simulation programs. J. Comput. Chem. 26, 1668–1688. doi:10.1002/jcc.20290

Chen, C.Y.-C., 2011. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening In Silico. PLoS ONE 6, e15939. doi:10.1371/journal.pone.0015939

Chen, Y.Z., Zhi, D.G., 2001. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins Struct. Funct. Genet. 43, 217–226. doi:10.1002/1097-0134(20010501)43:2<217::AID-PROT1032>3.0.CO;2-G

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A.R., Vailaya, A., Wang, P.-L., Adler, A., Conklin, B.R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G.J., Ideker, T., Bader, G.D., 2007. Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc. 2, 2366–2382. doi:10.1038/nprot.2007.324

Davis, A.P., Grondin, C.J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B.L., Wiegers, T.C., Mattingly, C.J., 2014. The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Res. doi:10.1093/nar/gku935

Dictionary of Natural Products [WWW Document], 2014. . Dict. Nat. Prod.

URL dnp.chemnetbase.com/ (accessed 11.13.14).

Dr Duke Phytochemical and Ethnobotanical Database [WWW Document], n.d. URL http://www.ars-grin.gov/duke/

Dunnick, J.K., Nyska, A., 2013. The toxicity and pathology of selected dietary herbal medicines. Toxicol. Pathol. 41, 374–386. doi:10.1177/0192623312466451

Ekins, S., Mestres, J., Testa, B., 2007. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br. J. Pharmacol. 152, 9–20. doi:10.1038/sj.bjp.0707305

Friesner, R.A., Murphy, R.B., Repasky, M.P., Frye, L.L., Greenwood, J.R., Halgren, T.A., Sanschagrin, P.C., Mainz, D.T., 2006. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. J. Med. Chem. 49, 6177–6196. doi:10.1021/jm051256o

GALAHAD Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144-2319. www.tripos.com. Contact company for pricing information., 2007. . J. Am. Chem. Soc. 129, 5780–5780. doi:10.1021/ja069815j

Gao, Z., Li, H., Zhang, H., Liu, X., Kang, L., Luo, X., Zhu, W., Chen, K., Wang, X., Jiang, H., 2008. PDTD: a web-accessible protein database for drug target identification. BMC Bioinformatics 9, 104. doi:10.1186/1471-2105-9-104

Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J.P., 2012. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40, D1100–D1107. doi:10.1093/nar/gkr777

Gausachs, R., 2008. Les Herbes Remeieres, Rafael Dalmau. ed. Rafael Dalmau, Editor, Barcelona.

Gu, S., Yin, N., Pei, J., Lai, L., 2013. Understanding molecular mechanisms of traditional Chinese medicine for the treatment of influenza viruses infection by computational approaches. Mol. Biosyst. 9, 2696–2700. doi:10.1039/c3mb70268e

He, M., Yan, X., Zhou, J., Xie, G., 2001. Traditional Chinese medicine database and application on the Web. J. Chem. Inf. Comput. Sci. 41, 273–277.

Huang, C., Zheng, C., Li, Y., Wang, Y., Lu, A., Yang, L., 2014. Systems pharmacology in drug discovery and therapeutic insight for herbal medicines. Brief. Bioinform. 15, 710–733. doi:10.1093/bib/bbt035

Huang, H.-J., Jian, Y.-R., Chen, C.Y.-C., 2014. Traditional Chinese medicine application in HIV: an in silico study. J. Biomol. Struct. Dyn. 32, 1–12.

doi:10.1080/07391102.2012.745168

Jain, A.N., 2007. Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. J. Comput. Aided Mol. Des. 21, 281–306. doi:10.1007/s10822-007-9114-2

Jensen, L.J., Bateman, A., 2011. The rise and fall of supervised machine learning techniques. Bioinforma. Oxf. Engl. 27, 3331–3332. doi:10.1093/bioinformatics/btr585

Jones, A.W., 2011. Early drug discovery and the rise of pharmaceutical chemistry. Drug Test. Anal. 3, 337–344. doi:10.1002/dta.301

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., Herwig, R., 2011. ConsensusPathDB: toward a more complete picture of cell biology. Nucleic Acids Res. 39, D712-717. doi:10.1093/nar/gkq1156

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2014. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42, D199-205. doi:10.1093/nar/gkt1076

Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J., Shoichet, B.K., 2007. Relating protein pharmacology by ligand chemistry. Nat. Biotechnol. 25, 197–206. doi:10.1038/nbt1284

Koehn, F.E., Carter, G.T., 2005. The evolving role of natural products in drug discovery. Nat. Rev. Drug Discov. 4, 206–220. doi:10.1038/nrd1657

Kuhn, M., Campillos, M., Letunic, I., Jensen, L.J., Bork, P., 2010. A side effect resource to capture phenotypic effects of drugs. Mol. Syst. Biol. 6, 343. doi:10.1038/msb.2009.98

Kuhn, M., Szklarczyk, D., Pletscher-Frankild, S., Blicher, T.H., von Mering, C., Jensen, L.J., Bork, P., 2014. STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Res. 42, D401-407. doi:10.1093/nar/gkt1207

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z.T., Han, B., Zhou, Y., Wishart, D.S., 2014. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 42, D1091-1097. doi:10.1093/nar/gkt1068

Li, B., Xu, X., Wang, X., Yu, H., Li, X., Tao, W., Wang, Y., Yang, L., 2012. A systems biology approach to understanding the mechanisms of action of chinese herbs for treatment of cardiovascular disease. Int. J. Mol. Sci. 13, 13501–13520. doi:10.3390/ijms131013501

Li, X., Xu, X., Wang, J., Yu, H., Wang, X., Yang, H., Xu, H., Tang, S., Li, Y., Yang, L., Huang, L., Wang, Y., Yang, S., 2012. A system-level

investigation into the mechanisms of Chinese Traditional Medicine: Compound Danshen Formula for cardiovascular disease treatment. PloS One 7, e43918. doi:10.1371/journal.pone.0043918

Li, Y., Han, C., Wang, J., Xiao, W., Wang, Z., Zhang, J., Yang, Y., Zhang, S., Ai, C., 2014. Investigation into the mechanism of Eucommia ulmoides Oliv. based on a systems pharmacology approach. J. Ethnopharmacol. 151, 452–460. doi:10.1016/j.jep.2013.10.067

Liu, H., Wang, J., Zhou, W., Wang, Y., Yang, L., 2013. Systems approaches and polypharmacology for drug discovery from herbal medicines: an example using licorice. J. Ethnopharmacol. 146, 773–793. doi:10.1016/j.jep.2013.02.004

Liu, J., Pei, M., Zheng, C., Li, Y., Wang, Y., Lu, A., Yang, L., 2013. A systems-pharmacology analysis of herbal medicines used in health improvement treatment: predicting potential new drugs and targets. Evid.-Based Complement. Altern. Med. ECAM 2013, 938764. doi:10.1155/2013/938764

Liu, T., Lin, Y., Wen, X., Jorissen, R.N., Gilson, M.K., 2007. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res. 35, D198-201. doi:10.1093/nar/gkl999

Luo, J., Ren, Y., Gu, H., Wu, Y., Wang, Y., 2013. dTGS: Method for Effective Components Identification from Traditional Chinese Medicine Formula and Mechanism Analysis. Evid.-Based Complement. Altern. Med. ECAM 2013, 840427. doi:10.1155/2013/840427

Ma, S., Feng, C., Zhang, X., Dai, G., Li, C., Cheng, X., Liu, P., Ju, W., Yu, H., 2013. The multi-target capabilities of the compounds in a TCM used to treat sepsis and their in silico pharmacology. Complement. Ther. Med. 21, 35–41. doi:10.1016/j.ctim.2012.12.001

McChesney, J.D., Venkataraman, S.K., Henri, J.T., 2007. Plant natural products: back to the future or into extinction? Phytochemistry 68, 2015–2022. doi:10.1016/j.phytochem.2007.04.032

Mishra, B.B., Tiwari, V.K., 2011. Natural products: an evolving role in future drug discovery. Eur. J. Med. Chem. 46, 4769–4807. doi:10.1016/j.ejmech.2011.07.057

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., Olson, A.J., 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J. Comput. Chem. 30, 2785–2791. doi:10.1002/jcc.21256

Naoghare, P.K., Song, J.M., 2010. Chip-based high-throughput screening of

herbal medicines. Comb. Chem. High Throughput Screen. 13, 923–931.

NAPRALERT [WWW Document], 2012. . NAPRALERT - Nat. Prod. Alert Database. URL http://www.napralert.org/ (accessed 11.13.14).

Newman, D.J., 2008. Natural products as leads to potential drugs: an old process or the new hope for drug discovery? J. Med. Chem. 51, 2589–2599. doi:10.1021/jm0704090

Nikolaou, P., Papoutsis, I., Stefanidou, M., Dona, A., Maravelias, C., Spiliopoulou, C., Athanaselis, S., 2012. Accidental poisoning after ingestion of "aphrodisiac" berries: diagnosis by analytical toxicology. J. Emerg. Med. 42, 662–665. doi:10.1016/j.jemermed.2011.03.023

Online Mendelian Inheritance in Man, OMIM [WWW Document], n.d. URL http://omim.org/ (accessed 11.13.14).

Paulke, A., Kremer, C., Wunder, C., Achenbach, J., Djahanschiri, B., Elias, A., Schwed, J.S., Hübner, H., Gmeiner, P., Proschak, E., Toennes, S.W., Stark, H., 2013. Argyreia nervosa (Burm. f.): receptor profiling of lysergic acid amide and other potential psychedelic LSD-like compounds by computational and binding assay approaches. J. Ethnopharmacol. 148, 492–497. doi:10.1016/j.jep.2013.04.044

Peri, S., Navarro, J.D., Kristiansen, T.Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T.K.B., Chandrika, K.N., Deshpande, N., Suresh, S., Rashmi, B.P., Shanker, K., Padma, N., Niranjan, V., Harsha, H.C., Talreja, N., Vrushabendra, B.M., Ramya, M.A., Yatish, A.J., Joy, M., Shivashankar, H.N., Kavitha, M.P., Menezes, M., Choudhury, D.R., Ghosh, N., Saravana, R., Chandran, S., Mohan, S., Jonnalagadda, C.K., Prasad, C.K., Kumar-Sinha, C., Deshpande, K.S., Pandey, A., 2004. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res. 32, D497-501. doi:10.1093/nar/gkh070

Qi, Y., 2012. Random Forest for Bioinformatics, in: Ensemble Machine Learning: Methods and Applications. p. 18.

Qiao, X., Hou, T., Zhang, W., Guo, S., Xu, X., 2002. A 3D structure database of components from Chinese traditional medicinal herbs. J. Chem. Inf. Comput. Sci. 42, 481–489.

Ru, J., Li, P., Wang, J., Zhou, W., Li, B., Huang, C., Li, P., Guo, Z., Tao, W., Yang, Y., Xu, X., Li, Y., Wang, Y., Yang, L., 2014. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. J. Cheminformatics 6, 13. doi:10.1186/1758-2946-6-13

Sathishkumar, N., Sathiyamoorthy, S., Ramya, M., Yang, D.-U., Lee, H.N., Yang, D.-C., 2012. Molecular docking studies of anti-apoptotic BCL-2, BCL-XL, and MCL-1 proteins with ginsenosides from Panax ginseng. J.

Enzyme Inhib. Med. Chem. 27, 685–692. doi:10.3109/14756366.2011.608663

Schwede, T., Kopp, J., Guex, N., Peitsch, M.C., 2003. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. 31, 3381–3385.

Sheng, S., Wang, J., Wang, L., Liu, H., Li, P., Liu, M., Long, C., Xie, C., Xie, X., Su, W., 2014. Network pharmacology analyses of the antithrombotic pharmacological mechanism of Fufang Xueshuantong Capsule with experimental support using disseminated intravascular coagulation rats. J. Ethnopharmacol. 154, 735–744. doi:10.1016/j.jep.2014.04.048

Shi, S., Cai, Y., Cai, X., Zheng, X., Cao, D., Ye, F., Xiang, Z., 2014. A network pharmacology approach to understanding the mechanisms of action of traditional medicine: Bushenhuoxue formula for treatment of chronic kidney disease. PloS One 9, e89123. doi:10.1371/journal.pone.0089123

Skozen, S., Bussman, R.W., 2006. International Ethnobotany Database. Lyonia: A Journal of Ecology and Application 11, 11.

Song, J., Zhang, F., Tang, S., Liu, X., Gao, Y., Lu, P., Wang, Y., Yang, H., 2013. A Module Analysis Approach to Investigate Molecular Mechanism of TCM Formula: A Trial on Shu-feng-jie-du Formula. Evid.-Based Complement. Altern. Med. ECAM 2013, 731370. doi:10.1155/2013/731370

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M., 2006. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34, D535-539. doi:10.1093/nar/gkj109

Tao, W., Xu, X., Wang, X., Li, B., Wang, Y., Li, Y., Yang, L., 2013. Network pharmacology-based prediction of the active ingredients and potential targets of Chinese herbal Radix Curcumae formula for application to cardiovascular disease. J. Ethnopharmacol. 145, 1–10. doi:10.1016/j.jep.2012.09.051

Tian, S., Li, Y., Li, D., Xu, X., Wang, J., Zhang, Q., Hou, T., 2013. Modeling compound-target interaction network of traditional Chinese medicines for type II diabetes mellitus: insight for polypharmacology and drug design. J. Chem. Inf. Model. 53, 1787–1803. doi:10.1021/ci400146u

Tou, W.I., Chang, S.-S., Lee, C.-C., Chen, C.Y.-C., 2013. Drug design for neuropathic pain regulation from traditional Chinese medicine. Sci. Rep. 3, 844. doi:10.1038/srep00844

Traditional Medicine Growing Needs and Potential - WHO Policy Perspectives on Medicines, No. 002, May 2002, 2002.

Venkatachalam, C.M., Jiang, X., Oldfield, T., Waldman, M., 2003. LigandFit: a

novel method for the shape-directed rapid docking of ligands to protein active sites. J. Mol. Graph. Model. 21, 289–307.

Vidal, D., Garcia-Serna, R., Mestres, J., 2011. Ligand-based approaches to in silico pharmacology. Methods Mol. Biol. Clifton NJ 672, 489–502. doi:10.1007/978-1-60761-839-3_19

Wang, X., Ren, Z., He, Y., Xiang, Y., Zhang, Y., Qiao, Y., 2014. A combination of pharmacophore modeling, molecular docking and virtual screening for iNOS inhibitors from Chinese herbs. Biomed. Mater. Eng. 24, 1315–1322. doi:10.3233/BME-130934

Wang, X., Sun, H., Zhang, A., Sun, W., Wang, P., Wang, Z., 2011. Potential role of metabolomics apporoaches in the area of traditional Chinese medicine: As pillars of the bridge between Chinese and Western medicine. J. Pharm. Biomed. Anal. 55, 859–868. doi:10.1016/j.jpba.2011.01.042

Wang, X., Xu, X., Tao, W., Li, Y., Wang, Y., Yang, L., 2012. A systems biology approach to uncovering pharmacological synergy in herbal medicines with applications to cardiovascular disease. Evid.-Based Complement. Altern. Med. ECAM 2012, 519031. doi:10.1155/2012/519031

Wang, Y., 2004. Entity grammar systems: a grammatical tool for studying the hierarchical structures of biological systems. Bull. Math. Biol. 66, 447–471. doi:10.1016/j.bulm.2003.08.011

Whirl-Carrillo, M., McDonagh, E.M., Hebert, J.M., Gong, L., Sangkuhl, K., Thorn, C.F., Altman, R.B., Klein, T.E., 2012. Pharmacogenomics knowledge for personalized medicine. Clin. Pharmacol. Ther. 92, 414–417. doi:10.1038/clpt.2012.96

Xue, R., Fang, Z., Zhang, M., Yi, Z., Wen, C., Shi, T., 2013. TCMID: Traditional Chinese Medicine integrative database for herb molecular mechanism analysis. Nucleic Acids Res. 41, D1089-1095. doi:10.1093/nar/gks1100

Yamanishi, Y., 2013. Chemogenomic approaches to infer drug-target interaction networks. Methods Mol. Biol. Clifton NJ 939, 97–113. doi:10.1007/978-1-62703-107-3_9

Yang, D.H., Kang, J.H., Park, Y.B., Park, Y.J., Oh, H.S., Kim, S.B., 2013. Association rule mining and network analysis in oriental medicine. PloS One 8, e59241. doi:10.1371/journal.pone.0059241

Yanuar, A., Mun'im, A., Lagho, A.B.A., Syahdi, R.R., Rahmat, M., Suhartanto, H., 2011. Medicinal Plants Database and Three Dimensional Structure of the Chemical Compounds from Medicinal Plants in Indonesia. Int J Comp Sci 8, 180–183.

Yanuar, A., Suhartanto, H., Munim, A., Anugraha, B.H., Syahdi, R.R., 2014.

Virtual Screening of Indonesian Herbal Database as HIV-1 Protease Inhibitor. Bioinformation 10, 52–55. doi:10.6026/97320630010052

Yao, Y., Zhang, X., Wang, Z., Zheng, C., Li, P., Huang, C., Tao, W., Xiao, W., Wang, Y., Huang, L., Yang, L., 2013. Deciphering the combination principles of Traditional Chinese Medicine from a systems pharmacology perspective based on Ma-huang Decoction. J. Ethnopharmacol. 150, 619–638. doi:10.1016/j.jep.2013.09.018

Ye, H., Ye, L., Kang, H., Zhang, D., Tao, L., Tang, K., Liu, X., Zhu, R., Liu, Q., Chen, Y.Z., Li, Y., Cao, Z., 2011. HIT: linking herbal active ingredients to targets. Nucleic Acids Res. 39, D1055–D1059. doi:10.1093/nar/gkq1165

Zhang, G.-B., Li, Q.-Y., Chen, Q.-L., Su, S.-B., 2013. Network pharmacology: a new approach for chinese herbal medicine research. Evid.-Based Complement. Altern. Med. ECAM 2013, 621423. doi:10.1155/2013/621423

Zhang, H.-P., Pan, J.-B., Zhang, C., Ji, N., Wang, H., Ji, Z.-L., 2014. Network understanding of herb medicine via rapid identification of ingredient-target interactions. Sci. Rep. 4, 3719. doi:10.1038/srep03719

Zhao, M., Zhou, Q., Ma, W., Wei, D.-Q., 2013. Exploring the ligand-protein networks in traditional chinese medicine: current databases, methods, and applications. Evid.-Based Complement. Altern. Med. ECAM 2013, 806072. doi:10.1155/2013/806072

Zheng, C.-S., Xu, X.-J., Ye, H.-Z., Wu, G.-W., Li, X.-H., Xu, H.-F., Liu, X.-X., 2013a. Network pharmacology-based prediction of the multi-target capabilities of the compounds in Taohong Siwu decoction, and their application in osteoarthritis. Exp. Ther. Med. 6, 125–132. doi:10.3892/etm.2013.1106

Zheng, C.-S., Xu, X.-J., Ye, H.-Z., Wu, G.-W., Xu, H.-F., Li, X.-H., Huang, S.-P., Liu, X.-X., 2013b. Computational pharmacological comparison of Salvia miltiorrhiza and Panax notoginseng used in the therapy of cardiovascular diseases. Exp. Ther. Med. 6, 1163–1168. doi:10.3892/etm.2013.1291

Zhu, F., Shi, Z., Qin, C., Tao, L., Liu, X., Xu, F., Zhang, L., Song, Y., Liu, X., Zhang, J., Han, B., Zhang, P., Chen, Y., 2012. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. Nucleic Acids Res. 40, D1128-1136. doi:10.1093/nar/gkr797

III.3: Exploring polypharmacology of bioactive compounds

Joaquim Olivés, and Jordi Mestres*

Systems Pharmacology, Research Programme on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute and University Pompeu Fabra, Parc de Recerca Biomèdica, 08003 Barcelona, Catalonia, Spain

Abstract

Polypharmacology is known as the ability of small molecultes to bind to multiple proteins. It has been reported for drugs in several articles, but not much for other bioactive groups, like metabolites or natural products. In this respect, here we are presenting a comparison of the activity data available of many groups of compounds, such as Natural Products Libraris, Metabolites, Plant Compounds and Synthetic chemicals. To performing this study we are using virtual profiling methodology. Results pointed to a relatively high average of targets proteins in most of the groups, suggesting the importance of study more in deep the activity of less study compounds like human metabolites and plant compounds.

* Corresponding autor. *E-mail address:* jmestres@imim.es

Introduction

From few years ago, with the growing understanding of diseases, one drug-one target premise has been shifted away to a new 'multi-target, multi-drug' model. Nowadays it is widely recognised that selective drugs are more exception rather than the rule and that most therapeutically effective molecules tend to interact with multiple proteins (Anighoro et al., 2014). It has been reported that only 15% of drugs are currently known to interact only with one single target, whereas 50% of them interact with more than 5 targets (Jalencas and Mestres, 2013). The ability of a molecule to interact with multiple proteins has been popularly referred to as 'polypharmacology'.

Polypharmacalogy is the result of chemical and biological sources. Two main aspects are highlighted as chemical sources linked to polypharmacology, molecular properties and fragment composition, both much related to each other. From a decade ago it has prevailed the idea that simple molecules are more likely to bind to multiple proteins than complex molecules. However, despite of the increasing of molecular complexity tend to limit drug polypharmacology, it has been reported that within a given range of Molecular Weight values, promiscuity tends to increase with hydrophicity (Jalencas and Mestres, 2013). About biological sources, if a small molecule is binding to a protein, there are chances for it to bind also to other proteins related by sequence identity and/or binding site similarity (Liscio et al., 2013).

Polypharmacology is not only a characteristic of drug compounds; natural products, phytochemicals and metabolites, may also be interacting with more than one protein target. In this part of the thesis we have compared the relative degree of polypharmacology between 5 types of compounds. We will explore whether bioactive compounds (natural compounds and drugs) differ from synthetic compounds by their intrinsic level of polypharmacology.

Methods

Data collection

The different compounds used in this study have been collected from many different databases:

- Drugs: Drugbank.
- Metabolites: HMDB (Wishart et al., 2013)
- Plant compounds: DUKE (Duke, 2016), KNApSacK (Afendi et al., 2012), CVDHD (J. Gu et al., 2013), SWEETLEAD (Novick et al., 2013), and TCMSP (Ru et al., 2014).
- Synthetic compounds: Synthetic compounds libraries from InterBioScreen, Analyticon, Indofine and TimTec.
- Natural Products: Natural product libraries from AnalyticonNP, Arbonova, PrincetonNP, SelleckChem, SpecsNP, Sequoia Research Products, TimTec

Chemical structures were downloaded directly from these databases. To avoid repeated structures, for each compound set it was generated InChIKeys. The quantity of compound structures collected for each bioactive group can be observed in Table 1.

Chemical groups	Total
Drugs	2671
Metabolites	10577
Plant compounds	55116
Synthetic compounds	488497
Natural Product Libraries	70142

Table 1: Compounds collected for each group

Target profiling

For all the compounds we performed a ligand-based virtual profiling using CTLink. Target profiles obtained are composed of experimental and predicted interactions. In both cases we are considering only those with activity values >6. For predicted interactions we add a cuttoff of 0.3 on confidence score.

Similarity calculation

In this study we have calculated the similarity of structures. This was performed for compounds of each group, and between group pairs. Similarity values have been calculated using phrag as molecular descriptors (Vidal et al., 2011)). For each was calculated the similarity for compounds pairs, between compounds inside each group, and between compounds of differents groups.

Results

In general we have found experimental activity for a low number of compounds (Figure 1). Drugs are the group with a higher percentage of chemicals with reported experimental activity, 36%. It is followed by metabolites group, with 482 metabolites, a 4.56%. While for the other groups these percentages are 1.9% for plants compounds, 0,47% for natural products, and 0,28% for synthetic compounds.

Regarding the amount of compounds with experimental data, there is a relatively similar number of chemicals in most of the groups. For metabolites and plant compounds there are 482 and 635 chemicals. While for drugs, natural products libraries and synthetic chemicals we have found interactions for 908, 1028 and 1388 compounds, respectively.

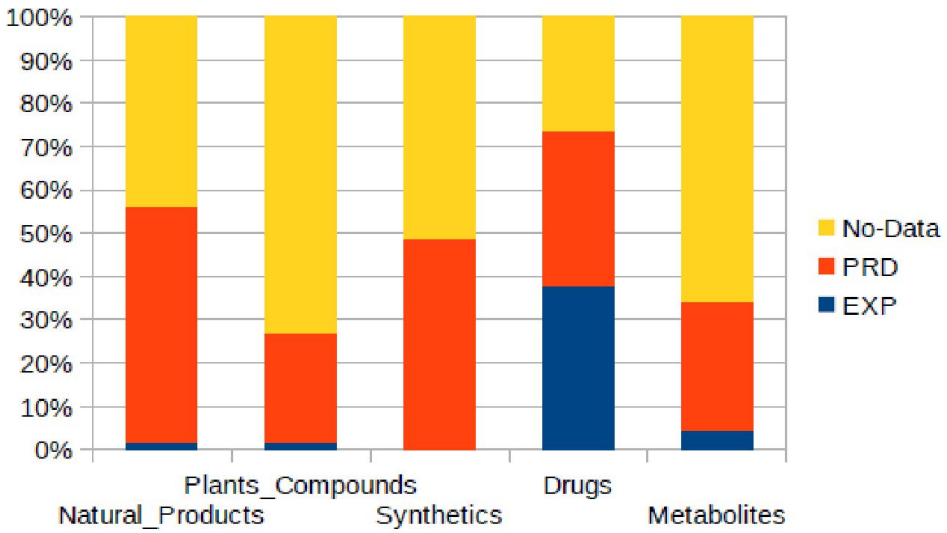


Figure 1: Graphic with the percentage of compounds for which we find experimental, for which we are able to predict some activy, and the % of compounds for we are not able to obtain activity data.

After adding predicted activities reported by CTLink, the number of compounds with interaction data is increasing significantly in all the groups (figure 1). The group with a lower % of compounds with predicted interaction data is phytochemicals, 27%. While the chemical groups with higher % are drugs (73,6%), that are followed by Natural Products with 56,3% (Figure 1).

Figure 2: Similarity heatmap, right values are representing the similarity values assigned to each color on the heatmap. D=Drugs, M=Metabolites;PC=Plant Compounds; NP=Natural Products; S=Synthetic chemicals; C=CTLink

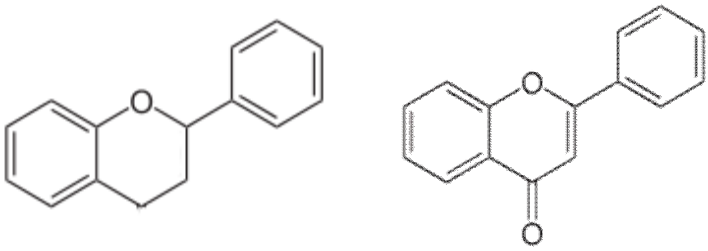


Figure 3: Molecular structure of flavone (left) and flavan (right)

To understand prediction percentages we have to look on the similarity values of each chemical group (figure 2). We could assume that for groups with higher similarities with CTLink chemicals, it would be easier to predict their target profile. Metabolites group, one of the groups with a lower % , is the group with a lower similarity to CTLink chemicals. As well, the opposite is happening with synthetic chemicals, where, moreover, we find a higher in-group similarity. The same is happening with natural products, that in addition have a high average

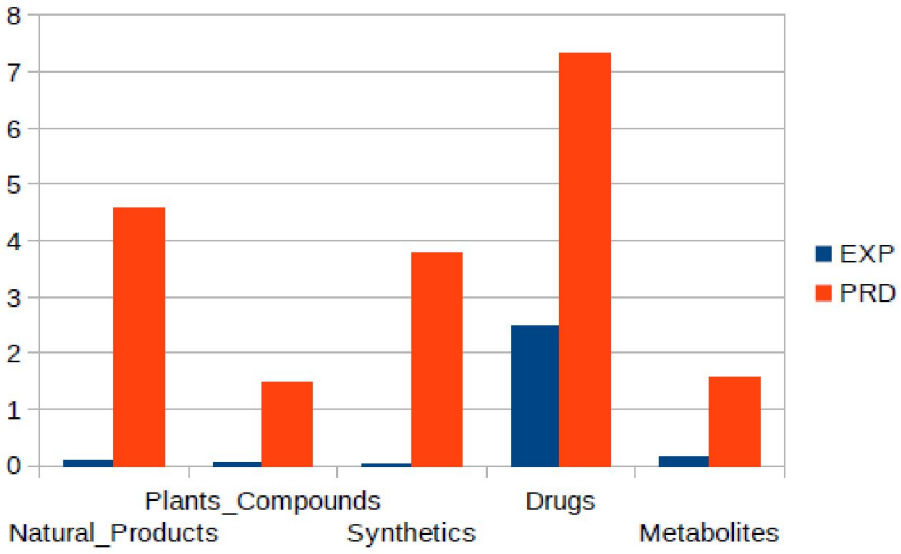


Figure 4: Promiscuity average values of each compound group. In blue there are promiscuity values from experimental data, in red with experimental and predicted

similarity value with synthetic chemicals (0.33).

On the other hand, lowest in-group similarities are found in Plant Compounds (0.16), Drugs (0.19) and Metabolites (0.25). The variability in drugs can be related with the use of Natural products on drug discovery to increase drugs chemical space.

Oppositly to natural products, we observe a high similarity between natural products. It could be consequence of the high number of compounds sharing a basic chemical structure, like flavonoids (Figure 3).

Polypharmacology analysis

For each group we measured the average number of target proteins for the compounds, promiscuity values (figure 4). Since the number of chemicals with experimental data is low in some groups like natural products and synthetic chemicals, we are obtaining very low average values.

We observe many differences on the promiscuity of each group of compounds; being drugs the chemicals with a higher promiscuity, in average they are targeting 2.48 proteins. Followed by Metabolites, with 0.14. For the other chemicals groups this value is lower, around 0.10.

After adding CTLink predicted data, the number of chemicals with target profile available is increasing in all chemical groups (Figure 4). Natural products and Synthetic chemicals are the groups with a most significant increase. It results on an average promiscuity of 4.57 for natural products, and 3,78 for synthetic chemicals. On the other hand, drugs promiscuity is even higher, increasing to 7.33 targets per compound

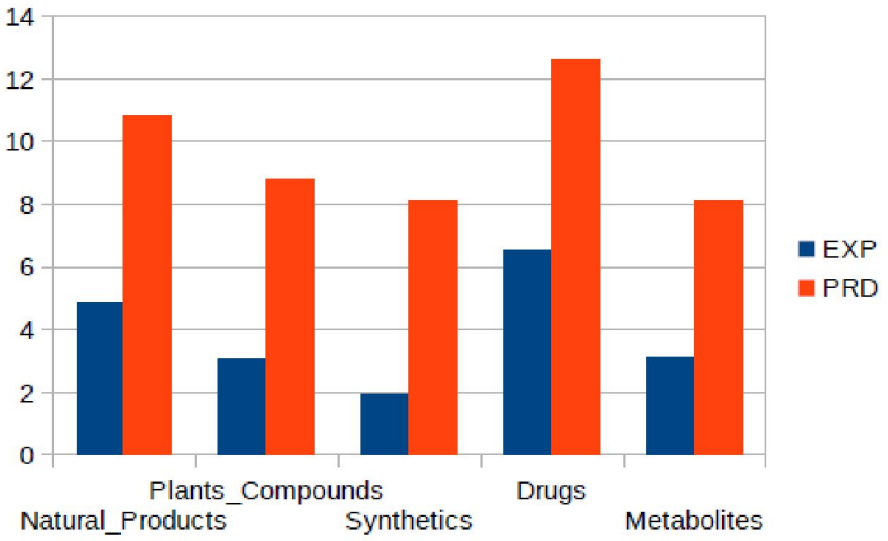


Figure 5: Promiscuity average values of each compound group. In blue there are promiscuity values from experimental data, in red with experimental and predicted

In order to analyse how vary the promiscuity value on each group of compounds after adding predicted data, we performed the same analysis but selecting only those compounds for which we already have experimental activity (figure 5). In general it can be appreciated an important variation in promiscuity values, especially in plant compounds and metabolites, that now have promiscuity values of 8.82 and 8.10 respectively. These values are very similar to that of synthethic chemicals, which is 8.12. So afterall we can see that there is a high predicted promiscuity in all groups. Natural products and drugs are also increasing their promiscuity, being 10.81 and 12.59, respectively.

Conclusions

As expected, drugs are the most well known compounds, and the group with a higher promiscuity (Tan et al., 2016), in average they are targeting more than 6 proteins, increasing to 12 after predicting the full target profile with CTLink.

CTLink virtual profiling is increasing substantially the promiscuity value for all chemical groups. In all them, for chemicals with known experimental activity, values are reaching average promiscuity values that are over 8. Thus, not only drug may be targeting a high number of proteins, also other compounds such metabolites. Thus, metabolic networks may have a high complexity.

Usually metabolic networks are represented in a linear way, with a few numbers of targets for each metabolite, but many are interacting with more than 2 proteins. Metabolite polypharmacology may be related with enzyme promiscuity (Gololobov et al., 1994). Recent literature shows that enzyme promiscuity may result on enzyme side-reactions that are used for obtaining diverse kinds of molecules (Piedrafita et al., 2015).

About natural products, they use to have a wide chemical space (Gu et al., 2013), and be structurally complex (Morrison and Hergenrother, 2014), which could be associated to a low promiscuity (Jalencas and Mestres, 2013). However, here we have found that they have high promiscuity average values in all the cases. They are the second compound group with a highest promiscuity, which is very similar to the drugs one.

High promiscuity values are found also on synthetic compounds, it is the biggest group of compounds but also the group with a highest in-group similarity, and the most similar group to CTLink chemicals. With all these factors we have been able to predict the activity profile of a high number of them. So, for Synthetic chemicals, as well as for Natural products, promiscuity is increasing after adding predicted data. This value is related with the fact that many of these synthetic chemicals may have been used as lead compounds, so they may have a high structural similarity to drugs.

On the other side, there are some other groups with low inter-group similarity values to CTLink compounds, such as Plant Compounds and specially HMDB metabolites. This is reflected on the low proportion of chemicals from these

groups for which we predict their target profile.

Otherwise, despite of there are many compounds with reported experimental activity (Figure 1), there's still many information to be added on their target profile, as can be observed in figure 5,

Finally, for human metabolites and plant compounds there is a low amount of chemicals with reported experimental data. But like for other groups, with CTLink we are able to perform the target profile for a significant number of compounds. There we are finding high promiscuity values. This suggest that chemical space of plant compounds could be higher than those of Natural Product Libraries. So it should be of interest to go in deep on these chemical activities research, in order to use their structures to expand to identify potential lead compounds which would allow to expand drug's chemical space. In a similar way, metabolite activities have yet to be explored for a high number of compounds, despite of knowing some reactions where they are taking part, they are interacting with many other proteins, and probably altering the metabolic flux.

References

Afendi, F.M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K., Ikeda, S., Takahashi, H., Altaf-Ul-Amin, M., Darusman, L.K., Saito, K., Kanaya, S., 2012. KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. Plant Cell Physiol. 53, e1. doi:10.1093/pcp/pcr165

Anighoro, A., Bajorath, J., Rastelli, G., 2014. Polypharmacology: Challenges and Opportunities in Drug Discovery: Miniperspective. J. Med. Chem. 57, 7874–7887. doi:10.1021/jm5006463

Duke, J.A., 2016. Dr. Duke's Phytochemical and Ethnobotanical Databases.

Gololobov, M.Y., Stepanov, V.M., Voyushina, T.L., Morozova, I.P., Adlercreutz, P., 1994. Side reactions in enzymatic peptide synthesis in organic media: Effects of enzyme, solvent, and substrate concentrations. Enzyme Microb. Technol. 16, 522–528. doi:10.1016/0141-0229(94)90024-8

Gu, J., Gui, Y., Chen, L., Yuan, G., Xu, X., 2013. CVDHD: a cardiovascular disease herbal database for drug discovery and network pharmacology. J. Cheminformatics 5, 51. doi:10.1186/1758-2946-5-51

Gu, S., Yin, N., Pei, J., Lai, L., 2013. Understanding molecular mechanisms of traditional Chinese medicine for the treatment of influenza viruses infection by computational approaches. Mol. Biosyst. 9, 2696–2700. doi:10.1039/c3mb70268e

Jalencas, X., Mestres, J., 2013. On the origins of drug polypharmacology. Med Chem Commun 4, 80–87. doi:10.1039/C2MD20242E

Kenny, B., Ballard, S., Blagg, J., Fox, D., 1997. Pharmacological Options in the Treatment of Benign Prostatic Hyperplasia. J. Med. Chem. 40, 1293–1315. doi:10.1021/jm960697s

Liscio, P., Camaioni, E., Carotti, A., Pellicciari, R., Macchiarulo, A., 2013. From polypharmacology to target specificity: the case of PARP inhibitors. Curr. Top. Med. Chem. 13, 2939–2954.

Morrison, K.C., Hergenrother, P.J., 2014. Natural products as starting points for the synthesis of complex and diverse compounds. Nat Prod Rep 31, 6–14. doi:10.1039/C3NP70063A

Novick, P.A., Ortiz, O.F., Poelman, J., Abdulhay, A.Y., Pande, V.S., 2013. SWEETLEAD: an in silico database of approved drugs, regulated chemicals, and herbal isolates for computer-aided drug discovery. PloS One 8, e79568. doi:10.1371/journal.pone.0079568

Piedrafita, G., Keller, M.A., Ralser, M., 2015. The Impact of Non-Enzymatic Reactions and Enzyme Promiscuity on Cellular Metabolism during (Oxidative) Stress Conditions. Biomolecules 5, 2101–2122. doi:10.3390/biom5032101

Ru, J., Li, P., Wang, J., Zhou, W., Li, B., Huang, C., Li, P., Guo, Z., Tao, W., Yang, Y., Xu, X., Li, Y., Wang, Y., Yang, L., 2014. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. J. Cheminformatics 6, 13. doi:10.1186/1758-2946-6-13

Tan, Z., Chaudhai, R., Zhang, S., 2016. Polypharmacology in Drug Development: A Minireview of Current Technologies. ChemMedChem 11, 1211–1218. doi:10.1002/cmdc.201600067

Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., Scalbert, A., 2013. HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res. 41, D801-807. doi:10.1093/nar/gks1065

Vidal, D., Garcia-Serna, R., Mestres, J., 2011. Ligand-based approaches to in silico pharmacology. Methods Mol. Biol. Clifton NJ 672, 489–502. doi:10.1007/978-1-60761-839-3_19

III.4: Completeness of metabolic databases

Joaquim Olivés, and Jordi Mestres*

Systems Pharmacology, Research Programme on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute and University Pompeu Fabra, Parc de Recerca Biomèdica, 08003 Barcelona, Catalonia, Spain

Abstract

Currently there are several chemicals databases containing information about metabolomes, This databases provides scientist with the most current and comprehensive metabolic data. Some of the most important databases are HMDB, KEGG and BioCyc. These resources have facilitated the research for thousands of studies, which, otherwise, may allow the expansion of these databases. An actualized maintenance of this data is difficult because of the quantity of published studies about metabolomics. As consequence, despite sharing a high amount of data, these database may contain also different data. On this study we have explored the completeness of these databases, analysing the amount on metabolites activity data that can be added to HMDB from other resources, KEGG, BioCyc and ChEMBL. In addition, we have applied an In silico methodolosy for predicting activity of these metabolites. This article described how metabolic network complexity is increasing after expanding data from HMDB

* Corresponding autor. *E-mail address:* jmestres@imim.es

Introduction

Metabolome can be defined as the complete collection of small molecule metabolites found in the human body. These molecules include peptides, amino acids, nucleic acids, carbohydrates, organic acids, vitamins, minerals, food additives, drugs, toxins, pollutants and any other chemical that humans ingest, metabolize, catabolise or come into contact with. Metabolome, in contrast to genome and proteome, is not easily defined, since it is not solely dictated by our genes. Metabolome consists of a mix of genetic and external factors. Some external factors that are contributing to metabolome are environment (like what we eat, breathe and drink), and microflora (the bacteria that live in our intestinal tract). So it is composed of both endogenous and exogenous compounds. Endogenous metabolites are small molecules synthesised by the enzymes encoded by genome or microbial genomes, while exogenous metabolites are foreign or xenobiotic chemicals consumed as foods or other consumables (Wishart et al., 2013; Zamboni et al., 2015).

A comprehensive knowledge of metabolism is essential for a better understanding of diseases, and exploring characteristics of organisms. In consequence, metabolomics has emerged as a functional methodology in a wide range of research areas such as toxicology, pharmacology, food technology, nutrition, microbial biotechnology, systems biology, and plant biotechnology. Because many metabolic mechanisms are yet to be well characterized, many studies focused on metabolome prediction have been performed (Cesare Marincola et al., 2015).

HMDB is a resource dedicated to provide the most current and comprehensive coverage of the human metabolome. It was released in 2007 and since then it has facilitated research for nearly 1000 published studies in metabolomics. Currently it has annotated more than 40000 metabolites, which includes both

'detected' metabolites (those with measured concentrations or experimental confirmation of their existence) and 'expected' metabolites (those for which biochemical pathways are known or human intake/exposure is frequent but the compounds has yet to be detected in the body) (Wishart et al., 2013).

HMDB annotations provide detailed compound description which includes compound synonyms, biofluid concentration, tissue location data and synthesis records. Furthermore, HMDB also contains data of proteins, diseases and pathways linked to metabolites.

In this part of the thesis we've tried to complement the data available in HMDB with the data available other databases with information available in other public databases, KEGG and HumanCyc, which are metabolome databases, and ChEMBL (Caspi et al., 2014; Gaulton et al., 2012; Kanehisa et al., 2014); and predicted data from CTLink, a similarity based virtual profiling software.

Methods

HMDB is composed by 41993 metabolites where we find 41681 different chemical structures. Some of the information available about metabolites in HMDB is their 'origin', which can be endogenous or exogenous. We can find 9 different origins and each compound can have more than one (Table 1). To perform these completeness study, from HMDB compounds we will use only endogenous compounds.

Origin	Metabolites	Metabolites MW<800; nºC>2; rotors<40 CTLink – Applicability domain
Endogenous	29208	11024
Food	32416	14044
Drug	1501	1392
Drug metabolite	920	879
Microbial	171	156
Toxin/Pollutant	163	132
Plant	149	139
Drug or steroid metabolite	32	32
Cosmetic	17	14

Table 1: Number of metabolites classified according to their origin. The total number and total number of metabolites within the applicability domain of CTLink

On the other hand, in KEGG we find 13329 compounds, while in BioCyc (HumanCyc) we have 1577 metabolites, from all these compounds, 4777 and 983, respectively, are associated to some enzyme. Finally, ChEMBL contains activity data for more than 400.000 chemicals with drug-like properties.

From each database we have downloaded the compound structures and selected those which have a molecular weight lower than 800 Armstrongs, more than 2 carbons, and less than 40 rotors, in order to avoid minerals, water and other small molecules that are not in the applicability domain of CTLink.

In CTLink and ChEMBL data, it is available the activity value in most of the reported interaction. For this study we have collected those interactions from ChEMBL with an activity higher than 5, while from CTLink we've selected those interactions with an activity higher than 5 in experimental interactions, and higher than 7 in predicted interactions.

Once data have been filtered we proceeded to associate metabolite with enzymes using the data available in each database. In BioCyc (HumaCyc) and KEGG metabolites are directly associated to EC (Enzyme Comission) number. However in HMDB, ChEMBL and CTLink compounds are linked to proteins, uniprot ID. In these cases, using the protein Uniprot ID we are able to associate these metabolites to EC numbers.

Results

After filtering the metabolites available in HMDB, KEGG and HumanCyc/MetaCyc, we have 11024, 12221 and 1415 metabolites, respectively. In table 2 it can be observed the data provided by each database.

Source	Compounds	Proteins	Enzymes (EC)
HMDB	11024	4140	1161
KEGG	12221	-	3529
BioCyc (HumanCyc)	1415	3217	1421
ChEMBL	424075	2881	798

Table 2: Drug-like compounds from each database, and the number of proteins and enzymes associated to these compounds.

From 11024 endogenous metabolites in HMDB, only 2051 are present in some of the others databases. The distribution of these compounds can be observed in the venndiagram of figure 1. KEGG is the database sharing a higher number of metabolites with HMDB, 1626. It is followed by ChEMBL, with 792 compounds, and finally HumanCyc, that is sharing almost the half of compounds with HMDB, 634. On the other hand, CTLink is able to predict some interaction for 3313 endogenous metabolites from HMDB, 951 of them included in other databases.

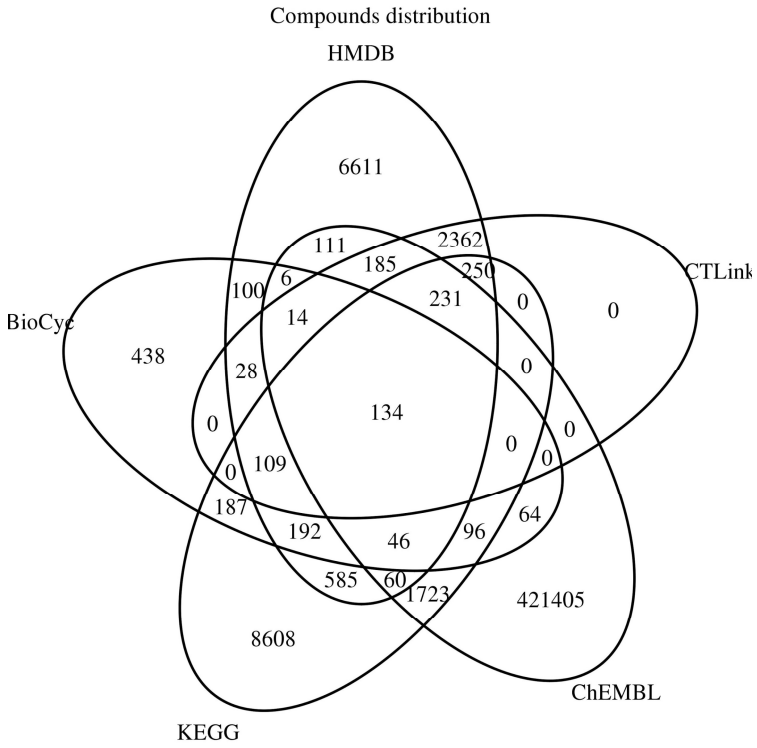


Figure 1

Distribution of endogenous druglike compounds in databases.

From 11024 endogenous metabolites in HMDB, only 2051 are present in some of the others databases. The distribution of these compounds can be observed in the venndiagram of figure 1. KEGG is the database sharing a higher number of metabolites with HMDB, 1626. It is followed by ChEMBL, with 792 compounds, and finally HumanCyc, that is sharing almost the half of compounds with HMDB, 634. On the other hand, CTLink is able to predict some interaction for 3313 endogenous metabolites from HMDB, 951 of them included in other databases.

Regarding on these 2051 HMDB metabolites we observe that only 913 are associated to some enzyme protein. They are linked to 1099 different EC numbers, and there is in total 6321 interactions with EC (Table 3).

Source	Compounds (*)	Compounds(*) (cumulative)	EC	EC (cumulative)	Interactions	Interactions (cumulative)
HMDB	913	913	1099	1099	6321	6321
HumanCyc	491	982 (+69)	973	1289 (+190)	2288	6946 (+625)
KEGG	783	1079 (+97)	865	1302 (+13)	2998	7308 (+362)
ChEMBL	219	1189 (+110)	148	1319 (+17)	386	7613 (+305)
CTLink	489	1319 (+130)	219	1333 (+14)	1005	8214 (+601)

Table 3: Compounds from each database that are also present in HMDB, the number of enzymes associated with, and the number of interactions between these metabolites and enzymes. Cumulative columns show the number of interacting compounds and enzymes after adding each database. (*)Compounds associated to some enzyme protein.

After adding the data available in HumanCyc, KEGG and ChEMBL, the number of interacting compounds and enzymes has increased 30.2% and 20%, respectively, while the number of interactions both components increases a 20.4%. After adding predicted data from CTLink this percentages are increasing to 44.5%, 21.3% and 30%, respectively (Figure 2).

Metabolites associated to a higher number of new interactions with enzymes added are coenzymes like ATP, ADP, NADH, NADPH, NAD, ADP and AMP. ATP and ADP are essential nucleoside phosphates to the flow of energy in living cells. Energy transfer is the result of dephosphorylation of ATP. While AMP is a nucleoside phosphate that can be produced during the synthesis of ATP or the hydrolysis of ADP or ATP (Berg et al., 2009). On the other hand, NAD+ and NADH are coenzymes involved in redox reactions, carrying electrons from one reaction to another. NAD+ is an oxidizing agent, accepts electrons from other molecules, and NADH is used as reducing agent to donate

electrons. Finally NADPH is a cofactor used in anabolic reactions as reducing agent (Ying, 2008). HumanCyc is the database adding more data about these metabolites. All them are necessary in a several quantity of reactions, so their activity is well known, as consequence it would be expected that their activity was completely reported. However, in many cases, in reactions data cofactors use to be excluded, including only the main substrates and products.

Some other compounds for which we are adding activity data are dopamine, glutamate and 3,3',4',5-Tetrahydroxystilbene. The first 2 are classified as "Detected and Quantified" in HMDB, while last one is only 'Expected'. From these 3 metabolites, dopamine and glutamate are synthesized in the body (Elsworth and Roth, 1997; Watford, 2015). 3,3',4',5-Tetrahydroxystilbene is a metabolite of resveratrol, which is found in wine (Maggiolini et al., 2005). Glutamate is one of the non-essential aminoacids, meaning that it is synthesized in the body, it is a key molecule in cellular metabolism, with many functional roles (Berg et al., 2002; Vazana et al., 2016)

Metabolites	HMDB	+ HumanCyc	+ KEGG	+ ChEMBL	+ CTLink
ATP	229	269 (+40)	272 (+3)	272 (+0)	173 (+1)
NADH	122	148 (+26)	164 (+16)	164 (+0)	165 (+1)
NADP	153	181 (+28)	191 (+10)	191 (+0)	191 (+0)
NAD	146	175 (+29)	182 (+7)	182 (+0)	183 (+1)
ADP	171	197 (+26)	200 (+3)	203 (+3)	203 (+0)
AMP	95	113 (+18)	113 (+0)	115 (+2)	116 (+1)
Dopamine	28	30 (+2)	30 (+0)	37 (+7)	58 (+21)
3,3',4',5-Tetrahydroxystilbene	0	0 (+0)	1 (+1)	3 (+2)	17 (+14)
L-Glutamic acid	57	67 (+10)	69 (+2)	69 (+0)	69 (+0)
Succinic acid	23	28 (+5)	30 (+2)	30 (+0)	30 (+0)

Table 4: Table with some of the compounds with more enzyme interactions added, and their sources. Data is cumulative

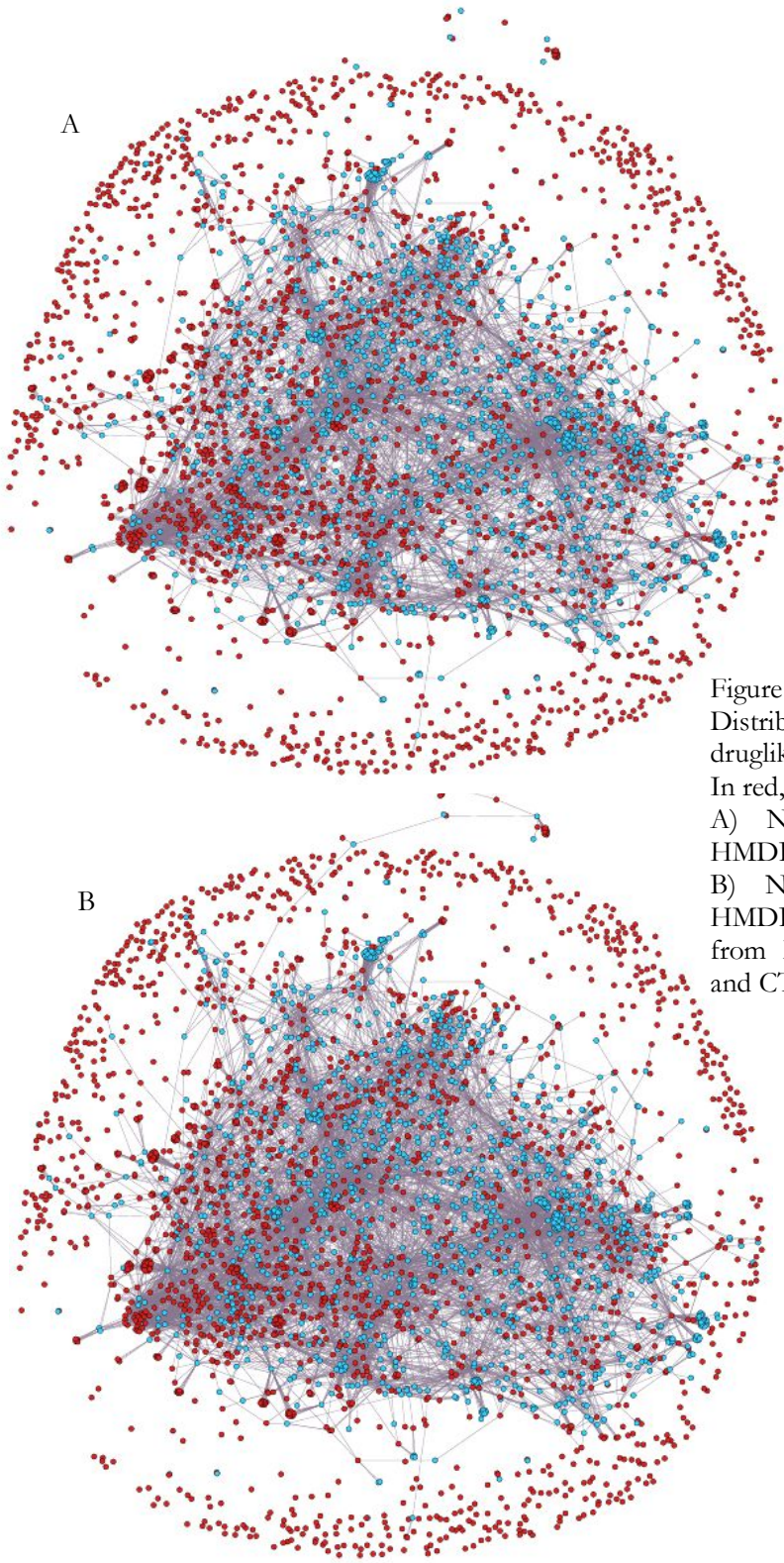


Figure 2:
Distribution of endogenous druglike compounds in databases. In red, metabolites, in blue EC.
A) Network with data from HMDB
B) Network with data from HMDB completed with new data from KEGG, BioCyc, ChEMBL and CTLink

On the other hand, table 5 shows some of the properties of both networks in figure 2. There is a significant increase on the average degree, from 3,716 to 4,833. With the increase of the average degree we would expect a decrease of network diameter; however it is increasing from 12 to 16 as consequence of the addition of new nodes to the network, which reduce the number of isolated components.

	HMDB	HMDB + HumanCyc + KEGG + ChEMBL + CTLink
Avg Degree	3,716	4,833
Network Diameter	12	16
Connected components	1397	756
Avg Path Length	4,549	4,529

Table 5: Properties of the networks from Figure 2

The initial number of components is 1397 and it is reduced to 756 after adding data and connecting many of them to the main network. Despite of this, the average path length decreases as result of the new connections. This new interactions can be from data of reactions not reported in HMDB or metabolite inhibitory activities. The first case would be important in the designing of synthetic metabolism, to optimize synthesis pathways (Bilgin and Wagner, 2012).

Conclusions

Metabolomics are a difficult omics science to be investigated because of its expensive experiments, but it has been extensively studied for some model organisms, such as Human, Escherichia coli, or Saccharomyces cerevisiae (Guo et al., 2013; Jewison et al., 2012). Nowadays, there are many databases collecting available metabolic data (Caspi et al., 2014; Kanehisa et al., 2014; Wishtart et al.,

2013). They are sharing an important amount of data, however, none of them is complete. One of the reasons is the difficulty of maintaining them completely actualized with recently published data. In this study we observe how metabolic data is varying between these databases and how it is increasing after checking overlapped data between several databases and adding specific data from other databases.

In this study we have focused on the available data in HMDB (Wishtart et al., 2013). Where a 25% of the compounds classified as endogenous are found in other sources. Their activity data could be extended a 20% using other metabolic databases, which demonstrates the incompleteness of this database. On the other hand, the other data sources could also be extended using HMDB database.

Despite of focus our study on the activity data of few HMDB metabolites found in other databases, other data sources, like HumanCyc, are also including human metabolites not reported in HMDB, such as 7alpha,12alpha-dihydroxy-4-cholesten-3-one and Lipoyl-AMP, which is found in HMDB but with 'origin' field not defined. In the same way, there are many metabolites found in HMDB that are not present in BioCyc. So, both databases are clearly complementary.

On the other hand, ChEMBL is focused in drugs and drug-like compounds (Gaulton et al., 2012), so we cannot find there many metabolites activity data. However, CTLink allow us to predict not reported activity for a few more metabolites, in total 2071. Through virtual profiling we are increasing the number of interactions of these metabolites around 30%.

This study reflects the high amount of missing data about human metabolism, one of the most studied metabolomes. Moreover, relations between metabolites and enzymes are not only substrate/product-enzyme, but metabolites may also have inhibitory activities. All these interactions are increasing the complexity of metabolic systems and their dynamics.

References

Berg, J., Hung, Y.P., Yellen, G., 2009. A genetically encoded fluorescent reporter of ATP/ADP ratio. Nat. Methods 6, 161–166. doi:10.1038/nmeth.1288

Berg, J.M., Tymoczko, J.L., Stryer, L., 2002. The First Step in Amino Acid Degradation Is the Removal of Nitrogen.

Bilgin, T., Wagner, A., 2012. Design Constraints on a Synthetic Metabolism. PLoS ONE 7. doi:10.1371/journal.pone.0039903

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Weerasinghe, D., Zhang, P., Karp, P.D., 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 42, D459–D471. doi:10.1093/nar/gkt1103

Cesare Marincola, F., Dessì, A., Corbu, S., Reali, A., Fanos, V., 2015. Clinical impact of human breast milk metabolomics. Clin. Chim. Acta Int. J. Clin. Chem. 451, 103–106. doi:10.1016/j.cca.2015.02.021

Elsworth, J.D., Roth, R.H., 1997. Dopamine synthesis, uptake, metabolism, and receptors: relevance to gene therapy of Parkinson's disease. Exp. Neurol. 144, 4–9. doi:10.1006/exnr.1996.6379

Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J.P., 2012. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40, D1100–D1107. doi:10.1093/nar/gkr777

Guo, A.C., Jewison, T., Wilson, M., Liu, Y., Knox, C., Djoumbou, Y., Lo, P., Mandal, R., Krishnamurthy, R., Wishart, D.S., 2013. ECMDB: The E. coli Metabolome Database. Nucleic Acids Res. 41, D625–D630. doi:10.1093/nar/gks992

Jewison, T., Knox, C., Neveu, V., Djoumbou, Y., Guo, A.C., Lee, J., Liu, P., Mandal, R., Krishnamurthy, R., Sinelnikov, I., Wilson, M., Wishart, D.S., 2012. YMDB: the Yeast Metabolome Database. Nucleic Acids Res. 40, D815–D820. doi:10.1093/nar/gkr916

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2014. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42, D199-205. doi:10.1093/nar/gkt1076

Maggiolini, M., Recchia, A.G., Bonofiglio, D., Catalano, S., Vivacqua, A., Carpino, A., Rago, V., Rossi, R., Andò, S., 2005. The red wine phenolics piceatannol and myricetin act as agonists for estrogen receptor alpha in human breast cancer cells. J. Mol. Endocrinol. 35, 269–281. doi:10.1677/jme.1.01783

Vazana, U., Veksler, R., Pell, G.S., Prager, O., Fassler, M., Chassidim, Y., Roth, Y.,

Shahar, H., Zangen, A., Raccah, R., Onesti, E., Ceccanti, M., Colonnese, C., Santoro, A., Salvati, M., D'Elia, A., Nucciarelli, V., Inghilleri, M., Friedman, A., 2016. Glutamate-Mediated Blood-Brain Barrier Opening: Implications for Neuroprotection and Drug Delivery. J. Neurosci. Off. J. Soc. Neurosci. 36, 7727–7739. doi:10.1523/JNEUROSCI.0587-16.2016

Wang, Y., Xu, N., Ye, C., Liu, L., Shi, Z., Wu, J., 2015. Reconstruction and in silico analysis of an Actinoplanes sp. SE50/110 genome-scale metabolic model for acarbose production. Front. Microbiol. 6. doi:10.3389/fmicb.2015.00632

Watford, M., 2015. Glutamine and glutamate: Nonessential or essential amino acids? Anim. Nutr. 1, 119–122. doi:10.1016/j.aninu.2015.08.008

Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., Scalbert, A., 2013. HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res. 41, D801-807. doi:10.1093/nar/gks1065

Ying, W., 2008. NAD+/NADH and NADP+/NADPH in cellular functions and cell death: regulation and biological consequences. Antioxid. Redox Signal. 10, 179–206. doi:10.1089/ars.2007.1672

Zamboni, N., Saghatelian, A., Patti, G.J., 2015. Defining the Metabolome: Size, Flux, and Regulation. Mol. Cell 58, 699–706. doi:10.1016/j.molcel.2015.04.021

Zou, W., Zhou, M., Liu, L., Chen, J., 2013. Reconstruction and analysis of the industrial strain Bacillus megaterium WSH002 genome-scale in silico metabolic model. J. Biotechnol. 164, 503–509. doi:10.1016/j.jbiotec.2013.01.01

III.5: Construction of genome-scale metabolic networks of Mycoplasma pneumoniae

Joaquim Olivés, and Jordi Mestres*

Systems Pharmacology, Research Programme on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute and University Pompeu Fabra, Parc de Recerca Biomèdica, 08003 Barcelona, Catalonia, Spain

Abstract

Mycoplasma pneumoniae is a pathogenic bacterium belonging to the class Mollicutes. Due to its small genome and its ability to grow in vitro, it has become and interesting model organism for systems biology approaches. In this study we have reconstructed its metabolic network using a genomic-scale approach that has been validated with the data of previous experimental study. In our genome-scale approach we are implementing also virtual profiling, which may contribute to get more activity data, increasing the complexity of the metabolic network. According to the results, 837 metabolites and 481 enzymes have been identified. This model would be useful on the mycoplasma metabolome research.

* Corresponding autor. *E-mail address:* jmestres@imim.es

Introduction

Mycoplasma pneumoniae is a human parasite that colonizes the lung epitelium, it is involved in several diseases, amongst them walking pneumonia (Waites and Talkington, 2004). It is a bacterium belonging to *Mollicutes,* a class of bacteria lacking cell walls and typically having small genomes under 1000kb. This bacterium is surrounded by a cytoplasmic membrane only, which contains cholesterol as indispensable component (Baseman and Tully, 1997).

Mycoplasma causes up to 40% of the community acquired pneumonias. Despite of the infection is mild in most of the cases; it is a significant cause of hospitalisation in elderly population and immunocompromised patients. Manifestations of M. pneumoniae infections can range from self-limiting upper respiratory illness to severe pneumonia (Chaudhry et al., 2016). In addition, 20% of these infections can be accompanied by extrapulmonary complications, due to the direct infection of other organs, or infection-associated autoimmune phenomena (Xiao et al., 2015).

M. pneumoniae has been highly investigated, and several virulence mechanisms have been identified. As surface parasite, since it is incapable of de novo synthesis of aminoacids due to its highly reduced genome, it requires close association with the host cell to survive. Moreover, mycoplasma also requires host cholesterol in their growth. The adherence to host respiratory epitelium initiate the cell injury, tissue disruption and cytotoxic effects (Himmelreich et al., 1996).

Because of its growth requirements, initially it was not considered as an organism suitable for basic studies, until 1990s, with the appearance of new methods of molecular biology (Dybvig, 1990). Its exceptional simplicity and reduced genomic complexity has led to mycoplasma to become appreciated for researchers. Beyond dealing with diseases that it causes, it have been used as simplified model for cell structure and genome analysis (Balish, 2014). So it has

become a model organism for bacterial systems biology.

Mycoplasma pneumoniae is the most comprehensively analysed specie of Mycoplasma, there are many recent studies characterizing its transcriptome, proteome and metabolome (Xiao et al., 2015). Yus et al (2009) constructed a manually curated metabolic network for *Mycoplasma pneumoniae*, consisting of 189 reactions catalysed by 129 enzymes, and 225 metabolites. Compared with more complex bacteria, *M. pneumoniae* metabolic network results to have a more linear topology and contained a higher fraction of multifunctional enzymes.

Currently, genomic-scale metabolic network reconstruction has become and important tool for studying system biology of metabolism. This process extract biochemical data from genome annotations to computational interconnect it with genomic data available for another organism. For dozens of organisms a genomic-scale metabolic model has been constructed, such as *Haemophilus influenza Rd* (Edwards and Palsson, 1999)*, Saccharomyces cerevisiae* (Nookaew et al., 2011)*,* rice (Liu et al., 2013) or human (Ryu et al., 2015). Metabolic networks allow a better understanding of metabolism and cellular behaviour, which facilitates biological studies in a variety of applications, including network properties, metabolic engineering and drug discovery (Cazzaniga et al., 2014; Kell and Goodacre, 2014; Simeonidis and Price, 2015).

The aim of this study is to predict the metabolites present in organisms using the data available in BioCyc. This database is an assortment of more than 1700 organism specific Pathway/Genome Databases. They provide metabolites, enzymes, reactions and metabolic pathways. One of the databases found in BioCyc is MetaCyc, which provides experimentally determined metabolic pathways and enzymesof organisms from all domains of life (Caspi, 2015). Besides we have applied virtual profiling on the BioCyc metabolites before the metabolome prediction. On the other hand, we are also applying virtual profiling to predict chemical off-targets. This may increase metabolic networks

complexity.

In this study we have predicted the metabolome of *Mycoplasma pneumoniae* through a genome-scale metabolic reconstruction framework. Since metabolome of M. pneumoniae has been already constructed with experimental methodologies by Yus *et al,* we are able to check the recall and precision of our prediction method. Moreover, we will analyse how is increasing the complexity of the metabolic network.

Methods

To reconstruct metabolic network of *Mycoplasma pneumoniae* we are using available data from BioCyc. Our framework is divided in 3 steps:

1- Retrieving from MetaCyc: all reported data from Mycoplasma is retrieved.

2- Genome-scale reconstruction framework (MProjection): Mycoplasma genome is projected to BioCyc metabolomic data using OrthoMCL (Fischer et al., 2011).

3- Virtual profile and projection (CTLink): For the metabolites predicted for mycoplasma we perform a ligand-based virtual profil using CTLink. Using the same methodology, we are projecting genomic data on virtual profile results.

-Metabolome and reactions

For metabolome reconstruction framework we've used the data available in BioCyc and CTLink. From BioCyc we downloaded the data from databases belonging to Tier 1, which are the databases that have received at least one year literature based manual curation (Caspi et al., 2010). Currently it is composed of

7 databases. Among them, the major database is MetaCyc, which contains experimental information of 2332 organisms from 1086 species. The other 6 databases are Humancyc (*Homo sapiens*), which contain metabolic information of humans, EcoCyc (*Escherichia coli*), AraCyc (*Arabidopsis thaliana*), YeastCyc (*Saccharomyces cerevisiae*), LeishCyc (*Leishmania major* Friedlin) and TrypanoCyc (*Trypanosoma brucei*) (Caspi et al., 2012; Shameer et al., 2015).

From each organism of BioCyc we are retrieving their metabolites, and the proteins (uniprot ID) and enzymes that are interacting with. The structure of metabolites is downloaded from PubChem through the PubChem ID available in BioCyc.

BioCyc compounds are identified inside each database with a BioCyc_ID, and have associated a PubChem ID. We have found that the same BioCyc ID can be associated to different PubChem ID in different databases; moreover one PubChem ID can be associated to different BioCyc ID. In most of these cases they are the same compounds with a different protonation or deprotonation.

-**Genomes**

BioCyc uses its own ID for proteins, but many of them are linked to Uniprot ID. In these cases we have used the genomic sequences from Uniprot Database. While for species where BioCyc data is not associated to Uniprot DB, such as TrypanoCyc and LeishCyc, we used the genomic data of its own database.

Identification of metabolites has been performed through the projection by orthology of proteins associated to the metabolites onto the genome of the organism whose metabolome we are predicting.

-Virtual profiling

For the virtual profiling we have used CTLink. It is a software for large off-target pharmacology and predictive safety of small molecule pharmaceuticals, cosmeceuticals and agrochemicals. It creates chemical pharmaceutical profiles using ligand-based approaches and cross pharmacology. The data used for these predictions is extracted from several public databases including ChEMBL (Gaulton et al., 2012), DrugBank (Wishart et al., 2006), BindingDB (Gilson et al., 2016), IUPHARdb (Southan et al., 2016), PDSP (Roth et al., 2004) and affinDB (Block et al., 2006).

The activities obtained from CTLink are filtered considering its confidence score, which is related with predicted activity value. The cut-off used for confidence score is 0.7.

-Orthology

Orthologous mapping of genes was calculated with OrthoMCL-DB Version 5. We use the service of the OrthoMCL-DB Website, which maps our proteins to OrthoMCL-DB groups. In the mapping process, this tool performs a BLASTP against all the proteins in OrthoMCL-DB, using a cut-off of 1e-5 and 50% match. Proteins are assigned to the group containing its best hit. If the best matching protein isn't assigned to any group, it is assigned to NO_GROUP (Fischer et al., 2011).

Results and Discussion

BioCyc Data

From Biocyc database it has been collected 12604 metabolites associated to 10158 InchiKeys. 7297 of them are part of some reaction. Reactions may be

directly linked to proteins and their uniprot ID or BioCyc enzyme IDs (that may be linked to Uniprot IDs). In total there are 5768 metabolites associated to some uniprot ID. On the other hand, through CTLink we are able to predict or collect target proteins for 2828 metabolites.

	Compounds			Enzymes	Proteins	Interactions	
	Total	Interacting					
		Enzymes	Proteins			Enzymes	Proteins
MetaCyc	12135	7210	5335	5643	19821	29231	80526
YeastCyc	1151	928	891	849	1190	4096	5465
HumanCyc	2014	1415	1419	1421	3217	6130	12250
TrypanoCyc	944	757	463	758	978	3448	3596
LeishCyc	722	507	464	526	952	2341	3659
AraCyc	2822	2248	2053	1508	5895	8280	31069
EcoCyc	3007	1030	1101	1103	1341	4928	5554
TOTAL	12604	7297	5768	5704	27844	32833	101137
CTLink	2828	-	2828	-	2292	-	18835
TOTAL	12604	7297	7173	5704	28868	26149	119312

Table 1: Data extracted from each BioCyc Database and the results from the profile in CTLink.

Since we are also using CTLink to predict the activities, the number of metabolites associated to some reaction or protein is increased to 7173, a 56.9% of the compounds from BioCyc.

From MetaCyc we extract metabolic information for about 1086 different organism from many different kingdoms and orders. Bacteria is the kingdom where we found more different species, 286, followed by Viridiplantae (green plants) which includes 265 different species. On the other hand, MetaCyc contains 230 species from Metazoa kingdom. Moreover, we can find also metabolic data for many strains. Bacteria kingdom is also where we find a

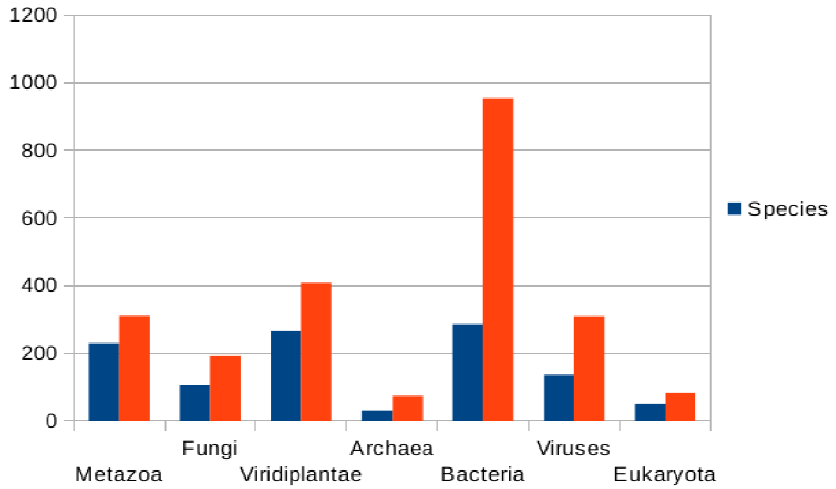


Figure 1: Distribution by kingdoms of species and organisms found in MetaCyc. In blue there is the number of species, and in red the total number of strains

higher number of strains (Figure 1).

From MetaCyc we extract metabolic information for about 1086 different organism from many different kingdoms and orders. Bacteria is the kingdom where we found more different species, 286, followed by Viridiplantae (green plants) which includes 265 different species. On the other hand, MetaCyc contains 230 species from Metazoa kingdom. Moreover, we can find also metabolic data for many strains. Bacteria kingdom is also where we find a higher number of strains (Figure 1).

Regarding on the average number of compounds and enzymes from each kingdom; Archaeas are the organisms in MetaCyc with a higher average number of metabolites linked to reactions, 18,38; followed by Viridiplantae, Bacteria, Fungi and Metazoa, with and average number of 14,79, 14,55, 13,16 and 11,94 metabolites per organism, respectively. These 5 kingdom are also those with a higher average number of EC and proteins, which are found between 6,25 and

9,59, and 29,33 and 25,51 respectively. On the other hand, Virus kingdom is where we find a lower quantity of data in general, the average number of compound is 2,26, and in EC and proteins is 1,62 and 1,4 respectively.

Organism	Kingdom	Metabolites*	EC	Protein	Metabolite-EC interactions	Metabolite-Protein interactions
Arabidopsis thaliana	Viridiplantae	1209	636	1047	4109	4739
Escherichia coli K-12	Bacteria	1030	995	1303	5828	4656
Homo sapiens	Metazoa	961	767	1054	4356	4098
Saccharomyces cerevisiae s288c	Fungi	657	550	739	3097	2917
Bacillus subtilis subsp. subtilis str. 168	Bacteria	430	363	404	1330	1396
Rattus norvegicus	Metazoa	389	317	454	1334	1433
Mus musculus	Bacteria	298	237	352	771	950
Haemophilus influenzae Rd KW20	Metazoa	291	273	284	981	972
Methanocaldococcus jannaschii DSM 2661	Archaea	268	199	224	818	826
Schizosaccharomyces pombe 972h	Fungi	248	189	241	718	770
...						

Table 2: Table with the organism for which there is more available data in MetaCyc. (*) Number of metabolites directly associated to some reaction.

These numbers are representative of the little knowledge of most of the organisms; since, for example, for *Escherichia coli* it is reported the activity for around 1000 metabolites, but the average number of metabolites for organisms from Bacteria Kingdom is 14,55.

Organisms with a largest amount of metabolic data available in MetaCyc are *Arabidopsis thaliana, Escherichia coli, Homo sapiens, Saccharomyces cerevisiae, Bacillus subtilis and Rattus norvegicu*s (Table 2)

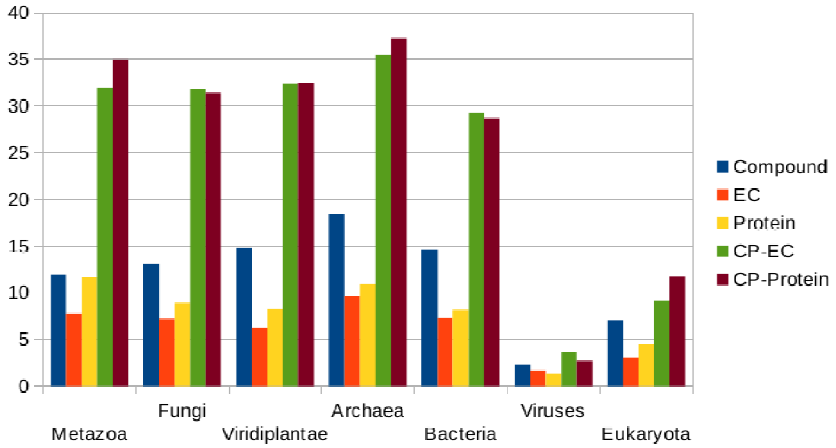


Figure 2: Distribution of data in MetaCyc classified by kingdoms. Each column is the average amount of data of the organisms iof each kingdom. CP=Compound

Prediction of Mycoplasma pneumoniae metabolism.

Mycoplasma pneumoniae is one of the organisms that are present in MetaCyc. In this database there are reported 89 compounds associated to 78 enzymes (EC). In total there is 268 interactions compounds-EC, which is a lower number than the total number of interactions reported by Yus *et al*, 629. Almost 70% of MetaCyc interactions are found in Yus *et al* study, and most of the compounds and EC are shared. So, many of the remaining 30% of interactions may be between metabolites and/or enzymes reported by Yus *et al*. These data can be observed in table 3, where the data obtained in each step of the framework is deconvulated.

	Metabolites	EC	Proteins		Interactions (CP-EC)	Interactions (CP-Protein)
			Enzymes	No-Enzymes		
Yus *et al*	225	134	0	0	629	0
MetaCyc	89	76	0	0	268	0
MProjection	830	467	0	9	1822	114
MetaCyc U MProjection	831	470	0	9	1840	114
CTLink	156	112	0	22	270	55
MProjection U CTLink	836	478	0	24	1975	162
MetaCyc U MProjection U CTLink	837	481	0	24	1992	162

Table 3: *Mycoplasma pneumoniae* data obtained from each source, and from each step of the prediction framework.

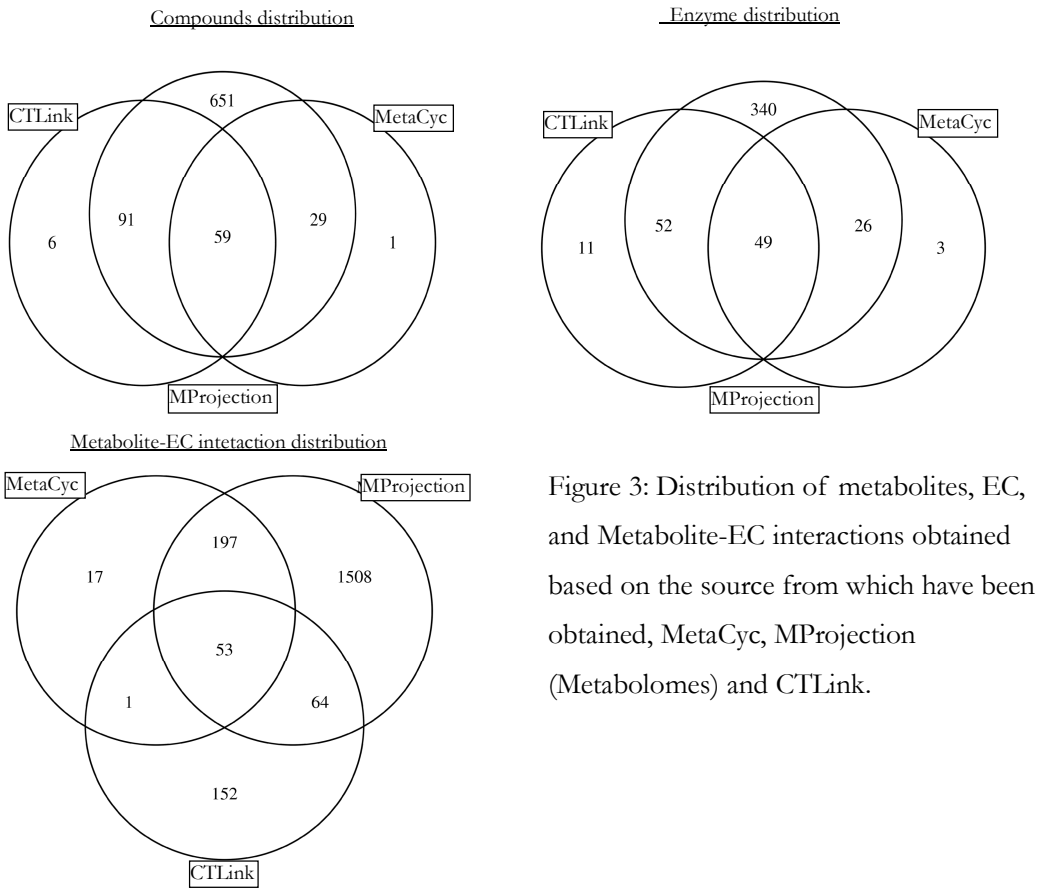


Figure 3: Distribution of metabolites, EC, and Metabolite-EC interactions obtained based on the source from which have been obtained, MetaCyc, MProjection (Metabolomes) and CTLink.

Through the projection of Mycoplasma genomic data to organisms from BioCyc and MetaCyc we are predicting 830 metabolites and 467 enzymes in 1822 interactions (Table 3). These predictions are recovering most of the experimental data reported in MetaCyc for Mycoplasma, almost 100% of the metabolites, EC and interactions EC-Metabolite are recovered (Figure 3). On the other hand, CTLink reports new interactions for 156 of the 831 metabolites obtained from MetaCyc and predicted metabolites. Using CTLink we don't obtain many new EC, only 3, but we get new interactions between metabolites and EC collected in the previous steps.

In total we have obtained 837 metabolites and 481 EC in 1992 interactions. Almost a 20% of these data is matching with the experimental data of Yus *et al.* MetaCyc data is where we have a higher number of recovered data, which is a

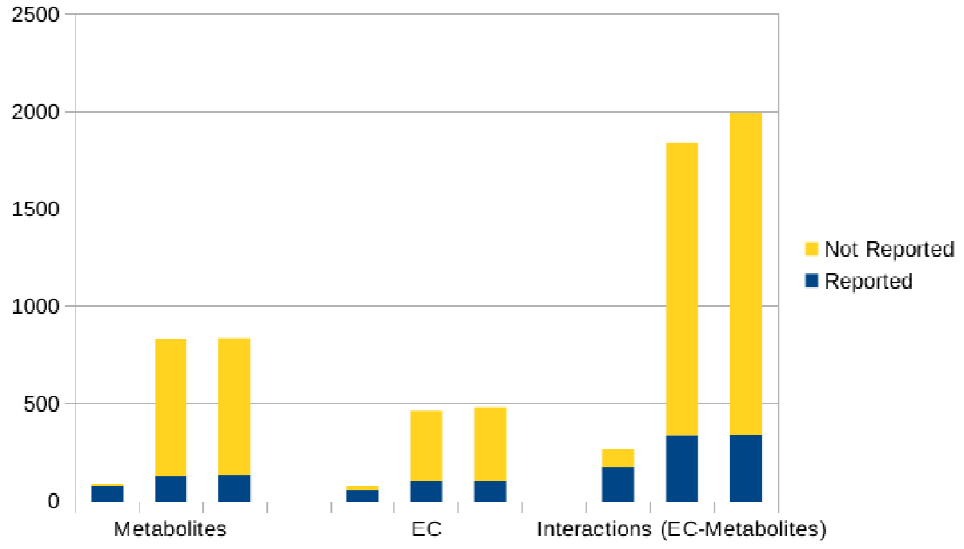


Figure 4: Evolution of the data obtained after each step of the framework (MetaCyc, MProjection and CTLink), and how it increase respect experimental data from *Yus et al* (reported).

89% of metabolites, 78% of enzymes and 66% of interactions. With genome projection we are increasing the recall, but the % of recovered data in respect the total predicted (precision) is decreasing.

Some examples of predicted interactions are represented in table 4. They are the interaction of 5,10-methenyltetrahydrofolate with 3.5.4.9 (Methenyltetrahydrofolate cyclohydrolase), 1.5.1.5 (Methylene tetrahydrofolate dehydrogenase) and 6.3.3.2 (5-Formyltetrahydrofolate cyclo-ligase), and the interactions of 3-phospho-D-glycerate with 2.7.2.3 (phosphoglycerate kinase), 5.4.2.12 (phosphoglycerate mutase) and 5.3.1.1 (triose-phosphate isomerase). 3 of these 6 predicted interactions are reported by Yus *et al.*

5,10-methenyltetrahydrofolate is a form of tetrahydrofolate that is an intermediate in metabolism. It is a coenzyme that accepts and donates methenyl groups. In BioCyc we find this compound in all the 7 databases, but it is associated to some reaction in only 4 of them, YeastCyc, LeishCyc, TrypanoCyc and AraCyc. In *Saccharomyces cerevisiae* and *Arabidopsis thaliana* we have found that 3 of the associated proteins are orthologs of 2 *Mycoplasma pneumoniae* genes. Specifically, yeast genes CT1TM_YEAST and C1TC_YEAST and Arabidopsis genes FOLD4_ARATH and FOLD2_ARATH are orthologs of FOLD_MYCPN, from *Mycoplasma pneumoniae*. All these genes are associated to the EC number 1.5.1.5 and 3.5.4.9, except C1TC_YEAST and FOLD2_ARATH, which are only associated to 1.5.1.5. So we have linked this metabolite to both ECs through FOLD_MYCPN. These interactions are confirmed by Yus *et al.* On the other hand, this metabolite is also linked to FTHC_YEAST and SFCL_ARATH, associated to EC number 6.3.3.2. They are orthologs of MTHFS_MYCPN, so we are also linking 5,10-methenyltetrahydrofolate to 6.3.3.2.

About 3-phospho-D-glycerate, in a similar way, it is reported to be associated with proteins from 212 different organisms. For 8 of these organisms,

Arabidopsis thaliana, Candida boidinii, Escherichia coli, Homo sapiens, Lactobacillus delbrueckii, Saccharomyces cerevisiae, Spinacia oleracea and *Thermotoga maritima*, some of the associated proteins are orthologs of PGK_MYCPN from *Mycoplasma pneumoniae*. These proteins are associated to EC-2.7.2.3. We also find a group of proteins associated to EC-5.4.2.12 that are orthologs of GMPI_MYCPN. Finally, from CTLink we obtain that this compound is associated to human gene TPIS_HUMAN, EC-5.3.1.1, an ortholog of TPIS_MYCPN, which is associated to the same EC number. So we linked 3-phospho-D-glycerate to 2.7.2.3, 5.4.2.12 and 5.3.1.1. From these interactions, the first one is confirmed in Yus *et al.*

Metabolite	Protein	EC	Organism	Mycoplasma Ortholog
5,10-Methenyl-tetrahydrofolate	YBR084W-MONOMER C1TM_YEAST	1.5.1.5, 3.5.4.9	*Saccharomyces cerevisiae*	FOLD_MYCPN
	YGR204W-MONOMER C1TC_YEAST	1.5.1.5		
	AT4G00620-MONOMER FOLD4_ARATH	1.5.1.5, 3.5.4.9	*Arabidopsys thaliana*	
	AT3G12290-MONOMER FOLD2_ARATH	1.5.1.5		
	YER183C-MONOMER FTHC_YEAST	6.3.3.2	*Saccharomyces cerevisiae*	MTHFS_MYCPN
	AT5G13050-MONOMER SFCL_ARATH		*Arabidopsis thaliana*	
3-phospho-D-glycerate	HS02359-MONOMER PGK1_HUMAN	2.7.2.3	*Homo sapiens*	PGK_MYCPN
	HS10215-MONOMER PGK2_HUMAN			
	MONOMER-13169 Q5KTR2_CANBO		*Candida boidinii*	
	PGK PGK_ECOLI		*Escherichia coli K-12*	
	CPLX-1864 PGKT_THEMA		*Thermotoga maritima MSB8*	
	MONOMER-12711 PGKH_SPIOL		*Spinacia oleracea*	
	AT3G12780-MONOMER		*Arabidopsis thaliana*	

	PGKH1_ARATH			
	AT1G56190-MONOMER PGKH2_ARATH			
	PGK_LACDE		*Lactobacillus delbrueckii*	
	MONOMER-9165 PMGI_MAIZE	5.4.2.12	*Zea Mays*	GMPI_MYCPN
	AT1G09780-MONOMER PMG1_ARATH		*Arabidopsis thaliana*	
	AT3G08590-MONOMER PMG2_ARATH			
	PMGI-MONOMER GMPI_ECOLI		*Escherichia coli*	
	TPIS_HUMAN	5.3.1.1	*Homo sapiens*	TPIS_MYCPN

Table 4: Example of some of the results obtained. In the first column there is the compound predicted, and in the last column the Mycoplasma Protein to which it have been associated. Protein column contains the BioCyc protein to which this compound in associated, EC contains the EC of the reactions associated to these proteins, while organism column contains the source organism of this data.

In total 521 organisms from BioCyc are contributing to Mycoplasma metabolism prediction. 477 of them are reporting some data from Yus *et al.* In figure 5 it can be observed the evolution of cumulative metabolic data obtained from each organism sorted randomly. In this plot we observe how data is increasing progressively after each projection, however, we see that it isn't increasing continuously. Usually the species with more available data are those which are adding a higher number of new data. The higher peaks observed in figure 5 are corresponding to *Arabidopsis thaliana, Escherichia coli K-12, Homo sapiens*, *Saccharomyces cerevisiae s288c* and *Bacillus subtilis subsp. subtilis str. 168*.

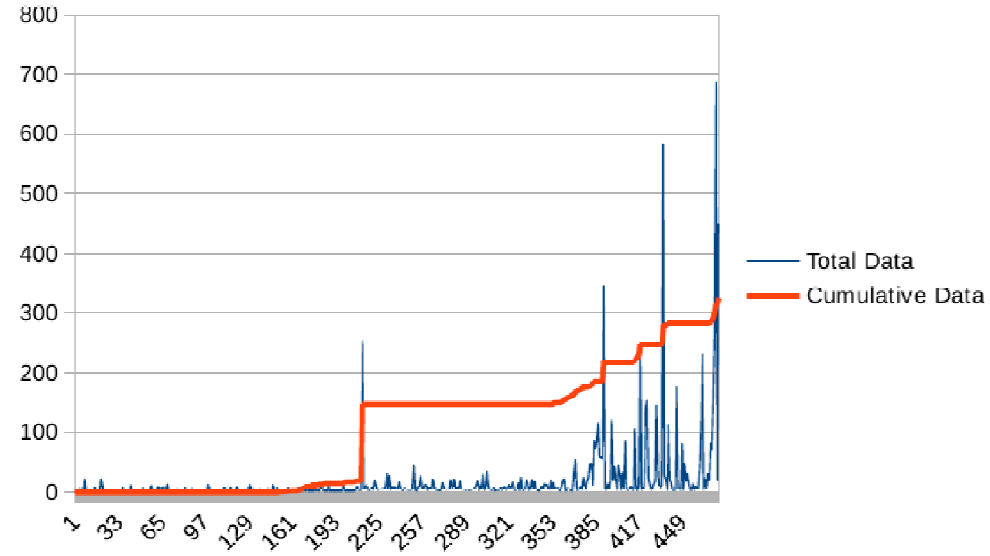


Figure 5: Contribution of each organism to the data obtained. Sorting randomly this species.

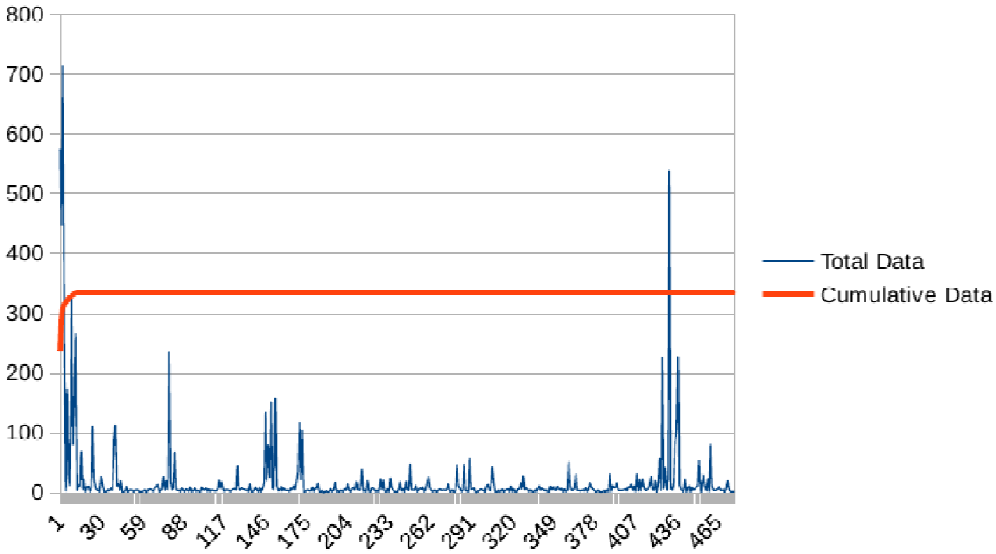


Figure 6: Contribution of each organism to the data obtained. Sorting species through their contribution on adding new data.

If we sort the species by the quantity of data that they are contributing to the metabolome prediction, we see only 13 organisms can be necessary to obtain it. These organisms are *Escherichia coli K-12, Homo sapiens, Arabidopsis thaliana*, *Leishmania major*, *Mycopllasma arginini, Lactococcus lactis subsp. Lactis*, *Pseudomonas aeruginosa*, *Lactobacillus sakei, Saccharomyces cerevisiae, Geobacillus stearothermophilus*, *Synechocytis sp. PCC 6803*, *Bacillus subtilis subsp. subtilis* and *Haemophilus ducreyi*. Their individual data contribution can be observed in more detail on figure 6 and table 5, with *Escherichia coli* we are already obtaining more than the 70% of the total data from Yus *et al,* and together with *Homo sapiens* and *Arabidopsis thaliana* is almost 93%.

	Total Data (Metabolite-EC)	Mycoplasma data (cumulative)	New data apported (Metabolite-EC)
Escherichia coli K-12	575	241	241
Arabidopsis thaliana	714	299	58
Homo sapiens	448	322	23
Leishmania major	252	328	6
Saccharomyces cerevisiae	332	332	4
Mycoplasma arginini	3	335	3
Geobacillus stearothermopilus	81	337	2
Synechocytis sp PCC 6803	143	338	1
Bacillus subtilis	266	339	1

Table 5: Table with the organisms which are adding more new data in a progressive form to the Mycoplasma metabolome predicted. In this case we are representing only data about interactions Metabolite–EC

Despite of most of the data can be obtained from *Escherichia coli,* there are several other organisms from which most of these data can be obtained. In figure 7 we can observe how many organisms are the sources of each interaction Compound-EC. Despite there are many interactions which are reported only in 1 or 2 organisms, most of them can be obtained from between 3 organisms and 40. In this histogram it can also be observed some interactions obtained from almost 60 and 90 organisms, they are those of the reaction catalysed by enzyme 1.2.1.12 (glyceraldehyde-3-phosphate dehydrogenase), *D-glyceraldehyde 3-phosphate + NAD^+ + Phosphate = 3-phospho-D-glyceroyl phosphate + NADH + H^+* and 2.7.2.3 (phosphoglycerate kinase), *ATP + D-Glyceraldehyde 3-phosphate = ADP + 3-phospho D-glyceroyl phosphate.*

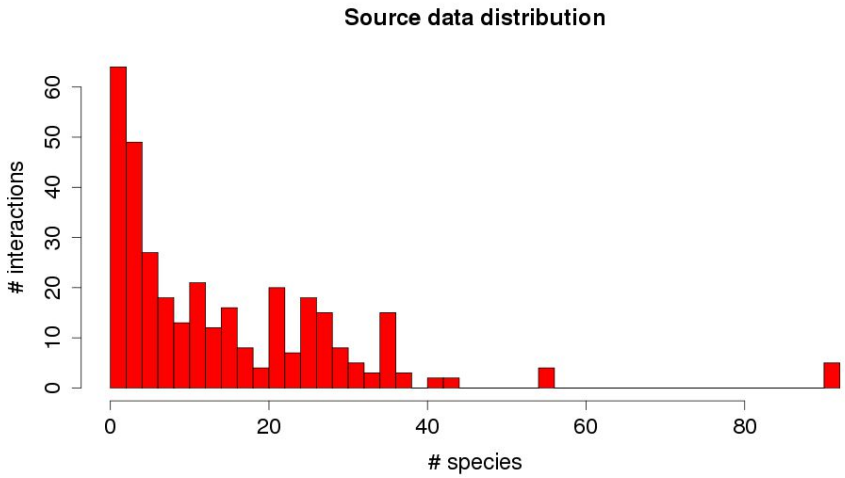


Figure 7: Histogram with the distribution of the interactions predicted. Each column represent the quantity of interactions predicted by the values of x-axe.

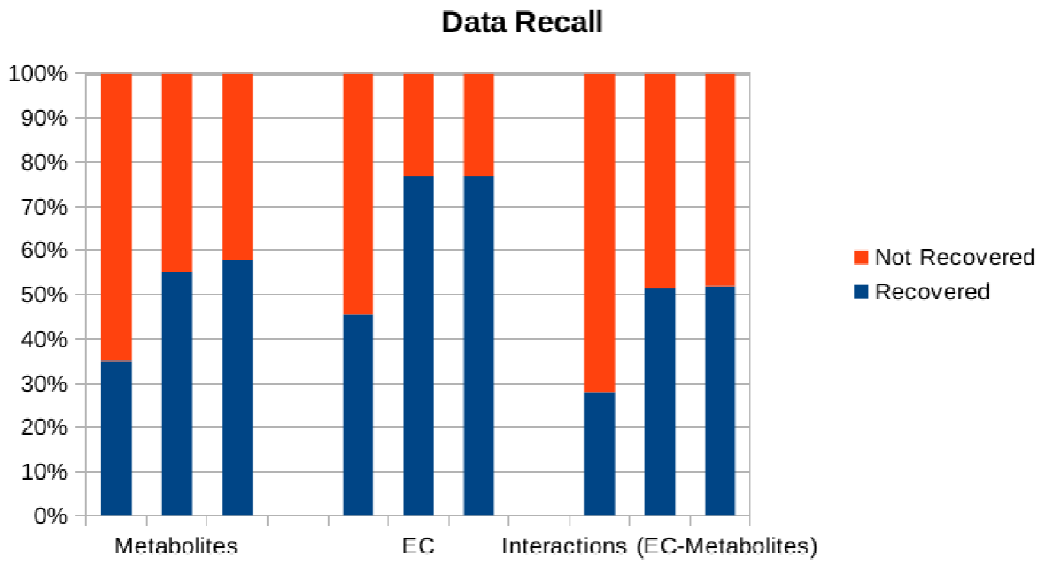


Figure 8: Grafic with the % of data recovery from *Yus et al.* Each column of each type of data (Metabolites, EC, Interactions) represents de data added after each step of the prediction: BioCyc data, Prediction using metabolomes, Predictions using CTLink.

On the other hand, looking on the recall of the data extracted from Yus *et al.*, we can observe how it increases after each prediction step. Enzyme field is where we recover a higher quantity of data, a 78%, followed by metabolites, where we recover 136 out of 225.

There are many interactions that we have not been able to predict. From 286 not recovered interactions, only in 41 of them we were recovering both metabolites and enzymes.

Some of the reasons why we couldn't recovered some interactions are that BioCyc enzyme associated to the reaction is not assigned to any OrthoMCL group (80 interactions), while some other interactions could not be recovered because they are missing in BioCyc (165 interactions). As result, excluding those 165 interactions that are impossible to be recovered, recall % would increase to 73%.

We have delved into these metabolites and enzymes that we are not able to recover through our methodology. The main reasons preventing their recovering are listed subsequently:

- The interaction is reported in BioCyc, however we don't find any ortholog to the protein associated to this interaction because:
    - BioCyc protein is not associated to any orthology group.
- The interaction is not present in BioCyc. In these cases, when the enzyme is reported in BioCyc, it may or not has an ortholog in Mycoplasma genome.
    - Both components are present in BioCyc, but the interaction is not reported.
    - Some of the components is not reported in BioCyc, metabolite or EC.
    - None of the components are present in BioCyc.

Between the interactions not recovered in the prediction, for 8 of them both components (metabolite and EC) are found in BioCyc, and we have a Mycoplasma ortholog for the EC. So, if the interaction between both components had been present in BioCyc we would have been able to predict it. Moreover, there are also 109 interactions for which despite we have found also mycoplasma ortholog of the enzyme, but the metabolite of the interaction is not reported in BioCyc.

On the other hand, there are 42 interactions missing in BioCyc, but where we didn't find neither any mycoplasma ortholog (if the enzyme is reported in BioCyc).

So, concluding, there are 117 interactions which we would have been able to predict if they had been present in BioCyc databases. While, 80 interactions are not recovered because in Mycoplasma genome we haven't found any ortholog.

One example of these not recovered interactions is the interaction between D-Fructose 6-Phosphate and EC-2.7.1.11 (6-phosphofructokinase). This metabolite is reported in all the databases except LeishCyc, whereas the enzyme reaction is found in all them. Despite both components are present in almost every database, their interaction is not reported on anyone. On the other hand, despite of being reported in all the databases, this enzyme reaction is associated to specific proteins in only 4 of them, MetaCyc, LeishCyc, AraCyc and EcoCyc. All these associated proteins where grouped in the OrthoMCL orthology group OG5_126758, where we find *Mycoplasma pneumoniae* protein PFKA_MYCPN. So, if the interaction of D-Fructose 6-Phosphate with 2.7.1.11 had been reported in BioCyc, we would have been able to predict it.

Completing *Mycoplasma pneumoniae* metabolome.

On Mycoplasma metabolome reconstruction we are obtaining a good recall. So, the confidence of our predictions is high enough to assume that a significant percentage of our predictions may be correct.

In total we have predicted 1649 new interactions which have not been reported in Yus *et al.* They could be divided between those interactions of compounds and EC from Yus *et al* (118), and those interactions where some of the components (or both) are predicted (1531).

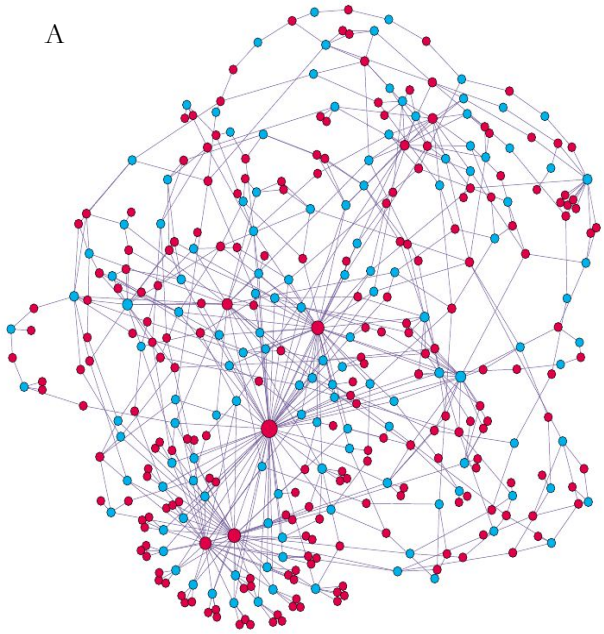
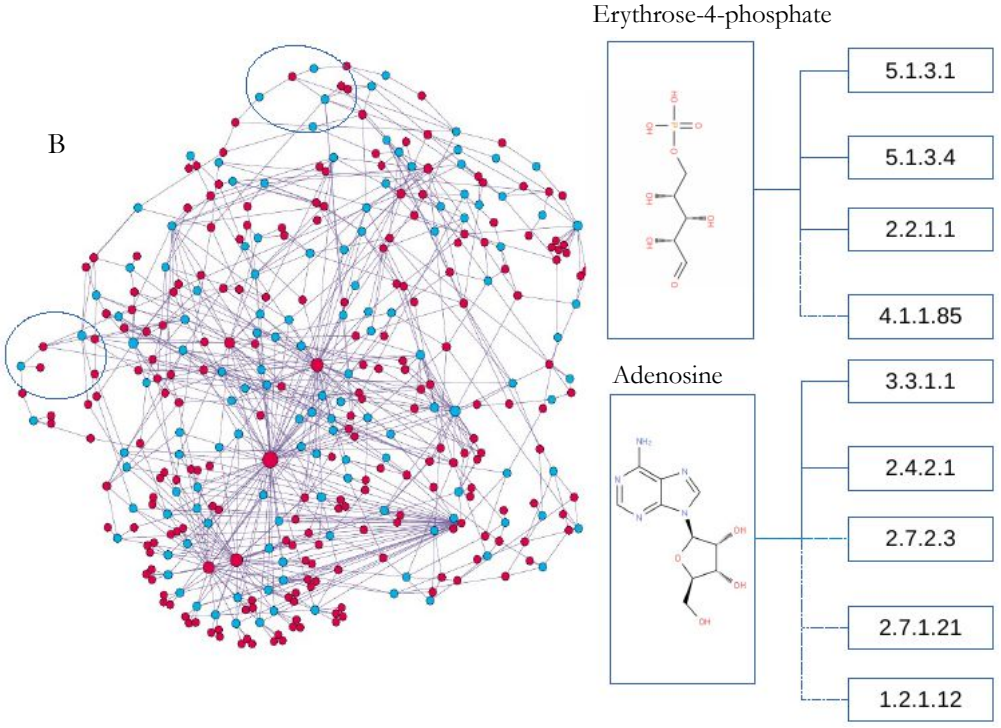


Figure 9:

A) Network created with the data reported in *Yus et al.*

B) Network created with the nodes from *Yus et al* adding the new interactions predicted for these nodes.

In circles there is some of the metabolites for which we are prediction new interactions. These metabolites and their interactions are showd on the right of the graphic. Continuous line are interactions from *Yus et al*, while discontinuous ones are new interactions.



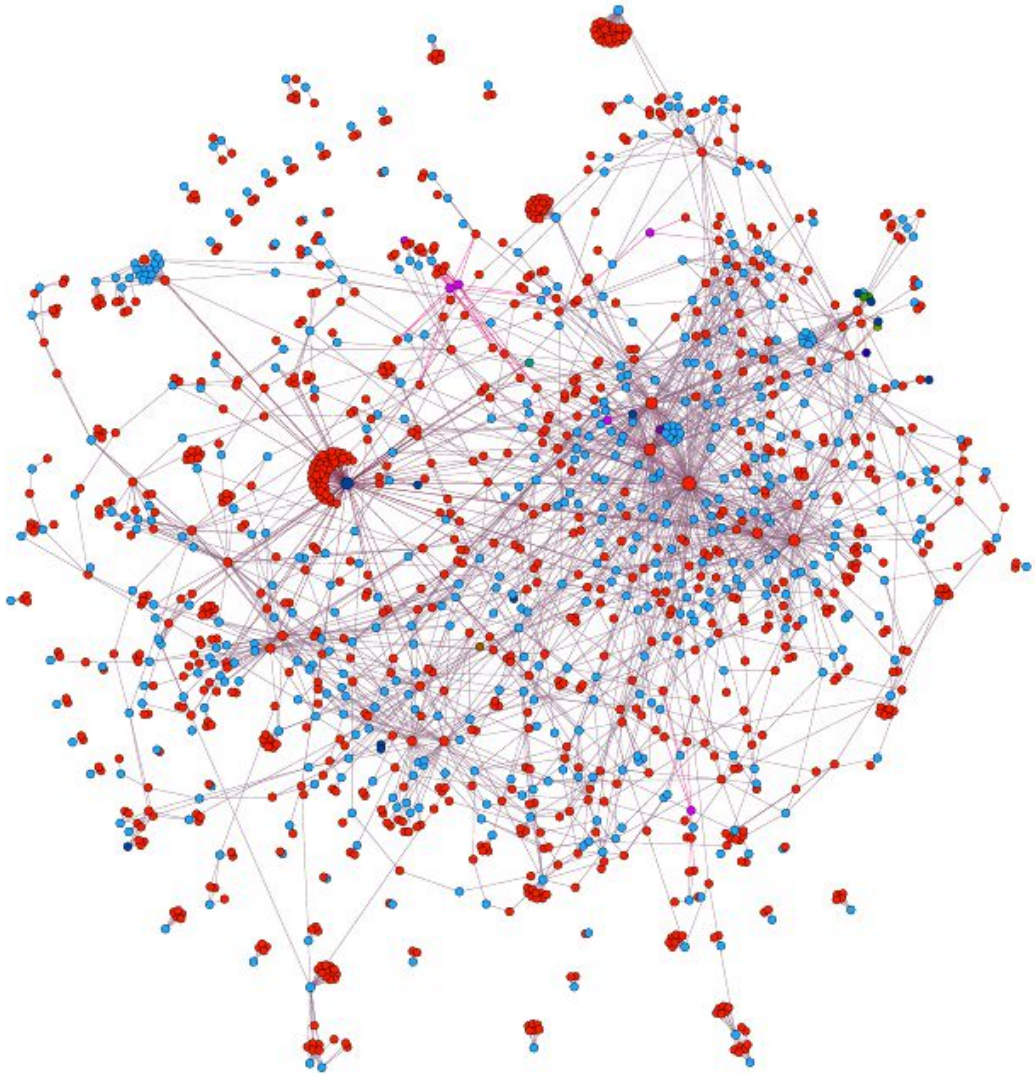


Figure 10: Network created adding all the data predicted to *Yus et al* reported data. In red there is metabolites, in blue EC and in dark blue it is represented those proteins that are not associated to any EC

In figure 9 we can observe the metabolic networks before and after adding predicted interactions. This new interactions are giving insights about the complexity of metabolic networks. Focusing on these 118 new interactions between Yus *et al* metabolites and EC, we are predicting new affinities for 54

metabolites. Pyrophosphate is the compound for which it is predicted a higher number of new interactions, 32. Between the other metabolites we find nucleotides and nucleotide phosphates like Cytidine, Deoxythymidine Triphosphate or Uridine monophosphate; aminoacids like L-Serine, cofactors like NADH, FAD and FMN, and other metabolites such as D-mannose 6-phosphate, 3-phosphoglycerate or d-erythrose 4-phosphate.

After adding predicted interactions, the average degree increase from 3,5 to 4,15, while diameter decrease from 12 to 11, reflecting an increase in the metabolic network connectivity.

Finally, in figure 10 it is represented the complete metabolic network of *Mycoplasma pneumoniae*. This network is composed with metabolic data from Yus *et al*, and data predicted by our framework. We are increasing more than 3 times the experimental data from Yus *et al,* obtaining 926 metabolites, 511 enzymes, and 2278 interactions. Our framework is not only predicting metabolic enzymes associated to EC number but proteins from genetic processing family, environmental information processing family and cellular processes family. In total we predict 24 proteins that are not associated to EC-number, between them we find transporters, cytoskeleton proteins and proteins related to genetic information processing.

Conclusions

Our framework has been able to recover an important part of *Mycoplasma pneumoniae* metabolome reported by Yus *et al*. Despite not having a high recall of the interactions Compound-EC, we recover almost 60% of metabolites and 80% of EC. Many of not recovered data have been impossible to be predicted since it was not available in the metabolic data source, BioCyc.

On the other hand we are increasing the complexity of metabolic network. We are not only adding new metabolites and proteins to the network, but we are

are also adding new interactions between the components previously reported by Yus *et al.*

Despite showing some limitations on the prediction capacity, our methodology have resulted to be useful on the prediction of metabolomes and completing the existing ones. In this case only a few species have been required for this reconstruction because of the simplicity of *Mycoplasma pneumoniae*, but for bigger genomes we should expect that there would be a higher number of organisms used on the metabolome reconstruction.

References

Balish, M.F., 2014. Mycoplasma pneumoniae, an Underutilized Model for Bacterial Cell Biology. J. Bacteriol. 196, 3675–3682. doi:10.1128/JB.01865-14

Baseman, J.B., Tully, J.G., 1997. Mycoplasmas: sophisticated, reemerging, and burdened by their notoriety. Emerg. Infect. Dis. 3, 21–32. doi:10.3201/eid0301.970103

Block, P., Sotriffer, C.A., Dramburg, I., Klebe, G., 2006. AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. Nucleic Acids Res. 34, D522-526. doi:10.1093/nar/gkj039

Caspi, R., Altman, T., Dale, J.M., Dreher, K., Fulcher, C.A., Gilham, F., Kaipa, P., Karthikeyan, A.S., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Paley, S., Popescu, L., Pujar, A., Shearer, A.G., Zhang, P., Karp, P.D., 2010. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 38, D473–D479. doi:10.1093/nar/gkp875

Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Pujar, A., Shearer, A.G., Travers, M., Weerasinghe, D., Zhang, P., Karp, P.D., 2012. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 40, D742-753. doi:10.1093/nar/gkr1014

Cazzaniga, P., Damiani, C., Besozzi, D., Colombo, R., Nobile, M.S., Gaglio, D., Pescini, D., Molinari, S., Mauri, G., Alberghina, L., Vanoni, M., 2014. Computational strategies for a system-level understanding of metabolism. Metabolites 4, 1034–1087. doi:10.3390/metabo4041034

Chaudhry, R., Ghosh, A., Chandolia, A., 2016. Pathogenesis of Mycoplasma pneumoniae: An update. Indian J. Med. Microbiol. 34, 7–16. doi:10.4103/0255-0857.174112

Dybvig, K., 1990. Mycoplasmal Genetics. Annu. Rev. Microbiol. 44, 81–104. doi:10.1146/annurev.mi.44.100190.000501

Edwards, J.S., Palsson, B.O., 1999. Systems properties of the Haemophilus influenzae Rd metabolic genotype. J. Biol. Chem. 274, 17410–17416.

Fischer, S., Brunk, B.P., Chen, F., Gao, X., Harb, O.S., Iodice, J.B., Shanmugam, D., Roos, D.S., Stoeckert, C.J., 2011. Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups, in: Goodsell, D.S. (Ed.), Current Protocols in Bioinformatics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J.P., 2012. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40, D1100–D1107. doi:10.1093/nar/gkr777

Gilson, M.K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., Chong, J., 2016. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 44, D1045-1053. doi:10.1093/nar/gkv1072

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C., Herrmann, R., 1996. Complete Sequence Analysis of the Genome of the Bacterium Mycoplasma Pneumoniae. Nucleic Acids Res. 24, 4420–4449. doi:10.1093/nar/24.22.4420

Kell, D.B., Goodacre, R., 2014. Metabolomics and systems pharmacology: why and how to model the human metabolic network for drug discovery. Drug Discov. Today 19, 171–182. doi:10.1016/j.drudis.2013.07.014

Liu, L., Mei, Q., Yu, Z., Sun, T., Zhang, Z., Chen, M., 2013. An integrative bioinformatics framework for genome-scale multiple level network reconstruction of rice. J. Integr. Bioinforma. 10, 223. doi:10.2390/biecoll-jib-2013-223

Nookaew, I., Olivares-Hernández, R., Bhumiratana, S., Nielsen, J., 2011. Genome-scale metabolic models of Saccharomyces cerevisiae. Methods Mol. Biol. Clifton NJ 759, 445–463. doi:10.1007/978-1-61779-173-4_25

Roth, B.L., Lopez, E., Beischel, S., Westkaemper, R.B., Evans, J.M., 2004. Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. Pharmacol. Ther. 102, 99–110. doi:10.1016/j.pharmthera.2004.03.004

Ryu, J.Y., Kim, H.U., Lee, S.Y., 2015. Reconstruction of genome-scale human metabolic models using omics data. Integr. Biol. Quant. Biosci. Nano Macro 7, 859–868. doi:10.1039/c5ib00002e

Shameer, S., Logan-Klumpler, F.J., Vinson, F., Cottret, L., Merlet, B., Achcar, F., Boshart, M., Berriman, M., Breitling, R., Bringaud, F., Bütikofer, P., Cattanach, A.M., Bannerman-Chukualim, B., Creek, D.J., Crouch, K., de Koning, H.P., Denise, H., Ebikeme, C., Fairlamb, A.H., Ferguson, M.A.J., Ginger, M.L., Hertz-Fowler, C., Kerkhoven, E.J., Mäser, P., Michels, P.A.M., Nayak, A., Nes, D.W., Nolan, D.P., Olsen, C., Silva-Franco, F., Smith, T.K., Taylor, M.C., Tielens, A.G.M., Urbaniak, M.D., van Hellemond, J.J., Vincent, I.M., Wilkinson, S.R., Wyllie, S., Opperdoes, F.R., Barrett, M.P., Jourdan, F., 2015. TrypanoCyc: a community-led biochemical pathways database for Trypanosoma brucei. Nucleic Acids Res. 43, D637-644. doi:10.1093/nar/gku944

Simeonidis, E., Price, N.D., 2015. Genome-scale modeling for metabolic engineering. J. Ind. Microbiol. Biotechnol. 42, 327–338. doi:10.1007/s10295-014-1576-3

Southan, C., Sharman, J.L., Benson, H.E., Faccenda, E., Pawson, A.J., Alexander, S.P.H., Buneman, O.P., Davenport, A.P., McGrath, J.C., Peters, J.A., Spedding, M., Catterall, W.A., Fabbro, D., Davies, J.A., NC-IUPHAR, 2016. The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. Nucleic Acids Res. 44, D1054-1068. doi:10.1093/nar/gkv1037

Waites, K.B., Talkington, D.F., 2004. Mycoplasma pneumoniae and its role as a human pathogen. Clin. Microbiol. Rev. 17, 697–728, table of contents. doi:10.1128/CMR.17.4.697-728.2004

Xiao, L., Ptacek, T., Osborne, J.D., Crabb, D.M., Simmons, W.L., Lefkowitz, E.J., Waites, K.B., Atkinson, T.P., Dybvig, K., 2015. Comparative genome analysis of Mycoplasma pneumoniae. BMC Genomics 16. doi:10.1186/s12864-015-1801-0

Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W.-H., Wodke, J.A.H., Güell, M., Martínez, S., Bourgeois, R., Kühner, S., Raineri, E., Letunic, I., Kalinina, O.V., Rode, M., Herrmann, R., Gutiérrez-Gallego, R., Russell, R.B., Gavin, A.-C., Bork, P., Serrano, L., 2009. Impact of genome reduction on bacterial metabolism and its regulation. Science 326, 1263–1268. doi:10.1126/science.117726

Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, J., 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 34, D668-672. doi:10.1093/nar/gkj067

III.5: Insight on systemic metabolic prediction

Joaquim Olivés, and Jordi Mestres*

Systems Pharmacology, Research Programme on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute and University Pompeu Fabra, Parc de Recerca Biomèdica, 08003 Barcelona, Catalonia, Spain

Abstract

Metabolic networks are consisting of compounds, enzymes and their interactions. They are very in important for several applications in pharmacology. However, its experimental study is expensive and time-consuming, so computational modeling techniques become essential. One of the most used methodologies to model metabolomes is genome scale reconstruction. In this study we propose a methodology based on this strategy for the reconstruction of metabolic networks from genomic data. From experimental data from other organisms we predict their orthologs and construct a network with the reactions catalized. In our methodology we are also applying In silico profiling methodologies to predict the metabolites activity profile. This way we are perfoming metabolic networks with were it is reported the interaction of all the metabolites with enzymes and other proteins. We demonstrated the usefulness of the methodology to provide insight on the possible organisms metabolic networks by applying it on 6 organisms with its metabolic network already reported. The proposed methodology allows the prediction of primary and secondary metabolism which may accelerate metabolic research.

* Corresponding autor. *E-mail address:* jmestres@imim.es

Introduction

The establishment of complete genome sequences allowed the development of genome-scale models for the reconstruction of metabolic models (Edwards and Palsson, 2000). Metabolic models are an important tool in systems biology, they provide snapshots of the global metabolism given genetic and environmental conditions. These models enable the construction of metabolic models for different types of cells, from photosynthetic microorganisms to human cell types (Ryu et al., 2015). Genome-scale metabolic reconstruction use genome sequence for the integration of biochemical metabolic reactions. In general, the more information about biochemistry and genetics is available for the target organism, the better the predictive capacity. The first genome-scale metabolic model was generated in 1999 for *Haemophilus influenzae* (Edwards and Palsson, 1999). While in 2000 it was modeled for first time the metabolic network of the most widely used bacterium, *Escherichia coli* (Edwards and Palsson, 2000). The first model of eukaryotic metabolism was from *Saccharomyces cerevisiae* in 2003 (Förster et al., 2003). And in 2009 it was pubblished for first time a genome scale model for *Arabidopsis thaliana*, used as a model organism in plant biology (Poolman et al., 2009).

Metabolic networks can be viewed as lists of those molecular mechanisms (reactions) and associated molecular components (enzymes, substrates, and products) that are most directly related to the metabolic capabilities (Durot et al., 2009). The knowledge of the physiology of the whole organism has many applications. For instance, initially the reconstruction of metabolic models for microbial organisms allowed to know which metabolites are they able to produce, which chemical nutrients requires or how efficient is it converting chemicals (Durot et al., 2009; Kim et al., 2012). While metabolic models of mammals have been employed to study various human diseases and develop strategies for potential treatments (Lewis et al., 2010).

On the other hand, for drugs emerged the polypharmacology paradigm, consistent on the philosophy of "one drug multiple targets" (Reddy and Zhang, 2013). Polypharmacology refers to the ability of a molecule to interact with multiple proteins. To attempt the prediction of links between the chemical structures of bioactive molecules and which proteins, chemogenomics approaches have emerged in the recent years. These *In silico* approaches have been widely used in drug discovery since they have great potential predicting the pharmacological profile of bioactive compounds and identifying potential targets (Lavecchia and Cerchia, 2016).

Polypharmacology paradigm can also be applied in metabolic networks, where most enzymes accept multiple substrates and possess affinity for a wide range of compounds. This phenomenon can be referred as substrate promiscuity (Piedrafita et al., 2015). The chemical reactivity of metabolites and unspecific enzyme function give rise to a number of side reactions and side products that are not part of canonical pathways. The knowledge of these molecules is important since they may affect metabolic efficiency and play a potential role in diseases (Khersonsky and Tawfik, 2010).

The aim of this work is to present and evaluate prediction capacity of a genome-scale metabolic reconstruction framework using MetaCyc and BioCyc databases (Caspi et al., 2014). To include this substrate promiscuity in our metabolic networks predictions, we have applied *In silico* methodologies in our framework. BioCyc is a collection of 3000 pathway/genome databases, were each database is dedicated to one organism. While MetaCyc is an encyclopedia of experimentally defined metabolic pathways, it contains 2100 metabolic pathways and 114000 metabolic reactions (Caspi et al., 2014). These databases are used as starting point for the reconstruction of some of the most complete metabolic models, *Escherichia coli, Homo sapiens, Saccharomyces cerevisiae, Arabidopsis thaliana, Trypanosoma brucei and Leishmania major.*

Methods

We have used the data available in MetaCyc and BioCyc to predict metabolome of 6 organisms using their genome. MetaCyc is composed of metabolic data reported experimentally; while from BioCyc includes several genomes classified in 3 Tiers according to the curation level. We have used only databases from Tier-1, which are composed by experimental and predicted metabolic data withmore than one year curation. It includes 6 databases, humancyc (*Homo sapiens*), ecocyc (*Escherichia coli K-12 substr. M1655*), yeastcyc (*Saccharomyces cerevisiae*), aracyc (*Arabidopsis thaliana*), trypanocyc (*Trypanosoma brucei*) and leishcyc (*Leishmania major strain Friedlin*).

In this study we have predicted the metabolome of the previously mentioned most studied species in BioCyc. For the prediction of their metabolome we have projected genomic data to metabolic data from other species.

Genomic data have been downloaded mainly from Uniprot and BioCyc database, and it has been classified in orthology groups using the OrthoMCL-DB Version 5. Using the web application of OrthoMCL-DB we map our proteins on OrthoMCL-groups. This tool performs a BLASTP against all the proteins in OrthoMCL-DB using a cut-off of 1e-5 and 50% match. Proteins are assigned to the group containing the best hit. If the best match is not assigned to any group, it is assigned to NO_GROUP (Fischer et al., 2011).

On the other hand, for the target profiling of metabolites we are using CTLink, that use ligand-based approaches and cross pharmacology. This program extract information from several public databases to perform the prediction. The databases included are ChEMBL (Gaulton et al., 2012), DrugBank (Wishart et al., 2006), BindingDB (Gilson et al., 2016), IUPHARdb (Southan et al., 2016), PDSP (Roth et al., 2004), and affinDB (Block et al., 2006). From the activities obtained from CTLink we select those with a confidence score >0.7 and a predicted activity higher than 4.

Results

In general we are recovering more than the half of the metabolites found in BioCyc, obtaining recall values higher than 0,8 in most of the species. *Saccharomyces cerevisiae* is the organism with a higher number of recovered metabolites, 623 out of 638 (0.98), followed by *Escherichia coli, Leishmania major*, and *Homo sapiens*, with recall values of 0.83, 0.819 and 0.803, respectively. On the other hand, for *Trypanosoma brucei* and *Arabidopsis thaliana* we could recovered about 60% of the total number of metabolites. About precision,

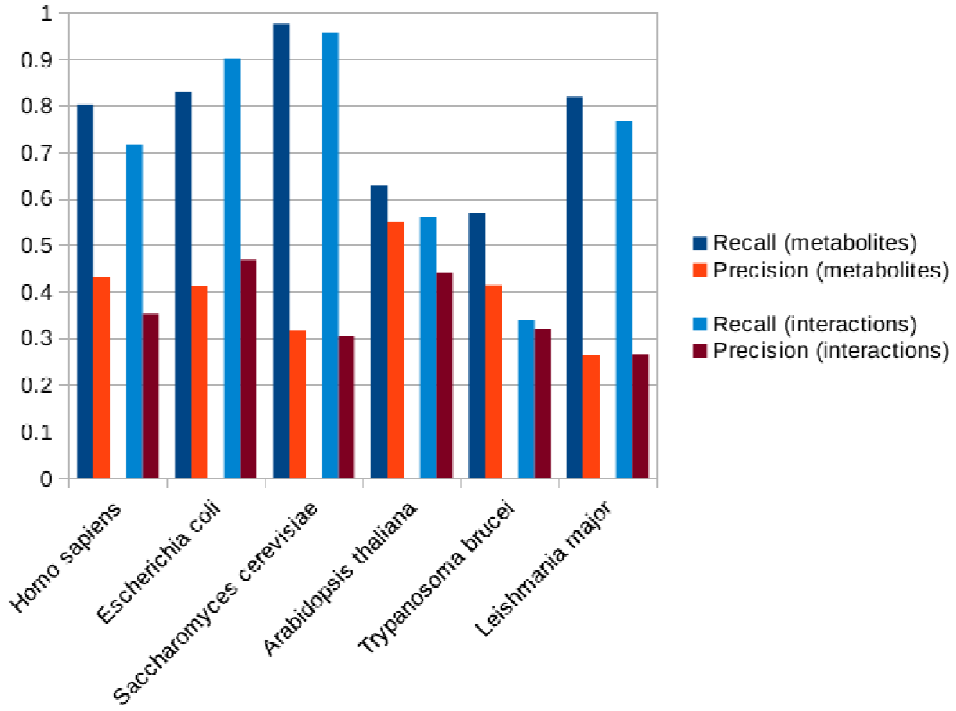


Figure 1: Graphic with the recall and precisions obtained for each predicted organism metabolome. Calculated for metabolites and interactions between EC and metabolites.

most of the values are between 0.26 and 0.44. The highest precision is found in Arabidopsis, 0.55.

Regarding interactions between metabolites and enzymes, recall values doesn't use to be high. However they are still higher than 70% for most of the organisms. *Saccharomyces cerevisiae* is the organism with the highest recall, 0.959, followed by *Escherichia coli, Leishmania major* and *Homo sapiens*, with recall values of 0.902, 0.767 and 0.717, respectively.

Since metabolomes are predicted using metabolic data from several organisms belonging to all kingdoms, we compared the contribution of species from each kingdoms to the reconstruction (figure 2).

For *Escherichia coli* and *Saccharomyces cerevisiae*, most of the recovered data is obtained from organisms belonging to their kingdom, Bacteria and Fungi, respectively. However, the precision of this data is not the highest.

On the other hand, for *Homo sapiens* and Leishmania, where the highest recall values are much lower than those of E.coli and yeast, there is a relatively similar % of recovered data (recall) from most of the kingdoms. However, for the precision values, we can be clearly observe higher values on data obtained from living organisms belonging to the same kingdom of the studied organism, Metazoa for *Homo sapiens*, and Eukaryota for *Leishmania major*. Otherwise, for *Trypanosoma brucei* many of the recovered data is obtained from organisms from Eukaryota kingdom. Moreover, in data predicted from Eukaryota we observe a high precision, it is over 0.7. Finally, for *Arabidopsis thaliana*, there is a similar contribution from most of the kingdoms. Precision values have a high variability, there is no relation with the source organisms.

So, for most of living organisms, the most important contributors to the recovered data are those organisms belonging to the same kingdom. In many cases, those belonging to the same kingdom are not only adding an important amount of data to metabolic information, but also it has a high reliability.

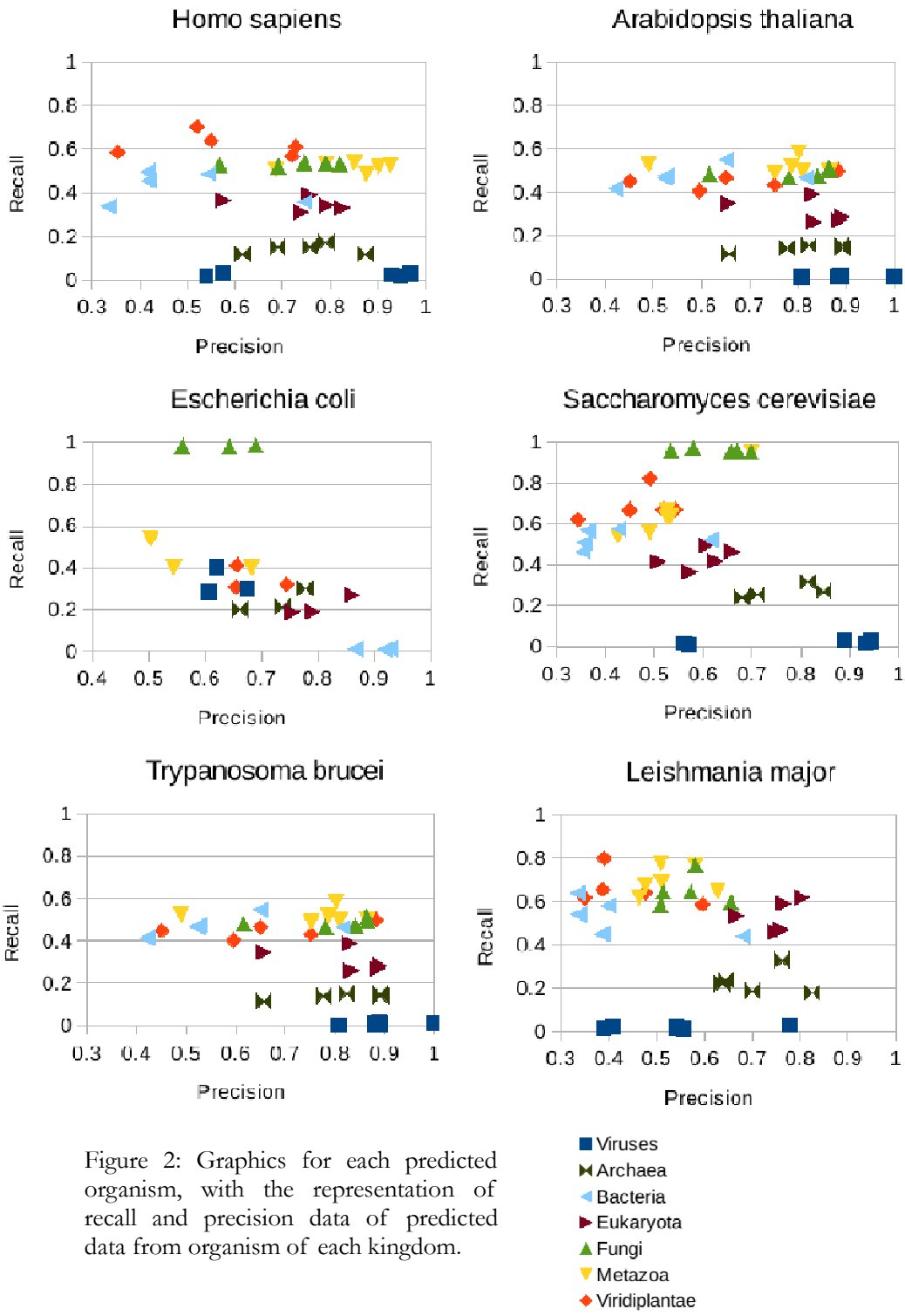


Figure 2: Graphics for each predicted organism, with the representation of recall and precision data of predicted data from organism of each kingdom.

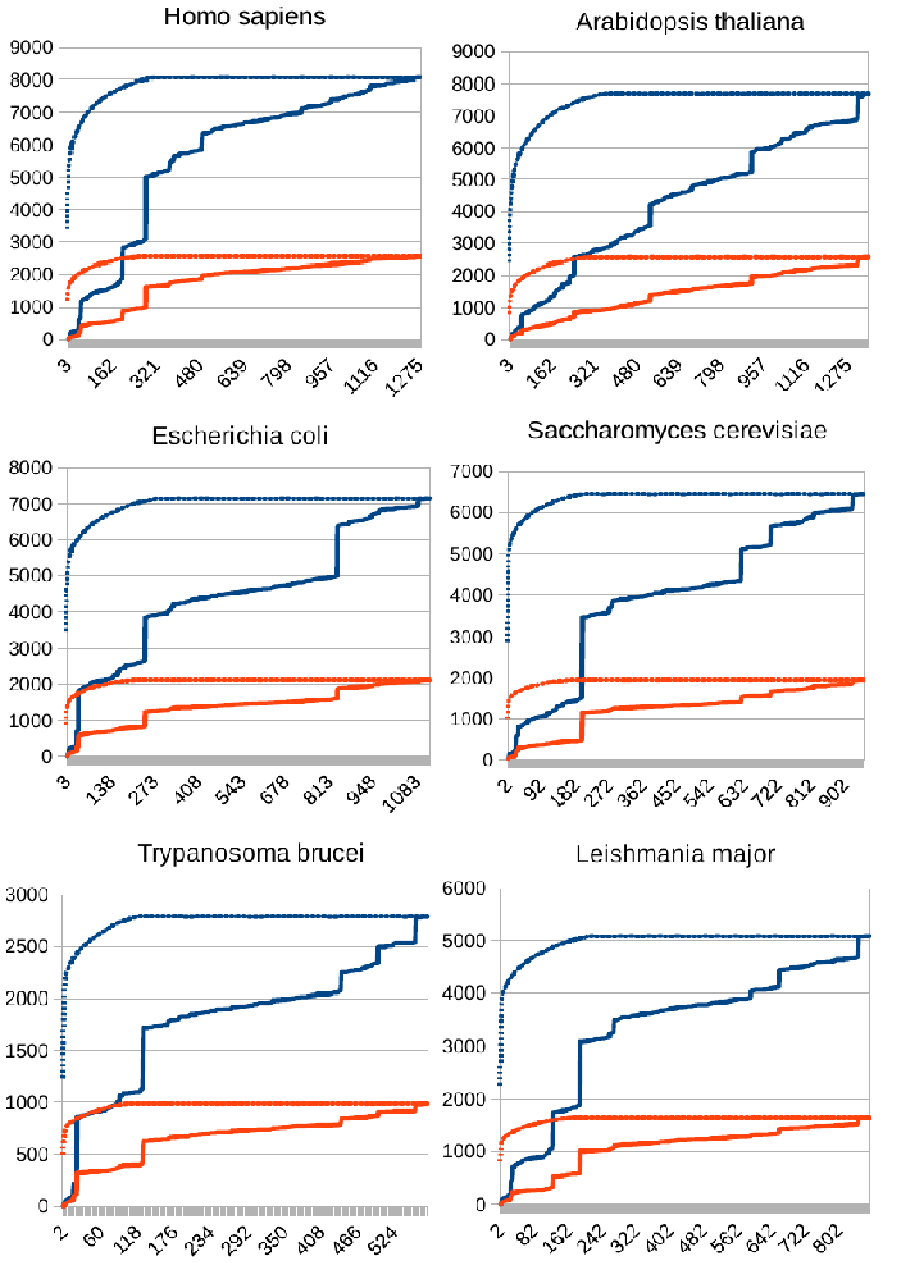


Figure 3: Cumulative graphics with metabolic data obtained from each organism. Red is for metabolites, and blue for interactions EC-metabolite.Pointed line is for cumulative data sorting organisms which are apporting more new data to metabolic data each time, and continue is for unsorted cumulative data.

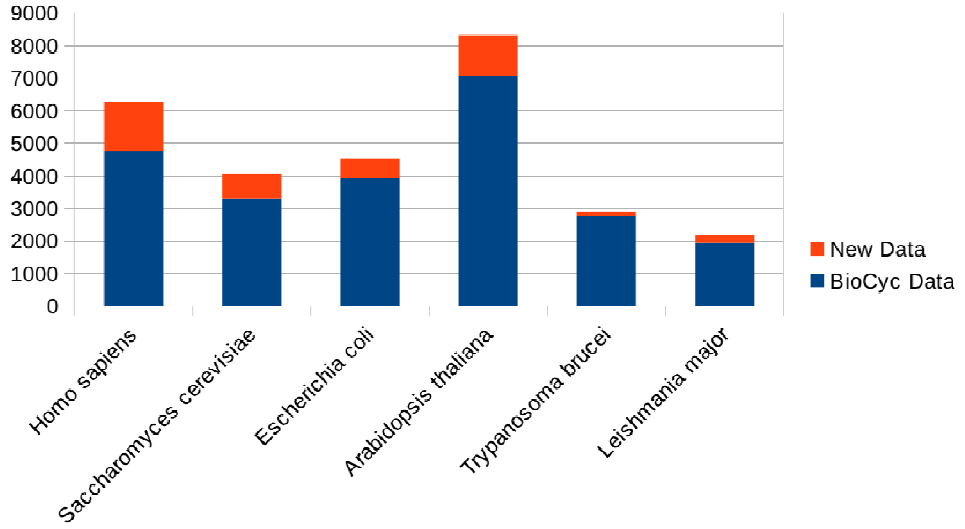


Figure 4: Quantity of new interaction data between metabolites and EC found in BioCyc database. In blue there is the experimental data reported in BioCyc, and in red the new predicted data.

Moreover, we have analysed the individual contribution of each organism to each predicted metabolome (Figure 3), obtaining in all cases a high number of species. But, not all them have the same contribution, as expected, organisms with a more complete metabolome are contributing the most (E. coli, Yeast, Homo sapiens,..).

On the other hand, if we reorder data sources, sorting them according to new data contribution, we can observe the minimum quantity of organisms (and which ones) required to perform each prediction. In all cases, the highest data contributions are from *Arabidopsis thaliana, Homo sapiens, Escherichia coli* and *Saccharomyces cerevisiae*; specially the first 2, whose complexity is higher.

So a higher number of data on the source metabolome may allow a higher number of data for the reconstruction. Taking only the first 2 species we are able to obtain more than 50% of the total data. It is a huge amount of data, but there is still an important quantity that must be obtained from other organisms.

The quantity of required organisms for collecting all the predicted metabolites is 17-22% of the total number of contributing organisms. While interactions metabolites-EC, it is necessary a 20-26%.

In general we have a high recall, while precision value is not high for any specie. We predict many interactions not reported in MetaCyc/BioCyc. This data includes new compounds and enzymes, or new interactions between metabolites and enzymes. These new interactions increase the number nodes and connectivity of metabolic networks. The quantity of new interactions between confirmed enzymes and metabolites are shown in figure 4. Homo sapiens is the organism with a higher number of new interactions between components from the BioCyc/MetaCyc network, 1503, increasing it a 31.5%. It is followed by yeast and Arabidopsis thaliana, with 23.1% and 17.5%, respectively.

Among those predicted interactions on human, we could find experimental evidences of many of them. Some of them are listed in table 1:

Metabolite	EC	Prediction Method*	Predicted Activity	Activity	Reference
17beta-estradiol	1.2.3.1	CTLink (EXPrd, 5.2)	Ligand	Inhibitor	Barr and Jones, 2013
4-hydroxybenzaldehyde	2.6.1.19	CTLink (EXPrd 4.8)	Ligand	Inhibitor	Tao et al., 2009
Acetaldehyde	1.2.1.4	Metabolomes	Product/ Reactant	Substrate	Patel et al., 2008
Acetate	1.2.1.4	Metabolomes	Product	Product	Patel et al., 2008
Alpha-D-glucose-6-phosphate	2.4.1.1	CTLink (EXPrd 4.6)	Ligand	Inhibitor	Ercan-Fang et al., 2005
benzoate	1.4.3.3	CTLink (EXPrd 5.3)	Ligand	Inhibitor	Caldinelli et al., 2010
cholesterol	3.6.3.44	CTLink	Ligand	Substrate	Garrigues et

		(EXPrd 5.1)			al., 2002
D-erythrose-4-phosphate	5.3.1.8	CTLink (EXP 4.0)	Ligand	Inhibitor	Proudfoot et al., 1994
D-gluconate 6-phosphate	5.3.1.9	CTLink (EXPrd 4.4)	Ligand	Inhibitor	Proudfoot et al., 1994
D-glyceraldehyde	1.1.1.21	CTLink (EXP 8.0)	Ligand	Substrate	Rakowitz et al., 2007
dihydrolipoamide	1.8.1.4	Metabolomes	Reactant	Substrate	Liu et al., 1995
geranylgeranyl diphosphate	2.5.1.60	CTLink (EXPrd 9.1)	Ligand	Substrate	Baron and Seabra, 2008
glycolate	1.1.3.15	Metabolomes	Reactant	Substrate	Vignaud et al., 2007
glyoxylate	1.1.3.15	Metabolomes	Product	Product	Vignaud et al., 2007
L-canavanine	3.5.3.1	Metabolomes	Reactant	Inhibitor	Colleluori and Ash, 2001
N-butanal	*1.1.1.21*	*Metabolomes*	*Product*	*Substrate*	Endo et al., 2009
N-butanol	*1.1.1.21*	*Metabolomes*	*Substrate*	*Product*	Endo et al., 2009
Nomega-hydroxy-L-arginine	3.5.3.1	CTLink (EXPrd 5.0)	Ligand	Inhibitor	Colleluori and Ash, 2001
Nomega-hydroxy-L-arginine	1.14.13.39	CTLink (EXPrd 5.3)	Ligand	Product/ Reactant	de Visser and Tan, 2008
Oleoyl-CoA	2.3.1.22	Metabolomes	Reactant	Reactant	Lockwood et al., 2003
2-oleoylglycerol	2.3.1.22	Metabolomes	Reactant	Reactant	Lockwood et al., 2003
4-coumarate	1.1.1.21	CTLink (EXPrd 5.5)	Ligand	Inhibitor	Chethan et al., 2008
phytosphingosine	2.7.1.91	Metabolomes	Reactant	Substrate	Kee et al., 2005
progesterone	3.6.3.44	Metabolomes CTLink (EXP 4.2)	Reactant / Ligand	Substrate	Wan et al., 2006

progesterone	2.3.1.26	CTLink (EXP 4.8)	Ligand	Inhibitor	Simpson and Burkhart, 1980
quinine	1.14.14.1	CTLink (EXP 5.0)	Ligand	Substrate	Diczfalusy et al., 2008
sphingosine	2.7.1.91	CTLink (EXPrd 5.8)	Reactant	Substrate	Meacci et al., 2004
Thymidine	2.7.4.9	CTLink (EXPrd 4.6)	Ligand	Inhibitor	Chen et al., 2001

Table 1: Table with new predicted interactions between human metabolites and enzymes previously reported experimentally in other studies. There is annotated the prediction method and the possible role of metabolite in this reaction, together with the real and confirmed role.
(*)For CTLink interactions there is reported the pActivity value of the original interaction, and is indicated if it predicted (PRD), experimental (EXP), or both (EXPrd)

Between these interactions we find metabolic reactions that are not reported or totally reported in metacyc/biocyc, such as 1.1.3.15, 1.2.1.4 and 1.1.1.21. For them we are collecting substrate and products, as can be observed in table 1.

However, for other reactions illustrated in table 1, we are only showing some of the components (reactant or substrate), because not all them are found in humancyc. In example, for the reaction associated to EC-2.7.1.91, we find that phytosphingosine is the substrate and phytosphingosine-1-phosphate the product, however only the first metabolite is reported in humancyc.

From CTLink we are also able to predict some substrate of many reactions, like D-glyceraldehyde, Nomega-hydroxy-L-arginine and progesterone. For the first of them we find a high activity value for aldehyde reductase (EC-1.1.1.21). D-glyceraldehyde is one of the substrates for this reaction, together with NADPH. We are also predicting the interaction of NADPH with this enzyme in CTLink, but the confidence score value is lower than the cut-off applied. On the other hand, the main product of this reaction is glycerol, but his interaction with

aldose reductase is not predicted by CTLink.

About Nomega-hydroxy-L-arginine, it interacts with Nitric-oxide synthase, EC-1.14.13.39, which mainly catalyzes the production of Nitric oxide (NO) from L-arginine. However, it is also reported that L-arginine can be hydroxylated to N(omega)-hydroxy-L-arginine.

While for progesterone and cholesterol we have found affinity for P-glycoprotein, EC-3.6.3.44. It is an ATP dependant protein of the cell membrane that pumps many xenobiotic substances out of the cell. Despite this transport reaction is found in HumanCyc, it doesn't include any compound that it can be pumped out of the cell.

On the other hand, in many of the new interactions obtained from CTLink, metabolites are acting as inhibitor. This would increase the number of ligands for some enzymes. Also the pathways may have a higher interaction between them, since the inhibitory activity of some metabolites produced may affect other pathway dynamics. We have found reported information about the inhibitory activity of 12 metabolites. One of them is D-erythrose-4-phosphate, synthetized in pentose phosphate cycle by transaldolase (EC-2.2.1.2). For this metabolite it has been reported inhibitory activity against Mannose-6-phosphate isomerase (EC-5.3.1.8). So the accumulation of D-erythrose-4-phosphate could be altering the dynamics of pathways that contains this enzyme, like D-mannose degradation and GDP-mannose biosynthesis.

Conclusions

We have presented a metabolic network reconstruction genome-based framework, integrating also an *in silico* profiling methodology. This approach used protein sequences to project them into experimental and curated metabolic data. Framework is able to predict metabolites of an organism and the reactions where they are involved, as substrate or product. Moreover we are

predicting inhibitory activity with enzymatic proteins of the system.

Our method is also able to recover almost all the data available in BioCyc for many of the studied organisms. Despite we obtain a low accuracy when we are comparing the predicted data against BioCyc data, we find potential reactions among predited metabolites. Some of the predicted new interactions between metabolites and enzymes have been already reported in literature.

Despite of the framework is not recovering the full metabolome; it is capable to recover a high amount of the original data from BioCyc databases. So it can help on the creation of a preliminar metabolome which may support on its experimental elucidation. Since this method is highly dependent of the experimental data available, experimental research and computational frameworks can complement each other for a better and faster metabolomic networks construction.

Moreover, the use of *in silico* methodologies is giving new insights about the interaction between different pathways, and how the products originated in a pathway may be altering the flux of other pathways.

References

Baron, R.A., Seabra, M.C., 2008. Rab geranylgeranylation occurs preferentially via the pre-formed REP-RGGT complex and is regulated by geranylgeranyl pyrophosphate. Biochem. J. 415, 67–75. doi:10.1042/BJ20080662

Barr, J.T., Jones, J.P., 2013. Evidence for Substrate-Dependent Inhibition Profiles for Human Liver Aldehyde Oxidase. Drug Metab. Dispos. 41, 24–29. doi:10.1124/dmd.112.048546

Block, P., Sotriffer, C.A., Dramburg, I., Klebe, G., 2006. AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. Nucleic Acids Res. 34, D522-526. doi:10.1093/nar/gkj039

Caldinelli, L., Molla, G., Bracci, L., Lelli, B., Pileri, S., Cappelletti, P., Sacchi, S., Pollegioni, L., 2010. Effect of ligand binding on human D-amino acid

oxidase: implications for the development of new drugs for schizophrenia treatment. Protein Sci. Publ. Protein Soc. 19, 1500–1512. doi:10.1002/pro.429

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Weerasinghe, D., Zhang, P., Karp, P.D., 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 42, D459–D471. doi:10.1093/nar/gkt1103

Chen, J.-J., Wang, P.S., Chien, E.-J., Wang, S.-W., 2001. Direct inhibitory effect of digitalis on progesterone release from rat granulosa cells. Br. J. Pharmacol. 132, 1761–1768. doi:10.1038/sj.bjp.0703974

Chethan, S., Dharmesh, S.M., Malleshi, N.G., 2008. Inhibition of aldose reductase from cataracted eye lenses by finger millet (Eleusine coracana) polyphenols. Bioorg. Med. Chem. 16, 10085–10090. doi:10.1016/j.bmc.2008.10.003

Colleluori, D.M., Ash, D.E., 2001. Classical and slow-binding inhibitors of human type II arginase. Biochemistry (Mosc.) 40, 9356–9362.

de Visser, S.P., Tan, L.S., 2008. Is the bound substrate in nitric oxide synthase protonated or neutral and what is the active oxidant that performs substrate hydroxylation? J. Am. Chem. Soc. 130, 12961–12974. doi:10.1021/ja8010995

Diczfalusy, U., Miura, J., Roh, H.-K., Mirghani, R.A., Sayi, J., Larsson, H., Bodin, K.G., Allqvist, A., Jande, M., Kim, J.-W., Aklillu, E., Gustafsson, L.L., Bertilsson, L., 2008. 4Beta-hydroxycholesterol is a new endogenous CYP3A marker: relationship to CYP3A5 genotype, quinine 3-hydroxylation and sex in Koreans, Swedes and Tanzanians. Pharmacogenet. Genomics 18, 201–208. doi:10.1097/FPC.0b013e3282f50ee9

Durot, M., Bourguignon, P.-Y., Schachter, V., 2009. Genome-scale models of bacterial metabolism: reconstruction and applications. Fems Microbiol. Rev. 33, 164–190. doi:10.1111/j.1574-6976.2008.00146.x

Edwards, J.S., Palsson, B.O., 2000. The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. Proc. Natl. Acad. Sci. U. S. A. 97, 5528–5533.

Edwards, J.S., Palsson, B.O., 1999. Systems properties of the Haemophilus influenzae Rd metabolic genotype. J. Biol. Chem. 274, 17410–17416.

Endo, S., Matsunaga, T., Mamiya, H., Ohta, C., Soda, M., Kitade, Y., Tajima,

K., Zhao, H.-T., El-Kabbani, O., Hara, A., 2009. Kinetic studies of AKR1B10, human aldose reductase-like protein: endogenous substrates and inhibition by steroids. Arch. Biochem. Biophys. 487, 1–9. doi:10.1016/j.abb.2009.05.009

Ercan-Fang, N., Taylor, M.R., Treadway, J.L., Levy, C.B., Genereux, P.E., Gibbs, E.M., Rath, V.L., Kwon, Y., Gannon, M.C., Nuttall, F.Q., 2005. Endogenous effectors of human liver glycogen phosphorylase modulate effects of indole-site inhibitors. Am. J. Physiol. Endocrinol. Metab. 289, E366-372. doi:10.1152/ajpendo.00264.2004

Fischer, S., Brunk, B.P., Chen, F., Gao, X., Harb, O.S., Iodice, J.B., Shanmugam, D., Roos, D.S., Stoeckert, C.J., 2011. Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups, in: Goodsell, D.S. (Ed.), Current Protocols in Bioinformatics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Förster, J., Famili, I., Fu, P., Palsson, B.Ø., Nielsen, J., 2003. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res. 13, 244–253. doi:10.1101/gr.234503

Garrigues, A., Escargueil, A.E., Orlowski, S., 2002. The multidrug transporter, P-glycoprotein, actively mediates cholesterol redistribution in the cell membrane. Proc. Natl. Acad. Sci. U. S. A. 99, 10347–10352. doi:10.1073/pnas.162366399

Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J.P., 2012. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40, D1100–D1107. doi:10.1093/nar/gkr777

Gilson, M.K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., Chong, J., 2016. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 44, D1045-1053. doi:10.1093/nar/gkv1072

Kee, T.H., Vit, P., Melendez, A.J., 2005. Sphingosine kinase signalling in immune cells. Clin. Exp. Pharmacol. Physiol. 32, 153–161. doi:10.1111/j.1440-1681.2005.04166.x

Khersonsky, O., Tawfik, D.S., 2010. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu. Rev. Biochem. 79, 471–505. doi:10.1146/annurev-biochem-030409-143718

Kim, T.Y., Sohn, S.B., Kim, Y.B., Kim, W.J., Lee, S.Y., 2012. Recent advances in reconstruction and applications of genome-scale metabolic models. Curr. Opin. Biotechnol. 23, 617–623. doi:10.1016/j.copbio.2011.10.007

Lavecchia, A., Cerchia, C., 2016. In silico methods to address polypharmacology: current status, applications and future perspectives. Drug Discov. Today 21, 288–298. doi:10.1016/j.drudis.2015.12.007

Lewis, N.E., Schramm, G., Bordbar, A., Schellenberger, J., Andersen, M.P., Cheng, J.K., Patel, N., Yee, A., Lewis, R.A., Eils, R., König, R., Palsson, B.Ø., 2010. Large-scale in silico modeling of metabolic interactions between cell types in the human brain. Nat. Biotechnol. 28, 1279–1285. doi:10.1038/nbt.1711

Liu, T.C., Korotchkina, L.G., Hyatt, S.L., Vettakkorumakankav, N.N., Patel, M.S., 1995. Spectroscopic studies of the characterization of recombinant human dihydrolipoamide dehydrogenase and its site-directed mutants. J. Biol. Chem. 270, 15545–15550.

Lockwood, J.F., Cao, J., Burn, P., Shi, Y., 2003. Human intestinal monoacylglycerol acyltransferase: differential features in tissue expression and activity. Am. J. Physiol. Endocrinol. Metab. 285, E927-937. doi:10.1152/ajpendo.00179.2003

Meacci, E., Cencetti, F., Donati, C., Nuti, F., Becciolini, L., Bruni, P., 2004. Sphingosine kinase activity is required for sphingosine-mediated phospholipase D activation in C2C12 myoblasts. Biochem. J. 381, 655–663. doi:10.1042/BJ20031636

Patel, M., Lu, L., Zander, D.S., Sreerama, L., Coco, D., Moreb, J.S., 2008. ALDH1A1 and ALDH3A1 expression in lung cancers: correlation with histologic type and potential precursors. Lung Cancer Amst. Neth. 59, 340–349. doi:10.1016/j.lungcan.2007.08.033

Piedrafita, G., Keller, M.A., Ralser, M., 2015. The Impact of Non-Enzymatic Reactions and Enzyme Promiscuity on Cellular Metabolism during (Oxidative) Stress Conditions. Biomolecules 5, 2101–2122. doi:10.3390/biom5032101

Poolman, M.G., Miguet, L., Sweetlove, L.J., Fell, D.A., 2009. A Genome-Scale Metabolic Model of Arabidopsis and Some of Its Properties. Plant Physiol. 151, 1570–1581. doi:10.1104/pp.109.141267

Proudfoot, A.E., Payton, M.A., Wells, T.N., 1994. Purification and characterization of fungal and mammalian phosphomannose isomerases. J. Protein Chem. 13, 619–627.

Rakowitz, D., Piccolruaz, G., Pirker, C., Matuszczak, B., 2007. Novel aldose reductase inhibitors derived from 6-[[(diphenylmethylene)amino]oxy]hexanoic acid. Arch. Pharm. (Weinheim) 340, 202–208. doi:10.1002/ardp.200600166

Reddy, A.S., Zhang, S., 2013. Polypharmacology: drug discovery for the future.

Expert Rev. Clin. Pharmacol. 6, 41–47. doi:10.1586/ecp.12.74

Roth, B.L., Lopez, E., Beischel, S., Westkaemper, R.B., Evans, J.M., 2004. Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. Pharmacol. Ther. 102, 99–110. doi:10.1016/j.pharmthera.2004.03.004

Ryu, J.Y., Kim, H.U., Lee, S.Y., 2015. Reconstruction of genome-scale human metabolic models using omics data. Integr. Biol. Quant. Biosci. Nano Macro 7, 859–868. doi:10.1039/c5ib00002e

Simpson, E.R., Burkhart, M.F., 1980. Acyl CoA:cholesterol acyl transferase activity in human placental microsomes: inhibition by progesterone. Arch. Biochem. Biophys. 200, 79–85.

Southan, C., Sharman, J.L., Benson, H.E., Faccenda, E., Pawson, A.J., Alexander, S.P.H., Buneman, O.P., Davenport, A.P., McGrath, J.C., Peters, J.A., Spedding, M., Catterall, W.A., Fabbro, D., Davies, J.A., NC-IUPHAR, 2016. The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. Nucleic Acids Res. 44, D1054-1068. doi:10.1093/nar/gkv1037

Tao, Y.-H., Xu, H.-B., Yang, X.-L., 2009. Inactivation of GABA transaminase by 3-chloro-1-(4-hydroxyphenyl)propan-1-one. Bioorg. Med. Chem. Lett. 19, 731–734. doi:10.1016/j.bmcl.2008.12.033

Vignaud, C., Pietrancosta, N., Williams, E.L., Rumsby, G., Lederer, F., 2007. Purification and characterization of recombinant human liver glycolate oxidase. Arch. Biochem. Biophys. 465, 410–416. doi:10.1016/j.abb.2007.06.021

Wan, C.-K., Zhu, G.-Y., Shen, X.-L., Chattopadhyay, A., Dey, S., Fong, W.-F., 2006. Gomisin A alters substrate interaction and reverses P-glycoprotein-mediated multidrug resistance in HepG2-DR cells. Biochem. Pharmacol. 72, 824–837. doi:10.1016/j.bcp.2006.06.036

Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, J., 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 34, D668-672. doi:10.1093/nar/gkj067

III.6: Human Metabolome Interference to Drug Polypharmacology

Cristina Leal, Irene Sanz, Joaquim Olivés, and Jordi Mestres*

Systems Pharmacology, Research Programme on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute and University Pompeu Fabra, Parc de Recerca Biomèdica, 08003 Barcelona, Catalonia, Spain

Abstract

Motivation: Treatment failure and undesirable drug-related adverse events remain to be a current issue nowadays. Therefore, gaining a better understanding of how drugs and singular pathophysiological conditions are associated at a systems level is of utmost importance. This study aims at establishing the interference that endogenous metabolites offer to the mechanism of action of drugs.

Results: Endogenous metabolites reflect each individual physiological conditions and thus they offer opportunities for precision medicine from a chemical point. In this work, we demonstrate that metabolite competition with drugs for pharmacological annotated targets is a valuable unexploited asset for better understanding drug efficacy and safety, as well as for the identification of specific metabolite markers for each anatomical and therapeutic drug type. The set of anti-Parkinsonian drugs is taken as an illustrative example.

* Corresponding autor. *E-mail address:* jmestres@imim.es

Introduction

One of the major challenges in drug discovery today is to provide more effective therapies to individual patients by maximizing drug efficacy and minimizing drug-related adverse effects for a given individual. Reasons for the success or failure of any therapeutic treatment are many fold and the ultimate outcome depends very much on the subject's pathophenotype dictating the likely outcome (i.e., age, gender, stress, disease), as well as environmental factors (i.e., diet, lifestyle, exposure to pollution) (Wilson, 2009).

The novel paradigm shift and the recent application of "one drug multiple targets" in the sector of drug discovery is referred to as polypharmacology and is of current interest for human health due to its potentially wide implications, going from anticipation of drug side effects to identification of drug repurposing opportunities (Jalencas and Mestres, 2012; Mestres *et al.*, 2009; O'Hagan and Kell, 2015). Since then, several approaches have already been undertaken in order to investigate the molecular basis and complexity of drug actions. For instance, *drug* – *disease* interactions (Zhao and Li, 2012), *drug* – *disease* – *gene* correlations (Sun, 2015), and *drug* – *drug* interaction (Ai *et al.*, 2015) at network based levels. Advances in all those areas are likely to be of paramount importance for enabling progress in the two aspects outlined above, namely, the early prediction of adverse drug reactions and the discovery of novel uses for existing drugs.

On the other hand, the recent and rapidly evolving discipline of metabolomics offers great potential to contribute significantly to biomedical research in general, as well as to the drug discovery process in particular. Metabolite behavior closely reflects the actual cellular environment better than gene and protein expression do (Nicholson *et al.*, 2012), such that much of the activity that happens in the cell occurs at the metabolite level: cell signaling, energy transfer, cell-to-cell communication, while genomics and proteomics set the

stage for what happens next in the cell (Perspectives, 2004). Almost a decade ago, metabolic profiling was already suggested to be an essential component, together with other more exploited 'omics', to permit a more personalized medicine leading to a better understanding of drug efficacy and toxicity (Schnackenberg and Beger, 2007). For example, the comparison of the metabolite levels between diseased and non-diseased individuals, has already been used as disease biomarkers (Shah *et al.*, 2015). Likewise, metabolic profiling between good-responders and poor-responders under a specific therapy could also be used as treatment *chemomarkers*.

The large-scale analysis presented in this work introduces a new computational framework to investigate the interference of the endogenous human metabolomes in the action of drugs based on current experimentally confirmed target profiles of chemical entities and explores its potential use in precision medicine for the identification of treatment biomarkers.

Material and Methods

Metabolite, drug and protein databases

Structures for metabolite and drug molecules were retrieved and processed respectively from the Human Metabolome Database (HMDB) version 3.6 (www.hmdb.ca, downloaded on October 2015) (Wishart *et al.*, 2013) and the DrugBank database version 4.3 (www.drugbank.ca, downloaded on October 2015) (Knox *et al.*, 2011), both in XML and SDF format.

The python module Pybel was used to generate the corresponding InChIkeys of the different chemical structures contained in the downloaded SDF files through the OpenBabel API v2.3.0 (O'Boyle *et al.*, 2008). In such way, InChIKeys were used to map both, metabolites and drugs.

With respect to metabolites, only organic and endogenous metabolites were considered (all inorganic molecules were discarded). In this regard, a good portion of metabolites were found to be wrongly annotated as endogenous compounds by HMDB. Accordingly, we cross-checked the contents of HMDB with the endogeneity criteria applied in the HumanCyc database version 18.5, (www.humancyc.org, downloaded on December 2014) (Romero *et al.*, 2004), followed by a final manual curation of all remaining molecules.

As regards to drugs, only the FDA-approved drugs present in the Anatomical Therapeutic Chemical (ATC) classification system were taken into account. The final set of drugs was filtered from all molecules presenting a structural collision with any metabolite structure (i.e., Levodopa, Mannitol). This was determined by comparing the first part of the InChIKey, which is 14 characters long and is based on the connectivity layer in the InChI string. Drug molecules were stratified according to the ATC classification system (WHO, 2015). Under this classification, a drug may actually belong to more than one group depending on the organ or system they are intended to act and/or their therapeutic and chemical characteristics. See Supplementary Table 1 for the complete list of the anatomical therapeutic chemical classification system drug classes analyzed. A total of 9 drug groups of the ATC anatomical-level were considered: (A) Alimentary tract and metabolism system (n=70), (C) Cardiovascular system (n=121), (D) Dermatological drugs (n=62), (G) Genitourinary system and reproductive hormones (n=56), (L) Antineoplastic and immunomodulating agents (n=69), (M) Musculoskeletal system (n=33), (N) Nervous system (n=163), (R) Respiratory system (n=81), and (S) Sensory organs (n=61). Drugs exclusively belonging to any of the (B) Blood and blood forming organs (n=16), (H) Systemic hormonal preparations – excluding reproductive hormones and insulins (n=10), (P) Antiparasitic products (n=14), (V) Various ATC structures (n=7), and (J) Antiinfectives for systemic use (n=34) were excluded from the study because of an insufficient number of drugs ($n < 20$)

(B-V), and because its pharmacological nature does not aim to be primarily directed towards endogenous human targets and we are just considering human proteins at the moment (P and J). Therapeutic groups corresponding to the second level of the ATC system classification of the 9 drug classes aforementioned were also inspected.

Experimental affinity values (pKi, pKd, pIC50, or pEC50) for *metabolite – target* and *drug – target* interaction data were collected from public domain databases, namely, ChEMBL v19 (Gaulton *et al.*, 2012), PubChem (imported from ChEMBL v19) (Bolton *et al.*, 2008), IUPHAR-DB (downloaded on June 2014) (Sharman *et al.*, 2011), BindingDB (downloaded on September 2014) (Liu *et al.*, 2007).

As the study is limited to the human metabolome, we focused our analysis on human proteins. Moreover, protein subunits were collapsed into unique consensus UniprotKB to avoid protein redundancies. This made a final set of 803 unique protein entities.

Finally, only those *metabolite – protein* and *drug – protein* interactions with experimental affinity better than 1μM (pAct ≥ 6; pAct = pKi, pKd, pIC50, pEC50) were considered. Under this activity threshold, a total of 633 *metabolite – protein* and 4,227 *drug – protein* interactions involving 194 metabolites and 604 drugs, respectively, were finally used in this study. For any given molecule (metabolite or drug), the interaction having the highest pAct value, as well as all additional interactions within 1 log unit of it, were labelled as primary/on-target interactions; all other interactions were treated as secondary/off-target interactions.

Analysis and visualization of *metabolite – protein* and *drug – protein* interaction networks

Networks were constructed to visually illustrate the interaction between drugs, metabolites, and proteins. Interaction networks were visualized using the Gephi

open source package version 0.9.0 (www.gephi.org) (Bastian *et al.*, 2009). Both systems were exemplified as bipartite undirected networks, where nodes stood for the elements of the different sets (drugs, metabolites, and proteins) and edges connecting two nodes from different sets having an interaction affinity value above a certain threshold (in our case, pAct ≥ 6). Nodes were positioned using the Force Atlas 2 algorithm (Jacomy *et al.*, 2011, 2014). Force Atlas 2 is a simple and very fast force-directed algorithm. It uses classic force-vectors, providing a generic and intuitive way to spatialize networks to allow a visual interpretation of their structure, turning structural proximities into visual proximities. Accordingly, highly connected nodes present higher attractive forces and are thus positioned at the center, whereas weaker nodes with lesser interactions are placed on the periphery.

Node size reflects degree centrality (min size = 8 – max sixe = 45), which measures the number of ties a node has to other nodes. Thereby, large nodes are related to wide promiscuities, whereas smaller ones are associated to higher selectivity. Nodes were also colored with respect to the different chemical entities (blue = drugs, yellow = metabolites and red = proteins). Finally, edges were colored according to the primary/on-target (black) and secondary/off-target (grey) label of the interactions.

Construction of the *drug – metabolite* interference networks

The original *metabolite – protein* and *drug – protein* interaction networks described above were connected through the common protein nodes to obtain a new *drug – metabolite* interference network. Given the set of drugs and metabolites for which the interaction for proteins is experimentally known, we consider that a metabolite interferes with the interaction of a drug for a protein if the pAct value of the *metabolite – protein* interaction is higher than the pAct value of the *drug – protein* interaction. These interferences are the edges of the *drug – metabolite* interference network.

Impact of metabolite interference to *drug – target* interaction networks

Edges and nodes of the *drug – target* interaction network were filtered based on the previously described *drug – metabolite* interference network.

The impact of metabolic interference to a *drug – target* interaction network is calculated as the relative number of *drug – target* interactions being interfered by at least one metabolite among all the original *drug – target* interactions in the network. This is quantitatively reflected by what we refer to as the relative metabolic interference (RMI). Using the ATC as the reference framework, the RMI values were evaluated at the different Anatomical and Therapeutic levels under study, and within each network, split into on-target and off-target interferences.

Identification of metabolite markers within drug classes

We used the statistical Fisher's exact test and the two-sided P-values associated to them in order to assess whether metabolites presented any level of interference-specificity towards the different drug classes under study. An alpha value of 0.05 was set to establish the significance of P-values.

Then, be *MI*, the set of metabolites participating in drug interferences, being $\{MI_1, \ldots, MI_n\}$ the different metabolites (n=116), and *DI* the set of interfered drug interactions, with $\{DI_1, \ldots, DI_n\}$ the subset of different anatomical/therapeutic drug interactions subset (n=9/n=50), the contingency table for the metabolite MI_1 at the anatomical subset DI_1 is defined as:

DI_1 interfered by MI_1	DI-$\{DI_1\}$ interfered by MI_1
DI_1 interfered by MI-$\{MI_1\}$	DI-$\{DI_1\}$ interfered by MI-$\{MI_1\}$

The fact that one drug may actually belong to more than one group did not exclude them from the DI-$\{DI_1\}$ set.

Specificity was represented in a heatmap using $-\log_{10}$(P-values). Significant metabolites (P-value ≤ 0.05) for each given anatomical or therapeutic drug class were colored using a yellow-to-black range colors. Non-colored metabolites have non-significant P-values for specificity.

In addition, Pearson correlation coefficients were used to evaluate the hierarchical relationship between metabolites and the different drug classes. This was represented with two dendrograms along the metabolite and a drug type dimensions. This analysis was carried out in the R platform v3.2.1. (www.r-project.org) using 'RColorBrewer', 'pheatmap', and 'gplots' packages.

Results and Discussion

Common protein spaces between drugs and metabolites

We rely on the hypothesis that common target profile spaces between drugs and metabolites might be an important source of poor drug efficacy and undesirable adverse events. Endogenous metabolites can affect directly the pharmacodynamic and pharmacokinetic properties of drugs by the means of binding affinities and thus, become real competitors for a common protein.We start with 194 manually curated organic and endogenous metabolite structures from the Human Metabolome Database (Wishart *et al.*, 2013) and cross-check them with HumanCyc database (Romero *et al.*, 2004) to ensure its intrinsic endogeneity. This subset of endogenous metabolites represents a tiny little portion of what human metabolome is believed. Estimate approaches have reported ranges from 3,000 essential metabolites to approximates of 20,000 unidentified metabolites that are not essential for growth and development but could be of significant importance for prognosis, diagnosis and for the identification of surrogate markers for different disease conditions as well as for a better understanding of applied translational systems biology (Kouskoumvekaki and Panagiotou, 2011; Perspectives, 2004). However, as

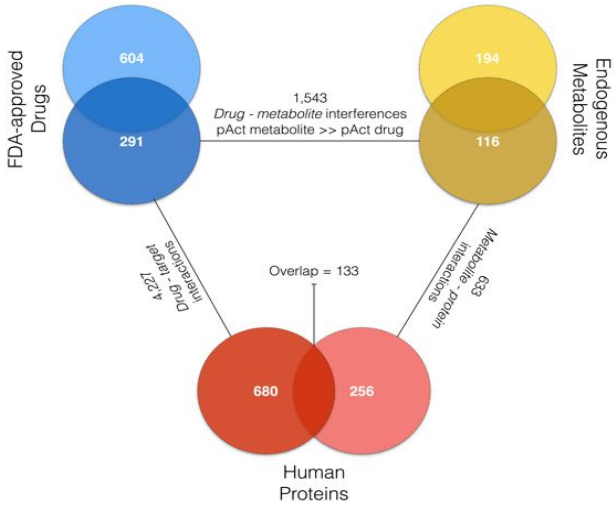


Figure 1. Illustration of our different network models. The drug – target and the metabolite – protein networks are projected into a drug – metabolite network when metabolite's activity values are greater than the drug ones towards a common protein.

above mentioned, metabolomics is still a young field and the real number of different human metabolites is to date unknown since a huge part of them have still not been identified and quantified and neither target-profiled. So far, current HMDB statistics report a total of 2,721 detected and quantified endogenous metabolites. This number barely represents the 9% of a total of 29,284 endogenous metabolites (http://www.hmdb.ca/statistics, downloaded on June 2016). However, not all these identified and quantified compounds have been experimentally target-profiled, hence reducing down to ~1% the portion of the human metabolome we have knowledge of for this study.

Likewise, 604 FDA-approved drug structures are gathered from DrugBank database (Wishart *et al.*, 2006, 2008) and ensured not to be metabolite-like (see Material and Methods section).

In total we globally collect 4,227 *drug* – *target* interactions between 604 drugs and 680 protein targets (Figure 1). This results in an average number of targets

per drug of 7 and an average of drugs per target of 6.2, which resembles a previously reported projected value of 6.3 targets per drug (Mestres *et al.*, 2008). However, this average fluctuates over the different types of anatomical and therapeutic drugs (see Supplementary Table 2). For instance, Antineoplastic and immunomodulating agents (L) as well as Nervous system drugs (N) present the widest level of promiscuity with their respective average of targets per drug of 15 and 10.2. In contrast, Cardiovascular system drugs (C) seems to be a more target-directed group with an average of 3.8 targets per drug.

On the other hand, we collect a total of 633 interactions between 194 metabolites and 256 proteins (Figure 1), resulting in an average number of proteins per metabolite of 3.2 and an average of metabolites per protein of 2.6. As we work only with annotated interaction data results may show an inherent bias towards therapeutic relevant targets in the case of drugs, and relevant metabolic pathways in the case of metabolites (Mestres *et al.*, 2009). Noteworthy to mention the huge difference of interaction data collected in regard with drugs. This is explained by the novelty of metabolomics as a field together with the interest of drugs for being profiled due to tight toxicity and efficacy regulation concerns by the FDA.

Up to 52% of the proteins interacting with metabolites are also found to be drug targets (overlap n=133) between 436 drugs and 134 metabolites. A total of 5,474 *drug – protein – metabolite* associations are estimated by collapsing the drug together with the metabolite interaction profiles. This tripartite association set represents our basis for the further identification of potential competitive points between external agents, in our case drugs, and endogenous human metabolites.

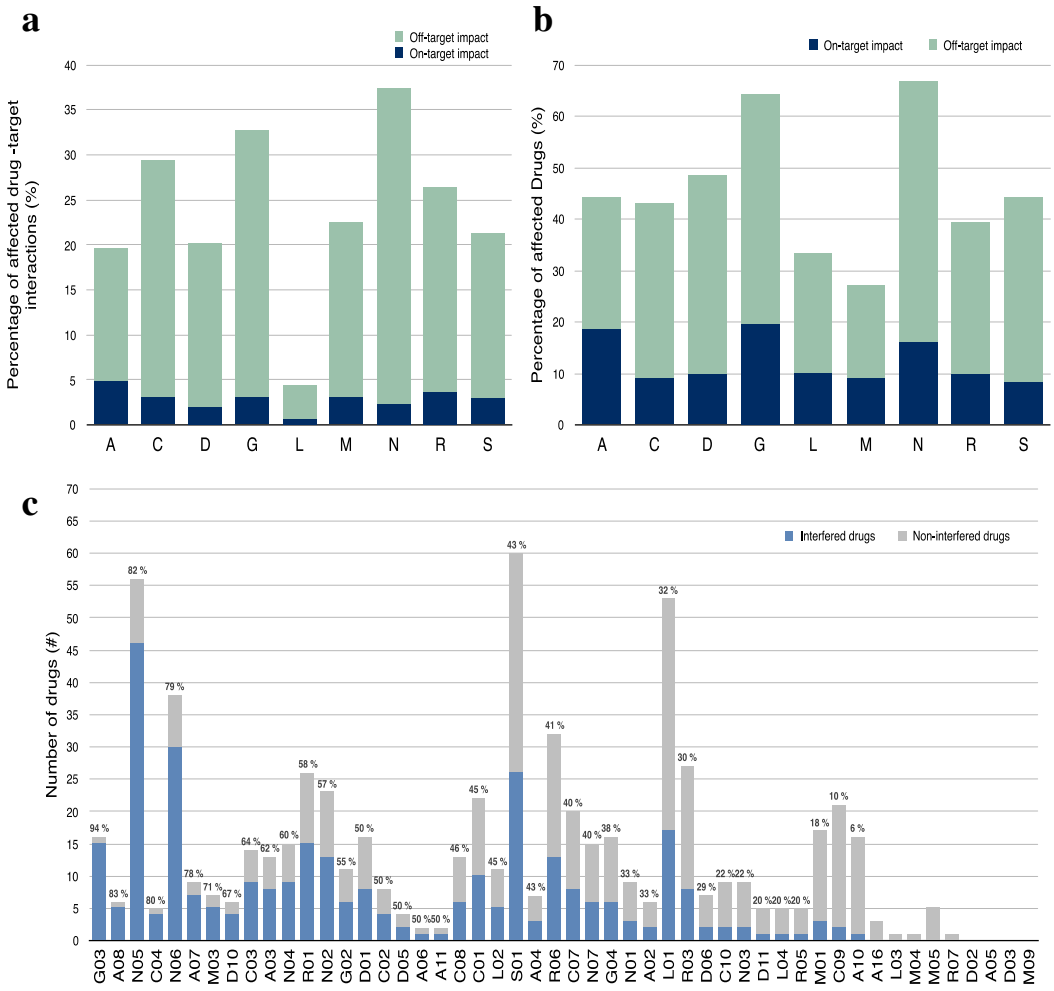


Figure 2. Metabolite interference in drug Polypharmacology. (a) Bar height reflects the percentage of interactions interfered by the affinities of endogenous metabolites within each anatomical drug type. Blue portion corresponds to the percentage of primary targets affected and green portion to the percentage of off-targets affected. (b) Bar height reflects the percentage of drugs interfered by the affinities of endogenous metabolites within each anatomical drug type. Blue portion corresponds to the percentage of primary targets affected and green portion to the percentage of off-targets affected. (c) Bar height reflects the number of drugs for each therapeutical drug type of the secondary-level of the ATC system. The different groups are sorted according to the percentage of interfered drugs (top label). Blue portion represents the number of drugs being interfered, and grey portion represents non-interfered drugs.

Human metabolome interference to drug polypharmacology

We define the term *drug – metabolite* interference as the competitive effect between a drug and a metabolite when a metabolite has a greater annotated affinity towards at least one of the drug profile targets. Under this premise, we assume that: (I) both chemical entities have similar kinetics towards a common target; (II) target and site of action is reachable in all of the cases without considering any kind of body compartmentalization; (III) and lastly, metabolites are present in equilibrium with all the targets considered in their active pharmacological profile, hence, reaching at least a minimum concentration equivalent to the activity threshold set in this work (1μM or pAct = 6) (see Material and Methods section).

After applying this criterion, we reduced down to 2,527 *drug – protein – metabolite* interferences (46.2%). This subset of *drug – protein – metabolite* associations is then projected into unique pairwise interferences between drugs and metabolites with a total of 1,543 *drug – metabolite* pairs between 291 drugs and 116 metabolites, being connected through one or diverse protein targets (Figure 1).

Figure 2a shows the relative metabolic impact of the *drug – metabolite* associations at the different anatomical drug types considered (see Material and Methods section). The 25.17% of the FDA-approved *drug – target* interactions network is competitively prone to be compromised by metabolite affinities, of which 9.3% affecting at the on-target level (blue) and 90.7% affecting at the off-target level (green). When inspecting at the different drug types, nervous system (N) and genitourinary system (G) drugs are the most susceptible to metabolite's endogenous competitiveness with respective percentages of 37.4% and a 32.7%. However, only 4.4% of the profile for antineoplastic drugs (L) shows such metabolic vulnerability. If considering interferences inferring at the primary profile, alimentary tract class (A) is the most sensitive type with a

relative on-target impact of the 24.5%, whereas the smallest percentage corresponds to nervous system drugs (6.1%) (see Supplementary Table 2).

In parallel, Figure 2b represents the percentage of affected drugs within each anatomical class. Following the same trend as before, most part of the drug types are mainly affected at their secondary profile (green), with on-target impacts ranging from 8 to 20% (blue). Although, in this case, we consider a drug with on-target impact if its primary target is interfered by any metabolite action, without taking into account any other impact at the secondary profile, while drugs classified as secondary impact are just interfered at their secondary profile. Nervous and genitourinary groups are once again on the top, with more than the 60% of their drugs exposed to metabolite's activity. By contrast, only a 27.3% of drugs belonging to the musculoskeletal (M) class are interfered. In regard to drugs primarily affected at their primary target, genitourinary group is the one with the greatest percentage (19.6%), followed by alimentary tract (18.6%) and nervous system (16%) classes. In contrast, sensory organs (S) group shows simply an 8.2% of its drugs being compromised on their primary target (see Supplementary Table 2).

Lastly, Figure 2c represents the number of drugs included at each therapeutic drug subtype (n=50) decreasingly sorted according to the relative percentage of drugs affected. Blue portion of the bars indicates the number of interfered drugs. At the uppermost we find again genitourinary and nervous subclasses. Concretely, sex hormones and modulators of the genital system (G03), psycholeptics (N05), and psychoanaleptics (N06) with 94%, 82%, and 60% of their drugs affected, respectively.

In general, there is a strong metabolic interference at drug off-targets rather than at specific on-targets. Moreover, some anatomical drug classes seem to have more predisposition to compete with metabolite's action, such as nervous and genitourinary therapeutic strategies. However, the take home message here

is that human metabolism *per se* constitutes an innate protection against exogenous agents, such as drugs, explained somehow by this competitive impact. Nonetheless, the insufficient amount of experimental annotated interaction data, specially for metabolites, directly affects these results, which may be biased towards well-known metabolic and therapeutic pathways. This immediate deficit needs to be taken under consideration and thus we cannot extract any deciding metabolic impact patterns throughout the different anatomical and therapeutic drug classes. For instance, antineoplastic drug class might reasonably be the less affected by metabolite's action due to this lack of interaction data in contrast with nervous and genitourinary classes, which have been of great relevance and consequently further studied.

Identification of specific metabolite markers

As before mentioned, a total of 116 metabolites representing the 59.8% of the initial profiled set, participates at least at one drug interference. This percentage greatly differs when inspecting the different drug anatomical classes (19.1% A, 32.5% C, 22.7% D, 25.8% G, 34% L, 16% M, 39.2% N, 19.1% R, and 29.9% S). We could say that, the greater the number of metabolites potentially inflicting a therapeutic group, the greater metabolically divergent it is. But again, it is important to remember the implicit limitation of annotated interaction data we are carrying with throughout all this work. In such way, results point towards nervous system drugs presenting the greatest participation of metabolites, to be a profoundly divergent and persuadable system.

In general, an average of 13.3 drugs are pointed by a metabolite and an average of 5.3 metabolites are linked to a drug. Again, this average diverges when particularly looking at the different groups. More specifically, an average of 2.81 drugs/metabolite and 3.35 metabolites/drug for A drugs; 3.56 drugs/metabolite and 4.31 metabolites/drug for C drugs; 2.98 drugs/metabolite and 4.37 metabolites/drug for D drugs; 5.04 drugs/metabolite and 7

metabolites/drug for G drugs; 1.48 drugs/metabolite and 4.26 metabolites/drug for L drugs; 1.23 drugs/metabolite and 4.22 metabolites/drug

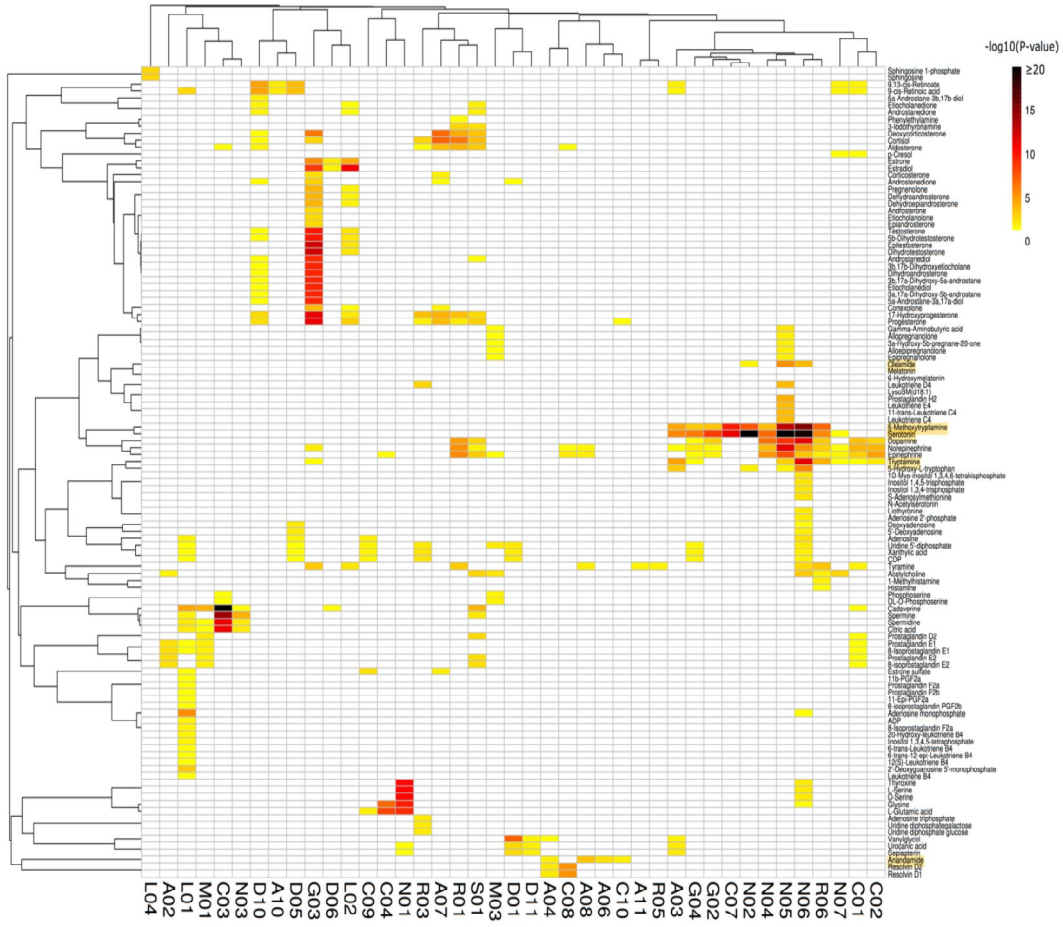


Figure 3. Heatmap for identification of specific metabolite markers. Heatmap representation for the identification of specific enrichments of each of the metabolites shown to be interfering at drug's profile. X-dimension represents the different secondary therapeutical drug classes (n=41), and y-dimension represents the 116 metabolite interferers. Fisher's exact test and its significant two-sided P-values (≤ 0.05) are colored according to the legend (P-values are displayed in negative logarithmic scale from yellow - weak significant evidence - to black - strong significant evidence, whereas uncolored boxes represent no metabolite specificity for the corresponding drug subclass. A total of 9 drug subclasses are excluded for the heatmap representation due to no significant P-values for any metabolite (A05, A16, D02, D03, L03, M04, M05, M09, R07). Drug classes and metabolites are hierarchically distributed based on their relationship according to Pearson correlation coefficients. Metabolites mentioned in section 3.3 are highlighted in yellow.

for M drugs; 9.11 drugs/metabolite and 6.35 metabolites/drug for N drugs; 3.86 drugs/metabolite and 4.47 metabolites/drug for R drugs; and 2.05 drugs/metabolite and 4.41 metabolites/drug for S drugs. Consistent with the previous findings, nervous and genitourinary drugs are in average pointed by a large number of metabolites.

Reached this point, our endeavor now is to see whether there is an inherent specificity for these metabolite interferers towards the different drug anatomical and therapeutical classes, and whether there is a defined relationship between these drug classes according to their metabolites. Figure 3 shows the heatmap representation of specific metabolite markers (y-dimension, n=116) acting at the different therapeutic drug subclasses (x-dimension, n=41). Drug subclasses showing no significant enrichment for any metabolite (P-value > 0.05) were excluded of the heatmap (A05, A16, D02, D03, L03, M04, M05, M09, and R07). Along the y- and x-dimensions, the different therapeutical drug subclasses together with their metabolites are hierarchically distributed according to their Pearson correlation coefficients (P-value ≤ 0.05). We can observe from it that Respiratory (R) and nervous anatomical classes are placed together, which emphasizes the already reported relationship between these type of drugs. Similarly, genitourinary drugs are closely related with alimentary tract system drugs and dermatological drugs (D) (see Supplementary Figure 1). These relationships show somehow the inherent tight connection these systems have within the body. As an aside, there is a high portion of major steroid hormones specifically enriched for genitourinary drugs (i.e. Progesterone, Cortexolone, Testosterone, Androsterone, and Pregnenolone), that are also specific to the nervous system. This reflects in some way how hormones down- and up-regulate a grand number of biological processes within different systems in the body and that they are not exclusives to one system (Barth *et al.*, 2015). In a comparable way, Allopregnanolone, which is a neuroactive metabolite of progesterone and a modulator of the central gamma-aminobutyric acid

receptors, appears to be a specific *chemomarker* perturbing psycholeptics (N05) and muscle relaxant drugs (M03). Moreover, recent studies considered neurosteroids of great potential for the treatment of diverse central nervous system disorders (i.e., epilepsy, pre-menstrual syndrome, infantile spasms, fragile X syndrome, chronic pain, Alzheimer's disease, bipolar disorder, and smoking and alcohol dependencies)(Reddy, 2010; Reddy and Estes, 2016). Conversely and as we could expect, the neurotransmitters Tryptamine and Serotonin, known to be involved in many physiological cycles such as the sleep-wake cycle (Portas *et al.*, 2000), appear to specifically impact at different drug subclasses of the nervous system, particularly anti-parkinsonian agents (N04), psycholeptics and psychoanaleptics. Moreover, serotonin also is identified as a marker for cardiovascular, respiratory, andgenitourinary drugs. Another interesting example is the metabolite Anandamide. This compound is a fatty acid neurotransmitter which effects are mediated primarily by CB1 and CB2 cannabinoid receptors. Among its different functions, there is the regulation of the feeding behavior (Mahler *et al.*, 2007). Here, it appears to be a specific marker of alimentary tract drugs, particularly to the therapeutical subclasses: antiemetics and antinauseants (A04), antiobesity preparations (A08), drugs for constipation (A06), and lipid modifying agents (C10). Finally, last remarkable example worth to bring up is the endogenous metabolite Oleamide. This metabolite is structurally similar to the previous cannabinoid Anandamide, and has been suggested to play an important role during sleep deprivation by inducing sleep (Reyes Prieto *et al.*, 2012). We find it to be a specific marker for nervous system drugs, concretely at psycholeptics, psychoanaleptics, and analgesics (N02).

The identification of such markers should be considered when looking at the metabolic profile of patients under treatment. For example, a recent study found that patients suffering from major depression under Sertraline therapy

with lower levels of serotonin and higher levels of 5-Methoxytryptamine lead to a better outcome (Zhu *et al.*, 2013).

Metabolic Impact at Anti-Parkinsonian drugs polypharmacology

Parkinson's disease (PD) is a neurodegenerative disorder of the central nervous system that leads to chronic and progressive malfunction of the motor system. Current available treatment leads to the occurrence of short and long term undesirable side effects (Müller, 2012) and new therapeutic strategies implying lower toxicity and better efficacy are of actual necessity. Our set of Anti-Parkinsonian drugs is mainly formed by anticholinergic (N04A) and dopaminergic (N04B) agents. Specifically, the 15 PD drugs under study are: Ropinirole (N04B), Seleginile (N04B), Pramipexole (N04B), Rotigotine (N04B), Pergolide (N04B), Bromocriptine (N04B, G02C), Cabergoline (N04B, G02C), Apomorphine (N04B, G04B), Benzatropine (N04A), Procyclidine (N04A), Trihexyphenidyl (N04A), Metixene (N04A), Ethopropazine (N04A), Biperiden (N04A), and Tolcapone (N04B).

The derived *drug – target* network shown in Figure 4a consists of 131 interactions between these 15 drugs and 42 target proteins. The average number of targets per drug is of 8.7, and the average of drugs per target is of 3.1. Distribution of the two main groups of anti-parkinsonians is well defined in the network, bottom part for anticholinergic agents and upper part for dopaminergic agents with Benzatropine appearing to be in the middle sharing some of the dopaminergic targets.

The derived *metabolite-protein* network (See Supplementary Figure 2) is formed by a total of 66 interactions between 8 metabolites (Serotonin, Dopamine, 5-Methoxytryptamine, Epinephrine, Norepinephrine, Tryptamine, 5-Hydroxy-L-tryptophan and Oleamide) and 33 proteins.

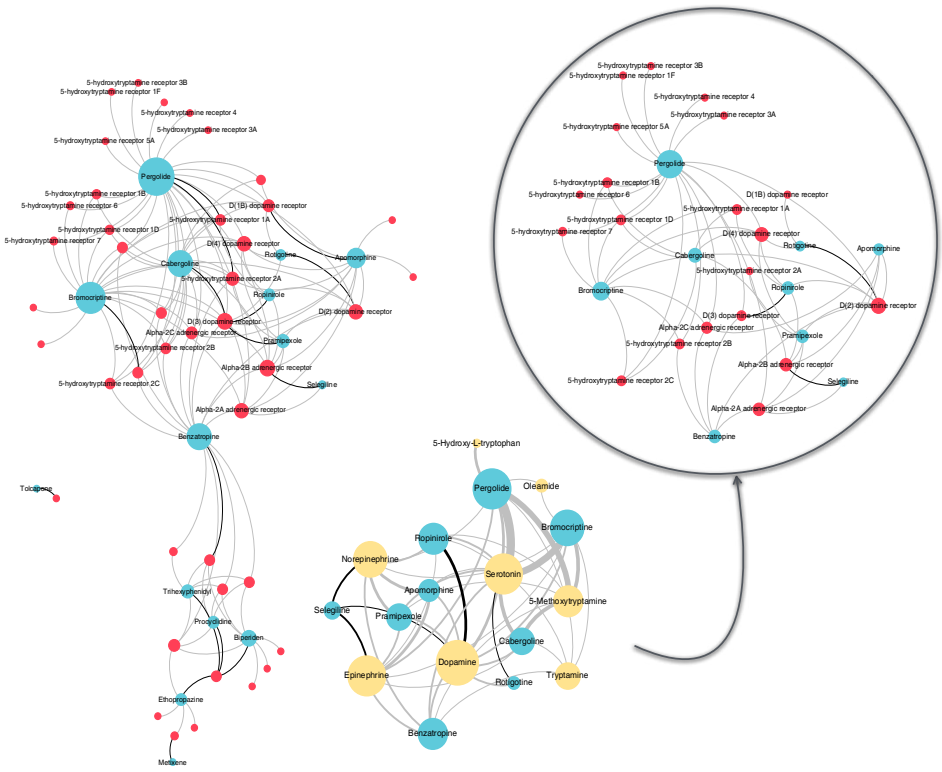


Figure 4. Potential versus actual polypharmacology of anti-Parkinson drugs (N04). (a) Actual *drug – protein* interaction network. Magenta nodes stand for proteins and Blue nodes stand for drugs. Black edge color corresponds to primary/on-target interactions and grey edge color to secondary/off-target intractions. Proteins nodes without any label do not participate in *Drug – metabolite* interferences. (b) *Drug – metabolite* interference network. Blue nodes stand for drugs and yellow nodes for endogenous metabolites. Black edge color corresponds to primary/on-target interferences and grey edge color to secondary/off-target interferences. Edge width correlates with the number of proteins mediating for the interference. (c) Potential *drug – protein* interaction network susceptible of metabolic interference. Black edge color corresponds to primary/on-target interactions and grey edge color to secondary/off-target interactions.

The extracted *drug – metabolite* network accounting for the interferences is formed by these 8 metabolites and 9 drugs with a total of 46 interferences shown in Figure 4b. A total of 20 drug targets (47.62%) are considered as potential competitive points to metabolite affinities (i.e., 5-HT_7 receptor, 5-HT_4 receptor, 5-HT_{1D} receptor, 5-HT_{1B} receptor and 5-HT_{1F} receptor).

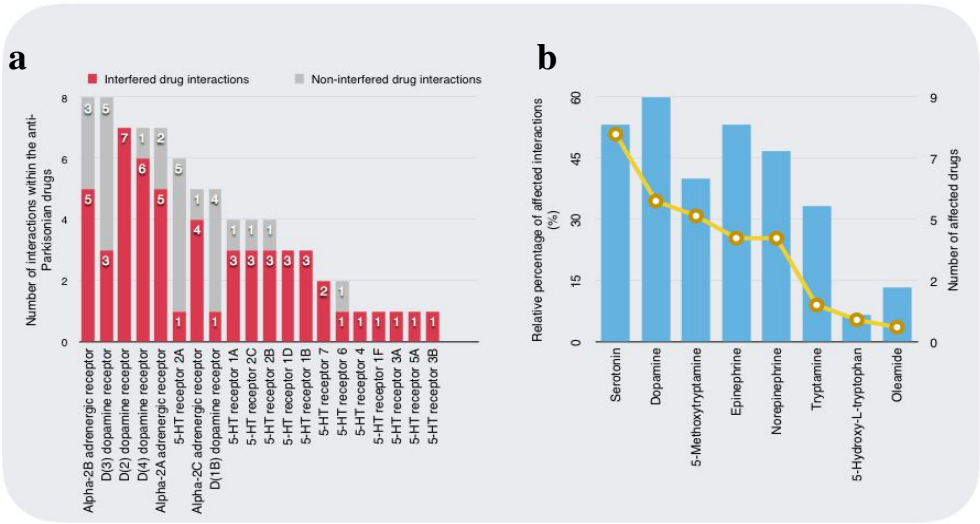


Figure 5. Proteins and Drugs susceptible to Metabolite's action. (a) Ranking by drug occurrence of the proteins participating in interferences. Bar height reflects the number of drug – protein interactions within the anti-Parkinsonian working set of drugs, magenta portion indicates the number of interactions susceptible to metabolite's action, and grey portion indicates the portion of interactions not affected. (b) Metabolites participating in the interference. Bar height reflects the number of affected drugs (right y-axis), and the yellow trend line the percentage of affected interactions within the whole affected anti-Parkinsonian network (left y-axis). Metabolites are decreasingly sorted according to the relative number of affected interactions.

Figure 5a shows these proteins ranked according to their occurrence at the *drug – target* network. In a simplistic way we could say that the most frequently targeted proteins by drugs are also the most relevant ones, so being mainly adrenergic and dopaminergic receptors. The magenta portion indicates the number of affected annotated interactions among the unaffected ones (grey).

Five of these interfering metabolite appear to be specific markers to the anti-parkinsonian group, namely Serotonin, 5-Methoxytryptamine, Epinephrine, Norepinephrine, and Dopamine (Figure 3). Serotonin, the most drug-frequented metabolite, appears to be interfering with a total of 8 drugs (Figure 5b), with interestingly 7 of them belonging to the class of dopaminergic agents, excluding Benzatropine. On the other hand, Dopamine is ranked as the second

most common target for anti-parkinsonian agents and also as the one affecting more number of drugs (n=9) with a total of 19 interfered *drug – protein* associations.

In an overall, the relative metabolic impact for the PD treatment is of the 41.98% and affects to 60% of its drugs. Concretely, with a relative on-target impact of the 5.5%, and a relative off-target impact of the 94.6%. Interestingly, all drugs showing more than a 50% impact in their profiles belong to dopaminergic agents: Ropinorle (100%), Selegiline (100%), Pramipexole (83.3%), Rotigotine (75%), and Pergolide (64%), with exception of Tolcapone that shows no interference to its profile. Actually, also many adverse effects reported in PD patients come from dopaminergic therapies, and even some of the drugs have been withdrawn in some countries, i.e., Pergolide (Elangbam, 2010; Chaudhari *et al.*, 2004). By contrast, the less affected ones are in grand part of anticholinergic nature: Biperiden (0%), Ethopropazine (0%), Metixene (0%), Trihexyphenidyl (0%), Procyclidine (0%), Benzatropine (31.3%), Apomorphine (36.4%), Cabergoline (37.5%), and Bromocriptine (45%).

Metabolite's effect potentially leads to a considerable reduction of the annotated *drug – target* interactions network free of endogenous competitiveness. This prominent reduction also affects to the average number of targets per drug, which turns to 5.9, and a respective average of drugs per target of 2.3. Figure 4c shows the potential *drug – target* annotated interactions susceptible to be interfered by metabolites. Notice that this network exclusively shows now dopaminergic agents (except of Benzatropine), in turn also being those ones resulting in a great number of side effects (Müller, 2012). May be metabolites the ones modulating such adverse reactions in respect to the different physiological status and stage of the disease? Development of motor and non motor symptoms, efficacy and tolerability of therapeutic approaches for this affliction vary from one subject to another. Our message with this Parkinson exemplification is that potential metabolically perturbed drug

interactions should be further considered for assessing a better outcome treatment and a reduction of undesirable drug-related side effects for individual patients. PD, as well as other individual conditions, should be idyllically treated in an individual and holistic manner and not in a population-based pharmacology.

Conclusions

In this study, we propose a novel interpretation of *drug – metabolite* relationship by identifying potential metabolic activities action to drug polypharmacology. To best of our knowledge, this is the first study to investigate endogenous metabolite's relationship with drugs. We also assessed the degree of specificity of such metabolites among different anatomical and therapeutic drug classes. We further exemplified some of our findings by representing the network for anti-Parkinsonian drugs, and how it is affected by such metabolic effect.

Of course, our approach also has some limitations. For example, it does not consider compartmentalization within the body, nor metabolite's or drug's titration, both implied in the reachability of the target. In a more realistic way, such *drug – metabolite* interferences would just take place as long as both chemical entities were available at the site of action, as well as in a sufficient concentration to displace drug's action. However, this strategy aims to qualitatively highlight the existence of *hotspots for interference* between drugs and endogenous physiological conditions, characterized by metabolites activities.

Metabolomics until recently has lacked the data reference resources such as electronic databases equivalent of GenBank and Uniprot for compound identification as other more exploited "omics" as genomics, transcriptomics and proteomics have. The current development and progress of the Human Metabolome Database and the Human Metabolome Project are being of great

importance for the growth of metabolomics as a field, allowing for a better metabolite identification and quantification (Psychogios *et al.*, 2011). Despite just considering a microscopic portion of the human metabolome (~1%) our approach already reveals an inherent interference that human endogenous metabolites offer to any external agent, such as drugs. But this only represents the tip of the iceberg. The different interference trends among the various anatomical classes reported in this work show nervous, genitourinary, and cardiovascular drug classes as the ones with a significantly higher RMI, in contrast with the notable lower metabolic impact at antineoplastic and immunomodulating agents. We cannot draw any plausible interpretation for these results, but emphasize the fact that results may be biased towards important metabolic targets in where nervous, genitourinary and cardiovascular systems and pathways have been well-studied for decades, while antineoplastic targets are yet not connected with many targets and thus, its lower metabolic interference percentage clearly reflects this lack of data completion. The representation of the Parkinson network was chosen because it is a widely studied affliction and it is directly connected with nervous system.

There is much more effort to be done in order to metabolomics be at the same level as the other "omics" by standardization and normalization of Laboratory Information Management System (LIMS) techniques, software analysis development and other *in-silico* approaches, metabolic modeling, and expansion of comprehensive databases. These new footsteps within computational large-scale metabolomics are likely to catch up genomics, transcriptomics and proteomics approaches, and then their final integration will lead us to a better understanding and application of the complexity of systems biology. The application of metabolomics to biological consequences will fill important gaps in system biology. Metabolites have already been used as disease biomarkers. Their up- and down-regulation under specific conditions (diet, sleep patterns, age, smoking, etc.), disease and treatments will allow a better descriptions of

individual's drug action, a better identification of novel therapeutic strategies for different patients, and potentially predict safety and more effective drugs development and prescription for precision medicine.

References

Ai,N. *et al.* (2015) In silico methods for predicting drug-drug interactions with cytochrome P-450s, transporters and beyond. *Adv. Drug Deliv. Rev.*, **86**, 46–60.

Barth,C. *et al.* (2015) Sex hormones affect neurotransmitters and shape the adult female brain during hormonal transition periods. *Front. Neurosci.*, **9**, 1–20.

Bastian,M. *et al.* (2009) Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third Int. AAAI Conf. Weblogs Soc. Media*, 361–362.

Bolton,E.E. *et al.* (2008) PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annu. Rep. Comput. Chem.*, **4**, 217–241.

Chaudhari,K.R. *et al.* (2004) Valvular heart disease and fibrotic reactions may be related to ergot dopamine agonists, but non-ergot agonists may also not be spared. *Mov. Disord.*, **19**, 1522–1523.

Elangbam,C.S. (2010) Drug-induced valvulopathy: an update. *Toxicol. Pathol.*, **38**, 837–848.

Gaulton,A. *et al.* (2012) ChEMBL : a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, 1100–1107.

Jacomy,M. *et al.* (2011) Force Atlas 2, a graph layout algorithm for handy network visualization. … *Http//Www. Medialab. Sci. Fr/* …, 1–21.

Jacomy,M. *et al.* (2014) ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One*, **9**, 1–12.

Jalencas,X. and Mestres,J. (2012) On the Origins of Drug Polypharmacology. *Medchemcomm*, 80–87.

Knox,C. *et al.* (2011) DrugBank 3 . 0 : a comprehensive resource for ' Omics ' research on drugs. *Nucleic Acids Res.*, **39**, 1035–1041.

Kouskoumvekaki,I. and Panagiotou,G. (2011) Navigating the human metabolome for biomarker identification and design of pharmaceutical molecules. *J. Biomed. Biotechnol.*

Liu,T. *et al.* (2007) BindingDB : a web-accessible database of experimentally

determined protein – ligand binding affinities. *Nucleic Acids Res.*, **35**, 198–201.

Mahler,S. V *et al.* (2007) Endocannabinoid hedonic hotspot for sensory pleasure: anandamide in nucleus accumbens shell enhances 'liking' of a sweet reward. *Neuropsychopharmacology*, **32**, 2267–2278.

Mestres,J. *et al.* (2008) Data completeness--the Achilles heel of drug-target networks. *Nat. Biotechnol.*, **26**, 983–984.

Mestres,J. *et al.* (2009) The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol. Biosyst.*, **5**, 1051–1057.

Müller,T. (2012) Drug therapy in patients with Parkinson's disease. *Transl. Neurodegener.*, **1**, 1–13.

Nicholson,J.K. *et al.* (2012) Longitudinal pharmacometabonomics for predicting patient responses to therapy: drug metabolism, toxicity and efficacy. *Expert Opin. Drug Metab. Toxicol.*, **8**, 135–9.

O'Boyle,N.M. *et al.* (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.*, **2**, 5.

O'Hagan,S. and Kell,D.B. (2015) Understanding the foundations of the structural similarities between marketed drugs and endogenous human metabolites. *Front. Pharmacol.*, **6**, 105.

Perspectives,E.H. (2004) What ' s Happening Downstream of DNA. *Environ. Heal.*, **112**, 411–415.

Portas,C.M. *et al.* (2000) Serotonin and the sleep/wake cycle: Special emphasis on microdialysis studies. *Prog. Neurobiol.*, **60**, 13–35.

Psychogios,N. *et al.* (2011) The human serum metabolome. *PLoS One*, **6**.

Reddy,D. (2010) Neurosteroids: Endogenous Role in the Human Brain and Therapeutic Potentials. *Prog Brain Res*, **186**, 113–137.

Reddy,D.S. and Estes,W.A. (2016) Clinical Potential of Neurosteroids for CNS Disorders. *Trends Pharmacol. Sci.*, **xx**, 1–19.

Reyes Prieto,N.M. *et al.* (2012) Oleamide restores sleep in adult rats that were subjected to maternal separation. *Pharmacol. Biochem. Behav.*, **103**, 308–312.

Romero,P. *et al.* (2004) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.*, **6**, 1–17.

Schnackenberg,L.K. and Beger,R.D. (2007) Metabolomic biomarkers: their role in the critical path. *Drug Discov. Today Technol.*, **4**, 13–16.

Shah,N.J. *et al.* (2015) Metabolomics: A Tool Ahead for Understanding

Molecular Mechanisms of Drugs and Diseases. *Indian J. Clin. Biochem.*, **30**, 247–254.

Sharman,J.L. *et al.* (2011) IUPHAR-DB : new receptors and tools for easy searching and visualization of pharmacological data. *Nucleic Acids Res.*, **39**, 534–538.

Sun,P.G. (2015) The human Drug–Disease–Gene Network. *Inf. Sci. (Ny).*, **306**, 70–80.

WHO (2015) Guidelines for ATC Classification and DDD assignment.

Wilson,I.D. (2009) Drugs, bugs, and personalized medicine: pharmacometabonomics enters the ring. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 14187–14188.

Wishart,D.S. *et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.

Wishart,D.S. *et al.* (2008) DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*

Wishart,D.S. *et al.* (2013) HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res.*, **41**, D801–7.

Zhao,S. and Li,S. (2012) A co-module approach for elucidating drug-disease associations and revealing their molecular basis. *Bioinformatics*, **28**, 955–961.

Zhu,H. *et al.* (2013) Pharmacometabolomics of response to sertraline and to placebo in major depressive disorder - possible role for methoxyindole pathway. *PLoS One*, **8**, e68283.

Supplementary Material

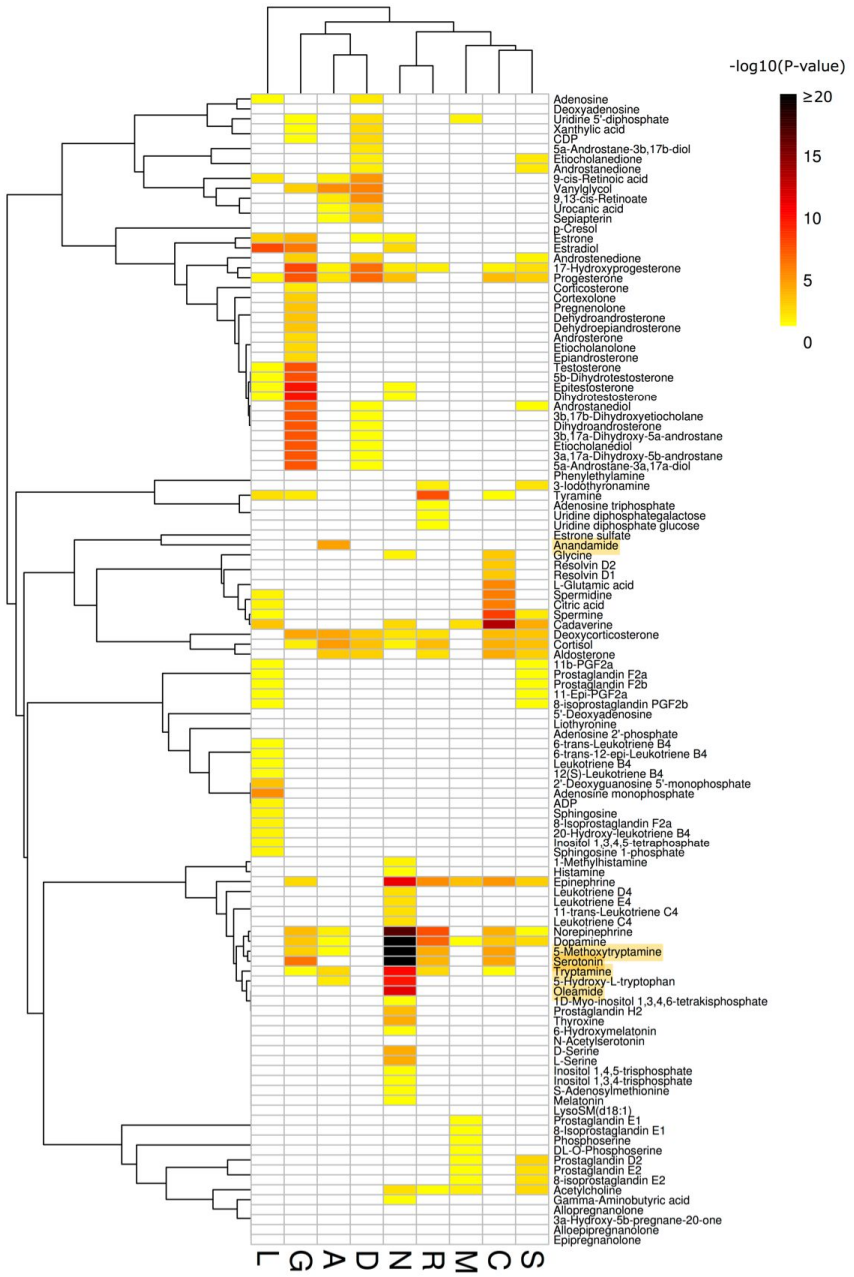
Supplementary Table 1. Anatomical Therapeutic Chemical Classification system (ATC) codes of the drugs under study

Anatomical group	Code	Therapeutic group
Genitourinary system and reproductive hormones (G)	G01	Gynecological infectives and antiseptics
	G02	Other gynecological drugs
	G03	Sex hormones and modulators of the genital system
	G04	Urological drugs
Antineoplastic and immunomodulating agents (L)	L01	Antineoplastic agents
	L02	Endocrine therapy
	L03	Immunostimulants
	L04	Immunosuppressive agents
Alimentary tract and metabolism (A)	A01	Stomatological preparations
	A02	Drugs for acid related disorders
	A03	Drugs for functional gastrointestinal disorders
	A04	Antiemetics and antinauseants
	A05	Bile and liver therapy
	A06	Laxatives
	A07	Intestinal anti-infective agents
	A08	Antiobesity preparations, excluding diet products
	A09	Digestives, including enzymes
	A10	Drugs used in diabetes
	A11	Vitamins
	A12	Mineral supplements
	A13	Tonics
	A14	Anabolic agents for systemic use
	A15	Appetite stimulants
Cardiovascular system (C)	C01	Cardiac therapy
	C02	Antihypertensive drugs
	C03	Diuretic drugs
	C04	Peripheral vasodilators
	C05	Vasoprotective drugs
	C07	Beta blocking agents
	C08	Calcium channel blockers
	C09	Agents acting on the renin-angiotensin system
	C10	Lipid modifying agents
Dermatological drugs (D)	D01	Antifungals for dermatological use
	D02	Emollients and protectants
	D03	Treatment of wounds and ulcers
	D04	Antipyretic drugs
	D05	Antipsoriatic drugs
	D06	Antibiotics and chemotherapeutics for dermatological use
	D07	Topical dermatological corticosteroids
	D08	Antiseptic and disinfectant drugs
	D09	Medicated dressings
	D10	Acne drugs
	D11	Other dermatological drugs
Musculoskeletal system (M)	M01	Anti-inflammatory and antirheumatic products
	M02	Topical products for joint and muscular pain
	M03	Muscle relaxants
	M04	Antigout preparations
	M05	Drugs for treatment of bone diseases
	M09	Other musculoskeletal drugs
Nervous system (N)	N01	Anesthetics
	N02	Analgesics
	N03	Antiepileptics
	N04	Anti-parkinson drugs
	N05	Psycholeptics drugs
	N06	Psychoanaleptics drugs
	N07	Other nervous system drugs
Antiparasitic products, insecticides and repellents (P)	P01	Antiprotozoals
	P02	Anthelmintics
	P03	Ectoparasiticides, incl. scabicides, insecticides and repellents
Respiratory system (R)	R01	Nasal preparations
	R02	Throat preparations
	R03	Drugs for obstructive airway diseases
	R05	Cough and cold preparations
	R06	Antihistamines for systemic use
Sensory organs (S)	S01	Ophthalmologicals
	S02	Otologicals
	S03	Ophthalmological and otological preparations
Various (V)	V01	Allergens
	V03	All other therapeutic products
	V04	Diagnostic agents
	V06	General nutrients
	V07	All other non-therapeutic products
	V08	Contrast media
	V09	Diagnostic radiopharmaceuticals
	V10	Therapeutic radiopharmaceuticals

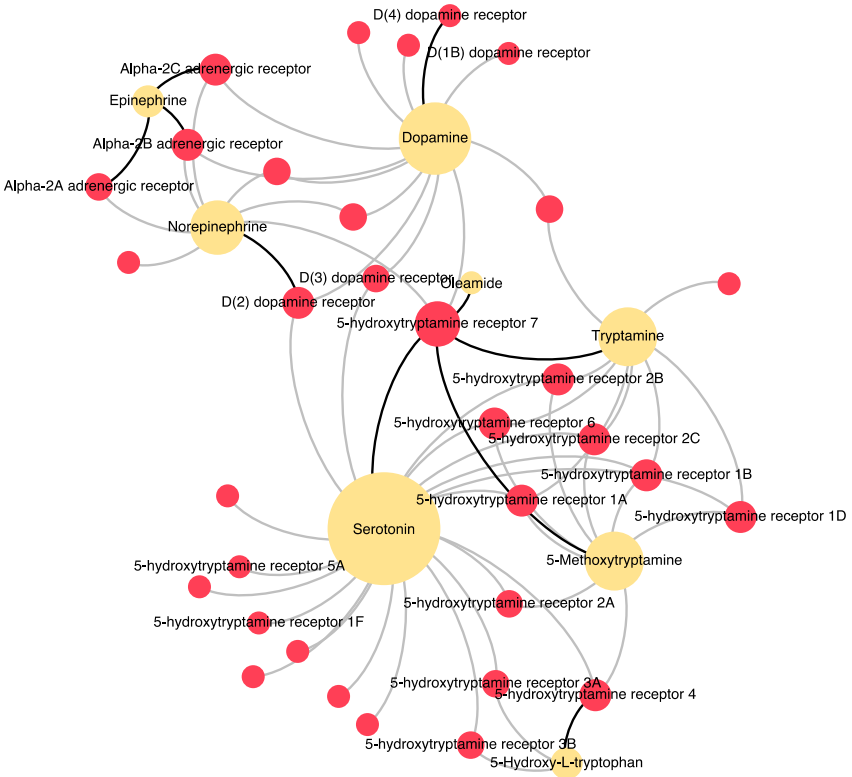
Supplementary Table 2. Properties of the Anatomical drug group networks

Anatomical Classification of drugs	Average targets/drug	Average drugs/target	RMP (%)	On-target (%)*	Off-target (%)*	Affected drugs (%)**	Number of metabolite interferers
All FDA-approved	7.0	6.2	25.2	9.3	90.7	48.2	116
A	3.8	2.3	19.7	24.5	75.5	44.3	37
C	3.8	3.4	29.5	10.5	89.6	43.0	63
D	4.8	4.8	20.1	10.0	90.0	48.4	44
G	6.3	3.4	32.7	9.5	90.5	64.3	50
L	15.0	2.5	4.4	15.6	84.4	33.3	66
M	3.0	1.7	22.5	13.6	86.4	27.3	31
N	10.2	11.2	37.4	6.1	93.9	66.9	76
R	4.5	3.9	26.3	13.7	86.3	39.5	37
S	4.4	2.9	22.8	13.1	80.3	44.3	58

* On-target and Off-target impacts in respect to the general relative metabolic impact (RMP).
** Percentage of affected drugs in respect to the number of drugs for each anatomical set.



Supplementary Figure 1. Heatmap for the identification of specific metabolite markers. Heatmap representation for the identification of specific enrichments of each of the metabolites shown to be interfering at the drug profile. X-dimension represents the different secondary anatomical drug classes (n=9), and y-dimension represents the 116 metabolite interferers. Fisher's exact test and its significant two-sided P-values (≤ 0.05) are colored according to the legend (P-values are displayed in negative logarithmic scale from yellow - weak significant evidence - to black - strong significant evidence, whereas uncolored boxes represent no metabolite specificity for the corresponding drug type. Drug classes and metabolites are hierarchically distributed based on their relationship according to Pearson correlation coefficients. Metabolites mentioned in section 3.3 are highlighted in yellow.



Supplementary Figure 2. Metabolite – protein interaction network. Magenta nodes stand for proteins and yellow nodes for endogenous metabolites. Proteins nodes without any label do not participate in *Drug – metabolite* interferences. Black edge color corresponds to primary/on-target interactions and grey edge color to secondary/off-target intractions.

III.7: Food Interference to Human Endogenous Metabolome

Irene Sanz, Cristina Leal, Joaquim Olivés and Jordi Mestres*

Systems Pharmacology, Research Programme on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute and University Pompeu Fabra, Parc de Recerca Biomèdica, 08003 Barcelona, Catalonia, Spain

Abstract

Motivation: Food has influence in human metabolism and, therefore, in health. Here, we aim to analyze the possible interference of food components to our body. For that purpose, a *metabolite–food* space is built by identifying competitive events between endogenous metabolites and food ingredients towards common protein targets. New *metabolite–food* interactions are sought in order to gain a better understanding of the diet and lifestyle impact over human metabolome.

Results: *Metabolite–food* interfered pairs were associated. Despite food substances seem to not greatly interfere to human metabolome (27,49% interferences among total *metabolite–protein* interactions), interference network is highly connected, showing a non-specific but spread food influence to human metabolome.

* Corresponding autor. *E-mail address:* jmestres@imim.es

Introduction

In the past 20 years, the scientific community has advance significantly in different fields due tot the rising of high-throughput, "omics", technologies. To date, a wide variety of omics disciplines has emerged, and thanks to them researchers are now facing the possibility of connecting food substances, food entities, diet, health, diseases, drugs and metabolism (Capozzi and Bordoni, 2013).

Metabolomics is the systematic study of the chemical processes involving small molecules (metabolites) that characterize the metabolic pathways of biological systems. It can be regarded as the end point of the "omics" cascade, as metabolites are the end products of cellular regulatory processes (Dettmer, K.; Hammock, 2004). Then, changes in the metabolome are the ultimate answer of an organism to genetic or environmental alterations. Consequently, the study of metabolism at a global level has the potential to contribute significantly to biomedical research, clinical medical practice, as well as drug discovery (Manach *et al.*, 2009; German *et al.*, 2005; Trujillo *et al.*, 2006)

Awareness that disease susceptibility is not only dependent on genetic make-up but can be affected by lifestyle decisions, has brought more attention to the role of diet and food (Jensen *et al.*, 2014). To date, scientific literature has reported associations between diet or food/beverage groups and diseases. Indeed, the transition from a traditional diet towards a diet composed of more industrialized, refined and energy-dense foods has led to the well-known worldwide epidemics of obesity and type II diabetes (Fardet and Boirie, 2014). As a result, there are some recommended diets (i.e., Mediterranean and Okinawa diets) (Sofi *et al.*, 2010; Kastorini *et al.*, 2011; Willcox *et al.*, 2009) or food groups (i.e., fruits, vegetables and fish) (He *et al.*, 2007; Smith-Warner *et al.*, 2003) that prevent risk factors for several chronic diseases such as cancer, diabetes, cardiovascular diseases, obesity, etc. On the other hand, the intake of

particular food groups can lead to higher prevalence of these diseases (i.e., high consumption of red/processed meat over many years) (Larsson and Wolk, 2012). Nevertheless, we cannot forget that prevalence of certain diseases, dietary habits, lifestyle or many other factors, differ among populations. Thus, giving a general recommendation to this heterogeneity seems not feasible.

Considerable information on the chemistry and biological properties of dietary phytochemicals has been gathered over the past three decades. Conversely, other food components have been less explored. Furthermore, the ideal information system on food composition is still challenging (electronic resources are particularly scarce in the field of nutrition) (Scalbert *et al.*, 2011; Manach *et al.*, 2009).

In the present work, we aim to study food interference to human metabolism by building a *metabolite–food* space from comparing well-known data of protein associations. A pair of *metabolite–food* is going to be linked through common protein associations, taking into account affinity values. We expect to discover new *metabolite–food* interactions to gain a better understanding of the influence that food and diet may have in our bodies.

Material and Methods

Data retrieval

Metabolite and food compound structures

Available structures for metabolite and food compound molecules were retrieved and processed respectively from Human Metabolome Database (HMDB) version 3.6 (www.hmdb.ca, downloaded on October 2015) (Wishart *et al.*, 2009) and Food Database (FooDB) version 1.0 (www.foodb.ca, downloaded on October 2015), both in XML and SDF format. Then, each structure was mapped by its corresponding InChiKey according to pybel OpenBabel module

(O'Boyle *et al.*, 2008).

Regarding metabolites, only organic and endogenous metabolites were considered. In this respect, a great number of metabolites were found to be wrongly annotated in HMDB. Accordingly, structures were filtered by comparing them to HumanCyc database version 18.5, (www.humancyc.org, downloaded on December 2014) (Romero *et al.*, 2005), followed by a final manual curation of all remaining molecules. Correspondingly, only food compound structures with a reported food association were considered.

Protein targets and association data

Molecule–protein interactions including only experimental reported data were collected from public domain databases, namely, ChEMBL DB version 19 downloaded in 2014 (Gaulton *et al.*, 2012), PubChem imported from ChEMBL version 19 (Bolton *et al.*, 2008), IUPHAR-DB downloaded on June 2014 (Sharman *et al.*, 2011), and BindingDB downloaded on September 2014 (Liu *et al.*, 2007).

Affinity value (pKi, pKd, pIC50 or pEC50) for association data could be active, inactive, undetermined or a quantitative number. However, only experimental *molecule–protein* interactions with an affinity value of pACT $\geq$ 6 (1 μM; pACT = pKi, pKd, pIC50, pEC50) were finally considered.

Among all the protein targets showing an affinity value equal or greater than the abovementioned threshold, only human proteins were selected. Moreover, some protein subunits were collapsed into unique consensus UniProt codes to avoid redundancy. Consequently, a final set of 698 unique protein entities was obtained.

Furthermore, we classified the resulting associated proteins into primary/on-targets and secondary/off-targets for each molecule by sorting their activity values, assuming that the highest pACT corresponds to the primary target.

Final molecule working sets

Initially we started with 18,182 metabolite structures and 23,100 food compound structures. However, as we have declared, not all metabolites were truly endogenous and not all the available food compound structures were pointed to a food source. In this regard, 10,759 ingredient structures could not be related to any food and 30 foods without including any compound were found.

In addition to the different filtering criteria abovementioned, structural collision derived from the overlap between metabolites and food compounds was avoided by discarding the common structures from the food dataset according to the first part of the InChIKey, which is 14 characters long.

After removing all the structures that did not fulfill the several imposed requirements (i.e., truly endogenous metabolites, food related substances, experimental *molecule–protein* associations with pACT $\geq$ 6, human proteins and non structural collision), reduced molecule sets were acquired. Therefore, we ended up with 194 metabolites associated to 256 protein targets and 344 food compounds (encompassed in 838 foods from 23 different food groups) with 336 protein targets.

The number of different metabolites in the human is unknown; experts believe there are at least 2,000–3,000 essential metabolites for normal growth and development (primary metabolites) and thousands more unidentified (around 20,000, compared to an estimated 30,000 genes and 100,000 proteins) that are not essential for growth and development but could represent prognostic, diagnostic, and surrogate markers for a disease state and a deeper understanding of mechanisms of disease (secondary metabolites) (Perspectives, 2004; Kouskoumvekaki and Panagiotou, 2011). However, current HMDB statistics report a total of 2,721 detected and quantified endogenous metabolites, representing the 9.3% among a total of 29,284 endogenous

metabolites (http://www.hmdb.ca/statistics, downloaded on June 2016). Nevertheless, it has not been experimentally found a target-profile for all of these identified and quantified metabolites. Consequently, our final working set represents the 0.66% of the human endogenous metabolome (counting 29,284 as the total) and the 7.13% of the detected and quantified human endogenous metabolome (counting 2,721 as the total).

Metabolite- Food interferences

Comparing the different target profiles from each data source (metabolites and foods), a *metabolite–food* interference space can be built. Common proteins act as the linking points providing a *metabolite–food* association when both molecules share a target protein. Hence, a pair of *metabolite–food* can be associated by one or several proteins.

For each *metabolite-protein-food* connection, the pACT difference (dtpACT) is calculated by subtracting the food compound affinity value for the involved protein to its corresponding metabolite pACT (*Eq. 1*).

We talk about *interference* when dtpACT of a molecule pair for a same protein is negative, unraveling a food compound competition for that metabolite target due to higher affinity (*Eq. 2*). Conversely, when dtpACT values are equal or greater than 0, we talk about *potential interactions*, since metabolites still present higher affinities for their targets (*Eq. 3*). The sum of the two constitutes the total number of *metabolite–food* interactions.

(1) $$dtpACT_{protein} = metabolite\ pACT_{protein} - food\ pACT_{protein}$$

(2) $$if\ dtpACT < 0 \rightarrow food\ pACT > metabolite\ pACT;\quad Interference$$

(3) $$if\ dtpACT \geq 0 \rightarrow metabolite\ pACT > food\ pACT;\quad Potential\ interaction$$

First of all, three different protein target spaces were constructed depending on the classification targets: (a) primary metabolite targets versus all food

compound targets, (b) secondary metabolite targets versus all food compound targets and (c) all metabolite targets versus all food compound targets. Nevertheless, only last case (c) was selected for further study, distinguishing by primary and secondary targets. From this common target space, we classified the resulting *metabolite–food* associations into *interferences* or *potential interactions* according to the dtpACT value, as previously described. We focused on the *interference* set, since we are interested in identifying metabolism alterations due to food intake.

Interferences were examined at three different levels: (1) *metabolites–food compounds*, (2) *metabolites–foods* and (3) *metabolites–food groups*. We scaled up from 1 to 3 by selecting those foods in which food interfering compounds were included (2) and by looking for their pertaining food group (3), such as vegetables, herbs and spices, fruits, etc. Thus, the inference is always settled at the food compound level and ultimately projected to food groups.

Visualization of the Networks

Interaction networks were constructed to visually illustrate *molecule–protein* and *metabolite–food* interference spaces. They were visualized using the Gephi open source package version 0.9.0 (www.gephi.org) (Bastian *et al.*, 2009). Both systems are undirected bipartite networks composed of two sets of nodes depending on each approach (*metabolite–protein* or *food–protein* for target profile networks or *metabolite–food* for interference network). Two nodes are connected by edges when they are interacting.

Ordering and clustering can be processed according to the data. Then, graphical modules like size gradient or color are applied to modify the network display. Node size reflects degree centrality (min size: 10, max sixe: 60), which measures the number of ties a node has to other nodes. Thereby, large nodes are related to wide promiscuities, whereas smaller ones are associated to higher selectivity. Nodes were also colored with respect to the different chemical

entities (yellow=metabolites, blue=food groups and pink=proteins). Finally, edges were colored according to the primary/on-target (black) and secondary/off-target (grey) label of the interactions.

Moreover, highly configurable layout algorithms can be run. We implemented ForceAtlas2 (a force-directed layout) by activating "Dissuade Hubs", "Prevent Overlap" and "Approximate Repulsion" as setting options (Jacomy *et al.*, 2014). It uses classic force-vectors, providing a generic and intuitive way to spatialize networks to allow a visual interpretation of their structure, turning structural proximities into visual proximities. Accordingly, highly connected nodes present higher attractive forces and are thus positioned at the center, whereas weaker nodes with lesser interactions are placed on the periphery.

When needed, we also applied the "k-core" topology filter to select specific nodes and/or edges.

Results and Discussion

Food compounds, food entities and food groups

Each chemical food compound is included in one or more food sources and, in the same way, each food encompasses several food ingredients. On the other hand, each food source points to a single food group, but a same group includes different foods. Additionally, some food groups present a subgroup classification (Fig. S1).

Relationship from food compound level to food group is assessed. In total, there are 23,100 unique compound structures, 888 different food entities and 23 food groups annotated at FooDB. However, not all chemical compounds could be linked to a food source nor to a food group. Consequently, a reduced working set of 12,341 food compound structures (53,42% from the annotated data), 858 food sources and 23 food groups is obtained. It is noteworthy to

mention that the number of food components included within a food can greatly differ among the different ingredients, as well as the number of food entities pointed by a food compound. In principle, we can uncover which food chemicals have a greater impact in metabolism by looking in how many different foods are contained. The more number of foods containing a same compound, the greater the expected impact on metabolism.

We will go in depth with food and food group involvement when talking about *metabolite–food* interference set (*see "Food interference to human metabolome" section*).

Molecule – Protein target spaces and common proteins

We aim to study how diet and food intake may alter the human metabolism. For that purpose, our starting point relies on building individual and independent *metabolite–protein* and *food–protein* target spaces, in order to compare them and identify common protein targets. A shared target profile will link a metabolite to a food compound, showing molecule's mechanism of action over same proteins (enzymes, transporters…). As a result, food ingredients may affect metabolite activities in terms of protein binding affinity and can be considered as possible metabolite competitors for common targets.

After removing unsuitable structures, 194 organic endogenous metabolites and 344 food compounds were captured. Then, experimental protein association profiles for both metabolites and foods were collected from public domain databases (*see "Material and Methods"*).

Regarding metabolite set, we collect 633 *metabolite–protein* associations accounting for 194 unique metabolites and 256 proteins. On the other hand, 841 *food compound–protein* interactions are gathered. These associations are translated to foods and further to food groups by relating each chemical compound to its food source. Then, we obtain 17,341 *food–protein* and 1,722 *food group–protein* associations constituted of 344 unique food compounds, 838 foods, 23 food groups and 336 proteins (Fig. 1). The average number of protein

targets per metabolite is 3.26 (SD ± 3.25) while for food components is 2.46 (SD ± 3.43) (Fig. S2).

Next, proteins from the different data sources are compared. The intersection implies the linking points to relate metabolites to foods. The resulting *metabolite–food* interactions represent potential competitive hotspots between endogenous metabolites and external elements, such as food compounds. Hence, they are classified into *potential interactions* or *interferences* depending on the difference affinity value towards a common protein (*see "Material and Methods"*). However, as we are talking about food ingredients (surely present in our diet) we understand that a *metabolite–food* interaction will suppose a relevant metabolism alteration at certain food component concentrations.

106 protein targets are obtained from the overlap between the two target spaces, meaning that up to 41.41% proteins already known to be associated with metabolites are also found as food compound targets. Actually, 1,006 unique *metabolite–protein–food compound* tripartite interactions are accounted. At this intersection point, we have 125 metabolite structures and 206 food compounds involved.

Food interference to human metabolome

Since we are interested in evaluate how diet and food ingredients may perturb our metabolism, we define a *metabolite–food* interference as the food association to a metabolite when the first one shows a higher affinity towards at least one of the metabolite's profile targets. In such cases, foods and metabolites compete for a same protein target. For that reason, an interference is set down when dtpACT < 0 (food compound pACT > metabolite pACT) (*see "Material and Methods"*).

Following this, 43.74% of the *metabolite–protein–food compound* tripartite interactions are interferences (n=440). From them, 27.73% interfere through primary metabolite targets (n=122) and 72.27% through secondary metabolite

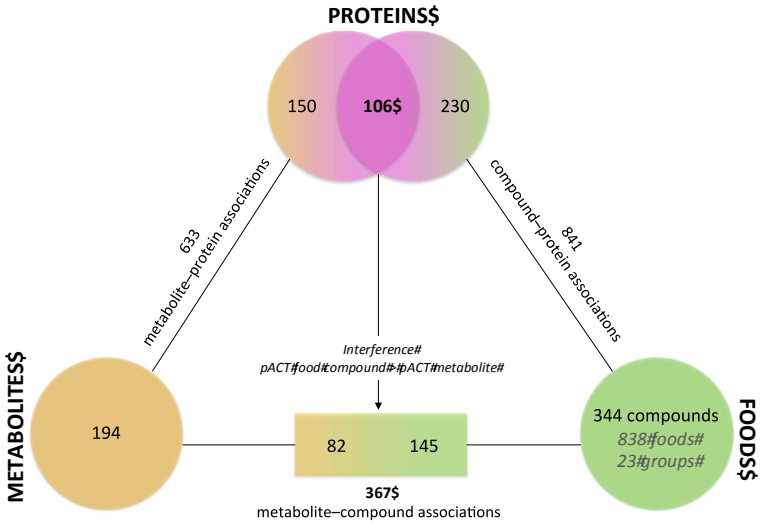


Figure 1. Schema of the different connected elements. A metabolite–food space is built from the protein target intersection of the independent metabolite–protein and food compound–protein networks

targets (n=318). In total, 145 food compounds and 82 metabolites model these interferences, matching into 367 *metabolite–food compound* interfering pairs (Fig. 1).

Further analysis is done based on this interference set, although the remaining interactions might also be significant when food ingredient concentration is higher enough to displace metabolite activity.

Interactions are always established at food compound level, but they are further projected to food groups. Thus, a better representation is achieved allowing us to identify outstanding clusters or interactions of interest. Once focused, we take steps backwards towards food compounds and food sources levels, to perform some statistics and point out more specific conclusions.

In order to examine how and which specific foods interfere to the human metabolism, a computational framework to build *molecule–target* and *metabolite–food* connectivity maps is developed. To do so, we integrate molecule and

molecule/protein connectivity information based on protein interactions, as previously described (Fig. 2).

All networks were exemplified as bipartite undirected networks. We represent proteins, metabolites and food groups as nodes. Edges linking *metabolite–protein* associations distinguish between primary and secondary targets according to different line color (black for primary and grey for secondary). On the other hand, in *food group–protein* network, edges width correlates with the number of foods (and therefore food compounds) mediating the association. Similarly, edge width in *metabolite–food group* network stands for the amount of foods (and ultimately common proteins) taking part in the association (Fig. 2 and 3).

From a total of 633 *metabolite–protein* associations, 174 are affected by food components (acting over common proteins). 145 food compounds (accounting for 831 food sources from 23 food groups), conjointly with 82 metabolite structures and 65 proteins are the key responsible elements in the interferences. Hence, 27.49% of the *metabolite–protein* network is interfered by food. Figure S3 shows the relative food interference impact to the overall *metabolite–protein* interactions.

A cleaned *metabolite–protein* version of the network can be observed when collapsing *metabolite–protein* interactions with *food–protein* ones. As a result of food interference to metabolome, *metabolite–protein* affected associations are split from the original target space leading to unaffected *metabolite–protein* interactions network. Regarding the affected interactions map, two remarkable clusters are noticed. We name them as *cluster 1* and *cluster 2* (Fig. 3).

Cluster 1 is mainly constituted of steroid hormones acting over 6 proteins. In total, 25 metabolites and 19 food compounds share these proteins as common targets (Fig. 3a). Food substances taking part in these interferences show a primary metabolite target competition of 21.62%, while the remaining 78.38% of interactions are through metabolite secondary targets.

As for metabolites grouped in cluster 2, they mainly belong to the nervous system. There are a total of 31 metabolites, 69 food compounds and 38 shared protein targets (Fig. 3b). Among the interferences, 21.95% of them are towards primary metabolite targets while 78.05% are towards secondary targets.

On the other hand, assessment of the most relevant interferences is done in two ways. Primarily, we rank order the metabolites according to how often they are affected by food substances. Secondly, we do the same for food compounds but according to how often they interfere a *metabolite–protein* interaction. Then, the first top 10 for each case are represented (Fig. 4a-b). First ranking order provides information about the most affected metabolites in interferences, while second one shows the food compounds presenting greater interferences. Furthermore, we also look at the related food number profile for each compound participating in interferences (n=145). We ranked the substances in a decreasing order according to their related number of foods, and we finally select the first top 25 substances to be represented (Fig. 4c).

Most interfering food compounds

As we have said, we expect that a food substance would imply a greater metabolome impact as far as it is contained in more foods. Thus, we compare the top 25 food ingredients shown to be present in a greater number of different foods, with the obtained top 10 most interfering substances, to see if there exists a correlation. Certainly, we reach three remarkable coincidences: ergocalciferol, 17alpha-ethynylestradiol and quercetin.

Ergocalciferol, also called vitamin D_2, is the sixth compound most included in foods (n=322) and it is found in all 23 different food groups (Fig. 4c). According to our results, it is the third most interfering substance by the number of affected interactions (interfering to a 2.58% of the total *metabolite–protein* interactions) and the first one most related to different metabolites (n=17) (Fig. 4b). This compound may be used as a vitamin D supplement and

it is officially regarded as equivalent and interchangeable with cholecalciferol (vitamin D_3), which is produced naturally by the skin when exposed to ultraviolet light (Holick *et al.*, 2011). However, conflicting evidence exists for how similarly D_2 and D_3 behave in the body and whether they are equally active or effective (Houghton and Vieth, 2006).

Second example, 17alpha-ethynylestradiol (also called ethinyl estradiol) can be found in 82 different foods classified in 8 different groups: coffee, cocoa, teas, nuts, herbs and spices, vegetables, cereals and fruits (Fig. 4c). It ranks fifth in most interfering ingredients by affected metabolite interactions (2.05%) but third by its metabolite associations (n=12) (Fig. 4b). Ethinyl estradiol is a derivative of 17β-estradiol (E2), the major endogenous estrogen in humans. In pharmacology, it is an orally bioactive estrogen used in the estrogen-progestin combination preparations of oral contraceptives.

Lastly, quercetin is a flavonol included in 194 different food sources (Fig. 4c), projected to 13 different groups (such as herbs and spices, pulses, vegetables and fruits). It is the last compound of the top 10 most interfering substances (affecting to a 0.95% of the total protein metabolic profile) and it is found to be related to 6 different metabolites. This compound is largely used as a nutritional supplement and as a phytochemical remedy for a variety of diseases like diabetes, obesity and circulatory dysfunction (D'Andrea, 2015).

Comparing the interfered metabolite profile of these food compounds, we find that all three interact with epitestosterone and testosterone, which are steroid hormones from the androgen group. In addition, general metabolite profile for ergocalcalciferol, 17alpha-ehynylestradiol and quercetin is related to steroid hormone metabolites. Particularly, as abovementioned, we find an interfered cluster (cluster 1) composed of steroid hormones in which these compounds may have an important role (Fig. 3a).

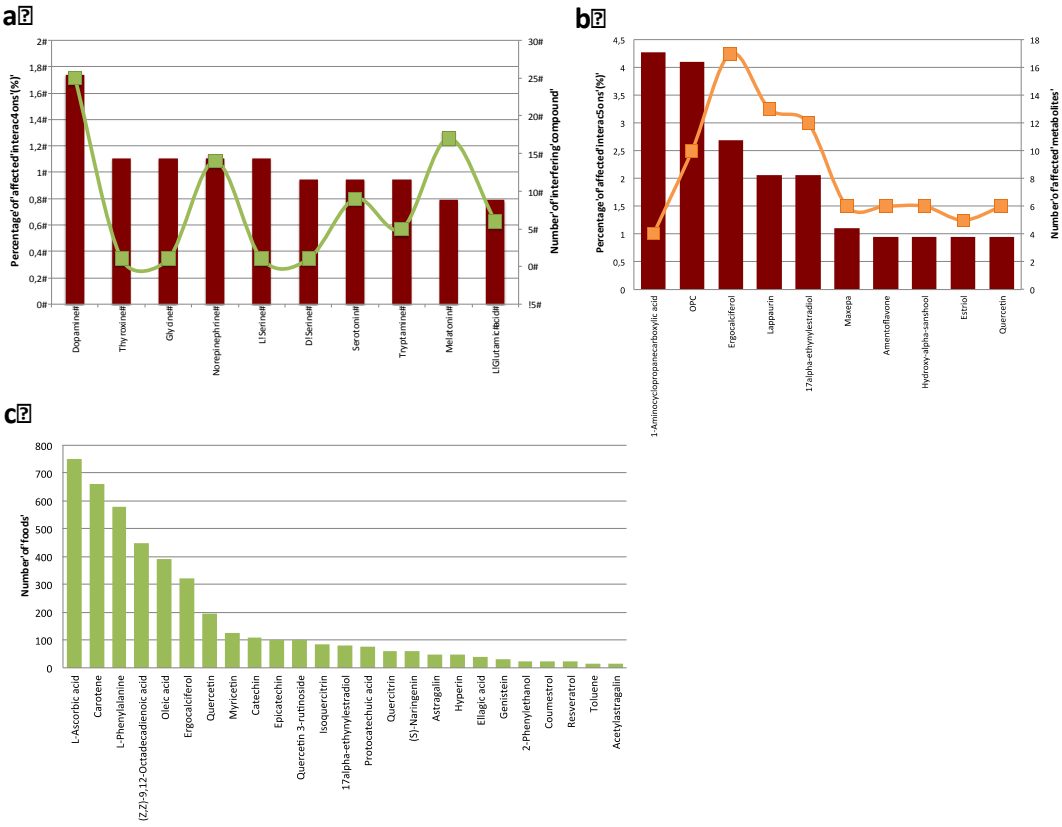


Figure 4. Ranking orders. (a) Top 10 most affected metabolites participating in interferences. Bar height represents the percentage of affected interactions regarding the total of interactions. On the other hand, the green trend line points to the number of food compounds interfering to the metabolite. Metabolites are decreasingly ranked according to the number of affected interactions. (b) Top 10 most interfering food compounds participating in interferences. Bar height represents the percentage of interfering interactions regarding the total of interactions. On the other hand, the orange trend line points to the number of metabolites being interfered by the food compound. Ingredients are decreasingly ranked according to the number of interfering interactions. (c) Top 25 most contained compounds. Bar height stands for the number of foods including each compound.

Most affected metabolites

First of all, we can see that the major part of the 10 most affected metabolites are found in cluster 2, providing an attractive connectivity map of interactions to be further evaluated (Fig. 3b).

Dopamine holds the first position by affected interactions (1,74%) as well as for food chemical associations (n=25) (Fig. 4a). Conjointly with norepinephrine, serotonin and melatonin neurotransmitters, all four share OPC as a common interfering food compound. Certainly, OPC is the second most interfering substance, accounting for 4.11% of the total *metabolite–protein* interactions (Fig. 4b). This chemical compound is found in cinnamon, which has been explored due to its beneficial effects in Parkinsons, diabetes, blood and brain (Kawatra and Rajagopalan, 2015).

Interestingly, glycine, thyroxine and D/L-serine are interfered by a single food compound: 1-Aminocyclopropanecarboxylic acid (ACC), which has been shown to be the most interfering food chemical by affected interactions (4.27%) (Fig. 4b). Moreover, these 4 metabolites stand for the whole metabolite profile of ACC. As a result, possible biological role relies on synaptic plasticity and memory function, since ACC and the four mentioned metabolites competes for NMDA receptor (Inanobe *et al.*, 2005).

Conclusions

In this study we propose a novel interpretation of *metabolite–food* relationship by identifying common mechanism of action between food chemicals and metabolites. In conclusion, we have observed that food and diet may play an important role in human metabolism. However, we have estimated a small interference impact (27.49%), yet highly connected networks.

These findings suggest that influence of diet over metabolism is mainly due to the variety and quantity of food intake, meaning that even food substances by themselves do not show a great impact to displace metabolite activities, they can imply an important effect since a same food chemical can be found in many different food entities, and therefore, in different food groups. Consequently, we observed highly connected interference networks when relating food groups

to metabolites. Despite using food groups to represent and identify the most relevant clusters, our main findings rely on food substances level in view of the specific conclusions that can be gathered, as opposed to the ambiguous results we could obtain by looking at food groups level.

In comparison to other fields, such as pharmacology and drug mechanisms, nutrition has been less explored. In fact, collected information to date is limited to few, well-studied compounds (such as polyphenols, lipids and nutrients). For that reason, we cannot ensure that our findings are reliable consequence of the observed interferences.

On the other hand, our approach has some limitations. First, it does not consider compartmentalization within the body, nor metabolite or food compound concentration levels. Second, we assume similar kinetics between metabolites and food substances towards a common target. Third, we also assume that all food substances are at chemical equilibrium with all their protein targets, meaning that there is the enough concentration of food compound so that it can reach its target and trigger the corresponding response. Hence, we are assuming that all food substances are at least at 1μM of concentration, since we only took into account *molecule–protein* associations with pACT $\geq$ 6 (= 1μM). As a result, our study is not quantitative but qualitative, revealing only the tip of the iceberg of what it can really involve. Then, our strategy aims to highlight the existence of susceptible hotspots for competitive events between food ingredients and endogenous metabolites.

Further study will be assessed to find linking points with diseases pathology, as well as evaluate possible food supplementation or replacement of known drugs, by identifying common activity actions between food compounds and drugs.

References

Bastian,M. *et al.* (2009) Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third Int. AAAI Conf. Weblogs Soc. Media*, 361–362.

Bolton,E.E. *et al.* (2008) PubChem: Integrated Platform of Small Molecules and Biological Activities. In, *Annual Reports in Computational Chemistry*., pp. 217–241.

Capozzi,F. and Bordoni,A. (2013) Foodomics: A new comprehensive approach to food and nutrition. *Genes Nutr.*, **8**, 1–4.

D'Andrea,G. (2015) Quercetin: A flavonol with multifaceted therapeutic applications? *Fitoterapia*, **106**, 256–271.

Dettmer, K.; Hammock,B.D. (2004) Metabolomics — A New Exciting Field within the ' omics ' Sciences. *Environ. Heal. Perspect.*, **112**, A396–397.

Fardet,A. and Boirie,Y. (2014) Associations between food and beverage groups and major diet-related chronic diseases: an exhaustive review of pooled/meta-analyses and systematic reviews. *Nutr. Rev.*, **72**, 741–62.

Gaulton,A. *et al.* (2012) ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**.

German,J.B. *et al.* (2005) Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics*, **1**, 3–9.

He,F.J. *et al.* (2007) Increased consumption of fruit and vegetables is related to a reduced risk of coronary heart disease: meta-analysis of cohort studies. *J. Hum. Hypertens.*, **21**, 717–728.

Holick,M.F. *et al.* (2011) Evaluation, treatment, and prevention of vitamin D deficiency: an Endocrine Society clinical practice guideline. *J. Clin. Endocrinol. Metab.*, **96**, 1911–30.

Houghton,L.A. and Vieth,R. (2006) The case against ergocalciferol (vitamin D2) as a vitamin supplement. *Am. J. Clin. Nutr.*, **84**, 694–697.

Inanobe,A. *et al.* (2005) Mechanism of partial agonist action at the NR1 subunit of NMDA receptors. *Neuron*, **47**, 71–84.

Jacomy,M. *et al.* (2014) ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One*, **9**.

Jensen,K. *et al.* (2014) Integrated Text Mining and Chemoinformatics Analysis Associates Diet to Health Benefit at Molecular Level. *PLoS Comput. Biol.*, **10**, e1003432.

Kastorini,C.M. *et al.* (2011) The effect of mediterranean diet on metabolic syndrome and its components: A meta-analysis of 50 studies and 534,906

individuals. *J. Am. Coll. Cardiol.*, **57**, 1299–1313.

Kawatra,P. and Rajagopalan,R. (2015) Cinnamon: Mystic powers of a minute ingredient. *Pharmacognosy Res.*, **7**, 1.

Kouskoumvekaki,I. and Panagiotou,G. (2011) Navigating the human metabolome for biomarker identification and design of pharmaceutical molecules. *J. Biomed. Biotechnol.*, **2011**.

Larsson,S.C. and Wolk, a (2012) Red and processed meat consumption and risk of pancreatic cancer: meta-analysis of prospective studies. *Br. J. Cancer*, **106**, 603–7.

Liu,T. *et al.* (2007) BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**.

Manach,C. *et al.* (2009) The complex links between dietary phytochemicals and human health deciphered by metabolomics. *Mol. Nutr. Food Res.*, **53**, 1303–1315.

O'Boyle,N.M. *et al.* (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.*, **2**, 5.

Perspectives,E.H. (2004) What ' s Happening Downstream of DNA. *Environ. Heal.*, **112**, 411–415.

Romero,P. *et al.* (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.*, **6**, R2.

Scalbert,A. *et al.* (2011) Databases on food phytochemicals and their health promoting effects. *J. Agric. Food Chem.*, **59**, 4331–4348.

Sharman,J.L. *et al.* (2011) IUPHAR-DB: New receptors and tools for easy searching and visualization of pharmacological data. *Nucleic Acids Res.*, **39**.

Smith-Warner,S. a *et al.* (2003) Fruits, vegetables and lung cancer: a pooled analysis of cohort studies. *Int. J. Cancer*, **107**, 1001–1011.

Sofi,F. *et al.* (2010) Accruing evidence on benefits of adherence to the Mediterranean diet on health: an updated systematic review and meta-analysis. *Am. J. Clin. Nutr.*, **92**, 1189–96.

Trujillo,E. *et al.* (2006) Nutrigenomics, proteomics, metabolomics, and the practice of dietetics. *J. Am. Diet. Assoc.*, **106**, 403–413.

Willcox,D.C. *et al.* (2009) The Okinawan diet: health implications of a low-calorie, nutrient-dense, antioxidant-rich dietary pattern low in glycemic load. *J. Am. Coll. Nutr.*, **28 Suppl**, 500S–516S.

Wishart,D.S. *et al.* (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, **37**, D603–10.

Part IV: Discussion

IV.0: Endogenous Metabolites in Drug Discovery: from Plants to Humans

In this Thesis, I have pursued the main objective of studying the activity of exogenous and endogenous compounds on human metabolism. On one hand, I have demonstrated that herbal medicines are performing their therapeutic activity through many different strategies, and many modes of action have been elucidated. On the other hand, we found that polypharmacology is not only common in drugs, but also in plant compounds and metabolites. Moreover it has been developed a framework to predict organism metabolic networks using genomic data. This framework aims to be useful on the elucidation of metabolomes, the identification of microorganisms that synthesize compounds of interest, and also on the construction of synthetic metabolism. Finally, we studied the interferences between metabolites and exogenous compounds, including drugs and food. These interferences could affect on drugs activity, reducing its therapeutic action, or preventing drugs side-effects. On the other hand, food substances may also interfere on metabolome dynamics.

IV.1:Ethnopharmacology

This thesis started with the objective of elucidating the mode of action of medicinal herbs. Many ethnopharmacology studies were performed during last years because of the declining of drug discovery. Several of them used computational methodologies such as data mining, and *in silico* target-based and

ligand-based approaches to identify active compounds and their therapeutic target. However they usually didn't test experimentally their predictions.

In this thesis it has been applied a large-scale *in silico* ligand-based profiling software (CTLink) on a set of compounds found in Catalan herbal medicines. Target profiling allowed the identification of compounds that are active against therapeutic target proteins associated to the diseases treated by these plants.

Methodology has been successfully applied, being able to collect experimental and predicted activity data of herbal compounds supporting the therapeutic use of many traditional medicines. Nevertheless, only with reported experimental data, it was also possible to elucidate, partially or totally, a great amount of herbal therapeutic uses, a 55.8% of the total. One example of therapeutic action validated retrospectively is *Papaver somniferum* and its analgesic and sedative activity. For this plant we have recovered its main active compound, morphine, which is interacting with opioid receptors for producing the therapeutic effect.

Therapeutic actions can be the result of the activity of only one natural product, like morphine in opium poppy, or the result of the synergistic interaction of many compounds. Usually they may be targeting one therapeutic target, or several proteins related with a disease. For bilberry, for example, we have found that its antiatherosclerosis activity is related with the activity of cyanidin, delphinidin and malvidin against 5-arachidonate lypoxigenase.

In our study, experimental data has been complemented with predicted activity data. It provided an increase of the amount of herbal therapeutic applications hypothetically elucidated. 6 predicted interactions were tested experimentally, and 3 were confirmed, validating the methodology applied. They confirmed the

interactions of zeatin riboside with adenosine receptors A1 and 3, isorhamnetin with D(4) dopamine receptor. Trough these interactions zeatin riboside would have a therapeutic effect on the treatment of dysrhythmia and ischemia, and isorhamnetin on hypertension.

Methodology has shown to be useful on the identification of active herbal chemicals and their therapeutic target proteins. However, since compound concentrations may be varying on each herb, it was not possible a completely elucidation of mechanisms of action. Moreover composition may vary in species depending on where they are growing.

Nevertheless, once we know the activity of certain herbal active compound of interest, it is possible to identify the plants from where it can be extracted, and calculate the necessary concentration of the compound to exert the therapeutic action.

These results contribute to give support to the necessity of understanding the combinatiorial action of compounds on herbal medicines, which could be also applied to drug discovery.

IV.2: Polypharmacology in bioactive chemical groups

After predicting the activity of herbal medicine compounds, their promiscuity was compared with those from drugs. Natural products showed a lower promiscuity, indicating they have a higher specificity. After it, the analysis was extended to other bioactive chemical groups, including natural product libraries,

synthetic chemicals, human metabolites and plant compounds. The amount of plant compounds from previous analysis was increased with data from other herbal databases.

In this study it was expected to find higher diversity and lower promiscuity on natural chemicals groups (including Natural Product Libraries, Plant compounds and Metabolites). However, in Natural Product Libraries we found a high in-group structural similarity, and a relatively high promiscuity compared to other groups.

Results revealed drugs as the most promiscuous group, followed by natural product libraries. But promiscuity is increasing in all groups after target profiling.

For plant compounds we obtained a lower average number of targets than for natural products. Moreover, we find lower in-group similarity values. So they can be a source of chemical diversity for the drug discovery, allowing the development of more specific drugs. On the other hand, comparing the quantity of chemicals against those of natural products libraries we can conclude that there's a high amount of compounds that are still missing to be included in these libraries.

Finally, human metabolites usually are only associated to those protein enzymes of the reactions where they are involved, and metabolic reaction networks use to be very linear, implying an average of 2 enzymes for each metabolic. However, in our results the average number of protein targets was higher, suggesting the existence of polypharmacological capacity in metabolites. This may give a new insight on metabolomics, reflected on the creation of more

complex metabolic networks.

IV.3: Human metabolic networks, current state

Thesis followed with the study of the current state of metabolomics. It was analysed the completeness of 3 of the most important metabolomic databases, HMDB, KEGG, and BioCyc, all them including information about metabolites and the reactions that are related with. In this analysis we added to HMDB activity data from other databases.

Some of the compounds with a higher quantity of missing activity data were phosphate and cofactors such as ATP, pyruvic acid or NADPH. Besides, for other intermediate metabolites like succinic acid and oxalacetic acid, KEGG and BioCyc were also adding activity data.

On the other hand, also in ChEMBL and CTLink are adding data to many metabolites found in human, increase metabolic network connectivity.

In the last years some studies have reported data about the possible occurrence of side-reactions en enzymes (Piedrafita et al., 2015). These side-reactions may give as result the production of new metabolites, which moreover, could be active against other proteins, increasing even more network complexity.

Not only the metabolism may be affected by metabolites polypharmacology, but exogenous compounds like drugs activity may also be affected. The existence of a metabolites targeting the same protein of a drug may reduce drug activity, and as consequence its efficacy or side-effects. These interferences between metabolome and drugs has been studied in Chapter III.7.

IV.4: Metabolome reconstruction

Thesis continued with the study of side-reactions and network complexity. It has been perfomed a genome-scale framework for predicting organisms metabolome.

Genome-scale methodology has been applied since some years ago on metabolism reconstruction. They took advantadge of the high number of genome-sequencing projects that appeared during the last 15 years thanks to the advances in sequencing technology. Currently thousands of genomes have been sequenced. On the other hand, these methodologies are also depending on the current quantity of metabolic data available. So they have become more useful along time, being able to perform more complete predictions about metabolomes for a high number of organisms.

Taking as a starting point the state of the art on metabolic reconstruction, it has been developed a genome-scale metabolic network reconstruction framework that includes virtual profiling methodologies. Virtual profiling allows to obtain activity data about metabolites that otherwise wouldn't be predicted. In contrast to other approaches we are predicting the activity of metabolites against other enzymes, where they may act, usually, as substrate or inhibitor. In both cases they are interfiring to enzyme activity, and as consequence cell metabolism. Moreover, when these metabolites are acting as substrate on the reactions, they may result on new metabolites which also may interfere on other reactions.

This method has demonstrated to succesfully predict the metabolome of any

organisms whose genome is available. We didn't obtain high precision values, but recall is high in most of the cases. On the other hand, like other genome-scale methods, it's missing the ability to predict metabolites that are not found in other metabolomes. Nevertheless, it is useful to perform a metabolome reconstruction to support the experimental metabolome research, and moreover understanding relations between metabolites. Aditionally, if some interesting metabolite was identified in some organism, we would be able to look for other organisms that can synthesize it.

For these reconstructions it was observed that most of the predicted data came from the most complete metabolomes, like those of *Homo sapiens* or *Escherichia coli*. But the integration of data from incomplete metabolomes allows to perform a better and more complete prediction. Moreover, the most realiable data is obtained from organisms belonging to the same kingdom as target organism.

IV.5: Inerference between human endogenous metabolome and exogenous compounds

As last part of the Thesis, it was investigated the relation of metabolites with exogenous compounds. On one hand we focused on its role in disease treatments. We studied the interference of metabolome on drugs activity, which may drive to treatment failure or, oherwise, protect against side-effects.

In our results we have identified strong interfence on drug off-targets, meaning that metabolome is offering an intrinsic protection against exogenous agents.

Because of the use of experimental data, results have been affected by the amount of available data. The quantity of metabolic interferences found on each therapeutic pathways are directly related with the knowledge about them, being nervous and genitourinary system drugs some of those which are more affected.

We have represented an example of metabolome interference on anti-parkisonian drugs, with 9 drugs and 8 metabolites competing on 20 targets. Metabolites competition is potentially leading to a reduction of the drug - target interactions, from an average of 5.9 per drug, to 2.3, which may reduce adverse reactions.

This study suggests the considering of metabolic conditions for assessing a appropiated treatments, with better therapeutic efficacy and a reduction of undesired effects, for individual patients.

On the other hand, we investigated the interference impact of food chemicals on metabolome. It was observed little influence of food substance by themselves; however, this impact is related with the amound of food substance intake, implying that a same chemical can has a greater effect if it is ingested from different food entities. Results revealed the existence of susceptible hotspots for competitive events between food ingredients and endogenous metabolites.

Currently, nutrition is unexplored compared to other fields, but from this investigation we can highlight the existence of intereference between diet and metabolome. Knowledge of composition and activity of food groups would allow evaluating food supplementation or replacements of known drugs.

IV.6 Future directions of research

From this thesis I can consider several future research lines. The first one focused on combinatorial medicine, related with the research of synergistic activities. The second one, on metabolic reconstruction, where it would be interesting to improve the methodology to have a better predictive capacity. And a third research line about metabolome protective capacity against drugs.

Designing synergistic drugs

Side effects produced by drugs are common in pharmacology, usually they are caused by polypharmacology. A possible way to avoid them could be the use of synergistic compounds; it would help to avoid side-effects and toxicity produced by high doses of single drugs. Synergistic combination of two or more agents could overcome this undesired effect through biological compensation, sparing doses on each compound, or accessing multi-target mechanisms. On the other hand there are many multi-target diseases whose treatment would improve making use of synergistic drugs. Biological network become very useful on the study of these diseases and the research of effective drugs combinations.

Despite of the prediction of synergistic combinations is complex, as it is reflected in current literature about this field, the development of a tool able to predict synergies would allow giving a step forward on combinatorial pharmacology. Traditional medicine knowledge may have an important role on

this research. But not only traditional medicines are useful, plants in general are a large source of chemical diversity. Systems biology may support the research of potential therapeutic uses of plants, and in the same way the research of synergistic compounds.

Complete metabolome prediction

Genome-scale metabolic network reconstruction framework has proven to be useful on the prediction of a general view of metabolic networks. But it has some limitations, for example on the prediction of metabolites not found in source organisms. Currently the number of metabolites that can be predicted is limited. Despite of the difficulty of predicting new metabolites, there is already methodologies doing this, and virtual profiling could allow us predicting them.

On the other hand, currently we are only using BioCyc data, as observed in Chapter III.4, the inclusion of other databases such as HMDB and/or KEGG would allow improving the methodology. However, the addition of databases can be something complex and time-consuming because they are organized differently and use their own ID codes on data. Otherwise, in Chapter III.3 we observed that chemical diversity of metabolites is higher than drugs. Since currently CTLink is using data from ChEMBL and other drugs activity databases to perform predictions, the inclusion to CTLink of experimental metabolic data would improve the prediction capacity for metabolites activity.

Interference of endogenous metabolome and diet on drug discovery

Endogenous metabolites are commonly present on the organism in different concentrations. Results in Chapter III.6 revealed that metabolome can be affecting on drugs activity; which may decrease its activity or improve

therapeutic action and reduce side-effects. In future studies it would be interesting to go further on this research line, focusing on some specificic drugs classified in specific Anatomical Therapeutic Chemical (ATC) groups, like those from cardiovascular or nervous system, that has been more investigated. Moreover, increasing the metabolic data on this study may allow a deeper research. On the other hand, it would also be of interest to study the interefence between foods and drugs. The identification of common activity actions between food compounds and drugs in order to evaluate possible food complements during drug treatment.

Part V: Conclusions

The main contributions of this Thesis can be summarized as follows:

i) An extensive review covering the current state of research in medicinal plants has been performed. It provides an up-to-date useful overview of current computational methodologies used for the elucidation of the likely mechanisms of action associated with the therapeutic use of plants.

ii) Following on the above, we constructed an integrated database linking plants, molecules, proteins, and diseases and put together a systems protocol to generate mechanistic hypotheses for the therapeutic uses of plants. We were able to predict, first, and experimentally confirm the *in vitro* activity of some chemicals, ribosylzeatin and isorhamnetin, for proteins associated with the therapeutic use of some plants, namely, *Ginkgo biloba, Vitis vinifera*, and *Citrus limon*.

iii) Usually medicinal plants are not exerting their therapeutic action through the activity of a single compound but they are provided as compound mixtures and administered raw or as extracts. In this respect, their therapeutic action can be viewed as the result of the synergism of a collection of compounds that are present at different concentrations in different plants. Therefore, not only the active ingredients may vary, but also the ensemble of interacting proteins. One example is the combination of delphinidin, cyanidin, and malvidin of *Vaccinium myrtillus* (blueberry) to treat atherosclerosis. The combined activity of these compounds against 5-arachidonite lipoxygenase may contribute to the ultimate therapeutic action of the plant.

iv) The amount of pharmacological data available for small molecules

in public repositories is increasing at enormous pace. However, testing millions of small molecules against thousands of proteins seems, at present, unattainable. Consequently, there will always be an issue about data completeness. It is in this aspect that computational methodologies able to predict the pharmacological profile of small molecules are expected to have an increasing impact as well in our quest for elucidating the mechanism of action linked to the therapeutic use of plants.

v) Drug discovery changed its paradigm from the 'one drug – one target' to 'one drug – multiple targets' already more than a decade ago. The investigation of synergistic mechanisms with multiple small molecules at low concentration found in herbal medicines may actually provide insights for investigating a new generation of efficient synergistic medicines with improved safety profiles.

vi) Currently, there are several commercial vendors that supply chemical libraries of natural products. These libraries contain only a small part of the large amount of plant chemicals and certainly do not cover all their chemical diversity.

vii) A metabolome prediction framework that takes the genome of species as input has been developed and integrated with virtual profiling methodologies. This framework can be useful for the creation of an *ab initio* version of an organism metabolome. This initiative can help metabolomics research in the identification of metabolites in samples and their involvement in metabolic reactions. The framework developed obtained high recall values for the predicted metabolomes of several organisms in BioCyc.

viii) Metabolic networks are very complex. This is in part due to the fact that metabolites do have polypharmacology, acting on their native

enzyme with potent affinity but having biologically relevant affinities for multiple other proteins.

ix) The metabolome is an intrinsic protection mechanism of organisms against external compounds. It is interfering with drugs polypharmacology, thus reducing the number of drug-target interactions through competition with the endogenous metabolites. This is a highly novel aspect of this thesis that is yet to be fully exploited in drug discovery.

x) Diet may be also influencing our metabolism. Despite food ingredients being present at low concentrations and offering low levels of competition with endogenous human metabolites, they can have some effects if taken more often than necessary or combined with drugs acting synergistically on similar targets. Hotspots of competitive events between endogenous metabolites and food ingredients were highlighted.

i)

References

(±)-Propranolol hydrochloride | Sigma-Aldrich [WWW Document], n.d. URL http://www.sigmaaldrich.com/catalog/substance/propranololhydrochloride2958031898911?lang=es&region=ES (accessed 8.17.16).

Alomar, M.J., 2014. Factors affecting the development of adverse drug reactions (Review article). Saudi Pharm. J. 22, 83–94. doi:10.1016/j.jsps.2013.02.003

Altman, L.K., 1987. Who Goes First?: The Story of Self-experimentation in Medicine. University of California Press.

Aminov, R.I., 2010. A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future. Front. Microbiol. 1. doi:10.3389/fmicb.2010.00134

Anderson, G.D., 2008. Gender differences in pharmacological response. Int. Rev. Neurobiol. 83, 1–10. doi:10.1016/S0074-7742(08)00001-9

Anighoro, A., Bajorath, J., Rastelli, G., 2014. Polypharmacology: Challenges and Opportunities in Drug Discovery: Miniperspective. J. Med. Chem. 57, 7874–7887. doi:10.1021/jm5006463

Ara, T., Enomoto, M., Arita, M., Ikeda, C., Kera, K., Yamada, M., Nishioka, T., Ikeda, T., Nihei, Y., Shibata, D., Kanaya, S., Sakurai, N., 2015. Metabolonote: A Wiki-Based Database for Managing Hierarchical Metadata of Metabolome Analyses. Front. Bioeng. Biotechnol. 3. doi:10.3389/fbioe.2015.00038

Bent, S., 2008. Herbal Medicine in the United States: Review of Efficacy, Safety, and Regulation. J. Gen. Intern. Med. 23, 854–859. doi:10.1007/s11606-008-0632-y

Bentley, R., 2009. Different roads to discovery; Prontosil (hence sulfa drugs) and penicillin (hence β-lactams). J. Ind. Microbiol. Biotechnol. 36, 775–786. doi:10.1007/s10295-009-0553-8

Black, J., 1999. Claude Bernard on the action of curare. BMJ 319, 622.

Bono, H., Ogata, H., Goto, S., Kanehisa, M., 1998. Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. Genome Res. 8, 203–210.

Bosch, F., Rosich, L., 2008. The Contributions of Paul Ehrlich to Pharmacology: A Tribute on the Occasion of the Centenary of His Nobel Prize. Pharmacology 82, 171–179. doi:10.1159/000149583

Brenner, G.M., Stevens, C.W., 2010. Principles of Pharmacology, in: Pharmacology. Saunders/Elsevier, Philadelphia, PA, p. 9.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Weerasinghe, D., Zhang, P., Karp, P.D., 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 42, D459–D471. doi:10.1093/nar/gkt1103

Catterall, W.A., Perez-Reyes, E., Snutch, T.P., Striessnig, J., n.d. Voltage-gated calcium channels | Ion channels | IUPHAR/BPS Guide to PHARMACOLOGY [WWW Document]. URL http://www.guidetopharmacology.org/GRAC/FamilyDisplayForward?familyId=80 (accessed 1.28.16).

Cragg, G.M., Newman, D.J., 2005. Biodiversity: A continuing source of novel drug leads. Pure Appl. Chem. 77. doi:10.1351/pac200577010007

Csermely, P., Korcsmáros, T., Kiss, H.J.M., London, G., Nussinov, R., 2013. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. Pharmacol. Ther. 138, 333–408. doi:10.1016/j.pharmthera.2013.01.016

de Bolòs, O., Vigo, J., Masalles, R.M., Ninot, J.M., 1997. Flora MANUAL DELS PAÏSOS CATALANS. Editorial Pòrtic, Barcelona.

Dias, D.A., Urban, S., Roessner, U., 2012. A Historical Overview of Natural Products in Drug Discovery. Metabolites 2, 303–336. doi:10.3390/metabo2020303

Drews, J., 2000. Drug discovery: a historical perspective. Science 287, 1960–1964.

Ehrlich, P., Guttmann, P., 1891. Ueber die Wirkung des Methylenblau bei Malaria. Berl. Klin. Wochenschr. 953–956.

Enrique Raviña Rubira, 2011. Evolution of Drug Discovery, in: The Evolution of Drug Discovery: From Traditional Medicines to Modern Drugs. Wiley-VCH, Weinheim, p. 22.

Etkin, N.L., Elisabetsky, E., 2009. Ethnopharmacology: An Overview, in: Ethnopharmacology. EOLSS Publishers Co Ltd, p. 12.

Faber, M.S., Fuhr, U., 2004. Time response of cytochrome P450 1A2 activity on cessation of heavy smoking. Clin. Pharmacol. Ther. 76, 178–184. doi:10.1016/j.clpt.2004.04.003

Ferner, R.E., Butt, T.F., 2012. Adverse drug reactions. Medicine (Baltimore), Clinical Pharmacology 40, 366–370. doi:10.1016/j.mpmed.2012.05.001

Firenzuoli, F., Gori, L., 2007. Herbal Medicine Today: Clinical and Research Issues. Evid.-Based Complement. Altern. Med. ECAM 4, 37–40. doi:10.1093/ecam/nem096

Fitzgerald, M.A., McCouch, S.R., Hall, R.D., 2009. Not just a grain of rice: the quest for quality. Trends Plant Sci. 14, 133–139. doi:10.1016/j.tplants.2008.12.004

Flora, G., Gupta, D., Tiwari, A., 2012. Toxicity of lead: A review with recent updates. Interdiscip. Toxicol. 5, 47–58. doi:10.2478/v10102-012-0009-2

Fulton, M.M., Riley Allen, E., 2005. Polypharmacy in the elderly: A literature review. J. Am. Acad. Nurse Pract. 17, 123–132. doi:10.1111/j.1041-2972.2005.0020.x

Gausachs, R., 2008. Les Herbes Remeieres, Rafael Dalmau. ed. Rafael Dalmau, Editor, Barcelona.

Gelpi, A., Gilbertson, A., Tucker, J.D., 2015. Magic bullet: Paul Ehrlich, Salvarsan and the birth of venereology. Sex. Transm. Infect. 91, 68–69. doi:10.1136/sextrans-2014-051779

Gertsch, J., 2011. Botanical drugs, synergy, and network pharmacology: forth and back to intelligent mixtures. Planta Med. 77, 1086–1098. doi:10.1055/s-0030-1270904

Goerig, M., Schulte am Esch, J., 1991. Friedrich Wilhelm Adam Sertürner - dem Entdecker des Morphins zum 150. Todestag. AINS - Anästhesiol. · Intensivmed. · Notfallmedizin · Schmerzther. 26, 492–498. doi:10.1055/s-2007-1000624

Grundmann, E., 2004. Gerhard Domagk: The First Man to Triumph Over Infectious Diseases. LIT Verlag Münster.

Gurib-Fakim, A., 2006. Medicinal plants: traditions of yesterday and drugs of tomorrow. Mol. Aspects Med. 27, 1–93. doi:10.1016/j.mam.2005.07.008

Habermann, E.R., 1974. Rudolf Buchheim and the Beginning of Pharmacology as a Science. Annu. Rev. Pharmacol. 14, 1–9. doi:10.1146/annurev.pa.14.040174.000245

Hajar, R., 2011. Animal Testing and Medicine. Heart Views Off. J. Gulf Heart Assoc. 12, 42. doi:10.4103/1995-705X.81548

Heinrich, M., 2014. Ethnopharmacology: quo vadis? Challenges for the future. Rev. Bras. Farmacogn. 24, 99–102. doi:10.1016/j.bjp.2013.11.019

Heinrich, M., 2010. 3.12 - Ethnopharmacology and Drug Discovery A2 - Liu, Hung-Wen (Ben), in: Mander, L. (Ed.), Comprehensive Natural Products II. Elsevier, Oxford, pp. 351–381.

Heinrich, M., Edwards, S., Moerman, D.E., Leonti, M., 2009. Ethnopharmacological field studies: A critical assessment of their conceptual basis and methods. J. Ethnopharmacol. 124, 1–17. doi:10.1016/j.jep.2009.03.043

Heinrich, M., Kufer, J., Leonti, M., Pardo-de-Santayana, M., 2006. Ethnobotany and ethnopharmacology--interdisciplinary links with the historical sciences. J. Ethnopharmacol. 107, 157–160. doi:10.1016/j.jep.2006.05.035

Hopkins, A.L., 2007. Network pharmacology. Nat. Biotechnol. 25, 1110–1111. doi:10.1038/nbt1007-1110

Jalencas, X., Mestres, J., 2013. On the origins of drug polypharmacology. Med Chem Commun 4, 80–87. doi:10.1039/C2MD20242E

Ji, H.-F., Li, X.-J., Zhang, H.-Y., 2009. Natural products and drug discovery. Can thousands of years of ancient medical knowledge lead us to new and powerful drug combinations in the fight against cancer and dementia? EMBO Rep. 10, 194–200. doi:10.1038/embor.2009.12

Jiang, M., Xiao, C., Chen, G., Lu, C., Zha, Q., Yan, X., Kong, W., Xu, S., Ju, D., Xu, P., Zou, Y., Lu, A., 2011. Correlation between cold and hot pattern in traditional Chinese medicine and gene expression profiles in rheumatoid arthritis. Front. Med. 5, 219–228. doi:10.1007/s11684-011-0133-y

Journal of Ethnopharmacology, n.d.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30.

Karp, P.D., Krummenacker, M., Paley, S., Wagg, J., 1999. Integrated pathway–genome databases and their role in drug discovery. Trends Biotechnol. 17, 275–281. doi:10.1016/S0167-7799(99)01316-5

King, V.F., Garcia, M.L., Himmel, D., Reuben, J.P., Lam, Y.K., Pan, J.X., Han, G.Q., Kaczorowski, G.J., 1988. Interaction of tetrandrine with slowly inactivating calcium channels. Characterization of calcium channel modulation by an alkaloid of Chinese medicinal herb origin. J. Biol. Chem. 263, 2238–2244.

Kotera, M., Tabei, Y., Yamanishi, Y., Muto, A., Moriya, Y., Tokimatsu, T., Goto, S., 2014. Metabolome-scale prediction of intermediate compounds in multistep metabolic pathways with a recursive supervised approach. Bioinforma. Oxf. Engl. 30, i165-174. doi:10.1093/bioinformatics/btu265

Kotera, M., Tabei, Y., Yamanishi, Y., Tokimatsu, T., Goto, S., 2013. Supervised de novo reconstruction of metabolic pathways from metabolome-scale compound sets. Bioinforma. Oxf. Engl. 29, i135-144. doi:10.1093/bioinformatics/btt244

Koulman, A., Lane, G.A., Harrison, S.J., Volmer, D.A., 2009. From differentiating metabolites to biomarkers. Anal. Bioanal. Chem. 394, 663–670. doi:10.1007/s00216-009-2690-3

Krauss, M., Singer, H., Hollender, J., 2010. LC-high resolution MS in environmental analysis: from target screening to the identification of

unknowns. Anal. Bioanal. Chem. 397, 943–951. doi:10.1007/s00216-010-3608-9

Leonti, M., Sticher, O., Heinrich, M., 2002. Medicinal plants of the Popoluca, México: organoleptic properties as indigenous selection criteria. J. Ethnopharmacol. 81, 307–315. doi:10.1016/S0378-8741(02)00078-8

Lewis, P.R., Karpa, K.D., Felix, T.M., 2015. Adverse Effects of Common Drugs: Adults. FP Essent. 436, 23–30.

Ligon, B.L., 2004. Sir Alexander Fleming: Scottish researcher who discovered penicillin. Semin. Pediatr. Infect. Dis., Problems and Solutions for Antimicrobial Resistance among Pediatric Respiratory Tract and Nosocomiall Pathogens 15, 58–64. doi:10.1053/j.spid.2004.02.002

Liu, Q.-Y., Karpinski, E., Pang, P.K.T., 1992. Tetrandrine Inhibits Both T and L Calcium Channel Currents in Ventricular Cells: J. Cardiovasc. Pharmacol. 20, 513–519. doi:10.1097/00005344-199210000-00001

Lloyd, N.C., Morgan, H.W., Nicholson, B.K., Ronimus, R.S., 2005. The Composition of Ehrlich's Salvarsan: Resolution of a Century-Old Debate. Angew. Chem. 117, 963–966. doi:10.1002/ange.200461471

Ma'ayan, A., 2011. Introduction to Network Analysis in Systems Biology. Sci. Signal. 4, tr5. doi:10.1126/scisignal.2001965

macognogy, P., n.d. Tetrandra root (Fangji)-Stephania tetrandra-Radix Stephaniae tetrandrae. Pharmacogn. Plants Herb. Herb Tradit. Med. Altern. Bot.

Maehle, A.-H., Prüll, C.-R., Halliwell, R.F., 2002. The emergence of the drug receptor theory. Nat. Rev. Drug Discov. 1, 637–641. doi:10.1038/nrd875

Malinovska, A., Bingisser, R., Nickel, C.H., 2015. [Adverse drug events of older patients presenting in the emergency department]. Ther. Umsch. Rev. Thérapeutique 72, 673–677. doi:10.1024/0040-5930/a000736

Morimoto, T., Gandhi, T.K., Fiskio, J.M., Seger, A.C., So, J.W., Cook, E.F., Fukui, T., Bates, D.W., 2004. An evaluation of risk factors for adverse drug events associated with angiotensin-converting enzyme inhibitors. J. Eval. Clin. Pract. 10, 499–509. doi:10.1111/j.1365-2753.2003.00484.x

Muscholl, E., 2001. Schmiedeberg, Oswald, in: eLS. John Wiley & Sons, Ltd.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M., 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 27, 29–34.

Paton, W.D., 1976. An experiment of Claude Bernard on curare: the origins of the Physiological Society [proceedings]. J. Physiol. 263, 26P–29P.

Pelkonen, O., Pasanen, M., Lindon, J.C., Chan, K., Zhao, L., Deal, G., Xu, Q., Fan, T.-P., 2012. Omics and its potential impact on R&D and regulation of complex herbal products. J. Ethnopharmacol., Good Practice in Traditional Chinese Medicine Research in the Post-genomic Era 140, 587–593. doi:10.1016/j.jep.2012.01.035

Peters, J.-U., 2013. Polypharmacology – Foe or Friend? J. Med. Chem. 56, 8955–8971. doi:10.1021/jm400856t

Peters, J.-U., 2012. Polypharmacology in Drug Discovery. John Wiley & Sons.

Piedrafita, G., Keller, M.A., Ralser, M., 2015. The Impact of Non-Enzymatic Reactions and Enzyme Promiscuity on Cellular Metabolism during (Oxidative) Stress Conditions. Biomolecules 5, 2101–2122. doi:10.3390/biom5032101

propranolol intravenous□: Uses, Side Effects, Interactions, Pictures, Warnings & Dosing [WWW Document], n.d. . WebMD. URL http://www.webmd.com/drugs/2/drug-20278/propranolol-intravenous/details (accessed 8.17.16).

Pubchem, n.d. tubocurarine | C37H41N2O6+ - PubChem [WWW Document]. URL https://pubchem.ncbi.nlm.nih.gov/compound/tubocurarine (accessed 8.17.16).

QUIRKE, V., 2006. Putting Theory into Practice: James Black, Receptor Theory and the Development of the Beta-Blockers at ICI, 1958–1978. Med. Hist. 50, 69–92.

Raviña Rubira, E., 2011. The evolution of drug discovery: from traditional medicines to modern drugs. Wiley-VCH, Weinheim.

Reddy, A.S., Zhang, S., 2013. Polypharmacology: drug discovery for the future. Expert Rev. Clin. Pharmacol. 6, 41–47. doi:10.1586/ecp.12.74

Rollinger, J.M., Schuster, D., Danzl, B., Schwaiger, S., Markt, P., Schmidtke, M., Gertsch, J., Raduner, S., Wolber, G., Langer, T., Stuppner, H., 2009. In silico target fishing for rationalized ligand discovery exemplified on constituents of Ruta graveolens. Planta Med. 75, 195–204. doi:10.1055/s-0028-1088397

Roth, B.L., Sheffler, D.J., Kroeze, W.K., 2004. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. Nat. Rev. Drug Discov. 3, 353–359. doi:10.1038/nrd1346

Rubin, R.P., 2007. A Brief History of Great Discoveries in Pharmacology: In Celebration of the Centennial Anniversary of the Founding of the American Society of Pharmacology and Experimental Therapeutics. Pharmacol. Rev. 59, 289–359. doi:10.1124/pr.107.70102

Scheindlin, n.d. A brief history of pharmacology [WWW Document]. URL http://pubs.acs.org/subscribe/archive/mdd/v04/i05/html/05timeline.html (accessed 2.18.16).

Shoshi, A., Ogultarhan, V., Hoppe, T., Kormeier, B., Müller, U., Hofestädt, R., 2015. Identifying adverse drug reactions and drug-induced diseases using network-based drug mapping. J. Bioinform. Comput. Biol. 13, 1540007. doi:10.1142/S0219720015400077

Sibley, J.T., 2015. The Way of the Wise: Traditional Norwegian Folk and Magic Medicine. Xlibris Corporation.

Soejarto, D.D., Fong, H.H.S., Tan, G.T., Zhang, H.J., Ma, C.Y., Franzblau, S.G., Gyllenhaal, C., Riley, M.C., Kadushin, M.R., Pezzuto, J.M., Xuan, L.T., Hiep, N.T., Hung, N.V., Vu, B.M., Loc, P.K., Dac, L.X., Binh, L.T., Chien, N.Q., Hai, N.V., Bich, T.Q., Cuong, N.M., Southavong, B., Sydara, K., Bouamanivong, S., Ly, H.M., Thuy, T.V., Rose, W.C., Dietzman, G.R., 2005. Ethnobotany/ethnopharmacology and mass bioprospecting: issues on intellectual property and benefit-sharing. J. Ethnopharmacol. 100, 15–22. doi:10.1016/j.jep.2005.05.031

Strebhardt, K., Ullrich, A., 2008. Paul Ehrlich's magic bullet concept: 100 years of progress. Nat. Rev. Cancer 8, 473–480. doi:10.1038/nrc2394

Tubbs, R.S., Loukas, M., Shoja, M.M., Shokouhi, G., Oakes, W.J., 2008. François Magendie (1783-1855) and his contributions to the foundations of neuroscience and neurosurgery. J. Neurosurg. 108, 1038–1042. doi:10.3171/JNS/2008/108/5/1038

van Bronswijk, P., Cohen, A.F., 2008. The first recordings of pharmacological effects. Br. J. Clin. Pharmacol. 66, 588–593. doi:10.1111/j.1365-2125.2008.03316.x

Vickers, A., Zollman, C., Lee, R., 2001. Herbal medicine. West. J. Med. 175, 125–128.

Wachtel-Galor, S., Benzie, I.F.F., 2011. Herbal Medicine: An Introduction to Its History, Usage, Regulation, Current Trends, and Research Needs, in: Benzie, I.F.F., Wachtel-Galor, S. (Eds.), Herbal Medicine: Biomolecular and Clinical Aspects. CRC Press/Taylor & Francis, Boca Raton (FL).

Wagner, H., 2011. Synergy research: Approaching a new generation of phytopharmaceuticals. Fitoterapia, Papers from the 2010 DSHEA Symposium, Chicago, Il, USADSHEA 2010 Symposium Chicago 82, 34–37. doi:10.1016/j.fitote.2010.11.016

Wang, C.-Z., Calway, T., Yuan, C.-S., 2012. Herbal medicines as adjuvants for cancer therapeutics. Am. J. Chin. Med. 40, 657–669. doi:10.1142/S0192415X12500498

Wang, H., Zou, H., Kong, L., Ni, J., 2000. Analysis of Bioactive Components in Traditional Chinese Medicines by Molecular Biochromatography with α1-Acid Glycoprotein Stationary Phase. J. Basic Clin. Physiol. Pharmacol. 11. doi:10.1515/JBCPP.2000.11.2.155

Weisse, A.B., 2012. Self-Experimentation and Its Role in Medical Research. Tex. Heart Inst. J. 39, 51–54.

Wills, R.B., Bone, K., Morgan, M., 2000. Herbal products: active constituents, modes of action and quality control. Nutr. Res. Rev. 13, 47–77. doi:10.1079/095442200108729007

Xie, L., Li, J., Xie, L., Bourne, P.E., 2009. Drug Discovery Using Chemical Systems Biology: Identification of the Protein-Ligand Binding Network To Explain the Side Effects of CETP Inhibitors. PLoS Comput. Biol. 5, e1000387. doi:10.1371/journal.pcbi.1000387

Xie, L., Xie, L., Kinnings, S.L., Bourne, P.E., 2012. Novel Computational Approaches to Polypharmacology as a Means to Define Responses to Individual Drugs. Annu. Rev. Pharmacol. Toxicol. 52, 361–379. doi:10.1146/annurev-pharmtox-010611-134630

Zhang, J., Onakpoya, I.J., Posadzki, P., Eddouks, M., Zhang, J., Onakpoya, I.J., Posadzki, P., Eddouks, M., 2015. The Safety of Herbal Medicine: From Prejudice to Evidence, The Safety of Herbal Medicine: From Prejudice to Evidence. Evid.-Based Complement. Altern. Med. Evid.-Based Complement. Altern. Med. 2015, 2015, e316706. doi:10.1155/2015/316706, 10.1155/2015/316706

Zhao, J., Jiang, P., Zhang, W., 2010. Molecular networks for the study of TCM pharmacology. Brief. Bioinform. 11, 417–430. doi:10.1093/bib/bbp063