Universidade do Minho

Development of a tool for partial automated annotation of metabolic reactions

Pedro Miguel Teixeira Queirós

M.Sc. Thesis on Bioinformatics

September of 2018

School of Engineering

University of Minho

Supervisor:

Ines Thiele, Associate professor at University of Luxembourg

Co-supervisor:

Miguel Rocha, Associate professor at University of Minho

Mentor:

Alberto Noronha, Research associate at the Luxembourg Centre for Systems Biomedicine at University of Luxembourg

Acknowledgements

I would like to thank my M.Sc. thesis supervisor, associate professor Ines Thiele, for the opportunity to do my M.Sc. thesis in the Molecular Systems Physiology Group. Her guidance and support to pursue my own ideas was crucial. The freedom to pursue my personal interests truly allowed the internship period to be professionally and personally rewarding.

I would also like to thank my co-supervisor, associate professor Miguel Rocha, for his availability whenever I had any question. Being the director of my M.Sc. in Bioinformatics and a teacher in several classes should also receive some praise. This course pushed and challenged me to gain invaluable skills in the field of Bioinformatics.

In addition I would like to thank my mentor, research associate Alberto Noronha, his guidance and feedback were essential in the work developed during this thesis. He also aided in getting settled into the group, adding value to the more personal aspect of the present internship.

I would like to thank the several members of the Molecular Systems Physiology Group, their support, feedback, and friendship truly made it feel like the group works as a team guided towards a common goal.

Last, but not least, I would like to thank my family and friends for their unfaltering support. Both the old and new connections were essential to maintaining a healthy personal life.

To add that I am thankful for the financial support provided by the University of Luxembourg and the Erasmus programme.

Abstract

Development of a tool for partial automated annotation of metabolic reactions

Metabolic network reconstructions provide the mathematical in-silico framework for the study of metabolism through the simulation of generic or specific metabolic pathways.

Polyphenols are niche dietary compounds with a growing field of interest in the scientific community. Modelling with these compounds is a step forward in the completion of metabolic reconstructions.
As polyphenols are metabolized by both the gut microbiome and the human host, it is essential to understand the organism-specific metabolic mechanism behind their degradation. Since information is spread out through several information sources and their metabolism is highly complex, it is extremely challenging and time-consuming to manually annotate their metabolism.

The focus of the work here developed was thus the creation of a tool that could speed up the data collection process for posterior manual curation, with a focus on the addition of polyphenol metabolism into the largest and most comprehensive reconstruction of human metabolism, "Recon". This resulted in the creation of the Database Reaction Automatic eXtraction (DRAX) tool, a biological database web scraper. DRAX was initially targeted at polyphenols but it also allows the collection of reactions for other metabolites and drugs.

DRAX allows the comprehensive extraction of metabolite reactions through metabolic pathway-based iterative web scraping. It will provide researchers with a starting point for metabolism reconstruction, allowing a more efficient addition of novel metabolic pathways.

Resumo

Desenvolvimento de uma ferramenta para anotação parcialmente automática de reações metabólicas

Reconstruções metabólicas fornecem um framework matemático in-silico para o estudo de metabolismo através de simulações genéricas ou específicas de pathways metabólicos. Um modelo metabólico é a união de uma reconstrução metabólica com restrições fisiologicamente coerentes.
Polifenóis são substâncias dietéticas com um crescente interesse na comunidade científica. Modelação com estes compostos é portanto um importante passo a dar na completação de reconstruções metabólicas.

Como polifenóis são metabolizados pelo microbioma intestinal e pelo hospedeiro humano, é essencial perceber o metabolismo, específico a cada organismo, por trás da sua degradação. Como a informação para anotar o seu metabolismo está espalhada em várias fontes, este processo torna-se extremamente desafiante e moroso.

O objetivo desta tese foi então o desenvolvimento de uma ferramenta que acelere a coleção desta mesma informação, para que esta possa depois ser curada manualmente por investigadores. Isto levou à criação da ferramenta Database Reaction Automatic eXtraction (DRAX). Inicialmente esta ferramenta focava-se na adição de polifenóis à reconstrução metabólica mais compreensiva, "Recon", contudo também permite a adição de outros metabolitos e drogas.

DRAX permite então a extração de reações metabólicas através de web scraping iterativo baseado em pathways metabólicos. Irá então disponibilizar investigadores com um ponto de começo para reconstrução do metabolismo específico a diversos compostos previamente desconhecido.

Index

List of Figures

List of Tables

List of abbreviations

COBRA - Constraint-Based Model Reconstruction and Analysis

DRAX - Database Reaction Automatic eXtraction

DNA - Deoxyribonucleic acid

kcal - a thousand calories

LPH - Lactase-Phlorizin Hydrolase

GALT - Gut-Associated Lymphoid Tissue

SCFAs - Short-Chain Fatty Acids

AGORA - Assembly of Gut Organisms through Reconstruction and Analysis

MFA - Metabolic Flux Analysis

MNA - Metabolic Network Analysis

GSMMs - Genome-Scale Metabolic Models

FBA - Flux Balance Analysis

sIECs - small Intestinal Epithelial Cells

ORF - Open Reading Frame

ATP - Adenosine TriPhosphate

VMH - Virtual Metabolic Human

API - Application Programming Interface

REST - REpresentational State Transfer

XML - Extensible Markup Language

HTML - Hypertext Markup Language

EC - Enzyme Commission

LCSB - Luxembourg Centre for Systems Biomedicine

SWOT - Strengths, Weaknesses, Opportunities, and Threats

Chapter 1

Introduction

1.1 Context and motivation

The field of nutrition is a relatively recent field in biological sciences which, in recent years, has been under a spotlight, both from a commercial as well as a scientific perspective. Thus, the development of tools and scientific research that allow both nutritionists/dietitians and the scientific community to move past the general nutritional guidelines has become an attractive field of work. With this in mind, both nutrigenomics, nutrigenetics, and surrounding nutritional sciences fields have seen a substantial increase in scientific publications. Nutrigenomics can be defined as a field that studies the effect of food or specific food items on gene expression, while nutrigenetics is defined as a field that studies how our unique genome affects our response to a nutrient [1].

With these definitions and the complexity of human nutrition in mind, it is clear that these two fields would greatly benefit from a better understanding of how different combinations and amounts of nutrients affect the metabolism.

Metabolic network reconstructions provide the mathematical in-silico framework for the study of metabolism through the simulation of generic or specific metabolic pathways. A model is the union of a metabolic network reconstruction with coherent physiological constraints. The main objective is the reconstruction of all or part of the metabolic and transport reaction network within an organism. Such a model for human metabolism (Recon) has already been developed and while still incomplete, it already allows the exploration and comprehension of the complexity underlying human metabolism [2]. Constraint-based modeling has been applied to numerous biomedical questions, including the phenotypic consequences of dietary regimes [3, 4]. However, metabolic modeling of more niche nutritional compounds within the COBRA [5] framework still presents some limitations. The main goal of this thesis will be to assist in overcoming some of these limitations, with the introduction of these niche nutritional compounds into Recon.

The work developed in this thesis will contribute to the current challenges faced in the expansion of increasingly comprehensive metabolic reconstructions. While further work and research are still needed until nutritionists and other clinicians can apply this knowledge in their day-to-day functions, this project contributes towards this ultimate goal. In the future, we can hopefully have a system where food intake contributes to the physical and psychological well-being, while committing itself to the prevention of this century's emerging noninfectious diseases, such as cancer or metabolic syndrome.

1.2 Objectives

This thesis will have the concrete following objectives:

- Review of the state of the art in building and using Constraint-Based Metabolic Model Reconstructions;
- Review the methodology at which metabolic reactions are added to metabolic reconstructions;
- Development of a tool that provides a starting point in the reaction annotation of different compounds, with an emphasis on dietary compounds;
- Develop understanding of polyphenol metabolism through literature review and results provided by aforementioned tool.

1.3 Document structure

This document starts with the chapter "*Introduction*", which begins by defining the MSc thesis goals and objectives, followed by a thorough review of the state of the art. The chapter "*State of the art*" includes a few specific fields of Nutrition, going into more details regarding polyphenols and gut microbiota, in order to provide the reader with enough theoretical background and allowing them to understand the relevance of the work developed. As the work here developed is intrinsically related to metabolic modelling, this topic will also be briefly introduced. The final subchapter will concisely familiarize the reader with more technical aspects of this MSc thesis – web scraping. The "*Methodology and results*" chapter will elaborate on the materials, methodologies and algorithms developed. The following chapter "*Results and discussion*" will show and discuss the results of the work developed. Since the main outcome of this thesis is in providing a service, this chapter will display examples to evaluate functionality. In the "*Conclusion*" a summary of the main outcomes will be discussed, and in the "*Final thoughts and future work*",some ideas for future implementation will be presented.

Chapter 2

State of the art

2.1 Dietary intake across the globe

The human diet is influenced by many factors and interactions: income, food prices, individual preferences and beliefs, cultural traditions, as well as geographical, environmental, social and economic factors all interact in an intricate manner to shape dietary consumption patterns [6]. Food consumption is a key variable in quantifying and assessing the evolution of global and regional food situation. It is important to understand that this assessment comes from national food balance sheets and might not represent actual food consumption. As can be seen in Table 1, the availability of calories per capita has been increasing, with diverse increments across different regions.

Table 1: Global and regional per capita food consumption (kcal per capita per day) , reproduced from [6]

Region	**1964 - 1966**	**1974 - 1976**	**1984 - 1986**	**1997 - 1999**	**2015**	**2030**
World	2358	2435	2655	2803	2940	3050
Developing countries	2054	2152	2450	2681	2850	2980
Near East and North Africa	2290	2591	2953	3006	3090	3170
Sub-Saharan Africa	2058	2079	2057	2195	2360	2540
Latin America and the Caribbean	2393	2546	2689	2824	2980	3140
East Asia	1957	2105	2559	2921	3060	3190
South Asia	2017	1986	2205	2403	2700	2900
Industrialized countries	2947	3065	3206	3380	3440	3500
Transition countries	3222	3385	3379	2906	3060	3180

It seems that there has been a significant progress in raising food availability, however this same evolution has become a double-edged sword. An increased food availability aligned with a progress in urbanization and other socio-economic factors, cause various changes in dietary patterns. Whilst remaining somewhat heterogeneous in their food items, these changes lead to eating patterns that mimic the industrialized Western Diet.
This diet is richer in red meat, refined grains, high-fat dairy products, processed food, high-sugar drinks and fried food [7]. The adoption of this dietary pattern comes in detriment of diets richer in vegetables, fruit, and whole grains. Figure 1 depicts global dietary pattern quality, showing what was previously discussed, that industrialized countries possess worse dietary quality scores.

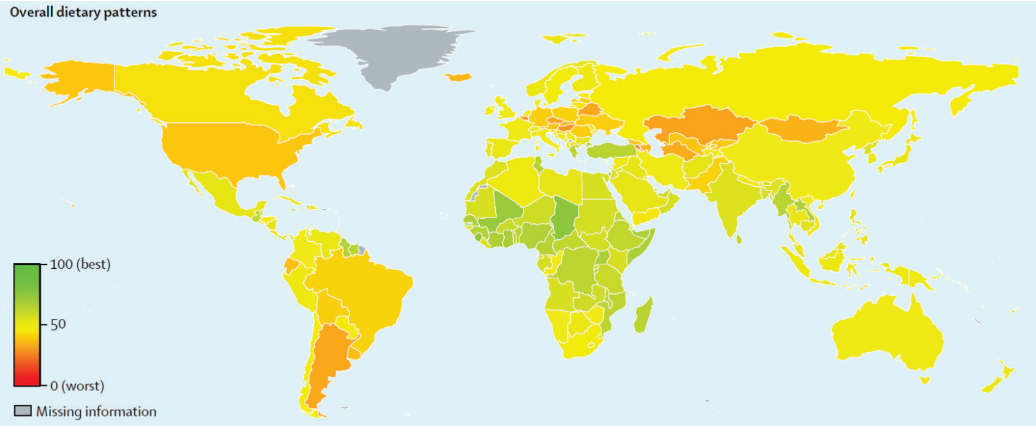


Figure 1: Global dietary patterns in 187 countries in 2010, reproduced from [8]. Diets of better quality have a higher score and vice versa.

This unhealthy dietary pattern, along with other factors, such as tobacco use, harmful use of alcohol and physical inactivity, leads to an increase on the incidence of obesity and non-communicable diseases, illnesses like metabolic syndrome, diabetes, cardiovascular diseases, cancers and chronic respiratory diseases [9, 10, 11, 12, 13]. As previously mentioned, there is a downward trend in the consumption of fiber due to the globalization of dietary patterns poorer in fiber-rich food. There is a systematic replacement of non-processed higher fiber foods for their counterpart processed version, which in most cases have a lower fiber content.

While there's an obvious correlation between an unhealthy diet and the emergence of diseases, there's still a deep lack in the understanding of how specific dietary compounds influence human physiology. Polyphenols are an important and largely misunderstood group of dietary compounds with a wide impact on human health.

2.2 Polyphenols

The contribution of polyphenols in human physiology is an emerging and exciting field in nutrition research. Their importance is however not easy to measure, as in-vivo studies can't fully capture their complex metabolism. Gaining a better understanding of polyphenol metabolism and integrating this knowledge into in-silico tools is one of the main goals of this thesis.

Polyphenols are secondary metabolites that constitute one of the most common groups of substances in plants [14]. Secondary metabolites are part of plants defense mechanism against a variety of organisms (herbivores, microbes, viruses, competing plants, etc), they also act as signaling compounds and offer protection against ultraviolet radiant and oxidants [14]. Secondary metabolites have very diverse biological functions which has led to an extensive genomic reservoir of secondary metabolites transcribing genes. By extension this equates into a colossal variety of these same secondary metabolites [14].

While sharing a common characteristic - the presence of one or more phenols, polyphenols are a highly diverse group of substances [14]. Phenols are a class of chemical compounds that consist of an aromatic hydrocarbon, more specifically a phenyl ring, bonded with a hydroxyl group substituent [15].

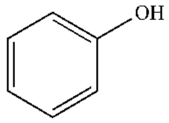


Figure 2: Chemical structure of a phenol

The number of phenols, the structural elements that bind them and the substituents linked to these rings [16] lead to an immense diversity and thus a necessity to separate them into severa sub- classes.

Polyphenols are frequently classified by the number of phenol rings they possess and the structural elements that bind these rings to one another or even by the chemical structure of their aglycones [17, 18, 19]. An aglycone is the non-sugar compounds remaining after the replacement of the glycosyl group from a glycoside by a hydrogen atom [20].

These divisions are usually concordant with one another, albeit more importance will be given to the classification system present in Phenol-Explorer [19], as this database provides a very comprehensive resource of known polyphenols. This database provides polyphenol content in foods for numerous different polyphenols in over 400 foodstuffs. The data contained in this database was collected from more than 1300 scientific publications. Phenol-Explorer divides polyphenols into several classes:

- Phenolic acids:
  - o Benzoic acid derivatives
  - o Cinnamic acid derivatives
- Flavonoids :

o Flavonols

o Flavones

o Isoflavones

o Flavanones

o Anthocyanidins

o Flavanols / Flavan-3-ol

- Stilbenes
- Lignans
- Other polyphenols

Polyphenols remain a highly diverse group of substances. Several thousand have been identified [17] but the focus will be on those that can be found on edible plants, as these are the ones that can potentially alter human physiology through dietary intake. These still go onto the several hundred [17].

2.2.1 Polyphenols in Human Physiology

Humans have had contact with polyphenols for millennia through the dietary intake of plant-based foodstuff, their effect on human health and physiology are unquestionable, albeit hard to measure. This difficulty can mostly be attributed to the fact that polyphenols mostly act chronically and at a low-concentrations. The diseases they may help prevent are also chronical by nature.

The same pathways that metabolize certain xenobiotics or endogenous hormones may be responsible for the metabolism of polyphenols thus exerting a competitive inhibition physiological effect. This may cause an indirect effect on health.

Since most polyphenol metabolites are rapidly eliminated from plasma, in order for these to exert a physiological effect, a constant consumption of polyphenols-rich foodstuff is required. In addition, since polyphenol absorption is limited, their effects on health are usually chronical rather than acute, at least at an average dietary consumption.

Most research on polyphenols is either in-vitro or uses extremely high polyphenol dosage (which misrepresents dietary intake) in order to cause a physiological response. Despite a necessity, this is not the best practice, as any conclusions drawn cannot be simply extrapolated into human metabolism and physiology.

Polyphenols are commonly known for their anti-oxidant properties. However, these compounds display numerous other physiological effects [14], such as:

- Anti-oxidant properties;
- Pro-oxidant properties - in specific circumstances, flavonoids can auto-oxidize and promote hydroxyl-radical formation. Reactive oxygen species can also metabolize flavonoids into highly reactive products.;
- Interaction with signaling cascades;
- Enzyme inhibition;
- Effect on vascular function and angiogenesis;
- Non-covalent binding to proteins;
- Activity with microbial and tissue metabolites;
- Alteration gene expression- by interacting with nuclear receptors that control transcription or inhibiting or activating non-nuclear proteins which can then alter gene expression.

Since polyphenols act chronically and are an ever-present element in the human diet, as knowledge grows, more effects are bound to be found.

While this mechanistic analysis of physiological alterations is important, the ultimate goal is to improve health, and indeed a number of polyphenols have shown some positive health-related benefits, such as the reduction of cancer incidence and tumor growth, lower risk of acute coronary events, prevention of osteoporosis and several more [21]. The array of effects is usually specific to each polyphenol.

2.2.2 Polyphenols as dietary compounds

Polyphenols can be attained through the consumption of plants or foodstuff containing plants. The presence and quantity of polyphenols throughout the plethora of consumed foodstuff are staggeringly diverse, which warrants the creation of databases such as Phenol-Explorer. However, capturing the type and quantity of polyphenols in all in all plant-based foodstuffs remains a challenge.

The remaining content of this chapter will be a brief consolidation of a previously published article by Manach et al. [17], which is a useful review on polyphenols food sources.

To note that each foodstuff has a particular array of polyphenols, their quantity being highly dependent on a number of factors:

- Environmental – pedoclimatic (soil type, sun exposure, rainfall) or agronomic (culture in greenhouses or fields, biological culture, hydroponic culture, fruit yield per tree, etc);
- Storage;
- Ripening;
- Food processing – removal of skin, cooking method, etc;
- Polyphenols in food. Some of the most common food sources for the various polyphenols will be discussed in the following sub-chapters.

Phenolic acids

The hydroxybenzoic acid content of some common edible plants is usually very low, with the exception of some red fruits (eg: strawberries, raspberries and blackberries), black radish and onions. Tea, red wine, some berries and nuts are also a good source of a specific trihydroxybenzoic acid- galliac acid [17, 22].

On the other hand, hydroxycinnamic acids are relatively more common, with the most common ones being p-coumaric, caffeic, ferulic and sinapic acids. These usually can't be found in their free forms, instead, they are typically bound to other compounds forming chlorogenic acids. Blueberries, kiwis, plums, cherries, and apples are amongst the fruits with the highest content of hydrocynamic acids. Caffeic acid is the most predominant hydrocynnamic acid in fruit and ferulic acid in whole cereal grains.

Flavonoids

Flavonols are the most commonly found flavonoids in food, especially the flavonols quercetin and kaempferol. They are commonly found in their glycosylated forms, usually bound to glucose or rhamnose. The richest source of **flavonols** are onions, curly kale, leeks, broccoli, and blueberries. Parsley and celery are the most significant sources of **flavones**; citrus fruits of **flavonones**; soy and other leguminous plants of **isoflavones**; green tea, beans, and chocolate of **flavanols**; and **anthocyanins** in fruit (as they are pigments their content in food is associated with color intensity).

Lignans

There are several sources of lignans, the main source, however, linseed contains around a thousand times more lignans than other minor sources such as algae, leguminous plants (lentils), cereals (triticale and wheat), vegetables (garlic, asparagus, and carrots) and fruit (pears and prunes).

Stillbenes

These polyphenols are found in low quantities in the human diet. One of the most mainstream stilbenes is found in red wine, grapes, peanuts, pistachios and berries [22] - resveratrol, which research suggests to have, among other properties, a protective effect against heart disease, cancer, Alzheimer's and diabetes [23, 24, 25, 26, 27].

In conclusion, polyphenol concentration in foodstuff and its consumption highly varies and is hard to evaluate, the consumption of each polyphenol class even more so. There are country-specific studies on polyphenols consumption but even these have a high standard deviation [17, 28, 29, 30, 31, 32, 33]. However, obvious predictors of polyphenol consumption are the consumption of fruits, berries, vegetables, legumes, cereals, chocolate, and beverages such as tea, coffee and wine.

2.2.3 Polyphenols metabolism

It was previously discussed which foodstuff contains the highest levels of polyphenols, however, one should take heed, as, while a polyphenol might be present in high quantities in foodstuff, it doesn't necessarily mean a high activity within the human body, thus polyphenol activity is always dependent on its own bioavailability. In addition, while the original polyphenol might have a specific biological function, its metabolic derivatives might not share the same function [17]. This is important to note in order to avoid erroneously presuming a polyphenol's physiological function.

Most polyphenols cannot be absorbed in their original form, requiring first metabolic processing by either endogenous or gut microbiota's enzymes.

Polyphenols are fairly active substances and, as such, can generate metabolites in their journey through the human digestive tract.

1. Saliva - hydrolysis of flavonoid glycosides in the mouth is limited to glucose conjugates, this process also being highly determined by the nature of the food matrix, as well as other confounding factors, thus the importance of oral hydrolysis is difficult to assess [14, 17, 22]. To add that food spends little time in the oral cavity, hence polyphenol modification is fairly low.

2. Stomach - this organ is responsible for the mechanical and chemical digestion of the bolus. Isolated polyphenols can be depolymerized and hydrolyzed, however, the mixture of the food matrix confers protection against chemical digestion, as such polyphenols remain fairly stable in the stomach [14, 17, 22].

Saliva and gastric secretions play a rather weak role in the modification of polyphenolic structures.

3. Liver - the liver is responsible for the metabolism of several polyphenols, such as quercetin, luteolin, caffeic acid, and curcumin, to name a few [14, 17].

4. Small intestine - this organ is the primary absorption site of specific flavonoid glucosides. Human lactase-phlorizin hydrolase (LPH) is capable of hydrolyzing glucosides of flavones, flavonols, flavanones, and isoflavones. An alternative pathway is catalyzed by a cytosolic β- glucosidase [14, 17, 22]. The aglycones released are then absorbed by the enterocyte, presumably by passive diffusion.

Before entering systemic circulation, polyphenols undergo phase II enzymatic detoxification, first occurring in the small intestine but then are further metabolized in the liver. Only a small percentage of dietary polyphenols (5-10%)

are initially absorbed in the small intestine [22].

5. Colon - Around 90-95% of dietary polyphenols are too big to be absorbed by the small intestine, and/or cannot by degraded into smaller molecules by the enzymes that are present in the small intestine [22]. Thus these compounds reach the colon, where they can be extensively modified by the gut microbiota [22]. These transformations include multiple and interweaved stages of ester and glycoside hydrolysis, demethylation, dihydroxylation and decarboxylation [34].

It is interesting to note that polyphenols permanence in the human body can vary extremely, some are excreted within 2-3 hours of consumption while others can circulate for more than 3 days [22]. An overview of the polyphenols path through the digestive tract is depicted in Figure 3.

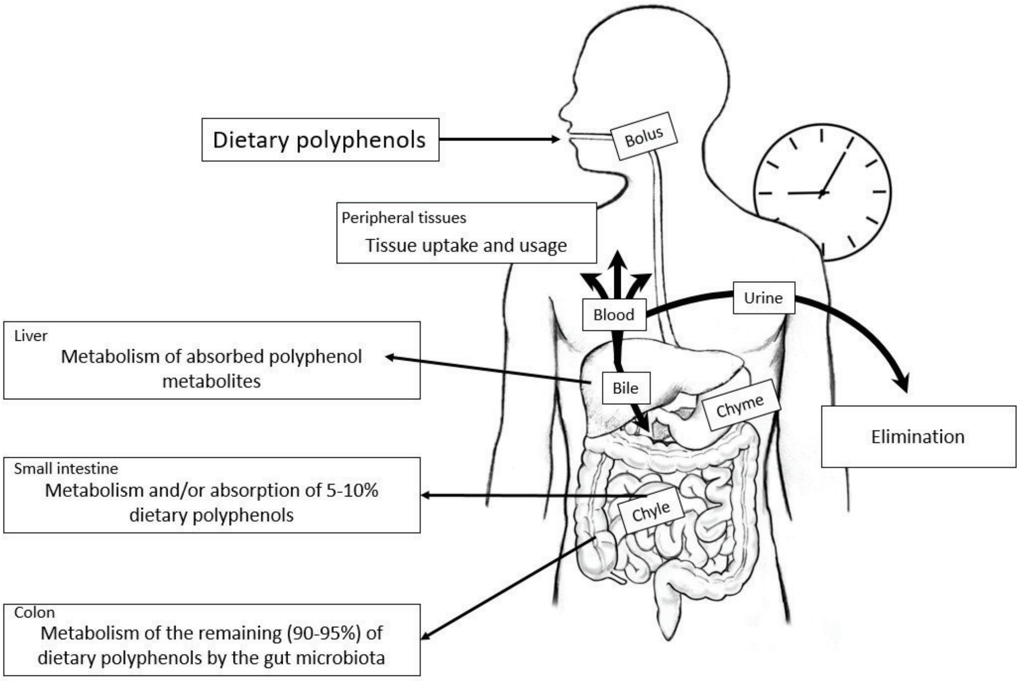

Dietary polyphenols
Bolus
Peripheral tissues
Tissue uptake and usage
Blood
Urine
Liver
Metabolism of absorbed polyphenol metabolites
Bile
Chyme
Elimination
Small intestine
Metabolism and/or absorption of 5-10% dietary polyphenols
Chyle
Colon
Metabolism of the remaining (90-95%) of dietary polyphenols by the gut microbiota

Figure 3: Polyphenols through the digestive tract

Glycosides are hydrolyzed into aglycones, which in turn are then metabolized into several aromatic acids. According to their chemical structure, aglycones are split at different points:

- Flavonols primarily produce hydroxyphenylacetic acids;
- Flavones and flavanones primarily produce hydroxyphenylpropionic acids;
- Flavanols primarily produce phenylvalerolactones and hydroxyphenylpropionic acids.

For future reference, note that metabolism of xenobiotics can be split into 3 phases:

1. Phase I - Modification
2. Phase II - Conjugation
3. Phase III - Excretion

The metabolites produced in the colon then undergo phase II metabolism locally or in the liver. Polyphenols are subjected to 3 main types of conjugation: methylation, sulfation, and glucuronidation. The physiological importance of each type depends on the nature and dosage of the substrate. Sulfation is usually a higher-affinity and lower-capacity pathway than glucuronidation. The ratio of these two types may also be affected by species, gender and food deprivation [35].

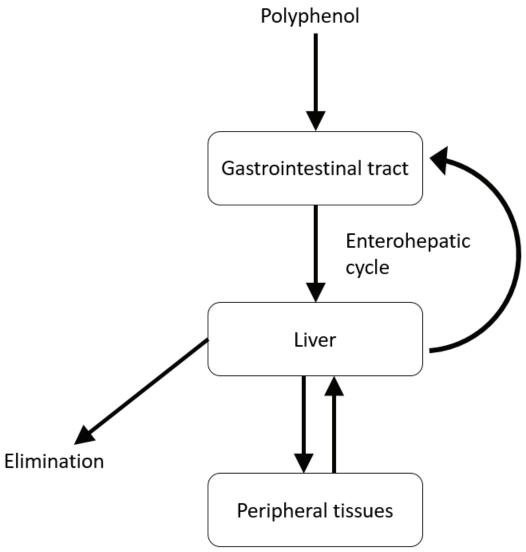


Figure 4: Polyphenols through the digestive tract

After conjugation in the liver, metabolites can reach peripheral tissues or be excreted through the bile or urine. This cycle, named "enterohepatic cycle", is depicted in Figure 4.

During systemic circulation, polyphenols metabolites travel bound to other complexes, albumin being the most common [17], however, to add that polyphenols affinity to albumin varies with chemical structure. The *postprandium* plasma concentration of each polyphenol also depends on the nature of the polyphenol and food source. To note that while polyphenol availability remains a critical criterion for tissue uptake, polyphenol plasma concentration is not directly correlated with tissue uptake. Thus, it remains crucial to identify the organ/tissue-specific polyphenol uptake. In peripheral tissues, polyphenol metabolites can possibly be further metabolized. The metabolites not used in peripheral tissues then travel back into the liver where they are excreted in bile to the small intestine or completely eliminated in urine. Large, extensively conjugated metabolites are more likely to be eliminated through the bile, while small conjugates in urine.

In conclusion, several points in polyphenol metabolism should be taken into consideration:

- Polyphenol metabolism highly fluctuates between classes and even between compounds of the same class. While analyzing polyphenol metabolism one should analyze each polyphenol separately;
- Tissue uptake is not uniform, thus polyphenol concentration must be measured accross the various human body tissues;
- Gut microbiota is essential for polyphenol absorption, thus constituting an important factor for polyphenol metabolism and consequently polyphenol bioavailability. This is yet another perfect example of the symbiotic relationship between humans and their gut microbiota;
- A polyphenol's dietary abundancy, is a poor indicator of physiological impact. An abundant polyphenol may have poor bioavailability and therefore have low physiological impact;
- Conjugated polyphenols are usually not as readily bioavailable as free-form polyphenols, their absorption is very dependent in endogenous or gut microbiota's enzymes;
- Food matrix composition may greatly alter and limit polyphenol absorption. This may be due to an alteration of digestion time, enzymatic competition, and other factors. Consequently, research that studies polyphenol bioavailability by supplementing a high quantity of a singular compound should be taken with a grain of salt, as they do not mirror the average dietary intake.

2.2.4 Gut microbes and polyphenol metabolism

As previously discussed, the gut microbiota plays a very important role in polyphenol metabolism, around 90-95% of dietary polyphenols are metabolized by gut microbes [**BOWEY2003631**]. Table 2 displays which microbe metabolizes each polyphenol class.

Table 2: Microbes responsible for polyphenols metabolism, adapted from [36]

Polyphenolic group	**Polyphenolic compound**	**Gut bacteria**
Flavanols	Kaempferol, Quercetin, Myricetin	Bacteroides distasonis, Bacteroides uniformis, Enterococcus casseliflavus and Eubacterium ramulus
Flavanones	Hesperetin, Naringenin	Clostridium sps and E. ramulus
Flavan-3-ols	Catechin, Epicatechin, Gallocatechin	Bifidobacterium infantis and Clostridium coccides
Anthocyanidins	Cyanidin, Pelagonidin, Malvidin	Lactobacillus plantarum, L. casei, L. acidophilus and Bifidobacterium longum
Isoflavones	Daidzein, Geinstein, Formononentin	Lactobacillus and Bifidobacterium
Flavones	Luteolin, Apigenin	C.orbiscinden, Enterococcus avium
Tannins	Gallo tannins, Ellagitannins	Butyrivibrio sps
Lignin	Secoisolariciesinol, metaresinol, pinoresinol, larciresinol, isolarciresinol, syringiresinol	Species of Bacteroides, Clostridium, Peptostreptococcus and Eubacterium

To remember though that the host likewise plays a role (even if a smaller one), so it should also be considered.

2.3 Gut microbiota

2.3.1 Composition of the gut microbiota

It seems somewhat counterintuitive that the human body hosts a greater number of non-human cells than human's. The human body hosts a plethora of bacteria, archaea, viruses and unicellular eukaryotes, what one often calls microbiota. The microbiota is sustained by its host, coexisting with it, and in some cases even providing benefits [37, 38, 36].

In this thesis, the main focus will be the gastrointestinal tract and its microbiota –gut microbiota, which is by far the heaviest colonized environment. The gut microbiota is composed mainly of strict anaerobes, followed by facultative anaerobes and aerobes [39, 40, 41]. Although it contains several tens of bacterial phyla, it is dominated by two of them: Bacterioidetes and Firmicutes, while Proteobacteria, Verrucomicrobia, Actinobacteria, Fusobacteria, and Cyanobacteria are present in minor proportions [42].

Even though this general profile remains constant, gut microbiota has shown temporal and spatial differences in distribution at the genus level and beyond. From the esophagus to the rectum, a pronounced difference in diversity and number of bacteria can be seen. Frank et al. [43] have reported that different bacterial groups are enriched at different sites when comparing biopsy samples of the small intestine and colon from healthy individuals. Samples from the small intestine were enriched with the Bacilli class of the Firmicutes and Actinobacteria. On the other hand, Bacteroidetes and the Lachnospiraceae family of the Firmicutes were more prevalent in colonic samples [43]. It is also interesting to note that even in the same gastrointestinal area, the microbiota can change between the intestinal epithelium layers. For example, some bacteria can be found close to the epithelium, while other species can be found in the mucus layer and epithelial crypts [44].

Figure 5 below demonstrates the variation and dominant bacteria of human microbiota throughout the digestive system.

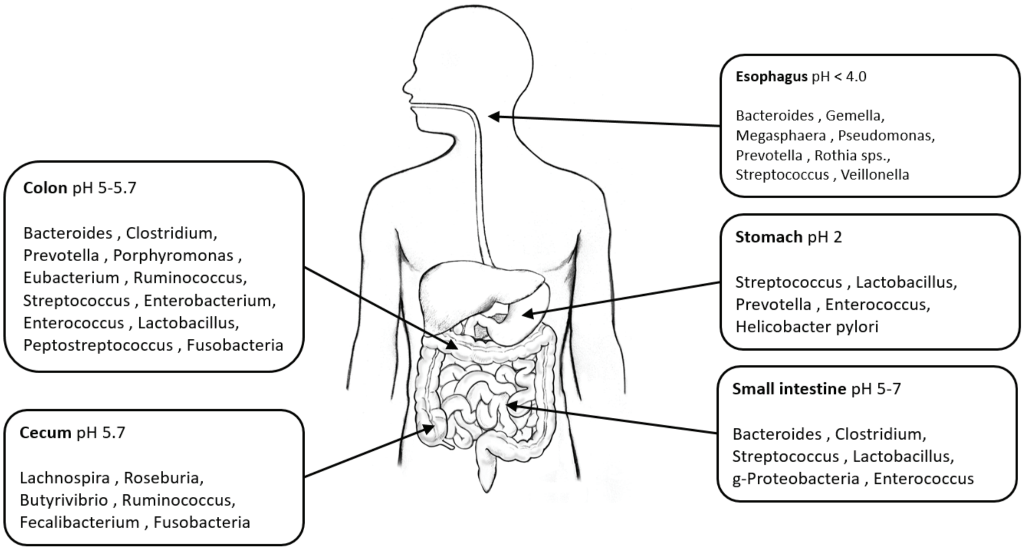

Esophagus pH < 4.0
Bacteroides , Gemella, Megasphaera , Pseudomonas, Prevotella , Rothia sps., Streptococcus , Veillonella
Colon pH 5-5.7
Bacteroides , Clostridium, Prevotella , Porphyromonas , Eubacterium , Ruminococcus, Streptococcus , Enterobacterium, Enterococcus , Lactobacillus, Peptostreptococcus , Fusobacteria
Stomach pH 2
Streptococcus , Lactobacillus, Prevotella , Enterococcus, Helicobacter pylori
Small intestine pH 5-7
Bacteroides , Clostridium, Streptococcus , Lactobacillus, g-Proteobacteria , Enterococcus
Cecum pH 5.7
Lachnospira , Roseburia, Butyrivibrio , Ruminococcus, Fecalibacterium , Fusobacteria

Figure 5: Human gut microbiota distribution. adapted from [36]

2.3.2 Factors affecting gut microbiota

Age

There is growing evidence that the colonization of the human gut microbiota starts in the uterus [45]. It is at birth, however, that microbiota is largely shaped and the species that colonize the gut are highly influenced by the method of conception. Infants that come in contact with their mother's vaginal tract share similarities between their gut microbiota and their mother's vaginal microbiota [46]. Interestingly, infants born through cesarean section and infants born through vaginal delivery have different gut microbiota [47]. Before the first year of life, the gut microbiota is unstable, varying between individuals (due to the inoculation at birth) and with time. After this first year, it starts to resemble the gut microbiota of a young adult and stabilizes [48, 49]. When the infant is 3 years old his/her microbiota is 40-60% similar to that of an adult [50]. In addition to microbiota acquired during birth, several factors have been found to affect gut microbiota. Genetics, for example, can play a part [51, 52] albeit a somewhat indirect one.

During infancy, there is a high prevalence of infancy-related complications which usually lead to the administration of antibiotics. This is not only a common occurrence during infancy though, indiscriminate antibiotics administration is recurrent even during adulthood. Antibiotics can produce temporary, long-lasting or even permanent detrimental effects on gut microbiota diversity [53, 54, 55, 56, 57]. This reason, along with the development of antibiotic-resistant strain[58, 59, 60] , highlights the importance of controlled and cautious antibiotics administration.

Diet

Unsurprisingly, dietary intake also has a role in the composition of microbiota [37]. Breast-milk is considered the gold standard for human nutrition during infancy, having an important role in gut microbiota composition [61]. It is known that it possesses unrivaled properties and, therefore, provides unique benefits during infancy. During this critical stage, breast milk provides the necessary nutritional and physiological demands of the infants, since it contains several bioactive compounds which play a role in nutrient digestion and absorption, immune protection and antimicrobial defense [62, 63]. Breast-fed milk also possesses microbes which can offer benefits to the infant, formula, on the contrary, is sterilized and does not contain any of these beneficial microbes [61].

Another of its benefits is the development of gut microbiota, which will differ if the infant is fed formula instead of breast milk [64]. This superiority is well documented, but some examples will be given. Breast milk is bifidogenic [65], enhancing the production of short-chain fatty acids [66]. Breastfed infants also suffer from fewer allergies and gastrointestinal infections in comparison to formula-fed infants [61]. They also possess a lower intestinal pH than formula-fed infants [65], which inhibits the growth and activity of pathogenic bacteria [67].

A high-fat, high-sugar diet has been associated with reduced gut microbial diversity, whereas a diet rich in fruits, vegetables, and high fiber content is associated with increases in bacterial richness [3]. For example, rural African children (agrarian diet) had a higher proportion of Prevotella, while children from Europe (western diet) had higher abundance of Bacteroides [68]. These species are commonly used as dietary biomarkers [69]. Interestingly, even seasonal dietary variation caused small changes in gut microbiota, highlighting once more its elasticity [70].

While sudden dietary alterations might only cause modest, but nonnegligible, alterations in gut microbiota, long-term dietary modifications can play a more significant role [71].

In the end, given the wide range of contributing factors, it is unexpected that gut microbiota remains fairly stable, with the major microbial phyla that constitute it being conserved [42, 72]. What can be modulated, however, is the proportion of species within phyla which is undoubtedly important given their importance in human health [38, 36]. Thus, both short term and long term dietary interventions should be considered when one aims to induce beneficial modifications on the gut microbiota.

2.3.3 Functionality of the gut microbiota

The human gut microbiome maintains a symbiotic relationship with the gut mucosa, providing sizeable metabolic, immunological and gut protective functions. It obtains its nutrition from the host's dietary intake and can be considered an organ by itself with extensive metabolic capabilities and substantial functional plasticity [36, 73]. The several functions of the gut microbiota will be described in the following sub-sections.

Immunomodulation and protection

Given that the intestinal mucosa is the largest surface area to come in contact with antigens while remaining permeable for nutrient absorption, it is of no surprise that it possesses a well-developed immune system. This gut mucosal immune system is known as the gut-associated lymphoid tissue (GALT). The main sites of GALT include the tonsils and adenoids at the back of the mouth, the Peyer's Patches of the small intestine, the appendix and solitary lymphoid follicles of the large intestine and rectum [74]. This system needs to fulfill two somewhat conflicting tasks, it needs to tolerate the commensal gut microorganisms to avoid an excessive and detrimental systemic immune response, whilst being able to control the gut microbiota or resident pathogens to control their overgrowth and translocation to other systemic locations [38]. In turn, the composition of gut microbiota can influence the regulatory signals leading to an immune response and, therefore, trigger the development of autoimmune or inflammatory diseases [75].

Gut microbiota provides the host with a physical barrier to deter pathogens by competitive exclusion, like the occupation of attachment sites, consumption of nutrients and production of antimicrobial substances. It also stimulates the production of antimicrobial compounds by its host [38].

Xenobiotic and drug metabolism

Gut microbiota can influence xenobiotics metabolism, which can impact therapeutic procedures for treating several diseases[76, 77, 78]. Interindividual and interpopulation differences in gut microbiomes can account for differences in the toxicity of xenobiotics [79].

Structure and Function

There is increasing evidence that gut microbiota contributes to the maintenance and structural and functional development of the gastrointestinal tract. More explicitly, gut microbiota is involved in peristalsis, surface maturation, barrier fortification and regenerative capacity. These functions are mainly accomplished by the production of proteins that mediate gut cells functionality, as well as inducing specific gene expression of the host's gastrointestinal tract cells [38].

Nutrition and metabolism

The contribution of human metabolic processes essential for homeostasis is incredibly small compared to those provided by gut microbiota [38]. In the healthy gut microbiota, many of these processes offer tremendous benefits to the host, in particular in nutrient acquisition. The symbiotic partnership between the human host and its microbiota is essential for the maintenance of human health [38].

Dietary carbohydrates are the main nutrition source of the gut microbiota. The fermentation of indigestible oligosaccharides and undigested carbohydrates produces short-chain fatty acids (SCFAs) such as butyrate, propionate, and acetate, which can later be used as an energy source by the host [80, 81]. These SCFAs are not only important as an extra energy source but also for other tasks, such as preventing the accumulation of toxic byproducts, such as D-lactate [82], anti-cancer activity and role in gluconeogenesis [83].

The main organisms responsible for carbohydrate metabolism are the members of the genus Bacteroides, accomplishing this task by expressing several enzymes such as glycosyltransferases, glycoside hydrolases and polysaccharide lyases [84].

Gut microbiota has also been shown to have a positive impact on lipid metabolism by promoting lipoprotein lipase (responsible for the hydrolyzation of triglycerides in lipoproteins – control of triglyceridemia) activity in adipocytes [36]. It has also been shown that Bacteroides thetaoiotaomicron can increase the efficiency of lipid hydrolysis by up-regulating expression of a colipase required by pancreatic lipase for lipid digestion [85]. In conjunction with human proteinases and their own microbial proteinases and peptidases, gut microbiota is also capable of metabolizing dietary protein [36]. Protein fermentation capabilities depend on the region of the colon, the proximal colon is majorly saccharolytic, with protein fermentation increasing distally in the colon [83].

Another metabolic task of the gut microbiota lies in the synthesis of vitamin K and components of the B vitamin group. Members of the genus Bacteroides have been shown to be able to synthesize conjugated linoleic acid which is known for several properties: antidiabetic, antiatherogenic, antiobesogenic, hypolipidemic and immunomodulatory [86, 87, 88].

A healthy gut microbiota has also been shown to increase energy metabolism and to have systemic effects on host lipid metabolism, reinforcing its role as a metabolism modulator [89].

As previously mentioned gut microbiota is also involved in the breakdown of numerous dietary polyphenols. Some species within gut microbiota are also responsible for the conversion of primary bile acids (derived from cholesterol in the liver) into secondary bile acids (produced by bacteria), deoxycholic acid and lithocholic acid [90, 91]. Bile acids are important in the emulsification and digestion of dietary fats, however high levels of secondary bile acids within the intestinal lumen can be detrimental to human health since they can damage the intestinal epithelium, induce apoptosis and damage DNA. These factors are strong inductors of tumorigenesis [91, 92]. This further reinforces the idea that it is of extreme importance to maintain an equilibrium in the number of gut microorganisms.

Beyond the gut

A simple speculation in the past – "Death sits in the bowls (. . .)" [93], has now become quite evident. Unhealthy gut microbiota has been associated with several diseases, ranging from metabolic diseases such as obesity and diabetes, inflammatory bowel diseases and irritable bowel syndrome [36] to influencing the immune and nervous systems [94]. Recognizing the importance of a healthy gut microbiota and understanding the extent of its influence on human health will surely bring about a new perspective into preventive medicine.

2.4 Nutritional Genomics

Reproduction and procurement of "fuel" to ensure reproduction feasibility is the drive of all living organisms. This "fuel" comes in many forms and from many sources, but one fact remains: it is essential for the maintenance of life. Nutrition is, therefore, a field deeply connected with life itself and understanding it remains a task of both extreme importance and extreme complexity. Up to this date, nutritionists, dietitians, and other health practitioners have been mainly focused on the task of developing and applying general dietary guidelines that prevent disease and promote health. These guidelines do not take into account the inter-individual response variability to dietary intake.

The basis for Nutrition research has been, for several decades, longitudinal observation followed by the creation of simple nutrient-disease connections. There exists, however, a need to acquire a deeper understanding and thus gain the ability to knowledgeably influence the metabolic processes that make up the Human metabolism. A transition of focus from epidemiology and observational studies to molecular biology and genomics is therefore crucial. Nutritional Genomics was born from this necessity, along with: **1**. significant advances in the several different omics; **2**. the recognition that genetics play a vital role in response to diet, and **3**. that dietary compounds can be treated as dietary signals which modulate transcription factors and hence cellular metabolism.

Advances in bioinformatics techniques and systems biology have allowed us to decipher biological complexity using a systems approach. This more comprehensive understanding of how pieces work together to form a whole has made several biological fields, including nutrition genomics, extremely attractive fields academically and commercially. Through a combination of multidisciplinary tools and methodologies, in the future, it may become possible to recommend a genome-specific diet, preventing cardiovascular diseases, metabolic diseases or cancer, whilst promoting optimal health and ageing. Perhaps, we can finally move on from curing preventable diseases to avert their onset altogether.

Two main fields exist in nutritional genomics, nutrigenomics which studies how our diet influences DNA structure and gene expression, and nutrigenetics which studies how genetic polymorphisms influence the response to a specific diet, hence helping in the formulation of an optimal genome specific diet [1, 95]. Figure 6 shows how these two fields complement each other and work together to reach a similar goal, personalized nutrition.

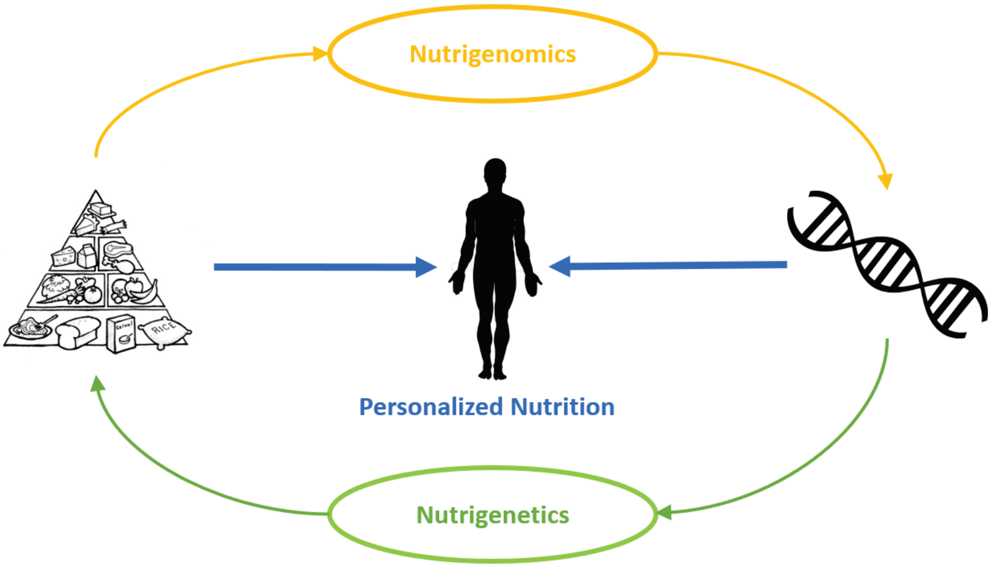


Figure 6: Nutrigenomics and nutrigenetics are distinct but work towards the same goal, adapted from [96]

In the context of this thesis, nutrigenomics is the more relevant field.

2.4.1 Nutrigenomics

Nutrigenomics is a highly complex field, still in its infancy. It utilizes the advanced omics techniques of systems biology to generate tremendous amounts of data whilst at the same time utilizing bioinformatics techniques to analyze such data. In practice, nutrigenomics focus on the use of omics tools to examine how nutrients influence metabolic pathways and homeostatic control [96].

The human body handles a large number of different nutrients and dietary substances, with highly varying concentrations. Each substance can have several targets with different affinities and specificities [97]. Unlike pharmacogenomics which is highly specific, nutrigenomics works with a complex mixture of substances (dietary intake) and studies how these modulate metabolic pathways. One possible approach to understand the metabolic function of nutrients is to treat dietary substances (it is imperative to understand the significance of using this nomenclature in detriment of the widely used "nutrient" since even non-essential substances / non-nutrients can be of metabolic consequence) as signaling molecules, transmitting and translating dietary signals in gene, protein and metabolite expression [1]. The main agents through which dietary substances influence gene expression are transcription factors [97]. Figure 7 depicts this approach and how dietary substances can be integrated into the several omics.

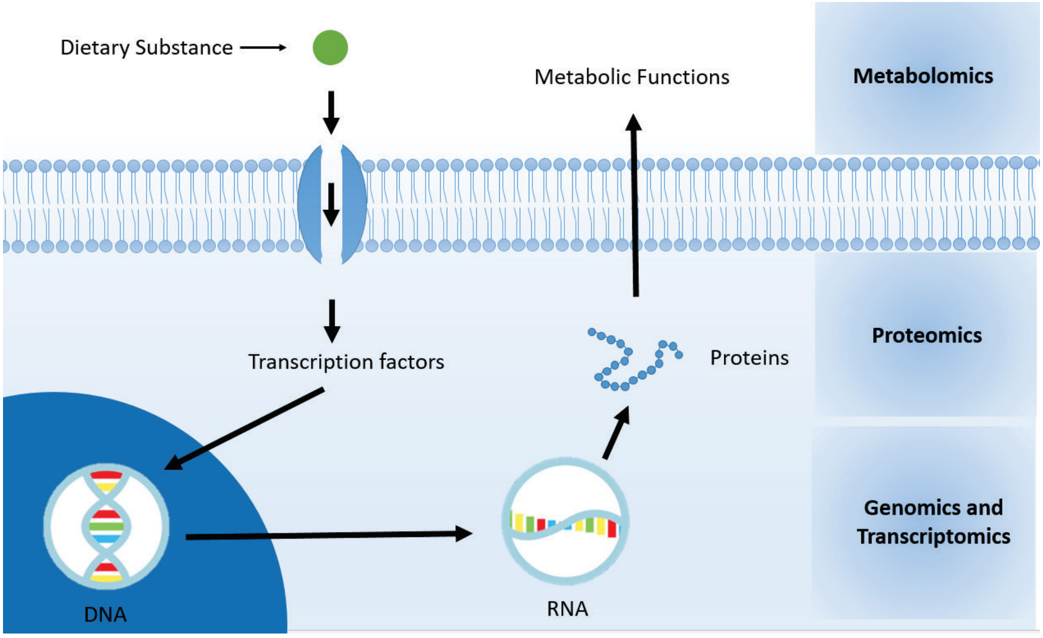


Figure 7: Dietary substances as signaling molecules, adapted from [98]. DNA and RNA image from [99]

As can be seen below in Figure 8, ideally, by combining the output of the various omics (genomics, transcriptomics, proteomics, and metabolomics) and focusing on common endpoints, it should be possible to eliminate the "noise" of each technique [96] and reach a conclusion.

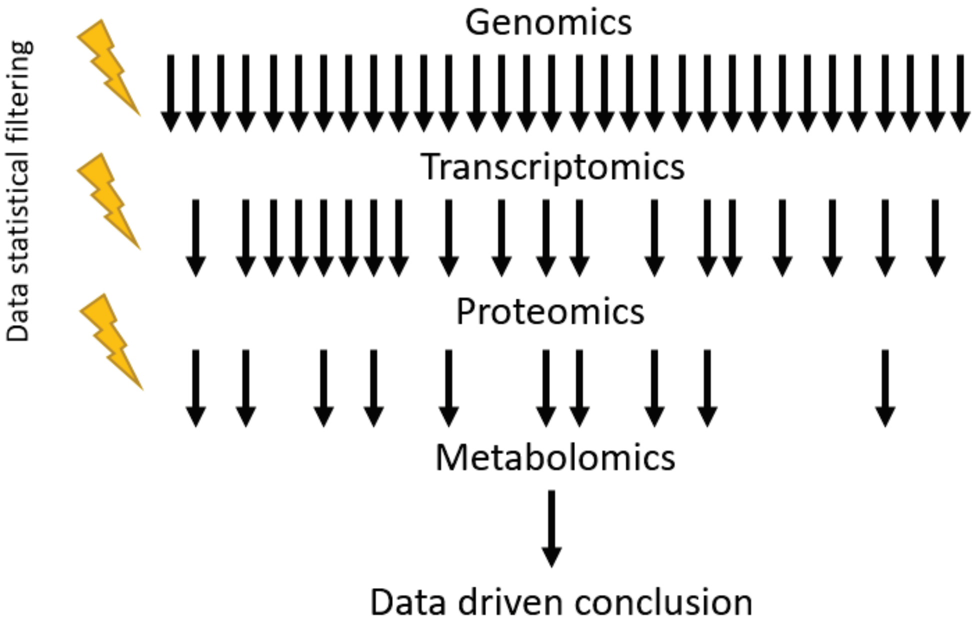


Figure 8: Approach to nutrigenomics research

Genomics is a well-established field with efficient and relatively cheap techniques. However, it cannot always represent what is actually happening *in vivo*, which is why the other fields, especially metabolomics, are better suited to this task. In any case, pooling the resources of these fields will ultimately lead to a better comprehension of the human metabolism. It is also crucial to create universal and replicable research guidelines and to properly document and identify relevant research in a tractable fashion.

2.4.2 Research in Nutrigenomics

Research in humans is desirable, but often arduous, complex and in some cases impossible. There are several problems inherent to human research:

- Problems of ethical nature;
- Sample collection- studying the effect of dietary intake on one or several

organs would require invasive tissue collection;

- Manipulation and control of dietary intake - it is extremely hard to supervise and control dietary intake in humans, a requirement in this field of research;
- Inadequate biomarkers – in *in vitro* certain biomarkers might be relevant, but that might not be the case *in vivo*, due to compensatory mechanisms [100].

Thus, an easier alternative is to apply the previous techniques and study simpler and more accessible organisms, like *in vitro* yeast and Caenorhabditis elegans or *in vivo* with the fruit fly Drosophila, which can serve as model systems [97]. Another possibility is to use *in silico* metabolic models to simulate and examine how dietary substances influence metabolic pathways. The latter topic will be discussed next.

2.5 Metabolic modelling

With an understanding of Nutrigenomics and the complexity of Human Nutrition in mind, it is clear that it would be of great benefit to gain a better understanding on how different combinations and amounts of nutrients affect the metabolism.

Metabolic network modeling allows the creation of generic or specific metabolic simulations within a mathematical in-silico framework. The core of this framework is the reconstruction of all or part of the metabolic and transport reaction network within an organism. As of the 20th of July 2018, besides the human metabolism reconstruction Recon, the Virtual Metabolic Human database [101] has gathered 818 reconstructions microbe individual reconstructions (668 different species). ”AGORA”, the collection of these reconstructions, [102] will later be discussed. The majority of these organisms belong to the bacteria domain, which is to be expected since bacteria are simpler organisms than eukaryotes and of high interest in scientific and industrial research.

This thesis will aim to provide a tool that aids in the expansion of these reconstructions, which in turn will allow the exploration and comprehension of the complexity underlying human [2] and gut microbes metabolism.

Metabolic models can be divided in dynamic models and stoichiometric models, the latter being also divided in models combined with measurements (metabolic flux analysis (MFA) and metabolic network analysis (MNA); and without measurements. Constraint-based modeling is a commonly used framework for metabolic simulations. Flux balance analysis is a frequently used algorithm for these simulations, but there are plenty other algorithms [103] in constraint-based modeling, but choosing one type over the other depends mostly on the resources available and purpose of the simulation.

2.5.1 Structure of metabolic reconstruction

Metabolic reconstruction's structure is based on the graph theory, metabolic networks are represented as graphs where the nodes represent metabolites and the edges represent reactions mediated by specific enzymes. The graph will be directed, with the origin node of each edge representing the substrate and the target node the product [104]. An example of this structure is shown in Figure 9.

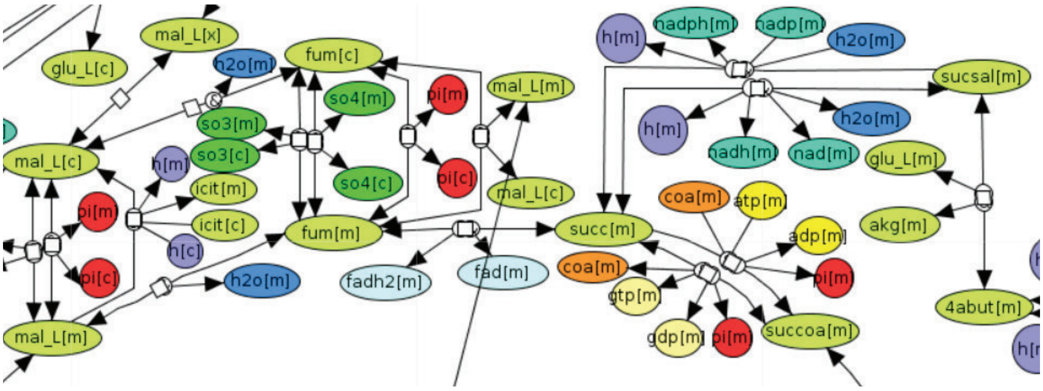


Figure 9: Recon Map, a visual representation of a genome-scale metabolic model [105]

2.5.2 Approaches to metabolic modelling

There are two distinct approaches in metabolic modeling, bottom-up modeling and top-down modeling. The bottom-up approach is the most commonly used and, to this date, the most developed. This method iteratively pieces simple parts together until a reconstruction with the desired complexity is created. On the other hand, the top-down approach filters out irrelevant data collected from singular or various omics research. The construction of models by this second approach can be considered a process of reverse engineering [106]. Figure 10 shows how these two very different metabolic modeling approaches reach similar goals.

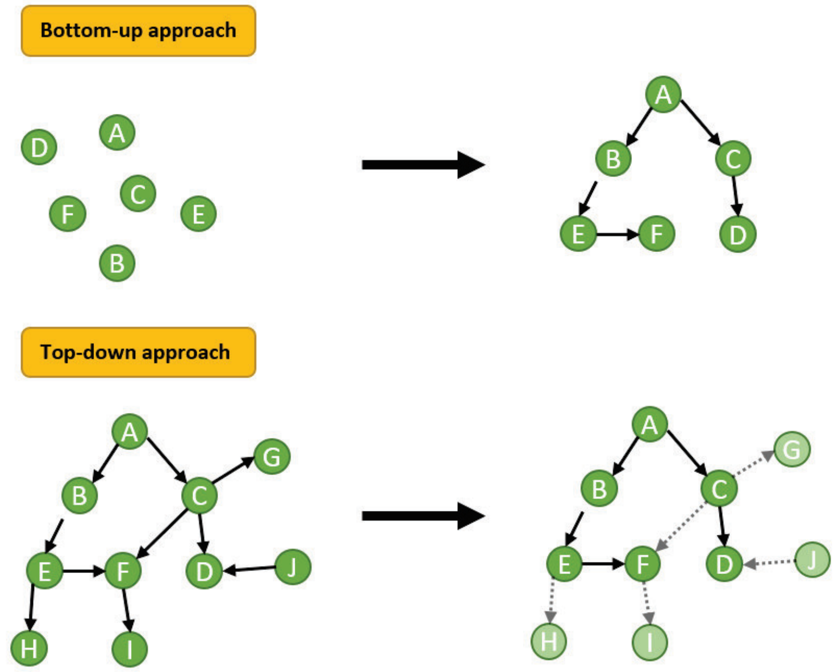


Figure 10: Approaches to metabolic modeling, adapted from [106]

The work developed in this thesis will primarily focus on a bottom-up approach to modelling.

2.5.3 Scaling the dimension of metabolic models

A bottom-up approach was initially used to generate kinetic models, which are relatively small and independent models. These are usually well characterized in stoichiometry and kinetics. These models are very useful when trying to manipulate a specific metabolic pathway. However,when scaling the model one can find that the enzymatic rates of all enzymes may not be available; and when they are available, they are usually derived from *in vitro* experiments, which most likely do not accurately represent physiological conditions. Their specificity and a poor reflection of *in vivo* conditions limit their usefulness when simulating the *in vivo* physiology of highly complex organisms.

Genome-Scale Metabolic Models (GSMMs) are stoichiometric models which aim to represent the metabolism of an organism. They contain thousands of reactions and consequently, due to a scarcity of data and proportion of the model, measurements or estimations of reaction rates are typically not available. Still, by maintaining stoichiometric structure, whilst sacrificing kinetic data, these models can cover thousands of metabolites, genes, and reactions, thus producing a highly informative metabolic network despite this kinetic data loss. Moreover, they rely on several assumptions, for example, the assumption of steady-state for internal metabolites. These assumptions are extremely useful as they allow feasible simulations. They are, however, assumptions and may prove erroneous when directly applied to physiological conditions.

2.5.4 Metabolic network simulation

Constraint-Based Model Reconstruction and Analysis (COBRA) is the most commonly used method for the quantitative prediction of cellular and multicellular biochemical networks with constraint-based modeling. COBRA provides a molecular mechanistic framework for the integrative analysis of experimental data and quantitative prediction of physicochemically and biochemically feasible phenotypic states [107]. To understand the mathematical process behind the creation of a metabolic model, an exemplary small model (Figure 11) will be presented and discussed.

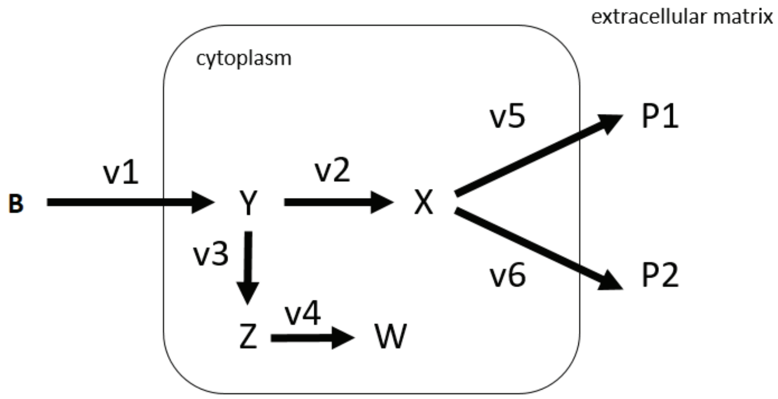


Figure 11: Example model

A very common approach to this kind of problems is Flux Balance Analysis (FBA) and will hence be the method described hereafter.

The model in Figure 11 has four internal metabolites Y, X, Z and W, three external metabolites B, $P1$ and $P2$ and 6 reaction fluxes, $v1$, $v2$, $v3$, $v4$, $v5$, and $v6$. This model can be defined by a stoichiometric matrix S where the rows denote metabolites and columns the reactions. The variation, over time, in the concentration of the X metabolite can then be determined by calculating the reaction rates/fluxes v. This is defined as a simple mathematical equation:

$$\frac{dX}{dt} = S \times v$$

Considering that metabolite X is dependent on the enzyme fluxes $v2$, $v5$ and $v6$, where $v2$ produces X and $v5$ and $v6$ consume it:

$$\frac{dX}{dt} = S \times v = v2 - v5 - v6$$

Since in larger scale models no data is available on the reaction fluxes, to make simulations on the concentration of X for any given moment in time, certain constraints need to be placed on the model's fluxes.

With constraint-based modelling, two major constraint groups exist. The first is the assumption of steady-state: a biological system, given an appropriate period of time, will reach a state where the production and consumption of all the internal metabolites are balanced. This constraint of balance translates into $S \times v = 0$ for mass and$\triangle E = 0$ for energy [108]. The second comprises the definition of bounds [108]:

• An enzyme/ transporter has an associated capacity. Hence, for each individual rate v, an upper limit α and lower limit β are established, such that: $\beta_j \leq v_j \leq \alpha_j$;

• Every given reaction is also defined by its directionality $d \leq v_j \leq +\infty$, where d will be 0 for irreversible reactions and $-\infty$ for reversible reactions;

• A biological system will also be bound by its solvent capacity, defined as $\Sigma_i c_i \leq c_{max}$.

Without constraints, no simulation would be possible since our solution space would be "infinite". By adding physiologically logical constraints, one can reduce the dimension of the solution space and hence outline an area of possible solution spaces for a set of given fluxes. For example, given the equation $\frac{dX}{dt} = v2 - v5 - v6$ for the concentration of metabolite X, a solution space (Figure 12) can be generated:

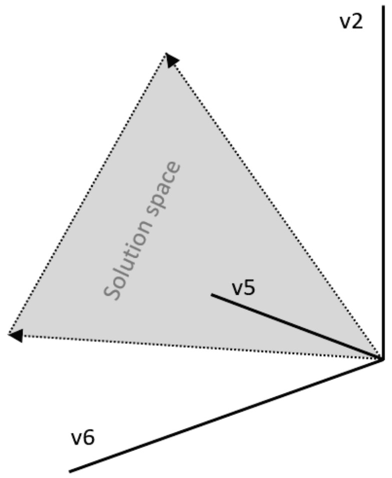


Figure 12: Solution space for metabolite X, adapted from [76]

Not only do these constraints reduce the solution search area, but they also restrain simulations that could otherwise produce biologically irrelevant solutions.

In summary, FBA is a mathematical algorithm used for simulating metabolism by treating this simulation as an optimization problem [**what˙is˙fba**]. As an optimization problem the goal is to quantify intracellular flux distribution and solves it by optimizing an objective function with linear programming [103]. For example, if one would like to optimize the production of metabolite X (objective function), by running the FBA method, it would find a possible optimal solution for our objective function.

Maximizing cellular growth and proliferation (commonly referred to as the maximization of biomass production) is one of the most common objectives of metabolic modelling. For example, a metabolic model has three fluxes ($v1$, $v2$, and $v3$), where one of them leads to cellular growth and proliferation ($v1$) and the other two deter cellular growth and proliferation.

In order to maximize biomass production, it would simply be a matter of defining an objective function which attempts to maximize $v1$ whilst minimizing $v2$ and $v3$.
This process is usually much more intricate, more so in highly complex metabolic models, but the idea remains the same. For other objective functions, other parameters may be optimized.

Overall, FBA works quite well when maximizing biomass production, there are nonetheless modified versions of FBA as well as other mathematical algorithms that, for other problems, might, or not, be able to find better solutions. Utilizing data from other omics may also be able to improve metabolic predictions, still, this is not guaranteed as various biological factors might lead to inconsistent results. The inclusion of metabolomics data, being an ”end of the line” omic, usually leads to good results [106]. In conclusion, there is no “best” methodology for better metabolic predictions, one has to approach the problem critically, always taking into consideration the organism and conditions one is working with.

2.5.5 Building a Genome-Scale Metabolic Model

There are already extensive protocols for the creation of metabolic models [108, 109]. Thiele and Palsson [108] provide an excellent overview of this procedure, which will be summarized below.

The **first** step in the reconstruction of a GSSM is based on the genome annotation of the model's organism and several biochemical databases, for example, KEGG [110, 111, 112], BRENDA [113, 114], and MetaCyc [115]. This first draft is usually generated automatically and is prone to errors. Pathway Tools [116], SEED [117] and Merlin [118] are three of the software tools that allow this automation.

In the **second** step, the previously generated draft will be reviewed and refined. During this manual curation, experts, supported by organism-specific literature, review the reconstruction for omissions, wrong assignments, gaps, and inconsistencies.

In the **third** step, the reconstruction is converted into a mathematical model where an objective function is defined and constraints are applied. A common tool used during this process is the COBRA toolbox [107].

The **fourth** step ensures network verification, evaluation, and validation. The model can also be tested against experimental high-throughput data. Amid other verifications, the model is tested for the synthesis of precursors required for biomass production, mass and charge balance, biomass growth rate etc. It is through the several time-consuming and arduous iterations of the second to fourth steps that the model is improved and refined, increasing its metabolic predictive capabilities.

2.5.6 Applications of GSMMs

While a GSMM requires plenty of time investment, it can serve a multitude of purposes [119, 120], which is why it has gained so much attention in the last years. Its construction is a costly process, both in time and resources, but extremely rewarding for future experiments. Some of the benefits are as follows:

- Validation of research, if one can compare literature results against a model, assuming the model has been proven accurate, inconsistencies in research can be detected and eliminated;
- Metabolic engineering for the production of desirable metabolic outputs;
- Prediction of metabolic adaptations and lethality of pharmaceutics, important for pathological organisms;

- Drug target identification for the study of the effect of drugs on human health, important in many fields such as oncology, cardiovascular and metabolic diseases, etc.

The possibilities are in fact endless. One field of great interest, is personalized healthcare, which will later be discussed.

2.6 Metabolic network reconstructions

Now that the topic of metabolic modeling has been introduced, some metabolic network reconstructions of consequence will be presented, in particular, Recon and AGORA.

2.6.1 Recon

Recon is a human genome-scale metabolism reconstruction first published in 2007, which has ever since been continuously changing and evolving. This model is a community effort and all versions of Recon are freely available. It was initially fleshed out by collecting and manually curating over 50 years' worth of literature on human biology and drafted into an *in silico* mathematical model using a bottom-up approach. It is the embodiment of decades of research neatly packed into a bioinformatics tool that computes acceptable network states under chemical and genetic constraints [121]. To note that Recon is a global reconstruction, and as such includes all human metabolic tasks, regardless of tissue type. In the following sections, the various versions of Recon will be chronologically presented and compared.

Recon 1

The first version of Recon, fittingly named "Recon 1", contained 1496 open reading frames (ORF), 2004 proteins, 2766 metabolites and 3111 metabolic and transport reactions. It was validated by simulating 288 known metabolic functions. This model was submitted to five iterative rounds of reconstruction and validation. It has numerous hierarchical levels of detail, specifically:

- Formulation of metabolites and reactions with known reaction stoichiometry, substrate/cofactor specificity, thermodynamics, conservation of mass and charge-based metabolite ionization states at pH 7.2;
- Compartmentalization of metabolites and their exchange between seven intracellular locations and extracellular space;
- Boolean description of gene-protein relationships;
- Confidence scores and literature references based on the available biological evidence for each gene, protein, and reaction.

Recon 2

Up until the creation of Recon 2, various research groups were working towards the creation of different human metabolism reconstructions, which led to the existence of numerous reconstructions with only partially overlapping content. Some cell type-specific models were also available (hepatocytes, macrophages and small intestinal enterocytes for example).

So, the community decided to pool their resources together and expand upon Recon 1, thus creating Recon 2, a community-driven global reconstruction of human metabolism [2].

With this update Recon, now contained 1789 enzyme-encoding genes, 7440 reactions and 2626 unique metabolites compartmentalized over eight intracellular locations. Recon 2 also included the mapping of known drug targets which should arrange for opportunities in the simulation of drug actions.

Overall, Recon 2 has shown to be a comprehensive metabolic tool with good predictive power, proving advantageous for future experiments in various fields, yet remaining open to further iterations of expansion and refinement.

Recon 2.2

In 2016 Recon 2.2 [122] was published, this version was built upon Recon 2 and several other intermediary versions (these provided better definition of transport proteins, wider deliberation on drug metabolism, improved carbon balancing, and some errors corrections). This version featured 5324 compartmentalized metabolites, 7785 reactions and 1675 genes (genes are now represented with HGNC [123] identifiers) and focused on charge balancing all reactions, producing a reconstruction that correctly predicts ATP flux on different carbon sources, giving biologically realistic results for simulations of growth and energy metabolism. This version was developed by a semi-automatic methodology which permits easier and faster future updates with improved traceability.

Recon 3D

In February 19th, 2018, Recon3D was published, bringing another update to the already comprehensive human metabolic reconstruction. It is available on the website Virtual Metabolic Human (VMH) [105]. This version is the first network reconstruction to include protein and metabolite structures along with as atom-atom mappings. It features 3288 open reading frames, 13543 metabolic reactions with 4140 unique metabolites, and 12890 protein structures [124]. This model will later be discussed in further detail.

ReconMap

ReconMap [125] is a comprehensive and manually curated map based on Recon that uses Google Maps application programming interface for interactive navigation. It is available at the Virtual Metabolic Human Database [101]. It allows "basic" Recon visualization as well as an overlaying of Recon-derived simulations and omics data.

2.6.2 AGORA

AGORA [102] is a resource of semi-automatically generated genome-scale metabolic reconstructions for 773 human gut microbes (205 genera and 605 species).
Conventional methods of research cannot identify the contribution of each gut bacterial species to the metabolic repertoire of the whole gut microbiome, neither can they evaluate the effects of gut microbiota composition on the host's metabolism. Metabolic modeling can overcome these problems by allowing the creation of models for each human gut microbe. This then allows the investigation of genotype-phenotype relationships, inter-microbial interactions, and host-microbe interactions. Since the tools used to automatically generate these models usually contain errors and are incomplete, AGORA was developed as a method that allows the refinement of one metabolic model to be propagated to other subsequent models. This is possible since all gut microbes share the same environmental conditions.

To add that the models were created using the previously described pipeline by Thiele and Palsson, where anaerobic growth was enabled since the human gut is generally anaerobic or micro-aerobic [126] (to note that inflammation of the digestive tract can increase oxygen levels [127]). The

AGORA reconstructions used a generalized microbial mass reaction, which nonetheless predicted average microbial doubling close to those reported in the literature [128]. As seen in Figure 13, compared to draft models, AGORA showed much higher sensitivity to carbon source, known fermentation products, and growth requirements as well as gene essentiality accuracy.

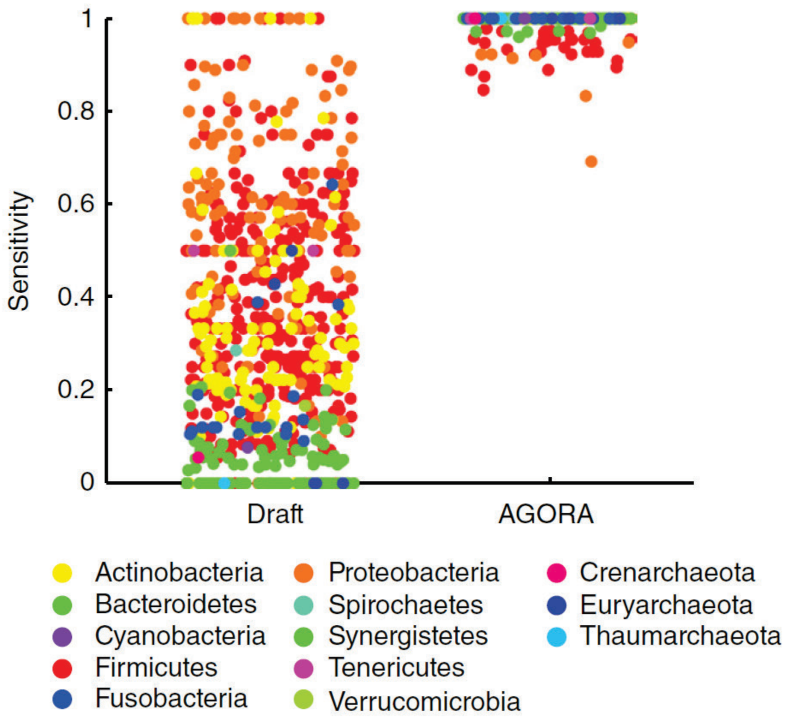


Figure 13: Sensitivity to carbon sources, known fermentation, reproduced from [102]

Metabolic pathway enrichment was found at different taxonomic levels, emphasizing the functional importance of different species for dietary substances degradation.

Both Recon and AGORA serve as a good example of how metabolic reconstructions can be used to predict and analyze metabolism, being by themselves very powerful in-silico tools but also serving as a baseline in formulating hypotheses for in-vivo or in-vitro research.

2.7 Personalized healthcare

Human health can be greatly improved by the many possibilities offered by metabolic modeling. It is a field still in its infancy, but with enormous potential.

In the field of xenobiotics, there is, for example, the case of Levodopa for the treatment of Parkinson's, in which the creation of a sIECs metabolic models led to the optimization of dietary recommendations in quantity, composition and intake times. This improved Levodopa efficiency and more importantly improved the patient's quality of life [129]. Indeed, with the aid of metabolic models, taking into account inter-individual response to xenobiotics, due to both microbiota and genetics, will surely play an important role in drug administration [130].

In this day and age, when dietary supplementation is trending and commercially attractive, using metabolic modeling to identify supplementation fallacies and opportunities is also a possbility [131].

Another recent paper reviewed the importance of patient-specific dietary interventions (quantity of supplemented fiber types) in the modulation of gut microbiota and the production of short-chain fatty acids in Crohn's disease patients [132].

In conclusion, metabolic modelling holds great promise in personalized healthcare. The future of this field not only relies on the creators of metabolic reconstructions themselves but also on those that are able to explore and derive biologically significant conclusions. As this is not an easily accessible field, perhaps it is time to expand health practitioners' educational curriculum to a degree at which they can offer their invaluable input.

2.8 Technical background

In the previous chapters, the state of the art for metabolic modelling and relevant fields of Nutrition were discussed. In the current chapter, more technical aspects will be presented.

2.8.1 Web mining

Web mining aims to find and extract relevant information or knowledge from a web page's hyperlink structure, page content, and/or usage data [133]. The 3 different categories of web mining are used within different applications, for this thesis web mining of page content will be the sole focus. Web mining follows four sequential steps [134] :

1. **Web data collection** – comparable to a user, the web mining program sends a request to a server which then sends back a response as the requested web page;

2. **Web data processing** - content is processed and cleaned, thus removing unwanted data. This step can be very time-consuming, fortunately there are tools that can simplify the pre-processing;

3. **Pattern discovery** - processed data is interpreted and patterns or meaningful knowledge are identified;

4. **Analysis of patterns** – not all patterns and knowledge are useful, during this step these are evaluated and the relevant ones are stored for later direct user interpretation, visualization, and any other tasks. In conclusion, web mining focuses on collecting and processing data, and then deriving knowledge out of this data. This thesis will focus on a specific field within web mining, **web scraping**.

2.8.2 Web scraping

Semantically, web scraping is the automated process of extracting data from websites. Like web mining, it involves collecting and extracting data, however it does not include deriving and processing knowledge from web pages. There is, of course, a tenuous line in these intermingled ideas and it is often the case that web scrapers borrow characteristics from the more wide-ranging web mining.

With the increased amount of information in public databases, it becomes increasingly difficult to manually extract and interpret all the data available. This is especially true for biological databases [135], which are exponentially growing in size, due to the uprising of more and better omics and other experimental technologies.

In a specific research field, a scientist may find that only a small part of the information gathered in these databases is actually relevant to their goals. Consequently, it is essential to gather information with precision and thoroughness.

This task is usually manual, researchers sift through websites and literature looking for the information they need, spending a large amount of time collecting data that they can then analyze.

Web scraping thus offers a solution to the time-consuming and often overwhelming task of collecting data.

Wrappers are the tools with which a web scraper extracts information structured data [133]. There are three main approaches to extraction:

- **Manual approach** – programmer analyses the webpage and its source code, finding patterns (HTML tags for example) and extracting relevant data.

- **Wrapper induction approach** – supervised approach in which manually labeled webpages are provided to establish extraction rules. Similarly, formatted pages can then be extracted using each format specific set of rules.

- **Automatic extraction approach** – unsupervised approach in which patterns are automatically identified.

A web application programming interface (**API**) is a useful web service provided by a growing number of websites. It is a web interface that allows third-party applications to access a specific server using HTTP protocol in an automated, efficient, and legal manner. Web APIs usually follow the Representational State Transfer (**REST**) [136] architectural style which provides easy integration within a third-party application. When available using this web service should be the preferred method for web scraping.

2.8.3 Legality of web scraping

The legality of web mining is a grey area, it is a still somewhat "new" field in terms of legislation and it is therefore not easy to determine which actions are permissible or not. In any case, being **polite** and acting **ethically** is essential to any developer using someone else's resources.

One thing to take into consideration when web scraping is to implement a polite scraper. A polite web scraper is one that avoids making too many requests to a specific webpage in order to not overload the host server. This not only ensures that the server remains usable by human users but also that the automated requests receive responses more quickly.

Another important point is to comply with the **Robot Exclusion Protocol** [137]. Each webpage normally has a robots.txt file stored within the webpage's root, which should be consulted prior to launching a web scraping bot. Compliance is not mandatory but it is advised to comply. Constraining data collection to public content only is also important to avoid any legal issues.

Chapter 3

Methodology and results

An exciting part of research is its volatile nature, a set goal can change or evolve in an instant. This was the case for this thesis. In this chapter the main work developed during the duration of this work will be described.

Due to the absence of polyphenols metabolism in Recon, the objective of this thesis was to start the manual annotation of their metabolism, so that someone could then continue this work.

3.1 Reaction annotation of dietary compounds

There are several biological databases that provide researchers with a plethora of information to answer a multitude of biological questions. In the case of extending a metabolic reconstruction, the main goal is to select a metabolite of interest and add it and its metabolism to the desired reconstruction. Some databases of interest include KEGG, Biocyc, Rhea [138], Reactome [139], Brenda, HMDB [140] – for human metabolism, and Drugbank [141] - for drug metabolism. The data on these databases is either collected from other databases, created through manual literature review or other methods.

Metabolism reconstruction is not an easy feat, it requires expertise and extensive database consulting and cross-checking. The previously mentioned databases are extremely comprehensive, offering information that may be irrelevant for certain tasks. Identifying relevant data also adds to the challenge.

Polyphenols metabolism is quite complex and involves the circulation and degradation of the primary ingested polyphenol and its many secondary derivatives. In addition, polyphenol diversity is astounding. With a necessity to define a starting point, a brief literature review pinpointed "quercetin" as a polyphenol with plenty of scientific literature and high interest in the scientific community.

After manually annotating quercetin reactions, with a better understanding of how reaction annotation works, it became clear that several steps of the annotation process could be extensively automated. Manual curation remains crucial, however a large portion of the work involves consulting databases and annotating relevant information , both of which can be automated.

3.2 Partial automation of metabolic reactions annotations

A general pipeline for the manual annotation of metabolic reactions can be seen in Figure 14 .

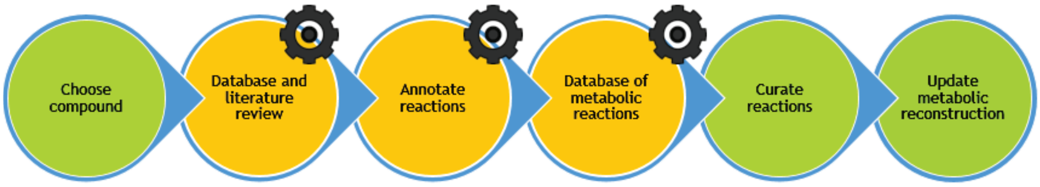

Choose compound
Database and literature review
Annotate reactions
Database of metabolic reactions
Curate reactions
Update metabolic reconstruction

Figure 14: Manual annotation of metabolic reactions

Not only is this process very time-consuming and requires high biochemical expertise, there's also no standardized methodology for the collection of metabolic reactions data.

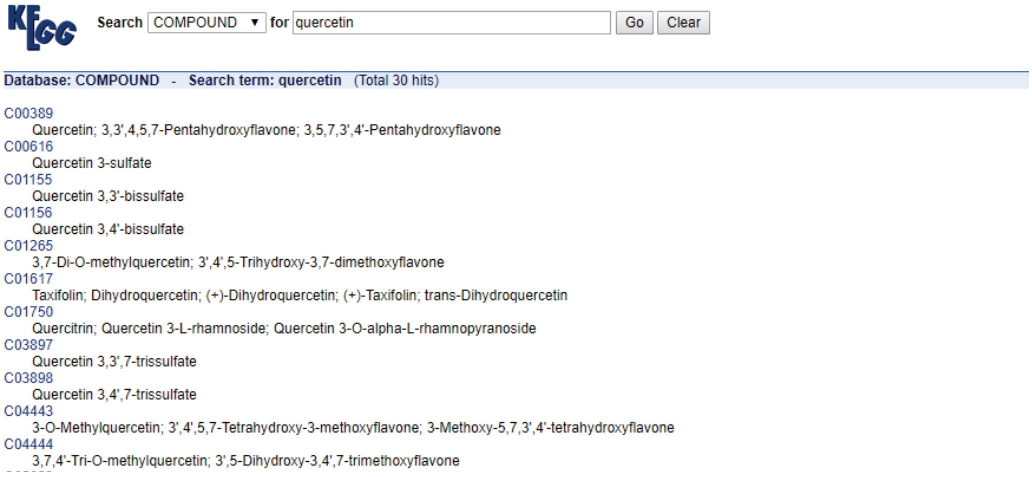
This realization led to the creation of a novel tool, the **Database Reaction Automated eXtraction** (**DRAX**), a database scrapper that aims at improving the efficiency at which new compounds are added to metabolic reconstructions. By standardizing data collection, DRAX also has the advantage to provide researchers with a universal reaction annotation starting point. The obvious benefit of this standardization is the increased ease at which a more novice researcher can start accurately annotating reactions. And, for more experienced researchers, providing a comprehensive starting point for an "unknown" compound can extensively reduce time spent on annotation while also aiding in comprehending the metabolic pathway for the "unknown" compound – a very important factor during manual curation of reactions.

3.3 Developing DRAX

3.3.1 DRAX overview

DRAX is in its essence a web scraper, it retrieves all relevant information from several databases and stores it in a spreadsheet. It was coded in *Python* and is dependent on several packages, the main ones being *Requests* and *Beautiful Soup*, the first retrieves data, the second formats it and provides tools for easier selection of the sought information.

The first step is the introduction of a word that the user wants to query, initially, it was targeted at polyphenols but after some refinement and addition of additional databases, the input of drugs and more general compounds also became possible. In this step, the user defines the principal compound of the query which will serve as the basis for all posterior searches.
This query word is then searched in several databases and the information for both the main compound and other compounds, which share the same root word is then retrieved. This step can be understood as an automation of query in a search bar.



Search COMPOUND for quercetin

Database: COMPOUND - Search term: quercetin (Total 30 hits)

```
C00389
    Quercetin; 3,3',4,5,7-Pentahydroxyflavone; 3,5,7,3',4'-Pentahydroxyflavone
C00616
    Quercetin 3-sulfate
C01155
    Quercetin 3,3'-bissulfate
C01156
    Quercetin 3,4'-bissulfate
C01265
    3,7-Di-O-methylquercetin; 3',4',5-Trihydroxy-3,7-dimethoxyflavone
C01617
    Taxifolin; Dihydroquercetin; (+)-Dihydroquercetin; (+)-Taxifolin; trans-Dihydroquercetin
C01750
    Quercitrin; Quercetin 3-L-rhamnoside; Quercetin 3-O-alpha-L-rhamnopyranoside
C03897
    Quercetin 3,3',7-trissulfate
C03898
    Quercetin 3,4',7-trissulfate
C04443
    3-O-Methylquercetin; 3',4',5,7-Tetrahydroxy-3-methoxyflavone; 3-Methoxy-5,7,3',4'-tetrahydroxyflavone
C04444
    3,7,4'-Tri-O-methylquercetin; 3',5-Dihydroxy-3,4',7-trimethoxyflavone
```

Figure 15: KEGG example query, reproduced from [142]

DRAX applies the same methodology as KEGG, which, with the query "quercetin", retrieves both quercetin and many of its derivatives. DRAX however, does so for several databases: KEGG, Biocyc, HMDB, Brenda, and Drugbank. All the information for these compounds is then retrieved and the data from several databases is intersected and joined together to create an initial list of compounds that DRAX has to search reactions for.

The next step is the secondary and more intensive database search, where DRAX crawls around multiple web pages to retrieve information about reactions and enzymes.

At the end of this extensive search, all information is neatly stored in a spreadsheet. Besides the reactions themselves, exported data includes compounds identifiers, literature, organisms in which either the reaction or enzyme are present, and much more information that the user might find relevant during the final manual curation of reactions.

3.3.2 DRAX technical details and algorithms

DRAX technical requirements

The requirements for the development of DRAX are mostly related to software and tools and not on physical material.

DRAX was implemented in Python 3.6. It is also reliant on several packages, some of which are present in the standard Python library. Packages not included in the Python's standard library are listed below, along with a brief explanation on why they are required:

- *BeautifulSoup* [143];
- *Pandas* [144];
- *Numpy* [145];

- *Requests* [146];
- *Selenium* [147].

Beautiful soup and *ElementTree XML* are two very useful packages for formatting and fetching information within HTML and XML files respectively, which are previously retrieved by packages like *requests* or *selenium.* The last one is used to overcome some restrictions found during some of the query requests. *Multiprocessing* permits multithreading databases queries so that the query becomes much faster and subprocess allows using a command line based tool like cxcalc in Python. Multithreading significantly increased DRAX performance but there is still room for improvement.
Keep in mind that DRAX also requires the installation of Marvin by ChemAxon [148] which contains cxcalc, the tool required for the correct calculation of compounds protonation state. ChemAxon provides a license for academic use.

DRAX compliance to Robot Exclusion Protocol

In order to develop a polite web scraper, one must ensure it complies with each web pages rules. This can be evaluated by accessing the robots.txt file stored in the root page of each website. Out of the several websites queried only Biocyc prohibits automated access through the use of robots. Before publication of DRAX, Biocyc will be contacted to request authorized access to their data.

DRAX robustness

A common implementation issue in web mining is the lack of robustness. Unresponsive webpages and unexpected download outputs can quickly crash a web scraper.

In order to resolve this a special connection function "*try_until_catch*" was implemented. This function keeps requesting a valid response from any website until it either gets a valid response or a certain amount of requests is reached. This ensures DRAX seeks out all information while not crashing in invalid or non-existent webpages, and temporary server connection errors. By default, connection retries are made with a 5 second interval.

Another common issue is overcoming errors in databases, which are fairly common in databases that rely on human data insertion. Part of the challenge in creating a robust web scraper is also fixing exceptional erroneous outputs without changing correctly inserted data. As these are hard to predict, a big portion of this process is continuous trial and error.

Query method

As previously mentioned DRAX allows the user to input a user query which will then be searched through several databases. This is a simple interpretation since the actual process can be executed in different ways, depending on the user-input:

1. Query with compound IDs;
2. Query with the compound name;
3. Query with a list of compounds.

In the first method, the user inputs one or more IDs from different databases and DRAX will then try to find the same compound in other databases and start the main query. In the second method, the compound name is searched around several databases. This method is quite good to find derivatives of a compound, since it uses the already implemented search function of each database. The third method is a simple iteration of the second method. Overall, the first method offers a more specific search while the second and third offer a thorough database search by using the root word the user inputs.

DRAX internal structure

To allow for better organization of the code, DRAX was split into several py modules. The Figure 16 illustrates the central modules and their workflow.

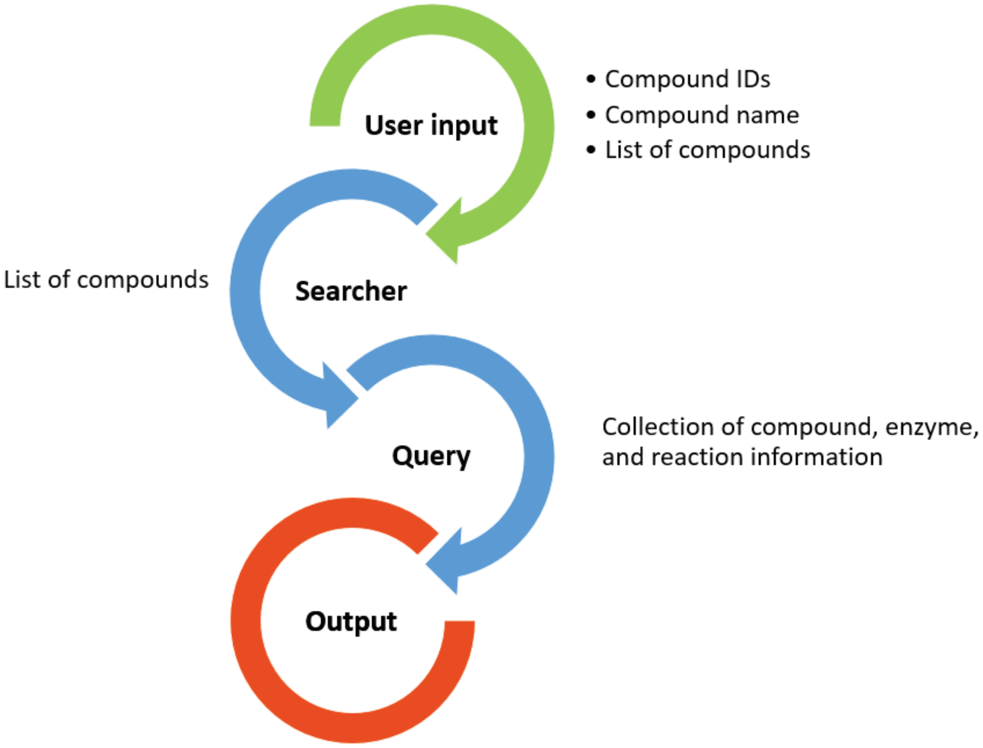


Figure 16: DRAX workflow

Each module is explained below:

• **Util.py**- a general utitilities module which stores non-specific functions that may be used across the different modules;

• **Fetcher.py** and **Parser.py** - these modules make requests and parse fetched information, respectively. In order to control the amount and frequency of requests done in each database, requests are done through a database-specific Fetcher class instance. This ensures that each database has an unique control point, which in turn results in a more polite and robust web scraper. To ensure politeness, time between connections is dynamically limited based on the latency of each database;

- **__init__.py** - serves as the front-end for user interaction. The current implementation relies on a console-based interaction, in the future this may be developed further;
- **Searcher.py** - provides methods for finding specific compounds and compound derivatives by querying several databases. It uses the built-in search method of each database and retrieves several compounds which, by default, contain the user query. It then intersects all collected information so that compounds are not repeated. It can either retrieve all of the search hits (derivatives search) or a singular hit (compound search). This class can work as a standalone tool;
- **Derivatives.py** - this class inherits from the Searcher class and is responsible for the integration of the previous class into the main Query.py. It provides the main Query.py with the the derivatives of a specific root word;
- **Query.py** - this is the main class, which handles most of the searching. This is the class that fetches information from databases, parses it and joins it together into a spreadsheet. This is therefore the central class which contains most of the algorithms for web scraping, parsing and information intersection;
- **Compound.py** - this class stores information regarding the compounds, providing also several methods for retrieving IDs as well as providing integration with VMH's compound database;
- **SMILES.py** - this module provides functions for fetching the SMILES of each compound from several databases (ChEMBL, CheBI, PDB, Pubchem and Food Database);
- **Identifier.py** - this class stores all possible IDs of a specific compound.

It acts mostly as an ID counter which updates the database specific compound ID as the most frequent ID in a list of several IDs found for that same compound;

- **Synonyms.py** - this class is similar to the Identifier class, providing somewhat similar methods for counting compound synonyms;
- **bash_cxcalc.py** - provides integration of Marvin's cxcalc tool and contains all the methods for the calculation of the chemical formula of a correctly protonated compound;
- **Mass_Balancing.py** - stores functions regarding the calculation of mass balance of a given reaction;
- **Reaction.py** - this module stores several types of information regarding the reactions, such as the enzyme instance that catalyzes the reaction, the database where this reaction was fetched from, the mass balance of the reaction, and the reaction in three different formats - reaction as a string, reaction as a string with compound IDs, and reaction with compound instances;
- **Enzyme.py** - this class stores information regarding the enzymes and provides methods for converting enzyme IDs into the enzyme commission number.

Access to databases and data retrieval

DRAX uses two ways to access database information. Whenever a database provides a programmatic access web service, DRAX makes use of it, otherwise, it will fetch the HTML file of the required webpage and parse it. The databases that offer programmatic access include:

- PubChem
- ChEMBL

- ChEBI
- KEGG
- Biocyc
- VMH

This feature improves the speed at which web scraping is done as it provides faster connection speed and more straightforward data parsing (especially since most of them can output data in an easy-to-parse format), being obviously the preferred method.

Information retrieval

DRAX mainly uses Beautiful Soup to examine data from HTML and smaller XML files.

Larger XML files are partitioned into small blocks of structured data, each block is iteratively read, relevant data is extracted and the block is then deleted in order to preserve memory. DRAX does not handle free text and always relies on a website-specific tag structure in order to find relevant information. At the current DRAX implementation, only Biocyc benefits from this type of parsing, however in the future this method will be applied to other databases.

Since a considerably low amount of databases were used and these had different format structures, the rules of extraction for the wrapper were manually defined.

Also to note that the same compound may have different IDs in different reactions, which may cause issues in compound matching during reaction fetching. This was solved by adding a new class *"Identifier"* which will set a compound's ID based on the most frequent compound-specific ID, while still storing low-frequency IDs, and consequently allow compound matching with more or less frequent IDs.

Selection of organisms

Since metabolic reactions are organism-specific, DRAX allows the user to provide a list of organisms that they want reactions for. Organisms in databases are usually reported in the genus and species level, however, some databases also include the specific strain. Indeed some reactions might be strain-specific, however since the amount of information is not uniform at this level, DRAX only selects at genus and species level.

Reactions not observed in the user-provided list of organisms can be catalyzed by enzymes that these organisms may possess. If that is the case, the organism may also be able to catalyze this reaction, but it is possible that there is a gap in scientific research. Consequently, it is possible to include these reactions as they may help in metabolic pathways gap-filling.

Selection of organisms is crucial, as it is known that gut microbiota plays a large role in human metabolism. This feature allows the integration of DRAX not only into human metabolic reconstructions such as Recon but also microbe-specific or microbe community reconstructions as AGORA. As the Molecular Systems Physiology Group works in both these areas, this flexibility is a simple and straightforward solution to incorporate DRAX into both projects.

Reaction scoring

DRAX provides a reaction scoring system, where the score ranges from 0-1 and takes into account:

- Presence and quantity of literature references;
- Whether the reaction was proven experimentally or predicted;

- Whether the reaction is mass balanced (an indicator that it was most likely already curated) - unbalance in the number of protons will not be included, to avoid penalizing reactions due to the protonation state of compounds.
- Whether the reaction is known to be catalyzed by the organism or it is only known whether the organism has the enzyme responsible for catalyzing this reaction.

Additional points may be considered in the future in an effort to refine the scoring system. To note that each point may have an associated weight, for example, the frequency of the reaction in the scraped databases will have a low weight in order to compensate for replication of information through data sharing across databases.

$$Reaction\ score = \frac{B \times Bw + C \times Cw + D \times Dw + E \times Ew}{N}$$

Where each variable is:

- B = Existence (1) or absence (0) of literature (**binary** variable)
- Bw = Weight % for variable "Existence or absence of literature"
- C = Experimentally proven (1) or predicted reaction (0) (**binary** variable)
- Cw = Weight % for variable "Predicted or experimentally proven reaction"
- D = Whether reaction is mass balanced (1) or not (0) (**binary** variable)
- Dw = Weight % for variable "Whether reaction is mass balanced or not"
- E = Whether the reaction is known to be catalyzed by the organism (1) or it is only known that it possesses the enzyme responsible for the reaction

(0) (**binary** variable)

- Ew = Weight % for the variable "Whether the reaction is known to be catalyzed by the organism or it is only known that it possesses the enzyme responsible for the reaction"
- N = Number of scoring variables

Since not all databases have the information regarding being experimentally proven or predicted, B is not always included, consequently N may be either 3 or 4. Weight % can be user-defined.

This functionality takes advantage of compound matching and intersection across the databases. In this manner, DRAX not only retrieves each database specific reaction, but also provides the user with a list of unique reactions along with their respective scores. In conclusion, this functionality will advise the user on the confidence of each scraped reaction.

Connecting compounds

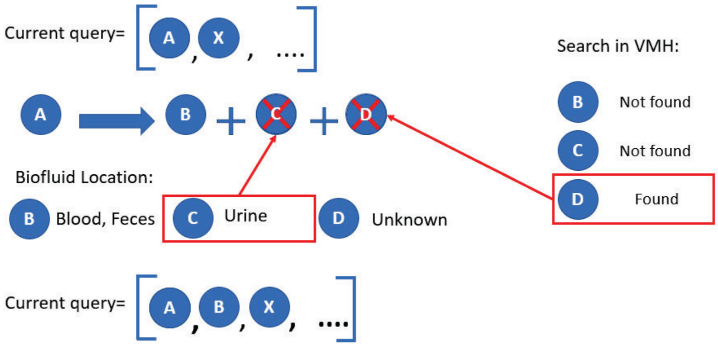
DRAX initial query ensures that compounds that share the root word of the user query are also retrieved. This is a sound approach for polyphenols as these are usually named around a certain root word. However, for other compounds, this is often not the case, which is why the "Connecting compounds" algorithm was developed. This algorithm lengthens the query beyond compounds found during the initial query.

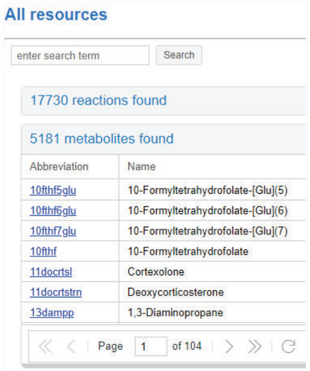
While retrieving a reaction, DRAX extends the query for all the metabolites in this reaction (with some exceptions). This allows a thorough reaction pathway reconstruction of the compound and its derivatives. While metabolism can be clustered into different functional groups, most metabolites are physiologically intertwined. Should the query extension be left unconstrained, DRAX would "endlessly" add new metabolites to the query.

Thus, the “Connecting compounds” algorithm has to follow certain restrictions. To ensure that the query doesn’t keep adding metabolites a three-step algorithm was implemented:

1. **Presence in VMH** - During reaction retrieval, each compound in a reaction is searched for in VMH’s metabolite database (Figure 17), if it is not found the compound is added to the query of compounds that DRAX has to search for. Since we are mapping new reactions, if it’s present in VMH there’s a high chance that the compound is a reaction cofactor or its reactions have already been mapped. This, of course, is prone to errors, since a metabolite might be present in VMH but its pathway may be incomplete or not even mapped. To extend upon these flaws further conditions were established;

2. **Biofluid location** - If a metabolite is located in either urine, feces or both then this metabolite is meant to be excreted. Hence, it is considered a “terminal” metabolite as no consequent pathway will originate from it. Consequently, these are also not added to the list of compounds to search for (Figure 17). However, biofluid location is not always available;




All resources

enter search term | Search

17730 reactions found

5181 metabolites found

Abbreviation	Name
10fthf5glu	10-Formyltetrahydrofolate-[Glu](5)
10fthf6glu	10-Formyltetrahydrofolate-[Glu](6)
10fthf7glu	10-Formyltetrahydrofolate-[Glu](7)
10fthf	10-Formyltetrahydrofolate
11docrtsl	Cortexolone
11docrtstrn	Deoxycorticosterone
13dampp	1,3-Diaminopropane

Page 1 of 104

Figure 17: Connecting compounds algorithm - Tree Growth, presence in VMH and biofluid location

3. **Maximal depth** - To limit the number of compounds added to the query, the query can be constrained by a maximum depth. The simplest

interpretation (Figure 18) of this idea is to think of a hierarchical tree made of nodes, where each node is a compound and as the tree grows so does the depth increase. Until a user-defined depth (default is a depth of 3) the tree can grow by adding a reaction's metabolites to the query as new nodes. Beyond that point, the tree cannot grow and therefore metabolites are not added anymore. The depth of each compound is defined upon the creation of the respective compound instance.

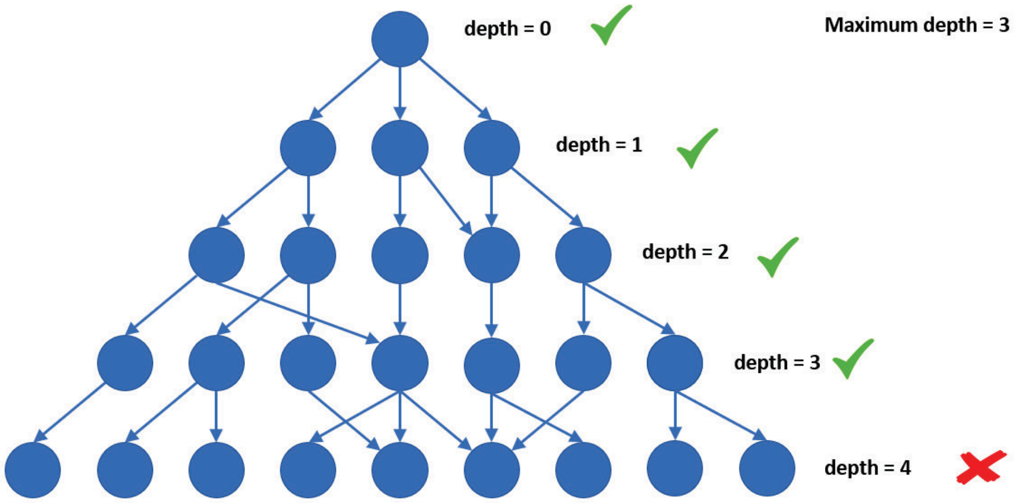


Figure 18: Connecting compounds algorithm - Tree Growth, maximal depth restriction

In a more practical sense, this means that DRAX will find reactions for a compound and add new compounds as long as the compounds remain within the maximal depth:

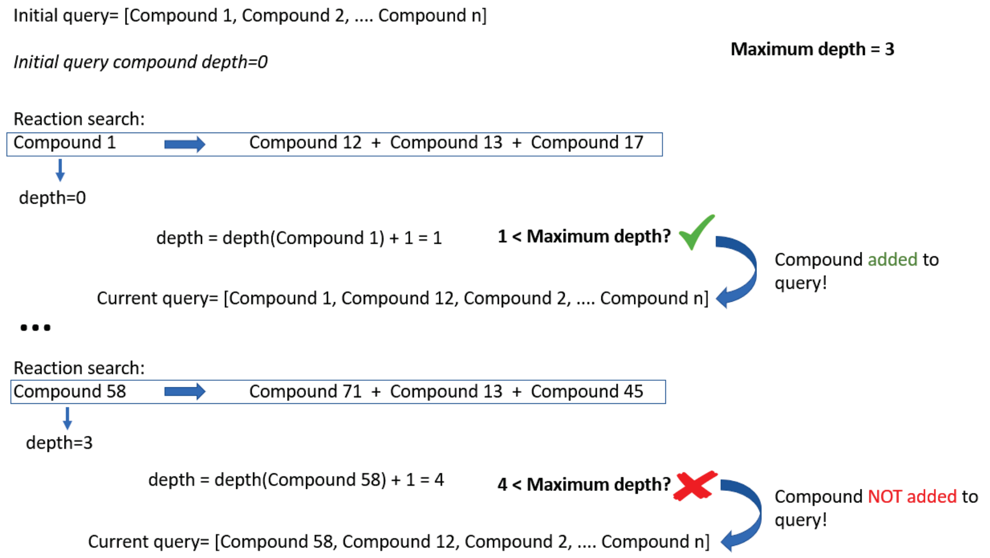

Initial query= [Compound 1, Compound 2, ... Compound n]
Initial query compound depth=0
Maximum depth = 3
Reaction search:
Compound 1
Compound 12 + Compound 13 + Compound 17
depth=0
depth = depth(Compound 1) + 1 = 1
1 < Maximum depth?
Compound added to query!
Current query= [Compound 1, Compound 12, Compound 2, ... Compound n]
...
Reaction search:
Compound 58
Compound 71 + Compound 13 + Compound 45
depth=3
depth = depth(Compound 58) + 1 = 4
4 < Maximum depth?
Compound NOT added to query!
Current query= [Compound 58, Compound 12, Compound 2, ... Compound n]

Figure 19: Connecting compounds - Example

The combination of these three conditions ensures not only that DRAX has a stopping point but it also takes advantage of the information produced by the host group and the various databases it fetches information from.

If possible, it would be interesting to extend this algorithm, so that it better removes common biological cofactors that, in most cases, won't be of interest to the researcher.

Protonation

DRAX also calculates the mass balance of reactions by adding up the chemical formula of the substrates and products. A simple calculation with the fetched chemical formulas is not accurate though, as most organic molecules can gain or lose protons/hydrogens in aqueous solutions. Therefore, to better represent the mass balance in physiological conditions it is necessary to take the protonation state into account.

The ratio of a compound's ionized and neutral forms depend on the pH, temperature and ion activity of the bulk phase. The ionization constant K_a is calculated with the following formula:

$$K_a = \frac{[A^-][H^+]}{[AH]}$$

Which is the same as saying that $K_a = (conjugated\,base \div conjugated\,acid) \times [H^+]$.
Since this value is usually extremely low or extremely high, it is common to use the logarithm of this constant, pKa. The pH in the various locations of the human body highly fluctuates, for example, through the initial parts of the digestive tract it is more acidic, while in the lower parts it is more basic. Despite this variation, it is important to note that the human body is quite sensitive to pH and thus it possesses well-developed mechanisms to preserve homeostasis. The physiological pH of the human body stands at 7.2 with minor variation. This value will be used as the default for the calculation of the major compound form during mass balancing.

The protonation state - the amount of protons/hydrogens of a given compound; depends on its structure and since it's not the purpose of this thesis to develop such a tool, a commonly used calculator was integrated into DRAX. Several tools are available but Marvin, from ChemAxon, was chosen as it provides a free academic license. It can also be more easily implemented since it can be also used as a command line-based tool – "cxcalc". For a more thorough explanation of how pKa is calculated, please refer to ChemAxon's documentation [149].

DRAX automatically checks whether cxcalc and Marvin's license are available. If one of these is not available, DRAX requires user confirmation to continue the query and will then calculate mass balance without taking compounds protonation state into account.

Data persistence

A query for a more complex pathway can take be quite time consuming, therefore it is important to allow for data persistence should the script crash due to some unexpected event. As of this moment, these crashes are quite rare as DRAX has been thoroughly tested, but in any case, web scrapping a large amount of different data may result in an unforeseeable event, which is why implementing data saving checkpoints is important. These can even be connection errors due to an irresponsive database. Otherwise, hours of web scrapping can be lost in an instant. With this in mind, two save checkpoints were implemented in DRAX:

1. **Metabolite checkpoint**- After the initial query to find derivatives is finished, DRAX will add these compounds to a file which contains information of all the compounds found during the current and past queries. Since this file will persist through different DRAX runs, it will iteratively increase in size as DRAX finds information for more and more compounds. In the end, this will allow DRAX to make fewer database requests the more times it runs;

2. **Query checkpoint**- this checkpoint occurs when DRAX saves all the information found into a spreadsheet. DRAX looks for the reactions of a certain compound and saves it into a spreadsheet, repeating this process repeatedly until the list of compounds to search for is empty. This list of compounds that DRAX has to search for is also saved into a different file as it can keep changing in each iteration – due to the connecting compounds functionality explained above.

The current implementation is only used as a failsafe, for this reason, it uses a quite simple saving system, it exports part of the data to a .txt file. A better system , using the Django framework is currently under-development.

DRAX file exportation

DRAX's output is exported into a .xlsx spreadsheet, which can be read by several spreadsheet management tools. After each compound query, the output file is either created (if it is the first query) or updated. DRAX is also able to export a txt file with a list of unique reactions along with their database internal reaction IDs and reaction confidence score. This confidence score is calculated with the reaction scoring formula previously defined.

Chapter 4

Results and discussion

In this chapter, the main outcomes of this work, along with practical examples, will be discussed.

4.1 DRAX usage and output

DRAX is, at the moment, a shell-based tool, which requires the user to input several arguments. These include:

- User query - either a compound, a list of compounds or a compound identifier/s;
- List of organisms (can be empty);
- VMH compound spreadsheet;
- Marvin's installation path;
- pH - for calculation of protonation state;
- Whether *connect metabolites* algorithm should be active or not;
- Max depth of the connect metabolites algorithm.

With the __init.py__ module this is handled in the following way:

```
Would you like to search for a compound name or ID?
Press 1 for name, 2 for ID. Default value is ”1”.
1
Enter the compound you want to query. You can also input the location of a new-line separated txt list
of compounds.
caffeine
Input a file with a list of organisms you want reactions for, one organism per line.
If you want to include all organisms, please leave blank.
list_desired_organisms.txt
Please provide the name of the VMH metabolites csv spreadsheet
metabolites_VMH.csv
Please provide the location cxcalc executable (within Marbin’s installation folder).
Leave blank if you do not want to take protonation into account.
C:/Program Files/ChemAxon/MarvinSuite/bin
Please provide the pH for the calculation of protonation state. Default is 7.2.
7.5
Would you like DRAX to include tree-like search?, 1 for yes, 2 for no. Default value is ”no”.
1
Input the depth for the search, high values will result in a longer query. Default value is ”3”.
5
Single query will now start!
All check-ups passed! Query will now start
```

DRAX exports all the information found in a single Excel spreadsheet, thus providing an easy-to-read standardized output. The current implementation of DRAX includes all the information in Table 3, which should aid in manual curation of reactions.

Table 3: Data exported to the output file

Compound information	Reaction information	Enzyme information
Name	**Database**	**Enzyme ID**
Synonyms	Reaction DB ID	Enzyme names
ID VMH	Mass Balance	Organisms with enzyme
ID Food Database	Reaction	Enzyme details
ID knapsack	Reaction with IDs	Enzyme Status
ID ChemSpider	Reaction with VMH IDs	Enzyme Comments
ID KEGG	Reaction organisms	Enzyme summary
ID Biocyc	Reaction references	Enzyme References
ID Bigg	Pathways with reaction	Enzyme Class
ID Metlin	Reaction comments	
ID PubChem		
ID Chebi		
ID HMDB		
ID Brenda		
ID DrugBank		
ID Phenol Explorer		
ID PDB		
InChI Key		
Location		

Note that the information regarding the compound includes VMH's compound abbreviations since DRAX also takes advantage of this resource. The remaining information may also be useful to the end-user.

The reaction information includes the database the reaction was extracted from, the database's internal reaction ID, the mass balance calculation, and the reaction itself as a simple string, with database internal compound IDs and VMH's internal compound IDs (these are often not available when searching for compounds not annotated in Recon). Reaction organisms provide the organisms in which the reaction was observed experimentally or predicted. The remaining information may help the end-user in "categorizing" the reaction and successfully curating it.

Reaction enzyme information is also included. The enzyme ID will correspond to either the database internal enzyme ID or the standard Enzyme Commission (EC) number. The EC number can sometimes be extracted directly from the queried database or, given that enough information is available, fetched from UniProt.

DRAX also exports a list of unique reactions, along with their confidence score (formula previously shown) and database internal reaction IDs. This will also aid the end-user in understanding how thoroughly the several queried databases have curated each reaction and whether there is enough confidence to add these reactions to the metabolic reconstruction.

With all this information available, researchers should be able to speed up the addition of new reactions to metabolic reconstructions.

4.2 DRAX testing

This work has the main objective to provide a starting point in reaction annotation, thus its output will never be, by itself, the last step of annotation. However, it is no less essential to control DRAX for the quality of its web scraping. This will ultimately provide some confidence in its usage. DRAX was evaluated with two different tests:

- Collection of reactions for new compounds;
- Comparison between DRAX and Recon.

These tests provide information regarding the depth of the search, efficiency gains and the quality of information specific to each reaction. Measuring how much manual curation is required afterward can also aid in assessing the usefulness of this tool. The members of LCSB's Molecular Systems Physiology group also aided in testing DRAX and their feedback was implemented into it. When a final version of DRAX is published, more researchers will hopefully provide additional feedback.

4.2.1 Collection of reactions for new compounds

As an example, the polyphenol “quercetin” was queried. In this case, the query method was “Query with the compound name”, which was previously described. Across the several databases, 122 compounds were retrieved, which DRAX will then find reactions for. Reactions were selected for both *Homo sapiens* and a list of gut microbes (list extracted from VMH’s database). For the initial query and its derivatives, 92 unique reactions were found. Out of these 92 reactions, 17 had unknown products.

To note that out of the 122 compounds found during the initial query, reactions occuring in the selected organisms were only found for 10 of these compounds. This is both due to the fact that many of the initial compounds were derivatives formed during the biosynthesis of the polyphenols (specific to plants) and also to the fact that polyphenol degradation is a quite niche subject, with room for new research. Still, DRAX was able to find reactions for quercetin and some other important quercetin methylated and glucoronidated conjugates such as rutin, isoquercetin, isorhamnetin and taxifolin.

A maximum score of 0.75/1 was obtained across 14 reactions, in order to show how DRAX can be quickly implemented in metabolism reconstruction, some of these will be discussed:

- From HMDB, the reaction, with ID ”4643”, is catalyzed by Pirin, a human protein known to be involved in quercetin degradation (*in vitro* evidence) [150]

- From Biocyc, the reaction with ID ”QUERCETIN-23-DIOXYGENASE-RXN”, is catalyzed by the enzyme quercetin 2,3-dioxygenase, a Pirin homolog.*Escherichia coli* is known to have the gene yhhW, which is responsible for the encoding of this enzyme.

- From Brenda, the reaction with ID 406540, is relative to the inhibitory effect of flavonoids against myeloperodidase, which may be implicated in the oxidation of protein/lipoprotein during the development of cardiovascular diseases [151].

- From Brenda, the reactions with IDs 322246, 378159, 402156, catalyzed by human glucuronosyltransferase , an enzyme responsible for the glucuronidation of exogenous compounds such as the polyphenol quercetin [152, 153, 154]. This enzyme is thus responsible for the formation of glucuronidated derivatives within the metabolic pathway of polyphenols.

The remaining reactions should also be analysed further but the examples above show how DRAX can retrieve information that can lead to a more efficient expansion of metabolic models.

A second test was then executed, with the connecting metabolites algorithm (described previously) enabled at a max depth of 2. A fairly low max depth was defined as it was previously established (at least for polyphenols) that most derivatives are found during the initial derivative search. This may not be the case for compounds which were named following a different nomenclature practice.

In this case, DRAX captured 245 unique reactions, of which 102 were excluded. These excluded reactions mostly belonged to starch, a fairly common compound. Since starch has not been extensively annotated in VMH and was captured within a fairly initial reaction ”$starch + rutin = glycosylrutin$”, it was added to the list of compounds to search for. This is one of the flaws of DRAX, and it is one that can be addressed by the improving the ”connecting compounds” exclusion criteria. In addition, it is relevant to note that out of all the compounds which were searched for, only reactions for 15 compounds

(13, if excluding compounds which are not of interest) were extracted. This highlights the issue with the annotation of compounds which metabolism has not been well researched or not yet added to biological databases.

These results demonstrate DRAX's ability to capture a compound's metabolism either through using the search tools available in the queried databases (derivatives retrieval through shared root word) as well as the "connecting compounds" algorithm. To add that the amount of captured information might also be a flaw itself. When constructing metabolic models, it may be preferable to only include the main metabolic pathways, rather than capturing all pathways that may not even be used. This is due to the fact that constraint-based metabolic modelling is based on a mathematical framework, which relies on optimization calculations. In some cases it might be more important to gain a faster execution time, at the expense of a minor model completion loss.

4.2.2 Comparison between DRAX and Recon

DRAX is a very straightforward web scraper and does not rely on prediction, however, a quality control of its scraping capabilities remains indispensable. In this sub-chapter, compounds were queried for different organisms. The first three rows are compounds metabolized specifically by the gut microbes (butyrate) and by the human host (cholesterol and calcitriol). To understand whether DRAX is also capable of capturing drug metabolism, which is usually more specific to certain databases (such as Drugbank), three drugs (Buspirone, dextromethorphan, and methamphetamine) were also tested. The number of reactions in Recon (excluding transport and exchange reactions) and those found during DRAX's query can be seen below in Table 4. To add that the list of gut microbes was extracted from VMH and the connecting compounds algorithm was disabled.

Table 4: Comparison between Recon and DRAX

Compounds	**Recon**	**DRAX**
Fibers (selection for gut microbes)	35	65
Cholesterol (selection for Homo Sapiens)	125*	99
Calcitriol & derivatives (selection for Homo Sapiens)	6	20
Buspirone (selection for Homo Sapiens and gut microbes)	11	5
Dextromethorphan (selection for Homo Sapiens and gut microbes)	8	9
Methamphetamine (selection for Homo Sapiens and gut microbes)	15	12

Results across the manual annotation and DRAX are dissimilar. Manual annotation may find more reactions than DRAX if reactions are buried in literature, or DRAX may find more reactions than manual annotation if it captures previously unknown reactions or incorrect reactions (somewhat common in more rare compounds).

Dietary fiber are carbohydrates resistant to digestion and absorption in the human small intestine. Being resistant to human enzymes,fibers can be completely or partially fermentated in the large intestine by the gut microbiota. [155]. Thus humans rely on the gut microbiota for their partial or total degradation. In this context, a list of fibers (previously annotated in Recon) was used as input and reactions were selected for gut microbes. The list of fibers is enumerated below:

arabinotriose; melibiose; mannotriose; raffinose; stachyose; larch arabinogalactan; kestose; cellulose; xylan; arabinoxylan; galactomannan; alpha-mannan; arabinan; alpha-dextrin; glucomannan; omogalacturonan; pectic galactan; hamnogalacturonan; xyluglucan; Laminarin; Lichenin; inulin; levanbiose.

For these fibers, reactions with undefined compounds were not accounted for, as well as the reactions which were not specific to gut microbiota organisms. To keep in mind that a lot of these reactions were more general than what is generally acceptable within metabolic annotation, which suggests a clear lack of information in databases regarding fiber metabolism. While more reactions were found with DRAX, the overall quality was lacking.

It is also worth to mention the results obtained in the compound cholesterol: Recon has 278 reactions where cholesterol participates, however, if one also excludes reactions where non-specific terms for compounds are used (for example "fatty acid"), then Recon's numbers are reduced to 125 reactions, which is considerably closer to DRAX results.

For example, DRAX found 38 reactions for calcitriol and derivatives (out of these, DRAX was able to reduce the number to 20 unique reactions). However, posterior manual curation of the 38 reactions provided a better overview of calcitriol metabolism.
General (n=12, e.g. *" (...) reduced adrenal ferredoxin (...)"*) and wrong (n=3, incoherent reaction products) reactions were removed, resulting in 23 reactions. Afterwards, each reaction was analyzed and 7 metabolically coherent reactions were found:

- Hydroxylation of 25-hydroxycholecalciferol (calcidiol) into 1alpha,25-dihydroxyvitamin D3 (calcitriol);

- Dehydration of 1alpha,25-dihydroxyvitamin D3 D3 into 25-hydroxycholecalciferol (calcidiol);

- Hydroxylation of1alpha,25-dihydroxyvitamin D3 into 1alpha,24,25-trihydroxyvitamin D3;

- Hydroxylation 1alpha,24,25-trihydroxyvitamin D3 into 1alpha-hydroxy-23-carboxy-24,25,26,27-tetranorvitamin D3 (calcitroic acid);

- Degradation of 1alpha,24(R)-dihydroxyvitamin D3 , chemically synthesized analogue of 25-dihydroxyvitamin D3;

- Formation of metabolite 1alpha,25-dihydroxyvitamin D3-26,23-lactone, from 1alpha,25-dihydroxyvitamin D3;

- Glucuronidation of 1alpha,25-dihydroxyvitamin D3, forming the metabolite 1alpha,25-dihydroxyvitamin D3 25-O-beta-D-glucuronoside.

In summary, the reactions found mostly involve trade of hydroxyl groups (a common reaction during phase I of xenobiotics metabolism) as well as formation of vitamin D3 metabolites and analogues.

As seen from the previous results, biological databases display a comprehensive interpretation of published scientific research and thus are able to provide a significant amount of metabolic reactions (albeit not so much regarding more niche metabolic subsystems or metabolites) , without requiring one to extensively browse through literature. Curation of these reactions is still essential to ensure an accurate representation of metabolism. This then directly reflects DRAX capabilities as a database web scraper.

DRAX has also shown a good capacity at capturing metabolic pathways, both through the initial derivatives search and the "connecting compounds" algorithm. This will give researchers an outlook on more complex metabolic pathways whilst also providing a good basis for annotation of novel compounds.

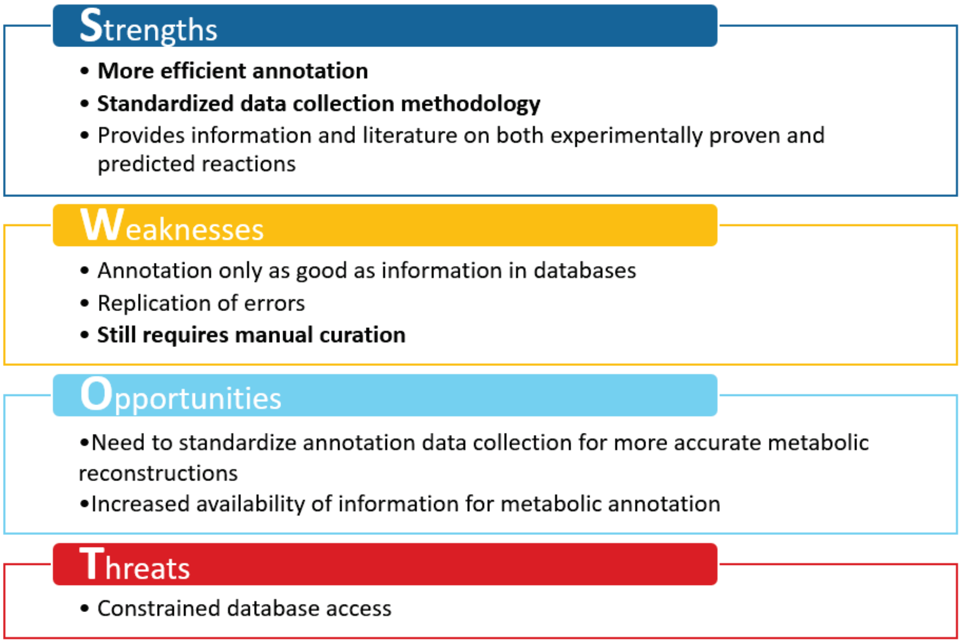
Chapter 5

Conclusion

DRAX provides a standardized framework for the annotation of metabolic reactions, especially novel compounds, such as polyphenols and other niche dietary compounds.
With this tool, human and microbe-specific metabolic reconstructions can be expanded at a faster pace and hopefully, in a few years or decades, we will have more comprehensive metabolic reconstructions that can be used in the treatment of the increasingly common nutrition-related diseases and syndromes, as well as aiding in developing treatments for other clinical issues. With a highlighted grasp on the metabolic pathways promoted or disrupted through dietary intake, identifying pathways that can affect gene expression will also become easier, thus driving Nutritional Genomics forward.

There are some tools [116, 117, 156] that allow the automated construction of metabolic reconstructions, some of which were mentioned in the state of the art. These generate draft reconstructions which can be used as a starting point for metabolic modeling. However DRAX does not aim to replace these, instead, DRAX provides a framework for extension and refinement of previously generated metabolic reconstructions, being, therefore, a novel tool in the field of metabolic modeling.

To summarize, a SWOT analysis of this work is provided in Figure 20 :



Strengths

- **More efficient annotation**
- **Standardized data collection methodology**
- Provides information and literature on both experimentally proven and predicted reactions

Weaknesses

- Annotation only as good as information in databases
- Replication of errors
- **Still requires manual curation**

Opportunities

•Need to standardize annotation data collection for more accurate metabolic reconstructions

•Increased availability of information for metabolic annotation

Threats

- Constrained database access

Figure 20: SWOT analysis of DRAX

It is important to directly mention the constrained database access and how researchers found ways to circumvent established systems, which prevent free access to knowledge [157, 158]. This is an issue with deep roots in scientific research and it is something that should be addressed by the scientific community. Publishing of papers and non-sensitive data on open-access platforms should be encouraged by senior and junior researchers, so that knowledge can be equally available to scientists, students, and the general public. Peer review, provided by academic publishers and other sources, is crucial as it ensures scientific validity and promotes collaborative and constructive criticism. However centralizing and allowing a monopoly of commercially oriented publishers is not the way forward, it does not benefit researchers, only allows the festering of an ideology that aims to hide knowledge behind a paywall. This is even more problematic for countries, universities, researchers, or students with less monetary means.

Chapter 6

Final thoughts and future work

The work developed during the M.Sc. and thesis in Bioinformatics has helped in improving research, programming, and analytical skills and opened up new possibilities in computational biology. It has provided tools and knowledge to put into practice novel methodologies and solutions to implement new systems or optimize already established ones.

While already stable, DRAX will remain in development past the period of the M.Sc. internship. As its development continues, as the sole developer, my main hope is that it will start being used by the community of researchers that do the annotation of metabolic reactions. I also hope that I can keep pursuing its development and remain able to implement current and future ideas.

Despite subject to change, ideas for future work include:

- Improve saving system for data persistence;
- Improve efficiency;
- Graphical user interface;
- Scientific papers text mining for increased output reliability.

The complexity of each bullet point is progressively more complex, thus whether or not the implementation of these features would be possible would also depend on future projects.

DRAX can be optimized, multi threading in particular requires future improvements. As of this moment each thread will fetch information from a single database. While this is more efficient than a streamlined/single-threaded implementation, it would be ideal to be able to exchange threads between databases, should a database query finish before the others.

In addition, it is important to note that DRAX should be updated regularly in order to keep up to date with any database structural change. In this regard, it would be important for databases to provide comprehensive APIs that would replace the need to parse web pages and review parsing criteria whenever interface changes occur.

In conclusion, DRAX is a functional and useful tool, yet still a work in progress and open to improvements and updates.

Chapter 7

References

[1] Cristiana Pavlidis, George P. Patrinos, and Theodora Katsila. "Nutrigenomics: A controversy". In: *Applied & Translational Genomics* 4 (Supplement C Mar. 1, 2015), pp. 50–53. ISSN: 2212-0661. DOI: 10.1016/j.atg.2015.02.003. URL: http://www.sciencedirect.com/science/article/pii/S2212066115000058.

[2] Ines Thiele et al. "A community-driven global reconstruction of human metabolism". In: *Nature Biotechnology* 31.5 (May 2013), pp. 419–425. ISSN: 1546-1696. DOI: 10.1038/nbt.2488.

[3] Saeed Shoaie et al. "Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome". In: *Cell Metabolism* 22.2 (Aug. 4, 2015), pp. 320–331. ISSN: 1932-7420. DOI: 10.1016/j.cmet.2015.07.001.

[4] Swagatika Sahoo and Ines Thiele. "Predicting the impact of diet and enzymopathies on human small intestinal epithelial cells". In: *Human Molecular Genetics* 22.13 (July 1, 2013), pp. 2705–2722. ISSN: 1460-2083. DOI: 10.1093/hmg/ddt119.

[5] Jan Schellenberger et al. “Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0”. In: *Nature Protocols* 6.9 (Aug. 4, 2011), pp. 1290–1307. ISSN: 1750-2799. DOI: 10.1038/nprot.2011.308.

[6] *Global and regional food consumption patterns and trends*. Nov. 21, 2017. URL: http://www.fao.org/docrep/005/ac911e/ac911e05.htm.

[7] Thomas L. Halton et al. “Potato and french fry consumption and risk of type 2 diabetes in women”. In: *The American Journal of Clinical Nutrition* 83.2 (Feb. 1, 2006), pp. 284–290. ISSN: 0002-9165, 1938-3207. URL: http://ajcn.nutrition.org/content/83/2/284.

[8] Fumiaki Imamura et al. “Dietary quality among men and women in 187 countries in 1990 and 2010: a systematic assessment”. In: *The Lancet Global Health* 3.3 (Mar. 1, 2015), e132–e142. ISSN: 2214-109X. DOI: 10.1016/S2214-109X(14)70381-X. URL: http://www.thelancet.com/journals/langlo/article/PIIS2214-109X(14)70381-X/abstract.

[9] GBD 2015 Risk Factors Collaborators. “Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015”. In: *Lancet (London, England)* 388.10053 (Oct. 8, 2016), pp. 1659–1724. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(16)31679-8.

[10] Lukas Schwingshackl, Berit Bogensberger, and Georg Hoffmann. “Diet Quality as Assessed by the Healthy Eating Index, Alternate Healthy Eating Index, Dietary Approaches to Stop Hypertension Score, and

Health Outcomes: An Updated Systematic Review and Meta-Analysis of Cohort Studies". In: *Journal of the Academy of Nutrition and Dietetics* 118.1 (Jan. 1, 2018), 74–100.e11. ISSN: 2212-2672. DOI: 10.1016/j.jand.2017.08.024. URL: http://www.sciencedirect.com/science/article/pii/S2212267217312601.

[11] M. L. McCullough et al. "Adherence to the Dietary Guidelines for Americans and risk of major chronic disease in men". In: *The American Journal of Clinical Nutrition* 72.5 (Nov. 2000), pp. 1223–1231. ISSN: 0002-9165.

[12] M. L. McCullough et al. "Adherence to the Dietary Guidelines for Americans and risk of major chronic disease in women". In: *The American Journal of Clinical Nutrition* 72.5 (Nov. 2000), pp. 1214–1222. ISSN: 0002-9165.

[13] Daphne P. Guh et al. "The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis". In: *BMC public health* 9 (Mar. 25, 2009), p. 88. ISSN: 1471-2458. DOI: 10.1186/1471-2458-9-88.

[14] Vincenzo Lattanzio(eds.) *Recent Advances in Polyphenol Research, Volume 1*. Wiley-Blackwell, 2008. ISBN: 978-1-4051-5837-4. URL: http://gen.lib.rus.ec/book/index.php?md5=d21bc3899a02670981d1d8251344ce29.

[15] *Illustrated Glossary of Organic Chemistry - Phenol; phenolate; phenoxide*. URL: http://www.chem.ucla.edu/~harding/IGOC/P/phenol.html.

[16] *Polyphenols: definition, chemical structure, classification*. Tuscany Diet. Jan. 12, 2014. URL: http://www.tuscany-diet.net/2014/01/12/polyphenols-definition-structure-classification/.

[17] Claudine Manach et al. "Polyphenols: food sources and bioavailability". In: *The American Journal of Clinical Nutrition* 79.5 (May 1, 2004), pp. 727–747. ISSN: 0002-9165, 1938-3207. URL: http://ajcn.nutrition.org/content/79/5/727.

[18] Rong Tsao. "Chemistry and Biochemistry of Dietary Polyphenols". In: *Nutrients* 2.12 (Dec. 10, 2010), pp. 1231–1246. ISSN: 2072-6643. DOI: 10.3390/nu2121231. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3257627/.

[19] *Database on Polyphenol Content in Foods - Phenol-Explorer.* URL: http://phenol-explorer.eu/.

[20] International Union of Pure {and} Applied Chemistry. *IUPAC Gold Book - aglycon (aglycone).* URL: https://goldbook.iupac.org/html/A/A00185.html.

[21] Anna-Marja Aura. "Microbial metabolism of dietary phenolic compounds in the colon". In: *Phytochemistry Reviews* 7.3 (Oct. 1, 2008), pp. 407–429. ISSN: 1568-7767, 1572-980X. DOI: 10.1007/s11101-008-9095-3. URL: https://link.springer.com/article/10.1007/s11101-008-9095-3.

[22] Ilaria Zanotti et al. "Atheroprotective effects of (poly)phenols: a focus on cell cholesterol metabolism". In: *Food & Function* 6.1 (Jan. 2015), pp. 13–31. ISSN: 2042-650X. DOI: 10.1039/c4fo00670d.

[23] Stéphane Bastianetto, Caroline Ménard, and Rémi Quirion. "Neuroprotective action of resveratrol". In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease.* Resveratrol: Challenges in translating pre-clinical findings to improved patient outcomes 1852.6 (June 1, 2015), pp. 1195–1201. ISSN: 0925-4439. DOI: 10.1016/j.bbadis.

2014.09.011. URL: http://www.sciencedirect.com/science/article/pii/S0925443914002920.

[24] Silvia Bradamante, Livia Barenghi, and Alessandro Villa. "Cardiovascular protective effects of resveratrol". In: *Cardiovascular Drug Reviews* 22.3 (2004), pp. 169–188. ISSN: 0897-5957.

[25] Sanghyun Cho et al. "Cardiovascular Protective Effects and Clinical Applications of Resveratrol". In: *Journal of Medicinal Food* 20.4 (Apr. 2017), pp. 323–334. ISSN: 1557-7600. DOI: 10.1089/jmf.2016.3856.

[26] Justine Renaud and Maria-Grazia Martinoli. "Resveratrol as a protective molecule for neuroinflammation: a review of mechanisms". In: *Current Pharmaceutical Biotechnology* 15.4 (2014), pp. 318–329. ISSN: 1873-4316.

[27] Heng Zhang et al. "The protective effects of Resveratrol against radiation-induced intestinal injury". In: *BMC Complementary and Alternative Medicine* 17 (Aug. 16, 2017). ISSN: 1472-6882. DOI: 10.1186/s12906-017-1915-9. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5559783/.

[28] Lidwine Johannot and Shawn M. Somerset. "Age-related variations in flavonoid intake and sources in the Australian population". In: *Public Health Nutrition* 9.8 (Dec. 2006), pp. 1045–1054. ISSN: 1475-2727, 1368-9800. DOI: 10.1017/PHN2006971. URL: https://www.cambridge.org/core/journals/public-health-nutrition/article/agerelated-variations-in-flavonoid-intake-and-sources-in-the-australian-population/466B443DD364A1FE7E428DCC8DA88AC1.

[29] *Natural Antioxidants and Food Quality in Atherosclerosis and Cancer Prevention - 1st Edition*. URL: https://www.elsevier.com/books/

natural-antioxidants-and-food-quality-in-atherosclerosis-and-cancer-prevention/kumpulainen/978-1-85573-794-5.

[30] M. G. Hertog et al. "Intake of potentially anticarcinogenic flavonoids and their determinants in adults in The Netherlands". In: *Nutrition and Cancer* 20.1 (1993), pp. 21–29. ISSN: 0163-5581. DOI: 10.1080/01635589309514267.

[31] Jorma T. Kumpulainen. *Intakes of flavonoids, phenolic acids and lignans in various populations.* Kuopion yliopiston painatuskeskus, 2001. ISBN: 978-951-781-846-9. URL: http://jukuri.luke.fi/handle/10024/451487.

[32] U. Justesen, P. Knuthsen, and T. Leth. "Determination of plant polyphenols in Danish foodstuffs by HPLC-UV and LC-MS detection". In: *Cancer Letters* 114.1 (Mar. 19, 1997), pp. 165–167. ISSN: 0304-3835.

[33] Laura Sampson et al. "Flavonol and flavone intakes in US health professionals". In: *Journal of the American Dietetic Association* 102.10 (Oct. 2002), pp. 1414–1420. ISSN: 0002-8223.

[34] María V. Selma, Juan C. Espín, and Francisco A. Tomás-Barberán. "Interaction between phenolics and gut microbiota: role in human health". In: *Journal of Agricultural and Food Chemistry* 57.15 (Aug. 12, 2009), pp. 6485–6501. ISSN: 1520-5118. DOI: 10.1021/jf902107d.

[35] M. K. Piskula. "Soy isoflavone conjugation differs in fed and food-deprived rats". In: *The Journal of Nutrition* 130.7 (July 2000), pp. 1766–1771. ISSN: 0022-3166.

[36] Sai Manasa Jandhyala. "Role of the normal gut microbiota". In: *World Journal of Gastroenterology* 21.29 (2015), p. 8787. ISSN: 1007-9327.

DOI: 10.3748/wjg.v21.i29.8787. URL: http://www.wjgnet.com/1007-9327/full/v21/i29/8787.htm.

[37] Ting-Chin David Shen. “Diet and Gut Microbiota in Health and Disease”. In: 88 (2017), pp. 117–126. DOI: 10.1159/000455220. URL: http://www.karger.com/Article/FullText/455220.

[38] I. Sekirov et al. “Gut Microbiota in Health and Disease”. In: *Physiological Reviews* 90.3 (July 1, 2010), pp. 859–904. ISSN: 0031-9333, 1522-1210. DOI: 10.1152/physrev.00045.2009. URL: http://physrev.physiology.org/cgi/doi/10.1152/physrev.00045.2009.

[39] M A Harris, C A Reddy, and G R Carter. “Anaerobic bacteria from the large intestine of mice.” In: *Applied and Environmental Microbiology* 31.6 (June 1976), pp. 907–912. ISSN: 0099-2240. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC169855/.

[40] Dwayne C. Savage. “Associations of Indigenous Microorganisms with Gastrointestinal Mucosal Epithelia”. In: *The American Journal of Clinical Nutrition* 23.11 (Nov. 1, 1970), pp. 1495–1501. ISSN: 0002-9165, 1938-3207. URL: http://ajcn.nutrition.org/content/23/11/1495.

[41] J. H. Gordon and R. Dubos. “The anaerobic bacterial flora of the mouse cecum”. In: *The Journal of Experimental Medicine* 132.2 (Aug. 1, 1970), pp. 251–260. ISSN: 0022-1007.

[42] Paul B. Eckburg et al. “Diversity of the human intestinal microbial flora”. In: *Science (New York, N.Y.)* 308.5728 (June 10, 2005), pp. 1635–1638. ISSN: 1095-9203. DOI: 10.1126/science.1110591.

[43] Daniel N. Frank et al. “Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases”. In: *Proceedings of the National Academy of Sciences of the United States of America* 104.34 (Aug. 21, 2007), pp. 13780–13785. ISSN: 0027-8424. DOI: 10.1073/pnas.0706625104. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1959459/.

[44] Alexander Swidsinski et al. “Spatial organization of bacterial flora in normal and inflamed intestine: a fluorescence in situ hybridization study in mice”. In: *World Journal of Gastroenterology* 11.8 (Feb. 28, 2005), pp. 1131–1140. ISSN: 1007-9327.

[45] Maria Carmen Collado et al. “Human gut colonisation may be initiated *in utero* by distinct microbial communities in the placenta and amniotic fluid”. In: *Scientific Reports* 6 (Mar. 22, 2016), srep23129. ISSN: 2045-2322. DOI: 10.1038/srep23129. URL: https://www.nature.com/articles/srep23129.

[46] V. Redondo-Lopez, R. L. Cook, and J. D. Sobel. “Emerging role of lactobacilli in the control and maintenance of the vaginal bacterial microflora”. In: *Reviews of Infectious Diseases* 12.5 (Oct. 1990), pp. 856–872. ISSN: 0162-0886.

[47] Anu Huurre et al. “Mode of delivery - effects on gut microbiota and humoral immunity”. In: *Neonatology* 93.4 (2008), pp. 236–240. ISSN: 1661-7819. DOI: 10.1159/000111102.

[48] R. I. Mackie, A. Sghir, and H. R. Gaskins. “Developmental microbial ecology of the neonatal gastrointestinal tract”. In: *The American Journal of Clinical Nutrition* 69.5 (May 1999), 1035S–1045S. ISSN: 0002-9165.

[49] R. Mändar and M. Mikelsaar. “Transmission of mother’s microflora to the newborn at birth”. In: *Biology of the Neonate* 69.1 (1996), pp. 30–35. ISSN: 0006-3126.

[50] Tanya Yatsunenko et al. “Human gut microbiome viewed across age and geography”. In: *Nature* 486.7402 (May 9, 2012), pp. 222–227. ISSN: 1476-4687. DOI: 10.1038/nature11053.

[51] Ruth E. Ley et al. “Obesity alters gut microbial ecology”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.31 (Aug. 2, 2005), pp. 11070–11075. ISSN: 0027-8424. DOI: 10.1073/pnas.0504978102.

[52] Chenhong Zhang et al. “Interactions between gut microbiota, host genetics and diet relevant to development of metabolic syndromes in mice”. In: *The ISME journal* 4.2 (Feb. 2010), pp. 232–241. ISSN: 1751-7370. DOI: 10.1038/ismej.2009.112.

[53] Shigemitsu Tanaka et al. “Influence of antibiotic exposure in the early postnatal period on the development of intestinal microbiota”. In: *FEMS immunology and medical microbiology* 56.1 (June 2009), pp. 80–87. ISSN: 1574-695X. DOI: 10.1111/j.1574-695X.2009.00553.x.

[54] Sonja Löfmark et al. “Clindamycin-induced enrichment and long-term persistence of resistant Bacteroides spp. and resistance genes”. In: *The Journal of Antimicrobial Chemotherapy* 58.6 (Dec. 2006), pp. 1160–1167. ISSN: 0305-7453. DOI: 10.1093/jac/dkl420.

[55] Cecilia Jernberg et al. “Long-term ecological impacts of antibiotic administration on the human intestinal microbiota”. In: *The ISME*

journal 1.1 (May 2007), pp. 56–66. ISSN: 1751-7362. DOI: 10.1038/ismej.2007.3.

[56] Les Dethlefsen et al. "The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing". In: *PLoS biology* 6.11 (Nov. 18, 2008), e280. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0060280.

[57] M. F. De La Cochetière et al. "Effect of antibiotic therapy on human fecal microbiota and the relation to the development of Clostridium difficile". In: *Microbial Ecology* 56.3 (Oct. 2008), pp. 395–402. ISSN: 0095-3628. DOI: 10.1007/s00248-007-9356-5.

[58] Chris S. Smillie et al. "Ecology drives a global network of gene exchange connecting the human microbiome". In: *Nature* 480.7376 (Oct. 30, 2011), pp. 241–244. ISSN: 1476-4687. DOI: 10.1038/nature10571.

[59] H. Ochman, J. G. Lawrence, and E. A. Groisman. "Lateral gene transfer and the nature of bacterial innovation". In: *Nature* 405.6784 (May 18, 2000), pp. 299–304. ISSN: 0028-0836. DOI: 10.1038/35012500.

[60] Laura S. Frost et al. "Mobile genetic elements: the agents of open source evolution". In: *Nature Reviews. Microbiology* 3.9 (Sept. 2005), pp. 722–732. ISSN: 1740-1526. DOI: 10.1038/nrmicro1235.

[61] Kiera Murphy et al. "The Composition of Human Milk and Infant Faecal Microbiota Over the First Three Months of Life: A Pilot Study". In: *Scientific Reports* 7 (Jan. 17, 2017), p. 40597. ISSN: 2045-2322. DOI: 10.1038/srep40597. URL: https://www.nature.com/articles/srep40597.

[62] Eric M. Brown, Manish Sadarangani, and B. Brett Finlay. “The role of the immune system in governing host-microbe interactions in the intestine”. In: *Nature Immunology* 14.7 (July 2013), pp. 660–667. ISSN: 1529-2916. DOI: 10.1038/ni.2611.

[63] Lindsey G. Albenberg and Gary D. Wu. “Diet and the intestinal microbiome: associations, functions, and implications for health and disease”. In: *Gastroenterology* 146.6 (May 2014), pp. 1564–1572. ISSN: 1528-0012. DOI: 10.1053/j.gastro.2014.01.058.

[64] Federica Guaraldi and Guglielmo Salvatori. “Effect of Breast and Formula Feeding on Gut Microbiota Shaping in Newborns”. In: *Frontiers in Cellular and Infection Microbiology* 2 (Oct. 16, 2012). ISSN: 2235-2988. DOI: 10.3389/fcimb.2012.00094. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3472256/.

[65] W. Allan Walker and Rajashri Shuba Iyengar. “Breast milk, microbiota, and intestinal immune homeostasis”. In: *Pediatric Research* 77.1 (Oct. 13, 2014), p. 220. ISSN: 1530-0447. DOI: 10.1038/pr.2014.160. URL: https://www.nature.com/articles/pr2014160.

[66] Věra Bunešová et al. “Bifidobacteria, Lactobacilli, and Short Chain Fatty Acids of Vegetarians and Omnivores”. In: *Scientia Agriculturae Bohemica* 48.1 (2017), pp. 47–54. DOI: 10.1515/sab-2017-0007. URL: https://www.degruyter.com/view/j/sab.2017.48.issue-1/sab-2017-0007/sab-2017-0007.xml.

[67] Michael A. Conlon and Anthony R. Bird. “The Impact of Diet and Lifestyle on Gut Microbiota and Human Health”. In: *Nutrients* 7.1 (Dec. 24, 2014), pp. 17–44. ISSN: 2072-6643. DOI: 10.3390/nu7010017. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4303825/.

[68] Carlotta De Filippo et al. “Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.33 (Aug. 17, 2010), pp. 14691–14696. ISSN: 0027-8424. DOI: 10.1073/pnas.1005963107. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2930426/.

[69] Anastassia Gorvitovskaia, Susan P. Holmes, and Susan M. Huse. “Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle”. In: *Microbiome* 4 (Apr. 12, 2016). ISSN: 2049-2618. DOI: 10.1186/s40168-016-0160-7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4828855/.

[70] Emily R. Davenport et al. “Seasonal variation in human gut microbiome composition”. In: *PloS One* 9.3 (2014), e90731. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0090731.

[71] Gary D. Wu et al. “Linking long-term dietary patterns with gut microbial enterotypes”. In: *Science (New York, N.Y.)* 334.6052 (Oct. 7, 2011), pp. 105–108. ISSN: 1095-9203. DOI: 10.1126/science.1208344.

[72] Steven R. Gill et al. “Metagenomic analysis of the human distal gut microbiome”. In: *Science (New York, N.Y.)* 312.5778 (June 2, 2006), pp. 1355–1359. ISSN: 1095-9203. DOI: 10.1126/science.1124234.

[73] Justin L. Sonnenburg et al. “Glycan foraging in vivo by an intestine-adapted bacterial symbiont”. In: *Science (New York, N.Y.)* 307.5717 (Mar. 25, 2005), pp. 1955–1959. ISSN: 1095-9203. DOI: 10.1126/science.1109051.

[74] Jr Charles A Janeway et al. “The mucosal immune system”. In: (2001). URL: https://www.ncbi.nlm.nih.gov/books/NBK27169/.

[75] Mariagrazia Valentini et al. *Immunomodulation by Gut Microbiota: Role of Toll-Like Receptor Expressed by T Cells*. Journal of Immunology Research. 2014. URL: https://www.hindawi.com/journals/jir/2014/586939/.

[76] T. Andrew Clayton et al. "Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism". In: *Proceedings of the National Academy of Sciences of the United States of America* 106.34 (Aug. 25, 2009), pp. 14728–14733. ISSN: 1091-6490. DOI: 10.1073/pnas.0904489106.

[77] J. R. Saha et al. "Digoxin-inactivating bacteria: identification in human gut flora". In: *Science (New York, N.Y.)* 220.4594 (Apr. 15, 1983), pp. 325–327. ISSN: 0036-8075.

[78] Bret D. Wallace et al. "Alleviating cancer drug toxicity by inhibiting a bacterial enzyme". In: *Science (New York, N.Y.)* 330.6005 (Nov. 5, 2010), pp. 831–835. ISSN: 1095-9203. DOI: 10.1126/science.1191175.

[79] Jeremy K. Nicholson, Elaine Holmes, and Ian D. Wilson. "Gut microorganisms, mammalian metabolism and personalized health care". In: *Nature Reviews. Microbiology* 3.5 (May 2005), pp. 431–438. ISSN: 1740-1526. DOI: 10.1038/nrmicro1152.

[80] Sandra Macfarlane and George T. Macfarlane. "Regulation of short-chain fatty acid production". In: *The Proceedings of the Nutrition Society* 62.1 (Feb. 2003), pp. 67–72. ISSN: 0029-6651. DOI: 10.1079/PNS2002207.

[81] R. Balfour Sartor. "Microbial influences in inflammatory bowel diseases". In: *Gastroenterology* 134.2 (Feb. 2008), pp. 577–594. ISSN: 1528-0012. DOI: 10.1053/j.gastro.2007.11.059.

[82] Sylvia H. Duncan, Petra Louis, and Harry J. Flint. “Lactate-Utilizing Bacteria, Isolated from Human Feces, That Produce Butyrate as a Major Fermentation Product”. In: *Applied and Environmental Microbiology* 70.10 (Oct. 2004), pp. 5810–5817. ISSN: 0099-2240. DOI: 10.1128/AEM.70.10.5810-5817.2004. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC522113/.

[83] Ian Rowland et al. “Gut microbiota functions: metabolism of nutrients and other food components”. In: *European Journal of Nutrition* (Apr. 9, 2017). ISSN: 1436-6215. DOI: 10.1007/s00394-017-1445-8.

[84] Brandi L. Cantarel, Vincent Lombard, and Bernard Henrissat. “Complex carbohydrate utilization by the healthy human microbiome”. In: *PloS One* 7.6 (2012), e28742. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0028742.

[85] L. V. Hooper et al. “Molecular analysis of commensal host-microbial relationships in the intestine”. In: *Science (New York, N.Y.)* 291.5505 (Feb. 2, 2001), pp. 881–884. ISSN: 0036-8075. DOI: 10.1126/science.291.5505.881.

[86] Estelle Devillard et al. “Differences between human subjects in the composition of the faecal bacterial community and faecal metabolism of linoleic acid”. In: *Microbiology (Reading, England)* 155 (Pt 2 Feb. 2009), pp. 513–520. ISSN: 1350-0872. DOI: 10.1099/mic.0.023416-0.

[87] Estelle Devillard et al. “Metabolism of Linoleic Acid by Human Gut Bacteria: Different Routes for Biosynthesis of Conjugated Linoleic Acid”. In: *Journal of Bacteriology* 189.6 (Mar. 2007), pp. 2566–2570. ISSN: 0021-9193. DOI: 10.1128/JB.01359-06. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1899373/.

[88] A. Baddini Feitoza et al. “Conjugated linoleic acid (CLA): effect modulation of body composition and lipid profile”. In: *Nutricion Hospitalaria* 24.4 (Aug. 2009), pp. 422–428. ISSN: 0212-1611.

[89] Vidya R. Velagapudi et al. “The gut microbiota modulates host energy and lipid metabolism in mice”. In: *Journal of Lipid Research* 51.5 (May 2010), pp. 1101–1112. ISSN: 1539-7262. DOI: 10.1194/jlr.M002774.

[90] Satoru Fukiya et al. “Conversion of cholic acid and chenodeoxycholic acid into their 7-oxo derivatives by Bacteroides intestinalis AM-1 isolated from human feces”. In: *FEMS microbiology letters* 293.2 (Apr. 2009), pp. 263–270. ISSN: 1574-6968. DOI: 10.1111/j.1574-6968.2009.01531.x.

[91] Hana Ajouz, Deborah Mukherji, and Ali Shamseddine. “Secondary bile acids: an underrecognized cause of colon cancer”. In: *World Journal of Surgical Oncology* 12 (May 24, 2014), p. 164. ISSN: 1477-7819. DOI: 10.1186/1477-7819-12-164. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4041630/.

[92] Harris Bernstein et al. “Bile acids as endogenous etiologic agents in gastrointestinal cancer”. In: *World Journal of Gastroenterology : WJG* 15.27 (July 21, 2009), pp. 3329–3340. ISSN: 1007-9327. DOI: 10.3748/wjg.15.3329. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2712893/.

[93] Hippocrates of Kos. ““Death sits in the bowels and bad digestion is the root of all evil.””.

[94] Anastasia I. Petra et al. “Gut-Microbiota-Brain Axis and Its Effect on Neuropsychiatric Disorders With Suspected Immune Dysregula-

tion". In: *Clinical Therapeutics* 37.5 (May 2015), pp. 984–995. ISSN: 01492918. DOI: 10.1016/j.clinthera.2015.04.002. URL: http://linkinghub.elsevier.com/retrieve/pii/S014929181500226X.

[95] Jose M. Ordovas and Vincent Mooser. "Nutrigenomics and nutrigenetics". In: *Current Opinion in Lipidology* 15.2 (Apr. 2004), p. 101. ISSN: 0957-9672. URL: http://journals.lww.com/co-lipidology/Abstract/2004/04000/Nutrigenomics_and_nutrigenetics.2.aspx.

[96] David M. Mutch, Walter Wahli, and Gary Williamson. "Nutrigenomics and nutrigenetics: the emerging faces of nutrition". In: *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* 19.12 (Oct. 2005), pp. 1602–1616. ISSN: 1530-6860. DOI: 10.1096/fj.05-3911rev.

[97] Lydia Afman and Michael Müller. "Nutrigenomics: from molecular nutrition to prevention of disease". In: *Journal of the American Dietetic Association* 106.4 (Apr. 2006), pp. 569–576. ISSN: 0002-8223. DOI: 10.1016/j.jada.2006.01.001.

[98] Michael Müller and Sander Kersten. "Nutrigenomics: goals and strategies". In: *Nature Reviews. Genetics* 4.4 (Apr. 2003), pp. 315–322. ISSN: 1471-0056. DOI: 10.1038/nrg1047.

[99] *Khan Academy*. Khan Academy. Jan. 18, 2018. URL: http://www.khanacademy.org.

[100] Ben van Ommen and Rob Stierum. "Nutrigenomics: exploiting systems biology in the nutrition and health arena". In: *Current Opinion in Biotechnology* 13.5 (Oct. 2002), pp. 517–521. ISSN: 0958-1669.

[101] Alberto Noronha et al. “The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease”. In: *bioRxiv* (May 15, 2018), p. 321331. DOI: 10.1101/321331. URL: https://www.biorxiv.org/content/early/2018/05/15/321331.

[102] Stefanía Magnúsdóttir et al. “Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota”. In: *Nature Biotechnology* 35.1 (Jan. 2017), pp. 81–89. ISSN: 1546-1696. DOI: 10.1038/nbt.3703.

[103] Sang Yup Lee, Jong Myoung Park, and Tae Yong Kim. “Application of metabolic flux analysis in metabolic engineering”. In: *Methods in Enzymology* 498 (2011), pp. 67–93. ISSN: 1557-7988. DOI: 10.1016/B978-0-12-385120-8.00004-8.

[104] Georgios A. Pavlopoulos et al. “Using graph theory to analyze biological networks”. In: *BioData Mining* 4 (2011), p. 10. DOI: 10.1186/1756-0381-4-10. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3101653/.

[105] *VMH.* Dec. 26, 2017. URL: https://vmh.uni.lu/#reconmap.

[106] Tunahan Çakır and Mohammad Jafar Khatibipour. “Metabolic Network Discovery by Top-Down and Bottom-Up Approaches and Paths for Reconciliation”. In: *Frontiers in Bioengineering and Biotechnology* 2 (Dec. 3, 2014). ISSN: 2296-4185. DOI: 10.3389/fbioe.2014.00062. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4253960/.

[107] Laurent Heirendt et al. “Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0”. In: *arXiv:1710.04038 [q-*

bio] (Oct. 11, 2017). arXiv: 1710.04038. URL: http://arxiv.org/abs/1710.04038.

[108] Ines Thiele and Bernhard Ø. Palsson. “A protocol for generating a high-quality genome-scale metabolic reconstruction”. In: *Nature protocols* 5.1 (2010), pp. 93–121. ISSN: 1754-2189. DOI: 10.1038/nprot.2009.203. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125167/.

[109] Adam M. Feist et al. “Reconstruction of biochemical networks in microorganisms”. In: *Nature Reviews. Microbiology* 7.2 (Feb. 2009), pp. 129–143. ISSN: 1740-1534. DOI: 10.1038/nrmicro1949.

[110] *KEGG: Kyoto Encyclopedia of Genes and Genomes*. Dec. 26, 2017. URL: http://www.genome.jp/kegg/.

[111] Minoru Kanehisa et al. “KEGG: new perspectives on genomes, pathways, diseases and drugs”. In: *Nucleic Acids Research* 45 (D1 Jan. 4, 2017), pp. D353–D361. ISSN: 1362-4962. DOI: 10.1093/nar/gkw1092.

[112] Minoru Kanehisa et al. “KEGG as a reference resource for gene and protein annotation”. In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D457–462. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1070.

[113] Sandra Placzek et al. “BRENDA in 2017: new perspectives and new tools in BRENDA”. In: *Nucleic Acids Research* 45 (D1 Jan. 4, 2017), pp. D380–D388. ISSN: 0305-1048. DOI: 10.1093/nar/gkw952. URL: https://academic.oup.com/nar/article/45/D1/D380/2290911.

[114] *BRENDA: The Comprehensive Enzyme Information System*. URL: www.brenda-enzymes.org.

[115] Ron Caspi et al. “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases”. In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D471–D480. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1164. URL: https://academic.oup.com/nar/article/44/D1/D471/2502657.

[116] Peter D. Karp, Suzanne Paley, and Pedro Romero. “The Pathway Tools software”. In: *Bioinformatics (Oxford, England)* 18 Suppl 1 (2002), S225–232. ISSN: 1367-4803.

[117] Scott Devoid et al. “Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED”. In: *Methods in Molecular Biology (Clifton, N.J.)* 985 (2013), pp. 17–45. ISSN: 1940-6029. DOI: 10.1007/978-1-62703-299-5_2.

[118] Oscar Dias et al. “Reconstructing genome-scale metabolic models with merlin”. In: *Nucleic Acids Research* 43.8 (2015), pp. 3899–3910. DOI: 10.1093/nar/gkv294. eprint: /oup/backfile/content_public/journal/nar/43/8/10.1093_nar_gkv294/2/gkv294.pdf. URL: http://dx.doi.org/10.1093/nar/gkv294.

[119] Cheng Zhang and Qiang Hua. “Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine”. In: *Frontiers in Physiology* 6 (Jan. 7, 2016). ISSN: 1664-042X. DOI: 10.3389/fphys.2015.00413. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4703781/.

[120] Radhakrishnan Mahadevan et al. “Applications of Metabolic Modeling to Drive Bioprocess Development for the Production of Value-added Chemicals”. In: *Biotechnology and Bioprocess Engineering* 10 (Oct. 1, 2005), pp. 408–417. DOI: 10.1007/BF02989823.

[121] Natalie C. Duarte et al. “Global reconstruction of the human metabolic network based on genomic and bibliomic data”. In: *Proceedings of the National Academy of Sciences* 104.6 (Feb. 6, 2007), pp. 1777–1782. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0610772104. URL: http://www.pnas.org/content/104/6/1777.

[122] Neil Swainston et al. “Recon 2.2: from reconstruction to model of human metabolism”. In: *Metabolomics: Official Journal of the Metabolomic Society* 12 (2016), p. 109. ISSN: 1573-3882. DOI: 10.1007/s11306-016-1051-4.

[123] *HGNC database of human gene names — HUGO Gene Nomenclature Committee*. Dec. 28, 2017. URL: https://www.genenames.org/.

[124] Elizabeth Brunk et al. “Recon3D enables a three-dimensional view of gene variation in human metabolism”. In: *Nature Biotechnology* 36.3 (Mar. 2018), pp. 272–281. ISSN: 1546-1696. DOI: 10.1038/nbt.4072. URL: https://www.nature.com/articles/nbt.4072.

[125] Alberto Noronha et al. “ReconMap: an interactive visualization of human metabolism”. In: *Bioinformatics (Oxford, England)* 33.4 (Feb. 15, 2017), pp. 605–607. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw667.

[126] Michael Graham Espey. “Role of oxygen gradients in shaping redox relationships between the human intestine and its microbiota”. In: *Free Radical Biology & Medicine* 55 (Feb. 2013), pp. 130–140. ISSN: 1873-4596. DOI: 10.1016/j.freeradbiomed.2012.10.554.

[127] Lionel Rigottier-Gois. “Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis”. In: *The ISME journal* 7.7 (July 2013), pp. 1256–1261. ISSN: 1751-7370. DOI: 10.1038/ismej.2013.80.

[128] R J Gibbons and B Kapsimalis. "Estimates of the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice." In: *Journal of Bacteriology* 93.1 (Jan. 1967), pp. 510–512. ISSN: 0021-9193. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC315031/.

[129] Marouen Ben Guebila and Ines Thiele. "Model-based dietary optimization for late-stage, levodopa-treated, Parkinson's disease patients". In: *npj Systems Biology and Applications* 2 (June 16, 2016), p. 16013. ISSN: 2056-7189. DOI: 10.1038/npjsba.2016.13. URL: https://www.nature.com/articles/npjsba201613.

[130] Ines Thiele et al. "Quantitative systems pharmacology and the personalized drug–microbiota–diet axis". In: *Current Opinion in Systems Biology* 4 (Supplement C Aug. 1, 2017), pp. 43–52. ISSN: 2452-3100. DOI: 10.1016/j.coisb.2017.06.001. URL: http://www.sciencedirect.com/science/article/pii/S2452310017300847.

[131] Christopher D. Nogiec and Simon Kasif. "To Supplement or Not to Supplement: A Metabolic Network Framework for Human Nutritional Supplements". In: *PLOS ONE* 8.8 (May 8, 2013), e68751. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0068751. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0068751.

[132] Eugen Bauer and Ines Thiele. "From metagenomic data to personalized computational microbiotas: Predicting dietary supplements for Crohn's disease". In: *arXiv:1709.06007 [q-bio]* (Sept. 18, 2017). arXiv: 1709.06007. URL: http://arxiv.org/abs/1709.06007.

[133] Bing Liu (auth.) *Web data mining: Exploring hyperlinks, contents, and usage data*. 2nd ed. Data-centric systems and applications. Springer-Verlag Berlin Heidelberg, 2011. ISBN: 978-3-642-19459-7. URL: `http://gen.lib.rus.ec/book/index.php?md5=A92526B73B878BEB9B8F0286C79F8593`.

[134] Rajesh R Gawali. "Web Mining techniques, process and applications in Ecommerce". In: (2013), p. 7.

[135] Emilio Ferrara et al. "Web data extraction, applications and techniques: A survey". In: *Knowledge-Based Systems* 70 (Nov. 1, 2014), pp. 301–323. ISSN: 0950-7051. DOI: `10.1016/j.knosys.2014.07.007`. URL: `http://www.sciencedirect.com/science/article/pii/S0950705114002640`.

[136] Roy Thomas Fielding. "Architectural Styles and the Design of Network-based Software Architectures". PhD thesis. Irvine: University of California, 2000. URL: `https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm`.

[137] *The Web Robots Pages*. URL: `http://www.robotstxt.org/`.

[138] Anne Morgat et al. "Updates in Rhea – an expert curated resource of biochemical reactions". In: *Nucleic Acids Research* 45 (D1 Jan. 4, 2017), pp. D415–D418. ISSN: 0305-1048. DOI: `10.1093/nar/gkw990`. URL: `https://academic.oup.com/nar/article/45/D1/D415/2333936`.

[139] Antonio Fabregat et al. "The Reactome Pathway Knowledgebase". In: *Nucleic Acids Research* 46 (D1 Jan. 4, 2018), pp. D649–D655. ISSN: 1362-4962. DOI: `10.1093/nar/gkx1132`.

[140] David S. Wishart et al. “HMDB 4.0: the human metabolome database for 2018”. In: *Nucleic Acids Research* 46 (D1 Jan. 4, 2018), pp. D608–D617. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1089.

[141] David S. Wishart et al. “DrugBank 5.0: a major update to the DrugBank database for 2018”. In: *Nucleic Acids Research* 46 (D1 Jan. 4, 2018), pp. D1074–D1082. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1037.

[142] Kanehisa Laboratories. *KEGG*. URL: https://www.kegg.jp.

[143] *Beautiful Soup Documentation*. URL: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[144] *Python Data Analysis Library — pandas: Python Data Analysis Library*. URL: https://pandas.pydata.org/.

[145] *NumPy — NumPy*. URL: http://www.numpy.org/.

[146] *Requests: HTTP for Humans — Requests 2.19.1 documentation*. URL: http://docs.python-requests.org/en/master/.

[147] *Selenium with Python — Selenium Python Bindings 2 documentation*. URL: http://selenium-python.readthedocs.io/.

[148] *ChemAxon - Software Solutions and Services for Chemistry & Biology*. URL: https://chemaxon.com/.

[149] *pKa calculation*. URL: https://chemaxon.com/marvin-archive/5_2_0/marvin/help/calculations/pKa.html.

[150] Melanie Adams and Zongchao Jia. “Structural and biochemical analysis reveal pirins to possess quercetinase activity”. In: *Journal of Biological Chemistry* 280.31 (2005), pp. 28675–28682.

[151] Yuko Shiba et al. "Flavonoids as substrates and inhibitors of myeloperoxidase: molecular actions of aglycone and metabolites". In: *Chemical research in toxicology* 21.8 (2008), pp. 1600–1609.

[152] Christopher D King et al. "The glucuronidation of exogenous and endogenous compounds by stably expressed rat and human UDP-glucuronosyltransferase 1.1". In: *Archives of biochemistry and biophysics* 332.1 (1996), pp. 92–100.

[153] Nikhil K Basu et al. "Differential and special properties of the major human UGT1-encoded gastrointestinal UDP-glucuronosyltransferases enhance potential to control chemical uptake". In: *Journal of Biological Chemistry* 279.2 (2004), pp. 1429–1441.

[154] Barry D Davis and Jennifer S Brodbelt. "Regioselectivity of human UDP-glucuronosyl-transferase 1A1 in the synthesis of flavonoid glucuronides determined by metal complexation and tandem mass spectrometry". In: *Journal of the American Society for Mass Spectrometry* 19.2 (2008), pp. 246–256.

[155] *Dietary Fiber*. Nov. 20, 2017. URL: `https://www.aaccnet.org/initiatives/definitions/Pages/DietaryFiber.aspx`.

[156] Esa Pitkänen et al. "Comparative Genome-Scale Reconstruction of Gapless Metabolic Networks for Present and Ancestral Species". In: *PLOS Computational Biology* 10.2 (Feb. 6, 2014), e1003465. ISSN: 1553-7358. DOI: `10.1371/journal.pcbi.1003465`. URL: `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003465`.

[157] John Bohannon. *Who's downloading pirated papers? Everyone*. 2016.

[158] Paige Mann. “Sci-Hub provides access to nearly all scholarly literature”. In: *The Idealis* (2018).