# Bioinformatic methods for species-specific metabolome inference

Pablo Moreno

European Bioinformatics Institute

St. John's College, University of Cambridge

A thesis submitted for the degree of

*Doctor of Philosophy*

14 of August, 2012

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation does not exceed the word limit of 60,000 words.

to my beloved wife Claudia
and our darling little Mia.

# Acknowledgements

I would like to acknowledge my supervisor, Dr. Christoph Steinbeck, for the energy and enthusiasm that he put into guiding me, and for being supportive about many personal circumstances I had during my PhD. Chris gave me interesting insights about myself which allowed me to grow and improve during this time.

To my TAC members, for their time and invaluable feedback at each of our meetings.

The support from St. John's College was essential for our stay in Cambridge. The social environment was always pleasant and living next to Jesus Green for so long was incredible.

To the many people at the genome campus that I had the chance to share with. You all made my time at EBI a very exciting one. Particularly, to my colleagues of the Cheminformatics and Metabolism group, for their support, help, and encouragment.

To my PhD brothers Stephan, John, Luis, and our sister Kalai. They made my PhD very exciting and estimulating, they are fantastic friends. I'm grateful for the chats, advice, help proof-reading, and the so many things that I have learnt from all you guys.

To my collaborators, at the EBI and away. Specially to Prof. Joern Piel, Dr. Asad Rahman, Joe Foster, and Eric Helfrich, for the interesting scientific chats and access to their work. A special thank to my friends at the text mining group, who were always happy to help me.

To my friends in Cambridge, and to my friends from Chile who made the effort to visit us. You all made us feel so many times like if we were at home.

To the people of this country, who were always welcoming and happy to have us here. I owe a lot to this country and it's people.

To our parents and family, who were always supportive and loving, and visited us everytime they could. I'm specially in debt with my brother Diego, for all his help back at home.

None of what I did would have been possible without the constant love and support from my beloved wife Claudia, who left so much behind to be here with me. She gave me the most precious gift I could ever receive: our little Mia, who being so young has taught me so much.

Above all, I thank our Lord, for all His blessings, and for giving us the strength to overcome periods of difficult times during the past years.

# Abstract

The metabolome is the complete set of small molecules (<1,500 Da) present in a biological sample or organism. Metabolomics studies metabolomes through technologies such as Nuclear Magnetic Resonance and Mass Spectrometry.

Metabolomics poses a major bioinformatics challenge: to identify the large number of metabolites detected, whose structure and biochemistry is unknown. This thesis contributes towards solving this challenge by developing methods to predict organism-specific metabolomes: metabolism database integration, text mining, and chemical enumeration.

Metabolism data integration – through a novel merge method – shows that merging metabolism resources significantly increases the size of the metabolite catalogue. The integrated metabolite collection covers ~15% of the Human Metabolome Database (HMDB); the main difference is accounted for by the large lipid collection in the HMDB.

The text mining pipeline built – analyzing PubMed abstracts – produces some thousands of additional metabolites and relations between tissues and small molecules. This method retrieves an additional 6% of what remained undiscovered in the HMDB after the database integration part. Results retrieved only through text mining have a bias towards exogenous small molecules.

On enumerating generic reactions from the previous sets, the number of small molecules generated grows exponentially and only a few paths lead to known metabolites. To narrow down the results, I explore methods which rely on thermodynamic feasibility, catalogue

lookups, and reaction similarity. While this part produces only little overlap with what remained uncovered of the HMDB, the selection methods restrict the results to ~5,000 connectivities resembling known molecules, out of ~67,000 connectivities produced.

Polyketides are an example of a more complicated case within metabolism, where reactions steps are defined in a very particular and non-obvious order. I contribute towards the elucidation of polyketide structures through the development of tools to improve our understanding of a rarely studied class of polyketides.

The methods in this thesis aim to be a start-up point for the semi-automatic generation of species-specific metabolomes, producing a result which is in-between existing metabolism resources and highly curated databases such as the HMDB – including as well many molecules that are not part of the HMDB. These methods produce a richer organism-specific catalogue of small molecules, compared to what can be accessed by use of existing metabolism databases.

# Contents

# List of Figures

# Chapter 1

# Introduction

## 1.1 Motivations

Before life, there was chemistry. Proteins, RNAs, and their blue-prints in the genome, enable the cell to act on a chemical world that encompass it. The living cell obtains its building blocks and energy from the surrounding chemical entities through complex processes. The cell aims to harness the chemical world around it to its favour. Yet there is so much to be discovered about the complex interplay between the chemistry and the cell. Thousands of genomes are available today, but only few extensive species-specific collections of small molecules – or metabolomes – are currently compiled [69; 121; 166]. We need more knowledge about the small chemical molecules that are found in specific organisms, tissues, and cell types to improve our understanding of the interplay between organisms and their surrounding chemistry, and of the internal mechanisms of the cell. Such is the importance of small molecules in biology that some have even labeled them as "the missing piece of the central dogma of molecular biology"[132].

Genomes tell us very little about the particular phenotype of a cell or organism at a particular condition[1]. Gene and protein expression take us closer to the phenotype of a cell. Descending through the "omics" cascade [26] (Figure 1.1), the metabolome – the set of small chemical molecules that can be found in a cell or biological sample – gives the most sensitive and accurate representation of a

[1]There are of course some genes, or SNPs in them, that determine phenotype directly, like those for eye colour for instance.

phenotype [26; 37].

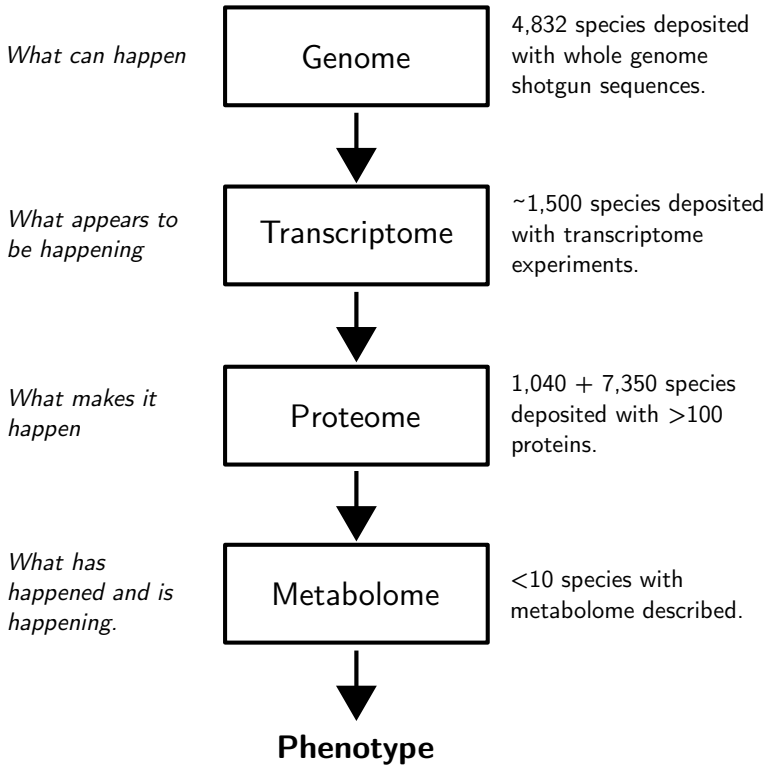



Figure 1.1: The "omics cascade" as proposed in [26], represents only the main flux of information through the "omics". To the right we find the number of organisms that have whole genome sequences (deposited in the European Nucleotide Archive, ENA [90]), of organisms with transcriptomic experiment results (in the ArrayExpress database[116]), and the number of organisms with more than 100 proteins known (1,040 in SwissProt and 7,350 in trEMBL[19]). Metabolomics is behind in developing databases with an adequate coverage of species, compared to other "omics".

Metabolomics, the area that experimentally catalogues and measures small chemical molecules in biological samples, has been around for hundreds of years – shaped to how we know it today, mainly in the past 50 years[157]. However, the task of resolving and identifying small molecules is of such complexity, that still only less than 40% of the small molecules detected in a sample can be adequately identified on average using well established technologies[154; 161]. There are

many experimental challenges that are responsible for this. In the bioinformatics side, the difficulty partly lies in the lack of good resources, such as the Human Metabolome Database (HMDB [166]), that compile existing reference evidence of small molecule occurrence in biological containers (species, organs, tissues, cell types, etc.). Having more resources in this area will improve the chances of identifying more complete sets of small molecules in Metabolomics experiments.

Compiling metabolomes as reference data sets is complex and time consuming because the potential sources of data are not properly integrated. One approach, is the collection in the form of an archive, of as many Metabolomics studies as possible. Examples of these are KnApSacK database [135], BinBase/SetUpX[40], GMD [85], METLIN [136], or the MetaboLights Database at the European Bioinformatics Institute (EBI). This approach, provides a consolidated evidence bank, that allows the comparison of multiple results and cross confirmation of observations. However, it does not solve the problem of identifying unknowns in the data set. For this, a reference biology data base is required, like the HMDB, that provides background knowledge on the extensive list of small molecules that can be found in the biological sample under study. This kind of resource requires a massive investment of man power, so methods that derive this kind of resources in a more automated way would be useful towards the generation of more reference species-specific metabolomes data sets.

## 1.2 Introduction to small molecules

Small molecules are organic chemical compounds with a mass of up to 1,500 Daltons [166].

A small molecule can be classified into many different categories according to physicochemical properties, biological roles (activator, inhibitor, metabolite, etc.), chemical properties (acid, base, alcohol, oxide, etc.), applications (pesticide, anti tumoral, indicator, etc.), substructural features (polycyclic, carboxylic acid part, etc.), or provenance (secondary metabolism, endogenous, natural, synthetic, etc.), to name some. While far from completely covering the whole spectrum of possibilities, chemical ontologies and information hierarchies provide ways of classifying small molecules. I discuss these ontologies later.

Small molecules are connected to the cell mostly through metabolism: the process of uptake of mass and energy which the cell couples with the production of highly ordered structures that allow it to replicate and perpetuate life. Most of the small molecules present in the cell at some point interact with enzymes, to be subject of biologically catalyzed chemical reactions, or at least will interact with a protein for some purpose. Another relevant biochemical process which involves small molecules is osmoregulation, which actively modulates the concentration of solutes, to avoid either very diluted or too viscous solutions.

Depending on their size and other properties, such as polarity, small molecules will either be able to go through biological membranes or require some form of protein mediated transport[27].

Small molecules play a central role in inter-cellular communication, as in chemotaxis and quorum sensing in the case of bacteria, hormone signalling in higher eukaryotes, and biological warfare between bacteria, fungi, insects, and plants. This is mostly achieved by secondary metabolism: the production of normally more complex small molecules that are not necessary for the cell's direct survival, but which in many cases make it more fit. Bacteria, fungi, and plants through their evolution have developed highly sophisticated secondary metabolites. Figure 1.2 shows some examples: a bacterial polyketide, a fungal aflatoxin, and a plant terpene.

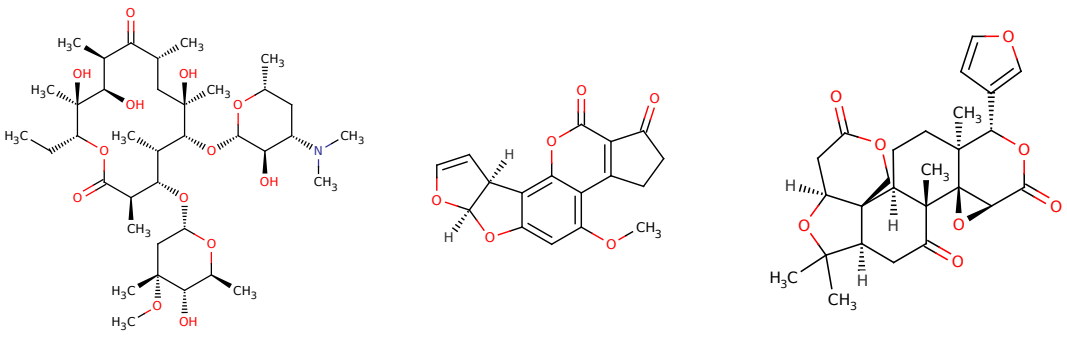



Figure 1.2: The structure of a bacterial polyketide, an *Aspergillus* aflatoxin, and of a citric terpene.

During the past years, open chemical databases such as PubChem [130], KEGG [71], and ChEBI [25] have consolidated as important repositories of reference chemistry data. PubChem receives direct electronic submissions of small molecules, these structures are subject to automatic normalization and merging steps, and then made available to the community. In manually curated resources such as ChEBI, submitted chemical structures are additionally manually checked and corrected by expert chemists. While PubChem has a larger coverage of small molecules, ChEBI has higher quality data. These chemical repositories are relevant, as they provide persistent identifiers for chemical molecules and reference chemical information. They also provide ways of organizing molecules, through hierarchies in the case of PubChem (NCBI MeSH chemical branch) and KEGG(KEGG BRITE chemical categories), and through the ChEBI Ontology.

Detection of small molecules to understand the phenotype of a cell or living organisms has been long used. In ancient cultures, detection of certain small molecules by their smell or taste was a major part of medical diagnosis[157].Currently many diagnostic tests rely on determining concentrations of certain small molecules. This is because concentrations of different small molecules within a cell or biofluid are good snapshots of the active phenotype.

## 1.3 Introduction to Metabolomes

The metabolome refers to the complete set of small molecules (< 1500 Daltons) present in a biological sample or organism [165]. The term was first coined by Steve Oliver in the context of *S. cerevisiae* omics studies[113]. The most common way to investigate metabolomes nowadays is through the use of a wide range of Metabolomics technologies, most of them branching from Nuclear Magnetic Resonance Spectroscopy and Mass Spectrometry.

Estimations on the number of species on earth, range from a few millions to a few tenths of millions [32]. According to [105], ~1.2 million species have been catalogued. Within those known species, a few thousands have been fully sequenced. At the time of this writing, we are approaching a rate of 8,000 million base pairs deposited monthly[1]. The metabolome is considered to be one of

[1]During the first half of 2012, more than 7,900 million base pairs from whole genome shotgun

the main links between genotype and phenotype [37], yet even after completing thousands of genome sequences, we still barely have a few metabolomes available [69; 121; 166].

Metabolomes size estimations are much larger than metabolism reconstructions based on genome sequences. While the metabolism reconstructions of *B. subtilis*, based on the enzymes annotated on its genome, estimated ~800 small molecules, experimental results showed 1,692 possible metabolites[139]. For *H. sapiens*, while the major metabolic resources can explain close to ~2,000 small molecules, the HMDB estimates ~8,000 small molecules. This again is evidence of the knowledge gap between what is known of metabolism and what experimental methods find.

Estimating the size of a metabolome requires to set boundaries to which small molecules are considered as part of a metabolome. While mass constraints can be easy to set, the main problem comes when considering endogenous and exogenous molecules. For a non-versatile single cell microorganism – able to process only few sources of carbon and nitrogen – or an autotroph this might seem straightforward. However, what about a complex higher eukaryote, which is food-fed, has a number of bacterial communities as microflora of different organs, can be affected by disease, and uses drugs for different purposes. All these oddities influence the set of small molecules present in that organism. Additionally, having many different tissues and cell types, each of those have its own localized metabolome. Is the metabolome of such an organism only the small molecules that its own enzymes can process or produce? Boundaries here are suddenly extremely blurry.

The Lipidome, the set of all lipids within the metabolome, has been given special attention in the past decades, due to its impact in a number of diseases[115]. In 2007 the LipidMAPS[146] database was released, which contains structure and annotations for relevant lipids. As of the first half of 2012, the database holds more than 30,000 unique lipid structures. In [168] ~180,000 theoretical lipids have been estimated to cover the major lipid classes (fatty acids, fatty acid acyl-CoA, monoacyl glycerols, diacyl glycerols, triacyl glycerols, phospholipids,

projects were deposited each month in GenBank. Calculation based on section 2.2.8 "Growth of GenBank" of the NCBI-GenBank Flat File Release 190.0 release notes.

ceramides/sphingomyelin, glycosphingolipids, and cholesterols).

Aside experimental Metabolomics, the following sources of metabolome knowledge can be identified:

**Metabolism databases:** These are one of the biggest repositories of metabolome knowledge, where enzymatic reactions with chemical structures for reactants and products are stored.

**Literature:** According to some experts, approximately 20% of knowledge is deposited in structured databases, the rest is hidden in thousands of articles in the literature[62]. Given this, there is probably plenty of metabolome knowledge that is not found in metabolism databases and that is only found in literature.

**Chemical databases:** While chemical databases do not provide an association of chemical entities with organisms, they do provide reference chemical data for the small molecules, which through other resources can be associated with particular biological containers.

**Whole genome metabolic models:** These are sets of reactions and molecules obtained after the annotation of the enzymes of an organism[151]. Although they rely mostly on metabolism databases, sometimes they also undergo manual curation, which can bring new molecules into the models. However, as one of the concerns when building them is adequate connectivity, most of the times these models avoid small molecules that are not appropriately connected. Most of these models do not include chemical structures, and often not even an adequate cross reference to a chemical database.

## 1.4 Introduction to Metabolomics

Metabolomics can be defined as the comprehensive and quantitative survey of all small molecules present in a biological sample, tissue or cell type [26; 50]. This survey can be achieved by a number of experimental methods, followed by data analysis.

Currently, the most common detection methods used are Mass Spectrometry (MS) and Nuclear Magnetic Resonance (NMR) Spectroscopy [29]. Normally, MS detectors are coupled with a previous separation step, such as gas (GC-MS) or liquid chromatography (LC-MS)[48], which gives a further dimension (of retention time) to the mass over charge ($m/z$) result. While generating in silico $m/z$ predictions for a molecule is relatively straightforward, the retention time estimation is the most tricky part to predict, as it depends on the separation column used and on a number of experimental factors (such as temperature, mobile phase, and matrix of the sample among others)[21]. Through fragmentation trees, where a selected ion is fragmented iteratively, MS can shed light on the structure of a small molecule[75]. This is a time consuming task, and so it can be only done for few compounds. Today the main challenges in MS are the identification of compounds and the absolute quantification of concentrations. Mostly relative quantities are reported, as absolute quantities require the use of standards.

NMR spectroscopy is a lot less sensitive than MS, and normally able to detect less than a hundred metabolites in most samples [50], restricted to high concentration metabolites. NMR based methods have the advantage of being non-destructive, less contaminating, and provide information to elucidate parts of the structure of the molecule directly. NMR spectra can be predicted from a structure.

Results provided by these experimental setups can be used in different ways. In Metabolic Profiling[30; 38], the resulting spectra are used to identify, and hopefully quantify, particular small molecules. In this case, it becomes relevant to have good reference collections of metabolites to be able to match the spectra against.

A different approach is Metabolic Fingerprinting[30], where the spectra/peaks are used as patterns, neglecting the identification of particular small molecules in the spectra. These patterns are normally compared to other stored patterns to identify or characterize the condition of the newer sample. In this case, there is no need for good reference layers of small molecules to compare against.

Another relevant approach is Metabolic Footprinting[77], which looks for the change in the composition of small molecules in the media, where the studied cells are growing or are in contact with, the chemical impact of the cellular system

in its environment, rather than intracellular changes. For instance, metabolomic studies of blood or urine samples are normally considered "Metabolic Footprinting" studies, as they are inspecting the outcome of many tissues metabolism in these biofluids.

Metabolomics studies can be classified as either targeted or untargeted[26]. In targeted Metabolomics, a small number of specific metabolites are monitored and quantified, which is particularly useful for clinical applications as diagnosis tests. In untargeted Metabolomics, there is no predefined hypothesis of metabolites to pursue and the aim is to identify relevant metabolites and the global biochemical change within the conditions under study. This second case requires good metabolome databases to aid in the identification of putative metabolites.

The field of Metabolomics has seen the emergence during the past 10 years of a number of experimental data repositories, such as METLIN [136], KnApSacK [135], GMD [85], and Madison Metabolomics Database [23], among others. These are relevant steps towards an accumulation of small molecule knowledge, however there is a lack of a well funded and highly centralized resources such as those for nucleotide (NCBI RefSeq, ENA, KEGG), and protein deposition (UniProt). These central resources store the nucleotides and proteins depositions, annotate them, and link them with their biological background.

### 1.4.1 Challenges in Metabolomics

In 2004, Goodacre and colleagues [48] described six different types of databases that would be relevant for Metabolomics; among them they included: "Databases listing all known metabolites for each biological species. With suitable metadata, these databases could be extended to contain temporal and spatial information."

Pedro Mendes [102] stated in 2006 three major bioinformatic challenges posed by Metabolomics: to identify the large number of metabolites detected whose chemistry is unknown, to identify the active areas of metabolism and the need for data standards. While the second challenge has been partly addressed by various studies in systems biology (specially including mRNA and protein expression studies) and the third challenge has been tackled by a number of data standards initiatives (MSI [39; 51], MIAMET [6], ArMet [66], mzData [114], mzXML [117],

etc.), very little advance has been made in the first one.

Experimentalists in Metabolomics are able to identify only a small fraction of the metabolome in a sample. With the existing chemical and metabolism databases, researchers in mass spectrometry doing profiling leave most of the detected metabolites without assignment to a known small molecule. This is the main niche that this work attempts to tackle, the study of existing and new methods for generating extensive collections of small molecules, that can accurately represent the metabolome of a sample or a species. Figure 1.3 illustrates in which part of the complete metabolomics pipeline this work lies.

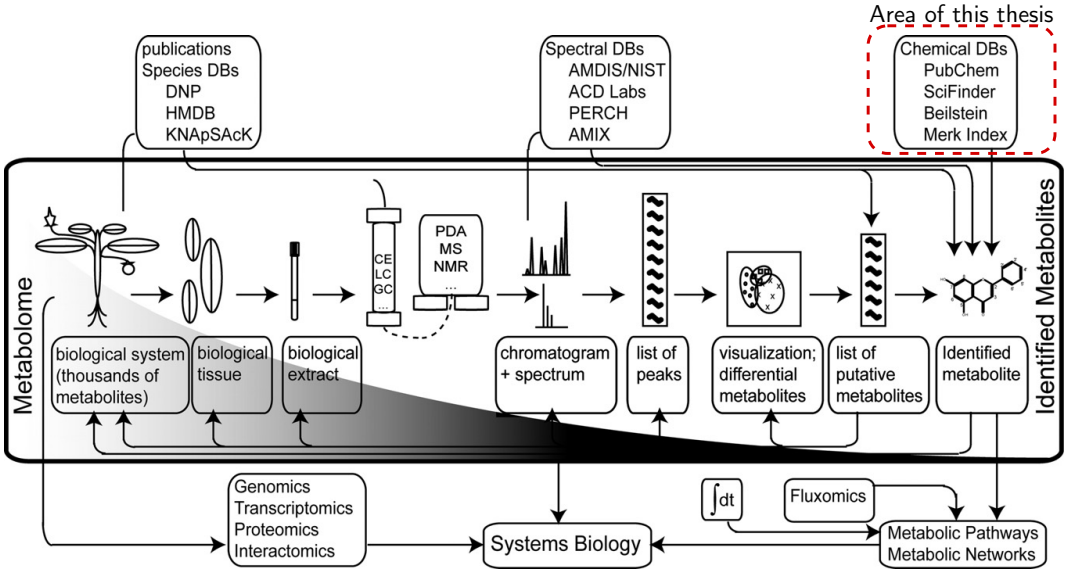



Figure 1.3: Diagram showing the main components of the metabolomics pipeline, from metabolome to identified metabolites: this work lies in the Chemical Databases part, used to identify putative metabolites, irrespective of the detection technology. The Chemical Databases part could as well have included databases like KEGG, BioCyc, and ChEBI, among others. This work focuses on generating species-specific chemical (metabolome) databases. Figure from [104], used with permission.

This work aims to generate resources, based on different types of knowledge, that can bridge the gap between Metabolomics results – putative metabolites – and the small molecules that they should be assigned to.

## 1.5 Research objectives

The main objective of this thesis is the investigation of methods to generate species-specific collections of metabolomes. Considering the huge man power that initiatives like HMDB require, an important component of our study is the generation of automatic methods that can alleviate this requirement, as there are so many organisms for which metabolomes are unknown. Today, the most common resource of small molecules, on a species-specific basis, are metabolism databases, like the Kyoto Encyclopedia of Genes and Genomes (KEGG) or BioCyc, among many others. However, this type of resources does not hold more than two thousands small molecules when restricted to particular organisms (probably, in most cases, below a thousand small molecules). In contrast, the HMDB reaches nearly ~8,000 small molecules for *H. sapiens*. Although it might be difficult to reach the amount of molecules in HMDB, a reasonable objective would be to have species-specific collections that double or triple the size of what is offered by the main metabolic resources. In other words, collections that lie in between those derived from metabolism resources and the HMDB, but that can be easily replicated in many organisms.

Towards this aim, the areas of metabolism databases integration, text mining, and chemical enumeration are visited. Initially a backbone of data is extracted and merged from the main metabolism databases. This backbone is further enriched by the use of text mining tools to add relations to further small molecules, proteins apparently related to them and the tissues and cell types in which these entities could be found. Starting from known reactions and metabolites, new putative small molecules are added through methods of enumeration. Additionally, this work explores a more elaborated case in metabolism: the production of polyketides. These interesting natural products are examples of how the combination of simple chemical transformations in non obvious orders can generate extremely complex molecules.

It is important to stress again that the work in this thesis lies – as Figure 1.3 shows – in the layer of small molecule collections, as small molecule, metabolome, and metabolism databases could be considered. Hence this work neither deal with the data analysis that each different experimental metabolomic

technology require, nor how the raw output of each technology is compared to known molecules for assignment or annotation, but one step after that, as Figure 1.3 illustrates.

## 1.6 Guide to chapters

The first work chapter, deals with merging different major metabolism resources in a species-specific manner, towards more complete, species-specific, catalogues of small molecules. I explore the difficulty of integrating different metabolism resources, decide which are the main resources to be integrated and choose a framework for doing this. The software written builds Species-specific metabolome sets for different organisms, which I compare to find commonalities and particularities between the organisms. Results for *H. sapiens* are compared against the HMDB.

The second work chapter, explores the use of text mining techniques for generating species-specific metabolomes. Following a short exploration of the area, the chapter shows the strategy implemented – based on named entity recognition and co-occurrences – to find small molecules in the literature and their relationships with organisms, proteins, tissues, and cell types. This approach relies on dictionaries of terms built from relevant resources, such as databases and ontologies. Classification methods are used to increase the reliability of the results obtained. Again, results for *H. sapiens* are compared against the HMDB.

The first two work chapters deal with small molecule knowledge which is deposited in existing resources, such as metabolism databases or chemical databases. The final work chapter, explores methods of enumerating, or predicting, new possible small molecules that under certain rationality could be part of a defined metabolome. This chapter visits the enumeration of generic reactions (reactions that have a variable part), as a learning stepping stone towards the general enumeration of reaction mechanisms; and bacterial polyketides, as an example of a difficult case were molecules are relatively unique to each organism and are "directly" encoded in their genomes.

# Chapter 2

# Metabolism database integration

## 2.1 Introduction

In this chapter I focus on the methods used to merge different databases of metabolism in an organism specific manner. I discuss briefly some minimum aspects that a metabolomes database should include, to step into a survey of databases and frameworks that could be used for the proposed integration, explaining the main features that lead to the inclusion of certain resources, the use of particular technologies and the blessing of gold standards for comparison.

The main aim of this part of the work is to generate a method to consolidate as much available knowledge as possible for the metabolome of an organism, starting with what I think is the obvious first stop: metabolism databases. I gather data for four organisms (*H. sapiens*, *M. musculus*, *E. coli*, and *S. cerevisiae*), exemplify and assess the consolidation process with one of them, and then compare this merged result with a gold standard.

Finally, I compare the merged sets of the different organisms, looking for differences and commonalities between them.

## 2.2 Desired elements in a metabolomes database

A database that hold metabolomes should include a number of other biological entities besides mere small chemical molecules. A biological context to these

small molecules is essential to integrate the knowledge existing in other fields of biology into assembling complete metabolomes. At least the following entities should be present in some form:

**Biochemical reactions:** They link small molecules to a particular enzyme, that needs to be present in the organism's genome so that the reaction is considered to occur in the organism. Biochemical reactions are normally indexed by the IUBMB Enzyme Nomenclature number, or EC Number [152]. Biochemical reactions are essential as they encode the chemical mechanisms that the reactome is able to exert on the metabolome – allowing future enzyme promiscuity studies – and because they link small molecules to enzymes, for which there is so much localization knowledge that could be then transferred to chemical entities.

**Enzymes:** The proteins in charge of catalyzing the reactions are invaluable sources of knowledge for assembling a complete metabolome. They are a gateway to gene/protein expression experiments, that allow one to characterize a metabolome not only at the level of the organism, but also at the level of tissues, cell types and/or conditions. For each enzyme, an external identifier (such as a UniProt [19] identifier) should be included.

**Tissues:** One of the main aggregated containers in higher eukaryotes and the target of many metabolomics and gene expression experiments. Tissues should be vocabulary controlled through a hierarchical organization, probably an ontology such as the BRENDA Tissue Ontology [49], to allow for general overviews as well as detailed granularity analysis of the data composition.

**Cell types:** Although of different granularity, serves for this purpose similarly to tissues, and should be controlled likewise.

**Organisms:** The aim is to build organism specific collections, which can then be compared at different taxonomic level. Organisms are naturally organized through resources like the NCBI Taxonomy [36].

**Conditions:** Metabolism is highly dynamic, probably the fastest changing of all omes. As such it is relevant to be able to store meta data about the treat-

ment, disease, environments, etc. to which a biological sample is exposed. As with Tissues and Cell types, which could be considered a part of the conditions vector, conditions should also be controlled through an ontology where ever possible.

### 2.2.1 Small molecules

The electronic collection of small molecules ($m < 1500Da$) from the different databases visited during this work divides data in three levels according to completeness: small molecules with chemical structure, small molecules with generic structures and small molecules with no structures.

The first and ideal scenario happens when both the database contains chemical structures for the referenced small molecules and the particular molecule is completely defined. Example of this is pyruvate in databases like ChEBI or KEGG.

Generic structures represent molecules that have variable parts in their structures. Databases that have chemical structures for their referenced small molecules normally include them. "2-oxo monocarboxylic acid" in ChEBI or "a long chain fatty acid" in KEGG are examples of these type of molecules. Most databases store these generic structures in a representation called "markush structures", which places pseudoatoms or "R-groups" – denoted normally by R or other letters not used for any element – to mark variable parts of a molecule. Figure 2.1 shows an example of a markush structure for 3-sn-phosphatidate, from BRENDA.

Finally, chemical entities with no structure are present in databases with chemical structures, when the represented molecule is a class of molecules that is too general to be represented by a markush structure. Examples would be "an oxoacid" from ChEBI or "Lipid" from KEGG. However, this also happens when databases simply do not include chemical structures, like IntEnz [43], or, in the worst scenario, have missing structures for some molecules, as it happens in many cases with BRENDA.

Small molecules stored should include structure and chemical identifiers when ever possible. Identifiers allow quick search and comparison within a database.

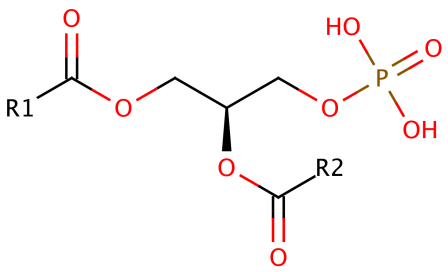



Figure 2.1: A markush structure representation for 3-sn-phosphatidate, from BRENDA

The preferred chemical identifier should be the InChI string. InChIs are a one-line text representation of the small molecule's structure, which allows to identify a molecule with a single text line without the need to interpret its structure. Given that for database indexing purposes it is useful to have fixed-length fields, the InChI string can be encoded, through a hashing function, into a 27 length string called the InChI Key. It is not possible to go from an InChI Key back to a complete InChI, but as long as the association is maintained within the database this is not a problem. It is certainly faster to search (and index) a table through InChI keys than through complete InChIs. Optionally, a third component, the AuxInfo string, accompanies the InChI string. To produce the original chemical structure from an InChI string, both the InChI string and the AuxInfo string are required.

The only drawback of the InChI identifier is that it requires a completely defined structure to be calculated, so InChIs cannot be calculated for generic molecules (markush structures), which include an unknown or variable portion of the molecule.

For generic molecules where InChI cannot be calculated, the SMILES identifier is the next best choice, taking care of calculating SMILES always with the same implementation. SMILES can be calculated for generic molecules (having an R-group). There is an extension of SMILES – known as SMARTS – capable of a higher level of expression for generic structures, allowing to set restrictions to the R groups.

For each small molecule stored, references to external databases, like ChEBI [25], KEGG[72], PubChem Compounds [160] or ChemSpider [118], among others, should also be stored. This is particularly useful for molecule classes with generic structures, as the classifications in these databases might allow to find instances of them, that could be latter checked to see whether they could belong to the organism in question.

### 2.2.2 Reactions

Reactions are a central part of the data found in metabolism databases. They are of extreme importance as they join the chemical world with the biological world, linking small molecules and enzymes. Biochemical reactions have been systematically classified since 1961 through the International Union of Biochemistry and Molecular Biology (IUBMB) Enzyme Commission number[152], widely known as EC number. Metabolism database use EC numbers to classify reactions, often as well storing proposed EC numbers that the Enzyme commission blesses when enough experimental evidence exists. Although widely used for classification of reactions, EC numbers many times do not refer to a unique reaction but to small classes of reactions (producing one EC to many reactions relations).

Reactions found in metabolism resources can be classified in different types according to their participants. Some of them have implications in the way that data needs to be handled to make use of them.

We distinguish normal reactions, where all small molecules are defined, from generic reactions, where markush structures represent certain participants. Some generic reactions also represent polymeric reactions (like DNA bases elongating a DNA stretch, or single sugars being added to a polysaccharide). Polymeric reactions pose the issue that they are normally represented with the same participant at both sides, denoting its length in often unconventional ways.

Although not exactly generic structures, sometimes reactions with protein participants (as substrates and products, not as catalyzers) are also represented through the use of generic reactions, where markush structures represent the portion of the protein that reacts with small molecules. This is particularly difficult to handle as frequently databases do not provide any semantic indication that

those participants are actually proteins, and that the R-groups do not represent really a variable chemistry, but the rest of the protein involved. This leaves to the user the responsability to tell protein participants apart from regular generic chemical molecules. It is also common that protein participants are only represented with a name, which at least does not produce the false artifact of a generic molecule.

When retrieving small molecules from metabolism databases of biochemical reactions, normally the approach is to look for reactions that are catalyzed by known enzymes in the organism of interest. One exception to this rule are spontaneous reactions, which are known to take place at biochemically relevant rates without the need of a catalyzing enzyme. The transformation of Trypanothione to Trypanothione disulfide in the presence of dehydroascorbate[1], or the dissociation of carbamate into carbon dioxide and ammonia[2], are examples of spontaneous reactions. To consider a spontaneous reaction part of a biochemical system, the main requirement would be that at least one side of the reaction participants (small molecules) are known to be part of the system. In the whole of KEGG reactions (that is, including all organisms), there are 64 reactions labeled as spontaneous. The MetaCyc database (compilation of all the BioCyc databases reactions with experimental support) contains 188 spontaneous reactions.

Spontaneous reactions should not be confused in metabolism databases with reactions that, for a given organism, do not have an enzyme assigned. This normally happens when the enzymatic activity is known to exist in the organism due to experimental studies, but no enzyme from this organism has been isolated as responsible for this enzymatic activity. Reaction acetoacetate decarboxylase (EC 4.1.1.4) – where acetoacetate is broken to form acetone and carbon dioxide – is an example of this. In the words of the Enzyme Commission: *"While no acetoacetate decarboxylase (AADase) gene has been identified in the human genome, an AADase enzyme activity has been purified from human serum"*. Some resources, like the BioCyc databases, assign pathways (from a relevant taxonomic range) to an organism if the organism has most of the enzyme for those pathways. This can as well lead to reactions in the organism with no enzyme assigned.

[1] Reaction R07316 in KEGG
[2] Reaction R08358 in KEGG

When a reaction occurs through a cell or organelle membrane and there is a mass transfer through it, this is called a transport reaction. In general there is much less knowledge about transport reactions than there is of regular, same compartment, reactions. EC numbers do not deal with multi compartimentalization of reactions. A more recent classification, the Transport Classification Database[10], or TCDB, can be used for this. The TCDB is far less widely adopted than the EC number classification.

## 2.3 Methods for database integration

### 2.3.1 Metabolism database selection

There are many metabolism databases from where pathways, reactions and small molecules can be obtained. However, one should be selective, as each resource requires additional integration and validation efforts, and not all of them provide additional data, as some resources are built by simply merging other databases. In the following sections I explore a number of resources to decide which ones to integrate.

#### 2.3.1.1 KEGG

The Kyoto Encyclopedia of Genes and Genomes[72], better known as KEGG, is one of the most widely used bioinformatics databases available[1]. Developed since 1995 by the Kanehisa Laboratories in Japan, it is a multi organism database that holds genome sequences, annotated genes and proteins, assignments from enzymes to reactions, small molecules, a comprehensive collection of pathways and ontologies to organize these entities. The knowledge in KEGG is stored in 15 different internal databases, eight of which are directly connected to metabolism (PATHWAY, DRUG, COMPOUND, GLYCAN, REACTION, RPAIR, RCLASS and ENZYME), giving KEGG a strong bias towards metabolism. On its version 57, of March 2011, KEGG included 1,389 different organisms, nearly 6 million coding sequences (CDSs), 144 pathways, 8,393 biochemical reactions and 16,413

[1] According to Web of Knowledge, the 4 more cited papers about KEGG have together more than 440 citations per year

small molecules[1], among others. KEGG was one of the first metabolism databases to include actual chemical structures for its compounds. Data in KEGG receives a combination of automatic processing and manual curation. Most of the manual curation in KEGG is at the gene and protein annotation level. Once this is done, the metabolic information is generated automatically, using EC Numbers as the main key between the genome and biochemical reactions.

Besides the data resources, the Kanehisa laboratories have made available through KEGG a number of very useful and interesting tools for the study of metabolism. For instance KEGG E-zyme [167], predicts new biochemical reactions between any given two small molecules, using the knowledge of reaction pairs mapping stored in the RPAIR database.

Until July 2011, the data in KEGG was completely available for download and redistribution. Unfortunately, a major funding crisis lead to a closure of some of the services provided by KEGG, terminating the bulk data access through FTP. This work, as many other initiatives, was left then only with what was made available by KEGG before the shutdown. There is still a programmatic access through SOAP web services that allow certain level of bulk querying, but for the size of a resource like KEGG, massive integrations like the one aimed on this thesis cannot be done through web services access only.

#### 2.3.1.2 MetaCyc/BioCyc

Following in importance to KEGG are the MetaCyc/BioCyc[11][2] collection of databases. These databases are organism-specific (one database per organism), with the exception of MetaCyc, which is an aggregated database comprised of all the experimentally verified or manually collected elements existing in the organism specific databases.

A Cyc database is normally constructed starting from the annotated genome sequence of an organism. In the coding sequences (CDSs) annotation, the software used to build these databases, Pathway Tools[74], recognizes EC numbers

[1] This only includes the COMPOUND database section, it does not include GLYCANS, DRUGS and other sections

[2] According to Thompson-Reuters Web of Science, the four most cited BioCyc articles make together an average of 104 citations per year

and enzyme names. These annotations are then compared to MetaCyc, where enzymes and EC numbers have mappings to biochemical reactions. In this way, Pathway Tools associates biochemical reactions to the protein products of the organism. Using the assignment of reactions, the tool transfers pathways to the new organism whenever reactions of that pathway were mapped. This also introduces new candidate reactions that could be present in the organism.

Given the construction process, a Cyc database contains annotated genome sequences, annotated coding sequences, annotated proteins (and particularly, enzymes), reactions, small molecules (including structures) and metabolic pathways. Additionally, some of the manually curated Cycs include allosteric enzyme regulation and transcriptional regulation pathways. EcoCyc[78], for instance, is a good example of a database with such a level of completeness.

The BioCyc databases are separated into 3 tiers, or categories, according to the level of human curation that they have undergone. Tier 1 databases, comprising MetaCyc, *E. coli* (EcoCyc), *H. sapiens* (HumanCyc), *A. thaliana* (AraCyc) and *S. cerevisiae* (YeastCyc) represent the highest level of quality, some of them reaching a man-decade time of manual curation. Tier 2 databases are the second category of data sets, which have a moderate level of manual curation. By the end of 2011, SRI reported to have 34 databases at Tier 2 level. Finally Tier 3 groups only automatically generated databases, approximately 1,653 databases. In some rare exceptions, these databases can consist of aggregations of organisms, like the case of PlantCyc, an aggregation of 350 plant species, probably a very valuable resource for the study of natural products.

Pathway Tools is distributed under an academic license, allowing groups around the world to create their own Cycs or PGDBs, as called in the context of the local installation. This PGDBs can be easily submitted back to the main SRI repository.

The Pathway Tools software is able to load downloaded BioCyc databases[1] and export them in plain text files for later parsing. However, there are certain aspects, like small molecules structures, that are not made available by the export mechanism, and are only accessible through the Lisp API provided by Pathway Tools. The need to access certain parts of the data through Lisp can be seen as

[1] Academic license required for the download

a disadvantage for BioCyc and Pathway Tools, as this is a rare expertise in most bioinformatics groups. Until very recently, the BioCyc databases could not be accessed through a programmatic web service[1].

#### 2.3.1.3 BRENDA

The Braunschweig Enzyme Database[131], or BRENDA, is a comprehensive multi-organism metabolism resource that focuses on enzyme kinetic parameters and biochemical reactions. BRENDA is recognized within the community[2] due to its extensive catalogue of enzyme kinetic parameters, reactions and their variations. These are mostly collected manually from the literature, and hence of very high quality.

By the end of 2011, BRENDA housed 5,536 reaction types with different EC numbers, including 297 preliminary BRENDA EC Numbers, more than 100,000 $K_M$ values[3] and more than a million small molecules (enzyme ligands), more than 85,000 with chemical structures. BRENDA has been developed since the early 2000's by the group led by Prof. Dietmar Schomburg.

Unfortunately the BRENDA general distribution policy is poor, and the only way to get the bulk data set is through a monolithic, and badly constructed, text file which has no internal identifiers for small molecules. This leaves the task of associating the small molecule name given to an actual structure to the researcher. This file is seldom updated. Furthermore, chemical structures are not easily available for download.

During the last part of 2011, the BRENDA SOAP web service was improved, exposing much more data. The web service allows one to clarify parts that are ambiguous in the downloaded file.

[1] A programmatic web service exposes an Application Programming Interface (API) through the web, which means that programs can be written that interact through the network with the database that it is exposing this service. Normally used for bulk querying or massive data access.

[2] According to Web Of Knowledge, the 4 most cited BRENDA articles have a combined average of nearly 50 citations per year

[3] $K_M$ is the concentration of substrate at which the enzyme achieves a reaction rate of half its maximum possible velocity, $V_{max}$.

#### 2.3.1.4 Reactome

Reactome[99] is a high quality, manually curated, metabolic, information transfer and signalling pathway database focused specifically on *H. sapiens*, and to some extent in *mammals*. It has been extended to other taxonomic ranges but only in an automated fashion. This produces a quality gap between its main organisms (*H. sapiens*) and the ones belonging to other taxonomic ranges, as the main source of knowledge for the automatic reconstruction tends to be very narrow from the phylogenetic point of view, in contrast to what happens in the case of MetaCyc, which is derived from many organism's experimental results. Reactome has been developed by a consortia of life sciences research institutes and universities continuously since 2004.

From all the resources presented, Reactome features the most extensive, encyclopedic articles on most of the topics that covers (mostly for *H. sapiens*). The database is extensively cross referenced to key resources such as UniProt (for proteins) and ChEBI (for small molecules) among others.

Reactome provides a wide range of bulk access modes, from MySQL dumps[1], different pathway exchange format standards (SBML, BioPax) and also a SOAP web service access. It is worth mentioning that a text book version of all of the articles can be retrieved. The web pages of Reactome are of very high quality in terms of information content: Reactome offers an excellent view of metabolism for the interested researcher. Most of the diagrams displayed are drawn in CellDesigner[45], one of the best tools for assembling SBML models with compliant layout, and include cellular compartmentalization when adequate.

However, due to the narrowness of the phylogenetic spectrum of this database, I did not to include it. Although other organisms are provided, this data is mainly automatic mappings to *H. sapiens*, decreasing reliability for other species. It was not included as a gold standard model for validation of *H. sapiens* metabolome since the coverage of small molecules that it has is much lower than the Human Metabolome Database, HMDB. However, if the aim would be to build a good metabolism resource for *H. sapiens*, Reactome should be definitely included, due to the depth in which metabolic processes are annotated.

[1] A database dump is an entire copy of the database in a flat file, that can be used with the same engine to recreate the database completely in another location

#### 2.3.1.5 BioSystems

NCBI BioSystems [46] is a relatively young resource in the metabolism database landscape. It is mainly an aggregated resource that stores entries of genes, proteins, metabolic pathways, diseases, reactions and small molecules existing in the main metabolic resources mentioned. Entries are linked by the resource to biochemical objects present at the NCBI internal databases: organisms are normalized with NCBI Taxonomy[36] ids; genes and proteins to Entrez Genes and Entrez Proteins respectively; diseases to the OMIM database; literature entries to PubMed and Entrez Books; and finally small molecules to PubChem. As of December 2011, approximately 5.4 million proteins, 1.9 million genes and 9,241 small molecules were annotated as part of the BioSystems collection. To the best of my knowledge, to this date, this have been derived mostly from KEGG, EcoCyc and Human Reactome. These entries are linked to presumably[1] 1,664 different organisms.

The NCBI BioSystems compendium can be bulk accessed through the NCBI E-Utils web services[150] and through FTP access. However, since this resource is mostly a compendium and not a database that generates its own entries, I decided to leave it out of the main metabolism database integration. It shall prove useful in the later text mining chapter.

#### 2.3.1.6 Pathway Commons

Pathway Commons [12], another aggregated database, gathers primarily data sets of interactions between the main biochemical macromolecules. As of early 2012, it provides data coming from 9 different sources (of interest for this work, only 2, HumanCyc and Reactome, the rest of them mostly concerning protein and gene interactions). The resource has an important bias towards *H. sapiens* data sets and networks. It includes biochemical reactions, but this is not the focus.

The resource provides adequate bulk access to its data through a SOAP web service and downloads. All the data aggregated are made available in BioPax and

[1] They are linked to 1,664 different NCBI Taxonomy identifiers, however, some of these might refer to a genus instead of a species, which might change this figure, although in general it is in line with the number of sequenced organisms to date

SIF[1], which is convenient if the data includes what one is after, since it reduces considerably the parsing efforts.

Pathway Commons is mainly built using cPath[13], an open source solution for parsing, storing and querying most of the data sets made available at Pathway Commons. This would be a very useful tool for a project which requires integration of different interaction data sets.

Given the main focus on protein-protein and gene interactions, the fact that this is a database merger which does not add further data of its own and that is heavily biased towards *H. sapiens* data sets, I decided that this resource is not adequate for the integration of metabolism resources described in this work.

#### 2.3.1.7 WikiPathways

Surprisingly, WikiPathways [120] has a reasonably high visibility in the community[2], despite being relative recent. The resource is built by the collaborations of it members, as in any Wiki platform, which to January 2012 was of around 2,000 members. Currently the platform includes 19 organisms, which are linked to a total of 45,000 UniProt entries, 974 distinct ChEBI and 1,730 PubChem Compounds small molecule entries.

Although a promising database to be watched in the coming years, this resource still is small in comparison to the other major metabolism projects, which is partly reflected in the amount of organisms and small molecules referenced. WikiPathways has the tremendous value though to be a potential repository of many new pathways that might not be present in other resources, as it operates by manual input of pathways. Still, given its size, it is probably not worth to integrate it in the current project.

#### 2.3.1.8 Rhea

Rhea[3] is an organism-independent database of biochemical reactions. The database is derived from an earlier resource, the Integrated relational Enzyme Database (IntEnz)[43], which is the source of the Swiss Institute of Bioinformat-

[1] Simple Interaction Format

[2] The two main articles considered, the resource has an average of 20 citations per year

ics (SIB) ENZYME database[5]. The ENZYME database reflects the decisions of the Nomenclature Commission of the International Union of Biochemistry and Molecular Biology (NC-IUBMB), which are the definitions of EC Numbers and the reactions they represent.

Rhea has most of its molecules referenced to the ChEBI database, providing high quality chemical structures for more than 17,000 reactions as of December 2011. These reactions are cross referenced to 3,890 different small molecules from ChEBI, and nearly 140,000 different proteins present in UniProt. Reactions are also cross referenced against major metabolism resources such as KEGG, MetaCyc, and Reactome, among others.

Rhea contains reactions standardized to pH 7.4, their EC number assignments, and the small molecules that participate in the reaction. Rhea does not contain pathways information. The resource is manually curated by researchers at the Swiss Institute of Bioinformatics (SIB), who also add new reactions to Rhea. The European Bioinformatics Institute (EBI) develops the software layer of Rhea.

#### 2.3.1.9 The Human Metabolome Database, HMDB

Researchers at the Wishart Group in Alberta, Canada, manually compiled the Human Metabolome Database (HMDB)[166], comprising ~8,000 small molecules with evidence to be present in the human body. They derived this resource mostly from peer-reviewed literature results, text books and experimental results.

The authors manually annotated the molecules with several fields regarding the biological source of the samples and with a detailed description of context and role of the chemical entity within the Human body. There is no other resource that matches the HMDB both in its coverage of the Human Metabolome, or in its level of annotation.

Given that one of my main aims is to build a resource in the most automatic possible way, so that it can be replicated to many organisms, we need a validation data set, a gold standard. Throughout this work I explore different methods of generating collections of metabolites in an organism specific manner. For these reasons, and for the quality of the resource, is that I use HMDB as the gold standard for a single-species Metabolome. All the results from our methods are

at some point compared against this resource.

### 2.3.2 Schema decision

I explored a number of alternative database schemas for housing metabolomes with adequate connection to their biological context. A first need is to be able to store Enzymes, Reactions and their participants, as metabolism is deemed as an important connection between small molecules and organisms. It also needs to be flexible enough to house other relationships, like evidence that a protein or a small molecule are known to occur in certain cell types and/or tissues. It is important as well to be able to store sets of data belonging to different organisms, having a way to separate them, and different sources, even for the same organism.

From the technical perspective, such a resource should have its storage backend in a relational database engine, such as MySQL, PostgreSQL or Oracle. It should have a persistence layer[1] based on a technology such as Hibernate or equivalent, in a language such as Java, Perl, Python or C++. Preferably in Java, which allows a better integration with cheminformatics tools such as the Chemistry Development Kit (CDK) [141] or Chemaxon's JChem library[22], required for handling small molecule structures, and bioinformatics packages like BioJava[56], SBML[59] and MIRIAM[70], that aid respectively in sequence manipulation, metabolic model imports/exports, and external database identifiers consolidation. It should also be a project relatively tested by the community and with some level of visibility, for instance a reasonable number of citations or an active user community.

Most of the Warehousing solutions for bioinformatics before 2009 tend to focus on genomics and proteomics, giving little attention to metabolism, and virtually none to chemical entities participating in metabolism.

One of the first widely used warehouse systems was FlyMine[94], developed within the *D. melanogaster* community to house data from this model organism and other insects. This initiative focused on genomics and proteomics data, including sequence annotations, microarray results, protein-protein interaction net-

[1] A persistence layer, in this context, refers to a software layer that encapsulates the interaction with the database for client programs, exposing functionality through an application programming interface, API

works, diseases information and transcriptional modules. FlyMine did not include particular information about enzyme catalyzed reactions nor small molecules.

Another warehousing system that had attention from the community was Atlas[134], which had a much broader scope in terms of organisms. Like FlyMine, it was mainly focused on genomic and proteomic data. It had a particular focus on protein-protein interactions. Atlas again had no explicit support for biochemical reactions nor small molecules.

The GMOD bioinformatics community released in 2007 the CHADO[106] warehouse system. As most of the other systems, focused in genome annotation and features data. This system had the particularity of handling data in an ontology oriented fashion, which should allow it to handle new types of data without major schema modifications. This has not been translated, to the best of my knowledge, into a support for metabolism related data.

The BNDB[87], the Biochemical Network Database, was one of the first warehouse solutions to consider metabolic pathways data as well as genomic and proteomic data. It also had a strong focus on protein - protein interaction networks and general network visualization. Its main persistence layer access was written in C++, although a Java API is also available, but apparently only for its visualization package. Unfortunately, BNDB was scarcely tested by the community, averaging less than two citations per year since its release[1] and the release history of the software shows little activity for the past years.

ONDEX[84] is a graph-based visualization analysis system which has a backend that in many respects resembles the database warehouse concept that this work requires. Much more oriented to networks than all the previous efforts, this system apparently manages to integrate metabolic pathways data with other network types and expression results, with the aim of visualizing it in an integrated manner. Unfortunately, there is little technical documentation regarding the database back end and the persistence layer API that would allow to make a reasonable judgement regarding how much of our problem it can cover.

BioWarehouse[89] is a bioinformatics warehouse that focuses primarily on pathway-centric resources, focusing on integrating resources such as KEGG or BioCyc. As most of the mentioned resources, BioWarehouse needs to be deployed

[1] According to Thompson-Reuter's Web of Knowledge Citation report for its main paper

on a RDBMS system, in this case either MySQL or Oracle database engines. It has explicit support for pathways, reactions, enzymes and chemical entities. It has loaders for various databases, including KEGG, BioCyc databases and NCBI Taxonomy, allowing an adequate organization of sets of data by organisms. It has a very good visibility[1], nearly as good as FlyMine, and the project has been actively developed since 2006[2]. The persistence layer, although available in Java and a few other languages, is not based in hibernate or other suitable standard, but can be reasonably extended as it is open source and well documented. Furthermore, BioWarehouse is released by a group with wide experience in metabolism resources, Peter Karp's group at SRI, the same group responsible for Pathway Tools and the BioCyc/MetaCyc suite of databases.

For the reasons and characteristics discussed for each of these warehouse alternatives, I decided to use BioWarehouse as our central repository for metabolome collections.

### 2.3.3 BioWarehouse Schema Structure

The BioWarehouse database schema is relatively large, and exploring it completely would go beyond the needs of the chapter. For this reason, Figure 2.2 presents the main aspects of the BioWarehouse database schema that are useful to this work, explained in the following paragraphs.

BioWarehouse organizes different sources of data through the DataSet table. A DataSet can hold multi organism or single organism sets of data, depending on the loading process. Each source of data is uniquely identified with a DataSet WID (WID stands for Warehouse Identifier), which is a numeric index.

"Object" tables hold different biological and meta data objects, such as proteins, chemicals or reactions. As each object has different attributes, there is one table per each type of "object" represented. All the "object" tables reference each of its entries to a DataSet (so there is a non-identifying one to many relation going from DataSet table to each object table, connecting the WID of the DataSet

[1] According to Web Of Knowledge, BioWarehouse has an average of nearly 7 citations per year, the highest we found among these type of resources after FlyMine, with more than 8.

[2] Development seems to be interrupted in terms of funding by the end of 2010, reaching version 4.6.1

with the DataSetWID field of each object), as each instance in an object table comes from a source data. Elements in an "object" table are uniquely identified as well by a WID (WIDs are unique database wide, so no two objects, regardless of the source have the same WID). Any table that has WID and DataSetWID fields is an "object" table. Besides the object tables Protein, Reaction, Chemical and ChemicalStructure, which are self explanatory, other relevant objects are EnzymaticReaction, BioSource and Taxon. EnzymaticReaction associates Reaction objects and Protein objects (normally the catalyzing enzyme) to represent catalyzed biochemical reactions. BioSource table stores biological containment units (tissues, cell types, organisms, diseased cell types, cellular locations, etc). The Taxon table stores organisms taxonomic organization, which is essentially derived from NCBI Taxonomy.

A different type of table are the linking tables, which store relations between objects or additional vectors of attributes for objects. An example of this is the very important CrossReference table, which links to any "object" table through its OtherWID field. This table stores, for instance, references to ChEBI or PubChem Compounds identifiers (among others) for elements of the Chemical table or UniProt identifiers for elements of the Protein table. The CrossReference can also bind this external identifier to its object in the warehouse if it is present. For instance, one could load the entire UniProt database, then a protein for a set in KEGG that has a UniProt identifier could be linked through the CrossReference table to the actual protein object in Protein table from UniProt that is loaded in the warehouse. Other examples of linking tables are DBID, which stores the identifier of the object from its source database (like the KEGG COMPOUND identifier, C00005, for NADPH in the KEGG DatSet), or the SynonymTable, which stores name synonyms for objects of the different types.

The following section details which tables of the schema were part of the improvements I did.

#### 2.3.3.1 Improvements and modifications to BioWarehouse

BioWarehouse is accompanied by a number of parsers for different knowledge bases (KEGG ligand and the BioCyc collections among others). The aim of these

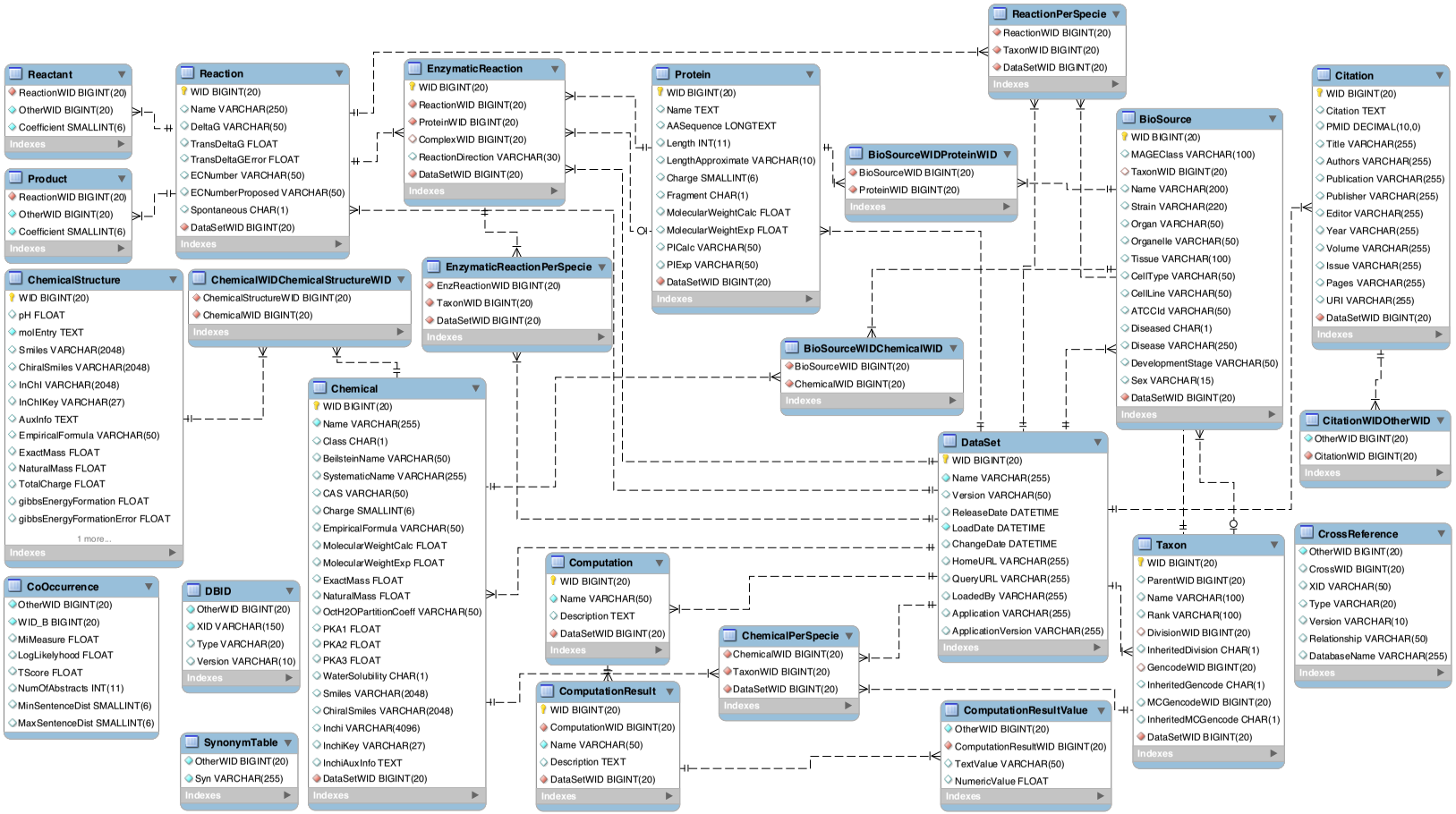



Figure 2.2: The schema shows the used portion of biowarehouse, including the additional tables and fields added, detailed in Section 2.3.3.1.

parsers is much wider than biochemical reactions and small molecules, leaving behind some Chemical detail that falls out of the scope of the resource. For this reason, I improved some of these parsers to include additional features of interest that were not considered by the original authors. For some other databases, such as the HMDB, which was loaded for results validation, or BRENDA, I wrote parsers in Java, as none were available.

The schema lacked some attributes that were desirable for storing metabolomes:

**Chemical Structures** An additional table added to store the molecules as complete MDL MOL V2000 representations. All the databases to be unified export chemical structures in this format. This enables a quick loading into CDK molecule objects, for any necessary cheminformatics computation. The MDL MOL V2000 data were saved in the database as it came from different database sources. In this way, parsing errors only influence computations (such as molecular formula, descriptors, mass, protonation states, etc.), but the stored structure remains as obtained from the source and calculations can be redone if parsing errors[1] are detected. I added a many-to-many relation table to link the entry in Chemical table with the entry in ChemicalStructure table. A many to many relation gives the flexibility of having multiple Chemical table entries pointing to the same molecule and also to store standardized versions of molecules at particular pH values.

**InChI** The original schema did not consider the use of InChI and InChI Keys, which are useful for searching for molecules in a text-only manner. InChI has the advantage, compared to other forms of one line molecule serializations, that there is only one implementation of the algorithm to calculate them. In other cases, such as SMILES, the output might depend on the implementation used. The only possible variation of InChIs is on the version of the software used and the parameters. I added InChI, InChI Keys

[1] During the course of the work at least 3 bugs were submitted to the CDK project about parsing errors of the MDL MOL V2000 reader module.

and the AuxInfo string fields to the original table Chemical and to ChemicalStructure.

**PerSpecies tables** For multi organism data sets like KEGG or Rhea, we need the ability to retrieve all the reactions and small molecules for a particular organism. Depending on how this different data sources are built by their original authors, the steps required to enumerate those reactions and chemical entities for a particular organism will vary. These tables (EnzymaticReactionPerSpecies, ReactionPerSpecies, ChemicalPerSpecies) allow to store the relations once computed, between those objects and the organism, simplifying and standardizing any later queries per organism.

**Transformed Gibbs Energy of Reaction** In the Reaction table, I add a field to store the Standard Transformed Gibbs energy of reaction. This is suitable for the thermodynamic analysis of a set of biochemical reactions, as it considers for the calculation of the energy the effect of pH and ionic strength. Gibbs energy calculations that do not incorporate these effects produce an energy value that is correct for zero ionic strength aqueous solutions, but they do not reflect the true state of reactions within the cell.

**Text mining co-occurrences** In the later chapter of text mining, we shall add a data set of small molecules that are retrieved for the organisms of interest. This will include relations to containers (cell types and/or tissues), represented as BioSources within BioWarehouse, and have associated scores. This could not be stored in the original schema.

Table 2.1 shows data sets I loaded into this modified schema, and after the considerations presented in 2.3.1.

For each type of data source, in this case depending more on the database type than on the organism variety, different strategies had to be used and different shortcomings solved. I inspect the different particularities of each type of data set in the following part.

| Data set | Version | URL | Comment |
|---|---|---|---|
| NCBI Taxonomy | 2009-04 | www.ncbi.nlm.nih.gov/Taxonomy | |
| KEGG | 57 | www.genome.jp/kegg | LIGAND |
| HumanCyc | 15 | www.biocyc.org | |
| EcoCyc | 15 | www.biocyc.org | |
| MouseCyc | 1.36 | www.biocyc.org | |
| YeastCyc | 14 | www.biocyc.org | |
| BRENDA Human | 2010-02 | www.brenda-enzymes.org | |
| BRENDA E.coli | 2010-02 | www.brenda-enzymes.org | |
| BRENDA Mouse | 2010-02 | www.brenda-enzymes.org | |
| BRENDA Yeast | 2010-02 | www.brenda-enzymes.org | |
| HMDB | 2.5 | www.hmdb.ca | |

Table 2.1: Data sets loaded into BioWarehouse.

### 2.3.4 Database loading

#### 2.3.4.1 KEGG Data set

The included BioWarehouse loader for KEGG parsed and loaded KEGG ligand v57, after minor fixes to its source code, which only supported slightly older releases of KEGG. Considered data includes all the chemical entities present in KEGG COMPOUND as well as those in DRUG and/or GLYCAN that participated in reactions.

I linked each organism in KEGG, represented in BioWarehouse as entries in the BioSource table, to the appropriate NCBI Taxonomy species, which are represented as entries of the table Taxon. At the same time, I linked each protein in KEGG to the its organism BioSource table entry. Proteins relate to biochemical reactions in BioWarehouse through the EnzymaticReaction table. Through this chain of links, one can obtain most of the reactions (and small molecules) that should be part of the metabolism of a particular organism stored in KEGG. This process accounts for all the reactions and small molecules for which KEGG has direct citations or experimental evidence that links them to a protein (an enzyme normally).

In the case of KEGG, reactions can be additionally assigned to enzymes through orthology families. An orthology or protein family gathers together proteins from diverse organisms that show a certain degree of conservation at the

level of sequence and functional domains, given certain constraints (for particular implementations of the orthology concept see [67; 71; 149]). For enzymes, these families will have assigned an EC number (associated to a reaction), because there is evidence of this for some members of the family. For enzymes that do not have a direct assignment of a reaction due to lack of experimental evidence, KEGG assigns a reaction through orthology families. Using orthology families is easy to make incorrect assignments of EC numbers to an organism. For instance, for *H. sapiens* as many as a hundred EC numbers could be assigned that do not correspond to the mammalian taxonomic range, so these should always be checked against a reference resource such as UniProt, to ensure that the EC number has been annotated to a protein of the organism. I managed to increase in more than ~700 metabolites the KEGG *H. sapiens* collection of small molecules through orthology, checking that the EC number and orthology assigned to the enzymatic reaction exist in *H. sapiens*.

Multi organisms data sets require special handling of spontaneous reactions when assembling single organisms metabolomes. Spontaneous reactions in the database cannot be assigned directly to the organism of interest, as they might be in the data set due to other organisms. To add the adequate spontaneous reactions, and avoid adding chemical entities for which there is no real evidence, I only add those spontaneous reactions where all participating chemical entities of one of the sides of the reaction has been previously identified as part of the organism (because of direct evidence or orthology). To address the unlikely case of consecutive spontaneous reactions, I repeated this process until no further spontaneous reaction is added, although for the data sets tried so far, I have not found consecutive spontaneous reactions.

I store all the identifiers for Reactions, Enzymatic reactions (which link reactions and enzymes), and participating small molecules assigned to an organism of interest by any of the mentioned methods in the PerSpecie set of tables, for easy retrieval and standardization to other data sets.

#### 2.3.4.2 BioCyc organisms databases

As mentioned previously, I added to the warehouse the BioCyc collections for *H. sapiens* (HumanCyc), *E. coli* (EcoCyc), *M. musculus* (MouseCyc), and *S. cerevisiae* (YeastCyc). Differently to KEGG, each of these are separate data sets, simplifying the treatment of them at the organism level. I linked each data set to its NCBI Taxonomy identifier for the adequate organism.

For BioCyc databases, the chemical structure of the participating chemical entities is not part of the regular export files. I wrote Lisp scripts that interacted with Pathway Tools[1] through its API to extract the chemical structures in MDL MOL V2000 file format. The same was done for generic molecules present in the reactions, which in the normal exported data files are not registered as regular chemical entities.

#### 2.3.4.3 BRENDA organisms databases

The BRENDA database had no parser available to be uploaded to BioWarehouse. I wrote a parser that uploads all enzymes, reactions, small molecules and tissues/cell types occurring for a specified organism in the BRENDA data file. Figure 2.3 shows the execution order that the parser follows.

BRENDA is the only one of these resources that adds information about localization within the organism, mostly organs, tissues and cell types. I normalize[2] the localization vocabulary using the BRENDA Tissue Ontology (BTO) [49].

Free text tissues and cell types names present in the BRENDA data file many times had slight differences to the names in the ontology, so I used string normalization and Levenstein distances to compare these and make assignments to elements of the ontology. The program accepted automatically matches with $d_{Levenstein} = 0$, and left results with $0 < d_{Levenstein} < 10$ for later manual revision. A file stores the outcome of the manual curation, which the parser consults when running into those cases, so that parsing and loading goes uninterrupted.

[1] Pathway Tools is the software used by SRI to create the different BioCyc databases, starting from genome sequences of the organism and comparing against the data in the multi organism MetaCyc database

[2] In the context of information retrieval and text mining, normalizing means to associate a noun or entity in the text to an identifier of a database or controlled vocabulary.

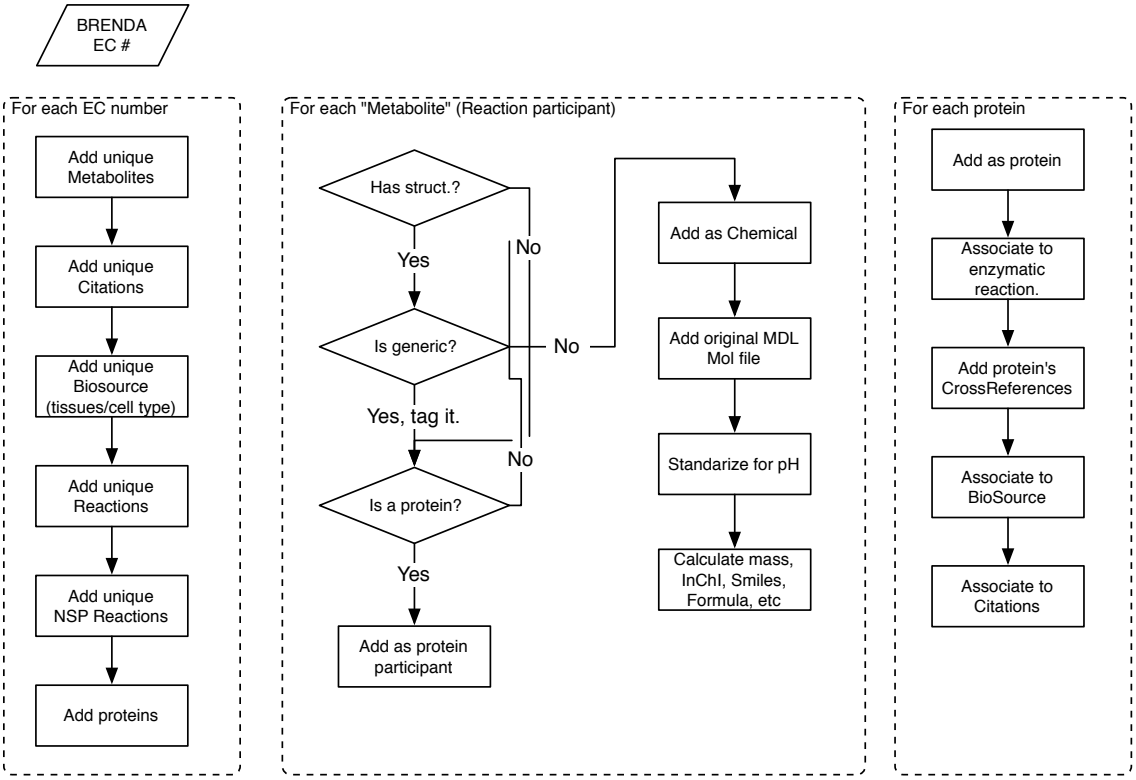



Figure 2.3: For loading BRENDA organism specific databases, the loader written executes the pipeline shown. The main object in BRENDA are enzymes, hence the pipeline starts by parsing the reaction associated to each enzyme (codified by the EC Number), as can be seen in the left block. The central block shows the detailed process for loading small molecules, obtained from the reactions. The rightmost block shows the process that the loader follows for the proteins. NSP Reactions stands for natural substrate product reactions, as opposed to artificial catalysis achieved using the enzyme with exogenous molecules.

For some cases, the process could not map the tissues and cell types present in the BRENDA data file to elements in BRENDA Tissue Ontology. The mapping process relied on the Ontology LookUp Service (OLS) [20], which allowed to easily switch the target ontology, enabling to fall into other ontologies dealing with tissues such as the Experimental Factor Ontology (EFO)[97] among others.

Chemical structures, as MDL MOL files or any other format, were not available directly for download from BRENDA, so a Web Bot was configured as described in [24] to retrieve MDL MOL files for all the structures in BRENDA. The loader stores chemical structures as obtained from BRENDA and standardized to pH 7.

#### 2.3.4.4 HMDB: Human Metabolome Database

There was no parser available for the HMDB, so I wrote a parser that uploads all the small molecules, tissues and cell types present in the HMDB metabo cards file, which is the main downloadable file from HMDB, and in the SDF chemical structure files.

HMDB links small molecules directly to tissues and cell types, in contrast to BRENDA, which links proteins to these containments. The loader produces links between the Chemical table and the BioSource table to account for this. As with BRENDA, tissues and cell types in HMDB are normalized through the BRENDA Tissue Ontology.

#### 2.3.4.5 General post-processing of data sets

Loaded data sets contained various errors like missing references or duplicated Chemical entries (specially KEGG), which I corrected through Java-based post processors. These post processors, or helpers as I sometimes refer to them, also handled the upload of structures (MDL MOL files) into the database, the derivation of the major microspecies at pH 7 and the computation of the following properties and identifiers:

**Major microspecies pH 7:** Using Chemaxon JChem, this module calculates the most abundant microspecies at pH 7 for each existing structure, storing

it in the ChemicalStructure table and linking it to the original chemical entity in the Chemical table.

**CrossReference normalizer:** This module normalizes cross references to external databases, so that the name of the external resource and the format of the accession or identifier is the same across different data sets. For this standardization, I used the MIRIAM catalogue of databases and resources, as well as synonyms added manually.

**Chemical cross references:** This module enriches chemical cross references to other databases through existing names, synonyms and existing cross references. The helper searches these against previously prepared Lucene indexes for ChEBI, PubChem Compounds (a sub set of it), and KEGG Compounds. When using names and synonyms, the tool only accepts the results if they are exact and unique, so if two different ChEBI IDs are retrieved after searching for a name, these identifiers are neglected. All used IDs are primary identifiers. The tool checks every new cross reference to have the same connectivity than the molecule to be annotated in the database, unless that the molecule has no structure. Generally speaking, around 5% of the assignments by name get rejected due to differences in connectivity.

**Redirection to primary identifiers:** For ChEBI and PubChem Compounds, there are many entries that are equivalent within the databases even though they have different ID numbers. A normal ID number comparison would yield these entries as different. To avoid this, the tool replaces all equivalent identifiers by the primary identifier in the case of ChEBI, or the parent CID in the case of PubChem Compounds.

**Molecular formula:** This module calculates the empirical formula using the CDK, after transforming implicit to explicit hydrogens, both for the original structure and its major micro species[1] at pH 7.

**Exact mass:** The module computes the mass of the most abundant isotopic

[1]The major micro specie at a particular pH is the predominant structure, in terms of protonation, at that pH

isomer through the CDK, after transforming implicit to explicit hydrogens, both for the original structure and its major micro species at pH 7.

**Natural mass:** The module computes the average mass, weighted by the abundances of the different isotopic forms, calculated with the CDK.

**InChI:** This module calculates Standard InChI string, Standard InChI key and AuxInfo using IUPAC InChI 1.03 through the CDK.

**SMILES:** The module computes the "simplified molecular input line entry specification" using the ChemAxon's JChem SMILES generator[1].

**Gibbs Energy:** The Standard Transformed Gibbs Energy of reaction is calculated for each reaction, through the calculation of Gibbs energies of formation based on group contribution (Section 4.1.4.1 on Chapter 4).

For all the operations that required the chemical structure of the molecules, the tool filters the structures to get rid of smaller disconnected components that can be sometimes found with the structures (like metal ions or salts components for instance). All SMILES, InChIs, masses and formulas hence reflect the largest connected component appearing on each molecule structure. The MDL MOL V2000 structures stored in the database contains all the data, the program applies the filtering only when executing calculations. This is specially important for the later steps of molecule comparison towards a unified data set.

### 2.3.5 Overview of loaded data sets after post processing

With the construction of a database of metabolomes being the main aim of this work, the most important objects of the loaded data sets should be the Chemical structures and their meta data. Figure 2.4 shows the amount of small molecules within each data set, where chemical entities are separated into the three mentioned categories: with structures, with generic structures and chemical entries with no structure. The number of different structures tends to be similar between data sources within a certain organism, with some exceptions. BioCyc

[1] Originally I used the CDK SMILES generator, however I found a high amount of incorrectly generated SMILES, when comparing the same molecules from different data sources.

database EcoCyc is probably the most intensively curated metabolism resource for *E. coli*, this generates the higher number of different small molecules with structure compared to BRENDA or KEGG. In the other cases, KEGG has a slightly higher amount of molecules, backing the general impression that it is the widest metabolism resource, although it achieves this by reactions and small molecules added through the KEGG orthology. The graph also shows the number of molecules for the HMDB database, our gold standard collection of small molecules for *H. sapiens*. The HMDB molecule count goes beyond the limit of the chart, as it has ~8,000 small molecules, which is a big difference compared to regular metabolism resources for *H. sapiens*, with normally no more than 1,500 small molecules. One of the technical challenges of this thesis is to be able to generate organism specific collections with sizes that lie in between the existing metabolism resources and metabolome databases like HMDB.

One would expect that two model mammal organisms such as *H. sapiens* and *M. musculus* would show similar number of chemical structures, the differences shown however are mostly due to the level of curation that the related databases receive.

Given the difficulty of comparing small molecules, it is useful to have as many cross references to other small molecules databases as possible. Figure 2.5 shows the number of cross references to KEGG, ChEBI and PubChem Compounds for all the chemical entities on each of the loaded data sets.

At this point we can also compare the mass distributions of small molecules for the different organisms and data sources. Figure 2.6 shows the exact mass distribution for each data set by organism. The exact mass is calculated using the most abundant isotope for each element in the molecule. As it would be expected, the aggregation of metabolites masses tend to be relatively similar from data set to data set within organism, but also, relatively similar from organism to organism. Only the mass distribution of HMDB strikes as different to the other *H. sapiens* databases. I inspected the small molecules at peaks from 700 to 900 Daltons and at 1,500 Daltons, where HMDB mostly differs, and found that it is due to the high number of Lipid species included in HMDB. Details on the numbers and

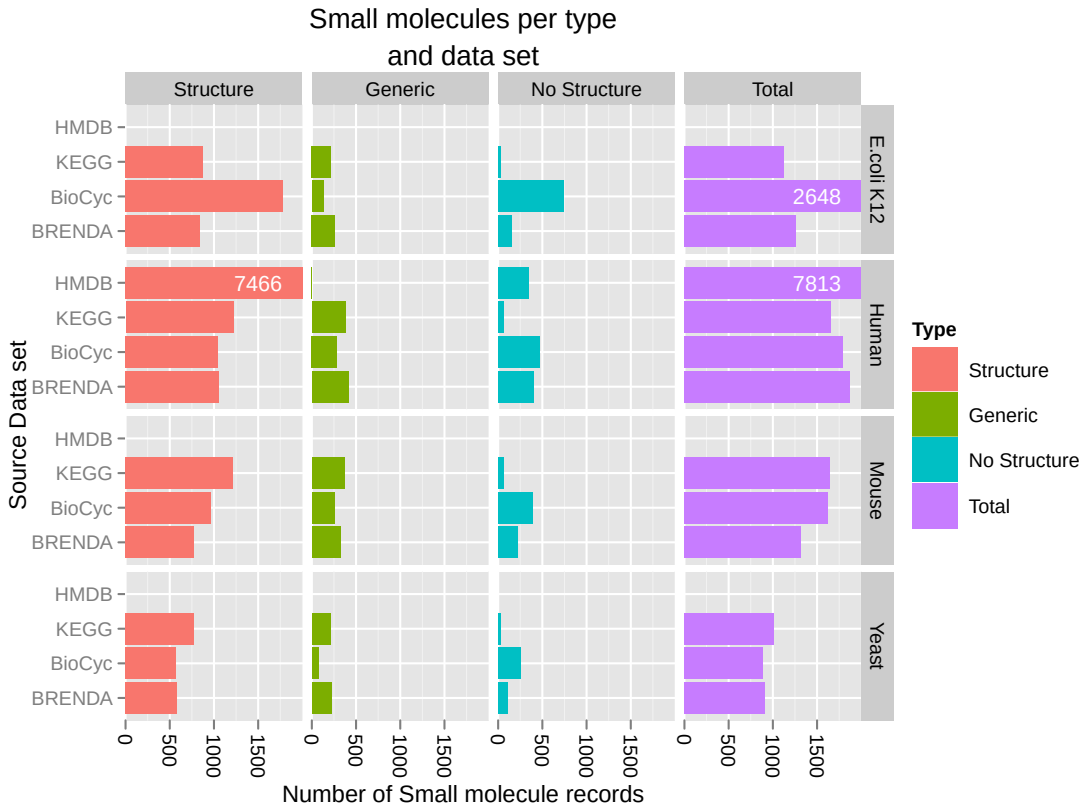



Figure 2.4: Summary of the number of molecules per data set loaded, dissected by organism. The bars for complete structure and generic structures represent counts of different structures (different InChI and different SMILES respectively), where the bars for no structure are just different entries in the database.

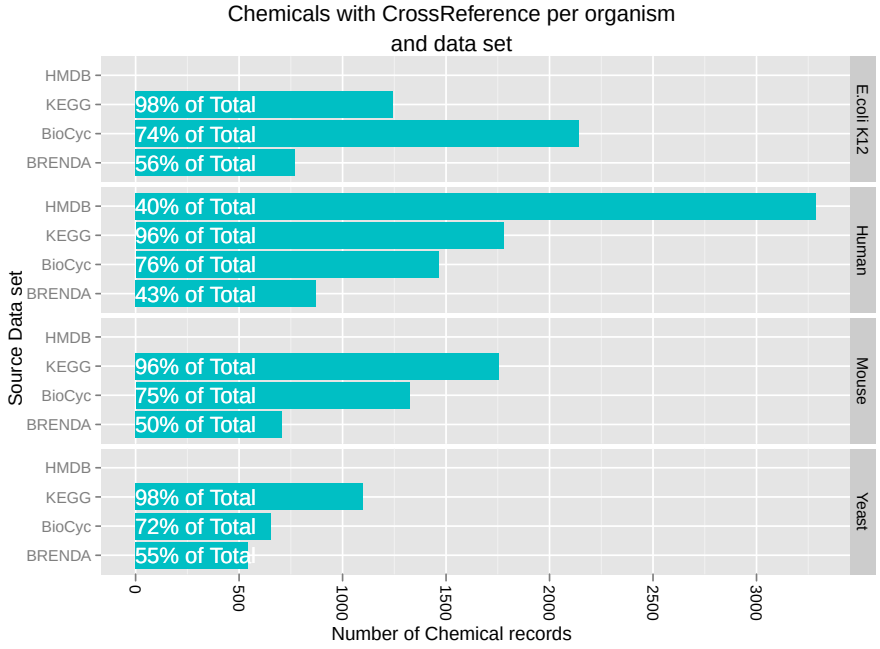



Figure 2.5: Summary of the number of molecules with cross references per data set loaded, dissected by organism. This only considers cross references to ChEBI, PubChem Compounds or KEGG. The counts shown are after applying the modules for adding cross references and normalizing them in format and identifier (primary/secondary). The text within the bars indicates what % of the total number of chemical entries for the organism/data source combination has a cross-reference.

types of Lipids are given in Figure 2.28.

Figure 2.6 also shows – for the four organisms – that the masses of molecules distribute in a common first peak right below the 200 Daltons and then in a second, much smaller peak, close to 800 Daltons. This behaviour seems very interesting, so I analyze these peaks compositions per organism later on, once organisms are unified.

Another important aspect of metabolism resources is the biological context they provide, mostly through reactions and enzymes. Figure 2.7 shows the number of reactions for each organism and data source. I trust reactions to be different within each data source, so each entry is counted. The only requirement is that the reaction involves at least one chemical entry (regardless of whether it has a structure, a partial structure or no structure at all). I separate reactions in two groups, the first one with a direct assignment of a protein to catalyze the reaction in the database, or "catalyzed", and a second one with reactions with no protein assignment to catalyze it, or "not catalyzed". The fact that a reaction appears in a database without assigned enzyme does not mean necessarily the reaction is spontaneous, but most of the time the resource does not have evidence to support the assignment of an enzyme to that reaction in that organism. It is true that spontaneous reactions fall in this category, but they represent only a minor fraction (overall spontaneous reaction counts can be seen in page 18).

As with the count of chemical entities, the number of reactions is similar across different databases for a given organism. KEGG tends to have fewer "catalyzed" reactions but normally gets to the same numbers as BioCyc or BRENDA by additional "not catalyzed" reactions. In the case of KEGG these are derived through the orthology families as explained in section 2.3.4.1. In the case of BioCyc databases, this happens mostly due to the assignment of pathways of the adequate taxonomic range when most, but not all, the enzymes are annotated in the organism.

As we mentioned before, enzymatic reactions are normally classified through the Enzyme Comission number, or EC number. This classification is widely used as well for assigning reactions to the enzymes of newly sequenced genomes, once

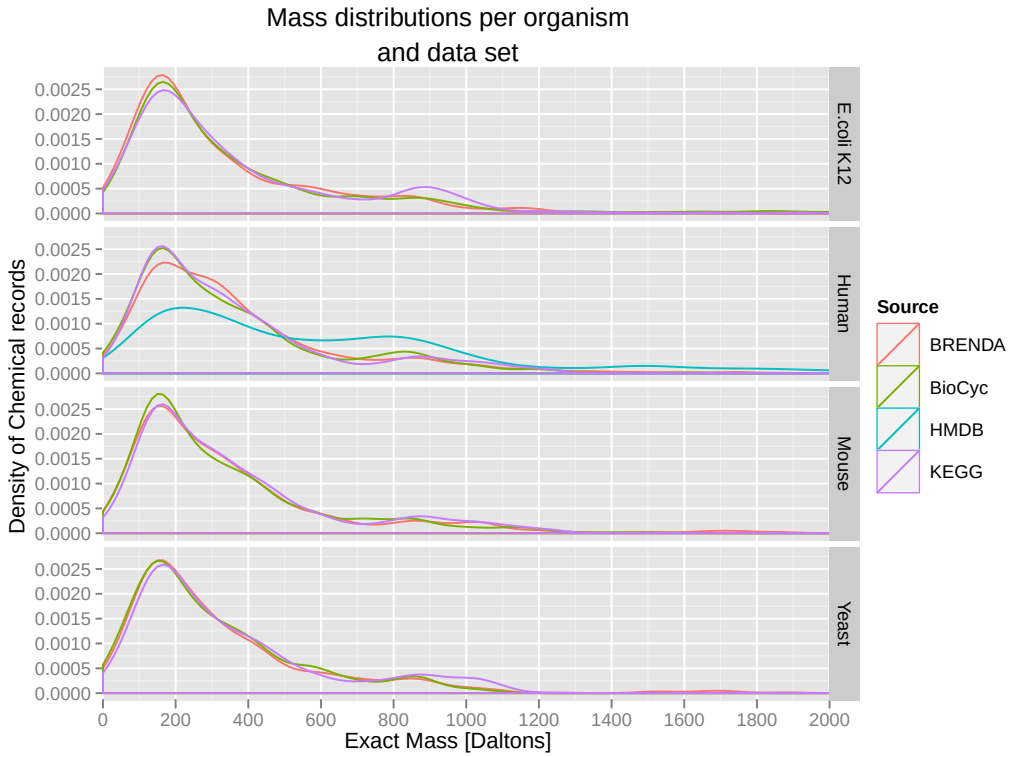



Figure 2.6: Distribution of chemical entities mass per data set loaded, dissected by organism. Density curves always have equal areas, so overall size of each data set is not reflected, the aim of the plot is to compare the positions of the peaks and the distributions of the data set, normalized by size. The mass corresponds to the exact mass, calculated from the most abundant isotopes. Metabolites in all the organisms align reasonably well in terms of mass distribution. The distribution for HMDB is remarkably different to the other databases for *H. sapiens* metabolism. The lower peak between 100 and 200 Daltons for HMDB and the higher peak at 700 to 900 and at 1,500 Daltons is explained by the inclusion of thousands of Lipids in HMDB that are not present in the other databases.

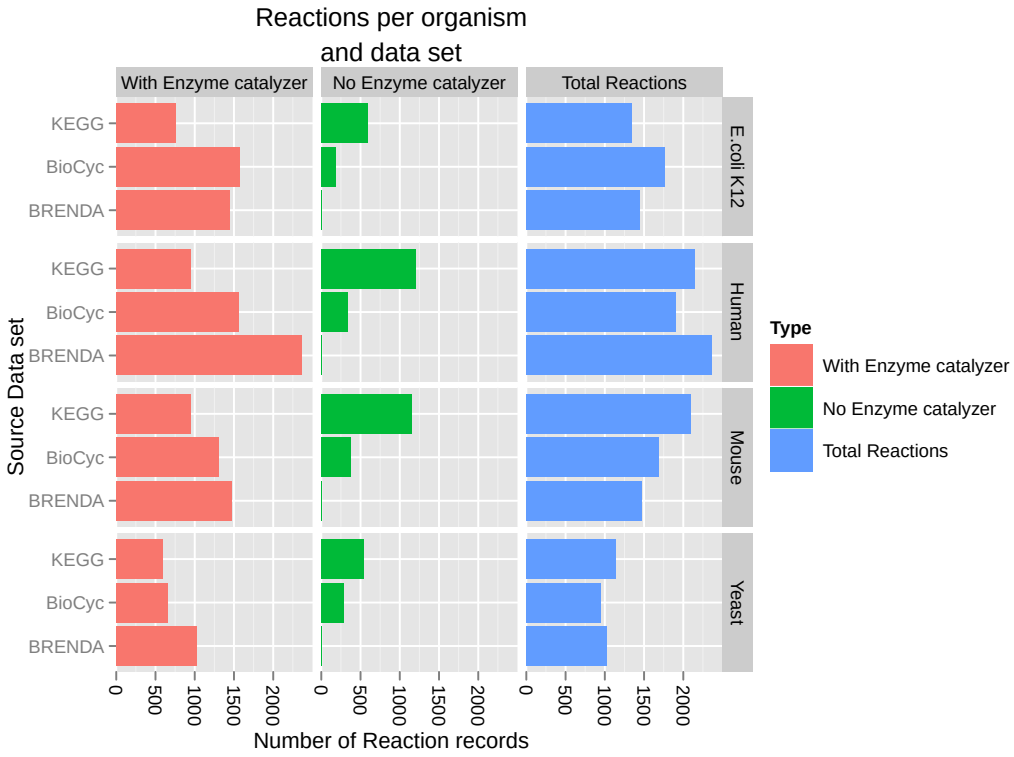



Figure 2.7: Summary of the number of reactions per data set loaded, dissected by organism. The total number of reactions is the sum of reactions with an assigned enzyme catalyst and reactions without an assigned enzyme catalyst. The number of reactions without an enzyme catalyst assigned does not mean that the organism has that many spontaneous reactions, but that the data set lacks evidence to make some of the reaction-enzyme assignments. In the case of KEGG, many catalyst-to-reaction assignments are done indirectly through the orthology groups, which explains why this database has so many more reactions without an assigned enzyme. In the case of BRENDA, as the resource focuses on enzymes directly, it does not have reactions without a catalyst assigned.

these enzymes have been annotated with EC numbers. The number of different EC numbers annotated in a genome will normally tell us something about the chemical diversity of the metabolism of that organism, and hence we can use it as a non-quantitative predictor of the organisms metabolome. Figure 2.8 shows the number of different EC numbers according to organism and source data set. Only looking at the amount of different EC numbers per organism given by the different databases, it might be tempting to conclude that the databases have essentially the same reaction content. However, upon closer inspection of the data, some differences show up that the following sections reveal.

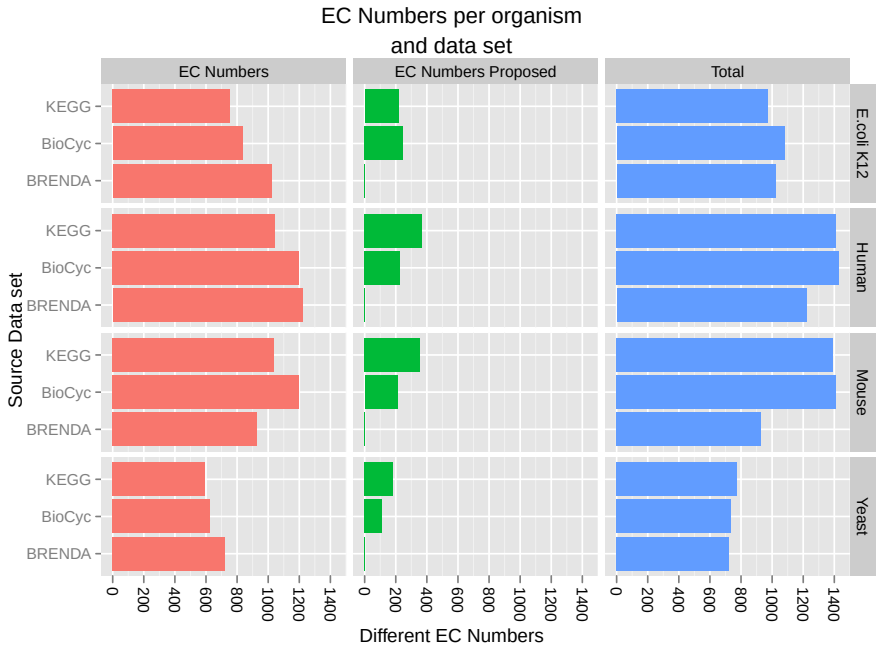



Figure 2.8: Summary of the number of different EC numbers per data set loaded, dissected by organism. Total number of EC numbers is the sum of officially assigned EC numbers and proposed EC numbers.

The number of proteins that each data set is linking against small molecules is also relevant, as proteins are a gateway for additional contextual knowledge. Figure 2.9 shows the count of proteins available per data set, both as a count of entries and as a count of unique UniProt identifiers, which is the most useful

representation of proteins to retrieve localization information from many other resources. The different count of proteins for each organism and data source is produced by the fact the some resources, like BioCyc, index the full proteome of the organism, where as others like BRENDA only include enzymes. Also some issues of data completeness play against some databases, specially KEGG. In the case of *E. coli* an inconsistency of data in KEGG release v57 makes the automatic mapping of proteins impossible. KEGG release v57 for *E. coli* points to the taxonomy identifier of the *E. coli* K-12 substrain MG1655, NCBI Taxonomy TaxID 511145, however previously it used to point towards *E. coli* K-12, NCBI Taxonomy TaxID 83333. UniProt indexes this bacterium proteins under NCBI Taxonomy TaxID 83333, and hence probably in the automatic generation of release files at KEGG all UniProt identifiers linked were lost. For generating the enzymatic reactions, reactions and small molecules for *E. coli* in KEGG the PerSpecies helper – explained in section 2.3.4.5 – uses the KEGG orthology information deposited in UniProt to map up to 2,589 *E. coli* proteins to known KEGG reactions. Similar thing happens with *S. cerevisiae* in KEGG, for which approximately 1,333 proteins are potentially mapped to KEGG reactions. This illustrates one of the many difficulties that integrating multiple databases for multiple organisms can have, and that many times can go unnoticed if they only happen in a few cases.

Through proteins, we can relate small molecules to different biological containers: tissues, cell types and even cellular organelles. BRENDA is the only metabolic database that provides this kind of data out of the three major resources. BRENDA provides this information for proteins, not for small molecules directly. The HMDB has assignments of small molecules to some biologcal containers as well. These data are only useful if they are indexed by a controlled vocabulary or even better, a proper ontology. As I mentioned before, during the loading of both BRENDA and HMDB, I normalized the biological containers present through the BRENDA Tissue Ontology and the Experimental Factor Ontology ontologies among others. Figure 2.10 shows an overview of the amount of different tissues and cell types linked by BRENDA and HMDB, and the number

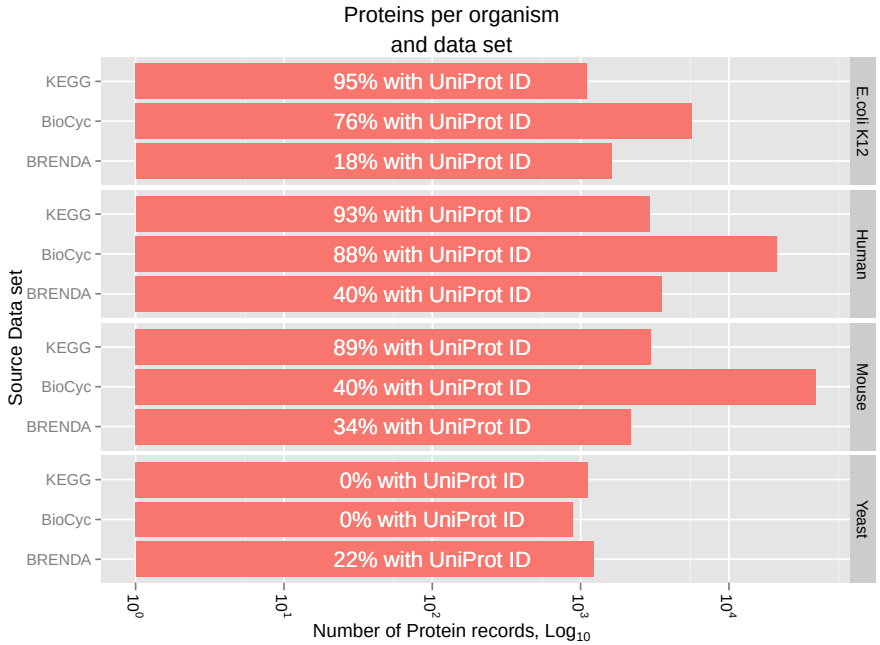



Figure 2.9: Summary of number of proteins per data set loaded, dissected by organism. Number of proteins linked vary between data sets within an organism due to the focus of the data source: BioCyc databases tend to include most of the proteome of the organism whereas BRENDA includes only those that are enzymes. An adequate cross reference to UniProt is very important, as this allows to obtain localization data through proteins.

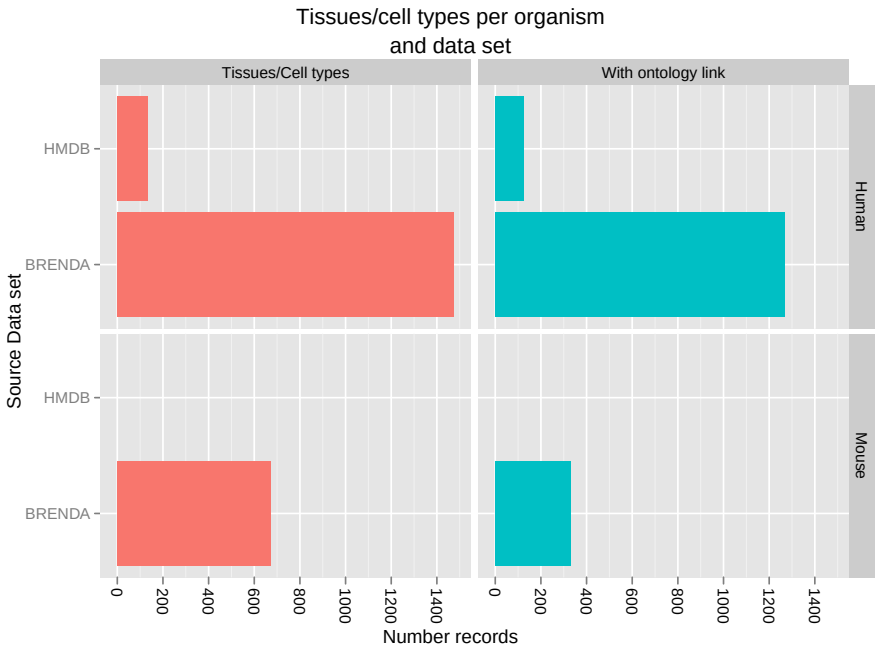



Figure 2.10: Summary of number of tissues/cell types and their mappings to ontologies per data set loaded, dissected by organism. Data of tissue localization is of interest for multi cellular organisms mainly, so we leave out *E. coli* and *S. cerevisiae* (although for unicellular organisms tissues and cell types can be analogous to media types or host cells). The HMDB tends to make use of a much smaller and general set of tissues and cell types, but makes direct assignments of small molecules to them. The data from BRENDA is much richer given the number of individual protein studies and high-throughput expression studies.

of links to ontology terms.

In general both the numbers of reactions and chemical entries suggest that the three main metabolism resources should have equivalent collections of small molecules for each organism. However, to really assess this, a proper molecule comparison is required. In general, comparisons existing in the literature tend to show very low overlaps between the small molecules in one resource and the other for the same organism, but much of this is probably due to the technical challenge of the comparison rather that the collections being so different.

### 2.3.6 Comparing small molecules from one database to another

Among different biochemical entities (genes, proteins, etc.), it has been recognized that the comparison of small molecules between different data sources is particularly difficult [92] due to stereochemistry, tautomerism, different protonation, and errors present in some structures. In essence, small molecule comparison is a graph isomorphism problem, which is normally very expensive to solve and scales non-polynomially[1]. Through isomorphism, two molecules are said to be the same if there is a complete bijection between all their bonds. A bijection between any pair of bonds – one bond belonging to each molecule compared – is only possible if they both connect the same atoms in their respective molecules. Because of the high computational cost of this problem, a number of alternatives have been generated during the past decades to compare chemical molecule's representations in the computer.

A common approach is to calculate fingerprints for a set of molecules and then compare the molecules through their fingerprints. Fingerprints attempt to capture a defined finite set of structural/physicochemical properties or characteristics of molecules, but they are not able to capture all the structural information. However, for two molecules to be the same, necessarily, their fingerprints (that are a simplified representation of them) need to be the same. Hence, when we want to find the unique subset of a set of molecules, we do not need to compute

[1]That a problem scales non-polynomially means that the execution time of known algorithms to solve it increase very quickly with the size of the input.

the graph isomorphism between every pair of molecules, but only for those pairs that have equal fingerprints.

A different approach is the calculation of line notation representations of the chemical graph, converting the chemical graph into a variable length ASCII string. The idea in this case is to have an algorithm that is able to traverse molecule graphs in a canonical way (within that algorithm at least) and then generate an ASCII string representation of that traversal, considering of course atoms symbols, bond types, charges, etc. Examples of this are SMARTS, SMILES and InChI among others. Different to the fixed size fingerprints, this type of representation aims to capture all the structural information of a chemical graph to a certain level of detail, and some of the quoted examples are often used for direct comparison of small molecules.

SMILES, which stands for "simplified molecular input line entry specification", was developed originally in the late 80's by Arthur and David Weininger [164]. During the past decades it has been extended and implemented in a number of cheminformatics packages, being widely used by the research and commercial communities. Probably due to historical reasons, maybe because the open source movement was not so strong by the early 90's, different closed source implementations of SMILES proliferated, giving rise to its main problem today: SMILES is highly implementation dependent, equal molecules can have different SMILES representation if they are computed by different software packages. However, if one can make sure that SMILES are always calculated with the same implementation, then it should be safe to compare molecules. Again the problem of multiple implementations means as well that any bugs found in one of them is probably still present in the others, which also makes SMILES slightly less reliable than other options. On the other hand, one of the nice features about SMILES is that it supports generic structures.

The InChI identifier [140], which stands for "International Chemical Identificator", is a more modern line notation representation developed by IUPAC and NIST during 2000-2005, after that mainly by IUPAC, and from 2010 also supported by the InChI Trust. Differently to SMILES, InChI has always been open source, which has allowed the entire community to share a single implementation. This makes InChI strings a much more stable way of comparing small

molecules across different databases. Initial versions of InChI provided a number of options regarding how to interpret various aspects of a molecule's representation (whether to consider or neglect stereo chemistry, mobile hydrogens, to add protons according to present valencies, etc.). By request of the community, later versions of InChI (starting from 1.02) introduced the Standard InChI, which is an InChI string generated with certain default options that cannot be modified, making easier the comparison of InChI strings across databases.

When comparing small molecules, there are an additional number of issues that arise from the chemistry being represented. A first problem is generated by protons: different databases have sometimes their molecules at different protonation states. Many times as well, they represent the molecule by a different tautomeric variant, and often stereo chemistry is far from being reliable or comparable [123]. These issues affect the generated SMILES and InChIs depending on the way they are generated, specially different protonation states and different tautomer forms.

#### 2.3.6.1 Method for comparing small molecules from different databases

Molecules contained in metabolism databases like BRENDA, KEGG and BioCyc, among others, can be operationally separated in four types according to the level of data they contain:

**Complete Structure** These are entries in the database, that represent small molecules, for which the complete chemical structure is available (normally as an MDL MOL V2000 file). This excludes entries that have molecules with undefined generic groups (-R groups).

**Generic Structure** These entries, that represent a small molecule, have a chemical structure available (normally in the form of a MDL MOL V2000 file) that is only partially specified, as they contain a variable regions. This kind of structures are normally known as generic structures. These generic structures represent "classes" of molecules like "alcohols" or "aldehydes". Generic structures sometimes fail to describe the richness of the chemical entity portrayed due to the simplicity of the representation (for instance,

restrictions on the type of structures that could replace the variable part have no way of being represented).

**Empty Structures representing small molecules** Although most of the entries present in these databases should fall in either of the first two types described previously, databases contain some chemical entities with no structure associated. This can be either product of an error, lack of completeness or even the explicit decision to avoid using generic structures to represent certain classes (because of the poor representation they might provide).

**Empty Structures representing non-small molecules** In many biochemical reactions there are proteins and other biomolecules that can participate and are as such included in the databases. Many times the transference of a database from format to format might loose semantic indications that those elements are not small molecules.

Given the different richness of these types of entries, they are initially compared separately. In the case of entities with structure, the comparison is done through Standard InChI strings. In a first approach, I used the JChem $pK_a$ plugin [22] to standardize all molecules across databases to pH 7, to solve differences in protonation states, and then compare the molecules across the databases through Standard InChI key. As mentioned before, the database schema of BioWarehouse was modified so that it could hold multiple chemical structures (at different pHs) for a defined chemical entry.

The results of this comparison showed that the protonation standardization still left many molecules with the same scaffold and cross references in different protonation states. Molecules from HMDB had particularly many differences in the hydrogen layer when compared to the other databases, for molecules that a human curator would consider the same. Overall, I identified nearly 80 cases like these in the *H. sapiens* metabolism database comparison. Figure 2.11 shows one example for reduced riboflavin from KEGG and BioCyc; Figure 2.12 shows the same for hypotaurine, where actually the molecule coming from BioCyc has also an error in the sulfur configuration.

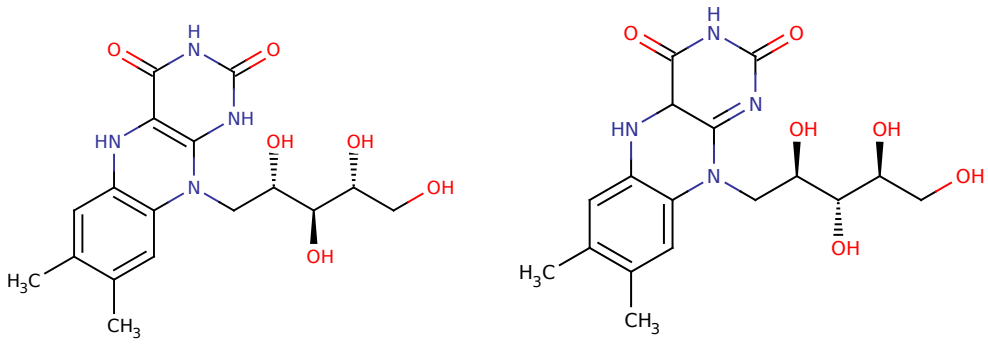



Figure 2.11: Molecules from KEGG and BioCyc for reduced riboflavin, where the BioCyc version has a nitrogen less protonated, and a corresponding double bond in the upper ring. Both molecules remained unchanged when protonated to pH 7 using Chemaxon JChem, so original versions are shown only. The reduced riboflavin from BRENDA has the same structure as the BioCyc version, hence is not shown here. Below the names, the Standard InChI hydrogen layers – which are different – can be seen for both molecules.

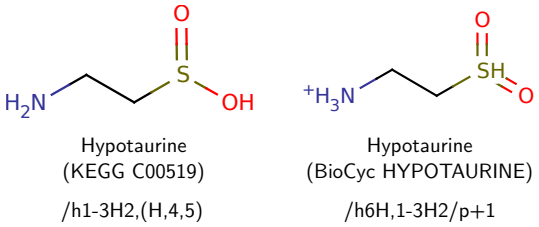



Figure 2.12: Molecules from KEGG and BioCyc for hypotaurine, with different protonation for the nitrogen atom, and with an error in the sulfur configuration in the case of BioCyc. By the time this work was submitted, hypotaurine was fixed in BioCyc, but this is the state of the molecules when it was downloaded. Both molecules remained unchanged when protonated to pH 7 using Chemaxon JChem, so original versions are shown only. Hypotaurine from BRENDA has the same structure as the KEGG version, hence is not shown here. Below the names, the Standard InChI hydrogen layers – which are different – can be seen for both molecules.

I re-tried the same comparison avoiding the last character of the Standard InChI key, which is reserved for the protonation state, however many incorrect mismatches still appeared due to the differences in hydrogen distributions, even after protonation standardization. Even though the last character in the Standard InChI key is reserved for the protonation state, the attached protons still influence the computations of other parts of the key (as they modify saturations and other features for instance).

To overcome the protonation problems, I turned to the complete Standard InChI string in a second approach. I transformed the Standard InChI taking advantage of its formation by layers, to remove the effect of protonation. The transformation implemented replaces the Hydrogen in the empirical formula layer (second layer of an Standard InChI) with a wildcard recognized by the database search engine, and deletes all layers appearing after the main connectivity layer (which only considers "heavy" atoms, that is all the non hydrogens). We call this an InChI Connectivity, as it represents essentially the connectivity of the molecule. For instance, for the Standard InChI of Serine:

```
InChI=1S/C3H7NO3/c4-2(1-5)3(6)7/h2,5H,1,4H2,(H,6,7)
```

the corresponding InChI Connectivity would be:

```
InChI=1S/C3%NO3/c4-2(1-5)3(6)7/
```

Comparisons through the InChI Connectivity tend to be very generous, sometimes grouping stereo isomers or species with different saturation levels, which one would want separated in a database. This demanded the implementation of subsequent separation steps, based on meta data (chemical names and cross references to other databases mostly) and structure, that could process a group formed by chemicals from different sources but with the same InChI Connectivity, and separate them in adequate sub groups of small molecules that are equivalent. So first the method relies on connectivity, and then within each connectivity class, uses human annotation of the molecules to discriminate which of those chemical entries should be considered the same.

For chemical entities with generic structures, the InChI string cannot be calculated. Because of this, it is necessary to rely on a second option, compare

small molecules by SMILES. Identification of small molecules by SMILES is of lesser quality than by InChI. Even after protonation correction, out of a few thousands equivalence classes of small molecules generated through InChIs, one finds a few hundred of these classes where the molecules within (which have the same InChI) have different SMILES (and hence, if classified by SMILES in the first place, they would have been deemed different when they are not). I computed SMILES for each molecule as Section 2.3.4.5 explains (protonation to pH 7, detecting aromaticity, removing hydrogens and only keeping the largest connective component).

I implemented a method to unify different sets of small molecules which makes use of structure and meta data. The method is illustrated in Figures 2.13, 2.14, 2.16 and 2.17. It is composed of three main steps. For defined structures, the first step is a very generous way of comparison that builds groups of molecules with equal connectivities, but that might include molecules that are not exactly equivalent. In a second step the method dissects these groups, using mostly meta data (cross references, names and synonyms mainly) and occasionally stereo chemistry, into sub groups where molecules should be equivalent. The same applies for generic structures. For molecules with no structure associated, the initial comparison relies on an index built from the name, which normalizes the words avoiding trivial differences, as explained in Appendix A. Finally, the method compares the unified groups formed in the structure and no structure parts using cross references and normalized name matches, to try to complete groups in the structure (and generic structure) part with elements that do not have a structure due to database completeness errors.

Figure 2.13 shows the initial step of group formation, that starts by generating keys that allow to join the data sets. Completely defined chemical structures use InChI Connectivity as keys. Generic structures use SMILES as keys. The pipeline retrieves the unique InChI Connectivity and SMILES from the data sets and puts them together, generating a set of unique InChI Connectivity and SMILES. For molecules with no structure, the program generates an index based on our Chemical name fingerprinter. For *H. sapiens*, considering KEGG, BioCyc and BRENDA, there are 1,543 unique InChI Connectivity, 914 unique SMILES (for generic molecules) and 885 different normalized names for the remaining chemical

entries that do not have a structure.

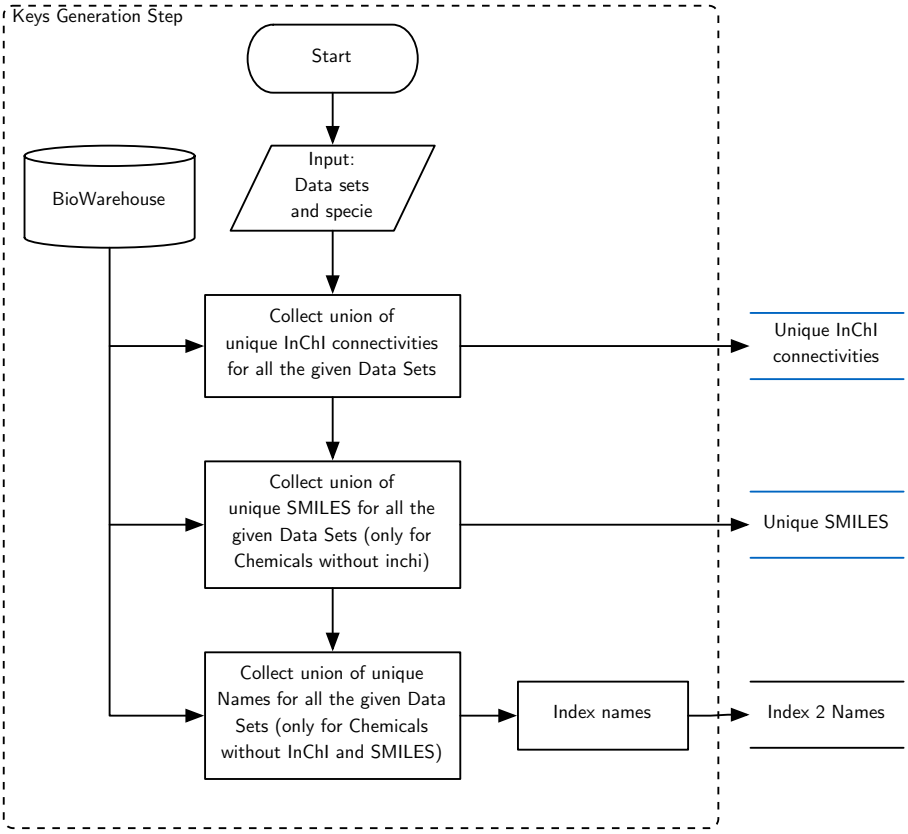



Figure 2.13: Initial part of the consolidation algorithm, where keys are generated to join the chemical entries in the different data sets. The method obtains InChI connectivities for all the complete molecules in the given data sets and generates a unique set of InChI connectivities. The same is done with generic molecules, but the method uses SMILES instead of InChI connectivities. Finally, it retrieves a list of unique names from the chemical entries in the data sets that do not have a structure.

Figure 2.14 illustrates the second phase of the consolidation. Using the compiled set of keys, the method queries each data set, to obtain the representative chemical entities in each data set that match each key. All the elements from the different data sets that match a defined key are left together in one group. For example, the InChI Connectivity:

InChI=1S/C11%N2O4/c12-8(11(16)17)5-10(15)7-3-1-2-4-9(7)13-6-14/

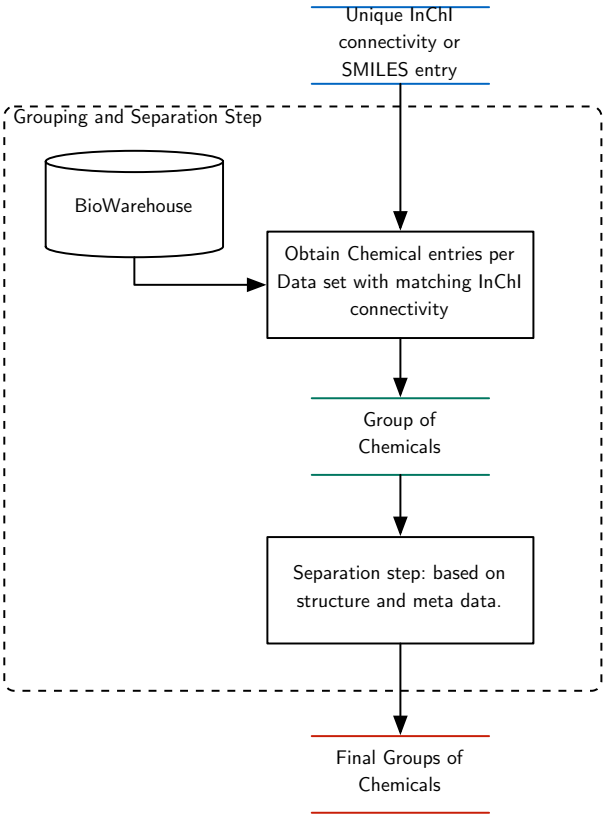



Figure 2.14: Second part of the consolidation algorithm, the method queries each key generated to the different data sets, to find representatives of that key in each data set. All the results of a single key accross data sets are considered a Group of Chemicals (in green). The method then attempts to break up the group into sub groups if it is appropiate. This separation step relies on structure and meta data, and can be seen in more detail in Figure 2.16.

used as a key retrieves the results shown in Table 2.2, for the formyl kynurenine related compounds. In that case, we can see that the group should be latter split in two groups, the N-Formyl-D and the N-Formyl groups. Even though N-Formylkynurenine and L-Formylkynurenine, in BioCyc and KEGG respectively, mean the same molecule, their InChIs have different hydrogen layers (this was not corrected either by protonation adjustment). Because of this, through an

ordinary InChI consolidation, they would end as different molecules. Figure 2.15 shows that these molecules have different structures in KEGG and BioCyc both before and after protonating to pH 7.

| DataSet | Name | InChI (hydrogen layer) |
|---|---|---|
| KEGG | N-Formyl-D-kynurenine | /h1-4,6,8H,5,12H2,(H,13,14)(H,16,17) |
| BioCyc | N-formyl-D-kynurenine | /h1-4,6,8H,5,12H2,(H,13,14)(H,16,17) |
| KEGG | L-Formylkynurenine | /h1-4,6,8H,5,12H2,(H,13,14)(H,16,17) |
| BioCyc | N-formylkynurenine | /h1-4,8,13-14H,5-6,12H2,(H,16,17) |

Table 2.2: Initial group formed for InChI=1S/C11%N2O4/c12-8(11(16)17)5-10(15)7-3-1-2-4-9(7)13-6-14/ InChI Connectivity, corresponding to Formylkynurenine. L-Formylkynurenine and N-formylkynurenine refer to the same molecule, however the hydrogen layer differs (initial layers are the same), leaving them as different in a plain InChI search. Unfortunately in this case, the representative from BRENDA has different connectivity, rendering them directly different.

If the method queries for InChI Connectivity

```
InChI=1S/C3%NO3/c4-2(1-5)3(6)7/
```

the results are the two different stereo isomers of Serine and an undefined stereochemistry Serine. This is another example where the method generates groups that include different stereo isomers, which are later separated. The "(L-or-D) Serine" result shows the complication generated by unexpected constructs, which will probably be left on a group of its own, increasing the number of molecules artificially.

Table 2.3 shows a more complicated case in which more molecules are obtained for the InChI Connectivity

```
InChI=1S/C3\%O2/c1-3(5)2-4/
```

In this case there are several sub groups that the separation step needs to isolate. This is normally as complicated, in terms of separation, as it can get. This also shows a number of molecules for which BRENDA either does not have

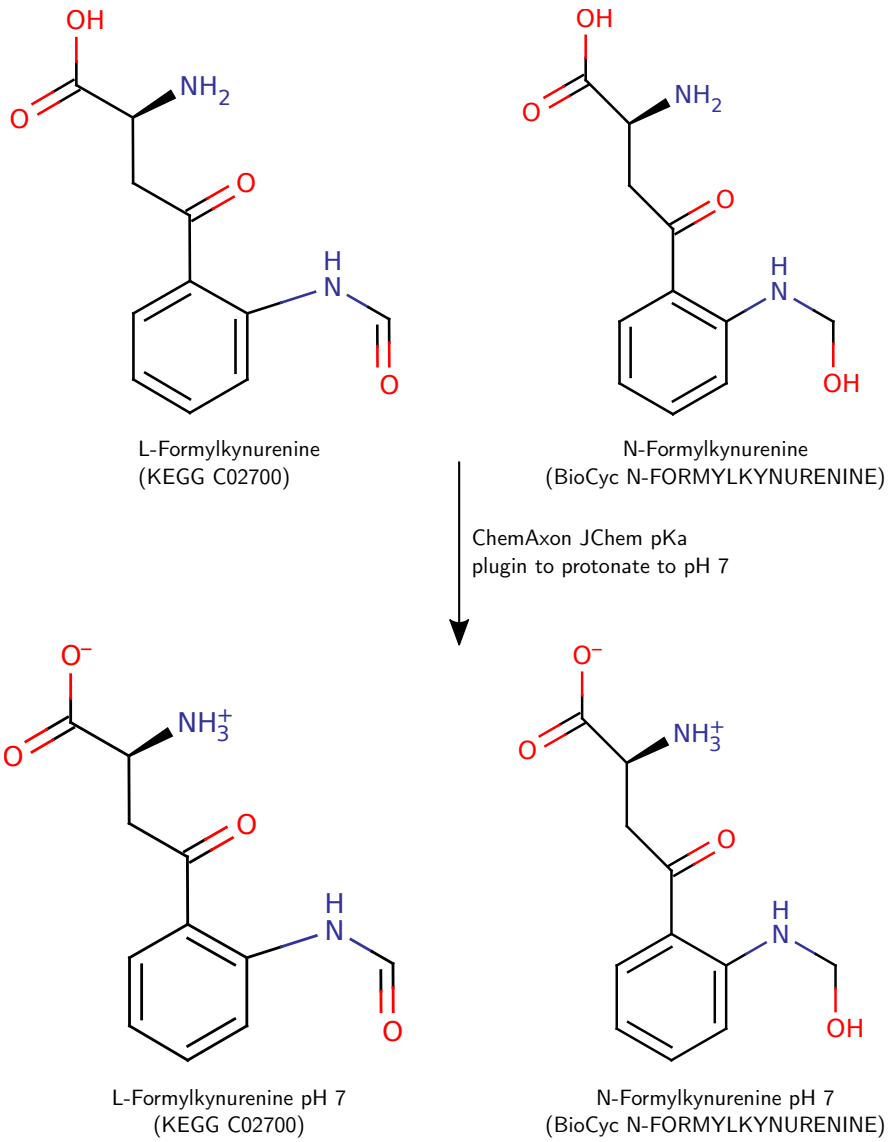



Figure 2.15: N-Formylkynurenin (or L-Formylkynurenine) from KEGG and BioCyc, before and after being protonated. Molecules, which are supposed to be the same – since the BioCyc entry links to the KEGG entry and they have the same synonyms – are different both before and after protonating to pH 7 with Chemaxon JChem.

| DataSet | Name | InChI |
|---|---|---|
| KEGG | Methylglyoxal | InChI=1S/C3H4O2/c1-3(5)2-4/h2H,1H3 |
| BioCyc | methylglyoxal | InChI=1S/C3H4O2/c1-3(5)2-4/h2H,1H3 |
| BRENDA | methylglyoxal | InChI=1S/C3H4O2/c1-3(5)2-4/h2H,1H3 |
| KEGG | Lactaldehyde | InChI=1S/C3H6O2/c1-3(5)2-4/h2-3,5H,1H3 |
| BioCyc | D-lactaldehyde | InChI=1S/C3H6O2/c1-3(5)2-4/h2-3,5H,1H3 |
| KEGG | (R)-Lactaldehyde | InChI=1S/C3H6O2/c1-3(5)2-4/h2-3,5H,1H3 |
| KEGG | (S)-Lactaldehyde | InChI=1S/C3H6O2/c1-3(5)2-4/h2-3,5H,1H3 |
| KEGG | Hydroxyacetone | InChI=1S/C3H6O2/c1-3(5)2-4/h4H,2H2,1H3 |
| BioCyc | acetol | InChI=1S/C3H6O2/c1-3(5)2-4/h4H,2H2,1H3 |
| KEGG | (S)-Propane-1,2-diol | InChI=1S/C3H8O2/c1-3(5)2-4/h3-5H,2H2,1H3 |
| BioCyc | L-1,2-propanediol | InChI=1S/C3H8O2/c1-3(5)2-4/h3-5H,2H2,1H3 |
| KEGG | Propane-1,2-diol | InChI=1S/C3H8O2/c1-3(5)2-4/h3-5H,2H2,1H3 |
| KEGG | (R)-Propane-1,2-diol | InChI=1S/C3H8O2/c1-3(5)2-4/h3-5H,2H2,1H3 |
| BioCyc | D-1,2-propanediol | InChI=1S/C3H8O2/c1-3(5)2-4/h3-5H,2H2,1H3 |

Table 2.3: Initial group formed after querying for InChI=1S/C3%O2/c1-3(5)2-4/ InChI Connectivity, corresponding to a number of different compounds in terms of protonation. In this case, stereo isomers and different protonation states are confounded in the same chemical group.

a molecule, or it has with a different connectivity which excludes it from this result.

Table 2.4 shows the results of searching for the connectivity of pyruvate. Here we see that all three databases show molecules for the different compounds that have that same connectivity. In the same table we find examples of differences in the charge layer which would render a normal InChI consolidation unusable without handling protonation/charges.

Finally, Table 2.5 shows another example of the same molecule in two different databases with different InChI hydrogen layers. This is an example where even adjusting the pH does not solve the differences, separating the molecules in a consolidation by mere InChI. Queuine is also a case where the databases even show different connectivities, as in BRENDA Queuine is a 7 carbon molecule with a different structure. This has been fixed in the current online version of BRENDA. The InChI shown for BRENDA actually corresponds to 7-aminomethyl-7-deazaguanine or Pre$Q_1$, which is related to Queuine. This confusion apparently has been dragged to other databases, as Pre$Q_1$ in BioCyc for *B. subtilis* links to Queuine in

| DataSet | Name | InChI |
|---|---|---|
| BRENDA | pyruvate | InChI=1S/C3H4O3/c1-2(4)3(5)6/h1H3,(H,5,6) |
| KEGG | pyruvate | InChI=1S/C3H4O3/c1-2(4)3(5)6/h1H3,(H,5,6) |
| BioCyc | pyruvate | InChI=1S/C3H4O3/c1-2(4)3(5)6/h1H3,(H,5,6)/p-1 |
| BRENDA | (R)-lactate | InChI=1S/C3H6O3/c1-2(4)3(5)6/h2,4H,1H3,(H,5,6) |
| KEGG | (R)-Lactate | InChI=1S/C3H6O3/c1-2(4)3(5)6/h2,4H,1H3,(H,5,6) |
| BioCyc | (R)-lactate | InChI=1S/C3H6O3/c1-2(4)3(5)6/h2,4H,1H3,(H,5,6)/p-1 |
| KEGG | Lactate | InChI=1S/C3H6O3/c1-2(4)3(5)6/h2,4H,1H3,(H,5,6) |
| BRENDA | (S)-lactate | InChI=1S/C3H6O3/c1-2(4)3(5)6/h2,4H,1H3,(H,5,6) |
| KEGG | (S)-Lactate | InChI=1S/C3H6O3/c1-2(4)3(5)6/h2,4H,1H3,(H,5,6) |
| BioCyc | (S)-lactate | InChI=1S/C3H6O3/c1-2(4)3(5)6/h2,4H,1H3,(H,5,6)/p-1 |

Table 2.4: Initial group formed after querying forInChI=1S/C3%O3/c1-2(4)3(5)6/ InChI Connectivity, corresponding to pyruvate and lactate. In this case, stereo isomers and different protonation states are confounded in the same chemical group. A plain InChI search aiming for pyruvate would have missed BioCyc's pyruvate, which has a different InChI (charge in this case).

ChEBI and KEGG. Recovering from these differences goes beyond the capacity of our pipeline, and is a quality issue of the resources in which the method bases the consolidation.

The consolidation process proceeds then with each one of the groups obtained by querying the connectivity keys and the name indexes. For the structure-based groups, the method attempts to break them into sub groups if appropiate, as Figure 2.16 shows in more detail. Groups in Tables 2.2 and 2.3 would be splitted in that step.

Figure 2.17 shows the third and last part of the unification process. The method compares the groups formed from chemical entries with no structure with groups of chemicals with structure through names and cross references, when exact matches are found and the merge increases the database coverage, the chemical entries with no structure are added to the groups with chemical structures.

The exposed way of merging small molecules was the most generous way that I could implement (generous in the sense of providing the higher overlap between

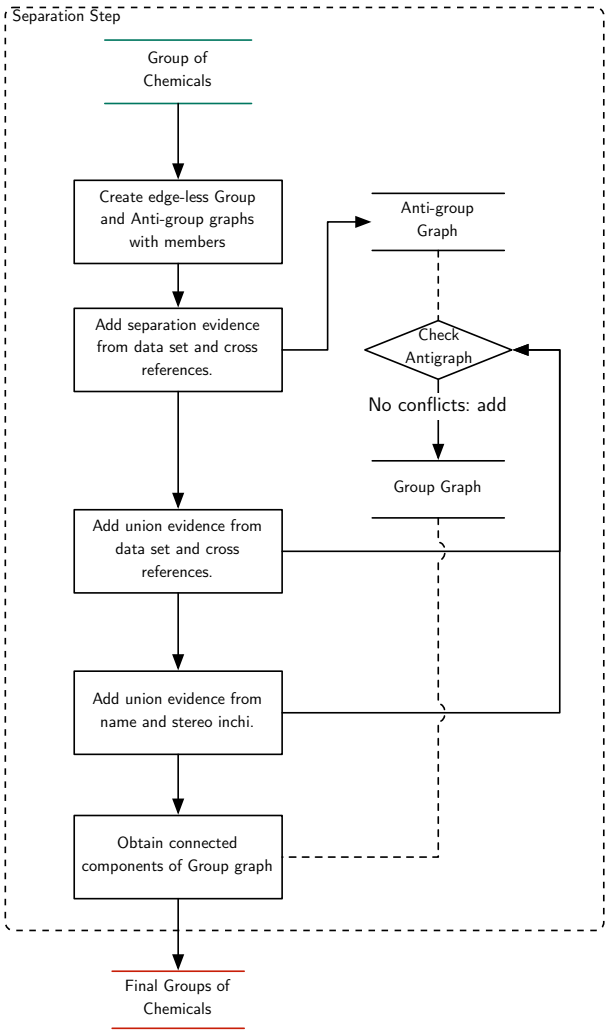



Figure 2.16: Diagram of the separation step that dissects groups of connectivities into more accurate groups. The method relies on cross references, names, synonyms and stereo InChI, which is only trusted under particular conditions. As the method visits the evidence, two graphs store the knowledge retrieved from the evidence: a normal graph stores evidence that two molecules should be considered the same (nodes are molecules, an edge between them means they are equivalent); an anti-graph stores evidence that two molecules (nodes) are different (direct edge between them). Each time that new evidence is inspected, these graphs are updated transitively and checked to see whether no previous evidence – of higher priority – is being contradicted.

| DataSet | Name | InChI |
|---|---|---|
| BioCyc | Queuine | InChI=1S/C12H13N5O3/c13-12-16-10-8(11(20)17-12)5(4-15-10)3-14-6-1-2-7(18)9(6)19/h1-2,4,6-7,9,14,18-19H,3H2,(H2,13,17,20)/p+1 |
| KEGG | Queuine | InChI=1S/C12H15N5O3/c13-12-16-10-8(11(20)17-12)5(4-15-10)3-14-6-1-2-7(18)9(6)19/h1-2,4,6-7,9,14,18-19H,3H2,(H4,13,15,16,17,20) |
| BRENDA | Queuine | InChI=1S/C7H9N5O/c8-1-3-2-10-5-4(3)6(13)12-7(9)11-5/h2H,1,8H2,(H4,9,10,11,12,13)/p+1 |

Table 2.5: Initial group formed after querying for InChI=1S/C12%N5O3/c13-12-16-10-8(11(20)17-12)5(4-15-10)3-14-6-1-2-7(18)9(6)19/ InChI Connectivity, corresponding to Queuine. Even though both database refer to the same molecule, the InChIs calculated from the provided structures have different hydrogen layers. Even though of different connectivity, we show the entry for BRENDA with the same name, as an example of data quality issues.

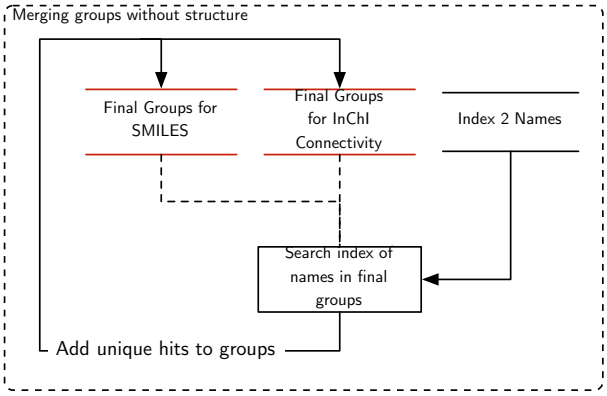



Figure 2.17: Diagram shows the third part of the consolidation algorithm implemented to consolidate sets of small molecules. Once the program generates the groups of molecules for complete structure, generic structure and no structure, it searches the indexes of the names of elements in the group of no structure against the names of defined structure groups and generic structure groups. When unique exact hits are found and the addition of the element with no structure completes a group (meaning that it adds an element from a database that was not previously in the group), then the new element with no structure is added.

the different data sets). However, this approach also looses resolution in certain cases. The InChI connectivity does not include any form of stereo chemistry, which is an advantage, since stereo chemistry is known to be problematic for merging data sets. However, from the standpoint of the database, is probably not desirable to have (R)-2-hydroxystearate and (S)-2-hydroxystearate as the same molecule. Neglecting bond orders and protonation can lead to molecules with very different properties being considered in the same class, but it is necessary to resolve artificial differences due to forced protonation of molecules in the source databases. For these reasons, I included the additional separation steps described to the molecule merge process, which solve most of those resolution problems by using molecule's meta data.

#### 2.3.6.2 Assessment of the method

To assess the quality of the whole database consolidation algorithm, I manually inspected a sample of results, checking whether the groups of molecules formed were correct. In the case of complete groups, this meant assessing whether all the members were equivalent. For incomplete groups, this implied to check whether I could find entries in the missing resources that could be added to the group. After many manual inspections of the results during the development and testing phase, I roughly estimated that the overall probability of success of consolidating correctly a group of molecule should be in the order of ~80% for the method, or a success probability estimate of $\hat{p} \geq 0.8$. Considering that the method treats data differently for molecules with structure, generic structure and no structure, I stratified the sampling accordingly. With this initial estimate we derive a sample size $n$, through a confidence interval, to have a better estimate for the true probability of success, $p$, of the method.

$$p = \hat{p} \pm Error \tag{2.1}$$

$$Error = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{2.2}$$

Solving for $n$, the sample size, and then assuming $\hat{p} = 0.8$, $z_{\frac{\alpha}{2}} = 1.96$ for a

95% confidence and an error of 10%

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{Error}\right)^2 \hat{p}(1-\hat{p}) \tag{2.3}$$

$$n = \left(\frac{1.96}{0.1}\right)^2 0.8(1-0.8) \tag{2.4}$$

$$n = 61.46 \tag{2.5}$$

Equation 2.2 is used normally to derive sample sizes for population proportion problems (in this case, which proportion of the groups formed are correct). This requires that the sample size is sufficiently large and that sample means are distributed normally around the population mean, which is what we want an estimate for. For choosing the sample size, a common assumption would be to set the probability estimate $\hat{p}$ of success at $\hat{p} = 0.5$, where the largest variance is expected, however, as we have a previous estimate, we can use it to reduce the number of samples required. Given $n = 61.46$ we need at least 62 samples from each stratum. I used the sample function from R to obtain random samples for each stratum. Each one of the selected groups was inspected to decide whether it was correctly formed or not.

Manual inspection of 97 groups randomly selected for the completely defined molecules gives a total of 86 correct groups. Calculating a proportionality test with continuity correction[1], generates a 95% confidence interval of $[0.80, 0.94]$ for the probability of correct assignment of groups of molecules with complete structure. The same results for the generic molecules and molecules with no structure can be seen in Table 2.6.

Overall, the manual assessment shows that for groups formed by chemical entities with structure, in the worst case scenario, up to a 20% of error can be expected. This error increases to 30% in the worst case scenario for groups of generic molecules and for groups of molecules with no structure. However, these error rates have more chance of being at ~10% and ~20%, respectively.

Chemical fingerprints and isomorphism comparison are independent checks

[1]The proportionality test was calculated with R using prop.test() method

| Category | $\hat{p}$ | 95% CI |
|---|---|---|
| Defined | 0.89 | $[0.80, 0.94]$ |
| Generic | 0.79 | $[0.69, 0.86]$ |
| No Struct. | 0.79 | $[0.67, 0.88]$ |

Table 2.6: Estimated probability of success as confidence intervals for each of the three types of data that the consolidation method handles. In the case of chemical entities with no structure, we find that between 54% to 74% represent proteins or RNAs names and not small molecules, so these were not considered for the success chance estimation. This reduces the quality of our estimate, producing a wider interval.

that can be applied to the chemical groups consolidated, as I did not use them in the merge process. I calculated the MACCS fingerprint (using CDK) for all molecules in the integration, and then, based on that fingerprint, the Tanimoto similarity between all the molecules that belonged to different regions of the integration. If two molecules are the same, but have not been unified by the approach, they should present the exact same fingerprints, and the similarity between them will be one. Of course, that two molecules have a fingerprint similarity of one is not a sufficient condition for them to be the exact same molecule; for every pair that presented maximum fingerprint similarity, I calculated an ismorphism similarity (using SMSD [124] isomorphism toolkit from the CDK). However, neither fingerprint similarity nor isomorphism similarity account for stereo chemistry (at least in the implementations that I have access to).

The total possible comparisons between different group pairs belonging to disjoints regions[1] in the *H. sapiens* chemical integration amounts to ~3.5 million. Out of these, only 361 group pairs have perfect isomorphism similarity. Table 2.7 summarizes the finding per regions compared. Appendix B shows representatives of the group pairs compared, for the 361 pairs and also for some additional ones that had perfect fingerprint similarity but not perfect isomorphism similarity.

Out of the 361 perfect isomorphism matches, I managed to classify 204 automatically as correctly separated pairs of groups through some very simple stereo chemistry checks (comparing up and down bonds). A manual revision of a sam-

[1]By region I mean each of the sections of the Venn diagram shown in Figure 2.20 ahead.

ple 40 of the remaining 157 pairs shows that ~50% of the pairs are different due to stereo chemistry, and hence correctly separated by the method. The remaining 20 pairs are badly separated groups and should be merged together. Given the sample size, the confidence interval for this probability of erroneously separating groups (for these 157 cases) is $[35\%, 65\%]$. If we assume the worst case for the error of separation at the ~65% estimation for the 157 cases, then 103 pairs of groups should be merged, reducing in 103 groups the total count of small molecules, out of ~2100 groups with structure. This means that the fingerprint and isomorphism check shows that there are at the most ~5% of errors in the groups formation, for groups with structure. Even assumming that the ~65% error applies to the 361 pairs of groups, that would yield 235 pairs of groups to be merged, meaning an error of ~11% of the groups in the chemical consolidation process. These numbers (~5% and ~11%) are in agreement with the 89% probability of success estimated before for the case of completely defined structures, supporting the previous assesment and the estimated rate of error for the chemical unification.

Overall, both assessment methods – manual inspection of generated groups and comparison of structures between groups representing unique small molecules – account at the most for an error rate of ~20% for the unification method.

### 2.3.7 Overview of molecule integration results

The three different types in which the method separates the data have different completeness and consistency levels. The richest stratum (molecules with structures) allows a better integration of the data sets, as it is shown in Figure 2.18, where more groups of chemical entities with representatives from the 3 databases can be formed. Figure 2.19 shows that BRENDA is the database that generates the highest number of singletons and participates in the lowest number of groups with two members (intersection between two databases), where as BioCyc and KEGG tend to form more intersections. BRENDA shows more redundancy between its molecules and has higher numbers of chemical entities without a

| Region | | Pairs | | |
|---|---|---|---|---|
| A | B | Total | DS | Generic |
| BR | BC-BR | 4 | 3 | 0 |
| BR | KG | 10 | 6 | 8 |
| BR | KG-BR | 17 | 10 | 5 |
| BC-BR | KG | 2 | 0 | 1 |
| BC | BR | 8 | 4 | 7 |
| BC | BC-BR | 7 | 1 | 1 |
| BC | KG | 16 | 12 | 12 |
| BC | KG-BR | 8 | 5 | 6 |
| KG-BR | BC-BR | 18 | 12 | 3 |
| KG-BR | KG | 6 | 3 | 2 |
| KG-BC-BR | BR | 72 | 47 | 22 |
| KG-BC-BR | BC | 18 | 3 | 6 |
| KG-BC-BR | BC-BR | 28 | 19 | 1 |
| KG-BC-BR | KG | 24 | 16 | 3 |
| KG-BC-BR | KG-BR | 29 | 20 | 5 |
| KG-BC-BR | KG-BC | 25 | 8 | 12 |
| KG-BC | BR | 18 | 9 | 14 |
| KG-BC | BC | 13 | 7 | 7 |
| KG-BC | BC-BR | 7 | 1 | 4 |
| KG-BC | KG | 15 | 11 | 6 |
| KG-BC | KG-BR | 16 | 7 | 9 |

Table 2.7: Pairs of chemicals with complete isomorphism similarity from the different regions of the unification. BR : BRENDA; BC : BioCyc; KG : KEGG; BR-BC stands for BRENDA-BioCyc only intersect region. DS stands for Different Stereo chemistry, and shows the number of pairs within the total of that row that have different stereo chemistry according to some simple automatic check. More pairs could have different stereo chemistry. The column Generic stands for the count of pairs in each row that are formed by generic molecules.


Figure 2.18: The bar plots summarize how much of the resources could be unified in terms of types chemical entities. A group is a set of chemical entries suspected to represent the same molecule, so the more groups with the 3 databases represented (meaning that the group has a chemical entry from each of the databases), the better. The Y axis shows the number of databases per group, 1 (singleton groups, with an entry only from BioCyc, KEGG, or BRENDA), 2 (groups conformed by representatives of two of the databases), or 3 (groups with a molecule from each of the three databases). The method separates the data in completely defined chemical structures, generic structures and chemical entries with no structure. The leftmost part shows that the chemical entities with structure tend to form a bigger proportion of groups with the 3 databases represented, nearly half of the number singletons (groups with just one database represented), compared to nearly a fourth and less than a fourth in the case generic structures and no structures, respectively. The same tends to happen with groups in which at least two databases are represented.

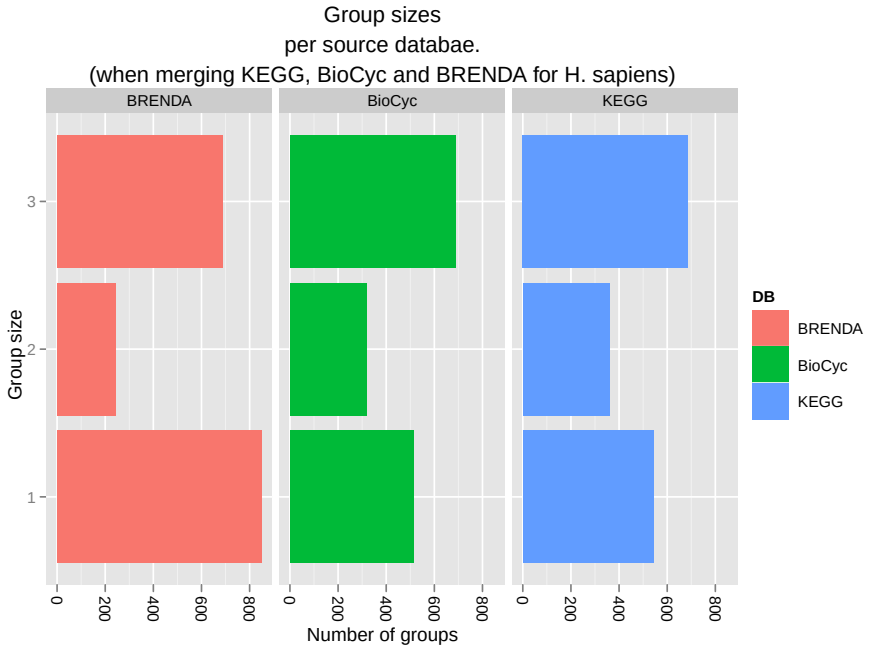



Figure 2.19: The bar plots summarize the participation of each databases in groups with the 3 databases represented (total intersection, obviously same amount for the 3 databases), with 2 databases represented (the database plus one of the two others intersected) and 1 database represented (that database forming a singleton for that chemical entry). BRENDA shows by far the highest number of singletons, while KEGG and BioCyc tend to form more intersections.

chemical structure, which partly explains this.

The Venn diagram in Figure 2.20 shows the intersection for *H. sapiens* of the three metabolism resources integrated. It illustrates that BRENDA is the resource with the highest number of singletons and that KEGG and BioCyc tend to integrate better. The amount that each database adds on its own to the human metabolome collection is considerable, even if we assume errors as high as 20%.

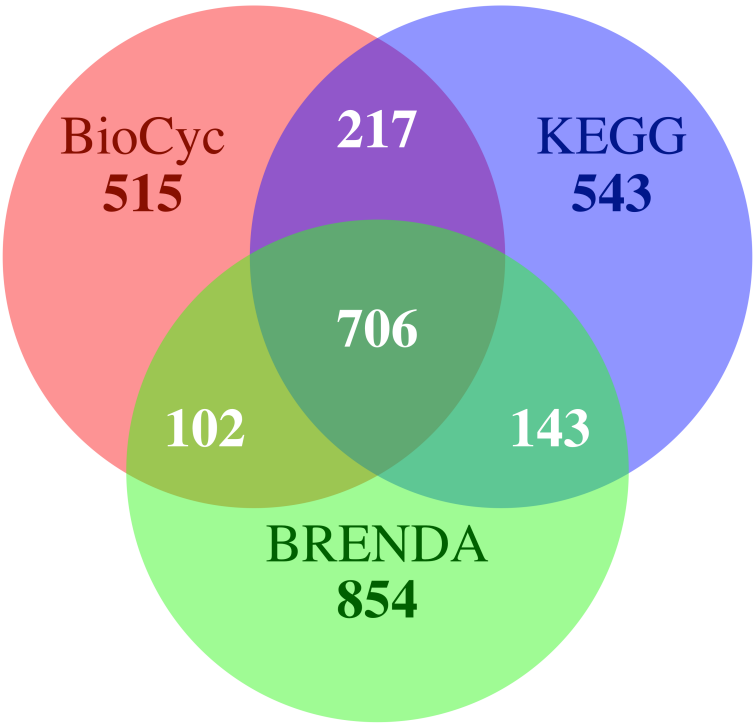



Figure 2.20: Venn diagram (generated with R package VennDiagram [17]) of the intersection of chemical entries in KEGG, BioCyc and BRENDA for *H. sapiens*. Numbers denote the small molecule counts.

Beyond the manual checks that we performed on particular samples of the data in the assessment Section 2.3.6.2, it is also useful to check the overall behaviour of it. Figure 2.21 shows that the distributions of mass of the chemical entities that could not be integrated (singleton groups) has marked differences between the three data sets. Particularly, BioCyc singletons concentrate in the 800 Daltons region, when compared to the other two resources. Looking closer,

this range reveals 44 entries in BioCyc, 24 in KEGG and 22 in BRENDA (which represents 6.6% of the collection of singletons with structures and 8% of the BioCyc collection in this same category). BRENDA and KEGG collections for *H. sapiens* present other biases as well that are unique. Among these, we find 11 compounds with 9 to 11 length prenyl units, most of them involved in the biosynthesis of ubiquinol-10 pathway, that seem to be only present in the BioCyc collection. This happens because both BRENDA and KEGG handle the elements in this pathway as variable length units, and hence do not provide entirely defined structures for these compounds, leaving only the ones in BioCyc available. Also in this mass range we find 8 compounds in BioCyc involved in the thyroid hormone metabolism pathway. This pathway has some steps without assigned enzymes in *H. sapiens*, which would normally avoid our procedure to retrieve results for KEGG and BRENDA, as in the first case the concept of pathways is too broad and not organism specific, and the second case, no pathway grouping is exposed. Additionally, the treatment given to the pathway in KEGG and in BioCyc differs, being the second one much more detailed. None of the compounds selected that participate in the BioCyc thyroid hormone metabolism pathway had links in the HumanCyc web site to KEGG either.

### 2.3.8 Comparing reactions

After comparing small molecules from different datasets, the next challenge is to unify reactions from the different data sets. This is useful since it allows to connect small molecules with proteins, mostly enzymes, through reactions (either as catalysts or reactants). Connecting small molecules to proteins opens the possibilities of stating whether a small molecule can be present in a particular biological compartment (cell type, tissue, etc.) based on a plethora of existing experimental results (gene expression, proteomics, RNA-Seq, etc.) and resources.

Deciding whether two reactions are equal poses a number of problems. The easier case is just to assume that if the reactants, stoichiometries and products are the same, then the reactions are the same. This leaves a number of equivalent reactions still separated. Lets go through some examples:

Reactions which might have been adjusted for a reference pH in a database,

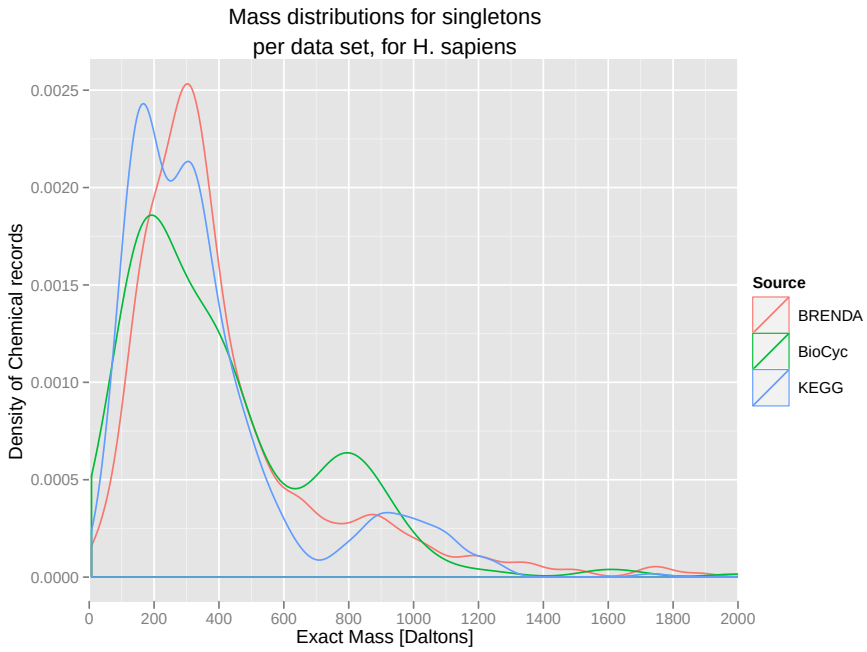



Figure 2.21: The density plot shows the distribution of mass for all the chemical entries, from the three different databases integrated, that do not find any intersection with the other resources (singletons or groups with just one member). As it is a density plot, is very good at keeping the distribution of each set of small molecules, yet the areas under the curve are only comparable within each set and not between them, so the relevant data is the location of the peaks. This result supports the quality of the unification, as it shows that chemicals that could not be consolidated do present marked differences in their mass distributions. Singletons from BioCyc show a bias for a peak centered at 800 Daltons, suggesting that a collection of compounds in this database, centered around this mass, might not be present in the other resources. The same happens with BRENDA at 300 Daltons and KEGG at 280 Daltons, approximately. Identical distributions within the singletons would have been less suggestive that they are actually different molecules. Again this figure shows the two main peaks where molecules seem to concentrate: in the vicinity of 200 Daltons and 800 Daltons.

compared to other sources, might have additional protons and water participating. Such is the case of the serotonin degradation reaction (EC 1.4.3.4), which in BioCyc has an added proton compared to the same equivalent reaction in KEGG. Sometimes these added protons are accompanied by water for automatic balancing purposes. To avoid considering these cases as different, we neglect any differences in protons and water in the reactions. In our *H. sapiens* reaction consolidation we find ~400 classes of unified reactions with these kind of artifacts, that would be otherwise separated. Of these cases, 379 have an additional proton in one of the databases, and 26 show differences due to added water. Ten reaction classes show both water and protons added in some of the databases. Most of the proton additions to reactions come from BioCyc, where reactions are protonated to pH 7. In 357 reaction classes the only difference between members of the class is a single proton, being all the other reactants the same (except for protein participants and tRNA participants).

Both protons and water are readily available to any biochemical reaction inside the cell. A biochemical reaction normally takes place not between single state protonation molecules, but between pseudoisomer groups [1, page 63]. The pseudoisomer group of a molecule (for instance ATP) represents the equilibrium concentration of all the possible protonation states of that molecule at the pH of operation ($ATP^{4-}$, $HATP^{3-}$, $H_2ATP^{2-}$ for ATP). This drives metabolism databases to implement differently the inclusion of protons and water in some reactions. This is because databases sometimes pick different major microspecies (the most abundant protonation form within a pseudoisomer group) for use in their reactions. This can also happen when databases choose a different operation pH.

Another difference might arise from balancing (which can be accumulated with the previous example), where a database shows a version of the same reaction balanced, where others do not, exhibiting different stoichiometries. Example of this is the oxidation 6-Mercaptopurine to 6-thiouric acid, EC number 1.17.3.2 as portrayed in KEGG, under reaction identifier R08235, and BRENDA. In KEGG 2 molecular oxygens and 2 water molecules are consumed, where as in BRENDA only one of each is consumed. This kind of difference is less common, but nevertheless happens. We solve this by setting all stoichiometric coefficients to one for

the comparison (in the comparison objects, not the database).

Reversibility of a reaction can also complicate the comparison of reactions. The truth is that for most reactions, we do not really know the preferred direction (unless a particular pathway context can be used as reference). In most cases, biochemical reactions are also close to equilibrium, which means that there is no preferred direction at all, as the reaction flux is equal in both directions. To understand reaction directions, we need to calculate Gibbs energies of reaction and know the participants concentrations. Because of this lack of knowledge, some databases might represent the reaction in one direction while others in the opposite. This can be solved by comparing the two sets of participants, regardless of the side they are given (but still keeping them separate).

Participating proteins are not always well defined in reactions, and as such, we are left only with a name matching case. We neglect participating proteins at the comparison level, considering them as the same object that we label as "protein participant".

For the sake of this exercise, we do not deal with internal cellular compartments, so reaction participant's sub cellular localizations are neglected for this comparison. These kind of data is seldom available in any case.

Our reaction comparison algorithm starts from the chemical entries unified in the previous part, merging reactions that share small molecules that have been consolidated. The method neglects protein participants (as substrates or products), ignores the direction of the reactions, allows differences in a group of selected molecules (water, protons, oxygen) and neglects differences in stoichiometric coefficients. Given that that many times EC numbers are not one to one mappings to reactions, we avoid using EC numbers for merging reactions, leaving them as a tool for prospective curators to asses the consolidation.

Figure 2.22 shows the intersection of reactions for *H. sapiens* after application of the method. Previous published work [143] claimed that the KEGG, BioCyc, the Edinburgh Human Metabolic Network and the Human Metabolism reconstruction in BiGG only had approximately 3% of intersect for its reactions. Through our method of chemical entities comparison, we achieve more than 10% of total intersection of reactions and more than 23% of reactions with some level of intersection. Given that our chemical integration scheme can have errors, we

identified approximately 50 chemical unification conflicts (that is, chemical entities deemed as different when in fact they are the same) that could increase by a few hundred the number of intersected reactions if they were manually corrected. For a fair comparison, considering that these authors integrate more and different databases, we compared the number of reactions integrated between BioCyc and KEGG for *H. sapiens*, which are two databases used in both studies. According to their supplementary material, the authors manage to match 579 reactions between HumanCyc and KEGG, whereas our Venn diagram in Figure 2.22 shows a total intersection of 670 reactions (422 + 248 regions). For the small molecule match between these two databases, the authors calculate the ratio between intersection and union of both small molecule sets, reporting 28% between BioCyc and KEGG for *H. sapiens*. From our Venn diagram in Figure 2.20 the ratio of intersection to union between BioCyc and KEGG for *H. sapiens* is 41%. Our method achieves higher levels of overlap between metabolism databases for both small molecules and reactions.

The method found 466 different EC numbers annotated in reactions that were unique to each database. These EC numbers did not appear either in any of the unified reactions (either two or three database intersections). The EC Wheels in Figure 2.23 display the distribution of EC numbers that are unique to each of the resources. This again is evidence that the different resources hold a number of unique small molecules that are not shared by the others.

It is useful to compare the number of unique reactions from each database (~2,900, the singletons or zones of no intersection in the Venn diagram) with the numbers of unique EC numbers from each of the databases (466). The general assumption that one EC number maps to one reaction leads to a confusion here, as one would expect these two numbers to be the similar. In fact each EC number can have more than one reaction assigned. HumanCyc has 1,255 EC numbers, 235 map to more than one reaction. These 235 EC numbers map to a total of 742 reactions, so more than 3 reactions per EC number on average. In BRENDA for *H. sapiens* 499 EC numbers (that each map to more than one reaction) map to a total of 1676 reactions, again more that 3 reactions per EC number on average.

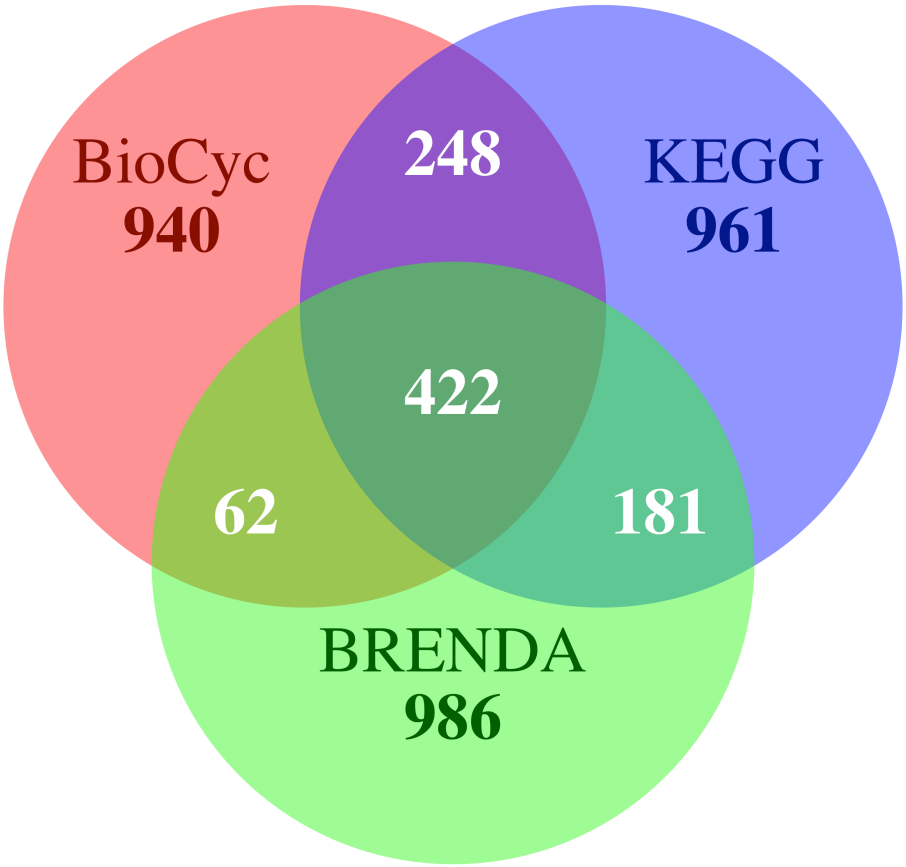



Figure 2.22: Intersection of Reactions in *H. sapiens*.

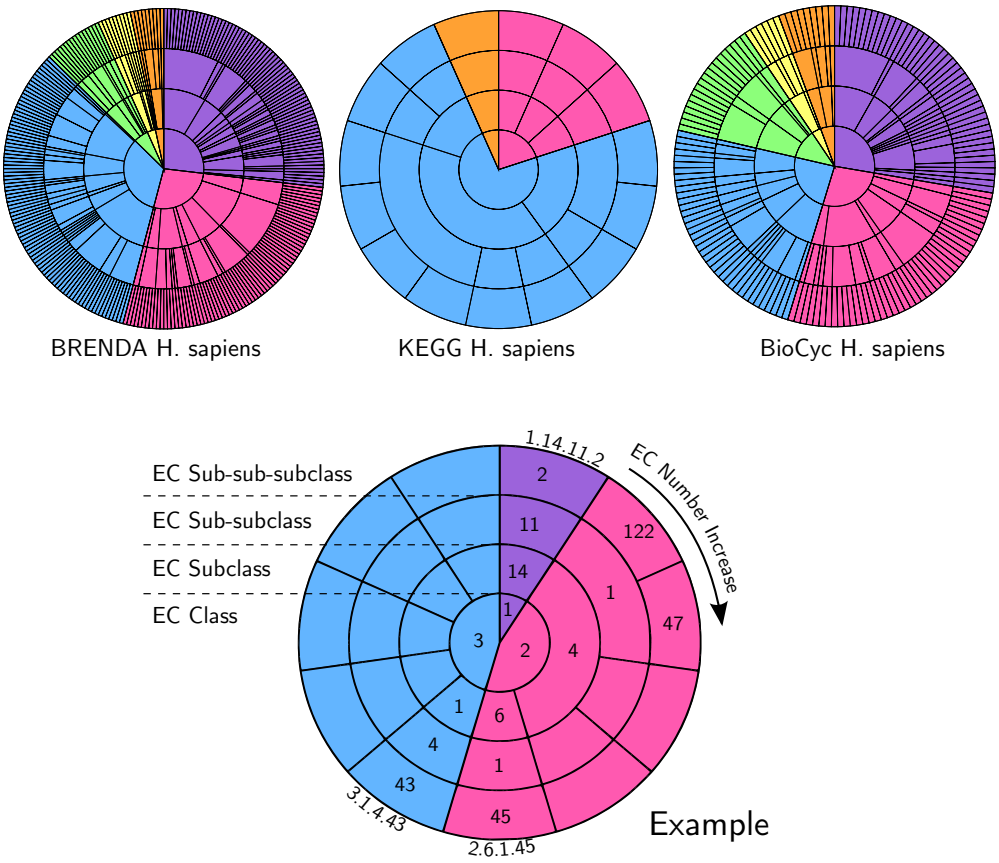



Figure 2.23: EC Wheels representing the distribution of unique EC numbers for BRENDA, KEGG and BioCyc for *H. sapiens*. The EC Wheel starts from angle 0 in clockwise manner, as the example shows. Each subdivision of the inner most circle, and each color, corresponds to a different EC class: purple, pink, blue, green, yellow and orange represent classes 1 (oxidoreductases), 2 (transferases), 3 (hydrolases), 4 (lyases), 5 (isomerases), and 6 (ligases) respectively. Concentric belts represent sub classes. The separated regions in the outer concentric belt represents a complete EC number (see EC number 1.14.11.2 in the example). Overall, there are 466 distinct EC numbers not shared by the databases (out of a total of 1588), 292 coming from BRENDA (out of 1221), 159 from BioCyc (out of 1255) and 15 from KEGG (out of 1056). Many of these unique EC numbers did not have an entry in UniProt, in many cases described activities but with no protein catalyst known. The amount of unique ECs on each database supports the fact that they contribute with many different metabolites to a metabolome collection.

Many reactions as well do not have an EC number assigned. However, the number of unique reactions is still high in the unification, as we tend to have 6 reactions per unique different EC on average. This higher reaction to EC ratio is probably due to errors in the unification process. If we assume an approximate 15% of error at the chemical entity unification level, and each reaction has an average of 3.7 participants without including water, then this 15% error at small molecule unification level translates to as much as 48% error at the reaction level (if we assume a 20% error, this translates, on average for a 4 participant reaction to 59% chance of error).

As bad as they seem, all these error estimations are upper bounds for the error for a few reasons. First not all metabolites participate in the same number of reactions, and those that tend to be more connected tend to be better annotated (correct structures, more cross references, more synonyms), leading surely to a lower error rate of chemical unification for them. Also, most reactions have only completely defined structures, which means their unification error is probably closer to 10% than to 15%, again lowering the overall reaction unification probability error.

For example EC number 3.1.4.11 appears twice in the *H. sapiens* unification, one version from BioCyc and one from BRENDA. The reactions are the same, yet in BioCyc participant 1,2-diacyl-sn-glycerol has a generic structure, where as in BRENDA the equivalent entry diacylglycerol does not have a structure. None of them have a matching synonym or cross reference, so they are left as different entities, generating two versions of reaction 3.1.4.11.

From the unique EC numbers for each data set, I mapped UniProt identifiers, which I used for protein function enrichment analysis[1] using DAVID[2] [58]. The Cellular compartment branch of the gene ontology showed BioCyc with a bias towards localization in Golgi, Nucleus and other intracellular parts, whereas the BRENDA related proteins had a bias towards localization in the extracellular region. There are few EC numbers found only in KEGG, which map to only a handful of UniProt identifiers, making these enrichment analyses worthless for

[1] Section C. 1 in Appendix C shows the statistical explanation of all enrichment analyses done.

[2] DAVID is a widely used enrichment analysis web tool for proteins, which incorporates a number of protein functional classification systems.

KEGG in this case.

The biological process branch of the gene ontology enrichment showed a bias in BioCyc unique ECs for information processing (DNA, RNA processing and post translational modifications), which makes sense with the compartmentalization enrichment. These ECs also had presence in lipid and glycan metabolism. For BRENDA, this same part of the ontology shows enrichment for protein catabolism and protein modification biological processes.

Inspection of the molecular function branch of the gene ontology shows enrichment in helicases and other enzyme activities which relate to information processing for BioCyc EC numbers. For BRENDA, this branch shows enrichment in peptidases enzymatic activity, which make sense with the biological processes enriched for the set of ECs unique to BRENDA.

These enrichment analyses show again that the different databases sometimes focus on particular domains of knowledge with different levels of details, which further supports the fact that we find small molecules that are unique to each database.

## 2.4 Comparison against the Human Metabolome Database

Using the same merge process described in Section 2.3.6.1, I compared the HMDB with the merge of BRENDA, KEGG and BioCyc for human. Figure 2.24 shows the results of the merge in a Venn diagram.

A high proportion of HMDB records are not matched with elements from any of the other databases, only 1288 records from HMDB are matched with records from other databases. KEGG has the higher overlap with the HMDB (959 chemical entries), followed by BioCyc (913 chemical entries). Figure 2.25 shows the mass distributions for the unique entries to each of the data sets, and indicates that more than two thirds of the unique entries from HMDB are lipids.

Using the Metabolite Biological Role (MBRole) tool [14], I did enrichment analysis using different mass sub sets of the unique HMDB entries, according to

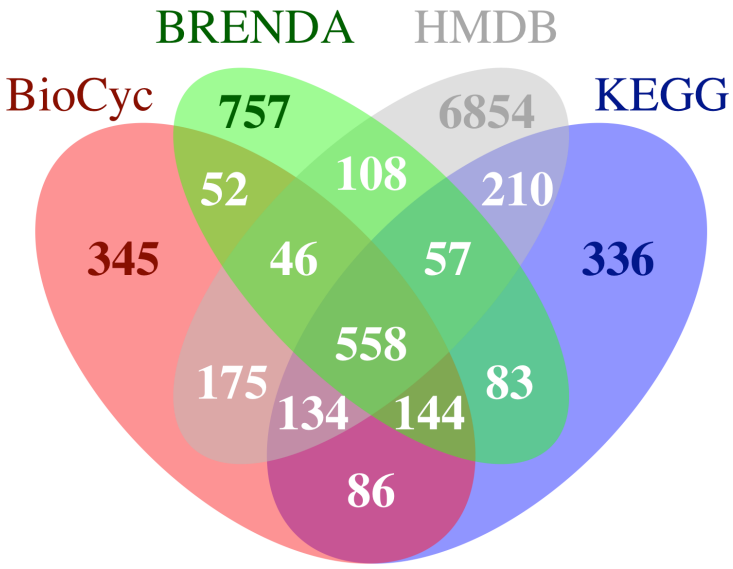



Figure 2.24: Venn diagram showing the intersect between the different *H. sapiens* metabolism databases and the HMDB. Numbers denote the small molecule counts.

the main peaks in Figure 2.25. Through these analyses, Figure 2.28 supports the high concentration of lipids by showing that most of the unique HMDB entries in the range of 500 to 1000 Daltons belong to lipid categories in the HMDB chemical taxonomy classification. In contrast, Figure 2.26 shows that the unique 1,654 HMDB entries in the range 0 to 500 Daltons are less dominated by lipid categories. Nevertheless, Figure 2.27 shows the main categories of lipids present in this range, which include a total of 380 different HMDB entries, ~23% of the HMDB entries in the 0 to 500 Daltons region. The remaining 713 unique HMDB entries above the 1000 Daltons are ~90% lipids according to enrichment analysis, where ~56% are glycolipids.

Overall, enrichment analysis of the unique elements of HMDB shows that out of the 6,533 elements of HMDB that have a chemical structure and are not found in the other databases, only ~25% are non lipid small molecules, 1,635 HMDB

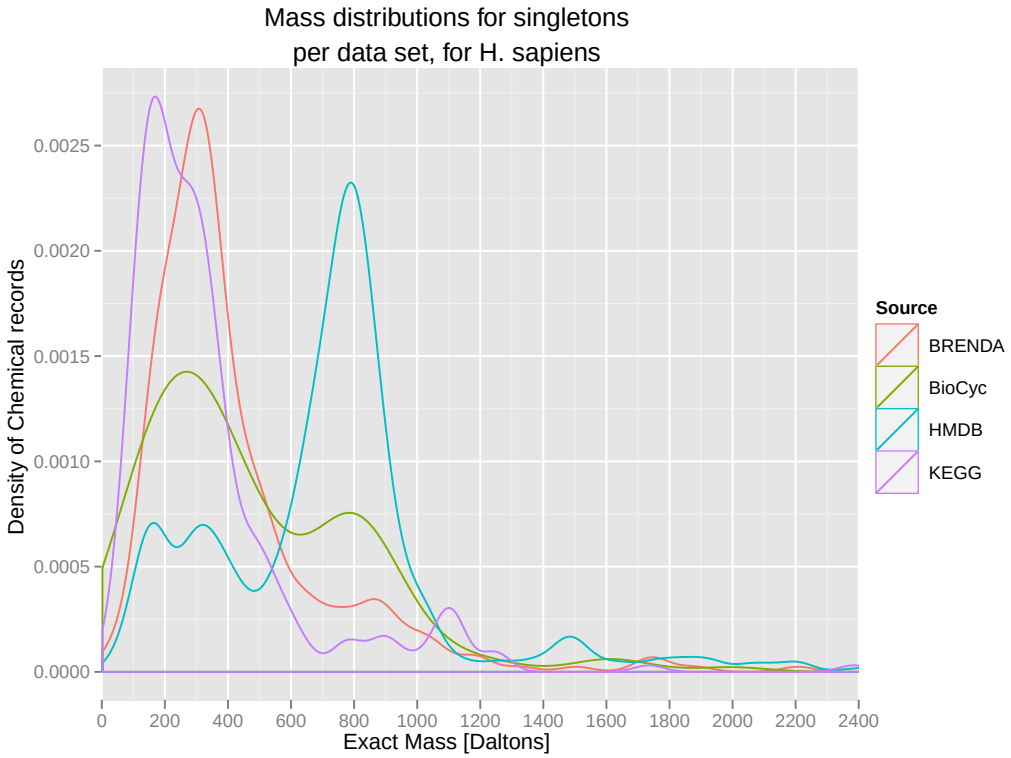



Figure 2.25: Mass distributions of all the singleton regions (small molecules unique to each database). The peak between 500 and 1,000 Daltons for HMDB consists of 4,166 molecules not found in the other resources, at least 3,876 are lipids (recognized by naming nomenclature), Figure 2.28 shows details. The region between 0 and 500 Daltons contains 1,654 entries unique to HMDB, ~23% of these are lipids. Figures 2.26 and 2.27 show the 1,654 unique HMDB entries in the range 500 to 1000 Daltons classified according to the HMDB Chemical Taxonomy.

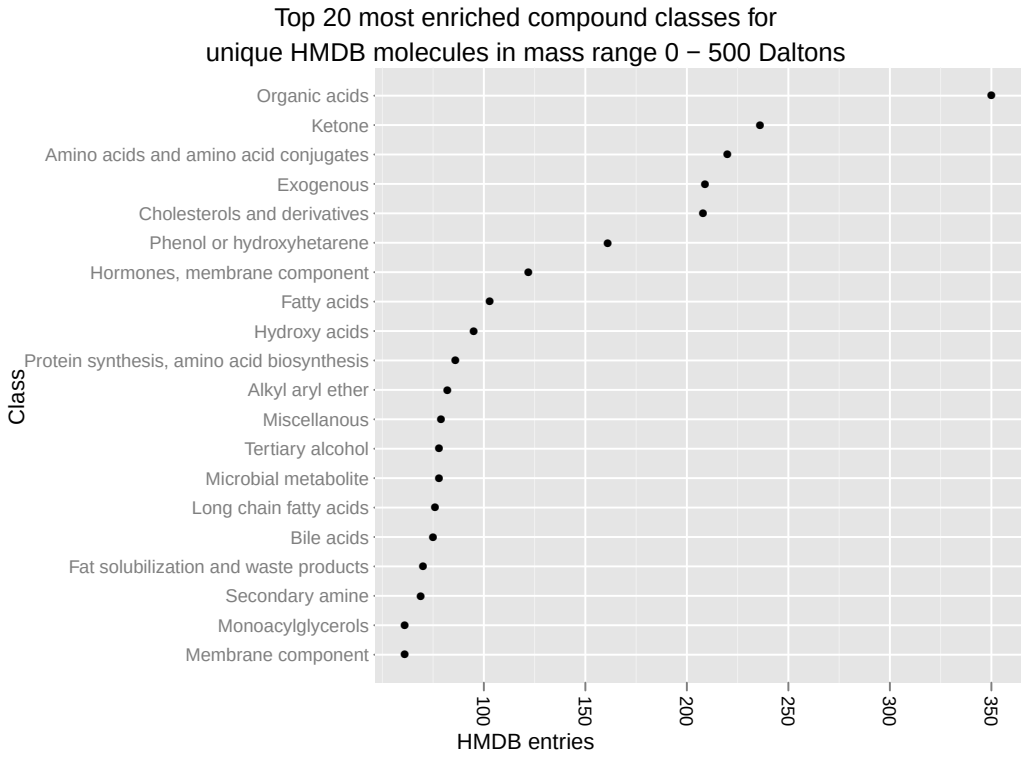



Figure 2.26: Top 20 most enriched categories of the HMDB Chemical Taxonomy for the unique HMDB entries in the mass range 0 to 500 Daltons. This mass range concentrates most of the non lipid elements from HMDB that are not found in the other databases by the unification method. The categories shown are not exclusive, a small molecule can be assigned to more than one. All these categories have significant p-values for enrichment ($p\text{-}value < 10^{-5}$), but given the high similarity of enrichment p-values, they are sorted by the number of molecules in each category.

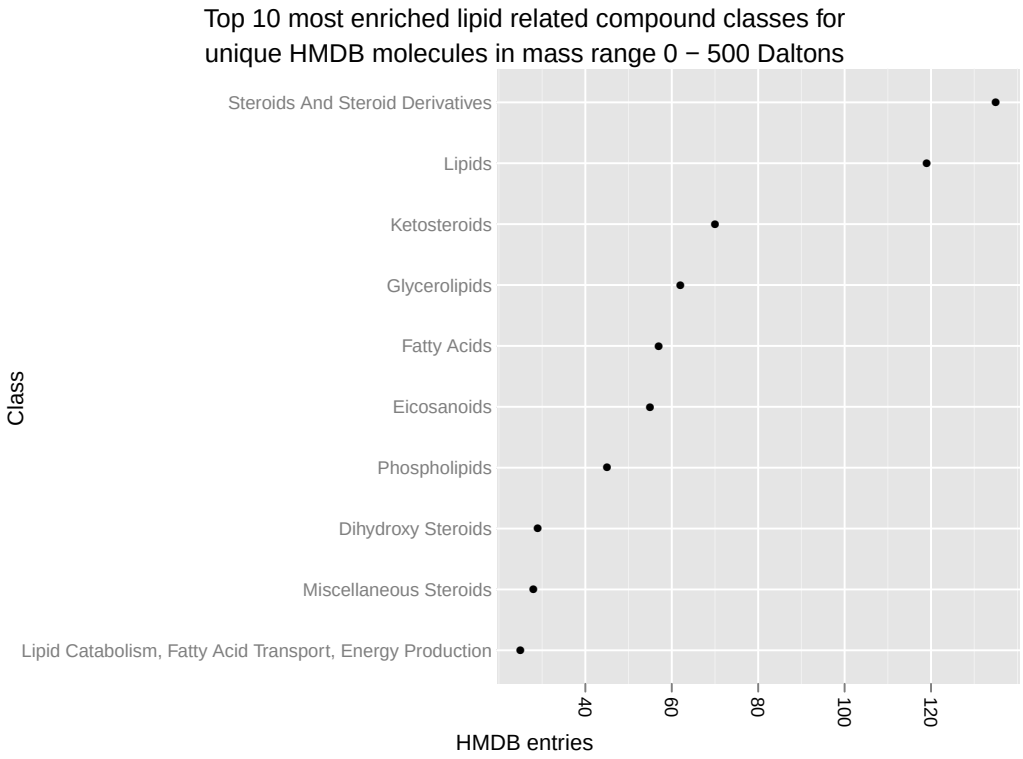



Figure 2.27: Top 10 most enriched categories of the HMDB Chemical Taxonomy related to lipids for the unique HMDB entries in the mass range 0 to 500 Daltons. Counting unique entries, these categories include a total of 380 different HMDB entries.

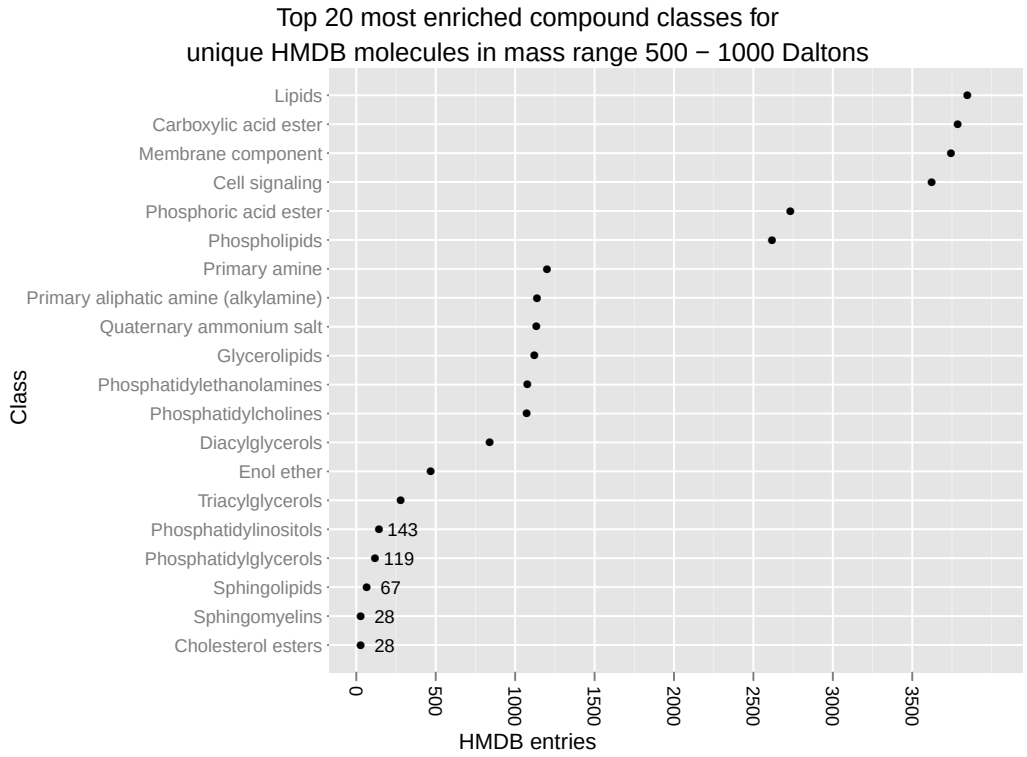



Figure 2.28: Top 20 most enriched categories of the HMDB Chemical Taxonomy related to unique HMDB entries in the mass range 500 to 1,000 Daltons. As the graph shows, most categories are lipid related, and the general category "lipids" holds more than ~3,700 HMDB entries for the 500 to 1,000 mass range.

entries.

Metabolism databases add 1,803 small molecules that do not appear in HMDB. Even in the worst case scenario of a 20% error, this would still mean ~1,400 small molecules that do not seem to be present in HMDB.

## 2.5 Comparison of metabolism databases based species metabolomes

Using the chemical and reaction unification methods, I built consolidated data sets in BioWarehouse for *H. sapiens*, *M. musculus*, *S. cerevisiae* and *E. coli* based on BRENDA, BioCyc and KEGG. Using the same unification method, I compare these consolidated data sets. Figure 2.29 shows the intersection of small molecules for these organisms.

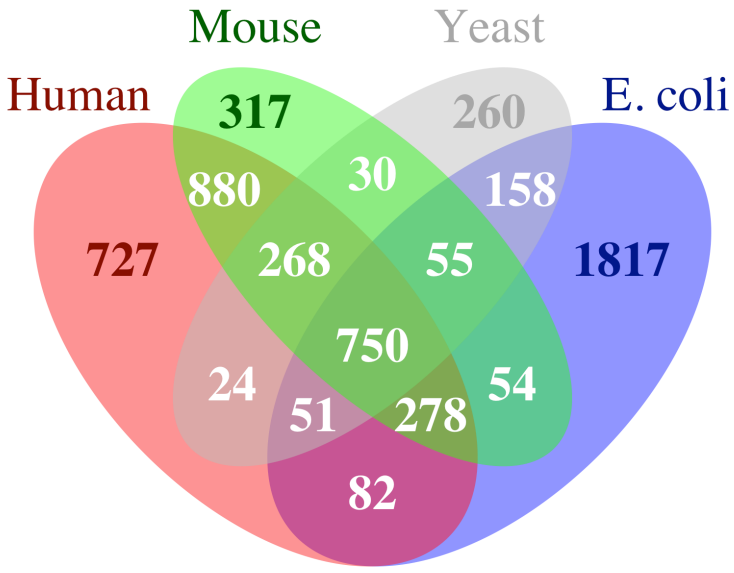



Figure 2.29: Venn diagram for multi species comparison of small molecules between *H. sapiens*, *M. musculus*, *E. coli* and *S. cerevisiae*. Numbers denote the small molecule counts.

As it could be expected, the biggest overlap for *H. sapiens* in terms of small

molecules is with *M. musculus*. *S. cerevisiae* has a bigger overlap with the two other eukaryotes ($24 + 268 + 30 = 322$ common small molecules that are not present in *E. coli*) rather than to its unicellular counterpart, *E. coli* (only 158 common small molecules present only in *S. cerevisiae* and *E. coli*), suggesting that the chemistry of *S. cerevisiae* might be more associated to its eukaryotic character rather than its unicellular nature.

Figure 2.30 shows the distribution of masses for the unique small molecules of each organism (that means, regions 727, 317, 260 and 1817 that belong respectively to unique *H. sapiens*, *M. musculus*, *S. cerevisiae* and *E. coli* small molecules), where we can see that both higher eukaryotes tend to be distributed towards slightly bigger metabolites (in the 200 to 400 Daltons range), where as both unicellular organisms display distributions towards smaller metabolites (in the 150 to 200 Daltons). This might suggest some increase of chemical complexity when we step from unicellular to multi cellular organisms. Using the union of *H. sapiens* and *M. musculus* that does not intersect *S. cerevisiae* and *E. coli* also confirms this trend.

To dive into the singleton regions, I obtained from the database the unique set of ChEBI identifiers available for the small molecules in each region that is unique to each organism. This search retrieves 106 ChEBI identifiers for *H. sapiens*, 80 for *M. musculus*, 101 for *S. cerevisiae*, and 753 for *E. coli*. Unfortunately not all molecules have a ChEBI cross references, but the retrieved ones can well be considered as samples of each region.

With ChEBI identifiers, and using the MBRole tool, I performed enrichment analysis to characterize through the ChEBI ontology these different sets of molecules. Unfortunately MBRole maps the given sets of ChEBI entries to very few roles, and poor results are obtained. This is partly because MBRole restricts the search space to ChEBI 3 stars entries only.

BiNGO[96] is a Cytoscape[138][1] plug-in for enrichment analysis of gene sets through the Gene Ontology. I modified BiNGO version 2.42 to run the same type of enrichment analysis, but using the ChEBI Ontology, including both the

[1] Cytoscape is a widely used network visualization program, with a special emphasis in biological data and protein protein interactions

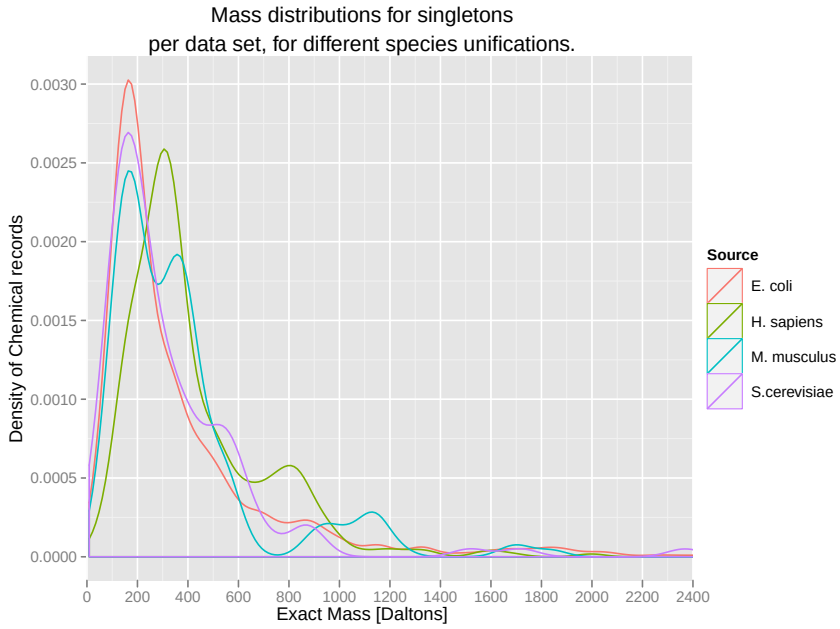



Figure 2.30: Mass distribution for the different species. Each curve is a density curve, so the area for each of the curves is one. This means that each curve is normalized by the number of molecules in each of the sets, so the distribution differences are well appreciated, but areas under each curve represent different count of molecules. There are 1,359 different molecules with mass for *E. coli* that could not be mapped in the other organisms, 400 in *H. sapiens*, 169 in *M. musculus* and 115 in *S. cerevisiae*.

chemical ontology and the role ontology. Unfortunately as well, the sparseness of roles annotation and the low number of unique elements in each region of the Venn diagram that had ChEBI annotations, produce poor results for many of the small molecule sets. In most cases, only high level chemical categories, such as organic molecular entity or oxygen molecule entity, showed high levels of enrichment, but they are of little use to understand the biological context as they are very broad. I explore here, organism by organism, the enrichment analysis results of the singleton and combined regions, looking for relevant patterns.

### 2.5.1 *H. sapiens* enrichment analyses

In the case of *H. sapiens*, some lipid classes show a certain level of enrichment, this is based on 34 molecules (out of the total 106 different ChEBI entries for *H. sapiens*, approximately a third of it). Figure 2.31 shows these enriched lipid categories as a result of the analysis done with BiNGO. This might be indication that a good amount of the 727 small molecules from *H. sapiens* that could not be merged with the sets from other organisms could be lipid related.

Through the same analysis for *H. sapiens*, Figure 2.32 shows the main biological roles uncovered in this data set of 727 unique molecules, where we can see biases towards enzyme inhibitors, and drugs.

Another way of analyzing these sets of unique molecules is to trace back from them to the enzymes catalyzing the reactions that produce/consume them. The database can map the 727 unique *H. sapiens* molecules to 256 UniProt protein identifiers, where only 136 are unique. This reduction has a number of causes. Many of the unique molecules come from BRENDA, which many times implies that the enzyme does not have a UniProt identifier. Also BioCyc databases add a number of small molecules as enzyme inhibitors only, specially for highly curated organisms like *E. coli* or *H. sapiens*, but only specifying the type of enzyme they inhibit, not the particular protein entry in UniProt. Additionally, there are number of proteins that appear recurrently, 40% of them are linked to more than one small molecule, and even some of them, like Diphosphoinositol polyphosphate phosphohydrolase (O95989) or Sulfotransferase 1A1 (P50225), are

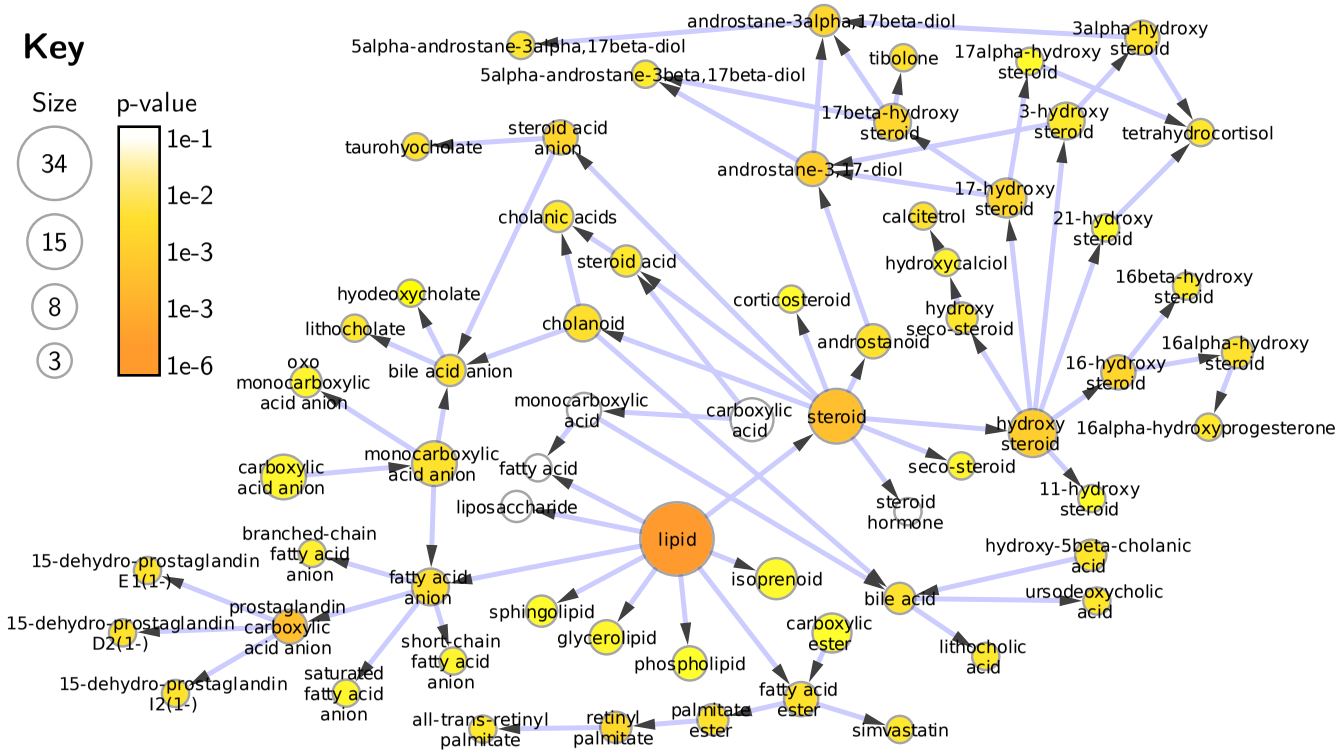



Figure 2.31: Enrichment analysis graph depicting the lipids categories found enriched in the set of *H. sapiens* unique molecules. Many of the lipids shown appear as unique to *H. sapiens*, when they should be part of *M. musculus* as well, because they are not annotated in the other mammal data sets. Colors towards orange are more enriched (lower p-value), bigger circles mean more ChEBI entities from the analyzed data set involved. This only reflects though the presence of 34 ChEBI entries, out of the unique 106 different ChEBI entries annotated in the set of 127 unique *H. sapiens* metabolites.

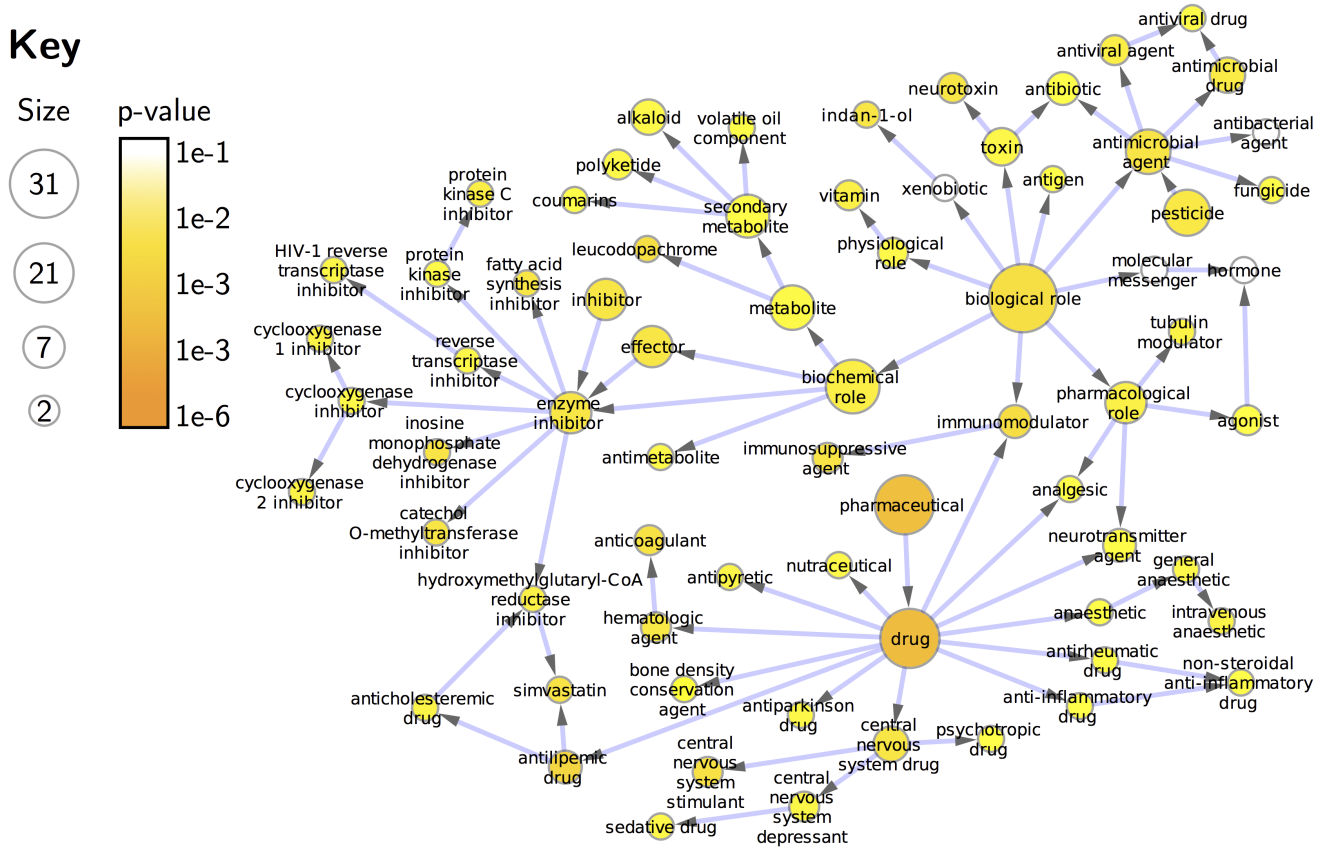



Figure 2.32: Enrichment analysis graph showing the main ChEBI roles enriched in the set of small molecules only found in *H. sapiens* metabolism databases.

| Biological Process | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| Phosphoinositide metabolic process | 9.5 | 1.9E-11 | 19.4 |
| Glycerophospholipid metabolic process | 10.3 | 2.5E-10 | 13.0 |
| Glycerolipid metabolic process | 11.1 | 8.6E-10 | 10.2 |
| Lipid phosphorylation | 5.6 | 2.2E-09 | 51.7 |
| Fucose metabolic process | 5.6 | 5.1E-09 | 46.0 |
| Phospholipid metabolic process | 11.1 | 6.1E-09 | 8.7 |
| Glycoprotein biosynthetic process | 10.3 | 7.5E-09 | 9.7 |
| Steroid metabolic process | 11.1 | 1.3E-08 | 8.2 |
| Protein amino acid glycosylation | 8.7 | 1.1E-07 | 10.2 |
| Biopolymer glycosylation | 8.7 | 1.1E-07 | 10.2 |
| Fatty acid metabolic process | 9.5 | 7.9E-07 | 7.2 |
| Lipid modification | 6.3 | 1.7E-06 | 13.7 |
| Carboxylic acid metabolic process | 14.3 | 3.3E-06 | 3.8 |
| Cellular carbohydrate catabolic process | 6.3 | 6.8E-06 | 11.1 |

Table 2.8: Biological processes from Gene Ontology enriched for proteins linked to unique *H. sapiens* small molecules. The % refers to the percentage of proteins associated to that biological process.

linked to as many as 9 small molecules. Table 2.8 shows the most enriched categories of the biological process of Gene Ontology using DAVID. Many lipid metabolism related processes are enriched, supporting the result shown by the ChEBI ontology enrichment analysis.

### 2.5.2 *E. coli* enrichment analyses.

In the case of *E. coli* we see a number of metal ions included by the metabolism databases. This is probably due to the existing research in metals resistance by bacteria through efflux pumps, which as unicellular organisms are exposed to a variety of environments. In contrast with *H. sapiens*, *E. coli* shows low enrichment of lipids, but higher enrichment in sugars, other carbohydrates derivatives, and antibiotics. The latter are probably added to some of the databases as inhibitors of bacterial enzymes, in some cases because they are produced by the bacteria, or because of the inactivation reactions that the bacteria has evolved to neutralize them. Table 2.9 shows some of these enriched categories, along with those that seem more interesting from the functional point of view. For the

most enriched ChEBI categories in *E. coli* and for more detail on some of the categories, such as the low enrichment of lipids, see Section C. 2 in Appendix C.

Table 2.10 shows the result of the enrichment analysis of *E. coli* protein identifiers related to the unique *E. coli* small molecules. The protein enrichment strongly supports the association of this sets of unique small molecules to carbohydrates.

### 2.5.3 *S. cerevisiae* enrichment analyses

*S. cerevisiae* shows again a lower enrichment of lipids than *H. sapiens*, although some fungus fatty acids and steroids appear, such as ergosterols. *S. cerevisiae* shows enrichment as well in cyclic compounds, including some lactones, porphyrinogens, and many nucleobase containing cyclic structures. However, most of the nucleotides seen are errors of the integration, as those compounds should be part of the other organisms as well. One of the *S. cerevisiae* resources had a number of duplicated nucleotides, this generates problems in the integration. These nucleotides do appear in the other organisms. Table 2.11 summarizes the main interesting ChEBI classes that are enriched in the set of *S. cerevisiae* "unique" small molecules. Section C. 3 in Appendix C shows the most enriched ChEBI categories for reference.

Enrichment analysis of the proteins linked to these small molecules in *S. cerevisiae* produces a rather puzzling result: many of the proteins enriched are linked to viral activities. This analysis as well shows that these small molecules are also related to the sporulation process and the Ehrlich pathway, both unique to *S. cerevisiae* within the organisms analyzed. Table 2.12 shows these enrichments in the Biological process branch of the Gene Ontology.

### 2.5.4 *M. musculus* enrichment analyses

For *M. musculus*, the 80 ChEBI entries mapped show very little enrichment of roles or relevant chemical categories like carbohydrates, lipids or sugars. This is

| ChEBI class | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| Heterocyclic antibiotic | 3.3 | 7.6E-09 | 5.4 |
| Organonitrogen heterocyclic antibiotic | 3.2 | 1.2E-08 | 5.5 |
| beta-lactam antibiotic | 3.0 | 3.9E-08 | 5.5 |
| beta-lactam | 3.0 | 6.4E-08 | 5.3 |
| Antibiotic | 5.4 | 4.4E-07 | 3.0 |
| Lactam | 3.0 | 7.2E-07 | 4.6 |
| Cyclic amide | 3.0 | 7.3E-07 | 4.6 |
| Toxin | 5.8 | 2.2E-06 | 2.7 |
| Lipid A oxoanion | 0.9 | 4.5E-06 | 25.9 |
| Antibacterial agent | 2.9 | 5.0E-06 | 4.3 |
| Metal cation | 2.2 | 5.2E-06 | 5.7 |
| Cephalosporin | 1.6 | 5.2E-06 | 8.4 |
| Carbohydrate acid derivative anion | 1.9 | 7.7E-06 | 6.5 |
| Carbohydrate | 8.4 | 2.0E-04 | 1.9 |
| Monosaccharide derivative | 7.5 | 1.2E-03 | 1.8 |
| Organophosphorus compound | 13.5 | 2.3E-02 | 1.4 |
| Sulfur molecular entity | 8.4 | 7.5E-02 | 1.3 |
| Carbohydrate derivative | 16.8 | 1.2E-01 | 1.2 |

Table 2.9: Interesting ChEBI classes enriched in the set of unique *E. coli* molecules. Even though these are not the most enriched categories, they still have relevant corrected p-values, and they tend to be more informative than high level ChEBI classes. This list covers 283 ChEBI entities of the 753 unique ChEBI entities mapped to the unique ChEBI small molecules, so it is a very reasonable sample of it. The rest of the entities tend to widespread into many different high level categories.

| Biological Process | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| Polysaccharide biosynthetic process | 11.3 | 2.9E-31 | 4.4 |
| Carbohydrate catabolic process | 12.1 | 4.9E-31 | 4.1 |
| Carbohydrate biosynthetic process | 12.0 | 1.4E-30 | 4.1 |
| Polysaccharide metabolic process | 12.5 | 1.4E-28 | 3.7 |
| Oxidation reduction | 22.5 | 2.9E-27 | 2.3 |
| Cellular carbohydrate biosynthetic process | 8.8 | 6.7E-21 | 3.9 |
| Cellular polysaccharide metabolic process | 8.2 | 9.6E-21 | 4.0 |
| Cellular cell wall macromolecule metabolic process | 3.9 | 9.2E-20 | 9.4 |
| Peptidoglycan biosynthetic process | 3.7 | 3.1E-19 | 9.5 |
| Aminoglycan biosynthetic process | 3.7 | 3.1E-19 | 9.5 |
| Glycosaminoglycan biosynthetic process | 3.7 | 3.1E-19 | 9.5 |
| Cellular polysaccharide biosynthetic process | 7.7 | 3.1E-19 | 4.0 |
| Cellular component macromolecule biosynthetic process | 3.7 | 8.7E-19 | 9.1 |
| Cell wall macromolecule biosynthetic process | 3.7 | 8.7E-19 | 9.1 |
| Anaerobic respiration | 4.8 | 1.9E-18 | 6.4 |
| Lipid biosynthetic process | 8.4 | 3.5E-18 | 3.6 |
| Lipopolysaccharide metabolic process | 6.6 | 5.0E-18 | 4.3 |
| Lipopolysaccharide biosynthetic process | 6.5 | 8.2E-18 | 4.4 |
| Peptidoglycan-based cell wall biogenesis | 3.7 | 1.1E-17 | 8.4 |
| Cell wall biogenesis | 3.7 | 1.1E-17 | 8.4 |

Table 2.10: Biological processes from Gene Ontology enriched for proteins linked to unique *E. coli* small molecules. The % refers to the percentage of proteins associated to that biological process. The analysis shows a strong bias towards carbohydrates, polysaccharides and cell wall related processes. Bacterial cell walls, rich in peptidoglycans, polysaccharides and other polymers are massive in term of material demand, and unique to *E. coli* in the context of the four model organisms compared.

| ChEBI class | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| Fatty acid ester | 4.0 | 4.2E-03 | 21.0 |
| 3-beta-sterol | 4.0 | 7.6E-03 | 14.6 |
| Nucleobase-containing molecular entity | 10.9 | 4.5E-02 | 2.4 |
| Phosphoric acid derivative | 20.8 | 4.5E-02 | 1.8 |
| Lactone | 5.0 | 5.3E-02 | 3.6 |
| Ergosterol | 1.0 | 4.9E-02 | 94.4 |
| Fecosterol | 1.0 | 2.6E-02 | 472.1 |
| Carbohydrate derivative | 19.8 | 1.6E-01 | 1.4 |

Table 2.11: Interesting ChEBI classes enriched in the set of unique *S. cerevisiae* molecules. Even though these are not the most enriched categories, they still have relevant corrected p-values, and they tend to be more informative than high level ChEBI classes. This list covers 41 ChEBI entities of the 101 unique ChEBI entities mapped to the unique ChEBI small molecules, so it is a very reasonable sample of it. The rest of the entities tend to widespread into many different high level categories.

really to be expected, as the organism is being compared to *H. sapiens*, so very few, if not none, unique small molecules should be expected. Section C. 4 in Appendix C includes the Top 15 most enriched ChEBI categories for reference.

### 2.5.5 Enrichment analysis of joint regions

In the comparison that Figure 2.29 shows, one would expect to see a higher overlap between the two mammalian species, and the differences seen must probably be artifacts of annotation levels and unification defects. A much robust set should be the union of the singletons regions of *H. sapiens* and *M. musculus* (regions sized 727 and 317) and the intersect between them that is not shared with *S. cerevisiae* and *E. coli* (region size 880). This would be a set of metabolites unique to mammals when compared to *S. cerevisiae* and *E. coli*.

From this set of 1,924 small molecules, 630 had a ChEBI cross reference, reaching 699 different ChEBI identifiers. The ChEBI enrichment analysis has a much more consistent result this time, where lipids represent nearly one third of the whole set, and steroids play a major part of this enrichment. Table 2.13

| Biological Process | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| Developmental maturation | 10.9 | 8.4E-45 | 12.9 |
| Virion assembly | 10.9 | 8.4E-45 | 12.9 |
| Viral assembly, maturation, egress, and release | 10.9 | 8.4E-45 | 12.9 |
| Viral procapsid maturation | 10.9 | 8.4E-45 | 12.9 |
| Cellular component assembly involved in morphogenesis | 10.9 | 8.4E-45 | 12.9 |
| Viral infectious cycle | 10.9 | 8.4E-45 | 12.9 |
| Viral reproductive process | 10.9 | 8.4E-45 | 12.9 |
| Viral capsid assembly | 10.9 | 8.4E-45 | 12.9 |
| Viral reproduction | 10.9 | 1.1E-43 | 12.6 |
| DNA integration | 10.9 | 7.5E-41 | 11.8 |
| Transposition, RNA-mediated | 11.4 | 2.4E-27 | 7.0 |
| Transposition | 11.4 | 2.9E-26 | 6.7 |
| Protein amino acid dephosphorylation | 6.5 | 1.0E-19 | 9.5 |
| Dephosphorylation | 7.0 | 1.1E-16 | 7.0 |
| DNA recombination | 11.9 | 8.7E-14 | 3.4 |
| Phosphate metabolic proc. | 14.5 | 1.7E-11 | 2.6 |
| Phosphorus metabolic proc. | 14.5 | 3.6E-11 | 2.6 |
| Amino acid catabolic proc. to alcohol via Ehrlich pathway | 2.8 | 5.4E-11 | 13.2 |
| Amino acid catabolic proc. via Ehrlich pathway | 2.8 | 5.4E-11 | 13.2 |
| Oxidation reduction | 16.3 | 6.2E-11 | 2.4 |

Table 2.12: Biological processes enriched in the set of proteins related to the unique *S. cerevisiae* molecules. Surprisingly, a number of proteins are related to viral activities, that probably use *S. cerevisiae* as host cell. There are also enriched categories for information processing, coherent with the number of nucleotides found, but again part of the same error. The developmental maturation and cellular component assembly involved in morphogenesis are both associated to the *S. cerevisiae* sporulation process, which is unique when compared to the other organisms considered. The Ehrlich pathway for fuse alcohols is also very specific to *S. cerevisiae*.

| ChEBI class | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| Steroid | 17.8 | 2.4E-53 | 5.9 |
| Hydroxy steroid | 12.2 | 3.7E-42 | 7.3 |
| Lipid | 31.3 | 1.1E-34 | 2.5 |
| Organic polycyclic compound | 18.1 | 1.2E-31 | 3.5 |
| Polycyclic compound | 20.0 | 3.6E-25 | 2.8 |
| Organic hydroxy compound | 19.4 | 3.1E-22 | 2.6 |
| Alcohol | 17.1 | 1.6E-20 | 2.7 |
| Oxo steroid | 6.2 | 2.5E-19 | 6.5 |
| 3-hydroxy steroid | 5.3 | 1.6E-15 | 6.1 |
| Oxysterol | 2.1 | 1.5E-14 | 26.1 |
| 3-oxo steroid | 4.3 | 3.7E-14 | 7.1 |
| 7alpha-hydroxy steroid | 2.5 | 2.1E-12 | 13.2 |
| Steroidal acyl-CoA | 1.9 | 3.0E-12 | 21.4 |
| 7-hydroxy steroid | 2.5 | 7.7E-12 | 12.1 |
| Androstanoid | 1.9 | 7.0E-10 | 14.4 |

Table 2.13: Enrichment results with the ChEBI ontology for the combined molecule set of *H. sapiens* and *M. musculus*, that do not intersect *S. cerevisiae* nor *E. coli*. Data shows a strong bias towards lipids and particularly steroids. The enrichment signal becomes much more powerful when combining the *H. sapiens* and *M. musculus* regions, providing a better overview of the particularities of a mammalian metabolome. Contrary to what happens for most other regions of the Venn diagram of Figure 2.29, these ChEBI categories show even higher enrichment than the high level ChEBI classes, which is reflected in the very low p-values.

shows the most enriched categories (these are not the most interesting selected one, but the actual most enriched ones) for the joint *M. musculus* and *H. sapiens* regions of small molecules that do not intersect with *S. cerevisiae* and *E. coli*.

To find relevant small molecules associated with the "unicellular" character of *S. cerevisiae* and *E. coli*, I inspected the union of *S. cerevisiae* and *E. coli* that does not intersect with the two *mammals*. Out of the 2,235 small molecules in these regions, 1,005 had ChEBI identifiers. These 1,005 small molecules mapped to 1,018 different ChEBI identifiers. I found no relevant enrichment beyond that already seen in *E. coli*. This indicates that there is no common area of metabolism that differentiates these two unicellular organisms from the *H. sapiens* and *M. musculus*, or if it exists, it is not adequately reflected in ChEBI or in

the metabolism databases. This again supports the initial idea that *S. cerevisiae* metabolism is more influenced by its *eukaryota* lineage rather than its unicellular nature.

For an eukaryotic metabolism comparison to *E. coli*, I retrieved all small molecules from the union of *H. sapiens*, *M. musculus*, and *S. cerevisiae* that did not intersect with *E. coli*. Out of 2,506 small molecules in this set, 1,704 had a ChEBI identifier, mapping to a total of 1,061 ChEBI entities. This set of molecules did not exhibit any stronger signals than the already observed with the *H. sapiens* and *M. musculus* union. Lipids again were highly enriched.

The Venn diagram of Figure 2.29 shows a core region of 750 small molecules shared by all organisms, which presumably represents the core primary metabolism. Inspection of this region through enrichment analysis shows that there is strong enrichment of phosphorilated entities, specially sugars and nucleotides. This region shows little lipid enrichment in contrast to the *eukaryota* regions, and the enriched lipids tend to be fatty acids and glycerophospholipids, main membrane components of all living organisms, which makes sense for a core metabolism.

### 2.5.6 Analysis of the second main mass peak

Earlier in section 2.3.5, Figure 2.6 showed that the masses of molecules from the four organisms exhibit two main peaks: a first larger peak around 200 Daltons and a second smaller peak around 800 Daltons. While this shape in the distribution might still be an artifact of our lack of knowledge of metabolism, it is reasonable to ask whether there might be a reason for metabolism to have this distribution in terms of mass.

Through initial manual inspections of the molecules's names in the 800 to 1200 Daltons range for the *H. sapiens* unification, I noticed that 120 out of 190 molecules had a co-enzyme A (CoA) structural unit. This is even more pronounced for *M. musculus*. Table 2.14 summarizes the result of this exercise for the four species. Acetyl-CoA plays a key role in metabolism, as it is one of the main hubs connecting catabolic and anabolic processes[112], connected to amino acids, lipids, and secondary metabolites. This probably explains the variety of acyl-CoAs found. The high level of enrichment in Fatty acyl-CoA compounds

| DataSet | CoA-bound | Total |
|---|---|---|
| *H. sapiens* | 120 | 190 |
| *M. musculus* | 121 | 158 |
| *S. cerevisiae* | 76 | 93 |
| *E. coli* | 112 | 180 |

Table 2.14: Count of small molecules in the 800 to 1,200 Daltons range, for each database species unification. The table shows the count for molecules bound to a co-enzyme A and the total count for the mass range, showing that in most cases nearly two thirds are CoA bound small molecules.

reflects the relevance of fatty acids as building blocks.

Enrichment analysis through BiNGO – which Table 2.15 shows for *H. sapiens* – confirms the finding of a variety of acyl-CoAs, and allows to characterize it in two main groups: fatty acyl-CoA and steroidal-CoAs. After acyl-CoAs, the analysis finds a second much smaller group of nucleotides – which Table 2.16 shows. The same holds for *M. musculus* and *S. cerevisiae*, the result of the enrichment is very similar. In the case of *E. coli*, the main categories enriched are again the same, with the exception that pyrroles appear with higher enrichment than nucleotides.

| ChEBI class | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| Acyl-CoA | 48.3 | 8.5E-85 | 29.2 |
| Thiocarboxylic ester | 48.3 | 1.3E-83 | 27.9 |
| Fatty acyl-CoA | 37.6 | 4.8E-68 | 33.7 |
| Ester | 69.1 | 4.4E-46 | 4.4 |
| Unsaturated fatty acyl-CoA | 13.4 | 7.2E-22 | 29.8 |
| Steroidal acyl-CoA | 7.4 | 5.5E-17 | 81.9 |
| Hydroxy fatty acyl-CoA | 8.1 | 3.2E-13 | 31.0 |
| Medium-chain fatty acyl-CoA | 6.0 | 5.2E-13 | 67.0 |
| Long-chain fatty acyl-CoA | 6.7 | 1.8E-12 | 42.7 |
| Oxo-fatty acyl-CoA | 6.7 | 1.8E-12 | 42.7 |
| Acyl-CoA(4-) | 7.4 | 6.4E-12 | 28.9 |
| Cholestanoyl-CoA | 4.0 | 3.9E-09 | 76.8 |
| 3-oxo-fatty acyl-CoA | 4.7 | 8.2E-09 | 42.3 |
| Organophosphate oxoanion | 12.1 | 1.3E-08 | 6.4 |
| Coenzyme | 6.7 | 3.2E-08 | 15.2 |
| Phosphoric acid derivative | 29.5 | 7.2E-08 | 2.5 |
| Short-chain fatty acyl-CoA | 4.7 | 7.7E-08 | 30.3 |
| Phosphorus oxoacid derivative | 29.5 | 9.4E-08 | 2.5 |
| Phosphorus oxoacids and derivatives | 29.5 | 1.8E-07 | 2.4 |
| 3-hydroxy fatty acyl-CoA | 4.7 | 2.7E-07 | 24.9 |

Table 2.15: Top 20 most enriched ChEBI ontology classes and roles for small molecules in *H. sapiens*, ranging from 700 Daltons to 1200 Daltons. Some general classes or roles have been removed.

| ChEBI class | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| Adenosine 3',5'-bisphosphate | 4.7 | 3.2E-07 | 24.1 |
| Adenosine bisphosphate | 4.7 | 7.7E-07 | 20.9 |
| Purine nucleoside bisphosphate | 4.7 | 1.5E-06 | 18.7 |
| Dinucleotide | 4.7 | 1.5E-06 | 18.7 |
| Ribonucleoside bisphosphate | 4.7 | 1.7E-06 | 18.4 |
| Nucleoside bisphosphate | 4.7 | 2.4E-06 | 17.4 |
| Purine ribonucleotide | 6.0 | 1.7E-05 | 8.5 |
| Adenosine phosphate | 4.7 | 2.2E-05 | 12.3 |
| Adenyl ribonucleotide | 4.7 | 2.2E-05 | 12.2 |
| Nucleotide | 11.4 | 2.6E-05 | 3.8 |
| Ribonucleotide | 6.7 | 2.8E-05 | 6.8 |
| Adenyl nucleotide | 4.7 | 5.5E-05 | 10.5 |
| Nucleoside phosphate | 11.4 | 8.0E-05 | 3.5 |
| Purine nucleotide | 6.0 | 1.3E-04 | 6.4 |
| Ribose phosphate | 6.7 | 3.1E-04 | 5.0 |
| Adenosines | 4.7 | 3.1E-04 | 7.8 |
| Purine ribonucleoside | 5.4 | 3.4E-04 | 6.4 |
| Ribonucleoside 5'-tetraphosphate | 2.0 | 3.8E-04 | 43.6 |

Table 2.16: Most enriched ChEBI ontology classes related to nucleotides, for small molecules in *H. sapiens* ranging from 700 Daltons to 1200 Daltons. Some general classes or roles have been removed. These classes are not the most enriched (Table 2.15 present those), but those with the higher enrichment that are related to nucleotides.

## 2.6 Conclusion

In this chapter, I introduced a framework and algorithm for consolidating metabolism knowledge in an organism specific way. I show cases in which conventional comparisons through InChI or SMILES would fail, which requires then an additional effort in the way we integrate chemical datasets of metabolism. A combination of structural data and meta data provides an improved way of consolidating these data sets. With this framework, I integrated three major metabolism databases for four model organisms. I integrated a number of other resources as cross references and showed a number of post processing steps that are necessary to improve the quality of a chemistry based unification of metabolism databases. The unification algorithm has better results than previously published methods.

The integration of *H. sapiens* metabolism databases shows that the overlap of the main metabolism providers (KEGG, BioCyc and BRENDA) is not as high as expected, sharing each database from half to two thirds of it contents with the other resources. This leaves still a large amount of data that is unique to each resource, which implies that there is plenty to gain from putting together the knowledge of different metabolism databases.

Assessment of the database integration results through two independent methods shows evidence that the error should remain below 20%. Inspection of particular regions of the data gives evidence that supports the high level of complementarity of the metabolism resources suggested by the integration. The unification of chemical data sets of metabolism is a complex problem, and I think that, even though our method performs better than other published algorithms, there is still reasonable room for improvement. I would personally expect to lower the number of unique chemical entities in the unifications through some enhancements, both in the unification method and in the data loading and annotation parts.

A comparison of the main metabolism databases for *H. sapiens* against the HMDB – which is considered in this text as a metabolome database and not a metabolism database – reveals a massive amount of chemical entities that are unique to the HMDB. Inspection of the metabolites unique to HMDB that are not found in the *H. sapiens* metabolism databases reveals that the majority are lipids, many of them isomers with only different saturations and double bonds

positioning. This can be explained partly because metabolism databases describe many lipid related pathways as generic reactions. For instance, the first step of the cycle in the fatty acid $\beta$-oxidation[1] is a Medium-chain acyl-CoA dehydrogenase, which converts "a 2,3,4-saturated fatty acyl-CoA" into "a trans-2-enoyl-CoA". In "a 2,3,4-saturated fatty acyl-CoA", the fatty acid moiety normally has 14 to 24 carbons, and it will not contain double bonds in positions 2, 3, and 4. This means that metabolism databases only include generic molecules in this case, which represent a diversity of fatty acids, but do not contain all the fatty acid molecules which match the generic description. In general, most metabolism databases have the bias of including only chemical entities that are participating in known reactions, while HMDB only requires evidence that the molecule can be found in *H. sapiens*. It is also important to note that metabolism databases add more than a thousand small molecules that are not present in the HMDB.

Inspection of the four model organisms metabolomes assembled shows particularities and commonalities – like the concentration of acyl-CoAs in the 800 to 1,200 Daltons – between the chemistry of these organisms. Mass distributions show that higher eukaryotes, like *H. sapiens* and *M. musculus*, have slightly more complex chemical entities than simpler organisms, as masses tend to shift towards higher values on average. Enrichment analysis through the ChEBI ontology further shows that this complexity in *H. sapiens* and *M. musculus* grows mostly towards the lipids space. On the other hand, enrichment analysis for *E. coli* suggest it might specialize more in carbohydrate variants than in lipid complexities. These analyses are of course limited by the coverage that the ChEBI ontology achieves of the chemistry of the different organisms. It is clear as well that we need a wider and better ChEBI, and that its ontology is by far one of the most useful resources for classifying chemical entities and giving them a biological interpretation.

[1] Details for this pathway correspond to those found in HumanCyc

# Chapter 3

# Text mining methods for inferring metabolomes

## 3.1 Introduction

Although online databases, such as KEGG [72], BRENDA [16], Rhea [3] or BioCyc [11] among many other resources represent metabolism to a wide extent, there is still a lot of knowledge deeply buried in millions of scientific publications. Approximately 20% of biological knowledge is normally available in structured data repositories like databases, the remaining part is hidden in free text[1] scientific literature [62].

There are currently 20 million papers deposited in NCBI PubMed[2]. KEGG, BioCyc, and BRENDA for *H. sapiens* point to only 57.5 thousand of these. In contrast, small molecules dictionaries find hits in approximately 7 million abstracts. Considering that *H. sapiens* is by far the most studied organisms (more that 1.4 million abstracts mention it, more than doubling the next mentioned species[3]), one could expect an additional reasonable amount of knowledge of metabolism and small molecules to be uncovered from literature.

[1] Refers to unstructured text, written normally in a narrative manner, as opposed to structured text, as deposited organized in a database.

[2] NCBI PubMed is the main repository of life sciences literature in the world

[3] I obtained the number of mentions per organisms from the text mining infrastructure described later.

The HMDB database contains nearly 8,000 small molecules. Most metabolism databases contain less than two thousand small molecules that can be assigned through them to *H. sapiens*. This leaves at the least some thousands of small molecules that cannot be retrieved from metabolism databases but need to be mined from the literature or text books. Furthermore, the rate in which new literature is generated makes it difficult for these databases, let alone researchers, to cope with the vast amount of new relationships generated between proteins, small molecules, and the cell types or tissues in which they reside. Sometimes the level of knowledge about a particular small molecule is not sufficient to include it in metabolism databases. This normally happens when the reactions in which they participate are not well understood.

Most metabolism databases do not provide further biological context to small molecules than the reactions in which they participate and the enzymes that could catalyze those reactions. There is rarely any reference in them to cell types or tissues in which these small molecules can be found. This is of extreme importance for metabolomics, as the metabolome is different from tissue to tissue or from cell type to cell type[158]. These localization relations can be addressed through a text mining approach, in which dictionaries of tissues and cell types can be used alongside small molecules dictionaries to find relations in free text.

The need to complement what is found in existing metabolism databases with small molecules mentioned in literature motivates the use of text mining tools to fill this gap. Although the use of text mining tools is well established in bioinformatics, at the time this work started there were surprisingly few examples of application to metabolome assembly. In this chapter we introduce the development of tools that exploit text mining techniques for the proposed aim of inferring complete metabolomes.

## 3.2 Background

### 3.2.1 Text mining terms

Before going into the subject, this section presents a short overview of relevant terms that readers not familiar with text mining might find useful.

**Grammar:** A grammar is a set of rules that describe how certain components can form more complicated structures within a natural language, explaining how strings – as words in natural language – can be combined to form higher order structures – such as phrases or sentences – in that language. Having a grammar allows to have software that can decide whether a particular arrangement of strings is correct or not in that language.

**Parser:** A program or tool that scans a chunk of text for certain defined grammatical structures, normally retrieving them for further processing by other components. These structures are normally defined in a separate grammar, which tends to be simpler than the original text rules, like the encoding of simple ways of writing a particular interaction or rules governing how to write a IUPAC name.

**Shallow parser:** A tool that labels parts of a sentence with simple syntactic roles, such as verb phrase and noun phrase, without descending into additional detail. Figure 3.1 shows an example, generated with the Brat tag visualizer [142].

**Deep parser:** A tool that labels parts of a sentence, characterizing completely the syntactic roles of each word in the sentence, normally producing a syntactic tree. Figure 3.2 shows an example syntactic tree.

**Tagger:** A tool that processes text to find particular entities (normally words or multi-word tokens) according to a large terminology set (as a dictionary), a set of rules (such as a grammar), or an engine that allows to identify entities. The output is normally the same text given as input with the recognized entities "tagged" with the class to which they belong. Figure 3.3 shows an example, from the web version of the GENNIA Tagger, mentioned in Section 3.2.2.

**Part-of-speech tagging:** Assignment of a grammatical role (noun, verb, adverb, etc.) to each word of a sentence. Normally a deep parser requires to do this before generating a syntactic tree.

**Named entity recognition:** abbreviated to NER, is the act of assigning a recognized token to a class of elements: the word "ligase" found in a free text piece refers to a class of objects called "proteins" or even "enzymes".

**Normalization:** In this context, normalization means assigning a database identifier to a recognized entity name within a sentence. For instance, if a tagger recognizes a protein name, the normalization step would associate that name with an identifier from a database like UniProt.

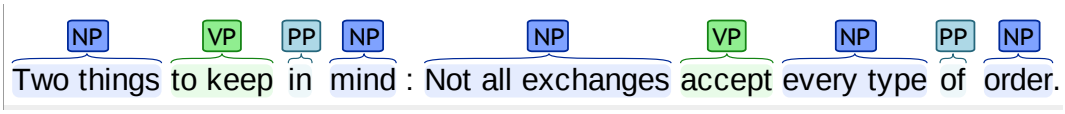



Figure 3.1: Example of a sentence processed by a shallow parser, where only the main syntactic structures are recognized: noun phrases (NP), preposition phrases (PP), and verb phrases (VP). This was produced with Brat [142]

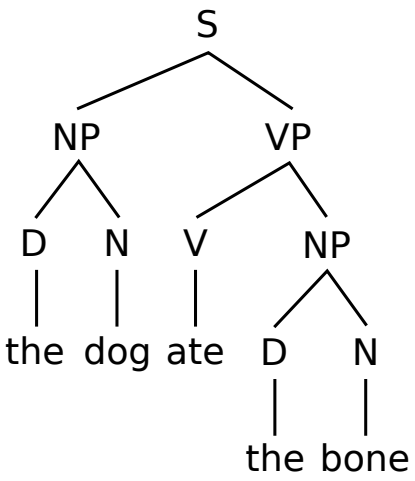



Figure 3.2: Syntax tree, showing the level of characterization that a deep parser would normally give to the sentence "the dog ate the bone", image from Wikipedia. Entities shown are sentences (S), noun phrases (NP), verb phrases (VP), articles (D), nouns (N), and verbs (V).

Cancer cells exhibit accelerated rates of metabolism favoring glucose over fatty acid (FA) utilization. For both energy substrates, protein-mediated transport plays an essential role in facilitating glucose or FA movement across plasma membrane into the cells . Scarce data exist regarding the expression of glucose and/or FA transporter in cancer tissue . Therefore , we examined glucose (GLUT-1 , GLUT-3 , GLUT-4) and FA (FAT/CD36, FABPpm, FATP-1) transporter expressions at the protein and post-transcript (mRNA) levels in 35 endometrial carcinomas (G1, type endometrioid , FIGO I) and compared them with normal endometrial mucosa (n=10) . Endometrial cancer tissue had significantly greater protein expression of GLUT-1, GLUT-3, and GLUT-4 and, conversely, lower fatty acid (FAT/CD36 and FATP-1) transporter expression. Interestingly, mRNA content closely mirrors the changes, but only for glucose transporters and not fatty acid transporters. These results suggest the presence of metabolic switch of energy utilization in endometrial cancers favoring glucose consumption as the major source of energy.

(entity types: protein, DNA, RNA, cell_line, cell_type)

Figure 3.3: Tagged text output visualization, generated by the GENNIA Tagger. Entities are only recognized as being a cell type, disease or protein, but there is no normalization (assignment of the identified entity to a database entry).

### 3.2.2 Existing resources: text mining metabolome knowledge

#### PolySearch

This tool [18] provides a web search for relating diseases, tissues, cell compartments, gene/protein names, single nucleotide polymorphisms, mutations, drugs and metabolites. It relies both on text mining algorithms, to process the whole of NCBI PubMed abstracts, and on existing databases, like NCBI Entrez or DrugBank[83], to improve and annotate results. The tool ranks results by statistical scores that evaluate the relevance of the relations. The user can impose cut-offs for the minimal number of citations where the relations should be found.

PolySearch has an interesting set of thumb rules to limit the extent to which two terms – like a protein and a disease – can be related. It allows the user to add specific verbs – such as "accumulates" or "depleted" – that need to be part of the sentence relating the two terms.

Although it relies on text mining to process all the abstracts in NCBI PubMed, the size of the small molecules collections is only limited to the Human Metabolome Database (HMDB). PolySearch is only web accessible, so it cannot be easily incorporated as a part of a local pipeline. Despite being an excellent tool due to the integration of many different biological entities, it fails to fill the gap addressed here, since it only considers a small set of ~3,000 small molecules obtained from

the HMDB. This also makes it highly biased towards *H. sapiens*, impairing its use with different organisms.

### OSCAR and OPSIN

OSCAR – Open Source Chemistry Analysis Routines[68] – is a Java library for text mining of chemical entities. Development of OSCAR started in the early 2000's and has been mainly aimed at retrieving entities from experimental sections of Chemistry journals and patents, specially chemical synthesis procedures. OSCAR works essentially as a shallow parser of chemical names. It combines dictionary based search with regular grammar based tokenization of chemical names.

This latter functionality, as well as its ability to produce chemical structures from the identified names (when written in IUPAC format), is provided by the package OPSIN [93]. OPSIN is the best open sources alternative for identifying chemical names. One of its main drawbacks when it comes to inferring metabolomes is that "... OPSIN does not yet support biochemical nomenclature, e.g., carbohydrate nomenclature, and hence will have very low recall when presented with such names." as the authors state [93].

This should not be a surprise given that OSCAR and OPSIN development has been historically biased towards the area of experimental chemistry rather than biochemistry. This however could be improved, as OPSIN's rules are extensible without any need to modify its code. OSCAR and OPSIN are available to be used locally, allowing high-throughput analysis.

When the work for this part of the thesis began, OSCAR 3 was available. This package was extremely difficult to integrate due to the need of complicated set ups and an overall unintuitive API. Fortunately much of this was solved in release 4 [68], but this came too late to be integrated into our workflow. The need for extending OPSIN for general metabolite names also made this tool less attractive for our endeavour. For these reasons I decided not to use OSCAR nor OPSIN.

### WhatIzIt Text Taggers

The WhatIzIt text taggers, available from the text mining group at the EBI, provide a number of different normalizer taggers for Small molecules, proteins, species

and other biological entities [125]. Most of these taggers rely on dictionaries, or large terminology sets, which contain various synonyms and regular expressions that recognize in free text terms that belong to a database or ontology.

WhatIzIt offers four dictionaries for small molecules: ChEBI, PubChem Compounds, HMDB and OSCAR. WhatIzIt uses UniProt as dictionary of protein names, and NCBI Taxonomy as dictionary of organisms and taxonomic classes. The text mining group at EBI has tools and expertise to aid in the generation of new dictionaries, based on databases or ontologies.

Handling large collections of regular expressions and plain text for concurrent search is a demanding task in terms of memory and processing. In the case of WhatIzIt, the underlying technology for handling large collections is provided by the monq.jfa Java library [82]. This library is a flexible and efficient framework for deployment of taggers as servers in a distributed environment (like a computer cluster). The ability to distribute text taggers and be able to use these tools locally at the EBI cluster are essential to process a corpus like the whole NCBI PubMed abstracts.

### EBIMed

The text mining group at EBI provides EBIMed [126], a web application that relies on co-occurrences of proteins, organisms, entries of the Gene Ontology and drugs. EBIMed normalizes occurrences of these biological objects in the NCBI PubMed abstracts through WhatIzIt text taggers, using UniProt for proteins, the three Gene Ontology branches (cellular component, biological process and molecular function), MedlinePlus for drugs and NCBI Taxonomy for organisms. Only sentence based co-occurrences are retrieved, ranking the pairs of co-occurring entities by the number of sentences that contain them.

The tool provides the user with a table of results, where the interacting biological entities are shown with a score for the interaction and with a link to the abstracts and sentences where the co-occurrence was found.

The aim of the tool is mainly oriented for biomedical queries, and is a really useful at relating proteins, organisms, functions and diseases. Unfortunately, it does not include small molecules.

### GENNIA Tagger

The GENNIA Tagger [155] provides part-of-speech tagging, shallow parsing and named entity recognition capabilities. The tagger is a maximum entropy model[1], trained with a corpus containing biomedical journals and newspaper articles.

The authors evaluated the tagger using the GENNIA Annotated corpus of biomedical abstract[81], as well as the PennBioIE Corpus[98], with very high accuracies (above 90%). The tagger recognizes protein, DNA, RNA, Cell line and Cell type entities, with an average F-Score of 71%. However, it does not tag small molecule names and tissues.

### ChemList Chemical Dictionary

ChemList[55] is a text mining oriented chemical dictionary built by integrating small molecule and drug data from ChEBI, UMLS (Unified Medical Language System [8]), NCBI MeSH, DrugBank, KEGG, HMDB and ChemIDPlus. The main purpose of it is to be used as a NER and normalization reference dictionary.

The post processing of the data from the text mining point of view seems impeccable: they apply seven different filtering rules to the dictionary and a final manual check of the highly occurring terms. Finally, when used to tag elements in a corpus, through their tool Peregrin, they apply disambiguation of chemical names.

The chemical unification was mainly done relying on InChI strings and CAS numbers. As I commented in Chapter 2, there are several difficulties in unifying chemical data sets, and using InChI strings does not solve them all. Furthermore, due to the publication date of this paper, it is very likely that the authors used plain InChI and not Standard InChI, which exposes this collection to an additional number of duplicates and artificial merges, as there is no guarantee that all the source resources used the same parameters for their InChI generations. This could have been partially alleviated if the authors would have calculated the InChI strings themselves, but they essentially collected them from the source

[1]Very often in the field of Natural Language Processing, models based on multinomial logistic regression are referred to as maximum entropy models. A multinomial logistic regression is a generalization of a logistic regression, which has a binary outcome, to get multiple outcomes.

databases. Assessing uniqueness of small molecules by CAS numbers leads to duplicate as well, in this case they are aware of this issue.

ChemList is available for download and can be used as a dictionary for NER and normalization of chemical names. However, I decided not to use it as a chemical dictionary given the consolidation process used. I could manually verify a number of undesired duplicates and class including compounds that should be deemed different. The text mining processing techniques that the authors of this work propose should be replicated when building chemical dictionaries.

### NaCTEM Tools

The National Centre for Text Mining, at the University of Manchester, provides a series of web based tools for shallow and deep parsing.

**TerMine:** is an automatic term recognition tool [44]. Parses a corpus to find relevant technical multi word terms through part-of-speech, linguistic filtering (mostly nouns, adverbs and some prepositions are accepted) and statistical analysis of occurrences. Given a corpus, it generates a list of possible relevant terms (multi word) found in the corpus, ranked by their relevance. This is useful for generating an ontology out of a corpus. For our application, this tool could be helpful to pre-filter abstracts and obtain terms, which later filtered by a chemical name recognizer/dictionary would yield small molecule names mentioned in the corpus. However, through a SOAP web service access is difficult to process the amount of data required.

**MEDIE:** is semantic search service [111] which uses a precomputed deep parse of NCBI PubMed. Enables search patterns in which the user must define the subject, verb and/or object of the query, retrieving abstracts that have that pattern. For instance, searching for subject "molecule", verb "inhibits" and subject "enzyme", the system finds results like "Omapatrilat is a single molecule that simultaneously inhibits neutral endopeptidase and angiotensin-converting enzyme", where the different biological terms are tagged. In order to resolve ambiguous terms like "molecule" or "enzyme", the processing of NCBI PubMed includes tagging with the GENNIA ontology[81]. MEDIE is developed by NaCTEM collaborators at the

Natural Language Processing Laboratory at the University of Tokyo. Unfortunately, this tool can only be used through a web interface.

**FACTA+:** is a tool to retrieve direct and indirect association facts between terms in the biomedical literature [156]. It detects biomolecular events between compound, protein/gene, disease, symptom, drug or enzyme through a machine learning approach, which is trained using a corpus with annotated events. Statistics between the associated terms are computed through a pre-calculated co-occurrence count between all possible terms in each of the listed classes, done through a dictionary (one dictionary per class) search. In the query the user defines a query term, a pivot concept and a target concept. A query term would normally be a particular compound, protein/gene, disease, symptom, drug or enzyme, pivot and target concept would be one of these classes of entity. FACTA+ is only available through a web interface, and does not include support for tissues or cell types.

Unfortunately, NaCTeM does not provide these tools as executables, which makes the analysis of huge amounts of text impossible.

During the final stages of this work, authors from NaCTEM published a very interesting study on extracting *S. cerevisiae* metabolites from literature [109], based mainly on name entity recognition and then machine learning to rule out non-metabolites, using a corpus of selected publications for *S. cerevisiae*. While this is a sensible way of avoiding the organism disambiguation problem, it is only applicable for those organisms where such collections of articles can be easily retrieved.

### U-Compare

U-Compare [73] is a workflow environment for Natural Language Processing tasks. It has a number of components to read corpora, break text into sentences, do shallow and deep parsing, named entity recognition and disambiguation. Although it could be used as a general purpose text mining tool, the main aim of it is the comparison of text mining workflows, and so most of its visualizers are tailored for comparison purposes (of running a corpus through one pipeline or the other).

## 3.3 Procedure followed

As much as I felt inclined to learn more and more about natural language processing tools, it was more clear that implementing a solution using syntactic parsing could be too complicated for a first approach to the problem. The problem has various types of relationships (organisms to proteins, organisms to tissues, organisms to small molecules, tissues to proteins, tissues to small molecules and protein to small molecules) and each one of them can be represented by a number of patterns in natural language. On top of this I aimed at tools that could process all the NCBI PubMed abstracts in a short time, and many of these tools did not have the capacity or were not available in a format that I could use for this purpose.

I decided to take a much simpler initial approach based on named entity recognition and normalization, that would enable to find all the different types of relationships in free text, with less details, but generating something that could be processable for the amount of data that I wanted to inspect. Other works in the future can dig deeper into particular types of relations and analyze them from a semantic point of view. My focus is on the generation of organism-specific small molecules collections that consider relations to biological containments, for aiding experimental metabolomics in the annotation of features.

Using the tools available from the text mining group at EBI, I built a pipeline of dictionary based taggers to process the entire NCBI PubMed abstracts collection. The pipeline includes the following tagger servers, used in this order:

**Sentencizer:** Recognizes text chunks within the abstract and title sections of the NCBI PubMed XML export file and breaks it into sentences.

**UniProt Tagger:** Recognizes protein name and synonyms through a dictionary based on UniProt. Contains ~230,000 entries pointing to ~130,000 different proteins. Out of these, ~50,000 point to more than one organism. The protein name tagger has higher priority than small molecule taggers to avoid recognizing parts of enzyme names as small molecules.

**BRENDA Tissue Ontology Tagger:** Tagging server that recognizes tissues and cell types through a dictionary based on the BRENDA Tissue Ontology.

This dictionary was not available from the WhatIzIt repository, and was built from the BRENDA Tissue Ontology[1]. Contains ~5,600 entries pointing to ~3,300 different tissues and cell types.

**ChEBI Tagger:** Tagging server that recognizes small molecules through a dictionary based on the ChEBI ontology. The dictionary contains approximately ~43,000 entries, pointing to ~11,000 different ChEBI entities.

**PubChem Compounds Tagger:** Tagging server that recognizes small molecules through a dictionary based on PubChem Compounds. The dictionary contains ~195,000 entries, pointing to ~104,000 different PubChem Compounds compounds.

**NCBI Taxonomy Tagger:** Tagging server that recognizes organisms and taxonomic names, through a dictionary based on the NCBI Taxonomy database. The dictionary contains ~208,000 entries, pointing to ~170,000 different NCBI Taxonomy identifiers.

Unless stated, all dictionaries and tagging servers are part of the WhatIzIt collection, built and validated by the text mining group at EBI and a wide range of external users.

The pipeline of tagging servers takes as input NCBI PubMed XML export files, and outputs the same NCBI PubMed XML with annotations of proteins, small molecules, tissues/cell types and organisms, normalized to the mentioned databases. A second program written in Java parses this output and stores all occurrences of these biological entities in a relational database that I designed for this purpose.

The pipeline runs on the EBI cluster, where I deploy each tagger server on a different node. The client program feeds the NCBI PubMed XML files to the pipeline, and then parses the output to load it into the database. The program runs in a separate cluster node, and parallelizes in six threads the feeding and recollection process. Figure 3.4 illustrates the complete process. Figure 3.5 shows

[1]The dictionary was built with partial assistance from Ms. Kalaivani Jayaseelan, who by that time was an intern student. All the work was carried out under my supervision and guidance.

the schema of the database designed to hold the co-occurrences results. All communication between client and different tagging servers within the cluster is through TCP sockets.

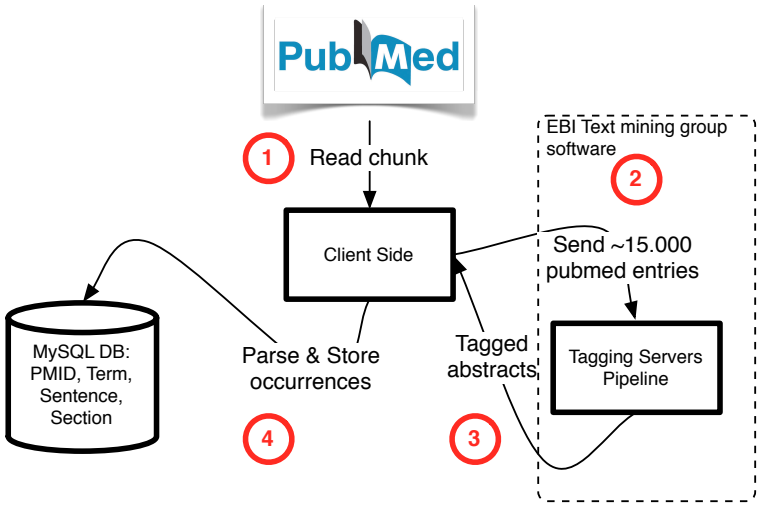



Figure 3.4: Diagram of the text mining pipeline built to retrieve small molecules and their co-occurrences with organisms, tissues/cell types and proteins. The pipeline starts with the client reading a chunk of entries from a directory containing NCBI PubMed XML export files which sends to the tagging servers (for proteins, small molecules, tissues/cell types and organisms). After going through all the tagging servers, the submitted text is tagged with these biological entities. The client application then parses these tags and stores them in the MySQL database designed. Information stored includes the database and identifier of the biological entity, the NCBI PubMed ID of the document where it was found, the section and the sentence number.

The pipeline processed all NCBI PubMed abstracts and titles until September 2009. The whole process took seven days, if it could run without interruptions. The main bottleneck of the process is the storage of data in the database, and interruptions of the whole process by communication errors or cluster issues, requiring the process to be restarted a few times, before a first complete and successful run. As the pipeline analyses six or more sets of NCBI PubMed abstracts at once, but only one process can handle the insertions (otherwise there are collisions when writing to the database), this produces the database bottleneck.

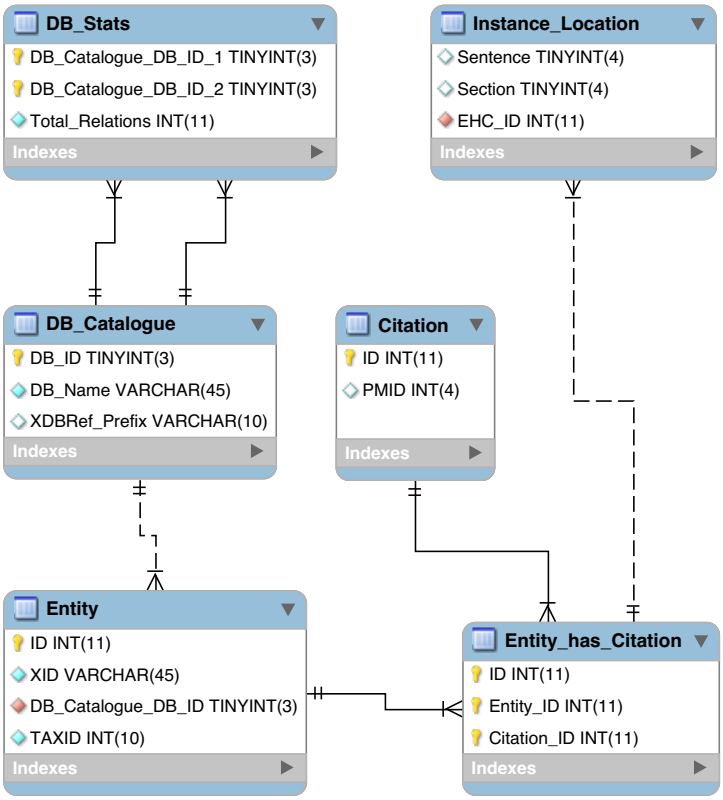



Figure 3.5: Schema for the database that stores co-occurrences between small molecules, organisms, tissues, cell types and proteins. The table Entity stores the biological entities mentioned, normalized by the corresponding databases (ChEBI and PubChem Compounds, NCBI Taxonomy, BRENDA Tissue Ontology, and UniProt respectively). The table Citation stores the identifier for a NCBI PubMed document. The table Entity_has_Citation stores a relation meaning that the entity referenced was mentioned in a particular citation and the table Instance_Location stores where that occurrence took place (Sentence number and section). A co-occurrence between any two biological entities in the database is computed by joining the Entity_has_Citation table with itself. Table DB_Catalogue holds the databases to which the different entities belong, and DB_Stats holds statistics of interactions between these types of databases, necessary for calculating significances of co-occurrences.

Several strategies – disabling binary logs, disabling keys, dropping constraints, partitioning, etc. – and tunes were necessary to achieve the seven days running time.

## 3.4 Co-occurrences results

I inspected the overall co-occurrences between the different biological entities that the process considered. Figure 3.6 shows the overall relations between the dictionaries used.

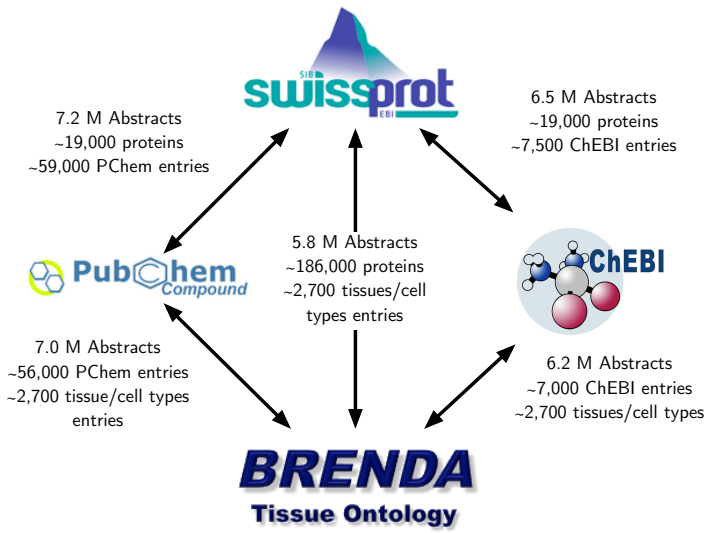



Figure 3.6: Diagram of the number of entities of each type related through co-occurrences for the different dictionaries. The diagram shows that approximately 7.2 million abstracts related 19,000 proteins to 59,000 small molecules in PubChem Compounds. The other relations (arrows) are read likewise.

Figure 3.7 shows an important aspect, the distribution of entity occurrences for the different dictionaries in the whole corpus annotated. The distribution of occurrences varies from dictionary to dictionary for different reasons.

The NCBI Taxonomy dictionary has a median number of occurrences much lower than the rest of the dictionaries. This is partly produced because NCBI

Taxonomy does not only have organism names, but all the range of taxonomic names available from super kingdom – like *bacteria*, *archaea* or *eukaryota* – to sub strains – like *Escherichia coli str. K-12 substr. MG1655*; sub strain is two levels lower than species. Most of these taxonomic ranges will be seldom mentioned and represent a big part of the NCBI Taxonomy tree, lowering the average. Also for NCBI Taxonomy, the supposed outliers with around a 100 thousands to a million occurrences, which in most dictionaries one would blame to a promiscuous synonym that should be removed, represent *H. sapiens*, *M. musculus*, *E. coli* and *bacteria* among few others. None of them had promiscuous synonyms when checked in the dictionary.

The ChEBI dictionary shows the highest median of co-occurrences per term among the chemical names dictionaries: slightly below the order of hundreds of co-occurrences, compared to the nearly one order of magnitude lower of PubChem Compounds or the HMDB. This is probably due to the higher number and quality of synonyms in ChEBI. The ChEBI dictionary has on average four synonyms per chemical entity.

PubChem Compounds is in general recognized to have bad quality and low number of synonyms [55], the box plot in Figure 3.7 reflects this, as entities are found only few times (lack of synonyms), has a very skewed distribution and a very thin and long tail, which is probably symptom of many synonyms being highly unspecific. This needs to be taken into consideration when later using the results, as high occurring terms might need to be ignored.

The HMDB has very good synonyms, however a high number of compounds as we know are lipids, which have historically been represented with different nomenclatures[1][34]. This makes their identification in free text difficult, lowering the overall occurrences average. The region above the upper whisker of HMDB shows a lower density of outliers compared to PubChem Compounds. Although this could well be due to the higher number of entries in the PubChem Compounds dictionary, knowing first hand the content of both data sets, and considering the number of times I have inspected results from both of them, I

[1]The last IUPAC-IUBMB recommendation for Lipid nomenclature came in 1976, since then a number of novel lipid classes have been elucidated with no proper systematic nomenclature covering them.

would dare to speculate that this is an unspecific synonyms artifact. In contrast to small molecule name dictionaries, the tissue and cell types dictionary built from BRENDA Tissue Ontology seems well distributed and with a high median of co-occurrences. This is probably due to the fact that tissues and cell types, differently to small molecules or proteins, are a much more constrained vocabulary and is based mostly on Latin roots that have been accepted for very long now in the life sciences community (so most authors refer to them in a relatively well controlled way). On top of this, it is rare to have abbreviations for tissues and cell types; abbreviations sometimes lead to unspecific synonyms.

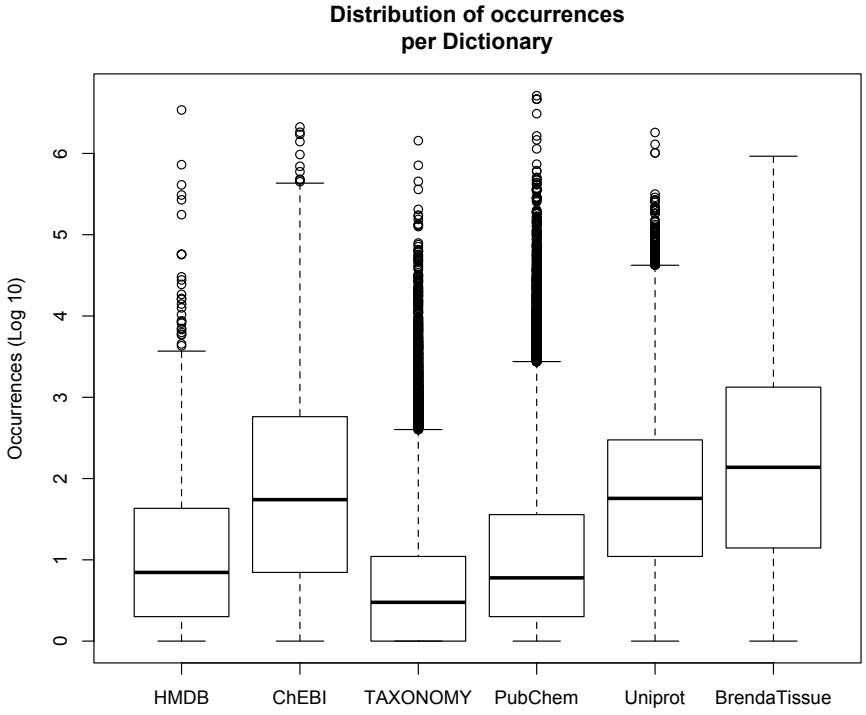



Figure 3.7: Box plots of distributions of occurrences of the terms in each dictionary on the NCBI PubMed abstracts until September 2009. Boxes in the box plot represent the observations in between the first and third quartile of the distribution, the whiskers represent 1.5 parts of the interquartile range below and above those quartiles respectively. Middle line of the box represents the median.

### 3.4.1 Entity ambiguity

The normalization process in the case of protein names will tag a name recognized as protein name with all UniProt identifiers that match that name, regardless of the species. One of the problems of normalizing results of named entity recognition is the assignment of the entity to a particular organism [162]. Although studies [159; 162] have produced results with good precision and recalls for deciding to which organism a particular entity (protein, gene, etc.) belongs, they have essentially relied on very good gold standard collections to train their methods. This was only done for a handful of model organisms.

For *H. sapiens* a high proportion of citations mention it as the only species, which simplifies a bit the disambiguation problem for proteins, tissues and cell types. This does not hold for the other model organisms presented in Figure 3.8, only half of the time they tend to be the only organism mentioned in the paper.

For this work, given that I only collect co-occurrences and not entire sentences, considering that I am mainly interested initially in a *H. sapiens* metabolome to compare against our gold standard, the programs written disambiguate proteins by the organisms mentioned in the citation. In the case of *H. sapiens* this is relatively safe as most of the abstracts tend to mention it as the only species (~70%). For the remaining ~30% of the citations that mention *H. sapiens* with other species, at least more than one third are always mammals.

### 3.4.2 Significance in text mining relations

I investigated statistics to rank the strength of co-occurrences of biological terms (proteins, small molecules, tissues, organisms). Works by [33; 119; 133] provide excellent coverage of the general statistic of co-occurrences, and explanations in this section reflect my understanding of their treatment.

Most of the statistical theory behind the significance of co-occurrences relies on a contingency table approach to the problem. What we essentially want to know is, given two terms $T_1 = m$ and $T_2 = n$, that belong to dictionaries $D_m$ and $D_n$, whether the amount of evidence in the corpus correlating the terms can

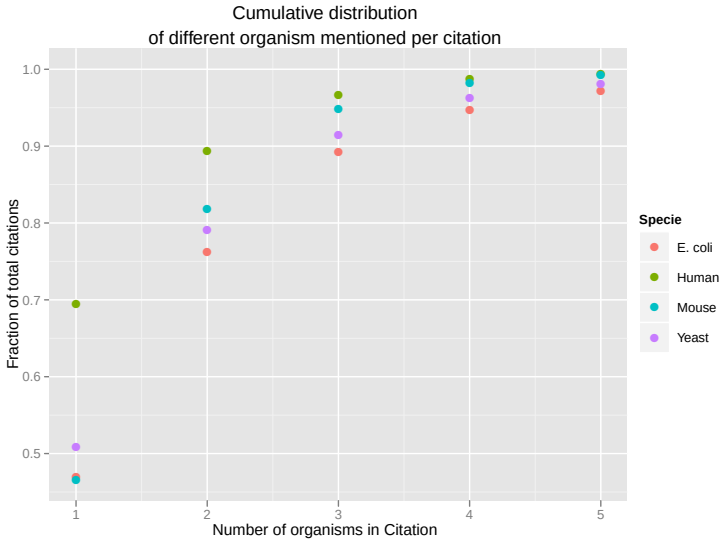



Figure 3.8: Cumulative distribution of the number of different organisms mentioned per document, shown for different organisms. For *H. sapiens*, ~70% of the abstracts mention a single organism, and ~90% mention one or two organisms (one of these *H. sapiens*).

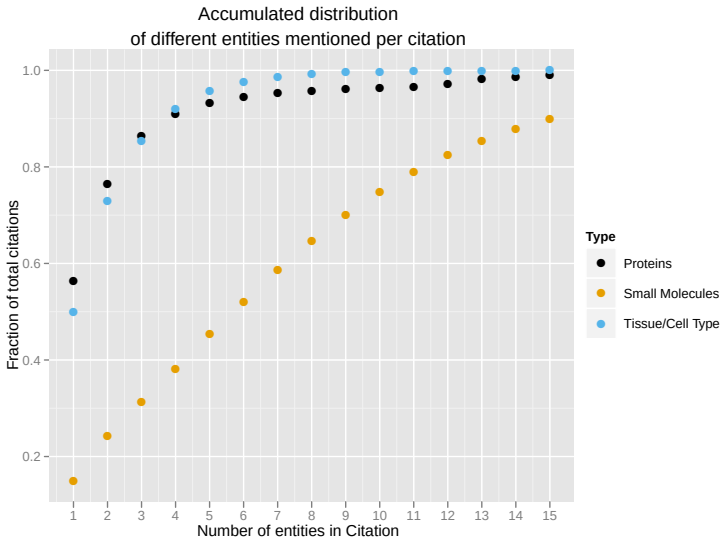



Figure 3.9: Cumulative distribution of the number of different tissues/cell types, proteins and small molecules mentioned per abstract. 85% of the abstracts mention four or less tissues/cell types, only ~50% mention just one tissue/cell types. Proteins (organism disambiguated) show a similar profile. The slower growth of the small molecule distribution is probably due to their lower frequency of occurrence, combined with the fact that chemical names are more difficult to detect compared to other biochemical entities.

| | $T_1 = m$ | $T_1 \neq m$ | |
|---|---|---|---|
| $T_2 = n$ | $O_{11}$ | $O_{12}$ | $= R_1$ |
| $T_2 \neq n$ | $O_{21}$ | $O_{22}$ | $= R_2$ |
| | $= C_1$ | $= C_2$ | $N$ |

Table 3.1: Contingency table of observed frequencies of co-occurrences. $R_1$ and $R_2$ denote row totals, $C_1$ and $C_2$ column totals. $O_{11}$ is the observed number of co occurrences of terms $m$ and $n$, $O_{12}$ the number of co-occurrences between term $n$ and any term from the dictionary of term $m$, but not $m$.

be explained by chance. If this is not the case, we suspect there is a meaningful relationship between $m$ and $n$.

Contingency Table 3.1 summarizes the possible outcomes of observing the co-occurrences of term $m$ with the elements of $D_n$, splitting the outcome in co-occurrence being with either $n$ ($T_2 = n$) or with another element from $D_n$ which is not $n$ ($T_2 \neq n$). The same applies for $n$ co-occurring with elements of $D_m$. $O_{11}$ in Table 3.1, corresponds to the observed number of documents, in this case NCBI PubMed abstracts and titles, where $m$ and $n$ co-occur. Conversely, $O_{12}$ corresponds to the observed number of documents where $n$ co-occurs with any other member of $D_m$, but $m$. Since the database stores all the occurrences of terms, we can obtain all $O_{ij}$ values for every pair of terms between any two dictionaries. Given that the computation of the $O_{22}$ is expensive as requires counting all co-occurrences of all $D_m$ members with all $D_n$ members, the software written computes this once and stores the value of this interactions in the database[1]. $R_1$, $R_2$, $C_1$, $C_2$ and $N$ are the sum of rows, columns and total of the table respectively.

For a particular co-occurrence of terms $m$ and $n$ then, the software can compute a contingency table of observed co-occurrences. Using this data it builds a second contingency table that reflects expected levels of co-occurrences of $m$ and $n$. Table 3.2 shows how expected co-occurrences $E_{ij}$ depend on the rows and columns sums, $R$ and $C$, from Table 3.1.

The significance of the co-occurrence of $m$ and $n$, as in most statistical tests, depends on how much the observed values deviate from the expected values.

[1]The exact value stored is $O_{22} + O_{11} + O_{12} + O_{21}$, and then each time the software looks for a particular co-occurrence, it subtracts the particular $O_{11} + O_{12} + O_{21}$ values, yielding $O_{22}$.

| | $T_1 = m$ | $T_1 \neq m$ |
|---|---|---|
| $T_2 = n$ | $E_{11} = \frac{R_1 C_1}{N}$ | $E_{12} = \frac{R_1 C_2}{N}$ |
| $T_2 \neq n$ | $E_{21} = \frac{R_2 C_1}{N}$ | $E_{21} = \frac{R_2 C_2}{N}$ |

Table 3.2: Contingency table of expected frequencies of co-occurrences, calculated from the values in the observed contingency table. $E_{11}$ stands for the expected number of co-occurrences between terms $m$ and $n$, given the observed totals in Table 3.1.

Different metrics of this significance make distinct uses of these expected and observed values, through compromises that give them their weaknesses and strengths.

The statistics picked were the Mutual Information Measure, Log likelihood score and the t-Score, as according to [33], these seem to be the most useful ones, in that order.

For the Mutual Information Measure, derived from Shannon's information content, Equation 3.1 shows how it is defined in terms of the contingency tables. This is the most popular statistic, and only suffers from giving high rankings when the number of observations is low ($< 3$). This is an issue in the noisy case of text mining, but can be a strength if we are after rare cases. Normally the approach is to use a cutoff for the minimal number of documents that show the co-occurrence, or have a second statistic and impose thresholds for both.

$$MiM = log \frac{O_{11}}{E_{11}} \tag{3.1}$$

Equation 3.2 defines the t-Score, named after the resemblance to a t-Test[1], in accordance to the contingency tables.

$$t\text{-}Score = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}} \tag{3.2}$$

Finally, Equation 3.3 defines the Log Likelihood (LLH). This statistic diminishes the effect of small $O_{11}$ by comparing through the complete table and multiplying by all observations over expectations.

$$LLH = 2 \sum_{ij} O_{ij} log \frac{O_{ij}}{E_{ij}} \tag{3.3}$$

[1]Even though this kind of test would not apply to these tables

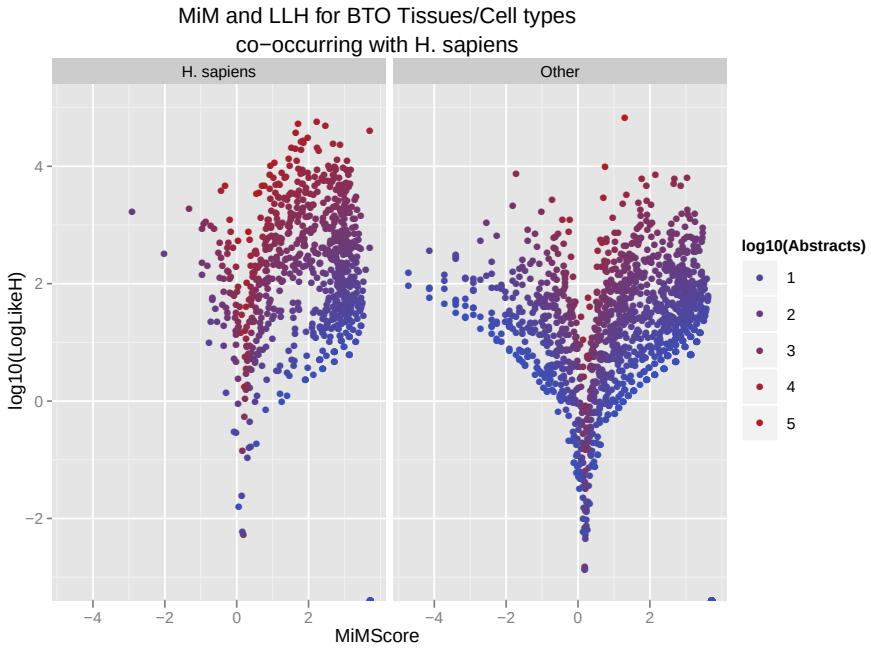



Figure 3.10: Log likelihood and mutual information measure scores for the co-occurrences of *H. sapiens* and tissues/cell types from the BRENDA Tissue Ontology. Using the data from BRENDA database, one can select tissues/cell types that have links in the database to *H. sapiens* enzymes (left panel) and tissues/cell types that do not (right panel), which might indicate that they do not belong to *H. sapiens*. The color shows the order of magnitude of abstracts that include that co-occurrence. The graphs show that co-occurrences of tissues/cell types that have links in the database to *H. sapiens* enzymes are more concentrated in areas of higher mutual information and higher log likelihood, with higher number of abstracts. The left graph shows that the Log likelihood allows to separate the cases of low number of abstracts that the minimum information measure would normally rank with good scores (the weakness of this scheme).

Setting a threshold for these statistics, or for the number of documents where the relations should be found, is not straightforward. While some authors make recommendations on particular cut-offs, these recommendations are normally based on the use of particular dictionaries and corpus. In other words, the diversity of terms of a given dictionary and the ubiquity of their occurrences in the corpora analyzed will influence the behaviour of expected $E_{ij}$ and observed $O_{ij}$

values.

This work bases its cut-off on the behaviour of the organisms and tissues/cell-types dictionaries interaction, measured by the $MiM$ and $LLH$ statistics (Figure 3.10). These dictionaries are suitable to use for this comparison, as a list of tissues/cell-types that belongs to a particular species can be extracted from the whole dictionary. The co-occurrences of that list of tissues/cell-types with that source species can be compared against those of tissues/cell-types that are not part of the species.

I implemented the scores calculation within MySQL as stored procedures. This permits to obtain results as shown in Table 3.3, which shows the most significant co-occurrences between the small molecule Kynurenic acid and tissues/cell types from BRENDA Tissue Ontology, with just one query to the database. Kynurenic acid is indeed a relevant metabolite in the nervous systems, which is what the table expresses. In a single search we ask for a query entity (in this case a particular small molecule) and a target dictionary (in this case, BRENDA Tissue Ontology), and using thresholds for minimal MiM, LLH and sentence distance. Appendix F contains more examples like this, for both ubiquitous and specific metabolites. These examples show that ubiquitous metabolites – such as Pyruvate, Lactate or Taurine – associate with a wider variety of tissues/cell types compared to more specific metabolites – such as Ursodeoxycholic acid, Pregnenolone, or GABA – which associate to functionally related tissues/cell types.

These encapsulated queries, given the size of the main table Entity_has_Citation which holds more than 200 million records and needs to be joined with itself, can take between a few seconds to a few minutes depending on the dictionaries used in the query. I optimized queries and indices, but this seems to be at the limit of MySQL capacity.

To improve the response time of the queries involving the UniProt dictionary, the largest one in term count and occurrences count, I indexed each UniProt entry in the dictionary with the corresponding NCBI Taxonomy identifier for the organism – each UniProt entry belongs to one particular organism – so whenever the search is limited by organism, it only retrieves the correct UniProt identifiers.

| Sample | Obs. | Exp. | Total | MiM | LLH |
|---|---|---|---|---|---|
| Cerebral ganglion | 222 | 17.71 | 5.0E5 | 2.20 | 738 |
| Neuron | 146 | 8.29 | 2.3E5 | 2.69 | 572 |
| Hippocampus | 44 | 1.97 | 5.5E4 | 3.03 | 190 |
| Astrocyte | 25 | 0.94 | 2.6E4 | 3.29 | 116 |
| Corpus striatum | 26 | 1.09 | 3.1E4 | 3.12 | 115 |
| Excretion | 32 | 3.80 | 1.1E5 | 1.62 | 80 |
| Spinal cord | 25 | 2.95 | 8.3E4 | 1.64 | 63 |
| Locus ceruleus | 11 | 0.24 | 6.8E3 | 4.06 | 62 |
| Nerve | 37 | 8.06 | 2.3E5 | 0.74 | 55 |
| Cerebral cortex | 15 | 1.22 | 3.4E4 | 2.17 | 47 |

Table 3.3: Table of Best ranked biological samples (Brenda Tissue Ontology[49] entries) co-occurring with Kynurenic Acid, a nervous system related metabolite. In this case, results are sorted by Log Likelihood and are asked to have mutual information score $MiM > 0$, log likelihood score $LLH > 10$ and $t - Score > 0$. The column Obs. stands for observed number of co-occurrences (between the sample and Kynurenic Acid), Exp. for expected number of co-occurrences and Total for the total number of co-occurrences of that sample (Brenda Tissue Ontology entry) with all other small molecules. Only the term "Excretion" is not related to the nervous system.

## 3.5 Obtaining species specific co-occurrences towards a text-mining metabolome

The data of co-occurrence between the different biological entities represents a starting point for the generation of collections of metabolites for a species. In order to build such a collection we need to extract the relevant relations between the key objects. The database holds data for as many organisms as they are available in the NCBI Taxonomy dictionary used, as many proteins as the UniProt dictionary includes and so on for tissues/cell types and small molecules. However not all of these will be part of a metabolome, not all of these represent context to small molecules. Figure 3.11 shows the pipeline built to obtain the relevant objects and relations for a single organism.

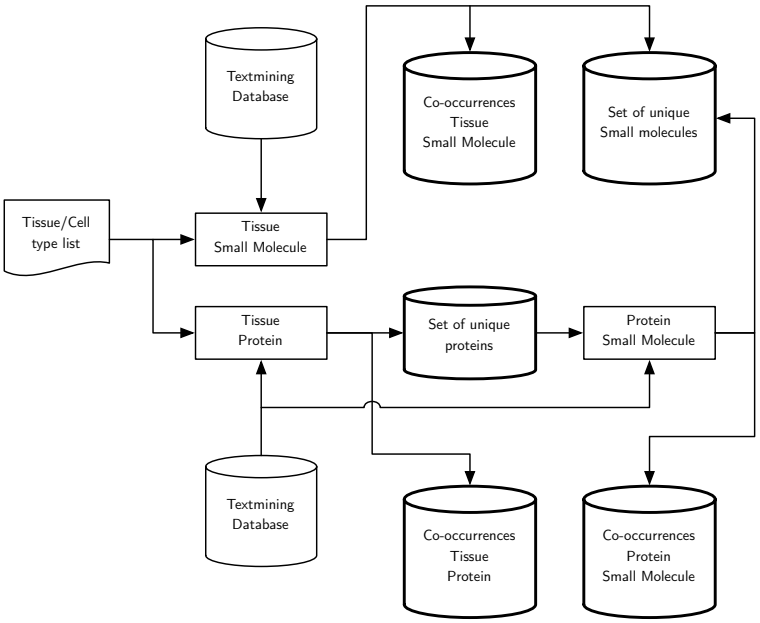



Figure 3.11: Diagram of selection of co-occurrences for a single species. The process starts with a set of tissues and/or cell types known to that organism.

The method starts with a comprehensive list of tissues and cell types that belong to the organism of interest, in this case *H. sapiens*. The software queries this list of tissues against the database of co-occurrences, limiting by the desired

organism, to obtain tissues and small molecules co-occurrences and tissues and proteins co-occurrences. These queries are limited with thresholds in MiM ($>$ 0), LLH ($>$ 10) and sentence distance ($\leq$ 1). These sets of co-occurrences are stored for later upload to BioWarehouse, both as co-occurrences and as proteins, tissues and small molecule entries. The set of protein-tissue yields a unique list of proteins, which the tools queries against the co-occurrences database to retrieve protein-small molecule co-occurrences, again limited by organism, score thresholds and sentence distance. Relations and small molecules retrieved are again stored for later upload to BioWarehouse. The pipeline yields ~14,000 small molecules entries from ChEBI and PubChem Compounds.

The co-occurrences approach only gives an idea on how often two terms are related, but does not imply directly that a small molecule will be a metabolite of the desired organism. To narrow down the solution to a set that can be more confidently described as metabolites of the organism, we need to filter out small molecules which should not be considered metabolites, such as chemicals of industrial use, plastics, adhesives, and others. The case of a *H. sapiens* metabolome is particularly difficult, as there are many varied interests in the use or application of different small molecules.

## 3.6 Filtering of text mining results

The number of retrieved chemical records (~14,000, Figure 3.11), impairs a complete manual classification of them, at least in the time frame of this thesis. Furthermore, the spirit of the project is to generate something that can be easily applied on other organisms, having to manually classify such a number of chemical records would make this difficult.

To classify this set of molecules into highly probable metabolites and not probable metabolites, I used a mixed approach of manual curation and machine learning methods. I collected a number of meta data features for each PubChem Compounds or ChEBI entry, relying on the NCBI MeSH chemical branch, the ChEBI Role ontology branch, some chemical classes of the ChEBI Ontology and on the appearance of the record in a number of "niche" databases (such

Figure 3.12: Diagram of the annotation of chemical entities obtained through text mining, previous to the filtering steps. The pipeline shown retrieves chemical structures and as much meta data (including cross references to a number of niche databases) for each of the ChEBI and PubChem Compounds identifiers obtained from the source databases. Functional classification information is retrieved from the ChEBI ontology role branch, from the KEGG BRITE chemical categories and from the PubChem Compounds entries annotated with chemical related NCBI MeSH terms. The tool calculates MACCS fingerprints for each structure and compares this against MACCS fingerprints for the small molecules obtained in the metabolism database part, storing the minimal distance and the average of the ten minimal distances. With all these data, the pipeline proceeds to the filtering step through machine learning methods.

as databases specializing in drugs, commercial chemicals, pesticides, bacterial secondary metabolites, etc.). I used a few chemical descriptors to increase the resolution power between metabolites and non-metabolites. These were natural product likeness score, number of carbon atoms and number of non C, H, N, O, S, or P atoms. I computed Tanimoto chemical similarities, through the MACSS fingerprint, against the set of metabolites unified in Chapter 2. This process also filters small molecules that have $> 500,000$ occurrences in the corpus, to avoid unspecific tags.

Figure 3.12 shows the annotation procedure (where meta data are collected) for the chemical entities resulting from the previous pipeline (Figure 3.11). This annotation process, comprising not only simple meta data but functional and ontological classification, is necessary not only for the classification process, but for later inclusion of the selected set of metabolites and surrounding context into BioWarehouse. The consolidation process implemented in Chapter 2 relies heavily on adequate meta data to consolidate small molecule sets. The following sections explore the detail of this annotation in terms of the technologies and resources used.

### 3.6.1 Annotating chemical entities through NCBI resources

NCBI MeSH is the Medical Subject Headings classification system of the NCBI, which is used by NCBI PubMed to organize and index life science literature. I explored the branch of NCBI MeSH corresponding to the Chemicals and Drugs Categories, that organizes the chemical knowledge into a tree of different categories and applications of the annotated PubChem Compounds entries. This branch has 16 major terms, to which many entities in PubChem Compounds are annotated. I selected a sub set of these topics (and sub topics, one level down, for increased granularity in same cases) as features for the metabolites classification step.

The aim is to have categories (in this case NCBI MeSH terms) that have a certain bias or probability of containing either metabolites or non-metabolites. None of the categories will be exclusive, but if they have a relevant bias, then the

| NCBI MeSH Term | Bias |
|---|---|
| Biomedical and Dental Materials | NM |
| Chemical Actions and Uses:Pharmacologic Actions | NM |
| Chemical Actions and Uses:Specialty Uses of Chemicals | NM |
| Complex Mixtures | NM |
| Hormone Antagonists | NM |
| Inorganic Chemicals | NM |
| Pharmaceutical Preparations | NM |
| Amino Acids, Peptides, and Proteins | M |
| Biological Factors | M |
| Carbohydrates | M |
| Enzymes and Coenzymes | M |
| Hormones | M |
| Lipids | M |
| Nucleic Acids, Nucleotides, and Nucleosides | M |
| Steroids | M |

Table 3.4: NCBI MeSH terms from the chemical branch that I expect to have biases towards either metabolites or non-metabolites. Column Bias shows the expected bias, either towards metabolite (M) or non-metabolite (NM). It is important to clarify that while I expect a certain bias from each NCBI MeSH term, this will be defined by the classification algorithm when using the training data that is provided to it.

later classification machine learning methods will weight them adequately. Table 3.4 contains the selected NCBI MeSH terms (topics and sub topics).

The annotation pipeline also retrieves a collection from PubChem Compounds called NCBI BioSystems. The NCBI BioSystems set of PubChem Compounds entries corresponds to small molecules present in the NCBI BioSystems project at NCBI (described in section 2.3.1.5, Chapter 2). The pipeline relies on the NCBI E-Utils web services to retrieve PubChem Compounds terms for each NCBI MeSH term and for the NCBI BioSystems collection.

The pipeline maps ChEBI small molecules to NCBI MeSH terms (and NCBI BioSystems) through cross references that the ChEBI entry might have to PubChem Compounds. Compounds in PubChem Compounds link to external database entities (such as ChEBI) through PubChem Substances. Using NCBI E-Utils, the pipeline retrieves PubChem Substances entries for each PubChem Compounds entry, and then for each PubChem Substances entry, cross refer-

ences to ChEBI and other databases detailed in the next sections. Lucene indices store the relations between PubChem Compounds entries and the resulting external database identifiers for later query. This allows to make the web service call to NCBI E-Utils just once (or once every some defined period of time, like a month for instance), speeding up the process, as querying local Lucene indices is much faster than querying a service through the web.

As mentioned, none of the headings or databases listed comprise only metabolites or non-metabolites, but each category will probably have a different tendency towards any of these two outcomes. For the same reason, we need to add more attributes that allow a trained machine learning method to resolve.

### 3.6.2 Annotating chemical entries with the ChEBI Ontology

The ChEBI Ontology organizes chemical knowledge for more than 27,000 chemical classes and small molecules, through different relationships. Besides its main chemical classification, the ChEBI Ontology has a branch of roles which is particularly useful for resolving metabolites from non-metabolites.

I selected high hierarchy elements of the Role branch of the ChEBI Ontology as features for the classification of chemical entries. The software classifies a ChEBI entry into one of these categories if the entry has a "has role" relationship to that role in the ontology, or if that relationship can be "reasoned" from the ontology. The pipeline extracts these relations for a given ChEBI entry, using the ChEBI OWL Ontology through the OWL API [57].

Assignments of roles within the ChEBI Ontology are unfortunately rather scarce, only ~15% of ChEBI entities with structure have a role assigned in the ontology. Figure 3.13 shows the accumulated distribution of the number of ChEBI entities that have roles assigned. I investigated ways to increase the number of assignments.

Software packages known as Semantic Reasoners can infer new relationships within an ontology that are not explicitly mentioned, starting from existing statements in that ontology. This process is called reasoning or semantic reasoning.

| Role/Class | # | Bias |
|---|---|---|
| Aetiopathogenetic role | | |
| Agrochemical | | NM |
| Antimicrobial agent | | NM |
| Buffer | | NM |
| Cofactor | | M |
| Detergent | | NM |
| Disinfectant | | NM |
| Drug metabolite | | M |
| Dye | | NM |
| Fixative | | NM |
| Flavouring agent | | NM |
| Food emulsifier | | NM |
| Fragrance | | NM |
| Fuel | | NM |
| Fuel additive | | NM |
| Hormone | | M |
| Hormone antagonist | | NM |
| Indicator | | NM |
| Label | | NM |
| Metabolite | | M |
| Mimotope | | |
| Pesticide | | NM |
| Pharmaceutical | | NM |
| Pharmacological role | | NM |
| Pheromone | | M |
| Photochemical role | | |
| Probe | | NM |
| Protein denaturant | | NM |
| Secondary metabolite | | M |
| Solvent | | NM |
| Surfactant | | NM |
| Sweetening agent | | NM |
| Tracer | | NM |
| Xenobiotic | | NM |

Table 3.5: Selection of ChEBI roles and classes that one could expect to have relevant biases towards metabolites and non-metabolites. Column Bias shows the expected bias, either towards metabolite (M) or non-metabolite (NM). It is important to clarify that while I expect a certain bias from each ChEBI role or class term, this will be defined by the classification algorithm when using the training data that is provided to it.

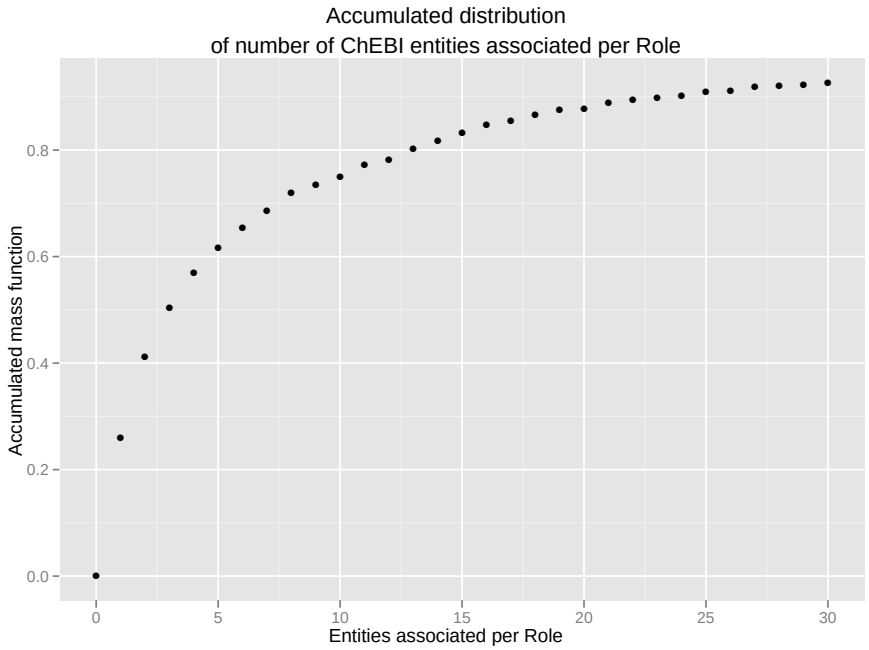



Figure 3.13: Accumulated distribution of the number of direct assignments to ChEBI entities with structure that the ChEBI roles have. The graph shows that ~75% of the roles have 10 or less ChEBI entities assigned. As of this version of ChEBI (Feb. 2012), there are 39 ChEBI roles with more than 29 assigned molecules with structures, some examples are: herbicide (30 molecules), opioid analgesic (33), local anaesthetic (36), mutagen (38), antibiotic (40), H1-receptor antagonist (53), secondary metabolite (65), metabolite (87), epitope (176), and finally fluorochrome (411).

Through reasoning, one can increase the number of assignments of Roles to ChEBI entities.

For instance, the ChEBI entity 1-phenanthrol (CHEBI:27528) does not have a role assignment. 1-phenanthrol is an instance of the class phenanthrenes (CHEBI-:25961), which has a role as xenobiotic. A direct query to ChEBI for a role for 1-phenanthrol yields no results, yet through the reasoner we can get the xenobiotic assignment[1].

When the reasoner fails to assign a role to a ChEBI entry, the pipeline uses a ChEBI entry to role assignment produced through Hearst patterns[54]. This assignment was provided by Adam Bernard, fellow PhD student at the text mining group at the EBI. One example of Hearst patterns for this type of assignment would be:

```
<small molecule noun> is a <role or class noun>
```

This pattern would be found in these phrases:

```
‘‘pyruvate is a metabolite’’
‘‘aspirin is a drug’’
```

When none of the previous attempts result in a role assignment, the pipeline searches against the original text mining database. In this case, it retrieves co-occurrences between the ChEBI record (or a PubChem Compounds record) and ChEBI roles. Only very frequent, high scoring and same sentence results are retrieved, to avoid false positive role assignments.

As the aim is to annotate roles to as many chemical records as possible, to improve the classification of molecules, I annotated PubChem Compounds entries in the result with a role if they had a ChEBI cross reference that can be assigned to the role by the same method.

### 3.6.3 Annotation of chemical entries using KEGG resources

The method uses KEGG in two ways to aid in the classification of small molecules into metabolites or non-metabolites. The KEGG COMPOUND database con-

[1] This example is the simplest case, called a direct assertion

| KEGG BRITE Term | Bias |
|---|---|
| Carbohydrates | M |
| Lipids | M |
| Nucleic Acids | M |
| Peptides | M |
| Cofactors | M |
| Steroids | M |
| Alkaloids | NM |
| Terpenoids | NM |
| Flavonoids | NM |
| Hormones/Transmitters | M |
| Antibiotics | NM |
| Pesticides | NM |
| Pesticides/Herbicides | NM |
| Plasticizers/Plastics | NM |
| Phytochemical Compounds | NM |

Table 3.6: KEGG BRITE terms selected with expected biases towards metabolites and non-metabolites.

tains mainly metabolites, and hence we considered the annotation against a compound in KEGG as a niche database biased for metabolites. It also uses KEGG DRUG and KEGG GLYCAN as niche database that could tell us something about the nature of the small molecules to classify.

Additionally, KEGG provides a hierarchy of biological entities named KEGG BRITE, which contains a few branches organizing chemical knowledge. I retrieved the KEGG COMPOUND, DRUG and GLYCAN identifiers for each of the KEGG BRITE categories in Table 3.6.

### 3.6.4 Annotation of chemical entries with niche databases

I selected a collection of niche databases to annotate chemical entries. Besides the text mining results classification, retrieving these cross references is useful for the unification step, which makes use of external identifiers for discriminating whether two molecules of equal connectivities are the same or not. The databases are listed and described in the following enumeration:

**EINECS:** The European Inventory of Existing Commercial chemical Substances. It is mainly cross referenced by PubChem Compounds.

**HSDB:** The Hazardous Substances Data Bank. It is mainly cross referenced by PubChem Compounds.

**HMDB:** The Human Metabolome Database. Although not used for the classification step, it is used for later validation and is useful as a general cross reference. It is accessible through both ChEBI and PubChem Compounds entries.

**ZINC:** an open database for virtual screening compounds. It is expected to be mainly a source of synthetic lead compounds, and is regularly used [28; 64] in the literature as source of non-metabolites small molecules. It is cross referenced by both ChEBI and PubChem Compounds.

**EPA Pesticide:** The US Environmental Protection Agency (EPA) has a collection of small molecules used as pesticides, herbicides and insecticides. This contains both synthetic and natural molecules. In the case of *H. sapiens*, probably none of the metabolites should be listed here, but in the case of some plants, bacteria, insects and fungus it might well be the case that some metabolites are part of this database. *H. sapiens* could be exposed to some of these metabolites in some conditions, and hence they might be detected in samples. However, for the sake of this exercise, I assume that this database mostly holds metabolites that should not be found in *H. sapiens*.

**BRN:** The Beilstein Record Number. This is mainly collected as a useful cross reference, and it is not used as a classification feature.

**KEGG Compound:** The main collection of small molecules in the Kyoto Encyclopedia of Genes and Genomes, the Compound collection (also called Ligand collection), holds mostly metabolites that can be mapped to biochemical reactions, with some exceptions.

**KEGG Drug:** Drug collection of the Kyoto Encyclopedia of Genes and Genomes. It can be accessed both from PubChem Compounds and ChEBI.

**CAS:** The Chemical Abstracts Service registry number. As with BRN, it is only kept as a useful cross reference, and not used as a classification feature.

**ChEMBL:** is a database of bio-active, drug-like, small molecules. It contains both synthetic and natural molecules, some of which could be metabolites of the organism in question (for instance, some hormones in *H. sapiens* are used as drugs).

**ChemIDplus:** A collection of nearly 380,000 small molecules, which includes compounds tested for toxicity, chemicals causing cancer and/or birth defects, hazardous pollutants, controlled substances, pesticides and in general substances that would be of interest for regulatory agencies (many parts of it are actually provided by regulatory agencies). This collection probably contains very few *H. sapiens* metabolites, but some plants and fungus secondary metabolites.

**KEGG Glycan:** A collection of experimentally determined glycan structures.

**LIPID MAPS:** Nature Lipid Maps is one of the largest collections of lipid structures that have been detected experimentally.

**BioCyc:** A collection of metabolism databases for several organisms. Chemical structures contained here are mostly, if not all, metabolites, as to be included they need to participate in a reaction.

**UM-BBD:** The University of Minnesota Microbial biocatalytic reactions and biodegradation pathways. Holds mostly xenobiotic chemicals and information on how they are degraded by microorganisms.

The presence of a molecule in one of these databases cannot guarantee, nor completely rule out, that a small molecule might be a metabolite, and further that it might be a *H. sapiens* metabolite. However, the fact that a molecule belongs to one of these collections gives an implicit probability of being a metabolite. The aim of using these collections is to serve as features in the classification, they add or subtract to the overall probability of a molecule being a metabolite or not.

### 3.6.5 Chemical descriptors used to classify

Although a big effort was done to collect relevant meta data for each of the entries, many of them will have none assigned for many reasons. ChEBI roles for instance, are sparsely annotated in the ChEBI database, let alone those roles be migrated to PubChem Compounds entries, which in many occasions do not exist in ChEBI.

There are probably a number of unknown chemical characteristics that would either make a small molecule a metabolite. However, it is difficult to come up with a good set of structural characteristics, as chemistry is varied and complex. Some attempts have been done in the past on classifying small molecules as either closer to lead compounds, drugs or metabolites [28], or to calculate how much a small molecule resembles natural products or synthetic molecules [31]. In [28], the major resolving features between Drugs and metabolites are LogD (the two phase partition coefficient), atom count and Tanimoto similarity through the MDL Public Keys fingerprint. In [64] an NP-Likeness scorer is built based on circular atom signatures. In [95] authors use 32 physicochemical descriptors from the MOE commercial chemical package to discriminate metabolites into different classes (within metabolism). However, I have no access to that software, and the documentation does not explain clearly how to derive them.

I calculate the number of carbons (non-metabolites tend to be distributed towards bigger molecules [28]), the number of non CHNOSP atoms (I expect non-metabolites to have more) and the NP-Likeness [64]. I calculate the Tanimoto similarity between each molecule and all the molecules in the metabolites retrieved in Chapter 2.

### 3.6.6 Machine learning for chemical entries classification

After manually and semi manually curating ~2,800 chemical entries (out of the 14,000), a machine learning approach was used to try to classify the remaining 11,000 entries as metabolites or non-metabolites. The classification was done using 101 attributes (annotation of ChEBI roles and classes; NCBI MeSH Terms; KEGG BRITE categories; presence in niche database; chemical descriptors). For those categories that were equivalent across the different classification systems,

the method used merged versions of them.

The first step is to choose a regression method to use with the data. In general, one should always start with linear methods, as they do not suffer from high dimensionality issues and over fitting as non linear methods. If a problem is reasonably linearly structured, a linear classification should normally work better than non linear methods. As a rule of thumb, only when linear methods fail, one should step into non linear methods.

I used the RapidMiner machine learning environment[103] to try a few different methods with the data set. Given the simplicity of the environment, it is easy to build several models, cross validate and compare the models validation to pick the best one. Most of my understanding of machine learning methods comes from the excellent treatment of the subject given in [7; 52]. I used the following methods:

**Decision Trees:** are highly interpretable classifiers, that work by successively splitting the data set in two parts according to conditions on the attribute values. This classifier yields very good results in certain data conditions. Having sparse categorical attributes (without a value set for many of the entries) or too many non categorical attributes sometimes can lead this method to poor results, mainly due to over fitting the data. A key decision in binary trees are the tree size and the splitting criteria. Even though this is a highly non linear method, it is so widely used in this kind of problems that is reasonable to try it.

**SVM (linear and non linear):** support vector machines, or SVM, generate hyperplanes in a multidimensional space, based on example points for each of the classes that one wants to classify, that produce the best separation of those classes. Once a good set of example points (that produce good separation hyperplanes) has been selected and hyperplanes generated, new values are localized in the space to see in which region (class) they fall into. Whether an SVM is linear or not, depends on the kernel function used for the internal products (which is in the linear case is a dot product), which can lead to have non linear boundaries in the original space (in the separation space, the boundaries will always be linear). Linear SVM

is different to Linear Discriminant Analysis (LDA) in that LDA computes the boundaries based on all the points of a class (using the centroid and co-variances), whereas SVM only uses sets of points from the class that are close to the boundaries, which locally gives a better solution, and has less computational cost.

**Logistic regression:** As the SVM, the logistic regression method also generates linear hyperplanes that act as boundaries between classes. This method transforms the output of a regular linear regression through a logit function[1], which maps the entire domain to a continuous range $[0, 1]$, understood as the probability or confidence of an event happening (for instance, the probability of being a metabolite). This makes logistic regression specially fit for binary classification problems, like the one we have.

**Random forests:** Is an ensemble method in which many random decision trees are generated, and a majority vote decision – from the trees – is taken for each of the submitted points to be classified. Random forests aim at reducing the high variance of noisy methods (like decision trees) through the averaging produced by voting, without changing the bias too much. The method grows trees by randomly selecting a subset of the attributes. When there are few relevant attributes and many noisy attributes, the probability that at each selection the method will choose mostly uninformative attributes is high, producing a poor performance of the method. There are some claims that random forests cannot over fit data by construction [52, p. 596].

**Linear perceptron:** It is the simplest form of a neural network, which maps an input vector to a binary output through a sigmoid function (and a dot product with a weight vector). This makes it very similar to the logistic regression, however the main difference is that the perceptron uses a scale factor accompanying the input value which controls the "activation rate". The scale factor is learnt from the data as well.

[1]The logit function is $f(x) = \frac{1}{1+e^x}$

I used 10-fold cross validation for every method on the ~2,800 examples, using stratified sampling. Alternatives to 10-fold cross validation is the leave one out method, which is more computationally demanding as it builds N models, and probably worse at judging the regression method on the data, as all the models will be relatively similar if N is big, as it is in this case.

Figure 3.14 shows the methods performance according to ROC curves for the 10-fold cross validation. The best method was SVM (linear and non-linear), closely followed by Logistic Regression. Decision trees, as expected, had a very high variance, and were probably over fitting the example data. Random forests performed even worse than random picking (lazy guessing in Figure 3.14).

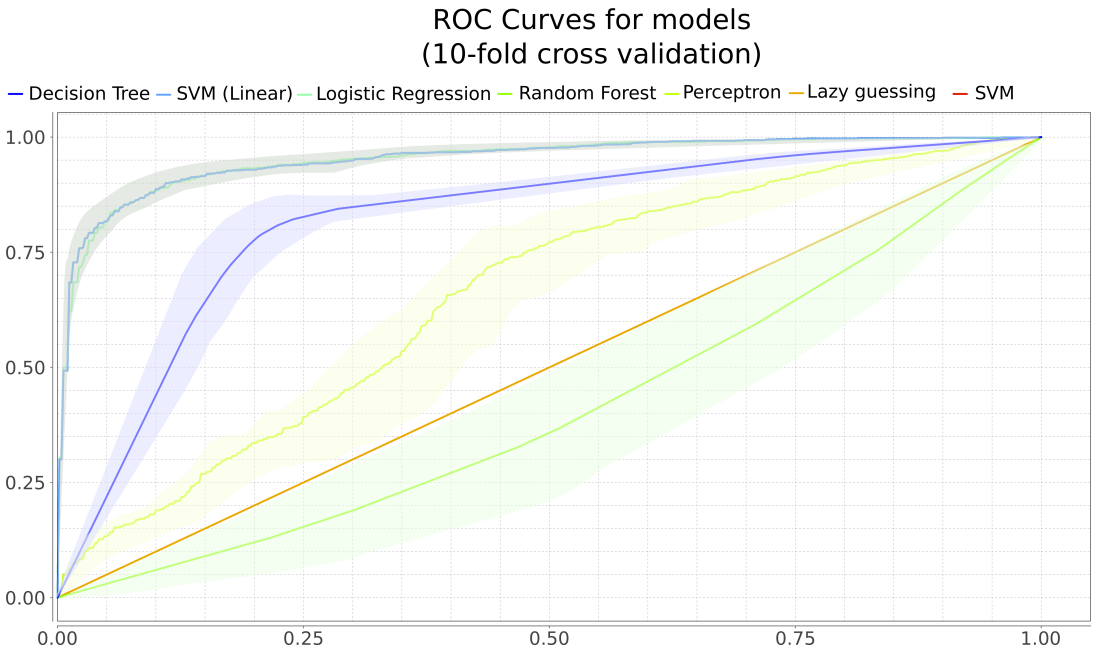



Figure 3.14: ROC comparison graph showing the performance of the different methods tried on the training and validation data set. I used a 10-fold cross validation strategy, with stratified sampling, to asses the methods capability of predicting the data set. The best performing methods were Support Vector Machines and Logistic regression. Random forests performed even worse than random choosing (lazy guessing), probably because many of the attributes chosen might not be that informative.

| Attribute | Weight |
|---|---|
| Tanimoto Similarity MACSS | -1.555 |
| Non CHNOSP count | 1.228 |
| BioSystems collection | -1.156 |
| Carbon count | 0.889 |
| ChemIDplus | 0.849 |
| Drug | 0.769 |
| NP-Likeness | -0.673 |
| BRITE Peptides | -0.61 |
| ChEBI Nucleoside | -0.592 |
| EPA PESTICIDE | 0.538 |
| ChEBI Carbohydrate | -0.529 |
| BiologicalMolecule | -0.5 |
| MeSH Enzymes & Coenzymes | -0.489 |
| Brite Hormones & Transmitters | -0.487 |
| Carbohydrates | 0.419 |

Table 3.7: Top 15 influential attributes in the logistic regression model. Tanimoto similarity is against the set of *H. sapiens* metabolites collected in Chapter 2. Negative weights indicate tendency towards metabolites.

Table 3.7 shows the most important attributes according to the Logistic regression model.

Table 3.8 shows the most relevant attributes according to the linear SVM model.

Using the HMDB as a final check, we find the best cut-off point for both regressions and for a union of both models (if either of them decided the molecule a metabolite, then it is classified as metabolite). Figure 3.15 shows the sensibility and specificity of the prediction compared to the HMDB for the logistic regression model, for different cut-off points. Figure 3.16 shows the same for the support vector machine regression.

| Attribute | Weight |
|---|---|
| Tanimoto Similarity MACSS | -0.546 |
| Carbon Count | 0.426 |
| BioSystems | -0.4 |
| Non CHNOSP count | 0.36 |
| Tanimoto 10 | -0.355 |
| NP-Likeness | -0.326 |
| ChemIDplus | 0.319 |
| Biological Molecule | -0.252 |
| EPA PESTICIDE | 0.237 |
| Tanimoto 1-10 | -0.219 |
| Drug | 0.203 |
| ChEBI Carbohydrate | -0.175 |
| ChEMBL | 0.166 |
| ChEBI Role aetiopathogenetic role | 0.166 |
| Brite Peptides | -0.164 |

Table 3.8: Top 15 influential attributes in the support vector machine regression model. Tanimoto similarity is against the set of *H. sapiens* metabolites collected in Chapter 2. Tanimoto 10 refers to the average distance calculated to the 10 more proximal elements in distance in the metabolites collection. Tanimoto 1-10 refers to the difference between the most proximal distance and the Tanimoto 10 average. Negative weights indicate tendency towards metabolites.

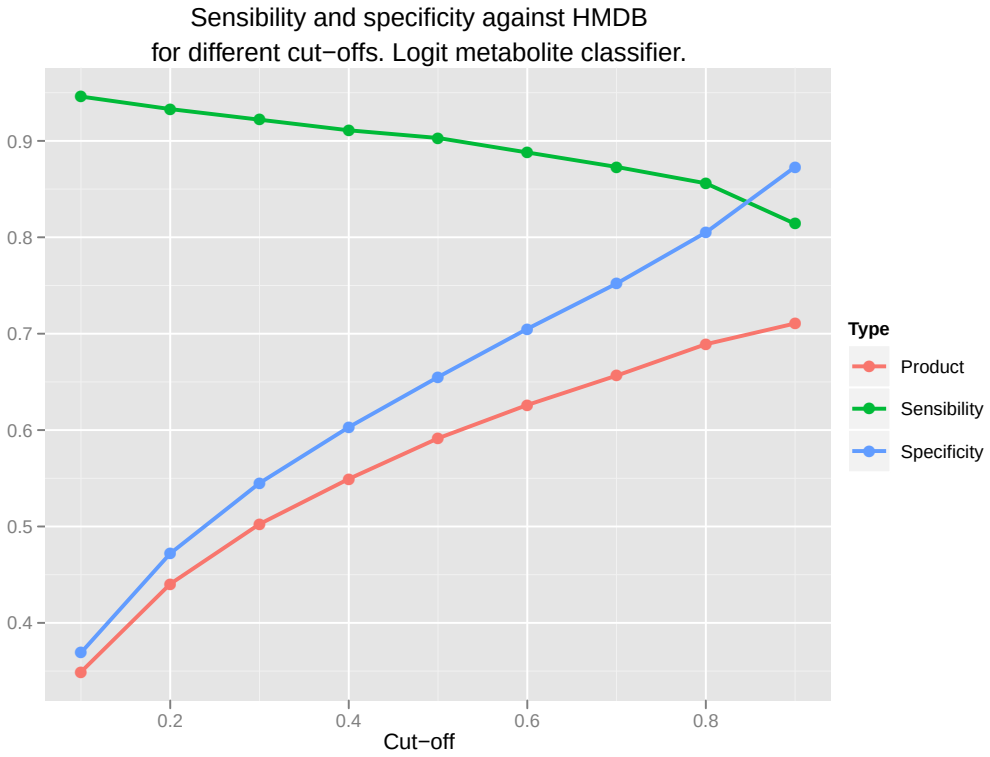



Figure 3.15: Sensibility and specificity for a varying cutoff of the logistic regression based classifier, when the result is compared against HMDB. A cut-off of 0.85 for the classifier (above this value is a metabolite, below is not) provides a reasonable compromise, where sensibility and specificity are above 80%.

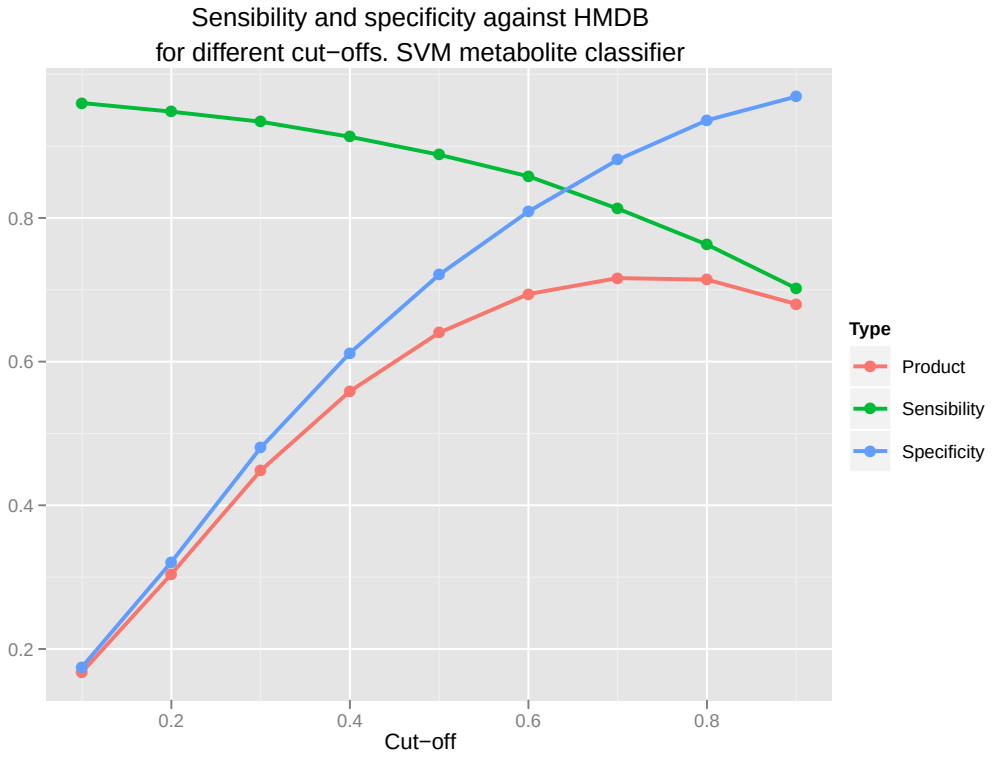



Figure 3.16: Sensibility and specificity for a varying cutoff for the SVM based classifier, when the result is compared against HMDB. A cut-off of 0.65 for the classifier (above this value is a metabolite, below is not) provides a reasonable compromise where both sensibility and specificity are above 80%.

## 3.7 Main Results

### 3.7.1 Comparison against the HMDB and database unification

Using the unification method proposed in Chapter 2, I compared the HMDB data set and the database unification for *H. sapiens* with the text mined generated metabolome. Figure 3.17 shows the intersection of chemical entities between these three sets. The text mining effort adds entries existing on HMDB that the database unification did not provide.

Enrichment analysis through BiNGO using the ChEBI mappings for the unique region of text mining small molecules (785 had ChEBI identifiers) reveals that this set of small molecules includes many exogenous natural products that are used as drugs or that are part of diet. Metabolism databases tend to concentrate more on endogenous metabolites, whereas the text mining derived metabolome produces more exogenous molecules that affect *H. sapiens*. As with other *H. sapiens* sets, this portion of small molecules also shows some level of enrichment of lipids and particularly steroids. Tables C.4, C.5, and C.6 in Appendix C show these results.

The same region, but through 1973 different PubChem Compounds identifiers, mostly complementary to the previous ChEBI subset, shows enrichment in nucleotide-like, biological factors, lipids, hormones, coenzymes, steroids, and pharmaceutical preparations. There is far less over representation in the PubChem Compounds set of exogenous natural products. Tables C.5 and C.6 from Appendix C present these results.

### 3.7.2 Small molecules and tissues/cell types

The distribution of small molecules with strong co-occurrences to tissues/cell types is long tailed: most small molecules show strong associations with less than 5 tissues/cell types; a very small number show strong associations with even hundreds of tissues/cell types. The same applies when the data are inspected from the tissues perspective.

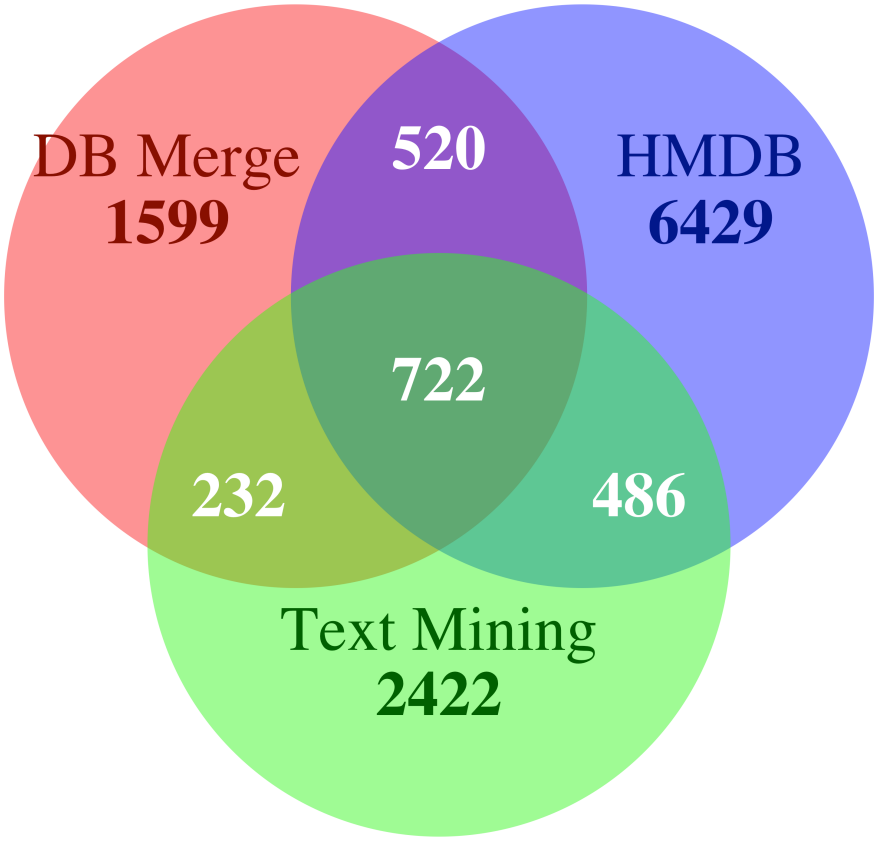



Figure 3.17: Small molecule intersections for the database consolidation, HMDB and Text mining results for *H. sapiens*. The text mining collection discovers nearly ~500 additional small molecules from HMDB not found in the databases. Text mining additionally provides more than 2,000 new small molecules that could be part of a *H. sapiens* metabolome.

Unfortunately the need to impose restrictions at the level of organism, and the need to require high scores to avoid very promiscuous terms, reduce the evidence of co-occurrences between small molecules and the biological containers. As a result of this, examples like the one of Kynurenic Acid, visible before the two pipelines and machine learning methods are applied, loses most of it signal once these filtering algorithms are applied.

Figure 3.18 shows two graphs of strong co-occurrence relations between tissues/cell types of the central nervous system[1] and small molecules. The graph to the left shows the relations when only direct small molecules to tissues (constrained by organisms and co-occurrences scores as explained previously) are used. In this graph there are in general few assignments per tissue, which does not make it wide enough for the generation of metabolomes at the tissue and cell type level. The box plot in the lower part of the same Figure 3.18 shows the distribution of the degree (number of connections) for this case (Direct), both for small molecules and tissues/cell types.

A way of alleviating the low level of direct connectivity is through the use of protein co-occurrence, which can relate small molecules to proteins and then proteins to tissues, producing a transitive relation of small molecules to tissues. The graph to the right in Figure 3.18 shows small molecules to tissues relations, when stepping through proteins co-occurrences, for the central nervous system related tissues/cell types. The effect, shown by the graph, is that the use of protein relations recruits an important number of additional small molecules, and further connects them more densely to tissues. This can be seen in the graph as many more blue squares, representing small molecules, can be seen now in the central region of the graph, compared to the other case where the blue squares are more densely located in the perimeter of the graph. The box plots in the same Figure show, for both small molecules and tissues, how the degree is considerably increased in the case were protein relations ("Through proteins") are used compared to the case without proteins ("Direct"). Another interesting outcome is that tissues that are related tend to get clustered together. In the right graph, there are two big clusters of red squares, one of them belongs to

[1]To retrieve all tissues/cell types within the central nervous system, I descend through the BRENDA Tissue Ontology starting from the term "Central nervous system".

different major brain parts, while the other belongs to cell types that derive from glial cells. In the same graph, another smaller red cluster forms, that contains the hypophysis, the adenohypophysis, and other glands of the central nervous system.

The inclusion of proteins in the relation, albeit increases the number of small molecules than can be associated to tissues/cell types, also tends to make the set of small molecules less specific to the set of tissues queried. Lets explore an example comparing liver related tissues/cell types and brain related tissues/cell types. The BRENDA Tissue Ontology allows to obtain tissues and cell types related to each of these two organs by descending through the ontology starting from "liver" and "brain" terms.

Table 3.9 contains comparison of the enrichment in ChEBI Ontology categories between small molecules related to brain (and derived tissues) and related to liver (and derived tissues), for the portion of small molecules that have a ChEBI identifier. The table shows nearly no difference when stepping through proteins. Only slight re-arrangements of the relevance of some classes are seen, but mostly the top 20 most enriched small molecule categories for both tissues/cell types collection are the same. This is indication of a shared core metabolism, and smaller unique sections of it for each tissue, as one would expect maybe. This is reasonable when compared to results from [144], where gene sets in *H. sapiens* that are completely tissue specific are normally below 100 genes; less than 1% considering a conservative estimate of 20,000 protein coding genes.

From this same set of molecules, using those with PubChem Compounds identifiers, there is a core of 963 shared entries (from both small molecules related to liver and brain), 126 entries unique to brain and 171 unique to liver. The same exercise with those small molecules that have ChEBI identifiers, yields a core of 700 shared ChEBI entries, 74 unique entries to brain and 130 unique to liver. Enrichment analysis of these ChEBI sets shows a slight enrichment of steroids in the unique liver collection, no major groups are enriched in the unique brain part. However, looking closer into the set of molecules, Kynurenic acid and dopachrome are examples of molecules that appear in the list of unique brain entries that might make some sense. These numbers again reflect the idea of a core metabolism and

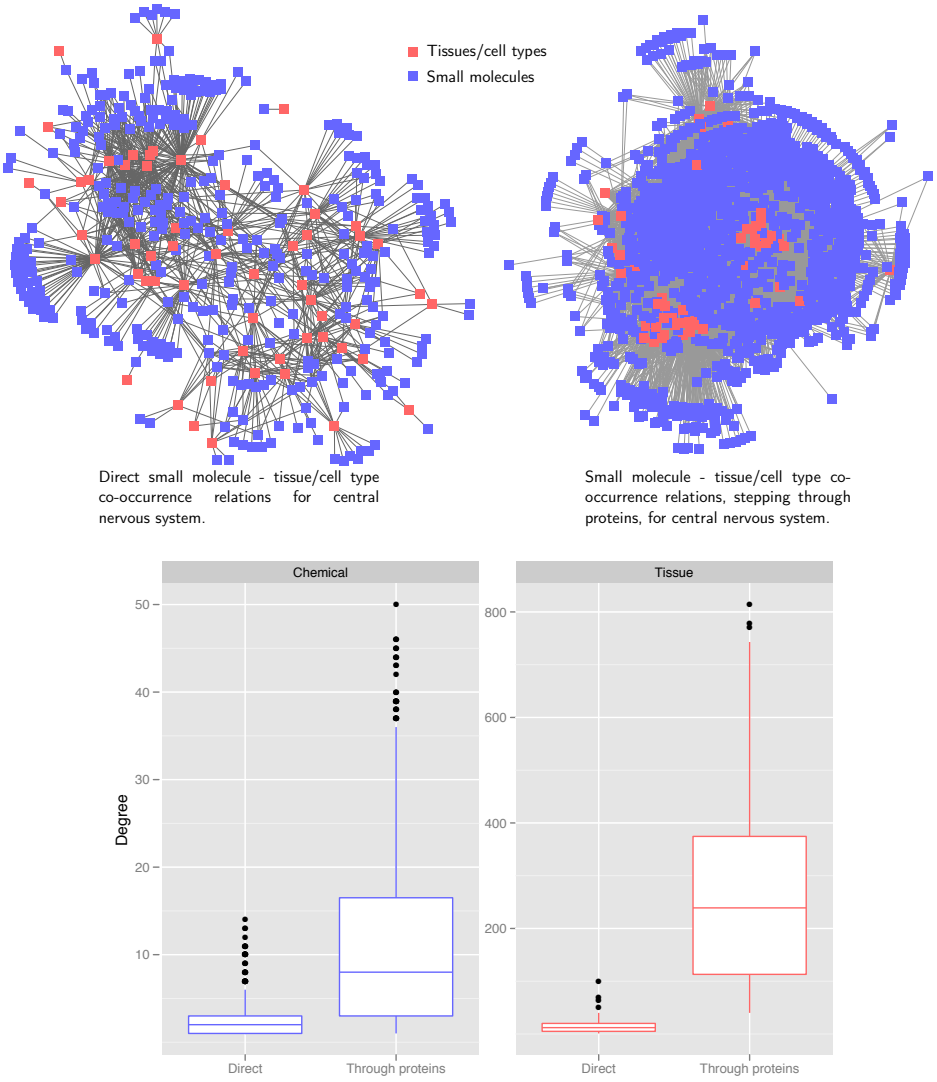



Figure 3.18: Graphs show co-occurrences between tissues/cell types belonging to the central nervous system, in red, and small molecules, in blue. Nodes that are farther from the centre tend to have lower degree (count of connections). *Left:* graph displays direct tissues to small molecules co-occurrences. *Right:* graph presents the interactions when tissues to proteins, and proteins to small molecules co-occurrences are considered. Stepping through proteins considerably increases the number of small molecules related to a tissue, and improves the clustering of tissues. The right graph main tissue clusters correspond to brain parts and glial-related cells. Box plots show how the level of connectivity of both chemicals and tissues increases when moving from direct co-occurrences to protein mediated co-occurrences.

| Class | # | | p-value | | Fold | | % | |
|---|---|---|---|---|---|---|---|---|
| | B | L | B | L | B | L | B | L |
| steroid | 1 | 1 | 1.2E-14 | 8.8E-19 | 3.1 | 3.4 | 9 | 10 |
| molecular messenger | 2 | 2 | 1.9E-12 | 1.9E-13 | 4.7 | 4.7 | 5 | 5 |
| hydroxides | 3 | 3 | 4.8E-12 | 1.1E-12 | 1.5 | 1.5 | 34 | 34 |
| alpha-amino acid | 4 | 4 | 7.8E-12 | 3.4E-12 | 4.3 | 4.2 | 5 | 5 |
| biological role | 5 | 5 | 1.2E-11 | 4.4E-12 | 1.7 | 1.7 | 26 | 26 |
| eicosanoid | 6 | 8 | 1.5E-11 | 5.8E-11 | 6.5 | 6.1 | 3 | 3 |
| fatty acid derivative | 7 | 9 | 1.6E-11 | 6.4E-11 | 6.5 | 6.1 | 3 | 3 |
| oxo steroid | 8 | 6 | 5.1E-11 | 1.8E-11 | 4.6 | 4.6 | 4 | 4 |
| steroid hormone | 9 | 10 | 2.7E-09 | 9.4E-11 | 8.0 | 8.3 | 2 | 2 |
| 3-oxo steroid | 10 | 11 | 1.3E-08 | 2.9E-10 | 5.2 | 5.4 | 3 | 3 |
| hydroxy steroid | 11 | 7 | 2.0E-08 | 3.4E-11 | 3.2 | 3.5 | 5 | 6 |
| chemical role | 12 | 17 | 7.9E-08 | 7.5E-08 | 1.5 | 1.4 | 29 | 28 |
| prostanoid | 13 | 19 | 3.6E-07 | 7.2E-07 | 6.7 | 6.3 | 2 | 2 |
| lipid | 14 | 13 | 4.1E-07 | 2.5E-08 | 1.6 | 1.6 | 20 | 20 |
| drug | 15 | 24 | 5.4E-07 | 7.4E-06 | 1.9 | 1.8 | 11 | 10 |
| physiological role | 16 | 21 | 6.1E-07 | 1.6E-06 | 5.0 | 4.7 | | |
| hormone | 17 | 16 | 7.4E-07 | 3.7E-08 | 4.0 | 4.2 | 3 | 3 |
| pharmaceutical | 18 | 31 | 8.6E-07 | 1.2E-05 | 1.9 | 1.8 | 11 | 10 |
| agonist | 19 | 14 | 1.1E-06 | 2.8E-08 | 3.5 | 3.8 | 4 | 4 |
| tetrahydrofuranol | 20 | 38 | 1.7E-06 | 4.8E-05 | 2.8 | 2.5 | 5 | 4 |

Table 3.9: Top 20 enriched ChEBI Ontology categories for small molecules related to brain, when using protein co-occurrence to produce a small molecule to tissue relation. The table shows the results for brain (B) and liver (L). # stands for the ranking of the category in the enrichment for either brain (B) or liver (L). Corrected p-value corresponds to the p-value after Benjamini-Hochberg correction for false discovery rate. Fold is the number of times that the category is enriched in the sample compared to the overall ontology. % shows the portion of the sample that falls in that category.

variable smaller parts for each type of tissue.

Results change markedly if direct small molecules to tissues/cell types are considered (instead of stepping through proteins), for both liver and brain. For liver and associated tissues/cell types, Table 3.10 shows that steroids, cholanoids, lipids and bile acids are among the most enriched ChEBI categories. In the case of brain and associated tissues/cell types, Table 3.11 enrichment shows neurotransmitters, molecular messengers, and nucleosides[1] among the most enriched

[1]There is research showing that nucleosides might play important roles in the neuronal functions of the brain: [76; 86], to name a few.

| ChEBI Class | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| steroid | 14 | 9.0E-12 | 4.6 |
| cholanoid | 5 | 1.1E-09 | 16.3 |
| organic polycyclic compound | 14 | 1.5E-06 | 2.8 |
| lipid | 25 | 2.5E-06 | 2.0 |
| hydroxy steroid | 7 | 1.0E-05 | 4.3 |
| oxysterol | 2 | 1.3E-05 | 27.2 |
| bile acid | 2 | 1.3E-05 | 43.0 |
| biochemical role | 18 | 1.9E-05 | 2.2 |
| organic cyclic compound | 46 | 2.0E-05 | 1.5 |
| oxacycle | 16 | 6.8E-05 | 2.2 |
| hydroxy-5beta-cholanic acid | 2 | 7.4E-05 | 27.8 |
| nutraceutical | 2 | 8.3E-05 | 17.8 |
| 1-benzopyran | 6 | 2.5E-04 | 4.0 |
| drug | 13 | 2.8E-04 | 2.3 |
| nucleoside | 6 | 2.8E-04 | 3.9 |
| benzopyran | 6 | 3.2E-04 | 3.8 |
| biological role | 26 | 3.6E-04 | 1.7 |
| antioxidant | 3 | 3.9E-04 | 9.9 |
| 5beta-cholanic acids | 2 | 3.9E-04 | 18.3 |
| pharmaceutical | 13 | 4.1E-04 | 2.2 |

Table 3.10: Top 20 ChEBI enriched categories for small molecules collected from direct co-occurrences to liver derived tissues and cell types. Steroid metabolism, bile acid generation (cholanoid is a parent category for bile acids), and nucleoside synthesis are known processes to occur in liver cells. Also the processing of drugs and other complex xenobiotics.

ChEBI classes and roles. Intersects, both at PubChem Compounds compounds as well as ChEBI entries are proportionally much smaller compared to the unique parts.

Using protein results and gene expression repositories, it is feasible to evaluate how much sense the co-occurrence results make. The list of proteins that co-occurs with liver related tissues/cell types, contains approximately 723 UniProt entries. Out of this collection, ~86% of them show either up or down regulated expression in at least one experiment in the ArrayExpress ATLAS gene expression repository for "liver" (Experimental Factor Ontology entry EFO_0000887). Using the gene expression module from DAVID with the same set of proteins, ~21%

| ChEBI Class | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| molecular messenger | 8 | 1.4E-07 | 7.6 |
| neurotransmitter | 4 | 1.3E-06 | 23.0 |
| organonitrogen compound | 47 | 1.3E-06 | 1.7 |
| nucleoside | 8 | 2.0E-05 | 5.3 |
| tetrahydrofuranol | 8 | 2.6E-05 | 4.7 |
| pharmaceutical | 16 | 2.6E-05 | 2.7 |
| drug | 16 | 3.6E-05 | 2.7 |
| oxolanes | 9 | 6.9E-05 | 4.1 |
| 2'-deoxyribonucleoside | 3 | 1.7E-04 | 16.6 |
| heteroarene | 20 | 2.0E-04 | 2.2 |
| pyrimidine nucleoside | 4 | 2.4E-04 | 9.2 |
| N-glycosyl compound | 8 | 2.7E-04 | 4.0 |
| reactive oxygen species | 2 | 2.7E-04 | 37.2 |
| organic amino compound | 20 | 2.7E-04 | 2.1 |
| pharmacological role | 9 | 2.7E-04 | 3.4 |
| agonist | 5 | 5.7E-04 | 5.3 |
| deoxyribonucleoside | 3 | 6.8E-04 | 11.9 |
| pyrimidine 2'-deoxyribonucleoside | 2 | 1.6E-03 | 22.2 |
| reactive nitrogen species | 1 | 1.6E-03 | 46.5 |
| tryptamines | 2 | 1.8E-03 | 20.7 |

Table 3.11: Top 20 ChEBI enriched categories for small molecules collected from direct co-occurrences to brain derived tissues and cell types.

of the mapped proteins had high levels of expression in liver according to the GNF Human Affymetrix Array U1333A [145], and ~40% have expression in liver according to the UniProt tissue annotation.

Repeating the same analysis for brain related tissues/cell types, which includes 564 proteins with UniProt entries, ~92% of them showing expression changes in at least one experiment in the ArrayExpress ATLAS gene expression repository for "brain" (Experimental Factor Ontology entry EFO_0000302). UniProt tissue annotation shows no major enrichment for brain or nervous systems tissues/cell types (but for plasma, blood, liver, lung, and placenta). Slight enrichments in UniProt annotation are seen for fetal brain cortex, Cajal-Retzius cells (which are a type of neuron), cerebellum, fetal astrocytes, and Alzeimer's patients cortex. The GNF genes to tissues arrays [145] show high expression of these proteins in a number of brain related tissues, such as globuspallidus (25%), sub thalamic nucleus (27%), cerebellum (73%), and occipital lobe (22%), among others.

Figure 3.19 shows a complete overview look at the small molecules to tissue/cell types relations selected for *H. sapiens*. This is a hierarchical clustering of both tissues/cell types and the related small molecules, through their co-occurrences. While the data of co-occurrences classifies together nearly half of the participating tissues/cell types, in aggregated categories such as "Digestive", "Nervous system", or "Reproductive system", it does not seem to have the same classification power to gather related small molecules together. Only a few small clusters, one of "eicosanoids" and another of "fatty acids" could be identified.

Figure 3.20 shows the same approach of clustering applied to the small molecules to tissue/cell types relations when stepping through proteins. While in this case the amount of relations is much higher, the clusters that can be formed for tissues/cell types are much smaller. Apparently, the use of proteins tends to generalize the data, losing some specificity.

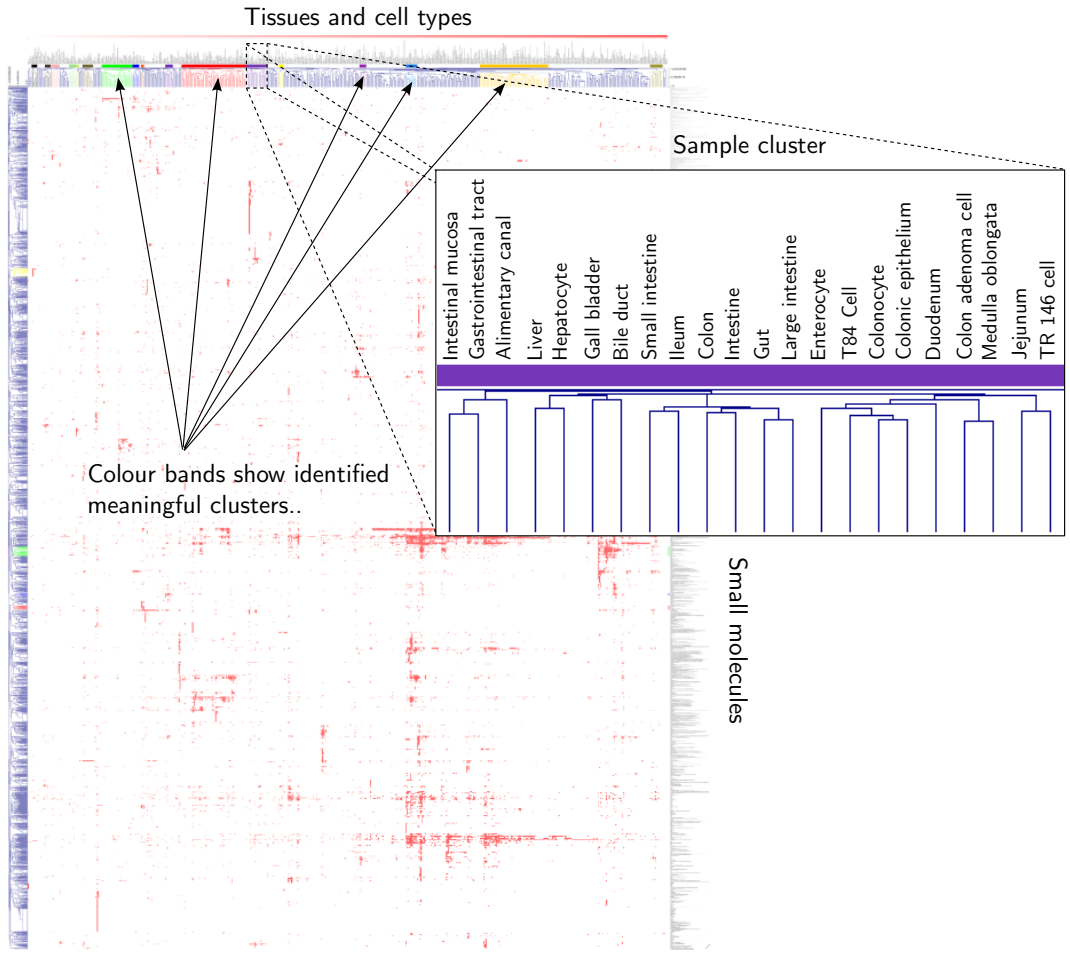



Figure 3.19: Clustering of small molecules and tissues/cell types using their co-occurrences relations. Direct small molecule to tissue co-occurrence relations are informative enough to allow clusterization of approximately half of the tissues/cell types into adequate biological clusters. The largest clusters belong to different cancer cells (71 entries, yellow band), related to reproductive system (67 entries, red band) and nervous system (32 entries, green band). There are approximately 14 additional clusters of tissues/cell types, having on average approximately 9 tissues/cell types each. In contrast, there are only four small clusters of small molecules that have some relation between its participants: two clusters with eicosanoids (25 and 6 molecules), a cluster of iodine-thyronine related molecules (20 molecules), and a cluster of fatty acids (10 molecules). In many cases, these clusters include exceptions, but most of the participants belong to the theme of the cluster.

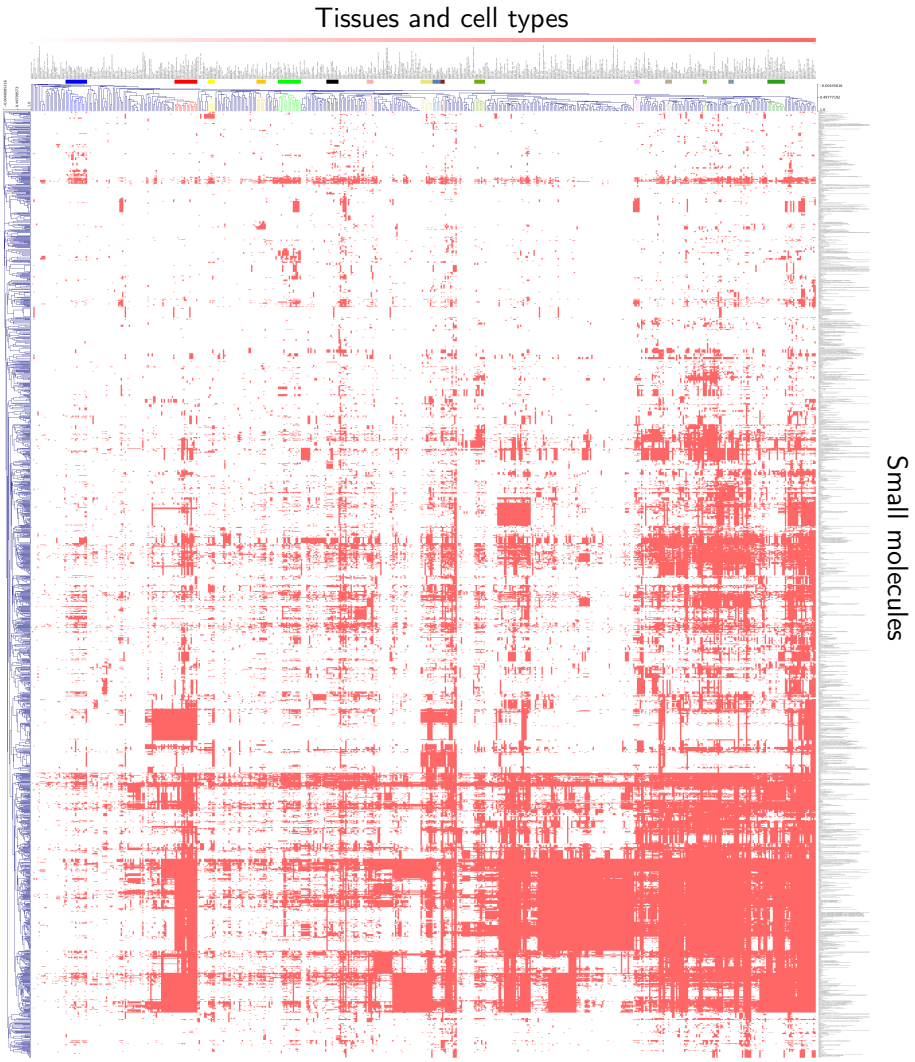



Figure 3.20: Clustering of small molecules and tissues/cell types using protein co-occurrences relations. Protein mediated small molecule to tissue co-occurrence relations are less informative than in the direct case to allow clusterization of the tissues/cell types into adequate biological clusters, as less color bands and of less length can be seen in the tissues clusters, compared to the direct case shown in Figure 3.19. Adding proteins to the method increases the number of relations that can be obtained between small molecules and tissues/cell types. This is shown by a heatmap that has many more correlations (red points), compared to the previous heatmap in Figure 3.19.

## 3.8 Conclusions

I developed a text mining software infrastructure, based on WhatIzIt dictionaries for Proteins (UniProt), Small Molecules (ChEBI and PubChem Compounds), Tissues (BRENDA Tissue Ontology) and Organisms (NCBI Taxonomy). The software retrieves occurrences of these terms from NCBI PubMed abstracts and titles until September 2009. The infrastructure stores occurrences in a relational database designed for this purpose. Information on the entity seen – either a protein, small molecule, tissue/cell type or organism – and its location – article identifier, section and sentence number – are stored on this database. Co-occurrences of these terms are obtained through queries to the database. Nearly 200 million appearances of terms in the literature were captured in the database (which roughly equates to 10 terms annotated per article). The whole process takes nearly a week, where the bottleneck in terms of speed is the insertion of data into the MySQL database.

Dictionaries of small molecules, proteins and tissues show different behaviours in terms of the distribution of their co-occurrences. This is due to a number of factors, as how often they are really mentioned but also how hard can it be to detect each kind of name and how rich are the different dictionaries in terms of synonyms. Some of these dictionaries show as well many promiscuous entries. There is an important need for better dictionaries of small molecules, with wider coverage and better selection of synonyms.

I wrote a pipeline in Java to extract an organism's small molecule collection in the least possible time, including the co-occurrence between small molecules and proteins, tissues and the organism name. The pipeline also captures co-occurrences of protein to tissues and tissues to organisms. A second pipeline annotates the retrieved text mining metabolome with external references and functional categories.

Finally, I used machine learning methods to classify results of the second pipeline as either metabolites and non-metabolites. I selected a confidence cut-off by comparison with the HMDB. I loaded resulting metabolites and their interactions into BioWarehouse.

This effort retrieves approximately 6% of the missing HMDB entries, that

the unification of metabolism databases could not explain. While it still leaves a significant part of the HMDB unexplained, it generates a number of additional candidates, albeit of lesser quality, but that could prove valuable if they are identified by experimentalists in the described organism.

Small molecules retrieved through text mining tend to be much more biased towards external molecules to the organism, such as natural products used as drugs or through dietary intake, more than towards endogenous molecules. This is reflected by the enrichment analysis of the small molecules that are provided by the text mining solution and that are not part of metabolism databases or the HMDB.

The use of text mining towards metabolome inference has still a number of unsolved problems and issues that need to be improved. Both co-occurrence analysis and more sophisticated methods, rely heavily on the use of dictionaries of terms, which in some cases can be very deficient in terms of specificity. It is essential to improve the quality of dictionaries, specially for chemical molecule names and protein names, which show many spurious results. In this work I decided to have as starting point already built dictionaries, as they have been used and published, and it was clear to me through my time researching the field that building them was a laborious and demanding task, which required as well a level of experience with those resources that I did not have back then. However, as I advanced in this work, and confronted the many situations in which the dictionaries produced spurious results, it became more and more clear to me that the biggest improvements in the retrieval of data, currently would come from the improvement of the dictionaries. Especially since this would allow to relax other restrictions in terms of co-occurrence quality that emanate from the poor specificity of the dictionaries.

Another important problem is deciding to which organism a particular abstract/paper belongs to. The inability to decide this produces ambiguity in the biological entities associations to organisms, as many times abstracts mention more than one organism. Here, natural language processing tools would be useful, to either decide by document or sentence, and assign an organism or taxonomic range. While for *H. sapiens* nearly ~70% of the abstracts that mention it have it as the only organism, this number drops dramatically under ~50% for most other

organisms.

The use of proteins to generate more small molecules to tissues relations has interesting particularities. For what the data analyzed showed, the use of proteins allowed to greatly increase the number of relations between small molecules and tissues/cell types. However, this comes at the cost of generalizing the small molecule set. The small molecules set harvested through the use of proteins to molecules and proteins to tissues relations showed a bigger general core region, a shared region of small molecules by most tissues, and smaller, tissue/cell type specific small molecule collections. On the other hand, direct small molecule to tissues co-occurrences produced set of small molecules that seemed to be mostly unique or more relevant to those tissues.

In this chapter we can see once again the enormous relevance of ontologies and classification systems, as vehicles to the understanding and interpretation of the big piles of data that the methods used can generate. There is a strong need for better chemical molecule knowledge organization, that can cover more molecules and hence make this type of analysis more robust.

# Chapter 4

# Constrained chemical enumeration

So far I have introduced methods that rely on the chemistry of existing databases. However, to date, many experimentalist in metabolomics find thousands of peaks that cannot be assigned to known small molecules in any of the databases. While this has a number of reasons, it is also partly because of the many molecules that are yet to be discovered. In this chapter I attempt to find candidates for these unknowns.

The first approach is metabolic neighbourhood. It relies partly on the fact that many enzymes are promiscuous to accept variants of their most commonly known substrates. The idea is taken to the extreme to produce as many small molecules as possible in a few chemical transformation steps, starting from an initial metabolome. The method uses generic reactions known to be present in the reactome and enumerates possible instances based on small molecules known as well to be part of the reactome.

Bacterial secondary metabolism produces a wide range of chemical compounds with a number of interesting properties. Providers of survival advantages in the biochemical warfare, many of these small molecules have a range of applications as antibiotics, anti-tumoral activity, anti-fungals, and herbicides among other uses. An important class of these bacterial secondary metabolites are polyketides, which are synthesized in huge enzymatic factories called polyketide synthethases

(PKS). I explore the problem of PKSs, as they have a very nice "co-linearity" property, in which the order of the domains in the long PKS sequence can be used to predict the structure of the resulting polyketide. I step into a particular type of PKS for which this "co-linearity" rule does not apply, due to evolutionary differences, and work towards contributing to produce a rule set for this type of PKS. This is an example of a complicated case in metabolism, where very particular and non-obvious reaction rules need to be followed in an exact order to actually reach such small molecules. This kind of molecules could be hardly achieved by following the first reaction enumeration process that I propose in this Chapter, as it would require many iterations, which is hard to achieve with the first method.

Generated small molecules by metabolic neighbourhood are compared against what is found through other approaches, as these molecules have an organism specific bias. The other case is shown here as example of more complicated generation scenario which requires a more targeted approach.

## 4.1 Reaction enumeration

### 4.1.1 Enzyme promiscuity

Enzyme promiscuity goes against what is classically taught in courses and text books: that enzymes are mostly specific and can accept only one or a few substrates to catalyze their conversion. Classically, only a small portion of enzymes are acknowledged to catalyze different activities or accept many different substrates, considering promiscuity as an exception to the rule. However experts in the field are claiming that enzyme promiscuity is more a rule than an exception [79; 80].

The concept of enzyme promiscuity refers to the ability of many enzymes to catalyze transformations that differ from those for which they evolved to catalyze. According to [60], enzyme promiscuity can be classified in **substrate promiscuity**, where enzymes can accept a range of substrates without changing the applied transformation, **catalytic promiscuity**, where chemical transformations exerted are different to the native activity, and **contextual or conditional promiscu-**

**ity**, which means that enzymes change their targets and activities in a different environment.

Historically, enzyme promiscuity has been studied from two different perspectives: evolutionary and application. The first approach was the evolutionary one, where enzyme promiscuity was related to the evolutionary path of enzymes. In this field, experts argue that, initially, the primordial enzymes must have been very versatile, allowing to catalyze many different activities with the same machinery. These primordial catalyzers evolved by duplication, mutation and selection, into more constrained catalyzers that at the expense of losing generality became very efficient at particular transformations and were positively selected for them. Research in enzyme evolution also suggests that there is a mechanism operating that supports promiscuity starting from specialized enzymes. A model is proposed in [80] which suggests that specialist enzymes are duplicated, lose the selection pressure for their specialist function, became generalists again through mutations, and then acquire novel catalytic activities for the host organism.

Generally, researchers in evolution tend to neglect the **substrate promiscuity**, classifying it as multi-substrate specificity [80], rather focusing on the **catalytic promiscuity**. Those interested in the enzymatic applications of promiscuity – such as [60] – tend to be more open when it comes to definitions, and have less trouble embracing **substrate promiscuity** and **conditional promiscuity** as feasible categories of enzyme promiscuity.

From a kinetic point of view, there is probably no question about the existence of promiscuity, but rather when do we start accepting an activity or substrate processing in terms of the reaction rates. Normally, what is seen is that native substrates show consumption or turn-over rates that are various orders of magnitudes higher than promiscuously accepted substrates [80]. The net effect is that the flux of those natives reactions will be higher than those relying on alternative substrates, in other words, enzymes prefer some small molecules over others. This does not mean that the promiscuous activities will be totally neglected, specially if the native substrates are scarce, and alternative substrates are abundant. Even though compared to native substrates, promiscuous substrates are much more slowly catalyzed, there is still a huge gain compared to the spontaneous scenario.

I propose methods of assessing the likelihood of the new molecules to actually

exist, although the definitive answer lies on what experimentalist might find to prove or disprove these molecules.

In the reaction enumeration work, I centre mostly in the more conservative view of promiscuity – **substrate promiscuity** – where the reaction mechanism stays the same, but where the enzyme is able to process many different substrates within the same reaction mechanism. Using the reaction mechanism encoded in generic reactions associated to the organisms, and using small molecules known to be present in the organism, I build a system to enumerate new small molecules which result from applying the reaction mechanisms to the known small molecules. This process is iterated, using again the same reaction mechanisms and the new molecules produced in the previous iteration.

### 4.1.2 Existing tools

#### 4.1.2.1 KEGG RPAIR and RDM Patterns

The RPAIR part of KEGG consists of reaction pair mappings through atom-atom mapping of the participants of a reaction. Reaction pairs associate substrate molecules to product molecules where there is at least one atom of the substrate molecule that is present in a product molecule. For instance, Figure 4.1 presents the reaction pairs for reaction ATP:D-hexose 6-phosphotransferase (EC number 2.7.1.1). Using the knowledge represented in the reaction pairs (constant part, reaction center, changing part), different transformations can be applied to molecules that match the relevant regions of the applied reaction pair. This is how KEGG E-zyme[167] works to generate new reaction paths between any two small molecules provided.

For doing the atom-atom mapping, KEGG first assigns atom types to each atom. Atom types reflect the element and the valence state that an atom is currently in, given its bonding circumstances. Atom types are not shown in Figure 4.1 to simplify the diagram.

RDM Patterns are a codification of the reaction pair mappings concerning the atom type[1] changes of the reaction centre (R), of the difference atom (D),

[1]KEGG atom type definitions can be inspected online at the web address *http://www.genome.jp/kegg/reaction/KCF.html*

and for the matched atom (M). Each reaction pair has one RDM pattern (many reaction pairs will have the same RDM pattern). For instance, for reaction pair 1 in Figure 4.1, the RDM pattern is O2c-O1c:P1b-*:P1b-P1b. This represents that the reaction center – oxygen in red – goes from a P-O-P configuration (coded as O2c) to a P-OH configuration (coded as O1c); the difference atom – phosphate in blue – goes from a P-O configuration (coded as P1b) and leaves the pair (coded as *); the matched atom – phosphate in yellow – goes from a P-O configuration and remains in the same configuration. RDM patterns are more general than reactions pairs, they imply less restrictions on what a molecules needs (just the reaction centre and neighbouring atoms) to participate in a reaction.

#### 4.1.2.2 BNICE

BNICE [53] is a reaction enumerator scheme which uses accumulated reaction knowledge from KEGG. Through the use of Ugi matrices to encode reactions, BNICE generalizes reactions according to the EC Number classification, up to the third level. In many occasions, the authors find that reactions classified through EC Number do not fit well in the generalization and create new classes to accommodate them. According to the publication, they summarize the chemical reaction diversity housed in KEGG (for the 2005 version) through less than 250 generalized reactions.

The authors of BNICE use it to obtain new biosynthetic pathways for amino acids biosynthesis. In general, they observe that the natural pathways tend to be more favourable in terms of Gibbs Energies than the inferred pathways (set of reactions produced by BNICE), for the same pairs of initial precursor to amino acids.

The tool is an excellent approach for exploiting the enzyme promiscuity problem towards the generation of novel molecules that could be present in a reactome. Unfortunately, BNICE is not available for use, neither as a web application nor as a downloadable application.

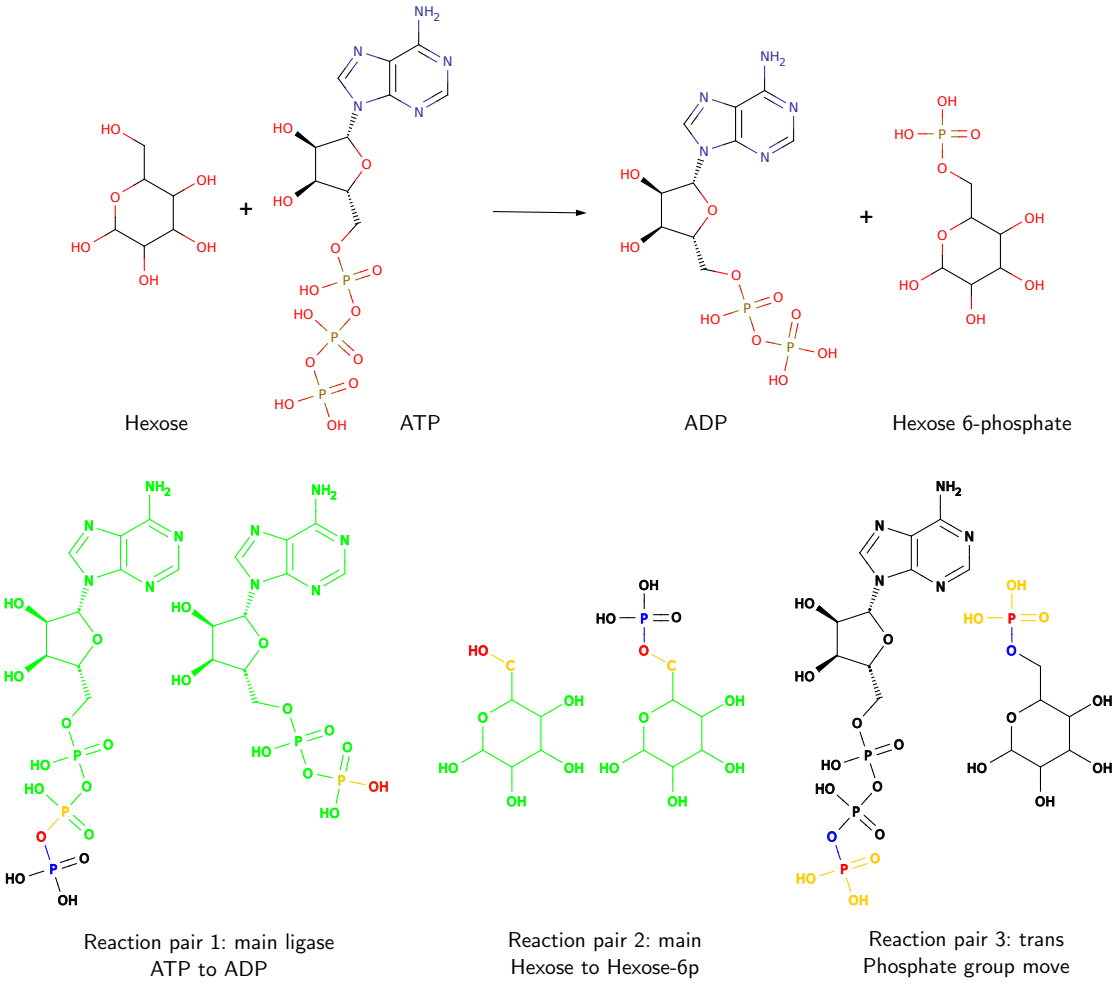



Figure 4.1: Diagram of reaction pairs for reaction ATP:D-hexose 6-phosphotransferase (EC number 2.7.1.1). The reaction can be expressed with three different reaction pairs. The first reaction pair shows the transition from ATP to ADP by removal of the last phosphate group. The red atom – the linking oxygen between second and third phosphate groups – is the reaction centre for this reaction pair. The phosphate in yellow next to it shows the beginning of the constant part (in green) in the reaction pair, while the phosphate in blue shows the beginning of the non-constant region in the reaction pair. The same applies to the other two reaction pairs: red atom is the reaction center, yellow is the first common or matched atom, blue is the first different or unmatched atom. Each of these reaction pairs are used to explain many different reactions.

#### 4.1.2.3 Mariboes

MaRiBoEs[24] is an engine for reaction generalization and enumeration which builds considerably on top of the reaction pair knowledge existing in KEGG. MaRiBoEs improves the codification of reaction rules in RDM patterns introduced by [110] (Figure 4.1 shows an example), by adding a simple stereo chemical check and fingerprint distance check to decide whether a new compound can be modified by a reaction rule encoded in the RDM pattern of a reaction. Furthermore, they extend the RDM pattern concept to include all the unmatched atoms in the reaction pair, and not only those that are directly connected to the reaction centre. Authors use BRENDA to generate their extended RDM patterns, due to the richness of the reaction mechanism deposited there.

The stereo chemistry check proposed consists of rotating the molecules so that the major axes of both molecules are aligned in the x - y plane, and then checking the z coordinates of the atoms to see if they are the same.

MaRiBoEs is implemented in MATLAB, which is a major drawback for using it, as it does not integrate well with the rest of Java based tools and APIs used for other parts of the project. MATLAB is also a commercial package, and hence running MaRiBoEs requires to have a MATLAB license.

#### 4.1.2.4 ChemAxon Metabolizer

The Metabolizer package from ChemAxon is a Java software that relies on JChem library to predict metabolic fate of a given small molecule, given a library of reaction mechanisms. Metabolizer is a commercial software, ChemAxon provides a preview version with a reduced library of reactions. The software presents a GUI that asks the user to submit up to 100 small molecules. Apparently it can calculate a few levels of iterative applications of the reaction mechanisms of the supplied library. It also has some way of assessing the metabolic stability of the generated metabolites. There is no technical documentation available that describes how Metabolizer works.

Metabolizer also offers a command line access and an API access for Java. It is unclear whether Metabolizer could cope with dealing with several hundreds structures and few hundred reactions, as it is required in the application that I

intend to give to it.

#### 4.1.2.5 CDK Enumerator Library

The CDK Enumerator Library is a Java library which relies on the CDK to enumerate generic reactions products for defined substrates. The library can be used also as a standalone application with a graphical user interface that allows the user to load a generic reaction and a small set of small molecules. The CDK Enumerator Library is also exposed as a module in the CDK-Taverna workflow environment.

CDK Enumerator Library heavily relies on the use of substructure searches for detecting repeated features and variable length regions, as described in [153]. While this is fine for a user interface based use, where the user loads only a handful of generic reactions and structures to enumerate, this can be a major issue when running at the whole metabolome level. These substructure comparisons are handled with the CDK isomorphism module.

The CDK Enumerator Library is open source and available to be integrated into other Java packages. It is however not directly suitable to run problems at the complete metabolome level, but being open source, improvements can be made towards this end.

### 4.1.3 Method: Generation of metabolites

Given a set of generic reactions (reactions with at least one generic reactant and one generic product) that belong to an organism or biological containment, and a set of non-generic molecules from that same biological containment, I want to list all possible new metabolites that those reactions can produce when applying them to those metabolites. To apply a generic reaction to a non-generic small molecule, the only requirement is that the non-generic small molecule fits into the template provided by one of the generic reactants of the reaction.

Starting from the CDK Enumerator library [153], I built a pipeline for the enumeration of generic reactions. This pipeline consumes generic reactions, either from a BioWarehouse data set or a provided MDL RXN file, and non-generic molecules, either from a BioWarehouse data set or a provided MDL MOL V2000 file.

Originally, the authors of the CDK Enumerator Library conceived it for a lighter use in which a user would load a generic reaction and a few non-generic molecules into a GUI, and the GUI then shows the enumeration result. For the number of reactions and molecules a user can analyze through a GUI, the design of the CDK Enumerator Library is adequate. However, when hundreds of generic reactions and hundreds of non-generic molecules are to be crossed, the CDK Enumerator is too slow. I introduced a number of improvements, both inside the CDK Enumerator Library and around it, to make it dramatically faster (and feasible) for the task at hand.

#### 4.1.3.1 Use of fingerprints to limit substructure search

The original implementation of the CDK Enumerator Library used the CDK Universal Isomorphism Tester, to check whether the generic molecule template is a substructure of the non-generic molecule. For a few reactions and a small number of molecules, this is reasonable, however, for the scale of the problem proposed this is prohibitive as each substructure search is expensive. Because of this, I introduced a filtering step that would pre-screen the whole collection of non-generic metabolites against the generic elements of the reactions being enumerated, reducing the number of expensive substructure searches.

However, in the CDK there is no Fingerprint method that can be used to decide, given a generic and a non-generic molecule, whether the generic molecule can be a substructure of the non-generic molecule, as all the fingerprint implementations make use of the variable region, assigning it to calculations that would change one or more bits of fingerprint. For this reason, I modified the clique-path based CDK Fingerprint to make it generic molecule friendly, allowing to check whether a generic molecule could be a substructure of a non-generic molecule given. I named this fingerprinter, MarkushAwareFingerprinter. This was used to dramatically reduce the number of substructure searches done.

Given the bits resulting from applying the MarkushAwareFingerprinter to a generic and to a non-generic molecule, we say that the generic molecule is contained in the non-generic molecule if all "1" bits of the generic molecule are turned on in the non-generic molecule. Figure 4.2 illustrates how the fingerprinter

works, and how it is used to rule out structures that cannot be compatible with the generic structure that the pipeline is processed.

#### 4.1.3.2 Use of a better substructure search library

Even though I reduced the number of substructure searches for the enumeration problem, still each time that the fingerprint screening accepts a new molecule for applying a generic reaction, substructure searching is invoked. On top of this, for some cases, this search can be particularly expensive (for instance, very big query molecule with many repeated parts that also appear in the target). Given the number of substructure searches done, the complexity of them, and considering that they are one of the most processor intensive parts of the enumeration, it makes sense to find faster ways of performing it.

Recently the Small Molecule Subgraph Detector (SMSD) [124], an improved library for substructure and isomorphism search, was released. The authors claim that it can be as much as 5 times faster than the CDK original implementation, and even solve some cases that the CDK Universal Isomorphism Tester cannot resolve. In order to make my reaction enumeration faster, I changed the CDK Enumerator Library to use the SMSD substructure search capabilities, instead of the CDK solutions.

The SMSD Substructure search library implements a series of algorithms for the substructure problem, from which the library chooses according to the characteristics of the problem presented. Additionally, the SMSD can rank the atom atom mapping solutions provided as result of the substructure search through different criteria, such as bond energies and stereo chemistry.

#### 4.1.3.3 Better control of structures accepted

The CDK Enumerator Library had no control on whether a non-generic small molecule would have protruding parts, in regions other than the variable markush region, that did not appear on the generic structure. Figure 4.3 shows an example of a generic structure, and two non-generic structures, where one of them fits perfectly with the generic specification and the other is almost identical but for

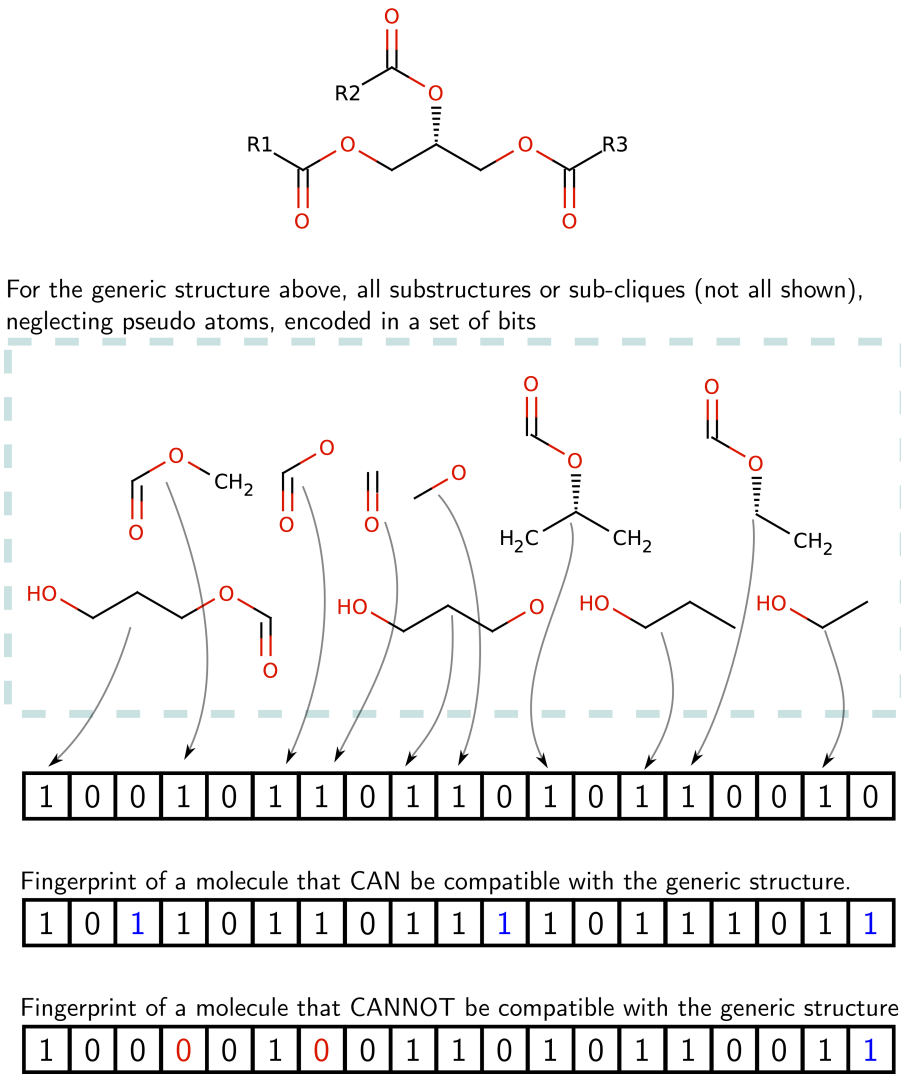



Figure 4.2: Diagram for the markush-aware fingerprinter, which is a modification of the classic CDK Fingerprinter, but that neglects R regions (the R regions are not included in the fingerprint or set of bits). The Fingerprinter produces all the cliques of a given molecule, without adding any R groups found. The fingerprint signals with bits turned on the appearance of certain sub structures. For a new molecule to contain the template generic structure, a necessary but not sufficient condition is that when the same fingerprinter is applied, all the "1's" are conserved, as in the second bit set shown. If there is a "1" that changes to "0", then the proposed non-generic molecule cannot be compatible with the generic structure. If fingerprints are compatible, then a substructure search makes the final deciding comparison.

an additional phosphate group that is not part of the R-group. I implemented additional checks to control this feature, enabling both the enumeration with extra protruding parts or exact matches.

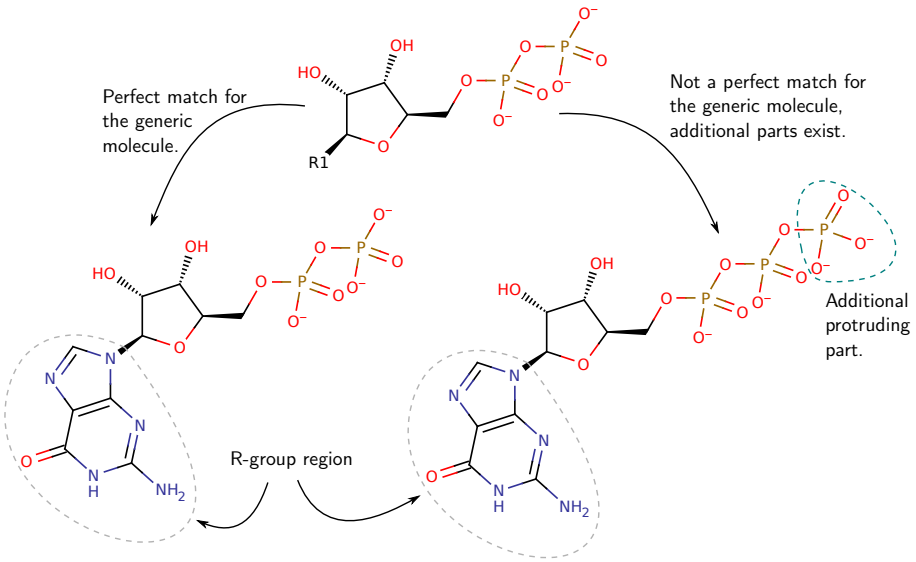



Figure 4.3: Diagram of a generic molecule with a match without any additional groups but those located in the R group, and with a partial match due to a protruding phosphate group in the non-generic molecule that is not present in the generic molecule, nor is part of the R-group region.

#### 4.1.3.4 Mapping of multiple markush regions

In reactions where there is more than a markush region, the CDK Enumerator Library makes the match between different markush regions by the label of the R-group: the library assumes that "R1" in the reactants should be mapped to "R1" in the products, "R2" in the reactants to "R2" in the products and so on. While this might be a reasonable assumption for reactions that are loaded to a GUI and inspected by the user, this is problematic for an automated pipeline, as in many cases this mapping assumption does not hold. For instance, Figure 4.4 shows a generic reaction from HumanCyc where the direct mapping by labels

cannot be used.

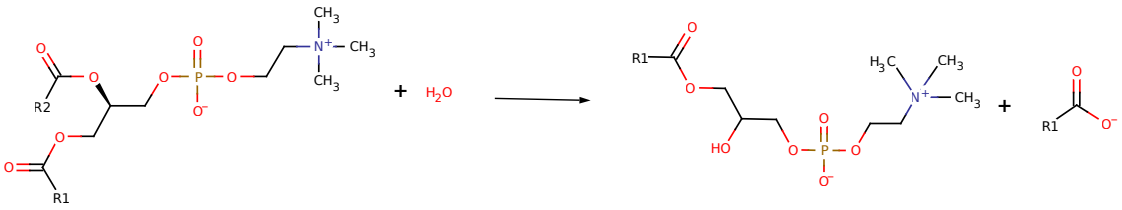


Figure 4.4: Generic reaction for which the generic atom labeling does not allow a correct mapping.

Using chemical signatures [35], I implemented a mapper for generic reactions that solves this problem. Figure 4.5 illustrates the algorithm used – for the same reaction shown in Figure 4.4 – to map the generic regions between reactants and products. For each generic group in the reactants, the mapper calculates the signatures at a minimum height of the non-generic atom attached to the generic group. Signatures for reactants and product atoms are compared, towards finding a unique mapping. If the mapper finds a unique map for an atom attached to a markush group, the mapping is accepted. For all those reactant atoms attached to markush groups for which the mapper cannot find a unique match at the product side, the mapper recalculates their signatures increasing the height by one, the same with the remaining unmapped product atoms attached to markush groups. The mapper compares, keeps the unique matches, and repeats signature comparison for the unmapped increasing the height. It is important to start with low initial heights, because a markush group close to a reaction bond break or creation might not be mapped. Not mapping the generic atoms in a reaction means it cannot be used in the enumeration.

All mappings are stored in the reaction by means of modifying the markush group label to the adequate numbering (R1, R2, R3, etc), so that the CDK Enumerator Library can deal with them adequately. When there are elements that cannot be mapped, the reaction is neglected.

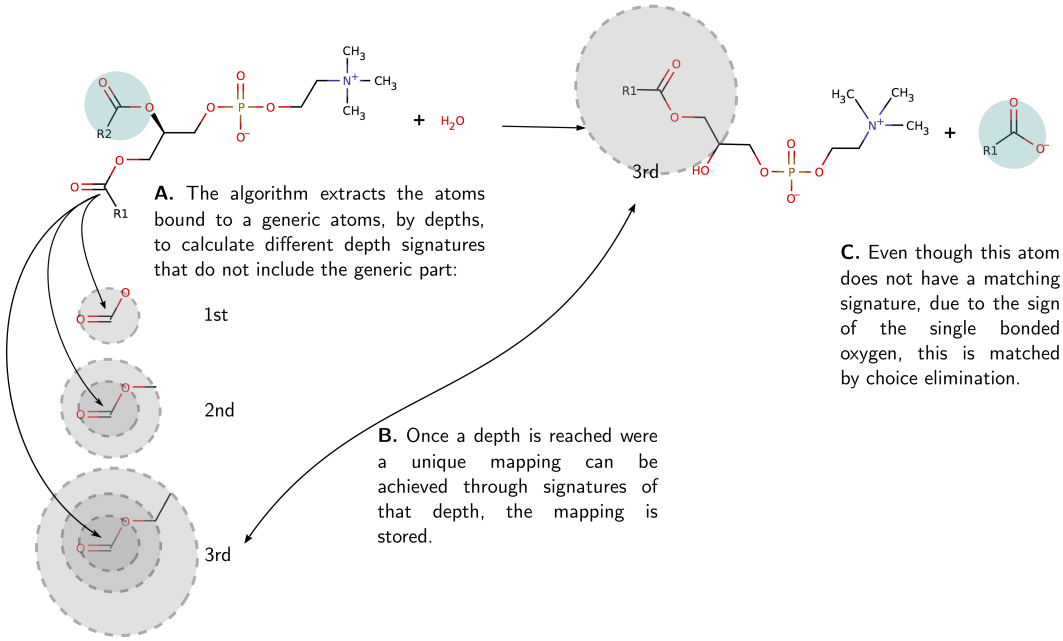



Figure 4.5: Algorithm used to map generic reactions, when more than one R-group is found in the reactants.

#### 4.1.3.5 Multi-processor support and distribution

Given that I introduce a number of steps for filtering results, as following sections show, there are many algorithms that can be executed in parallel. Making use of multi-core processors, I make the whole enumeration process faster by designing the pipeline in a concurrent manner. The enumerator parallelizes the fingerprint calculations and comparisons, the mapping of multiple markush regions, the expensive substructure searches within the main enumerator, and each of the filtering steps detailed later.

However, making the process parallel is not enough to produce results in reasonable time (less than five days per iteration), so I built this pipeline in a distributable fashion. Normally, to distribute the execution of a program it is necessary to split the data. In this case, both the set of generic reactions and non-generic small molecules can be partitioned, so that each different node in the cluster can handle a subset of the problem, and try all the non-generic small molecules on them. Shell scripts are in charge of distributing Java jobs across the

cluster, segmenting the generic reactions and non-generic small molecules. Given that each reaction must get the chance of being applied on each small molecule, the number of distributed jobs is the product of the number small molecules and reactions partitions, for each iteration. As the number of molecules grow with the iterations, this generates an increasing number of separate execution jobs. Figure 4.6 illustrates the distribution process across iterations.

The only case in which reactions and small molecules cannot be partitioned is when a generic reaction contains more than one generic reactant. All these reactions are submitted separately with the complete set of small molecules.

### 4.1.4 Methods: limiting the generated results

Stretching the concept of enzyme promiscuity can generate many spurious results. There is a need to narrow down or at least rank the resulting molecules and reactions, to distinguish those molecules that might have more chances of existing in nature. As part of this pipeline, I introduce three methods that could be used for filtering or ranking reactions and small molecules.

#### 4.1.4.1 Gibbs Energies of Formation and Reaction

The Gibbs Energy of Reaction predicts whether a reaction can proceed in a particular direction. For a reaction

$$aA + bB \rightarrow cC + dD \tag{4.1}$$

the Gibbs Energy of Reaction is calculated from the Gibbs Energy of Formation of each of the participants

$$\Delta_R G^{0'} = \sum_i v_i G^{0'}_{f,i} \tag{4.2}$$

The stoichiometric factor $v_i$ is negative for reactants and positive for products. In [100], a group contribution method was developed to calculate the Standard Gibbs Energy of Formation of any chemical structure, provided it could be completely described by the groups. Later on, [63] further extended the method to

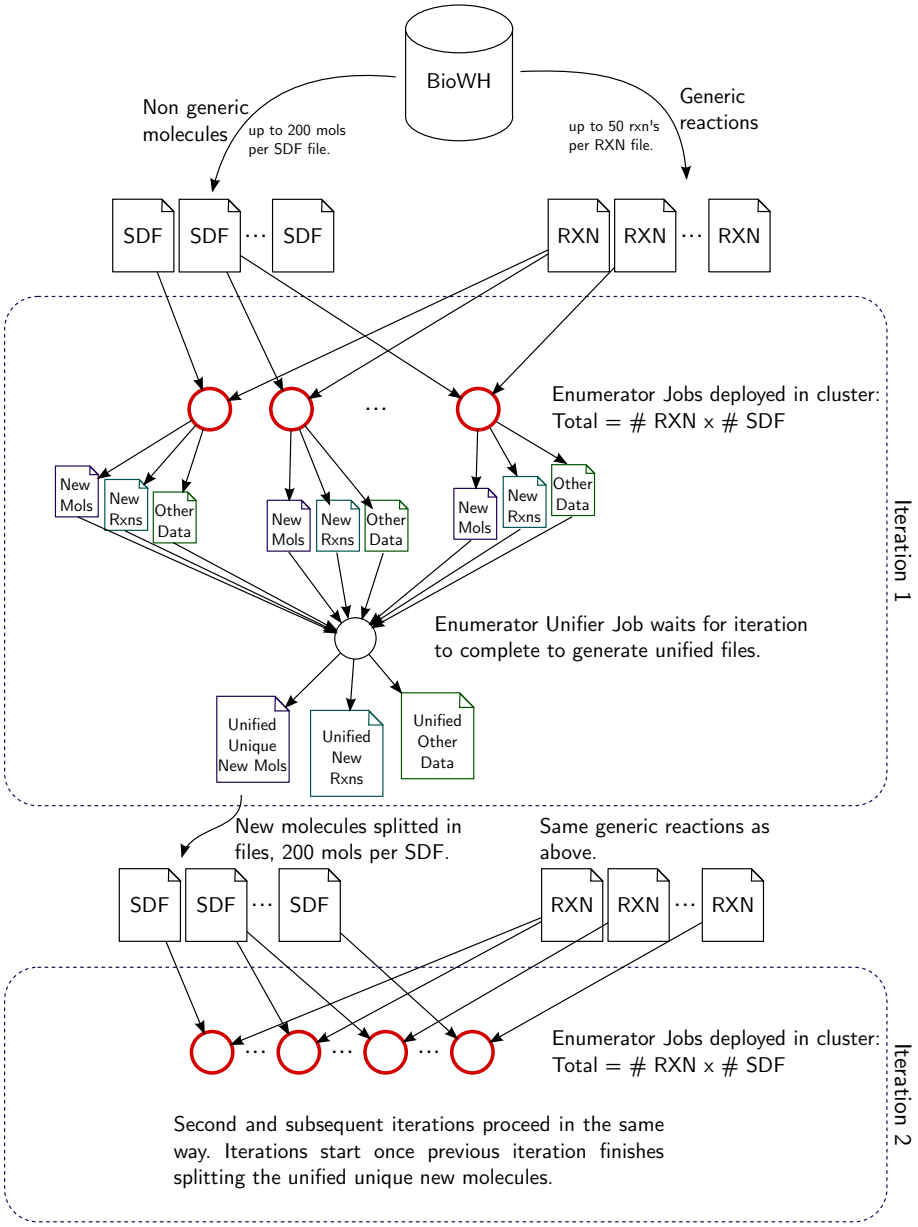



Figure 4.6: Diagram of the distribution of the reaction enumeration jobs across the clusters, by iteration. Initially generic reactions and non-generic small molecules are retrieved from the database, to be split into smaller sets. A shell script generates the number of proper enumerator jobs submissions to the number of molecules and reaction files, for each iteration as files are ready.

include more groups. These works were only possible due to the existence of a plethora of Gibbs Energy measurements for individual compounds, mostly due to the works of [1; 2; 47], among others.

For biochemical reactions, it is important to consider the pH and ionic strength of the solution, as reactions do not occur really at zero ionic strength as it is done in most measurements. The Standard Transformed Gibbs Energy of Formation [1; 2] accounts for these variations. The methods of [100] and [63] do not allow to calculate the Transformed Gibbs Energies. For a biochemical reaction to proceed, the Transformed Gibbs Energy of Reaction needs to be below zero, or, be coupled to other reactions where the net output is a Transformed Gibbs Energy below zero.

Once Gibbs Energies of Formation are calculated, there are certain precautions when calculating Gibbs Energies of Reaction. The reaction needs to be balanced, hopefully avoiding scaling issues, as the stoichiometric coefficient play a role in the reaction energy. The Transformed Gibbs Energy of Reaction considers the protonation equilibrium within the different pseudoisomers, so it is not necessary to balance protons for this type of energy. Gaseous carbon dioxide also needs to be treated as the dissolved species that conform it in solution. None of these considerations are solved by the works of [100] and [63].

Unfortunately, at the time of this work, there was no implementation available of the group contribution method for the calculation of Gibbs Energies of Formation for small molecules. By extension, there was no available software to deal with chemical reactions for calculating Gibbs Energies of Reaction (neither Standard or Standard Transformed).

For this task, I completed and improved work started by Dr. Kai Hartmann at the Cologne University Bioinformatics Center, a Gibbs Energy calculator which was based on the group contribution method of [100]. I extended the method to incorporate the additional data and groups in [63], generated an improved multiple linear regression for the group contribution and completed the ability of the software of converting Gibbs Energies to Transformed Gibbs Energies, among others. Appendix D gives further details of the Gibbs Energy calculator implementation.

The software calculates the Gibbs Energy of Formation of any given molecule

provided that all of the atoms in the molecule can be mapped to one of the groups used. Having calculated the Gibbs Energy of Formation, we devised – with Dr. Hartmann – a mathematical way to calculate the Transformed Gibbs Energy of Formation, relying on $pK_a$ estimations produced by the JChem library. Additionally, I implemented in the package routines for dealing with complete reactions, including any balance or species adjustment necessary to calculate the Standard Transformed Gibbs Energy of a reaction.

#### 4.1.4.2 Known connectivities

For each new molecule generated, the pipeline calculates its InChI Connectivity and compares it against Lucene indexes prepared for ChEBI, HMDB, PubChem Compounds and KEGG, to find matches in these databases. If the enumeration produces a molecule that has known connectivity, then it might be more likely to exist. The pipeline records as result the number of equivalent connectivities found on each database.

#### 4.1.4.3 Downstream transformations

Using the PredictTransform engine – developed by Asad Rahman at Janet Thornton's group at the EBI – I look for small molecules generated that might be further metabolized once produced. Based mostly on KEGG reaction pairs – which discover reaction centers as Figure 4.1 shows – and the SMSD, this tool predicts which reactions could further consume a given metabolite. This prediction is based on structural comparisons and the identification of reaction centers – of the eventual reaction to be applied – in the queried molecule. The engine produces a list of EC numbers of the reactions that are most likely able to metabolize the small molecule presented.

I couple the output of PredictTransform with a look up of the EC numbers produced in a Lucene index that contains EC numbers to organism mappings, to accept only those EC numbers that are known to occur in the organism of interest. I generated this index by parsing the complete UniProt (SwissProt + trEMBL) protein dump file. The rationale here is that if the metabolite can be further processed by other enzymes in the organism, then is more likely to be

real.

### 4.1.5 Method: Main procedure

Starting from all the unique small molecules – in terms of Standard InChI – gathered from the Text mining and database unification produced in the previous chapters, and from a set of generic reactions obtained from the database unification, the engine iterated three times. The first iteration starts from the mentioned reactions and small molecules, the second iteration again uses the same reactions, but on the unique small molecules generated on the first iteration and that are not part of the original set of small molecules. The same process repeats on the third iteration, starting from the molecules generated on the second iteration.

The pipeline separates small molecules into groups of up to 200 molecules, leaving each group in a file. The same occurs with reactions, producing groups of up to 50 reactions, left each group in a file as well. While these files are produced from data in the database, the need for distributing the reaction enumeration in the cluster prohibits the direct extraction of the small molecules and reactions from the database by each distributed job. This is because the database can only accept a limited number of connections, which constrained the number of distributed jobs that could be generated. For this reason, reactions and small molecules are divided into files, which are accessed in parallel by the different instances of the enumerator that run on different nodes of the cluster.

As an iteration ends, a helper application written in Java collects all the results from the finished distributed jobs, and unifies them. This includes identifying molecules that are equivalent – in terms of Standard InChI – and assigning them the same identifier, so that they are not duplicated when running the following iteration. The pipeline neglects molecules for the next iteration if they have the same Standard InChI of a molecule that has been already used in previous iterations.

All results – generated small molecules, the reactions that generate them, and meta data about small molecules and reactions including the results of the processes mentioned in sections 4.1.4.1, 4.1.4.2, and 4.1.4.3 – were analyzed using

| | Molecules | | | Reactions | | |
|---|---|---|---|---|---|---|
| Iteration | Fed | Used | Produced | Fed | Used | Produced |
| 1 | 3,920 | 2,965 | 8,378 | 253 | 44 | 11,568 |
| 2 | 8,378 | 7,962 | 22,881 | 253 | 64 | 46,949 |
| 3 | 22,881 | 21,627 | 73,358 | 253 | 70 | 205,696 |

Table 4.1: Summary of results for the three iterations executed with reaction enumeration pipeline. Fed are all molecules or reactions given to the engine; Used are only those that the engine actually utilized on each iteration; Produced stands for the newly generated molecules or reactions (considering as a new reaction the use of a particular non-generic molecule in a generic reaction). While molecules used and produced are unique, reactions produced are not unique.

R to find potential useful relations between the generated molecules, reactions, and the methods that could help in filtering molecules generated.

### 4.1.6 Reaction enumeration results

Enumeration of small molecules based on known rules can lead to an exponential rate of generation of small molecules. Table 4.1 shows the details of molecules and reactions generated.

Given that a metabolome has normally associated a mass maximum threshold, Figure 4.7 shows how the mass of generated molecules tends to distribute towards higher masses as iterations proceed. Still, most of the generated small molecules within three iterations tend to remain mostly <1,500 Daltons. Box plots in Figure 4.8 display the same effect, indicating where the median, first and third quartiles of each set of molecules are located. Both graphs include only unique molecules generated on each iteration.

Also, as the iterations proceed we see an increase not only of mass of the molecules, but of their complexity. As each iterations generate new molecules, more reactions with different EC numbers are able to match the generated molecules, introducing reaction mechanisms in posterior iterations that were absent in previous iterations. Bar plots in Figure 4.9 point this out: as iterations advance,

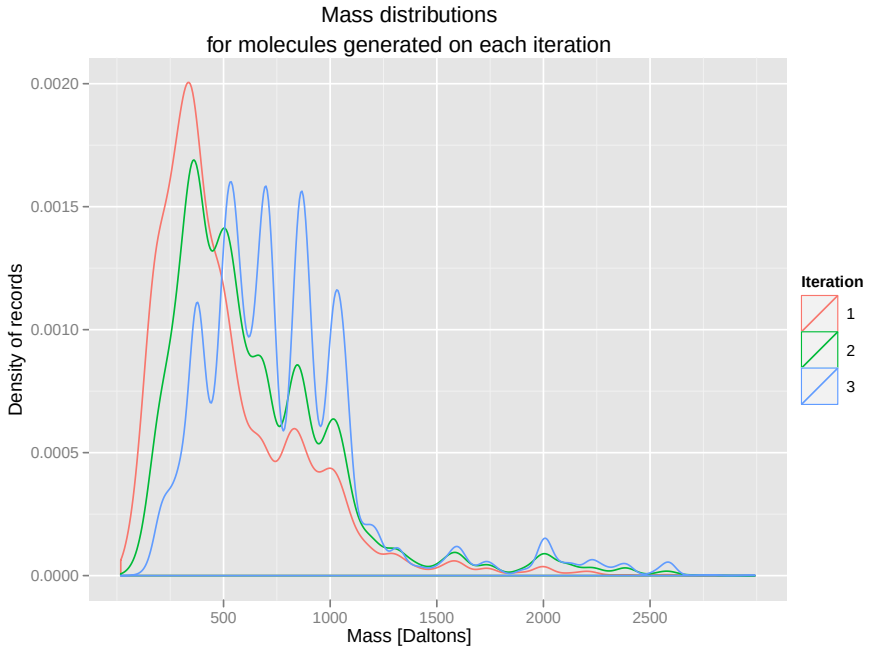



Figure 4.7: Density of the masses for molecules generated during each iteration. As iterations proceed, generated molecules distribute towards heavier masses.

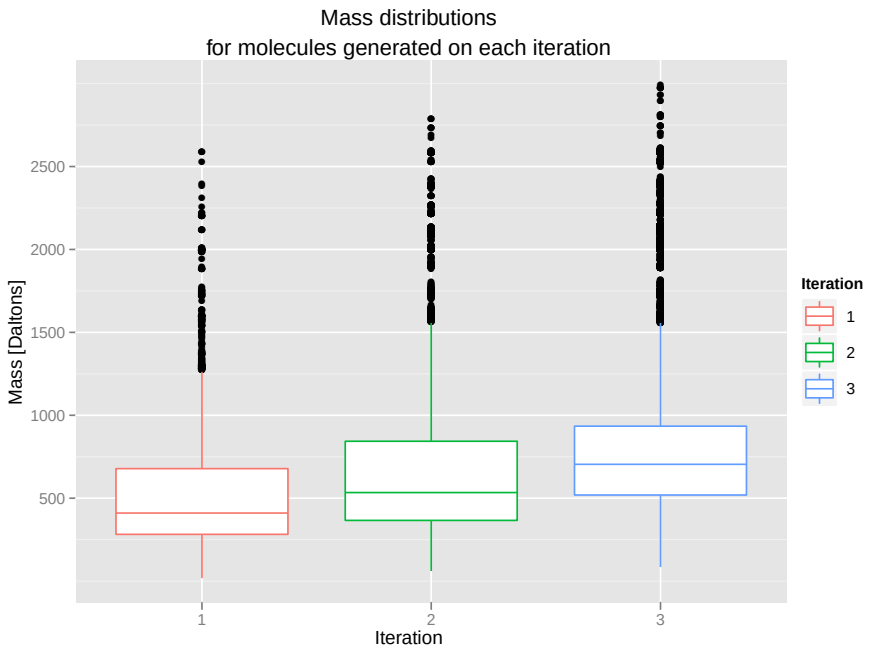



Figure 4.8: Box plots of the masses for molecules of each iteration.

new reaction mechanisms under the categories of transferases (EC Group 2.-.-.-), hydrolases (EC Group 3.-.-.-) and isomerases (EC Group 5.-.-.-) are seen to act on the small molecules of the previous iterations. Interestingly, none of the iterations see the action of lyase mechanisms. It is also notable that while the iterations show a tendency to increase the average mass of molecules that they produce, none of the iterations display more than one different ligase mechanism, which would normally be associated with mass increases, leaving most of this responsibility of mass increases to the present transferases (as both oxidoreductases and hydrolases would not normally generate much heavier molecules).

The increment in molecule complexity also makes the produced structures less similar to known molecules, or to molecules that could be further metabolized by the enzymes in the organism. The bar plot in Figure 4.10 shows that as iterations advance, less and less molecules generated resemble those in ChEBI, PubChem Compounds, KEGG, HMDB, and less molecules are predicted to be target of

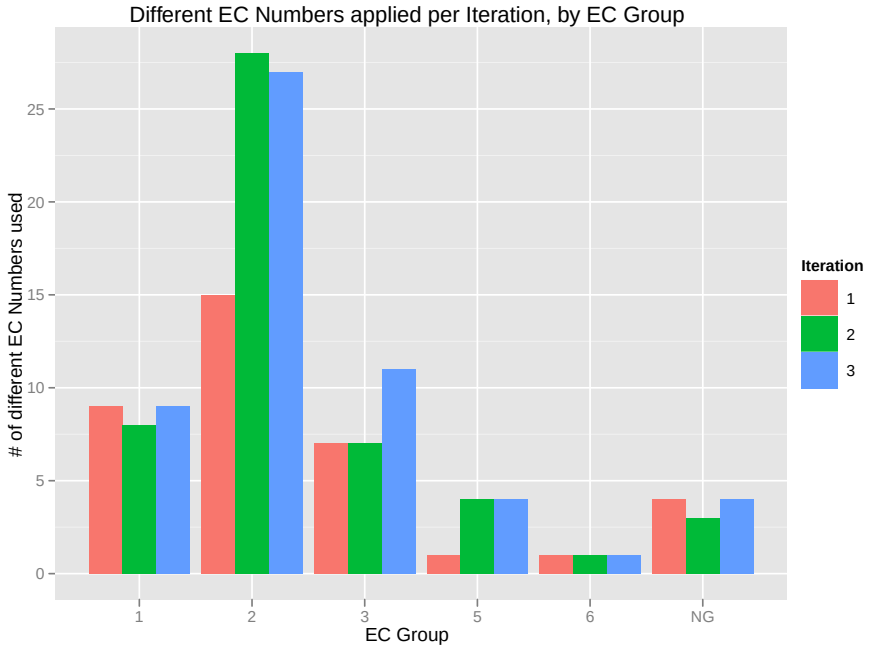



Figure 4.9: Bar plot shows the number of different EC numbers that each iteration uses, grouped by the EC group that the EC numbers belong too. The first EC group corresponds to oxidoreductases, the second group to transferases, the third to hydrolases, the fourth group – absent in the iterations – corresponds to lyases, the fifth group to isomerases, and the sixth group to ligases. NG stands for no EC group. Transferases, hydrolases, and isomerases show relevant increases of unique EC numbers used between iterations, representing an increase in the complexity of the pool of small molecules as the iterations proceed.

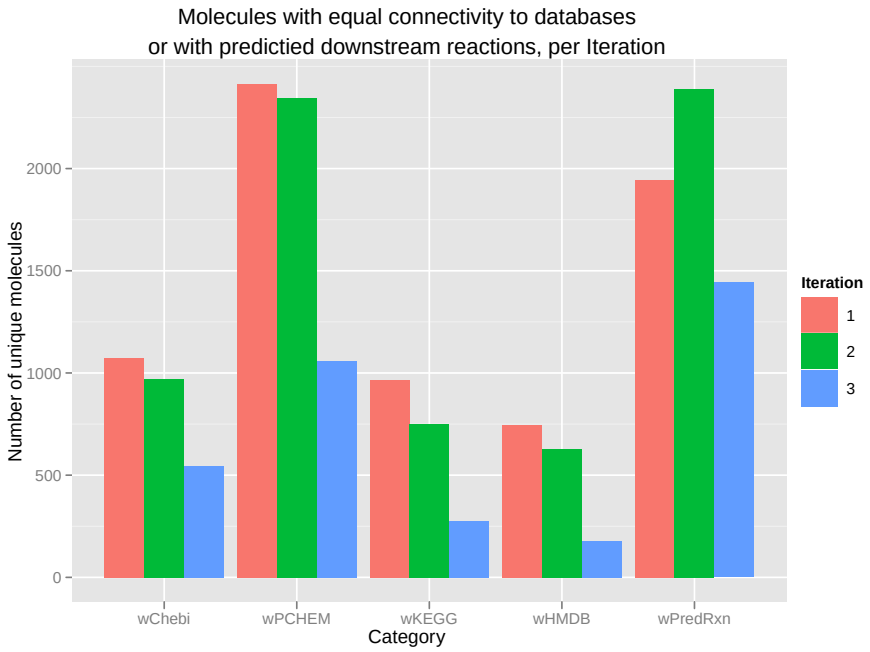



Figure 4.10: Bar plot shows how the number of links to databases – through equal connectivity – diminishes as iterations advance. This also applies to the number of molecules for which downstream reactions could be predicted in the same organism. Even though the number of molecules with predicted downstream reaction seems to increase from Iteration 1 to Iteration 2, normalizing by the total number different molecules generated per iteration would illustrate a proportional decrease from Iteration 1 to Iteration 2.

further processing by enzymes in the system.

Beyond the participation of unique EC numbers, it is important to know how many times each EC number is participating on each iteration. Certain EC numbers with simpler generic reactions probably participate more often than others with more complex requirements in their markush structures. Table 4.2 shows the different EC numbers, and the numbers of times they participate on each iteration, from the higher to lower frequency. Data in the table illustrates that EC number 1.1.1.1 – the oxidoreductase reaction where an alcohol group is transformed to an aldehyde group – is responsible for nearly half of the reactions enumerated on each of the iterations. Reaction EC number 1.1.1.2 closely follows, responsible for ~20% of the reactions enumerated on each iteration. The remaining ~30% of reactions tend to be distributed within the remaining 57 EC numbers seen in the enumeration. This happens because there are a few generic reactions which do not impose a lot of restrictions on potential molecules (the Markush structure is simple), like reactions where "an alcohol" – R-OH – participates, which matches any molecule with an hydroxyl group. On the other hand, many generic reactions use complex Markush structures – such as an $\alpha$-L-fucosyl-(1,2)-$\beta$-D-galactosyl – which impose more restrictions on potential molecules.

As iterations proceed, there is also a phenomenon of decreased Gibbs Energies of Formation, as Figure 4.11 shows. There is normally no correlation between the Gibbs Energies of Formation of a small molecule and its mass, so this must mean that the reactions being used in the iterations tend to form more substructures in the molecules which show low energies. While it is true that Gibbs Energies of Formation are relative terms, having molecules with lower relative energies can imply that the reactions leading to them find favourable equilibriums towards products formation, and would not require much coupling from the thermodynamic point of view.

If one considers the match in connectivity against one of the databases, or the fact that the metabolite could be further processed by other reactions in the same reactome, as evidences of a molecule being real, then reactions that can lead to small molecules with evidence can be compared to those reactions that do

| | Iteration 1 | | Iteration 2 | | Iteration 3 | |
|---|---|---|---|---|---|---|
| EC | # | % | # | % | # | % |
| 1.1.1.1 | 5,819 | 49.2 | 21,618 | 45.9 | 108,623 | 52.8 |
| 1.1.1.2 | 3,040 | 25.7 | 8,967 | 19.1 | 37,256 | 18.1 |
| 1.2.1.3 | 227 | 1.9 | 6,498 | 13.8 | 27,047 | 13.1 |
| 2.4.1.1 | 89 | 0.8 | 1,243 | 2.6 | 7,049 | 3.4 |
| 3.2.1.2 | 87 | 0.7 | 1,048 | 2.2 | 6,617 | 3.2 |
| 2.1.1.49 | 907 | 7.7 | 2,464 | 5.2 | 5,619 | 2.7 |
| 3.2.1.20 | 74 | 0.6 | 848 | 1.8 | 4,940 | 2.4 |
| 1.1.1.21 | 229 | 1.9 | 666 | 1.4 | 2,802 | 1.4 |
| 2.4.1.144 | 75 | 0.6 | 314 | 0.7 | 1,378 | 0.7 |
| 6.2.1.3 | 389 | 3.3 | 457 | 1.0 | 761 | 0.4 |

Table 4.2: Number of occurrences (#) of EC numbers in the reactions of each of the 3 iterations, including the percentage (%) that each EC number covers – in terms of reactions – for each iteration. Few EC classified reactions – most of them oxidoreductases that generate more that ~70% of the enumerated reactions – dominate the enumeration. The whole process is represented by 57 EC numbers, this table shows only the Top 10 in terms of participation (Table G.1 in Appendix G shows all EC numbers). EC numbers are sorted by the number of occurrences (#) in the third iteration.

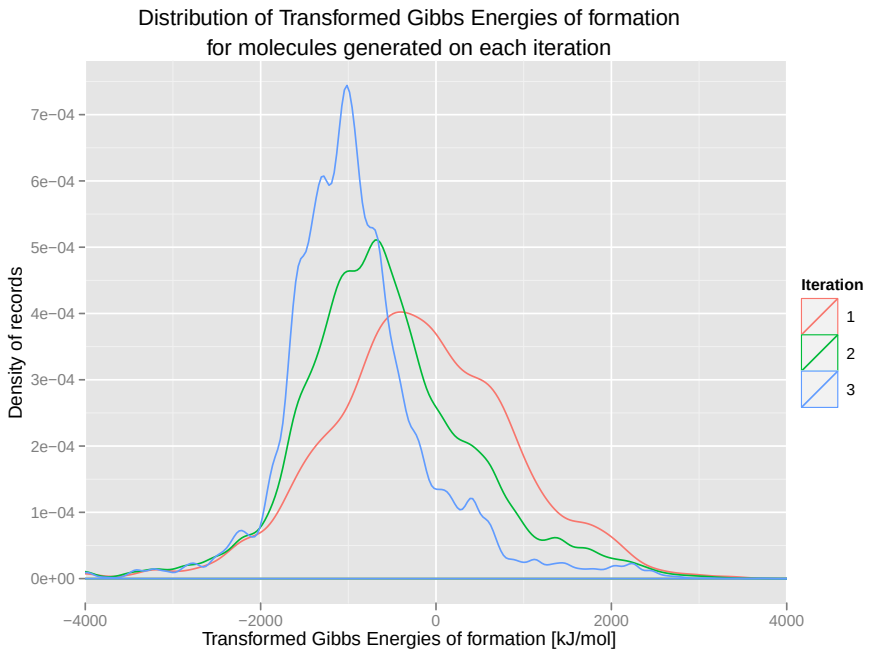



Figure 4.11: Density of transformed Gibbs Energies of Formation for molecules of each iteration. As iterations advance, there is a clear shift towards lower energies.

not lead to small molecules with evidence. Separating by iterations, Figure 4.11 illustrates a displacement towards lower Transformed Gibbs Energies for reactions that lead to small molecules that show evidence of being real. This seems to be consistent between all three iterations.

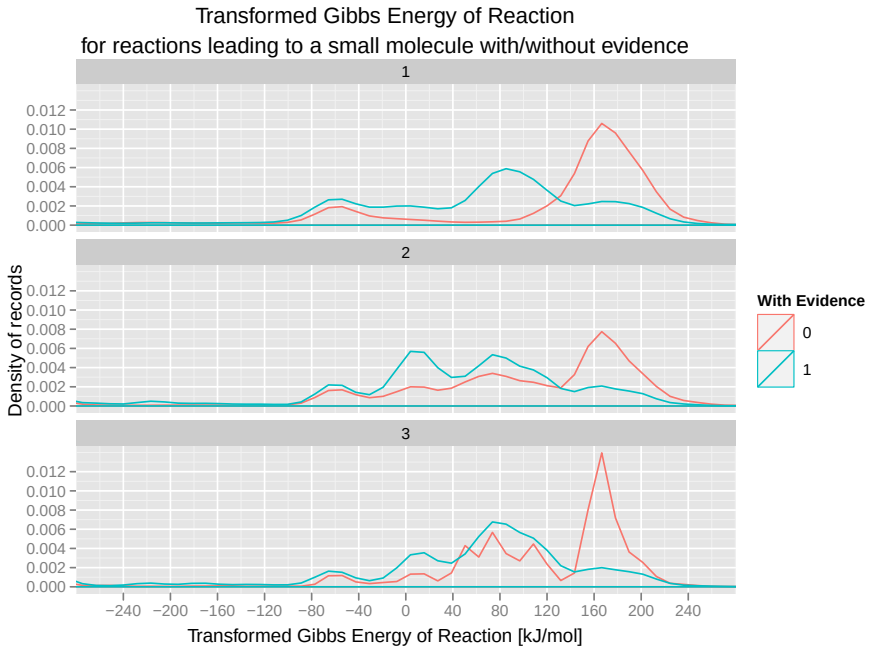



Figure 4.12: Density plot shows the distribution of Transformed Gibbs Energies of Reaction for reactions leading to small molecules with evidence – either a connectivity match with a database or a prediction of downstream transformation – of being real molecules, and without it. Reactions generating molecules that could exist display a slight tendency to be in regions of lower energy, compared to reactions leading to small molecules for which there is no evidence. Plots are divided by iteration.

While it would be tempting to generalize this result into "reactions that produce molecules existing in nature tend to have lower Transformed Gibbs Energies", further inspection shows that this shift might be an artifact of the participating reactions in the enumeration. The consistent peak at ~165 kJ/mol is due mostly to the many reactions enumerated from EC number 1.1.1.1 (82,997 fall in 140 to 200 kJ/mol, overall iterations, 1.9% of them lead to a molecule

that could be real), and then followed by reactions enumerated from EC number 1.1.1.2 (18,008 reactions with between 140 kJ/mol and 200 kJ/mol, 6.3% leading to a possibly real molecule).

Again using the distinction between molecules that show some evidence of potentially being real (database connectivity and downstream transformation), and sorting reactions according to the ranking of participation given in Table 4.2, Figure 4.13 shows the fraction of molecules with evidence that each reaction class has per iteration. As iterations progress, each reaction has fewer chances of producing a molecule that resembles existing molecules. Also, the more reactions are enumerated for an EC number (lower rank number implies that the reaction generated a higher number of enumerated reactions, EC 1.1.1.1 for instance is ranked 1st), the lower the fraction of reactions leading to molecules resembling existing molecules. This slight tendency is gradually lost as iterations progress.

The most important motivation of the application of generic reactions is to produce new molecules that could be considered as part of the metabolome of an organism. To asses this, I compared the generated results of the three iterations to the molecules in the metabolism database unification metabolome, text mining metabolome, and the small molecules present in the HMDB. While the iterations start from molecules in the metabolism database unification and text mining, at each step only the molecules that were not present in the initial step were collected.

Figure 4.14 summarizes these comparisons – using connectivities – through Venn diagrams. In the first diagram to the left, it can be seen that while there is a big overlap of approximately ~900 molecules to both the database unification and text mining collections, the unique overlap with HMDB is much lower. The three overlaps only represent a minor fraction of the total connectivities generated by the enumeration part. The diagram to the right in Figure 4.14 shows that by limiting the enumeration result to molecules with evidence of existing, the total number of different connectivities generated by the enumeration pipeline can be reduced to 6.2% of its size, probably to a set of molecules with higher confidence of belonging to the metabolome. Figure 4.15 gives more detail on how these same overlaps are when splitting the data by iterations, showing that

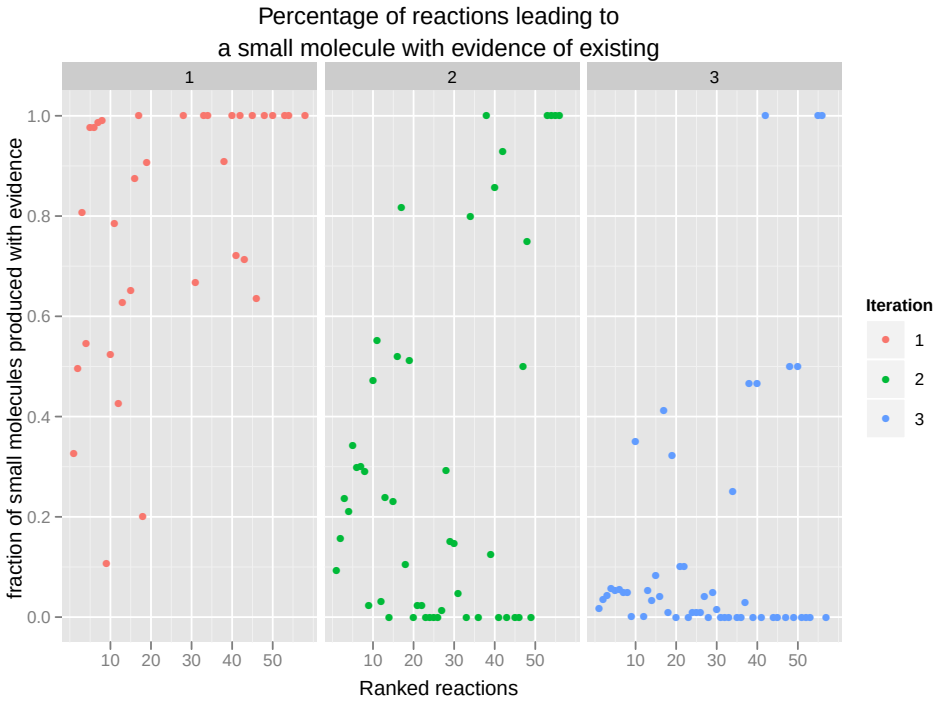



Figure 4.13: Scatter plots display the fraction of small molecules produced by the different reactions – as ranked in Table 4.2 – that show evidence of being real small molecules; either that they have a connectivity match against a database or a predicted downstream transformation. A lower rank number means more reactions enumerated (EC number 1.1.1.1 is ranked 1st).

as iterations proceed, the number of intersections with the previous resources decreases.

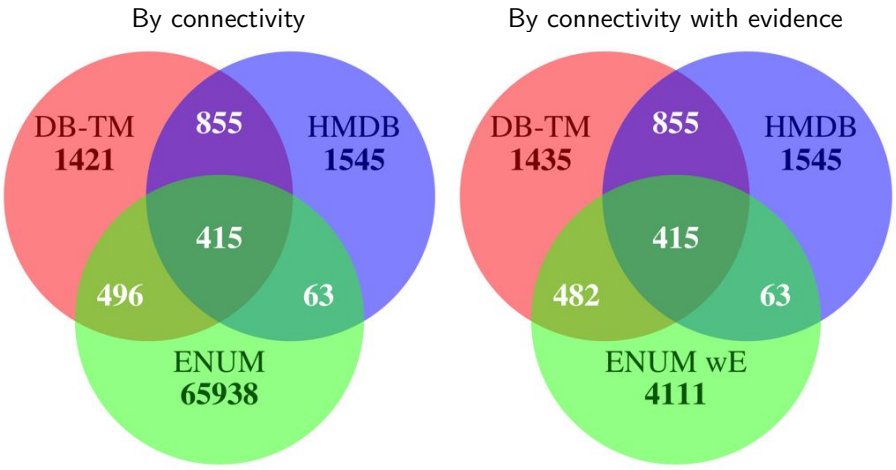
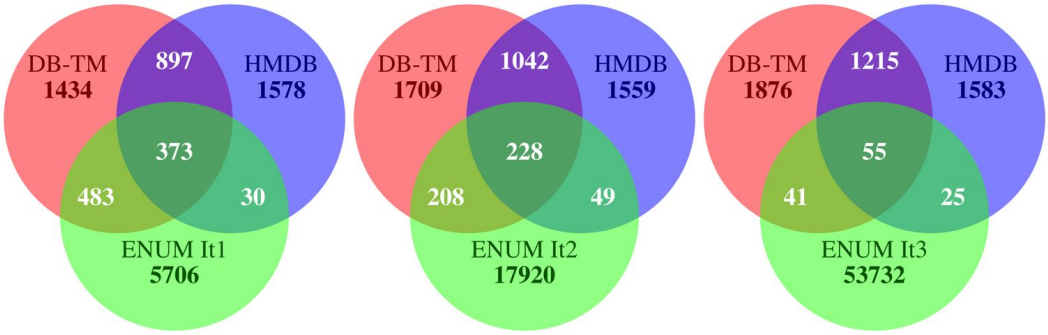



Figure 4.14: Venn diagrams of small molecules produced by the database unification, text mining, HMDB and the reaction enumeration scheme, using connectivities only. *Left:* Diagram built using the complete enumeration result. *Right:* Diagram built using the enumeration result limited to small molecules that show either similar connectivity to known molecules or predicted downstream reactions – which are labeled as "With evidence". Limiting results to these small molecules "with evidence", reduces the set of unique molecules generated by the enumeration to 6.2% of its original size, presumably to molecules with higher chances of being real. Figure G.1 in Appendix G shows the same exercise using Standard InChI instead of connectivities.

### 4.1.7 Reaction enumeration conclusions

Enumeration of reactions to produce new molecules from a starting metabolome is a highly intensive task from a computational point of view. This is mostly due to the high number of substructure comparisons required and the use of PredictTransform. On top of this, while the enumeration possibilities increase heavily at each enumeration step, nature only uses very few of them, mostly constrained probably by enzyme specificity. While there is evidence that enzymes

**A** Joint DB unification-Text mining set, HMDB and molecules resulting from enumeration, separated by iteration.




**B** Joint DB unification-Text mining set, HMDB and molecules resulting from enumeration that show evidence, separated by iteration.

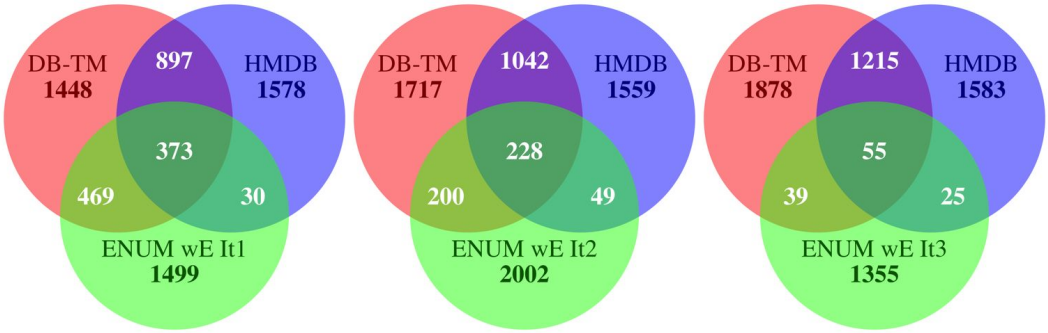



Figure 4.15: *A:* Decomposition of the left Venn diagram in Figure 4.14 by iterations of the enumeration. *B:* Decomposition of the right Venn diagram in Figure 4.14 by iterations of the enumeration. This shows that Iterations 1 and 2 generate most of the intersections with the joint database unification and text mining set, and with the HMDB set, while Iteration 3 produces the higher ratio of unique to intersected molecules. Figure G.2 in Appendix G shows the same exercise using Standard InChI instead of connectivities.

can behave promiscuously, this level of promiscuity must be very low compared to the overall possibilities that the chemical enumeration constrains, as it becomes very difficult to find known molecules through the application of transformation rules on existing molecules. This is more and more clear as the enumeration proceeds through the iterations – departing from the original molecule set – and less and less known connectivities can be obtained.

Partly the lack of verifiable results that the pipeline obtains from enumerating reactions can be due to the choice of transformations made. Limiting only to transformations that are present as generic reactions was a first step towards the problem. This allowed me to focus on the construction of the software infrastructure and on exploring methods for limiting results. This same infrastructure could be used with generalization of reaction mechanisms – encoded as markush structures – within in a metabolic system. This requires an in-depth analysis of reaction's structures. The legacy of this work is a software that is highly parallelized and distributable to face this challenge at the complete metabolome and reactome level.

Better rules are needed for limiting the application of certain reactions that only require minimal substructures to be present, as in the case of the oxido reductases EC numbers 1.1.1.1 and 1.1.1.2, which produce so many enumerated molecules that are not viable and that consume so much of the computing power in posterior iterations. The reduction of these enumerations through the introduction of more advanced, manually introduced, chemical rules could give room for faster iterations and hence deeper analysis.

With all these caveats, the pipeline for enumeration still generate in 3 iterations ~5,000 different chemical connectivities with some evidence that could syndicate them as being real molecules. Out of these set of ~5,000 connectivities, nearly a thousand show equivalents in metabolism databases, the text mining based metabolome and the HMDB overall. This represents a reduction of nearly 20 times from the original size of total different connectivities generated by the enumeration. This set of molecules is feasible to be incorporated in a metabolome database, as a very low confidence data set that offers more alternatives for detected but unidentified molecules.

## 4.2 Polyketides

Polyketides are complex small molecules, produced mainly by secondary metabolism in bacteria and fungi, and have a wide variety of applications. Huge modular enzymes, called polyketide synthases (PKS), assemble polyketides through a series of elongation steps, where malonyl-CoA derivatives are added (but only a C2-unit is incorporated due to decarboxylation), similar in a way to fatty acid synthesis. Examples of well known polyketides are erythromycin or tetracyclines. Polyketides in general have found applications as antibiotics, anti-tumoral agents, anti-fungals, insecticides and growth factors, among others.

*I decided to study polyketides as an example of a complicated case in metabolome enumeration.*

Polyketide synthases (type I, modular) are divided in modules, each of them produces an elongation and modification to the growing backbone of the polyketide [65]. These modules are composed of a set of domains, as Figure 4.16 shows, that characterize the type of modification produced by the module. These domains – valid for modular and iterative PKS – are:

**KS:** Ketosynthase domain, receives the growing chain from the previous module's ACP domain. It then interacts with the ACP domain of its own module through a Claisen condensation in which the polyketide is elongated with the building block (normally acetyl-CoA or malonyl-CoA) bound to the ACP. This reaction leaves the KS domain free and the growing chain attached to the ACP domain of the same module.

**AT:** Acyl-transferase domain, binds the acyl building blocks to be added to the ACP domain.

**KR:** Ketoreductase domain, optional domain which follows to KS domain action, reduces the keto structure left by KS by protonating the double bonded oxygen introduced as part of the previous added acyl module.

**DH:** Dehydratase domain, which is optional, removes an hydroxyl – producing a double bond in the chain – from the building block that the previous module inserts. Normally follows the action of a KR domain in the same module that generates the hydroxyl.

**ER:** Enoylreductase domain – another optional domain which follows the action of a DH domain – removes the double bond introduced in the previous step in the chain.

**MT:** Methyltransferase domain, methylates the linking carbon (-$CH_2$-) between the newer and older groups added.

**ACP:** Acyl-carrier protein domain, always present, holds the growing chain during most of the synthesis process. All other domains act on the chain while this is attached to the ACP domain. Other than the ACP, only the KS domain holds the growing chain at some point. This happens when the previous module's ACP hands over the chain to the following module's KS domain, which in turns transfers it to its own module ACP domain (producing the elongation). The ACP domain, before receiving the growing chain from the KS domain, accepts the building block provided by the AT domain.

The close relation observed by the order of modules, and their domains, with polyketide chemical structures known to be synthesized by some of these enzymes have driven the community to a "co-linearity" rule, which can predict the structure of the polyketide starting from the sequence of the PKS enzyme. This is however mostly valid for a certain type of PKS – type I modular *cis*-AT PKS – and for other related synthases (such as fatty acid synthases).

There are many different types of polyketide synthases. In type II these domains are not in the same open reading frame and act as separate proteins. In contrast, in type I – our main interest – the domains are part of the same protein. Type I can be further divided in iterative and modular; iterative PKS reuse the same module on the same molecule, over and over again; modular PKS – our main interest within type I – have several modules and the growing polyketide chain traverses them in linear order. While type I iterative PKS are known to be present in fungus – responsible of producing molecules such as aflatoxin B1 – and *eubacteria*, type I modular are mostly present in *eubacteria*.

Within type I modular PKS, there are two variants: *cis*-AT PKS and *trans*-AT PKS, a relatively less studied variant than the first one. *trans*-AT PKS are

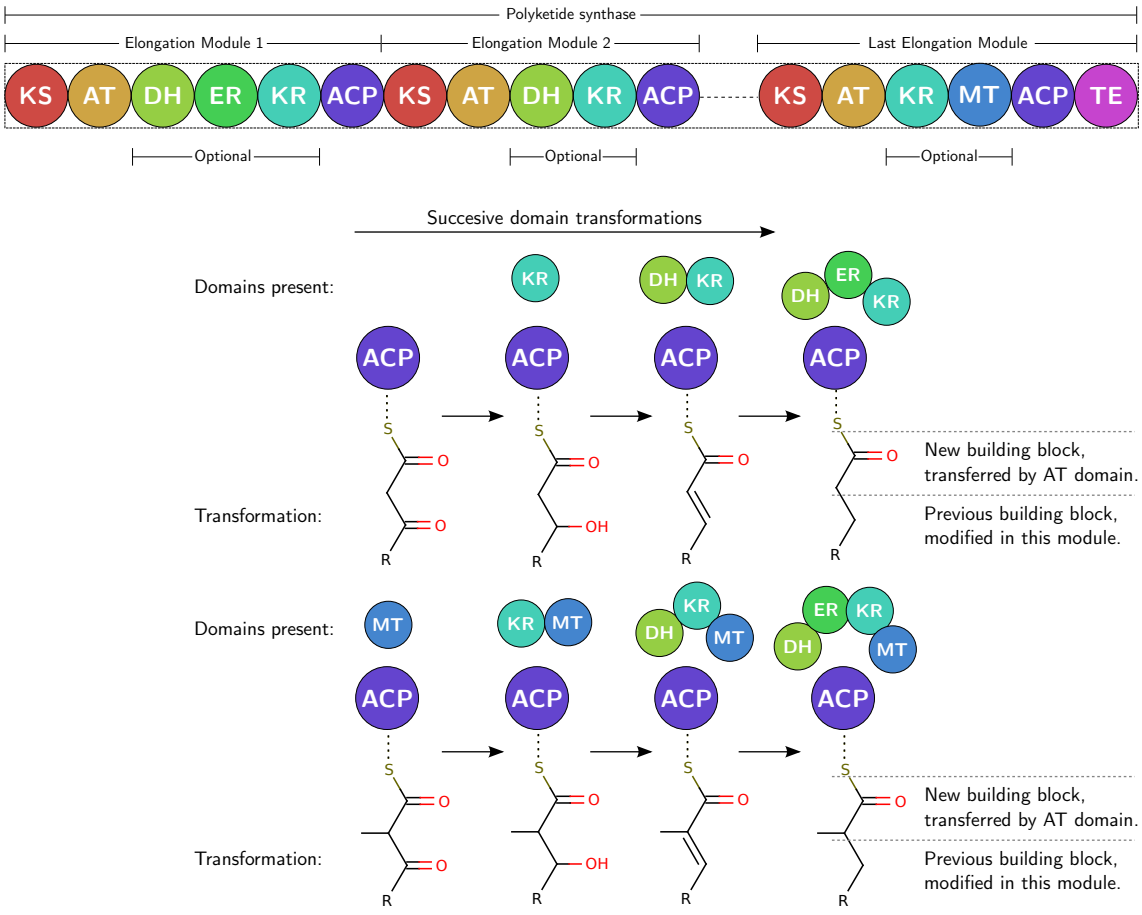



Figure 4.16: General polyketide synthase (PKS) domain architecture. The modular PKS is divided in modules which contain domains, these determine the transformations that the PKS produces on the growing polyketide carbon chain. This figure is an adaptation from Figures found in [65] and [108]. These rules are applicable to *cis*–AT PKS systems mainly.

the main focus of this work.

### 4.2.1 *trans*-AT Polyketide synthases

*trans*-AT PKS enzymes differ from other PKSs in that they do not have acyltransferase (AT) domains within its modules, instead this function – transference of the acyl groups to the ACPs – is performed by external enzymes that have this AT domain [65]. Because it is not performed by the same enzyme *in situ*, it is said to be a transference in *trans*.

While the widely studied *cis*-AT PKS evolved mostly through vertical gene transfer[1] *trans*-AT PKS evolved through horizontal transfer[2] of domains and modules[108] between different organisms. When classified phylogenetically, *cis*-AT PKS domains tend to cluster together by the final polyketide molecule they produce (Figure 4.17 from [128]), while *trans*-AT PKS domains tend to cluster together by the substrate they receive from the upstream module (Figure 4.18 from [108]).

These two parallel evolutionary pathways imply a number of differences between these two classes of type I polyketide synthases. One of the first practical consequences is that the "co-linearity" rule used for *cis*-AT PKS (Figure 4.16) is not directly applicable to *trans*-AT PKS. There are as well other domains that do not appear in *trans*-AT PKS, besides the AT domain, such as the ER domain[108]. Also the *trans*-AT PKS has unique variants of KS domains, such as the KS0 domain, which only makes the chain grow without any modification (KS0 domain transfers the growing chain to the downstream module without modifying it).

In the case of *trans*-AT PKS, as mentioned previously, our collaborators recently showed [108] that the evolution of KS domains is coupled to substrate specificity. Because of this, they implied that to undercover the sequence of substrates used to build the polyketide, it would only require to inspect the KS domains present in the PKS sequence, in contrast to what happens in the *cis*-AT

[1] The most common gene transfer method, by inheritance in reproduction

[2] In horizontal transfer, genetic material is transferred from one organism to another unrelated one through plasmids or other mobile elements, using mechanisms such as transduction, transformation, or bacterial conjugation among others.

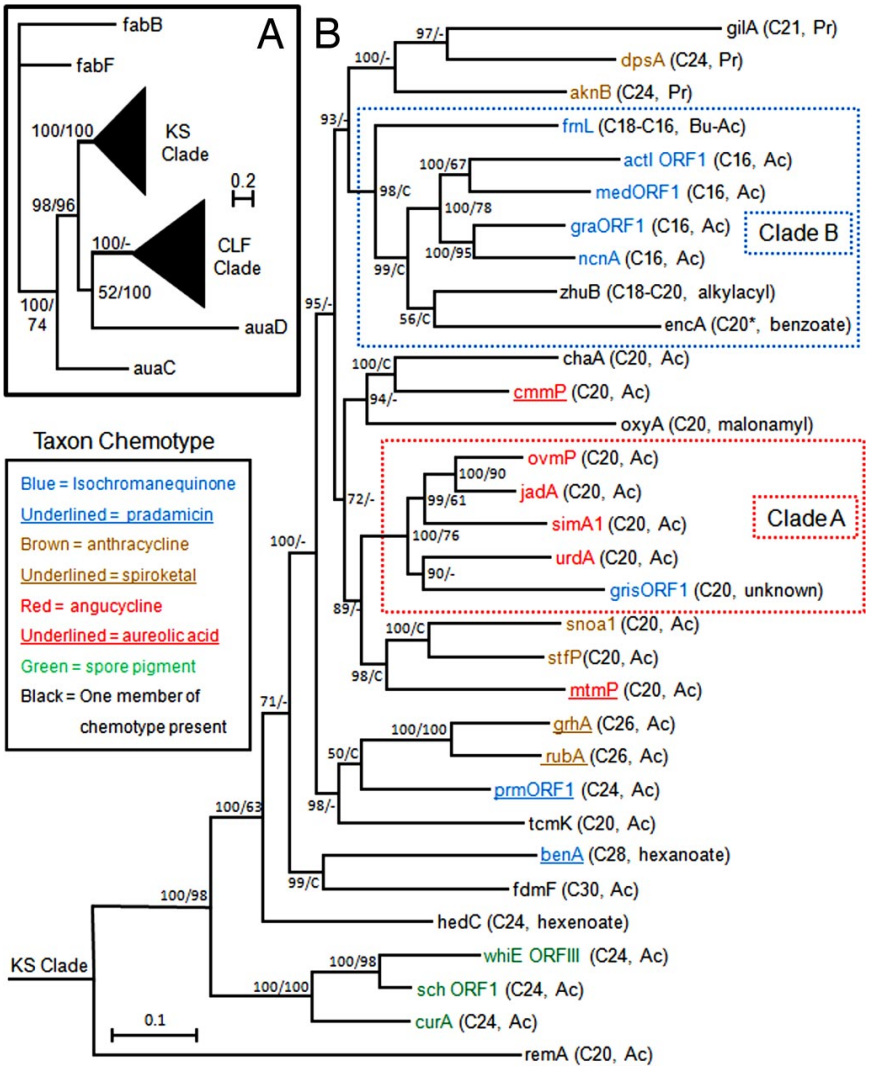



Figure 4.17: Phylogenetic tree for KS domain sequences from *cis*-AT PKS, from [128], used with permission from the author. The cladogram shows that the KS domains cluster by the final polyketide they produce, in contrast to the case of *trans*-AT PKS, where KS domains cluster by the chemical substrate they process (Figure 4.18). The cladogram illustrates the association between polyketide produced and clades with colours and underlines in the Taxon Chemotype box and in the tree. "Copyright (2008) National Academy of Sciences, USA."

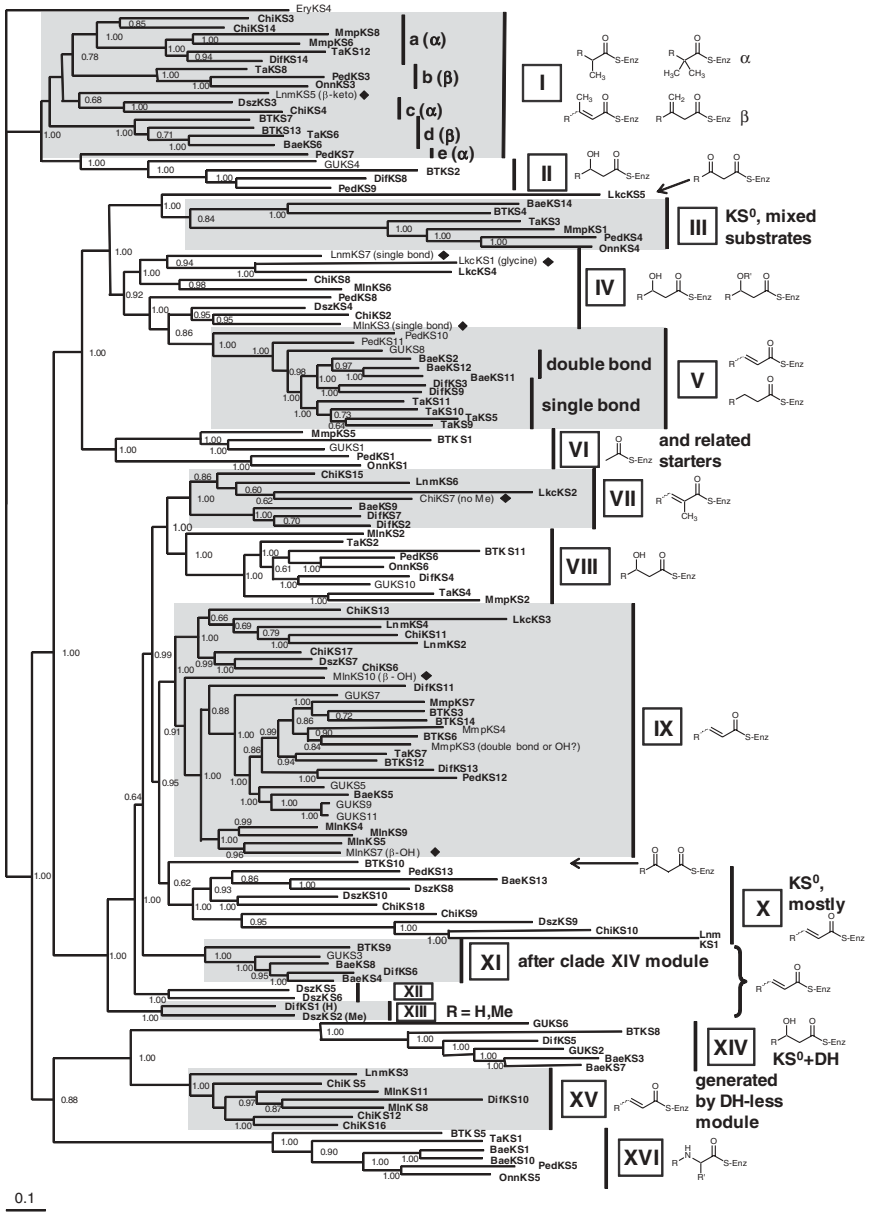



Figure 4.18: Phylogenetic tree for KS domain sequences from *trans*-AT PKS, from [108], used with permission from the publisher. The cladogram shows that the KS domains cluster by the chemical substrate they process, in contrast to the case of *trans*-AT PKS, where KS domains cluster by the final polyketide they produce. Labels with roman numerals represent each of the clades, which in turn are accompanied by the upstream module substrate structures that each KS domain clade accepts.

PKS scenario, which requires reading all the domains of each module. Uncovering the sequence of substrates that the subsequent KS domains accept in the PKS sequence allows to hypothesize the final polyketide generated, much in the same way that the "co-linearity" rule can be used in the case of *cis*-AT PKSs.

### 4.2.2 Marine sponge's endosymbiont's *trans*-AT PKSs

Marine sponges, such as *Theonella swinhoei*, have been noticed to produce a rich array of polyketides[41]. In this project, together with our experimental collaborators, Prof. Joern Piel and his research assistant Eric Helfrich from the University of Bonn, we were interested in unravelling the rules that govern the synthesis of polyketides by the symbiont bacterial[1] flora of marine sponges. We focus on *trans*-AT polyketide synthases.

The Piel-group assembled initially a collection of 138 KS domain sequences of bacterial *trans*-AT PKS [108]. In the case of trans-AT type PKS, the authors showed that these KS domain variants recognize only specific substrates, constraining the transformations that can occur modules upstream in the synthesis of a growing polyketide. The authors subdivided the 138 KS domains in a phylogenetic tree, which clades[2] matched almost entirely with the different chemical specificity for substrates. This means that only knowing the order of KS domains, and neglecting the rest of the participating domains, it is possible to predict the structure of the produced polyketide. Hence, given a new polyketide synthase sequence, recognizing the different types of KS domains, module after module, allows to hypothesize the structure of the resulting molecule.

Later on, the Piel-group assembled a second collection of 423 KS domains from *trans*-AT PKSs, which produced initially 28 KS types, and later through further analysis, more than 40 KS types – or clades in the phylogenetic tree – associated to different substrate specificities. Through my work I mostly used this set of 423 KS domains aligned.

[1] While there is evidence that the main polyketide producers in the marine sponges are *bacteria*, this has not been completely proved; yet it is the most probably scenario, according to our collaborators

[2] A clade is a subgroup of sequences or objects that belong all to a common branch in a phylogenetic tree

Given that it is expected that most polyketide structure are yet to be discovered, specially in the case of marine sponges endosymbionts, one of the first tasks was to aid in the discovery of new PKS sequences in metagenomic samples of marine sponges. This was achieved through the design of degenerate PCR primers, for the amplification of PKS sequences by recognition of unique sequences present in the KS domains, similar to those already collected.

A second sub-project consisted on the identification of key residue positions of the *trans*-AT KS domains, that could be strongly related to the chemical specificity of each KS type defined. This serves two purposes: to aid in the protein engineering modifications of the known sequences, to produce new variant KS domains with desired properties; and generate a simple code – as presented in [15] for non ribosomal protein synthases – that aids researchers in the identification of the KS domain of interest.

### 4.2.3 PCR Amplification of novel PKS sequences

Using the alignment of KS domains, we were interested in producing degenerate primers for nested PCR that would amplify variants of KS domain from bacterial meta genomic samples of marine sponges. Given the huge diversity of species that can be found on this kind of sample, there is no definitive nucleotide sequence for the KS domain. Because of this, we designed primer candidates starting from the protein consensus sequence for each KS domain variant.

A degenerate primer does not have a totally defined sequence, where certain positions of the primer can have different bases. This is normally achieved at the synthesis process, whenever the design has a degenerated position, the chemical synthesis will add any of the specified bases, producing a mixture of different primers which represent the degenerate primer. This has of course a dilution effect, as the signal of amplification is more and more diluted as more degenerate positions are used (because the mixture of primers becomes more and more heterogeneous). Also, as more degenerate primers are used, there are higher chances of unspecific binding. Because of this, the degeneracy needs to be minimized as much as possible. Degenerate primers are normally used in microbial ecology, or where primers are designed starting from an amino acid sequence, as in this case.

After an initial search, I found a few tools that could be eventually suitable for designing the required primers.

CODEHOP[129] is the most widely used program for generation of degenerate PCR primers. It introduced the idea of making a more conserved 5' end of the primers to increase binding chances, accepting most of the degeneracy after half or two thirds of the length of the primer. This tool relies strongly on having a good codon usage table for the organism being pursued (even distinguishing from different probabilities of codons for the same amino acid, which normally requires a reasonable sample of coding sequences). Unfortunately this level of information is not available in our problem (we do not know which bacterial species we are amplifying for). CODEHOP makes no specificity check against negatives through sequence alignments.

DePiCT[163] builds on the work done in CODEHOP and incorporates the clustering process to obtain consensus sequences. DePiCT builds clusters according to the length of the primers required, and hence would modify the clustering previously obtained, which maps very well to chemical specificity. This tool introduces an interesting idea of splitting primers to reduce the degeneracy. DePiCT is implemented in Perl, using BioPerl. A copy of the source-code of DePiCT was not provided on request.

MAD-DPD[107] is a degenerate primer generator which tackles the problem from an optimization point of view. MAD-DPD works on alignments of nucleotides instead of protein, and hence does not deal with the back translation problem, and cannot be used directly in our case. This tool does not consider sequence specificity issues outside of the given source sequences, and makes no thermodynamic considerations either.

HYDEN[91] is another degenerate primer generator tool. As in MAD-DPD, HYDEN authors make a strong mathematical development of the complexity behind the problem, and focus on the coverage provided by the primers. They however neglect important technical aspects such as thermodynamics of the primers and their specificity when compared against negatives. HYDEN works on genomic sequences and not protein sequences as our problem required.

### PKS Primers Methods

Given that none of the tools completely fitted the problem to solve, I implemented a program in Java, which relies on BioJava[56] and other libraries, to produce the degenerate primers starting from consensus KS protein sequences. The following paragraphs explain the algorithm of the program, which Figure 4.20 illustrates as well.

The group of Prof. Piel provided a multiple alignment of 423 keto-acyl synthase (KS) protein domains, representing the phylogenetic classification of them. This alignment was given separated in 28 clades, each of them matching most of the time a defined chemical substrate specificity.

From each clade, using the alignment, I generated a consensus string – which Figure 4.19 illustrates – with support above 70%. This means that the consensus sequence only has an amino acid in a position, if the multiple alignment for the clade showed that same amino acid for more than 70% of the sequences that participated in the alignment. If there was no consensus for a position, a gap is introduced in the consensus string.

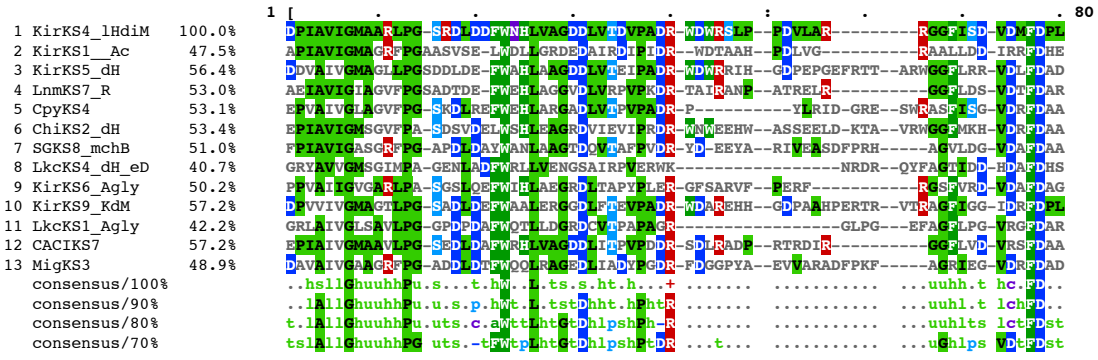



Figure 4.19: Section of a multiple alignment of proteins sequences – belonging to a particular KS domain type – which illustrates the concept of a consensus string. Each column of the protein alignment is inspected, if the same residue is repeated beyond a certain threshold, then the consensus (at that threshold) includes that residue in that position. Consensus sequences for 70 to 100% are shown below the main alignment. Image generated with MView [9]

For each 70% consensus sequence corresponding to each clade, the software

visits the entire sequence through a sliding window of a defined length (between 5 and 6 amino acids), discarding windows with gaps.

Using the standard IUPAC translation table for bacteria (table id 11 according to BLAST nomenclature), for each window with no gaps, the software calculates the degeneracy, which is simply the product of number of codons at each position of the amino acid sequence. If the degeneracy was above a set threshold (for primers of length 5 aa, we accepted degeneracy up to 512; for length 6 aa, we accepted degeneracy up to 600), the window was discarded.

The software then compares each accepted window against a sequence library generated using gapped $> 50\%$ consensus sequences corresponding to each clade, but the one being used. I use the Smith-Waterman [137] local sequence alignment algorithm – through the water implementation in the EMBOSS[127] bioinformatics package – for doing the sequence comparison. The pipeline uses the actual algorithm instead of heuristics like BLAST[4] as both the query sequence (window of at the most 6 amino acids) and the library of sequences (normally a few hundreds amino acid sequences of length ~400 aa) are reasonably small. Using the algorithm guarantees finding the best solution.

If the short amino acid window did not match any of the other clades (identity $> 70\%$), then the software keeps the amino acid sequence, to be used as a clade resolving primer. On the other hand, if the sequence similarity to other clades is above the threshold, then the tool makes a further sequence comparison, but this time against gapped 70% consensus sequences corresponding to each of the other clades. If the amino acid window shows in this case complete sequence similarity to other clades, then the amino acid window is kept as a candidate for a general KS primer (otherwise, discarded). At this point is good to remember that for a specific PCR amplification, only one of the two primers used has to be totally specific, the other one needs to hybridize of course, but can be less specific.

Having visited all 70% consensus sequences (one per clade), the tool has a set of amino acid windows, both specific and general, that can be used for primer generation for each clade. However, not all pairs of amino acid windows are compatible. Melting temperature, specificity, secondary structure and distance restrictions apply.

Distance restrictions are easy to tackle: the two windows separation needs to

be within a range, so that the PCR amplified region is between a minimal and maximal length. Specificity restrictions are also straight forward: at least one of the two amino acid windows needs to be specific for the clade, or, for less specific pairs of windows, the specificity intersect of the two primers has to be only the desired clade.

Melting temperatures for primers are calculated at the nucleotide level, where the actual binding occurs. This means that for each amino acid window, the program computes all degenerate nucleotide back translations. For each nucleotide back translation, the written tool uses MELTING [88] to estimate its $T_m$, or melting temperature. The melting temperature of a probe is defined as the temperature at which half of the probe will be binded to its reverse complementary, and half will be free in solution.

Given that each amino acid window has a set of $T_m$ estimations for each possible back translation, the tool produces a score for each primer pair – for pairs that fulfilled the distance and specificity constraints – which depends on the level of overlap of the distributions of $T_m$. The higher the distribution overlap, the higher the score. I designed an additional score that combined the melting temperature equivalency score, specificity and level of degeneracy of the primer pair, allowing to rank all the primer pairs for each clade.

### PKS Primers results

With the different restrictions for the primer generation agreed with our collaborators (length of the primers, maximum degeneracy, and specificity thresholds), for an alignment of 28 total clades, we managed to generate primer pairs that covered 9 clades.

Using the provided degenerate primer pairs for the different clades of KS domains, my collaborators pursued the identification in the wet lab of new PKS sequences, that contained the previously identified clades, in meta genomic samples of marine sponges.

Previously my collaborators designed the degenerate PCR primers manually. According to their results, the introduction of the automatic method increased in 30% the rate of successful amplifications of KS sequences, specific for the desired

clades. The automatic generation also eliminated completely the non specific (to a clade) amplifications.

### 4.2.4 Identification of relevant residues in KS domains

In [15] researchers defined a simple way of coding domains for a non-ribosomal peptide synthases (NRPS) based on the most relevant amino acid positions. These were amino acid positions that were close to the catalytic centre and seemed to be specific for the function of each of the sub members of the protein family that coded for this NRPS. So, different combinations of residues in those positions map to different types of NRPS.

Inspired in the NRPS work from [15], and given the richness of the multiple alignment of KS domains, we were interested in identifying residues in the different clades that would be responsible for specific substrate specificity. My central hypothesis here was to aim at positions that were highly conserved within the clades, and in the largest number of clades possible, but not conserved in the overall alignment. These should be the residues that explain the particular specificity of each KS domain variant, as they are conserved within the clade (so all known versions of that KS domain have a particular amino acid in that position), but essentially different across diverse clades (so they show what is different between clades).

#### Relevant KS residues methods

The first approach I followed to identify relevant positions was based on information content. The information content for a column $j$ in a sequence alignment, defined as Equation 4.3 shows, reflects how much the occurrences of symbols in the permitted alphabet deviate from expected occurrences in that column $j$. In Equation 4.3, $i$ is the index for the alphabet, which contains $N_a$ letters, $P_{i,j}$ is the frequency of symbol $i$ in column $j$, and $Q_i$ the expected frequency for symbol $i$. A column in the alignment, has a relevant information content if it deviates ($\sum_i P_{i,j} >> \sum_i Q_i$) from the expected occurrences. Normally a column with high information content in an alignment is considered important as it has a high level

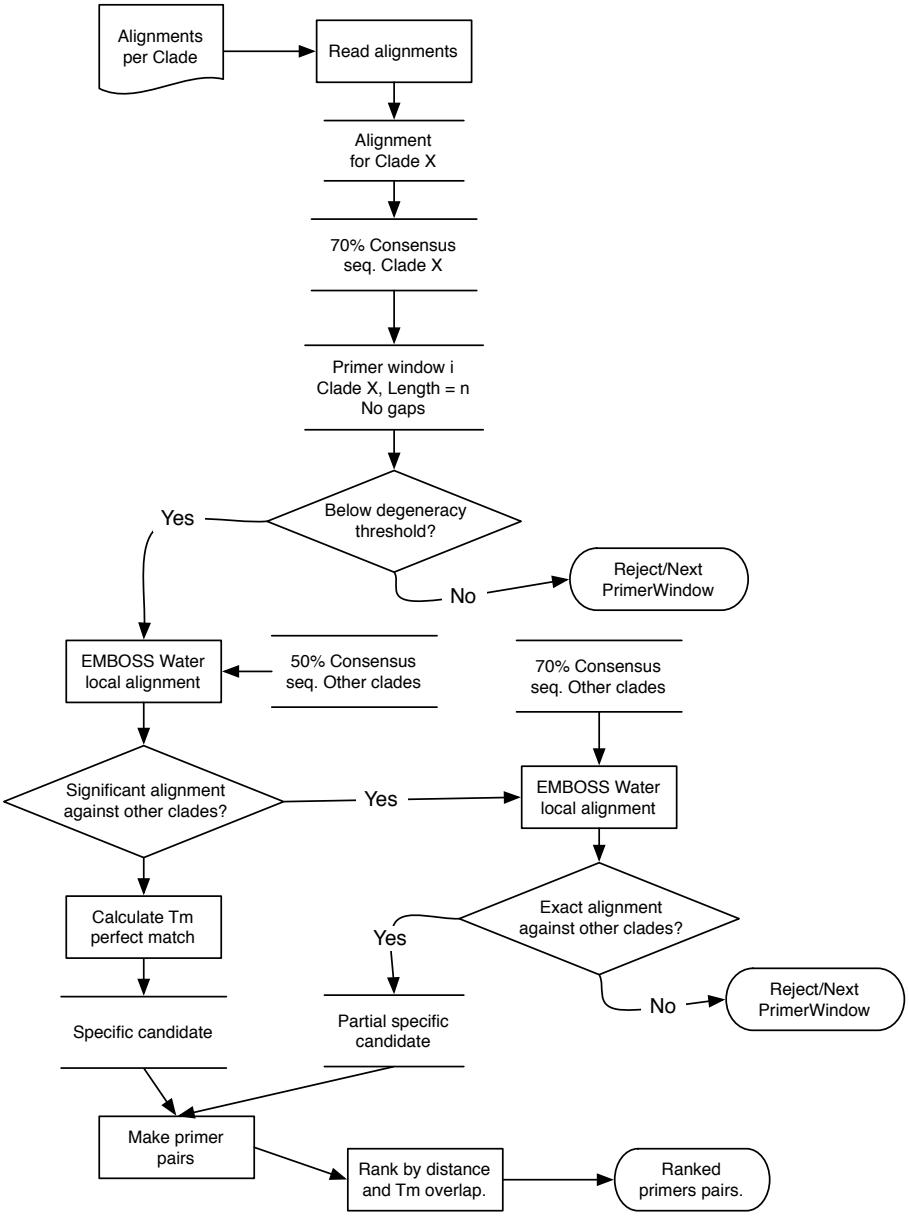



Figure 4.20: Pipeline for primers generation, based on the KS domain variants alignments. Sequences that had exact matches with several clades were paired with specific ones to increase the number of clades resolved. Pairs required minimal distances and were ranked according to the mean and variances of their calculated melting temperatures, as each amino-acid primer has several nucleotide representatives.

of conservation.

$$IC_j = \sum_{i}^{N_a} P_{i,j} log \frac{P_{i,j}}{Q_i} \tag{4.3}$$

I implemented a Python script, using BioPython, to retrieve all positions of the alignment that showed conservation above 70 % in at least 10 clades. For all those positions, information content is calculated for each clade that does not present a gap for the position and for the overall alignment. The top 20 positions of the alignment that maximize the average clade information content over general information content were selected.

However, this method did not pick, among the highest ranked, those positions that were separating the most clades, nor those that were closest to the catalytic centre.

Through an R program, I implemented a second approach consisting on iteratively choosing the residue position (or column in a sequence alignment) $j$ that maximized the diversity of conserved sequences, which I defined as shown in Equation 4.4:

$$D_j = \sum_{i \in R_j} C_{i,j} \sum_{i' \in R_j, i' > i} C_{i',j} \tag{4.4}$$

where $C_{i,j}$ is the number of clades with residue $i$ in position $j$ of the alignment, and $R_j$ is the ordered set of all residues appearing in the alignment in position $j$. I call this method simply Diversity by position.

SDPClust [101] is a published method that aims to identify specificity determining residues from protein families through phylogeny methods. I applied this software to the same data set.

### Relevant KS residues results

I evaluated the performance of the methods in two ways. The first was to asses how good was each method in distinguishing between different clades (resolving clades by finding positions in the alignment that would tell them apart). The less positions of the alignment required to resolve between the highest amount

of clades, the better. Figure 4.21 presents the results for this evaluation for the different methods.

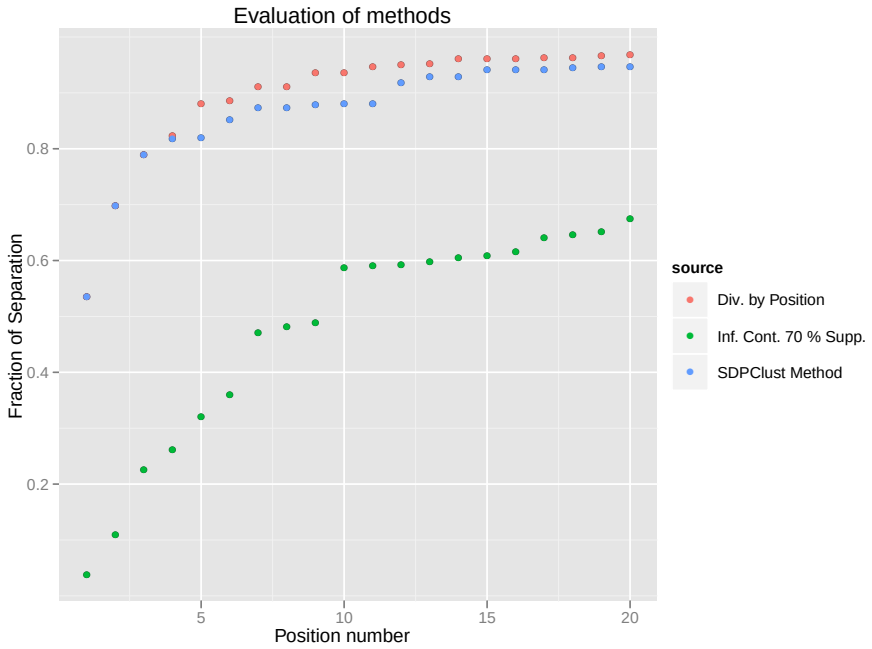



Figure 4.21: Graph of the number of clades that are separated by the first 20 positions picked of each method. The closer the points approach a fraction of separation of one, the better. *Inf. Cont.* stands for the information content method, which uses 70% consensus sequences; *SDPClust* corresponds to the published method. The Position Number refers to the order in which they are drawn by the methods, so Position Number one would be the first position (or column number) of the alignment that each method draws (which can be a different position for each method)

The second validation compared the chosen positions with the expected catalytic centre, the more chosen positions in the catalytic centre, or close to it, the higher score. The catalytic centre region for the KS domains alignments was hypothesized by my collaborators at Bonn – based on previous work by [147; 148] – visualizing the 3D structures deposited for a cis-AT KS domain with a protein structure viewer that allowed the selection of residues that were at a defined distance from a point in the model.

Each of the positions picked by the different methods were compared to the positions of the alignment expected to be within a 10 Angstrom vicinity of the catalytic centre. If the residue was in the 10 Angstrom vicinity, a score of 10 was awarded to the position. If the residue was within a window of variable length in the sequence of a residue known to be in the 10 Angstrom vicinity, a score of 1 was awarded. With this scoring scheme, and variable window length from 1 to 5, I compared the methods. Figure 4.22 displays the results of scoring the different solutions based on the proximity of the chosen residues to the catalytic centre, where we can see that the information content based picks are the worst performing, but that the Division by position method implemented outperforms the published method (SDPClust method).

Currently our collaborators at the University of Bonn are making use of these selected residues for protein engineering tasks. There were however no relevant results at the time of this writing.

### 4.2.5 Annotation of *trans*-AT KS domains through hidden-markov models

With the aim of enlarging our database of *trans*-AT KS domains, and advancing towards a software that predicts a polyketide structure starting from a *trans*-AT PKS gene (or protein) sequence, I proposed a method for finding more *trans*-AT KS domains in existing protein sequences, and be able to assign them to the clades classifications that we have.

Currently, protein domain resources like PFam[122] and InterPro[61] group KS domains in families that are very general. These domain families encompass both *cis* and *trans*-AT type KS domains, and also include within the same domain description all the variability of chemical substrate specificity. In contrast, the clades from the KS domain multiple alignment that I have access to nicely capture chemical substrate specificity. KS domains existing in prime protein resource like PFam and InterPro, cannot characterize new PKS sequences to the level of detail that our clades permit.

To harness this predictive power, the knowledge existing in our KS domain

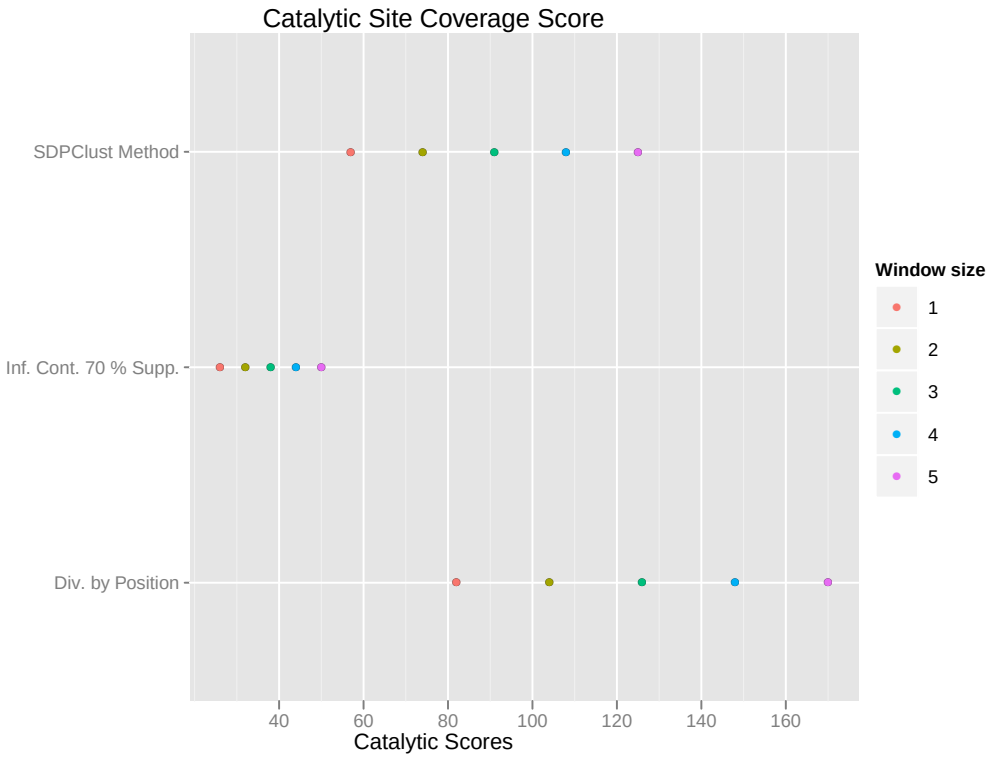



Figure 4.22: Graph shows the scores that the different methods get when compared against the list of residues that were at 10 Angstrom distance of the catalytic centre of the KS domain. Higher scores are better. The window size refers to the maximum distance in amino acids were the method would award score for proximity to the catalytic site. For all windows sizes, the *Diversity by position* outperforms the *SPDClust* method.

alignment needs to be re-encoded in a format that is amenable to sequence similarity. The most widely used way of characterizing protein domains is through the use of hidden Markov models, particularly through the HMMER[42] implementation. Both PFam and InterPro heavily rely on HMMER built models.

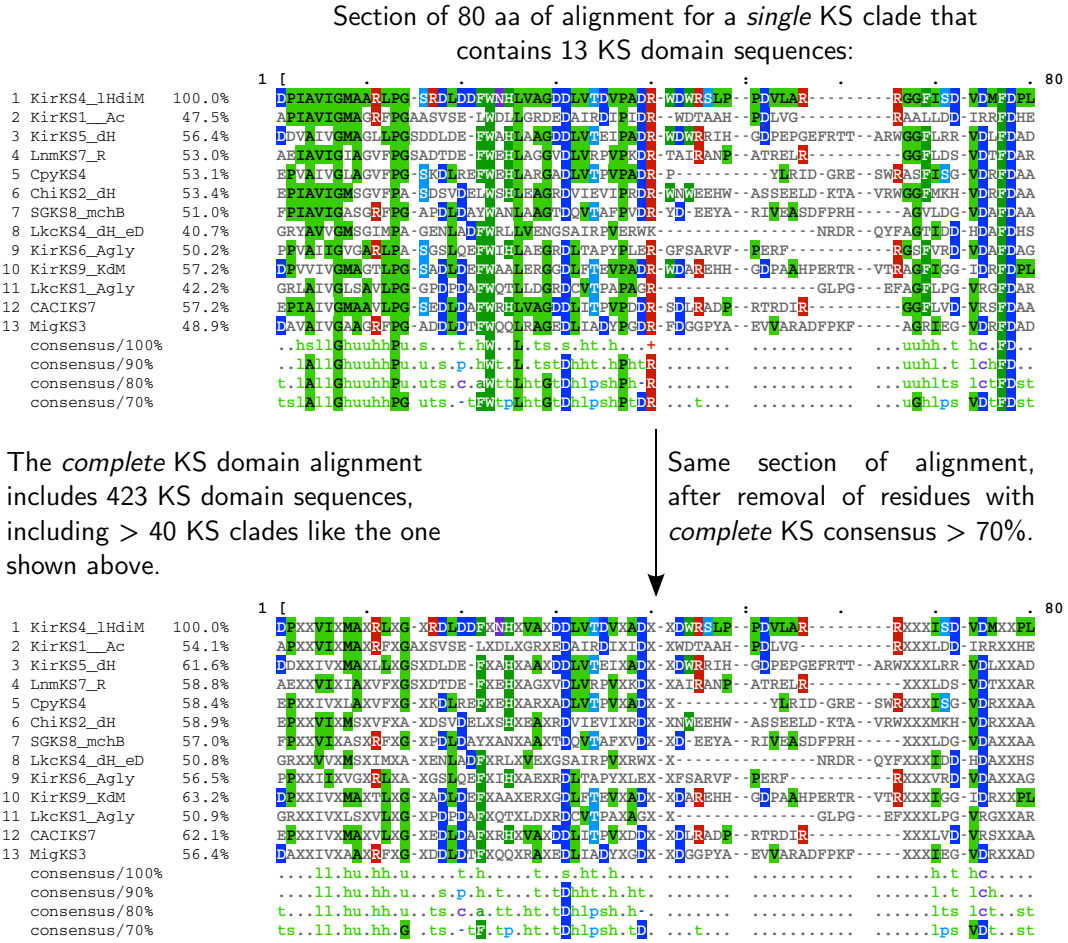
However, as I intend to separate different KS domains by the substrate specificity that they might have, it is important to separate the individual KS clade signal – coming from the amino acid positions that are relevant to each individual KS clade – from the general KS signal, coming from conserved residues that are important for every or most KS clades. Removing this general signal allows the method to score a new domain, increasing the sequence similarity only due to specific residues. Using too many general KS residues in the model might as well produce many results that are linked to either *cis*-AT PKS, non-ribosomal peptide synthases (NRPS), or even fatty acid synthases.

This also allows one to differentiate three types of scenarios for a new KS domain in study. The first scenario would be that the newly detected KS domain can be considered part of the *trans*-AT family and belong to a specific known clade, getting a high score both for the general *trans*-AT KS domain model and for a specific KS clade domain model. A second scenario, where the new sequence has a KS domain that belongs to the *trans*-AT group but forms part of an undiscovered clade, should score high in the general model but low in the different clade models. Finally, a third scenario, where the new domain is a KS domain but does not belong to the *trans*-AT group, with low scores for both the general KS domain model and even lower for every clade model.

### HMM-based annotation of KS domains: Methods

I removed the residues from the individual clade alignments (transformed the symbol into an X) every time that the residue showed a conservation above a certain threshold in the overall or general KS domain alignment (423 KS alignment provided). Figure 4.23 illustrates the reduction of general KS domain signal for a particular KS domain clade, for a section of the KS domain alignment.

To check that removing residues that were well conserved in the overall alignment does not affect the signal of the individual clades, I did a leave-one-out




Positions left (no X's replaced) can have high consensus for the *single* KS domain clade alignment, but they do not exhibit high consensus for the *complete* KS domain alignment. Hence they could be relevant for the specific functionality of the clade.

Figure 4.23: Multiple alignments illustrate the effect of removing residues from a KS clade domain alignment, for positions of the alignment that have a general or complete KS domain sequence consensus above a certain threshold

cross validation with the hidden Markov models of each clade (having removed the general KS signal).

To choose the level of support of the general alignment that should be used as cut-off to remove general signal, I inspected the distribution of sensitivities when the cut-off for conservation changes from 40% to 90%.

Using the HMM models with the reduced general KS signal, I wrote a program in Python, named KSDomainPredictor, which receives a sequence – either nucleotide or protein – and uses HMMER to search these sequences against these models. The program first uses the general HMM model for KS domains (an HMM model built from the alignment of the 423 KS domains sequences) to define regions in the submitted sequence that should be KS domains. Then – using the reduced signal models – assigns each of these regions to the particular *trans*-AT KS domain with the best HMMER alignment. KSDomainPredictor produces a GenBank file with the submitted sequence annotated with the assigned KS domains.

To define cut-offs for the E-value of each of the HMM models to be used by KSDomainPredictor, I compared the results of aligning the built models against known *trans*-AT PKS sequences and *cis*-AT PKS sequences, for each KS domain HMM model. To this end, I selected sequences from UniProt which contained general KS domains, and either contained (for *cis*-AT PKS) or not contained (for *trans*-AT PKS) AT domains. I excluded sequences with other domains associated to fatty acid synthases and non-ribosomal peptide synthases. I selected all these domains from the InterPro database; Table 4.3 details these domains. Additionally, the query restricted results by taxonomy to *bacteria* – TAX ID 2 from NCBI Taxonomy – and to sequences of at least 1,000 residues. These queries yielded 815 putative UniProt *trans*-AT PKS sequences and 3,577 putative UniProt *cis*-AT PKS sequences[1]. These sequences are the complete protein sequences, and not just the KS domains.

I built BLAST databases from the two sets of sequences retrieved. Through BlastP I compared the 423 KS domain sequences to the 815 putative *trans*-AT PKS sequences. I selected all those PKS sequences that did not have 100%

[1]These numbers are correct for July 2012

| | | | Query | |
|---|---|---|---|---|
| InterPro ID | Name | Type | *trans* | *cis* |
| IPR020841 | Polyketide synthase, beta-ketoacyl synthase domain | KS | Req. | Req. |
| IPR018201 | Beta-ketoacyl synthase, active site | KS | Req. | Req. |
| IPR014030 | Beta-ketoacyl synthase, N-terminal | KS | Req. | Req. |
| IPR014031 | Beta-ketoacyl synthase, C-terminal | KS | Req. | Req. |
| IPR016038 | Thiolase-like, subgroup | KS | Req. | Req. |
| IPR018201 | Beta-ketoacyl synthase, active site | KS | Req. | Req. |
| IPR016035 | Acyl transferase/acyl hydrolase/lysophospholipase | AT | Avoid | Req. |
| IPR020801 | Polyketide synthase, acyl transferase domain | AT | Avoid | Req. |
| IPR001227 | Acyl transferase domain | AT | Avoid | Req. |
| IPR014043 | Acyl transferase | AT | Avoid | Req. |
| IPR016181 | Acyl-CoA N-acyltransferase | AT | Avoid | Req. |
| IPR013114 | Beta-hydroxyacyl (acyl-carrier protein) dehydratase, FabA/FabZ | FA S | Avoid | Avoid |
| IPR013624 | Non-ribosomal peptide synthetase | NRPS | Avoid | Avoid |

Table 4.3: List of InterPro domains that constrained queries for UniProt putative *trans*-AT PKS and UniProt putative *cis*-AT PKS sequences. All the KS domains are included in an *OR* statement, so that at least one had to appear, the same for AT domains in the *cis* case. "Req." stands for Required; "FA S" for Fatty Acid synthesis; "NRPS" for Non-ribosomal peptide sequence.

similarity to the provided KS domains, to avoid validating with the same PKS sequences from where the provided domains came from. This search left 604 putative *trans*-AT PKS sequences for positive validation; named this set UniProt *trans*-AT PKS selection.

Also from my collaborators, I obtained a list of 587 KS domains belonging to *cis*-AT PKSs. BlastP of these 587 KS domains produced zero hits with 100% sequence similarity to the 815 putative *trans*-AT PKS sequences (actually, only one of the domains had more than 75% sequence similarity to one of the 815 PKS proteins). BlastP of these 587 KS domains aligned with 100% sequence similarity to 520 of the 3,577 putative UniProt *cis*-AT PKS sequences, which were selected as a negative validation set; named UniProt *cis*-AT PKS selection.

### HMM-based annotation of KS domains: Results

The leave-one-out cross validation analysis shows through Figure 4.24 that there was no major loss of sensitivity for most of the 45 KS domain clades when removing the general KS signal. The HMM models could still detect successfully most other members of its own clade, even when highly conserved residues at the general KS level were removed.

Regarding the threshold for removing general KS signal, there is a gain in moving the threshold from 40% to ~70%, but then there is no major gain in sensitivity when moving the consensus threshold from 70% to 90%. Graphs in Figure 4.25 show this, both for the average sensitivities (over the different clades), which tend to be always very high, but also for the third quartile, the lower whisker ($Q3 + 1.5IQR$) and lower outliers. So 70% seems a reasonable cut-off, which increases specificity (as the models rely less in residues that are probably conserved across all of them) and does not loose too much sensibility compared to the cases where nearly no residues are removed (90%).

KSDomainPredictor executed on both set of PKS proteins: UniProt *trans*-AT PKS selection and UniProt *cis*-AT PKS selection. Figure 4.26 shows the E-value distribution of the *trans*-AT KS domains assignment through HMM models for both sets. The distribution of the exponent of the E-values for most of the

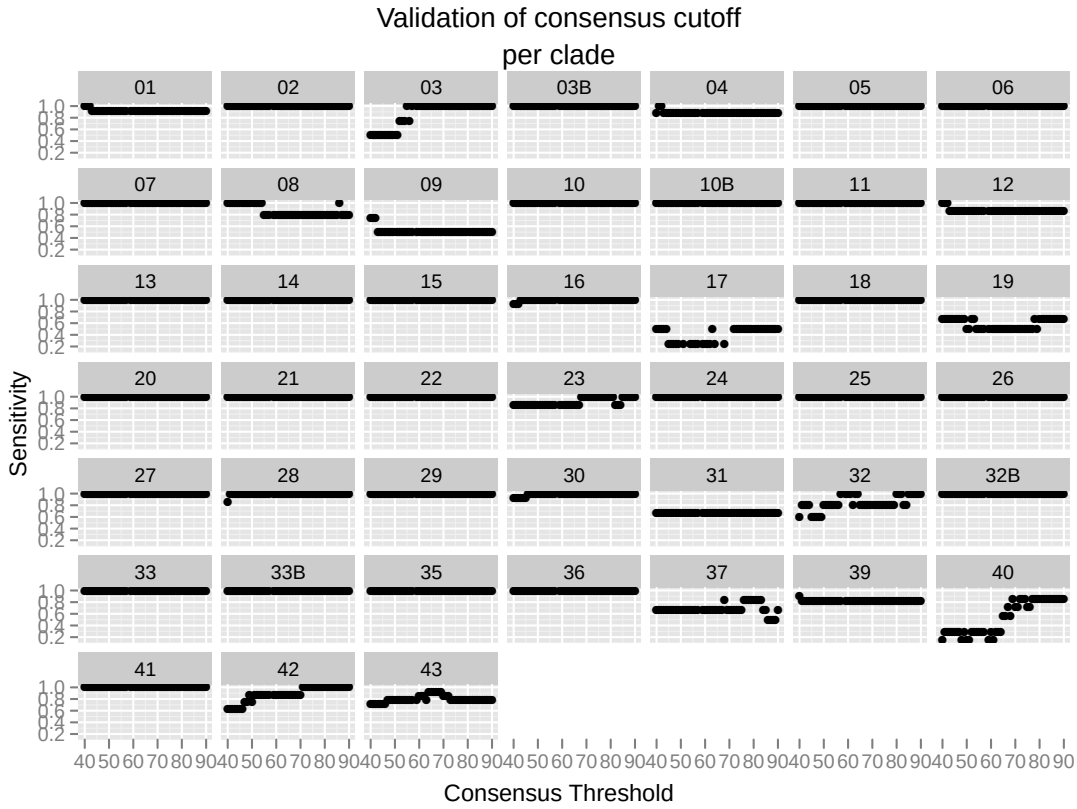



Figure 4.24: Graphs of the impact on sensitivity of HMM models by removal of residues that have above the threshold conservation in the general KS domain alignment. In most cases, removing KS wide conserved residues does not impact in the ability of the HMM models of most clades to recognize sequences that should belong to that clade as first results.

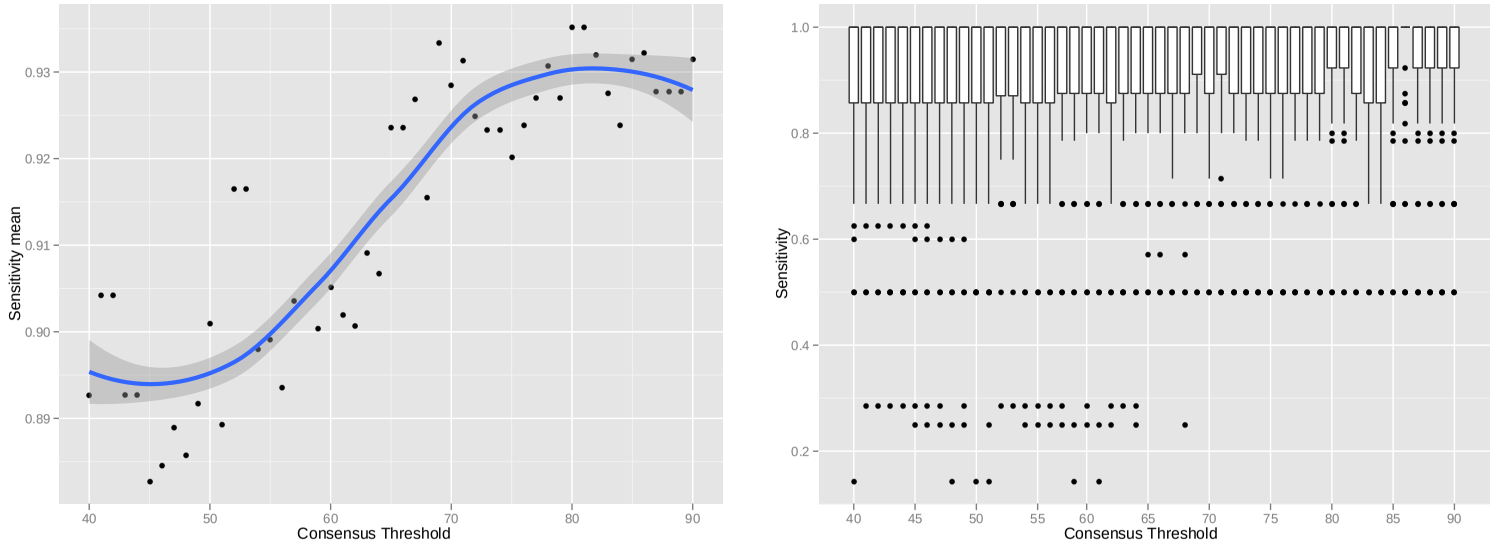



Figure 4.25: Study of the distribution of sensitivities for different consensus thresholds, cut-offs for removing residues from the individual clades that showed above consensus in the general KS domain alignment. *Left:* The mean of sensitivities graph shows that, after a region of steady gain of sensitivity between ~50% and ~70%, sensitivities enter a plateau. The trend is generated through a local polynomial regression (also known as LOESS). *Right:* Box plots illustrating the complete distribution of sensitivities across the different clades. These show that above ~70%, the lower parts of the distributions tend to be higher (third quartile, lower whisker and outliers).

KS clades is higher for the UniProt *trans*-AT PKS selection – meaning better alignment results – compared to the UniProt *cis*-AT PKS selection, which in some cases even do not show up for some particular clades.

This is a very positive result, as our software preferentially recognizes *trans*-AT KS domains over *cis*-AT KS domains, which would be the closest sequences that could confound the method. This result is an important milestone for our future work on mapping the identified KS domains in putative *trans*-AT PKS sequences into an actual chemical structure produced by the polyketide synthase.

The next step is a detailed phylogenetic analysis to incorporate the best hits produced by the HMMER models for each of the clades, making their signal more robust, and eventually splitting some of the clades into more specific substrate specificity. Once this is done, the step from the domain predictor written during this work, to an actual polyketide chemical structure generator is direct, as most of the clades have been mapped to their preferred substrates. This will be achieved through a combination of our KSDomainPredictor tool and a software based on the CDK to produce a chemical structure based on the found KS domains for new sequences.

### 4.2.6 Polyketides conclusions

The field of polyketides is of immense interest for the community due to the applications that these complex small molecules have. Our increasing ability to understand how polyketides are synthesized generates exciting new opportunities for the engineering and synthesis of molecules that otherwise it would be very complex to generate.

My work, on the automated generation of PCR primers for use in the amplification of new PKS sequences, contributes towards the discovery of new polyketide variants in marine sponge symbiont's meta-genome. In this area, our collaborators have had increases in productivity due to the tools that I developed. We are currently planning a revised version of this software that incorporate more features, inspired in existing solutions that were left for future releases, and by what our collaborators have learned from using the initial primers.

The identification of relevant KS domain residues in a clade specific fashion

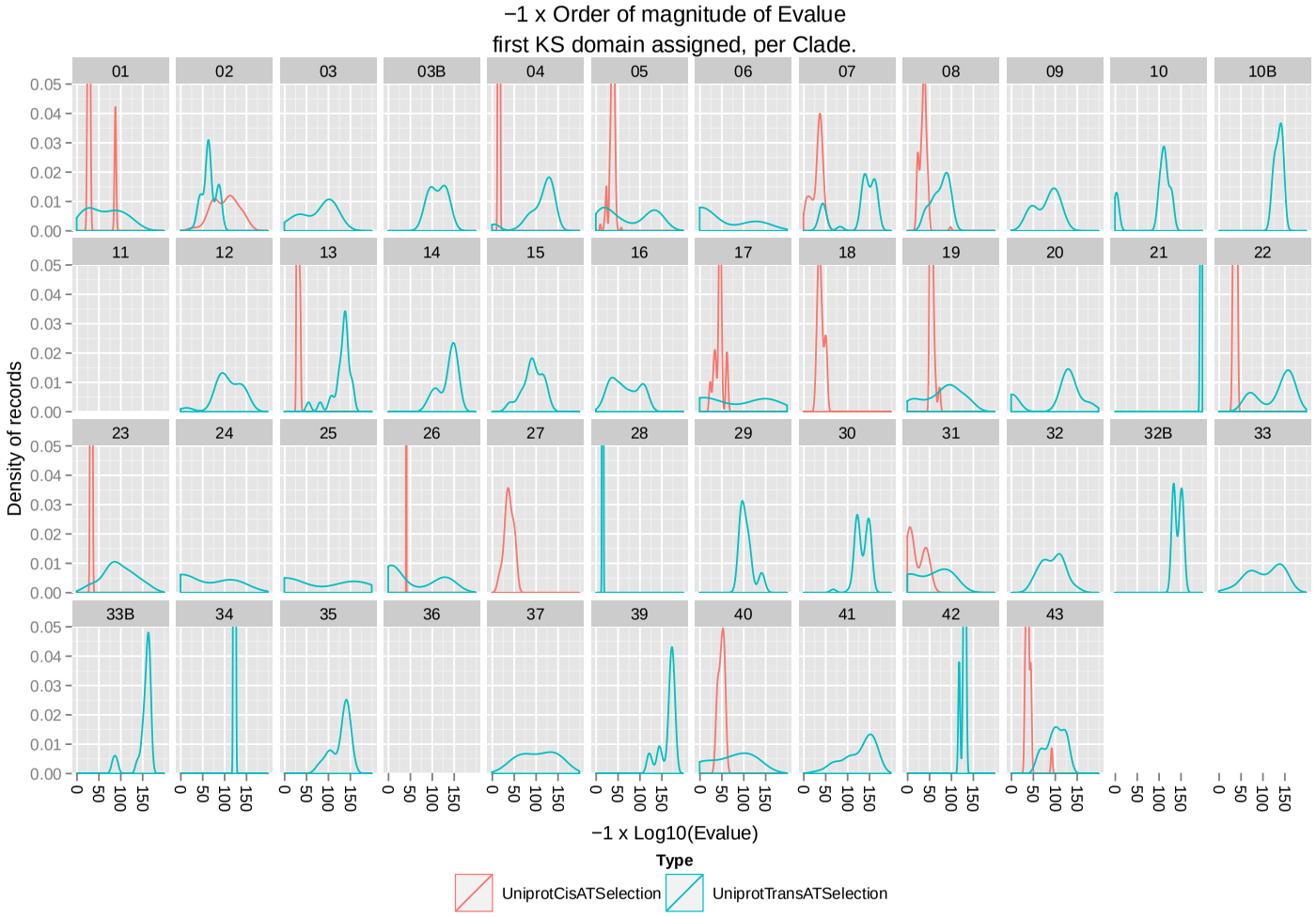



Figure 4.26: Comparison graphs for E-values of HMMER *trans*-AT KS models search against a UniProt dataset of *trans*-AT PKS proteins and a negative control dataset of *cis*-AT PKSs, for each KS clade. Most of the *cis*-AT PKSs either do not match or show $E_{value} > 10^{-50}$, while in most cases *trans*-AT PKSs get $E_{value} < 10^{-50}$.

has potential to guide – by allowing to focus in a few residue positions – protein engineering studies towards new synthesis possibilities. The methods I implemented proved to have better chances – compared to a published method – of identifying residues that were both close to the putative reactive center of the KS domain and that also show evidence of being relatively specific for each of the KS domains.

My work on the generation of general-signal-reduced KS HMM models builds towards an improved model that allows to map newly discovered *trans*-AT PKS sequences to the polyketide that they can synthesize. *trans*-AT PKS have been far less studied that their *cis* counter parts, which means that our work in this area could be of great impact in the diversity of polyketides that could be discovered, understood and even at some point synthesized through recombination of PKS modules.

# Chapter 5

# Conclusions

## 5.1 Summary and Conclusions

Mapping metabolomics results with known small molecules is a difficult problem. This is due to a wide range of obstacles: different technologies address different parts of the metabolome and have different sensitivities, the stability of small molecules is varied, the metabolome itself changes very rapidly, the range of concentrations spans various orders of magnitude, etc. These obstacles affect the experimental side and/or the posterior bioinformatics analysis of the results.

In the bioinformatics side, we recognized three major challenges posed by metabolomics: the need for data standards, a better integration with other omics, and the need for better reference metabolome databases. My work centered on that last challenge. In the past years, the major landmark in that part of the field was the introduction of the Human Metabolome Databases, HMDB, which increased the number of small molecules that could be easily associated to *H. sapiens* by four times, compared to other resources previously available.

Producing resources like the HMDB for more organisms is very important for the advancement of metabolomics and its integration with the other omics. However, these type of resources are very intensive in manual curation, and for the same reason are both expensive and slow to develop. In this work I explored methods to come up with extensive collections of small molecules in an organism-specific fashion and in the most automatic possible way. While my results might not have the in-depth annotation that entries in HMDB have, it is a good starting

point for curation efforts, as it integrates various important sources of data, and the procedure can be applied to many other organisms. Results are highly cross referenced to relevant chemical and metabolism databases, and include many synonyms and abbreviations. The pipeline annotates small molecules with tissues and cell types.

The exploration of the bioinformatic possibilities for generating complete metabolomes visited three areas: unification of metabolism databases; text mining of small molecules, proteins, tissues, and organisms; enumeration of reaction mechanisms; and a revision of a complicated example in metabolism: polyketides.

The results can be divided in different quality levels, ranging from high quality data for metabolism databases, medium quality for text mining, and finally lower quality for enumerated small molecules. As with any automatic method, there is a trade off between human intervention, scope, and quality of the results: more automatic methods yield more results, but of lower quality. For that reason, both text mining and reaction enumeration results are filtered to produce a selection. Candidate small molecules in the medium and lower quality set could be considered of better quality, if found in metabolomics experiments. The methods I proposed, generate a few thousand additional candidate small molecules to those present in the HMDB: ~1,600 from metabolism databases, ~2,400 from text mining and more than ~4,100 from reaction enumeration.

### 5.1.1 Integration of Metabolism Databases

Through the integration of the small molecules from different metabolism resources, we learn that there are many parts of them that are complementary. For this reason, there is an important gain in including all the resources, instead of using any one of them. Even though it is true that the merge process has errors, which might lead to higher levels of intersection between the resources, boundaries found for the level of error showed that, at the most, 20% of the unified groups of molecules could be wrong, and probably much less than that.

Metabolism databases unified for *H. sapiens* manage to explain ~15% of the HMDB small molecules. Additionally, these databases together produce more than a thousand small molecules for *H. sapiens* that are not part of HMDB.

Comparisons against the HMDB, of the three unified metabolism databases for *H. sapiens*, shows that their main difference in terms of coverage is the high amount of lipids in HMDB (nearly 3/4 of it are lipids). The fact that metabolism databases include only small molecules that participate in known reactions, and that many reactions involving lipids in these databases are generic, explain partly why metabolism databases miss so many lipids.

A closer look on different organism's metabolites – using enrichment analysis – reveals that higher eukaryotes like *H. sapiens* and *M. musculus* possibly have a higher complexity of small molecules towards lipids. In contrast, *E. coli* shows a higher complexity expanding to carbohydrate variants. Analyses here are partly limited by the coverage of the ChEBI ontology and the NCBI MeSH chemical annotations for small molecules in PubChem Compounds. There is a strong need to improve these resources, to be able to make better enrichment analysis.

### 5.1.2 Text mining for metabolomes

Through the use of dictionaries for organisms, tissues, proteins, and small molecules, the text mining module built recovers on its own ~15% of the entries in HMDB, including ~6% of the missing HMDB entries (486 small molecules) that the database unification of chapter 2 was not able to explain. The text mining approach covers approximately one third of the entries generated by the database merge. The approach generates 2,422 small molecules that could be part of a *H. sapiens* metabolome, but that are not part of neither the HMDB nor the metabolism database unification. Many of these molecules seem to be exogenous, in contrast to what happens in the case of database merge.

The use of protein co-occurrences, increases the number of small molecules that can be associated to tissues. This increase however comes at an expense. The inclusion of proteins makes the small molecule to tissue associations less specific, producing a wider and more general set of small molecules, given a tissue. Direct small molecule to tissue relations generate smaller sets of small molecules that are more specific to the tissue. Particular examples inspected – liver and brain related small molecule sets – show that small molecule sets obtained through proteins are closer to a core metabolome shared by tissues, while direct tissue to

small molecule retrieval generates tissue specific molecule sets. Also, direct small molecule to tissue co-occurrences form better thematic clusters of tissues than protein mediated co-occurrences.

An important legacy from this part of the work is the software that executes the entity name recognition, tagging, normalization, and analysis of co-occurrences for all NCBI PubMed abstracts. This is a starting point for improvements in later work.

The results shown are obtained after a series of filters and scores that attempt to cope with the high level of noise of the text mining data. The main component of this noise seems to be the existence of unspecific synonyms – mostly in small molecules and protein dictionaries. This problem can be resolved by building better dictionaries, where complete names can be semantically distinguished from abbreviations, and promiscuous synonyms filtered out, among other improvements. Another important difficulty, is to decide to which species an abstract is referring. This probably requires the use of natural language processing methods. These improvements should be part of more specialized follow up work in text mining. Probably the greatest gain, in terms of quality of the results, would come from improving the used dictionaries.

### 5.1.3 Reaction enumeration

The process of reaction enumeration is very demanding in terms of computational power, due to the algorithms used and the combinatorial explosion of the chemical space. To tackle it at a metabolome-wide level, I have built a framework based on existing open source technologies and cheminformatics tools, to make it more efficient, parallel, and distributable across a computer cluster.

The reaction enumerator generates more than 100,000 different new small molecules in three iterations, starting from molecules in the text mining collection and in the metabolism databases integration (both for *H. sapiens*). It uses generic reactions from *H. sapiens*, as found in the integration of metabolism databases. As iterations proceed, generated small molecules deviate more and more from the original set of molecules. The complexity of molecules also increases, as new reaction mechanisms participate in iterations two and three, that were absent in

the previous iterations. Higher masses are seen as the iterations advance, most of them still within the small molecule range.

The fraction of the HMDB that the reaction enumeration manages to match is much lower than the two previous cases – database unification and text mining. Only considering small molecules that have some evidence of being real, this method produces the largest amount of candidate small molecules to form part of the metabolome, but of lower quality than the other two approaches.

### 5.1.4 Polyketides discovery

The work presented in this part, contributes towards understanding better the rules governing *trans*-AT Polyketide synthases, which have been less studied than their *cis*-AT PKSs counter parts. *trans*-AT PKSs followed a different evolutionary path, which makes most of the rules known for *cis*-AT PKSs not applicable to them.

The pipeline I wrote for the generation of primers for *trans*-AT PKS KS domains proved very valuable to our collaborators, increasing the number of working specific primers by ~30%. These primers allow our collaborators to discover novel variants of KS domains within PKSs sequences, in the symbiont bacterial flora of the marine sponge *Theonella swinhoei*, as well as in other metagenomes. Improvements such as higher clade coverage, and newer features inspired by other solutions are part of an upcoming version.

Another source for the discovery of new *trans*-AT PKS associated KS domain sequences are the many new bacterial sequences deposited everyday in the sequence repositories. To exploit this, I wrote a software relying on hidden Markov models to identify particular variants of *trans*-AT PKS associated KS domain sequences. The identification resolution of this software goes beyond the capabilities of the established protein domain databases, and it is a milestone towards both a semi-automatic *trans*-AT PKS associated KS domain database and an automated molecule generator for *trans*-AT PKS sequences.

## 5.2 Field-related conclusions

While chemical ontologies and classification schemes are potentially very useful, chemical databases have a serious deficit in the amount of small molecules that are adequately annotated with these ontologies. For instance, a high proportion of the ChEBI molecules do not have an assignment to a ChEBI role. There is an important need not only to increase the size of these collections, towards the identification of more unknowns, but also of a denser population of these annotations of roles and functions.

There are still many closed source technologies that need to be opened or emulated in the open source software world in cheminformatics. Different from bioinformatics, cheminformatics has been dominated by commercial interests – especially due to its high impact in pharmacological development – leading to many key developments happening within companies. This has led to important algorithms (like SMILES) having various competing closed implementations that do not nurture each other. It also makes relevant proprietary cheminformatics algorithms unavailable to academia. Luckily I was able to obtain a license for the calculations of protonation micro-species and pKa. This work, as many others, would be benefited if more of these technologies – which are mostly pre-competitive – would have been open and accessible. This is specially true for the classification step of text mining results and for the reaction enumeration part, where technologies that are currently unavailable could have been used to characterize better the generated molecules.

This lack of openness also reaches databases. The recent closure of KEGG ftp access was a major loss for the world of metabolism research. On top of that, most metabolomics databases only allow searches of spectra or mass ranges, but do not grant easy access to their structures in bulk format. Only during the last period of this work, it was revealed that there was a way of obtaining the structures from KnApSacK, after accepting a license in Japanese. METLIN, a major metabolomics resource, does not provide its mapped structural data (even upon request). This is not exclusive to metabolomics databases: BRENDA does not provide access to small molecule's structures, and the main data download presents many technical obstacles for bulk usage and parsing. The key players in

the field should realise that not only html web access to their resources is important, but that also bulk access is, through adequate downloads and programmatic access (web services). Most of these resources are publicly funded, and as such, their data should be truly public as well.

There are important lessons learnt regarding databases integration, where minor structural details can mean a number of molecules deemed as different when they are not. These minor differences make InChI based comparisons fail in many cases. Different protonation states and different stereo-chemical representations are among the most common structural problems. It is also important to avoid hydrogen balance within biochemical reactions in metabolism databases, as biochemical reactions do not balance protons, and this makes reaction integration more complicated. There is a need for better cross talk between the main metabolism databases, to reduce these differences, which otherwise require elaborated solutions to be circumvented.

Another important point is to really understand the resources used, specially larger databases like ChEBI and PubChem Compounds, where the diversity of compounds leads to the use of primary and secondary (or parent and children) identifiers, which researchers normally will compare as different identifiers.

There is a need for chemical databases to provide equivalency tables within their own identifiers, both through web services and as downloadable data for local use. This equivalence classes could be generated under different assumptions, such as equivalent connectivity or micro-species with different protonation states. This will help researchers who compare small molecules only by cross references, which is fairly common.

Metabolism databases abuse of the generic molecule concept for some compound classes, leading sometimes to totally different uses for the same generic structure. Many times the same generic structure, specially very vague ones, are used for more than one different metabolite class, making the definition of these metabolite classes something very diffuse. BioCyc defines the generic molecule "a phenol" as an R group with an hydroxyl group bound, while also using the same structural definition for "an alcohol". In KEGG the supposedly equivalent class "a phenol" has the same structure as "phenol", a non-generic molecule.

There is a lack of semantic annotation for proteins participating in reactions,

which are normally encoded as generic molecules, and are left to the final user to classify them as proteins, based only on the name. These should be at least linked to a proper ontology term (Gene Ontology) or database identifier (UniProt). The same holds true for RNAs participating in biochemical reactions.

There is a need for further specification of generic reactions in metabolism databases, as this hampers their enumeration in a useful way. Reactions could be accompanied by a SMARTS string that encodes restrictions for the generic substrates that participate. This could largely limit the promiscuous enumeration of highly unspecified generic reactions such as EC numbers 1.1.1.1 and 1.1.1.2.

Towards the use in text mining efforts, there is a need to semantically annotate or distinguish small molecule names, synonyms, and abbreviations, and to get rid of those that are unspecific. This is particularly important for abbreviations that have meanings both in small molecule and in protein vocabularies, as they create many false positives. Also, it is important to choose carefully which synonyms for a particular molecule should be included in dictionaries. Another source of error is the incorrect enzyme name tagging as a small molecule, when the complete enzyme is not recognized by a protein dictionary.

## 5.3 Future work

Important improvements can be obtained from enhancing the dictionaries used for text mining, and generating an engine that allows to decide to which organism a particular abstract or sentence refers to. These improvements in the text mining part would readily reduce the number of false positives and reduce the need for stringent filtering.

The text mining part relied on article's titles and abstracts; it would be interesting to use full text articles instead. This will certainly generate technical challenges in the data handling part, as it was already demanding with abstracts and titles only.

Future work could include the addition of the structures associated to organisms available in many metabolomics databases. While many of these associations are drawn from the same metabolism databases that I use, there could be as well

unique associations from experimental data that would very valuable. Cross referencing these associations with molecules discovered through enumeration could as well serve to support the existence of these hypothetical molecules.

Future work in the reaction enumeration part should be the application of the same framework but with generalized reactions (up to the third level of the EC numbers organization) belonging to an organism. These will probably capture a wider mechanism diversity than the initial pick of generic reactions. Another important improvement should be to add an additional layer of rules for reactions that require minimal substructure in their markush structures. This will limit the enumeration of those reactions to more realistic scenarios, narrowing down the paths taken.

The Gibbs Energy calculator introduced can have an important impact in the future integration of metabolomics with systems biology modeling. Once metabolomics has gone past the difficulty of generating absolute concentrations, it would be natural to integrate this data in whole-genome metabolic models (flux balance analysis type). These models conserve fluxes, but do not incorporate small molecule concentrations. Through the use of Gibbs energies that account for reaction direction, these concentrations could be integrated into these models, as restrictions for the reaction directions.

Recurrently I mention the important portion of lipids present in the HMDB. While not included on this thesis, I am currently working on a lipid enumerator which sees its output through the LipidHome[1] initiative at the EBI. Given a head group and size ranges for fatty acid chains, my software efficiently generates all possible lipids according to rules given by a group of lipidomics experts. These rules constraint the possible fatty acids that can be attached to the head group. The main difference with LIPID MAPS is that they constraint lipids to those with experimentally detected fatty acids, our database generates all theoretically possible fatty acids given rules for the saturation patterns. Together with my collaborators, we hope to represent as many known lipids within our theoretical lipid framework as possible.

[1] http://www.ebi.ac.uk/apweiler-srv/lipidhome/

## 5.4 Concluding remarks

Small molecule data are complex to integrate, and hence a large amount of effort was necessary to produce an adequate integration of the data sets. The complexity of this kind of data comes from the fact that it has various layers, being the structural one particularly amenable for interpretation by the different data producers.

The thesis work has reached the objective of generating small molecule catalogues that lie in size between metabolism databases and the HMDB for *H. sapiens*. This organism-specific catalogue is integrated by ~3,000 small molecules belonging to metabolism databases, an additional ~3,000 small molecules retrieved through text mining, and more than ~4,000 additional small molecules generated through reaction enumeration with some degree of evidence to be considered as real molecules. While the size of the collection seems comparable to the HMDB, it only covers approximately ~21% of it. While many molecules could be potentially ruled out by manual curation, there is still a chance that many of them belong to the *H. sapiens* metabolome.

Finding so many small molecules through the database integration and the text mining effort that do not map to the HMDB is also an indication that even this metabolome is far from complete. The integration method was validated manually in two different ways, so this cannot be directly blamed on the integration. This implies that more work is required to have a *H. sapiens* metabolome, as many of these missing molecules are part of existing resources, not just completely unknown molecules.

Technical lessons learnt will be an input to initiatives like the MetaboLights database, particularly for the creation of its reference biology data set. This resource will require the integration of data from metabolism databases, text mining, and other sources for its reference biology layer.

# Appendix A

# Chemical name normalization

Chemical names use a number of lexicographical resources, such as punctuation, hyphenatation and parenthisation among other, that can make comparison of names a complicated problem. Some databases, like BioCyc often include HTML tags in the chemical names to produce subscripts or greek letters. There are as wells other issues of nomenclature, but those probably escape the scope of this humble solution.

Google Refine is a tool designed for aiding in manual data merges, where the source data sets can be noisy and requires manual intervention to be merged. Imagine for instance the task of unifying two directories of people, where some of them might be in both directories, and where the main index is something like the surname and name. Done for a multinational city like Cambridge, one would find many diverse names, some with unusual characters and even ocassionally mispelled. Google Refine implements a number of algorithms to aid in these type of merges.

One of the algorithms implemented in Google Refine is a word fingerprinter, which essentially takes a word, applies a number of substitutions and transformations, and produces out of it a key which is much more robust for string comparison. Transformations classically applied by the fingerprinter method of Google Refine are (please note that this list is taken from the Google Refine documentation):

- Remove leading and trailing whitespace
- Change all characters to their lowercase representation
- Remove all punctuation and control characters
- Split the string into whitespace-separated tokens
- Sort the tokens and remove duplicates
- Join the tokens back together
- Normalize extended western characters to their ASCII representation (for example “gödel” to “godel”)

Considering the nature of chemical names, we made a few changes to the source code that implements those steps, to generate a chemical name fingerprinter. The steps our fingerprinter takes are:

- Remove leading and trailing whitespace.
- Change all characters to their lowercase representation.
- Remove generic chemical name beginning, changing “a lipid” to “lipid” or “an alcohol” to “alcohol”.
- Remove plural endings.
- Remove HTML tags.
- Remove all punctuation and control characters, including dashes “-” and internal spaces.
- Split the string into individual characters, producing two groups of tokens, one for letters and one for numbers. Do not remove duplicates.
- Sort the group of letters, but not the numbers.
- Join the tokens back together
- Normalize extended western characters to their ASCII representation (for example “gödel” to “godel”).

# Appendix B

# Fingerprint and Isomorphism comparison of different regions of the chemical unification

This appendix consists of all the comparisons between the groups in the different regions of the Venn diagram in Figure 2.20, for *H. sapiens*. For each pair of groups the best fingerprint and isomorphism comparison is shown.

For reasons of formatting, number of pages, and printing quality, this appendix had to be included electronically. It is as well better viewed as a PDF than as a printed page. It is available at:

`http://www.ebi.ac.uk/steinbeck-srv/suppmat/metabolome-inference/AppB.pdf`

## B. 1 Captions explained

Similarity equals to 1 means perfect Fingerprint similarity. Isomorphism sim., as it is abbreviated in the captions of the appendix, equals to 1 means complete

isomorphism similarity. SBS in the caption stands for the Stereo chemistry Basic Similarity, where "Yes" means that the the program finds them to be equivalent, and "No" means that the program finds evidence that the stereo chemistry is different. "Yes" in this case is not sufficient to say that the actual stereo chemistry of the show components is the same, but that needs to be checked. "No" on the other hand should be sufficient to say that the two molecules have different stereo chemistry.

# Appendix C

# Enrichment analyses results

## C. 1 Statistical basis

The enrichment analyses presented on this thesis rely on the hypergeometric distribution for the null model. The hypergeometric probability is calculated as equation C.1 shows, for obtaining $k$ successes in $n$ trials, where there is a maximum of $m$ possible successes in a population of size $N$.

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{C.1}$$

Considering the enrichment of small molecules in classes of an ontology, this would be translated in the following way: given a sample of $n$ molecules, $N$ the total number of molecules in the ontology and $m$ the total elements in the ontology in the class examined, then $P(X = k)$ represents the probability – according to the hypergeometric distribution – of obtaining $k$ molecules from the sample size $n$ that belong to the ontology class examined (size $m$). A very high $P(X = k)$

means that there is probably no enrichment as this scenario could be given just by chance (the null hypothesis model, in this case the hypergeometric distribution). A very low $P(X = k)$ means that there must be a meaningful enrichment, as this is unlikely happening by chance.

The p-value reported in the tables is derived from this $P(X = k)$, but after a false discovery rate correction, to deal with the multiple hypothesis testing problem (derived from testing many ontology/hierarchy classes within the same set of small molecules).

As each of the hypothesis tested has a 5% chance of false positives (for $\alpha = 0.05$), that means that each time that we asses whether a group of small molecule is enriched in ONE of the classes in the ontology, there is a 5% chance that our group is deemed as significantly enriched in that class when in fact is not. While this might be fine for one single class, after assessing for 100 ontology classes or roles, we get an accumulated of 5 false (positive) significant enrichments only by chance. A false positive is still a significant result (it has a p-value below the decided $\alpha$), but known to be incorrect. Benjamini-Hochberg false discovery rate allows to shift the attention from the probability of false positives among all the results (significant and non significant, the 100 classes in the example), emphasis put by the common p-value, to a probability of false positives restricted only within significant results (for the chosen $\alpha$, that is, a subset of those 100 experiments). This limits with more stringency the number of false positives obtained, which is translated in a corrected p-value, which is what it is shown in all the enrichment analyses.

A common way of calculating the Benjamini-Hochberg FDR is by using the following procedure:

1. Sort all the original $N$ p-values from low to high.

2. Starting from the highest p-value (index $i = N$) and going to the lowest one (index $i = 1$), check whether

$$p_i\text{-}value < \frac{i}{N}\alpha \tag{C.2}$$

   is true, where $i$ is the index of the current p-value. Stop when the inequality is true, denoting this $i$ as $i_{stop}$.

3. From $i = i_{stop}$ to $i = 1$, all the p-values are accepted.

The Benjamini-Hochberg FDR method only gives a threshold after which all p-values are accepted so that the probability of false positive within significant results is below the set $\alpha$. One might want as well to calculate, for each p-value, at what $\alpha$ it would still be rejected under the Benjamini-Hochberg FDR. The FDR-adjusted p-value gives this probability, which for each $p_i\text{-}value$ is calculated as equation C.3 shows.

$$p_{i,corr}\text{-}value = \min_{h \geqslant i} \left\{ \frac{N}{h} p_h\text{-}value \right\} \tag{C.3}$$

The minimum is taken to keep the new corrected p-values monotonous in the same order as the original p-values. Equation C.3 is nothing but to solve $\alpha$ for each $i$ in equation C.2. This corrected p-value is the one shown in tables and figures on the thesis.

| ChEBI class | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| Anion | 24.9 | 2.7E-28 | 2.6 |
| Organic anion | 23.2 | 2.7E-28 | 2.7 |
| Organic ion | 23.5 | 2.2E-26 | 2.6 |
| Ion | 27.8 | 3.8E-26 | 2.3 |
| Organophosphate oxoanion | 8.6 | 4.0E-19 | 4.5 |
| Oxide | 17.0 | 3.4E-15 | 2.3 |
| Heteroatomic molecular entity | 67.2 | 7.6E-15 | 1.3 |
| Phosphorus oxoacids and derivatives | 23.3 | 2.1E-14 | 1.9 |
| Carboxylic acid anion | 14.6 | 2.3E-14 | 2.4 |
| Oxoanion | 15.8 | 3.0E-14 | 2.3 |
| Polyatomic anion | 16.1 | 4.2E-14 | 2.3 |
| Molecular entity | 99.6 | 1.1E-13 | 1.1 |
| Phosphorus molecular entity | 23.3 | 1.3E-13 | 1.9 |
| Phosphoric acid derivative | 22.3 | 1.9E-13 | 1.9 |
| Phosphorus oxoacid derivative | 22.3 | 4.5E-13 | 1.9 |

Table C.1: Top 15 ChEBI classes enriched in the set of unique *E. coli* molecules. Most of the categories are high level ChEBI classes, which are very broad and do not give much information on the biological relevance of the set of molecules.

## C. 2 Enrichment analysis for unique *E. coli* molecules

Table C.1 shows the Top 15 ChEBI categories in terms of enrichment for the unique molecules of *E. coli*, as the Venn diagram of Figure 2.29 shows in Chapter 2. I performed this analysis with a modified version of BiNGO, a Cytoscape plugin originally deviced for Gene Ontology enrichment analysis.

The Top 15 ChEBI classes enriched in this case tell little about the biological implication of the set of small molecules, and even as chemical information are too general to be of much use. Table 2.9 shows some selected categories, that even

though are not among the most enriched classes, are still enriched to a reasonable degree and give more usable information.

Figure C.1 shows details of the enrichment for the carbohydrate derivatives ChEBI class (up to a two level depth from this ChEBI class). Figure C.2 shows details of the enrichment for the carbohydrate ChEBI class.

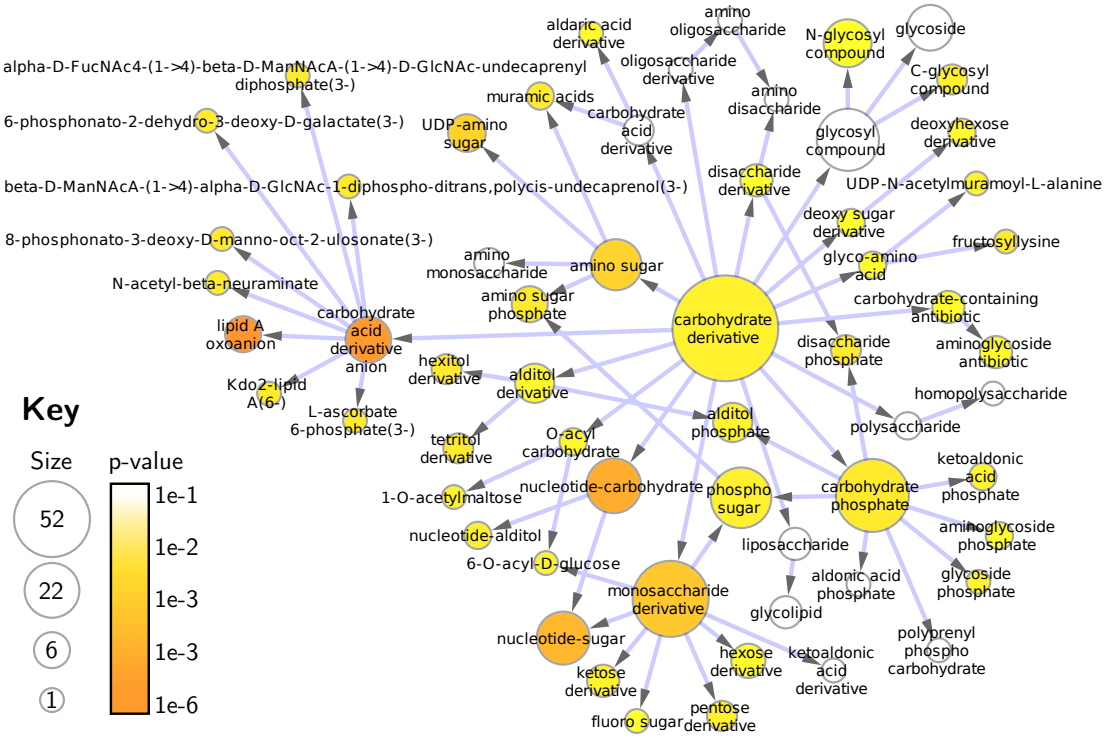



Figure C.1: ChEBI carbohydrate derivative class enrichment shown with two layers of depth (that is, descending up to two levels in the hierarchy). Bigger circles mean more molecules of this class present. Colours towards orange mean more significant enrichments, as the Key shows. Main enriched subclasses are nucleotide-carbohydrate, monosaccharide derivative, carbohydrate phosphate, amino sugar, and carbohydrate acid derivatives.

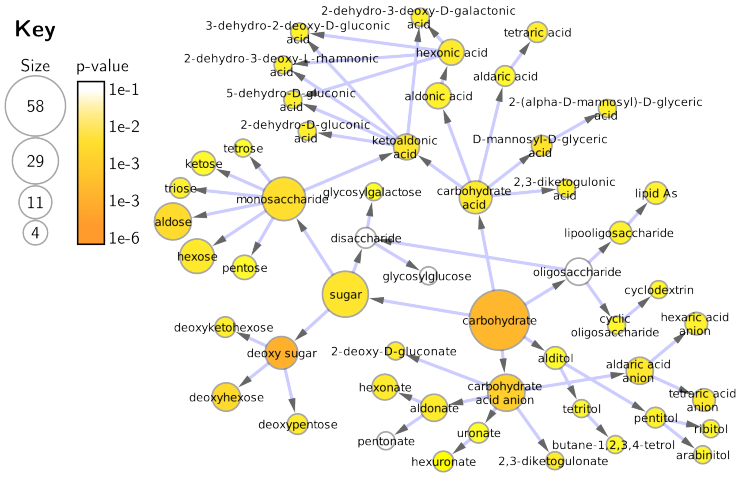



Figure C.2: ChEBI carbohydrate class enrichment shown with three layers of depth (that is, descending up to three levels in the hierarchy). Bigger circles mean more molecules of this class present. Colours towards orange mean more significant enrichments.

## C. 3 Enrichment analysis for *S. cerevisiae*

Table C.2 shows the top 15 most enriched ChEBI classes in the set of unique molecules from *S. cerevisiae*. No particular area of metabolism has an important enrichement in this case.

## C. 4 Enrichment analysis for *M. musculus*

Considering that *M. musculus* is being compared to *H. sapiens*, there should be very small fraction of unique small molecules belonging to *M. musculus*. Most what could appear as unique to *M. musculus* is probably either errors of unification or strange artifacts introduced by the base databases, or areas of study that

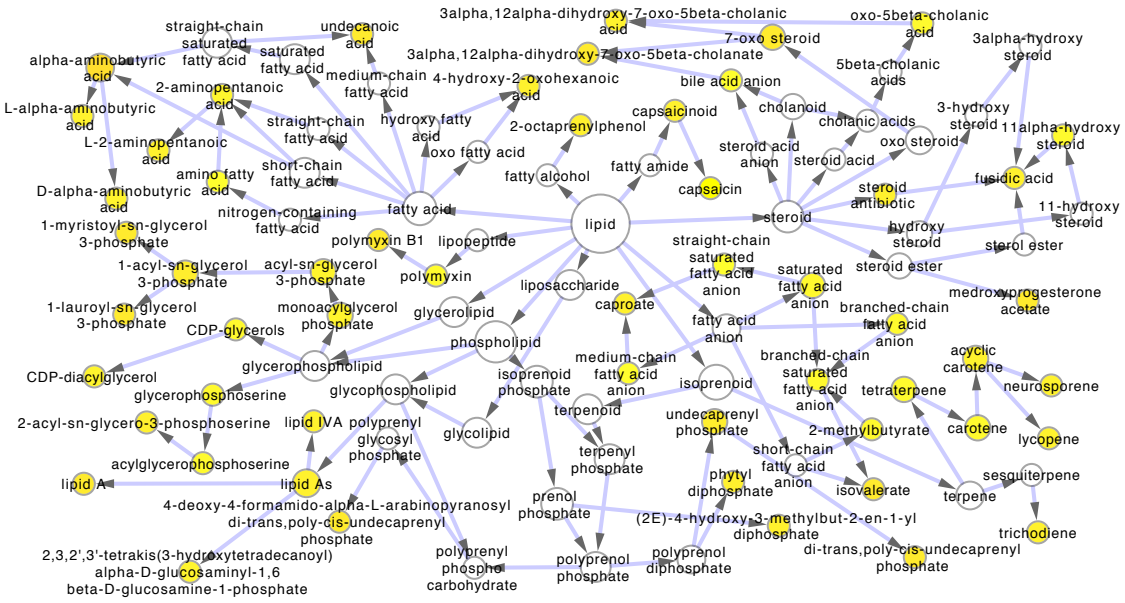



Figure C.3: ChEBI lipid class enrichment shown with maximum depth (that is, descending to each leave of the hierarchy). Bigger circles mean more molecules of this class present. Colours towards orange mean more significant enrichments. In this case, enrichment p-values are in a range of 1.4E-2 to 1, being most classes at p-value = 0.1, so the enrichment level is rarely significant. Sizes of nodes, which mean number of molecules in each ChEBI class, range from 1, in most of the cases, to 32 for the lipid class. This shows that there is seldom enrichment of lipid in the unique set of *E. coli*, or at least that it can be represented with the ChEBI ontology.

| ChEBI class | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| pyrimidines | 10.9 | 1.2E-03 | 6.1 |
| diazines | 10.9 | 1.2E-03 | 5.6 |
| thiamine | 3.0 | 1.2E-03 | 83.3 |
| fatty acid ester | 4.0 | 4.2E-03 | 21.0 |
| tricarboxylic acid trianion | 4.0 | 4.2E-03 | 21.0 |
| carboxylic acid trianion | 4.0 | 4.2E-03 | 20.3 |
| 2'-deoxycytidine phosphate | 3.0 | 4.2E-03 | 41.7 |
| homoisocitrate(3-) | 2.0 | 4.2E-03 | 157.4 |
| deoxycytidine phosphate | 3.0 | 4.2E-03 | 38.3 |
| p-block molecular entity | 99.0 | 4.2E-03 | 1.1 |
| tricarboxylic acid anion | 4.0 | 7.6E-03 | 15.2 |
| ethyl 3-hydroxyhexanoate | 2.0 | 7.6E-03 | 104.9 |
| 3beta-sterol | 4.0 | 7.6E-03 | 14.6 |
| main group molecular entity | 99.0 | 7.6E-03 | 1.1 |
| sterol | 4.0 | 1.1E-02 | 12.8 |

Table C.2: Top 15 most enriched ChEBI classes in the set of unique molecules from *S. cerevisiae* that had ChEBI mappings.

| ChEBI class | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| Cytokinin | 6.3 | 6.92E-06 | 64.8 |
| Aminopurine | 6.3 | 9.64E-04 | 21.1 |
| Phytohormone | 6.3 | 1.64E-03 | 17.4 |
| Hormone | 7.5 | 5.15E-03 | 9.3 |
| 6-isopentenylaminopurine | 3.8 | 5.15E-03 | 39.7 |
| Zeatin | 2.5 | 5.15E-03 | 149.0 |
| N-oxide | 3.8 | 5.15E-03 | 36.5 |
| UDP-sugar | 5.0 | 7.60E-03 | 15.4 |
| Phenylpyridine | 2.5 | 7.60E-03 | 108.4 |
| Agonist | 7.5 | 7.60E-03 | 7.3 |
| Organic molecular entity | 96.3 | 7.60E-03 | 1.2 |
| Carbon group molecular entity | 96.3 | 8.43E-03 | 1.2 |
| Molecular messenger | 7.5 | 8.43E-03 | 6.9 |
| Quinoline N-oxide | 2.5 | 1.21E-02 | 70.1 |
| Phorbol ester | 2.5 | 1.47E-02 | 51.8 |

Table C.3: Top 15 most enriched ChEBI classes in the set of unique molecules from *M. musculus* that had ChEBI mappings. Phytohormones (like cytokinin and zeatin) appear as they are included in BioCyc for *M. musculus*.

have been focused in *M. musculus* and less studies in *H. sapiens*, but I am not aware of any such a case. Enrichment analysis through BiNGO, for the 80 unique ChEBI entries that the unique molecules to *M. musculus* map, does not produce any relevant results. Even emblematic categories as lipids or carbohydrates show less than 10 small molecules. Table C.3 shows the Top 15 most enriched ChEBI classes.

| ChEBI class | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| Terpenoid fundamental parent | 2.2 | 8.7E-15 | 20.2 |
| Natural product fundamental parent | 4.1 | 3.0E-14 | 6.8 |
| Molecular entity | 99.0 | 1.1E-10 | 1.1 |
| Main group molecular entity | 96.2 | 2.4E-09 | 1.1 |
| Metabolite | 11.3 | 1.5E-08 | 2.1 |
| s-block molecular entity | 37.5 | 1.9E-08 | 1.4 |
| Secondary metabolite | 10.6 | 2.2E-08 | 2.2 |
| Terpene | 3.6 | 1.0E-07 | 4.2 |
| Hydrogen molecular entity | 35.0 | 1.5E-06 | 1.4 |
| Ketone | 8.3 | 2.2E-06 | 2.1 |
| Biochemical role | 14.0 | 9.9E-06 | 1.7 |
| Alkaloid | 4.6 | 9.9E-06 | 2.8 |
| 3-oxo steroid | 2.5 | 1.6E-05 | 4.2 |
| Chromenone | 2.8 | 2.6E-05 | 3.8 |
| Organic fundamental parent | 6.6 | 3.5E-05 | 2.2 |
| Organic hydride | 6.6 | 3.5E-05 | 2.2 |
| p-block molecular entity | 93.9 | 4.6E-05 | 1.1 |
| Hormone antagonist | 1.1 | 6.0E-05 | 9.4 |
| Benzopyran | 4.1 | 6.6E-05 | 2.7 |
| Hydrides | 7.4 | 7.7E-05 | 2.0 |

Table C.4: Top 20 most enriched ChEBI categories from the unique section of small molecules of the text mining metabolome. The enrichment uses 785 unique ChEBI entries, which are all the ChEBI identifiers that the database provides for the unique set of small molecules from the text mining *H. sapiens* metabolome.

## C. 5 Enrichment analysis for Text mining

### C. 5. 5 Unique Text mining small molecules

Table C.4 shows the top 20 most enriched ChEBI categories, after enrichment analysis of 785 ChEBI entries retrieved from the region of unique elements provided by the text mining approach, which Figure 3.17 depicts.

| NCBI MeSH Topic | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| Organic Chemicals | 17.4 | 1.2E-31 | 0.6 |
| Nucleic Acids, Nucleotides, and Nucleosides | 7.7 | 1.6E-24 | 2.7 |
| Heterocyclic Compounds | 14.3 | 1.3E-17 | 0.6 |
| Biological Factors | 5.2 | 6.0E-17 | 2.7 |
| Lipids | 6.6 | 3.7E-12 | 2.0 |
| Hormones, Horm. Substitutes, and Horm. Antagonists | 4.5 | 3.6E-08 | 2.0 |
| Enzymes and Coenzymes | 1.5 | 2.3E-04 | 2.4 |
| Pharmaceutical Preparations | 0.4 | 3.0E-03 | 4.9 |
| Chemical Actions and Uses | 12.4 | 3.7E-02 | 0.9 |
| Inorganic Chemicals | 1.5 | 3.7E-02 | 0.7 |
| Amino Acids, Peptides, and Proteins | 7.0 | 3.7E-02 | 1.1 |
| Polycyclic Compounds | 10.8 | 5.3E-02 | 1.1 |
| Complex Mixtures | 0.3 | 5.3E-02 | 2.6 |
| Carbohydrates | 4.2 | 6.8E-02 | 0.9 |
| Macromolecular Substances | 0.5 | 1.4E-01 | 1.3 |
| Biomedical and Dental Materials | 0.3 | 2.0E-01 | 0.9 |

Table C.5: Enrichment of NCBI MeSH topics that are direct children of Chemicals and Drugs Category (the source NCBI MeSH topic for small molecules). The enrichment uses 1973 unique PubChem Compounds entries, which are all the PubChem Compounds identifiers that the database provides for the unique set of small molecules from the text mining *H. sapiens* metabolome. This complements mostly the ChEBI portion of Table C.4.

Table C.4 presents the top 20 most enriched NCBI MeSH topics, after enrichment analysis of 1973 PubChem Compounds entries retrieved from the region of unique elements provided by the text mining approach, which Figure 3.17 depicts. This tends to be complementary to the previous ChEBI analysis.

| NCBI MeSH Topic | % | Enrichment | |
|---|---|---|---|
| | | p-value | Fold |
| Nucleosides | 5.1 | 3.6E-23 | 3.4 |
| Heterocyclic Compounds, 1-Ring | 4.9 | 2.9E-21 | 0.4 |
| Ribonucleosides | 2.8 | 3.3E-15 | 4.0 |
| Steroids | 7.8 | 3.8E-15 | 2.1 |
| Purine Nucleosides | 2.5 | 6.3E-14 | 4.0 |
| Inflammation Mediators | 2.5 | 2.8E-12 | 3.7 |
| Azoles | 0.9 | 8.1E-12 | 0.3 |
| Hydrocarbons, Cyclic | 2.8 | 4.4E-11 | 0.4 |
| Purines | 4.6 | 1.2E-10 | 2.3 |
| Hydrocarbons | 6.2 | 1.8E-10 | 0.6 |
| Eicosanoids | 2.1 | 2.6E-10 | 3.6 |
| Autacoids | 2.1 | 3.9E-10 | 3.5 |
| Hydrocarbons, Aromatic | 2.1 | 1.2E-09 | 0.4 |
| Androstanes | 1.9 | 2.5E-09 | 3.6 |
| Amines | 1.3 | 6.7E-09 | 0.4 |
| Noxae | 2.3 | 1.6E-08 | 3.0 |
| Amides | 0.5 | 2.4E-08 | 0.2 |
| Fatty Acids | 4.0 | 3.0E-08 | 2.1 |
| Adrenal Cortex Hormones | 1.8 | 7.7E-08 | 3.3 |
| Prostaglandins | 1.4 | 1.8E-07 | 3.9 |
| Pregnanes | 2.4 | 1.9E-07 | 2.6 |

Table C.6: Top 20 most enriched NCBI MeSH topics from the unique section of small molecules of the text mining metabolome. The table skips the direct children of Chemicals and Drugs Category (the source NCBI MeSH topic for small molecules). The enrichment uses 1973 unique PubChem Compounds entries, which are all the PubChem Compounds identifiers that the database provides for the unique set of small molecules from the text mining *H. sapiens* metabolome. This complements mostly the ChEBI portion of Table C.4.

# Appendix D

# Gibbs Energy calculator supplementary material

This appendix contains the supplementary material for the Gibbs Energy calculator built.

For reasons of formatting, this appendix had to be included as a document itself, and hence all its Figures and pages are not indexed as part of the main thesis. It also contains its own sections and references.

## D. 1 Derivation of equation to obtain non-apparent standard energy of least protonated isomer

From [1] we obtain equation D.1 (4.5-6 in [1]) and the partition function D.2 (4.5-7 in [1]), which explains how to obtain the isomer's Gibbs energy of formation[1], $\Delta_f G_i'^0$ (the apparent value, measured experimentally, which is used in the regression) from the least protonated specie's Gibbs energy of formation[2], $\Delta_f G_{j=1}'^0$.

$$\Delta_f G_i'^0 = \Delta_f G_{j=1}'^0 - RTlnP \tag{D.1}$$

$$P = 1 + \frac{[H^+]}{K_1} + \frac{[H^+]^2}{K_1 K_2} + \cdots + \frac{[H^+]^n}{K_1 K_2 \cdots K_n} \tag{D.2}$$

Besides, we know that a specie $j$ (part of an isomer group) can be specified at a desired $pH$ and ionic strength $I$ using equation D.3

$$\Delta_f G_j'^0 = \Delta_f G_j^0(I=0) + N_{H,j} RTln(10) pH - \frac{\alpha RT(z_j^2 - N_{H,j})\sqrt{I}}{1 + 1.6\sqrt{I}} \tag{D.3}$$

Using D.3 in D.1 we get

$$\Delta_f G_i'^0 = \Delta_f G_{j=1}^0(I=0) + N_{H,j=1} RTln(10) pH - \frac{\alpha RT(z_1^2 - N_{H,j=1})\sqrt{I}}{1 + 1.6\sqrt{I}} - RTlnP$$

$$\Delta_f G_1^0(I=0) = \Delta_f G_i'^0 - N_{H,1} RTln(10) pH + \frac{\alpha RT(z_1^2 - N_{H,1})\sqrt{I}}{1 + 1.6\sqrt{I}} + RTlnP$$

$$\Delta_f G_1^0(I=0) = \Delta_f G_i'^0 + RTln([H^+]^{N_{H,1}} P) + \frac{\alpha RT(z_1^2 - N_{H,1})\sqrt{I}}{1 + 1.6\sqrt{I}}$$

$$\Delta_f G_1^0(I=0) = \Delta_f G_i'^0 + RTln([H^+]^{N_{H,1}} + \sum_{k=1}^{k=N_s-1} \frac{[H^+]^{N_{H,1}+k}}{\prod_{j=1}^{k} K_j}) + \frac{\alpha RT(z_1^2 - N_{H,1})\sqrt{I}}{1 + 1.6\sqrt{I}} \tag{D.4}$$

If we are considering that the experimental values were measured probably at zero ionic strength, then $I$ associated to $\Delta_f G_i'^0$ is $I = 0$, which simplifies the equation to

$$\Delta_f G_1^0(I=0) = \Delta_f G_i'^0(I=0) + RTln([H^+]^{N_{H,1}} + \sum_{k=1}^{k=N_s-1} \frac{[H^+]^{N_{H,1}+k}}{\prod_{j=1}^{k} K_j}) \tag{D.5}$$

Hence, using equation D.5 we can obtain the standard (non-apparent) Gibbs energy of formation, $\Delta_f G_1^0(I=0)$ at zero ionic strength for the least protonated specie of the isomer group for which we had an apparent energy of formation, $\Delta_f G_i'^0(I=0)$ (derived from the regression). From here, one can obtain the energies for the rest of the species in the isomer group as described in [1] and then obtain finally the standard transformed Gibbs energy for the chemical entity in question, all starting from the values obtained from the regression and without needing experimental values.

[1] Please note that in $\Delta_f G_i'^0$ the $i$ stands for isomer.

[2] Please note that in $\Delta_f G_j'^0$ the $j$ index stands for the specie number within the isomer group, in ascending order of protonation, so $j = 1$ is the least protonated specie of the isomer group.

## D. 2 Improvement of the regression for apparent standard Gibbs energies

### D. 2. 2 The need for a new regression

The work by [5] provided most of the data directly used (much of it is derived by them from [4, 2, 3]). However, in the implementation of recognition of Chemical groups, some errors were noted when validating our group recognition implementation against their results. This resulted of course in the correction of many errors in our implementation, yet the discovery of certain mistakes in the recognition of certain groups in [5] required to make a new regression that would include this fixes. The main errors found and fixed were:

**Chloride groups** Cls are said to be attached to tertiary carbons (no other Cls attached) when they are actually attached to secondary carbons. The same happens with secondary and primary carbons. This can be seen in molecules 624, 659, 660 and 661, according to their supplementary material. This incorrect detection produces an error of 2.82 kcal/mol and 1.5 kcal/mol per group misdetected, respectively.

**=C< in ring** The regression from [5] makes a difference between two types of =C< (carbon with a double bond and two single bonds) participating in a ring: in the first case the two single bonds participate in the ring, in the second one single and one double bond participate in the ring. According to our group identification, in many cases one of this groups was identified by [5] when the other was the one actually present. This derives in a systematic error of 20.4 kcal/mol every time that one of this mistakes is made. This can be seen in mol 108, 403 from their supplementary material.

**-$SO_3$ group value** Molecules 678, 540, 679, 696 and 272 show that there is an inconsistency with the -SO3 group value. According to the manuscript, it is -156 kcal/mol, however the actual calculated energies for the molecules implies that should be -123.75 kcal/mol.

**$N^+$ badly typed** A number of molecules were found in the data set of the supplementary material in which N had four bonds but wasn't charged positively, this cases were corrected looking at the correct configurations in the ChEBI database. This impacted on certain groups being recognized.

Approximately 50 molecules from the data set of [5] were modified to correct some minor chemical issue.

### D. 2. 2 Work towards a new regression

Using the data provided by [5], our group identification implementation based on the CDK [6] and the corrections mentioned, a new multiple linear regression was derived using the lm (linear model) package from R. The best regression achieved, after fixing errors of implementation that introduced biases in different regions, was used in the end for our group contribution method. We see a slight improvement in the residuals compared to the regression from [5], but most importantly, groups are correctly recognized for the mentioned cases. The values for each of the groups resulting from the regression can be seen in the following R summary:

```
Residuals:
     Min       1Q   Median       3Q      Max
-14.3411  -0.6351   0.0019   0.9525  10.1824

Coefficients:
                                          Estimate Std. Error  t value Pr(>|t|)
CCCCconjugation                          -5.72340    0.40474  -14.141  < 2e-16 ***
CCCNconjugation                          -4.11169    0.83237   -4.940 8.28e-07 ***
=CH2                                      6.74411    0.30363   22.211  < 2e-16 ***
=CH-                                     13.41627    0.24577   54.588  < 2e-16 ***
=C<                                      16.62884    0.39344   42.265  < 2e-16 ***
=NH                                     -19.09548    1.45281  -13.144  < 2e-16 ***
=NH2                                    -17.59795    1.29785  -13.559  < 2e-16 ***
=N-                                      13.61456    3.11327    4.373 1.27e-05 ***
OCCCconjugation                          -1.79649    0.22200   -8.092 8.63e-16 ***
OCCNconjugation                          -3.11462    0.53873   -5.781 8.22e-09 ***
OCCOconjugation                           2.06671    0.17822   11.597  < 2e-16 ***
SG_ClO2                                   4.10000    1.82062    2.252 0.024401 *
SG_ClO3                                  -0.80000    1.82062   -0.439 0.660398
SG_H                                     -7.24227    0.16303  -44.424  < 2e-16 ***
SG_H2                                     3.12466    0.18519   16.872  < 2e-16 ***
SG_H2O2                                 -32.16839    0.95126  -33.817  < 2e-16 ***
SG_acetaldehyde                         -32.14846    0.38380  -83.763  < 2e-16 ***
SG_acetate                              -88.78636    0.20863 -425.578  < 2e-16 ***
SG_acetone                              -38.29296    0.40499  -94.553  < 2e-16 ***
SG_ammonia                              -18.03161    0.37048  -48.670  < 2e-16 ***
SG_ammonium_car=amate                   -91.28624    0.82866 -110.161  < 2e-16 ***
SG_ch4                                  -10.18738    0.45051  -22.613  < 2e-16 ***
SG_chloromethane                        -12.60000    1.82062   -6.921 5.55e-12 ***
SG_co2                                  -90.22668    0.39895 -226.159  < 2e-16 ***
SG_dichloromethane                      -15.80000    1.82062   -8.678  < 2e-16 ***
SG_diphosphate                         -483.10282    0.68600 -704.234  < 2e-16 ***
SG_ethane                                -5.04792    1.06058   -4.760 2.04e-06 ***
SG_ethanol                              -41.77921    0.39733 -105.150  < 2e-16 ***
SG_fe2                                  -20.20078    1.32787  -15.213  < 2e-16 ***
SG_fe3                                    0.25078    1.32787    0.189 0.850219
SG_formaldehyde                         -31.46352    0.58372  -53.902  < 2e-16 ***
SG_formate                              -85.13234    0.33747 -252.269  < 2e-16 ***
SG_formyl_phosphate                    -297.58271    1.87495 -158.715  < 2e-16 ***
SG_h2o                                  -57.10078    0.23898 -238.935  < 2e-16 ***
SG_h2s                                   -6.66000    1.82062   -3.658 0.000259 ***
SG_hco3                                 -140.93192    0.30347 -464.408  < 2e-16 ***
SG_hydroxylamine                        -40.29593    0.81131  -49.668  < 2e-16 ***
SG_methanol                             -43.33079    0.44060  -98.346  < 2e-16 ***
SG_methylamine                          -10.21721    0.79476  -12.856  < 2e-16 ***
SG_n2                                     1.70001    0.99822    1.703 0.088672 .
SG_n2h4                                  30.60000    1.82062   16.807  < 2e-16 ***
SG_n2o                                   20.08229    2.09299    9.595  < 2e-16 ***
SG_no                                    14.95428    1.14133   13.103  < 2e-16 ***
SG_no2                                   -9.54929    0.70621  -13.522  < 2e-16 ***
SG_no3                                  -26.63806    0.81956  -32.503  < 2e-16 ***
SG_o2                                    4.34799    0.97617    4.454 8.75e-06 ***
SG_o2Minus                               8.12208    1.12869    7.196 7.92e-13 ***
SG_oxalate                            -161.08500    1.28737 -125.127  < 2e-16 ***
SG_pi                                 -267.96003    0.37897 -707.072  < 2e-16 ***
```

| | | | | | |
|---|---|---|---|---|---|
| SG_s | 5.65561 | 0.68123 | 8.302 | < 2e-16 | *** |
| SG_s2o3 | -131.25738 | 0.82113 | -159.849 | < 2e-16 | *** |
| SG_s2o4 | -143.50000 | 1.82062 | -78.819 | < 2e-16 | *** |
| SG_s3o6 | -226.74576 | 1.30737 | -173.436 | < 2e-16 | *** |
| SG_so2 | -71.87100 | 1.82062 | -39.476 | < 2e-16 | *** |
| SG_so3 | -118.47451 | 0.54884 | -215.863 | < 2e-16 | *** |
| SG_sulfate | -178.23359 | 0.71093 | -250.705 | < 2e-16 | *** |
| SG_tetrachloromethane | -10.80000 | 1.82062 | -5.932 | 3.36e-09 | *** |
| SG_trichloromethane | -15.90000 | 1.82062 | -8.733 | < 2e-16 | *** |
| SG_urea | -52.20692 | 0.95344 | -54.756 | < 2e-16 | *** |
| -CH2- | 1.73001 | 0.09239 | 18.725 | < 2e-16 | *** |
| -CH3 | -3.79204 | 0.14562 | -26.040 | < 2e-16 | *** |
| -CH=O | -29.93351 | 0.17332 | -172.709 | < 2e-16 | *** |
| -COO | -83.80153 | 0.15674 | -534.667 | < 2e-16 | *** |
| -CO-OPO3- | -293.33745 | 0.38851 | -755.041 | < 2e-16 | *** |
| -NH2 | -1.58569 | 0.44034 | -3.601 | 0.000322 | *** |
| -NH2- | 4.28922 | 0.82403 | 5.205 | 2.08e-07 | *** |
| -NH3 | -5.29332 | 0.30603 | -17.297 | < 2e-16 | *** |
| -NH< | 15.47413 | 1.13077 | 13.685 | < 2e-16 | *** |
| -NH- | 8.51985 | 0.50985 | 16.710 | < 2e-16 | *** |
| -N< | 21.16259 | 1.10382 | 19.172 | < 2e-16 | *** |
| -OH | -41.79636 | 0.15882 | -263.166 | < 2e-16 | *** |
| -OPO2 | -213.50172 | 0.56356 | -378.846 | < 2e-16 | *** |
| -OPO2- | -212.68954 | 0.35651 | -596.584 | < 2e-16 | *** |
| -OPO3 | -250.53039 | 0.34551 | -725.111 | < 2e-16 | *** |
| -OPO3- | -237.83039 | 0.50690 | -469.186 | < 2e-16 | *** |
| -OSO3minus1 | -154.46849 | 0.92317 | -167.325 | < 2e-16 | *** |
| -O- | -23.84818 | 0.39506 | -60.366 | < 2e-16 | *** |
| -O-CO- | -73.31272 | 0.44087 | -166.292 | < 2e-16 | *** |
| -Oneg | -37.15324 | 0.77463 | -47.963 | < 2e-16 | *** |
| -SH | -1.36061 | 0.63466 | -2.144 | 0.032132 | * |
| -SO3 | -123.23769 | 0.86285 | -142.826 | < 2e-16 | *** |
| -S< | 24.16641 | 1.98341 | 12.184 | < 2e-16 | *** |
| -S- | 7.90539 | 0.73740 | 10.721 | < 2e-16 | *** |
| -S-OH | 21.08688 | 3.38117 | 6.237 | 5.15e-10 | *** |
| -S-S- | 4.55842 | 1.19203 | 3.824 | 0.000134 | *** |
| -Sneg | 8.28676 | 2.68875 | 3.082 | 0.002076 | ** |
| =CH | 51.96352 | 4.44996 | 11.677 | < 2e-16 | *** |
| =C- | 37.22676 | 2.17793 | 17.093 | < 2e-16 | *** |
| =N | -23.35352 | 4.06048 | -5.751 | 9.81e-09 | *** |
| =CH- | 5.52149 | 0.16915 | 32.643 | < 2e-16 | *** |
| =C< | 7.82110 | 0.32783 | 23.857 | < 2e-16 | *** |
| =C=O | -27.67383 | 0.21852 | -126.644 | < 2e-16 | *** |
| =N< | 62.11248 | 1.88270 | 32.991 | < 2e-16 | *** |
| amide | -12.14147 | 0.36021 | -33.707 | < 2e-16 | *** |
| aromatic_ring =CH- | 4.69418 | 0.13942 | 33.670 | < 2e-16 | *** |
| aromatic_ring =C< | 8.11175 | 0.31993 | 25.354 | < 2e-16 | *** |
| aromatic_ring fused_to_nonaromatic_ring >C= | 8.43421 | 0.59123 | 14.266 | < 2e-16 | *** |
| aromatic-Br | 2.09777 | 1.21489 | 1.727 | 0.084328 | . |
| aromatic-F | -43.43556 | 1.21489 | -35.753 | < 2e-16 | *** |
| aromatic-I | 16.19777 | 1.21489 | 13.333 | < 2e-16 | *** |
| dbl_sgl_ring =N< | 14.74436 | 0.81136 | 18.172 | < 2e-16 | *** |
| heteroaromaticring | -3.55028 | 0.52938 | -6.706 | 2.40e-11 | *** |
| hydrocarbon | 4.15856 | 0.84849 | 4.901 | 1.01e-06 | *** |
| primary Cl2 | -8.05969 | 0.30808 | -26.161 | < 2e-16 | *** |

```
primary Cl3                              -5.69035    0.28457  -19.996  < 2e-16 ***
primary Cl                              -11.34913    0.38795  -29.254  < 2e-16 ***
ring >C<                                  8.18267    0.39495   20.718  < 2e-16 ***
ring >C=                                 31.51928    2.07075   15.221  < 2e-16 ***
ring >C=O                               -29.75670    0.30554  -97.390  < 2e-16 ***
ring =N-                                  6.10061    0.58077   10.504  < 2e-16 ***
ring =CH-                                 8.78084    0.28398   30.921  < 2e-16 ***
ring =C<                                 12.05322    0.37387   32.239  < 2e-16 ***
ring =NH-                                 7.11203    0.87432    8.134 6.15e-16 ***
ring -CH2-                                2.77125    0.24392   11.362  < 2e-16 ***
ring -CH<                                 5.44459    0.20895   26.057  < 2e-16 ***
ring -NH-                                 4.43847    0.51844    8.561  < 2e-16 ***
ring -N<                                 19.83668    0.60056   33.031  < 2e-16 ***
ring -OPO2-                            -194.18371    0.96634 -200.947  < 2e-16 ***
ring -O-                                -37.32142    0.85990  -43.402  < 2e-16 ***
ring -O-CO-                             -70.67898    0.79034  -89.429  < 2e-16 ***
ring -S-                                  0.65833    0.70165    0.938 0.348192
secondary-Cl                             -8.24988    0.41186  -20.031  < 2e-16 ***
thioesterFactor                         -10.12517    0.45413  -22.296  < 2e-16 ***
threering                                14.61372    1.49429    9.780  < 2e-16 ***
two_fused_aromatic_rings >C=             -0.02287    0.88666   -0.026 0.979427
two_fused_rings >C=                      16.78700    0.87552   19.174  < 2e-16 ***
two_fused_rings >CH-                      3.58945    0.73718    4.869 1.18e-06 ***
two_fused_rings >C<                      -1.92597    2.94102   -0.655 0.512608
two_fused_rings =N<                       6.00654    1.25923    4.770 1.94e-06 ***
two_fused_rings -N<                      12.53668    1.06413   11.781  < 2e-16 ***
vicinalCl                                 1.50652    0.34212    4.403 1.11e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.821 on 2804 degrees of freedom
Multiple R-squared: 0.9988,Adjusted R-squared: 0.9987
F-statistic: 1.789e+04 on 131 and 2804 =F,  p-value: < 2.2e-16
```

As it can be seen in the regression summary, only three coefficients had low significance, but they are nevertheless left in the model as it allows a higher coverage of molecules. The regression can be further evaluated through diagnostic graphs shown in Fig. D.1. As it can be seen in the Residuals vs Fitted graph, most of the residuals (difference between fitted value and data value) lie within $[-5, 5]$ kcal/mol. The scale location graph reveals certain tendency to higher errors for molecules and reactions with higher fitted energy values. This is to be expected in an additive method, as the bigger the mass of a molecule (or the side of a reaction), more groups will be found, accumulating error additively. The Residuals vs. Leverage graph shows that there are no points (either reactions or small molecules) with cook's distance $> 0.5$ (and there are none actually with cook's distance $> 0.2$). High cook's distances are associated to points that have a very high influence in the regression and at the same time have high residuals, meaning that they affect the model more than other points and are not fitting correctly, "deforming" the model normally. Most of the points with leverage $> 0.4$ have very low residuals. Overall, these graphs and the multiple R-squared of 0.9988 indicate a very good fit to the data. This further validates the quality of the groups recognition implementation based on the CDK.

Compared to the regression made by [5], our regression seems to improve on the overall residues obtained. This can be seen in Fig. D.2. Here it can be seen that the present regression has a higher proportion of its residuals towards zero (black line), compared to the distribution

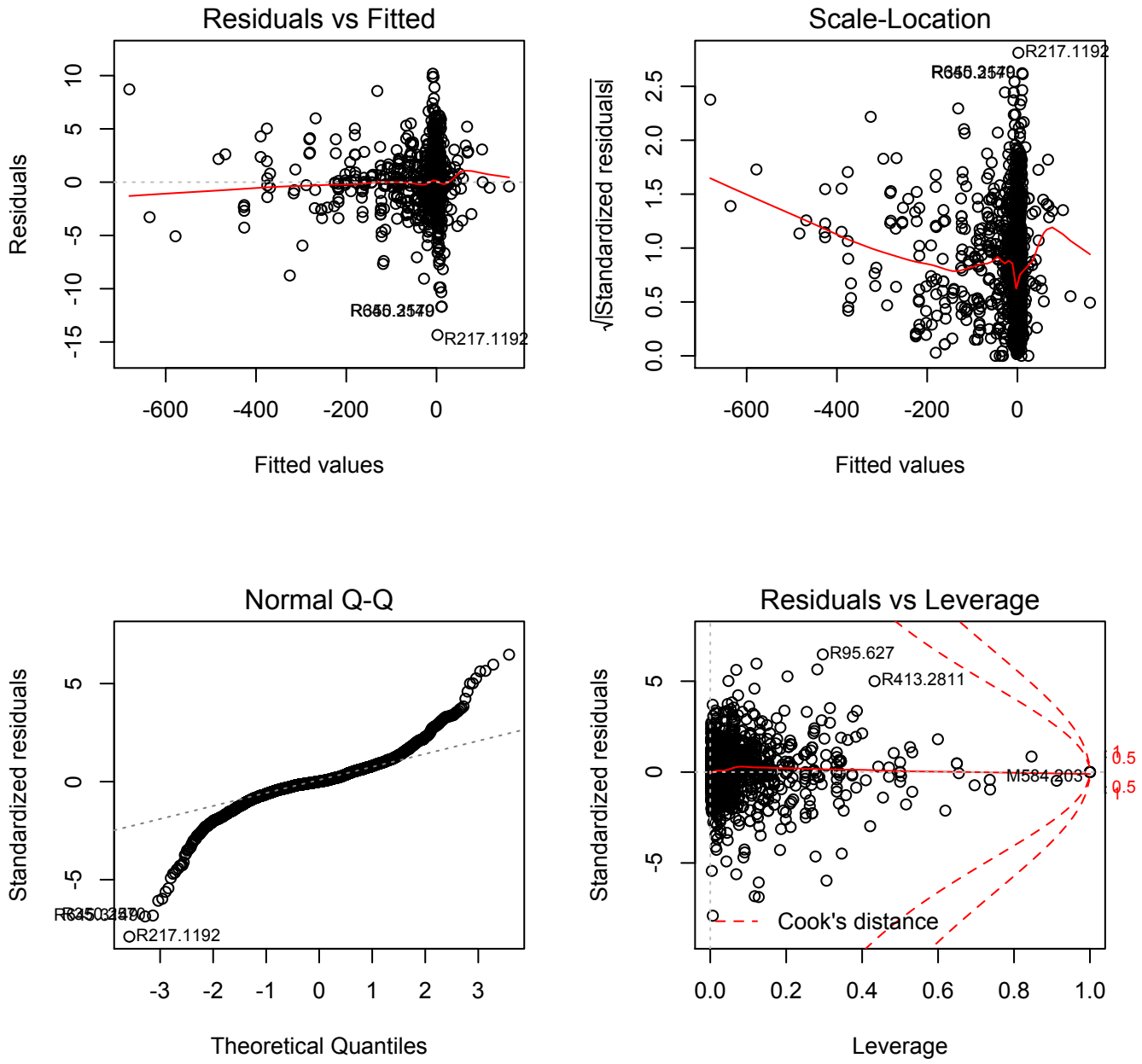



Figure D.1: Standard multiple linear regression diagnostic graphs

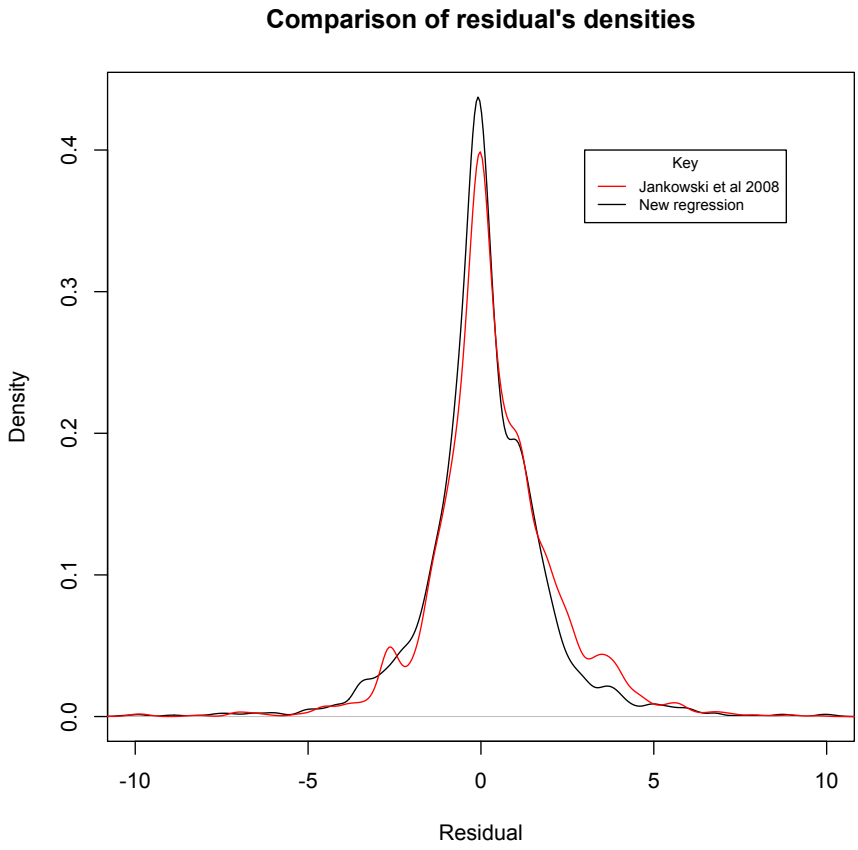



Figure D.2: Comparison of residuals distributions for the regression provided by [5] and the new one provided in this work.

in [5]. As we don't have the exact linear model used by these authors (it could be derived from their data, but might well not be the exact same model), it is difficult to do a proper comparison of the models.

## D. 3 Validation of Standard Transformed Gibbs Energies

Calculated Standard Transformed Gibbs Energies obtained through the group contribution method and the procedure described with equations 4 and 5 in the paper were compared against the values provided by [1], derived from experimentally measured values. Although most of the differences $|\Delta_f G'_{Alberty} - \Delta_f G'_{GroupContrib}|$ were in the range of 0 to 10 kcal/mol, as shown in Fig. D.3, there are a number of entities that show big differences between the transformed energy derived from the regression and from experimental values as described by [1].

A closer look into the values with higher differences (above 100 kcal/mol), shows that most of them, if not all, correspond to chemical entities for which conventions are defined for their base transformed energies (or chemical entities closely related to them). This can be seen in Fig. D.4, were it can be appreciated that related chemical entities have similar differences when comparing the Alberty given Transformed Gibbs Energy value compared to the one obtained from our regression (and posterior correction as explained using a method derived from the equations presented by [1]), which means probably that most of the observed difference is due

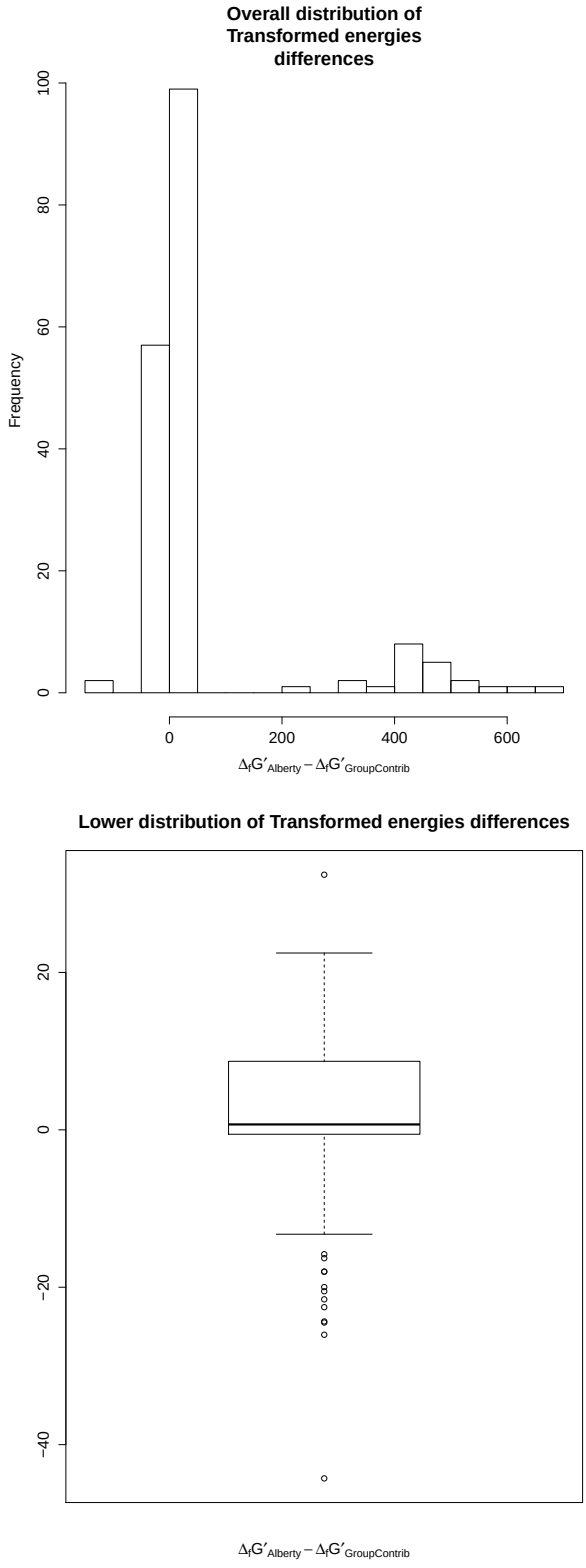



Figure D.3: Distribution of Transformed energies differences between values obtained from the Alberty catalogue and from Group contribution.

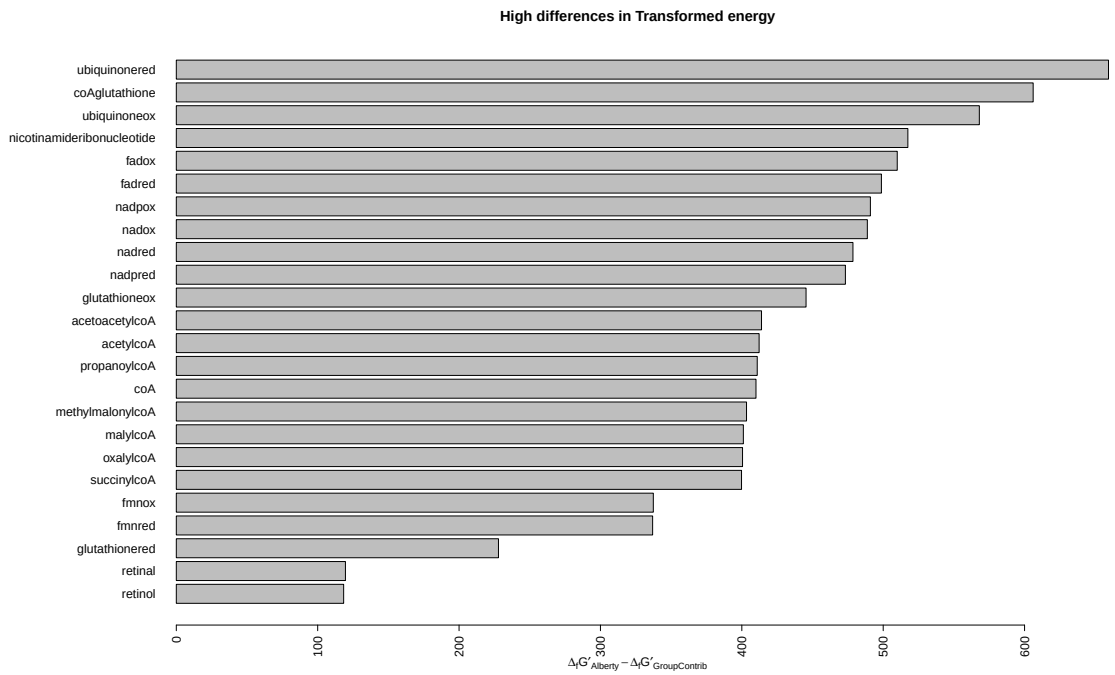



Figure D.4: Chemical entities with high differences in Standard Transformed Gibbs Energies: All the entities that showed a high energy difference in Transformed energy between the group contribution method and the Alberty catalogue have their base energies either defined by convention in zero or depend directly on an entity with such a convention.

to the convention of value taken by Alberty in his tables. Table D.1 shows which of these species form part of the convention or are directly connected by a reaction to one of these convention-set energy values.

Although this is a note of alert to use Transformed Gibbs energy values with caution, revision of the Transformed energies of reactions where these chemical entities[3] participate shows that the energy balance cancels out this error, as the effect of the conventions are cancelled out in the subtraction of energies between products and reactants. Hence the differences seen for these compounds, with energy values set by convention or closely related, doesn't translate into errors for the Transformed Gibbs Energies of reaction. This can be seen in Fig. D.5 , where it can be seen that most of the differences between Transformed Gibbs Energies of reaction calculated through the Alberty catalogue or through our regression lie within -5,5 kcal/mol, most of them very close to zero. Furthermore, the two small peaks that can be seen below -5 kcal/mol and above 5 kcal/mol correspond to variations introduced by phosphate groups, which deviate in our regression by 5.9 kcal/mol of Standard Gibbs Energy from the value found experimentally.

## D. 4 Reaction balancing

When balancing reactions, many times an acceptable result can be achieved from the perspective of the mass and charge balance by adding balancing species and changing stoichiometric coefficients. However, if not done carefully, this can lead to problems in the calculation of any property of the reaction that depends on this balance. The excesive augmentation of coefficient

[3]This included all the small molecules that could be found in the Alberty catalogue and all the 471 reactions approximately were at least one of the participated.

| Entity | Std. GE. [kJ/mol] | Conv. | Distance |
|---|---|---|---|
| Ubiquinone red. | -89.92 | No | 1 |
| Ubiquinone ox. | 0.00 | Yes | 1 |
| Nicotinamide ribonuc. | 840.08 | No | 1 |
| FAD ox. | 0.00 | Yes | 0 |
| FAD red. | -38.88 | No | 1 |
| NADP ox. | 0.00 | Yes | 0 |
| NAD ox. | 0.00 | Yes | 0 |
| NAD red. | 22.65 | No | 1 |
| NADP red | -809.19 | No | 1 |
| Glutathione ox. | 0.00 | Yes | 1 |
| Glutathione-coA | -35.85 | No | 1 |
| Glutathione Red. | 34.17 | No | 1 |
| Acetoacetyl CoA | -285.32 | No | 1 |
| Acetyl CoA | -188.52 | No | 1 |
| Propanoyl CoA | -179.14 | No | 1 |
| CoA | 0.00 | Yes | 0 |
| MethylmalonylCoA | -502.48 | No | 1 |
| MalylCoA | -663.44 | No | 1 |
| OxalylCoA | -509.96 | No | 1 |
| SuccinylCoA | -509.59 | No | 1 |
| FMN ox. | 0.00 | Yes | 0 |
| FMN red. | -38.88 | No | 1 |
| Retinal | 0.00 | Yes | 0 |
| Retinol | -27.91 | No | 1 |

Table D.1: Chemical entities with high differences: most of the entities are either defined by convention to zero or are directly related to them.

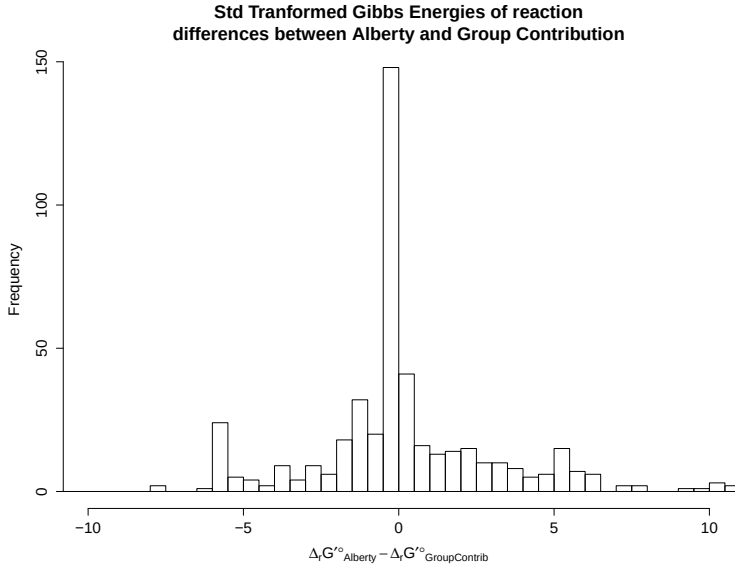



Figure D.5: Distribution of differences in Transformed energies between using the catalogue of Alberty and our regression.

or the addition of really unnecessary entities adds more groups than required in the calculation of group contributions. Hence, balancing must be done step by step and always observing how the overall coefficient sum changes, as big changes will produce big deviations from the real $\Delta_r G$. We suggest the following thumb rules:

**Define the purpose** It should be very clear for what is the balance required. In our case, it depends on whether the Standard or Standard Transformed Gibbs energy (or even Further Transformed) is going to be calculated. In the first case, Hydrogens need to be balanced, in the second case, they are neglected. In general it is very easy to introduce errors in the Standard Gibbs energy of reaction calculation because of the need of balancing protons, which is something that biochemical reactions don't really do.

**Balancing additions** Balancing molecules should be added one at a time, to avoid adding more than strictly necessary. Adding $H_2O$, $O_2$ and H at the same time can lead to results in which all of these are added without a real need.

**Minimize change** Always start by avoiding changes in the original reactant's and product's coefficients, and only allowing changes on the balancing additions. Once this doesn't work, then try with bounded changes in the coefficients of reactants and products.

**Nature's preference** In general, and observing databases of biochemical reactions, it can be seen that biochemical reactions rarely tend to deviate from one-to-one reactant to product ratios.

# Bibliography


[1] R. A. Alberty. *Thermodynamics of biochemical reactions*. John Wiley & Sons, June 2003.

[2] R. A. Alberty. *Biochemical thermodynamics: applications of Mathematica*. John Wiley & Sons, January 2006.

[3] R. N. Goldberg, Y. B. Tewari, and T. N. Bhat. Thermodynamics of enzyme-catalyzed reactions–a database for quantitative biochemistry. *Bioinformatics (Oxford, England)*, **20**[16]:2874–2877, November 2004.

[4] R. N. Goldberg, Y. B. Tewari, and T. N. Bhat. Thermodynamics of Enzyme-Catalyzed Reactions: Part 7—2007 Update. *Journal of Physical and Chemical Reference Data*, **36**[4]:1347, October 2007.

[5] M. D. Jankowski, C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis. Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks. *Biophysical journal*, **95**[3]:1487–1499, May 2008.

[6] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E. L. Willighagen. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Current pharmaceutical design*, **12**[17]:2111–2120, January 2006.

# D. GIBBS ENERGY CALCULATOR SUPPLEMENTARY MATERIAL

# Appendix E

# Permissions of use

This appendix contains information regarding permissions obtained for the use of Figures 1.3, 4.17, and 4.18, which are originals of other publications.

For the use of Figure 1.3, I obtained a license for electronic and paper distribution to be used solely within this thesis, through the Rightslink service. License number 2961871209474, dated Aug 04, 2012. The license requires that the following text is included in the thesis:

"Reprinted from TrAC Trends in Analytical Chemistry, Vol 26(9), S. Moco, J. Vervoort, R. Bino, R. De Vos, Metabolomics Technologies and metabolite identification, Pages No. 855-866, Copyright 2007, with permission from ELSEVIER".

For the use of Figure 4.18, I obtained a license for electronic and paper distribution to be used solely within this thesis, through the Rightslink service. License number 2942061417749, dated Jul 04, 2012. The license requires that the following text is included in the thesis:

"Reprinted by permission from Macmillan Publisher Ltd: Nature Biotechnology, T. Nguyen, K. Ishida, H. Jenke-Kodama, E. Dittmann, C. Gurgui, T.

Hochmuth, S. Taudien, M. Platzer, C. Hertweck, and J. Piel. Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. Nature Biotechnology, 26[2]:225233, February 2008, copyright 2008."

Figure 4.17 is part of a [128], which is a paper published in *Proceedings to the National Academy of Sciences of the United States of America*, which states in its "Rights and Permissions"[1] that:

"Anyone may, without requesting permission, use original figures or tables published in PNAS for noncommercial and educational use (i.e., in a review article, in a book that is not for sale) provided that the original source and the applicable copyright notice are cited."

It also states that:

"Authors whose work will be reused should be notified. PNAS cannot supply original artwork. Use of PNAS material must not imply any endorsement by PNAS or the National Academy of Sciences. The full journal reference must be cited and, for articles published in Volumes 90105 (19932008), "Copyright (copyright year) National Academy of Sciences, USA." "

I wrote to Prof. Chaita Khosla at Stanford University, asking for permission to use the Figure, to which he agreed in an e-mail dated July 4, 2012 16:53:40 GMT+01:00, from his University email address khosla@stanford.edu. The text requested by PNAS was included in the Figure and the paper is cited.

[1] http://www.pnas.org/site/misc/rightperm.shtml

# Appendix F

# Text mining supplementary material

## F. 1 Tissues related to some example metabolites

This section includes additional examples as the one shown for Kynurenic acid in Table 3.3, page 131. This part distinguishes two types of small molecules in terms of their expected distribution across tissues: widespread molecules and specific molecules.

In the case of widespread molecules, such Pyruvate (Table F.1), Taurine (Table F.2), and Lactate (Table F.3), the respective tables show that they do not have a strong bias to any particular major physiological system. On the contrary, specific molecules, such as GABA (Table F.6), Pregnenolone (Table F.5), and Ursodeoxycholic acid (Table F.4), show through their tables well defined

biases towards particular body systems.

| Sample | Obs. | Exp. | Total | MiM | LLH |
|---|---|---|---|---|---|
| Liver | 1432 | 201.71 | 4.9E5 | 1.38 | 3230 |
| Culture medium | 1094 | 112.95 | 2.8E5 | 1.83 | 3056 |
| Blood | 1451 | 378.52 | 9.2E5 | 0.49 | 1812 |
| Heart | 870 | 198.63 | 4.9E5 | 0.67 | 1250 |
| Hepatocyte | 330 | 24.53 | 6.0E4 | 2.30 | 1111 |
| Muscle | 623 | 136.18 | 3.3E5 | 0.74 | 933 |
| Cerebral ganglion | 673 | 202.82 | 5.0E5 | 0.28 | 685 |
| Skeletal muscle | 249 | 30.15 | 7.4E4 | 1.60 | 617 |
| Adipocyte | 115 | 9.02 | 2.2E4 | 2.22 | 375 |
| Adipose tissue | 117 | 11.38 | 2.8E4 | 1.91 | 335 |
| Blastocyst | 83 | 4.79 | 1.2E4 | 2.67 | 318 |
| Kidney | 270 | 89.78 | 2.2E5 | 0.14 | 236 |
| Skin fibroblast | 54 | 3.51 | 8.6E3 | 2.50 | 195 |
| Cardiac muscle | 65 | 6.40 | 1.6E4 | 1.90 | 184 |
| Mesophyll | 34 | 0.92 | 2.3E3 | 3.76 | 180 |
| Embryo | 161 | 47.27 | 1.2E5 | 0.30 | 168 |
| Erythrocyte | 147 | 40.52 | 9.9E4 | 0.41 | 167 |
| Astrocyte | 74 | 10.77 | 2.6E4 | 1.33 | 159 |
| Pectoral muscle | 29 | 0.95 | 2.3E3 | 3.49 | 143 |

Table F.1: Table of the 20 Best ranked biological samples (Brenda Tissue Ontology[49] entries) co-occurring with **Pyruvate**, a metabolite relevant for central metabolism, particularly glycolysis and energy metabolism. In this case, results are sorted by Log Likelihood and are asked to have mutual information score $MiM > 0$, log likelihood score $LLH > 10$ and $t - Score > 0$. The column Obs. stands for observed number of co-occurrences (between the sample and Pyruvate), Exp. for expected number of co-occurrences and Total for the total number of co-occurrences of that sample (Brenda Tissue Ontology entry) with all other small molecules. In this case there is no strong bias towards a particular major system of the mammalian body.

| Sample | Obs. | Exp. | Total | MiM | LLH |
|---|---|---|---|---|---|
| Cerebral ganglion | 644 | 82.92 | 5.0E5 | 1.51 | 1556 |
| Excretion | 289 | 17.79 | 1.1E5 | 2.58 | 1078 |
| Retina | 219 | 7.86 | 4.7E4 | 3.35 | 1041 |
| Liver | 467 | 82.47 | 4.9E5 | 1.05 | 868 |
| Culture medium | 346 | 46.18 | 2.8E5 | 1.45 | 805 |
| Hippocampus | 179 | 9.24 | 5.5E4 | 2.83 | 725 |
| Astrocyte | 124 | 4.40 | 2.6E4 | 3.37 | 591 |
| Animal | 411 | 105.75 | 6.3E5 | 0.51 | 517 |
| Cerebellum | 116 | 5.70 | 3.4E4 | 2.90 | 480 |
| Neuron | 226 | 38.84 | 2.3E5 | 1.09 | 426 |
| Heart | 322 | 81.21 | 4.9E5 | 0.53 | 413 |
| Corpus striatum | 95 | 5.13 | 3.1E4 | 2.77 | 376 |
| Cerebral cortex | 96 | 5.72 | 3.4E4 | 2.62 | 362 |
| Glia | 70 | 4.41 | 2.6E4 | 2.54 | 257 |
| Photoreceptor | 59 | 2.86 | 1.7E4 | 2.92 | 245 |
| Central nervous system | 127 | 22.85 | 1.4E5 | 1.02 | 229 |
| Hepatocyte | 78 | 10.03 | 6.0E4 | 1.51 | 185 |
| Retinal pigment epithelium | 38 | 1.32 | 7.9E3 | 3.41 | 183 |
| Cardiac myocyte | 50 | 3.49 | 2.1E4 | 2.39 | 173 |
| Neutrophil | 81 | 11.96 | 7.1E4 | 1.32 | 172 |

Table F.2: Table of the best ranked biological samples (Brenda Tissue Ontology[49] entries) co-occurring with **Taurine**, an organic acid found across many different animal tissues, and an important constituent of bile acid. It is important to notice that this results are not limited to a particular taxonomic range. Results are sorted by Log Likelihood and are asked to have mutual information score $MiM > 0$, log likelihood score $LLH > 10$ and $t - Score > 0$. These boundaries limit the results to only 15 samples. The column Obs. stands for observed number of co-occurrences (between the sample and Taurine), Exp. for expected number of co-occurrences and Total for the total number of co-occurrences of that sample (Brenda Tissue Ontology entry) with all other small molecules. In this case there is no strong bias for a particular major animal/plant body system, although nearly half of the entries are related to the nervous system.

| Sample | Obs. | Exp. | Total | MiM | LLH |
|---|---|---|---|---|---|
| Blood | 10393 | 1075.66 | 9.2E5 | 1.82 | 30162 |
| Heart | 3420 | 564.47 | 4.9E5 | 1.15 | 6767 |
| Muscle | 2868 | 387.00 | 3.3E5 | 1.44 | 6650 |
| Culture medium | 2017 | 320.99 | 2.8E5 | 1.20 | 4081 |
| Liver | 2579 | 573.22 | 4.9E5 | 0.73 | 3822 |
| Hepatocyte | 768 | 69.70 | 6.0E4 | 2.02 | 2306 |
| Cerebral ganglion | 1913 | 576.37 | 5.0E5 | 0.28 | 1951 |
| Skeletal muscle | 735 | 85.67 | 7.4E4 | 1.65 | 1874 |
| Erythrocyte | 558 | 115.14 | 9.9E4 | 0.82 | 881 |
| Astrocyte | 272 | 30.61 | 2.6E4 | 1.70 | 709 |
| Stolon | 144 | 5.80 | 5.0E3 | 3.19 | 653 |
| Leukocyte | 455 | 112.77 | 9.7E4 | 0.57 | 588 |
| Pleural fluid | 94 | 4.39 | 3.8E3 | 2.98 | 399 |
| Fast muscle | 78 | 2.39 | 2.1E3 | 3.59 | 396 |
| Vein | 416 | 136.12 | 1.2E5 | 0.16 | 372 |
| Sertoli cell | 105 | 9.08 | 7.8E3 | 2.08 | 323 |
| Excretion | 371 | 123.69 | 1.1E5 | 0.14 | 322 |
| Neutrophil | 284 | 83.13 | 7.1E4 | 0.33 | 297 |
| Semen | 235 | 65.41 | 5.6E4 | 0.39 | 263 |
| Cardiac myocyte | 138 | 24.29 | 2.1E4 | 1.05 | 253 |

Table F.3: Table of the highest 20 ranked biological samples (Brenda Tissue Ontology[49] entries) co-occurring with **Lactate**, a widespread metabolite, closely linked metabolically to Pyruvate. Results are sorted by Log Likelihood and are asked to have mutual information score $MiM > 0$, log likelihood score $LLH > 10$ and $t{-}Score > 0$. The column Obs. stands for observed number of co-occurrences (between the sample and Lactate), Exp. for expected number of co-occurrences and Total for the total number of co-occurrences of that sample (Brenda Tissue Ontology entry) with all other small molecules. In this case there is no clear bias for a particular major animal/plant body system.

| Sample | Obs. | Exp. | Total | MiM | LLH |
|---|---|---|---|---|---|
| Liver | 566 | 25.81 | 4.9E5 | 3.01 | 2533 |
| Gall bladder | 122 | 1.36 | 2.6E4 | 5.04 | 862 |
| Hepatocyte | 56 | 3.14 | 6.0E4 | 2.71 | 218 |
| Bile duct | 44 | 1.67 | 3.2E4 | 3.27 | 204 |
| Colon | 43 | 4.58 | 8.7E4 | 1.78 | 116 |
| Secretion | 65 | 11.92 | 2.3E5 | 1.00 | 115 |
| Excretion | 28 | 5.57 | 1.1E5 | 0.88 | 46 |
| Intestine | 28 | 5.89 | 1.1E5 | 0.79 | 43 |
| Biliary epithelial cell | 4 | 0.04 | 6.7E2 | 5.38 | 30 |
| Hep-G2 cell | 5 | 0.34 | 6.6E3 | 2.41 | 17 |
| Peripheral blood cell | 4 | 0.22 | 4.1E3 | 2.76 | 16 |
| Ileum | 7 | 1.21 | 2.3E4 | 1.08 | 13 |
| Duodenum | 7 | 1.22 | 2.3E4 | 1.08 | 13 |
| Small intestine | 9 | 2.11 | 4.0E4 | 0.64 | 12 |
| Jejunum | 5 | 0.76 | 1.5E4 | 1.26 | 10 |

Table F.4: Table of the best ranked biological samples (Brenda Tissue Ontology[49] entries) co-occurring with **Ursodeoxycholic acid**, a bile acid found in the bile of bears as conjugate with Taurine, which is used therapeutically. Results are sorted by Log Likelihood and are asked to have mutual information score $MiM > 0$, log likelihood score $LLH > 10$ and $t - Score > 0$. These boundaries limit the results to only 15 samples. The column Obs. stands for observed number of co-occurrences (between the sample and Ursodeoxycholic acid), Exp. for expected number of co-occurrences and Total for the total number of co-occurrences of that sample (Brenda Tissue Ontology entry) with all other small molecules. In this case there is a bias towards components of the digestive system.

| Sample | Obs. | Exp. | Total | MiM | LLH |
|---|---|---|---|---|---|
| Adrenal gland | 306 | 4.84 | 7.5E4 | 4.54 | 1965 |
| Leydig cell | 86 | 0.77 | 1.2E4 | 5.37 | 644 |
| Testis | 77 | 2.63 | 4.1E4 | 3.42 | 373 |
| Adrenal cortex | 54 | 0.71 | 1.1E4 | 4.80 | 362 |
| Placenta | 72 | 2.33 | 3.6E4 | 3.50 | 356 |
| Granulosa cell | 51 | 0.59 | 9.1E3 | 5.00 | 356 |
| Ovary | 69 | 4.15 | 6.5E4 | 2.61 | 260 |
| Secretion | 110 | 14.62 | 2.3E5 | 1.46 | 256 |
| Lutein cell | 28 | 0.17 | 2.6E3 | 5.96 | 232 |
| Culture medium | 109 | 17.71 | 2.8E5 | 1.18 | 216 |
| Cerebral ganglion | 131 | 31.81 | 5.0E5 | 0.59 | 176 |
| Zona glomerulosa | 19 | 0.15 | 2.4E3 | 5.50 | 145 |
| COS-1 cell | 17 | 0.15 | 2.4E3 | 5.34 | 126 |
| Corpus luteum | 21 | 0.42 | 6.5E3 | 4.21 | 124 |
| Zona fasciculata | 11 | 0.08 | 1.2E3 | 5.74 | 88 |
| Y-1 cell | 9 | 0.03 | 4.8E2 | 6.76 | 85 |
| MA-10 cell | 8 | 0.02 | 2.4E2 | 7.57 | 84 |
| Chorion | 11 | 0.16 | 2.5E3 | 4.63 | 71 |
| Theca cell | 8 | 0.07 | 1.2E3 | 5.31 | 59 |
| Hippocampus | 24 | 3.55 | 5.5E4 | 1.30 | 51 |

Table F.5: Table of the highest 20 ranked biological samples (Brenda Tissue Ontology[49] entries) co-occurring with **Pregnenolone**, a human endogenous steroid hormone. It is the initial precursor to different prostagens, androgens, mineralocorticoids, glucocorticoids, and estrogens. Results are sorted by Log Likelihood and are asked to have mutual information score $MiM > 0$, log likelihood score $LLH > 10$ and $t - Score > 0$. The column Obs. stands for observed number of co-occurrences (between the sample and Pregnenolone), Exp. for expected number of co-occurrences and Total for the total number of co-occurrences of that sample (Brenda Tissue Ontology entry) with all other small molecules. In this case there is are biases towards components of the reproductive system and nervous system components.

| Sample | Obs. | Exp. | Total | MiM | LLH |
|---|---|---|---|---|---|
| Neuron | 6777 | 174.46 | 2.3E5 | 3.83 | 37785 |
| Cerebral ganglion | 5373 | 372.41 | 5.0E5 | 2.41 | 19413 |
| Hippocampus | 1219 | 41.51 | 5.5E4 | 3.43 | 5947 |
| Pyramidal neuron | 812 | 12.48 | 1.7E4 | 4.58 | 5237 |
| Central nervous system | 1372 | 102.63 | 1.4E5 | 2.29 | 4630 |
| Cerebral cortex | 872 | 25.70 | 3.4E4 | 3.64 | 4493 |
| Substantia nigra | 650 | 10.60 | 1.4E4 | 4.49 | 4112 |
| Corpus striatum | 774 | 23.03 | 3.1E4 | 3.63 | 3972 |
| Retina | 861 | 35.31 | 4.7E4 | 3.16 | 3881 |
| Spinal cord | 845 | 62.03 | 8.3E4 | 2.32 | 2871 |
| Cerebellum | 623 | 25.60 | 3.4E4 | 3.16 | 2802 |
| Granule cell | 393 | 6.70 | 8.9E3 | 4.43 | 2449 |
| Globus pallidus | 308 | 7.06 | 9.4E3 | 4.00 | 1736 |
| Cerebellar Purkinje cell | 304 | 6.98 | 9.3E3 | 4.00 | 1713 |
| Nerve | 940 | 169.46 | 2.3E5 | 1.02 | 1698 |
| Amygdala | 343 | 14.01 | 1.9E4 | 3.17 | 1544 |
| Thalamus | 315 | 15.50 | 2.1E4 | 2.90 | 1305 |
| Glia | 330 | 19.80 | 2.6E4 | 2.61 | 1242 |
| Neocortex | 224 | 7.36 | 9.8E3 | 3.48 | 1103 |
| Ganglion | 279 | 15.76 | 2.1E4 | 2.70 | 1082 |

Table F.6: Table of the highest 20 ranked biological samples (Brenda Tissue Ontology[49] entries) co-occurring with **GABA**, a neurotransmitter. Results are sorted by Log Likelihood and are asked to have mutual information score $MiM > 0$, log likelihood score $LLH > 10$ and $t - Score > 0$. The column Obs. stands for observed number of co-occurrences (between the sample and GABA), Exp. for expected number of co-occurrences and Total for the total number of co-occurrences of that sample (Brenda Tissue Ontology entry) with all other small molecules. In this case there is a strong bias towards components of the nervous system.

# Appendix G

# Constrained chemical enumeration supplementary material

## G. 1 Reaction enumeration

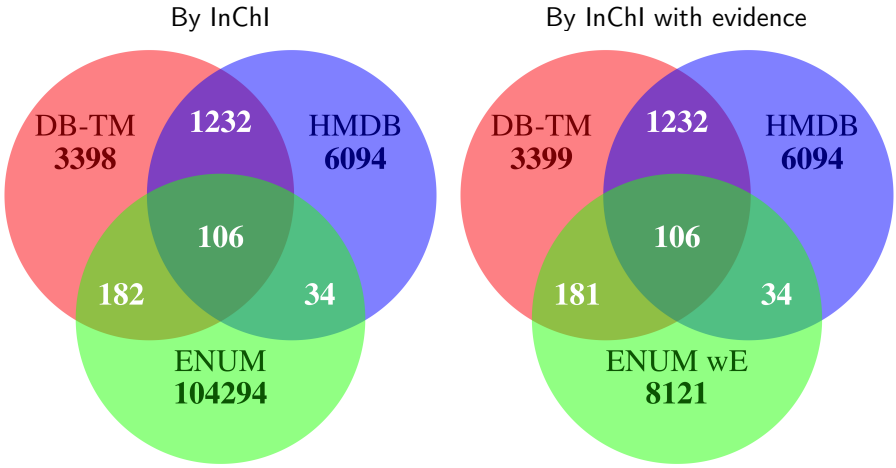
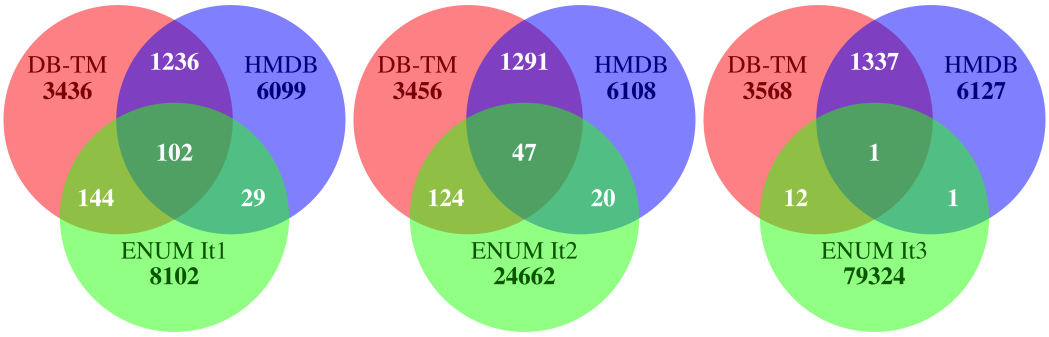



Figure G.1: Venn diagrams of small molecules produced by the database unification, text mining, HMDB and the reaction enumeration scheme, using Standard InChI. *Left:* Diagram built using the complete enumeration result. *Right:* Diagram built using the enumeration result limited to small molecules that show either similar connectivity to known molecules or predicted downstream reactions – which are labeled as "with evidence". Limiting results to these small molecules "with evidence", reduces the set of unique molecules generated by the enumeration to 8.1% of its original size, presumably to molecules with higher chances of being real.

**A** Joint DB unification-Text mining set, HMDB and molecules resulting from enumeration, separated by iteration.




**B** Joint DB unification-Text mining set, HMDB and molecules resulting from enumeration that show evidence, separated by iteration.

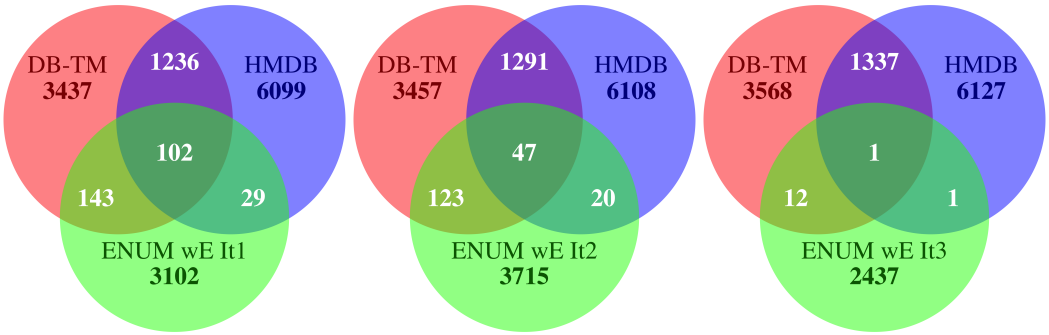



Figure G.2: *A:* Decomposition of the left Venn diagram in Figure G.1 by iterations of the enumeration. *B:* Decomposition of the right Venn diagram in Figure G.1 by iterations of the enumeration. This shows that Iterations 1 and 2 generate most of the intersections with the joint database unification and text mining set, and with the HMDB set, while Iteration 3 produces the higher ratio of unique to intersected molecules.

Table G.1: Number of occurrences (#) of EC numbers in the reactions of each of the 3 iterations, including the percentage (%) that each EC number covers – in terms of reactions – for each iteration. Few EC classified reactions – most of them oxidoreductases that generate more that ~70% of the enumerated reactions – dominate the enumeration. The whole process is represented by 57 EC numbers. EC numbers are sorted by the number of occurrences (#) in the third iteration.

| | Iteration 1 | | Iteration 2 | | Iteration 3 | |
|---|---|---|---|---|---|---|
| EC | # | % | # | % | # | % |
| 1.1.1.1 | 5819 | 49.2 | 21618 | 45.9 | 108623 | 52.8 |
| 1.1.1.2 | 3040 | 25.7 | 8967 | 19.1 | 37256 | 18.1 |
| 1.2.1.3 | 227 | 1.9 | 6498 | 13.8 | 27047 | 13.1 |
| 2.4.1.1 | 89 | 0.8 | 1243 | 2.6 | 7049 | 3.4 |
| 3.2.1.2 | 87 | 0.7 | 1048 | 2.2 | 6617 | 3.2 |
| 2.1.1.49 | 907 | 7.7 | 2464 | 5.2 | 5619 | 2.7 |
| 3.2.1.20 | 74 | 0.6 | 848 | 1.8 | 4940 | 2.4 |
| 1.1.1.21 | 229 | 1.9 | 666 | 1.4 | 2802 | 1.4 |
| 2.4.1.144 | 75 | 0.6 | 314 | 0.7 | 1378 | 0.7 |
| 6.2.1.3 | 389 | 3.3 | 457 | 1.0 | 761 | 0.4 |
| 3.5.1.4 | 54 | 0.5 | 156 | 0.3 | 537 | 0.3 |

Continued...

| | Iteration 1 | | Iteration 2 | | Iteration 3 | |
|---|---|---|---|---|---|---|
| EC | # | % | # | % | # | % |
| 2.8.2.– | 0 | 0.0 | 205 | 0.4 | 506 | 0.2 |
| 2.1.1.9 | 86 | 0.7 | 165 | 0.4 | 301 | 0.1 |
| 2.3.1.87 | 209 | 1.8 | 247 | 0.5 | 285 | 0.1 |
| 3.1.1.– | 0 | 0.0 | 35 | 0.1 | 224 | 0.1 |
| 2.4.1.212 | 0 | 0.0 | 27 | 0.1 | 145 | 0.1 |
| 5.1.3.17 | 0 | 0.0 | 44 | 0.1 | 140 | 0.1 |
| 5.1.3.19 | 0 | 0.0 | 44 | 0.1 | 140 | 0.1 |
| 1.3.1.74 | 67 | 0.6 | 196 | 0.4 | 126 | 0.1 |
| 2.8.2.17 | 0 | 0.0 | 43 | 0.1 | 121 | 0.1 |
| 2.8.2.35 | 0 | 0.0 | 44 | 0.1 | 109 | 0.1 |
| 2.8.2.5 | 0 | 0.0 | 42 | 0.1 | 108 | 0.1 |
| 2.4.1.69 | 70 | 0.6 | 161 | 0.3 | 99 | 0.0 |
| No EC | 136 | 1.2 | 179 | 0.4 | 97 | 0.0 |
| 2.4.1.38 | 0 | 0.0 | 34 | 0.1 | 69 | 0.0 |
| 2.8.2.33 | 0 | 0.0 | 0 | 0.0 | 67 | 0.0 |
| 3.2.1.21 | 75 | 0.6 | 125 | 0.3 | 65 | 0.0 |
| 2.3.1.81 | 3 | 0.0 | 21 | 0.0 | 56 | 0.0 |
| 2.8.2.8 | 1 | 0.0 | 5 | 0.0 | 55 | 0.0 |
| 5.2.1.8 | 15 | 0.1 | 58 | 0.1 | 51 | 0.0 |
| 2.4.1.37 | 0 | 0.0 | 78 | 0.2 | 48 | 0.0 |
| 2.4.1.40 | 0 | 0.0 | 66 | 0.1 | 40 | 0.0 |

Continued. . .

| | Iteration 1 | | Iteration 2 | | Iteration 3 | |
|---|---|---|---|---|---|---|
| EC | # | % | # | % | # | % |
| 3.1.6.4 | 0 | 0.0 | 0 | 0.0 | 35 | 0.0 |
| 3.1.6.12 | 0 | 0.0 | 0 | 0.0 | 34 | 0.0 |
| 2.4.1.143 | 0 | 0.0 | 7 | 0.0 | 27 | 0.0 |
| 2.4.1.179 | 0 | 0.0 | 16 | 0.0 | 16 | 0.0 |
| 2.4.1.152 | 0 | 0.0 | 0 | 0.0 | 16 | 0.0 |
| 1.11.1.12 | 11 | 0.1 | 7 | 0.0 | 15 | 0.0 |
| 1.11.1.15 | 9 | 0.1 | 7 | 0.0 | 15 | 0.0 |
| 2.8.2.30 | 1 | 0.0 | 2 | 0.0 | 11 | 0.0 |
| 2.4.1.101 | 0 | 0.0 | 4 | 0.0 | 8 | 0.0 |
| 3.5.1.23 | 0 | 0.0 | 0 | 0.0 | 8 | 0.0 |
| 3.1.6.14 | 0 | 0.0 | 0 | 0.0 | 7 | 0.0 |
| 3.6.3.47 | 27 | 0.2 | 5 | 0.0 | 4 | 0.0 |
| 2.3.2.1 | 0 | 0.0 | 5 | 0.0 | 4 | 0.0 |
| 1.17.4.2 | 4 | 0.0 | 0 | 0.0 | 4 | 0.0 |
| 3.1.2.2 | 1 | 0.0 | 14 | 0.0 | 2 | 0.0 |
| 2.3.1.5 | 18 | 0.2 | 5 | 0.0 | 2 | 0.0 |
| 2.4.2.30 | 3 | 0.0 | 4 | 0.0 | 2 | 0.0 |
| 1.1.1.145 | 2 | 0.0 | 1 | 0.0 | 2 | 0.0 |
| 2.4.1.145 | 0 | 0.0 | 0 | 0.0 | 2 | 0.0 |
| 2.7.1.60 | 0 | 0.0 | 1 | 0.0 | 1 | 0.0 |
| 5.1.3.8 | 0 | 0.0 | 1 | 0.0 | 1 | 0.0 |

Continued. . .

| | Iteration 1 | | Iteration 2 | | Iteration 3 | |
|---|---|---|---|---|---|---|
| EC | # | % | # | % | # | % |
| 2.7.8.27 | 70 | 0.6 | 882 | 1.9 | 0 | 0.0 |
| 2.7.8.2 | 14 | 0.1 | 2 | 0.0 | 0 | 0.0 |
| 2.7.8.1 | 11 | 0.1 | 2 | 0.0 | 0 | 0.0 |
| 2.7.1.94 | 1 | 0.0 | 1 | 0.0 | 0 | 0.0 |
| 3.1.3.4 | 1 | 0.0 | 0 | 0.0 | 0 | 0.0 |

# G. CONSTRAINED CHEMICAL ENUMERATION SUPPLEMENTARY MATERIAL

# References


[1] R. A. Alberty. *Thermodynamics of biochemical reactions*. John Wiley & Sons, June 2003. 76, 183

[2] R. A. Alberty. *Biochemical thermodynamics: applications of Mathematica*. John Wiley & Sons, January 2006. 183

[3] R. Alcántara, K. B. Axelsen, A. Morgat, E. Belda, E. Coudert, A. Bridge, H. Cao, P. de Matos, M. Ennis, S. Turner, G. Owen, L. Bougueleret, I. Xenarios, and C. Steinbeck. Rhea–a manually curated resource of biochemical reactions. *Nucleic Acids Research*, **40**[D1]:D754–D760, January 2012. 25, 107

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, **215**[3]:403–410, October 1990. 210

[5] A. Bairoch. The ENZYME database in 2000. *Nucleic Acids Research*, **28**[1]:304–305, January 2000. 26

[6] R. Bino, R. Hall, O. Fiehn, J. Kopka, K. Saito, J. Draper, B. Nikolau, P. Mendes, U. Roessnertunali, and M. Beale. Po-

tential of metabolomics as a functional genomics tool. *Trends in Plant Science*, **9**[9]:418–425, September 2004. 9

[7] C. Bishop. *Pattern Recognition And Machine Learning*. Information Science and Statistics. Springer, 2006. 145

[8] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, **32**[90001]:267D–270, January 2004. 114

[9] N. P. Brown, C. Leroy, and C. Sander. MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics (Oxford, England)*, **14**[4]:380–381, 1998. xxxiii, 209

[10] W. Busch and M. H. Saier. The transporter classification (TC) system, 2002. *Crit. Rev. Biochem. Mol. Biol.*, **37**:287–337, 2002. 19

[11] R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang, and P. D. Karp. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, **36**[Database issue]:D623–31, January 2008. 20, 107

[12] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, **39**[Database]:D685–D690, December 2010. 24

[13] E. G. Cerami, G. D. Bader, B. E. Gross, and C. Sander. cPath: open source software for collecting, storing, and querying biological pathways. *BMC Bioinformatics*, **7**[1]:497, 2006. 25

[14] M. Chagoyen and F. Pazos. MBRole: enrichment analysis of metabolomic data. *Bioinformatics (Oxford, England)*, **27**[5]:730–731, February 2011. 82

[15] G. L. Challis, J. Ravel, and C. A. Townsend. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chemistry & Biology*, **7**[3]:211–224, March 2000. 207, 212

[16] A. Chang, M. Scheer, A. Grote, I. Schomburg, and D. Schomburg. BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Research*, **37**[Database]:D588–D592, January 2009. 107

[17] H. Chen and P. C. Boutros. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics*, **12**:35, 2011. xix, 73

[18] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, and D. S. Wishart. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research*, **36**[Web Server issue]:W399–405, July 2008. 111

[19] U. Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research*, **38**[Database issue]:D142–8, January 2010. xiii, 2, 14

[20] R. G. Côté, P. Jones, R. Apweiler, and H. Hermjakob. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, **7**:97, 2006. 38

[21] D. J. Creek, A. Jankevics, R. Breitling, D. G. Watson, M. P. Barrett, and K. E. V. Burgess. Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. *Analytical Chemistry*, **83**[22]:8703–8710, November 2011. 8

[22] F. Csizmadia. JChem: Java applets and modules supporting chemical database handling from web browsers. *Journal of Chemical Information and Computer Sciences*, **40**[2]:323–324, March 2000. 27, 54

[23] Q. Cui, I. A. Lewis, A. D. Hegeman, M. E. Anderson, J. Li, C. F. Schulte, W. M. Westler, H. R. Eghbalnia, M. R. Sussman, and J. L. Markley. Metabolite identification via the Madison Metabolomics Consortium Database. *Nature Biotechnology*, **26**[2]:162–164, February 2008. 9

[24] M. de Groot, R. van Berlo, W. van Winden, P. Verheijen, M. Reinders, and D. de Ridder. Metabolite and reaction inference based on enzyme specificities. *Bioinformatics (Oxford, England)*, **25**[22]:2975, November 2009. 38, 173

[25] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, **36**[Database issue]:D344–50, January 2008. 5, 17

[26] K. Dettmer, P. A. Aronov, and B. D. Hammock. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, **26**[1]:51–78, January 2007. xiii, 1, 2, 7, 9

[27] P. D. Dobson and D. B. Kell. Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nature Reviews Drug Discovery*, **7**[3]:205–220, March 2008. 4

[28] P. D. Dobson, Y. Patel, and D. B. Kell. 'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discovery Today*, **14**[1-2]:31–40, January 2009. 142, 144

[29] W. B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J. D. Knowles, A. Halsall, J. N. Haselden, A. W. Nicholls, I. D. Wilson, D. B. Kell, and R. Goodacre. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols*, **6**[7]:1060–1083, June 2011. 8

[30] W. B. Dunn, D. I. Broadhurst, H. J. Atherton, R. Goodacre, and J. L. Griffin. Systems level studies of mammalian metabolomes: the

roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews*, **40**[1]:387–426, 2010. 8

[31] P. Ertl, S. Roggo, and A. Schuffenhauer. Natural product-likeness score and its application for prioritization of compound libraries. *Journal of Chemical Information and Modeling*, **48**[1]:68–74, January 2008. 144

[32] T. L. Erwin. THE TROPICAL FOREST CANOPY: The Heart of Biotic Diversity. In E. Osborne Wilson and F. M. Peter, editors, *Biodiversity, Part 3*. National Academy Press, Washington, D.C., February 2003. 5

[33] S. Evert. *The statistics of word cooccurrences : word pairs and collocations*. PhD thesis, University of Stuttgart, Germany, August 2004. 124, 128

[34] E. Fahy, S. Subramaniam, H. A. Brown, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. H. Raetz, D. W. Russell, Y. Seyama, W. Shaw, T. Shimizu, F. Spener, G. van Meer, M. S. VanNieuwenhze, S. H. White, J. L. Witztum, and E. A. Dennis. A comprehensive classification system for lipids. *Journal of Lipid Research*, **46**[5]:839–861, May 2005. 122

[35] J. L. Faulon, D. P. Visco, and R. S. Pophale. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *Journal of Chemical Information and Modeling*, **43**[3]:707–720, May 2003. 179

[36] S. Federhen. The NCBI Taxonomy database. *Nucleic Acids Research*, **40**[Database Issue]:D136–D143, December 2011. 14, 24

[37] O. Fiehn. Metabolomics–the link between genotypes and phenotypes. *Plant Molecular Biology*, **48**[1-2]:155–171, January 2002. 2, 6

[38] O. Fiehn, J. Kopka, P. Dörmann, T. Altmann, R. N. Trethewey, and L. Willmitzer. Metabolite profiling for plant functional genomics. *Nature Biotechnology*, **18**[11]:1157–1161, November 2000. 8

[39] O. Fiehn, D. Robertson, J. Griffin, M. Werf, B. Nikolau, N. Morrison, L. W. Sumner, R. Goodacre, N. W. Hardy, C. Taylor, J. Fostel, B. Kristal, R. Kaddurah-Daouk, P. Mendes, B. Ommen, J. C. Lindon, and S.-A. Sansone. The metabolomics standards initiative (MSI). *Metabolomics*, **3**[3]:175–178, August 2007. 9

[40] O. Fiehn, G. Wohlgemuth, M. Scholz, T. Kind, D. Y. Lee, Y. Lu, S. Moon, and B. Nikolau. Quality control for plant metabolomics: reporting MSI-compliant studies. *The Plant Journal*, **53**[4]:691–704, February 2008. 3

[41] L. Fieseler, U. Hentschel, L. Grozdanov, A. Schirmer, G. Wen, M. Platzer, S. Hrvatin, D. Butzke, K. Zimmermann, and J. Piel. Widespread Occurrence and Genomic Context of Unusually Small Polyketide Synthase Genes in Microbial Consortia Associated with Marine Sponges. *Applied and Environmental Microbiology*, **73**[7]:2144, April 2007. 206

[42] R. D. Finn, J. Clements, and S. R. EDDY. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, **39**[Web Server issue]:W29–37, July 2011. 218

[43] A. Fleischmann, M. Darsow, K. Degtyarenko, W. Fleischmann, S. Boyce, K. B. Axelsen, A. Bairoch, D. Schomburg, K. F. Tipton, and R. Apweiler. IntEnz, the integrated relational enzyme database. *Nucleic Acids Research*, **32**[Database issue]:D434–7, January 2004. 15, 25

[44] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, **3**[2]:115–130, August 2000. 115

[45] A. Funahashi, M. Morohashi, H. Kitano, and N. Tanimura. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, **1**:159–162, 2003. 23

[46] L. Y. Geer, A. Marchler-Bauer, R. C. Geer, L. Han, J. He, S. He, C. Liu, W. Shi, and S. H. Bryant. The NCBI BioSystems database. *Nucleic Acids Research*, **38**[Database Issue]:D492–D496, December 2009. 24

[47] R. N. Goldberg, Y. B. Tewari, and T. N. Bhat. Thermodynamics of Enzyme-Catalyzed Reactions: Part 7—2007 Update. *Journal of Physical and Chemical Reference Data*, **36**[4]:1347, October 2007. 183

[48] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell. Metabolomics by numbers: acquiring and understanding

global metabolite data. *Trends in Biotechnology*, **22**[5]:245–252, May 2004. 8, 9

[49] M. Gremse, A. Chang, I. Schomburg, A. Grote, M. Scheer, C. Ebeling, and D. Schomburg. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Research*, **39**[Database]:D507–D513, December 2010. 14, 36, 131, 274, 275, 276, 277, 278, 279

[50] J. Griffin. Understanding mouse models of disease through metabolomics. *Current Opinion in Chemical Biology*, **10**[4]:309–315, August 2006. 7, 8

[51] J. L. Griffin, H. J. Atherton, C. Steinbeck, and R. M. Salek. A Metadata description of the data in "A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human.". *BMC Research Notes*, **4**[1]:272, 2011. 9

[52] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, 2008. 145, 146

[53] V. Hatzimanikatis, C. Li, J. Ionita, C. Henry, M. Jankowski, and L. Broadbelt. Exploring the diversity of complex metabolic networks. *Bioinformatics (Oxford, England)*, **21**[8]:1603, April 2005. 171

[54] M. Hearst. Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545, 1992. 140

[55] K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. M. Hendriksen, B. J. A. Schijvenaars, E. M. v. Mulligen, J. Kleinjans, and J. A. Kors. A dictionary to identify small molecules and drugs in free text. *Bioinformatics (Oxford, England)*, **25**[22]:2983–2991, November 2009. 114, 122

[56] R. Holland, T. Down, M. Pocock, A. Prlic, D. Huen, K. James, S. Foisy, A. Drager, A. Yates, M. Heuer, and M. Schreiber. BioJava: an open-source framework for bioinformatics. *Bioinformatics (Oxford, England)*, **24**[18]:2096, September 2008. 27, 209

[57] M. Horridge and S. Bechhofer. The owl api: A java api for owl ontologies. *Semantic Web*, **2**[1]:11–21, 2011. 137

[58] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4**[1]:44–57, December 2008. 81

[59] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi,

M. Tomita, J. Wagner, J. Wang, and Unknown. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)*, **19**[4]:524–531, March 2003. 27

[60] K. Hult and P. Berglund. Enzyme promiscuity: mechanism and applications. *Trends in Biotechnology*, **25**[5]:231–238, May 2007. 168, 169

[61] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, E. de Castro, P. Coggill, M. Corbett, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Muellenet, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A. F. Quinn, C. Rivoire, A. Sangrador-Vegas, J. D. Selengut, C. J. A. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajanarthanan, P. D. Thomas, C. H. Wu, C. Yeats, and S.-Y. Yong. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research*, **40**[Database issue]:D306–12, January 2012. 216

[62] D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, B. Gabrys, R. J. Howlett, and L. C. Jain, ed-

itors. *Unification of Protein Data and Knowledge Sources*, **4251** of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2006. 7, 107

[63] M. D. Jankowski, C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis. Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks. *Biophysical Journal*, **95**[3]:1487–1499, May 2008. 181, 183

[64] K. V. Jayaseelan, P. Moreno, A. Truszkowski, P. Ertl, and C. Steinbeck. Natural product-likeness score revisited: an open-source, open-data implementation. *BMC Bioinformatics*, **13**[1]:106, 2012. 142, 144

[65] H. Jenke-Kodama, A. Sandmann, R. Müller, and E. Dittmann. Evolutionary implications of bacterial polyketide synthases. *Molecular Biology and Evolution*, **22**[10]:2027–2039, October 2005. xxxii, 200, 202, 203

[66] H. Jenkins, N. Hardy, M. Beckmann, J. Draper, A. R. Smith, J. Taylor, O. Fiehn, R. Goodacre, R. J. Bino, R. Hall, J. Kopka, G. A. Lane, B. M. Lange, J. R. Liu, P. Mendes, B. J. Nikolau, S. G. Oliver, N. W. Paton, S. Rhee, U. Roessner-Tunali, K. Saito, J. Smedsgaard, L. W. Sumner, T. Wang, S. Walsh, E. S. Wurtele, and D. B. Kell. A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnology*, **22**[12]:1601–1606, December 2004. 9

[67] L. J. Jensen, P. Julien, M. Kuhn, C. von Mering, J. Muller, T. Doerks, and P. Bork. eggNOG: automated construction and anno-

tation of orthologous groups of genes. *Nucleic Acids Research*, **36**[Database issue]:D250–4, January 2008. 35

[68] D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy, and P. Murray-Rust. OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, **3**[1]:41, 2011. 112

[69] T. Jewison, C. Knox, V. Neveu, Y. Djoumbou, A. C. Guo, J. Lee, P. Liu, R. Mandal, R. Krishnamurthy, I. Sinelnikov, M. Wilson, and D. S. Wishart. YMDB: the Yeast Metabolome Database. *Nucleic Acids Research*, **40**[D1]:D815–D820, December 2011. 1, 6

[70] N. Juty, N. Le Novère, and C. Laibe. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Research*, **40**[Database issue]:D580–6, January 2012. 27

[71] M. Kanehisa. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, **32**[90001]:277D–280, January 2004. 5, 35

[72] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **28**[1]:27–30, January 2000. 17, 19, 107

[73] Y. Kano, W. A. Baumgartner, L. McCrohon, S. Ananiadou, K. B. Cohen, L. Hunter, and J. Tsujii. U-Compare: share and compare text mining tools with UIMA. *Bioinformatics (Oxford, England)*, **25**[15]:1997–1998, July 2009. 116

[74] P. D. Karp, S. Paley, and P. Romero. The Pathway Tools software. *Bioinformatics (Oxford, England)*, **18**[Suppl 1]:S225–S232, July 2002. 20

[75] P. T. Kasper, M. Rojas-Chertó, R. Mistrik, T. Reijmers, T. Hankemeier, and R. J. Vreeken. Fragmentation trees for the structural characterisation of metabolites. *Rapid Communications in Mass Spectrometry*, **26**[19]:2275–2286, August 2012. 8

[76] K. A. Kékesi, Z. Kovács, N. Szilágyi, M. Bobest, T. Szikra, Á. Dobolyi, G. Juhász, and M. Palkovits. Concentration of Nucleosides and Related Compounds in Cerebral and Cerebellar Cortical Areas and White Matter of the Human Brain. *Cellular and Molecular Neurobiology*, **26**[4-6]:831–842, August 2006. 157

[77] D. B. Kell, M. Brown, H. M. Davey, W. B. Dunn, I. Spasic, and S. G. Oliver. Metabolic footprinting and systems biology: the medium is the message. *Nature Reviews Microbiology*, **3**[7]:557–565, June 2005. 8

[78] I. M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muñiz-Rascado, C. Bonavides-Martínez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A. G. Shearer, A. Mackie, I. Paulsen, R. P. Gunsalus, and P. D. Karp. EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Research*, **39**[Database issue]:D583–90, January 2011. 21

[79] O. Khersonsky, C. Roodveldt, and D. S. Tawfik. Enzyme promiscuity: evolutionary and mechanistic aspects. *Current Opinion in Chemical Biology*, **10**[5]:498–508, October 2006. 168

[80] O. Khersonsky and D. S. Tawfik. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annual Review of Biochemistry*, **79**:471–505, January 2010. 168, 169

[81] J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. GENIA corpus–a semantically annotated corpus for bio-textmining. *Bioinformatics (Oxford, England)*, **19**[Suppl 1]:i180–i182, July 2003. 114, 115

[82] H. Kirsch, S. Gaudan, and D. Rebholz-Schuhmann. Distributed modules for text annotation and IE applied to the biomedical domain. *International Journal of Medical Informatics*, **75**[6]:496–500, June 2006. 113

[83] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research*, **39**[Database]:D1035–D1041, December 2010. 111

[84] J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Rüegg, C. Rawlings, P. Verrier, and S. Philippi. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics (Oxford, England)*, **22**[11]:1383–1390, June 2006. 28

[85] J. Kopka, N. Schauer, S. Krueger, and C. Birkemeyer. GMD@ CSB. DB: the Golm Metabolome Database. *Bioinformatics (Oxford, England)*, January 2005. 3, 9

[86] Z. Kovács, Á. Dobolyi, G. Juhász, and K. A. Kékesi. Nucleoside Map of the Human Central Nervous System. *Neurochemical Research*, **35**[3]:452–464, October 2009. 157

[87] J. Küntzer, C. Backes, T. Blum, A. Gerasch, M. Kaufmann, O. Kohlbacher, and H.-P. Lenhof. BNDB – The Biochemical Network Database. *BMC Bioinformatics*, **8**[1]:367, 2007. 28

[88] N. Le Novère. MELTING, computing the melting temperature of nucleic acid duplex. *Bioinformatics (Oxford, England)*, **17**[12]:1226–1227, December 2001. 211

[89] T. J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D. W. J. Stringer-Calvert, J. D. Tenenbaum, and P. D. Karp. BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics*, **7**:170, January 2006. 28

[90] R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tarraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, G. Hoad, M. Jang, N. Pakseresht, S. Plaister, R. Radhakrishnan, K. Reddy, S. Sobhany, P. Ten Hoopen, R. Vaughan, V. Zalunin, and G. Cochrane. The European Nucleotide Archive. *Nucleic Acids Research*, **39**[Database]:D28–D31, December 2010. xiii, 2

[91] C. Linhart and R. Shamir. The degenerate primer design problem: theory and applications. *Journal of Computational Biology : a journal of computational molecular cell biology*, **12**[4]:431–456, May 2005. 208

[92] A. Lourenco, S. Carneiro, M. Rocha, E. C. Ferreira, and I. Rocha. Challenges in integrating Escherichia coli molecular biology data. *Briefings in Bioinformatics*, pages 1–13, November 2010. 51

[93] D. M. Lowe, P. T. Corbett, P. Murray-Rust, and R. C. Glen. Chemical name to structure: OPSIN, an open source solution. *Journal of Chemical Information and Modeling*, **51**[3]:739–753, March 2011. 112

[94] R. Lyne, R. Smith, K. Rutherford, M. Wakeling, A. Varley, F. Guillier, H. Janssens, W. Ji, P. Mclaren, P. North, D. Rana, T. Riley, J. Sullivan, X. Watkins, M. Woodbridge, K. Lilley, S. Russell, M. Ashburner, K. Mizuguchi, and G. Micklem. FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biology*, **8**[7]:R129, 2007. 27

[95] A. Macchiarulo, J. M. Thornton, and I. Nobeli. Mapping human metabolic pathways in the small molecule chemical space. *Journal of Chemical Information and Modeling*, **49**[10]:2272–2289, October 2009. 144

[96] S. Maere, K. Heymans, and M. Kuiper. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)*, **21**[16]:3448–3449, August 2005. 89

[97] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson. Modeling sample variables with an Experimental Factor

Ontology. *Bioinformatics (Oxford, England)*, **26**[8]:1112–1118, April 2010. 38

[98] M. A. Mandel. Integrated annotation of biomedical text: Creating the pennbioie corpus. In *Proceedings of the Workshop on Text Mining, Ontologies and Natural Language Processing in Biomedicine (2006).*, March 2006. 114

[99] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, **37**[Database]:D619–D622, January 2009. 23

[100] M. L. Mavrovouniotis. Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnology and Bioengineering*, **36**[10]:1070–1082, December 1990. 181, 183

[101] P. V. Mazin, M. S. Gelfand, A. A. Mironov, A. B. Rakhmaninova, A. R. Rubinov, R. B. Russell, and O. V. Kalinina. An automated stochastic approach to the identification of the protein specificity determinants and functional subfamilies. *Algorithms for Molecular Biology*, **5**[1]:29, 2010. 214

[102] P. Mendes. Metabolomics and the challenges ahead. *Briefings in Bioinformatics*, **7**[2]:127–127, March 2006. 9

[103] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. YALE: Rapid Prototyping for Complex Data Mining Tasks. In L. Ungar, M. Craven, D. Gunopulos, and T. Eliassi-Rad, editors, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, New York, NY, USA, August 2006. ACM. 145

[104] S. Moco, R. J. Bino, R. C. H. De Vos, and J. Vervoort. Metabolomics technologies and metabolite identification. *TrAC Trends in Analytical Chemistry*, **26**[9]:855–866, 2007. xiii, 10

[105] C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm. How Many Species Are There on Earth and in the Ocean? *PLoS Biology*, **9**[8]:e1001127, August 2011. 5

[106] C. J. Mungall, D. B. Emmert, and The FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics (Oxford, England)*, **23**[13]:i337–i346, July 2007. 28

[107] H. S. Najafabadi, A. Saberi, N. Torabi, and M. Chamankhah. MAD-DPD: designing highly degenerate primers with maximum amplification specificity. *BioTechniques*, **44**[4]:519–20, 522, 524–6, March 2008. 208

[108] T. Nguyen, K. Ishida, H. Jenke-Kodama, E. Dittmann, C. Gurgui, T. Hochmuth, S. Taudien, M. Platzer, C. Hertweck, and J. Piel. Exploiting the mosaic structure of trans-acyltransferase polyke-

tide synthases for natural product discovery and pathway dissection. *Nature Biotechnology*, **26**[2]:225–233, February 2008. xxxii, xxxiii, 202, 203, 205, 206

[109] C. Nobata, P. D. Dobson, S. A. Iqbal, P. Mendes, J. Tsujii, D. B. Kell, and S. Ananiadou. Mining metabolites: extracting the yeast metabolome from the literature. *Metabolomics*, **7**[1]:94–101, March 2011. 116

[110] M. Oh, T. Yamada, M. Hattori, S. Goto, and M. Kanehisa. Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *Journal of Chemical Information and Modeling*, **47**[4]:1702–1712, January 2007. 173

[111] T. Ohta, T. Matsuzaki, N. Okazaki, M. Miwa, R. Sætre, S. Pyysalo, and J. Tsujii. Medie and Info-pubmed: 2010 update. *BMC Bioinformatics*, **11**[Suppl 5]:P7, 2010. 115

[112] D. J. Oliver, B. J. Nikolau, and E. S. Wurtele. Acetyl-CoA—Life at the metabolic nexus. *Plant Science*, **176**[5]:597–601, May 2009. 101

[113] S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz. Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, **16**[9]:373–378, September 1998. 5

[114] S. Orchard, H. Hermjakob, C. Taylor, P.-A. Binz, C. Hoogland, R. Julian, J. S. Garavelli, R. Aebersold, and R. Apweiler. Autumn 2005 Workshop of the Human Proteome Organisation Proteomics

Standards Initiative (HUPO-PSI) Geneva, September, 4–6, 2005. *Proteomics*, **6**[3]:738–741, February 2006. 9

[115] M. Oresic, V. A. Hänninen, and A. Vidal-Puig. Lipidomics: a new window to biomedical frontiers. *Trends in Biotechnology*, **26**[12]:647–652, December 2008. 6

[116] H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwan, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S.-A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*, **37**[Database issue]:D868–72, January 2009. xiii, 2

[117] P. G. A. Pedrioli, J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, and R. Aebersold. A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology*, **22**[11]:1459–1466, November 2004. 9

[118] H. E. Pence and A. Williams. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education*, **87**[11]:1123–1124, 2010. 17

[119] S. Petrovic, J. Snajder, B. D. Basic, and M. Kolar. Comparison of Collocation Extraction Measures for Document Indexing. *Journal of Computing and Information Technology*, **4**[14]:321–327, 2006. 124

[120] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo. WikiPathways: pathway editing for the people. *PLoS Biology*, **6**[7]:e184, July 2008. 25

[121] N. Psychogios, D. D. Hau, J. Peng, A. C. Guo, R. Mandal, S. Bouatra, I. Sinelnikov, R. Krishnamurthy, R. Eisner, B. Gautam, N. Young, J. Xia, C. Knox, E. Dong, P. Huang, Z. Hollander, T. L. Pedersen, S. R. Smith, F. Bamforth, R. Greiner, B. McManus, J. W. Newman, T. Goodfriend, and D. S. Wishart. The Human Serum Metabolome. *PLoS ONE*, **6**[2]:e16957, February 2011. 1, 6

[122] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. EDDY, A. Bateman, and R. D. Finn. The Pfam protein families database. *Nucleic Acids Research*, **40**[Database issue]:D290–301, January 2012. 216

[123] K. Radrich, Y. Tsuruoka, P. Dobson, A. Gevorgyan, N. Swainston, G. Baart, and J.-M. Schwartz. Integration of metabolic

databases for the reconstruction of genome-scale metabolic networks. *BMC Systems Biology*, **4**:114, 2010. 53

[124] S. A. Rahman, M. Bashton, G. L. Holliday, R. Schrader, and J. M. Thornton. Small Molecule Subgraph Detector (SMSD) toolkit. *Journal of Cheminformatics*, **1**[1]:12, 2009. 68, 176

[125] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno. Text processing through Web services: calling Whatizit. *Bioinformatics (Oxford, England)*, **24**[2]:296, January 2008. 113

[126] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P. Stoehr. EBIMed–text crunching to gather facts for proteins from Medline. *Bioinformatics (Oxford, England)*, **23**[2]:e237, January 2007. 113

[127] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics : TIG*, **16**[6]:276–277, June 2000. 210

[128] C. P. Ridley, H. Y. Lee, and C. Khosla. Evolution of polyketide synthases in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **105**[12]:4595–4600, March 2008. xxxiii, 203, 204, 272

[129] T. M. Rose, J. G. Henikoff, and S. Henikoff. CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. *Nucleic Acids Research*, **31**[13]:3763–3766, June 2003. 208

[130] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **39**[Database issue]:D38–51, January 2011. 5

[131] I. Schomburg. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Research*, **32**[90001]:431D–433, January 2004. 22

[132] S. L. Schreiber. Small molecules: the missing link in the central dogma. *Nature Chemical Biology*, **1**[2]:64–66, July 2005. 1

[133] V. Seretan. *Collocation extraction based on syntactic parsing*. PhD thesis, University of Geneva, February 2008. 124

[134] S. P. Shah, Y. Huang, T. Xu, M. M. Yuen, J. Ling, and B. F. Ouellette. Atlas - a data warehouse for integrative bioinformatics. *BMC Bioinformatics*, **6**[1]:34, 2005. 28

[135] Y. Shinbo, Y. Nakamura, M. Altaf-Ul-Amin, H. Asahi, K. Kurokawa, M. Arita, K. Saito, D. Ohta, D. Shibata, and S. Kanaya. *Biotechnology in Agriculture and Forestry*, **57** of *Biotechnology in Agriculture and Forestry*. Springer-Verlag, Berlin/Heidelberg, 2006. 3, 9

[136] C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, and G. Siuzdak. METLIN: a metabolite mass spectral database. *Therapeutic Drug Monitoring*, **27**[6]:747–751, December 2005. 3, 9

[137] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**[1]:195–197, March 1981. 210

[138] M. E. Smoot, K. Ono, J. Ruscheinski, P. L. Wang, and T. Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)*, **27**[3]:431–432, January 2011. 89

[139] T. Soga, Y. Ohashi, Y. Ueno, H. Naraoka, M. Tomita, and T. Nishioka. Quantitative Metabolome Analysis Using Capillary Electrophoresis Mass Spectrometry. *Journal of Proteome Research*, **2**[5]:488–494, October 2003. 6

[140] S. E. Stein, S. R. Heller, and D. Tchekhovskoi. An Open Standard for Chemical Structure Representation: The IUPAC Chemical Identifier. *In Proceedings of the 2003 International Chemical Information Conference (2003), pp. 131-143.*, pages 131–143, 2003. 52

[141] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *Journal of Chemical Information and Computer Sciences*, **43**[2]:493–500, February 2003. 27

[142] P. Stenetorp, S. Pyysalo, G. Topi c, T. Ohta, S. Ananiadou, and J. Tsujii. Brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Demonstrations Session at EACL 2012*, Avignon, France. European Chapter of the Association for Computational Linguistics. xxiii, 109, 110

[143] M. D. Stobbe, S. M. Houten, G. A. Jansen, A. H. van Kampen, and P. D. Moerland. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Systems Biology*, **5**[1]:165, 2011. 77

[144] A. I. Su, M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker, T. Wiltshire, A. P. Orth, R. G. Vega, L. M. Sapinoso, A. Moqrich, A. Patapoutian, G. M. Hampton, P. G. Schultz, and J. B. Hogenesch. Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, **99**[7]:4465–4470, April 2002. 155

[145] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the Na-*

*tional Academy of Sciences of the United States of America*, **101**[16]:6062–6067, April 2004. 160

[146] M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. H. Raetz, D. W. Russell, and S. Subramaniam. LMSD: LIPID MAPS structure database. *Nucleic Acids Research*, **35**[Database]:D527–D532, January 2007. 6

[147] Y. Tang, A. Y. Chen, C.-Y. Kim, D. E. Cane, and C. Khosla. Structural and mechanistic analysis of protein interactions in module 3 of the 6-deoxyerythronolide B synthase. *Chemistry & Biology*, **14**[8]:931–943, August 2007. 215

[148] Y. Tang, C.-Y. Kim, I. I. Mathews, D. E. Cane, and C. Khosla. The 2.7-Angstrom crystal structure of a 194-kDa homodimeric fragment of the 6-deoxyerythronolide B synthase. *Proceedings of the National Academy of Sciences of the United States of America*, **103**[30]:11124–11129, July 2006. 215

[149] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**[1]:41, 2003. 35

[150] T. Tatusova. *Genomic Databases and Resources at the National Center for Biotechnology Information*, **609** of *Methods in Molecular Biology*. Humana Press, Totowa, NJ, October 2009. 24

[151] I. Thiele and B. Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, **5**[1]:93–121, July 2010. 7

[152] K. Tipton and S. Boyce. History of the enzyme nomenclature system. *Bioinformatics (Oxford, England)*, **16**[1]:34–40, January 2000. 14, 17

[153] A. Truszkowski, K. V. Jayaseelan, S. Neumann, E. L. Willighagen, A. Zielesny, and C. Steinbeck. New developments on the cheminformatics open workflow environment CDK-Taverna. *Journal of Cheminformatics*, **3**:54, 2011. 174

[154] H. Tsugawa, Y. Tsujimoto, M. Arita, T. Bamba, and E. Fukusaki. GC/MS based metabolomics: development of a data mining system for metabolite identification by using soft independent modeling of class analogy (SIMCA). *BMC Bioinformatics*, **12**:131, 2011. 2

[155] Y. Tsuruoka, Y. Tateishi, J. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii. Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics, Proceedings*, pages 382–392. Japan Sci & Technol Agcy, CREST, Kawaguchi, Saitama 3320012, Japan, 2005. 114

[156] Y. Tsuruoka, M. Miwa, K. Hamamoto, J. Tsujii, and S. Ananiadou. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics (Oxford, England)*, **27**[13]:i111–9, July 2011. 116

[157] J. van der Greef and A. K. Smilde. Symbiosis of chemometrics and metabolomics: past, present, and future. *Journal of Chemometrics*, **19**[57]:376–386, 2005. 2, 5

[158] S. G. S. Villas-Bôas, S. S. Mas, M. M. Akesson, J. J. Smedsgaard, and J. J. Nielsen. Mass spectrometry in metabolome analysis. *Mass Spectrometry Reviews*, **24**[5]:613–646, January 2005. 108

[159] X. Wang and M. Matthews. Distinguishing the species of biomedical named entities for term identification. *BMC Bioinformatics*, **9**[Suppl 11]:S6, 2008. 124

[160] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, **37**[Web Server]:W623–W633, June 2009. 17

[161] C. Warren. Use of chemical ionization for GC–MS metabolite profiling. *Metabolomics*, pages 1–11, 2011. 2

[162] Q. Wei and N. Collier. Towards classifying species in systems biology papers using text mining. *BMC Research Notes*, **4**[1]:32, 2011. 124

[163] X. Wei, D. N. Kuhn, and G. Narasimhan. Degenerate primer design via clustering. *Proceedings / IEEE Computer Society Bioinformatics*

*Conference IEEE Computer Society Bioinformatics Conference*, **2**:75–83, January 2003. 208

[164] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, **28**[1]:31–36, February 1988. 52

[165] D. Wishart. Current Progress in computational metabolomics. *Briefings in Bioinformatics*, **8**[5]:279, September 2007. 5

[166] D. Wishart, D. Tzur, C. Knox, R. Eisner, A. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.-A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. Duggan, G. Macinnis, A. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. Sykes, H. Vogel, and L. Querengesser. HMDB: the Human Metabolome Database. *Nucleic Acids Research*, **35**[Database issue]:D521, January 2007. 1, 3, 6, 26

[167] Y. Yamanishi, M. Hattori, M. Kotera, S. Goto, and M. Kanehisa. E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. *Bioinformatics (Oxford, England)*, **25**[12]:i179–86, June 2009. 20, 170

[168] L. Yetukuri, K. Ekroos, A. Vidal-Puig, and M. Oresic. Informatics and computational strategies for the study of lipids. *Molecular BioSystems*, **4**[2]:121–127, February 2008. 6