UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA


# Structural and semantic similarity metrics for chemical compound classification

João Diogo Silva Ferreira

MESTRADO EM BIOQUÍMICA

2010

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA


# Structural and semantic similarity metrics for chemical compound classification

João Diogo Silva Ferreira

MESTRADO EM BIOQUÍMICA

Dissertação orientada pelo Prof. Doutor Francisco José Moreira Couto

2010

# Resumo

Ao longo das últimas décadas, tem-se assistido a um grande aumento na quantidade de dados produzidos e disponibilizados em química, em especial após a introdução de métodos de análise mecanizados. Devido a este crescimento no número de dados, existe cada vez mais uma necessidade de implementar sistemas automáticos computacionais capazes de armazenar, estudar e interpretar estes dados de forma eficiente. Uma das tarefas mais importantes em quimio-informática é, de facto, a utilização dos dados obtidos em laboratório em sistemas de comparação e classificação de compostos químicos. Os métodos actuais mais eficazes baseiam-se na premissa de que a função de um composto químico está intimamente relacionada com a sua estrutura. Apesar de esta premissa estar geralmente correcta, como comprovam os métodos actuais, eles podem falhar, especialmente quando moléculas parecidas desempenham funções diferentes (como acontece com os L- e D-aminoácidos) ou moléculas diferentes desempenham uma função biológica semelhante (como acontece com inúmeros exemplos de inibidores).

O trabalho proposto neste documento apresenta uma solução para resolver este problema através da utilização de uma métrica híbrida que integre no seu núcleo informação não só estrutural mas também semântica, ou seja, o sistema desenvolvido tem a capacidade de explorar a informação acerca do significado das moléculas num contexto bioquímico. Para este efeito, utilizei o ChEBI como fonte de informação semântica, tendo criado uma ferramenta denominada Chym (Chemical Hybrid Metric) que é capaz de lidar com problemas de classificação de compostos químicos. Resumidamente, para decidir se um composto químico possui uma determinada característica, por exemplo se atravessa a barreira hematoencefálica, este sistema atribui ao composto um coeficiente de actividade que é calculado com base nos compostos químicos que se sabe possuírem a característica; por comparação com um valor de corte, o

Chym classifica o composto em estudo como possuidor ou não dessa característica.

A ferramenta que resultou do trabalho desta tese foi aqui explorada e validada. Assim, o trabalho apresentado mostra evidências substanciais que suportam a eficácia do Chym, uma vez que este apresenta melhores resultados do que todos os modelos com os quais foi comparado. Particularmente, para três problemas seleccionados, o Chym decide correctamente qual a classificação de um composto 90.9%, 87.7% e 84.2% das vezes: pela ordem apresentada, esses valores referem-se à classificação de compostos como permeáveis à barreira hematoencefálica, como substratos da glicoproteína-P, ou como ligandos de um receptor de estrogénio. Para efeitos de comparação, estes três problemas foram anteriormente resolvidos com exactidão de 81.5%, 80.6% e 82.8% respectivamente. Comprova-se, portanto, a hipótese da tese, ou seja, que a integração de informação semântica em sistemas de comparação e classificação de compostos químicos aumenta, por vezes de forma substancial, a fidelidade do método.

Desta forma, o objectivo da tese foi bem sucedido em duas frentes. Por um lado a tese serviu para validar a hipótese, e por outro culminou na criação de uma ferramenta de classificação de compostos químicos que pode vir a ser usada no futuro em projectos mais abrangentes, nomeadamente no estudo da evolução das vias metabólicas, na área de desenvolvimento de fármacos ou na análise preliminar da toxicidade de compostos químicos.

**Palavras chave:** Aprendizagem automática, Ontologias, Semelhança de compostos químicos, Semelhança semântica

# Abstract

Over the last few decades, there has been an increasing number of attempts at creating systems capable of comparing and classifying chemical compounds based on their structure and/or physicochemical properties. While the rate of success of these approaches has been increasing, particularly with the introduction of new and ever more sophisticated methods of machine learning, there is still room for improvement. One of the problems of these methods is that they fail to consider that similar molecules may have different roles in nature, or, to a lesser extend, that disparate molecules may have similar roles.

This thesis proposes the exploitation of the semantic properties of chemical compounds, as described in the ChEBI ontology, to create an efficient system able to automatically deal with the binary classification of chemical compounds. To that effect, I developed Chym (Chemical Hybrid Metric) as a tool that integrates structural and semantic information in a unique hybrid metric.

The work here presented shows substantial evidence supporting the effectiveness of Chym, since it has outperformed all the models with which it was compared. Particularly, it achieved accuracy values of 90.9%, 87.7% and 84.2% when solving three classification problems which, previously, had only been solved with accuracy values of 81.5%, 80.6% and 82.8% respectively. Other results show that the tool is appropriate to use even if the problem at hand is not well represented in the ChEBI ontology. Thus, Chym shows that considering the semantic properties of a compound helps solving classification problems.

Therefore, Chym can be used in projects that require the classification and/or the comparison of chemical compounds, such as the study of the evolution of metabolic pathways, drug discovery or in preliminary toxicity analysis.

**Keywords:** Chemical compound similarity, Machine learning, Ontologies, Semantic similarity

# Acknowledgments

This work would not have been possible without the precious help of a number of people, particularly my supervisor, Prof. Dr Francisco José Moreira Couto. He always had some fresh new insights about the work being developed, and was particularly helpful during the writing of this document. Comments, ideas and observations from my colleagues at the XLDB were also invaluable; after our tertulias, I always felt I had something new to add or modify in this project.

My colleagues Diogo Vila Viçosa and João Freire, from the Biochemistry Master's Degree, also deserve a "thank you". With them, especially during our Bioinformatics project, I gained plentiful of experience in the subject of programming.

I would also like to thank my parents, whose unconditional support always comforted me, even when my decisions seemed incomprehensible.

Last, but by no means least, I am thoroughly thankful to all the people that supported this project, even if only with a little pat on my back. Sandra, for the afternoons of work, and for not being able to stroll in the park with you even with allowing weather; Simão, for all those times when I couldn't play video games with you because I was busy writing this up; my aunts and uncles, for listening attentively when I tried to explain what it means to compare two chemical compounds (even if sometimes you wouldn't completely understand); and those that had to listen to me talking about my work, during lunch, even though I was usually received with bored looks on your faces.

A big "Thank You" to you all.

*Lisboa*, April 30$^{\text{th}}$, 2010
João Ferreira

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| **ANN** | Artificial Neural Network |
| **BBB** | Blood-Brain Barrier |
| **ChEBI** | Chemical Entities of Biological Interest |
| **ChEMBL** | A database of bioactive drug-like small molecules |
| **Chym** | Chemical Hybrid Metric |
| **DAG** | Directed Acyclic Graph |
| **EBI** | European Bioinformatics Institute |
| **FN** | False Negative |
| **FP** | Fingerprint format / False Positive |
| **GO** | Gene Ontology |
| **IC** | Information Content |
| **InChI** | IUPAC International Chemical Identifier |
| **IUPAC** | International Union of Pure and Applied Chemistry |
| **KEGG** | Kyoto Encyclopedia of Genes and Genomes |
| **MCC** | Matthews Correlation Coefficient |
| **MDL** | A file format used to store chemical structures |
| **OBO** | Open Biomedical Ontologies |
| **P-gp** | P-glycoprotein |
| **QSAR** | Quantitative Structure-Activity Relationship |
| **SAR** | Structure-Activity Relationship |
| **SMILES** | Simplified molecular input line entry specification |
| **simGIC** | Semantic similarity calculated through an hybrid Graph-based and IC-based method |
| **simUI** | Semantic similarity calculated through a graph-based method, where UI stands for Union & Intersection |
| **SVM** | Support Vector Machines |
| **TN** | True Negative |
| **TP** | True Positive |

# Chapter 1

# Introduction

With the current amount of chemical data being published and produced, it has become increasingly necessary to devise automatic systems capable of handling this information. The term "*automatic*" has been used for a long time in chemistry, but the concept that comes to mind today is relatively different from the one that would be evoked by an average scientist 40 years ago. One has only to open one of the first issues of the *Journal of Automated Methods and Management in Chemistry*, published since 1978 (when the name was actually *Journal of Automatic Chemistry*), to see that the term's meaning was more close to *mechanization* (see, for example, Mitchell, 1978). Nowadays, with the increasing use of computers in the biochemical, biological and biomedical sciences, the use of the word *automatic* progressively tends to suggest the use of computation for the organization, study and production of new information and knowledge. Specifically in chemistry, computers are mostly used to automatically study molecules or molecular interactions (such as the ones between proteins and ligands) (Yılmaz and Göktürk, 2009). The field of knowledge that uses computers and computer science in general to handle, study and solve chemical problems is known as *Cheminformatics*.

In a sense, it was the first meaning that pushed chemical science to start using computers as a way to produce new information: the mechanization induced an extremely rapid increase in the amount of data produced and made it much more difficult to manually validate and process it. This raised the need to create programs specifically designed to deal with chemical data.

This work is inserted in the context of cheminformatics and aims to provide a tool capable of dealing with that need. I propose the creation of an effec-

tive system to automatically compare and classify chemical compounds. This system serves a number of different applications. While the following is by no means an exhaustive list, I tried to collect and expose three examples of applications that would benefit from such a chemical compound comparison tool.

## 1.1 Applications of chemical compound similarity

The first example is the use of chemical compound similarity in metabolic pathway comparison. This is a problem which has seen some interesting implementations. The most usual methods try to align two metabolic pathways such that the number of modifications that need to be done to go from one to the other is minimal. Usual methods to compute the *weight*, or *cost*, of each modification, use protein similarity, calculated based on EC number (Pinter et al., 2005; Heymans and Singh, 2003), protein sequence (Shlomi et al., 2006) or other more sophisticated similarity measures, such as semantic similarity based on the Gene Ontology (Clemente et al., 2005). However, more recently, works have been published where the alignment of the pathways is done not based on enzymes but on metabolites. Tohsato and Nishimura (2008) present an approach for the alignment of chemical reactions based on their substrates and products, which is used to create a similarity measure for pathways. Using metabolites instead of enzymes has some advantages:

(i) Even if enzymes are much more widely studied than small molecules, the first elements of a pathway to be completely known are the metabolites.

(ii) Macromolecules, such as enzymes, are harder to study than the small molecules that act as metabolites. For instance, sequencing and three dimensional studies usually take a long time and are not always readily available (as in the case of membrane proteins).

A system capable of identifying and aligning similar metabolic pathways would enable other, larger projects, such as the comparison of metabolic networks. Such a system can be applied, for instance, in the study of several strains of a single species. If some strains possess an interesting characteristic, as virulence, for example, this system could help determine the most important metabolic pathways responsible for that characteristic.

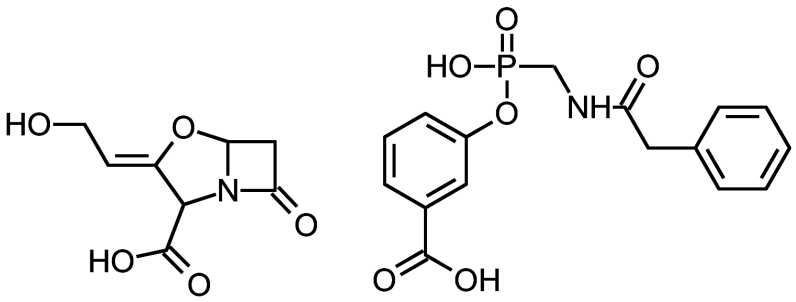
A second application of chemical similarity is in the study and development of pharmacophores, also known as drug discovery. While the subject of

drug discovery exists since the beginning of the 20$^{\text{th}}$ century (Drews, 2000), the advances in computer science and the increase of computational power seen in more recent years impelled the rapid growth of this field. For example, the use of computers to model the structure of proteins and small chemical molecules enabled the detection of specific interactions between proteins and ligands. This information can be used to understand the mechanism of action of a drug and to determine whether other molecules are expected to behave similarly (Wolber et al., 2008; Fukunishi et al., 2006). Penzotti et al. (2002) used a system capable of differentiating amongst substrates of the P-glycoprotein based on a similarity measure taken from structural properties of the molecules. This protein transports a variety of metabolites across membranes and is very important in drug-resistance pathways (Cordon-Cardo et al., 1990; Schinkel et al., 1995).

The third and last example is a variation of the previous one. Since molecular similarity measures can be used to estimate whether a small chemical compound is ligand of a protein, it can also be used to estimate whether it is toxic or not. Richard et al. (2006) present an interesting review on this subject. With a system capable of comparing and classifying chemical compounds, the task of screening a full database to search for potentially toxic substances becomes much easier. For instance, the use of a hypothetical toxicology analyzer could help reduce the cost of drug development through estimation of whether a given chemical compound is or has the potential to be harmful to animals or humans before attempting an *in vivo* experiment (Muster et al., 2008).

## 1.2 Problem

Today, several methods are used to compare chemical compounds. The best approaches to date are usually based on the structure-activity relationship premise (SAR), which states that the biological activity of a molecule is strongly related to its structural or physicochemical properties (Patani and LaVoie, 1996). While these methods are proof that this assumption generally holds, it is not always the case that structure is a good indicative of biological function. For instance, L-amino acids are used by cells to synthesize proteins, but their stereo-isomers, D-amino acids, are much less frequent in nature and their role is totally different. D-Serine, for instance, is a mediator in several physiological and pathological processes, including plasticity and neurotoxicity



**(a)** *Clavulanic Acid*

**(b)** *3-carboxyphenyl phenylacetamidomethylphosphonate*

**Figure 1.1: Chemical structure of two semantically related compounds.** The structures of these two compounds are very different, and yet they both inhibit β-lactamase.

(Wolosker et al., 2008). From a biological point of view, these are two distinct molecules, but because they share an almost absolute structural similarity, the methods mentioned above would fail to clearly distinguish them.

On the other hand, both *clavulanic acid* and *3-carboxyphenyl phenylacetamidomethylphosphonate* are β-lactamase inhibitors (Reading and Cole, 1977; Pratt, 1989), despite the differences between their structures, (see **Figure 1.1**). To be accurate, it should be mentioned that the mechanisms of inhibition are not the same: clavulanic acid is a competitive inhibitor of the β-lactamase (Todd and Benfield, 1990), while the 3-carboxyphenyl phenylacetamidomethylphosphonate seems to phosphorylate the active center of the enzyme (Pratt, 1989). However, it is a fact that these two completely different molecules share a very specific role in biological processes, and a robust comparison tool should take this fact into consideration.

## 1.3 Objective

In the examples presented above, a purely structural similarity measure will not reflect the biological activity of the molecules. A robust comparison method used in the context of biological sciences (in which the three applications described above are included) should take these discrepancies into

account. To address this problem, this thesis proposes the use of the *semantics* of a chemical compound in the context of biological relevance, i.e., its role in biological processes, which I used in conjunction with existing methods in an attempt to produce a better chemical compound comparison method that improves on the existing ones. This was achieved through the development of a novel hybrid metric that takes into account both structural and semantic information. I dubbed this approach Chym, for **C**hemical **Hy**brid **M**etric. Semantic information was extracted from ChEBI, the Chemical Entities of Biological Interest ontology, an ontology that contains more than 500,000 terms at the time of writing, which can be used as the core of semantic similarity (Degtyarenko et al., 2007).

Therefore, this thesis's **hypothesis** is that compounds sharing a biological role should have a similarity measure higher than the one obtained using only structural or physicochemical properties, reflecting that fact; conversely, the measure should also reflect different biological relevance even if both molecules are similar in structure. This goal was achieved through the integration of semantic similarity with other comparison methods in a hybrid metric.

## 1.4 Methodology

The methodology followed here is partly derived from Pesquita et al. (2008), at least where semantic similarity is concerned. That paper details several methods to calculate semantic similarity between two proteins described with Gene Ontology terms. Some of these methods were adapted to be used with chemical compounds described under the ChEBI ontology. The part of the hybrid metric that integrates a semantic similarity approach follows closely that work. The structural similarity is computed based on a fingerprint approach. These two methods are combined into a single formula, which is the Chym similarity measure. To assess the effectiveness of Chym, I used it as a classification tool on three data sets of chemical compounds extracted from previous works, and compared these results with the ones obtained in those works.

## 1.5 Results

After the validation process, Chym obtained accuracy values as high as 90.9%. This value was obtained when Chym was used to predict the permeability to

the blood-brain barrier of molecules in one of the sets. On the other two sets, accuracy values were 87.7% and 84.2% for prediction of whether compounds are P-glycoprotein substrates and whether they bind to an estrogen receptor, respectively. These values are clear improvements over the previous results obtained in the same data sets, which were 81.5%, 80.6% and 82.8% respectively.

## 1.6 Structure of the thesis

This document is structured in several chapters that should be read in order. It starts with some terminology definition and a detailed background section in chapter 2, explaining the terms and ideas that are needed to understand the rest of the work. Terms like *semantics*, mentioned above, are given a precise meaning, while the data sources are also mentioned and briefly analyzed. Chapter 3 delineates previous attempts at classifying and comparing chemical compounds, mentioning and detailing some of the existing methods and the results that were achieved with them. Those results are critical, since Chym is compared against them. Chapter 4 delineates the methods underlining Chym's approach and the major algorithms and formulas developed are presented and explained. This chapter also mentions the software used to implement Chym. The evaluation of the effectiveness of this approach is the topic of chapter 5, where the data and results of previous studies are compared to the outcome of Chym. This chapter is divided in four sections: the first two describe the setting used to validate Chym; then the results obtained are laid out; and finally the results are discussed. The thesis proceeds with chapter 6, where I make some final conclusions in the context of possible applications and briefly mention the future work that could be constructed upon this project. Finally, three appendices finish the thesis, where less important remarks are mentioned. Appendix A describes the road map that gave origin to the Chym tool, presenting the main decisions that had to be made and showing the results that support those decisions. Appendix B includes a brief description of the steps followed to create the local Chym database, and mentions programming languages and software used. Appendix C proves a mathematical statement made during the main text.

# Chapter 2

# Background

To fully understand the work developed to create and validate Chym, there are a number of terms, concepts and data sources that the reader should be familiar with. This chapter deals with those definitions and details.

## 2.1 Terminology

Throughout this work, I will often use three terms that were at the very base of the development of Chym.

The first term is the adjective *structural*. I refer to structural properties, structure and other related words when dealing with the composition of a molecule: atoms, bonds, charge, etc. The structure can usually be represented as a two dimensional graph, although sometimes it is important, for stereochemical reasons, to represent the relative position of atoms in three dimensional space, in which case three dimensional coordinates should be used (for instance L-amino acids and D-amino acids, can only be distinguished if spatial arrangement is taken into account).

The second term is the adjective *physicochemical*. It is common, in the field of cheminformatics, to refer to structural properties as a much wider concept, that includes not only the composition of the molecule but also some of the properties that are not immediately obvious from the structure. For instance, many papers mention the structure-activity relationship as the hypothesis that the biological activity of a molecule is a direct consequence of its structural and/or physicochemical properties. In the terminology I used in this thesis, it is important to distinguish between these two sets of prop-

erties. Examples of physicochemical properties include (a) the octanol-water partition coefficient, a very common property used in classification studies that make use of physicochemical properties, and which reflects the ratio of the concentration of a molecule in the two phases of the mixture of those two immiscible solvents, (b) the molar refractivity, which is a measure of the total polarizability of the molecule, or (c) the acid dissociation constant, or its logarithmic counterpart, $pK_a$, which measures the strength of an acid in an aqueous solution. All these properties are difficult to predict from the structure of the molecule, and they are generally gathered from literature by automatic data mining approaches, especially when the number of molecules involved is too big to allow the data mining to be done manually.

According to Oxford Advanced Learner's Dictionary (Crowther, 1995), the definition of *semantics* is "the branch of linguistics dealing with the meaning of words in sentences." While this term seems utterly connected to the linguistics filed, in has been transported to computer science mainly due to the study of artificial intelligence. Computer science has usually used this term as opposed to syntax (which refers to the rules that govern the correct arrangement of the words in the sentence); more recently, it has often been found in the expression *semantic web*, which is an attempt to introduce the concept of metadata into the information that travels through the internet. The semantics of a term is, therefore, its meaning in a predetermined context, the concept for which it stands. Specifically for chemical compounds, their semantics reflect what is known about their structure, their properties, and their role in nature. As such, it can be seen that semantic properties are a more abstract and generalized concept than both structural and physicochemical properties, since the semantic information can include both these properties.

Structural properties are directly computable from the structure and, because the semantics of a chemical compounds is described in a database (see below, under subsection 2.5), the retrieval of this information is also easily done. For this reason, the implementation of Chym, which uses only structural and semantic properties, was feasible. However, since the semantic information may include physicochemical information, this approach is not blatantly ignoring properties, but is rather exploiting them from a different perspective.

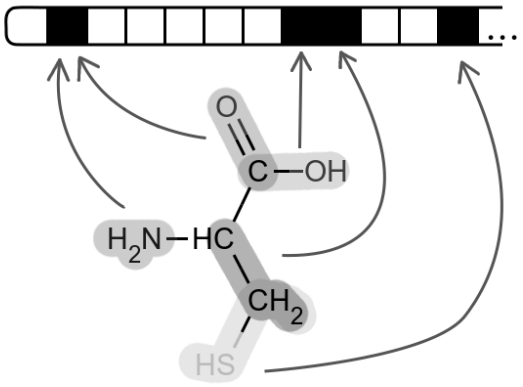



**Figure 2.1: Fingerprint example.** This is an example of a simple fingerprint. All the fragments of a molecule (in this case cysteine, ChEBI:15356) were retrieved from the structure, and then each fragment was used once to set one of the bits to 1, here represented with black squares. The fingerprint goes on to the right, where other bits may have been set to 1 as well. Only five fragments are represented in the figure, but it is possible to generate more than 40 from that molecule. As can be seen, sometimes more than one fragment sets the same bit to 1, in which case some information will be lost.

## 2.2 Fingerprints

A fingerprint, in the context of chemical compound similarity, is a bit-string, a sequence of 0's and 1's, where each bit represents the presence or absence of a given feature or substructure. There are several ways to construct the fingerprint. For instance, there is a class of fingerprints commonly used in cheminformatics usually called Daylight fingerprints (Daylight is the name of the company that first used the concept here described) (Daylight Chemical Information Systems, Inc., 2008). To construct this fingerprint, all distinct linear fragments, up to a certain size, are identified from the graph and then converted into numbers $n_i$: usually, a hash function is applied to the fragment, followed by a modulo function, effectively obtaining a number smaller than the size of the fingerprint. The $n_i{}^{\text{th}}$ bits in the fingerprint are then set to 1 (see **Figure 2.1** for a very simple example) (Flower, 1998; Raymond and Willett, 2002). One of the disadvantages of this method is the possibility of overlaps. Since all substructures up to a limit size are considered, the hash function will

have to assign the same bit to more than one linear fragment. This means a loss of resolution, but the general nature of the method somehow compensates for this problem.

Another method is to assign a particular substructure to each one of the bits of the fingerprint; one of the bits may represent the presence of sulfur atoms in the molecule or the existence of hydroxyl groups, for example. Here, no overlap will happen, but the descriptors must be wisely chosen to accurately represent the differences and similarities that one wants to catch.

Whatever the algorithms used to create the fingerprints, two molecules can then be quickly compared based on the number of common bits in their fingerprints, for example, through the Tanimoto (also called Jaccard-Tanimoto) (Jaccard, 1901; Willett et al., 1998; Flower, 1998; Martin et al., 2002) or the cosine (Willett et al., 1998) coefficients. These coefficients (and several other; see Willett et al. (1998) for an interesting review on this subject) assign the number of 1-bits in fingerprint A and B to variables $a$ and $b$, and the number of common 1-bits in both fingerprints to variable $c$. The Tanimoto coefficient is defined in equation 2.1 and the cosine coefficient in equation 2.2.

$$\mathrm{sim}_{\mathrm{Tanimoto}} = \frac{c}{a+b-c} \tag{2.1}$$

$$\mathrm{sim}_{\mathrm{Cosine}} = \frac{c}{\sqrt{ab}} \tag{2.2}$$

Of course, both fingerprints must have been calculated with the same method, or the comparison will be meaningless.

## 2.3 Semantic similarity

The semantic information of an object, i.e., its meaning in a predetermined context, is not easily handled by computers, mainly because meaning is a subjective concept and is often described in terms of natural language. For this reason, comparing the semantics of two objects (in this case, two chemical compounds), is not a straightforward task, and is only possible if the semantics of both objects are described under a common schema (Lord et al., 2003). In this work, I exploit the ChEBI ontology (see subsection 2.5 for a description of this ontology, and the next paragraph for an explanation of what an ontology is) to semantically describe chemical compounds. Under that common schema, it was possible to derive a semantic similarity metric.

An ontology is a representation of terms and the relationship between them, and is usually visualized as a directed graph where nodes are the terms and edges are the relationships (Chandrasekaran et al., 1999). A common type of relationship in ontologies is the *is a* relationship. It expresses the fact that one term (the child) can be classified as a subclass of another term (the parent). Given the non-cyclic nature of this relationship (a term can never be a subclass of one of its subclasses), some ontologies can be interpreted as directed acyclic graphs (DAG), where a term can have several parents and several children; in such a graph, the deeper a term is, the more specific its meaning is. For instance, if there is a single top term, it is completely unspecific, since it is a superclass of all the other terms. But while an ontology with only *is a* relationships is a DAG, a DAG can contain other relationship types (as long as they are non-cyclic). As will be seen later, ChEBI has a number of distinct non-cyclic relationship types, which enriches the ontology and enables the expression of a much wider spectrum of relations between the terms.

Given their structured an organized nature, ontologies are a common schema chosen to annotate biological entities like proteins, diseases and chemical compounds. For example, the Gene Ontology (GO) contains terms that can be used to annotate proteins. These terms include functions, biological processes etc. and one annotation with a GO term is a means to show that a particular protein possesses that function or participates in that process.

In the context of ontologies, a semantic measure between two terms reflects their proximity in the ontology. One of the simplest ways to compare two terms is to count the minimum number of relations that must be traversed to get from one compound to the other (Resnik, 1999). Another approach used in DAGs is to find the closest common ancestor of both terms; then, the distance between them is the maximum number of relations from one of the two terms being compared to the common ancestor.

It is worth noting that a measure can be a *distance* (as the terms get closer, the distance decreases) or a *similarity* (as the terms get closer, the similarity increases). Chym considers only similarity measures, but could be easily adjusted to use distances instead.

## 2.4 Information Content

Information content (IC) is an abstract concept that reflects the specificity of a particular object (Resnik, 1999). From information theory, the information content of an object in a particular context can be evaluated as the negative logarithm of the probability of finding that object on that context (Ross, 1994):

$$\mathrm{IC}(x) = -\log \Pr(x) \tag{2.3}$$

Intuitively, equation 2.3 means that a very frequent term is considered to be less informative and vice-versa.

When calculating information content, it should be noted that a function is only meaningful if each term's occurrence contains all its children's occurrences too. In an ontology like ChEBI, this means that for more abstract terms (terms closer to the top of the ontology) this probability includes many terms, decreasing its information content, which, in turn, reflects its low specificity.

The probability function that Chym uses is based on the number of pathways each compound participates in. KEGG databases will be used to determine this function, as will be explained in chapter 4.

The information content is one of the main principles of semantic information. It is not immediately obvious why the probability of finding a particular term should be useful in semantic similarity measures, but its purpose becomes a little more evident when considering one of the most common semantic metrics. From equation 2.3, the more specific a chemical compound is, the higher is its information content. Consider that the similarity between two compounds is the information content of the most informative common ancestor of the two terms (Resnik, 1999). From the nature of an ontology, if the most informative common ancestor of two very specific terms is almost as deep in the DAG as the two terms, then their similarity should be high, since they share a lot of their ancestry; and because the common ancestor is also very specific, the similarity between the two terms is high as requested. On the contrary, two distinct terms will have a very far, unspecific common ancestor, making their similarity small. This is a good term-based method to measure semantic similarity. Pesquita et al. (2008) have shown this and other interesting results on term-based similarity measures. Chym, however, uses two more sophisticated methods (also adapted from the work of Pesquita et al. (2008)), which are detailed in chapter 4.

## 2.5 Chemical Entities of Biological Interest

**Chemical Entities of Biological Interest** (ChEBI) is a freely available database of small molecular entities (distinct isotopes, atoms, ions, molecules etc.). These entities may be products of nature or synthetic products used to intervene in biological processes (Degtyarenko et al., 2007). The ontology also includes classes of molecular entities and partial molecular entities, enabling ChEBI to be organized as an ontology, structuring chemical entities into classes and defining the relations between them. This database was constructed to enable biological relevance to be extracted from its terms.

Unlike other biomedical ontologies, ChEBI does not objectively separate between classes and instances. For example, GO is a DAG of abstract classes that can be attributed to proteins as annotation. On the contrary, ChEBI terms can be classes, instances or even both. *Cysteine*, for example, is an instance of *α-amino acids* and *sulfur-containing amino acid*, but is also the superclass of L-*cysteine*, D-*cysteine*, or other not so common compounds such as *cystine* or *cysteine radical*. Although this approach creates some disadvantages, it is fitting to a database of chemical compounds, since compounds are generally mentioned without full detail and specificity. When biologists refer to cysteine, they are probably referring to L-cysteine. However, it is sometimes desirable not to be so specific, which happens, for instance, in racemic mixtures (a mixture with equal amounts of L- and D-isomers), in which case the less specific term is used.

One of the consequences of not separating classes from instances is that ChEBI includes the elements it is trying to classify. In the following paragraphs, I will use the words *class* and *subclass* to refer to the parent and child of a relationship, respectively.

ChEBI defines six non-cyclic relationship types (Degtyarenko et al., 2007):

(i) **is a** implies that the subclass is an instance of the class, as described above. Subclasses are always more specific than their classes.

(ii) **has part** indicates that a part of the subclass's structure is equal to the class's structure.

(iii) **has functional parent** indicates that the subclass can be derived from the class by functional modification.

(iv) **has parent hydride** indicates that the subclass can be derived from the class by substitution of one or more of its hydrogens (a hydride is "an unbranched acyclic or cyclic structure or an acyclic/cyclic structure having a semi-systematic or trivial name to which only hydrogen atoms are attached" (IUPAC Commission on Nomenclature of Organic Chemistry, 1993)).

(v) **is substituent group from** indicates that the subclass is formed by loss of one or more protons or simple functional groups of the class.

(vi) **has role** indicates that the subclass's behavior is described by the class.

The "is a" relationship is defined by Smith et al. (2005), and is part of the Open Biomedical Ontologies (OBO) community effort to standardize the relationships used in biomedical ontologies. The "has part" relationship is not specified by OBO, but OBO does specify its reversal, a "is part of" relationship. Since chemical compounds with a given substructure are usually more specific than the compound defined by that substructure, the subclasses are also more specific than the superclasses. The other four relationships are unique to ChEBI, since they are specifically related to the properties of chemical compounds, although the "has functional parent", "has parent hydride" and "is substituent group from" fall into the OBO "derives from" relationship.

The other four relationship types ("is conjugate acid of", "is conjugate base of", "is enantiomer of" and "is tautomer of") are cyclic and, as such, unsuitable to use in a DAG structure, which by definition cannot contain cycles.

The ontology is subdivided into three partially overlapping sub-ontologies:

(i) **Molecular structure**, in which the entities are classified according to composition and structure.

(ii) **Role**, in which entities are classified according to their role within a biological context.

(iii) **Subatomic particle**, which classifies particles smaller than atoms.

Each of these sub-ontologies starts as a node in the top of ChEBI, and consists of all the terms that can be connected to that node through a path of non-cyclic relations.

As of the time of the computations (January 2010), the graph of this ontology contains more than 500,000 nodes representing terms (release 64). As stated above, some terms are not chemical compounds but parts of compounds, such as functional groups, that make the ontology structure possible. Also, for each chemical compound, there may be several ChEBI identifiers, resulting from different annotations that were later identified as the same compound.

Besides the ontology, the ChEBI database is enriched with an extensive list of synonyms and manually curated cross-references to other non-proprietary databases, as well as a list of chemical structures.

## 2.6 Kyoto Encyclopedia of Genes and Genomes

**Kyoto Encyclopedia of Genes and Genomes** (KEGG) is a collection of databases categorized into systems, genomic and chemical information. The different KEGG databases are highly integrated in an effort to be an efficient and accurate computer representation of the biological system (Kanehisa et al., 2006).

One of the main components of KEGG is the PATHWAY database, which contains a collection of graphical representations of known pathways. Each metabolic pathway entry integrates information from other databases in KEGG, such as the intervening enzymes (KEGG ENZYME), chemical reactions (KEGG REACTION) and chemical compounds present in the pathway (KEGG COMPOUND).

KEGG COMPOUND is a chemical structure database for metabolic compounds and other chemical substances that are relevant to biological systems. Chym uses entries in the KEGG COMPOUND database to discover the compounds present in the metabolic pathways (KEGG PATHWAY entries). The existence of a mapping between ChEBI and KEGG COMPOUND makes it possible to integrate information from both databases, which in turn enabled the calculation of the frequency of each chemical compound defined in ChEBI. This is the step required to determine the information content of a compound, as discussed previously (see section 2.4).

# Chapter 3

# Current approaches in chemical similarity

This section tries to bring into focus the previous work done in the field of chemical similarity by mentioning some of the previous contributions to this area of research. This section is also important because Chym was developed and validated as an improvement of these previous works, and the results obtained with this method are actually compared to the ones obtained previously.

The comparison of chemical compounds has recently been gaining some focus on the scientific community. Most methods implemented currently use either (a) the chemical structure as the foundation of the comparison, or (b) a combination of structural and physicochemical properties, like the molecular weight and the octanol-water partitioning coefficient etc. (Doniger et al., 2002; Svetnik et al., 2003; Tong et al., 2003). Both these approaches are further discussed below.

## 3.1 Direct structure comparison methods

The great advantage of approach (a), which uses structural properties to derive a chemical similarity measure, is its ability to compare two or more molecules *on demand*, i.e., one can theoretically draw an arbitrary molecule and compare it to a whole database of structures without any prior knowledge of its relevance or physicochemical properties. This is because structural properties are directly derived and easily calculated from the structure of a chemical

molecule, when it is known. It is disadvantageous when the structure of a molecule is not known, but with advances in spectroscopy (Herzberg, 2008) it is becoming increasingly rare to not know the structure of the molecule being studied. Furthermore, the structure of biologically relevant molecules can also sometimes be obtained from crystallography X-ray studies of the protein to which it binds (Ghosh et al., 2008).

There have been attempts to use graph comparison algorithms applied to the chemical structure of two molecules, but since the problem of comparing two graphs is computationally expensive (Raymond et al., 2002), heuristics are used. One way of doing it is to restrict the similarity problem to the search for the maximum common sub-graph (Le et al., 2004). The topology of the molecule can also be used as the base of chemical similarity measures, where, for instance, a molecule can be represented as the matrix of the number of bonds between any two atoms and compared based on those matrices with algebraic methods (Fukunishi and Nakamura, 2009). More often, though, structural similarity is calculated with the aid of fingerprints (see section 2.2).

From the several coefficients that can be used to measure the similarity between two fingerprints, the Tanimoto coefficient is more widely used, at least in chemical similarity approaches, since it was shown (Willett et al., 1986; Salim et al., 2003) that is performs better than the cosine coefficient and is faster to calculate because it does not involve a square root (cf. equations 2.1 and 2.2).

## 3.2 Comparison from physicochemical properties

The approach (b) consists in using structural and physicochemical properties to create a similarity measure between two chemical compounds. As is evident, there is a big disadvantage here when compared to the previous approach. Since most of the physicochemical properties are difficult to estimate computationally, one has to gather them from literature and/or databases, or to conduct experiments to obtain them. There is an increasing number of systems capable of retrieving chemical compound properties through data mining (Cheng et al., 2001; King et al., 2001; Teixeira et al., 2009), but even these systems sometimes fail. However, this is still the approach most commonly used to compare and classify chemical compounds.

Doniger et al. (2002) used Artificial Neural Network (ANN) and Support Vector Machine (SVM) to distinguish compounds capable of crossing the blood-brain barrier (**BBB**) from those that do not cross it. Each compound is described as a 9-dimensional vector, where each element is a physicochemical property of the molecule. An ANN is composed by a number of artificial neurons (a conceptual object that receive several inputs and combines them to produce a single output) arranged in layers, where the first layer gets as input the descriptors of the molecule and the last layer outputs the classification; the SVM method consists in finding the hyper-surface that best separates the vectors of the compounds that cross the BBB from the vectors of the other compounds (Cortes and Vapnik, 1995), in this case in the 9-dimensional space.

Penzotti et al. (2002) used a three dimensional representation of molecules and applied an approach named "four-point pharmacophore". This approach builds millions of descriptors, each being a different spatial arrangement of 4 features with the respective distances, and then determines whether the compound contains each of the descriptors, effectively constructing a big bit-string which can be used like fingerprints, just as previously described. In their work, the "four-point pharmacophore" model was used to predict whether compounds are substrates of the P-glycoprotein (**P-gp**). An SVM approach was also attempted on this set by Xue et al. (2004).

Tong et al. (2003) applied the concept of decision trees to predict whether a chemical compound binds to an **estrogen** receptor. A decision tree consists of several if-then statements, operating over the descriptors, which ultimately come together to create a tree with several branches. The last limbs of the tree classify the compound as positive or negative. In their work, they defined and used decision forests, which are ensembles of several decision trees, where each tree is constructed from the set of descriptors not used in any of the previous trees, in order to minimize the number of misclassifications, and the final output is a combination of the outputs of the trees.

Random forests also use decision trees as its base, as shown by Svetnik et al. (2003). In their work, they used random forests to classify compounds as positive or negative in several sets, including the BBB, P-gp and estrogen sets described above. Unlike the decision tree approach, however, the descriptors used in each tree are randomly drawn from the set of all descriptors, rather then drawn from the set of unused descriptors.

These previous works (as well as the study that is in the origin of this

**Table 3.1: Previous works.** This table summarizes the efficiency of several classification methods.

| Testing set | Classification system | Accuracy | Reference |
|---|---|---|---|
| BBB | Artificial Neural Networks | 75.7% | (Doniger et al., 2002) |
| | Random Forest | 80.9% | (Svetnik et al., 2003) |
| | Support Vector Machines | 81.5% | (Doniger et al., 2002) |
| P-gp | Four-point Pharmacophore | 62.7% | (Penzotti et al., 2002) |
| | Support Vector Machines | 79.4% | (Xue et al., 2004) |
| | Random Forest | 80.6% | (Svetnik et al., 2003) |
| estrogen | Decision Forest | ∼80% | (Tong et al., 2003) |
| | Random Forest | 82.8% | (Svetnik et al., 2003) |

thesis) validate their approaches by using the comparison algorithms as classification systems. **Table 3.1** presents the accuracy values obtained from those systems.

# Chapter 4

# Methodology

In order to develop and validate the hybrid similarity metric for chemical compounds, the **C**hemical **hy**brid **m**etric (Chym), I built a model based both on fingerprints and on the semantic similarity measures developed for the Gene Ontology (GO) (Pesquita et al., 2008). The backbones of these two methods were laid down on previous sections of the Background chapter (section 2.2 for fingerprints and sections 2.3 and 2.4 for semantic similarity) and are now further detailed, but now with Chym in mind.

## 4.1 Structural similarity

To calculate the structural similarity between two molecules, Chym needs a representation of their structures. Because ChEBI contains a list of structures in SMILES, MDL and InChI chemical file formats, these were the formats used, in that order of preference. The rationale for this choice is the wide use of SMILES over MDL and InChI.

For each structure, three fingerprints were calculated. These formats were computed with the OpenBabel software (Guha et al., 2006; Open Babel Project, 2009):

**FP2** All non-branched (linear or possibly circular) fragments of up to 7 atoms are calculated from the initial structure. Each fragment is assigned a number from 0 to 1020 by means of a hash function and the corresponding bit in the fingerprint is set to 1. This is an approach similar to the Daylight fingerprints previously discussed (section 2.2), but uses a different algorithm.

**FP3** The molecule is analyzed and, if a specific pattern is identified, its corresponding bit in the fingerprint is set. The patterns are detailed in a file that is part of the OpenBabel software.

**FP4** This is the same as the FP3 format, but the patterns are defined in a different file, also part of the OpenBabel software.

The reason to use more than one fingerprint format is the flexibility of Chym. For each classification problem, one of these formats will perform better than the others, but the best format is not always the same. For example, the FP2 format covers a wider range of substructures than the other formats, but as a result some substructures will be assigned to the same bit, and the format loses resolution. FP3 and FP4 specify different descriptors, which means the similarity calculated through these two methods will reflect distinct groups of differences. Any good classification scheme must be able to adapt to different problems and different training sets. Having different fingerprint formats (and different semantic similarity formats as well, as can be seen under section 4.2) enables this required flexibility to Chym.

Given two molecules and the corresponding fingerprints $(a_i)$ and $(b_i)$, the similarity score between them is calculated according to the Tanimoto coefficient. This coefficient can be redefined with equation 4.1, which is equivalent to equation 2.1:

$$\mathrm{sim}_{\mathrm{structural}} = \frac{\#\{i|a_i = 1 \wedge b_i = 1\}}{\#\{i|a_i = 1 \vee b_i = 1\}} \tag{4.1}$$

where $a_i$ and $b_i$ are the $i^{\mathrm{th}}$ bit in each of the fingerprints. That fraction is always defined unless $\#\{i|a_i = 1 \vee b_i = 1\} = 0$. In that case, neither fingerprint has a 1-bit, which means they are equal and Chym then forces $\mathrm{sim}_{\mathrm{structural}} = 1$.

From equation 4.1, it can be seen that the structural similarity will be 0 when no single bit is set to 1 on both fingerprints (total disparity) and will be 1 when the 1-bits in the two fingerprints are the same.

## 4.2 Semantic similarity

Following the application of semantic measures for the GO (Pesquita et al., 2008), I developed a similar approach; instead of proteins, however, Chym works with chemical compounds. As has already been stated, there are a number of different ways to measure semantic similarity based on an ontology. I chose to use simUI and simGIC, which seem the most sophisticated

measures, encompassing more information in a single measure. Other possibilities included term-based measures like the Resnik similarity measure (Resnik, 1999), which was discussed in section 2.4. simUI and simGIC are a less time-consuming, because it is not necessary to determine which of the common ancestors is the most informative, and they showed a greater resolution in GO than other term-based methods. As there is no objective reason to suspect that the same will not be true in ChEBI, they were the preferred measures.

Having two methods for calculating semantic similarity plus several complementing sub-ontologies gives Chym flexibility in the semantic side of the approach, in parallel to what happens with structural similarity (see previous section).

For the rest of this section, consider $c$, $c_1$ and $c_2$ as chemical compounds and $\text{asc}(c)$ as the set of ancestors of the chemical compound $c$, including $c$ itself. Furthermore, **Figure 4.1** represents a very simplified view of an example ontology. While reading this section, it is useful to refer to it in order to understand the concepts and the usefulness of the formulas. It is worth noting that the figure represents only an exemplification, not a real ontology.

### 4.2.1 simUI

simUI is a graph-based measure, which means that it considers the terms and all of their ancestors in the graph of the ontology. It is defined as follows (Gentleman, 2005):

$$\text{simUI}(c_1, c_2) = \frac{\#\{t \mid t \in \text{asc}(c_1) \cap \text{asc}(c_2)\}}{\#\{t \mid t \in \text{asc}(c_1) \cup \text{asc}(c_2)\}} \tag{4.2}$$

This measure is purely graph-based. Only the nodes and edges are considered, without any information content being used in the calculation. Intuitively, simUI measures the amount of common ancestry between two terms. When two terms share most of their ancestry, they share most of their meaning, and the simUI similarity measure returns a high value to reflect this fact. On the other hand, two dissimilar terms will share few ancestor terms, and simUI will be low.

### 4.2.2 simGIC

It is known, however, that for ontologies where term specificity is not well correlated with term depth (the minimum number of relations between the

term and the root of the DAG), methods based on information content (IC) are preferable (Pesquita et al., 2008). simGIC is a combination of the graph-based simUI metric with the information content properties of compounds through equation 4.3 (Pesquita et al., 2008). In fact, both definitions are very similar, but simGIC weights each ancestor with its information content:

$$\text{simGIC}(c_1, c_2) = \frac{\sum\{\ \text{IC}(t) \mid t \in \text{asc}(c_1) \cap \text{asc}(c_2)\}}{\sum\{\ \text{IC}(t) \mid t \in \text{asc}(c_1) \cup \text{asc}(c_2)\}} \tag{4.3}$$

where IC is the information content, as calculated through the IC equation:

$$\text{IC}(c) = -\log \Pr(c) \tag{4.4}$$

The corpus used to calculate the probability of finding a compound is KEGG, and the probability of finding a compound is the fraction of pathways in which it participates.

Equation 4.3 shows that unspecific ancestors (like the root of the ontology or other early compounds) contribute very little to the similarity measure. This has two main consequences:

(i) simGIC similarity measure is generally lower than simUI (see why in appendix C), and

(ii) two completely equal compounds share a similarity of 1, but the similarity decreases rapidly when the compounds draw apart, which increases the resolution of this method for similar compounds.

For both simUI and simGIC, the similarity value is between 0 and 1 because the intersection of two sets is always a subset of their union, and therefore the numerator is always smaller then the denominator.

## 4.3 Hybrid metric

Until this point, I presented two orthogonal metrics able to measure the similarity between two chemical compounds, each using a different set of properties. My intent, however, is to join them together to produce a hybrid metric that takes into account both structural and semantic information.

Since all metrics detailed above fall in the closed interval $[0, 1]$, I propose the following definition for the hybrid similarity:

$$\text{sim}_{\text{hybrid}} = \alpha \cdot \text{sim}_{\text{structural}} + (1 - \alpha) \cdot \text{sim}_{\text{semantic}} \tag{4.5}$$

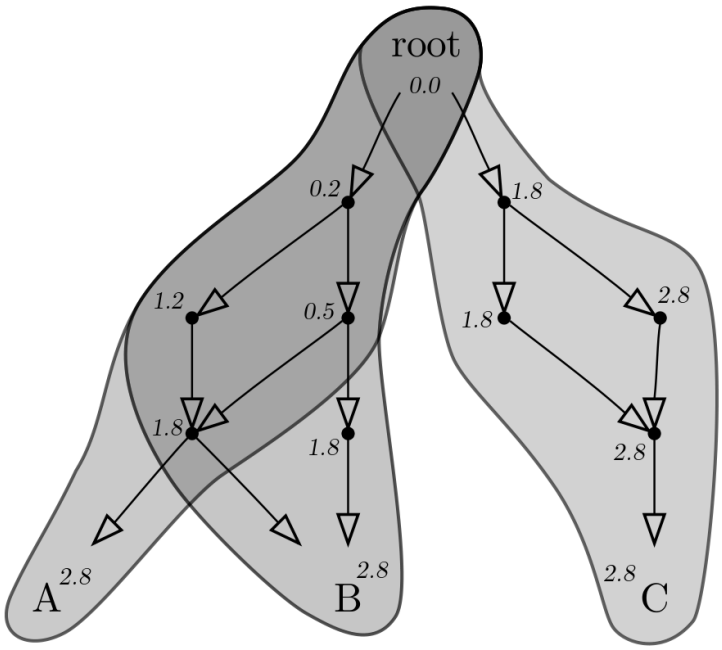



**Figure 4.1: Semantic similarities in action.** An example ontology with terms specified as dots or words and relationships represented with arrows. Information content of each term was fictitiously calculated and written near it. From this figure, it is possible to calculate, for instance, $\text{simUI}(A, B) = \frac{5}{8} = 0.625$ or $\text{simGIC}(B, C) = \frac{0.0}{20.3} = 0.0$.

where $\alpha$ is a real number from 0 to 1. When $\alpha = 0$, the identity degenerates into pure semantic similarity and when $\alpha = 1$, into pure structural similarity. Furthermore, with the several fingerprint formats and semantic similarity measures presented, equation 4.5 is not a metric but a collection of metrics.

In the introduction (chapter 1), I mentioned that the main objective of this thesis is to create a metric that reflects the biological relevance of a compound based on its semantics in such a way that the similarity measure for two biologically related compounds is higher than what would be obtained with purely structural methods. The equation given in equation 4.5 integrates both semantic and structural information, but if semantic similarity is lower than structural similarity, the definition above will return a value lower than the purely structural similarity. This is not in agreement with the objective delineated above. But the absolute value of similarity is not the important

parameter. Consider a list of pairs of compounds ranked according to similarity: the metric is considered successful (it will conform to the hypothesis) if the rank of a pair of semantically related compounds increases, or, in other words, if its position in the list goes up.

## 4.4 The Chym approach to classification

To assess the performance of Chym when it comes to conforming to the hypothesis rewritten above, it was crucial to use it in a way that could be compared and evaluated based on previous works.

One of the possible uses of Chym is the application of this similarity metric to classify compounds. The ideal system gets as input a set of chemical compounds that possess a common property (*positive* compounds) and a set of chemical compounds without that property (*negative* compounds), and then determines whether other chemical compounds also possess the same property. This is also the approach used by SVM and random forests, for example, where the input serves as a training set that is used to create a classification model. In Chym, the model consists of the set of positive compounds and a threshold that is used to decide whether a compound is positive or negative.

Given the training set of positive and negative compounds, the algorithm used to predict whether another compound is positive or negative is the following (**Figure 4.2** contains a visual representation of the process):

1. Compare each compound in the training set with the positive compounds in that group. The comparison of a positive compound with itself is excluded, since this value (which is always 1) could introduce a bias into the rest of the algorithm.

2. For each of these compounds, determine its *activity coefficient*, which is the unweighted average of the results in step 1. A compound will be classified as positive if its activity coefficient is above a threshold, which still needs to be calculated.

3. Determine the *threshold of activity*, $t$. To do this, Chym uses all the coefficients calculated in step 2 as potential thresholds, and classifies the compounds in the training set as positive or negative accordingly. The coefficient that minimizes the number of misclassifications in the input set is chosen. This ends the training step.

4. For all compounds in the validation set, Chym calculates their activity coefficient (average of similarities between a compound and all positive compounds in the training set), and classifies it as positive if the activity coefficient is greater than or equal to the threshold of activity $t$, and as negative otherwise.

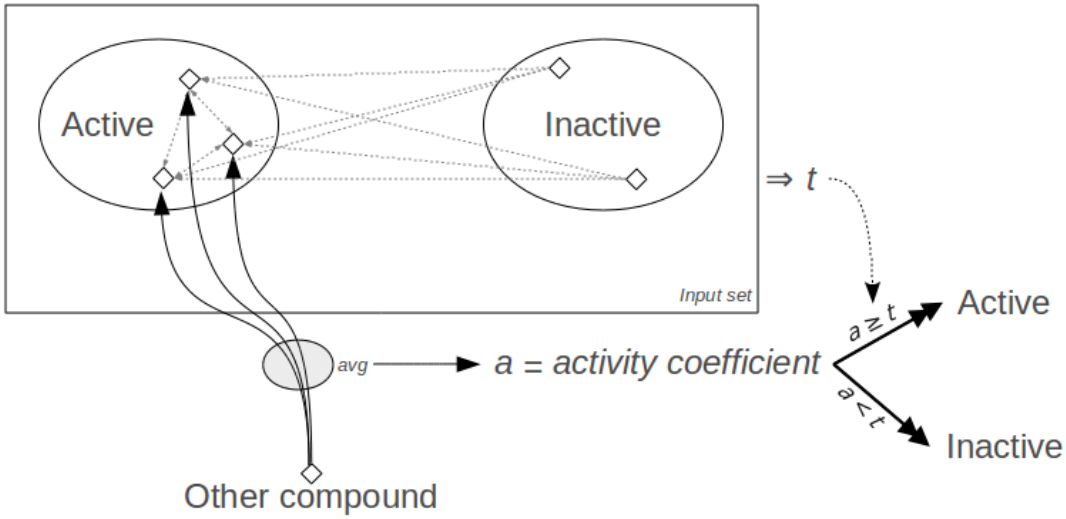



**Figure 4.2: Visual explanation of Chym's classification process.** To classify a compound that is not in the input set, Chym starts by comparing all the compounds in the input set with the known positive compounds (dotted, straight arrows) in order to train the model and obtain a threshold value. This training step is represented here in the rectangle, which yields $t$, the threshold of classification. Then, the new compound is compared with all the positive compounds in the input set (continuous, curved arrows) and the activity coefficient $a$ is calculated as the average of these values (here represented by the ellipsis). The compound is classified as positive or negative based on the comparison of the activity coefficient and the threshold. Compounds are represented with the diamond-shaped symbols.

## 4.5 Assessment of the quality of Chym

Since the performance of the previous systems was assessed based on their accuracy (cf. **Table 3.1**), I will report mainly the accuracy of Chym. However, accuracy may not be the best parameter to evaluate whether a classification tool is good. It is very common in bioinformatics to construct what is called a *confusion matrix*, which attributes to each classification one of four labels: true positive (TP), true negative (TN), false positive (FP) or false negative (FN), respectively to positive compounds being classified as positive, negative compounds as negative, negative compounds as positive and positive compounds as negative (the last two are the misclassifications). Refer to **Table 4.1** for a better visualization of a confusion matrix.

There is no single number that can perfectly describe a confusion matrix. The accuracy of a prediction is the fraction of correct classifications, but it fails to assess the true quality of a system in some cases, particularly when the distribution of positive and negative compounds is not balanced:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4.6}$$

F-measure is also very common in bioinformatics, and is mainly used in data mining operations. It does not take into account the number of false negatives. It is defined in terms of precision ($p$) and recall ($r$), which, in turn, are defined in terms of the variables in the confusion matrix:

$$\begin{aligned} p &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ r &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{F-measure} &= \frac{2pr}{p + r} \end{aligned} \tag{4.7}$$

These two coefficients are in the interval $[0, 1]$, with 0 corresponding to a totally unsatisfying tool and 1 to the perfect prediction tool.

**Table 4.1: A confusion matrix.**

| **Experimental** | **Real** | |
|---|---|---|
| | Positive | Negative |
| Positive | TP | FP |
| Negative | FN | TN |

When the sizes of positive and negative groups are disparate, neither of these two coefficients describe accurately the quality of the prediction. For instance, if 90% of the compounds are positive, and the tool assigns "positive" to all compounds, the accuracy would be 90%, but as a prediction tool, this would correspond to an unsuccessful attempt.

The Matthews correlation coefficient (MCC) is the least used of these three measures, but it is also the one that more faithfully reflects the quality of the tool when the number of positive compounds is very different from the number of negative compounds (Baldi et al., 2000):

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{4.8}$$

Unlike the other two, this coefficient measures a correlation, and as such varies from $-1$ to 1, with $-1$ reflecting a tool that fails all predictions, 0 a seemingly random prediction tool and 1 a perfect prediction.

## 4.6 Implementation

The methods used to structurally compare compounds are implemented by the software OpenBabel (Guha et al., 2006; Open Babel Project, 2009), an open sourced chemical toolbox containing a number of different functions, including the ability to store and analyze data from molecular modeling, particularly, to this work, from the structure of chemical compounds. Version 2.2.3 of the OpenBabel software was downloaded and installed locally on December 2009.

The semantic similarity demanded for the application of semantic tools on the ontology used, ChEBI. As has been previously done in the work of Grego et al. (2010), it was necessary to reorganize the ChEBI ontology so that it could fulfill Chym's purposes. All cyclic relationships ("*is tautomer of*" etc.) were removed, and the other relationships were merged into a single "*is a*"-like relationship. ChEBI identifiers pointing to the same chemical compounds were also merged into a single node. This resulted in the production of three independent DAGs, one for each branch of the main ontology (*structure*, *role* and *subatomic*), and a forth DAG resulting from joining the other three (*all*). With this modification, it is possible to directly calculate simUI similarities with equation 4.2.

To calculate the IC needed for the simGIC metric, Chym needs a corpus where chemical compounds are referenced, KEGG PATHWAY. With this corpus,

the value of $\Pr(c)$ from equation 4.4 is the fraction of pathways where the compound $c$ or any of its descendants are present. To map ChEBI identifiers into KEGG identifiers, I used the ChEBI cross-references. Sometimes, however, these references were ambiguous (some ChEBI ids point to two or more KEGG COMPOUND ids, as happens, for instance, with compound ChEBI:16218, which points both to C:00663 and C:01135 from KEGG COMPOUND). For this reason, whenever a ChEBI id $c$ had more than one KEGG COMPOUND reference, Chym uses all of them to determine the number of pathways in which $c$ participants.

The information described above was stored locally in an manner that can be easily accessed:

- All ChEBI compounds, with their names as defined in that database;
- The set of cross references between ChEBI and KEGG COMPOUND;
- The four DAGs in closed form, where a row in the table is of the form $c_1, c_2$ and determines that $c_1$ is part of the ancestry of $c_2$;
- The IC of the compounds present in at least one of the KEGG pathways.

Since there are 3 fingerprint formats (FP2, FP3 and FP4), and semantic similarity can be calculated based on 4 different DAGs (all, role, structure and subatomic) and with 2 different methods (simUI and simGIC), the Chym's approach consists of $3 \times 4 \times 2 = 24$ different similarity metrics, each tuned with a real parameter $\alpha$.

With this setting, Chym is now ready to be validated. In the next chapter, I will show the effectiveness of Chym, based on its performance when used as a binary classification tool.

# Chapter 5

# Assessment

Up until now, this document presented the background information needed to understand Chym, the current approaches and their results, and the methodology followed in the implementation of Chym. This chapter now shows the results obtained with this tool. To validate the effectiveness of Chym, I used it as a binary classification tools in the BBB, P-gp and estrogen sets described in section 3.2 and **Table 3.1**.

This chapter is the direct application of the methodology described in chapter 4 to real world problems. Here, the positive compounds are the positive compounds of each of the three data sets used: in the BBB set they are the molecules that cross the blood-brain barrier; in the P-gp set they are the substrates of P-glycoprotein, and in the estrogen set they are the estrogen-binding molecules. The negative compounds are, therefore, the compounds from each set that do not possess the corresponding biological activity.

The chapter is divided in four sections. First, the three data sets are analyzed and converted to something that Chym can use. Then, I delineate the steps followed to validate Chym. The last two sections are the most critical of the chapter, since they first show the results achieved by the validation of Chym in section 5.3 and then proceed to discuss their meaning, in section 5.4.

## 5.1 Sources of data sets

In all the three input sets retrieved from the previous works, BBB, P-gp and estrogen, the compounds are listed by name. Therefore, the first step in the assessment of Chym was to translate that list of names into ChEBI identifiers,

so that the semantics of each compound can be retrieved. The task of getting a ChEBI id from a name was accomplished by a string matching technique that I have called "C-match", for Chemical Match.

To this end, each ChEBI name was converted into a list of words, where a word is a string containing only letters or only numbers. It is worth mentioning that a ChEBI compound may have several distinct names. For instance, the names of ChEBI:33813 include *((18)O)water*, *Water-(18)O* and *Heavy-oxygen water*. Those three names are thus converted into the lists {18, O, water}, {Water, 18, O} and {Heavy, oxygen, water}, where the order and case are unimportant. This means that C-match considers the first and second lists equal.

A name from the list of compounds in the input set is matched to a ChEBI compound $c$ if the list of words in the name is equal to at least one of the lists of words of the compound $c$. In case names from several ChEBI compounds meet this requirement, the difference between the non-alphanumeric characters is used to create a scoring system among the compounds, and those with the absolute minimum distance are considered. This can still lead to two or more matches. Because ChEBI is continually growing, I estimate that older compounds in the ontology are usually more correctly annotated and tend to have smaller identifiers. Therefore, in case of more than one match, C-match chooses the smallest ChEBI id. Only compounds with a molecular structure in the ChEBI database were considered.

Since the ontology does not contain all the possible molecules, it was impossible to get a full mapping between the names in the three sets and ChEBI compounds, which means that the sets used by Chym were shorter versions of the original ones. From now on, I refer to these smaller sets as purged versions and denote them as $\text{BBB}_\text{p}$, $\text{P-gp}_\text{p}$ and $\text{estrogen}_\text{p}$. **Table 5.1** shows the fraction of compounds in each of the three sets that are present in the ontology.

The results presented in the table show a significant reduction in the size of all three sets after converting the names into ChEBI identifiers. Since an exhaustive validation process includes the comparison of the results obtained by Chym with previous results, I had to compare Chym's performance with those previous classification systems'. However, facing these values, I chose to directly compare Chym's results only to the ones obtained with the blood-brain barrier, because (a) the BBB set is the one with higher percentage of coverage,

**Table 5.1: Fraction of compounds in the ChEBI ontology.** This table summarizes the fraction of names in each of the three input sets that C-match was able to map into the ChEBI ontology. Coverage for positive and negative compounds is detailed.

| Testing set | ChEBI coverage | | |
|---|---|---|---|
| | positive | negative | overall |
| BBB | 74/180 | 79/144 | 47.2% |
| P-gp | 57/109 | 24/87 | 41.3% |
| estrogen | 42/132 | 59/101 | 43.3% |

(b) after purging, it remains the biggest set, and as such is fitter to be broken into training and validation sets without losing too much information, (c) it is the set where the distribution of positive and negative compounds in the purged version is more faithful to the distribution in the original set, and finally (d) it has a more balanced distribution of positive vs. negative compounds. These are the reasons that led me to choose the BBB set as the main validation set for Chym. Nonetheless, it is important to observe its performance with other, not so *well behaved* sets. As such, I applied Chym's classification algorithm to those sets as well, but the analysis of those results was not as deep.

## 5.2 Validation process

The BBB set was first described in Doniger et al. (2002), where the authors use an artificial neural network (ANN) and a support vector machine (SVM) to classify several chemical compounds as either able to cross the blood-brain barrier (positive) or unable to do so (negative). The paper showed that SVMs are more efficient in this particular classification problem than ANNs (see **Table 3.1**). The set was further used by Svetnik et al. (2003), where a random forest (RF) was grown. The authors of this work showed the efficiency of this system in several chemical compound sets, but the results obtained for the BBB set in particular were not better than the ones obtained with SVM. For this reason, I compared Chym's results to the 81.5% accuracy obtained in the work of Doniger et al. (2002).

In order to make an unbiased comparison between Chym and SVMs, I addressed the validation process in three steps, which were devised so that for every two consecutive steps, only one specification of the process changed:

1. The SVM model described in Doniger et al. (2002) was used to replicate the results reported in that paper with the original set;
2. The same SVM model was used in the purged set, $\mathrm{BBB_p}$;
3. Finally, the SVM model was replaced with Chym. It must be noted that there are 24 metrics, each having a real parameter $\alpha$. For this task, Chym used $\alpha$ values running from 0 to 1 in steps of 0.01, making a total of $24 \times 101 = 2424$ metrics, and considered the one with higher accuracy value the best one to classify the compounds.

For the SVM approach, I retrieved the compounds' properties from the article as 9-dimensional vectors and used the SVMlight software with a radial basis function kernel (Joachims, 1999), as is described in the original paper.

Moreover, to decrease the potential bias in my analysis, I implemented several validation methods. The first one is a 30-fold cross-validation-like process, described in Doniger et al. (2002), which I dub "Cross25", and which consists of these steps:

1. First, 25 positive compounds and 25 negative compounds are randomly removed from the input set; they are now the validation set;
2. The remaining set was used to train the model, as described in the algorithm on section 4.4;
3. The 50 compounds in the validation set are classified according to the model learned in the previous step;
4. Steps 1–3 are repeated 30 times, and the averages of the accuracy values and the Matthews correlation coefficient (MCC) values are recorded.

Because the set is reduced to 47.2% of the original size (cf. **Table 5.1**), I further implemented a Cross12 method, equal in everything to that previous method, except that instead of picking 25 positive and 25 negative compounds from the set to form the validation set, I picked $25 \times 0.472 \approx 12$ positive and negative compounds.

The third validation method is called leave-one-out, a method widely used and well documented (Baldi et al., 2000). The difference from the above is that in each repetition only one compound stands as the validation set, and all compounds are used as validation set exactly once. This has the advantages of not depending on randomness and of using the maximum information possible from the primary set, since only one compound is being withdrawn from the training set.

To complement this information, two other processes of validation were tested: bootstrap, where the training set is randomly withdrawn from the primary set, *with reposition*, until a training with the same size as the original set size is obtained, and the validation set contains all the non-selected compounds. This procedure is repeated 30 times and an average of the performance is calculated. The fifth process is named $k$-fold validation, where the original set is partitioned into $k$ smaller sets, and each of the partitions is used once as the validation set. Again, this is performed 30 times and the reported value is simply an average of the individual performances obtained.

The last step in the assessment of Chym was to predict some new positive compounds in each of the three sets. I calculated the activity coefficient of all compounds in the ChEBI ontology that are annotated with a structure (based on all positive compounds in each of the purged sets) and retrieved the ones whose coefficient was higher.

## 5.3 Results

With Chym built as detailed in the previous section, its performance as a binary classification tool can be assessed. In chapter 4, I delineated the main procedure used to validate Chym. Here I will show some of the most important and representative results. As discussed in section 4.5, accuracy is the value that other systems usually report, and as such it is also the value that should be used to compare Chym with these systems. Moreover, since the positive and negative groups of the $BBB_p$ set have approximately the same size, accuracy is as good as the Matthews correlation factor when used to measure the quality of Chym. This is not valid for the P-$gp_p$ and the $estrogen_p$ sets, however. As such, I will report the Matthews correlation coefficient as well. For an homogeneous and complete comparison between the previous systems and Chym, **Table 5.2** presents the Matthews correlation coefficient (MCC) as it

would have been reported in the previous works. These values were computed based on the true positive, true negative etc. values reported in the original papers. For the random forest approach, I used the code provided by the authors of the paper to replicate their results but then used them to compute the MCC, which means that the values presented are an average of 50 5-fold cross-validation processes. For the ANN and SVM approaches of the BBB problem, the authors report only the mean values of 30 true positives, true negatives etc. These were the values used to compute the MCC, which means that they are estimates, at best. Finally, the MCC calculated for the SVM approach in the P-gp problem is an average of the 5 experiments the authors reported.

**Table 5.3** shows the main results of the validation process, including the attempt to replicate the results of Doniger et al. (2002). Given that Chym has 24 different metrics, each one tuned with a real parameter $\alpha$, I had to select one of the possibilities. The best combination for the BBB problem was FP3 fingerprint format with semantic similarity calculated for all the ChEBI ontology with a simGIC method, with 30% of weight to structure and 70% to semantics ($\alpha = 0.30$). Actually, this is not exactly the best approach for some of the rows in that table. For instance, using a bootstrap validation approach,

**Table 5.2: Previous works – Matthews correlation coefficient.** This table summarizes the efficiency of the classification methods presented in **Table 3.1**, but with the Matthews correlation coefficient. An asterisk $*$ marks the work where insufficient information was given, making the coefficient impossible to calculate. Values marked with the $\sim$ sign are only estimated. See the text for an explanation.

| Testing set | Classification system | MCC |
|---|---|---|
| BBB | Artificial Neural Networks | $\sim$0.549 |
| | Random Forest | 0.605 |
| | Support Vector Machines | $\sim$0.628 |
| P-gp | Four-point Pharmacophore | 0.315 |
| | Support Vector Machines | 0.591 |
| | Random Forest | 0.591 |
| estrogen | Decision Forest | $*$ |
| | Random Forest | 0.647 |

**Table 5.3: Replication of the results of BBB.** For the Cross25 and Cross12 methods, the values are the mean of 30 experiments, as explained in the previous section. The Chym results were obtained for FP3 fingerprint format, simGIC semantic method using the entire ontology, and $\alpha = 0.30$. The bootstrap and $k$-fold validation processes were only attempted with the Chym approach.

| Set | Approach | Validation method | Accuracy | MCC |
|---|---|---|---|---|
| BBB | SVM | Cross25 | 81.3% | 0.608 |
| $BBB_p$ | SVM | Cross25 | 72.7% | 0.469 |
| $BBB_p$ | Chym | Cross25 | 88.3% | 0.772 |
| BBB | SVM | Cross12 | 80.4% | 0.593 |
| $BBB_p$ | SVM | Cross12 | 73.6% | 0.451 |
| $BBB_p$ | Chym | Cross12 | 90.0% | 0.804 |
| BBB | SVM | Leave one out | 80.9% | 0.612 |
| $BBB_p$ | SVM | Leave one out | 73.2% | 0.464 |
| $BBB_p$ | Chym | Leave one out | 90.2% | 0.809 |
| $BBB_p$ | Chym | Bootstrap | 88.1% | 0.766 |
| $BBB_p$ | Chym | 10-fold | 90.9% | 0.826 |

the best combination of parameters was FP3, simGIC, role, $\alpha = 0.24$, but from the 2424 hybrid metrics tested (each metric, with $\alpha$ values from 0 to 1 in steps of 0.01), the parameters FP3, simGIC, all $\alpha = 0.30$ are the $15^{\text{th}}$ best. Therefore, I specifically chose these parameters because they are always close to the best one and this allows me to directly compare the results with each other. The table is separated in several sections, where each section is the result of applying a different validation method.

These results show the superiority of Chym when compared to the SVM approach, not only when using the accuracy values but also the Matthews correlation coefficient. Moreover, when all the sections of the table are compared, it is clear that the validation method does not affect significantly the results. Since the "leave one out" approach is widely used (Baldi et al., 2000), at least when compared to the Cross25, a method tailored to the BBB problem by Doniger et al. (2002), I performed the remainder analysis of Chym's results with this method. The second reason behind this choice is that it is in between the other two methods, with the worse accuracy and correlation

**Table 5.4: Classification system derived from the Chym comparison method.** For each problem, the Chym parameters that yielded the best results are reported. The validation process used was leave one out.

| Set | Chym Parameters | Accuracy | MCC |
|---|---|---|---|
| $BBB_p$ | FP3, simGIC, all, $\alpha = 0.28$ | 90.9% | 0.821 |
| $P\text{-}gp_p$ | FP4, simUI, all, $\alpha = 0.66$ | 87.7% | 0.704 |
| $estrogen_p$ | FP4, simGIC, role, $\alpha = 0.42$ | 84.2% | 0.691 |

values corresponding to the bootstrap approach and the best corresponding to the 10-fold approach.

The second row in the same table (along with the fifth and the eighth) show that the accuracy and MCC of the SVM method used previously decrease significantly when some of the compounds in the set are removed. This means that, at least for the 9 descriptors used, the SVM classification is very dependent on the set size. However, the same purged set can be used by Chym, and still achieve an accuracy almost 10% superior to the original and with MCC approximately 0.2 units higher then the original. This seems to indicate that Chym is less sensitive to the size of the training set and performs well even with small sets. **Table 5.4** reinforces this conclusion, since the performance of Chym with the $P\text{-}gp_p$ and $estrogen_p$ sets, which are also about 60% shorter than the original ones, is still higher than (for the P-gp set) or comparable to (for the estrogen set) the value obtained with the random forest approach, the best method applied so far to those sets.

When overlaid to the values on **Table 3.1** and **Table 5.2**, the values on this table also reflect the superiority of Chym compared to those previous methods. Not only is there an increase in accuracy but also in MCC, mostly evident in the BBB and P-gp problems. The accuracy and MCC for the estrogen problem do not increase much, but the fact that the set used by Chym is less than half of the one used by the random forest and decision forest approaches seems to indicate that these values would in fact increase if more of the compounds were present in the ChEBI ontology.

**Table 5.5** contains the values for Chym's prediction accuracy and MCC, calculated for different $\alpha$ values. For each set, the parameters used with Chym are those which reached maximum accuracy for some value of $\alpha$, which means that the parameters are the same as the ones in **Table 5.4**, minus the $\alpha$

**Table 5.5: Effect of the $\alpha$ parameter in Chym performance.** The Chym parameters used are the ones in **Table 5.4**, except that instead of a single $\alpha$ value, the table shows the performance for several values. Accuracy values are reported with MCC in parenthesis. Validation was performed with a leave one out approach. Bold values are the maximum for each column.

| **Alpha** | **BBB$_p$** | **P-gp$_p$** | **estrogen$_p$** |
|---|---|---|---|
| 0.0 | 81.0% (0.628) | 74.1% (0.317) | 73.3% (0.452) |
| 0.1 | 86.9% (0.746) | 74.1% (0.356) | 74.3% (0.497) |
| 0.2 | 88.9% (0.782) | 79.0% (0.491) | 74.3% (0.490) |
| 0.3 | **90.2% (0.809)** | 76.5% (0.431) | 79.2% (0.577) |
| 0.4 | 88.2% (0.766) | 81.5% (0.561) | **84.2% (0.691)** |
| 0.5 | 85.0% (0.699) | 84.0% (0.620) | 83.2% (0.605) |
| 0.6 | 83.0% (0.660) | 85.2% (0.654) | 78.2% (0.560) |
| 0.7 | 83.0% (0.660) | **86.4% (0.679)** | 81.2% (0.622) |
| 0.8 | 81.0% (0.622) | 82.7% (0.576) | 76.2% (0.515) |
| 0.9 | 77.1% (0.546) | 84.0% (0.611) | 71.3% (0.411) |
| 1.0 | 71.9% (0.437) | 85.2% (0.637) | 79.2% (0.580) |

part. It is visible that, in all the three Chym systems, the accuracy and MCC start by increasing at first, reaching a maximum at some point in the table (highlighted in bold), and then decrease again. This shows that using the hybrid measure is better than using only purely structural or purely semantic metric. When this same analysis is applied to other Chym parameters, the same behavior is observed, which confirms the idea that, the integration of structural and semantic information in a single metric helps to increase the prediction power.

Finally, **Table 5.6** shows the most positive ChEBI compounds, as defined by the activity coefficient, retrieved for each set. For each compound, the table gives a reference that shows that the compound is indeed positive. These results make it clear that many of the compounds with predicted activity are, in fact, positive compounds (they cross the blood-brain barrier, are substrates to P-glycoprotein or ligand to the estrogen receptor), which also contributes to the idea that the Chym method is effective. The only false positive in that list is ChEBI:5078, flavonol (Zand et al. (2000) showed that this compound is not an estrogen receptor ligand). However, the class of compounds named flavonoids, into which flavonol is classified, is known to contain several com-

**Table 5.6: Discovery of new positive compounds in each set.** The activity coefficients of the most positive compounds in ChEBI was calculated through compared to the positive compounds in each set. For each compound, a reference showing that the compound is indeed positive is given.

| Compound | | Set | Coefficient | Ref. |
|---|---|---|---|---|
| ID | Name | | | |
| 1015 | orthanilic acid | $BBB_p$ | 0.503 | (Gupta, 2006) |
| 2654 | aminoglutethimide | $BBB_p$ | 0.489 | (Unger et al., 1986) |
| 2089 | O-methylserotonin | $BBB_p$ | 0.477 | (Kaminka, 1971) |
| 3638 | chloroquine | $BBB_p$ | 0.475 | (Ohtsuki and Terasaki, 2007) |
| 2430 | aconitine | $P\text{-}gp_p$ | 0.474 | (Chen et al., 2009) |
| 1883 | 4-hydroxystyrene | $estrogen_p$ | 0.577 | (Dall'Acqua et al., 2009) |
| 5078 | flavonol* | $estrogen_p$ | 0.577 | (Zand et al., 2000) |
| 5262 | galangin | $estrogen_p$ | 0.577 | (So et al., 1997) |

* This compound is a false positive. The reference shows the compound inactivity.

pounds that bind to the estrogen receptor (Markiewicz et al., 1993; Miksicek, 1993). Because these compounds share a strong similarity, both structural and semantic, the activity coefficient turned out to be the same, which makes it a false positive.

## 5.4 Discussion

Apart from the discussion of the results, it is important to mention that Chym's construction was not as straightforward as may seem from the reading of this document. Some decisions had to be made. For instance, Chym's classification algorithm (the steps followed to classify a compound not in the input set as positive or negative, as described in section 4.4) uses the average of the similarities between the new compound and the positive compounds of the input set. This was not the first choice. However, it was the choice that gave Chym the ability to predict with as much as 90.9% accuracy. The results that led to this and other decisions are detailed and explained in appendix A.

Moreover, there are mainly four other points of discussion that should be addressed, now that all the results have been analyzed.

Firstly, it is important to note that it is not artificial to select the best

metric for a specific case. Each classification problem is different from the others, and the molecular characteristics that are important to distinguish between positive and negative molecules are not the same. For instance, while the FP3 fingerprint format may be good at detecting some substructures that are important in the BBB problem, it may miss the relevant structures in other problems. This actually happens, since Chym's parameters to solve the BBB problem include the FP3 format. Furthermore, some problems may be better solved with a stronger focus on the semantic information, in which case the $\alpha$ value will reflect this; this also happens in the BBB problem, where only 28% of the metric is structure-based. Choosing one of the 24 metrics and an $\alpha$ value is no different than choosing the correct descriptors in a SVM, random forest or any other approach presented so far, and is in fact what makes Chym so powerful.

The second issue is the coverage of compounds present in the ChEBI ontology. For instance, to address the estrogen problem, Chym had to reject almost 60% of the compounds in the estrogen data set because there was no mapping between them and the chemical ontology. Even so, the accuracy of prediction is slightly higher than the one obtained with the random forest approach, which uses the whole data set. This seems to suggest that, as the ChEBI ontology grows, Chym's prediction power will increase. Since ChEBI is a database developed and maintained by EBI, a very prominent institute in the Bioinformatics field, the growth of ChEBI is almost assured. As a side note, in October 2009, ChEBI has integrated in its database all entries in ChEMBL, a database of bioactive drug-like small molecules, and because the manual curation process is a long one, the ontology is bound to have missing compounds and erroneous relations. With time, ChEBI will mature, providing Chym with a better ontology which will probably improve these results.

Third, it is important to discuss some of the results obtained when Chym used other classification algorithms. As a matter of fact, after analyzing several algorithm to predict activity, the parameters that maximize Chym's prediction power remain very close to each other (cf. appendix A and **Table A.3**). This means that Chym achieves approximately the same result even with different algorithms, which, in turn, contributes to the idea of stability of Chym.

Finally, Chym's high accuracy values could be due to a possible term in the ontology that classified compounds as able to cross the blood-brain barrier, as substrates to the P-glycoprotein or as estrogen receptor ligands. Admittedly,

if there were such terms in the ChEBI ontology, Chym would be biased and would report high accuracy values because it was using the information it was trying to validate as a means to prove its efficiency. As it turns out, no term in the ontology refers to the words "brain", "P-glycoprotein" or "permeability" (the meaning of the P in P-glycoprotein). "Estrogen receptor" appears twice, in "estrogen receptor modulator" and "estrogen receptor antagonist", but these two terms have only a total of 5 descendants in the ontology, and none of them is present in the set $\text{estrogen}_{\text{p}}$. This fact suggests that Chym can be used in many classification and similarity problems, even if they are not well represented in the ontology.

# Chapter 6

# Conclusions

Over the last few decades, there has been a shift in the term *automatic* in the field of chemical science: while 40 years ago it meant the mechanization used to help in the laboratory, it now refers to the computational effort that can be expended to analyze and organize existing data and to create new information based on the currently existing one. Specifically in the field of chemical compound similarity, cheminformatics has been applied to a certain extent to data sets in order to compare and classify those compounds. In general, these existing methods make use of structural and/or physicochemical properties of the molecules in order to compare them, which is a good method because of the structure-activity relationship (SAR) premise, which states that the biological role of a molecule is a function of its structure. While previous studies contribute to the idea that this premise is correct, it may fail in cases where similar molecules have different roles, or different molecules have the same role (as happens with L- and D-amino acids, for instance).

With the work of this thesis, I presented a method that tries to solve this problem, through a novel approach to the chemical similarity problem, namely, the use of a hybrid metric that encompasses both structural and semantic information. The tool developed, which I named Chym, implements a system capable of handling this information. It is based on fingerprints for structural comparison and on an adaptation of a previous work that used semantic information on proteins.

The results reported in this thesis are a compelling evidence for the effectiveness of Chym as a classification system. The validation process consisted in using three data sets previously described and used with other classifica-

tion methods. Used on those sets, Chym achieves high performance, measured through accuracy and Matthews correlation coefficient, predicting the correct blood-brain barrier permeability up to 90.9% of the time. Other results include correct classification of 87.7% of the compounds in the P-gp set and 84.2% in the estrogen set, which correspond to the classification of compounds as substrates of the P-glycoprotein or as ligands to an estrogen receptor respectively. Parallel to these results, I also showed that the use of a hybrid metric that uses both structural and semantic information is better suited for this kind of problems than a system which uses only one of these data. Finally, Chym was used to correctly predict new positive compounds in each of the three sets.

These results provide substantial evidence for validating the proposed hypothesis: compounds sharing a biological role should have a similarity measure higher than the one obtained using only structural or physicochemical properties, reflecting that fact; conversely, the measure should also reflect different biological relevance even if both molecules are similar in structure.

In the future, it would be interesting to apply Chym to real world problems, such as the ones mentioned in the Introduction. For example, Chym could be used in the comparison of metabolic networks to detect the metabolic characteristics responsible for the virulence of some *Streptococcus pneumoniae* strains, or it could be used in a drug development project to determine chemical compounds predicted to have a certain impact in the organism of humans.

As a complement to this work, and to further improve Chym as a classification tool, I think that trying other hybrid metrics, especially other structural comparison algorithms, would yield a fascinating project, probably with very good results. For instance, since SVM and random forests seem to perform very well on the sets they were used in, perhaps a Chym-like system, where the structural part of the comparison is replaced with one of those methods will, probably outperform Chym.

Moreover, while not mentioned during the main text of the thesis, the SAR premise has recently been complemented with the QSAR premise (*Quantitative* Structure-Activity Relationship). Instead of a binary classifier, Chym-could have been implemented as a tool that predicts a continuous variable. For instance, it could have been used to predict the concentration needed to reach half of the maximum inhibition of and enzyme. It would make an exceptional study to understand if the activity coefficient, as defined in this work, is in any way correlated to these continuous variables.

# Bibliography


Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. and Nielsen, H.: 2000, Assessing the accuracy of prediction algorithms for classification: an overview, *Bioinformatics* **16**(5), 412.

Chandrasekaran, B., Josephson, J. R. and Benjamins, V.: 1999, What are ontologies, and why do we need them?, *IEEE Intelligent systems* **14**(1), 20–26.

Chen, L., Yang, J., Davey, A. K., Chen, Y., Wang, J. and Liu, X.: 2009, Effects of diammonium glycyrrhizinate on the pharmacokinetics of aconitine in rats and the potential mechanism, *Xenobiotica* **39**(12), 955–963.

Cheng, A., Diller, D. J., Dixon, S., Egan, W., Lauri, G. and Merz Jr, K.: 2001, Computation of the physio-chemical properties and data mining of large molecular collections, *Journal of computational chemistry* **23**(1), 172–183.

Clemente, J. C., Satou, K. and Valiente, G.: 2005, Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and the gene ontology, *GENOME INFORMATICS SERIES* **16**(2), 45.

Cordon-Cardo, C., O'brien, J., Boccia, J., Casals, D., Bertino, J. and Melamed, M.: 1990, Expression of the multidrug resistance gene product (P-glycoprotein) in human normal and tumor tissues, *Journal of Histochemistry and Cytochemistry* **38**(9), 1277.

Cortes, C. and Vapnik, V.: 1995, Support-vector networks, *Machine learning* **20**(3), 273–297.

Crowther, J.: 1995, *Oxford Advanced Learner's Dictionary*, fifth edition edn, Oxford University Press. "Semantic".

Dall'Acqua, S., Tomè, F., Vitalini, S., Agradi, E. and Innocenti, G.: 2009, In vitro estrogenic activity of Asplenium trichomanes L. extracts and isolated compounds, *Journal of Ethnopharmacology* **122**(3), 424–429.

Daylight Chemical Information Systems, Inc.: 2008, Daylight theory manual. Version 4.9.
http://www.daylight.com/dayhtml/doc/theory/index.html

Degtyarenko, K., Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M. and Ashburner, M.: 2007, ChEBI: a database and ontology for chemical entities of biological interest, *Nucleic acids research* .

Doniger, S., Hofmann, T. and Yeh, J.: 2002, Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms, *Journal of computational biology* **9**(6), 849–864.

Drews, J.: 2000, Drug discovery: a historical perspective, *Science* **287**(5460), 1960.

Flower, D. R.: 1998, On the properties of bit string-based measures of chemical similarity, *Journal of Chemical Information and Computer Sciences* **38**(3), 379–386.

Fukunishi, Y., Mikami, Y., Takedomi, K., Yamanouchi, M., Shima, H., Nakamura, H. et al.: 2006, Classification of Chemical Compounds by Protein-Compound Docking for Use in Designing a Focused Library, *J. Med. Chem* **49**(2), 523–533.

Fukunishi, Y. and Nakamura, H.: 2009, A Similarity Search Using Molecular Topological Graphs, *Journal of Biomedicine and Biotechnology* **2009**.

Gentleman, R.: 2005, Visualizing and distances using GO.
http://www.bioconductor.org/docs/papers/2003/Compendium/GOvis.pdf

Ghosh, A. K., Kumaragurubaran, N., Hong, L., Kulkarni, S., Xu, X., Miller, H., Srinivasa Reddy, D., Weerasena, V., Turner, R., Chang, W. et al.: 2008, Potent memapsin 2 ([beta]-secretase) inhibitors: Design, synthesis, protein-ligand X-ray structure, and in vivo evaluation, *Bioorganic & medicinal chemistry letters* **18**(3), 1031–1036.

Grego, T., Ferreira, J. D., Pesquita, C., Bastos, H., Vila Viçosa, D., Freire, J. and Couto, F.: 2010, Chemical and Metabolic Pathway Semantic Similarity, *Technical Report 1*, Faculty of Sciences, University of Lisbon.

Guha, R., Howard, M. T., Hutchison, G., Murray-Rust, P., Rzepa, H., Steinbeck, C., WEGNER, J. and Willighagen, E.: 2006, The Blue Obelisk-interoperability in chemical informatics, *Journal of chemical information and modeling* **46**(3), 991–998.

Gupta, R. C.: 2006, Taurine analogues and taurine transport: Therapeutic advantages, *Advances in experimental medicine and biology* **583**, 449–467.

Herzberg, G.: 2008, *Molecular Spectra and Molecular Structure*, Vol. 1, second edn, Reitell Press. "Introduction".

Heymans, M. and Singh, A. K.: 2003, Deriving phylogenetic trees from the similarity analysis of metabolic pathways, *Bioinformatics* **19**(Suppl 1), i138.

IUPAC Commission on Nomenclature of Organic Chemistry: 1993, *A Guide to IUPAC Nomenclature of Organic Compounds (Recommendations 1993)*, Blackwell Scientific publications.

Jaccard, P.: 1901, Étude comparative de la distribution florale dans une portion des Alpes et des Jura, *Bull. Soc. Vaudoise Sci. Nat* **37**, 547–579.

Joachims, T.: 1999, SVMLight: Support Vector Machine, *Technical report*, University of Dortmund.
http://svmlight.joachims.org/

Kaminka, M.: 1971, Serotonin preparations, *Pharmaceutical Chemistry Journal* **5**(5), 301–304.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M.: 2006, From genomics to chemical genomics: new developments in KEGG, *Nucleic acids research* **34**(Database Issue), D354.

King, R. D., Srinivasan, A. and Dehaspe, L.: 2001, Warmr: a data mining tool for chemical data, *Journal of Computer-Aided Molecular Design* **15**(2), 173–181.

Le, S. Q., Ho, T. and Phan, T.: 2004, A novel graph-based similarity measure for 2D chemical structures, *Genome Infommatics Series* **15**(2), 82.

Lord, P. W., Stevens, R., Brass, A. and Goble, C.: 2003, Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, *Bioinformatics* **19**(10), 1275.

Markiewicz, L., Garey, J., Adlercreutz, H. and Gurpide, E.: 1993, In vitro bioassays of non-steroidal phytoestrogens, *Journal of Steroid Biochemistry and Molecular Biology* **45**(5), 399–405.

Martin, Y. C., Kofron, J. and Traphagen, L.: 2002, Do structurally similar molecules have similar biological activity?, *J. Med. Chem* **45**(19), 4350–4358.

Miksicek, R. J.: 1993, Commonly occurring plant flavonoids have estrogenic activity, *Molecular Pharmacology* **44**(1), 37.

Mitchell, F.: 1978, Automation in clinical chemistry: developments and recent trends, *The Journal of Automatic Chemistry* **1**(1), 7.

Muster, W., Breidenbach, A., Fischer, H., Kirchner, S., Müller, L. and Pähler, A.: 2008, Computational toxicology in drug development, *Drug discovery today* **13**(7-8), 303–310.

Ohtsuki, S. and Terasaki, T.: 2007, Contribution of carrier-mediated transport systems to the blood–brain barrier as a supporting and protecting interface for the brain; importance for CNS drug discovery and development, *Pharmaceutical research* **24**(9), 1745–1758.

Open Babel Project: 2009, The Open Babel Package. Version 2.2.3, downloaded in December 2009.
http://openbabel.sourceforge.net/

Patani, G. and LaVoie, E.: 1996, Bioisosterism: a rational approach in drug design, *Chemical reviews* **96**(8), 3147–3176.

Penzotti, J. E., Lamb, M., Evensen, E. and Grootenhuis, P.: 2002, A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein, *J. Med. Chem* **45**(9), 1737–1740.

Pesquita, C., Faria, D., Bastos, H., Ferreira, A., Falcao, A. and Couto, F.: 2008, Metrics for GO based protein semantic similarity: a systematic evaluation, *BMC bioinformatics* **9**(Suppl 5), S4.

Pinter, R. Y., Rokhlenko, O., Yeger-Lotem, E. and Ziv-Ukelson, M.: 2005, Alignment of metabolic pathways, *Bioinformatics* **21**(16), 3401.

Pratt, R.: 1989, Inhibition of a class C beta-lactamase by a specific phosphonate monoester, *Science* **246**(4932), 917.

Raymond, J. W., Gardiner, E. and Willett, P.: 2002, Rascal: Calculation of graph similarity using maximum common edge subgraphs, *The Computer Journal* **45**(6), 631.

Raymond, J. W. and Willett, P.: 2002, Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases, *Journal of computer-aided molecular design* **16**(1), 59–71.

Reading, C. and Cole, M.: 1977, Clavulanic acid: a beta-lactamase-inhibiting beta-lactam from Streptomyces clavuligerus, *Antimicrobial Agents and Chemotherapy* **11**(5), 852.

Resnik, P.: 1999, Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, *Journal of Artificial Intelligence* **11**(11), 95–130.

Richard, A. M., Gold, L. and Nicklaus, M.: 2006, Chemical structure indexing of toxicity data on the internet: moving toward a flat world, *Current Opinion in Drug Discovery and Development* **9**(3), 314.

Ross, S.: 1994, A first course in probability, *New York* .

Salim, N., Holliday, J. and Willett, P.: 2003, Combination of fingerprint-based similarity coefficients using data fusion, *J. Chem. Inf. Comput. Sci* **43**(2), 435–442.

Schinkel, A., Wagenaar, E., Van Deemter, L., Mol, C. and Borst, P.: 1995, Absence of the mdr1a P-Glycoprotein in mice affects tissue distribution and pharmacokinetics of dexamethasone, digoxin, and cyclosporin A., *Journal of Clinical Investigation* **96**(4), 1698.

Shlomi, T., Segal, D., Ruppin, E. and Sharan, R.: 2006, QPath: a method for querying pathways in a protein-protein interaction network, *BMC bioinformatics* **7**(1), 199.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A. and Rosse, C.: 2005, Relations in biomedical ontologies, *Genome Biology* **6**(5), R46.

So, F. V., Guthrie, N., Chambers, A. and Carroll, K.: 1997, Inhibition of proliferation of estrogen receptor-positive MCF-7 human breast cancer cells by flavonoids in the presence and absence of excess estrogen, *Cancer letters* **112**(2), 127–133.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. and Feuston, B.: 2003, Random forest: a classification and regression tool for compound classification and QSAR modeling, *J. Chem. Inf. Comput. Sci* **43**(6), 1947–1958.

Teixeira, A. L., Santos, R. and Couto, F.: 2009, ThermInfo: Collecting and Presenting Thermochemical Properties, *Technical report*, Faculty of Sciences, University of Lisbon.
http://xldb.di.fc.ul.pt/xldb/publications/Teixeira.etal:ThermInfoCollectingAnd:2009_document.pdf

Todd, P. and Benfield, P.: 1990, Amoxicillin/clavulanic acid: an update of its antibacterial activity, pharmacokinetic properties and therapeutic use, *Drugs(Basel)* **39**(2), 264–307.

Tohsato, Y. and Nishimura, Y.: 2008, Metabolic pathway alignment based on similarity between chemical structures, *Information and Media Technologies* **3**(1), 191–200.

Tong, W., Hong, H., Fang, H., Xie, Q. and Perkins, R.: 2003, Decision forest: combining the predictions of multiple independent decision tree models, *J. Chem. Inf. Comput. Sci* **43**(2), 525–531.

Unger, C., Eibl, H., Heyden, H. W., Kim, D. and Nagel, G.: 1986, Aminoglutethimide, *Investigational New Drugs* **4**(3), 237–240.

Willett, P., Barnard, J. M. and Downs, G.: 1998, Chemical similarity searching, *J. Chem. Inf. Comput. Sci* **38**(6), 983–996.

Willett, P., Winterman, V. and Bawden, D.: 1986, Implementation of nearest-neighbor searching in an online chemical structure search system, *Journal of Chemical Information & Computer Sciences* **26**(1), 41.

Wolber, G., Seidel, T., Bendix, F. and Langer, T.: 2008, Molecule-pharmacophore superpositioning and pattern matching in computational drug design, *Drug discovery today* **13**(1-2), 23–29.

Wolosker, H., Dumin, E., Balan, L. and Foltyn, V. N.: 2008, d-Amino acids in the brain: d-serine in neurotransmission and neurodegeneration, *FEBS Journal* **275**(14), 3514–3526.

Xue, Y., Yap, C., Sun, L., Cao, Z., Wang, J. and Chen, Y.: 2004, Prediction of P-glycoprotein substrates by a support vector machine approach, *J. Chem. Inf. Comput. Sci* **44**(4), 1497–1505.

Yılmaz, B. and Göktürk, M.: 2009, Interactive Data Mining for Molecular Graphs, *Journal of Automated Methods and Management in Chemistry* **2009**.

Zand, R. S., Jenkins, D. and Diamandis, E.: 2000, Steroid hormone activity of flavonoids and related compounds, *Breast cancer research and treatment* **62**(1), 35–49.

# Appendix A

# The construction of Chym

As happens with every tool, particularly the ones developed with a scientific purpose in mind, the creation of Chym was not a linear process; on the contrary, it was actually filled with experimentation and try-error cycles. This appendix presents the road that I had to travel when developing and validating Chym, and documents the reasons that were in the origin of some of the decision made.

## A.1 The possible choices for Chym

The first "problem" encountered was the extraction of ChEBI molecular structures from the ChEBI database. As stated above, ChEBI has SMILES, MDL and InChI formats, but not all chemical entities have the three structures. The preference of SMILES over MDL over InChI was set based on a rough estimation of usage. This estimate was done through Google, as detailed in **Table A.1**.

InChI was actually discarded, because all chemical compounds with a structure in this format had a structure in one of the other formats. Since the structure is only used to compute the fingerprints, and the three fingerprint formats in the base of Chym's structural similarity use only two dimensional properties, I expect that either one of the remaining formats will return very similar (if not equal) fingerprints. For instance, the compound *octane-1,8-diol*, ChEBI:44630, exposes a SMILES structure and an MDL structure. The fingerprints obtained for each structure are equal when using either one of the three fingerprint formats. With this in mind, SMILES was preferred because

**Table A.1: Choice of SMILES over MDL over InChI.** This table details the approximate number of hits returned by the Google search engine, when the query is "⟨*term*⟩ chemical structure" (without the quotes). These numbers refer to a search made on April 9$^{\text{th}}$, 2010.

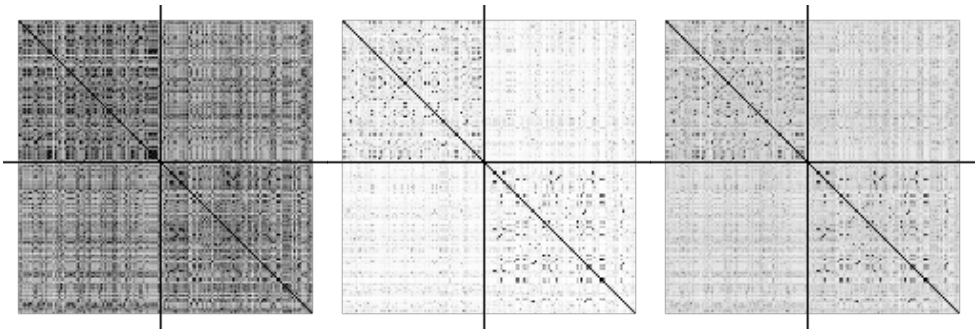
| ⟨*term*⟩ | **#results** |
|---|---|
| SMILES | 4,670,000 |
| MDL | 1,050,000 |
| InChI | 158,000 |

it is a simpler language than MDL, and takes much less space to store.

As for the classification algorithm chosen, there are mainly two decisions that had to be made to ascertain Chym's performance as a classification tool. The most obvious parameter to be tuned is the criteria used to decide whether a compound should be classified as positive or negative. As stated in the algorithm presented in section 4.4, Chym classifies a compound based on its activity threshold, which is the mean of the similarity values of the compound with the positive compounds in the input set. Prior to the establishment of this principle, I tried other approaches.

Before discussing the first attempts, I think the understanding of this section is facilitated by the visual representations on **Figure A.1**, which exhibits the similarity matrices obtained for the BBB classification problem, with Chym parameters detailed in the sub-captions. Three conclusions can be retrieved from that figure. First, there is a visible difference in the average shade of each of the quadrants. Second, each format leads to different averages of similarity (as supposed). Third, and more important for the rest of the assessment, the mean similarity of the negative vs. negative compounds is lower than the positive vs. positive. This reflects the wider range of chemical entities present in the negative group, which is expected, since the negative compounds are a heterogeneous group. **Table A.2** is a numeric overview of those matrices.

We now go back to the attempts made before the final establishment of Chym's classification algorithm. The first attempt was the use of a spectral cluster technique. This method gets, as input, the matrix of similarities between all compounds and tries to distribute the compounds among $k$ non-



**(a)** Purely structural similarity, FP3 format

**(b)** Purely semantic similarity, simGIC method, ontology 'all'

**(c)** FP3, simGIC, all, $\alpha = 0.29$

**Figure A.1: Similarity matrices for several Chym details.** All the matrices refer to the BBB classification problem. These images represent the similarity between all pairs of compounds in the $BBB_p$ set. Starting from the top left quadrant, and going clockwise, the quadrants contain: the similarities between positive compounds; between positive and negative; between negative; between negative and positive. The horizontal and vertical dark lines are not part of the matrix, but instead stand as the separation between positive compounds and negative compounds. Darker shades of gray reflect higher similarity values, and the top-left-to-bottom-right diagonal is completely black because $\text{sim}(c, c) = 1$. Since the matrices are symmetric, the figures are also symmetric, with the symmetry axis being the black diagonal.

**Table A.2: Details of some similarity matrices of the BBB problem.** In this table, P means positive compounds, N means negative and the values are the average of all the similarities between all the compounds in the two groups. The values on each row can be seen as the mean shade of gray in the top left, top right and bottom left quadrants respectively.

| **Chym parameters** | **Average of similarities** | | |
|---|---|---|---|
| | P × P | P × N | N × N |
| FP3, $\alpha = 1.00$ | 0.524 | 0.423 | 0.468 |
| simGIC, all, $\alpha = 0.00$ | 0.121 | 0.046 | 0.079 |
| FP3, simGIC, all, $\alpha = 0.29$ | 0.238 | 0.155 | 0.192 |

overlapping clusters of compounds such that intra-similarity inside a cluster is high and inter-similarity between compounds in different clusters is low. In all attempts, I used $k = 2$, since there are two classes of compounds: positive and negative. The clusters techniques tended to fail, because there is not enough power to separate the positive compounds from the negative. I think this happens because, while the values in **Table A.2** show evidence of greater similarity among positive compounds, each line in the image (corresponding to the different similarities between a compound and and every other compound) contains a series of dark and light pixels, and many of the dark pixels are actually in the "wrong" quadrant (wrong here means in the right quadrant for lines of the top, or left quadrant for lines on the bottom). The same happens with light pixels. This leads to a general indecision as to whether to classify the compound as positive or negative. Actually, after applying spectral cluster to those matrices, the algorithm usually puts a single compound in one of the clusters and the remaining compounds in the other cluster, which is a totally inappropriate result for Chym. Besides, this classification system is not efficient, because it is not able to classify a new compound in a straightforward manner.

From this, it was clear that the path to follow did not made use of clustering techniques. The second approach attempted was, according to these previous finding, the approach that ended up being implemented in the actual definition of Chym: a compound is classified as positive or negative based on whether an *activity coefficient*, which is calculated through its similarity with the compounds in the primary set, is above or below a certain threshold. Since **Table A.2** uses average values, I tried an average approach, where the activity coefficient is the average of the similarities between a compound and the known positive compounds. This is the current definition, because it was the approach which achieved higher performance, and leads to the results shown below. This is called the Average classification algorithm. In case this was still not the best algorithm, I continued the analysis.

The third attempt was similar to the second one, but instead of taking the average of the similarities between a compounds and all the positive compounds, I took the maximum of these similarities. This is the Maximum classification algorithm.

The forth and final attempt is the most sophisticated. A compound is classified as positive if among the $m$ most similar compounds there are at

least $n$ positive compounds. The pair $(n, m)$ may have several values, like $(8, 10)$ or $(10, 20)$. This approach is dubbed Fraction $n/m$.

Finally, since Chym uses the model where a compound is classified as positive if its activity coefficient is higher than a threshold, there is still a parameter from the classification algorithm that must be decided. Section 4.4 refers a training step which takes the known positive and negative compounds and determines the best threshold as the value that minimizes the number of misclassification in the training group, thus maximizing the accuracy. This process could be achieved not only through accuracy, but also through the optimization of other descriptors, like the the Matthews correlation coefficient or the F-measure (see section 4.5). After using either one of those three, the chosen threshold was approximately the same, which means that the parameter to be optimized does not influence significantly the performance of Chym. Thus, Chym uses accuracy, since it is simpler, quicker to calculate, and more intuitive.

## A.2 Choosing the correct options

**Table A.3** shows what happens to the accuracy of Chym when some of the alternative approaches mentioned in the previous section are used. The table contains the results obtained by Chym when applied to the BBB problem. In this table, the Chym parameters used are those that lead to the best accuracy (not necessarily the best MCC, though). These results show that any of the attempts produce good results, with the best one being the Average classification algorithm.

It is also worth mentioning here that no matter what classification algorithm is used (cf. **Table A.3**), the best parameters of Chym do not change much. The greatest deviation is that the Average and Fraction 12/20 algorithms perform better in the whole ontology, while the other perform better in the "role" branch of the ontology. As for the $\alpha$ values, there is a slightly wider variation, with the Fraction 8/10 and Fraction 10/20 algorithms deviating from the values of the other attempts, which are all approximately 0.30. I suspect that this happens because the Fraction $n/m$ algorithm is very different from the other two, and this may introduce an unpredictable variation. On the other hand, the Fraction 12/20 algorithm also uses the whole ontology and $\alpha = 0.30$, which is almost equal to the Average algorithm. Lastly, it is mostly

**Table A.3: Chym with different classification and training algorithms.** When a parameter is given in parenthesis, it is the parameter that was maximized during the training step of the classification. As such, it is not applicable to to the algorithm Fraction $n/m$. All validation was done with a leave one out approach.

| Classification algorithm | Chym Parameters | Accuracy | MCC |
|---|---|---|---|
| Average (MCC) | FP3, simGIC, all, $\alpha = 0.29$ | 90.9% | 0.821 |
| Average (accuracy) | FP3, simGIC, all, $\alpha = 0.28$ | 90.9% | 0.821 |
| Average (F-measure) | FP3, simGIC, all, $\alpha = 0.28$ | 90.9% | 0.821 |
| Maximum (MCC) | FP3, simGIC, role, $\alpha = 0.34$ | 86.9% | 0.744 |
| Maximum (accuracy) | FP3, simGIC, role, $\alpha = 0.32$ | 86.9% | 0.741 |
| Maximum (F-measure) | FP3, simGIC, role, $\alpha = 0.30$ | 86.9% | 0.741 |
| Fraction 8/10 | FP3, simGIC, role, $\alpha = 0.15$ | 88.2% | 0.766 |
| Fraction 12/20 | FP3, simGIC, all, $\alpha = 0.30$ | 87.6% | 0.752 |
| Fraction 10/20 | FP3, simGIC, role, $\alpha = 0.17$ | 84.3% | 0.700 |

evident that the maximizing parameter chosen has minimal influence in the outcome of Chym, which, as stated previously, is not surprising, because the thresholds that maximize accuracy, Matthews correlation coefficient or the F-measure are bound to be close to each other. All together, these results are a strong evidence of a kind of stability that enriches Chym with a sense of credibility in its prediction power.

# Appendix B

# Technical Details

Chym is a tool built mainly on top of the ChEBI database, since the semantic similarity is intricately related to the ontology produced by that database. Since this is the novelty presented with this work, I provide in this appendix an insight in some technical details that allowed the development of the semantic similarity.

Chym's internal database has a lot of data directly imported from ChEBI, although some of the data used comes from KEGG, or even from the OpenBabel software. Here I show the twelve steps needed to create (not to validate, which has been detailed in the main text of the thesis) a stable database from which the structural and semantic similarities of Chym can be extracted:

1. Extract ChEBI's database dump from the ChEBI website, http://www.ebi.ac.uk/chebi/. This was done on January 2010, release number 64. This database is freely accessible and available to anyone, since it contains only non-proprietary data.

2. Create empty tables in Chym's database, where the information extracted from the data sources or computed from this information will be stored.

3. Populate the database with ChEBI compounds. To do this, Chym distinguishes between primary compounds and secondary compounds. Some ChEBI compounds refer to the same chemical entity and as soon as this was discovered, they were merged into a single compound (named primary), but for compatibility reasons, the previous identifier remained in the database as a pointer to the primary compound. When a reference

to a compound is ever made in Chym, it first checks whether this is a secondary compound and, if so, converts it to the corresponding primary identifier.

4. The next step was to extract the ChEBI ontology relations into one of Chym's tables. As documented in the main text, only non-cyclic relations were considered. The relations in the original table are only direct relations, i.e., if the relations $c_1 \rightarrow c_2$ and $c_2 \rightarrow c_3$ exist in the ontology (where the right side of the arrow represent the ancestor of the relation), they are present in the original table and are extracted to Chym's table; but the fact that $c_1$ and $c_3$ are connected through a path of relations is not directly inserted in this table.

5. Now Chym computes the ancestry of all compounds. To do this, I used a transitive closure algorithm on the universe of all relations. This means that, using the previous example, Chym now constructs a table with a row stating $c_1 \xrightarrow{2} c_3$, which means that $c_3$ is an ancestor of $c_1$ and the distance between them, in number of relations, is 2. Since Chym does not differentiate between non-cyclic relations, it is not relevant if the two relations are of the same type or not. Notice that Chym uses several ontologies, with one transitive closure table for each ontology.

6. From KEGG PATHWAY, Chym then extracts the metabolic pathways that are used as the basis of the IC calculations. This information was retrieved directly from a webpage, http://www.genome.jp/kegg-bin/get_htext?htext=br08901.keg&filedir=%2ffiles&hier=2, which was parsed to get all the pathways of the data source. KEGG was also accessed on January 2010, which corresponds to version 53.0.

7. The KEGG web-services (http://www.genome.jp/kegg/soap/) were further used to get the compounds of each pathway, as KEGG COMPOUND entries.

8. ChEBI's cross references are extracted to Chym's internal database.

9. Using ChEBI's cross-references, Chym then converts each KEGG COMPOUND entry into a ChEBI identifier, thus calculating the number of distinct pathways in which each ChEBI primary compound participates.

As stated in the main text, this step is keen to lead to some ambiguity. In the case of one KEGG COMPOUND identifier being mapped into two ChEBI compounds, both the compounds inherit the KEGG COMPOUND entry's pathways; when one ChEBI compound references two KEGG COMPOUND entries, it inherits both their pathways.

10. Finally, the information content of each compound is calculated. Chym uses four ontologies, and as such it has to calculate four information content values for each compound. To do this, Chym refers to the corresponding closure graph, determines the descendants of the compound and the number of distinct pathways in which any one of these compounds participates. When divided by the total number of pathways, this is the value of $\Pr(x)$ from equation 4.4, and the value is stored for all compounds in each one of the four ontologies. With these steps, semantic similarity is a matter of selecting the appropriate rows of the tables and combining them through equation 4.2 and equation 4.3.

11. For the structural metrics, ChEBI structures where extracted from the original database and inserted into Chym's internal database. There are various formats, as mentioned earlier, but Chym prefers SMILES to MDL. Only one structure per primary identifier was considered.

12. To calculate the structural similarity, OpenBabel was then used to compute the three structural fingerprints of each compound. With this, Chym is completed, and is now able to calculate the hybrid similarity metric for any two compounds which can be mapped to ChEBI and have a structure.

The information calculated in these steps is stored locally in a MySQL database. All the steps were coded with Python 2.4, and the files are stored in one of the Informatics Department internal servers.

# Appendix C

# Mathematical proof

In section 4.2.2, I mentioned that simGIC *usually* has a smaller value than simUI:

$$\text{simUI} > \text{simGIC} \tag{C.1}$$

Consider two compounds $c_1$ and $c_2$ and the sets of their ancestry, $\text{asc}(c_1)$ and $\text{asc}(c_2)$. Let $a_1, \ldots, a_n$ be the information content of the compounds that belong to *both* ancestries and $b_1, \ldots, b_m$ the information content of the compounds that belong to either $\text{asc}(c_1)$ or $\text{asc}(c_2)$ but *not to both*. $n$ is the size of the intersection set and $n + m$ is the size of the union set. Remember that $\text{asc}(c)$ contains $c$ itself.

This new notation allows the redefinition of the semantic measures:

$$\text{simUI} = \frac{n}{n+m} \tag{C.2}$$

$$\text{simGIC} = \frac{a_1 + \cdots + a_n}{a_1 + \cdots + a_n + b_1 + \cdots + b_m} \tag{C.3}$$

and with this, inequality C.1 can be expressed in terms of these new variables:

$$\begin{aligned} & \frac{n}{n+m} > \frac{a_1 + \cdots + a_n}{a_1 + \cdots + a_n + b_1 + \cdots + b_m} \\ \Leftrightarrow \quad & \frac{a_1 + \cdots + a_n + b_1 + \cdots + b_m}{n+m} > \frac{a_1 + \cdots + a_n}{n} \\ \Leftrightarrow \quad & \text{avg}(a_i, b_i) > \text{avg}(a_i) \end{aligned} \tag{C.4}$$

It is a fact that a term's IC is never lower than the value of each of its ancestors' IC. Moreover, while not always true, it is at least reasonable to expect that the IC of the compounds in the intersection of the ancestries is smaller then the IC of the compounds not in the intersection. This is because

the compounds that are in both ancestries are less specific than the others (cf. **Figure 4.1**, where none of the common ancestors of $A$ and $B$ has an IC value higher than the not-common ancestors). Because of that, $b_i > a_j$, for most values of $i$ and $j$. This means that the average of all IC values, $\text{avg}(a_i, b_i)$, is *generally* higher than the average of the less specific compounds, $\text{avg}(a_i)$, which proves equation C.4 and, thus, the initial inequality.