Similarity-based Virtual Screening: Effect of the Choice of Similarity Measure

A study submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

at

The University of Sheffield

by

Hua Xiang

Supervisor: Prof. Peter Willett, PhD

Co-Supervisor: John Holliday, PhD

Information School

October 2013

Acknowledgments

Foremost, I would like to thank my supervisors Prof. Peter Willett and Dr. John Holliday for their support and guidance throughout my study. I am very grateful for their patience, motivation, enthusiasm, and immense knowledge in Chemoinformatics that, taken together, make them great mentors.

I want to thank past and present members of Computational Information Systems Research Group during my time in the Michael Lynch Research Lab: Prof. Val Gillet, Dr. Eleanor Gardiner, Iain Mott, Richard Martin, Chia-Wei Chu, Georgios Papadatos, Christoph Mueller, Shereena Arif, Nurul Malim, Ben Allen, Richard Sherhod, Sonny Gan, Jorge Valencia, James Wallace, Nor Sani, Simon Hand, Edmund Duesbury and Matthew Seddon for their intellectual discussions and interesting group meetings. In particular, massive thanks to Chia-Wei Chu and Nurul Malim for their encouragement and moral support all the time; many thanks to Shereena Arif, Georgios Papadatos, Christoph Mueller and Jorge Valencia for their help and resources sharing.

I am deeply thankful to my family and friends outside of the CISRG group for their love, support, and sacrifices. Without them, this thesis would never have been written.

Finally, I am grateful to The University of Sheffield for giving me the opportunity and funding.

Abstract

The aim of the research was to identify novel similarity measures for similarity-based virtual screening. Similarity-based virtual screening is at the lead identification stage of drug discovery process and normally requires explorations in large scale databases. Thus, the improvement of accuracy of the methods employed could result in a significant enhancement of effectiveness of the whole process of drug discovery. There are three key components involved in similarity-based virtual screening, i.e., structural representations, similarity coefficients and weighting schemes. The research focuses on the choice of similarity coefficient and weighting scheme.

Three investigations have been conducted: investigation of interactions between weighting schemes and similarity coefficients; comparison of binary coefficients and evaluation of similarity coefficients using weighted fingerprints. Four chemical databases were used, i.e., MDDR, WOMBAT, MUV and ChEMBL. The results show that there are strong, and often quite subtle, interactions between the similarity coefficient and the weighting scheme comprising a similarity measure. They also exhibit that, although the Tanimoto coefficient remains one of the most practical coefficients for use in similarity-based virtual screening on binary representations, it may not be the coefficient of choice when weighting schemes are applied. In addition, other coefficients were identified as favorable for similarity-based virtual screening when weighted fingerprints are available. The findings indicate that the study of the combinations of weighting schemes and similarity coefficients could make a significant contribution to similarity-based virtual screening.

Table of Contents

Acknowledgments ..... *i*

Abstract ..... *iii*

Table of Contents ..... *v*

List of Figures ..... *ix*

Index of Tables ..... *xi*

Chapter 1: Introduction *1*

Chapter 2: Similarity-Based Virtual Screening *5*

2.1 Introduction **5**

2.2 Representations of Molecular Structures **10**

2.2.1 Representations of 2D Molecular structures 11

2.2.1.1 Linear Notations 11

2.2.1.2 Connection Table 12

2.2.1.3 Reduced Graphs 14

2.2.1.4 Fingerprints 15

2.2.2 3D Molecular Representations 19

2.3 Weighting Scheme **22**

2.4 Similarity Coefficients **24**

2.4.1 Distance Coefficients 25

2.4.2 Association Coefficients 26

2.4.3 Correlation Coefficients 27

2.4.4 Probabilistic Coefficients 28

2.4.5 Choice of Coefficient 28

2.5 Data Fusion **29**

2.6 Conclusion **34**

Chapter 3: Methodology ***35***

3.1 Introduction **35**

3.2 Data **35**

3.2.1 Chemical Databases 35

3.2.1.1 MDDR 36

3.2.1.2 WOMBAT 37

3.2.1.3 MUV 38

3.2.1.4 ChEMBL 39

3.2.1.5 Comparison of Databases' Diversity 44

3.2.2 Molecular Representation 45

3.3 Procedure of Similarity Search **48**

3.4 Evaluation Method **50**

3.4.1 The Kendall W Test of Concordance 52

3.4.2 The Wilcoxon Signed-rank Test 53

3.5 Clustering Method **54**

3.6 Summary **57**

Chapter 4: Evaluation of Interactions between Weighting Scheme and Similarity Coefficient in Similarity-based Virtual Screening ***58***

4.1 Introduction **58**

4.2 Method **59**

4.3 Results **65**

4.3.1 MDDR Results 66

4.3.2 WOMBAT Results 71

4.4 Discussion **74**

4.4.1 Comparison of Coefficients 74

4.4.1.1 Symmetric or Asymmetric Weighting 76

4.4.1.2 W3 Involved or Non-W3 Involved Weighting 83

4.4.1.3 Conclusion 85

4.4.2 Comparison of Weighting Schemes 90

4.5 Further Analysis and Evaluation 93

4.5.1 Effect of Fingerprint's Density 93

4.5.2 Results and Analysis 97

4.6 Conclusion 100

Chapter 5: Comparison of Established Level of Binary Coefficients for Chemical Similarity Search 102

5.1 Introduction 102

5.2 Coefficients for Binary Variables 103

5.3 Classification and Rescaling of Coefficients 105

5.3.1 Classes of Coefficient 105

5.3.2 Rescaling of Coefficients 106

5.4 Method 107

5.5 Results of Initial Investigation 112

5.5.1 Ranks of Retrieval Abilities 112

5.5.2 Comparison of Retrieval Rate of Active Compounds 119

5.5.2.1 Retrieval Rate of Active Compounds 119

5.5.2.2 Grouping Coefficients using Retrieval Rate of Active Compounds 119

5.5.2.3 Comparison of Coefficients based on Nature of Classes 125

5.5.3 Conclusion 128

5.6 Validation Experiments 128

5.7 Conclusion 139

Chapter 6: Comparison of Similarity Coefficients using Weighted Chemical Data 140

6.1 Introduction 140

6.2 Selection of Coefficients 140

6.2.1 Identification of Coefficients 143

6.2.2 More Coefficients 146

6.3 Method 148

6.4 Results and Discussion 149

6.5 Conclusion..... 163

Chapter 7: Conclusion.....166

Bibliography: 170

Appendix A: Results of Chapter 4 180

Appendix B: Results of Chapter 5 186

Appendix C: Results of Chapter 6 196

List of Figures

Figure 1.1 The drug design process *1*
Figure 2.1 Typical virtual screening processes *7*
Figure 2.2 Structure, name, InChI, SMILES and connection table for aspirin (from (Holliday and Willett, 2011)). *14*
Figure 2.3 Two examples of different types of reduced graphs. *15*
Figure 2.4 An example of how a dictionary-based fingerprint is generated (Based on Digital Chemistry, 2011). *17*
Figure 2.5 An example of how a hashed-based fingerprint is generated. *18*
Figure 2.6 A 3D model of the Aspirin structure (http:// www.3dchem.com). *19*
Figure 2.7 3D coordinates and conformation of pharmacophore points. *21*
Figure 2.8 A typical fusion process (Leach and Gillet, 2007). *31*
Figure 2.9 Schematic outline of a turbo similarity search (Gardiner et al., 2009). *33*
Figure 3.1 A molecule from the ChEMBL database (left) and its corresponding Murcko scaffold (right). *37*
Figure 3.2 Comparison of MPS values of four databases using boxplot *45*
Figure 3.3 Schematic representation for generating extended connectivity descriptors (Rogers and Hahn, 2010). *47*
Figure 3.4 Procedure of Similarity Search. *50*
Figure 4.1 Median numbers of actives retrieved in searches of the MDDR database using the Tanimoto, cosine and MinMax similarity coefficients. *88*
Figure 4.2 Median numbers of actives retrieved in searches of the WOMBAT database using the Tanimoto, cosine and MinMax similarity coefficients. *89*
Figure 4.3 Median value of active molecules retrieved by 25 measures (average over MDDR and WOMBAT databases and three coefficients) *91*
Figure 4.4 Comparison of activity classes. (a) on MDDR, (b) on WOMBAT. *95*

Figure 4.5 Comparison of three coefficients on MUV database (averaged over 25 similarity measures).....99
Figure 4.6 Comparison of three coefficients on MUV database (averaged over 17 active datasets).....99
Figure 5.1 Heatmap of the ranks of coefficients in MDDR......117
Figure 5.2 Heatmap of the ranks of coefficients in WOMBAT......118
Figure 5.3 Heatmap of the retrieval rates of coefficients in MDDR......122
Figure 5.4 Heatmap of the retrieval rates of coefficients in WOMBAT......123
Figure 5.5 Comparison of similarity coefficients from Symmetric measures on MDDR......124
Figure 5.6 Comparison of similarity coefficients from Symmetric measures on WOMBAT......124
Figure 5.7 Heatmap of the ranks of coefficients in ChEMBL.....130
Figure 5.8 Heatmap of the retrieval rates of coefficients in ChEMBL......131
Figure 5.9 Correlations of symmetric coefficients in 50 activity classes of ChEMBL. 132
Figure 6.1 Comparison of the top 1% retrieval rates of active compounds in MDDR. (a) W4 weighted, (b) W5 weighted......153
Figure 6.2 Comparison of the top 1% retrieval rates of active compounds in WOMBAT. (a) W4 weighted, (b) W5 weighted......154
Figure 6.3 Comparison of the top 1% retrieval rates of active compounds in ChEMBL. (a) W4 weighted, (b) W5 weighted......155
Figure 6.4 Heatmaps of the top 1% retrieval rates of active compounds in MDDR. Upper, W4 weighted; Lower, W5 weighted......157
Figure 6.5 Heatmaps of the top 1% retrieval rates of active compounds in WOMBAT. Upper, W4 weighted; Lower, W5 weighted......158
Figure 6.6 Heatmaps of the top 1% retrieval rates of active compounds in ChEMBL. (a), W4 weighted; (b), W5 weighted......160
Figure 6.7 Mean retrieval rates of the 13 coefficients.....161

Index of Tables

Table 2.1 Typical coefficients commonly used in similarity search..... 30
Table 3.1 MDDR 11 selected activity classes 41
Table 3.2 WOMBAT 14 selected activity classes 41
Table 3.3 MUV 17 activity classes 42
Table 3.4 ChEMBL 50 activity classes 43
Table 4.1 Statistical data describing the MDDR, WOMBAT and MUV datasets using ECFC_4 fingerprints..... 62
Table 4.2 Average numbers of actives molecules retrieved in the top 1% of searches of the MDDR database using the Tanimoto coefficient. 68
Table 4.3 Average numbers of active molecules retrieved in the top 1% of searches of the MDDR database using the cosine coefficient. 69
Table 4.4 Average numbers of active molecules retrieved in the top 1% of searches of the MDDR database using the MinMax coefficient. 70
Table 4.5 Rankings of the 25 measures for combinations of database and similarity coefficient. 72
Table 4.6 Screening effectiveness of 25 combined weighting schemes in similarity searches of the MDDR and WOMBAT databases using three similarity coefficients. ... 74
Table 4.7 The Wilcoxon signed-ranks test analysis for all results on MDDR and WOMBAT. 76
Table 4.8 Mean actives (a) and Median actives (b) results of using symmetric (both reference molecule and database structure are weighted by using the same weighting scheme) similarity measures. 77
Table 4.9 Mean actives (a) and Median actives (b) results of using asymmetric (reference molecule and database structure are weighted by using different weighting schemes) similarity measures..... 81

Table 4.10 Upper-bound values for the self-similarity of two single reference structures from MDDR and WOMBAT databases, using the Tanimoto, cosine and MinMax coefficients. 82
Table 4.11 Mean actives (a) and Median actives (b) results of using W3 involved similarity measures. 83
Table 4.12 Mean actives (a) and Median actives (b) results of using non-W3 involved similarity measures. 84
Table 4.13 The Wilcoxon signed-ranks test analysis for the comparison of pairs of similarity coefficients: (a) five symmetric schemes; (b) 20 asymmetric weighting schemes; (c) W3 involved weighting schemes; (d) Non-W3 involved weighting schemes. Significant p values (p<=0.05) are bolded.....87
Table 4.14 Rankings of 25 combined weighting schemes in similarity searches of the MDDR and WOMBAT using three similarity coefficients. The best result in each column is shaded.....92
Table 4.15 ECFC_4 fingerprints' analysis for each activity classes in MDDR and WOMBAT, respectively.....96
Table 4.16 Statistical p values for the comparison of pairs of similarity coefficients in the Wilcoxon signed-ranks test on MUV: 97
Table 5.1.Contingency table of quantities 104
Table 5.2. List of the binary similarity coefficients. 108
Table 5.3 Kendall's test of concordance results. 112
Table 5.4 Rank positions of each of the 44 coefficients when averaged over all of the activity classes for each of the two databases. 115
Table 5.5 Comparison of coefficients on different activity classes based on their top 1% retrieval rates.....127
Table 5.6 Comparison of coefficients on different activity classes based on their top 1% retrieval rates in ChEMBL.....133
Table 5.7 Rank positions of the 44 coefficients when averaged over all of the activity classes in the ChEMBL dataset.....134
Table 5.8 Coefficients analysis of the retrieval rates in 50 activity classes of ChEMBL. 137

Table 5.9 The p-values of Wilcoxon signed rank test. All the p-values that turned out to be less than the 0.05 significance level were bolded and italic. 138
Table 6.1. Coefficients ranked in top 50% among databases using binary fingerprints. 143
Table 6.2. Ten Non-binary similarity coefficients..... 145
Table 6.3 Additional Non-binary similarity coefficients..... 147
Table 6.4 The Kendall W test results for the combinations of weighting schemes and databases..... 149
Table 6.5 Rank positions of each of the 13 coefficients when averaged over all of the activity classes for each of the combination with databases and weighting schemes. . 152
Table 6.6 Mean retrieval rates of four high-achieving coefficient(group)s. 163

Chapter 1: Introduction

Launching a new drug requires considerable labor, time and expense. Twelve to fifteen years is generally recognized as the time required for developing a new drug from initial inception to final production. On average, it can cost approximately US$750 million and the tests up of a million compounds to identify a new molecule (Atkinson and Jones, 2009). As a global industry, drug discovery is a crucial area of research.

The modern drug design process can be generally described as five main steps as follows:

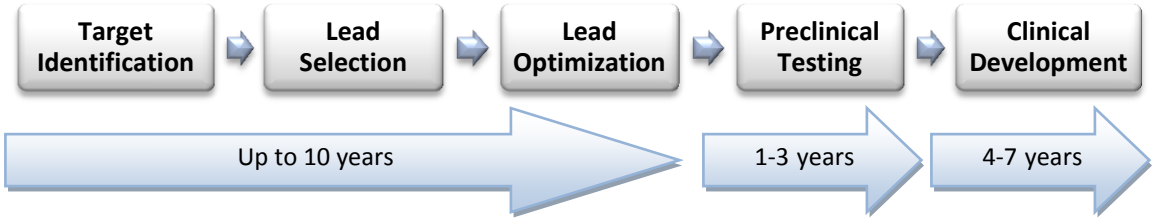


Figure 1.1 The drug design process

In the initial stages, a target (e.g., a protein) is identified as the focus of a certain disease from genetic information. The next two steps focus on the identification of new lead structures (i.e., drug candidates) that can block or activate the target, and the optimization of these structures in order to increase the biological activity and ADMET. The fourth step usually involves *in vitro* and *in vivo* tests which include toxicity assessments and animal tests. The final step focuses on human testing.

In the drug design process, the identification and optimization of drug candidates are complicated, due to the complexity of molecular structure and features. Computational

methods thus are often used. The use of such methods to support lead identification and optimization is referred to as "chemoinformatics".

Brown (1998) first introduced the term, "chemoinformatics", as:

"*The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information, and information into knowledge, for the intended purpose of making better decisions faster in the area of drug lead identification and optimization.*"

Gasteiger (2006) also provides a much broader definition of chemoinformatics: "*The application of informatics methods to solve chemical problems*". Hence, chemoinformatics covers many areas such as "*chemical structure representation, chemical reaction manipulation, data processing and data analysis, property prediction, chemometrics, data mining, structure elucidation, and synthesis design*" (Gasteiger, 2006).

More recently, Brown (2009) emphasizes that chemoinformatics is not only an essential component of chemical discovery, but also the field affected by mathematics, statistics, biology and computer science.

With the capabilities of generating compounds increased, it is expected that the drug discovery process can be significantly accelerated. The new field of chemoinformatics provides an invaluable tool in these efforts. As a crucial component of chemoinformatics, similarity-based virtual screening is the main focus of this thesis and is discussed in the next chapter.

The objective of this thesis is to further develop previous research findings on similarity coefficients and weighting schemes, and identify the most effective and appropriate approaches. To achieve this aim, the thesis is organized in the following way:

Chapter 2 begins by laying out the main components of molecular similarity search, and then introduces the most relevant and recent research findings. Chapter 3 describes the experimental design of this study, which includes the experimental databases, the methods of similarity search and the methods to evaluate the experimental results. Chapter 4 investigates the interactions between similarity coefficients and weighting schemes in similarity-based virtual screening. Three similarity coefficients and five weighting schemes are involved. Chapter 5 compares a large number of coefficients (i.e., 44) based on their performance on binary chemical similarity search. Chapter 6 explores the high performing coefficients in Chapter 5 and evaluates their performance when working with weighting schemes. The final chapter provides a conclusion of the thesis and suggests areas for future research based on this study.

Chapter 2: Similarity-Based Virtual Screening

2.1 Introduction

In this chapter, aspects associated with the concept of similarity-based virtual screening are reviewed. The review starts with an introduction of how computational technology influenced modern drug design as well as main approaches in virtual screening. Then the three principal components in similarity-based virtual screening are discussed in detail. Applications of similarity search and several relevant studies are also reviewed. Therefore, this chapter presents a theoretical basis of the studies reported in this thesis.

Traditional approaches to drug discovery rely on a step-wise synthesis and screening program for large numbers of compounds to optimize activity profiles. Since the 1980s, rational drug design has become the standard methodology for drug discovery. It is the inventive process which aims to find new medications based on the knowledge of the biological target (Guldbrandt *et al.*, 2002). The drug is usually recognised as a small organic molecule (sometimes also referred to as a ligand) that can activate or inhibit functions of a biomolecule, such as a protein, that benefits patients.

As shown in Figure 1.1, the first step of drug design process is the identification of a molecular target critical to a disease process or an infectious pathogen. The next step of drug design is the determination of the molecular structure of the target. In order to find suitable drug candidates, large numbers of compounds are tested to investigate how they interact with a certain biological target (Bajorath, 2002). This real screening process is known as *high-throughput screening* (HTS) and the most potential compounds obtained

are called *hits*. HTS can rapidly select those substances that affect the target; however, it is extremely expensive.

Since the 1960s, computers have been used in drug discovery in order to reduce the cost and increase the effectiveness (Ekins, 2006). Richon (1994) indicated that: "*…computational chemistry/molecular modeling is the science (or art) of representing molecular structures numerically and simulating their behaviour with the equations of quantum and classical physics.*" He claimed that in computational chemistry programs scientists can "*generate and present molecular data including geometries (bond lengths, bond angles and torsion angles), energies (heat of formation, activation energy, etc.), electronic properties (moments, charges, ionization potential and electron affinity), spectroscopic properties (vibrational modes, chemical shifts) and bulk properties (volumes, surface areas, diffusion, viscosity, etc.).*" At this stage, finding a suitable ligand is not the only factor of concern. In addition, according to Hubbard (1997), some other properties, bioavailability, metabolic half-life, lack of side effects etc. should be optimized before the ligand becomes an efficient drug .

Armed with this information, researchers in the pharmaceutical industry can use powerful computational technology to search through databases containing the structures of many different chemical compounds. The computer can select compounds that are most likely to interact with the receptor, and these can be tested in the laboratory. If an interacting compound cannot be found, other programs could be used that try to build molecules that are likely to interact with the receptor. Further programs can search databases to identify compounds with similar properties to a known compound. Thus, the idea is to narrow down the search as much as possible to avoid the expense of large-scale lab tests. This searching process also is known as virtual screening (VS).

The purpose of virtual screening is to narrow the range of target molecules by scoring, ranking and/or filtering molecular datasets with computational methods and data mining technologies. As defined by Walters *et al.* (1998), virtual screening is a process "*automatically evaluating very large libraries of compounds*" using computer programs.

After virtual screening, a manageable number of compounds can be targeted for synthesis, testing or purchase.

The process of virtual screening can be shown as a flowchart as follows (Leach and Gillet, 2007):

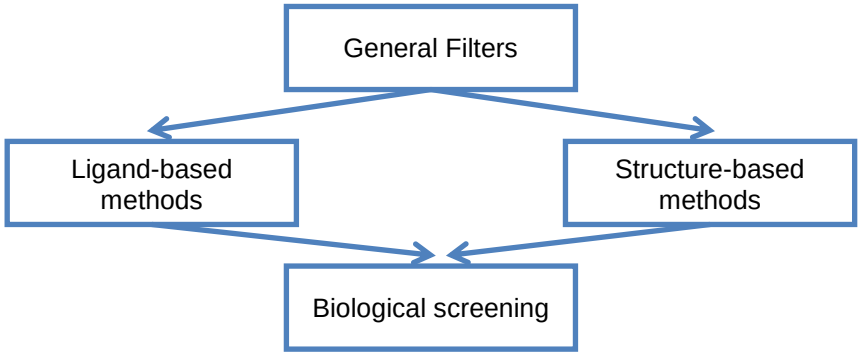


Figure 2.1 Typical virtual screening processes

As Figure 2.1 shows, typically the first step of virtual screening is to filter out the structures unlikely to be drug like molecules using some criteria, e.g., molecular weight, calculated value of logP, number of hydrogen bond donors and hydrogen bond acceptors (Lipinski *et al.*, 1997). After the general filters, the virtual screening techniques can be divided into two categories: ligand-based (if the ligand is known) and structure-based (if the target structure is known).

For ligand-based virtual screening approaches, three different techniques could be used in terms of the number of actives known. If only a single active molecule is known, such as a competitor's compound or a natural product, then similarity searching can be used. In addition, if several structurally related actives are available, then pharmacophore mapping can be applied to identify common patterns of features. Moreover, if it is difficult to identify common patterns and significant numbers of both active and inactive molecules are available, then machine learning techniques can be adopted to verify the structures that suitable for virtual screening, e.g., neural network (Leach and Gillet, 2007).

For structure-based virtual screening approaches, once the 3D structure of protein(s) is known then protein-ligand docking can be employed. Although structure-based design methods have been studied for many years and numerous methods have been suggested for protein-ligand docking (Taylor *et al.*, 2002), some drawbacks of structure-based approaches still need to be considered. Generally, there are two components to the docking problem, the method of identifying the poses of possible protein-ligands and the method of scoring the poses so as to find the binding mode for each compound. Due to the complexity and the flexibility of compounds, however, obtaining three-dimensional coordinates of the protein structure and docking ligands into the binding pocket of a target protein for large datasets are time consuming and difficult. Moreover, there are still problems regarding the ability of docking methods to predict the affinity or the rank of structures (Warren *et al.*, 2006).

Kuhn and colleagues (Schnecke and Kuhn, 2000; Zavodszky *et al.*, 2009) summarized challenges of structure-based virtual screening. Firstly, for each binding site, many ligand candidates in many orientations need to be evaluated. Secondly, normally as many as a hundred low-energy conformations exist for one candidate. Thirdly, usually many thousands of candidates need to be screened to identify several lead compounds. In addition, the process of scoring and screening is too long due to computational intensity. More recently, Cheng *et al*. (2012) reviewed a number of successful applications in structure-based virtual screening. They also highlighted the aspects which are crucial to a successful implementation, i.e., the in-depth knowledge of target-ligand interactions, optimized scoring function and the application of machine learning techniques.

Compared with structure-based virtual screening methods, one of the most widely used and simplest methods is using chemical similarity analysis (or molecular similarity search) methods to scan all the molecules from a dataset against one active structure.

As one of the most important topics in chemoinformatics, similarity search approaches have been intensively used and been considered to enhance the drug discovery process

(Geppert *et al.*, 2010; Jaworska and Nikolova, 2004; Johnson and Maggiora, 1990; Willett, 2008). Similarity search is the process of identifying molecules in a database which are structurally similar to a reference molecule, where a reference molecule is one that has been shown to have the biological activity of interest. It is based on Johnson and Maggiora's similar property principle: "*similar compounds have similar properties*" (Johnson and Maggiora, 1990). It has been demonstrated that structurally similar compounds do have similar biological activity, and that the biological similarity increases with the increasing structural similarity (Martin *et al*., 2002).

Similarity search is a simple and straight-forward method for retrieving chemical information. It thus becomes important when applied at the beginning of the drug discovery process. With the increase in computational ability and storage capacity, it was assumed that increasing the chemical diversity of compound libraries would enhance the drug discovery process. Similarity search is therefore also widely adopted in molecular diversity analysis and compound clustering (Dean and Lewis, 1999). The early studies of similarity search were conducted by Carhart *et al.* (1985) and Willett and Winterman (1986).

The number of reported chemical substances is over 71 million (CAS, 2013). This is a huge and consistently increasing (over one million new compounds are found in each year) amount of chemical data. Before any virtual screening task could be undertaken, database availability is the first consideration. These databases are specifically designed to store chemical information, e.g., information about chemical and crystal structures, spectra, reactions and thermophysical data. The information enables users to obtain the required results within seconds (Leach and Gillet, 2007; Raymond *et al.*, 2003).

The effectiveness of similarity searching, i.e., its ability to identify bioactive molecules, is determined by the similarity measure that determines the degree of resemblance between the reference structure and each of the database structures. Willett *et al*. (1998) identified three key components involved in similarity searching. They are structural representation, similarity coefficient and weighting scheme. With this discipline, the

similarity searching of chemical databases can be interpreted as: using a structure that is known to be active, comparing it with the rest of the structures of the database, measuring the quantity of likeness between the selected structure and each structure in the entire chemical database using a similarity coefficient. Different weighting schemes can be applied to the selected structure and/or the database to enhance similarity results.

2.2 Representations of Molecular Structures

A molecular structure represents a vast amount of information. The information can be as simple as the count of elements or as sophisticated as descriptions of its shape or electrostatic field. Therefore, molecules can be represented in various ways.

Basically, molecules are formed from collections of atoms and can be represented symbolically in several different ways. The molecular structures are represented by languages of chemistry which contain fundamental information on these molecules. It is difficult, however, to determine an optimum approach to represent molecular structures which is suitable for varied applications.

Traditionally, molecules are represented by molecular formulas, structural formulas and line drawings. Molecular formulas, also called chemical formulas, indicate the actual numbers of atoms of different elements in one molecule of a compound. One example of this is the chemical formula H_2O which indicates that there are two hydrogen atoms and one oxygen atom in a water molecule. In most cases, a molecular formula alone does not represent a unique molecule. For example, in the case of isomers, molecules with the same molecular formula have different arrangements of atoms. Structural formulas depict the structure of a molecule. They designate individual bonds between the atoms within a molecule represented as lines. There are several ways to achieve this, namely symbolic structural formula, graphical depiction as ‘ball and stick’ model or space-filled model represents a molecule in terms of the approximate size of atoms in three dimensional way.

These representations are more easily constructed and interpreted by people, but are less suitable for computers. Thus, finding a way to specify the features of molecules, identify the molecular structure even in the three-dimensional manner in which the atoms are bonded together is crucial. Moreover, they should be readable, computable and retrievable by computers, i.e., molecular representations.

Molecular representations can represent chemical structures and their properties in different format. They play a fundamental role in molecular similarity search. Their importance has been described by Todeschini and Consonni as follows: "*It is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment*" (Todeschini and Consonni, 2000).

2.2.1 Representations of 2D Molecular structures

In 2D, most representations are generated on the basis of graph theoretical methods (Biggs *et al.*, 1976; Varnek and Baskin, 2011). They account for topological properties. Since the 1960s when computers were first suggested for processing chemical information, different computer readable molecular representations have been developed.

2.2.1.1 Linear Notations

Linear notations represent a molecular structure in the form of a linear sequence of alphanumeric characters. They are simple and compact, and consequently are suitable for manipulation such as storing and retrieval of large numbers of molecules or compounds in a chemical information system (Leach and Gillet, 2007). The early line notations include the Dyson/IUPAC notation and the Wiswesser Line Notation (WLN) during 1960s to 1970s (Willett, 2009).

Later, the Simplified Molecular Input Line Entry Specification (SMILES) notation was widely used. It consists of a series of characters to specify how the non-hydrogen atoms

are arranged. Initially, it was designed as an input method. It was then found that SMILES was relatively easy and explicit for storing molecular structures.

Recently, a more advanced and increasingly-used line notation, called InChI (IUPAC International Chemical Identifier) was proposed by IUPAC (International Union of Pure and Applied Chemistry) and NIST (National Institute of Standards and Technology). It characterizes chemical structures by strings, and also contains more information than traditional line notations, such as the atoms and their bond connectivity, tautomeric information, isotopic information, stereo chemical and electronic charge information (Engel, 2006; IUPAC, 2011). Consider ethanol, CH3CH2OH, it can be represented as CCO in SMILES, InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3 in InChI. Comparing SMILES and InChl, the former was developed for use in in-house industrial chemoinformatics systems, while the latter was developed for use in open-source chemoinformatics software (Holliday and Willett, 2011).

Linear notations can be used for storing compounds as a compact molecular representation. They can also be used for computational manipulation. They cannot, however, provide explicit information of 2D arrangement that the chemoinformatics systems require for some cases. Thus, the connection table, a data structure that records the molecular topology information, i.e., the atoms within a molecule and the ways that bonds link those atoms together, need to be introduced.

2.2.1.2 Connection Table

The connection table is another notable format of chemical structure representation in a computer system and is also a suitable approach for representing molecules as graphs (Engel, 2006). A connection table is an example of a graph. It describes a set of objects and their relationships as nodes and edges in a mathematical construct (Diestel, 2000; Wilson, 1996). It is a 2D matrix containing chemical information about all the atoms and bonds in a 2D structure. In comparison with SMILES notation, a connection table provides the same information but in a different format. Each row represents information about a particular atom such as the atom number, symbol, and number of

atoms to which it is directly bonded. Each atom is numbered as an index forming an atom list; moreover, each row in the bond list shows the indices of two atoms connected by a particular bond type.

Morgan (1965) introduced an algorithm involving a node labeling technique in order to obtain unique machine descriptions of chemical compounds. The algorithm can be described as follows (Leach and Gillet, 2007): First, all nonhydrogen atoms are assigned numbers according to the number of heavy atoms to which they are attached. Second, for each atom, calculate a new value as sum of each of its neighbors and assign the new value to the atom. Repeat this process until the numbers of each atom are unique and no longer increase. Finally, assign the highest labeled atom as number "1" and then assign its highest labeled neighbour atom as "2", second highest as "3", etc. Then, number the atoms which are attached to atom "2" in order of label values, and so on, until all atoms are numbered.

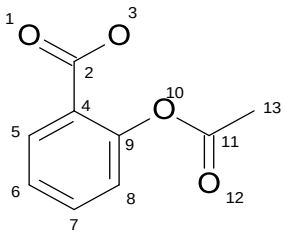
By adopting the Morgan algorithm (Leiter *et al.*, 1965), the connection tables have a unique coding of the inter-connections between the atoms in a molecule. For storing and arranging molecules in databases, Chemical Abstracts Service set another standard which gives each compound a unique number to identify them with their connection table molecular representation in the CAS Registry System (Leiter *et al*., 1965).

Figure 2.2 shows different representations of the structure of aspirin, where nonhydrogen atoms are numbered from 1 to 12. In the connection table representation, each row illustrates the way that the corresponding atoms are connected to the others. For example, the first row demonstrates that atom number 1 (Oxygen) is connected by a double bond (D) to atom number 2; the second row shows that atom number 2 (Carbon) is connected by a double bond (D) to atom number 1 and is connected by single bonds (S) to atom number 3 and atom number 4.

2.2.1.3 Reduced Graphs

As mentioned at the beginning of Section 2.2.1, a chemical structure can be represented as a topological graph, where the nodes of the graph correspond to the atoms, and the edges represent the bonds. In a reduced graph, each node represents a group of connected atoms. An edge links two nodes, if there is a bond between any two atoms from the two separate groups.

Reduced graphs were initially designed for structure and substructure searching (Gillet *et al.*, 1991). The subsequent studies also demonstrated they can be employed for similarity searching, see Figure 2.3 (Gillet *et al.*, 2003; Takahashi *et al.*, 1992).



InChI: 1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)

Smiles: CC(=O)Oc1ccccc1C(=O)O Name: 2-acetoxybenzoic acid

Connection Table:

1 O D 2
2 C D 1 S 3 S 4
3 O S 2
4 C S 2 D 5 S 9
5 C D 4 S 6
6 C S 5 D 7
7 C D 6 S 8
8 C S 7 D 9
9 C S 4 D 8 S 10
10 O S 9 S 11
11 C S 10 D 12 S 13
12 O D 11
13 C S 11

Figure 2.2 Structure, name, InChI, SMILES and connection table for aspirin (from (Holliday and Willett, 2011)).

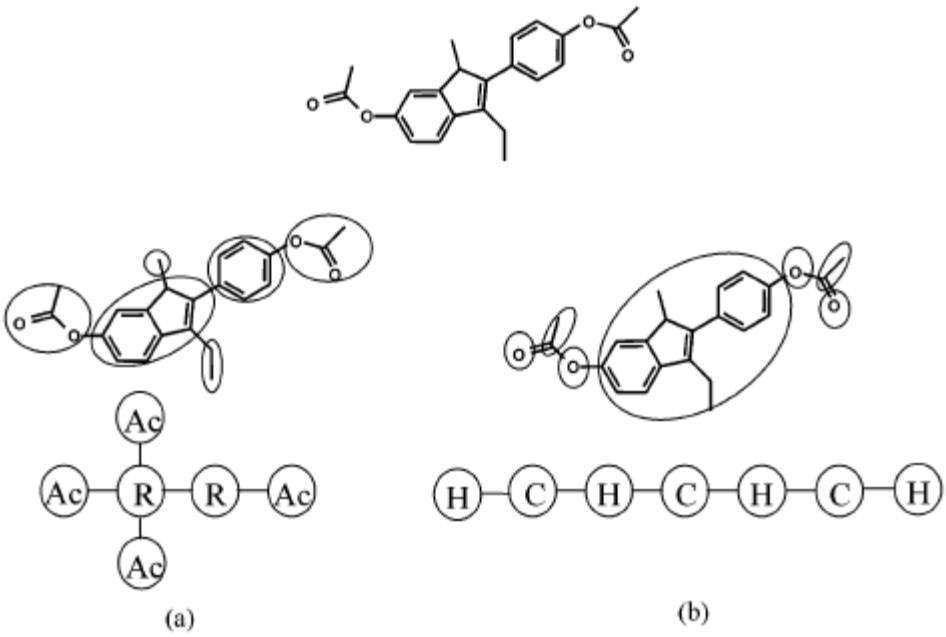


Figure 2.3 Two examples of different types of reduced graphs.

According to the top original structure: (a) nodes represent to ring systems Ⓡ and connected acyclic components Ⓐ; (b) nodes correspond to connected carbon components Ⓒ and connected heteroatom components Ⓗ (from (Gillet *et al.*, 2003)).

2.2.1.4 Fingerprints

Neither linear notations nor connection tables, however, are suitable for performing similarity search in large scale databases efficiently due to computational complexity. Thus, fingerprints, a form of molecular representation which simply uses bit strings to indicate occurrence of molecular features, provide more rapid system for similarity search (Bajorath, 2002). The early study of this machine readable format was recorded by Adamson *et al* in the 1970s (Adamson and Bush, 1975; Adamson *et al.*, 1973). Initially, fingerprints were designed to support chemical database substructure searching; currently, they are also widely used to carry out other applications, such as similarity searching, clustering, and classification (Rogers and Hahn, 2010).

In chemoinformatics, molecules can be characterized as structures consisting of different substructures or fragments. Thus, a fingerprint X of molecule A can simply be defined as below which basically is a sequence of numbers, see **Error! Reference source not found.**.

$$X_A = \{x_1, x_2, \ldots x_n\}$$

Equation 2.1 A fingerprint X of molecule A

Where x_i specifies the *i*-th structural unit contained in the molecule A, e.g., atoms, bonds or fragments. The value n relates to the length or the size of the fingerprints, i.e., the number of properties the molecule has.

Rogers and Hahn stated a fingerprint of a molecule can be described as a sequence of bits or integers (Rogers and Hahn, 2010). For the sequence consisting of bits, each bit represents the presence or absence of a pre-defined substructure or fragment. If a substructure is present, then its relevant bit is set to "1" which is similar to the "on" bit in binary code. In contrast, if a feature is absent, then the relevant bit is set to "0". For the sequence consisting of integers, each non-zero number represents the frequency of occurrence of a fragment or molecular feature and zero indicates absence (Leach and Gillet, 2007).

Commonly, there are two different types of fingerprints based on the ways used to generate them: dictionary-based fingerprints and hashed fingerprints (Leach and Gillet, 2007).

For generating dictionary-based fingerprints, the most important task is creating a dictionary. The dictionary must contain a set of structural fragments that is used to decide whether each element in the fingerprints is set on or off. Figure 2.4 illustrates how dictionary-based fingerprints are generated. In this example, there is a simple dictionary consisting of just five fragments (usually dictionaries contain around 500 to 5000 fragments or substructures). Four fragments are specific and one is generalized,

representing two substructures: one contains sulphur and the other contains oxygen. Thus, the chemical structure in the left can be represented as a fingerprint as "11100" which shows that the first three substructures are present but not the last two.

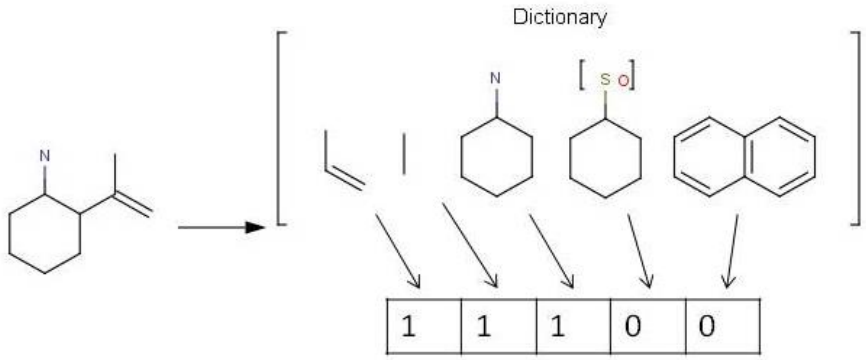


Figure 2.4 An example of how a dictionary-based fingerprint is generated (Based on Digital Chemistry, 2011).

Unlike the dictionary-based fingerprints, hashed fingerprints do not rely on pre-defined substructure dictionaries. Take Daylight fingerprints as an example, the process of generating hashed-based fingerprints can be described as: "*an algorithm is used for generating paths throughout a compound, with all the elements on the path being represented; then, a hashing function is used to create the binary fingerprints*" (Daylight Chemical Information Systems, 2011). As shown in Figure 2.5, bit collision is allowed, i.e., various smaller fragments from different bigger fragments may share the same fingerprint position(s). The relationship between fragments and fingerprint position in hashed-based fingerprints, therefore, is many to many. This also means that once the fingerprint has been generated, it is not possible to identify the fragment(s) on each position.

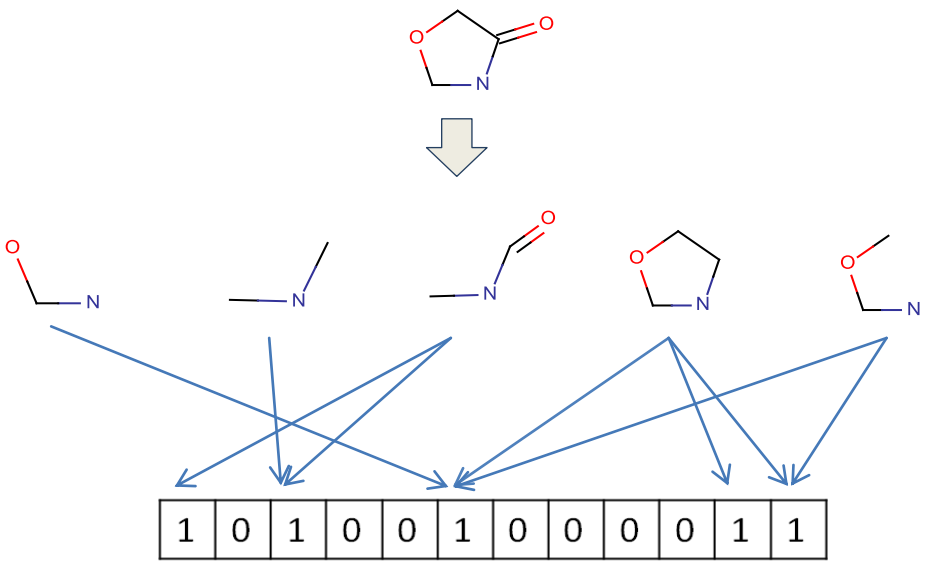


Figure 2.5 An example of how a hashed-based fingerprint is generated.

Regardless of the differences between dictionary-based fingerprints and hashed-based fingerprints, both of them have drawbacks. For the former, each bit maps to a specific fragment which is pre-defined in a structural fragment dictionary. The dictionary has to be created in the first place and then once the dictionary is changed, all the fingerprints should be changed as well. For the latter, each bit is a set of fragment combinations rather than a specific structural features, but this method may lead to ambiguity in which case it may be impossible to interpret bits back to fragments (Bajorath and Eckert, 2006).

To avoid the shortcomings of either dictionary-based or hashed-based fingerprints, some fingerprints take advantages from both of them, e.g., structural keys and path-generating algorithm(s). An example is the Extended-connectivity fingerprints (ECFP), a newly developed fingerprint methodology. Since the ECFPs is the descriptors which are used in this study, all details are discussed in Chapter 3.

Binary fingerprints, as discussed above, are one of the most popular molecular descriptors and have been intensively used in similarity-based virtual screening (Carhart *et al.*, 1985; Willett *et al.*, 1986). Recently, occurrence-based fingerprints started attracting more interests. Molecular hologram (Hurst and Heritage, 1998; Mezey, 2001)

is one of the representations known as occurrence-based descriptors. It consists of integers to encode the frequency a fragment occurs in a molecule rather than using 0s and 1s to encode the absence and presence of a fragment. Alternatively, occurrence-based fingerprints can be considered as a sort of weighted fingerprints. A further discussion of weighting is in Section 2.3.

2.2.2 3D Molecular Representations

In comparison to 2D molecular descriptors, 3D molecular representations also consider molecular geometries. They encode spatial relationships between atoms, ring centroids, and planes. Thus, 3D molecular representations can characterize molecules in a more accurate and unambiguous way, e.g., the difference between isomers can be easily found by 3D molecular representations but not so easy by 2D representations.

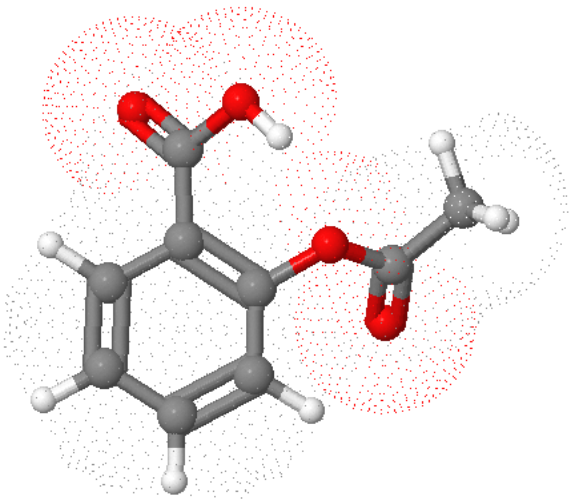


Figure 2.6 A 3D model of the Aspirin structure (http:// www.3dchem.com).

Figure 2.6 shows how the atoms are located geometrically in relation to each another in aspirin. The 3D structure can be generated from experimental data, e.g., X-ray crystallography, electron diffraction, and microwave spectroscopy. Typically, a 3D molecular descriptor consists of values of molecular surface area, volume, shape, spatial

distribution of atoms, potential energy, molecular properties, substructural groups or electrostatics.

Most 3D molecular descriptors are calculated from a 3D connection table or chemical graph, which can be obtained either experimentally or computationally. The Cambridge Structural Database (CSD, 2011) and the Protein Data Bank (Berman *et al.*, 1999; PDB, 2011) are the databases which contain large numbers of experimentally generated molecular structures. Alternatively, 3D molecular representations can also be generated by structure generation software programs, such as CONCORD and CORINA. They are usually used to build computationally determined 3D molecular structures from a molecular graph (Pearlman, 1987).

In the various descriptors, one of the most popular 3D molecular descriptors is 3D fingerprints which are similar to the 2D fingerprints described in the previous section. The major difference is that the structural units of 3D descriptors are based on 3D information, e.g., 3D distance-based, angle-based or pharmacophore-based.

The distance-based fingerprints are based on the distance of features, such as atoms, atom types, ring centroids and planes (Pepperrell and Willett, 1991). In this case, all the distances are assigned to a distance range which could be equidistant or not, if the specific distance between two features is present. Then, the corresponding bit in the fingerprints is set to 1; otherwise it will be set to 0.

The angle-based fingerprints are based on valence or torsion angles without considering if the corresponding atoms are connected (Bath *et al.*, 1994).

An alternative approach to 3D fingerprints is based on the concept of a pharmacophore, as the portion of a ligand molecule which binds to the receptor site, i.e., a spatial arrangement of chemical groups (e.g., atoms or the centroid of an aromatic ring). IUPAC defines a pharmacophore to be "*an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific*

biological target and to trigger (or block) its biological response" (Wermuth *et al.*, 1998). The feature types include: Hbond donors, Hbond acceptors, positive charge, negative charge, positive ionizable, negative ionizable, aromatic ring and hydrophobic. Pharmacophore-based fingerprints can be specified as 2D pharmacophore fingerprints and 3D pharmacophore fingerprints: the former is defined by molecular feature or pharmacophore points and the corresponding inter-feature distances; the latter is similar to the former but using Euclidean distances based on 3D coordinates instead of topological distances. Figure 2.7 illustrates 3D coordinates and conformation of pharmacohpore points.

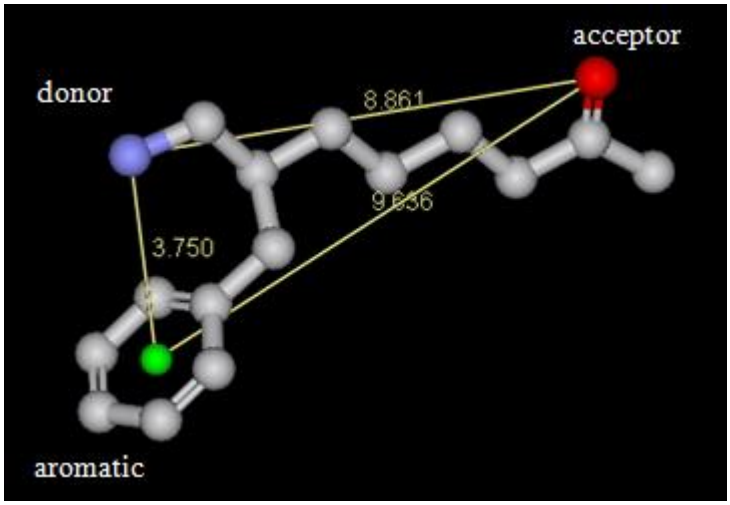


Figure 2.7 3D coordinates and conformation of pharmacophore points.

As a pharmacophore can be described as the spatial arrangement of functional groups required for binding, it can therefore be used to identify the features of one or more molecules with the same biological activity, i.e., similarity search; or used to do molecular dissimilarity search (Beno and Mason, 2001; Leach and Gillet, 2007). Typically, the pharmacophore-based descriptors consider only the heavy atoms of a molecule.

Recently, 3D fingerprints have been suggested in similarity searching, e.g., 3D pharmacophore fingerprints and 3D fingerprints generated by protein-ligand interaction information (Baroni *et al.*, 2007; Deng *et al.*, 2004; Kelly and Mancera, 2004; Mason *et al.*, 1999). Several investigations tried to combine 3D information into 2D fingerprints for similarity searching. Tan *et al.* (2008) attempted to do 2D fingerprint-based similarity searching with 3D interaction information and still keep the 2D fingerprint format. Based on their results, they indicated that the conventional structural fingerprints-based similarity search can be further improved using interacting fragments which capture much compound class-specific information.

As stated above, due to the complexity of molecular structures and features, many studies showed it is very difficult to design a universal molecular representation suitable for different demands. In similarity search, the key should be the chemical libraries that cover as much chemical substructures or fragments as possible.

2.3 Weighting Scheme

Binary 2D fingerprints consist of "0"s and "1"s to show the incidence of fragments. They can, however, lose important information regarding how frequently they occur. This may result in a considerable impact on similarity search. For instance, if the key information to distinguish molecules from one class to another is that a certain fragment has to occur twice, then using binary fingerprints could not provide the correct information. In addition, a fragment with a high weight occurring in both a reference structure and a database structure should make a greater contribution to the overall degree of inter-molecular similarity than will a fragment in common that has a lesser weight. Therefore, fingerprint weighting schemes need to be considered in similarity search (Cosgrove and Willett, 1998; Ormerod *et al.*, 1989; Willett and Winterman, 1986).

In a study of simulation of property-prediction, Willett and Winterman suggest three types of weighting: "*weighting based on the frequency of a fragment's occurrence in an individual compound; weighting based on the frequency of a fragment's occurrence in an entire database and weighting based on the total number of fragments within a compound*" (Willett and Winterman, 1986). The former two types of weighting have been studied in work by Arif *et al.* (2009b, 2010), who found that the first type of weighting could bring about notable increases in screening effectiveness in some circumstances, but that the second type was of less general applicability. The study reported in this thesis hence focuses on the first approach, i.e., on exploiting information on how frequently fragments occur within individual molecules.

Experiments have found that occurrence-based representations which show the nature of a molecule may be superior to the incidence-based (binary) representations either on small scale datasets or in large scale simulated virtual screening (Arif *et al.*, 2009a; Baldi *et al.*, 2007; Chen and Reynolds, 2002; Oprea *et al.*, 2004; Willett and Winterman, 1986). It is not always the case, however, as argued by Arif *et al.* (2010) that more occurrences may cause lower similarity.

There are some weighting schemes successfully applied in other domains which can be implied in molecular similarity search, e.g., tf-idf in information retrieval and text mining (Sparck Jones, 1972). Tf-idf weight stands for *term frequency- inverse document frequency*. It has been intensively used to evaluate how important a word is to a document in a collection or corpus. The importance of a word to a document increases corresponding to the number of times it appears in the document, but offset by the frequency of it in the corpus. For example, consider a document containing 100 words wherein the term A appears 7 times. Thus, the tf for A is (7/100) = 0.07. Assume the term A appears in one hundred of ten thousand documents. Then the idf is computed as log(10,000/100) = 2. Therefore, the tf-idf weight is the product of the two quantities: 0.07*2=0.14.

Many variations of the tf or idf weighting scheme have been adopted in chemical similarity search. These weighting schemes enable differentiation of more important features from less important features in a representation. Several studies have shown weighted fingerprints gave far better results than did un-weighted fingerprints (Grier *et al.*, 1988). Other studies, however, also suggested weighting schemes show effectiveness only with appropriate similarity coefficients (Arif *et al.*, 2010). This suggestion is discussed in detail in Chapter 4.

Many other weighting schemes have been applied to fingerprint weighting, e.g., square root to reduce the impact of high occurrence features, log to ignore the features which occurred only once (Arif *et al.*, 2009a). As a result of their successful applications to text categorization, some term weighting schemes may be able to be adopted in similarity search, e.g., tf.rf, tf, logtf, ITF (Lan *et al.*, 2005; Lan *et al.*, 2007).

2.4 Similarity Coefficients

Regardless of varieties of molecular descriptions, similarity search follows a standard rule as described earlier in this chapter. It is necessary to compare the reference molecule and database molecules in a pair-wise manner and rank the similarity results in decreasing order by using a similarity coefficient.

Hence, once the (possibly the weighted) molecular descriptions of both the reference molecule and molecules from the database are clarified, similarity coefficients can be used to quantify the similarity between them. Similarity coefficients are used in a wide range of disciplines such as biology, information retrieval, multivariate statistics, numeric taxonomy and marketing (Willett *et al.*, 1998). In similarity search, a similarity coefficient is used to measure the grade of likeness between pairs of objects. Each object can be described by some number of attributes or descriptors (Holliday *et al.*, 2002; Leach and Gillet, 2007).

There are four categories of coefficient described, i.e., distance, association, correlation, and probabilistic (Sneath and Sokal, 1962). Most of these coefficients can be used for either binary or continuous (e.g., real value vectors) descriptions, whereas the first two coefficients are commonly discussed in the literature and widely used in similarity searching (Holliday *et al.*, 2002; Willett, 1987). For binary molecular descriptors, usually three different quantities are involved to measure the similarity between the two bit vectors: the number of bits set on in both molecules (a); the number of bits set on only in molecule A (b) and the number of bits set on only in molecule B (c). For continuous descriptors, the equivalent three quantities are defined as:

$$a = \sum_{i=1}^{n} x_{iA} x_{iB}$$

$$b = \sum_{i=1}^{n} (x_{iA})^2 - \sum_{i=1}^{n} x_{iA} x_{iB}$$

$$c = \sum_{i=1}^{n} (x_{iB})^2 - \sum_{i=1}^{n} x_{iA} x_{iB}$$

Equation 2.2 The a, b and c parameters for the calculation of similarity coefficient when using continuous descriptors.

where x_i represents the value of *i*-th element (property) of molecule *A* which has *n* properties. A few examples of most commonly used coefficients are listed in Table 2.1., more coefficients are described in Chapter 5 and Chapter 6.

2.4.1 Distance Coefficients

Distance coefficients are widely used for their simple geometric representation. In chemoinformatics, they are commonly applied to measure the dissimilarity between structures in a molecular space or the distance between two compounds. The value of distance indicates the discrepancy between compounds under consideration (Sneath and Sokal, 1962). Two objects are identical if their positions coincide, which means that the distance value between them, is 0; consequently, as the distance increases the probability of the two objects to be similar decreases.

Distance coefficients are referred to as metrics if they obey the following criteria (Leach and Gillet, 2007), if there are three objects, A, B and C:

- The distance values are zero or positive, and the distance between identical structures is zero: $D_{AB} \geq 0$, $D_{AA} = D_{BB} = 0$;
- The distance values are symmetric: $D_{AB} = D_{BA}$;
- The distance values obey the triangular inequality rule: $D_{AB} + D_{BC} \geq D_{AC}$;
- The distance value between non-identical structures is greater than zero: if $A \neq B$, then $D_{AB} > 0$.

If a distance coefficient only obeys the first three criteria, then it is referred to as "pseudo-metric", and a distance coefficient that does not satisfy the third criterion is called non-metric (Willett *et al.*, 1998). Examples of metric distance coefficients include the Hamming, the Euclidean and the Soergel coefficients. The Euclidean coefficient is also one of the most popular distance coefficients and is broadly used in nearest neighbour algorithms or in clustering.

2.4.2 Association Coefficients

Association coefficients measure the agreement between pairs of compounds (Sneath and Sokal, 1962). They work well with both binary fingerprints and continuous fingerprints. Compared with distance coefficients, binary association coefficients measure the similarity between two compounds in which the value ranges from 0 and 1, denoting no similar features in common and an identical description, respectively (Salim *et al.*, 2003). The higher the value of the coefficient, the more similar the two objects are. As one of association coefficients, the Tanimoto coefficient is the most commonly used similarity coefficient in 2D fingerprint similarity searching. It is a simple and intuitive coefficient and certainly most used in chemoinformatics (Chen and Reynolds, 2002; Willett, 2006):

Other association coefficients are also popular, i.e., Dice, cosine, Fossum, Rusell-Rao and Forbes coefficients (Leach and Gillet, 2007). While the cosine coefficient has been

widely used in information retrieval with good performance, it has not been used in molecular similarity search as often as the Tanimoto coefficient.

Another example of the association coefficient is the Tversky coefficient (Tversky, 1977), which is:

$$S_{AB} = \frac{a}{\alpha b + \beta c + a}$$

Equation 2.3 The Tversky coefficient for binary variables

It allows the user to bias the similarity calculation with one of the structures, e.g., when $\alpha = 1$, and $\beta = 0$, the Tversky similarity value $a/(a+b)$ represents the fraction of features of A that are present in B; if the similarity value is 1, then A is a substructure of B. If $\alpha = \beta = 1$ then the Tversky coefficient is identical to the Tanimoto coefficient and when $\alpha = \beta = \frac{1}{2}$ then it is identical to the Dice coefficient which is widely used in information retrieval. Recently, the Tversky coefficient has been recommended for perceiving features and similarity of images. In chemoinfomatics, some studies showed the Tversky coefficient can enhance similarity search by optimizing the effect of the size of molecular structures, e.g., similarity calculation using MACCS keys on the MDDR database (Wang and Bajorath, 2008; Wang *et al.*, 2007).

2.4.3 Correlation Coefficients

Correlation coefficients measure the degree of correlation, i.e., proportionality and independence, between the sets of values that describe the pair of objects. Typical examples are the Spearman Rank coefficient, Pearson, Stiles and Yules correlation coefficients (Salim *et al.*, 2003). The Spearman rank coefficient was first applied in chemoinformatics as a measure of electrostatic similarity (Manaut *et al.*, 1991). It is illustrated below:

$$\rho_{AB} = 1 - \frac{6\sum_{i=1}^{n}(x_{iA} - x_{iB})^2}{n(n^2 - 1)}$$

Equation 2.4 The Spearman rank coefficient

The use of correlation coefficients in chemoinformatics is problematic. Jardine and Sibson (1971) indicated that correlation coefficients might not be an appropriate measurement for similarity comparisons. Hubalek (1982) also pointed out that the values in a comparison cannot show the similarity of two objects because it can either be correlated or not. Normally, for correlation coefficients, the used properties should have independent distributions. In chemical similarity search, however, chemical information is usually not independent and the properties are correlated with each other. Thus, the correlation coefficients might not be suitable for similarity search.

2.4.4 Probabilistic Coefficients

Probabilistic coefficients are based on the frequency distribution of the descriptors in a database. One study found that they can yield poor performances when applied in chemistry and required extensive computations (Adamson and Bush, 1975). They have therefore not been investigated to any extent in molecular similarity measures.

2.4.5 Choice of Coefficient

There are various types of coefficients available for similarity search. In order to choose the most appropriate one, certain factors need to be considered. First, different similarity coefficients denote similarity on different ranges. For example, the result measured by the Tanimoto coefficient ranges by zero to one whereas some other coefficients such as Euclidean provide a wider range of zero to infinity. Therefore, a standardization procedure may be required to convert the attribute value to a range of zero to one (Holliday *et al.*, 2002; Leach and Gillet, 2007). Second, the molecular size may also

affect the calculation of similarity especially on the binary representations, as many conventional similarity coefficients only take bits into account that are set to 1, e.g., the Tanimoto coefficient (Holliday and Haranczyk, 2008; Salim *et al.*, 2003). For example, in a binary fingerprint similarity search with the Tanimoto coefficient, the small molecules usually have lower similarity scores since they are likely to have fewer bits set in fingerprints than large molecules. By contrast, small molecules tend to be more similar when using the Hamming distance (Leach and Gillet, 2007). With such biases of coefficients on different sized molecules, it requires some degree of size standardization to avoid such problems.

It has been established that there is no single coefficient which consistently performs better than others (Holliday and Haranczyk, 2008; Willett *et al.*, 1998). Further research has suggested that using mixed indices combining two or more standard measures may exhibit better performance on similarity searching (Leach and Gillet, 2007). In brief, it may be true that there is still a need to find the most appropriate coefficient or combination of coefficients for specific similarity searching applications.

2.5 Data Fusion

Although many types of descriptor have been used in similarity searching, by far the best established is the 2D fingerprint (Willett, 2006). As one of the traditional methods for chemical database mining, the similarity measure using 2D molecular fingerprints is normally used with the Tanimoto coefficient. However, this conventional similarity measure might not be the most suitable choice for similarity searching. Thus, developing optimized similarity measures has been an important topic in chemoinformatics for a long time and continues to be an interesting subject in drug discovery.

Table 2.1 Typical coefficients commonly used in similarity search.

S_{AB} and D_{AB} represent the value of similarity and the value of distance of two molecules A and B, respectively. a is defined as the number of bits set on in both molecule A and B; b as the number of bits set on only in molecule A; c is the number of bits both set on only in B; and x_i represents the value of *i*-th element (property) of a molecule. Table based on (Leach and Gillet, 2007; Willett *et al.*, 1998).

Name	Formula for continuous variables	Formula for binary variables
Tanimoto coefficient	$S_{AB} = \frac{\sum_{i=1}^{n} x_{iA}x_{iB}}{\sum_{i=1}^{n}(x_{iA})^2 + \sum_{i=1}^{n}(x_{iB})^2 - \sum_{i=1}^{n} x_{iA}x_{iB}}$	$S_{AB} = \frac{a}{a+b+c}$
Dice coefficient	$S_{AB} = \frac{2\sum_{i=1}^{n} x_{iA}x_{iB}}{\sum_{i=1}^{n}(x_{iA})^2 + \sum_{i=1}^{n}(x_{iB})^2}$	$S_{AB} = \frac{2a}{2a+b+c}$
Euclidean distance	$D_{AB} = [\sum_{i=1}^{n}(x_{iA} - x_{iB})]^{\frac{1}{2}}$	$D_{AB} = [b + c]^{\frac{1}{2}}$
Hamming distance	$D_{AB} = \sum_{i=1}^{n}\lvert x_{iA} - x_{iB}\rvert$	$D_{AB} = b + c$

A number of studies as discussed at the end of Section 2.2 focus on the enhancement of fingerprints rather than considering other key factors in similarity search. There have been many comparisons of fingerprints and similarity coefficients for similarity searching, e.g., the recent detailed studies by Duan *et al.* (2010) and Sastry *et al.* (2010). However, selecting an appropriate fingerprint for a given problem is still a challenging topic that requires ongoing research and improvement. The same also goes for choosing the appropriate coefficients and weighting schemes.

Considering that no single similarity measure will be consistently the best, data fusion has been increasingly used in similarity search for years. This idea is derived from

information retrieval in which the integration of multiple sources of evidence has been shown to greatly enhance the retrieval performance (Willett, 2006a; Croft, 2000)

In the domain of chemoinformatics, this technique is also known as "consensus scoring" in ligand docking and scoring algorithms. Two main approaches are widely adopted: using one reference structure with different descriptors or similarity coefficients for each search, combining all of the results; using multiple reference structures with the same descriptor and similarity measure, combining all the results. For each case, the key factor is the fusion rule which is a function to calculate the final score.

A typical fusion process can be described as follows (from Leach and Gillet, 2007).

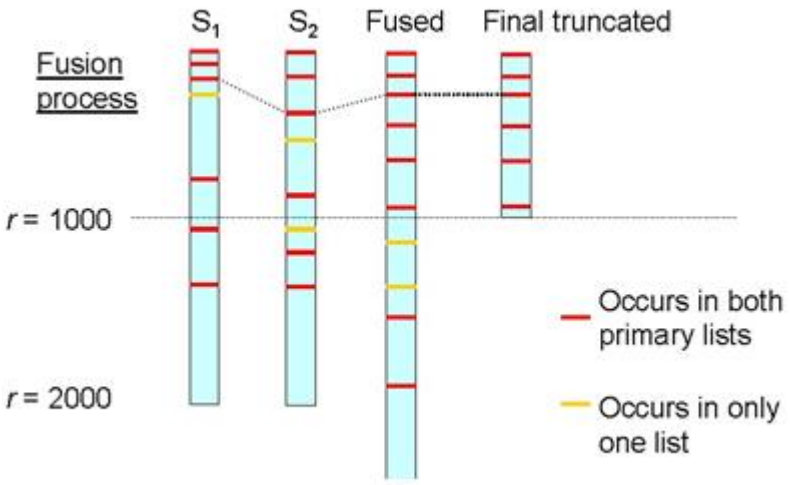


Figure 2.8 A typical fusion process (Leach and Gillet, 2007).

As shown in Figure 2.8, S1 and S2 are the results which are produced when using different similarity measures. Then, a fusion rule involves combining the results, e.g., MIN, MAX or SUM. The final step is truncating the top (usually 1% or 5%) from the combination list.

Many studies found data fusion methods can enhance the similarity search. Ginn *et al.* (2000) investigated the application of similarity data fusion methods on the

combinational ranking results calculated by various types of descriptors and datasets. They studied three fusion rules, MAX, MIN and SUM. Their findings showed that data fusion performed more consistently than any single similarity measure; besides, the SUM rule was the most effective and robust rule. Salim and co-workers (2003) studied the SUM rule in similarity searches using the same descriptor for the compounds in the database, but varied the number of fused coefficients. The results showed that data fusion improved the overall results of similarity searches compared with the use of a single coefficient. However, they also found none of the combinations was consistent enough across all the searches.

The subsequent studies (Hert *et al.*, 2004a; Whittle *et al.*, 2004) of data fusion discussed the potential problems in previous studies, which is the fact that different similarity measures cannot be directly fused due to incompatibilities among them. They proposed an alternative to the previous similarity fusion approach which is called "group fusion". The main idea of group fusion is using multiple reference structures rather than multiple similarity measures. Furthermore, they calculated the actual similarity scores instead of the rankings results. It showed that the fusion of scores was much more effective than the fusion of ranks, especially in the case of more heterogeneous drug classes. Moreover, the MAX rule was superior to SUM (Hert *et al.*, 2004a).

More recently, many further investigations of group fusion have been carried out (Chen *et al.*, 2009; Gardiner *et al.*, 2009; Hert *et al.*, 2006; Medina-Franco *et al.*, 2007; Muchmore *et al.*, 2008; Whittle *et al.*, 2006; Williams, 2006). According to these studies, data fusion was recommended to be utilized in similarity search either using multiple reference structures or using multiple similarity measures.

Gardiner *et al.* (2009) reported a detailed evaluation of an extension of similarity searching, namely, turbo similarity searching (TSS). TSS makes the assumption that the nearest neighbors of a reference structure are truly active and these can then be used as reference structures in addition to the original reference structure, for group fusion. The process of TSS is shown as Figure 2.9, where the user provides an original reference

structure to match against each of the database structures, the top ranked structures are the nearest neighbors and are then used as the reference structures. Therefore, for k nearest neighbors, k rankings of similarity searching are produced in addition to the ranking results from the initial similarity searching. Then the k+1 rankings are fused into a single ranking as the ranked output by using group fusion rule. The results showed that TSS can provide enhancements in screening performance but that this is normally achieved only if the initial similarity searching has already achieved some reasonable level of search effectiveness. They (Gardiner *et al.*, 2009) also concluded that: the ECFP_4 fingerprints would be the choice of structure representation for similarity-based virtual screening; TSS is likely to provide notable enhancement in screening performance if the actives are tightly grouped, but is unlikely be effective for heterogeneous sets of actives.

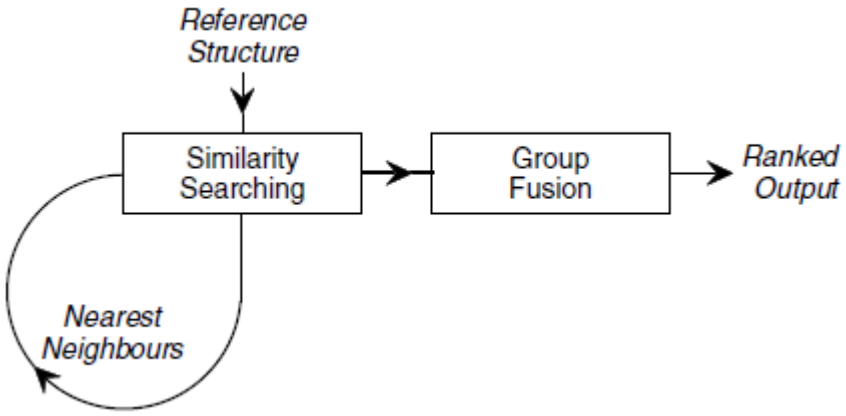


Figure 2.9 Schematic outline of a turbo similarity search (Gardiner *et al.*, 2009).

There are many parameters which could affect the results of fusion based similarity search, i.e., the type of structure representations, the selection of databases, the choice of fusion rules, the selection of weighting schemes or similarity coefficients etc. It can thus be suggested that developing an optimized similarity measure is the precondition for the implementation of data fusion. This is also the aim of the study reported in this thesis.

2.6 Conclusion

This chapter discussed the three key components involved in similarity search: molecular descriptors which are used to represent and differentiate compounds; similarity coefficients which measure the degree of similarity between compounds; and the weighting schemes which can be used in similarity measurements. It also introduced the common approach of similarity search and the relative works of this thesis.

To sum up, the performance of similarity search is determined by combinations of similarity measures used. As discussed before, molecular representations are the most important element in similarity search. However, molecular structures are very complex. A little variance may cause significant differences of features. This is why significant improvement in similarity search can be achieved through studying weighing schemes design and comparing different coefficients. Thus, the main objective of this PhD study is to investigate the use of different coefficients and weighting schemes in similarity-based virtual screening.

Chapter 3: Methodology

3.1 Introduction

This chapter will outline the experimental design used for the three investigations reported in this thesis. The three investigations are: first, the evaluation of interactions between weighting scheme and similarity coefficient in similarity-based virtual screening (reported in Chapter 4); second, the comparison of established level of binary coefficients for chemical similarity search (reported in Chapter 5); third, the comparison of similarity coefficients using weighted chemical data (reported in Chapter 6). Since the experimental background was mostly in common to all three chapters, this chapter provides the details of methodology in terms of the databases and evaluation methods used.

3.2 Data

3.2.1 Chemical Databases

There are several databases now available for evaluation purposes, and four were used in this study, so as to ensure that the results obtained are not overly dependent on the nature of the test data. These databases were: the MDL Drug Data Report and WOrld of Molecular Bioactivity databases (MDDR and WOMBAT, as described in detail by Gardiner et al. (2009)); the Maximum Unbiased Validation database (MUV, as described in detail by Rohrer and Baumann (2009)); and the ChEMBL database (Gaulton *et al.*, 2012; Heikamp and Bajorath, 2011).

In each database, there are sets of molecules with some specific biological activity and the remainder of the database is assumed to be inactive. The effectiveness of similarity

searching can hence be evaluated by the extent to which it is possible to retrieve the known active molecules from the database.

3.2.1.1 MDDR

The MDDR database (available from Accelrys Inc. at http://www.http://accelrys.com/), developed in collaboration with Prous Science, provides essential information from new patent applications about drugs recently launched or under development. It contains the structures and pharmacological class information for molecules that have been reported in patents, journals and conference proceedings as exhibiting biological activity, i.e., over 180,000 biologically relevant compounds and well-defined derivatives. All of the information can be found in Prous Science's Drug Data Report. Each year, Prous Science publishes about 10,000 new compounds including primary compounds, derivatives and formulations (Accelrys Software, 2009). The activity data is qualitative: a molecule is noted as exhibiting a specific activity, and it is assumed to be inactive if that is not the case. The dataset utilized in this study was the version from 1995 which contained 102,540 molecules.

The searches of MDDR in chapters 4-6 were carried out for the 11 classes of active compounds that were first described by Hert *et al.* (2004b) and that were devised in collaboration with Novartis (Novartis, 2012). It has been used in several subsequent studies both at the University of Sheffield and in many other research groups.

Table 3.1 summarizes the 11 selected activity classes with different diversities that belong to MDDR. Each row of the table contains an activity class, the number of active molecules in the class, the number of distinct scaffolds present in the class and the mean pairwise similarity (MPS) values.

In Table 3.1, the numbers of scaffold for each activity class are also presented. The term "Scaffold" is widely used to describe the core structure that is the central component of a molecule, i.e., the substantial substructure that contains the molecular material necessary to ensure that the functional groups are in a desired geometric arrangement

and therefore bioisosteric. An extensively used definition of scaffold in chemoinformatics was given by Bemis and Murcko (1996). Based on their definition, the scaffold of a molecule is a reduced molecular framework, maintaining the original atom typing and bond orders but not side chains. Therefore, scaffolds can be applied in molecular classification or molecular diversity selection (Brown, 2009). The type of scaffold defined in this study is Murcko scaffold, shown as an example in Figure 3.1.

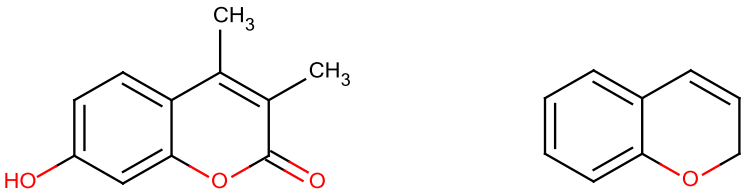


Figure 3.1 A molecule from the ChEMBL database (left) and its corresponding Murcko scaffold (right).

The MPS value in each row was calculated by comparing each member of an activity class with all of the other members of that class, calculating the inter-molecular similarities using the standard UNITY 2D fingerprints (available from Tripos Inc. at http://www.tripos.com) and the Tanimoto coefficient, and then computing the mean intra-set similarity (Gardiner *et al.*, 2009; Hert *et al.*, 2004b). MPS values indicate the diversity of activity classes: it is easier for a similarity search method to retrieve molecules from the activity classes which has a high MPS value and vice versa. These activity classes were used throughout the study (Chapter 4-6).

3.2.1.2 WOMBAT

The WOrld of Molecular BioAcTivity database (WOMBAT) was released by Sunset Molecular and is a leading small molecule chemogenomics database (World of Molecular Bioactivity, 2011) which contains data (structures) extracted from important drug-discovery journals such as the *Journal of Medicinal Chemistry* and *Bioorganic & Medicinal Chemistry*, etc. The bioactivity data for WOMBAT is quantitative, e.g.,

having an IC50 (the half maximal inhibitory concentration). Gardiner *et al.*(2009) converted the activity data to qualitative using the drug potency, expressed as pIC50 (known as the higher the value of the –log IC50), to determine a particular compound as active for a class. They set the threshold of pIC50 at 5.0. For each activity class, molecules with pIC50 >= 5.0 are marked as active for that class, and molecules with pIC50 < 5.0 are removed from that class. The resulting database contained a total of 138,127 molecules reduced from the original version which has 186,117 molecules by removing duplicated molecules and compounds that do not possess the desired effect. The WOMBAT dataset used in this study is described in Gardiner *et al.* (2009).

14 activity classes were selected from this database. They are similar to the 11 activity classes selected from MDDR with several additional activity classes. Table 3.2 presents the 14 selected activity classes derived from WOMBAT with the same structure of Table 3.1. These 14 activity classes are used throughout this study (Chapter 4-6).

Both of the MDDR and WOMBAT databases have been extensively used for virtual screening, as well as in all of the initial experiments reported in the following three experimental chapters, but they do have a limitation. Since the molecules which are considered as active are because of they have been shown to be active against a biological target. Thus, the limitation with these two databases (MDDR and WOMBAT) is that: the molecules which have not been tested against the target were assumed as inactive for the screenings; however, they might in fact be active. For that reason, two following databases were used to attempt to overcome this limitation.

3.2.1.3 MUV

The Maximum Unbiased Validation (MUV) dataset (available by download from http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html, (MUV, 2011)) is rather different in nature from the MDDR and WOMBAT databases, since it has been designed specifically for the evaluation of virtual screening systems, using structure-activity data from the PubChem database. It was designed to overcome the problem of analog bias, i.e., active molecules are too similar to each other, and artificial enrichment,

i.e., actives are too dissimilar from the inactives. The design strategy comprised of three major steps: 1. For a collection of potential actives, compounds with a potential for unspecific bioactivity are removed. 2. Actives devoid of decoys are removed and the rest are required to be well embedded in decoys. 3. Subsets are selected with a spatially random distribution of actives and decoys regarding simple molecular properties, in which a set of 30 actives and 15000 decoys are contained (Rohrer and Baumann, 2009).

The MUV consists of 17 datasets, 30 actives with 15,000 decoys for each dataset. All those datasets have "a non-clumpy, spatially random topology" (Rohrer and Baumann, 2009). The 17 datasets are named as aid466, aid548, aid600, aid644, aid652, aid689, aid692, aid712, aid713, aid733, aid737, aid810, aid832, aid846, aid852, aid858 and aid859, which come from the Assay IDs referring to the bio-assays in PubChem that were used for the assignment of bioactivities.

Table 3.3 shows the activity classes of the MUV database which were applied throughout studies in Chapter 5 and Chapter 6. The last column makes clear the high diversity (i.e., low MPS values) of the MUV data set when compared with the MPS values for the other three databases.

3.2.1.4 ChEMBL

ChEMBL (available by download from https://www.ebi.ac.uk/chembldb) is relatively new, and is one of the largest publicly available databases of curated compound activity data chosen from medicinal chemistry sources. It is an Open Data database containing binding, functional and ADMET information for a large number of drug-like bioactive compounds. It contains 2D structures, calculated properties (e.g., logP, Molecular Weight, Lipinski Parameters, etc.) and abstracted bioactivities (e.g., binding constants, pharmacology and ADMET data). These data are manually abstracted from the main published literature on a regular basis, then further extracted and standardized to maximize their value and utility across a wide range of chemical biology and drug-discovery research problems. Currently, the database contains more than 1 million compounds and 5200 protein targets (Gaulton *et al.*, 2012).

Heikamp and Bajorath (2011) provided some directions for 2D similarity search on publicly available activity classes. In their study, 266 activity classes were extracted from ChEMBL (Version 9) and all compounds were represented as MACCS structural keys and ECFP_4 fingerprints. For each class, a number of reference compounds were randomly selected and the similarity values were calculated by comparing those reference compounds with the rest of the compounds in the database. Based on their similarity values, compound recovery rates (RRs) were generated to show the ratio of active compounds obtained in each activity class. Eventually, 50 activity classes were identified based on the selection criteria, which states that the minimum compound recall yielded for MACCS is more than 30% while the maximum compound recall obtained for ECFP_4 is less than 80%. The difference between the relative search results is more than 20%, i.e., $|(\text{recall obtained for ECFP\_4}) - (\text{recall yielded for MACCS})| > 20\%$, thus reflecting the overall performance range (Heikamp and Bajorath, 2011).

The ChEMBL database used here is version 9 which contains 657,733 molecules, and searches were carried out for 50 classes of 11,561 active compounds in total. The 50 activity classes are presented in Table 3.4. The MPS values presented in the last column were calculated by using UNITY 2D fingerprints and the Tanimoto coefficient. The ChEMBL database was employed throughout studies in Chapter 5 and Chapter 6.

Table 3.4 shows 50 activity classes from ChEMBL. The first column contains the ids of activity classes and the structure of the rest of the table is similar to Table 3.1 to Table 3.3.

Table 3.1 MDDR 11 selected activity classes

Activity class (with abbreviations)	Number of Actives	Number of Scaffolds	MPS
5HT3 antagonists (5HT3)	752	417	0.35
5HT1A agonists (5HT1A)	827	450	0.34
5HT Reuptake inhibitors (5HT)	359	181	0.35
D2 antagonists (D2)	395	258	0.35
Renin inhibitors (Renin)	1130	554	0.57
Angiotensin II AT1 antagonists (AT1)	943	464	0.40
Thrombin inhibitors (Thrombin)	803	425	0.42
Substance P antagonists (SubP)	1246	586	0.40
HIV protease inhibitors (HIVP)	750	461	0.45
Cyclooxygenase inhibitors (COX)	636	282	0.27
Protein kinase C inhibitors (PKC)	453	171	0.32

Table 3.2 WOMBAT 14 selected activity classes

Activity class (with abbreviations)	Number of Actives	Number of Scaffolds	MPS
Renin inhibitors (RENIN)	474	253	0.59
Protein kinase C inhibitors (PKC)	142	31	0.57
Matrix metalloprotease inhibitors (MMP1)	694	280	0.44
Angiotensin II AT1 antagonists (ANG)	724	253	0.44
HIV protease inhibitors (HIVP)	1128	473	0.44
Substance P antagonists (SUBP)	558	186	0.43
Thrombin inhibitors (THR)	421	196	0.42
5HT1A antagonists (5HT1A)	592	224	0.40
Factor Xa inhibitors (Fxa)	842	328	0.39
5HT3 antagonists (5HT3)	220	117	0.38
Acetylcholine esterase inhibitors (AChE)	503	220	0.37
D2 antagonists (D2)	910	324	0.37
Phosphodiesterase inhibitors (PDE4)	596	270	0.36
Cyclooxygenase inhibitors (COX)	965	220	0.32

Table 3.3 MUV 17 activity classes

Activity class (with aid key)	Number of Actives	Number of Scaffolds	MPS
Sphingosine-1-phosphate 1 receptor potentiators (aid 466)	30	28	0.29
Protein kinase A inhibitors (aid 548)	30	27	0.29
Steroidogenic factor 1 inhibitors (aid 600)	30	24	0.29
Rho kinase 2 inhibitors (aid 644)	30	27	0.27
HIV reverse transcriptase RNase (aid 652)	30	27	0.26
Ephrin type-A receptor 4 antagonist inhibitors (aid 689)	30	29	0.27
Steroidogenic factor 1 activators (aid 692)	30	30	0.25
Heat shock protein 90kDa alpha inhibitors (aid 712)	30	27	0.26
Estrogen receptor-alpha coactivator binding inhibitors (aid 713)	30	26	0.26
Estrogen receptor-beta coactivator binding inhibitors (aid 733)	30	28	0.27
Estrogen receptor-alpha coactivator binding potentiators (aid 737)	30	28	0.30
Focal adhesion kinase inhibitors (aid 810)	30	28	0.28
Cathepsin G (aid 832)	30	24	0.32
Factor XIa (aid 846)	30	21	0.28
Factor XIIa (aid 852)	30	24	0.30
Dopamine receptor D1 allosteric modulators (aid 858)	30	24	0.25
Muscarinic receptor M1 allosteric modulators (aid 859)	30	29	0.28

Table 3.4 ChEMBL 50 activity classes

ChEMBL tid	Activity class	Number of Actives	Number of Scaffolds	MPS
Target_no_4	Phosphodiesterase 4D	152	60	0.43
Target_no_8	Thymidylate synthase	103	44	0.38
Target_no_9	Ghrelin receptor	493	228	0.42
Target_no_10	Tyrosine-protein kinase ABL	170	64	0.43
Target_no_12	Tyrosine-protein kinase SRC	442	229	0.40
Target_no_13	Tyrosine-protein kinase receptor FLT3	122	49	0.38
Target_no_14	Serine/threonine-protein kinase Aurora-A	124	66	0.47
Target_no_16	Insulin-like growth factor I receptor	303	124	0.46
Target_no_21	C-Jun N-terminal kinase 1	208	51	0.42
Target_no_35	Carbonic anhydrase XII	119	60	0.39
Target_no_42	Glucocorticoid receptor	485	169	0.37
Target_no_44	Progesterone receptor	330	99	0.39
Target_no_52	Beta-2 adrenergic receptor	150	88	0.47
Target_no_54	Muscarinic acetylcholine receptor M3	252	140	0.40
Target_no_57	Dopamine D3 receptor	388	214	0.39
Target_no_59	Serotonin 1d (5-HT1d) receptor	67	45	0.45
Target_no_81	Neuropeptide Y receptor type 5	367	182	0.38
Target_no_86	G protein-coupled receptor 44	427	132	0.39
Target_no_95	Cyclooxygenase-2	349	117	0.33
Target_no_98	Renin	550	183	0.47
Target_no_105	Beta-secretase 1	536	246	0.45
Target_no_112	Glycine transporter 1	174	66	0.41
Target_no_113	Vasopressin V1a receptor	188	110	0.47
Target_no_115	Oxytocin receptor	161	55	0.42
Target_no_120	Somatostatin receptor 5	130	67	0.46
Target_no_121	Neuropeptide Y receptor type 1	174	66	0.40
Target_no_129	C5a anaphylatoxin chemotactic receptor	170	67	0.46
Target_no_140	C-C chemokine receptor type 4	142	87	0.40
Target_no_142	C-C chemokine receptor type 2	605	178	0.46
Target_no_143	Sodium channel protein type IX alpha subunit	200	58	0.42
Target_no_146	Leukotriene A4 hydrolase	160	87	0.45
Target_no_147	Phosphodiesterase 4A	73	38	0.39
Target_no_148	Cathepsin S	625	298	0.41
Target_no_152	Voltage-gated potassium channel subunit Kv1.5	201	97	0.42
Target_no_163	Cathepsin L	161	67	0.37
Target_no_168	Cytochrome P450 2C9	50	31	0.37
Target_no_171	Orexin receptor 2	100	43	0.45

Target_no_181	Nicotinic acid receptor 1	170	80	0.38
Target_no_186	Serine/threonine-protein kinase B-raf	144	73	0.48
Target_no_195	Cathepsin B	105	56	0.44
Target_no_196	P2X purinoceptor 7	137	69	0.44
Target_no_210	Inhibitor of nuclear factor kappa B Kinase beta subunit	103	46	0.42
Target_no_211	Interleukin-8 receptor B	274	76	0.39
Target_no_213	Sphingosine 1-phosphate receptor Edg-1	133	51	0.37
Target_no_220	Urotensin II receptor	120	74	0.53
Target_no_230	Melatonin receptor 1B	166	52	0.48
Target_no_234	Liver glycogen phosphorylase	347	104	0.48
Target_no_238	Metabotropic glutamate receptor 1	188	84	0.42
Target_no_241	Estradiol 17-beta-dehydrogenase 3	106	39	0.36
Target_no_250	Macrophage colony stimulating factor receptor	117	59	0.44

3.2.1.5 Comparison of Databases' Diversity

As noted in the previous sections, the MPS values signify the diversity of activity classes, which is important for evaluating similarity search approaches. Therefore, the four databases used in this study are compared based on their MPS values.

As shown in Figure 3.2, the MUV dataset is obviously the most diverse one with MPS value in the range from 0.25 to 0.32. The median MPS values of MUV is the lowest at 0.28 compared with >0.35 for the other three databases.

Observation from the other three databases indicates that, MDDR and WOMBAT have outlier(s), i.e., the MPS values of some activity classes are far more than of the others, e.g., Renin from MDDR (0.57), Renin from WOMBAT (0.59) and PKC from WOMBAT (0.57). The three outliers here indicate the most homogeneous classes. The most heterogeneous activity class in MDDR is COX with the lowest MPS value 0.27, which is also the most diverse class in WOMBAT with a MPS value of 0.32. The MPS values of ChEMBL are in the range from 0.33 to 0.53, with no outliners. Generally, WOMBAT and ChEMBL are similar in terms of their median MPS values, minimum MPS values and 3rd quartiles values. MDDR is more diverse than WOMBAT and

ChEMBL. This is shown diagrammatically in Figure 3.2 and suggests that searching MUV activity classes will be much more difficult than for the other three databases.

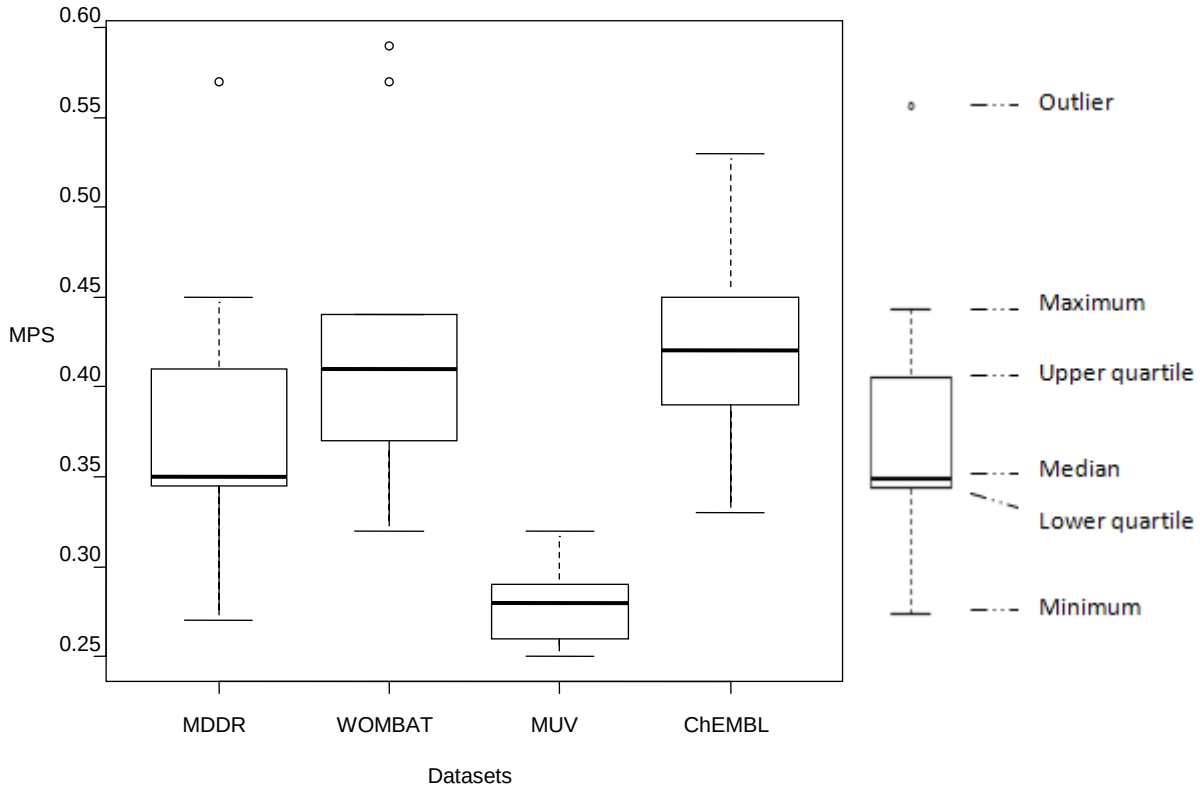


Figure 3.2 Comparison of MPS values of four databases using boxplot.

The construction of a box is shown on the right hand side of this figure. The dark thick segment in the box represents the median MPS value; the lower quartile and the upper quartile are equal to the first and third quartiles MPS values. The outlier is shown as an empty circle.

3.2.2 Molecular Representation

As noted in Chapter 2, in 2D similarity searching, the approach of comparing fingerprints is based on the assumption that the similarity of two fingerprints also indicates a similarity between two molecules in terms of their structures and activities. An appropriate fingerprint, therefore, is one of the most important components in similarity search.

Hert *et al.* (2004b) provided an in-depth analysis and comparison of different fingerprints. A total of 15 fingerprint types were evaluated and classified into four categories: structural keys, hashed fingerprints, circular substructures and pharmacophores. Based on their study, circular substructures have been demonstrated to be the most effective molecular descriptors in similarity searching. A typical type of circular substructure-based descriptors, the Extended Connectivity Fingerprints (ECFPs), have been given a detailed description of their generation in comparison to other fingerprints by Rogers *et al.* (2010).

The ECFPs are not based on pre-defined substructural keys and are designed to capture molecular features related to molecular activity. They were first released by Pipeline Pilot (Accelrys Software, 2009) and have been widely utilized since then. ECFPs are recognized as novel class of topological fingerprints for molecular description. They are popular in ligand-based virtual screening and have obtained very good performances in chemoinformatics (Rogers and Hahn, 2010). The generation process of ECFPs systematically records the neighborhood of each non-hydrogen atom into multiple circular layers up to a given diameter. These atom-centered substructural features are then mapped into integer codes using a hashing procedure.

The ECFPs generation process starts by assigning an integer to each nonhydrogen atom of a molecule. The value of the integer relates to the atom properties, e.g., atomic number and connection count. A number of iterations can then take place to combine the initial atom identifier(s) with the identifier(s) of the neighbor atom(s) in a given diameter. All of the iterations can encode a list which consists of integer(s) that are calculated by a suitable hashing function. Figure 3.3 illustrates the process of generating extended connectivity descriptors (Morgan, 1965; Rogers and Hahn, 2010).

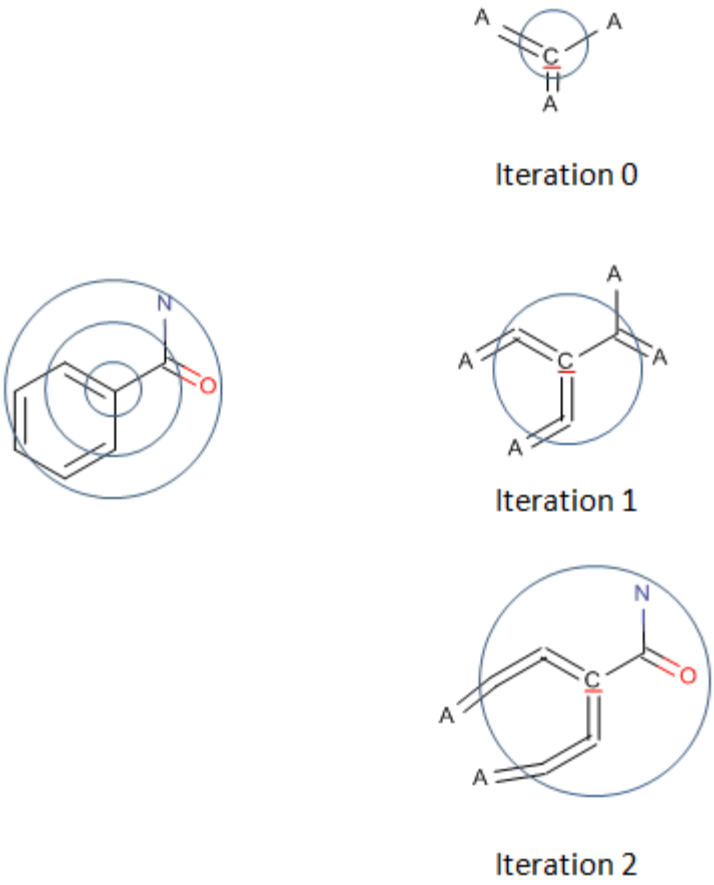


Figure 3.3 Schematic representation for generating extended connectivity descriptors (Rogers and Hahn, 2010).

The ECFPs used in this study refer to Pipeline Pilot's Extended-Connectivity Fingerprints. They encode the central atom and the neighboring atoms within a diameter of 2 (i.e., ECFP(C)_2), 4 (i.e., ECFP(C)_4) or 6 (i.e., ECFP(C)_6) atoms. Where, the last letter "C" in ECFC refers to count. Hence, the ECFP consists of binary strings while the ECFC consists of counts (numbers). They characterize a much bigger set of features than is common for other fingerprints that may be valuable for molecular comparison, e.g., a typical molecule may generate fingerprints containing hundreds or thousands of features; a typical molecular catalog may contain several thousand or millions of different features (Accelrys Software, 2009).

In this study, the molecules were characterized by ECFP_4 and ECFC_4 extended connectivity fingerprints, generated using Pipeline Pilot software (Accelrys Software, 2009). The ECFP_4 and ECFC_4 integer codes representing a circular substructure were hashed to give a fixed-length fingerprint containing 1024 elements, in which the ECFP_4 were applied to experiments in Chapter 4 and Chapter 5 and the ECFC_4 were applied to experiments in Chapter 4 and Chapter 6.

3.3 Procedure of Similarity Search

In 2D similarity search, a fingerprint can be considered as a vector with every element denoting a fragment occurring any number of times in a molecule. The common approach for similarity search can be described briefly as: matching a known bioactive molecule (often called the reference structure) against each of the structures in a database, computing the degree of similarity in each case, and then ranking the database structures in order of decreasing similarity. As noted in Chapter 2, the similar property principle indicates that molecules that are structurally similar have similar properties. Thus, the top-ranked structures from a similarity search are most likely to exhibit the required bioactivity (Sheridan, 2007; Stumpfe and Bajorath, 2011; Willett, 2009).

The workflow of similarity search experiments in this work is shown as Figure 3.4, that：

Step 1: For each database, ten active molecules are selected from each activity class as the reference structures. In the studies on the MDDR and WOMBAT databases, each group of ten active molecules was selected by the MaxMin method. The MaxMin algorithm was reported as suitable for dissimilarity-based compound selection and also for large database processing (Holliday *et al.*, 1995; Snarey *et al.*, 1997). The algorithm starts with a subset containing a single randomly selected structure. This structure can then be used to compare against the rest of structures from the database, and the most dissimilar structure (the one

which has the smallest value of similarity) will be added into the subset. In that case, the two structures in the subset can be used to compare the rest of structures from the database; the most dissimilar structure will be identified and added into the subset, and so on until ten reference molecules are selected. Due to the high diversity of the MUV database and the large scale of ChEMBL database, the reference molecules of these two databases were selected randomly.

Step 2: A similarity coefficient is selected, and, the number of active molecules that occurred in the top 1% of the database when ranked in order of decreasing similarity to each reference molecule is calculated.

Step 3: For each reference molecule, there is a result of a number of active molecules from the same active class as the reference molecule.

Step 4: The mean active number achieved for the certain class by averaging over the ten results.

After selecting reference molecules, different weighting schemes were applied in some of the experiments to weight both the reference molecules and the whole database. In the study in Chapter 4, five weighting schemes were employed which resulted in a total of 25 weighting combinations. In the study in Chapter 6, two kind of weighting combinations were adopted.

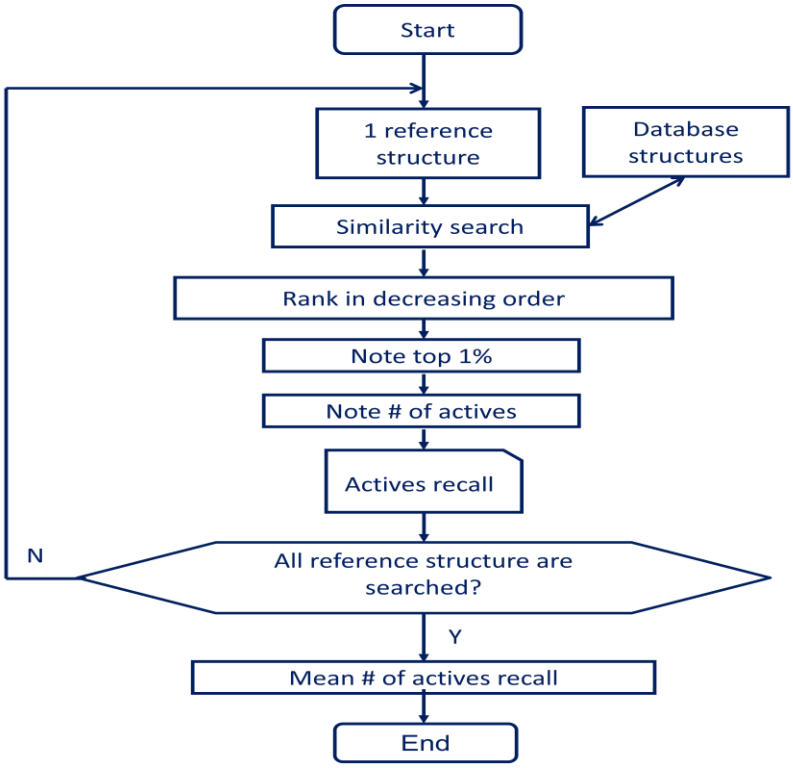


Figure 3.4 Procedure of Similarity Search.

3.4 Evaluation Method

Evaluating similarity search methods focus on two aspects: the efficiency and the effectiveness (Edgar *et al.*, 2000). The efficiency refers to the computational requirements for searching. The effectiveness, on the other hand, measures if the actual output meets the desired output. There are a number of studies interested in criteria for the evaluation of virtual screening experiments (Edgar *et al.*, 2000; Jain and Nicholls, 2008; Truchon and Bayly, 2007). The studies in this thesis are concerned with criteria for measuring the effectiveness of similarity search, i.e., the number of the active molecules which have been retrieved at a cut-off threshold in the ranking. For example, setting the cut-off point in the rank at 5%, if 30% of active molecules have been retrieved, then it indicates a six-fold enrichment of the output compared with a random screening of the database. Amongst the thresholds of cut-off, top 1% is simple to compute and to understand and widely used. Hence, in this study, the effectiveness of a

similarity search for a reference structure can be determined by the number of structures from the same active class as the reference structure contained within the top 1% of the ranking (see Section 3.3).

In the studies reported in this thesis, either weighting combination or similarity coefficients are involved in similarity search approaches. Based on the large amount of experimental results that were obtained, statistical tests are used for determining if differences in performance between similarity search methods are significant, i.e., if the observed differences are meaningful or simply due to chance (Hull, 1993). The most common significance tests are known as parametric tests which were pointed out as being not suitable for discrete measures (Van Rijsbergen, 1979). Therefore, non-parametric tests that do not require stringent distributional assumptions are valid for evaluating similarity search approaches.

Here, Kendall's *W* test and Wilcoxon Signed Rank test were selected to evaluate the statistical significance of the experimental results.

Kendall's *W* test (also known as Kendall's coefficient of concordance) is a non-parametric statistic. It makes no assumptions regarding the nature of the probability distribution and can cope with any number of distinct outcomes. The value of *W* indicates the degree of unanimity among a group of judges, such as the various different activity classes associated with one of the four databases described in Section 3.2.1. For example, the calculated *W* value can help us to identify the performance of different weighting schemes in Chapter 4, and the performance of different similarity coefficients in Chapter 5 and Chapter 6.

The Wilcoxon signed-rank test is also a non-parametric test of statistical significance that can be used to compare two related samples when the data is not known to satisfy the assumptions inherent in the more common *t*-test (interval or ratio scale data with a normal distribution) (McDonald, 2009; Siegel and Castellan, 1988). There are two nominal variables and one measurement variable, one of the nominal variables has only

two values, such as "before" and "after," and the other nominal variable often represents individuals. For instance, in Chapter 4 of this study, a pair of coefficients can be considered as one of the nominal variables and the weighting schemes can be regarded as another. The Wilcoxon signed-rank test can give the level of significance of the difference between the screening effectiveness of two similarity coefficients.

3.4.1 The Kendall W Test of Concordance

The Kendall *W* test of concordance is a measure of the agreement among several (*k*) judges that are assessing a given set of *N* objects (Siegel and Castellan, 1988). For *k* judges, a set of rankings, in which each ranking has N objects are generated. The null hypothesis of Kendall's test is: the *k* judges produced independent rankings of the objects. Thus, the higher the calculated W values are, the higher is the concordance of the ranked objects by the different judges.

The concordance level W can be calculated based on the ranks of corresponding objects, shown as **Error! Reference source not found.**.

$$W = \frac{12\sum_{i=1}^{N} R_i^2 - 3k^2N(N+1)^2}{k^2N(N^2-1) - kT}$$

Equation 3.1 Formula of Kendall's *W* statistic test

Where, R_i is the sums of ranks received by the *i*-th object, N is the number of objects, k is the number of judges. T is a correction factor for tied ranks (**Error! Reference source not found.**):

$$T = \sum_{j=1}^{m} \left(t_j^3 - t_j\right)$$

Equation 3.2 Formula of T

Where t_j is the number of tied ranks in each j of m groups of ties. Then the T value can be computed adding over all groups of ties found in all k judges. For the obtained tied ranks, the average of their ranking scores will be re-assigned.

When N >7, χ^2 value has to be obtained from W (**Error! Reference source not found.**). This quantity is asymptotically distributed like chi-square with $(N-1)$ degrees of freedom. It is then used to derive the corresponding probability value and to test W for statistical significance.

$$\chi^2 = k(N-1)W$$

Equation 3.3 Formula of χ^2

The Kendall W test in this study is used to determine whether or not the agreement occurred on ranking different methods, e.g., 25 combinations of weighting schemes (Chapter 4), or the 44 different coefficients (Chapter 5). The typical significance level (p value) of 0.05 was chosen as a threshold which means that a set of rankings is related and has not just occurred by chance if the p value is less than 0.05. If a significant value is achieved in the Kendall's W test then Siegel and Castellan (1988) suggest that one can obtain a ranking of the N objects. For instance, in Chapter 4, combinations of weighting schemes can be compared using their mean rank value after averaged over their ranks of all activity classes, if a significant W value was obtained.

3.4.2 The Wilcoxon Signed-rank Test

Another non-parametric between pairs test, Wilcoxon signed-rank test (Wilcoxon, 1946), has been used to test the significance of the differences between pairs of the coefficients. Wilcoxon indicated the possibility of using ranking methods so as to attain a quick rough plan of the significance of the differences in experiments.

The Wilcoxon signed- rank test is based on the signed differences which are obtained in each pair of observations. Then a ranking can be calculated by the absolute values of the differences. The absolute value of the differences between observations are ranked from smallest to biggest, with the smallest difference getting a rank of 1, then next larger difference getting a rank of 2, etc. Ties are assigned the mean rank value. The rank values of all positive and negative differences are summed separately. For each pair i, d_i is used to denote the difference. According to the values of d_i, two statistics $T^+$and T^- can be produced, in which T^+ is the sum of the ranks of positive d_i and T^- indicates the sum of the ranks of negative d_i. If the difference between $T^+$and T^- is too small then it is more possible that there is no statistical difference between methods.

The Wilcoxon signed-rank test has been used in this study to test if there is any evidence that one coefficient performs better than another. In this case, the smaller of these two sums $T^+$and T^- is the test statistic, which equals to min(T^+, T^-), e.g., the number of times that the Tanimoto coefficient gives a markedly better result than the cosine coefficient is about the same as the number of times that the converse applies.

In this study, the Wilcoxon signed-rank tests were carried out on the Tanimoto-cosine comparison, Tanimoto-MinMax comparison and the cosine-MinMax comparison in Chapter 4. As for the Kendall *W* test the significance levels for the statistics were measured at the thresholds of 5%.

3.5 Clustering Method

According to the number and range of coefficients tested in Chapter 5 and Chapter 6, hierarchical cluster analysis was adopted to classify the coefficients based on their retrieval abilities.

In chemoinformatics, cluster analysis has been widely used to partition a set of structures into clusters (Varin *et al.*, 2009). The structures in each cluster can hence exhibit high degree of both intra-cluster similarity and inter-cluster dissimilarity (Downs

and Barnard, 2002; Raymond *et al.*, 2003; Willett, 1987). As one of the more practical methods of cluster analysis, the hierarchical cluster analysis is the main method for finding relatively homogeneous clusters based on measured characteristics. Basically, there are two types of algorithms for hierarchical clustering: agglomerative hierarchical clustering and divisive hierarchical clustering. The process of agglomerative hierarchical clustering can be described as: at first, each object is considered a separate cluster; two most similar clusters are then combined sequentially until only one cluster is left. Divisive hierarchical clustering is an inverse procedure of agglomerative hierarchical clustering, i.e., it starts from a single cluster and finally divides it into object number of clusters.

Hierarchical cluster analysis produces a dendogram (tree) to show the hierarchy of the clusters. The clustering method uses the similarities/dissimilarities or distances between objects when forming the clusters. In cluster analysis, the clustering algorithm is crucial as the rule that measures distances between objects and determines cluster membership. In this study, Ward's method was selected as the most appropriate clustering algorithm. It is an agglomerative clustering method proposed by Ward (Ward, 1963). Distinct from other hierarchical methods, it uses an analysis of variance approach to evaluate the distances between clusters and it is less susceptible to noise and outliers. This algorithm is particularly useful and has been widely used in chemoinformatics (Bocker *et al.*, 2005; Downs *et al.*, 1994; Schuffenhauer *et al.*, 2007; Varin *et al.*, 2009).

The Ward's method is based on the error sum of squares within a cluster. Let x_{li}^{k} denotes the value of the *k*th element of *i*th examples in a cluster *l* and n_l denotes the number of examples in *l*. Therefore, the error sum of square of cluster *l* can be defined as:

$$S_l = \sum_{i=1}^{n_l} \sum_{k=1}^{p} (x_{li}^{k} - \overline{x}_l^{k})^2$$

Equation 3.4

Where $\bar{x}_l^k$ is:

$$\bar{x}_l^k = \frac{1}{n_l}\sum_{i=1}^{n_l} x_{li}^k$$

Equation 3.5

For another cluster *m*, the error sum of square can be obtained by:

$$S_m = \sum_{i=1}^{n_m}\sum_{k=1}^{p}(x_{mi}^k - \bar{x}_m^k)^2$$

Equation 3.6

Then the two clusters *l* and *m* can be merged as a new cluster *lm* when ΔS_{lm} is minimum with respect to all the clusters. Where, ΔS_{lm} can be given by:

$$\Delta S_{lm} = \frac{n_l n_m}{n_l + n_m}\sum_{k=1}^{p}(\bar{x}_l^k - \bar{x}_m^k)^2$$

Equation 3.7

and for cluster *lm*, S_{lm} is defined as:

$$S_{lm} = S_l + S_m + \Delta S_{lm}$$

Equation 3.8

This process is repeated until all of the initial clusters are merged into a single cluster. In this study, the hierarchical structures of clusters are visualized as dendrograms with heatmaps, see details in Chapter 5 and Chapter 6.

3.6 Summary

This chapter represents the methods that are involved in the studies reported in this thesis. It contains all the databases that have been tested; the experimental design and also the statistics methods that have been adopted to evaluate the experimental results. However, for the next chapters, the experimental details may vary. Thus, specific clarification will be given separately.

Chapter 4: Evaluation of Interactions between Weighting Scheme and Similarity Coefficient in Similarity-based Virtual Screening

4.1 Introduction

In this chapter, there is an analytical elaboration of the interactions between weighting scheme and similarity coefficient in similarity-based virtual screening. Three similarity coefficients and five weighting schemes were investigated. Through the process of comparing and contrasting their respective features, the central objective was to evaluate and comment on their interactions with each other.

Of particular interest to this study are the findings of Arif *et al.* (2009b). Arif *et al.* discussed the interactions between structural representation, weighting scheme and similarity coefficient when a chemical similarity measure is produced. In their study, five weighting schemes were applied to fragment occurrence data to identify their effectiveness in similarity search. One similarity coefficient was selected according to its known success and extensive use for binary similarity search, namely, the Tanimoto coefficient.

Given its widespread usage with binary fingerprints, Arif *et al.* (2009b) used the Tanimoto coefficient in their experiments on frequency-based weighting, but found that problems could arise that were absent when conventional binary fingerprints were being compared. Specifically, they found that even quite small variations in the weighting

scheme could affect the magnitudes of the Tanimoto coefficients that are calculated in a similarity search; most notably they found that if there is a large discrepancy in the weights computed for the reference structure and for the database structure then screening effectiveness is likely to be markedly less than if the two weights are of comparable magnitude. This behavior was ascribed to the precise mathematical form of the Tanimoto coefficient, and it was suggested that other types of coefficient might be less affected by changes in the weighting scheme that was being used. Thus, the study reported in this chapter attempted to determine whether other coefficients may be preferable to the Tanimoto coefficient when frequency-weighted fingerprints are used for similarity-based virtual screening.

Another finding from Arif *et al.* (2009b) is, out of the above five weighting schemes in their research, one was found to be superior to the others, namely, W4, i.e., the square root weighting scheme. Since their primary interest was in the incidence and occurrence representations, and the measures where both the reference structures and the database structures are weighted the same. Thus, they analysed 19 rather than a total of 25 possible weighting combinations which are introduced in next section.

Therefore, the study reported in this chapter focused on determining whether other coefficients maybe superior to the Tanimoto coefficient, and combinations of weighting schemes and coefficients that could enhance similarity searching.

4.2 Method

As described in Chapter 3, ECFC_4 fingerprints were chosen as the molecular descriptors throughout this study. All molecules were characterized by 1,024 fixed-length ECFC_4 fingerprints generated using Pipeline Pilot software (Accelrys Software, 2009). Three publicly available databases were investigated, i.e., the MDDR databases, the WOMBAT databases and the MUV database. The former two databases were utilized in comparison with Arif *et al.*'s (2009b) study and the MUV database was used in this study to mirror the results obtained from the former two databases.

The experimental process is described in Chapter 3. Before completing the similarity search, illustrated by Figure 3.4, selected weighting scheme(s) was(were) applied to weight both the reference molecules and the molecules from a database. Each search involved computing the similarity between one of the ten reference structures chosen from each activity class and all of the database structures, ranking the database structures in decreasing order of the computed similarities, and then checking how many of them in the top-1% of the ranked list belonged to the same activity class as the reference structure. This was repeated for each of the ten chosen structures in an activity class, and then the mean number of actives retrieved was calculated to describe the effectiveness of screening. Searches were carried out for each of the eleven MDDR activity classes, for each of the 14 WOMBAT activity classes and for each of the 17 MUV activity classes, using the combinations of weighting schemes.

Following Arif *et al*. (2009b), five types of weighting schemes were adopted in this study, i.e., W1 to W5. Details of these weighting schemes are given later in this section. In order to identify the possible 25 combinations of weighting schemes, Mab has been introduced (Arif *et al.*, 2009b) where *a* represents the weighting scheme applied to the molecules from database, and *b* represents the weighting scheme applied to a reference molecule. For example, M15 represents the similarity measure using W1 as the weighting scheme for molecules in the database, and using W5 as the weighting scheme for a reference molecule.

The ECFC_4 fingerprint can be considered as a vector, X (where X can denote either the reference structure or a database structure), with the *i*-th element denoting a fragment occurring f_i times in a molecule ($f_i \geq 0$). If $f_i > 0$, i.e., if a fragment occurs at least once in a molecule, then the f_i value may then be weighted. In this experiment, the f_i value was weighted with five different weights (denoted here by W1- W5).

W1: the incidence weight, in which the element is set to one, i.e., as with a conventional binary fingerprint. The value of non-zero *i*th element:

W1: $x(i) = 1$

W2: the occurrence weight, i.e., using the raw occurrence counts. In this study, ECFC_4 fingerprints were used as molecular descriptors, in which the elements are '0's, or counts, that stand for a certain feature (fragment)'s absence or frequency of occurrence, respectively.

W2: $x(i) = f_i$

These two weighting schemes, W1 and W2, are the obvious weights, and the ones that are normally meant when binary and weighed fingerprints are referred to in the chemoinformatics literature. Following Arif *et al*., three further weights were included in this study, of which the first two are standard normalizations in data analysis, and the final weight is a normalization which has been widely used to weight terms in text search engines (Salton and Buckley, 1988).

W3: the natural logarithm. Due to the fact that many fragments in a molecule occur only once, the natural logarithm was applied to the non-zero elements in order to emphasize the high-frequency occurrence elements. It is expected to result in very sparse fingerprints containing much smaller numbers of non-zero elements than for the other weights. The value of non-zero *i*th element:

W3: $x(i) = ln(f_i)$

W4: the square root. In order to reduce the contribution of high-frequency occurrence in a molecule, square root was used on the non-zero elements. Then the value of the non-zero elements:

W4: $x(i) = \sqrt{f_i}$

W5: In chemoinformatics, it has been proven by studies that the method which works well in Information Retrieval (IR) can be applied to retrieve desired molecules from chemical databases effectively, e.g., scoring, term weighting and vector space model etc. (Willett, 2000; Willett, 2009). By taking molecular size into account, the weighting scheme below was used to normalize the non-zero *i*-th element's value to generate a number in the range of [0.5, 1]. This scheme is derived from the same formula which has been used as an effective method in automatic text retrieval (Boyce, 1990; Salton, 1986; Salton and Buckley, 1988). It expresses the raw occurrence frequency as a fraction of the frequency of the most frequently occurring fragment in the molecule. Here, $max\{f_i\}$ stands for the largest value on *i-t*h element of fingerprints in the whole molecule.

W5: $x(i) = 0.5 + 0.5\frac{f_i}{max\{f_i\}}$

Table 4.1 provides quantitative data for the coding of the three datasets using the weights W1-W5. Many fragments in a molecule will occur only once and hence the use of the logarithmic W3 weight results in sparse fingerprints containing much smaller numbers of non-zero elements than for the other weights.

Table 4.1 Statistical data describing the MDDR, WOMBAT and MUV datasets using ECFC_4 fingerprints.

		MDDR	**WOMBAT**	**MUV**
Number of molecules		102,540	138,127	255,510
Mean non-zero elements per fingerprint		52.43	50.32	44.46
Mean non-zero elements (W3) per fingerprint		15.15	15.21	12.60
Mean value of the non-zero elements	*W1*	1.00	1.00	1.00
	W2	1.70	1.76	1.56
	W3	1.07	1.08	1.02
	W4	1.22	1.24	1.19
	W5	0.61	0.61	0.61

As denoted before, structures can be represented as vectors. Each structure therefore has its own axis in a vector space. The similarity of any two structures can be calculated based on the Euclidean distance or/and the angle formed between these two structures. When two structures are similar, they will be relatively close in space and the angle formed between them will be relatively small. If a similarity metric is based on the angle, theoretically, it is less affected by the way of the employed weighting schemes, e.g., the two structures are weighted the same or not. For example, when calculating the similarity between a structure and its weighted form, the similarity metric based on their angle can provide higher similarity values than the similarity metric based on their distance. Therefore, the two ways of weighting, i.e. the two structures weighted equally and non-equally, both are considered in this study.

In this chapter, three similarity coefficients were used to measure the similarity *S* between the vectors *X* and *Y*, representing a reference structure and a database structure, respectively. If x_i is used to represent the value of a certain element of the reference molecule and y_i for the value of the same element of a database molecule, then the similarity coefficients applied are:

The first of these was the Tanimoto coefficient S_T (Willett *et al.*, 1998), as used in the previous study of occurrence-based weighting schemes. The formula of the Tanimoto coefficient was listed in Table 2.1, where x_{iA} denotes x_i and x_{iB} denotes y_i and the summations are over all of the elements in each fingerprint. In previous study (Arif *et al.*, 2009b), this performs effectively when *X* and *Y* are weighted in the same way, i.e., using M11, M22, M33, M44 or M55 (a situation refers to as a symmetric weighting scheme), but has been found to be capable of giving highly variable levels of screening effectiveness when *X* and *Y* are weighted using different weights (e.g., M14 or M23, a situation referred to as an asymmetric weighting scheme). Their results suggested that a positive bias given to the features in the reference structures or a negative bias given to the features in the database structures tends to improve active analogue search. This indicates the situation that the reference and the database structures weighted non-equally need to be considered in similarity search. Arif *et al.* (2009b) also showed that

this variation in performance is due to the second and third components of the denominator of the Tanimoto coefficient. The first component relates to the reference structure and hence makes a constant contribution in a database search, but the relative contributions of the second and third components of the denominator are highly dependent on the precise weights that are used. The experiments in this study have hence used an alternative coefficient that does not involve the third factor in the Tanimoto's denominator. This is the cosine (or Ochiai) coefficient S_C(Willett *et al.*, 1998).

$$S_C = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt[2]{\sum_{i=1}^{n} x_i^2 \times \sum_{i=1}^{n} y_i^2}}$$

Equation 4.1

which is clearly similar in form to the Tanimoto coefficient and which has been widely used in information retrieval (IR) with excellent performance (Korenius *et al.*, 2007). It has also been found to offer comparable levels of performance when binary fingerprints are used (Holliday *et al.*, 1995; Willett, 2006), but not been proved superior to the Tanimoto coefficient in occurrence-weighted similarity searching.

As a further, rather different, the MinMax coefficient S_M was tested. It has been introduced and used in mutagenicity and toxicity prediction by Swamidass *et al.* (2005):

$$S_M = \frac{\sum_{i=1}^{n} min\{x_i, y_i\}}{\sum_{i=1}^{n} max\{x_i, y_i\}}$$

Equation 4.2

which reduces to the Tanimoto coefficient in the case of binary fingerprints.

To evaluate the statistical significance of the results obtained, Kendall's *W* test and Wilcoxon Signed Rank test were used in this study.

As noted in Chapter 4, Kendall's *W* test (Siegel and Castellan, 1988) evaluates the significance of agreement between *k*, different classes, and *N*, objects from each class. For *N*>7,

$$W = \frac{12\sum_{i=1}^{N} R_i^2 - 3N(N+1)^2}{N(N^2-1)}$$

Equation 4.3

χ^2, also known as chi-square value, can be given by Equation 3.3.

Here, R_i is the average of the ranks assigned to the *i*-th object, e.g., M11. In this study, for the MDDR searches, *k*=11 and *N*=25 and for the WOMBAT searches, *k*=14 and *N*=25.

The other statistical test, Wilcoxon signed-rank test (Wilcoxon, 1946), was used to test the significance of the differences between pairs of the three coefficients, i.e., paired results of 25 similarity measures on MDDR and WOMBAT with three coefficients. Thus, the Wilcoxon signed-rank test was carried out on the Tanimoto-cosine comparison, Tanimoto-MinMax comparison and the cosine-MinMax comparison.

4.3 Results

In Arif *et al.* (2009b), 19 weighting combinations were employed and M44 has been found is superior to other similarity measures with the Tanimoto coefficient. Their study was conducted in MDDR and WOMBAT datasets. Hence, the results presented and discussed in this section and the next were from the MDDR and the WOMBAT databases so as to compare with Arif *et al.*'s. The investigation carried out on the MUV database will reported in Section 4.5 as a further validation.

Tables 4.2 to Table 4.4 illustrate the average number of active molecules retrieved in each class using the three coefficients and different similarity measures in the MDDR

databases. The results of the WOMBAT databases are attached to Appendix A, Table A.1 to Table A.3. For each table, the antepenultimate columns show the mean of active molecules averaged over different activity classes and the penultimate columns show the corresponding median values (eleven activity classes in MDDR and fourteen activity classes in WOMBAT). The last columns show the mean of average rank over different activity classes (eleven classes in MDDR and fourteen activity classes in WOMBAT) in descending order. In each column except the last, the dark-shaded cell indicates the largest value; in the last column the dark-shaded cell indicates the smallest value.

The penultimate columns show the median value of active molecules averaged over different activity classes. The previous study (Arif *et al.*, 2009b) compared weighting schemes using the numbers of Mean actives retrieved from activity classes. However, for evaluating data, in some cases, the median value is more resistant to outliers than is the mean, so this study used both the numbers of Mean actives and the numbers of Median actives from activity classes. The numbers of Median actives were calculated to mitigate the effects of some other activity classes (most obviously the Renin and AT1 activity classes) which have far greater numbers of actives than for the other classes.

After averaging the rank results of 11 activity classes in MDDR (5HT3, 5HT1A, 5HT, D2, Renin, AT1, Thrombin, SubP, HIV P, COX and PKC), 14 in WOMBAT (5HT3, 5HT1A, AChE, D2, Renin, PDE4, Thrombin, SubP, HIV P, COX, PKC, ANG, FXa and MMP1) and ten reference structures for each activity class, the mean rank result for each similarity measure is shown on the last column of Table 4.2 to Table 4.4 and Table A.1 to Table A.3. The top rank is dark shaded. Kendall's *W* values and chi-square values are shown in the captions of all the tables.

4.3.1 MDDR Results

From Table 4.2, M54 retrieved the largest mean number of active molecules in MDDR. M44 and M14 were ranked first. M23 performed badly on mean actives, median actives

and mean rank, with the mean active molecules only retrieving 13.48, while the median actives value was 6.60.

The *W* value is 0.58 and the chi-square value yielded is 154.02; both are significant at the 0.001 level of statistical significance. If a significant value is achieved in the Kendall's *W* test then Siegel and Castellan (1988) suggest that one can obtain a ranking of the *N* objects. Thus, the rank of the 25 measures is:

M14>M44>M55>M41>M51>M54>M12>M11>M52>M22>M15>M45>M42>M35>M33>M24>M31>M34>M21>M53>M43>M32>M25>M13>M23

The mean ranks from this analysis are shown in the final column of Table 4.2. They are based on the ranks for each of the classes. Correspondingly, the mean actives result is:

M54>M51>M12>M52>M14>M44>M55>M11>M41>M42>M35>M31>M15>M22>M34>M45>M33>M24>M32>M53>M21>M43>M13>M25>M23

and the median actives result is:

M11>M55>M14>M51>M41>M44>M54>M12>M15>M52>M22>M33>M42>M45>M53>M35>M24>M21>M34>M31>M43>M13>M32>M25>M23

Table 4.2 Average numbers of actives molecules retrieved in the top 1% of searches of the MDDR database using the Tanimoto coefficient.

(W=0.58, chi-square=154.02, p<= 0.001)

Similarity measure	Activity class											Mean actives	Median actives	Mean rank
	5HT3	5HT1A	5HT	D2	Renin	AT1	Thrombin	SubP	HIV P	COX	PKC			
M11	89.60	81.30	24.60	27.20	420.20	235.70	56.60	120.60	86.70	28.20	35.40	109.65	81.30	7.36
M12	86.60	72.10	24.70	28.90	513.60	241.20	52.30	136.50	94.00	23.70	31.90	118.68	72.10	6.91
M13	40.20	54.10	11.90	10.00	50.60	8.30	2.30	79.60	23.30	17.40	20.40	28.92	20.40	21.82
M14	91.70	79.00	23.90	28.80	460.50	238.60	54.60	135.60	88.20	27.50	36.30	114.97	79.00	5.36
M15	89.00	79.00	22.00	24.50	302.30	179.60	42.70	99.10	66.50	29.50	33.00	87.93	66.50	11.18
M21	94.90	66.90	17.80	21.40	169.00	65.10	39.40	33.20	19.10	18.20	13.20	50.75	33.20	16.91
M22	76.20	62.90	23.30	27.60	283.50	180.70	36.10	128.40	57.90	29.90	41.10	86.15	57.90	10.55
M23	45.80	41.30	6.60	4.80	9.50	0.40	5.00	13.90	3.30	11.80	5.90	13.48	6.60	23.45
M24	98.60	71.00	21.50	24.60	196.00	91.40	35.50	76.40	28.30	27.30	19.10	62.70	35.50	14.82
M25	82.80	54.70	13.80	14.20	43.60	14.30	17.10	7.90	4.80	14.70	7.40	25.03	14.30	21.00
M31	15.50	45.40	5.80	16.40	471.50	125.90	15.20	138.10	96.80	8.20	30.60	88.13	30.60	16.00
M32	5.00	17.90	2.00	6.90	255.90	86.30	11.40	122.00	70.20	5.60	22.30	55.05	17.90	20.64
M33	73.50	61.10	16.10	26.90	193.70	106.00	30.50	131.50	57.30	28.20	35.10	69.08	57.30	14.18
M34	13.40	36.50	3.70	14.90	417.20	136.10	15.70	138.70	106.50	9.20	32.70	84.05	32.70	16.36
M35	39.20	62.10	13.80	27.40	468.20	175.10	40.60	129.80	57.00	20.20	35.20	97.15	40.60	13.82
M41	90.00	77.60	23.70	28.70	418.60	239.50	56.20	124.20	75.40	31.20	36.10	109.20	75.40	6.82
M42	56.50	59.90	17.40	22.00	420.50	202.30	32.10	139.00	84.40	22.30	37.20	99.42	56.50	12.09
M43	40.70	57.10	16.30	13.80	65.80	23.90	7.90	78.40	20.50	27.40	25.00	34.25	25.00	19.64
M44	80.50	74.60	23.10	28.90	449.10	247.20	49.50	143.20	88.90	31.90	43.20	114.55	74.60	5.36
M45	90.30	75.30	22.50	24.10	252.40	169.40	43.20	88.70	48.40	31.90	32.60	79.89	48.40	11.73
M51	84.70	76.10	23.40	27.30	501.70	259.40	61.30	133.40	93.30	24.80	33.50	119.90	76.10	6.82
M52	81.40	66.10	21.20	27.70	525.00	246.90	51.20	130.50	66.60	21.00	32.10	115.43	66.10	10.45
M53	45.50	63.00	15.30	20.70	155.00	50.20	8.20	112.80	44.90	23.00	34.00	52.05	44.90	17.64
M54	86.20	73.50	22.10	28.20	535.50	259.00	58.40	136.90	85.90	25.30	33.90	122.26	73.50	6.82
M55	90.00	80.20	23.30	27.60	446.90	240.30	53.60	127.90	84.30	29.30	38.90	112.94	80.20	6.64

Table 4.3 Average numbers of active molecules retrieved in the top 1% of searches of the MDDR database using the cosine coefficient.

(W=0.57, chi-square=150.96, p<= 0.001)

Similarity measure	Activity class											Mean actives	Median actives	Mean rank
	5HT3	5HT1A	5HT	D2	Renin	AT1	Thrombin	SubP	HIV P	COX	PKC			
M11	88.90	80.50	23.90	26.50	425.40	233.30	55.40	120.40	88.80	27.60	34.60	109.57	80.50	8.73
M12	86.90	73.70	21.20	22.00	310.40	123.00	33.20	98.80	51.90	27.40	31.70	80.02	51.90	16.91
M13	47.90	60.70	14.10	14.70	111.80	19.80	4.30	91.90	27.30	19.20	24.30	39.64	24.30	23.36
M14	90.00	80.00	23.00	26.40	401.50	209.10	48.00	121.80	79.20	28.10	37.60	104.06	79.20	9.82
M15	90.40	80.30	23.50	26.90	418.10	227.40	53.30	120.90	84.80	27.90	36.10	108.15	80.30	8.82
M21	76.90	72.80	21.00	29.40	500.30	224.50	55.50	133.80	85.70	21.50	30.00	113.76	72.80	10.18
M22	57.30	58.70	18.20	26.40	340.20	207.80	30.90	122.60	62.80	21.70	40.40	89.73	57.30	15.82
M23	38.10	55.00	12.10	23.50	272.10	160.00	24.30	130.40	63.20	19.50	38.30	76.05	38.30	19.00
M24	73.00	68.90	21.50	30.40	469.70	253.10	46.60	139.30	92.00	24.20	42.10	114.62	68.90	8.00
M25	76.50	71.00	21.80	29.80	496.30	250.50	51.20	140.50	93.40	24.10	39.10	117.65	71.00	7.55
M31	32.80	58.90	9.80	22.70	478.20	144.30	31.70	133.30	82.90	13.10	28.60	94.21	32.80	18.64
M32	66.70	57.50	17.80	24.10	285.90	113.00	25.50	102.80	35.50	28.30	32.00	71.74	35.50	19.00
M33	66.70	59.90	16.90	26.10	237.40	108.30	29.10	120.10	43.80	27.00	32.20	69.77	43.80	18.73
M34	51.10	63.30	14.80	27.00	449.00	188.10	36.90	135.30	69.00	20.70	36.10	99.21	51.10	14.27
M35	43.60	61.60	12.60	25.50	479.90	174.30	34.30	137.40	83.20	17.20	34.80	100.40	43.60	15.55
M41	81.30	76.20	22.60	27.00	481.50	258.70	55.10	137.10	100.00	28.00	37.90	118.67	76.20	5.82
M42	66.90	67.10	21.00	25.60	359.60	203.60	35.00	119.30	63.80	26.60	40.60	93.55	63.80	14.36
M43	37.70	58.20	13.10	22.20	237.30	100.40	22.10	122.00	57.20	22.70	36.80	66.34	37.70	20.45
M44	76.90	73.30	21.80	28.00	464.30	252.10	47.40	138.90	93.80	30.00	42.90	115.40	73.30	6.27
M45	79.90	75.20	22.30	27.90	478.50	258.20	52.10	140.90	99.30	29.50	40.80	118.60	75.20	5.09
M51	87.80	80.30	23.50	26.90	448.90	243.10	55.70	126.20	89.60	28.60	37.30	113.45	80.30	6.82
M52	81.40	71.80	22.80	23.10	326.30	156.40	33.40	108.30	52.80	30.40	36.90	85.78	52.80	13.91
M53	46.40	60.70	15.40	19.30	158.70	40.80	10.00	102.00	32.70	22.70	31.70	49.13	32.70	21.73
M54	88.40	78.00	22.50	27.70	430.20	227.30	47.90	127.30	79.10	30.10	40.40	108.99	78.00	8.18
M55	89.20	79.10	23.70	27.00	446.60	238.90	53.40	127.40	85.20	29.00	39.20	112.61	79.10	6.55

Using the cosine coefficient on MDDR (see Table 4.3), M45 and M41 achieved a very good performance both on the number of active molecules and the order of rank. The worst measure is M13 which scored 39.64 mean active molecules and 24.30 median actives; however, it is three times the number of actives which M23 gained with the Tanimoto coefficient.

The *W* value is calculated to be 0.57 and the chi-square value is yielded to be 150.96; both are significant at the 0.001 level of statistical significance.

Table 4.4 Average numbers of active molecules retrieved in the top 1% of searches of the MDDR database using the MinMax coefficient.

(W=0.56, chi-square=147.95, p<= 0.001)

Similarity measure	Activity class											Mean actives	Median actives	Mean rank
	5HT3	5HT1A	5HT	D2	Renin	AT1	Thrombin	SubP	HIV P	COX	PKC			
M11	89.60	81.30	24.60	27.20	420.20	235.70	56.60	120.60	86.70	28.20	35.40	109.65	81.30	8.82
M12	88.50	79.30	24.70	27.60	453.40	246.60	58.60	128.00	94.60	27.50	34.20	114.82	79.30	7.45
M13	44.70	59.90	13.70	13.50	62.90	17.20	3.10	88.10	27.70	18.80	23.60	33.93	23.60	21.36
M14	89.60	81.10	24.40	27.30	435.70	240.90	57.50	123.20	89.80	27.80	35.10	112.04	81.10	8.18
M15	90.80	80.40	22.90	26.20	361.30	206.70	47.30	110.40	74.90	29.70	35.50	98.74	74.90	11.27
M21	94.70	72.00	23.60	25.50	286.90	200.80	54.30	76.30	55.50	24.90	22.20	85.15	55.50	14.73
M22	92.50	75.00	25.30	30.00	411.20	235.20	51.30	154.30	80.30	33.20	45.90	112.20	75.00	6.73
M23	41.10	47.00	8.60	5.80	15.30	1.90	2.50	26.10	6.40	11.10	8.00	15.80	8.60	24.00
M24	102.70	75.40	23.40	28.20	336.70	194.80	54.10	107.70	58.00	25.90	26.60	93.95	58.00	12.36
M25	94.80	68.00	21.60	23.20	184.90	153.60	42.50	55.10	41.20	24.00	18.90	66.16	42.50	17.36
M31	24.30	55.30	8.30	19.70	570.60	123.40	21.40	141.00	83.10	10.20	25.90	98.47	25.90	16.82
M32	5.40	30.10	2.40	9.50	332.20	100.10	14.00	124.20	80.90	6.60	22.80	66.20	22.80	20.82
M33	82.30	63.90	17.20	28.10	242.80	112.40	34.20	143.00	60.10	29.60	36.10	77.25	60.10	13.27
M34	12.60	39.30	4.90	15.00	456.70	130.80	16.10	136.50	110.10	9.50	29.10	87.33	29.10	16.82
M35	48.70	68.00	12.00	24.20	519.10	159.20	36.00	136.50	47.00	16.00	26.20	99.35	47.00	15.73
M41	92.60	79.10	24.90	27.10	378.50	229.20	57.70	106.90	73.70	28.60	30.90	102.65	73.70	10.00
M42	78.40	77.50	21.50	26.60	467.00	252.30	52.20	146.50	108.80	30.00	42.80	118.51	77.50	7.64
M43	44.20	57.40	14.40	10.20	48.70	11.90	3.70	65.80	17.20	17.40	20.30	28.29	17.40	22.27
M44	92.40	81.70	24.80	29.20	447.60	248.10	55.70	144.80	89.20	31.30	42.20	117.00	81.70	4.27
M45	94.40	77.20	23.80	25.20	300.10	198.00	49.30	91.50	58.80	29.70	29.30	88.85	58.80	12.73
M51	91.30	78.60	22.70	27.70	470.00	245.90	59.10	126.30	85.90	26.80	34.00	115.30	78.60	8.55
M52	88.40	75.80	22.90	27.10	494.70	255.40	62.20	131.30	88.20	24.40	30.40	118.25	75.80	9.00
M53	51.10	65.90	14.50	20.70	133.80	43.70	6.30	111.10	38.20	22.70	31.20	49.02	38.20	18.82
M54	91.20	77.20	23.30	27.30	481.00	249.50	60.40	128.50	87.40	26.70	32.70	116.84	77.20	7.91
M55	90.40	81.40	24.40	28.10	435.20	243.90	55.60	128.20	86.60	29.70	38.50	112.91	81.40	7.00

On the MDDR database (see Table 4.4), M44 demonstrated a good performance working with the MinMax coefficient. The inferior measure is M23, which scored 15.80

mean active molecules and 8.60 median actives. This result is very similar to that achieved with the Tanimoto coefficient.

The *W* value is calculated to be 0.56 and the chi-square value is yielded to be 147.95, both are significant at the 0.001 level of statistical significance. Notably, the top three similarity measures are symmetric similarity measures.

4.3.2 WOMBAT Results

Table A.1 to Table A.3 illustrate the performance of the three coefficients on the WOMBAT databases.

When the similarity search was performed on the WOMBAT database with the Tanimoto coefficient, M54 achieved the best performance, and then M12. Still, M23 only retrieved 8.24 active molecules and 7.05 of median actives was the worst. The *W* value is 0.71 and the chi-square value yielded is 239.79; both are significant at the 0.001 level of statistical significance.

Very similar to the results obtained from the MDDR database, M41 and M45 were shown to be the best measures in the WOMBAT database with the cosine coefficient. M13 was the worst one but yielded more than four times of active molecules compared to M23 with the Tanimoto coefficient. The *W* value is 0.60 and the chi-square value yielded is 202.83; both are significant at the 0.001 level of statistical significance.

When the similarity search was carried out with the MinMax coefficient on the WOMBAT database, M44 achieved the best performance, and then M22. M23 worked poorly: it only retrieved 11.76 active molecules and the median value of 11.45 actives was the worst. The *W* value 0.73 and the chi-square value yielded is 245.40; both are significant at the 0.001 level of statistical significance.

Generally, as shown in Table 4.5, symmetric measures performed better than asymmetric measures, e.g., M44, M55. This is in line with Arif *et al.*'s finding.

Table 4.5 Rankings of the 25 measures for combinations of database and similarity coefficient.

(based on the results from the last three columns, Table 4.2 to Table 4.4 and Table A.1 to Table A.3)

Database	Coefficient	Ranking type	Ranking result
MDDR	Tanimoto	rank	M14>M44>M55>M41>M51>M54>M12>M11>M52>M22>M15>M45>M42>M35>M33>M24>M31>M34>M21>M53>M43>M32>M25>M13>M23
		mean actives	M54>M51>M12>M52>M14>M44>M55>M11>M41>M42>M35>M31>M15>M22>M34>M45>M33>M24>M32>M53>M21>M43>M13>M25>M23
		median actives	M11>M55>M14>M51>M41>M44>M54>M12>M15>M52>M22>M33>M42>M45>M53>M35>M24>M21>M34>M31>M43>M13>M32>M25>M23
	cosine	rank	M45>M41>M44>M55>M51>M25>M24>M54>M11>M15>M14>M21>M52>M34>M42>M35>M22>M12>M31>M33>M23>M32>M43>M53>M13
		mean actives	M41>M45>M25>M44>M24>M21>M51>M55>M11>M54>M15>M14>M35>M34>M31>M42>>M22>M52>M12>M23>M32>M33>M43>M53>M13
		median actives	M11>M51>M15>M14>M55>M54>M41>M45>M44>M21>M25>M24>M42>M22>M52>M12>M34>M33>M35>M23>M43>M32>M31>M53
	MinMax	rank	M44>M22>M55>M12>M42>M54>M14>M51>M11>M52>M41>M15>M24>M45>M33>M21>M35>M34>M31>M25>M53>M32>M13>M43>M23
		mean actives	M42>M52>M44>M54>M51>M12>M55>M22>M14>M11>M41>M35>M15>M31>M24>M45>M34>M21>M33>M32>M25>M53>M13>M43>M23
		median actives	M44>M55>M11>M14>M12>M51>M42>M54>M52>M22>M15>M41>M33>M45>M24>M21>M35>M25>M53>M34>M31>M13>M32>M43>M23

Database	Coefficient	Ranking type	Ranking result
WOMBAT	Tanimoto	rank	M54>M12>M14>M51>M52>M11>M55>M44>M41>M15>M22>M45>M35>M42>M24>M33>M21>M53>M34>M31>M25>M43>M13>M32>M23
		mean actives	M54>M12>M51>M14>M52>M11>M55>M44>M41>M22>M15>M42>M35>M45>M33>M24>M53>M34>M31>M21>M43>M13>M32>M25>M23
		median actives	M54>M52>M55>M44>M14>M12>M11>M41>M51>M15>M45>M22>M42>M35>M33>M53>M31>M24>M34>M21>M43>M13>M25>M32>M23
	cosine	rank	M41>M45>M51>M55>M25>M11>M44>M24>M54>M15>M14>M21>M42>M52>M22>M34>M12>M35>M23>M33>M43>M32>M53>M31>M13
		mean actives	M41>M45>M51>M25>M55>M11>M24>M44>M15>M54>M21>M14>M42>M22>M34>M52>M35>M23>M12>M33>M43>M31>M32>M53>M13
		median actives	M55>M45>M25>M24>M44>M54>M15>M21>M51>M41>M11>M14>M42>M52>M12>M22>M34>M35>M23>M33>M43>M31>M53>M32>M13
	MinMax	rank	M44>M22>M54>M12>M52>M51>M14>M55>M42>M11>M41>M24>M15>M45>M21>M33>M35>M25>M53>M31>M34>M13>M32>M43>M23
		mean actives	M44>M12>M52>M54>M22>M42>M51>M14>M55>M11>M41>M15>M24>M45>M21>M33>M35>M25>M31>M34>M53>M32>M13>M43>M23
		median actives	M22>M44>M55>M42>M14>M11>M51>M54>M12>M52>M24>M41>M15>M45>M21>M33>M35>M25>M31>M53>M34>M32>M13>M43>M23

4.4 Discussion

Compared with the previous study (Arif *et al.*, 2009b), this study carried out a further investigation using different coefficients. Furthermore, the impact from typical activity classes also was noticed and analyzed.

The general results show that the W4 weighting scheme yields more actives and the W3 weighting scheme performed the worst. The average number of actives retrieved either by the cosine coefficient or by the MinMax coefficient is more than the number achieved using the Tanimoto coefficient. Some particular classes contributed more than other classes in the similarity search, for example, Renin and AT1 in MDDR; Renin, PKC and ANG in WOMBAT. More detailed discussion is given in this section.

4.4.1 Comparison of Coefficients

For each of the three coefficients, the mean and standard deviation values for all of the 25 combined weighting schemes were calculated, based on the results from Table 4.2 - 4.4 and Table A.1 – A.3, which is shown in Table 4.6.

Table 4.6 Screening effectiveness of 25 combined weighting schemes in similarity searches of the MDDR and WOMBAT databases using three similarity coefficients.

		Mean actives			Median actives		
		S_T	S_C	S_M	S_T	S_C	S_M
MDDR	Mean	82.06	94.84	89.55	50.66	59.22	57.00
	S.D.	33.50	22.37	30.33	23.59	18.71	24.21
WOMBAT	Mean	72.81	84.32	80.83	59.60	69.73	66.93
	S.D.	31.79	18.83	29.69	25.76	15.13	23.26

Table 4.6 demonstrates that the Tanimoto coefficient yielded a lower mean and a greater standard deviation than did the cosine coefficient. It is evident that the latter was hence both more effective (in that it retrieved more active molecules) and more robust (in that there is much less variation across the range of weighting schemes). Similar comments

apply when it is compared with the MinMax coefficient (though to a lesser extent). The variability of the Tanimoto and the MinMax results is illustrated in Figure 4.1, which shows the median numbers of actives retrieved for the 25 different weighting schemes using the MDDR dataset. The figures highlight the poor performance of several of the schemes involving W3 (e.g., W13 and W23) for the Tanimoto and MinMax coefficients, whereas the cosine coefficient is far less affected.

From Table 4.2 to Table 4.4, pairs of observations in the MDDR database were obtained based on the mean actives retrieved and the median actives retrieved. Thus, the Wilcoxon signed-rank test was employed to judge whether there is a difference between the observations that were taken using the Tanimoto coefficient and observations taken using the cosine coefficient; the observations that are taken using the cosine coefficient and observations taken using the MinMax coefficient; the observations that are taken using the Tanimoto coefficient and observations taken using the MinMax coefficient, respectively. For example in Table 4.2 and 4.3, the data from the columns headed "Mean actives" were tested to see whether there is a significant difference between the two sets of 11 values retrieved by two coefficients, i.e., the Tanimoto coefficient and the cosine coefficient. Pairs of observations in the WOMBAT database can be inspected in Table A.1 to Table A.3.

For the Tanimoto–cosine comparison, the Wilcoxon's *W* value is 91 and 80.5 on MDDR pairs; 107 and 99.5 on WOMBAT pairs. The p values on MDDR are significant with p=0.05 on the Mean actives pair and p=0.03 on the Median actives. However, the p values on WOMBAT are not significant showing p=0.15 on the Mean actives pair and p=0.09 on the Median actives pair. These statistical results can help us draw a rough conclusion: the cosine coefficient was superior to the Tanimoto coefficient on MDDR but this was not proved when they were applied on WOMBAT.

The Wilcoxon signed-ranks test analyses of the pairs on two databases are depicted in Table 4.7.

Table 4.7 The Wilcoxon signed-ranks test analysis for all results on MDDR and WOMBAT.

Here, S_T-S_C indicates the Tanimoto – cosine comparison, S_T-S_M indicates the Tanimoto- MinMax comparison and S_C-S_M indicates the cosine – MinMax comparison. *W* stands for the Wilcoxon's *W* value and P for the P value. Significant p values (p<=0.05) are bolded.

Coefficients	MDDR				WOMBAT			
	Mean actives		Median actives		Mean actives		Median actives	
	W	P	W	p	W	p	W	p
S_T-S_C	91.0	**0.050**	80.5	**0.030**	107.0	0.150	99.5	0.090
S_T-S_M	72.0	**0.026**	49.0	**0.004**	59.0	**0.010**	63.0	**0.013**
S_C-S_M	133.0	≥0.200	138.5	≥0.200	149.0	≥0.200	152.0	≥0.200

Observation from Table 4.7 indicates that the cosine coefficient is significantly better (i.e., p≤0.05) than the Tanimoto coefficient for MDDR (both mean and median number of actives), and that the MinMax coefficient is significantly better than the Tanimoto coefficient for both MDDR and WOMBAT (both mean and median numbers of actives). The cosine coefficient is not significantly better than the MinMax coefficient in any set of experiments (despite the former's better screening figures in Tables 4.2 - 4.4 and A.1 - A.3).

As described in Section 4.2, weighting schemes can impact on the density of fingerprints, e.g., W3 changed the values of all the elements that occurred once from 1 to 0. Thus, in the subsequent sections, the detailed comparison of the three coefficients are based on the weighting schemes applied, i.e., if the weighting schemes are *Symmetric or asymmetric*, and *W3 involved or non-W3 involved.* The Wilcoxon signed-rank test analysis for all the results for these schemes are illustrated in Table 4.13.

4.4.1.1 Symmetric or Asymmetric Weighting

It was reported that symmetric measures (both the reference molecule and the molecule from the database are weighted by using the same weighting scheme. e.g., M11, M33) performed better than the asymmetric measures (e.g., M13, M45) using the Tanimoto

coefficient (Arif *et al.*, 2009b). Therefore, these two observations are discussed separately.

Table 4.8 Mean actives (a) and Median actives (b) results of using symmetric (both reference molecule and database structure are weighted by using the same weighting scheme) similarity measures.

Similarity measure	MDDR			WOMBAT		
	S_T	S_C	S_M	S_T	S_C	S_M
M11	109.65	109.57	109.65	103.15	101.66	103.15
M22	86.15	89.73	112.20	86.04	83.12	106.36
M33	69.08	69.77	77.25	71.77	65.45	79.28
M44	114.55	115.40	117.00	103.01	100.79	108.64
M55	112.94	112.61	112.91	103.11	101.84	104.36
Averaged over five symmetric similarity measures	98.47	99.42	105.80	93.42	90.57	100.36

(a)

Similarity measure	MDDR			WOMBAT		
	S_T	S_C	S_M	S_T	S_C	S_M
M11	81.30	80.50	81.30	83.90	80.60	83.90
M22	57.90	57.30	75.00	68.50	66.75	91.25
M33	57.30	43.80	60.10	60.50	56.55	72.00
M44	74.60	73.30	81.70	85.80	83.65	90.10
M55	80.20	79.10	81.40	86.05	86.15	86.45
Averaged over five symmetric similarity measures	70.26	66.80	75.90	76.95	74.74	84.74

(b)

Table 4.8 illustrates that the MinMax coefficient performed the best of the three coefficients and there was no significant difference between the Tanimoto coefficients and the cosine coefficient.

From Table 4.8(a) and Table 4.13(a), the Tanimoto coefficient with the cosine coefficient in MDDR pairs and WOMBAT pairs, p values of the Wilcoxon signed-rank test's are $p>0.2$, $p<0.001$ for MDDR pairs and $p<0.001$, $0.1<p<0.2$ for WOMBAT pairs based on five participants. The result shows no evidence that the two coefficients are different.

By comparing the cosine coefficient with the MinMax coefficient, all the p values of the four pairs are less than 0.001, which shows the MinMax coefficient performed better than the cosine coefficient both on MDDR dataset and on WOMBAT dataset when symmetric measures are used.

For the Tanimoto- MinMax comparison, the p value on MDDR- Mean actives is 0.2, which is insignificant; the p values on MDDR- Median pair and pairs of WOMBAT are less than 0.001, significant.

For the symmetric similarity measures, as shown in Table 4.8(a) and Table 4.8(b), the MinMax coefficient retrieved more actives than did the Tanimoto coefficient and the cosine coefficient on the two databases, while the Tanimoto coefficient performed slightly better than the cosine coefficient. From Table 4.8, the performances of the three coefficients are in the order: MinMax > Tanimoto > cosine. However, it must be emphasized that only five observations are available for each comparison, which makes it difficult to draw conclusions.

Table 4.9 demonstrate the results of the 20 asymmetric similarity measures. It is obvious that the cosine coefficient performed the best on MDDR of the three coefficients. According to the averaged Mean (Median) actives, the performances of the three coefficients are in the order: cosine > MinMax > Tanimoto.

Inspection of the values in Tables 4.8 and 4.9 suggests that the differences between the Tanimoto coefficient and the other two coefficients are quite small for the five symmetric weighting schemes (e.g., M11 or M44) but markedly greater for the remaining asymmetric weighting schemes (e.g., M21 or M24). The Wilcoxon probability values for the asymmetric weighting schemes (in Table 4.13 (a)) are analogous to those in Table 4.13 (b) illustrate that both the cosine and MinMax coefficients often out-performing the Tanimoto coefficient.

A rationale for the robustness of the cosine coefficient in the face of changes in the weighting scheme can be obtained by using an upper-bound analysis. The basic form of the cosine coefficient between a reference structure, X, and a database structure, Y, see Equation 4.21.

Here, regarding as the combination of two weights, *a* and *b*, to form a weighting scheme *Mab* and make two simplifying assumptions. First, that the reference structure is matched with itself, i.e., that X=Y, in which case all of the fragment substructures in the two fingerprints are identical. Second, and more thoroughly, that all of the fragments that are present in the reference structure occur the same number of times, and are thus assigned the same weight; let this weight be *Mnz*, the mean value of the non-zero elements in a fingerprint (as listed in Table 4.1). Then for a weighting scheme *Mab* with mean weight values *Mnz(a)* and *Mnz(b)*, the cosine self-similarity of the reference structure with itself will be

$$S_C = \frac{\sum_{i=1}^{n} Mnz(a)\, Mnz(b)}{\sqrt[2]{\sum_{i=1}^{n} Mnz(a)^2 \sum_{i=1}^{n} Mnz(b)^2}}$$

Equation 4.4

Similarly, the self-similarities using the Tanimoto and MinMax coefficients will be

$$S_T = \frac{\sum_{i=1}^{n} Mnz(a)\, Mnz(b)}{\sum_{i=1}^{n} Mnz(a)^2 + \sum_{i=1}^{n} Mnz(b)^2 - \sum_{i=1}^{n} Mnz(a)\, Mnz(b)}$$

Equation 4.5

and

$$S_M = \frac{\sum_{i=1}^{n} min\{Mnz(a),\ Mnz(b)\}}{\sum_{i=1}^{n} max\{Mnz(a),\ Mnz(b)\}}$$

Equation 4.6

If a symmetric weighting scheme (e.g., M11 or M44) is used then *Mnz(a) = Mnz(b)* for all of the fragments, resulting in all three of the coefficients having the value of unity, which is the upper-bound value for these coefficients. This will also be the case if an asymmetric weighting scheme (e.g., M14 or M23) is used with the cosine coefficient, since the numerator and the denominator terms cancel each other out. For the other coefficients, however, the upper-bound value may not be unity when an asymmetric

weighting scheme is used. The value of the resulting similarity in each case can be calculated by using the data in Table 4.1. For example, using the MDDR database and the w1 and w2 weights, the values of *Mnz* from Table 4.1 are 1.00 and 1.70, respectively; substituting these into the upper-bound formulae above, the computed values (rounded to two decimal places) of the cosine, Tanimoto and MinMax coefficients for matching the reference structure in the W1 representation with itself in the W2 representation are 1.00, 0.78 and 0.59, respectively. Table 4.1 can be used to compute analogous upper-bound values for all of the combined weighting schemes, and the resulting values are listed in Table 4.10 (since self-similarities are being considered here, *M21*, for example, will give the same value as that listed in the table for *M12*). Very similar values are obtained in the WOMBAT database, shown in Table 4.10.

Inspection of Table 4.10 shows that the cosine upper-bound is unity for all the weighting schemes; indeed, this is evident from inspection of the formula. However, the Tanimoto and MinMax upper-bounds vary considerably, meaning that these coefficients are markedly less robust in the face of variations in the weighting schemes used for similarity searching. In saying that, one must remember that the values in this table are for reference-structure self-similarity (whereas in virtual screening, the reference structure is matched against each of the database structures in turn), and are upper-bounds (rather than the actual values that would be obtained in practice) based on the assumption that all the fragments occur the same number of times in the reference structure. Even so, the figures demonstrate clearly the very different character of the cosine coefficient, a characteristic that supports the use of this coefficient for weighted similarity searching. It must be emphasized that the cosine coefficient will not give the best performance in all circumstances (see, e.g., the results in Table 4.2 for the *M12* scheme); however, when averaged over multiple activity classes and weighting schemes, it was superior to the other two coefficients.

Table 4.9 Mean actives (a) and Median actives (b) results of using asymmetric (reference molecule and database structure are weighted by using different weighting schemes) similarity measures.

Similarity measure	MDDR			WOMBAT		
	S_T	S_C	S_M	S_T	S_C	S_M
M12	118.68	80.02	114.82	108.08	72.94	107.15
M13	28.92	39.64	33.93	26.16	39.14	33.75
M14	114.97	104.06	112.04	105.84	95.05	105.15
M15	87.93	108.15	98.74	84.69	99.81	93.71
M21	50.75	113.76	85.15	50.02	97.49	86.22
M23	13.48	76.05	15.80	8.24	73.87	11.76
M24	62.70	114.62	93.95	62.06	100.91	91.87
M25	25.03	117.65	66.16	22.65	102.71	68.99
M31	88.13	94.21	98.47	55.19	65.03	65.96
M32	55.05	71.74	66.20	24.99	62.51	34.53
M34	84.05	99.21	87.33	55.95	82.74	54.29
M35	97.15	100.40	99.35	80.43	76.97	79.00
M41	109.20	118.67	102.65	100.86	104.27	98.91
M42	99.42	93.55	118.51	82.01	84.70	106.23
M43	34.25	66.34	28.29	31.55	65.42	26.31
M45	79.89	118.60	88.85	77.04	104.06	86.72
M51	119.90	113.45	115.30	107.19	102.92	105.66
M52	115.43	85.78	118.25	104.75	77.91	106.51
M53	52.05	49.13	49.02	56.04	49.10	49.83
M54	122.26	108.99	116.84	109.46	97.67	106.49
Averaged over 20 asymmetric similarity measures	77.96	93.70	85.48	67.66	82.76	75.95

(a)

Similarity measure	MDDR			WOMBAT		
	S_T	S_C	S_M	S_T	S_C	S_M
M12	72.10	51.90	79.30	84.90	68.85	82.80
M13	20.40	24.30	23.60	22.75	35.05	30.95
M14	79.00	79.20	81.10	85.10	78.25	84.45
M15	66.50	80.30	74.90	74.50	82.05	78.70
M21	33.20	72.80	55.50	40.95	81.35	72.05
M23	6.60	38.30	8.60	7.05	60.20	11.45
M24	35.50	68.90	58.00	44.65	84.05	81.35
M25	14.30	71.00	42.50	21.30	84.65	59.45
M31	30.60	32.80	25.90	45.95	49.95	50.15
M32	17.90	35.50	22.80	17.10	45.70	31.15
M34	32.70	51.10	29.10	42.45	65.60	42.65
M35	40.60	43.60	47.00	63.75	60.45	63.45
M41	75.40	76.20	73.70	83.30	80.65	80.55
M42	56.50	63.80	77.50	67.70	74.70	85.20
M43	25.00	37.70	17.40	27.15	51.65	22.40
M45	48.40	75.20	58.80	69.10	85.60	74.35
M51	76.10	80.30	78.60	80.55	80.85	83.70
M52	66.10	52.80	75.80	86.35	69.85	82.25
M53	44.90	32.70	38.20	51.05	46.55	49.50
M54	73.50	78.00	77.20	89.65	83.65	83.00
Averaged over 20 asymmetric similarity measures	45.77	57.32	52.28	55.27	68.48	62.48

(b)

Table 4.10 Upper-bound values for the self-similarity of two single reference structures from MDDR and WOMBAT databases, using the Tanimoto, cosine and MinMax coefficients.

Similarity measure	MDDR			WOMBAT		
	S_T	S_C	S_M	S_T	S_C	S_M
M12	0.78	1.00	0.59	0.75	1.00	0.57
M13	1.00	1.00	0.93	0.99	1.00	0.93
M14	0.96	1.00	0.82	0.96	1.00	0.81
M15	0.80	1.00	0.61	0.80	1.00	0.61
M23	0.82	1.00	0.63	0.80	1.00	0.61
M24	0.90	1.00	0.72	0.89	1.00	0.70
M25	0.47	1.00	0.36	0.45	1.00	0.35
M34	0.98	1.00	0.88	0.98	1.00	0.87
M35	0.76	1.00	0.57	0.75	1.00	0.56
M45	0.67	1.00	0.50	0.66	1.00	0.49

4.4.1.2 W3 Involved or Non-W3 Involved Weighting

According to the depiction in Section 4.2, W3 is dissimilar with the other four weighting schemes. It can change all the elements equal to 1, i.e., features set on these potions which occurred only once, into 0. Thus, the fingerprints weighted by using W3 consist of more zeros, which could produce a big impact on the results, shown in Table 4.1, evidentially. Therefore, the results are listed separately based on W3 involved or non-W3 involved weighting.

As shown in Table 4.11, in general, the cosine coefficient yielded more actives than did the other two coefficients. It performed even better on a particular measure, i.e., M23, where the coefficient retrieved five to nine-fold more actives than did the other two. However, its good performance is not always the case, i.e., both the MinMax coefficient and the Tanimoto coefficient can retrieve more with M33, compared to the other W3-measures.

The Wilcoxon signed-ranks test analysis in Table 4.13 (c) shows only the Tanimoto-cosine comparison in MDDR-Mean has a significantly difference.

Table 4.11 Mean actives (a) and Median actives (b) results of using W3 involved similarity measures.

Similarity measure	MDDR			WOMBAT		
	S_T	S_C	S_M	S_T	S_C	S_M
M13	28.92	39.64	33.93	26.16	39.14	33.75
M23	13.48	76.05	15.80	8.24	73.87	11.76
M31	88.13	94.21	98.47	55.19	65.03	65.96
M32	55.05	71.74	66.20	24.99	62.51	34.53
M33	69.08	69.77	77.25	71.77	65.45	79.28
M34	84.05	99.21	87.33	55.95	82.74	54.29
M35	97.15	100.40	99.35	80.43	76.97	79.00
M43	34.25	66.34	28.29	31.55	65.42	26.31
M53	52.05	49.13	49.02	56.04	49.10	49.83
Averaged over nine W3 involved similarity measures	58.02	74.05	61.74	45.59	64.47	48.30

(a)

Similarity measure	MDDR			WOMBAT		
	S_T	S_C	S_M	S_T	S_C	S_M
M13	20.40	24.30	23.60	22.75	35.05	30.95
M23	6.60	38.30	8.60	7.05	60.20	11.45
M31	30.60	32.80	25.90	45.95	49.95	50.15
M32	17.90	35.50	22.80	17.10	45.70	31.15
M33	57.30	43.80	60.10	60.50	56.55	72.00
M34	32.70	51.10	29.10	42.45	65.60	42.65
M35	40.60	43.60	47.00	63.75	60.45	63.45
M43	25.00	37.70	17.40	27.15	51.65	22.40
M53	44.90	32.70	38.20	51.05	46.55	49.50
Averaged over nine W3 involved similarity measures	30.67	37.76	30.30	37.53	52.41	41.52

(b)

Table 4.12 Mean actives (a) and Median actives (b) results of using non-W3 involved similarity measures.

Similarity measure	MDDR			WOMBAT		
	S_T	S_C	S_M	S_T	S_C	S_M
M11	109.65	109.57	109.65	103.15	101.66	103.17
M12	118.68	80.02	114.82	108.08	72.94	107.15
M14	114.97	104.06	112.04	105.84	95.05	105.15
M15	87.93	108.15	98.74	84.69	99.81	93.71
M21	50.75	113.76	85.15	50.02	97.49	86.22
M22	86.15	89.73	112.20	86.04	83.12	106.36
M24	62.70	114.62	93.95	62.06	100.91	91.87
M25	25.03	117.65	66.16	22.65	102.71	68.99
M41	109.20	118.67	102.65	100.86	104.27	98.91
M42	99.42	93.55	118.51	82.01	84.70	106.23
M44	114.55	115.40	117.00	103.01	100.79	108.64
M45	79.89	118.60	88.85	77.04	104.06	86.72
M51	119.90	113.45	115.30	107.19	102.92	105.66
M52	115.43	85.78	118.25	104.75	77.91	106.51
M54	122.26	108.99	116.84	109.46	97.67	106.49
M55	112.94	112.61	112.91	103.11	101.84	104.36
Averaged over 16 non-W3 involved similarity measures	95.59	106.54	105.19	88.12	95.49	99.13

(a)

Similarity measure	MDDR			WOMBAT		
	S_T	S_C	S_M	S_T	S_C	S_M
M11	81.30	80.50	81.30	83.90	80.60	83.90
M12	72.10	51.90	79.30	84.90	68.85	82.80
M14	79.00	79.20	81.10	85.10	78.25	84.45
M15	66.50	80.30	74.90	74.50	82.05	78.70
M21	33.20	72.80	55.50	40.95	81.35	72.05
M22	57.90	57.30	75.00	68.50	66.75	91.25
M24	35.50	68.90	58.00	44.65	84.05	81.35
M25	14.30	71.00	42.50	21.30	84.65	59.45
M41	75.40	76.20	73.70	83.30	80.65	80.55
M42	56.50	63.80	77.50	67.70	74.70	85.20
M44	74.60	73.30	81.70	85.80	83.65	90.10
M45	48.40	75.20	58.80	69.10	85.60	74.35
M51	76.10	80.30	78.60	80.55	80.85	83.70
M52	66.10	52.80	75.80	86.35	69.85	82.25
M54	73.50	78.00	77.20	89.65	83.65	83.00
M55	80.20	79.10	81.40	86.05	86.15	86.45
Averaged over 16 non-W3 involved similarity measures	61.91	71.29	72.02	72.02	79.48	81.22

(b)

For the non-W3 involved similarity measures, the average value of mean actives and median values show that both the cosine coefficient and the MinMax coefficient retrieved more actives than the Tanimoto coefficient produced.

The values of Wilcoxon signed-ranks test in Table 4.13 (d) show that, the Tanimoto-MinMax comparisons in WOMBAT and in MDDR-Median have significantly difference.

A rationale for the robustness of the cosine coefficient can also be obtained by using the upper-bound analysis (refer to Section 4.4.1.1 and Table 4.10) .

4.4.1.3 Conclusion

At the end of the comparison of the three coefficients, two figures (Figure 4.1 and Figure 4.2) are plotted to compare the three coefficients over the 25 measures. As stated at the beginning of the results section, the median value is more resistant to outliers than is the mean value. In this study, the median value can mitigate the effects of some

activity classes (most clearly the Renin and AT1 classes). Therefore, the median values are used to give a general comparison of the three coefficients.

These two figures demonstrate that both the cosine coefficient and the MinMax coefficient yielded better results than did the Tanimoto coefficient. Additionally, the cosine coefficient is distinctly less affected by changes in the weighting scheme that is used, while the Tanimoto coefficient and the MinMax coefficient give very low levels of performance with some types of weighting schemes, e.g., W3; compared with the Tanimoto coefficient, the MinMax coefficient provided good performance with symmetric similarity measures, i.e., M22, M33, M44.

Here, conclusions can be drawn that: The cosine coefficient and the MinMax coefficient often yield larger numbers of actives than are obtained with the Tanimoto coefficient when the frequency-based fingerprints and fragment weighting approaches are used; Moreover, the performance of the cosine coefficient is better than the MinMax coefficients when averaged over 25 weighting schemes. In addition, the cosine coefficient is noticeably less affected by the precise nature of the weighting function that is used, whereas the Tanimoto coefficient and the MinMax coefficient give relatively poor performance when several types of weighting schemes are used.

Table 4.13 The Wilcoxon signed-ranks test analysis for the comparison of pairs of similarity coefficients: (a) five symmetric schemes; (b) 20 asymmetric weighting schemes; (c) W3 involved weighting schemes; (d) Non-W3 involved weighting schemes. Significant p values (p<=0.05) are bolded.

Coefficients	MDDR				WOMBAT			
	Mean actives		Median actives		Mean actives		Median actives	
	W	p	W	p	W	p	W	p
S_T-S_C	3	0.200	0	**0.001**	0	**0.001**	1	0.150
S_T-S_M	1	0.200	0	**0.001**	0	**0.001**	0	**0.001**
S_C-S_M	0	**0.001**	0	**0.001**	0	**0.001**	0	**0.001**

(a) five symmetric schemes

Coefficients	MDDR				WOMBAT			
	Mean actives		Median actives		Mean actives		Median actives	
	W	p	W	p	W	p	W	p
S_T-S_C	54	0.063	35	**0.008**	54	0.063	46.5	**0.030**
S_T-S_M	57	0.081	39	**0.013**	51	**0.046**	54	**0.080**
S_C-S_M	68	0.190	65	0.150	75	0.200	71	0.200

(b) 20 asymmetric weighting schemes

Coefficients	MDDR				WOMBAT			
	Mean actives		Median actives		Mean actives		Median actives	
	W	p	W	p	W	p	W	p
S_T-S_C	2	**0.015**	10	0.200	6	0.060	7	0.070
S_T-S_M	9	0.150	19	0.200	12	0.200	11	0.200
S_C-S_M	9	0.150	11	0.200	11	0.200	12	0.200

(c) W3 involved weighting schemes

Coefficients	MDDR				WOMBAT			
	Mean actives		Median actives		Mean actives		Median actives	
	W	p	W	p	W	p	W	p
S_T-S_C	54	0.200	38.5	0.140	59	0.200	56.5	0.200
S_T-S_M	31	0.120	2	**0.001**	25	**0.050**	25	**0.050**
S_C-S_M	54	0.200	38.5	0.140	53	0.200	53	0.200

(d) Non-W3 involved weighting schemes

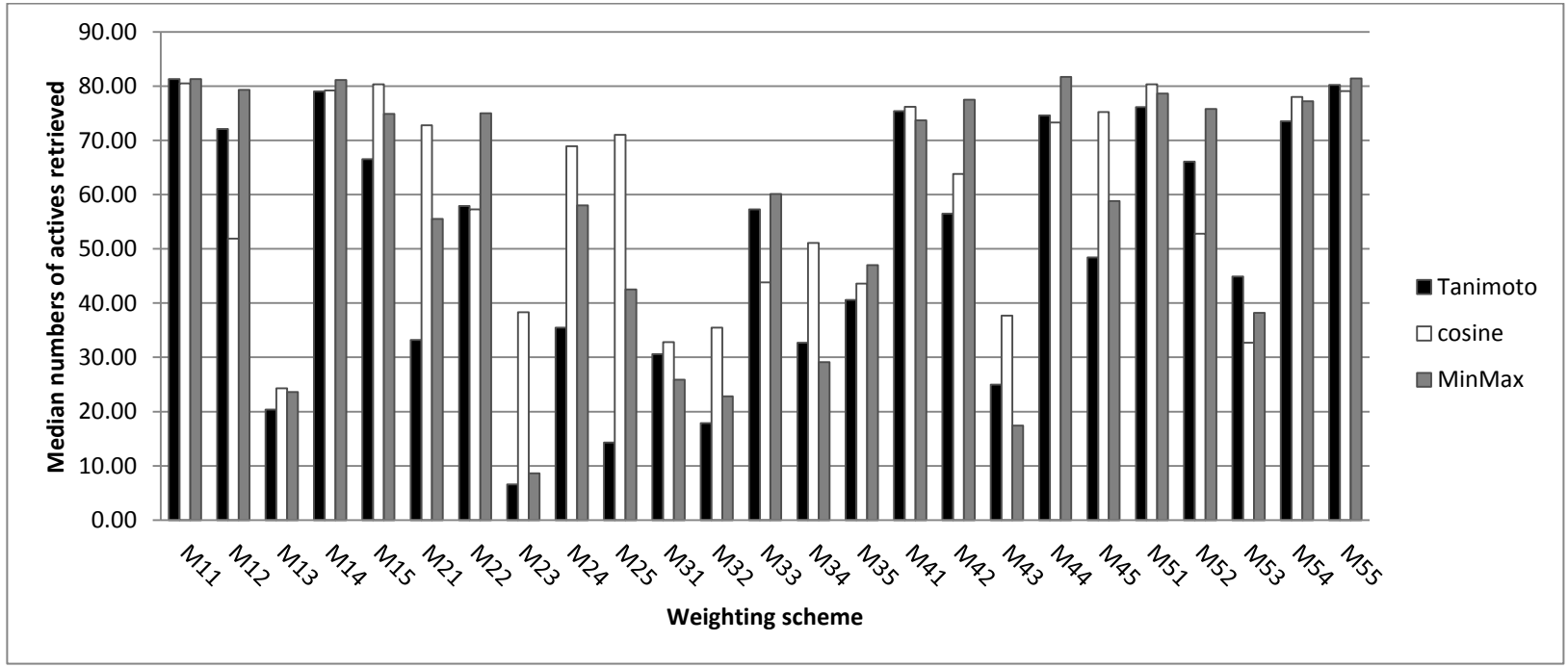


Figure 4.1 Median numbers of actives retrieved in searches of the MDDR database using the Tanimoto, cosine and MinMax similarity coefficients.

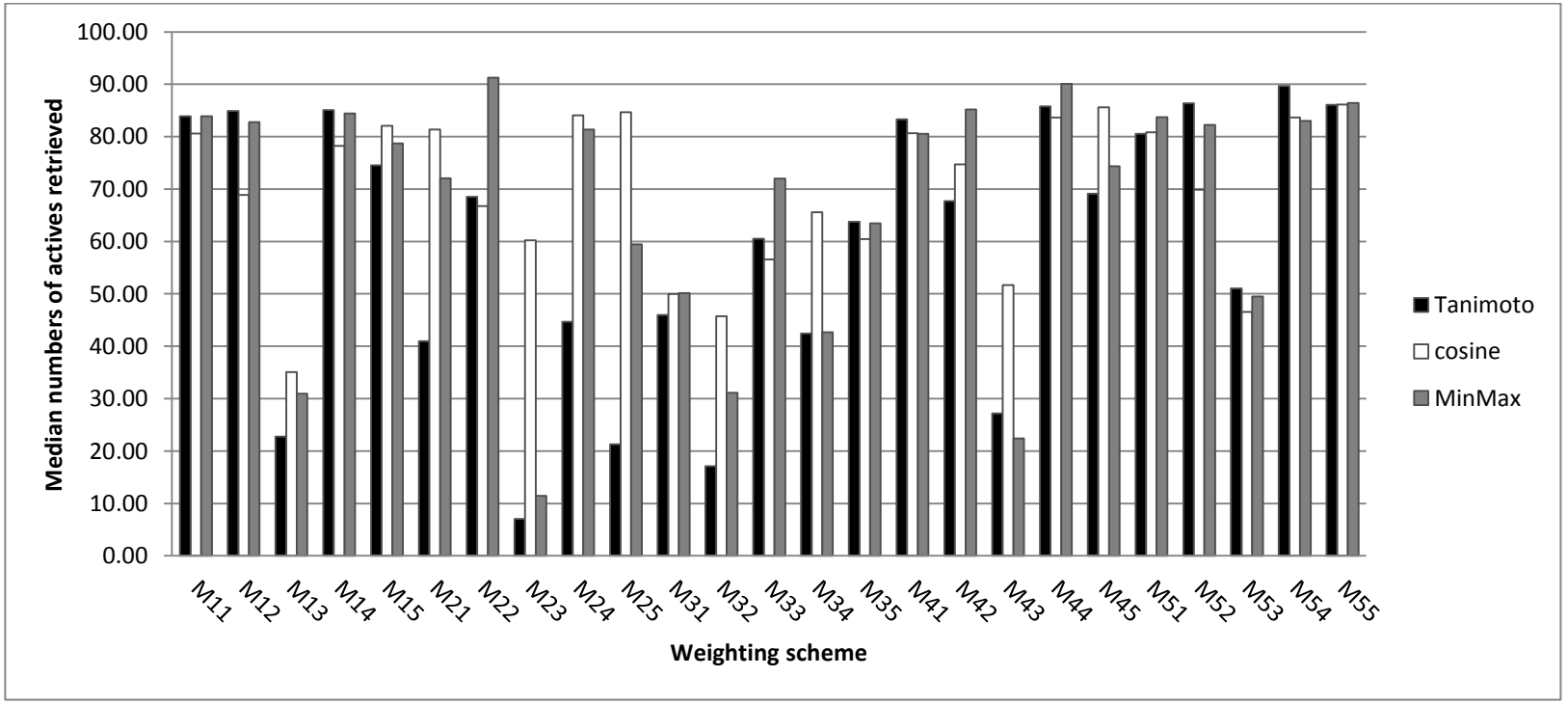


Figure 4.2 Median numbers of actives retrieved in searches of the WOMBAT database using the Tanimoto, cosine and MinMax similarity coefficients.

4.4.2 Comparison of Weighting Schemes

To investigate the effect of the five weighting schemes working on occurrence frequency of fingerprints, the median value of actives are plotted in Figure 4.3 (averaged over the two databases). As Figure 4.3 illustrates, it is apparent that weighting schemes involving W3 performed poorly. Since W3 has been discussed in the previous section, the results are not surprising at all.

The decreasing order of median value of actives molecules retrieved by 25 weighting schemes is:

M55>M11>M44>M14>M54>M51>M41>M15>M12>M52>M42>M22>M45>M24>M21>M33>M35>M25>M34>M53>M31>M43>M32>M13>M23

It is clear that three symmetric measures ranked the top three positions, i.e., M55, M11 and M44. In general, the W4 weighting schemes provided high performance, followed by W5, W1 and W2. To enhance this observation, ranks of the 25 weighting combinations are listed in Table 4.14.

Inspection of Table 4.14, after averaging two databases and three coefficients, the decreasing order of the 25 weighting schemes is:

M44>M51>M55>M54>M41>M14>M11>M12>M52>M45>M15>M22>M42>M24>M21>M25>M35>M33>M34>M31>M53>M43>M32>M13>M23

The result shows the top five rankings are related with W4 and W5. This demonstrates that it is notably better when weighting schemes are applied in similarity measures, rather than with a binary representation (W1). This is coincident with the conclusion from Arif *et al.* (2009b).

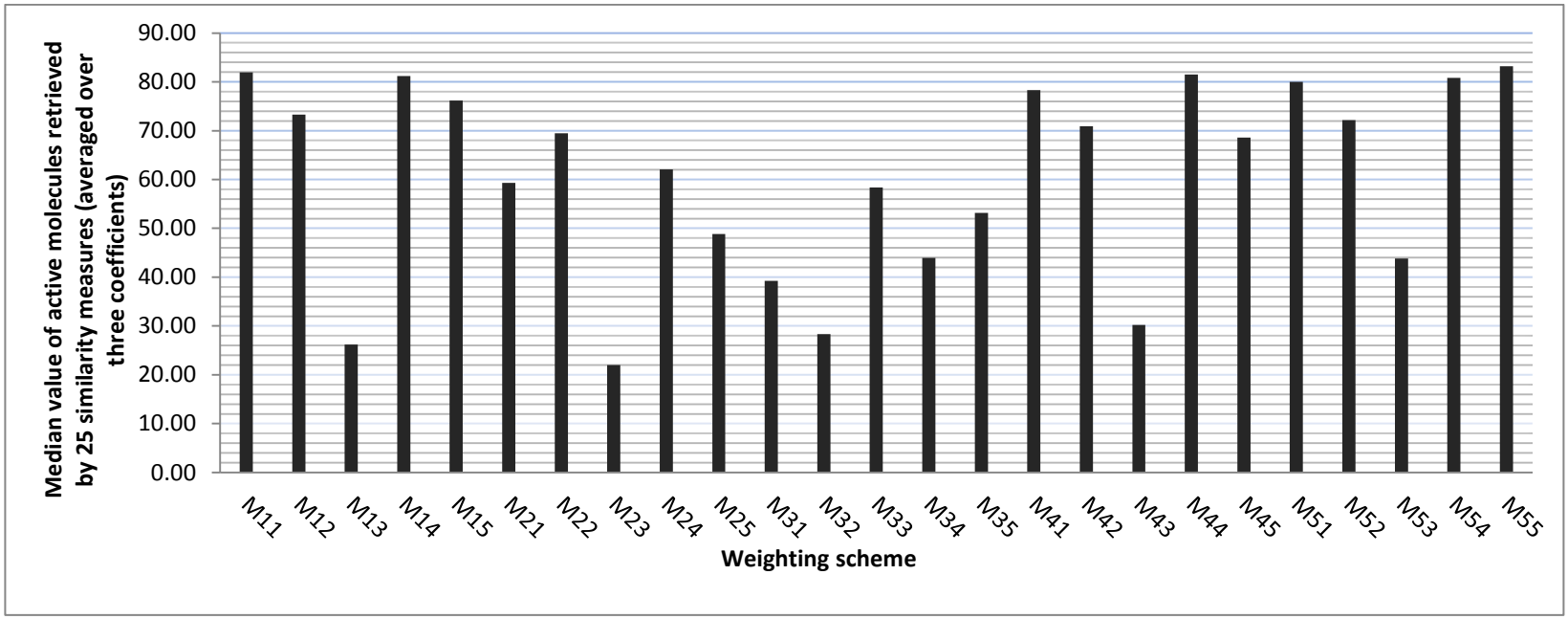


Figure 4.3 Median value of active molecules retrieved by 25 measures (average over MDDR and WOMBAT databases and three coefficients)

Table 4.14 Rankings of 25 combined weighting schemes in similarity searches of the MDDR and WOMBAT using three similarity coefficients. The best result in each column is shaded.

Combined weighting scheme	Similarity coefficient						Mean
	S_T		S_C		S_M		
	MDDR	WOMBAT	MDDR	WOMBAT	MDDR	WOMBAT	
M11	7.36	6.79	8.73	7.36	8.82	8.57	7.94
M12	6.91	5.21	16.91	15.50	7.45	6.21	9.70
M13	21.82	21.86	23.36	23.14	21.36	20.93	22.08
M14	***5.36***	5.64	9.82	10.07	8.18	7.36	7.74
M15	11.18	10.64	8.82	8.57	11.27	11.21	10.28
M21	16.91	16.43	10.18	10.43	14.73	14.00	13.78
M22	10.55	11.00	15.82	14.21	6.73	5.29	10.60
M23	23.45	24.43	19.00	18.29	24.00	24.43	22.27
M24	14.82	14.07	8.00	8.00	12.36	10.79	11.34
M25	21.00	20.21	7.55	7.07	17.36	17.00	15.03
M31	16.00	18.93	18.64	21.79	16.82	19.14	18.55
M32	20.64	22.57	19.00	20.79	20.82	22.79	21.10
M33	14.18	14.50	18.73	18.57	13.27	14.71	15.66
M34	16.36	18.43	14.27	14.86	16.82	20.86	16.93
M35	13.82	12.57	15.55	17.64	15.73	15.79	15.18
M41	6.82	7.36	5.82	***5.64***	10.00	10.43	7.68
M42	12.09	12.71	14.36	12.79	7.64	7.64	11.21
M43	19.64	21.21	20.45	19.29	22.27	23.07	20.99
M44	***5.36***	7.21	6.27	7.64	***4.27***	***4.71***	***5.91***
M45	11.73	12.14	***5.09***	5.86	12.73	13.64	10.20
M51	6.82	5.71	6.82	6.64	8.55	6.93	6.91
M52	10.45	6.57	13.91	13.71	9.00	6.57	10.04
M53	17.64	16.57	21.73	21.29	18.82	18.43	19.08
M54	6.82	***4.57***	8.18	8.36	7.91	6.07	6.99
M55	6.64	7.07	6.55	6.86	7.00	7.50	6.94

4.5 Further Analysis and Evaluation

4.5.1 Effect of Fingerprint's Density

Shown as Figure 4.4, it is apparent that several particular classes contributed more than did other classes in similarity search, i.e., Renin and AT1 in MDDR; Renin and ANG in WOMBAT yielded many more actives than the other classes.

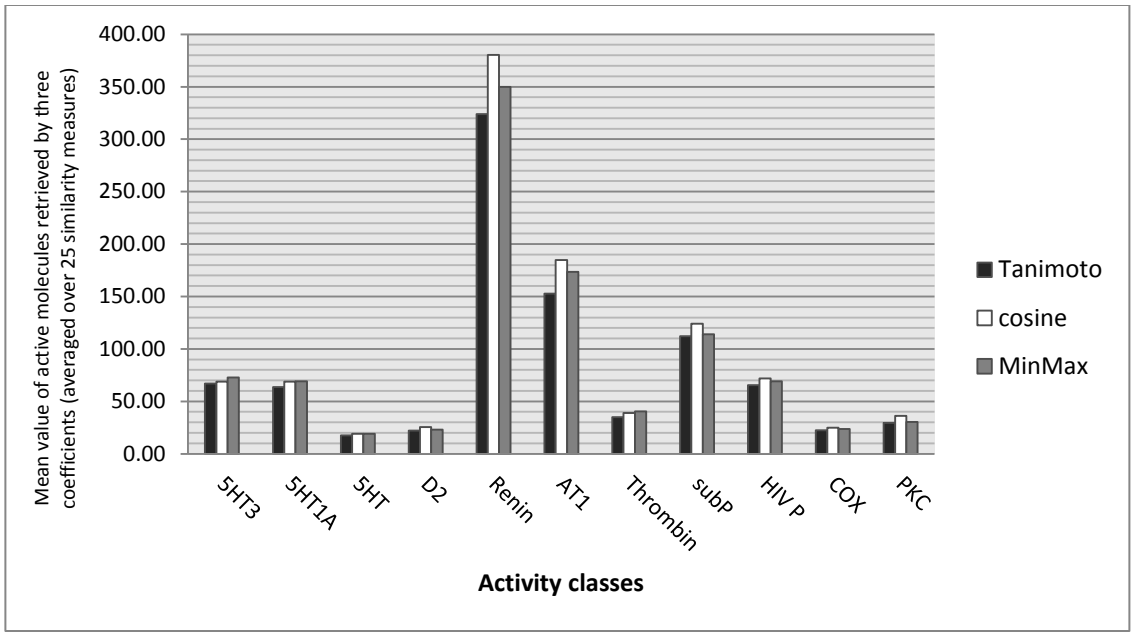
As noted in previous sections, weighting schemes can change the density of fingerprints, i.e., W3 replace all the values of elements that equal to 1 with 0. According to this observation, the fingerprints of the activity classes used in this study are analysed by computing their elements density. Here, the density of fingerprints indicates statistical data for the coding of the ECFC_4 fingerprints in a particular activity class. For example, after averaging over all actives in the class, the fingerprints' density of activity class Renin in MDDR database can be described using four quantitative values, number of elements valued zero (947.41), number of elements valued one (49.27), number of elements valued rather than zero and one (27.32), and the maximum value occurred in the fingerprints of the class. Inspection of Table 4.15 indicates that it is evident that the higher contributed activity classes have higher fingerprints density, e.g., in MDDR database, fingerprints of a structure in activity class Renin normally consists of 34 (981.53-947.41) more non-zero elements than the fingerprints in COX; when W3 weighting scheme applied, although the fingerprints from both two activity classes become sparse, fingerprints' density from Renin still about three times higher than it from COX, i.e., non-zero elements in the fingerprints from Renin is 27.32 compare to 10.54 from COX. Combined with the MPS analysis in Section 3.2.1, the higher fingerprints density classes also provide higher MPS values.

Three coefficients are compared based on the outcomes from individual activity classes. As shown in Figure 4.4, averaging all 25 weighting schemes, activity class Renin contributed the most in MDDR and activity classes Renin and ANG obtained the best performance in WOMBAT. It can be seen that the cosine coefficient performed better

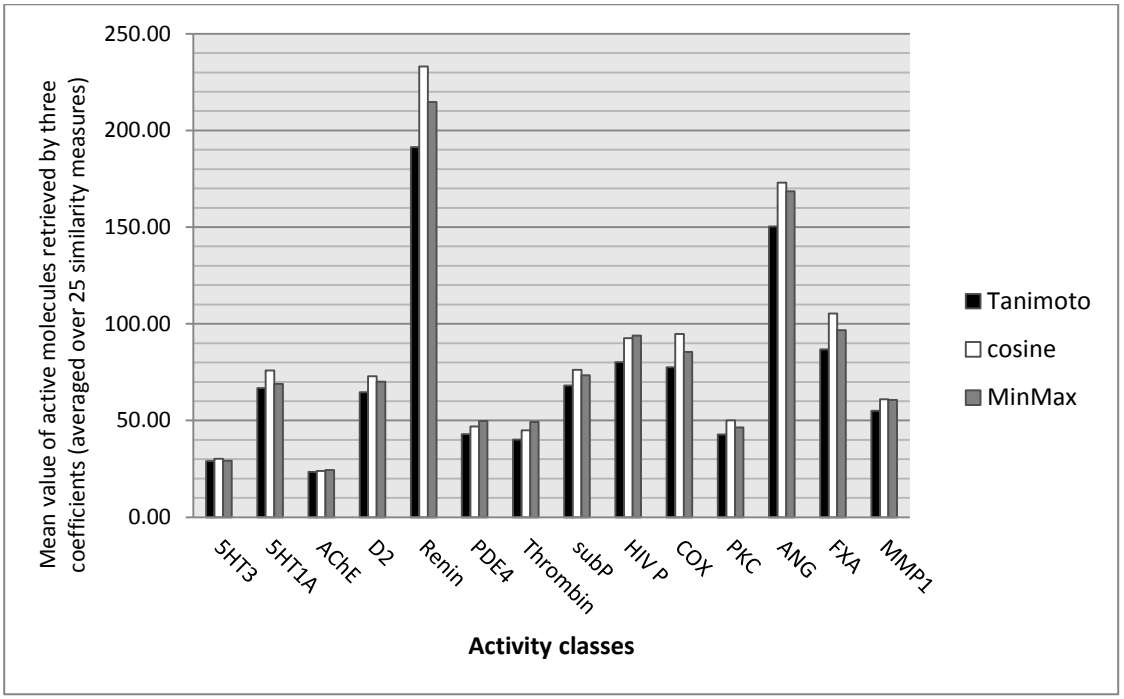
than the other two coefficients when it working with the high density activity classes, i.e., Renin. The performance of the three coefficients were similar when working with the low density activity classes.

Motivated by these observations, a further evaluation needed to be carried out on different databases. Thus, the Maximum Unbiased Validation (MUV) dataset (available by download from http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html, (MUV, 2011)) was selected. This dataset was designed to overcome the problem of analog bias and specifically for the evaluation of virtual screening systems (Rohrer and Baumann, 2009). It is rather different in nature from the MDDR and WOMBAT databases.

Therefore, to further validate this study, the same methods were applied to the Maximum Unbiased Validation (MUV) database.



(a)



(b)

Figure 4.4 Comparison of activity classes. (a) on MDDR, (b) on WOMBAT.

Table 4.15 ECFC_4 fingerprints' analysis for each activity classes in MDDR and WOMBAT, respectively.

Columns 3-5 stand for fingerprints' distribution of activity classes (by averaging over all actives in the class). Column 6 stands for the maximum value occurred in the class.

Dataset	Activity classes	Fingerprints' distribution			
		Mean zero elements per fingerprint	*Mean one elements per fingerprint*	*Mean none-zero and non-one elements*	*Maximum value in fingerprints*
MDDR	Renin	947.41	49.27	27.32	25.00
	AT1	962.74	45.71	15.56	22.00
	Thrombin	961.14	45.24	17.62	36.00
	HIVP	960.65	41.29	22.05	28.00
	SubP	965.91	39.38	18.72	39.00
	D2	972.90	36.59	14.50	21.00
	5HT	976.55	35.70	11.74	17.00
	5HT3	978.87	33.45	11.68	19.00
	PKC	974.31	33.03	16.66	30.00
	5HT1A	977.16	32.25	14.59	16.00
	COX	981.53	31.94	10.54	18.00
WOMBAT	Renin	945.51	50.49	28.00	24.00
	ANG	961.36	46.08	16.56	28.00
	THR	961.95	45.43	16.61	21.00
	FXa	965.18	43.09	15.73	21.00
	HIVP	960.17	40.13	23.70	31.00
	5HT1A	972.64	36.57	14.78	16.00
	D2	974.06	35.83	14.11	23.00
	MMP1	972.54	35.74	15.72	17.00
	PDE4	973.16	35.57	15.27	20.00
	SubP	969.92	34.84	19.24	24.00
	5HT3	978.98	34.41	10.60	14.00
	AChE	975.90	32.30	15.80	26.00
	COX	982.82	30.37	10.81	18.00
	PKC	980.04	26.06	17.91	28.00

4.5.2 Results and Analysis

The methods applied to MUV are the same as used on MDDR and WOMBAT databases. Table A.4 shows the average numbers of active molecules retrieved in the top 1% of searches of the MUV database using the three coefficients. Although the results are poor due to the low MPS value of the MUV database, they are still comparable with those for MDDR and WOMBAT.

Table 4.16 demonstrates the Wilcoxon signed-ranks test analysis of the three coefficients based on the median values.

Table 4.16 Statistical p values for the comparison of pairs of similarity coefficients in the Wilcoxon signed-ranks test on MUV:

column (a) all 25 schemes; (b) 5 symmetric schemes; (c)20 asymmetric weighting schemes; (d) W3 involved weighting schemes; (e) Non-W3 involved weighting schemes. Significant p values (p<=0.05) are bolded.

Coefficients	(a)	(b)	(c)	(d)	(e)
S_T-S_C	***0.01 < P < 0.02***	*P > 0.2*	***0.02 < P < 0.05***	***0.02 < P < 0.05***	*P > 0.2*
S_T-S_M	*P > 0.2*	*P > 0.2*	*P > 0.2*	*0.10 < P < 0.20*	*P > 0.2*
S_C-S_M	***0.01 < P < 0.02***	*P > 0.2*	***0.01 < P < 0.02***	***P < 0.001***	*P > 0.2*

From Table 4.16, it is clear that the difference between S_T-S_C and the difference between S_C-S_M are significant. As the conclusions have been drawn in Section 4.4 that: in general, the cosine coefficient performs better than the Tanimoto coefficient when non-binary fingerprints are used; the cosine coefficient is noticeably less affected by the selection of different weighting schemes, whereas the Tanimoto coefficient and the MinMax coefficient give relatively poor performance when some types of weighting schemes are used; similarity search with weighted fingerprints can retrieve more actives than the search with binary fingerprints, e.g., W4, W5.

Figure 4.5 and Figure 4.6 are plotted to demonstrate the comparison of the three coefficients. As shown in Figure 4.5, it is obvious that the differences between activity classes are not as much as it on MDDR and WOMBAT. Moreover, it is also clear that

the cosine coefficient retrieved more active molecules than did the Tanimoto coefficient and the MinMax coefficient after averaging over 25 similarity measures. This observation supports the conclusion that has been drawn from MDDR and WOMBAT.

Based on Figure 4.5, the difference among the three coefficients' performance in each active datasets are very similar. Thus, the three coefficients working with the 25 similarity measures are compared averaging over 17 activity classes, plotted in Figure 4.6. As shown in Figure 4.6, the cosine coefficient yielded better results than the other two in most cases. However, the MinMax coefficient performed the best when symmetric similarity measures are used, i.e., M22, M33, M55. These observations are corresponding to the previous conclusion from the investigation on MDDR and WOMBAT databases.

From Figure 4.6, it is clear that the weighting schemes related with W4 and W5 obtain better performance, rather than binary (W1), which is also corresponding to the conclusion in Section 4.4.2. Rankings of the 25 weighting schemes is (averaged over three coefficients):

The MinMax coefficient:

M41>M44>M42>M55>M45>M22>M15>M11>M21>M14>M12>M24>M25>M54>M51>M52>M33>M53>M35>M13>M43>M31>M34>M32>M23

The Tanimoto coefficient:

M44>M41>M45>M55>M11>M12>M14>M15>M51>M54>M42>M22>M24>M52>M21>M25>M33>M35>M53>M43>M34>M32>M31>M13>M23

The cosine coefficient:

M44>M45>M41>M51>M15>M11>M55>M54>M14>M42>M25>M24>M52>M21>M12>M22>M32>M33>M23>M34>M43>M35>M53>M31>M13

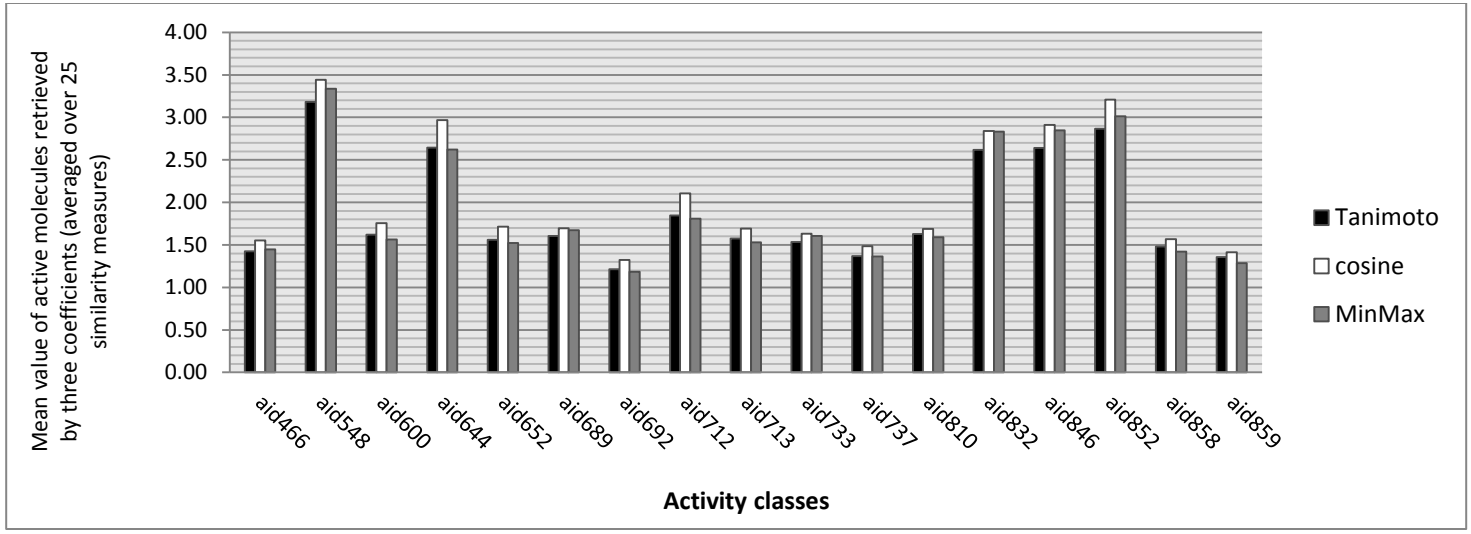


Figure 4.5 Comparison of three coefficients on MUV database (averaged over 25 similarity measures)

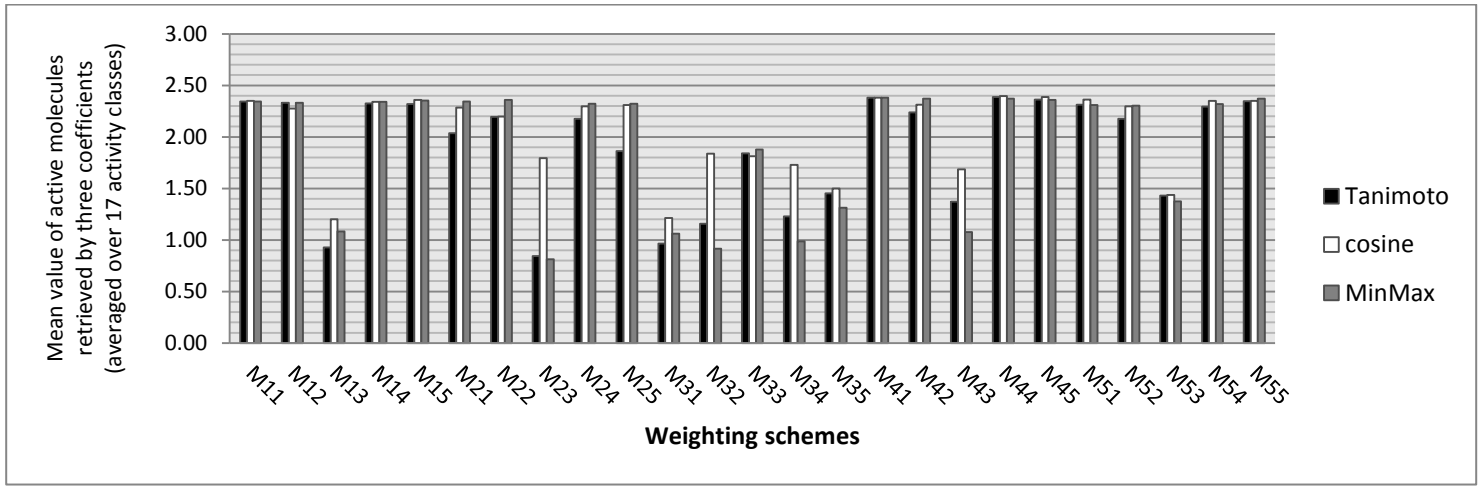


Figure 4.6 Comparison of three coefficients on MUV database (averaged over 17 active datasets)

4.6 Conclusion

In this chapter, detailed investigations were carried out into the interactions between weighting scheme and similarity coefficient in similarity-based virtual screening.

These experiments clearly demonstrate that generally both the cosine coefficient and the MinMax coefficient tended to retrieve greater numbers of active molecules than the Tanimoto coefficient when weighted fingerprints are used. Figure 4.1 and Figure 4.2 demonstrate that the cosine coefficient is more robust than the Tanimoto coefficient in that its screening abilities are much less affected by the precise nature of the weights applied to the fingerprints for the reference structure and the database structures in a similarity search, i.e., the cosine coefficient exhibited an enhanced retrieval performance than the Tanimoto coefficient and the MinMax coefficient when asymmetric measures were applied. For symmetric measures, when compared with the Tanimoto coefficient, the MinMax coefficient was particularly effective. Generally, the cosine coefficient is noticeably less affected by changes in the nature of the employed weighting scheme, whereas both the Tanimoto coefficient and the MinMax coefficient indicated reduced levels of performance with some types of weighting schemes, e.g., W3. However, with the diversity of different activity classes, more research is required, as the findings so far might not be transferable to all activity classes.

Another finding was that W4 (square root) and W5 weighting schemes were superior in 2D molecular similarity search, compared to W3, W1. The results showed that increases in performance can be achieved by weighting the bits in a fingerprint, indicating the presence or absence of 2D substructural fragments.

These findings are hence suggested as the coefficient of choice for similarity-based virtual screening when weighted fingerprints are available, e.g., if the characters (weight/size) of the structures (reference or database) are unknown, then the cosine coefficient might be appropriate for similarity-based virtual screening; if the activity classes are known less diverse, then the cosine and the MinMax coefficients can be the choice for similarity-based virtual screening and W4 and W5 can be used to enhance the performance.

The findings are also suggested further investigation on the interactions between coefficients and weighting schemes, i.e, the coefficients and weighting schemes need to be both considered rather than be considered independently. In addition, more coefficients and weighting schemes need to be identified and be adopted in similarity searching.

Chapter 5: Comparison of Established Level of Binary Coefficients for Chemical Similarity Search

5.1 Introduction

Given the conclusions in Chapter 4, it is evident that certain weighting schemes can enhance the effectiveness of similarity-based virtual screening. The outcome revealed when the weighting schemes were applied, the cosine and the MinMax coefficients exhibited more effective performance. It was also found that when the cosine and the MinMax coefficients were applied to non-weighted data, they yielded similar or identical results as the Tanimoto coefficient.

The results from Chapter 4, however, are based on the investigation of three coefficients. There is a wide variety of coefficients can be adopted in similarity-based virtual screening. Before testing and evaluating the interactions between weighting schemes and a wide range of coefficients, it is necessary to test their performance with non-weighted data, and then select the best performing coefficients. These will be further analysed in Chapter 6.

As explained in Chapter 2, a fingerprint can be considered as a vector with the *i*-th element indicating whether a fragment is present or absent in a molecule. The coefficients studied in the previous chapter focused on the fragment presences rather than their absences, i.e., the Tanimoto, the cosine and the MinMax coefficients. Thus, in this chapter, coefficients which take the fragment absences into account are also investigated.

The study reported in this chapter is part of a collaboration with the Milano Chemometrics and QSAR Research Group (see http://michem.disat.unimib.it/chm/), which compared 44 coefficients using simulated data. An extended comparison of the 44 coefficients was carried out in this study to determine which of these coefficients were suitable for similarity search using real data. The nature and extent of the collaboration is made explicit by Todeschini *et al.*(2012), where the sections on virtual screening are based on research carried out during the thesis and reported in this chapter (which also contains much additional material).

This chapter reports detailed investigations of 44 binary coefficients. Although there have been many previous comparisons of association coefficients for similarity searching (Holliday and Haranczyk, 2008; Holliday *et al.*, 2002; Holliday *et al.*, 2003; Salim *et al.*, 2003; Willett, 2006), none have involved either the number or the range of coefficients considered here.

5.2 Coefficients for Binary Variables

With the intention of choosing the right coefficient, the different coefficients and their nature need to be better understood. Based on the study of Batagelj *et al* (1995) on set theory, a vector can be represented as a unit A which has n properties and each property is of a binary type indicating presence/absence. Unit A has thus the form $A = [x_{1A}, x_{2A}, x_{3A}, \cdots \cdots x_{nA}]$, $x_{iA} \in C = \{0,1\}$ where $x_{iA} = 1$, if unit A has the i-th property in set C, and $x_{iA} = 0$, if A lacks the i-th property in set C, $1 \leq i \leq n$. Thus, the scalar product $AB = \sum_{i=1}^{n} x_{iA}x_{iB}$ of units $A, B \in C$, and with $\overline{A}$, $\overline{B}$ the complementary vectors of A : $\overline{A} = 1 - A = [1 - x_{iA}]$ and B : $\overline{B} = 1 - B = [1 - x_{iB}]$. Therefore, for any pair of vectors A and B, the resemblance can be described by four quantities (a, b, c, and d), see Table 5.1.

Table 5.1.Contingency table of quantities

	$x_{iB} = 1$	$x_{iB} = 0$	
$x_{iA} = 1$	$\boldsymbol{a}$	$\boldsymbol{b}$	$a + b$
$x_{iA} = 0$	$\boldsymbol{c}$	$\boldsymbol{d}$	$c + d$
	$a + c$	$b + d$	$\boldsymbol{n}$

In this table, $a = AB$ indicates the number of common properties present in A and B and $d = \overline{AB}$ the number of common properties absent. $a + b$ and $a + c$ are the number of properties present in A and B, respectively. $b = A\overline{B}$ and $c = \overline{A}B$ stand for the number of properties that appear only in a single side. n is the total of the four quantities, equal to $a + b + c + d$, which is the length of the binary vectors. In other words, a is the proportion of *1*s that A and B share in the same positions (i.e., common "presences"), d is the proportion of *0*s that both A and B share in the same positions (i.e., common "absences").

Therefore, the diagonal entries a and d express the similarity between the two vectors while the entries b and c provide information on their dissimilarity.

As illustrated in the following example, two objects are represented by two following vectors A and B.

A: 1 0 1 1 1 0 0 1 0 1

B: 1 1 0 1 0 1 1 0 0 0

Thus, the a, b, c and d values can be calculated as:

$a = 2 \quad b = 4$

$c = 3 \quad d = 1$

and n = 10, the total number of binary variables.

Here, in this study, all structures are represented as 1,024 length ECFP_4 fingerprints. Therefore, value of n is fixed at 1,024. The actual chemical data used here are sparse databases, i.e., fingerprints consist of a large number of 0s. The value of parameter d could therefore be the largest compared with the other three. For example, for a pair of structures from activity class 5HT3 in MDDR, the d value is around 940 while a is about 10, b is around 40 and c is about 40.

5.3 Classification and Rescaling of Coefficients

5.3.1 Classes of Coefficient

Concerning the importance of property occurrence and nonoccurrence, many similarity measures have been discussed (Baulieu, 1989; Gower and Legendre, 1986; Harris and Lahey, 1978; Hubalek, 1982; Liebetrau, 1983; Snijders *et al.*, 1990; Sokal and Sneath, 1963). Sokal and Sneath (1963), for example, argued that quantity d (i.e., certain properties both absent in two objects) does not contribute to similarity. In their analysis, d is improper when calculating the similarity between two species, e.g., the absence of wings would surely be an absurd character of the affinity between a camel and nematode. Many conventional coefficients such as the Jaccard (1912), the Dice (1945), the Kulczynski (1927) and the DK (Driver and Kroeber, 1932) do not take d into account. Goodman and Kruskal (1954), however, asserted that coefficients should be based on $a + d$ in general. In cases where 1 and 0 stand for two mutually exclusive attributes, e.g., correct and incorrect, the quantities a and d should be equally weighted. The typical coefficient of this sort is the Simple Matching coefficient (Sokal and Michener, 1958), which has also been proposed by Rand (1971), for comparing two clustering algorithms.

In this study, a major distinction of classifying coefficients focuses on those that do, and those that do not, include d: in symmetric coefficients both a and d counts are equally considered; in asymmetric coefficients only a count is considered in measuring the

similarity. There are also intermediate coefficients where both a and d counts are considered, but d is underweighted with respect to a count. The class of correlation-based coefficients is also considered here (Although the correlation coefficients have been argued as being not appropriate for similarity search, as discussed in Section 2.4.3, this has not been proven via large scale screening test).

The 44 coefficients adopted here, therefore, can be classified as belonging to one of the following classes: symmetric (S), asymmetric (A), intermediate (I) and correlation-based (Q). It should be noted that each coefficient will be given an abbreviated symbol in italic, e.g., coefficient Sokal-Michener is given as *SM*. In later sections of this chapter, abbreviated symbols are used to represent the coefficients, abbreviations and formulas as detailed in Table 5.2.

5.3.2 Rescaling of Coefficients

As noted in Table 5.2, many similarity coefficients are defined as fractions. Therefore, the denominator of a coefficient may become zero but it cannot provide appropriate similarity values. For example, when calculating the similarity of two identical objects which means $b = c = 0$, the *Mou* coefficient gives $S = \frac{2a}{ab+ac+2bc} = \frac{2a}{0}$; if the two identical objects consist of only '1's which indicates $a = n$ and $b = c = d = 0$, then the *RG* coefficient gives $S = \frac{a}{2a+b+c} + \frac{d}{2d+b+c} = \frac{n}{2n} + \frac{0}{0}$. For these critical cases, the value of the coefficient is redefined under specific conditions, i.e., when $a = n$ or $d = n$, the similarity value of two objects (S) is set to 1; when the value of the denominator equals zero, $S = 0$. Another case is when the calculated similarity value can extend beyond the range [0, 1]. To solve this case, the similarity measure was rescaled using linear transform:

$$S' = \frac{S + \alpha}{\beta}$$

Equation 5.1 Formula for rescaling coefficients

Where S is the original similarity measure, S' the rescaled function yielding a similarity value between [0,1] and α and β are numerical parameters. Obviously, $\alpha = 0$ and $\beta = 1$ indicate no transformation. Examples of some coefficients can be rescaled using above formula: the symmetric coefficient *Phi* and the correlation-based coefficients *Mic*, *CO1*, *CO2*, *Yu1* and *Yu2*.

5.4 Method

This chapter is devoted to the examination of the performance of different coefficients in virtual screening. Four databases were studied, i.e., MDDR, WOMBAT, MUV and ChEMBL, and all molecules were represented as 1024-length ECFP_4 fingerprints.

The experimental process is described in Chapter 3 and the Kendall W test of concordance has been applied to determine which coefficient is in fact superior compared with the others. Hierarchical cluster analysis has also been used to classify coefficients based on their retrieval abilities, and the hierarchical structures of clusters are visualized as dendrograms with heatmaps.

In this chapter, the initial investigation was carried out on three databases, MDDR, WOMBAT and MUV; then a validation experiment was accomplished on the much larger ChEMBL database.

Table 5.2. List of the binary similarity coefficients.

The first column indicates the abbreviations of coefficients; the second column lists the used coefficients' name(s); the formulas are listed in the third column and the last column shows the symbols of classes the coefficients belong to. Two columns *α and β* provide parameters used for rescaling coefficients. For each coefficient, the corresponding class is shown in the last column, namely S (symmetric), A (asymmetric), I (intermediate) and Q(correlation-based).

Symbol	Name(s)	Formula	α	β	Class
SM	Sokal-Michener, Rand, simple matching	$\frac{a+d}{n}$	0	1	S
RT	Rogers-Tanimoto	$\frac{a+d}{n+b+c}$	0	1	S
JT	Jaccard,Tanimoto	$\frac{a}{a+b+c}$	0	1	A
Gle	Gleason, Dice, Sorenson	$\frac{2a}{2a+b+c}$	0	1	A
RR	Russel-Rao	$\frac{a}{n}$	0	1	A
For	Forbes	$\frac{na}{(a+b)(a+c)}$	0	$\frac{n}{a}$	A
Sim	Simpson	$\frac{a}{min\{(a+b),(a+c)\}}$	0	1	A
BB	Braun-Blanquet	$\frac{a}{max\{(a+b),(a+c)\}}$	0	1	A
DK	Driver-Kroeber, Ochia, cosine	$\frac{a}{\sqrt{(a+b)(a+c)}}$	0	1	A
BUB	Baroni_Urbani-Buser	$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	0	1	I
Kul	Kulczynski	$\frac{1}{2}\cdot\left[\frac{a}{a+b}+\frac{a}{a+c}\right]$	0	1	A

SS1	Sokal-Sneath	$\frac{a}{a+2b+2c}$	0	1	A
SS2	Sokal-Sneath	$\frac{2a+2d}{n+a+d}$	0	1	S
Ja	Jaccard	$\frac{3a}{3a+b+c}$	0	1	A
Fai	Faith	$\frac{a+0.5\cdot d}{n}$	0	1	I
Mou	Mountford	$\frac{2a}{ab+ac+2bc}$	0	2	A
Mic	Michael	$\frac{4\cdot(ad-bc)}{(a+d)^2+(b+c)^2}$	+1	2	Q
RG	Rogot-Goldberg	$\frac{a}{2a+b+c}+\frac{d}{2d+b+c}$	0	1	S
HD	Hawkins-Dotson	$\frac{1}{2}\cdot\left[\frac{a}{a+b+c}+\frac{d}{d+b+c}\right]$	0	1	S
Yu1	Yule	$\frac{ad-bc}{ad+bc}$	+1	2	*Q*
Yu2	Yule	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	+1	2	Q
Fos	Fossum in Holiday et al.	$\frac{n\cdot(a-0.5)^2}{(a+b)(a+c)}$	0	$\frac{(n-0.5)^2}{n}$	A
Den	Dennis in Holiday et al.	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$	$\frac{n}{2\sqrt{n}}$	$\frac{n-1}{\sqrt{n}}$	Q
Co1	Cole	$\frac{ad-bc}{(a+c)(c+d)}$	*n-1*	*n*	Q
Co2	Cole	$\frac{ad-bc}{(a+b)(b+d)}$	*n-1*	*n*	Q

dis	dispersion in Choi et al.	$\frac{ad-bc}{n^2}$	1/4	1/2	Q
GK	Goodman-Kruskal	$\frac{2\cdot min(a,d)-b-c}{2\cdot min(a,d)+b+c}$	+1	2	S
SS3	Sokal-Sneath	$\frac{1}{4}\cdot\left[\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right]$	0	1	S
SS4	Sokal-Sneath	$\frac{a}{\sqrt{(a+b)(a+c)}}\cdot\frac{d}{\sqrt{(b+d)(c+d)}}$	0	1	S
Phi	Pearson-Heron	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$	+1	2	Q
Di1	Dice, Wallace, Post-Snijders	$\frac{a}{a+b}$	0	1	A
Di2	Dice, Wallace, Post-Snijders	$\frac{a}{a+c}$	0	1	A
Sor	Sorgenfrei	$\frac{a^2}{(a+b)(a+c)}$	0	1	A
Coh	Cohen	$\frac{2\cdot(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$	+1	2	Q
Pe1	Peirce	$\frac{ad-bc}{(a+b)(c+d)}$	+1	2	Q
Pe2	Peirce	$\frac{ad-bc}{(a+c)(b+d)}$	+1	2	Q
MP	Maxwell-Pilliner	$\frac{2\cdot(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)}$	+1	2	Q

HL	Harris-Lahey	$\frac{a \cdot (2d + b + c)}{2 \cdot (a + b + c)} + \frac{d \cdot (2a + b + c)}{2 \cdot (b + c + d)}$	0	*n*	S
CT1	Consonni-Todeschini	$\frac{ln(1 + a + d)}{ln(1 + n)}$	0	1	S
CT2	Consonni-Todeschini	$\frac{ln(1 + n) - ln(1 + b + c)}{ln(1 + n)}$	0	1	S
CT3	Consonni-Todeschini	$\frac{ln(1 + a)}{ln(1 + n)}$	0	1	A
CT4	Consonni-Todeschini	$\frac{ln(1 + a)}{ln(1 + a + b + c)}$	0	1	A
CT5	Consonni-Todeschini	$\frac{\ln(1 + ad) - \ln(1 + bc)}{\ln(1 + \frac{n^2}{4})}$	0	1	S
AC	Austin-Colwell	$\frac{2}{\pi} \cdot \sin^{-1}\sqrt{\frac{a + d}{n}}$	0	1	S

5.5 Results of Initial Investigation

In the following Section 5.5.1, the 44 coefficients are compared using their rank positions of each activity class. For each coefficient, a median value of retrieved active compounds was calculated over 10 runs. These very extensive tests are summarized in Appendix B. In each activity class, all of the coefficients can therefore be ranked using the median values in decreasing order to show their retrieval abilities from 1 to 44. These ranks can provide a general view of the coefficients' performance.

In Section 5.5.2, the 44 coefficients are compared using the actual results. The impact from the nature of activity classes can therefore be determined, and the variation of the coefficients' performance can be quantitatively measured.

5.5.1 Ranks of Retrieval Abilities

According to the median value of retrieved active compounds, in each activity class, the 44 coefficients were ranked in decreasing order. Thus, for each coefficient in an activity class, corresponding rank positions can be numbered from 1 to 44.

Table 5.3 gives the statistically significant levels of concordance of number of actives across the activity classes which were observed for the MDDR, WOMBAT and MUV databases. The *W* values of the three databases are significant. Hence, the overall rankings for the three databases can be generated and these are shown in Table 5.4. It should be noted that even though the *W* value of MUV database is statistically significant, it is in the weak range ($W = 0.185$).

Table 5.3 Kendall's test of concordance results.

	Statistic		
	W	χ^2	*P*
MDDR	0.353	167	$p \leq 0.001$
WOMBAT	0.504	303	$p \leq 0.001$
MUV	0.185	135	$p \leq 0.05$

Based on the outcomes in Table B.1 to Table B.3 (see Appendix B), for each similarity coefficient, Table 5.4 presents the ranks of retrieval ability in the three databases using Kendall's *W* test. Note that the lower the average rank, the greater the capability of a coefficient to retrieve compounds belonging to a specific active class. The MDDR and WOMBAT columns in Table 5.4 reveal a very high degree of resemblance throughout the entire ranked list. The ranking for the MUV dataset is rather different, with the top-ranked coefficients for MDDR and WOMBAT appearing lower down the ranked list; however several of the clusters of coefficients that are apparent in the MDDR and WOMBAT columns are also apparent here (e.g., the clusters containing coefficients *JT, Gle, SS1, Ja* and coefficients *Yu1, Yu2, CT5*). Moreover, observation from Table B.3 indicates that it is probably not appropriate to read too much into the MUV column of Table 5.4.

According to Table 5.4, the results of coefficients *JT, Gle, SS1, Ja* and *CT4* are clearly evident, with all ranked first equal over the 11 activity classes comprising the MDDR data set, and coefficient *Phi* and *HL* ranked fifth and numerically very near to the first-placed group. Among these performance indices, there is the Tanimoto coefficient (*JT*), which is the coefficient of choice in most operational similarity searching systems (Leach and Gillet, 2007). Also ranked high, along with others, is the coefficient *Fos*. While for coefficients *SM, RT, SS2, CT1, CT2, AC* the ranks observed are very low, both in MDDR and WOMBAT. By detecting the coefficients' classes, *JT, Gle, SS1, Ja* and *CT4* belong to the asymmetric class in which the value *d* is not involved in calculating the similarity value, while *SM, RT, SS2, CT1, CT2, AC* are from the symmetric class in which *d* is an equally treated factor when calculating similarity values.

From Table 5.4, it can be seen that a number of coefficients provide identical ranks, i.e., coefficients *{JT, Gle, SS1, Ja}*, coefficients *{dis, Pe1}*, coefficients *{For, DK, Sor}*, coefficients *{Yu1, Yu2, CT5}*, coefficients *{RR, Di1, CT3}*, coefficients *{Co2, Di2}* and coefficients *{SM, RT, SS2, CT1, CT2, AC}*. Based on these observations, the diversity of the coefficients and the relationship between type of coefficient and screening

effectiveness should be considered. This is analysed and discussed using a cluster analysis method in more detail later in this chapter.

Figure 5.1 and Figure 5.2 illustrate the hierarchical structures of clusters by similarity coefficients based on the ranks of coefficients' retrieval abilities in different activity classes. For each figure, the intersection of a column (coefficient) and a row (activity class) represents the rank of retrieval ability among the 44 coefficients in a corresponding activity class. All cells are coloured from red to blue to indicate their retrieval abilities from strong to weak (the colour keys are shown in the left corner scale) in a certain class, as a heatmap. For example, in Figure 5.1, the first cell (intersection of the first column and the first row) is blue which indicates that the *AC* coefficient's retrieval ability was ranked about the 40th among all coefficients in activity class 5HT. When comparing other cells in the first column, the first cell also demonstrates that the *AC* coefficient is much less effective in 5HT compared to activity class COX in the first column coloured red. In both figures, the class of coefficients are labelled after their names, as shown in the x axes. All coefficients were clustered and were linked at increasing levels of dissimilarity, as shown the dendrograms in the top of the heatmaps. The clustering method was introduced in Section 3.5. For example, in Figure 5.1, the clustering process starts out with all examples (coefficients) in 44 clusters of size 1 each, and each coefficient consists of 11 elements (their ranks in 11 activity classes). Then the pairs (group) of coefficients that yield the smallest error sum of squares (refer to Section 3.5) will form a new cluster. This process stops when all coefficients are combined into a single large cluster of size 44. The heatmaps clearly shows that the coefficients are grouped in terms of ranks of their retrieval abilities.

Table 5.4 Rank positions of each of the 44 coefficients when averaged over all of the activity classes for each of the three databases.

Rank	MDDR	Average ranks	WOMBAT	Average ranks	MUV	Average ranks
1	JT, Gle, SS1, Ja, CT4	13.86	HL	10.46	BUB	18.53
2			CT4	10.75	Fai	19.71
3			JT, Gle, SS1, Ja	12.96	RG, Den, SS4, Coh, MP	19.76
4						
5						
6	Phi, HL	15.36				
7			BB	13.14		
8	Fos	15.50	Fos	14.11	Fos	19.85
9	SS4	15.73	GK	14.64	HD	20.38
10	dis, Pe1	15.96	RG, dis, Pe1	15.07	CT4	20.59
11					SS3	20.74
12	For, DK, Sor	16.09			For, DK, Phi, Sor, HL	20.79
13			For, DK, Sor	15.54		
14						
15	Den	17.14				
16	MP	17.18	SS4	15.61		
17	Coh	17.27	Coh	15.79	GK	21.18
18	GK	17.64	MP	16.07	BB	21.26
19	Co1	18.36	Phi	17.75	JT, Gle, Kul, SS1, Ja	21.71
20	Kul, RG	18.91	Kul	18.61		
21			SS3	18.93		
22	BB	19.64	Den	19.04		
23	SS3	20.41	HD	19.93		
24	Mic	20.64	Co1	21.96	Yu1, Yu2, CT5	21.76
25	BUB	21.09	BUB	22.21		
26	Yu1, Yu2, CT5	23.50	Mic	23.36		
27			Yu1, Yu2, CT5	24.75	Sim	23.35
28					RR, Di1, CT3	23.71
29	RR, HD, Di1, CT3	25.73				
30			RR, Di1, CT3	26.29		
31					Mic, Co1, dis, Pe1	23.74
32						
33	Pe2	28.68	Fai	28.61		
34	Mou	29.14	Sim	28.86		
35	Sim	29.91	Mou	30.93	Pe2	24.00
36	Fai	30.23	Pe2	32.96	SM, RT, SS2, Co2, Di2, CT1, CT2, AC	24.29
37	Co2, Di2	31.05	Co2, Di2	33.93		
38						
39	SM, RT,SS2, CT1,CT2, AC	36.32	SM, RT,SS2,CT1, CT2, AC	38.61		
40						
41						
42						
43						
44					Mou	43.38

From Figure 5.1 and Figure 5.2, it can be seen that the best performing coefficients in MDDR and WOMBAT columns from Table 5.4 all achieved high ranks in the majority of activity classes, e.g., coefficients *JT, Gle, SS1, Ja, CT4* and *For* ranked high in many activity classes, both in MDDR and WOMBAT. After clustering the coefficients, most of the coefficients from the asymmetric class, and those from the correlation-based class, were well clustered. In MDDR, asymmetric coefficients *CT3, RR, Di1* and correlation-based coefficients *dis, Pe1, Co1, Mic* were tightly clustered together. In another cluster, asymmetric coefficients *Kul, CT4, Fos, For, DK, Sor, BB, JT, Gle, SS1, Ja* and correlation-based coefficients *Phi, Den, MP, Coh* were also clustered together. Five symmetric coefficients *SS3, HL, SS4, GK, RG* were also grouped in this cluster. All the coefficients in these two clusters yielded high ranks. The poorer performing coefficients were clustered together, consisting of two intermediate coefficients *Fai* and *BUB,* eight symmetric coefficients *AC, CT2, CT1, SS2, SM, RT, CT5, HD,* four correlation-based coefficients *Yu1, Yu2, Pe2, Co2* and three asymmetric coefficients *Sim, Mou, Di2.* Similarly, the poorer performing cluster in WOMBAT consisted of the same coefficients of those in MDDR, except *BUB* and *HD.*

Inspection of Figure 5.1 indicates that activity class COX is anomalous among the 11 activity classes. The top ranked six coefficients (*RT, SM, SS2, CT1, CT2, AC*) in COX all have complementary performance in the other 10 activity classes. In COX, the top ranked six coefficients are symmetric measures and the last four are asymmetric measures. To investigate the differences between COX and the other activity classes, MPS of activity classes were considered (see Chapter 3 Table 3.1). These indicate the class's diversity (i.e., its degree of structural heterogeneity). COX is the most diverse class in MDDR with the MPS value at 0.27, the lowest among all 11 activity classes. Another interesting observation is that the performance of asymmetric coefficients *Di1, RR* and *CT3* are very effective in less diverse activity classes HIVP, Thrombin, AT1 and Renin compared to the highly diverse activity classes COX, PKC, 5HT3, 5HT1A and 5HT. Similarly, correlation-based coefficients *dis, Pe1, Mic* and *Co1* that yielded high ranks in less diverse activity classes and performed poorly in the highly diverse activity classes.

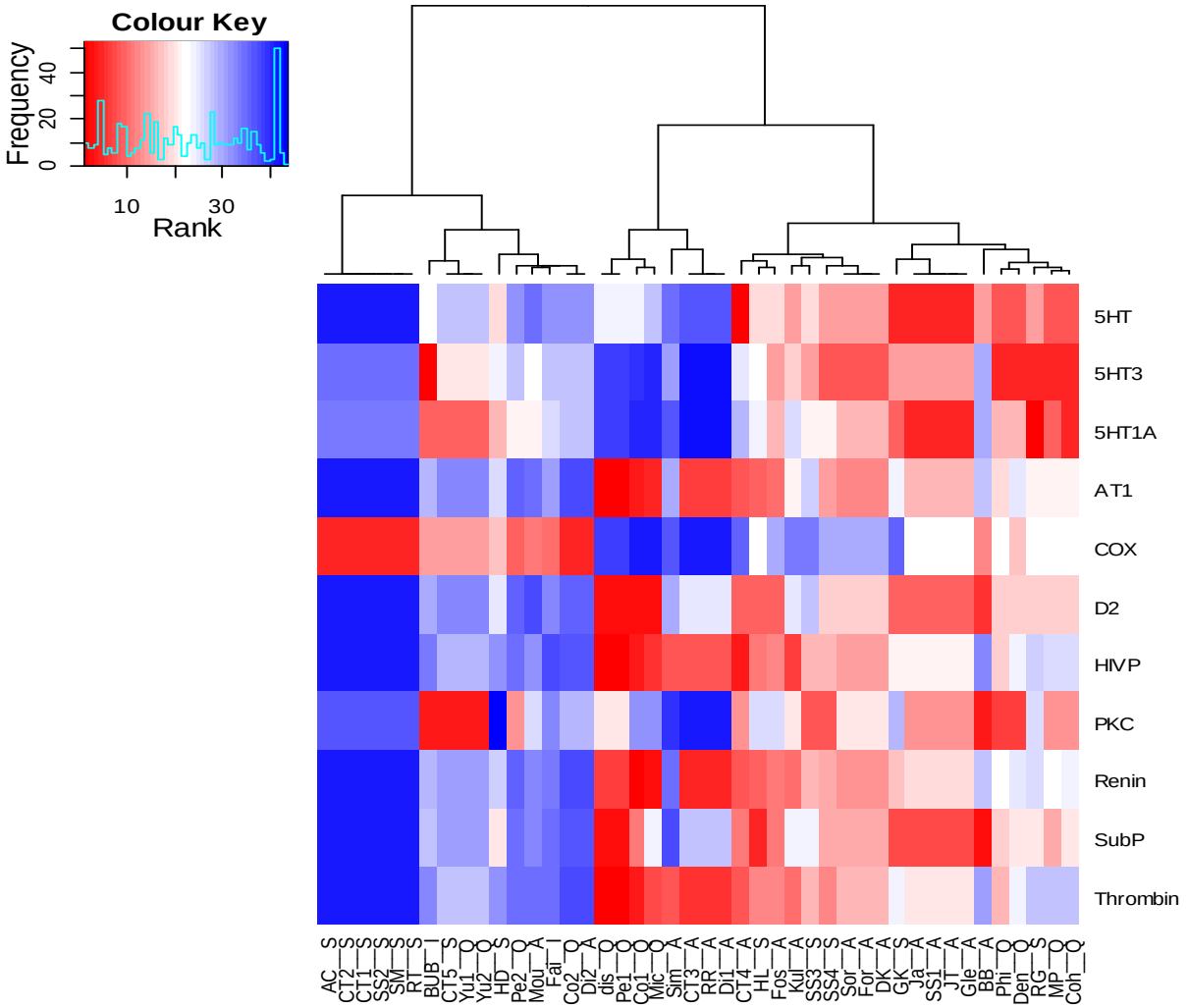


Figure 5.1 Heatmap of the ranks of coefficients in MDDR.

The columns (coefficients) are clustered using Ward's method. Ranks are coloured from red to blue which represents coefficients' retrieval abilities from strong to weak. In the left top corner legend, the X axis scales the ranks where red represents higher positions and blue denotes lower positions; the Y axis measures the frequency of corresponding ranks where the histographic curve shows the number of the ranks.

Similar observations also detected in Figure 5.2. Coefficients' ranks in COX and PKC were different from the ranks of the other 12 classes. For example, in PKC, poor performing symmetric coefficients *RT, SM, SS2, CT1, CT2* and *AC* obtained high ranks while high achieving coefficients *Pe1*, *dis*, *CT4* and *HL* performed badly. Similar to

PKC, ranks of COX were also visibly different, e.g., *Yu1*, *Yu2* and *CT5* ranked very high only in PKC and COX but not in the others.

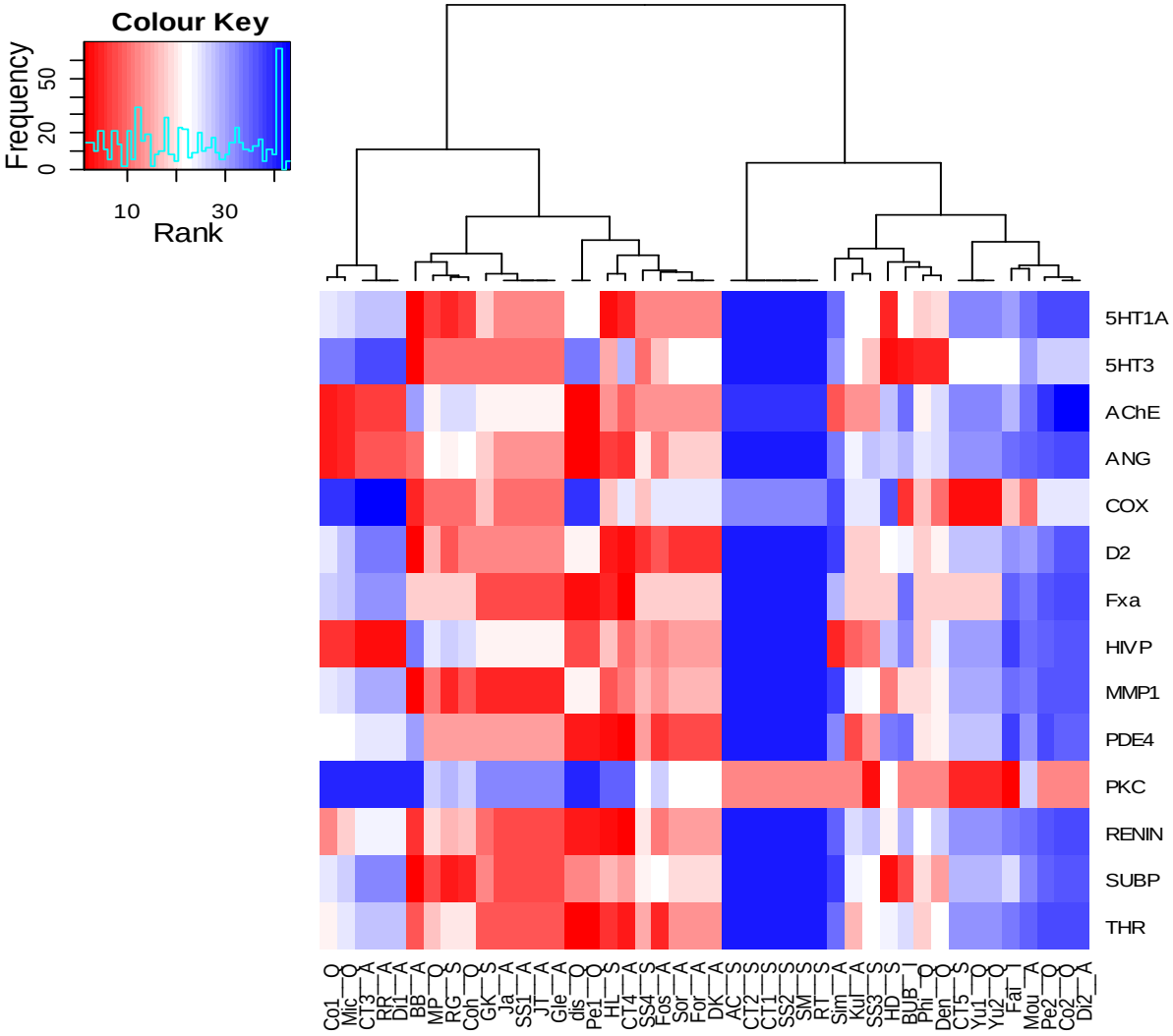


Figure 5.2 Heatmap of the ranks of coefficients in WOMBAT.

The columns (coefficients) are clustered using Ward's method. Ranks are coloured from red to blue which represents coefficients' retrieval abilities from strong to weak. In the left top corner legend, the X axis scales the ranks where red represents higher positions and blue denotes lower positions; the Y axis measures the frequency of corresponding ranks where the histographic curve shows the number of ranks.

5.5.2 Comparison of Retrieval Rate of Active Compounds

5.5.2.1 Retrieval Rate of Active Compounds

The experiment did not simply classify coefficients by their mode of construction described in Section 5.3.1 (based on the involvement of variable *a*, *b*, *c* and *d*). The retrieval rate of active compounds was used here for grouping coefficients. It is the ratio of retrieved active compounds to the total active compounds in the corresponding activity class. Different from using the actual number of retrieved actives, using retrieval rate of actives can reduce the influence of the nature of activity classes, i.e., the difference of the size of activity classes. For instance, activity classes M and N consist of 100 and 1000 actives, respectively. Coefficient A retrieved 50 actives from M and 500 actives from N. Coefficient B retrieved 80 actives from M and 450 from B. Thus, the total amount of actives retrieved by A was 550 and that retrieved by B was 530. In this case, the total amount of retrieved actives is not suitable for evaluating the performance of the two coefficients. While, their performance can be compared using the percentage of actives retrieved, as shown below:

$$Retrieval\ rate_{top-x\%} = \frac{\#\ active\ componds\ retrieved}{\#\ total\ compounds\ in\ a\ certain\ activity\ class} \cdot 100$$

Equation 5.2 Formula of Retrieval rate

Here, $Retrieval\ rate_{top-x\%}$ refers to the percentage of active compounds retrieved at a certain intercept x. In this study, x equals 1.

Table B.4 in Appendix B is an example to represent the median retrieval rate out of 10 runs of 11 activity classes in MDDR.

5.5.2.2 Grouping Coefficients using Retrieval Rate of Active Compounds

In Section 5.5.1, coefficients were clustered using their ranks. Most of the coefficients from the same class can be clustered together, e.g., correlation-based coefficients *dis*,

Pe1, *Co1*, *Mic* were tightly clustered together. However, a small number of coefficients from different classes were also grouped, e.g., symmetric coefficients *SS3*, *SS4* were grouped with asymmetric coefficients *Kul*, *CT4*, *For*, *Sor* and *DK*. In order to mirror these observations, the values of retrieval rate of active compounds were used for cluster analysis.

Figure 5.3 illustrates the top 1% retrieval rates of active compounds in MDDR. Similar to Figure 5.1, coefficients' retrieval rates are coloured blue to red to show the values of retrieval rates from low to high. For example, the first cell (intersection of the first column and the first row) is blue which indicates that the *Pe2* coefficient's retrieval rate in 5HT class was low. The first column shows clearly that the *Pe2* coefficient performed well (with retrieval rate at 17.23%) on the AT1 activity class, to a lesser extent on the 5HT3 activity class with retrieval rate at 9.77%, and poorly on the remaining nine activity classes (retrieval rates less than 5%). The outcomes are scaled in the left top corner legend where the X axis values refer to the percentage of retrieval rates and the Y axis represents the frequency of corresponding retrieval rates. It is evident that, all of the 44 coefficients' retrieval rates were below 20% in class HIVP, 5HT3, SUBP, COX, Thrombin, D2, 5HT, PKC and 5HT1A. In SUBP and COX, most of their outcomes were below 5%. Two activity classes, however, were distinct from the other nine classes, i.e., AT1 and Renin which are less diverse classes with MPS value 0.40 and 0.57, respectively. In AT1, all coefficients' retrieval rates were between 20% and 40% and symmetric coefficients *RT*, *SM*, *SS2*, *CT1*, *CT2* and *AC* worked less well than the others. In Renin, half of the coefficients achieved more than 40% retrieval rates. Symmetric coefficients *RT*, *SM*, *SS2*, *CT1*, *CT2* and *AC* were ineffective with retrieval rates under 10%. Generally, asymmetric coefficients and correlation-based coefficients performed better than other coefficients. Similar findings were also found in WOMBAT, as illustrated in Figure 5.4.

After clustering, from the dendogram in Figure 5.3, most of the asymmetric coefficients and correlation-based coefficients with higher retrieval rates were clustered together. In these, three asymmetric coefficients *CT3*, *RR*, *Di1* and four correlation-based

coefficients *Mic*, *Co1*, *dis*, *Pe1* were tightly clustered together, which is similar to the observations in Figure 5.1. Five symmetric coefficients *SS3*, *HL*, *SS4*, *GK*, *RG* yielding high retrieval rates also grouped in this cluster which is in agreement with Figure 5.1. Compared to the low performance cluster in Figure 5.1, one more asymmetric coefficient *BB* was grouped in the low retrieval rates cluster.

According to Table B.4, the average retrieval rates of asymmetric, symmetric, correlation-based and intermediate coefficients over 11 activity classes in MDDR can be ordered as follows: asymmetric coefficients > correlation-based coefficients > intermediate coefficients > symmetric coefficients with average values of 10.57%, 10.21%, 8.01% and 7.70% respectively. For WOMBAT, the average retrieval rates of the four classes of coefficients over 14 activity classes are ordered accordingly: asymmetric coefficients > correlation-based coefficients > intermediate coefficients > symmetric coefficients with average values of 13.63%, 12.98%, 11.86% and 10.32% respectively.

Observations from Figure 5.3 and Figure 5.4 support well the classification of coefficients in Section 5.5.1. They also show that coefficients from the same class often yield similar retrieval rates of active compounds. Coefficients from the symmetric class, can, however, also produce very different results, i.e., coefficients *RG*, *SS3*, *SS4* and *HL* performed very well compare to coefficients *SM*, *RT*, *SS2*, *CT1*,*CT2* and *AC*, both in MDDR and WOMBAT.

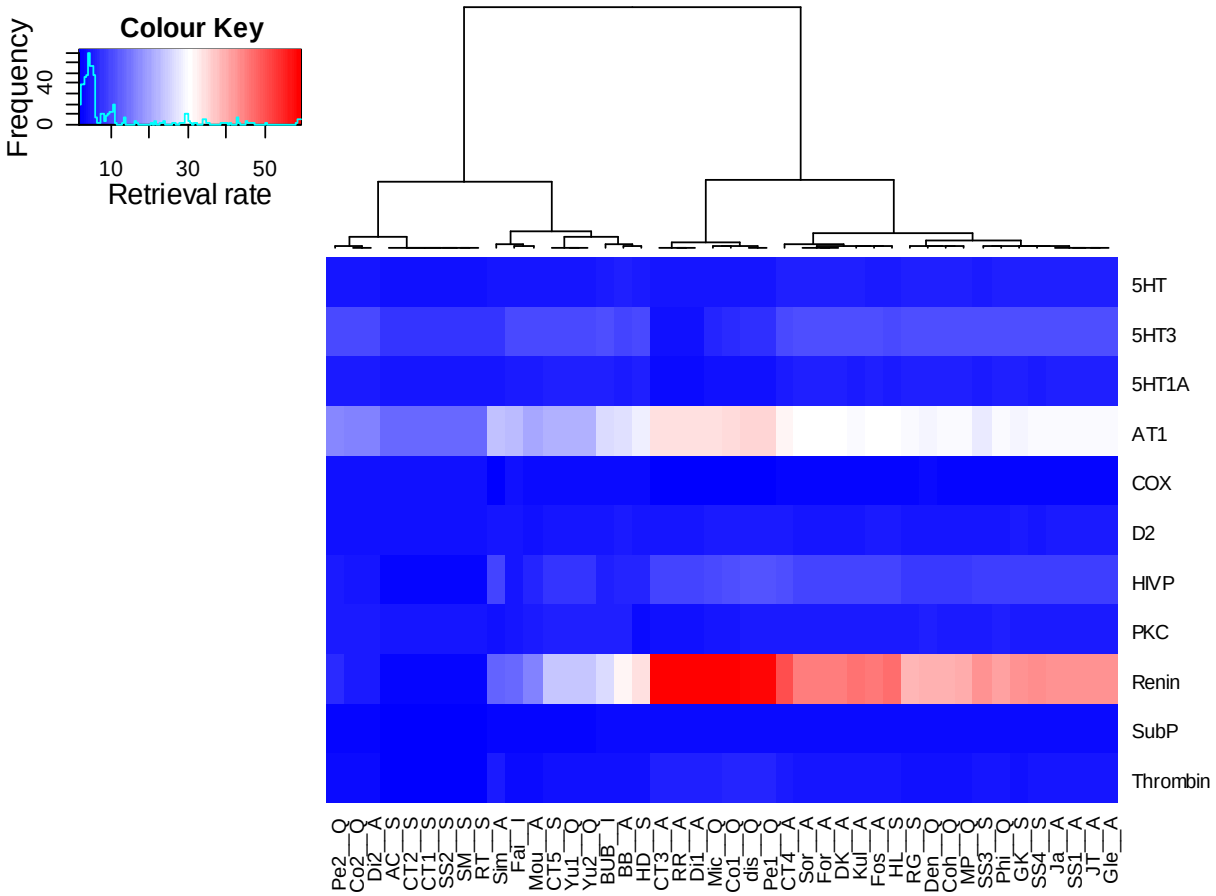


Figure 5.3 Heatmap of the retrieval rates of coefficients in MDDR.

The columns (coefficients) are clustered using Ward's method. Retrieval rates are coloured from red to blue which represents coefficients' retrieval abilities from strong to weak. In the left top corner legend, the X axis scales the retrieval rates where red represents higher rates and blue stands for lower rates; the Y axis measures the frequency of corresponding rates where the histographic curve shows the number of that rates.

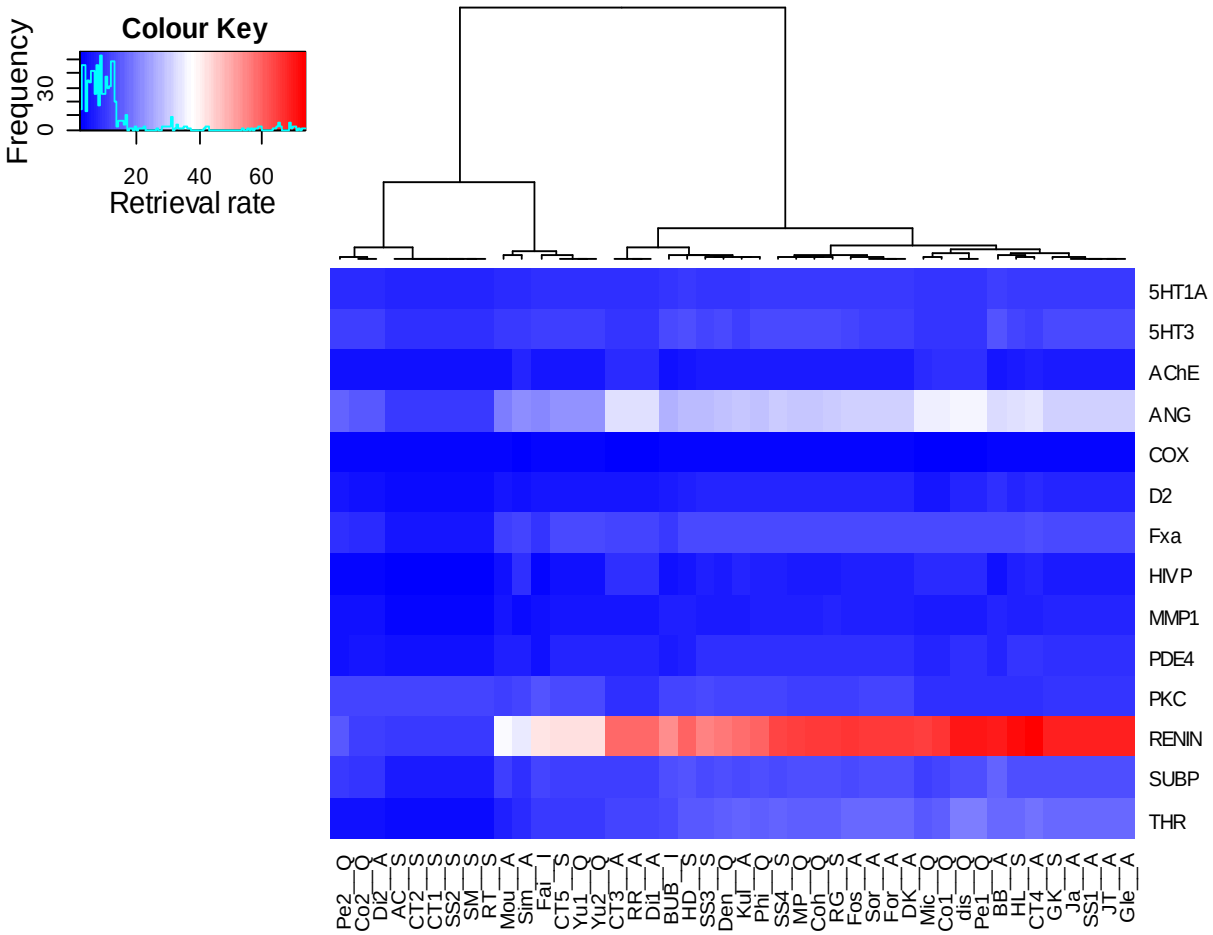


Figure 5.4 Heatmap of the retrieval rates of coefficients in WOMBAT.

The columns (coefficients) are clustered using Ward's method. Retrieval rates are coloured from red to blue which represents coefficients' retrieval abilities from strong to weak. In the left top corner legend, the X axis scales the retrieval rates where red represents higher rates and blue stands for lower rates; the Y axis measures the frequency of corresponding rates where the histographic curve shows the number of that rates.

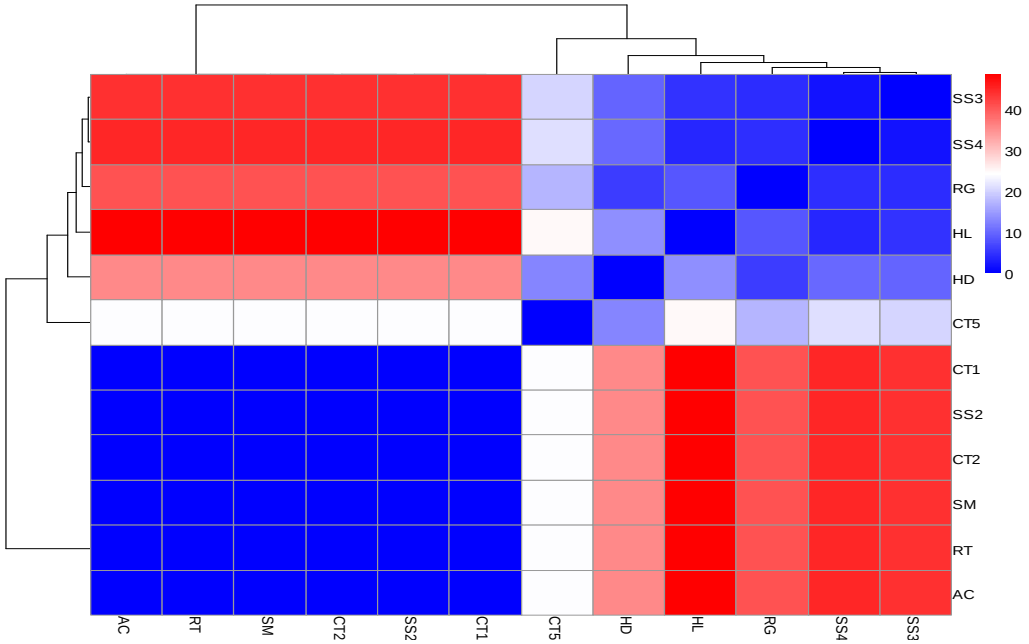


Figure 5.5 Comparison of similarity coefficients from Symmetric measures on MDDR.

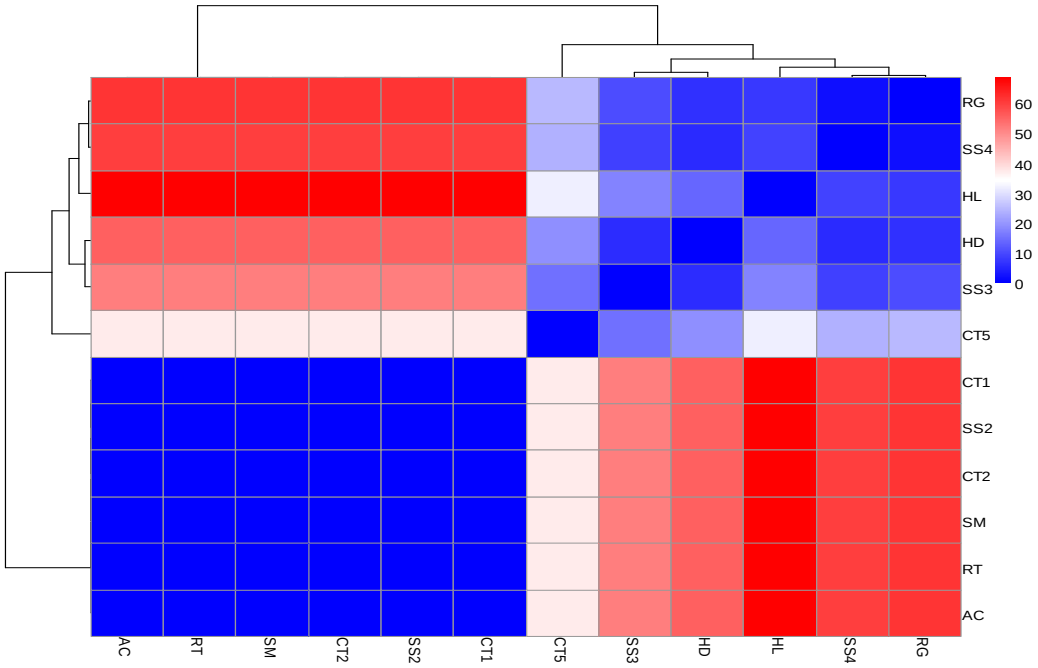


Figure 5.6 Comparison of similarity coefficients from Symmetric measures on WOMBAT.

The distances between coefficients are scaled from blue (high correlated) to red (low correlated) as shown in the right hand label.

5.5.2.3 Comparison of Coefficients based on Nature of Classes

As discussed in Section 5.5.1 and Section 5.5.2, a number of coefficients appear to be affected by the nature of activity classes. For example, in WOMBAT, coefficients *Pe1*, *dis*, *CT4* and *HL* obtained high ranks in activity classes PDE4, Fxa, ANG, RENIN and THR but not in activity classes PKC, 5HT3 and COX; coefficients *RT*, *SM*, *SS2*, *CT1*, *CT2* and *AC* worked poorly in all activity classes except PKC. One more example, asymmetric coefficients *Gle, JT*, *SS1* and *Ja* achieved good performance in class MMP1 compared to the results obtained in PKC, etc. It is also very clear that a coefficient can yield very different rank results in specific activity classes, e.g., the performance of coefficients *JT* and *Gle* in activity classes COX and PKC, both in MDDR and WOMBAT databases.

As noted in previous sections, in MDDR, the most diverse activity class COX (with MPS value 0.27) resulted in very low retrieval rate (2.89% averaged over 44 coefficients and the max retrieval rate is only 3.93%). However, the less diverse activity class Renin (with MPS value 0.57), retrieved 33.67% when averaged over all coefficients and the highest retrieval rate reached 59.16%. It is obvious that the nature of classes can affect the results of similarity search. Thus, the 11 activity classes in MDDR were divided into two groups according to their MPS values (threshold value was set at 0.40, see Chapter 3 Table 3.1), homogeneous and heterogeneous. The homogeneous group consists of activity classes AT1, HIVP, Renin, SubP and Thrombin, and the heterogeneous group contains activity classes 5HT, 5HT3, 5HT1A, COX, D2 and PKC.

Thus, in MDDR, the coefficients' retrieval rates from homogeneous classes can be ranked as:

Pe1, dis (Q 22.79%) > *Co1* (Q 22.70%) > *Mic* (Q 22.33%) > *RR, Di1, CT3* (A 22.20%) > *CT4* (A 20.07%) > *HL* (S 18.84%) > *Fos* (A 18.56%) > *Kul* (A 18.49%) > *For, DK, Sor* (A 18.34%) > *SS4* (S 17.84%) > *JT, Gle, SS1, Ja* (A 17.65%) > *GK* (S 17.54%) > *SS3* (S 17.34%) > *Phi* (Q 17.31%) > *MP* (Q 16.87%) > *Coh* (Q 16.78%) > *Den* (Q 16.75%) > *RG* (S 16.64%) > *HD* (S 15.05%) > *BB* (A 14.29%) > *BUB* (I

12.92%) > *Yu1*, *Yu2* (Q 11.97%), *CT5* (S 11.97%) > *Sim* (A 10.48%) > *Mou* (A 9.80%) > *Fai* (I 9.37%) > *Pe2* (Q 6.88%) > *Co2* (Q 6.25%), *Di2* (A 6.25%) > *SM, RT, SS2, CT1, CT2, AC* (S 4.40%)

It is obvious that the best performing coefficients yielded five-fold more retrieval rates than coefficients that ranked the last.

In heterogeneous classes, the performance of all coefficients is poor compared to their retrieval rates from homogeneous classes. The retrieval rates obtained from the heterogeneous classes are:

Den (Q 5.68%) > *Phi* (Q 5.66%), *JT, Gle , SS1, Ja* (A 5.66%) > *Coh* (Q 5.65%) > *MP* (Q 5.64%) > *RG* (S 5.60%) > *BUB* (I 5.59%) > *SS4* (S 5.58%) > *GK* (S 5.55%), *For, DK, Sor* (A 5.55%) > *Fos* (A 5.52%) > *SS3* (S 5.49%) > *BB* (A 5.47%) > *CT4* (A 5.45%) > *HL* (S 5.44%), *Kul* (A 5.44%) > *Yu1, Yu2* (Q 5.41%), *CT5* (S 5.41%) > *Pe2* (Q 5.33%) > *Co2* (Q 5.26%), *Di2*(A 5.26%) > *Mou*(A 5.22%) > *Fai* (I 5.19%) > *HD* (S 5.14%) > *dis, Pe1* (Q 4.59%) > *SM, RT, SS2, CT1, CT2, AC* (S 4.54%) > *Sim* (A 4.44%) > *Co1* (Q 4.31%) > *Mic* (Q 4.18%) > *RR, Di1, CT3* (A 3.50%)

It can be seen that the retrieval rates obtained from the four groups of coefficients are very close. There is no big difference when the coefficients are applied to heterogeneous class.

Based on the outcomes above, average retrieval rates can be calculated from different activity classes. Averaged over all coefficients, the retrieval rates of homogeneous classes and heterogeneous classes are 13.84% and 5.16%. Results obtained from four main groups of coefficients are presented in Table 5.5 (a).

In WOMBAT, 14 activity classes were grouped based on the threshold (0.40) of their MPS values. Thus, 5HT1A, ANG, HIVP, MMP1, PKC, RENIN, SUBP and THR are clustered as homogeneous classes; 5HT3, AChE, COX, D2, Fxa and PDE4 are grouped as heterogeneous classes. Averaging the retrieval rates over all coefficients, the results

for homogeneous classes and heterogeneous classes are 16.06% and 7.10%. Table 5.5 (b) gives the outcomes based on the four classes of coefficients.

From Table 5.5, in general, working on homogeneous classes, asymmetric coefficients and correlation-based coefficients obtained notably better retrieval rates than symmetric coefficients. However, there are some exceptions. For example, in MDDR, some symmetric coefficients performed better than asymmetric coefficients and correlation-based coefficients, i.e., symmetric coefficient *HL* yield an excellent retrieval rate, more than 18%, while asymmetric coefficients *Di2* and *Mou* achieved less than 10% retrieval rates, and correlation-based coefficients *CO2* and *Pe2* only achieved less than 10%. The same observations were also found in WOMBAT.

Table 5.5 Comparison of coefficients on different activity classes based on their top 1% retrieval rates.

Last row indicates the outcomes by averaged over all 44 coefficients. (a) is for MDDR; (b) is for WOMBAT.

	Homogeneous classes	***Heterogeneous classes***
Asymmetric coefficients	17.06	5.09
Symmetric coefficients	10.89	5.03
Correlation-based coefficients	16.28	5.14
Intermediate coefficients	11.14	5.39
	13.84	5.16

(a)

	Homogeneous classes	***Heterogeneous classes***
Asymmetric coefficients	18.22	7.71
Symmetric coefficients	13.22	6.46
Correlation-based coefficients	17.07	7.53
Intermediate coefficients	15.72	6.71
	16.06	7.10

(b)

5.5.3 Conclusion

The results of the initial investigation show that there are a number of coefficients which are suitable for chemical similarity search. Generally, asymmetric coefficients and correlation-based coefficients performed better than symmetric coefficients and intermediate coefficients. The hierarchical cluster analysis revealed that most of the coefficients from same class can yield similar results. The analysis based on the nature of activity classes indicates that the performance of coefficients may vary when applied to homogeneous classes, and that asymmetric coefficients yielded the best results.

However, in the initial investigation (Section 5.5.1 and 5.5.2), the three databases (the MDDR, WOMBAT and MUV databases) had been extensively used in previous studies of ligand-based virtual screening (Nasr *et al.*, 2009). Additionally, as the results shows in Table 5.3, MUV might not favorable for fingerprint-based similarity search. Given diversity of databases, validation experiments are needed, as the findings here might not be transferable to other chemical databases.

5.6 Validation Experiments

To extend the work described previously in this chapter (Section 5.5), 50 activity classes extracted from all compound data sets of ChEMBL (Heikamp and Bajorath, 2011) were employed as the dataset for validation experiments. The detail of this dataset is in Chapter 3, Section 3.2.1.4.

The experiments carried out here mirror those reported in Section 5.4.

The Kendall *W* test resulted in a value of W=0.597 and $p < 0.001$ which indicates the results are highly statistically significant. Table B.5 illustrates the top 1% retrieval rates of active compounds yielded by different coefficients. On the basis of the results, heatmaps are plotted to demonstrate the retrieval abilities of coefficients. Figure 5.7 illustrates the ranks of coefficients and Figure 5.8 demonstrates the values of retrieval

rates. The descriptions of these heatmaps can be found in Section 5.5.1 and Section 5.5.2, referring to Figure 5.1 and Figure 5.3.

After clustering, the poor performing clusters in Figure 5.7 and Figure 5.8 consist of 15 coefficients which are the same as observed in the WOMBAT dataset (Figure 5.2 and Figure 5.4) and most of the asymmetric coefficients are tightly clustered together. Compared to Figure 5.3 and Figure 5.4, Figure 5.8 shows that the variation in coefficient retrieval abilities on different activity classes is not great in ChEMBL, where the majority of coefficient retrieval rates range from 20% to 40% compared to 10% in Figure 5.3 (MDDR) and 15% in Figure 5.4 (WOMBAT).

Six symmetric coefficients, *HL*, *GK*, *SS4*, *SS3*, *RG* and *HD* still achieved relatively good results in ChEMBL, and were grouped in a high performance cluster. The other seven symmetric coefficients were continuously ranked last, as with their performance in MDDR and WOMBAT. Figure 5.9 shows the correlations of symmetric coefficients based on their retrieval rates in 50 activity classes. It gives similar results to those of Figure 5.5 and Figure 5.6.

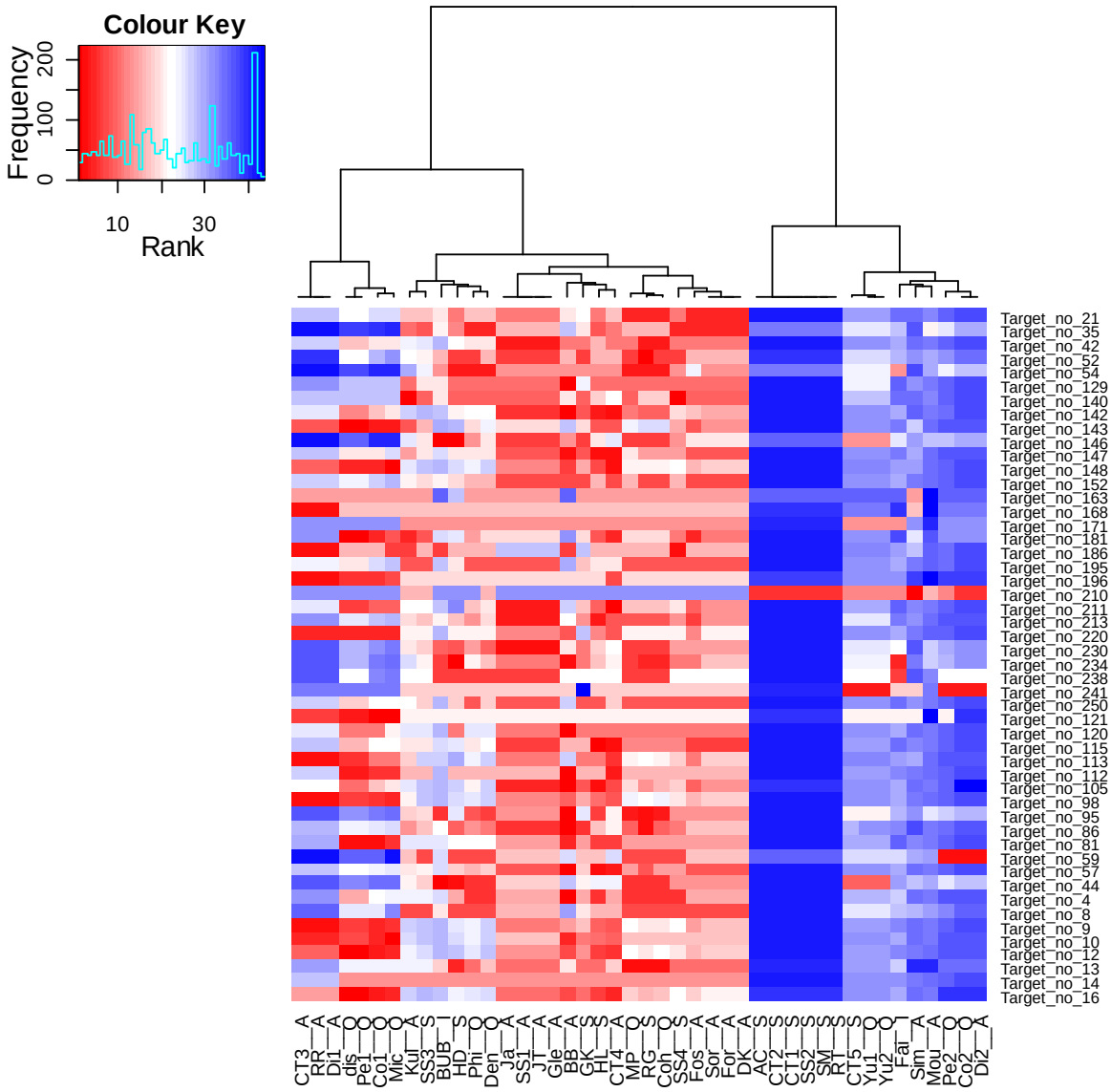


Figure 5.7 Heatmap of the ranks of coefficients in ChEMBL.

The columns (coefficients) are clustered using Ward's method. Ranks are coloured from red to blue which represents coefficients' retrieval abilities from strong to weak. In the left top corner legend, the X axis scales the ranks where red represents higher positions and blue denotes lower positions; the Y axis measures the frequency of corresponding ranks where the histographic curve shows the number of the ranks.

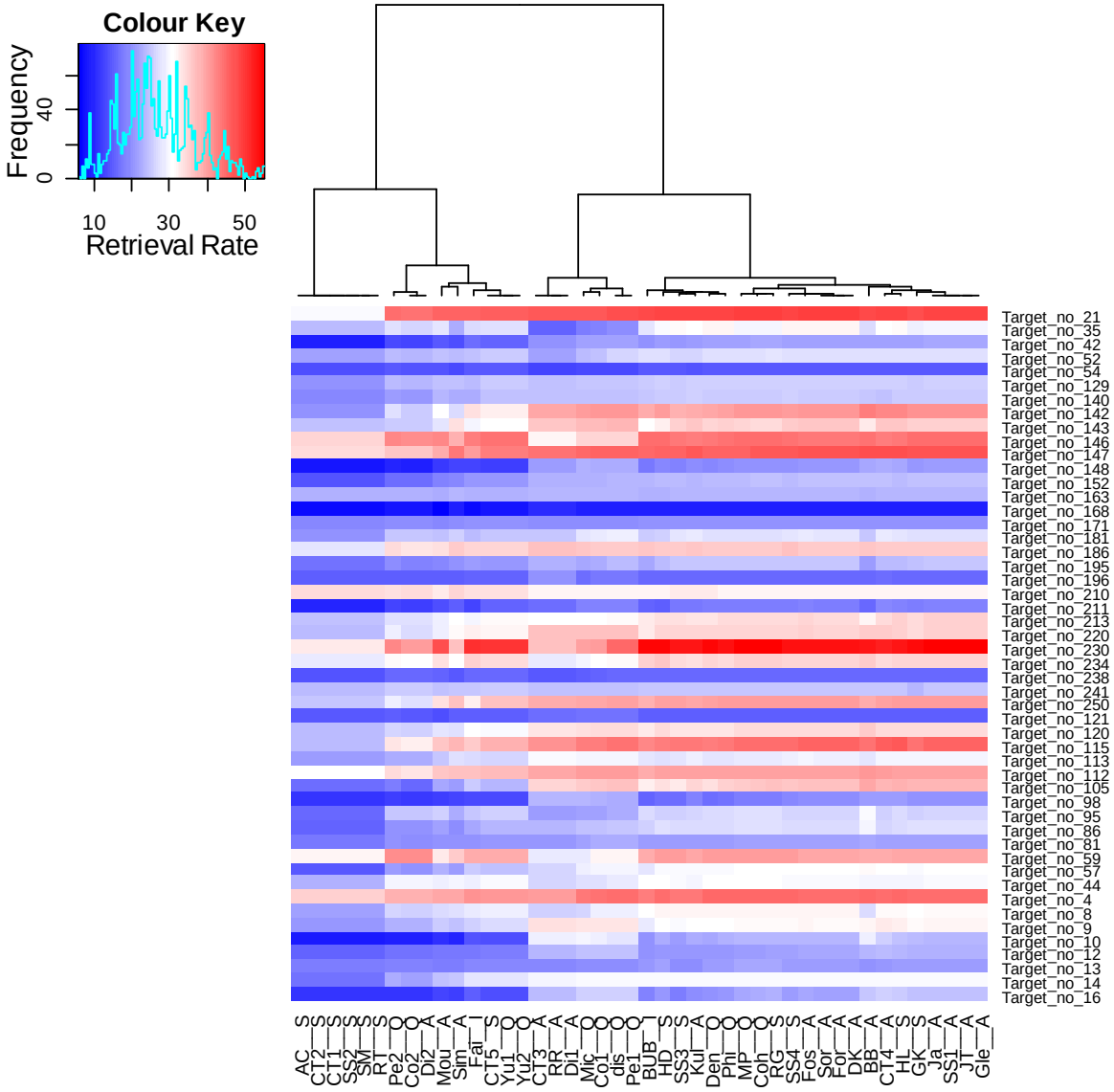


Figure 5.8 Heatmap of the retrieval rates of coefficients in ChEMBL.

The columns (coefficients) are clustered using Ward's method. Retrieval rates are coloured from red to blue which represents coefficients' retrieval abilities from strong to weak. In the left top corner legend, the X axis scales the retrieval rates where red represents higher rates and blue stands for lower rates; the Y axis measures the frequency of corresponding rates where the histographic curve shows the number of that rates..

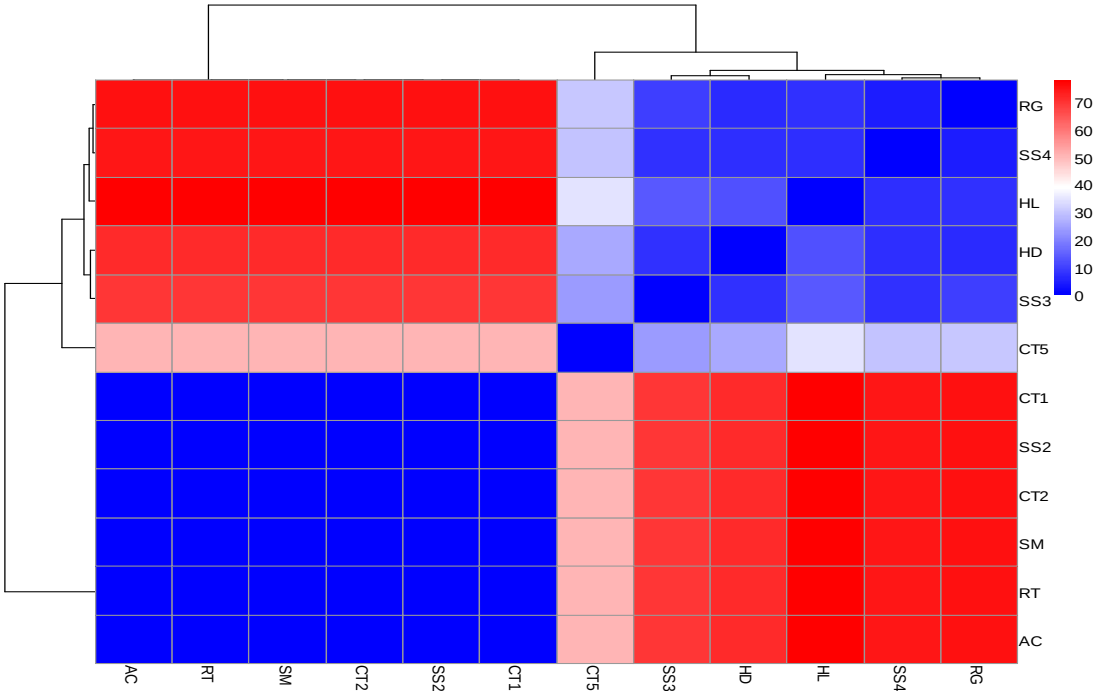


Figure 5.9 Correlations of symmetric coefficients in 50 activity classes of ChEMBL.

The distances between coefficients are scaled from blue (high correlated) to red (low correlated) as shown in the right hand label.

Figure 5.8 shows the activity classes in ChEMBL are less diverse as most of their retrieval rates are between 20% and 40%. Averaged over all 50 activity classes, the percentages of retrieval rates yielded by four classes of coefficients are: asymmetric coefficients (28.33%) > correlation-based coefficients (27.68%) > intermediate coefficients (27.19%) > symmetric coefficients (25.11%). Symmetric *coefficients RT, SM, SS2, CT1, CT2* and *AC* gave worse outcomes among the 44 coefficients. Checked against the MPS values of 50 activity classes in Chapter 3 Table 3.4 the values are in range of 0.33 to 0.53. Therefore, the 50 activity classes can also be divided into two groups for further comparison. Based on the threshold (0.42) of their MPS values, 50 activity classes are grouped as 25 homogeneous classes and 25 heterogeneous classes.

Table 5.6 Comparison of coefficients on different activity classes based on their top 1% retrieval rates in ChEMBL.

Last row indicates the outcomes by averaged over all 44 coefficients.

	Homogeneous class	***Heterogeneous class***
Asymmetric coefficients	32.21	24.44
Symmetric coefficients	28.13	22.08
Correlation-based coefficients	31.36	24.01
Intermediate coefficients	30.84	23.54
	30.64	23.52

Shown as Table 5.6, the finding is in line with the observations on MDDR and WOMBAT in which asymmetric coefficients and correlation-based coefficients yielded more actives than did symmetric coefficients in homogeneous classes, but this is less notable due to the lower diversity of the activity classes in ChEMBL. In heterogeneous classes, four classes of coefficients worked quite equally.

Table 5.7 presents the rank positions of the 44 coefficients based on their retrieval rates and ranks averaged over all the 50 activity classes. Results from this table can be compared with the outcomes in Table 5.4 which shows that a number of coefficients have as good retrieval ability as the *JT* (Tanimoto) coefficient, e.g., coefficients CT4, HL, Fos, etc. Details of the results are presented in Table 5.8, where Min., Max., Median, 1st Qu., 3st Qu., Mean refer to the minimum, maximum, median, first quartile, third quartile and mean value of retrieval rates over the 50 activity classes, respectively.

The results obtained from all of the three databases show that coefficients can yield identical similarity ranking with other coefficients, with different similarity values. These pair/group coefficients are defined as being monotonic to each other (Gasteiger and Engel, 2003; Leach and Gillet, 2007; Willett *et al.*, 1998). For example, coefficients *JT*, *Ja*, *SS1* and *Gle* are monotonic to each other.

Table 5.7 Rank positions of the 44 coefficients when averaged over all of the activity classes in the ChEMBL dataset.

The second and third columns present the rank positons based on the average retrieval rates; the last two columns provide the rank positions according to the average ranks. Coefficients' class are shown in brackets after their names.

Rank	Average retrieval rates		Average ranks	
1	*HL (S)*	29.85%	*JT, Gle, SS1, Ja (A)*	12.20
2	*CT4 (A)*	29.82%		
3	*JT, Gle, SS1, Ja (A)*	29.74%		
5			*HL (S)*	12.39
6			*RG (S)*	12.58
7	*Fos (A)*	29.64%	*CT4 (A)*	12.80
8	*RG (S)*	29.57%	*Coh (Q)*	13.08
9	*For , DK, Sor (A)*	29.56%	*Fos (A)*	13.34
10			*MP (Q)*	13.55
11	*BB (A)*	29.54%	*SS4 (S)*	14.30
12	*Coh (Q)*	29.53%	*For , DK, Sor (A)*	14.31
13	*SS4 (S)*	29.51%		
14	*MP (Q)*	29.50%	*BB (A)*	14.66
15	*GK (S)*	29.49%	*GK (S)*	15.89
16	*Phi (Q)*	29.26%	*Phi (Q)*	15.99
17	*Den (Q)*	29.14%	*Den (Q)*	17.03
18	*Kul (A)*	29.07%	*HD (S)*	17.35
19	*HD (S)*	29.04%	*dis, Pe1 (Q)*	17.90
20	*SS3 (S)*	28.88%		
21	*dis, Pe1 (Q)*	28.79%	*Kul (A)*	18.83
22			*SS3 (S)*	19.65
23	*Co1 (Q)*	28.42%	*Co1 (Q)*	19.84
24	*BUB (I)*	28.29%	*Mic (Q)*	20.84
25	*Mic (Q)*	28.16%	*BUB (I)*	21.80
26	*RR, Di1, CT3 (A)*	27.43%	*RR, Di1, CT3 (A)*	24.41
30	*CT5 (S), Yu1, Yu2 (Q)*	26.43%	*CT5 (S)*	27.58
33	*Fai (I)*	26.09%	*Yu1, Yu2 (Q)*	27.64
34	*Sim (A)*	24.97%	*Fai (I)*	28.79
35	*Mou (A)*	24.94%	*Sim (A)*	31.37
36	*Pe2 (Q)*	24.11%	*Pe2 (Q)*	32.07
37	*Co2 (Q) , Di2 (A)*	23.65%	*Co2 (Q), Di2 (A)*	33.70
38				
39	*SM, RT,SS2,CT1, CT2, AC (S)*	20.60%	*Mou (A)*	33.79
40			*SM, RT,SS2,CT1, CT2, AC (S)*	39.84

Some coefficients can be mathematically identical under certain conditions, i.e., the rescaled *GK* coefficient is identical to the *Gle* coefficient when satisfied $d>a$, where,

$$S'_{GK} = \left[\frac{2 \cdot min(a,d) - b - c}{2 \cdot min(a,d) + b + c} + 1\right] \cdot \frac{1}{2} = \left[\frac{2a - b - c}{2a + b + c} + 1\right] \cdot \frac{1}{2} = \left[\frac{4a}{2a + b + c}\right] \cdot \frac{1}{2} = \frac{2a}{2a + b + c}$$
$$= S_{Gle}$$

the *Sim* coefficient is identical to the *Di1* coefficient if $b<c$, thereby,

$$S_{Sim} = \frac{a}{min\{(a+b),(a+c)\}} = \frac{a}{(a+b)} = S_{Di1}$$

one more example, the *BB* coefficients is identical to the *Di2* coefficient when $b>c$, thus,

$$S_{BB} = \frac{a}{max\{(a+b),(a+c)\}} = \frac{a}{(a+c)} = S_{Di2}$$

It is apparent from the above transformations that although the *GK* coefficient was defined as a symmetric coefficient due to the fact that quantity *d* is included, however, it can be changed to an asymmetric coefficient when the number of common occurrence elements is less than the number of elements that did not occur in both objects. In chemical similarity search, quantity *d* normally obtained a higher value than quantity *a*, but it is not always the case, thus, the *GK* coefficient and the *Gle* coefficient are not monotonic to each other. In addition, the transformation indicates the *GK* coefficient's relatively high achievement as the good performance of the *Gle* coefficient.

In some cases, two coefficients may not be completely monotonic but may give highly correlated similarity results. The Wilcoxon signed rank test (as described in Section 3.4.2) was employed here to compare coefficients which produced the best results in Table 5.8, i.e., *JT*, *Gle*, *For*, *BB*, *DK*, *Kul*, *Ja*, *RG*, *Fos*, *GK*, *SS4*, *MP*, *HL*, *SS1* and *CT4*. For each coefficient, 50 results of retrieval rate were applied to do head-to-head comparisons.

Table 5.9 demonstrates the *p*-values of the Wilcoxon signed rank test. The *p*-values are bolded and italic if they are less than the 0.05 significance level. Since coefficients *JT*,

Ja, *SS1* and *Gle* are monotonic to each other and the *For, DK* and *Sor* coefficients are monotonic to each other, the *p*-values of the pairs of them were marked as NA. At 0.05 significance level, it can be concluded that the retrieval rates yield of most of the coefficients from ChEMBL data set are highly correlated, e.g., *JT* with *RG, Fos*, *BB*, *HL* and *CT4.*

According to the analysis data above, it is clear that coefficient *CT4* and *HL* are as good as *JT*, the Tanimoto coefficient. Moreover, many other coefficients can give similar performance to *JT*. The finding here confirmed the assumption in the beginning of this chapter that there exist other coefficients which are suitable for similarity search in virtual screening.

Table 5.8 Coefficients analysis of the retrieval rates in 50 activity classes of ChEMBL.

	SM	*RT*	*JT*	*Gle*	*RR*	*For*	*Sim*	*BB*	*DK*	*BUB*	*Kul*
Min.	7.00	7.00	9.00	9.00	10.00	9.00	9.00	9.00	9.00	9.00	9.00
1st Qu.	15.40	15.40	23.27	23.27	21.38	21.86	18.31	22.43	21.86	20.34	20.94
Median	20.29	20.29	28.46	28.46	25.61	28.47	24.22	28.52	28.47	27.63	28.76
Mean	20.60	20.60	29.74	29.74	27.43	29.56	24.97	29.54	29.56	28.29	29.07
3rd Qu.	24.38	24.38	35.40	35.40	33.69	35.83	31.05	35.82	35.83	33.08	35.00
Max.	35.53	35.53	55.42	55.42	46.63	53.61	45.67	53.01	53.61	54.52	53.01

	SS1	*SS2*	*Ja*	*Fai*	*Mou*	*Mic*	*RG*	*HD*	*Yu1*	*Yu2*	*Fos*
Min.	9.00	7.00	9.00	7.00	6.00	9.00	9.00	9.00	8.00	8.00	9.00
1st Qu.	23.27	15.40	23.27	19.42	18.13	22.15	22.33	21.27	19.14	19.14	22.03
Median	28.46	20.29	28.46	25.23	24.47	27.14	28.23	27.66	25.40	25.40	28.89
Mean	29.74	20.60	29.74	26.09	24.94	28.16	29.57	29.04	26.43	26.43	29.64
3rd Qu.	35.40	24.38	35.40	32.49	30.23	34.48	35.34	34.81	32.38	32.38	35.94
Max.	55.42	35.53	55.42	50.90	47.89	46.88	55.12	54.82	50.00	50.00	53.61

	Den	*Co1*	*Co2*	*Dis*	*GK*	*SS3*	*SS4*	*Phi*	*Di1*	*Di2*	*Sor*
Min.	9.00	9.00	8.00	9.00	9.00	9.00	9.00	9.00	10.00	8.00	9.00
1st Qu.	21.20	22.50	16.61	22.61	23.29	20.60	21.56	21.34	21.38	16.61	21.86
Median	27.64	28.05	23.26	28.28	28.33	28.08	28.47	27.93	25.61	23.26	28.47
Mean	29.14	28.42	23.65	28.79	29.49	28.88	29.51	29.26	27.43	23.65	29.56
3rd Qu.	34.84	34.57	27.42	34.81	35.23	34.90	35.60	35.19	33.69	27.42	35.83
Max.	54.52	46.88	44.23	47.60	54.22	53.31	53.61	53.61	46.63	44.23	53.61

	Coh	*Pe1*	*Pe2*	*MP*	*HL*	*CT1*	*CT2*	*CT3*	*CT4*	*CT5*	*AC*
Min.	9.00	9.00	8.00	9.00	9.00	7.00	7.00	10.00	9.00	8.00	7.00
1st Qu.	22.21	22.61	17.40	21.99	23.64	15.40	15.40	21.38	23.67	19.14	15.40
Median	28.20	28.28	23.61	28.20	28.75	20.29	20.29	25.61	29.03	25.40	20.29
Mean	29.53	28.79	24.11	29.50	29.85	20.60	20.60	27.43	29.82	26.43	20.60
3rd Qu.	35.34	34.81	29.14	35.23	35.94	24.38	24.38	33.69	36.04	32.38	24.38
Max.	55.12	47.60	44.71	55.12	53.31	35.53	35.53	46.63	52.11	50.00	35.53

Table 5.9 The *p*-values of Wilcoxon signed rank test. All the p-values that turned out to be less than the 0.05 significance level were bolded and italic.

	JT	*Gle*	*For*	*BB*	*DK*	*Sor*	*Kul*	*Ja*	*RG*	*Fos*	*GK*	*SS4*	*MP*	*HL*	*CT4*	*SS1*
JT																
Gle	NA															
For	***0.0335***	***0.0335***														
BB	0.5087	0.5087	0.9076													
DK	***0.0335***	***0.0335***	NA	0.9076												
Sor	***0.0335***	***0.0335***	NA	0.9076	NA											
Kul	***0.0003***	***0.0003***	***<0.0001***	0.1025	***<0.0001***	***<0.0001***										
Ja	NA	NA	***0.0335***	0.5087	***0.0335***	***0.0335***	***0.0003***									
RG	0.0603	0.0603	0.5925	0.9005	0.5925	0.5925	***0.0011***	0.0603								
Fos	0.1573	0.1573	***0.0029***	0.8847	***0.0029***	***0.0029***	***<0.0001***	0.1573	0.7421							
GK	***0.0001***	***0.0001***	0.4456	0.8179	0.4456	0.4456	***0.0177***	***0.0001***	0.2619	0.1145						
SS4	***0.0421***	***0.0421***	0.2989	0.9166	0.2989	0.2989	***<0.0001***	***0.0421***	0.2231	***0.0223***	0.6335					
MP	***0.0421***	***0.0421***	0.9655	0.9502	0.9655	0.9655	***0.0015***	***0.0421***	***0.0418***	0.2429	0.4010	0.7672				
HL	0.2329	0.2329	***0.0146***	0.3050	***0.0146***	***0.0146***	***<0.0001***	0.2329	0.0870	0.0774	***0.0008***	***0.0425***	0.0678			
CT4	0.3654	0.3654	0.1325	0.4006	0.1325	0.1325	***0.0007***	0.3654	0.2232	0.3911	***0.0126***	0.1683	0.1315	0.5115		
SS1	NA	NA	***0.0335***	0.5087	***0.0335***	***0.0335***	***0.0003***	NA	0.0603	0.1573	***0.0001***	***0.0421***	***0.0421***	0.2329	0.3654	

5.7 Conclusion

In this chapter, 44 binary similarity coefficients were analysed by an extensive comparison of their retrieval abilities in similarity-based virtual screening, both by comparison of their ranks in each class and their retrieval rates of active compounds. The Ward's method was used to cluster coefficients based on their retrieval rates.

There are four main findings. First, in heterogeneous activity classes, the performances of coefficients from different classes were often similar. Second, the asymmetric coefficients and the correlation-based coefficients worked very well with less diverse activity classes. Third, working on homogeneous classes, a number of coefficients performed better than the Tanimoto coefficient which is the conventional coefficient in similarity search, i.e., correlation-based coefficients *Pe1*, *dis*, *Co1* and *Mic* in MDDR and correlation-based coefficients *Pe1* and *dis* in WOMBAT. Finally, the symmetric coefficient *HL*, and the new asymmetric coefficient *CT4* yielded very good results in both homogeneous and heterogeneous classes and performed superior to the Tanimoto coefficient in WOMBAT and ChEMBL. Therefore, it can be suggested that when working with un-weighted fingerprints, if the characters of the structures are unknown, then coefficients *HL* and *CT4* might be appropriate for similarity-based virtual screening instead of the standard Tanimoto coefficient.

In published work (Todeschini *et al.*, 2012), it has been shown that it is generally more important in chemoinformatic applications to take account of the properties that are present rather than those which are absent. It also suggested several coefficients may be worthy of further study for applications in chemoinformatics. In this study, in addition to the small-scale similarity results, the results of the performed experiments showed that a number of coefficients had better/similar retrieval ability to the Tanimoto coefficient. It is also possible to apply these suggested coefficients, e.g., the *HL* coefficient and the *CT4* coefficient, to large-scale virtual screening. In addition, data fusion approach can be adopted to optimise similarity search using selected high performing coefficients.

Chapter 6: Comparison of Similarity Coefficients using Weighted Chemical Data

6.1 Introduction

In the previous two chapters, investigations were carried out to evaluate interactions between weighting schemes and similarity coefficients in similarity-based virtual screening (Chapter 4) and the comparison of binary coefficients for chemical similarity search (Chapter 5). The results have shown that both weighting schemes and the choice of similarity coefficient can affect similarity search. In addition, a number of binary coefficients were detected which exhibit better retrieval abilities than the Tanimoto coefficient in similarity search. Thus, in this chapter, the emphasis is on comparing the performance of coefficients when applied to weighted data, and on identifying coefficients which provide consistently high performance when different weighting schemes are involved.

6.2 Selection of Coefficients

In Chapter 5, 44 similarity coefficients were compared. These similarity coefficients are specific for dichotomic (binary) variables, and are based on comparisons between co-occurring elements and non-co-occurring elements. The results demonstrate the importance of similarity coefficient choice, so that the most effective ones in each specific situation can be employed. Among these coefficients, many have been widely utilized for different applications. The majority of studies, however, tend not to justify their preferences for any one particular model. The most common coefficient, the

Tanimoto/Jaccard coefficient, has been investigated and compared with different coefficients in many chemoinformatics studies (Chen *et al.*, 2009; Chen and Reynolds, 2002; Duarte *et al.*, 1999; Holliday *et al.*, 2002; Salim *et al.*, 2003; Sesli and Yegenoglu, 2010; Snijders *et al.*, 1990; Whittle *et al.*, 2003). Holliday *et al.* (2002) compared the performance of thirteen coefficients in similarity searches of chemical databases. In subsequent research, data fusion techniques were applied to investigate the combinations of coefficients as well as their relative (Holliday *et al.*, 2002) individual performance (Chen *et al.*, 2009; Salim *et al.*, 2003). Their studies illustrated that some other coefficients may be less affected by the compound bit-density that occurs with the Tanimoto coefficient. Their results also indicated that combining coefficients does improve the performance of similarity searches when compared with the use of a single measure, in particular the industry standard Tanimoto measure. They also concluded that no single combination showed a consistently high performance across all types of activities.

As well as studies in chemoinformatics, comparisons of similarity coefficients have also been carried out in studies of genetic divergence. Meyer et al. (2004) compared eight coefficients to evaluate whether different similarity coefficients used with dominant markers can influence the results of cluster analysis. Their results revealed that the Anderberg (Anderberg, 1973), Sorensen-Dice (Sørensen, 1948) and the Tanimoto/Jaccard coefficient had almost identical results. These two coefficients were also studied and provided identical results to the Tanimoto coefficient in Chapter 5, namely, the *SS1* coefficient and the *Gle* coefficient. Meyer *et al.* (2004) also proposed that the choice of similarity coefficient can be based on excluding the negative co-occurred properties in the similarity measure, i.e., the symmetric coefficients in Chapter 5.

The results of above studies are in agreement with the findings in Chapter 5 that coefficients which do not consider the negative co-occurrences (parameter *d*) generally performed better. Furthermore, there were other coefficients performed better than the Tanimoto coefficient. According to the results in Chapter 4, showing that the choice of

weighting scheme can enhance similarity search, it seems reasonable to apply the high-achieving coefficients in Chapter 5 to the weighted data.

A few studies, however, have reported the alternative forms of binary coefficients that can be used to quantify the degree of similarity between non-binary data (Ellis *et al.*, 1993; Holliday *et al.*, 2009; Whittle *et al.*, 2003). Ellis *et al.* (1993) reviewed 27 similarity coefficients applied to binary fingerprints and, in all but ten coefficients, these have an equivalent non-binary form. Among the remaining 17 non-binary coefficients, three coefficients arguably cannot produce appropriate rankings because the self-similarity value is not always the highest, i.e., two molecules, A and B can be more similar to each other, than each is to itself, when a non-binary fingerprint is used. These are coefficients Russell-Rao, Kulczynski(2) and Forbes (Whittle *et al.*, 2003).

More recently, Al Khalifa *et al.* (2009) investigated the effects of applying similarity coefficients to a set of compounds from MDDR. These compounds were characterized by 378 real-valued structure-based property descriptors. They measured 12 coefficients, from those by Ellis *et al.* (1993) and Whittle *et al.*(2003). Al Khalifa *et al.* (2009) concluded that there is no single coefficient which worked consistently well across all methodologies. The Tanimoto, Dice, Kulczynski(1), and Sokal/Sneath(1) coefficients, which were found to show near identical performance when applied to similarity searches using binary fingerprints, exhibited good performances when applied on continuous descriptors. Two other coefficients found to be near identical in the binary case, the Cosine and Fossum, also achieved good results on non-binary descriptors. Al Khalifa *et al.* concluded that the Cosine and Fossum coefficients would be more favorable for nonhierarchical clustering. Their study, however, was carried out on a small subset containing just 20,000 compounds from the MDDR database. Their conclusion would be enhanced by the use of more appropriate databases and alternative characterizations.

In the following sections, selected similarity coefficients are tested to determine whether coefficients that work well with binary fingerprints can function effectively when applied to weighted fingerprints.

6.2.1 Identification of Coefficients

In Chapter 5, 44 similarity coefficients were compared. Apart from the results achieved in MUV, coefficients were ranked in decreasing order by their retrieval abilities in each dataset. The top 22 (50%) of these are listed in Table 6.1. The shaded cells indicate the coefficients ranked in the top 22 in all three databases.

Table 6.1. Coefficients ranked in top 50% among databases using binary fingerprints.

	MDDR	WOMBAT	ChEMBL
1	***dis***	***CT4***	***HL***
2	*Pe1*	***dis***	***CT4***
3	*Co1*	***Pe1***	***JT***
4	*Mic*	***HL***	***Gle***
5	***CT4***	*BB*	***SS1***
6	*RR*	***JT***	***Ja***
7	*Di1*	***Gle***	***Fos***
8	*CT3*	***SS1***	*RG*
9	***HL***	***Ja***	***For***
10	***Fos***	***GK***	***DK***
11	***Kul***	***Fos***	***Sor***
12	***For***	***For***	*BB*
13	***DK***	***DK***	*Coh*
14	***Sor***	***Sor***	***SS4***
15	***SS4***	*RG*	*MP*
16	***JT***	*Co1*	***GK***
17	***Gle***	***SS4***	***Phi***
18	***SS1***	*Coh*	*Den*
19	***Ja***	*MP*	***Kul***
20	***GK***	*Mic*	*HD*
21	***Phi***	***Phi***	*SS3*
22	*SS3*	***Kul***	***dis***

As shown in Table 6.1, 15 coefficients were identified. Based on the studies of Ellis *et al.* (1993), nine coefficients' formulae for continuous variables can be verified. Two of these coefficients, the *DK* coefficient and the *Sor* coefficient, are monotonic to each other, i.e., they produce identical similarity rankings against a specified target, even though their similarity values are different. Thus, only one of them was studied in this chapter, i.e., the *DK* coefficient.

Two other coefficients of these nine were previously described: the Forbes coefficient and the Kulczynski(2) coefficient. Ellis *et al.* (1993) argued, however, that the above cited coefficients cannot provide appropriate rankings with continuous representation (Whittle *et al.*, 2003). The Forbes coefficient, however, even working on binary data, was regarded as not a suitable coefficient for calculating similarity degree (Cole, 1957; Michael, 1920). Nevertheless, in chemoinformatics, it was found to be effective in similarity search (Holliday *et al.*, 2003; Salim *et al.*, 2003; Willett, 2006). According to Michael's deduction (1920), another coefficient, the Fossum coefficient, which is very similar to the Forbes coefficient, is considered unsuitable for measuring the degree of similarity with non-binary descriptors. Al Khalifa *et al.* (2009), however, showed that the Fossum coefficient performed well in similarity search and clustering as a non-binary coefficient. Motivated by the success for binary fingerprints, the two debatable coefficients, the Fossum and Forbes coefficients, have hence been included in the coefficients list for subsequent study.

The *CT4* coefficient, which was derived by applying the logarithm transformation to the Tanimoto coefficient (Consonni and Todeschini, 2012), can be transformed to continuous format based on the original formula applied to binary data. The *Ja* (Jaccard) coefficient, which is not listed in Ellis *et al*.'s (1993) report, can also be transformed to non-binary form by analogy with the formula of the *JT* coefficient.

Thus far, 10 non-binary similarity coefficients forms have been identified. The formulas are shown in Table 6.2.

Table 6.2. Ten Non-binary similarity coefficients.

x_{iA} is the value of a descriptor in compound A at attribute i; x_{iB} the value of a descriptor in compound B at attribute i; n is the total number of descriptors used for each compound.

Name	Formula for continuous variables	Formula for binary variables
JT (Jaccard/ Tanimoto)	$S_{AB} = \frac{\sum_{i=1}^{n} x_{iA} \cdot x_{iB}}{\sum_{i=1}^{n}(x_{iA})^2 + \sum_{i=1}^{n}(x_{iB})^2 - \sum_{i=1}^{n} x_{iA} \cdot x_{iB}}$	$S_{AB} = \frac{a}{a+b+c}$
DK (Driver-Kroeber /cosine)	$S_{AB} = \frac{\sum_{i=1}^{n} x_{iA} \cdot x_{iB}}{\sqrt{\sum_{i=1}^{n}(x_{iA})^2 \cdot \sum_{i=1}^{n}(x_{iB})^2}}$	$S_{AB} = \frac{a}{\sqrt{(a+b)(a+c)}}$
Gle (Gleason /Dice)	$S_{AB} = \frac{2\sum_{i=1}^{n} x_{iA} \cdot x_{iB}}{\sum_{i=1}^{n}(x_{iA})^2 + \sum_{i=1}^{n}(x_{iB})^2}$	$S_{AB} = \frac{2a}{2a+b+c}$
For (Forbes)	$S_{AB} = \frac{n\sum_{i=1}^{n} x_{iA} \cdot x_{iB}}{\sum_{i=1}^{n}(x_{iA})^2 \cdot \sum_{i=1}^{n}(x_{iB})^2}$	$S_{AB} = \frac{na}{(a+b)(a+c)}$
Kul (Kulczynski (2))	$S_{AB} = \frac{\sum_{i=1}^{n} x_{iA} \cdot x_{iB} \cdot (\sum_{i=1}^{n}(x_{iA})^2 + \sum_{i=1}^{n}(x_{iB})^2)}{2\,(\sum_{i=1}^{n}(x_{iA})^2 \cdot \sum_{i=1}^{n}(x_{iB})^2)}$	$S_{AB} = \frac{1}{2} \cdot [\frac{a}{a+b} + \frac{a}{a+c}]$
SS1 (Sokal-Sneath)	$S_{AB} = \frac{\sum_{i=1}^{n} x_{iA} \cdot x_{iB}}{2\sum_{i=1}^{n}(x_{iA})^2 + 2\sum_{i=1}^{n}(x_{iB})^2 - 3\sum_{i=1}^{n} x_{iA} \cdot x_{iB}}$	$S_{AB} = \frac{a}{a+2b+2c}$
Fos (Fossum)	$S_{AB} = \frac{n(\sum_{i=1}^{n} x_{iA} \cdot x_{iB} - 0.5)^2}{\sum_{i=1}^{n}(x_{iA})^2 \cdot \sum_{i=1}^{n}(x_{iB})^2}$	$S_{AB} = \frac{n \cdot (a-0.5)^2}{(a+b)(a+c)}$
Phi (Pearson)	$S_{AB} = \frac{\sum_{i=1}^{n}(x_{iA} - \overline{x_A})(x_{iB} - \overline{x_B})}{\sqrt{\sum_{i=1}^{n}(x_{iA} - \overline{x_A})^2 \cdot \sum_{i=1}^{n}(x_{iB} - \overline{x_B})^2}}$	$S_{AB} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$
CT4 (Consonni - Todeschini)	$S_{AB} = \frac{\ln(1 + \sum_{i=1}^{n} x_{iA} \cdot x_{iB})}{\ln(1 + \sum_{i=1}^{n}(x_{iA})^2 + \sum_{i=1}^{n}(x_{iB})^2 - \sum_{i=1}^{n} x_{iA} \cdot x_{iB})}$	$S_{AB} = \frac{ln(1+a)}{ln(1+a+b+c)}$
Ja (Jaccard)	$S_{AB} = \frac{3\sum_{i=1}^{n} x_{iA} \cdot x_{iB}}{\sum_{i=1}^{n}(x_{iA})^2 + \sum_{i=1}^{n}(x_{iB})^2 + \sum_{i=1}^{n} x_{iA} \cdot x_{iB}}$	$S_{AB} = \frac{3a}{3a+b+c}$

6.2.2 More Coefficients

Apart from the 44 coefficients in Chapter 5, a large number of similarity and distance coefficients have been defined and extensively used in different domains. Many of them are closely related to each other or identical, e.g., they have been discovered and rediscovered by different authors. A number of coefficients, which are equivalent when applied to binary variables, generate different results when applied to continuous variables. A number of coefficients are defined as being monotonic to each other. In some cases, two coefficients may not be completely monotonic and may give highly correlated rankings. Hence, the high performing coefficients that give the same or similar performance on binary data need to be considered in this study.

Even though numerous binary similarity measures have been described in the literature, only a few comparative studies have collected a wide variety of binary similarity measures. For example, Cha (2008) enumerated 45 coefficients classified in seven groups. Out of these 45, a number of distance coefficients were reported as giving similar results to the Jaccard/Tanimoto coefficient, e.g., the Sorensen distance. This coefficient has been widely used in ecology (Looman and Campbell, 1960). A number of coefficients have distance forms which are identical to others, e.g., the distance form of the Czekanowski coefficient (Campbell, 1978) is identical to the Sorensen coefficient (Sørensen, 1948). Several coefficients are proportional to others, e.g., half of the Czekanowski coefficient is called the Motyka similarity or is known as the Kulczynski similarity. A number of coefficients have different names, e.g., the cosine coefficient has other names including the Ochiai coefficient and the Carbo coefficient; The Dice coefficient is occasionally called Sorensen, Czekanowski, Hodgkin-Richards or Morisita.

More recently, Choi *et al.* (2010) reviewed 76 binary similarity and distance measures used over the last century. The definitions of binary similarity and distance measures are expressed by Operational Taxonomic Units in a 2 x 2 contingency table. The coefficients were analysed and classified through hierarchical clustering, and the close relationships among several of the measures were observed.

Based on these studies, several more coefficients can be considered, since they exhibited identical or very similar performance with the Jaccard/Tanimoto coefficient in binary case. They are listed in Table 6.3.

Table 6.3 Additional Non-binary similarity coefficients.

x_{iA} is the value of a descriptor in compound A at attribute i; x_{iB} the value of a descriptor in compound B at attribute i; n is the total number of descriptors used for each compound.

Name	Formula for continuous variables	Formula for binary variables
Soe (Soergel)	$D_{AB} = \frac{\sum_{i=1}^{n}\lvert x_{iA} - x_{iB}\rvert}{\sum_{i=1}^{n} max(x_{iA}, x_{iB})}$	$D_{AB} = \frac{b+c}{a+b+c}$
MR (MinMax /Ruzicka)	$S_{AB} = \frac{\sum_{i=1}^{n} min(x_{iA}, x_{iB})}{\sum_{i=1}^{n} max(x_{iA}, x_{iB})}$	$S_{AB} = \frac{a}{a+b+c}$
Hel (Hellinger)	$D_{AB} = 2\sqrt{(1 - \frac{\sum_{i=1}^{n} x_{iA} \cdot x_{iB}}{\sqrt{\sum_{i=1}^{n}(x_{iA})^2 \sum_{i=1}^{n}(x_{iB})^2}})}$	$D_{AB} = 2\sqrt{(1 - \frac{a}{\sqrt{(a+b)(a+c)}})}$
Cze (Czekanowski)	$S_{AB} = \frac{2\sum_{i=1}^{n} min(x_{iA}, x_{iB})}{\sum_{i=1}^{n}(x_{iA} + x_{iB})}$	$S_{AB} = \frac{2a}{2a+b+c}$

In Table 6.3, the *Soe* coefficient and the *Hel* coefficient are distance/dissimilarity coefficients. The capital letter 'D' represents the coefficients. The calculated similarity values range from 1 to 0 in both binary and continuous data. A high value indicates high dissimilarity, and a low value indicates high similarity, i.e., 0 represents the condition that two structures are identical and vice versa. In this study, distance coefficients were transformed directly to similarity coefficients taking the complement D = 1−S, e.g., the *Soe* coefficient takes the form that complements the Tanimoto coefficient if it is used with dichotomous (binary) data.

Checking against their continuous formulae, it was found that:

$$1 - D_{Soe} = 1 - \frac{\sum_{i=1}^{n}|x_{iA} - x_{iB}|}{\sum_{i=1}^{n} max(x_{iA},\ x_{iB})} = \frac{\sum_{i=1}^{n} max(x_{iA},\ x_{iB}) - \sum_{i=1}^{n}|x_{iA} - x_{iB}|}{\sum_{i=1}^{n} max(x_{iA},\ x_{iB})}$$
$$= \frac{\sum_{i=1}^{n} max(x_{iA},\ x_{iB}) - \{\sum_{i=1}^{n} max(x_{iA},\ x_{iB}) - \sum_{i=1}^{n} min(x_{iA},\ x_{iB})\}}{\sum_{i=1}^{n} max(x_{iA},\ x_{iB})}$$
$$= \frac{\sum_{i=1}^{n} min(x_{iA},\ x_{iB})}{\sum_{i=1}^{n} max(x_{iA},\ x_{iB})} = S_{MR}$$

The *MR* coefficient is the MinMax coefficient used in Chapter 4. It is effective when weighting schemes are applied. For further comparision with other high performing coefficients in a larger chemical dataset (ChEMBL), the *MR* coefficient was selected, while the complementary coefficient *Soe* was excluded.

In Table 6.2 and 6.3, certain coefficients are identical in binary case with the same formula. When working on continuous variables, they may be different, i.e., the *Cze* coefficient and the *Gle* coefficient.

Thus, three coefficients from Table 6.3 were selected in this study, i.e., the coefficients *MR*, *Hel* and *Cze*, with those in Table 6.2 gives a total of 13 available for evaluation.

6.3 Method

In this study, experiments were carried out on three databases, MDDR, WOMBAT and ChEMBL. The representation of all structures is based on the frequency of fragment occurrence, i.e., ECFC_4 fingerprints. According to the results in Chapter 4, two high performing weighting schemes were adopted to weight both the reference structure and the structures from database. They are:

W4: $\sqrt{(fi)}$

W5: $0.5 + 0.5\frac{fi}{max\{fi\}}$

Where, fi represents the frequency of occurrence of the i-th element and $max\{fi\}$ is the largest fi value for a whole molecule, see details in Section 4.2.

Based on the conclusion in Chapter 4, in general, the symmetric weighing schemes performed better than the asymmetric ones. In this chapter, therefore, both reference structures and database structures were weighted equivalently, i.e., M44, M55. The experimental process is described in Chapter 3 Figure 3.4.

6.4 Results and Discussion

For each coefficient, a median value of retrieved active compounds was calculated over 10 runs (the results calculated from the 10 reference structures). In each activity class, therefore, all of the coefficients can be ranked using their median values to show their retrieval abilities from 1 to 13.

In Table 6.4, the statistically significant levels of concordance of number of actives (raw results are attached in Appendix C, Table C.1 to Table C.6) across the activity classes which were observed for the MDDR, WOMBAT and ChEMBL databases. As the *W* values are significant for these three databases, it is possible to generate overall rankings for MDDR, WOMBAT and ChEMBL databases combined with two weighting schemes.

Table 6.4 The Kendall *W* test results for the combinations of weighting schemes and databases.

	W4 weighting scheme		W5 weighting scheme	
	W	*p*	*W*	*p*
MDDR	0.156	0.056	0.217	<0.005
WOMBAT	0.461	<0.001	0.240	<0.001
ChEMBL	0.425	<0.001	0.297	<0.001

Observation of the six combinations from Table 6.5 reveals a very high degree of correspondence throughout the entire ranked list. For each combination, coefficients have been ranked when averaged over all of the activity classes. For example, coefficient *CT4* was ranked first in MDDR with the W4 weighting scheme, followed by coefficients *MR* and *Cze* then the next coefficient *Fos* was ranked the fourth. The first

positions of all six combinations were occupied by Coefficient *CT4* and coefficients *MR* and *Cze*, alternately. The Tanimoto coefficient (*JT*) ranked in the middle among all of the combinations.

According to Table 6.5, three groups of monotonic coefficients can be identified. They are coefficients (*MR, Cze*), coefficients (*DK, Hel*) and coefficients (*JT, Gle, SS1, Ja*). Of these, four coefficients *JT, Gle, SS1* and *Ja* were also found to be monotonic with each other when applied to binary representations. When compared to the formulae in Table 6.2 and Table 6.3, the *MR* coefficient is identical to the *JT* coefficient in binary form but monotonic with the *Cze* coefficient when using the weighting schemes.

The total rank is calculated averaging all the average ranks based on Table 6.5. As shown below:

MR, Cze (5.09) < *CT4* (5.19) < *JT, Gle, SS1, Ja* (6.35) < *Fos* (6.40) < *DK, Hel* (7.29) < *Phi* (8.26) < *Kul* (9.32) < *For* (11.67)

The rank above shows that the high-performing coefficient from Chapter 5 also performed well when the weighting schemes applied, i.e., coefficient *CT4*. The *CT4* coefficient was highly ranked: higher than the *JT* coefficient. The other two coefficients *MR* and *Cze* performed the best.

Figure 6.1 to Figure 6.3 illustrate the details of the 13 coefficients' top 1% retrieval rates in MDDR, WOMBAT and ChEMBL databases. In each Figure, plot (a) illustrates the results when the W4 weighting scheme was applied and plot (b) displays the outcomes when the W5 weighing scheme was used. For each plot, the retrieval abilities of coefficients are displayed as bars which indicate the retrieval rates of all activity classes over the databases. Activity classes are differentiated using varying colours as shown in the right hand side legends but not in Figure 6.3 due to the large number of activity classes. The Y axis scaled the value of total retrieval rates.

It is clear that the *For* coefficient performed poorly in all of the three databases. It performed even worse when the W4 weighting scheme was applied, i.e., in MDDR, all of the other coefficients yielded three times more results. Based on the study of Whittle *et al*. (2003), the performance of the *For* coefficient is more strongly affected by molecular size or the density of the bits set than the other coefficients. In Table C.1, the *For* coefficient retrieved similar results, as did the other coefficients, on activity classes 5HT1A, D2, COX, 5HT3 and 5HT in MDDR. It performed extremely poorly, however, on the activity class Renin with retrieval rates of 1.42% and 4.65% compared to those over 43% and 46% that were retrieved by other coefficients with W4 and W5 weighting schemes in MDDR, respectively. As shown, the difference in performance between coefficient *For* and the others is not notable in ChEMBL compared with MDDR. Whittle *et al*. (2003) suggests that the *For* coefficient is a suitable coefficient in similarity search for small actives but it may not be appropriate for continuous variables. They also analysed the possible reason theoretically. Although the theoretical explanation cannot support another coefficient's success, i.e., the *Fos* coefficient, it is still worth to have a further investigation based on their conclusion.

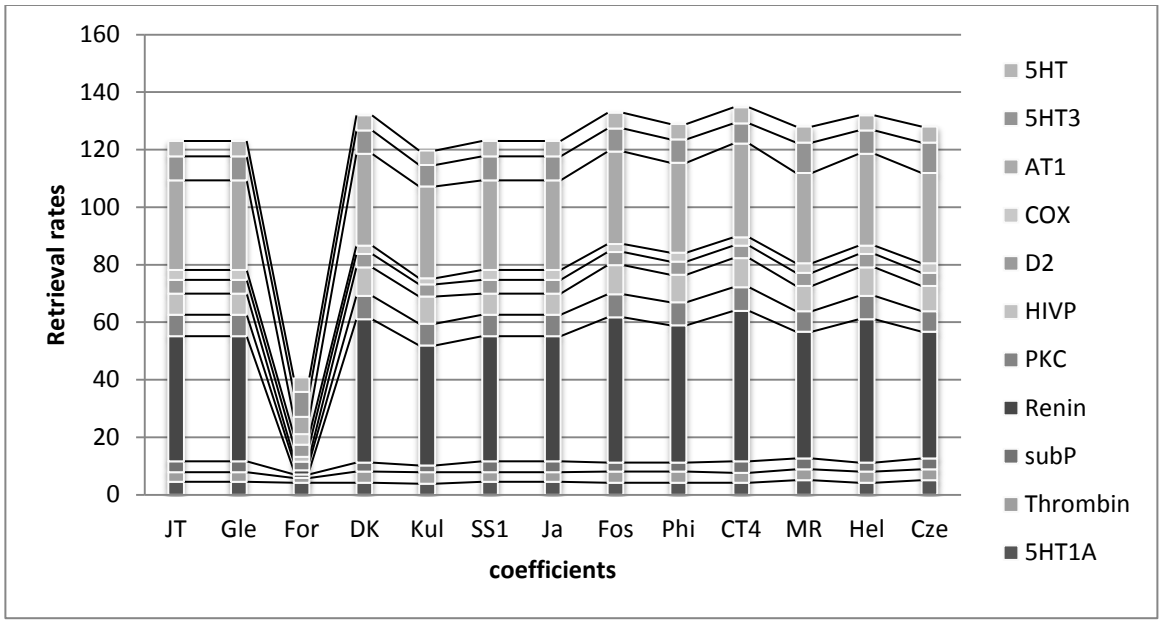
Generally, the performances of coefficients were more diverse when the W4 weighting scheme were employed, rather than cases where the W5 weighting scheme was used, e.g., more coefficients produced similar results in Figure 6.3 (b), compared with Figure 6.3 (a). It is also evident that some coefficients performed better than the Tanimoto coefficient with different combinations, e.g., the *DK*, *Fos*, *Phi*, *CT4* and *Hel* coefficients in the MDDR when using the W4 weighting scheme.

The same methods are used to analyse results in this chapter as in Chapter 5. First, the 13 coefficients are presented and roughly clustered using heatmaps. Then, they are compared by averaging over the activity classes to detect the effect of the choice of weighting schemes.

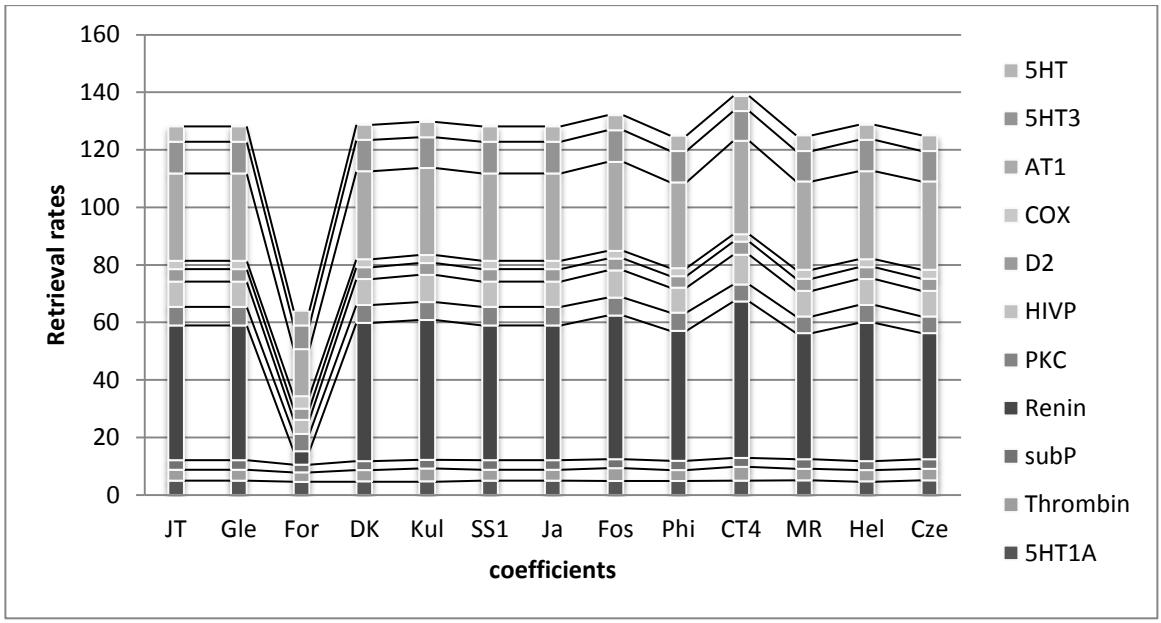
Table 6.5 Rank positions of each of the 13 coefficients when averaged over all of the activity classes for each of the combination with databases and weighting schemes.

For each combination, coefficients were ranked of their retrieval abilities with the averaged ranks, where, lower the average rank greater the retrieval ability.

Rank	MDDR_W4		MDDR_W5		WOMBAT_W4		WOMBAT_W5		ChEMBL_W4		ChEMBL_W5	
1	*CT4*	*5.18*	*CT4*	*5.73*	*MR,Cze*	*4.07*	*CT4*	*3.61*	*MR, Cze*	*3.83*	*MR, Cze*	*5.29*
2	*MR, Cze*	*5.73*	*MR, Cze*	*5.91*			*MR, Cze*	*5.71*				
3					*CT4*	*4.18*			*JT,Gle,SS1,Ja*	*5.96*	*CT4*	*5.62*
4	*Fos*	*6.23*	*JT,Gle,SS1,Ja*	*5.95*	*JT,Gle,SS1,Ja*	*6.25*	*Fos*	*5.79*			*Fos*	*6.00*
5	*DK, Hel*	*6.45*					*JT,Gle,SS1,Ja*	*6.68*			*JT,Gle,SS1,Ja*	*6.23*
6												
7	*Phi*	*6.95*							*CT4*	*6.81*		
8	*JT,Gle,SS1,Ja*	*7.05*	*Fos*	*6.18*	*Fos*	*6.71*			*DK,Hel*	*7.38*		
9			*DK,Hel*	*7.59*	*DK,Hel*	*7.36*	*DK,Hel*	*7.54*			*DK,Hel*	*7.42*
10									*Fos*	*7.47*		
11			*Kul*	*7.64*	*Phi*	*8.39*	*Phi*	*8.46*	*Phi*	*8.13*	*Kul*	*8.54*
12	*Kul*	*9.64*	*Phi*	*9.00*	*Kul*	*11.57*	*Kul*	*8.64*	*Kul*	*9.87*	*Phi*	*8.59*
13	*For*	*10.45*	*For*	*11.64*	*For*	*12.29*	*For*	*11.29*	*For*	*12.46*	*For*	*11.91*

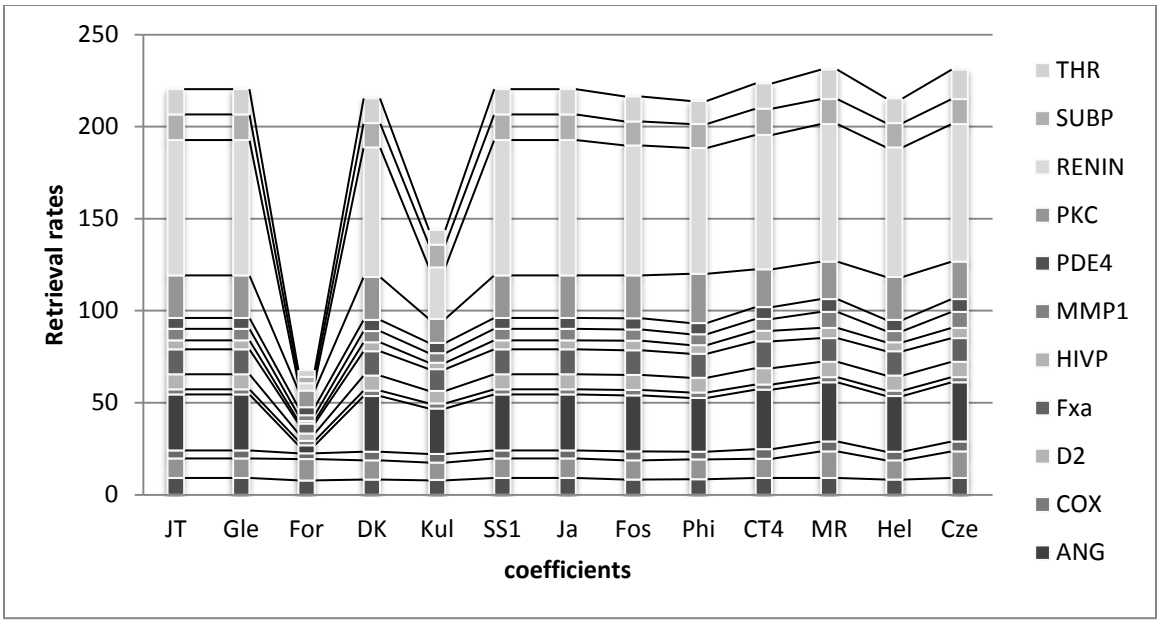


(a)

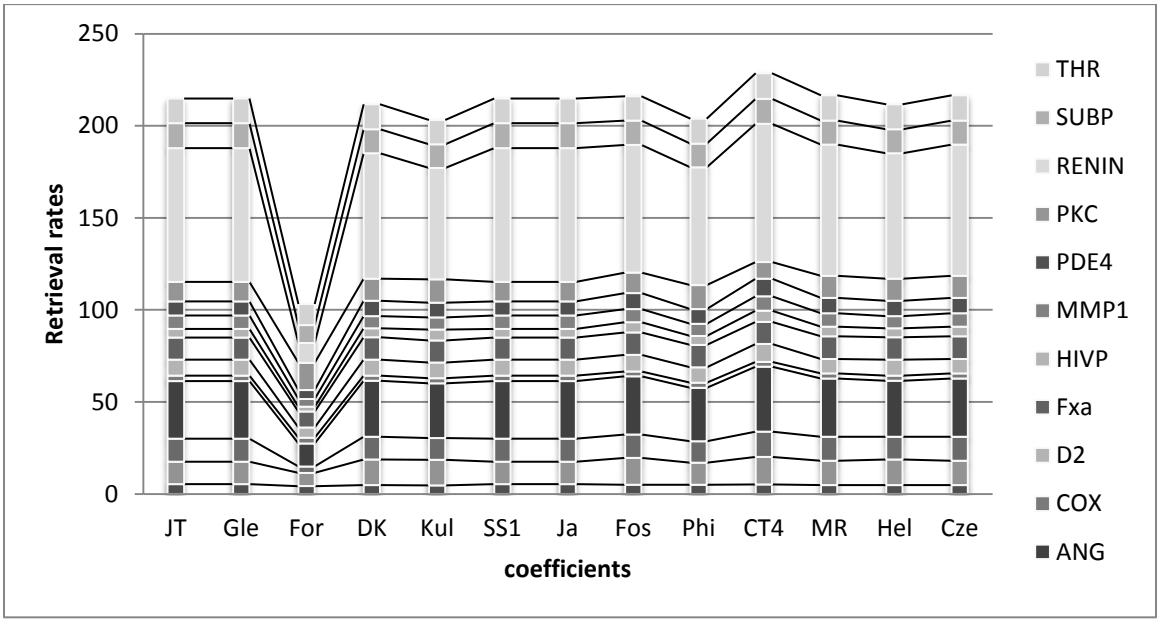


(b)

Figure 6.1 Comparison of the top 1% retrieval rates of active compounds in MDDR. (a) W4 weighted, (b) W5 weighted.

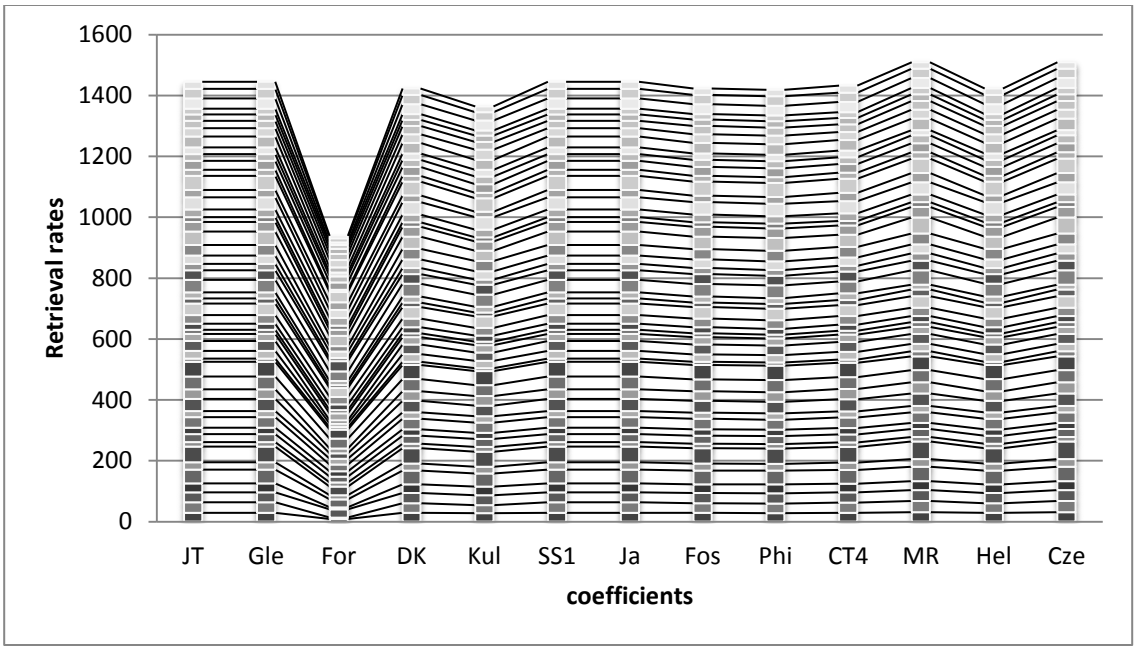


(a)

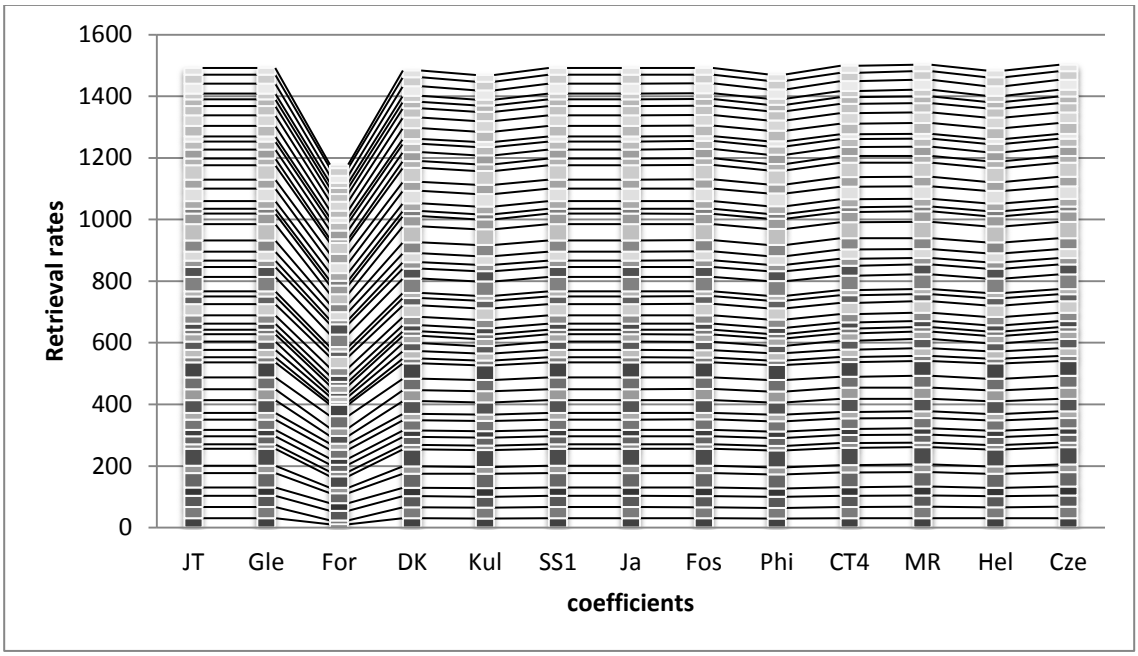


(b)

Figure 6.2 Comparison of the top 1% retrieval rates of active compounds in WOMBAT. (a) W4 weighted, (b) W5 weighted.



(a)



(b)

Figure 6.3 Comparison of the top 1% retrieval rates of active compounds in ChEMBL. (a) W4 weighted, (b) W5 weighted.

Figure 6.4 to Figure 6.6 illustrate the details of coefficients' performance on MDDR, WOMBAT and ChEMBL. For each figure, coeffcients' retrieval abilities are demonstrated using blue to red, representing the retrieval ability from low to high. Rows present the activity classes, columns illustrate the coefficients. An intersection of a specified row and a specified column shows the outcome of the corresponding coefficient working on the specified class. For each figure, the left top corner legend scaled the frequency of the correlated retrieval rates, where the X axis value refers to the value of retrieval rates and the Y axes represents the counts (frequency) of corresponding retrieval rates.

All of the coefficients provided superior outcomes on ChEMBL with a large number of retrieval rates ranging from 20% to 50%, compared with MDDR and WOMBAT, in which most of the outcomes are less than 10% and 20%, respectively. The results are similar to those in Chapter 5.

Inspection of Figure 6.4 shows that when applying the W4 weighting scheme, the *Kul* coefficient and coefficients *(JT, Gle, SS1, Ja)* are tightly clustered. Coefficients *Fos, Kul, DK* and *Hel* and coefficients *(JT, Gle, SS1, Ja)* are highly correlated when the W5 weighting scheme is applied.

CoeffIcient *CT4*, which is derived from the Tanimoto coefficient, does not show high correlations to coefficient *JT* in Figure 6.4 and 6.6, while in Figure 6.5, the *CT4* coefficient is highly correlated with coefficients *JT, Gle, SS1* and *Ja*.

Figure 6.4 to Figure 6.6 also illustrate the retrieval abilities are affected by the nature of activity classes. Coefficients often retrieved more actives in less diverse activity classes, e.g., the retrieval rates obtained by most of the coefficients are around 50% in activity class Renin both in the MDDR and WOMBAT databases. This is in agreement with the observation from Chapter 5.

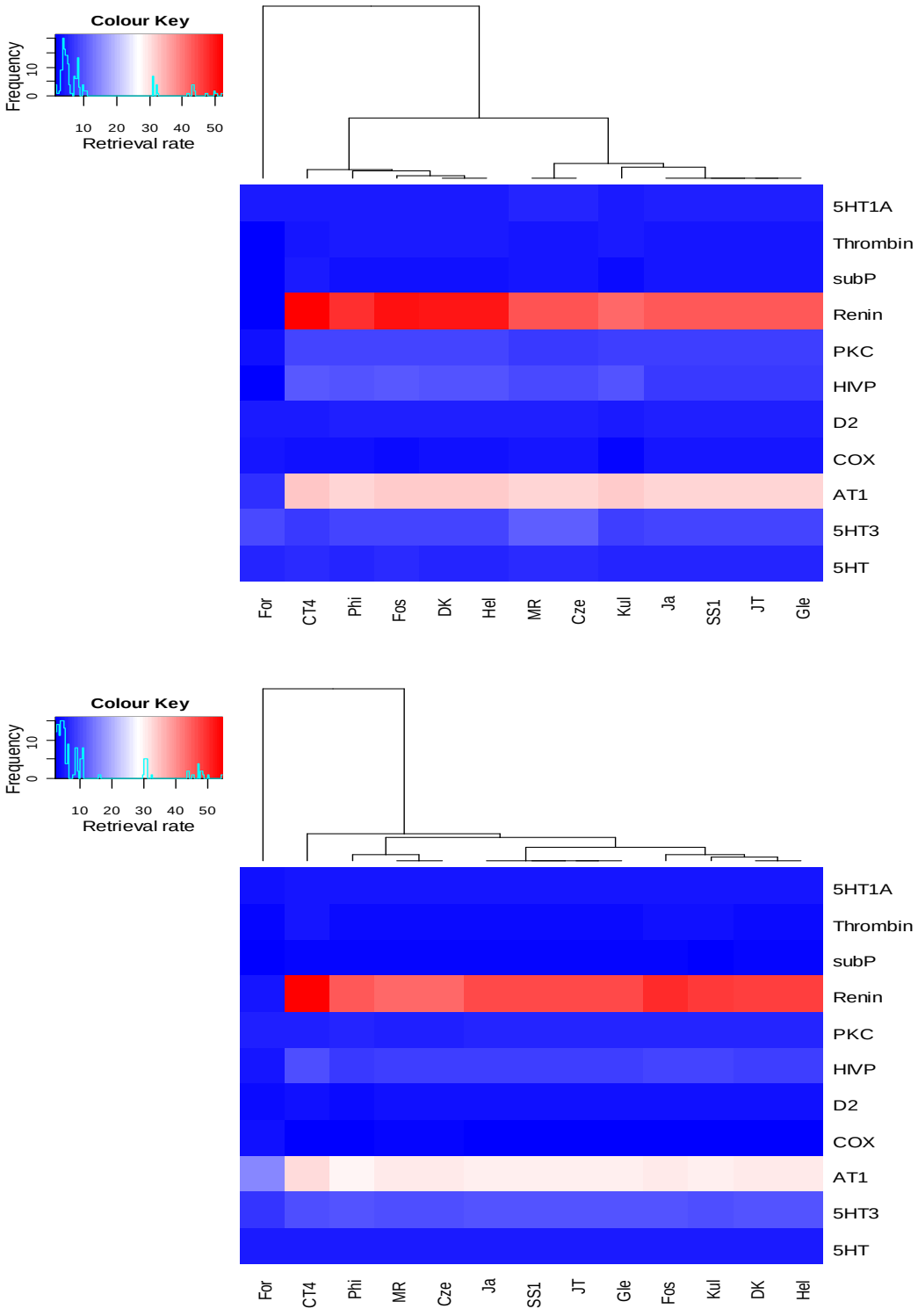


Figure 6.4 Heatmaps of the top 1% retrieval rates of active compounds in MDDR. Upper, W4 weighted; Lower, W5 weighted.

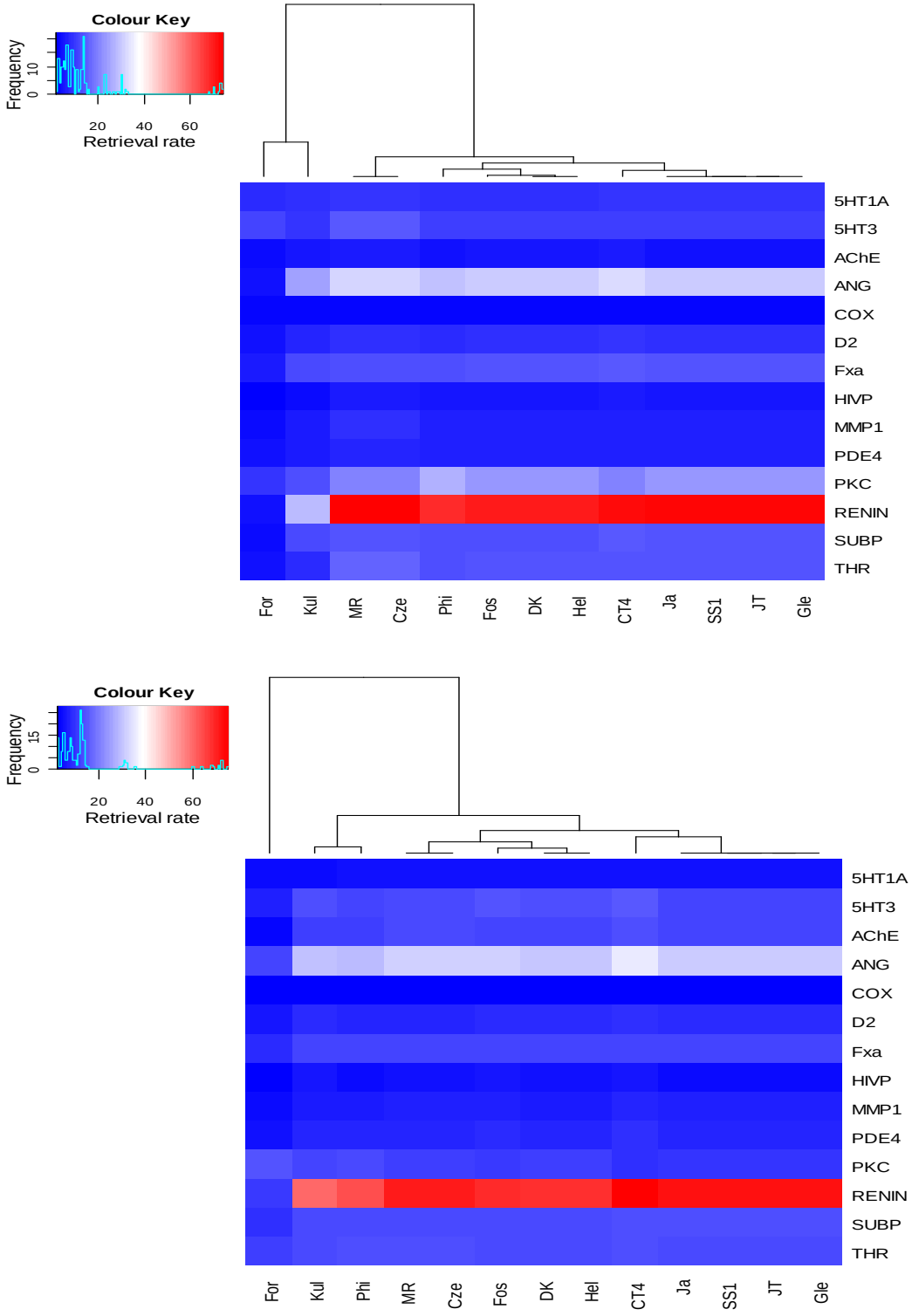
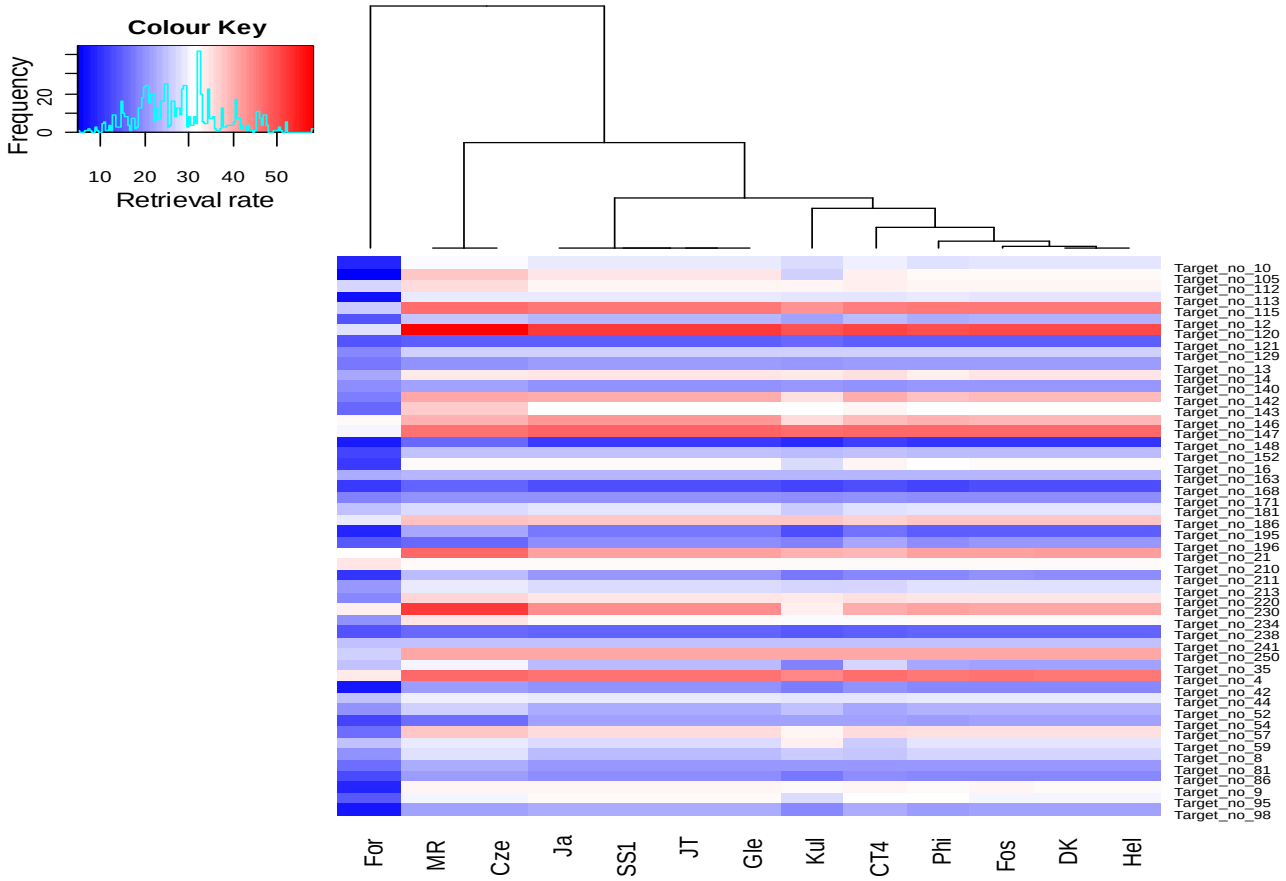
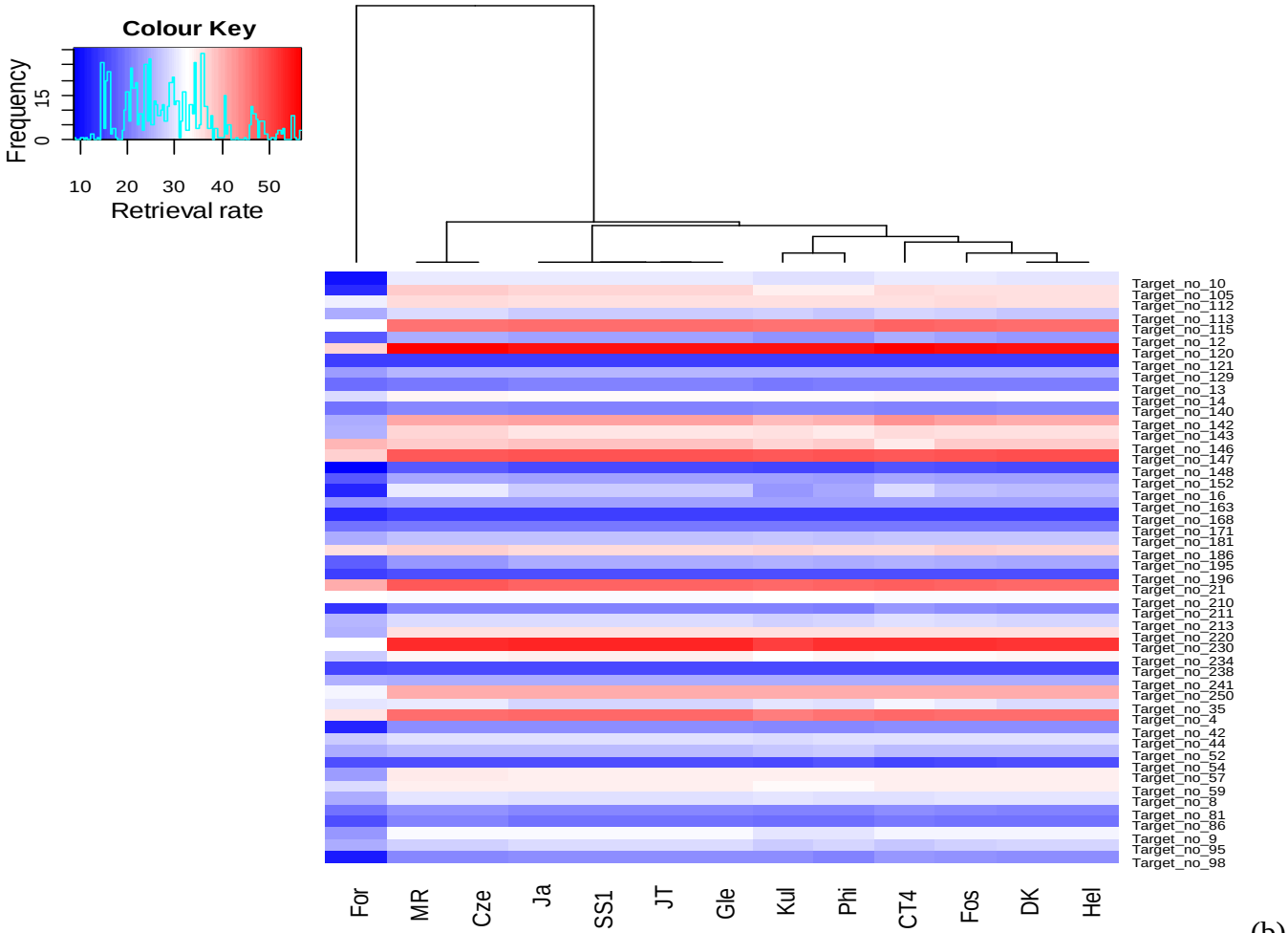


Figure 6.5 Heatmaps of the top 1% retrieval rates of active compounds in WOMBAT. Upper, W4 weighted; Lower, W5 weighted.


Colour Key
Frequency
0
20
10
20
30
40
50
Retrieval rate
Target_no_10
Target_no_105
Target_no_112
Target_no_113
Target_no_115
Target_no_12
Target_no_120
Target_no_121
Target_no_129
Target_no_13
Target_no_14
Target_no_140
Target_no_142
Target_no_143
Target_no_146
Target_no_147
Target_no_148
Target_no_152
Target_no_16
Target_no_163
Target_no_168
Target_no_171
Target_no_181
Target_no_186
Target_no_195
Target_no_196
Target_no_21
Target_no_210
Target_no_211
Target_no_213
Target_no_220
Target_no_230
Target_no_234
Target_no_238
Target_no_241
Target_no_250
Target_no_35
Target_no_4
Target_no_42
Target_no_44
Target_no_52
Target_no_54
Target_no_57
Target_no_59
Target_no_8
Target_no_81
Target_no_86
Target_no_9
Target_no_95
Target_no_98
For
MR
Cze
Ja
SS1
JT
Gle
Kul
CT4
Phi
Fos
DK
Hel

(a)



(b)

Figure 6.6 Heatmaps of the top 1% retrieval rates of active compounds in ChEMBL. (a), W4 weighted; (b), W5 weighted.

As described in Chapter 4, weighting schemes can enhance similarity search, the outcomes are hence plotted by averaging the coefficients' retrieval rates over all activity classes to give an overview of the effect from the weighting schemes.

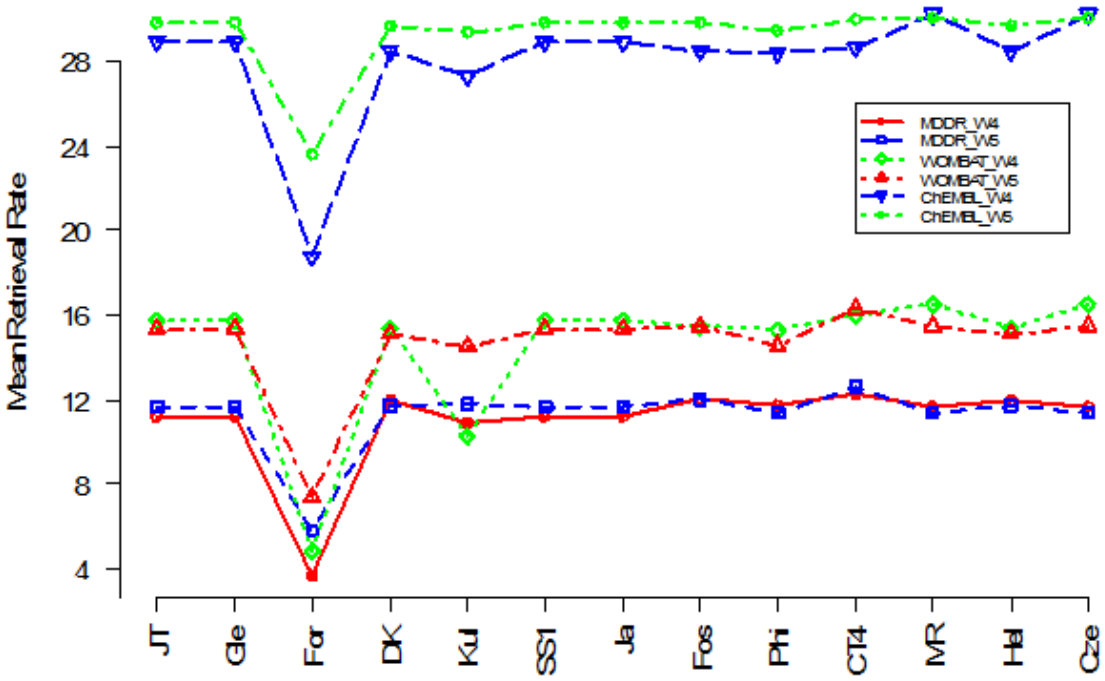


Figure 6.7 Mean retrieval rates of the 13 coefficients.

As shown in Figure 6.7, the mean retrieval rates of coefficients are calculated by averaging over the activity classes in the database. Thus, six groups of mean retrieval rates were obtained, e.g., MDDR_W4 refers to the combination of data from the MDDR database using the W4 weighting scheme.

All of the coefficients performed well in the ChEMBL database with mean retrieval rates of approximately 30%. The mean retrieval rates obtained from the MDDR and WOMBAT databases are rather poor at about 12% and 16%. In ChEMBL, nearly all of the coefficients worked better with the W5 weighting scheme compared to the W4 weighting scheme.

The Wilcoxon signed rank test was employed to compare the two weighting schemes which produced the best results in Figure 6.7. The null hypothesis is that, on a specified database, the retrieval rates of the 13 coefficients using the two weighting schemes are identical. Thus, the two weighting schemes were compared in three pairs according to the three databases. The *p*-values were 0.092, 0.261 and 0.003 from the MDDR, WOMBAT and ChEMBL databases, respectively. The p-values indicate that the two weighting schemes have significant differences working on the ChEMBL database, but not with the MDDR and WOMBAT dataset.

From Figure 6.7, it is apparent that some coefficients produced almost identical results when different weighting schemes were adopted, i.e., the *DK, SS1, Ja, Fos, Phi, CT4* and *Hel* coefficients in MDDR and WOMBAT; the *MR* and *Cze* coefficients in ChEMBL. This suggests that those coefficients are less affected by the change of weighting schemes in specific databases.

According to the observations above, four high performing coefficients(group) were selected to detect the effect of the choice of weighting schemes. Hence, the *(JT, Gle, SS1, Ja), CT4, (MR, Cze)* and *Fos* coefficients were compared with their outcomes from Chapter 5.

Table 6.6 illustrates the comparison of coefficients working with the two weighting schemes. The results in the W1 rows are from Chapter 6 where W1 is used to specify the binary (non-weighted) data. It is apparent that the two weighting schemes can sometimes improve similarity search in the MDDR and WOMBAT databases. The W5 weighting scheme provide improvement when employed by the *CT4* coefficient. The W4 weighting scheme shows preference when applied to the coefficients *MR* and *Cze.*

From Table 6.6, the *MR, Cze* coefficient(s) gave notable improvement over the *JT* coefficient when applying weighting schemes and the *CT4* coefficient consistently over-achieved the *JT* coefficient.

Table 6.6 Mean retrieval rates of four high-achieving coefficient(group)s.

		JT, Gle, SS1, Ja	*CT4*	*MR, Cze*	*Fos*
MDDR	W1	11.11	12.09	11.11	11.45
	W4	11.18	12.25	11.64	12.07
	W5	11.64	12.61	11.36	12.01
WOMBAT	W1	15.04	15.78	15.04	14.94
	W4	15.74	15.97	16.50	15.45
	W5	15.35	16.32	15.47	15.43
ChEMBL	W1	29.74	29.82	29.74	29.64
	W4	28.90	28.64	30.19	28.47
	W5	29.84	29.98	30.04	29.84

6.5 Conclusion

The experiments in this chapter investigated whether there were other coefficients that could perform better than the Tanimoto coefficient in similarity search when weighting schemes were applied. Most of the coefficients investigated here were taken from the high performing coefficients list based on the outcomes from Chapter 5. Coefficients from the other studies with similar characteristics with these high performing coefficients were also included.

The outcomes showed that there are a number of coefficients which perform well on binary descriptors and can be applied to non-binary variables. Several of them can consistently perform well, with different weighting schemes. Moreover, the experiments confirm that weighting schemes can enhance similarity search, i.e., coefficients *CT4*, *MR* and *Cze* yielded marked improvements when weighted.

On average, the overall best performing coefficients are *CT4*, *MR* and *Cze*. Of these, the *CT4* coefficient consistently provided better performance over the *JT* coefficient. These findings therefore suggested that when weighting schemes applied, *CT4*, *MR* and *Cze* might be appropriate for similarity-based virtual screening and *CT4* might be the choice

when applied on un-weighted and weighted fingerprints. This study has found that although the *JT* (Tanimoto) coefficient is strong in the use of similarity search, some coefficients can be viable alternatives to the *JT* (Tanimoto) coefficient using both binary and weighted fingerprints. The finding also ascertained that the interactions between coefficients and weighting schemes can be considered in similarity search. This is hence suggested that future study can investigate more combinations of similarity measures by adopting different data fusion rules.

Chapter 7: Conclusion

The aim of the research reported in this thesis was to identify novel similarity measures for ligand-based virtual screening. Ligand-based virtual screening belongs to the lead identification stage of drug discovery process, as shown as Figure 1.1 in Chapter 1. It normally requires explorations in large scale databases. Thus, a slight improvement of accuracy of the methods employed can result in a significant enhancement of effectiveness of the whole process of drug discovery.

The theoretic foundations of the research reported in this thesis are Johnson and Maggiora's similar property principle (Johnson and Maggiora, 1990), and the three key components in similarity search were introduced by Willett *et al*. (1998). As the similar property principle states, structurally similar compounds do have similar biological activity, and the biological similarity increases with the increasing structural similarity. Therefore, the three principal components involved in similarity search can be used to measure the correlation of molecular similarity. They are structural representation, similarity coefficient and weighting scheme.

The experimental section of the thesis started with investigating the effect of the combinations of weighting schemes and similarity coefficients. The idea of fragment weighting schemes was derived from information retrieval, and the five weighting schemes investigated in this thesis were from a previous study (Arif *et al.*, 2009b). Since weighting schemes can be applied on reference structures, database structures and on both, five weighting schemes results in 25 weighting combinations in total. As reported in Chapter 4, in Arif *et al*.'s study, 19 out of 25 weighting schemes were chosen and the experiments were conducted in two databases, i.e., MDDR and WOMBAT. According to their results, they concluded that some weighting schemes could enhance similarity search when the Tanimoto coefficient was employed. The Tanimoto coefficient, one of

the most conventional coefficients has been widely utilized in similarity search as well as other aspects in chemoinformatics. In order to verify if the weighting schemes can consistently enhance similarity search when other coefficients were applied, in the first part of the investigation, experiments were carried out in the same databases, i.e., MDDR and WOMBAT. The results confirmed that some weighting schemes can boost similarity search, i.e., W4 and W5. For all weighting combinations, the symmetric weighting combinations performed better than the asymmetric weighting combinations.

Another finding was that when weighting schemes are applied, the cosine coefficient performed in a more stable manner than the Tanimoto coefficient. When averaged over all results of 25 weighting combinations, the cosine coefficient outperformed the Tanimoto coefficient. Due to the limitations of the two databases that were used, a further evaluation was carried out on the MUV database. The results were similar to those from the MDDR and WOMBAT databases.

The main conclusion from the first investigation was that there are strong, and often quite subtle, interactions between the similarity coefficient and the weighting scheme comprising a similarity measure. These interactions indicate that the Tanimoto coefficient may not be the coefficient of choice when weighted fingerprints are used. It is therefore suggested some other coefficients may be favorable for similarity-based virtual screening when weighted fingerprints are available.

The second investigation focused on the identification of binary coefficients which can be used for similarity search. Based on the findings from the first investigation, the similar performed coefficients in non-weighted case can exhibit different performance when weighting schemes were applied, i.e., the MinMax coefficient was identical to the Tanimoto coefficient but yielded better results with weighting schemes. There are many similarity coefficients that have been employed in other domains. It is possible to identify a number of them which might provide similar/better performance to the Tanimoto coefficient. For this purpose, 44 binary coefficients were extensively analysed and compared, most of those have so far not been studied in similarity-based virtual

screening. The initial investigation was carried on the MDDR, WOMBAT and MUV databases. The statistical values indicated, however, the MUV database might not be favorable for fingerprint-based similarity searching. Therefore, the ChEMBL database has been used for further validation.

The comparison of the 44 coefficients was implemented by using their retrieval abilities, i.e., their ranks in each activity class and their retrieval rates of active compounds. The Ward's method was used to cluster coefficients based on their retrieval abilities.

The outcome showed that there are a number of coefficients which are suitable for similarity search. Generally, the asymmetric coefficients and the correlation-based coefficients performed better than the symmetric coefficients and the intermediate coefficients. The hierarchical cluster analysis revealed that most of the coefficients from the same class can yield similar results. The analysis based on the nature of activity classes indicates that the performance of coefficients may vary when applied to homogeneous classes, and that asymmetric coefficients yielded the best results. Working on homogeneous classes, a number of coefficients performed better than the Tanimoto coefficient. Therefore, it is possible to apply the identified high performing coefficients to large-scale virtual screening, and, these identified might also provide consistently good performance with weighting schemes, i.e., *CT4, DK, Gle, For, Kul, SS1, Fos, Phi, Ja, Sor.*

The final investigation hence focused on the interactions between the high performing coefficients and weighting schemes. In the first investigation, there was an analytical elaboration of five weighting schemes. Two of these are considered appropriate and effective systems for similarity-based virtual screening, i.e., W4 and W5. Apart from the high performing coefficients from the second investigation, coefficients from the other studies with similar characterises with these coefficients were also included. The outcomes showed that there are a number of coefficients which perform well on binary descriptors and can be applied to weighted fingerprints. Several of them can consistently perform well, with different weighting schemes. The experiments also confirmed that

weighting schemes can enhance similarity search. In addition, this study has found that although the Tanimoto coefficient is strong in the use of similarity search, some coefficients can be viable alternatives to the Tanimoto coefficient using both binary and weighted fingerprints, i.e., *CT4*, *MR* and *Cze*.

The research reported in this thesis has shown that although the Tanimoto coefficient remains one of the most practical coefficients for use in similarity searching on binary representations, it may not be the coefficient of choice when weighting schemes were applied. Therefore, the further study in the interactions of similarity coefficient and weighting schemes is expected to provide supplementary consequential contribution to similarity-based virtual screening.

The future research would further investigate possibilities to optimize these similarity measures through adding data fusion rules such as MAX and TSS discussed in Section 2.5. Other weighting schemes and different types of fingerprints can also be involved in the future study and provide more solid conclusions in various circumstances.

Bibliography:

Accelrys Software, I. (2009). *Pipeline Pilot Software* [Online]. Available: http://accelrys.com/ [Accessed 2010].

Adamson, G. W. & Bush, J. A. (1975). "Comparison of Performance of Some Similarity and Dissimilarity Measures in Automatic Classification of Chemical Structures". *Journal of Chemical Information and Computer Sciences,* 15**,** 55-58.

Adamson, G. W., Cowell, J., Mclure, A. H. W., Town, W. G., Yapp, A. M. & Lynch, M. F. (1973). "Strategic Considerations in Design of a Screening System for Substructure Searches of Chemical Structure Files". *Journal of Chemical Documentation,* 13**,** 153-157.

Al Khalifa, A., Haranczyk, M. & Holliday, J. (2009). "Comparison of Nonbinary Similarity Coefficients for Similarity Searching, Clustering and Compound Selection". *Journal of Chemical Information and Modeling,* 49**,** 1193-1201.

Anderberg, M. (1973). *Clustering Analysis for Applications,* London, Academic Press.

Arif, S. M., Hert, J., Holliday, J. D., Malim, N. & Willett, P. (2009a). "Enhancing the Effectiveness of Fingerprint-Based Virtual Screening: Use of Turbo Similarity Searching and of Fragment Frequencies of Occurrence". *Pattern Recognition in Bioinformatics, Proceedings,* 5780**,** 404-414.

Arif, S. M., Holliday, J. D. & Willett, P. (2009b). "Analysis and Use of Fragment-Occurrence Data in Similarity-Based Virtual Screening". *Journal of Computer-Aided Molecular Design,* 23**,** 655-668.

Arif, S. M., Holliday, J. D. & Willett, P. (2010). "Inverse Frequency Weighting of Fragments for Similarity-Based Virtual Screening". *Journal of Chemical Information and Modeling,* 50**,** 1340-1349.

Atkinson, J. D. M. & Jones, R. (2009). "Intellectual Property and Its Role in the Pharmaceutical Industry". *Future Medicinal Chemistry,* 1**,** 1547-1550.

Bajorath, F. (2002). "Integration of Virtual and High-Throughput Screening". *Nature Reviews Drug Discovery,* 1**,** 882-894.

Bajorath, J. & Eckert, H. (2006). "Design and Evaluation of a Novel Class-Directed 2D Fingerprint to Search for Structurally Diverse Active Compounds". *Journal of Chemical Information and Modeling,* 46**,** 2515-2526.

Baldi, P., Azencott, C. A., Ksikes, A., Swamidass, S. J., Chen, J. H. & Ralaivola, L. (2007). "One- to Four-Dimensional Kernels for Virtual Screening and the Prediction of Physical, Chemical, and Biological Properties". *Journal of Chemical Information and Modeling,* 47**,** 965-974.

Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F. & Mason, J. S. (2007). "A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands and Proteins (Flap): Theory and Application". *Journal of Chemical Information and Modeling,* 47**,** 279-294.

Batagelj, V. & Bren, M. (1995). "Comparing Resemblance Measures". *Journal of Classification,* 12**,** 73-90.

Bath, P. A., Poirrette, A. R., Willett, P. & Allen, F. H. (1994). "Similarity Searching in Files of 3-Dimensional Chemical Structures - Comparison of Fragment-Based Measures of Shape Similarity". *Journal of Chemical Information and Computer Sciences,* 34**,** 141-147.

Baulieu, F. B. (1989). "A Classification of Presence Absence Based Dissimilarity Coefficients". *Journal of Classification,* 6**,** 233-246.

Bemis, G. W. & Murcko, M. A. (1996). "The Properties of Known Drugs.1. Molecular Frameworks". *Journal of medicinal chemistry,* 39**,** 2887-2893.

Beno, B. R. & Mason, J. S. (2001). "The Design of Combinatorial Libraries Using Properties and 3d Pharmacophore Fingerprints". *Drug Discovery Today,* 6**,** 251-258.

Berman, H. M., Westbrook, J., Arzberger, P., Bourne, P., Gilliland, G. & Fagan, P. (1999). "Our Vision for the New Protein Data Bank". *Biophysical Journal,* 76**,** A200-A200.

Biggs, N., Lloyd, E. K. & Wilson, R. J. (1976). *Graph Theory 1736-1936,* Oxford Eng., Clarendon Press.

Bocker, A., Derksen, S., Schmidt, E., Teckentrup, A. & Schneider, G. (2005). "A Hierarchical Clustering Approach for Large Compound Libraries". *Journal of Chemical Information and Modeling,* 45**,** 807-815.

Boyce, B. R. (1990). "Automatic Text-Processing - the Transformation Analysis, and Retrieval of Information by Computer - Salton,G". *Journal of the American Society for Information Science,* 41**,** 150-151.

Brown, F. K. (1998). "Chemoinformatics: What Is It and How Does It Impact Drug Discovery". *Annual Reports in Medicinal Chemistry***,** 33.

Brown, N. (2009). "Chemoinformatics – an Introduction for Computer Scientists". *ACM Computing Surveys (CSUR),* 41**,** Article No. 8.

Campbell, B. M. (1978). "Similarity Coefficients for Classifying Releves". *Vegetatio,* 37**,** 101-109.

Carhart, R. E., Smith, D. H. & Venkataraghavan, R. (1985). "Atom Pairs as Molecular-Features in Structure Activity Studies - Definition and Applications". *Journal of Chemical Information and Computer Sciences,* 25**,** 64-73.

CAS. (2013). *Chemical Abstracts Service* [Online]. Available: http://www.cas.org/ [Accessed 2013].

Cha, S. H. (2008). "Taxonomy of Nominal Type Histogram Distance Measures". *Recent Advances on Applied Mathematics***,** 325-330.

Chen, J., Holliday, J. & Bradshaw, J. (2009). "A Machine Learning Approach to Weighting Schemes in the Data Fusion of Similarity Coefficients". *Journal of Chemical Information and Modeling,* 49**,** 185-194.

Chen, X. & Reynolds, C. H. (2002). "Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients". *Journal of Chemical Information and Computer Sciences,* 42**,** 1407-1414.

Cheng, T. J., Li, Q. L., Zhou, Z. G., Wang, Y. L. & Bryant, S. H. (2012). "Structure-Based Virtual Screening for Drug Discovery: A Problem-Centric Review". *The AAPS Journal,* 14**,** 133-141.

Choi, S.-S., Cha, S.-H. & Tappert, C. C. (2010). "A Survey of Binary Similarity and Distance Measures". *Journal of Systemics, Cybernetics and Informatics,* 8(1)**,** 43-48.

Cole, L. C. (1957). "The Measurement of Partial Interspecific Association". *Ecology,* 38**,** 226-233.

Consonni, V. & Todeschini, R. (2012). "New Similarity Coefficients for Binary Data". *MATCH Communications in Mathematical and in Computer Chemistry,* 68**,** 581-592.

Cosgrove, D. A. & Willett, P. (1998). "Slash: A Program for Analysing the Functional Groups in Molecules". *Journal of Molecular Graphics & Modelling,* 16**,** 19-32.

CSD. (2011). *Chemical Structure Database* [Online]. Available: http://www.ccdc.cam.ac.uk/products/csd [Accessed 2011].

Daylight Chemical Information Systems, I. (2011). *Daylight Theory Manual* [Online]. Laguna Niguel, CA: Daylight Chemical Information Systems, Inc. [Accessed 2011].

Dean, P. M. & Lewis, R. A. (1999). *Molecular Diversity in Drug Design,* Dordrecht ; London, Kluwer Academic.

Deng, Z., Chuaqui, C. & Singh, J. (2004). "Structural Interaction Fingerprint (Sift): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions". *Journal of Medicinal Chemistry,* 47**,** 337-344.

Dice, L. R. (1945). "Measures of the Amount of Ecologic Association between Species". *Ecology,* 26**,** 297-302.

Diestel, R. (2000). *Graph Theory,* New York, Springer-Verlag.

Downs, G. M. & Barnard, J. M. (2002). "Clustering Methods and Their Uses in Computational Chemistry". *Reviews in Computational Chemistry, Vol 18,* 18**,** 1-40.

Downs, G. M., Willett, P. & Fisanick, W. (1994). "Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data". *Journal of Chemical Information and Computer Sciences,* 34**,** 1094-1102.

Driver, H. E. & Kroeber, A. L. (1932). "Quantitative Expression of Cultural Relationship.". *The University of California Publications in American Archaeology and Ethnology,* 31**,** 211-256.

Duan, J. X., Dixon, S. L., Lowrie, J. F. & Sherman, W. (2010). "Analysis and Comparison of 2D Fingerprints: Insights into Database Screening Performance Using Eight Fingerprint Methods". *Journal of Molecular Graphics & Modelling,* 29**,** 157-170.

Duarte, J. M., dos Santos, J. B. & Melo, L. C. (1999). "Comparison of Similarity Coefficients Based on Rapd Markers in the Common Bean". *Genetics and Molecular Biology,* 22**,** 427-432.

Edgar, S. J., Holliday, J. D. & Willett, P. (2000). "Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures". *Journal of Molecular Graphics & Modelling,* 18**,** 343-357.

Ekins, S. (2006). *Computer Applications in Pharmaceutical Research and Development,* Hoboken, N.J., Wiley-Interscience ; Chichester : John Wiley [distributor].

Ellis, D., Furner-Hines, J. & Willett, P. (1993). "Measuring the Degree of Similarity between Objects in Text Retrieval Systems". *Perspectives in Information Management,* 3**,** 128-149.

Engel, T. (2006). "Basic Overview of Chemoinformatics". *Journal of Chemical Information and Modeling,* 46**,** 2267-2277.

Gardiner, E. J., Gillet, V. J., Haranczyk, M., Hert, J., Holliday, J. D., Malim, N., Patel, Y. & Willett, P. (2009). "Turbo Similarity Searching: Effect of Fingerprint and Dataset on Virtual-Screening Performance". *Statistical Analysis and Data Mining,* 2**,** 103–114.

Gasteiger, J. (2006). "Chemoinformatics: A New Field with a Long Tradition". *Analytical and Bioanalytical Chemistry,* 384**,** 57-64.

Gasteiger, J. & Engel, T. (2003). *Chemoinformatics : A Textbook,* Weinheim, Wiley-VCH.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. & Overington, J. P. (2012). "Chembl: A Large-Scale Bioactivity Database for Drug Discovery". *Nucleic Acids Research,* 40**,** D1100-D1107.

Geppert, H., Vogt, M. & Bajorath, J. (2010). "Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation". *Journal of Chemical Information and Modeling,* 50**,** 205-216.

Gillet, V. J., Downs, G. M., Holliday, J. D., Lynch, M. F. & Dethlefsen, W. (1991). "Computer-Storage and Retrieval of Generic Chemical Structures in Patents .13. Reduced Graph Generation". *Journal of Chemical Information and Computer Sciences,* 31**,** 260-270.

Gillet, V. J., Willett, P. & Bradshaw, J. (2003). "Similarity Searching Using Reduced Graphs". *Journal of Chemical Information and Computer Sciences,* 43**,** 338-345.

Goodman, L. A. & Kruskal, W. H. (1954). "Measures of Association for Cross Classifications". *Journal of the American statistical association,* 49**,** 732-764.

Gower, J. C. & Legendre, P. (1986). "Metric and Euclidean Properties of Dissimilarity Coefficients". *Journal of Classification,* 3**,** 5-48.

Grier, D., Hounshell, W. D., Moock, T. & Grethe, G. (1988). "Similarity Searching in Reaction Databases". *Abstracts of Papers of the American Chemical Society,* 196**,** 32-Comp.

Guldbrandt, M., Johansen, T. N., Frydenvang, K., Brauner-Osborne, H., Stensbol, T. B., Nielsen, B., Karla, R., Santi, F., Krogsgaard-Larsen, P. & Madsen, U. (2002). "Glutamate Receptor Ligands: Synthesis, Stereochemistry, and Enantiopharmacology of Methylated 2-Aminoadipic Acid Analogs". *Chirality,* 14**,** 351-363.

Harris, F. C. & Lahey, B. B. (1978). "Method for Combining Occurrence and Non-Occurrence Interobserver Agreement Scores". *Journal of Applied Behavior Analysis,* 11**,** 523-527.

Heikamp, K. & Bajorath, J. (2011). "Large-Scale Similarity Search Profiling of Chembl Compound Data Sets". *Journal of Chemical Information and Modeling,* 51**,** 1831-1839.

Hert, J., Willett, P. & Wilton, D. J. (2004a). "Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures". *Journal of Chemical Information and Computer Sciences,* 44**,** 1177-1185.

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E. & Schuffenhauer, A. (2004b). "Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures". *Organic & Biomolecular Chemistry,* 2**,** 3256-3266.

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E. & Schuffenhauer, A. (2006). "New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching". *Journal of Chemical Information and Modeling,* 46**,** 462-470.

Holliday, J., Al Khalifa, A. & Haranczyk, M. (2009). "Comparison of Nonbinary Similarity Coefficients for Similarity Searching, Clustering and Compound Selection". *Journal of Chemical Information and Modeling,* 49**,** 1193-1201.

Holliday, J. & Haranczyk, M. (2008). "Comparison of Similarity Coefficients for Clustering and Compound Selection". *Journal of Chemical Information and Modeling,* 48**,** 498-508.

Holliday, J. & Willett, P. (2011). Representation and Searching of Chemical Structure Information in Patents. *In:* Lupo, M., Mayer, K., Tait, J. & Trippe, A. J. (eds.) *Current Challenges in Patent Information Retrieval.* 1st ed**,** 343-356. Springer.

Holliday, J. D., Hu, C. Y. & Willett, P. (2002). "Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity Using 2D Fragment Bit-Strings". *Combinatorial Chemistry & High Throughput Screening,* 5**,** 155-166.

Holliday, J. D., Ranade, S. S. & Willett, P. (1995). "A Fast Algorithm for Selecting Sets of Dissimilar Molecules from Large Chemical Databases". *Quantitative Structure-Activity Relationships,* 14**,** 501-506.

Holliday, J. D., Salim, N., Whittle, M. & Willett, P. (2003). "Analysis and Display of the Size Dependence of Chemical Similarity Coefficients". *Journal of Chemical Information and Computer Sciences,* 43**,** 819-828.

Hubalek, Z. (1982). "Coefficients of Association and Similarity, Based on Binary (Presence Absence) Data - an Evaluation". *Biological Reviews of the Cambridge Philosophical Society,* 57**,** 669-689.

Hubbard, R. E. (1997). "Can Drugs Be Designed?". *Current Opinion in Biotechnology,* 8**,** 696-700.

Hull, D. A. (1993). "Using Statistical Testing in the Evaluation of Retrieval". *Research and Development in Information Retrieval - SIGIR.* PA, USA.

Hurst, J. R. & Heritage, T. W. (1998). *Molecular Hologram Qsar*. USA patent application 022252. 27/03/2001.

IUPAC. (2011). *International Union of Pure and Applied Chemistry* [Online]. Available: http://iupac.org/ [Accessed 2011].

Jaccard, P. (1912). "The Distribution of the Flora in the Alpine Zone". *The New Phytologist,* 11**,** 37-50.

Jain, A. N. & Nicholls, A. (2008). "Recommendations for Evaluation of Computational Methods". *J Comput Aided Mol Des,* 22**,** 133-9.

Jardine, N. & Sibson, R. (1971). *Mathematical Taxonomy,* London, Wiley.

Jaworska, J. & Nikolova, N. (2004). "Approaches to Measure Chemical Similarity - a Review". *Qsar & Combinatorial Science,* 22**,** 1006-1026.

Johnson, M. & Maggiora, G. (1990). *Concepts and Applications of Molecular Similarity*, Wiley New York.

Kelly, M. D. & Mancera, R. L. (2004). "Expanded Interaction Fingerprint Method for Analyzing Ligand Binding Modes in Docking and Structure-Based Drug Design". *Journal of Chemical Information and Computer Sciences,* 44**,** 1942-1951.

Korenius, T., Laurikkala, J. & Juhola, M. (2007). "On Principal Component Analysis, Cosine and Euclidean Measures in Information Retrieval". *Information Sciences,* 177**,** 4893-4905.

Kulczynski, S. (1927). "Die Pflanzenassociationen Der Pienenen". *Bulletin International de L'Acad´emie Polonaise des Sciences et des Letters, Classe des Sciences Mathematiques et Naturelles, Serie B, Suppl´ement II,* 2**,** 57-203.

Lan, M., Sung, S. Y., Low, H. B. & Tan, C. L. (2005). "A Comparative Study on Term Weighting Schemes for Text Categorization". *Proceedings of the International Joint Conference on Neural Networks (IJCNN), Vols 1-5***,** 546-551.

Lan, M., Tan, C. L., Su, J. & Low, H. B. (2007). "Text Representations for Text Categorization: A Case Study in Biomedical Domain". *2007 IEEE International Joint Conference on Neural Networks, Vols 1-6***,** 2556-2561.

Leach, A. R. & Gillet, V. J. (2007). *An Introduction to Chemoinformatics,* Dordrecht; London, Springer.

Leiter, D. P., Morgan, H. L. & Stobaugh, R. E. (1965). "Installation and Operation of a Registry for Chemical Compounds". *Journal of Chemical Documentation,* 5**,** 238-&.

Liebetrau, A. M. (1983). *Measures of Association,* Beverly Hills, Sage Publications.

Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. (1997). "Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings". *Advanced Drug Delivery Reviews,* 23**,** 3-25.

Looman, J. & Campbell, J. B. (1960). "Adaptation of Sorensens-K (1948) for Estimating Unit Affinities in Prairie Vegetation". *Ecology,* 41**,** 409-416.

Manaut, F., Sanz, F., Jose, J. & Milesi, M. (1991). "Automatic Search for Maximum Similarity between Molecular Electrostatic Potential Distributions". *Journal of Computer-Aided Molecular Design,* 5**,** 371-380.

Martin, Y. C., Kofron, J. L. & Traphagen, L. M. (2002). "Do Structurally Similar Molecules Have Similar Biological Activity?". *Journal of medicinal chemistry,* 45**,** 4350-4358.

Mason, J. S., Morize, I., Menard, P. R., Cheney, D. L., Hulme, C. & Labaudiniere, R. F. (1999). "New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures". *Journal of Medicinal Chemistry,* 42**,** 3251-3264.

McDonald, J. H. (2009). *Handbook of Biological Statistics* Baltimore, Maryland, Sparky House Publishing.

Medina-Franco, J. L., Maggiora, G. M., Giulianotti, M. A., Pinilla, C. & Houghten, R. A. (2007). "A Similarity-Based Data-Fusion Approach to the Visual Characterization and Comparison of Compound Databases". *Chemical Biology & Drug Design,* 70**,** 393-412.

Meyer, A. D., Garcia, A. A. F., de Souza, A. P. & de Souza, C. L. (2004). "Comparison of Similarity Coefficients Used for Cluster Analysis with Dominant Markers in Maize (Zea Mays L)". *Genetics and Molecular Biology,* 27**,** 83-91.

Mezey, P. G. (2001). "The Holographic Principle for Latent Molecular Properties". *Journal of Mathematical Chemistry,* 30**,** 299-303.

Michael, E. L. (1920). "Marine Ecology and the Coefficient of Association - a Plea in Behalf of Quantitative Biology". *Journal of Ecology,* 8**,** 54-59.

Morgan, H. L. (1965). "Generation of a Unique Machine Description for Chemical Structures-a Technique Developed at Chemical Abstracts Service". *Journal of Chemical Documentation,* 5**,** 107-&.

Muchmore, S. W., Debe, D. A., Metz, J. T., Brown, S. P., Martin, Y. C. & Hajduk, P. J. (2008). "Application of Belief Theory to Similarity Data Fusion for Use in Analog Searching and Lead Hopping". *Journal of Chemical Information and Modeling,* 48**,** 941-948.

MUV. (2011). *Maximum Unbiased Validation Datasets* [Online]. Available: http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html [Accessed 2011].

Nasr, R. J., Swamidass, S. J. & Baldi, P. F. (2009). "Large Scale Study of Multiple-Molecule Queries". *Journal of Cheminformatics,* 1.

Novartis. (2012). *Novartis* [Online]. Available: http://www.novartis.com [Accessed 2012].

Oprea, T. I., Olah, M. & Bologa, C. (2004). "An Automated Pls Search for Biologically Relevant Qsar Descriptors". *Journal of Computer-Aided Molecular Design,* 18**,** 437-449.

Ormerod, A., Willett, P. & Bawden, D. (1989). "Comparison of Fragment Weighting Schemes for Substructural Analysis". *Quantitative Structure-Activity Relationships,* 8**,** 115-129.

PDB. (2011). *Protein Data Bank* [Online]. Available: http://www,rcsb.org/pdb/home [Accessed].

Pearlman, R. S. (1987). "Rapid Generation of High Quality Approximate 3d Molecular Structures". *Chemical Design Automation News,* 2**,** 5-6.

Pepperrell, C. A. & Willett, P. (1991). "Techniques for the Calculation of 3-Dimensional Structural Similarity Using Inter-Atomic Distances". *Journal of Computer-Aided Molecular Design,* 5**,** 455-474.

Rand, W. (1971). "Objective Criteria for the Evaluation of Clustering Methods". *Journal of the American Statistical Association,* 66**,** 846-850.

Raymond, J. W., Blankley, C. J. & Willett, P. (2003). "Comparison of Chemical Clustering Methods Using Graph- and Fingerprint-Based Similarity Measures". *Journal of Molecular Graphics & Modelling,* 21**,** 421-433.

Richon, A. B. (1994). *An Introduction to Molecular Modeling* [Online]. Available: http://www.netsci.org/Science/Compchem/feature01.html [Accessed 2010].

Rogers, D. & Hahn, M. (2010). "Extended-Connectivity Fingerprints". *Journal of Chemical Information and Modeling,* 50**,** 742-754.

Rohrer, S. G. & Baumann, K. (2009). "Maximum Unbiased Validation (Muv) Data Sets for Virtual Screening Based on Pubchem Bioactivity Data". *Journal of Chemical Information and Modeling,* 49**,** 169-184.

Salim, N., Holliday, J. & Willett, P. (2003). "Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion". *Journal of Chemical Information and Computer Sciences,* 43**,** 435-442.

Salton, G. (1986). "On the Use of Knowledge-Based Processing in Automatic Text Retrieval". *Proceedings of the American Society for Information Science,* 23**,** 277-287.

Salton, G. & Buckley, C. (1988). "Term-Weighting Approaches in Automatic Text Retrieval". *Information Processing & Management,* 24**,** 513-523.

Sastry, M., Lowrie, J. F., Dixon, S. L. & Sherman, W. (2010). "Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments". *Journal of Chemical Information and Modeling,* 50**,** 771-784.

Schnecke, V. & Kuhn, L. A. (2000). "Virtual Screening with Solvation and Ligand-Induced Complementarity". *Perspectives in Drug Discovery and Design,* 20**,** 171-190.

Schuffenhauer, A., Brown, N., Ertl, P., Jenkins, J. L., Selzer, P. & Hamon, J. (2007). "Clustering and Rule-Based Classifications of Chemical Structures Evaluated in the Biological Activity Space". *Journal of Chemical Information and Modeling,* 47**,** 325-336.

Sesli, M. & Yegenoglu, E. D. (2010). "Comparison of Similarity Coefficients Used for Cluster Analysis Based on Rapd Markers in Wild Olives". *Genetics and Molecular Research,* 9**,** 2248-2253.

Sheridan, R. P. (2007). "Chemical Similarity Searches: When Is Complexity Justified?". *Expert Opinion on Drug Discovery,* 2**,** 423-430.

Siegel, S. d. & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences,* New York ; London, McGraw-Hill.

Snarey, M., Terrett, N. K., Willett, P. & Wilton, D. J. (1997). "Comparison of Algorithms for Dissimilarity-Based Compound Selection". *Journal of Molecular Graphics & Modelling,* 15**,** 372-385.

Sneath, P. H. A. & Sokal, R. R. (1962). "Numerical Taxonomy". *Nature,* 193**,** 855-860.

Snijders, T. A. B., Dormaar, M., Vanschuur, W. H., Dijkmancaes, C. & Driessen, G. (1990). "Distribution of Some Similarity Coefficients for Dyadic Binary Data in the Case of Associated Attributes". *Journal of Classification,* 7**,** 5-31.

Sokal, R. R. & Michener, C. D. (1958). "A Statistical Method for Evaluating Systematic Relationships". *University of Kansas Science Bulletin,* 38**,** 1409-1438.

Sokal, R. R. & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy,* San Francisco ; London, W.H. Freeman and Co.

Sørensen, T. (1948). *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content, and Its Application to Analyses of the Vegetation on Danish Commons*, København.

Sparck Jones, K. (1972). "Statistical Interpretation of Term Specificity and Its Application in Retrieval". *Journal of Documentation,* 28**,** 11-21.

Stumpfe, D. & Bajorath, J. (2011). "Similarity Searching". *Wiley Interdisciplinary Reviews: Computational Molecular Science,* 1**,** 260-282.

Swamidass, S. J., Chen, J., Bruand, J., Phung, P., Ralaivola, L. & Baldi, P. (2005). "Kernels for Small Molecules and the Prediction of Mutagenicity, Toxicity and Anti-Cancer Activity". *Bioinformatics,* 21 (Supplement 1)**,** 359-368.

Takahashi, Y., Sukekawa, M. & Sasaki, S. (1992). "Automatic Identification of Molecular Similarity Using Reduced-Graph Representation of Chemical-Structure". *Journal of Chemical Information and Computer Sciences,* 32**,** 639-643.

Tan, L., Lounkine, E. & Bajorath, J. (2008). "Similarity Searching Using Fingerprints of Molecular Fragments Involved in Protein-Ligand Interactions". *Journal of Chemical Information and Modeling,* 48**,** 2308-2312.

Taylor, R. D., Jewsbury, P. J. & Essex, J. W. (2002). "A Review of Protein-Small Molecule Docking Methods". *Journal of Computer-Aided Molecular Design,* 16**,** 151-166.

Todeschini, R. & Consonni, V. (2000). *Handbook of Molecular Descriptors,* Weinheim ; Chichester, Wiley-VCH.

Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M. & Willett, P. (2012). "Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets". *Journal of Chemical Information and Modeling,* 52**,** 2884-2901.

Truchon, J. F. & Bayly, C. I. (2007). "Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem". *Journal of Chemical Information and Modeling,* 47**,** 488-508.

Tversky, A. (1977). "Features of Similarity". *Psychological Review,* 84**,** 327-352.

Van Rijsbergen, C. J. (1979). *Information Retrieval,* London ; Boston, Butterworths.

Varin, T., Bureau, R., Mueller, C. & Willett, P. (2009). "Clustering Files of Chemical Structures Using the Szekely-Rizzo Generalization of Ward's Method". *Journal of Molecular Graphics & Modelling,* 28**,** 187-195.

Varnek, A. & Baskin, I. I. (2011). "Chemoinformatics as a Theoretical Chemistry Discipline". *Molecular Informatics,* 30**,** 20-32.

Walters, W. P., Stahl, M. T. & Murcko, M. A. (1998). "Virtual Screening - an Overview". *Drug Discovery Today,* 3**,** 160-178.

Wang, Y. & Bajorath, J. (2008). "Balancing the Influence of Molecular Complexity on Fingerprint Similarity Searching". *Journal of Chemical Information and Modeling,* 48**,** 75-84.

Wang, Y. A., Eckert, H. & Bajorath, J. (2007). "Apparent Asymmetry in Fingerprint Similarity Searching Is a Direct Consequence of Differences in Bit Densities and Molecular Size". *ChemMedChem,* 2**,** 1037-1042.

Ward, J. H. (1963). "Hierarchical Grouping to Optimize an Objective Function". *Journal of the American Statistical Association,* 58**,** 236-244.

Warren, G. L., Andrews, C. W., Capelli, A. M., Clarke, B., LaLonde, J., Lambert, M. H., Lindvall, M., Nevins, N., Semus, S. F., Senger, S., Tedesco, G., Wall, I. D., Woolven, J. M., Peishoff, C. E. & Head, M. S. (2006). "A Critical Assessment of Docking Programs and Scoring Functions". *Journal of Medicinal Chemistry,* 49**,** 5912-5931.

Wermuth, G., Ganellin, C. R., Lindberg, P. & Mitscher, L. A. (1998). "Glossary of Terms Used in Medicinal Chemistry (Iupac Recommendations 1998)". *Pure and Applied Chemistry,* 70**,** 1129-1143.

Whittle, M., Gillet, V. J., Willett, P., Alex, A. & Loesel, J. (2004). "Enhancing the Effectiveness of Virtual Screening by Fusing Nearest Neighbor Lists: A Comparison of Similarity Coefficients". *Journal of Chemical Information and Computer Sciences,* 44**,** 1840-1848.

Whittle, M., Gillet, V. J., Willett, P. & Loesel, J. (2006). "Analysis of Data Fusion Methods in Virtual Screening: Similarity and Group Fusion". *Journal of Chemical Information and Modeling,* 46**,** 2206-2219.

Whittle, M., Willett, P., Klaffke, W. & van Noort, P. (2003). "Evaluation of Similarity Measures for Searching the Dictionary of Natural Products Database". *Journal of Chemical Information and Computer Sciences,* 43**,** 449-457.

Wilcoxon, F. (1946). "Individual Comparisons of Grouped Data by Ranking Methods". *Journal of Economic Entomology,* 39**,** 269-270.

Willett, P. (1987). *Similarity and Clustering in Chemical Information Systems*, Research Studies Press Letchworth.

Willett, P. (2000). "Chemoinformatics–Similarity and Diversity in Chemical Libraries". *Current Opinion in Biotechnology,* 11**,** 85-88.

Willett, P. (2006). "Similarity-Based Virtual Screening Using 2D Fingerprints". *Drug Discovery Today,* 11**,** 1046-1053.

Willett, P. (2008). "From Chemical Documentation to Chemoinformatics: Fifty Years of Chemical Information Science". *Journal of Information Science,* 34**,** 477-499.

Willett, P. (2009). "Similarity Methods in Chemoinformatics". *Annual Review of Information Science and Technology,* 43**,** 3-71.

Willett, P., Barnard, J. & Downs, G. (1998). "Chemical Similarity Searching". *Journal of Chemical Information and Computer Sciences,* 38**,** 983-996.

Willett, P. & Winterman, V. (1986). "A Comparison of Some Measures for the Determination of Intermolecular Structural Similarity Measures of Intermolecular Structural Similarity". *Quantitative Structure-Activity Relationships,* 5**,** 18-25.

Willett, P., Winterman, V. & Bawden, D. (1986). "Implementation of Nearest-Neighbor Searching in an Online Chemical-Structure Search System". *Journal of Chemical Information and Computer Sciences,* 26**,** 36-41.

Williams, C. (2006). "Reverse Fingerprinting, Similarity Searching by Group Fusion and Fingerprint Bit Importance". *Molecular Diversity,* 10**,** 311-332.

Wilson, R. (1996). *Introduction to Graph Theory,* Harlow, Longman.

Zavodszky, M. I., Rohatgi, A., Van Voorst, J. R., Yan, H. G. & Kuhn, L. A. (2009). "Scoring Ligand Similarity in Structure-Based Virtual Screening". *Journal of Molecular Recognition,* 22**,** 280-292.

Appendix A: Results of Chapter 4

Table A.1 Average numbers of active molecules retrieved in the top 1% of searches of the WOMBAT database using the Tanimoto coefficient. (W=0.71, chi-square=239.79, p<= 0.001)

similarity measure	Activity class														Mean actives	Median actives	Mean rank
	5HT3	5HT1A	AChE	D2	Renin	PDE4	Thrombin	SubP	HIV P	COX	PKC	ANG	FXa	MMP1			
M11	34.90	78.30	29.70	90.00	274.60	67.60	71.00	89.50	127.90	103.90	48.10	226.90	127.40	74.30	103.15	83.90	6.79
M12	36.70	78.90	32.10	90.90	302.70	68.00	68.60	107.60	124.90	99.80	48.90	246.40	135.50	72.10	108.08	84.90	5.21
M13	13.50	24.90	14.90	20.50	49.00	21.50	3.60	53.80	24.00	41.00	41.10	15.10	17.50	25.80	26.16	22.75	21.86
M14	36.70	80.00	28.90	90.20	288.10	67.20	69.00	107.40	125.20	107.10	48.70	229.00	129.90	74.30	105.84	85.10	5.64
M15	35.40	71.80	26.00	77.20	211.20	58.10	46.10	83.00	98.80	102.80	49.50	161.90	97.00	66.90	84.69	74.50	10.64
M21	39.90	75.20	29.60	65.20	123.50	40.70	27.80	20.70	35.30	43.00	41.20	71.90	31.20	55.10	50.02	40.95	16.43
M22	34.10	81.60	24.80	71.90	248.20	36.80	39.10	65.10	73.00	118.70	56.40	174.20	120.90	59.70	86.04	68.50	11.00
M23	24.50	17.90	11.90	9.20	4.70	13.60	1.10	4.90	1.10	10.90	2.70	1.30	2.40	9.20	8.24	7.05	24.43
M24	44.10	81.30	29.20	66.00	168.80	41.00	37.30	40.30	44.60	60.70	44.70	103.60	37.80	69.50	62.06	44.65	14.07
M25	35.50	53.50	28.60	34.60	24.50	29.20	10.10	8.50	8.50	18.10	10.20	15.00	14.80	26.00	22.65	21.30	20.21
M31	5.00	48.80	13.50	41.70	160.10	21.10	24.00	50.70	94.80	50.00	43.10	128.70	70.70	20.40	55.19	45.95	18.93
M32	1.60	23.60	8.10	14.40	35.70	11.50	6.10	63.60	39.70	13.00	19.80	64.30	36.70	11.70	24.99	17.10	22.57
M33	33.80	88.80	28.90	62.50	198.20	29.10	33.70	55.60	77.90	79.50	38.40	129.40	90.50	58.50	71.77	60.50	14.50
M34	5.10	44.30	12.50	33.10	140.60	22.00	22.30	80.60	91.30	43.30	41.60	131.60	86.40	28.60	55.95	42.45	18.43
M35	18.20	87.30	16.10	67.90	254.10	33.70	41.70	59.60	102.20	79.80	50.90	168.00	90.70	55.80	80.43	63.75	12.57
M41	36.50	82.60	27.30	92.60	263.40	57.60	69.20	84.00	117.80	108.40	48.80	225.90	121.20	76.80	100.86	83.30	7.36
M42	26.50	68.60	21.30	66.80	226.00	42.80	39.00	98.50	85.50	75.40	48.80	166.80	136.40	45.80	82.01	67.70	12.71
M43	28.00	37.50	10.60	26.30	66.10	21.00	6.20	41.70	22.60	65.50	35.90	26.20	22.30	31.80	31.55	27.15	21.21
M44	33.00	83.50	26.30	88.10	271.80	51.10	62.10	100.10	119.20	117.10	52.30	223.00	140.90	73.70	103.01	85.80	7.21
M45	37.60	74.20	24.00	72.50	188.20	47.90	42.90	63.30	72.40	102.10	49.10	149.50	89.00	65.80	77.04	69.10	12.14
M51	34.60	83.30	33.20	100.10	287.20	69.90	75.00	75.10	126.90	99.60	47.80	257.60	132.60	77.80	107.19	80.55	5.71
M52	34.90	84.70	32.70	100.70	280.80	62.40	58.50	88.00	107.90	95.70	48.60	264.50	122.80	84.30	104.75	86.35	6.57
M53	25.10	53.30	15.50	40.40	154.10	30.70	12.20	72.60	42.50	89.70	55.60	82.00	62.00	48.80	56.04	51.05	16.57
M54	35.10	85.70	33.20	103.20	290.00	67.60	70.00	93.60	122.30	101.50	48.50	269.30	128.30	84.10	109.46	89.65	4.57
M55	36.30	80.80	28.20	91.30	273.10	62.30	66.70	96.50	119.00	109.90	48.50	226.00	127.10	77.90	103.11	86.05	7.07

Table A.2 Average numbers of active molecules retrieved in the top1% of searches of the WOMBAT database using the cosine coefficient. (W=0.60, chi-square=202.83, p<=0.001)

Similarity measure	Activity class														Mean actives	Median actives	Mean rank
	5HT3	5HT1A	AChE	D2	Renin	PDE4	Thrombin	SubP	HIV P	COX	PKC	ANG	FXa	MMP1			
M11	34.90	77.00	30.10	89.60	266.90	70.10	68.80	84.20	126.60	102.10	48.60	222.60	127.70	74.10	101.66	80.60	7.36
M12	35.80	69.20	24.80	70.30	210.70	54.10	27.20	85.50	68.50	88.20	50.80	103.20	76.80	56.10	72.94	68.85	15.50
M13	21.60	35.60	17.70	29.80	93.50	32.50	5.90	69.30	27.60	59.30	46.20	35.90	38.50	34.50	39.14	35.05	23.14
M14	36.50	74.80	26.90	81.70	252.50	65.90	55.30	100.60	108.50	105.10	49.90	190.30	113.60	69.10	95.05	78.25	10.07
M15	35.90	76.40	28.30	87.70	263.60	69.00	65.00	93.70	118.60	104.50	48.80	211.40	122.10	72.30	99.81	82.05	8.57
M21	32.50	95.30	29.20	100.40	263.10	51.90	61.30	58.20	116.20	95.20	47.80	238.10	108.20	67.50	97.49	81.35	10.43
M22	27.30	75.30	22.70	63.80	241.40	35.50	39.50	70.80	69.70	99.00	56.30	181.90	125.20	55.30	83.12	66.75	14.21
M23	22.20	69.60	14.70	49.90	208.60	32.30	30.50	67.40	67.70	93.30	52.50	159.80	112.70	53.00	73.87	60.20	18.29
M24	31.10	91.10	28.60	86.30	263.10	47.30	61.10	81.80	120.30	107.00	52.10	237.10	129.30	76.60	100.91	84.05	8.00
M25	31.90	92.30	30.30	98.20	268.00	48.10	63.40	75.40	125.80	104.30	49.30	250.30	123.70	77.00	102.71	84.65	7.07
M31	14.20	70.80	14.60	54.40	202.70	26.30	30.60	45.50	101.30	59.50	44.90	144.00	73.90	27.70	65.03	49.95	21.79
M32	28.90	75.30	19.70	59.10	225.40	28.90	25.30	43.20	39.80	62.00	43.80	108.70	67.40	47.60	62.51	45.70	20.79
M33	32.10	85.70	25.00	58.80	201.10	28.00	31.40	49.80	60.10	69.50	35.00	110.00	75.50	54.30	65.45	56.55	18.57
M34	24.50	88.20	19.80	64.10	240.40	35.00	42.20	67.10	109.10	82.30	52.00	171.50	101.70	60.50	82.74	65.60	14.86
M35	20.60	81.50	17.10	58.40	229.80	31.80	38.40	62.50	108.10	76.60	49.80	166.40	89.80	46.80	76.97	60.45	17.64
M41	32.60	81.90	30.90	93.30	269.70	61.30	69.40	79.40	130.40	105.60	48.70	247.00	135.70	73.90	104.27	80.65	5.64
M42	31.40	74.40	22.40	75.00	244.90	41.20	36.10	82.50	75.40	112.50	58.90	152.50	124.50	54.10	84.70	74.70	12.79
M43	25.50	56.30	11.80	43.20	192.40	32.10	18.40	78.90	48.40	104.50	54.90	105.90	99.90	43.70	65.42	51.65	19.29
M44	31.90	81.00	27.00	86.30	265.10	52.70	60.60	95.40	116.80	113.70	52.80	218.50	137.80	71.40	100.79	83.65	7.64
M45	31.60	81.20	29.80	90.00	269.30	56.50	66.60	90.70	127.30	110.40	49.30	238.80	140.30	75.10	104.06	85.60	5.86
M51	35.10	79.50	29.80	93.80	268.90	67.50	69.20	82.20	125.00	104.70	48.30	231.40	129.80	75.70	102.92	80.85	6.64
M52	37.60	73.80	23.80	72.40	227.00	47.50	27.20	81.40	67.30	105.30	57.00	123.90	83.50	63.10	77.91	69.85	13.71
M53	27.60	49.20	17.20	37.00	133.70	30.90	8.30	68.30	30.80	83.00	51.90	54.80	50.80	43.90	49.10	46.55	21.29
M54	37.10	79.80	26.40	87.50	257.90	60.70	55.80	97.00	107.40	111.20	52.40	200.80	118.10	75.30	97.67	83.65	8.36
M55	36.30	80.00	29.10	92.30	265.60	64.70	65.90	93.50	118.10	108.50	48.60	221.10	125.80	76.20	101.84	86.15	6.86

Table A.3 Average numbers of active molecules retrieved in the top1% of searches of the WOMBAT database using the MinMax coefficient. (W=0.73, chi-square=245.40, p<= 0.001)

Similarity measure	Activity class														Mean actives	Median actives	Mean rank
	5HT3	5HT1A	AChE	D2	Renin	PDE4	Thrombin	SubP	HIV P	COX	PKC	ANG	FXa	MMP1			
M11	34.90	78.30	29.70	90.00	274.60	67.60	71.00	89.50	127.90	103.90	48.10	226.90	127.40	74.30	103.15	83.90	8.57
M12	35.20	80.40	31.40	92.00	292.40	71.40	74.10	85.20	132.80	101.50	47.40	244.40	136.10	75.80	107.15	82.80	6.21
M13	15.90	31.50	16.70	24.80	56.90	27.10	5.50	63.10	30.40	60.00	47.20	27.50	32.60	33.30	33.75	30.95	20.93
M14	34.90	79.30	29.90	91.00	283.40	69.30	71.80	89.60	131.00	103.00	48.20	233.40	131.70	75.60	105.15	84.45	7.36
M15	36.50	74.90	26.20	82.50	243.10	61.90	57.40	91.70	110.90	105.50	48.80	191.80	109.70	71.00	93.71	78.70	11.21
M21	38.10	75.40	28.70	81.00	223.80	59.10	63.80	54.90	81.00	92.20	46.70	193.00	100.70	68.70	86.22	72.05	14.00
M22	40.80	89.70	32.20	92.80	284.70	52.80	59.60	107.60	111.20	122.90	55.20	220.30	136.60	82.60	106.36	91.25	5.29
M23	15.00	21.00	13.60	11.10	8.90	17.30	1.20	10.30	9.80	13.30	21.50	5.20	4.60	11.80	11.76	11.45	24.43
M24	41.60	85.20	30.00	80.80	250.50	57.80	62.40	75.40	83.80	94.00	47.70	194.20	100.90	81.90	91.87	81.35	10.79
M25	37.90	70.10	28.40	66.40	156.60	52.00	45.50	41.00	59.60	85.50	46.60	139.60	77.40	59.30	68.99	59.45	17.00
M31	6.70	60.60	13.60	50.80	251.50	28.10	29.20	53.10	111.90	49.50	43.90	127.90	69.00	27.60	65.96	50.15	19.14
M32	2.70	31.90	11.50	22.80	43.00	14.80	13.80	48.70	73.30	30.40	38.70	80.80	54.40	16.60	34.53	31.15	22.79
M33	35.20	91.50	30.70	70.20	225.10	32.60	38.90	73.80	82.40	93.70	46.00	132.40	91.20	66.20	79.28	72.00	14.71
M34	4.60	45.70	10.50	37.40	147.80	21.40	22.10	61.50	101.00	43.40	41.90	119.60	79.20	24.00	54.29	42.65	20.86
M35	19.00	73.60	16.20	59.60	285.00	38.90	39.30	67.30	92.30	73.90	50.00	160.90	80.40	49.60	79.00	63.45	15.79
M41	37.00	78.30	26.80	88.90	263.60	65.20	71.00	82.80	106.70	101.90	47.90	219.50	121.80	73.40	98.91	80.55	10.43
M42	32.00	83.50	28.70	86.90	271.10	58.10	69.10	104.70	141.90	107.20	50.50	229.60	147.40	76.50	106.23	85.20	7.64
M43	18.40	29.50	12.20	18.90	46.00	23.10	4.90	40.80	19.70	46.40	38.50	21.70	19.00	29.30	26.31	22.40	23.07
M44	36.90	85.80	28.70	94.40	286.40	61.30	72.30	112.60	126.30	115.30	51.60	231.50	135.00	82.80	108.64	90.10	4.71
M45	38.40	74.40	25.20	77.90	219.30	58.00	57.00	74.30	86.10	100.90	48.10	183.90	102.00	68.60	86.72	74.35	13.64
M51	36.30	84.80	30.50	100.20	283.20	66.80	73.00	82.60	119.60	103.10	47.80	244.80	125.50	81.00	105.66	83.70	6.93
M52	35.60	84.10	32.60	102.40	288.30	69.90	73.70	74.40	123.40	99.40	46.80	254.00	126.20	80.40	106.51	82.25	6.57
M53	24.60	49.10	16.50	37.40	117.90	35.20	9.50	72.10	40.50	82.20	54.40	54.30	54.00	49.90	49.83	49.50	18.43
M54	35.90	85.10	31.60	101.50	287.40	68.50	74.10	79.40	120.90	102.00	47.60	249.30	126.70	80.90	106.49	83.00	6.07
M55	35.60	81.50	29.10	91.40	277.60	64.90	70.70	98.70	121.90	107.30	49.00	228.60	128.90	75.80	104.36	86.45	7.50

Table A.4 Average numbers of active molecules retrieved in the top1% of searches of the MUV database using the Tanimoto coefficient (a), the cosine coefficient (b) and the MinMax coefficient (c).

Similarity measure	Activity class																	Mean actives	Median actives
	aid466	aid548	aid600	aid644	aid652	aid689	aid692	aid712	aid713	aid733	aid737	aid810	aid832	aid846	aid852	aid858	aid859		
M11	1.77	3.97	1.90	2.77	1.73	2.13	1.40	2.07	1.77	1.93	1.60	1.87	4.00	4.03	3.97	1.57	1.37	2.34	1.90
M12	1.73	3.80	1.83	3.10	2.07	1.90	1.43	2.23	1.80	1.93	1.63	1.73	3.67	3.57	4.07	1.63	1.50	2.33	1.90
M13	0.53	2.03	0.73	2.30	0.60	0.67	0.70	0.97	0.80	0.73	0.50	0.83	0.43	1.00	0.93	0.70	1.30	0.93	0.73
M14	1.77	3.87	1.80	3.00	1.93	2.00	1.43	2.10	1.87	1.90	1.57	1.80	3.83	3.77	3.90	1.60	1.37	2.32	1.90
M15	1.73	4.07	1.83	2.87	1.73	2.03	1.43	1.90	1.80	1.93	1.50	1.80	3.80	3.83	4.17	1.57	1.40	2.32	1.83
M21	1.50	3.97	1.70	2.40	1.43	1.60	1.33	1.50	1.70	1.80	1.33	1.83	2.90	3.30	3.60	1.60	1.10	2.04	1.70
M22	1.67	3.50	1.93	3.30	1.90	1.50	1.30	2.47	1.80	1.57	1.73	1.80	2.90	3.17	3.50	1.70	1.60	2.20	1.80
M23	0.40	1.70	0.60	2.23	0.40	0.30	0.80	1.10	0.80	0.80	0.20	1.27	1.00	0.70	0.67	0.73	0.63	0.84	0.73
M24	1.43	3.93	1.93	3.23	1.57	1.47	1.40	1.67	1.73	1.70	1.53	1.80	3.10	3.27	4.20	1.73	1.27	2.17	1.73
M25	1.23	3.43	1.73	2.50	1.13	1.53	1.43	1.37	1.53	1.50	1.17	1.67	2.57	3.13	2.97	1.73	1.03	1.86	1.53
M31	0.83	1.27	0.97	1.03	0.97	1.27	0.53	1.17	1.03	0.83	1.30	1.30	0.67	0.60	0.70	0.90	1.03	0.96	0.97
M32	1.47	1.07	1.17	1.27	1.13	1.60	0.57	1.73	1.27	0.97	1.00	1.43	0.77	0.17	0.60	1.40	2.07	1.16	1.17
M33	1.53	3.20	1.97	2.90	1.97	1.43	1.33	1.83	1.47	1.43	1.77	1.53	1.70	2.00	2.20	1.47	1.53	1.84	1.70
M34	1.20	1.57	1.37	1.47	1.37	1.43	0.63	1.43	1.23	1.00	1.33	1.43	1.00	0.53	0.87	1.40	1.63	1.23	1.37
M35	1.10	2.50	1.70	1.97	1.37	1.27	1.03	1.50	1.43	1.10	1.43	1.47	1.17	1.80	1.50	1.13	1.20	1.45	1.43
M41	1.80	4.13	1.83	3.00	1.83	2.07	1.40	2.17	1.90	1.93	1.70	1.73	3.87	3.93	4.20	1.60	1.37	2.38	1.90
M42	1.63	3.00	2.00	2.90	2.03	2.10	1.43	2.80	1.80	1.80	1.63	1.77	3.33	3.00	3.30	1.80	1.70	2.24	2.00
M43	0.87	2.73	1.10	2.97	1.10	1.03	1.13	1.70	1.27	1.17	0.70	1.40	1.20	1.10	1.33	1.33	1.20	1.37	1.20
M44	1.80	3.97	1.87	3.30	1.93	1.93	1.47	2.40	1.87	1.90	1.60	1.83	3.70	3.60	4.23	1.80	1.40	2.39	1.90
M45	1.77	4.17	1.83	3.23	1.83	1.87	1.47	2.00	1.83	1.93	1.43	1.83	3.73	3.87	4.33	1.73	1.30	2.36	1.83
M51	1.80	3.73	1.93	2.73	1.90	2.00	1.40	2.07	1.80	1.97	1.77	1.83	3.87	3.77	3.77	1.60	1.37	2.31	1.93
M52	1.67	3.37	1.97	2.73	1.93	1.77	1.43	1.87	1.87	1.73	1.70	1.83	3.33	3.17	3.63	1.67	1.30	2.17	1.87
M53	0.93	2.97	1.00	3.07	1.30	1.23	1.00	1.83	1.23	1.03	0.90	1.30	1.27	1.27	1.13	1.37	1.47	1.43	1.27
M54	1.67	3.70	1.97	2.83	1.97	2.00	1.40	2.10	1.87	1.87	1.63	1.87	3.70	3.57	3.83	1.67	1.37	2.29	1.97
M55	1.77	3.93	1.87	2.97	1.83	2.00	1.43	2.17	1.90	1.87	1.57	1.77	3.93	3.83	4.07	1.60	1.40	2.35	1.87

(a)

Similarity measure	Activity class																		Mean actives	Median actives
	aid466	aid548	aid600	aid644	aid652	aid689	aid692	aid712	aid713	aid733	aid737	aid810	aid832	aid846	aid852	aid858	aid859			
M11	1.80	3.93	1.90	2.87	1.73	2.10	1.40	2.03	1.77	1.97	1.60	1.83	4.00	4.07	4.03	1.57	1.37		2.35	1.90
M12	1.67	3.90	1.70	3.40	1.73	1.83	1.43	1.87	1.80	1.87	1.33	1.80	3.57	3.57	4.17	1.60	1.47		2.28	1.80
M13	0.77	2.67	0.83	2.90	0.93	0.80	0.97	1.30	1.07	1.00	0.73	0.97	0.90	1.17	1.07	0.97	1.37		1.20	0.97
M14	1.77	3.93	1.87	3.03	1.87	1.97	1.43	2.13	1.87	1.93	1.50	1.77	3.80	3.77	4.07	1.63	1.47		2.34	1.87
M15	1.77	4.00	1.87	2.90	1.83	2.07	1.40	2.10	1.87	2.00	1.57	1.83	3.87	3.97	4.10	1.57	1.40		2.36	1.87
M21	1.70	3.80	2.00	2.90	1.93	1.87	1.40	2.00	1.80	1.83	1.77	1.67	3.47	3.63	4.13	1.60	1.33		2.28	1.87
M22	1.73	3.17	1.87	3.27	1.90	1.57	1.27	2.87	1.83	1.57	1.60	1.90	3.03	3.10	3.27	1.97	1.47		2.20	1.90
M23	1.43	2.97	1.87	2.97	1.73	1.30	1.27	2.60	1.63	1.30	1.53	1.50	1.60	1.63	1.87	1.67	1.63		1.79	1.63
M24	1.77	3.50	1.90	3.33	1.90	1.83	1.37	2.37	1.87	1.73	1.60	1.93	3.47	3.30	4.07	1.73	1.40		2.30	1.90
M25	1.77	3.67	1.93	3.00	2.00	1.90	1.40	2.27	1.80	1.73	1.73	1.73	3.57	3.57	4.10	1.73	1.37		2.31	1.90
M31	1.03	1.87	1.23	1.63	1.13	1.27	0.87	1.30	1.20	1.07	1.23	1.23	0.97	1.27	1.33	0.93	1.03		1.21	1.23
M32	1.50	3.23	1.97	3.00	1.60	1.33	1.33	1.67	1.53	1.50	1.63	1.80	1.73	2.07	2.33	1.60	1.40		1.84	1.63
M33	1.47	3.20	1.83	2.83	1.73	1.33	1.33	1.77	1.53	1.37	1.60	1.67	1.70	2.00	2.43	1.53	1.47		1.81	1.67
M34	1.30	3.00	1.87	2.57	1.57	1.47	1.30	1.77	1.60	1.37	1.60	1.70	1.50	1.90	1.97	1.47	1.43		1.73	1.60
M35	1.17	2.50	1.77	2.03	1.43	1.30	1.07	1.63	1.50	1.17	1.43	1.50	1.27	1.63	1.67	1.20	1.23		1.50	1.43
M41	1.80	3.90	1.87	3.00	1.97	2.13	1.37	2.30	1.83	1.93	1.73	1.83	3.87	3.90	4.07	1.60	1.40		2.38	1.93
M42	1.57	3.60	1.90	3.53	1.87	1.73	1.47	2.77	1.83	1.83	1.43	1.83	3.33	3.33	3.93	1.83	1.53		2.31	1.83
M43	1.23	3.03	1.43	3.07	1.63	1.57	1.37	2.47	1.43	1.23	1.17	1.37	1.47	1.50	1.50	1.60	1.57		1.68	1.50
M44	1.77	3.87	1.87	3.30	1.93	2.00	1.43	2.47	1.87	1.87	1.60	1.93	3.83	3.60	4.27	1.77	1.40		2.40	1.93
M45	1.80	3.83	1.87	3.13	1.93	2.00	1.43	2.40	1.93	1.90	1.67	1.87	3.87	3.73	4.17	1.67	1.37		2.39	1.93
M51	1.87	3.97	1.87	2.90	1.80	2.13	1.40	2.13	1.83	1.93	1.63	1.83	4.00	3.87	4.00	1.60	1.40		2.36	1.87
M52	1.67	3.87	1.90	3.50	1.77	1.73	1.40	2.13	1.93	1.83	1.37	1.87	3.37	3.37	4.17	1.77	1.43		2.30	1.87
M53	0.97	2.93	1.10	3.13	1.20	1.17	1.13	1.83	1.23	1.07	0.93	1.27	1.23	1.37	1.13	1.27	1.47		1.44	1.23
M54	1.73	3.83	1.83	3.10	1.93	1.93	1.43	2.23	1.87	1.90	1.53	1.80	3.73	3.67	4.23	1.70	1.47		2.35	1.90
M55	1.77	3.87	1.87	2.93	1.83	2.03	1.43	2.23	1.90	1.87	1.60	1.77	3.90	3.83	4.10	1.60	1.43		2.35	1.87

(b)

Similarity measure	Activity class																	Mean actives	Median actives
	aid466	aid548	aid600	aid644	aid652	aid689	aid692	aid712	aid713	aid733	aid737	aid810	aid832	aid846	aid852	aid858	aid859		
M11	1.77	3.97	1.90	2.77	1.73	2.13	1.40	2.07	1.77	1.93	1.60	1.87	4.00	4.03	3.97	1.57	1.37	2.34	1.90
M12	1.83	3.80	1.87	2.77	1.73	2.10	1.37	2.00	1.77	1.93	1.67	1.83	4.03	4.03	4.00	1.57	1.33	2.33	1.87
M13	0.67	2.47	0.80	2.50	0.70	0.77	0.73	1.37	0.93	0.83	0.67	0.97	0.63	1.03	1.07	0.93	1.33	1.08	0.93
M14	1.77	3.87	1.90	2.80	1.73	2.13	1.40	2.07	1.77	1.93	1.63	1.83	4.00	4.10	3.93	1.57	1.33	2.34	1.90
M15	1.77	4.07	1.80	2.83	1.83	2.13	1.40	2.07	1.87	1.93	1.53	1.77	4.00	3.87	4.17	1.57	1.40	2.35	1.87
M21	1.80	4.13	1.80	2.87	1.77	2.07	1.47	1.90	1.67	2.03	1.57	1.80	3.97	3.97	4.13	1.67	1.23	2.34	1.80
M22	1.73	4.40	1.83	3.40	1.93	1.70	1.40	2.23	1.77	1.90	1.50	1.97	3.53	3.37	4.17	1.83	1.43	2.36	1.90
M23	0.43	1.90	0.63	2.40	0.53	0.43	0.73	0.80	0.63	0.57	0.13	0.97	0.67	0.77	0.77	0.70	0.73	0.81	0.70
M24	1.70	4.13	1.87	3.13	1.83	1.80	1.37	1.90	1.70	2.03	1.50	1.90	3.70	3.77	4.03	1.70	1.40	2.32	1.87
M25	1.87	4.17	1.77	2.93	1.73	2.00	1.47	1.83	1.63	2.00	1.47	1.83	3.77	3.80	4.20	1.70	1.30	2.32	1.83
M31	0.90	1.67	1.00	1.30	0.97	1.33	0.63	1.20	1.10	1.10	1.20	1.10	0.77	1.07	0.90	0.87	0.93	1.06	1.07
M32	0.77	1.07	0.80	1.00	0.90	1.43	0.47	1.33	0.93	0.93	1.00	1.17	0.77	0.27	0.60	0.93	1.17	0.91	0.93
M33	1.57	3.77	1.93	2.97	1.93	1.40	1.33	1.83	1.57	1.43	1.67	1.43	1.77	2.13	2.30	1.47	1.43	1.88	1.67
M34	0.87	1.23	0.97	0.97	1.07	1.33	0.50	1.30	1.17	0.87	1.17	1.33	0.80	0.40	0.73	0.93	1.13	0.99	0.97
M35	1.00	2.43	1.30	1.73	1.00	1.23	0.90	1.33	1.27	1.23	1.20	1.10	1.03	1.53	1.90	1.03	1.07	1.31	1.23
M41	1.87	4.03	1.87	2.97	1.77	2.03	1.47	2.03	1.70	2.03	1.60	1.90	4.07	4.07	4.17	1.60	1.33	2.38	1.90
M42	1.77	3.83	1.80	3.07	1.87	2.13	1.30	2.37	1.83	1.83	1.77	1.90	3.90	3.70	4.17	1.70	1.40	2.37	1.87
M43	0.57	2.13	0.93	2.83	0.70	0.73	0.87	1.17	1.00	0.87	0.40	1.07	0.77	0.97	1.13	1.03	1.13	1.08	0.97
M44	1.77	4.03	1.83	3.10	1.90	1.93	1.37	2.23	1.90	1.90	1.67	1.80	3.87	3.83	4.10	1.73	1.37	2.37	1.90
M45	1.83	4.07	1.77	2.93	1.87	2.00	1.43	2.03	1.77	2.03	1.50	1.77	4.00	3.90	4.20	1.63	1.37	2.36	1.87
M51	1.73	3.87	1.90	2.77	1.83	2.00	1.40	2.10	1.77	1.90	1.70	1.80	3.83	3.80	3.83	1.63	1.40	2.31	1.90
M52	1.70	3.73	1.90	2.73	1.93	1.93	1.40	2.03	1.80	1.93	1.73	1.80	3.93	3.73	3.83	1.63	1.40	2.30	1.93
M53	0.87	2.90	1.13	3.07	1.10	1.07	1.03	1.67	1.23	1.03	0.87	1.20	1.10	1.33	1.10	1.23	1.43	1.37	1.13
M54	1.73	3.77	1.90	2.80	1.87	2.00	1.40	2.10	1.80	1.93	1.73	1.80	3.93	3.77	3.83	1.63	1.40	2.32	1.90
M55	1.90	3.93	1.90	2.90	1.80	2.00	1.37	2.20	1.87	1.97	1.67	1.83	4.00	3.93	4.07	1.63	1.33	2.37	1.90

(c)

Appendix B: Results of Chapter 5

Table B.1 Median numbers of active molecules retrieved in the top 1% of searches of the MDDR database using the 44 coefficients

	5HT	5HT3	5HT1A	AT1	COX	D2	HIVP	PKC	Renin	SubP	Thrombin
SM	12.50	57.00	38.00	126.50	25.00	14.00	20.00	18.50	26.00	21.50	15.00
RT	12.50	57.00	38.00	126.50	25.00	14.00	20.00	18.50	26.00	21.50	15.00
JT	20.00	79.50	44.00	280.50	18.00	18.50	65.00	22.50	481.00	40.00	32.50
Gle	20.00	79.50	44.00	280.50	18.00	18.50	65.00	22.50	481.00	40.00	32.50
RR	14.50	29.50	27.00	320.50	11.50	17.50	72.00	16.00	665.00	36.00	45.50
For	19.00	80.00	43.00	282.00	17.50	18.00	70.00	22.00	511.00	39.00	33.00
Sim	15.00	59.00	37.50	218.50	13.50	16.50	72.00	17.00	143.00	29.50	37.00
BB	19.00	71.00	39.50	252.00	19.50	19.00	44.00	24.50	358.50	41.50	30.50
DK	19.00	80.00	43.00	282.00	17.50	18.00	70.00	22.00	511.00	39.00	33.00
BUB	17.50	81.50	43.50	248.50	19.00	16.50	42.00	24.50	296.00	36.00	28.50
Kul	19.00	78.00	41.50	277.00	17.00	17.50	73.00	22.00	521.00	36.50	34.50
SS1	20.00	79.50	44.00	280.50	18.00	18.50	65.00	22.50	481.00	40.00	32.50
SS2	12.50	57.00	38.00	126.50	25.00	14.00	20.00	18.50	26.00	21.50	15.00
Ja	20.00	79.50	44.00	280.50	18.00	18.50	65.00	22.50	481.00	40.00	32.50
Fai	15.50	73.50	41.50	214.50	23.50	16.00	33.00	19.50	154.50	32.50	27.50
Mou	15.00	76.50	42.50	194.00	21.00	15.00	46.00	21.50	183.00	33.50	27.50
Mic	16.00	45.00	29.00	323.00	11.50	19.50	74.50	20.00	666.00	36.50	45.00
RG	19.00	81.00	44.50	277.00	18.00	18.00	61.00	21.50	438.50	38.00	31.00
HD	18.50	75.50	43.00	268.00	18.50	17.50	46.00	14.00	383.00	38.00	30.00
Yu1	16.00	77.50	43.50	204.50	19.00	16.00	58.50	24.50	268.00	34.50	31.00
Yu2	16.00	77.50	43.50	204.50	19.00	16.00	58.50	24.50	268.00	34.50	31.00
Fos	18.50	79.50	43.00	285.00	17.50	18.50	70.50	21.50	518.00	39.50	33.50

Den	19.50	81.00	43.00	274.00	18.50	18.00	64.00	23.50	442.00	38.00	32.00
Co1	16.50	49.50	29.00	328.50	11.50	19.50	79.00	20.00	668.50	39.50	46.50
Co2	15.50	73.50	41.00	155.00	25.00	15.50	33.50	21.00	54.00	31.50	24.50
dis	16.50	55.50	31.00	330.50	13.00	19.50	86.00	22.00	658.00	41.50	47.00
GK	20.00	79.50	43.50	275.50	16.50	18.50	65.00	21.00	481.50	40.00	32.00
SS3	18.50	79.50	42.50	264.00	17.00	17.00	67.00	23.00	483.00	36.50	33.00
SS4	19.00	80.00	42.50	281.00	17.50	18.00	67.00	23.00	487.50	39.00	33.50
Phi	19.50	81.00	43.00	278.00	18.00	18.00	65.50	23.50	465.00	38.50	33.00
Di1	14.50	29.50	27.00	320.50	11.50	17.50	72.00	16.00	665.00	36.00	45.50
Di2	15.50	73.50	41.00	155.00	25.00	15.50	33.50	21.00	54.00	31.50	24.50
Sor	19.00	80.00	43.00	282.00	17.50	18.00	70.00	22.00	511.00	39.00	33.00
Coh	19.50	81.00	44.00	277.00	18.00	18.00	62.00	22.50	444.50	38.00	31.00
Pe1	16.50	55.50	31.00	330.50	13.00	19.50	86.00	22.00	658.00	41.50	47.00
Pe2	15.50	73.50	42.50	162.50	24.50	15.50	34.50	22.50	74.00	32.50	27.50
MP	19.50	81.00	43.50	277.00	18.00	18.00	62.00	22.50	449.00	39.00	31.00
HL	18.50	76.50	42.00	288.00	18.00	18.50	71.00	21.50	528.00	41.00	33.50
CT1	12.50	57.00	38.00	126.50	25.00	14.00	20.00	18.50	26.00	21.50	15.00
CT2	12.50	57.00	38.00	126.50	25.00	14.00	20.00	18.50	26.00	21.50	15.00
CT3	14.50	29.50	27.00	320.50	11.50	17.50	72.00	16.00	665.00	36.00	45.50
CT4	20.50	74.00	40.50	299.00	16.50	18.50	79.00	22.50	568.50	39.50	37.00
CT5	16.00	77.50	43.50	204.50	19.00	16.00	58.50	24.50	268.00	34.50	31.00
AC	12.50	57.00	38.00	126.50	25.00	14.00	20.00	18.50	26.00	21.50	15.00

Table B.2 Median numbers of active molecules retrieved in the top 1% of searches of the WOMBAT database using the 44 coefficients

	5HT1A	5HT3	AChE	ANG	COX	D2	Fxa	HIVP	MMP1	PDE4	PKC	RENIN	SUBP	THR
SM	44.50	18.50	21.50	71.00	26.50	34.00	42.50	26.00	22.00	24.50	16.50	46.50	33.50	16.50
RT	44.50	18.50	21.50	71.00	26.50	34.00	42.50	26.00	22.00	24.50	16.50	46.50	33.50	16.50
JT	58.50	26.50	30.00	224.00	28.00	67.00	102.50	68.50	48.50	51.50	13.50	325.00	73.50	69.00
Gle	58.50	26.50	30.00	224.00	28.00	67.00	102.50	68.50	48.50	51.50	13.50	325.00	73.50	69.00
RR	53.50	20.50	40.00	238.50	18.50	43.50	99.50	100.00	37.00	44.50	12.00	278.50	59.00	47.50
For	58.50	24.50	30.50	222.50	27.00	68.00	102.00	72.50	45.50	52.50	16.00	309.50	71.00	68.50
Sim	49.00	22.50	35.50	156.50	24.50	42.00	100.00	98.50	27.50	37.50	16.50	162.50	47.50	32.50
BB	65.50	30.00	25.00	233.00	29.00	77.00	102.00	49.50	49.00	42.00	12.00	331.00	89.00	69.00
DK	58.50	24.50	30.50	222.50	27.00	68.00	102.00	72.50	45.50	52.50	16.00	309.50	71.00	68.50
BUB	56.50	27.50	23.50	192.00	28.50	56.00	88.00	51.00	44.50	36.50	16.50	254.00	73.50	52.00
Kul	56.50	24.50	30.50	212.50	27.00	63.50	102.00	84.00	41.50	52.50	16.50	274.50	67.00	66.00
SS1	58.50	26.50	30.00	224.00	28.00	67.00	102.50	68.50	48.50	51.50	13.50	325.00	73.50	69.00
SS2	44.50	18.50	21.50	71.00	26.50	34.00	42.50	26.00	22.00	24.50	16.50	46.50	33.50	16.50
Ja	58.50	26.50	30.00	224.00	28.00	67.00	102.50	68.50	48.50	51.50	13.50	325.00	73.50	69.00
Fai	51.00	24.50	25.50	151.50	27.50	45.00	78.00	34.00	31.50	26.00	19.00	197.50	66.00	41.50
Mou	49.00	23.00	23.50	142.50	28.00	45.50	89.50	46.00	34.00	40.50	15.50	173.00	59.00	28.00
Mic	55.00	21.00	41.50	253.00	23.00	47.00	100.50	91.50	38.50	45.50	12.00	306.00	62.50	59.50
RG	59.50	26.50	28.00	216.00	28.00	67.50	102.00	64.00	48.50	51.50	15.00	308.50	74.50	64.00
HD	59.50	28.50	26.00	202.50	26.00	60.50	102.00	58.00	47.50	37.00	16.00	283.00	75.50	60.50
Yu1	50.50	24.50	24.50	161.00	30.50	47.00	102.00	53.50	37.00	44.00	17.00	201.50	61.00	42.00
Yu2	50.50	24.50	24.50	161.00	30.50	47.00	102.00	53.50	37.00	44.00	17.00	201.50	61.00	42.00
Fos	58.50	25.00	30.50	225.50	27.00	67.50	102.00	73.50	46.50	53.00	15.50	313.00	70.00	70.50
Den	57.00	27.00	28.00	205.50	28.00	63.00	102.00	68.00	42.50	50.00	16.50	268.00	72.00	62.50
Co1	56.00	21.00	43.00	255.50	23.00	49.00	101.50	91.50	39.00	45.50	12.00	312.50	66.50	63.50
Co2	48.00	24.00	21.00	104.00	27.00	43.00	70.00	36.00	29.00	28.50	16.50	52.50	52.50	17.50
dis	56.50	21.00	43.50	262.00	23.00	63.00	106.00	91.00	42.50	53.50	12.00	333.00	73.00	82.50

GK	58.00	26.50	30.00	222.50	27.50	67.00	102.50	68.50	48.50	51.50	13.50	324.50	73.00	69.00
SS3	56.50	25.00	30.50	200.50	27.00	63.50	102.00	76.50	42.00	51.50	17.50	262.00	68.50	62.00
SS4	58.50	26.50	30.50	216.50	27.50	68.00	102.00	72.50	45.50	51.50	16.00	301.50	70.50	66.50
Phi	58.00	27.00	30.00	210.00	27.50	63.50	102.00	70.50	44.50	51.00	16.50	281.50	71.00	64.50
Di1	53.50	20.50	40.00	238.50	18.50	43.50	99.50	100.00	37.00	44.50	12.00	278.50	59.00	47.50
Di2	48.00	24.00	21.00	104.00	27.00	43.00	70.00	36.00	29.00	28.50	16.50	52.50	52.50	17.50
Sor	58.50	24.50	30.50	222.50	27.00	68.00	102.00	72.50	45.50	52.50	16.00	309.50	71.00	68.50
Coh	59.00	26.50	28.00	215.50	28.00	67.00	102.00	64.50	48.00	51.50	15.50	307.50	74.00	64.00
Pe1	56.50	21.00	43.50	262.00	23.00	63.00	106.00	91.00	42.50	53.50	12.00	333.00	73.00	82.50
Pe2	48.00	24.00	21.50	112.50	27.00	43.50	72.00	37.00	29.00	28.00	16.50	70.00	55.50	17.50
MP	59.00	26.50	30.00	215.00	28.00	64.50	102.00	66.50	47.50	51.50	15.50	305.00	73.50	65.00
HL	60.00	26.00	30.50	239.50	27.50	68.50	104.00	71.50	48.00	54.50	12.50	338.50	71.50	70.00
CT1	44.50	18.50	21.50	71.00	26.50	34.00	42.50	26.00	22.00	24.50	16.50	46.50	33.50	16.50
CT2	44.50	18.50	21.50	71.00	26.50	34.00	42.50	26.00	22.00	24.50	16.50	46.50	33.50	16.50
CT3	53.50	20.50	40.00	238.50	18.50	43.50	99.50	100.00	37.00	44.50	12.00	278.50	59.00	47.50
CT4	59.50	23.50	32.50	245.00	27.00	71.00	107.00	81.00	47.50	57.50	12.50	348.00	72.00	75.00
CT5	50.50	24.50	24.50	161.00	30.50	47.00	102.00	53.50	37.00	44.00	17.00	201.50	61.00	42.00
AC	44.50	18.50	21.50	71.00	26.50	34.00	42.50	26.00	22.00	24.50	16.50	46.50	33.50	16.50

Table B.3 Median numbers of active molecules retrieved in the top 1% of searches of the MUV database using the 44 coefficients

	aid466	aid548	aid600	aid644	aid652	aid689	aid692	aid712	aid713	aid733	aid737	aid810	aid832	aid846	aid852	aid858	aid859
SM	1.00	3.00	2.00	3.00	1.00	2.00	1.00	1.00	1.00	1.50	1.00	2.00	4.00	3.00	3.50	1.00	1.00
RT	1.00	3.00	2.00	3.00	1.00	2.00	1.00	1.00	1.00	1.50	1.00	2.00	4.00	3.00	3.50	1.00	1.00
JT	1.00	2.50	2.00	2.00	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	3.50	1.00	1.00
Gle	1.00	2.50	2.00	2.00	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	3.50	1.00	1.00
RR	2.00	2.00	2.00	2.00	2.00	2.00	1.00	1.00	2.00	2.00	1.50	1.50	4.00	3.00	2.50	1.00	1.00
For	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	3.50	1.00	1.00
Sim	1.50	2.00	2.00	2.50	2.00	2.00	1.00	1.00	1.00	2.00	1.00	2.00	4.00	3.00	3.50	1.00	1.00
BB	1.00	3.00	1.50	2.00	2.00	2.00	1.00	2.00	2.00	2.00	2.00	1.50	4.00	3.50	3.50	1.00	1.00
DK	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	3.50	1.00	1.00
BUB	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.50	2.00	2.00	1.00	2.00	4.00	4.00	4.00	1.00	1.00
Kul	1.00	2.00	2.00	2.50	2.00	2.00	1.00	1.00	1.50	2.00	1.00	2.00	4.00	4.00	4.00	1.00	1.00
SS1	1.00	2.50	2.00	2.00	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	3.50	1.00	1.00
SS2	1.00	3.00	2.00	3.00	1.00	2.00	1.00	1.00	1.00	1.50	1.00	2.00	4.00	3.00	3.50	1.00	1.00
Ja	1.00	2.50	2.00	2.00	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	3.50	1.00	1.00
Fai	1.00	3.00	1.50	3.00	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.50	3.50	3.50	1.00	1.00
Mou	0.00	2.00	1.00	1.50	1.00	1.00	0.00	0.00	0.00	1.00	0.00	1.00	3.00	2.50	2.50	0.00	0.00
Mic	1.50	2.00	2.00	2.00	2.00	2.00	1.00	1.00	2.00	2.00	1.50	1.50	4.00	3.00	3.00	1.00	1.00
RG	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	4.00	1.00	1.00
HD	1.00	2.50	2.00	2.00	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	4.50	1.00	1.00
Yu1	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	1.50	2.00	1.00	2.00	4.00	4.00	3.50	1.00	1.00
Yu2	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	1.50	2.00	1.00	2.00	4.00	4.00	3.50	1.00	1.00
Fos	1.00	2.50	2.00	3.00	2.00	2.00	1.00	1.00	1.50	2.00	1.00	2.00	4.00	4.00	4.00	1.00	1.00
Den	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	4.00	1.00	1.00
Co1	1.50	2.00	2.00	2.00	2.00	2.00	1.00	1.00	2.00	2.00	1.50	1.50	4.00	3.00	3.00	1.00	1.00
Co2	1.00	3.00	2.00	3.00	1.00	2.00	1.00	1.00	1.00	1.50	1.00	2.00	4.00	3.00	3.50	1.00	1.00
dis	1.50	2.00	2.00	2.00	2.00	2.00	1.00	1.00	2.00	2.00	1.50	1.50	4.00	3.00	3.00	1.00	1.00

GK	1.00	2.50	2.00	3.00	2.00	2.00	1.00	1.00	2.00	2.00	1.00	1.50	4.00	4.00	3.50	1.00	1.00
SS3	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	1.50	2.00	1.00	2.00	4.00	4.00	4.00	1.00	1.00
SS4	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	4.00	1.00	1.00
Phi	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	3.50	1.00	1.00
Di1	2.00	2.00	2.00	2.00	2.00	2.00	1.00	1.00	2.00	2.00	1.50	1.50	4.00	3.00	2.50	1.00	1.00
Di2	1.00	3.00	2.00	3.00	1.00	2.00	1.00	1.00	1.00	1.50	1.00	2.00	4.00	3.00	3.50	1.00	1.00
Sor	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	3.50	1.00	1.00
Coh	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	4.00	1.00	1.00
Pe1	1.50	2.00	2.00	2.00	2.00	2.00	1.00	1.00	2.00	2.00	1.50	1.50	4.00	3.00	3.00	1.00	1.00
Pe2	1.00	3.00	1.50	3.00	1.50	2.00	1.00	1.00	1.00	2.00	1.00	2.00	4.00	3.00	3.50	1.00	1.00
MP	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	4.00	1.00	1.00
HL	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	3.50	1.00	1.00
CT1	1.00	3.00	2.00	3.00	1.00	2.00	1.00	1.00	1.00	1.50	1.00	2.00	4.00	3.00	3.50	1.00	1.00
CT2	1.00	3.00	2.00	3.00	1.00	2.00	1.00	1.00	1.00	1.50	1.00	2.00	4.00	3.00	3.50	1.00	1.00
CT3	2.00	2.00	2.00	2.00	2.00	2.00	1.00	1.00	2.00	2.00	1.50	1.50	4.00	3.00	2.50	1.00	1.00
CT4	1.50	2.00	2.00	2.50	2.00	2.00	1.00	1.00	2.00	2.00	1.00	2.00	4.00	4.00	3.50	1.00	1.00
CT5	1.00	2.50	2.00	2.50	2.00	2.00	1.00	1.00	1.50	2.00	1.00	2.00	4.00	4.00	3.50	1.00	1.00
AC	1.00	3.00	2.00	3.00	1.00	2.00	1.00	1.00	1.00	1.50	1.00	2.00	4.00	3.00	3.50	1.00	1.00

Table B.4 Top 1% retrieval rates of the 44 coefficients in MDDR

	5HT	5HT3	5HT1A	AT1	COX	D2	HIVP	PKC	Renin	SubP	Thrombin
SM	3.48	7.58	4.60	13.42	3.93	3.54	2.67	4.08	2.30	1.73	1.87
RT	3.48	7.58	4.60	13.42	3.93	3.54	2.67	4.08	2.30	1.73	1.87
JT	5.57	10.57	5.32	29.75	2.83	4.68	8.67	4.97	42.57	3.21	4.05
Gle	5.57	10.57	5.32	29.75	2.83	4.68	8.67	4.97	42.57	3.21	4.05
RR	4.04	3.92	3.27	33.99	1.81	4.43	9.60	3.53	58.85	2.89	5.67
For	5.29	10.64	5.20	29.91	2.75	4.56	9.33	4.86	45.22	3.13	4.11
Sim	4.18	7.85	4.53	23.17	2.12	4.18	9.60	3.75	12.66	2.37	4.61
BB	5.29	9.44	4.78	26.72	3.07	4.81	5.87	5.41	31.73	3.33	3.80
DK	5.29	10.64	5.20	29.91	2.75	4.56	9.33	4.86	45.22	3.13	4.11
BUB	4.88	10.84	5.26	26.35	2.99	4.18	5.60	5.41	26.20	2.89	3.55
Kul	5.29	10.37	5.02	29.37	2.67	4.43	9.73	4.86	46.11	2.93	4.30
SS1	5.57	10.57	5.32	29.75	2.83	4.68	8.67	4.97	42.57	3.21	4.05
SS2	3.48	7.58	4.60	13.42	3.93	3.54	2.67	4.08	2.30	1.73	1.87
Ja	5.57	10.57	5.32	29.75	2.83	4.68	8.67	4.97	42.57	3.21	4.05
Fai	4.32	9.77	5.02	22.75	3.70	4.05	4.40	4.31	13.67	2.61	3.43
Mou	4.18	10.17	5.14	20.57	3.30	3.80	6.13	4.75	16.20	2.69	3.43
Mic	4.46	5.98	3.51	34.25	1.81	4.94	9.93	4.42	58.94	2.93	5.60
RG	5.29	10.77	5.38	29.37	2.83	4.56	8.13	4.75	38.81	3.05	3.86
HD	5.15	10.04	5.20	28.42	2.91	4.43	6.13	3.09	33.89	3.05	3.74
Yu1	4.46	10.31	5.26	21.69	2.99	4.05	7.80	5.41	23.72	2.77	3.86
Yu2	4.46	10.31	5.26	21.69	2.99	4.05	7.80	5.41	23.72	2.77	3.86
Fos	5.15	10.57	5.20	30.22	2.75	4.68	9.40	4.75	45.84	3.17	4.17
Den	5.43	10.77	5.20	29.06	2.91	4.56	8.53	5.19	39.12	3.05	3.99
Co1	4.60	6.58	3.51	34.84	1.81	4.94	10.53	4.42	59.16	3.17	5.79
Co2	4.32	9.77	4.96	16.44	3.93	3.92	4.47	4.64	4.78	2.53	3.05
dis	4.60	7.38	3.75	35.05	2.04	4.94	11.47	4.86	58.23	3.33	5.85

GK	5.57	10.57	5.26	29.22	2.59	4.68	8.67	4.64	42.61	3.21	3.99
SS3	5.15	10.57	5.14	28.00	2.67	4.30	8.93	5.08	42.74	2.93	4.11
SS4	5.29	10.64	5.14	29.80	2.75	4.56	8.93	5.08	43.14	3.13	4.17
Phi	5.43	10.77	5.20	29.48	2.83	4.56	8.73	5.19	41.15	3.09	4.11
Di1	4.04	3.92	3.27	33.99	1.81	4.43	9.60	3.53	58.85	2.89	5.67
Di2	4.32	9.77	4.96	16.44	3.93	3.92	4.47	4.64	4.78	2.53	3.05
Sor	5.29	10.64	5.20	29.91	2.75	4.56	9.33	4.86	45.22	3.13	4.11
Coh	5.43	10.77	5.32	29.37	2.83	4.56	8.27	4.97	39.34	3.05	3.86
Pe1	4.60	7.38	3.75	35.05	2.04	4.94	11.47	4.86	58.23	3.33	5.85
Pe2	4.32	9.77	5.14	17.23	3.85	3.92	4.60	4.97	6.55	2.61	3.43
MP	5.43	10.77	5.26	29.37	2.83	4.56	8.27	4.97	39.74	3.13	3.86
HL	5.15	10.17	5.08	30.54	2.83	4.68	9.47	4.75	46.73	3.29	4.17
CT1	3.48	7.58	4.60	13.42	3.93	3.54	2.67	4.08	2.30	1.73	1.87
CT2	3.48	7.58	4.60	13.42	3.93	3.54	2.67	4.08	2.30	1.73	1.87
CT3	4.04	3.92	3.27	33.99	1.81	4.43	9.60	3.53	58.85	2.89	5.67
CT4	5.71	9.84	4.90	31.71	2.59	4.68	10.53	4.97	50.31	3.17	4.61
CT5	4.46	10.31	5.26	21.69	2.99	4.05	7.80	5.41	23.72	2.77	3.86
AC	3.48	7.58	4.60	13.42	3.93	3.54	2.67	4.08	2.30	1.73	1.87

Table B.5 Top 1% retrieval rate of the 44 coefficients in ChEMBL dataset. First 25 classes (a), latter 25 classes (b).

	Target_no_21	*Target_no_35*	*Target_no_42*	*Target_no_52*	*Target_no_54*	*Target_no_129*	*Target_no_140*	*Target_no_142*	*Target_no_143*	*Target_no_146*	*Target_no_147*	*Target_no_148*	*Target_no_152*	*Target_no_163*	*Target_no_168*	*Target_no_171*	*Target_no_181*	*Target_no_186*	*Target_no_195*	*Target_no_196*	*Target_no_210*	*Target_no_211*	*Target_no_213*	*Target_no_220*	*Target_no_230*
SM	30.05	23.95	9.07	21.67	13.69	20.29	19.01	20.25	24.50	35.00	34.25	8.08	13.93	22.98	7.00	19.00	20.29	28.13	17.14	15.33	34.47	9.49	24.44	24.17	33.13
RT	30.05	23.95	9.07	21.67	13.69	20.29	19.01	20.25	24.50	35.00	34.25	8.08	13.93	22.98	7.00	19.00	20.29	28.13	17.14	15.33	34.47	9.49	24.44	24.17	33.13
JT	48.80	29.41	21.86	27.67	14.68	25.88	25.70	41.57	35.25	44.69	47.26	20.88	24.38	23.60	9.00	20.00	27.35	35.76	25.24	16.06	32.04	18.80	35.34	35.42	55.42
Gle	48.80	29.41	21.86	27.67	14.68	25.88	25.70	41.57	35.25	44.69	47.26	20.88	24.38	23.60	9.00	20.00	27.35	35.76	25.24	16.06	32.04	18.80	35.34	35.42	55.42
RR	46.63	15.55	20.31	21.67	12.90	24.71	24.30	39.34	36.50	32.19	44.52	21.28	23.13	23.60	10.00	19.50	25.00	36.81	22.86	20.07	32.04	16.79	30.45	36.67	36.75
For	49.04	31.93	21.75	27.33	14.68	25.88	25.70	40.74	36.25	43.75	47.26	20.64	24.38	23.60	9.00	20.00	27.94	36.11	25.24	16.06	32.04	18.25	34.21	35.00	53.61
Sim	45.67	22.69	13.61	24.00	13.29	24.71	22.54	27.19	34.00	38.44	43.84	12.48	19.15	23.60	9.00	19.50	26.18	35.76	20.00	15.33	34.95	15.33	30.45	31.25	36.75
BB	48.32	26.89	21.75	27.33	14.68	26.18	25.00	43.22	32.75	44.06	47.95	22.24	24.38	22.98	9.00	20.00	25.00	36.46	20.95	16.06	32.04	16.06	31.95	35.83	53.01
DK	49.04	31.93	21.75	27.33	14.68	25.88	25.70	40.74	36.25	43.75	47.26	20.64	24.38	23.60	9.00	20.00	27.94	36.11	25.24	16.06	32.04	18.25	34.21	35.00	53.61
BUB	48.32	28.15	19.90	27.33	14.68	25.59	25.00	38.76	30.75	45.00	45.21	17.60	23.63	22.98	9.00	20.00	25.29	36.46	21.90	16.06	32.04	15.69	33.08	32.92	54.52
Kul	48.56	31.09	20.93	26.33	14.09	25.88	26.06	39.01	36.50	43.13	46.58	19.36	23.63	23.60	9.00	20.00	28.53	36.46	25.24	16.06	33.01	17.70	33.46	35.00	53.01
SS1	48.80	29.41	21.86	27.67	14.68	25.88	25.70	41.57	35.25	44.69	47.26	20.88	24.38	23.60	9.00	20.00	27.35	35.76	25.24	16.06	32.04	18.80	35.34	35.42	55.42
SS2	30.05	23.95	9.07	21.67	13.69	20.29	19.01	20.25	24.50	35.00	34.25	8.08	13.93	22.98	7.00	19.00	20.29	28.13	17.14	15.33	34.47	9.49	24.44	24.17	33.13
Ja	48.80	29.41	21.86	27.67	14.68	25.88	25.70	41.57	35.25	44.69	47.26	20.88	24.38	23.60	9.00	20.00	27.35	35.76	25.24	16.06	32.04	18.80	35.34	35.42	55.42
Fai	45.67	26.89	16.70	24.67	14.68	24.12	22.54	33.72	29.50	43.13	40.41	12.48	20.65	22.98	7.00	20.00	23.82	35.07	20.95	15.69	33.98	12.77	29.32	32.50	50.90
Mou	45.91	28.15	14.74	24.33	14.09	24.41	23.59	30.58	27.75	42.19	39.04	11.60	18.41	22.36	6.00	19.00	24.41	33.68	19.05	14.96	33.01	13.69	28.95	29.17	47.89
Mic	46.88	18.07	21.44	24.00	13.10	25.00	24.30	40.50	37.50	34.69	45.21	23.04	22.89	23.60	9.00	19.50	28.24	36.46	21.90	16.79	32.04	18.43	30.83	36.67	39.76
RG	49.04	29.41	21.86	28.00	14.88	25.88	25.35	41.07	34.50	44.69	46.58	20.24	24.38	23.60	9.00	20.00	27.35	36.11	25.24	16.06	32.04	18.07	34.59	35.00	55.12
HD	48.80	29.83	21.03	27.67	14.88	25.88	25.70	40.41	32.50	45.00	45.89	19.12	23.88	23.29	9.00	20.00	26.18	36.11	24.29	16.06	32.04	15.33	33.83	35.00	54.82
Yu1	46.39	27.31	15.77	25.33	14.48	25.29	24.30	32.48	30.25	44.06	43.15	12.40	20.40	22.98	8.00	20.00	25.00	34.72	20.48	15.69	33.98	15.51	31.58	32.08	50.00
Yu2	46.39	27.31	15.77	25.33	14.48	25.29	24.30	32.48	30.25	44.06	43.15	12.40	20.40	22.98	8.00	20.00	25.00	34.72	20.48	15.69	33.98	15.51	31.58	32.08	50.00
Fos	49.04	31.93	21.75	27.33	14.48	25.88	25.70	40.66	36.50	43.75	47.26	20.72	24.38	23.60	9.00	20.00	28.53	36.11	25.24	16.06	32.04	18.43	34.96	35.42	53.61
Den	48.56	31.93	21.44	27.33	14.88	25.88	25.70	39.59	34.75	43.75	45.89	19.04	24.13	23.60	9.00	20.00	27.35	36.11	25.24	16.06	33.01	17.88	33.46	35.00	54.52
Co1	46.88	18.49	21.44	24.67	13.10	25.00	24.30	40.66	37.50	34.69	45.89	22.56	23.13	23.60	9.00	19.50	28.53	36.11	22.38	17.52	32.04	18.43	31.20	36.67	40.36
Co2	44.23	26.47	12.47	23.67	14.29	23.53	20.77	25.54	25.75	41.88	36.30	9.44	16.42	22.98	8.00	19.50	25.00	33.33	18.57	15.33	34.47	12.04	27.44	27.08	40.36
dis	47.60	19.75	21.55	26.33	13.29	25.00	24.30	40.91	37.75	35.00	45.89	22.56	23.38	23.60	9.00	19.50	28.82	36.11	24.29	17.52	32.04	18.61	31.58	36.67	45.48
GK	47.60	28.99	21.75	27.67	14.68	25.29	25.70	41.40	35.25	44.69	47.26	20.80	24.38	23.60	9.00	20.00	27.35	35.76	24.76	16.06	32.04	18.07	34.21	34.58	54.22
SS3	48.56	31.51	20.41	26.67	14.48	25.59	25.70	38.35	35.50	43.75	45.89	18.64	23.88	23.60	9.00	20.00	28.24	36.11	25.24	16.06	33.01	17.70	33.83	35.00	53.31
SS4	48.80	31.93	21.75	27.67	14.68	25.88	26.06	40.58	35.75	44.06	46.58	20.16	24.13	23.60	9.00	20.00	27.94	36.81	25.24	16.06	32.04	17.88	34.21	35.00	53.61
Phi	48.56	31.93	21.44	27.67	14.88	25.88	25.70	39.92	35.25	44.06	45.89	19.76	24.13	23.60	9.00	20.00	27.65	36.11	25.24	16.06	32.04	18.07	34.21	35.00	53.61
Di1	46.63	15.55	20.31	21.67	12.90	24.71	24.30	39.34	36.50	32.19	44.52	21.28	23.13	23.60	10.00	19.50	25.00	36.81	22.86	20.07	32.04	16.79	30.45	36.67	36.75
Di2	44.23	26.47	12.47	23.67	14.29	23.53	20.77	25.54	25.75	41.88	36.30	9.44	16.42	22.98	8.00	19.50	25.00	33.33	18.57	15.33	34.47	12.04	27.44	27.08	40.36
Sor	49.04	31.93	21.75	27.33	14.68	25.88	25.70	40.74	36.25	43.75	47.26	20.64	24.38	23.60	9.00	20.00	27.94	36.11	25.24	16.06	32.04	18.25	34.21	35.00	53.61
Coh	49.04	29.41	21.86	27.67	14.88	25.88	25.35	41.07	34.50	44.69	46.58	20.24	24.38	23.60	9.00	20.00	27.35	36.11	25.24	16.06	32.04	18.07	34.59	35.00	55.12
Pe1	47.60	19.75	21.55	26.33	13.29	25.00	24.30	40.91	37.75	35.00	45.89	22.56	23.38	23.60	9.00	19.50	28.82	36.11	24.29	17.52	32.04	18.61	31.58	36.67	45.48
Pe2	44.71	27.31	13.40	24.00	14.48	24.12	21.13	27.27	25.75	42.19	36.30	9.68	16.67	22.98	8.00	19.50	25.00	34.38	19.52	15.33	33.98	12.77	27.44	29.17	42.47
MP	49.04	29.41	21.75	27.67	14.88	25.88	25.70	40.99	34.75	44.69	45.89	20.24	24.13	23.60	9.00	20.00	27.35	36.11	25.24	16.06	32.04	18.07	34.59	35.00	55.12
HL	48.80	31.51	21.65	27.33	14.68	25.88	25.35	42.15	36.50	43.75	47.95	22.08	24.13	23.60	9.00	20.00	28.24	36.11	24.29	16.06	32.04	18.43	35.34	35.42	53.31
CT1	30.05	23.95	9.07	21.67	13.69	20.29	19.01	20.25	24.50	35.00	34.25	8.08	13.93	22.98	7.00	19.00	20.29	28.13	17.14	15.33	34.47	9.49	24.44	24.17	33.13
CT2	30.05	23.95	9.07	21.67	13.69	20.29	19.01	20.25	24.50	35.00	34.25	8.08	13.93	22.98	7.00	19.00	20.29	28.13	17.14	15.33	34.47	9.49	24.44	24.17	33.13
CT3	46.63	15.55	20.31	21.67	12.90	24.71	24.30	39.34	36.50	32.19	44.52	21.28	23.13	23.60	10.00	19.50	25.00	36.81	22.86	20.07	32.04	16.79	30.45	36.67	36.75
CT4	48.56	30.67	21.75	27.33	14.68	25.88	24.65	42.56	36.50	43.13	47.95	22.64	24.38	23.60	9.00	20.00	28.53	36.11	24.29	16.79	32.04	19.34	34.59	35.83	52.11
CT5	46.39	27.31	15.77	25.33	14.48	25.29	24.30	32.48	30.25	44.06	43.15	12.40	20.40	22.98	8.00	20.00	25.00	34.72	20.48	15.69	33.98	15.51	31.58	32.08	50.00
AC	30.05	23.95	9.07	21.67	13.69	20.29	19.01	20.25	24.50	35.00	34.25	8.08	13.93	22.98	7.00	19.00	20.29	28.13	17.14	15.33	34.47	9.49	24.44	24.17	33.13

(a)

	Target_no_234	*Target_no_238*	*Target_no_241*	*Target_no_250*	*Target_no_121*	*Target_no_120*	*Target_no_115*	*Target_no_113*	*Target_no_112*	*Target_no_105*	*Target_no_98*	*Target_no_95*	*Target_no_86*	*Target_no_81*	*Target_no_59*	*Target_no_57*	*Target_no_44*	*Target_no_4*	*Target_no_8*	*Target_no_9*	*Target_no_10*	*Target_no_12*	*Target_no_13*	*Target_no_14*	*Target_no_16*
SM	28.53	14.10	24.06	24.79	14.66	24.23	23.91	21.01	30.75	16.79	11.00	16.19	15.69	17.57	32.09	14.69	22.88	35.53	21.36	20.59	8.53	15.61	18.03	16.94	11.22
RT	28.53	14.10	24.06	24.79	14.66	24.23	23.91	21.01	30.75	16.79	11.00	16.19	15.69	17.57	32.09	14.69	22.88	35.53	21.36	20.59	8.53	15.61	18.03	16.94	11.22
JT	35.16	16.22	25.00	40.17	14.94	34.23	45.65	29.26	39.94	38.06	20.09	26.65	27.52	21.39	39.55	30.41	29.85	44.74	31.55	32.05	23.53	23.19	21.31	29.84	24.75
Gle	35.16	16.22	25.00	40.17	14.94	34.23	45.65	29.26	39.94	38.06	20.09	26.65	27.52	21.39	39.55	30.41	29.85	44.74	31.55	32.05	23.53	23.19	21.31	29.84	24.75
RR	28.67	14.89	24.53	38.89	16.67	33.46	41.30	30.05	39.37	35.07	23.55	20.92	23.54	20.57	28.36	26.42	26.67	40.46	26.21	33.77	29.12	23.30	18.85	29.44	23.93
For	34.87	15.96	25.00	40.17	14.94	34.23	45.65	28.99	39.94	36.85	19.55	26.65	26.81	21.39	39.55	30.41	30.00	44.74	32.04	31.44	23.53	22.17	21.31	29.84	21.45
Sim	30.40	14.89	25.00	36.75	14.94	27.31	37.89	27.66	36.78	21.08	12.91	22.78	19.79	20.30	35.82	21.13	29.24	39.14	27.67	24.44	9.71	17.53	18.03	25.00	12.38
BB	35.73	15.96	25.00	38.89	14.94	35.77	44.41	27.93	40.80	39.37	21.18	29.94	29.27	21.93	38.81	31.44	29.24	45.39	27.18	31.74	29.12	23.64	20.08	29.84	25.25
DK	34.87	15.96	25.00	40.17	14.94	34.23	45.65	28.99	39.94	36.85	19.55	26.65	26.81	21.39	39.55	30.41	30.00	44.74	32.04	31.44	23.53	22.17	21.31	29.84	21.45
BUB	35.30	16.22	25.00	38.46	14.94	33.08	42.55	27.93	39.08	32.46	15.82	27.22	26.23	20.98	38.81	28.09	30.76	43.42	31.07	28.40	20.29	19.91	20.49	29.84	18.48
Kul	34.15	15.96	25.00	39.32	14.94	33.46	43.79	28.99	39.94	34.98	18.45	26.36	26.93	20.98	39.55	29.90	29.70	43.75	32.04	30.43	22.65	20.81	19.67	29.84	19.80
SS1	35.16	16.22	25.00	40.17	14.94	34.23	45.65	29.26	39.94	38.06	20.09	26.65	27.52	21.39	39.55	30.41	29.85	44.74	31.55	32.05	23.53	23.19	21.31	29.84	24.75
SS2	28.53	14.10	24.06	24.79	14.66	24.23	23.91	21.01	30.75	16.79	11.00	16.19	15.69	17.57	32.09	14.69	22.88	35.53	21.36	20.59	8.53	15.61	18.03	16.94	11.22
Ja	35.16	16.22	25.00	40.17	14.94	34.23	45.65	29.26	39.94	38.06	20.09	26.65	27.52	21.39	39.55	30.41	29.85	44.74	31.55	32.05	23.53	23.19	21.31	29.84	24.75
Fai	35.45	16.22	25.00	32.48	14.94	30.38	36.02	27.13	37.07	25.47	13.18	24.79	22.25	19.89	38.81	25.64	28.94	41.45	28.64	26.06	14.12	17.76	19.26	28.23	15.02
Mou	33.86	15.43	24.53	33.76	13.79	28.08	36.34	25.27	36.49	22.11	12.36	25.79	21.43	20.16	32.84	23.07	29.09	39.47	27.18	25.35	10.59	17.42	18.03	24.60	12.87
Mic	29.68	15.16	24.53	39.32	17.24	33.85	43.48	29.26	40.23	36.01	23.45	21.49	24.59	21.80	28.36	28.99	27.58	43.75	26.70	33.57	29.41	24.32	19.67	29.84	25.91
RG	35.45	16.22	25.00	40.17	14.94	34.23	45.03	28.46	39.94	36.85	18.09	27.65	27.63	21.39	40.30	30.41	30.30	45.07	32.04	31.34	23.53	21.49	22.13	29.84	22.94
HD	36.17	16.22	25.00	39.74	14.94	34.23	43.17	27.66	39.94	33.96	16.73	25.79	27.40	21.12	40.30	30.03	30.45	44.74	32.04	30.63	22.65	20.36	21.72	29.84	20.79
Yu1	34.29	15.96	25.47	36.75	14.94	30.00	38.51	26.60	37.36	23.41	13.09	26.07	23.42	20.57	38.81	24.48	30.15	40.79	29.13	27.08	13.82	18.21	18.85	28.23	13.70
Yu2	34.29	15.96	25.47	36.75	14.94	30.00	38.51	26.60	37.36	23.41	13.09	26.07	23.42	20.57	38.81	24.48	30.15	40.79	29.13	27.08	13.82	18.21	18.85	28.23	13.70
Fos	35.16	15.96	25.00	40.17	14.94	34.23	45.65	29.26	39.94	37.13	19.82	26.65	26.81	21.39	39.55	30.41	30.00	44.74	32.04	31.74	23.53	22.29	21.31	29.84	21.95
Den	34.87	16.22	25.00	39.74	14.94	33.46	43.48	27.93	39.94	33.96	16.64	26.93	27.05	21.12	40.30	30.03	30.30	45.07	32.04	30.43	22.06	20.36	20.90	29.84	20.30
Co1	30.40	15.43	24.53	39.32	17.24	34.23	43.48	29.26	40.23	36.47	23.00	21.63	24.71	22.48	32.09	29.25	27.58	44.08	29.13	33.67	29.12	24.66	19.67	29.84	26.07
Co2	31.12	15.96	25.47	27.35	14.66	26.15	32.61	22.61	33.62	16.14	11.73	24.79	19.91	19.35	41.79	20.10	29.24	38.16	26.21	22.82	9.12	17.19	18.44	21.77	11.22
dis	31.56	15.96	24.53	39.74	16.95	34.23	44.41	29.52	40.52	36.94	22.36	22.78	24.94	22.48	32.09	29.51	28.33	44.74	29.13	33.27	28.24	24.77	19.67	29.84	26.24
GK	35.16	15.96	23.58	40.17	14.94	34.23	44.41	29.26	39.94	37.78	20.00	25.79	27.52	21.39	39.55	30.03	29.70	44.74	31.07	31.44	23.82	23.19	21.31	29.84	24.75
SS3	34.58	15.96	25.00	39.74	14.94	33.46	43.79	27.93	39.94	33.12	16.45	26.22	26.81	20.98	40.30	29.77	29.85	43.75	32.04	29.82	21.18	20.48	19.67	29.84	18.32
SS4	35.16	15.96	25.00	40.17	14.94	34.23	45.03	28.99	39.94	36.19	19.09	26.65	27.05	21.39	40.30	30.28	30.00	45.07	32.04	31.24	23.53	21.49	21.31	29.84	21.12
Phi	34.58	16.22	25.00	40.17	14.94	33.46	43.79	28.19	39.94	35.26	17.09	26.79	26.93	21.12	40.30	30.15	30.30	45.07	32.04	30.73	22.94	20.59	21.31	29.84	20.46
Di1	28.67	14.89	24.53	38.89	16.67	33.46	41.30	30.05	39.37	35.07	23.55	20.92	23.54	20.57	28.36	26.42	26.67	40.46	26.21	33.77	29.12	23.30	18.85	29.44	23.93
Di2	31.12	15.96	25.47	27.35	14.66	26.15	32.61	22.61	33.62	16.14	11.73	24.79	19.91	19.35	41.79	20.10	29.24	38.16	26.21	22.82	9.12	17.19	18.44	21.77	11.22
Sor	34.87	15.96	25.00	40.17	14.94	34.23	45.65	28.99	39.94	36.85	19.55	26.65	26.81	21.39	39.55	30.41	30.00	44.74	32.04	31.44	23.53	22.17	21.31	29.84	21.45
Coh	35.45	16.22	25.00	40.17	14.94	34.23	45.03	28.72	39.94	36.57	18.00	27.36	27.40	21.39	40.30	30.28	30.30	45.07	32.04	31.34	23.53	21.27	22.13	29.84	22.44
Pe1	31.56	15.96	24.53	39.74	16.95	34.23	44.41	29.52	40.52	36.94	22.36	22.78	24.94	22.48	32.09	29.51	28.33	44.74	29.13	33.27	28.24	24.77	19.67	29.84	26.24
Pe2	31.56	15.96	25.47	29.06	14.94	26.54	33.54	22.61	34.20	18.00	12.00	25.21	20.26	19.89	41.79	21.01	29.39	38.16	26.70	23.23	9.12	17.19	18.44	22.58	11.22
MP	35.30	16.22	25.00	40.17	14.94	34.23	45.03	28.72	39.94	36.10	18.00	27.36	27.40	21.39	40.30	30.28	30.30	45.07	32.04	31.24	23.53	21.15	22.13	29.84	21.95
HL	34.87	15.96	25.00	40.17	14.94	34.23	46.58	29.26	39.94	38.25	20.91	26.93	26.81	21.53	39.55	30.67	29.55	45.07	31.55	32.25	24.41	23.76	21.31	29.84	25.74
CT1	28.53	14.10	24.06	24.79	14.66	24.23	23.91	21.01	30.75	16.79	11.00	16.19	15.69	17.57	32.09	14.69	22.88	35.53	21.36	20.59	8.53	15.61	18.03	16.94	11.22
CT2	28.53	14.10	24.06	24.79	14.66	24.23	23.91	21.01	30.75	16.79	11.00	16.19	15.69	17.57	32.09	14.69	22.88	35.53	21.36	20.59	8.53	15.61	18.03	16.94	11.22
CT3	28.67	14.89	24.53	38.89	16.67	33.46	41.30	30.05	39.37	35.07	23.55	20.92	23.54	20.57	28.36	26.42	26.67	40.46	26.21	33.77	29.12	23.30	18.85	29.44	23.93
CT4	34.58	15.96	25.00	39.74	14.94	34.23	46.27	29.52	40.52	37.97	21.27	26.07	26.35	21.53	38.81	30.54	29.55	44.74	31.55	32.96	25.88	23.87	20.49	29.84	25.91
CT5	34.29	15.96	25.47	36.75	14.94	30.00	38.51	26.60	37.36	23.41	13.09	26.07	23.54	20.57	38.81	24.48	30.15	40.79	29.13	27.08	13.82	18.21	18.85	28.23	13.70
AC	28.53	14.10	24.06	24.79	14.66	24.23	23.91	21.01	30.75	16.79	11.00	16.19	15.69	17.57	32.09	14.69	22.88	35.53	21.36	20.59	8.53	15.61	18.03	16.94	11.22

(b)

Appendix C: Results of Chapter 6

Table C.1 Top 1% retrieval rates in MDDR with W4 weighting scheme.

	JT	*Gle*	*For*	*DK*	*Kul*	*SS1*	*Ja*	*Fos*	*Phi*	*CT4*	*MR*	*Hel*	*Cze*
5HT1A	4.53	4.53	4.23	4.17	3.87	4.53	4.53	4.17	4.23	4.23	5.08	4.17	5.08
Thrombin	3.36	3.36	1.62	3.92	3.99	3.36	3.36	3.92	3.92	3.42	3.80	3.92	3.80
subP	3.81	3.81	1.28	3.13	2.33	3.81	3.81	3.13	3.09	4.05	3.81	3.13	3.81
Renin	43.45	43.45	1.42	49.87	41.77	43.45	43.45	50.49	47.65	52.26	43.98	49.87	43.98
PKC	7.51	7.51	2.98	8.06	7.51	7.51	7.51	8.06	8.06	8.17	7.17	8.06	7.17
HIVP	7.27	7.27	1.73	9.93	9.47	7.27	7.27	10.13	9.53	10.20	8.73	9.93	8.73
D2	4.81	4.81	4.18	4.68	4.18	4.81	4.81	4.56	4.56	4.30	4.56	4.68	4.56
COX	3.46	3.46	3.62	2.91	2.20	3.46	3.46	2.75	3.14	2.91	3.38	2.91	3.38
AT1	31.18	31.18	6.10	31.87	31.87	31.18	31.18	32.13	31.23	32.61	31.34	31.87	31.34
5HT3	8.31	8.31	8.64	8.11	7.45	8.31	8.31	7.98	8.18	6.98	10.57	8.11	10.57
5HT	5.29	5.29	5.01	5.29	5.01	5.29	5.29	5.43	5.29	5.57	5.57	5.29	5.57

Table C.2 Top 1% retrieval rates in MDDR with W5 weighting scheme.

	JT	*Gle*	*For*	*DK*	*Kul*	*SS1*	*Ja*	*Fos*	*Phi*	*CT4*	*MR*	*Hel*	*Cze*
5HT1A	4.96	4.96	4.59	4.78	4.78	4.96	4.96	4.90	4.90	4.96	5.08	4.78	5.08
Thrombin	3.80	3.80	3.24	3.92	4.42	3.80	3.80	4.42	3.80	4.86	3.99	3.92	3.99
subP	3.33	3.33	2.69	3.05	2.97	3.33	3.33	3.09	3.13	3.09	3.33	3.05	3.33
Renin	46.81	46.81	4.65	48.05	48.63	46.81	46.81	50.00	45.18	54.38	43.81	48.05	43.81
PKC	6.51	6.51	6.07	6.29	6.29	6.51	6.51	6.18	6.40	5.85	5.85	6.29	5.85
HIVP	8.80	8.80	4.93	8.93	9.40	8.80	8.80	9.47	8.60	10.40	8.93	8.93	8.93
D2	4.30	4.30	3.80	4.18	4.18	4.30	4.30	4.18	4.05	4.56	4.18	4.18	4.18
COX	2.91	2.91	4.40	2.75	2.75	2.91	2.91	2.59	2.75	2.52	3.07	2.75	3.07
AT1	30.33	30.33	16.38	30.59	30.28	30.33	30.33	30.97	29.80	32.50	30.75	30.59	30.75
5HT3	11.04	11.04	8.18	10.90	10.70	11.04	11.04	10.97	10.97	10.31	10.57	10.90	10.57
5HT	5.29	5.29	5.15	5.29	5.29	5.29	5.29	5.29	5.29	5.29	5.43	5.29	5.43

Table C.3 Top 1% retrieval rates in WOMBAT with W4 weighting scheme.

	JT	*Gle*	*For*	*DK*	*Kul*	*SS1*	*Ja*	*Fos*	*Phi*	*CT4*	*MR*	*Hel*	*Cze*
5HT1A	9.12	9.12	7.85	8.36	8.19	9.12	9.12	8.36	8.53	9.12	9.21	8.36	9.21
5HT3	10.68	10.68	11.59	10.45	9.32	10.68	10.68	10.45	10.68	10.45	14.32	10.45	14.32
AChE	4.27	4.27	3.08	4.67	4.67	4.27	4.27	4.77	4.27	5.27	5.47	4.67	5.47
ANG	30.46	30.46	4.35	30.18	24.65	30.46	30.46	30.46	29.21	32.25	32.18	30.18	32.18
COX	2.85	2.85	2.44	2.90	2.80	2.85	2.85	2.90	2.90	2.85	2.80	2.90	2.80
D2	8.08	8.08	4.01	8.24	6.76	8.08	8.08	8.19	8.02	8.79	8.24	8.24	8.24
Fxa	13.60	13.60	5.40	13.24	11.82	13.60	13.60	13.48	13.00	14.61	12.89	13.24	12.89
HIVP	4.83	4.83	1.46	4.88	3.46	4.83	4.83	5.10	4.65	5.81	5.63	4.88	5.63
MMP1	6.27	6.27	3.10	6.05	5.26	6.27	6.27	6.05	5.91	6.48	8.57	6.05	8.57
PDE4	5.87	5.87	4.28	6.12	5.54	5.87	5.87	6.12	6.12	6.38	6.96	6.12	6.96
PKC	23.24	23.24	9.15	23.24	13.03	23.24	23.24	23.24	26.76	20.42	20.42	23.24	20.42
RENIN	73.42	73.42	3.90	70.46	28.06	73.42	73.42	70.57	68.25	73.10	74.68	70.46	74.68
SUBP	13.89	13.89	3.49	13.08	12.37	13.89	13.89	13.08	13.08	14.25	13.53	13.08	13.53
THR	13.78	13.78	3.68	13.42	7.96	13.78	13.78	13.54	12.59	13.78	16.03	13.42	16.03

Table C.4 Top 1% retrieval rates in WOMBAT with W5 weighting scheme.

	JT	*Gle*	*For*	*DK*	*Kul*	*SS1*	*Ja*	*Fos*	*Phi*	*CT4*	*MR*	*Hel*	*Cze*
5HT1A	5.32	5.32	4.22	4.90	4.65	5.32	5.32	4.98	4.98	5.15	4.81	4.90	4.81
5HT3	12.27	12.27	7.27	13.86	13.86	12.27	12.27	14.55	12.05	15.00	13.18	13.86	13.18
AChE	12.52	12.52	3.48	12.33	11.93	12.52	12.52	12.72	11.53	13.72	13.12	12.33	13.12
ANG	31.15	31.15	12.50	30.46	29.70	31.15	31.15	31.63	29.14	35.22	31.63	30.46	31.63
COX	3.01	3.01	3.11	2.69	2.75	3.01	3.01	2.69	2.75	2.69	2.69	2.69	2.69
D2	8.63	8.63	5.55	8.74	8.46	8.63	8.63	9.01	8.41	9.62	7.91	8.74	7.91
Fxa	12.11	12.11	8.61	12.11	12.11	12.11	12.11	12.11	12.11	12.05	12.29	12.11	12.29
HIVP	4.65	4.65	2.62	4.88	5.90	4.65	4.65	5.67	4.79	6.12	5.19	4.88	5.19
MMP1	7.28	7.28	4.25	6.70	6.56	7.28	7.28	7.13	6.70	7.78	7.42	6.70	7.42
PDE4	7.72	7.72	4.87	8.22	8.05	7.72	7.72	8.56	8.05	9.56	8.31	8.22	8.31
PKC	10.56	10.56	14.79	11.97	12.68	10.56	10.56	11.27	13.03	9.15	11.97	11.97	11.97
RENIN	72.68	72.68	10.76	68.14	60.44	72.68	72.68	69.20	63.82	75.00	71.20	68.14	71.20
SUBP	13.53	13.53	9.77	13.17	12.81	13.53	13.53	13.26	12.90	13.62	13.17	13.17	13.17
THR	13.42	13.42	11.40	13.42	13.18	13.42	13.42	13.30	13.66	13.78	13.66	13.42	13.66

Table C.5 Top 1% retrieval rates in ChEMBL with W4 weighting scheme.

	JT	*Gle*	*For*	*DK*	*Kul*	*SS1*	*Ja*	*Fos*	*Phi*	*CT4*	*MR*	*Hel*	*Cze*
Target_no_10	29.12	29.12	8.82	28.53	27.65	29.12	29.12	28.53	27.94	29.41	30.59	28.53	30.59
Target_no_105	34.24	34.24	4.85	32.28	26.59	34.24	34.24	32.18	32.00	33.40	37.50	32.28	37.50
Target_no_112	33.05	33.05	26.72	33.05	32.76	33.05	33.05	33.05	33.05	33.33	35.63	33.05	35.63
Target_no_113	28.99	28.99	6.91	28.72	28.46	28.99	28.99	28.72	28.99	28.72	29.26	28.72	29.26
Target_no_115	45.65	45.65	26.09	45.34	43.17	45.65	45.65	45.34	45.34	45.03	46.89	45.34	46.89
Target_no_12	23.64	23.64	13.69	23.19	21.72	23.64	23.64	23.19	22.85	24.32	25.57	23.19	25.57
Target_no_120	51.92	51.92	28.08	50.38	49.23	51.92	51.92	50.38	49.62	50.77	58.08	50.38	58.08
Target_no_121	14.94	14.94	13.79	14.94	15.52	14.94	14.94	14.94	14.94	14.94	14.94	14.94	14.94
Target_no_129	26.18	26.18	19.12	26.18	26.18	26.18	26.18	26.18	26.18	26.18	26.18	26.18	26.18
Target_no_13	20.90	20.90	17.21	20.90	20.49	20.90	20.90	20.90	20.90	20.90	20.08	20.90	20.08
Target_no_14	34.27	34.27	22.18	34.27	33.87	34.27	34.27	34.27	33.47	35.08	33.87	34.27	33.87
Target_no_140	20.07	20.07	19.37	20.77	20.77	20.07	20.07	20.77	20.77	20.07	21.48	20.77	21.48
Target_no_142	40.50	40.50	17.77	38.51	34.96	40.50	40.50	38.43	38.10	40.33	40.66	38.51	40.66
Target_no_143	31.25	31.25	15.75	31.75	31.50	31.25	31.25	31.75	31.25	33.00	37.00	31.75	37.00
Target_no_146	42.19	42.19	32.19	39.06	35.31	42.19	42.19	39.06	39.69	38.75	39.69	39.06	39.69
Target_no_147	47.95	47.95	30.14	47.26	46.58	47.95	47.95	47.26	47.26	47.26	45.89	47.26	45.89
Target_no_148	11.20	11.20	7.52	10.32	9.60	11.20	11.20	10.32	10.24	11.44	16.00	10.32	16.00
Target_no_152	24.88	24.88	12.19	24.38	24.13	24.88	24.88	24.38	24.38	24.88	24.88	24.38	24.88
Target_no_16	32.51	32.51	10.73	32.51	27.56	32.51	32.51	32.51	31.68	32.67	32.01	32.51	32.01
Target_no_163	23.60	23.60	22.67	23.60	23.60	23.60	23.60	23.60	23.60	23.60	23.91	23.60	23.91
Target_no_168	13.00	13.00	11.00	13.00	12.00	13.00	13.00	13.00	12.00	13.00	15.00	13.00	15.00
Target_no_171	20.00	20.00	18.50	19.50	19.50	20.00	20.00	19.50	19.50	20.00	20.00	19.50	20.00
Target_no_181	28.53	28.53	24.71	28.53	25.88	28.53	28.53	28.53	28.53	27.94	27.35	28.53	27.35
Target_no_186	37.85	37.85	28.82	37.85	37.50	37.85	37.85	37.85	37.85	36.46	38.19	37.85	38.19
Target_no_195	17.14	17.14	9.05	14.76	12.86	17.14	17.14	14.76	14.76	16.67	21.90	14.76	21.90
Target_no_196	19.34	19.34	14.23	20.44	18.61	19.34	19.34	20.44	19.71	21.90	15.69	20.44	15.69
Target_no_21	41.35	41.35	31.01	41.59	39.66	41.35	41.35	41.11	41.35	39.42	47.36	41.59	47.36
Target_no_210	32.04	32.04	34.47	32.04	32.52	32.04	32.04	32.04	32.04	32.04	32.04	32.04	32.04

Target_no_211	20.62	20.62	10.40	19.71	17.34	20.62	20.62	19.89	19.16	19.16	24.27	19.71	24.27
Target_no_213	27.44	27.44	20.30	27.82	26.69	27.44	27.44	28.20	28.20	26.69	29.32	27.82	29.32
Target_no_220	34.58	34.58	18.75	34.58	33.75	34.58	34.58	34.58	34.58	35.00	35.83	34.58	35.83
Target_no_230	43.67	43.67	33.13	40.66	33.43	43.67	43.67	40.66	41.57	40.06	52.11	40.66	52.11
Target_no_234	32.13	32.13	20.03	32.28	30.84	32.13	32.13	32.28	32.28	30.55	34.44	32.28	34.44
Target_no_238	15.43	15.43	13.83	15.16	14.10	15.43	15.43	15.16	15.16	14.63	15.96	15.16	15.96
Target_no_241	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00
Target_no_250	40.60	40.60	26.50	40.60	40.60	40.60	40.60	40.60	40.60	40.60	40.60	40.60	40.60
Target_no_35	24.37	24.37	24.79	21.85	18.49	24.37	24.37	21.85	22.27	26.89	30.25	21.85	30.25
Target_no_4	46.05	46.05	33.88	45.72	44.08	46.05	46.05	46.05	45.72	46.38	47.37	45.72	47.37
Target_no_42	20.10	20.10	7.32	18.97	17.63	20.10	20.10	18.97	18.97	20.10	21.03	18.97	21.03
Target_no_44	28.94	28.94	24.70	29.09	28.79	28.94	28.94	29.09	29.24	28.33	29.39	29.09	29.39
Target_no_52	22.67	22.67	20.00	23.00	24.67	22.67	22.67	23.00	23.00	22.00	26.67	23.00	26.67
Target_no_54	21.63	21.63	12.30	21.43	21.83	21.63	21.63	21.43	21.23	21.83	16.47	21.43	16.47
Target_no_57	35.44	35.44	16.49	34.92	32.99	35.44	35.44	34.92	34.79	35.31	37.76	34.92	37.76
Target_no_59	27.61	27.61	24.63	28.36	33.58	27.61	27.61	28.36	28.36	26.12	29.10	28.36	29.10
Target_no_8	24.27	24.27	19.90	26.70	25.73	24.27	24.27	26.70	26.70	25.24	28.16	26.70	28.16
Target_no_81	20.71	20.71	16.21	20.71	20.57	20.71	20.71	20.71	20.57	20.44	22.62	20.71	22.62
Target_no_86	19.56	19.56	12.65	18.97	17.21	19.56	19.56	18.85	18.97	19.32	20.84	18.97	20.84
Target_no_9	32.66	32.66	8.82	32.25	32.25	32.66	32.66	32.56	32.05	33.06	32.76	32.25	32.76
Target_no_95	32.38	32.38	14.18	30.23	27.51	32.38	32.38	30.09	30.95	30.95	30.23	30.23	30.23
Target_no_98	22.73	22.73	7.45	21.45	18.82	22.73	22.73	21.45	21.27	22.73	21.73	21.45	21.73

Table C.6 Top 1% retrieval rates in ChEMBL with W5 weighting scheme.

	JT	*Gle*	*For*	*DK*	*Kul*	*SS1*	*Ja*	*Fos*	*Phi*	*CT4*	*MR*	*Hel*	*Cze*
Target_no_10	30.59	30.59	10.59	30.00	29.71	30.59	30.59	30.29	29.71	30.59	30.59	30.00	30.59
Target_no_105	36.66	36.66	12.78	35.63	34.51	36.66	36.66	35.54	34.42	36.29	37.50	35.63	37.50
Target_no_112	35.63	35.63	31.03	35.92	35.92	35.63	35.63	36.21	35.63	35.92	36.21	35.92	36.21
Target_no_113	27.66	27.66	24.73	27.39	28.19	27.66	27.66	28.19	27.39	28.46	29.26	27.39	29.26
Target_no_115	46.27	46.27	32.61	46.27	45.96	46.27	46.27	46.58	45.65	47.20	45.96	46.27	45.96
Target_no_12	23.76	23.76	17.31	22.96	22.17	23.76	23.76	23.64	22.29	24.77	24.89	22.96	24.89
Target_no_120	55.00	55.00	36.92	55.00	55.00	55.00	55.00	55.38	55.00	56.15	56.54	55.00	56.54
Target_no_121	14.94	14.94	14.94	14.94	14.94	14.94	14.94	14.94	14.94	14.94	14.94	14.94	14.94
Target_no_129	25.88	25.88	23.24	25.88	25.88	25.88	25.88	25.88	25.88	25.59	25.88	25.88	25.88
Target_no_13	20.90	20.90	19.26	20.49	20.08	20.90	20.90	20.49	20.49	20.49	20.08	20.49	20.08
Target_no_14	33.47	33.47	29.03	33.47	33.47	33.47	33.47	33.87	33.47	33.87	33.87	33.47	33.87
Target_no_140	21.13	21.13	19.72	21.48	21.48	21.13	21.13	21.13	21.48	21.13	21.48	21.48	21.48
Target_no_142	41.65	41.65	24.63	40.66	39.09	41.65	41.65	41.65	39.92	42.73	40.83	40.66	40.83
Target_no_143	35.50	35.50	25.50	35.75	35.75	35.50	35.50	36.00	35.00	36.25	36.50	35.75	36.50
Target_no_146	38.75	38.75	39.69	37.50	36.56	38.75	38.75	37.50	37.50	35.00	37.81	37.50	37.81
Target_no_147	48.63	48.63	36.99	49.32	47.95	48.63	48.63	48.63	48.63	47.95	47.95	49.32	47.95
Target_no_148	15.84	15.84	8.80	15.68	15.28	15.84	15.84	16.00	15.04	16.72	16.96	15.68	16.96
Target_no_152	23.88	23.88	17.16	23.63	23.63	23.88	23.88	23.88	23.38	24.13	24.38	23.63	24.38
Target_no_16	27.56	27.56	12.21	26.40	23.10	27.56	27.56	26.90	24.09	28.88	30.36	26.40	30.36
Target_no_163	23.60	23.60	22.98	23.60	23.60	23.60	23.60	23.91	23.60	23.91	23.60	23.60	23.60
Target_no_168	15.00	15.00	13.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
Target_no_171	20.00	20.00	19.50	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00
Target_no_181	26.76	26.76	25.00	27.06	27.06	26.76	26.76	27.06	26.76	27.06	26.76	27.06	26.76
Target_no_186	36.46	36.46	35.76	36.81	36.81	36.46	36.46	37.15	36.46	36.11	37.15	36.81	37.15
Target_no_195	24.76	24.76	17.62	24.29	25.24	24.76	24.76	24.76	24.76	25.24	22.86	24.29	22.86
Target_no_196	16.06	16.06	14.96	16.06	16.06	16.06	16.06	16.06	16.06	16.06	16.06	16.06	16.06
Target_no_21	47.12	47.12	40.63	46.88	46.88	47.12	47.12	47.12	47.12	47.60	48.32	46.88	48.32
Target_no_210	32.04	32.04	33.01	32.04	32.52	32.04	32.04	32.04	32.52	32.04	32.04	32.04	32.04

Target_no_211	21.17	21.17	14.05	21.35	21.17	21.17	21.17	22.08	20.26	22.99	20.80	21.35	20.80
Target_no_213	28.95	28.95	25.56	28.57	28.20	28.95	28.95	28.95	28.57	29.70	28.95	28.57	28.95
Target_no_220	35.83	35.83	25.42	35.83	35.83	35.83	35.83	35.83	35.83	35.83	35.83	35.83	35.83
Target_no_230	53.01	53.01	33.13	51.51	50.60	53.01	53.01	52.11	51.81	52.11	52.71	51.51	52.71
Target_no_234	34.44	34.44	27.67	33.72	33.00	34.44	34.44	33.57	33.72	33.57	34.01	33.72	34.01
Target_no_238	15.96	15.96	15.43	15.96	15.96	15.96	15.96	15.96	15.96	15.96	15.96	15.96	15.96
Target_no_241	25.00	25.00	25.47	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00
Target_no_250	40.60	40.60	31.62	40.60	40.60	40.60	40.60	40.60	40.60	40.60	40.60	40.60	40.60
Target_no_35	28.57	28.57	30.25	28.99	30.25	28.57	28.57	30.67	29.41	31.51	30.67	28.99	30.67
Target_no_4	46.71	46.71	35.53	46.05	45.07	46.71	46.71	46.38	45.72	46.71	46.05	46.05	46.05
Target_no_42	21.96	21.96	12.58	21.75	21.24	21.96	21.96	21.96	21.24	21.96	21.96	21.75	21.96
Target_no_44	29.39	29.39	27.88	29.39	29.24	29.39	29.39	29.39	29.39	29.39	29.70	29.39	29.70
Target_no_52	26.00	26.00	25.00	26.33	27.33	26.00	26.00	26.00	27.67	26.00	26.00	26.33	26.00
Target_no_54	16.27	16.27	16.07	16.07	15.87	16.27	16.27	15.87	16.47	15.08	16.07	16.07	16.07
Target_no_57	34.54	34.54	23.45	34.54	34.15	34.54	34.54	34.28	34.28	34.54	34.66	34.54	34.66
Target_no_59	34.33	34.33	29.10	34.33	33.58	34.33	34.33	34.33	33.58	34.33	34.33	34.33	34.33
Target_no_8	29.61	29.61	24.76	30.10	30.10	29.61	29.61	30.10	29.61	29.61	30.10	30.10	30.10
Target_no_81	21.39	21.39	19.75	21.12	21.12	21.39	21.39	21.53	20.98	22.07	22.34	21.12	22.34
Target_no_86	19.67	19.67	16.28	19.44	18.85	19.67	19.67	19.56	19.09	19.79	20.96	19.44	20.96
Target_no_9	31.85	31.85	23.12	31.34	30.02	31.85	31.85	31.44	30.22	31.64	32.15	31.34	32.15
Target_no_95	29.23	29.23	24.79	28.80	27.79	29.23	29.23	28.37	28.65	27.36	28.08	28.80	28.08
Target_no_98	21.91	21.91	11.45	21.82	21.82	21.91	21.91	22.18	21.00	23.09	21.36	21.82	21.36