SHAPE-BASED TWO DIMENSIONAL DESCRIPTOR FOR SEARCHING MOLECULAR DATABASE

HENTABLI HAMZA

UNIVERSITI TEKNOLOGI MALAYSIA
“I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of the degree of Master of Science (Computer Science)”

Signature : 
Name : Naomie Salim 
Date : May 26, 2014
SHAPE-BASED TWO DIMENSIONAL DESCRIPTOR FOR SEARCHING MOLECULAR DATABASE

HENTABLI HAMZA

A thesis submitted in fulfilment of the requirements for the award of the degree of Master of Science (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

MAY 2014
I declare that this thesis entitled “Shape-Based Two Dimensional Descriptor for Searching Molecular Database” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : 
Name : Hentabli Hamza
Date : May 26, 2014
This research study is dedicated to my wife, to my mother, and in the memory of my father

"If we knew what it was we were doing, it would not be called research, would it?"
Albert Einstein

Research is to see what everybody else has seen, and to think what nobody else has thought
Albert Szent-Gyorgyi
ACKNOWLEDGEMENT

Alhamdulillah, all praises to Allah S.W.T., The Most Greatest and The Most Merciful for His guidance and blessing, because without him I cant finished this research. I also wish to express my gratitude to my thesis supervisor, Prof. Dr. Naomie Salim for her enthusiastic guidance, valuable help, encouragement and patience for all the aspect of the thesis progress. Her numerous comments, criticisms and suggestion during the preparation of this project are greatly appreciated. Also, for her patience on problems that occurred during the process of completing this thesis.

Furthermore, special thanks go to Dr. Faisal Saeed, for many days of scientific discussions, this work would not have been possible without his support. Many thanks goes to the people who supported me in write this thesis, especially amer tawfiq, Dr.Ali Ahmed, Dr.Djamel. I also would like to thank all my friends that give me support and help during the writing of the thesis. Their support and help always give me the motivation and energy I need to finish the thesis. My appreciation also extended to all academic and non-academic member of the Faculty of Computing for their warm heart cooperation during my stay in Universiti Teknologi Malaysia.

My heartfelt acknowledgement are expressed to my family, especially My mother and Wife, without their guidance, support, encouragement and advises I may never have overcome this long journey in my studies. When I felt down, their love always give me the strength to face the challenges of the research. I would also like to thank people that directly or indirectly help me in finishing the thesis. Thank you very much.
ABSTRACT

Biological functions of compounds can be predicted from similarity of their chemical structures to discover new compounds for drug development. Molecular similarity can also be used to infer unknown functions and side effects of existing drugs. A multitude of molecular similarity methods based on different molecular representations have been used to perform virtual screenings. The molecules are transformed into descriptors to create a chemical database which allows mathematical manipulation and searching of the chemical information contained in the molecules. In this research, a new Shape based Descriptor of Molecule (SBDM) was developed based on the 2-dimensional shape of a chemical compound. The outline shape of a molecule is split into parts that are related in graph connectivity. The first atom in the molecule is determined using the Morgan algorithm. The molecular features, such as atom name, bond type, angle and rings are represented using specific symbols based on some specification rules. Subsequent atoms are scanned in a clockwise direction with respect to the first atom. The scan is repeated until the first atom is reached again. Two similarity measures were used to evaluate the performance of the molecular descriptors, which are the Basic Local Alignment Search Tool (BLAST) and the Tanimoto coefficient. The performance of the SBDM is compared with six standard molecular descriptors. Simulation of virtual screening experiments with the MDL Drug Data Report database show the superiority of the shape-based descriptor, with 19.32 % and 34.13 % in terms of average recall rates for the top of 1 % and 5 % retrieved molecules, respectively, compared to the six standard descriptors mentioned earlier.
**ABSTRAK**

# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECLARATION</td>
<td>ii</td>
<td></td>
</tr>
<tr>
<td>DEDICATION</td>
<td>iii</td>
<td></td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>iv</td>
<td></td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>v</td>
<td></td>
</tr>
<tr>
<td>ABSTRAK</td>
<td>vi</td>
<td></td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>vii</td>
<td></td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xi</td>
<td></td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xiii</td>
<td></td>
</tr>
</tbody>
</table>

1 INTRODUCTION | 1 |
| 1.1 Introduction | 1 |
| 1.2 Problem Background | 2 |
| 1.3 Problem Statement | 4 |
| 1.4 Research Question | 5 |
| 1.5 Research Aim | 5 |
| 1.6 Objectives | 6 |
| 1.7 Research Scope | 6 |
| 1.8 Significance and justification of study | 7 |
| 1.9 Organization of the report | 7 |
| 1.10 Summary | 7 |

2 LITERATURE REVIEW | 8 |
| 2.1 Introduction | 8 |
| 2.2 Chemical Database | 9 |
| 2.3 Chemical Compounds | 9 |
| 2.4 Molecular Similarity and Molecular searching | 10 |
| 2.5 Representation of Chemical Structures | 11 |
| 2.5.1 Fingerprint Descriptors | 13 |
2.5.2 String Representation and Canonical SMILES 14

2.5.3 Graph Representation 21

2.6 Molecular Representation 22

2.6.1 1D Descriptors 23

2.6.2 2D Descriptors 24

2.6.3 3D descriptors 26

2.7 Searching in Chemical Database 27

2.7.1 Structure Searching 28

2.7.2 Substructure Searching 28

2.7.3 Similarity Searching 29

2.7.4 Molecular Descriptors for Similarity Searching 31

2.7.5 Similarity Coefficients 32

2.7.5.1 Distance Coefficients 32

2.7.5.2 Association Coefficients 34

2.7.5.3 Correlation Coefficients 36

2.7.5.4 Probabilistic Coefficients 37

2.8 Shape Representation and Description Techniques 37

2.8.1 Classifications of Shape Representation Techniques 38

2.8.2 Chain Code Representation 39

2.8.3 Syntactic Analysis 41

2.9 Discussion 42

2.10 Summary 45

3 RESEARCH METHODOLOGY 46

3.1 Introduction 46

3.2 Research Design 46

3.3 General Research Framework 47

3.4 Molecular Dataset Preparation 49

3.5 Generating Shape-based Descriptors 54

3.5.1 Language for Writing Descriptors of Outline Shape of Molecules 54

3.5.2 2D Fingerprint Descriptors of Outline Shape of Molecules 55

3.5.3 Canonical Representation for Descriptors of Outline Shape of Molecules 56

3.6 Generating Other Molecular Descriptors 57
3.6.1 Extended Connectivity Fingerprints (ECFP) 59
3.6.1.1 Properties of ECFP 59
3.6.1.2 Applications 59
3.6.1.3 Representation and Generation 60
3.6.1.4 The Generation Process 61

3.7 Evaluate the Proposed Descriptors 63
3.7.1 Structure-based approaches 65
3.7.2 Ligand-based approaches 66

3.8 Evaluation Measures of Similarity Performance 69
3.8.1 Tanimoto Coefficient 69
3.8.2 Recall Calculation 70
3.8.3 Kendall rank correlation coefficient τ 71
3.8.4 Basic Local Alignment Search Tool 72

3.9 Summary 75

4 LANGUAGE FOR WRITING DESCRIPTORS OF OUTLINE SHAPE OF MOLECULES 76
4.1 Introduction 76
4.2 Materials and Methods 77
4.2.1 Specification Rules 78
4.3 Experimental Design 80
4.4 Validation of the Implemented System 81
4.5 Results and Discussion 85
4.6 Summary 92

5 2D FINGERPRINT DESCRIPTORS OF OUTLINE SHAPE OF MOLECULES 93
5.1 Introduction 93
5.2 Materials and Methods 95
5.3 Experimental Design 99
5.4 Results and Discussion 105
5.5 Summary 113

6 CANONICAL REPRESENTATIONS FOR DESCRIPTORS OF OUTLINE SHAPE OF MOLECULES 114
6.1 Introduction 115
6.2 Materials and Methods 116
6.3 Experimental Design 124
6.4 Validation of the implemented System 125
6.5 Results and Discussion 127
6.6 Summary 136

7 CONCLUSION AND FUTURE WORK 137
7.1 Conclusion 137
7.2 Research Contributions 138
7.3 Future Works 139

REFERENCES 140
## LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Example of SMILES structure of different molecules illustrating the design rules of SMILES-sturcture.</td>
<td>16</td>
</tr>
<tr>
<td>2.2</td>
<td>Examples of descriptors classified according to dimensionality, demonstrated on the saccharin molecule.</td>
<td>23</td>
</tr>
<tr>
<td>2.3</td>
<td>Comparison of the three type of searching method in chemical database</td>
<td>31</td>
</tr>
<tr>
<td>2.4</td>
<td>Examples of distance coefficients</td>
<td>34</td>
</tr>
<tr>
<td>2.5</td>
<td>Examples of association coefficients</td>
<td>35</td>
</tr>
<tr>
<td>2.6</td>
<td>Examples of correlation coefficients</td>
<td>37</td>
</tr>
<tr>
<td>3.1</td>
<td>MDDR Activity Classes for DS1 Data Set</td>
<td>53</td>
</tr>
<tr>
<td>3.2</td>
<td>MDDR Activity Classes for DS2 Data Set</td>
<td>53</td>
</tr>
<tr>
<td>3.3</td>
<td>MDDR Activity Classes for DS3 Data Set</td>
<td>54</td>
</tr>
<tr>
<td>3.4</td>
<td>Interpretation of the correlation coefficients size for positive values.</td>
<td>72</td>
</tr>
<tr>
<td>4.1</td>
<td>Validation of the implemented System</td>
<td>82</td>
</tr>
<tr>
<td>4.2</td>
<td>Retrieval results of top 1% for data set DS1</td>
<td>86</td>
</tr>
<tr>
<td>4.3</td>
<td>Retrieval results of top 5% for data set DS1</td>
<td>87</td>
</tr>
<tr>
<td>4.4</td>
<td>Retrieval results of top 1% for data set DS2</td>
<td>87</td>
</tr>
<tr>
<td>4.5</td>
<td>Retrieval results of top 5% for data set DS2</td>
<td>88</td>
</tr>
<tr>
<td>4.6</td>
<td>Retrieval results of top 1% for data set DS3</td>
<td>88</td>
</tr>
<tr>
<td>4.7</td>
<td>Retrieval results of top 5% for data set DS3</td>
<td>89</td>
</tr>
<tr>
<td>4.8</td>
<td>Rankings of various types of descriptors Based on Kendall W Test Results</td>
<td>90</td>
</tr>
<tr>
<td>4.9</td>
<td>Numbers of Shaded Cells for Mean Recall of Actives molecules</td>
<td>91</td>
</tr>
<tr>
<td>5.1</td>
<td>List of atoms name that contain two characters and their new labeling</td>
<td>96</td>
</tr>
<tr>
<td>5.2</td>
<td>Sample of Lingo-DOSM descriptor</td>
<td>101</td>
</tr>
<tr>
<td>5.3</td>
<td>Fragmentation of LWDOSM Molecule A</td>
<td>102</td>
</tr>
</tbody>
</table>
5.4 Fragmentation of LWDOSM Molecule B

5.5 Intermolecular Similarity Calculation for the molecule A and B

5.6 Retrieval results of top 1% for data set DS1

5.7 Retrieval results of top 5% for data set DS1

5.8 Retrieval results of top 1% and 5% for data set DS1 comparing with LWDOSM

5.9 Retrieval results of top 1% for data set DS2

5.10 Retrieval results of top 5% for data set DS2

5.11 Retrieval results of top 1% and 5% for data set DS2 comparing with LWDOSM

5.12 Retrieval results of top 1% for data set DS3

5.13 Retrieval results of top 5% for data set DS3

5.14 Retrieval results of top 1% and 5% for data set DS3 comparing with LWDOSM

5.15 Rankings of various types of descriptors Based on Kendall W Test Results: Top 1% and 5%

5.16 Numbers of Shaded Cells for Mean Recall of Actives Using Different Descriptors: DS1-DS3 Top 1% and 5%

6.1 Validation of the implemented System

6.2 Retrieval results of top 1% for data set DS1

6.3 Retrieval results of top 5% for data set DS1

6.4 Retrieval results of top 1% and 5% for data set DS1 comparing with LWDOSM and Lingo-DOSM

6.5 Retrieval results of top 1% for data set DS2

6.6 Retrieval results of top 5% for data set DS2

6.7 Retrieval results of top 1% and 5% for data set DS2 comparing with LWDOSM and Lingo-DOSM

6.8 Retrieval results of top 1% for data set DS3

6.9 Retrieval results of top 5% for data set DS3

6.10 Retrieval results of top 1% and 5% for data set DS3 comparing DS2 comparing with LWDOSM and Lingo-DOSM

6.11 Rankings of various types of descriptors based on Kendall W Test Results

6.12 Numbers of Shaded Cells for Mean Recall of Actives Using Different Descriptors: DS1-DS3 Top 1% and 5%
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Three main step for molecular descriptor</td>
<td>11</td>
</tr>
<tr>
<td>2.2</td>
<td>The chemical structure of aspirin</td>
<td>12</td>
</tr>
<tr>
<td>2.3</td>
<td>Fragment code representation for three molecules</td>
<td>13</td>
</tr>
<tr>
<td>2.4</td>
<td>Examples of the SMILES Strings for a variety of molecules</td>
<td>14</td>
</tr>
<tr>
<td>2.5</td>
<td>Examples of the Wisswesser Line Notation Strings of molecule</td>
<td>17</td>
</tr>
<tr>
<td>2.6</td>
<td>Examples of the InChI Strings descriptor of molecule</td>
<td>18</td>
</tr>
<tr>
<td>2.7</td>
<td>Examples of Canonical versus non-Canonical SMILES</td>
<td>19</td>
</tr>
<tr>
<td>2.8</td>
<td>Generation of SMILES: A)Break cycles, B)write as branches off a main backbone C)spanning tree D) Final Canonical SMILES</td>
<td>20</td>
</tr>
<tr>
<td>2.9</td>
<td>The connection table for aspirin in the MDL format (hydrogen-suppressed form), along with its 2D structure</td>
<td>21</td>
</tr>
<tr>
<td>2.10</td>
<td>Representations of a molecular structure in 1D, 2D, and 3D</td>
<td>27</td>
</tr>
<tr>
<td>2.11</td>
<td>Classification of shape representation and description techniques</td>
<td>39</td>
</tr>
<tr>
<td>2.12</td>
<td>(a) The directions of the eight connected chain code (K=8). (b) A sample object, a square, and (c) chain code presentation of the square</td>
<td>40</td>
</tr>
<tr>
<td>2.13</td>
<td>Structural description of chromosome shape</td>
<td>41</td>
</tr>
<tr>
<td>3.1</td>
<td>The general research operational framework</td>
<td>48</td>
</tr>
<tr>
<td>3.2</td>
<td>Examples of low diversity molecules in MDDR dataset</td>
<td>50</td>
</tr>
<tr>
<td>3.3</td>
<td>Examples of high diversity molecules in MDDR dataset</td>
<td>51</td>
</tr>
<tr>
<td>3.4</td>
<td>The mean pairwise similarity (MPS) across each set of active molecules</td>
<td>52</td>
</tr>
<tr>
<td>3.5</td>
<td>Framework for generating of Language for Writing Descriptors of Outline Shape of Molecules</td>
<td>55</td>
</tr>
</tbody>
</table>
3.6 2D Fingerprint Descriptors of Outline Shape of Molecules. 56
3.7 Canonical Representation for Descriptors of Outline Shape of Molecules. 57
3.8 Generating ALOGP using Pipeline Pilot software. 58
3.9 Molecular Descriptor generation. 58
3.10 Illustration of the effect of iterative updating for a selected atom in a sample molecule 62
3.11 ECFP generation process 62
3.12 Generation of the fixed-length bit string 63
3.13 Schematic representation of virtual screening approaches within early-phase drug discovery. 64
3.14 An example workflow for 3D shape-based flexible alignment: 1) 2D input structures; 2) 3D conformer is generated and the shape is colored by atomic types; 3) the volume intersection, which maximized during the alignment, is shown along with the resulting pose 65
3.15 Overview of the Ligand Based Virtual Screening (LBVS) process. (A) Query and database compounds are encoded by pharmacophore patterns or descriptor vectors. (B) Chemical similarity between query and database compounds is quantified. (C) The library compounds are ranked according to their similarity with the query. 68
3.16 Tanimoto in similarity searching 70
3.17 Generating word seeds. 73
3.18 Examples of how alignment seeds work. Larger word sizes usually mean that there are fewer regions to evaluate (e.g., part a versus part b). Once an alignment is seeded, BLAST extends the alignment according to a threshold (T), which is set by the user. When performing a BLAST query, the computer extends words with a neighborhood score greater than T (part c). 74
4.1 The external and internal visiting movement in 2D graph 78
4.2 The representation of internal regions (cyclic structures) 79
4.3 The Angle Direction 79
4.4 The disconnect parts representation 80
5.1 LWDOSM calculation process 95
5.2 The method to establish the q-letter query word list 99
5.3 LINGO-DOSM fragmentation 99
6.1 The adjacency matrix of the Graph (G) to represent the aspirin molecule 117
6.2 Example of the Morgan algorithm 119
6.3 Example of canonical enumerations for atoms 121
6.4 Graphical representation of Tree T extracted from aspirin Molecule 122
6.5 Process of generating the shape descriptor (SBMD) 123
CHAPTER 1

INTRODUCTION

1.1 Introduction

Chemoinformatics is the collection, representation and organization of chemical data to create chemical information in which it can be applied to create chemical knowledge. In pharmaceutical and agrochemical industry, Chemoinformatics has been used for identification of novel compounds with useful and commercially valuable biological properties. The drug discovery process is very complex and it is a multi-disciplinary task with many stages to be performed in a long time. However, the molecule that has the potential to become drugs may cause unexpected long-term side effects and the drug discovery process can take about 12 years and the costs may be up to $350 million per drug [1].

The high costs of bringing a drug into the market have increased the pressure on the pharmaceutical industries. Therefore, attention is given to research and development in order to develop faster and more effective way to produce chemical compounds that can react to the disease and also produce antibodies towards the disease. This has encouraged the study of chemoinformatics and drug discovery as one of the new areas in Malaysia’s research and development (R&D) [1].

The primary task of the similarity calculation and molecular searching in chemical database is to find and quantify the relationship between structures of compounds with the hope that the quantity represents similarity in physical, chemical, or biological properties. Thus, an appropriate description (molecular descriptor)
of chemical structure for the property to be represented is an essential task. Molecular descriptors are numerical values (or vectors of numbers) that characterize properties of compounds. They can be generated from a machine-readable structure representation like a 2D connection table or a set of experimental or calculated 3D coordinates. Molecular descriptors should be effective (can differentiate between different compounds) and efficient (fast to calculate) by playing a vital role in similarity calculations especially when a large database is used. However, there is a conflict between these two requirements since the most effective descriptors tend to be the least efficient to calculate, and vice versa. Molecular descriptors can be classified into 1D, 2D and 3D descriptors. Molecule representations based on the real shape of the molecules are important because it represents the physical interaction between molecules, especially protein-ligand bindings, instead most of the descriptors are developed based on the molecule features or topological indices.

1.2 Problem Background

In organic chemistry research the chemists synthesize novel compounds that exhibit the desired properties and perform better than existing compounds. The commercial objective is to sell these new and superior compounds directly or in new products containing them. During the process from the initial creation of a new compound until it is commercialized, a lot of data related to it is generated such as physiochemical properties, analytical and toxicological data. All this data needs to be stored and scientists must be able to search it by chemical structure [2].

The chemist performs a chemical reaction and enters the relevant chemical structures, the exact procedure and outcome into this electronic lab notebook (ELN) so that other chemists can find it and repeat the reaction if necessary. After this new compound is synthesized, a new entry for it is generated in the chemical registration system. Unfortunately there is not much scientific literature concerning how and what such registration systems do [2]. The main purpose is to assign each compound a unique in-house reference number that then can be used to cross-reference it. There are many commercial software solutions for ELNs and chemical registrations such as those from PerkinElmer (CambridgeSoft), Accelrys, ChemAxon or Dotmatics to name a few. However these solutions may be costly especially if they also require a commercial relational database management system (RDBMS). An exception is
ChemAxon Compound Registration [2] [3] [4] which runs on MySQL. Another issue is that they are closed-source and if such systems are used in the context of scientific experiments, the experiment is not fully reproducible by other scientists unless they also have access to a license for the same system. Because these systems need to be highly configurable, they are also very complex. Configuring and administrating such a system requires a lot of very specific expertise. Consequentially, it can be advantageous to build your own solution especially for organizations that already have an in-house software development department [5] [6] [7]. A custom solution does not have to be highly configurable as it can be tailored to your needs. This includes administrative interfaces for user management and specific tasks like informing all users about new features. Such tools can greatly reduce administrative overhead [2].

The foundation of a chemical information system is the ability to represent molecules in a computer and to compare a molecule’s structure with another. Molecular comparison has been used in the early chemical information systems, e.g. structure and substructure searching [8] [2]. Structure searching involves searching a chemical database for a particular query structure with aims to retrieve all molecules with an exact match to the query structure whereas substructure searching retrieves all molecules that contain the query structure as a subgraph. The equivalence (similarity) between two structures can be achieved by using a graph and subgraph isomorphism algorithms. Isomorphism algorithms are time consuming because it is a combinatorial problem. Various isomorphism algorithms have been developed for efficient performance but they are still too slow for large chemical databases [10] [11] [12] [13]. However, structure and substructure searching were later complemented by another searching mechanism called similarity searching.

Similarity searching methods may be the simplest tools for ligand based virtual screening. The basic idea in similarity searching is the similar property principle, which states that structurally similar molecules will exhibit similar physicochemical and biological properties [14]. Over the years, many ways of measuring the structural similarity of molecules have been introduced [15] [16]. 2D similarity methods can be divided into two classes, the first class is the graph-based similarity methods and the second class is the fingerprint-based similarity methods. The graph-based similarity methods directly compare the molecular structures with each other and identify the similar (or common) substructures. These methods relate parts of one molecule to parts of the other molecule, they generates a mapping or alignment between molecules. Maximum common sub-graph method (MCS) is an example of the graph-based
Another example of the graph-based similarity methods is the feature trees. The feature trees were introduced by Rarey and Dixon [17], which are the most abstract way of representing a molecule by means of a graph. A feature tree represents hydrophobic fragments and functional groups of the molecule and the way these groups are linked together. Each node in the tree is labelled with a set of features representing chemical properties of the part of the molecule corresponding to the node. The comparison of feature trees is based on matching subtrees of two feature trees onto each other. Feature trees allow similarity searching to be performed against large database, when combined with a fast mapping algorithms [18]. However, the most common similarity approaches use molecules characterized by fingerprints that encode the presence of fragments features in a molecule. The similarity between two molecules is then computed using the number of sub-structural fragments that is common to a pair of structures and a simple association coefficient [19].

1.3 Problem Statement

The shape similarity between two molecules can be determined by comparing the shapes of those molecules; find the overlap volume between them and then use similarity measure (e.g. Tanimoto) to calculate the similarity between the molecules. However, most of the works in shape-based similarity approaches depended on the 3D molecular shape [20, 21]. Recently, shape-based similarity approaches have been used more. The shape comparison program Rapid Overlay of Chemical Structures (ROCS) [22] is used to perceive similarity between molecules based on their 3D shape. The objective of this approach is to find molecules with similar bioactivity to a target molecule but with different chemotypes, i.e., scaffold hopping. However, a disadvantage of 3D similarity methods is that the conformational properties of the molecules should be considered and therefore these methods are more computationally intensive than methods based on 2D structure representation. The complexity increases considerably if conformational flexibility is taken into account. There are many 2D structure representations in a numerical form integer or real. The simplest 2D descriptors are based on simple counts of features such as hydrogen donors, hydrogen bond acceptors, ring systems (such as aromatic rings) and rotatable bonds, whereas the complex 2D descriptors are computed from complex mathematical equations such as 2D fingerprints and topological indices. Topological indices are integer or real value numbers (single value) that represent the constitution of the molecules and can be calculated from the 2D graph representation of molecules and may contain
additional property information about the molecule [23, 24, 25]. They characterize molecular structures according to their size, degree of branching and overall shape where the structural diagram of molecules is considered as a mathematical graph, but not the contour of molecule shape. In this thesis, we introduced a new shape-based molecular descriptor that have been inspired by research in information retrieval on the use of contour based shape descriptor for image retrieval systems [26, 27, 28]. Shape-based molecular descriptor is a new method to obtain rough description of 2D molecular structure from its 2D outline shape of its 2D diagram. shape-based molecular descriptor is a textual descriptor that allows rigorous structure specification by the use of a very small and natural grammar.

1.4 Research Question

This study will focus on the following questions:

1. How can we develop new descriptor based on 2D shape of molecule to be used in similarity searching?

2. How can this descriptor be modeled to represent the molecule in fingerprint format?

3. How can graph theory be applied to canonical representation of molecule for similarity searching?

1.5 Research Aim

Molecular descriptors play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, and health researches. The molecules are transformed into numbers, allowing some mathematical treatment of the chemical information contained in the molecule. The aim of this research is to propose and develop a new descriptor for chemical compounds based on their 2D shape used in
similarity searching and molecular retrieval.

1.6 Objectives

The objectives of this research are as follows:

1. To develop a new descriptor based on 2D molecular structure shape for similarity searching.

2. To develop a new fingerprint descriptor for shape based representation of chemical compound.

3. To develop canonical representation based on 2D shape of molecule and graph theory for molecular similarity searching.

1.7 Research Scope

In order to achieve the objectives stated above, the scope of this study is limited to the following:

1. Develop the new descriptor based on the shape of molecules extracted from 2D connection table representation.

2. To evaluate the new descriptor and compared with the six standard descriptors (ALOGP, MACCS, EPFP4, CDKFP, PCFP, GRFP), then evaluate using Tanimoto coefficient.

3. The database aimed to be used in this study is only limited to chemical data from MDL Drug Data Report (MDDR).
1.8 Significance and justification of study

The field of molecular descriptors is strongly interdisciplinary and involves a mass of different theories. For the definition of molecular descriptors, knowledge of algebra, graph theory, information theory, computational chemistry, theories of organic reactivity and physical chemistry is usually required. This study is going to produce a descriptor based on shape of molecules, which can be used in similarity searching in chemical database for drug design and discovery.

1.9 Organization of the report

This thesis will be organized into seven Chapters. Chapter 2 will describe the relevant literature review, the fundamentals of chemoinformatics from the molecule descriptors and the molecule searching. Chapter 3 introduces the methodology that is used to build up the proposed descriptor. Chapter 4 introduces the development of the first model labeled as "Language for Writing Descriptor or Outline Shape of Molecule (LWDOsM)". The development of the second model 2D Fingerprint Descriptors of Outline Shape of Molecules using the stepwise fragmentation of the first language is presented in Chapter 5 and finally Chapter 6 presents the third and the last model of our work shape-based molecular descriptor SBDM. Chapter 7 discusses and concludes this thesis, highlights the contribution and finding of this work, and provide suggestions and recommendations for future research.

1.10 Summary

The molecular descriptor 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. This Chapter gives an overview of the problems involved in the representation of chemical database that store the molecular structures. It highlights the importance of the problem, especially for similarity searching and drug discovery.
REFERENCES


8. Craig and Ebert. Eleven years of structure seraching using the skf (smith,


39. Dmitrii Filimonov, Vladimir Poroikov, Yulia Borodina, and Tatyana Gloriotova. Chemical similarity assessment through multilevel neighborhoods of atoms: Definition and comparison with the other descriptors. *Journal of*


50. David J. Wild and Peter Willett. Similarity searching in files of


86. Tiejun Cheng, Qingliang Li, Zhigang Zhou, Yanli Wang, and Stephen H


