NEW SIMILARITY MEASURES FOR LIGAND-BASED VIRTUAL SCREENING

MUBARAK HUSSEIN IBRAHIM HIMMAT

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

Faculty of Computing
Universiti Teknologi Malaysia

AUGUST 2017
I dedicate this work to my beloved parents, my wife, my brothers, my sisters, and to my lovely sons.
ACKNOWLEDGMENTS

In the Name of Allah, Most Gracious, Most Merciful. First and foremost, all praise and thanks to Allah (SWT), then I would like to extend thanks to the many people, who so generously contributed to the work presented in this thesis.

Firstly, I would like to express my sincere gratitude to my supervisor, Prof. Dr. Naomie Salim, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I greatly appreciate her vast knowledge and skill in many areas of research, she added considerably to my graduate experience; we learned from her how we could acquire knowledge in the better manner and without her assistance and her full supervision this work will never be a reality.

My sincere thanks also go to all members of our research group and especially chemoinformatics group, and many thanks to the University of Technology Malaysia and its community for their support and assistance during research years.

Finally, but by no means least, a gratitude thanks go to my all family especially my parents, my wife, my sons, my sisters and brothers for their patience, encouragement, and support, thank you so much for your Prayers throughout my Ph.D. and my life in general.

Last but not the least, I would like to thank my family: my parents and to my wife, brothers and sister for supporting me throughout my Ph.D. and my life in general.
ABSTRACT

The process of drug discovery using virtual screening techniques relies on “molecular similarity principle” which states that structurally similar molecules tend to have similar physicochemical and biological properties in comparison to other dissimilar molecules. Most of the existing virtual screening methods use similarity measures such as the standard Tanimoto coefficient. However, these conventional similarity measures are inadequate, and their results are not satisfactory to researchers. This research investigated new similarity measures. It developed a novel similarity measure and molecules ranking method to retrieve molecules more efficiently. Firstly, a new similarity measure was derived from existing similarity measures, besides focusing on preferred similarity concepts. Secondly, new similarity measures were developed by reweighting some bit-strings, where features present in the compared molecules, and features not present in both compared molecules were given strong consideration. The final approach investigated ranking methods to develop a substitutional ranking method. The study compared the similarity measures and ranking methods with benchmark coefficients such as Tanimoto, Cosine, Dice, and Simple Matching (SM). The approaches were tested using standard data sets such as MDL Drug Data Report (MDDR), Directory of Useful Decoys (DUD) and Maximum Unbiased Validation (MUV). The overall results of this research showed that the new similarity measures and ranking methods outperformed the conventional industry-standard Tanimoto-based similarity search approach. The similarity measures are thus likely to support lead optimization and lead identification process better than methods based on Tanimoto coefficients.
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>ACKNOWLEDGMENT</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>xii</td>
<td></td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xv</td>
<td></td>
</tr>
<tr>
<td>LIST OF ABBREVIATION</td>
<td>xix</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1.2 Problem Background</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1.3 Problem Statement</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>1.4 Research Questions</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>1.5 Research Objectives</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>1.6 Importance of the Study</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>1.7 Scope of study</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>1.8 Thesis outline</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>1.9 Summary</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>
# Molecular Representation and Similarity Concepts

## 2.1 Introduction

## 2.2 Molecular Representations

## 2.3 Molecular Descriptors

### 2.3.1 1D Descriptors

### 2.3.2 2D Descriptors

#### 2.3.2.1 2D Fingerprints

### 2.3.3 3D Descriptors

## 2.4 Virtual Screening

## 2.5 Chemical Database Search

### 2.5.1 Discussion of Similarity Searching

## 2.6 Basic Similarity Concepts

### 2.6.1 Measuring

### 2.6.2 Compared objects

### 2.6.3 Objects Characteristics

### 2.6.4 Similar Property Principle

## 2.7 Compounds similarity searching

## 2.8 Similarity Coefficients

## 2.9 Similarity Search Practice

## 2.10 Conventional VS Similarity Coefficients

## 2.11 Bit string Reweighting

## 2.12 Similarity Measure Proprieties in information Retrieval

## 2.13 Standard Quantum-Based Similarity Model

## 2.14 Discussion on Similarity Coefficients

## 2.15 Non-Linear Similarity Methods

## 2.16 Molecule Sub-structural Analysis

### 2.16.1 Fragment Reweighting and

### 2.16.2 Explicit Feedback
2.16.3 Implicit Feedback 67
2.16.4 Pseudo Feedback 68
2.17 Ranking Approaches 68
2.17.1 Probability Ranking Principle 69
2.17.2 The Maximal Marginal Relevance (MMR) 73
2.18 Discussion 74
2.19 Summary 77

3 RESEARCH METHODOLOGY 78
3.1 Introduction 78
3.2 Research Design 79
3.3 Research Framework 81
  3.3.1 Phase 1: Preliminary Phase 83
  3.3.2 Phase 2: Constructing algorithm for Virtual Screening 83
  3.3.3 Phase 3: Similarity Based-Virtual Screening Using Bit-string Reweighting 84
  3.3.4 Phase 4: Adapting document similarity measure for ligand-based virtual screening Comparing the retrieval with conventional similarity methods 84
  3.3.5 Phase 5: Using Maximal Marginal Relevance in Ligand-Based-Virtual Screening 85
  3.3.6 Phase 6: Report Writing 85
3.4 The Database 86
3.5 Evaluation Measures of the Performance 90
  3.5.1 Kendall's W Significance Tests 91
  3.5.2 ROC Curve 92
3.6 Summary 93
4 SIMILARITY-BASED VIRTUAL SCREENING USING BIT-STRING REWEIGHTING

4.1 Introduction 95
4.2 The ASMSC Algorithm 96
4.3 Experimental Design 100
4.4 Experimental Results 101
4.5 Discussion 116
4.6 Summary 121

5 ADAPTING DOCUMENT SIMILARITY MEASURES FOR LIGAND-BASED VIRTUAL SCREENING

5.1 Introduction 122

5.2 New similarly measure for similarity-based Virtual Screening 123
5.2.1 Introduction 123
5.2.2 The proposed similarity measure 125
5.2.3 Experimental Design 129
5.2.4 Experimental Results 129
5.2.5 Discussion 134

5.3 Adapting Document Similarity Measures For Ligand-Based Virtual Screening 135
5.3.1 The Adapted Similarity Measure of Text Processing (ASMT) 135
5.3.2 Experimental Design 138
5.3.3 Experimental Results 139
5.3.4 Discussion 140

5.4 Summary 153
6 MAXIMAL MARGINAL RELEVANCE IN LIGAND-BASED VIRTUAL SCREENING

6.1 Introduction 155
6.2 The Maximal Marginal Relevance for VS 156
   6.2.1 MMR Calculation Steps 157
6.3 Experimental Design 161
6.4 Experimental Results 164
6.5 Discussion 174
6.6 Summary 178

7 CONCLUSION AND FUTURE WORKS 180

7.1 Introduction 180
7.2 Summary of Results 181
7.3 Research contributions 182
7.4 Future Work 183

REFERENCES 185

Appendix A 207
### LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Summarization of problem background</td>
<td>8</td>
</tr>
<tr>
<td>2.1</td>
<td>Example of SMILES string for some molecules</td>
<td>23</td>
</tr>
<tr>
<td>2.2</td>
<td>Examples of similarity search in different aspects</td>
<td>34</td>
</tr>
<tr>
<td>2.3</td>
<td>Common Distance and Correlation Coefficients</td>
<td>42</td>
</tr>
<tr>
<td>2.4</td>
<td>Common similarity measures most used in text retrieval</td>
<td>53</td>
</tr>
<tr>
<td>3.1</td>
<td>MDDR activity classes for DS1 dataset</td>
<td>87</td>
</tr>
<tr>
<td>3.2</td>
<td>MDDR activity classes for DS2 dataset</td>
<td>88</td>
</tr>
<tr>
<td>3.3</td>
<td>DUD Selected 11 activity classes</td>
<td>89</td>
</tr>
<tr>
<td>3.4</td>
<td>MUV activity classes</td>
<td>89</td>
</tr>
<tr>
<td>4.1</td>
<td>Retrieval results of top 1% and 5% for DS1 (ECFP_4) dataset</td>
<td>102</td>
</tr>
<tr>
<td>4.2</td>
<td>Retrieval results of top 1% and 5% for DS2 (ECFP_4) dataset</td>
<td>103</td>
</tr>
<tr>
<td>4.3</td>
<td>Retrieval results of top 1% and 5% for DS1 dataset (ALOGP)</td>
<td>104</td>
</tr>
<tr>
<td>4.4</td>
<td>Retrieval results of top 1% and 5% for DS2 dataset (ALOGP)</td>
<td>105</td>
</tr>
<tr>
<td>4.5</td>
<td>Retrieval results of top 1% and 5% for DS1 dataset (PubChem).</td>
<td>106</td>
</tr>
<tr>
<td>4.6</td>
<td>Retrieval results of top 1% and 5% for DS2 dataset (PubChem)</td>
<td>107</td>
</tr>
<tr>
<td>4.7</td>
<td>Retrieval results of top 1% and 5% for DUD dataset</td>
<td>108</td>
</tr>
<tr>
<td>4.8</td>
<td>Number of shaded cells for mean recall of actives using different search models for DS1, DS2, and DS3 Top 1% and 5%</td>
<td>117</td>
</tr>
</tbody>
</table>
4.9 Comparison results of enrichment values of (BEDROC $\alpha = 20$) and (EF 1%) using MMR-VS on MDDR1, MDDR2, and DUD data sets

4.10 Rankings of TAN, SM and ASMSC approaches Based on Kendall W Test Results: DS1, DS2, and DUD at top 1% and top 5%

5.1 The recall is calculated using the top 1% and top 5% of the DUD 12 selected activity classes.

5.2 The recall is calculated using the top 1% and top 5% of the MUV activity classes

5.3 T- Test Results: DUD, MUV at top 1% and top 5%

5.4 Retrieval results of top 1% and 5% for MDDR-DS1 dataset

5.5 Retrieval results of top 1% and 5% for MDDR-DS2 dataset

5.6 The recall is calculated using the top 1% and top 5% of the MUV activity classes

5.7 Number of shaded cells for mean recall of actives using different search models for DS1, DS2, and DS3 Top 1% and 5%

5.9 Comparison results of enrichment values of (BEDROC $\alpha = 20$) and (EF 1%) using ASMTP on MDDR1, MDDR2, and MUV data sets.

5.10 Rankings of TAN, SQB and ASMTP approaches Based on Kendall W Test Results: DS1, DS2, and MUV at top 1% and top 5%

6.1 Obtained similarity values results for MMR example

6.2 The recall is calculated using the top 1% and top 5% of the DS1 data set for Tanimoto and MMR

6.3 The recall is calculated using the top 1% and top 5% of the DS2 dataset for Tanimoto and MMR

6.4 The recall is calculated using the top 1% and top 5% of the
MUV data sets for Tanimoto and MMR

6.5 Retrieval results of top 1% and 5% of a DS1 data set for ASMTP and MMR.

6.6 The recall is calculated using the top 1% and top 5% of the DS2 data set for ASMTP and MMR.

6.7 The recall is calculated using the top 1% and top 5% of the MUV 17 activity class data sets for ASMTP and MMR.

6.7 Comparison results of enrichment values of (BEDROC $\alpha = 20$) and (EF 1%) using MMR-VS on MDDR1, MDDR2, and MUV data sets

6.8 Rankings of TAN and MMR approaches based on Kendall W Test Results: DS1, DS2, and MUV at top 1% and top 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>Representation of Caffeine using Adjacency Matrix</td>
<td>18</td>
</tr>
<tr>
<td>2.2</td>
<td>Connection table for benzoic</td>
<td>20</td>
</tr>
<tr>
<td>2.3</td>
<td>connection table for benzoic acid</td>
<td>21</td>
</tr>
<tr>
<td>2.4</td>
<td>Virtual screening approaches</td>
<td>30</td>
</tr>
<tr>
<td>2.5</td>
<td>Representation of Isopropanol structure 2D and 3D</td>
<td>32</td>
</tr>
<tr>
<td>2.6</td>
<td>Similarity search general processes</td>
<td>33</td>
</tr>
<tr>
<td>2.7</td>
<td>The three different types of feedback</td>
<td>67</td>
</tr>
<tr>
<td>3.1</td>
<td>The brief description of research design</td>
<td>80</td>
</tr>
<tr>
<td>3.2</td>
<td>Research framework</td>
<td>82</td>
</tr>
<tr>
<td>4.1</td>
<td>Flow chart of the process of adapting of SMC metric.</td>
<td>98</td>
</tr>
<tr>
<td>4.2</td>
<td>Comparison of the average percentage of active compounds retrieved at cut-off 1% for DS1 (1024-bit ECFC_4).</td>
<td>109</td>
</tr>
<tr>
<td>4.3</td>
<td>Comparison of the average percentage of active compounds retrieved at cut-off 5% for DS1 (1024-bit ECFC_4).</td>
<td>109</td>
</tr>
<tr>
<td>4.4</td>
<td>Comparison of the average percentage of active compounds retrieved at cut-off 1% for DS2 (1024-bit ECFC_4).</td>
<td>110</td>
</tr>
<tr>
<td>4.5</td>
<td>Comparison of the average percentage of active compounds retrieved at cut-off 5% for DS2 (1024-bit ECFC_4).</td>
<td>110</td>
</tr>
<tr>
<td>4.6</td>
<td>Comparison of the average percentage of active compounds retrieved at cut-off 1% for DS1(120-bit</td>
<td></td>
</tr>
</tbody>
</table>
4.7 Comparison of the average percentage of active compounds retrieved at cut-off 5% for DS1 (120-bit ALOGP).-DS1.

4.8 Comparison of the average percentage of active compounds retrieved at cut-off 1% for DS2 (120-bit ALOGP).

4.9 Comparison of the average percentage of active compounds retrieved at cut-off 5% for DS2 (120-bit ALOGP)

4.10 Comparison of the average percentage of active compounds retrieved at cut-off 1% for DS1 (881-bit PubChem).

4.11 Comparison of the average percentage of active compounds retrieved at cut-off 5% for DS1 for (881-bit PubChem).

4.12 Comparison of the average percentage of active compounds retrieved at cut-off 1% for DS2(881-bit PubChem).

4.13 Comparison of the average percentage of active compounds retrieved at cut-off 5% for DS2 for DS2(881-bit PubChem).

4.14 Comparison of the average percentage of active compounds retrieved at cut-off 1 % for DUD dataset using Tanimoto, SM and ASMSC.

4.15 Comparison of the average percentage of active compounds retrieved at cut-off 1 % for DUD dataset using Tanimoto, SM and ASMSC.

4.16 ROC and AUCs at 5% cutoff for DS1,DS2 and DUD datasets

5.1 Comparison of the average percentage of active compounds retrieved in 1%, for DUD.
5.2 Comparison of the average percentage of active compounds retrieved in 5%, for DUD.

5.3 Comparison of the average percentage of active compounds retrieved in 1%, for MUV.

5.4 Comparison of the average percentage of active compounds retrieved in 5%, for MUV.

5.5 Comparison of the average percentage of active compounds retrieved in 1% for DS1.

5.6 Comparison of the average percentage of active compounds retrieved in 5%, for DS1.

5.7 Comparison of the average percentage of active compounds retrieved in 1%, for DS2.

5.8 Comparison of the average percentage of active compounds retrieved in 5%, for DS2.

5.9 Comparison of the average percentage of active compounds retrieved in 5%, for MUV.

5.10 Comparison of the average percentage of active compounds retrieved in 5%, for MUV.

5.11 ROC curves and AUCs at 5% cutoff of DS1 data set.

5.12 ROC curves and AUCs at 5% cutoff of DS2 data set.

5.13 ROC curves and AUCs at 5% cutoff of MUV data set.

6.1 The general way of MMR ranking method

6.2 Comparison of the average percentage of active compounds retrieved in the top 1% and top 5% of the DS1 dataset for Tanimoto and MMR

6.3 Comparison of the average percentage of active compounds retrieved in the top 1% and top 5% of the DS2 data set for Tanimoto and MMR

6.4 Comparison of the average percentage of active compounds retrieved in the top 1% and top 5% of the DS1 dataset for ASMTP and MMR.
6.5 Comparison of the average percentage of active compounds retrieved in the top 1% and top 5% for a DS2 dataset for ASMTP and MMR. 172

6.6 Comparison of the average percentage of active compounds retrieved in the top 1% and top 5% of the MUV dataset 173

6.7 Comparison of the average percentage of active compounds retrieved in the top 1% and top 5% of the MUV dataset 173

6.8 ROC curves and AUCs at 5% cutoff of the DS1 data set 176

6.9 ROC curves and AUCs at 5% cutoff of the DS2 data set 176

6.10 ROC curves and AUCs at 5% cutoff of MUV data set 177
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D</td>
<td>One Dimension</td>
</tr>
<tr>
<td>2D</td>
<td>Tow Dimension</td>
</tr>
<tr>
<td>3D</td>
<td>Three Dimension</td>
</tr>
<tr>
<td>AIM</td>
<td>Atoms-in-Molecules</td>
</tr>
<tr>
<td>AM</td>
<td>Adjacency Matrix</td>
</tr>
<tr>
<td>ASMTOP</td>
<td>Adapted Similarity Measure of Text Processing</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under the Curve</td>
</tr>
<tr>
<td>BEDROC</td>
<td>Boltzmann enhanced discrimination of receiver operating characteristic</td>
</tr>
<tr>
<td>CML</td>
<td>Chemical Markup Language</td>
</tr>
<tr>
<td>DUD</td>
<td>Directory of Useful Decoys</td>
</tr>
<tr>
<td>ECFC</td>
<td>Atom Type Extended-Connectivity Fingerprint</td>
</tr>
<tr>
<td>EEFC</td>
<td>Atom Type Atom Environment Fingerprint</td>
</tr>
<tr>
<td>EHFC</td>
<td>Atom Type Hashed Atom Environment Fingerprint</td>
</tr>
<tr>
<td>FCFC</td>
<td>Functional Class Extended-Connectivity Fingerprint</td>
</tr>
<tr>
<td>FEFC</td>
<td>Functional Class Atom Environment Fingerprint</td>
</tr>
<tr>
<td>FHFC</td>
<td>Functional Class Hashed Atom Environment Fingerprint</td>
</tr>
<tr>
<td>HTS</td>
<td>High-throughput Screening</td>
</tr>
<tr>
<td>IPRP</td>
<td>Interactive Probability Ranking Principle</td>
</tr>
<tr>
<td>k-NN</td>
<td>K-Nearest Neighbors</td>
</tr>
<tr>
<td>LBVS</td>
<td>Ligand-Based Virtual Screening</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>MCS</td>
<td>Maximal Common Substructure</td>
</tr>
<tr>
<td>MDDR</td>
<td>MDL Drug Data Report</td>
</tr>
<tr>
<td>MDL</td>
<td>Molecular Design Limited</td>
</tr>
<tr>
<td>MMR</td>
<td>Maximal Marginal Relevance</td>
</tr>
<tr>
<td>MUV</td>
<td>Maximum Unbiased Validation</td>
</tr>
<tr>
<td>PRP</td>
<td>Probability Ranking Principle</td>
</tr>
<tr>
<td>PT</td>
<td>Portfolio Theory</td>
</tr>
<tr>
<td>QBE</td>
<td>Query By Example</td>
</tr>
<tr>
<td>QSAR</td>
<td>Quantitative Structure-Activity Relationship</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>SBVS</td>
<td>Structure-Based Virtual Screening</td>
</tr>
<tr>
<td>SMILES</td>
<td>Simplified Molecular Input Line System</td>
</tr>
<tr>
<td>SMTP</td>
<td>Similarity Measure for Text Classification and Clustering</td>
</tr>
<tr>
<td>SQB</td>
<td>Standard Quantum Based</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>TAN</td>
<td>Tanimoto</td>
</tr>
<tr>
<td>TSS</td>
<td>Turbo Similarity Searching</td>
</tr>
<tr>
<td>VS</td>
<td>Virtual Screening</td>
</tr>
<tr>
<td>WLN</td>
<td>Wiswesser Line Notation</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Introduction

In chemical and pharmaceutical research, computers have been used for many years to decrease the cost of drug discovery (Todeschini and Consonni, 2009). Many different computer techniques and methods have been applied, and the data mining methods and information retrieval methods have been widely used in chemical, biomedical, and other medical fields. The actual laboratory drug discovery process can take between 12 and 15 years and can cost approximately more than one million dollars (Rollinger et al., 2008); for that, considerable effort has been made to cover research into this area. This has taken years and cost in excess of $1 billion. It is complex and costly and consumes a lot of time in laboratory experiments. These two above-mentioned reasons have attracted the attention of researchers in different aspects to solve and reduce the long drug discovery time and its high cost. One of the rich science areas within the last decades is chemoinformatics, which is a multi-disciplinary area that combines many older different disciplines such as computational chemistry, chemometrics and Quantitative Structure–Activity Relationship (QSAR). The term chemoinformatics has some synonyms in literature, as it is also known as Chemical Informatics and Chemical Information. Its general definition is “the use of computer and informational techniques applied to a range of problems in the field of chemistry” (Brown, 1998). Another definition is “The mixing of different information resources for the purpose of transforming 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” (Brown, 1998). Another general definition was given newly by Gasteiger
chemoinformatics is the use of informatics methods to solve chemical problems”.

The process of discovering new drugs using computational screening methods is being continuously developed, and improved as it is one of the most important tools for drug discovery. Virtual screening now becomes an alternative to High-throughput Screening (HTS). HTS was considered the basic and main method for drug candidate development, but virtual screening (VS) with its various techniques and search methods is becoming a reliable method for drug discovery.

Virtual screening methods can be used in many aspects of chemistry, such as molecule ranking, clustering, docking and virtual screening; as a result, this is now used as a complementary tool to HTS in drug discovery, because the rational drug discovery requires fast and computationally straightforward methods that distinguish active ligands from inactive molecules in huge molecular databases. Huge databases can be screened easily and successfully in a short time. VS, or screening as described here, is the process of selecting molecules to help in bioactivity testing. This screening is applied automatically by computer methods that select molecules; this is generally referred to as VS, and the Ligand-based virtual screening extrapolates from known active compounds used as input information and aims at identify structurally diverse compounds having similar bioactivity, regardless of the methods that are applied.

The screening methods conducted by computers are employed to rank the molecules according to their structures and put the most promising structures at the top of the list(Brown, 1998; Chen and Reynolds, 2002); this gives a high ranking to those molecules with structures that may be similar to structures that have already been tested. The screening methods and concept of molecular similarity are closely related to those used in information retrieval. Researchers have found most of the existing ligand–based similarity methods and similarity measures to be unsatisfactory, and consider the Tanimoto as the better similarity measure (Dávid Bajusz, 2015). However, some new similarity measures for information retrieval
have recently been proposed (Lin et al., 2014; Todeschini et al., 2012) as well as some proposed for virtual screening that outperformed the Tanimoto, refuting the claim that only Tanimoto could achieve better results (Al-Dabbagh et al., 2015).

Our general hypothesis for this work is that although considerable enhancements could be achieved in ligand-based virtual screening, more effort needs to be provided to help accelerate the drug discovery process and some of its major pitfalls and challenges that still need to be solved in order to handle the exponentially increased volume of molecule data (Cereto-Massagué et al., 2014; Muegge and Mukherjee, 2016). As mentioned above, the general belief is that the Tanimoto similarity measure is the best similarity measure for virtual screening in spite of many similarity measures that have been proposed and applied in other aspects of science. This belief has led researchers to ignore the recently proposed similarity measures, and at the same time reduce the determination of researchers in cheminformatics to use and modify the similarity measures that could outperform the existing similarity measures for virtual screening.

This thesis, primarily focus on ligand-based virtual screening. Different algorithms are proposed based on bit-strings and fragment-based that enhanced ligand-based virtual screenings. The rest of this chapter discusses the background of the problem, the importance of the study, the objectives and scope of this research. The last section will describe the organization and outline of the thesis.

1.2 Problem Background

Great efforts have been made to provide new drugs to the market, and there are considerable investments in the research regarding this issue. The development of a new drug consumes very long timeframes and high cost as mentioned earlier in this chapter. In chemoinformatics, researchers try to help the industry and chemists to make the drug discovery process less risky and less costly and accelerate the processing time, which takes years (DiMasi et al., 2016; Wang et al., 2016). Virtual
screening provides many tools and methods to provide considerable influence in drug discovery and in the process of obtaining a drug candidate. Recently, many new techniques have been proposed in chemoinformatics to be used as a substitute for old, traditional, synthesized laboratories testing a New Chemical Entity (NCE) approaches, high-throughput screening (HTS), combinatorial chemistry (CC)\cite{Li2016}. With HTS screening, millions of chemical, pharmacological, or genetic tests could be conducted in a short time by using computer aids that could execute a million processes in a few seconds. Although there is no doubt that considerable progress has been made in the field of computational drug discovery and ligand prediction\cite{Chen2016, DeVivo2016}, the commonly used methodology is still far from perfect, and it needs more work to satisfy chemists. According to some studies, the estimated time to produce a new drug to the market is twelve years, at an estimated cost ranging from US$92 million to US$883 million \cite{DiMasi2016, Morgan2011}. Differences in methods, data sources, and timeframes explain some of the variation in estimates. As a result, the focus of most researchers in cheminformatics is twofold: reducing the cost and time of drug discovery process, and avoiding the failure rates in later stages of drug development. Hence, the time and cost of finding and testing new chemical entities can be considered the main objective in drug discovery. For virtual screening, researchers strive for ways to find new active compounds and to bring these compounds to the market as quickly as possible.

The huge chemical compound libraries provide a good source of new potential drugs that can be randomly or methodically tested or screened to find good drug compounds. It is now possible to test hundreds of thousands of compounds in a short time using high-throughput screening techniques. Therefore, virtual chemical libraries that are done by computer systems become useful supporters that aid this process of drug discovery \cite{XuHagler2002}.

Chemists have always struggled with the difficult problem of deciding which chemical structures to synthesize among large numbers of compounds. However, this is still a small percentage of the total number that could be synthesized. Therefore, in recent years the techniques of chemical search have been
called virtual screening, which encompasses a variety of computational techniques that are used to test a large number of compounds by computer instead of experience (Bajorath, 2013; Muegge and Mukherjee, 2015; Stumpfe and Bajorath, 2011; Walters et al., 1998). These computational methods can be used for searching chemical libraries to filter out the unwanted chemical compounds, and these methods allow chemists to reduce a huge virtual library, and make it more manageable size to assess the probability that each molecule will exhibit the same activities against a specific biological target. The approaches of virtual screening can be categorized into structure-based virtual screening (SBVS) approaches (Ono et al., 2014; Vuorinen et al., 2014), and ligand-based virtual screening (LBVS) approaches. The SBVS approaches can be used when the 3D structure of the biological target is available, such as ligand-protein docking and de novo design. The LBVS approaches are applicable in the case of absence of such structural information, such as machine learning methods and similarity methods.

The similarity methods may be the simplest and most widely-used tools for LBVS of chemical databases (Cereto-Massagué et al., 2015a; Willett, 2009; Willett et al., 1998). The increased importance of similarity searching applications is mainly due to its role in lead optimization in drug discovery programs, where the nearest neighbors for an initial lead compound are sought in order to find better compounds. There are many studies in the literature associated with the measurement of molecular similarity (Bender and Glen, 2004; Maldonado et al., 2006; Nikolova and Jaworska, 2003). Similarity searching aims to search and scan chemical databases to identify those molecules that are most similar to a user-defined reference structure using some quantitative measures of intermolecular structural similarity. However, the most common approaches are based on 2D fingerprints, with the similarity between a reference structure and a database structure computed using association coefficients such as the Tanimoto coefficient (Dávid Bajusz, 2015; Deng et al., 2015; Johnson and Maggiora, 1990; Todeschini et al., 2012). The similarity measures methods play a significant role in detecting the rate for pairwise molecular similarity (Lynch and Ritland, 1999). These methods can be employed to find the most similar molecules among thousands of compounds, and then organize these similar molecules in decreasing order depending on the probability ranking.
principle that only relies on the values of probability between the molecules and molecular target.

In general, the processes of a similarity measure for molecules have two stages, which are similarity stage and ranking stage. At similarity level, the performance of conventional similarity methods has been enhanced in various ways. Some studies have used the weighting scheme (Abdo and Salim, 2010; Ahmed et al., 2012; Jaghoori et al., 2015; Kar and Roy, 2013; Klinger and Austin, 2006), while others have employed the techniques of data fusion (Ahmed et al., 2014; Salim et al., 2003; Willett, 2013b). The relevant feedback has also been applied and used in LBVS to improve the performance of similarity methods (Abdo et al., 2012; Abdo et al., 2011). However, the effectiveness of any similarity method has been found to vary greatly from one biological activity to another in a way that is difficult to predict (Gasteiger, 2016; Sheridan and Kearsley, 2002). In addition, the use of any two methods has been found to retrieve different subsets of actives from the chemical library, so it is advisable to utilize several search methods where possible.

Considerable effort has been expended in finding the appropriate similarity measures in virtual screening among such available of choices of similarity measures, and this has attracted the attention of researchers from the early time of High Throughput Screening, and cheminformatics.

Many similarity measures have been applied in cheminformatics for virtual screening. These similarity measures have contributed in screening performance. Some other similarity measures have been adapted and derived from existing similarity measures and achieved good results in other areas, but haven’t been applied in virtual screening. In addition, many similarity measures have been proposed for text (Lin et al., 2014), and could be adapted for virtual screening due to many similar aspects between the text and chemical information retrieval. Thus, the algorithms that have been applied in text information retrieval can also be applied in chemical information retrieval (Obaid et al., 2017; Willett, 2000a).
The fragment bases and bit-strings similarity method has gained attention from researchers in chemoinformatics and especially in virtual screening (Abdo and Salim, 2010; Ahmed et al., 2012; Chen and Reynolds, 2002; Holliday et al., 2002; Zoete et al., 2009), and many types of research are focused on it. The molecules databases (fingerprint) contain a large number of bit-strings that represent the molecules features (Bajorath, 2017; from Structure, 1997; Todeschini and Consonni, 2009; Todeschini et al., 1994), and considering all these features as the same and giving them same weight features in similarity calculations is not fair. This is because most proposed methods usually assume that all molecular features are equal in importance. On the other hand, all weighting schemes calculate the weight for each feature independently with no relation to all other features, in general, The summarization of the all mentioned problem background are demonstrated in Table 1.1. For all these mentioned cases, in order to enhance the virtual screening effectiveness, feature reweighting using important bit-strings calculations can enhance the recall of similarity measure.

In order to enhance the effectiveness of the similarity measure, the primary aim of this research is to propose ligand-based similarity methods, and propose a ranking method based on bit-strings and fragment-based reweighting. Additional aims include adapting an existing similarity measure, adapting text similarity measure and proposing alternative ranking method to be used for ligand-based virtual screening.
Table 1.1: Summarization of problem background

<table>
<thead>
<tr>
<th>Issue</th>
<th>What have been done in LR</th>
<th>Why not enough</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity Method</td>
<td><strong>Enhancement of similarity measures using:</strong>&lt;br&gt;Similarity coefficients (Consonni and Todeschini, 2012; Dávid Bajusz, 2015; Lin et al., 2014; Rognan and Bonnet, 2014; Todeschini et al., 2012).&lt;br&gt;• Data Fusion (Chen et al., 2010; Sastry et al., 2013; Willett, 2013a).&lt;br&gt;• Relevance feedback (Abdo et al., 2012; Agarwal et al., 2010; Chen et al., 2009b).&lt;br&gt;Weighting functions (Ahmed et al., 2012; Arif et al., 2010; Holliday et al., 2013).&lt;br&gt;Machine Learning (Cereto-Massagué et al., 2015b; Durrant and Amaro, 2015; H Haga and Ichikawa, 2016; Lavecchia, 2015).&lt;br&gt;</td>
<td>Although several similarity coefficients and techniques have been applied to enhance VS, but the area of VS still requires more investigation to determine whether other coefficients might yield a higher level of screening effectiveness than those which been used for virtual screening.</td>
<td>Enhance the effectiveness of the Ligand-based similarity searching method by adapting several similarity measures from information retrieval field. <strong>Adapted Similarity Measure of Text Processing (ASMTP)</strong></td>
</tr>
<tr>
<td>Fragment Reweighting</td>
<td><strong>Finding new weighting schemes or functions</strong> (Abdo and Salim, 2010; Arif et al., 2010; Holliday et al., 2013).&lt;br&gt;</td>
<td>There are other weighting features methods need to be investigating to assign more weights to the bit-strings for improving the effectiveness of LBVS.</td>
<td>Enhance the effectiveness of similarity measure by reweighting molecular bit strings. <strong>Adapted Simple Matching Similarity Coefficient (ASMSC)</strong></td>
</tr>
<tr>
<td>Molecular Ranking Principle</td>
<td>• Enhancement of PRP (Text IR &amp; Chemical IR)&lt;br&gt;• Classification methods (Dörr 2015, Chen 2014, Rathke 2010)&lt;br&gt;• Regression methods ( Li 2011, Hasegawa 2010)&lt;br&gt;• Data Fusion (Willett, 2013)&lt;br&gt;• Alternative ranking approaches (Text IR)&lt;br&gt;• QPRP Quantum probability ranking principle (Zuccon, 2012)&lt;br&gt;• IPRP Interactive Probability Ranking Principle (Sheridan ,2008)&lt;br&gt;</td>
<td>One of the key controversial issues of PRP is the independence among ranking compounds, which prevents molecule’s ranking position from the effect of other molecules.</td>
<td>Enhance the effectiveness of similarity measure by using Maximal Marginal Relevance (MMR) ranking principle of molecules that is inspired from text and document retrieval domain. <strong>Maximal Marginal Relevance (MMR) for LBVS</strong></td>
</tr>
</tbody>
</table>

Notes: The text describes various methods used in similarity searching, with a focus on enhancing the effectiveness of virtual screening techniques in drug discovery programs. It highlights the use of similarity measures, data fusion, relevance feedback, weighting functions, and machine learning approaches. The table compares what has been achieved in computational methods with the reasons why more effort is needed, and proposes new approaches to improve screening effectiveness.
1.3 Problem Statement

In general, the aim of virtual screening is developing new drugs, in addition, its significance is to decrease the consumption of times and cost which is considered a big challenge in drug discovery process, where the estimated cost of drug discovery exceeding millions and years to discover new drugs, and virtual screening reduce this cost to be very low compared to conducting experiments in real laboratory screening.

By understanding the problem background that has been discussed in the previous section, it can be concluded that the needs of many chemical similarity search methods is considered one of the continuing challenges in cheminformatics (Sheridan and Kearsley, 2002), the ligand-based virtual similarity methods have been under development for decades, and the ligand-based virtual screening field still needs more investigation. In addition, in coming up with a new proposed similarity measure and a similar information retrieval field for improvement, there are limitations of the currently used similarity measures.

The aim of this study research is to develop a ligand-based similarity method based on developing algorithms that emphasize the common structural features (bit-strings) and give high priority in similarity calculations, and reweighting some bit-strings when conducting the search on chemical databases to retrieve the active compounds with the most similar biological activity to the specific reference structure.

Recently, many studies in text information retrieval have proved that retrieval models are based on some new similarity measures and have provided significant improvements in retrieval performance compared to conventional models, and this could be adapted for ligand-based virtual screening.
The all developed similarity methods as well as benchmark similarity coefficients have used the classical ranking approach when ranking the chemical structures, and this study will also investigate the most popular common ranking methods used in information retrieval and propose an alternative method to conventional probability ranking principle (PRP) (Robertson, 1977).

The proposed algorithms apply different approaches to fingerprint data fragment reweighting; this approach is based on fragment reweighting factors. Fragment reweighting here is the process of adding some constant weight to the original weight in order to improve retrieval performance in information retrieval systems. This approach has been derived from document retrieval filed.

The core of virtual screening is to develop anew drugs that decrease the consumption of times and cost. will help in development of representation of time spent on the virtual screening experiments is not taken as a big issue when it has been compared to the high cost and long duration of screening of molecules in a real laboratory. For that this research does not concern the time of virtual screening as an important factor.

1.4 Research Questions

Referring to the problem background, the main questions of this research are:

- Can some similarity measures from document retrieval be adapted to improve ligand-based virtual screening?
- How can new similarity matrices be developed for virtual screening using some preferred similarity measure properties used in document retrieval areas?
Can the ligand-based virtual screening performance be improved by reweighting some bit-strings of the features?

How can other ranking method be proposed to improve the effectiveness of virtual screening?

1.5 Research Objectives

The main goal of this research is to develop a similarity-based virtual screening approach using reweighted fragments or the bit-strings, with the ability to improve the retrieval effectiveness and provide an alternative to existing tools for ligand-based virtual screening. Therefore, our general hypothesis for this study is How could constructing and adapting similarity measures and ranking methods from document retrieval can help improve the retrieval performance of molecular similarity? To achieve this goal, the following objectives have been set:

- To investigate some molecule features (bit-strings) to be reweighting for enhance retrieval effectiveness of VS.
- To formulate and adapt new similarity metric for ligand-based virtual screening. Virtual screening.
- To formulate a similarity-based virtual screening method for molecular similarity searching based on text and document retrieval similarity measure concepts.
- To formulate and develop alternative ranking method for ligand-based virtual screening instead of conventional probability ranking principle (PRP).
1.6 Importance of the Study

This study introduced some ligand-based virtual screening algorithms that incorporate adaptation and modification of some similarity measures in order to enhance the efficiency of ligand-based virtual screening. It is also suggested an alternative ranking method that could outperform probability-ranking principle (PRP), which is considered the most popular ranking theory for current similarity searching methods in LBVS. The study rely on the believe that some modification of the existing methods could provide valuable enhancement.

1.7 Scope of the Study

This study will focus on ligand-based virtual screening, especially on similarity-based virtual screening using 2D fingerprint representations of molecular structure. The 2D fingerprint is a vector that encodes the presence and absence of the topological structure that represents the typical atoms, bond, or ring-canter fragment. The proposed screening methods mentioned before will be used to quantify the degree of structural resemblance between a pair of molecules characterized by 2D fingerprints. Most methods are applied with both binary and non-binary 2D fingerprints descriptors. The study focuses on the fragment, bit-string and reweighting methods and similarity coefficients and ranking methods to present an enhancement of molecular retrieval. The bit-strings emphasize the common structural features (bit-strings) and give high priority in similarity calculations. The reweighting factor here will take some similarity concepts to reweight some bit-string values.

The proposed virtual screening enhancement solutions in this study have been evaluated by simulated virtual screening experiments that were conducted on large benchmark datasets which have been derived from MDL Drug Data Report (MDDR) database ("Symyx Technologies. MDL drug data report: Sci Tegic Accelrys Inc., the
MDL Drug Data Report (MDDR). Database is available at http://www.accelrys.com/), Maximum Unbiased Validation (MUV) (Rohrer and Baumann, 2009;), the MDL Drug Data Report (MDDR) and the Directory of Useful Decoys (DUD) (Rohrer and Baumann, 2009) where single and multiple reference structures are available. The performance of this method is evaluated against the performance of conventional 2D similarity measure Tanimoto.

1.8 Thesis Outline

This section describes the organization of the thesis. There are seven chapters in this thesis, which are:

Chapter 1, Introduction: this chapter gives a general introduction to chemoinformatics, drug discovery, and virtual screening topic of the proposed research work. There are brief overviews of some of the issues concerning the virtual screening research area, and it briefly discusses the following topics: problem background, the problem statement, objectives of the study, research scope, and significance of the study.

Chapter 2, Molecular representations and Similarity concepts: this chapter begins with an overview of computer representations of chemical structures and various types of searching mechanisms offered by chemical information systems. In the third section, we present molecular representations that can be employed for molecular similarity searching as well as for molecular analysis and clustering. The chapter describes in detail the 2D fingerprint-based similarity methods and different types of similarity coefficients. The chapter also briefly discusses the implementation of machine learning techniques to molecular similarity and similarity measures of text and document areas. At the end of the chapter there is a conclusion that summarizes the applicability of the discussed methods to molecular similarity searching and the best ways to improve the performance of these methods.
Chapter 3, *Research Methodology*: this chapter describes the overall methodology adopted in this research to achieve the objectives of this thesis; it presents the methodology used in this research. A methodology is generally a guideline for solving a research problem. It contains the generic framework of the research and the steps required to carry out the research systematically, and it discusses in detail the datasets that will be used to conduct the experiments of the proposed methods. This includes discussion on the research components such as the phases, techniques, and tools involved. At the end of the chapter, we will conclude with a summary.

Chapter 4, *Enhancing Ligand-based Virtual Screening Using Bit-strings Reweighting*: this chapter introduces the new ligand-based virtual screening ranking algorithm, called Adapted Simple Matching Similarity Coefficient (ASMSC) that emphasizes the common molecular structural features (bit-strings) to be given a high priority in similarity calculations. The chapter describes the construction of the algorithm and experiments done to evaluate the proposed coefficient. In the results and discussion section, the results are presented and discussed.

Chapter 5, *constructing new similarity metric and Adapting Document Similarity Measures for Ligand-based Virtual Screening*: the study investigates the newly documented similarity measure and adapts it for ligand-based virtual screening. The adapted SMTP algorithm focuses on the preferred selected similarity properties. In the results and discussion section, all experiments conducted on different datasets are discussed, and the chapter also discusses comparison of the achieved results with the standard coefficient of VS, and discusses the investigation of the effectiveness of proposed adapted similarity measure. At the end of the chapter we will conclude with a summary.

Chapter 6, *Using Maximal Marginal Relevance in Ligand-based Virtual Screening*: the chapter investigates the susceptibility of using the concepts of MMR in order to enhance the efficiency of ligand-based virtual screening. We will examine the use of MMR with different datasets to investigate its capability to
improve virtual screening. The chapter discusses some ranking methods that have been applied in information retrieval, and it covers a comparison of the achieved results with the standard coefficient of VS. It also discusses the investigation of the effectiveness of proposed adapted similarity ranking. At the end of the chapter, we will conclude with a summary.

Chapter 7, *Conclusion and Future Work*: this is the last chapter, and it provides a conclusion of the overall work of this thesis. It highlights the findings and contribution made by this study and provides suggestions and recommendations for future research.

### 1.9 Summary

In this chapter, we give a broad overview of the problems involved in the molecular similarity. This chapter serves as an introduction to the research problem set out earlier in this thesis. The goal, objectives, the scope and the outline of this thesis are also presented.


Ginn, Claire MR, Willett, Peter, & Bradshaw, John. (2002). Combination of molecular similarity measures using data fusion *Virtual Screening: An Alternative or Complement to High Throughput Screening?* (pp. 1-16). Netherlands: Springer.


Han, LY, Ma, XH, Lin, HH, Jia, J, Zhu, F, Xue, Y, . . . Chen, YZ. (2008). A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor. *Journal of Molecular Graphics and Modelling, 26*(8), 1276-1286.


Cheminformatics studies to analyze the therapeutic potential of phytochemicals from Rhazya stricta. *Chemistry Central Journal, 11*(1), 11.


Whittle, Martin, Willett, Peter, Klaffke, Werner, & van Noort, Paula. (2003). Evaluation of similarity measures for searching the dictionary of natural


LIST OF PUBLICATIONS


