

**FUSION OF SIMILARITY MEASURES USING GENETIC ALGORITHM  
FOR SEARCHING CHEMICAL DATABASE**

**YAHYA ALI ABDELRAHMAN ALI**

**UNIVERSITI TEKNOLOGI MALAYSIA**

# UNIVERSITI TEKNOLOGI MALAYSIA

## DECLARATION OF THESIS / UNDERGRADUATE PROJECT PAPER AND COPYRIGHT

Author's full name : **YAHYA ALI ABDELRAHMAN ALI**

Date of birth : **10 NOVEMBER 1975**

Title : **FUSION OF SIMILARITY MEASURES USING GENETIC ALGORITHM FOR SEARCHING CHEMICAL DATABASE**

Academic Session: **2007/2008**

I declare that this thesis is classified as :

**CONFIDENTIAL**

(Contains confidential information under the Official Secret Act 1972)\*

**RESTRICTED**

(Contains restricted information as specified by the organization where research was done)\*

**OPEN ACCESS**

I agree that my thesis to be published as online open access (full text)

I acknowledged that Universiti Teknologi Malaysia reserves the right as follows:

1. The thesis is the property of Universiti Teknologi Malaysia.
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Certified by :

\_\_\_\_\_  
SIGNATURE

\_\_\_\_\_  
SIGNATURE OF SUPERVISOR

**B0443950**  
\_\_\_\_\_  
(NEW IC NO. /PASSPORT NO.)

**PM.DR.NAOMIE BINTI SALIM**  
NAME OF SUPERVISOR

Date : \_\_\_\_\_

Date : \_\_\_\_\_

**NOTES :** \* If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction.

“I declare that I have read this project and in my opinion this project report has satisfied the scope and quality for the award of the degree of Master of Science (Computer Science)”.

Signature : \_\_\_\_\_

Name of Supervisor : Assoc. Prof. Dr. Naomie Bt Salim

Date : \_\_\_\_\_

FUSION OF SIMILARITY MEASURES USING GENETIC ALGORITHM FOR  
SEARCHING CHEMICAL DATABASE

YAHYA ALI ABDELRAHMAN ALI

A project report submitted in partial fulfillment of the  
requirements for the award of the degree of  
Master of Science (Computer Science)

Faculty of Computer Science and Information System  
University Technology Malaysia

NOVEMBER 2007

“I declare that this project report is the result of my own research except as cited in references.”

Signature : \_\_\_\_\_

Name of Candidate: YAHYA ALI ABDELRAHMAN ALI

Date : \_\_\_\_\_

## DEDICATION

*“This thesis dedicated to my parents, beloved family and to whoever serves the truth  
for the truth itself.”*

## ACKNOWLEDGEMENTS

First, I would like to thank ALLAH S.W.T. for all the achievements that I have gained today. Next, I wish to extend my grateful appreciation to all those who have contributed directly and indirectly to the preparation of this study. I would like to take this opportunity to thank my supervisor, Associate Professor Dr. Naomie BT Salim for attention, encouragement and guidance throughout the length of this study. Not forgetting my beloved family for all supports and understandings they gave me. Not to forget also, my examiners Associate Professor Dr. Siti Mariyam Binti HJ. Shamsudin and Dr. Siti Zaiton BT. Mohd Hashim for many helpful suggestions.

Second, I would like to thank my evaluators for their critics and comments. All have helped me improve my work as well as my report writing.

Lastly, I am grateful to all my colleagues, friends, staff, and lecturers in Faculty of Computer Science and Information System, University Technology Malaysia and Sudan University of Science and Technology for their help and support at every step during this study.

## ABSTRACT

Similarity searching is a process to find compounds that are similar to a target compound, which is useful in discovering potential drugs. The main objective for this study is to optimize weights of different similarity measures in data fusion for searching chemical database by applying genetic algorithm (GA). Comparisons of different coefficient fusions were carried out. The results show that the Tanimoto, Cosine, Kulcznski(2) and Fossum coefficients are the best single coefficient. Cosine and Fossum coefficients gave the best combination for 2-coefficient fusion with weightings of 0.960 and 0.937 respectively. For 3-coefficient fusion Russell-Rao, Tanimoto and Cosine coefficients of weightings 0.972, 0.960 and 0.960 respectively gives the best result. The combinations Tanimoto and Cosine coefficients perform well and give large number of actives. Using combination with weights ranging between 0.0 and 1.0 generated by genetic algorithm, gave a better number of active than the non-weighted combination. Cosine and Fossum coefficients combined without weights yields an average 21.89% among the top 10% compound; whereas when genetic algorithm (GA) is used to combine Cosine and Fossum Coefficients with weights of 0.960 and 0.937 respectively, an average of 22.16% among the top 10% compound is obtained. Generally, the combinations of coefficients performed better than the single coefficients.



## ABSTRAK

Pencarian keserupaan merupakan suatu proses untuk mencari sebatian-sebatian yang adalah menyerupai satu sebatian sasaran, yang adalah berguna dalam penemuan berpotensi ubat. Matlamat utama untuk kajian ini ialah untuk mengoptimumkan pemberat dari pengukuran keserupaan yang berbeza dalam lakuran data untuk penggeledahan pangkalan data kimia oleh penyelenggaraan algoritma genetik (GA). Perbandingan-perbandingan perpaduan koefisien yang berbeza telah dijalankan. Hasil-hasil menunjukkan bahawa Tanimoto, Cosine, Kulcznski(2) dan koefisien-koefisien Fossum adalah koefisien tunggal dan terbaik. Kosinus dan koefisien-koefisien Fossum memberi gabungan terbaik untuk 2 koefisien lakuran dengan pemberat 0.960 dan 0.937 untuk masing-masing. Untuk 3 koefisien lakuran Russell-Rao, Tanimoto dan koefisien-koefisien Cosine elaan sara hidup 0.972, 0.960 dan 0.960 masing-masing memberi keputusan terbaik. Gabungan-gabungan Tanimoto dan koefisien-koefisien Cosine berjalan dengan baik dan memberi sejumlah besar aktif. Menggunakan gabungan dengan pemberat berjarak di antara 0.0 dan 1.0 dijana dengan algoritma genetik, telah memberi sejumlah aktif yang lebih baik daripada gabungan tidak menggunakan pemberat. Kosinus dan koefisien-koefisien Fossum digabungkan tanpa pemberat hasil purata 21.89% di antara bahagian teratas 10% sebatian; manakala algoritma genetik (GA) adalah digunakan untuk menggabungkan Cosine dan Fossum Coefficients dengan pemberat 0.960 dan 0.937 untuk masing-masing, puratanya adalah 22.16% di antara 10% teratas sebatian telah diperolehi. Umumnya, gabungan-gabungan koefisien memberikan hasil lebih baik daripada koefisien-koefisien tunggal.

## TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
	TITLE	i
	DECLARATION	ii
	DEDICATION	iii
	ACKNOWLEDGEMENTS	iv
	ABSTRACT	v
	ABSTRAK	vi
	TABLE OF CONTENTS	vii
	LIST OF TABLES	xi
	LIST OF FIGURES	xiii
	TABLE OF SYMBOLS	xvi
	LIST OF ABBREVIATIONS	xvii
	LIST OF TERMINOLOGIES	xviii
	LIST OF APPENDICES	xix
1	INTRODUCTION	1
	1.1 Introduction	1
	1.2 Problem background	3
	1.3 Problem statement	5
	1.4 Objectives	5
	1.5 Project Scope	6
	1.6 Justifications for Employing GA	6

1.7	Report Organization	7
<b>2</b>	<b>LITERATURE REVIEW</b>	<b>8</b>
2.1	Chemical Database	8
2.1.1	Chemical compounds	8
2.1.2	Representation of Chemical Structures	9
2.1.2.1	Fragmentation Codes	10
2.1.2.2	Linear Notations	11
2.1.2.3	Connection Tables	11
2.2	Retrieval of Data from Chemical Database	13
2.2.1	Structure Searching	13
2.2.2	Substructure Searching	14
2.2.3	Similarity Searching	15
2.3	Molecular Descriptors For Similarity Searching And Similarity Coefficient	16
2.3.1	Molecular Representation	17
2.3.2	Similarity Coefficients	21
2.4	Data Fusion	23
2.5	Data Fusion in Chemical Compound	26
2.5.1	Fusions of Molecular Representation	27
2.5.2	Fusions of Several Molecules in a Single Query	28
2.5.3	Fusions of Similarity Coefficients	28
2.6	Genetic Algorithm	31
2.6.1	Components of the Genetic Algorithm	34
<b>3</b>	<b>METHODOLOGY</b>	<b>42</b>
3.1	Introduction	42
3.2	Operational Framework	42
3.2.1	Research Plan and Review of Literature	44
3.2.2	Collection of Chemical Data Set	44
3.2.3	Converting Molecule Data to Barnard Chemical	

	Information (BCI)	45
3.2.4	Converting Molecule Data to Topological Indices	47
3.2.5	Data Fusion Process	50
3.2.6	Design Basic Genetic Algorithm (GA)	53
3.2.7	Optimization of Fusion Process	54
3.2.8	Applying the Genetic Algorithm (GA) for Optimization of Fusions	55
3.3	Evaluation of the GA	58
3.4	System Requirements	59
3.5	Summary	59
<b>4</b>	<b>EXPERIMENTAL RESULTS AND DISCUSSIONS</b>	<b>60</b>
4.1	Introduction	60
4.2	Representation of Chemical Compounds	60
4.3	Compounds Retrieval	62
4.4	Similarity Calculation	64
4.5	Combination of Coefficients Using Fusion Process	69
4.6	Optimization Result of Similarity Coefficients Fusions by Using GA Weights	79
<b>5</b>	<b>CONCLUSION</b>	<b>83</b>
5.1	Introduction	83
5.2	Recommendation	84
5.3	Project Advantages	85
5.4	Summary	85
	<b>REFERENCE</b>	<b>86</b>
	<b>APPENDICES</b>	<b>91</b>
	<b>Appendix A</b>	
	<b>Appendix B</b>	

**Appendix C**

**Appendix D**

**Appendix E**

**Appendix F**

**LIST OF TABLES**

<b>TABLE NO.</b>	<b>TITLE</b>	<b>PAGE</b>
2.1	2D versus 3D structure representation	21
2.2	Association coefficients (1-16), correlation coefficients (17-21), and distance coefficient (22) (Ellis et al., 1994). Each coefficient computes the similarity between two molecular	23
3.1	Topological indices descriptor generated using dragon software	49
3.2	The 13 groups of coefficients (from salim el at,2003), In this study the 13 groups of coefficients will be used	51
4.1	First Data and its Activities.	62
4.2	Second Data and its Activities.	62
4.3	Similarity Values of Different Coefficients	67
4.4	Similarity Values normalization with $Z\_score$	68
4.5	Similarity Values normalization with $X'$ between(0 and 1)	68
4.6	Summary of single coefficient of Average percentage top 10% of all Actives	68
4.7	The Average Percentage of all Actives (Second Data) Using Single Coefficient	69
4.8	Summary of Fusion of 2-Coefficients and the Average	71

	Percentage of top 10% for all actives	
4.9	Summary of Fusion of 3-Coefficients and the Average Percentage of top 10% for all actives	72
4.10	Shows Similarity Value After Order Descending And Ranking Positions	73
4.11	The Average Percentage of all Actives (Second Data) Using Single Coefficient and Ranking Positions	75
4.12	The Average Percentage of all Active top 10%	76
4.13	Summarization the Average Percentage of all Active top 10% depends on fusions of coefficients.	77
4.14	The Average Percentage of all Active top 10% on GA based fusions of coefficients (GA weights)	80

**LIST OF FIGURES**

<b>FIGURE NO.</b>	<b>TITLE</b>	<b>PAGE</b>
2.1	Connection table of molfile type, which generated by the XyM2Mol application. Where the top row of the connection tables gives brief pieces of information, each row in the middle part is concerned with the attributes of a nod.	12
2.2	The Tanimoto Similarity Measures	24
2.3	Classes of search techniques	34
2.4	GA components	35
2.5	One point crossover	39
2.6	Two point crossover	40
2.7	Cut and splice	40
2.8	Mutation	41
3.1	Flow chart of the Framework	44
3.2	Generate BCI bit string descriptor by using MAKEBITS software (a) BCI dictionary could generate and (b) Two dimensional Input Vector containing input data, where row represented as the molecular and column is compounds.	47
3.3	Show the mechanism to searching	50



3.4	Flow chart of basic genetic algorithm	54
3.5	Flow chart of optimizations of the fusion process	55
3.6	Chromosome Representation	57
3.7	Population of Chromosomes (random weights)	57
3.8	Single-point Crossover process	58
3.9	Inversing Mutation process	59
4.1	Similarity Measure program for obtaining the value of a, b, c, d and n.	64
4.2	Single Coefficient Analysis using program to Find Similarity	66
4.3	The percentage Average of top 10% Actives versus the single coefficients	76
4.4	The Average Percentage of Top 10% Actives versus the Fusions of Coefficients	78
4.5	The Percentage Average of Top 10% Actives versus the GA based fusions of coefficients (GA weights)	81

**LIST OF SYMBOLS**

$\Sigma$	-	Sum
$\mu$	-	Mean
$\sigma$	-	Standard Deviation
$X, Y$	-	Bit-string representations
$n$	-	Maximum length of chemical database
$a$	-	number of bits set in both $X$ and $Y$
$b$	-	number of bits set exclusively in $X$
$c$	-	number of bits set exclusively in $Y$
$d$	-	number of bits set in neither $X$ or $Y$
$x$	-	unnormalized similarity values
$x'$	-	normalized similarity values
$S$	-	Similarity value
$M$	-	Molecule
$Q$	-	Query(target)
$W$	-	Weights

**LIST OF ABBREVIATION**

1D	One Dimensional
2D	Two Dimensional
3D	Three Dimensional
BCI	Barnard Chemical Information
WLN	Wiswesser Line Notation
SMILES	Simplified Molecular Input Line Entry System
MDL	MDL Molecular Design Limited
MDDR	MDL Drug Database Report
SDF	Structure Data Format
GA	Genetic Algorithm
MCS	Maximal common substructure

## LIST OF TERMINOLOGIES

- Chemical Database - Chemical database is a database specifically designed to store chemical compound attributed data.
- Molecule - The simplest structural unit of a substance that retains the properties of the substance, and is composed of one or more atoms.
- Molecular Descriptors - Vector of number based on predefined attributes. It is used to represent a chemical structure.
- Similarity Coefficient - Measure used to quantify the degree of structural resemblance between target structure and each structure in the database.
- Similarity Searching - Standard tool for drug discovery .A chemical compound retrieval system that requires specification of an entire target structure, which is compared with the corresponding sets of features for each of the database structures. The database molecules are then sorted in decreasing order of similarity with the target.
- Physicochemical properties - It is also known as global molecular properties, which is also used as a molecular descriptor. Examples of these properties are mol refractivity, molecular weight, principal axes etc.

**LIST OF APPENDIX**

<b>APPENDIX</b>	<b>TITLE</b>	<b>PAGE</b>
A	Details on MDDR data activity	91
B	Percentage of actives using different single coefficient	93
C	Percentage average of actives using different query single coefficients second dataset	98
D	Percentage of actives using different fusion coefficients	102
E	Percentage average of actives using different 10 queries for single coefficients	110
F	Percentage Of Actives Fusions Of Coefficient Non- Weights (10 Queries) Combination Compared With GA Weights Combination For Each Active Fusion	118

## CHAPTER 1

### INTRODUCTION

#### 1.1 Introduction

Chemical database is a database specifically designed to store chemical compound attributed data. The process of information retrieval is commonly used to retrieve chemical compounds. A filtering retrieval process called Data Fusion process has been recently used to combine compound results from multiple chemical data resources. Similarity measures were used as tools, in chemical database such as retrieval, clustering, diversity analysis, which has two main components of molecular representation and similarity coefficients.

Most chemical databases store information on stable molecules. Chemical structures are traditionally represented using lines indicating chemical bonds between atoms and drawn on paper (2D structural formulae). While these are ideal visual representations for the chemists, they are unsuitable for computational use and, especially, for search and storage. Large chemical databases are expected to handle the storage and searching of information on millions of molecules taking terabytes of physical memory.

In most cases, only molecular representation and similarity coefficient can be used in chemical data retrieval. However, combination of multiple similarity measures and data fusion can produce better results.

Genetic algorithms are searching technique used in computing to find true or approximate solutions to optimization and search problems. Genetic algorithms are categorized as heuristics. They are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover (also called recombination). Furthermore, genetic algorithms are implemented as a computer simulation in which a population of abstract representations (chromosomes, genotype or the genome) of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible.

Similarity searches are now a standard tool for drug discovery. The idea behind such searches is that, given a compound with an interesting biological activity is compared to other compounds. Compounds that are “similar” to it in structure are likely to have a similar activity. In practice, an investigator provides a chemical structure as a “probe”, searches over a database of sample-available compounds, and finds those that are most similar, which are then submitted for testing. Similarity searching can be done on the basis of 2D or 3D structure. 2D similarity searches, especially those based on comparing lists of pre-computed substructure descriptors, are computationally inexpensive.

Similarity measures are a converse of the distance function. Similarity functions take pair of points where they return the large similarity value for the nearby points, and the small similarity value for the distant points. One way to transform between a distance function and a similarity measure is to take the reciprocal. Such transformation is the standard method for transforming between resistance and conductance in physics and electronics. Similarity measures use two basic tools of molecular representation and similarity coefficient to quantify the similarity between the representations of two molecules.

Data fusion is an approach where data, evidence, or decisions coming from or based on multiple sources about the same set of objects, are integrated to increase the

quality of decision making under uncertainty about the objects. It exists in nature where living things combine information from multiple resources to create a reliable recognition of their surroundings. Fusion has been used for various purposes like detection, tracking and decision making. It has been applied in areas like military, robotics, medicine and information retrieval. Fusions can improve confidence in decisions due to the use of complementary information. The use of data fusion can also improve performance if, for example, a sensor were to become damaged or ineffective there would still be information coming in from the other sensors. Data fusion leads to extended coverage since if there is more than one sensor they can cover disparate areas, times and qualities.

A weighting scheme is used to differentiate between different features in a molecule, based on how important they are in determining the similarity of that molecule with another molecule.

Genetic Algorithm is used to find the best linear combination of weights assigned to the scores of different matching functions. It is found that his GA based system outperforms any of the individual expert matching functions on the performance measures. The system also outperforms the best of the individual expert matching functions.

## **1.2 Problem background**

In process of chemical compound information retrieval, much data fusion efforts have been made to combine results from multiple similarities searching systems. In similarities searching, a query involves the specification of the entire structure of molecules. This specification is in the form of one or more structural descriptors and is compared with the corresponding set of descriptors for each molecule in the database. A measure of similarity is then calculated between the target structure and every database structure.



Molecule graph representations can also be used for the representation and searching of databases of 3D structures (Martin and Willett, 1998). The pharmaceutical industry makes very extensive use of highly sophisticated systems for the storage, retrieval and processing of information describing the chemical structures of molecules.

The similarity calculations between molecules have been used not only in similarity searching, but also in applications like compounds selection and molecular diversity analysis. The results of the similarity measure were then used to sort the database structures into the order of decreasing similarity with the target.

One of the ways to improve the performances of molecular similarity searching is the combining of the result of different similarity measures; which, known as data fusion process. How to optimize this combined result has become an interesting research area in chemoinformatics.

Several methods have been used to further optimize the measure of similarity between molecules. These methods include weighting and data fusion. A weighting scheme is used to differentiate between different features in a molecule, based on how important they are in determining the similarity of that molecule with another molecule.

In recent work by Salim (2002) it is shown that fusion does give improvement over the use of single coefficients. Ginn (1997) has found that the use of data fusion on two types of ranking resulted in combined rankings that contained very different sets of nearest neighbors and often performed better in simulated property prediction than did the individual measures.

### 1.3 Problem statement

Data fusion has been applied in chemical compound retrieval, in order to combine results from multiple similarities searching system. However, one problem associated with the data fusion approach is how to optimally combine the results obtained from various retrieval systems since there is no known guideline on the best fusion model that works for all type of data and activity. A Genetic Algorithm (GA)-based approach might be employed to find the best linear combination of weights assigned to the scores of different retrieval system to get the most optimal retrieval performance.

Therefore, the focus of this study is to get the most optimal weights to combine the similarity measures in order to discover different similarity measures with deferent characteristic. That is well suited for various activities, database and the type of molecular. To optimize better result, and this study will also apply GAs which is a search technique used in computing to find true or approximate solution to optimization and search problem.

This study is aimed to optimize weights of different similarity measures in data fusion by applying of genetic algorithm (GA) in a chemical database.

### 1.4 Objectives

The objectives of this study are summarized as follows:

- a) To retrieve compounds from chemical databases using different similarity measures based on different similarity coefficient and molecular representations.
- b) To apply a Genetic Algorithm based data fusion for optimizing combination of similarity measures for chemical database retrieval.

- c) To investigate whether the optimized fusion process does give improvement in chemical data retrieval process based on the coefficients used.

## 1.5 Project Scope

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

1. The databases aimed to be used in this study are only limited to chemical data from MDDR and ID databases will be used.
2. Applying fusion of only similarity coefficients and representations.
3. Only compound data representation of BCI-based and topological indices will be used.

## 1.6 Justifications for Employing GA

Genetic algorithms are well suited to with problems involving chemical database due to their adaptability and their effectiveness at searching large spaces. The reason for genetic algorithms success at a wide and ever growing range of optimization problems is a combination of power and flexibility. The power derives from the empirically proven ability of evolutionary algorithms to efficiently find globally competitive optima in large and complex search spaces. The favorable scaling of evolutionary algorithms as a function of the dimension of the search space makes them particularly effective in comparison with other search algorithms for the large search spaces typical of real world scheduling. Besides the flexibility, genetic algorithms effectively multiple facets.

Many optimization problems in parallel As pointed by Goldberg (1989) these are some of the differences between GA and other optimization and search methods, which make GA much favorable to be implemented in chemical database searching :

1. GAs work with a coding of the parameter set, not the parameter themselves.
2. GAs search from a population of points, not a single point.
3. GAs use objective function information, not derivatives or other auxiliary knowledge.
4. GAs use probabilistic transition rules, not deterministic rules.

In this project we use GA for generated deferent weights (in the range of 0.0 to 1.0) to each similarity searching and combined the similarity value normalized to get the combination

## **1.7 Report Organization**

This report is mainly divided into five chapters. The first chapter provides an introduction and brief overview of the project including the problem background, problem statement, objective, scope, and justifications for employing GA. Chapter 2 reviews the literature chemical database chemical compounds and retrieval . This includes the background knowledge on the terms that are involved in the project mainly on similarity searching, data fusion and genetic algorithms. Chapter 3 covers the methodology of the research, the techniques that are involved are discussed which are data fusion using genetic algorithm. The hardware and software requirements for this project are also discussed in this section. Chapter 4 presents the results from applying data fusion and genetic algorithm, findings and discusses in this project. Chapter 5 is the conclusion of the project based on the four previous chapters that has been discussed. There are also recommendation and advantage of this project.

## CHAPTER 2

### LITERATURE REVIEW

#### 2.1 Chemical Database

Chemical database is specifically designed to store chemical information of stable molecules. Chemical structures are traditionally represented using lines indicating chemical bonds between atoms and drawn on paper (2D structural formulae). While these are ideal visual representations for the chemist, they are unsuitable for computational usage and especially for search and storage (Wikipedia, 2007).

##### 2.1.1 Chemical compounds

The chemical compound is a chemical substance formed by chemically combining two or more elements, with a fixed ratio determining the composition. A compound can consist either of atoms covalently bonded together (i.e. molecules, such as water) or of ions bonded together by the attraction of their opposing charges (eg sodium chloride). Compounds may have a number of possible phases. All compounds can exist as solids. Molecular compounds may also exist as liquids, gases or plasma. All compounds decompose to smaller compounds or individual atoms if heated to a certain temperature (called the decomposition temperature).

For example, water ( $\text{H}_2\text{O}$ ) is a compound consisting of two hydrogen atoms for every oxygen atom. A defining characteristic of a compound is that it has a chemical formula. Formulas describe the ratio of atoms in a substance and the number of atoms in a single molecule of a substance. The formula does not indicate that a compound is composed of molecules; for example, sodium chloride ( $\text{NaCl}$ ) is an ionic compound.

Compounds may have a number of possible phases. All compounds can exist as solids, liquids or gases. All compounds may decompose to smaller compounds or individual atoms if heated to a certain temperature called the decomposition temperature.

The atoms within a compound can be held together by a variety of interactions, ranging from covalent bonds to electrostatic forces in ionic bonds. A continuum of bond polarities exist between the purely covalent bond (as in  $\text{H}_2$ ) and ionic bonds. For example  $\text{H}_2\text{O}$  is held together by polar covalent bonds. Sodium chloride is an example of an ionic compound.

### **2.1.2 Representation of Chemical Structures**

Many chemistry and drug related organizations have their publicly accessible and proprietary databases of chemical compounds, containing large number of molecules; several hundred thousand is a common figure and some have even billions of compounds. Many have virtual libraries of compounds generated using computational techniques that can be converted to chemicals using combinatorial chemistry techniques.

Although there are 2-dimensional and 3-dimensional representation techniques available for chemical molecules, only the 2-D representation techniques are more popular. The 2D chemical structure diagrams are not suitable for storage and retrieval in a computerised chemical database system. A different kind of

representations has been used for storing chemical structures. These representations model the entire molecule by listing every atom that makes up the molecule. Four types of 2-D representation techniques have been used extensively in chemical information systems. These techniques are the systematic nomenclature, fragmentation codes, line notations and connection tables.

The systematic nomenclatures primarily used in manual information retrieval systems; however, due to the lack of flexibility nomenclatures often require translation automatically to another type of representation if it is to be used in computerized systems (Stouw and Elliott, 1974, Willet, 1980).

#### **2.1.2.1 Fragmentation Codes**

Fragmentation codes were the first to be used as structural representation in chemical retrieval systems such as the structure and substructure searching, and are still in use. A fragment code is a set of pre defined substructure attributes, where the presence or absence of which is used as characterization of a molecule.

A fragment code representation has several problems due to its subjective nature. The system is yet to be standardized and every organization has its own specific code order for its own database of compounds. The coding of new compounds is normally a manual task and sometimes all the molecules need to be recoded if the code changed because of the addition of a new compound (Craig, 1969). Fragmentation codes also result in ambiguous representation for the molecule as the set of codes assigned to a molecule might interconnect in a different number of ways.

### **2.1.2.2 Linear Notations**

It represents a structure by string of alphanumeric characters. Each character represents an atom, a bond or a small group of atoms and bonds. It is very short as compared to the connection table; therefore, it is well suited for storing and transmission of molecules. The inter-connections of the atoms and groups are not stated explicitly, but are implicit in the ordering of the symbols within the string. The meaning of the symbols may be context dependent, and thus quite detailed analysis may be necessary to determine the exact form of the structure represented by the notations (Willett, 1987). Winner in 1988 has developed a large number of notational schemes before the simplified molecular input line entry specification (SMILES). The SMILES notation has gained a wide spread acceptance because of its easiness to use and is more comprehensive than the wiswesser line notation (WLN) ,which had been in use for more than three decades since its development in 1954.

### **2.1.2.3 Connection Tables**

Connection tables list all of the atoms present in a molecule together with details of how each atom is connected to its neighbors. The atoms are numbered, with the atom types usually being represented by their atom symbols and the bond types are indicated by code. The present connection tables are the primary means of representation for the chemical structures in both public and in house chemical information systems.



molfile	Number of		The first atom is																								
	Bonds		carbon																								
Number of Atoms	9	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.0000	3.0000	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1.2990	2.2500	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1.2990	0.7500	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.0000	0.0000	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	-1.2990	0.7500	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	-1.2990	2.2500	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.0000	4.5000	0.0000	F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2.5981	3.0000	0.0000	Cl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.0000	-1.5000	0.0000	OH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	2	1	0	0	0																					
	2	3	2	0	0	0																					
	3	4	1	0	0	0																					
	4	5	2	0	0	0																					
	5	6	1	0	0	0																					
	6	1	2	0	0	0																					
	1	7	1	0	0	0																					
	2	8	1	0	0	0																					
	4	9	1	0	0	0																					
	M	END																									

**Figure 2.1** Connection table of molfile type, which is generated by the XyM2Mol application. Where the top row of the connection table gives brief pieces of information, each row in the middle part is concerned with the attributes of a node. The bottom part shows data of bonds (Kei ITO,TANAKA and FUJITA,2005).

A connection table contains all necessary information required for plotting its 2-D graph by any program like SMILES. The simplest connection table consists of at least two sections: first, a list of the atomic numbers of the atoms; second, a list of the pair wise atomic bonds in a molecule. More Tables that are sophisticated contain the bonding angles information for plotting the bonds. Since connection tables contain an explicit representation of the inter-connections between the atoms in a molecule, they are particularly well suited to manipulations involving such topological information as atom-by-atom searching, graphical structure input and display, structure property correlation and reaction indexing. There are a large number of standardized file formats based on connection tables such as the Molecular Design Limited's (MDL), MOL and Structure Data Format (SDF) formats.

Connection table representation can be expanded to 3D structure representation when the related information is available, either through experiments (Allen et al., 1991) or through calculations (Ricketts et al., 1993). There must be some indication of the distances between all atoms, and not just the bonds between

them as in 2D structure representations. This can be done by storing the inter-atomic distances within a molecule. 3D structures can have various conformations, each of which may represent a different shape.

The connection tables have proved to be the most flexible and generally useful representation due to which it forms the basis of most present day computer libraries and systems.

## **2.2 Retrieval of Data from Chemical Database**

Retrieval of data from chemical structure search system can offer three principle types of searching facility of retrieval mechanisms these are structure, substructure and similarity searching. This expands the capabilities of the existing systems by capitalizing on the strengths of relational database technology. The system allows the user to optimally store and search chemical structure information including information related to multi-valued atoms and multi-typed bonds.

### **2.2.1 Structure Searching**

Structure searching is the first principle of the retrieval mechanisms. Structure search involves searching a molecule database for a specified query molecule. The user to the database to search for a compound that matches perfectly with the target structure supplies the complete structure of a molecule. This type of search is use to get some data about a particular compound. Another use to this type of search is during registration process, for new molecule; a structure file is used, in which a single and unique record of each compound is maintained and known as a registry file (Ash et al., 1985).

In the connection tables, the equivalence of two structures can be demonstrated by generating all possible numbering of atoms in a query and then comparing the resulting connection tables with those stored in the database using a graph isomorphism algorithm. A chemical structure has treated as a graph, where there is a finite set of vertices (atoms) and edges (bonds) (Tarjan, 1977). This approach takes a lot of time due to the number of different tables that can be constructed for a compound, which is  $N!$  For an  $N$ -atom molecule. A scheme called the Morgan algorithm can be used to produce a unique numbering of the set of atoms in a connection table (Morgan, 1965).

### 2.2.2 Substructure Searching

Substructure searching is the second principal retrieval mechanisms. It refers to the capability to use a structure as a search term and locate chemical structures containing that structural skeleton. They involve the retrieval of molecules from the database that contain a user-defined query substructure. Substructure search is especially useful for finding structures containing a specified functional group, thus allowing the properties common to that group to be observed.

Substructure search identifies all the molecules in the database that contain a specified substructure. Another use of substructure search is in the implementation of pharmacophoric pattern searching, where compounds containing a specific 3D substructure were identified in a molecular modelling study. The subgraph isomorphism of graph-matching operation is an NP-complete problem, in which no polynomial time algorithm is known (Garey and Johnson, 1977).

Even if heuristics of various sorts are adopted to rapidly reject non-mappings, the operation is still very time-consuming and thus most research on substructural retrieval has focussed upon the development of efficient and effective screening methodologies. Furthermore, there are types of substructure searching system which do not employ this two-stage searching technique (Barnard, 1993).

According to Willett et al. (1998) Substructure searching has a few limitations. The first limitation is that the user has to specify the structural constraints required in the molecules that are retrieved. Second limitation is that the user normally cannot control the size of the output produced and there is no direct mechanism to rank the output in the order of decreases the similarity. These characteristics have led to the development of another access mechanism known as similarity searching.

### 2.2.3 Similarity Searching

Similarity is a degree of symmetry in either analogy or resemblance between two or more concepts or objects. The notion of similarity rests either on exact or approximate repetitions of patterns in the compared items.

Similarity searching in chemical database was first introduced in the mid-1980s by (Carhart et al.1985; Willett et al.1986). Similarity searching offers complementary alternative to substructure and 3D pharmacophore searching. A query compound is used to search a database to find those compounds that are most similar to it; this involves comparing the query with every compound in the database. The database is then sorted in order of decrease similarity to query. A measure of the similarity is then calculated between the target structure and every database structure.

Similarity measures quantify the relatedness of two molecules with a large number if their molecular descriptions are closely related and with a small number (large negative or zero) when their molecular descriptions are unrelated. There are many measures available to quantify the degree of similarity between a pair of molecules. The computational requirements of these measures vary depending on the level of detail used to represent the molecules that are being compared.

The maximal common substructure (MCS) is the largest set of atoms or bonds from the target structure that can be superimposed exactly onto another

structure, and is identified by using a maximal common subgraph isomorphism algorithm. Garey and Johnson 1977 due to its NP-complete computational requirement, MCS algorithms have not been widely used for similarity searching to date.

Molecular similarity can be defined in many ways depending on the information used to represent the molecules and the measures employed to quantify the degree of similarity between two molecules (Johnson and Maggiora, 1990; Dean, 1999).

Similarity searching offers several advantages. First, there is no need to define a precise substrate, 3D substructure or pharmacophore query since a single active compound is sufficient to initiate a search. Second, the user has control over the size of output as every compound in the database is given a numerical score that can be used to generate a complete ranking. Finally, similarity searching facilitates iterative approach to searching chemical database since the top-scoring compounds resulting from one search can be used as queries in subsequent similarity search.

Molecular similarity measure has two principal components: (i) the structural representation, used to characterize the molecules, and (ii) the similarity coefficient, used to compute the degree of resemblance between pairs of such representations (Willett, 2003).

### **2.3 Molecular Descriptors for Similarity Searching and Similarity Coefficient**

The manipulation and analysis of chemical structural information is made possible through the use of molecular descriptors, these are numerical values that characterise properties of molecules each of which is based on some pre-defined attributes. Many different molecules have been described and used for wide variety

of purposes. The machine-readable structure was generation representation like a 2D connection table or a set of experimental or calculated 3D co-ordinates.

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 experiments.

### **2.3.1 Molecular Representation**

The Molecular descriptors can be divided into 1D, 2D (such as topological indices and 2D fingerprint) and 3D descriptors (such as pharmacophore keys).

#### **I. 1D descriptors**

1D descriptors are aspects of molecules and they are physicochemical properties. Several similarity measures make use of the global physicochemical properties of the whole molecules. Examples of these types of properties were investigated by CAS and molar refractivity workers (Fisanick et al., 1992). The whole-molecule properties include principal moments of inertia, principal axes, volume of the inertia ellipsoid and molecular weights. Some of these properties can take a lot of time to calculate, most obviously properties that are calculated using quantum mechanics packages. Molecular properties are sometimes combined with structural descriptors as will be described later, to characterise molecules.

#### **II. 2D descriptors**

These types of descriptors are based on information derived from the traditional 2D structure diagrams. Examples of 2D descriptors are the topological indices and 2D screens. The **2D screens** were initially developed for substructure

search systems in which bit strings are used to represent molecules. There are two types of 2D screens: dictionary-based bit strings and hashed fingerprints.

In the **dictionary-based** bit strings, a molecule is split up into fragments of specific functional groups or substructures. The fragments (structural keys) used are recorded in a predefined fragment dictionary that specifies the corresponding bit positions (screen number) of the fragments in the bit string. Bits, either individually or as a group, represent the absence or presence of fragments. Substructural fragment descriptors can involve atoms, bonds and rings. Examples of these types of fragments are augmented atoms, atom sequences, ring sequence, ring fusion sequence, atom pair and topological torsion (Dittmar et al., 1983; Carhart et al., 1985; Nilakantan et al., 1987).

In **hashed fingerprints**, all the unique fragments that exist in a molecule are hashed using some hashing function to fit into the length of the bit string. This approach allows for more generalisations because it does not depend on a predefined list of structural fragments. The fingerprints generated are characterised by the nature of the chemical structures in the database rather than by the fragments in some predefined list. Instead of using a fragment dictionary, this method defines a set of patterns to index.

**Topology indices** are single-value descriptors that can be calculated from the 2D graph representation of molecules. The Topological indices characterise the bonding pattern of a molecule by a single value integer or real number, obtained from mathematical algorithms applied to the chemical graph representation of the molecules. Each index, thus, contains information not about fragments or some locations on the molecule, but rather about the molecule as a whole. Simpler descriptors include the number of atoms and bonds and the number of rotatable bonds.

They are various topological indices like the molecular connectivity indices such as topological state indices, electrotopological state indices, hydrogen

electrotopological state indices, atom type electrotopological state indices, hydrogen bonding descriptor indices, bond type electrotopological state indices, the difference chi indices, the delta chi indices and vertex eccentricity (Hall and Kellogg, 1999). The topological relationship is based on the graph distance to each other atom. The electronic aspect is based on an intrinsic state plus perturbation due to intrinsic state differences between atoms in the molecule.

Molecular connectivity or chi indices quantify molecular structure by counts of substructure fragments like branching, heteroatom content and cyclicity. The topological state indices are numerical values that encode information about the topological environment of an atom, based on the encoding of atom information in all paths emanating from that atom. The electrotopological state indices are numerical values that encode information about both the topological environment of each atom in a molecule and the electronic interactions due to all other atoms in that molecule.

### **III. 3D descriptors**

3D descriptors are modeling environment of molecules. They have the ability to model the biological activity of molecules because the binding of a molecule to a receptor site is a 3D event. Generating a 3D structure, handling conformational flexibility and deciding which conformers to include can all make 3D descriptors computationally more expensive than 2D descriptors. Examples of 3D descriptors are 3D screens, potential-pharmacophore-point descriptors, affinity fingerprints, 3D atom environment for use in atom mapping similarity searching and 3D molecular fields for use in field-based similarity searching. Shape descriptors, like surface area and volume, also use the 3D shape of the whole molecule.

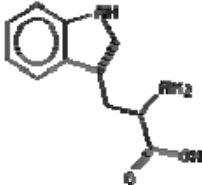

3D screens were initially designed for 3D substructure searching. The screening methods used in these systems encode spatial relationships, most usually distances and/or angles, between features in a molecule such as atoms, ring centroids and planes. Examples of the distance-based descriptors are distance distributions that are the count of distance ranges in a molecule, individual-distances descriptor



which encodes inter-atomic distances between pairs of elemental types, and a descriptor based on the sum of the squared distances between each triplet of atoms (Willett et al., 1998).

In chemistry, 2D structural representations are still extensively in use. Table 2.1 below shows some relative merits and problems of 2D and 3D structural representations.

**Table 2.1 : 2D versus 3D structure representation**

2D		3D	
	<p><b>Pros:</b></p> <ul style="list-style-type: none"> <li>• Show complete structure</li> <li>• Easy recognition of patterns</li> <li>• Chemists know how good structures look like</li> </ul> <p><b>Cons:</b></p> <ul style="list-style-type: none"> <li>• Removes too much information from the real structure</li> <li>• Make impossible spatial matching of structures</li> </ul>		<p><b>Pros:</b></p> <ul style="list-style-type: none"> <li>• All available structural information are present</li> <li>• Understand shapes</li> <li>• See what would be hidden in a 2D view</li> </ul> <p><b>Cons:</b></p> <ul style="list-style-type: none"> <li>• Limited to viewing part of structure</li> <li>• Unsuitable for quick comparisons</li> <li>• Needs interaction to avoid ambiguities</li> </ul>

### 2.3.2 Similarity Coefficients

Similarity coefficients are used to obtain a numeric quantification to the degree of similarity between a pair of structures (Willett, 1990). There are many types of similarity measures that are in use. As an example, edit distance, which is a string-based measure of the number of operations to transform the representation of a structure to the representation of another structure, has been used to measure the similarity between two 3D molecular structures (Wang and Wang, 2001).

There are four main types of similarity coefficients: distance coefficients, association coefficients, correlation coefficients and probabilistic coefficients (Sneath and Sokal, 1973; Willett, 1987; Ellis et al., 1994) table 2.2 shows some similarity coefficients.

Distance coefficients are used to measure the distance between structures in a molecular space. Association coefficients are pair-functions that can measure the agreement between the binary, multi-state or continuous character representations of two molecules (Sneath and Sokal, 1973). Some association coefficients that have been used to measure the similarity between compounds. Correlation coefficients are generally used to measure the degree of correlation between sets of values repercentage the molecules, like the proportionality and independence between pairs of real-valued molecular descriptors. Probabilistic coefficients, whilst not much used in measuring molecular similarity, focus on the distribution of the frequencies of descriptors over the members of a data set, giving more importance to a match on an infrequently occurring variable.

**Table 2.2** : Association coefficients (1-16), correlation coefficients (17-21), and distance coefficient (22) (Ellis et al., 1994). Each coefficient computes the similarity between two molecular fingerprints,  $X$  and  $Y$ , of length  $n$ , in which  $a$  is the number of bits set in both  $X$  and  $Y$ ,  $b$  is the number of bits set exclusively in  $X$ ,  $c$  is the number of bits set exclusively in  $Y$  and  $d$  is the number of bits set in neither  $X$  or  $Y$ .

No	Coefficient	Formula	No	Coefficient	Formula
1	Jaccard/Tanimoto	$\frac{a}{a+b+c}$	12	Ochiai/Cosine	$\frac{a}{\sqrt{(a+b)(a+c)}}$
2	Dice	$\frac{2a}{2a+b+c}$	13	Kulczynski(2)	$\frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)}$
3	Russell/Rao	$\frac{a}{n}$	14	Forbes	$\frac{n \times a}{(a+b)(a+c)}$
4	Sokal/Sneath (1)	$\frac{a}{a+2b+2c}$	15	Fossum	$\frac{n \left( a - \frac{1}{2} \right)^2}{(a+b)(a+c)}$
5	Kulczynski(1)	$\frac{a}{b+c}$	16	Simpson	$\frac{a}{\min(a+b, a+c)}$
6	Simple Matching	$\frac{a+d}{n}$	17	Pearson	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
7	Hamann	$\frac{a+d-b-c}{n}$	18	Yule	$\frac{ad-bc}{ad+bc}$
8	Sokal/Sneath(2)	$\frac{2a+2d}{a+d+n}$	19	McCon-naughey	$\frac{a^2-bc}{(a+b)(a+c)}$
9	Rogers/Tanimoto	$\frac{a+d}{b+c+n}$	20	Stiles	$\log_{10} \frac{n \left(  ad-bc  - \frac{n}{2} \right)^2}{(a+b)(a+c)(b+d)(c+d)}$
10	Sokal/Sneath(3)	$\frac{a+d}{b+c}$	21	Dennis	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$
11	Baroni-Urbani/Buser	$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	22	Mean Manhattan	$\frac{b+c}{n}$

Similarity coefficients were examined in order to group together those with comparable performance when applied to searches of binary representations. Example coefficients from each of the groups were then used to determine whether the performance of single coefficients could be improved by combining coefficient using data fusion.

For example, a fingerprint is a binary sequence of Boolean values where 1 indicates a presence of a feature and a 0 indicates absence. Feature examples are “200 < molecular weight <= 250” or “aromatic ring count = 1”. When comparing these fingerprints similarity scores are used. A much-applied method is the Tanimoto similarity (Figure 2.2). Which for molecule A and B is defined as  $S_{AB} = a / (a + b + c)$  where  $a$  bit are set common to both strings,  $b$  bits are set in the comparison string,  $c$  bit are set in the reference-structure string, and  $d$  bits are set in neither string.

<b>A</b>	1	0	1	1	1	0	1	1	0	0	1	1	<b>b = 3</b>
<b>B</b>	0	0	1	1	0	0	1	0	1	0	1	1	<b>c = 1</b>

$S_{AB} = 5 / (5 + 3 + 1) = \mathbf{0.56}$

**Figure 2.2** The Tanimoto Similarity Measures

## 2.4 Data fusion

Data fusion is the process of combining inputs from several similarity measures with information from other similarity measures comprising information processing blocks, databases or knowledge bases, into one representational format. The data fusion can give an overall estimate of the similarity according to the characteristics mentioned.

The process of data fusion involves computing several types of similarity measures, and combining the results using one of several fusion rules. The combined scores output by the fusion rule are then used to re-order the compounds to give the final ranked output. Holliday et al. (2002) have concluded that data fusion results in an increase in search effectiveness. In some cases where the use of a fusion rule results in the assignment of the same score to two or more items, a further sort key is specified for the tied compounds. An example would be to sort the canonicalised connection tables of the tied compounds alphabetically. Weights can also be allocated to individual rankings based on some statistical observations of the coefficients' historical performances.

The fusion process is often categorized as low, intermediate or high-level fusion, depending on the processing stage at which fusion takes place (Ng and Kantor, 1998). At the low or primary level, fusion (also called **data fusion**), all the raw data available to the detecting systems were considered together in the fusion to produce new raw data that is an overall estimate and expected to be more informative than the inputs. For example, in image processing, images from several spectral bands of the same scene are fused to produce a new image that ideally contains a single image comprising all information available in the various spectral bands.

At the intermediate or attribute level fusion, (also called feature level fusion), primary signals from the detecting systems are processed to produce a set of specific attributes, and decisions about the objects are made according to an optimal decision rule based on all such attributes. An example of this would be using several different features of extraction methods on image data to get a set of features of the object. Relevant features of the object to be detected can then be obtained from this set of features.

In high or decision level fusion, each detecting system individually makes its own partial decision about the objects using its own data and according to its own criteria. A final decision combining all these partial decisions is then made. When the partial decisions are in the form of a confidence or score, the fusion is called hard fusion. If the partial decision is in the form of a decision, the fusion is called soft

fusion. Methods of decision fusion include voting methods, statistical methods and fuzzy logic based methods.

Fusion usually follows a similar set of procedures. These are the collection of primary observable data from each source, performing some preliminary filtering of the data, scoring or ranking the data against some ideals. The processing the data so that it is in the form suitable for fusion with other data, weighting the data based on its importance and then combining the different sources of data using a fusion scheme (Hall, 1992).

Fusion may be useful for several objectives such as detection, recognition, identification, tracking, change detection, decision-making, etc. These objectives may be encountered in many application domains as Defense, Robotics, Medicine, Space, etc.

Several data fusion algorithms have been developed and applied, individually and in combination, providing users with various levels of informational detail. The U.S. Defense Department's Joint Directorate of Laboratories Data Fusion Sub panel has developed three basic categories or levels of data fusion (Linn and Hall, 1991). These fusion levels are differentiated according to the amount of information they provide. The most basic level involves the fusion of multi-sensor data to determine the position, velocity and identity of a target. At this level, however, only raw, uncorrelated data are provided to the user. In comparison, level two data fusion provides a higher level of inference and delivers additional interpretive meaning suggested from the raw data. Level three data fusion is designed to make assessments and provide recommendations to the user, much as occurs in knowledge-based expert systems (KBES). Thus, each jump between data fusion levels represents a corresponding leap in technological complexity to produce increasingly valuable informational detail.

Using an efficient fusion scheme significant advantages are expected as given in the points bellow:

- Improved confidence in decisions due to the use of complementary information.
- Improved performance to countermeasures.
- Improved performance in adverse environmental conditions. Typically smoke or fog cause bad visible contrast and some weather conditions (rain) cause low thermal contrast (Infra Red imaging), combining both types of sensors should give better overall performance.

The fusion procedures are categorized by their input/output characteristics in five categories as proposed by Dasarathy (1994) they are, Data in-Data out, Data in-Feature out, Feature in-Feature out, Feature in-Decision out, Decision in-Decision out.

Fusion methods may also be categorized by the extent to which they make use of learning from examples. A part from what is learned; knowledge about the optimum fusion is fed into the system either explicitly or implicitly. Implicit feeding, however, needs some expert knowledge that defines the optimal fusion as an and/or combination. A scheme often used that allows the introduction of various and complex expert knowledge is based on the fuzzy logic theory. On the other hand neural networks are often presented as good candidates for learning from examples. The best fusion scheme should include both knowledge sources.

## **2.5 Data Fusion in Chemical Compound**

There is a high possibility that there does not exist a single, best, measure of molecular similarity that can uniquely represent biological similarity of molecules. This possibility has led to the consideration of combining several similarity measures with the expectation that a more descriptive measure of biological similarity will result compared to when only a single measure is used. Many of the fusion ideas applied to molecular similarities originate from the idea of combining several

independent information retrieval system components to get a performance over and above that of the best individual component (Willett, 2000). Combining several similarity measures may be desirable because different measures may make use of different sources of evidence. A particular similarity measure may return active molecules that another does not, or it may give a better estimation of the biological similarity of a particular type of molecule. Fusing different measures could result in a value that is a more comprehensive measure of similarity.

### **2.5.1 Fusions of Molecular Representation**

There are many studies that involve the combination of similarity measures based on different molecular descriptors including the study conducted by Kearsley et al. (1996). They performed some linear combinations of pairs of similarity scores generated by different 2D descriptors such as atom pairs, topological torsions and their binding property versions. Two combination functions were used. The first combination function took the mean of the similarity values of the descriptors as the combined similarity scores. The second combination took the minimum between the two ranks obtained by a compound using two descriptors as the new rank.

Matter (1997) have showed that combining 2D fingerprint with other descriptors did not increase the coverage of actives in a cluster-based selection, where as the performance of 3D descriptors can be improved if they are combined with a valid 2D descriptors.

On other studies involving fusion of similarity rankings based on different molecular descriptors have also been conducted by Ginn et al. (1997 and 2000). They have followed several steps where the fusion rules originated from the rules used by an earlier study in text retrieval by Fox and Shaw(1994) have been implemented . These rules are MAX- maximum (individual rankings), SUM- sum (individual rankings), MIN- minimum (individual rankings), and SUMN- sum



(individual rankings) / count (rankings less than  $n$  nearest neighbours), where  $n$  is the nearest neighbours of interest (Ginn et al, 1997; Ginn et al., 2000).

### **2.5.2 Fusions of Several Molecules in a Single Query**

The efforts of the fusions of queries in text retrieval are to combine more than one molecule in a single query. A modal fingerprint, which constructed from the common bits found in the molecular fingerprints of an input dataset, been used to search different databases (Shemetulskis et al., 1996). Some user-defined thresholds were commonly determined. It has been found that potentially interesting compounds, that could not be found by direct similarity searching against structures closest in representation to the features contained in the modal fingerprint, can be extracted from a commercial database using this kind of query. Combined chemical target has also been used in an iterative similarity searching using an approach analogous to relevance feedback in the text retrieval area .

Another effort made by Nachbar(2000) in combining the molecular descriptors of several molecules to construct a hybrid molecule for use as a target in similarity searching has also resulted in a diverse set of structurally reasonable molecules .

The descriptor of the joint chemical target was averaging from the descriptors of its member. The joint target was used again to rank the compounds in the database. It was shown that the use of the joint targets in addition to the use of single targets could significantly enhance the retrieval of active compounds.

### **2.5.3 Fusions of Similarity Coefficients**

Many different types of similarity coefficient have been described previously and most of them can be grouped into three broad classes: distance coefficients,

association coefficients and correlation coefficients. Distance coefficients quantify the degree of difference between two objects and have been extensively used in many applications of multivariate statistics, probably due to the simple geometric interpretation that is attached to many of them. With a distance coefficient, the greater the degree of similarity between two objects the smaller the value of the coefficient (and vice versa). Association coefficients, conversely, are most commonly used with binary data (i.e., variables denoting the presence or absence of descriptors in an object) and are often normalised to lie within the range of zero (no similarity at all) and unity (identical sets of descriptors). That said, association coefficients can be used with non-binary data, in which case other ranges of values may apply (e.g., the lowerbound of the well-known Tanimoto coefficient is  $-1/3$  when used with such data). Finally, correlation coefficients measure the degree of correlation between the sets of values characterising each of a pair of objects (rather than their more conventional use in multivariate analyses to probe the relationships between pairs of variables). Group fusion involves combining the results of similarity searches based on multiple reference structures and a single similarity measure (Willett 2006).

There have been several studies comparing the merits of different chemical similarity measures. It has been found that the Tanimoto coefficient provides a generally effective approach to molecular property prediction and similarity searching, and this coefficient is now widely used for measuring the similarity between pairs of 2D bit-strings (although some limitations that have recently become apparent Holliday et al.( 2002).

Data fusion rules were used in virtual screening and to a multiple integral formalism. Many of the fusion ideas applied to molecular similarities originate from the idea of combining several independent information retrieval system components to get and enhancing performance over that of the best individual component (Willett, 2000). Combining several similarity measures may be desirable because different measures may make use of different sources of evidence. They examine several cases of similarity fusion using different coefficients and different

representations and consider the reasons for positive or negative results in terms of the similarity distributions.

Three rankings based on 2D fragments, 3D fragments and physical properties, were calculated using the Tanimoto similarity measures, and fused using the SUM, MIN and MAX fusion algorithms (Ginn et al., 2000). Whilst the best similarity measure varies across activity subclasses, it was found that fusion using SUM and MAX generally gives a higher number of activities among the top rank positions compared to the individual measures. SUM was also found to perform better than any individual measure in terms of giving smaller rankings to active compounds. Another performance measure is the hamming distance between the bit strings repercentage activity subclasses of a database compound and a target compound at each rank. It is also proved that the SUM fusion is significantly better than each of the three original similarity methods, followed closely by MAX.

Using a different database, Ginn et al.(2000) also tried fusing 2D fingerprint, 3D atom-mapping and field based similarity searching measures. They found that the fused results give a generally high level of effectiveness compared to the best individual results. The SUM fusion also proved to be the best fusion rule compared to other fusion rules tested. The generally good results of fusion support its applicability in similarity searching, and the experiments, the best measures were found to vary from target to another.

Twenty two different binary similarity coefficients were fused based on the number of common top ranking compounds (Holliday et al.2002). Eleven similarity coefficients were identified as representative of the full range of coefficients. For each compound, the rankings produced by all possible combinations of any number of coefficients were averaged to give a new ranking. Based on this new ranking, it was found that the best performing 2 to 8 coefficient combinations returned more actives among the top 400 compounds compared to the best individual coefficient. However, the mean and median sum of actives returned by combinations never exceeded the mean and median sum of actives returned by individual coefficients. The low number of mean and median sum of actives and also the low number of

actives returned by best performing combinations as more coefficients are combined reflected the effect of poorly performing coefficients in combinations. In general, an individual coefficient that performs well on its own is also likely, but not guaranteed, to perform well in combinations. Most good performing combinations involved the Russell/Rao, Simple Matching, Stiles, Jaccard/Tanimoto, Ochiai/Cosine, Baroni-Urbani/Buser and Kulczynski(2) coefficients. Holliday et al. concluded that fused rankings give more actives among top ranking compounds compared to rankings based on individual coefficients if an appropriate combination of coefficients is chosen for the fusion.

Daut (2004) using neural network algorithms on finding best coefficient and fusion of coefficients for similarity searching. They found that can concluded that from MDDR database ,the stiles coefficients it the best used for large MDDR database, and for Z-score values  $<0$  .the baroni coefficient is best can be used for both Z-score values  $<0$  and  $>0$  .for ID alert database ,the Kulczynski(2)coefficient is best used for large database and z-score values  $<0$  while the Russell-Roa coefficient is best used for large database and for z-score value  $>0$  .for AIDS dataset, combination of Russell-Roa with stiles and Forbes coefficient is best used to perform similarity searching .

Welmina (2004) apply comparison of the effectiveness of probability model with vector space model for compound similarity searching. Result conducting a series of simulated similarity searching, it is concluded that PM approaches really did perform better than the existing similarity searching. It gave better result in all evaluation criteria to confirm this statement. In terms of which probability model performs better, the BD model shown improvement over the BIR model.

## **2.6 Genetic Algorithm**

In this section, an overview of genetic algorithms (GAs) will be presented that includes the origin of genetic algorithms; the basic understanding on what is

genetic algorithm and also the components of genetic algorithms, which include the various selection methods, crossover and mutation operators.

The GAs is general purpose optimization algorithms developed by Holland (1975). They are based on principles of natural evolution. In these algorithms, a population of individuals (chromosomes) undergoes a sequence of transformation by means of genetic operators to form a new population. Two operators of mutation and crossover were used. Mutation creates new individuals by a small change in a single individual and the crossover creates new individuals by combining parts of two individuals.

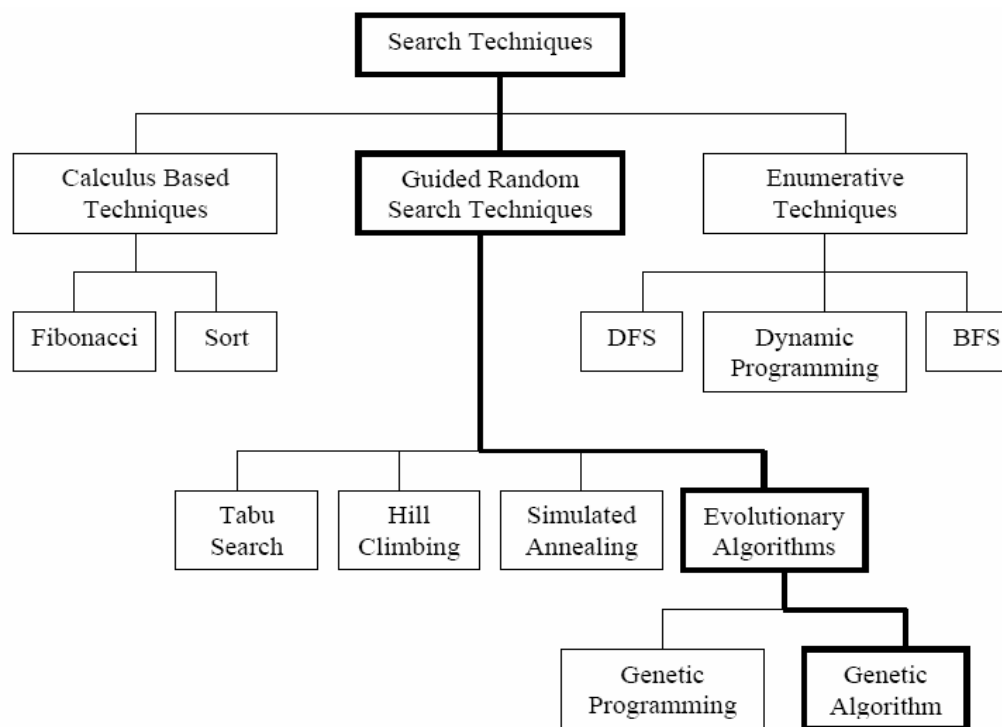
Genetic algorithms are well suited for chemical data searching problems due to their adaptability and their effectiveness at searching large spaces. The reason for genetic algorithms success at a wide and ever growing range of scheduling problems is a combination of power and flexibility. The power derives from the empirically proven ability of evolutionary algorithms to efficiently find globally competitive optima in large and complex search spaces. The favorable scaling of evolutionary algorithms as a function of the dimension of the search space makes them particularly effective in comparison with other search algorithms for the large search spaces typical of real world scheduling. The flexibility of the genetic algorithms has multiple facets.

The characteristics listed below are the main essence of Darwin's theory.

- (i) Each individual tends to pass on its characteristic to its descendants.
- (ii) Nature nevertheless produces individuals with different characteristics.
- (iii) Individuals with most favorable characteristics tend to have more descendents compared to those having lesser favorable characteristics. Thus, this drives the population towards favorable characteristics.

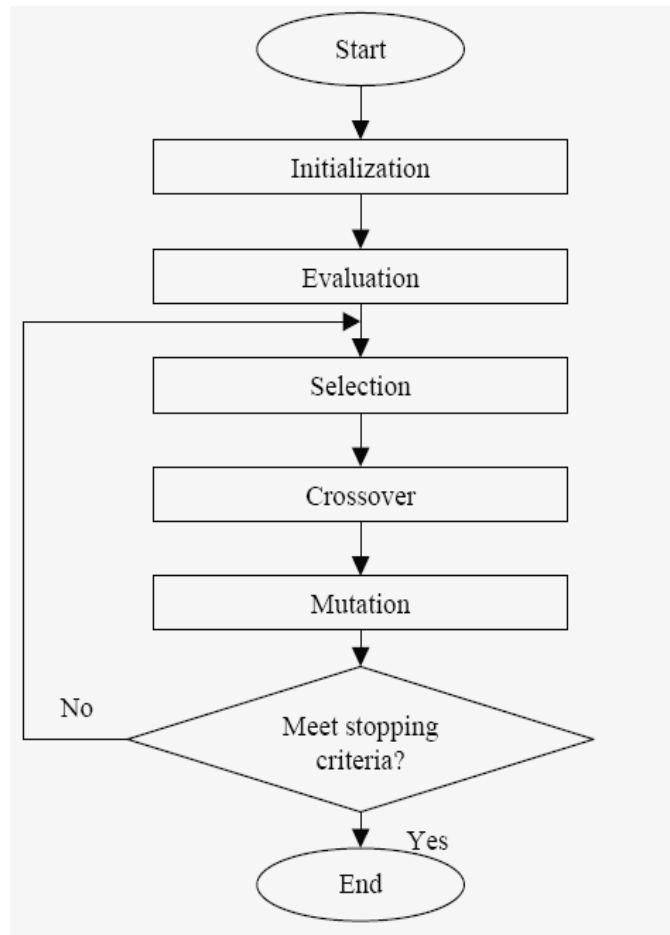
- (iv) Over the long period, variation can accumulate, producing entirely new species whose characteristics made them suitable for a certain or particular ecological life.

Genetic algorithms can be considered as a part of evolution algorithms for its usage of gene transmission and mutation mechanisms as an optimization technique. Figure 2.2 shows the classes of search techniques and it clearly highlights that genetic algorithms is a sub technique of evolutionary algorithms.



**Figure 2.3** Classes of search techniques

Once the genetic representation and the fitness function are defined, GA proceeds to initialize a population of solutions randomly, and then improve it through repetitive application of mutation, crossover and selection operators as shown figure 2.3.



**Figure 2.4** GA components

Each phase or components in GA plays an important role towards producing an optimal solution. These components will be explained in the next section.

### 2.6.1 Components of the Genetic Algorithm

The components of GA are listed as below:

(i) Initialization

There are many ways to initialize and to encode the initial generation through binary or non-binary, fixed or variable length strings, and others. At

the initial stage, the system generates randomly valid chromosomes and evaluates each one. Encoding in GA means transforming a chromosome into a string of symbols that can be any cardinality. Binary symbol is the most frequently used in chromosome encoding due to the fact that it is the simplest and it can represent maximum information with minimum number of bits. It is also because a binary chromosome can be operated with common genetic operators found in most packages of GA. However, But in the case of real-time scheduling problems, it is impossible to represent the chromosomes in the form of binary. Therefore, it is wise to represent the chromosome in the form of non-binary values, such as integers.

#### (ii) Reproduction

There are two types of reproduction, general reproduction and steady state reproduction. In general reproduction, the whole population can be potentially replaced at each generation. The most often used procedure is to loop  $N/2$  times, where  $N$  is the population size. For example, selecting two chromosomes each time according to the current selection procedure, producing two children from the two parents selected, and finally producing  $N$  new chromosomes. Using this method can lead to best chromosome in the population being replaced by a least fit chromosome.

Unlike the general reproduction, the steady state method selects one chromosome at a time according to the current selection procedure and performs crossover and perhaps mutation on them to obtain one or two children. The result is reinstall back to the population whereby the least fit of the population will be destroyed. Thus, it ensures that the best string found so far will always remain in the population. This result in a more aggressive search that in practice is quite often effective.

Goldberg and Dep (1991) have shown that replacing the worst member in the population much higher selective pressure compared to



random replacements. Therefore, the focus will be given to steady state reproduction as the reproduction method that will be used in this project.

### (iii) Selection

The purpose of the selection is to return the probabilistic of a selected parent. This procedure is a stochastic type of procedure, but it does not mean that GA employs a directionless search. The chances of each parent being selected very much depend on its fitness. There are six types of selection:

#### 1. Spatially-oriented selection

This selection is more towards local search, rather than a global search. This means that selection competition only occurs between several small neighbouring chromosomes, instead of the whole population. The neighbourhood is defined by the structure in which the population is distributed. The smaller the size of the neighbourhood is better, though the isolation distance between individuals in the population is bigger. But this will ensure the exchange of information between all individuals due to overlapping neighbourhoods.

#### 2. Tournament selection

In tournament selection, a number of individuals are chosen randomly from the population and the best individual from this group is selected as parent. This process is repeated as often as individuals that need to be chosen. The selected parent will then undergo crossover and/or mutation to produce new offspring.

### 3. Rank-based selection

In rank-based selection, the population is sorted according to the objective values. The fitness assigned to each individual depends only on its position in the individuals rank and not on the actual objective value. Rank-based selection overcomes the scaling problems of the fitness proportionate selection. The reproductive range is limited, so that no individuals generate an excessive number of offspring. Ranking introduces a uniform scaling across the population and is the method of choice where behaves in a more robust manner compared to fitness proportionate selection.

### 4. Fitness proportionate selection (Roulette wheel selection)

This is a standard and original method for parent selection, which is also known as the Roulette Wheel selection. It is the simplest selection scheme. In this kind of selection, each chromosome has a chance of selection that is directly proportional to its fitness. The effect of this depends on the range of fitness values in the current population. The technique provides a zero bias but does not guarantee a minimum spread.

### 5. Stochastic universal sampling

Stochastic universal sampling provides zero bias and guarantees a minimum spread compared to fitness-based selection. The individuals are mapped to contiguous segments of a line, such that each individuals segment is equal in size to its fitness exactly as in fitness-based selection. Equally, spaced pointers are placed over the line, as many as there are individuals to be selected. This kind of selection ensures a selection of offspring that is closer to what it deserves compared to fitness-based selection.

## 6. Truncation selection

Compared to the previous selection methods, modeling natural selection truncation selection is an artificial selection method. It is used by breeders for the purpose of large populations or mass selection. In truncation selection, the ideals' are sorted according to their fitness. Only the best individuals are selected for parents.

### (iv) Crossover

This is the most important operator in GA where it is a process of yielding recombination of bit strings via the exchange of segments between pair of chromosomes. In another way, crossover simply means a method for sharing information between two parents. The intention is to extract relevant features in the mating of the parents to produce better-fitted offspring. There are four kinds of crossover:

- One-point crossover

Parent	Chromosome 1	11011   00100110110
Parent	Chromosome 2	11011   11000011110
	Offspring 1	11011   11000011110
	Offspring 2	11011   00100110110

**Figure 2.5** One- point crossover

- Two-point crossover

Parent	Chromosome 1	11011   00100   110110
Parent	Chromosome 2	11011   11000   011110
	Offspring 1	11011   11000   110110
	Offspring 2	11011   00100   011110

**Figure 2.6** Two- point crossover

- N-point crossover
- Cut and splice

Parent	Chromosome 1	11011   00100110110
Parent	Chromosome 2	1101111   000011110
	Offspring 1	11011   000011110
	Offspring 2	1101111   00100110110

**Figure 2.7** Cut and splice

- Uniform crossover

In uniform crossover scheme (UX), individual bits in the string are compared between two parents. The bits are swapped with a fixed probability, typically 0.5. However, the bits are randomly copied from the first or from the second parent

(v) Mutation

By applying mutation on the chromosomes, it ensures that all possible chromosomes are reachable. There are several choices of mutation operators that can be applied onto the genes in a chromosome. For example, one can

apply simple mutation operators, order mutation operators or directed mutation operators.

Original offspring 1	1101111000011110
Original offspring 2	1101100100110110
Mutated offspring 1	1100111000011110
Mutated offspring 2	1101101100110110

**Figure 2.8** Mutation

(vi) Inversion

The inversion technique is somehow similar with reordering, whereby; it operates on a single chromosome and inverts the order of the elements between two randomly chosen points on the chromosome. However, a biological process inspired this operator and it requires additional overhead.

(vii) Migration

By allowing several populations to be run at the same time in each processor separately without increasing the total processing time, it allows a chromosome to be passing from one population to another population occasionally. In other word, it allows migration to be done. At each migration, a chromosome is chosen from one population according to the current selection procedure and is copied to other populations. The populations will remain in size, so the least fit in a population will be destroyed to allow insertion of another chromosome.

In most cases, most search methods prematurely converge to a suboptimal feasible or infeasible solution. Since a proper choice of penalty parameters are the key aspects of the working of such a scheme, most researchers experiment with

different values of penalty parameter values and find a set of reasonable values. The following criteria are always enforced:

- Any feasible solution will have a better fitness than any infeasible solution.
- Two feasible solutions are compared only based on their objective function values.
- Two infeasible solutions are compared based on the amount of constraint violations.

(viii) Termination of GA

The process of the GA is stopped by more than one methods such as the preset number of generations is reached and if there is no more improvement in total population average. The chromosomes in the last generation are chosen as the best individuals to solve the problem at hand. Therefore, the setting of termination affects also the performance of GA.

## 2.7 Summary

In this chapter, extensive literature review on the chemical database and its compounds and the representation of these databases in a computer and the process of fusions chemical compounds was represented highlights were made on the classical of representation of chemical structures, retrieval data from chemical database (molecule) and how the Molecular Descriptors for Similarity Searching and Similarity Coefficient. The data fusions that have been used in combination with molecular representation, Molecules in a Single Query and Similarity Coefficient, were also overview. Then finally, how GAs basically work for solving problems such as the one in hand have been covered

## **CHAPTER 3**

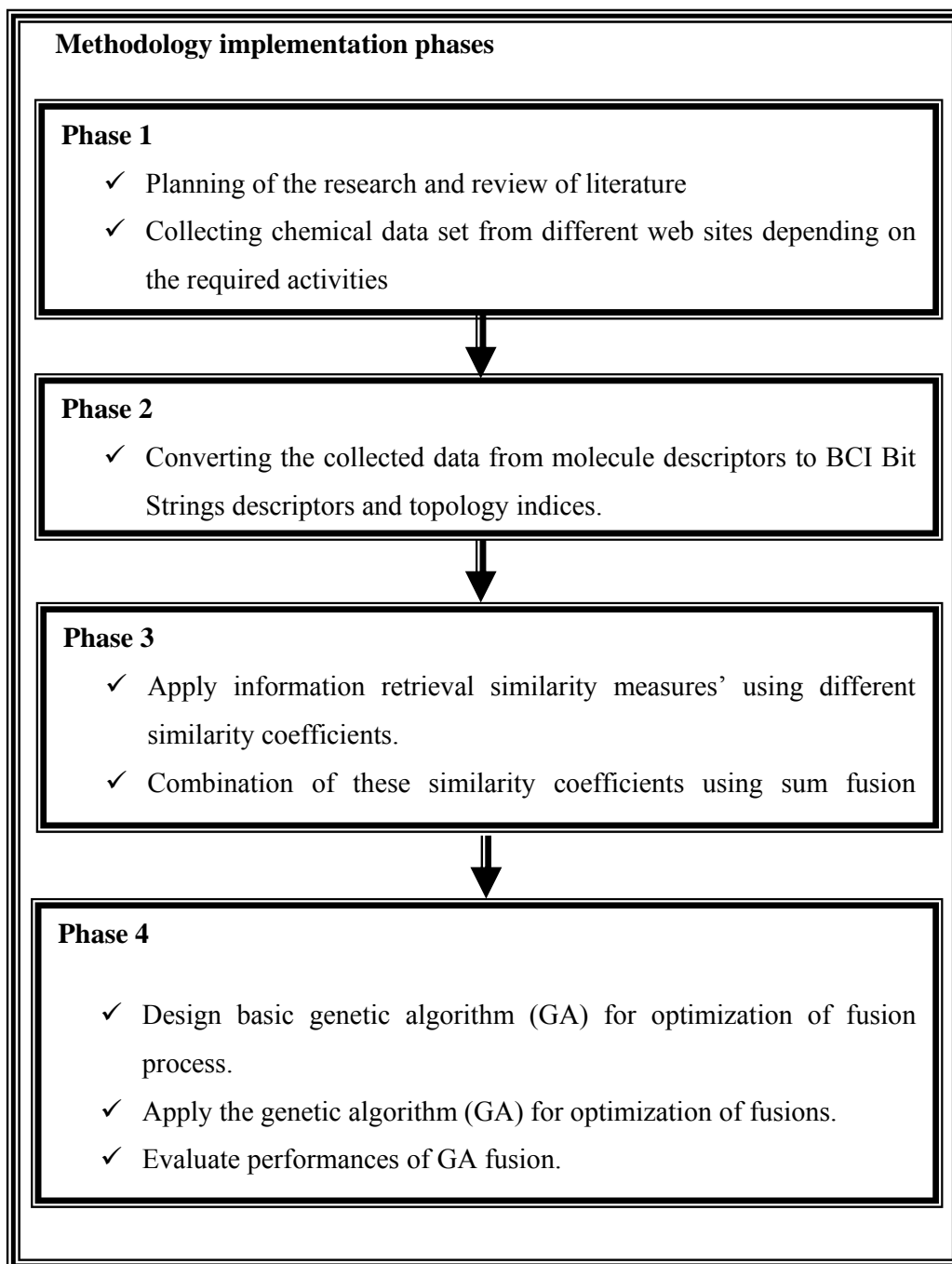
### **METHODOLOGY**

#### **3.1 Introduction**

This chapter presents the methods that will be used to carry out optimization of weights from different similarity coefficients for their data fusion in chemical database through the implementation of GAs. The methodology employed will provide a systematic framework of procedures and principles to achieve the objective of this study.

#### **3.2 Operational Framework**

The study methods will be conducted according to the workflow process as illustrated in Figure 3.1.



**Figure 3.1** Flow Chart of the Framework



### 3.2.1 Research Plan and Review of Literature

The work plan include understanding the representation of chemical structures, retrieval of data from chemical structure, molecular descriptors for similarity searching and using similarity coefficients to get similarities between the molecules. Chemical database containing 1360 compounds have been used. Data needed to Preparing to retrieval from chemical database as well as how to combine or fuse between coefficients. Furthermore, data was prior for using the genetic algorithm (GAs) to optimize the weights in order to find the better combination between the coefficients.

### 3.2.2 Collection of Chemical Data Set

Chemical databases containing huge number of compounds are available in many web sites. In this study, the required data is MDL Drug Database Report (MDDR) (MDDR is a database covering the patent literature, journals, meetings and congresses). Produced by Molecular Design Limited's (MDL) and Prous Science, the database contains over 100,000 compound biologically relevant compounds and well-defined derivatives, with updates adding about 10,000 a year to the database. The MDDR Finder allows searching the database by structure or across relevant data fields. MDL also offers MDDR-3D collected from the Discovery Gate web site (URL:<https://www.discoverygate.com>).

Although the MDDR (MDL Drug Database Report) contains many chemical compounds in groups of activities with different biological similarities, the data collection is focused on selecting some compounds of different activities.

### 3.2.3 Converting Molecule Data to Barnard Chemical Information (BCI)

The data available in molecule graph format is converted using MAKEBITS software from BCI (Barnard Chemical Information) into bit string format where the compound is represented as series of 0's and 1's without spacing between them. The result of this conversion is a text file containing all the data in BCI bit strings format. Each bit represents the existence or absence of certain fragment in the molecule. Bit string descriptors are chosen because of their ability to distinguish between actives and inactive better than the other descriptors (Brown and Martin, 1996). They have also been found to be the most effective descriptors in selecting representative subsets of the bioactive compounds (Matter, 1997).

BCI is a 1052-bit structural key-based bit string generated based on the presence and absence of fragments in the BCI's standard 1052 fragment dictionary, which encodes augmented atoms, atom sequences, atom pairs, ring components and ring fusion descriptors (Dittmar *et.al*, 1983). BCI dictionary could generate thousands of keys, resulting in molecular fingerprint bit lengths of approximately 5,000 bits (MacCuish and MacCuish, 2003).

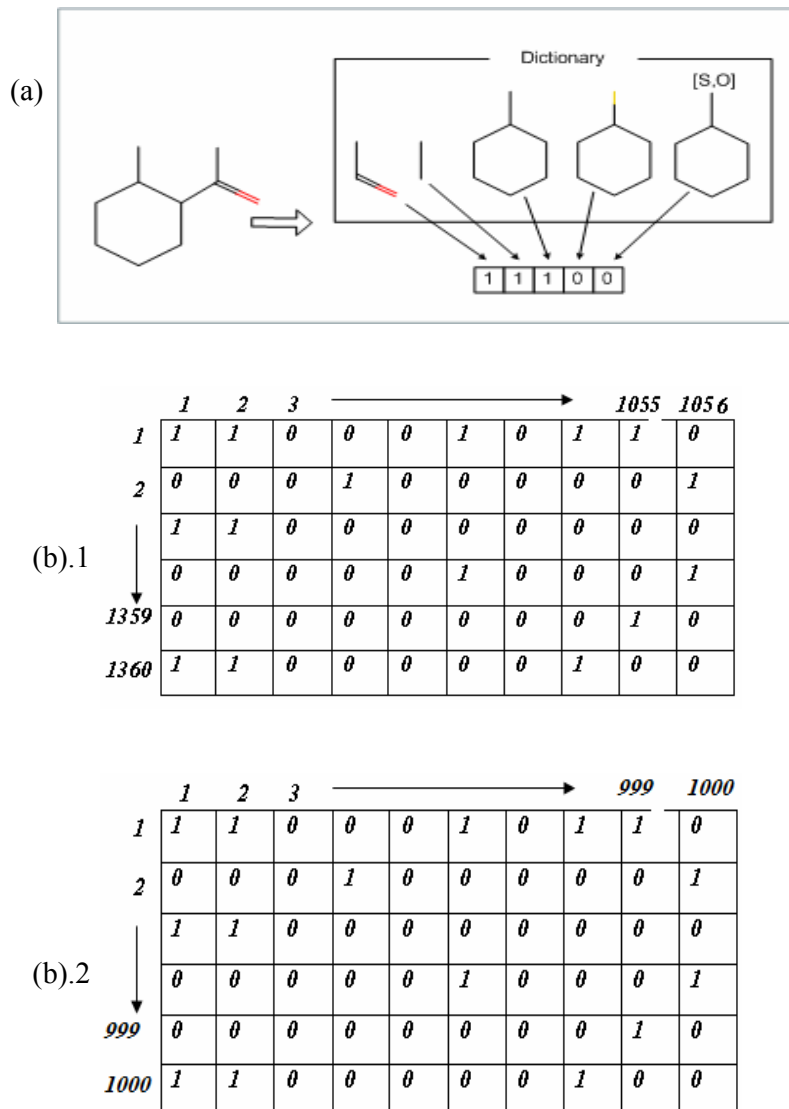
Bit strings are strings of 1 s and 0 s. They can be used to store or visualize bit masks. They can be used also to represent sets or to manipulate binary data.

#### I. Input data file formats and extension

- Sybyl © MOL2 files (.mol, .ml2, mol2) by Tripos, Inc.Sybyl © Molfiles (.sm2) as provided by ChemOffice, CambridgeSoft Corp.Sybyl © multiple Molfiles (.mol, .ml2) by Tripos, Inc.Molfiles (.mol) by Molecular Design Ltd. (MDL).Multiple SD files (.sdf) by Molecular Design Ltd. (MDL).HyperChem © files (.hin) by Hypercube, Inc.SMILES notations (.smi).MacroModel © files (.dat, .out) by Schrodinger

#### II. Output data formats (BCI Fingerprint File .bci)

Bellow is an example how MAKEBITS works to convert molecule data into BCI bit string data as shown in figure 3.2.



**Figure 3.2** Generate BCI bit string descriptor by using MAKEBITS software (a) BCI dictionary could generate and (b).1 and (b).2 Two Dimensional Input Vector containing input data, where row represented as the molecular and column is compounds.

### 3.2.4 Converting Molecule Data to Topological Indices

Topological indices have been used to discriminate “drug-like” compounds from “non-drug-like” ones. Topological descriptors support high-speed technological processes that use combinatorial syntheses and high-throughput screening of toxicity of large number of compounds. These descriptors do not require time-consuming computation procedure.

The topological indices characterize the bonding pattern of a molecule by a single value integer or real number, obtained from mathematical algorithms applied to the chemical graph representation of the molecules. Each index, thus, contains information not about fragments or some locations on the molecule, but rather about the molecule as a whole. Simpler descriptors include the number of atoms and bonds and the number of readable bonds (Kunal, 2004). Table 3.1 shows sample data generated using DRAGON software.

To run the DRAGON software molecular structure files previously obtained by other specific molecular modeling software are used. The most commonly accepted molecular file formats are:

1. Sybyl © MOL2 files (.mol, .ml2, mol2) by Tripos, Inc.
2. Sybyl © Molfiles (.sm2) as provided by ChemOffice, CambridgeSoft Corp.
3. Sybyl © multiple Molfiles (.mol, .ml2) by Tripos, Inc.
4. Molfiles (.mol) by Molecular Design Ltd. (MDL)
5. Multiple SD files (.sdf) by Molecular Design Ltd. (MDL)
6. HyperChem © files (.hin) by Hypercube, Inc.
7. SMILES notations (.smi)
8. MacroModel © files (.dat, .out) by Schrodinger

To make full use of DRAGON calculations, 3D optimized structures with hydrogen's should be used. However, DRAGON can also deal with H-depleted molecules and 2D-structures, but some restrictions to descriptor calculation should

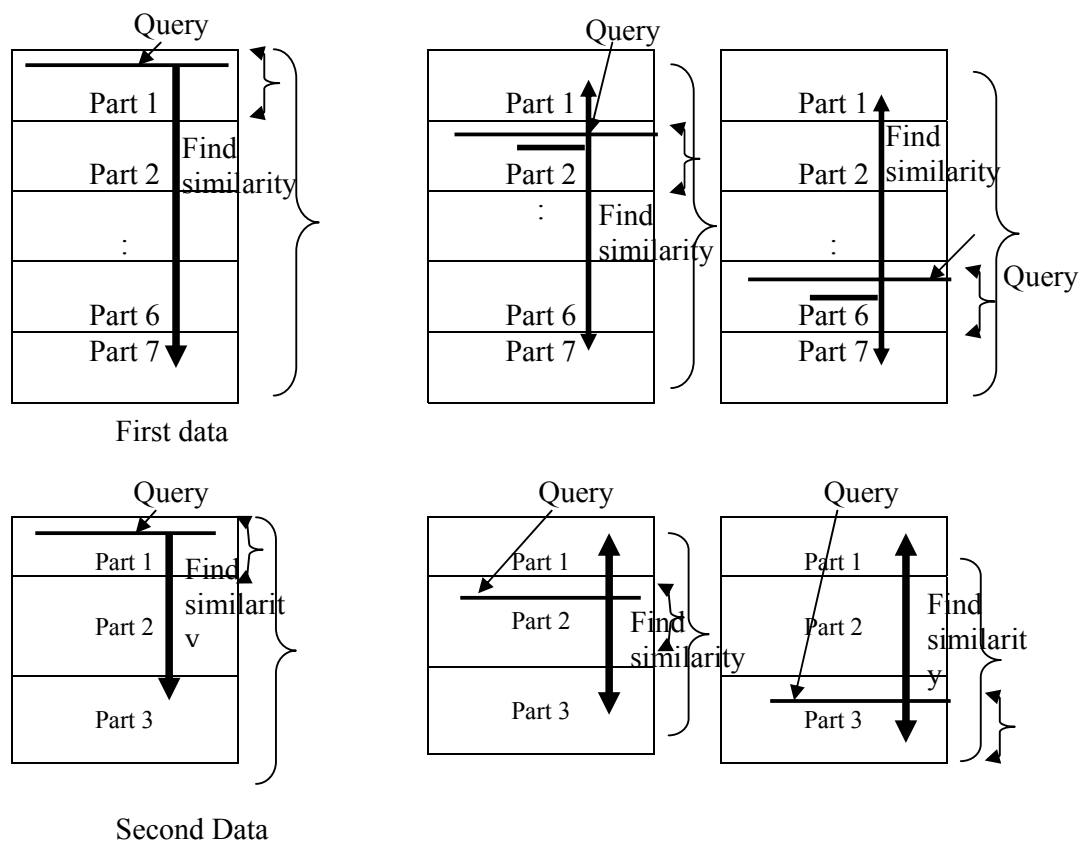
be applied. Using DRAGON software, most of the common organic and inorganic compounds, both charged and uncharged, are correctly processed. DRAGON cannot process molecules containing atoms for which some physicochemical properties are undefined, disconnected structures such as salts, and molecules with radicals.

**Table 3.1:** Topological indices descriptors generated using dragon software

<b>Molecule ID</b>	<b>Molecule Properties or Parameters'</b>									
<b>ID</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>
1	1.43	-0.3	-0.17	-0.1	1.6	1.62	1.38	-1.01	-0.32	-0.42
2	1.78	-0.21	-0.35	-0.11	1.81	1.73	1.15	-1.01	-0.36	-0.42
3	1.78	-0.21	-0.37	-0.11	1.81	1.73	1.15	-1.01	-0.36	-0.42
4	1.68	-0.33	-0.27	-0.1	1.66	1.68	1.46	-1.01	-0.32	-0.42
5	1.43	-0.3	-0.27	-0.1	2.32	1.68	1.15	-1.01	-0.33	-0.42
6	2.17	-0.15	-0.47	-0.13	1.5	1.68	1.7	0.88	-0.46	-0.42
7	1.78	-0.21	-0.3	-0.12	1.3	1.62	1.93	0.88	-0.43	-0.42
8	2.05	-0.27	-0.37	-0.11	1.3	1.62	1.77	-1.01	-0.38	-0.42
9	0.78	-0.03	-0.3	-0.11	1.19	1.47	1.15	0.88	-0.35	-0.42
10	0.4	-0.33	0.04	-0.06	0.22	0.37	0.13	0.88	-0.11	-0.42
11	0.84	0.11	-0.44	-0.13	1.5	1.31	0.91	-1.01	-0.44	-0.42
:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:
1360	1.48	0.14	-0.6	-0.14	0.22	0.84	0.6	0.88	-0.57	2.41

Each index, thus, contains information not about fragments or some locations on the molecule, but rather about the molecule as a whole. Simpler descriptors include the number of atoms and bonds and the number of rotatable bonds.

The data represented here composed from active compounds with different degree of the activity for each of them. There are two data's as shown in figure 3.3; the first data has seven actives. The second has three actives, the first compound from each part selected as active target (query), all other compounds are assumed to be inactive, to find the similarity to that target. The process will continue the some way for other parts.



**Figure 3.3** Show the Mechanism to Searching

The molecular data converted to BCI bit-strings will be used for similarity searches on the MDDR database through the implementation of the similarity coefficient given in Table 2.2 Salim and co-workers in (2003) they have clustered the mentioned 22 coefficient into 13 groups separated of coefficient in order to enhance the similarity searching as well as combination between groups. To calculate the similarity of retrieval compound depend on the 13 groups we use the equation (3.3) shows under this table. The 13 groups of coefficients are shown in Table 3.2.

**Table 3.2:** The 13 Groups Of Coefficients (From Salim El At, 2003), In This Study the 13 Groups of Coefficients Will Be Used.

NO	THE GROUPING OF COEFFICIENTS	The Number Related To Table 2.2
1	Group A : {Jaccard/Tanimoto, Dice, Sokal/Sneath(1), Kulczynski(1)}	{1 2 4 5}
2	Group B : {Russell/Rao}	{3}
3	Group C : {Simple Matching, Hamann, Sokal/Sneath, Rogers/Tanimoto, Sokal/Sneath(3), Mean Manhattan}	{6 7 8 9 10 22}
4	Group D : {Baroni-Urbani/Buser}	{11}
5	Group E : {Ochiai/Cosine}	{12}
6	Group F : {Kulczynski(2), McConnaughey}	{13 19}
7	Group G : {Forbes}	{14}
8	Group H : {Fossum}	{15}
9	Group I : {Simpson}	{16}
10	Group J : {Pearson}	{17}
11	Group K : {Yule}	{18}
12	Group L : {Stiles}	{20}
13	Group M : {Dennis}	{21}

### 3.2.5 Data Fusion Process

Similarity measures in the chemical information field have, in the main, been limited to a few single measures such as the Tanimoto, Cosine, and Euclidean Distance. Indeed, many of those shown in Table 3.2 have rarely been used in connection with chemical structure similarity calculations. Several have been used in comparative studies, seeking to determine which was best single measure for performing a particular task, a similarity search for instance. The variety of performance of the similarity measures available, has led to an interest in data fusion

methods for combining more than one similarity measure with a goal to improving the performance of the single measure. A study by Ginn et al (2002). suggests that combining the rankings of searches using more than one coefficient would give improved performance, and this was further tested by (Holliday et al, 2002).

Two procedures were employed in order to carry out the required data fusion process.

1. Application of information retrieval similarity measures' using different similarity coefficients (weight score, rank).

The retrieval of similarity coefficients is used to obtain a numeric quantification to the degree of similarity between a pair of structures. The coefficients were examined in order to group together those with comparable performance when applied to searches of binary representations. Example coefficients from each of the groups were then used to determine whether the performance of a single coefficient could be improved by combining coefficients using data fusion.

A set of 13 similarity coefficients, as shown in table 3.2 (chapter2), was selected as has been done by salim et al (2003). The formulas shown in table 3.2 indicate the similarity (or dissimilarity in the case of coefficient 13), between two bit-string representations, molecules  $M_1$  and  $M_2$  of length  $n$

$$S(M,Q) = \sum_{i=1}^{13} S_i(M,Q) \quad (3.1)$$

Where  $S(M,Q)$  is the similarity,  $M$  is molecules,  $Q$  is query (target),  $n$  is the number of molecules and  $CO$  is list of coefficient from 1 to 13 formulas.

2. Combination of similarity coefficients using sum fusion method.

Similarity measures in the chemical information field have mainly, been limited to a few single measures such as the Tanimoto, Cosine, and Euclidean



Distance. Many of those shown in table 3.2 have rarely been used in connection with chemical structure similarity calculations.

Fusion was carried out using representative coefficients selected from each of the 13 groups resulting from clustering 22 similarity coefficients. The rank-positions from the coefficients were summed to give a new similarity ranking for each compound when compared to a target. The SUM fusion function was used as it was found to be the most effective in an earlier study.

Data fusion was based on a summing procedure on the rankings produced by the similarity searches. The combination of similarity rankings using data fusion was found to be the most effective method for similarity searching in chemical databases according to the following steps:

1. Execute a similarity search of a chemical database for some particular target structure using two, or more, different measures of inter-molecular structural similarity.
2. Note the rank position,  $r_i$ , of each database structure in the ranking resulting from use of the  $i$ -th similarity measure.
3. Combine the various rankings using one of the fusion rules (SUM), giving a new combined score for each database structure.
4. Rank the resulting combined scores, and then use this ranking to calculate a quantitative measure of the effectiveness of the search for the chosen target structure.

The sum of fusion rules for combining  $n$  ranked lists are given by:

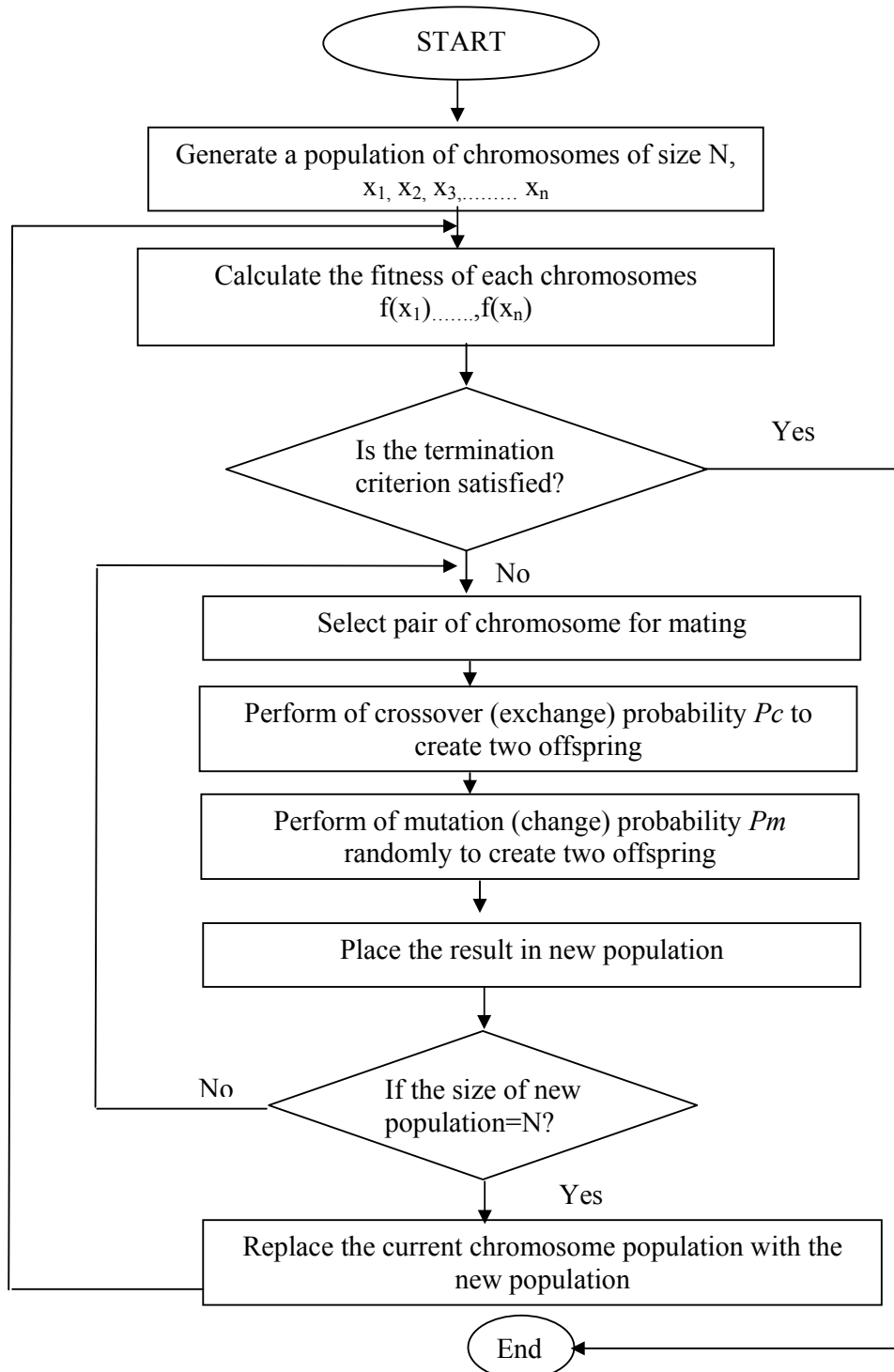
$$SUM_{FUS} = \sum_{i=1}^n r_i \quad (3.2)$$

Where  $r_i$  denotes the rank position of a specific database structure in the  $i$ -th

( $1 \leq i \leq n$ ) ranked list,

### 3.2.6 Design Basic Genetic Algorithm (GA)

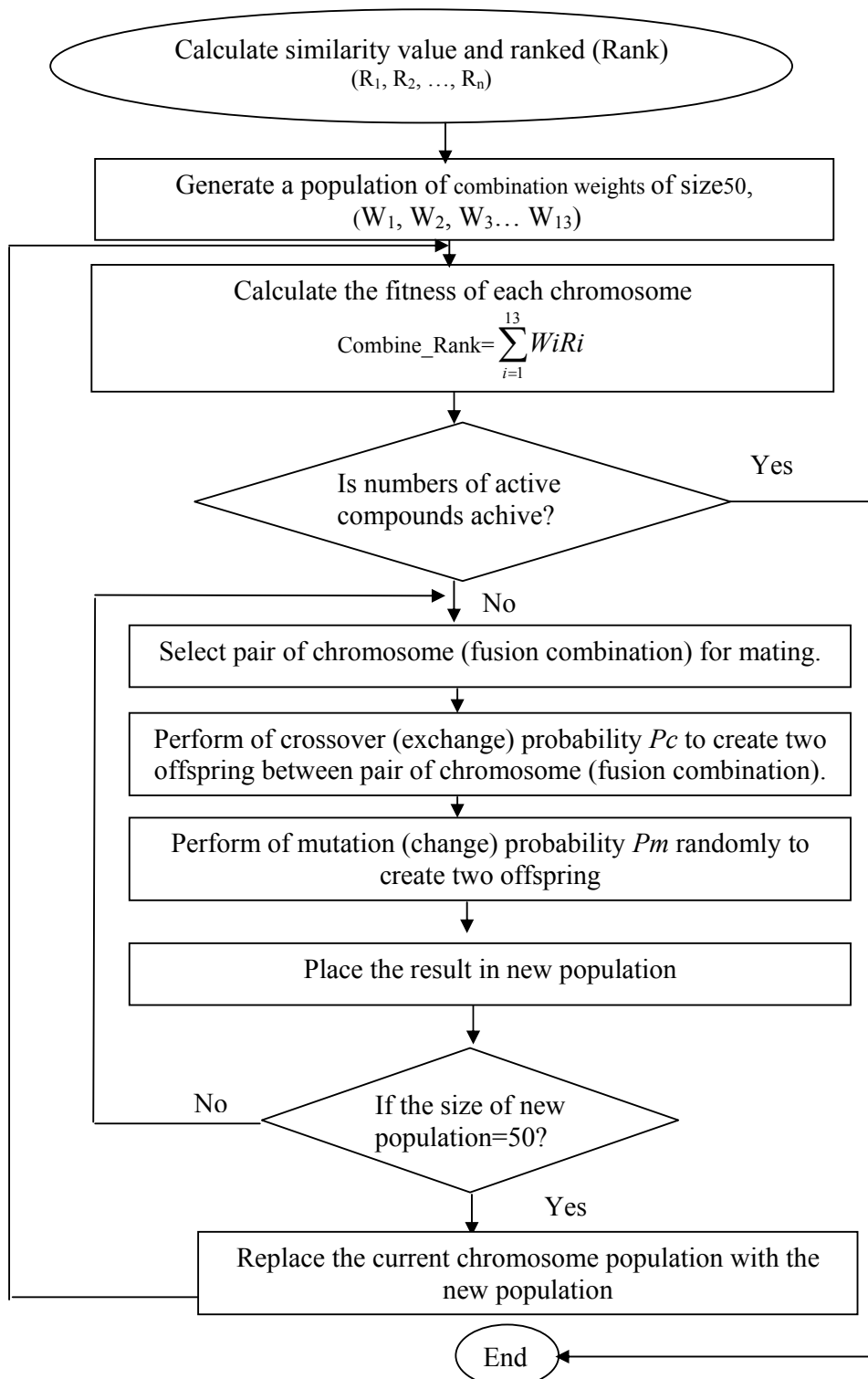
The Genetic algorithm (GA) is used and the implementations flow chart is depicted figure 3.3.



**Figure 3.4** Flow Chart of the Basic Genetic Algorithm

### 3.2.7 Optimization of Fusion Process

For the purpose of the optimization of fusion process the genetic algorithm is used and the implementations flow chart is depicted in figure 3.4.



**Figure 3.5** Flow Chart of the Optimizations of the Fusion Process

### 3.2.8 Applying the Genetic Algorithm (GA) for Optimization of Fusions

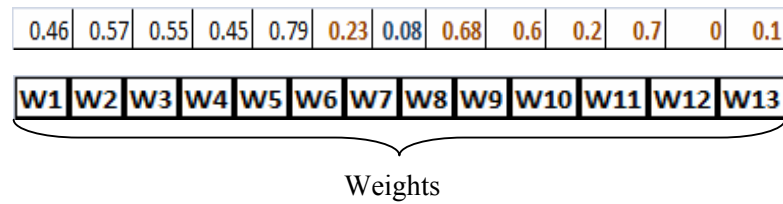
The Preprocessing procedure to apply the GA, for optimization weight purposes begins with the calculation the different similarity measures based on similarity coefficient using the formula given in table 2.2. The method is based on linear combinations of ranking from different similarity measures instead of similarity values, as a way to standardize the data. For each target structure, the combinations were sorted into decreasing order of the number of actives retrieved. Then utilize Genetic Algorithms (GA) to explore the search space of the weights. GA emulates the process of evolution of species to search for more 'fit' individuals. These algorithms are very well suited to explore complicated multidimensional space. We assign weights (in the range of 0.0 to 1.0) to each similarity searching and combined the rankings to get a combined ranking for each structure. A negative weight attached to a value signifies a reduced role in retrieval for the particular similarity searching that produced ranking. A positive weight, on the other hand, signifies an increased role in retrieval. The structures are then ordered in increasing order of this combined ranking and then presented to the user for evaluations. A proper selection of weights, thus tries to exploit such complementarities.

The input data:

1. Auto dimensional input vectors which used is consist from 1 to 1360 molecules. It contains 1056 column and 1360 rows, each of which repercentage only one chemical compound.
2. The value for crossover ( $P_c$ ) is taken to be (0.6 to 0.9). This value generally produces good result, and for mutation ( $P_m$ ) best value is between 0.001and 0.1 due to being quite small and kept quite low for using GAs.

The overall process could be completed in ten main steps.

**Step 1:** Represent the weight as chromosome (parent), and any chromosome consist of 13 genes (any gene  $\equiv$  Single coefficient (13 coefficients)).



**Figure 3.6** Chromosome Representation

**Step 2:** Generate an initial chromosomes population the perfect population size is about 50.

Chromosomes (groups)  $\equiv$   $ch_1, ch_2, ch_3 \dots ch_n$

Where chi is generated randomly weights (chromosomes)

ch1	0.0	0.0	0.0	0.2	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ch2	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ch3	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.2	0.2	0.2	0.2	0.6	0.2	0.3
ch4	0.0	0.2	0.0	0.2	0.2	0.4	0.0	0.4	0.0	0.4	0.0	0.8	0.0	0.5
ch5	0.2	0.4	0.2	0.4	0.4	0.6	0.2	0.6	0.2	0.6	0.2	1.0	0.2	0.7
ch6	1.0	0.6	0.4	0.6	0.0	0.8	0.0	0.8	0.4	0.2	0.3	0.0	0.4	0.9
:														
chn														

**Figure 3.7** Population of Chromosomes (random weights)

**Step 3:** Calculation of the fitness of each individual chromosome, using the formula:

$$f(w) = \sum_{i=1}^{13} W_i R_i, \text{ and select top 10\% of actives} \quad (3.3)$$

where  $R_i$  similarity value normalized or ranking positions from similarity searching for the structure in the collection.  $W_i$  is the associated weight

generated by GA. 'i' varies from 1 to the number of similarity measure used in the experiment.

**Step 4:** Select pair of chromosomes from the current population to apply the genetic algorithm operators (crossover and mutation)

**Step 5:** Creation of a pair of offspring chromosomes to apply the genetic operators' **Crossover**. In my work I use two-point crossover

- Single-point crossover

0.66	0.77	0.75	0.65	0.99	0.43	0.28	0.88	0.8	0.4	0.9	0.2	0.3
0.46	0.57	0.55	0.45	0.79	0.23	0.08	0.68	0.6	0.2	0.7	0	0.1

New two offspring or Childs

0.46	0.57	0.55	0.45	0.79	0.23	0.08	0.68	0.6	0.2	0.7	0	0.1
0.66	0.77	0.75	0.65	0.99	0.43	0.28	0.88	0.8	0.4	0.9	0.2	0.3

**Figure 3.8** Single-point Crossover process

**Step 6:Mutation** creation of offspring chromosomes, which change the bit randomly one bit or two to the chromosome already crossover.

In mutation operation we used inversion technique to change randomly chosen elements. This process we ensures no falling in local optima .the probability was used is 0.01%.

- Mutation operation of one bit change

0.46	0.57	0.55	0.45	0.79	0.23	0.08	0.68	0.6	0.2	0.7	0	0.1
------	------	------	------	------	------	------	------	-----	-----	-----	---	-----

New offspring or child

↓

0.46	0.57	0.55	0.45	0.79	0.23	0.01	0.68	0.6	0.2	0.7	0	0.1
------	------	------	------	------	------	------	------	-----	-----	-----	---	-----

**Figure 3.9** Inversing Mutation process

Most the mutation  $P_m$  (between 0.001 and 0.1).

**Step 7:** Placing the created offspring chromosome in the new population.

**Step 8:** Repeat step 4 until the size of the new chromosome population =  $N$ .  
 $N=50$ .

**Step 9:** Replace the initial (parent) chromosomes population with the new (offspring) population.

**Step 10:** Go to step 3, and repeat until the termination criterion is satisfied (until to achieve the number of iteration).

### 3.3 Evaluation of the GA

Fitness is a numerical score assigned to each chromosome. It is expected that the more fit (the higher the fitness number) chromosome the better is the utility of the chromosome in solving the problem at hand. Thus the selection of fitness function is vital for the effectiveness of GA; and might provide an indication of how good the solution is and survival.

The process of the GA is stopped when preset number of generations is reached which are 1500.

In this study, different number of compound from either small or large data was being tried with different database parameters, through long or short queries.

### 3.4 System Requirements

The requirements involved in developing the system can divide into two types, which is hardware and software:

- i) Hardware:
  - a) Processor Intel Pentium 4.
  - b) Memory 512 MB.
  - c) Hard disk 80GB.
  - d) Monitor 15”.
  
- ii) Software:
  - a) Matlab 7.4
  - b) Microsoft Visual C++ 6.0.
  - c) Microsoft Word (Office 2007).
  - d) Microsoft Work
  - e) Windows XP/Vista.

### 3.5 Summary

This chapter discussed the methodology which is used in implementing this project, it has four Phases, starting from the first phase which has: planning of the research and review of literature and collections of the chemical data sets. Second phase is converting the collected data sets from Molecule descriptors to BCI and topological indices. Third Phase apply information retrieval similarity measures using different similarity coefficient, and combination of these similarity coefficients using sum fusions. Finally design basic genetic algorithms (GA) and design to optimization of fusion process. Then the steps to apply GA for optimization of fusions. Each stage of these stages plays an important role in accomplishing this study.



## CHAPTER 4

### EXPERIMENTAL RESULTS AND DISCUSSIONS

#### 4.1 Introduction

This chapter primarily presents the results obtained by retrieving the chemical compounds from chemical databases using different similarity measures based on different similarity coefficients and molecular representations.

Several query structures were selected and used in the similarity search. These similarity searches were carried out against the respective test database using 13 coefficients, discussed in chapter 3. The database molecules were ranked in a decreasing order of the calculated similarity coefficient. The rankings of two coefficient search for the same query were compared by counting the number of similarity active compounds to the activity of the query compound in the top *ranked* structures.

#### 4.2 Representation of Chemical Compounds

Two groups of chemical data sets are used in this project. The first group contains 1360 compounds divided into seven groups (Activities) depending on its

biological similarities as represented in Table 4.1. The second group contains 1000 compounds and is divided into three Activities (Table 4.2).

The chemical compounds are available in molecule format in chapter 3 (Figure 3.2) and converted to BCI bit Strings.

**Table 4.1:** First data and it's activities.

<b>NO</b>	<b>Activity</b>	<b>Start</b>	<b>End</b>	<b>No. Compounds</b>
1	Interacting on 5HT receptor	0	270	271
2	Antidepressants	271	502	232
3	Antiparkinsonians	503	636	134
4	Antiallergic/antiasthmatic	637	859	223
5	Agents for Heart Failure	860	959	100
6	AntiArrhythmics	960	1159	200
7	Antihypertensives	1160	1359	200

For the first group, the 1360-molecules are from the MDL Drug Data Report (MDDR) database, containing molecules of drugs launched. The main groups, their subgroups and their aggregate activity are summarized see appendix A.

**Table 4.2:** Second data and it's activities.

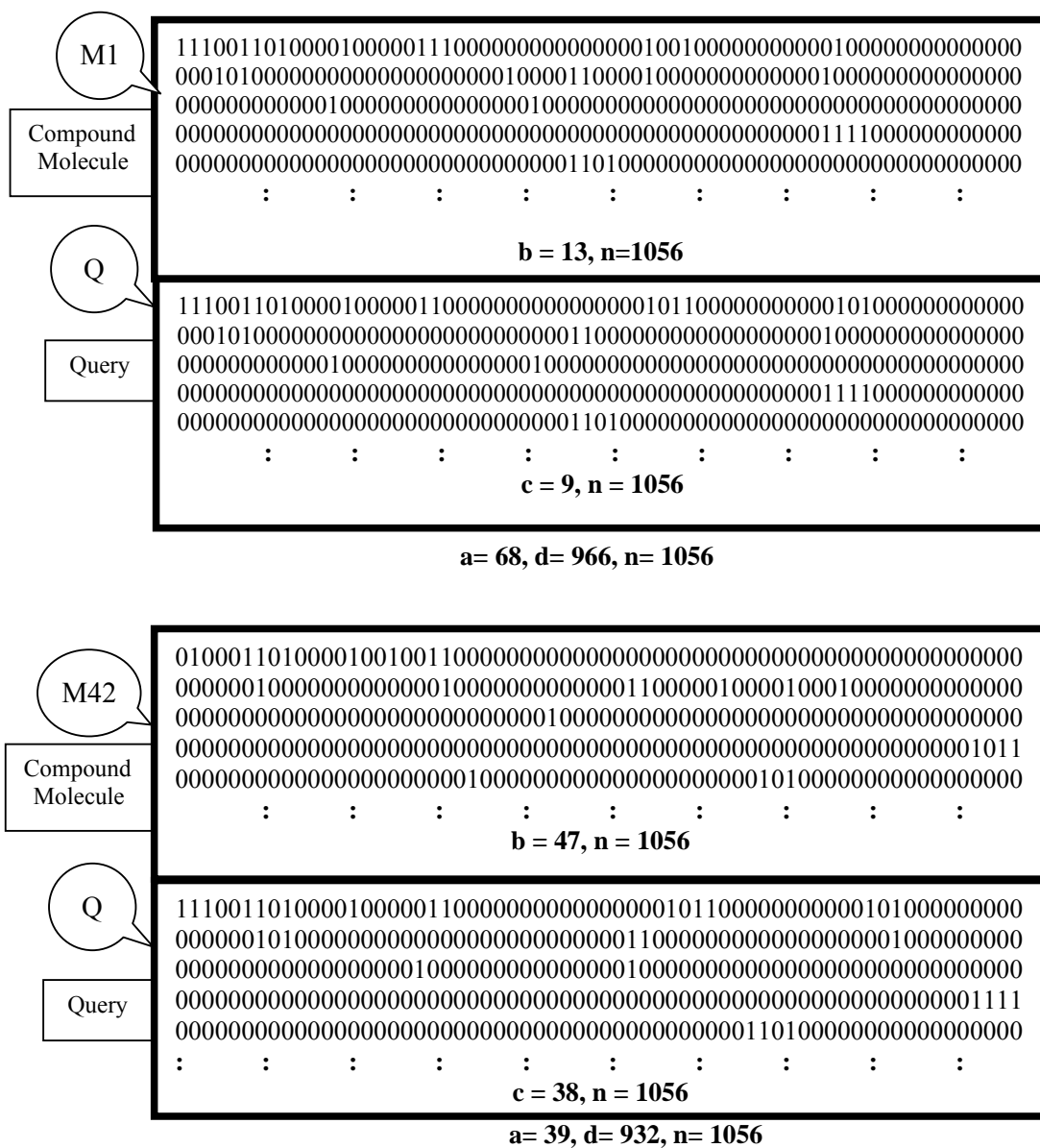
<b>NO</b>	<b>Start</b>	<b>End</b>	<b>No. Compounds</b>
1	1	247	247
2	248	500	253
3	501	1000	500

The input data contains the first group, which consists of 1360 compounds and each compound is represented by binary vector containing 1056 columns.

### 4.3 Compounds Retrieval

The compound retrieval from chemical database using different similarity a measure is based on different similarity coefficients .As can be seen It below, clearly shows how compounds are retrieved from chemical databases by applying the query Q (for M1 and M42) and thus; obtaining the similarity using formula in chapter 3 (Table 3.2).

A matlab program is written to demonstrate how chemical compounds are retrieved and similarity of compute using the coefficient formula. The example below shows how to define the variables used as coefficients are defined.



The values for variables a, b, c, d and n discussed earlier in chapter 2 are computed and thus are prerequisites for calculating the coefficients.

```

clc
co2=0;
co3=0;
co4=0;
m=data(1,:);
N=length(m);
for i=1:1360
    a(i,:)=data(i,*)&m;
end

for i=1:1360
    for j=1:1056
        if ((data(i,j)==1)&(m(1,j)==0))
            b(i,j)=co2+1;
        end;
    end;
end;

for i=1:1360
    for j=1:1056
        if ((m(1,j)==1)&(data(i,j)==0))
            c(i,j)=co3+1;
        end;
    end;
end;

for i=1:1360
    for j=1:1056
        if ((data(i,j)==0)&(m(1,j)==0))
            d(i,j)=co2+1;
        end;
    end;
end;

countd=sum(d.');
counta=sum(a.');
countb=sum(b.');
countc=sum(c.');

```

**Figure 4.1** Similarity Measure program for obtaining the value of a, b, c, d and n.

#### 4.4 Similarity Calculation

In this section, the formula of coefficients is applied to find similarity between molecules and the query. Table 4.3 present examples of the similarity values after calculation of the 13 group different coefficients such as Table 3.2 in chapter3. From each group the one coefficient representative is obtained. Figure 4.2 shows the Matlab program developed to applying the Similarity using single coefficient equations.

```
JT=counta./(counta+countb+countc);JT_f=JT.';
RR=counta./N;RR_f=RR.';
SM=(counta+countd)./N;SM_f=SM.';
BB=((sqrt(counta.*countd))+counta)/((sqrt(counta.*countd))+counta+countb+countc);
    BB_f=BB.';
OC=counta./sqrt((counta+countb).*(counta+countc));OC_f=OC.';
KU2=(counta/2).*((2*counta+countb+countc))/((counta+countb).*(counta+countc));
    KU2_f=KU2.';
FOR=(N*counta)/((counta+countb).*(counta+countc)); FOR_f=FOR.';
FOS=(N*((counta-(1/2)).^2))/((counta+countb).*(counta+countc));FOS_f=FOS.';
SIM=counta./(min(counta+countb,counta+countc));SIM_f=SIM.';
PE=(counta.*countd -countb.*countc)/
(sqrt((counta+countb).*(counta+countc).*(countb+countd).*(countc+countd))); PE_f=PE.';
Yu=(counta.*countd - countb.*countc)/(counta.*countd + countb.*countc);Yu_f=Yu.';
Stiles=log10(N*((abs(counta.*countd-countb.*countc)-(N/2)).^2)/
((counta+countb).*(counta+countc).*(countb+countd).*(countc+countd)));
    Stiles_f=Stiles.';
Den=(counta.*countd - countb.*countc)/(sqrt(N*((counta+countb).*(counta+countc))));
Den_f=Den.';
```

**Figure 4.2** Single Coefficient Analysis using program to Find Similarity

**Table 4.3** : Similarity Values of Different Coefficients

Molecules	Tanim	Ruse	Simple	Bar	Cosine	Kuls	Forbes
M1	1	0.0767	1	1	1	1	13.04
M2	0.7556	0.0644	0.97917	0.93647	0.861	0.8613	11.51
M3	0.62	0.0587	0.96402	0.88936	0.7654	0.7654	9.979
M4	0.604	0.0578	0.96212	0.88316	0.7531	0.7531	9.818
M5	0.6327	0.0587	0.96591	0.89465	0.7751	0.7751	10.23
M6	0.5192	0.0511	0.95265	0.84882	0.6838	0.684	9.143
M7	0.5769	0.0568	0.95833	0.87172	0.7318	0.7318	9.424
M8	0.5289	0.0521	0.9536	0.85277	0.692	0.6921	9.193

Molecules	Fossum	Simpson	Pear	Yule	Stiles	Dennis
M1	1043	1	1	1	3.0178	30.004
M2	771.43	0.88312	0.84983	0.99644	2.8753	25.55
M3	608.76	0.76543	0.74594	0.98789	2.7613	22.381
M4	589.12	0.75309	0.73257	0.98636	2.7454	21.98
M5	624.17	0.78481	0.75663	0.98918	2.7736	22.725
M6	484.61	0.7013	0.65822	0.97613	2.6513	19.789
M7	556.08	0.74074	0.70919	0.98323	2.717	21.256
M8	496.45	0.70513	0.66688	0.97742	2.6629	20.04

After performing similarity searching with different coefficients, the similarity values retrieved are not in the same range in terms of number of bit set where the values are large for certain coefficients and small for others. To standardize these similarity values, normalization was performed. There are many ways to achieve normalization. The process of normalization was achieved by calculating the z-scores. To calculate the z-scores, the standard deviation  $\sigma$  must be calculated first. The equation for z-scores follows:-

$$Z\_score = \frac{x - \mu}{\sigma} \quad (4.1)$$

Where

- $x$  is a raw score to be standardized
- $\sigma$  is the standard deviation
- $\mu$  is the mean

Table 4.4 shows the similarity value normalized after apply the Z-score equation.

**Table 4.4:** Similarity Values normalization with Z\_score

Tanim	Ruse	Simple	Bar	Cosine	Kuls
-0.277	-0.294	-0.2774	-0.27741	-0.2774	-0.2774
-0.282	-0.294	-0.2778	-0.27854	-0.2799	-0.2799
-0.284	-0.294	-0.2781	-0.27939	-0.2816	-0.2816
-0.284	-0.294	-0.2781	-0.2795	-0.2818	-0.2818
-0.284	-0.294	-0.278	-0.27929	-0.2814	-0.2814
-0.286	-0.294	-0.2783	-0.28011	-0.2831	-0.2831
-0.285	-0.294	-0.2782	-0.2797	-0.2822	-0.2822
-0.286	-0.294	-0.2782	-0.28004	-0.2829	-0.2829

The second equation were used in this project it is  $x'$  equation, normalization step is first consisted. It considered of on the normalization of the domain of the variables and it is applied before the file is partitioned (usual normalization in the (0, 1) interval was applied:  $x' = (x - \min) / (\max - \min)$ ).

$$x' = \frac{(x - \min)}{(\max - \min)} \quad (4.2)$$

Table 4.5 shows the similarity value normalized after apply the  $x'$  equation.

**Table 4.5:** Similarity Values normalization between (0 and 1)

Tanim	Ruse	Simple	Bar	Cosine	Kuls	Forbes
1	1	1	1	1	1	1
0.731595	0.816897	0.86422	0.871107	0.831863	0.830103	0.856126
0.582745	0.732388	0.765465	0.775528	0.716178	0.712649	0.711308
0.565132	0.718303	0.75308	0.762949	0.701247	0.697533	0.696109
0.596635	0.732388	0.777785	0.786261	0.72783	0.72452	0.735192
0.472096	0.619709	0.69135	0.693278	0.617361	0.612872	0.632375
0.535441	0.704218	0.728375	0.739739	0.675439	0.671477	0.658951
0.482659	0.633794	0.697543	0.701292	0.62727	0.622783	0.637086

<b>Fossum</b>	<b>Simpson</b>	<b>Pear</b>	<b>Yule</b>	<b>Stiles</b>	<b>Dennis</b>
1	1	1	1	1	1
0.732259	0.850004	0.828128	0.992371	0.923276	0.829599
0.571883	0.698968	0.709224	0.97405	0.861896	0.70836
0.55252	0.683132	0.693921	0.970772	0.853335	0.693019
0.587076	0.723839	0.721459	0.976814	0.868519	0.721521
0.449484	0.616668	0.608827	0.94885	0.802671	0.609196
0.519946	0.667283	0.667163	0.964065	0.838044	0.66532
0.461157	0.621583	0.618738	0.951615	0.808916	0.618798

The conceptual different between the Z-score and  $x'$  equations is that Z-score equation normally yielded negative value, while  $x'$  equation yields the values between 0 and 1 and it is settable for this project.

The first data represented here is composed of active compounds with different degree of the activity for each (1360 rows). As shown in chapter 3 Figure 3.3, each compound from each part selected as active target (query), all other compounds in other parts are assumed to be inactive, in order to determine the similarity. The procedure is repeated for other parts. The summarization of single coefficients of average percentage of the top 10% for all actives by using 10 different queries are shown in Table 4.6 is the percentages of actives obtained.



**Table 4.6:** Summary of Single Coefficient of Average Percentage Top 10% of all Actives

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES							Average
	Active 1	Active 2	Active 3	Active 4	Active 5	Active 6	Active 7	
	10%	10%	10%	10%	10%	10%	10%	
Jaccard/tanimoto	17	16	15	0.45	6	6	21.5	11.7
Russel/rao	17	3.9	13	1.35	10	17	20	11.8
<b>Simple</b>	<b>28</b>	<b>12</b>	<b>16</b>	<b>0.45</b>	<b>6</b>	<b>4.5</b>	<b>21</b>	<b>12.6</b>
Baroni	20	16	18	0.45	6	3	15.5	11.3
Ochiai/cosine	17	15	14	0.45	6	7.5	21.5	11.6
Kulczynski(2)	17	10	13	0.45	6	11	21	11.2
<b>Forbes</b>	<b>28</b>	<b>12</b>	<b>17</b>	<b>0.45</b>	<b>6</b>	<b>5</b>	<b>15.5</b>	<b>12</b>
Fossum	17	15	14	0.45	6	8	21.5	11.7
Simpson	23	3.9	12	0.9	7	18	21.5	12.3
Pearson	18	16	14	0.45	6	7.5	21	11.9
Yule	26	14	16	0.4	6	7.5	21.5	13.1
Stile	18	15	14	0.45	6	6	21	11.5
Dennis	20	<b>16</b>	14	0.45	6	6	21	11.9

The Table above shows the summary of single coefficient among the average percentage of top 10% of all actives. The Simple and Forbes are the best single coefficients for all 7 actives among the among the top 10% compounds. For details of single coefficient of percentage of active by using 10 targets see appendix B.

The second data represented here is composed from active compounds with different degree of the activity for each of them (1000 rows) as shown in Table 4.2. Table 4.7 shows the single coefficients percentages of actives obtained. For more details on single coefficient of percentage of active second data, see appendix C.

**Table 4.7:** The Average Percentage of all Actives (Second Data) Using Single Coefficient

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES			Average
	Active 1	Active 2	Active 3	
	10%	10%	10%	
Jaccard/tanimoto	29.96	14.2	11	18.4
Russel/rao	28.74	15	9.2	17.6
Simple	29.55	13.4	12	18.3
Baroni	30.77	13.8	11	18.5
Ochiai/cosine	30.36	14.2	11	18.5

<b>Kulczynski(2)</b>	<b>29.96</b>	<b>15</b>	<b>12</b>	<b>19</b>
Forbes	22.67	13	13	16.2
Fossum	30.36	14.2	11	18.5
Simpson	24.29	13.8	13	17
Pearson	30.36	14.2	11	18.5
<b>Yule</b>	<b>30.36</b>	<b>14.2</b>	<b>13</b>	<b>19.2</b>
Stile	30.36	15	11	18.8
Dennis	30.77	13.8	11	18.5

#### 4.5 Combination of Coefficients Using Fusion Process

The 13 groups discussed in Chapter 3 (Table 3.2) were used. Fusion was carried out using representative coefficients selected from each of the 13 groups based on the study of Salim (2003). 1360 molecules were selected from the database and matched in order to determine their similarities. After calculating similarity among target (query), the similarity values normalized from the coefficients were summed to give a new similarity for each compound when compared to a target and find the number of active at the top 10% on whole data they are percentage for each combination Table 4.8 show the percentage of active thus obtained for each combination.

The fusions between coefficients based on linear combinations of similarity values after normalizations from different similarity measures were used as a way to standardize the data. Although results of some test retrieval experiments have shown that the use of similarity values can give slightly better retrieval effectiveness than rank values, the fusion using similarity values can be applicable when sources combined have similar rank-similarity curves.

The fusion used was based on summing procedure which normalized similarity values produced by the searches. For each target structure, the combinations were sorted into decreasing order of the number of actives retrieved. Then the best combination of coefficients on top 10% is obtained. The SUM fusion function was used and found to be more effective and efficient. Table 4.8 below

shows the summarization of combinations of different selection of 2, 3 and 4 coefficients on different of all actives presented on Table 4.8. For more details information about the combination of different coefficients of the percentage of actives obtain for each actives with fusions 2, 3 coefficients, see appendix E.

Table 4.8 below shows the summarization of combinations of different selection of 2-coefficients (non-weighted) of all actives that presented on Table 4.1. As details the combination of different coefficients of the percentage of actives obtain for each actives with fusions 2, 3 coefficients, see appendix E.

**Table 4.8:** Summary of Fusion of 2-Coefficients and the Average Percentage of top 10% for all actives

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES							Average
	Active 1	Active 2	Active 3	Active 4	Active 5	Active 6	Active 7	
	10%	10%	10%	10%	10%	10%	10%	
BarFos	19	15	17	0.45	6	5.5	21	12
TanCos	17	16	15	0.45	6	7	22	11.9
CosSti	17	15	15	0.45	6	8	22	11.9
RusFor	16	10	15	0.45	6	11	21	11.4
TanRus	15	5.6	16	0.45	6	15	21	11.3
RusCos	15	5.2	17	0.45	6	15	22	11.5
RusSti	15	5.2	15	0.9	6	15	22	11.3
ForTan	27	14	17	0.45	6	4.5	19	12.6
<b>CosFor</b>	<b>27</b>	<b>16</b>	<b>17</b>	<b>0.45</b>	<b>6</b>	<b>4</b>	<b>21</b>	<b>13.1</b>
ForStil	15	15	17	0.45	6	4.5	20	11.1
<b>TanSimple</b>	<b>22</b>	<b>12</b>	<b>17</b>	<b>0.45</b>	<b>6</b>	<b>17</b>	<b>20</b>	<b>13.5</b>
CosFos	17	15	14	0.45	6	7.5	22	11.7
TanBar	19	16	17	0.45	6	4.5	21	12
RusSimple	16	15	14	0.45	6	18	20	12.8
RusKul	15	5.2	15	0.9	6	16	22	11.4
TanFos	17	16	15	0.45	6	7	22	11.9
RusFos	15	5.2	15	0.9	6	18	20	11.4
RusBar	14	11	15	0.45	6	12	21	11.4

From Table 4.8, it can be seen that the best combination for 2-coefficients fusion the Tanimoto and Simple matching coefficients (TanSimple) are satisfies the best average percentage for top 10% of all actives (7) by average value 13.5%.

Table 4.9 below is the summarization of combinations of different selection of 3-coefficients (non-weighted) of all actives. As details the combination of different coefficients of the percentage of actives obtain for each actives with fusions 3-coefficients, see appendix E.

**Table 4.9 :** Summary of Fusion of 3-Coefficients and the Average Percentage of top 10% for all actives

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES							Average
	Active 1	Active 2	Active 3	Active 4	Active 5	Active 6	Active 7	
	10%	10%	10%	10%	10%	10%	10%	
TanBarFos	18.8	14.7	15.79	0.45	6	5.5	21.5	11.8
<b>RusForCos</b>	17.7	11.3	17.29	0.45	6	14	21	<b>12.5</b>
RusForTan	17.3	12.1	14.29	0.45	6	14	21	12.2
RusForSti	17.7	11.3	16.54	0.45	6	14	21	12.4
RusTanCos	14.8	9.96	14.29	0.45	6	13	21	11.4
RusCosSti	14.8	9.09	15.04	0.45	6	13	20.5	11.3
ForTanCos	24	15.6	15.04	0.45	6	4	21	12.3
ForCosSti	24.7	14.3	15.04	0.45	6	4	21	12.2

From Table 4.9, it can be seen that the best combination for 3-coefficients fusion the Russell-Rao, Forbes and Cosine Simple coefficients (RusForCos) are satisfies the best average percentage for top 10% of all actives (7) by average value 12.5%.

The fusion between coefficients based on linear combinations by using ranking from different similarity measures was used, instead of similarity values, as a

way to standardize the data. Although results of some text retrieval experiments have shown that use of similarity values can give slightly better retrieval effectiveness than rank values, the fusion using similarity values is only appropriate when sources combined have the similar rank-similarity curves.

For each target structure, the combinations were sorted into descending order of the number of actives retrieved and assigned an ordinal value from 1(best ranking) down to last (worst ranking). Consequently, after calculating the similarity for each of them, their positions were ranked accordingly. The rank-positions from the coefficients were summed to give a new similarity ranking for each compound when compared to a target. The SUM fusion function was used as it was found to be the most effective. Table 4.10 shows similarity value after order descending and ranking positions, the rank-positions from the coefficients were summed to give a new similarity ranking for each compound when compared to a target.

**Table 4.10:** (a ,b ,c and d) shows similarity value after order descending and ranking positions, The rank-positions from the coefficients were summed to give a new similarity ranking for each compound when compared to a target.

(a) Shows the Tan,Rus,Simple and Bar coefficients after order descending and ranking positions

Molecules	Tan similarity values	Rank Position	Molecules	Rus similarity values	Rank Position	Molecules	Simple similarity values	Rank Position	Molecules	Bar similarity values	Rank Position
<b>M1</b>	1	1	<b>M1</b>	0.07	1	<b>M1</b>	1	1	<b>M1</b>	1	1
<b>M2</b>	0.756	2	<b>M2</b>	0.06	2	<b>M2</b>	0.979	2	<b>M2</b>	0.936	2
<b>M5</b>	0.633	3	<b>M3</b>	0.05	3	<b>M5</b>	0.966	3	<b>M5</b>	0.895	3
<b>M3</b>	0.62	4	<b>M5</b>	0.05	4	<b>M3</b>	0.964	4	<b>M3</b>	0.889	4
<b>M27</b>	0.62	5	<b>M27</b>	0.05	5	<b>M27</b>	0.964	5	<b>M27</b>	0.889	5
<b>M4</b>	0.604	6	<b>M4</b>	0.04	6	<b>M4</b>	0.962	6	<b>M4</b>	0.883	6
<b>M7</b>	0.577	7	<b>M7</b>	0.04	7	<b>M7</b>	0.958	7	<b>M7</b>	0.872	7
<b>M8</b>	0.529	8	<b>M10</b>	0.04	8	<b>M8</b>	0.954	8	<b>M8</b>	0.853	8

(b) Shows the Cos,Kuls,Forbes and Fossum coefficients after order descending and ranking positions

Molecules	Cos similarity values	Rank Position	Molecules	Kuls similarity values	Rank Position	Molecules	For similarity values	Rank Position	Molecules	Fos similarity values	Rank Position
<b>M1</b>	1	1	<b>M1</b>	1	1	<b>M1</b>	13.04	1	<b>M1</b>	1043	1
<b>M2</b>	0.861	2	<b>M2</b>	0.861	2	<b>M2</b>	11.51	2	<b>M2</b>	771.4	2
<b>M5</b>	0.775	3	<b>M5</b>	0.775	3	<b>M5</b>	10.23	3	<b>M5</b>	624.2	3
<b>M3</b>	0.765	4	<b>M3</b>	0.765	4	<b>M3</b>	9.979	4	<b>M3</b>	608.8	4
<b>M27</b>	0.765	5	<b>M27</b>	0.765	5	<b>M27</b>	9.979	5	<b>M27</b>	608.8	5
<b>M4</b>	0.753	6	<b>M4</b>	0.753	6	<b>M120</b>	9.899	6	<b>M4</b>	589.1	6
<b>M7</b>	0.732	7	<b>M7</b>	0.732	7	<b>M4</b>	9.818	7	<b>M7</b>	556.1	7
<b>M8</b>	0.692	8	<b>M8</b>	0.692	8	<b>M1107</b>	9.545	8	<b>M8</b>	496.5	8

(c) Shows the Simpson,Pear,Yule and Stiles coefficients after order descending and ranking positions

Molecules	Simp similarity values	Rank Position	Molecules	Pear similarity values	Rank Position	Molecules	Yul similarity values	Rank Position	Molecules	Stil similarity values	Rank Position
<b>M1</b>	1	1	<b>M1</b>	1	1	<b>M1</b>	1	1	<b>M1</b>	3.018	1
<b>M2</b>	0.883	2	<b>M2</b>	0.85	2	<b>M2</b>	0.996	2	<b>M2</b>	2.875	2
<b>M5</b>	0.785	3	<b>M5</b>	0.757	3	<b>M5</b>	0.989	3	<b>M5</b>	2.774	3
<b>M3</b>	0.765	4	<b>M3</b>	0.746	4	<b>M3</b>	0.988	4	<b>M3</b>	2.761	4
<b>M27</b>	0.765	5	<b>M27</b>	0.746	5	<b>M27</b>	0.988	5	<b>M27</b>	2.761	5
<b>M4</b>	0.759	6	<b>M4</b>	0.733	6	<b>M120</b>	0.986	6	<b>M4</b>	2.745	6
<b>M7</b>	0.753	7	<b>M7</b>	0.709	7	<b>M4</b>	0.983	7	<b>M7</b>	2.717	7
<b>M8</b>	0.741	8	<b>M8</b>	0.667	8	<b>M1107</b>	0.977	8	<b>M8</b>	2.663	8

(d) Shows the Dennis coefficient after order descending and ranking position and summation on ranking position.

Molecules	Den similarity values	Rank Position	Sum of Ranking	Rank Position
<b>M1</b>	30	1	<b>13</b>	1
<b>M2</b>	25.55	2	<b>26</b>	2
<b>M5</b>	22.73	3	<b>40</b>	3
<b>M3</b>	22.38	4	<b>47</b>	4
<b>M27</b>	22.38	5	<b>65</b>	5
<b>M4</b>	21.98	6	<b>80</b>	6
<b>M7</b>	21.26	7	<b>94</b>	7
<b>M8</b>	20.04	8	<b>119</b>	8

The single coefficient in Table 4.11 shows the percentage of actives by using ranking positions.

**Table 4.11:** The Average Percentage of all Actives (Second Data) Using Single Coefficient and Ranking Positions

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES			Average
	Active 1	Active 2	Active 3	
	10%	10%	10%	
Jaccard/tanimoto	17	0.862	14.9	10.9
Russel/rao	17.3	2.586	12.7	10.9
Simple	28.8	0	15.7	14.8
Baroni	20.3	0.431	17.9	12.9
Ochiai/cosine	16.6	0.431	14.2	10.4
Kulczynski(2)	17	0.431	12.7	10
<b>Forbes</b>	<b>28.4</b>	<b>0.862</b>	<b>17.2</b>	<b>15.5</b>
Fossum	16.6	0.431	14.2	10.4
Simpson	23.2	0.862	11.9	12
Pearson	17	0.431	14.2	10.5
Yule	26.2	0.431	15.7	14.1
Stile	17.7	0.431	14.2	10.8
Dennis	19.6	0.431	14.2	11.4

From Table 4.11, it can be seen that the best single coefficient by using ranking position the Forbes coefficients is satisfies the best average percentage for top 10% of all actives by average value 15.5%.

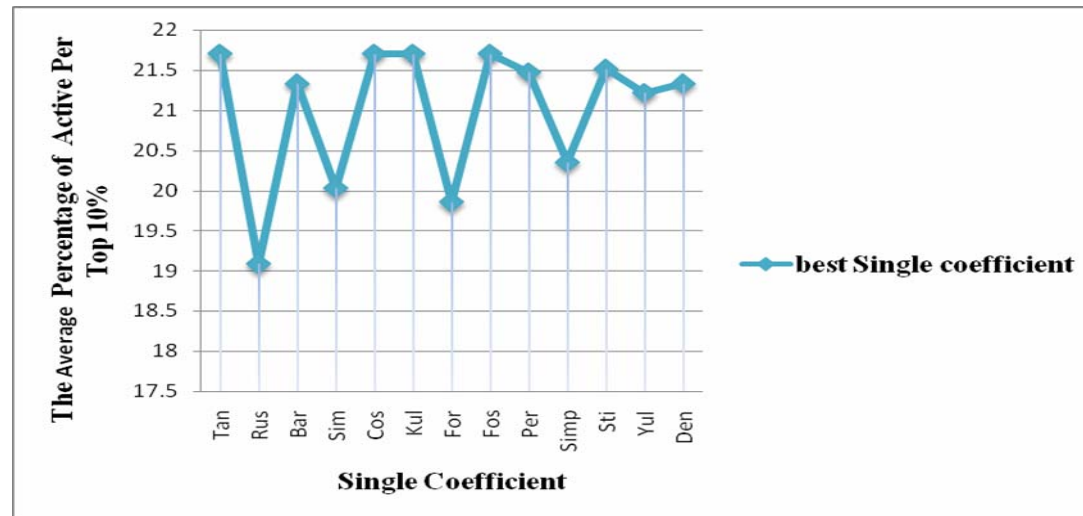
In this section, 10 different queries (targets) non-weights were taken for each Active top10% average of actives and it was used to obtain the best single coefficients. For more details about single coefficients with 10 different queries and the percentage of actives obtains for each active, see appendix E.

Table 4.12 and Figure 4.3 below show the best single coefficient of active top 10% their selected similarity.



**Table 4.12:** The Average Percentage of all Active top 10% of single coefficient

Active	THE PERCENTAGE OF ACTIVES												
	Tan	Rus	Bar	Sim	Cos	Kul	For	Fos	Per	Simp	Sti	Yul	Den
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
Active 1	15.9	13.9	18.1	25.8	16.3	17.2	26.3	16.2	17	21.7	17	22.4	18.3
Active 2	24.5	21	25	21.9	24.4	23.8	24	24.2	24.5	24.8	24.4	24.5	24.6
Active 3	22.5	17.5	23	20.7	22.2	21.4	21	22.2	21.6	17.3	21.5	20.3	21.6
Active 4	19.2	15.4	19.2	14.3	19	18.9	15.3	19.1	18.8	15.4	18.8	18.8	18.4
Active 5	19.9	24.6	16.2	11.8	19.4	17.7	9.4	19.3	18.4	12.2	18.5	13.2	16.3
Active 6	13.9	10.4	12.2	13.1	14.9	17.6	13	15.2	14.4	19.8	14.6	14.4	14.2
Active 7	35.7	30.8	35.6	32.6	35.4	35.6	30	35.4	35.6	31.3	35.8	34.9	35.9
Average	21.7	19.1	21.3	20	21.7	21.7	19.9	21.7	21.5	20.4	21.5	21.2	21.3

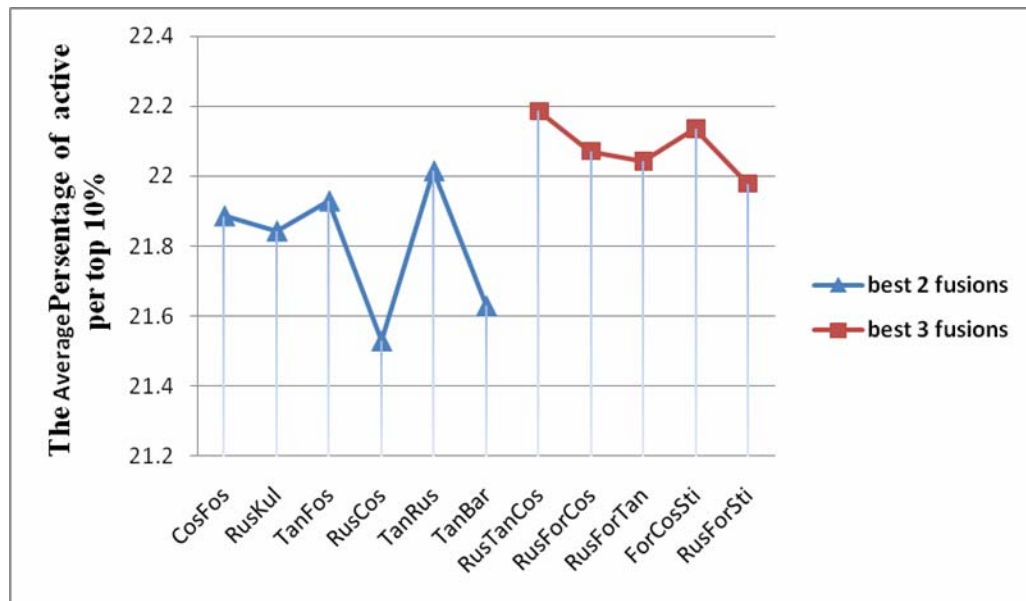
**Figure 4.3** The Average percentage of top 10% Actives versus the single coefficients

In Table 4.12 and Figure 4.3, it clearly shows that Tanimoto, Cosine, Kulczynski(2) and Fossum have the best average percentage of top 10% among all the Actives (7). The single coefficient has an average value 21.7.

Table 4.13 and Figure 4.4 below shows the Summarization of the Average Percentage of all Active among the top 10% on different combination of non-weights coefficients.

**Table 4.13:** Summarization the Average Percentage of all Active top 10% depends on fusions of coefficients.

Actives	THE PERCENTAGE OF ACTIVES											
	Fusion2						Fusion3					Fusion4
	CosFos	RusKul	TanFos	RusCos	TanRus	TanBar	RusTanCos	RusForCos	RusForTan	ForCosSti	RusForSti	TanBarCos Kul
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
<b>Active 1</b>	14.6	14.7	14.5	15.5	<b>14.6</b>	16	<b>15.1</b>	15.1	15.6	15.1	15.1	17
<b>Active 2</b>	22.3	22.3	22.7	24.7	<b>22.5</b>	24.3	<b>22.7</b>	23	22.6	22.9	22.6	24.4
<b>Active 3</b>	20.7	20.4	19.3	22.9	<b>19.5</b>	22.2	<b>21.7</b>	21.9	20.9	21.7	21.1	21.9
<b>Active 4</b>	17.8	18	18.3	19.8	<b>18.8</b>	19.5	<b>19.6</b>	19.4	19.4	20	19.3	19.7
<b>Active 5</b>	23.3	23	22.7	18.9	<b>22.8</b>	19.6	<b>21.6</b>	21.5	21.5	21.8	21.5	19
<b>Active 6</b>	19.7	19.8	20.6	13.3	<b>21.1</b>	14.2	<b>19</b>	18.5	19.2	18.5	19.3	14.3
<b>Active 7</b>	34.8	34.7	35.4	35.6	<b>34.8</b>	35.6	<b>35.6</b>	35.1	35.1	34.95	34.9 5	35.1
<b>Average</b>	<b>21.89</b>	<b>21.84</b>	<b>21.93</b>	<b>21.53</b>	<b>22.01</b>	<b>21.63</b>	<b>22.19</b>	<b>22.1</b>	<b>22</b>	<b>22.14</b>	<b>21.98</b>	<b>21.63</b>



**Figure 4.4** The Average Percentage of Top 10% Actives versus the Fusions of Coefficients

In Figure 4.4, it clearly shows that the best combinations of coefficient for 2-coefficient fusions is the Tanimoto and Russell-Rao(TanRus). It has the best percentage of top 10% among the all 7 Actives and average value of 22.01%. The worst is Russell-Rao & Cosine (RusCos) with an average of 21.53%. The graph was shows that the best combination for 3-coefficients fusion is obtain from Russell-Rao, Tanimoto and Cosine (RusTanCos), having the best percentage of top 10% for all 7 Actives and an average value of 22.19%. The worst is Russell-Rao,Forbes and Stile(RusForSti) which both have an average of 21.98%.

#### **4.6 Optimization Result of Similarity Coefficients Fusions by Using GA Weights**

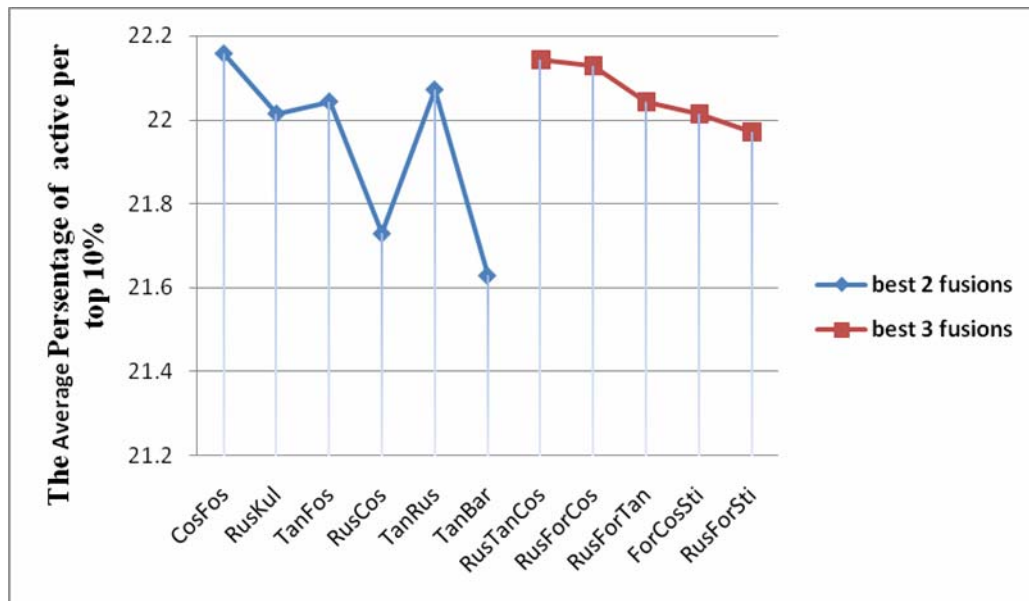
In this part, different combination or fusion of coefficients were used with the GA combinations weights among the top 10% of each Active. For GA validation, several single coefficients and combinations of non-weighted coefficients are generated. The generated fusions were compared with GA optimized fusions of top 10% of each Active see appendix F.

In this section, 10 different queries were taken and the averages of all Actives top10% were used to obtain the best combinations of coefficients. Genetic algorithm (GA) is used to combine the coefficients with weights.

The summary combinations of coefficients with weights that generated by GA as shown in Table 4.14 and Figure 4.5 below are the average percentage of actives obtains on top 10%.

**Table 4.14:** The Average Percentage of all Active top 10% on GA based fusions of coefficients (GA weights)

		<b>THE PERCENTAGE OF ACTIVES</b>											
		<b>Fusion2</b>					<b>Fusion3</b>					<b>Fusion 4</b>	
<b>Fusion of Coefficients</b>	<b>Weights</b>	<b>CosFos</b>	<b>RusKul</b>	<b>TanFos</b>	<b>RusCos</b>	<b>TanRus</b>	<b>TanBar</b>	<b>RusTanCos</b>	<b>RusForCos</b>	<b>RusForTan</b>	<b>ForCosSti</b>	<b>RusForSti</b>	<b>TanBarCosKul</b>
		0.960 and 0.937	0.972 and 0.960	0.960 and 0.960	0.972 and 0.960	0.960 and 0.972	0.960 and 0.1440	0.972, 0.960 and 0.960	0.972, 0.960 and 0.960	0.972, 0.960 and 0.960	0.960, 0.972 and 0.960	0.972, 0.960 and 0.960	0.972, 0.1440, 0.960 and 0.960
<b>Actives</b>	<b>10%</b>	<b>10%</b>	<b>10%</b>	<b>10%</b>	<b>10%</b>	<b>10%</b>	<b>10%</b>	<b>10%</b>	<b>10%</b>	<b>10%</b>	<b>10%</b>	<b>10%</b>	<b>10%</b>
<b>Active 1</b>	<b>14.8</b>	14.8	14.3	16.8	14.6	16	15.1	15.1	15.6	15.1	15	17	
<b>Active 2</b>	<b>22.3</b>	22.4	22.7	24.7	22.5	24.3	22.6	23	22.6	22.8	22.6	24.4	
<b>Active 3</b>	<b>20.6</b>	20.4	19.4	23.1	19.6	22.2	21.6	21.9	20.9	21.7	20.7	21.9	
<b>Active 4</b>	<b>18.8</b>	19	19.1	19.8	18.8	19.5	19.4	19.4	19.4	19.3	19.3	19.7	
<b>Active 5</b>	<b>23.3</b>	23	22.8	18.9	22.9	19.5	21.7	21.9	21.5	21.8	21.6	19	
<b>Active 6</b>	<b>19.7</b>	19.8	20.6	13.3	21.2	14.2	19	18.5	19.2	18.4	19.3	14.3	
<b>Active 7</b>	<b>35.6</b>	34.7	35.4	35.5	34.9	35.7	35.6	35.1	35.1	35	35.3	35.5	
<b>Average</b>	<b>22.16</b>	<b>22.01</b>	<b>22.04</b>	<b>21.73</b>	<b>22.1</b>	<b>21.6</b>	<b>22.14</b>	<b>22.13</b>	<b>22.04</b>	<b>22.01</b>	<b>21.97</b>	<b>21.7</b>	



**Figure 4.5** The Average Percentage of Top 10% Actives versus the GA based fusions of coefficients (GA weights)

In Figure 4.5, it clearly shows the best combinations of coefficients by using GA based fusion weights that for 2-Coefficients fusions Cosine and Fossum (CosFos) are satisfies the best percentage of top 10% Actives (7) by average value 22.16%. and the worst is Tanimoto and Baroni (TanBar) by 21.6%. And for the best combination of 3-coefficients that Russell-Rao, Tanimoto and Cosine (RusTanCos) are satisfies the best percentage of top 10% Actives (7) by average value 22.14% and the worst Russell-Rao, Forbes and Stile (RusForSti) by 21.97%. Therefore, instead of using combinations of 3-coefficients fusion for the best value we can use combination of 2-Coefficients fusion as shown in the peak points. Comparisons for different coefficient fusions were carried out. The result show that the Tanimoto, Cosine, Kulczynski(2) and Fossum coefficients are the best single coefficient. A cosine and Fossum coefficient yields the best combination for 2-coefficient fusion with the weights of 0.960 and 0.937 respectively. For 3-coefficient fusion Russell-Rao, Tanimoto and Cosine coefficients of weightings 0.972, 0.960 and 0.960 respectively gives the best result. The combinations Tanimoto and Cosine coefficients perform well and eventually results in large number of actives. Using combination with weights ranging between 0.0 and 1.0 generated by genetic algorithm, gave the results in better number of active than the non-weighted

combination. Cosine and Fossum coefficients combined without weights yields an average 21.89% among the top 10% compound; whereas when genetic algorithm (GA) is used to combine Cosine and Fossum Coefficients with weights of 0.960 and 0.937 respectively, an average of 22.16% among the top 10% compound is obtained.

## CHAPTER 5

### CONCLUSION

#### 5.1 Introduction

This chapter primarily focuses on general findings of the results, suggestion and recommendation for future works and overall summary about this project. Preliminary result on the retrieval chemical compounds from chemical databases using different similarity measures based on different similarity coefficients and molecular representations have been discussed.

Firstly, after applying retrieval from chemical database to find the active compounds using similarity coefficient, it was found that Tanimoto, Cosine, Kulczynski(2) and Fossum coefficient are the best single coefficient. The Russell-Rao, Forbes, Simple matching and Simpson coefficient were found to be the worst. This clearly indicates that the similarity measures between molecules depend on different factor and each factor has different of degree of effect. As such, the need for combination with other coefficients is essential, therefore; fusion will get better similarity.



Secondly, the power of GA is that it provides diversity of visible solutions combination which will eventually leads to high accurate result for the problem at hand .However; GA needs a proper parameter tuning to achieve a good result.

GA is used to find the best combination of coefficients. The result is based on the input data which are the similarity value normalized or ranking position and the output data which are the number of actives that represents coefficient and combination of several coefficients based on the number of actives yield for each coefficient.

It has been proven that the best combination can be satisfied by using 2-coefficients fusion Cosine and Fossum coefficients with weights (0.960 and 0.937), and the same for 3-coefficients fusion Russell-Rao , Tanimoto and Cosine with weights (0.972, 0.960 and 0.960).

Finally, the result of this research shows that GA optimizes a combination which increases the number of active that strengthens accuracy of the solution. By comparing GA weighting approach with non-weighting, it was found that the GA improved the result up to 10% on average for all the 7 actives. GA locates suitable weights by 150 generations with a little improvement in weights achieved by 1500 generation.

## **5.2 Recommendation**

Based on the above facts, it is recommended that different sets of GA parameters tuning is tested. For instance, by introducing the concept of migrations that will share population establishment by loading it into different machine. In addition, using different method for crossover (uniform) operator keeps the probability between (0.60 – 0.90), will subsequently enhance the searching capability for suboptimal weights.

More input data can be considered to find more effective result in training and testing data using GA such as the size of database and number of actives in database.

### **5.3 Project Advantage**

The followings are some advantages that can be found in this project:

- i. To give some exposure to other researcher on applying genetic algorithm in finding ideal weights for combination of coefficients to be used in performing similarity searching.
- ii. It has clearly demonstrated that it is absolutely possible to use genetic algorithm (GA) in determining the ideal weights to be in performing similarity searching.
- iii. To give ideas to other researcher to do more research on the potential of intelligent techniques in the field of chemoinformatics.

### **5.4 Summary**

Overall, this project meets the aim, objective and scope that have been outlined. Genetic algorithm can be used for optimizing combination of similarity measures for chemical database retrieval and perform similarity searching. This is done when there are many selection of good coefficient to use and need to find the most perfect and suitable coefficient to be used.

## REFERENCES

- Allen F.H., Davies J.E., Galloy J.J., Johnson O., Kennard O., Macrea C.F., Mitchell E.M., Mitchell G.F., Smith J.M. and Watson D.G. (1991) *J. Chem. Inf. Computer. Science*, 31, 187–204.
- Ash, J., Chubb, P., Ward, S., Welford, S., & Willett, P. (1985). *Communication, storage, and retrieval of chemical information*. Chichester, England: Ellis Horwood Limited.
- Barnard, J.M. (1993). "Substructure searching methods: Old and new", *Journal of Chemical Information and Computer Science*, 33, 532-538.
- Brown, R. D. *Perspect. Drug DiscoV. Des.* 1996, 7/8, 31
- 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.
- Craig P.N. and Ebert, H.M. (1969). "Eleven years of structure searching using the SKF (Smith, Kline, and French) fragment codes, *journal of chemical documentation*, 9, 141-146.
- Dasarthy, *Decision Fusion*, IEEE Computer Society Press, (1994 )
- Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. (1983), *The CAS Online Search System. 1. General System Design and Selection, Generation, and Use of Search Screens*. *J. Chem. Inf. Computer. Science*. 23, 93.

- Ricketts, E. (1993). Comparison of Conformations of Small Molecule Structures from the Protein Data Bank with Those Generated by Concord, Cobra, C ChemDBS-3D, and Converter and Those Extracted from the Cambridge Structural Database, American Chemical Society, pp. 905-925 vol.33 No.6.
- Ellis, D.; Furner-Hines, J.; Willett, P. *Perspect.* (1994) Measuring the degree of similarity between objects in text retrieval systems. *Inf. Manag.* 3, 128-149.
- Fisanick, W. et al., (1992) "Similarity Searching on CAS Registry Substances 1: Global . . ." *Journal of Chemical Information & Computer Sciences*, vol. 32, No. 6, pp. 664-674.
- Garey M. R, Johnson (1977). The Rectilinear Steiner Tree Problem is NP-Complete, and D. S. Johnson, *SIAM J. Appl. Math.*, 32.
- Ginn, C.M.R, Turner D.B., Willett, P., (1997). "Similarity Searching in Files of Tree-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion", *Journal of Chemical Information and Computer Science*, 37, pp. 23-37.
- Ginn, C.M.R., Willett, P. and Bradshaw, J. *Perspect* (2000) "Combination of molecular similarity measures using data fusion." *Perspectives in Drug Discovery and Design*, submitted for publication. 20, 1-6
- Goldberg D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading: Addison-Wesley.
- Goldberg D. E. and Dep K. (1991). A Comparative Analysis on Selection Schemes Used in Genetic Algorithms. *Foundations of Genetic Algorithms*, Rawlins G., ed. Morgan Kaufmann. 69 - 93.
- Hall, D.L. (1992) *Mathematical Techniques in Multisensor Data Fusion*. Northwood, MA: Artech House.
- Hall. L. B. Kier (1999). *Molecular Structure Description: The Electrotopological State and L.*, Academic Press, San Diego,

- Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*, University of Michigan Press.
- Holliday, J. D.; Hu, C.-Y.; Willett, P. Comb. (2002)Chem. High Throughput Screening, 5, 155-166.
- [http://en.wikipedia.org/wiki/Chemical\\_database](http://en.wikipedia.org/wiki/Chemical_database) (2007)
- <http://www.sic.rma.ac.be/Research/Fusion/Intro/content.html> (1995)
- Johnson MA(1990), Maggiora GM: *Concepts and Applications of Molecular Similarity*. New York: John Wiley,.
- Kearsley, Simon K.; Sallamack, Susan; Fluder, Eugene M.; Andose, Joseph D.; Mosley, Ralph T.; Sheridan, Robert P. (1996). *Chemical Similarity Using Physiochemical Property Descriptors*. *Journal of Chemical Information and Computer Sciences*. 36 (1): 118-127. ISSN: 0095-2338; CODEN: JCISD8.
- Kei ITO, N.TANAKA and S. FUJITA (2005). *Development and Application of XyM2Mol System for Converting Structural Data by XyM Notation into Connection Tables*. *Journal of Computer Chemistry, Japan*, Vol. 4, No. 3
- Kunal Roy (2004) *Topological descriptors*. *Molecular Diversity*, 8: 321–323, Drug Theoretics & Cheminformatics Lab, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India
- Leach A, Gillet V(2003), ‘*An Introduction to Chemoinformatics*’, Kluwer Academic Publishers
- Linn R. J., Hall D. L. (1991) *A Survey of Multi-sensor Data Fusion Systems*, In: *Proceedings of the SPIE - The International Society for Optical Engineering SPI*, Orlando, Vol. 1470, 13-29
- Martin, Y.C. and Willett, P. (1997) (editors) *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*. American Chemical Society, Washington.

- Morgan, H. L. (1965). The Generation of a Unique Machine Description for Chemical Structures - a Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* 5, 107-113)
- MacCuish, J.D. & MacCuish, N.E. (2003). Clustering Ambiguity and Binary Descriptors. MESA Analytics and Computing, New York.
- Ng, Kantor et al ( 1998). active data fusion in IR: A pilot study (context)
- Ramaswamy Nilakantan,\* Norman bauman,\* j. Scott dixon,t and R. Venkataraghavan Lederle Laboratories, Pearl River, New York (1987)
- Salim, N, (2002). Analysis and Comparison of Molecular Similarity Measures, University of Sheffield: Ph. D Thesis.
- Salim, N., Holliday, J., and Willet, P., "Combination of Fingerprint-Based Similarity Coefficient Using Data Fusion", *Journal of Chemical Information and Computational Science*, 2003, 43 pp. 435-442.
- Shemetulskis, N.E.; Weininger, D.; Blankley, C. J.; Yang. J. J.; Humblet, C. Stigmata: an algorithm to determine structural commonalties in diverse datasets. *J. Chem. Inf. Comput. Sci.* 1996, 36, 862-871.
- Sneath PHA and Sokal RR (1973) *Numeric taxonomy: the principles and practice of numerical classification.* W.H. Freeman, San Francisco, 573 pp.
- Tarjan, R. E. (1977). In. "Algorithms for Chemical Computation"; Christoffersen., R. E., Ed.; American Chemical Society: Washington D.C.,
- Vander Stouw, Elliott, P.M., and Isenberg, A.C. (1974). "Automated conversion of chemical substance names to atom-bond connection tables", *Journal of Chemical documentation* 14, 185-193.
- Wang, R.; Wang, S. *J. Chem. Inf. Computer. Science.* 2001, 41, 1422-1426.
- Willet, P. (1980). "Nomenclature processing and the interconversion of chemical structure representations", *Proc. Of the CNA (UK) seminar on chemical Structure Searching of the published literature, chemical notation association*

(UK), London.

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

Willett, P., Winterman, V., Bawden, D. (1986) "Implementation of nearest neighbour searching in an online chemical structure search system 26, 36-41.

Willett, P.; Barnard, J. M.; Downs, G. M. J. *Chem. Inf. Computer. Science*.1998, 38, 983-996.

Willett, Peter (2000)."Textual and chemical information processing: different domains but similar algorithms" *Information Research*.

Willett, P. (2003).” *Structural Biology in Drug Metabolism and Drug Discovery*”, Biochemical Society. Volume 31, 603-606.

Willett, P. (2006).” *Similarity-based virtual screening using 2D fingerprints and Drug Discovery*”, Biochemical Society. Krebs Institute for Bimolecular Research and Department of Information Studies, University of Sheffield, 211 Portobello, Sheffield S1 4DP 11(23-24):1046-53.

Dean, (1999).*Molecular DiVersity in Drug Design*; P. M.; Lewis, R. A., Eds.;Kluwer: Amsterdam.

Dauat, N.(2004) *Finding best coefficient and fusion of coefficients for similarity searching using neural network algorithms*. University Technology Malaysia (U.T.M): MS Thesis

Welmina P,(2004) *Comparison of the effectiveness of probability model with vector space model for compound similarity searching*. University Technology Malaysia (U.T.M): MS Thesis

**APPENDIX A**

**DETAILS ON MDDR DATA ACTIVITY**



**Table 1: GROUPS AND ACTIVITIES OF THE DATA**

<b>S.No</b>	<b>Activity</b>	<b>No. molecules</b>
1	<b>Interacting on 5HT receptor</b>	
	5HT Antagonists	48
	5HT1 agonists	66
	5HT1C agonists	57
	5HT1D agonists	100
2	<b>Antidepressants</b>	
	Mao A inhibitors	71
	Mao B inhibitors	161
3	<b>Antiparkinsonians</b>	
	Dopamine (D1) agonists	32
	Dopamine (D2) agonists	102
4	<b>Antiallergic/antiasthmatic</b>	
	Adenosine A3 antagonists	73
	Leukotine B4 antagonists	150
5	<b>Agents for Heart Failure</b>	
	Phosphodiesterase inhibitors	100
6	<b>AntiArrhythmics</b>	
	Potassium channel blockers	100
	Calcium channel blockers	100
7	<b>Antihypertensives</b>	
	ACE inhibitors	100
	Adrenergic (alpha 2) blockers	100
	<b>TOTLE</b>	<b>1360</b>

**APPENDIX B**

**PERCENTAGE OF ACTIVES USING DIFFERENT SINGLE COEFFICIENT**

**Table 1:** The percentage of Active 1 using single coefficient

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
Jaccard/tanimoto	17	36	51	76	97	100
Russel/rao	17	24	37	57	89	100
Simple	28	43	58	76	97	100
Baroni	20	38	54	80	99	100
Ochiai/cosine	17	37	55	76	97	100
Kulcznski(2)	17	38	55	75	97	100
Forbes	28	43	57	80	99	100
Fossum	17	37	55	76	97	100
Simpson	23	41	53	73	93	100
Pearson	18	38	55	78	98	100
Yule	26	41	58	80	99	100
Stile	18	37	55	78	98	100
Dennis	20	39	55	78	98	100

**Table 2:** The percentage of Active 2 using single coefficient

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
Jaccard/tanimoto	16	25	30	48	88	100
Russel/rao	3.9	20	30	52	88	100
Simple	12	30	39	49	85	100
Baroni	16	26	31	48	87	100
Ochiai/cosine	15	23	30	47	88	100
Kulcznski(2)	10	22	29	45	88	100
Forbes	12	25	34	49	87	100
Fossum	15	24	30	47	88	100
Simpson	3.9	21	31	52	88	100
Pearson	16	24	30	47	88	100
Yule	14	23	29	47	89	100
Stile	15	24	29	47	89	100
Dennis	16	24	30	48	89	100

**Table 3:** The percentage of Active 3 using single coefficient

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
Jaccard/tanimoto	15	37	53	65	83	100
Russel/rao	13	38	44	62	87	100
Simple	16	29	39	58	83	100
Baroni	18	35	47	66	84	100
Ochiai/cosine	14	39	52	67	83	100
Kulcznski(2)	13	39	47	67	83	100
Forbes	17	32	47	66	84	100
Fossum	14	39	51	67	83	100
Simpson	12	38	43	62	87	100
Pearson	14	37	53	66	83	100
Yule	16	37	51	66	84	100
Stile	14	38	54	65	83	100
Dennis	14	38	54	65	83	100

**Table 4:** The percentage of Active 4 using single coefficient

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
Jaccard/tanimoto	0.45	0.9	1.79	8.52	61	100
<b>Russel/rao</b>	<b>1.35</b>	2.24	4.93	21.5	67.7	100
Simple	0.45	1.35	3.14	17.9	62.8	100
Baroni	0.45	0.45	1.35	8.07	59.6	100
Ochiai/cosine	0.45	0.9	1.79	8.07	61.4	100
Kulcznski(2)	0.45	0.45	1.79	8.52	61.9	100
Forbes	0.45	1.35	2.24	13.5	60.5	100
Fossum	0.45	0.9	1.79	8.52	61	100
Simpson	0.9	1.35	2.24	8.07	59.6	100
Pearson	0.45	0.45	1.79	7.17	60.5	100
Yule	0.4	0.45	1.79	8.52	58.3	100
Stile	0.45	0.45	1.79	7.17	60.5	100
Dennis	0.45	0.45	1.35	7.62	60.1	100

**Table 5:** The percentage of Active 5 using single coefficient

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
Jaccard/tanimoto	6	16	31	50	66	100
Russel/rao	10	28	47	54	90	100
Simple	6	13	17	25	37	100
Baroni	6	16	24	43	60	100
Ochiai/cosine	6	16	31	50	67	100
Kulcznski(2)	6	16	31	50	61	100
Forbes	6	13	18	31	57	100
Fossum	6	16	31	50	67	100
Simpson	7	19	32	51	58	100
Pearson	6	16	27	47	61	100
Yule	6	16	27	47	61	100
Stile	6	14	20	42	57	100
Dennis	6	16	25	46	60	100

**Table 6:** The percentage of Active 6 using single coefficient

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
Jaccard/tanimoto	6	20	26.5	44	71.5	100
Russel/rao	17	27	30	42.5	76	100
Simple	4.5	9	15	34.5	64.5	100
Baroni	3	11.5	24.5	44	71	100
Ochiai/cosine	7.5	22	28	42.5	70	100
Kulcznski(2)	11	22.5	29.5	42.5	69	100
Forbes	5	8	14	37	70	100
Fossum	8	22	28	43	69.5	100
Simpson	18	25	32	43	68.5	100
Pearson	7.5	19.5	27	43	69	100
Yule	7.5	19.5	27	43.5	69	100
Stile	6	12.5	25	45.5	73	100
Dennis	6	20	26.5	44	71.5	100

**Table 7:** The percentage of Active 7 using single coefficient

<b>SINGLE COEFFICIENT</b>	<b>THE PERCENTAGE OF ACTIVES</b>					
	<b>10%</b>	<b>20%</b>	<b>30%</b>	<b>50%</b>	<b>80%</b>	<b>100%</b>
Jaccard/tanimoto	21.5	20	60	49	75	100
Russel/rao	20	20	52	42	74	100
Simple	21	23	69	58	87	100
Baroni	15.5	22	58	52	76	100
Ochiai/cosine	21.5	22	60	51	75	100
Kulcznski(2)	21	22	62	53	79	100
Forbes	15.5	25	72	58	85	100
Fossum	21.5	21	60	51	75	100
Simpson	21.5	25	72	58	85	100
Pearson	21	22	60	53	78	100
Yule	21.5	24	71	55	82	100
Stile	21	22	61	53	78	100
Dennis	21	23	62	54	78	100

**APPENDIX C****PERCENTAGE OF ACTIVES USING DIFFERENT SINGLE COEFFICIENT  
SECOND DATASET AND RANKING POSITION**

**Table 1:** The Percentage of Active 1(Second Data) Using Single Coefficient

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
Jaccard/tanimoto	29.96	34	41.7	64.8	87.9	100
Russel/rao	28.74	34.4	44.5	68.8	90.7	100
Simple	29.55	37.2	41.3	50.2	76.1	100
Baroni	30.77	35.2	40.9	61.5	84.2	100
Ochiai/cosine	30.36	34.8	40.1	62.3	85	100
Kulcznski(2)	29.96	34.4	40.5	60.3	83	100
Forbes	22.67	33.6	40.1	49.4	76.9	100
Fossum	30.36	34.8	40.1	62.3	85.4	100
Simpson	24.29	34.8	39.3	48.6	75.7	100
Pearson	30.36	34.8	40.9	60.7	83.4	100
Yule	30.36	35.2	40.9	51	79.8	100
Stile	30.36	34.8	40.1	60.7	83.4	100
Dennis	30.77	35.2	40.9	59.5	83	100

**Table 2:** The percentage of Active 2(second data) using single coefficient.

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
Jaccard/tanimoto	14.2	26.5	39.1	61.26	86.2	100
Russel/rao	15	23.3	37.5	61.26	87	100
Simple	13.4	23.7	34.4	59.29	86.6	100
Baroni	13.8	26.1	38.7	59.29	87	100
Ochiai/cosine	14.2	26.9	38.3	60.08	86.6	100
Kulcznski(2)	15	27.7	39.1	61.26	87	100
Forbes	13	24.5	38.3	58.89	87	100
Fossum	14.2	26.9	37.9	60.08	87	100
Simpson	13.8	26.1	37.9	59.68	84.6	100
Pearson	14.2	26.9	38.3	61.66	87.7	100
Yule	14.2	26.5	38.3	61.66	87.4	100
Stile	15	26.1	38.7	59.29	87	100
Dennis	13.8	26.5	38.7	61.26	88.1	100

**Table 3:** The Percentage of Active 3(Second Data) Using Single Coefficient.

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
Jaccard/tanimoto	11	20	30	49	75	100
Russel/rao	9.2	20	26	42	74	100
Simple	12	23	35	58	87	100
Baroni	11	22	29	52	76	100
Ochiai/cosine	11	22	30	51	75	100



Kulcznski(2)	12	22	31	53	79	100
Forbes	13	25	36	58	85	100
Fossum	11	21	30	51	75	100
Simpson	13	25	36	58	85	100
Pearson	11	22	30	53	78	100
Yule	13	24	36	55	82	100
Stile	11	22	30	53	78	100
Dennis	11	23	31	54	78	100

**Table 4 :** The Percentage of Actives in Active1 by using Ranking Positions

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
Jaccard/tanimoto	17	36.5	50.6	75.6	96.7	100
Russel/rao	17.3	24.4	36.9	56.8	89.3	100
Simple	28.8	43.9	57.9	76.8	97	100
Baroni	20.3	38.7	54.6	80.4	99.3	100
Ochiai/cosine	16.6	37.3	55	76.4	96.7	100
Kulcznski(2)	17	39.1	55.4	75.3	97	100
Forbes	28.4	43.5	56.8	80.1	98.9	100
Fossum	16.6	37.6	54.6	76.4	96.7	100
Simpson	23.2	41	53.1	73.1	92.6	100
Pearson	17	38	55	78.2	97.8	100
Yule	26.2	40.6	57.9	80.1	99.3	100
Stile	17.7	37.6	55	78.2	97.8	100
Dennis	19.6	39.5	55	78.2	98.2	100

**Table 5:** The Percentage of Actives in Active2 by using Ranking Positions

SINGLE COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
Jaccard/tanimoto	0.862	11.2	21.1	42.67	76.3	100
Russel/rao	2.586	11.6	25.9	49.57	75.9	100
Simple	0	6.03	23.3	47.41	86.6	100
Baroni	0.431	11.2	20.7	43.53	78	100
Ochiai/cosine	0.431	8.62	19.8	42.67	76.7	100
Kulcznski(2)	0.431	6.47	19.4	42.67	74.6	100
Forbes	0.862	5.17	18.1	49.57	85.8	100
Fossum	0.431	8.62	20.3	47.41	76.7	100
Simpson	0.862	4.31	10.8	43.53	79.3	100
Pearson	0.431	8.19	19.8	38.79	76.3	100
Yule	0.431	5.6	20.3	43.97	81.9	100
Stile	0.431	8.19	19.8	45.26	76.3	100
Dennis	0.431	8.19	20.7	45.69	78	100

**Table 6:** The Percentage of Actives in Active3 by using Ranking Positions

<b>SINGLE COEFFICIENT</b>	<b>THE PERCENTAGE OF ACTIVES</b>					
	<b>10%</b>	<b>20%</b>	<b>30%</b>	<b>50%</b>	<b>80%</b>	<b>100%</b>
Jaccard/tanimoto	14.9	36.6	53	64.9	82.8	100
Russel/rao	12.7	37.3	43.3	61.2	86.6	100
Simple	15.7	29.1	38.8	57.5	82.8	100
Baroni	17.9	34.3	47	65.7	83.6	100
Ochiai/cosine	14.2	38.8	51.5	66.4	82.8	100
Kulcznski(2)	12.7	38.8	47	66.4	82.1	100
Forbes	17.2	31.3	46.3	65.7	83.6	100
Fossum	14.2	38.8	50.7	66.4	82.8	100
Simpson	11.9	38.1	41.8	61.2	86.6	100
Pearson	14.2	36.6	53	65.7	82.8	100
Yule	15.7	35.8	50	64.9	83.6	100
Stile	14.2	37.3	53	64.2	82.8	100
Dennis	14.2	37.3	53	64.2	82.1	100

**APPENDIX D**

**PERCENTAGE OF ACTIVES USING DIFFERENT FUSION  
COEFFICIENTS**

**Table 1 : Fusion of 2-Coefficients on Active1**

FUSION COEFFICIENTs	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
BarFos	19	37	55	80	99	100
TanCos	17	37	53	76	97	100
CosSti	17	38	56	77	97	100
RusFor	16	28	43	62	91	100
TanRus	15	28	43	61	91	100
RusCos	15	28	43	62	92	100
RusSti	15	29	45	65	93	100
ForTan	27	41	57	82	99	100
CosFor	<b>27</b>	40	58	81	99	100
ForStil	15	28	43	61	91	100
TanSimple	22	40	57	81	99	100
CosFos	17	37	55	76	97	100
TanBar	19	36	55	80	99	100
RusSimple	16	37	51	78	99	100
RusKul	15	28	43	63	93	100
TanFos	17	37	53	76	96	100
RusFos	15	28	43	61	91	100
RusBar	14	33	48	68	94	100

**Table 1 : Fusion of 2-Coefficients on Active2**

FUSION COEFFICIENTs	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
BarFos	15	23.8	30.74	48.92	89.18	100
<b>TanCos</b>	16	24.7	30.3	48.05	88.31	100
CosSti	15	24.2	29.87	46.75	88.31	100
RusFor	10	22.9	30.3	46.75	88.31	100
TanRus	5.6	24.2	29.44	47.62	87.45	100
RusCos	5.2	24.2	29.44	47.62	87.88	100
RusSti	5.2	23.8	29	46.75	88.31	100
ForTan	14	24.2	30.74	48.48	88.31	100
<b>CosFor</b>	16	23.4	30.74	48.05	89.18	100
<b>ForStil</b>	15	24.2	30.74	48.05	89.18	100
TanSimple	12	25.5	34.63	51.52	82.25	100
CosFos	15	23.8	29.87	46.75	88.31	100
<b>TanBar</b>	16	25.1	31.6	49.35	89.18	100
RusSimple	15	24.7	30.74	48.48	87.45	100
RusKul	5.2	23.8	30.3	47.19	87.45	100
<b>TanFos</b>	16	24.2	29.87	47.62	88.31	100
RusFos	5.2	23.4	29.87	47.19	87.88	100
RusBar	11	22.9	29	46.75	88.31	100

**Table 3:** Fusion of 2-Coefficients on Active3

FUSION COEFFICIENTs	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
<b>BarFos</b>	17	37	52	66	85	100
TanCos	15	38	54	65	83	100
CosSti	15	39	54	67	83	100
RusFor	15	38	42	63	88	100
TanRus	16	38	45	65	86	100
<b>RusCos</b>	17	38	46	66	86	100
RusSti	15	38	45	67	86	100
<b>ForTan</b>	17	35	50	66	85	100
<b>CosFor</b>	17	37	51	65	84	100
<b>ForStil</b>	17	36	51	66	84	100
<b>TanSimple</b>	17	31	45	64	84	100
CosFos	14	39	52	67	83	100
<b>TanBar</b>	17	36	51	66	85	100
RusSimple	14	38	53	65	84	100
RusKul	15	38	42	64	86	100
TanFos	15	37	54	65	83	100
RusFos	15	38	42	63	88	100
RusBar	15	38	48	67	83	100

**Table 4:** Fusion of 2-Coefficients on Active4

FUSION COEFFICIENTs	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
BarFos	0.45	0.45	1.79	7.62	59.6	100
TanCos	0.45	0.9	1.79	8.52	61.4	100
CosSti	0.45	0.9	1.79	8.07	61.4	100
RusFor	0.45	0.45	1.35	8.07	61.4	100
TanRus	0.45	1.35	3.59	14.8	63.7	100
RusCos	0.45	1.35	3.59	13.5	61.9	100
RusSti	0.9	1.35	4.04	14.8	61.9	100
ForTan	0.45	0.45	1.79	9.42	60.1	100
CosFor	0.45	0.9	1.79	8.07	61.4	100
ForStil	0.45	0.9	1.79	8.07	61.4	100
TanSimple	0.45	0.45	1.35	11.2	62.8	100
CosFos	0.45	0.9	1.79	8.07	61.4	100
TanBar	0.45	0.45	1.79	7.17	60.1	100
RusSimple	0.45	0.9	1.79	7.17	58.3	100
RusKul	0.9	1.35	3.14	12.6	61.4	100
TanFos	0.45	0.9	1.79	8.52	61	100
RusFos	0.9	1.35	3.59	14.8	63.7	100
RusBar	0.45	1.35	3.14	10.8	61	100

**Table 5:** Fusion of 2-Coefficients on Active5

FUSION COEFFICIENTs	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
BarFos	6	16	26	46	60	100
TanCos	6	16	31	50	67	100
CosSti	6	16	31	48	61	100
RusFor	6	16	28	49	61	100
<b>TanRus</b>	6	30	44	52	85	100
<b>RusCos</b>	6	29	43	51	78	100
RusSti	6	27	43	52	70	100
ForTan	6	14	20	35	58	100
CosFor	6	14	20	39	58	100
ForStil	6	14	20	38	58	100
TanSimple	6	15	20	32	52	100
CosFos	6	16	31	50	61	100
TanBar	6	16	26	46	62	100
RusSimple	6	16	26	44	58	100
<b>RusKul</b>	6	29	42	51	76	100
TanFos	6	16	31	50	67	100
<b>RusFos</b>	6	29	44	52	85	100
RusBar	6	21	42	52	70	100

**Table 6:** Fusion of 2-Coefficients on Active6

FUSION COEFFICIENTs	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
<b>BarFos</b>	5.5	16	27	44	70	100
TanCos	7	21	27.5	44	70.5	100
CosSti	8	21	27.5	44	69.5	100
RusFor	11	22	28.5	43	69	100
TanRus	15	25	31.5	42	77	100
<b>RusCos</b>	15	25	31	44	74	100
RusSti	15	24	31.5	45	72	100
<b>ForTan</b>	4.5	7.5	17	43	71.5	100
<b>CosFor</b>	4	8	20.5	43	71.5	100
<b>ForStil</b>	4.5	7.5	20	43	72	100
<b>TanSimple</b>	17	24	30.5	42	69	100
CosFos	7.5	22	28	43	69.5	100
<b>TanBar</b>	4.5	15	26	44	70	100
RusSimple	18	28.5	32.5	41	70	100
RusKul	16	25	31.5	44	73.5	100
TanFos	7	21	27.5	44	70.5	100
RusFos	18	27	30	43	76	100
RusBar	12	22.5	31.5	46	71.5	100

**Table 7:** Fusion of 2-Coefficients on Active7

FUSION COEFFICIENTs	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
<b>BarFos</b>	21	29	36.5	52	84	100
TanCos	22	30	37.5	53	83	100
CosSti	22	31	38	53	81.5	100
RusFor	21	31.5	38.5	54	82	100
TanRus	21	31.5	37.5	56	79	100
<b>RusCos</b>	22	31	38	55	79	100
RusSti	22	31	38	56	78.5	100
<b>ForTan</b>	19	30	36	49	83.5	100
<b>CosFor</b>	21	30.5	37	52	85.5	100
<b>ForStil</b>	20	30.5	36	51	86	100
<b>TanSimple</b>	20	31	38	52	83.5	100
CosFos	22	31.5	38.5	55	81.5	100
<b>TanBar</b>	21	30.5	37.5	52	85	100
RusSimple	20	33.5	39.5	55	77	100
RusKul	22	32	38.5	56	78.5	100
TanFos	22	29.5	37	53	84	100
RusFos	20	33.5	39.5	55	74.5	100
RusBar	21	31.5	37.5	56	79	100

**Table 8:** Fusion of 3-Coefficients on Active1

FUSION COEFFICIENTs	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
TanBarFos	18.8	36.2	54.61	79.34	98.89	100
RusForCos	17.7	39.1	55.72	76.01	97.42	100
RusForTan	17.3	38	55.72	76.38	97.42	100
RusForSti	17.7	39.1	54.98	77.49	97.42	100
RusTanCos	14.8	32.5	46.49	66.05	92.99	100
RusCosSti	14.8	32.1	47.6	70.48	93.73	100
ForTanCos	24	39.9	56.83	79.34	99.26	100
<b>ForCosSti</b>	24.7	39.1	56.46	79.7	99.26	100

**Table 9:** Fusion of 3-Coefficients on Active2

FUSION COEFFICIENTs	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
TanBarFos	14.7	24.2	30.74	48.92	88.31	100
RusForCos	11.3	22.9	29.44	45.02	87.88	100
RusForTan	12.1	22.9	29	45.45	87.88	100
RusForSti	11.3	22.9	29.44	45.89	88.31	100
RusTanCos	9.96	24.2	29	47.19	87.88	100
RusCosSti	9.09	24.2	29.44	47.19	88.31	100
<b>ForTanCos</b>	15.6	23.8	30.74	48.48	89.61	100
ForCosSti	14.3	23.8	29.87	48.05	90.48	100

**Table 10:** Fusion of 3-Coefficients on Active3

FUSION COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
TanBarFos	15.79	37.59	45.11	65.4	86.5	100
RusForCos	17.29	30.83	45.11	64.7	84.2	100
RusForTan	14.29	39.1	51.88	66.9	83.5	100
RusForSti	16.54	36.09	51.13	66.2	85	100
RusTanCos	14.29	37.59	52.63	65.4	84.2	100
RusCosSti	15.04	37.59	42.11	64.7	86.5	100
<b>ForTanCos</b>	15.04	36.84	54.14	65.4	83.5	100
<b>ForCosSti</b>	15.04	37.59	42.11	63.2	88	100



**Table 11:** Fusion of 3-Coefficients on Active4

FUSION COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
TanBarFos	0.45	0.45	1.79	8.07	61	100
RusForCos	0.45	1.35	3.14	10.8	60.5	100
RusForTan	0.45	1.35	3.14	11.2	61	100
RusForSti	0.45	1.35	3.59	11.7	60.5	100
RusTanCos	0.45	1.35	3.14	10.8	60.5	100
RusCosSti	0.45	1.35	3.59	10.8	60.5	100
<b>ForTanCos</b>	0.45	1.35	3.59	11.7	60.5	100
<b>ForCosSti</b>	0.45	0.45	1.79	8.07	61	100

**Table 12:** Fusion of 3 Coefficients on Active5

FUSION COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
TanBarFos	6	16	26	46	62	100
RusForCos	6	23	42	52	77	100
RusForTan	6	23	41	51	78	100
RusForSti	6	23	41	52	70	100
RusTanCos	6	23	42	59	77	100
RusCosSti	6	21	41	51	69	100
<b>ForTanCos</b>	6	15	21	43	59	100
<b>ForCosSti</b>	6	14	21	43	59	100

**Table 13:** Fusion of 3-Coefficients on Active6

FUSION COEFFICIENT	THE PERCENTAGE OF ACTIVES					
	10%	20%	30%	50%	80%	100%
TanBarFos	5.5	18	27	44	69.5	100
RusForCos	14	23.5	32	45	79	100
RusForTan	14	23.5	32	45.5	74	100
RusForSti	14	23	32	45	71.5	100
RusTanCos	13	23	32	45	73	100
RusCosSti	13	23	31.5	45	72	100
<b>ForTanCos</b>	4	11.5	24.5	44.5	70.5	100
<b>ForCosSti</b>	4	12	24.5	44	70	100

**Table 14:** Fusion of 3-Coefficients on Active7

<b>FUSION COEFFICIENT</b>	<b>THE PERCENTAGE OF ACTIVES</b>					
	<b>10%</b>	<b>20%</b>	<b>30%</b>	<b>50%</b>	<b>80%</b>	<b>100%</b>
TanBarFos	21.5	30	35	52.5	82	100
RusForCos	21	32	38.5	55.5	79	100
RusForTan	21	32	38.5	56	79	100
RusForSti	21	31.5	37.5	51.5	84.5	100
RusTanCos	21	31.5	37.5	51.5	84.5	100
RusCosSti	20.5	32	37.5	55	79	100
<b>ForTanCos</b>	21	31.5	37.5	51.5	85	100
<b>ForCosSti</b>	21	31.5	38.5	52	86	100

**APPENDIX E**

**PERCENTAGE AVERAGE OF ACTIVES USING DIFFERENT 10 QUERIES  
FOR SINGLE COEFFICIENTS**

**Table 1:** The Percentage of Actives Obtain For Each Active with Average of 10 Target(Active 1)

Different Query	THE PERCENTAGE OF ACTIVES												
	Tan	Rus	Bar	Sim	Cos	Kul	For	Fos	Per	Simp	Sti	Yul	Den
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
<b>Query 1</b>	13.7	17.3	19.9	28.4	16.6	17	28.4	16.6	17.7	23.2	17.7	26.2	19.6
<b>Query 2</b>	18.5	17.3	19.2	23.2	16.2	16.6	22.9	16.2	17.7	21.4	17.3	22.1	19.6
<b>Query 3</b>	14	14.4	17	26.9	18.1	17.7	26.9	18.1	17.7	22.1	17.7	21.8	17.7
<b>Query 4</b>	14.4	14.4	16.6	26.2	17.3	18.5	26.6	17.3	17.7	21	17.7	20.3	17.7
<b>Query 5</b>	13.7	14	18.1	29.9	17.3	18.1	29.5	17.3	16.6	21.4	16.6	23.2	17
<b>Query 6</b>	16.6	12.5	18.1	25.8	14.8	15.9	26.6	14.4	15.5	19.2	15.5	21.4	18.1
<b>Query 7</b>	17	12.2	17	26.9	14.4	16.2	28.4	14.4	15.9	23.6	15.9	22.5	17.3
<b>Query 8</b>	17.3	11.4	19.9	26.6	14.4	17.7	26.6	14	17.3	21	17.3	21.8	20.3
<b>Query 9</b>	17	13.3	20.3	24.7	18.8	18.5	27.7	18.8	19.2	23.6	19.2	24.4	19.9
<b>Query 10</b>	17	12.5	15.1	18.8	15.1	15.5	19.2	14.7	14.8	19.9	15.1	20.3	15.5
<b>Average</b>	<b>15.9</b>	<b>13.9</b>	<b>18.1</b>	<b>25.8</b>	<b>16.3</b>	<b>17.2</b>	<b>26.3</b>	<b>16.2</b>	<b>17</b>	<b>21.7</b>	<b>17</b>	<b>22.4</b>	<b>18.3</b>

**Table 2:** The Percentage of Actives Obtain For Each Active with Average of 10 Target(Active 2)

Different Query	THE PERCENTAGE OF ACTIVES												
	Tan	Rus	Bar	Sim	Cos	Kul	For	Fos	Per	Simp	Sti	Yul	Den
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
<b>Query 1</b>	16.5	3.9	<b>15.6</b>	3.9	14.7	10.8	11.7	14.7	15.6	11.7	15.2	13.4	15.6
<b>Query 2</b>	26	32	<b>25.1</b>	32	26.8	26.8	25.5	26.4	26.4	26.8	26.4	26.4	26.8
<b>Query 3</b>	39.8	30.3	<b>43.7</b>	32.5	39.4	38.1	38.5	39.4	40.7	39.4	40.7	44.2	42
<b>Query 4</b>	29.9	19.9	<b>21.6</b>	19.9	21.2	21.2	22.9	21.2	21.2	23.8	21.2	21.2	21.2
<b>Query 5</b>	21.2	32	<b>25.1</b>	32	26.8	26.8	25.5	26.4	26.4	26.8	26.4	26.4	26.8
<b>Query 6</b>	26	26	<b>30.3</b>	26.8	29.9	30.7	28.6	29.9	29.9	29.4	29.9	29.9	29.9
<b>Query 7</b>	29	23.4	<b>28.6</b>	25.5	29	29	25.5	29	29	27.7	29	28.6	28.6
<b>Query 8</b>	16	10.4	<b>17.3</b>	10.4	15.2	15.2	15.6	14.7	15.2	14.7	15.2	15.2	15.2
<b>Query 9</b>	22.9	16	<b>23.4</b>	19.9	22.5	22.5	26	22.5	22.5	28.6	22.5	22.9	22.5
<b>Query 10</b>	17.7	15.6	<b>19</b>	15.6	17.3	16.5	19.9	17.3	17.7	19.5	17.7	16.9	17.7
<b>Average</b>	<b>24.5</b>	<b>21</b>	<b>25</b>	<b>21.9</b>	<b>24.4</b>	<b>23.8</b>	<b>24</b>	<b>24.2</b>	<b>24.5</b>	<b>24.8</b>	<b>24.4</b>	<b>24.5</b>	<b>24.6</b>

**Table 3 :** The Percentage of Actives Obtain For Each Active with Average of 10 Target (Active 3)

Different Query	THE PERCENTAGE OF ACTIVES												
	Tan	Rus	Bar	Sim	Cos	Kul	For	Fos	Per	Simp	Sti	Yul	Den
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
<b>Query 1</b>	<b>14.9</b>	12.7	17.9	15.7	14.2	12.7	17.2	14.2	14.2	11.9	14.2	15.7	14.2
<b>Query 2</b>	<b>5.22</b>	11.9	6.72	5.22	5.97	6.72	5.22	5.97	5.22	11.9	5.22	5.22	5.22
<b>Query 3</b>	<b>8.21</b>	15.7	11.2	5.97	8.96	8.21	7.46	8.96	8.96	15.7	5.97	8.21	8.96
<b>Query 4</b>	<b>31.3</b>	17.2	29.9	28.4	32.1	33.6	29.1	32.8	30.6	18.7	32.1	26.9	30.6
<b>Query 5</b>	<b>34.3</b>	20.9	32.1	30.6	34.3	30.6	28.4	34.3	32.8	21.6	32.8	28.4	32.8
<b>Query 6</b>	<b>34.3</b>	21.6	34.3	27.6	33.6	32.8	27.6	33.6	32.8	21.6	32.8	30.6	32.1
<b>Query 7</b>	<b>20.1</b>	13.4	20.1	20.1	20.1	19.4	20.9	20.1	20.1	11.9	20.1	19.4	19.4
<b>Query 8</b>	<b>26.9</b>	23.9	28.4	24.6	25.4	24.6	25.4	25.4	24.6	23.9	25.4	23.9	25.4
<b>Query 9</b>	<b>26.9</b>	22.4	27.6	26.1	24.6	23.9	26.1	25.4	24.6	20.1	24.6	23.1	25.4
<b>Query 10</b>	<b>21.6</b>	14.9	21.6	23.1	21.6	21.6	23.1	21.6	21.6	15.7	21.6	21.6	21.6
<b>Average</b>	<b>22.5</b>	<b>17.5</b>	<b>23</b>	<b>20.7</b>	<b>22.2</b>	<b>21.4</b>	<b>21</b>	<b>22.2</b>	<b>21.6</b>	<b>17.3</b>	<b>21.5</b>	<b>20.3</b>	<b>21.6</b>

**Table 4:** The Percentage of Actives Obtain For Each Active with Average of 10 Target(Active 4)

Different Query	THE PERCENTAGE OF ACTIVES												
	Tan	Rus	Bar	Sim	Cos	Kul	For	Fos	Per	Simp	Sti	Yul	Den
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
<b>Query 1</b>	0.45	1.35	<b>0.45</b>	0.45	0.45	0.45	0.45	0.45	0.45	0.9	0.45	0.45	<b>0.45</b>
<b>Query 2</b>	10.3	5.38	<b>9.87</b>	7.17	10.3	9.87	7.17	10.3	9.87	6.73	9.87	7.62	<b>9.42</b>
<b>Query 3</b>	22.4	17	<b>23.3</b>	17.9	23.3	23.8	19.7	23.8	22.9	17.5	22.9	23.8	<b>23.3</b>
<b>Query 4</b>	2.24	1.79	<b>2.69</b>	3.14	2.24	1.79	4.48	2.24	2.24	2.69	2.24	2.69	<b>2.69</b>
<b>Query 5</b>	23.8	22	<b>23.8</b>	18.4	22.4	22.9	19.3	22.9	22.4	22.4	22	22.9	<b>23.8</b>
<b>Query 6</b>	20.2	14.3	<b>20.2</b>	16.1	19.7	19.3	17	19.3	19.7	14.8	19.7	18.4	<b>20.2</b>
<b>Query 7</b>	28.3	26	<b>28.7</b>	19.3	28.3	28.3	20.6	28.3	28.3	23.8	28.3	28.7	<b>28.7</b>
<b>Query 8</b>	29.6	25.1	<b>28.7</b>	20.6	29.1	29.1	22	29.1	29.1	23.8	30	30	<b>29.6</b>
<b>Query 9</b>	29.6	26	<b>30</b>	22	29.6	30.5	22.9	29.6	29.6	23.3	29.6	30.5	<b>30</b>
<b>Query 10</b>	23.8	15.2	<b>24.2</b>	17.9	24.2	23.3	19.7	25.1	23.8	18.4	23.3	22.9	<b>24.2</b>
<b>Average</b>	<b>19.2</b>	<b>15.4</b>	<b>19.2</b>	<b>14.3</b>	<b>19</b>	<b>18.9</b>	<b>15.3</b>	<b>19.1</b>	<b>18.8</b>	<b>15.4</b>	<b>18.8</b>	<b>18.8</b>	<b>18.4</b>

**Table 5 :** The Percentage of Actives Obtain For Each Active with Average of 10 Target(Active 5)

Different Query	THE PERCENTAGE OF ACTIVES												
	Tan	Rus	Bar	Sim	Cos	Kul	For	Fos	Per	Simp	Sti	Yul	Den
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
<b>Query 1</b>	6	<b>10</b>	6	6	6	6	6	6	6	7	6	6	6
<b>Query 2</b>	6	<b>11</b>	5	4	5	4	4	5	5	5	5	4	4
<b>Query 3</b>	4	<b>9</b>	5	3	4	3	3	3	4	4	4	3	4
<b>Query 4</b>	49	<b>40</b>	44	24	49	45	17	49	48	23	48	33	4
<b>Query 5</b>	13	<b>25</b>	10	7	12	11	7	12	10	10	11	8	45
<b>Query 6</b>	19	<b>23</b>	20	16	19	19	12	19	20	13	20	12	10
<b>Query 7</b>	14	<b>25</b>	10	7	14	10	7	14	11	10	11	9	19
<b>Query 8</b>	18	<b>25</b>	15	6	19	17	5	19	16	6	16	11	10
<b>Query 9</b>	23	<b>33</b>	19	10	23	22	11	23	22	21	22	17	22
<b>Query 10</b>	46	<b>45</b>	28	35	42	40	22	43	42	23	42	29	39
<b>Average</b>	<b>19.9</b>	<b>24.6</b>	<b>16.2</b>	<b>11.8</b>	<b>19.4</b>	<b>17.7</b>	<b>9.4</b>	<b>19.3</b>	<b>18.4</b>	<b>12.2</b>	<b>18.5</b>	<b>13.2</b>	<b>16.3</b>



**Table 6 :** The Percentage of Actives Obtain For Each Active with Average of 10 Target(Active 6)

Different Query	THE PERCENTAGE OF ACTIVES												
	Tan	Rus	Bar	Sim	Cos	Kul	For	Fos	Per	Simp	Sti	Yul	Den
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
<b>Query 1</b>	6	17	4.5	3	7.5	11	5	8	7.5	<b>17.5</b>	7.5	6	7
<b>Query 2</b>	16	8.5	11.5	14.5	16	16	12	16	15.5	<b>14</b>	15.5	13.5	15.5
<b>Query 3</b>	17	10	12	13.5	16.5	15	12	16.5	15	<b>14</b>	15.5	12	14
<b>Query 4</b>	11	15	12.5	13.5	10.5	15	13	11.5	10	<b>26.5</b>	10	12.5	11
<b>Query 5</b>	9.5	24	13.5	11.5	12	16.5	13	12.5	12	<b>25</b>	12	15	11.5
<b>Query 6</b>	16	12	13.5	14.5	17	18.5	13.5	17	17	<b>28.5</b>	17	17.5	17
<b>Query 7</b>	17	6.5	16	17.5	18	20.5	17	18	18	<b>6.5</b>	18	17.5	18
<b>Query 8</b>	15	9	14	14	15	15	15.5	15	15	<b>20.5</b>	15	14.5	14.5
<b>Query 9</b>	15.5	1	12	14.5	18	24	14.5	18.5	17	<b>22.5</b>	17.5	17.5	16.5
<b>Query 10</b>	15.5	1	12	14.5	18	24.5	14.5	18.5	17	<b>22.5</b>	17.5	17.5	16.5
<b>Average</b>	<b>13.9</b>	<b>10.4</b>	<b>12.2</b>	<b>13.1</b>	<b>14.9</b>	<b>17.6</b>	<b>13</b>	<b>15.2</b>	<b>14.4</b>	<b>19.8</b>	<b>14.6</b>	<b>14.4</b>	<b>14.2</b>

**Table 7:** The Percentage of Actives Obtain For Each Active with Average of 10 Target (Active 7)

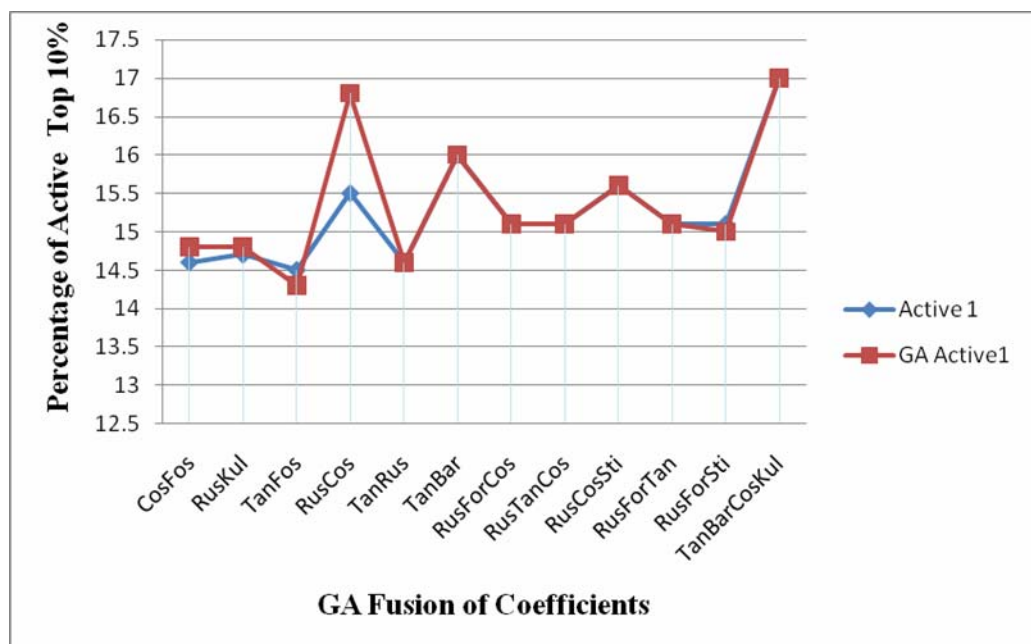
Differnet Query	THE PERCENTAGE OF ACTIVES												
	Tan	Rus	Bar	Sim	Cos	Kul	For	Fos	Per	Simp	Sti	Yul	Den
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
<b>Query 1</b>	21.5	20	21	15.5	21.5	21	15.5	21.5	21.5	21	21.5	21	<b>21</b>
<b>Query 2</b>	39	34.5	41.5	39	39	42	36.5	39	42	37	42	40	<b>42.5</b>
<b>Query 3</b>	44.5	40	45	41	44.5	44	39	44.5	44.5	40.5	44.5	43	<b>45</b>
<b>Query 4</b>	25	24	26	22.5	25	24	22	25	24.5	19	24.5	23	<b>25</b>
<b>Query 5</b>	34	28.5	31.5	28	33.5	33.5	26	33.5	33	27	33	31.5	<b>32.5</b>
<b>Query 6</b>	40	35	38	35	39	40	34	39	38	36.5	39.5	37.5	<b>39</b>
<b>Query 7</b>	48	39.5	47	44	47.5	47.5	41.5	47.5	47.5	43.5	47.5	48.5	<b>47.5</b>
<b>Query 8</b>	42	34.5	42	40.5	42	42.5	36.5	41.5	42.5	37.5	42.5	42.5	<b>43</b>
<b>Query 9</b>	31.5	26	32	30	31.5	30	25.5	31.5	31.5	25.5	31.5	31.5	<b>31.5</b>
<b>Query 10</b>	30.5	26	31.5	30.5	30.5	31	23.5	30.5	31	25.5	31	30.5	<b>31.5</b>
<b>Average</b>	<b>35.7</b>	<b>30.8</b>	<b>35.6</b>	<b>32.6</b>	<b>35.4</b>	<b>35.6</b>	<b>30</b>	<b>35.4</b>	<b>35.6</b>	<b>31.3</b>	<b>35.8</b>	<b>34.9</b>	<b>35.9</b>

**APPENDIX F**

**PERCENTAGE OF ACTIVES FUSIONS OF COEFFICIENT NON-WEIGHTS (10 Queries) COMBINATION COMPARED WITH GA WEIGHTS (10 Queries) COMBINATION FOR EACH ACTIVE FUSION.**

**Table 1:** The Percentage of Active top 10% on Active1 fusions of coefficient Compared with GA Active1 Fusions.

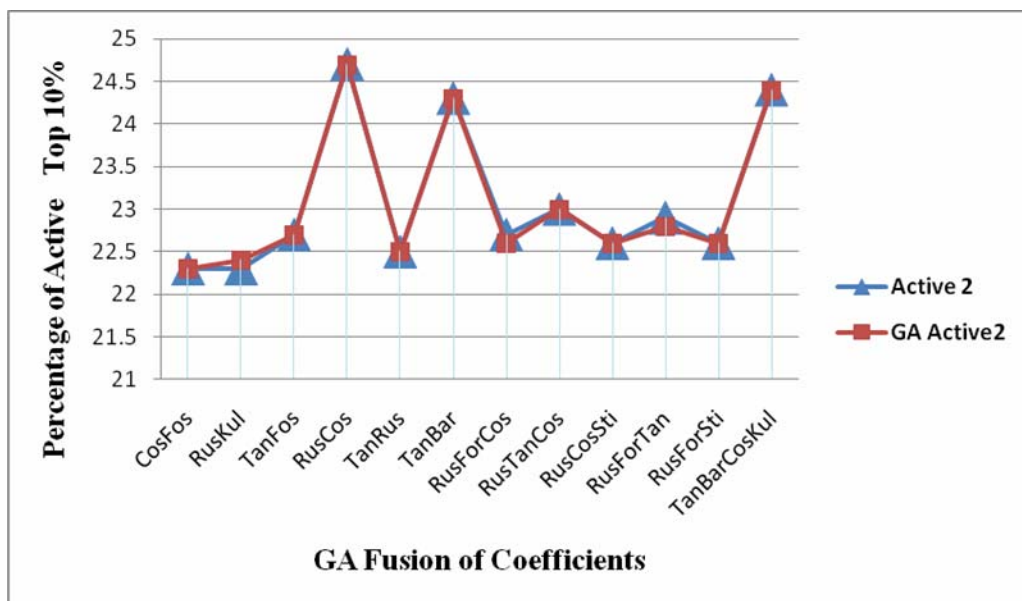
THE PERCENTAGE OF ACTIVES												
	Fusion2					Fusion3					Fusion 4	
	CosFos	RusKul	TanFos	RusCos	TanRus	TanBar	RusForCos	RusTanCos	RusCosSti	RusForTan	RusForSti	TanBarCosKul
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
<b>Active1</b>	14.6	14.7	14.5	15.5	14.6	16	15.1	15.1	15.6	15.1	15.1	<b>17</b>
<b>GA Active1</b>	14.8	14.8	14.3	16.8	14.6	16	15.1	15.1	15.6	15.1	15	17



**Figure 1** The Percentage Average of Top 10% Active1 versus the GA Active1 Fusions of Coefficients

**Table 2:** The Percentage of Active top 10% on Active2 fusions of coefficients Compared with GA Active2 Fusion.

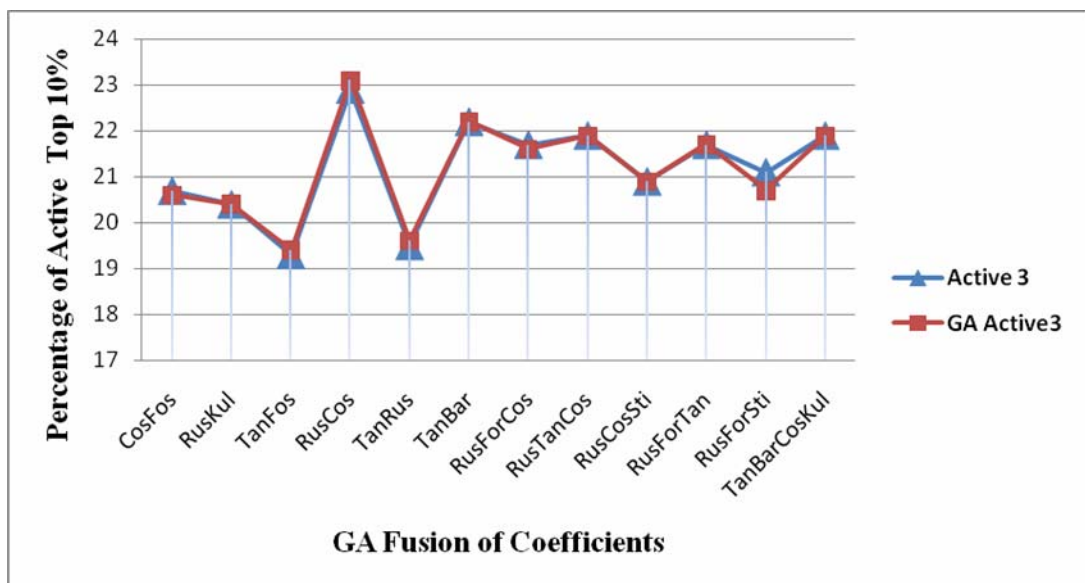
THE PERCENTAGE OF ACTIVES												
	Fusion2					Fusion3					Fusion 4	
	CosFos	RusKul	TanFos	RusCos	TanRus	TanBar	RusForCos	RusTanCos	RusCosSti	RusForTan	RusForSti	TanBarCosKul
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
<b>Active2</b>	22.3	22.3	22.7	24.7	22.5	24.3	22.7	23	22.6	22.9	23	24.4
<b>GA Active2</b>	22.3	22.4	22.7	24.7	22.5	24.3	22.6	23	22.6	22.8	23	24.4



**Figure 2** The Percentage Average of Top 10% Active2 versus the GA Active2 Fusions of Coefficients

**Table 3 :** The Percentage of Active top 10% on Active3 fusions of coefficients  
Compared with GA Active3 Fusion.

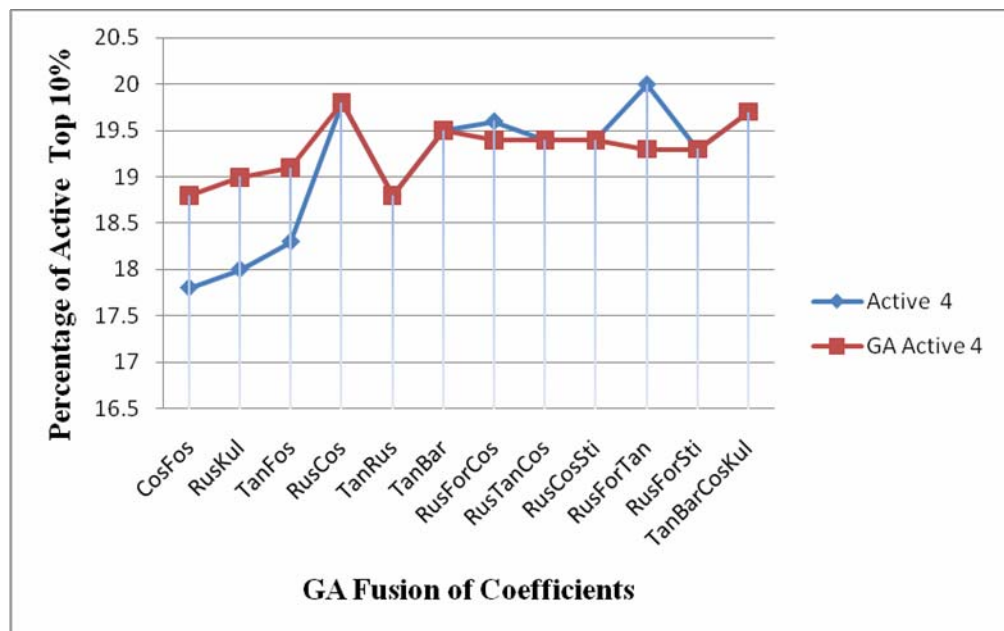
THE PERCENTAGE OF ACTIVES												
	Fusion2					Fusion3					Fusion 4	
	CosFos	RusKul	TanFos	RusCos	TanRus	TanBar	RusForCos	RusTanCos	RusCosSti	RusForTan	RusForSti	TanBarCosKul
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
<b>Active3</b>	20.7	20.4	19.3	22.9	19.5	22.2	21.7	21.9	20.9	22	21.1	21.9
<b>GA Active3</b>	20.6	20.4	19.4	23.1	19.6	22.2	21.6	21.9	20.9	22	20.7	21.9



**Figure 3** The Percentage Average of Top 10% Active3 versus the GA Active3 Fusions of Coefficients

**Table 4:** The Percentage of Active top 10% on Active4 fusions of coefficients Compared with GA Active4 Fusion.

THE PERCENTAGE OF ACTIVES												
	Fusion2					Fusion3					Fusion 4	
	CosFos	RusKul	TanFos	RusCos	TanRus	TanBar	RusForCos	RusTanCos	RusCosSti	RusForTan	RusForSti	TanBarCosKul
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
<b>Active4</b>	17.8	18	18.3	19.8	18.8	19.5	19.6	19.4	19.4	20	19.3	19.7
<b>GA Active4</b>	18.8	19	19.1	19.8	18.8	19.5	19.4	19.4	19.4	19.3	19.3	19.7



**Figure 4** The Percentage Average of Top 10% Active4 versus the GA Active4 Fusions of Coefficients

**Table 5:** The Percentage of Active top 10% on Active5 fusions of coefficients compared with GA Active5 Fusion.

THE PERCENTAGE OF ACTIVES												
Fusion2						Fusion3					Fusion 4	
CosFos	RusKul	TanFos	RusCos	TanRus	TanBar	RusForCos	RusTanCos	RusCosSti	RusForTan	RusForSti	TanBarCosKul	
10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	
<b>Active5</b>	23.3	23	22.7	18.9	22.8	19.6	21.6	21.5	21.5	21.8	21.5	19
<b>GA Active5</b>	23.3	23	22.8	18.9	22.9	19.5	21.7	21.9	21.5	21.8	21.6	19

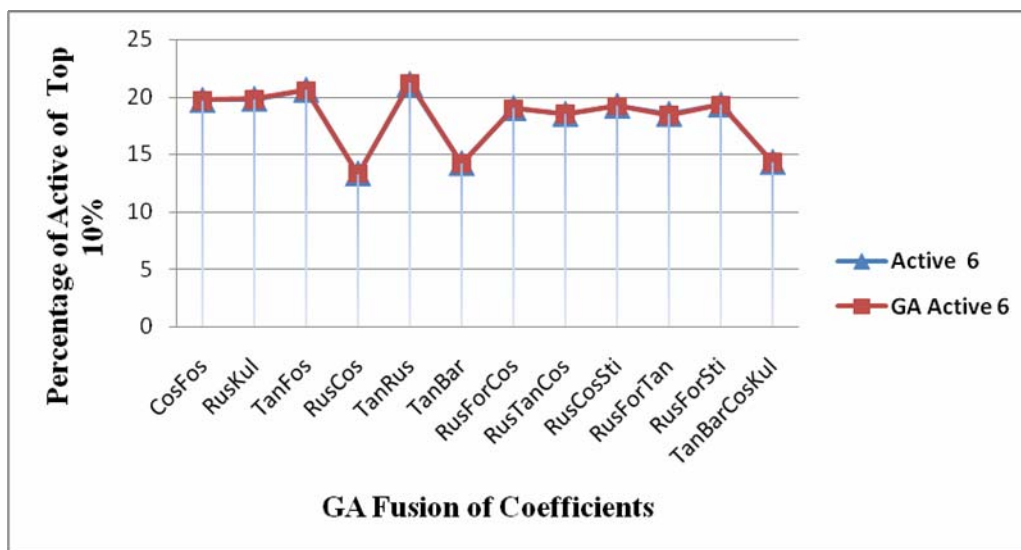


**Figure 5** The Percentage Average of Top 10% Active5 versus the GA Active5 Fusions of Coefficients.



**Table 6 :** The Percentage of Active top 10% on Active6 fusions of coefficients  
Compared with GA Active6 Fusion.

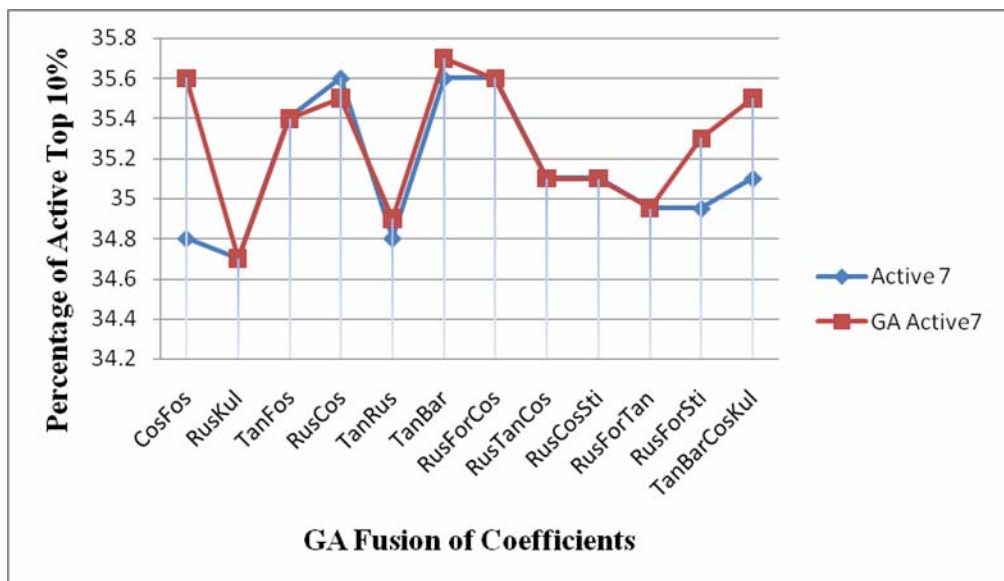
THE PERCENTAGE OF ACTIVES												
Fusion2						Fusion3					Fusion 4	
CosFos	RusKul	TanFos	RusCos	TanRus	TanBar	RusForCos	RusTanCos	RusCosSti	RusForTan	RusForSti	TanBarCosKul	
10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	
<b>Active6</b>	19.7	19.8	20.6	13.3	21.1	14.2	19	18.5	19.2	18.5	19.3	14.3
<b>GA Active6</b>	19.7	19.8	20.6	13.3	21.2	14.2	19	18.5	19.2	18.4	19.3	14.3



**Figure 6** The Percentage Average of Top 10% Active6 versus the GA Active6 Fusions of Coefficients

**Table 7 :** The Percentage of Active top 10% on Active7 fusions of coefficients compared with GA Active7 Fusion.

THE PERCENTAGE OF ACTIVES												
	Fusion2					Fusion3					Fusion 4	
	CosFos	RusKul	TanFos	RusCos	TanRus	TanBar	RusForCos	RusTanCos	RusCosSti	RusForTan	RusForSti	TanBarCosKul
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
<b>Active7</b>	34.8	34.7	35.4	35.6	34.8	35.6	35.6	35.1	35.1	35	34.95	35.1
<b>GA Active7</b>	35.6	34.7	35.4	35.5	34.9	35.7	35.6	35.1	35.1	35	35.3	35.5



**Figure 7** The Percentage Average of Top 10% Active7 versus the GA Active7 Fusions of Coefficients