Amended Data Fusion Similarity Measurement based on Genetic Algorithm for Chemical Database Retrieval

Yahya Ali Abdelrahman Ali [1†], **Ahmed Hamza Osman** [2††] **and Suad Mohammed** [1†]

[1†] Department of Information System, Faculty Computer Science and information System Najran University, Kingdom of Saudi Arabia

[2††] Faculty of Computing and Information Technology, King Abdulaziz University, P.O. Box 344, 21911, Jeddha Rabigh, Saudi Arabia

Summary

Virtual screening (VS) is a computer scheme used in the study of medicine development. VS is often used in computer-aided searches for novel lead compounds based on chemical similarity. Similarity retrieving is a technique for identifying molecules that are architecturally matched to a target chemical, which is beneficial in the discovery of new medicines. In the majority of traditional similarity methods, the molecular characteristics of biological and non-biologically linked activities are given equal weight. However, it has been shown that some distinguishing characteristics are more significant than others, depending on the chemical structure. As a result, this distinction should be considered when assigning a higher weight to each significant piece. The main objective for this study is to optimize weights of different similarity measures in data fusion for searching chemical database by applying a genetic algorithm (GA). In this paper, comparisons of various 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, A Tanimoto and Cosine coefficient, of weightings 0.972, 0.960 and 0.960 respectively, give the best result. Combinations of Tanimoto and Cosine coefficients perform well and give a 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. Combining Cosine and Fossum coefficients without weights yields an average of 21.89% among the top 10% of compounds, whereas when a 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% of compounds is obtained. Generally speaking, combinations of coefficients performed better than single coefficients.

Keywords: *Compounds; Data Fusion; Similarity; Chemical Database; Genetic Algorithm; coefficients*

1. Introduction

A biochemical database is a collection of data that is particularly intended to hold information about chemical compounds and their attributes. The process of information retrieval is commonly used to retrieve chemical compounds. A filtering retrieval process called the data fusion process has recently been used to integrate compound results from multiple chemical data resources [1]. Similarity measures were used as tools in the chemical database – such as retrieval, clustering, diversity analysis – which have two main components of molecular representation and similarity coefficients [2]. The majority of chemical databases include information on compounds that are stable throughout time. Chemical structures have historically been expressed on paper (2D structural formulae) [3] by lines denoting chemical bonds between atoms [4] and plotted on paper (2D structural formulae] using chemical bond lines. As perfect visual representations for chemists, they are not appropriate for computational usage, particularly in the context of research and storage. In order to store and search for information on millions of molecules, large chemical databases are anticipated to need physical memory space equivalent to terabytes of space [27].

Genetic algorithms look for actual or estimated answers to optimization and search issues. GA are categorized as universal search heuristics, which means they may be used in any situation. In evolutionary algorithms, they are a specific class of processes that make use of approaches enthused by evolutionary biology, such as descent with modification, crossover, selection and mutation (also termed recombination). Moreover, in computer simulations, genetic algorithms are implemented as sets of abstract representations (chromosomes, genotypes or genomes) of candidate solutions (referred to as individuals, creatures or phenotypes) that are subjected to an optimization problem in order to arrive at better solutions. Alternative encodings of solutions are available, in addition to the traditional binary representation of solutions as strings of 0s and 1s. Similarity

Manuscript received January 5, 2022
Manuscript revised January 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.1.69

measurements are the polar opposite of distance measures. Similarity functions take two points and return the high similarity value for the points that are close together and the small similarity value for the points that are far apart [5, 6]. The reciprocal method is one way of converting between a distance function and a similarity measure. This is the usual technique in physics and electronics for converting between resistance and conductance. Similarity measures assess the similarity between two molecules' representations using two fundamental tools: molecular representation and similarity coefficient [2].

Data fusion is a process in which data, evidence or judgments regarding the same set of objects are combined from, or based on, various sources in order to enhance the 3 superiority of decision-making under conditions of ambiguity about the objects [4, 7-9]. 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 the military, robotics, medicine and information retrieval. Fusions can improve confidence in results due to the use of balancing data [2, 8, 10, 11]. The use of data fusion may also enhance performance if, for example, a sensor becomes damaged or useless, since information from the other sensors will continue to flow in. Data fusion results in increased coverage, since many sensors may cover different regions, timeframes and quality. A weighting system is used to classify various properties of a molecule according to their importance in determining the molecule's resemblance to another.

The GA is used to determine the optimal linear integration of weights for the scores of various corresponding functions. On performance metrics, a GA [31,32] based system beats any of the separate expert matching algorithms. Additionally, the system beats the best individual expert matching algorithms.

In summary, this article makes the following significant contributions:

1- An enhanced GA optimize weights technique for looking for molecular similarity that makes use of chemical compound properties.

2- By focusing greater weights on essential characteristics, the introduction of a GA data fusion technique for feature weighting is suggested.

3- When compared to benchmark techniques, the suggested method demonstrated promising performance results.

The remainder of this work is divided into the following sections: Section II discusses related studies. The explanations of the planned Framework, Martial and Method are provided in Section III. The experimental results as well as the data set Section V contains a summary of the findings, analysis and discussion. Section VI details the study's Summary and Future Work.

2. Related work

Chemoinformatics has been a thriving interdisciplinary field of study in recent years, using a variety of techniques and technologies to benefit chemistry and drug development. In chemoinformatics, the use of virtual screening (VS) is deemed necessary to examine records of molecules and select those structures that are more likely to be linked to a pharmacological target. VS is categorized into two comprehensive classes: target-based and ligand-based [3]. Recently, many approaches based on both structure and ligand have been developed [12, 13]. All ligands are rated according to their maximum score in chemical databases, and the one with the highest score is then investigated further. The VS is based on architectural matching, comparing known and potentially active ligands, and emphasizing the molecular similarity principle, which states that compounds with similar structures may have comparable activity. Similarity searching is a widely utilized method for ligand-based VS. This method searches a chemical database for molecules that are the most comparable to a user-defined reference structure [14]. All similarity measures have three fundamental components [30]: (a) the representation, which depicts the structures to be considered; (b) the weighting scheme, which assigns significance weights to various sections of the structural representation; and (c) the similarity coefficient, which quantifies the degree of similarity between two appropriately weighted recursive structures [4]. To aid in the discovery of prospective (new-)SVHC compounds, we have devised a chemical similarity technique that determines if a novel chemical is structurally related to an existing SVHC compound [15]. Wassenaar et al. [16] examined the system performance of generated similarity modelling using a pseudo-external evaluation on a collection of compounds that had purportedly been classified as SVHC or non-SVHC using expert elicitations. When compared to the experts' views, the findings show that carcinogenic, mutagenic or reprotoxic (CMR) and endocrine disrupting (ED) chemicals performed well statistically, whereas (very) persistent, (very) bio-accumulative and toxic (PBT/vPvB) substances performed poorly[16].

Numerous data fusion attempts have been undertaken in the process of chemical compound information retrieval in order to integrate findings from various similarity searching systems [17, 18]. A query in similarity searching entails specifying the complete structure of molecules. It is necessary to describe this specification in terms of one or more structural descriptors, which are then compared to the set of structural descriptors associated with each molecule in the database. After that, a measure of similarity between the target structure and each database structure is computed. For molecular graph representations, graph representations may also be used to represent and explore databases of three-dimensional structures [3, 19]. The pharmaceutical business makes significant use of extremely complex technologies for storing, retrieving and analyzing data about the chemical structure of molecules. Not just for similarity searches, but also for compound selection and molecular diversity analysis, similarity computations between molecules were utilized. The similarity measure's findings were then utilized to rank the database structures in decreasing order of resemblance to the goal. One of the ways to improve the performances of molecular similarity retrieving is to combine the consequences of different measures of similarity, which is known as the data fusion process. How to optimize this combined result has become an interesting research area in chemoinformatics. Numerous techniques have been tried to improve the measurement of molecular similarity. Weighting and data fusion are two of these techniques. A weighting system is used to classify various properties of a molecule according to their importance in determining the molecule's resemblance to another. Recent work has shown that fusion outperforms the usage of single coefficients [29] Willett discovered that combining two kinds of ranking results in composite rankings that include significantly diverse groups of closest neighbours and often outperform the separate measures in simulated property prediction [4]. Similarity coefficients are used to quantify the degree of similarity between two structures numerically [17]. There are many different kinds of similarity metrics in use. As an example, edit distance, which is a string-based measure of the number of operations to modify the structure representation to another representation structure, has been used to measure the similarity between two 3D molecular structures [5]. Similarity coefficients are classified into four groups: association coefficients, probabilistic coefficients, correlation coefficients and distance coefficients [8, 20, 21]. Table 2 summarizes several coefficients of similarity. The distance coefficients refer to the quantities used to quantify the distance between structures in a molecular space. Association coefficients are pair-functions that may be used to quantify the degree of agreement between two molecules' binary, multi-state or continuous character representations [20]. A number of association coefficients have been employed to quantify the similarity of substances. Connection coefficients are often used to quantify the degree of correlation between sets of values associated with molecules, such as the independence of couples of real-rate molecular descriptors and proportionality. While probabilistic coefficients are seldom used to quantify molecular similarity, they concentrate on the distribution of descriptor frequencies among members of a data collection, emphasizing a matching on a rarely occurring variables. 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 [2]. This is well suited for various activities, databases and the type of molecular. To achieve a higher level of optimization, this research will also use GAs, a search method used in computing to determine the real or approximate answer to optimization and search problems.

3. Martial and Method

This paper's study methods will be conducted according to the workflow process, which consists of four phases as illustrated in Figure 1.

A. *Research Plan*

The work plan includes a 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. The similarity coefficients were clustered into 13 groups, then data fusion was employed to combine the results of the similarity to get a better result; this could be through the fusion of 2, 3 or 4 coefficients.

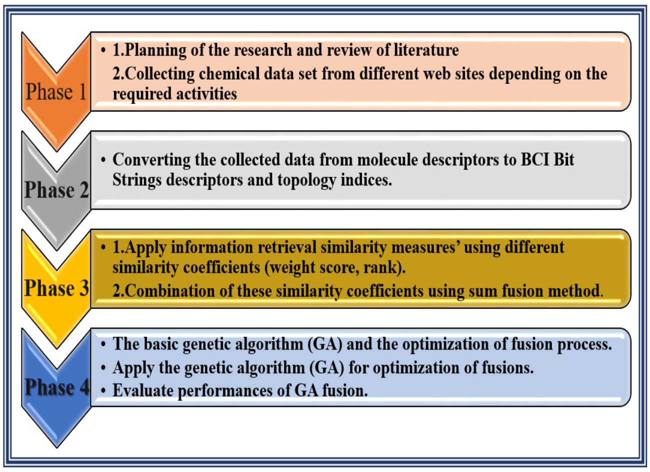
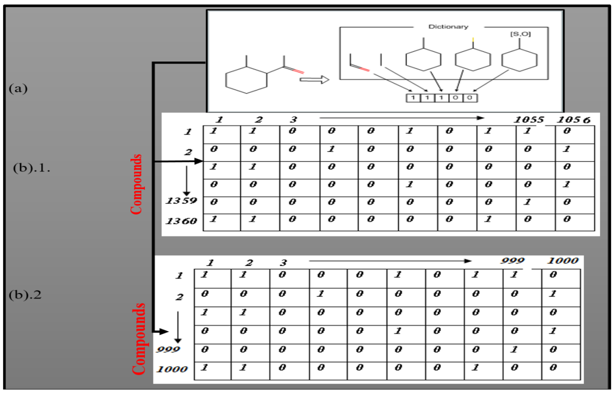

Phase 1
• 1.Planning of the research and review of literature
2.Collecting chemical data set from different web sites depending on the required activities
Phase 2
• Converting the collected data from molecule descriptors to BCI Bit Strings descriptors and topology indices.
Phase 3
• 1.Apply information retrieval similarity measures' using different similarity coefficients (weight score, rank).
2.Combination of these similarity coefficients using sum fusion method.
Phase 4
• The basic genetic algorithm (GA) and the optimization of fusion process.
• Apply the genetic algorithm (GA) for optimization of fusions.
• Evaluate performances of GA fusion.

Figure 1. Flow Chart of the Framework

B. *Collection of Chemical Data Set*

Chemical databases containing a huge number of compounds are available on many web sites. The necessary data for this research is the MDL Drug Database Report (MDDR). The database, which was created by Molecular Design Limited (MDL) and Prous Science, includes over 100,000 physiologically relevant chemicals and well-defined derivatives, with updates adding about 10,000 compounds each year [22][28]. The MDDR Finder enables one to carry out searches inside the database or across relevant data fields. MDL also suggests MDDR-3D, are collected from the Discovery Gate website (URL:https://www.discoverygate.com). The data available in molecule graph format is converted using MAKEBITS software from BCI (Barnard Chemical Information)[12, 13] into bit string format, where the compound is represented as a series of 0's and 1's without spacing between them. BCI is a 1052-bit structural key-based bit string that is produced based on the presence and absence of fragments in the standard 1052 fragment dictionary of the BCI, which contains enhanced atoms, atom sequences, atom pairings, ring components and ring fusion descriptors [23]. Below is an example of how MAKEBITS works to convert molecule data into BCI bit string data, as shown in Figure 2.



	1	2	3						1055	1056
1	1	1	0	0	0	1	0	1	1	0
2	0	0	0	1	0	0	0	0	0	1
	1	1	0	0	0	0	0	0	0	0
	0	0	0	0	0	1	0	0	0	1
1359	0	0	0	0	0	0	0	0	1	0
1360	1	1	0	0	0	0	0	1	0	0

	1	2	3						999	1000
1	1	1	0	0	0	1	0	1	1	0
2	0	0	0	1	0	0	0	0	0	1
	1	1	0	0	0	0	0	0	0	0
	0	0	0	0	0	1	0	0	0	1
999	0	0	0	0	0	0	0	0	1	0
1000	1	1	0	0	0	0	0	1	0	0

Figure 2. BCI bit string descriptor example

Figure 2 shows the BCI bit string descriptor generated by MAKEBITS software: (a) the 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.

The data represented here are composed from active compounds with different degrees of activity for each of them. There are two types of data, as shown in Figure 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 same way for other parts.

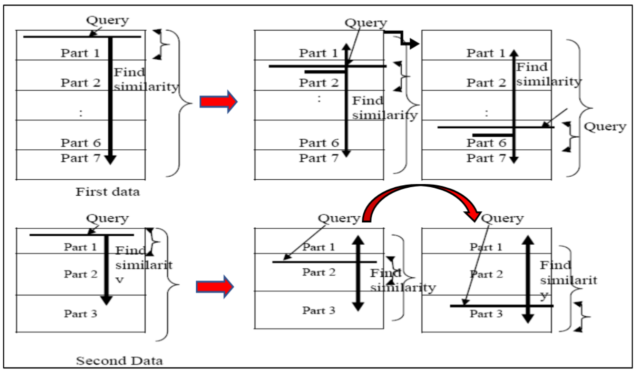


Figure 3. Mechanism of Searching Similarity

C. Calculate the Similarity of Retrieval Compound

To calculate the similarity of the retrieval compound depend on the 13 groups we use the equation (3.3) as shown under this table. Salim and co-workers [10] have clustered the 22 coefficients mentioned into 13 groups separated of coefficient in order to enhance the similarity searching as well as the combination between groups [10, 20].

Table1. Main 13 groups of coefficients measure

The grouping of coefficients	
Group A :	{ Sokal/Sneath(1), Jaccard/Tanimoto, Kulczynski(1), Dice}
Group B :	{Russell/Rao}
Group C :	{Simple Matching, Hamann, Sokal/Sneath , Rogers/Tanimoto, Sokal/Sneath(3), Mean Manhattan}
Group D :	{Baroni-Urbani/Buser}
Group E :	{Ochiai/Cosine}
Group F :	{Kulczynski(2), McConnaughey}
Group G :	{Forbes}
Group H :	{Fossum}
Group I :	{Simpson}
Group J :	{Pearson}
Group K :	{Yule}
Group L :	{Stiles}
Group M :	{Dennis}

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

Application of information retrieval similarity measures using different similarity coefficients (weight score, rank) between two bit-string representations, molecules M1 and M2 of length n

$$\text{Rank } M(I,j) = \sum_{i=1}^{n} M(I,j) \qquad (1)$$

Where i is molecule, j is query and n is the number of molecules.

The sum fusion technique is used to combine similarity coefficients. In the area of chemical information, similarity measurements have mostly been confined to a few single metrics, such as the Cosine, Euclidean Distance and Tanimoto. Numerous items listed in Table 2.2 have seldom been employed in conjunction with estimates of chemical structural similarity. The molecular data converted to BCI bit-strings will be used for similarity retrieves on the MDDR databases through the implementation of the similarity coefficient given in Table 2. Naomie and co-workers (2003) clustered the 22 coefficients mentioned into 13 groups separated of coefficient in order to enhance the similarity searching as well as the combination between groups.

The data fusion method was based on a summation of the rankings generated by the similarity queries. According to the following stages, the combining of similarity rankings through data fusion was determined to be the most efficient technique for similarity retrieving in chemical databases [24]:

- Conduct a similarity search [25] of a chemical database using two or more distinct metrics of intermolecular structural similarity to locate a specific target structure.

- Take note of each database structure's rank position, ri, in the ranking produced by the i-th similarity metric.
- Using one of the fusion algorithms (SUM), aggregate the different ranks, resulting in a new combined score for each database structure.
- Sort the resultant combined scores and then use them to generate a quantitative measure of the search's efficacy for the selected target structure.

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

$$\text{SUMFUS} = \Sigma \text{ni=1 ri} \qquad (2)$$

Where ri denotes the rank position of a specific database structure in the i-th (1≤i≤n) ranked list.

D. Optimization of Fusion Process

For the purpose of the optimization of the fusion process, the genetic algorithm is used and the implementations flow chart is depicted in Figure 4. The input data:

- The auto dimensional input vectors used consist of 1 to 1,360 molecules. These contain 1,056 columns and 1,360 rows, each of which re-percentage only one chemical compound.
- The value for crossover (Pc) is taken to be (0.6 to 0.9). This value generally produces a good result, and for mutation (Pm) the best value is between 0.001and 0.1, due to its being quite small and kept quite low for using GAs.

The overall process could be completed in ten calculations of the fitness of each individual chromosome, using the formula:

$$f(w) = \sum_{i=1}^{13} WiRi \text{ ,and select top 10\% of actives} \qquad (3)$$

Where f(w) is the calculation of the fitness of each individual chromosome, using the formula, Ri similarity value normalized or ranking positions from searching for the structure in the collection based on similarities, Wi is the associated weight generated by GA, 'i' variables range from 1 to the number of similarity measures that were used in the experiment.

4. Experimental Results and Discussions

The similarity search was conducted using a variety of query formats. These similarity searches were conducted using 13 coefficients against the relevant test database. The database molecules were ordered according to their estimated similarity coefficient in decreasing order. We compared the ranks of two coefficient searches for the same query by counting the number of compounds that appeared in the top-ranked structures. Two groups of chemical data sets are used in this project. The first group contains 1,360 compounds divided into seven groups (activities) depending on their biological similarities, as represented in Table 2. The second group contains 1,000 compounds and is divided into three activities (Table 4). The main difference between these two groups is that while group one has smaller bioactivity of similarity, group two has big differences in the bioactivity of similarity among the compounds. The chemical compounds are available in molecule format (Figure 2) and they are converted to BCI Bit String format.

Table 2. First tranche of data and their activities.

Activity	Start	End	No. Compounds
1	0	270	271
2	271	502	232
3	503	636	134
4	637	859	223
5	860	959	100
6	960	1159	200
7	1160	1359	200

For the purpose of optimizing the fusion process, the genetic algorithm is used along with the implementations flow chart, as shown below:

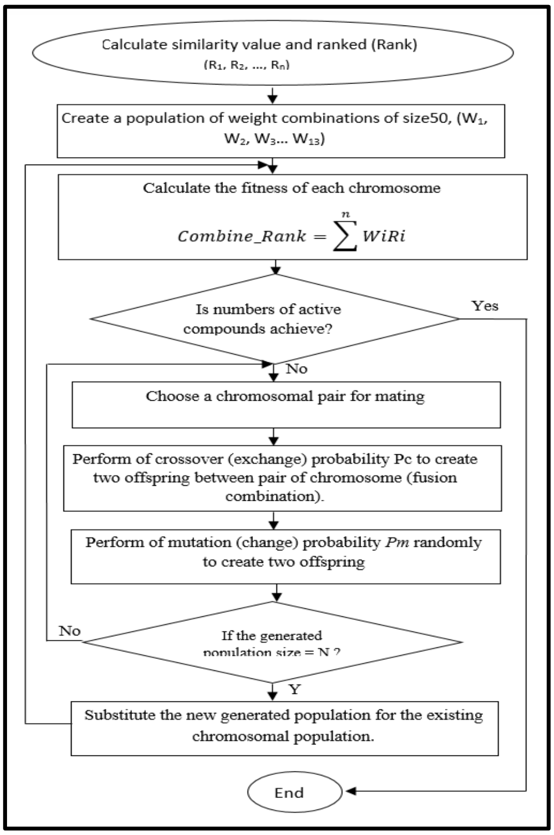

Calculate similarity value and ranked (Rank)
(R₁, R₂, ..., Rₙ)
Create a population of weight combinations of size50, (W₁, W₂, W₃... W₁₃)
Calculate the fitness of each chromosome
Combine_Rank = ∑ⁿ WiRi
Is numbers of active compounds achieve?
Yes
No
Choose a chromosomal pair for mating
Perform of crossover (exchange) probability Pc to create two offspring between pair of chromosome (fusion combination).
Perform of mutation (change) probability Pm randomly to create two offspring
No
If the generated population size = N ?
Y
Substitute the new generated population for the existing chromosomal population.
End

Figure 4. Flowchart of the proposed method

Figure 4 illustrates Execute, a similarity search of a chemical database for a specific target structure using two or more different measures of intermolecular structural similarity, and recording the rank position, ri, of each database structure in the ranking, resulting from the application of the i-th similarity measure in the search results table. This puts the results in order, and then uses that order to create a quantitative measure of how successful the search was for the target structure decided upon. Step 1 represents the weight as chromosome (parent), and any chromosome consists of 13 genes (any gene ≡ single coefficient (13 coefficients)). Step 2 generates a population of combination weights of size 50 (W1, W2, W3… W13). Step 3 calculates the fitness of each individual chromosome, using the formula in Equation 3. Step 4 chooses a chromosomal pair for matching. Step 5 performs crossover, mutation and then places the created offspring chromosome in the new population. Step 8 repeats step 4 until the size of the new chromosome population=N ~ N=50. Step 9 replaces the initial (parent) chromosomes population with the new (offspring) population. Step 10 goes to step 3, and repeats until the termination criterion is satisfied (until the number of iterations is achieved).

The input data for similarity and fusions are vector contains the first group, which consists of 1,360 compounds, and each compound represented as having a binary vector contains 1,056 columns. The Different queries (targets) were taken for each Active top 10% average of actives was used to obtain the best single coefficients. The single coefficients, as shown in Table 3, is the percentage of actives obtains for each active with 10 targets.

Table 3. Percentage of actives obtained for each active with 10% targets using single coefficients

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**

As shown in Table 3, the Tanimoto, Cosine, Kulcznski (2) and Fossum have the best average percentages of top 10% among all the actives (7). The single coefficient has an average value of 21.7. The best single coefficients of top 10% actives is represents in more shape in Figure 5.

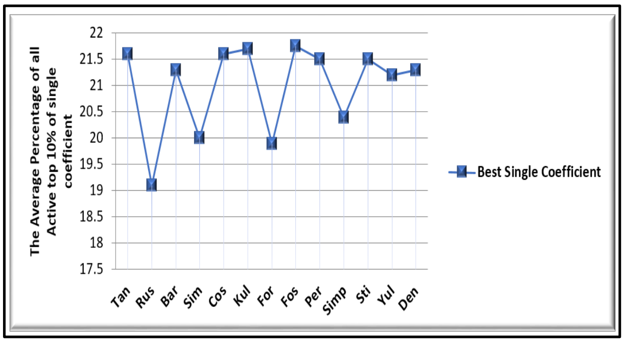


Figure 5. Average percentage of top 10% actives versus using single coefficients

Figure 5 shows that the Tanimoto, Cosine, Kulcznski (2) and Fossum measures have the best percentages of top 10% actives (7) for single coefficient by an average value of 21.7. In contrast, the Rus measure achieved the worst results in percentages among the top 10% actives. The second tranche of data represented here is composed from active compounds with different degrees of activity for each of them (1,000 rows), as shown in Table 4, divided into three Activities (Table 4) depending on their biological similarities as represented.

Table 4. Second tranche of data and their activities

Activity	Start	End	No. Compound s
1	1	247	247
2	248	500	253
3	501	1000	500

Table 5. Average percentage of all actives top 10% depends on fusion of coefficients

Actives	THE PERCENTAGE OF ACTIVES											
	Fusion2						Fusion3					Fusion4
	CosFos	RusKul	TanFos	RusCos	TanRus	TanBar	RusTanCos	RusForCos	RusForTan	ForCosSti	RusForSti	TanBarCosKul
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
Active 1	14.6	14.7	14.5	15.5	**14.6**	16	**15.1**	15.1	15.6	15.1	15.1	**17**
Active 2	22.3	22.3	22.7	24.7	**22.5**	24.3	**22.7**	23	22.6	22.9	22.6	**24.4**
Active 3	20.7	20.4	19.3	22.9	**19.5**	22.2	**21.7**	21.9	20.9	21.7	21.1	**21.9**
Active 4	17.8	18	18.3	19.8	**18.8**	19.5	**19.6**	19.4	19.4	20	19.3	**19.7**
Active 5	23.3	23	22.7	18.9	**22.8**	19.6	**21.6**	21.5	21.5	21.8	21.5	**19**
Active 6	19.7	19.8	20.6	13.3	**21.1**	14.2	**19**	18.5	19.2	18.5	19.3	**14.3**
Active 7	34.8	34.7	35.4	35.6	**34.[illegible]**	[illegible]5.6	**35.6**	35.1	35.1	35	34.95	**35.1**
Average	**21.9**	**21.8**	**21.9**	**21.5**	**22.01**	**21.6**	**22.19**	**22.1**	**22**	**22.1**	**21.98**	**21.63**

We noted that in Table 5 the average percentage of all actives among the top 10% on different combination of non-weights coefficients on fusions of coefficients. In addition, we noted that the best average results in combinations of coefficients – fusion2 - Tanimoto & Russell-Rao (TanRus), fusion3 - Russell-Rao & Cosine (RusCos), and fusion4 - Tanimoto & CosKul – have been achieved with scores of 22.01, 22.19 and 21.63, respectively. The average percentage of all actives top 10% depends on fusion-1, fusion-2 and fusion-3 represents in more shape in Figure 6.

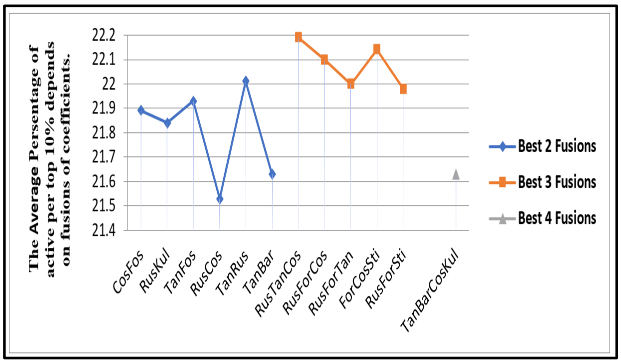


Figure 6. Average percentage of top 10% actives versus using Fusion Coefficients

The results presented in Figure 6 show the best combination for three coefficients: Russell-Rao, Tanimoto and Cosine (RusTanCos) with the best percentage of top 10% by average value 22.19% and the worst Russell-Rao, Forbes and Stile (RusForSti) by 21.98%.

In addition, the experimental result of similarity coefficient fusions has been optimized using GA weights. The result is based on the input data – which are the molecular size factors – and the output data, which are the values between 0 and 1 that represent coefficients and a combination of several coefficients based on the number of actives yield for each coefficient. After applying the comparison of top 10% of each active (7) for different combinations or fusion coefficients with the GA top 10% of each active (7). Ten different queries (targets) were taken after the average for each active top 10% average of actives was used to obtain the best combinations of coefficients. The combinations of coefficients, as shown in Table 4, are the percentage average of actives obtained. Table 6 demonstrates the Percentage Average of all active top 10% on GA-based fusions of coefficients GA weights.

Table 6 Average percentage of all active top 10% on GA-based fusions of coefficients

	THE PERCENTAGE OF ACTIVES											
	Fusion2						Fusion3					Fusion4
Fusion of Coefficients	CosFos	RusKul	TanFos	RusCos	TanRus	TanBar	RusTanCos	RusForCos	RusForTan	ForCosSti	RusForSti	TanBarCosKı
Weights	0.960 and 0.93	0.972 and 0.96	0.960 and0.96	0.972 and 0.96	0.960 and 0.972	0.960 and 0.1440	0.972, 0.960and 0.960	0.972, 0.960and 0.960	0.972, 0.960and 0.960	0.960,0.972 and 0.960	0.972, 0.960and 0.960	0.972, 0.1440,0.960 nd 0.960
Actives	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
Active 1	14.8	14.8	14.3	16.8	14.6	16	15.1	15.1	15.6	15.1	15	17
Active 2	22.3	22.4	22.7	24.7	22.5	24.3	22.6	23	22.6	22.8	22.6	24.4
Active 3	20.6	20.4	19.4	23.1	19.6	22.2	21.6	21.9	20.9	21.7	20.7	21.9
Active 4	18.8	19	19.1	19.8	18.8	19.5	19.4	19.4	19.4	19.3	19.3	19.7
Active 5	23.3	23	22.8	18.9	22.9	19.5	21.7	21.9	21.5	21.8	21.6	19
Active 6	19.7	19.8	20.6	13.3	21.2	14.2	19	18.5	19.2	18.4	19.3	14.3
Active 7	35.6	34.7	35.4	35.5	34.9	35.7	35.6	35.1	35.1	35	35.3	35.5
Average	22.16	22.01	22.04	21.73	22.1	21.6	22.14	22.13	22.04	22.01	21.97	21.7

The genetic algorithm having maximum and minimum value of target compound. Comparisons of different coefficient [26] fusions were carried out in Table 6. The result shows that the Tanimoto, Cosine, Kulczneki (2) and Fossum are the best single coefficients. Cosine and Fossum gave the best combination for two fusion coefficients weighting 0.960 and 0.937, respectively. For combinations three fusion coefficients, Russell-Rao, Tanimoto and Cosine gave the best combinations of weighting with 0.972, 0.960 and 0.960, respectively. However, the Russell-Rao, Forbes, simple matching and Simpson were found to be the worst for finding the similarity of actives. The combination. Tanimoto, Cosine coefficient were to perform well for large active. Using combination with weights ranging between 0.0 and 1.0, generated by genetic algorithm, gave a better number of actives than the manual combination. This was observed when Cosine and Fossum were combined without weights, thus yielding 21.89% actives on average, whereas weights generated by genetic algorithm (GA) achieved 22.16% actives on average with combine Cosine and Fossum, having weights of 0.960 and 0.937, respectively. The average percentages of top 10% actives versus the GA-based fusions of coefficients (GA weights) are demonstrated in Figure 7.

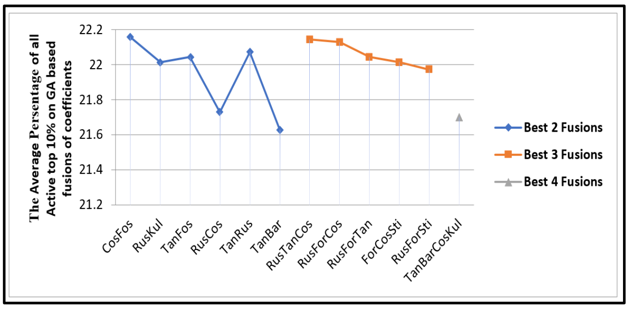


Figure 7. Average percentages of top 10% actives versus the GA-based fusions of coefficients (GA weights)

Figure 7 clearly shows the best combinations of coefficients' by using GA-based fusion weights that for 2-Coefficients fusions Cosine & Fossum (CosFos) are satisfies the best percentage of top 10% actives (7) with an average value of 22.16%, and the worst is Tanimoto and Baroni (TanBar) with 21.6%. The best combination of 3-coefficients that Russell-Rao, Tanimoto & Cosine (RusTanCos) are satisfies the best percentage of top 10% actives (7) with an average value of 22.14% and the worst Russell-Rao, Forbes and Stile (RusForSti) with 21.97%. Therefore, instead of using combinations of 3-coefficients' fusion for the best value, we can use a combination of 2-coefficients fusion, as shown in the peak points.

5. Conclusion and Future Work

The purpose of this research was to demonstrate the use of the GA technique for investigating methods to improve similarity searches in virtual screenings. Additionally, this research examined the application of GA optimal weights in conjunction with the idea of data fusion. The purpose of this research was to ensure that reliable reconstruction weights for all molecular features [20] were available in a variety of molecular descriptors in order to reweight molecular features and select only the most important ones, i.e., those with a higher weight and lower error rates, as well as to remove outlier features. These outlier features are those with significant reconstruction errors, which may be discovered by studying the reconstruction error distribution. The experimental findings demonstrated that optimizing the weights of various similarity measures in data fusion increased the efficiency of searching in a chemical database, demonstrating that GA weights may be effectively used to improve similarity search performance. The tests were performed using the MDDR benchmark dataset, which was shown to be more successful than the other techniques examined. The result of this research shows that GA optimizes a combination, which increases the number of actives and strengthens the accuracy of the solution. Also, it was found that the increase of GA from the ideal after doing 7 tests is 40% on average. It has been proved that the best combination can be satisfied by using 2 fusion coefficients, Cosine and Fossum, and the same for 3 fusion coefficients, Russell-Rao, Tanimoto and Cosine. The GA locates suitable weights by 150 generations with a little improvement achieved by 1,500 generations. This is because of the nature of chromosomes, which is weights. The evaluation of the screening results revealed that the proposed measure improved performance and, more specifically, that GA optimize weights data fusion with architecturally heterogeneous data sets (MDDR -DS1 and MDDR -DS3) obtained superior results when compared to the coefficients searching measures.

In future work, it is recommended that different sets of GA parameter tuning are tested; for instance, by introducing the concept of migrations that will share population establishment by loading it into a different machine. In addition, using a different method for the crossover (uniform) operator keeps the probability between (0.60 – 0.90). This will subsequently enhance the searching capability for suboptimal weights. More input data can be considered to find more effective results in training and testing data using GA, such as the size of database and number of actives in the database.

References

[1] W. Liang, L. Xiao, K. Zhang, M. Tang, D. He, and K.-C. Li, "Data fusion approach for collaborative anomaly intrusion detection in blockchain-based systems," IEEE Internet of Things Journal, 2021.

[2] C. M. Ginn, D. B. Turner, P. Willett, A. M. Ferguson, and T. W. Heritage, "Similarity searching in files of three-dimensional chemical structures: evaluation of the EVA descriptor and combination of rankings using data fusion," Journal of Chemical Information and Computer Sciences, vol. 37, pp. 23-37, 1997.

[3] F. Sirci, L. Goracci, D. Rodríguez, J. van Muijlwijk-Koezen, H. Gutiérrez-de-Terán, and R. Mannhold, "Ligand-, structure-and pharmacophore-based molecular fingerprints: a case study on adenosine A 1, A 2A, A 2B, and A 3 receptor antagonists," Journal of computer-aided molecular design, vol. 26, pp. 1247-1266, 2012.

[4] P. Willett, "Combination of similarity rankings using data fusion," Journal of chemical information and modeling, vol. 53, pp. 1-10, 2013.

[5] R. Wang and S. Wang, "How does consensus scoring work for virtual library screening? An idealized computer experiment," Journal of Chemical Information and Computer Sciences, vol. 41, pp. 1422-1426, 2001.

[6] H. Öztürk, E. Ozkirimli, and A. Özgür, "A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction," BMC bioinformatics, vol. 17, pp. 1-11, 2016.

[7] C. M. Ginn, P. Willett, and J. Bradshaw, "Combination of molecular similarity measures using data fusion," Virtual Screening: An Alternative or Complement to High Throughput Screening?, pp. 1-16, 2000.

[8] D. L. Hall and S. A. McMullen, Mathematical techniques in multisensor data fusion: Artech House, 2004.

[9] M. Liggins II, D. Hall, and J. Llinas, Handbook of multisensor data fusion: theory and practice: CRC press, 2017.

[10] N. Salim, J. Holliday, and P. Willett, "Combination of fingerprint-based similarity coefficients using data fusion," Journal of chemical information and computer sciences, vol. 43, pp. 435-442, 2003.

[11] R. J. Linn, D. L. Hall, and J. Llinas, "Survey of multisensor data fusion systems," in Data Structures and Target Classification, 1991, pp. 13-29.

[12] C. Chen, T. Wang, F. Wu, W. Huang, G. He, L. Ouyang, et al., "Combining structure-based pharmacophore modeling, virtual screening, and in silico ADMET analysis to discover novel tetrahydro-quinoline based pyruvate kinase isozyme M2 activators with antitumor activity," Drug design, development and therapy, vol. 8, p. 1195, 2014.

[13] M. N. Drwal and R. Griffith, "Combination of ligand-and structure-based methods in virtual screening," Drug Discovery Today: Technologies, vol. 10, pp. e395-e401, 2013.

[14] P. Willett, "Similarity methods in chemoinformatics," Annual review of information science and technology, vol. 43, pp. 3-71, 2009.

[15] P. N. Wassenaar, E. Rorije, N. M. Janssen, W. J. Peijnenburg, and M. G. Vijver, "Chemical similarity to identify potential Substances of Very High Concern–An effective screening method," Computational Toxicology, vol. 12, p. 100110, 2019.

[16] P. N. Wassenaar, E. Rorije, M. G. Vijver, and W. J. Peijnenburg, "Evaluating chemical similarity as a measure to identify potential substances of very high concern," Regulatory Toxicology and Pharmacology, vol. 119, p. 104834, 2021.

[17] A. Bender, H. Y. Mussa, R. C. Glen, and S. Reiling, "Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance," Journal of chemical information and computer sciences, vol. 44, pp. 1708-1718, 2004.

[18] M. Krallinger, O. Rabal, A. Lourenco, J. Oyarzabal, and A. Valencia, "Information retrieval and text mining technologies for chemistry," Chemical reviews, vol. 117, pp. 7673-7761, 2017.

[19] P. Willett, J. M. Barnard, and G. M. Downs, "Chemical similarity searching," Journal of chemical information and computer sciences, vol. 38, pp. 983-996, 1998.

[20] P. H. Sneath and R. R. Sokal, Numerical taxonomy. The principles and practice of numerical classification, 1973.

[21] J.-. Willett, Similarity and clustering in chemical information systems: John Wiley & Sons, Inc., 1987.

[22] K. KAWAI, K. YOSHIMARU, and Y. TAKAHASHI, "MDL Drug Data Report MDL Drug Data Report, 2001," Journal of computer chemistry, Japan, vol. 10, pp. 79-87, 2011.

[23] P. G. Dittmar, N. A. Farmer, W. Fisanick, R. C. Haines, and J. Mockus, "The CAS ONLINE search system. 1. General system design and selection, generation, and use of search screens," Journal of Chemical Information and Computer Sciences, vol. 23, pp. 93-102, 1983.

[24] R. Kunimoto and J. r. Bajorath, "Combining similarity searching and network analysis for the identification of active compounds," ACS omega, vol. 3, pp. 3768-3777, 2018.

[25] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas, "Molecular fingerprint similarity search in virtual screening," Methods, vol. 71, pp. 58-63, 2015.

[26] P. Welmina, "Comparison of the effectiveness of probability model with vector space model for compound similarity searching," University Technology Malaysia (UTM): MS Thesis, 2004.

[27] Balabantaray, R. C., et al. (2015). "Document clustering using k-means and k-medoids." arXiv preprint arXiv:1502.07938.

[28] Accelrys Inc: San Diego, CA, USA. MDL Drug Data Report (MDDR). Available online:http://www.accelrys.com (accessed on 15 January 2020).

[29] Salim, N, (2002). Analysis and Comparison of Molecular Similarity Measures, University of Sheffield: Ph. D Thesis.

[30] Shirkhorshidi AS, Aghabozorgi S, Wah TY (2015) A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. PLoS ONE 10(12): e0144059. https://doi.org/10.1371/journal.pone.0144059

[31] Goldberg D. E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Reading: Addison-Wesley.

[32] Goldberg D. E. and Dep K. (1991). A Comparitive Analysis on Selection Schemes Used in Genetic Algorithms. Foundations of Genetic Algorithms, Rawlins G., ed. Morgan Kaufmann. 69 - 93.

Dr. Yahya Ali Abdelrahman Ali graduated with a Bachelor of Computer Science from University of Science and Technology, He obtained his Master's Degree in (Computer Science) University Technology Malaysia (UTM) 2007 and his PhD in Computer Science with excellent academic achievements in Sudan University of Science and Technology, He was the Head of Programming department at the Computer Center in Sudan University of Science and Technology , Currently he works as Assistant Professor in Najran University (NU) College of Computer Science and information system Arabia KSA. His research interest includes Natural Language Processing (NLP), Information Retrieval (IR), Plagiarism Detection, and Data Mining, Natural Language Processing and Text Mining.

Assoc. Prof. Dr. Ahmed Hamza Osman graduated with a Bachelor of Computer Science from International University of Africa. He obtained his Master's Degree in Computer Science from Sudan University of Science and Technology, Sudan and his PhD in Computer Science with excellent academic achievements from Universiti Teknologi Malaysia (UTM). He was the Head of Computer Science department at the Faculty of Computer Studies at international University of Africa. Currently he works as Associate Professor in King Abduaziz University (KAU) Saudi Arabia. His research interest includes Information Retrieval, Plagiarism Detection, Soft Computing, and Data Mining, Natural Language Processing and Text Summarization.

Suad Mohammed Fadalmola Dafallh graduated with a **B.S.C (**Honors) in computer science, Sudan University of Science and Technology ,Faculty of Computer Science andTechnology and information System,Department of Computer Sciencer (October 2002).she is btained his Master in computer science, Sudan University of Science and Technology ,Faculty of Computer Science and Technology and information System, complete curses (SUST DEC 2009)**2002).** She was Teaching Assistance at the Computer Center in Sudan University of Science and Technology, currently he works as Lecture in Najran University (NU) College of Computer Science and information system Arabia KSA.