A Framework For Genetic-Based Fusion Of Similarity Measures In Chemical Compound Retrieval

Naomie Salim
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia
naomie@fsksm.utm.my

Mercy Trinovianti bt. Mulyadi
Program Pengajian Diploma UTM KL
mercy@citycampus.utm.my

Abstract

In chemical compound retrieval, much data fusion effort has been made to combine results from multiple similarities searching system. A fundamental problem in 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 .This paper proposes a framework of data fusion approach based on linear combinations of retrieval status values obtained from Vector Space Model and Probability Model system. A Genetic Algorithm(GA)-based approach is used to find the best linear combination of weights assigned to the scores of different retrieval system to get the most optimal retrieval performance.

1. Introduction

Similarity searching is a tool which is widely used in the pharmaceutical and agrochemical industries, with database searching, design of combinatorial library and bioactivity prediction being among its most common use. The basic hypothesis for similarity searching in chemical databases is the “similarity property principal”, which states that compounds that are structurally alike in some way will have similar biological activities. Similarity is a subjective quality and attempts to quantify it, so that it can be processed by a computer, can only be incomplete. However, it can be said that two compounds are similar with respect to a particular descriptor or a particular feature. There are many ways in which the structural similarities between pairs of molecules can be calculated, and there has been much debate as to which similarity measure is the best for this purpose. Variation of similarity measures can be in terms molecular descriptors used, and the calculation of similarity between those descriptors.

Most existing chemical compound similarity searching systems apply the Vector Space Model (VSM). Even though this approach has acceptable retrieval effectiveness [1], the VSM only considers structural similarity, ignoring both activity and inactivity. Other than that, the evaluation order of the query and the database compound was not taken into account. It also assumes that fragments are independent of all other fragments, which is not necessarily true [2]. Recently, a similarity searching method based on the Probability Model (PM) has been developed [3] to overcome the VSM limitations. Among the PM approaches, the Binary Independence Retrieval (BIR) and Binary Dependence Model (BD) has been used for chemical compound retrieval. Apart from having a strong theoretical basis, PM is more realistic approach in retrieval system. It will rank chemical compounds in decreasing order of their probability of being similarly active to target compound. According to the Probability Ranking Principle (PRP), if the ranking of the compounds is in decreasing probability of usefulness to the user, then the overall effectiveness of the system to its users will be the best [4].

The variations in similarity measures as explained above give rise to questions on how these measures can be chosen, optimized and combined in order to best reflect the biological and physicochemical similarity between molecules.

2. Optimization of Similarity Measures

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. Certain molecular features can be emphasized by associating higher weights with them when calculating similarity. Many of the weightings used in chemical information systems are derived from the general information retrieval literature, such as the term-frequency and inverse document frequency. For example, higher weights can be given to attributes that occur frequently in a molecule, attributes that occur in small molecules and also attributes that occur less frequently in a data set.

Initial work of fusion techniques in the field of chemical information has been reported by Kearsley *et al*. [5] and Ginn *et al*. [6][7], both of whom have carried out similarity searches for drug molecules in the Standard Drug File database. Ginn *et al*., for instance, used data fusion to combine rankings of chemical compounds that have been generated using several different measures of inter-molecular structural similarity [7]. The result shows that the fused similarity measures can enable better predictions to be made of the cell-staining activities of the molecules than can the original measures. In another study, Kearsley *et al.* [5] found that for each of the two-descriptor combinations they investigated, approximately half of the fused searches were better than the original, individual measures. It was never the case that any combination of descriptors was less successful than the worst descriptor in the combination considered on its own. Ginn *et al.*, have reported database searching experiments in which rankings based on the EVA descriptor were fused, which is based on information derived from infra-red vibrational spectra, and on 2D Fingerprints [6]. They found that the use of data fusion on the 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.

More recently, Salim did a study to ensure that fusion does give improvement over the use of single coefficients [1]. She also deduces how much improvement over the use of single coefficients is possible when more similarity coefficients are included in the combinations and the optimal number and combination of similarity coefficients that could be used to give such improvement. 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 at it was found to be the most effective in an earlier study [6][7]. She found that although combinations are generally better alternatives than single coefficient, the practicality of their use remains questionable as no particular combinations of coefficients showed consistent high performance across all types of actives.

Another data fusion research have been carried out by Daut [8] to find the best coefficients or combined coefficients to be used in similarity searching by using Neural Network algorithm with molecular size factors as inputs. Among the size factors considered are average molecular size of the target actives and of the compounds in the databases. From the results of the experiment, it can be concluded that there is no specific coefficient or combination of coefficients that is best for all cases. However, results from the second experiment show a pattern when choosing the best coefficient or combination of coefficients to perform similarity searching based on the database attributes. Although there are many studies in similarity searching [9][10][11], no specific study shows how to choose the best coefficient or combination of coefficients for use in similarity searching.

ON the other hand, ways to optimize the choice and combination of different retrieval systems has been done in other types of information retrieval research. For instance, Fan *et al*. proposes an optimized genetic-algorithm-based data fusion approach based on linear combinations of retrieval status (RSV) obtained from four different matching functions or expert [12]. 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.

In this paper, we proposed a framework to optimize fusion function for molecular similarity searching using ideas derived from this genetic-algorithm based approach. The basic summation-based fusion approach [1][3][8][12] is extended to include fusion optimization of different similarity coefficients among Vector space Model and with different similarity searching using the Probability Model.

3. Optimization of Similarity Measures Fusion Based on Genetic Algorithms

The method is based on linear combinations of ranking from different similarity measures 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 lightly better retrieval effectiveness than rank values, fusion using similarity values is only appropriate when sources combined have similar rank-similarity curves [13]. Study [14] shows that some similarity coefficients generate quite different rank-similarity curves, confirming the appropriateness of using rank fusion instead of similarity fusion when combining coefficient. For any given query, each similarity searching assigns a score for each structure. The structures in the collection are ordered in the decreasing order of their similarity scores. These rankings, adjusted by certain weights, will then be summed.

We assign weights (in the range of –1.0 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. Mathematically the combined ranking can be expressed as follows:

$$Combined\_Rank_m = \sum_{i=1}^{n} W_i S_i \qquad (1)$$

Here *'n'* is the number of similarity measures used. S_i is the ranking produced by the *i*th similarity searching for the *m*th structure in the collection. W_i is the associated weight. *'i'* varies from 1 to the number of similarity measure used in the experiment. *'m'* varies from 1 to the maximum number of structures in the collection. By proper selection of weights *'W'* , it should be possible to increase the retrieval performance. This is so because the similarity searching are complementary to each other in terms of their weighting strategies for clues offered in the structures and queries. A proper selection of weights, thus tries to exploit such complementarities.

We will 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. GA starts with a population of individuals known as chromosomes. Each chromosome represents a possible solution to the problem, and in this case, the weights of each similarity measures. The initial population is either randomly generated or it can also be generated using some known characteristics from earlier results [1][3][8], for example, giving higher weights to coefficients that performed better in earlier experiments. The individuals in the population change with successive iterations of the algorithm (known as generations) following the process of selection, crossover and mutation. Selection is based on the fitness of each chromosome in the population. Fitness is a numerical score assigned to each chromosome. It is expected that the more fit (the higher the fitness number) a 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 performance of the GA. The GH Score and the number of actives at top 5% of the list will be used as the fitness function, as will be explained later. These are the two most important performance measures used in chemical retrieval [3]. GH Score gives an indication of how good the retrieved list is with respect to a compromise between maximum yield and maximum percent of active retrieved. Mathematically, GH Score can be expressed as:

$$GH = \frac{H_a\left(A + H_t\right)}{2AH_t}$$

Where A is the number of actives structures in the database,
H_t is the number of structure in a retrieved list.
H_a is the number of active structures in a retrieved list,

A high number of actives at top 5% of the list denotes a good similarity searching system. This performance measure is important when the user is interested in looking at more than a few structures presented to the user.

Crossover operator is used to transfer more fit building blocks from one generation to another, while the mutation operator is used to introduce random diversity in the population so that the population does not get stuck in a local optimum. The process is stopped when either the preset number of generations is reached or if there is no improvement in the performance. The chromosomes in the last generation are chosen as the best individuals to solve the problem at hand.

4. Experimental Design

In this section, the design of the experiment to test the viability of the algorithm will be explained. First we start with description of the similarity searching we used for combinations. Then we describe the chemical data used in the experiments, the fitness functions used to train the GA and finally a detailed description of the training, validation, and testing phase used during the genetic process.

4.1 Similarity Searching

We decided to use two very well known model similarity searching that was explained before this. Experiment for these two models will be conducted separately because of their different characteristics. Their choice was motivated by the fact that these similarity searching have performed very well in the recent studies [1][3][8].

Here we will consider two type of model in similarity searching. First is the Vector Space Model, with 13 similarity coefficients considered in the study from a previous experiment [1] and second is 2 similarity searching of Probability Model from a previous study [2], the Binary Dependence Model and the Binary Independence Model. The probability based models have probability as the basis for rankings, whereas the Vector Space Model uses the notion of similarity. These different warrant for the use of rankings instead of actual scores for combinations.

4.2. Experimental Data

The experiment involves two databases: the first was a set of 5772 compounds from the NCI AIDS database. The second is a set of 11607 compounds from the IDALERT database and the third is a subset of 113842 molecules MDL Drug Data Report (MDDR) database. All structures in three databases were characterized by three types of real bit strings: Barnand Chemical Information (BCI) bit strings, the Daylight fingerprints and the UNITY 2D bit strings. The BCI bit string is a 1052-bit structural key-based bit string generated based on the presence or 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, similar to those in the CAS Online Dictionary. The Daylight fingerprint, on the other hand is a 20048 bit hashed fingerprint that encodes each atom's type, augmented atoms and paths of length 2-7 atoms. Meanwhile, UNITY 2D bit string, unlike Daylight fingerprint that hashes all recorded information over the whole length of the fingerprint, keeps information from different-length paths distinct. Different parts of the bit string recorded information of fragments of length 2 to 6. A few generic structural keys are added for some common atoms and bond types, producing a bit string of 992 bits.

We use the residual collection method to divide the entire data into three parts: training (50%), validation (20%) and test data (30%). The training data, along with the relevance information for queries, is used by GA-based system to generate a set of "candidate" weighting schemes. The validation data is used to choose the candidate scheme that has the best generalization capability for new data. The performance comparisons of all systems are based on the results on the test data only. Structure cut-off value was set to 400 i.e. the user is presented with top 400 structures. The fitness calculations and other performance measures are based on the top 400 structure retrieved. Leave one out cross validation will be used to validate the results.

The experiment is conducted in three phases. In the first step, the training data is used to train the weights associated with the chosen similarity searching. The validation phase is used to choose the best weighting scheme that generalizes well on the validation data set, while the test phase applies these weights on the test data set.

4.3. The Training Phase

The training phase uses the training data. based on Figure 1, it starts with the generation of random weights associated with each of the 13 similarity searching and two based on probability searching for all of the chromosomes/individuals in the population. A sample chromosome/individual in the population is given in Table 1. A chromosome is a series of real numbers in the range -1.0 to 1.0. Each of the real number in the chromosome is the weight associated with an individual similarity searching function.

Table 1: Sample Chromosome for Similarity Searching Vector Space Model

Chromosome Weight	0.213	. .	. .	-0.198
Associated Similarity Searching	Russell coefficient			Fossum coefficient

The fitness of each individual is calculated using the chosen fitness function and the individuals are sorted in the decreasing order of their fitness values. We store the top individuals for later analysis. The next step is to performed genetic modifications to generate a new population. We copy the top 10 of the individuals in a generation into next generation. The remaining individuals are selected using tournament selection. The crossover rate is chosen as 70%. We use Blx-crossover operator as it has proved very effective in other evaluation studies with real-valued genes [12].

In this crossover method the idea is to first get the maximum (c_{max}) and minimum (c_{min}) of the current parents for each of the fitness functions. Letting I= (c_{max}-c_{min}), the crossover is done by randomly selecting a child from the [c_{min}-α*I,c_{max}+α*I] where α is the crossover rate. Mutation are performed by introducing Gaussion noise in randomly selected genes, according to mutation rate of 15%. Training will be done for 20 generations. The parameter involved in genetic operation (the elitist rate, crossover rate, mutation rate, and the number of generations) will be chosen after initial exploratory analysis. At the end of the training phase, we have information about 200 individuals (as stated earlier we store top 10 individuals in each of the 20 generations).

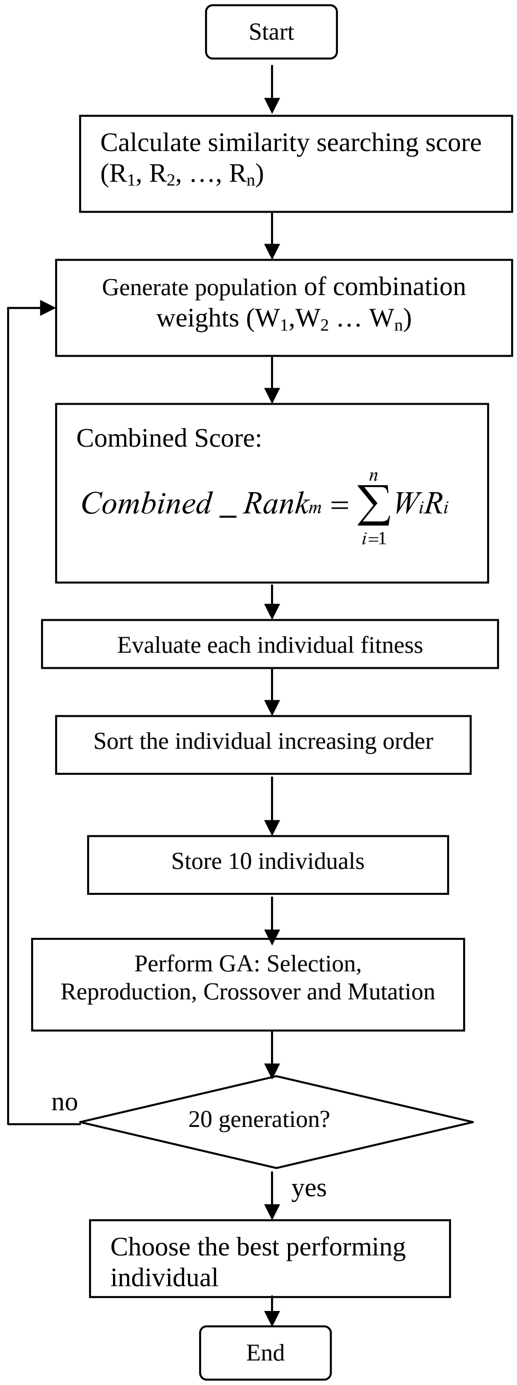
Start
Calculate similarity searching score (R_1, R_2, …, R_n)
Generate population of combination weights (W_1,W_2 … W_n)
Combined Score:
$Combined\_Rank_m = \sum_{i=1}^{n} W_i R_i$
Evaluate each individual fitness
Sort the individual increasing order
Store 10 individuals
Perform GA: Selection, Reproduction, Crossover and Mutation
20 generation?
no
yes
Choose the best performing individual
End

Figure 1: Flow of Training Phase

Out of the 200 individuals available, the best performing individual on the validation data set is used in the testing phase.

4.4. The Testing Phase

The last phase of the experiment is the testing phase (see Figure 3). Test data set is used in this phase. We use the chosen individual from the training phase. Equation (1) is used to calculate ranking for each structure in the test data set. The structures are arranged according to rankings and the performance measures are calculated (with structure cut-off of 400). We use three performance measures: GH Score, Initial enhancement, which refers to a number of chemical structure retrieved before half of the actives are found and the number of actives at top 5% of the list. At the end of the test phase, we will compare the performance results obtained by the methodology with those obtained by the standalone similarity searching result [1][3][8] .

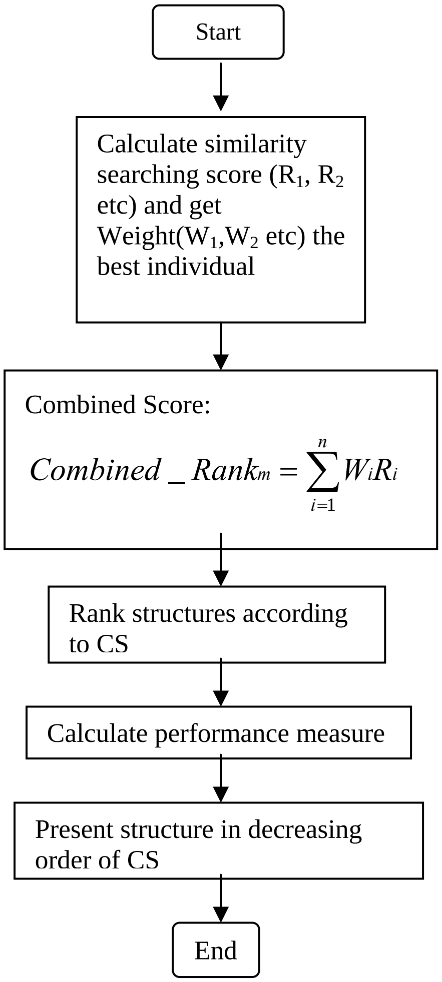
Start
Calculate similarity searching score (R_1, R_2 etc) and get Weight(W_1,W_2 etc) the best individual
Combined Score:
$Combined\_Rank_m = \sum_{i=1}^{n} W_i R_i$
Rank structures according to CS
Calculate performance measure
Present structure in decreasing order of CS
End

Figure 2: Flow of Test Phase

5. Summary

The fusion method is independent of the similarity searching used. There is no restriction on the form of the similarity searching that can be combined using the approach. All that is required of a similarity searching is that produces a retrieval ranking for a structure. The method is also scalable in the sense that there is no restriction on the number of similarity searching that can be used. This method can incorporate any fitness functions while evolving the weights associated with matching functions, with the hope to have better significance in retrieval performance over conventional retrieval models.

References

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

[2] Yates, R.B and Neto, B.R, *Modern Information Retrieval*, England: ACM Book Press, 224-34, 1999.

[3] Godfrey, W.W., *Comparison of the Effectiveness of Probability Model with Vector Space Model for Compound Similarity Searching*, University Teknologi Malaysia: M.Sc. Thesis, 2004.

[4] Cooper, W.S., "The formalism of probability theory in IT: A foundation or an encumbrance?" *Proceeding of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin Ireland: ACM, 1994, pp. 242-247.

[5] Kearsley S.K, Sallamack, S. and Fluder, E.M., "Chemical Similarity Using Physiochemical Descriptors", *Journal of Chemical Information and Computer Science*, 1996, 36.pp. 118-127.

[6] Ginn, C.M.R, Turner D.B., Willett, P., "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,* 1997, 37.pp. 23-37.

[7] Ginn, C.M.R, Ranade, S.R, Willett, P., Bradshaw, J, "The Application of Data Fusion To Similarity Searching In Chemical Databases", 1998, Available Online at http://www.daylight.com.

[8] Daut, N., *Finding Best Coefficient and Fusion of Coefficient for Similarity Searching Using Neural Network Algorithms*, University Teknologi Malaysia: M.Sc, Thesis, 2004.

[9] Bender, A., Mussa, H.Y. and Glen R.C., "Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection and a Naïve Bayesian Classifier*" Journal of Chemical Information and Computational*, 2004, 44. pp. 170-178.

[10] Gillet, V.J., Willett, P. and Bradshaw, J., "Similarity Searching Using Reduced Graphs", *Journal of Chemical Information and Computational,* 2003, 43 pp 338-345.

[11] Rhodes, N., Willett, P., Calvet, A., Dunbar, J.B. and Humblet, C., "CLIP: Similarity Searching of 3D Databases Using Clique Detection" *Journal of Chemical Information and Computational*, 2003, 43.pp 443-448.

[12] Fan, W., Gordon, M. and Pathak, P., "On Linear Mixture of Expert Approaches to Information Retrieval", *in press, Decision Support System*, 2004.

[13] Lee J.H., "Analysis of Multiple Evidence Combination", *Proceeding of 20th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, ACM Press, 1997, pp. 267-276.

[14] 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.