Inference Networks for Molecular Database Similarity Searching

Ammar Abdo* ammar_utm@yahoo.com
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia
Johor Bahru, Skudai, 81310, Malysia
**Corresponding author : Tel : +6- 0143123054, +6-07- 5532637, Fax : +6-07-5532210*

Naomie Salim naomie@utm.my
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia
Johor Bahru, Skudai, 81310, Malysia

Abstract

Molecular similarity searching is a process to find chemical compounds that are similar to a target compound. The concept of molecular similarity play an important role in modern computer aided drug design methods, and has been successfully applied in the optimization of lead series. It is used for chemical database searching and design of combinatorial libraries. In this paper, we explore the possibility and effectiveness of using Inference Bayesian network for similarity searching. The topology of the network represents the dependence relationships between molecular descriptors and molecules as well as the quantitative knowledge of probabilities encoding the strength of these relationships, mined from our compound collection. The retrieve of an active compound to a given target structure is obtained by means of an inference process through a network of dependences. The new approach is tested by its ability to retrieve seven sets of active molecules seeded in the MDDR. Our empirical results suggest that similarity method based on Bayesian networks provide a promising and encouraging alternative to existing similarity searching methods.

Keywords: Bayesian networks, molecular similarity searching, chemical databases, inference network, drug discovery.

1. INTRODUCTION

The term chemoinformatics was coined only a few years ago, but it rapidly gained widespread use. Chemoinformatics is the use of informatics methods to solve chemical problem [42]. Chemoinformatics is now being extensively used by pharmaceutical and agrochemical companies. The pressure to find new active compounds and bring them to market as quickly as possible has led many pharmaceutical and agrochemical companies to use information technology in their product discovery and development processes. Database searching can be divided into three distinct classes of problem: exact-match searching for the database record that is identical to the query record, partial-match searching for those database records that contain the query and best-match searching for those database records that are most similar to the query

record. In chemoinformatics, the first two classes correspond to structure searching and substructure searching, respectively. The provision of best-searching facilities for chemical database is normally referred to as similarity searching, which involves quantifying the similarity of a target molecule with all others in the chemical database in terms of a chosen descriptor or set of descriptors. It is used whenever a potential drug compound, a lead, has been found. The lead can be further optimised by finding similar compounds to it, with the hope that a similar, but better drug can be synthesised.

The virtual screening (VS) is widely used to enhance the cost-effectiveness of drug-discovery programmes by ranking database of chemical structures in decreasing probability of activity, this prioritisation then means that biological testing can be focused on just those few molecules that have significant a priori probabilities of activity. There are many different ways in which a database can be prioritized, here we focus on similarity searching methods. Similarity searching is one of the most widely used VS approaches. The basic idea underlying similarity searching based VS is a very simple idea that similar property principle states that structurally similar molecules tend to have similar properties [1]. According to this principle, any molecule that has not been tested for biological activity but is structurally similar to a target molecule that is exhibit the interest activity is also expected to be active. Furthermore the molecules will be ranked in decreasing order, so that first molecule is more expected to be active than others and so on.

One objective of the computational tools which applied in chemoinformatics was to finding leads early in a drug discovery project. The effectiveness of any similarity method can vary greatly from one biological activity to another in a way that is difficult to predict. Moreover, any two similarity methods tend to select different subsets of actives from a database, consequently it is advisable to use several similarity search methods where possible [2].

In essence, most of the molecular similarity measures used originates from areas outside chemoinformatics, particularly from text retrieval. Although chemical structures differ greatly from other entities that are commonly stored in database, some parallels can be drawn between chemical database searches and searches on words or documents [3]. The many similarities between information retrieval and chemoinformatics that have already been identified suggest that chemoinformatics is a domain of which information retrieval researchers should be aware when considering the applicability of new techniques that they have developed [4]. During last two decades many researches has been done to develop different textual information retrieval techniques. Currently, Bayesian network the best approach to managing probability and to solve the uncertainty problem in textual information retrieval.

2. MOLECULAR SIMILARITY SEARCHING

In similarity searching, a query involves the specification of an entire structure of a molecule. This specification is in the form of one or more structural descriptors and this is compared with the corresponding set of descriptors for each molecule in the database [5]. A measure of similarity is then calculated between the target structure and every database structure. Similarity measures quantify the relatedness of two molecules with a large number (or one) if their molecular descriptions are closely related and with a small number (large negative or zero) when their molecular descriptions are unrelated. The results of the similarity measure will be used to sort the database structures into the order of decreasing similarity with the target. The resulting ranked list of structures will then be returned to the user. There is an extensive and continuing debate about what sorts of measures are most appropriate [46]. The similarity measure based on the number of substructural fragments common to a pair of molecules and a simple association coefficient are the most common at least until now [46]. The performance of different similarity coefficients with regard to their use in molecular similarity searching has earlier been analyzed. Several methods have been used to further optimise the measures of similarity between molecules, which include weighting [49], standardisation [47] and data fusion [46, 48]. Probability-based similarity

searching [50] has also been developed on top of the industry-standard vector-space models (VSM).

A common application of similarity searching is in the rational design of new drugs and pesticides where the nearest neighbours for an initial lead compound are sought in order to find better compounds. Similarity searching is also used for property prediction purposes [7], where the properties of an unknown compound are estimated from those of its nearest neighbours. Underpinning these applications of molecular similarity measure is the similar property principle [1], which states that structurally similar molecules will exhibit similar physiochemical and biological properties. Related to the similar property principle is the concept of neighbourhood behavior [8], which states that compounds within the same neighbourhood or similarity region have the same activity. Unknown biological or physicochemical properties of a molecule can be predicted from the properties of molecules that lie within the same neighbourhood region. In lead finding, selection of compounds whose neighbourhood regions overlap one another should be avoided. In lead optimisation, if a particular compound is found to be active, compounds that lie in the same neighbourhood region can be tested to find one with the most optimum activity.

The first reports on similarity searches appeared in the mid-1980s, based on the work carried out at Lederle Laboratories [7] and Pfizer [9]. In the Lederle study, molecules were represented by their constituent atom pairs, where an atom pair is a substructural fragment comprising two non-hydrogen atoms together with number of intervening bonds. The similarity search allowed users to request either some number of the top-ranked molecules or all those that had a similarity with the target structure greater than a minimal value. In the Pfizer system, together with a conventional substructural query, a user can submit a target molecule typical of the type of the structure that was required. The conventional screen search and atom-by-atom search were used to identify matches in the substructure searching, after which a similarity measure based on the screens common to the target and the matches was used to rank the substructure search output. The subsequent development of a faster, inverted-file-based, nearest neighbour search algorithm allowed the ranking of the entire database against the target structure in real time, without the need for the specification of the initial substructural query. Since the Lederle and Pfizer systems, similarity searching has undergone further development. An example is Hagadone's work on substructure similarity searching [10]. Substructure similarity searching is used to identify molecules containing a substructure similar to a target structure or substructure. Another extension of similarity search was described by Fisanick et al. [11] on facilities developed for Chemical Abstracts Service (CAS) Registry File. It focuses on different types of similarity relationships that can be identified between a structure in the query and a database structure. This study found that different representations could give different measures of structural resemblances between compounds, which suggest that a further analysis into a combined approach could give a more comprehensive similarity measure between them. The use of similarity calculations between molecules have since been used not only in similarity searching, but also in applications like compounds selection [12, 13] and molecular diversity analysis [14, 15, 16]. Three principal tools used for the similarity calculations are the representation that is used to characterize the molecules that are being compared, the weighting scheme that is used to assign differing degrees of importance to the various components of these representations, and the coefficient that is used to determine the degree of relatedness between two structural representations [17].

2.1 Molecular descriptors

Molecular descriptors are vectors of numbers, each of which is based on some pre-defined attributes. They are generated from a machine-readable structure representation like a 2D connection table or a set of experimental or calculated 3D co-ordinates. Molecular descriptors can be classified into 1D descriptors, 2D descriptors and 3D descriptors. 2D descriptors are based on information derived from the traditional 2D structure diagram. Examples of 2D descriptors are 2D fingerprint and topological indices, which are our focus as they play a prominent role in the experimental work of this paper.

2D fingerprints are the most commonly used descriptors. These descriptors were initially developed to provide a fast screening step in substructure search systems in which bit strings are used to represent molecules. They have also proved very useful for similarity searching. There are two different types of 2D fingerprints: dictionary-based bit strings and hashed fingerprints. In dictionary-based bit strings, a molecule is split up into fragments of specific functional groups or substructures. The fragments used are recorded in a predefined fragment dictionary that specifies the corresponding bit positions of the fragments in the bit string. Bits either individually or as a group represent the absence or presence of fragments. Examples of dictionary-based assignment are the CAS ONLINE Screen Dictionary for substructure searching [18], Barnard Chemical Information system [19, 20] and MDL MACCS key system [21, 22]. 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. This approach is used in the Daylight Chemical Information Systems [24] and Tripos systems [23].

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.

Similarity measures based on bit strings are currently the most widely used approach for database searching [25]. One of the principal applications of bit string based searching is in the selection of compounds for inclusion in biological screening programs. This is largely due to the low processing requirements needed to calculate the similarities between a target structure and a large number of structures.

2.2 Weighting schemes

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 emphasised by associating higher weights with them when calculating similarity. Different types of statistical information can be extracted from computerised representations of molecules to form the basis for a fragment weighting schemes. These are follows, (a) Fragment Frequency (*ff*), is the number of occurrence of a particular fragment within a molecule, with high frequently occurring fragments being given a greater weight than those that occur less frequently. (b) Inverse Fragment Frequency (*iff*), is the frequency of the fragment in the molecule collection, with less frequently occurring fragment being given a greater weight than those that occur high frequently throughout the molecule collection. (c) Molecule size (*mz*), is the number of the fragments assigned to a molecule, with a fragment in small molecule being assigned a greater weight than the same fragment in a large molecule. One more weighting scheme can be used whenever we can differentiate between active and inactive molecules within dataset. Unfortunately, limited studies have been done on the effect of applied weighting schemes on molecular similarity searching methods. All of the above mentioned considerations have been used for assigning weights at the National Cancer Institute [26]. Willett and Winterman have found that giving more weight to fragments that occur more frequently in a molecule did seem to give good results, but other weighting schemes had little significance [27].

2.3 Similarity Coefficients

Similarity coefficients are used to obtain a numeric quantification to the degree of similarity between a pair of structures [28]. There are four main types of similarity coefficients [29, 30, 31] : distance coefficients, association coefficients, correlation coefficients and probabilistic coefficients. Association coefficients are commonly used with binary representations and are often normalized to lie within the range of zero (no similar features in common) and unity (identical representations). However, they can be used with non-binary representations, in which

case the range may be different. Correlation coefficients measure the degree of correlation between sets of values characterizing a pair of objects. Distance coefficients quantify the degree of dissimilarity between two objects and, when normalized and using binary data, range between zero (identity) and unity (no similar features in common). 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. Examples of these coefficients can be found elsewhere [29]. Assume $\mathbf{S}_{K,L}$ is the similarity between molecules *K* and *L*, both molecules described by binary representation. For bit string descriptors, **n** is the total bit positions in the bit strings representing the two molecules compared. **b** is the number of bit positions set in only one of the two molecules whilst ***c*** is the number of bit positions set in only the other molecule. **d** of the **n** bits are not set in either one of the molecules and **a** is the number of bits set in both molecules. Thus, **n = a + b + c + d**. The origins of the coefficients can be found in a review paper by Ellis et al. [31]. Examples of some of the coefficients that were used are listed in Table 1.

Coefficient	Continuous		Binary	
	Formula	Range	Formula	Range
Tanimoto	$\frac{\sum_{j=1}^{M}(w_{jk}w_{jl})}{\sum_{j=1}^{M}(w_{jk})^2+\sum_{j=1}^{M}(w_{jl})^2-\sum_{j=1}^{M}(w_{jk}w_{jl})}$	-0.3 to 1	$\frac{a}{a+b+c}$	0 to 1
Cosine	$\frac{\sum_{j=1}^{M}(w_{jk}w_{jl})}{\sqrt{\sum_{j=1}^{M}(w_{jk})^2\sum_{j=1}^{M}(w_{jl})^2}}$	0 to 1	$\frac{a}{\sqrt{(a+b)(a+c)}}$	0 to 1
Forbes	$\frac{n\sum_{j=1}^{M}(w_{jk}w_{jl})}{\sum_{j=1}^{M}w_{jk}^2\sum_{j=1}^{M}w_{jl}^2}$	-∞ to ∞	$\frac{n\times a}{(a+b)(a+c)}$	0 to ∞
Russell-Rao	$\frac{\sum_{j=1}^{M}w_{jk}w_{jl}}{n}$	-∞ to ∞	$\frac{a}{n}$	0 to 1
Dice	$\frac{2\sum_{j=1}^{M}(w_{jk}w_{jl})}{\sum_{j=1}^{M}(w_{jk})^2+\sum_{j=1}^{M}(w_{jl})^2}$	0 to 1	$\frac{2a}{2a+b+c}$	0 to 1

TABLE 1: Examples of Association Coefficients.

Tanimoto coefficient in Eq. 1 is the most popular coefficient used by similarity methods. If two molecules *K* and *L* have **b** and **c** bits set in their fragment bit-strings, with **a** of these bits being set in both of the fingerprints, then the similarity between these two molecules using Tanimoto coefficient is defined to be:

$$S_{K,L}=\frac{a}{a+b+c} \quad (1)$$

The Tanimoto coefficient gives values in the range of zero (no bits in common) to unity (all bits the same). The Tanimoto coefficient gives the best result than the other coefficients. Currently,

The Tanimoto coefficient is widely used in molecular similarity methods and was becomes the best choice in both in-house and commercial software systems for chemical information management.

3. BAYESIAN NETWORKS

Recent research in information retrieval has proved that retrieval models based on Bayesian network give significant improvements in retrieval performance compare to conventional models [36, 37, 38, 43]. It is therefore likely that Bayesian network is able to represent the main (in)dependence relationships between molecular descriptors as conditional probabilities with the degree of resemblance between pairs of such descriptors computed to represent the probability. Molecular similarity will be regarded as an inference or evidential reasoning process in which the probability that a given compound met the requirements of a query is estimated and used as evidence. Network representations have show promise as mechanisms for inferring these kinds of relationships. In this paper, we explore the possibility and effectiveness of using such networks for similarity searching.

A Bayesian network (BN) is graphical model of a probability distribution [33]. A Bayesian network is a directed acyclic graph (DAG) in which the nodes represent random variables and the arcs show causality, relevance or dependency relationships between them. The variables and their relationships comprise the qualitative knowledge stored in a Bayesian network. The strength of the relationships, measured by means of probability distributions, is also stored in the DAG. Associated with each node is a set of conditional probability distributions, one for each possible combination of values that its parents can take. A Bayesian network can be considered an efficient representation of a joint probability distribution that takes into account the set of independent relationships represented in the graphical component of the model. In general terms, given a set of variables $\{X_1, \ldots, X_n\}$ and a Bayesian network G, the joint probability distribution in terms of local conditional probabilities is obtained as follows:

$$P(X_1,...X_n) = \prod_{i=1}^{n} P(X_i \left| \pi(X_i) \right.)$$

where $\pi(X_i)$ is any combination of the values of the parent set of X_i. If X_i has no parents, then the set $\pi(X_i)$ is empty, and therefore $P(X_i|\pi(X_i))$ is just $P(X_i)$. Once completed, a Bayesian network can be used to derive the posterior probability distribution of one or more variables in the network, or to update previous conclusions when new evidence reaches the system.

4. SIMILARITY INFERENCE NETWORK MODEL

The basic model for similarity inference network, shown in Fig.1, consists of two component networks: a compound network and a query network. The compound network represents the compound collection. The compound network is built once for a given collection and its structure does not change during query processing. The query network consists of a single node, which represents the target molecule and one or several query molecules, which express the target molecule. A query network is built for each target molecule and modified during query processing as the query is refined or additional representations are added in an attempt to better characterize the target molecule. The compound and query networks are connected though links between their descriptor nodes.

4.1 Compound Network

The compound network shown in Fig. 1 is a simple direct acyclic graph (DAG) consisting of compound nodes (c_j) as roots, and descriptor nodes (d_i) as leaves. Each compound node represents a compound in the collection. Each compound node has a prior probability associated

with it that describes the probability of observing that compound. This prior probability will generally be set to *1/(collection size)* and this probability will be small for real collections.

Compound nodes have one or more descriptor nodes as children. The descriptor nodes can be divide into several subsets, each corresponding to a single descriptor technique that has been applied to the compound. When 1052 bits are used to describe the compounds using BCI fingerprint, 1052 nodes are used to represent these bits. If 10 topological indices are used to describe the compounds, 10 nodes are used to represent these numerical values. We represent the assignment of a specific descriptor to a compound by draw a directed arc to the descriptor node from each compound node corresponding to a descriptor node. Each descriptor node contains a specification of the conditional probability associated with the node given its set of parent compound nodes. This specification incorporates the effect of any weighting scheme associated with the descriptors node.

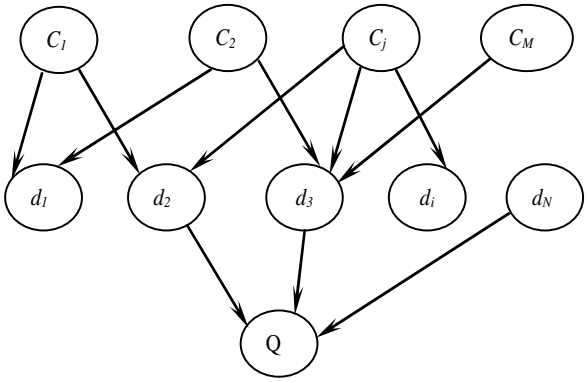

FIGURE 1: Similarity inference network model.

4.2 Query Network

The query network is an "inverted" DAG with a single leaf that corresponds to a target molecule and multiple roots that correspond to the descriptors that express the target. If there is only one query molecule, the target molecule node and query molecule node coincide. In addition, the query network is intended to allow us to combine several query molecules to form a single query molecule. The roots of the query network are query descriptors; they correspond to the descriptors used to express the target molecule. A single query descriptor node has a single compound descriptor node as parent. Each query descriptor node contains a specification of its dependence on a single parent compound descriptor node. The query descriptor nodes define the mapping between the descriptor layer used to represent the compound collection and the descriptor layer used to describe target molecule. In our model, the relation between query and compound descriptors is 1:1 and completely depends. Thus, in order to simplify and reduce our model, the query descriptors are the same as the compound descriptors. The attachment of the query descriptors nodes to the compound network has no effect on the basic structure of the compound network. None of the existing links needs change and none of the conditional probability specifications stored in the nodes are modified.

To produce a ranking of the compounds in the collection with respect to a given target molecule *T*, we compute the probability that this target molecule is satisfied given that compound c_j has been observed, $P(T|c_j)$. This is referred to as instantiating c_j and corresponds to attaching evidence to the network, by stating that $c_j = true$, whereas the rest of the compound nodes are set to false. When the probability $P(T|c_j)$ is computed, this evidence is removed and a new compound c_j, $i \neq j$, is instantiated. By repeating this computation for the rest of the compounds in the collection, the ranking is produced.

The similarity inference network is intended to capture all of the significant probabilistic dependencies among the random variables represented by nodes in the compound and query networks. If these dependencies are characterised correctly, then the results provided are good estimates of the probability this target molecule is met. Given the prior probabilities associated with the compounds (roots) and the conditional probabilities associated with the interior nodes (descriptor nodes), we can compute the posterior probability associated with each node in the network. Further, if the value of any variable represented in the network becomes known we can use the network to recompute the probabilities associated with all remaining nodes based on this "evidence". The query network is first built and attached to the compound network, and then the belief associated with each node in the query network computed. All compounds are equally likely (or unlikely).

4.3 Probabilities Estimation

For any of the non-root nodes ***A*** of the network, the dependency on its set of parent nodes $\{P_1, P_2,\ldots,P_n\}$, quantified by the conditional probability $P(\mathbf{A}|P_1,P_2,..,P_n)$, must be estimated and encoded. Link matrices are used to encode the probability value assigned to a node ***A*** given any combination of values of its parent nodes. However, all the random variables (d_i, q, T), represented by the non-root nodes in the network, are binary and therefore, when a node has n parents, the link matrix associated with it is of size 2×2^n.

Canonical link matrix forms allow us to compute for ***A*** any value $L_{\mathbf{A}}[i, j]$ of its link matrix $L_{\mathbf{A}}$, where $i \in \{0,1\}$ and $0 \leq j \geq 2^n$, will be used [36, 40]. The row number {0,1} of the link matrix corresponds to the value assigned to the node ***A***, whereas the binary representation of the column number is used so that the highest order bit reflects the value of the first parent, the second highest order bit the value of the second parent and so on. The weighted-sum canonical link matrix form [36] allows us to assign a weight to the child node ***A***, which is, in essence, the maximum belief that can be associated with that node. Furthermore, weights are also assigned to its parents, reflecting their influence on the child node. Consequently, our belief in the node is determined by the parents that are true. For instance if node ***A*** has two nodes as parent P_1, P_2 and that the weight assigned to them w_1, w_2 respectively and w_A is weight for node ***A***, now suppose $P(P_1=true)=p_1$ and $P(P_1=true)=p_2$, then the link matrix L_A is as follows:

$$L_A = \begin{bmatrix} 1 & 1-\frac{w_2 w_A}{w_1+w_2} & 1-\frac{w_1 w_A}{w_1+w_2} & 1-\frac{(w_1+w_2)w_A}{w_1+w_2} \\ 0 & \frac{w_2 w_A}{w_1+w_2} & \frac{w_1 w_A}{w_1+w_2} & \frac{(w_1+w_2)w_A}{w_1+w_2} \end{bmatrix} \tag{2}$$

The evaluation for this link matrix is as following:

$$P(A = true) = \frac{(w_1 p_1 + w_2 p_2) w_A}{w_1 + w_2} \tag{3}$$

$$P(A = false) = 1 - \frac{(w_1 p_1 + w_2 p_2) w_A}{w_1 + w_2} \tag{4}$$

In the more general and complicated case of the node ***A*** having n parents, the link matrix at Eq. 2 cannot be evaluated because become NP hard, therefore the derived link matrix can be evaluated using the following closed form expression:

$$bel(A) = \frac{w_A \sum_{i=1}^{n} w_i p_i}{\sum_{i=1}^{n} w_i} \tag{5}$$

For our similarity inference network model, estimates for the *(d_j, q, T)* random variables that characterise the following three dependencies are provided

- The dependence of the descriptor nodes upon the compound nodes which containing them
- The dependence of the query molecule nodes upon the descriptor nodes which containing them.
- The dependence of the target molecule upon the different query node.

In case one query molecule node is used in the model, then the target molecule node coincide with query molecule node. Therefore, we only need to estimate the first two probabilities. The only roots in Fig. 1 are the compound nodes, therefore the prior probability associated with these nodes is set to *1/(collection size)*. Compound and query descriptor nodes are viewed as identical under the assumption that the user knows the set of compound descriptors and can formulate queries using the compound descriptors directly.

To estimate the probability that a descriptor node is good for discriminating a chemical compound's structure, a weighting function can be incorporated in the weighted-sum link matrix. We will use the weighting schemes mentioned in section 2.2 above and difference between values of descriptors nodes for compound and query as weighting function. For instance, molecular descriptors such as topological indices values and bit frequency of fingerprints can be used for weighting function. For normalized topological indices descriptor, this estimate is given by:

$$P(d_i \left| c_j = true \right.) = \alpha + (1-\alpha) \times (1 - \left| d_i - d_i^{'} \right|^2) \tag{6}$$

where $\boldsymbol{\alpha}$ is a constant and experiments using the inference network show that the best value for $\boldsymbol{\alpha}$ is 0.4 [36, 40], d_i is the value of compound descriptor and d_j' is the value of query descriptor. For bit string molecular descriptors, the molecule size (*mz*) and inverse fragment frequency (*iff*) as weighting functions. This estimate is given by:

$$P(d_i \left| c_j = true \right.) = \alpha + (1-\alpha) \times \frac{k_{jq}}{mz_j} \times iff_i \tag{7}$$

For both descriptors,

$$P(d_i \left| all \quad parent \quad false \right.) = 0 \tag{8}$$

Where k_{jq} is the no of common bits between *q* and c_j, mz_j is the size of compound c_j and iff_i is the inverse fragment frequency of fragment *i* in the compound collection.

The target molecule can be expressed as a small number of queries. These can be combined using a weighted-sum link matrix in Eq. 3 with weights adjusted to reflect any user judgments about the importance or completeness of the individual queries. We only have one query node, so the w_A in probability function in Eq. 5 will omit and w_i is set to 1 that's for topological indices and incorporated with weighting function given below for bit strings

$$bel\ (Q) = \frac{\sum_{i=1}^{n} (\frac{k_{jq}}{mz_q} \times iff_i \times p_i)}{\sum_{i=1}^{n} (\frac{k_{jq}}{mz_q} \times iff_i)} \tag{9}$$

where k_{jq} is same as in Eq. 7, mz_q is the size of query *q* and iff_i is the inverse fragment frequency of fragment *i* in the compound collection. The k_{jq} factor is normalizing to the range [0, 1] by

dividing k_{jq} by the maximum possible k_{jq} value (mz_j and mz_q are the maximum values of k_{jq} in Eq. 7 and Eq. 9 respectively). The inverse fragment frequency is given by

$$iff = \log(\frac{collection\ size}{fragment\ \ frequency}) \tag{10}$$

We will normalize *iff* to the range [0, 1] by dividing *iff* by the maximum possible *iff* value in the collection (the *iff* score for a fragment that's occurs once).

$$iff = \frac{\log(\frac{collection\ size}{fragment\ \ frequency})}{\log(collection\ size)} \tag{11}$$

5. EXPERIMENTAL DESIGN

In this study a subset of the MDDR database comprised of around 15 biologically active groups of compounds have been used. Most of the activities chosen are highly diverse whereas the first four categories can be regarded as the most heterogeneous as compared to the rest of the compounds. The experiments were conducted using a collection of 1360 compounds from the MDL's Drug Data Report (MDDR) database [44]. For the first experiment developed to test our similarity inference model with 2D fingerprint descriptors. We used bit string descriptors from Barnard Chemical Inc (BCI) fingerprint generation software based on BCI dictionaries bci1052 [41] for 1052 bit-strings. Unfortunately this type of fingerprint only represents the fragment presence without frequency counts. Therefore, fragment frequency for any fragment in the compound is set to 1. We used 9 targets molecules as queries for each of the 7 activity groups. The main groups, their subgroups and their aggregate activity are summarized in Table 2

S.No	Activity	No. Molecules
1	Interacting on 5HT receptor	
	5HT Antagonists	48
	5HT1 agonists	66
	5HT1C agonists	57
	5HT1D agonists	100
2	Antidepressants	
	Mao A inhibitors	84
	Mao B inhibitors	148
3	Antiparkinsonians	
	Dopamine (D1) agonists	31
	Dopamine (D2) agonists	103
4	Antiallergic/antiasthmatic	
	Adenosine A3 antagonists	73
	Leukotine B4 antagonists	150
5	Agents for Heart Failure	
	Phosphodiesterase inhibitors	100
6	AntiArrythmics	
	Potassium channel blockers	100
	Calcium channel blockers	100
7	Antihypertensives	
	ACE inhibitors	100
	Adrenergic (alpha 2) blockers	100
Total molecules		1360

TABLE 2: Groups and activities of the dataset.

For the second experiment developed to test our similarity inference model with topological indices, we generated around 100 topological indices using the Dragon software [45], out of which only 10 have been selected, accounting for around 98% of the variance in the dataset. A list of the 10 topological indices selected is shown in Table 3. Results were compared with the industry standard Tanimoto measure [46].

TI	Description
Gnar	Narumi geometric topological index
Xt	Total structure connectivity index
Dz	Pogliani index
SMTI	Schultz Molecular Topological Index
PW3	path/walk 3 – Randic shape index
PW4	path/walk 4 – Randic shape index
PW5	path/walk 5 – Randic shape index
PJI2	2D Petitjean shape index
CSI	eccentric connectivity index
D/Dr03	distance/detour ring index of order

TABLE 3: Selected Topological Indices.

6. RESULT AND DISCUSSION

Our similarity inference approach and industry standard Tanimoto measures conducted on the same database and queries. Same evaluation method used for both. Result from the first experiment is shown in Fig. 2, which shows the average number of similarly active compounds to the target structures among the top 5% compounds retrieved. We found that our approach was surpasses the industry standard Tanimoto measure in Antidepressants, Antiallergic/antiasthmatic, AntiArrythmics and Antihypertensives activity groups tested. In Interacting on 5HT receptor, Antiparkinsonians and Agents for Heart Failure activity groups our approach was found inferior to the industry standard Tanimoto measures.

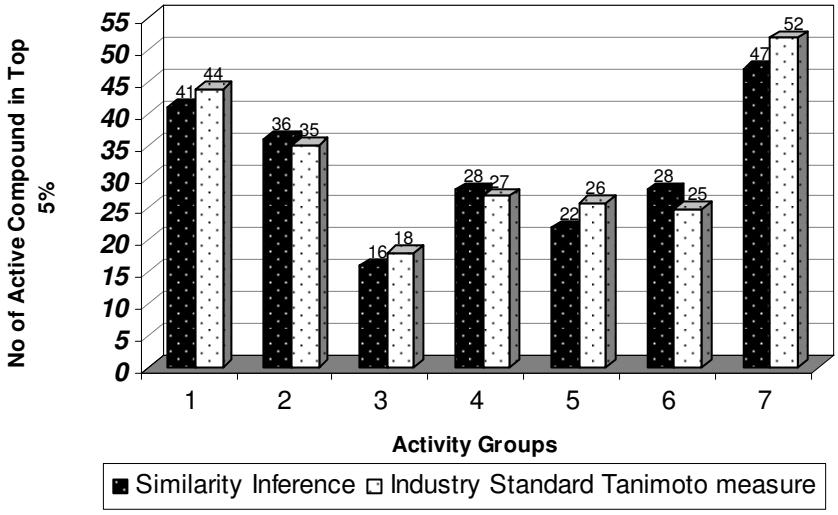

FIGURE 2: Performance of Similarity Inference Network Compared to Performance of Industry Standard Tanimoto Measure using BCI 2D bit string.

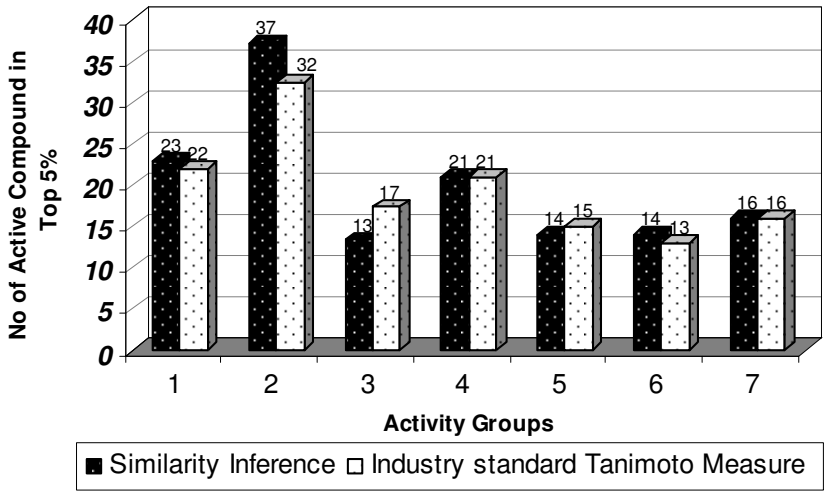

FIGURE 3: Performance of Similarity Inference Network Compared to Performance of Industry Standard Tanimoto Measure using Topological Indices.

Fig. 3 shows result from the second experiment. We found that our approach was surpasses the industry standard Tanimoto measures in Interacting on 5HT receptor, Antidepressants and AntiArrythmics activity groups tested. In Antiparkinsonians and Agents for Heart Failure activity groups our approach was found inferior to the industry standard Tanimoto measures. In Antiallergic/antiasthmatic and Antihypertensives activity groups, we found that both of the approaches perform similarly.

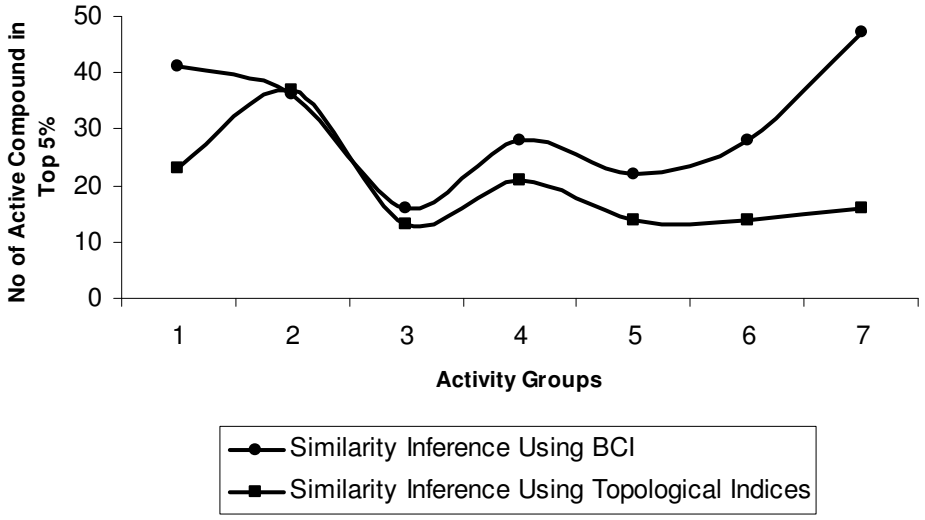

FIGURE 4: Performance of Similarity Inference Network Using BCI Compared to Performance of Similarity Inference Network Using Topological Indices.

Fig. 4 shows the average number of similarly active compounds to the target structures among the top 5% compounds retrieved. We found that our approach with bit-string descriptors from BCI was performing better than when used with topological indices.

There are two distinct factors influence on the result produced by our approach. For 2D bit-string, the no of common bits between compound and query (k_{jq}), and the inverse fragment frequency

(*iff*) of the fragment in the collection. For topological indices, the distance between descriptors values of query and compound, and weight of query descriptor nodes (w_i).

These factors constitute the weighting functions used in our approach. These weighting function are intended to increase the influence of fragments and descriptors that are believed to be important on quantifying the similarity. The basic ideas are that

- Many bits share by compound and query lead to increase the similarity score of this compound
- Those fragments that occurs infrequently in the collection are more likely to be important than frequent fragments and increase the similarity score of this compound.
- Slight distance between descriptor values lead to increase the similarity score of this compound

7. CONSLUSION & FUTURE WORK

We have notice that the existing molecular similarity searching methods suffer from problems like instability, unstandardize and poor results. The instability appears because no judgment can be made about which best coefficients can be used for all biological activities. The similarity method can start with little information, and as a general rule, the molecular similarity concept is most often applied when knowledge of the system is sparse. This one of the advantage of molecular similarity method but at the same time is disadvantage to these methods.

In this work we are proposing Bayesian inference networks for molecular similarity searching. We have developed a novel approach for molecular similarity based on Bayesian inference networks, which can resolve these problems. Our approach can comprise belief, weights and any other evidences in the problem of molecular similarity. Overall results show the networks performed slightly improvement than industry standard Tanimoto measures. We foresee that the result can be much better when a better weighting function can be devised. Currently, we are working on developing new weighting functions which include the frequency of each fragment in compound to use in our similarity inference network.

8. REFERENCES

1. M. A. Johnson and G. M. Maggiora. "*Concepts and Application of Molecular Similarity*", John Wiley & Sons, New York (1990)

2. R. P. Sheridan and S. K. Kearsley. "*Why do we need so many chemical similarity search methods?*". Drug Discov. Today, 7, 903–911, 2002

3. M. A. Miller. "*Chemical Database Techniques in Drug Discovery*". Nature Reviews Drug Discov.,1, pp. 220-227, 2002

4. P. Willett. "*Chemoinformatics: an application domain for information retrieval techniques*". In Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval SIGIR '04. ACM, New York, NY, 393-393, 2004

5. P. Willett, J. M. Barnard and G. M. Downs. "*Chemical similarity searching"*. Journal of Chemical Information and Computer Sciences, 38:983-996, 1998

6. P. M. Dean. "*Molecular Similarity In Drug Design*". Blackie Academic & Professional, London, 1995

7. R. E. Carhart, D. H. Smith and R. Venkataraghavan. "*Atom pairs as molecular features in structure-activity studies: definitions and applications*". Journal of Chemical Information and Computer Science, 25:64-73, 1985

8. D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark and L. E. Weinberger. "*Neighborhood behavior: as useful concept for validation of molecular diversity descriptors*". Journal of Medical Chemistry, 39:3060-3069, 1996

9. P. Willett, V. Winterman and D. Bawden. "*Implementation of nearest neighbour searching in an online chemical structure search system*". Journal of Chemical Information and Computer Science, 26:36-41, 1986

10. T. R. Hagadone. "*Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases*". Journal of Chemical Information and Computer Science. 32:515-521, 1992

11. W. Fisanick, K. P. Cross and A. Rusinko. "*Similarity searching on CAS Registry Substances. 1. Global molecular property and generic atom triangle geometric searching".* Journal of Chemical Information and Computer Sciences, 32:664-674, 1992

12. D. Bawden. "*Molecular dissimilarity in chemical information systems*". In Chemical Structures Vol. 2: The International Language of Chemistry (W. A. Warr, ed.), Springer-Verlag, Hiedelberg, pp. 383-388, 1993

13. M. S. Lajiness. "*Dissimilarity-based compound selection techniques*". Perspectives in Drug Discovery and Design, 7/8:65-84, 1997

14. E. J. Martin, J. M. Blaney, M. A. Siani, D. C. Spellmeyer, A. K. Wong and W. H. Moos. "*Measuring diversity: Experimental design of combinatorial libraries for drug discovery*. Journal of Medicinal Chemistry, 38:1431-1436, 1995

15. J. D. Holliday and P. Willett. "*Definitions of "dissimilarity" for dissimilarity-based compound selection*". Journal of Biomolecular Screening, 1:145-151, 1996

16. V. J. Gillet, P. Willett and J. Bradshaw. "*The effectiveness of reactant pools for generating structurally diverse combinatorial libraries*". Journal of Chemical Information and Computer Science. 37:731-740, 1997

17. P. Willett. "*Similarity-based virtual screening using 2D fingerprints*". Drug Discov. Today, 1046-1053, 2006

18. 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, 23:93-102, 1983

19. Barnard Chemical Information Ltd., "*Barnard Chemical Information Fingerprint Software Documentation*". MAKEBITS version 3.3, p. 1-5, 1997

20. Barnard Chemical Information Ltd., "*Barnard Chemical Information Fingerprint Software Documentation*". MAKEFRAG version 3.3, Sheffield, p. 1, 1997

21. J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse. "*MDL keys revisited*". 2nd Joint Sheffield Conference on Chemoinformatics: Computational Tools For Lead Discovery, University of Sheffield, Sheffield, 2001

22. J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse. "*Reoptimization of MDL keys for use in drug discovery*". Journal of Chemical Information and Computer Science, 42:1273-1280, 2002

23. Tripos Inc. UNITY Reference Guide version 4.1. Tripos, St. Louis, Missouri, 1999

24. C. A James, D. Weininger and J. Delany. "*Daylight Theory Manual*" http://www.daylight.com/dayhtml/doc/theory/index.html

25. G. M. Downs and P. Willett. "*Similarity searching in databases of chemical structures*". In: K. B. Lipkowitz and D. B. Boyd (Eds.), Reviews in Computational Chemistry, VCH Publishers, New York, Vol. 7, pp. 1-66, 1996

26. L. Hodes. "*Clustering a large number of compounds. 1. Establishing the method on an initial sample*". Journal of Chemical Information and Computer Science, 29:66-71, 1989

27. P. Willett and V. Winterman. "*A comparison of some measures of intermolecular structural similarity"*. Quantitative Structure-Activity Relationships, 5, 18–25, 1986

28. P. Willett. "*Algorithms for calculation of similarity in chemical structure databases*". In Concepts and Application of Molecular Similarity, M. A. Johnson and G. M. Maggiora, Eds., John Wiley and Sons, New York. pp. 43-61, 1990

29. P. H. A. Sneath and R. R. Sokal. "*Numerical Taxanomy*". Freeman, San Francisco, 1973

30. P. Willett. "*Similarity And Clustering In Chemical Information Systems*", Research Studies Press, Letchworth, (1987)

31. D. Ellis, J. Furner-Hines and P. Willett. "*Measuring the degree of similarity between objects in text retrieval systems*". Perspective in Information Management. 3:128-149, 1993

32. G. W. Adamson and J. A. Bush. "*A method for the automatic classification of chemical structures*". Information Storage and Retrieval, 9:561-568,1973

33. J. Pearl. "*Probabilistic reasoning in intelligent systems: Networks of plausible inference*", Morgan Kaufmann Publishers, (1988)

34. G. Salton and M. J. McGill. "*Introduction to Modern Information Retrieval*", McGraw-Hill, NewYork, (1983)

35. C. J. Van Rijsbergen. "*Information Retrieval*", 2nd ed., University of Glasgow, 87-110 (1979)

36. H. Turtle. "*Inference Networks for Document Retrieval*". PhD Thesis, University of Massachusetts, 1990

37. H. Turtle and W. Croft. "*A comparison of text retrieval models*". Comput. Journal, 35, 279-290, 1992

38. B. A. N. Ribeiro and R. Muntz. "*A belief network model for IR*". In: Proceedings of the 19th ACM SIGIR Conference, pp. 253–260,1996

39. S. K. M. Wong and Y. Y Yao. "*On modeling information retrieval with probabilistic inference*". ACM Transactions on Information Systems, Vol. 13, No. 1, pp. 38-68, 1995

40. H. Turtle and W. Croft. "*Evaluation of an inference network-based retrieval model*". ACM Transactions on Information Systems, 9:187-222, 1991

41. Barnard Chemical Information Ltd., "*Barnard Chemical Information Fingerprint*". http://www.bci.gb.com

42. J. Gasteiger and T. Engel. *"Chemoinformatics*", VCH-Wiley, New York, Vol. 1, pp. 3-5 (2003)

43. L. M. De Campos, J. M. Fernández and J. F. Huete. "*The BNR model: foundations and performance of a Bayesian network- based retrieval model*". Int. J. Approx. Reasoning, 3, pp. 265–285, 2003

44. Molecular Design Ltd., MDDR "*MDL Drug Data Report Database".* http://www.mdli.com

45. Melano Chemoinformatics. "*Dragon software*". http://www.talete.mi.it

46. N. Salim, J. Holliday and P. Willet. "*Combination of fingerprint-based similarity coefficients using data fusion*". J. Chem. Inf. Comput. Sci., 43, pp. 435-442, 2003

47. P.A. Bath, C. A. Morris and P. Willett. "*Effect of standardisation of fragment-based measures of structural similarity*". Journal of Chemometrics, 7, pp. 543, 1993.

48. N. Daut, R. Mohemad and N. Salim. "*Finding Best Coefficients for Similarity Searching Using Neural Network Algorithm*". International Conference in Artificial Intelligence in Engineering & Technology (ICAIET), 2006.

49. Downs, G.M., Poirrette, A.R., Walsh, P. and Willett, P. "*Evaluation of similarity searching methods using activity and toxicity data*". In Chemical Structures Vol. 2: The International Language of Chemistry (W. A. Warr, ed), Springer Verlag, Heidelberg, pp. 409-421, 1993

50. N. Salim and W. W. P. Godfrey. "*Effectiveness of Probability Models for Compound Similarity Searching*". Journal of Advancing Information Management Studies, 2(1): pp. 56-74, 2005.