

Combination of Structure and Physicochemical Properties Measures Using Data Fusion for Chemical Similarity Search

Jianmei Fan

University of Illinois at Urbana-Champaign

jmfan@illinois.edu

ABSTRACT

This is a project report for UIUC CS512 Data Mining class. Current similarity search based on structure graph such as fingerprint, Maccs keys, cyclic patterns and tree, substructure, and closure tree are reviewed. This report also introduced chemical similarity search based on different types of physicochemical descriptors. Data fusion technique is used to combine closure tree and some physicochemical descriptors. Experiment shows this method can be used. However, a more comprehensive and better designed experiment is needed.

Keywords

Chemical similarity, closure-tree, physicochemical property descriptors, data fusion

1. INTRODUCTION

Today more and more pharmaceutical companies realize the importance of information technology in the process of drug discovery. They use all kinds of bio- and cheminformatics tools to improve quantity and quality of drug candidates, and to speed up the process [1]. Typically after high throughput screening (HTS), one or several hit compounds show some interesting activity. Then chemists try to identify other molecules that have better activity and physicochemical property based on these lead structures. Sometimes they use chemical similarity search to find other interesting compounds in the database and compare their structure and activity relationships (SAR). Then they synthesize compounds based on SAR theory they developed and test them in the biology assay. There are two major different ways to calculate similarity. One is based on molecule graph, such as paths, fragments, and substructure etc. One is based on physicochemical properties, such as binding property – hydrophobicity, hydrogen bond donor and acceptor potentials, and molecular electrostatic potentials etc. [2, 3]. Each of them has their own advantages and pitfalls. Using physicochemical property descriptor alone will give more diversified structures of molecule, then it will be hard for chemist to compare these structures and draw SAR conclusion. Meanwhile using molecule graph alone will only give similar structures but it may overlook some very important binding effect. “It is as if we have a set of imperfect windows through which to view Nature” [4]. People have used data fusion technique [5] to combine different molecular similarity measures to improve the performance of similarity search in chemical databases. Kearsley et al. [6] combine different physicochemical property descriptors. Ginn et al. [5] combine 2D fingerprint measure with physical properties. Since fingerprints are typically path based approach, global structure information is lost. However, Closure-tree captures the entire structure of constituent graphs very well [7]. In this

project, we propose the combination of closure-tree based structure similarity with physicochemical property to improve similarity search.

2. Similarity Search Based on Structure Graph

2.1 Fingerprints

Wale et. al [8] mentioned several types of structure descriptors. Most popular commercial system use fingerprints (fp-n) as their descriptor. In this method, fingerprints are generated by enumerating paths and hashing them into a fixed bit-string. Similarity is defined by how many these bit-fixed strings two compounds are in common vs. total number of these strings. This method is efficient due to its computation inexpensive. Many variants of these fingerprints exist. Some use predefined structural fragments in conjunction with the fingerprints (Unity fingerprints), some count the number of times a bit position is set (hologram), and others generate fingerprints based on the extended connectivity of each atom (ECFP), etc. But in all, they represent a very large number of paths into a compact form and then compare this form.

2.2 Maccs Keys

Molecular Design Limited (MDL) has used structural keys which are based on general molecule properties such as atom, bond, and aromatic etc. These keysets were originally constructed and optimized for substructure searching. Durant, J. el. [9] reoptimized them for use in similarity search. MDL ISIS calculates the similarity value by comparing the structural keys in the query to the structural keys in the target molecule. Keys are weighted differently. ISIS uses predefined keys to define the structures stored in the database. When a similarity search is conducted, ISIS compares the keys registered by the query with the keys registered for each compound in the database. If the calculated similarity is equal to or greater than the similarity value entered, the molecule is retrieved. There are two common MDL keysets: one containing 960 keybits and the other containing a subset of 166 keybits.

2.3 Cyclic Patterns and Trees

Horovath et al [10] use a set of cycles and trees to represent compounds. They first identify biconnected blocks of a chemical graph and generate simple cycles for the blocks. Then the blocks are deleted. The resulting left over graph is a set of trees. Each cycle and tree is used as structure descriptor to do chemical graph search.

2.4 Frequent Substructure

Yan et al [11] proposed Grafil to do substructure similarity search. They first use data mining method to find frequent small substructures, and then use them as features (descriptors) for the compounds in the database. The similarity search can be performed in the following 4 steps: 1. Index construction by selecting small substructures as features in the graph database. 2. Determine the indexed features belonging to the query graph. Calculate the upper bound of features can be missed. 3. Calculate the difference in the number of features between each graph in the database and query. If the number is greater than the upper bound, discard graph. Otherwise it belongs to the candidate answer set. 4. Calculate substructure similarity using the existing algorithms and prune the false positive in the candidate set. Comparing with path based approach, substructures based approach maintain more chemical structure information.

2.5 Graph Closure

Compare with other methods, graph closures capture the entire structure of constituent graph, which means keep more information about a graph. Closure tree organizes graph hierarchically where each node summarizes its descendants by a graph closure. They define Graph similarity based on edit distance. The edit distance between two graphs is the cost of transforming one to the other, which includes both vertex distance and the edge distance respectively. Since computing exact similarity is expensive, they compute approximate graph distance using heuristic graph mapping method.

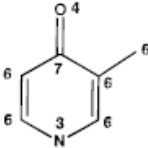
The properties of a chemical are implicit in its molecular structure. For example, N, O, and C etc. represent different electron property atom in the molecule, while single, double, and triple bond represent neighbors of atom distance and angle. Although in this report, I separate similarity chemical search based on graph from physiochemical property, this is just one point of view. In reality, these two are not absolutely separated. For example, similarity search based on Maccs keys can be also thought as similarity search based on molecule atom property.

3. Chemical Similarity Using Physiochemical Property Descriptors

The structural similarity has been heavily exploited. However current research shows that it does not always imply similarity in activity. Similar compounds can have very different physiochemical properties (see figure 2) and the application of similarity by structure has to be justified in every specific case.

Based on the chemical physiochemical properties, there are many different types of descriptors. The atom pair (ap) [12] and topological torsion (tt) [13] are two very early ones that are used for chemical similarity search. For ap and tt, atoms are distinguished by the element, number of non-hydrogen neighbors, and number of pi electrons. Generally in these descriptors based similarity searching, chemical compounds are parsed into descriptors and these descriptors are stored in a descriptor database. At search time, query is also parsed into descriptors. The similarity between the probe and each database entry is calculated by comparing the list of descriptors in the probe with the list for the entry. Since it is computationally inexpensive to compare lists of descriptors, search can be done quickly. But the information of which descriptor corresponds to which molecular

feature is lost, and one can not obtain an equivalence between features in the query and a database entry. Sheridan et al [14] have proposed 3D variants of the atom pair, referred as geometric atom pairs and geometric binding property pairs. In geometric atom pairs, the atom types are defined as for 2D standard atom pairs, but the distance between them is the through-space distance rather than the through-bond distance. In geometric binding property pairs, the distance is through-space distance and the atom type is generalized to one of seven binding classes (cation, anion, H-bond donor, H-bond acceptor, polar, hydrophobic, and other). Brown et al. [15] have described two descriptors based on potential pharmacophore points (PPPs). These points includes H-bond donor, H-bond acceptor, positively charged, negatively charged, and hydrophobic. The two descriptors are PPP-pairs, which is similar to atom pairs, and PPP-triangles, which is triplets of PPPs and their associated distances. Fisanick et al [16] have proposed eight 3D descriptors. They are atom pair distance, three-bonded atoms angle, three atoms and one bond vector, four-bonded atoms, four atoms and two bond vectors, atom triangle, atom triangle three-slot, and atom triangle five slot. These descriptors provide an effective similarity search for 3D molecule size and shape. They also investigated the use of calculated molecular properties, such as ClogP, molar refractivity, ionization potential, HOMO, and LUMO. The values can be used directly as descriptors.



unique ap	frequency
1 CX3-(2)-CX2	3
2 CX3-(1)-CX2	2
3 OX1-(3)-CX2	2
4 NX2-(1)-CX2	2
5 CX3-(3)-CX2	1
6 CX2-(2)-CX2	1
7 CX2-(1)-CX2	1
8 CX2-(3)-CX2	1
9 CX3-(1)-CX3	1
10 NX2-(3)-CX3	1
11 NX2-(2)-CX3	1
12 OX1-(2)-CX3	1
13 OX1-(4)-NX2	1
14 NX2-(2)-CX2	1
15 OX1-(1)-CX3	1
16 OX1-(2)-CX2	1
17 CX3-(1)-CX1	1
18 CX2-(4)-CX1	1
19 CX3-(2)-CX1	1
20 OX1-(3)-CX1	1
21 CX2-(3)-CX1	1
22 CX2-(2)-CX1	1
23 NX2-(3)-CX1	1
28 total	

unique bp	frequency
1 4-(3)-6	3
2 6-(1)-6	3
3 6-(2)-6	3
4 6-(3)-6	3
5 6-(2)-7	3
6 3-(1)-6	2
7 3-(2)-6	2
8 4-(2)-6	2
9 6-(1)-7	2
10 3-(4)-4	1
11 3-(3)-6	1
12 3-(3)-7	1
13 4-(1)-7	1
14 6-(4)-6	1
28 total	

unique tt	frequency
1 OX1-CX3-CX3-CX2	1
2 OX1-CX3-CX3-CX1	1
3 CX3-CX3-CX2-CX2	1
4 CX2-CX3-CX3-CX2	1
5 CX2-CX3-CX3-CX1	1
6 NX2-CX2-CX3-CX3	1
7 OX1-CX3-CX2-CX2	1
8 CX3-CX2-NX2-CX2	1
9 NX2-CX2-CX2-CX3	1
10 NX2-CX2-CX3-CX1	1
11 CX2-NX2-CX2-CX2	1
11 total	

unique bt	frequency
1 4-7-6-6	3
2 6-6-7-6	3
3 3-6-6-7	2
4 6-3-6-6	2
5 3-6-6-6	1
11 total	

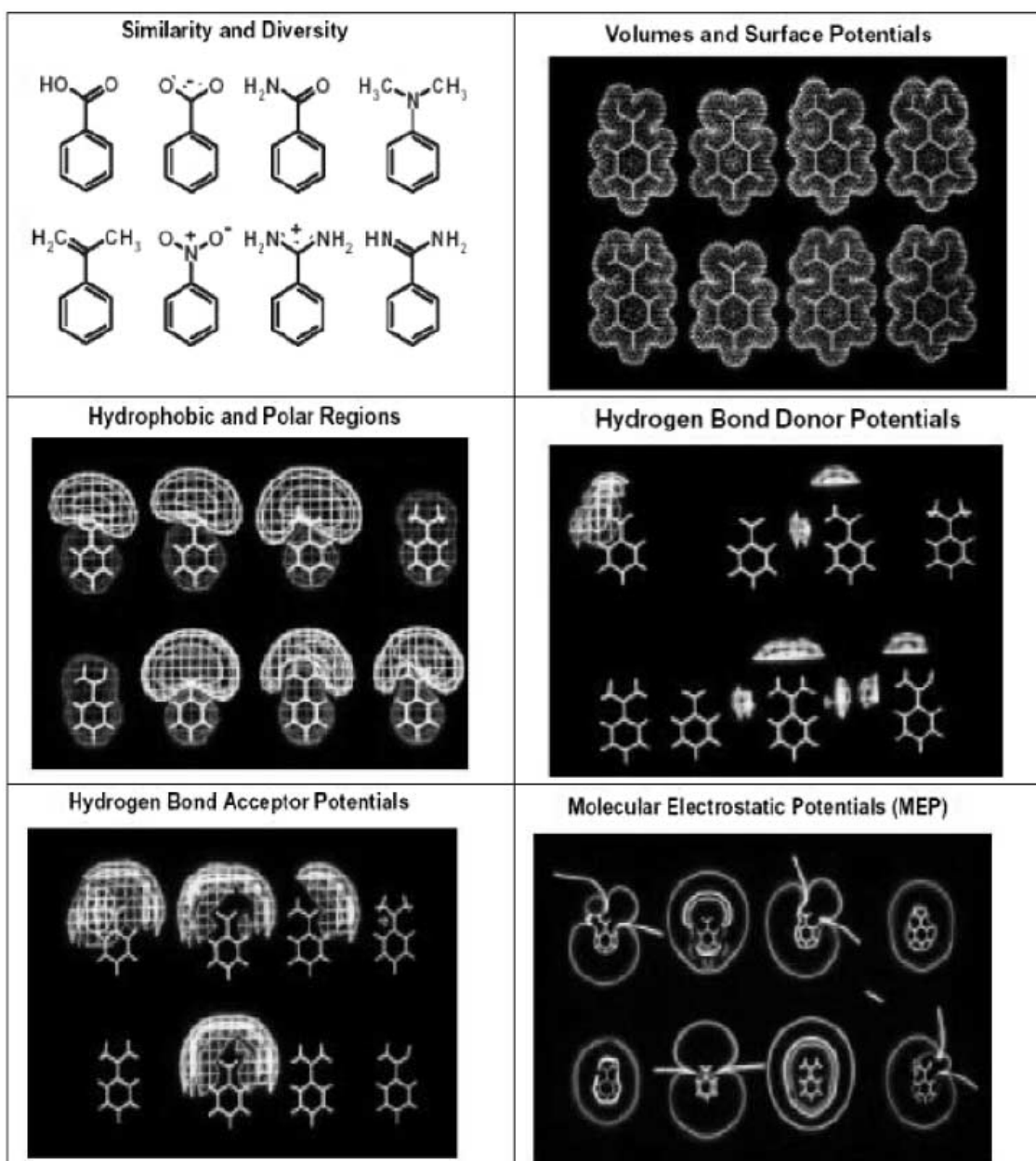
Figure 1. Sample molecule parsed into descriptors.

Kearsley et al. [6] have extended the ap and tt descriptors. The new descriptors are based on binding property class, atomic log P contribution, and partial atomic charges. In binding atomic pairs (bp) and binding property torsions (bt), the atom are assigned to one of seven binding property classes for non-hydrogen atoms: cations, anions, neutral hydrogen bond donors, neutral hydrogen bond acceptors, polar atoms, hydrophobic atoms, and other. In figure 1, the original atom pairs (ap) and topological torsions (tt) are shown as well as binding property pairs (bp) and binding property torsions (bt). In ap and tt, "OX1." is an oxygen with one neighbor and one pi electron. The ap "NX2-(4)-OX1." is an sp³ nitrogen with two neighbors four bonds away from a carbonyl oxygen. The tt "NX2-CX2.-CX2.-CX3." represents four consecutively bonded atoms. The bp and bt descriptors are constructed analogously to ap and tt. 3 is neutral H-bond donor, 4 is neutral hydrogen bond acceptor, 6 is hydrophobic, and 7 is other. For instance, 4-(3)-6 is atom 4 and 6 with 3 bonds between them. 4-7-6-6 has frequency 3.

There are infinite varieties of potential descriptors. Many other similar descriptors have been investigated by different group of people. In all, we need to choose those most appropriate ones to a given application.

There is a general conflict in chemical similarity search between effective and efficient. Compare to ap and tt, quantum-mechanical descriptors such as the electron probability density function proposed by Carbo et al [17] capture molecular property very well. In principle, quantum mechanical wave function contains all the information about a given chemical. All the other descriptors can be regarded as direct manifestations of the underlying wave equations that describe the molecule. However it takes too long to calculate. But with the new developments of algorithm and the increase of computer power, this method becomes more and more attractive.

Figure 2. Structure similarity compounds can have very different physiochemical properties.



4. Data Fusion

4.1 Definition

“Data fusion is generally defined as the use of techniques that combine data from multiple sources and gather that information in order to achieve inferences, which will be more efficient and potentially more accurate than if they were achieved by means of a single source. Data fusion is a low-level fusion process. Fusion processes are often categorized as low, intermediate or high, depending on the processing stage at which fusion takes place. Low level fusion, (Data fusion) combines several sources of raw data to produce new raw data. The expectation is that fused data is more informative and synthetic than the original inputs.” from Wikipedia, the free encyclopedia

4.2 Combination of molecular similarity measures using data fusion

Data fusion method has yet not been fully explored in the chemical similarity search area. However, data fusion is not completely novel to chemist. Long time ago, chemists have combined different spectrum such as IR, proton NMR and carbon NMR to identify the structure of compounds. Recently, Kearsley et al [6] uses their in-house system TOPOSIM to combine different physicochemical property descriptors. The idea behind the combination is that since people can not predict how well a descriptor can do to the similarity search, using two or more might increase chance that better results might be obtained. They calculate for each molecular the similarity based on each of the eight descriptors. Then they all sorted from high to low score. Ranks are then assigned for each descriptor. The molecule with highest score is rank 1, the next highest rank 2, etc. They only use rank because absolute scores vary from one descriptor to another. It is hard to combine these similarity measures directly. For example, after rank numbers are generated for ap and tt, they define a new score for each compound as its rank in the ap or tt list, whichever is smaller. Then the compounds are sorted by the new score. Compound with smallest score is rank 1. The resulting compound list is the union of the top scoring compounds for each descriptor.

Ginn et al. [5] combine 2D fingerprint measure with physical properties. They use EVA descriptor, which characterizes a molecule by its fundamental vibrational fingerprint. Comparable search were carried out using the 2D similarity searching in the UNITY chemical information management system, and using data fusion to combine the two individual types of ranking. On average, the fused rankings appeared to be better than the original 2D and EVA rankings. The study provided at least some evidence that data fusion could be used to improve the performance of similarity searching in chemical databases. They use ranks instead of directly fusing two set of data because their distribution quite different and it is unwise to fuse them directly. They used SUM and MAX rule to fuse data, but SUM performs consistently better.

Computer scientists are very good at developing graph algorithm, and using these algorithms for the chemical similarity search. However they tend to ignore the underlying molecule chemistry property. While chemists like to do similarity search based on molecule chemistry property. However, this approach limit them to view a whole molecule as many small fragments, hence didn't catch whole graph of molecule. It will be very useful if we can combine these two approaches together. On one hand, computer scientists will understand more what other properties are under the chemical structure. On the other hand, chemists can use these powerful graph algorithm tools in the real application.

5. Combination Chemical Physical Property with Closure Tree Similarity Search

There are several interesting similarity search based on graph path, substructure, or closure tree. Most popular commercial system use fingerprints which we regard it as path. Substructure and closure tree, especially closure-tree hasn't been combined with physical properties to do chemical similarity search. One of the unique properties of closure tree is that it captures the whole graph of a molecule.

As I mentioned before, using only physicochemical descriptors alone may cause lost of information of which descriptor corresponds to which molecular feature, one can not obtain an absolute equivalence between features in the query and a chemical compound. Closure tree might be able to improve this shortcoming. In He's paper, they use K-NN query to finds K nearest graph. Here is their procedure: A priority queue maintains C-tree nodes and they are visited according to their similarity to the query. Each time the top entry is chosen from the priority queue. They check whether the entry is a node. If the entry is a node, then its children are inserted into the priority queue, otherwise the entry is reported. The procedure stops after k database graphs have been reported. In this report, I choose c-tree m as 20, and K as 11.

There are variety descriptors can be chosen. However, many of them are not available and need to parse molecule structure to get them. Because of time, in this report I just use a few simple descriptors such as logP and logS that can be readily calculated using exiting web tool. (<http://146.107.217.178/lab/alogps/>)

Table 1 shows one example of experiments I did, query compound is NCI 340333. LogP rank is based on $|\text{LogP}-\text{LogPq}| = |\text{LogP}-2.6|$, the molecule with lowest score is rank 1, the next lowest rank 2, and etc. LogS rank is based on $|\text{LogS}-\text{LogSq}| = |\text{LogS}+3.26|$, the molecule with lowest score is rank 1, the next lowest rank 2, and etc. Sum Rank = C-tree Rank + LogP rank + LogS rank. In figure 3, the original closure tree ranking is on the left, rank high to low from top to bottom. The fusion ranking is on the right, rank high to low from top to bottom. The experiment is feasible. However, because of limited data, it is hard to draw conclusion whether this join approach is more effective or not.

Table 1. Sample list of 11 similar NCI compounds

NCI ID	C- treeRank	LogP LogP	LogP rank	LogS LogS	LogS rank	SUM Rank
340333	1	2.6	1	-3.26	1	3
340324	2	1.66	6	-1.64	7	15
340334	3	2.84	3	-3.26	1	7
340332	4	2.32	4	-2.95	2	10
340325	5	1.67	5	-1.65	6	16
340329	6	2.37	2	-2.6	3	11
337243	7	1.34	10	-1.42	8	25
340327	8	1.51	7	-2.08	5	20
340320	9	1.36	9	-1.24	9	27
340341	10	1.39	8	-2.48	4	22
340323	11	1.24	11	-1.18	10	32

6. Discussion and future plan

In this project, data fusion technique is used to combine two different similarity methods. Maybe it is too simple and not the best way to go. It will be interesting to see if we can push these chemical property constraints deeper into the structure search process. For example, in closure tree, they only use edit distance to measure graph similarity, maybe we can add chemical molecule property.

In this project, closure tree is combined with a very few chemistry descriptor. It will be interesting to combine with other descriptors such as binding atomic pairs (bp), binding property torsions (bt), and etc. It will also be interesting to combine substructure with different descriptors. Then these different combinations can be compared.

Form class CS512 we have learned information network analysis. Maybe we can also view all the molecule information as a network. We might even be able to apply link and rank analysis to this field and improve search result.

“The notion of similarity is used mainly in early stages of the development of a particular science, and it may be quantified and explained accurately later as the theory of this science develops...Chemical compounds activity has been traditionally modeled using a variety of topological, physicochemical and electronic descriptor. This has provided the grounds for evaluating similarity between compounds by comparing numerical values of these descriptors. However, the most informative description of a molecule is its quantum mechanical wave function. In principle, it contains all the information about a given chemical. The structure diagram, 3D coordinates and some numerical descriptors can be regarded as direct manifestations of the underlying wave equations that describe the molecule” [2]. For now, we will choose best windows suitable to our need, combine these pieces together, and hope they can be as informative as possible.

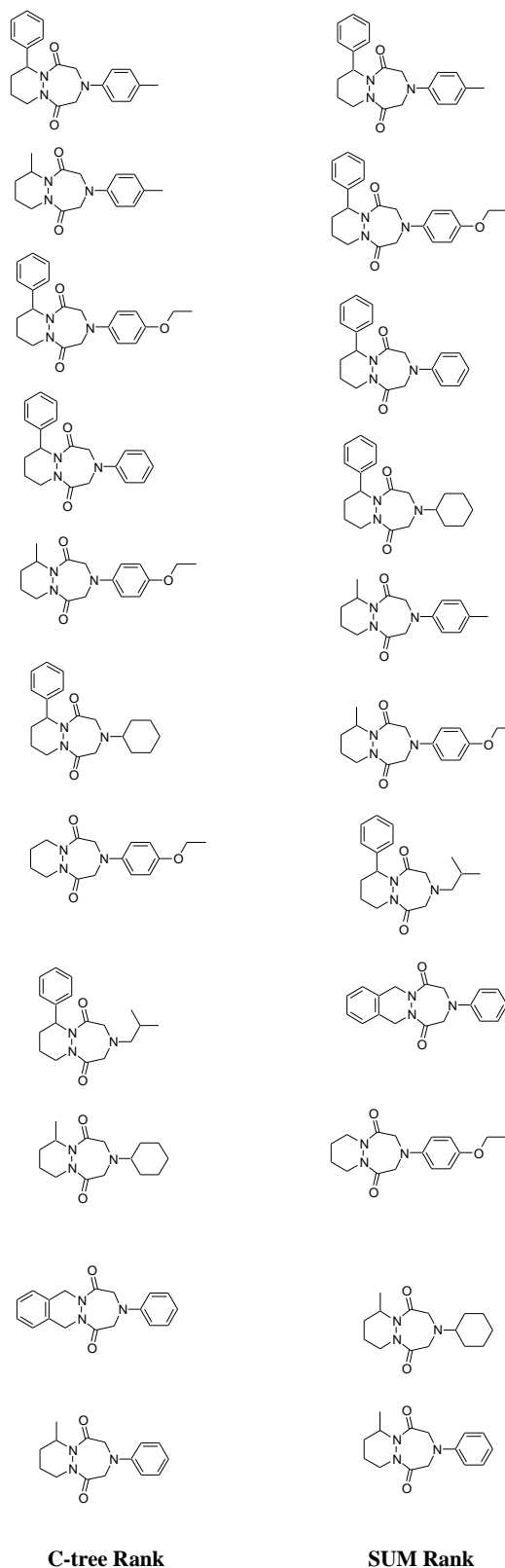


Figure 3. Sample similarity compounds. Rank from 1 to 11 from top to down.

7. ACKNOWLEDGMENTS

Thanks to Huahai He to offer closure-tree executable file.

Thanks to professor Han and TA Jing. Before this class, I was always wondering what is behind those search tools for chemical. I feel I can not use them effieicently. After this class and this project, I even can think about building these tools. I have learned a lot about data mining, especially chemcial structural search.

8. REFERENCES

- [1] Xu, J., and Hagler, A. 2002. Chemoinformatics and Drug Discovery. *Molecules* 2002, 7, 566-600.
- [2] Nikolova, N., and Jaworska, J. 2003. Approaches to Measure Chemical Similarity – a Review. *QSAR Comb. Sci.* 22, 1006-1026.
- [3] Willett, P., Barnard, J. M., and Downs, G. M. 1998. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.*, 1998, 38(6), 983-996.
- [4] Sheridan, R. P., and Kearsley, S. K. 2002. Why do we need so many chemical similarity search methods? *DDT7* (17), 2002, 903-911.
- [5] Ginn, C. M. R., Willett, P. and Bradshaw, J. 2000. Combination of molecular similarity measures using data fusion. *Perspectives in Drug Discovery and Design*, 20 (1), 1-16.
- [6] Kearsley, S. K., Sallamack, S. and Fluder, E. M. 1996. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.*, 1996, 36(1), 118-127.
- [7] He, H., and Singh, A. K. Closure-Tree: An Index Structure for Graph Queries. *ICDE'06* 2006.
- [8] Wale, N., Watson, I. A., Karypis, G. 2006. Comparison of Descriptor Spaces for Chemical Compound Retrieval and Classification. *ICDM'06*, 678-689.
- [9] Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. 2002. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* 2002, 42, 1273-1280.
- [10] T. Horvath, T. Grtner, and S. Wrobel. Cyclic pattern kernels for predictive graph mining. 2004. *SIGKDD*, 158-167.
- [11] Yan, X., Yu, P. S., and Han, J. Substructure similiarity search in graph database. 2005. *SIGMOD'05*
- [12] Carhart, R.E., Smith, D. H., Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: Definition and application. *J. Chem. Inf. Comput. Sci.* 1985, 25, 64-73
- [13] Nilakantan, R., Bauman, N., Dixon, J. S., Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* 1987, 27, 82-85.
- [14] Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Pair Descriptors. *J. Chem. inf. Comput. Sci.* 1996, 36, 128-136
- [15] Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* 1996, 36, 572-584.
- [16] Fisanick, W.; Cross, K. P.; Rusinko, A. Similarity Searching on CAS Registry Substances. 1. Global Molecular Property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci.* 1992, 32, 664-674.
- [17] Carbo, R., Calabuig, B. Quantum similarity measures, molecular cloud description and structure-property relationships. *J. Chem. Inf. Comput. Sci.* 1992, 32, 600-606.