A Method for Substructure Search by Atom-Centered Multilayer Code

Yunde Xiao, Yuanyuan Qiao, Jinpei Zhang, Shaofan Lin,* and Weidong Zhang

Centre Laboratory, Nankai University, Tianjin, 300071, People's Republic of China

Received November 4, 1996⊗

Atom-Centered Multilayer Code (ACMC) is a novel coding algorithm by characterizing the sphere environment of a non-hydrogen atom, called center atom. This algorithm starts from the center atom and then develops to its environment layer by layer automatically without predefining specific fragments, and the number of layers is determined on the structure of the coded compound without limit. Fast searching strategy at the atomic level is obtained by using this code. The coding process and the global structure and substructure searching are discussed with examples.

INTRODUCTION

In recent decades, a great deal of work has been done in the field of structure search since the method of structure search is important and complicated. It is indispensable for any chemical information system. In a structure search the principal task is the substructure search because global structure search is perceived as a specialization of substructure search. When a compound has been synthesized, scientists often have interest in the information on certain parts of the compound, such as functional groups. In chemical information systems, such as systems of computer-aided synthesis design and systems of molecular design for new drugs, substructure search plays a very important role.

The method of a substructure search has very close relations with the method of structure coding. Besides the linear notation developed by Wiswesser,[1] various other structure codes have been proposed. Most of them are based on two-dimensional connection table and topology.[2] The DARC system of Dubois[3] describes substructures contained in the molecule using the idea of parameterization of atoms. The CIDS Chemical Search Keys[4] allows the identification of functional groups with the help of a digital code. The HOSE and HORD[5] proposed by Bremser characterizes spherical environment of single atoms and complete ring systems.

The method of structure coding proposed in this paper is an algorithm of structure coding starting from an atom and then developing to its environment layer by layer automatically without predefining specific fragments, and the number of layers is determined on the structure of compound without limit.

THE CODING ALGORITHM

Substructure is a part of a compound. It may be a functional group or the combination of some functional groups or a fragment, so it might be meaningful or meaningless in chemistry. All compounds conist of atoms by some sequence, on the other hand, atoms can be considered as the smallest structure units of compounds. Therefore an atom can be considered as a substructure. The atom-centered multilayer code for an atom can be generated by the following coding process.

Coding Process. 1. Select a non-hydrogen atom in the compound, as a center atom. It consistitutes the first layer, and its code is the given atomic weight of the atom. The given atomic weights of elements in the periodic table are defined in Table 1. The given atomic weights for the elements in the same column of the periodic table have as close values as possible, and the most common elements in organic compounds have lower values. To avoid generating the same code for different environments, all given atomic weights are prime numbers, except halogens.

The first layer code can be represented mathematically as

$$C_1 = W_1 \times W_{a1}$$

In this equation, C_1 is the first layer code; W_1 is the given layer weight (all layer weight values are defined in Table 2), i.e., $W_1 = 1$, W_{a1} is the given atom weight of the center atom. If the center atom is a carbon, $W_c = 59$, $C_1 = 59$.

2. Consider the nearest neighborhood of the center atom as the second layer, i.e., the non-hydrogen atoms and bonds linking these atoms with the center atom. Then, the second layer code can be represented as follows:

$$C_2 = C_1 + W_2 \sum_{i=1}^{n} W_{bi} \times W_{ai} \tag{1}$$

In the equation, C_1 is the first layer code, W_2 is the second layer weight, n is the number of all non-hydrogen atoms linked with the center atom, W_{ai} is the center atom, W_{ai} is the given atom weight of the ith atom linked with the center atom, W_{bi} is the bond weight of the bond between the ith atom and the center atom. Bond weights are defined in Table 3.

3. Up to layer k, consider the spherical environment of layer $k-1$, i.e., all non-hydrogen atoms and bonds linked with the atoms of layer $k-1$, that have not been calculated. The code of layer k can be calculated by the following equation:

$$C_k = C_{k-1} + W_k \sum_{i=1}^{n} W_{bi} \times W_{ai} \tag{2}$$

In this case, C_{k-1} is the layer code of layer $k-1$, n is the number of the non-hydrogen atoms (have not been calculated so far) linked with all atoms at layer $k-1$, W_{ai} is the given atom weight of the ith atom linked with atoms at layer $k-1$, and W_{bi} is the bond weight of the bond between the ith atom and the layer $k-1$ atom.

⊗ Abstract published in *Advance ACS Abstracts,* May 1, 1997.

S0095-2338(96)00145-X CCC: $14.00

4. Check all atoms of layer k, if the spherical environment of an atom has been calculated go to step 5, else go to step 3 to calculate the code for layer $k+1$.

5. Here, a set of codes have been obtained, corresponding to various layers on the selected non-hydrogen atom as center atom. If all non-hydrogen atoms in the compound have been selected as center atom go to step 6, else select another nonselected atom as new center atom and go to step 1.

6. Arrange all codes based on center atoms in the compound in the order of value on layers, the equal codes in same layer are degenerated, and the number of equivalence is called degree of degeneration.

A set of codes and degrees of degeneration can be obtained by the above procedure and is described as the structure features of the compound.

Illustration. An example (compound **1**) from ref 5 is selected, and the structure codes by the above algorithm are calculated. Its structure and the arbitrary atomic number are shown in Figure 1.

When atom 1 is selected as center atom, the first layer atom is C, i.e., $W_{a1} = 59$, $W_1 = 1$, then the first layer code is $C_1 = W_1 \times W_{a1} = 59$. For the second layer, there are three atoms linked with atom 1 and they are

$$\left.\begin{array}{lll} 2 & C(59) & \text{aromatic bond}(47)\\ 6 & C(59) & \text{aromatic bond}(47)\\ 7 & C(59) & \text{single bond}(17) \end{array}\right\}$$

$$n_2 = 3, \quad k = 2, \quad W_k = 3, \quad \text{and} \quad C_1 = 59$$

$$C_2 = C_1 + \sum_{i=1}^{n_2} W_{bi} W_{ai} = 59 + 3(47 \times 59 + 47 \times 59 + 17 \times 59) = 19706$$

For the third layer, the sphere environment contains four atoms (see Figure 1)

$$\left.\begin{array}{lll} 3 & C(59) & \text{aromatic bond}(47)\\ 5 & C(59) & \text{aromatic bond}(47)\\ 8 & C(59) & \text{single bond}(17)\\ 16 & 0(67) & \text{double bond}(19) \end{array}\right\}$$

$$n_3 = 4, \quad k = 3, \quad W_k = 5$$

$$C_3 = C_2 + \sum_{i=1}^{n_3} W_{bi} + W_{ai} = 19706 + 5(47 \times 59 + 47 \times 59 + 17 \times 59 + 19 \times 67) = 58816$$

Because of the sphere of atom 16(O) in the third layer has been calculated, the coding procedure for atom 1 as the center atom finishes. Perform the same procedure for other atoms and arrange all codes in Table 4.

By ordering and degenerating the results in Table 4, the final codes of compound **1** can be obtained (shown in Table 5).

INDEX FILES

All compounds in a database can be coded by using the above steps. It can be seen from the coding method that the first layer code corresponds to the types of atoms constructing the molecular and the degree of degeneration is the number of atoms for each type. After all codes of compounds in the database have been calculated an index file can be created by arranging these codes in value order on layers and degrees of degeneration. The global structure of a compound can be considered as a special one of substructures; however, in order to get a high search speed, an independent global structure index file can be created by just taking the outermost codes of various center atoms and arranging them. From Table 4, the global structure codes of compound **1** can be obtained (shown in Table 6).

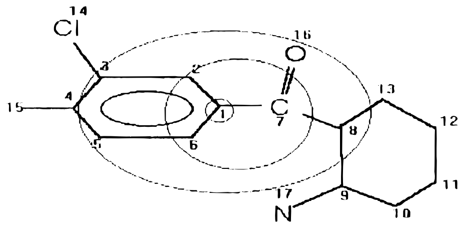

Figure 1. Compound **1**.

Table 1. The Given Atom Weights of Elements in the Periodic Table

H 1																	
Li 113	Be 41											B 43	C 59	N 61	O 67	F 10
Na 143	Mg 179											Al 53	Si 79	P 71	S 97	Cl 20
K 149	Ca 181	Sc 199	Ti 211	V 223	Cr 227	Mn 229	Fe 233	Co 239	Ni 241	Cu 251	Zn 257	Ga 73	Ge 89	As 101	Se 107	Br 30
Rb 163	Sr 191	Y 263	Zr 269	Nb 271	Mo 277	Tc 281	Ru 283	Rh 293	Pd 307	Ag 311	Cd 313	In 83	Sn 109	Sb 131	Te 127	I 40
Cs 167	Ba 193	La 317	Hf 331	Ta 337	W 347	Re 349	Os 353	Ir 359	Pt 367	Au 373	Hg 379	Tl 103	Pb 139	Bi 151	Po 137	At 50
Fr 173	Ra 197	Ac 463														

La 317	Ce 383	Pr 389	Nd 397	Pm 401	Sm 409	Eu 419	Gd 421	Tb 431	Dy 433	Ho 439	Er 443	Tm 449	Yb 457	Lu 461
Ac 463	Th 467	Pa 479	U 487	Np 491	Pu 499	Am 503	Cm 509	Bk 521	Cf 523	Es 541	Fm 547	Md 557	No 563	Lr 569

SEARCH AND MATCH

We can code all compounds in a database and create index files by the above steps, but the ultimate aim is to realize structure search and match. As mentioned above, structure search and substructure search uses different index files.

Global Structure Search. The search of structure uses the index files of the outermost codes in the database. The query compound can be coded in the same way as the compounds in the database. Comparing the outermost codes and the degeneration of the compound with the index files on layers, the search procedure can be stated as follows.

Suppose there are n atoms and n outermost codes in the target compound, and m unequal codes and degrees of degeneration have been left after being degenerated, which are C_i and k_i ($i = 1$ to m), respectively. If the set of compounds (each compound has an ID number) in the database containing the outermost code of C_i and the degree of degeneration of k_i is A_i, that is, $A_i = \{C_i, k_i | \text{ID}(C_i, k_i) =$ ID number of compounds containing the outermost code C_i and the degree of degeneration $k_i\}$, the search result is a set A, intersection of all members of A_i, i.e.,

$$A = mIi=1A_i$$

Table 2. Layer Weights

layer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
weight	1	3	5	7	11	13	17	19	23	29	31	37	41	43	47	49	53	57	59	61

Table 3. Bond Weights

	bond										
	single	double	triple	cyc-single	cyc-double	cyc−triple	coor bond	ionic bond	aromatic bond	hashed bond	wedged bond
weight	17	19	23	29	31	37	41	43	47	49	53

Table 4. The Codes on Various Atoms in Compound **1**

atomic no.	layer 1	layer 2	layer 3	layer 4	layer 5
1	59	19706	58816		
2	59	16697	51142		
3	59	17717			
4	59	19706			
5	59	16697	49442		
6	59	16697	49442	111217	
7	59	9896			
8	59	13334	47009		
9	59	13436			
10	59	10325	32620		
11	59	10325	27435	58648	
12	59	10325	27435	58410	94853
13	59	10325	32450	79595	
14	20	3029	30759	76602	
15	59	3068	30798	72000	
16	67	3430	13460	76236	186291
17	61	3070	20180	51155	113833

Table 5. The Atom-Centered Code of Compound **1**

layer 1		layer 2		layer 3		layer 4		layer 5	
code	DD[a]	code	DD[a]	code	DD[a]	code	DD[a]	code	DD[a]
20	1	3029	1	13460	1	51155	1	94853	1
59	14	3068	1	20180	1	58410	1	113833	1
61	1	3070	1	27435	2	58648	1	186291	1
67	1	3430	1	30759	1	72000	1		
		9896	1	30798	1	76236	1		
		10325	4	32620	1	76602	1		
		13334	1	47009	1	79595	1		
		13436	1	49442	2	111217	1		
		16697	3	51142	1				
		17717	1	58816	1				
		19706	2						

[a] Degree of degeneration.

Table 6. The Global Structure Code of Compound **1**

layer 2		layer 3		layer 4		layer 5	
code	DD[a]	code	DD[a]	code	DD[a]	code	DD[a]
9896	1	37620	1	58648	1	94853	1
13435	1	47009	1	72000	1	113833	1
17717	1	49442	1	76602	1	186291	1
19706	1	51142	1	79595	1		
		58816	1	111217	1		

[a] Degree of degeneration.

If the structure code is unique, set *A* should contain an element or be empty. An empty set *A* shows that the database does not contain the target compound, while one element means that the compound of the ID number in the database is just the target compound and the interesting information can be retrieved by the ID number from the database. In general, the conclusion is considered correct because we have not met the same code for different structures in plenty of applications, but so far the uniqueness of the structure code has not yet been proven in theory. Therefore, if *A* contains more than one element, other treatments have to be taken, such as atom-by-atom matching, to determine the target compound.

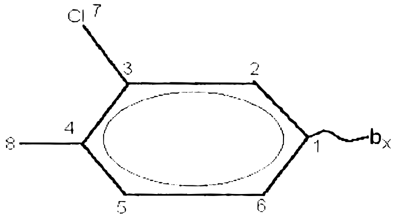

Figure 2. Query substructure **2** (b_x is an undefined bond).

Table 7. The Structure Codes of Query Substructure **2**

atom no.	layer 1	layer 2	layer 3	layer 4
1	59			
2	59	16697		
3	59	17717		
4	59	19706		
5	59	16697	49442	
6	59	16697		
7	20	3029	30759	76602
8	59	3068	30798	72000

Table 8. The Final Codes of Query Substructure **2**

layer 1		layer 2		layer 3		layer 4	
code	DD[a]	code	DD[a]	code	DD[a]	code	DD[a]
59	1	16697	2	49442	1	72000	1
		17717	1			76602	1
		19706	1				

[a] Degree of degeneration.

Substructure Search. Substructure is a structure or part of a structure. In substructure search, the aim is to determine whether the structure in the database is same as the query substructure or includes it. For this purpose, the definition of substructure (query substructure) and its coding method have to be defined at first. Based on the method of atom-centered multilayer code, each layer first links bonds, except for the first layer, so the query substructure should includes one or more undefined bonds, such as substructure II (see Figure 2). The coding method of query substructure is the same as the above except step 4 is modified such that if the spherical environment of an atom in the *k*th layer contains an undefined bond go to step 5. The codes of substructure 2 can be obtained by such steps (see Table 7).

The codes and degrees of degeneration by ordering and degenerating the data in Table 7 are listed in Table 8. They are the final codes for the query substructure.

Comparing the query codes and its degrees of degeneration with the substructure index files on layers of the database, the search algorithm is similar to the global structure except the definition of set A_i. Besides, if the outermost code of the center atom is from the first layer, and the code is 59,

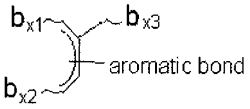

Figure 3. Special query substructure.

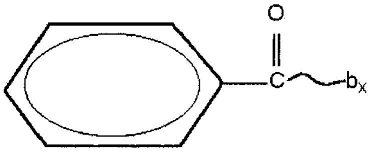

Figure 4. Query substructure.

the comparison would be omitted on the center atom because every compound in the database contains at least one carbon atom. $A_i = \{C_i, K_i|\mathrm{ID}(C_i, \geqslant k_i) = \mathrm{ID}$ number of compounds in the database containing layer code C_i and degree of degeneration, equal or bigger than $k_i\}$.

The result set A is the intersection of A_i. If A contains n ($n \geqslant 0$) elements, it is considered that there are n compounds matching the query substructure in the database.

APPLICATION AND COMPARISON WITH OTHER METHODS

ACMC and BCMC[6] were used as structure search tools in the organic reaction database of our laboratory in which more than 60 000 reactions and more than 200 000 compounds were included. From the application results, we noted that ACMC has some special characteristics. As the fragment that contains one or more than one atom can be considered as a substructure, it should be noted that a fairly wide range of query substructures could be used. It is very useful for some special applications, e.g., we can select the query substructure of Figure 3. But some other algorithms that need to define substructure fragments previously are powerless to do so. On the other hand, application results show that substructure search speed is very fast by ACMC. For example, for the query substructure Figure 4, it took less than 1 s to get the target compounds in our reaction database using a PC-486/100. A great number of query substructures including some special fragments were selected in applications and tests, and no one substructure search exceeded 4 s. For the Beilstein Current Facts in Chemistry 95 System, search of the query substructure of Figure 4 took about 15 s and actually used several minutes for some other query substructures. The reason for fast search speed has relations with the coding method and the search manner of ACMC, because the result codes are all numbers (32 bits) and the process of search and comparison is almost reading sequential files. In comparison with HOSE and HORD, each layer codes in ACMC are managed separately and organized into their own index files individually, so query substructure codes only match with related layer codes instead of all atomic environments in the database, which is the way of HOSE and HORD. Thus, ACMC seems to be faster than HOSE and HORD. However, we have not seen the related data about search speed of HOSE and HORD; therefore, actual comparisons could not be carried out.

CONCLUSION

We present here a general algorithm (atom-centered multilayer code, abbreviated ACMC) to code structures of organic compounds. When coding structures with this algorithm there is no need to define substructure or fragment previously. It only needs the linked relation between atoms in the compound to automatically generate the structure or substructure codes. Although the index files are big by using this method, it is no longer a problem with the current rapid development of computer technology. Therefore the primary task is to increase search speed, especially substructure search speed. With this algorithm, the process of search and comparison is almost reading sequential files, which means that the match method is simple and the performance is excellent. Since the smallest unit is an atom in this coding method, a fragment that includes one or more than one atom can be considered as a substructure; therefore, the range of substructure search is enlarged.

ACKNOWLEDGMENT

We are grateful to the National Science Foundation of China and Tianjin Science Foundation for financial support.

REFERENCES AND NOTES

(1) Wiswesser, W. J. *A Line-formula Chemical Notation*; Crowell: New York, 1954.
(2) Xu, L.; Guo, C. J. *Computer Chemistry: Method and Application*; Chemical Engineering Press (China): BeJing, 1990.
(3) Attias, R. DARC Substructure Search System: A New Approach to Chemical Information. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102−108.
(4) Handbook of CIDS Chemical Search Keys; Fein-Marquart Assoc., Inc., Towson, 1973.
(5) Bremser, W. Horse-A Novel Substructure Code. *Anal. Chem. Acta* **1978**, *103*, 355−365.
(6) Qiao, Y.; Xiao, Y.; Zhang, J.; Lin, S.; Fast Matching of Organic Compound Structure: A New Techniques for Structure Coding. *J. Comput. Chem.* In press.

CI960145I