

## A novel program for computer-aided generation of 2D chemical structures

B. P. Stoyanov<sup>1</sup>, E. P. Petrov<sup>1</sup>, N. T. Kochev<sup>2</sup>, I. P. Bangov<sup>3\*</sup>

<sup>1</sup>Department of Computer Informatics, Faculty of Mathematics and Informatics, Konstantin Preslavski University of Shumen, 115 Universitetska str., 9712 Shumen, Bulgaria

<sup>2</sup>Plovdiv University "Paisii Hilendarski", Faculty of Chemistry, Department of Analytical Chemistry and Computer Chemistry, 24 Tsar Assen str., 4000 Plovdiv, Bulgaria

<sup>3</sup>Department of Chemistry, Faculty of Natural Sciences, Konstantin Preslavski University of Shumen, 115 Universitetska str., 9712 Shumen, Bulgaria

Received April 30, 2014; Revised July 01, 2014

*Dedicated to Acad. Dimiter Ivanov on the occasion of his 120<sup>th</sup> birth anniversary*

A novel software product STRGEN-2D for generation of 2D chemical structures from a gross formula is implemented. It is based on an approach developed by one of the authors (IB). Various functionalities and options of the program are discussed. The output of STRGEN-2D software is compared with the generated structures by MOLGEN software. The generation efficiency and correctness of obtained structure sets are proved.

**Key words:** structure generation, 2D chemical structures, generation tree, depth-first approach

### INTRODUCTION

2D computer-aided structure-generation plays a special role in QSAR/QSPR investigations. Structure-generator software provides new prospective candidate-structures for both biological activity and chemical and physical properties studies in these fields.

The implementation of 2D chemical structure generator programs comes across on some specific mathematical problems. Its algorithm is exponential which leads to the so called combinatorial explosion. This explosion is conducive to generation of enormous number of structures with the increase of number of atoms participating in the generation process. Hence it is challenging to be processed by the most powerful computer equipment on the one hand, and on the other hand it is virtually impossible to be inspected by the users. Furthermore, the generation is a blind process which leads to so called isomorphism problem i.e. generation of isomorphic (duplicated) structures. Thus, if we have  $N$  atoms a permutation of 2 atoms will produce  $(N-2)!$  duplicated structures.

During the 80's a novel approach to the solution of these problems was devised by one of the authors (IB) [1-12]. This approach was oriented towards the development of a computer-assisted structure elucidation based on <sup>13</sup>C spectral information [8]. Our new development is oriented

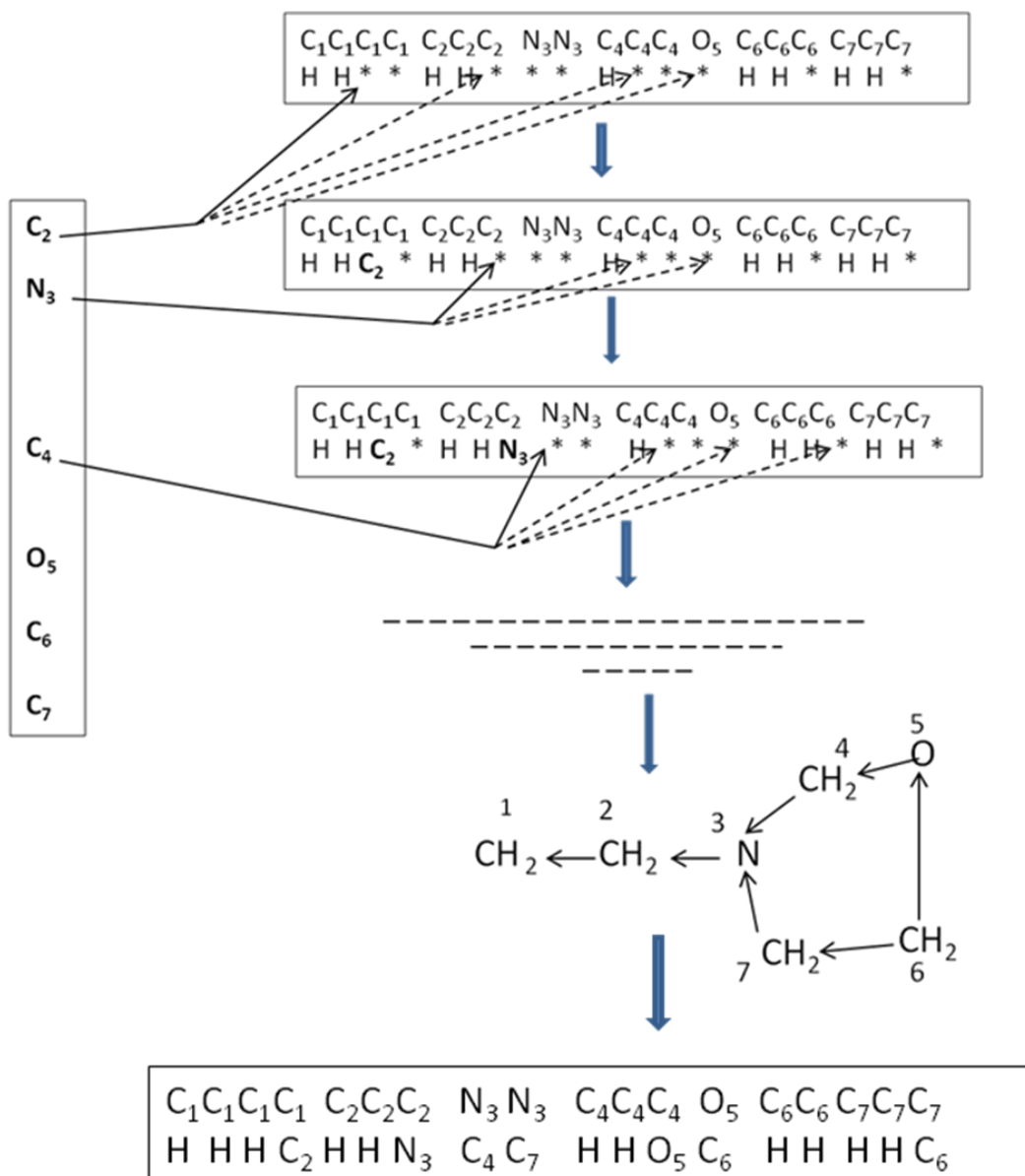
toward QSAR/QSPR problems. The software was implemented in C++ programming language. The aim of this paper is to study the efficiency and correctness of our new software by comparing the generated structures with structures from other sources.

### ALGORITHM DESCRIPTION

A chemical structure is mathematically considered a chemical graph  $G(A,B)$ . Here  $A$  is a set of atoms (vertices of the graph), and  $B$  is a set of bonds (edges of the graph). Efficient two-row representation of the chemical structure is used [1-3] as it is illustrated in Fig. 1. In order to alleviate the combinatorial process we devised the following separation of the chemical graph vertices. Consider a chemical structure represented by a directed graph (the orientations of the edges are given by arrows) depicted in Fig. 1.

Hence, we separate the vertices into two groups, vertices presented in Fig. 1 with arrows pointing to other vertices (we call them Saturating Valences - SVs), and vertices which accept the arrows, called Saturation Sites, SSs). One can see that each atom except the first has 1 SV and  $n-1$  SSs. Here  $n$  is the atom valency. The first atom has all its  $n$  valencies of SS type. In case of a cycle closure (see atom 6 from Fig. 1) a SS is transformed into SV. Accordingly the generation process is carried out by level by level saturation of the SS with the SVs, each level represented by one SV as shown in Fig. 1.

\* To whom all correspondence should be sent:  
E-mail: ivan.bangov@gmail.com



**Fig. 1.** Two-dimensional matrix structure representation and depiction of level-by-level structure generation.

Each saturation produces a fragment which is practically an extension to the formation of complete chemical structures. Thus, a structure generation tree shown in Fig. 2 is formed.

The process of saturation can be carried in two ways: either depth-first with backtracking. In this case the formation of a structure goes to its end and then follows backtracking over the generation tree and generation of a new structure. The other approach is breath-first approach. In this case all the SSs of a given level are saturated producing fragments the their SSs are saturated at the next level, and so on, to the generation of the full set of structure at the highest level.

For our QSAR/QSPR we have found the latter approach to be more appropriate. This allows some further developments towards pruning any chemically and physically inappropriate extensions, on the one hand, and the introduction of parallel processing on the other [9].

Our approach toward the isomorphism problem lies on the following principles:

- ❖ as it was suggested [4-6] the isomorphism is a consequence of the automorphism (vertex equivalence) the equivalence of the different vertices at each extension is studied by using the following local charge-related index (LCI):

$$Li = n_v + q - N_H \quad (1)$$

- ❖ where  $n_v$  is the atom valence,  $q$  is the atom charge and  $N_H$  is the number of attached hydrogen atoms. The equivalent vertices (atoms) give the same value  $L_i$ s within the computer-word accuracy we accept up to 6th sign after decimal point. Hence, permutations between equivalent atoms are forbidden. Such permutations are generated if two equivalent SVs saturate two equivalent SSs and then exchange their SSs. As shown in Fig. 1. Hence, in case of equivalent SVs saturating equivalent SSs, each next level starts from the atom next to the atom that has been saturated in the previous level.
- ❖ Besides the previous principle some duplications still remain because of the mirror symmetry of some structures cannot be predicted during the course of generation.

Further, the  $CTI$  index for each extension has been calculated and compared with the  $CTI$  indices [3,13-15] of the extensions at the current level. The  $CTI$  has the following form:

$$CTI = 1/2 \sum_i \sum_j \frac{L_i L_j}{D_{ij}} \quad (2)$$

Here  $L_i$  are the local indices (1) and  $D_{ij}$  are topological distance matrix elements. In a recent paper [15] it was shown by us that  $CTI$ s provide extremely good discriminating power, the equivalent (isomorphic) structures have the same  $CTI$ , whereas the different structures possess different  $CTI$ s within a computer word precision up to 7<sup>th</sup> sign after the decimal point. Thus, all the remaining branches containing isomorphic structures are pruned from the generation tree.

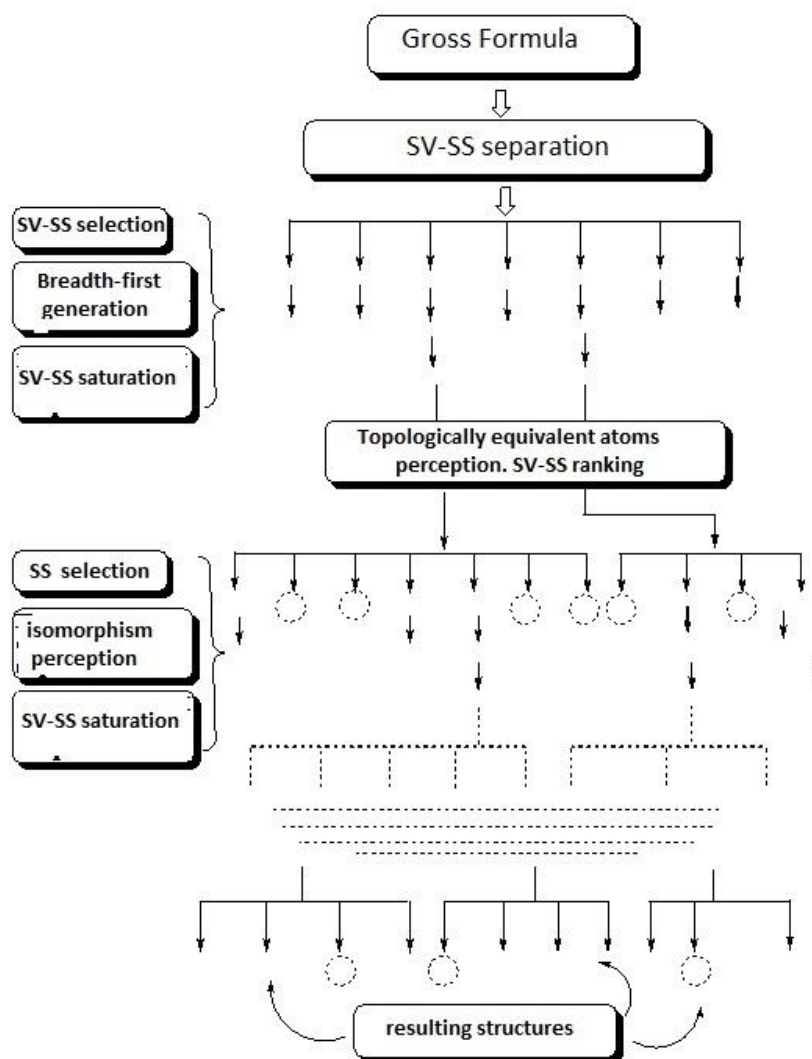


Fig. 2. Structure generation tree with depiction of the steps of structure generation algorithm.

## TEST RESULTS

In order to test the correctness of the result structure set from STRGEN-2D software we performed analogous calculation with MOLGEN [16,17] method for various gross formulas. The results from MOLGEN were exported as \*.MOL files and additionally converted into canonical linear notations SMILES. The STRGEN-2D program transforms internally the structures into SMILES forms. The generated from the two programs structures were compared on the base of both the SMILES canonical linear notation and with the *CTI* index mentioned above. The comparison results for some gross formulas are provided in Table 1. One can see that the numbers of generated structures by our method are the same with these generated by MOLGEN method.




**Table 1.** Number of generated structures by MOLGEN and STRGEN-2D programs and times of generation by STRGEN-2D program.

Gross formula	MolGen (number of structures)	STRGEN 2D	
		Number of structures	Time in seconds
C6N1H11	1111	1111	4.85
C6N2H14	2338	2338	11.82
C6N3H17	1395	1395	7.56
C6N2O1H14	31984	31984	2478.30
C6N3O1H17	20368	20368	1754.04
C6O2H12	1313	1313	4.72
C6O3H12	6171	6171	83.27
C6O4H12	24562	24562	2059.28

The times of generation from the STRGEN-2D software are also provided in Table 1. However they cannot be compared with these from the MOLGEN program for several reasons. Our program runs on a 64bit processor while the MOLGEN program on a 32bit processor. Furthermore, the two programs have different output which produces different times. And more important, our breadth-first approach with exploring each level makes the program slower. Instead of simple combinatorial process the program generates SMILES codes of the fragments at each level. However, we need this approach for our further QSAR/QSPR developments. Accordingly, it was clear for us that the STRGEN-2D is slower than the other programs for structure enumeration, only.

Additionally some results generated by using fragments which obviously reduce the number of generated structures are provided in Table 2.

**Table 2.** Number of generated structures having certain fragments.

Gross Formula	Fragments		Number of generated structures
C6N2O2H14	CC	C-C	32192
	C=N	C=N	
	CO		
C6N1H11	C=NC	C=N-C	141
C6N1H11	C=C=N	N=C=C	18
C6O3H12	C=O	C=O	703
C6N3H17	NN	N-N	708
	CN	C-N	
C6N2H14	C1CC1		63
	NN	N-N	
C6N2H14	C1CCN1		198

## CONCLUSION

A software product for two dimensional structure generation was created. It was developed towards the perception of novel chemical structures for a further QSAR/QSPR processing. It is based on a multilevel breath-first approach, at each level a set of substructures (fragments) is generated and transformed in SMILES codes.

**Acknowledgements:** We acknowledge the Bulgarian Science Fund for its financial support (Grant SFSI I01/7).

## REFERENCES

1. I. P. Bangov, *MATCH Commun. Math. Comput. Chem.*, **14**, 235 (1983).
2. I. P. Bangov, K. D. Kanev, *J. Mathem. Chem.*, **2**, 31 (1988).
3. I. P. Bangov, *J. Chem. Inform. Comput. Sci.*, **30**, 277 (1990).
4. I. P. Bangov, *J. Chem. Inf. Comput. Sci.*, **32**, 167 (1992).
5. I. P. Bangov, *MATCH Commun. Math. Comput. Chem.*, **27**, 3 (1992).
6. I. P. Bangov, *J. Chem. Inf. Comput. Sci.*, **34**, 318 (1994).
7. I. P. Bangov, M. I. Spassova, *Bulgarian Chem. Commun.*, **28**, 443 (1997).
8. T. Laidboeur, I. Laude, D. Cabrol-Bass, I. P. Bangov, *J. Chem. Inf. Comput. Sci.*, **34**, 171 (1994).
9. I. Bangov, S. Simova, I. Laude, D. Cabrol-Bass, *J. Chem. Inf. Comput. Sci.*, **34**, 546 (1994).
10. I. P. Bangov, *Discrete Appl. Math.*, **67**, 27 (1996).

11. I. P. Bangov, in: Handbook of Chemoinformatics, **1**, 178 (2003).
12. I. P. Bangov, *Annual of Konstantin Preslavski Shumen University, Natural Sciences, Chemistry*, **XXI B1**, 29 (2011).
13. P. A. Demirev, A. Dyulgerov, I. P. Bangov, *J. Math. Chem.*, **8**, 367 (1991).
14. N. Kochev, V. Monev, I. P. Bangov, in: Chemoinformatics. A Textbook, 291 (2003).
15. E. Petrov, B. Stoyanov, N. Kochev, I. Bangov, *MATCH Commun. Math. Comput. Chem.*, **71**, 645 (2014).
16. C. Benecke, R. Grund, R. Hohberger, A. Kerber, R. Laue, T. Wieland, *Anal. Chim. Acta*, **314**, 141 (1995).
17. MOLGEN (Molecule Structure Generation), <http://www.molgen.de/> (last accessed 25 April, 2014).

## ЕДНА НОВА ПРОГРАМА ЗА ГЕНЕРИРАНЕ НА 2D ХИМИЧНИ СТРУКТУРИ С ПОМОЩТА НА КОМПЮТРИ

Б. П. Стоянов<sup>1</sup>, Е. П. Петров<sup>1</sup>, Н. Т. Кочев<sup>2</sup>, И. П. Бангов<sup>3\*</sup>

<sup>1</sup>Катедра по компютърна информатика, Факултет по математика и информатика, Шуменски университет "Константин Преславски", ул. Университетска 115, 9712 Шумен, България

<sup>2</sup>Катедра по аналитична химия и компютърна химия, Пловдивски университет "Паисий Хилендарски", ул. Цар Асен 24, 4000 Пловдив, България

<sup>3</sup>Катедра Обща химия, Факултет по природни науки, Шуменски университет "Константин Преславски", ул. Университетска 115, 9712 Шумен, България

Постъпила на 30 април 2014 г.; Коригирана на 01 юли 2014 г.

(Резюме)

Един нов софтуерен продукт STRGEN-2D за генериране на 2D химични структури от брутна формула бе създаден. Той се основава на един подход разработен от един от авторите (ИБ). Различните функции и опции са дискутирани. Изходните резултати от софтуера STRGEN-2D са сравнени с генерираните структури със софтуерния продукт MOLGEN. Ефективността и правилността на получените структури при генерацията са доказани.