Computer Science Journal of Moldova, vol.**1**, no.3(3), 1993

Applied program system for the prognosis of the biological activity of chemical compounds: development and use

N.Shvets

Abstract

An applied program system for the biological activity investigation and new compounds with the activity needeed creation is presented. Based on the electron–topological (ET–) method, it has a high predictive rate, about 95 %, and proved to be especially appropriate for the fragrance activity prediction. Here the ET–method is described in short, the structure of the system in view is shown and an example of the concrete task, sandalwood smell investigation, solved using the system is given.

1 Introduction

Computers are now a widespread tool for investigations in different areas of sciences. Computer chemistry may serve not only as an example of many–purpose use of computers but as a source of new applications of the mathematical and computing methods known also.

Biological activity may be considered as a reaction of a biological system (in the whole or of its parts) on the effect of chemical compounds, expressed by means of quantitative characteristics. It is a part of more common "structure–property" problem in chemistry concerning all kinds of interactions in the nature. Biology, agriculture, medicine, food and parfume industry represent a list of areas, far from being full enough, where this problem is considered a central one. Due to this fact a new branch has appeared in the theoretical chemistry,

called molecular engineering, the main aim of which is to solve the problem of the new compounds with the given properties creation.

In principle, computer assisted design of molecules with the properties given beforehand (especially of biologically active compounds) should be based on the knowledge of the micromechanism of interaction between a molecule of active compound and corresponding bioreceptor. The direct study of such micromechanisms is rather difficult and they are hardly known in advance in the most cases,even qualitatively. Therefore many attempts were made to develop indirect methods based on the study of molecular structures in connection with their activities, unrelated to the mechanism mentioned, and to reveal the main features of the molecules responsible for the activity. Such methods are known in literature as quantitative structure–activity relation (QSAR) methods and use extensively mathematical techniques for the compound structure description and programming tools familiar enough for the methods implementation.

The first problem to be solved for any QSAR–method is the problem of selecting the appropriate media for the compound structure description (or the appropriate language of description). As far as molecules are spatially dispersed descrete constructs, mathematical objects as vectors, graphs and matrices are used usually as linguistic means for the molecular structure of compounds description. Their elements have their own sense depending on the method chosen and determine, as a rule, computational tools involved in the QSAR–method in view. But some choice of the appropriate computing program from the class of analogous programs may be done.

The majority of QSAR–methods uses some kinds of formal description appropriate for conformational calculations of organic molecules. Much less attention, if any, is paid to the electronic structure of the molecules. Meanwhile in the chemical interaction with the bioreceptor the electronic structure of the molecule together with its spatial topology seems to be of primary importance. Description of the electronic structure combined with molecular conformation (topology) seems to be much more informative to represent the biologic activity, than the formal geometrical descriptors only. This was achieved in the so–called

electron–topological (ET–) method of prediction and design of the biologically active compounds [1]. The essence of the ET–method and its program implementation are described below.

2 The ET–method description

Consider a series of compounds tested on the activity under consideration which are known to be either active, or inactive, or their activity is known quantitatively. The main advantage of the ET–method before the other QSAR–methods viewing the compounds with analogous structure is that the compounds of the series to be studied may be from any class of compounds without any constraints. For each molecule its conformational and electronic structures are determined and arranged as a set of m matrices of the order n, for some m chosen, where n is the number of atoms in the molecule (m is fixed for all the compounds from the same series). So, we have multidimensional matrices called electron–topological matrices of conjunction (ETMC), which serve a language for the description of the molecular structure of chemical substances, being as full and informative as possible.

Every such matrix corresponds to a graph where a few values are prescribed to every vertix and every edge. As a rule, the diagonal elements are presented by the local atomic characteristics and the non–diagonal ones by the different quantum–chemical and physical–chemical characteristics of bonds between the atoms.

When forming the ordinary, $n \times n$ ETMC, we may use a set of values taken from the different layers of the multidimensional matrix and to choose the most appropriate results of the subsequent processing of these matrices (the fragments responsible for the given kind of activity, represented by the submatrices of ETMC). To find these fragments, we compare every ETMC as control compound with the ETMCes for the rest of compounds. In the case when the activity is known quantitatively, ETMCes should be compared within each group of molecules which have the same value of activity.

The advantage of the ET–method is that the compounds structure may differ substantially while in other QSAR–methods it is to be simi-

lar and that it solves to a great extent the problem of interaction "active molecule — bioreceptor" as the "key — lock" problem (complementary structures).

The computational part of the method is presented by the following main classes of calculations:

- conformational analysis
- quantum–chemical calculations
- ETMC formation
- ETMC processing and activity features selection.

The first two classes are traditional enough, others reflect peculiarities of the ET–method.

A general scheme of the ET–method is shown in Fig.1. (On the picture IF is to be read as "input file" and OF as "output file").

æ

3 Program implementation of the ET-method

As it is seen from Fig.1, a program system implementing ET–method, being the system for prognosing new compounds with activity needed, unifies some known programs with the programs specific for the method itself. Besides it must be taken into account that

- there exists a number of programs for conformational analysis and for quantum–chemical calculations
- the search of activity features (i.e. fragments, or submatrices) may be based on different pattern recognition methods, by the user's desire
- the sequence of main steps is fixed, although every step may be fulfilled for one compoumd omly, for a few compounds or for all compounds at once (except the step for the activity features selection).

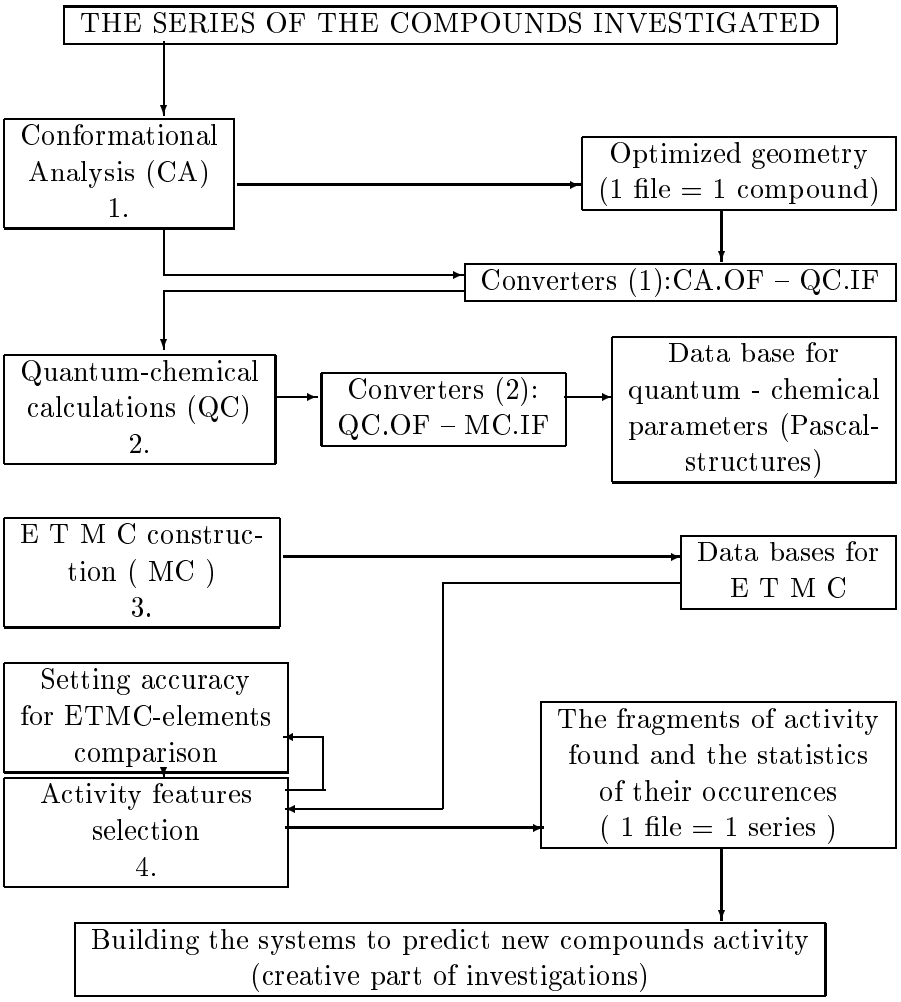

Figure 1 General scheme of the electron-topologic approach.

So, to determine the sequence of steps, the system is built in a form of main menu with the entries–submenues, excluding the entry for the ETMC–formation. There are also entries for service functions such as viewing files, help etc. The submenu presence means that the system may be extended, by the user's demand, at the account of several similar programs as at the first two steps, or in the process of improving the method at the step of the feature selection.

Ready programs inclusion causes many problems as to their call and input/output data representation. To solve the last problem, different data transforming programs (convertors, formatters) are to be written to serve as the intermediates among the neighbouring steps of menu. In that number enter the programs of packing and unpacking the data kept as the result of every previous step (packing is necessary for saving the main memory of computer, unpacking is needed for the intermediate files viewing and subsequent processing).

So, besides the four classes of functions mentioned above,two more classes arise: service functions and data transformers.

A few words is to be said about the program seeking for the activity features covering all the active compounds as much as possible. This search is controlled by a few parameters, some of them determine the limits of acceptability for the values of atomic and quantum–chemical characteristics to be considered the same (initial parameters), the others show the extent of coverage of the classes of active and nonactive compounds by the fragments found (resulting parameters). So, the process of the activity features selection is a cyclic one, and the intervention of the researcher is needed after every step of the cycle to change the values of the initial parameters and to examine the resulting ones, calculated for every fragment found. That is why this program is writtem as a dialogue one.

Other methods which may be used instead are the search of the maximal clique of a graph and the self–teaching program for the method of logical solving rules.

Some aspects of ET–method implementation and use were presented in [2, 3, 4].

4 An example of ET–method application

Quite a lot of problems was solved by the ET–method to predict and design the compounds of different activities and proved it to be very effective. To illustrate the method,the results for a problem of odourant biological activities, sandalwood odour, is given below in view of especially good prognosing ability of the system for such kind of activity [5].

As an example, ETMC for relatively small active compound N1 is given, formed of effective charges on atoms (Qii), Wieberg's indices (Wij), and optimized distances between atoms in the molecule (Rij) (the H–atoms are not given here for short) (see Fig.2). The electron characteristics are given in electron charge units (e), and the distances are given in Å. Ci for all i are carbon atoms enumerated. In view of ETMC symmetry, only its upper triangle is kept in memory of computer and processed.

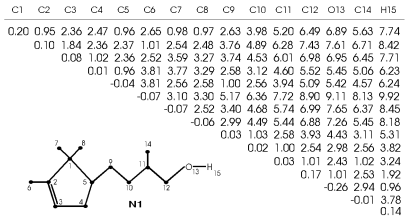

Figure 2. ETMC for a small active compound with sandalwood odour

Computer processing of an array of ETMC for 60 compounds with sandalwood odour allowed to single out a structure fragment of molecules realized in any active compound which takes part in training and is absent in the class of nonactive compounds, as a submatrix (Fig.3). Here *di* for $i = 1, 2, 3, 4$ mean the deviations permitted for every kind

of parameter.

0.15+d1	0.95+d1	5.15+d2	6.45+d2	5.55+d2	6.60+d4	7.74+d4
	0.05+d1	6.30+d2	7.40+d2	6.85+d2	7.50+d4	8.42+d4
		0.01+d1	1.01+d1	1.02+d1	2.43+d2	3.24+d2
			0.15+d1	2.52+d2	1.01+d1	1.92+d1
				-0.01+d1	3.40+d4	3.78+d4
					-0.26+d1	0.96+d1
						0.14+d1

d1 = 0.05
d2 = 0.15
d3 = 0.20
d4 = 0.40

Figure 3. ETMC of the activity fragment singled out

As it is seen from the ETMC fragment obtained, the definite group of hydrogen (H–) atoms belonging to the different parts of molecule and a hydroxil group are to be present for sandal odour revealing. The necessary condition to be fulfilled is that there must be the electronic state of atoms dispersed in the space on definite distances, formed in a fixed way.

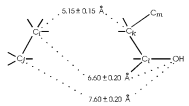

Figure 4. The activity fragment with the geometric parameters

The atoms C_i and C_j are tertiary and quaternary carbon atoms, as

a rule, with ordinary bond between them (Wij = 0.95 + 0.05) and small positive charges. The nearest to them atom C_k (tertiary or quaternary carbon, as a rule) is situated at the distance of 5.15 + 0.15 A. The oxygen atom has negative charge Q = -0.26 + 0.05 e. The realization of the activity fragment is shown in Fig.5 for the molecular skeletons of two active compounds, possessing sandalwood odour, N2 and N3.

All these quantum–chemical characteristics together with the fragment found represent necessary conditions for the sandalwood odour revealing, and being used in the frames of the program system considered, form an expert system for the new active compounds prognosing.

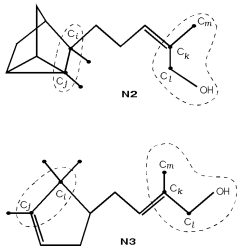

Figure 5. Molecular sceletons of two active compounds known with the fragment found

5 Conclusion

About 20 problems of prediction, screening and disign of the biologically active compounds, including drugs (sulfonilamids, antiallergens, inhibitors of ferments, antitumor medicines), odorants (musk, amber,

sandalwood, garlic, meat, smoke smell, etc.), plant growth regulators, defoliants, food aromatizers,etc. have been successfully solved, and they proved the ET–method to be efficient. Most of them were solved by means of the system being a program implementation of ET–method on personal computers, IBM PC.

PASCAL was taken as the language of implementation in the Turbo-PASCAL environment, and the Turbo Professional means were used for the service functions implementation. The programs for ET-method itself are compact enough (about 150 K), but a great amout of maim memory is needed for the intermediate files keeping, for Turbo Professional modules and for the commonly used quantum-chemical and comformational analysis programs. The memory volume depends also on the number of atoms in a molecule allowed in the system. Now the number is equal to 100.

Because of possibility to enlarge the system with new programs and for the case of a greater number of atoms a new implementation is needed for using more modern programming tools which allow to work with large volume of information, to keep it in a compact form, to include the programs written in different languages and to organize services in a natural way. The C, C++ languages and the environments for them based on the object–programming principles seem to be the most appropriate ones to solve all the problems of the more powerful user–friendly systems design for the new compounds with the activity needed prediction.

References

[1] Anatoly S. Dimoglo. Compositional approach to the electronic structure description of chemical compounds, oriented on the computer analysys of structure–activity relation. Khimiko–pharmazevticheskii zhurnal, 1985, No 4, pp.438–444.

[2] Anatoly S. Dimoglo, Natalie M. Shvets. Development of the electron–topologic approach software for the investigation of the

structure–activity relations. The I–st All–union WATOC conference, 30th of Sept. — 5th of October, 1991, Volgograd, p.530.

[3] Anatoly S. Dimoglo, Natalie M. Shwets. Software implementation of SAR–investigations in the series of compounds with interactive planning of experiment. Proceedings of the IXth All–union conference "Chemical information", Chernogolovka, 11–15th of Jan., 1992, p.226.

[4] Anatoly S. Dimoglo , Natalie M. Shvets. Software development for prognosis and construction of biologically active compounds. 9–th European Simposium on QSAR, Strassburg, Swiss., Sept., 1992.

[5] Anatoly S. Dimoglo, Alexey A. Beda, Natalie M. Shvets, L.A. Heyfits,I.S. Aulchenko. Investigations on the structure — sandalwood odour relations. The electron–topological approach. Reports of Russian Academy of Sciences, 1993, v.328, No 5, pp.570–572.

N.Shvets
Institute of Mathematics,
Academy of Sciences of Moldova,
5 Academiei str.,
Kishinev, 277028, Moldova
phone: +7(0422) 738073
e–mail: 22*nata@math.moldova.su*

Received April 18, 1993