Quantitative Comparison of the Shape of Bio-organic Molecules

Ioan Motoc

Max-Planck-Institut für Strahlenchemie, Mülheim a. d. Ruhr

Z. Naturforsch. **38 a**, 1342–1345 (1983); received November 2, 1983

The paper proposes and illustrates a new method for quantitatively comparing the degree of relatedness of the shape of bio-organic molecules which exhibit the same biological action but which may vary widely in their shape. The method is of interest for drug research in connection with the design and selection of test compounds.

1. Introduction

A revival of interest in the topological aspects of chemistry is apparent from numerous papers and review articles published during the last decade. The interest ranges from general considerations about topological chemistry [1], to topological indices and their correlation with physicochemical properties and bioactivities [2], to topological effects in quantum chemistry [3].

The aim of this paper is to develop a simple, topological-type description of the structure of bio-organic molecules which allows the quantitative comparison of the molecular shape of widely differing structures constituting a data base. The (topographic) concept of investigated receptor space (IRS) previously defined [4] is used with the information energy method [5] to make available an objective criterion to measure the shape (dis)similarity of bio-organic molecules (i.e., the information correlation coefficient R).

2. The Investigated Receptor Space

Consider a data base consisting of the compounds $M_1, M_2, \ldots, M_n$ which exhibit the same biological action. The M_i may be conveniently written as $Y_i F$, where F is the pharmacophoric group and Y_i is a variable part. Let M_r be the most active compound in the data base; then, M_r is the best available "copy" of the receptor site features (including the topographic ones), and it will be considered as the reference structure.

Reprint request to Dr. Ioan Motoc, Max-Planck-Institut für Strahlenchemie, D-4330 Mülheim a. d. Ruhr 1.

One proceeds to superimpose the n molecules upon the reference structure so that the pharmacophore F occupies the same position, and the superposition of the variable fragments upon Y_r is maximal. These are equivalent to the key-inlock theory [6] and the usual assumption [7] that the pharmacophore of each M_i interacts with the same active site residues. In order to simplify the problem, the hydrogen atoms, the variations of bond lengths, as well as the difference between the trigonal and tetrahedral bond angles, are neglected. If the Y_i's are conformationally flexible, it is reasonable to suppose [8] that the drug is recognized by the receptor in one of its major solution conformations. Accordingly, one selects from among the low energy conformations of the Y_i's, $i = 1, 2, \ldots, n$, that conformation which allows the maximal superposition upon the selected conformation of Y_r.

The resultant pattern of vertices (representing the nonhydrogen atoms) and edges (representing the covalent bonds between the non-hydrogen atoms in the molecules considered) is called [4] the investigated receptor space, abbreviated by IRS. The IRS reflects approximately the topography of the receptor space explored by the molecules in the data base considered.

The IRS offers a very convenient basis from which the shape of the M_i's are described numerically and compared quantitatively. One ascribes to each M_i the m-dimensional vector $\boldsymbol{X}_i = [X_{ij}]$, $j = 1, 2, \ldots, m$, as: $X_{ij} = 1$ if the vertex j is occupied by a non-hydrogen atom in M_i, and $X_{ij} = 0$ if it is empty; here, m stands for the number of the IRS vertices. It is easy to observe that being given the IRS and the numbering of its vertices, the $\boldsymbol{X}_i$ vector has the same meaning as the adjacency matrix associated with M_i.

0340-4811 / 83 / 1200-1342 $ 01.3 0/0. – Please order a reprint rather than making your own copy.

3. Quantitative Comparison of the Shape of Bio-organic Molecules

One may associate with each M_i, *via* the $\boldsymbol{X}_i$ vector, the finite probability scheme P_i:

$$P_i = (P_{ij}),\ j = 1, 2, \ldots, m\,; \qquad P_{ij} = S_{ij} X_{ij} / \sum_{j=1}^{m} S_{ij} X_{ij}\,, \tag{1}$$

where S_{ij} is a measure of the size of the atom j in molecule i. P_{ij} is the probability that M_i will occupy the receptor space centered around the vertex j.

The quantity $E(i)$, given by

$$E(i) = \sum_{j=1}^{m} P_{ij}^2\,; \qquad \frac{1}{m} \leqq E(i) \leqq 1\,, \tag{2}$$

is called [5] the information energy content of P_i, and it is a measure of the uniformity of the system describes by P_i.

Because P_i is related to the shape of the molecule M_i, the degree of relatedness of the probability schemes P_α and P_β will characterize the degree of the shape relatedness of the molecules M_α and M_β. The information correlation coefficient $R(\alpha, \beta)$,

$$R(\alpha, \beta) = \sum_{j=1}^{m} P_{\alpha j} P_{\beta j} / [E(\alpha) \cdot E(\beta)]^{1/2} \tag{3}$$

expresses quantitatively the relationship between P_α and P_β and, accordingly, the relationship between the shapes of the molecules M_α and M_β.

The information correlation coefficient is $R(\alpha, \beta) = 1$ if P_α and P_β are identical repartized (i.e., M_α and M_β have the same shape), and $R(\alpha, \beta) = 0$ if P_α and P_β are indifferent (i.e., the shapes of M_α and M_β are not related). One judges the intermediate values $0 < R(\alpha, \beta) < 1$ using the criteria for the significance of the correlation coefficient r, namely: the correlations having $r > 0.99$ are excellent, $r > 0.95$ satisfactory, $r > 0.90$ fair, $r < 0.90$ poor [9], and $r < 0.65$ insignificant [10].

4. Application and Discussion

Consider the data base consisting of the sixteen sulfamyl benzoyl ester inhibitors of carbonic anhydrase collected in Table 1.

The IRS depicted in Fig. 1b was constructed according to the procedure described in Sect. 2, with the compound no. 6 as the reference structure,

Table 1. Sulfamyl benzoyl esters: Affinity constants[a] (AC, $mole^{-1}$).

I	*R*	log AC (*I*)
1	I – Me	7.98
2	I – Et	8.50
3	I – *n*-Pr	8.77
4	I – *n*-Bu	9.11
5	I – *n*-Pent	9.39
6	I – *n*-Hex	9.39
7	II – Me	6.16
8	II – Et	6.21
9	II – *n*-Pr	6.44
10	II – *n*-Bu	6.95
11	II – *n*-Pent	6.86
12	III – Me	4.41
13	III – Et	4.80
14	III – *n*-Pr	5.28
15	III – *n*-Bu	5.76
16	III – *n*-Pent	6.18

[a] The affinity constants are for carbonic anhydrase taken from [11].

and the zig-zag conformation for the *n*-hexyl moiety (the vertices 15–20). The other molecules were superimposed upon the reference structure seeking maximal superposition. The compounds 7–11 and 15–16 (Table 1) were superimposed upon M_r using the path 15–20 (Figure 1b). As S_{ij} values, we use here the k values calculated by Austel et al. [12] by regression analysis from E_s(Taft) and v(Charton) steric parameters: $k_H = 0.0$, $k = 1.0$ for the 2-nd period elements except for $F(k_F = 0.8)$; $k = 1.2$, 1.3 and 1.7 for the 3-rd, 4-th and 5-th period elements, respectively.

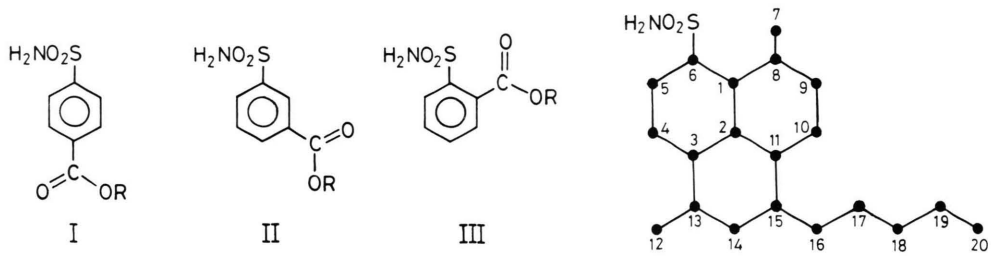
Using the IRS shown in Fig. 1, one ascribes the $\boldsymbol{X}$ vectors to the molecules of the data base. For example, the vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ below correspond to 4-sulfamyl benzoyl *n*-butyl ester, and to 2-sulfamyl benzoyl methyl ester, respectively.

$$\boldsymbol{X}_1 = [1\,1\,1\,1\,1\,1\,0\,0\,0\,0\,0\,1\,1\,1\,1\,1\,1\,1\,0\,0]$$

$$\boldsymbol{X}_2 = [1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0]$$

The mutual information correlation coefficients among the compounds in Table 1 are collected in Table 2. Because the intercorrelation matrix is symmetrical, only the lower triangular part is displayed.

Inspection of Table 2 shows, as expected, significant correlations among the near terms of the three homologous series (the diagonal blocks in Table 2). Comparison of affinity constants (Table 1) and the R values (Table 2) shows that compounds with



(a) (b)

Fig. 1. Sulfamyl benzoyl esters inhibitors of carbonic anhydrase (a) and the corresponding investigated receptor space, IRS (b).

Table 2. Sulfamyl benzoyl esters: the information intercorrelation matrix.

I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1.000															
2	0.954	1.000														
3	0.913	0.957	1.000													
4	0.877	0.920	0.961	1.000												
5	0.845	0.886	0.926	0.964	1.000											
6	0.817	0.856	0.894	0.931	0.966	1.000										
7	0.700	0.763	0.730	0.702	0.676	0.653	1.000									
8	0.667	0.727	0.783	0.753	0.725	0.701	0.954	1.000								
9	0.639	0.696	0.750	0.801	0.772	0.745	0.913	0.957	1.000							
10	0.614	0.669	0.721	0.769	0.815	0.788	0.877	0.920	0.961	1.000						
11	0.592	0.645	0.694	0.741	0.786	0.828	0.845	0.886	0.926	0.964	1.000					
12	0.600	0.572	0.548	0.526	0.507	0.490	0.700	0.667	0.639	0.614	0.592	1.000				
13	0.572	0.546	0.522	0.502	0.484	0.467	0.763	0.727	0.696	0.669	0.645	0.954	1.000			
14	0.639	0.609	0.583	0.560	0.540	0.522	0.822	0.783	0.750	0.721	0.694	0.913	0.957	1.000		
15	0.614	0.669	0.641	0.615	0.593	0.573	0.877	0.836	0.801	0.769	0.741	0.877	0.920	0.961	1.000	
16	0.592	0.645	0.694	0.667	0.643	0.621	0.845	0.886	0.849	0.815	0.786	0.845	0.886	0.926	0.964	1.000

similar shape exhibit similar affinity constants (e.g., R (5, 6) = 0.966 and log AC (5) = 9.39, log AC (6) = 9.39; R (4, 5) = 0.964 and log AC (4) = 9.11, log AC (5) = 9.39; R (8, 9) = 0.957 and log AC (8) = 6.21, log AC (9) = 6.44 etc.), while for the unrelated compounds the affinity constants are greatly different (e.g., R (3, 13) = 0.522 and log AC (3) = 8.77, log AC (13) = 4.80; R (4, 14) = 0.560 and log AC (4) = 9.11, log AC (14) = 5.28 etc.).

Because the steric effects are related to the shape and size of the molecule (or parts of it), it is probable that a good parallelism between the biological response and the information correlation coefficient indicates a sterically controlled biological interaction (for the sulfonamide binding to carbonic anhydrase see [13]). The method we suggest to relate the shape of bio-organic molecules is of interest for drug research in connection with design and selection of test compounds. In order to maximize the information-expense ratio, it is essential that test compounds are chosen to be mutually dissimilar. Our method is complementary to existing criteria for expressing the similarity of chemical compounds in respect to certain physico-chemical properties [14], e.g., lipophilicity (represented by π) and electronic properties (represented by σ); further our method, unlike previous ones, is not restricted to structurally congeneric compounds.

Acknowledgements

The author thanks Prof. Dr. O. E. Polansky and Prof. Dr. J. N. Silverman for valuable discussions.

[1] H. L. Frisch and E. J. Wasserman, J. Amer. Chem. Soc. **83,** 3789 (1961); D. M. Walba, R. M. Richards, and R. C. Haltiwanger, ibid. **104,** 3219 (1982).
[2] L. B. Kier and L. H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York 1976; I. Motoc, A. T. Balaban, O. Mekenyan, and D. Bonchev, Math. Chem. **13,** 369 (1982); A. T. Balaban, I. Motoc, D. Bonchev, and O. Mekenyan, in "Steric Effects in Drug Design", eds. M. Charton and I. Motoc, Top. Curr. Chem., Vol. 114, Springer, Berlin 1983.
[3] O. E. Polansky and M. Zander, J. Mol. Struct. **84,** 361 (1982); O. E. Polansky, M. Zander, and I. Motoc, Z. Naturforsch. **38a,** 196 (1983); W. Fabian, I. Motoc, and O. E. Polansky, ibid. **38a,** 916 (1983); I. Motoc, J. N. Silverman, and O. E. Polansky, Chem. Phys. Lett. (in press).
[4] I. Motoc and O. Dragomir-Filimonescu, Math. Chem. **12,** 117, 127 (1981); I. Motoc, I. Muscutariu, and O. Dragomir-Filimonescu, Kexue Tongbao **27,** 1333 (1982); I. Motoc, Quant. Struct. – Act. Relat. (in press).
[5] O. Onicescu, C. R. Acad. Sci. Paris A**263,** 841 (1966); St. Cerc. Mat. **18,** 1419 (1966).
[6] W. Bartmann and G. Snatzke, eds., "Structure of Complexes Between Biopolymers and Low Molecular Weight Molecules", Wiley-Heyden, Chichester 1982.
[7] K. B. Gibson and H. A. Scheroga, Proc. Natl. Acad. Sci. USA **58,** 1317 (1967); R. Rein, Adv. Quant. Chem. **7,** 335 (1973).
[8] L. B. Kier, Pure Appl. Chem. **35,** 509 (1973); T. M. Bustard and Y. C. Martin, J. Med. Chem. **15,** 1101 (1972); J. P. Green, C. L. Johnson, and S. Kang, Ann. Rev. Pharmacol. **14,** 319 (1974).
[9] P. R. Wells, Linear Free Energy Relationships, Academic Press, London 1968.
[10] J. G. Topliss and R. J. Costello, J. Med. Chem. **15,** 1066 (1972).
[11] R. W. King and A. S. V. Burgen, Proc. Roy. Soc. London B **193,** 107 (1976).
[12] V. Austel, E. Kutter, and W. Kalbfleisch, Arzneim.-Forsch. **29,** 585 (1979).
[13] B. Testa and W. P. Purcell, Eur. J. Med. Chem. **13,** 509 (1978).
[14] V. Austel, Eur. J. Med. Chem. **17,** 9 (1982); J. G. Topliss, J. Med. Chem. **15,** 1006 (1972); ibid. **20,** 463 (1977); P. N. Craig, Adv. Chem. Ser. **114,** 115 (1972); F. Darvas, J. Med. Chem. **17,** 799 (1974).