Journal of Mathematical Chemistry 4(1990)207–215

THE OASIS CONCEPT FOR PREDICTING THE BIOLOGICAL ACTIVITY OF CHEMICAL COMPOUNDS

O. MEKENYAN, S. KARABUNARLIEV and D. BONCHEV
Higher School of Chemical Technology, 8010 Burgas, Bulgaria

Abstract

The existing QSAR approaches are critically assessed. The OASIS methodology is outlined as a generalization of the Hansch method. A large set of calculable geometric (topological, steric) and electronic indices is used to characterize molecular structure. The number of descriptors is reduced stepwise in a preliminary screening procedure, thus strongly diminishing the risk of a chance correlation. The non-formal parameters included in the ultimate mathematical model provide opportunities for shedding light on the biological interaction mechanism. Implemented as an IBM PC pack, the OASIS system is applied to various series of drugs, arriving at successful regression models.

1. Introduction

During recent decades, it has become evident that the classical "trial and error" approach to drug design is ineffective and expensive. A new branch of theoretical pharmacology – quantitative structure–activity relationship (QSAR) – has arisen and proven its fruitfulness in the search for new drugs, pesticides, herbicides, etc. Many accurate predictions of biological activity from QSAR studies have been made [1]. This has prompted scientists to further efforts in developing new QSAR approaches and in improving the existing ones.

Like most of the existing QSAR methods for structurally similar compounds, the OASIS approach (Optimized Approach based on Structural Indices Set) is based on two main assumptions [1,2]. The first states that there is an objective relationship between molecular structure and its biological activity which can be described mathematically. This relationship, derived for a test series of compounds, can then be extrapolated to new compounds. The next assumption is that one can adequately quantitate those global and local properties of molecular structure of significance to the potency of the compound. These two assumptions need to be completed with details on the specific character of the SAR, as well as on the approaches to molecular structure description.

Generally, drugs take part in a large number of interactions in the organism, which in many cases cannot be controlled experimentally. Therefore, the biological activity is an integrated effect of all these interactions. Thus, SAR inevitably has a statistical character and this assumption can be treated as a third main postulate in theoretical pharmacology. On the other hand, the biomacromolecule receptor structure is usually unknown. For structurally similar molecules, the constant receptor cavity

structure implies the possibility of modelling biological interaction statistically, and it is described implicitly by the empirical variables of the model.

The widely used QSAR approach to modelling biological activity of structurally similar molecules is the physicochemical approach. The classical physicochemical method is the linear free energy extrathermodynamic method of Hansch [3]. According to this method, the substituent(s) effect (δ_x) on the interaction rate or equilibrium constant (K) is factored into the substituent effect on the hydrophobic, electronic and geometric characteristics of the unsubstituted compound and can be expressed in a first-order approximation by the following relationship:

$$\delta_x \log K = \delta_x G_{\text{hydrophobic}} + \delta_x G_{\text{electronic}} + \delta_x G_{\text{geometric}} \tag{1}$$

Free and Wilson [4] have developed the other basic approach – the additive method technique. The contribution of each substituent to the overall biological activity of the molecule is expressed by a specific constant, calculated by means of the least-squares fit for a set of linear equations:

$$\text{Biol. activ.} = \text{mean biol. activ.} + \text{sum subst. contributions.} \tag{2}$$

2. OASIS approach

2.1. MOLECULAR STRUCTURE DESCRIPTION

Two different aspects of molecular structure are distinguished: geometric and electronic. The geometric structure is in its turn characterized by both molecular topology and 3D-molecular geometry. Two levels of molecular topology are considered:

(1) Atom–atom connectivity, as described by the molecular graph concept. The following topological indices [5–9], based on different graph features, are used: the Randić connectivity index [10,11], the total distance of the graph (the Wiener number) [12], the Hosoya non-adjacency number [13], the Balaban centric [14] and distance connectivity indices [15], the Zagreb group indices [16], etc.

(2) The combination of atom–atom connectivity with atom or/and bond types as described by the molecular weighted graph concept. Indices of this class are, for example, the extended connectivities of Kier and Hall [17], I'Haya electropy and bondtropy indices [18], neighborhood indices of Ray et al. [19].

The 3D-molecular geometry is characerized by the atom distributions of stable conformations. Various steric indices are of use here: the Wiener number metric analogue [20], the largest interatomic distance, and other indices based on the matrix of interatomic Euclidean distances, the Verloop sterimol indices [21], etc.

The electronic structure of molecules is described for their ground state within the MO LCAO approximation. Different quantum chemical indices are calculated, such as frontier orbital energies, atomic charges and superdelocalizability indices [22], etc.

All molecular descriptors given in the foregoing are calculable from the molecular structure.

2.2. MAIN ASSUMPTION. LOCAL AND GLOBAL DESCRIPTORS

The OASIS method deals with series of congeneric compounds, including an unsubstituted compound and its derivatives. Thus, all compounds of the series under study incorporate a common substructure which is termed a reference or "parent" structure. The compounds display the same kind of biological activity. Hence, we associate the interaction sites that may take part in the biological interaction with positions on the reference structure only, but not with the substituents attached to it. The role of the latter is to promote or deactivate the parent molecule. For this reason, as far as specific biological action is concerned, we consider local descriptors of the reference structure atoms only.

Local descriptors refer to the parent structure sites. They either depend on molecular structure as a whole (e.g. path numbers, charges, superdelocalizability indices), or characterize only the substituent effect (the Hammett and Taft constants, Verloop indices, hydrophobicity index, molecular refraction, etc. [21,23–25]).

The alterations of the overall molecular structure can also condition the variance of biological activity within the series of congeners. Such effects are assessed by global parameters describing molecular topology, 3D-molecular geometry, and electronic structure (graph invariants, steric indices, electronic indices such as frontier orbital energies, dipole moments, etc.). Physicochemical properties of the compounds may also be regarded as a potential factor related to biological activity.

Proceeding from structural parameters, which have a physicochemical interpretation, the OASIS approach can shed light on the interaction mechanism.

2.3. PARAMETER SELECTION AND REGRESSION MODEL

In an endeavor to account for any structural factor that may affect biological activity, we proceed from a large set of topological, steric, and electronic parameters. Stepwise multivariate regression techniques are impractical for so many parameters, due to both the combinatorial complexity and the high risk of a chance correlation. Instead, we have developed a stepwise selection procedure to diminish such a risk. Along this avenue, we have followed the Topliss and Edwards recommendation [26] that "A good approach in correlation studies where a large number of potential variables could be considered, would be to initially select for the correlation study, where possible, a limited group of preferred variables. Any correlation which emerged would then be unlikely to be clouded by chance factors".

The first stage of the OASIS selection procedure partitions the initial set of parameters P into disjoint subsets (clusters) on the basis of the intercorrelation graph IG

(introduced in ref. [27]). Each vertex of this graph denotes one of these parameters. Two vertices i and j in $IG(X)$ are connected by an edge if and only if the correlation coefficient r for the respective parameters P_i and P_j characterizing a certain series of congeners is higher than a specified threshold level X ($0 \leq X < 1$):

$$(i, j) \in E(IG(X)) \quad \text{for} \quad r(P_i, P_j) > X, \tag{3}$$

where E is the set of edges of $IG(X)$.

Some of the $r(P_i, P_j)$ are not calculated. Rather, they are treated as having pairwise correlation values exceeding X, based on logical considerations such as pairs of parameters which belong to the same class of parameters describing the same molecular feature. Thus, there are at least ten topological indices derived from the distance matrix of the graph which will always belong to the same cluster, there are several electronic indices characterizing electron donor (acceptor) properties, etc.

Each cluster of parameters corresponds to a component of $IG(X)$. A prescribed number of clusters is obtained by systematically varying the threshold X value.

The second stage of the procedure produces individual linear correlations between each of the parameters and the biological activity under examination. Then for each of the clusters several best correlating parameters are selected as cluster representatives for the modelling procedure.

The search for the best structure–activity regression model is performed in the third stage of the procedure. Two-parameter models are first obtained after examining pairwise the representatives of each pair of clusters. They are compared with the best one-parameter model, and the procedure is terminated when no statistically significant improvement is found. In the case where the two-parameter models have better statistical estimates, the best models are examined for a possible improvement after adding a third parameter out of all the representatives of the remaining clusters. Again, the procedure would terminate if any of the examined third parameters are found to be insignificant. In the opposite case, the best three-parameter models are selected and a search for a fourth parameter is performed, etc.

In the last stage of the method, the validation of the OASIS model is tested by the cross-validation, i.e. the "leave-one-out" procedure [2,28]. The obtained best models are re-derived by omitting consecutively each one of the N-observables. All the equations should converge, and the deviation intervals of the coefficients should be smaller than the model confidence intervals. The model prognostic capabilities are thus also assessed.

2.4. OASIS CONCEPT ADVANTAGES

A number of points should be mentioned in summarizing the OASIS approach outline. Evidently, the selection of a reasonable number of clusters and their representatives reduces the risk of arriving at models with a chance correlation higher than 1%, a limit usually accepted in QSAR [26]. A convenient selection procedure, however,

makes it possible to proceed initially from a large set of parameters which account for all potential factors influencing biological activity. This is advantageous as compared to both the Hansch method, which utilized only three fixed parameters, and the stepwise multivariate technique, which cannot examine such large sets of parameters without arriving at a very high risk of chance correlation. Moreover, being based on non-formal parameters, the OASIS approach provides opportunities for a closer look at the mechanism of biological interaction, which is not possible for such powerful techniques as the DARC–PELCO [29] and Free–Wilson methods, the principal component analysis [30], etc. There is a good chance of arriving at physically reasonable mechanism hypotheses within the OASIS concept, due to its main assumption of biological activity being produced basically by the reference molecule and only additionally activated (or deactivated) by different substituents.

3. OASIS program pack

The OASIS approach was implemented as an IBM PC program pack. The latter comprises several programs which exchange information by means of data files and a database for substituent constants.

The first program provides input of structural information, as well as storage and retrieval of experimental data in the database. It is assumed that the reference structure geometry and that of the substituents do not alter significantly within the series. Hence, each fragment is entered once regardless of the manner in which it combines in forming a compound. Compounds and substituents are identified within the database by a unique coding of their molecular structure [31–35].

Another program calculates local and global topological and steric indices for all compounds of the series.

Two quantum chemical programs are included in the program pack to calculate electronic parameters. The first implements the Pariser–Parr–Pople [36] semi-empirical method taking into account the molecular sigma skeleton [37–39], while the second uses either the MNDO [40], CNDO/2 or MINDO/3 quantum chemical approaches. Both programs can be used alternatively, depending on the number of compounds in the series, their size and accuracy requirement.

The last program of the pack is used to derive the structure–activity model. It provides the parameter clustering, parameter elimination according to their individual correlation with biological activity, and stepwise parameter selection using a single parameter from each cluster. Various training subsets of compounds can be used within the series in order to validate the optimal regression model or to predict the property values for unknown compounds which also belong to the examined class of compounds.

4. Applications

The OASIS methodology was applied to the modelling of various types of biological activities. Here, we summarize some of the results obtained.

4.1. PURINE ANTITUMOR ACTIVITY [41–43]

The OASIS model for the *in vivo* interaction of 2- and 6-substituted purine derivatives with murine solid tumor adenocarcinoma CA 755 is presented below, together with the test 95% confidence intervals:

$$\log(1/C) = 3.69(\pm 0.14) + 0.51(\pm 0.14)S_6^E + 0.24(\pm 0.14)\pi_6,$$

$$n = 17,\ r = 0.920,\ s = 0.265,\ F = 38.6,\ s' = 0.324, \quad (4)$$

where C is the concentration in mol.kg^{-1} which produces a tumor mass regression of 80%, S_6^E and π_6 are the electron donor superdelocalizability index and hydrophobicity index at position 6 in the purine fragment, respectively. n, r, F, s, and s' denote the number of compounds, the correlation coefficient, the Fisher F-statistics and standard deviations for the model itself and the average one for the 17 "leave-one-out" models, respectively.

4.2. TOXICITY OF PHENOLS TO ALGAE "LEMNA MINOR" [44]

This kind of toxicity was modelled by the molecular hydrophobic index π and the acceptor superdelocalizability indices S_1^N(or S_7^N):

$$\log(1/C) = 3.50(\pm 0.09) + 0.78(\pm 0.20)\pi + 0.36(\pm 0.20)S_1^N,$$

$$n = 25,\ r = 0.983,\ s = 0.219,\ F = 315.4,\ s' = 0.242. \quad (5)$$

4.3. ANTITUMOR ACTIVITY OF SUBSTITUTED TRIAZENES [45]

The activity of 1-phenyl-3, 3-dimethyltriazenes was found to be best modelled by a net electronic charge of the γ–N atom q_3 and an electron acceptor superdelocalizability index of the α– or β–N atom from the triazene fragment, S_1^N(or S_2^N):

$$\log(1/C) = 1100(\pm 460) + 5480(\pm 2270)q^3 + 6810(\pm 2800)q_3^2 - 1.02(\pm 0.55)S_1^N,$$

$$n = 28,\ r = 0.833,\ s = 0.177,\ F = 18.2,\ s' = 0.206. \quad (6)$$

4.4. BENZYLAMINE AND AMPHETAMINE PNMT INHIBITORY POTENCY [45]

The phenyletanolamine N-methyltransferase (PNMT) regulates the epinephrine–norepinephrine ratio, suppressing the formation of norepinephrine. The OASIS studies have revealed that both the benzylamine and amphetamine inhibitory potencies are conditioned by the same types of parameters: the hydrophobic index for one of the meta positions in the phenyl fragment π_3 and the electron donor properties expressed by the

global electronic index E_{HOMO} (or by the donor superdelocalizability indices of the phenyl C-atoms at positions 1, 3, and 5). One of the best models found is:

$$pI_{50} = -17.46(\pm 3.29) - 2.19(\pm 0.35)E_{HOMO} + 1.99(\pm 0.55)\pi_3^2,$$

$$n = 52,\ r = 0.914,\ s = 0.437,\ F = 123.7,\ s' = 0.466, \qquad (7)$$

where pI_{50} is the negative logarithm of the inhibitor concentration, producing 50% inhibition of PNMT when using norepinephine as a substrate.

4.5. SOME REMARKS ON THE OBTAINED OASIS MODELS

The best OASIS models reported in the foregoing are characterized by better statistical estimates than the competitive physicochemical methods. As an example, the Neiman and Quinn model [46] for purine antitumor activity (recalculated for the same correlation sample of 17 compounds as in model (4)) provides $r = 0.836$ (versus the OASIS $r = 0.920$) and $s = 0.371$ (versus 0.265).

The validation statistic s' of our models by the "leave-one-out" procedure in all cases approximates the validation statistic s of the basic model. An additional indication for the model's validation is the small difference between the standard deviation of the model and the mean standard deviation for all $(n - 1)$ models. Thus, $s' - s \approx 0.03$ for models (6) and (7) and ≈ 0.02 for model (5), etc. These results suggest that there is a low risk of a chance correlation and that the OASIS models have good predictive capabilities.

It should be noted that some QSAR methods such as DARC–PELCO [29] in general provide more accurate regression models than the OASIS method. However, these methods are based on topological parameters only, which makes them incapable of elucidating the biological interaction mechanisms. On the contrary, proceeding from molecular descriptors having a physicochemical interpretation, the OASIS models can shed some light on these mechanisms. Thus, in studying the antitumor activity of the substituted triazenes we have revealed the major role of the triazene tail γ-atom, which may be related to the alkylation effect of its neighbouring methyl groups.

Acknowledgement

This work was partially supported by the National Committee for Science (Grants nos. 233 and 385).

References

[1] Y.C. Martin, J. Med. Chem. 24(1981)229.

[2] V.E. Golender and A.B. Rosenblit, *Logikokombinatornye metod'i v construirovanii lekarstv* (Zinatne, Riga, 1983), p. 22 (in Russian).

[3] C. Hansch, J. Med. Chem. 19(1976)1;
C. Hansch and T. Fujita, J. Amer. Chem. Soc. 86(1964)1616.
[4] S.M. Free and J.M. Wilson, J. Med. Chem. 7(1964)395.
[5] A.T. Balaban, I. Motoc, D. Bonchev and O. Mekenyan, Topics Curr. Chem. 114(1984)114.
[6] D. Bonchev, *Information Theoretic Indices for the Characterization of Chemical Structures* (Research Studies Press, Chichester, UK, 1983).
[7] A. Sabljić and N. Trinajstić, Acta Pharm. Yugosl. 31(1981)189.
[8] D.H. Rouvray, in: *Mathematics and Computational Concepts in Chemistry*, ed. N. Trinajstić (Horwood, Chichester, UK, 1986), p. 295; Sci. Amer. 254(1986)40.
[9] M.I. Stankevich, I.V. Stankevich and N.S. Zefirov, Usp. Khim. 57(1988)337.
[10] M. Randić, J. Amer. Chem. Soc. 97(1975)6609.
[11] L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research* (Academic Press, New York, 1976).
[12] H. Wiener, J. Amer. Chem. Soc. 69(1947)17; J. Phys. Chem. 52(1948)1082.
[13] H. Hosoya, Bull. Chem. Soc. Japan 44(1971)2332; J. Chem. Docum. 12(1972)181.
[14] A.T. Balaban, Theor. Chim. Acta (Berlin) 53(1979)335.
[15] A.T. Balaban, Pure. Appl. Chem. 55(1983)199; Chem. Phys. Lett. 89(1982)399.
[16] I. Gutman and N. Trinajstić, Chem. Phys. Lett. 17(1972)535.
[17] L.B. Kier, J. Pharm. Sci. 70(1981)930;
L.B. Kier and L.H. Hall, *Molecular Connectivity in Structure–Activity Analysis* (Research Studies Press, Letchworth, UK, 1986).
[18] W.T. Yee, K. Sakamoto and Y.J. I'Haya, Rep. Univ. Electr. Commun. 27(1976)53; ibid. 27(1977)227.
[19] S.K. Ray, S.C. Basak, C. Roychaudhury, A.B. Roy and J.J. Ghosh, Ind. J. Chem. 20B(1981)894;
S.C. Basak, D.K. Harris and V.R. Magnuson, J. Pharm. Sci. 73(1984)429.
[20] O. Mekenyan, D. Peitchev, D. Bonchev, N. Trinajstić and I. Bangov, Drug Res. 36(1986)176.
[21] A. Verloop, W. Hoogenstraten and J. Tipker, Drug Design 7(1976)165.
[22] F. Peradejordi, A. Martin and A. Cammarata, J. Pharm. Sci. 60(1971)576;
F. Peradejordi, A. Martin, O. Chalvet and R. Daudel, ibid. 61(1972)909.
[23] C. Hansch and A. Leo, *Substituent Constants for Correlation Analysis in Chemistry and Biology* (Wiley, New York, 1979).
[24] C. Hansch, A. Leo, S. Unger, K. Kim, D. Nikaitani and E. Lien, J. Med. Chem. 16(1973)1207.
[25] R.W. Taft, Jr., J. Amer. Chem. Soc. 74(1952)3120.
[26] J.G. Topliss and R.B. Edwards, J. Med. Chem. 22(1979)1238.
[27] I. Motoc, A.T. Balaban, O. Mekenyan and D. Bonchev, Math. Chem. (MATCH) 13(1982)369.
[28] R.D. Cramer III, J.D. Bunce, D.A. Patterson and I.E. Frank, QSAR 7(1988)18.
[29] J.-E. Dubois, in: *Computer Representation and Manipulation of Chemical Information*, ed. W.T. Wipke, S. Heller, R. Fellmann and E. Hyde (Wiley, London, 1974), p. 239;
J.-E. Bubois, in: *Chemical Applications of Graph Theory*, ed. A.T. Balaban (Academic Press, London, 1976), p. 333.
[30] G.K. Menon and A. Cammarata, J. Pharm. Sci. 66(1977)304;
T. Kubota, J. Hanamura, K. Kano and B. Uno, Chem. Pharm. Bull. 33(1985)488.
[31] A.T. Balaban, O. Mekenyan and D. Bonchev, J. Comput. Chem. 6(1985)538, 562.
[32] O. Mekenyan, A.T. Balaban and D. Bonchev, J. Comput. Chem. 6(1985)552.
[33] N. Ralev, S. Karabunarliev, O. Mekenyan, D. Bonchev and A.T. Balaban, J. Comput. Chem. 6(1985)587.
[34] D. Bonchev, O. Mekenyan and A.T. Balaban, in: *Mathematics and Computational Concepts in Chemistry*, ed. N. Trinajstić (Horwood, Chichester, UK, 1986), p. 34.
[35] S. Karabunarliev, O. Mekenyan and A. Dobrinin, Vychislitel'nye Systemy 103(1984)74.
[36] R. Pariser and R.G. Parr, J. Chem. Phys. 21(1965)466, 767;
R. Pariser, ibid. 21(1953)568; 24(1956)250;
J.A. Pople and D.L. Beveridge, *Approximate MO Theory* (McGraw-Hill, New York, 1970).

[37] J. Gasteiger and M. Marsili, Tetrahedron 46(1980)3219.
[38] R.J. Abraham, L. Griffits and P. Loftus, J. Comput. Chem. 3(1982)407;
R.J. Abraham and B. Hudson, ibid. 5(1984)562.
[39] M.J.S. Dewar, *The Molecular Orbital Theory of Organic Chemistry* (McGraw-Hill, New York, 1969).
[40] M.J.S. Dewar and W. Thiel, J. Amer. Chem. Soc. 99(1977)4899;
W. Thiel, J. Amer. Chem. Soc. 103(1981)1413, 1420.
[41] O. Mekenyan, D. Bonchev, D. Rouvray, D. Peitchev and I. Bangov, Eur. J. Med. Chem., in press.
[42] O. Mekenyan and D. Bonchev, Acta Pharm. Yugosl. 36(1986)225.
[43] O. Mekenyan and D. Bonchev, in: *Proc. 3rd Int. Conf. on Chemistry and Biotechnology of Biologically Active Natural Products*, Vol. 3, Sofia (1985), p. 385.
[44] O. Mekenyan, D. Bonchev and V. Entchev, Quant. Struct.-Act. Relat. 7(1988)240.
[45] C. Mercier, O. Mekenyan, J.-E. Dubois and D. Bonchev, Eur. J. Med. Chem., submitted.
[46] Z. Neiman and F.R. Quinn, J. Pharm. Sci. 70(1981)425; 71(1982)618.