ISSN 0012-5008, Doklady Chemistry, 2009, Vol. 429, Part 1, pp. 273–276. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © N.I. Zhokhova, I.I. Baskin, D.K. Bakhronov, V.A. Palyulin, N.S. Zefirov, 2009, published in Doklady Akademii Nauk, 2009, Vol. 429, No. 2, pp. 201–205.

CHEMISTRY

Method of Continuous Molecular Fields in the Search for Quantitative Structure–Activity Relationships

N. I. Zhokhova, I. I. Baskin, D. K. Bakhronov, V. A. Palyulin, and Academician N. S. Zefirov

Received June 24, 2009

DOI: 10.1134/S0012500809110056

The rapid development of methods of design of new pharmaceuticals calls for new efficient computational approaches that can reliably predict various types of biological activity of organic compounds to be synthesized. This is due to the fact that the available methods widely used to search for quantitative structure–activity relationships (QSARs) have significant drawbacks. In particular, common methods of constructing 3D QSARs, such as comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA), underlying the state-of-the-art approaches to the design of new pharmaceuticals are very sensitive to the dimensions, resolution, and spatial alignment of the hypothetical grid constructed around a molecule and used for approximating the electrostatic, steric, and hydrophobic molecular fields, the potentials of the latter being calculated at the grid points as molecular structure descriptors [1–3]. This leads to the ambiguity of the resulting 3D QSAR models and, hence, to the unreliability of the prediction based on these models.

In this work, we propose a new method for constructing 3D QSAR models, namely, the method of continuous molecular fields (MCMF). The basic idea of this approach is in direct analysis of continuous molecular fields rather than a discrete array of their potentials calculated at the points of the discrete grid of finite size (as in the standard CoMFA and CoMSIA methods). Such a description better corresponds to the physical nature of molecular fields; therefore, we can expect better statistical characteristics of 3D QSAR models upon such a substitution. Until recently, it was impossible to use continuous molecular fields in the framework of statistical analysis since common statistical procedures are intended to operate only with finite and limited number of molecular descriptors. Therefore, we were interested to realize this idea on the basis of the latest statistical approaches, for example, the support vector machines (SVM), which is free of this limitation and can operate with an infinite number of variables [4]. This is achieved by using so-called kernels [5]. As is known, for any kernel, there must exist a linear vector space (referred to as the reproducing kernel Hilbert space) in which the former can be uniquely represented as the scalar product of the corresponding vectors. It is evident that the scalar product of the molecular field potential values at the grid points is also a kernel (by definition). Inasmuch as an increase in the grid dimensions and a decrease in the grid cell size do not violate this property, the integral of the product of molecular fields taken over the entire physical space also remains a kernel, which can be used in appropriate statistical methods, such as support vector regression. Thus, this offers possibilities for constructing statistical models based on the description of molecular objects as continuous molecular fields.

The basic element of the MCMF proposed in this work is the procedure of calculation of kernels. The use of these kernels will be exemplified by constructing QSARs in the framework of the statistical method of support vector regression [5].

In the proposed method, the kernel $K(\mathrm{M}_i, \mathrm{M}_j)$, which describes the similarity between molecules M_i and M_j, is calculated as a linear combination of kernels calculated for each of the N_f types of molecular fields:

$$K(\mathrm{M}_i, \mathrm{M}_j) = \sum_{k=1}^{N_f} h_k K_k(\mathrm{M}_i, \mathrm{M}_j), \qquad (1)$$

where h_k is the mixing coefficient of molecular fields, $K_k(\mathrm{M}_i, \mathrm{M}_j)$ is the kernel describing the similarity between the molecular fields of the kth type of the ith and jth molecules. That $K(\mathrm{M}_i, \mathrm{M}_j)$ is a correctly constructed kernel follows from the fact that a linear combination of kernels is a kernel. To assign to the mixing parameters the meaning of the contribution made by a

Moscow State University, Moscow, 119991 Russia

definite type of molecular field, we calculate the normalized version of the kernel:

$$K'_k(\mathrm{M}_i, \mathrm{M}_j) = \frac{K_k(\mathrm{M}_i, \mathrm{M}_j)}{\sqrt{K_k(\mathrm{M}_i, \mathrm{M}_i)K_k(\mathrm{M}_j, \mathrm{M}_j)}}. \quad (2)$$

In this case, the kernel value is represented by a linear combination of the values of normalized kernels with modified mixing coefficients:

$$K(\mathrm{M}_i, \mathrm{M}_j) = \sum_{k=1}^{N_f} h'_k K'_k(\mathrm{M}_i, \mathrm{M}_j). \quad (3)$$

For each kth type of molecular field, the kernel is calculated by summation of the corresponding kernels for each pair of atoms of the ith and jth molecules:

$$K_k(\mathrm{M}_i, \mathrm{M}_j) = \sum_{l=1}^{N_i}\sum_{m=1}^{N_j} K_k(\mathrm{A}^i_l, \mathrm{A}^j_m), \quad (4)$$

where $K_k(\mathrm{A}^i_l, \mathrm{A}^j_m)$ is the kernel that shows the resemblance between the molecular fields of the kth type of the lth atom in the ith molecule and the mth atom in the jth molecule; N_i is the number of atoms in the ith molecule; N_j is the number of atoms in the jth molecule. As in the previous case, that $K_k(\mathrm{M}_i, \mathrm{M}_j)$ is a correctly constructed kernel follows from the fact that a linear combination of kernels is a kernel. We suggest to calculate the $K_k(\mathrm{A}^i_l, \mathrm{A}^j_m)$ kernel by integration of the product of the molecular fields of given atoms over the entire physical space $\Re^3$:

$$K_k(\mathrm{A}^i_l, \mathrm{A}^j_m) = \int\int\int_{\Re^3} \rho^k_{il}(x, y, z)\rho^k_{jm}(x, y, z)dxdydz, \quad (5)$$

where $\rho^k_{il}(x, y, z)$ is the potential of the molecular field of the kth type induced by the lth atom of the ith molecule at the (x, y, z) point of the physical space, and $\rho^k_{jm}(x, y, z)$ is the same for the mth atom of the jth molecule. To simplify integration, we approximate the molecular field by the Gaussian function as is done in the CoMSIA method:

$$\rho^k_{il}(x, y, z) = w^k_{il}\exp\left(-\frac{1}{2}\alpha_k((x-x_{il})^2 + (y-y_{il})^2 + (z-z_{il})^2)\right), \quad (6)$$

where x_{il}, y_{il}, z_{il}) are the Cartesian coordinates of the lth atom in the ith molecule, α_k is the fitting parameter for molecular fields of the kth type, and w^k_{il} is the weight of the contribution of the lth atom of the ith molecule to the molecular field of the kth type. The w^k_{il} is taken to be the partial charge on the lth atom of the ith molecule in the case of an electrostatic field, the Lennard-Jones potential parameters in the case of a steric field, and the contribution of a given atom to the total hydrophobicity of a molecule in the case of a lipophilic field. In this case, the above integral is calculated analytically:

$$K_k(\mathrm{A}^i_l, \mathrm{A}^j_m) = \int\int\int_{\Re^3} \rho^k_{il}(x, y, z)\rho^k_{jm}(x, y, z)dxdydz$$

$$= w^k_{il}w^k_{jm}\int\int\int_{\Re^3}\exp\left\{-\frac{1}{2}\alpha_k[(x-x_{il})^2 + (y-y_{il})^2 + (z-z_{il})^2]\right\}$$

$$\times\exp\left\{-\frac{1}{2}\alpha_k[(x-x_{jm})^2 + (y-y_{jm})^2 + (z-z_{jm})^2]\right\}dxdydz$$

$$= w^k_{il}w^k_{jm}\sqrt{\frac{\pi^3}{8\alpha^3_k}}$$

$$\times\exp\left\{-\frac{\alpha_k}{2}[(x_{il}-x_{jm})^2 + (y_{il}-y_{jm})^2 + (z_{il}-z_{jm})^2]\right\}. \quad (7)$$

Thus, taking into account the above expressions, we can suggest the general formula for the calculation of the kernel for the pair of the M_i and M_j molecules, $K(\mathrm{M}_i, \mathrm{M}_j)$:

$$K(\mathrm{M}_i, \mathrm{M}_j) = \sum_{k=1}^{N_f} h'_k \frac{\sum_{l=1}^{N_i}\sum_{m=1}^{N_j} w^k_{il}w^k_{jm}\exp\left\{-\frac{\alpha_k}{2}[(x_{il}-x_{jm})^2 + (y_{il}-y_{jm})^2 + (z_{il}-z_{jm})^2]\right\}}{\sqrt{\sum_{l=1}^{N_i}(w^k_{il})^2}\sqrt{\sum_{m=1}^{N_j}(w^k_{jm})^2}}. \quad (8)$$

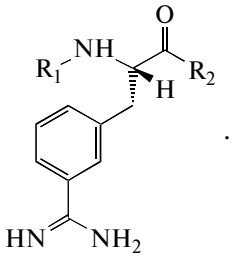
In this case, the value of the predicted property y for a new molecule M can be found by the formula

$$y = \sum_{j=1}^{N_j}(\lambda^-_j - \lambda^+_j)K(\mathrm{M}, \mathrm{M}_j) + b. \quad (9)$$

In addition to the set of parameters $\lambda^-_j - \lambda^+_j$ and b determined by the method of support vector regression, the MCMF method has a number of hyperparameters. First of all, the method of support vector regression ν-SVR itself contains two hyperparameters (ν and C), and their values should be optimized to

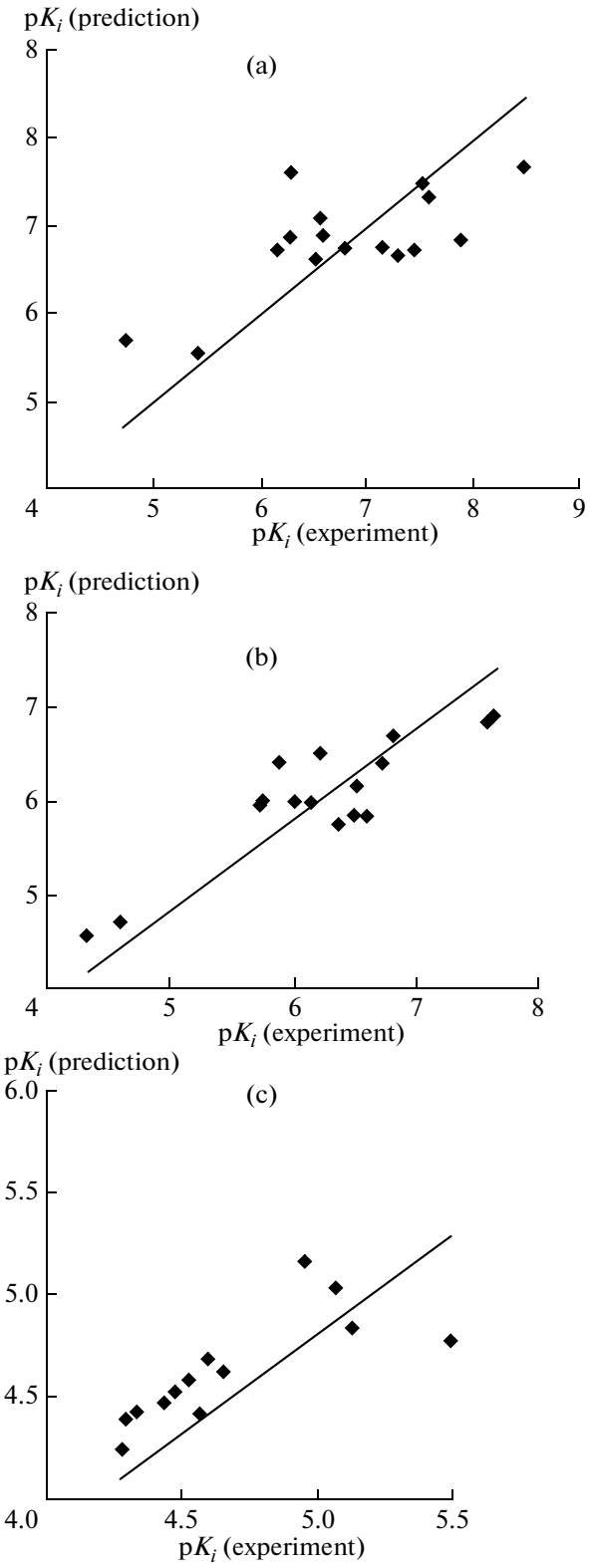
improve the predictive power of the model [5]. In addition, for each molecular field of the kth type, two hyperparameters are introduced: α_k (the fitting parameter that shows the molecular field "decay" rate, which is related to the width of the Gaussian function), h'_k is the mixing coefficient that has the meaning of the relative contribution of a molecular field of a given type. In this work, the optimal values of this set of hyperparameters were found by maximizing the q^2 value by successive use of two nonlinear programming methods: the simulated annealing method [6] and the Nelder–Mead simplex method [7]. The computational procedure of the MCMF method was implemented by us as a program package in C and Python languages.

To check the possibilities of the proposed method for constructing 3D QSAR models, we used a database containing information on 72 structures of 3-amidinophenylalanine derivatives, as well as the data on their inhibiting activity with respect to three serine protease enzymes—trypsin, thrombin, and factor Xa [8]. The general structure of 3-amidinophenylalanine derivatives is



The proteins used as targets for modeling the inhibiting activity of the set have similar structural features: in the active site, the hydroxyl group of the serine amino acid functions as a catalyst. These proteins are often used as potential targets for design of physiologically active compounds. The structures of organic compounds were selected with taking into account their ability to inhibit all three enzymes. In this work, we used the experimental inhibition constants determined by the Dixon method [9] and expressed as the logarithm of the dissociation constant pK_i ($-\log K_i$), which characterizes the degree of binding of the inhibitor with the enzyme (K_i, mol/L).The predictive power of the models was estimated by the leave-one-out cross-validation procedure. The quality of the models was estimated on the basis of the statistical characteristics q^2 and *RMSE*, which were calculated by Eqs. (10) and (11).

$$q^2 = 1 - \frac{PRESS}{SS} = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i^{\text{pred}})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \qquad (10)$$



Correlations of the experimental and predicted logarithms of inhibition constants (pK_i) obtained for the independent test set with respect to (a) thrombin, (b) trypsin, and (c) factor Xa with the use of the MCMF.

Statistical characteristics of the models obtained by the CoMFA, CoMSIA, and MCMF methods for the compounds of the training set

Parameter	Thrombin			Trypsin			Factor Xa		
	CoMFA	CoMSIA	MCMF	CoMFA	CoMSIA	MCMF	CoMFA	CoMSIA	MCMF
q^2	0.697	0.757	0.805	0.635	0.754	0.794	0.429	0.590	0.600
RMSE	0.378	0.244	0.210	0.204	0.159	0.137	0.294	0.194	0.189

where *PRESS* is the predictive sum of squares of differences between the experimental (y_i) and predicted (y_i^{pred}) values of the *i*th biological activity over all *n* compounds included in the cross-validation procedure; *SS* is the sum of squared deviations of the experimental values (y_i) from their arithmetic mean ($\bar{y}$) for each *i*th biological activity.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^{\text{pred}})^2}{n}}, \qquad (11)$$

where y_i is the experimental biological activity of the *i*th compound, y_i^{pred} is the predicted biological activity of the *i*th compound, and *n* is the number of compounds in the set.

The table presents the statistical characteristics of the models obtained for the inhibiting activity of the compounds of the training set with respect to thrombin, trypsin, and factor Xa by traditional 3D QSAR methods and the MCMF. For the CoMFA and CoMSIA methods, the parameters of the models constructed at the optimal grid cell size of 1.5, 1 and 1 Å for thrombin, trypsin, and factor Xa, respectively, are presented.

As follows from the table, all models obtained by the MCMF have a better predictive power than the models obtained by the common CoMFA and CoMSIA methods. The largest contribution to the models is made by steric molecular fields for thrombin, electrostatic and hydrophobic molecular fields for trypsin, and hydrophobic molecular fields for factor Xa. For factor Xa, the range of the experimental pK_i values is rather small, which leads to significantly lower q^2 values for this enzyme as compared with analogous models for thrombin and trypsin.

To independently estimate the predictive power of the models obtained by the proposed MCMF method, we used the test set comprising 16 3-amidinophenylalanine derivatives. These compounds were not included into the training set and were not used for constructing 3D QSAR models. The scatter plots for the experimental and predicted values of the inhibiting activity for the compounds of the independent test set with respect to thrombin, trypsin, and factor Xa are shown in the figure.

Thus, the models constructed with the use of the MCMF, especially when the thrombin- and trypsin-inhibiting activity is predicted, have a rather high predictive power and allow one to reliably differentiate high-active compounds from low-active compounds. This is evidence that the models can be used for efficient design of novel biologically active structures. The proposed method is free from some limitations inherent in the common 3D QSAR methods; i.e., the quality of the models is independent of the alignment inside a 3D grid and of the grid dimensions and grid cell size (the discreteness of estimated fields). Further development of the method will be focused on solving the problem of structure alignment and on choosing the biologically active conformation of compounds.

REFERENCES

1. *3D QSAR in Drug Design. Theory, Methods and Applications*, Kubinyi, H., Ed, New York: Kluwer, 1997.
2. *3D QSAR in Drug Design*, vol. 2: *Ligand–Protein Interactions and Molecular Similarity*, Kubinyi, H., Folkers, G., and Martin, Y.C., Eds, New York: Kluwer, 1998.
3. *3D QSAR in Drug Design*, vol. 3: *Recent Advances,* Kubinyi,H., Folkers, G., and Martin, Y.C., Eds, New York: Kluwer, 1998.
4. Ivanciuc, O., in *Reviews in Computational Chemistry*, Lipkowitz, K.B. and Cundari, T.R., Eds., Weinheim: Wiley–VCH, 2007, vol. 23, pp. 291–400.
5. Vapnik, V., *Statistical Learning Theory*, New York: Wiley-Interscience, 1998.
6. Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P., *Science*, *New Ser.*, 1983, vol. 220, no. 4598, pp. 671–680.
7. Nelder, A. and Mead, R., *Comput. J.*, 1965, vol. 7, pp. 308–313.
8. Böhm, M., Stürzebecher, J., and Klebe, G., *J. Med. Chem.*, 1999, vol. 42, pp. 458–477.
9. Dixon, M., *Biochem. J.*, 1953, vol. 55, pp. 170–171.