Russian Journal of Organic Chemistry, Vol. 38, No. 11, 2002, pp. 1564–1575. Translated from Zhurnal Organicheskoi Khimii, Vol. 38, No. 11, 2002, pp. 1618–1629.
Original Russian Text Copyright © 2002 by Pogrebnyak, Oganesyan, Glushko.

MATRIX, a New Algorithm for Predicting Biological Activity of Organic Molecules Based on Multidimensional Analysis of Physicochemical Descriptors of Modern Pharmaceuticals: I. General Principles

A. V. Pogrebnyak, E. T. Oganesyan, A. A. Glushko

Pyatigorsk State Pharmaceutical Academy, Pyatigorsk, 357500 Russia

Received February 8, 2002

Abstract—The quantum-chemical calculation of structures of organic molecules belonging to 1067 modern pharmaceuticals was carried out by semiempirical (AM1, PM3, MNDO, CNDO/2, MINDO/3) and ab initio (6-31G) procedures taking into account the hydration effects. Each molecule was characterized by 149 topochemical and quantum-chemical descriptors. Basing on combination of multidimensional analysis procedures a new method was developed for forecasting the biological activity of organic compounds consisting in determination of proximity of the molecules on a surface of a potential function in the multidimensional space of descriptors (MATRIX algorithm).

The exponential growth of the number of organic compounds and the high cost of pharmaceutical tests call for development of efficient procedures for prediction of the biological activity, especially with respect to human body. The main requirements that should meet such procedures are their physical validity, universal character, pictorial quality of results obtained, and economical feasibility. Taking into account these requirements we attempted to construct a prediction algorithm based on the actual three-dimensional structure of organic molecule with consideration of maximum possible physicochemical descriptors characterizing the given molecule. In order to decrease as much as possible the cost of the prediction the parameters used were obtained by the modern theoretical methods which at the same time ensure a sufficient accuracy of results. A single scheme is known now for forecasting the biological activity, PASS [1]. It is based on the analysis of the frequency of appearance of certain sequences of atoms and functional groups in two-dimensional molecular graphs.

The learning set used for prediction of the biological properties of new molecules must include only medicinals thoroughly studied and actually used in the medical practice. The careful selection has revealed that in the contemporary medical practice with a sufficient statistical weight are used a little bit over a thousand individual organic compounds (molecules). The other drugs are their combinations or the names are synonyms, and therefore the overall number of pharmaceuticals according different estimations is from 8 to 12 thousand.

The molecular dimensions no more are a stumbling block for precise calculations; moreover, admittedly the success in the use of chemical theory for understanding of mechanism of action on molecular level is obvious [2, 3]. Therefore the main stage in constructing the forecasting algorithm MATRIX is the calculation of descriptors for 1067 chosen molecules with the use of modern quantum-chemical and topological methods. Therefore the given name of the algorithm is MATRIX, meaning a matrix "molecular descriptors–activity".

The formulation of the principle (algorithm) proper for prediction is the second stage. Two reasonable approaches lead to the goal: the first one consists in a known procedure of finding the distance between points or data files in the *n*-dimensional space (claster analysis); the second one is more flexible and accurate; it is combined, and we present it here in an original modification.

Anzali *et al.* [1] developed formerly a program for prediction of a spectrum of pharmacological activity that was not species-specific. The program used derivatives of proximity matrix of atoms in a molecule (molecular graph). However the descriptors derived by three-dimensional simulation of molecules a fortiori active with respect to human body with

1070-4280/02/3811-1564$27.00 © 2002 MAIK "Nauka/Interperiodica"

accounting for the hydration effect are more accurately determined and have a clear physical meaning. The latter provides a possibility to describe the results in terms understandable for a chemist and a pharmacologist.

The most efficient future forecasting system would be not a single algorithm but a combination of several methods or their successive application (similarly to the procedures used for proving the chemical structure).

The information on the chemical structure and pharmacological activity of 1072 modern medicinals was taken from handbooks [4, 5, 6], and the refining was performed by original publications. We did not include in the selection the biologically active substances that were not used as pharmaceuticals due to two reasons: the action of the drug should be species-specific with respect to human body, and also the drugs permitted for use show variety of activity and side effects, therefore one molecule may be utilized in prediction of several kinds of actions.

In selecting compounds for the learning database we took into account that the time of quantum-chemical calculations grew according geometric progression with increase of atoms number in the system, and after certain number the self-consistent solution became impossible (with commonly used double accuracy of the calculation). In this connection we included in the original set of compounds those containing in the molecule no more than 100 non-hydrogen atoms or with overall number of atoms no more than 200. Parameters of polymer molecules (insulin, blood substitutes etc.) were calculated for short fragments possessing most of the properties of the original molecules save the macroscopic characteristics (viscosity, critical points etc.). The latter were not used in the algorithm of forecasting.

Physicochemical calculation procedures and generation of molecular descriptors. The optimization of molecular geometry of medicinal substances was performed by iteration search of a local (in a lucky case, of a global) minimum on a potential energy surface (PES). The localization of the minimum on PES was carried out along two main calculation procedures: molecular mechanics (preliminary) and molecular orbitals method (MO).

The primary geometry optimization was performed by molecular mechanics procedure MMP2. Since this method is time-saving computation it was generally used for primary analysis of comformational occupancy (for selection of a conformer of the probable lowest potential energy).

As starting geometry for MO methods was applied that obtained in the field of force of MMP2. On the one hand the geometry optimization for a molecule of medium size (40 heavy atoms) in the basis of 6-31G** level takes months of computation time even at the use of a claster of four computers (900 MHz, 512 Mb). On the other hand, the molecular mechanics method provides at optimization the closer to real geometry as compared to semiempirical and *ab initio* (STO-3G) simulations [7, 10]. Therefore in calculation of physicochemical descriptors by *ab initio* method we utilized the geometry found by MMP2.

The geometry obtained by optimization along AM1 procedure was further used in calculating physicochemical descriptors by semiempirical methods (save PM3 where the geometry optimization was carried out independently), and for calculation of descriptors of molecular structure (AM1).

An important feature of this work was the fact that the geometry optimization of molecules by AM1 and 6-31G methods was carried out both for vacuum and with accounting for hydration effects at physiological temperature. In the latter case we used the polarizable continuum model (PCM) [7–13]. The geometry optimization by PM3 method was performed without considering the hydration effects. The list of the main physicochemical descriptors is given in Table 1.

The overall number of descriptors obtained by MM and MO method equaled 53. The heats of hydration were calculated by equation (1).

$$H_{hydr} = H_{fsolv} - H_{fvacuum.}, \quad (1)$$

where H_{hydr} is energy of hydration; H_{fsolv} is the heat of formation with accounting for hydration; $H_{fvacuum}$ is the heat of formation without the solvent.

The adiabatic ionization potential (I_p) was calculated using the data on heat of formation of the molecule in the ground state and of the corresponding cation (everywhere the hydration effect was taken in account). The calculation was done along equation (2).

$$I_p = H_{fcat} - H_{fsolv}, \quad (2)$$

where I_p is ionization potential; H_{fcat} is the heat of formation of the corresponding cation; H_{fsolv} is the heat of formation of the molecule of the substance.

The calculation of the adiabatic electron affinity (A_e) is performed analogously along equation (3).

$$A_e = H_{fan} - H_{fsolv}, \quad (3)$$

Table 1. Crucial physicochemical descriptors of organic molecules

Descriptor name	Calculation procedure
Adiabatic ionization potential	AM1
Adiabatic electron affinity	AM1
Alpha-polarizability	AM1
Beta-hyperpolarizability	AM1
Linear dimensions of molecule	AM1
Gamma-hyperpolarizability	AM1
Dipole moment	6-31G, AM1, CNDO, MINDO3, PM3
Molecular weight	AM1
Total energy	6-31G, AM1, CNDO, MINDO3, PM3
Volume of van der Waals molecular model	6-31G
Area of van der Waals molecular model	6-31G
Sterical energy	MMP2
Heat of formation "in a vacuum"	AM1, PM3, CNDO, MINDO3
Heat of formation with hydration considered	AM1
Heat of anion formation with hydration considered	AM1
Heat of cation formation with hydration considered	AM1
Torsional energy of strain	MMP2
Angular energy of strain	MMP2
Energy of van der Waals interactions	AM1
HOMO energy	6-31G, AM1, PM3, CNDO, MINDO3
LUMO energy	6-31G, AM1, PM3, CNDO, MINDO3
Energy of void formation in water	6-31G
Bond energy	MMP2
Hydration energy	6-31G, AM1
Energy of electrostatic interactions	MMP2
Energy of electrostatic interaction with water	6-31G

where A_e is electron affinity; $H_{f_{an}}$ is heat of formation of the corresponding anion; $H_{f_{solv}}$ is heat of formation of the molecule of the substance. All calculations were carried out taking into account the hydration effects.

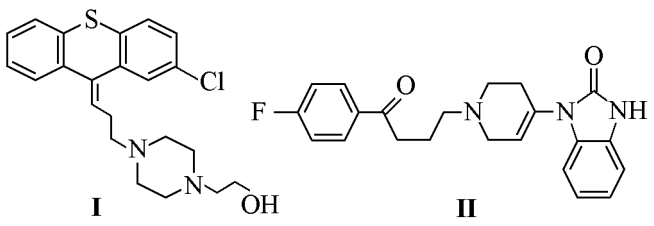
The use as descriptors of vertical ionization potentials (along Koopmans theorem, difference between HOMO and LUMO) gave as a rule worse results.

Topochemical descriptors. The name "topochemical descriptor" is used to distinguish logically the topological descriptors proper from those additive, geometric, and others [2]. The calculation methods for topological descriptors are outlined below. Into the prediction scheme have been included descriptors of the following groups (their overall number is indicated in parentheses): additive descriptors (11), topological descriptors (32), detour indices (10), geometric descriptors (6), 3D-MoRSE descriptors (10), WHIM descriptors (24), GETAWAY descriptors (3) [14–26]. To reduce the dimensionality of the data mass the descriptors having the same value for 10 and more molecules were excluded. The overall number of topochemical descriptors per one molecule amounted to 96.

Prediction of activity by claster analysis in the hyperspace of molecular descriptors. The prediction of relation "structure–activity" requires presentation of the molecular structure in a form invariant with respect to spatial position. This possibility is provided by utilization of descriptors of the molecular structure. It is possible by basing on this presentation to reveal the likeness of molecules and to construct models of "structure–activity" relation in the multidimensional space of molecular descriptors. On this foundation is built up a number of methods that permit forecasting biological activity of new compounds proceeding from the data on activity of the learning access.

The claster analysis allows selection of groups of substances most isolated in the descriptor space (clasters proper) [27–30]. Basing on the analysis results an assumption is advanced concerning the specific molecular activity mechanism characteristic of the molecules included in each cluster. The measuring of distances in the space of molecular descriptors provides an estimation of degree of likeness for the molecules. Thus the biological activity in the method of *k*-averages is forecasted by analysis of minimum distances of the compound under study from the molecules of the learning access in the common descriptor space.

Let us consider two medicinals from the same group of neuroleptics which have chemical structures



I **II**

in no way similar: zuclopentixol (a derivative of the thioxanthene, **I**), and droperidol (a derivative of butyrophenone, **II**) [4, 5].

The claster analysis by the method of *k*-averages [27–30] (the starting number of clusters 50, final 98) in the space of 149 descriptors placed both preparations in the same claster. Thus the clasterization of the physicochemical descriptors turned out to be an efficient selecting instrument. The calculation of the fine electronic structure in an extended basis demonstrated that the common element ensuring the similar activity was the localization of the highest occupied molecular orbital (Fig. 1).

However the prediction of the pharmacological activity by this method provides only qualitative data on the similarity of structures and disregards the quantitative characteristics of the activity. The claster analysis is founded on application of a matrix of distances in a metric descriptor space, and this notably limits the search for more sensitive models of the "structure–activity" relation.

Prediction of activity by algorithm MATRIX. From the viewpoint of the study of the quantitative relationships "structure–activity" the potential functions method [31] is more promising. We modified the potential function method to adjust it to the forecasting procedure (algorithm tentatively named MATRIX).

The numerical value of the pharmacological activities were approximated in the descriptor space by potential functions of (4) kind.

$$P_i = \sum_{i=1}^{m} k_j f_{ij}, \quad f_{ij} = f(r_{ij}), \tag{4}$$

$$r_{ij} = \sqrt{\sum_{c=1}^{\nu} W_c^2 (X_{ic} - X_{ic})^2}$$

Here P_i is the value of the potential function at the point of the *i*-th substance; k_j is the normalizing factor for the *j*-th distribution function; *m* is the number of distribution functions; *r* is the matrix of distances; *f* is the matrix of distribution function values; *W* is the vector of descriptors weights; *X* is the matrix of descriptors values; *f* is the distribution function; *i* is the substance number in the learning access; *j* is the substance number in the active subgroup of the learning access; *c* is the descriptor number; ν is the amount of descriptors.

It should be noted that the distribution functions are centered at the points corresponding to the active

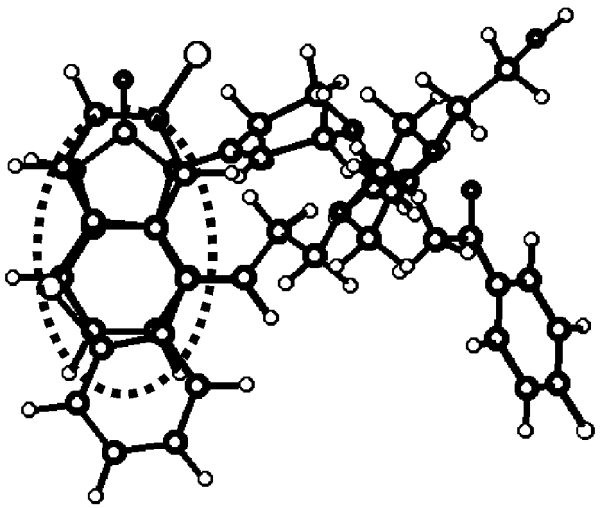


Fig. 1. Defining of the common active site of zuclopentixol and droperidol molecules.

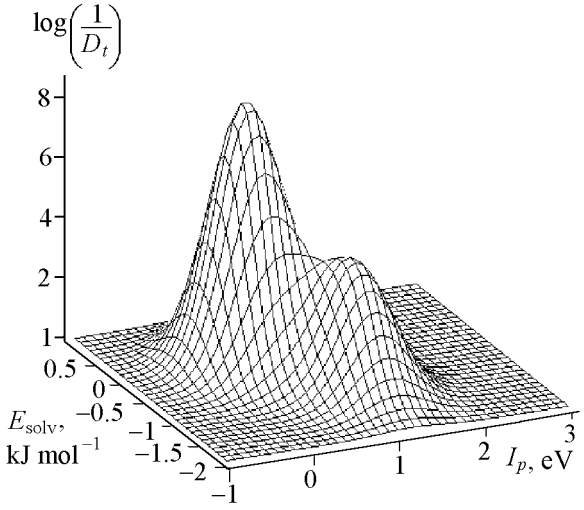


Fig. 2. Potential function surface for anticholinesterase activity approximating the Brigg's logarithm of the reciprocal average one-time therapeutic dose for an adult human of a mean weight (D_t).

substances of the learning access. We applied the Gauss function as distribution function. Function (5) is determined within the whole parameter space and takes values from 0 to 1. The function possesses a smoothing parameter that is subject to variation for the model optimization.

$$f(r) = e^{(-r^2/\sigma^2)} \tag{5}$$

Here *r* is a radial coordinate; σ is a smoothing parameter that characterizes the relative degree of

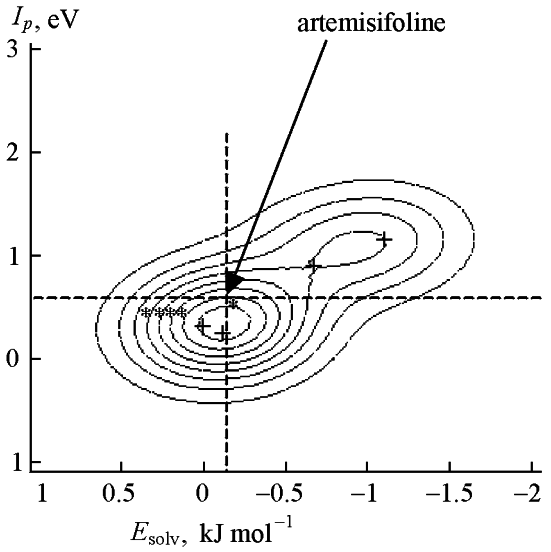


Fig. 3. Outline map (section) of a potential function proximating the anticholinesterase activity.

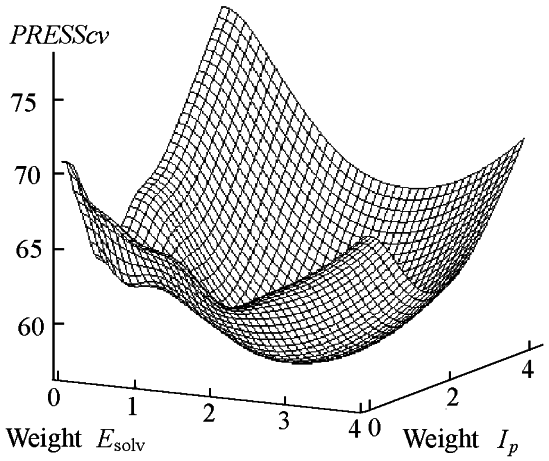


Fig. 4. Surface of dependence of *PRESSCv* on the weights of descriptors (hydration energy and adiabatic ionization potential).

scatter in the structural requirements for a given kind of activity.

The opportunity to apply different distribution functions with variation parameters therein supplies additional flexibility to the model. The option of including of sufficient number of parameters provides the algorithm with sufficient hospitality for investigating versatile kinds of activity. An example of a function of two variables usable as a potential function for predicting the pharmacological activity is given on Fig. 2. The function approximates the Brigg's logarithm of the reciprocal average one-time therapeutic dose for an adult human. The real potential function of this and all the other kinds of activity is naturally multidimensional.

The potential function is normalized for optimal reproducing of the quantitative values of pharmacological activity in the learning access. In this connection the prediction of the pharmacological activity is carried out by calculating the potential function value at the point of metric descriptor space corresponding to the molecule under study.

The final outcome of algorithm MATRIX operation is presented on Fig. 3: an outline map (section) of a potential function for anticholinesterase activity with positions of medicinals from the learning access and substances under study in the space of two descriptors (hydration energy and adiabatic ionization potential). As an example on Fig. 3 are indicated points of five sesquiterpene lactones separated from the plant ragweed (Ambrosia artemisiafolia) marked with an asterisk *, and of substances from the learning access with the closest activity values marked with a + sign. In the numerical presentation the forecast is a common logarithm of the reciprocal dose of this substance for an adult human (in grams).

For estimation of the quality of the model used were applied the standard error of prediction (S_{PRESS}) determined by procedure of the gliding inspection. The assumed model possesses a number of parameters: parameter of smoothing for the potential function (σ), and also statistical weights m of molecular structure descriptors (w_j, 1j = 1, …, m).

Since the error of prediction S_{PRESS} grows and decreases in keeping with the sum of squares of noncommunications in predictions (*PRESSCv*),the optimization of parameters has been directed to the minimization of *PRESSCv* (6).

$$PRESSCv = \sum_{i=1}^{n} \left[\left(\sum_{i=1}^{m} k_j f_{ij} \right) - A_i \right]^2, \qquad (6)$$

Here k_j is a normalization factor for j-th distribution function; m is the number of distribution functions; n is the number of preparations in the learning access; f is the matrix of distribution function values; i is the number of the preparation in the learning access; j is the number of the preparation in the active subgroup of the learning access; A_i is the pharmacological activity of the i-th preparation from the learning access.

An example of a surface of values *PRESSCv* in the space of the model parameters (weights of two descriptors) is given on Fig. 4. The surface has a

Table 2. Estimation of prediction error for some kinds of pharmacological activity (gliding inspection, 1067 molecules)

Pharmacological activity	S_{PRESS}	Relative error of prediction, %
Agonist 5-HT1 of receptors	0.137	3.367
Agonist of α_1-adrenoreceptors	0.346	5.332
Agonist of α_2-adrenoreceptors	0.402	5.228
Agonist of β_1-adrenoreceptors	0.372	5.499
Agonist of β_2-adrenoreceptors	0.637	8.319
Agonist of barbituric receptors	0.267	8.973
Agonist of benzodiazepine receptors	0.158	3.287
Agonist of GABA-receptors	0.059	4.763
Agonist of M-cholinoreceptors	0.284	5.500
Agonist of prostanoid EP-receptors	0.062	0.647
Activator of potassium channels	0.067	1.628
Activity of anabolic steroids	0.161	4.320
Androgenic activity	0.088	3.033
Antagonist of 5-HT_1 receptors	0.131	2.590
Antagonist of 5-HT_3 receptors	0.121	2.115
Antagonist of D_2-receptors	0.411	9.400
Antagonist of H_1-receptors	0.361	8.317
Antagonist of H_2-receptors	0.071	3.304
Antagonist of M-cholinoreceptors	0.440	10.289
Antagonist of α_1-adrenoreceptors	0.299	6.264
Antagonist of α_2-adrenoreceptors	0.136	2.745
Antagonist of β_1-adrenoreceptors	0.428	10.574
Antagonist of β_2-adrenoreceptors	0.388	9.198
Antagonist of H-cholinoreceptors	0.128	3.285
Antagonists of AT_1-receptors	0.191	5.678
Antiandrogenic activity	0.079	2.526
Antivitamin of folic acid	0.063	7.000
Antimetabolite, pyrimidines analog	0.025	3.619
Antimetabolite, purines analog	0.159	5.120
Blockader of calcium channels	0.141	3.443
Blockader of sodium channels in neuron membranes	0.294	12.668
Gestagenic activity	0.298	6.187
Glucocorticoid activity	0.148	2.642
Inhibitor of ACPase	0.307	6.236
Inhibitor of ergosterol biosynthesis	0.054	3.878
Inhibitor of histamine liberation	0.227	4.884
Inhibitor of GMG-CoA-reductase	0.019	0.467
Inhibitor of MAO	0.231	6.553
Inhibitor of reverse neuronal capture of monoamines	0.182	5.534
Inhibitor of reverse neuronal capture of serotonin	0.113	3.511
Inhibitor of potassium reabsorption in renal ductuli	0.269	6.462
Inhibitor of sodium reabsorption in renal ductuli	0.124	2.734
Inhibitor of cell wall synthesis in bacteria	0.153	26.203
Inhibitor of tubulin	0.018	0.299
Inhibitor of phosphodiesterase	0.190	8.503
Inhibitor of cholinesterase	0.157	3.081
Inhibitor of COG-1	0.261	14.237
Inhibitor of COG-2	0.204	10.863
Mineralocorticoid activity	0.017	0.572
Photosensibilizator	0.037	0.853
Estrogenic activity	0.179	3.148
Mean error of prediction for 255 activities	0.198	5.585

Table 3. Normalized values of molecular structure descriptors whose weights after optimization were larger than zero for sesquiterpene lactones and a series of acetylcholinesterase inhibitors (optimization of parameters for the model of anticholinesterase activity, σ 3.785)

Molecule	Heat of formation without hydration effect ($MINDO_3$) kcal mol^{-1}	Total energy (6-31G**), kcal mol^{-1}	Height of molecule with regard to hydration, Å (geometry of AM1)	Index of Zagreb group M2	Valence index of bonding	Sum of molecular detour of 4-th order	Sum of reverse detour of 10-th order	Kir's index	3D-MoRSE-signal 2/ weighed by van der Waals atomic volumes	Overall index of size V/weighed by van der Waals atomic volumes
Artemisifoline	–0.053	0.594	–0.189	–0.366	–0.217	–0.347	–0.599	–0.478	–1.074	–0.690
Peruvin	–0.055	0.434	0.197	0.335	0.317	0.485	0.588	–0.335	–0.790	–0.716
Dihydrocumanin	–0.057	0.429	0.253	0.238	0.566	0.370	0.610	1.918	–0.368	–0.629
Cumanin	–0.056	0.432	0.226	0.238	0.418	0.379	0.610	0.482	–0.564	–0.646
Psylostachyine	–0.056	0.434	0.153	0.160	0.202	0.282	0.610	0.349	–0.782	–0.679
Amiridin	–0.051	1.089	–0.833	–0.600	–0.566	–0.542	–0.642	–0.701	–0.283	–0.855
Deoxypeganine	–0.049	1.175	–1.111	–0.776	–0.875	–0.737	–0.663	–0.776	–0.529	–0.922
Psylostachyine	–0.051	0.405	–0.188	0.257	0.053	0.379	0.610	–0.160	–0.235	–0.384
Pyridostigmine	–0.045	1.016	–0.703	–1.185	–1.167	–1.180	–0.685	–0.719	–0.918	0.853
Stephaglabrine	–0.051	0.236	–0.008	0.628	0.126	0.760	0.631	–0.235	–0.476	–0.395
Tacrine	–0.049	1.010	–0.973	–0.522	–0.673	–0.480	–0.642	–0.675	–0.238	–0.845
Velnacrine	–0.051	0.860	–0.680	–0.366	–0.618	–0.312	–0.642	–0.642	–0.248	–0.817
Weights of descriptors	658.032	6.103	5.494	1.476	4.604	0.780	8.805	5.243	4.647	24.606

Table 4. Normalized values of molecular structure descriptors whose weights were larger than zero for sesquiterpene lactones and a number of officinal blockaders of M-cholinoreceptors (optimization of parameters for the model of M-cholinoblockading activity, σ 5.715)

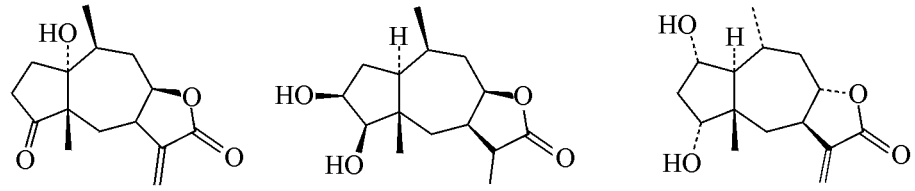
Molecule	Heat of formation without hydration effect ($MINDO_3$), kcal mol^{-1}	Energy of electrostatic interaction with water ($6–31G^{**}$), kcal mol^{-1}	Adiabatic ionization potential with regard to hydration (AM1),	Width of molecule with regard to hydration, Å (AM1 geometry)	Alpha-polarizability, cm^3	Beta-hyperpolarizability, cm^5	Centric Balaban's index	Deviation of eccentricity M1	Index of Zagreb group	Sum of reverse detour of 10-th order	Kir's index
Artemisifoline	–0.053	0.053	0.078	0.074	–0.515	–0.477	–0.157	–0.866	–0.495	–0.599	–0.478
Peruvin	–0.055	0.052	1.631	–0.167	–0.562	–0.600	0.251	–0.754	–0.254	0.588	–0.335
Dihydrocumanin	–0.057	0.052	1.974	–0.009	–0.586	–0.548	0.251	–0.708	–0.415	0.610	1.918
Cumanin	–0.056	0.052	1.552	–0.012	–0.536	–0.413	0.251	–0.710	–0.335	0.610	0.482
Psylostachyine	–0.056	0.052	1.635	–0.282	–0.472	–0.636	–0.157	–0.747	–0.254	0.610	0.349
Atropine	–0.053	0.053	–0.931	–0.404	–0.129	–0.656	–0.712	–0.234	–0.093	–0.578	1.176
Homatropine	–0.053	0.053	0.802	–0.848	–0.266	–0.644	–0.749	–0.340	–0.174	–0.599	–0.506
Scopolamine	–0.054	0.052	–0.475	0.573	–0.140	–0.580	–0.712	–0.099	0.228	0.631	–0.189
Ipratropium	–0.044	0.052	0.757	–0.355	0.304	–0.433	–0.008	0.216	0.108	0.631	–0.124
Troventol	–0.044	0.052	0.747	0.086	0.310	–0.518	–0.008	0.036	0.108	0.631	–0.224
Tropicamide	–0.050	0.053	0.690	0.062	0.038	–0.415	–0.601	–0.128	–0.013	–0.599	–0.023
Cyclozyl	–0.053	–0.626	0.727	0.918	–0.104	–0.596	–0.305	–0.003	–0.174	–0.578	–0.206
Weights of descriptors	0.587	0.379	0.367	6.565	2.559	6.404	4.751	21.069	30.467	9.952	12.675

pronounced minimum corresponding to the optimal value of parameters. The minimum point was determined in the space of parameters by the standard iteration procedure based on method of tangents.

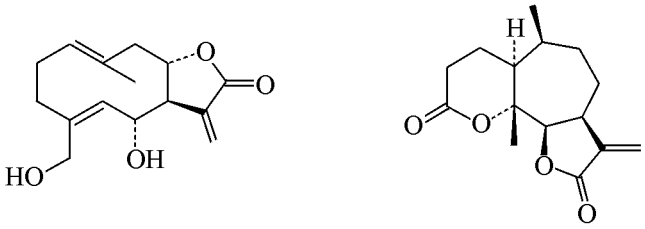
Comparative analysis of models concerning various kinds of pharmacological activity. Most effects of the modern pharmaceuticals result from action on certain receptor groups or enzymes of ionic channels in the human body. Proceeding from the method of potential functions we constructed model of pharmacological activity related to the effect on the main organs and systems of the human body. In Table 2 is illustrated the efficiency of prediction of various biochemical operation mechanisms with respect to human body and also with respect to microorganisms for the most common pharmaceuticals.

The prediction of biological activity for a number of new organic molecules by the developed method is in good agreement with experimental data. As a small illustration we give below the data on coincidence of activity forecasted along the developed procedure and experimentally determined for new molecules and known substances prepared at Pyatigorsk State Pharmaceutical Academy.

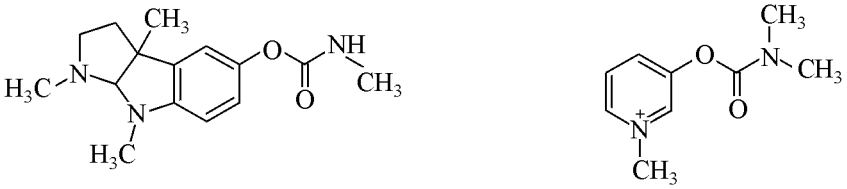
An example of forecast. To show the possibility provided by MATRIX algorithm we give the forecast examples of anticholinesterase activity, ability to block M-cholinoreceptors and D_2-receptors of dopamine for



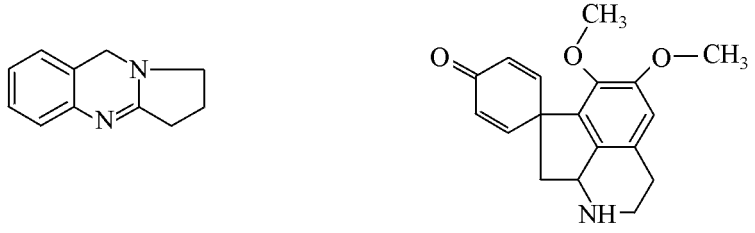
peruvin dihydrocumanin Cumanin



artemisifoline psylostachyine



physostigmine pyridostigmine



deoxypeganine stephaglabrine

a series of known drugs and sesquiterpene lactones from the plant ragweed (Ambrosia artemisiafolia).

At the first stage geometry optimization was carried out by MMP2 procedure followed by quantum-chemical calculations of molecules along a series of semiempirical and *ab initio* methods (GAMESS 6.0 taking around 2 h of machine time at the use of a claster of four PC of 833 MHz.

Table 5. Normalized values of molecular structure descriptors whose weights after optimization were larger than zero for sesquiterpene lactones and a series of neuroleptic medicinals with dopamine-blocking activity (optimization of parameters for the model of dopamine-blocking activity σ 5.982)

Molecule	Energy of electrostatic interactions (MM2), kcal mol^{-1}	Energy of electrostatic interaction with water (6-31G**), kcal mol^{-1}	Energy of HOMO disregarding dissolution (CNDO), eV	Energy of HOMO potential with regard to dissolution (AM1), eV	Vertical ionization with regard to dissolution (AM1), eV	Adiabatic ionization potential with regard to dissolution (AM1), kcal mol^{-1}	Alpha-polarizability, cm^3
Artemisifoline	0.273	0.053	–0.179	–0.331	–0.031	0.078	–0.515
Peruvin	0.498	0.052	–0.219	–0.329	1.472	1.631	–0.562
Dihydrocumanin	0.498	0.052	–0.405	–0.325	1.915	1.974	–0.587
Cumanin	0.273	0.052	–0.369	–0.286	1.652	1.552	–0.537
Psylostachyine	0.289	0.052	–0.300	–0.372	1.763	1.635	–0.473
Galanthamine	0.598	0.053	0.202	–0.143	–1.718	–1.648	1.139
Pyridostigmine	0.500	0.053	0.204	–0.239	–1.638	–1.670	0.415
Stephaglabrine	0.599	0.053	0.197	–0.232	–1.639	–1.672	0.456
Cyclozyl	0.504	0.053	0.236	–0.301	–1.797	–1.485	0.638
Tropicamide	0.389	0.052	0.189	–0.195	–0.824	–0.641	0.996
Troventol	0.271	0.053	0.164	–0.227	–0.712	–0.516	0.983
Aminazine	0.404	0.053	–0.163	–0.201	–0.124	–0.992	0.670
Chlorprothixene	0.375	0.052	–0.221	–0.204	–0.036	–0.900	0.652
Benperidol	–0.742	0.052	–0.038	0.075	–0.596	–0.310	1.759
Droperidol	0.203	0.052	–0.203	–0.149	–0.077	0.388	1.666
Haloperidol	0.455	0.052	0.177	–0.157	–0.822	–0.634	1.769
Trifluperidol	0.340	0.053	0.186	–0.270	–0.897	–0.661	0.713
Pimoside	0.723	0.053	0.079	–0.423	–0.952	–1.206	–2.729
Tiapride	–2.851	0.053	–0.130	–0.247	–0.180	–0.971	0.133
Sultopride	–0.760	0.052	–0.112	–0.408	–0.112	–1.242	0.163
Weights of descriptors	10.763	23.861	28.012	28.271	6.835	9.456	2.174

We performed a preliminary optimization of parameters for mathematical simulation of various kinds of biological activity with the goal of minimization of *PRESSc*v values. The estimation of prediction quality was done by gliding inspection method. The normalized values of molecules' descriptors, the weights for the descriptors, and smoothing parameters obtained by optimization are presented in Tables 3–5.

We obtained optimum parametric spaces of descriptors for various kinds of pharmacological activity, and data on three molecular mechanisms of action furnish illustrative examples. As seen from Table 3, the most descriptors in the optimum set for anticholinesterase activity correspond to the topological group [32] meaning that these characteristics are especially important for operation of this kind activity. With M-cholinoblocking activity (Table 4) the optimum set contains both topochemical and quantum-chemical descriptor. The descriptor set obtained for dopamine-blocking activity (Table 5) is fully comprised of physicochemical descriptors. Among the latter are the adiabatic and vertical ionization potentials, and also HOMO energy with and without the hydration effect. Apparently the electronic characteristics of molecular structure play an essential role in dopamine-blockaders' operation. Both for anticholinesterase and M-cholinoblocking activities sufficiently informative among the energetic characteristics was heat of formation calculated by MINDO/3 procedure. Interestingly the energy of electrostatic interaction with water (6-31G**) turned out to be common for the anticholinesterase and M-cholinoblocking activities.

The prediction of three kinds of pharmacological activity for compounds from the control access was

Tabke 6. Prediction of anticholinesterase activity, ability to block muscarine receptors and D_2-receptors of dopamine for sesquiterpene lactones and a series of pharmaceuticals (statistically meaningful values are in bold type)

Molecule	Real presence of activity			Prediction of activity		
	antagonist of D_2-receptors	Antagonist of M-cholinoreceptors	Inhibitor of cholinesterase	Antagonist of D_2-receptors	Antagonist of M-cholinoreceptors	Inhibitor of cholinesterase
Artemisifoline				0.002	0.006	0.157
Peruvin				0.002	0.003	0.064
Dihydrocumanin				0.002	0.003	0.007
Cumanin				0.002	0.003	0.080
Psylostachyine				0.002	0.003	0.062
Galanthamine[a]			+	0.000	0.007	**0.891**
Pyridostigmine			+	0.002	0.003	**0.539**
Stephaglabrine			+	0.005	0.009	**1.000**
Cyclozyl		+		0.000	**0.446**	0.006
Tropicamide		+		0.002	**0.050**	0.007
Troventol		+		0.000	**0.119**	0.007
Aminazine	+	+		**1.000**	**1.000**	0.006
Chlorprothixene	+	+		**0.844**	**0.578**	0.006
Benperidol	+			**0.117**	0.001	0.002
Droperidol	+			**0.063**	0.001	0.000
Haloperidol	+			**0.019**	0.003	0.006
Trifluperidol	+			0.009	0.000	0.000
Pimoside	+			**0.116**	0.000	0.000
Tiapride	+			0.005	0.004	0.005
Sultopride	+			0.008	0.003	0.002

[a] The known (officially used) medicinals are typed in italics.

carried out, and the results are given in Table 6. The values cited in Table 3 were obtained using optimal set of parameters corresponding to the given activity.

We forecasted the spectrum of pharmacological activity of lactones basing on the data of mean therapeutical doses of medicinals from the learning access. The results of prediction are presented as logarithms of reciprocal doses. Note that this prediction is valid for an adult person of mean weight, and it is easily recalculated into the units common for medical practice (mg kg^{-1}).

The forecasted biological activity of artemisifoline shows its low ability to activate barbituric receptors. The prediction with regard to psylostachyine indicates its possible capability to selective inhibition of cyclooxygenase-2. Therefore an antiphlogistic activity may be found in this substance. The biological tests showed that the diastolic pressure was reduced at intravenous infusion of terpenes from ragweed. The experimental data were in agreement with the performed forecast (the experiments were carried out at the Chair of pharmacology of the Pyatigorsk State Pharmaceutical Academy, head of the Chair Professor, Doctor of Medical Sciences M. N. Ivashev).

Thus or in the similar way the results of prediction can significantly help in the search for molecules as candidates for drugs. The prediction regards not only the desired activity but also the unwanted side effects: toxicity, embryotoxicity, carcinogenicity etc.

The checking of the learning data mass of drug molecules by the method of gliding inspection (when descriptors of each molecule from the access are one by one excluded from it, and the forecasting of activity for this molecule is carried out as if it be an unknown substance) has demonstrated that the correlation factor "prediction–experiment" for the algorithm developed amounts to 85.6%.

The described algorithm of predicting the pharmacological activity of organic molecules with respect to the human body provides a possibility to perform a purposeful search for biologically active compounds reducing to minimum the expensive and labor-consuming experimental investigations. The method can increase the productivity of laboratories

developing new drugs of the natural and synthetic origin, since it can show the direction of a target organic synthesis. The MATRIX algorithm permits forecast of the most existing kinds of activity and to consider also the probable side or toxic effect of the substances under study.

The method suggests disclosure of the molecular operation mechanism for it primarily is directed to forecasting mechanism of action and not the secondary biological effect.

The revealing of hidden physicochemical analogies, i.e., those not following directly from the similar elements in the chemical formulas, provides a possibility to explain quite a number of observations on similar pharmacokinetic and pharmacodynamic behavior of compounds differing in the chemical structure.

The authors are grateful to the staff of the laboratory of quantum chemistry of the Research Institute of Physical and Organic Chemistry at Rostov University, and personally to Professor, Doctor of Chemical Sciences R. M. Minyaev for long-standing consulting during performance of this work.

REFERENCES

1. Anzali, S., Barnickel, G., Cezanne, B., Krug, M., Filimonov, D., and Poroikov, V., *J. Med. Chem.*, 2001, vol. 15, p. 2432.
2. Raevskii, O.A., *Usp. Khim.*, 1999, vol. 68, no. 6, p. 555.
3. Karelson, M., Lobanov, V.S., and Katritzky, A.R., *Chem. Rev.*, 1996, vol. 96, p. 1027.
4. Mashkovskii, M.D., *Lekarstvennye sredstva* (Drugs), vol. 1, Moscow: Novaya Volna, 2000.
5. Mashkovskii, M.D., *Lekarstvennye sredstva* (Drugs), vol. 2, Moscow: Novaya Volna, 2000.
6. Doson, R., Elliot, D., Elliot, U., and Jhons, K., *Spravochnik biokhimika* (Handbook for Biochemist), Moscow: Mir, 1991.
7. Liotard, D.A., Hawkins, G.D., Lynch, G.C., Cramer, C.J., and Truhlar, D.G., *Report UMSI 94/143*, University of Minnesota Supercomputer Institute, 1994.
8. Olivares del Valle, F.J., Aguilar, M.A., and Tolosa, S., *J. Mol. Struct.*, 1993, no. 279, p. 233.
9. Tomasi, J., Bonaccorsi, R., Cammi, R., Olivares del Valle, F.J., *J. Mol. Struct.*, 1991, no. 234, p. 401.
10. Rzepa, H.S., Yi, M.Y., Karelson, M.M., and Zerner, M.C., *J. Chem. Soc., Perkin Trans.*, 1991, no. 2, p. 635.
11. Tapia, O. and Goscinski, O., *Molecular Physics*, 1975, no. 29, p. 1653.
12. Bottcher, C.J.F., *Theory Of Electrostatic Polarisation*, 1973, no. 105.
13. Scholte T.G., *Physics*, 1949, no. XV, p. 473.
14. Schuur, J. and Gasteiger, J., *Anal. Chem.*, 1997, vol. 83, pp. 2398–2405.
15. Kaliszan, R., Osmialowski, K., Tomellini, S.A., and Hsu, S.H., *Chromatographia,*, 1985, vol. 20, pp. 705–708.
16. Broto, P., Moreau, G., and Vandicke, C., *Eur. J. Med. Chem.*, 1984, vol. 19, pp. 66–70.
17. Galvez, J., Garcia, R., Salabert, M.T., and Soler, R., *J. Chem. Inf. Comp. Sci.*, 1994, vol. 34, pp. 520–525.
18. Randic, M., *J. Chem. Inf. Comp. Sci.*, 1995, vol. 35, pp. 372–382.
19. Todeschini, R., Lasagni, M., and Marengo, E., *J. Chemom.*, 1994, vol. 8, pp. 263–273.
20. Jug, K., *J. Org. Chem.*, 1983, vol. 48, p. 1344.
21. Clare, B.W. and Supuran, C.T., *J. Pharm. Sci.*, 1994, vol. 20, pp. 768–773.
22. Konstantinova, E.V., *J. Chem. Inf. Comp. Sci.*, 1997, vol. 36, pp. 54–57.
23. Rucker, G. and Rucker, C., *J. Chem. Inf. Comp. Sci.*, 1993, vol. 33, pp. 683–695.
24. Todeschini, R. and Gramatica, P., *QSAR 1998*, no. 16, pp. 113–119.
25. Burden, F.R., *QSAR*, 1997, no. 16, pp. 309–314.
26. Kruszewski, J. and Krygowski, T.M., *Tetrahedron Lett.*, 1972, vol. 36, pp. 3839–3842.
27. Hartigan, J.A., *Clustering algorithms*, New York: Wiley, 1975, p. 32.
28. Hartigan, J.A. and Wong, M.A., *Applied Statistics*, 1978, vol. 28, p. 100.
29. Tryon, R.C., *Cluster Analysis. Ann Arbor*, MI: Edwards Brothers, 1939.
30. Zupan, J., *Clustering of large data sets*, New York: Research Studies Press, 1982.
31. Tetko, I.V., Tanchuk, V.Yu., Vasil'ev, S.A., Khilya, V.P., Poda, G.I., and Luik, A.I., *Bioorganicheskaya khimiya*, 1995, vol. 21, no. 10, p. 809.
32. Stankevich, M.I., Stankevich, I.V., and Zefirov, N.S., *Usp. Khim.*, 1988, vol. 57, no. 3, p. 337.