One-class approach: models for virtual screening of non-nucleoside HIV-1 reverse transcriptase inhibitors based on the concept of continuous molecular fields*

P. V. Karpov,[a] *I. I. Baskin,*[a★] *N. I. Zhokhova,*[a] *M. B. Nawrozkij,*[b] *A. N. Zefirov,*[a]
A. S. Yablokov,[b] *I. A. Novakov,*[b] *and N. S. Zefirov*[a]

[a]*Department of Chemistry, M. V. Lomonosov Moscow State University,*
Build. 3, 1 Leninskie Gory, 119991 Moscow, Russian Federation.
Fax: +7 (495) 939 0290. E-mail: igbaskin@gmail.com; zhokhovann@gmail.com
[b]*Volgograd State Technical University,*
28 prosp. Lenina, 400131 Volgograd, Russian Federation.
Fax: +7 (8442) 23 8125. E-mail: kholstaedt@yandex.ru

One-class models for virtual screening of potent non-nucleoside HIV reverse transcriptase inhibitors were built for the first time in terms of the one-class approach using the support vector machine method. The training set included 786 structures of 2-substituted pyrimidinones and their inhibitory activity against the enzyme of wild-type and mutant (K103, IRLL98, Y188L) HIV-1 strains. The representation of molecular structures of organic ligands based on continuous molecular fields can be used to build classification models of higher quality compared to conventional approaches using Carhart fragment-based descriptors, molecular fingerprints, and spectrophores.

Key words: organic compounds, reverse transcriptase inhibitors, HIV, HIV-RT, one-class classification, virtual screening, continuous molecular fields, modeling of biological activity.

A search for novel highly efficient human immunodeficiency virus reverse transcriptase (HIV-RT) inhibitors is one of important problems of modern medicinal chemistry. Human immunodeficiency virus reverse transcriptase based on virion RNA synthesizes a complementary single-stranded DNA, which, after the second strand synthesis, is integrated into host DNA and damages host cells. The inhibition of the reverse transcription leads to the suppression of virus propagation, being an important stage of modern antiviral therapy. In the design of new lead compounds for anti-HIV drugs, researchers pay particular attention to a large group of organic compounds, which belongs to non-nucleoside HIV-RT inhibitors and includes various structural classes (HEPT,** DABO,*** *etc.*).[1,2] These compounds have low toxicity, high activity, and selectivity in the HIV-RT inhibition, and their use in complex highly active antiretroviral therapy substantially decreases the risk of AIDS death. However, despite the fact that the known compounds of this group cause a substantial decrease in the HIV replication rate, they do not fully suppress HIV, and the development of viral resistance raises the problem of the design of new more efficient non-nucleoside HIV-RT inhibitors with different resistance profiles.[3]

The virtual screening of libraries of organic ligands is an efficient tool in the search for and design of new lead compounds for drugs. One of the steps in the design of a virtual screening system involves the construction of models by 2D-3D QSAR* methods. Numerous regression QSAR models for the assessment of the inhibitory activity of narrow series of organic compounds against HIV-RT were described in the literature, and all these models have high statistical performance of the predictability of IC_{50} and EC_{50} ($q^2 \approx 0.8$).[4] However, these models do not allow one to perform virtual screening for new potent inhibitors in large ligand subclasses, while the use of special methods of combining highly specialized models[5] adds unnecessary difficulties and leads to a decrease in the accuracy of the prediction due to an additional classification error. As a rule, QSAR models built on large sets, which include different classes of organic structures, have much worse statistical performance leading to large prediction errors

* Dedicated to Academician of the Russian Academy of Sciences O. M. Nefedov on the occasion of his 80th birthday.

** HEPT is 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)-thymine.

*** DABO are dihydroalkoxybenzyloxopyrimidines.

* QSAR is the quantitative structure—activity relationship.

Published in Russian in *Izvestiya Akademii Nauk. Seriya Khimicheskaya,* No. 11, pp. 2371—2377, November, 2011.

and, consequently, the results of virtual screening are plagued by false positives.

A special problem of quantitative QSAR models is that it is difficult to correctly estimate their activity threshold, whose variation affects the set of compounds obtained by virtual screening. This may lead, as a result of the prediction error, to the loss of true active lead compounds.

An alternative approach, which allows researchers to avoid, in part, problems of regression QSAR models is based on the use of classification models for virtual screening. These models enable the qualitative prediction whether the compounds under examination would exhibit activity. In this case, the conclusion about the assignment of compounds to active or inactive structures is drawn with the use of special procedures for the assessment of whether the structure belongs to a particular modeled class. However, to correctly construct a two-class model, it is necessary to have at hand two representative sets for active and inactive compounds. Unfortunately, representative sets for inactive compounds are not always available in the case of the modeling of many types of biological activity.[6—8]

In the study,[9] we have employed a new approach to the construction of models for virtual screening based on the one-class classification, which requires only a set of active compounds. In chemoinformatics, the principle of the one-class classification is used for the determination of the applicability domains of QSAR/QSPR* models.[10,11]

In the present study, we used the one-class approach for the construction of models for virtual screening of potent non-nucleoside HIV-RT inhibitors. In the context of this problem, with the aim of finding the optimal way of building models, it was of interest to compare different representations of the structures of organic ligands (both with the use of conventional descriptors, such as molecular fingerprints, spectrophores,[12,13] and Carhart fragment-based descriptors,[14,15] and continuous molecular fields[16]).

Calculation procedure

The training set of organic ligands contained data on the inhibitory activity (EC_{50}) against HIV-RT of wild-type and mutant (K103, IRLL98, Y188L) HIV-1 strains for 786 compounds including derivatives of 2-substituted 6-arylmethyl)pyrimidin-4(3*H*)-ones (DABO). The molecular weights of compounds vary from 202 to 550 Da; the lipophilicity, from 3.4 to 8.0. The training set included both original data (supplied by the research group headed by M. B. Nawrozkij) and the data published in the literature.

To build models, the molecular structures of compounds were represented with the use of the following descriptors: molecular fingerprints, spectrophores, and Carhart descriptors. The first two types were calculated using standard functions of the freely distributed Open-Babel program.[17] Modified Carhart descriptors were calculated with the employment of the CARHART descriptor block developed by us earlier.[9] In this block, the principle of the generation of chain fragments according to Carhart, was combined with the scheme of encoding of the FRAGMENT block,[18] which takes into account the hybridization, the formal charge, the bonding environment, and the number of adjacent hydrogen atoms in the structure of the compound under consideration. This combination makes it possible to improve the quality of the description of the molecular structures.

One-class models based both on conventional descriptors and continuous molecular fields were built with the use of the MCMF* program package developed in our earlier study.[16] The support vector machine (1-SVM) method implemented in the LIBSVM** library was employed as the kernel-based machine learning method. The predictive ability of the models was assessed with the use of the cross-validation leave-one-out procedure. The quality of one-class models was estimated from the area under the receiver operating characteristic ROC*** curve.[19] To optimize the parameters of the support vector machine method, statistical kernels, and parameters of continuous molecular fields, we employed the empirical search for extrema of functions in specified ranges of changes in the parameters implemented in the NLopt library.****

The alignment of the organic structure database was performed with the use of the program, which we have designed based on the SEAL algorithm.[20]

Results and Discussion

In the construction of models with the use of the one-class classification, only active compounds are at one´s disposal and, consequently, the performance of the trained classifier can be assessed using the values for true positives (TP; active compounds predicted as active) or false negatives (FN, active compounds predicted as inactive). However, to assess the efficiency of the classifier using the area under the ROC curve, it is also necessary to calculate the values for false positives (FP, inactive compounds predicted as active) or true negatives (TN, inactive compounds predicted as inactive). The quality of the model is characterized by the following parameters: the sensitivity (TP/(TP + FN)) and the specificity (TN/(FP + TN)). In the present study, the library of so-called decoys from the DUD***** database[21] was used as negative examples of inhibitory activity against HIV-RT. The decoy set consisted

* QSPR is the quantitative structure—property relationship.

* MCMF is the method of continuous molecular fields.

** Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001; http://www.csie.ntu.edu.tw/~cjlin/libsvm.

*** ROC is the receiver operating characteristic.

**** http://ab-initio.mit.edu/nlopt.

***** DUD is the Directory of Useful Decoys.

of 1519 compounds, whose physicochemical properties are similar to those of the known HIV-RT ligands but which substantially differ from the latter in the chemical structure. It should be noted that the structures of decoys are used only for the calculations of the efficiency of the resulting one-class models as opposed to the two-class classification, where these structures are directly involved in the construction of the model.

Figure 1 shows the histogram characterizing the molecular similarity between the active ligands from the data set under consideration, as well as the ligands from the DUD test set for HIV-RT, and decoys from the DUD. The y axis represents the percentage of inactive ligands, for which the minimum values of the molecular similarity with active ligands fall in the range of changes in the Tanimoto coefficient (k_T)[22] shown on the x axis. As can be seen from Fig. 1, compounds belonging to the DABO class have a larger molecular similarity with decoys from the DUD data set than the ligands from the DUD test set for HIV-RT. Most of compounds from the set under study have the molecular similarity with the decoys of about ≤0.75, which characterizes this data set as a diverse one.

The model construction by the support vector machine method requires, in addition to the choice of the optimal parameters of statistical kernels, the optimization of the parameter ν, which limits the number of support vectors involved in the model construction. If the molecular structure is represented with the use of the combined molecular field as a linear combination of electrostatic, steric, and hydrophobic molecular fields, it is necessary to optimize seven parameters (mixing coefficients for three types of fields, three attenuation coefficients, and the parameter ν) for the model construction. This optimization was performed by maximizing the area under the ROC curve (AUC, area under curve). Figure 2 shows the scheme of the construction of one-class models using decoys. Each active compound was successively excluded from the training set, and its activity was predicted based on the model built with the use of the remaining structures. Then the model was constructed with the use of all activity examples, and predictions were made for decoys. The results of prediction were combined, and the area under the ROC curve (AUC) was calculated. The larger the AUC the better the classifier performance. The ideal classifier has AUC = 1.0.

We used the above-mentioned scheme to obtain a set of one-class models constructed with the employment of different approaches to the representation of molecular structures, such as conventional methods, including the modified version of Carhart fragment-based descriptors, molecular fingerprints, and spectrophores, as well as in terms of the method of continuous molecular fields, which we have developed earlier.[16] The conventional description of the structure in terms of the commonly used SAR/QSAR* methods implies the representation of the structure as a set of descriptors, *i.e.*, numbers describing particular characteristics of the compound under consideration.[23] Nowadays, fragment (substructure) descriptors[24]

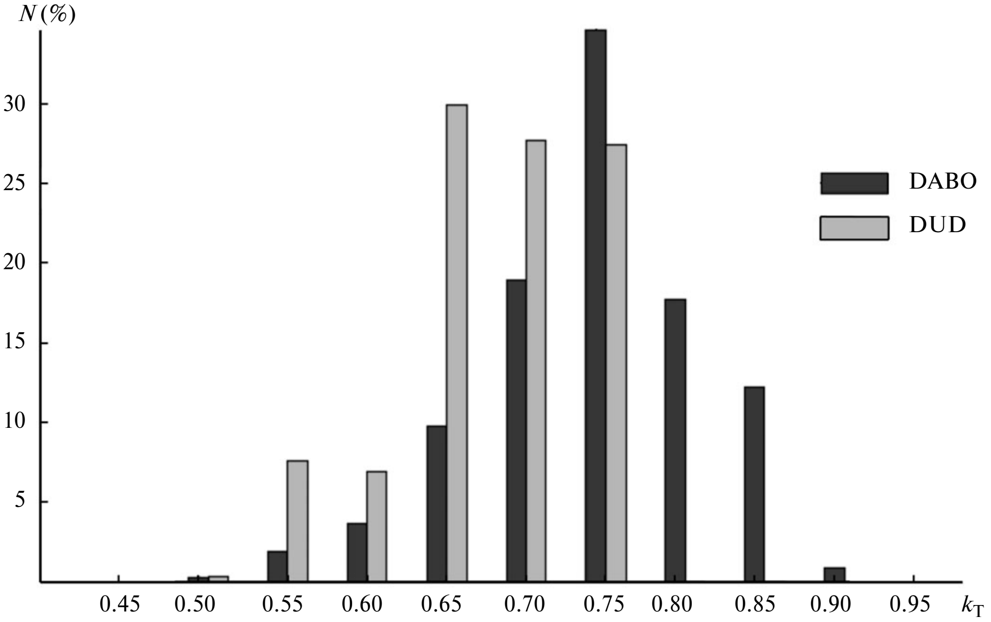

Fig. 1. Histogram of the molecular similarity between the active ligands from the data set under study, as well as the ligands from the DUD test set for HIV-RT, and decoys from the DUD data set (k_T is the Tanimoto coefficient and N (%) is the number of inactive molecules).

* SAR/QSAR is structure—activity relationships/quantitative structure—activity relationships.

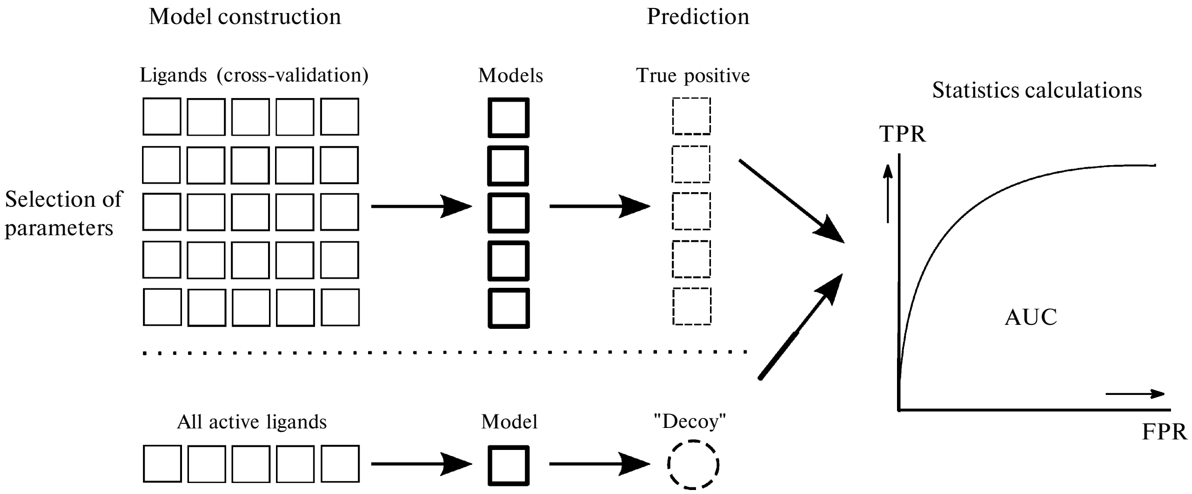

Fig. 2. Optimization scheme for the parameters of the one-class classifier using the fivefold sliding control and decoys; TPR is the true positive rate, and FPR is the false positive rate.

are widely used in the virtual screening. However, this representation has a number of substantial drawbacks, the main drawback being that it is impossible to suggest active compounds belonging to other structural types different from those used for the model construction. In this connection, non-descriptor-based methods for the structure description, *i.e.*, calculations of the modeled properties directly from the mathematical description of the structures of organic compounds, are extensively developed. For example, the structure is represented as a molecular graph, which serves as a basis for the construction of different molecular kernels.[25] We developed an alternative way of building kernels using continuous molecular fields.[16] This method was successfully applied to the construction of regression QSAR models for the prediction of biological activity, as well as to the building of one-class models for virtual screening. Thus, the use of this method for the construction of models using a standard DUD ligand set for the enzyme HIV-RT allows one to obtain models of higher quality compared to the simple similarity-based virtual screening.[26]

The method of the one-class classification, which we describe in the present study and which employs only samples of active ligands for the construction of prediction models, cannot be compared with two-class models, which are used primarily for virtual screening. In the study,[27] the DUD data set was investigated with the use of conventional similarity-based virtual screening employing different scoring functions to take into account 2D and 3D information. The similarity-based virtual screening takes into account only one active ligand (in the case under consideration, crystallographic), ranking all structures from the test set with respect to the active ligand. In essence, this procedure can be considered as the simplified one-class classification.

The statistical performance of the resulting one-class models of the ligands belonging to the DABO class are given in Table 1. The corresponding ROC curves are shown in Fig. 3. As can be seen from Table 1, the models constructed with the use of fragment-based descriptors (modified Carhart descriptors and molecular fingerprints) are characterized by high AUC values. For the model based on molecular fingerprints, the AUC value is 0.97 (the Tanimoto function as the kernel); for the model based on Carhart descriptors, 0.99 (the Gaussian kernel), which virtually corresponds to the ideal classifier. Table 1 illustrates the parameters of one-class models, which we built with the use of the standard set of HIV-RT ligands and decoys from the DUD data set. Due to large differences in the number of compounds in two sets (the standard DUD

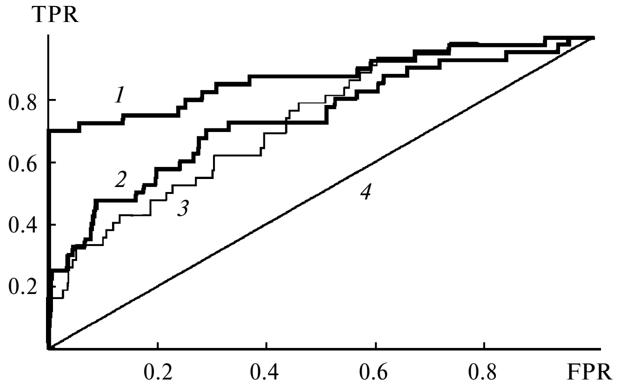

Fig. 3. ROC curves for one-class models for the standard DUD ligand and decoy sets for HIV-RT: *1*, the 1-SVM model, Gaussian kernel, spectrophore descriptors; *2*, the 1-SVM model, the steric molecular field as the description of structures; *3*, conventional similarity-based virtual screening using molecular fingerprints;[27] *4*, an arbitrary classifier.

Table 1. Statistical performance of one-class models constructed with the use of different representations of the structures for ligands belonging to the DABO class and the standard DUD data set for HIV-RT

Method	DABO ligands							Test set of ligands from the DUD data set for HIV-RT						
	AUC	TN	TP	FN	FP	Sensitivity	Specificity	AUC	TN	TP	FN	FP	Sensitivity	Specificity
Spectrophores	0.71	993	514	272	526	65.4	65.4	0.87	1100	31	9	345	76.1	77.5
Molecular fingerprints	0.98	1477	765	21	42	97.2	97.3	0.83	1157	32	8	288	80.1	80.8
Carhart descriptors	0.99	1489	770	16	30	98.0	98.0	0.80	1511	24	16	8	99.5	62.8
Electrostatic field	0.99	1505	779	7	14	99.1	99.1	0.60	831	23	17	614	57.5	57.5
Hydrophobic field	1.00	1519	786	0	0	100.0	100.0	0.75	1012	28	12	433	70.0	70.0
Steric field	1.00	1519	786	0	0	100.0	100.0	0.65	851	24	16	594	58.9	60.0

data set contains 40 ligands, whereas the data set under consideration includes 786 ligands), as well as in their diversity, it is impossible to correctly compare the results obtained with the use of these data sets. Nevertheless, the comparison of the ROC curves (see Fig. 3), which were constructed from the results of the prediction with the use of the standard DUD ligand set for HIV-RT employing the 1-SVM model based on Carhart descriptors, as well as employing the 1-SVM model combined with the steric molecular field for the description of molecular structures, and the data reported in the study[27] shows that the method of the one-class classification developed in the present study is more efficient than the simple similarity-based virtual screening based using Tanimoto indices. In the case of the steric field, approximately equal AUC for classifiers are observed. Thus, the spectrophore-based model (curve *1*) is characterized by the largest area under the ROC curve. Although the use of the steric molecular field for the description of ligand structures (curve *2*) is as efficient as the conventional similarity-based virtual screening (curve *3*), the former curve is located much higher in the initial step, which is of interest in the practical application of virtual screening. This is indicative of the higher efficiency of the 1-SVM model combined with the steric molecular field compared to the conventional similarity-based virtual screening in the most important early steps of the virtual screening. In both cases, the described method of the one-class classification is characterized by the larger area under the ROC curve.

As mentioned above, the use of the substructure description limits the structure set obtained by virtual screening to classes of compounds used for the model construction. Theoretically, this drawback can be avoided with the use of such descriptors as spectrophores, which quantify the surface field values at points of the cell artificially built around the ligand. However, in this study, the spectrophore-based model is characterized by the relatively low value AUC = 0.76 (Gaussian kernel).

The statistical performance of the models constructed with the use of continuous molecular fields are better than those of the models constructed using modified Carhart descriptors (AUC = 0.99), molecular fingerprints (AUC = = 0.97), and spectrophores (AUC = 0.76). The models constructed with the use of hydrophobic (AUC = 1.00) and steric (AUC = 1.00) fields have the maximum areas under the ROC curves. Taking into account high statistical performance of these models, they can be proposed as a tool for virtual screening for potent HIV-RT inhibitors.

Figure 4 shows the fields of coefficients (the direction cosines of the perpendicular to the separating hyperplane

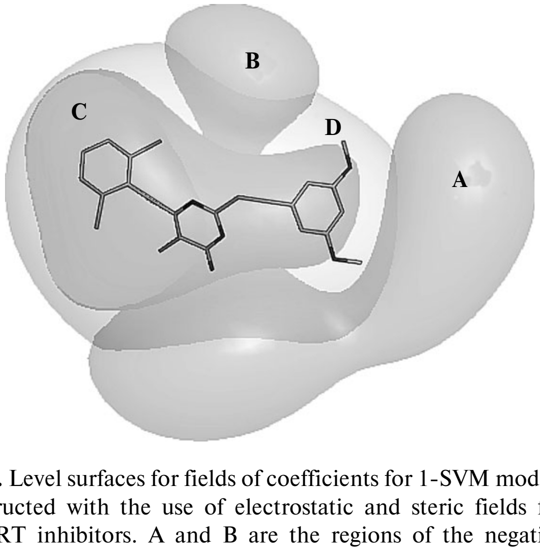

Fig. 4. Level surfaces for fields of coefficients for 1-SVM models constructed with the use of electrostatic and steric fields for HIV-RT inhibitors. A and B are the regions of the negative electrostatic potential, C is the region of the positive electrostatic potential, and D is the region of the steric potential determining the molecular shape.

in the feature space) for the 1-SVM models constructed for HIV-RT inhibitors with the use of electrostatic and steric fields. The fields of coefficients show the configuration of the molecular fields, which are necessary for a molecule of an organic compound to be the HIV-RT inhibitor.

The conventional molecular similarity-based virtual screening procedure involves the selection of a particular lead compound as the reference structure and the ranking of the tested structure data set with respect to this lead compound using Tanimoto indices. Formally, this method can be considered as a version of the one-class classifier. Hence, we compared the efficiency of the conventional similarity-based virtual screening with the one-class approach considered in the present study. For this purpose, we successively selected each structure from active ligands from the data set under consideration and calculated the molecular similarity for all other ligands and decoys based on this structure. As a result, we obtained a set of ROC curves, the areas under which provide an estimate of the efficiency of the similarity-based virtual screening with respect of the corresponding reference structure. Examples of such ROC curves are shown in Fig. 5 (curves *3*, *6*, and *7*). The smallest area under the ROC curve is observed for curve *7* (AUC = 0.31). Nevertheless, the average area under all ROC curves constructed according to this procedure is 0.97. Therefore, the proposed method is characterized by a somewhat higher statistical performance of the classifier (0.99 and 1.00 depending on the type of the molecular field). In addition, unlike the similarity-based virtual screening, the new method allows one, first, to take into account data on numerous active structures and, second, adjust the similarity measure to achieve the maximum efficiency of the virtual screening.

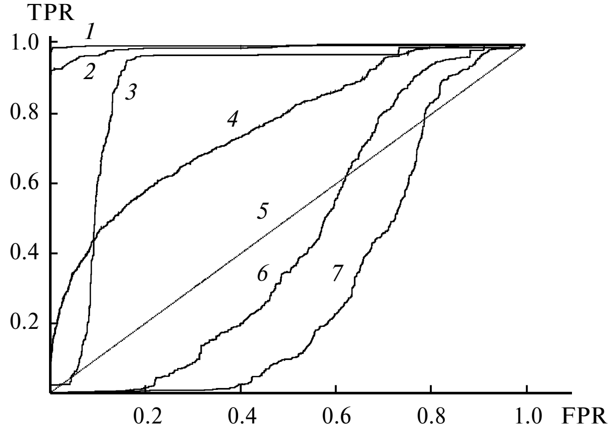

Fig. 5. ROC curves for one-class models constructed with the use of 1-SVM and electrostatic fields, as well as of molecular fingerprints and Carhart descriptors (*1*); 1-SVM and steric/hydrophobic fields (*2*); conventional similarity-based virtual screening using the Tanimoto index and reference structures (*3*, *6*, and *7*); 1-SVM and spectrophores (*4*); an arbitrary classifier (*5*). FPR (False Positive Rate) = 1 — [TN/(FP + TN)], TPR (True Positive Rate) = TP/(TP + FN), where TN/(FP + TN) is the specificity and TP/(TP + FN) is the sensitivity.

Therefore, in the present study we constructed for the first time models for virtual screening of potent inhibitors of HIV-1 reverse transcriptase in terms of the one-class approach with the use of the support vector machine method. It was shown that the representation of the molecular structures of organic ligands using continuous molecular fields enables one to obtain classification models of higher quality compared to approaches based on molecular fingerprints, spectrophores, and Carhart fragment descriptors. The best models constructed with the use of continuous molecular fields have parameters similar to those of the ideal classifier. These models can be recommended for a high-throughput virtual screening.

This study was financially supported by the Russian Foundation for Basic Research (Project No. 10-07-00201), the Council on Grants of the President of the Russian Federation (Program for State Support of Young Scientists of the Russian Federation, Grant MK-1351.2011.3), and the Ministry of Education and Science of the Russian Federation (Federal Target Program "Research and Development in Priority Fields of Science and Technology in Russia in 2007—2012," 2011-1.2- 512-055, Application 2011-1.2-512-055-011).

References

1. G. Barbaro, A. Scozzafava, A. Mastrolorenzo, C. T. Supuran, *Curr. Pharm. Design*, 2005, **11**, 1805.
2. E. De Clercq, *J. Med. Chem.*, 2005, **48**, 1.
3. M. B. Nawrozkij, D. Rotili, D. Tarantino, G. Botta, A. S. Eremiychuk, I. Musmuca, R. Ragno, A. Samuele, S. Zanoli, M. Armand-Ugón, I. Clotet-Codina, I. A. Novakov, B. S. Orlinson, G. Maga, J. A. Esté, M. Artico, A. Mai, *J. Med. Chem.*, 2008, **51**, 4641.
4. R. Garg, S. P. Gupta, H. Gao, M. S. Babu, A. K. Debnath, C. Hansch, *Chem. Rev.*, 1999, **99**, 3525.
5. G. Gini, M. V. Craciun, C. König, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1897.
6. L. Bruno-Blanch, J. Galvez, R. García-Domenech, *Bioorg. Med. Chem. Lett.*, 2003, **13**, 2749.
7. S. Rodgers, *J. Chem. Inf. Model.*, 2006, **46**, 569.
8. B. Su, M. Shen, E. X. Esposito, A. J. Hopfinger, Y. J. Tseng, *J. Chem. Inf. Model.*, 2010, **50**, 1304.
9. P. V. Karpov, I. I. Baskin, V. A. Palyulin, N. S. Zefirov, *Dokl. AN*, 2011, **437**, 642 [*Dokl. Chem.* (*Engl. Transl.*), 2011, **437**, 107].
10. I. I. Baskin, N. Kireeva, A. Varnek, *Mol. Inf.*, 2010, **29**, 581.
11. N. Fechner, A. Jahn, G. Hinselmann, A. Zell, *J. Chemin-form.*, 2010, **2**, 2.
12. P. Bultinck, W. Langenaeker, P. Lahorte, F. De Proft, P. Geerlings, C. Van Alsenoy, J. P. Tollenaere, *J. Phys. Chem. A*, 2002, **106**, 7895.
13. P. Bultinck, W. Langenaeker, R. Carbó-Dorca, J. P. Tollenaere, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 422.

14. R. E. Carhart, D. H. Smith, R. Ventkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 64.
15. P. M. Vasil´ev, A. A. Spasov, *Ros. Khim. Zh.,* 2006, No. 2, 108 [*Mendeleev Chem. J.*, 2006, No. 2 (in Russian)].
16. N. I. Zhokhova, I. I. Baskin, D. K. Bakhronov, V. A. Palyulin, N. S. Zefirov, *Dokl. AN*, 2009, **429**, 201 [*Dokl. Chem.* (*Engl. Transl.*), 2009, **429**, 273].
17. R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. K. Wegner, E. L. Willighagen, *J. Chem. Inf. Model.*, 2006, **46**, 991.
18. N. V. Artemenko, I. I. Baskin, V. A. Palyulin, N. S. Zefirov, *Dokl. AN*, 2001, **381**, 203 [*Dokl. Chem.* (*Engl. Transl.*), 2001, **381**, 317].
19. T. Fawcett, *Pattern Recognition Lett.*, 2006, **27**, 861.
20. S. K. Kearsley, G. M. Smith, *Tetrahedron Computer Methodology*, 1990, **3**, 615.
21. N. Huang, K. Shoichet, J. J. Irwing, *J. Med. Chem.*, 2006, **49**, 6789.
22. Pang-Ning Tan, M. Steinback, V. Kumar, *Introduction to Data Mining*, Publisher: Addision-Wesley, 2006, 769 pp.
23. R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000, 667.
24. I. Baskin, A. Varnek, in *Chemoinformatics Approaches to Virtual Screening*, RCS Publishing, 2008, **338**, 1.
25. P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, J.-P. Vert, *J. Chem. Inf. Model.*, 2005, **45**, 939.
26. P. V. Karpov, I. I. Baskin, N. I. Zhokhova, H. C. Zefirov, *Dokl. AN*, 2011, **440**, 480 [*Dokl. Chem.* (*Engl. Transl.*), 2011, 263].
27. V. Venkatraman, V. I. Perez-Nueno, L. Mavridis, D. W. Ritchie, *J. Chem. Inf. Model.*, 2010, **50**, 2079.

Received June 10, 2011;
in revised form September 1, 2011