

## Full Articles

### Fragmental descriptors in (Q)SAR: prediction of the assignment of organic compounds to pharmacological groups using the support vector machine approach

*E. P. Kondratovich, N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, and N. S. Zefirov\**

*Department of Chemistry, M. V. Lomonosov Moscow State University,  
1/3 Leninskie Gory, 119991 Moscow, Russian Federation.  
Fax: +7 (495) 939 0290. E-mail zhokhova@org.chem.msu.ru; zefirov@org.chem.msu.ru*

The structure–activity classification models for prediction of the assignment of organic compounds to 40 pharmacological groups were constructed in the framework of the fragmental approach using the support vector machine technique, the Platt–Wu probabilistic model, and resampling procedure. The models constructed allow one to predict possible types of the side pharmacological effects of drugs.

**Key words:** structure–activity model, pharmacological group, organic compounds, fragmental approach, fragmental descriptors, support vector machine.

At present, theoretical prediction of the spectrum of biological activity of organic compounds using the (quantitative) structure–activity relationships ((Q)SAR) methodology is an intrinsic component of novel drug design.<sup>1</sup> This allows one to assess in advance both the main and side pharmacological effects of compounds synthesized and to optimize the expenses for experimental assays. Recently, not only traditional (Q)SAR procedures, but also the machine learning classification methods based on the search for correlations between the molecular structure (represented by a set of descriptors) and activity (1 stands for active, 0 otherwise) of a compound have widely been used to this end. Among these methods,<sup>2–4</sup> the Support Vector Machine (SVM) technique becomes increasingly more popular. Its distinctive features include high flexibil-

ity and accuracy of the results of prediction.<sup>5–7</sup> A comparative assessment of modern methods of machine learning and their combinations by solving various classification tasks<sup>8</sup> suggested that the SVM approach often gives better results than artificial neural networks, the random forest and decision tree methods, the nearest neighbors method, and the naive Bayesian classifier (NBC). Various modifications of the last-mentioned technique serve as a basis for some modern software for qualitative prediction of biological activity, *e.g.*, PASS, developed to predict the spectrum of biological activities of organic compounds using the multilevel neighborhoods of atoms (MNA) descriptors.<sup>9,10</sup>

Recently, the SVM technique has efficiently been used for prediction of some types of biological activity of com-

pounds belonging to particular chemical classes<sup>11–13</sup> and for assignment of these compounds to potential pharmacological groups.<sup>14</sup> For instance, in a study<sup>15</sup> published in the course of preparation of this manuscript the SVM method and multiclass classification were used to investigate the spectrum of biological activity (up to 100 hierarchical subtypes) of a set of compounds from the MDDR database. Nevertheless, the authors of that study failed to estimate the probability of assignment of particular types of activity to organic compounds and to construct balanced models.

A universal method of the description of the structures of organic compounds in the framework of the (Q)SAR methodology is the fragmental approach. The advantages of fragmental descriptors (FD) include a clear meaning of each of them and the possibility of fast automated generation based on only the structural formula of a given compound. FD calculations do not require knowledge of the molecular geometry or electronic structure and therefore these descriptors can be used with ease in processing of large databases. The fragmental descriptors are widely used to describe the structures of organic compounds in (Q)SAR modeling of their biological activities. Often, the FD provide higher predictive power of models and better interpretability of the results compared to sets of other topological indices. Earlier,<sup>16–26</sup> we have pointed to the advantages of FD and used them for (Q)SAR prediction of biological activities and physicochemical properties including chromatographic retention indices,<sup>16</sup> boiling points,<sup>22</sup> enthalpies of sublimation,<sup>23</sup> polarizabilities,<sup>24</sup> flash points,<sup>25</sup> and the stability constants of complexes with  $\beta$ -cyclodextrin (see Ref. 26) for various classes of organic compounds.

In the present work, we studied the applicability of FD for prediction of the assignment of various classes of organic compounds to pharmacological groups using the SVM technique.

### Calculation Procedure

The database of chemical compounds and pharmacological groups for constructing the structure–activity classification models using the SVM technique was extracted from the KEGG DRUG database, a part of the KEGG (Kyoto Encyclopedia of Genes and Genomes, Kahenisa Lab, Bioinformatics Center, Kyoto University\*) database, which, in particular, contains structural information and characteristics of pharmacological groups for 6000 compounds, the active principles of main commercially available drugs registered on the world market. The database includes 120 pharmacological groups. The modeled property was the membership of an organic compound in a particular pharmacological group; therefore, all structures of the organic compounds from the database were characterized using the two-class approach (1 if the compound belongs to a particular pharmaco-

logical group, *i.e.*, it is active, and 0 otherwise). Then, in accordance with the results of preliminary modeling, we formed the training data sets for 40 target pharmacological groups, the number of compounds in each group exceeded 40 (Table 1). The total number of structures of organic compounds in the whole data set was 3450.

SVM-based SAR modeling was performed using the two-class classification with the LIBSVM software (Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001\*). To estimate the probability of assignment of a chemical compound to a particular class in the framework of the SVM methodology, the Platt method<sup>27</sup> modified by Wu<sup>28</sup> was applied. The chemical structures of the organic compounds from the working set were characterized by some sets of fragmental descriptors (occurrence numbers of structural fragments in the chemical structure) calculated using the FRAGMENT module of the NASAWIN software developed at the Department of Chemistry, M. V. Lomonosov Moscow State University.<sup>29</sup> The FD were generated using the algorithm of hierarchical classification of atoms.<sup>30</sup> A scheme that included the type of the chemical element, hybridization, bonding environment of atoms, formal atomic charge, and the number of hydrogen atoms in the nearest environment was used for assignment of fragments to different types. We also considered the atomic types of organogens (C, N, O, S, Se, P, As, Si, F, Cl, Br, I). To choose particular FD for the most accurate description of the topology of the organic structures under study, we analyzed fragments with different number of nonhydrogen atoms, namely, the 1–10 atomic chains, rings of size from 3 to 10 atoms, branched fragments comprising 4 to 6 atoms, and bicycles built of 6 to 10 atoms. In constructing the structure–activity classification models, for each pharmacological group we used the sets of descriptors corresponding to the number of inclusions of structural fragments containing 1 to 4, 1 to 8, and 1 to 10 nonhydrogen atoms. The predictive power of the models was assessed using the tenfold cross-validation procedure; the smallest error of prediction was provided by choosing optimum values of the variable SVM parameters (see Ref. 5) *C* (parameter characterizing the balance between the error in classification and complexity of the model) and *g* (Gaussian core parameter).

The following statistical characteristics were calculated for all models.

1. The accuracy of prediction (relative number of compounds with correctly predicted activity).
2. The balanced accuracy (arithmetic mean of the accuracy of prediction for active and inactive compounds). For a balanced set containing equal number of active and inactive compounds, the balanced accuracy equals the overall accuracy.

The optimum values of the parameters *C* and *g* were chosen in accordance with the maximum values of the balanced accuracy of the models.

### Results and Discussion

Practically, in most cases where the biological activity of organic compounds is predicted using the two-class approach (the assignment of a chemical compound to a particular class is determined by the occurrence or lack

\* <http://kegg.org/Kyoto/KEGG/DRUG>.

\* <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

**Table 1.** Accuracy of the SVM and NBC classification models used for assignment of organic compounds to 40 pharmacological groups

Group	Group code	Group name	Fragment size (1 to <i>N</i> nonhydrogenatoms)	Accuracy (%)	
				SVM	NBC
1	1	Neurotropic agents	1–4	75.2	68.9
2	11	Neurotropic agents for central nervous system	1–10	81.8	76.8
3	112	Anxiolytics, sedatives, and hypnotics	1–4	85.9	80.1
4	113	Antiepileptic agents	1–10	81.5	73.2
5	114	Antipyretics and non-narcotic analgesics, non-steroidal anti-inflammatory agents	1–8	78.6	73.9
6	116	Antiepileptic agents	1–8	70.6	76.3
7	117	Psychotropic agents	1–8	85.3	79.7
8	12	Neurotropic agents for peripheral nervous system	1–8	78.9	81.4
9	13	Neurotropic agents affecting individual sensory organs	1–10	62.4	62.8
10	131	Ophthalmic agents	1–4	66.8	62.4
11	2	Organotropic agents	1–8	72.4	68.7
12	21	Cardiovascular agents	1–4	77.1	—
13	211	Cardiotonics	1–8	77.8	69.9
14	212	Antiarrhythmic agents	1–10	85.0	73.9
15	213	Diuretics	1–4	89.3	75.7
16	214	Antihypertensive agents	1–4	80.9	72.2
17	217	Vasodilators	1–4	67.6	65.9
18	22	Respiratory organ agents	1–8	75.7	75.2
19	225	Bronchodilators	1–10	86.6	78.9
20	23	Gastrointestinal agents	1–10	73.9	70.8
21	232	Anti-gastric ulcer and anti-duodenal ulcer agents	1–4	72.9	72.0
22	24	Hormones	1–4	90.7	86.9
23	247	Estrogens and progesterones	1–4	91.7	90.9
24	25	Urogenital agents	1–4	74.9	64.6
25	26	Dermatotropic agents	1–4	76.0	67.1
26	265	Antiparasitic dermatosis agents	1–8	84.7	78.1
27	3	Metabolics	1–8	75.4	73.4
28	4	Cellular function agents	1–8	68.4	65.2
29	42	Antineoplastics	1–10	72.3	68.9
30	44	Allergic agents	1–8	75.1	66.5
31	6	Antimicrobial and antiparasitic agents	1–10	84.2	80.8
32	61	Antibiotics	1–10	94.3	90.6
33	611	Antibiotics acting on gram-(+) bacteria	1–4	91.3	84.0
34	613	Antibiotics acting on gram-(+)- and gram-(–) bacteria	1–4	95.7	92.0
35	6132	Cephem antibiotics	1–4	98.9	97.6
36	62	Chemotherapeutics	1–4	85.8	76.9
37	624	Synthetic antibacterials	1–4	94.5	86.7
38	625	Antivirals	1–8	82.9	74.1
39	64	Antiparasitic agents	1–8	70.0	64.4
40	8	Narcotics	1–8	84.0	84.1

of particular kind of biological activity), researchers deal with unbalanced sets. The set is considered unbalanced if the number of active compounds with respect to a particular property differs significantly from the number of inactive compounds. The quality of the SVM classification models is known to depend on the degree of balance of the database with respect to the modeled property.<sup>31</sup> In particular, in solving the problem of two-class classification using unbalanced databases the SVM-models constructed are also unbalanced, *i.e.*, they show a trend to distortion of

the results of prediction, giving preference to the class containing the larger number of compounds in the database. In other words, unbalanced models give different accuracy of prediction for the active and inactive compounds. In the limiting case, the unbalanced models may lead to trivial results where the most abundant class in the database is assigned to all compounds.

To avoid this situation, in the present work we studied two different strategies. One of them involved the use of a special modification of the SVM technique for un-

balanced databases.<sup>32</sup> In this case we assigned to active compounds the weight parameter ( $w$ ) equal to the ratio of the total number of inactive (with respect to the modeled property) compounds to the total number of active compounds ( $w = N_{ia}/N_a$ ). According to the intrinsic algorithm of the LIBSVM software, the introduction of the parameter  $w$  should reduce the unbalance of the models by using the weighting procedure. Indeed, our numerical simulation showed that this holds for a number of the pharmacological categories under study. Nevertheless, for some other pharmacological groups (metabolics, agents affecting cellular functions, neurotropic agents affecting individual sensory organs, dermatotropic agents, antiparasitic dermatosis agents, ophthalmic agents, antiarrhythmic agents, vasodilators, anti-gastric ulcer and anti-duodenal ulcer agents), for which the corresponding sets were significantly unbalanced while the number of non-assigned compounds much exceeded that of assigned compounds ( $N_{ia}/N_a \gg 1$ ), even using the weighting parameter we failed not only to construct balanced models, but also to avoid trivial results. Unbalanced character of classification models for these groups was a consequence of highly unbalanced character of the database with respect to a given pharmacological group. Consequently, the first approach was rejected because it provides no correct solution to the problem posed.

Contrary to this, the other approach appeared to be efficient. It is based on a procedure recommended<sup>31</sup> for constructing models for prediction of biological activity in the case of unbalanced sets. In the framework of this approach, we used the resampling procedure to reduce the effect of the degree of unbalance of the working sets in constructing models for all 40 target pharmacological groups. Based on the resampling, three balanced model sets were formed for each pharmacological group. Each model set contained information on (i) all active compounds and (ii) inactive ones that were randomly chosen from the initial database in such a manner that the numbers of the active and inactive compounds be equal. Next, for each pharmacological group we constructed classification models based on the corresponding three sets with subsequent averaging of the results obtained.

At first glance, the fact that the quality of a model is improved after rejection of some data may appear to be surprising because this contradicts the practice in application of statistical methods. Nevertheless, there are strong grounds supporting this strategy. First, since all models in the "ensemble" were constructed using different sets of inactive compounds, we reject a much smaller amount of information on the inactive compounds compared with the case of a single model. Moreover, almost no information is lost for a large number of models in the ensemble. Second, a feature of the biological activity databases for organic compounds is that they contain only information on the presence of a particular type of activity and do not

contain information that a given compound is inactive. It follows that by performing the resampling procedure we reject "digitized assumptions" (generally, a guess-work) rather than actual experimental data on the lack of activity. Again, the actual activity data are not rejected in the course of resampling.

The model construction procedure involved the choice of the optimum FD size (number of nonhydrogen atoms in the corresponding fragment of the chemical structure) to achieve the best description of the structures of all compounds assigned to each pharmacological group. Table 1 presents the averaged accuracies (equal to the balanced accuracy) of the best SVM models for assignment of organic compounds to 40 pharmacological groups constructed using the tenfold cross-validation procedure, resampling procedure, and optimum FD size for each group. For comparison, Table 1 also lists the accuracies of the corresponding NBC models (WEKA 3.5.8 software).<sup>33\*</sup>

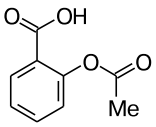
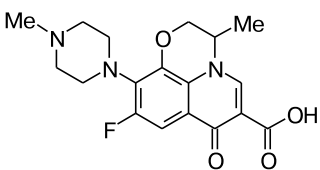
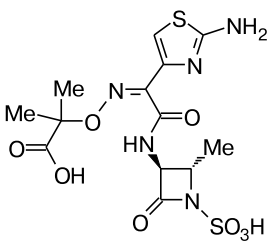
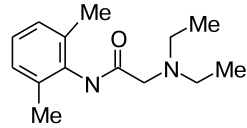
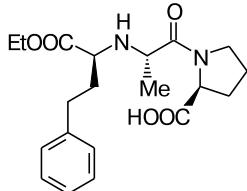
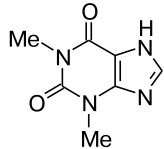
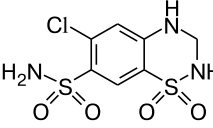
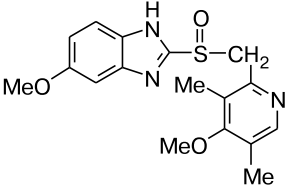
From Table 1 it follows that the accuracies of the SVM models for 38 pharmacological groups are better than those of the models based on the NBC algorithm. The exception is group 116 (anti-Parkinsonian agents). The spread in the accuracy values for the 40 SVM models is 62.3–98.9%. By and large, all models constructed for the six main pharmacological groups (neurotropic agents, organotropic agents, metabolics, agents affecting cellular functions, antimicrobial and antiparasitic agents, and narcotics) have a somewhat lower accuracy than the models for pharmacological groups at lower hierarchical levels. This can be due to the variety of chemical classes of compounds and, correspondingly, to a large number of structurally different organic compounds in each pharmacological group, which reduces the quality of the models. The best accuracy was obtained for the SVM-models for the following pharmacological groups (see Table 1): 6132 (98.9%), 613 (95.7%), 624 (94.6%), and 61 (94.3%). The least accurate are the models for the following pharmacological groups: 13 (62.3%), 131 (66.8%), and 217 (67.6%).

To additionally verify the quality of the classification models constructed, we estimated the probability of assignment of eight organic compounds (active component of the known drugs) including aspirin, aztreonam, levofloxacin, lidocaine, omeprazole, enalapril, theophylline, hypothiazide, and cefalexin to 40 pharmacological groups studied.<sup>34</sup> The results obtained are listed in Table 2.

From the analysis of the data in Table 2 it follows that in addition to the correct prediction of the main action of the drugs, the models constructed can also predict the side effects of the drugs, although in the latter case the database used for constructing the models contained no relevant information. For instance, levofloxacin was assigned to the main pharmacological group 624 with a probability of 0.740. However, this drug can also be assigned to the

\* <http://www.cs.waikato.ac.nz/ml/weka/>.

**Table 2.** Prediction of assignment of eight organic compounds to pharmacological groups based on the SVM models according to the hierarchy of the KEGG DRUG database (listed are the probabilities greater than 0.4)

Compound	Probability of assignment (KEGG DRUG)	Pharmacological group (KEGG DRUG) (prediction)	Pharmacological effects <sup>34*</sup>
Aspirin	 0.890	114. Antipyretic and non-narcotic analgesics, non-steroidal anti-inflammatory agents	Pain relieve, antipyretic and anti-inflammatory action
Levofloxacin	 0.740	624. Synthetic antibacterial agents	Antibacterial
	0.502	131. Ophthalmic agents	<i>Involuntary eye movements</i>
	0.403	112. Anxiolytics, sedatives, and hypnotics	<i>Drowsiness, apathy</i>
Aztreonam	 0.658	61. Antibiotics	Antibacterial
	0.916	6. Antimicrobial, antiparasitic, and antihelminthic agents	Antimicrobial, antiparasitic, antihelminthic
	0.927	613. Antibiotics affecting gram-(+) and gram(-) bacteria	Antibiotic affecting gram(-) bacteria
Lidocaine	 0.749	212. Antiarrhythmic agents	Antiarrhythmic
	0.481	112. Anxiolytics, sedatives, and hypnotics	<i>Drowsiness</i>
	0.541	22. Respiratory agents	<i>Vertigo, respiratory arrest</i>
Enalapril	 0.628	117. Psychotropic agents	<i>Euphoria, impairment consciousness</i>
	0.974	214. Antihypertensive agents	Antihypertensive
Theophylline	 0.753	225. Bronchodilators	Broncholytic
Hypothiazide	 0.775	25. Urogenital agents	Diuretic
	0.611	131. Ophthalmic agents	<i>Reduction of intraocular pressure</i>
	0.881	214. Antihypertensive agents	Antihypertensive
Omeprazole	 0.574	23. Gastrointestinal agents	Anti-ulcer
	0.425	112. Anxiolytics, sedatives, and hypnotics	<i>Drowsiness, weakness</i>

\* Listed are the main and side (in italics) effects.

pharmacological group 131, although with a lower probability of 0.502; this can be related to such side effects of levofloxacin as "involuntary eye movements".<sup>34</sup> Moreover, this compound can also be assigned to the pharmacological group 112 with a probability of 0.403, which can again be related to the side effects of the drug (drowsiness, apathy). From Table 2 it follows that the side effects of lidocaine ("drowsiness", "vertigo, respiratory arrest", and "euphoria, impairment of consciousness"), hypotiazide ("reduction of intraocular pressure"), and omeprazole ("drowsiness, weakness") are predicted in exactly the same manner.

Thus, in the present work we have shown that by combining the fragmental approach, the SVM technique, the Platt—Wu probabilistic model, and resampling procedure one can construct models for quite reliable assignment of organic compounds to pharmacological groups of registered drugs from the KEGG DRUG database. In the vast majority of cases, the models constructed have a much higher predictive power than the models constructed using the NBC approach, which is treated so far as a reference when solving this type of tasks. For a number of drugs, the results of prediction are consistent with the known, both main (desired) and side (undesired), pharmacological effects of the corresponding compounds.<sup>34</sup>

## References

1. N. S. Zefirov, *Vestn. Ros. Akad. Nauk*, 2004, **74**, 415 [*Herald of the Russ. Acad. Sci. (Engl. Transl.)*, 2004, **74**].
2. *Machine Learning, Neural and Statistical Classification*, Eds D. Michie, D. J. Spiegelhalter, C. C. Taylor, Ellis Horwood, London, 1994.
3. V. Svetnik, A. Liaw, C. Tong, J. C. Culbertson, R. P. Sheridan, B. P. Feuston, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1947.
4. P. Itskowitz, A. K. Tropsha, *J. Chem. Inf. Model.*, 2005, **45**, 777.
5. N. V. Vapnik, *IEEE Trans. Neural Networks*, 1999, **10**, 996.
6. O. Ivanciuc, in *Reviews in Computational Chemistry*, Weinheim, 2007, **23**, 291.
7. M. K. Warmuth, J. Liao, G. Ratsch, M. Mathieson, S. Putta, C. Lemmen, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 667.
8. R. Caruana, A. Niculescu-Mizil, *Proc. 23rd Intern. Conf. Machine Learning*, Pittsburgh, PA, 2006, 161.
9. D. A. Filimonov, V. V. Poroikov, *Ros. Khim. Zhurn. [Mendeleev Chem. J.]*, 2006, **2**, 68 (in Russian).
10. V. V. Poroikov, D. A. Filimonov, Y. V. Borodina, A. A. Lagunin, A. Kos, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1349.
11. X. J. Yao, A. Panaye, J. P. Doucet, R. S. Zhang, H. F. Chen, M. C. Liu, Z. D. Hu, B. T. Fan, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1257.
12. Y. Liu, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1823.
13. E. Byvatov, U. Fechner, J. Sadowski, G. Schneider, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1882.
14. *Designing Drugs and Crop Protectants: Processes, Problems and Solutions*, Eds M. Ford, D. Livingstone, J. Dearden, H. Waterbeemd, Blackwell Publishing, Bournemouth, 2002, 268.
15. K. Kawai, S. Fujishima, Y. Takahashi, *J. Chem. Inf. Model.*, 2008, **48**, 6, 1152.
16. N. S. Zefirov, V. A. Palyulin, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1112.
17. D. E. Petelin, V. A. Palyulin, N. S. Zefirov, G. McFarland, *Dokl. Akad. Nauk*, 1992, **327**, 1019 [*Dokl. Chem. (Engl. Transl.)*, 1992, **327**].
18. D. V. Sukhachev, T. S. Pivina, N. I. Zhokhova, N. S. Zefirov, S. I. Zeman, *Izv. Akad. Nauk, Ser. Khim.*, 1995, 1653 [*Russ. Chem. Bull. (Engl. Transl.)*, 1995, **44**, 1585].
19. D. V. Sukhachev, T. S. Pivina, N. I. Zhokhova, N. S. Zefirov, *Izv. Akad. Nauk, Ser. Khim.*, 1995, 1657 [*Russ. Chem. Bull. (Engl. Transl.)*, 1995, **44**, 1589].
20. D. V. Sukhachev, T. S. Pivina, N. I. Zhokhova, N. S. Zefirov, S. I. Zeman, *Izv. Akad. Nauk, Ser. Khim.*, 1995, 1661 [*Russ. Chem. Bull. (Engl. Transl.)*, 1995, **44**, 1594].
21. N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, N. S. Zefirov, *Dokl. Akad. Nauk*, 2007, **417**, 639 [*Dokl. Chem. (Engl. Transl.)*, 2007, **417**].
22. N. S. Zefirov, V. A. Palyulin, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1022.
23. N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov, N. S. Zefirov, *Zh. Prikl. Khim.*, 2003, 1966 [*Russ. J. Appl. Chem. (Engl. Transl.)*, 2003].
24. N. I. Zhokhova, V. A. Palyulin, I. I. Baskin, A. N. Zefirov, N. S. Zefirov, *Izv. Akad. Nauk, Ser. Khim.*, 2003, 1005 [*Russ. Chem. Bull., Int. Ed.*, 2003, **52**, 1061].
25. N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov, N. S. Zefirov, *Izv. Akad. Nauk, Ser. Khim.*, 2003, 1787 [*Russ. Chem. Bull., Int. Ed.*, 2003, **52**, 1885].
26. N. I. Zhokhova, E. V. Bobkov, I. I. Baskin, V. A. Palyulin, A. N. Zefirov, N. S. Zefirov, *Vestn. MGU, Ser. 2, Khimiya*, 2007, **48**, 5, 329 [*Moscow University Chem. Bull. (Engl. Transl.)*, 2007, **62**, 269].
27. J. Platt, in *Advances in Large Margin Classifiers*, Eds A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans, MIT Press, Cambridge, MA, 2000.
28. T.-F. Wu, C.-J. Lin, R. C. Weng, *J. Machine Learning Research*, 2004, **5**, 975.
29. I. I. Baskin, N. M. Halberstam, N. V. Artemenko, V. A. Palyulin, N. S. Zefirov, in *EuroQSAR 2002 Designing Drugs and Crop Protectants: Processes, Problems and Solutions*, Eds M. Ford, D. Livingstone, J. Dearden, H. Waterbeemd, Blackwell Publishing, Malden, 2003, 260.
30. N. V. Artemenko, I. I. Baskin, V. A. Palyulin, N. S. Zefirov, *Dokl. Akad. Nauk*, 2001, **381**, 203 [*Dokl. Chem. (Engl. Transl.)*, 2001].
31. J. J. Chen, C. A. Tsai, J. F. Young, R. L. Kodell, *SAR and QSAR in Environmental Research*, 2005, **16**, 517.
32. E. Osuna, R. Freund, F. Girosi, *AI Memo 1602*, Massachusetts Institute of Technology, Boston, 1997.
33. I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques (Second Ed.)*, Morgan Kaufmann Publ., San Francisco, 2005.
34. *Entsiklopediya lekarstv. Registr lekarstvennykh sredstv Rossii [Encyclopedia of Drugs. The Russian Register of Drugs]*, Ed. Yu. F. Krylov, 7th Ed., RLS-2000, Moscow, 2000 (in Russian),