SAR and QSAR in Environmental Research,
VOL. 14 (5–6), October–December 2003, pp. 339–347

Taylor & Francis
Taylor & Francis Group

PREDICTION OF BIOLOGICAL ACTIVITY SPECTRA VIA THE INTERNET*

A. SADYM, A. LAGUNIN, D. FILIMONOV and V. POROIKOV†

Institute of Biomedical Chemistry of Russian Academy of Medical Sciences, Pogodinskaya Street 10, Moscow, 119121, Russia

(Received 16 July 2003; In final form 1 September 2003)

The majority of biologically active compounds have both pharmacotherapeutic and side/toxic actions. To estimate general efficacy and safety of the molecules under study, their biological potential should be thoroughly evaluated. In an early stage of study, only information about structural formulae was available and was used as an input for computational prediction. Based on a structural formulae of compounds presented as SDF or MOL-files, computer program PASS predicts 900 pharmacological effects, mechanism of action, and specific toxicity. An average accuracy of prediction in leave-one-out cross-validation is about 85%. For evaluating new compounds, scientific community may use PASS via the Internet for free at URL: http://www.ibmh.msk.su/PASS. In the first 18 months of PASS Inet's use, approximately 1000 researchers from 60 countries have obtained predicted biological activity spectra for about 23,000 different chemical compounds. More than 64 million PASS predictions for almost 250,000 compounds from Open NCI database are available on the web site http://cactus.nci.nih.gov/ncidb2/. These predictions are used for selecting compounds with desirable and without unwanted types of biological activities among the NCI samples available for screening.

Keywords: Biological activity spectra; Computer-aided prediction; PASS; Internet; Free use; Drug R & D

INTRODUCTION

Biological activity of chemical compound may account for its application as a pharmaceutical or may cause side/toxic effects in the organisms. Since the selectivity of biologically active compounds is always limited, general biological potential of molecules under study should be evaluated in detail. Undoubtedly, it is not feasible to test hundreds of thousands of hits in thousands of screens currently available; therefore computer prediction is a "method-of-choice".

In the early stages of R & D no other information except structural formulae is available. Structural formulas should be used as an input for computational predictions.

Based on the structural formulas of compounds presented as SDF- or MOL-files, computer program PASS [1]‡ predicts 900 pharmacological effects, molecular mechanisms of action, mutagenicity, carcinogenicity, teratogenicity and embryotoxicity. An average accuracy of prediction in leave-one-out cross-validation is about 85%.‡ PASS predictions are based on

*Presented at CMTPI 2003: Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (Thessaloniki, Greece, September 17–19, 2003)

†Corresponding author e-mail: vvp@ibmh.msk.su

‡URL: [http://www.ibmh.msk.su/PASS]

ISSN 1062-936X print/ISSN 1029-046X online
DOI: 10.1080/1062936031000162393 5

the analysis of structure–activity relationships (SAR) for the training set consisting of about 46,000 biologically active compounds. Despite the incompleteness of the training set, PASS algorithm is robust enough to obtain the reliable SAR [2] or even to predict such "raw" characteristics of compound as "drug-likeness" [3]. It was shown that application of PASS increases the number of "actives" in the selected subset up to factor 17 [4]; therefore, it can be effectively applied for R & D of new pharmaceuticals.

A number of Russian and foreign research institutions are successfully applying PASS for several years. To introduce this facility to a broader circle of specialists involved in drug design and discovery, we had developed an Internet version of PASS that predicted 319 types of biological activity [5]. From year 1998 to 2001 more than 400 users from about 30 countries have performed predictions for about 3000 drug-like compounds using this old PASS Internet version.

This paper describes a new Internet version of PASS (PASS Inet). PASS Inet was introduced into practice since December 25, 2001 and enables the users to predict biological activity spectra with updated version of the program, which currently predicts 900 types of biological activity.

METHODS

General Description of PASS

Contrary to many other existing methods of SAR/QSAR/QSPR analysis focused on predicting a single type of biological activity within the same chemical series, PASS predicts the whole biological activity spectra of a molecule under study.

Technique of PASS is based on the analysis of SARs for the training set currently including about 46,000 drugs, drug-candidates and lead compounds whose biological activity is determined experimentally. These SARs are obtained during the training procedure and are stored in the knowledgebase called SAR Base. The so-called Multilevel Neighbourhoods of Atoms published elsewhere [6] are used in PASS chemical descriptors. The set of MNA descriptors is generated on the basis of structural formula (formulas) presented in MOL-file (SDF-files) form.[¶] Since MNA descriptors are generated for each compound *de novo*, new descriptors can be obtained upon presentation of a novel structural feature in the compound under study. A detailed description of mathematical algorithm has been published [3,5,7,8] and is also available on the web site.[‡]

The user obtains the results of prediction as a list of activity types, with the probabilities of presence (Pa) and absence (Pi) for each particular activity. Interpretation of the prediction results and selection of the most prospective compounds are based on flexible criteria, which depend on the purpose of particular investigation. If the user chooses rather higher value of Pa as a threshold for selection of probable activities, the chance to confirm the predicted activities by the experiment is also high, but many existing activities will be lost. For instance, if Pa $>$ 80% is used as a threshold, about 80% of real activities will be lost; for Pa $>$ 70%, the portion of lost activities is 70%, etc. By default, Pa $=$ Pi value determined at the training is used as a threshold that provides the mean accuracy of prediction about 85% in leave-one-out cross-validation for all $\sim$46,000 compounds and 900 activities from the PASS training set.

Due to the complex evaluation of biological potential of the molecules being studied by PASS, the user may select the predicted compounds to get the requested types of activity but not the unwanted ones. For instance, mutagenicity and carcinogenicity are always treated as

[¶]URL: [http://www.mdli.com]

unwanted specific toxicity, while sedative/hypnotic action may be considered as useful in hypnotics but undesirable in anxiolytics.

PASS Inet Peculiarities

PASS Internet version is designed to meet some special requirements. First of all, it should not require any special software and has to be used with usual Internet browsers like Internet Explorer or Netscape. No installation of additional programs (like Flash, etc.) on the user's computer is necessary. The user's interface is optimised for a screen resolution of 800 × 600 pixels. Although this resolution is not so popular among the Internet users today, the format is still used in many monitors both in Russia and abroad. The monitor, set to a higher resolution, reproduces such image correctly.

For effective operation of PASS, both the program and SAR Base have to be always residing in the server memory, and a client application (processing the users' requests) will access the loaded program. Among the possible client–server architectures we selected those based on sockets [9]. Finally, the following configuration of PASS Inet system is arranged: (1) Windows 2000 Server (operating system); (2) Apache (web server); (3) HTML (web-page language); (4) Common Gateway Interface (CGI) protocol [10] (program sending data from socket to server; this program and server are written in Delphi 5.0); (5) PHP scripts [11] (user registration, authentication, web page loading); (6) MySQL Server [11] (database management system).

Figure 1 shows an example of the initial page of PASS Inet system, providing brief information about PASS program and main options available. The page's top panel is divided into three parts, each providing link to other web pages. A click on the left- or right-side parts

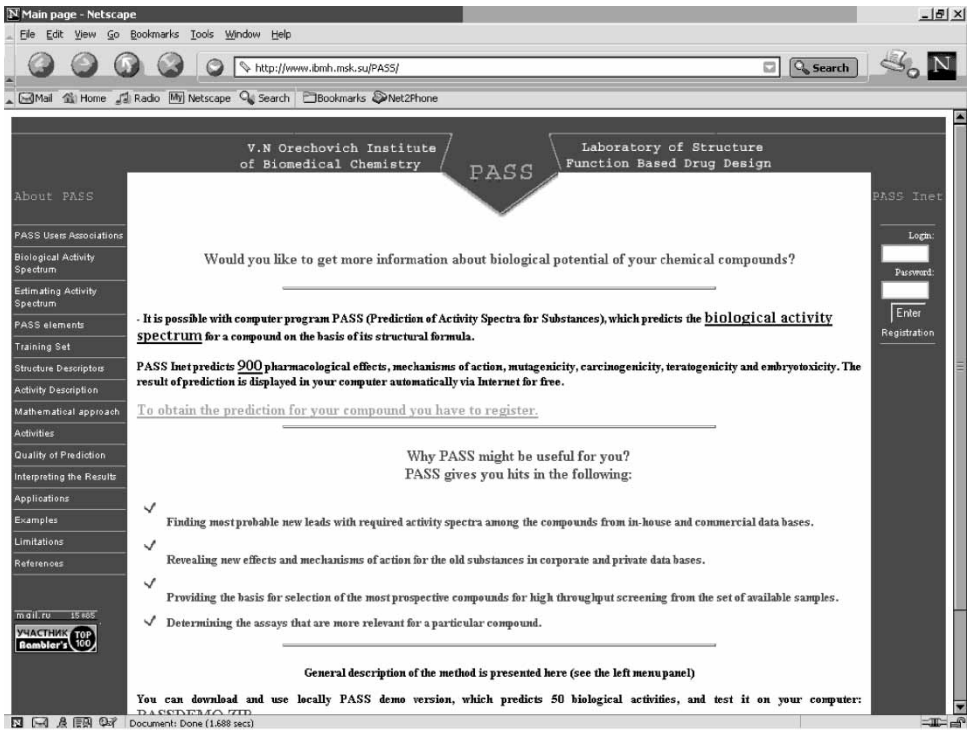

FIGURE 1 Starting page of PASS Inet system.

gives the user an access to the front pages of the Internet site of the Institute of Biomedical Chemistry of Russian Academy of Medical Sciences (Moscow) or the Laboratory for Structure–Function Based Drug Design, respectively. Clicking "PASS" implies transition to the front page of PASS Internet system.

The web site[‡] describes PASS program, including sections devoted to the concept of Biological Activity Spectrum, the list of predicted types of biological activity, the main elements of PASS, application examples and references to publications concerning PASS methodology and its applications. Full-text HTML- or PDF-files are available for some papers and can be downloaded by the user. PASS demo version, which predicts 50 types of biological activity, can be also downloaded from this web site.

On the left-side panel of the screen, grey buttons separated by orange lines represent the heads of all sections in a menu. The activated button turns green, while that indicated by the mouse marker becomes orange. Any desired section is entered upon clicking the selected (orange) cell. The right-side panel performs the functions of identification. If not previously registered, the user has to click the REGISTRATION button under the ENTER button. The field that has to be obligatory filled in by the user is indicated by an asterisk (*). If the necessary field remains unfilled or the user's name contains less than four letters, or if it coincides with the name of another user, the message about error will be obtained. Otherwise, the user will receive a notification of successful registration. During registration, the user submits his e-mail address. Once the registration is successfully accomplished, he will receive a message from pass@ibmh.msk.su with the registration name and password for entering the Internet PASS system.

Any user who synthesizes particular compounds, or is planning to do that, or performs biological activity testing of some compounds, may obtain PASS predictions free for having the Internet access and registering to PASS Internet system. Successful authentication provides the opening of prediction page (Figure 2).

One can use MOL-files prepared with the help of ISIS/Draw[¶] and saved in his computer as an input to PASS Inet or may directly introduce the structure using Marvin applet.[§] Using the prediction web page, one may define a preferable way of structure input by activating/inactivating the USE APPLET mark. If this mark is activated, the Marvin applet is loaded to draw a structure of chemical compound. By means of the applet, the user can introduce, edit or remove the structure. All changes introduced into the structure are also displayed in the auxiliary window. If the user introduces structure with ISIS/Draw chemical editor, the structure has to be highlighted, exported as a mol-file and saved on a disc. On accessing the Internet PASS system, the user clicks BROWSE and selects this file for prediction.

After introducing structure in applet form or by indicating file and pathway, the user can select the threshold value for prediction (default value, $Pa > 30\%$) and click on a button PREDICTION to activate prediction. The result of the prediction appears in a new window.

Interpretation of Prediction Results

Interpreting the results obtained, the user can rely upon the following principles:

(1) Some descriptors in the compound under the prediction may appear to be new compared to the compounds from the PASS training set. If all descriptors are of this kind, the given compound has no fragments in common with those from the training set, thus, no prediction is possible. If the number of new descriptors is relatively large (more than three), the results of prediction should be considered only as a rough estimation.

[§]URL: [http://www.chemaxon.com/marvin]

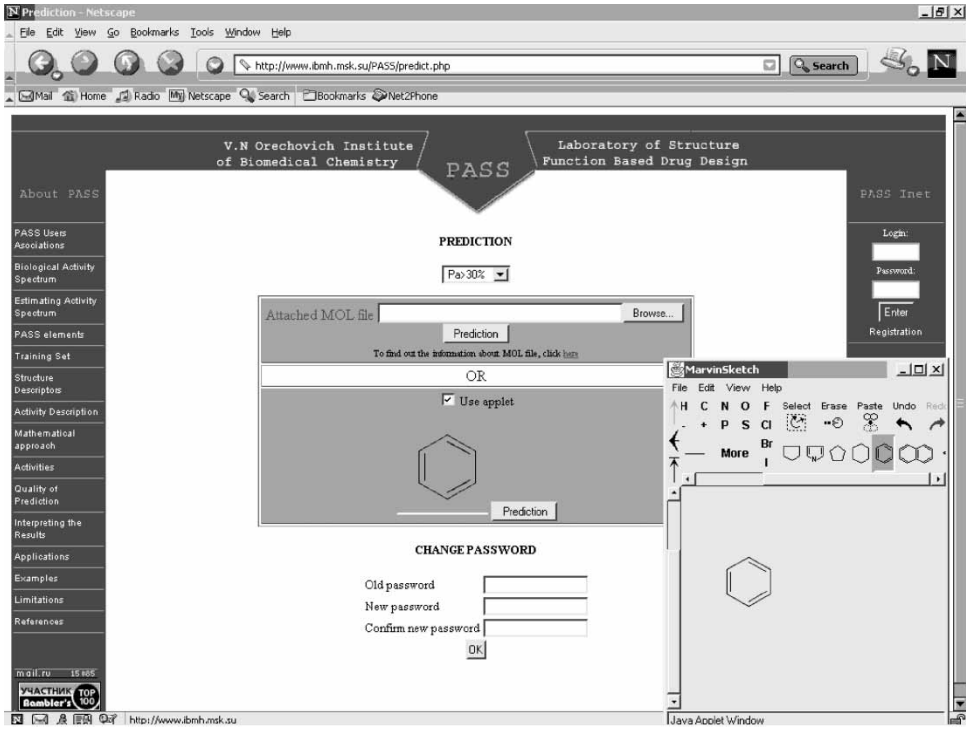

FIGURE 2 Prediction page of PASS Inet. Marvin applet is activated.

(2) If a certain type of activity was predicted for Pa $\geq$ 70%, the compound will most likely exhibit this activity in the experiment; however, the probability is also quite large that this compound is a close analogue of the known drugs.
(3) When 50% $<$ Pa $\leq$ 70%, there is still high probability that the compound will show the activity in the experiment, but a similarity with well-known pharmaceuticals is less probable.
(4) If Pa $<$ 50%, the probability that the given compound will show this activity in the experiment is less. However, if this activity is manifested, quite new lead compound will be found (New Chemical Entity, NCE).

It should also be mentioned that if a compound under prediction is equivalent to a particular one from the training set, the last one is excluded from the training set by default.

For instance, let us consider the prediction for Montelukast (1-((((1R)-1-(3-((1E)-2-(7-Chloro-2-quinolinyl)ethenyl)phenyl)-3-(2-(1-hydroxy-1-methylethyl)phenyl) propyl)thio)methyl)cyclopropaneacetic acid). In the prediction, this compound is excluded from the training set with the two known types of activity (Bronchodilator and Leukotriene antagonist). Sixty different MNA descriptors were generated for this molecule and no one descriptor is found to be new compared to the structures of the PASS training set. Twenty-four from 900 different types of biological activity are predicted with Pa $>$ 30%. Six types of biological activity are predicted with Pa $>$ 70%, most of which are associated with antiasthmatic/antiallergic actions. Some unknown activities are predicted with Pa $<$ 70%, in particular, GABA receptor antagonist, Antileishmanial, etc. These activities being confirmed by the experiment may become a reason for new indications of Montelukast. Prediction for activities with close Pa and Pi values (antiinflammatory ophthalmic, steroid synthesis inhibitor, etc.) cannot be considered as significant.

RESULTS AND DISCUSSION

Analysis of PASS Inet Use

About one thousand users from almost 60 countries were registered for the first 18 months of PASS Inet availability. The majority of users are from Russia (40.9%), USA (9.2%), India (8.6%), Ukraine (5.4%), France (2.5%), Australia (2.0%), Germany (1.9%), Italy (1.9%), UK (1.9%), South Korea (1.8%), Japan (1.4%), etc. Biological activity spectra of about 23,000 unique compounds has been predicted, of which 2000 compounds are included into the PASS training set. Since each new user wants to estimate PASS prediction accuracy by himself, many known pharmaceutical agents are submitted for prediction.

Analysing the structures submitted for prediction, we have found that most of compounds are drug-like and belong to different chemical series. Many users obtained predictions for several (or even more) derivatives from the same chemical class. It probably reflects the use of PASS for designing compounds with desirable and without unwanted properties, prior to the synthesis.

On average, 40 structures per day are submitted for prediction (Figure 3), while with the old PASS Internet version [5] this number was only about four. This statistics clearly demonstrates a significant increase in PASS Inet use after installing new Internet version.

Analysis of statistics of the predicted biological activity spectra demonstrates that a set of about 23,000 compounds submitted for prediction is rather diverse in the pharmacological space. If we choose the thresholds $Pa > Pi$, $Pa > 30\%$, $Pa > 50\%$ and $Pa > 70\%$, respectively, 900, 866, 825 and 747 different types of biological activities are predicted for at least one compound from the set. Statistics of the upper activities predicted at $Pa > 50\%$ is given in Table I.

Table I shows that at the threshold $Pa > 50\%$, the most probable are the following types of activity: Oxidoreductase inhibitor (predicted for 5209 compounds), Apoptosis agonist (predicted for 5164 compounds), Antiviral (herpes) (predicted for 4947 compounds), Convulsant (predicted for 4482 compounds), Neuroprotector (predicted for 4194 compounds), Vascular (peripheral) disease treatment (predicted for 4138 compounds). In general, a set of compounds submitted for prediction covers a wide pharmacotherapeutic area, including psychotropic, cardiovascular, immunomodulating, antineoplastic,

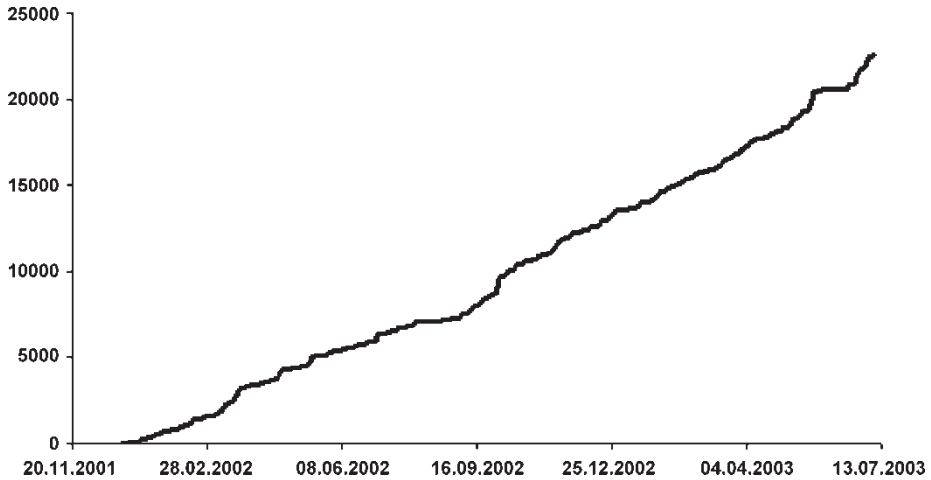

FIGURE 3 Total number of structures sent for prediction to PASS Inet system in Dec. 2001–June 2003.

TABLE I Top thirty types of biological activity predicted for about 23,000 structures submitted to Internet PASS system in the first 18 months of its use

No	*Pa > 50%*	*Pa > 70%*	*Types of activity*
1	5209	649	Oxidoreductase inhibitor
2	5164	2338	Apoptosis agonist
3	4947	303	Antiviral (herpes)
4	4282	1512	Convulsant
5	4194	1011	Neuroprotector
6	4138	970	Vascular (periferal) disease treatment
7	3903	1643	Arrhythmogenic
8	3893	120	Antihypoxic
9	3625	29	Antiviral (picornavirus)
10	3350	365	Fibrinolytic
11	3304	449	Membrane permeability inhibitor
12	3164	555	Myocardial ischemia treatment
13	3122	1819	Phosphatase inhibitor
14	3120	1340	Psychosexual dysfunction treatment
15	3108	1737	Membrane integrity agonist
16	2978	927	Nootropic
17	2964	399	Lysase inhibitor
18	2862	1184	Tyrosine phosphatase inhibitor
19	2775	777	Antineoplastic
20	2775	281	Antileishmanial
21	2743	588	Cardiotoxic
22	2537	204	Interleukin antagonist
23	2479	743	Antiepileptic
24	2448	332	Growth factor antagonist
25	2421	669	Lipid metabolism regulator
26	2397	135	Histamine N-methyltransferase inhibitor
27	2330	949	Colony stimulating factor agonist
28	2319	762	Antiseborrheic
29	2302	279	Autoimmune disorders treatment
30	2285	909	Cardiovascular analeptic

antibacterial, antiviral and many other actions. Thus, this collection of compounds may be used for computer-aided selection of hits with desirable but without unwanted properties.

Such a search can be performed with any set of desirable and unwanted types of biological activity used as criteria. For instance, preparing international project for INTAS[‖] we were looking for a collaborator team which could synthesize potential anxiolytics and cognition enhancers. Analysing compounds submitted for prediction by PASS Inet, we have found that the laboratory of Dr. Fliur Macaev from the Institute of Chemistry, National Academy of Sciences (Moldova) corresponds well to this purpose (preliminary results of this project were recently presented [12]).

Large Scale PASS Predictions Available via Internet

The largest collection of compounds for which PASS predictions are available via Internet is the Open NCI Database that includes about 250,000 highly diverse structures [13]. These compounds from both organic synthesis and natural source extracts have been collected and tested by the National Cancer Institute (NCI) in anticancer and anti-HIV assays for almost 50 years. Since about half of this collection is available as real samples for biological screening, these compounds are a valuable source for new hits and potential leads in many pharmacotherapeutic areas. However, it is extremely expensive to test all these compounds

[‖]URL: [http://www.intas.be]

against each of the thousands of known screens. Therefore, computer-aided prediction is the only method for pre-selection of compounds with desirable and without unwanted properties. On this basis, we have predicted biological activity spectra for all 250,000 compounds from the Open NCI database and presented more than 64 million predictions at the NCI web site.[#] These data are available for the analysis/download in a searchable mode. When the user of the Web service selects the query type "PASS Prediction Range", a separate selector popup window appears, in which the user can scroll through all possible predicted activities. A specific activity has to be selected, and the type of prediction has to be determined (probability of activity [Pa] or inactivity [Pi], respectively). PASS search criteria values have to be specified in probability ranges (in subintervals of 0.0–1.0). To form complex search strategies, PASS search can be combined with any of the numerous search criteria available on this site.[#] A detailed description of how to perform such searches has been published previously [14].

As was already mentioned, for about 43,000 compounds tested in anti-HIV assays from the Open NCI database the use of PASS predictions increases the number of potential anti-HIV agents by a factor from 2 to 17 (at threshold Pa $>$ 10% and Pa $>$ 90%, respectively), compared to the random screening [4]. Varying the queries, one may select the compounds from Open NCI database with desirable properties (biological activities, physical–chemical characteristics, etc.) and those without unwanted ones [4,15], significantly increasing the efficiency of drug R & D.

CONCLUSIONS

1. The new PASS Internet version gives scientists the chance to predict 900 pharmacotherapeutic effects, mechanisms of action, mutagenicity, carcinogenicity, teratogenicity and embryotoxicity for many structures, prior to their synthesis and/or biological testing. The predicted biological activity spectra can be used as criteria for selecting the most prospective compounds with desirable but without unwanted actions.
2. During the first 18 months of PASS Inet use, predictions for about 23,000 structures were obtained, demonstrating significant interest of the scientific community to this kind of service.
3. More than 64 million predictions for 250,000 compounds are available at the NCI web site. This information can be used for selecting most prospective compounds coincided with particular purpose of R & D project.
4. Successful predictions of protein kinase C and tyrosine kinase inhibiting activity [15], antileishmanial and antitrichomonal actions [16] with PASS Inet program have been demonstrated recently.

Acknowledgements

We are most grateful for the support of this work by the Russian Foundation of Basic Research (grant 03-07-90282), U.S. Civil Research and Development Foundation (grant RC1-2064), International Association for the promotion of cooperation with scientists from the New Independent States (grant INTAS 00-0711).

[#]URL: [http://cactus.nci.nih.gov/ncidb2/]

References

[1] Poroikov, V. and Filimonov, D. (2001) "Computer-aided prediction of biological activity spectra. Application for finding and optimization of new leads", In: Holtje, H.-D. and Sippl, W., eds, Rational Approaches to Drug Design (Prous Science, Barcelona), pp 403–407.

[2] Poroikov, V., Filimonov, D., Borodina, Yu., Lagunin, A. and Kos, A. (2000) "Robustness of biological activity spectra predicting by computer program PASS for non-congeneric sets of chemical compounds", *J. Chem. Inform. Comput. Sci.* **40**, 1349–1355.

[3] Anzali, S., Barnickel, G., Cezanne, B., Krug, M., Filimonov, D. and Poroikov, V. (2001) "Discriminating between drugs and nondrugs by Prediction of Activity Spectra for Substances (PASS)", *J. Med. Chem.* **44**, 2432–2437.

[4] Poroikov, V.V., Filimonov, D.A., Ihlenfeldt, W.-D., Gloriozova, T.A., Lagunin, A.A., Borodina, Yu.V., Stepanchikova, A.V. and Nicklaus, M.C. (2003) "PASS biological activity spectrum predictions in the enhanced open NCI database browser", *J. Chem. Inform. Comput. Sci.* **43**(1), 228–236.

[5] Lagunin, A., Stepanchikova, A., Filimonov, D. and Poroikov, V. (2000) "PASS: prediction of activity spectra for biologically active substances", *Bioinformatics* **16**, 747–748.

[6] Filimonov, D., Poroikov, V., Borodina, Yu. and Gloriozova, T. (1999) "Chemical similarity assessment through multilevel neighbourhoods of atoms: definition and comparison with the other descriptors", *J. Chem. Inf. Comput. Sci.* **39**, 666–670.

[7] Poroikov, V., Akimov, D., Shabelnikova, E. and Filimonov, D. (2001) "Top 200 medicines: can new actions be discovered through computer-aided prediction?", *SAR QSAR Environ. Res.* **12**, 327–344.

[8] Stepanchikova, A.V., Lagunin, A.A., Filimonov, D.A. and Poroikov, V.V. (2003) "Prediction of biological activity spectra for substances: evaluation on the diverse set of drugs-like structures", *Curr. Med. Chem.* **10**, 225–233.

[9] Coner, D. and Stevens, D. (1997) Internetworking with TCP/IP. Volume III. Client-server programming and applications–windows socket version (Prentice-Hall).

[10] Polyanskii, A. *A learning guide in CGI programming*, Moscow (Rus.)

[11] Tomson, L. and Welling, L. (2000) Development of web-applications on PHP and MySQL [Russian translation. Moscow: Diasoft, 2000].

[12] Geronikaki, A., Poroikov, V., Saloutin, V., Macaev, F., Babaev, E., Voronina, T., Proenca, F., Dearden, J. and Belzung, K (2003). "Computer-aided selection and biological testing of anxiolytics, anticonvulsants and cognition enhancers in diverse set of chemical compounds". *Abstr. International Symposium on Drud Discovery and Process Research (DDPR-2003)*, Kolhapur, India, I-3.

[13] Voigt, J.H., Bienfait, B., Wang, S. and Nicklaus, M.C. (2001) "Comparison of the NCI Open Database with Seven Large Chemical Structural Databases", *J. Chem. Inf. Comput. Sci.* **41**, 702–712.

[14] Ihlenfeldt, W.-D., Voigt, J.H., Bienfait, B., Oellien, F. and Nicklaus, M.C. (2002) "Enhanced CACTVS Browser of the Open NCI Database", *J. Chem. Inf. Comput. Sci.* **42**, 46–57.

[15] Manallack, D.T., Pitt, W.R., Gancia, E., Montana, J.G., Livingstone, D.J., Ford, M.G. and Whitley, D.C. (2002) "Selecting screening candidates for kinase and G protein-coupled receptor targets using neural networks", *J. Chem. Inf. Comput. Sci.* **42**, 1256–1262.

[16] Delmas, F., Di Giorgio, C., Robin, M., Azas, N., Gasquet, M., Detang, C., Costa, M., Timon-David, P. and Galy, J.-P. (2002) "*In vitro* activities of position 2 substitution-bearing 6-nitro- and 6-amino-benzothiazoles and their corresponding anthranilic acid derivatives against *Leishmania infantum* and *Trichomonas vaginalis*", *Antimicrob. Agents Chemother.* **46**, 2588–2594.