INTERNET SYSTEM PREDICTING THE SPECTRUM OF BIOLOGICAL ACTIVITY OF CHEMICAL COMPOUNDS

A. V. Sadym,[1] A. A. Lagunin,[1] D. A. Filimonov,[1] and V. V. Poroikov[1]

Translated from Khimiko-Farmatsevticheskii Zhurnal, Vol. 36, No. 10, pp. 21 – 26, October, 2002.

Original article submitted July 16, 2002.

Previously [1, 2], we have developed a computer program for predicting the activity spectrum of substances (PASS), which is capable of assessing with a sufficiently high accuracy the probability of manifestation of various types of biological activity of a given substance. The activity prognosis is made by the PASS program proceeding from the structural formula of a given compound and is based on an analysis of the data base including structural formulas and activity characteristics of many known biologically active substances. The program package of PASS Version 1.603 (April, 2002) is capable of predicting 783 activity types based on quantitative structure – activity relationship (QSAR) analysis for a teaching set including more than 45,000 substances.

The PASS program has been successfully employed for several years in a number of Russian and foreign research institutions In order to introduce this facility to a broader circle of specialists engaged in the drug design, we have developed a demonstration Internet version of PASS accessible since July 27, 1998 [3]. During the time period from that date to December 25, 2001, the site registered 435 users from more than 30 countries and performed activity predictions for more than 2800 chemical compounds, which corresponds to an average of 200 – 300 requests per month. The first Internet version of PASS possessed some disadvantages, being limited in the number of predicted activity types (a total of 319) and showing certain instability in the system functioning.

The interest of users in the first demonstration Internet version of PASS stimulated us to develop a new, updated variant of the Internet PASS System for predicting the activity spectrum of substances based on the latest version of the PASS program.

MATERIALS AND METHODS

The operation algorithm of PASS system is based on an analysis of the structural descriptors in the multilevel neighborhood of atoms (MNA descriptors [4]). The set of MNA descriptors is generated based on the structural formula of each compound, presented in the form of a list of atoms forming a given molecule and the list of bonds between these atoms. The results of the prediction are presented to the user in the form of a list of activity types, with probabilities of the presence (Pa) and the absence (Pi) of each particular activity.

An important element of the Internet PASS System is the user interface organized so as to meet the following requirements. First, the web pages have to be correctly reflected by the web-browsers of Internet Explorer (Version 4 and higher) and Netscape (Version 4 and higher). No technologies should be employed which require installing additional programs (such as Flash) on the user's computer. Second, the user interface must be optimized for a screen resolution of 800 × 600 pixels. Although this resolution is now far from being most popular among Internet users, this format is still set in many monitors in Russia and abroad. A site optimized for the 800 × 600 pixel image will always be correctly reproduced on a monitor set to a higher resolution.

In developing and realizing the prognostic algorithm [5], we took into account that the structure – activity relationship database (SAR Base or teaching set), which underlies the operation of PASS program and is loaded into the server memory, occupies in a compressed form a memory volume of about 13 Mbyte. Depending on the random access memory (RAM) volume available, SAR Base can be either rapidly (within a few seconds) loaded into the computer memory, or this process can take several minutes. Therefore, it was important to provide that both the PASS program and SAR Base were always residing in the server memory and that a client application (processing the requests of users) would access the loaded program. Among the possible client-server architectures, we have selected that based on sockets [6]. In determining requirements with respect to system software, the main point was that the products would be available either free (Apache, MySQL Server, PHP) or as a licensed product (Delphi 5.0).

Finally, we have selected the following configuration of the Internet PASS System:

[1] Orechovich Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow, Russia.

0091-150X/02/3610-0538$27.00

(1) Windows 2000 Server (operation system);
(2) Apache (web server);
(3) HTML (web page language);
(4) Common Getaway Interface (CGI) protocol (program sending data from Socket to server; this program and server are written in Delphi 5.0);
(5) PHP scripts (user registration, authentication, web page loading) [8];
(6) MySQL Server (database management system) [8].

RESULTS AND DISCUSSION

The number of predicted activities and the accuracy of prognosis in the Internet PASS System coincides with the corresponding parameters of the original PASS program. Presently, the Internet System (as well as the initial program) is capable of predicting 783 types of biological activity. The complete list of these activities is available on site [9]. Some of the types of predicted activities are as follows:

Biological activity	Prediction accuracy, %
1. Antagonist of urokinase-type-plasminogen activator receptors	99.4
2. Analog of vitamin D	92.2
3. Antibiotic, rifampicins	99.0
4. Antibiotic, taxanes	98.9
5. Antibiotic, carbapenems	98.7
6. Antibiotic, cephalosporins	98.7
7. Antibiotic, penicillins	98.6
8. Antibiotic, β-lactams	98.1
9. Antagonist of histamine H3 receptors	98.0
10. Antagonist of angiotensin AT1 receptors	97.9
11. Antagonist of purinergic P2Y receptors	97.8
12. Antibiotic, naphthyridines	97.8
13. Antibiotic, quinolones	97.8
14. Antimitotic, podophyllotoxin type	97.8
15. Antiviral (cytomegalovirus)	97.8
16. Corticosteroid	97.7
17. Agonist of glucocorticosteroids	97.7
18. Agonist of adenosine A3 receptors	97.7
...	
19. Modulator of cytokines	67.4
20. Nootrope	67.3
21. Histamine release stimulator	67.2
22. Dyskinesia treatment	67.2
23. Spermicide	67.2
24. Neurotrophic factor	66.8
25. Serotonin release factor	66.7
26. Antileishmaniatic agent	66.4
27. Calmodulin antagonist	66.3
28. Adenosinetriphosphatase inhibitor	66.2
29. Neruroprotector	65.8
30. Fibrinolytic	65.6
31. Antiischemic (myocardium)	65.6
32. Alzheimer's disease treatment	65.0
33. inhibitor of phospholipase C	64.9
34. Treatment of psychosexual disorders	64.8
35. Treatment of disseminated sclerosis	63.4
36. Antacid	62.6

The accuracy of prognosis, averaged over all compounds of the teaching set and all activity types, amounts to 85%. The quality of PASS predictions is evaluated based on the sliding control (leave-one-out) procedure, whereby the prognosis is performed for a given chemical compound excluded from the teaching set. For this purpose, the probabilities of errors of the first kind (active compound predicted to be inactive) and second kind (inactive compound predicted to be active) are calculated for each activity type at a selected decision threshold. Then, a threshold value is selected for which errors of the two kinds are equal. These equal values (if exist) characterize the aforementioned prediction accuracy.

An advantage of the Internet PASS System is that free access to the program is ensured for any interested user. There are no restrictions imposed on the operation system type: access to the Internet is the only necessary condition. Having visited the site [9] and being registered, any user can obtain a prognosis of the biological activity spectrum for a compound of interest. In contrast to the commercially available version of the PASS program, the Internet PASS System allows the user to obtain the prognosis for only one compound during each visit on site.

Figure 1 shows an example of the starting page of the Internet PASS System. This page provides brief information about the PASS program and allows the demo version to be loaded (click PASSDEMO. ZIP). The page top panel is divided into three parts, each being a reference. Clicks at the left- and right-hand parts allow a user to enter the front pages of Internet sites of the V. N. Orechovich Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences (Moscow) and the Laboratory of Structure – Function Based Drug Design, respectively. Clicking PASS implies transition to the front page of the Internet PASS System.

The Internet site [9] presents a description of the PASS program, including sections devoted to concepts such as biological activity spectrum and its description, the main elements of PASS, application examples, and references to publications employing PASS predictions and those devoted to the PASS system proper. Some of the articles are given in full variants which can be reloaded to the user's computer.

The left-hand side panel of the screen represents a menu listing the heads of all sections in gray cells separated by orange lines. The activated cell turns green, while that indicated by the mouse marker becomes orange. Any desired section is entered upon clicking at the selected (orange) cell. The right-hand side panel performs identification functions. Not previously registered, a user has to click the REGISTRATION cell situated under the ENTER button. The fields to be filled by the user are indicated by asterisk (*). Should the necessary fields remain unfilled, or the user's name contain less than four letters, or should it coincide with the name

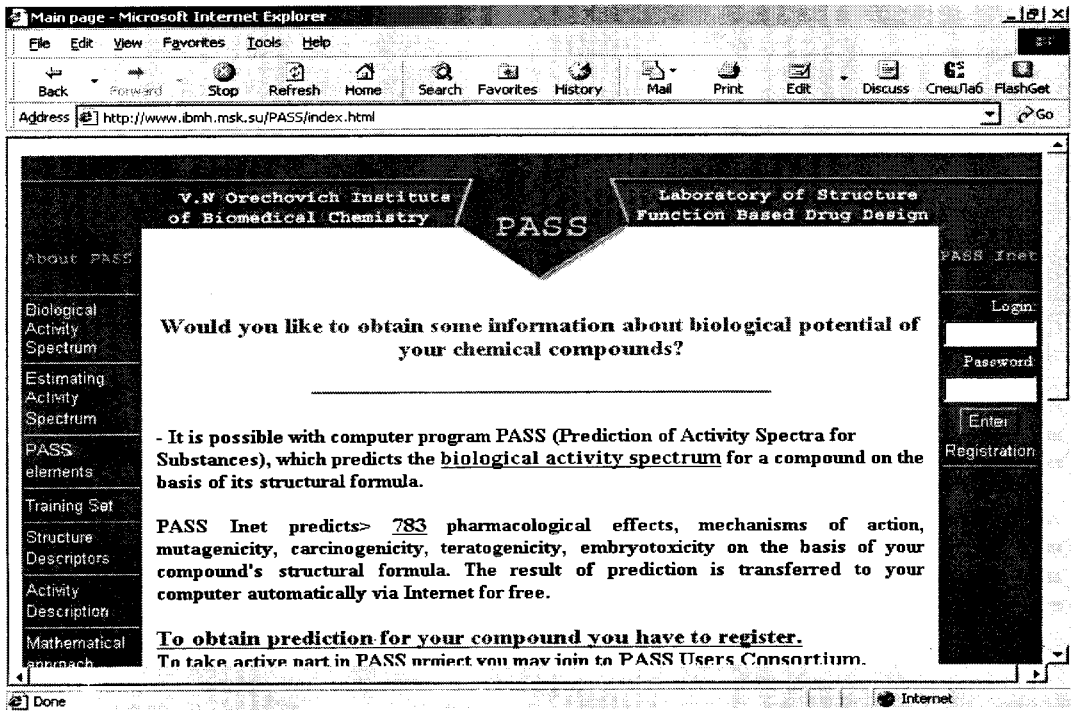
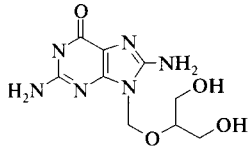


Fig. 1. Starting page of the Internet PASS System (see the text for explanations).

of another user or message will be obtained; otherwise the user will receive the notification of successful registration. During registration, the user gives the system his e-mail to which, 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 researcher having access to Internet can be registered in the Internet PASS System. In particular, this can be a chemist who has synthesized a new substance, or is planning such synthesis, or performs biological activity screening of some chemical compounds. For example, this user wants to check for possible biological activity of 2,8-diamino-1,9-dihydro-9-[[2-hydroxy-1-(hydroxymethyl)ethoxy]methyl]-6H-purin-6-one



For this purpose, the user has to enter the prediction page which is accessed upon introducing the name and password. Erroneous introduction leads to an error message. Successful authentication opens the prediction page depicted in Fig. 2.

In the present version, description of the structures of chemical compounds in the PASS system is based on two-dimensional structural formulas of substances. These formulas can be introduced into the computer, for example, with the aid of an ISIS/Draw formula editor (MDL Information Systems, Inc.). The standard representation of information for each molecule is provided by [*.mol] files. In the Internet PASS System, the user can either obtain a prediction for the prepared Mol file, or introduce the structure of interest in the form of Marvin-applet (developed by ChemAxon Ltd., http://www.chemaxoncom/marvin). The format of the Mol file must correspond to that of ISIS 2.0 [10].

If the user enters the prediction page for the first time, or the USE APPLET mark is activated, the Marvin-applet is loaded to provide for drawing of the structure of a chemical compound. By means of the applet, the user can introduce, edit, or remove the structures. Drawn by the user in the main window, all changes introduced into the structure are also reflected in an auxiliary window. Introduced in the form of ISIS/Draw chemical editor, the structure (e.g., 2,8-diamino-1,9-dihydro-9-[[2-hydroxy-1-(hydroxymethyl)ethoxy]methyl]- 6H-purin-6-one) has to be exported into Mol format. To this end, the structure is highlighted, the File → Export → Mol-file path is selected in Menu, the catalog and file name are indicated, and the file is stored. On access to the Internet PASS System, the user clicks BROWSE and selects the file.

Alternatively, the user can introduce the structure of the compound of interest (e.g., 2,8-diamino-1,9-dihydro-9-[[2-hydroxy-1-(hydroxymethyl)ethoxy]methyl]-6H-purin-6-one into the applet window. An advantage of using the applet form is the interactive dialog: upon changing the structure drawing, user need not perform the file save and open operations. It user wishes to employ only ready-to-use structures

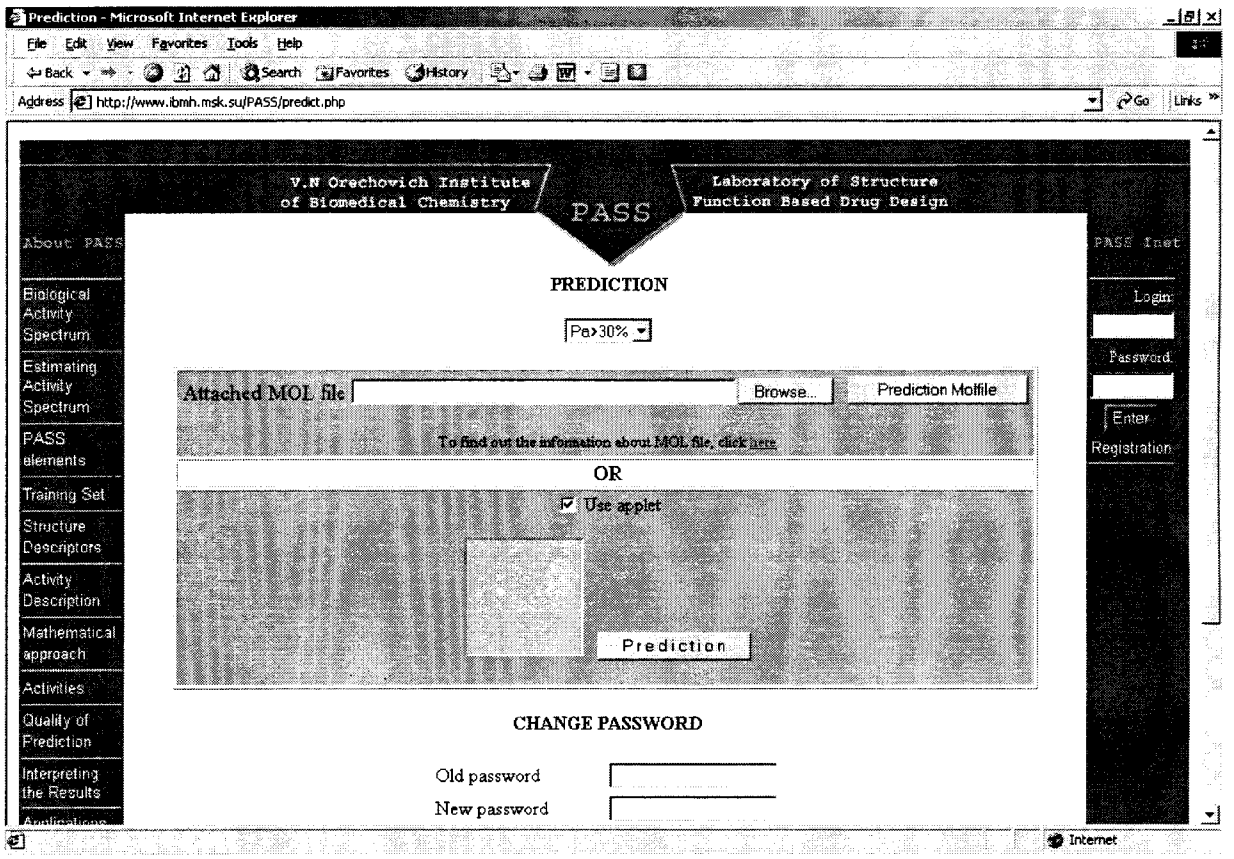


Fig. 2. Prediction page of the Internet PASS System (see the text for explanations).

stored in files, without recourse to the applet for drawing new structures, he has only to inactivate the USE APPLET mark. The page will be renewed and the applet will not be loaded in what follows. Once the applet is required again, the USE APPLET mark is activated to renew the page and load the drawing program.

After introducing a structure in the applet form or by indicating the file and pathway, the user can select a threshold value for the prognosis output in the list of activities (default value, Pa > 30%) and click a button necessary to activate the prognosis function: either PREDICTION (for the structure introduced via applet) or PREDICTION PROFILE (for the structure stored in a Mol file). The prediction is obtained in a new browser window (Fig. 3).

In interpreting the results obtained, the user can rely upon the following principles.

(i) Some descriptors of the chemical compound to be predicted can appear as new for the program dictionary (i.e., such fragments are not present in any molecule of the teaching set). If all descriptors are of this kind, the given compound has no fragments in common with compounds of the teaching set and, hence, the spectrum of biological activity cannot be predicted by the PASS system. Once the number of new descriptors is relatively large (exceeding three), the results of predictions should be considered only as rough orientation.

(ii) If a certain type of activity was predicted for Pa > 0.7, the given compound will most likely exhibit this activity in experiment; however, the probability that this substance is merely an analog of some well-known drug is also quite large.

(iii) In the case of $0.5 < Pa \leq 0.7$, there is still a rather large probability that the given compound will show the activity in experiment, but similarity to well-known compounds is less probable.

(iv) If $Pa \leq 0.5$, a probability that the given compound will show this activity in experiment decreases. However, should this activity be manifested, the compound can prove to be a principally new base structure (new chemical entity, NCE).

For example, the results of prognosis for 2,8-diamino-1,9-dihydro-9-[[2-hydroxy-1-(hydroxymethyl)ethoxy]methyl]-6H-purin-6-one (Fig. 4) allow the user to suggest that this compound possesses antiviral (Pa = 0.873) and antiinfectious (anti-HIV) (Pa = 0.875) properties. Indeed, the given compound was reported [11] to produce the antiviral effect – in agreement with the Pa = 87.3% probability predicted by the Internet PASS System.

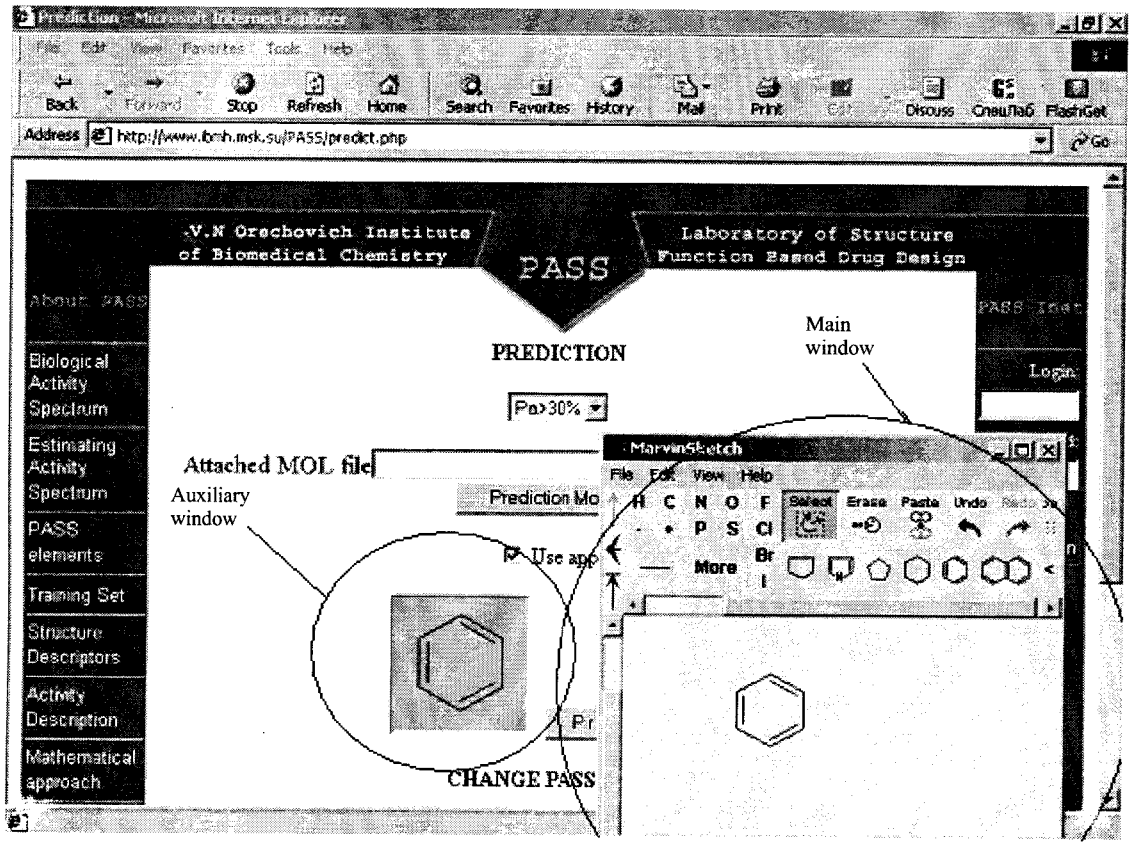
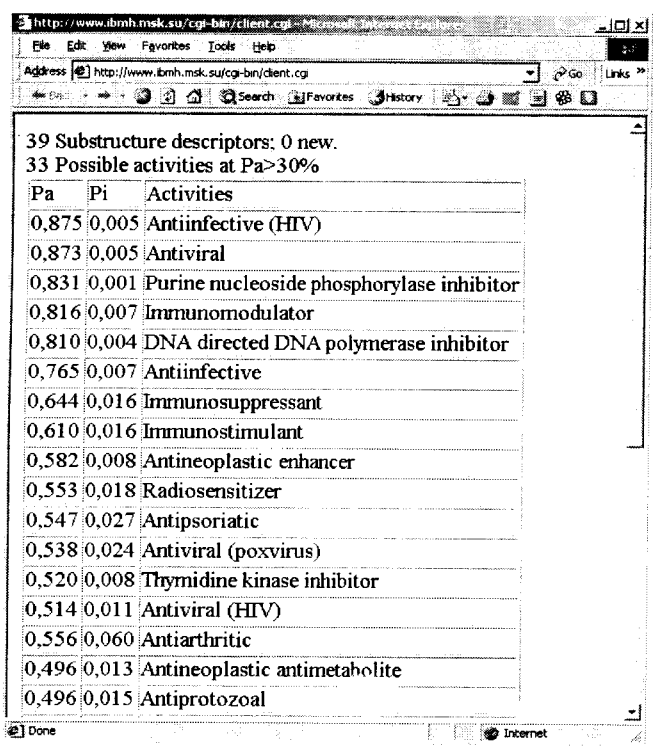


Fig. 3. Main and auxiliary windows of the Marvin-applet (see the text for explanations).



39 Substructure descriptors: 0 new.
33 Possible activities at Pa>30%

Pa	Pi	Activities
0,875	0,005	Antiinfective (HIV)
0,873	0,005	Antiviral
0,831	0,001	Purine nucleoside phosphorylase inhibitor
0,816	0,007	Immunomodulator
0,810	0,004	DNA directed DNA polymerase inhibitor
0,765	0,007	Antiinfective
0,644	0,016	Immunosuppressant
0,610	0,016	Immunostimulant
0,582	0,008	Antineoplastic enhancer
0,553	0,018	Radiosensitizer
0,547	0,027	Antipsoriatic
0,538	0,024	Antiviral (poxvirus)
0,520	0,008	Thymidine kinase inhibitor
0,514	0,011	Antiviral (HIV)
0,556	0,060	Antiarthritic
0,496	0,013	Antineoplastic antimetabolite
0,496	0,015	Antiprotozoal

Fig. 4. A typical result of the PASS prognosis (for 2,8-diamino-1,9-dihydro-9-[[2-hydroxy-1-(hydroxymethyl)ethoxy]methyl]-6H-purin-6-one).

In this way, the user can assess the biological activity of any chemical compound and, having obtained the results of prognosis, reject substances predicted with high probability to possess undesired properties (e.g., carcinogenicity). Such preliminary computer prognosis can significantly reduce the cost of R&D of new drugs and/or other biologically active compounds. Besides, new types of activity can be found in existing compounds [1 – 3, 12, 13]. This knowledge may significantly increase search efficiency for new biologically active substances. For example, an analysis of the results of PASS predictions for 42,689 compounds from the NCI (NIH, USA) database (http://dtp.nci.nih.gov/docs/aids/aids data.html) studied for anti-HIV activity showed that PASS increases the number of potential anti-HIV agents by a factor from 2 (for Pa > 10%) to 17 (for Pa > 90%) as compared to the case of random screening.

As can be seen from the visit statistics, the Internet PASS System is visited on the average by ten users per day. The total number of requests to the system exceeded 6000, which is about two times the analogous number (2800) for the first demo version after its functioning for two years. The average frequency of requests per week fluctuates from 70 to 850. During the first half year of the Internet PASS System functioning, more than 200 users from various countries have been registered.

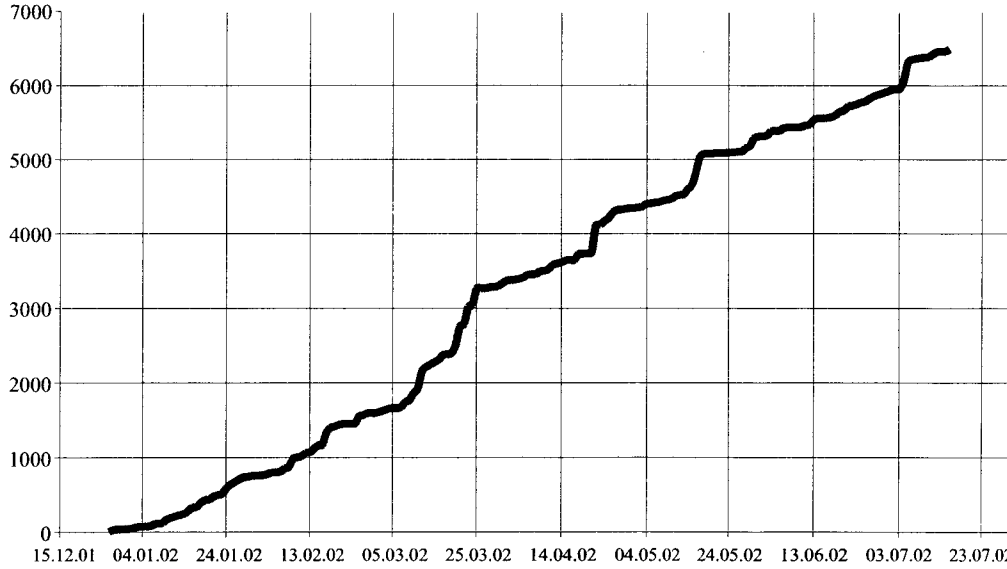


Fig. 5. Plot of total number of requests versus time of functioning (date) of the Internet PASS System.

There is still a probability that, accomplishing the PASS prognosis of the activity spectrum for a given compound, the user will not obtain a satisfactory result because the available teaching set contains no compounds possessing the biological activity type of interest. Developers of the PASS program spent considerable effort for many years to gain and add new information to the PASS teaching set. Obviously, it is a very difficult task to collect such information about all possible types of biological activity. Each user, dealing with a restricted number of substances and a limited number of activity types, collects information concerning this very special field of his interest. Therefore, if numerous PASS users would join the process of completing the SAR Base, the joint effort could create a much more powerful teaching set ensuring a higher quality of predictions. This is the idea of the Association of PASS Users, which would unite to provide for continuously increasing the quality of the teaching set. Association members will be regularly provided with updated PASS versions of increasing power, ensuring a higher quality of predictions of the biological activity spectra of the compounds studied. More detailed information about the Association of PASS Users can be found on site [9].

In parallel to the PASS program development, the Internet PASS System will also be continuously improved so that users of this system would be able to obtain more reliable predictions and improve the quality of their investigations.

REFERENCES

1. V. Poroikov and D. Filimonov, *Rational Approaches to Drug Design*, H.-D. Holtje and W. Sippl (eds.), Prous Science, Barcelona (2001), pp. 403 – 407, .
2. V. V. Poroikov and D. A. Filimonov, *Nitrous Heterocycles and Alkaloids 1* [in Russian], Iridium Press, Moscow (2001), pp. 123 – 129.
3. A. Lagunin, A. Stepanchikova, D. Filimonov, and V. Poroikov, *Bioinformatics*, **16**(8), 747 – 748 (2000).
4. D. Filimonov, V. Poroikov, Yu. Borodina, and T. Gloriozova, *J. Chem. Inf. Comput. Sci.*, **39**(4), 666 – 670 (1999).
5. D. Marca and D. McGowen, *Methodology of Structure Analysis and Design* [Russian translation], Meta Tekhnologiya, Moscow (1993).
6. D. Comer and D. Stevens, *Internetworking with TCP / IP. Volume III. Client-Server Programming and Applications-Windows Sockets Version*, Prentice Hall (1997).
7. A. Polyanskii, *A Learning Guide in CGI Programming* [in Russian], Moscow (2000).
8. L. Tomson and L. Welling, *Development of Web Applications on PHP and MySQL* [Russian translation], Diasoft, Moscow (2001).
9. http: // www.ibmh.msk.su / PASS.
10. CTfile Formats. MDL (http: // www.mdl.com / downloads / literature / ctfile.pdf).
11. M. Negwer and H-G. Scharnow, *Organic-Chemical Drugs and Their Synonyms*, Eighth, Extensively Enlarged Edition (Volume 1), p. 390 (2001).
12. K. Enselein, V. K. Gombar, and B. W. Blake, *Mutat. Res.*, **305**, 47 – 61 (1994).
13. J. E Ridings, M. D Barratt, R. Cary, et al., *Toxicology*, **106**, 267 – 279 (1996).