OPTIMIZATION OF THE SYNTHESIS AND PHARMACOLOGICAL CHARACTERIZATION OF SUBSTANCES ON THE BASIS OF COMPUTER AIDED PROGNOSIS OF BIOLOGICAL ACTIVITY SPECTRA

V. V. Poroikov,[1] D. A. Filimonov,[1] A. V. Stepanchikova,[1] A. P. Budunova,[2] E. V. Shilova,[2] A. V. Rudnitskikh,[1] T. M. Selezneva,[2] and L. V. Goncharenko[2]

Translated from Khimiko-Farmatsevticheskii Zhurnal, Vol. 30, No. 9, pp. 20 – 23, September, 1996.

Original article submitted April 23, 1996.

Total expenditures for the research and development (R&D) of a single drug are estimated abroad at 230 – 350 million US dollars [1, 2]. It is known that an attempt to develop a new drug is related to a high risk of obtaining a negative result caused by unpredicted side effects, toxicity, etc. The possibility to predict the main and side effects at early stages of the drug development process may allow us to optimize the work and reduce expenses and risks.

We have developed a computer system for prediction of the activity spectrum of substances (PASS) proceeding from their structural formulas, which offers new possibilities in assessing the pharmacological effects, mechanisms of action, and specific toxicities of substances [3, 4].

The prognosis is based on a "comparison" of the structure of a new chemical compound with the structures of known drugs and biologically active substances (BAS). The system operation resembles considerations of a chemist or pharmacist, based on the principle that "substances with like structures produce similar effects." However, formalization of the problem and the possibility of operation within large volumes of data must provide a much higher accuracy and reliability of the computer-aided analysis. A comparison of the quality of predictions made by the PASS computer system and the results of analyses made by ten expert specialists for the same sampling set of potential drugs showed that the computer prognosis accuracy was 3 times that of the human forecast [5].

At the same time, special experiments are required for the evaluation of accuracy, reliability, and efficiency of prognosis made for a large number of activities on the qualitative level (presence/absence of a given activity type). Below we describe the results of such experiments, showing a sufficiently high accuracy, reliability, and efficiency of the computer prognosis, which allows us to recommend the PASS system for wide practical use.

[1] Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow, Russia.
[2] All-Russia Scientific Center for Safety Testing of Biologically Active Substances, Ministry of Health of the Russian Federation, Kupavna, Moscow region, Russia.

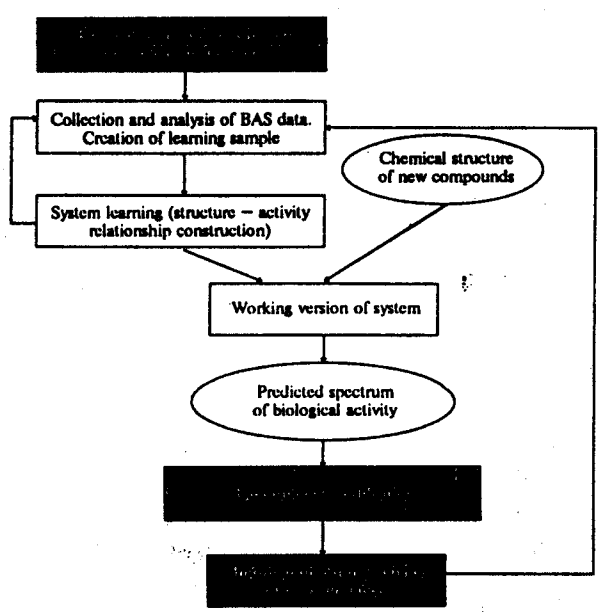

Fig. 1. Schematic diagram illustrating the functioning of a computer system predicting the biological activity of substances on the basis of structural formulas (shadowed parts refer to external stages).

GENERAL DESCRIPTION OF THE PASS SYSTEM

Figure 1 shows a schematic diagram illustrating functioning of the computer system for prediction of the activity spectrum of substances (PASS).

As is seen from the scheme, the PASS system must be permanently developed, at least in the part concerning updating of the learning sample. This is achieved by information search in the published literature and the chemical and biomedical computer data banks [6].

The idea of using computer prognosis for the screening purposes was originally suggested by Avidon [7]. The PASS system differs from analogous systems [8, 9] by the extended list of predicted activity types, the possibility to introduce chemical information in the form of structural formulas [10] convenient for the practical chemists, automatic coding of the chemical structure by fragment codes of the substructure superposition (FCSS) [11], and using a new, highly stable algorithms for the structure – activity relationship determination.

The information data bank, which serves as a basis for the learning sample, now contains data for more than 12,000 BAS including the overwhelming majority of both the established drugs and the preparations currently in the stage of clinical evaluation.

The PASS system is described in much detail elsewhere [3, 4] and some application results were reported previously [12, 13]. The current version is PASS 4.0. Figure 2 gives an example of the activity prognosis for 3-(2-propylamino)-1-phenoxy-2-propanol, a new β_1-adrenoreceptor antagonist [14], made with the aid of the PASS 4.0 system.

As is seen, the known activity type (β_1-adrenoreceptor antagonism) is predicted with a high probability. At the same time, our system predicts that the substance possesses a very broad activity spectrum that may lead to significant side effects. This circumstance does not allow 3-(2-propylamino)-1-phenoxy-2-propanol to be used as a drug. In particular, it is highly probable that this compound is a cardiodepressant (98%), antiarrhythmic agent (97%), peripheral vasodilatory agent (96%), etc.

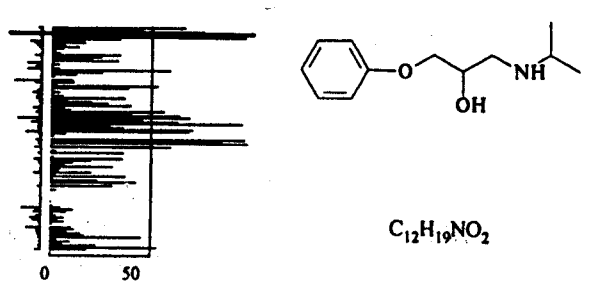

Fig. 2. Predicted spectrum of biological activity of 3-(2-propylamino)-1-phenoxy-2-propanol, a new β_1-adrenoreceptor antagonist [14].

ASSESSMENT OF RELIABILITY OF THE PASS PROGNOSIS

Criteria of the Prognosis Quality. Unlike the traditional quantitative structure – activity relationship (QSAR) analysis, the PASS system is based on the qualitative presentation of the activities of substances in the learning sample. The prognosis consists in evaluation of the probability that the compound under consideration exhibits each particular type of activity. The decision on recognizing the given substance as active or inactive with respect to each activity type is taken on the basis of referencing the probability estimates to the corresponding threshold values. Once the estimate exceeds the threshold level, the compound is referred to as active. At this stage, there are two possible types of error, whereby an active substance is classified as inactive (error of the 1st kind) or inactive substance is classified as active (error of the 2nd kind).

According to the theory of decision making, the threshold values must be determined on the basis of minimized risk of taking a wrong decision. Within the framework of the PASS system, the losses caused by errors of the 1st and 2nd kind are determined by specific features of the problem to be solved and cannot be predicted in advance. Nor do we know the a priori probabilities of the occurrence of each activity type, which can depend to a considerable extent on the assessed series of compounds. As a result, we have selected the threshold values of probability estimates on the basis of analysis of the quality of the activity spectrum predictions by PASS using the criterion of equal probability of the error of 1st and 2nd kinds. This is an extremal criterion in the sense that the probability of correct decision in each particular case can be only higher than that according to the criterion. For the evaluation of the quality of prognosis of the entire activity spectrum, we have determined an average error probability over all 114 activity types under consideration, and three error distribution quantiles corresponding to the 10, 50, and 90% probability levels. Below we present the results for the sliding and cross control of the prognosis quality, based on the criterion of equality of errors of the 1st and 2nd kinds.

Sliding Control. One of the methods widely employed for validation of the structure – activity relationships is the sliding control based on the leave-one-out cross validation, which is essentially as follows. Each compound of the series studied is sequentially excluded from the learning sample, and the activity spectrum is predicted by comparison of the given structure with those of the remaining substances. The results of this prognosis are compared to the known data on the biological activity of this compound.

The sliding control is carried out by the PASS system immediately in the course of learning [3, 4]. Theoretical estimates showed that this must provide statistical stability of the prognosis. The results of sliding control show that the average accuracy of predictions is 78%, the accuracy of predicting the "best" activity (thyroid hormone) amounts to 99%, and the accuracy of predicting the "worst" activity (teratogen or embryotoxic agent) is 60%. Note that 10% of the 114 predictable activity types are evaluated with an accuracy exceeding 88%, 50% are predicted with an accuracy better than 77%, and 90% are predicted with an accuracy not lower that 70%. The probability to accidentally "guess" any of the 114 predictable activities is less than 2%.

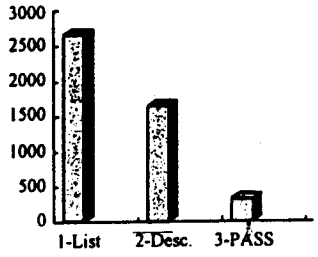

Fig. 3. Histogram of the number of tests necessary within three various strategies: *1*) by list; *2*) by decreasing activity occurrence frequency; *3*) by PASS prognosis.

Cross Control. As is known, the sliding control may give overstated rates for the quality of predictions, especially in the case of a large amount of "doubling" information in the learning sample. The problem is sometimes solved by passing from the "leave-one-out" procedure to the "leave-10%-out" validation method. However, for the learning sample of 9314 substances and 114 predictable activity types, used in the current PASS version, the latter procedure would require very large computation resources. For this reason, an independent estimate of the accuracy and reliability of prognosis was made on the basis of the cross control procedure, which was as follows.

All the learning sample is randomly (using the random number generator) divided into two approximately equal parts. Each of these subsamples is used as the learning sample, while the other is employed as a test sample. In our case, the first subsample contained 4862 substances, and the second, 4632 substances. Table 1 shows the results of the sliding and cross control for the two subsamples, in comparison with the sliding control data for the entire learning sample.

In Table 1, the values F10, F50, and F90 represent the quantiles of error distribution for all predictable activity types on the confidence levels of 10, 50, and 90%.

As is seen from Table 1, the average predicting accuracy is approximately the same for various combinations of substances, slightly increasing with the sample volume (this points to the necessity of constant renewal of the data bank). The data show evidence of the stability of algorithm (in the statistical sense) used in the PASS system. Thus, a high reliability of the prognosis based on this algorithm, which was theoretically justified previously [15], is also confirmed by the results of cross control.

TABLE 1. Data on the Sliding and Cross Control

Experiment	Error			
	Average	F10	F50	F90
Cross prognosis	0.251	0.138	0.256	0.352
Sliding control for two subsamples	0.248	0.141	0.254	0.344
Sliding control for the whole learning sample	0.224	0.120	0.233	0.300

ASSESSMENT OF THE EFFICIENCY OF THE PASS SYSTEM

The efficiency of predictions made by the PASS system was assessed for the previous version (3.05), which was based to a considerable extent on the learning sample corresponding to the handbook [16] published in 1987. We have used a more updated source [17] published in 1994 to form an independent test sample of 50 substances representing about 100 chemical classes, possessing 36 types of activity. Note that the test sample was taken from [17] randomly, the only requirement being the presence of activity predictable by the PASS 3.05 version. Thus, the learning and test samples were separated by a time period of 7 – 10 years, which assumed relative mutual independence of the two data.

Let us imagine that nothing is known of the biological activity of these 50 substances and consider the possibility to study this sample with respect to all the 114 predictable types of activity [4]. We may perform the test using various strategies: (1) systematic testing, according to the list of activity types available [4]; (2) descending method, by decreasing frequency of occurrence of various activity types in the learning sample [4]; and (3) PASS algorithm, by decreasing PASS-predicted probability of the manifestation of a given activity type. Then we may calculate the total number of trials necessary to establish the actual activity of a substance by each of the three methods. These estimates are presented in Fig. 3.

As is seen, using the PASS prognosis for determination of the activity of substances reduces the number of tests to 1/5 – 1/8 of that for the other strategies, thus saving both time and finances.

The results of experiments carried out in this work show that the accuracy, reliability, and efficiency of predictions of the spectrum of biological activity, offered by our computer-aided PASS system, is sufficiently high and the system can be recommended for a wide practical application.

Obviously, PASS cannot assess all the possible properties of any compound, but the system is open to further development, which allows a special learning sample to be created that would meet the needs of any organization or an individual specialist. Examples of such specialized systems are offered by systems predicting antiamnesic [18] and antiulcer [19] activities.

Our system is capable of predicting both the main and the side pharmacological effects, mechanisms of drug action, and specific toxicities (carcinogenic, mutagenic, teratogenic). Naturally, the final decision concerning the interpretation and use of predicted activity must be made by a qualified expert capable of taking into account additional information.

This work was partly supported by the State Scientific-Technological Program "Creation of New Drugs by Methods

of Chemical and Biological Synthesis," Direction 04 "Computer Design of New Drugs," Project 04.02.06.

REFERENCES

1. D. Bartling and H. Hadamic, *Development of a Drug. Its a Long Way from Laboratory to Patient*, Wurselen (1990).
2. J. G. Topliss, in: *Computer Aided Drug Design in Industrial Research*, Springer, Berlin (1994), pp. 11 – 38.
3. D. A. Filimonov, V. V. Poroikov, E. I. Karaicheva, et al., *Eksp. Klin. Farmakol.*, **58**(2), 56 (1995).
4. D. A. Filimonov and V. V. Poroikov, in: *Bioactive Compound Design: Possibilities for Industrial Use*, BIOS Sci. Publ., Oxford (1996),. pp. 47 – 56.
5. V. V. Poroikov, D. A. Filimonov, and A. P. Budunova, *Nauchno-Techn. Inform.*, Ser. 2, No. 6, 11 (1993); *Automatic Documentation and Mathematical Linguistics*, Allerton, **27**(3), 40 (1993).
6. V. V. Poroikov and E. I. Karaicheva, *Drugs: Economics, Technology, and Production Perspectives. A Review Information*, NPO Medbioekonomika, no. 4, Moscow (1990).
7. V. V. Avidon, *Khim.-Farm. Zh.*, **8**(8), 22 (1974).
8. V. V. Avidon, V. S. Arolovich, V. G. Blinova, et al., *Khim.-Farm. Zh.*, **16**(3), 321 (1983).
9. A. B. Rozenblit and V. E. Golender, *Logical-Combinatory Methods in Drug Design*, Zinatne, Riga (1984).
10. K. P. Khazanovskii, *Itogi Nauki Tekhn., Ser. Inform.*, Vol. 15, VINITI, Moscow (1991), p. 136.
11. A. I. Leibov, *Itogi Nauki Tekhn., Ser. Inform.*, Vol. 15, VINITI, Moscow (1991), p. 141.
12. E. M. Sergeeva, S. A. Minina, D. A. Filimonov, et al., *Byull. VNTs BAV*, No. 1, 64 (1993).
13. V. V. Poroikov, A. P. Budunova, V. P. Shamshin, et al., *Byull. VNTs BAV*, No. 1, 39 (1994).
14. S. Louis, T. L. Nero, D. Iakovidis, et al., *Abstracts of Papers, Conf. molecular Design Down Under*, Cairns (Australia), POS C-17 (1995).
15. D. A. Filimonov, *Abstracts of Papers, The 2nd Russian National Congress "Patient and Drug,"* Moscow (1995), p. 62.
16. M. Negwer, *Organic-Chemical Drugs and Their Synonyms*, Akademie, Berlin (1987), pp. 1 – 3.
17. J. Prous, *Drug Year Book*, Prous Sci. Publ., Barcelona (1994).
18. D. A. Filimonov, V. A. Trapkov, A. P. Budunova, et al., *Abstracts of Papers, The 2nd Russian National Congress "Patient and Drug,"* Moscow (1995), p. 62.
19. D. A. Filimonov, V. V. Poroikov, A. P. Budunova, et al., *Abstracts of Papers, Conf. "Design of Bioactive Compounds: Possibilities for Industrial Applications,"* Sci. Publ., London(1995), pp. 26.