BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 8 2000
Pages 747–748

PASS: prediction of activity spectra for biologically active substances

Alexey Lagunin, Alla Stepanchikova, Dmitrii Filimonov and Vladimir Poroikov

Laboratory of Structure-Function Based Drug Design, Institute of Biomedical Chemistry RAMS, Moscow, Pogodinskaya str., 10, 119832, Russia

Received on November 26, 1999; revised on December 11, 1999; accepted on March 21, 2000

Abstract

Summary: *The concept of the biological activity spectrum was introduced to describe the properties of biologically active substances. The PASS (prediction of activity spectra for substances) software product, which predicts more than 300 pharmacological effects and biochemical mechanisms on the basis of the structural formula of a substance, may be efficiently used to find new targets (mechanisms) for some ligands and, conversely, to reveal new ligands for some biological targets. We have developed a WWW interface for the PASS software. A WWW server for the on-line prediction of the biological activity spectra of substances has been constructed.*

Availability: *http://www.ibmh.msk.su/PASS/*

Contact: *pass@ibmh.msk.su*

Bioinformatics is progressing from the mere analysis of nucleic acid and amino acid sequences to the search for new targets and ligands as leads for new drugs. There are cases where a ligand, capable of binding to one protein, binds to a homologous (or similar) protein. The similarity of targets can be transferred to the similarity of drug-like ligands, and some new ligands can be found on the basis of structure–activity relationships.

We use the concept of 'biological activity spectrum' for a substance. It is the list of activity names, which reflects the result of the interaction of the chemical substance with different biological entities (Filimonov *et al.*, 1996).

Our goal is to provide full information on the biological activities of any substance. It is this aim that determined the choice of chemical substance representation and the mathematical methods on which the provided algorithm is based.

Structure descriptors, which we call 'multilevel neighbourhoods of atoms' (MNAs) (Filimonov *et al.*, 1999), have been suggested for chemical structure representation. This representation does not specify the bond types and includes the hydrogens according to valencies and partial charges of atoms. A set of MNA descriptors for a molecule is generated recursively:

1. The zeroth level of the MNA descriptor $D_i^0 = A_i$ is the designator A of the ith atom type;
2. The $(N+1)$th level of the MNA descriptor $D_i^{N+1} = A_i(D_{i1}^N D_{i2}^N \ldots D_{ij}^N)$ is the substructure notation, where D_{ij}^N is the Nth level of MNA descriptor for the jth neighbour atom of the ith atom.

This process can be continued iteratively for second, third, etc. neighbours of the atom. We use second-level descriptors in the present version of PASS.

The same set of descriptors is used for the analysis and prediction of all kinds of biological activity. MNA descriptors can also be used to predict the quantitative properties of a substance, such as boiling point or mutagenicity (Filimonov *et al.*, 1999).

The *Algorithm of prediction* is started by generating MNA descriptors for the substance. Then the primary estimates Pr_j of the probability of each activity j are calculated as

$$Pr_j = (1 + (s_j - s_{0j})/(1 - s_j s_{0j}))/2,$$

where

$$s_j = \mathrm{Sin}(u_j/m), \quad s_{0j} = \mathrm{Sin}(u_{0j}/m),$$
$$u_j = \Sigma_i \mathrm{ArcSin}\{r_i(2p_{ij} - 1)\},$$
$$u_{0j} = \Sigma_i \mathrm{ArcSin}\{r_i(2p_j - 1)\},$$

m is the number of descriptors for the compound under prediction;
$r_i = n_i/(n_i + 0.5/m)$ is the regulating factor;
$p_j = n_j/n$ is the estimate of the *a priori* probability of activity j;
$p_{ij} = n_{ij}/n_i$ is the estimate of the conditional probability of the activity j for the descriptor i;
n is the total number of compounds in the training set;
n_i is the number of compounds containing descriptor i;
n_j is the number of compounds revealing activity j;
n_{ij} is the number of compounds containing descriptor i and revealing activity j.

© Oxford University Press 2000

The prediction result is returned in the form of a table containing the list of biological activities with appropriate probability values-i.e. the likelihood for the given activity to be either revealed ($Pa_j = EF_j(Pr_j)$) or not revealed ($Pi_j = ES_j(Pr_j)$). $EF_j(CP)$ and $ES_j(CP)$ are estimates of the first and second kind of error probability respectively as a function of the cutting point CP.

The predictive accuracy varies between different biological activities. There is an appropriate table on our website (*Activities*) where the maximal error of prediction for each activity is shown. The quality of predictions is the main criterion of the program efficiency. The mean accuracy of the predictions is about 89% (leave-one-out cross-validation). In LOO cross-validation every compound was consequently excluded from the training set and its types of activity were predicted by PASS trained vs. the remaining compounds.

The Internet version of the program, PASS Inet, contains about 31 000 biologically active substances in the training set and predicts biological activity spectra for 319 types of pharmacological effects, action mechanisms and specific toxicities.

To obtain the biological activity spectrum of a substance, the user should send via the Internet a standard Molfile, which may be prepared with the ISIS/Draw chemical editor (http://www.mdli.com), or they may draw a structural formula using the Marvin applet (ChemAxon Ltd. http://www.chemaxon.com).

From the 'Prediction Results' window the user obtains the total number of chemical descriptors of the substance. Also reported are the number of descriptors which are completely new compared with the descriptors of substances from the PASS training set and comments on the interpretation of prediction results'.

If $Pa > 0.7$, the substance is very likely to exhibit the activity in experiment, but the chance of the substance being the analogue of a known pharmaceutical agent is also high.

If $0.5 < Pa < 0.7$, the substance is likely to exhibit the activity in experiment, but the probability is less, and the substance is unlike known pharmaceutical agents.

If $Pa < 0.5$, the substance is unlikely to exhibit the activity in experiment. However, if the presence of this activity is confirmed in the experiment the substance might be a new chemical entity.

Figure 1 shows the predicted biological activity spectrum for nipradilol as an example of PASS Inet usage. As shown in Figure 1, nipradilol has 50 different chemical descriptors including one new descriptor (which is absent in the PASS Inet training set) and the predicted biological activity spectrum contains 26 types of activities. Looking through the predicted biological activity spectrum of nipradilol one cannot help noticing that on the one hand nipradilol is a beta-1 adrenoreceptor antagonist, but on the other hand it is a beta-1 adrenoreceptor agonist. Such ambiguities indicate that the substance does interact with beta-1 adrenoreceptor; however, the mechanism of its intrinsic action cannot be elucidated by the PASS program.

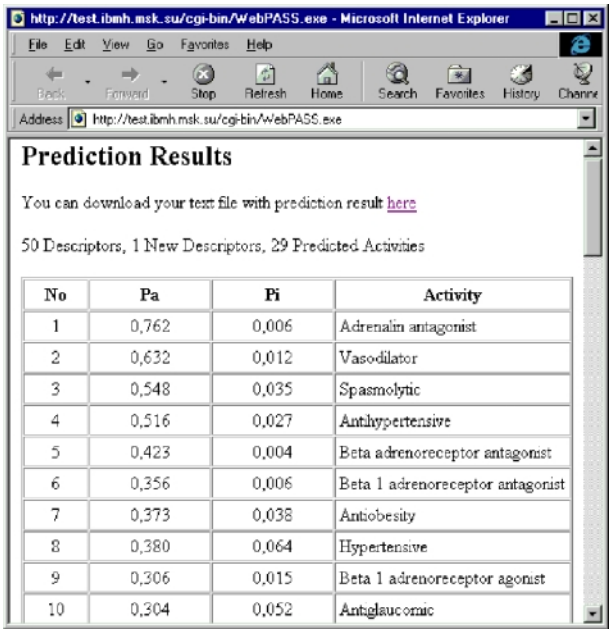

Fig. 1. Example output page of the PASS Inet on the WWW, showing the predicted biological activity spectrum for nipradilol.

References

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.

Filimonov,D. and Poroikov,V. (1996) Bioactive compound design: possibilities for industrial use. *PASS: Computerized prediction of biological activity spectra for chemical substances*. BIOS Scientific Publishers, pp. 47–56.