# PASS Biological Activity Spectrum Predictions in the Enhanced Open NCI Database Browser

Vladimir V. Poroikov,[‡] Dmitrii A. Filimonov,[‡] Wolf-Dietrich Ihlenfeldt,[#] Tatyana A. Gloriozova,[‡] Alexey A. Lagunin,[‡] Yulia V. Borodina,[‡] Alla V. Stepanchikova,[‡] and Marc C. Nicklaus*[,†]

Laboratory of Structure-Function Based Drug Design, V.N. Orekhovich Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences, 10 Pogodinskaya Street, Moscow 119121, Russia, Computer Chemistry Center and Institute for Organic Chemistry, University of Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany, and Laboratory of Medicinal Chemistry, Center for Cancer Research, National Cancer Institute, National Institutes of Health, NCI-Frederick, 376 Boyles Street, Frederick, Maryland 21702

Received August 2, 2002

The application of the program PASS (Prediction of Activity Spectra for Substances) to about 250 000 compounds of the NCI Open Database and the incorporation of over 64 million PASS predictions in the Enhanced NCI Database Browser are described. A total of 565 different types of activity are included, encompassing general pharmacological effects, specific mechanisms of action, known toxicities, and others. Application of this Web-based service to prediction of activities of the kinds "Angiogenesis inhibitor," "Antiviral (HIV)", and a set of activities that can be associated with antineoplastic action are reported. For this latter data set, a very substantial enrichment over random selection was found in the PASS predictions. It is shown how the user can conduct complex searches by combining ranges of PASS-predicted probabilities of compounds to be active or to be inactive, respectively, with, e.g., value ranges of physicochemical parameters, presence or absence of particular substructural fragment, and other search criteria.

## INTRODUCTION

More than half a million samples of compounds from both organic synthesis and natural source extracts have been collected and tested by the National Cancer Institute (NCI) since 1955.[1,2] For the synthetic compounds, which constitute the large majority of the samples, about half of the computer database (of both the structures and cell-based assay results) that grew concomitantly with these screening efforts is free of any disclosure and usage restrictions. These data, produced by NCI's Developmental Therapeutics Program (DTP), are thus entirely in the public domain and are commonly called the "Open NCI Database," or often, for short (and somewhat inaccurately) "the NCI Database" (NCI DB).[3] In the early decades, anticancer screening was predominantly carried out in animals, mostly in mouse models, the results of which have not (yet) been released by DTP. The publicly available assay data, used in the context of this project, comprise a subset of the Open NCI DB of about 38 000 compounds that have been tested since the late 1980s in cell-based anti-cancer assays (using approximately 60 different cancer cell lines) and a different (but not totally disjunct) subset of about 43 000 compounds tested in an anti-HIV screening assay.[4] Apart from the public and free availability of these large bodies of assay results, another distinguishing characteristic of the NCI DB is its large percentage of structures that are unique to it, i.e., do not seem to be present in most if not all of the other chemical databases.[5]

We have recently presented the "Enhanced CACTVS Browser of the Open NCI Database,"[6] a Web-based, graphical user interface allowing rapid searches by numerous criteria in all of the currently released more than 250 000 Open NCI DB compounds.[7] While a substantial part of the data presented in this Web service are the ca. 38 000 (times approximately 60) experimental anticancer and ca. 43 000 anti-HIV assay results, the majority of the data are, in fact, more than 64 million predictions of over 500 different biological activities calculated by the computer program PASS (Prediction of Activity Spectrum for Substances)[8−10] for most of the quarter-million compounds presented in this service.

Since many of the known biologically active compounds reveal several if not numerous kinds of biological activities when tested against a large number of targets, one might expect that compounds from the NCI DB may potentially possess additional biological activities not captured by the DTP assays, including activities not related to the anticancer and anti-HIV fields. Therefore, it is entirely reasonable to hypothesize that new leads for development of different classes of drugs might be found among these molecules. However, it would be extraordinarily expensive to test all these compounds against each of the thousands of known screens. Moreover, the size of the samples available in NCI's repository, which is in fact 0 mg for about half of the compounds, would pose another stringent limitation for such extensive screening. In light of this, computer-based screening is in all likelihood the only realistic method to preselect

* Corresponding author phone: (301)846-5903; e-mail: mn1@helix.nih.gov.
[‡] Laboratory of Structure-Function Based Drug Design.
[#] Computer Chemistry Center. Current address: ChemCodes Inc., Commercial Park West, 2300 Englert Drive, Suite G, Durham, NC 27713.
[†] Laboratory of Medicinal Chemistry.

10.1021/ci020048r CCC: $25.00 
Published on Web 12/13/2002

compounds for a wide spectrum of desirable biological activities (or to exclude undesirable activities).

Most of the currently available molecular modeling-type methods, such as docking, are designed to study the ligand−receptor interaction for one specific biological macromolecule at a time,[11] while quantitative structure−activity relationships (QSAR) analysis is generally only applicable to the optimization of lead compounds' properties within the same chemical series.[12,13] In contrast to both these techniques, the computer program PASS[8−10] is able to predict many kinds of biological activity for compounds from different chemical series on the basis of just their 2D structural formulas in a very rapid manner. The set of pharmacological effects, mechanisms of action, and specific toxicities that might be exhibited by a particular compound in its interaction with biological entities, and which is predicted by PASS, is termed the "biological activity spectrum" of this compound.

This approach was initiated in the beginning of the 1970s[14] in the framework of the State System for Registration of New Chemical Compounds Synthesized in the USSR.[15] As far as we are aware of, PASS currently has no direct equivalent in terms of its breadth of predicted activity spectrum. Pharmaceutical companies typically have only a few major therapeutical fields of research, which limits their interest in a wide spectrum of biological activities. As a consequence, tools of this breadth have not been the main focus of software developers. In contrast hereto, the fact that all research institutes in the Soviet Union were in the State's possession gave rise to the idea, and its subsequent implementation, to register all synthesized chemical compounds and to select the most promising ones via computer prediction.

The purpose of the project described in this paper was to apply this innovative technology for prediction of biological activity spectra to 250 000 compounds from the NCI DB and to present this information to the scientific community in a searchable way through the Enhanced NCI Database Browser.[6]

## METHODS

**Brief Description of PASS.** *Biological activity spectrum* is a concept that is crucial to PASS and that provides the rationale for predicting many biological activity types for different compounds. Within this concept, biological activity is considered to be an intrinsic property of the compound, depending only on its structure. Any "component" of this biological activity spectrum of a given compound is assumed to be detectable under suitable experimental conditions. For each activity, the quantitative characteristics of the assay results depend significantly on the particular experimental conditions, which makes it difficult to compare and mix activity data from different sources with the goal of obtaining robust quantitative models. By using a qualitative representation of biological activity, in contrast, it is possible to combine data collected from many different sources within the same training set.

*Chemical structure* is described in PASS by original descriptors called Multilevel Neighborhoods of Atoms (MNA). Their definition has been recently published.[16] It has been shown that the MNA descriptors are rather universal and are capable of representing various structure−property relationships, including many types of biological activity,[17] mutagenicity and carcinogenicity,[18] boiling point,[16] drug-likeness,[19] etc.

*Mathematical Approach*. The mathematical approach used in PASS was selected through the analysis and detailed comparison of the effectivity, for prediction, of about 100 different methods.[20] It has been described in detail elsewhere.[19,21,22]

*Training Set*. Compounds for the training set of PASS have been continuously collected from many papers and electronic databases since 1972. The version of PASS used in this study (1.41) was based on a training set that includes about 35 000 known biologically active compounds with 565 types of biological activity. The most current version of PASS as of the time of this writing is 1.603 (Release March 2002), which includes 45 466 known biologically active compounds and 783 types of biological activity.

*Input and Output of PASS*. PASS uses as input data a MOL- or SD-file[23] representing the structural information about the molecules under study. On the basis of these data, MNA descriptors are generated automatically. It is important to note that the program is open: New MNA descriptors are generated if a new structural feature, never found before in any of the compounds in the training set, appears in the compound being read in.

Based on the statistics of MNA descriptors for active and inactive compounds from the training set, two probabilities are calculated for each activity: $P_a$ − the probability of the compound being active and $P_i$ − the probability of being inactive. Being probabilities, the $P_a$ and $P_i$ values vary from 0.000 to 1.000 (with three relevant decimals being calculated), and in general[24] $P_a + P_i < 1$, since these probabilities are calculated independently. $P_a$ and $P_i$ can be considered to be measures of the compound under study belonging to the classes of active and inactive compounds, respectively, or can be seen as estimates for the first and second kinds of errors in the predicion.

All MNA descriptors influence the estimates in the activity prediction. Their influence can be either positive (if the descriptors are found in compounds with the particular activity), or negative (if the descriptors are found in compounds without the particular activity), or even neutral (if the descriptors are found in both active and inactive compounds). In the last case, they decrease the relative impact of the "positive" and "negative" descriptors.

*Interpretation of Predictions*. The PASS predictions can be interpreted, and used, in a flexible manner. The most probable activities, for a given compound, are characterized by $P_a$ values close to 1, and $P_i$ values close to 0. Let us first consider cases where the $P_a$ value is high and is much larger than $P_i$. If a statistically significant set of samples with predictions obtained with the threshold $P_a > 0.9$ is selected from a much larger database and assayed, one has to expect to lose 90% of the active compounds, but the fraction of false-positives will be very small. For a cutoff of $P_a > 0.8$, only 80% of the actives will be lost, but the fraction of false-positives will be a little bit higher. Finally, if one goes down to the criteria $P_a > P_i$, the probability of the first kind of error equals the probability of the second kind of error, i.e., one is just as likely to miss true actives as to find false-positives.

However, maximizing $P_a$ values for the desired activity is not the only criteria for selection of the most promising compounds. Another aspect might be the novelty of a compound. If $P_a$ is very high, the compound might be a close analogue of known pharmaceutical agents. Thus, if one is interested in finding new leads, especially New Chemical Entities (NCE), one may want to choose compounds for which the specified activity is predicted with lower probability, say, $0.5 < P_a < 0.7$. In this case, the probability of false positives is likely to be higher, but if the activity will be confirmed in the experiment, one has a higher chance of having obtained an NCE.

*Hardware, Software, and Performance*. PASS runs on PC-compatible computers under Windows 95/98/2000/NT/Me/XP or on Macintosh or SGI systems under a Windows emulator. The calculation of biological activity spectra for 10 000 compounds on a fast modern PC in the 1 GHz class takes only a few minutes. Therefore, PASS can be effectively used to analyze large databases, such as the NCI Open Database.

*Validation*. The validation of PASS was carried out by leave-one-out cross-validation (LOO CV) throughout all compounds and biological activities in the training set. It was shown that the average accuracy of prediction is about 85%, which demonstrates the applicability of PASS to the drug discovery process (the expected value of randomly guessing any one of the 783 types of biological activity less than 0.5%). To avoid the pitfall of LOO CV sometimes returning too optimistic an assessment of the prediction accuracy, especially if close congeners are present in the training set, more rigorous tests of PASS have been performed previously (e.g., half the training set, or up to 60% of the activity data, were removed from the training set),[25] which showed that the approach provides statistically robust structure−activity relationships and predictions. This robustness is extremely important because the data included in the training set are always incomplete in terms of both the active structures' diversity and biological information (i.e. none of the compounds was tested vs each known type of activity).

*PASS Predictions via Internet*. Recently, an Internet version of PASS has been made available at the PASS developers' Web site.[21] The users can submit a MOL-file of the molecule under study and obtain the predicted biological activity spectrum displayed on their computer immediately. This new Internet version of PASS provides access to the prediction of all 783 kinds of biological activity, in contrast to an earlier version that predicted 319 activities.[26] At this Web site,[21] one can also find a detailed description of the algorithm and chemical descriptors used in PASS, the list of predicted activities, some examples of PASS applications to the discovery of new lead compounds with antiulcer, hepatoprotective, antiamnestic, antihypertensive, and other actions as well as references of current publications.

*Implementation of PASS Predictions in the Enhanced NCI Database Browser*. Version 1.41 of PASS was used for the calculation of the $P_a$ and $P_i$ values that are implemented in the current version of the Enhanced NCI Database Browser for the quarter-million NCI DB compounds. The NCI DB was split into portions of about 50 000 compounds for more convenient processing on a Windows PC. On a 400 MHz Pentium II PC, processing of each of these files took less than an hour. (Re-running these jobs on a more modern computer, a 1 GHz AMD Athlon PC, allowed the entire file to be processed in under 2 h.) Out of 250 251 compounds, PASS was able to process 245 979. Out of those, 35 641 were reported by PASS to have produced two or more heretofore unknown MNA descriptors, which indicates that the predictions are to be used with care. The PASS runs showed that 5196 structures were present both in the NCI DB and in the PASS (1.41) training set. Such structures are by default excluded from the PASS training set during the prediction. The cumulative total number of different activities calculated for the NCI DB compounds was 565, although the average number of activities for each individual compound was much lower (130; minimum: 10; maximum: 346), since a threshold criteria of $P_a > P_i$ had been applied in the PASS runs for admittance of an activity to the output list.

The result files, which are SD files identical to the input files with the exception of an added—albeit large, multiline—field with the PASS predictions, were recombined to a single file. This file was then processed, using appropriate CACTVS scripts on a Linux server, to provide the input in the correct format for incorporation in the monolithic CACTVS binary database file that underlies the Enhanced NCI Database Browser Web service.[6]

The user has access, through the service's Help page, to the list of all activities occurring for the NCI DB compounds and, for each of these activities, the number of compounds that were predicted to exhibit it. Likewise, one can assess the quality of each predicted activity's SAR model by studying the SAR Base LOO cross-validation results made available in a file accessible from the Help page, too.

## RESULTS AND DISCUSSION

The total number of PASS predictions incorporated in the Enhanced NCI Database Browser is 64 188 212 as of now. All of these data are available for analysis in a searchable mode as well as for download with each individual activity selectable separately. 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 the 565 possible predicted activities. A specific activity has to be selected, and the type of prediction (probability of activity [$P_a$] or inactivity [$P_i$], respectively) has to be specified. PASS search criteria values have to be specified in probability ranges (in subintervals of 0.0−1.0). A PASS search can be combined with any other of the numerous search criteria available in the service to form complex search strategies. A detailed description of how to perform such searches has been published previously.[6]

*Application 1. Angiogenesis Inhibitors.* As an example, we want to search for probable angiogenesis inhibitors. For this, we specify $P_a \geq 0.9$ and $P_i < 0.2$ and molecular weight within 400−500 D. This query is submitted in the Query Form of the Enhanced NCI Database Browser shown in Figure 1. As can be seen from Figure 2, 10 compounds are found as satisfying these criteria. Detailed information about the structure and predicted biological activities of one of them, NSC 75532, is presented in Figure 3. For this compound, angiogenesis inhibiting activity is predicted with $P_a = 0.945$ and $P_i = 0.004$.

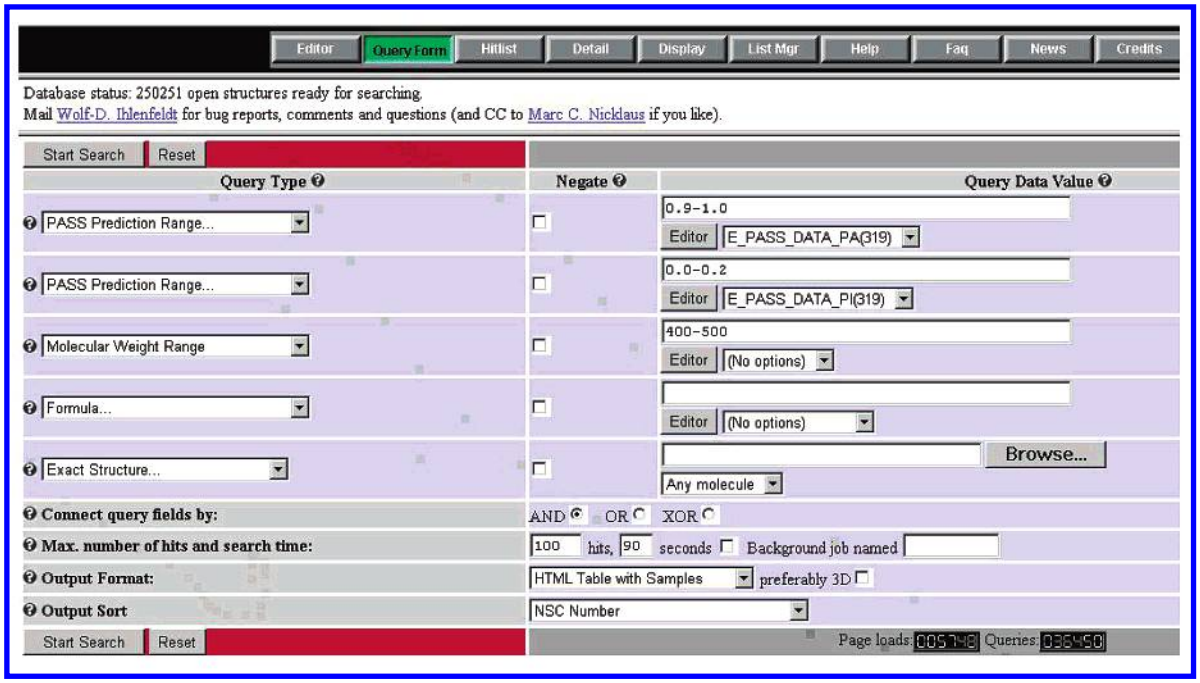



**Figure 1.** Search query. Search criteria: PASS Activity "Angiogenesis inhibitor" (internal code number 319), query data values: $P_a > 0.9$ and $P_i < 0.2$ and $400 < \mathrm{MW} < 500$.

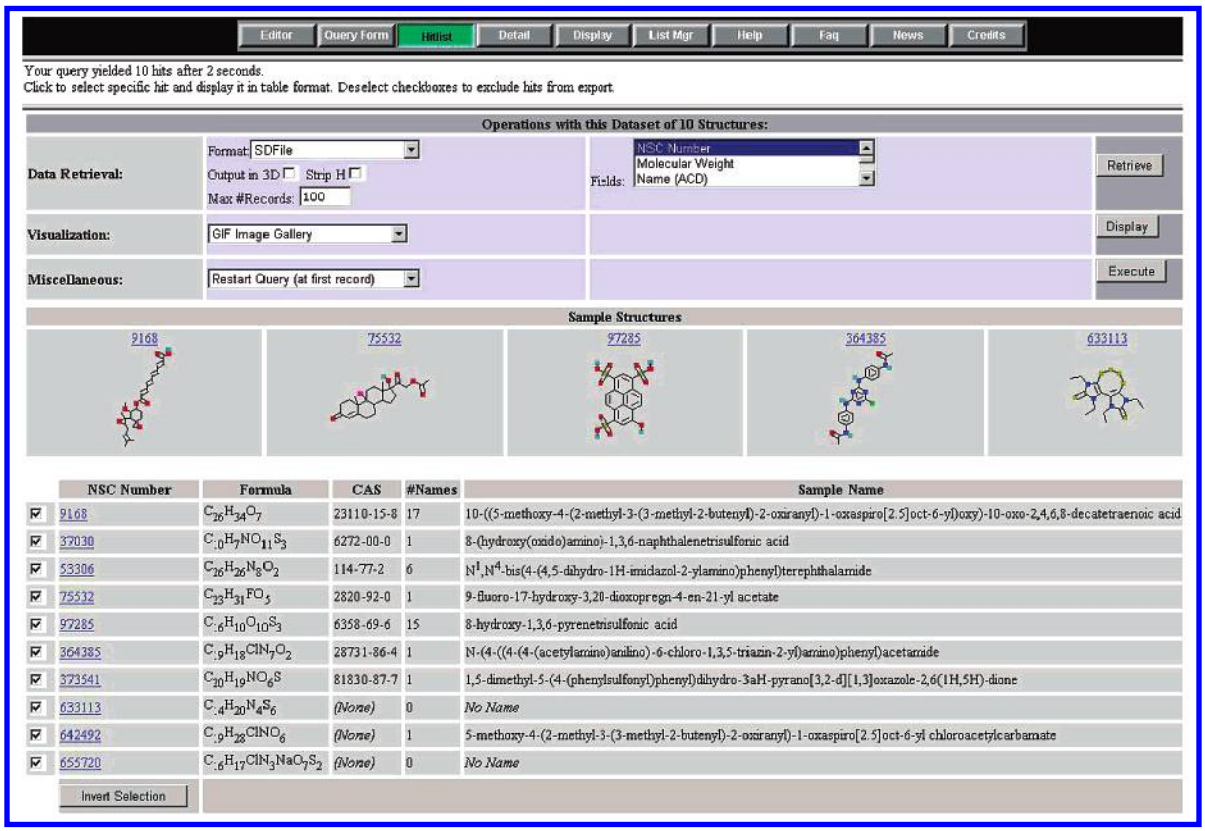
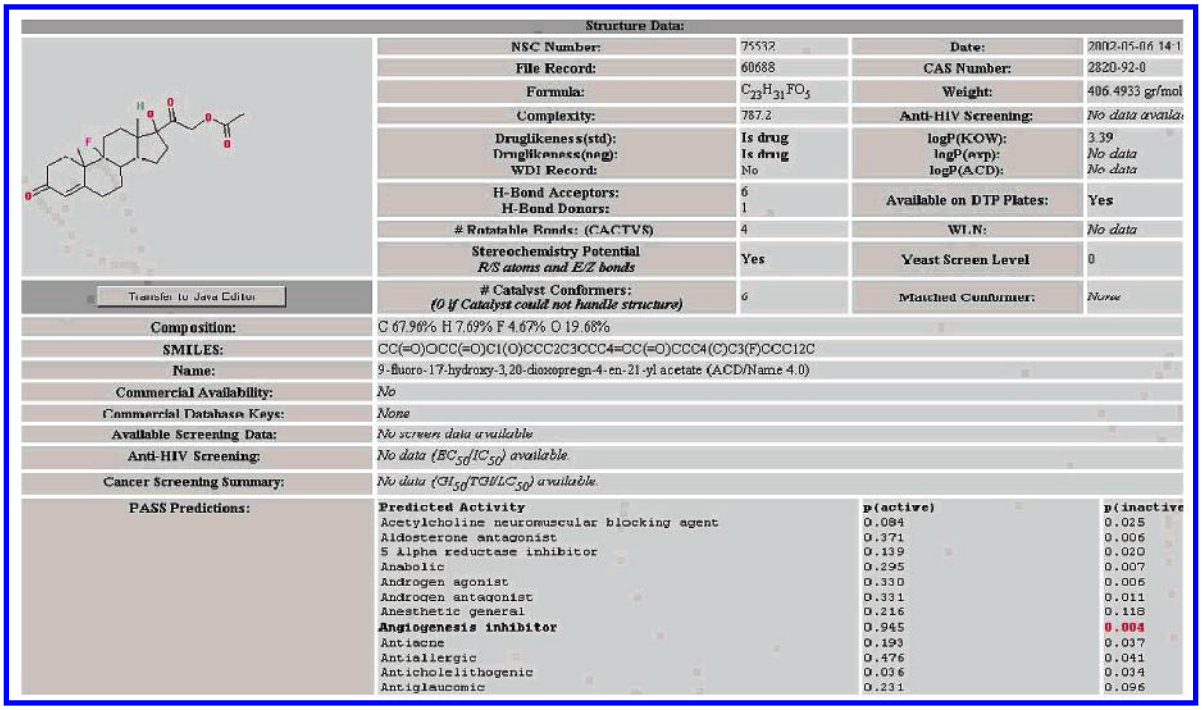



**Figure 2.** Hitlist of 10 structures that satisfied the search query shown in Figure 1.

As another example, we specify the same search criteria regarding predicted activity and inactivity as above, omitting the molecular weight restriction but adding one structural requirement—the absence of an amide group $-C(=O)NR_2$ (since many known angiogenesis inhibitors possess an amide functionality, and we want to find more novel structures). The number of structures that fulfill these criteria is 123. A partial list of the compounds found is shown in Figure 4.

If one wants to determine the total number of angiogenesis inhibitors that are predicted with varying probability for the 250 000 compounds of the Open NCI Database, one can easily do this varying the $P_a$ threshold values. One finds that the number of such compounds is 237 for $P_a > 0.9$; 983 for $P_a > 0.8$; 2988 for $P_a > 0.7$; 9280 for $P_a > 0.6$; and 24816 for $P_a > 0.5$; etc. Exploiting the fact that simple variation of the $P_a$ and/or $P_i$ thresholds allows one to greatly vary the



**Structure Data:**

| | | | | |
|---|---|---|---|---|
| | **NSC Number:** | 75532 | **Date:** | 2002-05-06 14:1 |
| | **File Record:** | 60688 | **CAS Number:** | 2820-92-0 |
| | **Formula:** | $C_{23}H_{31}FO_5$ | **Weight:** | 406.4933 gr/mol |
| | **Complexity:** | 787.2 | **Anti-HIV Screening:** | *No data availa* |
| | **Druglikeness(std):**<br>**Druglikeness(neg):**<br>**WDI Record:** | **Is drug**<br>**Is drug**<br>No | **logP(KOW):**<br>**logP(avp):**<br>**logP(ACD):** | 3.39<br>*No data*<br>*No data* |
| | **H-Bond Acceptors:**<br>**H-Bond Donors:** | 6<br>1 | **Available on DTP Plates:** | **Yes** |
| | **# Rotatable Bonds: (CACTVS)** | 4 | **WLN:** | *No data* |
| | **Stereochemistry Potential**<br>***R/S atoms and E/Z bonds*** | **Yes** | **Yeast Screen Level** | 0 |
| Transfer to Java Editor | **# Catalyst Conformers:**<br>***(0 if Catalyst could not handle structure)*** | 6 | **Matched Conformer:** | *None* |

| | |
|---|---|
| **Composition:** | C 67.96% H 7.69% F 4.67% O 19.68% |
| **SMILES:** | CC(=O)OCC(=O)C1(O)CCC2C3CCC4=CC(=O)CCC4(C)C3(F)CCC12C |
| **Name:** | 9-fluoro-17-hydroxy-3,20-dioxopregn-4-en-21-yl acetate (ACD/Name 4.0) |
| **Commercial Availability:** | *No* |
| **Commercial Database Keys:** | *None* |
| **Available Screening Data:** | *No screen data available* |
| **Anti-HIV Screening:** | *No data ($EC_{50}$/$IC_{50}$) available.* |
| **Cancer Screening Summary:** | *No data ($GI_{50}$/$TGI$/$LC_{50}$) available.* |

**PASS Predictions:**

| Predicted Activity | p(active) | p(inactive |
|---|---|---|
| Acetylcholine neuromuscular blocking agent | 0.084 | 0.025 |
| Aldosterone antagonist | 0.371 | 0.006 |
| 5 Alpha reductase inhibitor | 0.139 | 0.020 |
| Anabolic | 0.295 | 0.007 |
| Androgen agonist | 0.330 | 0.006 |
| Androgen antagonist | 0.331 | 0.011 |
| Anesthetic general | 0.216 | 0.118 |
| **Angiogenesis inhibitor** | 0.945 | **0.004** |
| Antiacne | 0.193 | 0.037 |
| Antiallergic | 0.476 | 0.041 |
| Anticholelithogenic | 0.036 | 0.034 |
| Antiglaucomic | 0.231 | 0.096 |

**Figure 3.** Detailed information for structure NSC 75532, one of the structures returned in the hit list shown in Figure 2. Note the specific PASS prediction used in the query (Figure 1), highlighted in red.

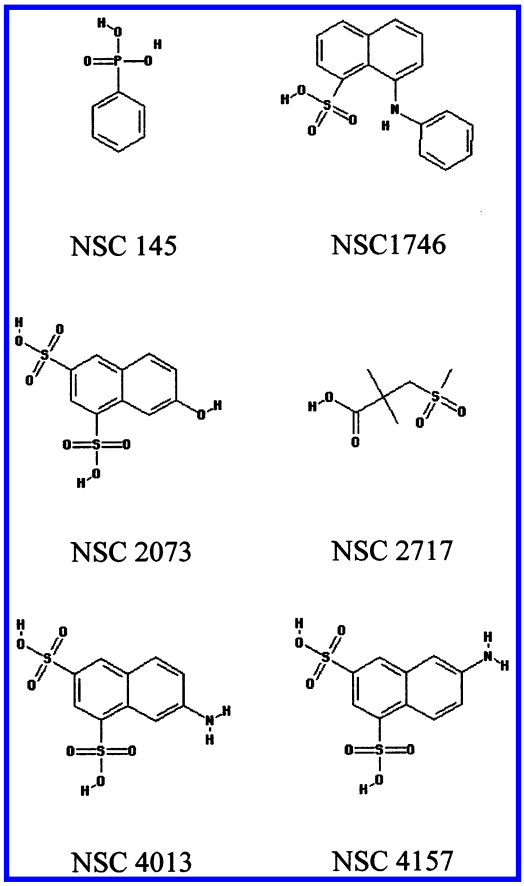



**Figure 4.** Results of the search for angiogenesis inhibitors that have no amide group in the structure.

numbers of PASS hits, one sees that one can fruitfully apply additional criteria, such as the ones discussed above, in a flexible manner to select compounds for testing according to the needs of the particular project.

In an effort closely related to this, similar searches that included PASS anti-angiogenesis queries were conducted, and a small number of the obtained hits subsequently assayed. Out of seven tested compounds, four showed inhibitory activity at the approximately 10−100 $\mu$M level. Although this set is too small to draw statistically valid conclusions from it, it may be exemplary of the usage, and usefulness, of this service: To provide, at very low cost (in fact, zero cost) and minimal learning curve, new ideas and possibly novel lead compounds in drug design projects.

*Application 2. Antiviral (HIV) Activity.* To analyze the power of the PASS predictions to enrich true positives in a database subset, we compared the results of anti-HIV activity prediction for the compounds from the Open NCI Database with the results of DTP's anti-HIV screening. We retrieved the compounds predicted by PASS as "Antiviral (HIV)" with varying $P_a$ thresholds ($P_a > 0.1$; $P_a > 0.2$; ... $P_a > 0.9$) and saved these results in separate hitlist files. A validation of these results can obviously only be performed with those compounds that have actually been antivirally tested. Within the 250k NCI Open Database, this is a subset of about 43 000 compounds. We therefore searched the database for "AIDS Screen Results" with two different criteria pertaining to the DTP AIDS screening, "Active" or "Moderately Active[27]" on one hand, and "Inactive" on the other hand, and saved the results in two separate hitlist files. The number of compounds in the first file (actives) was 1504, and the number of compounds in the second file (inactives) was 41 185, the total of all screened compounds thus being 42 689.

The appropriate intersection sets were then generated using the "INTERSECTION" option of the LIST MANAGER functionality in the Enhanced Open NCI Database Browser.[6] The percentage of actives in the tested subset of open NCI compounds is 3.52% (1504/42 689). A random selection would therefore preserve this ratio. Table 1 lists the sizes of the compound sets with PASS-predicted "Antiviral (HIV)"

**Table 1.** Comparison of PASS-Predicted "Antiviral (HIV)" Activity with DTP Assay Results

| $P_a$ threshold[a] | predicted in entire DB[b] | pred. in act.[c] | pred. in inact.[d] | pred. in screened[e] | predictivity[f] (%) | enrichment[g] |
|---|---|---|---|---|---|---|
| 0.1 | 43885 | 642 | 7288 | 7930 | 8.09584 | 2.29995 |
| 0.2 | 29758 | 559 | 5051 | 5610 | 9.96435 | 2.83078 |
| 0.3 | 12486 | 442 | 2354 | 2796 | 15.8083 | 4.49099 |
| 0.4 | 6462 | 330 | 1311 | 1641 | 20.10969 | 5.71298 |
| 0.5 | 3214 | 250 | 711 | 961 | 26.01457 | 7.3905 |
| 0.6 | 1659 | 207 | 392 | 599 | 34.5576 | 9.8175 |
| 0.7 | 970 | 174 | 231 | 405 | 42.96296 | 12.20539 |
| 0.8 | 530 | 129 | 124 | 253 | 50.98814 | 14.48527 |
| 0.9 | 215 | 76 | 53 | 129 | 58.91473 | 16.73714 |

[a] Allowed range [$P_{a\ cutoff}$,1.0] of predicted probability. [b] Compounds predicted as active in the entire 250k NCI Open DB. [c] Compounds predicted as active in the subset of NCI compounds screened and found to be antivirally active. [d] Compounds predicted as active in the subset of NCI compounds screened and found to be antivirally inactive. [e] Compounds predicted as active in the subset of NCI compounds screened (sum of columns 3 and 4). [f] Ratio (in percent)of column 3 to column 5. [g] Ratio of column 6 to percentage of actives in DTP-screened compound set (3.52%).

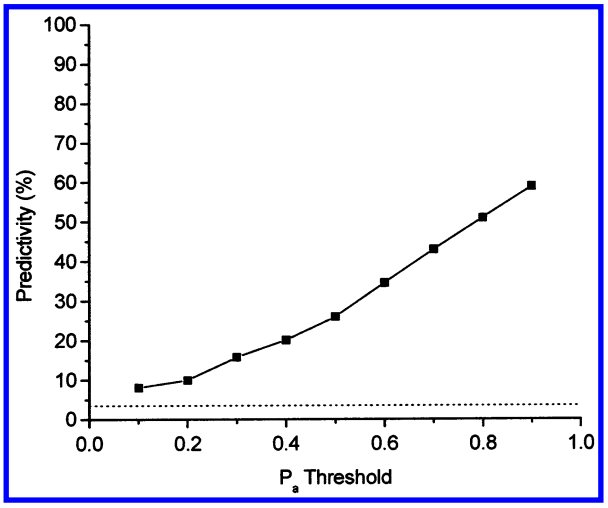



**Figure 5.** Percentage of the correctly predicted compounds with anti-HIV activity in the DTP-screened subset of the Open NCI Database intersected with the PASS-predicted subset, as a function of the threshold of predicted probability of activity. The horizontal dotted line is the percentage (3.52%) of active compounds (DTP label "Moderately Active" or "Confirmed Active") in the entire screened subset.

activity for the entire Open NCI DB as well as these sets' intersections with the active and inactive screened compounds, respectively, as a function of the lower boundary of the probability interval [$P_{a\ cutoff}$,1.0] chosen. A plot (Figure 5) of the predictivity ratio of the latter two set sizes, i.e., of

$$100 * \text{size\_of}\{\{\text{predicted}\}\cap\{\text{actives}\}\}/\text{size\_of}\{\{\text{predicted}\}\cap\{\text{screened}\}\}$$

shows that a very substantial enrichment is achieved even at the very lowest $P_a$ threshold values. At the highest $P_a$ threshold values, the predictivity ratio approaches 60%, i.e., the enrichment gets close to a factor of 17 (the theoretical maximum being 100/3.52 = 28.4). It is important to reiterate that, if a compound predicted by PASS happens to be in the program's training set, it is automatically excluded from the SAR model. Thus, none of the compounds predicted as actives in this analysis was simply "picked" from the training set.

*Application 3 and Analysis. Antineoplastic Activities.* PASS can be used for analyzing the occurrence, in a database, of compounds predicted to be active for a well-defined set of PASS activities. If the database, by design or by history, has a well-known "specialty" in its activity spectrum, then this can conversely be used to validate, to some degree, the PASS program. In this study, the obvious database to use was the NCI Database, and the activities to analyze were activities that can be linked with antineoplastic action. This approach, however, is of a more general nature, and can be applied to any large compound database to analyze its "propensity" for specific subsets of the PASS activity spectrum, thereby possibly revealing additional, hitherto unknown uses and characteristics.

The essence of this analysis is to compute the number of compounds, as a fraction of the entire database, that are predicted to be active for the chosen set of activities, as a function of the lower cutoff point for the admitted probability range. This can be done for either one of the predicted probabilities $P_a$ and $P_i$. Due to the nature of what is predicted, for the probability of *activity*, $P_a$, this histogram must be a decreasing curve (first steeply, then more gradually leveling of to zero as the probability cutoff is raised closer and closer to 1.0), whereas the equivalent curve for $P_i$ is monotonically increasing from 0 to 1.0 (all compounds have *some* inactivity probability $< 1.0$). Thus, the curve for $P_i$ is much closer to a linear relationship than that of $P_a$. Therefore its use and interpretation in the subsequent analyses is more straightforward than for $P_a$, hence all following curves are based on probability ranges $P_i = [P_{i\ cutoff},1.0]$. While the absolute distributions may not be that much interpretable, one can scale them by the equivalent histogram of the PASS training set, thereby deriving a measure of how much more the database under study is tilting toward (or away from) the chosen activity set than the training set. Calculating such scaled distributions for different databases then allows one to compare them as to how strongly the selected activity set is represented in one vs the other. If, for example, this is done for the subset of PASS activities that can be associated with antineoplastic action, then two databases can be compared as to how "cancer drug-like" they are relative to each other (see below).

Assuming that the distributions of $P_a$ and $P_i$ in our data set are *approximately* the same as in the PASS training set (i.e. we do not have any unusual, "bizarre" distribution), then the probability $h$ of recognition of a compound as active is

$$h \cong p\,(1 - P_a) + (1 - p)\,P_i \qquad (1)$$

where $p$ is the probability of having an active compound in

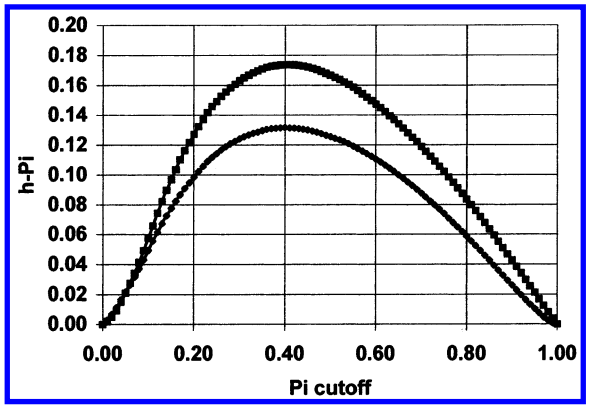



**Figure 6.** Preponderance $h-P_i$ of antineoplastic activities, as a function of the lower boundary $P_{i\ cutoff}$ of the admitted $P_i$ range [$P_{i\ cutoff}$ , 1.0] for compounds included, for 56 antineoplastic-associated PASS v. 1.41 activities (Table 2). Squares: NCI Open DB (upper curve); diamonds: ACD (lower curve).

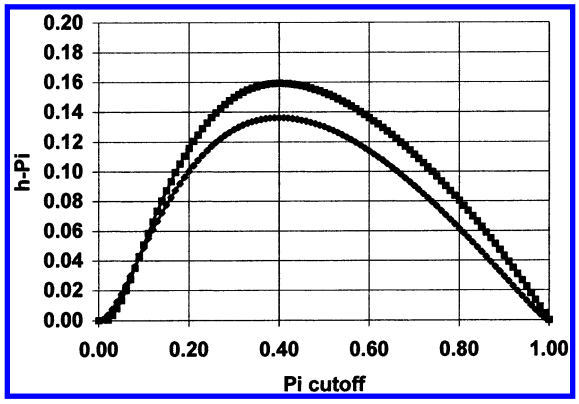



**Figure 7.** Preponderance $h-P_i$ of antineoplastic activities, as a function of the lower boundary $P_{i\ cutoff}$ of the admitted $P_i$ range [$P_{i\ cutoff}$ , 1.0] for compounds included, for 509 nonantineoplastic-associated PASS v. 1.41 activities. Squares: NCI Open DB (upper curve); diamonds: ACD (lower curve).

our data set. Simple transformation of (1) yields

$$h - P_i \cong p\,(1 - P_a - P_i) \qquad (2)$$

and this entity, which can be seen as a measure of the preponderance of a database for a given subset of activities, is plotted, for varying cutoff points $P_{i\ cutoff}$ of the $P_i$ range admitted, in Figures 6 and 7. These curves were calculated for both the subset of 56 activities that can be associated with antineoplastic action (Figure 6), which are listed in Table 2, and for the remaining 509 of the total 565 PASS v. 1.41 activities (Figure 7) that are included in the Web service. (The values that make up these curves could, in principle, be compiled with the help of the Web service; however, it was much faster to obtain them with a modified version of the PASS program itself.)

To compare the NCI DB with a similar-size but more generic small-molecule database, such curves were derived for 246 009 compounds of the NCI DB as well as for 237 787 compounds from the Available Chemicals Directory (ACD-3D 99.2).[23] The top curve in Figure 6, e.g., can be interpreted in such a way that, for the NCI DB, one has a nearly 18% higher chance of finding, for a $P_i$ interval of [0.4,1.0], a compound with one of the 56 antineoplastic-related compounds than in the training set. The equivalent value for the ACD is quite a bit lower, with the maximum $h-P_i$ value reaching only about 13%. For the nonantineoplastic-related 509 activities, the curves for NCI DB and ACD (Figure 7) are substantially closer together. Because of the very large numbers of compounds involved, these differences are highly statistically significant. This is a reassuring finding: The NCI DB is indeed more "cancer drug-like" than the ACD. For other activities, these two databases are more similar to each other, but the NCI DB still holds some advantage here. This is in fact what one would expect, since the NCI DB has undoubtedly been biased during sample collection toward drug-like compounds, which would be more likely to be associated with biological activities than the ACD's more general organic chemicals. Whether its even stronger pre-

**Table 2.** Activities of PASS 1.41 That Were Defined as Being Associated with Antineoplastic Action

| |
|---|
| 5 alpha reductase inhibitor |
| adenosine A3 receptor antagonist |
| alkylator |
| alkylphospholipid |
| aminopeptidase microsomal inhibitor |
| androgen antagonist |
| angiogenesis inhibitor |
| antibiotic anthracycline-like |
| antimetabolite |
| antimitotic |
| antimitotic podophyllotoxin-like |
| antineoplastic |
| antineoplastic alkaloid |
| antineoplastic antibiotic |
| antineoplastic enhancer |
| aromatase inhibitor |
| bombesin antagonist |
| cathepsin B inhibitor |
| collagenase inhibitor |
| DNA intercalator |
| dihydrofolate reductase inhibitor |
| dihydroorotate dehydrogenase inhibitor |
| estrogen agonist |
| estrogen antagonist |
| estrone sulfatase inhibitor |
| farnesyltransferase inhibitor |
| geranylgeranyltransferase inhibitor |
| glycinamide ribonucleotide formyltransferase inhibitor |
| histamine agonist |
| immunostimulant |
| interleukin agonist |
| luteinizing hormone-releasing hormone agonist |
| luteinizing hormone-releasing hormone antagonist |
| mannosidase inhibitor |
| microtubule formation inhibitor |
| neuropeptide antagonist |
| nitric oxide synthase inhibitor |
| nucleotide metabolism regulator |
| O6-alkylguanine-DNA alkyltransferase inhibitor |
| ornithine decarboxylase inhibitor |
| phospholipase C inhibitor |
| photosensitizer |
| progesterone antagonist |
| protein kinase C inhibitor |
| ribonucleoside diphosphate reductase inhibitor |
| ribonucleotide reductase inhibitor |
| S-adenosyl-L-homocysteine hydrolase inhibitor |
| S-adenosyl-L-methionine decarboxylase inhibitor |
| signal transduction pathways inhibitor |
| somatostatin agonist |
| thymidylate synthase inhibitor |
| topoisomerase II inhibitor |
| topoisomerase inhibitor |
| tyrosine kinase inhibitor |
| uridine phosphorylase inhibitor |
| vitamin D-like |

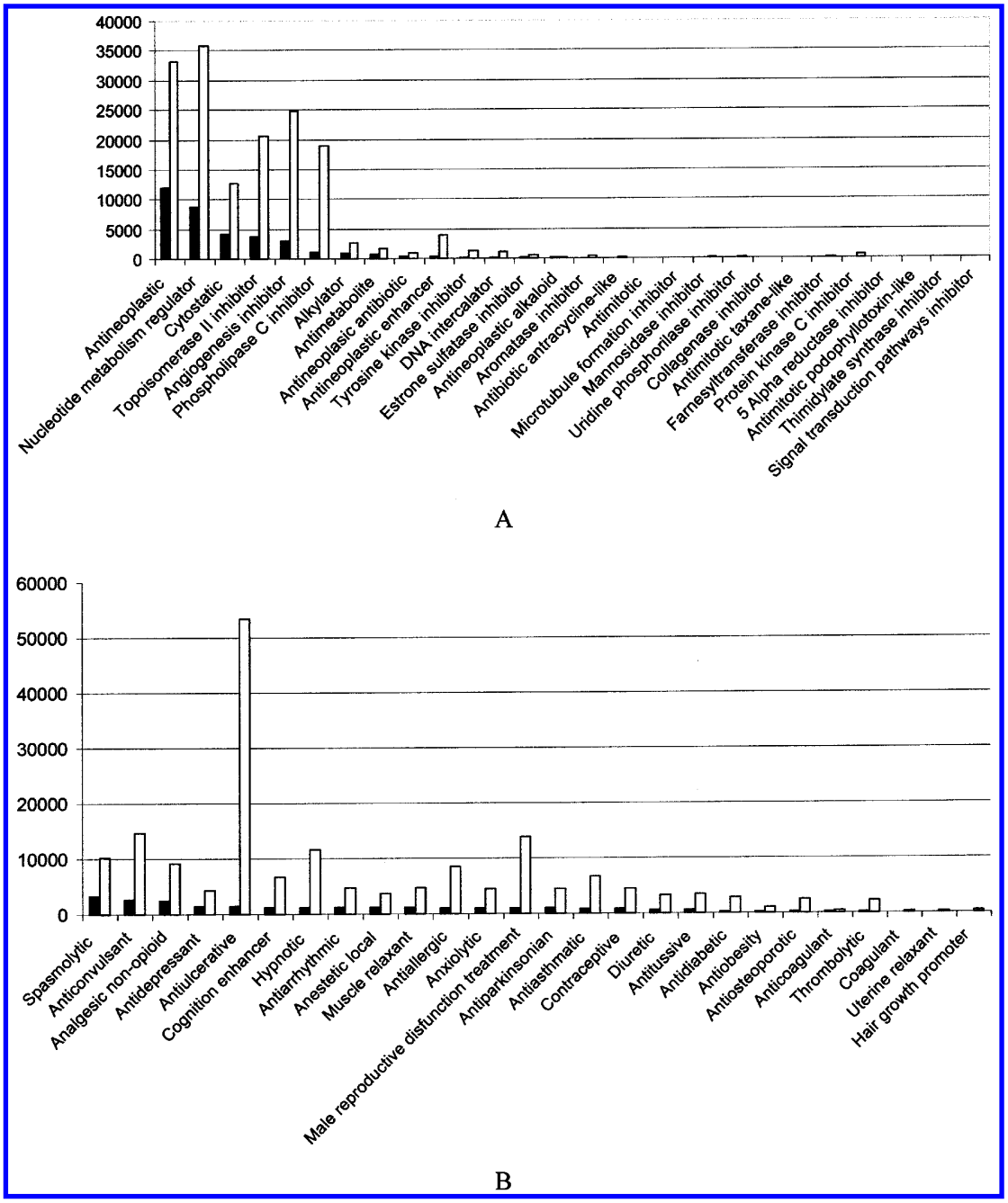



**Figure 8.** Number of compounds from the Open NCI database with predicted activities that are associated with anticancer action (A) and a random selection of those that are not associated with anticancer action (B). Dark bars: Threshold $P_a > 0.7$; light bars: $P_a > 0.5$.

ponderance for "cancer drug-like" molecules is cause or effect is debatable, given that NCI has been involved in the development, at one stage or another, of many of today's approved cancer chemotherapeuticals (many of which are by necessity part of the PASS training set), with the consequence that many of these compounds and related analogues were entered in the NCI DB. Whatever the relationship is, PASS correctly pinpoints this nature of the NCI DB.

To convey an impression of what specific activities are involved in the NCI DB's propensities for the above activity categories, and what some of the specific numbers of predicted compounds are, Figure 8 is included. It shows the number of hits predicted at $P_a > 0.7$ and $P_a > 0.5$ levels, respectively, predicted to be active against 28 kinds of biological activity associated with antineoplastic action, and the corresponding values for 26 kinds of biological activity (selected by random) that are not related to the anticancer field. For instance, 11 909 compounds are predicted as antineoplastic with $P_a > 0.7$ and 33 203 ones with $P_a > 0.5$; 8797 compounds are predicted as nucleotide metabolism regulators with $P_a > 0.7$ and 35 798 ones with $P_a > 0.5$; 4169 compounds are predicted as cytostatic with $P_a > 0.7$ and 12 716 ones with $P_a > 0.5$; etc. In contrast hereto, the most probable among the checked nonantineoplastic-related activities with $P_a > 0.7$ is spasmolytic, and this activity is predicted for 3193 compounds; the next is anticonvulsant that is predicted with $P_a > 0.7$ for 2696 compounds, etc.

Interestingly, when one lowers the $P_a$ threshold from 0.7 to 0.5, the number of predicted compounds with nonanticancer activities becomes much higher and can, in fact, surpass that for cancer-related activities (Figure 8).

For instance, antiulcerative activity is predicted for 53 457 compounds; anticonvulsant for 14 676 compounds, etc. This demonstrates that the PASS predictions for the Open NCI Database can be used as a flexible filter for selecting compounds for assaying against a wide variety of targets. Keeping in mind that the user has the possiblity to combine multiple search criteria in the Enhanced NCI Database Browser, such as selecting compounds with a desirable combination of "pharmacological effect + mechanism of action", plus, e.g., substructural search criteria, the range of applications that are possible seem primarily limited by the user's imagination.

## ACKNOWLEDGMENT

We gratefully acknowledge the support of this work by the U.S. Civil Research and Development Foundation (Grant # RC1-2064).

## REFERENCES AND NOTES

(1) Milne, G. W. A.; Miller, J. A. The NCI Drug Information System. 1. System Overview. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 154−159.
(2) Milne, G. W. A.; Nicklaus, M. C.; Driscoll, J. S.; Wang, S.; Zaharevitz, D. National Cancer Institute Drug Information System 3D Database. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1219−1224.
(3) http://dtp.nci.nih.gov/webdata.html.
(4) The DTP Human Tumor Cell Line screening data are available at http://dtp.nci.nih.gov/docs/cancer/cancer_data.html. The DTP AIDS Antiviral screening data are available at http://dtp.nci.nih.gov/docs/aids/aids_data.html. A very recent new release by DTP ("March 2002") has expanded this data set to about 41 000 compounds for the cancer screen and to about 44 000 compounds for the AIDS screen, respectively.
(5) Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI Open Database with Seven Large Chemical Structural Databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.
(6) Ihlenfeldt, W.-D.; Voigt, J. H.; Bienfait, B.; Oellien, F.; Nicklaus, M. C.; Enhanced CACTVS Browser of the Open NCI Database**,** *J. Chem. Inf. Comput. Sci*. **2002**, *42*(1), 46−57.
(7) The original source for structure information of the open NCI DB componds is http://dtp.nci.nih.gov/docs/3d_database/structural_information/structural_data.html. They can also be obtained, in several "repackaged" forms, at http://cactus.nci.nih.gov/ncidb2/download.html.
(8) Filimonov, D. A.; Poroikov, V. V. PASS: Computerized Prediction of Biological Activity Spectra for Chemical Substances. In *Bioactive Compound Design: Possibilities for Industrial Use;* BIOS Scientific Publishers: Oxford, 1996; pp 47−56.
(9) Gloriozova, T. A.; Filimonov, D. A.; Lagunin, A. A.; Poroikov, V. V. Evaluation of computer system for prediction of biological activity PASS on the set of new chemical compounds. *Chim.-Pharm. J. (Rus)*; **1998**, *32*(12), 32−39. (English translation by Consultants Bureau, New York: Pharmaceutical Chemistry Journal).
(10) Poroikov, V.; Filimonov, D. Computer-aided prediction of biological activity spectra. Application for finding and optimization of new leads. *Rational Approaches to Drug Design*; Prous Science: Barcelona, 2001; pp 403−407.
(11) *3D QSAR in Drug Design;* Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer: The Netherlands, 1997; Vol. II and Vol. III.
(12) Van de Waterbeemd, H. *Structure−Property Correlations in Drug Research;* Academic Press: 1996.
(13) *The Practice of Medicinal Chemistry;* Wermuth, C., Ed.; Academic Press: 1998.
(14) Avidon, V. V. The criteria of chemical structures similarity and the principles for design of description language for chemical information processing of biologically active compounds. *Chem. Pharmaceut. J. (Rus.)* **1974**, *No. 7*, 22−25.
(15) Burov, Yu. V.; Poroikov, V. V.; Korolchenko, L. V. National System for Registration and Biological Testing of Chemical Compounds: Facilities for New Drugs Search. *Bull. Natl. Center Biologically Active Compounds (Rus.)*. **1990**, *No. 1*, 4−25.
(16) Filimonov, D.; Poroikov, V.; Borodina, Yu.; Gloriozova, T. Chemical Similarity Assessment Through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 666−670.
(17) Poroikov V. V.; Filimonov D. A. Computer-assisted prediction of biological activity in a search for and optimization of new drugs. In *Nitrogen-containing heterocycles and alcaloides*; Iridium Press: Moscow, 2001; Vol. 1, pp 149−154.
(18) Suchkov, A. P.; Filimonov, D. A.; Stepanchikova, A. V.; Poroikov, V. V. *Abstr. 11th European Symposium on Quantitative Structure−Activity Relationships: Computer-Assisted Lead Finding and Optimisation*; Lausanne, Switzerland, 1996; P-32C.
(19) Anzali, S.; Barnickel, G.; Cezanne, B.; Krug, M.; Filimonov, D.; Poroikov V. Discriminating between Drugs and Nondrugs by Prediction of Activity Spectra for Substances (PASS). *J. Med. Chem.* **2001**, *44*, 2432−2437.
(20) Filimonov, D. A. *Abstr. II Rus. Natl. Congr. "Man and Drugs"* Moscow, 1995; pp 62−63.
(21) http://www.ibmh.msk.su/PASS.
(22) Poroikov V.; Akimov D.; Shabelnikova, E.; Filimonov, D. Top 200 medicines: can new actions be discovered through computer-aided prediction? *SAR QSAR Environ. Res.* **2001**, *12*(4), 327−344.
(23) MDL Information Systems, Inc., San Leandro, CA; http://www.mdli.com.
(24) If we have an "ideal system" that would give rise to a model with no errors, i.e., with the mean error of prediction (MEP) being zero, then we would strictly have $P_a > 0$ and $P_i = 0$, or $P_i > 0$ and $P_a = 0$. In fact, it is possible that there could be a region where $P_i = 0$ and $P_a = 0$. At the other extreme, we would have a system where MEP = 0.5. In this case, $P_a + P_i = 1$, but there would be no real information in the predictions.
(25) Poroikov, V. V.; Filimonov, D. A.; Borodina, Yu. V.; Lagunin, A. A.; Kos A. J. Robustness of biological activity predicting by computer program PASS for noncongeneric sets of chemical compounds. *J. Chem. Info. Comput. Sci.* **2000**, *40*, 1349−1355.
(26) Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V. PASS: Prediction of Activity Spectra for Biologically Active Substances. *Bioinformatics* **2000**, *16*, 747−748.
(27) http://dtp.nci.nih.gov/docs/aids/aids_data.html.

CI020048R