PASS

User Guide

V.V.Poroikov, D.A.Filimonov & Associates

© Copyright 2005 by V.V. Poroikov, D.A. Filimonov & Associates.
All rights reserved.

MDL is a registered trademark of MDL Information Systems, Inc.

ISIS is a trademark of MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, California, 94577.

All other product names are trademarks or registered trademarks of their respective holders.

No part of this document may be reproduced by any means except of permitted in written by V.V.Poroikov, D.A.Filimonov & Associates, Institute Biomedical Chemistry of Russian Academy of Medical Sciences, Pogodinskaya Sthreet, 10, Moscow, 119121, Russia.

CONTENTS

Abbreviations 3

1. Hardware and Software System Requirements 4

2. PASS system files and installation procedure. 5

3. Introduction to PASS. 6

3.1. Biological Activity Presentation 6

3.2. Chemical Structure 6

3.3. Equivalent Substances 7

3.4. External Files of Substances 7

3.5. Training Set 8

3.6 SAR Base 9

3.7. Prediction Results 10

4. PASS Interface 11

5. Opening SAR Base 15

6. Creating SAR Base 17

7. Adding New Data to SAR Base 19

8. Executing the Training Procedure 24

9. Selecting Activity Types to be Predicted 26

10. Defining Predicted Activity Spectra Selection Criterion. 31

11. Predicting the Biological Activity Spectra. 32

12. Viewing Prediction Results. 34

13. Viewing Basic Information 43

14. Interpreting Prediction Results 47

15. Suplements 49

15.1. Computer-aided methods in pharmaceutical R&D 49

15.2. Biological Activity Presentation. 50

15.3. Chemical Structure Description 51

15.4. Training Set. 53

15.5. Training Procedure. 55

15.6. PASS Validation 57

ABBREVIATIONS

DBMS - Database Management System

IEP - Invariant Error of Prediction

MNA - Multilevel Neighbourhoods of Atoms

PASS - Prediction of Activity Spectra for Substances

QSAR - Quantitative Structure-Activity Relationships

QSPR - Quantitative Structure-Property Relationships

SAR - Structure-Activity Relationships

MOLfile describes a single molecular structure format (MDL Information Systems, Inc.)

SDfile – Structure-data file. This format describes structure and data for any number of molecules (MDL Information Systems, Inc.)

1. HARDWARE AND SOFTWARE SYSTEM REQUIREMENTS

PASS is a software product for predicting the biological activity spectra for chemical substances on the basis of their structural formulas.

PASS is a commercially available software product.

Operating environment – Windows 98/NT/2000/XP.

PASS requires minimum 64 Mb RAM.

The installation procedure requires 40 Mb free hard disc space.

Chemical structure information is represented as SDfiles or MOLfiles (formats of MDL Information Systems, Inc.), which can be exported from many chemical Database Management Systems. The only requirement - these files should correspond with the ISIS/Base V2000 standard.

PASS variants:
PASS Professional – PASS environment, PASS SAR Base, SAR Base creating and editing options.
PASS – PASS environment, PASS SAR Base.
PASS Light – PASS environment, SAR Base creating and editing options.

If you have any questions about PASS program, please write us by E-mail: **pass@ibmc.msk.ru**

2. PASS SYSTEM FILES AND INSTALLATION PROCEDURE

*.**ADF** is an activity description file (only for supplying SAR base).

*.**DFW** is a configuration file (**D**irectories, **F**iles, **W**indows).

PASSACL.TXT file contains the list of pharmacotherapeutic effects, biochemical mechanisms, toxic & adverse effects and names of metabolising enzymes.

*.**HLP** is a help file.

***.HST** is a history type file. There are two types of such files:

> ***.HST** is history of SAR base creation and modification. It contains SAR base name; the list of errors; the number of substances in the created SAR base, etc.
>
> ***.HST** is prediction history file. It contains: (1) names of input and output files; (2) the identifiers of the compounds for which the equivalent structures are found; (3) errors in structures causing the failure of prediction.

***.LST** is the file, which contains SAR Base information; the number of compounds revealing each particular type of activity; the maximum error of prediction for each particular type of activity in percentage estimated in leave one out cross-validation procedure.

***.SAR** is SAR base file.

Installation procedure

1. Insert PASS CD to the CD-ROM drive.
2. Open INSTALL folder.
3. Run **setup.exe** file.
4. When Setup starts, follow the instructions on your screen.

3. INTRODUCTION TO PASS

PASS (*Prediction of Activity Spectra for Substances*) software product is developed as a tool for evaluation of general biological potential in a molecule under study. The latest version of PASS 2.005 (December 2005) predicts more than 2100 kinds of biological activity with the mean prediction accuracy of about 92%. The default list of predictable biological activities includes main and side pharmacological effects, (e.g., antihypertensive, hepatoprotective, sedative, etc.), mechanisms of action, (5-hydroxytryptamine agonist, acetylcholinesterase inhibitor, adenosine uptake inhibitor, etc.) and specific toxicities (mutagenicity, carcinogenicity, teratogenicity, etc.). Information about novel activities and new compounds can be straightforwardly included into PASS, and used for further prediction of biological activity spectra for new chemical compounds.

3.1. Biological Activity Presentation

In PASS biological activities are described qualitatively (active or inactive). Reflecting the result of chemical compound's interaction with a biological object, the biological activity depends on both the compound's molecular structure and the terms & conditions of the experiment. Therefore, structure-activity relationship analysis based on qualitative presentation of biological activity describes general "biological potential" of the molecule being studied. On the other hand, qualitative presentation allows integrating information concerning compounds tested under different terms and conditions and collected from many different sources as in the PASS training set.

Any property of chemical compounds, which is determined by their structural peculiarities, can be used for prediction by PASS. It is clear, that the applicability of PASS is broader than the prediction of biological activity spectra. To extend PASS application to other properties, the user needs the appropriate training set (see below).

3.2. Chemical Structure

The 2D structural formulae of compounds were chosen as the basis for description of chemical structure, because this is the only information available in the early stage of research (compounds may only be designed but not synthesized yet). Structure descriptors, which we

call "Multilevel Neighbourhoods of Atoms" (MNA), were designed for chemical structure representation (Filimonov D., et al. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 666-670).

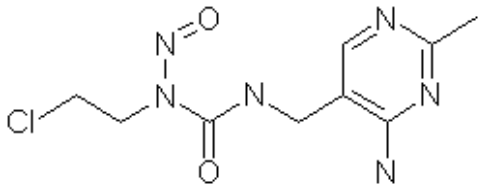
3.3. Equivalent Substances

The substances are considered equivalent in PASS if they have the same MNA descriptors set. Since MNA descriptors do not represent the stereochemical peculiarities of a molecule, the substances, which have only stereochemical differences in the structure, are formally considered equivalent.

3.4. External Files of Substances

PASS uses SDfile or MOLfile as external sources of structural and activity data to prepare both the SAR Base (see below) and set of substances to be predicted. SDfiles (*.sdf) can be exported either from ISIS/Base 2.0+ (MDL Information Systems, Inc.) or from any other Molecular Editor or DBMS, which has the option of SDfile's export.

The example of SDfile for Nimustine molecule:



Nimustine

```
  -ISIS-  10079915102D

 18 18  0  0  0  0  0  0  0  0999 V2000
    2.7250   -2.0667    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   -0.8417   -2.4833    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.5542   -2.0750    0.0000 N   0  0  3  0  0  0  0  0  0  0  0  0
    2.0125   -2.4792    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.3042   -2.0667    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.5583   -1.2500    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   -0.1250   -2.0708    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    2.0125   -0.8292    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    2.7250   -1.2417    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.3042   -1.2417    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.5875   -2.4833    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.8375   -3.3083    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.8417   -0.8333    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    2.0125   -3.3042    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   -2.2708   -2.4875    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -3.7000   -2.4875    0.0000 Cl  0  0  0  0  0  0  0  0  0  0  0  0
   -2.9833   -2.0750    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.4417   -0.8333    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  2  7  1  0  0  0  0
  3  2  1  0  0  0  0
  4  1  1  0  0  0  0
  5  4  2  0  0  0  0
  6  3  1  0  0  0  0
  7 11  1  0  0  0  0
  8  9  1  0  0  0  0
  9  1  2  0  0  0  0
 10  8  2  0  0  0  0
 11  5  1  0  0  0  0
 12  2  2  0  0  0  0
 13  6  2  0  0  0  0
 14  4  1  0  0  0  0
 15  3  1  0  0  0  0
 16 17  1  0  0  0  0
 17 15  1  0  0  0  0
 18  9  1  0  0  0  0
  5 10  1  0  0  0  0
M  END
>  <ID> (689)
689

>  <name_inn> (689)
Nimustine

>  <ACTIVITY> (689)
Antineoplastic, alkylator
DNA synthesis inhibitor
Teratogen

$$$$
```

MOLfile (*.mol) can be prepared with chemical editor ISIS/Draw 2.00 and higher (MDL Information Systems, Inc.). ISIS/Draw is available free for personal or non-commercial use from the MDL web site **http://www.mdl.com**.

3.5. Training Set

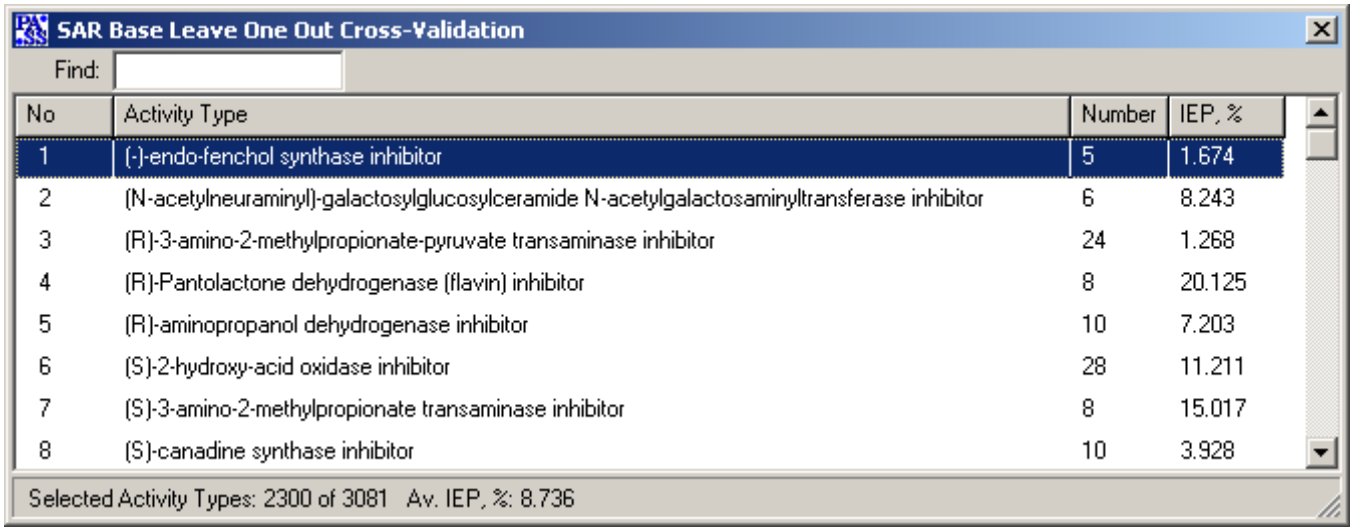
The current PASS training set includes 60722 substances. All these substances were thoroughly selected. For more information see supplements (15.3, 15.4). SAR Base is created on the basis of PASS training set.

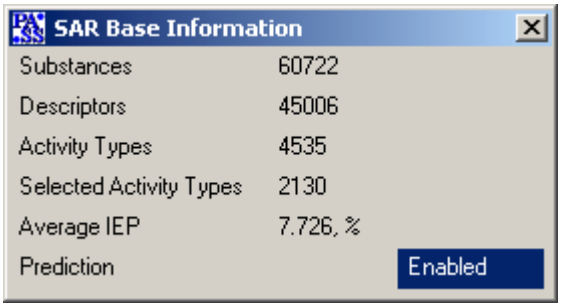
3.6. SAR Base

It is a complex knowledge base, containing vocabularies of MNA descriptors and activity names, the database of the substance structures presented by MNA descriptors, their biological activity types and data on SAR. Unfortunately it is currently impossible to collect sufficiently large number of active substances for all PASS activity types using available sources, that is why some activity types are presented by more than thousand substances, while some others are only represented by a few ones.

Before predicting you should decide what kind of the SAR Base you are going to use. You may use the SAR Base, supplied with PASS software, or create your own one (the last option is provided only by PASS Professional).

When the training procedure is over the following windows indicated this state of the SAR Base appears.





To get information about current SAR base use **View** menu command (see chapter 13).

3.7. Prediction Results

As a result of prediction PASS presents the biological activity spectra for each substance. It is the list of biological activity types for which the calculated probability to be revealed (**Pa**) is more than the calculated probability not to be revealed (**Pi**). Taking into account that some substances from the training set are considered inactive formally, the estimated value of **Pa** is more reliable. There are some other factors, which essentially influence on **Pa** absolute value: the number and diversity of substances revealing such activity in the training set, recall ratio, etc. In general, the higher **Pa** value is the higher probability for a studied substance to be structurally similar to known biologically active substances from the training set is.

The result of prediction is valuable at planning the experiment, but one should take into account some additional factors: particular interest to some kinds of activity, desirable novelty of a substance, available facilities for experimental testing, etc. Actually, each choice is always the compromise between the desirable novelty of studied substance and risk to obtain the negative result in testing.

You can examine prediction results saved in the appropriate SDfile. Use **File|Open** menu command (see chapter 12).

4. PASS INTERFACE

To start **PASS** - double-click PASS icon control (PASS shortcut) ;

or

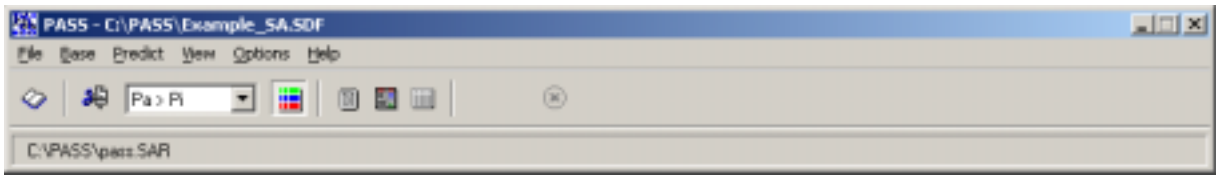
run **pass.exe** from PASS folder;

or

run **pass.exe** directly from PASS distributive CD.

The main window of PASS

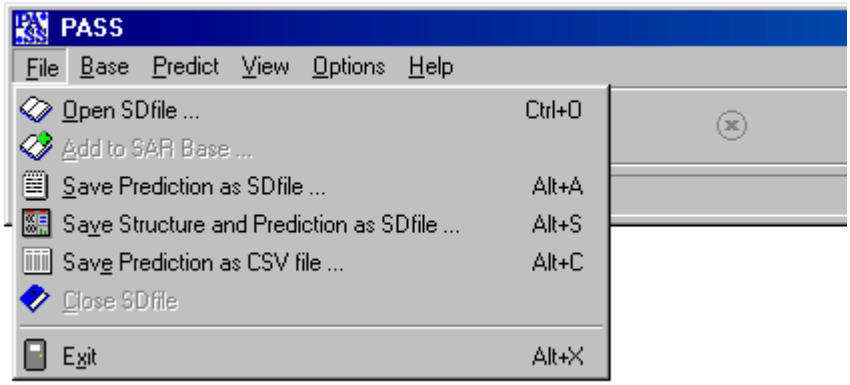
The main window contains menu, speed buttons, drop down list for the selection criterion, progress bar, status panel.



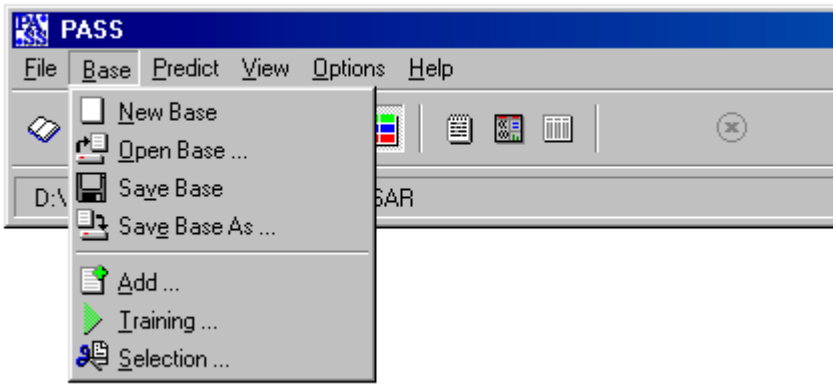
You can use either **menu commands** or the speed buttons to execute **PASS** system procedures.

The main menu items are:

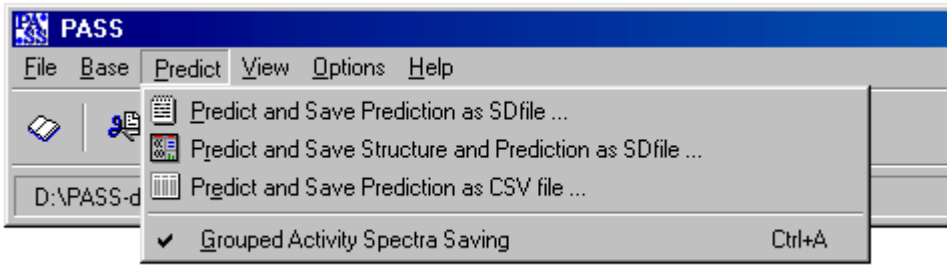
File



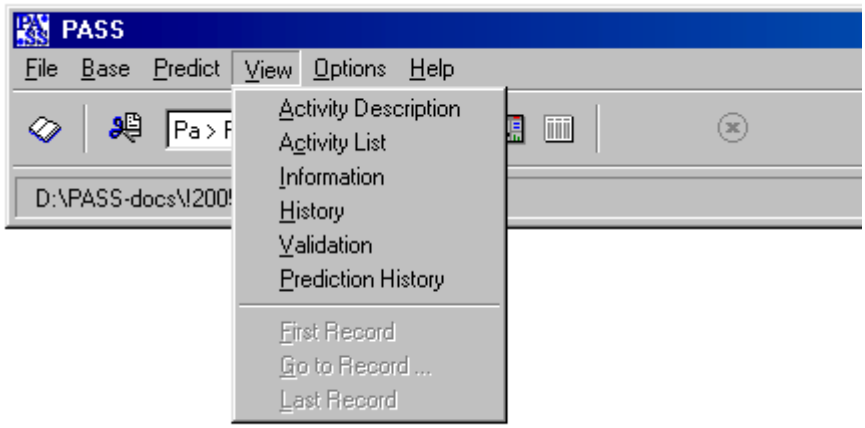
Base



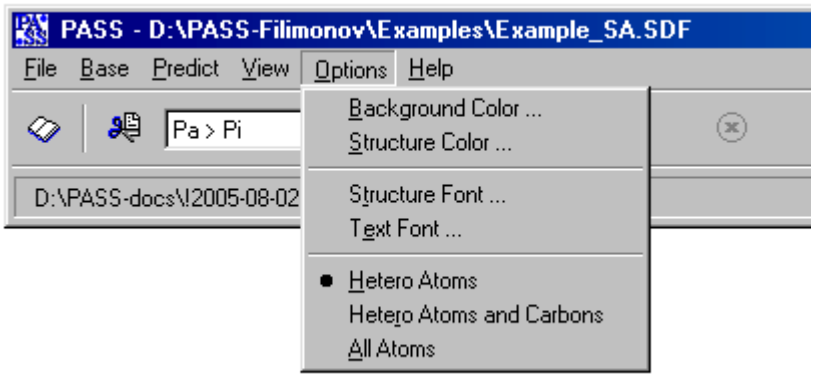
Predict



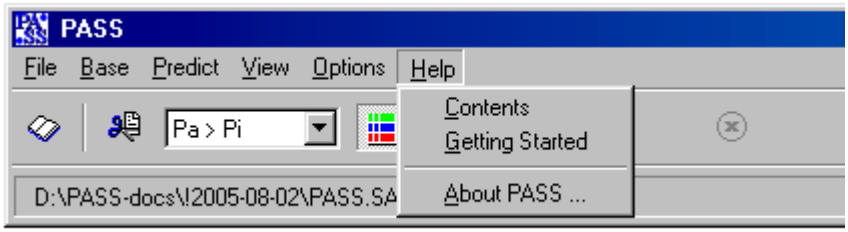
View



Options



Help



All these item menus are described in more detail below.

Use the mouse, the keyboard (**F10**) or key combinations: **Alt+F**, **Alt+B, Alt+P**, **Alt+V, Alt+O, Alt +H** to choose the particular menu item.

- Press F1 to call the context sensitive **Help** in the main window, in all dialogs and sub-windows of the PASS software.
- The speed buttons' hints will appear when the mouse cursor points to the button.
- Use Help menu command for more information you need.
- Use Alt+F4 shortcut, **File|Exit** (**Alt+X**) menu command or button in upper right corner of PASS main window to quit PASS.

An existing SAR Base is loading by default also as the last opened SDfile (for the first running of PASS the example.sdf file is opening). Now you can use this SAR Base for the prediction or updating. Don't forget to carry out the training procedure in case of SAR Base change. After that the most of the commands become available (see Figure below).

Each button has the following functions:

Opens *.SDf file or MOL file for viewing.

Selects activity types to be predicted.

It is a toggle. If it is active, you can save prediction results as SDfile in which activity names are divided into three groups: Effects, Mechanisms or Toxicity.

Predicts and Saves prediction results as a SDfile.

Predicts and Saves structure and prediction results as SDfile.

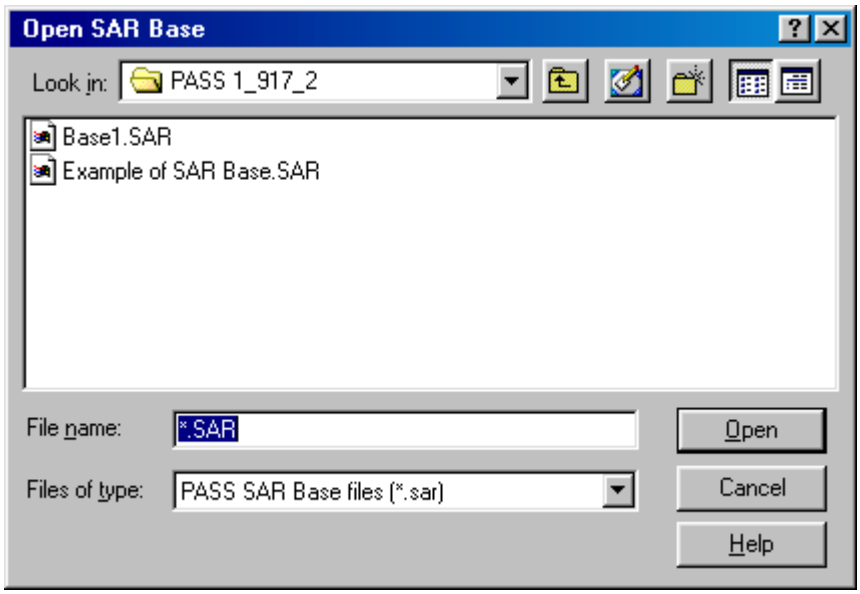
Predicts and Saves prediction results as CSV file.

Interruption of any current process (reading from file, saving into file, training, creating the set for the prediction, prediction).

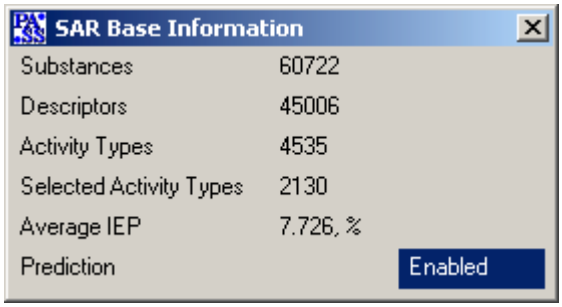
5. OPENING SAR BASE

Use **Base|Open Base** menu command to open an existing SAR Base.

The following window appears.



Use **View|Information** menu command to display statistical data about current SAR Base.



Where,

Substances the number of compounds in the SAR Base;

Descriptors the number of different MNA descriptors;

Activity Types the number of activity types in the SAR Base;

Selected Activity Types the number of activity types used for the biological activity spectra prediction;

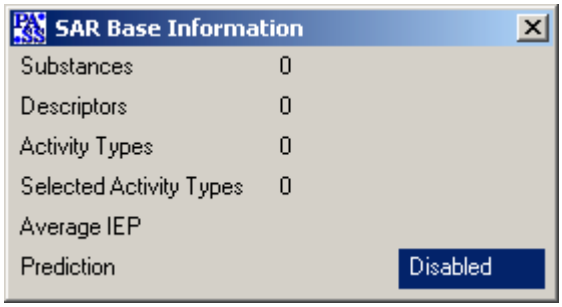
Average IEP the mean value of IEP (Invariant Error of Prediction) for all Selected Activity Types: $(\sum IEP_n)/n$;

Note! The average IEP (Invariant Error of Prediction) is calculated by leave-one-out cross-validation procedure.

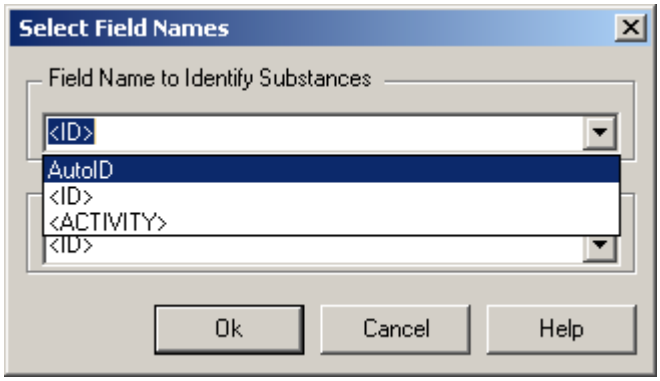
6. CREATING SAR BASE

(only for PASS Professional and PASS Light)

Use **Base |New Base** menu command to create a new SAR Base. **SAR Base Information** appears. Zero values of all parameters indicate that SAR base is empty at this moment.

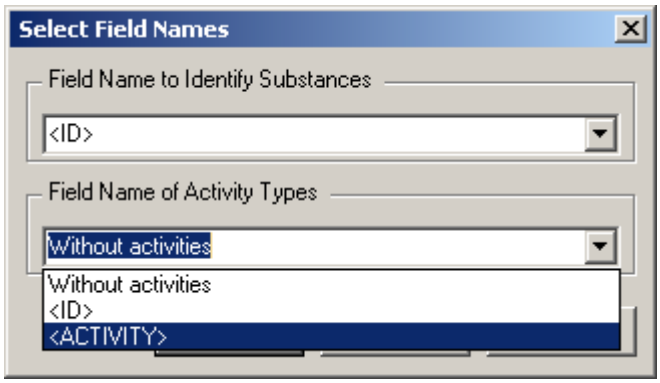


Then use **Base|Add** menu commands to select an SD file containing structures of compounds and names of their biological activity, which you want to have in your SAR Base. The example of SDfile is presented at the page 8. You may also use **File|Add to SAR Base** menu command to add to the new SAR Base information from the current SD file. The **Select Field Name** window will appear. Then you should select the substance identifier (**ID** as a rule).



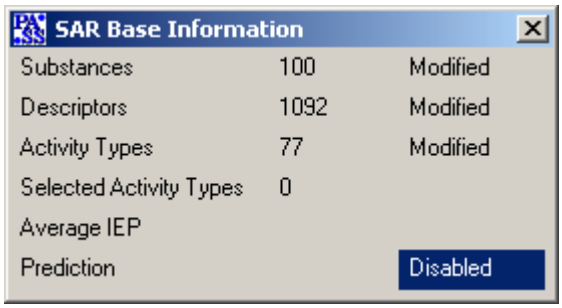
AutoID means that the substance identifier **(ID)** will be it's number of a record.

Then you should select the name of the field containing biological activity types (**Activity** as a rule).

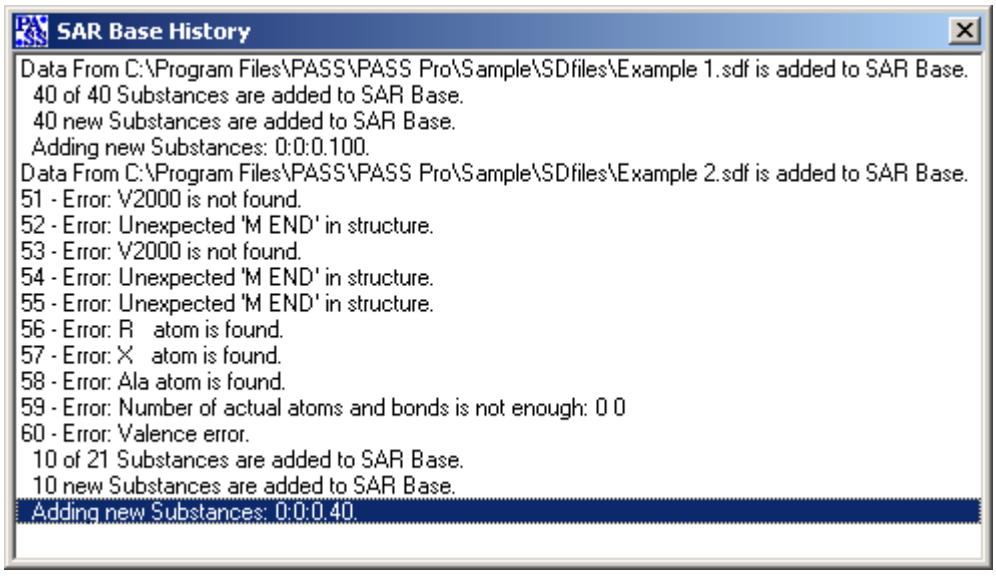


Without activities means that only information about structural formula will be added to **SAR Base.**

Click **OK** to start the procedure of the SAR Base creating. When it is over the **SAR Base Information** window is refreshed.



Attention! All structures of substances are checked. If the structure is not correct the substance is not included into the SAR Base. The errors are displayed in the corresponding window, for example:



Attention! Currently the maximum of activity types is equal to 16383, and maximum of substances in SAR Base is equal to 65520.

7. ADDING NEW DATA TO SAR BASE

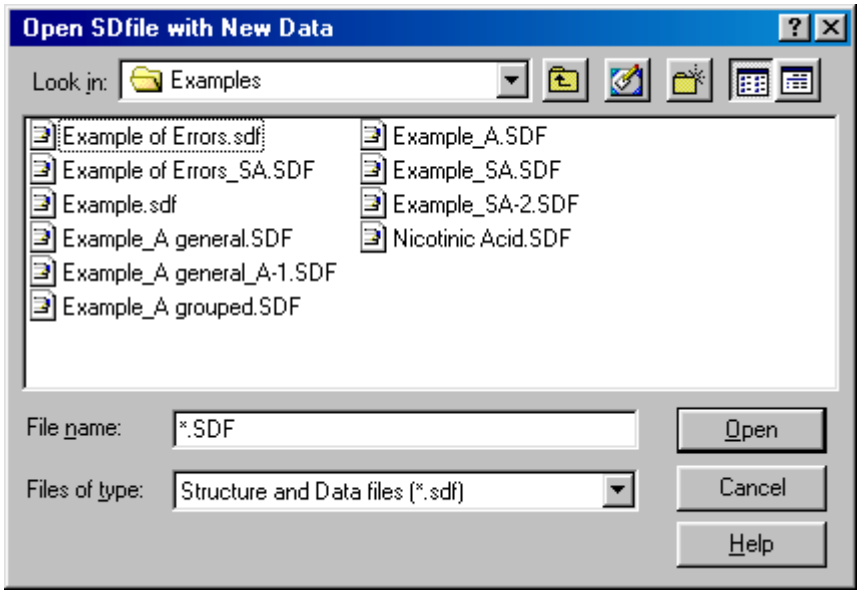
(only for PASS Professional and PASS Light)

The procedure is similar to the creation of SAR Base, but has some features.

First of all you should open SAR Base, which you are going to modify.

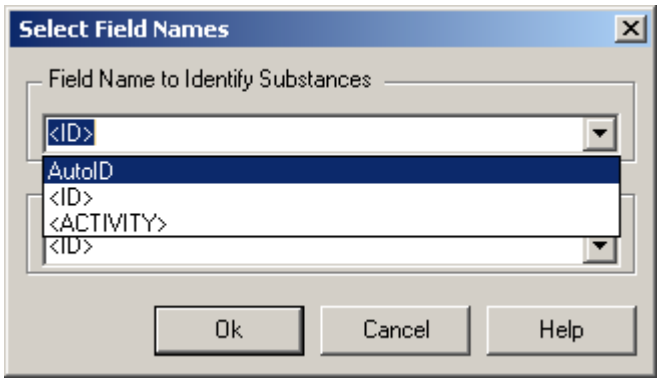
Use **Base|Open Base**.

Then Use **Base|Add** menu command to select the SDfile consisting substances, which you want to add to your SAR Base.

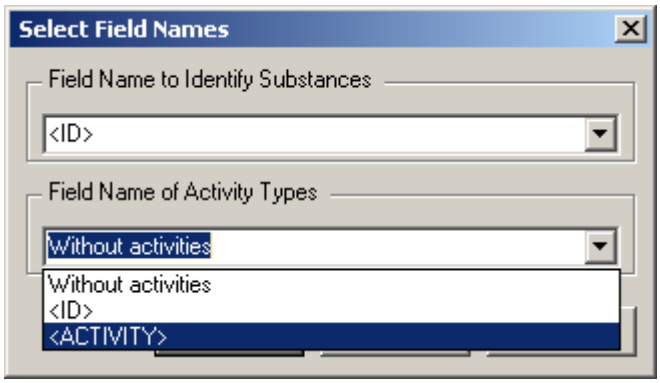


This SDfile should contain data on structure and biological activity types of substances. After opening this file you should select the substance's identifier field and the field containing the list of its biological activity names.

The **Select Field Name** window will appear. Then you should select the substance identifier (**ID** as a rule).



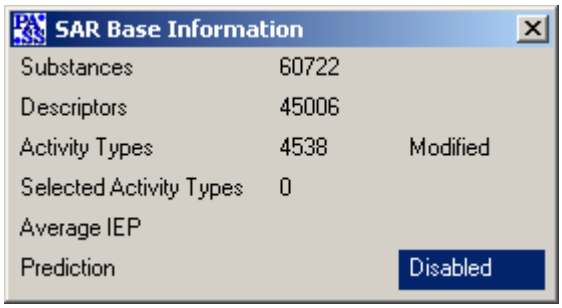
AutoID means that the substance identifier **(ID)** will be it's number of a record.



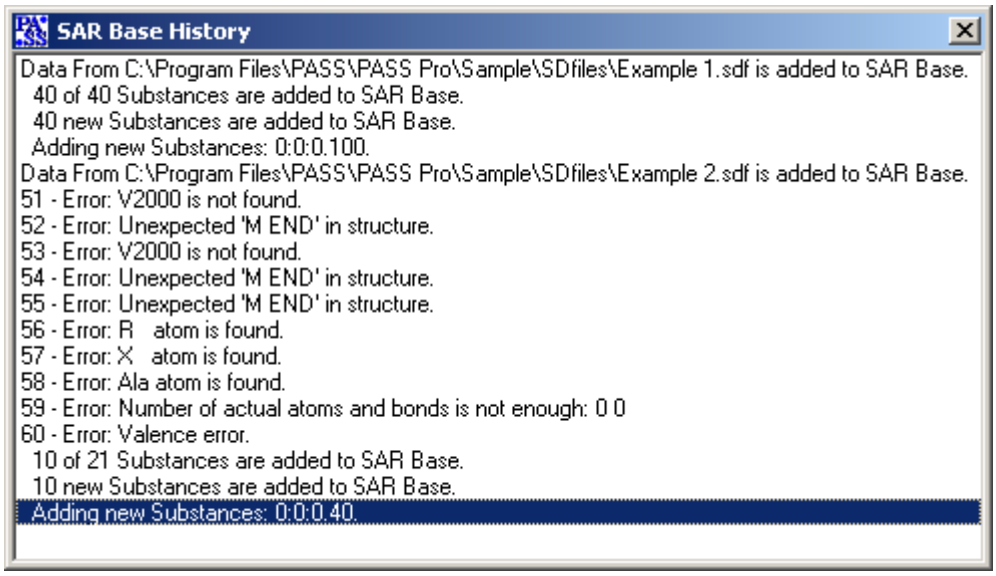
Then you should select the name of the field containing biological activity types (**Activity** as a rule).

Without activities means that only information about structural formula will be added to **SAR Base.**

Click **OK** to start the procedure of the SAR Base creating. When it is over the **SAR Base Information** window is renewed.



Attention! All structures of substances are checked. If the structure is not correct the substance is not included into the SAR Base. The errors are displayed in the corresponding window, for example:

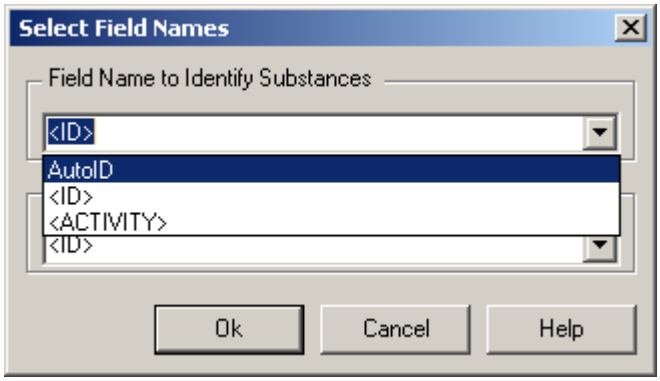


Attention! Currently the maximum of activity types is equal to 16383, and maximum of substances in SAR Base is equal to 65520.

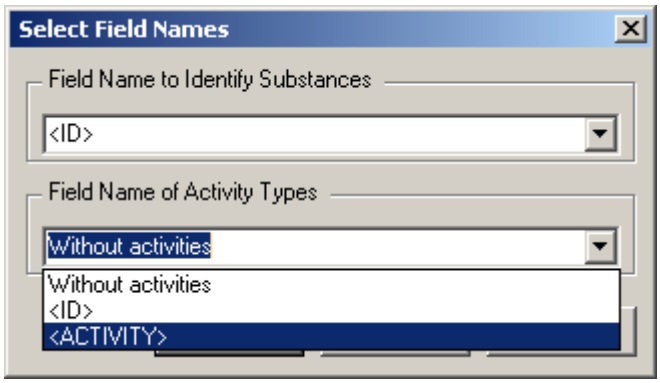
If you want to use old SAR Base later, you should save this new SAR Base as *.SAR file with the another name.

Another way to add new data to SAR Base:

1. Open SDFile (**File|Open SDfile …** or press button ) with data you want to add.
2. Then use **File|Add to SAR Base …** menu command.
3. The **Select Field Name** window will appear. Then you should select the substance identifier (**ID** as a rule).

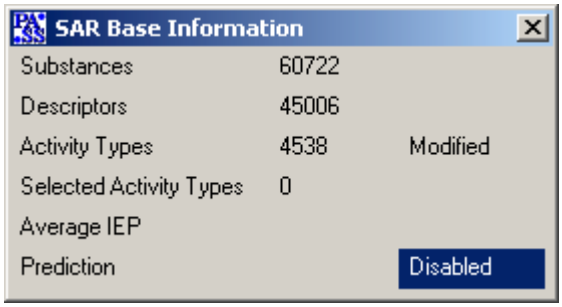


AutoID means that the substance identifier (**ID)** will be it's number of a record.
Then you should select the name of the field containing biological activity types (**Activity** as a rule).

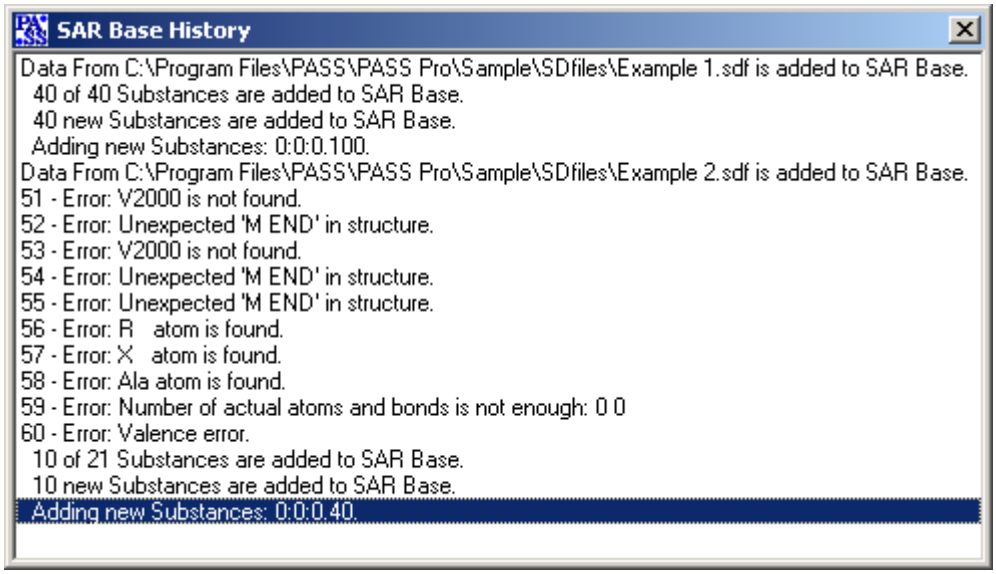


Without activities means that only information about structural formula will be added to **SAR Base.**

Click **OK** to start the procedure of the SAR Base creating. When it is over the **SAR Base Information** window is renewed.



Attention! All structures of substances are checked. If the structure is not correct the substance is not included into the SAR Base. The errors are displayed in the corresponding window, for example:



Attention! Currently the maximum of activity types is equal to 16383, and maximum of substances in SAR Base is equal to 65520.

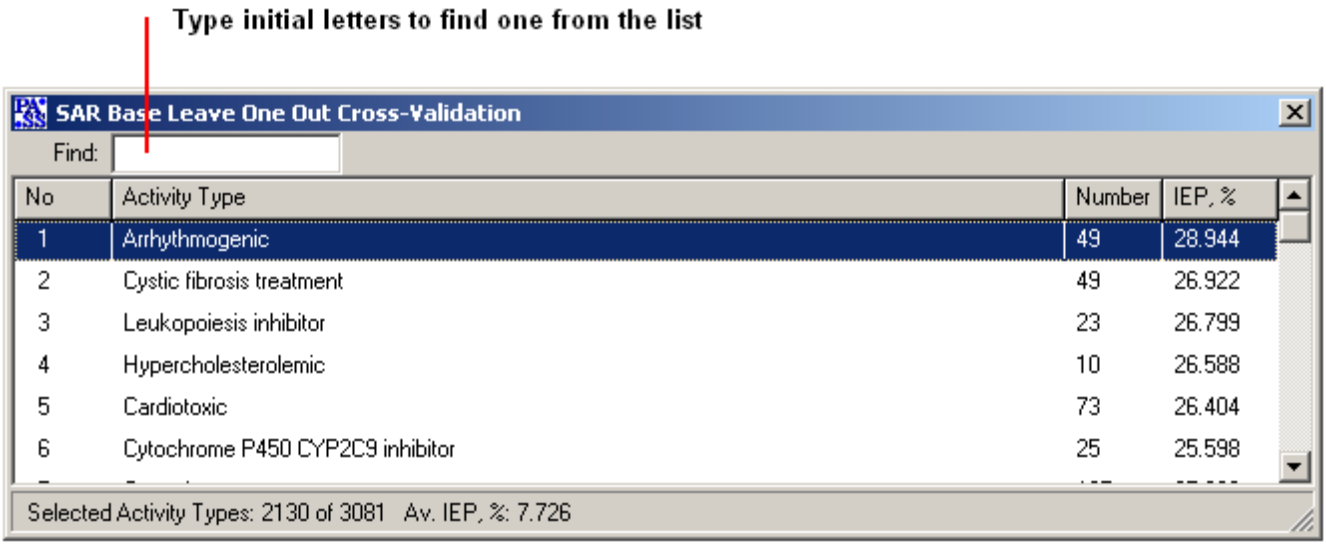
Select the substance's identifier field and the field containing the list of its biological activity names.

If you want to use old SAR Base later, you should save this new SAR Base as *.SAR file with the different name.

Attention! Please, pay attention to the using biological activity names. If your terms are the same as in the existing SAR Base, you must be sure that there is no equivocacy in your list of the biological activities. To avoid such situation, use different names for such activities.

8. EXECUTING TRAINING PROCEDURE

Use **Base|Training** menu command to start the training procedure for determining structure-activity relationships. During the training procedure **SAR Base Leave One Out Cross-Validation** window is displayed. You can interrupt the training procedure at any moment using button.



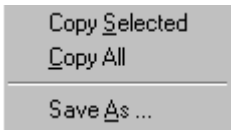
The **Activity Type** column displays an unsorted list of activity names.

The **Number** column displays the number of substances revealing a particular type of activity.

The **IEP** column displays the invariant error of the prediction. IEP value for every type of activity is obtained by Leave One Out Cross-Validation method (see Supplements).

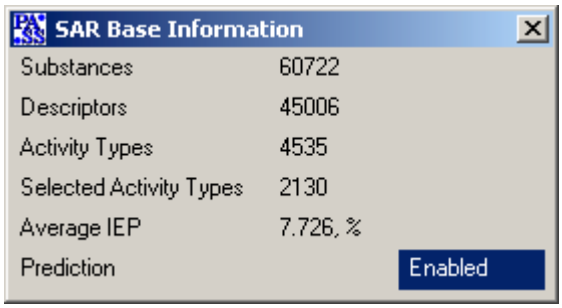
The bottom status bar displays the number of biological activity types from the SAR Base for which the training procedure has been executed and the average IEP for these activities.
When the training procedure is over you can look for a certain activity type using **Find** box.

Click right mouse button mouse to invoke the following menu.



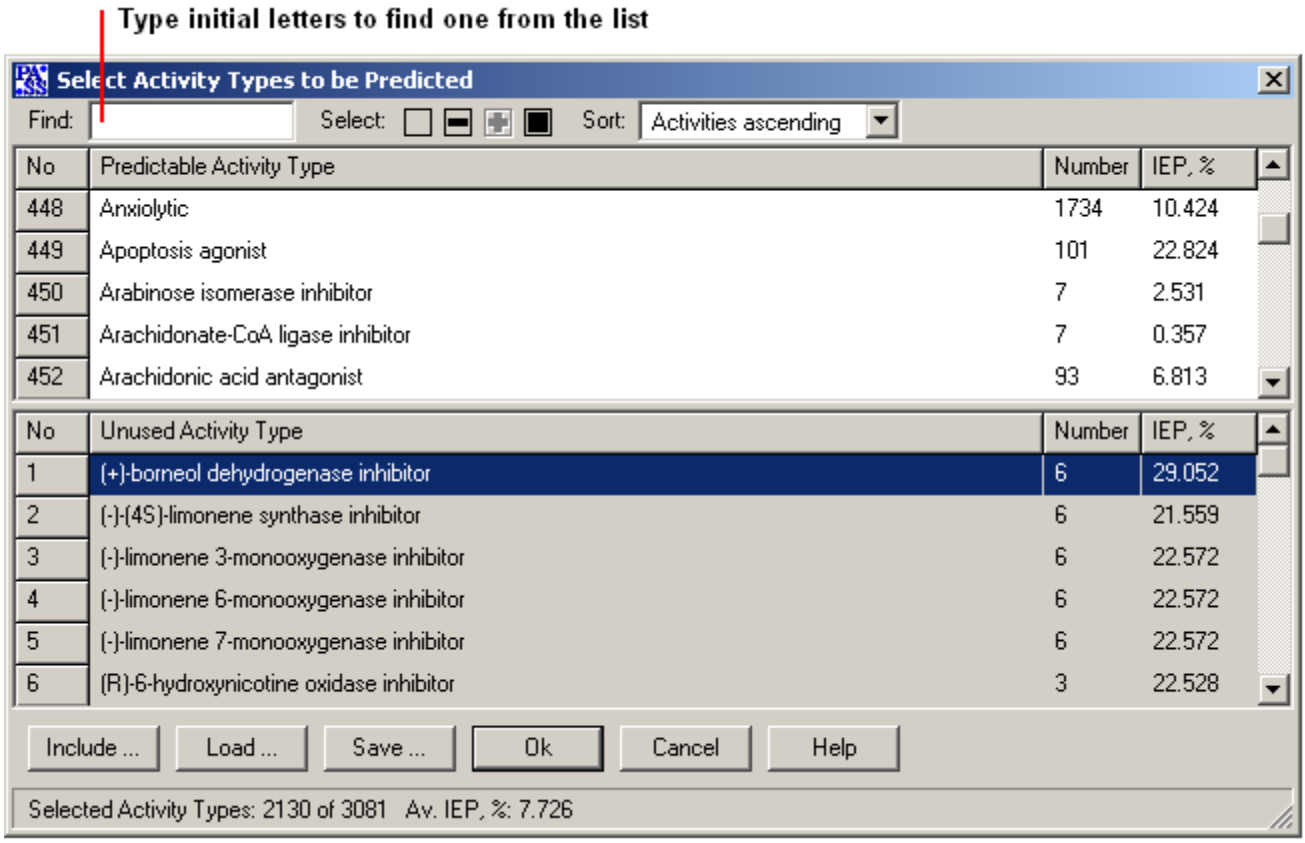
Use it to copy a selected string or all to the clipboard. **Save as** menu command allows you to save this information in a text file.

When the training procedure is over and new SAR base is saved the following window appears. It displays the statistics information on this new SAR base.



9. SELECTING ACTIVITY TYPES TO BE PREDICTED

Use **Base|Selection** (or ) menu command to perform the selection of activity types. The following window appears.



It displays the list of activity types, which will be predicted.

The **Predictable Activity Type** column displays activity names from SAR Base selected for activity spectra prediction.

The **Unused Activity Type** column displays activity names from SAR Base, which are not used for activity spectra prediction.

The **Number** column displays the number of substances containing in the training set and revealing a particular type of activity.

The **IEP** column displays the invariant error of the prediction. IEP value for every type of activity is obtained by leave one out cross-validation method.

The bottom status bar displays the number of selected biological activity types from SAR Base which will be used for prediction of biological activity spectra (2130), the number of all predictable biological activity types in the SAR Base (3081) and average IEP (7.726%) for selected activities.

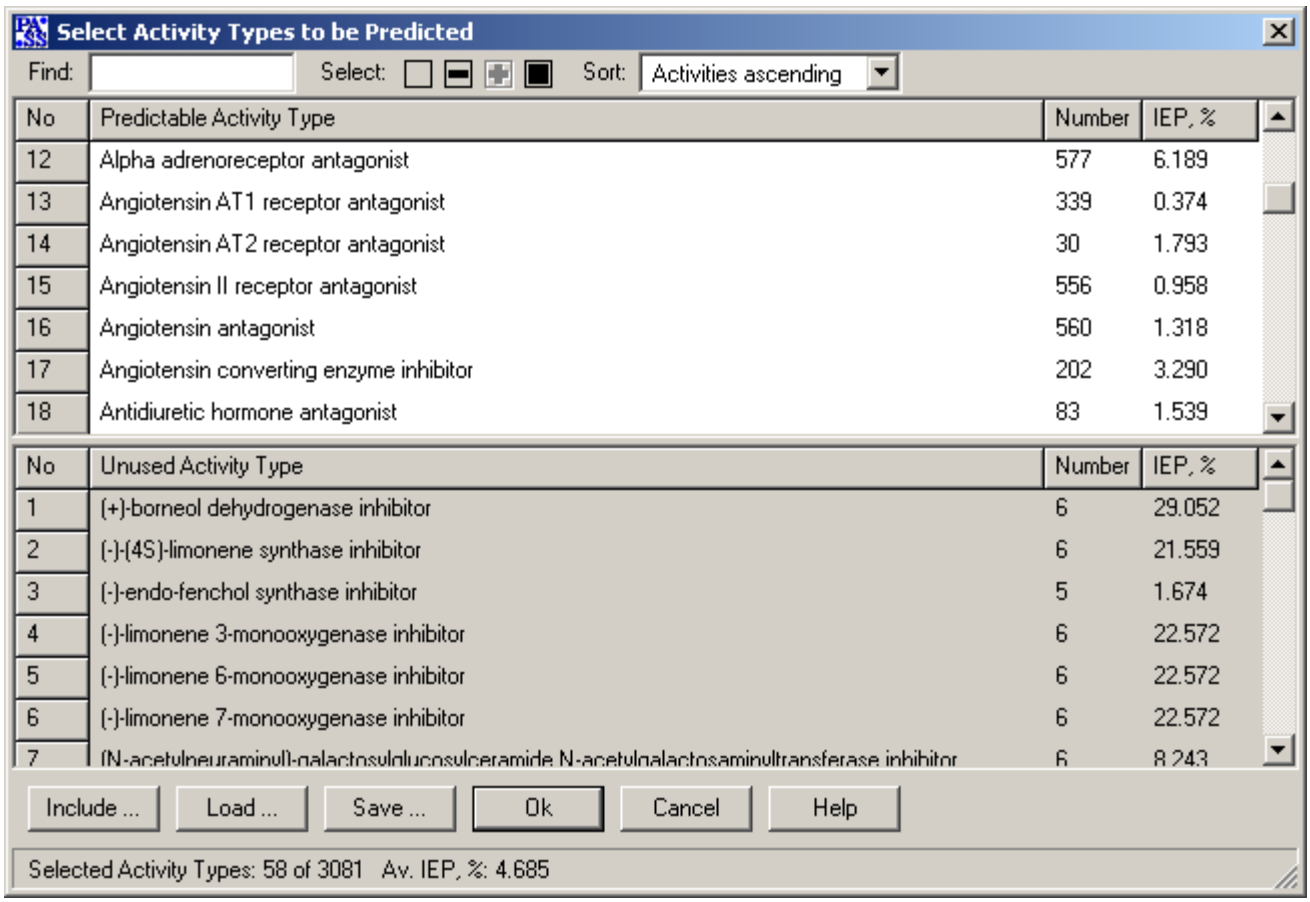
If you want to exclude some activity type(s) from the list of activities you should mark those activity(s) at the Predictable Activity Type column (press the left mouse button for the activity selection) and click  button or **Delete**. All these activity types will transfer to the Unused Activity Type table.

If you want to add some activity type(s) to the list of activities from the Unused Activity Type column you should mark those activity(s) at the Unused Activity Type column (press the left mouse button for the activity selection) and click  button or **Insert**. All these activity types will transfer to the Predictable Activity Type table.

You may sort activity names for more convenience according to your criteria. It will help you to make up selection promptly.
The names of activity types may be sorted in several ways:

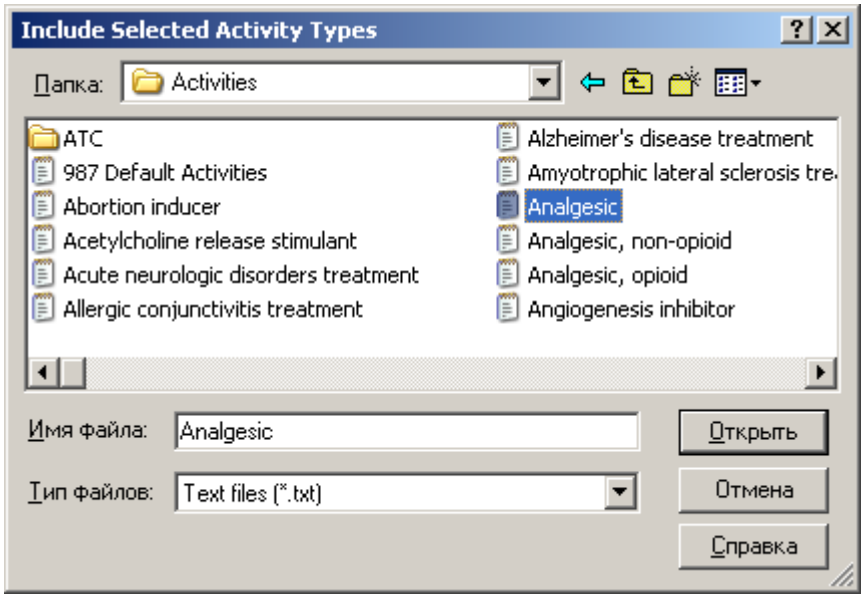
Activities ascending	Alphabetically.
Activities descending	Alphabetically (reverse order).
Number ascending	Ascending order of number of the active substances in the SAR Base revealing this activity type.
Number descending	Descending order of number of the active substances in the SAR Base revealing this activity type.
IEP ascending	Ascending order of invariant prediction error.
IEP descending	Descending order of invariant prediction error.

You may select only those activity types for prediction, which match the following criteria: (1) IEP is less than certain cut-off value (for example, IEP<20%); (2) the number of active compounds in the SAR Base is more than certain cut-off value (for example, N>5). Of course you can select only those activity types, which you are interested in (for example, the activity types associated with the treatment of essential hypertension):

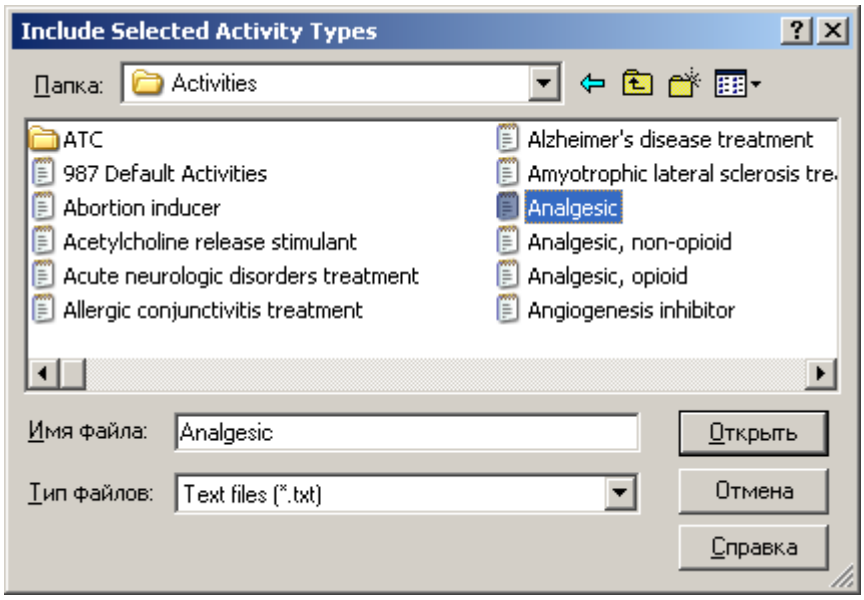


You may also select all activity types by clicking ■, or cancel selection of all activity types by clicking □.

- Click **Include…** button to add activity types to list of predicted activity types from one of your text file (*.txt) with a particular variant of selected activity list. Only those activity types, which names are coincided with activity names in SAR Base, will be selected (that are only those, which are from Unused list of activities).

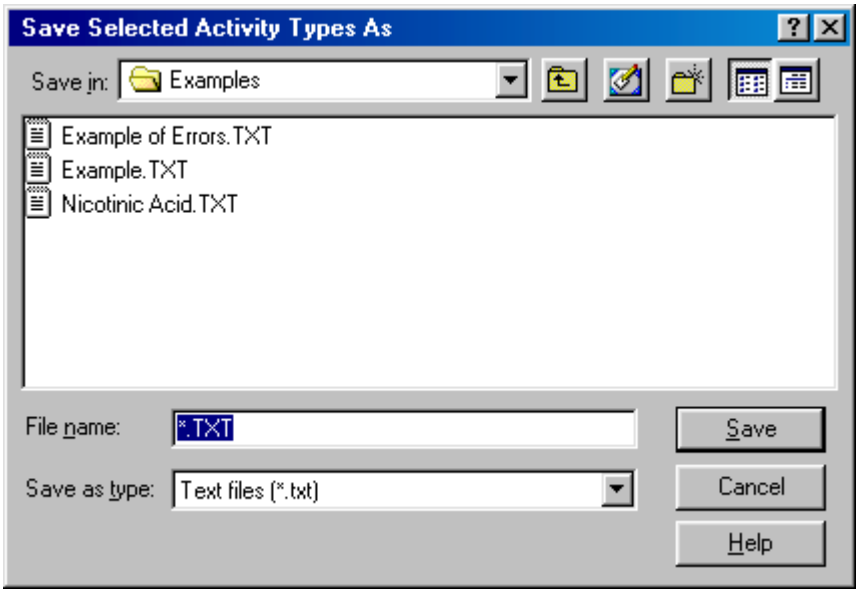


- Click **Load…** button to load one of your text file (*.txt) with a particular variant of selected activity list. Only those activity types, which names are coincided with activity names in SAR Base, will be selected (that are only those, which are from Unused list of activities).

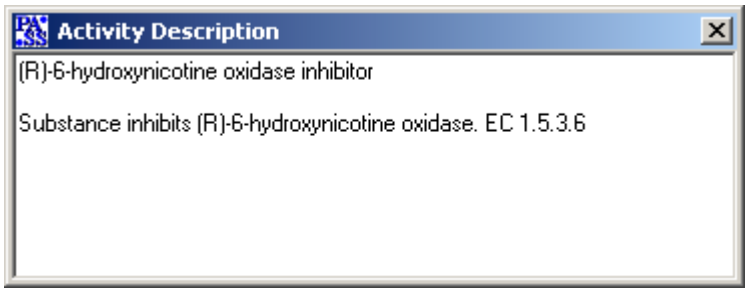


Attention! The lists of PASS activity types according to ATC classification and for the main pharmacotherapeutical effects are provided in .txt files. They are in the “activities” folder.

- Click **Save…** button to save the list of selected activity types in text file (*.txt).



Attention! Basic information (**Activity Description** window) about an activity type is appeared automatically when you point this activity out. This window is available only in case of standard SAR base.



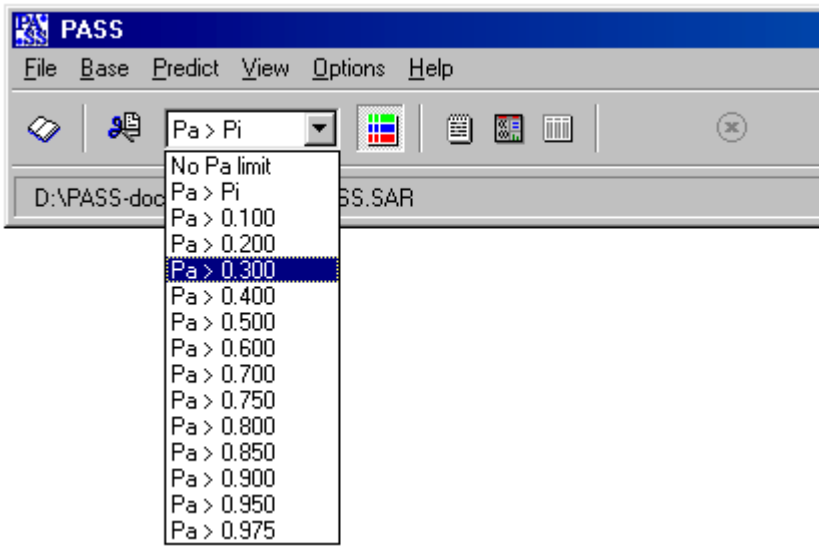
10. DEFINING PREDICTED ACTIVITY SPECTRA SELECTION CRITERION

During the PASS run **Pa** and **Pi** are calculated for every activity type from SAR Base list. **Pa** is a probability that the substance may reveal this kind of activity. **Pi** is a probability that the substance does not reveal this activity. Their values vary from 0.000 to 1.000.

The selection criterion is **Pa** > **Pi** by default. Only activity types with **Pa** > **Pi** are considered as possible.

You can specify the criterion according to which activity types are included in activity spectra. Only those activity types that meet the criterion you specified will be included in the predicted activity spectra.

Drop Down List for Selection Criterion

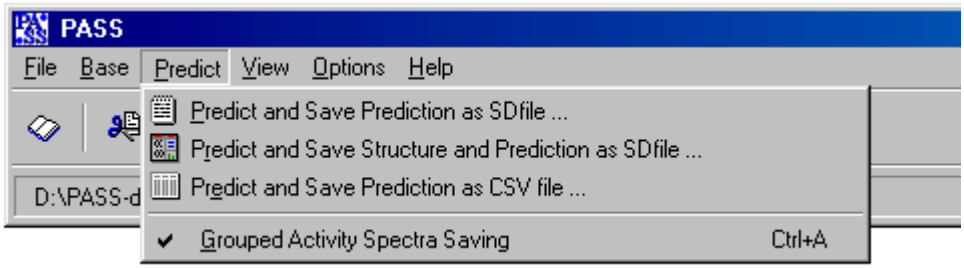


Choose the **Pa** value to define the selection criterion for predicted activities. Only activities with **Pa** value greater than the chosen threshold will be given in predicted activity spectra.

You may also display all activity types even with **Pa** < **Pi.** For this you should select “No Pa limit” threshold. This option can be useful for further clustering of compounds according to the predicted biological activity spectra, etc.

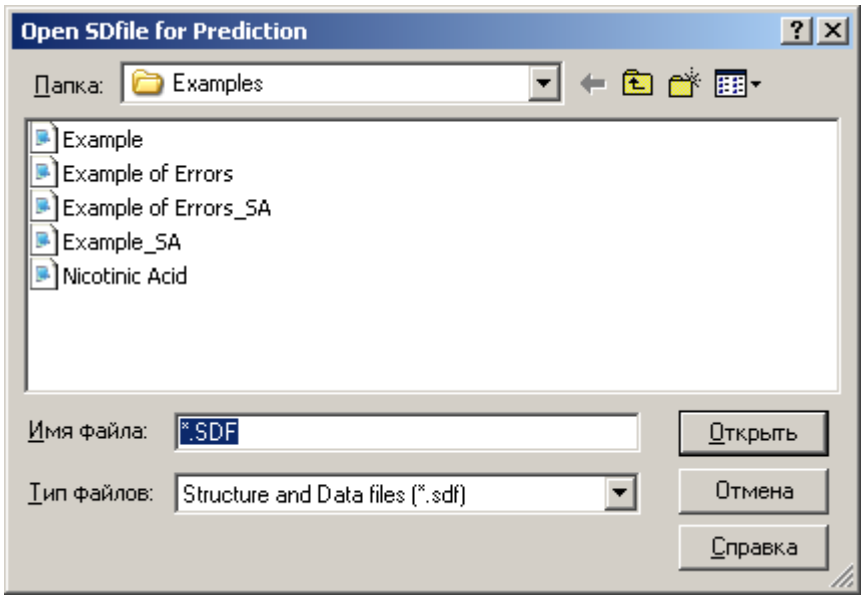
11. PREDICTING THE BIOLOGICAL ACTIVITY SPECTRA

Use **Predict** menu commands to predict and save the biological activity spectra for compounds. The information about these compounds should be prepared as SDfile or Molfile before the prediction.



Predict menu commands	Description
Predict|Predict and Save Prediction as SDFile…	Opens SD or Molfile for prediction. Predicts and saves the prediction results as Sdfile (without structures)
Predict and Save structure and Prediction as SDfile…	Opens SD or Molfile for prediction. Predicts and saves structures and prediction results as SDfile.
Predicts and Saves Prediction as CSV file…	Opens SD or Molfile for prediction. Predicts and Saves the prediction results as CSV file.
Grouped Activity Spectra Saving	Mark "**Grouped Activity Spectra Saving"** (or press button) to save the prediction results as SDfile in which activity names are divided into four groups: Effects, Mechanisms, Toxicity and Metabolism.

- If you want to save your results as SDFile use **Predict\|Predict and Save Prediction as SDfile…** menu command (or button). In this case you save prediction results in an appropriate field in addition to existing fields. After choosing this command you should open SD or MOL file in the **Open SDfile for Prediction** window, which appears at once.



Then use the **Save SDfile with Prediction as** dialog box to type or change the name of the prediction result file. Press **Save** to start prediction.

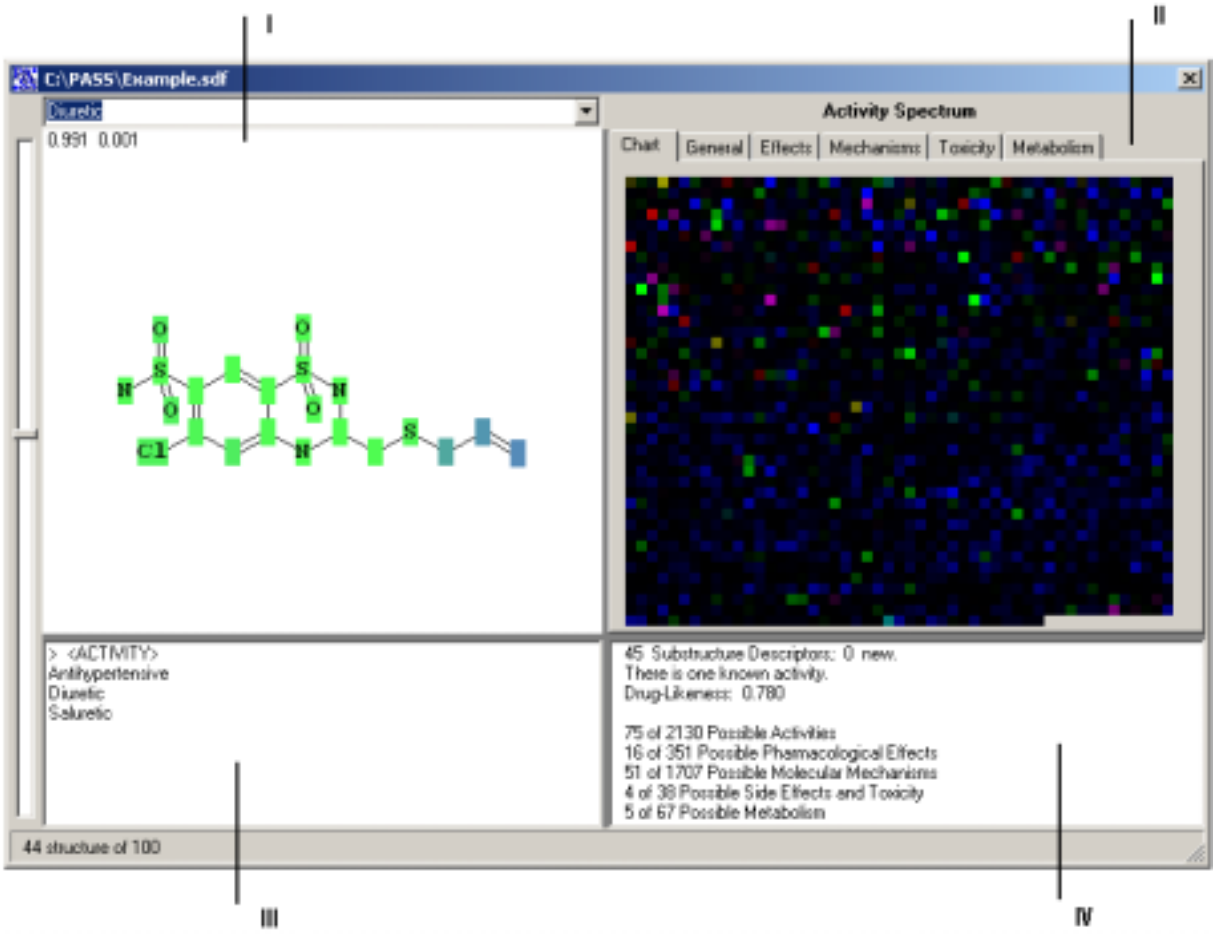
- If you want to save your prediction results and structures as SDFile use **Predict|Predict and Save structure and Prediction as SDfile…** menu command (or  button). In this case you save prediction results in an appropriate field and structure in "structure" field in addition to existing fields. After choosing this command you should open SD or MOL file in the **Open SDfile for Prediction** window, which appears at once. Then use the **Save SDfile with Prediction as** dialog box to type or change the name of the prediction result file. Press **Save** to start prediction.
- If you want to save your results as a CSV file use **Predict|Predict and Save Prediction as CSV file…** menu command (or  button). This format allows you to analyse prediction results by electronic spreadsheet (e.g., Microsoft Excel). In this case you should pay attention to some possible restrictions of such electronic spreadsheet. E.g., Microsoft Excel allows to analyse only 250 columns, so you should select no more then this number of activity types for prediction, using **Base/Selection…** menu item (or ). After choosing **Predict|Predict and Save Prediction as CSV file…** menu command you should open your SDfile in **Open SDfile for Prediction** window which appears at once. Then use the **Save CSV file with Prediction as** dialog box to type or change the name of the file of results. Press **Save** to start the prediction.

Mark "**Grouped Activity Spectra Saving"** (or press  button) to save the prediction results as SDfile in which activity names are divided into four groups: Effects, Mechanisms, Toxicity and Metabolism.

12. VIEWING PREDICTION RESULTS

To analyze prediction results **PASS** provides you with a special interactive tool. It allows you to view structure, text data and prediction results in several manners.

Use **Structure View** window to display structure(s) and the prediction results (**Activity Spectra**) for SD or Mol file which was opened by  button or **File|Open SDfile** menu command. In general you can use this option to look through any SD file.



The **Structure(s) File View** window is divided into four quadrants (I, II, III, IV).

I - a structural formula of the compound;

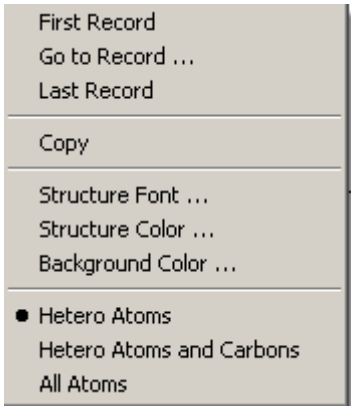
II - a predicted biological activity spectrum of the compound;

III – all additional (text) information presented in the appropriate SDfile for the current structure;

IV – general information about predicted results.

- The bottom status bar displays the current record number and the total number of records. (The main window status bar displays the size and the number of records in the file of the results).
- You can jump between records using left scroll bar or **Pg Up**, **Pg Down**, ↑,↓ .

I quadrant. Press the right mouse button on the structure frame or use the appropriate commands from **View** menu to navigate and change parameters of the structure representation.

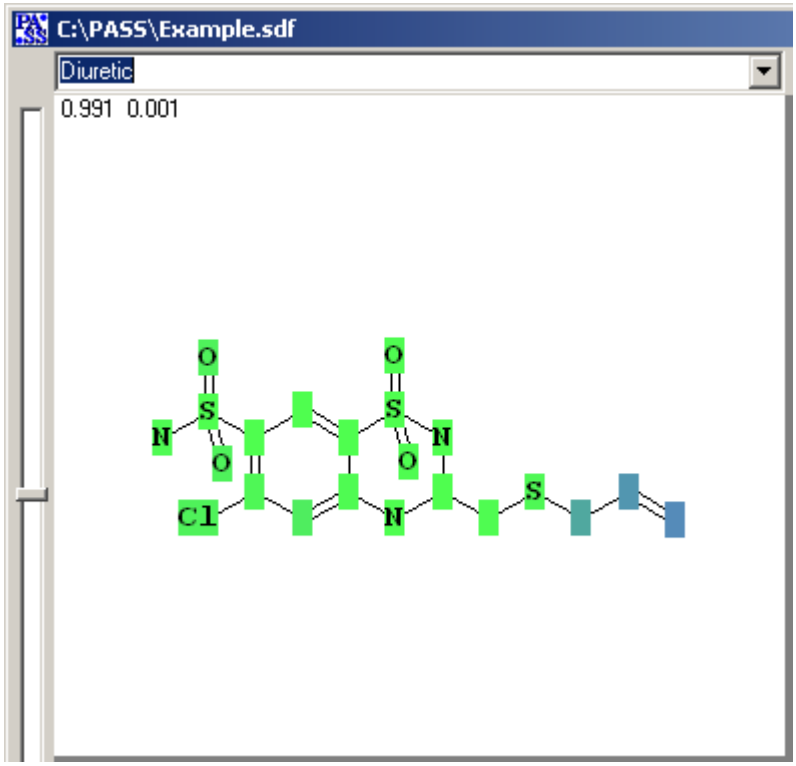


- ♦ Choose **First Record** command to go to the first record.
- ♦ Choose **Go to Record** command to go to the chosen record. The **Go to** window will appear. You should type the Record Number and press **OK** to go to this record.

- ♦ Choose **Last Record** command to go to the last record.
- ♦ Choose **Copy** command to copy the current structure to the clipboard. It is saved as a bitmap (.bmp file). This format is not used in ISIS Base.
- ♦ Choose **Structure Font …** command to change font of the atom symbols.
- ♦ Choose **Structure Color …** command to change color of the structure's bonds.
- ♦ Choose **Background Color …** command to change color of the background.

- Choose **Hetero Atoms** command to show only hetero atoms.
- Choose **Hetero Atoms and carbons** command to show hetero atoms and carbons.
- Choose **All Atoms** command to show all atoms of the compound.

PASS gives you an opportunity to evaluate the contribution of each atom of the structure to the estimation of biological activity. You should select a type of biological activity at the Activity Spectra Prediction or select it from the combo box, which is above the structure representation. Pa (0.991) and Pi (0.001) values for the selected activity are displayed over the structure.



The color of the atom depends from the contribution of the atom to the activity.

Red – Pa = 0, Pi=1;

Green – Pa=1, Pi=0;

Blue – Pa=0, Pi=0;

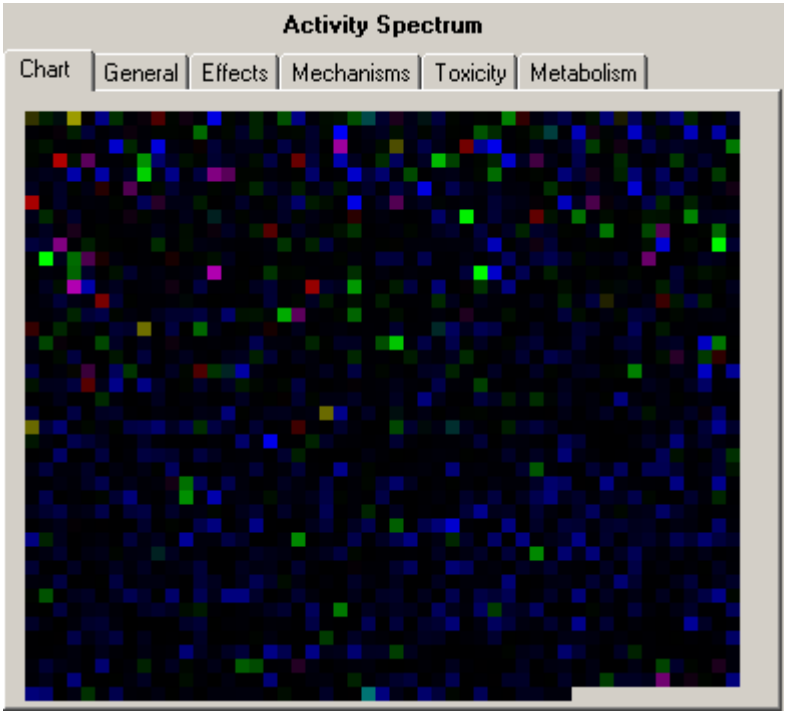
Grey – Pa=0.33, Pi=0.33.

Move the mouse cursor over the atom to display Pa and Pi values for the atom.

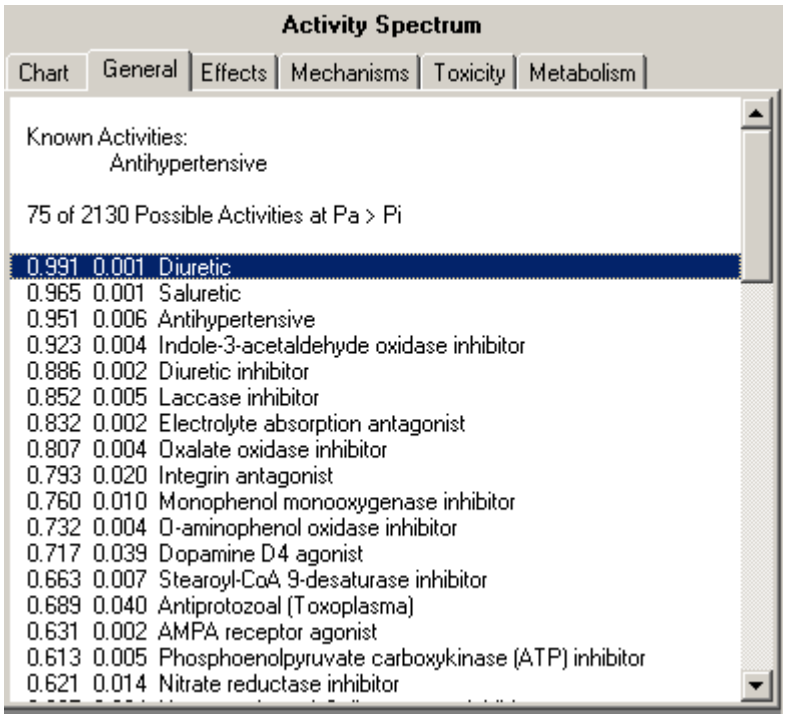
II quadrant

Biological Activity Spectrum Prediction is represented in different modes:

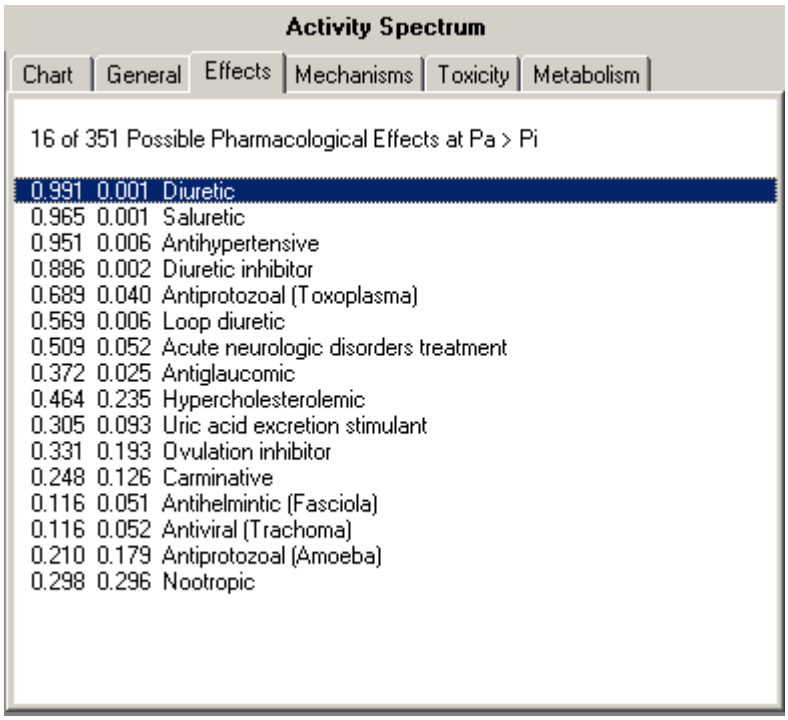
- **Chart** – the graphic representation of the biological activity spectra prediction, where each small square reflects a sertain biological activity type. Move the mouse cursor over an appropriate square to produce the hint with Pa and Pi values and the name of activity. The order of the squares corresponds with the order of activity types in the **Select Activity types to be Predicted** window. All activity types are divided into four classes: effects (■ - green squares), mechanisms (■ - blue squares), side and toxic effects (■ - red squares) and enzymes (■ - violet squares). There are combination of effects & mechanisms (■- turquoise squares) and combination of effects & toxic effects (■ - yellow squares). Brightness of the color corresponds with difference between Pa and Pi values for an appropriate activity.



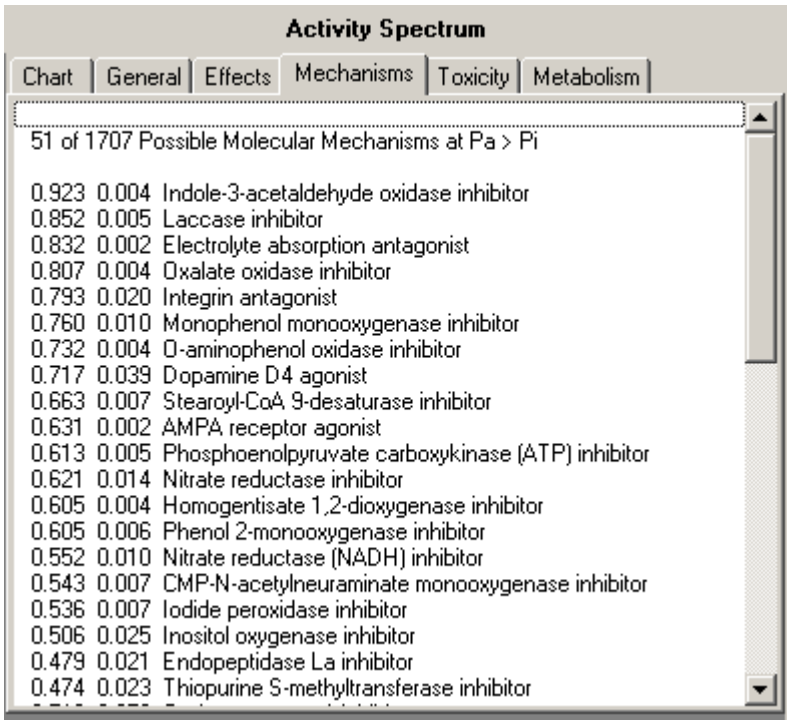
- **General** – all predicted biological activity types that meet the criteria specified by user. If a substance is equivalent to those from the SAR Base then the list of known activities appears for this substance in the top of appropriate window.



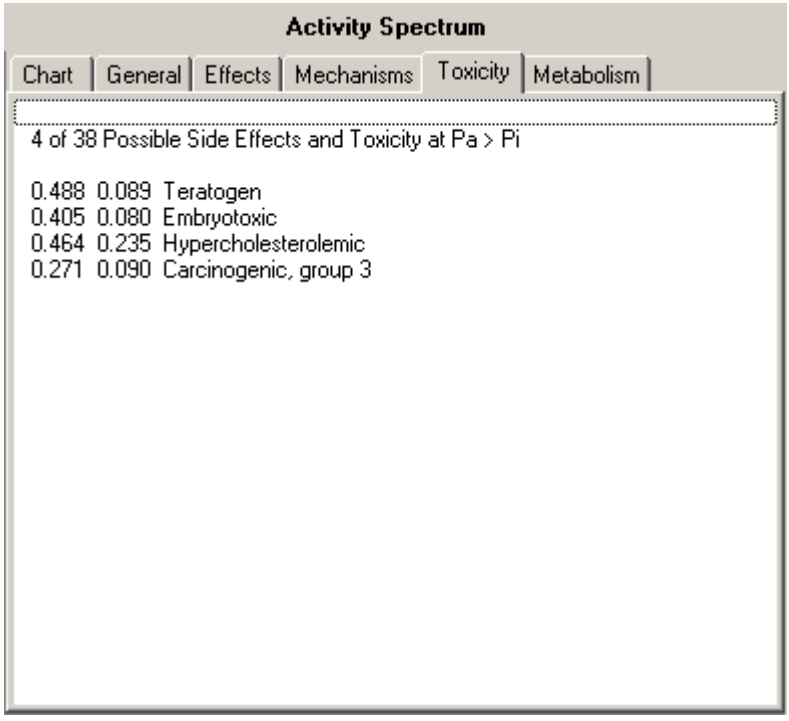
- **Effects** – all predicted pharmacotherapeutic effects that meet the criteria specified by user.



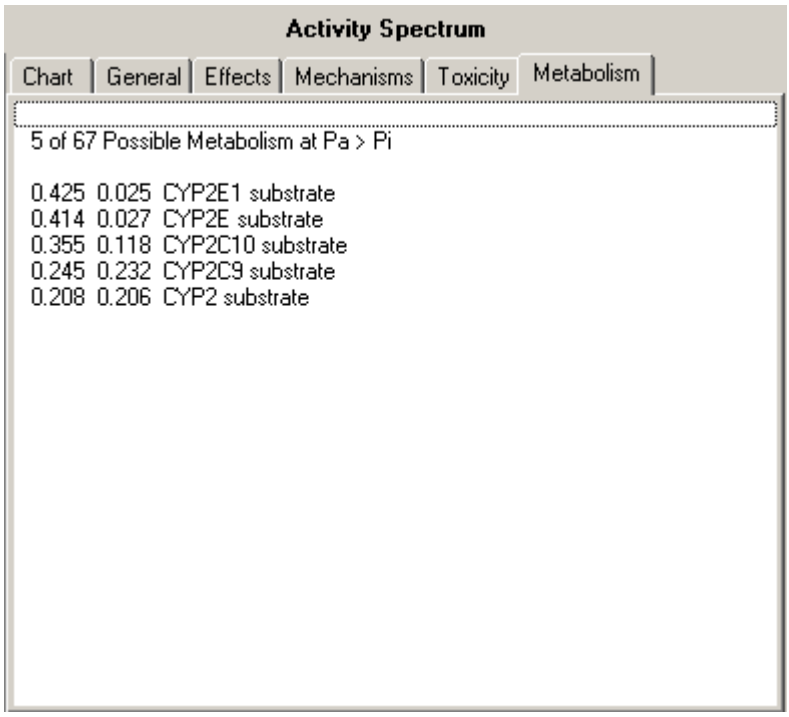
- **Mechanisms** – all predicted biochemical mechanisms of action that meet the criteria specified by user.



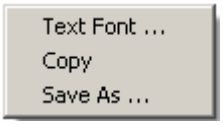
- **Toxicity** – all predicted adverse and toxic effects of action that meet the criteria specified by user.



- **Metabolism** - all predicted names of enzymes for which the substance is a substrate that meet the criteria specified by user.



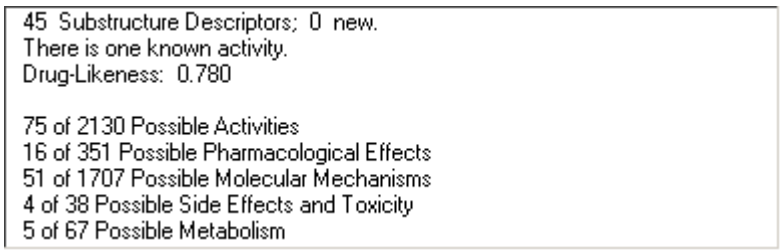
Press the right mouse button on the "**Activity Spectra**" frame to change the font of the **Structure View** window text, **copy** and **save** the prediction results for the current structure.



- ♦ Choose **Text Font …** command to change the font of the **Structure(s) and Prediction results View** window text.
- ♦ Choose **Copy** command to copy the content of the "Activity Spectrum" to the Windows' clipboard.
- ♦ Choose **Save as …** command to save the content of the current page (General, Effects, Mechanisms, Toxicity or Metabolism) in a TXT file.

III quadrant. All additional (text) information presented in the appropriate SDfile for the current structure:

IV quadrant displays the general information about predicted results:



First line represents the total number of descriptors (45) and the number of new descriptors (0) for a current structure (new descriptors are descriptors, which are not found in any substance from SAR base);

If a substance is equivalent to those from the SAR Base then the number of known activities for this substance appears (There is one known activity);

The drug-likeness value varies from 0 to 1. The more this value is the more drug-likeness of a substance is. The drug-likeness is 0.780 for the current structure.

Five last lines represent:

the number of all predicted activity types;

the number of predicted effects;

the number of predicted mechanisms;

the number of predicted toxic effects.

the number of predicted names of enzymes that may metabolise the current structure.

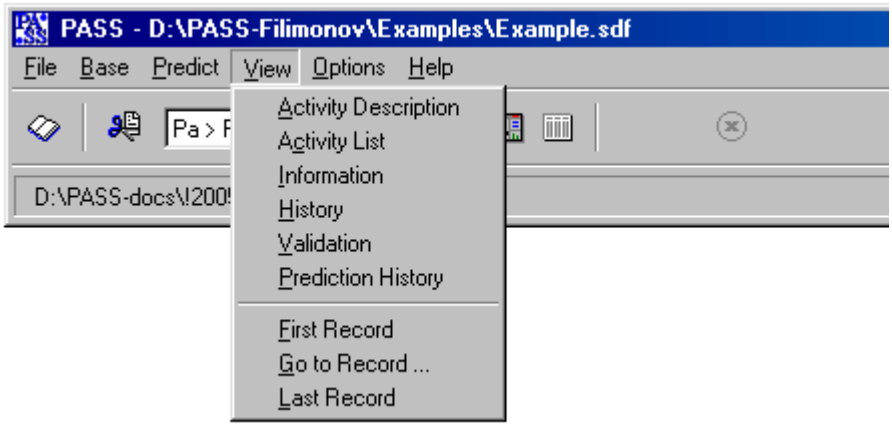
The string “75 of 2130 Possible Activities” means that from 2130 types of biological activity predicting by PASS only 75 activities met the criteria you specify for Cutting Point.

Attention!

The lists of effects, mechanisms, possible side effects and toxicity are saved in the appropriate files at PASS folder: effects.txt; mechanisms.txt and toxic.txt. The user could modify them with any text editor.

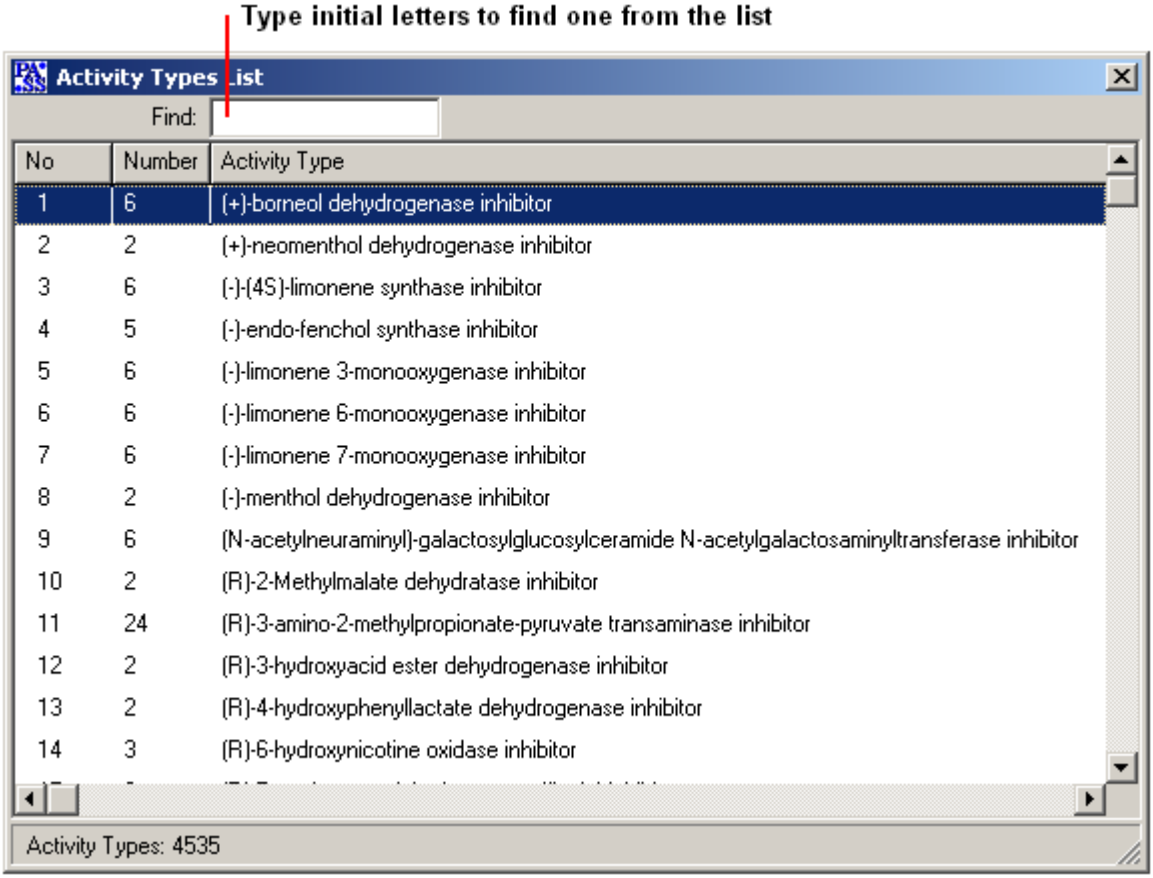
13. VIEWING BASIC INFORMATION

Use **View** menu commands to display the following information: the current status of the SAR Base, the history of creation and changes of SAR Base, the list of activities, which are predicted, the results of leave one out cross-validation procedure, the history of the latest prediction and prediction results. It contains the following items:



View menu options	Description
Activity Description	Displays the Activity Description window with basic information about an activity type.
Activity List	Displays the list of activity types.
Information	Displays the current status of SAR Base.
History	Displays all your previous steps.
Validation	Displays the results of leave one out cross-validation procedure.
Prediction History	Displays the history of the latest prediction.
First Record	Points out the first record in the **Structure(s) File View** window.
Go to Record...	Points out a chosen record in the **Structure(s) File View** window.
Last Record	Points out the last record in the **Structure(s) File View** window.

Use **Activity List** menu command to display the table of activity types presented in PASS training set.

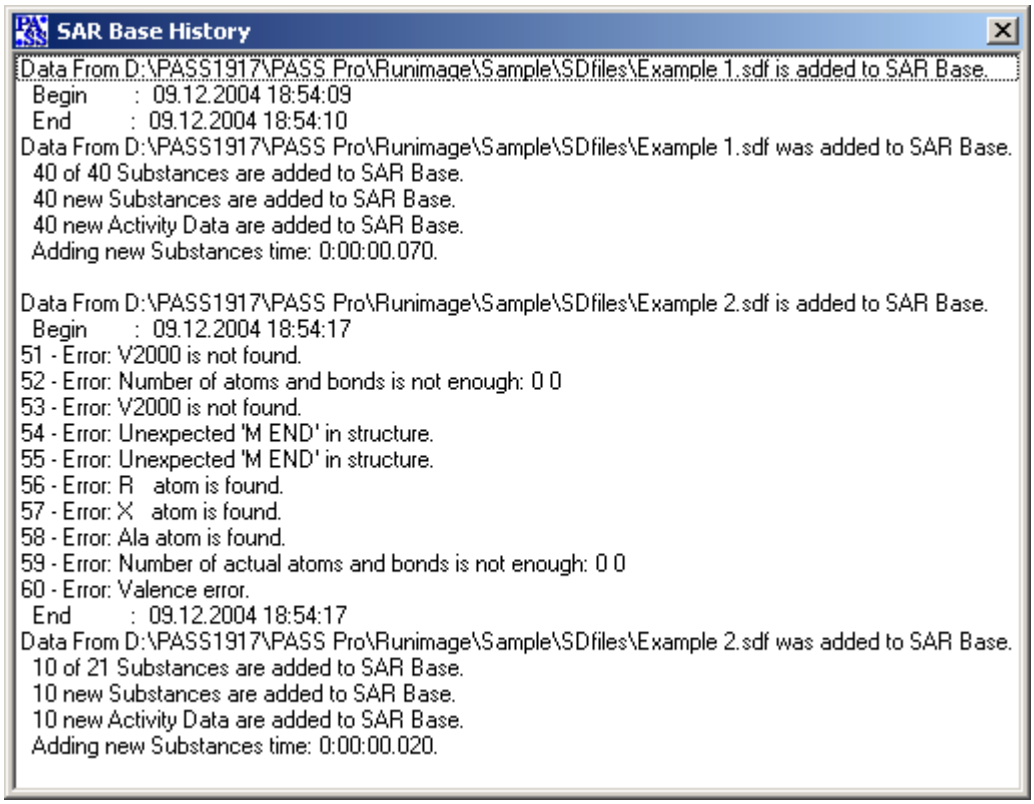


The **Number** column displays the number of compounds in the training set revealing this type of biological activity.

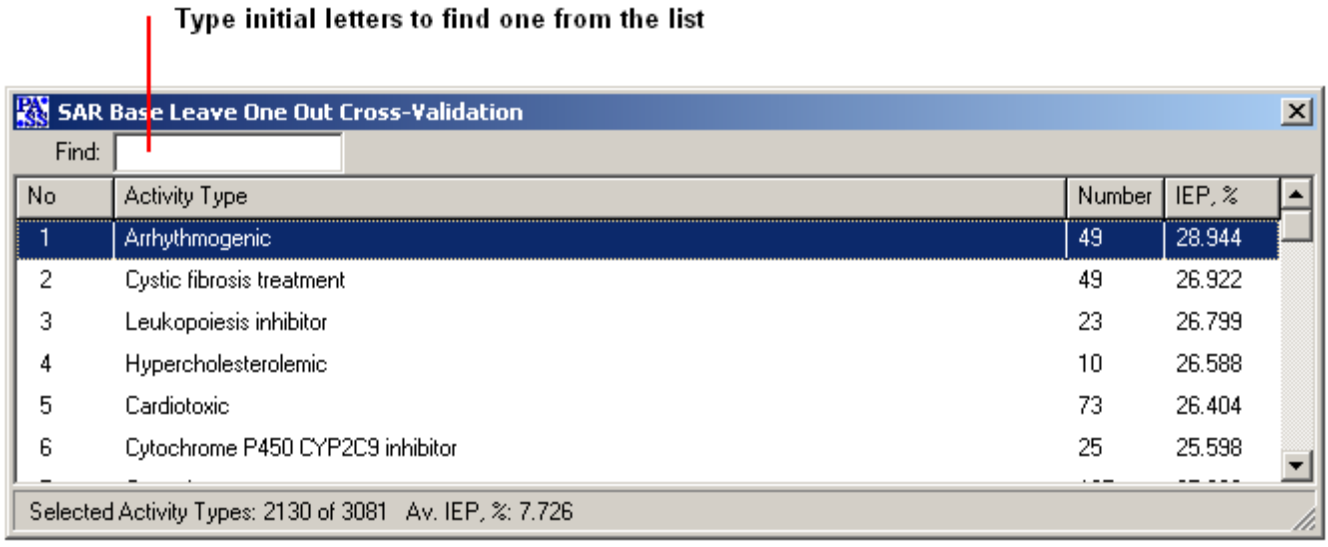
The bottom status bar displays the number of biological activity types in the SAR Base.

Use **View|Activity Description** command to open the **Activity Description** window with basic information about a chosen biological activity type.

Use **View|History** menu command to look at **SAR Base** Updating History. It contains: (1) names of input and output files; (2) the identifiers of the substances for which the equivalent structures are found; (3) errors in structures causing the failure of prediction.



Use **Validation** menu command to display the results of leave one out cross-validation procedure.

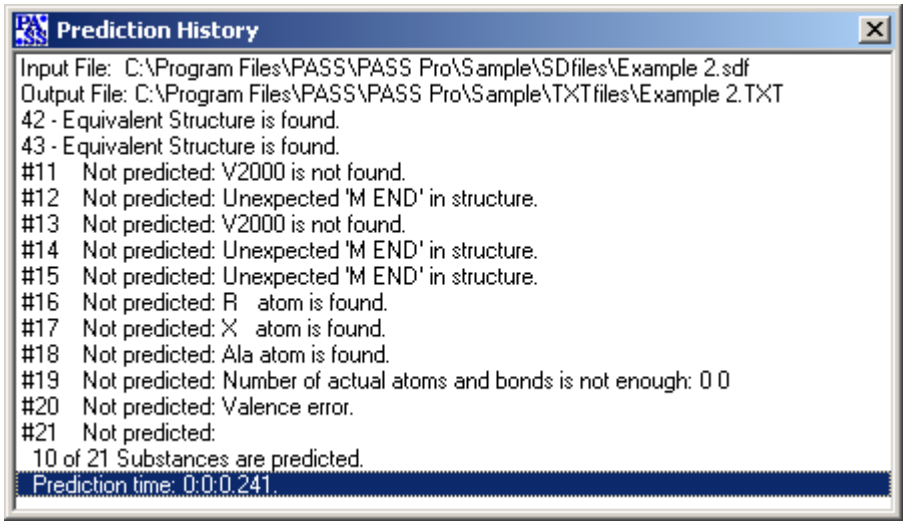


The **Number** column displays the number of compounds from the training set revealing this type of biological activity.

The **IEP** column displays the invariant error of prediction.

The bottom status bar displays the number of selected biological activity types from the SAR Base, which are used for biological activity spectra prediction and average IEP (Invariant Error of Prediction) for these activities.

Use **Prediction History** menu command to display the history of the latest prediction (the names of the input and output files).



14. INTERPRETING PREDICTION RESULTS

The result of prediction for the substance 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) or not revealed (Pi)**.** Their values vary from 0.000 to 1.000.

The more is Pa value, the less is the probability of false positives in the set of compounds selected for biological testing. For example, if one selects for testing only compounds for which a particular activity is predicted with Pa ≥ 90%, the expected probability to find inactive compounds in the selected set is very low, but about 90% of active compounds are missed. If only compounds with Pa ≥ 80% are chosen, the probability to find inactive compounds is also low, but about 80% of active compounds are missed; etc. By default, in PASS Pa=Pi value is chosen as a threshold, therefore all compounds with Pa>Pi are suggested to be active.

Another criteria for selection is the compounds' novelty. If Pa value is high, sometimes one may find close analogs of known biologically active substances among the tested compounds. For example, if Pa > 0.7 the chance to find the activity in experiment is high, but in some cases the compound may occur to be the close analogue of known pharmaceutical agents. If 0.5 < Pa < 0.7 the chance to find the activity in experiment is less, but the compound is not so similar to known pharmaceutical agents. If Pa < 0.5 the chance to find the activity in experiment is even more less, but if it will be confirmed the compound might occur to be a New Chemical Entity.

The quality of predictions is the main criterion of the program power. The mean accuracy of the prediction is about 88% (leave one out cross validation). There is an appropriate table where the maximum error of prediction for each type of activity is shown (see Supplement). In "Prediction Results" window a user also obtains the total sum of chemical descriptors of the substance. Reported are the number of descriptors, which are new compared with the PASS training set descriptors.

It should be mentioned that prediction of biological activity spectra is possible only for low molecular weight (drug-like) substances. Prediction of biological activity spectra for synthetic or biopolymers and inorganic substances will not provide reasonable results.

In some cases substance is predicted simultaneously as agonist and antagonist (stimulator and blocker, activator and inhibitor) for the same receptors (enzymes, etc.). It means that PASS could not make differentiation of intrinsic activity of substance and indicate only its affinity to this receptor (enzymes, etc.).

It is necessary to stress that PASS can’t predict if the concrete substance becomes drug, because it depends on many other factors. Prediction, however, can help to define what kind of tests are adequate for studying of biological activity of concrete chemical substance and which substances more probable will reveal the required effects.

Based on these criteria, you may choose which activities have to be tested in your compounds on the basis of compromise between the novelty of pharmacological action and the risk to obtain the negative result in experimental testing.

Certainly, you will also take into account your particular interest to some kinds of activity, experimental facilities, etc.

15. SUPPLEMENTS

15.1. Computer-aided methods in pharmaceutical R&D

Each pharmaceutical research and development (R & D) project is aimed at discovering new drug for the treatment of certain disease. The investigation of new pharmaceuticals is carried out in a stepwise manner. This is because drug discovery is a time consuming process involving enormous financial resources, manpower and substantially high risk factor. On an average it requires 12 years and approximately $800 million for introducing a new medicine to the market (*http://www.ifpma.org*) with a high risk of negative results (1 out of 10,000 substances studied is developed to a safe and potent drug). Drug research starts with identification of a "lead molecule" with required biological activity. Subsequently the lead molecule is developed to get more potent compounds with appropriate pharmacodynamic and pharmacokinetic properties that they qualify as drug candidates (*Wermuth C. The Practice of Medicinal Chemistry. New York: Academic Press, 2003*). General biological potential of any molecule under study is also evaluated in stages. The emphasis is first laid on testing for specific activity followed by general pharmacology & toxicology study, clinical trials, post-marketing registration of adverse effects etc. As a result, adverse/toxic actions are often discovered at a stage when a lot of time & money is already expended (*Poroikov VV, Filimonov DA. How to acquire new biological activities in old compounds by computer prediction. J Comput-Aided Mol. Des 2002, 16:819-824*). At the same time, it is practically impossible to test experimentally all compounds against each known kind of biological activity and possible toxic effects. So computer-aided prediction is "the method of choice" at the early stage of drug research. Relying on predicted results, one may establish the priorities for testing a particular compound and the basis for selecting the most prospective hits/leads/candidates from the set of compounds available for screening. Application of computational methods has significantly decreased the time required for obtaining a compound with the required properties with reduction in financial expenditures. In addition, it helps to obtain more effective and safety medicines.

Both computer-aided analysis of quantitative structure-activity/structure-property relationships (QSAR/QSPR) and molecular modeling are widely used for finding and optimizing lead compounds. However, the majority of such methods are constrained by studying a single targeted biological activity within the particular chemical *series (Holtje H.-D, Sippl W. Rational Approaches to Drug Design, Barcelona: Prous Science, 2001*.). Typically, they are applied step-by-step to analyze different activities/properties in

correspondence with the sequential study of biologically active compounds mentioned above. On the other hand, most of the known biologically active compounds demonstrate several or even many kinds of biological activity, which constitute the so-called “biological activity spectrum” of compound. Some components of biological activity spectrum may serve as a basis for the treatment of certain pathologies, while others may be a source for adverse/toxic effects. For instance, Thalidomide was prescribed worldwide (1950s to early 1960s) to pregnant women as a remedy for treatment of morning sickness. Subsequently it was discovered that Thalidomide was teratogenic (~12,000 babies were born with tiny or no limbs, flipper like arms and legs, with serious facial deformities and defective organs). Over this fact the drug was withdrawn from the market in 1962 (*Chu Y-H, Cheng CC. Affinity capillary electrophoresis in biomolecular recognition. Cell Mol Life Sci. 1998, 54:663–683*). However, now Thalidomide is again considered as a prospective pharmaceutical agent because of some newly discovered activities, e.g. angiogenesis inhibitor, tumor necrosis factor antagonist, and others (*Deplanque G, Harris AL. Anti-angiogenic agents: clinical trial design and therapies in development. European Journal of Cancer 2000, 36:1713-1724*). If, at the early stage of study, researchers could predict the most probable biological activities in drugs like Thalidomide, they might avoid the dramatic consequences of their adverse/toxic action and could suggest wider pharmacotherapeutic applications.

15.2. Biological Activity Presentation

In PASS biological activities are described qualitatively (active or inactive). Reflecting the result of chemical compound’s interaction with a biological object, the biological activity depends on both the compound’s molecular structure and the terms & conditions of the experiment. Therefore, structure-activity relationship analysis based on qualitative presentation of biological activity describes general “biological potential” of the molecule being studied. On the other hand, qualitative presentation allows integrating information concerning compounds tested under different terms and conditions and collected from many different sources as in the PASS training set.

Any property of chemical compounds, which is determined by their structural peculiarities, can be used for prediction by PASS. It is clear, that the applicability of PASS is broader than the prediction of biological activity spectra. To extend PASS application to other properties, the user needs the appropriate training set (see below).

15.3. Chemical Structure Description

The 2D structural formulae of compounds were chosen as the basis for description of chemical structure, because this is the only information available in the early stage of research (compounds may only be designed but not synthesized yet). Plenty of characteristics of chemical compounds can be calculated on the basis of structural formulae (*Guba W. Representation of Chemicals. In: Predictive toxicology. Ed. By Christoph Helma. Marcel Dekker, 2003*). Earlier (*Filimonov DA, Poroikov VV, Karaicheva EI, et al. Computer-Aided Prediction of Biological Activity Spectra of Chemical Substances on the Basis of Their Structural Formulae: Computerized System PASS. Experimental and Clinical Pharmacology (Rus). 1995, 58:56-62*) we applied the Substructure Superposition Fragment Notation (SSFN) codes (*Avidon VV, Pomerantsev IA, Rozenblit AB, Golender VE. Structure-activity relationship oriented language for chemical structure representation. J.Chem.Inform.Comput. Sci., 1982, 22:207-214*). But SSFN, like many other structural descriptors reflects rather abstraction of chemical structure by the human mind than the nature of the biological activity revealed by chemicals. The Multilevel Neighborhoods of Atoms (MNA) *descriptors (Lagunin A, Stepanchikova A, Filimonov D, Poroikov V. PASS: prediction of activity spectra for biologically active substances. Bioinformatics, 2000, 16:747-748*) have certain advantages in comparison with SSFN. These descriptors are based on the molecular structure representation, which includes the hydrogens according to the valences and partial charges of other atoms and does not specify the types of bonds. MNA descriptors are generated as recursively defined sequence:

- zero-level MNA descriptor for each atom is the mark ***A*** of the atom itself;
- any next-level MNA descriptor for the atom is the sub-structure notation

$A(D_1D_2..D_i\ldots)$,

where D_i is the previous-level MNA descriptor for ***i***–th immediate neighbor's of the atom ***A***.

The mark of atom may include not only the atomic type but also any additional information about the atom. In particular, if the atom is not included into the ring, it is marked by "-". The neighbor descriptors $D_1D_2...D_i\ldots$ are arranged in unique manner, e.g., in lexicographic order. Iterative process of MNA descriptors generation can be continued covering first, second, etc. neighbourhoods of each atom.

This process can be continued iteratively covering 2nd, 3rd, etc. neighbourhoods of the atom. We use 2nd level descriptors in the present version of PASS.

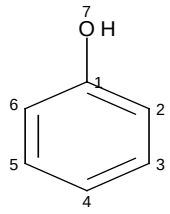
Multilevel Neighborhoods of Atoms (MNA) structure descriptors of a molecule are generated on the basis of connection table (C) and table of atoms types (A) presented the substance.

Connection table contains data on the valent bonds in a molecule. Various bond types are not specified (topological approximation). All hydrogens based on valencies and partial charges of atoms are taken into account. The types of atoms are specified according to the data presented in Table 1.

Table 1. Classification of different atom types used in calculation of descriptors

Class name	Elements
H	H
C	C
N	N
O	O
F	F
Si	Si
P	P
S	S
Cl	Cl
Ca	Ca
As	As
Se	Se
Br	Br
Li*	Li, Na
B*	B, Re
Mg*	Mg, Mn
Sn*	Sn, Pb
Te*	Te, Po
I*	I, At
Os*	Os, Ir
Sc*	Sc, Ti, Zr
Fe*	Fe, Hf, Ta
Co*	Co, Sb, W
Sr*	Sr, Ba, Ra
Pd*	Pd, Pt, Au
Be*	Be, Zn, Cd, Hg
K*	K, Rb, Cs, Fr
V*	V, Cr, Nb, Mo, Tc
Ni*	Ni, Cu, Ge, Ru, Rh, Ag, Bi
In*	In, La, Ce, Pr, Nd, Pm, Sm, Eu
Al*	Al, Ga, Y, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Tl
R*	R, He, Ne, Ar, Kr, Xe, Rn, Ac, Th, Pa, U, Np, Pu, Am, Cm, Bk, Cf, Es, Fm, Md, No, Lr, Db, Jl

Example of structure presentation by zero-, first- and second-levels MNA descriptors for the phenol's molecule is shown in Figure below.



Atom	MNA/0	MNA/1	MNA/2
1	C	C(CC-O)	C(C(CC-H)C(CC-H)-O(C-H))
2	C	C(CC-H)	C(C(CC-H)C(CC-O)-H(C))
3	C	C(CC-H)	C(C(CC-H)C(CC-H)-H(C))
4	C	C(CC-H)	C(C(CC-H)C(CC-H)-H(C))
5	C	C(CC-H)	C(C(CC-H)C(CC-H)-H(C))
6	C	C(CC-H)	C(C(CC-H)C(CC-O)-H(C))
7	-O	-O(C-H)	-O(C(CC-O)-H(-O))
8	-H	-H(C)	-H(C(CC-H))
9	-H	-H(C)	-H(C(CC-H))
10	-H	-H(C)	-H(C(CC-H))
11	-H	-H(C)	-H(C(CC-H))
12	-H	-H(C)	-H(C(CC-H))
13	-H	-H(-O)	-H(-O(C-H))

MNA descriptors for phenol.
MNA/0, MNA/1,MNA/2 - zero, first and second levels of MNA descriptors.

It is shown that usage of 1st & 2nd levels MNA descriptors provides the best accuracy of prediction. MNA descriptors are generated for each substance. Unique integer identifier is assigned to each particular descriptor according to the descriptors' dictionary.

The substances are considered to be equivalent in PASS if they have the same set of MNA descriptors. Since MNA descriptors do not represent the stereochemical peculiarities of a molecule, the substances whose structures differ only stereochemically, are formally considered as equivalent.

15.4. Training Set

The PASS estimations of biological activity spectra of new compounds are based on the Structure-Activity Relationships knowledge-base (SARBase), which accumulates the results of the training set analysis. The in-house developed PASS training set includes more than 60,000 known biologically active substances (drugs, drug-candidates, leads, and toxic compounds). Since new information about biologically active compounds is discovered regularly, we perform the special informational search and analyse the new information, which is further used for updating and correcting the PASS training set.

MNA descriptors $\{D_1, \ldots, D_m\}$ for each kind of activity A_k the following B_k values are calculated:

$$B_k = (S_k - S_{0k})/(1 - S_k \bullet S_{0k}),$$
$$S_k = Sin[\sum_i ArcSin(2P(A_k|D_i) - 1)/m],$$
$$S_{0k} = 2P(A_k) - 1,$$

where $P(A_k|D_i)$ is a conditional probability of activity of kind A_k if the descriptor D_i is present in a set of molecule's descriptors; $P(A_k)$ is a priori probability to find a compound with activity of kind A_k. For any kind of activity A_k, if $P(A_k|D_i)$ is equal to 1 for all descriptors of a molecule, then $B_k = 1$; if $P(A_k|D_i)$ is equal to 0 for all descriptors of a molecule, then $B_k = -1$; if there is no relationship between the molecule's descriptors and activity of kind A_k, and, so, $P(A_k|D_i) \approx P(A_k)$, then $B_k \approx 0$.

Up to the PASS version 1.703 the algorithm of prediction was based on the following data:
n is the total number of compounds in the SARBase;
n_i is the number of compounds containing descriptor D_i in the structure description;
n_k is the number of compounds containing the kind of activity A_k in the activity spectrum;
n_{ik} is the number of compounds containing both the kind of activity A_k and the descriptor D_i.
And the estimations of probabilities $P(A_k)$, $P(A_k|D_i)$ are given by:

$$P(A_k) = n_k/n, \quad P(A_k|D_i) = n_{ik}/n_i.$$

In PASS version 1.703 and later instead of integers n_i and n_{ik} the sums g_i and g_{ik} of descriptors weights w are used, where $w = 1/m$, and m is the number of MNA descriptors of individual molecule. This modification increases the accuracy of prediction significantly. So, right now the estimations of probabilities $P(A_k|D_i)$ are given by:

$$P(A_k|D_i) = g_{ik}/g_i.$$

The main purpose of PASS application is to predict the activity spectra for new substances. To provide more accurate predictions, if the compound under prediction has the equivalent structure in the SARBase, this structure is “excluded” from the SARBase during the prediction with all associated information about its biological activities. The calculations are done by using $n-1$, g_i-w, and, when the kind of activity A_k is contained in its activity spectrum in the SARBase, by using n_k-1 and $g_{ik}-w$. Here $w = 1/m$, and m is a number of MNA descriptors in molecule under prediction and its equivalent in the SARBase. The B_k values are calculated using MNA descriptors, which are found in SARBase, i.e., for descriptors of a molecule under prediction with $g_i > 0$ or $g_i-w > 0$, in the case of structure “exclusion”.

To take the "yes/no" qualitative prediction it is necessary to determine ***B***-statistics threshold values for each kind of activity A_k. Using theory of statistical decision this can be done on the basis of risk function's minimization. But nobody can a priori specify the risk functions for all activity kinds and all possible practical tasks. Therefore, the predicted activity spectrum in PASS is presented by the rank-order list of activities with probabilities "to be active" ***Pa*** and "to be inactive" ***Pi***, which are the functions of ***B***-statistics for a molecule under prediction. The ***B***-statistics functions ***Pa*** and ***Pi*** are the results of the training procedure described below. The list is arranged in descending order of ***Pa–Pi***; thus, the more probable activity kinds are at the top of the list. The list can be shortened at any desirable cutoff value, but ***Pa>Pi*** is used by default. If the user chooses rather higher value of ***Pa*** as a cutoff for selection of probable activities, the chance to confirm the predicted activities by the experiment is also high, but many existing activities will be lost. For instance, if ***Pa***>80% is used as a threshold, about 80% of real activities will be lost; for ***Pa***>70%, the portion of lost activities is 70%, etc.

15.5. Training procedure

For each compound from the training set MNA descriptors are generated and its known activity spectrum and set of descriptors are stored in the SARBase. If this compound has the equivalent structure in SARBase, only new activities are added to activity spectrum. After inclusion of all information from the training set(s) into SARBase the values n, g_i, n_k, g_{ik} are calculated. For each compound in the SARBase and for each activity kind A_k, values B_k of ***B***-statistics are calculated. Calculations are done taking into account the described above "exclusion" of processed compound. For each activity kind A_k, the calculated values B_k are subdivided into two samples: for active and inactive compounds. These obtained samples are used for calculation the smooth estimations of ***B***-statistics distribution functions on the following basis.

Suppose the sample $x_1, ..., x_n$ of n values of random variable X, which has an unknown distribution function $F(x)$. Using of an empirical step-function for approximation of F often faults because of small n. To provide the smooth estimation of $F(x)$, the inverse function $x(F)$ is calculated as the conditional expectation of random variable X:

$$x(F) = \sum_i (n-1)!\cdot F^{i-1}/(i-1)!\cdot(1-F)^{n-i}/(n-i)!\cdot x'_i,$$

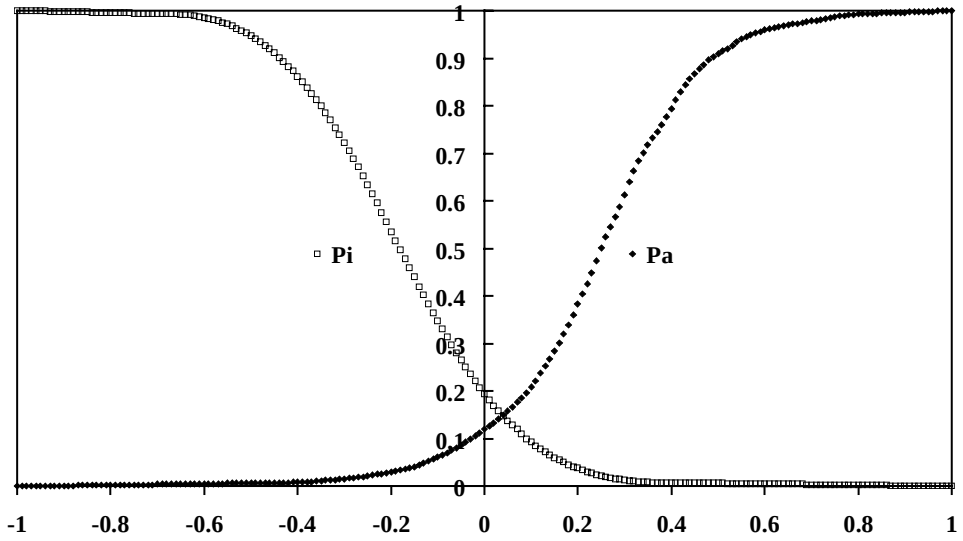
where $(n-1)!\cdot F^{i-1}/(i-1)!\cdot(1-F)^{n-i}/(n-i)!$ is the binomial distribution, and $x'_1, ..., x'_n$ ($x'_1 < x'_2 < ... < x'_n$) is the ranked sample $x_1, ..., x_n$. The distribution function $F(x)$ is given reciprocal function of quantiles $x(F)$.

Each sample of B values for active compounds is arranged in the ascending order; each sample of B values for inactive compounds is arranged in descending order. The described above quantiles $b(F)$ are calculated. As a result, for each appropriate kind of activity the probabilities Pa and Pi are given by:

$$b_{active}(Pa) = B, \qquad b_{inactive}(Pi) = B.$$

By definition the probabilities Pa and Pi are also the probabilities of the 1st and 2nd kinds of prediction error at the threshold B, respectively. They can be also interpreted as the measures of belonging to fuzzy subsets of "active" and "inactive" compounds. Both interpretations of probabilities Pa and Pi are equivalent and can be used for interpreting the results of prediction. They can also be used for construction of different criteria for prediction results' analysis corresponded to specific practical problems.

The example of the probabilities $Pa(B)$ and $Pi(B)$ for activity «Alpha adrenoreceptor antagonist» as functions of B-statistics value is shown in the figure.



PASS prediction accuracy is estimated using average ***IEP*** for all predictable activity kinds.
IEP is calculated for each type of activity in PASS prediction:

$$IEP = \#(B_0 > B_1)/(N_o N_1),$$

where $\mathbf{B_0}$ and $\mathbf{B_1}$ are values of ***B***-statistics for some pair of inactive and active compounds in the training set,

$\mathbf{N_o}$ is the number of inactive compounds in the training set,

$\mathbf{N_1}$ is the number of active compounds in the training set.

If the values of ***B***-statistics for any active compounds are higher than the values of ***B***-statistics for all inactive compounds, then **IEP=0**. It means that all "active" and "inactive" compounds for the current type of activity from the training set were devided absolutely correct during LOO CV procedure. If the values of ***B***-statistics for any active compounds are the same as the values of ***B***-statistics for all inactive compounds, then **IEP=0.5**. This means that prediction is not correct.

15.6. PASS Validation

Leave one out cross-validation for all ~4500 kinds of biological activity and ~60,000 substances provides the estimate of PASS prediction accuracy at the training procedure. Average accuracy of prediction is about 92% according to the LOO CV estimation, while that for particular kinds of activity varies from 77.7 % (antisecretoric) to 99.5 % (Melanocortin antagonist). Accuracy of prediction data for all kinds of biological activity predicted by PASS is presented at the web site (*http://www.ibmc.msk.ru/PASS)*.

The accuracy of PASS predictions depends on several factors, from which the quality of the training set seems to be the most important one. A perfect training set should include the comprehensive information about biological activities known or possible for each compound. In other words, the whole *biological activity spectrum* should be thoroughly investigated for each compound included into the PASS training set. Actually, no database exists with information about biologically active compounds tested against each kind of biological activity. Therefore, the information concerning known biological activities for any compound is always incomplete. We investigated the influence of the information's incompleteness on the prediction accuracy for new compounds. About 20,000 principal compounds from MDDR database (*http://www.mdl.com)* were used to create the heterogeneous training and evaluation

sets. At random 20, 40, 60, 80% of information were excluded from the training set. Either structural data or biological activity data were removed in two separate computer experiments. In both cases it was shown that even if up to 60% of information is excluded, the results of prediction are still satisfactory (*Poroikov VV, Filimonov DA, Borodina YuV, Lagunin AA, Kos A. Robustness of biological activity spectra predicting by computer program PASS for non-congeneric sets of chemical compounds. J.Chem.Inform.Comput.Sci., 2000, 40:1349-1355).* Thus, despite the incompleteness of information in the training set, the method used in PASS is robust enough to provide the reasonable results of prediction.

Incompleteness of data on biologically active compounds significantly restricts the possibilities for evaluation of PASS prediction abilities, because many biological activities, which are probable according to the predictions, were never tested. Fortunately, there exists the NCI database with 42,689 heterogeneous compounds each being tested in anti-HIV assays (*Voigt JH, Bienfait B, Wang S, Nicklaus MC. Comparison of the NCI Open Database with Seven Large Chemical Structural Databases. J.Chem.Inf.Comput.Sci. 2001, 41:702-712).* We used this database to estimate to which extent the application of PASS predictions enriches the population of active compounds in the subsets selected from the sample. It was shown in this experiment that depending on the cutoff value of ***Pa***, the enrichment of “actives” varies from 2.3 to 16.7 (*Poroikov VV, Filimonov DA, Ihlenfeldt W-D, Gloriozova TA, Lagunin AA, Borodina YuV, Stepanchikova AV, Nicklaus MC. PASS biological activity spectrum predictions in the enhanced open NCI database browser. J.Chem.Inform.Comput.Sci. 2003, 43:228-236).* Thus, PASS predictions significantly increase the probability of finding particular activity in compounds under study.