

BIOINFOQSAR – A SPECIALIZED SOFTWARE DEVELOPED FOR QSAR MODELS. PREDICTABLE CAPACITY TESTING FOR CONVENTIONAL ANTITUMOR DRUGS

LĂCRĂMIOARA POPA¹, DOINA DRĂGĂNESCU^{1*}, MĂDĂLINA GEORGIANA ALBU², ALINA ORTAN³, MIHAELA VIOLETA GHICA¹

¹“Carol Davila” University of Medicine and Pharmacy, Faculty of Pharmacy, Physical and Colloidal Chemistry Department, 6, Traian Vuia, 020956, Bucharest, Romania

² INCDTP – Division Leather and Footwear Research Institute, Collagen Department, 93 Ion Minulescu Str., 031215, Bucharest, Romania

³University of Agronomic Sciences and Veterinary Medicine, Faculty of Land Reclamation and Environment Engineering, Bucharest, Romania

*corresponding author: doinadraganescu@yahoo.com

Abstract

A new software for QSAR design and analysis was developed and is presented in this work. BIOINFOQSAR was designed for developing high predictable and robust models of correlation between molecular structure descriptors and biological properties. This study includes the presentation of the application interface, its modules and functionality. An algorithm for QSAR model generation and QSAR analysis as a structured program are disclosed.

The test of BIOINFOQSAR boosting capacity in high level prediction was conducted in a group of conventional antitumor agents. A model of correlation as a linear function between the biological property – IC₅₀ inhibitory concentration, and a sum of molecular descriptors was chosen, and various statistical tests were conducted. The highest level of predictability was obtained, even the number of terms in the equation was quite elevated. The objective of predictability level test for the BIOINFOQSAR was fulfill.

Rezumat

În această lucrare este prezentat un nou software de proiectare și analiză QSAR. Aplicația BIOINFOQSAR a fost realizată pentru obținerea unor modele robuste și cu o înaltă capacitate de predicție, de corelare între descriptori structural-moleculari și proprietăți biologice. Acest studiu include prezentarea interfeței aplicației, modulele și funcționalitatea sa. Se prezintă algoritmul de generare al modelelor QSAR și cel de analiză QSAR, ca program structurat.

Testarea aplicației BIOINFOQSAR privind capacitatea sa predictivă a fost realizată într-un grup de agenți antitumoralii convenționali. S-a ales un model QSAR de corelare, ca funcție liniară dintre proprietatea biologică IC₅₀ - concentrația inhibitoare și o sumă de descriptori moleculari, iar acest model a fost analizat din punct de vedere statistic. S-a obținut un înalt nivel de predictibilitate, chiar dacă numărul de termeni ai ecuației

QSAR a fost destul de mare. Obiectivul studiului privind testarea capacității de predicție pentru aplicația BIOINFOQSAR fost îndeplinit.

Keywords: BIOINFOQSAR, capacity testing, predictive models, antitumor drugs.

Introduction

All the techniques of predictive QSAR (quantitative relationships between structure and activity) are based on the assumption that the structure of molecules (geometrical, steric, charge and electron characteristics) contains responsible features for its physical, chemical and biological properties; also, the QSAR models are relied upon the ability that the chemical structure may be represented by one or more numeric descriptors [1-3]. Through QSAR models, biological activity (or property, reactivity, etc.) of a new chemical compound, previously not tested, can be deduced from the molecular structure of similar compounds whose activity has already been estimated [4,5]. When a property is modeled, the acronym QSPR (quantitative link between structure and property) is preferred [6-8].

There are more than 40 years since QSAR modeling was first used in drug design practice in toxicology, industrial and environmental chemistry [6,9]. Nowadays, the power of these techniques increased due to the rapid development of computational methodologies and techniques that allowed precise description and refinement of several variables used in this type of modeling [10-12]. Currently QSAR science is founded on the systematic use of mathematical models and is one of the basic tools in modern design of drugs, having an increasingly important role in environmental sciences [13-14].

QSAR models lies at the confluence of chemistry, statistics and biology in toxicological studies. Development of a QSAR model requires three components: (1) data set that provides experimental determination of the biological activity for a group (class) of chemical structures; (2) data about molecular structure and/or property (descriptors, variables, predictors) for each compound in this group; (3) statistical methods to determine the link between these two sets of data [15-17]. Limitations encountered in the development of reliable QSAR models are related to high-quality experimental data. In QSAR analysis, two aspects are imperative: both sets of input data have to be accurate, and the developed model should be relevant. The data used in QSAR models generation are obtained either from the literature or specifically obtained for QSAR analysis [18,19].

There are many software programs known for their ability to describe chemical structures, to calculate molecular descriptors, to generate

specifics algorithms, to generate correlation functions structure-activity and finally, to provide reliable, robust and valuable models of prediction [11,12].

Our study was focused on the new and specialized QSAR software named BIOINFOQSAR. It was developed as a methodology of QSAR studies with a specific algorithm for structure-activity relationships generation and analysis. For the BIOINFOQSAR a test was designed and conducted for its predicting capacity, in the class of conventional antitumor drugs. It had especially followed boosting capacity of predicting for this application. BIOINFOQSAR was designed for much more areas of application, not only QSAR studies.

Materials and Methods

The BIOINFOQSAR software was designed with a modular architecture, for proper management and implementation activities, and most of all for an adequate flexibility when applied to much more areas, not only QSAR models.

Description of data structures

The application works with several data structures, each with a specific role defined for application functionality. The main data structures are those with which the user interacts and can be divided into two categories:

Data structures used as input into the application: (a) molecular compound; (b) molecular descriptors; (c) properties of molecular compounds; (d) analysis of QSAR/QSPR models.

Data structures for the analysis resulted: models of correlation. Also, the application makes use of two other structures with administrative role in the application: work session, and case study.

The interface of the software application

The first version of the BIOINFOQSAR software runs under Windows operating systems 98/2000/XP/Vista. The user has one simple and well structured interface to provide all the information required for analysis and to obtain useful information during processing. The interface uses standard graphics (main application menu, dialog boxes, edit boxes, items for selection) and regular operating system specific events.

The main menu provides access to the main controls, grouped into three categories: (1) session –initiates a new work session, saves, opens previous sessions, closes session/application; (2) analysis - launches the process of correlations, orders the report; (3) help - operations support for the user (user manual, tutorials, application information), as illustrated in figure 1.

In the zone of work session management, there is the "Update Case Study." This command allows data updating for the current case study, based

on selections from the database. For each category of data, the management organization in the session could also be a copy from various databases.

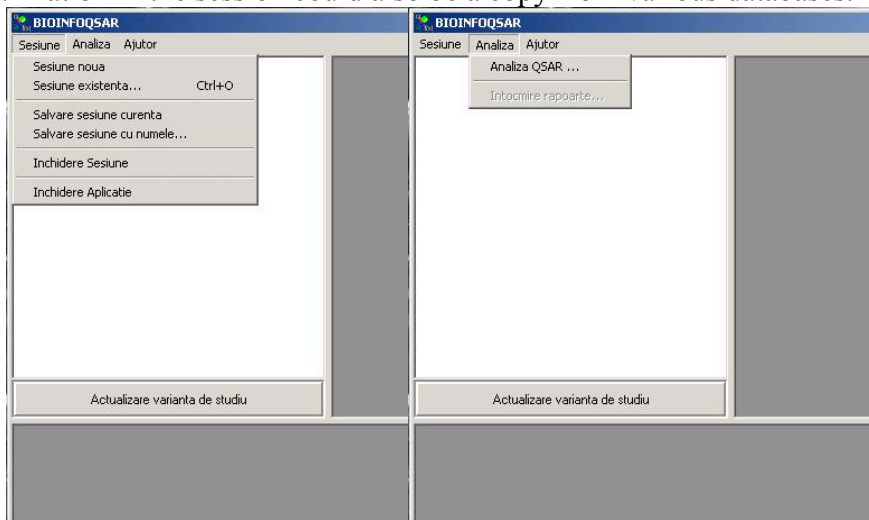


Figure 1

One first view of the BIOINFOQSAR interface

In this first version of the application, editing molecular modules are not implemented. The molecular structures of the compounds are imported from other sources, in a two- or three-dimensional view (Figures 2 and 3).

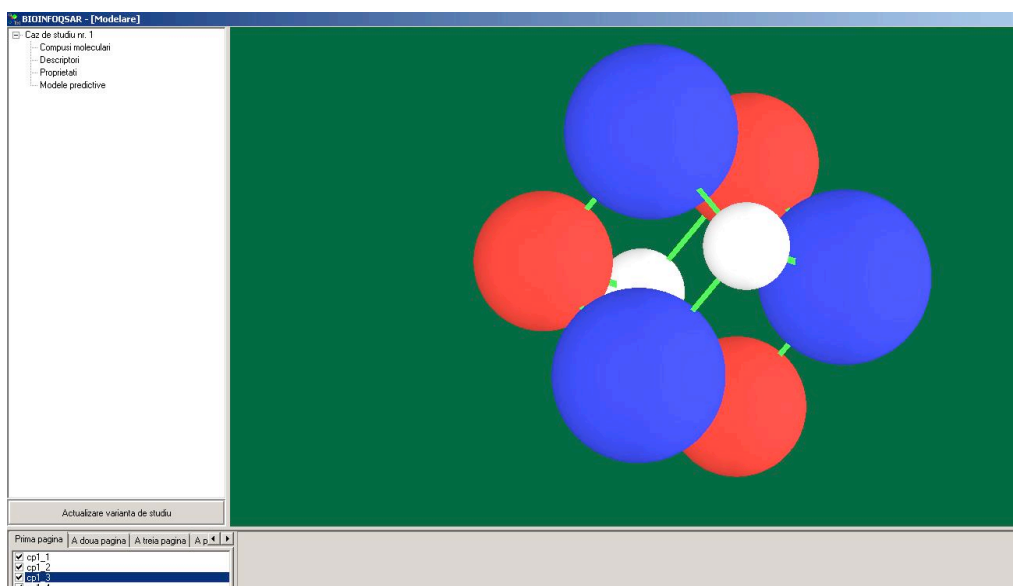


Figure 2

Typical three-dimensional view for a case of study

	Proprietate 1	Descriptor 2	Descriptor 3	Descriptor 4	Descriptor 5	Descriptor 6	Descriptor 7	Descriptor 8	Descriptor 9	Descriptor 10	Descriptor 11
Compus 1	0.098799	0.031300	0.961048	0.202581	0.272921	0.671654	0.318691	0.161795	0.372238	0.425674	0.082012
Compus 2	0.059724	0.293297	0.917285	0.367906	0.774665	0.327936	0.697675	0.844171	0.717983	0.308641	0.162626
Compus 3	0.825676	0.279029	0.491766	0.149185	0.874335	0.287294	0.772793	0.976460	0.492528	0.887936	0.827277
Compus 4	0.500666	0.021676	0.532335	0.009649	0.774478	0.656668	0.770489	0.709107	0.597908	0.206030	0.681086
Compus 5	0.998371	0.243790	0.676166	0.295695	0.085890	0.772322	0.495278	0.881694	0.509866	0.573323	0.953193
Compus 6	0.705351	0.988937	0.774045	0.749573	0.197707	0.143647	0.913699	0.794196	0.589281	0.864709	0.689371
Compus 7	0.280025	0.877185	0.161234	0.273188	0.543685	0.960792	0.174364	0.157070	0.268121	0.177457	0.576795

Figure 3
Calculation of descriptors for a case of study

A spreadsheet for a case of study is presented in figure 3, having the following structure: (1) a set of 7 molecular compounds; (2) a set of 14 molecular descriptors (only 11 visible here); (3) a property.

The algorithm for QSAR models generation and QSAR analysis of in congeneric series of standard therapeutic agents:

Identification of test compounds set (training set)

The first issue is the choice of the molecular structures used as testing set, by means of various molecular editing tools capable of building or import chemical structures (specific files, different formats);

Introduction of biological activity data

For each set of test molecules (training set) the values of certain biological properties specific to a certain therapeutic activity will be introduced (for example IC₅₀ - inhibitory concentration, GI₅₀ inhibitory concentration for human tumor cells, logP - partition coefficient, logD - distribution coefficient, etc.).

Generating conformations

In order to conduct a 3D analysis, conformational information is needed on chemical structures, followed by the development of specific methods for conformational generating;

Calculation of the molecular descriptors

Through the use of specific modules of the BIOINFOQSAR, a variety of molecular descriptors can be estimated: spatial, electronic,

topological, the atomic-content, thermodynamic, conformational, quantum mechanics, on molecular shape, etc.

Exploring Data

Charts and graphs can be generated, which can achieve an assessment of the distribution descriptors. It is also possible to visualize the correlation matrix to assist in identifying descriptors that are highly correlated, like histograms and run charts, to assist in data uniformity. Descriptive statistics is used for further characterization descriptors (conducting a principal components analysis - PCA or cluster) to adequately characterize the data.

Generate QSAR equation

After identifying the independent and dependent variables, the QSAR one or more equations are chosen, using different statistical methods. This may include multiple linear regression, partial least squares (PLS) method, simple linear regression, multiple linear stepwise regression, principal component regression (PCR), etc.;

Validation of QSAR equation

Validation techniques are applied in order to identify data points accumulation outliers (also using graphical analysis and cross analysis) to characterize the robustness of the chosen QSAR model.

Equation analysis

Appropriate graphical tools are used, in order to assess the distribution of biological activity observed from the foreshadowed theoretical identification outlier-PARTIES, or 3D graphics can be created for viewing 3D descriptors important positions in the Molecular Field Analysis (MFA) or Surface Receptor Analysis (RSA), in relation to molecules;

Saving QSAR equation

The calculated QSAR equation can be saved in the database for future use;

Prediction of the activity

The calculated QSAR equation can be used to predict the biological activity of the compounds. In a further development of QSAR analysis, a simple construction of the molecular structure of potentially active compound can calculate the associated biological activity, which is particularly important;

Saving the study

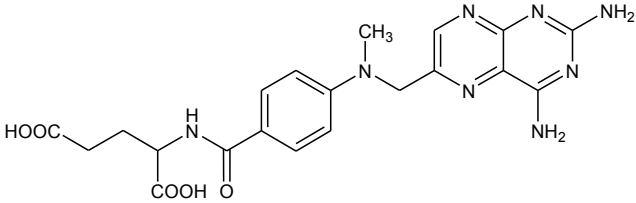
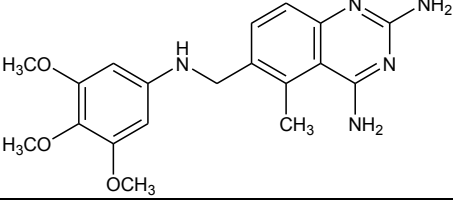
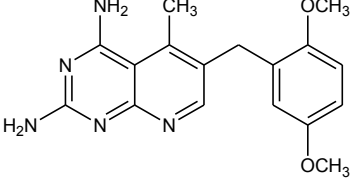
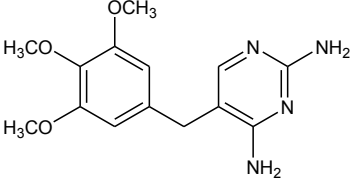
After completing a study, the QSAR analysis can be saved, including all necessary components and conformations structures for evaluation or subsequent use.

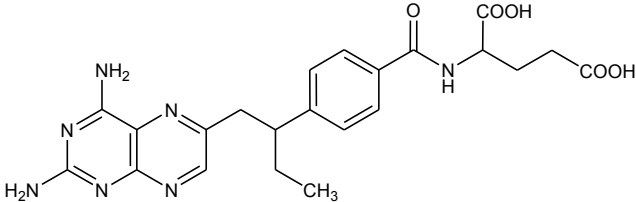
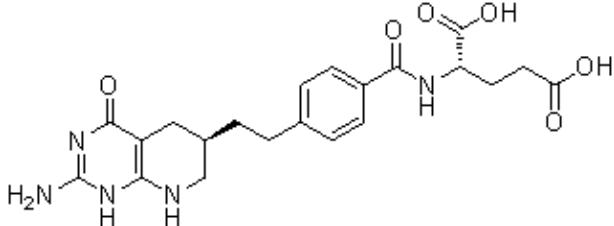
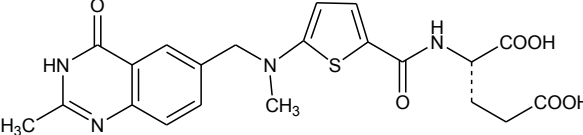
Results and Discussions

In accordance with the steps described in the structured program for QSAR modeling, the analysis was conducted for the conventional antitumor drugs - folic acid analogues (dihydrofolate reductase inhibitors, antifolates and folate antimetabolites): methotrexate and its derivatives, with the molecular structures and the values for biological property IC₅₀ - inhibitory concentration, presented in Table I.

For each of these seven compounds, 224 molecular descriptors were calculated [6,7,10].

Table I
Antitumour drug substances, chemical structures and biological property values

<i>Molecular compound/drug substance</i>	<i>Molecular structure</i>	<i>Value of biological property IC₅₀[*] (microM)</i>
Methotrexate		1.19
Trimetrexate		0.04
Piritrexim		0.03
Trimethoprim		27

<i>Molecular compound/drug substance</i>	<i>Molecular structure</i>	<i>Value of biological property IC₅₀[*] (microM)</i>
Edatrexate		2.01
Lometrexol		0.38
Raltitrexed		12.21

^{*} <http://dtp.nci.nih.gov/docs/cancer/>

The best correlation model for the tested property (the QSAR selected model) was a linear function:

$$\begin{aligned} \text{IC}_{50} = & 67.4429\langle \text{ALogP} \rangle + 2.286\langle \text{BCUTc-1h} \rangle + 0.8532\langle \text{BCUTp-1l} \rangle + \\ & 1.8065\langle \text{nA} \rangle + 3.7924\langle \text{nE} \rangle + 5.5321\langle \text{nG} \rangle + 40.0333\langle \text{ATSc2} \rangle + \\ & 135.1939\langle \text{ATSc4} \rangle + 0.4857\langle \text{khs.dssC} \rangle + 3.1713\langle \text{khs.aaN} \rangle + \\ & 34.1281\langle \text{khs.sssN} \rangle + 1.7041\langle \text{khs.dO} \rangle + 20.1159\langle \text{khs.ssO} \rangle + \\ & 81.2764\langle \text{LipinskiFailures} \rangle + 3.4849\langle \text{WTPT-4} \rangle + 1.75156. \end{aligned}$$

The significance for the molecular descriptors included to the model are presented in references 6, 7 and 10.

For this QSAR predictive model, statistical tests were:

- $R^2 = 0.99986$
- Standard deviation $s^2 = 0.01211$
- Fisher criterion $F = 48.2377$
- Akaike criterion $\text{AIC} = 21.23$

Even the number of descriptors was 16 in the QSAR equation (in a primary, raw and non-refined form), it's power of predictability was very high (Table II).

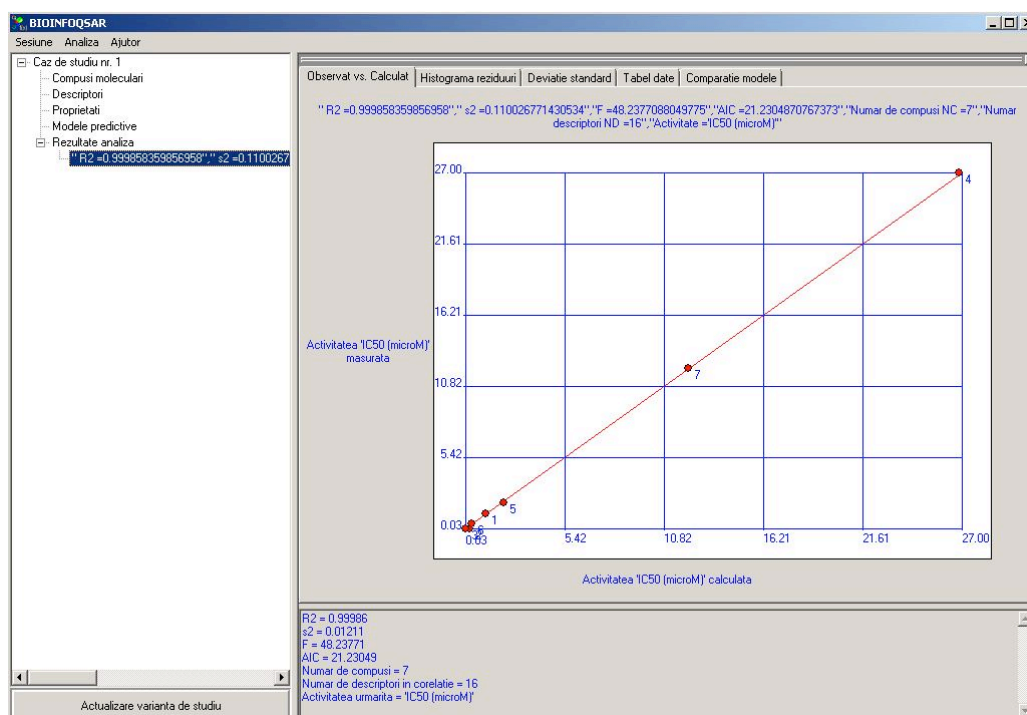
Table II

The observed and predicted values for IC50

<i>Molecular compound</i>	<i>Observed values for IC50 (microM)</i>	<i>Predicted values for IC50* (microM)</i>
Methotrexate	1.19	1.1089
Trimetrexate	0.04	0.2250
Piritrexim	0.03	0.0252
Trimethoprim	27	26.813
Edatrexate	2.01	2.0706
Lometrexol	0.38	0.3409
Raltitrexed	12.21	12.124

*<http://dtp.nci.nih.gov/docs/cancer/>

In figures 4 to 7, the dialog boxes of the interface in the statistical module are shown.

**Figure 4**

Observed values *versus* predicted values for IC50

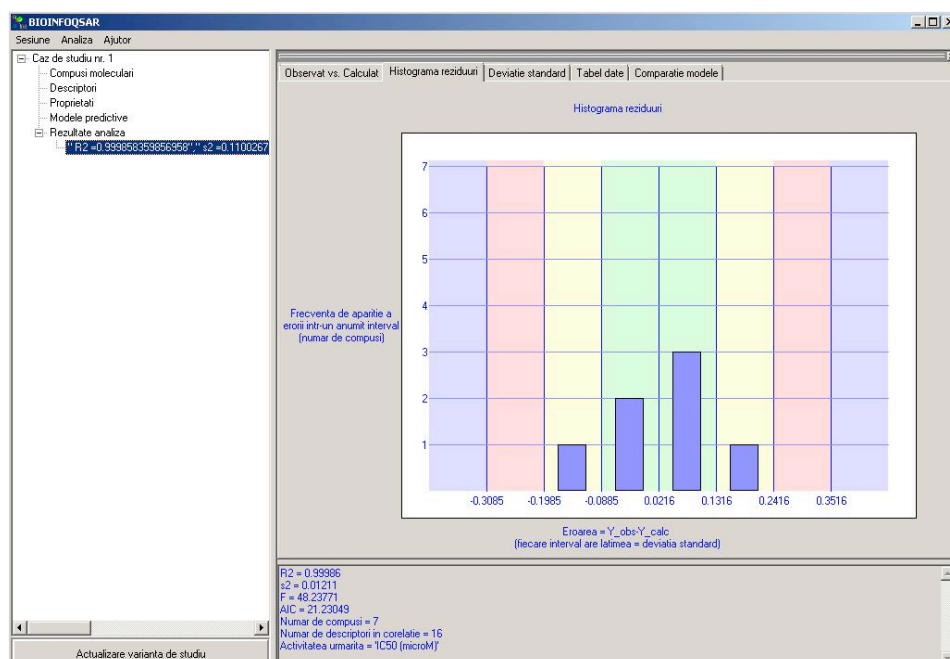


Figure 5
The histogram of the residuals

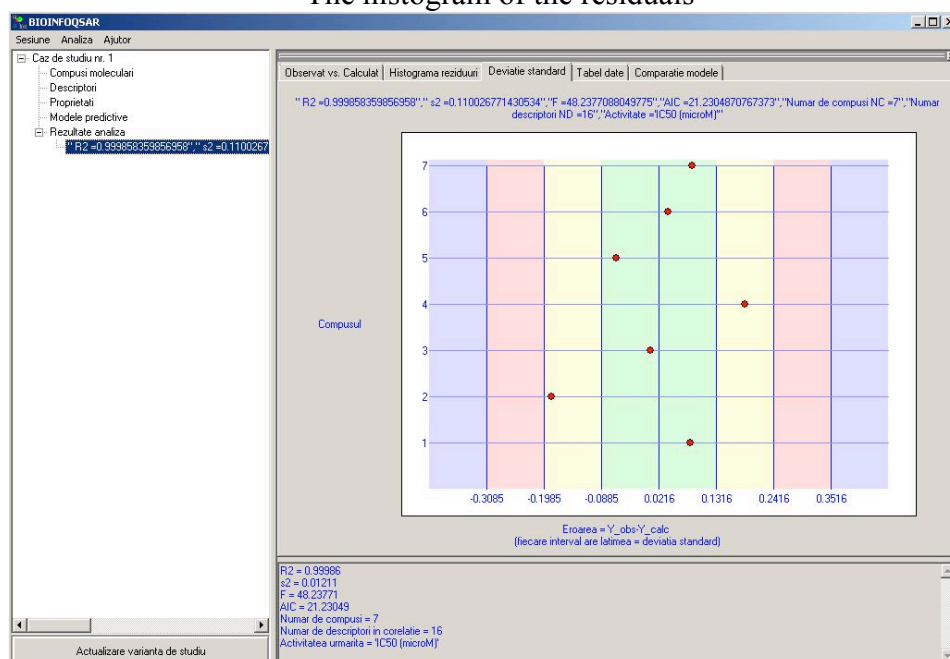


Figure 6
Standard deviation of the errors

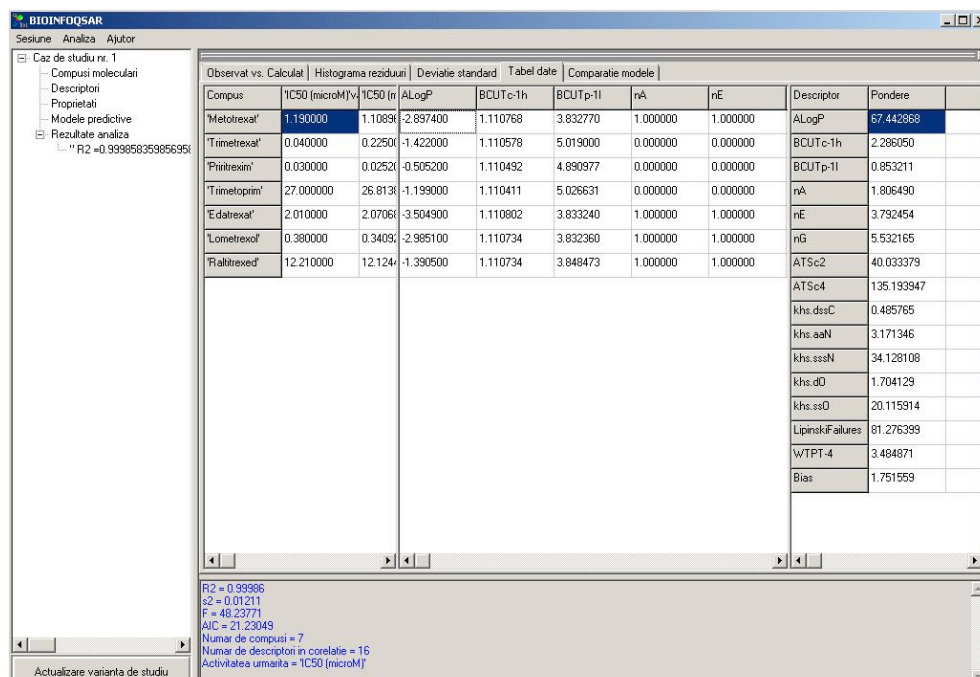


Figure 7
The results of the correlation model

Conclusions

A new software application - BIOINFOQSAR was presented and its functionality was tested.

The study included the development of QSAR analysis in a small series of antitumor drug substances. The algorithm proposed to follow in QSAR modeling has well structured steps and it was applied for the testing of BIOINFOQSAR boosting capacity of predicting. The chosen QSAR model and its statistical tests indicate a very high predictable capacity for this application, even in a raw form of the correlation function.

BIOINFOQSAR proves to have a great advantage: it is applicable in much more areas of research. One important field could be the correlation studies in clinical trials (patients described with quantitative indexes or specific illness degrees, instead of chemical/quantum descriptors, in relationship with their responses to a specific medication). The discovered patterns could be used as predictable tools in other population or data sets. This BIOINFOQSAR becomes a valuable tool in correlating and modeling of various data sets in many other fields of interest, not only chemical, pharmaceutical or other medical areas.

Acknowledgements

The authors thank to the team of Advanced Studies and Research Center that design, create and test the software application. The study was financially supported by the national research grants PNCDI 2, 61-037/2007.

References

1. D.J. Livingstone, The characterization of chemical structures using molecular properties. A survey, *J. Chem. Inf. Comput. Sci.*, 2000, 40, 195-209.
2. P. Gramatica, Principles of QSAR models validation: internal and external, *QSAR Comb.Sci.*, 2007, 26(5), 694-701.
3. A. Golbraikh, A. Tropsha, Beware of QSAR, *J. Mol.Graph Mod.* 2002, 20, 269-276.
4. J. H. van Drie, Pharmacophore discovery-lessons learned, *Curr. Pharm.Des.* 2003, 9, 1649-1664.
5. N.S. Zefirov, V.A. Palyulin, QSAR for boiling points of "small" sulfides. Are the "high-quality structure-property-activity regressions" the real high quality QSAR models?, *J. Chem. Inf. Comput. Sci.*, 2001, 41, 1022-1027.
6. M. Karelson, *Molecular Descriptors in QSAR/QSPR*. New York: Wiley-InterScience, 2000.
7. J. Devillers, A.T. Balaban, (Eds.) *Topological Indices and Related Descriptors in QSAR and QSPR*, Amsterdam: Gordon Breach Sci. Pub., 1999.
8. J. H. van Drie, in: *Computational Medicinal Chemistry for Drug Discovery*, P. Bultinck et al., Eds., Marcel Dekker, 2004.
9. L. B. Kier, in: *Computational Chemical Graph Theory*, D. H. Rouvray (Ed.), Nova Science Publishers, New York, 1990.
10. R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim (Germany), 2000.
11. R. Todeschini, V. Consonni, A. Mauri, M. Pavan, DRAGON—Software for the calculation of molecular descriptors. Ver. 5.4 for Windows, 2006, Talete SRL, Milan, Italy.
12. A.R., Katritzky, V.S., Lobanov, CODESSA, Version 5.3, University of Florida, Gainesville, 1994.
13. V. Vlaia, T. Olariu, L. Vlaia, M. Butur, C. Ciubotariu, M. Medeleanu, D. Ciubotariu, Quantitative structure-activity relationship (QSAR). V. Analysis of the toxicity of aliphatic esters by means of molecular compressibility descriptors, *Farmacia*, 2009, 57(5), 549-561.
14. V. Vlaia, T. Olariu, L. Vlaia, M. Butur, C. Ciubotariu, M. Medeleanu, D. Ciubotariu, Quantitative structure-activity relationship (QSAR). IV. Analysis of the toxicity of aliphatic esters by means of weighted holistic invariant molecular (WHIM) descriptors, *Farmacia*, 2009, 57(4), 511-522.
15. T.W. Schultz, M.T.D. Cronin, T.I. Netzeva, A.O. Aptula, Structure-toxicity relationships for aliphatic chemicals evaluated with *Tetrahymena pyriformis*, *Chem. Res. Toxicol.*, 2002, 15, 1602-1609.
16. R. Benigni, Structure-activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches, *Chem. Rev.*, 2005, 105, 1767-1800.
17. R. Mannhold, A. Petrauskas, Substructure versus Whole-molecule Approaches for Calculating Log P, *QSAR Comb. Sci.* 2003, 22, 466-475.
18. E. Benfenati, G. Gini, N. Piclin, A. Roncaglioni, M. R. Vari, Predicting logP of pesticides using different software, *Chemosphere*, 2003, 53, 1155-1164.
19. G.D. Veith, O.G. Mekenyan, A QSAR Approach for Estimating the Aquatic Toxicity of Soft Electrophiles [QSAR for Electrophiles], *Quant. Struct-Act. Rel.*, 1993, 12, 349-356.

Manuscript received: March 21st 2012