UNIVERSITY OF TARTU
Faculty of Physics and Chemistry
Institute of Chemical Physics

Oleksandr Kulshyn

Chemical Database of Optimized Molecular Structures

Thesis for Master's Degree

Supervisor: Professor Mati Karelson

Tartu 2006

CONTENTS

ABBREVIATIONS

ACD	Available Chemicals Directory
AM1	Austin Model 1
API	Application Programming Interface
BMLR	Best MultiLinear Regression
CAS	Chemical Abstracts Service
CODESSA	COmprehensive DEscriptors for Structural and Statistical Analysis
DBOG	DataBase of Optimized Geometries
DIPPR	Design Institute for Physical Property Data
HTML	HyperText Markup Language
InChI	International Chemical Identifier
ISIS	Integrated Scientific Information System
IUPAC	International Union of Pure and Applied Chemistry
JME	Java Molecular Editor
LAN	Local Area Network
LCAO	Linear Combination of Atomic Orbitals
MDL	Molecular Design Limited,
MO	Molecular orbitals
MOPAC	Molecular Orbital PACkage
NDDO	Neglect of Diatomic Differential Overlap
PHP	Personal Hypertext Preprocessor
PM3	Parameterized Model 5
QSAR	Quantitative Structure - Activity Relationship(s)
QSPR	Quantitative Structure - Property Relationship(s)
SMILES	Simplified Molecular Input Line Entry Specification
SQL	Structured Query Language

INTRODUCTION

Structure-based virtual screening has had several important successes in recent years [1, 2, 3] and is now a common technique in early stage of drug discovery at most pharmaceutical companies as well as some university groups. Unfortunately, virtual screening techniques continue to require expert knowledge and extensive infrastructure and remain out of reach for many medicinally and biologically oriented investigators who might otherwise be able to exploit them. Among the steepest barriers to entry is the lack of a suitable database of small molecules with which to screen. These databases are either expensive to acquire or time-consuming and difficult to prepare and curate. To be useful for structure based screening, 3D structures must be calculated for each available molecule. More difficult are the problems related to the calculation of manifold protonated, stereo- and regiochemical, tautomeric, and conformational states for the database molecules. Computing these multiple molecular species and states is challenging and is the focus of ongoing research [4].

QSPR methodology has been aided by new software tools, which allow chemists to elucidate and to understand how molecular structure influences properties. Very importantly, this helps researchers to predict and prepare structures with optimum properties. The software is also of great assistance for chemical and physical interpretation.

In the past fifteen years, multipurpose statistical analysis software in the form of the CODESSA (COmprehensive Descriptors for Structure and Statistical Analysis) program has been developed, recently updated as the CODESSA PRO program [5].

For a satisfactory QSAR treatment, it is essential that good quality input data are utilized: *i.e.* a set of structures and quantitative measurements of the property, measured under the similar conditions with satisfactory reproducibility and accuracy. The preparation of the input data in CODESSA PRO utilizes a molecular editor or direct import of the structures from a chemical database. The 3D-geometries are generated and optimized using molecular mechanics and semi-empirical quantum-chemical methods such as PM3, AM1 in MOPAC [6], etc.

Throughout the years, the computational chemistry groups at the Center for Heterocyclic Compounds at University of Florida and at University of Tartu had numerous projects dealing with a large number compounds, counting more than 20 000. These compounds had been used in the development of QSAR/QSPR models for numerous

physicochemical and biomedical properties. One of the main steps in QSAR/QSPR modeling is the optimization of the geometries and the descriptor calculation of the compounds. It is particularly important because a large part of the molecular descriptors are calculated from the quantum chemical wave function and energies of molecules.

The main goal of this work was to create comprehensive database that collects the compounds with already quantum-chemically optimized geometries for QSAR modeling. Thus, it is possible to avoid repetitive optimization of compounds, which overlap among the different projects. In addition, the working process for QSAR modeling would speed up greatly by using the flexibility of the database. Also, using such a database gives more reliability to the prediction of structures of newly developed compounds and the respective QSAR/QSPR models. It also assures that the projects based on the same optimized structures and results throughout different projects are comparable.

1. LITERATURE OVERVIEW

A number of databases have been already implemented to accommodate the virtual screening and QSAR/QSPR development needs. However, most of them contain only compounds belonging to a particular class of chemicals or include only few properties for a given compound. The development of a database that would store optimized geometries of compounds would give the ability to calculate hundreds of descriptors in short time. Numerous previous examples were used as models for the development of the database of optimized geometry (DBOG), the object of the present work.

A database, developed at the Brigham Young University, DIPPR (Design Institute for Physical Properties) contains more than 1800 compounds and lists 48 thermodynamic properties, 33 physical constant properties and 15 temperature-dependent properties for each compound [7, 8]. Compounds that are collected in the database are used by industries worldwide [9, 10].

MDL (Molecular Design Limited) has come up with their version of database for virtual screening. The MDL Available Chemicals Directory (MDL ACD) is the "grandfather" of chemical sourcing databases [10]. Trusted and in use by over 20,000 scientists at over 500 sites, for more than 20 years MDL ACD has been the *de facto* standard in pharmaceutical, biotechnology, chemical and agrochemical companies worldwide.(web mdli.com). The database contains about 480,000 purchasable compounds. Having great success with ACD and years of experience, MDL has developed Screening Compounds Directory formerly known as ACD-SC. SCD is basically an online version of ACD database. An important feature for this database is that scientists can access online to search for particular compound without the need of installing in-house ACD database, which requires Oracle, MDL ISIS/Host, MDL ISIS/Base and few updates per year.

Another free database of commercially available compounds for virtual screening is ZINC. It has used MDL ACD as the “golden standard”. The developers of the database had been focused on collecting several properties for each molecule, such as molecular weight, calculated LogP, number of rotating bonds. It also indicates the biologically relevant protonation states of molecules thus making them applicable for docking modeling using different popular docking programs. This database contains about 720,000 molecules with 3D structure and list of vendors that sell particular compound [12, 13].

All mentioned databases contain great amount of important information. However, they still have some essential drawbacks. One of the biggest shortcomings of them is that

the geometrical structures of the compounds are not optimized using any quantum-chemical method. In addition, they also do not contain information such as the calculated energies and optimization parameters applied (e.g. the gradient norm from the last optimization cycle) of the molecular structure.

Generally, database is a dynamic collection of information stored in a certain way. Hence, this information needs constant update. This is the case of MDL ACD database, which is commercial and thus requires constant update. Often the databases are used in conjunction with auxiliary software. For instance, MDL ACD contains 3D models for all compounds, produced using Corina, a software developed by Molecular Networks GmbH. However these compounds are not optimized using quantum-chemical methods, just a 3D structure derived form standard bond lengths and angles is given. Another example is DIPPR that contains many important properties but is too small to be used in large projects. Finally, ZINC is a free database, which is a big advantage but it only contains data that are mostly suitable for pharmaceutical companies.

2. DATABASE OF OPTIMIZED GEOMETRIES (DBOG)

2.1 Ideology of DBOG

In searching for suitable database on optimized molecular geometries we found that there are no databases that can properly meet all the needs of the QSAR/QSPR modeling. Since QSAR/QSPR is a vast area of modern computational chemistry, the researchers dealing with large number of compounds need to have easy and fast access to the database storages. Hence, one of the practical requirements for a good database is its accessibility. Thus, it should allow sharing the information among the geographically isolated groups involved in a given project. The respective remote nodes (computers) can access the database, perform queries to get the structure for specific compound and/or update a certain structure in the database. The general network connectivity of the DBOG server is given in Figure 1.

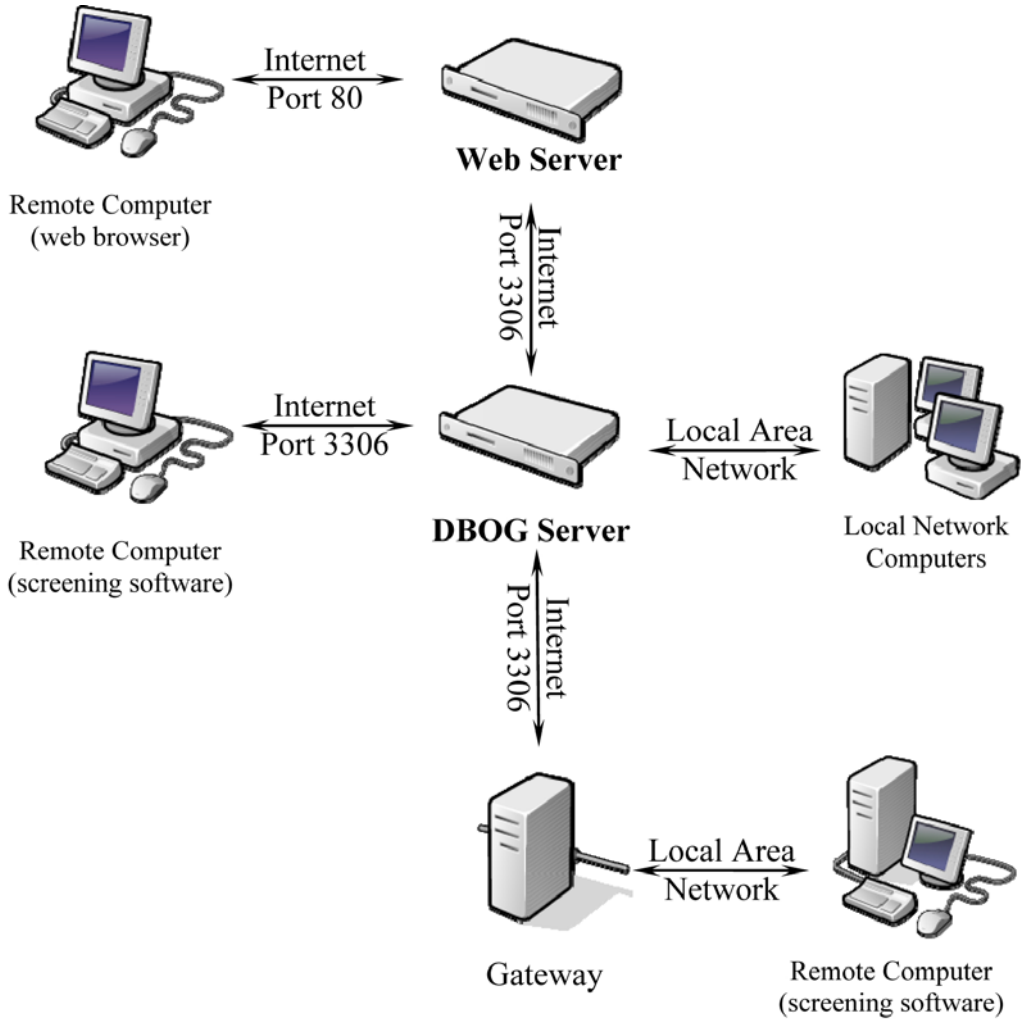


Figure 1. General layout of network connectivity

As shown in figure 1, DBOG is accessible remotely by two general nodes (computers), namely i) node connected to DBOG server via local area network (LAN) and ii) node (computer) connected to DBOG server via Internet. In the case i) the computers in the LAN are directly connected to the server. Thus, the accessibility speed depends on the LAN ability to transfer data and in most of the cases is the fastest way to reach the database records. As concerned to case ii), one should have necessary ports open in order to be able to establish connection. The DBOG database is running on default MySQL port 3306. If a remote computer is trying to establish connection using our screening software, it is necessary to have port 3306 for communication between application and server. If remote computer is located inside companies network which uses gateway server to connect to internet, port 3306 should be forwarded by the gateway server remote computer. There are also possibilities to perform search using web browser. This connection only requires port 80 to be open, which in many cases is the standard for web browsing.

2.2 Programming API of DBOG.

The development of a comprehensive database is certainly not an easy task. It requires rigorous scheme of different levels of consecutive steps connected with straightforward logic. Therefore, the choice of the building environment is of a great importance to establish the connections between these levels. In practice there are several application programming interfaces (API) for database building such as Oracle, Microsoft SQL Server, Microsoft Access, etc. All of these programming environments have their advantages and disadvantages. Our choice of programming environment for DBOG was MySQL. The reasons for this choice were as follows:

1) open source (in contrast to Microsoft SQL Server)
2) free (in contrast to Oracle and Microsoft SQL Server)
3) high database standards (in contrast to Microsoft Access)
4) high portability, runs on different operating systems (in contrast to Microsoft SQL Server)
5) high programming flexibility

Nowadays the criteria 1) and 2) are very important. Therefore, the DBOG has been developed as a free database on optimized molecular geometries available to use by other chemistry groups.

A standard structure of the database was used to accommodate the storage of the arrays of compounds and their chemical structures. It consists of two main tables: Molecule Table (MOL_TABLE) and Structure Table (STRUCT_TABLE) (see Figure 2). The Molecule Table contains common data for each molecule. Some of the fields are: Molecular Formula (mol_formula), Molecular Name (mol_name), Molecular Weight (mol_weight), CAS number (mol_cas) and a unique structural identifier InChI (mol_inchi) (see Figure 2). Besides InChI (see sect. 2.3), each molecule is identified by its own unique ID (mol_id) in the database. This ID is later used to connect the Structure Table to a Molecule Table. The Structure Table contains a separate unique ID number (struct_id) for each structure. In addition, it contains molecular ID, which helps to identify structure and other important fields: Quantum-Chemical (Semi-Empirical) Method (struct_method), Total Molecular Energy (struct_energy), Gradient Norm (struct_gradient), content of structure file (struct_file), file type (struct_format) and file extension (struct_file_ext). Later, it was decided to add additional table that stores the descriptor values for each structure. This table (DESCR_TABLE) contains descriptor ID and value, while another table (DESCR_NAME_TABLE) contains descriptor name connected by descriptor ID to DESCR_TABLE, see Figure 2.

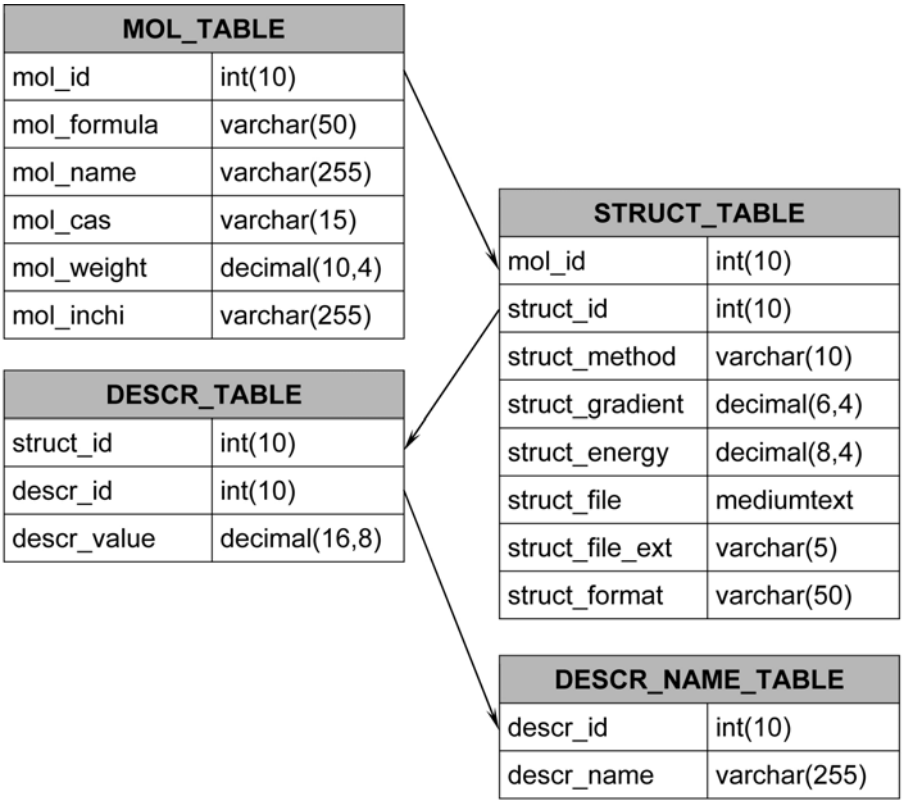


Figure 2. The scheme of the database tables

2.3 Characterization and implementation of DBOG

Since, the DBOG is a collection of many records (structures), it needs to possess straightforward criteria for input and output procedures. These criteria carry information for a given structure representation. All molecules in DBOG are two-dimensional and are expressed as InChI (International Chemical Identifier) code. InChI is a non-proprietary identifier for chemical substances that can be used in printed and electronic data sources thus enabling easier linking of diverse data compilations. It was developed under IUPAC Project 2000-025-1-800 during the period 2000-2004 [14]. Chemical structures are expressed in terms of five layers of information - connectivity, tautomeric, isotopic, stereochemical, and electronic. The InChI algorithm converts the input structural information into the identifier in a three-step process: normalization (to remove redundant information), canonization (to generate a unique set of atom labels), and serialization (to give a string of characters) [15, 16, 17, 18]. By using InChI each structure in the database can be correctly identified according to the InChI code in MOL_TABLE (see Fig. 2). Though most of the databases use SMILES (simplified molecular input line entry specification) for canonical serialization of molecular structure [19, 20, 21], it is not open source project as InChI is. This had led to many different conversion algorithms and different versions of SMILES for the same compound. As an example, seven different SMILES formulations can be found for caffeine (Figure 3). As can be seen from the figure the InChI presentation of the caffeine is unique, in contrast to the 7 versions of SMILES.

The DBOG is characterized by two processes:

A) Structure submission

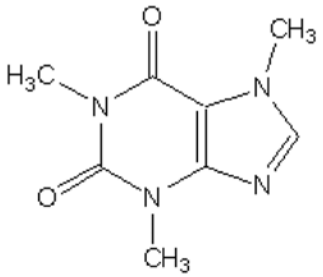
The process of submission allows the user to input own structure for a given compound according to the criteria for the total molecular energy and gradient norm. From the general physical considerations (variationally calculated quantum-chemical energies), the lower these criteria the better the optimized structure should be. However, this conjecture depends on the quantum-chemical method used for optimization, addressed to the different conformers of a given structure. In the DBOG case, these are the total molecular energy, obtained by MOPAC using AM1 or PM3 semi-empirical methods and the gradient norm used as stopping condition.

Structure submission process uses two methods for submission of data:

1) submission through a local node (LAN)
2) submission through a remote node (Internet)

1. [c]1([n+]([CH3])[c]([c]2([c]([n+]1[CH3])[n][cH][n+]2[CH3]))[O-])[O-]
2. CN1C(=O)N(C)C(=O)C(N(C)C=N2)=C12
3. Cn1cnc2n(C)c(=O)n(C)c(=O)c12
4. Cn1cnc2c1c(=O)n(C)c(=O)n2C
5. N1(C)C(=O)N(C)C2=C(C1=O)N(C)C=N2
6. O=C1C2=C(N=CN2C)N(C(=O)N1C)C
7. CN1C=NC2=C1C(=O)N(C)C(=O)N2C

Caffeine



InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

Figure 3. Different SMILES formulations found for caffeine on the Web

The two methods differ by their connections to the DBOG server (see Figure 1). Method 1) can be used in batch mode in contrast to 2). In addition, method 1) uses an auxiliary program written in C++ that implements InChI open source code libraries to convert a certain structure (in a given format, e.g. MDL MOL file) into InChI format string. Then by using MySQL C API, it allows multiple structures to be submitted to the database, whilst method 2) allows the user to upload only one structure at a time. Once the structures are uploaded, the molecular descriptor calculation for a QSAR/QSPR model development can start using suitable software (e.g. CODESSA PRO). Within CODESSA PRO, the open source of MOPAC has been used to calculate the descriptors and find the total molecular energy of the structure. The MOPAC (Molecular Orbital PACkage) is a semi-empirical quantum chemistry program based on Dewar and Thiel's NDDO approximation. [22, 23]. After the descriptor calculations are completed, the resulting data are returned directly to the database.

The second method is the submission of data over the internet. Before upload, the webpage asks for the quantum-chemical method used during optimization, calculated total molecular energy and gradient norm. The server will run a verification procedure by calculating the total molecular energy and gradient norm, using MOPAC. Next step is the

check for existence. If a structure already exists in the database, the results of calculations are compared to results for the same structure already stored in the database. By performing these steps, it guarantees that there are always structures with the lowest total molecular energy and gradient norm in DBOG. Because structures can be optimized using different conformers, it is possible to get better results by using a different starting point. If the new structure has lower total molecular energy and gradient norm or if there is no similar structure in the database, structure is passed into descriptor calculation step and then stored in the database. In this case, it is considered that all the descriptor calculations are performed on the remote server.

Figure 4 shows the general flowchart of the whole procedure during submission process.

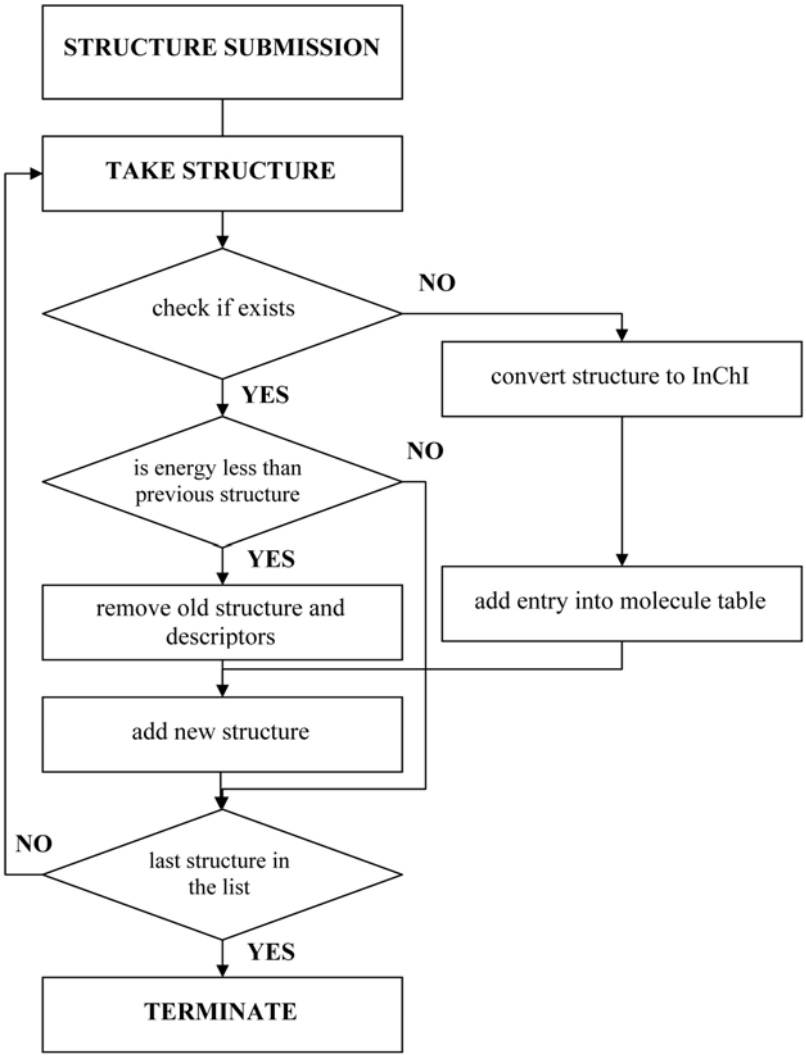


Figure 4. Flowchart of the data submission process

B) Structure retrieval

The retrieval of the data from the database can be also carried out using world wide web. The webpage for the search was built using PHP and JavaScript. It implements online sketcher using JME Molecular Editor Applet that allows drawing of a structure and later converting it into InChI for querying [24]. JME Molecular Editor is a Java applet which allows to draw / edit molecules and reactions (including the generation of substructure queries) and to depict molecules directly within an HTML page. This editor can generate Dayligh SMILES or MDL MOL files of created structures.

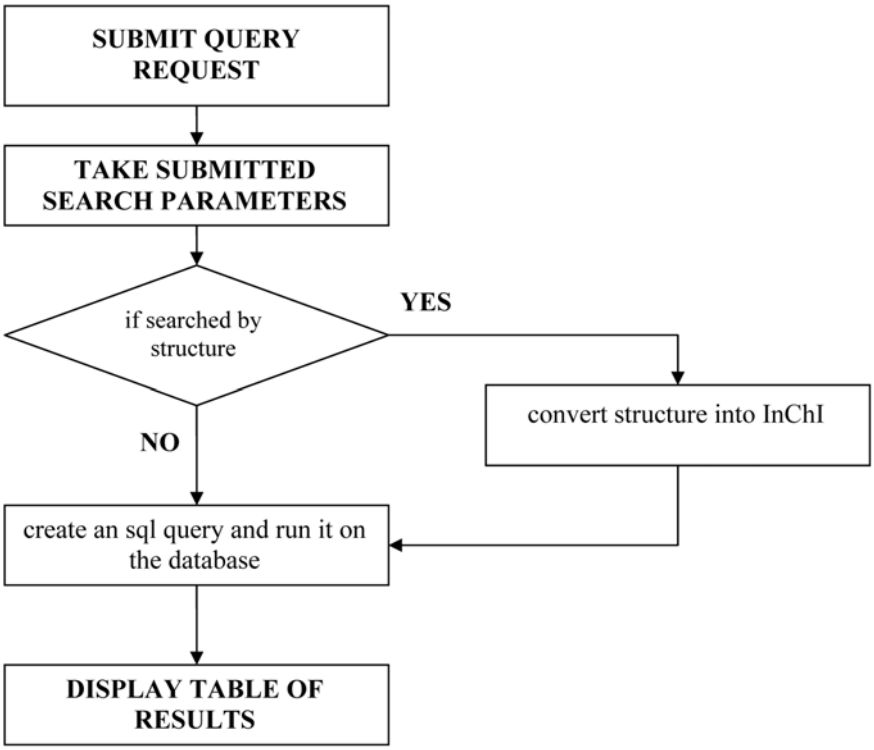


Figure 5. Flowchart of retrieval process

The applet has been developed by Peter Ertl as a part of web-based chemoinformatics and molecular modeling system at Novartis [24]. Due to many requests, the applet has been released to the public and become standard for molecular structure input on the web with more than 3500 installations worldwide.

After drawing 2D structures user can input the gradient norm and the quantum-calculated total molecular energy at the given level of theory. The database is designed to store only the best available structure so it automatically deletes structure with higher total molecular energy when a better one is submitted. Because of this database will return

structures, optimized using different quantum-chemical methods, for the same compound if no search criteria is defined. The result page returns the data, including the number of descriptors available for a given available structure. It is possible to view the descriptor values on the screen and download the structure into local computer.

Besides the web search, a standalone screening software has been developed by us for carrying out similar search functions. It was written in C++ and incorporates the quantum-chemical wave function calculation code of MOPAC. It works in the similar way as webpage. The input to the software includes the interactively submitted 2D structures, desired semi-empirical method or/and energy and the gradient norm. The results of the query give the number of structures available and the number of descriptors in the database for each structure. Alternatively, the information about the structure can be downloaded as a file and the retrieved descriptors can be viewed on the screen or saved as a tab-separated text file.

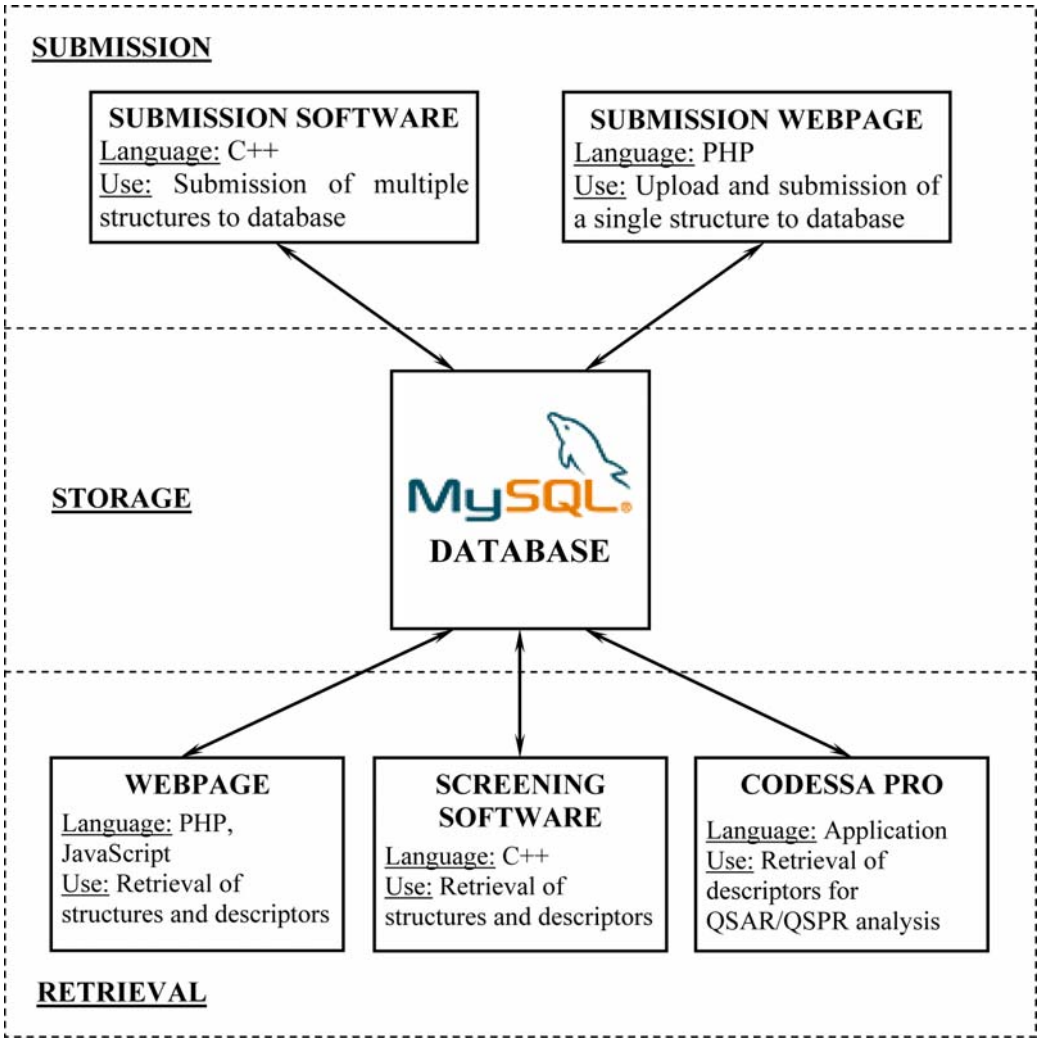


Figure 6. Database access availability

There is also a possibility to connect CODESSA PRO or any other software to DBOG database by creating an API connectivity package. This kind of solution would speed up preparation of the QSAR/QSPR models using different quantum-chemical software since the number of descriptors could be retrieved directly from the database. However, this is not implemented yet in the current version of CODESSA PRO.

The general implementation scheme of DBOG is given in Figure 6. From this figure, it can be seen that the main core storage interacts with the two levels of submission and retrieval by means of additional software programs written in C++, PHP and JavaScript. The software flexibility of MySQL database allows combining all molecules in one robust database which can interact with the user through internet or local computers.

3. EXPERIMENTAL UTILIZATION OF DBOG IN QSAR/QSPR

3.1 Antimalarial activities of compounds

The test application subjected to the DBOG was a QSAR study of antimalarial activity of chemical compounds [26] (Appendix). Malaria is well-known as an infectious lethal disease since ancient times, and remains a major cause of death. Spread by *soporoza* of the *genus plasmodium*, it is characterized clinically by periodic fever, anemia, and enlargement of the liver and spleen. Hundreds of millions of new clinical cases arise annually with a high percentage of fatalities, especially among children [25], in the tropical and subtropical countries of Asia, Africa, and South America.

A specific characteristic of the data for the antimalarial project was that only a limited number of drugs can prevent and cure malaria. Therefore, a careful selection of the compounds on which the QSAR modeling was based needed to be performed. In this stage of the QSAR building, DBOG was used to find and collect significant data set for the property under investigation (log IC50).

The general steps of working with DBOG for the QSAR investigation of the antimalarial activity were as follows:

I) Choice of initial search fields

This step includes searching by certain field criteria as shown in Figure 6 according to the compound data related to the property in question, in our case antimalarial activity (log IC50). The most straightforward way is to use the CAS number of the compounds. After literature search for drugs related to the antimalarial activity, we had collected more than 275 candidates with their CAS numbers. These CAS numbers were loaded into DBOG and checked for availability. Thus, the process of checking was less than five minutes to obtain the compounds that we have already optimized (by AM1) in our database. It was found that 174 drugs (out of 275) were readily available in DBOG.

II) General refinement

The structures of the selected compounds from step I) were later refined by checking alternative initial geometries and the compatibility of experimental data from different sources, and the final number of compounds chosen for QSAR treatment was 126. Hence, at this stage of the QSAR modeling, a significant data set with already optimized geometries of compounds was available. Also the molecular descriptors for them were calculated. In this data set, the drug molecules were rather large and the

geometry optimization process could have been time consuming. Thus, by using DBOG, we could skip the process of optimization of the molecules available in the DBOG and therefore shortened substantially the modeling timeframe. As the calculation of the molecular descriptors is also related to comparatively large amount of computing time when the structures are large, the use of DBOG gave additional savings in QSAR modeling time.

A specificity of this study was that the QSAR modeling was applied on two different datasets, regarding to two different malarial strains (D6 and NF54). Accordingly, the selection was performed by DBOG by splitting 126 compounds into the respective two datasets. As can be seen from Figures 7 and 8 DBOG interface allows carrying out fast screening and searching of the compounds based on their 2D structure, molecular weight, name of the drug.

III) Structure submission

During the process of refinement of 3D structures it is possible that the researcher may find structures in DBOG that are not satisfactory optimized according to the total molecular energy. In this case, DBOG allows him to update certain structure in the database. In the case of antimalarial project, several drugs were not optimized at the desired gradient norm (e.g. structures 58-62 in Table 2 of the article attached as Appendix). After proper optimizations at the desired level and descriptor calculations, these structures were submitted back to the DBOG by the procedure shown in Figure 4. This property of DBOG provides constant ability to update the database records.

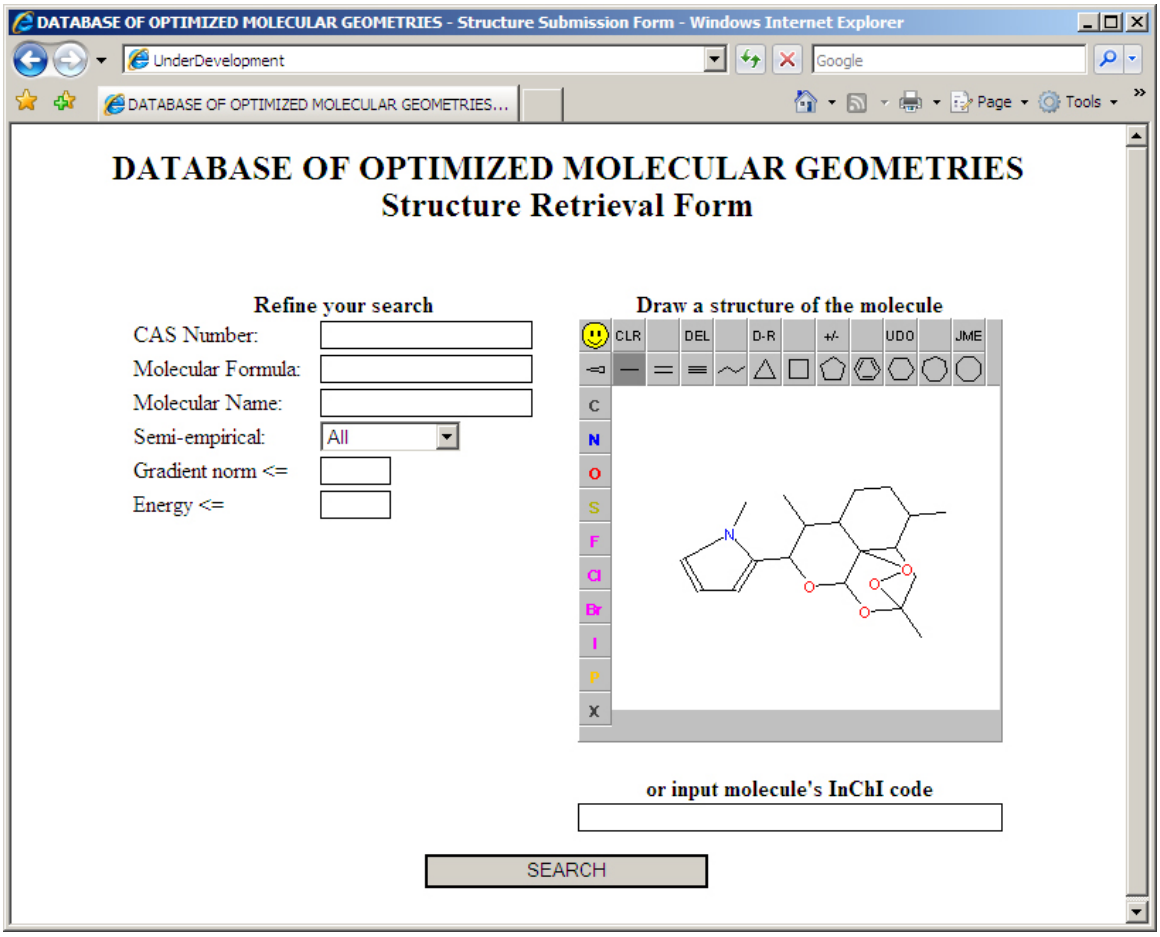


Figure 7. Screenshot of search page

IV) 3D structure extraction

From the results table, structures can be extracted one by one, by clicking on 'MOL' button or by doing batch download. Batch download allows downloading multiple structures at once by selecting them. As a result, a zip file that contains the structures was downloaded. All the structures were stored as MDL MOL files and can be used with CODESSA PRO without conversion.

ID	Structure	CAS	Molecular Formula	Molecular Weight	Molecular Name	Semi-empirical	Gradient Norm	Energy	Number of Descriptors	Download file
3427		204503-67-3	C20 H29 N O4	347.45	1H-Pyrrole, 2-[(3R,5aS,6R,8aS,9R,10R,12R,12aR)-decahydro-3,6,9-trimethyl-3,12-epoxy-12H-pyrano[4,3-j]-1,2-benzodioxepin-10-yl]-1-methyl- (9CI)	AM1	0.03127	-102510.1719	647	MOL
3632		220114-98-7	C26 H33 N O4	423.54	1H-Pyrrole, 2-[(3R,5aS,6R,8aS,9R,10R,12R,12aR)-decahydro-3,6,9-trimethyl-3,12-epoxy-12H-pyrano[4,3-j]-1,2-benzodioxepin-10-yl]-1-(phenylmethyl)- (9CI)	AM1	0.04362	-121492.3828	605	MOL
2877		220115-01-5	C24 H31 N O5	413.51	1H-Pyrrole, 2-[(3R,5aS,6R,8aS,9R,10R,12R,12aR)-decahydro-3,6,9-trimethyl-3,12-epoxy-12H-pyrano[4,3-j]-1,2-benzodioxepin-10-yl]-1-(2-furanylmethyl)- (9CI)	AM1	0.03705	-122315.3047	765	MOL

Figure 8. Screenshot of results page

The flexibility of DBOG (in conjunction with CODESSA PRO) interface allowed us to select very rapidly two data sets of 57 and 69 compounds for D6 and NF54 strains, respectively. After finishing the working step III) 3D structures of the selected compounds in certain format were extracted. DBOG supports easy to use interface to retrieve the desired 3D structure in MDL MOL files as shown in Figures 5 and 8. At this stage, the actual QSAR could start by building predictive equations that require the already available molecular descriptors in DBOG. However, the selected descriptors were reloaded into CODESSA PRO to carry out the statistical analysis and the QSAR model development.

The screening and searching by DBOG was executed on a local network computer as shown in Figure 1. Also, these procedures can be executed by a remote user via Internet (see Fig. 1). Therefore, provided that the remote user has access to the database server, he/she can use the DBOG for his/her research independently from their geographical location.

The total of 961 different molecular descriptors were refined and calculated. Derived solely from molecular structure, they were divided into the following classes: (i) constitutional, (ii) geometrical, (iii) topological, (iv) electrostatic, (v) quantum chemical, and (vi) thermodynamic. These descriptors are based on the molecular geometry, LCAO MO wave and thermodynamic functions calculated by using the MOPAC program package.

The best multilinear regression (BMLR) procedure was used to find the best correlation models from selected non-collinear descriptors [27]. The BMLR selects the best two-parameter regression equation, the best three-parameter regression equation etc., based on the highest R^2 value in the stepwise regression procedure.

By using the best multilinear regression method equations for the both strains were constructed with up to six descriptors. A simple rule (“breaking point” rule) was used to decide the optimum number of descriptors by considering the improvement of the R^2 by addition of a further descriptor to the model. If the difference between the models with n and n+1 descriptors is improved by a value of less than 0.04, then the optimum model is taken to have n descriptors. The selection of the optimum number of the descriptors is shown in Figure 1 of attached article. In addition, the Fisher criterion was also monitored for a significant improvement in the correlation coefficient value with respect to the number of the descriptors. The final QSAR models selected for the two malaria strains (D6 and NF54) are shown in Tables 3 and 4 of attached article, respectively.

3.2 Other applications

DBOG has been also used for the preparation of data in other QSAR/QSPR model development projects. As an example, the study “Neural Networks Convergence Using Physicochemical Data” article [28] dealt with a large number of compounds collected from different datasets concerning different physicochemical properties, namely:

i) 411 vapor pressures
ii) 298 boiling points
iii) 60 carcinogenic activities
iv) 115 milk/plasma ratios
v) 137 organic compounds with measured ozone tropospheric degradation rates
vi) 158 skin permeation rates
vii) 57 p-glycoprotein inhibitor activities
viii) 115 blood-brain partition coefficients.

In this study the DBOG was very useful since such a large number of compounds requires excessive computational time. By using DBOG it was possible to prepare five data sets for less than one hour (sets i, ii, iv, vi and viii). The reason for this fast collection was that these data had CAS numbers available and had been already accommodated into DBOG. However, the remaining datasets (v, vii and iii) searched by criteria molecular name and InChi code (see Fig. 7), showed that not all compounds were available in DBOG. Generally, 30 % of the structures were not available in the database. These structures were thus drawn manually and added to DBOG storages.

Importantly, the use of DBOG enabled to start the QSAR investigation in less than two days.

CONCLUSIONS

In this work, an open source database on optimized molecular structures (DBOG) was developed, applicable in QSPR/QSAR modeling. The optimization of the molecular geometrical data using quantum-chemical methods can be, depending on the size of molecule, excessively time-consuming. DBOG provides instant access to 3D structures optimized using different semi-empirical methods as well as descriptors calculated. The ability to store and view descriptors makes it even more useful for QSAR modeling.

A key feature of the database is the open source. Availability of the source code to public can lead to many improvements for certain needs of the researcher. The database can also be adapted for a specific scientific group. Companies can use it to store confidential data with limited access by setting up the Database Server inside their network.

To help speed up the research process we have made an easy to use interface. Both the screening software and web-based interface provide direct access to the data stored in the databases. The search function is straightforward but allows creating fairly complex search queries to narrow down the results. It also allows viewing the set of structures created for the same substructure. All these promising features of DBOG were applied, as an example, on a QSAR investigation of antimalarial activity and other QSAR/QSPR projects.

Uploading a structure into this database helps to share information between chemists. It also improves the data available and brings updates to a database on regular bases.

Therefore, DBOG is a helpful tool in virtual screening for many experts and scientists and it will enable more possibilities of high scale research in computational chemistry.

KOKKUVÕTE

Antud töös arendati välja avatud lähtekoodiga keemiline andmebaas (DBOG) molekulide optimeeritud struktuuride käsitlemiseks, mis on rakendatav kvantitatiivsete struktuur-omaduste/aktiivsuste sõltuvuste (QSPR/QSAR) modelleerimise protsessis. Sõltuvalt struktuuridest, võib molekulide geomeetria optimeerimine kasutades kvantkeemilisi meetodeid olla liigselt aeganõudev. DBOG pakub erinevate pool-empiiriliste meetoditega optimeeritud valmis 3D struktuure, ning samuti vastavatele struktuuridele arvutatud molekulaardeskriptoreid. Viimaste salvestamise ja visualiseerimise võimalus teeb andmebaasi veelgi mugavamaks QSPR/QSAR arendustele.

Andmebaasi eriomaduseks on avatud lähtekood. Algkoodi avalik kättesaadavus võimaldab kõigil andmebaasi täiendada vastavalt nende kasutajate vajadustele. Andmebaasi on võimalik kohandada ka vastavalt spetsiifilistele uurijate gruppidele. Ettevõtete puhul on samuti võimalus salvestada konfidentsiaalseid andmeid, seades üles andmebaasi serveri nende endi piiratud kasutusõigustega võrgus.

Et kiirendada teadustöö protsessi, on lisatud kergesti kasutatav liides. Nii skriinimistarkvara kui ka veebil baseeruv liides pakub vahetut juurdepääsu salvestatud andmetele. Otsingufunktsioon on otsene, kuid võimaldab ka koostada üsna keerukaid päringuid, vähendamaks vastete hulka. Samuti võimaldab ta vaadelda struktuuride seeriaid, mis on loodud ühise alamstruktuuri baasil. Kõik DBOG funktsioonid leidsid rakendamist malaariavastase aktiivsuse QSAR modelleerimisel ning teistes QSAR/QSPR projektides.

Struktuuride sisestamine andmebaasi aitab jagada informatsiooni teadlaste vahel. Samuti väldib ta dublikaatide teket ning parandab kirjeid andmebaasis automaatselt.

Seetõttu on DBOG kasulik vahend keemiliste ühendite virtuaalsel skriinimisel ning pakub erinevaid võimalusi kõrgetasemelisele arvutikeemiale.

ACKNOWLEDGMENTS

I would like to thank my supervisor Professor Mati Karelson for his excellent guidance throughout my research.

I would also like to express my sincere gratitude to Kenan Professor Alan R. Katritzky for his support and guidance.

REFERENCES

1. Schapira, M.; Abagyan, R.; Totrov, M. Nuclear hormone receptor targeted virtual screening. *J. Med. Chem.* **2003**, *46*, 3045-3059.
2. Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813-1818.
3. Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discov. Today* **2002**, *7*, 1047-1055.
4. Fornabaio, M.; Cozzini, P.; Mozzarelli, A.; Abraham, D. J.; Kellogg, G. E. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 2. Computational titration and pH effects in molecular models of neuraminidase-inhibitor complexes. *J. Med. Chem.* **2003**, *46*, 4487-4500.
5. http://www.codessa-pro.com
6. Stewart, James J. P. MOPAC : a semiempirical molecular orbital program. *J. Comput. Aided Mol. Des.* **1990**, *4(1)*, 1-105.
7. http://dippr.byu.edu/
8. Thomson, G. H. The DIPPR databases. *Int. J. Thermophys.* **1996**, *17(1)*, 223-32.
9. Wilding, W. Vincent; Rowley, Richard L.; Oscarson, John L. DIPPR Project 801 evaluated process design data. *Fluid Phase Equilib.* **1998**, *150*, 413-420.
10. Goodman, Benjamin T.; Wilding, W. Vincent; Oscarson, John L.; Rowley, Richard L. Use of the DIPPR Database for Development of Quantitative Structure-Property Relationship Correlations: Heat Capacity of Solid Organic Compounds. *J. Chem. Eng. Data* **2004**, *49(1)*, 24-31.
11. http://www.mdl.com/
12. Irwin, John J.; Shoichet, Brian K. ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Mod.* **2005**, *45(1)*, 177-82.
13. Irwin, John; Shoichet, Brian. *The ZINC database as a new research tool for ligand discovery*. Abstracts of Papers, 230th ACS National Meeting, Washington, DC, United States, Aug. 28-Sept. 1, 2005 (2005), CINF-023.
14. http://www.iupac.org/inchi/
15. Chemical 'Naming' Method Unveiled, *Chem. Eng. News*, 22 Aug 2005, volume 83 number 34 pp 39-40
16. Prasanna, M. D.; Vondrasek, Jiri; Wlodawer, Alexander; Bhat, T. N. Application of InChI to curate, index, and query 3-D structures. *Proteins* **2005**, *60(1)*, 1-4.

17. Coles Simon J.; Day Nick E.; Murray-Rust Peter.; Rzepa Henry S.; Zhang Yong Enhancement of the chemical semantic web through the use of InChI identifiers. *Org. Biomol. Chem.* **2005**, *3(10)*, 1832-4.
18. Heller, Stephen R.; Stein, Stephen E.; Tchekhovskoi, Dmitrii V. *InChI: Open access/open source and the IUPAC International Chemical Identifier.* Abstracts of Papers, 230th ACS National Meeting, Washington, DC, United States, Aug. 28-Sept. 1, 2005 (2005), CINF-060.
19. Sayle, Roger A.; Delany, John J., III. *Structure searching using SMILES and relational databases.* Abstracts of Papers, 222nd ACS National Meeting, Chicago, IL, United States, August 26-30, 2001 (2001), CINF-008.
20. Anderson, E.; Veith, G. D.; Weininger, D. SMILES (simplified molecular identification and line entry system): a line notation and computerized interpreter for chemical structures. Report (1987), (EPA/600/M-87/021; Order No. PB88-130034), 6 pp.
21. Weininger, David. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comp. Sci.* **1988**, *28(1)*, 31-6.
22. Liu, Xiang. Introduction of formats of MOPAC's input and output files and development of a file-managing system for MOPAC. *A. Gong. Dax. Xueb.* **2001**, *18(3)*, 239-241.
23. Stewart, James J. P. MOPAC: a semiempirical molecular orbital program. *J. Comput. Aided Mol. Des.* **1990**, *4(1)*, 1-105.
24. http://www.molinspiration.com
25. Grigorov, M.; Weber, J.; Tronchet, J. M. J.; Jefford, C. W.; Milhous, W. K.; Maric, D. A. QSAR Study of the Antimalarial Activity of Some Synthetic 1,2,4-Trioxanes *J. Chem. Inf. Sci.* **1997**, *37*, 124.
26. Katritzky, Alan R.; Kulshyn, Oleksandr V.; Stoyanova-Slavova, Iva; Dobchev Dimitar A.; Kuanar, Minati; Fara, Dan C.; Karelson, Mati Antimalarial activity: A QSAR modeling using CODESSA PRO software. *Bioorg. Med. Chem.* **2006**, *14*, 2333-2357
27. Katritzky, Alan R.; Mu, Lan; Lobanov, Victor S. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400-10407

28. Karelson, Mati; Dobchev, Dimitar A.; Kulshyn, Oleksandr V.; Katritzky, Alan R. Neural Networks Convergence Using Physicochemical Data. *J. Chem. Inf. Model.* **2006**, *46*, 1892-1897

APPENDIX

Katritzky, Alan R.; Kulshyn, Oleksandr V.; Stoyanova-Slavova, Iva; Dobchev Dimitar A.; Kuanar, Minati; Fara, Dan C.; Karelson, Mati Antimalarial activity: A QSAR modeling using CODESSA PRO software. *Bioorg. Med. Chem.* **2006** *14*, 2333-2357