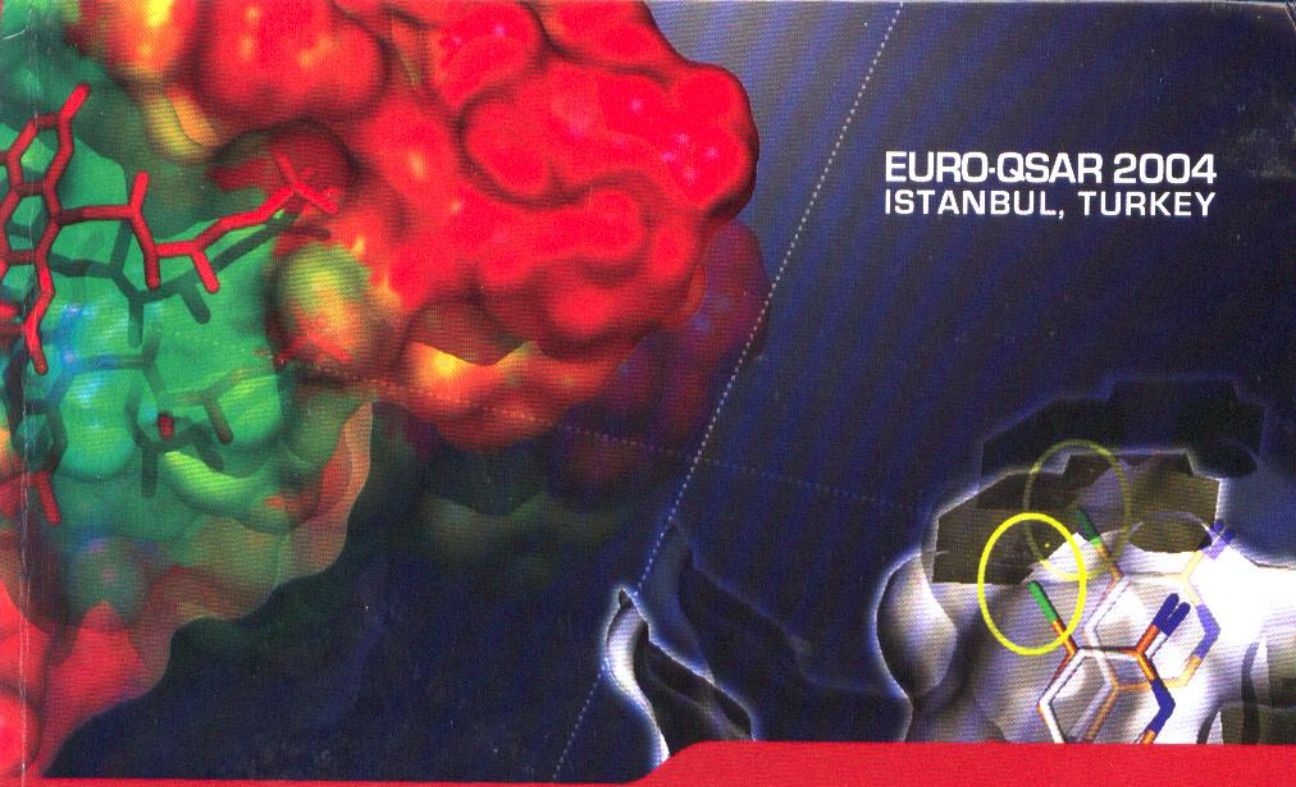

EURO-QSAR 2004
ISTANBUL, TURKEY

THE 15TH EUROPEAN SYMPOSIUM ON QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS & MOLECULAR MODELLING

# QSAR & Molecular Modelling in Rational Design of Bioactive Molecules

Editors:
Esin AKI (SENER)
Ismail YALCIN

# QSAR and Molecular Modelling in Rational Design of Bioactive Molecules

Proceedings of the 15th European Symposium on Structure-Activity Relationships (QSAR) and Molecular Modelling, Istanbul, September 05-10, 2004

Edited by

**Esin AKI (ŞENER)**

**Ismail YALÇIN**


COMPUTER AIDED DRUG DESIGN AND DEVELOPMENT SOCIETY IN TURKEY
ANKARA, TURKEY

*QSAR and Molecular Modelling in Rational Design of Bioactive Molecules:*
*Proceedings of the 15th European Symposium on Structure-Activity Relationships (QSAR) and Molecular Modelling, Istanbul, September 05-10.2004 / Eds by E. Akı (Şener) and I. Yalcin.*
Includes bibliographical references and index.

ISBN 975-00782-0-9 (hardbound)

Published by:
Computer Aided Drug Design & Development Society in Turkey
Ankara, Turkey
http://www.cadds.org



This book has been produced using printed, camera-ready manuscripts. Every effort has been made to reproduce the manuscripts as submitted.

The financial support of FARGEM, NOBEL İLAÇ is gratefully acknowledged.

Printed in Turkey by:
Başkent Klişe ve Matbaacılık
Phone: +90(312)4315490
Bayındır Sk. 30/E
06600 Kızılay, Ankara,
Turkey

*EuroQSAR 2004 Proceedings*, pp. 342-345
*QSAR and Molecular Modelling in Rational Design of Bioactive Molecules*
*Eds. E. Akı (Şener) and İ. Yalçın*

# *In silico* Design of New Compounds using Fragment Descriptors

Alexandre Varnek[1,*], Denis Fourches [1], Vitaly P. Solov'ev [2]

[1]*Laboratoire d'Infochimie, UMR 7551 CNRS, Université Louis Pasteur, 4, rue B. Pascal, 67000, Strasbourg, France,*
**e-mail : varnek@chimie.u-strasbg.fr*
*URL: http://infochim.u-strasbg.fr/recherche/index.html*
[2]*Institute of Physical Chemistry, Russ. Ac. Sci., 31 Leninskii prosp., 119991 Moscow, Russia*

## ISIDA System

"*In Silico*" **D**esign and data **A**nalysis (**ISIDA**) system has been developed to perform computer-aided design of new compounds using fragment descriptors. ISIDA represents an ensemble of structure-property tools including QSPR, clustering and combinatorial modules as well as the editors of 2D chemical structures and SD files. Fig. 1 illustrates links between these modules: information collected in the database is treated by the QSPR module which establishes relationships between structure of compounds and their properties. Then, structure-property models stored in the knowledgebase are applied to screen virtual combinatorial library. Below, we briefly describe these modules and give several examples of application of ISIDA for structure-property studies.

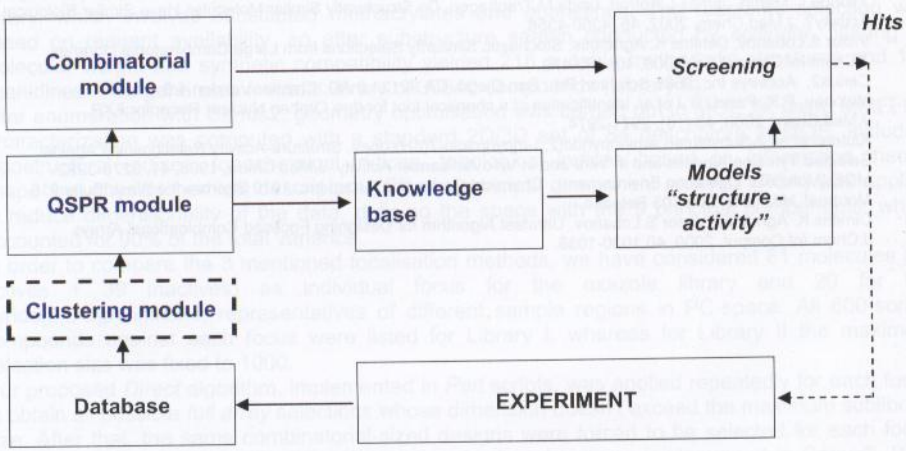


Fig. 1. Computer-aided design of new compounds using ISIDA system.

## Fragment Descriptors

Two different types of fragments are used as descriptors (Fig. 2): "sequences" (**I**) and "augmented atoms" (**II**). Three sub-types **AB**, **A** and **B** are defined for each class. For the fragments **I**, they represent sequences of atoms and bonds (**AB**), of atoms only (**A**), or of bonds only (**B**). Only shortest paths from one atom to the other are used. For each type of sequences, the minimal ($n_{min}$) and maximal ($n_{max}$) number of constituted atoms are defined. Thus, for the partitioning **I**(**AB**, $n_{min}$ - $n_{max}$), **I**(**A**, $n_{min}$ - $n_{max}$) and **I**(**B**, $n_{min}$ - $n_{max}$), the program generates "intermediate" sequences involving $n$ atoms ($n_{min} \le n \le n_{max}$). In the current version of the QSPR module, $n_{min} \ge 2$ and $n_{max} \le 6$. The number of sequences' types of different length corresponding to $n_{min} = 2$ and $n_{max} = 6$ is equal to 15 for each of three sub-types **AB**, **A** and **B**.

An "augmented atom" represents a selected atom with its environment including either neighbouring atoms and bonds (**AB**), or atoms only (**A**), or bonds only (**B**). Atomic hybridization (**Hy**) can be taken into account for augmented atoms of the **A**-type.

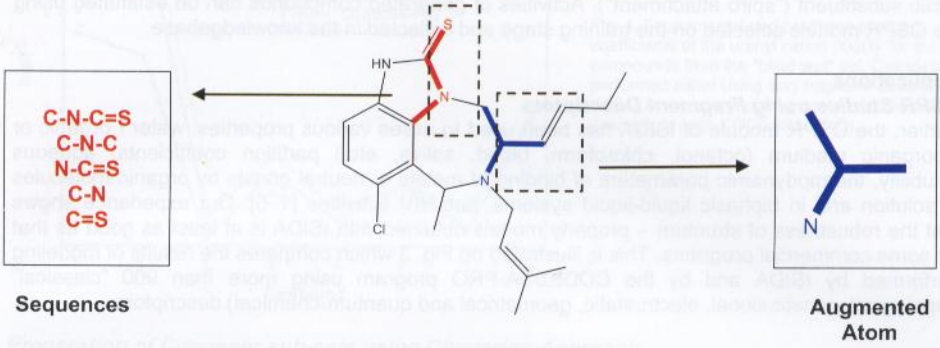


Fig. 2. Example of fragment descriptors: atom/bond sequences (left) and augmented atoms (right).

**QSPR Module**

Once a given molecular graph is represented by the ensemble of constitutive fragments (subgraphs), any of its quantitative physical or chemical property is calculated from the fragments contributions using linear or non-linear fitting equations. At the first step of the calculations, the QSPR module of ISIDA generates up to 196 models combining 49 fragmentation schemes with four linear and non-linear fitting equations. Successful models could then be selected according to statistical criteria. Optionally, any external descriptors can be combined with the fragment descriptors. On the test and screening stages, the program calculates the predicted values using the fitted fragments contributions obtained at the training stage.

The program takes into account the hybridization of atoms and distinguishes 9 different types of bonds: single, double, triple (in cycle or in chain), aromatic bonds and two types of coordination bonds. Optionally, the molecules containing the fragments of "rare occurrence" can be excluded from the training set; thus improving the statistical criteria of the models. If some fragment descriptors are linearly dependent, their linear combination forms a new variable. A procedure of reduction of the number of fragment variables according to *t*-test is implemented. Several QSPR models (instead of one model only) corresponding to different fragmentation schemes can be selected on the training stage. This allows user to calculate an average value of property, thus smoothing inaccuracies of particular individual models.

**Clustering Module**

The clustering module can be optionally used in order to split structurally diverse data set into several congener sub-sets more suitable for QSPR studies. Using fragment descriptors, the program builds either hierarchic or non- hierarchic clusters using Johnson and Jarvis-Patrick algorithms, respectively. The program can combine the both approaches, using first Jarvis-Patrick algorithm, then completing clustering with Johnson algorithm. The program uses different metrics to calculate the "distances" between molecules (Euclidian, Manhattan) or clusters (simple, complete and average links) as well as several normalization techniques. The property values for a set of molecules can be estimated using the *kNN* method.

**Knowledgebase**

The knowledgebase allows user to store selected models obtained with the QSPR module of ISIDA and to apply them for property assessment for any "external" set of molecules. For each model, the knowledgebase stores the type of structure-property equation, set of fragment descriptors and their contributions, statistical parameters of the models as well as chemical structures of the compounds from the training set. This program is based on server-client architecture.

**Combinatorial Module**

This module generates virtual combinatorial libraries using Markush structures. It uses its own editor of 2D structures allowing user to prepare the molecular core, to define the type of the

attachment "reaction", to select the attachment positions (atoms, bonds) and to prepare collections of substituents. Attachment of substituents to the molecular core can be performed by either connecting two atoms belonging to the two fragments ("atom-atom attachment"), by overlapping two bonds of these fragments ("bond-bond attachment") or by inclusion of an atom the core into a cyclic substituent ("spiro attachment"). Activities of generated compounds can be estimated using the QSPR models selected on the training stage and collected in the knowledgebase.

## Applications

### *QSPR Studies using Fragment Descriptors*

Earlier, the QSPR module of ISIDA has been used to asses various properties: water / organic or bioorganic medium (octanol, chloroform, blood, saliva, etc.) partition coefficients, aqueous solubility, thermodynamic parameters of binding of metals or neutral guests by organic molecules in solution and in biphasic liquid-liquid systems, anti-HIV activities [1–5]. Our experience shows that the robustness of structure – property models obtained with ISIDA is at least as good as that for some commercial programs. This is illustrated on Fig. 3 which compares the results of modeling performed by ISIDA and by the CODESSA-PRO program using more than 900 "classical" (topological, constitutional, electrostatic, geometrical and quantum-chemical) descriptors.

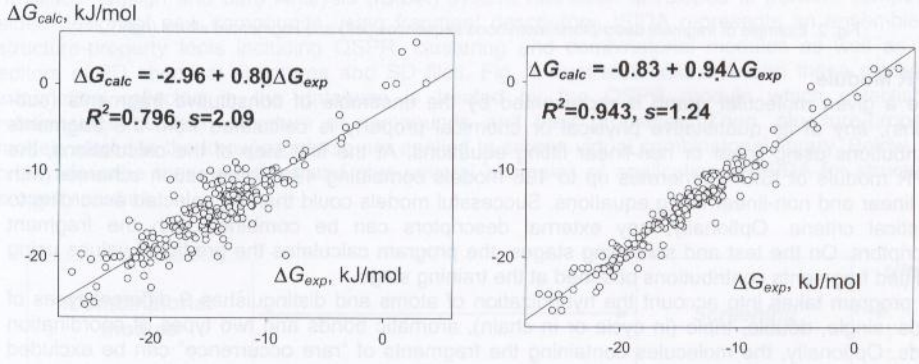


Fig. 3. Calculated vs experimental complexation free energies of β-cyclodextrin with neutral guests in water. Calculations were performed for the set of 218 compounds with CODESSA-PRO program using "classical" descriptors (left) and with ISIDA informational system using fragment descriptors (right).

### *Improvement of QSPR Models using Mixed Fragment/"Classical" Sets of Descriptors*

Mixing of fragment and "classical" descriptors is an interesting way to improve the robustness of the models. Thus, the modeling of water/octanol, water/chloroform, water/hexadecane, water/gas, octanol/gas, chloroform/gas and hexadecane/gas partition coefficients shows that the CODESSA-PRO models using molecular fragments from ISIDA as external descriptors correspond to better statistical parameters than the models using the "classical" descriptors.

### *"In Silico" Design of New Metal Binders*

The ensemble of ISIDA tools has been successfully used for computer-aided design of new extractants of $UO_2^{2+}$ cations in the biphasic water/dichloroethane system. First, the structure-property modeling has been performed on the set of 32 phosphoryl-containing podands (acyclic molecules with polyether spacer(s) linking 2 or 3 terminal phosphine oxide groups) for which the partition coefficients of the uranyl cation (log*D*) were available. Then a focused library of 2200 virtual molecules was prepared using the combinatorial model followed by the assessment of log*D* values using QSPR models stored in the knowledgebase. Selected from calculations 8 molecules were synthesized and experimentally studied [5] as uranyl extractants using the same protocol as in previous studies of the compounds from the parent set. Comparison of experimental and predicted log*D* values (Fig. 4) shows that the obtained QSPR models reasonably estimate log*D* for 7 from 8 compounds from the "blind test" set [5].

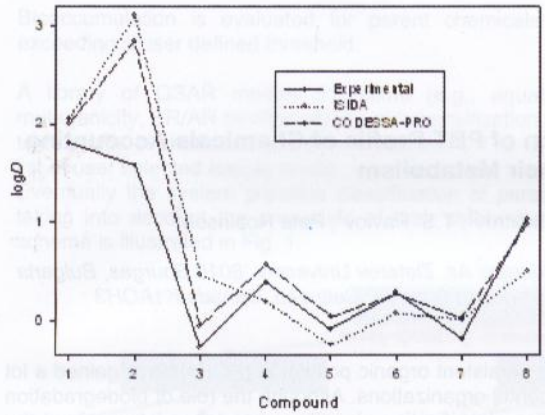


Fig. 4. Experimental and predicted partition coefficients of the uranyl cation (logD) for the eig compounds from the "blind test" set. Calculations performed either using only fragment descriptors ISIDA) or mixed set of fragment and "classical" descriptors (with CODESSA-PRO).

***Preparation of Congener sub-sets using Clustering Approach.***

The mixed non-hierarchic / hierarchic algorithm can be efficiently used to split a structurally diverse data set into several congener sub-sets more suitable for QSAR studies. Thus, the initial set on 1092 molecules for which the aqueous solubility data (logS) were available, was split into 3 clusters. The QSPR studies performed on these clusters led to models more robust than those obtained for the parent set: for the linear correlation $(logS)_{calc}$ *vs* $(logS)_{exp}$, $R^2$ = 0.930 - 0.952, $s$ = 0.37 - 0.48 for the clusters and $R^2$ = 0.917, $s$ = 0.70 for the entire set.

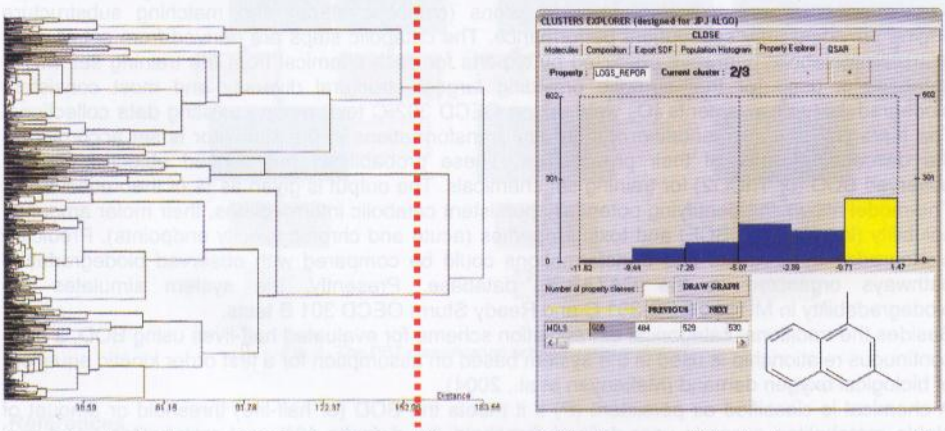


Fig. 5. Clustering the data set of 1092 molecules for which aqueous solubility data are available: dendrogram corresponding to hierarchic algorithm (left) and property explorer analysing content of a given cluster (right).

## References


[1] V. P. Solov'ev, A. Varnek, G. Wipff, Chem. Inf. Comp. Sci., 2000, 40 , 847-858.

[2] A. Varnek, G. Wipff, V. P. Solov'ev J. Solvent Extract. Ion. Exch., 2001, 19 (5), 791 - 837.
A. Varnek, G. Wipff, V. P. Solov'ev J. Chem. Inf. Comp. Sci., 2002, 42 , 812-829 .

[3] V. P. Solov'ev, A. Varnek J. Chem. Inf. Comp. Sci., 2003, 43 , 1703 – 1719

[4] A. R. Katritzky, D. C. Fara, H. Yang, M. Karelson, T. Suzuki, V. P. Solov'ev, A. Varnek, J. Chem. Inf. Comp. Sci., 2004, 44 , 529-541

[5] A. Varnek, D. Fourches, V. P. Solov'ev, V. E. Baulin, A. Turanov, V. Karandashev, A. R. Katritzky, D. Fara, J. Chem. Inf. Comp. Sci., 2004, 44, 1365-1382.

[6] CODESSA PRO project. http://www.codessa-pro.com/