INFORMATION SOCIETY TECHNOLOGIES
(IST)
PROGRAMME

OpenMolGRID

DESCRIPTION OF THE QUANTITATIVE STRUCTURE PROPERTY/ACTIVITY RELATION MODEL: MODEL BUILDING AND APPLICATION

Contract reference:	**IST-2001-37238**
Document identifier:	**OpenMolGRID-2-D2.4a-0102-1-1**
Date:	**08/10/2003**
Work package:	**WP2: Molecular Descriptor Generation and QSPR Model Building on the Grid**
Partner	**UT, UU, NEGRI, FZJ, CGX**
Lead Partner	**UT**
Document status:	**APPROVED**
Classification:	**PUBLIC**
Deliverable identifier:	**D2.4a**

Abstract: This document provides a brief description about the theoretical background of the quantitative structure property/activity relationship models.

Delivery Slip

	Name	Partner	Date
From	A Lomaka, U Maran	UT	18/09/2003
Verified by	S Sild	UT	08/10/2003
Approved by	GHF Diercksen (TC) R Ferenczi (QE)	OMC CGX	20/10/2003 07/04/2004

Document Log

Issue	Date	Comment	Author
0-0	12/09/2003	1st version	A Lomaka
0-1	15/09/2003	2nd stable version	U Maran
0-2	17/09/2003	3rd version	A Lomaka
0-3	18/09/2003	4th version	A Lomaka
1-0	08/10/2003	5th version	S Sild
1-1	16/04/2004	6th version	S Sild

Document Change Log

Issue	Item	Reason for Change
0-1	Updated and corrected chapters 1, 2, 3; added chapters 4 and 5;	Chapters required additional/new info and references were missing.
0-2	Minor changes	Incorporating review comments by FZJ
0-3	Minor changes	Incorporating review comments by UU
1-0	Minor corrections	Internal review process
1-1	Project information form	

Files

Files in this section relate to actual storage locations on the BSCW server located at https://hermes.chem.ut.ee/bscw/bscw.cgi. The URL below describes the location on BSCW from the root OpenMolGRID directory

Software Products	User files / URL
Word 2000/XP	OpenMolGRID/Workpackage 2/Deliverables/ OpenMolGRID-2-D2.4a-0102-1-1-ModelBuilding

Project information

Project acronym:	OpenMolGRID
Project full title:	Open Computing GRID for Molecular Science and Engineering
Proposal/Contract no.:	IST-2001-37238
European Commission:	
Project Officer:	Annalisa BOGLIOLO
Address:	European Commission - DG Information Society F2 - Grids for Complex Problem Solving B-1049 Brussels Belgium
Office	BU31 4/79
Phone:	+32 2 295 8131
Fax:	+32 2 299 1749
E-mail	annalisa.bogliolo@cec.eu.int
Project Coordinator:	Mathilde ROMBERG
Address:	Forschungszentrum Jülich GmbH ZAM D-52425 Jülich Germany
Phone:	+49 2461 61 3703
Fax:	+49 2461 61 6656
E-mail	m.romberg@fz-juelich.de

Contents

1. Introduction

1.1. Purpose and scope

The purpose of this document is to provide the general overview of the main components of a QSPR/QSAR study and their relationship.

1.2. Document Overview

The objective of the work package two (WP2) is for the adaptation of the existing quantitative structure-property/activity relationship (QSPR/QSAR) analysis software for the Grid environment.

The current deliverable provides an overview about the theoretical background of the QSPR/QSAR modelling and describes the sequence of necessary steps in the QSPR/QSAR analysis.

1.3. Document Structure

In addition to this section the document contains the following sections:

- Section 2 gives an overall description of the QSPR/QSAR modelling process
- Section 3 describes the individual steps of QSPR/QSAR study
- Section 4 application of the models
- Section 5 includes literature references

2. Overview of the QSPR/QSAR modelling

The high-level process flow of the QSPR/QSAR modelling in the Data Mining (DM) environment is summarised in ***Figure 1***. This process flow diagram outlines the two main tasks within this environment — the QSPR/QSAR model development and the deployment of developed QSPR/QSAR models. Both of these tasks involve cooperation between different software modules within the DM environment, such as quantum-chemical (QC) calculation, molecular descriptor calculation, QSPR/QSAR model development, and QSPR/QSAR prediction [1].

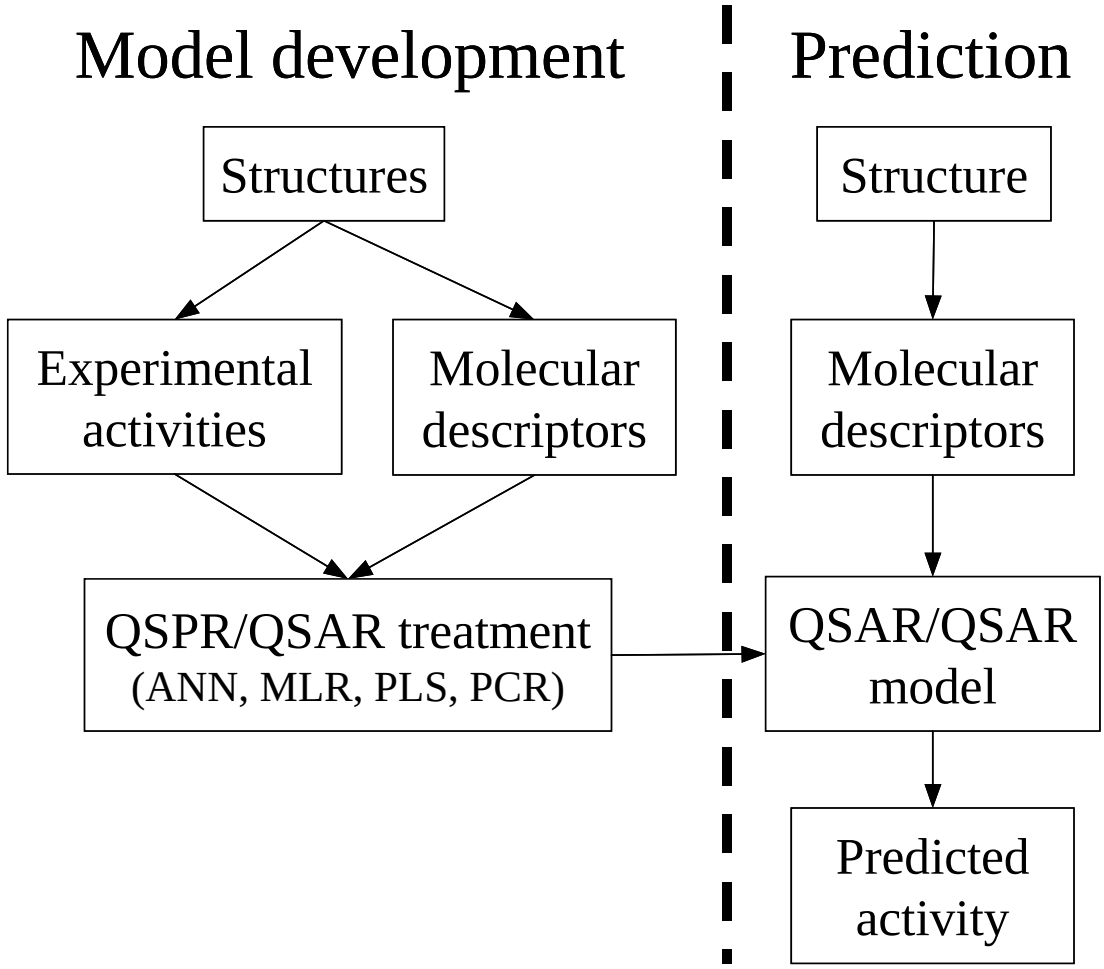

Figure 1: Overview of the QSPR/QSAR modelling.

The QSPR/QSAR model development process is considered well understood by the computational chemistry community. However, in practice it is still rather complicated task since it requires solid understanding in various scientific disciplines and involves combined application of multiple software programs [2]. The first step in the QSPR/QSAR model development is the preparation of a relevant data set, which contains a list of chemical structures and associated property or activity values. This task has to be carried out using the data warehouse tools, as the experimental data is often available from different sources and require therefore integration and pre-processing. The second step involves the calculation of the molecular descriptors for chemical structures. The last step is the application of statistical methods to derive QSPR/QSAR models for the investigated properties or activities. The detailed process flow for the QSPR/QSAR model development is depicted in the ***Figure 2***.

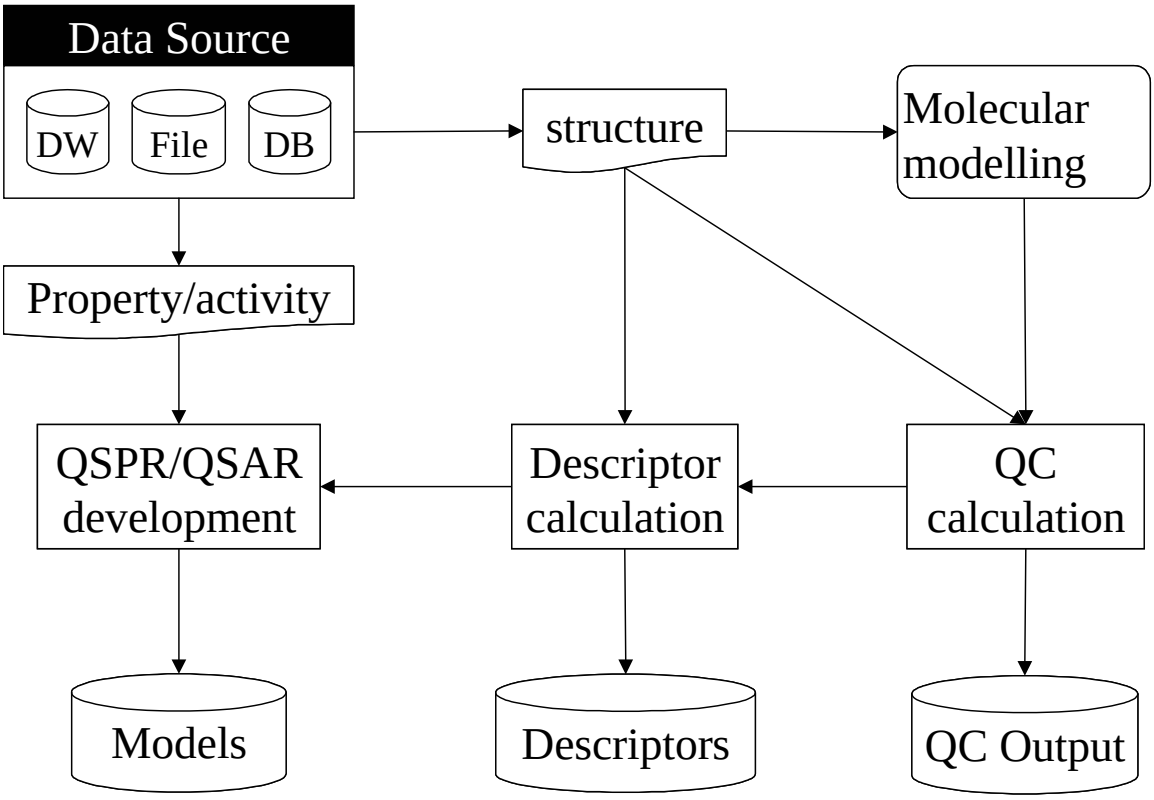

Figure 2: The QSPR/QSAR model development process.

The molecular modelling step in ***Figure 2*** generates the appropriate 3D representation(s) from a 2D structure that can be done automatically. However, this process is rather time consuming, because of the optimisation of the three-dimensional structure of molecules. This task will optionally involve computationally very intensive conformational search (cf. 3.3) combined with molecular mechanics calculations. It is desirable to store the calculated 3D structures in a database and load them later to the DW. This will avoid the most time consuming calculations in the future if this compound is used again later. In addition, the calculated molecular descriptors should be stored in a database, since usually more than one QSPR/QSAR method will be applied on the data set for the development of predictive models.

The last step in the QSPR/QSAR treatment is the development of the actual mathematical model. The input data set for the QSPR/QSAR modelling tools is generally a simple table with the values of dependent (property or activity) and independent (molecular descriptors) variables, and a set of control parameters for the model building tool.

The model development process is usually iterative, also a number of QSPR/QSAR models are generated using several different methods and parameters. The results are compared, analysed and the most predictive models are then selected for future deployment. The developed QSPR/QSAR models will be stored in a model database. This model database will be accessed from both the DM environment and the molecular engineering environment.

3. Steps in QSPR/QSAR

3.1. Pre-processing steps

Before the molecular descriptor values can be calculated, several pre-processing steps might be required, because chemical structure can be defined using different representations. These representations can be divided into following three classes:

- Chemical name and CAS number
- Two dimensional (2D) molecular structure
- Three dimensional (3D) molecular structure

The first kind of the structure representation cannot be used directly for the calculation of molecular descriptors and therefore it must be converted to 2D or 3D structure. The conversion from CAS number to 2D or 3D representation can be automated, if comprehensive database of CAS numbers is accessible. Similarly, the database can be used for the conversion of chemical name, but this is much more complicated, since one compound may have several different names [3]. In the ideal case, the molecular structures in the DW should contain information about the 2D or 3D structures.

3.2. Structure conversion from 2D to 3D

The conversion of a 2D molecular structure to a 3D representation (also known as '3D model building') is a common task in the QSPR/QSAR modelling. The 2D representation is very convenient to the end-user for sketching molecular structures and most chemical databases have only 2D representations available. However, all quantum chemical and most molecular descriptor calculation programs require the 3D representation of molecular structures as an input [4]. The calculation of the initial 3D coordinates from 2D coordinates or the equivalent connectivity information is performed by using standard bond lengths and hybridisation states of atoms.

3.3. Conformational analysis

Optionally, the 3D structure generated by 2D to 3D conversion step will be submitted to conformational analysis [5] since the guess structure (from 2D to 3D) may not be in the most optimal conformation. The need for conformational analysis is dependent on the flexibility of the compounds under study. In general, the lowest energy conformer is required for the QSPR/QSAR analysis, because it can be reproduced. It is desirable to store the calculated 3D structures in a database for reuse. This will avoid the most time consuming calculations in future if this compound is used again later.

3.4. Semi-empirical QC calculations

The semi-empirical QC [6] calculations in DM environment are needed for the calculation of quantum chemical descriptors. They expect guess molecular structures in 3D representation and require precise optimisation of chemical structures in order to provide correct description of structural characteristics. Some of the calculated characteristics of molecules are used directly as descriptors (e.g. dipole moment, heat of formation), while other output data will be used by the descriptor calculation module to calculate additional molecular descriptors (e.g. positively charged partial surface area, polarity parameter).

3.5. Calculation of Molecular Descriptors

Both the second and third class of structure representations can be directly used for the calculation of molecular descriptors. The descriptor calculation from 2D data does not require significant CPU resources. Thus, the Grid interaction in this part is not strictly required, unless we are dealing with a very large number (millions) of compounds. On the other hand, Grid integration is necessary in the

calculation of quantum-chemical (QC) descriptors, where 3D representations are required as input data for time-consuming QC calculations.

3.6. QSPR/QSAR model building

A variety of statistical structure-property correlation techniques can be used for the analysis of experimental data in combination with the calculated molecular descriptors. The following types of correlation models are most commonly used in QSAR/QSPR:

- Multilinear Regression Models (MLR) [7]
- Partial Least Squares (PLS) [8, 9]
- Principal Components Regression (PCR) [10]
- Artificial Neural Networks (ANN) [11]

Multilinear regression models can be obtained in the scalar space of the original descriptors. PCR and PLS apply in the principal-component orthogonalized descriptor space and in the target-transformed descriptor space, respectively. ANN, through its capability to model extremely complex nonlinear functions, is commonly used whenever linear approximation is unsufficient.

Several heuristic strategies [7] are available for descriptor selection in the effective search of the best (most informative) multiparameter correlations in the large space of the natural descriptors. The prediction capability of the model is judged by statistical parameters calculated for the model, various cross-validation techniques, internal and external validation sets.

4. Application of the models

The application of QSPR/QSAR model for the prediction is rather straightforward process. It involves following steps:

- The selection of developed QSPR/QSAR model from the database of the models.
- Following the above-described descriptor calculation step is performed for a compound under interest and the property or activity value is predicted by using selected QSPR/QSAR model. This step can be computationally very demanding, if predictions will be made to screen a huge virtual library or to validate a large number of candidate structures that were constructed in the molecular engineering environment.

Within current project the test application involves modelling and the prediction of cytotoxicity for the extensive library of compounds as described in the work package five (WP5) of the OpenMolGrid project. In practice, the scale of the application is very large involving various endpoints and characteristics in pharmaceutical and medicinal chemistry [4] and in chemical technology [12].

5. References

1. Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure, *Chem. Soc. Revs.* **1995**, *24*, 279-287.
2. Karelson, M. Molecular *Descriptors in QSAR/QSPR*. John Wiley & Sons, Inc, New York, USA, 2000.
3. IUPAC, Commission on Nomenclature of Organic Chemistry. *A Guide to IUPAC Nomenclature of Organic Compounds (Recommendations 1993)*, 1993, Blackwell Scientific Publications.
4. Katritzky, A. R.; Fara, D. C.; Petrukhin. R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Karelson, M. The Present Utility and Future Potential for Medicinal Chemistry of QSAR/QSPR with Whole Molecule Descriptors. *Curr. Top. Med. Chem.* **2002**, *2*, 1333-1356.
5. A.R. Leach, Molecular Modelling, Addison Wesley, 1997.
6. Stewart, J. J. P. Semiempirical Molecular Orbital Methods. In *Reviews in Computational Chemsitry*, 1990, vol. 1, 45-81
7. Draper, N. R.; Smith, H. *Applied Regression Analysis*, Wiley, New York, 1966
8. Hoskuldsson, A. PLS Regression Methods. *J. Chemometrics* **1988**, *2*, 211-228
9. Dunn, W. J.; Wold, S.; Edlund, U.; Hellberg, S. Multivariate Structure-Activity Relationships Between Data from a Battery of Biological Tests and an Ensemble of Structure Descriptors: The PLS Method. *QSAR*, **1984**, *3*, 131-137
10. Jolliffe, I. T. *Principal Component Analysis.* Springer-Verlag, New York, 1986.
11. Zupan, J. & Gasteiger, J. *Neural Networks for Chemists: An Introduction*. VCH, 1993.
12. Katritzky, A. R.; Maran, U., Lobanov, V. S.; Karelson, M. Structurally Diverse QSPR Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1-18.