My current topics in the field of QSPR

O. Skřehota
16.4.2012

Contents

Presentation contents

Motivation

Why is QSPR (QSAR) an interesting domain.

Why should we be interested?

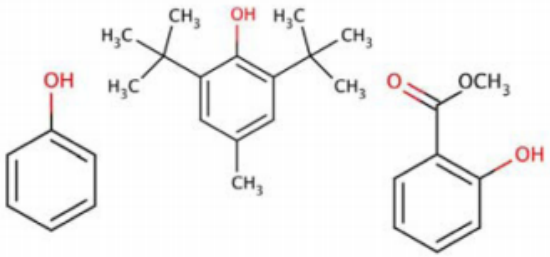
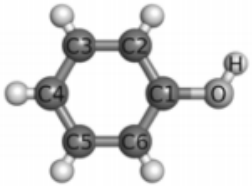
- Nowadays, a large amount of experimental and predicted data about the 3D structure of organic molecules and biomolecules is available.
  For drug discovery, it is very important to obtain physical and chemical properties of these molecules.

- It is very time consumming to measure the properties. Therefore approaches for their prediction are a topic of an intensive research. QSAR and QSPR models are very strong tools for predicting these properties.

Introduction into the QSPR

What is QSPR and why it is useful.

QSPR - Scheme

General scheme	Example	
Molecular structures	Phenols	
↓		
Descriptors	Charges for selected atoms	Theory levels (HF, MP2, B3LYP,…) Basis sets (STO-3G, 6-31G*, …) Population analysis (MPA, NPA, ESP, …) 
↓		
Suggestions of QSPR models	pK_a = f(charges)	$pK_a = p1_1.charge_H + p1_3$ $pK_a = p2_1.charge_H + p2_2.charge_O + p2_3$ …
↓		
Parameterization of QSPR models	MLR	$pK_a = -284.7.charge_H + 140.5$ $pK_a = -228.1.charge_H - 22.4.charge_O + 97.3$ …
↓		
Quality evaluation	Pearson coef. Root mean square error …	$R^2 = 0.965$ $RMSE = 0.424$ …

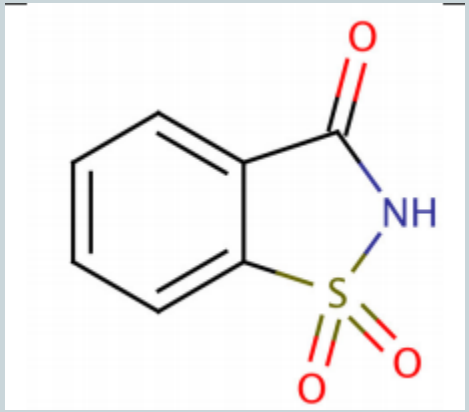
QSPR - Introduction

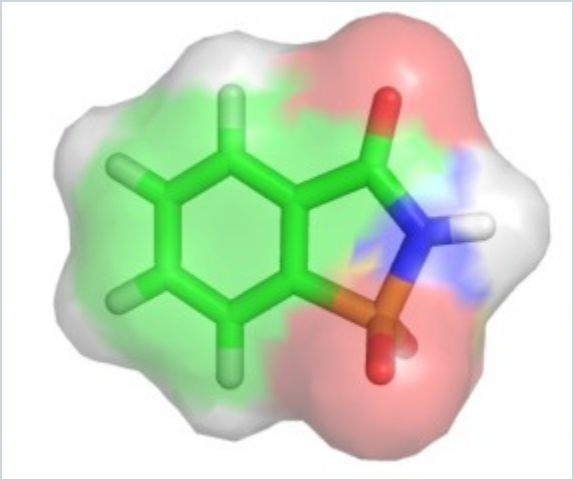
- Nowadays, a large amount of experimental and predicted data about the 3D structure of organic molecules and biomolecules is available.
- One of key methodologies for processing these data is **Quantitative Structure-Property Relationship** (QSPR) modeling.
- This methodology expresses molecules via various numerical values (called descriptors), which encode the structural characteristics of molecules.
- Afterwards the descriptors are employed to calculate the physicochemical properties of the molecules.
- QSPR provides an effective way to estimate physicochemical properties (e.g. dissociation constants, partition coefficients, solubility, lipophilicity, biological activity, . . .).
- The predecessors of QSPR models are the Quantitative Structure Activity Relationship (QSAR) models, which are focused on estimating of one particular property of a molecule – its biological activity.

Descriptors

- **Molecular descriptors**
  - Molecular descriptors are numerical values that characterise the properties of molecules.
  - Descriptors are frequently divided into 1D (types of atoms), 2D (bonding pattern), or 3D descriptors (spatial arangement of atoms).

$C_7H_5NO_3S$





Mathematical Methods Used in QSPR [1]

- **“Classical” methods:**
  - Multiple Linear Regression (MLR)
  - Partial Least Squares (PLS)
  - Neural Networks (NN)
  - Support Vector Machine (SVM)
- **“Emerging” methods:**
  - Gene Expression Programming (GEP)
  - Project Pursuit Regression (PPR)
  - Local Lazy Regression (LLR)

State of the Art

Currently available QSPR solutions

Currently Available Solutions

There are seveal solution currently available. Among the most well-known we count:

- MOE
- vLife QSARpro
- ChemBench
- OChem

Issues of Currently Available Solutions

Despite the fact, that software solutions available offer many interesting functionality, they do not provide sufficient extensibility. Each solution has one or several follwing issues:

- **Impossibility to define own descriptors**
  - Despite the fact, that a vast amount of descriptors are often supported “out of the box”, it is not possible to add own descriptors.
- **No means of creating own prediction methods**
  - It would be highly desirable to have a possibility to create own prediction methods, upload them into the application and use them directly there.
- **No possibility to create own quality indicators**
  - Even though a lot of indicators is often provided, own quality indicators could be an advantage.
- **Complicated installation or use**

Goal: QSPR Designer

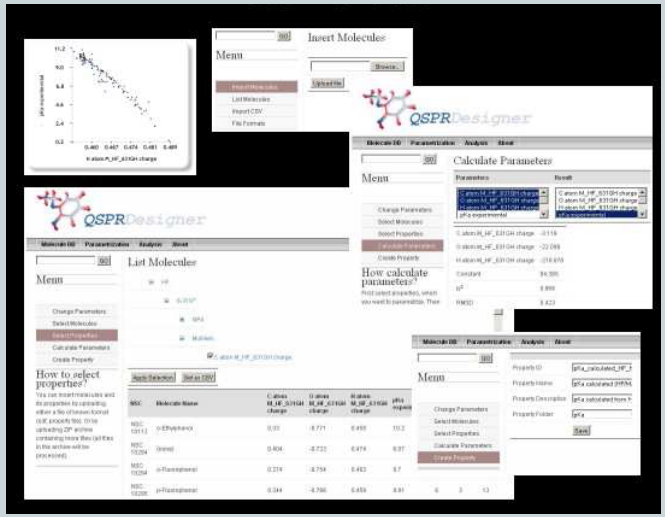
What is QSPR Designer, Descriptors, Calculation methods, Prediction and Quality Indicators

QSPR Designer - Introduction

- As already metioned, QSPR is relatively young approach. Many new molecular descriptors are being developed, new mathematical methods are being introduced and those already used are being improved.
- Currently available tools are in general not versatile enough to be able to react to this rapid development.
- QSPR Designer has been developed as versatile, yet easy to use application allowing user to develop and store molecular properties, predict their values and analyze them.
- It is a web application, therefore no installation is necessary.

QSPR Designer - Descriptors

- There are three types of descriptors available: stored (static), calculated (dynamic) and predicted.
- Value of dynamic descriptor can depend on values of other descriptors and is recalculated on demand.
- Each value of the descriptor can be related to molecule, atom in molecule or bond between atoms in molecule.



QSPR Designer - Prediction

- Prediction methods can be created programatically in the form of a plug-in (we do not support many methods out-of-box, but users can add their own).

- Multiple Linear Regression is supported out of the box.
- Simple interface to implement own prediction method.
- New prediction method must be uploaded to the application by the administrator.

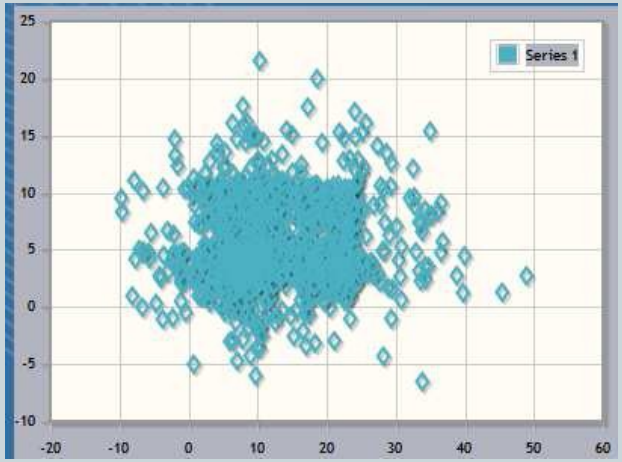
QSPR Designer – Quality Criteria

Currently supported are RMSE, RMSD and Pearson squared.

- Interface has been created in order to allow creation of other quality criteria in a form of pluggable modules.

QSPR Designer - Charting

- Charts are configurable and are created dynamically.



QSPR Designer – Current status & Future

- Currently new features are being tested and stabilized in order to be ready for the planned publication (see next slides). Release is planned for 1st week in May. This release should be robust enough to handle large amount of molecules.
- In the future we plan mainly to implement more mathematical methods for prediction and quality criteria.
- Releasing the application into the cloud could also be interesting – we would enable almost real-time (re)calculation of descriptor data for very large amount of molecules.

Publications

Current publications & conferences, Planned publications

Current publications & conferences

1. Skřehota, O. - Svobodová Vařeková, R. - Geidl, S. - Ionescu, C-M - Jan, Žídek - Koča, J. **QSPR Designer – Employ your own descriptors in the automated QSAR modeling process.** In 7th German Conference on Chemoinformatics. 2011.

2. Skřehota, O. - Svobodová Vařeková, R. - Geidl, S. - Kudera, M. - Sehnal, D. - Ionescu, C-M - Bouchal, T. - Koča, J. **QSPR modeling – algorithms, challenges and IT solutions.** In 9th Discussions in Structural Molecular Biology, Nové Hrady, Czech Republic. 2011.

3. Skřehota, O. - Svobodová Vařeková, R. - Geidl, S. - Kudera, M. - Sehnal, D. - Ionescu, C-M - Koča, J. **QSPR Designer - a program to design and evaluate QSPR models. Case study on pKa prediction.** In 6th German Conference on Chemoinformatics. 2010.

4. Geidl, S. - Beránek, Roman - Svobodová Vařeková, R. - Bouchal, T. - Brumovský, Miroslav - Kudera, M. - Skřehota, O. - Koča, J. **How the methodology of 3D structure preparation influences the quality of QSPR models?** In 7th German Conference on Chemoinformatics. 2011.

5. Geidl, S. - Svobodová Vařeková, R. - Skřehota, O. - Kudera, M. - Ionescu, C-M - Sehnal, D. - Bouchal, T. - Koča, J. **Predicting pKa values of substituted phenols from atomic charges.** In 9th Discussions in Structural Molecular Biology, Nové Hrady, Czech Republic. 2011.

6. Svobodová Vařeková, R. - Geidl, S. - Ionescu, C-M - Skřehota, O. - Kudera, M. - Sehnal, D. - Bouchal, T. - Abagyan, R. A. - Huber, H. J. - Koča, J. **Predicting pKa values of substituted phenols from atomic charges: Comparison of different quantum mechanical methods and charge distribution schemes.** Journal of Chemical Information and Modeling, 51, 8, od s. 1795–1806, 12 s. ISSN 1549-9596. 2011.

7. Ionescu, C-M - Svobodová Vařeková, R. - Raděj, T. - O., Skřehota - Koča, J. **Fast methods of atomic charge calculation: parameterization of EEM for applicability to proteins.** In 8th European Conference on Computational Chemistry. 2010.

8. Ionescu, C-M - Svobodová Vařeková, R. - Sehnal, D. - Skřehota, O. - Koča, J. **Fast methods of atomic charge calculation: the Electronegativity Equalization Method for proteins.** In 9th International Conference on Chemical Structures, Noordwijkerhout, The Netherlands. 2011.

Planned Publication

Article: **Different scenarios of pKa prediction implemented with QSPR Designer**
Authors: O. Skřehota, R. Svobodová, S. Geidl, C.M. Ionescu, J. Koča

Content:

- QSPR Designer & State of the Art in QSPR
- QSPR Designer
  - implementing own molecular descriptors
  - implementing own prediction methods
  - implementing own quality criteria used for analysis of results
- Case studies
  - Case 1: Inspect relation between QM charges and pKa of the molecule.
  - Case 2: Investigate the possibility of simple pKa prediction from molecules with similar charges.
  - Case 3: Employ MLR in order to predict pKa.

Planned submission: E05/2012

Questions

Thank you for your attention!