Stepwise D-Optimal design based on latent variables

Stefan Brandmaier,
Ullrika Sahlin, Tomas Öberg, Igor Tetko

Munich Interact,
Munich, 07.04.2011
What is experimental design?
Motivation

- **REACH legislation**: Each chemical compound produced in or imported into the EU in an amount of more than one ton has to be registered according to a number of endpoints.

- In case of hazardous, dangerous or toxic compounds, these endpoints contain toxicity and bio-accumulation.

- Experimental determination of all these values is often not possible, as experiments consume a lot of time, money - and in case of toxicity – life of animals!

- A valid approach to reduce experiments to a minimum is to test only a small subset of the compounds of interest and to build a reliable QSAR model from them.
QSAR / QSPR

- QSAR (Quantitative structure-activity relationship) modeling finds the quantitative correlation between molecular structures and a certain property.

- From molecule structure, so called descriptors are calculated (e.g. Molecular weight, number of benzene rings, energy)

- A machine learning algorithm is applied to these descriptors and calculates a model that can be used to predict the property for new compounds
Experimental design

- Given 600 compounds of interest and the limitation to test only 100 of them, the task is to find a 'good' subset to build a model from.
- But what does 'good' mean??
  - Avoidance of irrelevant information (outliers)
  - Avoidance of redundant information
  - Selected compounds should be representative
Standard solutions
Standard Solution

- Descriptor calculation for a molecule set
- Multivariate characterization of the compounds using PCA
  - Removal of linear dependencies
  - Decrease in the number of 'descriptors'
- Selection of testing subset (e.g. with D-Optimal algorithm)
- Testing of compounds
- Model building with linear regression algorithm
Problems

- Different descriptors deliver different outliers, as they are grouping molecules only by certain aspects.
- Globally relevant descriptors might be irrelevant for local models.
- There is no guarantee that principal components correspond to or correlate with the property focused on.
- Principal components can display noise, as long as it has just a high variance.
- Principal components are not specific for a certain endpoint.
- D-Optimal Design works like an outlier detector.
Stepwise PLS-based strategy
PLS-based adaptive strategy I

- Because of restricted capacities, labs usually do not test all compounds in parallel but in a stepwise procedure.
- The information gathered in each step can be used to refine the selection of compounds.
- PLS combines linear regression with PCA.
- Correlation to the target property is taken in consideration.
- PLS delivers so-called 'latent variables' instead of principal components to build a new vector base.
- Removal or at least decrease of noise.
PLS-based adaptive strategy II

- Based on D-Optimal design selection algorithm
- Utilizing Partial Least Squares (PLS) techniques to retrieve latent variables
  1) Select an initial set of compounds with a traditional D-Optimal approach, based on principal components
  2) Build a PLS regression model on the tested data
  3) Use this model to calculate the latent variables for all compounds
  4) Expand the selected set by applying the D-Optimal approach to the latent variables instead of principal components
  5) Repeat steps 2) – 4) as often as required
Validation
## Datasets

<table>
<thead>
<tr>
<th>Endpoint</th>
<th>Instances</th>
<th>Structural restrictions</th>
<th>Intricacy of endpoint</th>
<th>Model quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogKOC Partition coefficient</td>
<td>668</td>
<td>no</td>
<td>medium</td>
<td>average</td>
</tr>
<tr>
<td>Boiling Point</td>
<td>699</td>
<td>muted</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>Density Mass per volume</td>
<td>142</td>
<td>yes</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>IGC50 Toxicity on T. pyriformis</td>
<td>1158</td>
<td>no</td>
<td>high</td>
<td>good</td>
</tr>
<tr>
<td>LC50 Toxicity on fathead minnow</td>
<td>579</td>
<td>no</td>
<td>high</td>
<td>average</td>
</tr>
</tbody>
</table>
Validation pipeline

- **Descriptor calculation**
  - AlogPS, Estate indices, ISIDA fragments (length 2-5)

- **100 splits on each dataset**
  - 70% of compounds as operative set on which the design is performed
  - 30% of compounds as external validation set

- **Comparison of the performance**
  - **D-Optimal vs. Stepwise PLS-Optimal**
  - Comparison for a range of 20 to 200 selected compounds
  - 20 new compounds selected in each PLS-Optimal step
  - Validation on the operative set, the operative set without selected compounds and the external validation set
Results
Results I
Results II

- Improvement of performance is **highly significant** ($p(h_0) < 0.001$, binomial test) concerning
  - RMSE (up to 18%)
  - Q2 and R2
- for
  - all tested endpoints
  - both external and internal validations
  - each size of the datasets
  - the full range from 5% to 25% selected points
- Models of equal performance can be created with only 50% of compounds
Anything more ???
CADASTER

- FP7-funded EU project
- Implementation of REACH legislation to register chemical compounds
- Risk assessment for chemicals belonging to four compound classes
- Nine institutes from seven countries
Modeling platform Ochem.eu

- Containing 245398 measured values from literature
- Implementation of many machine learning algorithms
- Calculation of 20 different descriptor sets
- Various filtering options
- Application of peer reviewed published models
- Flexible management for endpoints
- All experimental conditions can be added to a record (unique feature)
Take home message

- Principal components are not necessarily correlated with the target property
- Stepwise approaches can be used to iteratively refine the design
- The usage of PLS latent variables instead of principal components can significantly improve the performance of experimental design

Experimental design makes three creatures happy:
Acknowledgement

Thanks to:

Ullrika Sahlin
Tomas Öberg
Igor V. Tetko

Ochem Team - Software to build models