Stepwise D-Optimal design based on latent variables

Stefan Brandmaier[1], Ullrika Sahlin[2], Thomas Öberg[2], Igor V. Tetko[1]

[1]Helmholtz-Zentrum München für Umwelt und Gesundheit
[2]Linnaeus University, Kalmar
E-mail contact: stefan.brandmaier@helmholtz-muenchen.de

1. Introduction

In the course of REACH, each chemical compound produced in or imported into the EU in amount of more than 1 ton has to be registered according to a number of environmental endpoints, including bioaccumulation and toxicity. Experimental determination of these properties requires a high number of animal tests. Apart from ethical reasons, animal experiments are expensive and time consuming. Therefore, the number of these tests should be kept as small as possible. This can be achieved by testing only a small representative subset of compounds, using them to build QSAR models and predict the remaining compounds.

There are several standard approaches for the selection of diverse sets of compounds for model purposes, such as factorial or D-Optimal design. The D-optimal design selects compounds using principal component analysis (PCA) of molecular descriptors. The analysis is done in one step and does not take into account the target property. Therefore, the selected compounds may not be optimal for modelling of the given property. Moreover, most labs, e.g. because of restricted capacities, test compounds not in parallel but in a stepwise procedure. The question is whether there is a better strategy that could provide better selection of compounds by taking into consideration the target property and available data.

We introduce a stepwise Partial Least Squares D-Optimal approach (PLS-Optimal design) to iteratively refine the chemicals space for the compound selection. The new approach utilizes the D-Optimal design but instead of PCA components, it selects compounds based on the PLS latent variables. We show that models developed with compounds selected using the PLS-Optimal design have significantly higher performance compared to those selected with the traditional approach.

2. Materials and methods

Two approaches were implemented. Firstly, a traditional D-Optimal design, selecting all compounds in one step, and based on principal components was implemented as a reference method. Its implementation was according to literature specifications. Secondly, the stepwise approach was implemented, utilizing the D-Optimal approach but using PLS latent variables instead of principal properties. These latent variables were retrieved from a PLS model built on all compounds that were considered as to be already tested. For the initial PLS model, a set of compounds was selected by chance. For all following steps, all initially selected compounds and the compounds suggested in the previous steps were used for model development.

The method performances were compared using four datasets, including endpoints for bioconcentration, lethal concentration, inhibition growth concentration and, soil organic partition coefficient.

In order to compare PLS-Optimal and D-Optimal design, 100 subsets (training sets), each containing 70% data points, were randomly selected for each endpoint. The detained 30% of the compounds were used as respective validation sets. Each of the training subset was used for the experimental design. Both approaches were used to select fixed numbers of compounds, which ranged from 25 to 200.

In order to compare qualities of sets selected using both methods, the selected compounds were used to build PLS models. These models (150 models per fixed number of compounds and per experimental design approach) were applied to predict molecules from the respective validation set. The mean values of RMSE, Q2 and R2 of models calculated for the validation sets were then used to compare quality of experimental designs of PLS-Optimal and D-Optimal methods.

3. Results and discussion

The results for all tested endpoints demonstrated a higher accuracy of models developed using compounds selected with PLS-optimal design compared to the D-optimal design. For the sake of simplicity and because of space limitations, the results section will only focus on the analysis of the LogKOC set. The LogKOC dataset contained 668 compounds. The training subsets used for experimental design and model development included 468 randomly selected compounds while the validation of models were done on remaining 200 compounds.

A comparison of D-Optimal versus PLS-Optimal design was done with 50, 75, 100, 125, 150, 175 and 200 compounds. Fig.1 depicts these results for RMSE. Firstly, with increase of the number of molecules in the training sets, the RMSE for the prediction of the validation set decrease for both approaches. This result is expected since larger number of molecules allows developing better models. Secondly, within the range from 50 to 150 selected compounds, the models developed with molecules selected using the stepwise approach provide significantly lower RMSE (*p* < 0.05 from the direct method using the Binomial distribution and 100 trials) compared to those developed using molecules selected with D-optimal design. On average, RMSE calculated using the PLS-Optimal were lower for about 0.05 log units (8%) compared to those developed the traditional method. In a similar way R2 and Q2 were significantly higher for models developed using PLS-Optimal design. These results indicate that sets of molecules selected using proposed method have significantly higher quality compared to those selected with traditional D-Optimal design approach.

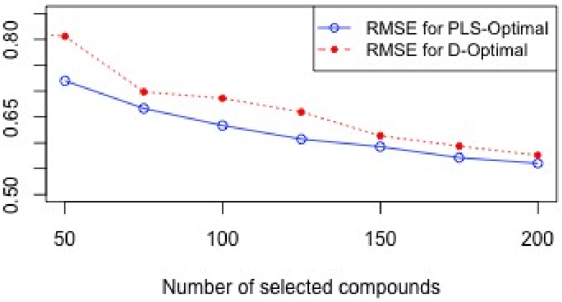

Figure 1: Comparison of performance of models calculated using both experimental designs for LogKOC model.

Although the models developed using 200 compounds, i.e. about 50% of the whole dataset, still had statistically significant better performances for PLS-Optimal, the difference in average RMSE was negligibly small. This result is not surprising, since models developed with these number of compounds reached saturation and approached RMSE value of models developed using full training subsets of 468 compounds. Thus the D-Optimal design provided better results compared to that of the D-Optimal design for 50-150 selected compounds, that is about 10%-30% training set compounds.

4. Conclusions

Our results show, that the performance of D-optimal experimental design can significantly be improved by taking into consideration the correlation between descriptors and property. The PLS-optimal design uses latent variables, which incorporates also information about the target property and descriptors. Thus it operates in the property-based space, contrary to the traditional method, which makes selection of molecules using only information about diversity of descriptors. The similar advantages of property-based space were demonstrated to assess accuracy of predictions for quantitative and qualitative models. The models developed using proposed PLS-optimal design provided significantly higher accuracy of prediction compared to the models developed using D-optimal design when using 10-30% training set compounds, the range that can be particular interesting for practical application.

Acknowledgement This study was supported with FP7 CADASTER project .