COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

Volume No: 7, Issue: 9, May 2013, e201305002, http://dx.doi.org/10.5936/csbj.201305002

CSBJ

# Robustness in experimental design: A study on the reliability of selection approaches

Stefan Brandmaier [a,*], Igor V Tetko [a,b,c]

**Abstract:** The quality criteria for experimental design approaches in chemoinformatics are numerous. Not only the error performance of a model resulting from the selected compounds is of importance, but also reliability, consistency, stability and robustness against small variations in the dataset or structurally diverse compounds. We developed a new stepwise, adaptive approach, DescRep, combining an iteratively refined descriptor selection with a sampling based on the putatively most representative compounds. A comparison of the proposed strategy was based on statistical performance of models derived from such a selection to those derived by other popular and frequently used approaches, such as the Kennard-Stone algorithm or the most descriptive compound selection. We used three datasets to carry out a statistical evaluation of the performance, reliability and robustness of the resulting models. Our results indicate that stepwise and adaptive approaches have a better adaptability to changes within a dataset and that this adaptability results in a better error performance and stability of the resulting models.

Research Article

## 1. Introduction

Experimental design techniques are crucial in terms of time and cost efficiency as well as to minimize the number of animal experiments. Reliable testing strategies are essential, especially in the course of the REACH legislation,[1] which includes the requirement that every chemical compound produced in/or imported into the European Union in an amount of more than one ton, has to be registered regarding a number of endpoints. But the application of selecting a representative and descriptive sub-sample from the chemical space of interest, and using it for the calculation of prediction models, is not only limited to risk assessment within REACH.[2],[3],[4] Also tasks as large scale scanning of chemical databases,[5] QSAR modeling,[6] drug target evaluation[7] or other pharmaceutical applications require systematic approaches to select representative subsamples.

The variety of concepts to address these problems in computational chemistry and QSAR modeling is widely spread,[8],[9] but most of them can be reduced to one of three basic ideas. Firstly, the selection of compounds with maximum dissimilarity, which is based on the theory that the most distinct compounds contain the most diverse information. This idea/theory is optimal for linear modeling. The D-Optimal criterion[10],[11],[12] and the Kennard-Stone algorithm[13] belong to this group of approaches. Secondly, the similarity selection aims to find compounds with high representativeness for the whole collection of relevant compounds. Approaches referring to this concept, e.g. the most descriptive compound selection (MDC),[14] usually select compounds from densely populated regions of the chemical space. Thirdly and lastly is an approach that aims to cover the whole chemical space of interest. The full factorial design[15] and space filling designs[16] are examples thereof. Recently, approaches that utilized hierarchical or density based clustering techniques were proposed.[9],[17] In our last study[18] we presented the advantages of an adaptive approach that combines a dissimilarity selection with an iteratively refined representation of the chemical space, by taking into consideration the information about the analyzed property that accumulates in the experimental process.

In QSAR modeling and chemoinformatics the focus within the evaluation of a novel approach is often exemplified on a particular dataset. Statistical evaluations, taking performance measures such as reliability and robustness of an approach into consideration are rare.[19] Due to chance correlations, this can result in misleading conclusions about the applicability of an approach. Furthermore stability, which is the ability of adapting small changes in a dataset, or to process structural outliers in a data collection, also needs to be taken into consideration. This is a quality criterion, which is as important as the performance itself.

In this study we present DescRep, a stepwise adaptive approach combining an iteratively refined descriptor selection with a sampling based on the concept of representative compounds. We compare this approach to experimental design strategies, which are commonly used in chemistry. An evaluation pipeline was implemented and applied to an ensemble of randomly selected subsets of three datasets, each with an endpoint relevant for REACH. We show that in comparison to the traditional approaches that select all compounds at the same time, DescRep performs significantly better.

We exemplify the importance of a statistical evaluation by investigating the effects of small changes in the underlying dataset on both the composition of the selected compounds and the performance of the resulting model. Furthermore, the collected datasets are extended with concerted structural outliers, to evaluate their influence on the selection approaches and the resulting models. Our results indicate that stepwise approaches, DescRep in particular, contribute to stability and reliability in experimental design.



[a]*Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Neuherberg D-85764, Germany*
[b]*Chemistry Department, Faculty of Science, King Abdulaziz University, P. O. Box 80203, Jeddah 21589, Saudi Arabia.*
[c]*eADMET GmbH, Ingolstaedter Landstrasse 1, Neuherberg D-85764, Germany*

* Corresponding author.
*E-mail address*: stefan.brandmaier@gmail.com (Stefan Brandmaier)

Supporting information for this article is available at http://www.csbj.org

OPEN ACCESS

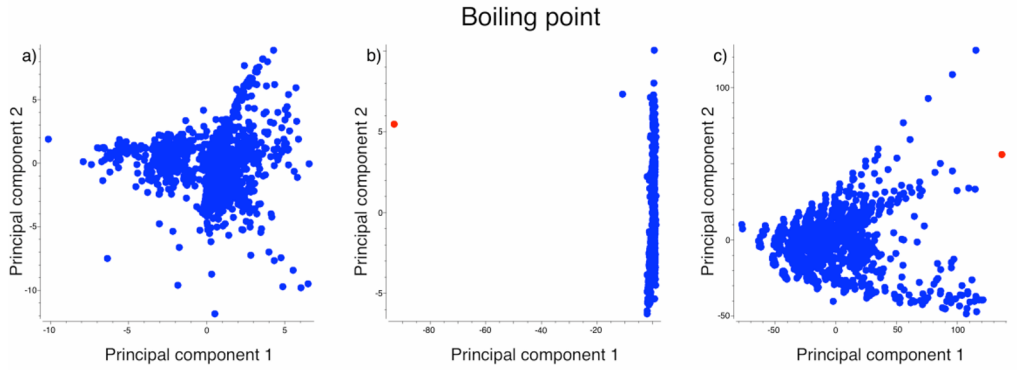


**Figure 1.** The change in the principal components view due to one structural outlier in the dataset. The principal components were calculated for the dataset with (b, c) and without (a) structural outlier. ALOGPS and E-State indices were used (a, b), as well as DRAGON descriptors (c). The protocol to calculate the principal components was always the same.

We investigate the benefits of a representation of the chemical space, which takes the correlation to the target property into consideration, and consequently arranges the compounds to a certain reference endpoint. Finally, we analyze our results with respect to the variability and adaptability of the examined approaches.

## 2. Materials and Methods

### *2.1. Materials*

To compare and evaluate the selection approaches, we collected three datasets, which vary in several criteria. The respective endpoints of these datasets, which were also used in our previous study,[18] were a physicochemical property, boiling point, a soil sorption coefficient, $\log K_{OC}$, and environmental aquatic toxicity against freshwater fish fathead minnow.

We extracted a collection of boiling point values from the Estimation Programs Interface (EPI) suite data.[20] The compounds within the dataset were restricted to halogenated compounds, containing bromine, chlorine and/or fluorine. As no further structural filters were applied, this set still provided a broad diversity regarding molecule size and chemical structures. We did not apply any kind of structural filter to the other datasets. The second dataset was based on the collection of $\log K_{OC}$ values by Meylan et al.[21] $\log K_{OC}$ is the log scale of the adsorption coefficient of a contaminant in the organic fraction of the soil. The endpoint for the toxicity dataset was the log scaled aquatic $LC_{50}$ value on the fathead minnow. The measurements were taken from the fathead minnow acute toxicity database[22] of the Environment Protection Agency (EPA).

All datasets were free of duplicate compounds. Measurements providing intervals of minimum or maximum values were excluded. In order to avoid problems in descriptor calculation, inorganic compounds, radicals, charged molecules and salts were filtered out. The final dataset for the boiling point contained 1198 compounds, the datasets for $\log K_{OC}$ and for toxicity on the fathead minnow contained 648 and 535 chemicals, respectively.

For each dataset, a collection of two types of descriptors was calculated. The first type was calculated using the ALOGPS 2.1 program[23] and contained two descriptors: solubility and lipophilicity of molecules. ALOGPS was the top-ranked model for prediction of logP.[24] The second type included E-State indices.[25],[26] These are electrotopological descriptors calculated for each atom and each bond in a compound and then summed according to their types over all atoms. The number of descriptors for the second type is determined by number of different chemical groups and thus it was not a fixed one. On our datasets, we calculated 179, 220 and 230 descriptors for $\log LC_{50}$, $\log K_{OC}$ and the boiling point dataset, respectively. The Online CHEmical database and Modeling environment (OCHEM)[27] was used for the calculation of the descriptors. To represent the chemical space of each dataset the descriptors were normalized to a [0,1] range. The rationale to use normalization instead of standardization is that standardization works on the underlying assumption that the objects are normally distributed. This assumption is not true for descriptors determined for chemical groups, e.g., in particular for the E-State indices. As they are linked to the presence of certain substructures, for most compounds, their value is just zero.

One of the aims of this study was to investigate the influence of structurally diverse compounds on the selection and accuracy of the resulting models. Therefore each of the three datasets was extended by the inclusion of a compound, which was characterized as a structural disrupter. We defined a structural disrupter as a data point that (a) influences the recalculated loadings of the first or the second principal component in such a manner that the principal properties represented by these components are changed and (b) results in one or more instances in the data set that are – according to the distribution of the instances in that principal component – at least five standard deviations from 97% of all other compounds.

Structural outliers like the ones used in this study are not artificial, but can result from several reasons, e.g. (a) from few compounds within the dataset, which have a specific chemical group that is different from other compounds and functionally is not relevant, (b) from the choice of a specific descriptor set, or (c) from a certain procedure within the multivariate analysis (centering or not the data, usage of raw, normalized or standardized data).

The structural outliers in our study were (a) ethyl 2-chloro-3-[2-chloro-5-[4-(difluoromethyl)-3-methyl-5-oxo-1,2,4-triazol-1-yl]-4-fluorophenyl]propanoate (carfentrazone-ethyl) for the boiling point dataset, (b) (1R,4aR,4bS,7S,10aR)-7-ethenyl-1,4a,7-trimethyl-3,4,4b,5,6,8,10,10a-octahydro-2H-phenanthrene-1-carboxylic acid (isopimaric acid) for the $\log LC_{50}$ dataset and (c) (1,2-dimethyl-3,5-diphenyl-pyrazol-1-yl) methyl sulfate for the $\log K_{OC}$ dataset. All these three compounds were retrieved from the same source as the rest of the respective dataset. Fig. 1a) shows the first two principal components of the boiling point dataset without outliers whereas Fig. 1b) shows the first principal components of the same dataset with the structural disruptor. The structural disrupter has a red color. The principal components were derived from the whole set of normalized ALOGPS descriptors and E-State indices and thus no variable selection was performed. Furthermore, the data were not centered before the orthogonal transformation.

Download English Version:

https://daneshyari.com/en/article/2079325

Download Persian Version:

https://daneshyari.com/article/2079325

Daneshyari.com