Déjà Vu - Réjà Vu : on knowledge-based approaches linking ligand and target information to bioactivity

Westen, G.J.P. van

Summary

This thesis focussed on the hypothesis that the combination of data from different research disciplines (here chemistry, biology and bioactivity) will have synergistic effects over methods focussing on a single discipline. We investigated this using several preclinical studies and one study using clinical data.

Chapter 1 introduced and defined common concepts from the world of computational chemistry (including: chemical space, chemical similarity, target space and target similarity). Furthermore the chapter highlights how '-informatics' based approaches have revolutionized the world of chemistry and biology. Similarly it contains a short introduction in structural methods, including the concept of X-ray crystallography. The chapter is concluded by a description of limitations in current methods and sketches why a need for novel approaches, like proteochemometrics, exists.

In **chapter 2** a review of the field of proteochemometrics was provided. We have provided a comprehensive overview of the concept and the full body of work using proteochemometric modelling until 2010 is shown in the primary table. This includes the data set modelled, the descriptors used and the machine learning technique applied. The chapter is concluded with a collection of possible pitfalls and a short outlook on novel machine learning approaches and application methods of proteochemometrics.

In **chapter 3** we introduced five novel descriptors to quantify target similarity and performed an extensive study of these and previously published amino acid descriptors. Amino acid similarity is quantified in a multi dimensional space where distance between amino acids correlates directly with similarity. Hence aromatic amino acids tend to cluster together as do charged amino acids. However this space is the result from a dimensionality reduction applied to a large input matrix and hence forms an approximation of the original input matrix. Therefore it was unknown which method would lead to a descriptor that performs optimal in proteochemometric modelling.

We concluded that the different descriptors all perform similar overall but large differences can occur for individual targets. Hence it is wise to sample different descriptors before embarking on final model training to achieve optimal performance. However we also observed that the inclusion of more information from the original input matrix affected performance of the descriptors in a negative way.

This is likely due to the nature of the data reduction approaches where the first factors tend to explain the majority of the variation in the dataset with subsequent factors explaining iteratively smaller fractions.

In **chapter 4** we performed a preclinical study with the goal to identify novel ligands for the human adenosine receptors. To obtain this goal we wanted to make optimal use of all data available to us in the public domain and combined experimental data obtained in bioassays incorporating human receptors with experimental data obtained on rat receptors. The combination of human and rat data resulted in a larger chemical space and hence our hypothesis was that this would translate to our model being able to identify novel ligands rather than analogues of existing compounds.

Of 54 compounds purchased, six novel high affinity adenosine receptor ligands were confirmed experimentally, one of which displayed an affinity of 7 nM on the human adenosine A_1 receptor. We concluded that our models perform better than current structure-activity modeling, as they were able to retrieve novel ligands (a low average tanimoto similarity to the training set) with a high hit rate (11 %)).

Another preclinical study forms the foundation for **chapter 5**, however here we targeted the application of proteochemometrics in a lead optimization project. Our models were trained on a data set, which was near complete (64 % of the possible compound – target interaction pairs had a pEC_{50} value). Through our model we could complete the missing 36 % with an accuracy that approached the assay accuracy, as we confirmed by a prospective experimental validation.

The high quality of this dataset allowed us to predict bioactivity spectra. In an antiviral drug discovery program, as modelled here, this allows for the selection of a compound that is not only active on the most frequently occurring mutant, but also on the majority of the other mutants. Hence the major contribution of our approach is that the optimal candidate can be selected without the need to perform all required 6,314 experiments. Finally, we demonstrated that we were able to define a model applicability domain based on target similarity.

In **chapter 6** we moved from preclinical studies to a clinical scenario. Here we used the largest dataset in literature to date that was subjected to proteochemometric modeling. The dataset consisted of thousands of unique HIV mutants (both protease and reverse transcriptase were included) on the target side and all clinically available drugs for these targets on the ligand side.

Our results demonstrated that we were able to create models capable of predicting a drug regimen for individual patients. Secondly we showed that PCM models performed better that sequence based approaches. Moreover, we demonstrated that our models are able to capture underlying relationships in the ligand – target space as we could predict the affinity of drugs on mutants not previously encountered. Additionally we could even predict the affinity of drugs on mixtures consisting of multiple unique mutants. Finally we confirmed the ability to define an applicability domain that can determine a reliability measure for each model prediction, an important feature in clinically applied models.

In **chapter 7** we applied a structure based approach rather than a machine learning method. We explored the use of novel techniques to mine the increasing amount of crystal structures available in the protein database. Using HIV reverse transcriptase as a case study we reached new insights on the dynamics of this protein and the changes induced by ligands binding to this target. Furthermore, using a three dimensional density based method deemed 'consensus structures' we identified novel features in a binding pocket that has been extensively studied since 1995. These features, currently unexploited by ligands, will improve the ability to inhibit HIV reverse transcriptase even in the presence of mutations.

Finally, in **chapter 8** we have drawn general conclusions from the thesis and proposed some future perspectives. The main hypothesis of this thesis was that linking information obtained from different disciplines (chemistry, biology, bioactivity) by computational approaches is synergistic. We have shown this to be true in the research chapters. However, the tools we used provided a framework, which relies on data from these disciplines to make predictions. Hence novel insights will lead to the ability to make novel predictions. An example is the case of drug target residence time. This concept is now actively being investigated and the results from these research programs will provide the foothold for the creation of possible residence time predicting models.

Furthermore, we proposed that any active drug discovery or research program should include a preliminary phase where all relevant data is gathered (even from related disciplines). Drawing conclusions in an organized fashion about chemistry active on related targets (which might be distantly related in the case of receptor deorphanization) or about the most efficient way to tune modeling parameters will improve the design of experiments. Hence a knowledge-based preliminary phase will help minimize the costs and time involved in the start of novel research projects.