PROJECT REVIEW

"Protein target prediction for identifying molecules with performance-enhancing potential"

J. Mitchell, H. Polaskova (University of St Andrews, UK)

In this project, we will develop protein target prediction software to allow the performance-enhancing potential of a molecule to be identified from its chemical structure. This will be made freely available to WADA and to national anti-doping agencies. A Web interface, suitable for intranet deployment, will be underpinned by state-of-the-art predictive software linking molecules to the protein targets against which they are active; both on-target and off-target activities are covered equally well.

We will use machine learning methods to identify substances with potential for use as doping agents. We will primarily employ the Random Forest algorithm, in which we have particular expertise and which has given us excellent results in previous studies. By means of hybrid descriptors combining the geometrical detail of Ultrafast Shape Recognition (UFS) with the chemical information provided by MACCS descriptors, we will encompass both key aspects of molecular recognition. We will use protein target prediction to obtain the on- and off-target bioactivities of molecules with known and unknown doping potential. The profile of activities across a representative panel of protein targets is the molecule's "bioactivity spectrum". We will use the bioactivity spectra of known performance-enhancing molecules to predict those of compounds whose performance-enhancing potential is unknown. Thus we will classify molecules, including both licensed pharmaceuticals and other drug-like compounds, as potentially performance-enhancing or otherwise.

While the use of illegal performance-enhancing substances continues to threaten both the integrity of sporting competition and the health of athletes, our software will allow early identification of potential doping molecules. These compounds can then be prioritised for experimental testing, while no further experiments need to be conducted on those with negative in silico predictions. The use of this computational technology will massively reduce the need for animal or human experiments.

"Protein target prediction for identifying molecules with performance-enhancing potential"

J. Mitchell, H. Polaskova (University of St Andrews, UK)

Results and Conclusion

During this two year project, we have worked on implementing a novel methodology to predict potential protein targets for performance enhancing molecules. During our project, we have designed and implemented a novel clustering algorithm (PFClust), used publicly available resources (ChEMBL, DrugBank and PubMed) and constructed a software standalone application.

In BMC Bioinformatics, 14:213 (2013), we presented the algorithm PFClust (Parameter Free Clustering), which is able automatically to cluster data and identify a suitable number of clusters to group them into without requiring any parameters to be specified. PFClust is heuristic in the sense that it cannot be described in terms of optimising any single simply-expressed metric over all possible clusterings. The algorithm partitions a dataset into a number of clusters sharing common attributes. Automatically determining the number of clusters present constitutes a significant challenge for clustering algorithms. Identifying the optimum number of clusters involves computing and evaluating a range of clusterings with different numbers of clusters. However, there is no agreed or unique definition of optimum in this context. Here, we tested PFClust on datasets for which an external gold standard of 'correct' cluster definitions exists. Results on synthetic datasets consisting of 2D vectors demonstrate that PFClust generates meaningful clusters, while our algorithm also showed excellent agreement with the correct assignments for three dimensional structures of protein domains, using a set of folds taken from the structural bioinformatics database CATH. We showed that PFClust is able to cluster the test datasets a little better, on average, than any of six other algorithms, even though five of the other methods are told in advance how many clusters to use.
It would be ideal to be able to identify all substances with one or more performance-enhancing pharmacological actions in an automated, fast and cost effective way. In Journal of Cheminformatics, 5:31 (2013), we use experimental data from the ChEMBL database (~7,000,000 activity records for 1,300,000 compounds) to build a model that takes into account both structure and experimental information, and use these data to predict both on-target and off-target interactions between these molecules and targets relevant to doping in sport. ChEMBL was screened and Ki, Kd, EC50, ED50, activity, potency, inhibition and IC50 were used for a rule-based filtering process to define quantitatively the labels "active" and "inactive". The "active" compounds for each ChEMBL family populated our bioactivity-based filtered families. A structure-based clustering step was subsequently performed with PFClust to split families with more than one distinct chemical scaffold. This produced refined families, whose members share both a common chemical scaffold and bioactivity against a common target in ChEMBL. We used the Parzen-Rosenblatt

machine learning approach to test whether compounds in ChEMBL can be correctly predicted to belong to their appropriate refined families. Validation using the refined families gave a significant increase in predictivity compared with the filtered or with the original families. Out of 61,660 queries in our Monte Carlo cross-validation, 41,300 (66.98%) had the parent family as the top prediction and 53,797 (87.25%) had the parent family in the top four hits. Having thus validated our approach, we used it to show that we could identify the ChEMBL protein targets associated with the WADA prohibited classes.

Finally the end result of the project as agreed in the contract was a software package. We have implemented "Predictor", which uses one of the most well-known and used bioactivity databases (ChEMBL). The software has been successfully delivered.