Toxicology and Applied Pharmacology 231 (2008) 225-234

Contents lists available at ScienceDirect

Toxicology and Applied Pharmacology

journal homepage: www.elsevier.com/locate/ytaap

Toxicological relationships between proteins obtained from protein target predictions of large toxicity databases

Florian Nigsch, John B.O. Mitchell *

Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

ARTICLE INFO

Article history:
Received 6 February 2008
Revised 3 April 2008
Accepted 5 May 2008
Available online 16 May 2008

Keywords:
Protein target prediction
Winnow algorithm
Computational toxicology
Toxicity

ABSTRACT

The combination of models for protein target prediction with large databases containing toxicological information for individual molecules allows the derivation of "toxiclogical" profiles, i.e., to what extent are molecules of known toxicity predicted to interact with a set of protein targets. To predict protein targets of drug-like and toxic molecules, we built a computational multiclass model using the Winnow algorithm based on a dataset of protein targets derived from the MDL Drug Data Report. A 15-fold Monte Carlo cross-validation using 50% of each class for training, and the remaining 50% for testing, provided an assessment of the accuracy of that model. We retained the 3 top-ranking predictions and found that in 82% of all cases the correct target was predicted within these three predictions. The first prediction was the correct one in almost 70% of cases. A model built on the whole protein target dataset was then used to predict the protein targets for 150 000 molecules from the MDL Toxicity Database. We analysed the frequency of the predictions across the panel of protein targets for experimentally determined toxicity classes of all molecules. This allowed us to identify clusters of proteins related by their toxicological profiles, as well as toxicities that are related. Literature-based evidence is provided for some specific clusters to show the relevance of the relationships identified.

Introduction

There are tangible reasons for the sustained interest in computational models for the prediction of protein targets of molecules. In pharmaceutical research it is of utmost importance to know as much as possible about the properties of a possible drug candidate as early as possible in the discovery/development process. The *in silico* prediction of protein targets allows the timely identification of potentially undesired off-target activities which are often related to adverse drug reactions (ADRs) observed in patients. But similarly, it also permits the determination of off-target activities which, instead of having to be regarded detrimental to the success of a compound, may actually be beneficial for its clinical use. The therapeutic outcome obtained by the simultaneous modulation of more than one target through the exploitation (and adaptation) of side effects is an approach that has been gaining interest in recent years (Morphy et al., 2004).

Knowledge about such off-target activities in addition to on-target activities can open up several routes to go down: 1) worst case scenario: the off-target activity is clinically incompatible with the primary activity, the compound is best discontinued; 2) not-too-bad scenario: the discovered off-target activity is serendipitously for a different target of similar clinical importance, so the compound may instead be engineered into a drug for this latter target; and 3) best case scenario: the activity profile of the compound hints at the possibility of allowing the modulation of more than one target at once, the compound may be designed into a multi-target drug (pharmacokinetics permitting). Not all off-target activities are necessarily bad, owing to their potential of being harnessed for increased therapeutic benefit. This is the case, for example, for the recent non-steroidal anti-inflammatory drugs (NSAIDs) which (try to) modulate cyclooxygenase-2 (COX-2) simultaneously with 5-lipoxygenase (5-LOX). Similarly, the lack of clinical efficacy of selective ligands has been realised in the case of antipsychotic drugs, where drugs with multiple endpoints are the current focus of research (Scripture and Figg, 2006).

As for other molecular properties of pharmaceutical importance, such as melting point, solubility and permeability, there is a considerable interest for computational models for target prediction (Nigsch et al., 2007). One of the first *in silico* models for the prediction of protein targets was developed by Lagunin et al. (2000) They used so-called "multilevel neighbourhoods of atoms" as descriptors in a model that ranks more than 300 pharmacological targets according to their relevance to the test molecule. In a leave-one-out cross-validation they achieved a mean accuracy of prediction of 89%. In a more recent application, they also used their methodology to draw inferences with respect to the molecular mechanism of action to explain certain toxic effects (Poroikov et al., 2007).

In 2005, Fliri et al. coined the term "biological activity spectrum" (Fliri et al., 2005). They defined the "biological activity spectrum" as the experimental activities of a given compound across a panel of

* Corresponding author. Fax: +44 1223 763 076.
E-mail address: jbom1@cam.ac.uk (J.B.O. Mitchell).

0041-008X/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.taap.2008.05.007

biological assays. They showed how this can be used for the clustering of biological activities, as well as for the prediction of untested compounds. Instead of experimentally determined activity spectra, Bender et al. used Bayes scores across a panel of target proteins ("Bayes affinity fingerprints") as descriptors for virtual screening experiments (Bender et al., 2006). A similar multi-class Bayesian approach was employed by Nidhi et al. to build a model for protein target prediction based on the WOMBAT (World of molecular bioactivity) chemogenomics database which they used to deconvolute therapeutic target annotations in the MDL Drug Data Report (MDDR) (Nidhi et al., 2006; Olah et al., 2004; Elsevier MDL, 2007). Steindl et al. developed a set of pharmacophore models for the simultaneous screening for multiple bioactivities (Steindl et al., 2006). Paolini et al. from Pfizer have undertaken an analysis of their large annotated in-house corporate dataset (Paolini et al., 2006). This analysis allowed them to establish relationships between and linkage maps of human pharmacological targets. Additionally, they used Bayesian statistics to build a predictive model. A different approach to interrelate human pharmacology targets was taken by Keiser et al. (2007). For all molecules belonging to a set of approximately 250 drug targets derived from the MDDR, they calculated all pairwise Tanimoto similarities. Incorporating a statistical correction for ligand sets of different sizes, they were able to link protein targets according to the similarity of their ligand sets. Other techniques to relate biological activities contained in the MDDR were reported by Sheridan and Shpungin (2004) as well as Schuffenhauer et al. (2003).

As it was pointed out above, potential off-target activities can either be desirable in the quest for multiple endpoint drugs, or be indicative of toxic effects associated with undesirable adverse drug reactions (ADRs). For time and cost reasons, there is substantial interest in accurate *in silico* models to predict such events. Bender et al. have recently presented a method that links biological activity information of molecules from WOMBAT with observed ADRs in patients, as obtained from the World Drug Index (WDI) (Bender et al., 2007). In the case of the discovery of desirable off-target effects, however, useful drugs may still be obtained. A concept called "selective optimization of side activities" (SOSA) can be applied to yield compounds that modulate other biological activities than the ones the molecule was originally intended/designed for (Wermuth, 2004). Numerous examples have been demonstrated in the literature (Morphy et al., 2004).

Models built on data about observed toxic effects lead to a field generally known as computational toxicology. This area of scientific research is about 20 years old, yet has only come to (a hesitant) fruition in the last couple of years. The main reasons for the slow development stem from the progressive integration of biological and chemical knowledge from different sources which has only begun to gain momentum within the last decade.

Apart from the pharmaceutical industry, regulatory bodies such as the European Medicines Agency (EMEA), US Food and Drug Administration (FDA) and (inter)national environmental protection agencies have a strong interest in the prediction of potential toxic effects. In the case of the agencies regulating the marketing of new drugs (e.g., EMEA or the FDA) possible benefits are, amongst others, a more efficient service preventing unnecessary review cycles. The rationale behind this is that if there are good computational models for the toxicities which are to be expected for a given class of compounds, new compounds may be approved more quickly. This would ultimately result in drugs reaching the market more rapidly (Benz, 2007). Various groups have taken different approaches to the problem of the prediction of toxic effects. There are models for specific receptors (Piotrowski et al., 2007, Vedani et al., 2007, Bhavani et al., 2006), as well as models indicating possibly toxic substructures (von Korff and Sander, 2006).

There are excellent articles discussing possible future developments, potential benefits and limitations in the field of computational toxicology to which we refer the reader for more information on this subject (Benz, 2007; Richard et al., 2006; Johnson and Rodgers, 2006).

The work that we present here is a combination of various aspects of what was mentioned above. First, we use a model for target prediction which was recently developed in our group. A first assessment of the underlying methodology showed that our method performs very well in the task of classifying molecules according to their biological activities (Nigsch and Mitchell, 2008). In previous work, however, we have used many fewer classes than in the current work where we use almost 250 different classes of biological activity. Second, our model for the prediction of protein targets is then used to predict the protein targets for molecules with experimentally determined toxicities. The combination of the known toxic effects of these molecules with their predicted protein targets—which may not always be known—allows us to infer relationships between toxicologically similar protein targets, as well as toxicities stemming from the modulation of a similar set of proteins. Overall, we use a method formerly not employed for the purposes of target prediction, and we use this method in order to relate protein targets according to their toxicological profiles.

The remainder of this article is organised as follows: In the next section (Methods), we will first outline the general strategy that will allow us to establish toxicological relationships between proteins, as well as the computational techniques that we employed. This is followed by a presentation of the Results that we obtained for the validation of the model, and the protein target predictions for toxic molecules. Subsequently, we will discuss these results and we show a specific example where our model identifies well the relationships of proteins implicated in breast cancer, even though such information was not explicitly included in the model. Finally, we will summarise and finish with our Conclusions.

Methods

Strategy. The work presented in this article relies on a model for protein target prediction to attribute toxic molecules to their most likely protein targets. To that end, we extended and used the computational framework that we presented in a recent paper (Nigsch and Mitchell, 2008).

Our target prediction model is built on a subset of the MDDR (see below), and we predict the biological activity profile of each molecule of the MDL Toxicity Database. This database contains experimentally determined toxicities for almost 150 000 molecules. As such, we obtain a toxicological profile (i.e., biological activity profile of a known toxic substance) for each toxic molecule. We do not intend to directly predict specific toxic effects for individual molecules from their structure. Instead, we use the predictions made to identify broader classes of toxicities which are related, as well as proteins which are related with respect to their toxicity profiles.

In order to do that, we only retained the most probable protein target for each molecule of experimentally determined toxic effect. In other words, we obtained a matrix R with all protein targets as columns, and the rows representing an experimentally observed toxicity. The individual entries of that matrix then represent how often a certain observed toxicity is associated with a specific protein, i.e., each matrix element r_{ij} specifies for how many molecules of toxicity i the highest score is found to be for class j. The columns of R then correspond to the toxicity profiles of the proteins included in the model, whereas the rows represent the toxicological profiles of the experimental toxicities. An analysis of the row-wise or column-wise correlation matrices permits the identification of relationships contained in the data: 1) relationships between toxicity classes (row-wise correlation matrix); and 2) relationships between protein targets according to the experimentally observed toxicities (column-wise correlation matrix). Additionally, a hierarchical clustering of the data can be used to identify clusters of proteins or toxicities which are closely related to each other.

We are aware of the fact that an analysis solely based on pairwise correlations of events has to be treated with care with respect to the drawing of any conclusions regarding the causality between these events. Nonetheless, in the analysis of the results (see Results) we found that the model provided concepts which have not been included *a priori* but corresponded to established facts found in the "real world". To a somewhat limited, yet appealing, degree, this constitutes a confirmation that our model is not producing meaningless information and that it can be used to reveal relationships between protein targets and different toxicity classes. Essentially, this is a statistical analysis *a posteriori* of the experimental data of approximately 90 000 drug-like and 150 000 toxic molecules.

Algorithm. The algorithm we used is known as Winnow, and works in the following way: each molecular feature i in every class c that the algorithm is aware of is associated with a weight $w_i^c > 0$. The score of a molecule in a certain class is the sum of the class-specific weights of the features present in that molecule; that permits a ranking of the predicted classes by their scores. By presenting a large number of training