Chapter 20
Learning Binding Affinity from Augmented High Throughput Screening Data

Nicos Angelopoulos
Edinburgh University, UK

Andreas Hadjiprocopis
Higher Technical Institute, Cyprus

Malcolm D. Walkinshaw
Edinburgh University, UK

ABSTRACT

In high throughput screening a large number of molecules are tested against a single target protein to determine binding affinity of each molecule to the target. The objective of such tests within the pharmaceutical industry is to identify potential drug-like lead molecules. Current technology allows for thousands of molecules to be tested inexpensively. The analysis of linking such biological data with molecular properties is thus becoming a major goal in both academic and pharmaceutical research. This chapter details how screening data can be augmented with high-dimensional descriptor data and how machine learning techniques can be utilised to build predictive models. The pyruvate kinase protein is used as a model target throughout the chapter. Binding affinity data from a public repository provide binding information on a large set of screened molecules. The authors consider three machine learning paradigms: Bayesian model averaging, Neural Networks, and Support Vector Machines. The authors apply algorithms from the three paradigms to three subsets of the data and comment on the relative merits of each. They also used the learnt models to classify the molecules in a large in-house molecular database that holds commercially available chemical structures from a large number of suppliers. They discuss the degree of agreement in compounds selected and ranked for three algorithms. Details of the technical challenges in such large scale classification and the ability of each paradigm to cope with these are put forward. The application of machine learning techniques to binding data augmented by high-dimensional can provide a powerful tool in compound testing. The emphasis of this work is on making very few assumptions or technical choices with regard to the machine learning techniques. This is to facilitate application of such techniques by non-experts.

DOI: 10.4018/978-1-4666-3604-0.ch020

INTRODUCTION

High throughput screening (HTS) is now a standard approach used in the pharmaceutical industry to identify potential drug-like lead molecules. Typically thousands, sometimes running up to a million, small molecule compounds are tested in a *binding* assay and affinity data is generated for each compound. More recently in a major initiative led by NCBI (the National Centre for Biotechnology Information, http://www.ncbi.nlm.nih.gov/) information on the results of more than 800 bioassays has been collated and disseminated. The analysis linking biological data with molecular properties is a major goal in both academic and pharmaceutical research.

One such assay is the protein pyruvate kinase (PYK), (Inglese et al., 2006). It is a potential anti-cancer and antiparasitic drug target. PYK acts as a tetramer and catalyses the last step in the breakdown of sugar to form pyruvate. This pathway is required for survival of the trypanasomatid parasites that cause diseases including sleeping sickness and leishmaniasis (Nowicki et al., 2008). Human PYK is also implicated in cancer pathogenesis and is crucial for the altered metabolism observed in tumour cells (Christofk et al., 2008). Inhibitors specific to the various pyruvate kinases are therefore of significant medical interest (Chan, Tan, & Sim, 2007).

We describe an approach for analysing this activity data in terms of *molecular descriptor* properties. This approach is based on *Bayesian model averaging* techniques. The results of the Bayesian approach are compared to that of two well-established machine learning techniques, namely Neural Networks and Support Vector Machines. A number of specific algorithms from the broad area of each technique are evaluated on one partition of the data. The best performing algorithm from each paradigm is selected and further, comprehensive comparison is carried out.

In order to realise our main objective of ligand discovery we retrained the best performing algorithms on all available known binder data and screened the 3.7 million unique compounds contained in the *Eduliss molecular database* for potentially active compounds.

BACKGROUND

Recent estimates suggest that there are about 3,000 possible druggable proteins (Zheng et al., 2006). These have structural features that allow the binding of small molecules. When bound, these adjust the biological function of their target. The object of ligand discovery is to identify molecules that will have a desired effect to the function of the protein.

Descriptor based approaches to ligand discovery have a long history in the field of QSAR (Hansch, Hoekman, & Gao, 1996). In early approaches simple models built by experts were proposed as predictors of *chemical activity.* The well known Lipinski *Rule of Fives* (Lipinski, Lombardo, Dominy, & Feeney, 2001) describes four properties (descriptors) common to the most successful *drug molecules*. The model depends on just four descriptors: solubility, molecular weight and the number of hydrogen bond donor and acceptor atoms in the molecule. There are however hundreds of characterising descriptors for every molecule (Todeschini, Consonni, Mannhold, Kubinyi, & Timmerman, 2000) covering calculated physical properties like polarizability or shape properties describing for example the presence of ring structures or particular chemical groups in the molecule. More modern approaches use statistical methods to build models based on large numbers of descriptors. In most cases preprocessing is used to do feature selection, that is to reduce the original number of descriptors to a subset that can be used to provide predictions which are approximately as good as those achieved by using the whole set of descriptors.

Descriptors such as the weighted holistic invariant molecular (WHIM) descriptors (To-