Chapter 2

Optimal Assignment Kernels for ADME in Silico Prediction

Holger Fröhlich
Bonn-Aachen International Center for IT (B-IT) 1, Germany

ABSTRACT

Prediction models for absorption, distribution, metabolic and excretion properties of chemical compounds play a crucial role in the drug discovery process. Often such models are derived via machine learning techniques. Kernel based learning algorithms, like the well known support vector machine (SVM) have gained a growing interest during the last years for this purpose. One of the key concepts of SVMs is a kernel function, which can be thought of as a special similarity measure. In this Chapter the author describes optimal assignment kernels for multi-labeled molecular graphs. The optimal assignment kernel is based on the idea of a maximal weighted bipartite matching of the atoms of a pair of molecules. At the same time the physico-chemical properties of each single atom are considered as well as the neighborhood in the molecular graph. Later on our similarity measure is extended to deal with reduced graph representations, in which certain structural elements, like rings, donors or acceptors, are condensed in one single node of the graph. Comparisons of the optimal assignment kernel with other graph kernels as well as with classical descriptor based models show a significant improvement in prediction accuracy.

INTRODUCTION

ADME in Silico Prediction

The development of a new drug is often compared with finding a needle in a haystack (Kubinyi, 2004). Therefore, rational approaches for drug design began to develop about 30 years ago with the aim to significantly reduce the amount of in vivo animal experiments. With the dramatic increase of computer performance during the last years there has been an increasing interest in virtual screening methods (HJ. Böhm & Schneider, 2000). The goal is to filter out a significant amount of “uninteresting” chemicals, that cannot be used as potential drugs, in silico in an early stage of
the drug discovery process. Thereby, especially the so-called ADME (Absorption, Distribution, Metabolism, Excretion) properties of a compound are of great interest (Kubinyi, 2002, 2003, 2004, Waterbeemd & Gifford, 2003): As most drugs are given orally for reasons of convenience, the compound is dissolved in the gastro-intestinal tract. It then has to be absorbed through the gut wall and pass the liver to get into the blood circulation. The percentage of the compound dose reaching the circulation is called the bioavailability. From there, the potential drug will have to be distributed to various tissues and organs in the body. The extend of distribution will depend on the structural and physico-chemical properties of the compound. For some drugs it will be further necessary to enter the central nervous system by crossing the blood-brain barrier. Finally, the chemical has to bind to its molecular target, for example, a receptor or ion channel, and exert its desired action.

The body will eventually try to eliminate a drug. Hence, for many drugs this requires metabolism or biotransformation. This takes place partly in the gut wall during absorption, but primarily in the liver. Traditionally, a distinction is made between phase I and phase II metabolism, although these do not necessarily occur sequentially. In phase I metabolism, a molecule is functionalized, for example, through oxidation, reduction or hydrolysis. In phase II metabolism, the functionalized compound is further transformed in so-called conjugation reactions, e.g. glucuronidation, sulfation or conjugation with glutathione.

The clearance of a drug from the body mainly takes place via the liver (hepatic clearance or metabolism, and biliary excretion) and the kidney (renal excretion). The half-life ($t_{1/2}$) of a compound is the time taken for its concentration in the blood plasma to be reduced by 50%. It is a function of the clearance and volume of distribution, and determines how often a drug needs to be administered.

QSPR (Quantitative Structure Property Relationship) methods try to predict in silico various ADME, but also physico-chemical properties, which have an important impact on a drug’s pharmacokinetic and metabolic fate in the body. Among others, today models for forecasting oral absorption, bioavailability, degree of blood-brain barrier penetration, clearance and volume of distribution are available. Additionally, there are methods for predicting physico-chemical properties, such as e.g. lipophilicity and water solubility (Waterbeemd & Gifford, 2003). Similarly, QSAR (Quantitative Structure Activity Relationship) methods are used to forecast the biological activity/inactivity of an untested ligand for a target protein (Kubinyi, 2002, 2003, 2004).

The basic assumption behind all QSAR/QSPR approaches is that the molecular properties in question can be derived from certain aspects of the molecular structure only. This implies that structurally similar compounds have similar biological or physico-chemical properties as well. In practice this supposition is often fulfilled, but there are also counter examples (Kubinyi, 2002, 2003).

Often, ADME models are derived via machine learning methods. Hence, one needs an abstract representation of a chemical compound in the computer. Classically, this is done by a large amount of descriptors (= features in machine learning language), which represent global molecular properties, like the polar surface area (Waterbeemd & Gifford, 2003), the distribution of certain physico-chemical properties, like the Radial Distribution Function (RDF) descriptor, the frequency of the occurrence of certain atomic patterns (fingerprints), invariances or characteristics of the molecular graph (topological indices) or others (Todeschini & Consonni, 2000). In conclusion, for each chemical compound one can calculate hundreds or even thousands of descriptors, which are of potential interest. The bottom line is that each molecule, which by itself is a complex three dimensional and dynamic object, is described in a simplified manner by a vector representation, which allows the easy use of classical machine learning algorithms.