Chapter 3

3D Ligand-Based Virtual Screening with Support Vector Machines

Jean-Philippe Vert
Mines ParisTech, Institut Curie and INSERM U900, France

ABSTRACT

The author reviews an approach, proposed recently by Mahé, Ralaivola, Stoven, and Vert (2006), for ligand-based virtual screening with support vector machines using a kernel based on the 3D structure of the molecules. The kernel detects putative 3-point pharmacophores, and generalizes previous approaches based on 3-point pharmacophore fingerprints. It overcomes the categorization issue associated with the discretization step usually required for the construction of fingerprints, and leads to promising results on several benchmark datasets.

INTRODUCTION

Computational models play an important role in early-stage drug discovery, in particular for lead identification and optimization. Starting from a list of molecules with experimentally determined binding affinity to a particular therapeutic target, as typically obtained by high-throughput screening (HTS), the goal of lead optimization is to find additional molecules with good binding affinity. The resulting leads are then further optimized, in particular to improve their pharmacokinetical and toxicological profiles, eventually leading to new candidate drugs.

Lead identification and optimization are usually performed by screening large databanks of small molecules, e.g., created by combinatorial chemistry, to find active molecules. Since experimental screening remains costly and time-consuming when large banks are concerned, and given the immensity of the space of small molecules which may be synthesized, in silico screening provides an interesting complementary approach to identify active molecules. An in silico screening is based on a model which can predict the activity of candidate molecules from their structure. Two general classes of models are often used. First, if the 3D structure of the target is known, then docking models predict whether
a small molecule can inhibit it by simulating its 3D structure and estimating the binding affinity of the protein-ligand complex. Docking involves difficult optimization problems to find the optimal 3D conformation of the molecule, binding configuration, and estimating its free energy. It is therefore rarely used on very large chemical databanks and is furthermore limited by the need to know in advance the 3D structure of the target.

A second common approach, sometimes used in parallel or as an alternative to docking, is ligand-based virtual screening. In that case an initial set of molecules with known binding affinity is used to build a predictive model that relates the structure of a molecule to its activity. The model can then be used to screen candidate molecules by ranking them in terms of their predicted activity by the model. This approach, often referred to as quantitative structure-activity relationship (QSAR), does not require the structure of the target, and usually results in computationally fast tools to predict the activity of candidate molecules. In this chapter we focus on this later, ligand-based virtual screening approach.

Ligand-based approaches usually involve statistical and machine learning procedure. Indeed, they can be formulated as the problem of estimating the attribute (activity) of patterns (molecules) given a set of patterns with known attributes. When the activity is considered as a real-valued attribute (e.g., free binding energy), we recognize a problem of regression, while when the attribute is categorical (e.g., active vs. not active) we are confronted with a problem of supervised binary classification, or pattern recognition. We note that, apart from ligand-based virtual screening, machine learning has been used to solve other problems in chemoinformatics such as toxicity or ADME in silico prediction, by simply modifying the definition of the attribute to be predicted. Both regression and pattern recognition have been much studied in statistics and machine learning, and a number of algorithms are available to attack them, ranging from linear models such as least-square regression, partial least-square (PLS) or linear discriminant analysis (LDA) to nonlinear methods such as neural networks, nearest neighbor or decision trees. A particularity of the ligand-based problem is that the patterns are molecules, while most algorithms for regression and pattern recognition work with vectors. Hence, in order to apply these algorithms to our problems, the molecules must first be converted to vectors of numerical features. This vectorization of the molecules turns out to be an important but difficult step. The problem of constructing numerical features, usually referred to as descriptors in chemoinformatics, remains one of the most debated and challenging issue in chemoinformatics. Indeed, a molecule is not easily and unambiguously described by a small set of numerical descriptors, and many descriptors have been proposed to describe various properties or features of the molecules (Todeschini & Consonni, 2002). Common descriptors include general properties of the molecules, such as the molecular weight, 2D descriptors with encode information about the 2D structure of the molecules, such as topological descriptors, hydrophobicity or substructure fingerprints, or 3D descriptors which capture geometric aspects of the molecule seen as a shape in the 3D space, such as quantum mechanical descriptors or shape indices (Figure 1).

The profusion of possible descriptors raises theoretical and practical issues. First, having many descriptors makes the dimension of the regression or pattern recognition problem large, which is often equivalent to making it hard to solve with a limited number of training examples. Second, computing and storing in memory many descriptors for large banks of molecules requires time and computational resources that may become limiting factors. Therefore, a common strategy in chemoinformatics is to select a limited number of “well-chosen” descriptors among the variety of possibilities. The criteria that should guide this selection for a given problem remain, however, largely debated.