Chemical Engineering Science 159 (2017) 31–42

Contents lists available at ScienceDirect

Chemical Engineering Science

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

# Developing an *in silico* pipeline for faster drug candidate discovery: Virtual high throughput screening with the Signature molecular descriptor using support vector machine models

Jonathan Jun Feng Chen [a], Donald Patrick Visco Jr. [b]

[a] Department of Biology, The University of Akron, 302 Buchtel Common, Akron, OH 44325, USA
[b] Department of Chemical and Biomolecular Engineering, The University of Akron, 302 Buchtel Common, Akron, OH 44325, USA

## HIGHLIGHTS

- Created SVM models using PCA as a filter and GA as a wrapper with the Signature molecular descriptor.
- Used Cathepsin-L as proof-of-concept for virtual high-throughput screening.
- Screened PubChem Compound Database and experimentally evaluated predicted inhibitors.
- First-pass through algorithm yielded a 19% hit rate.
- Second-pass through algorithm yielded a 75% hit rate.

## ARTICLE INFO

*Article history:*
Received 2 December 2015
Received in revised form
20 February 2016
Accepted 23 February 2016
Available online 2 March 2016

*Keywords:*
Virtual high throughput screening
QSAR
Drug discovery
CAMD
Signature

## ABSTRACT

Drug candidates make up a small portion of all possible compounds. To identify the candidates, traditional drug discovery methods like high-throughput screening test compound libraries against the target of interest. However, traditional high-throughput screening typically have a low efficiency, identifying $<1\%$ of the tested compounds as candidates and are costly because the majority of resources are spent testing compounds inactive towards a target of interest. To increase high-throughput screening efficiency, virtual high-throughput screening emerged as a way to focus compound libraries by removing unpromising drug candidates before bench-top testing is ever started. Virtual screens are usually based on energetics of a ligand-target complex, classification based on known ligands, or a combination of the two.

We propose a new ligand-based pipeline to reduce cost and increase efficiency: given a set of experimental data, the pipeline develops QSARs in the form of predictive SVM models and applies the models to virtually screen compound databases. The models obtained are based on a fragmental descriptor called Signature which has been previously shown as useful in virtual high-throughput screens.

For proof-of-concept, we used our pipeline to identify inhibitors for Cathepsin L, a receptor implicated in viral disease pathways. Our first pass virtual screen identified 16 compounds, 3 of which were experimentally confirmed as active, for a hit rate of 19%. Using the experimental data from the first-pass, we retrained the models to refine their predictive ability. Our second pass virtual screen identified 12 compounds, 9 of which experimentally confirmed as active, for a hit rate of 75%.



## 1. Introduction

As the human population and longevity increases, new pharmaceuticals are needed to treat the rise in communicable, age-related, and even drug-resistant illnesses. The process for procuring new pharmaceuticals is a long and difficult one and cannot keep up with demand as pharmaceuticals need to pass through rigorous drug trials, proving safety and efficacy, before they are made available to the public (Kapetanovic, 2008). But in order to start drug trials, drug leads need to be identified and there are many challenges in identifying those drug leads. The drug candidate chemical space is enormous, containing an conservative estimate of $10^{60}$ organic compounds (Bohacek et al., 1996). For a given target, only a small portion of the organic compounds are biologically

http://dx.doi.org/10.1016/j.ces.2016.02.037
0009-2509/

active and an even smaller portion constitute real leads (Dobson, 2004). If we consider all current and yet-to-be identified targets as one axis of a matrix and all organic compounds on the other, it would be an incomparably sparse matrix with few meaningful entries scattered throughout.

High-throughput screening (HTS) was the first method to systematically test and identify drug leads, experimentally evaluating a large numbers of compounds in a short period of time (Drews, 2000; Pereira and Williams, 2007; Triggle, 2007). Traditional HTSs are more effective than previous identification methods, testing more compounds than ever before, but are very costly, spending almost all resources testing compounds that are not active towards a target. While experts can preprocess compound libraries in the hopes of "focusing" the library, the lead discovery rate is still very small (Dobson, 2004). A better, more efficient method of identifying leads is needed.

While costly, traditional HTS do generate large amounts of data, coinciding with the advent of "big data" (Yan et al., 2006). Not only is more data generated than ever before, data is also shared more than ever before, resulting in libraries of both experimental (e.g. PubChem Bioassay (Wang et al., 2014) and CHEMBL (Gaulton et al., 2012)) and chemical (e.g. PubChem Compound (Kim et al., 2015) and ZINC (Irwin and Shoichet, 2005)) data. This data availability allows an alternate way of preprocessing a compound library. Instead of only using expert knowledge, it is now possible to build models and conduct HTS *virtually* to identify compounds that have a higher probability of being active. This alternate way increases efficiency since previous results are used to guide the search for new results, increasing the lead discovery rate and minimizing resources spent testing inactive compounds.

Current virtual HTS approaches fall under two categories: (1) molecular simulations, which include molecular dynamics and docking (e.g. AutoDock, DOCK, Flex, AMBER, GROMACS, CHARMM) (Cheng et al., 2012; Cornell et al., 1995; da Silva et al., 2010; Douguet et al., 2005; Durrant and McCammon, 2010, 2011; Kalyaanamoorthy and Chen, 2011; Lill and Danielson, 2011; Sinko et al., 2013; Wang et al., 2004; Wong and McCammon, 2003; Zeng and Wu, 2015) and (2) ligand-based screening, which vary in methods and metrics used (mainly classification and similarity) (Alvarsson et al., 2014a, 2014b; Bender et al., 2004; Zeng and Wu, 2015). It is noted that some researchers employ a mixture of both categories (Huang et al., 2015).

Molecular dynamics and docking simulations are numerical simulation experiments that look to find energetically-favorable conformations of ligand/substrate pairs through efficient exploration of the conformation space. The interaction energies of the system are calculated from thermodynamics models, but these approaches are computationally-expensive and are often dependent on starting conformations and/or multiple runs. A benefit of this approach, however, is that previous experimental information for known ligand/substrate pairs is not required.

Ligand-based screening approaches use known binding information for a given target (e.g. identities of ligands) and extrapolates that information to identify other possible ligands that bind with the substrate. Ligand-based screenings could be as simple as identifying compounds with similar structures to known ligands or as complex as classification based on a combination of chemical and structural properties. Ligand-based screenings can require less computational time compared to molecular simulations, but require binding information on ligand/substrate pairs.

A mixed approach can be used as well. Here, the number of compounds to be screened is reduced by the ligand-based approach and then these pairs are subjected to refinement using molecular simulations (dynamics or docking). While the mixture approach still requires previously obtained experimental data, it does shorten the computational time to obtain results. Note that such a mixed approach is highly-dependent on the pre-screen process: a useful pre-screen can enrich a compound library, while a poor pre-screen will miss promising compounds and spend time evaluating many inactive ones.

In our work, we use a ligand-based screen that combines three different approaches: similarity, classification, and regression. By using multiple methods, we maximize available data utilization. Our approach uses principal component analysis and the genetic algorithm to build support vector machine models specializing on predictive ability. These models are used to virtually screen large databases ($\sim$72 million compounds) quickly to create focused libraries of compounds predicted to be active. We use regression to predict quantitatively the activity of compounds as an additional way of discriminating among potentially active compounds.

The foundational core of our work is the Signature molecular descriptor. The Signature molecular descriptor, which is a fragmental descriptor, canonically describes the connectivity between atoms in a molecule in a tree-like fashion (Faulon et al., 2003b; Visco et al., 2002). It documents the structural features of a compound and allows their direct usage as variables in our models. Once compounds are fragmented into atomic Signatures, the fragments and their counts are used for the principal component analysis-genetic algorithm-support vector machine models for our method.

We tested our method by identifying previously unknown active compounds for the Cathepsin L inhibitor. Cathepsin-L is a receptor implicated in viral disease pathways, including malaria and Ebola (Grove and Marsh, 2011). Bioassay AID 825 assay from PubChem's Bioassay library contains experimental data on Cathepsin-L inhibitors. We used the $IC_{50}$ data found in PubChem's Bioassay database to train support vector machine models, screened PubChem's Compound database with our models, and ultimately identified and tested leads based on a confidence metric. As a loop-closing step, we included the newly obtained experimental data to the training dataset and retrained our models. We then rescreened PubChem's Compound database with this improved model and tested the new lead compounds to compare our new hit rate with our previous hit rate. All computations were performed on dual Intel Xeon processors (E5-2697, 2.7 GHz).

## 2. Materials and methods

### 2.1. Signature molecular descriptor

Molecular descriptors capture chemical or structural information in a quantitative fashion. Signature is a topological descriptor that captures the local connective environment of an atom, known as atomic Signatures, or a collection of atoms in a molecule, collectively known as a molecular Signature. The size of the environment captured is determined by a predefined distance or "height" (Faulon et al., 2003b). The Signature molecular descriptor was originally developed for structural elucidation studies (Faulon, 1994) but has since been used for solvent selection (Weis and Visco, 2010), substrate/inhibitor selection (Faulon et al., 2003a; Li et al., 2014; Weis et al., 2008), protein/protein interaction prediction (Martin et al., 2005), and molecular design (Chemmangattuvalappil and Eden, 2013; Chemmangattuvalappil et al., 2010; Churchwell et al., 2004; Dev et al., 2014; Weis et al., 2005; Weis and Visco, 2010). Signature is used in this work because of its history of creating QSAR models (Chemmangattuvalappil et al., 2010; Churchwell et al., 2004; Faulon et al., 2003a; Weis et al., 2005; Weis and Visco, 2010; Weis et al., 2008) and aiding virtual HTSs (Weis et al., 2008), it efficiently captures the structure and atomic connectivity of a compound, and it is amenable to molecular design. The process to obtain Signatures is illustrated in Fig. 1. For the purposes of our work, we used heights 0, 1, and