PRIORITIZING PROMISING COMPOUNDS IN VIRTUAL SCREENING CAMPAIGNS

Alexandre Victor Fassio, Rafaela Ferreira, Michael Keiser, Raquel Melo Minardi

UFMG

Abstract

Proteins are crucial macromolecules to all organisms as a whole, and countless diseases are associated with the proper functioning of proteins. Not surprisingly, proteins are the focus of numberless biological research whose focus is the development of new drugs able to modulate these macromolecules. However, the development and discovery of new lead compounds is a highly expensive and time-consuming program that takes up to 10-15 years. Thus, computational techniques like structural-based virtual screening (SBVS) and molecular docking contribute significantly to the early-stage drug discovery. A typical SBVS campaign consists of three major phases, namely the data preparation, the docking, and post-analysis. Commonly, a researcher starts with more than 20, 000 compounds, and after running a protocol of docking, 100-1000 candidate molecules remain to post-analysis. The latter is an essential procedure since scoring functions have several drawbacks and non-ligands might be prioritized first than true ligands, which is not desirable. Thus, the final step in SBVS strategies is a thorough manual process of hit selection, in which binding modes of hundreds of top-scoring compounds are inspected in molecular graphics programs. Nonetheless, the identification, prioritization, and automatic selection of promising HITs is still an open problem in the SBVS field.

Bearing this in mind, we propose MOTIF, a novel hashed interaction fingerprint that encodes interactions on molecular complexes both as binary or count fingerprint. Differently from other hashed fingerprints that are usually black-boxes, in which one has to design its own methods to interpret what each bit represents, MOTIF already provides several features to make the analysis straightforward and out-of-the-box.

Moreover, as an effort to validate and illustrate the applicability of MOTIF, we selected a recently published library consisting of 138 million molecules docked against Dopamine D4 as our case study. Herein, our goal was to train different machine learning models to reproduce Dock scores. In this scenario, we showed that MOTIF outperforms state of the art methods with an R-squared of 0.47.

Funding: