Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 26 (2008) 1237–1251
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Is it possible to increase hit rates in structure-based virtual screening by pharmacophore filtering? An investigation of the advantages and pitfalls of post-filtering

Daniel Muthas, Yogesh A. Sabnis, Magnus Lundborg, Anders Karlén *

Department of Medicinal Chemistry, Division of Organic Pharmaceutical Chemistry, BMC, Uppsala University, P.O. Box 574, SE-751 23 Uppsala, Sweden

Received 20 June 2007; received in revised form 16 November 2007; accepted 21 November 2007
Available online 18 January 2008

Abstract

We have investigated the influence of post-filtering virtual screening results, with pharmacophoric features generated from an X-ray structure, on enrichment rates. This was performed using three docking softwares, zdock+, Surflex and FRED, as virtual screening tools and pharmacophores generated in UNITY from co-crystallized complexes. Sets of known actives along with 9997 pharmaceutically relevant decoy compounds were docked against six chemically diverse protein targets namely CDK2, COX2, ERα, fXa, MMP3, and NA. To try to overcome the inherent limitations of the well-known docking problem, we generated multiple poses for each compound. The compounds were first ranked according to their scores alone and enrichment rates were calculated using only the top scoring pose of each compound. Subsequently, all poses for each compound were passed through the different pharmacophores generated from co-crystallized complexes and the enrichment factors were re-calculated based on the top-scoring passing pose of each compound. Post-filtering with a pharmacophore generated from only one X-ray complex was shown to increase enrichment rates in all investigated targets compared to docking alone. This indicates that this is a general method, which works for diverse targets and different docking softwares.

Keywords: Flo+; Surflex; FRED; Enrichment factor; Docking performance; Structure-based virtual screening; Pharmacophore filtering

1. Introduction

It is estimated that bringing a drug from idea to market takes approximately 12 years and costs as much as US$ 802 millions [1]. To cope with this high cost, more cost-efficient methods are required and various experimental and theoretical approaches have been developed. Virtual screening (VS) has arisen as an efficient method for rapidly identifying hits in terms of cost and time [2]. When the 3D-structure of the target is known, either by experimental or computational techniques, virtual screening is often performed by using structure-based docking [3]. With the advent of novel algorithms and faster computers it is now possible to screen millions of compounds in a matter of days. Virtual screening methods have been validated for their performance in several different studies (see for example Refs. [4–16]) and although these methods have been successfully used, they have some inherent limitations.

The so-called *docking problem* consists of correctly identifying the binding mode of a compound, i.e. finding the correct conformation and placement within the active site. The success of a docking is often compromised by the fact that the associated scoring functions often cannot resolve the most likely binding mode [17]. This highlights the importance of inspecting multiple conformations for the docked compounds and not only the highest scoring one. However, one can only visually inspect a much smaller number of compounds than the number of compounds usually contained in a screening library [18].

Database searching based on pharmacophore constraints is an alternative VS strategy [19–23]. One advantage of using pharmacophores is that they focus on specific key interactions for protein ligand binding. However, this approach does not perform optimally when used alone since little or no consideration of the shape of the binding site is taken into

* Corresponding author at: Department of Medicinal Chemistry, Uppsala University, Box 574, SE-751 23 Uppsala, Sweden. Tel.: +46 18 471 4293; fax: +46 18 471 4474.
E-mail address: anders.karlen@orgfarm.uu.se (A. Karlén).

1093-3263/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2007.11.005

account. Therefore, it is useful to have methods that make optimal use of both docking and pharmacophores to improve the selection of active molecules [21,24–28].

Docking and pharmacophores can be linked in different ways, e.g. by incorporating a pharmacophore constraint into the docking scheme, or by using pharmacophores as a post-filter of unconstrained docking poses [29]. A possible limitation with the introduction of pharmacophore constraints during the docking process is that if the specified features have to be adjusted, the docking has to be re-performed. In contrast, if the pharmacophore post-filtering approach is used, the features can easily be adjusted without re-performing the more time-consuming docking step. An additional benefit with pharmacophore post-filters is that they can easily be applied to results from any docking software.

The aim of this study is to combine the docking and pharmacophore-based approach in VS and to investigate the enrichment obtained after pharmacophoric post-filtering of docked poses and compared it to that after docking alone. We were interested to see if very limited prior information (one co-complexed ligand) could aid in identifying intrinsic binding features, and thereby help to retrieve a larger fraction of active molecules. The experiments were setup to simulate the situation faced by computational medicinal chemists early in a project where limited information about target–ligand interactions is known. To assess the generality of the approach we employed three different docking programs Flo+ [30], Surflex [31], and FRED [32] and six different data sets. We docked each database of actives and decoys to six pharmaceutically relevant targets retaining multiple poses. First, we calculated the enrichment rates using only the top-scoring pose of each compound for each target. Subsequently we passed all the poses of each compound through the pharmacophores generated from the co-crystallized complexes. After this process the enrichment rates were re-calculated based on the best scoring passing pose of each compound and compared to the enrichment rates obtained from docking alone (Fig. 1).

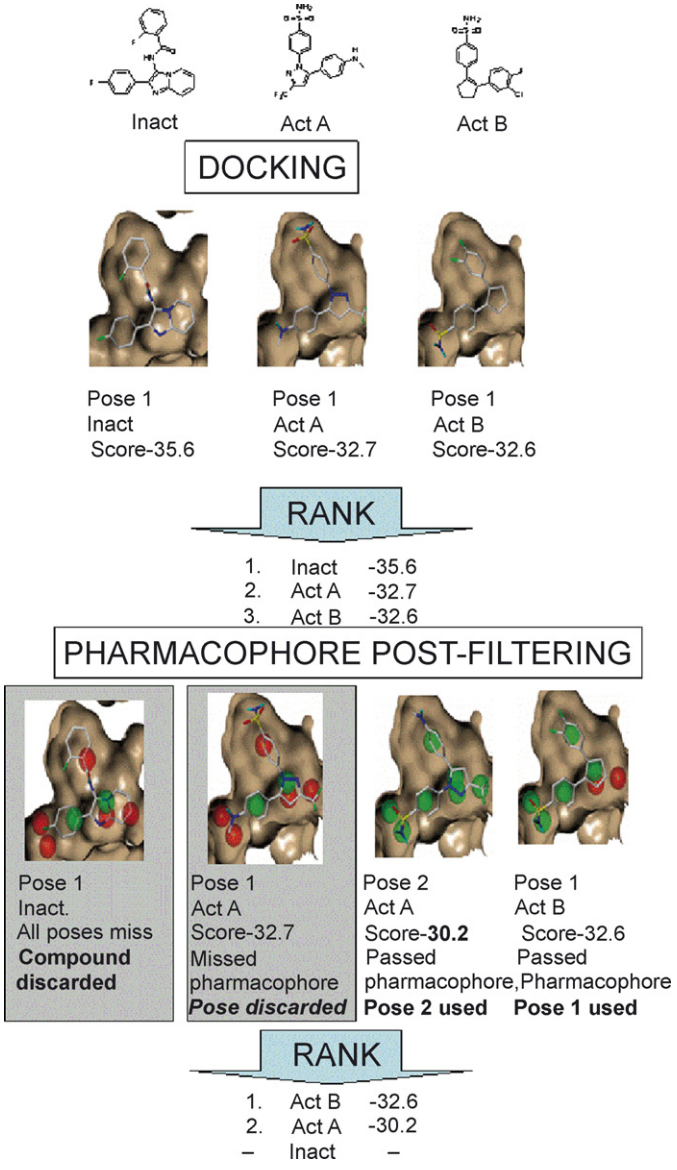

Fig. 1. Schematic representation of the approach using one decoy compound (Inact.) and two COX2 inhibitors (Act A and Act B). Compounds are first ranked according to scores alone and enrichment rates are calculated. All generated poses for all compounds are then passed through a pharmacophore filter. Those poses that do not fit the filter are discarded. The database is re-ranked and a new enrichment factor is calculated based on the best scored pose of each compound. A matched feature is colored green and a missed pharmacophore feature is colored red. Non-polar hydrogens are omitted for clarity.

2. Methods

2.1. Selection and preparation of protein complexes

The six reference protein complexes [33–38] used for docking were taken from the Brookhaven Protein DataBank [39] and are listed in Table 1. They were chosen to represent a variety of protein classes, differing in chemical characteristics and binding site shapes, encompassing both small-enclosed pockets such as ERα and large open pockets as in fXa. The selection of a particular crystal structure for docking among the available co-crystals was done by considering the resolution of the complex and the size of the co-crystallized ligand. Bound ligands, cofactors and metal ions were then removed except for the Zn^{2+} ion present in MMP3 (1CIZ), which has been implicated in ligand binding [37]. As the coordinates of the bound water molecules depend on the types of ligands in the active sites, all water molecules were removed from the protein complexes. Addition of hydrogen atoms and the definition of the protein active site were determined by the requirements of the respective programs and will be discussed later.

2.2. Dataset generation

The datasets used in this study has previously been compiled previously compiled by Jacobsson and Karlén [40], and consists of known actives (available in Supporting Information) taken from the literature [41–48] against each of the six proteins pooled with 10,000 decoy molecules. These decoys were selected from a database of commercially available compounds after having passed Lipinskís rule of 5 filter (allowing 1