Journal of Molecular Graphics and Modelling 32 (2012) 82–88

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Graphics and Modelling

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

Enrichment of virtual hits by progressive shape-matching and docking

Jiwon Choi, Ningning He, Nayoung Kim, Sukjoon Yoon*

Sookmyung Women's University, Department of Biological Sciences, Research Center for Women's Diseases, Hyochangwongil 52, Yongsan-gu, Seoul 140-742, Republic of Korea

ARTICLE INFO

Article history:
Received 15 July 2011
Received in revised form 1 October 2011
Accepted 3 October 2011
Available online 12 October 2011

Keywords:
Virtual screening
Lead hopping
Docking
Ligand-centric methods
Receptor-centric methods
Principal component analysis

ABSTRACT

The main applications of virtual chemical screening include the selection of a minimal receptor-relevant subset of a chemical library with a maximal chemical diversity. We have previously reported that the combination of ligand-centric and receptor-centric virtual screening methods may provide a compromise between computational time and accuracy during the hit enrichment process. In the present work, we propose a "progressive distributed docking" method that improves the virtual screening process using an iterative combination of shape-matching and docking steps. Known ligands with low docking scores were used as initial 3D templates for the shape comparisons with the chemical library. Next, new compounds with good template shape matches and low receptor docking scores were selected for the next round of shape searching and docking. The present iterative virtual screening process was tested for enriching Peroxisome proliferator-activated receptor and Phosphoinositide 3-kinase relevant compounds from a selected subset of the chemical libraries. It was demonstrated that the iterative combination improved the lead-hopping practice by improving the chemical diversity in the selected list of virtual hits.

1. Introduction

Fast, high-throughput virtual screening (HTVS) of chemical libraries has become a popular technique for the selection of a minimal number of compounds for experimental screening. HTVS also serves as an alternative to experimental high-throughput screening for the fast generation of lead compounds during drug discovery. The importance of HTVS is increasing simultaneously with the rapidly growing number of small molecules that are available in corporate and public compound libraries [1]. Despite the recent theoretical and technical improvements in virtual screening methods, there is still demand for the methodological development of more efficient and accurate virtual screening [2,3].

Virtual screening can be roughly divided into two categories: ligand-centric screening and receptor-centric screening. Ligand-centric methods essentially focus on the comparative analysis of the structural shape and chemical complementarity between compounds and known ligands. Therefore, knowledge of experimentally selected active compounds is a prerequisite for applying ligand-centric methods [4]. In contrast, receptor-centric methods predict the interaction of given compounds with a target receptor, which does not necessarily require experimental data on the active compounds. Molecular docking, which is a key method in receptor-centric virtual screening, is a technique that uses computers to predict the best binding mode of a given compound to a target receptor, resulting in a theoretically predicted binding affinity between the two molecules. Therefore, the docking step has become a primary component in many lead discovery practices [5–7].

It is likely that ligand-centric and receptor-centric virtual screening methods are complementary to one another. Thus, a well-tuned combination of these methods is required to maximize the effectiveness of virtual screening. In one of our previous studies, we presented a novel docking method that integrated ligand-centric and receptor-centric virtual screening methods to incorporate receptor flexibility with the time-efficiency of single conformation docking [8]. We elaborated on a distributed docking approach that took advantage of shape matching and multiple conformation docking methods, aiming to improve the activity of the hit enrichment in the selected list of virtual hits. The database compounds were classified in advance based on the shape similarities to one of the ligands that were complexed with the target protein in the available crystal structures. This classification enabled us to choose the appropriate receptor conformation for the single-receptor conformation docking of a given compound, thereby avoiding the time-consuming multiple docking approach. In the present study, we have improved the distributed docking by iterative combination of shape-search and docking methods.

Another important aspect of the application of virtual screening is lead hopping, which can be defined as the identification of isofunctional molecular structures with significantly different molecular backbones [9]. Lead hopping has become an important practice in the preparation of screening libraries and in maximizing the diversity of the chemical scaffolds in the potential hit list.

* Corresponding author. Tel.: +82 2 710 9415; fax: +82 2 2077 7322.
E-mail address: yoonsj@sookmyung.ac.kr (S. Yoon).

1093-3263/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2011.10.002

However, it has been recognized that 3D docking can be more effective than 2D topology search or 3D shape matching methods in lead hopping for the selection of a diverse subset of the chemical library [10–13]. Molecular docking is more efficient than shape-matching algorithms for enriching chemicals with diverse scaffolds in the list of virtual hits, although the computational speed is a critical issue in this case. Thus, one of our main goals has been to optimize virtual high-throughput screening approaches to improve the lead-hopping activity while searching for virtual hits. To this end, we designed a "progressive distributed docking" method that iteratively combined a shape-matching method with a multiple-receptor conformation docking method. This method was based on the idea that docking provides new receptor-relevant queries with diverse chemical structures for the next round of the shape search. The ultimate benefit of this method in virtual screening depends not only on the enriching virtual hits but also on the maximizing chemical diversity in selected compounds. In the present study, two targets, PPARγ (peroxisome proliferator-activated receptor γ) and PI3Kγ (phosphoinositide 3-kinase γ), were used for the evaluation of the proposed "progressive distributed docking" method. We comparatively analyzed the receptor relevance (docking score) and the chemical diversity of the selected virtual hits for the iterative and non-iterative methods.

2. Computational methods

2.1. Database construction

We prepared a set of diverse, drug-like compounds from commercially available chemical libraries retrieved from Zinc web site (http://zinc.docking.org). First, the molecules that contained any atoms other than H, C, N, O, F, S, Cl and Br were removed using FILTER (OpenEye, Inc.). Remaining compounds were further filtered to satisfy Lipinski's rules that excluded compounds with poor properties in druglikeness, such as AlogP, molecular weight (MW) and H-bond donor–acceptors (HBDA). The physicochemical properties were calculated using Pipeline Pilot (SciTegic Inc., San Diego, CA, USA). We used the default criteria defined in Pipeline Pilot for the filtering. The final size of compound library for the present analysis was ~1,500,000 non-redundant compounds. The compounds were converted from 2D to 3D SDF format using Ligprep, and the OMEGA program (OpenEye, Inc., v2.1.0) was used to generate multiple conformers for the library compounds. A maximum of 200 conformers was allowed for each compound based on a default RMSD cutoff (0.8 Å) and energy window cutoff (25 kcal/mol).

2.2. Ligand-centric virtual screening (shape comparison)

To make diverse compound subsets that were not biased to a specific region of the chemical space, we eliminated the neighboring compounds with a Tanimoto similarity of >0.5 at each step. The Tanimoto similarity was calculated using a Daylight fingerprint descriptor. Pairwise 3D shape similarities comparisons of the generated conformers against the crystal structure ligands were quantitatively calculated using the ROCS program (OpenEye, Inc., v2.2). This program performs shape-based overlays of the conformers of two molecules and compares a Gaussian-based overlap that is parameterized to reproduce hard-sphere volumes between a query molecule and a single conformation of a database molecule. The two scoring functions that are implemented in ROCS are the Shape Tanimoto score (the shape similarity) and the ColorScore (chemical pattern similarity). All of our experiments were performed with ROCS in color optimization mode. In this mode, ROCS is able to optimize the molecular overlay to maximize the molecular shape overlay and the chemical functionality overlap. ROCS provides two color force fields: the ImplicitMillsDean force field and the ExplicitMillsDean force field. To quantify the matching of the chemical functionalities, we used the ImplicitMillsDean color force field, which is a simple pK_a model that defines cations, anions, donors and acceptors. We then ranked the subset of compounds based on their combo scores (the sum of the ShapeTanimoto score for the overlay and the ScaledColor score), which ranged from 0 to 2, for which a value of 2 represented an exact match of the shape and functional groups to query the molecule and a value of 0 represented no 3D similarity. As a result, the top-scored 10,000, 5000 and 1000 molecules based on ROCS combo score from each three shape-based searches were prepared as each new subsets to obtain the overall recovery of the iterative distributed docking for PPARγ and PI3Kγ.

2.3. Receptor-centric virtual screening (molecular docking)

GLIDE (version 4.0) and ICM (version 3.3) with distinctive features in conformation search and scoring, were used for the docking studies of the receptors. For the molecular docking and scoring, PDB structures of PPARγ and PI3Kγ were prepared using the Schrödinger software package. All of the water molecules were removed, and the multimeric complexes were simplified from the PDB structures. If a PDB structure was missing side-chain atoms, Prime was used to predict their locations. Prior to the molecular docking, the receptor structures were preprocessed using protein preparation and refinement components in the GLIDE docking package. Hydrogen atoms were added using the all-atom force field. Side chains that were not close to the ligand binding site and that did not participate in salt bridges were neutralized. A restrained minimization was performed for the refinement of the complex structure using the OPLS-AA force field. This procedure reoriented the side-chain hydroxyl groups and alleviated potential steric clashes. This minimization continued until the average RMS deviation of the non-hydrogen atoms reached the specified limit of 0.3 Å. The GLIDE docking algorithm performed a series of hierarchical searches for the possible locations of the ligand in the binding site region of the receptor. The details of the GLIDE docking and scoring methods are described elsewhere. All of the compounds in the test sets were energy minimized using LigPrep and were then docked to the selected initial receptor structures using the standard mode of GLIDE docking (GLIDE SP 4.0). For the ICM docking, the initial structures were converted to ICM objects, and the grid maps were calculated with a grid spacing of 0.5 Å. The docking was performed with the default docking parameters. During the docking, either one of the torsional angles of the ligand was randomly changed or a pseudo-Brownian move was performed. Each random change was followed by 100 steps of local conjugate-gradient minimization. The new conformation was either accepted or rejected according to the Metropolis rule, using a temperature of 600 K. The number of Monte Carlo steps in the docking run and the length of the local minimization were determined automatically using an adaptive algorithm, depending on the size and number of the flexible torsions in a ligand.

2.4. Data analysis

To calculate the Tanimoto similarity, we chose a Laplacian-Modified Bayesian Classifier with Functional Connectivity Fingerprints (FCFP) that was implemented in Pipeline Pilot. The query templates for PPARγ and PI3Kγ were selected from this set using the SciTegic fingerprint-based clustering [14]. Cluster 3.0 software (developed based on Eisen Lab's Cluster and Tree View software) was used to generate the 3D queries templates for the initial library screening. The clusters were constructed with an intra-cluster radius of 0.5 (in SciTegic fingerprint tanimoto units), which