Methods 71 (2015) 146–157

Contents lists available at ScienceDirect

Methods

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

Benchmarking methods and data sets for ligand enrichment assessment in virtual screening

Jie Xia [a,b], Ermias Lemma Tilahun [b], Terry-Elinor Reid [b], Liangren Zhang [a,*], Xiang Simon Wang [b,*]

[a] State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences, Peking University, Beijing 100191, PR China
[b] Molecular Modeling and Drug Discovery Core for District of Columbia Developmental Center for AIDS Research (DC D-CFAR), Laboratory of Cheminformatics and Drug Design, Department of Pharmaceutical Sciences, College of Pharmacy, Howard University, Washington, DC 20059, USA

ARTICLE INFO

Article history:
Received 2 June 2014
Received in revised form 22 November 2014
Accepted 24 November 2014
Available online 3 December 2014

Keywords:
Benchmarking methodology
Decoy sets
Structure-based virtual screening
Ligand-based virtual screening
Artificial enrichment
Analogue bias

ABSTRACT

Retrospective small-scale virtual screening (VS) based on benchmarking data sets has been widely used to estimate ligand enrichments of VS approaches in the prospective (*i.e.* real-world) efforts. However, the intrinsic differences of benchmarking sets to the real screening chemical libraries can cause biased assessment. Herein, we summarize the history of benchmarking methods as well as data sets and highlight three main types of biases found in benchmarking sets, *i.e.* "analogue bias", "artificial enrichment" and "false negative". In addition, we introduce our recent algorithm to build maximum-unbiased benchmarking sets applicable to both ligand-based and structure-based VS approaches, and its implementations to three important human histone deacetylases (HDACs) isoforms, *i.e.* HDAC1, HDAC6 and HDAC8. The leave-one-out cross-validation (LOO CV) demonstrates that the benchmarking sets built by our algorithm are maximum-unbiased as measured by property matching, ROC curves and AUCs.

1. Introduction

Since the first seminal publication by Kuntz et al. [1], virtual screening (VS) has become an indispensable technique in the early-stage drug discovery to identify bioactive compounds against a specific target in a cost-effective and time-efficient manner [2]. A large collection of review-type literatures have discussed various VS approaches and provided perspectives of this technique [3–16,113]. In general, VS aims to filter out thousands of nonbinders *in silico* and ultimately to reduce the cost related to bioassay and chemical synthesis [9,17]. Depending on the availability of three-dimensional structures of biological targets, VS approaches are typically classified into structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS) [18]. The SBVS approaches, often referred to be molecular docking, employ the three-dimensional target structure to identify molecules that potentially bind to the target with appreciable affinity and specificity [10,16,19]. The latter is normally similarity-based, which identifies compounds of novel chemotypes but with similar activities by mining the information of known ligands [5,11,12,9,20–22].

To date, a wide variety of screening tools for both SBVS and LBVS have been developed [23–40]. Among them, DOCK [23], AutoDock [24], FlexX [25], Surflex [26], LigandFit [27], GOLD [28], Glide [29], ICM [30], and eHiTS [31] are popular tools for SBVS and updated regularly. For LBVS, QSAR modeling workflow [21] has been made publicly accessible to scientific communities by being incorporated into Chembench [32]. Catalyst [33], PHASE [34], and LigandScout [35] are classic algorithms for pharmacophore

Abbreviations: ACD, advanced chemical directory; AUC, area under curve; DEKOIS, demanding evaluation kits for objective *in silico* screening; DOE score, deviation from optimal embedding score; DUD, directory of useful decoys; DUD-E, DUD-enhanced; ECFP_4, extended-connectivity fingerprints of maximum diameter 4; Ed, Euclidean distance; ERα, estrogen receptor α; FC, formal charge; FCFP_6, function class fingerprints of maximum diameter 6; FL, final ligands; GDD, GPCR decoy database; GLIDA, GPCR-ligand database; GLL, GPCR ligand library; GPCR, G protein-coupled receptor; HBAs, hydrogen bond acceptors; HBDs, hydrogen bond donors; HDACs, histone deacetylases; LADS, latent actives in the decoy set; LBVS, ligand-based virtual screening; LOO CV, leave-one-out cross-validation; MDDR, MDL drug data report; ML, machine learning; MUV, maximum unbiased validation; MW, molecular weight; NRLiSt BDB, nuclear receptors ligands and structures benchmarking database; PCBioAssay, primary and confirmatory bioassays; *Psim*, property similarity; PSS, physicochemical similarity score; RBs, rotatable bonds; REPROVIS-DB, database of reproducible virtual screens; ROC, receiver operating characteristic; SBVS, structure-based virtual screening; *Tc*, Tanimoto coefficient; TK, thymidine kinase; *Tsim*, topological similarity; VDS, virtual decoy sets; VS, virtual screening; WOMBAT, world of molecular bioactivity.

* Corresponding authors at: State Key Laboratory of Natural and Biomimetic Drugs, Peking University School of Pharmaceutical Sciences, 38 Xueyuan Rd, Beijing 100191, PR China (L. Zhang). Howard University College of Pharmacy, 2300 4th St. NW, Washington, DC 20059, USA (X.S. Wang).

E-mail addresses: liangren@bjmu.edu.cn (L. Zhang), x.simon.wang@gmail.com (X.S. Wang).

http://dx.doi.org/10.1016/j.ymeth.2014.11.015
1046-2023/

modeling. Needless to say, similarity search based on 2D structural fingerprints also plays a pivotal role in LBVS [22]. To date, new approaches are still emerging at a rapid pace. The recent successes of integrating machine learning (ML) as well as other cheminformatic techniques to improve accuracy of scoring functions [15] are encouraging, e.g. SFCScore (RF) [36], libSVM plus Medusa [37], and the development of novel descriptors [38] or fingerprints [39,40].

With such a large number of VS approaches, it is of utmost importance for the users to learn which one is the optimal method for the specific target(s) under study. For this purpose, the objective assessments for all viable approaches become indispensable. Usually, the performance of each approach is measured by ligand enrichment from retrospective small-scale VS with a benchmarking set, as evidenced by numerous literatures [5,14,41–55]. Ligand enrichment is a metric to assess the capacity to place true ligands at the top-rank of the screen list among a pool of a large number of decoys, which are presumed inactives that are not likely to bind to the target [56,57]. The combination of true ligands and their associated decoys is known as the benchmarking set [58]. This type of assessment is expected to uncover the merits and deficits of each approach for a specific target/task, thus being able to provide advices on method selection for prospective VS campaigns. Particularly, when new algorithms are developed, an objective assessment is normally needed to compare with the prior ones, thus to decide the necessity of the update. Also, in SBVS the assessment can assist in the optimization of receptor structures as well as the selection of the best comparative model(s) for screening purpose [59]. In fact, these types of studies have become the normal practice in both SBVS and LBVS in recent years. Nevertheless, ligand enrichment assessment based on a highly-biased or unsuitable benchmarking set will not reflect the realistic enrichment power of various approaches for prospective VS campaigns. For example, as mentioned by Cleves and Jain, "2D-biased" data sets could cause questionable assessment when comparing SBVS and LBVS approaches [60]. In this way, the quality of the benchmarking sets becomes rather crucial for a fair and comprehensive evaluation.

In our opinion, benchmarking sets can be classified into two major types according to their initial designing purposes, *i.e.* the SBVS-specific and the LBVS-specific. Data sets such as directory of useful decoys (DUD) [56] and its recent DUD-enhanced (DUD-E) [57], virtual decoy sets (VDS) [61], G protein-coupled receptors (GPCRs) ligand library (GLL) and GPCRs decoy database (GDD) [62], demanding evaluation kits for objective *in silico* screening (DEKOIS) [63] and DEKOIS 2.0 [64], nuclear receptors ligands and structures benchmarking database (NRLiSt BDB) belong to SBVS-specific benchmarking sets. By contrast, only 3 data sets, *i.e.* DUD LIB VS 1.0 [65], database of *repro*ducible *vi*rtual *s*creens (REPROVIS-DB) [66] and maximum unbiased validation (MUV) [67] are specifically designed for the purpose of LBVS. A detailed introduction of each data set is given in Table 1. To date, DUD and DUD-E have been intensively employed as gold standard data sets among the community [37,68–73], while much fewer citations of DUD LIB VS 1.0 [55,74] and MUV [75,76] have been reported. In order to broaden the application domain of currently available LBVS-specific benchmarking sets, we recently proposed an unbiased method to build LBVS-specific benchmarking sets [77]. Herein, we review the development of both SBVS-specific and LBVS-specific benchmarking methods/sets and discuss their merits and deficits. In the end, we give a brief introduction to our in-house method and its application to build benchmarking sets for three human histone deacetylases (HDACs) isoforms which are under intensive studies.

2. Currently available benchmarking sets

2.1. SBVS-specific benchmarking sets and methods

2.1.1. Early-stage of benchmarking sets

The usage of benchmarking sets to evaluate docking approaches dates back to early 2000. The first pioneering benchmarking sets were created by Rognan et al. [78], and covered two popular targets: thymidine kinase (TK) and estrogen receptor α subtype (ERα). The data set for each target was composed of 10 antagonists (ligands) and 990 decoys. The method to build the benchmarking sets was relatively simple: First, 10 known ligands were collected for each target; then compounds in advanced chemical directory (ACD) v.2000-1 (Molecular Design Limited, San Leandro) were filtered to eliminate chemical reagents, inorganic compounds, and molecules with unsuitable molecular weights (MWs); at last, 990 compounds were randomly selected as decoys from the remaining compounds. Through these benchmarking sets, the authors addressed issues such as the performances of different docking programs and the accuracy of consensus scoring rationally. Because of the success of this study, parameter optimization based on benchmarking sets had been regarded as a necessity prior to the screening of large chemical libraries. Later on, the benchmarking set for TK was applied to the comparative evaluation of 8 docking tools [42] while similar method was adopted to build benchmarking sets for the assessment of GLIDE [79]. These decoy sets are available at http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html (accessed in Jun. 2014).

Besides ACD, MDL drug data report (MDDR) had also been employed as the main source of decoys during the years of 2002–2005. In Shoichet's group, MDDR was first processed by removing those compounds containing unwanted functional groups such as phosphine. Next, 95,000 remaining "drug-like" compounds were combined with a certain number of ligands for each target [80,81]. Diller et al. [82] selected 32,000 compounds randomly from MDDR as decoys and put them together with over 1,000 known kinase inhibitors across 6 targets. In their study, they kept properties of decoys such as MWs, number of rotatable bonds (RBs), H-bond acceptors (HBAs) and H-bond donors (HBDs) in line with those ligands. Though the criteria for property matching were not strict, this practice should be considered to be an improvement in benchmarking methods. These two types of benchmarking sets were not widely used, however, due to the commercial feature of MDDR.

In 2005, Jain's group at UCSF also released their own decoy sets of "ZINC negative set" to supplement then limited, not publicly-accessible benchmarking sets [83]. To build the sets, they retrieved 20 ligands for each target from PDBbind (http://sw16.im.med.umich.edu/databases/pdbbind/index.jsp) [84]. A total of 1000 decoys (also called "negative ligands") were randomly collected from ZINC drug-like subset. This particular benchmarking set was used to optimize the performance of Surflex-Dock and is available at http://www.jainlab.org/downloads.html. It should be noted that in the above benchmarking sets, the ratio of decoys per ligand was set arbitrarily and the physicochemical properties of ligands and decoys were not strictly matched. Besides, the issue that "drug-like" decoys can be true binders had not been carefully addressed. Those data sets were thus considered to be bias-uncorrected benchmarking sets [56] for comparison purpose.

2.1.2. DUD, DUD clusters and charge-matched DUD

Analysis of early-stage benchmarking sets indicated that molecular size may cause overoptimistic ligand enrichment from SBVS [86–88]. The similar situation applies for other low-dimensional physicochemical properties as well [87]. It is known that poor property matching between ligands and decoys causes the