Journal of Molecular Graphics and Modelling 33 (2012) 35–43

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Graphics and Modelling

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

Results of molecular docking as descriptors to predict human serum albumin binding affinity

Lijuan Chen [a,b], Xin Chen [a,b,*]

[a] State Key Laboratory of Plant Physiology and Biochemistry, Zhejiang University, Hangzhou 310058, PR China
[b] Department of Genetics and Bioinformatics, Zhejiang University, Hangzhou 310058, PR China

ARTICLE INFO

Article history:
Received 8 June 2011
Received in revised form 11 October 2011
Accepted 14 November 2011
Available online 23 November 2011

Keywords:
Molecular docking
Descriptor
QSAR
Serum albumin binding

ABSTRACT

Pharmacokinetic properties of a compound are important in drug discovery and development. These properties are most often estimated from the structural properties of a compound with a structural–activity relationship (QSAR) approach. Rapid advances in molecular pharmacology have characterized a number of important proteins that shape the pharmacokinetic profile of a compound. Previous studies have shown that molecular docking, which is capable of analyzing compound–protein interactions, could be applied to make a categorical estimation of a pharmacokinetic property. The present study focused on the binding affinity of human serum albumin (HSA) as an example to show that docking descriptors might also be useful to estimate the exact value of a pharmacokinetic property. A previously reported dataset containing 94 compounds with log K_{HSA} values was analyzed. A support vector regression model based on the docking descriptors was able to approximate the observed log K_{HSA} in the training and validation dataset with an $R^2 = 0.79$. This accuracy was comparable to known QSAR models based on compound descriptors. In this case study, it was shown that an account of protein flexibility is essential to calculate informative docking descriptors for use in the quantitative estimation of log K_{HSA}.

1. Introduction

Absorption, distribution, metabolism, and excretion (ADME) properties of compounds are important in pharmaceutical research. New drug discovery and development are time-consuming, expensive [1] and have a high attrition rate [2]. An evaluation of the reasons for attrition showed that poor pharmacokinetic properties accounted for nearly 40% of drug development failures [3]. Therefore, a substantial effort has been focused on the early estimation of ADME properties. The predicted properties of compounds have been increasingly considered in the design of combinatorial synthetic routes and high-throughput screening experiments and thus, have improved the quality of leading compounds that may later enter the development stages [4].

Computationally, ADME properties are most often estimated using a quantitative structure–activity relationship (QSAR) approach [5] that is based on the physico-chemical properties of a compound. Since the 1960s, QSAR approaches have successfully produced many classification and regression models that accurately predict a variety of ADME properties for a diverse array of compounds. Examples of ADME properties evaluated include the blood–brain barrier partition [6,7], oral bioavailability [8] and aqueous solubility [9,10]. The correlation coefficient (R) between the estimated and experimentally observed values was as high as 0.988 [11].

Rapid advances in molecular pharmacology have characterized a number of important proteins that shape the pharmacokinetic profile of a compound. Molecular docking, a computational approach capable of predicting the interactions between compounds and their binding proteins, has been used to analyze the behaviors, such as compound binding sites and substrate selectivity, of these proteins [12,13]. Intuitively, the interactions between compounds and pharmacokinetically important proteins may provide useful information for the estimation of pharmacokinetic properties. However, molecular docking results were seldom used for this purpose; docking results have only been reported for the rough categorical estimations of pharmacokinetic properties. For example, Bazeley et al. developed a neural network model using both compound descriptors and docking descriptors to classify the capability of 82 compounds to bind to CYP450 2D6 [14]. It remains unclear whether the docking descriptors are useful in the quantitative estimation of pharmacokinetic properties.

Plasma-protein binding is an important ADME property of drugs that affects their transport and release. Drugs primarily bind to

* Corresponding author at: Department of Genetics and Bioinformatics, Zhejiang University, Hangzhou 310058, PR China. Tel.: +86 571 88208595; fax: +86 571 88208595.
E-mail address: xinchen@zju.edu.cn (X. Chen).

1093-3263/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2011.11.003

three types of plasma proteins: human serum albumin (HSA), α1-acid glycoproteins and lipoproteins [15]. HSA accounts for 60% of the total plasma proteins. It binds a diverse array of drugs and influences their free concentration, solubility, transportation and metabolic clearance [16]. Given the importance of HSA, several QSAR studies have been performed to predict the $\log K_{HSA}$ using different compound descriptors [17–25].

The present study demonstrates that docking descriptors could be useful in the quantitative estimation of $\log K_{HSA}$ as compound structure descriptors. A support vector regression model based on docking descriptors was able to approximate the observed $\log K_{HSA}$ as precisely as other QSAR models based on compound descriptors. It was noted that an account of protein flexibility is essential for the calculation of informative docking descriptors for use in the quantitative estimation of $\log K_{HSA}$.

2. Materials and methods

2.1. Sample dataset

Colmenarejo et al. measured the retention time for 95 drug and drug-like compounds using high-performance affinity chromatography with an immobilized HSA column. The binding affinity constants ($\log K_{HSA}$) of these compounds were calculated based on the retention times [17]. One compound, captorial, displayed the same retention time as Na_2NO_3 and was thus considered as non-binding. The remaining 94 compounds and their $\log K_{HSA}$ values were analyzed by several previous QSAR studies using compound descriptors. These studies provided a sufficient groundwork to evaluate the usefulness of docking descriptors. The compound structures in 3D or 2D (when 3D structures were not available) SDF files were retrieved from the PubChem database (http://pubchem.ncbi.nlm.nih.gov/). When multiple entries were found for the same compound name, the entry provided in the DrugBank database [26] was selected. If a DrugBank entry could not be found for the compound, the PubChem entry without additional ion atoms or protonation states was selected. The 2D structures utilized were converted into 3D structures using Chem3D Ultra software (version 8.0, Cambridge Soft Corporation). The DrugBank ID, PubChem ID, name and $\log K_{HSA}$ value for each compound are shown in Table 1 .

2.2. Preparation of ligand structures

Two compounds in Table 1 (labeled with "a"), ebselen and digitoxin, were excluded in this analysis due to structural issues resulting from the presence of a Se atom and many heterocycles, respectively. The remaining 92 structures were optimized with the LigPrep program (v1.28.4.2, Schrodinger Co. Ltd.) using the default parameters. LigPrep produced one or more optimized structures for a compound after evaluating a wide range of potential protonation states, chiral isoforms, tautomers and ring conformations. The impact of ligand conformation optimization on the docking descriptor calculation is discussed further in later sections.

2.3. Preparation of the HSA structure

With the exception of fragmented or mutant structures, there are 30 monomer and 16 dimer HSA crystal structures available in the Protein Data Bank (PDB) database [27]. The ligands in these structures, if present, were deleted and all dimer structures were split into their monomeric structures. The monomer structures derived from a dimer structure were named using the original PDB ID followed by the chain names. For example, the dimer structure 1AO6 was split into two monomer structures 1AO6_A and 1AO6_B. Altogether, 62 structures were analyzed and superimposed based on sequence alignment using Discovery Studio™ (version 1.7). The pair-wise RMSD distances between these structures were calculated (Supplemental Table 1). Using *K*-means clustering, these structures could be grouped into four clusters. Four structures, 1E7A_A, 1N5U, 1O9X and 1UOR, were selected to represent each cluster. These representative structures were similar (RMSD < 2 Å) to every structure in their respective cluster. In the cases when multiple candidate structures were obtained for a single cluster, the structure with the longer peptide length or higher resolution was chosen. For each representative HSA structure, the water molecules were removed and hydrogen atoms were added using the Maestro™ software (version 9.0). The resultant structures were optimized using the Protein Preparation Wizard to fix their structural defects using the default parameters. Then the optimized structures were energy minimized for 1000 iterations using an OPLS-2005 force field that kept all non-hydrogen atoms frozen at their original positions.

2.4. Molecular docking and docking descriptors

The Glide™ program (version 5.5) [28] was used to predict the binding conformations between compounds and HSA structures using the protocols outlined below. HSA has an extraordinary binding capability for various ligands, such as hemin, fatty acids and drugs, at different sites. A report by Ghuman et al. summarized these binding sites [29]. To define the binding sites on each HSA structure, the SiteMap™ program (version 2.3) [30] was used to predict potential binding sites. The predicted sites supported by Ghuman et al. were considered correct and used for later docking analysis. Although the exact definitions of the sites (e.g., the specific locations of amino acids) in the different HSA structures were slightly different, overall, they were largely consistent. There are six sites (sites 1–6) on each structure and each site resides in roughly the same area across the different HSA structures (Fig. 1). Therefore, each site had four different conformations. These conformations were coded Site A-B, where A is the source PDB ID and B is the site number. According to these site definitions, enclosing grid boxes were computed with a size $\leq$20 Å. Then small compounds were docked into these sites using the Glide standard precision (SP) mode. For each pair of ligand conformations and HSA site conformations, Glide produced the 10 best binding poses. Each pose was associated with 12 scores (Table 2), which are provided in Supplemental Table 2. Based on these scores, 168 descriptors were calculated to represent the docking results between a compound and HSA. These descriptors summarized the results using different ligand conformations and site conformations (Supplemental Table 3).

2.5. Compound structure descriptors

The E-dragon program (version 1.0) [31,32] was used to compute a vector of compound structure descriptors to represent each compound. For each compound, a total of 1666 diverse descriptors were computed. These descriptors were grouped into 20 categories (Supplemental Table 4). Among the 1666 descriptors, the values of 19 descriptors were "−999" in all cases. These descriptors were removed, resulting in a final total of 1647 descriptors.

2.6. Selection of informative descriptors

Before applying the feature selection algorithm outlined below, the descriptors that were zero for >25% compounds were removed. A simple regression algorithm, the stepwise multiple linear regression (MLR) algorithm, was used to select informative descriptors that were associated with the $\log K_{HSA}$ values. For implementation, the PASW Statistics™ (version 18, SPSS, Inc.) software was used.