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.

Table 1
Compounds analyzed in this study.

Compound	PubChem_ID	DrugBank_ID	Exp.[b]	Pred.[c]	Res.[d]
Acetylsalicylic acid	CID_2244	DB00945	−1.39	−0.76	0.63
Cefuroxime	CID_5361202	DB01112	−1.33	−1.13	0.20
Amoxicillin	CID_33613	DB01060	−1.21	−0.83	0.38
Cephalexin	CID_27447	DB00567	−1.11	−0.73	0.38
5-Fluorocytosine	CID_3366	DB01099	−1.11	−1.03	0.08
Cromolyn	CID_27686	DB01003	−1.07	−0.61	0.46
Esbelen[a]	CID_3194		−1.04	–	–
Zidovudine	CID_35370	DB00495	−1.02	−0.80	0.22
Caffeine	CID_2519	DB00201	−0.92	−0.92	0.00
Acetaminophen	CID_1983	DB00316	−0.81	−0.86	0.05
l-Tryptophan	CID_6305	DB00150	−0.78	−0.51	0.27
Methotrexate	CID_126941	DB00563	−0.77	−0.62	0.15
Propylthiouracil	CID_657298	DB00550	−0.75	−0.69	0.06
Antipyrine	CID_2206	DB01435	−0.69	−0.05	0.64
Phenoxymethyl-penicillinic acid	CID_6869	DB00417	−0.69	−0.72	0.03
Salicylic acid	CID_338	DB00936	−0.66	−0.66	0.00
Cefuroxime axetil	CID_5361467	DB01112	−0.56	−0.63	0.07
Etoposide	CID_36462	DB00773	−0.49	−0.49	0.00
Atenolol	CID_2249	DB00335	−0.48	−0.32	0.16
Chloramphenicol	CID_298	DB00446	−0.46	−0.51	0.05
Chlorpropamide	CID_2727	DB00672	−0.44	−0.21	0.23
Cimetidine	CID_2756	DB00501	−0.44	−0.44	0.00
Sotalol	CID_5253	DB00489	−0.44	−0.29	0.15
Hydrochlorothiazide	CID_3639	DB00999	−0.42	−0.42	0.00
Tolazamide	CID_5503	DB00839	−0.42	−0.06	0.36
Nadolol	CID_39147	DB01203	−0.4	−0.16	0.24
Hydrocortisone	CID_5754	DB00741	−0.4	−0.16	0.24
Prednisolone	CID_5755	DB00860	−0.4	−0.18	0.22
Scopolamine	CID_5184	DB00747	−0.34	−0.05	0.29
Timolol	CID_33624	DB00373	−0.33	−0.37	0.04
Metoprolol	CID_4171	DB00264	−0.29	−0.29	0.00
Trimethoprim	CID_5578	DB00440	−0.26	−0.16	0.10
Dansylglycine	CID_70666		−0.26	−0.11	0.15
Lidocaine	CID_3676	DB00281	−0.23	−0.17	0.06
Tolbutamide	CID_5505	DB01124	−0.22	−0.70	0.48
Methylprednisolone	CID_6741	DB00959	−0.22	−0.06	0.16
Acebutolol	CID_1978	DB01193	−0.21	−0.07	0.14
Sulfaphenazole	CID_5335		−0.21	−0.18	0.03
Procaine	CID_4914	DB00721	−0.19	−0.27	0.08
Terazosin	CID_5401	DB01162	−0.16	0.04	0.20
Oxprenolol	CID_4631	DB01580	−0.15	−0.20	0.05
Clonidine	CID_2803	DB00575	−0.13	−0.13	0.00
Frusemide	CID_3440	DB00695	−0.13	−0.61	0.48
Lamotrigine	CID_3878	DB00555	−0.13	−0.25	0.12
Pindolol	CID_4828	DB00960	−0.13	−0.13	0.00
Carbamazepine	CID_2554	DB00564	−0.1	0.14	0.24
Ranitidine	CID_3001055	DB00863	−0.1	0.01	0.11
Camptothecin	CID_24360	DB04690	−0.08	0.26	0.34
Tetracycline	CID_5353990	DB00759	−0.08	0.23	0.31
Bupropion	CID_444	DB01156	−0.05	−0.04	0.01
Sumatriptan	CID_5358	DB00669	−0.05	0.02	0.07
Warfarin	CID_6691	DB00682	−0.04	0.06	0.10
Bumetanide	CID_2471	DB00887	−0.03	0.21	0.24
Oxyphenbutazone	CID_4641	DB03585	−0.02	0.32	0.34
Acrivastine	CID_5284514		−0.02	0.59	0.61
Phenytoin	CID_1775	DB00252	0	0.17	0.17
Doxiciclyne	CID_5281011	DB00254	0.01	0.48	0.47
Ketoprofen	CID_3825	DB01009	0.03	−0.15	0.18
Alprenolol	CID_2119	DB00866	0.04	−0.13	0.17
Prazosin	CID_4893	DB00457	0.06	−0.10	0.16
Digitoxin[a]	CID_441207	DB01396	0.13	–	–
Levofloxacin	CID_149096	DB01137	0.14	0.02	0.12
Ciprofloxacin	CID_2764	DB00537	0.14	0.16	0.02
Labetalol	CID_3869	DB00598	0.14	0.14	0.00
Norfloxacin	CID_4539	DB01059	0.14	−0.26	0.40
Phenylbutazone	CID_4781	DB00812	0.19	0.45	0.26
Minocycline	CID_5281021	DB01017	0.21	−0.06	0.27
Sancicline	CID_5351174		0.21	0.01	0.20
Naproxen	CID_1302	DB00788	0.25	−0.11	0.36
Clofibrate	CID_2796	DB00636	0.27	−0.22	0.49
Propranolol	CID_4946	DB00571	0.28	−0.07	0.35
Tetracaine	CID_5411		0.32	−0.24	0.56
Fusidic acid	CID_3000226	DB02703	0.33	−0.08	0.41
Novobiocin	CID_9346	DB01051	0.35	−0.35	0.70
Ondansetron	CID_4595	DB00904	0.37	0.39	0.02

Table 1 (*Continued*)

Compound	PubChem_ID	DrugBank_ID	Exp.[b]	Pred.[c]	Res.[d]
Droperidol	CID_3168	DB00450	0.43	0.42	0.01
Quinidine	CID_441074	DB00908	0.44	0.78	0.34
Indomethacin	CID_3715	DB00328	0.47	0.16	0.31
Quinine	CID_8549	DB00468	0.49	0.55	0.06
Verapamyl	CID_2520	DB00661	0.52	0.59	0.07
Sulfasalazine	CID_5353980	DB00795	0.56	−0.19	0.75
Progesterone	CID_5994	DB00396	0.59	0.64	0.05
Desipramine	CID_2995	DB01151	0.61	0.61	0.00
Glibenclamide	CID_3488	DB01016	0.68	0.68	0.00
Estradiol	CID_5757	DB00783	0.68	0.33	0.35
Testosterone	CID_6013	DB00624	0.74	0.61	0.13
Imipramine	CID_3696	DB00458	0.75	0.85	0.10
Ketoconazole	CID_47576	DB01026	0.84	0.57	0.27
Promazine	CID_4926	DB00420	0.92	0.68	0.24
Itraconazole	CID_55283	DB01167	1.04	0.96	0.08
Triflupromazine	CID_5568	DB00508	1.05	0.81	0.24
Chlorpromazine	CID_2726	DB00477	1.1	0.60	0.50
Terbinafine	CID_5402	DB00857	1.17	1.15	0.02
Clotrimazole	CID_2812	DB00257	1.34	0.85	0.49

[a] Excluded from this study because of structure issues.
[b] Experimental log K_{HSA}.
[c] Predicted log K_{HSA} by the SVR QSAR model.
[d] Residue.

The "stepping criteria" was set as "$p = 0.01$ for inclusion, $p = 0.05$ for exclusion". The MLR algorithm produced multiple models. Some of these models selected descriptors that were linearly correlated with each other. To reduce the number of selected descriptors, the variance inflation factor (VIF), condition index (CI) and tolerance (T) were used to detect co-linearity in the variables selected for each MLR model. Values of VIF larger than 5 [33,34], T values below 0.2 [35], and CI values larger than 30 [36] indicated co-linearity problems. The best model (in terms of adjusted R^2) without collinear problems was used for feature selection. In this study, the best models used for feature selection always exhibited R^2 coefficients no less than 90% of the highest R^2 coefficient. Therefore, the effort to reduce co-linearity problems did not significantly reduce the fitness of the analysis.

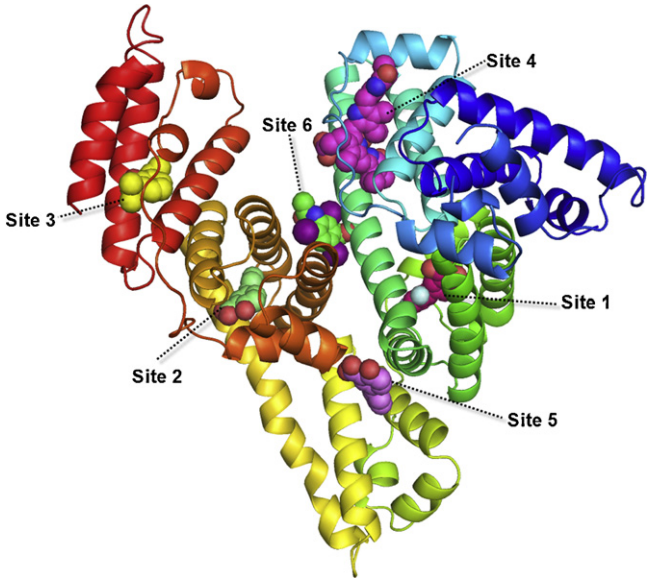

Fig. 1. Drug binding sites. The HSA structure is displayed as a ribbon diagram and colored by secondary structure. The drug binding site is depicted by ligands in space-filling representation (the HSA structure template is PDB ID: 1N5U).

2.7. Regression model training and evaluation

The support vector regression (SVR) algorithm that employed the radial basis function (RBF) kernel was used to build models to predict the log K_{HSA} values. For implementation, the LibSVM software package (version 2.9, http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was used. The dataset of 92 compounds was randomly divided into a training dataset (69 compounds) and an independent validation dataset (23 compounds). The distribution of binding affinity values was roughly the same in both datasets (Fig. 2). The parameters C, ε and γ in the prediction model were estimated from the training dataset using a grid-search with empirical ranges, $[2^{-10}, 2^{10}]$ for C, $[2^{-10}, 2^{-3}]$ for ε and $[2^{-10}, 2^{-3}]$ for γ. The squared correlation coefficient, R^2, for each set of parameters was estimated with a 10-fold cross-validation scheme using the training dataset. The performance of the optimal parameters was evaluated with a leave-one-out cross-validation (LOO-cv) scheme. Then the optimal parameters were used to train the final prediction model with the entire training dataset. The accuracy of this final model was further evaluated with an independent validation dataset that was not used during model training and testing. The consistency between the fitness R^2 values, obtained using LOO-cv based on the training dataset, and the R^2 values, estimated with the independent validation dataset, demonstrated that the prediction model was free of severe over-fitting problems.

Table 2
Definition of the 12 docking scores that Glide reports for predicted binding conformations [54].

Docking score	Definition
Glide gscore	GlideScore
Glide lipo	Lipophilic contact plus phobic attractive term in the GlideScore
Glide hbond	Hydrogen-bonding term in the GlideScore
Glide metal	Metal-binding term in the GlideScore
Glide rewards	Various reward or penalty terms
Glide evdw	van der Waals energy
Glide ecoul	Coulomb energy
Glide erotb	Penalty for freezing rotatable bonds in the GlideScore
Glide esite	Term in the GlideScore for polar interactions in the active site
Glide emodel	Model energy, Emodel
Glide energy	Modified Coulomb–van der Waals interaction energy
Glide einternal	Internal torsional energy

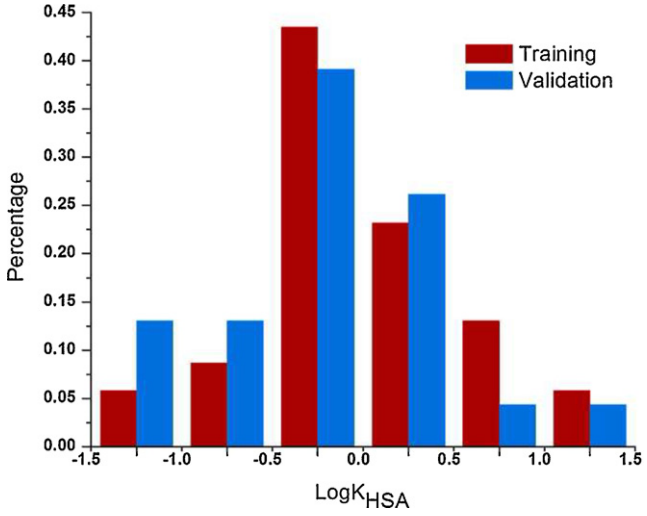

Fig. 2. The distribution of the binding affinities from the training and validation sets. Percentage is the ratio of the number of compounds within the log K_{HSA} range to the total number of compounds in the dataset.

3. Results

3.1. Estimation of the log K_{HSA} using compound structure descriptors

Using only compound structure descriptors, the stepwise MLR algorithm selected the best MLR model without co-linearity that had an adjusted $R^2 = 0.86$ with eight descriptors included (Table 3). These descriptors were used to build a SVR QSAR model. This model exhibited high accuracy without signs of over-fitting. The squared correlation coefficient (R^2) and mean square error (MSE) estimated using LOO-cv with the training dataset were 0.84 and 0.06, respectively. These numbers were similar to those observed using the independent validation dataset ($R^2 = 0.83$ and MSE = 0.06).

Our dataset has been previously analyzed in seven studies [17–23]. These seven studies reported an R^2 ranging from 0.61 to 0.94 during training, and an R^2 ranging from 0.69 to 0.89 in validation. Two of these studies showed similar R^2 values during training and validation; therefore, their estimated accuracies were more reliable. Colmenarejo et al. used six descriptors, H-bondDon (constitutional), AM-1 dipole moment, E_{HOMO} (quantum), ClogP, JursTPSA and X_{ring} (molecular connectivity) to achieve an $R^2 = 0.79$ during training and an $R^2 = 0.82$ in validation. Hall et al. also used four E-state descriptors, S^T(arom), S^T(CHsat), S^T(–F,Cl) and S^T(–OH), and two molecular connectivity descriptors, $^6\chi^v_{CH}$ and $^5\chi^v_{CH}$, to achieve an $R^2 = 0.70$ in training and an $R^2 = 0.74$ in validation [21].

Table 3
Compound structure descriptors selected by stepwise MLR.

Descriptor	Definition	Category
P2s	2nd component shape directional WHIM index/weighted by atomic electrotopological states	WHIM
Mor32v	3D-MoRSE – signal 32/weighted by atomic van der Waals volumes	Geometrical
GATS3e	Moran autocorrelation – lag 3/weighted by atomic Sanderson electronegativities	2D-autocorrelations
R7u	*R* autocorrelation of lag 7/unweighted	Gateway
nR06	Number of 6-membered rings	Constitutional
ALOGP	Ghose–Crippen octanol–water partition coeff.	Molecular
ALOGPS_logS	Aqueous solubility	Molecular
nO	Number of oxygen atoms	Constitutional

Table 4
Docking descriptors based on all receptor conformations and all ligand conformations that were selected by stepwise MLR.

Descriptor	Definition
1O9X_min_lipo	Minimum Glide lipo of ligands to 1O9X docking
aver_s2_ecoul	Average Glide ecoul of ligands to site 2 of all HSA conformation docking
1E7AA_min_esite	Minimum Glide esite of ligands to 1E7A_A docking
aver_s5_evdw	Average Glide evdw of ligands to site 5 of all HSA conformation docking
aver_s4_lipo	Average Glide lipo of ligands to site 4 of all HSA conformation docking
1O9X_topscore	Minimum Glide gscore of ligands to 1O9X docking

Our model accuracy was comparable to Colmenarejo's model and better than Hall's model.

3.2. Estimation of the log K_{HSA} using docking descriptors

Using only docking result descriptors, the stepwise MLR algorithm selected the best MLR model without co-linearity. This model had an adjusted $R^2 = 0.74$ and used six descriptors (Table 4). These descriptors were used to build an SVR QSAR model. This model exhibited high accuracy without signs of over-fitting. The R^2 and MSE estimated using LOO-cv with the training dataset were 0.75 and 0.09, respectively. These numbers were similar to those observed using the independent validation dataset ($R^2 = 0.78$ and MSE = 0.09). The correlation between predicted log K_{HSA} and experimental log K_{HSA} within the training and validation datasets is shown in Fig. 3.

The accuracy of the docking descriptor-based QSAR model was slightly lower than that of the compound structure-based QSAR model, but it was still comparable. Furthermore, as shown in Table 1, there were 11 compounds for which the predicted log K_{HSA} values were identical to the experiment results. The largest prediction error from our docking descriptor-based QSAR model was 0.75. This value was smaller than the more accurate (in terms of R^2) Colmenarejo's model, where the largest prediction error was −1.0. These results showed that the docking descriptors were able to predict log K_{HSA} as reliably as compound structure descriptors.

In previous studies, it was established that docking scores could only correlate well with binding affinities between a series of analog compounds and the same protein receptor. No correlation or a very weak correlation can be detected if heterogeneous classes of compounds were used. For example, Keserű docked 33 ligands into the crystal structure of CP450cam (cytochrome P450 camphor monooxygenase) and obtained an $R^2 = 0.88$ between the docking scores and experimental binding constants [37]. Kemp

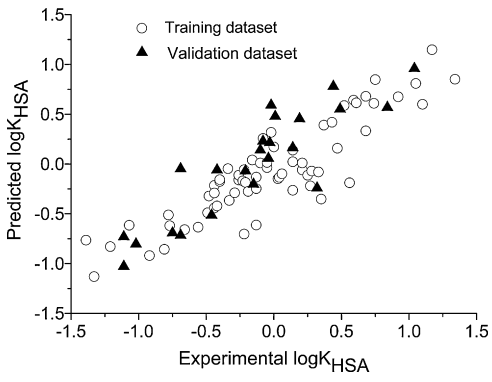

Fig. 3. Predicted and experimental log K_{HSA} values.

Table 5
Docking descriptors based on single protein conformation that were selected by stepwise MLR.

HSA	Descriptor	Definition
1E7A_A	1E7AA_min_lipo	Minimum Glide lipo of ligands to 1E7A_A docking
	1E7AA_min_hbond	Minimum Glide hbond of ligands to 1E7A_A docking
	1E7AA_min_rewards	Minimum Glide rewards of ligands to 1E7A_A docking
	1E7AA_min_esite	Minimum Glide esite of ligands to 1E7A_A docking
	1E7AA_min_ecoul	Minimum Glide ecoul of ligands to 1E7A_A docking
1N5U	1N5U_min_lipo	Minimum Glide lipo of ligands to 1N5U docking
	1N5U_max_emodel	Maximum Glide lipo of ligands to 1N5U docking
	1N5U_min_hbond	Minimum Glide hbond of ligands to 1N5U docking
1O9X	1O9X_min_lipo	Minimum Glide lipo of ligands to 1O9X docking
	1O9X_min_energy	Minimum Glide energy of ligands to 1O9X docking
	1O9X_min_esite	Minimum Glide esite of ligands to 1O9X docking
1UOR	1UOR_min_lipo	Minimum Glide lipo of ligands to 1UOR docking
	1UOR_min_rewards	Minimum Glide rewards of ligands to 1UOR docking
	1UOR_min_ecoul	Minimum Glide ecoul of ligands to 1UOR docking

et al. docked 33 compounds into a homology model of CYP2D6 and obtained a regression coefficient $R^2 = 0.60$ between the experimental $\log IC_{50}$ and the ChemScore [38]. However, Enyedy et al. noted that the Glide XP and Glide emodel scores of $R^2 = 0.05$ and $R^2 = 0.12$, respectively, calculated from 4904 heterogeneous compounds docking to the PDB structure 1YMN of KDR (kinase insert domain protein receptor) had nearly no correlation with experimental IC_{50} values [39]. Our results indicated that although docking scores may not correlate well with binding affinities involving heterogeneous classes of compounds, docking descriptors carrying more comprehensive information about the docking results could show a much better correlation with the binding affinities.

3.3. Estimation of the $\log K_{HSA}$ using both compound structure and docking descriptors

The possibility of combing compound structure and docking descriptors to achieve better prediction accuracy was investigated. The same stepwise MLR protocol was used to select features from a combined descriptor pool. This analysis yielded six descriptors, including five compound structure descriptors (ALOGP, nR06, R7u, Mor32v, nO) and one docking descriptor (1O9X_max_emodel). Unfortunately, the SVR QSAR model based on this set of mixed descriptors did not show improved accuracy. The R^2 values observed with the training dataset and the independent validation dataset were 0.78 and 0.76, respectively, which were comparable to those from the SVR QSAR model based on compound descriptors alone. Combining the compound structure and docking descriptors did not show the potential to increase prediction accuracy.

4. Discussion

4.1. The difference between compound structure and docking descriptors

Compound structure descriptors are usually "global" descriptors that are computed based on the entire compound structure. For example, physicochemical property descriptors, such as the octanol–water partition coefficient, are determined by the entire structure [40]. Functional descriptors such as nHDon, the number of donor atoms for H-bonds [41], do not describe how these hydrogen donors are distributed around the compound. Molecular indices, though reflecting substructure organization, do not carry information about the relative positions of these substructures. Therefore, it is not straightforward to use these descriptors to get a detailed description of a substructure in a compound. For example, it is not clear if a particular substructure simultaneously has a certain size and shape, enough accessibility, the desired number of hydrogen donors/acceptors and electron charge distribution to interact with a receptor-binding site. Fortunately, the long history of QSAR study has produced a large pool of compound structure descriptors that has alleviated this problem. In a limited structure space, a combination of global descriptors may provide sufficiently detailed, though implicit, descriptions of the overall properties of any substructure.

Alternatively, ligand–receptor docking is directly based on the detailed physics of the interaction between a ligand and a protein, and its results have explicit information on the substructures of compounds that are responsible for the selective binding. Therefore, docking results have the theoretical advantage of focusing on the actual structural determinants of binding affinity. However, ligand–protein binding is a process of induced fitting, and the computational need to accurately simulate this process is still prohibitive [42,43]. Available docking programs all use strategies to simplify the computation, and their results are approximate. It was found that the docking scores, the overall measurement of ligand–protein fitness, seldom correlate well with binding affinities [39,44,45]. One review even suggested that molecular docking was not appropriate for quantitative prediction of ligand binding affinity [46].

In other words, while molecular docking has a theoretical advantage to predict binding affinity more accurately, the difficulty in obtaining accurate predictions of binding conformation significantly offsets this advantage. Previous efforts attempting to predict binding affinities using single docking scores were generally unsuccessful. Our results suggested that by integrating multiple docking results, the problem of docking inaccuracy could be alleviated. The biggest challenge to accurate prediction of binding conformation is arguably the consideration of ligand and protein flexibility in docking [47,48]. Our docking descriptors summarized docking results using different starting ligand conformations and protein conformations. This method, which considered the ligand and protein flexibility, provided promising results. These docking descriptors were able to predict $\log K_{HSA}$ as accurately as the compound structure descriptors. The importance of ligand flexibility and protein flexibility in computing effective docking descriptors was further analyzed.

4.2. The effect of protein flexibility

This study used four HSA conformations, 1UOR, 1E7A_A, 1N5U and 1O9X, to represent typical HSA conformations known to bind different kinds of ligands. These structures are different on several levels from residue movements to rigid-body rotations [49–52]. To determine if using a diverse set of protein structures can improve the predictive capability of docking descriptors, we conducted four sets of experiments. In each experiment, only one HSA structure

Table 6
Prediction accuracy of SVR QSAR models based on docking descriptors that were calculated using single protein conformations.

HSA[a]	N[b]	MLR		SVM					
		R^2	s[c]	$R^2_{\text{LOO-cv}}$ [e]	$\text{MSE}_{\text{LOO-cv}}$ [d,e]	R^2_{training}	$\text{MSE}_{\text{training}}$ [d]	$R^2_{\text{validation}}$	$\text{MSE}_{\text{validation}}$ [d]
1E7A_A	5	0.61	0.36	0.59	0.14	0.62	0.13	0.57	0.16
1N5U	3	0.55	0.39	0.46	0.18	0.54	0.16	0.62	0.13
1O9X	3	0.60	0.37	0.55	0.15	0.61	0.13	0.68	0.12
1UOR	3	0.53	0.40	0.53	0.16	0.61	0.13	0.58	0.14
All	6	0.74	0.23	0.75	0.09	0.79	0.07	0.78	0.09

[a] HSA, the HSA conformation used to compute the docking descriptors.
[b] N, number of docking descriptors selected to build the SVR QSAR model.
[c] s, standardized error of the best MLR model.
[d] MSE, mean standardized error.
[e] LOO-cv, leave one out cross validation.

was used to compute the docking descriptors. In these experiments, some docking descriptors summarizing results from different protein conformations degenerated into simple docking results. For example, the maximum and average docking scores between a ligand and all four conformations of a ligand-binding site became the same when only one protein conformation was used. Descriptors that were redundant by definition were removed. In experiments using single HSA conformations, there were 79 non-redundant docking descriptors (Supplemental Table 5).

Using the same protocol, docking descriptors were selected using the stepwise MLR approach. SVR QSAR models were built with the selected descriptors (Table 5). As shown in Table 6, the R^2 of the best SVR model using a single HSA conformation was 0.62 during training and 0.57 during evaluation (using the 1E7A_A conformation). The average R^2 was 0.60 during training and 0.61 during evaluation. These numbers are noticeably lower than the accuracy of the QSAR model that used all four protein conformations ($R^2 = 0.79$ during training and $R^2 = 0.78$ during evaluation). These data demonstrated that summarizing docking results from different starting protein conformations noticeably improved the log K_{HSA} prediction accuracy. In other words, protein flexibility is an important factor to consider when computing docking descriptors for affinity estimation.

4.3. The effect of ligand flexibility

This study calculated docking descriptors using two types of ligand conformations: 3D SDF structures downloaded from the PubChem database and structures optimized by the LigPrep software. For each type of ligand conformation, the same set of docking descriptors was calculated as previously described (Supplemental Table 3). After removing the descriptors that were zero for >25% compounds, there were 144 and 146 non-redundant docking descriptors based on either the SDF ligand conformations or the LigPrep ligand conformations, respectively (Supplemental Table 6).

These docking descriptors were selected using the stepwise MLR approach. SVR QSAR models were built with the selected descriptors (Table 7). The resultant QSAR model that used the SDF ligand conformations utilized nine descriptors and yielded an $R^2 = 0.79$ during training and $R^2 = 0.82$ during evaluation (Table 8). The resultant QSAR model that used the LigPrep ligand conformations utilized six descriptors and yielded an $R^2 = 0.75$ during training and $R^2 = 0.71$ during evaluation (Table 8). As shown in Table 8, the accuracies of the models using only one type of ligand conformation did not drop sharply from the models that used all ligand conformations. These data showed that, compared to protein flexibility, ligand flexibility was less critical in the docking descriptor calculation.

This observation is consistent with the current situation in which ligand structures can be modeled to a similar quality; however, this is not the case for protein structures. There are various quality issues in protein structures, such as alternate side chain conformations, crystal packing interactions, poor electron density maps and uncertainty in ligand positions due to poor fittings to density maps or errors in preparations. In addition, in most docking software, ligand structure flexibility can be simulated with a conformation sampling algorithm, while proteins are usually treated as a rigid body or have only very limited conformational flexibility [53]. Therefore, the strategy of using a docking descriptor to summarize the results of docking from multiple protein conformations can be a practical approach to address the current inadequacy of docking programs in handling protein conformation issues.

Table 7
Docking descriptors based on single ligand conformations that were selected by stepwise MLR.

Ligand conformation	Descriptor	Definition
SDF	1E7AA_min_rewards	Minimum Glide rewards of "SDF" conformation to 1E7A_A docking
	1UOR_min_einternal	Minimum Glide einternal of "SDF" conformation to 1UOR docking
	aver_s1_ecoul	Average Glide ecoul of "SDF" conformation to site 1 of all HSA conformation docking
	aver_s4_esite	Average Glide esite of "SDF" conformation to site 4 of all HSA conformation docking
	1E7AA_min_lipo	Minimum Glide lipo of "SDF" conformation to 1E7A_A docking
	aver_s4_hbond	Average Glide hbond of "SDF" conformation to site 4 of all HSA conformation docking
	1N5U_max_einternal	Maximum Glide einternal of "SDF" conformation to 1N5U docking
	aver_s4_ecoul	Average Glide ecoul of "SDF" conformation to site 4 of all HSA conformation docking
	aver_s5_hbond	Average Glide hbond of "SDF" conformation to site 5 of all HSA conformation docking
LigPrep	1O9X_min_lipo	Minimum Glide lipo of "LigPrep" conformation to 1O9X docking
	1UOR_min_hbond	Minimum Glide hbond of "LigPrep" conformation to 1UOR docking
	1N5U_max_energy	Maximum Glide energy of "LigPrep" to 1N5U docking
	1E7AA_min_esite	Minimum Glide esite of "LigPrep" to 1E7A_A docking
	aver_s2_ecoul	Average Glide ecoul of "LigPrep" to site 2 of all HSA conformations docking
	1E7AA_max_evdw	Maximum Glide evdw of "LigPrep" to 1E7A_A docking

Table 8
Prediction accuracy of SVR QSAR models based on docking descriptors that were calculated using single ligand conformations.

Ligand[a]	N[b]	MLR		SVM					
		R^2	s[c]	$R^2_{\text{LOO-cv}}$[e]	$\text{MSE}_{\text{LOO-cv}}$[d,e]	R^2_{training}	$\text{MSE}_{\text{training}}$[d]	$R^2_{\text{validation}}$	$\text{MSE}_{\text{validation}}$[d]
SDF	9	0.75	0.29	0.72	0.10	0.79	0.08	0.82	0.06
LigPrep	6	0.72	0.30	0.69	0.11	0.75	0.10	0.71	0.10
Both	6	0.74	0.23	0.75	0.09	0.79	0.07	0.78	0.09

[a] Ligand, the type of ligand conformation used for calculating docking descriptors.
[b] N, number of docking descriptors selected to build the SVR QSAR model.
[c] s, standardized error of the best MLR model.
[d] MSE, mean standardized error.
[e] LOO-cv, leave one out cross validation.

5. Concluding remarks

Compound structure descriptors have been used for decades to predict binding affinities, and a number of programs have been developed to compute a wide range of compound structure descriptors. Molecular docking results, though possessing the theoretical advantage to directly describe the binding process, have failed to establish their usefulness in the quantitative estimation of binding affinities. This is because, among many reasons, current docking programs have a very limited capability to handle protein structure flexibility. Therefore, docking results are approximate and inappropriate for accurate estimation of binding affinities. This study, using the estimation of $\log K_{\text{HSA}}$ values as an example, showed that using a docking descriptor to summarize the results of docking with multiple protein conformations can be a practical approach to address the current inadequacy of docking programs in handling protein conformation issues. It is our hope that this limited case study will invite further effort to develop more delicate approaches to summarize docking results into docking descriptors. This might create a new category of descriptors for QSAR studies.

Acknowledgements

This work was supported by grants from the Natural Science Foundation of China (NSFC) [Nos. 30970690 and 31071161] and Zhejiang Provincial Natural Science Foundation of China [No. R207609].

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2011.11.003.

References

[1] D.S. Wishart, Improving early drug discovery through ADME modelling: an overview, Drugs R. D. 8 (2007) 349.
[2] I. Kola, J. Landis, Can the pharmaceutical industry reduce attrition rates? Nat. Rev. Drug Discov. 3 (2004) 711–716.
[3] R.A. Prentis, Y. Lis, S.R. Walker, Pharmaceutical innovation by the seven UK-owned pharmaceutical companies (1964–1985), Br. J. Clin. Pharmacol. 25 (1988) 387–396.
[4] J. Hodgson, ADMET – turning chemicals into drugs, Nat. Biotechnol. 19 (2001) 722–726.
[5] D. Butina, M.D. Segall, K. Frankcombe, Predicting ADME properties in silico: methods and models, Drug Discov. Today 7 (2002) S83–S88.
[6] T.J. Hou, X.J. Xu, ADME evaluation in drug discovery. 3. Modeling blood–brain barrier partitioning using simple molecular descriptors, J. Chem. Inf. Comput. Sci. 43 (2003) 2137–2152.
[7] M. Lobell, L. Molnar, G.M. Keseru, Recent advances in the prediction of blood–brain partitioning from molecular structure, J. Pharm. Sci. 92 (2003) 360–370.
[8] F. Yoshida, J.G. Topliss, QSAR model for drug human oral bioavailability, J. Med. Chem. 43 (2000) 2575–2585.
[9] J. Huuskonen, M. Salo, J. Taskinen, Aqueous solubility prediction of drugs based on molecular topology and neural network modeling, J. Chem. Inf. Comput. Sci. 38 (1998) 450–456.
[10] A. Cheng, K.M. Merz Jr., Prediction of aqueous solubility of a diverse set of compounds using quantitative structure–property relationships, J. Med. Chem. 46 (2003) 3572–3580.
[11] T.J. Hou, X.J. Xu, ADME evaluation in drug discovery. 2. Prediction of partition coefficient by atom-additive approach based on atom-weighted solvent accessible surface areas, J. Chem. Inf. Comput. Sci. 43 (2003) 1058–1067.
[12] F. Yamashita, M. Hashida, In silico approaches for predicting ADME properties of drugs, Drug Metab. Pharmacokinet. 19 (2004) 327–338.
[13] M.M. Ahlström, M. Ridderström, I. Zamora, CYP2C9 structure–metabolism relationships: substrates, inhibitors, and metabolites, J. Med. Chem. 50 (2007) 5382–5391.
[14] P.S. Bazeley, S. Prithivi, C.A. Struble, R.J. Povinelli, D.S. Sem, Synergistic use of compound properties and docking scores in neural network modeling of CYP2D6 binding: predicting affinity and conformational sampling, J. Chem. Inf. Model. 46 (2006) 2698–2708.
[15] P.A. Routledge, The plasma protein binding of basic drugs, Br. J. Clin. Pharmacol. 22 (1986) 499–506.
[16] G. Colmenarejo, In silico prediction of drug-binding strengths to human serum albumin, Med. Res. Rev. 23 (2003) 275–301.
[17] G. Colmenarejo, A. Alvarez-Pedraglio, J.L. Lavandera, Cheminformatic models to predict binding affinities to human serum albumin, J. Med. Chem. 44 (2001) 4370–4378.
[18] O. Deeb, B. Hemmateenejad, ANN-QSAR model of drug-binding to human serum albumin, Chem. Biol. Drug Des. 70 (2007) 19–29.
[19] E. Estrada, E. Uriarte, E. Molina, Y. Simon-Manso, G.W. Milne, An integrated in silico analysis of drug-binding to human serum albumin, J. Chem. Inf. Model. 46 (2006) 2709–2724.
[20] S.B. Gunturi, R. Narayanan, A. Khandelwal, In silico ADME modelling. 2: computational models to predict human serum albumin binding affinity using ant colony systems, Bioorg. Med. Chem. 14 (2006) 4118–4129.
[21] L.M. Hall, L.H. Hall, L.B. Kier, Modeling drug albumin binding affinity with e-state topological structure representation, J. Chem. Inf. Comput. Sci. 43 (2003) 2120–2128.
[22] K. Wichmann, M. Diedenhofen, A. Klamt, Prediction of blood–brain partitioning and human serum albumin binding based on COSMO-RS sigma-moments, J. Chem. Inf. Model. 47 (2007) 228–233.
[23] C.X. Xue, R.S. Zhang, H.X. Liu, X.J. Yao, M.C. Liu, Z.D. Hu, et al., QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine, J. Chem. Inf. Comput. Sci. 44 (2004) 1693–1700.
[24] P.J. Hajduk, R. Mendoza, A.M. Petros, J.R. Huth, M. Bures, S.W. Fesik, et al., Ligand binding to domain-3 of human serum albumin: a chemometric analysis, J. Comput. Aided Mol. Des. 17 (2003) 93–102.
[25] J.R. Votano, M. Parham, L.M. Hall, L.H. Hall, L.B. Kier, S. Oloff, et al., QSAR modeling of human serum protein binding with several modeling techniques utilizing structure–information representation, J. Med. Chem. 49 (2006) 7169–7181.
[26] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, et al., DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs, Nucleic Acids Res. 39 (2011) D1035–D1041.
[27] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, et al., The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.
[28] Glide 5.5, Schrodinger, LLC, New York, NY, 2009.
[29] J. Ghuman, P.A. Zunszain, I. Petitpas, A.A. Bhattacharya, M. Otagiri, S. Curry, Structural basis of the drug-binding specificity of human serum albumin, J. Mol. Biol. 353 (2005) 38–52.
[30] SiteMap 2.3, Schrodinger, LLC, New York, NY, 2009.
[31] I. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, et al., Virtual computational chemistry laboratory – design and description, J. Comput. Aided Mol. Des. 19 (2005) 453–463.
[32] VCCLAB, Virtual Computational Chemistry Laboratory, 2005. http://www.vcclab.org.
[33] R.D. Snee, Validation of regression models: methods and examples Technometrics, vol. 19, American Statistical Association and American Society for Quality, 1977, pp. 415–428.
[34] S. Chatterjee, A.S. Hadi, Regression Analysis by Example, Wiley-Interscience, 2006.
[35] S.W. Menard, Applied Logistic Regression Analysis, Sage Publications, 2002.
[36] D.A. Belsley, E. Kuh, R.E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, 2004.

[37] G.M. Keserű, A virtual high throughput screen for high affinity cytochrome P450cam substrates. Implications for in silico prediction of drug metabolism, J. Comput. Aided Mol. Des. 15 (2001) 649–657.
[38] C.A. Kemp, J.U. Flanagan, A.J. van Eldik, J.D. Marechal, C.R. Wolf, G.C. Roberts, et al., Validation of model of cytochrome P450 2D6: an in silico tool for predicting metabolism and inhibition, J. Med. Chem. 47 (2004) 5340–5346.
[39] I. Enyedy, W. Egan, Can we use docking and scoring for hit-to-lead optimization? J. Comput. Aided Mol. Des. 22 (2008) 161–168.
[40] J. Sangster, Octanol–Water Partition Coefficients: Fundamentals and Physical Chemistry, Wiley, 1997.
[41] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, 2000.
[42] V. Mohan, A.C. Gibbs, M.D. Cummings, E.P. Jaeger, R.L. DesJarlais, Docking: successes and challenges, Curr. Pharm. Des. 11 (2005) 323–333.
[43] M.L. Teodoro, G.N. Phillips Jr., L.E. Kavraki, Molecular docking: a problem with thousands of degrees of freedom, in: Robotics and Automation, 2001. Proceedings 2001 ICRA 1. IEEE International Conference on., 2001, pp. 960–965.
[44] P.M. Marsden, D. Puvanendrampillai, J.B.O. Mitchell, R.C. Glen, Predicting protein–ligand binding affinities: a low scoring game? Org. Biomol. Chem. 2 (2004) 3267–3273.
[45] R. Wang, Y. Lu, S. Wang, Comparative evaluation of 11 scoring functions for molecular docking, J. Med. Chem. 46 (2003) 2287–2303.
[46] C. de Graaf, N.P.E. Vermeulen, K.A. Feenstra, Cytochrome P450 in silico: an integrative modeling approach, J. Med. Chem. 48 (2005) 2725–2755.
[47] B. Coupez, R.A. Lewis, Docking and scoring – theoretically easy, practically impossible? Curr. Med. Chem. 13 (2006) 2995–3003.
[48] J.A. Erickson, M. Jalaie, D.H. Robertson, R.A. Lewis, M. Vieth, Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy, J. Med. Chem. 47 (2004) 45–55.
[49] X.M. He, D.C. Carter, Atomic structure and chemistry of human serum albumin, Nature 358 (1992) 209–215.
[50] M. Wardell, Z. Wang, J.X. Ho, J. Robert, F. Ruker, J. Ruble, et al., The atomic structure of human methemalbumin at 1.9 Å, Biochem. Biophys. Res. Commun. 291 (2002) 813–819.
[51] A.A. Bhattacharya, S. Curry, N.P. Franks, Binding of the general anesthetics propofol and halothane to human serum albumin, J. Biol. Chem. 275 (2000) 38731–38738.
[52] P. Zunszain, J. Ghuman, T. Komatsu, E. Tsuchida, S. Curry, Crystal structural analysis of human serum albumin complexed with hemin and fatty acid, BMC Struct. Biol. 3 (2003) 6.
[53] D.B. Kitchen, H. Decornez, J.R. Furr, J. Bajorath, Docking and scoring in virtual screening for drug discovery: methods and applications, Nat. Rev. Drug Discov. 3 (2004) 935–949.
[54] Glide 5.5 User Manual: Glide Overview, 2009, pp. 5–10.