MedusaScore: An Accurate Force Field-Based Scoring Function for Virtual Drug Screening

Shuangye Yin,[†] Lada Biedermannova,[‡] Jiri Vondrasek,[‡] and Nikolay V. Dokholyan*,[†]

Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, and Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, and Center for Biomolecules and Complex Molecular Systems, Prague, Czech Republic

Received April 15, 2008

Virtual screening is becoming an important tool for drug discovery. However, the application of virtual screening has been limited by the lack of accurate scoring functions. Here, we present a novel scoring function, MedusaScore, for evaluating protein−ligand binding. MedusaScore is based on models of physical interactions that include van der Waals, solvation, and hydrogen bonding energies. To ensure the best transferability of the scoring function, we do not use any protein−ligand experimental data for parameter training. We then test the MedusaScore for docking decoy recognition and binding affinity prediction and find superior performance compared to other widely used scoring functions. Statistical analysis indicates that one source of inaccuracy of MedusaScore may arise from the unaccounted entropic loss upon ligand binding, which suggests avenues of approach for further MedusaScore improvement.

INTRODUCTION

The current rate of emerging new pharmaceutical targets outpaces that of new drug leads,[1] posing a significant challenge for drug discovery. Traditional trial-and-error approach to drug discovery represents a substantial challenge due to the enormous dimensionality of the chemical space. Structure-based drug design is a promising approach of rational drug discovery, which takes advantage of the increasing amount of solved three-dimensional structures of target proteins.[2] By computational modeling of the target binding site, a ligand can be constructed de novo[3] or via screening over a large database of millions of chemical compounds (virtual screening).[4,5] In a typical virtual screening workflow, a library of small molecules are first computationally docked to the target protein and then ranked according to the predicted binding affinity. A scoring function is used throughout the process to (a) recognize the correct binding pose out of hundreds of computer-generated docking models (decoys) and to (b) predict the binding affinity for each molecule.

Docking algorithms have undergone substantial developments over the last two decades.[4,6−12] Early attempts treated receptors and ligands as rigid bodies and docking was only based on molecular surface matching.[13] Docking programs nowadays not only allow full flexibility of the ligands but also partially treat receptors as flexible objects. However, how to systematically treat protein flexibility and ligand-induced protein conformational changes remains a considerable challenge in the further development of docking algorithms.[4] Despite the progress in docking methodologies, recent studies have shown that scoring functions are becoming a bottleneck in structure-based virtual screening.[14−16] Benchmark studies have found that in many cases docking programs can generate native-like docking poses, but these are often missed at the scoring stage. Moreover, the binding scores predicted by the scoring functions may exhibit poor correlation with the actual binding affinity, resulting in a large percentage of false positives in the "hit list".[16]

Although several types of traditional scoring functions exist, all of them have major limitations. Physical force field-based approaches aim to describe protein−ligand interactions using elementary physical interactions. Combined with molecular dynamics (MD) and free energy perturbation (FEP) techniques,[17,18] the binding affinity can be reproduced to within an accuracy of 1 kcal/mol.[19,20] However, such approaches often involve intensive computation, especially in connection with the explicit solvent used to describe desolvation effects and hydrogen bonding. Because of the sampling insufficiency, such approaches are often limited to structurally similar targets and ligands. These speed and sampling limitations undermine the application of the physical force field-based scoring functions in virtual screening of vast molecular libraries. Alternatively, empirical and knowledge-based scoring functions[7,21−27] circumvent the speed and sampling problems by dissecting the protein−ligand interaction into statistical or empirical "potentials".[28] Such approaches however rely on parameter training using known protein−ligand binding structures or binding affinity measurements or both. Due to the limited size of the training set, the resulting scoring functions can be to too specialized (overtrained) thus having low transferability to targets and ligands that are structurally distinct from those in the training set.[29]

To eliminate these limitations, we report a novel scoring function MedusaScore for evaluating protein−ligand binding. MedusaScore describes the protein−ligand binding using a physical interaction model. MedusaScore includes an explicit

* Corresponding author e-mail: dokh@med.unc.edu.
[†] University of North Carolina at Chapel Hill.
[‡] Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, and Center for Biomolecules and Complex Molecular Systems.

10.1021/ci8001167 CCC: $40.75
Published on Web 08/02/2008

hydrogen-bonding model[30] and EEF1 pairwise implicit solvent model,[31] which allows accurate modeling of the hydrogen-bonding and desolvation effect without large-scale MD simulations. Additionally, unlike other statistical and empirical scoring functions, MedusaScore does not depend on any specific parameter training on protein−ligand data sets, thereby maintaining the transferability over a wide range of potential drug candidates in the chemical space.

We test the performance of MedusaScore for docking decoy recognition and binding affinity prediction. Using the docking decoys data sets generated by Wang et al.,[14] we find that the recognition success rate for docking decoys is 82%, higher than that of 11 other scoring functions that are currently widely used in virtual screening, including LigScore, PLP,[7] PMF,[24] LUDI,[21] F-Score,[8] G-Score,[9] D-Score,[11] ChemScore,[22] Autodock,[10] DrugScore,[25] and X-Score.[26] The success rate can be further improved to 85% by consensus scoring with DrugScore.[25] Using the PDBBind 2005 data set[32] for the binding affinity prediction, we find that the MedusaScore showed a correlation coefficient of 0.60 and 0.61 for the core set and refined set (see Methods), respectively. This correlation is higher than what has been reported for 14 other scoring function using the same database.[15] Statistical analysis suggests that the entropic contribution may be the key component for further improvement of the accuracy of the binding affinity prediction.

METHODS

Medusa Force Field. The Medusa force field[33] is a weighted sum of six energy terms

$$E = W_{vdw\_attr}E_{vdw\_attr} + W_{vdw\_rep}E_{vdw\_rep} + W_{solv}E_{solv} + W_{bb\_hbond}E_{bb\_hbond} + W_{sc\_hbond}E_{sc\_hbond} + W_{bb\_sc\_hbond}E_{bb\_sc\_hbond}$$

where E_{vdw_attr} and E_{vdw_rep} are the attractive and repulsive part of the van der Waals (VDW) interaction; E_{solv} is the solvation energy; and E_{bb_hbond}, E_{sc_hbond}, and $E_{bb\_sc\_hbond}$ are the hydrogen bond energies formed between backbone atoms, between side chains, and between backbone and side chains, respectively. The design of the force field is similar to that of the Rosetta force field,[34] which has also been widely used in protein folding and design. The VDW interaction model and parameters are adapted from CHARMM19.[35] The solvation model is the EEF1 implicit solvent model proposed by Lazaridis and Karplus.[31] We use the hydrogen bonding model proposed by Kortemme and Baker.[36] When evaluating the nonbonded interactions, we use a cutoff distance of 9.0 Å. The van der Waals repulsion (VDWR) potentials are implemented with linear extrapolation to dampen the fast increase of the potential as

$$E_{vdw\_rep} = \begin{cases} \sum_{i,j>i} 4\varepsilon_{ij}[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^{6}], \alpha_{cutoff}\sigma_{ij} < r_{ij} \le \sigma_{ij} \\ K_{slope}r_{ij} + 4\varepsilon_{ij}(\alpha_{cutoff}{}^{-12} - \alpha_{cutoff}{}^{-6}) - \\ \alpha_{cutoff}K_{slope}\sigma_{ij}, r_{ij} \le \alpha_{cutoff}\sigma_{ij}\ Here, \alpha_{cutoff} = \\ 0.92; K_{slope} = -24\varepsilon_{ij}(2\alpha_{cutoff}{}^{-13} - \alpha_{cutoff}{}^{-7})/\sigma_{ij} \end{cases}$$

$$\varepsilon_{ij} = \sqrt{\varepsilon_i\varepsilon_j}; \sigma_{ij} = \sigma_i + \sigma_j$$

Here, r_{ij} is the distance between two atoms i and j. The energy parameters ε and σ are taken from the CHARMM19 force field of united atoms.[35] Since the energy terms originate from different sources, a set of weighting parameters is assigned in order to balance their respective contributions. These weighting coefficients are trained to recapitulate the native amino acid sequence of 38 high-resolution crystal structures.[33] The force field and the coefficients have been tested in various studies including experimental validations.[33,37−40]

New Atom Types and Parametrizations. The original Medusa force field was designed to model atom types occurring in proteins. Thus, to model small molecules for virtual screening, we extend the number of atom types as follows:

1. For chemical groups that already exist in proteins, we keep the same atom types and parameters.

2. For new chemical groups, we define atom types based on (i) element types and hybridization, (ii) nearest-neighbor heavy atom types, and (iii) second-nearest neighbor heavy atom if they are charged. For the charged groups, we only consider carboxyl, phosphate, and sulfone groups in current implementation.

Overall, we define 23 new atom type in addition to the 38 existing atom types (Supporting Information Table S1).

The VDW parameters are assigned according to atom sizes. For oxygen atoms, the VDW parameters are assigned the same as that of the OC atom type in EEF1,[31] if they are charged, and the same as OH, if they are not charged. The VDW parameters for nitrogen and sulfur atom types are the same as those used in CHARM19.[35] The VDW parameters for P, F, Cl, Br, and I are taken for Tripos TAFF force field. There are no new types for carbon atoms.

The extension of hydrogen bonding parameters is not needed since the parametrization of the model only depends on the hybridization types, on the explicit bond coordinates and on whether the hydrogen bonding is related to backbone.[36] When applying the model, we treat all ligand atoms as protein side chain atoms, since usually there are no secondary structure constraints for ligand molecules.

There are two key parameters to extend in the EEF1 model: (1) the total solvation energy of the fully solvated atom ΔG_{free} and (2) the solvent volume V the atoms excludes.[31] The volume parameters are assigned by considering the full volume of the atom and subtracting the overlap with neighboring bonded heavy atoms. The original ΔG_{free} parameters are taken from experimental solvation free energy measurements.[31] Since such measurements are not available for most small molecules, we assign these values according to their polarities. Following the original EEF1 model, we assign a large ΔG_{free} value of −20 kcal/mol for all charged groups. For other polar atoms, we use a linear relationship between the known partial charge and ΔG_{free} to assign the new ΔG_{free} values.

We use the same weights W_X as the original Medusa force field. Therefore, there is no additional training involved in force field parameters for developing the MedusaScore. The force field parameters are listed in Supporting Information Table S1.

Scoring Protocol. The scores are obtained from calculating the binding energy between the protein and the ligand using the extended Medusa force field. The protein coordinates are provided in a Protein DataBank (PDB) format and the small molecule coordinates in SYBYL mol2 format. Hydrogen atoms are required in the mol2 file to enable hybridization type assignment. All nonpolar hydrogen atoms are ignored in the protein input file since we use united-

atom model. All polar hydrogen atoms are first reconstructed based on geometric bond constraints at physiological pH. Subsequently, the alternative positions of the rotatable polar hydrogen atoms are searched to optimize the protein−ligand hydrogen binding. To save computational time, the optimization is performed only for residues that are within 4 Å of the binding interface. All the hydrogen atoms in ligands are kept fixed during the optimization.

The computer program for MedusaScore is written in C++, and the simulation is performed on an Intel P4 2.4G Hz workstation running Gentoo Linux. The typical CPU time needed for the evaluation of a single protein−ligand complex is 0.11 s, of which 0.10 s is spent on parsing the input files.

Data Sets. To benchmark our force field and scoring protocol, we use publicly available data sets that have been previously used to benchmark other scoring functions. This choice of a third party data set allows comparison of MedusaScore with other scoring functions in an unbiased way. For discerning docking decoys, we use the data set generated by Wang et al.,[14] which has been used to compare docking accuracy for 11 scoring functions, including LigScore, PLP, PMF, LUDI, F-Score, G-Score, D-Score, ChemScore, Autodock, DrugScore, and X-Score. For scoring, we use the protein pocket model in PDB format and the ligand coordinates in mol2 format.

For binding affinity prediction, we use the PDBBind database[32] (version 2005), which contains 1296 high quality complex structures (the refined set) and a subset of 288 nonredundant complexes (the core set). The latter contains structurally unrelated targets and ligand and thus is more challenging for scoring function testing. Since our current force field does not take metal atoms into account, we eliminate all complexes that contain metal atoms within 4 Å of the ligand molecules (219 complexes total, Zn atoms in the carbonic anhydrase represent about one-half of all cases). We also exclude complexes where a third molecule (mostly phosphate and sulfate) is bound to the same pocket, which corresponds to 16 complexes in the data set. Finally, there are 4 complexes (PDB ID: 1duv, 1nw5, 1r6n, 2adm) containing atom types not modeled in the current force field implementation, which are also eliminated. After this filtering step, our data set for benchmarking binding affinity prediction consists of 1057 complexes in the refined set and 243 complexes in the core set.

RESULTS AND DISCUSSION

Docking Decoy Recognition. We apply the MedusaScore for the docking decoys of 100 protein−ligand complexes compiled by Wang et al.[14] Following the same criterion used by the authors,[14] we define a successful recognition when the best scoring (lowest binding energy) ligand decoy has a root-mean-square deviation (rmsd) less than 2 Å from the crystal structure (the rmsd is calculated only for ligand coordinates). We find that MedusaScore successfully recognizes the native-like poses for 82 of the entire 100 complexes, i.e., the success rate of MedusaScore is 82%. This success rate is higher than that for eleven other scoring functions studied by Wang et al.[14] (see Table 1). The success rates of these scoring functions vary from 26% (DSCORE) to 76% (PLP). Interestingly, the force field-based scoring functions generally feature lower success rates than the other types of functions (see Table 1). The inclusion of EEF1 solvation model and the explicit hydrogen bond model is likely to contribute to the improved accuracy of the MedusaScore. An example of using MedusaScore in docking decoy recognition is shown in Figure 1 for a thermolysin inhibitor (PDB ID: 1tlp). From the 100 docking decoys generated using Autodock, MedusaScore correctly recognizes the native-pose, while all other 11 scoring functions mistakenly picked non-native-like poses (rmsd > 6 Å).

We further test the consensus scoring using MedusaScore and the other 11 scoring functions. We use a "rank-by-rank" strategy,[41] where the ranking of the decoys using both MedusaScore and other scoring functions are calculated and the decoy with the highest average rank is selected as the

Table 1. Success Rates of MedusaScore and Other 11 Scoring Functions for Docking Decoy Recognition[a]

scoring function		success rate	
type	name	single scoring	consensus with Medusa
force field	*MedusaScore*	*82*	-
	Autodock	61	71
	GScore	42	74
	DScore	26	72
empirical	PLP	76	83
	FScore	74	79
	LigScore	74	74
	LUDI	67	82
	X-Score	65	78
	ChemScore	35	60
statistical potential	DrugScore	71	85
	PMF	52	70

[a] The scoring functions are tested on docking decoy data set consisting of 100 complexes.[14] The success rate is defined when the best scoring decoy ligand has rmsd less than 2 Å from the crystal structure. The MedusaScore has the highest success rate of all the tested functions (82%). Consensus scoring with the MedusaScore improves other scoring functions. The combination of the MedusaScore with PLP or with the DrugScore gives a success rate of 83% and 85%, respectively, higher than the MedusaScore alone.

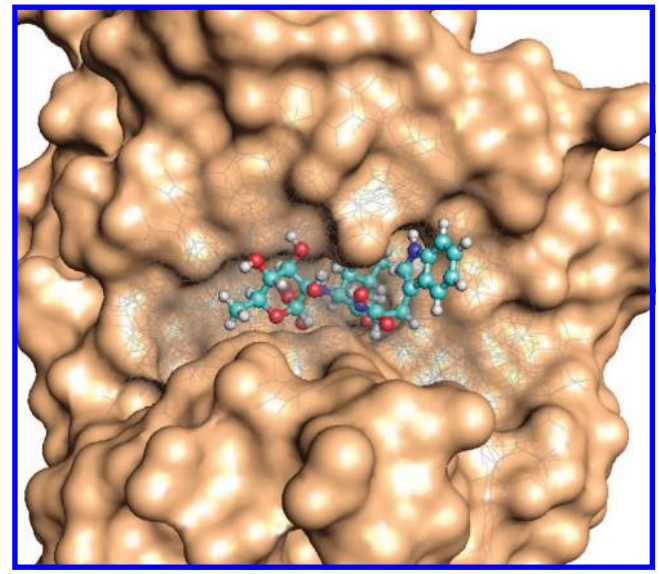

Figure 1. An example of docking decoy recognition using MedusaScore. The native docking pose (ball-and-stick) for an inhibitor against thermolysin protein (PDB ID: 1tlp) is correctly recognized using MedusaScore from the 100 docking decoys (gray lines) generated by Wang et al.[14] using Autodock.[1]

best scoring decoy. We find that consensus score with MedusaScore improves the decoy recognition rate for almost all other scoring functions except LIGSCORE, where the consensus scoring with MedusaScore has the same success rate as that of LIGSCORE alone.

On the other hand, we find that in most cases, consensus scoring decreases MedusaScore performance. The only exception is when MedusaScore is combined with PLP or DrugScore where the success rates are 83% and 85%, respectively. These success rates are higher than using MedusaScore alone (82%). We attribute this improvement to the fact that PLP and DrugScore have significantly different energy potentials from ours. For example, DrugScore's energy function comprises distance-dependent pairwise statistical potential and solvent-accessible surface dependent potential. None of these two potentials overlaps with the energy terms of MedusaScore. Due to these differences in the force field construction, some energy contributions that are ignored in MedusaScore may be realized in PLP or DrugScore, thereby improving the accuracy. Another important fact is that both PLP and DrugScore also feature relatively high success rates. The MedusaScore is less likely to improve by consensus scoring with scoring functions that exhibit low recognition accuracies themselves. We also find that consensus scorings of MedusaScore with other force field-based scoring functions in general have lower success rates (Table 1). This observation likewise can be attributed to lacking of complementarity between MedusaScore and other force field-based scoring functions.

Consensus using 3 scoring functions does not further improve the success rate. We find the highest consensus scorings have a success rate of 85%, which are achieved from consensus score using MedusaScore, DrugScore and one of the third scoring functions (FSCORE, LUDI, or HMSScore, which is a scoring protocol from X-Score). Therefore, the consensus accuracy seems to be ultimately limited by the inherent inaccuracy of the individual scoring functions, and no extra improvement can be obtained by further combination of the scoring functions.

Binding Affinity Prediction. Using the PDBBind database, we test the binding affinity prediction accuracy of MedusaScore. The correlation coefficient between the MedusaScore and the experimental dissociation constant (pK_d) is 0.55 for the refined set and 0.56 for the core set. If we exclude the repulsive part of the van der Waals energies (VDWR) from the total score, the correlation improves to 0.61 and 0.60 for the refined set and core set, respectively (Figure 2). Similar observations have been reported in earlier studies.[12] This improved correlation by excluding VDWR may be due to clashes in some of the complex structures in the PDBBind 2005 database. Besides subtracting the van der Waals repulsion directly, energy minimization of the complex structures should have the same effect. For simplicity, in the following analysis we only use the VDWR excluded MedusaScore.

Benchmarking studies have been reported[15] on an earlier version of PDBBind database (version 2002) for 14 scoring function including X-Score,[26] DrugScore,[25] D-Score,[11] PMF-Score,[24] G-Score,[9] ChemScore,[22] F-Score,[8] LigScore, PLP,[7] PMF,[24] LUDI,[21] GoldScore,[9] and HINT.[42] The highest correlation coefficient is 0.566 using X-Score, lower than

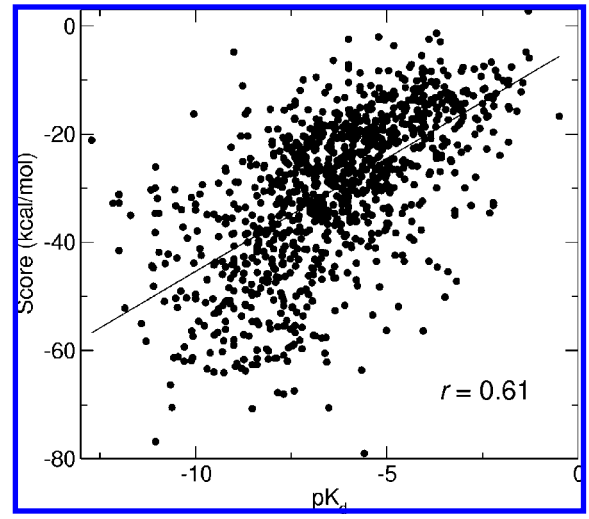

Figure 2. Scatter plot of the MedusaScore predictions (without VDWR) vs the experimental dissociation constant pK_d for the refined set. The Pearson correlation coefficient is 0.61. The solid line corresponds to a linear regression fit ($y = 4.18x - 3.59$).

what we obtained using MedusaScore, albeit with a newer PDBBind database (version 2005).[32] To make a more objective comparison, we also calculate the MedusaScore using the 2002 version of the PDBBind Database. Similarly, we also exclude the complexes with metal or other heterogeneous atoms near the binding pocket. This procedure eliminates 181 out of the 800 complexes in the refined set. Using the VDWR-excluded scoring protocol, we find a correlation of 0.63 between the MedusaScore and the experimental binding affinity. This prediction accuracy is also significantly higher than that of the other 14 scoring functions that have been tested on the same data set.

Although we do not use any binding affinity data for parameter training, MedusaScore still predicts the experimental values with reasonable accuracy. The robust performance over the various data sets suggests that MedusaScore likely grasps the crucial energetic component of protein−ligand binding. Since there is no training involved, MedusaScore should be applicable to a wide range of targets and ligands beyond those that have been tested in this study.

Size Dependence. We further examine if the performance of MedusaScore depends on the ligand size. It has been reported that scoring functions tend to predict higher binding affinity for larger ligands, due to the inherent "stickiness" of the molecules.[43] To avoid any bias in the data analysis, we use the core set (See Methods) because it contains diverse structures. We divide the core set into three subsets, based on the total number of heavy atoms (n) of the ligand molecule. The small, medium, and large ligand size subsets correspond to $6 \le n \le 19$, $20 \le n \le 29$, and $30 \le n \le 70$, respectively, and contain 83, 82, and 78 complexes, respectively. We find that for the subset of small ligands, our scoring function features the best correlation with experimental data−the Pearson correlation coefficient is 0.63, higher than the average correlation over the whole data set (Figure 3). The correlation decreases for larger ligands−the correlation coefficients for the medium and large ligands are 0.52 and 0.37, respectively, lower than the average over the whole data set. In general, our observations agree with the previously described tendency of larger molecules to have lower scores.[43]

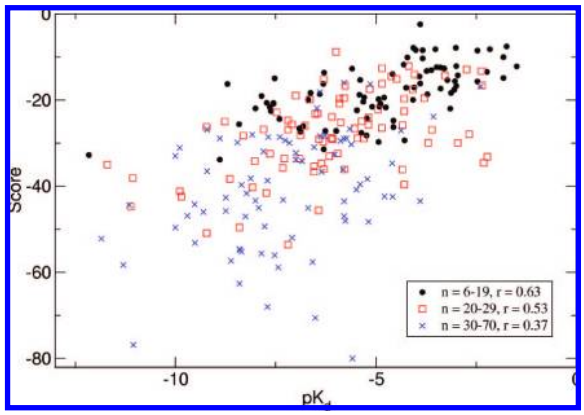

Figure 3. Scatter plot of the MedusaScore vs the experimental dissociation constant pK_d for the core set, categorized based on the number of heavy atoms in the ligand (n). The prediction accuracy is higher ($r = 0.63$) for small ligands than for medium and large sized ligands.

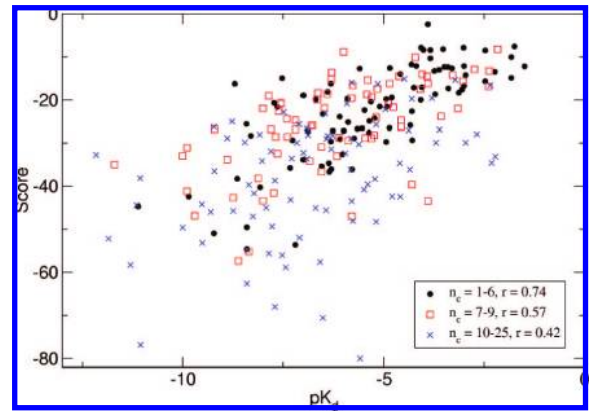

Figure 4. Scatter plot of the MedusaScore vs the experimental dissociation constant pK_d for the core set, categorized based on the number of ligand heavy atoms in contact with protein (n_c). We define a contact when a ligand heavy atom is within 3.5 Å of any protein heavy atom. The prediction accuracy is higher ($r = 0.74$) for ligands with fewer contacts with the proteins than for those with more extensive contact.

When we divide the data set according to the number of heavy atoms of the ligand that are in contact with the protein (n_c), we find a better correlation between MedusaScore estimates and the experimental data. The three subsets, defined by $1 \leq n_c \leq 6$, $7 \leq n_c \leq 9$, and $10 \leq n_c \leq 25$, contain 88, 72, and 83 protein−ligand complexes, respectively. The contact is defined when the ligand atom is within 3.5 Å of any protein heavy atom. We find that our scoring function is most accurate for the subset with the least contact atoms ($1 \leq n_c \leq 6$), for which the correlation coefficient is 0.74. The correlation decreases to 0.57 and 0.42 for the subsets of $7 \leq n_c \leq 9$ and $10 n_c \leq 25$, respectively. Similarly to the previous case, we observe systematic overestimation of the binding affinity for ligands with larger n_c.

Clearly MedusaScore can be utilized more confidently for smaller-sized ligands or ligands with a smaller number of contacts with proteins. This is partly due to the inaccuracies in the force field, which tend to accumulate with the increase of the ligand size or the number of contacts with the proteins. However, these force field inaccuracies alone cannot explain the systematic overestimation of binding affinities for ligands with larger n or n_c The experimental binding affinities have less significant dependence on n or n_c as shown in Figures 3 and 4, because the favorable binding enthalpies for larger molecules are often compensated by the larger entropic loss upon binding.[44] This enthalpy−entropy compensation effects are not considered in MedusaScore, which explains the overestimation of binding affinity.

It has been suggested that entropic contribution may play an important role in ligand binding and that entropic contribution can be as large as the enthalpy[45] and therefore may need to be properly evaluated to achieve a better accuracy in protein−ligand binding scoring. Furthermore, close contacts introduce significant fluctuations in the total energy due to the sensitivity/volatility of the VDWR energy term. These factors might be important for further MedusaScore improvement.

VDWR Terms and Energy Minimization. We have shown that excluding the VDWR term from MedusaScore in general results in slightly improved binding affinity prediction. We further test whether this observation is due to clashes in the molecular complexes and whether energy minimization can improve the prediction accuracy. We examine the VDWR energies in the core set and find that there are five complexes having exceptionally large VDWR energies (>25 kcal/mol), compared with the rest having an average VDWR energy of 4.7 kcal/mol and a standard deviation of 8.9 kcal/mol. After excluding these outliers from the core set, the correlation coefficient between the MedusaScore prediction (with all energy terms) and experimental binding affinity is 0.59. Further removal of the VDWR term only marginally improves the correlation to 0.60. We attempt to minimize the structures by allowing local rigid body movement of the ligand and side chain movement of the protein. Such relaxation reduces VDWR energies in general (by ∼3−6 kcal/mol, see Supporting Information Table S2) but fails to reduce the high VDWR energies for the outliners effectively. As a result, the overall correlation is not improved after the minimization.

Therefore, our results demonstrate that without performing explicit minimization, removing the VDWR term, in fact, makes the scoring function more robust to structure imperfections. This strategy might be necessary for virtual screening because it saves costly computational time for structure minimizations. Unphysically close atom contacts in the complex structures can be detected by applying a quick atomic distance filter before using the scoring function.

Comparison with RosettaLigand. It is interesting to compare the performance of MedusaScore with the scoring function of RosettaLigand[12] since the underlying force fields behind the two scoring functions have similar energy terms. To our best knowledge, the only published benchmark results for RosettaLigand are in the original publication,[12] where the RosettaLigand's scoring function was used for binding affinity prediction on the LPDB database.[46] The authors found a correlation of 0.63 between the experimental and predicted binding affinities.

In order to compare the performance of MedusaScore with RosettaLigand, we test the performance of MedusaScore using the LPDB database. The current LPDB database contains 262 complexes. Following the same protocol as described in Methods, we eliminate complexes with unsupported atom types or having metal and other additional hetero molecules near the binding interface. After the preprocessing,

we obtain 196 complexes for benchmarking. Applying MedusaScore on this data set results in correlations of 0.63 and 0.64 with and without the VDWR term, respectively. These correlations are comparable with the previously published results from RosettaLigand.[12]

Although RosettaLigand and MedusaScore feature similar accuracy in terms of binding affinity prediction, there are several notable differences between the two scoring functions: (1) RosettaLigand uses Tripos mol2 based atom types for ligands, while MedusaScore uses extension of CHARMM19 atom types. (2) The weighing coefficients have been retrained in RosettaLigands using 100 protein−ligand structures,[12] while we do not retrain any weighing coefficient in MedusaScore.

CONCLUSION

We have developed a scoring function for protein−ligand interaction by extending the Medusa force field and design suite. Benchmarking using available data sets shows superior performance of the MedusaScore for both docking decoy recognition and binding affinity prediction. Since the MedusaScore does not rely on parameter training using protein−ligand binding data, it is transferable to targets and small molecules beyond the tested data sets. Therefore, we expect the MedusaScore to have wide application in virtual screening of novel small molecule drug candidates.

ACKNOWLEDGMENT

The authors thank Dr. Feng Ding for stimulating discussions. This work is supported by the National Institutes of Health (Grant No. RO1-GM080742).

Supporting Information Available: Force field parameters and the MedusaScore benchmark results on the PDBBind core sets. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hughes, B. 2007 FDA Drug Approvals: A year of Flux. *Nat. Rev. Drug Discovery* **2008**, *7*, 107–109.
(2) Klebe, G. Recent Developments in Structure-Based Drug Design. *J. Mol. Med.* **2000**, *78*, 269–281.
(3) Schneider, G.; Fechner, U. Computer-Based de novo Design of Drug-Like Molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649–663.
(4) Schneider, G.; Bohm, H. J. Virtual Screening and Fast Automated Docking Methods. *Drug Discovery Today* **2002**, *7*, 64–70.
(5) Brooijmans, N.; Kuntz, I. D. Molecular Recognition and Docking Algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–373.
(6) Abagyan, R.; Totrov, M. et al. ICM - A New Method for Protein Modeling and Design - Applications to Docking and Structure Prediction from the Distorted Native Conformation. *J. Comput.-Aided Mol. Des.* **1994**, *15*, 488–506.
(7) Gehlhaar, D. K.; Verkhivker, G. M. et al. Molecular Recognition of the Inhibitor Ag-1343 by Hiv-1 Protease - Conformationally Flexible Docking by Evolutionary Programming. *Chem. Biol.* **1995**, *2*, 317–324.
(8) Rarey, M.; Kramer, B. et al. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.
(9) Jones, G.; Willett, P. et al. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727–748.
(10) Morris, G. M.; Goodsell, D. S. et al. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
(11) Ewing, T. J. A.; Makino, S. et al. DOCK 4.0: Search Strategies For Automated Molecular Docking of Flexible Molecule Databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411–428.
(12) Meiler, J.; Baker, D. ROSETTALIGAND: Protein-Small Molecule Docking with Full Side-Chain Flexibility. *Proteins: Struct., Funct., Bioinform.* **2006**, *65*, 538–548.
(13) Kuntz, I. D.; Blaney, J. M. et al. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.
(14) Wang, R. X.; Lu, Y. P. et al. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287–2303.
(15) Wang, R. X.; Lu, Y. P. et al. An Extensive Test of 14 Scoring Functions Using the PDBbind Refined Set of 800 Protein-Ligand Complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114–2125.
(16) Warren, G. L.; Andrews, C. W. et al. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
(17) Beveridge, D. L.; Dicapua, F. M. Free-Energy Via Molecular Simulation - Applications to Chemical and Biomolecular Systems. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431–492.
(18) Kollman, P. Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93*, 2395–2417.
(19) Bash, P. A.; Singh, U. C. et al. Free-Energy Calculations by Computer-Simulation. *Science* **1987**, *236*, 564–568.
(20) Dang, L. X.; Merz, K. M. et al. Free-Energy Calculations on Protein Stability - Thr-157-] Val-157 Mutation of T4 Lysozyme. *J. Am. Chem. Soc.* **1989**, *111*, 8505–8508.
(21) Bohm, H. J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein-Ligand Complex of Known Three-Dimensional Structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243–256.
(22) Eldridge, M. D.; Murray, C. W. et al. Empirical Scoring Functions 0.1. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.
(23) Wang, R. X.; Liu, L. et al. SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein-Ligand Complex. *J. Mol. Model.* **1998**, *4*, 379–394.
(24) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein-Ligand Interactions: a Simplified Potential Approach. *J. Med. Chem.* **1999**, *42*, 791–804.
(25) Gohlke, H.; Hendlich, M. et al. Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.
(26) Wang, R. X.; Lai, L. H. et al. Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11–26.
(27) Zhang, S.; Golbraikh, A. et al. Development of Quantitative Structure-Binding Affinity Relationship Models Based on Novel Geometrical Chemical Descriptors of the Protein-Ligand Interfaces. *J. Med. Chem.* **2006**, *49*, 2713–2724.
(28) Gohlke, H.; Klebe, G. Statistical Potentials and Scoring Functions Applied to Protein-Ligand Binding. *Curr. Opin. Struct. Biol.* **2001**, *11*, 231–235.
(29) Golbraikh, A.; Tropsha, A. Beware of q(2)! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.
(30) Kortemme, T.; Morozov, A. V. et al. An Orientation-Dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein-Protein Complexes. *J. Mol. Biol.* **2003**, *326*, 1239–1259.
(31) Lazaridis, T.; Karplus, M. Effective Energy Function for Proteins in Solution. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 133–152.
(32) Wang, R. X.; Fang, X. L. et al. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.
(33) Ding, F.; Dokholyan, N. V. Emergence of Protein Fold Families Through Rational Design. *PLoS. Comput. Biol* **2006**, *2*, e85.
(34) Kuhlman, B.; Dantas, G. et al. Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003**, *302*, 1364–1368.
(35) Brooks, B. R.; Bruccoleri, R. E. et al. Charmm - A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
(36) Kortemme, T.; Baker, D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14116–14121.
(37) Yin, S.; Ding, F. et al. Modeling backbone flexibility improves protein stability estimation. *Structure* **2007**, *15*, 1567–1576.
(38) Yin, S. Y.; Ding, F. et al. Eris: An automated estimator of protein stability. *Nat. Methods* **2007**, *4*, 466–467.
(39) Hao J.; Serohijos A. W. R. et al. Identification and Rational Redesign of Peptide Ligands to CRIP1, A Novel Biomarker for Cancers. *Public Library Sci. Comput. Biol.* 2008, in press.
(40) Ding, F.; Tsao D.; et al. Ab Initio Folding of Proteins Using All-Atom Discrete Molecular Dynamics. *Structure* 2008, in press.
(41) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422–1426.

(42) Cozzini, P.; Fornabaio, M. et al. Simple, Intuitive Calculations of Free energy of Binding for Protein-Ligand Complexes. 1. Models Without Explicit Constrained Water. *J. Med. Chem.* **2002**, *45*, 2469–2483.
(43) Coupez, B.; Lewis, R. A. Docking and Scoring-Theoretically Easy, Practically Impossible. *Curr. Med. Chem.* **2006**, *13*, 2995–3003.
(44) Dunitz, J. D. Win Some, Lose Some: Enthalpy-Entropy Compensation in Weak Intermolecular Interactions. *Chem. Biol.* **1995**, *2*, 709–712.
(45) Chang, C. E. A.; Chen, W. et al. Ligand Configurational Entropy and Protein Binding. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1534–1539.
(46) Roche, O.; Kiyama, R. et al. Ligand-protein database: linking protein-ligand complex structures to binding data. *J. Med. Chem.* **2001**, *44*, 3592–3598.

CI8001167