J|A|C|S
COMMUNICATIONS

Published on Web 01/09/2004

A Quantum Mechanics-Based Scoring Function: Study of Zinc Ion-Mediated Ligand Binding

Kaushik Raha and Kenneth M. Merz, Jr.*

152 Davey Laboratory, Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802-6300

Received September 12, 2003; E-mail: merz@psu.edu

The in silico "scoring" of protein−ligand interactions has been an area of intense research because of its potential impact on rational drug discovery and design. Because of the relentless pressure on the pharmaceutical industry to reduce drug discovery costs,[1] in silico structure-based screening is viewed as a very attractive and cost-effective alternative to traditional medicinal chemistry approaches. A number of methods can "dock" a small molecule into a biological receptor, and these include, for example, DOCK,[2] AutoDock,[3] FlexX,[4] and GOLD.[5] Docking is based on the principles of molecular recognition, and these programs sample the conformational space of a small molecule and "pose" them in the active site of the protein. To a large extent, the pose generation problem is solved, while the prediction of the free energy of binding (ΔG_{bind}) or "scoring" of the poses has proven to be a major challenge. Despite recent developments in this area, a physically based scoring function (SF) that is robust enough to evaluate the binding of ligands to proteins has been elusive.

Current-generation SFs can be grouped into three categories: empirical scoring functions (ESF), knowledge-based potentials (KBP), and force field (FF) methods. ESFs use empirically derived energetic contributions related to enthalpy (ΔH) and entropy (ΔS) and regression methods to fit it to a set of experimental observations. The problems with ESFs are that they can only be as discriminating as the overall potential function allows and they depend on the diversity of the training set.[6,7] In KBPs, ΔG_{bind} is represented as a potential of mean force calculated from frequencies of interatomic contacts from a database of structures.[8] Recently, KBPs have been shown to be successful in predicting binding affinities.[9] However, the accuracy of KBPs depends on the proper definition of the reference state and on the number of structures available to build the potential.[9,10] FF-based methods use potentials such as AMBER,[11] CHARMM,[12] OPLS,[13] and MMFF[14] to score poses, and they have been used frequently in free-energy perturbation (FEP) methods to evaluate relative ΔG_{bind}.[15] FF models have been extremely powerful in modeling biological systems but generally use simple electrostatic models (Coulomb potentials), which limit their ability to accurately model electrostatic energies. However within the FF framework, conformational sampling has been used in LIE[16,17] and MM/PBSA[18] methods. While all classes of SFs have shown success, none have been able to accurately and broadly score protein−ligand interactions. Moreover, metal-containing systems pose a challenge for these models due to the nature of the interactions between a small molecule and a metal ion in the active site. Indeed, many SFs deal with metals by ignoring them entirely in the scoring process. For example, the $\log_{10}$ value of atom pair occurrence of zinc and other heavy atom pairs is so low that the interaction is ignored in the KBP function PMF.[8] In the potentials that do explicitly model metal ions, charge-transfer (CT) interactions between the metal ion and the ligand (which reduces the charge on the ligand) are generally not accounted for.

Quantum mechanics (QM), although not new to the field of molecular interactions, has until now been used only to study smaller systems because of the computational cost associated with it. In recent work, our group has described the linear scaling divide and conquer (D&C) approach in conjunction with semiempirical Hamiltonians that can be used to study large molecular systems at the QM level.[19] We have also coupled this to a Poisson−Boltzmann (PB)-based self-consistent reaction field method (SCRF) for calculating solvation free energies.[20] The use of QM allows us to move away from FFs, especially when evaluating electrostatic interactions. FFs generally ignore QM effects such as polarization and CT.[21] In the first study of its kind, we report the use of QM for scoring protein−ligand interactions. We note that none of the previous studies have treated the complete protein−ligand complexes at such high levels of theory for ΔG_{bind} prediction. Herein, we briefly describe our method and present the results of its application to a set of 18 carbonic anhydrase (CA) inhibitors and 5 carboxypeptidase (CPA) inhibitors.

The 18 CA and 5 CPA complexes were downloaded from the Protein Data Bank (PDB).[22] These inhibitors and the experimental ΔG_{bind} are listed in the Supporting Information. Protons were added to heavy atoms of the protein using the LEAP module of AMBER 5.0.[23] Energy minimization was performed using constraints to relax the added protons using AMBER 5.0. All heavy atoms were fixed at the experimental coordinates during energy minimization. The active site of the uncomplexed protein was modeled with a zinc-bound water molecule. The interaction energy was calculated using the following thermodynamic cycle:

$$\Delta G_{\mathrm{bind}} = \Delta G_{\mathrm{b}}^{\mathrm{g}} + \Delta G_{\mathrm{solv}}^{\mathrm{protein-ligand}} - \Delta G_{\mathrm{solv}}^{\mathrm{protein}} - \Delta G_{\mathrm{solv}}^{\mathrm{ligand}}$$

$$\Delta G_{\mathrm{b}}^{\mathrm{g}} = \Delta H_{\mathrm{b}}^{\mathrm{g}} - T\,\Delta S_{\mathrm{b}}^{\mathrm{g}}$$

$$\Delta H_{\mathrm{b}}^{\mathrm{g}} = \Delta H_{\mathrm{f}} + (1/R^{6})LJ$$

$$\Delta S_{\mathrm{g}} = \Delta SA_{\mathrm{C,N,O,S}} + \mathrm{num(rot\_bonds)}$$

The solution-phase ΔG_{bind} was decomposed into the gas-phase interaction energy and solvation free energy. The gas phase $\Delta G_{\mathrm{b}}^{\mathrm{g}}$ is the sum of $\Delta H_{\mathrm{b}}^{\mathrm{g}}$ and $\Delta S_{\mathrm{b}}^{\mathrm{g}}$. The QM part of $\Delta H_{\mathrm{b}}^{\mathrm{g}}$ was calculated using DivCon[24] at the AM1[25] level as ΔH_{f}[complex] − ΔH_{f}[protein] − ΔH_{f}[ligand]. The dispersive part of the nonpolar interaction was calculated using the attractive part of the Lennard-Jones potential using the AMBER 96 force field.[11] The entropic contribution to binding was described by a solvent and conformational component. The solvent entropy, which is gained by water, on being displaced from the active site by the ligand during binding[26] was estimated based on the surface area burial for carbon, oxygen, nitrogen, and sulfur atoms during binding. The conformational entropy was evaluated from the number of degrees of freedom that was lost in the small molecule and side chains in the active site during

10.1021/ja038496i CCC: $27.50 © 2004 American Chemical Society

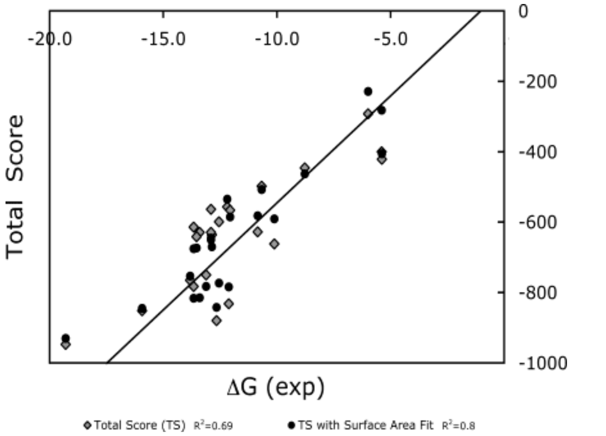

Figure 1. Plot of calculated total score versus the ΔG (exp) for the set of 23 complexes. (◆) Sum of the individual contributions from eq 1. The square of the correlation coefficient R^2 is 0.69. (●) Surface areas fitted against ΔG (exp) for the set of 23 complexes. The square of the correlation coefficient R^2 is 0.8.

binding.[2,7,9,27−30] We used a value of 1.0 kcal/mol for the loss of a rotatable bond. The free energy of solvation is well-known to play an important role in binding,[31−34] and we evaluated it using our PB/SCRF method.[20] CM2 charges[35] were used in the PB/SCRF calculation. Thus, the final total score is a sum of $\Delta G_{\mathrm{b}}^{\mathrm{g}}$ and $(\Delta G_{\mathrm{solv}}^{\mathrm{protein-ligand}} - \Delta G_{\mathrm{solv}}^{\mathrm{protein}} - \Delta G_{\mathrm{solv}}^{\mathrm{ligand}})$.

Figure 1 is a plot of experimental binding free energy (ΔG_{exp}) calculated as $-RT \ln K_i$ versus the calculated total score. The predictive capability of the method for scoring these inhibitors is shown in this plot. We achieve good agreement with ΔG_{exp} with a square of the correlation coefficient R^2 of 0.69 ($R = 0.83$) for this set of inhibitors. We calculated the standard error as:

$$\text{Std. Error} = \sqrt{(\Delta G_{\mathrm{exp}} - (mTS + c))^2}$$

where $mTS + c$ is the equation of the line fitted to ΔG_{exp} and TS is the total score. The mean standard error for this set is 1.5 kcal/mol. We note that we obtain an R^2 of 0.69 *without fitting any* of the contributions of the total score. Nonetheless, the solvent entropy term offers further opportunity for refinement since it is simply a difference of surface areas and not an energy term. We derive cross-validated parameters for the surface areas of atom types carbon, oxygen, nitrogen, and sulfur by maximizing the R^2 between the total score and the ΔG_{exp}. Applying the final set of surface area parameters results in an R^2 of 0.8 ($R = 0.90$) (Figure 1). The mean standard error in this case is 1.18 kcal/mol. We note that only the surface area dependent part of the total score has been fit without fitting any of the other terms. The difference in the magnitude of the total score and ΔG_{exp} arise from unweighted contributions to the total score. Fitting weighted contributions to ΔG_{exp} corrects this discrepancy and results in even higher R^2 (data not shown). In previous studies, we have reported on polarization and CT effects in biomolecular systems.[21] There is a significant amount of CT in these cases as well as between the ligand and the zinc ion in the active site. In the case of CA, the amount of CT is on the order of 1e, while in CPA, which has a carboxylate ligand, the amount of CT is reduced to about 0.5e (Table 2 in the Supporting Information). Recent studies have also pointed out the importance of polarization in binding.[36] The issue of sampling conformational microstates has not been addressed in this study but will be explored in future studies. We note that including sampling in other studies has resulted in improving prediction accuracy even further.[16,18]

In this paper, we have demonstrated the ability of a QM method to score known protein−ligand poses. In particular, we have focused on two classes of zinc metalloenzymes, which would be difficult to accurately model using ESFs, KBPs, or FFs because of the presence of the metal ion. As expected, we found that there is significant and variable metal−ligand CT among different families, a phenomenon that is difficult to capture using simpler scoring functions. Finally, by including two different families of proteins we show that this method has the ability to perform across different families.

Acknowledgment. We thank the NCSA for supercomputer time and the NSF via Grant MCB-0211639 for financial support.

Supporting Information Available: Tables of proteins with corresponding inhibitors and charge transfer values from proteins to inhibitors for 23 complexes (PDF). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Drews, J. *Science* **2000**, *287*, 1960−1964.
(2) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 380−397.
(3) Goodsell, D. S.; Olson, A. J. *Proteins* **1990**, *8*, 195−202.
(4) Kramer, B.; Rarey, M.; Lengauer, T. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 228−241.
(5) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727−748.
(6) Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309−323.
(7) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849−857.
(8) Muegge, I.; Martin, Y. C. *J. Med. Chem.* **1999**, *42*, 791−804.
(9) Ishchenko, A. V.; Shakhnovich, E. I. *J. Med. Chem.* **2002**, *43*, 2770−2780.
(10) Muegge, I. *Perspect. Drug Discovery Des.* **2000**, *20*, 99−114.
(11) Cornell, W. D.; Cieplak, P.; Baylay, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(12) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 182−217.
(13) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(14) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490−519.
(15) Kollman, P. A. *Chem. Rev.* **1993**, *7*, 2395−2417.
(16) Aqvist, J.; Luzhkov, V. B.; Brandsdal, B. O. *Acc. Chem. Res.* **2002**, *35*, 358−365.
(17) Carlson, H. A.; Jorgensen, W. L. *J. Phys. Chem.* **1995**, *99*, 10667−10673.
(18) Lin, J.-H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. *J. Am. Chem. Soc.* **2002**, *124*, 5632−5633.
(19) Dixon, S. L.; Merz, K. M., Jr. *J. Chem. Phys.* **1996**, *104*, 6643−6649.
(20) Gogonea, V.; Merz, K. M., Jr. *J. Phys. Chem. A* **1999**, *103* (26), 5171−5188.
(21) van der Vaart, A.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **1999**, *121*, 9182−9190.
(22) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(23) Case, D. A.; Caldwell, J. W.; Cheatham, T. E., II; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Ferguson, D. M.; Radmer, R. J.; Seibel, G. L.; Singh, U. C.; Kollman, P. A. *AMBER*, version 5.0; University of California: San Francisco, CA, 1997.
(24) Dixon, S. L.; van der Vaart, A.; Gogonea, V.; Vincent, J. J.; Brothers, E. N.; Suárez, D.; Westerhoff, L. M.; Merz, K. M., Jr. *DivCon*; The Pennsylvania State University: University Park, PA, 1999.
(25) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(26) Bardi, J. S.; Luque, I.; Freire, E. *Biochemistry* **1997**, *36*, 6588−6596.
(27) Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.
(28) DeWitte, R. S.; Shakhnovich, E. I. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11744.
(29) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(30) Gohlke, H.; Hendlich, M.; Klebe, G. *Perspect. Drug Discovery Des.* **2000**, *20*, 115−144.
(31) Arora, N.; Bashford, D. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 12−27.
(32) Polticelli, F.; Ascenzi, P.; Bolognesi, M.; Honig, B. *Protein Sci.* **1999**, *8*, 2621−2629.
(33) Hunenberger, P. H.; Helms, V.; Narayana, N.; Taylor, S. S.; McCammon, J. A. *Biochemistry* **1999**, *38*, 2358−2366.
(34) Schwarzl, S. M.; Tschopp, T. B.; Smith, J. C.; Fischer, S. *J. Comput. Chem.* **2002**, *23*.
(35) Li, J. B.; Zhu, T. H.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1998**, *102*, 1820−1831.
(36) Garcia-Viloca, M.; Truhlar, D. G.; Gao, J. *J. Mol. Biol.* **2003**, *372*, 549−560.

JA038496I