34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

A Multibody Atomic Statistical Potential for the Prediction of Enzyme-Inhibitor Binding Energy

Majid Masso, *Member, IEEE*

***Abstract*— Accurate prediction of enzyme-inhibitor binding energy has the capacity to speed drug design and chemical genomics efforts by helping to narrow the focus of experiments. Here a non-redundant set of three hundred high-resolution crystallographic enzyme-inhibitor structures was compiled for analysis, complexes with known binding energies (ΔG) based on the availability of experimentally determined inhibition constants (k_i). Additionally, a separate set of over 1400 diverse high-resolution macromolecular crystal structures was collected for the purpose of creating an all-atom knowledge-based statistical potential, via application of the Delaunay tessellation computational geometry technique. Next, two hundred of the enzyme-inhibitor complexes were randomly selected to develop a model for predicting binding energy, first by tessellating structures of the complexes as well as the enzymes without their bound inhibitors, then by using the statistical potential to calculate a topological score for each structure tessellation. We derived as a predictor of binding energy an empirical linear function of the difference between topological scores for a complex and its isolated enzyme. A correlation coefficient (r) of 0.79 was obtained for the experimental and calculated ΔG values, with a standard error of 2.34 kcal/mol. Lastly, the model was evaluated with the held-out set of one hundred complexes, for which structure tessellations were performed in order to calculate topological score differences, and binding energy predictions were generated from the derived linear function. Calculated binding energies for the test data also compared well with their experimental counterparts, displaying a correlation coefficient of $r = 0.77$ with a standard error of 2.50 kcal/mol.**

I. Introduction

EXPERIMENTAL high-throughput screening (HTS) is an effective method for discovering small molecular or peptide inhibitors that tightly bind a target enzyme [1]. Though the HTS process has the potential to be time consuming and expensive, the emergence of computer-based virtual HTS (vHTS) has led to more efficient identification of drug candidates from among a large collection of compounds [2]. A variety of computational tools have also been developed for estimating enzyme-inhibitor binding energy based on scoring functions that take into account physicochemical interactions, including X-Score [3], Lig-Score [4], DrugScore [5], SFCscore [6], AutoDock4 [7], ITScore [8, 9], and PHOENIX [10]. Here we develop an empirical scoring function for predicting enzyme-inhibitor binding affinity, one that relies on an all-atom, four-body statistical potential derived by implementing the computational geometry technique of Delaunay tessellation. Our atomic potential compares well with other atomic energy functions [11, 12] in identifying the native structure as a global minimum, extensive work to be reported elsewhere.

Two separate datasets of non-redundant, high-resolution crystallographic structures were complied for this study. One set includes over 1400 single and multi-chain proteins, many with bound ligands (inhibitors, substrates, or cofactors), whose tessellations were used to quantify the relative frequencies of atomic quadruplet interactions for the purpose of deriving the four-body statistical potential. The second group of macromolecular structures consists of three hundred enzyme-inhibitor complexes with experimentally measured inhibition constants (k_i), from which we determined their respective binding energies (ΔG). Two-thirds of the latter structures were randomly selected to train an empirical model for calculating ΔG, first by tessellating the enzyme-inhibitor complexes as well as the enzymes without their bound inhibitors, then by using the four-body statistical potential to calculate a topological score for each tessellation, and finally by calculating the difference between topological scores for the complex and the isolated enzyme in each case. Based on these data, we derived a linear function of the topological score difference as a predictor of binding energy, a model evaluated by using the remaining set of one hundred complexes for testing.

II. Materials and Methods

A. Datasets

In order to develop the four-body statistical potential, we selected for Delaunay tessellation a non-redundant set of 1417 high-resolution (≤ 2.2Å) crystallographic structures with atomic coordinate files deposited in the Protein Data Bank (PDB) [13], which additionally include protein chains that share a low (< 30%) sequence identity (http://proteins.gmu.edu/automute/tessellatable1417.txt). Structural diversity is highlighted by the fact that both single- and multi-chain proteins are represented, a majority of which are also complexed to small molecular or peptide ligands. The coordinates of hydrogen atoms and water molecules in all files were removed prior to tessellation.

A separate set of PDB files for three hundred enzyme-inhibitor complexes was compiled from the Binding MOAD [14, 15] database to develop a predictive model of binding

M. Masso is with the Laboratory for Structural Bioinformatics, School of Systems Biology, George Mason University, Manassas, VA 20110 USA (phone: 703-257-5756; fax: 703-993-8976; e-mail: mmasso@ gmu.edu).

energy (http://proteins.gmu.edu/automute/MOAD300ki.txt). The database contains all high-resolution (≤ 2.5Å) crystallographic structures of protein-ligand complexes from the PDB, and where available, the structures are annotated with experimentally determined binding data extracted from the literature. A non-redundant Binding MOAD is also available to avoid the bias introduced by over-represented proteins in the PDB, which is obtained by clustering proteins into families of 90% sequence identity and selecting a single representative for each cluster. For our dataset, we chose enzyme-inhibitor complexes from among PDB structure files in the non-redundant Binding MOAD that were additionally annotated with experimentally determined inhibition constants (k_i), and these values are included in the text file available from the above link. We randomly selected two-thirds of the enzyme-inhibitor complexes to train our model, which was tested using the held-out data, and the subset to which each structure belongs is also identified in the text file.

B. Software and Performance Measures

Qhull [16] was used to implement the Delaunay tessellation algorithm, while graphical depictions of the tessellated structures were produced with Matlab (Version 7.0.1.24704 (R14) Service Pack 1). Molecular graphics were generated with the UCSF Chimera software package [17]. Lastly, ad hoc Perl codes were written as needed for the purposes of data formatting and analyses.

The equation

$$\Delta G = RT \ln(k_i) = 0.592 \times \ln(k_i)$$

was used for obtaining the binding energy (ΔG, in units of kcal/mol) from the inhibition constant (k_i) for each enzyme-inhibitor complex, where R = 1.986 $\times 10^{-3}$ kcal K^{-1} mol^{-1} is the gas constant and T = 298° K is the absolute temperature. We evaluated agreement between experimental (x_i) and predicted (y_i) binding energy with the correlation coefficient

$$r = \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2} \times \sqrt{\frac{1}{n}\sum (y_i - \bar{y})^2}},$$

the standard error of the predictions

$$SE = \sqrt{\left[\frac{1}{n(n-2)}\right]\left[n\sum y_i^{\,2} - \left(\sum y_i\right)^2 - \frac{\left[n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)\right]^2}{n\sum x_i^{\,2} - \left(\sum x_i\right)^2}\right]},$$

and the equation of the regression line.

III. Results and Discussion

A. Four-body Statistical Potential

A six-letter alphabet (C, N, O, S, M = all metals, X = all other non-metals) was used for labeling the atoms contained in the 1417 PDB structure files. Delaunay tessellations of the structures were obtained by supplying their atomic coordinates to the Qhull program, which treats the points as vertices and generates a three-dimensional convex hull of space-filling, non-overlapping, irregular tetrahedra. Edges longer than 8Å were removed from each structure tessellation prior to analysis to avoid false-positive atomic interactions, which is consistent with that used by other researchers to generate an atomic pair potential [18], while shorter than a 12Å cutoff we used previously to develop a coarser-grained amino acid four-body potential [19, 20]. Table I provides summary data regarding the atoms whose coordinates were supplied to Qhull, as well as the total number of tetrahedra generated by all the tessellations.

TABLE I
Summary Data for the 1417 PDB Structure Files

Atom Types	Count	Proportion
(carbon) C	3612988	0.633193
(nitrogen) N	969253	0.169866
(oxygen) O	1088410	0.190749
(sulfur) S	28502	0.004995
(all metals) M	2529	0.000443
(all other non-metals) X	4299	0.000754
Total atom count:	5705981	
Total tetrahedron count:	34504737	

The four vertices of each tetrahedral simplex in a tessellation objectively identify an interacting atomic quadruplet, and based on a six-letter alphabet, there are 126 distinct alternatives (Table II). For each type of atomic quadruplet (i,j,k,l), we calculated an observed relative frequency of occurrence f_{ijkl} based upon the proportion of tetrahedral simplices, from among those generated by all the structure tessellations, for which the quadruplet appears at the four vertices. A rate expected by chance for the quadruplet was determined from a multinomial reference distribution, given by

$$p_{ijkl} = \frac{4!}{\prod_{n=1}^{6}(t_n!)} \prod_{n=1}^{6} a_n^{t_n}, \text{where } \sum_{n=1}^{6} a_n = 1 \text{ and } \sum_{n=1}^{6} t_n = 4.$$

In the above formula, a_n represents the proportion of atoms from all tessellated structures that are of type n (Table I), and t_n is the number of occurrences of atom type n in the quadruplet. Applying the inverted Boltzmann principle [21], we used the score s_{ijkl} = log (f_{ijkl} / p_{ijkl}) to quantify an interaction propensity for the atomic quadruplet. The set of 126 atomic quadruplet types with their respective scores defines the four-body statistical potential (Table II).

B. Topological Scores

After tessellating the enzyme-inhibitor complexes, the atomic coordinates for the inhibitors were removed from the PDB files, and the modified structures of the enzymes without their inhibitors were also tessellated (Fig. 1). All hydrogen atoms and water molecules were excluded from

TABLE II
ATOMIC FOUR-BODY STATISTICAL POTENTIAL

Quad	Count	s_{ijkl}	Quad	Count	s_{ijkl}
CCCC	4015872	-0.140244	MMNS	363	3.720958
CCCM	1592	-0.989223	MMNX	0	--
CCCN	4025206	-0.169866	MMOO	306	2.315530
CCCO	6202159	-0.032467	MMOS	104	3.127729
CCCS	293157	0.224008	MMOX	3	2.409325
CCCX	2796	-0.975047	MMSS	254	5.398477
CCMM	132	0.908235	MMSX	2	3.815151
CCMN	3318	-0.575981	MMXX	0	--
CCMO	5325	-0.420893	MNNN	1030	0.535960
CCMS	2293	0.795108	MNNO	1128	0.047955
CCMX	15	-0.567697	MNNS	561	1.326526
CCNN	1797552	-0.124635	MNNX	5	0.098041
CCNO	8233136	0.184864	MNOO	3744	0.518626
CCNS	124653	-0.053081	MNOS	314	0.723107
CCNX	2007	-1.024729	MNOX	29	0.510083
CCOO	3366568	0.047161	MNSS	793	3.008398
CCOS	198630	0.098905	MNSX	5	1.328573
CCOX	4626	-0.712426	MNXX	9	2.706383
CCSS	15288	0.868158	MOOO	5430	1.106856
CCSX	144	-0.637352	MOOS	156	0.669977
CCXX	143	0.482159	MOOX	168	1.523669
CMMM	23	3.480397	MOSS	210	2.380989
CMMN	144	1.216422	MOSX	4	1.181307
CMMO	256	1.415945	MOXX	55	3.442148
CMMS	662	3.410480	MSSS	62	3.910199
CMMX	1	1.411130	MSSX	2	2.763224
CMNN	2474	-0.132029	MSXX	0	--
CMNO	6267	-0.079754	MXXX	16	5.786451
CMNS	2588	1.118068	NNNN	3878	-0.869698
CMNX	26	-0.058415	NNNO	46665	-0.441730
CMOO	8481	0.302308	NNNS	460	-0.866046
CMOS	1010	0.659069	NNNX	34	-1.175817
CMOX	68	0.308765	NNOO	340620	0.195102
CMSS	2047	2.848813	NNOS	5637	-0.305233
CMSX	13	1.172117	NNOX	302	-0.754766
CMXX	6	1.958862	NNSS	311	0.319427
CNNN	102035	-0.623046	NNSX	6	-0.874705
CNNO	1995038	0.140679	NNXX	5	0.168652
CNNS	15892	-0.376176	NOOO	171147	0.021937
CNNX	578	-0.993919	NOOS	10697	-0.077374
CNOO	2734639	0.227273	NOOX	3102	0.206513
CNOS	95438	0.050981	NOSS	922	0.440012
CNOX	2168	-0.771173	NOSX	12	-0.925060
CNSS	4264	0.584024	NOXX	61	0.903627
CNSX	37	-0.957113	NSSS	33	1.052833
CNXX	61	0.382553	NSSX	0	--
COOO	524994	-0.062707	NSXX	0	--
COOS	34429	-0.141141	NXXX	3	2.475964
COOX	23801	0.520038	OOOO	34212	-0.125549
COSS	4380	0.545326	OOOS	4240	-0.052504
COSX	58	-0.812243	OOOX	9553	1.121777
COXX	65	0.359781	OOSS	300	0.203077
CSSS	285	1.417735	OOSX	36	-0.197264
CSSX	5	0.006247	OOXX	128	1.476181
CSXX	4	0.730845	OSSS	38	1.063748
CXXX	9	2.381656	OSSX	3	0.305472
MMMM	83	7.794725	OSXX	0	--
MMMN	37	4.258301	OXXX	2	2.249518
MMMO	29	4.102142	SSSS	6	2.446092
MMMS	379	6.800300	SSSX	0	--
MMMX	0	--	SSXX	0	--
MMNN	83	1.849597	SXXX	0	--
MMNO	102	1.587734	XXXX	0	--

the structure files prior to tessellation, and all edges longer than 8Å were removed from the structure tessellations prior to analysis. Next, the four-body potential was used to score each tetrahedron in a structure tessellation based on the

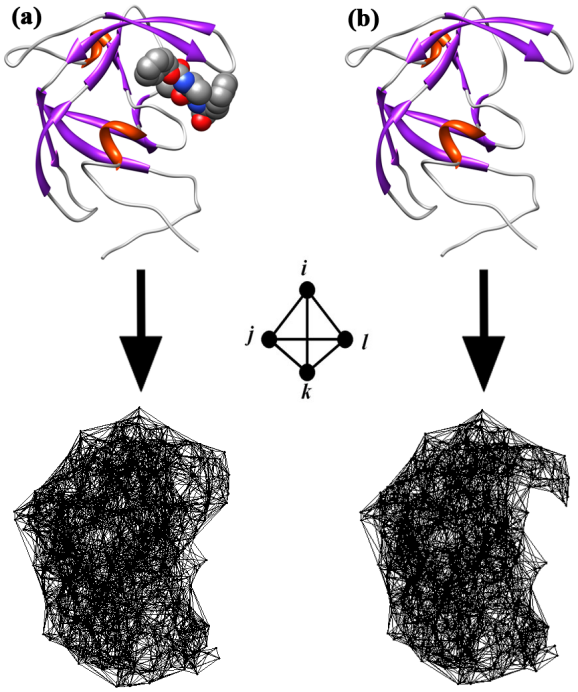

Fig. 1. Atomic Delaunay tessellation of the feline immunodeficiency virus protease enzyme (a) complexed with the C2 symmetric inhibitor 3TL and (b) with no bound inhibitor (PDB ID: 6FIV).

identity of the atomic quadruplet at its four vertices. A normalized topological score (*TS*) was calculated for each structure, defined as the sum of the scores for all the quadruplet interactions identified by the tetrahedra, divided by the total number of tetrahedra in the tessellation. Lastly, for each enzyme-inhibitor complex we computed the difference $\Delta TS = TS_{complex} - TS_{enzyme}$ in order to investigate its relationship to the experimental ΔG obtained from the inhibition constant (k_i). Calculated values of normalized topological scores, both for complexes and for enzymes without inhibitors, are tabulated in the text file of enzyme-inhibitor complexes (see MATERIALS AND METHODS).

C. Predictive Model of Binding Energy

We obtained a correlation coefficient of $r = 0.79$ between calculated ΔTS values and experimental ΔG measurements (ΔG_{exp}) for the two hundred enzyme-inhibitor complexes randomly selected for training a model. However, since the ΔTS data spanned both positive and negative real numbers that scaled differently from ΔG_{exp} values, they could not be used directly to represent predicted ΔG values (ΔG_{calc}). Both issues were addressed with the empirical linear function

$$\Delta G_{calc} = (1 / 0.0003) \times \Delta TS - 6.24,$$

which generated negative ΔG_{calc} values that scaled similarly to ΔG_{exp}. Given the linear transformation, ΔG_{calc} values and ΔG_{exp} measurements also displayed a correlation coefficient of $r = 0.79$, with a standard error of $SE = 2.34$ kcal/mol and a fitted regression line of $y = 0.98x - 0.41$ (Fig. 2). All the experimental and calculated ΔG values for the enzyme-inhibitor complexes are tabulated in the available text file

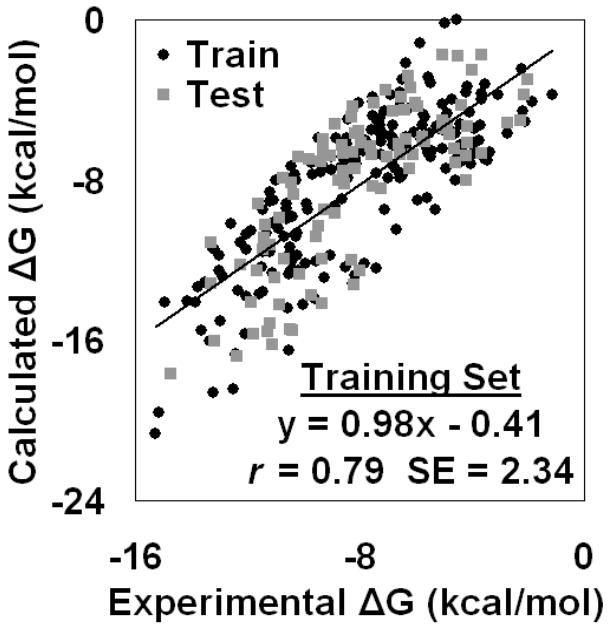

Fig. 2. Scatterplot of calculated versus experimental binding energies.

(see MATERIALS AND METHODS). Our results are comparable to those of other methods; for example, $r = 0.74$ and $SE = 1.34$ kcal/mol with the PHOENIX training set [10].

Structural tessellations were subsequently obtained for each of the remaining one hundred enzyme-inhibitor complexes held-out for testing as well as for each of the enzymes without their respective inhibitors. The four-body statistical potential was then used to compute a normalized topological score for each tessellation and a ΔTS value for each complex, from which ΔG_{calc} was computed with the linear transformation derived from the training data. Since the known experimental k_i values for these complexes were already used to determine their ΔG_{exp} measurements, the test data were also plotted in Fig. 2 superimposed over the training set. The correlation coefficient between ΔG_{calc} and ΔG_{exp} for the test data was slightly lower at $r = 0.77$, with a standard error for the predictions of $SE = 2.50$ kcal/mol and a fitted regression line of $y = 1.07x + 0.46$.

ACKNOWLEDGMENT

We are grateful to researchers affiliated with the Binding MOAD for creating a centralized database with access to all the structural and thermodynamic data used in this study.

REFERENCES

[1] A. Roy, P. R. McDonald, S. Sittampalam, and R. Chaguturu, "Open access high throughput drug discovery in the public domain: a Mount Everest in the making," Curr Pharm Biotechnol, vol. 11, 2010, pp. 764-778.

[2] S. Subramaniam, M. Mehrotra, and D. Gupta, "Virtual high throughput screening (vHTS)--a perspective," Bioinformation, vol. 3, 2008, pp. 14-17.

[3] R. Wang, L. Lai, and S. Wang, "Further development and validation of empirical scoring functions for structure-based binding affinity prediction," J Comput Aided Mol Des, vol. 16, 2002, pp. 11-26.

[4] A. Krammer, P. D. Kirchhoff, X. Jiang, C. M. Venkatachalam, and M. Waldman, "LigScore: a novel scoring function for predicting binding affinities," J Mol Graph Model, vol. 23, 2005, pp. 395-407.

[5] H. Gohlke, M. Hendlich, and G. Klebe, "Knowledge-based scoring function to predict protein-ligand interactions," J Mol Biol, vol. 295, 2000, pp. 337-356.

[6] C. A. Sotriffer, P. Sanschagrin, H. Matter, and G. Klebe, "SFCscore: scoring functions for affinity prediction of protein-ligand complexes," Proteins, vol. 73, 2008, pp. 395-419.

[7] R. Huey, G. M. Morris, A. J. Olson, and D. S. Goodsell, "A semiempirical free energy force field with charge-based desolvation," J Comput Chem, vol. 28, 2007, pp. 1145-1152.

[8] S. Y. Huang and X. Zou, "An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials," J Comput Chem, vol. 27, 2006, pp. 1866-1875.

[9] S. Y. Huang and X. Zou, "An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function," J Comput Chem, vol. 27, 2006, pp. 1876-1882.

[10] Y. T. Tang and G. R. Marshall, "PHOENIX: a scoring function for affinity prediction derived using high-resolution crystal structures and calorimetry measurements," J Chem Inf Model, vol. 51, 2011, pp. 214-228.

[11] C. M. Summa, M. Levitt, and W. F. Degrado, "An atomic environment potential for use in protein structure prediction," J Mol Biol, vol. 352, 2005, pp. 986-1001.

[12] F. Fogolari, L. Pieri, A. Dovier, L. Bortolussi, G. Giugliarelli, et al., "Scoring predictive models using a reduced representation of proteins: model and energy definition," BMC Struct Biol, vol. 7, 2007, pp. 15.

[13] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley, "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data," Nucleic Acids Res, vol. 35, 2007, pp. D301-303.

[14] L. Hu, M. L. Benson, R. D. Smith, M. G. Lerner, and H. A. Carlson, "Binding MOAD (Mother Of All Databases)," Proteins, vol. 60, 2005, pp. 333-340.

[15] M. L. Benson, R. D. Smith, N. A. Khazanov, B. Dimcheff, J. Beaver, et al., "Binding MOAD, a high-quality protein-ligand database," Nucleic Acids Res, vol. 36, 2008, pp. D674-678.

[16] C. B. Barber, D. P. Dobkin, and H. T. Huhdanpaa, "The quickhull algorithm for convex hulls," ACM Trans Math Software, vol. 22, 1996, pp. 469-483.

[17] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, et al., "UCSF Chimera--a visualization system for exploratory research and analysis," J Comput Chem, vol. 25, 2004, pp. 1605-1612.

[18] J. B. O. Mitchell, R. A. Laskowski, A. Alex, and J. M. Thornton, "BLEEP-Potential of mean force describing protein-ligand interactions: I. Generating potential," J Comput Chem, vol. 20, 1999, pp. 1165-1176.

[19] M. Masso and I. I. Vaisman, "Accurate prediction of enzyme mutant activity based on a multibody statistical potential," Bioinformatics, vol. 23, 2007, pp. 3155-3161.

[20] M. Masso and I. I. Vaisman, "AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements," Protein Eng Des Sel, vol. 23, 2010, pp. 683-687.

[21] M. J. Sippl, "Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures," Journal of Computer-Aided Molecular Design, vol. 7, 1993, pp. 473-501.