Majid Masso, School of Systems Biology, George Mason University, Manassas, VA, USA

Atomic Four-Body Statistical Potential for Macromolecular Structure Analysis

Over recent years, exponential growth of the Protein Data Bank (PDB) has facilitated selection of larger, non-redundant subsets of experimentally solved macromolecular structures at higher resolutions, which in turn have provided the data used in developing more effective knowledge-based statistical potentials for improved structure prediction. In contrast to physics-based energy functions, statistical potentials generally perform better and are more computationally efficient at identifying the native structure as a global minimum. Distance-dependent statistical potentials often focus on pairwise atomic contacts within macromolecular structures; however, such energy functions fail to consider important higher-order contributions based on multibody interactions. In the present work, we develop an all-atom four-body statistical potential and illustrate its applicability with a constructive example.

The potential was derived by analyzing coordinate data for 1417 high-resolution ($\leq$ 2.2Å) crystallographic structures selected from the PDB that contain protein chains sharing low (< 30%) sequence identity. Both single-chain and multimeric protein structures are represented in the dataset, the majority of which are also complexed to small molecular or peptide ligands (http://proteins.gmu.edu/automute/tessellatable1417.txt). Coordinates of hydrogen atoms and water molecules were excluded from these files, and a six-letter alphabet (C, N, O, S, M = all metals, X = all other non-metals) was used to designate the remaining heavy atom types in the structures. Atomic coordinates of each structure were supplied to the Qhull implementation of the Delaunay tessellation computational geometry algorithm, which treats the points as vertices and generates an aggregate of non-overlapping irregular tetrahedra. Edges longer than 12Å were removed from every tessellation prior to analysis, so that each remaining tetrahedron objectively identified at its vertices an interacting atomic quadruplet, of which there are 126 distinct possibilities.

An observed relative frequency of occurrence f_{ijkl} was calculated for each atomic quadruplet type (i,j,k,l) based upon the proportion of tetrahedra, from among those comprising all the tessellated structures, for which the quadruplet appears at the four vertices. A rate expected by chance was obtained with the multinomial reference distribution

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

In the above formula, a_n represents the proportion of all atoms in the tessellated structures that are of type n, and t_n is the number of occurrences of atom type n in the quadruplet. Through application of the inverted Boltzmann principle, the score $s_{ijkl} = \log(f_{ijkl} / p_{ijkl})$ quantified an energy of interaction for the atomic quadruplet. The set of 126 atomic quadruplet types with their respective energy scores defines the four-body potential, which can subsequently be used to compute a topological score for any structure as follows: first tessellate (subject to a 12Å edge-length cutoff), and then sum the scores of the atomic quadruplets identified on the four vertices of all constituent tetrahedra.

As a practical application, we analyzed structures of HIV-1 protease complexed to 140 distinct inhibitors, each with an experimentally known dissociation constant. The four-body potential was used to predict binding energy as the difference between the topological score of the complex and that of the target without the inhibitor (Fig. 1), and a correlation coefficient of $r^2 = 0.64$ was observed between experimental and predicted binding energies (Fig. 2).

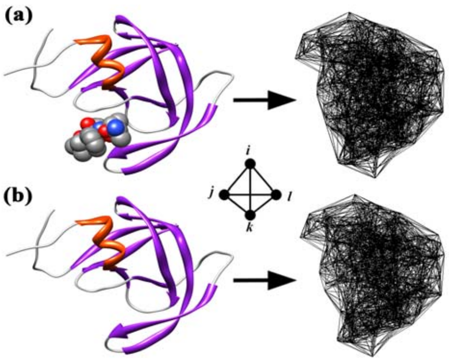

Fig. 1. Atomic Delaunay tessellations of HIV-1 protease.

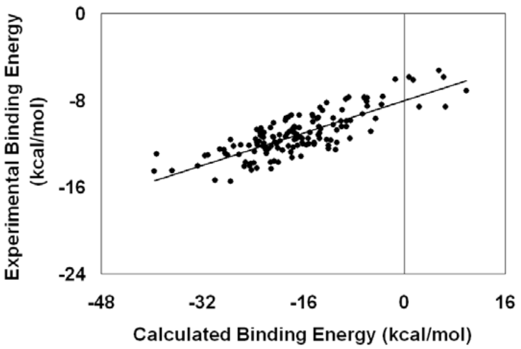

Fig. 2. Scatterplots of HIV-1 protease-inhibitor complexes.