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

Generation of Atomic Four-Body Statistical Potentials Derived from the Delaunay Tessellation of Protein Structures

Majid Masso, *Member, IEEE*

***Abstract*—Delaunay tessellation of the atomic coordinates for a crystallographic protein structure yields an aggregate of non-overlapping and space-filling irregular tetrahedral simplices. The vertices of each simplex objectively identify a quadruplet of nearest neighbor atoms in the protein. Here we apply Delaunay tessellation to 1417 high-resolution structures of single chains that share low sequence identity, for the purpose of determining the relative frequencies of occurrence for all possible nearest neighbor atomic quadruplet types. Alternative distributions are explored by varying two fundamental parameters: atomic alphabet selection and cutoff length for admissible simplex edges. The distributions are then converted to four-body potential functions by implementing the inverted Boltzmann principle, which requires calculating the distribution of the reference state. Two alternative definitions for the reference state are presented, which introduces a third parameter, and we derive and compare an array of such potential functions. These knowledge-based statistical potentials based on higher-order interactions complement and generalize the more commonly encountered atom-pair potentials, for which a number of approaches are described in the literature.**

I. Introduction

KNOWLEDGE-BASED statistical potentials have in recent years become tremendously popular as computationally efficient alternatives to physics-based energy functions for conducting large-scale analyses of protein structures [1, 2]. These statistical energy functions rely on information extracted from databases of known protein structures, whereby the observed relative frequency with which a feature (e.g., the interaction of a particular pair of residues or atoms) occurs in the known structures, f_{obs}, is related to its probability expected in the reference state, f_{exp}, in order to calculate an effective energy

$$E = \ln(f_{obs} / f_{\exp}) \qquad (1)$$

for that feature. An important underlying assumption is that observed relative frequencies satisfy the Bolzmann distribution with respect to feature energies, hence justifying the use of its inverted form for generating the potential [3].

Though pairwise statistical potentials were applied successfully at the residue [4-8] and atomic [9-11] levels, improvements were thought possible by including higher order cooperative interactions in the potentials [12-14]. Consequently, multibody potentials at the residue [15-17] and atomic [18, 19] levels were also investigated. In particular, four-body residue potentials were developed via the application of Delaunay tessellation, a tiling algorithm from computational geometry, to coarse-grained models of protein structures represented by their residue alpha-carbon coordinates [14, 20]. For each protein structure, Delaunay tessellation of the point-set yields a three-dimensional aggregate of space-filling non-overlapping irregular tetrahedra, or Delaunay simplices, whereby the vertices of each simplex objectively define four nearest neighbor residues via their alpha-carbons for the purpose of developing the four-body residue potential. Here we apply Delaunay tessellation at the atomic level, while considering variations to size of atomic alphabet, maximum length of simplex edges, and derivation of reference distribution, to generate an array of atomic four-body statistical potentials.

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).

II. Materials and Methods

A. Protein Dataset and Atom Types

Delauany tessellation was performed on each of 1417 single protein chains sharing low (< 30%) sequence similarity, whose structural coordinates were obtained from high-resolution (≤ 2.2Å) crystallographic files deposited in the Protein Data Bank (PDB) [21]. The dataset is available at http://proteins.gmu.edu/automute/tessellatable1417.txt.

Defining atom types necessitates a compromise between two opposing considerations: the need to fully describe the diversity of quadruplet atomic interactions (i.e., a larger alphabet) while ensuring that sufficient frequency data are collected for each type of quadruplet (i.e., a smaller alphabet) [22]. Coordinates of hydrogen atoms were not included in the tessellations, and three approaches were investigated for labeling each of the remaining heavy atom types. First, a simple four-letter alphabet (C, N, O, S) accounts for all atom types. Next, an eight-letter alphabet (Backbone: N, B = alpha-carbon, C, O; Side-chain: X = nitrogen, Z = carbon, U = oxygen, S) distinguishes residue backbone and side-chain atoms as well as residue backbone alpha- and carbonyl- carbon atoms. Lastly, a twenty-letter atomic alphabet obtained from Summa *et al.* [18] groups atoms based on common traits, including bonding pattern, partial charge, and hydrophobicity.

B. Four-Body Statistical Potential

Each Delaunay tessellation for a protein structure was obtained by supplying all constituent non-hydrogen atomic

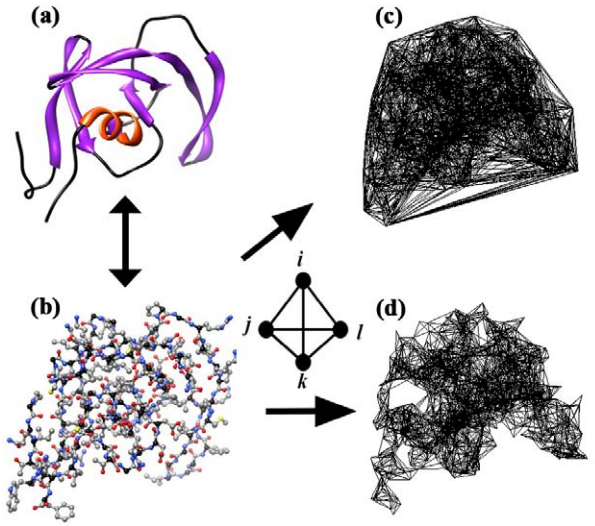

Fig. 1. HIV-1 protease (a) ribbon and (b) ball-and-stick diagrams (PDB ID: 3phv). The atomic coordinates of (b) are used for generating (c) the Delaunay tessellation of the protein chain, a convex hull of tetrahedral simplices defining quadruplets of nearest neighbor atoms. The modified Delaunay tessellation in (d) is obtained by removing all edges longer than 4.8 Å. Imposing such a strict edge-length cutoff significantly reduces the number of Delauany simplices.

coordinates to the Qhull program [23], which treats points as vertices and generates a convex hull of non-overlapping irregular tetrahedral simplices (Fig. 1). The vertices of each simplex objectively identify four nearest neighbor atoms; however, to ensure each simplex represents a quadruplet of interacting atoms that fall within a fixed distance of one another, all edges in the tessellations longer than a prescribed cutoff value may be subsequently removed prior to analysis (Fig. 1, Table I). For generating potentials, we considered quadruplet frequencies based on all observed simplices from the structure tessellations (i.e., no cutoff), as well as those based only on simplices whose edges all satisfied a specific length cutoff. Molecular structures of Fig. 1 were produced with Chimera [24], and tessellations were created in Matlab.

The number N of distinct subsets of size $r = 4$ letters that can be formed from an atomic alphabet of size K, excluding quadruplet permutations but allowing for the repeated occurrence of letters in a quadruplet, is given by the combinatorial formula

$$N = \binom{K+r-1}{r} = \binom{K+3}{4}. \qquad (2)$$

Using (2), we determined that the number of distinct atomic quadruplets that can possibly be enumerated based on atomic alphabets of size $K = 4$, 8, or 20 are $N = 35$, 330, and 8855, respectively. For each alphabet size, we calculated the observed relative frequency of occurrence f_{ijkl} for each of the N possible quadruplets (i,j,k,l) based upon the proportion of simplices, from among those generated by all the structure

TABLE I
SUMMARY DATA FOR THE 1417 PROTEIN CHAINS

Four-Letter Atom Types	Count	Proportion
(carbon) C	1572222	0.634149
(nitrogen) N	425874	0.171774
(oxygen) O	469869	0.189520
(sulfur) S	11299	0.004557
Total atom count:	2479264	
Total tetrahedron counts		
No edge-length cutoff:	16152638	
12 Å edge-length cutoff:	15497203	
4.8 Å edge-length cutoff:	9569503	

tessellations, for which the quadruplet appears at the four vertices. In cases where we applied an edge-length cutoff to tessellations prior to analysis, observed relative frequencies were based on a reduced total number of simplices.

Next, we calculated the rate expected by chance for each of the N quadruplets (i,j,k,l) from the multinomial reference distribution, given by

$$p_{ijkl} = \frac{4!}{\prod_{n=1}^{K}(t_n!)} \prod_{n=1}^{K} a_n^{t_n}, \text{where } \sum_{n=1}^{K} a_n = 1 \text{ and } \sum_{n=1}^{K} t_n = 4. \qquad (3)$$

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. A potential drawback to this reference state is the implicit assumption that the 1417 tessellated protein structures constitute a single molecular system from which any four atoms, either from the same structure or belonging to multiple protein chains, are able to serve as vertices of a simplex. Hence, we also defined a reference state separating atoms according to their structures, by calculating the expected rate for each quadruplet as a weighted average of multinomial probabilities obtained from each protein individually. Weights are based on protein size, and the formula for this alternative reference state is given by

$$p_{ijkl} = \sum_{m=1}^{1417} \frac{R_m}{R} p_{ijkl}^{m}, \qquad (4)$$

where R_m is the number of atoms in protein m, and R is the total number of all atoms in the 1417 protein chains.

Given any fixed selection of parameters (i.e., atomic alphabet, edge-length cutoff, and reference state), we applied the inverted Bolzmann principle (1) in order to calculate a score $s_{ijkl} = \log (f_{ijkl} / p_{ijkl})$ that quantifies the interaction energy for each atomic quadruplet (i,j,k,l), thus defining a four-body statistical potential function.

III. EXAMPLES

Comprehensive data regarding the development of a four-body statistical potential based on a four-letter alphabet, no edge-length cutoffs on the tessellations, and use of the

TABLE II
FOUR-BODY STATISTICAL POTENTIAL BASED ON A FOUR-LETTER ALPHABET, NO CUTOFF, AND UNWEIGHTED REFERENCE DISTRIBUTION

Quad	Count	f_{ijkl}	p_{ijkl}	s_{ijkl}
CCCC	1711740	0.105973	0.161720	-0.183570
CCCN	1823746	0.112907	0.175223	-0.190871
CCCO	2807489	0.173810	0.193325	-0.046213
CCCS	119435	0.007394	0.004649	0.201538
CCNN	832442	0.051536	0.071195	-0.140340
CCNO	3838549	0.237642	0.157100	0.179748
CCNS	53655	0.003322	0.003778	-0.055872
CCOO	1643096	0.101723	0.086665	0.069578
CCOS	86638	0.005364	0.004168	0.109530
CCSS	6408	0.000397	5.01E-05	0.898511
CNNN	64504	0.003993	0.012857	-0.507783
CNNO	961282	0.059512	0.042554	0.145664
CNNS	7628	0.000472	0.001023	-0.335839
CNOO	1380693	0.085478	0.046950	0.260215
CNOS	44097	0.002730	0.002258	0.082434
CNSS	2153	0.000133	2.71E-05	0.691035
COOO	336824	0.020853	0.017267	0.081946
COOS	17883	0.001107	0.001246	-0.051201
COSS	2068	0.000128	3.00E-05	0.630846
CSSS	214	1.32E-05	2.40E-07	1.741768
NNNN	4632	0.000287	0.000871	-0.482308
NNNO	36223	0.002243	0.003842	-0.233848
NNNS	407	2.52E-05	9.24E-05	-0.564301
NNOO	190771	0.011811	0.006359	0.268893
NNOS	3088	0.000191	0.000306	-0.204035
NNSS	236	1.46E-05	3.68E-06	0.599166
NOOO	129494	0.008017	0.004677	0.234026
NOOS	5426	0.000336	0.000337	-0.001929
NOSS	436	2.70E-05	8.11E-06	0.522015
NSSS	138	8.54E-06	6.50E-08	2.118466
OOOO	39551	0.002449	0.001290	0.278298
OOOS	1462	9.05E-05	0.000124	-0.137035
OOSS	158	9.78E-06	4.48E-06	0.339520
OSSS	22	1.36E-06	7.18E-08	1.278314
SSSS	50	3.10E-06	4.31E-10	3.855858

TABLE III
EFFECT OF REFERENCE STATE ON FOUR-BODY STATISTICAL POTENTIALS BASED ON A FOUR-LETTER ALPHABET AND 4.8 Å EDGE-LENGTH CUTOFF

Quad	Count	s_{ijkl} (unweighted)	s_{ijkl} (weighted)
CCCC	922742	-0.224573	-0.225145
CCCN	1229054	-0.134910	-0.134801
CCCO	1533444	-0.081509	-0.081269
CCCS	53175	0.077468	0.080517
CCNN	612799	-0.046021	-0.046302
CCNO	2695736	0.253612	0.254318
CCNS	28817	-0.098480	-0.096682
CCOO	796121	-0.017752	-0.017954
CCOS	36077	-0.043594	-0.040613
CCSS	3685	0.885580	0.763552
CNNN	18568	-0.821250	-0.822983
CNNO	647785	0.201598	0.201779
CNNS	3499	-0.446952	-0.447508
CNOO	757532	0.226873	0.227017
CNOS	22706	0.021519	0.022970
CNSS	1312	0.703279	0.575108
COOO	65646	-0.400894	-0.402722
COOS	4711	-0.403174	-0.401559
COSS	1001	0.543084	0.416103
CSSS	125	1.735618	1.308914
NNNN	133	-1.796871	-1.800957
NNNO	5948	-0.791107	-0.792435
NNNS	58	-1.183114	-1.187162
NNOO	97822	0.206171	0.205677
NNOS	1061	-0.440643	-0.441841
NNSS	79	0.351235	0.215194
NOOO	26079	-0.234579	-0.236150
NOOS	1832	-0.246129	-0.246333
NOSS	220	0.452305	0.318451
NSSS	48	1.887182	1.441123
OOOO	1542	-0.903421	-0.908012
OOOS	78	-1.182534	-1.183627
OOSS	32	-0.126633	-0.260776
OSSS	5	0.862215	0.415567
SSSS	31	3.875603	2.870350

unweighted reference distribution, are reported in Table II. In particular, for each of the 35 possible atomic quadruplet types, we provide the respective number of Delaunay simplices from the 1417 protein structure tessellations for which the quadruplet appears at the four vertices, the observed relative frequency, the rate expected by chance as calculated from the unweighted multinomial reference distribution, and finally the calculated interaction energy score. Note that quadruplet propensity for occurrence relative to chance is greatest for SSSS, while of all quadruplets that appear less often than by chance alone, NNNS is the most rarely observed.

Again based on a four-letter alphabet, Table III presents for comparison four-body statistical potentials based on a 4.8 Å edge-length cutoff imposed on all the structure tessellations, using both reference distribution formulations. Note the fewer number of observed simplices/quadruplets in all cases for Table III due to the cutoff relative to those in Table II, some more significantly reduced than others owing to biophysical considerations. For each quadruplet in Table III, the corresponding pair of scores based on both potentials have the same sign and are relatively similar in magnitude, with the exception of quadruplets that are composed of at least two sulfur (S) atoms, for which the differences in magnitude are more substantial. Comparing the unweighted reference state potential of Table III with that of Table II allows us to evaluate the impact of a strict edge-length cutoff. Here we observe not only more substantial differences in magnitudes between pairs of quadruplet scores, but also several cases involving quadruplets with multiple oxygen (O) atoms in which there are sign differences. Additionally, while SSSS has the most positive score in both potentials, NNNN is scored most negatively by the unweighted reference potential in Table III and nearly so in Table II as well, surpassed only by NNNS and CNNN.

Since four-body statistical potentials based on eight- and twenty-letter alphabets are too large to be tabulated, we generated one example of each in order to present a graphical depiction of their ordered distribution of scores (Fig. 2). Three quadruplets, BBBS, CCCC, and CCCX, were not observed at all as simplex vertices in the case of the eight-letter, 8 Å cutoff potential, and while the same is true even with no cutoff, the number of unobserved quadruplets increases to eight with a 4.8 Å cutoff. Similarly, 995 quadruplets were unobserved based on a twenty-letter, 12 Å cutoff potential. In both cases, the quadruplets with the most extreme scores were identical, with SSSS and BCCC having the largest positive and negative scores, respectively.

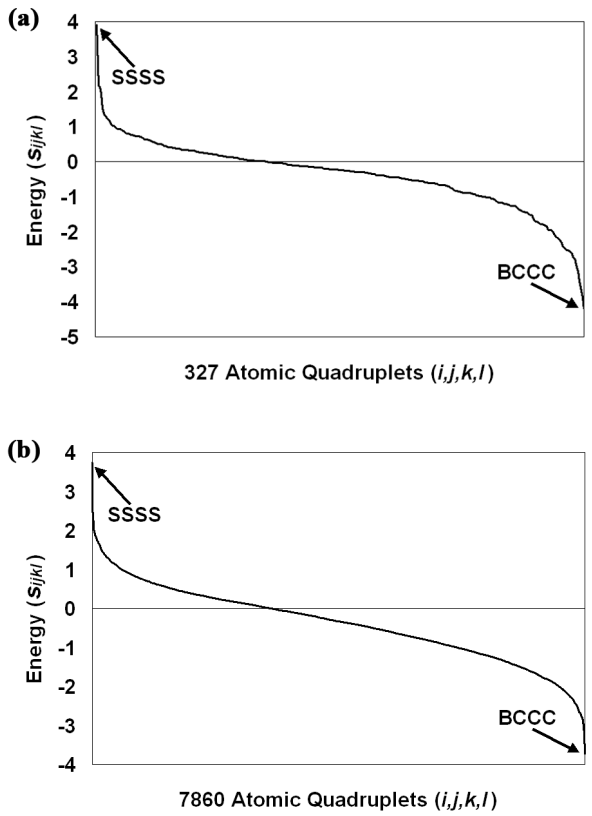

Fig. 2. Ordered atomic quadruplet interaction energies from four-body potentials based on (a) an eight-letter alphabet with an 8 Å edge-length cutoff, and (b) a twenty-letter alphabet with a 12 Å edge-length cutoff. The unweighted reference state is used in both cases, and the letters B = alpha-carbon, C = carbonyl-carbon, and S = sulfur (from either cysteine or methionine) represent the same atom types in both alphabets.

In conclusion, we have developed an approach based on atomic Delaunay tessellation of protein structures for generating an array of four-body statistical potentials. Evaluation of these potentials for their ability to discriminate native protein structures from non-native folds, and for their practical application to the analysis of protein structure and function, are the current focus of our research efforts.

APPENDIX

Tabulated four-body statistical potentials corresponding to the distribution plots depicted in Fig. 2 are available at http://proteins.gmu.edu/automute/origRef-8let8A-pot.txt and http://proteins.gmu.edu/automute/origRef-20let12A-pot.txt.

REFERENCES

[1] L. A. Clark and H. W. van Vlijmen, "A knowledge-based forcefield for protein-protein interface design," Proteins, vol. 70, 2008, pp. 1540-1550.

[2] A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Oldziej, et al., "Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins," J Phys Chem B, vol. 111, 2007, pp. 260-285.

[3] 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.

[4] M. R. Betancourt and S. J. Omovie, "Pairwise energies for polypeptide coarse-grained models derived from atomic force fields," J Chem Phys, vol. 130, 2009, pp. 195103.

[5] G. Zhao and H. Lu, "Development of a Grid-based statistical potential for protein structure prediction," Conf Proc IEEE Eng Med Biol Soc, vol. 6, 2005, pp. 6064-6067.

[6] S. Miyazawa and R. L. Jernigan, "An empirical energy potential with a reference state for protein fold and sequence recognition," Proteins, vol. 36, 1999, pp. 357-369.

[7] F. Melo and M. A. Marti-Renom, "Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets," Proteins, vol. 63, 2006, pp. 986-995.

[8] C. Zhang and S. H. Kim, "Environment-dependent residue contact energies for proteins," Proc Natl Acad Sci U S A, vol. 97, 2000, pp. 2550-2555.

[9] F. Melo and E. Feytmans, "Novel knowledge-based mean force potential at atomic level," J Mol Biol, vol. 267, 1997, pp. 207-222.

[10] R. Samudrala and J. Moult, "An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction," J Mol Biol, vol. 275, 1998, pp. 895-916.

[11] H. Lu and J. Skolnick, "A distance-dependent atomic knowledge-based potential for improved protein structure selection," Proteins, vol. 44, 2001, pp. 223-232.

[12] M. Lappe, G. Bagler, I. Filippis, H. Stehr, J. M. Duarte, et al., "Designing evolvable libraries using multi-body potentials," Curr Opin Biotechnol, vol. 20, 2009, pp. 437-446.

[13] E. Ferrada and F. Melo, "Effective knowledge-based potentials," Protein Sci, vol. 18, 2009, pp. 1469-1485.

[14] R. K. Singh, A. Tropsha, and I. I. Vaisman, "Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues," J Comput Biol, vol. 3, 1996, pp. 213-221.

[15] S. Mayewski, "A multibody, whole-residue potential for protein structures, with testing by Monte Carlo simulated annealing," Proteins, vol. 59, 2005, pp. 152-169.

[16] J. E. Fitzgerald, A. K. Jha, A. Colubri, T. R. Sosnick, and K. F. Freed, "Reduced C(beta) statistical potentials can outperform all-atom potentials in decoy identification," Protein Sci, vol. 16, 2007, pp. 2123-2139.

[17] X. Li and J. Liang, "Geometric cooperativity and anticooperativity of three-body interactions in native proteins," Proteins, vol. 60, 2005, pp. 46-65.

[18] 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.

[19] M. R. Betancourt, "A reduced protein model with accurate native-structure identification ability," Proteins, vol. 53, 2003, pp. 889-907.

[20] A. Tropsha, C. W. Carter, Jr., S. Cammer, and Vaisman, II, "Simplicial neighborhood analysis of protein packing (SNAPP): a computational geometry approach to studying proteins," Methods Enzymol, vol. 374, 2003, pp. 509-544.

[21] 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.

[22] 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.

[23] 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.

[24] 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.