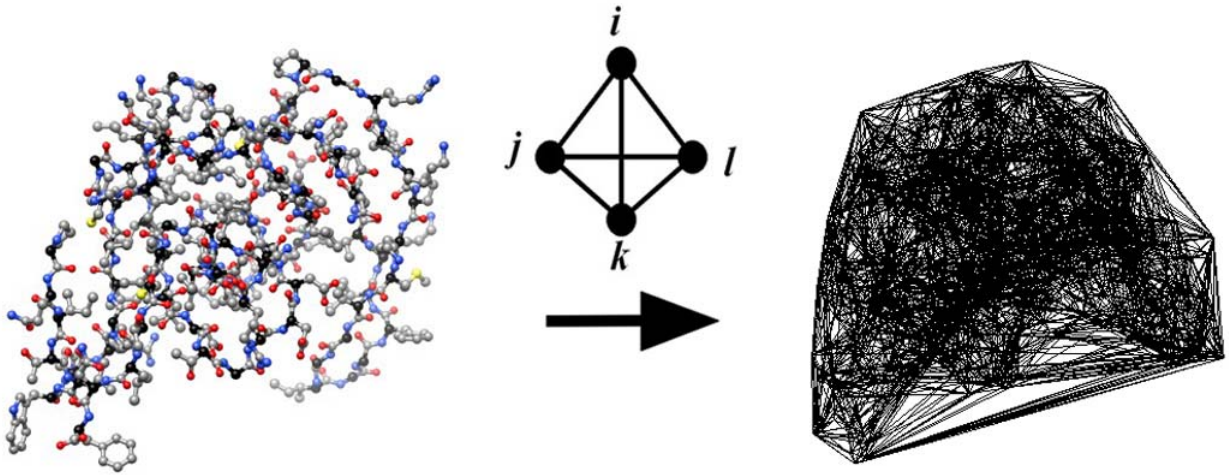
Generation of Atomic Four-Body Statistical Potentials Derived from the Delaunay Tessellation of Protein Structures



Majid Masso

School of Systems Biology, George Mason University

Manassas, Virginia 20110, USA

EMBC 2012, San Diego, California

Protein Data Bank (http://www.rcsb.org/pdb)

- PDB – repository of solved (x-ray, nmr, ...) structures
- Each structure file contains atomic 3D coordinate data

	Atom					X	Y	Z			
ATOM	1	N	PRO	E	1	-13.470	40.080	31.429	1.00	20.55	N
ATOM	2	CA	PRO	E	1	-13.130	40.167	30.000	1.00	18.37	C
ATOM	3	C	PRO	E	1	-13.948	39.156	29.184	1.00	19.60	C
ATOM	4	O	PRO	E	1	-14.613	38.254	29.681	1.00	14.33	O
ATOM	5	CB	PRO	E	1	-11.628	39.965	29.977	1.00	18.47	C
ATOM	6	CG	PRO	E	1	-11.271	39.349	31.318	1.00	16.39	C
ATOM	7	CD	PRO	E	1	-12.253	40.025	32.261	1.00	17.03	C
ATOM	8	N	GLN	E	2	-13.954	39.360	27.885	1.00	17.78	N
ATOM	9	CA	GLN	E	2	-14.612	38.615	26.843	1.00	18.54	C
ATOM	10	C	GLN	E	2	-13.519	37.852	26.032	1.00	21.37	C
ATOM	11	O	GLN	E	2	-12.525	38.462	25.599	1.00	18.10	O
ATOM	12	CB	GLN	E	2	-15.431	39.449	25.905	1.00	9.92	C
ATOM	13	CG	GLN	E	2	-16.976	39.087	25.986	1.00	21.78	C
ATOM	14	CD	GLN	E	2	-17.504	39.810	24.755	1.00	28.81	C
ATOM	15	OE1	GLN	E	2	-17.660	39.195	23.731	1.00	28.78	O
ATOM	16	NE2	GLN	E	2	-17.660	41.125	24.919	1.00	42.68	N
						:		:			
						:		:			

Knowledge-Based Potentials of Mean Force

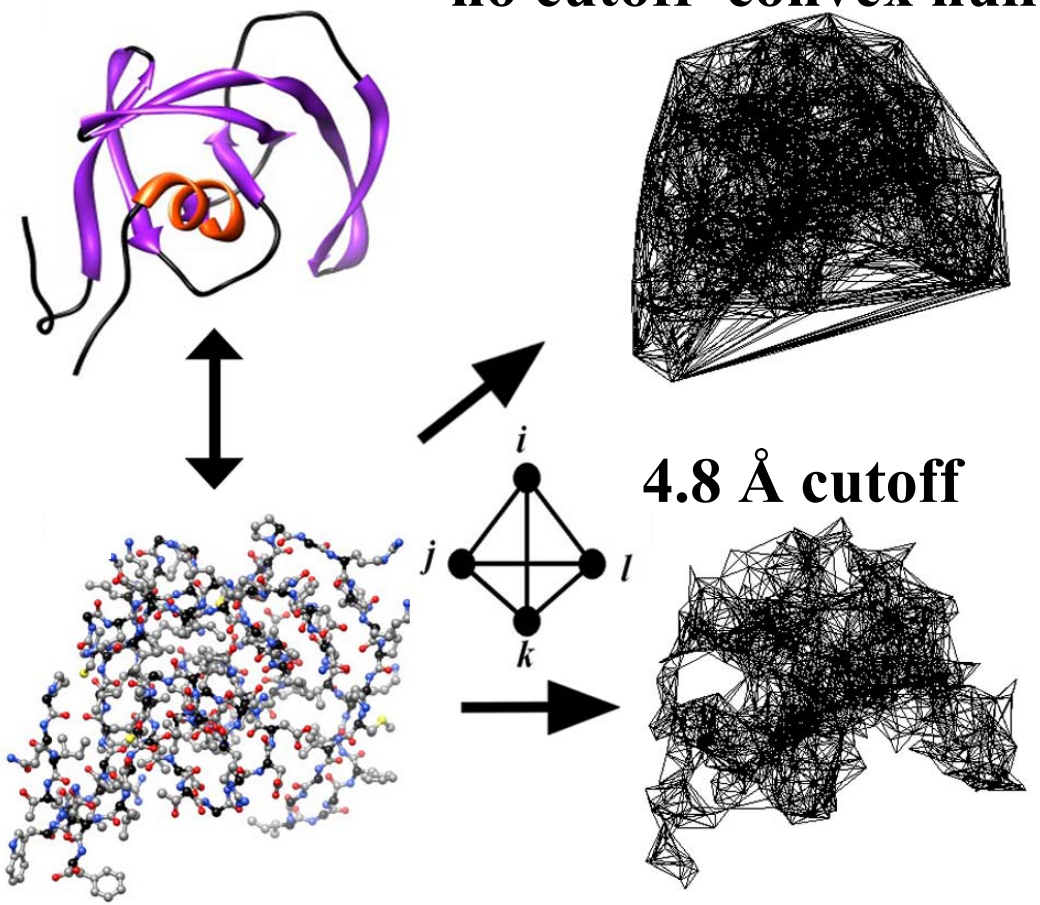
- Generated via statistical analysis of observed features in a diverse subset of protein structures from PDB (training set)
- Alternative to physics or molecular mechanics energy functions
- Assumption: observed features follow a Boltzmann distribution
- Examples:
  - Well-documented in the literature: distance-dependent pairwise interactions at the atomic or amino acid level
  - This study: inclusion of higher-order contributions by developing all-atom four-body statistical potentials
- Motivation (our prior work):
  - Four-body protein potentials at the amino acid level

All-Atom Four-Body Statistical Potential

- Diverse PDB dataset consisting of 1417 single protein chains
- Represent each structure as a collection of atomic points in 3D
- 1st adjustable parameter – atomic alphabet size for labeling points:
  - 4-letter (N, C, O, S), the focus of this talk
  - 8-letter (Backbone: N, Cα, C, O; Side-chain: N', C', O', S)
  - 20-letter (defined as in Summa *et al.*, *JMB*, 2005)

All-Atom Four-Body Statistical Potential

- Apply Delaunay tessellation to the atomic point coordinates of each PDB file – objectively identifies all nearest neighbor quadruplets of atoms in the structure
- 2nd adjustable parameter – tetrahedral simplex edge length cutoff, if any, to ensure all 4 atoms are interacting with one other



All-Atom Four-Body Statistical Potential

- $K = 4$ atomic alphabet size ➔ $N = 35$ possible atomic quadruplets that may appear on the four vertices of a tetrahedron
- Assumption: letters can be repeated in a quad, but the permutation of letters in any given quad does not constitute a new one
- Other cases: $K = 8$ ➔ $N = 330$; $K = 20$ ➔ $N = 8855$
- For each atomic quadruplet (i, j, k, l), calculate proportion f_{ijkl} of all the tetrahedra generated by the 1417 protein tessellations that have the quad appearing on the four vertices (i.e., the observed relative frequency of quad occurrence)

All-Atom Four-Body Statistical Potential

- Compute a rate p_{ijkl} expected by chance from a multinomial reference distribution:

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

- a_n = proportion of atoms from the proteins that are of type n
- t_n = number of occurrences of atom type n in the quad
- Apply inverted Bolzmann principle: score $s_{ijkl} = \log(f_{ijkl} / p_{ijkl})$ quantifies interaction propensity and is proportional to the energy of interaction (by a factor of '$-RT$')
- How to use it to calculate total potential energy of any protein? 1) Tessellate its atomic coordinates; 2) score each tetrahedron based on the atomic quad at its vertices; 3) add up the scores

Summary Data for the 1417 Protein Chains and their Delaunay Tessellations

Four-Letter Atom Types	Count	Proportion
(carbon) C	1,572,222	0.634149
(nitrogen) N	425,874	0.171774
(oxygen) O	469,869	0.189520
(sulfur) S	11,299	0.004557
Total atom count:	2,479,264	
Total tetrahedron counts		
No edge-length cutoff:	16,152,638	
12 Å edge-length cutoff:	15,497,203	
4.8 Å edge-length cutoff:	9,569,503	

3rd Adjustable Parameter – Reference Choice

- Potential drawback of previous reference distribution: implicitly assumes atoms from different protein chains are capable of forming the four vertices of a Delaunay simplex
- Alternative approach: weighted averaging by protein chain

$$p_{ijkl} = \sum_{m=1}^{1417} \frac{R_m}{R} p^m_{ijkl}$$

- R_m = number of atoms in protein chain m
- R = total number of atoms in all 1417 protein chains
- p^m_{ijkl} = multinomial dist applied only to atoms of protein chain m

All-Atom Four-Body Statistical Potential:

4-letter, no simplex edge cutoff, unweighted reference

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

Effect of Reference State on the Potential:

4-letter, 4.8 Å simplex edge 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

References and Acknowledgments

- PDB (structure DB): http://www.rcsb.org/pdb
- Qhull (Delaunay tessellation): http://www.qhull.org/
- UCSF Chimera (ribbon/ball-stick structure visualization): http://www.cgl.ucsf.edu/chimera/
- Matlab (tessellation visualization): http://www.mathworks.com/products/matlab/
- Perl programming language used exclusively for performing all data extraction, formatting, and analyses.