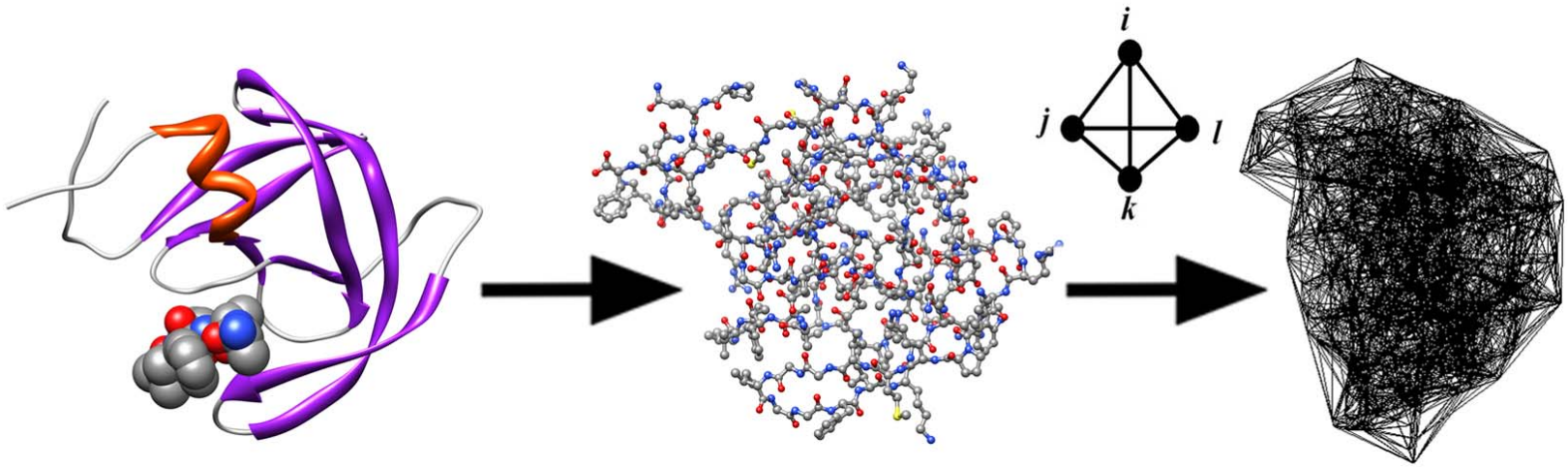
Atomic Four-Body Statistical Potential for Macromolecular Structure Analysis



Majid Masso

School of Systems Biology, George Mason University

Manassas, Virginia 20110, USA

SMB 2012 Annual Meeting, Knoxville, Tennessee

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
						:		:			
						:		:			

Macromolecular Modeling

- Native structure is conformation having lowest energy
- Physics-based energy calculations using quantum mechanics are computationally impractical
- Same for molecular mechanics-based potential energy functions (i.e., force fields): E(total) = E(bond) + E(angle) + E(dihedral) + E(electrostatic) + E(van der Waals)
- Alternative (our approach): knowledge-based potentials of mean force (i.e., generated from known protein structures)

Knowledge-Based Potentials of Mean Force

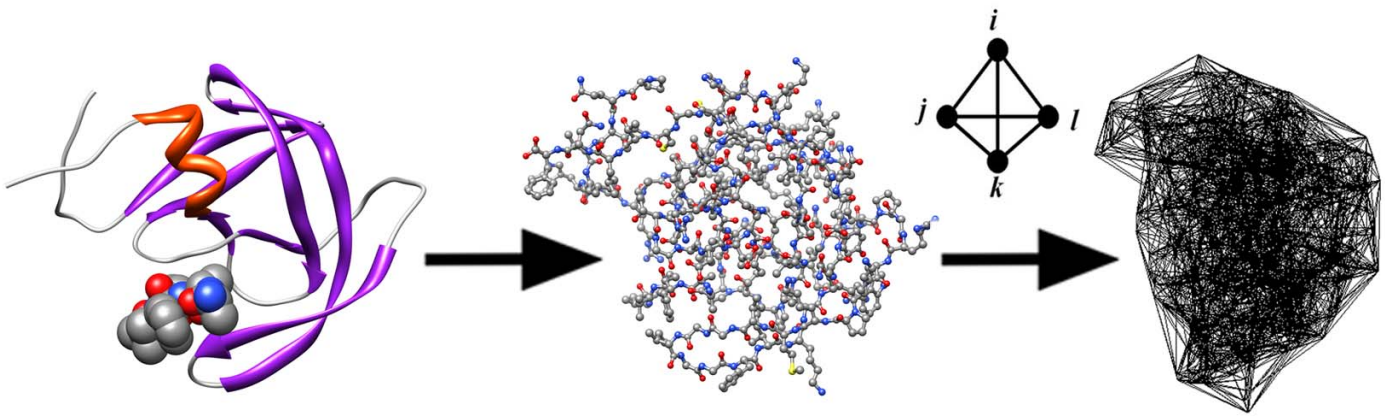
- Assumptions:
  - At equilibrium, native state has global free energy min
  - Microscopic states (i.e., features) follow Boltzmann dist
- Examples:
  - Well-documented in the literature: distance-dependent pairwise interactions at the atomic or amino acid level
  - This study: inclusion of higher-order contributions by developing an all-atom four-body statistical potential
- Motivation (our prior work):
  - Four-body protein potential at the amino acid level

Motivational Example: Pairwise Amino Acid Potential

- A 20-letter protein alphabet yields 210 residue pairs
- Obtain large, diverse PDB dataset of single protein chains
- For each residue pair (i, j), calculate the relative frequency f_{ij} with which they appear within a given distance (e.g., 12 angstroms) of each other in all the protein structures
- Calculate a rate p_{ij} expected by chance alone from a reference distribution (more later…)
- Apply inverted Bolzmann principle: $s_{ij} = \log(f_{ij} / p_{ij})$ quantifies interaction propensity and is proportional to the energy of interaction (by a factor of '$-RT$')

All-Atom Four-Body Statistical Potential

- Diverse PDB dataset includes 1417 single chain and multimeric protein structures, many complexed to ligands
- Six-letter atomic alphabet: C, N, O, S, M (metals), X (other)
- Apply Delaunay tessellation to the atomic point coordinates of each PDB file – objectively identifies all nearest-neighbor quadruplets of atoms in the structure (12 angstrom cutoff)



All-Atom Four-Body Statistical Potential

- A six-letter atomic alphabet yields 126 distinct quadruplets
- Calculate observed rate f_{ijkl} of quad (i, j, k, l) occurrence among all tetrahedra from the 1417 structure tessellations
- Compute rate p_{ijkl} expected by chance from a multinomial reference distribution:

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

- a_n = proportion of atoms from all structures that are of type n
- t_n = number of occurrences of atom type n in the quad

Summary Data for the 1417 Structure Files and their Delaunay Tessellations

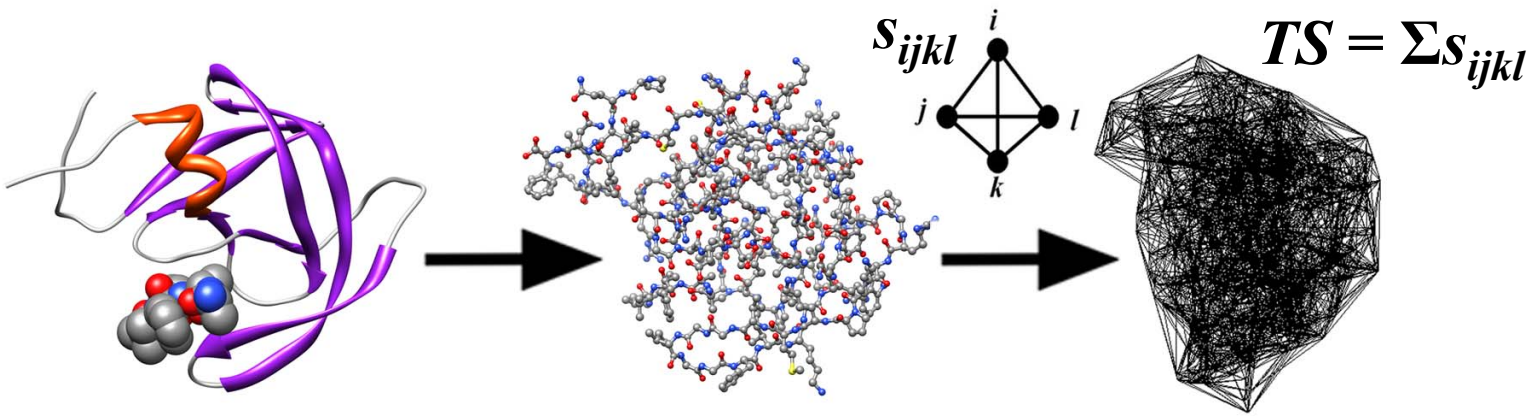
Atom Types	Count	Proportion
C	3,612,988	0.633193
N	969,253	0.169866
O	1,088,410	0.190749
S	28,502	0.004995
(all metals) M	2,529	0.000443
(all other non-metals) X	4,299	0.000754
Total atom count:	5,705,981	
Total tetrahedron count:	36,406,467	

All-Atom Four-Body Statistical Potential

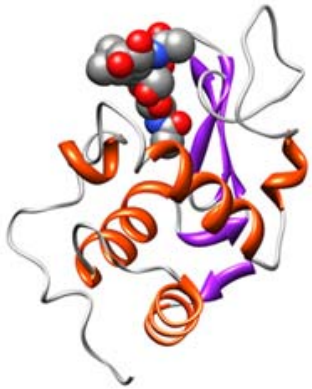
Quad	Count	s_{ijkl}	Quad	Count	s_{ijkl}	Quad	Count	s_{ijkl}	Quad	Count	s_{ijkl}
CCCC	4107297	-0.15377	CMOX	77	0.339447	MMNX	0	--	NNOS	6209	-0.28656
CCCM	1924	-0.93026	CMSS	2052	2.826573	MMOO	320	2.311659	NNOX	354	-0.70907
CCCN	4142684	-0.18067	CMSX	13	1.148817	MMOS	104	3.104429	NNSS	319	0.307157
CCCO	6462239	-0.03793	CMXX	6	1.935563	MMOX	3	2.386025	NNSX	6	-0.898
CCCS	297980	0.207795	CNNN	122810	-0.56586	MMSS	254	5.375177	NNXX	7	0.29148
CCCX	2996	-0.96834	CNNO	2117811	0.143315	MMSX	2	3.791851	NOOO	227156	0.121592
CCMM	157	0.96026	CNNS	16884	-0.37318	MMXX	0	--	NOOS	11871	-0.05545
CCMN	3758	-0.5452	CNNX	631	-0.97912	MNNN	1048	0.520184	NOOX	3214	0.198618
CCMO	6511	-0.35687	CNOO	2981894	0.241565	MNNO	1323	0.093906	NOSS	951	0.430162
CCMS	2320	0.776892	CNOS	99630	0.04635	MNNS	562	1.303999	NOSX	13	-0.9136
CCMX	15	-0.591	CNOX	2400	-0.75032	MNNX	6	0.153922	NOXX	66	0.914541
CCNN	1871781	-0.13036	CNSS	4318	0.56619	MNOO	4193	0.544515	NSSS	35	1.055088
CCNO	8544461	0.177683	CNSX	38	-0.96883	MNOS	352	0.74942	NSSX	0	--
CCNS	128008	-0.06485	CNXX	68	0.406432	MNOX	31	0.515747	NSXX	0	--
CCNX	2159	-1.01632	COOO	683049	0.028291	MNSS	793	2.985098	NXXX	3	2.452665
CCOO	3686844	0.063328	COOS	38976	-0.11057	MNSX	5	1.305273	OOOO	61473	0.105657
CCOS	205846	0.091103	COOX	24064	0.50151	MNXX	9	2.683083	OOOS	5019	-0.00255
CCOX	4995	-0.7024	COSS	4524	0.536074	MOOO	5790	1.111435	OOOX	9614	1.101242
CCSS	15467	0.849914	COSX	64	-0.79279	MOOS	167	0.676269	OOSS	331	0.222484
CCSX	148	-0.64875	COXX	84	0.447847	MOOX	171	1.508056	OOSX	45	-0.12365
CCXX	161	0.510349	CSSS	320	1.44474	MOSS	211	2.359752	OOXX	144	1.504034
CMMM	29	3.557768	CSSX	5	-0.01705	MOSX	4	1.158007	OSSS	38	1.040448
CMMN	164	1.249604	CSXX	4	0.707545	MOXX	55	3.418848	OSSX	3	0.282172
CMMO	293	1.451272	CXXX	12	2.483295	MSSS	62	3.8869	OSXX	0	--
CMMS	665	3.389144	MMMM	83	7.771426	MSSX	2	2.739925	OXXX	5	2.624158
CMMX	1	1.38783	MMMN	42	4.290048	MSXX	0	--	SSSS	11	2.686034
CMNN	2643	-0.12663	MMMO	31	4.107805	MXXX	16	5.763152	SSSX	0	--
CMNO	7243	-0.0402	MMMS	379	6.777	NNNN	5639	-0.7304	SSXX	0	--
CMNS	2610	1.098444	MMMX	0	--	NNNO	60175	-0.35461	SXXX	0	--
CMNX	30	-0.01957	MMNN	85	1.836638	NNNS	538	-0.82132	XXXX	0	--
CMOO	9551	0.33061	MMNO	113	1.608913	NNNX	39	-1.13953			
CMOS	1041	0.648899	MMNS	364	3.698853	NNOO	384854	0.224828			

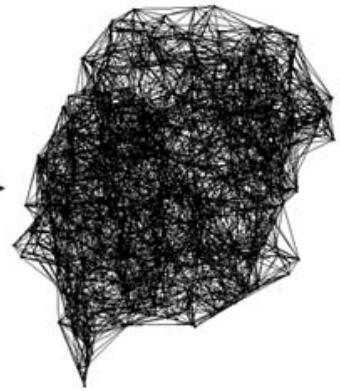
Topological Score (TS)

- Delaunay tessellation of any macromolecular structure yields an aggregate of tetrahedral simplices
- Each simplex can be scored using the all-atom four-body potential based on the quad present at the four vertices
- Topological score (or 'total potential') of the structure: the sum of all constituent simplices in the tessellation

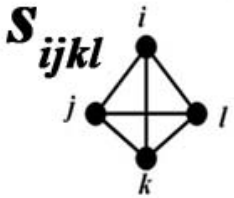


Topological Score Difference (ΔTS)

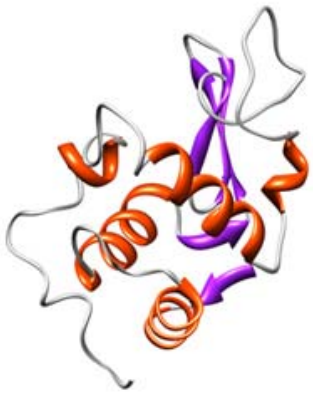


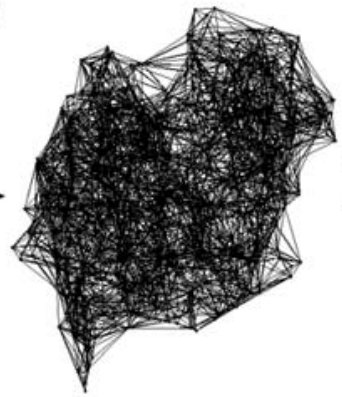


$$TS_{complex} = \Sigma s_{ijkl}$$



$$\Delta TS = TS_{complex} - TS_{protein}$$

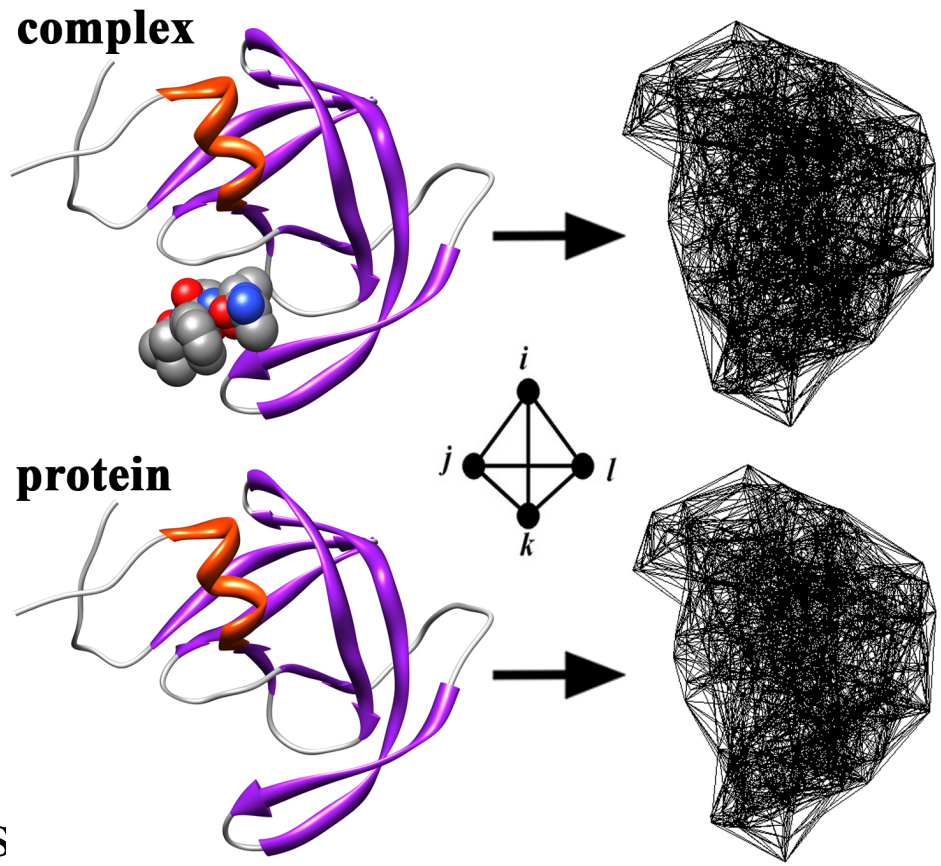




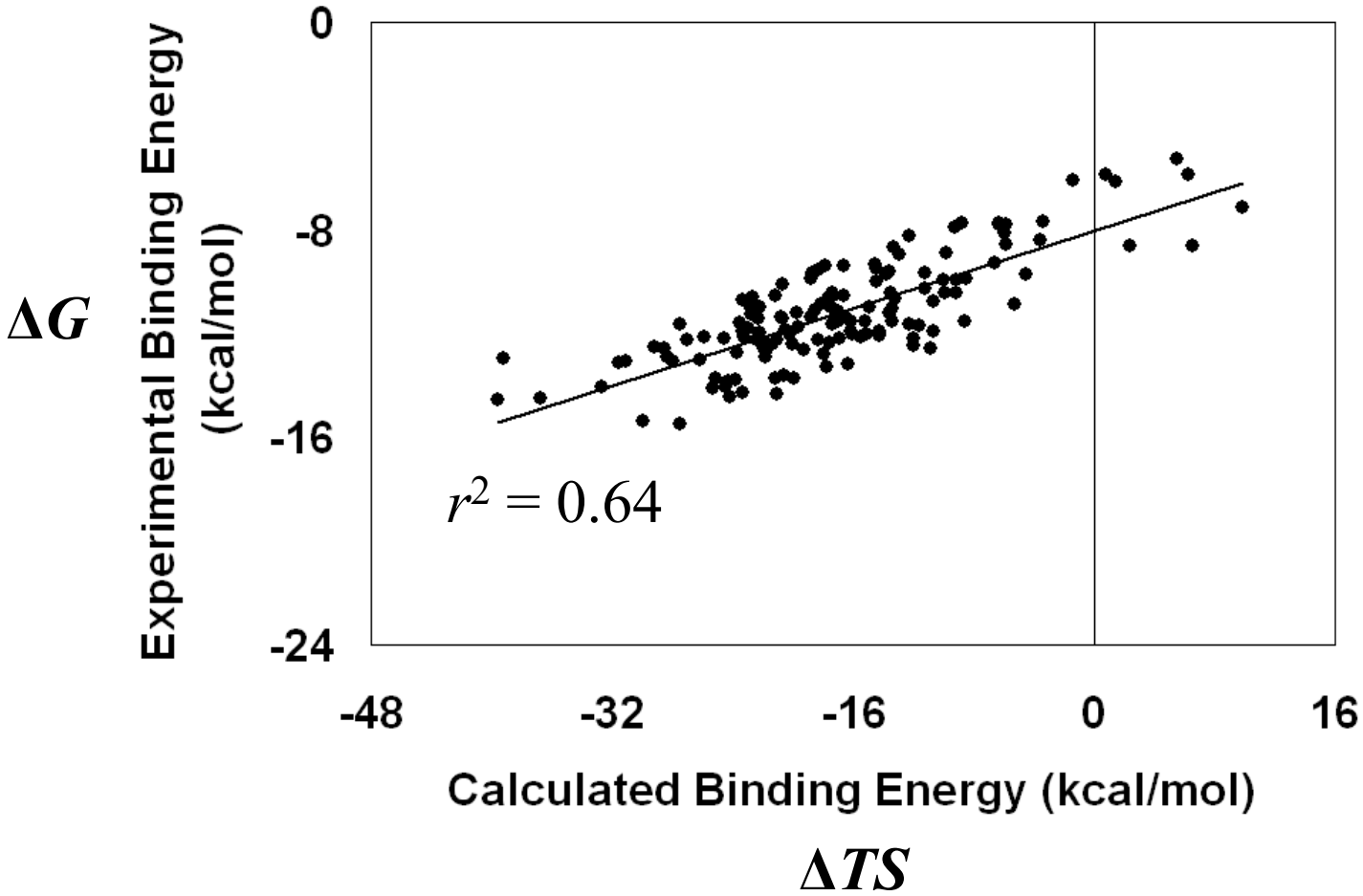
$$TS_{protein} = \Sigma s_{ijkl}$$

Application of ΔTS: Predicting Binding Energy of HIV-1 Protease Inhibitors

- MOAD – repository of exp. dissociation constants (k_d) for protein–ligand complexes whose structures are in PDB
- Found k_d values for 140 inhibitors of HIV-1 protease, with a PDB structure of the complex in each case
- Obtained exp. binding energy from k_d via $\Delta G = -RT\ln(k_d)$
- Calculated ΔTS for complexes



Predicting Binding Energy of HIV-1 Protease Inhibitors



References and Acknowledgments

- PDB (structure DB): http://www.rcsb.org/pdb
- MOAD (ligand binding DB): http://bindingmoad.org/
- 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/