Evaluation of Ligand-Receptor Binding Affinity with a Novel Statistical Scoring Function Based on Delaunay Tessellation of Protein-Ligand Interface

Alexander Tropsha

Laboratory for Molecular Modeling

University of North Carolina at Chapel Hill

OUTLINE

- Empirical scoring functions
- Statistical Geometry of Ligand-Receptor Interface
- Novel four-body scoring function based on Delaunay tessellation of protein ligand interface: correlation with experimental liagnd affinity

Scoring Functions for Ligand-Receptor Interaction

- Empirical: based on estimation of most important physical interactions at protein-ligand interface.
- Statistical, or knowledge-based: derived from the statistical analysis of protein-ligand complexes.

Typical Interaction Energy Terms in Empirical Scoring Functions

- Van der Vaals interaction
- Hydrogen bonding
- Solvation Effects
- Electrostatic interaction
- Conformational entropy

$$Score = \sum_{i=0}^{n} C_i F_i + C_0$$

Ci values are determined from regression analysis

Examples of empirical scoring functions

- VALIDATE (*J.Am.Chem.Soc.1996,118,3959-3969*)
  51 complexes, R^2=0.85, 12 terms

- SCORE (*J.Mol.Model.1998,4,379-394*)
  170 complexes, R^2=0.78, 11 terms

Knowledge-based (Statistical) Potentials

- Two Body distance dependent potentials

  PMF Muegge, I.; Martin, Y.C.; *J.Med.Chem.*1999, *42*, 791-804

  BLEEP Mitchell, J.B.; Laskowski R.A.; Alex A.; Thornton, J.M.; *J. Comp. Chem.* 1999, *20*,1165-1176

  DrugScore Gohlke, H.; Hendlich, M.; Klebe,G.; *J Mol Biol* 2000, *295*, 337-356
- Two Body contact potential

  SMoG DeWitte, R. S.; Shakhnovich, E.I. *J Am. Chem. Soc.* 1996, *118*,11733-11744

  SMoG2001 Ishchenko. A. V.; Shakhnovich, E. I.; *J. Med. Chem*. 2002, 45, 2770-2780
- Four-Body contact potential

Scoring functions derived from statistical analysis of L/R complexes

- Potential of Mean Force

$$PMF = -k_B T \sum_{ij=1}^{npairscutoff} \sum_{r=0} \ln\left[\frac{g_{ij}(r)}{f_{ij}}\right]$$

Distance-dependent scoring functions

- Muegge & Martin method (J.Med.Chem.1999,42,791-804).
  - training set: 697 x-ray structures (resolution below 2.5Å)
  - 16 atom types for proteins, 34 for ligands
  - 77 complexes: R^2~0.6
- BLEEP *(J.Comp.Chem. 1999, 1165-1176)*
  - training set: x-ray structures (resolution below 2.5Å)
  - 38 atom types for nonmetals, 18 metals
  - 90 complexes: R^2 ~0.3-0.5

Contact Potential

- Two body contact potential (SMoG)

$$B(s_p, s_l) = -\ln\left[\frac{p(s_p, s_l)}{p_{ref}}\right]$$

(s_p, s_l): protein-ligand contacts

$B(s_p, s_l)$: potential parameter of protein-ligand contacts (s_p, s_l)

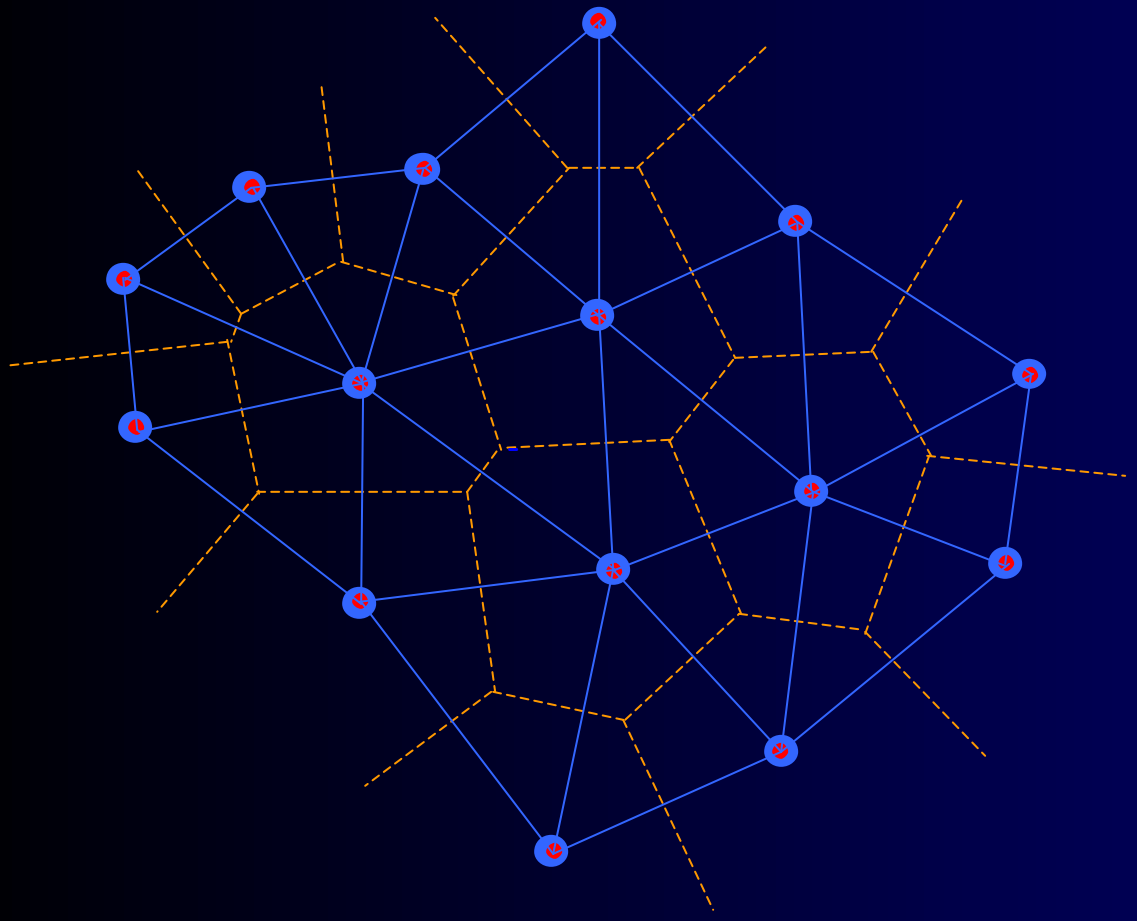
- Problems of most scoring functions
  - Relatively low correlation with experimental affinity
  - predictive ability is typically limited to compounds similar to those in the training set.

Higher-order Potentials?

- Interactions between two atoms can be affected by the presence of surrounding atoms.
- Inclusion of all distance-dependent multi-body terms is impossible.
  For 1000 atoms, there are 499,500 two body terms and 166,167,000 three body terms.

➢ Four-Body contact scoring function based on 3D Delaunay Tessellation of protein ligand interface.

Voronoi/Delaunay Tessellation in 2D



Delaunay Triangulation is Created as a Dual Tessellation

Nearest neighbors are unambiguously defined in sets of three (in 2D) and in sets of four (in 3D)

Some history: Three-Dimensional Tessellation of protein structure

- United residue (side chain centroid) representation
- No chain connectivity
- Protein structure represented as an aggregate of space filling, irregular tetrahedra
- Unique and objective recognition of nearest neighbor residues in sets of FOUR (Tetrahedral Quadruplets)

Four Body Compositional Propensities of Delaunay Simplices

$$q_{ijkl} = \log \frac{f_{ijkl}}{p_{ijkl}}$$

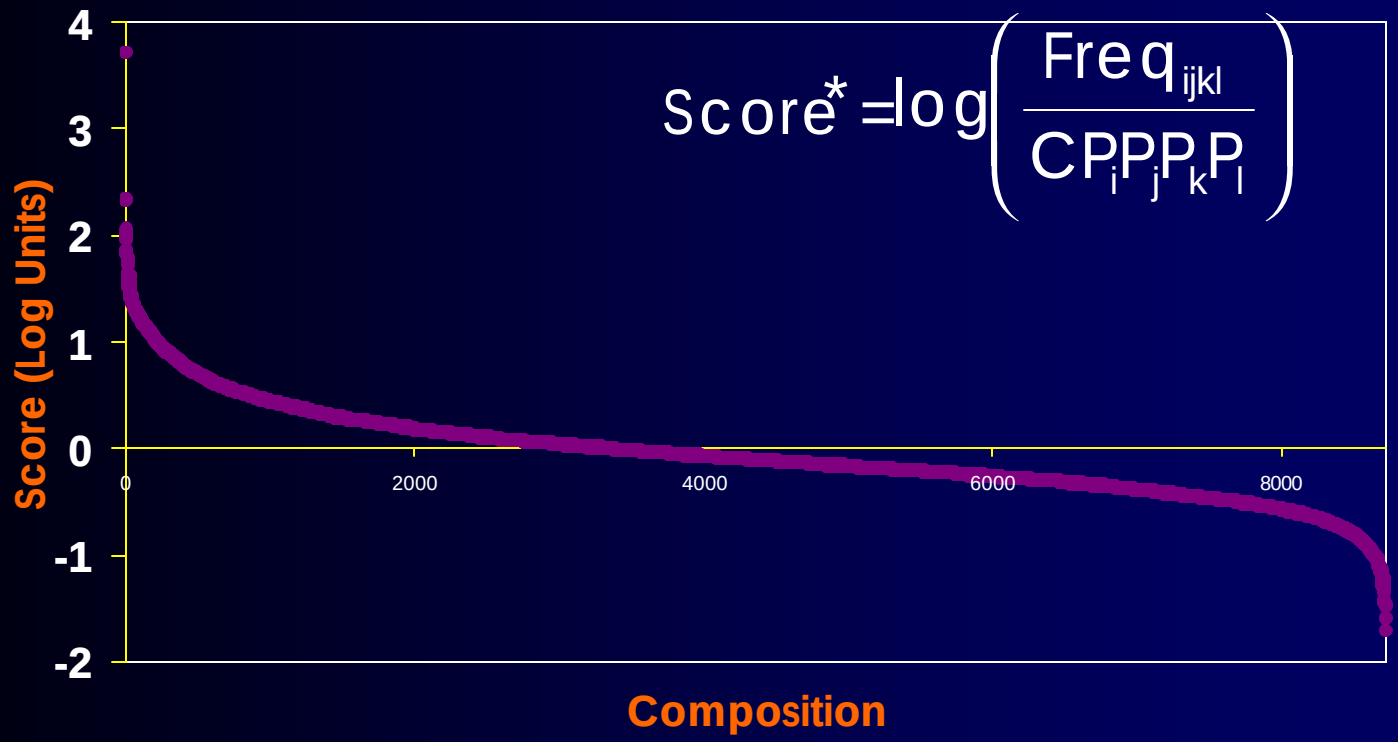
f- observed quadruplet frequency

$$p_{ijkl} = c a_i a_j a_k a_l$$

a- individual AA frequency

$$C = \frac{4!}{\prod_i^{\mathrm{n}} (t_i!)}$$

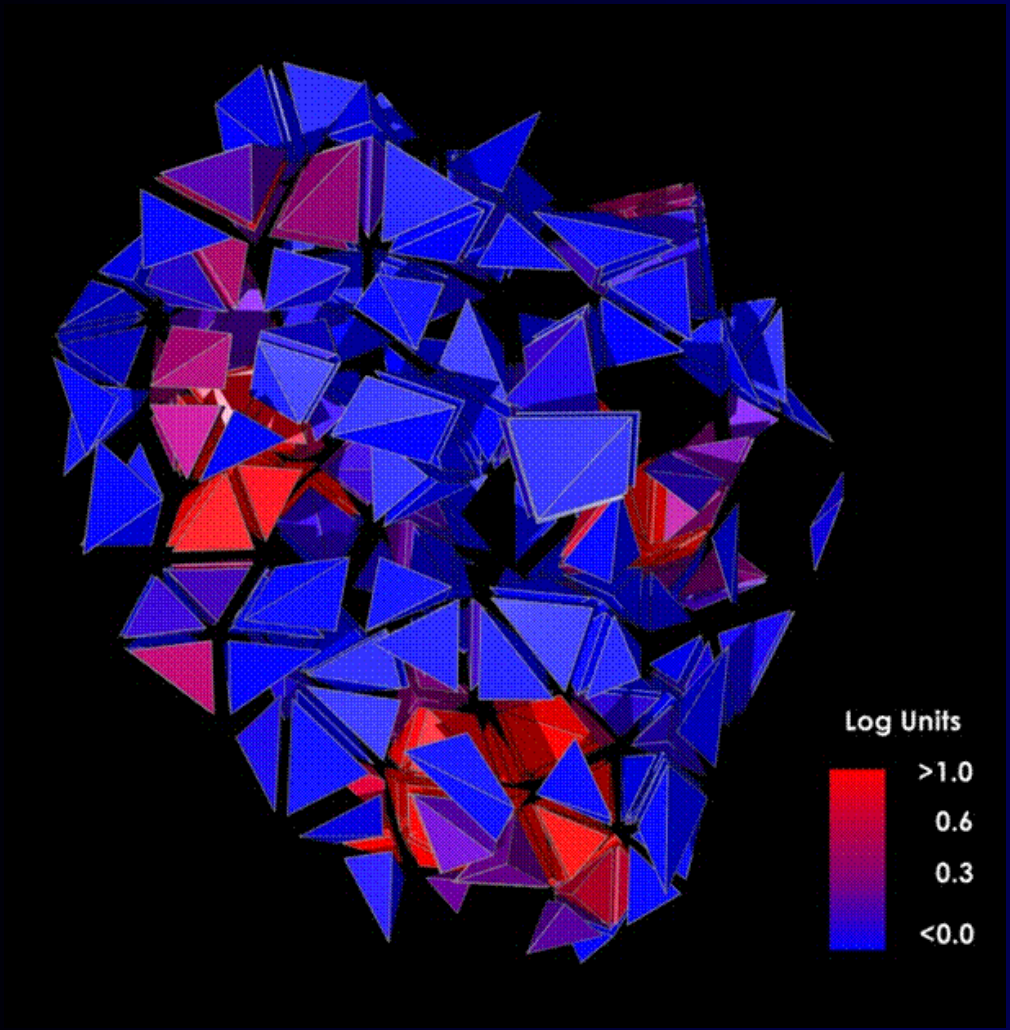
Likelihood Scores for 8724 Compositions



1100 Protein Chain Database <20% Id. 2.5 Å
254,013 Total Four-Body Contacts

* Tropsha A, Singh R, Vaisman I, Zheng W. *Pac Symp Biocomput* . 614-23 (1996)
Dunbrack, R. Culled PDB: http://www.fccc.edu/research/labs/dunbrack/culledpdb.html

Visualization of Likelihood Mapping

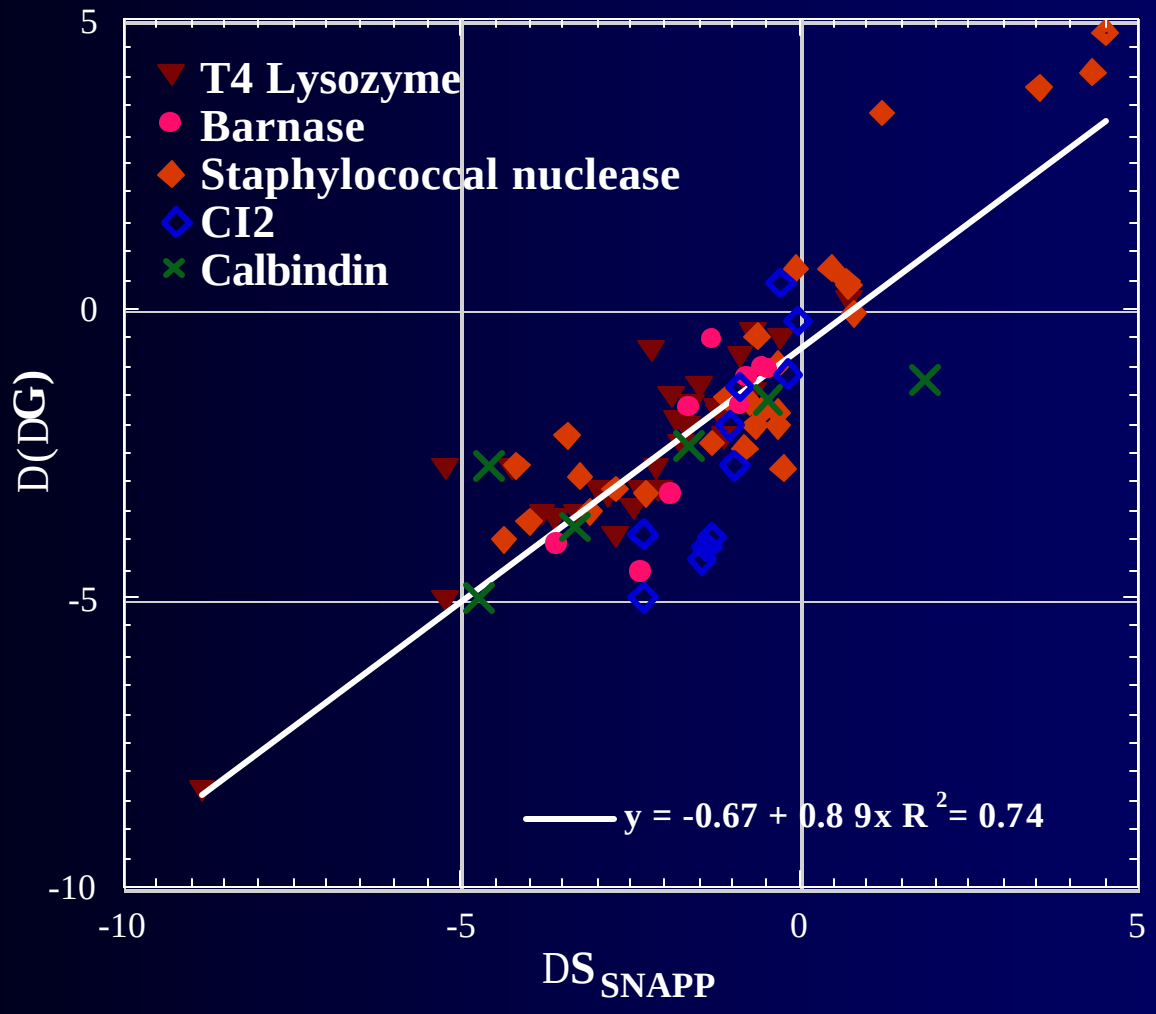


Protein *Tiles Painted* Using Log-Likelihood Scores

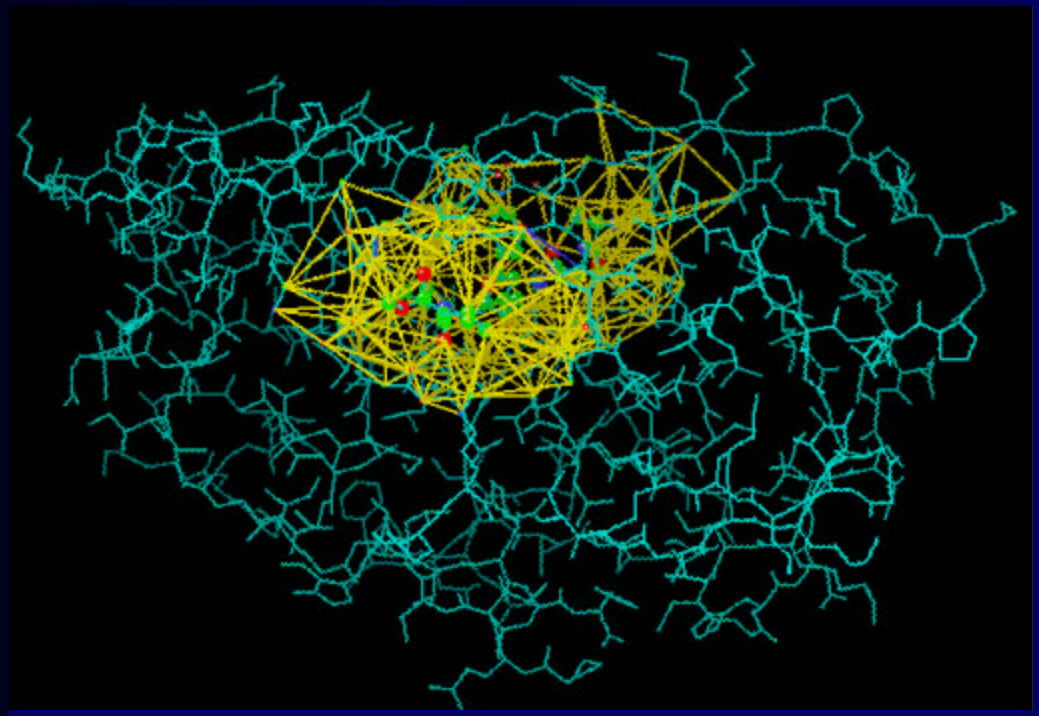
Hydrophobic Core Organization is Mapped

A Measure of *Hydrophobic Fitness* ? (Levitt)

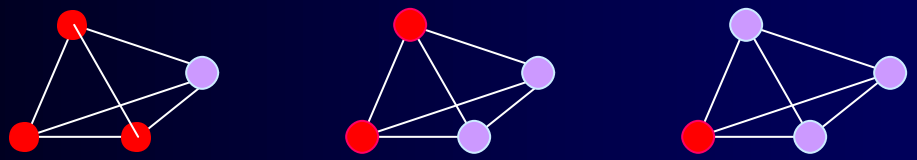
Correlation between calculated and predicted stability of hydrophobic core mutants (Carter et al., JMB, (2001) 311, 625-638)



3D Delaunay Tessellation in the Active Site



Three Types of Tetrahedra at Protein-ligand Interface



RLLL RRLL RRRL

RRRL: Formed by 3 receptor atoms and 1 ligand atoms
RRLL: Formed by 2 receptor atoms and 2 ligand atoms
RLLL: Formed by 1 receptor atoms and 3 ligand atoms

Atom Types

9 Ligand atom types

Cp	Polar carbon
Cnp	Non-polar carbon
Npl	Nitrogen without lone pair
Nlp	Nitrogen with lone pair
O3	Sp3 oxygen
O2	Sp2 oxygen
S	Sulfur
X	Halogen
P	Phosphorus

7 Receptor atom types

Cp_R	Polar carbon
Cnp_R	Non-polar carbon
N+_R	Nitrogen with formal charge
N_R	Nitrogen without formal charge
O-_R	Oxygen with formal negative charge
O_R	Oxygen without formal negative charge
S_R	Sulfur

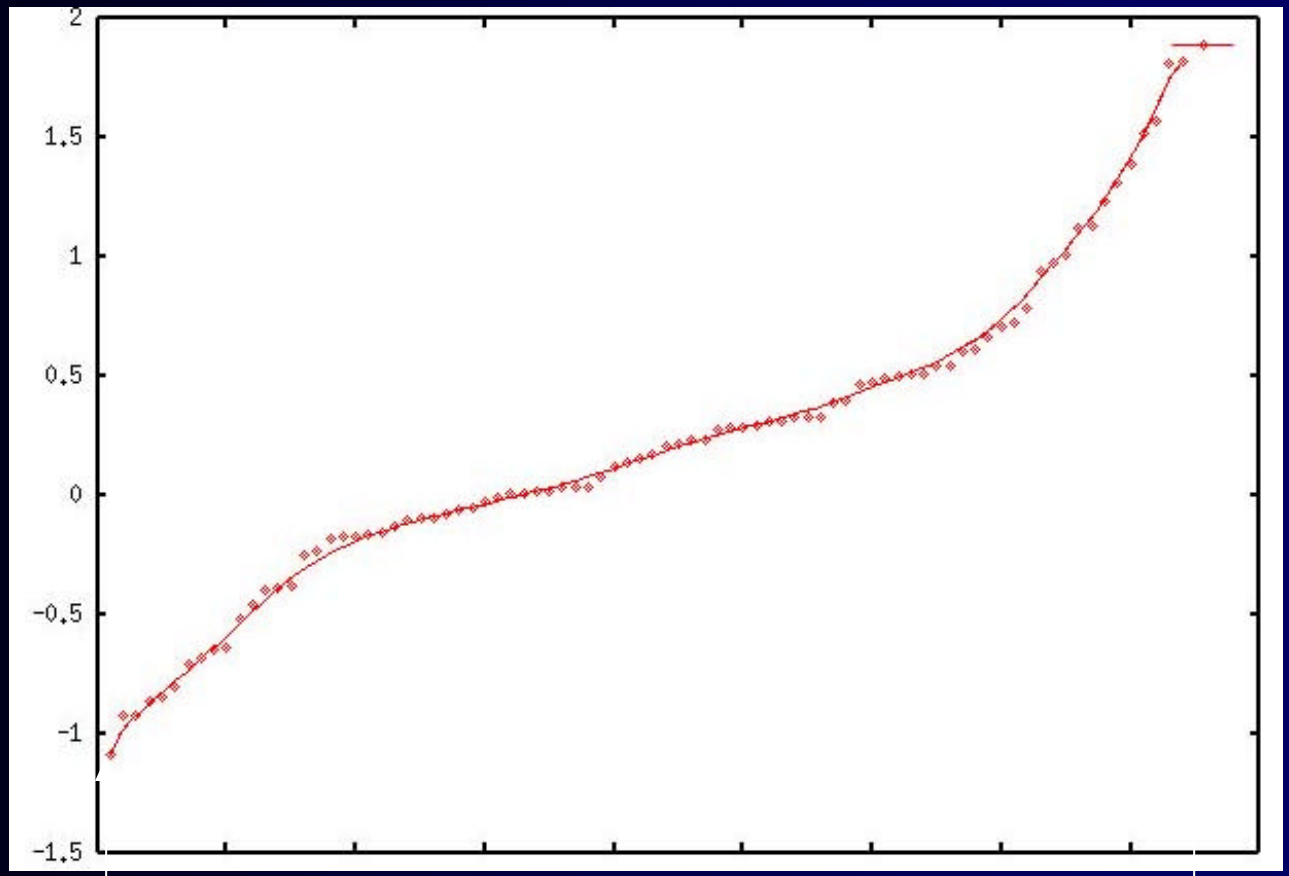
Four-Body Statistical Contact Scoring Function Based on Delaunay Tessellation

$$E_{RRRL} = -\ln \frac{f_{RRRL}}{f_{RRR} f_{L}}$$

$$E_{RRLL} = -\ln \frac{f_{RRLL}}{f_{RR} f_{LL}}$$

$$E_{RLLL} = -\ln \frac{f_{RLLL}}{f_{R} f_{LLL}}$$

Potential Distribution for RRRL type tetrahedra



Test sets

PMF test set: 68 compounds

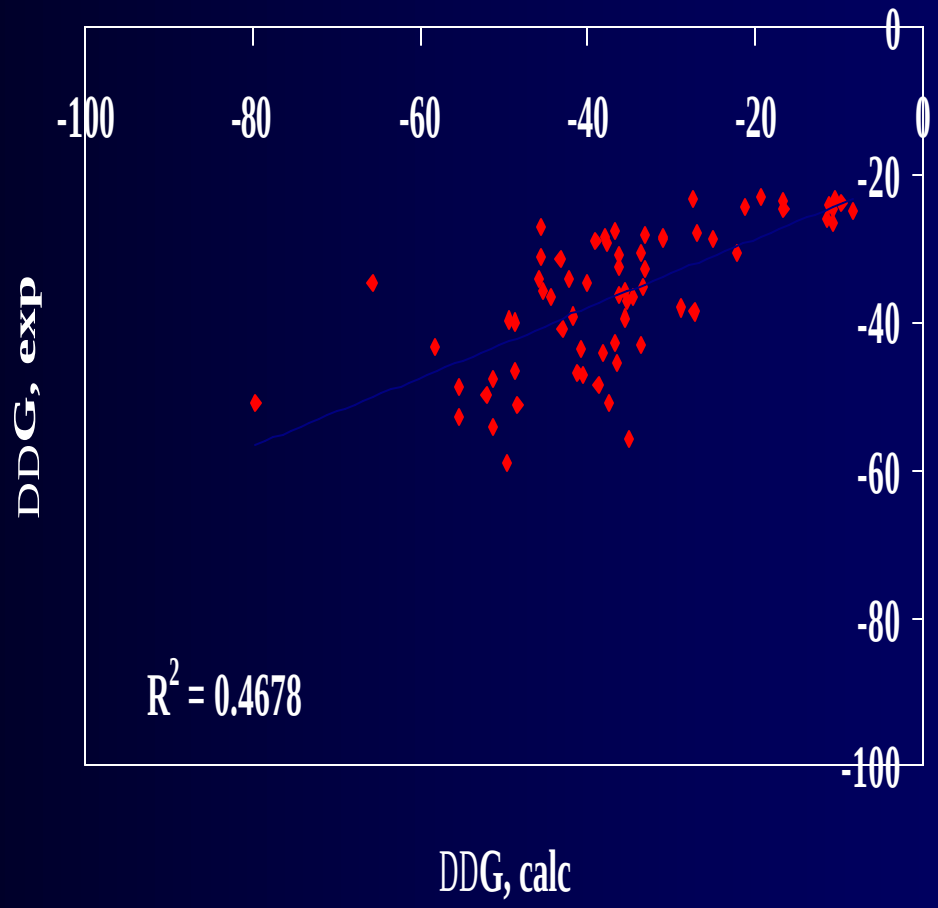
Serine Proteases: 16
Metalloproteases: 15
L-arabinose binding protein: 9
Endothiapepsin: 11
Other proteins: 17

SMoG2002 test set: 111 compounds

Aspartic Proteases: 18 complexes
Serine Proteases: 20
Metalloproteases: 22
Human Carbonic Anhydrase II (HCA, 19 Complexes) Sugar-Binding Proteins (14 Complexes) Endothiapepsin (11 Complexes) Purine Nucleoside Phosphatase (PNP, 5 Complexes) Subset 8: Other Proteins (10 Complexes)

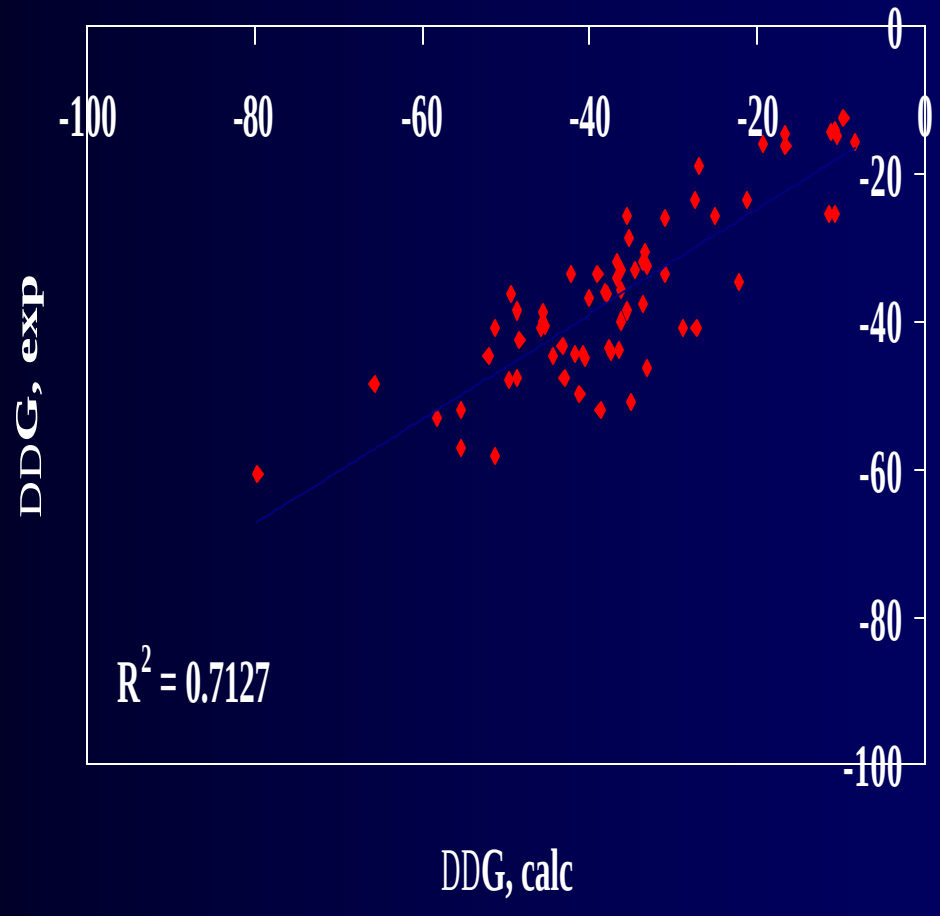
Correlation between experimental and calculated binding free energy for PMF dataset using four-body scoring function

$$E = a\sum E_{RRRL} + b\sum E_{RRLL} + g\sum E_{RLLL}$$



Correlation between experimental and calculated $\Delta\Delta G_b$ for PMF dataset using combined two- and four-body scoring function

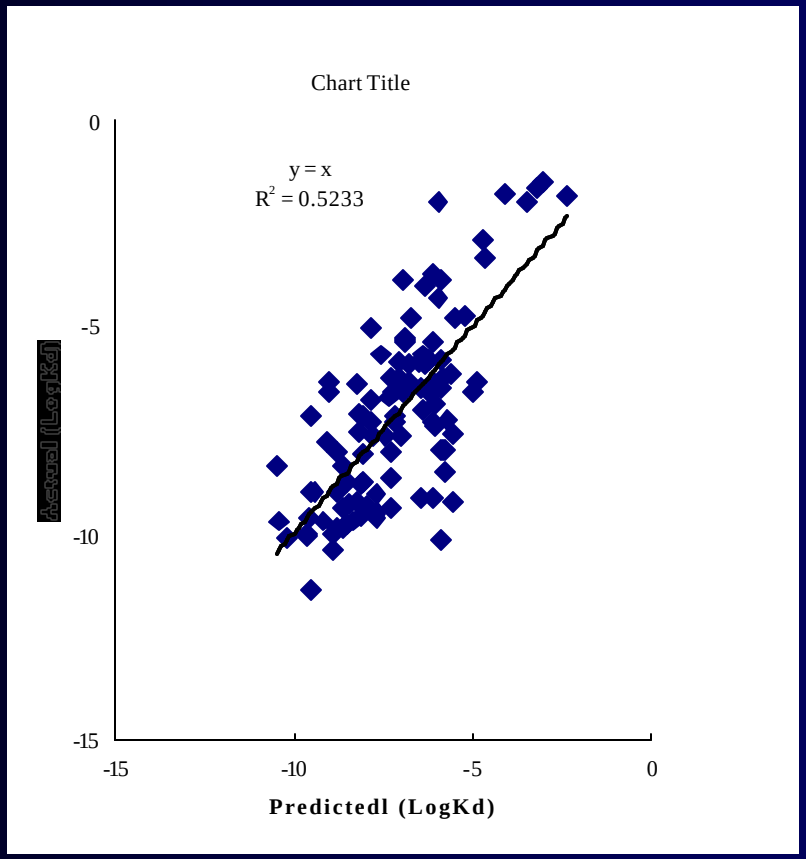
$$E = a\sum E_{ij}(r) + b_1\sum E_{RRRL} + b_2\sum E_{RRLL} + b_3\sum E_{RLLL}$$



Comparison between DT-based and two-body statistical potentials (For Muegge dataset, 67 complexes)

POTENTIAL	R^2
2-body contact	0.25
2-body distance-dependent	0.58
4-body contact	0.46
4-body contact + 2-body distance-dependent	0.71

Correlation between experimental and calculated $\Delta\Delta G_b$ for SMoG2001 dataset using combined two- and four-body scoring function (107 complexes)



Comparison of Current Scoring Functions

	Training Set size	Test Set size	Test Set R^2
BLEEP	351	90	0.53
PMF	697	77	0.61
SMoG96	120	46	0.42
SMoG2001	725	111	0.436
DT2001	319	67	0.71
DT2002	319	107	0.54

CONCLUSIONS

- Four-body contact statistical scoring function seems to outperform existing lower-order scoring schemes
- Combination with existing potentials improves correlation with experimental binding
- Modification of existing scoring function may improve the results:
  - Bigger training set
  - Different definitions of atom types
  - Additional terms (e.g., geometry of tetrahedra)

ACKNOWLEDGMENTS

- UNC associates

  Former:
  - Weifan ZHENG
  - Sung Jin CHO
  - Xin Chen
  - Stephen CAMMER

Protein folding group:

–John GRIER

–Jun FENG

–Shuxing ZHANG

–David BOSTICK

–Bala KRISHNAMOORTHY

–Ruchir SHAH

–Sagar KHARE

QSAR group:
- Alex GOLBRAIKH
- Yun-De XIAO
- Min SHEN
- Scott OLOFF
- Yuanyuan QIAO

- Collaborators:
  - C. Breneman (RPI)
  - D. Bonchev (Texas A&M)
  - R. Hormann (R&H)
  - C. Reynolds R&H→Ortho)
  - V. Gombar (GSK)
  - E. Gilford (Pfizer)
- Funding
  - NIH
  - NSF
  - Millenium
  - Rohm & Haas
  - Ortho McNail
  - GSK
  - Pfizer