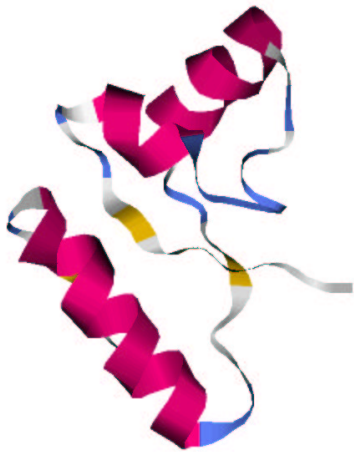


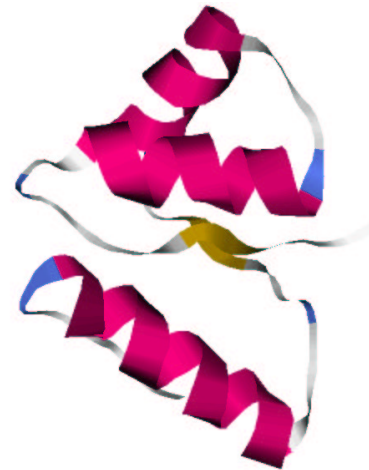
Statistical Potentials Based on Alpha-Shapes



Leo Guibas



Patrice Koehl
Stanford University



Afra Zomorodian



Focus

- Protein Structure Prediction
 - ★ Input: target sequence
 - ★ Output: three-dimensional structure
 - ★ Method:
 - * Generate a large number **decoys** (possible shapes)
 - * Select “best” and hopefully **native** fold
- Compare using **cRMS**

State of the Art

target	length	best decoy		best submitted	
		cRMS	length	cRMS	length
T087-A	192	5.3	214	6.5	128
T087-B	118	4.8	124	6.5	85
T091	109	3.1	90	6.1	85
T095	244	3.8	178	5.0, 2.9	139, 120
T096-B	160	4.9	123	5.7	63
T097	105	3.8	100	4.6	81

- Good methods for generating decoys
- Not so good at selection
 - ★ very hard problem
 - ★ not well understood
- We wish to replace cRMS with a **potential**

Potentials

- “Physical”
 - ★ Idea: native has minimum energy
 - ★ Problem: No ideal energy function
 - ★ Terms: van der Waals, electrostatics, hydrophobic effects (?)

Potentials

- “Physical”
 - ★ Idea: native has minimum energy
 - ★ Problem: No ideal energy function
 - ★ Terms: van der Waals, electrostatics, hydrophobic effects (?)
- Data-base derived
 - ★ Idea: native “looks” like a protein
 - ★ Problem: need to quantify “looks”
 - ★ Pairwise potential (1979)
 - ★ Residue-based (Sippl 1990)
 - ★ Atom-based (Samudrala et al. 1997)

Method

- Potential:

- ★ $E \propto -\ln f(r)$

- ★ $\Pi = -\sum_{i \neq j} \ln(\Pr \{X_{i,j} | Y_{i,j}\})$

- ★ X : pair of certain type, e.g. CYS-CYS

- ★ Y : pair is at distance r

- ★ From database:

$$\Pr \{X | Y\} = \frac{\Pr \{X\} \cdot \Pr \{Y | X\}}{\Pr \{Y\}}$$

Method

- Potential:

- ★ $E \propto -\ln f(r)$

- ★ $\Pi = -\sum_{i \neq j} \ln(\Pr \{X_{i,j} | Y_{i,j}\})$

- ★ X : pair of certain type, e.g. CYS-CYS

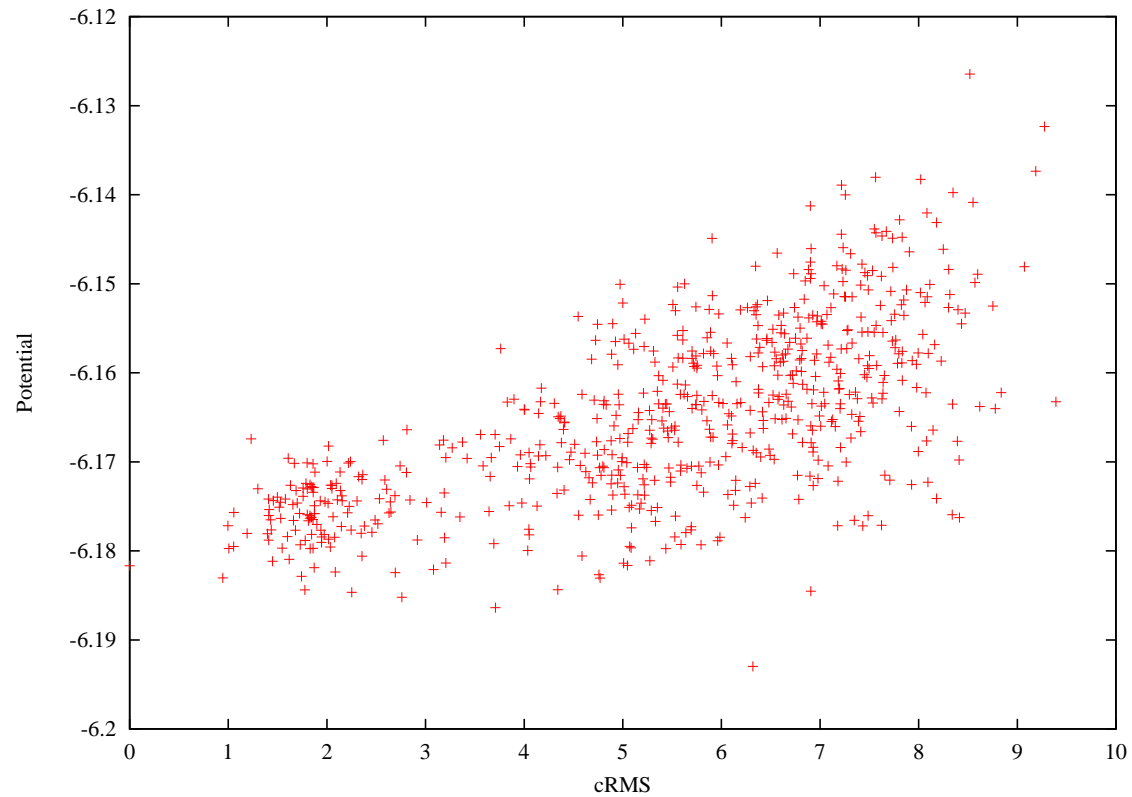
- ★ Y : pair is at distance r

- ★ From database:

$$\Pr \{X | Y\} = \frac{\Pr \{X\} \cdot \Pr \{Y | X\}}{\Pr \{Y\}}$$

- Verification

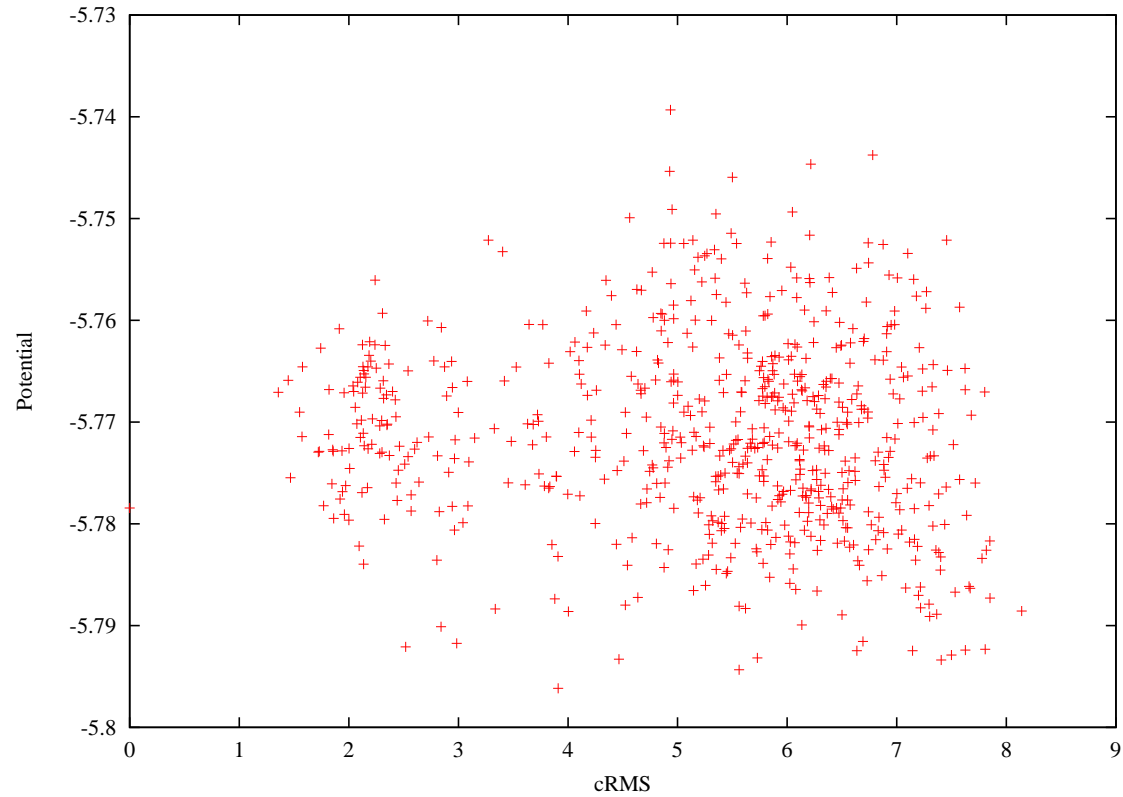
3icb



- 653 decoys, correlation 0.66¹

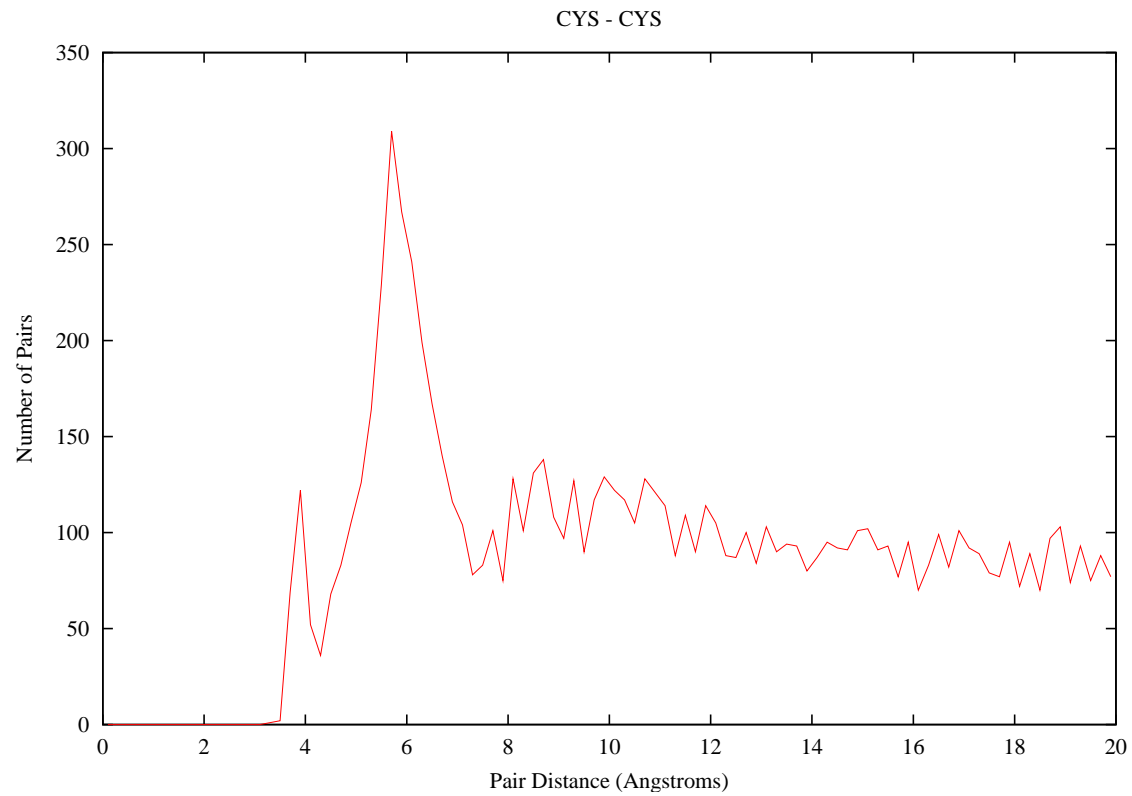
¹[Park & Levitt 1996]

4rxn

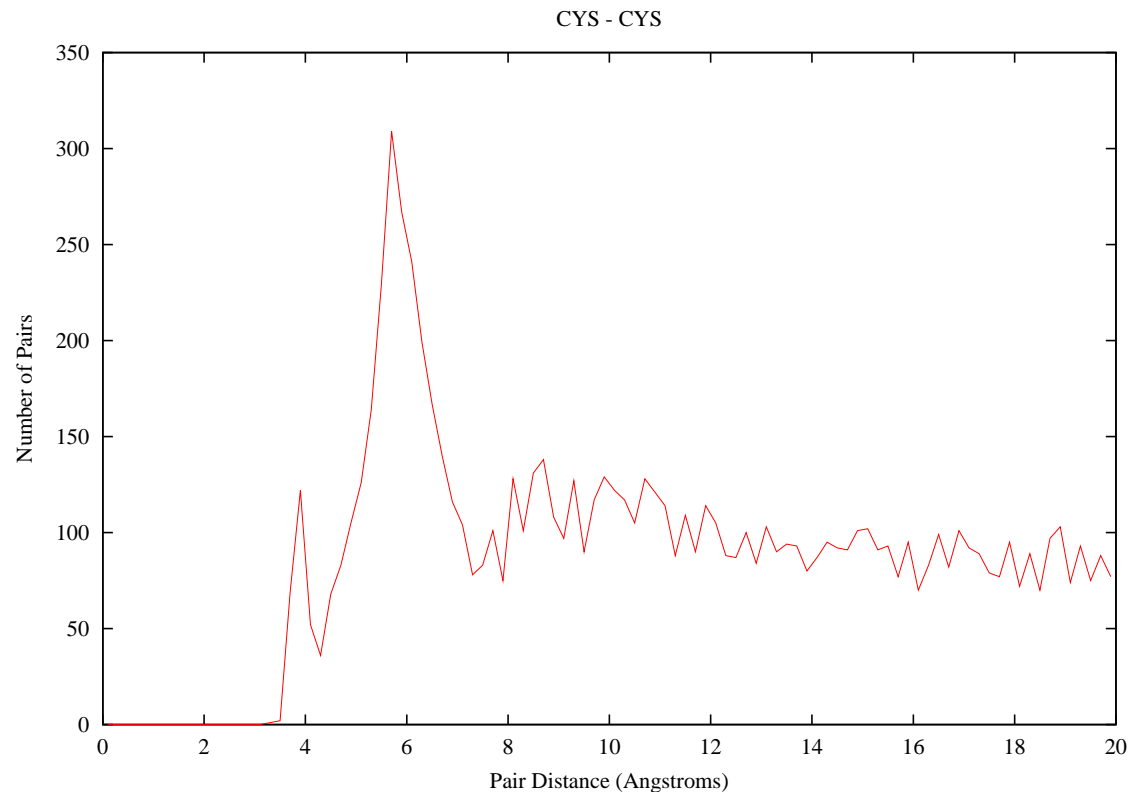


- 677 decoys, correlation: -0.08

Signal and Noise

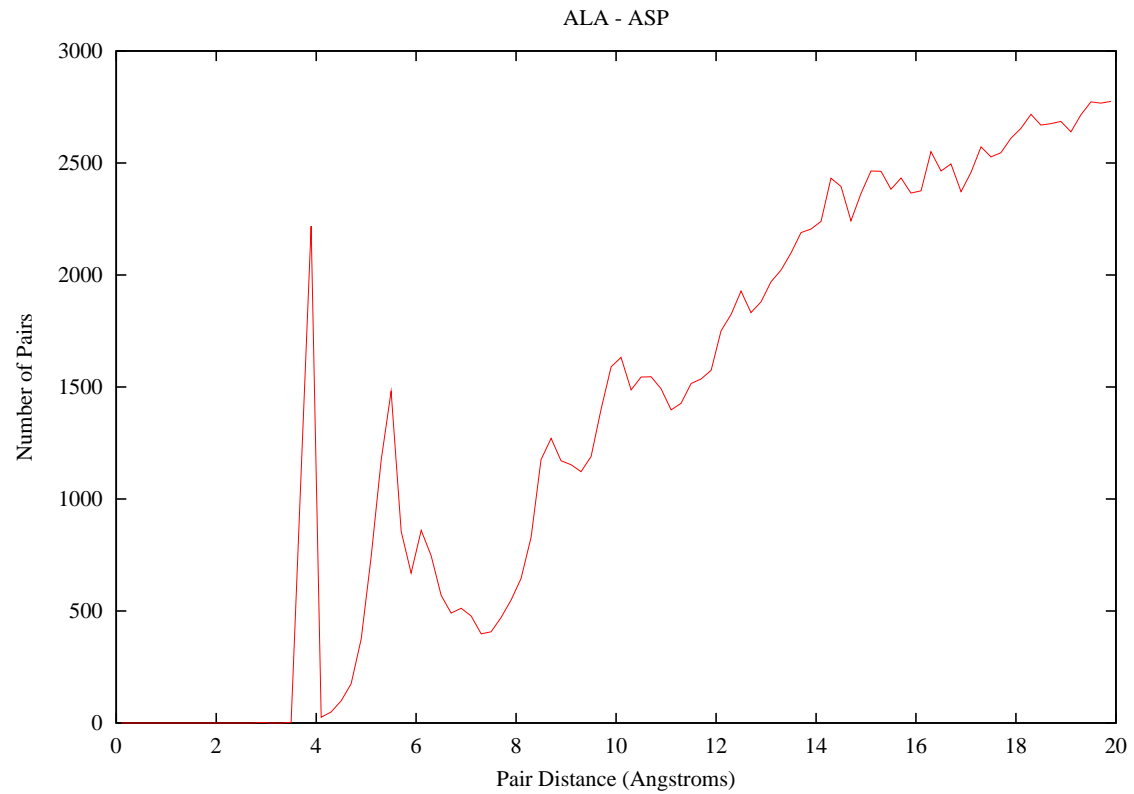


Signal and Noise

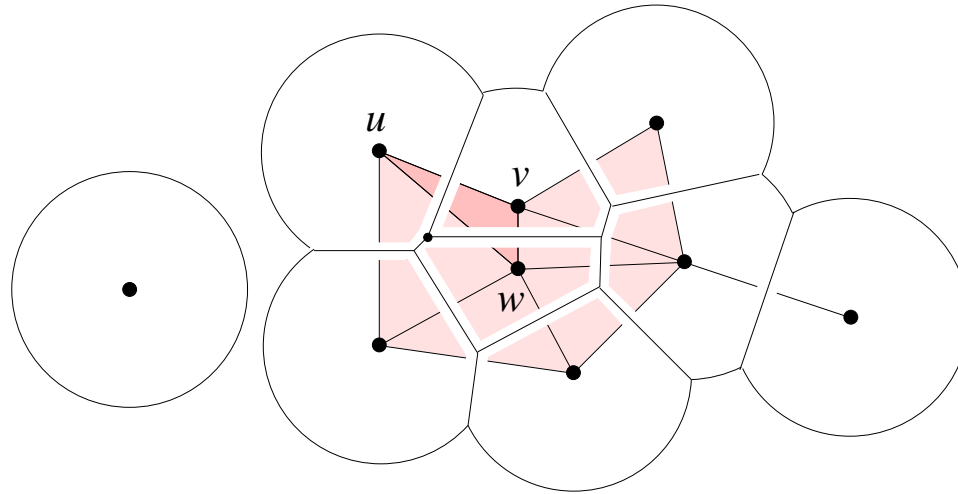


Problem: $O(n^2)$ distances, $O(n)$ matter

Signal and Noise

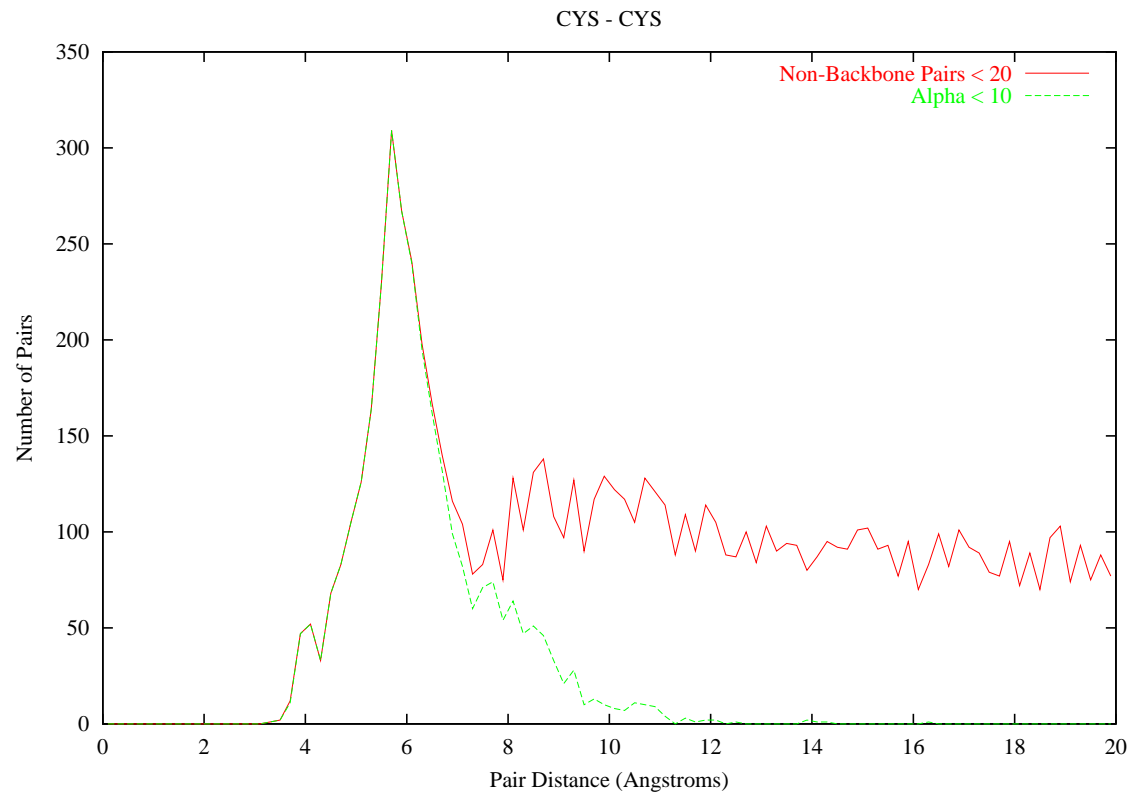


Alpha Complex

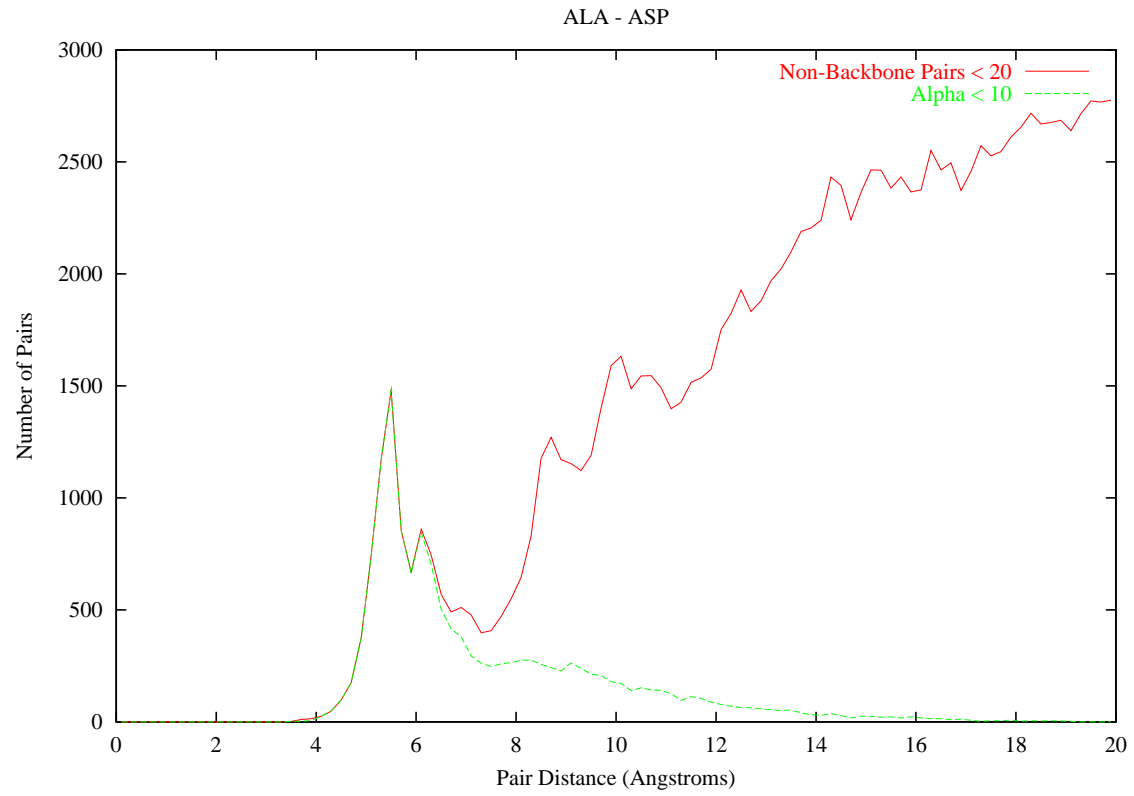


- van der Waals model
- Captures the topology of the ball set
- Subcomplex of Delaunay

Filtering



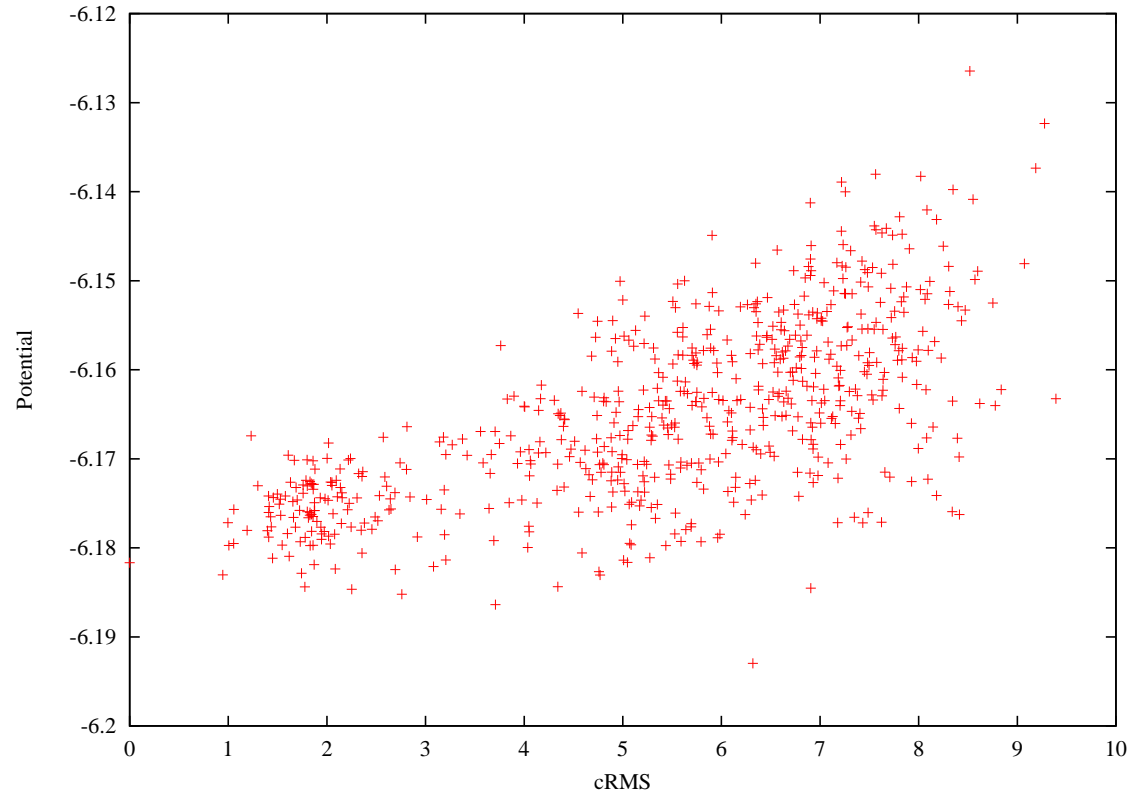
Filtering



Database

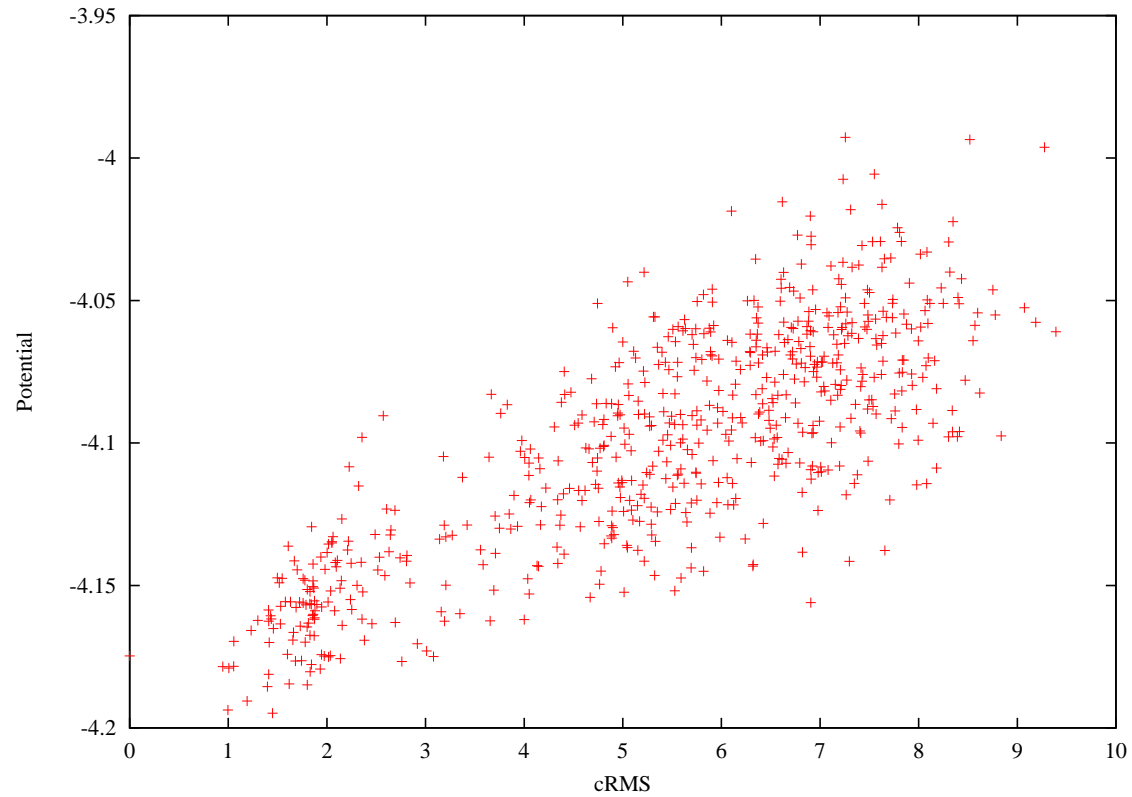
- 2,145 domains from SCOP
- 29,654,812 C^α pairs (**edges**) with distance $\leq 20\text{\AA}$
- Compute α -complex for all domains
- $\alpha = 10$ gives 3,643,018 C^α edges
- $\approx 12.3\%$ of possible non-backbone pairs
- 210 types

3icb – (all pairs)



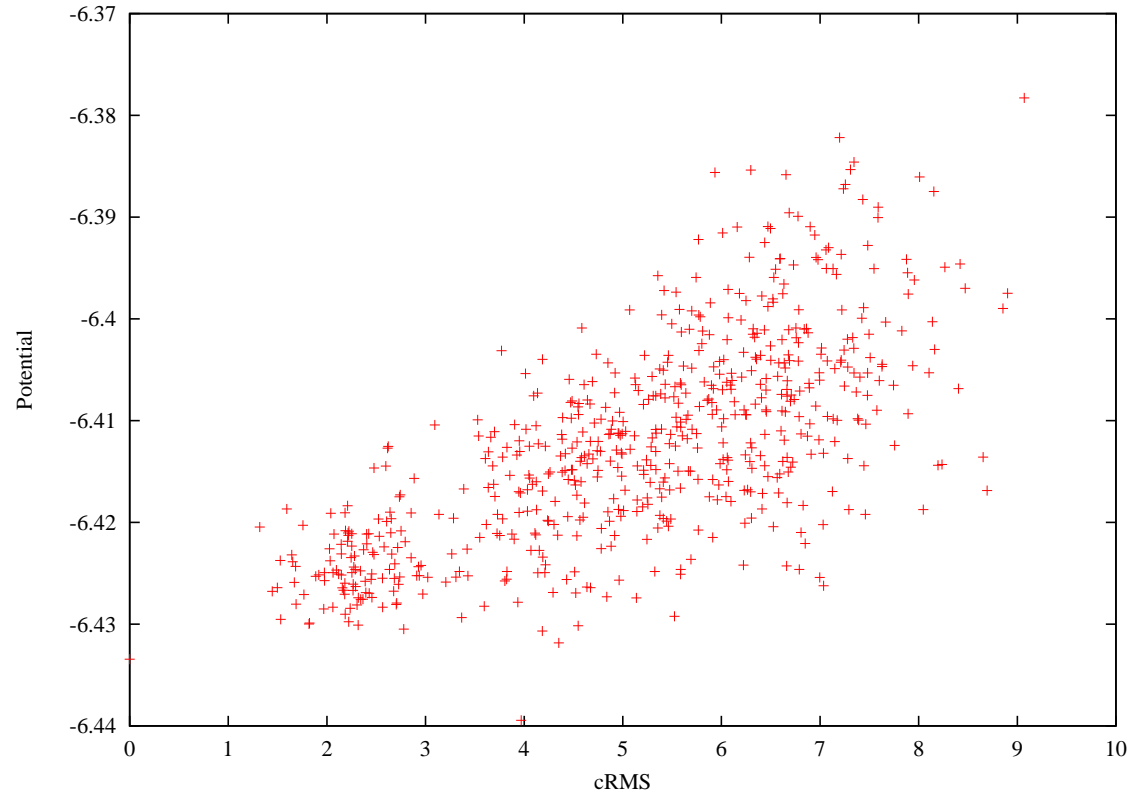
- 653 decoys, correlation 0.66

3icb – (alpha complex)



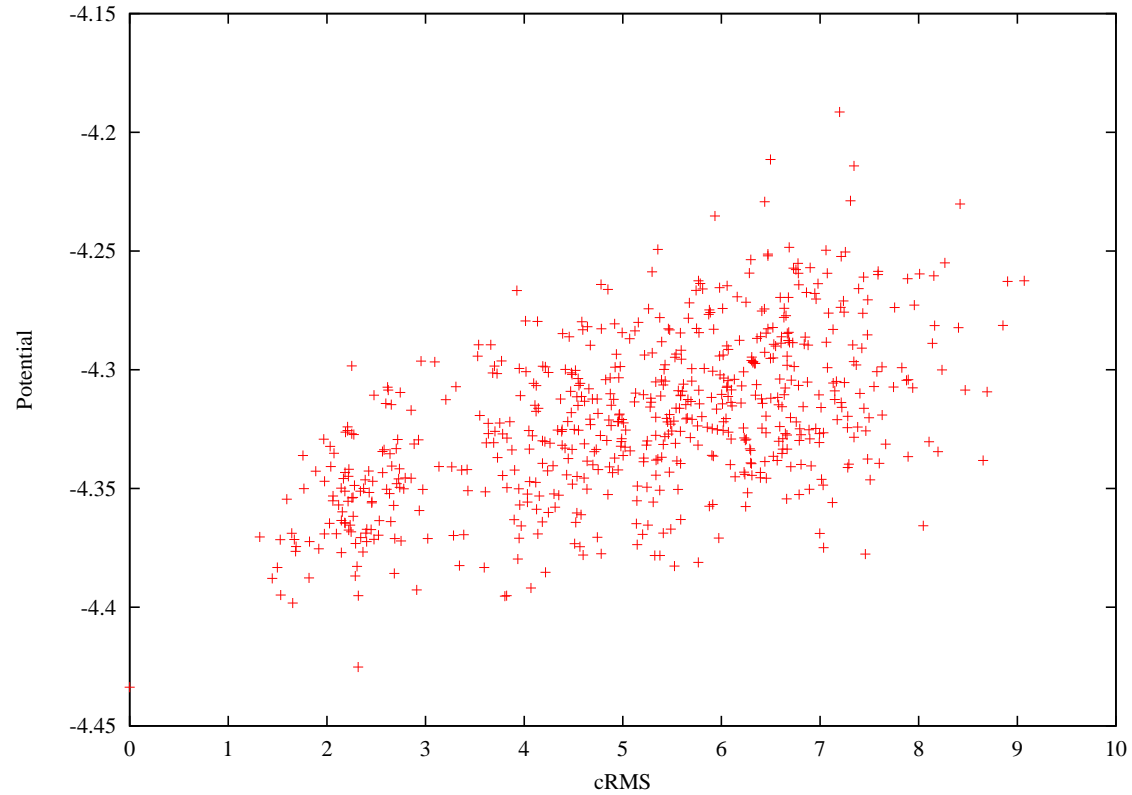
- 653 decoys, correlation 0.79

1ctf – (all pairs)



- 630 decoys, correlation 0.70

1ctf – (alpha complex)



- 630 decoys, correlation 0.56

Correlation

protein	size	type	correlation	
			ap	ac
1ctf	74	a+b	0.70	0.56
1r69	69	a	0.31	0.43
1sn3	65	a	-0.04	0.002
2cro	75	a	0.32	0.52
3icb	75	a	0.66	0.79
4pti	58	small	0.18	0.04
4rxn	54	small	-0.08	-0.22

Improved Selection

protein	type	best cRMS	cRMS		rank	
			ap	ac	ap	ac
1ctf	a+b	1.32	3.97	2.32	156	57
1r69	a	0.88	5.35	4.47	410	268
1sn3	a	1.31	6.56	6.36	428	383
2cro	a	0.81	4.19	1.87	246	24
3icb	a	0.94	6.32	1.45	387	16
4pti	small	1.41	6.28	4.75	424	169
4rxn	small	1.36	3.91	7.01	140	606

Experiments

- Decoy 'Я' Us datasets
- All atom databases
- Backbone databases
- 3 body, 4 body potentials
- CHARMM classification
- Significance measures

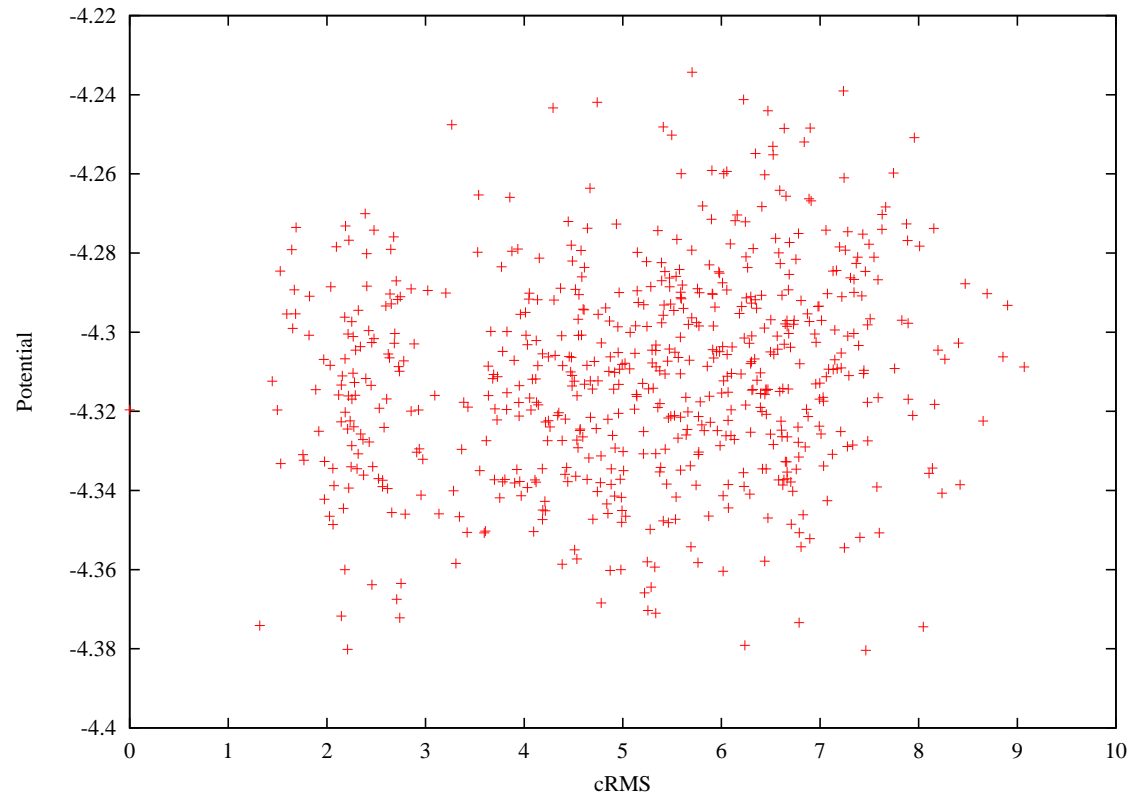
Conclusion

- Goal: discrimination
- Observation: not all pairs carry information
- Idea: filter using geometry
- Method: use alpha complex edges
- Result: better selection, but nothing spectacular

Discussion

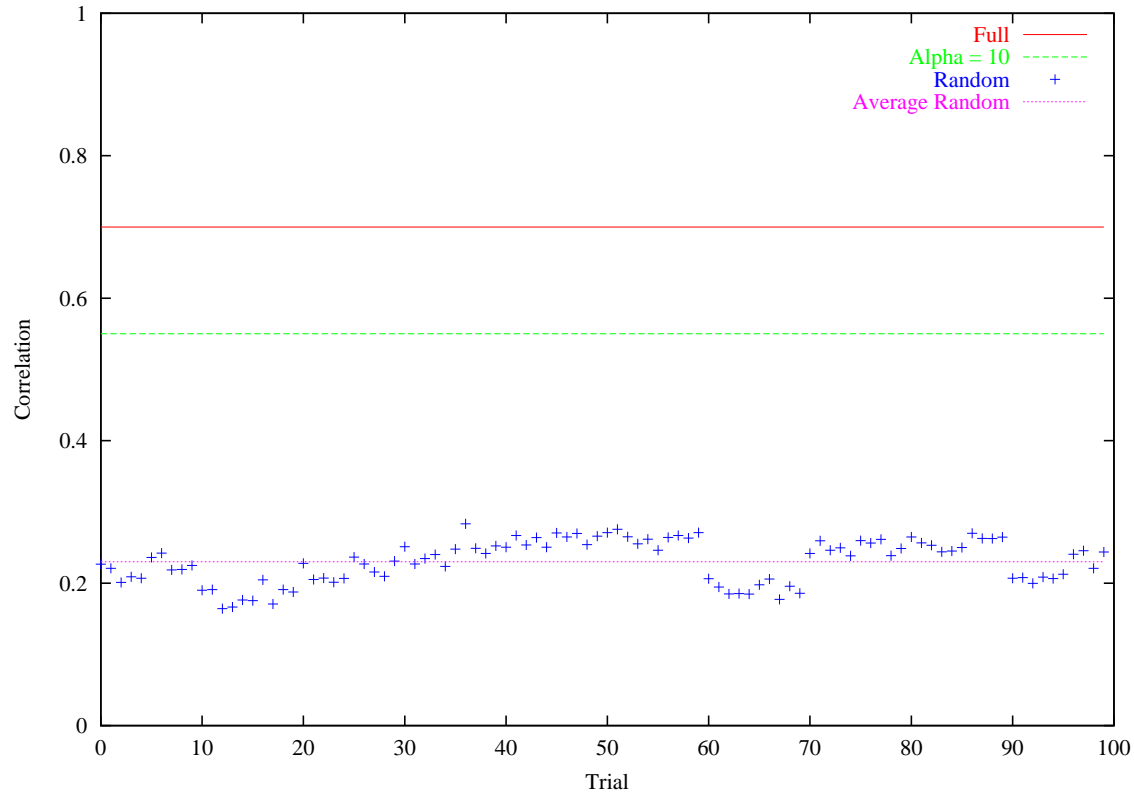
- Future work:
 - ★ Restrict to domains
 - ★ Alternate functions and Multivariate regression
 - ★ No distance dependence
- Issues:
 - ★ Why cRMS?
 - ★ Induced Problems: rigidity, local computation

1ctf – Random 12.3% Database



- 630 decoys, correlation 0.16

Random Trials



- 10 12.3% databases, 10 trials each: average correlation 0.23