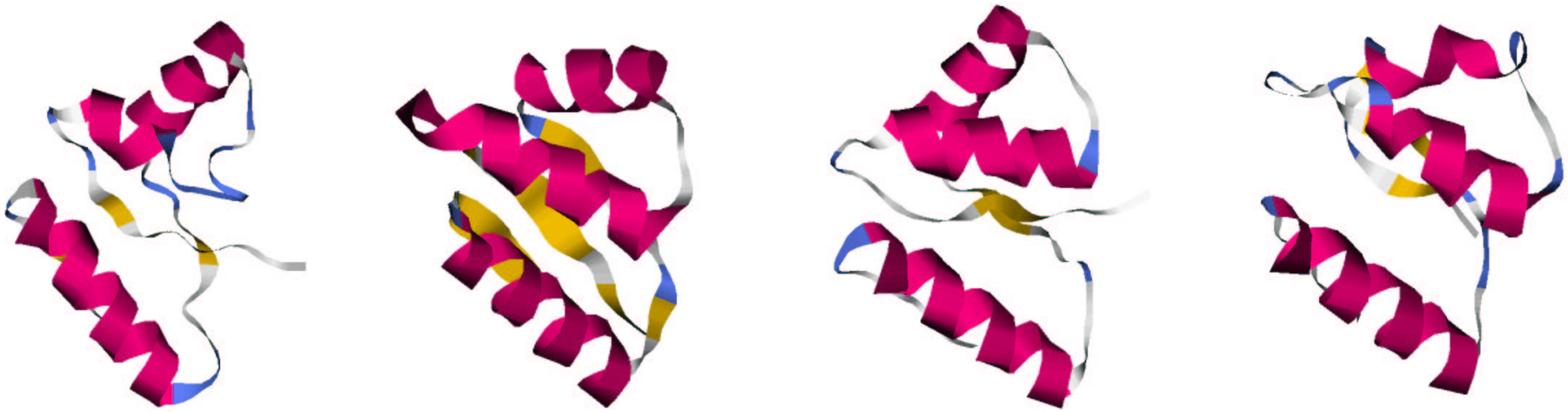
Statistical Potentials Based on Alpha-Shapes



Leo Guibas Patrice Koehl Afra Zomorodian
Stanford University

Focus

- Protein Structure Prediction
  - ⋆ Intellectual exercise
  - ⋆ Essential to structural genomics
- Classical scheme:
  - ⋆ generate models for the protein of interest
  - ⋆ select “best” and hopefully “native-like” model

Selection Problem

Target	Length	Best Decoy		Best Submitted	
		RMSD	Length	RMSD	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
T098	121	4.2	114	3.9	63
T102	70	3.2	70	3.56	70

[1]

[1] CASP 4 Experiment: Best results from D. Baker's Lab

State of the Art

- Good methods for generating decoys
- Not so good at selection
  - ⋆ very hard problem
  - ⋆ not well understood
- We need to predict cRMS(D)

State of the Art

- Good methods for generating decoys
- Not so good at selection
  - ⋆ very hard problem
  - ⋆ not well understood
- We need to predict cRMS(D)
- *This talk: Database-Derived Potentials*

Overview

- Potentials
- Alpha-Complex
- Random Databases
- Results
- Discussion

Physical Potentials

- Idea: native has minimum energy
- Problem: No ideal energy function
- VdW
- Electrostatics
- Hydrophobic effects (?)

Database-Derived Potentials

- Idea: native “looks” like a protein
- Problem: need to quantify “looks”
- Physics (Boltzmann law) or Information theory
- Pairwise potential (1979)
- Residue-based (Sippl 1990)
- Atom-based (Samudrala et al. 1997)

Method

- Potential
  - $\star$ $E \propto -\ln f(x)$
  - $\star$ X: Pair of certain type, e.g. CYS-CYS
  - $\star$ Y: Pair is at distance r
  - $\star$ $\Pr\{X \mid Y\} = \frac{\Pr\{X\} \cdot \Pr\{Y|X\}}{\Pr\{Y\}}$
  - $\star$ $\Pi = -\sum_{i \neq j} \ln(\Pr\{X_{ij} \mid Y_{ij}\})$

Method

- Potential
  - $\star$ $E \propto -\ln f(x)$
  - $\star$ X: Pair of certain type, e.g. CYS-CYS
  - $\star$ Y: Pair is at distance r
  - $\star$ $\Pr\{X \mid Y\} = \frac{\Pr\{X\} \cdot \Pr\{Y|X\}}{\Pr\{Y\}}$
  - $\star$ $\Pi = -\sum_{i \neq j} \ln(\Pr\{X_{ij} \mid Y_{ij}\})$
- Verification

1. Generate decoys for a known protein

Method

- Potential
  - $E \propto -\ln f(x)$
  - X: Pair of certain type, e.g. CYS-CYS
  - Y: Pair is at distance r
  - $\Pr\{X \mid Y\} = \frac{\Pr\{X\} \cdot \Pr\{Y|X\}}{\Pr\{Y\}}$
  - $\Pi = -\sum_{i \neq j} \ln(\Pr\{X_{ij} \mid Y_{ij}\})$
- Verification

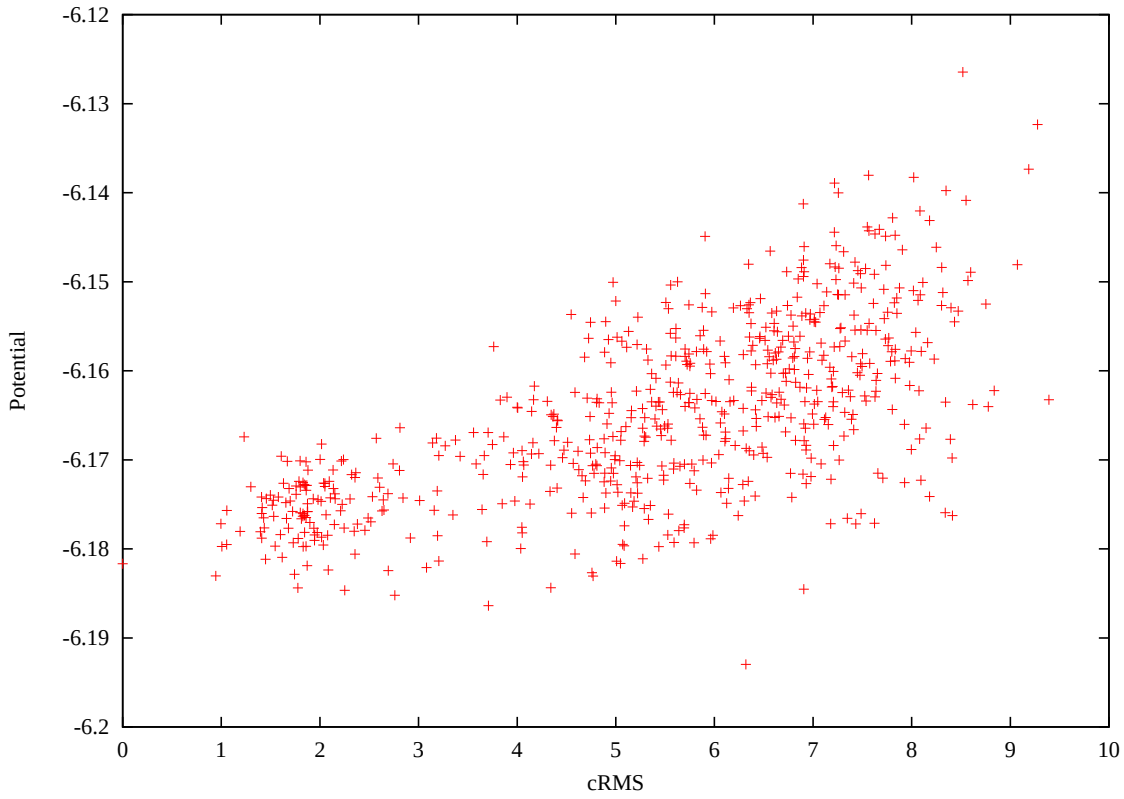
1. Generate decoys for a known protein
2. Compute cRMS to native and potential

Method

- Potential
  - $E \propto -\ln f(x)$
  - X: Pair of certain type, e.g. CYS-CYS
  - Y: Pair is at distance r
  - $\Pr\{X \mid Y\} = \frac{\Pr\{X\} \cdot \Pr\{Y|X\}}{\Pr\{Y\}}$
  - $\Pi = -\sum_{i \neq j} \ln(\Pr\{X_{ij} \mid Y_{ij}\})$
- Verification

1. Generate decoys for a known protein
2. Compute cRMS to native and potential
3. Compare cRMS with potential

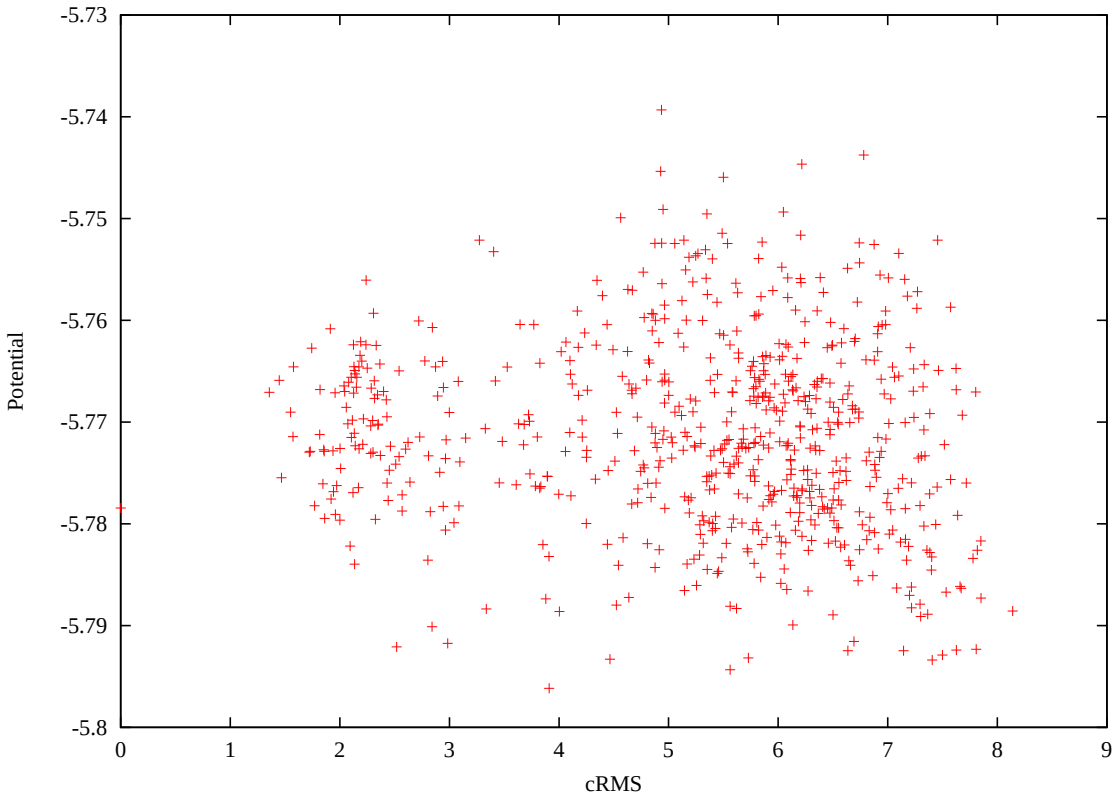
3icb



- 653 decoys, correlation 0.66[2]

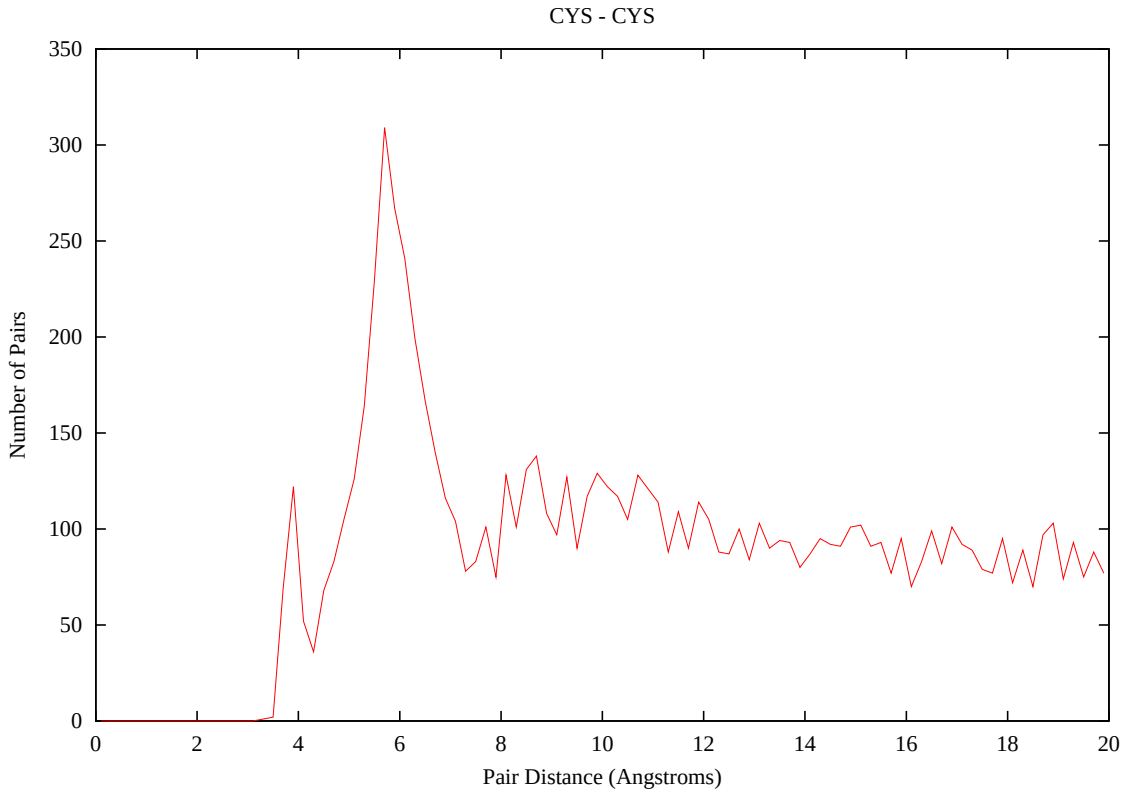
[2][Park & Levitt 1996]

4rxn

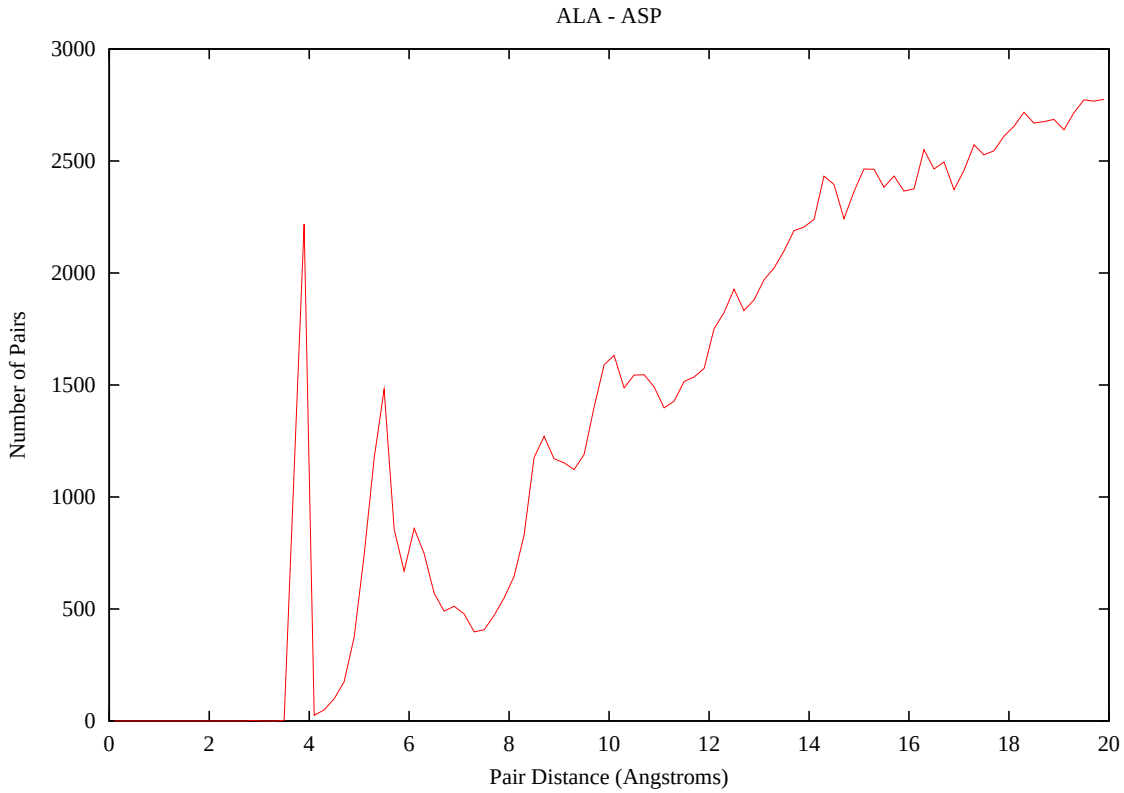


- 677 decoys, correlation: -0.08

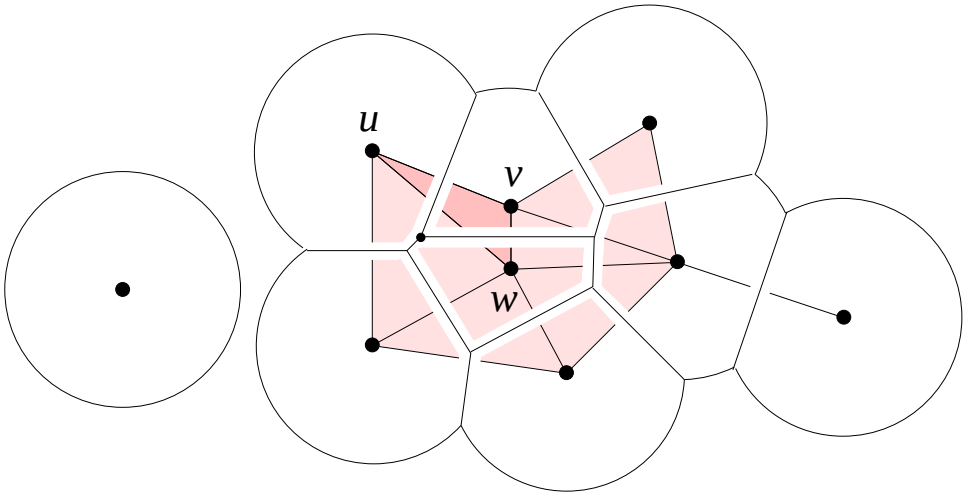
Signal and Noise



Signal and Noise

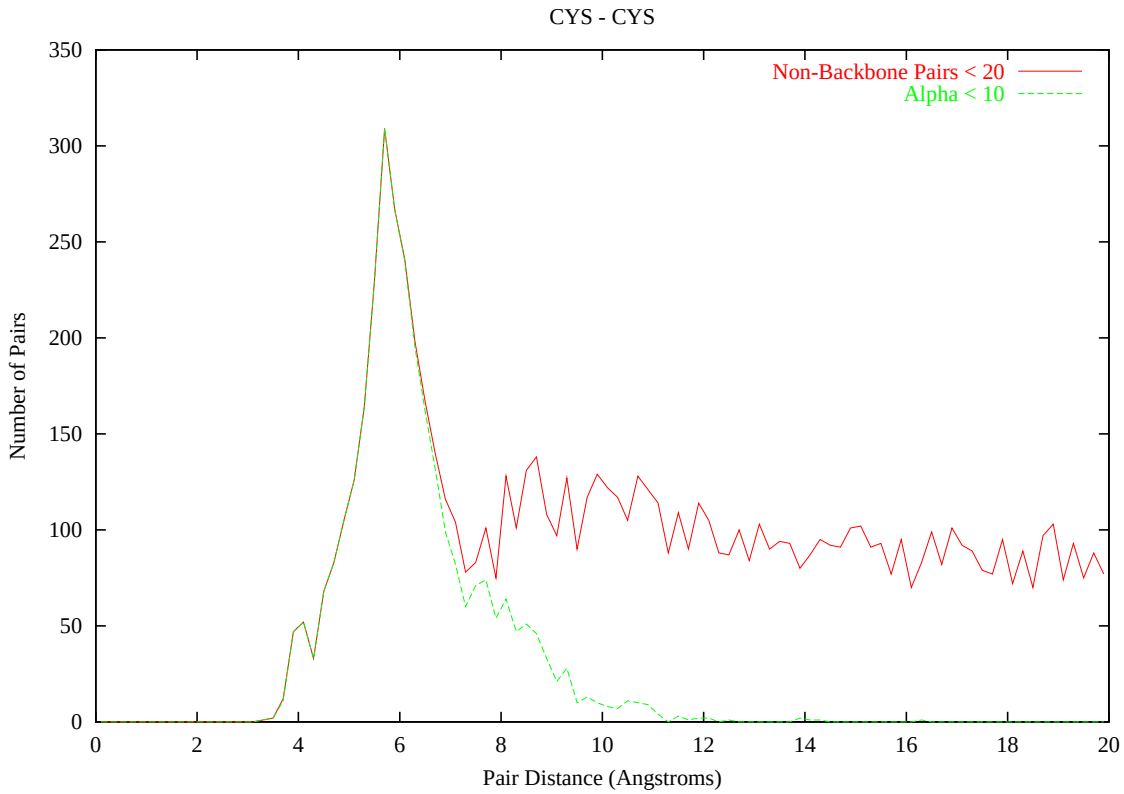


Alpha Complex

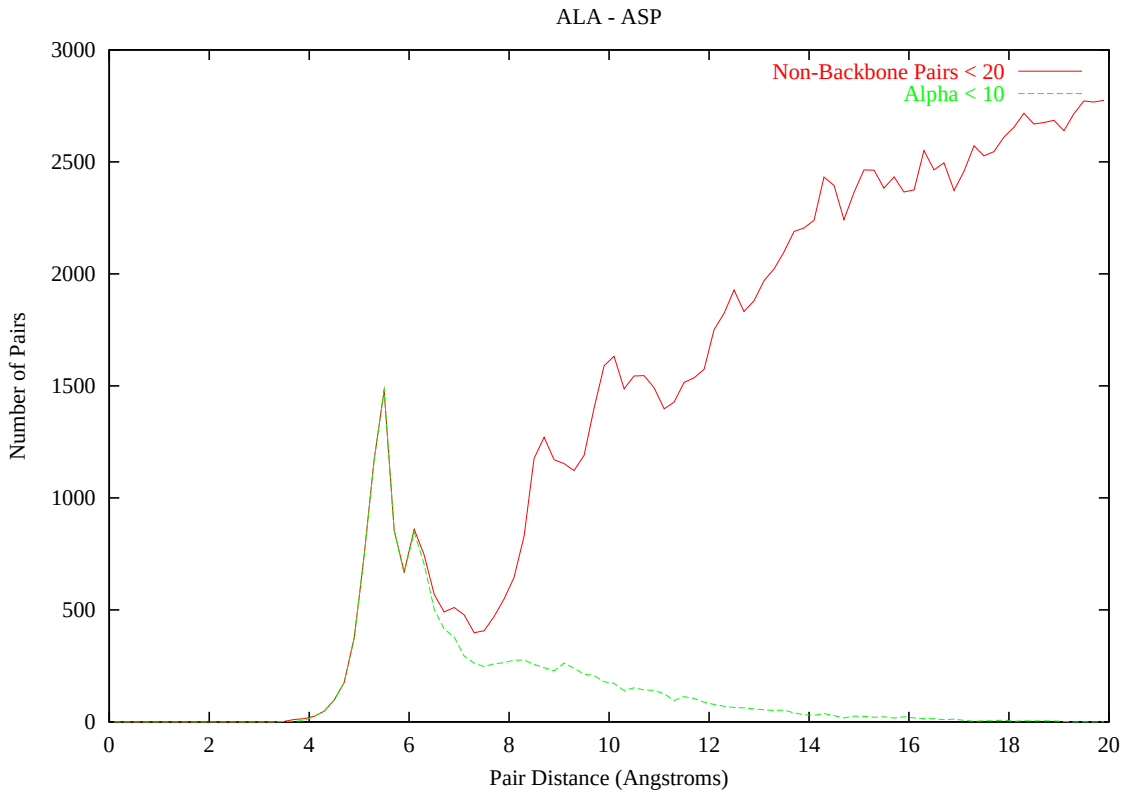


- van der Waals model
- Captures the topology of the set of ball set
- Subcomplex of the Delaunay complex

Pruning


CYS - CYS
Non-Backbone Pairs < 20
Alpha < 10
Number of Pairs
Pair Distance (Angstroms)
0
50
100
150
200
250
300
350
0
2
4
6
8
10
12
14
16
18
20

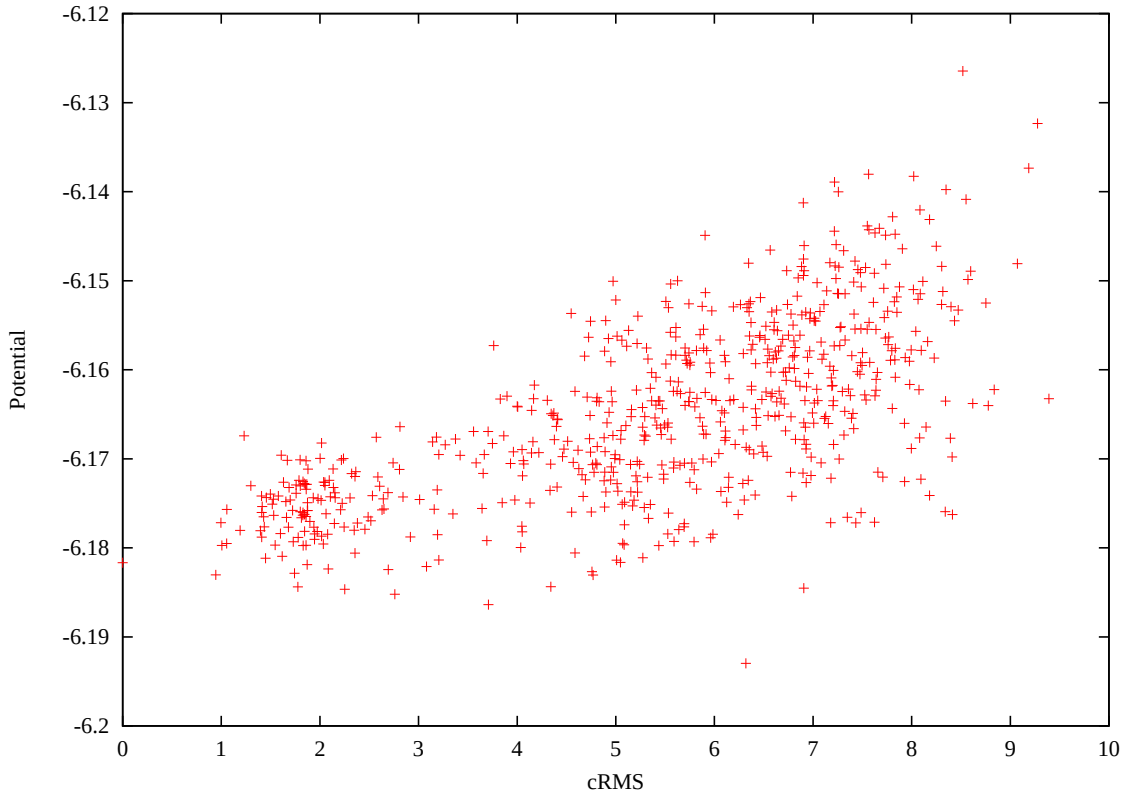
Pruning


ALA - ASP
Non-Backbone Pairs < 20
Alpha < 10
Number of Pairs
Pair Distance (Angstroms)
0
500
1000
1500
2000
2500
3000
0
2
4
6
8
10
12
14
16
18
20

Database

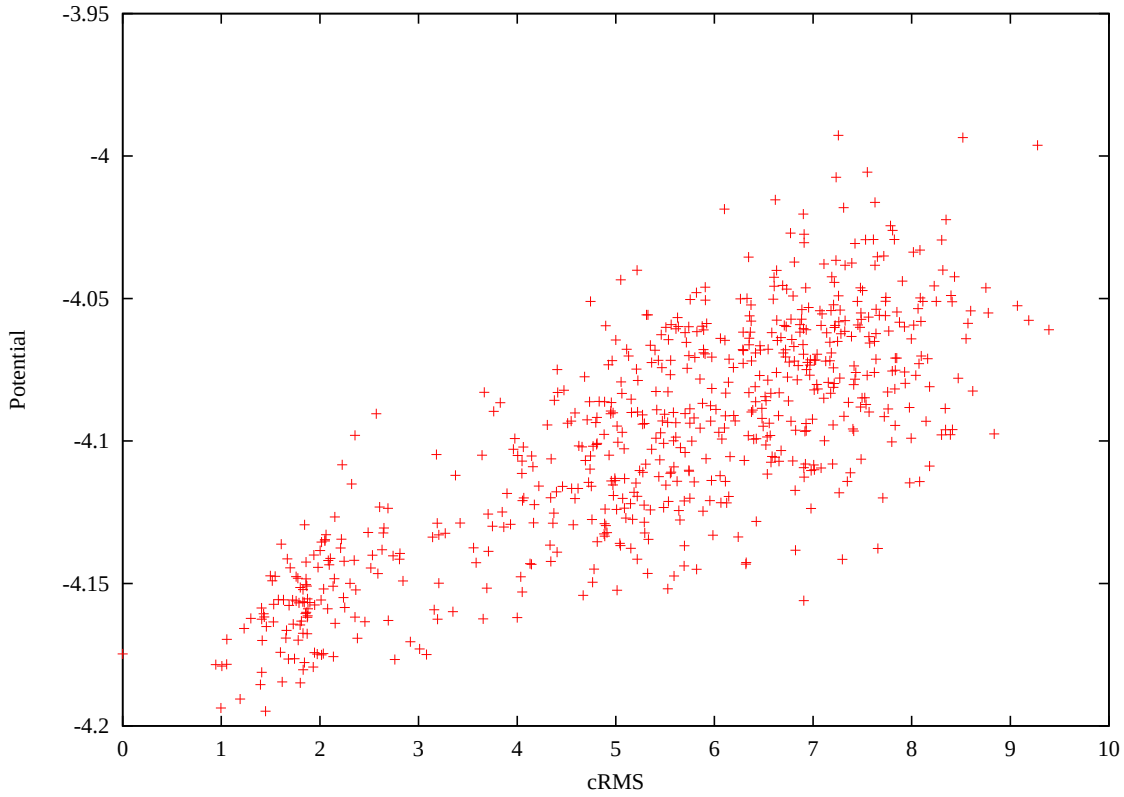
- 2,145 domains from SCOP
- 29,654,812 C^{α} pairs (*edges*) with distance $\leq$ 20Å
- $\alpha = 10$ gives 3,643,018 C^{α} edges
- $\approx$ 12.3% of possible non-backbone pairs
- 211 types

3icb – Full Database



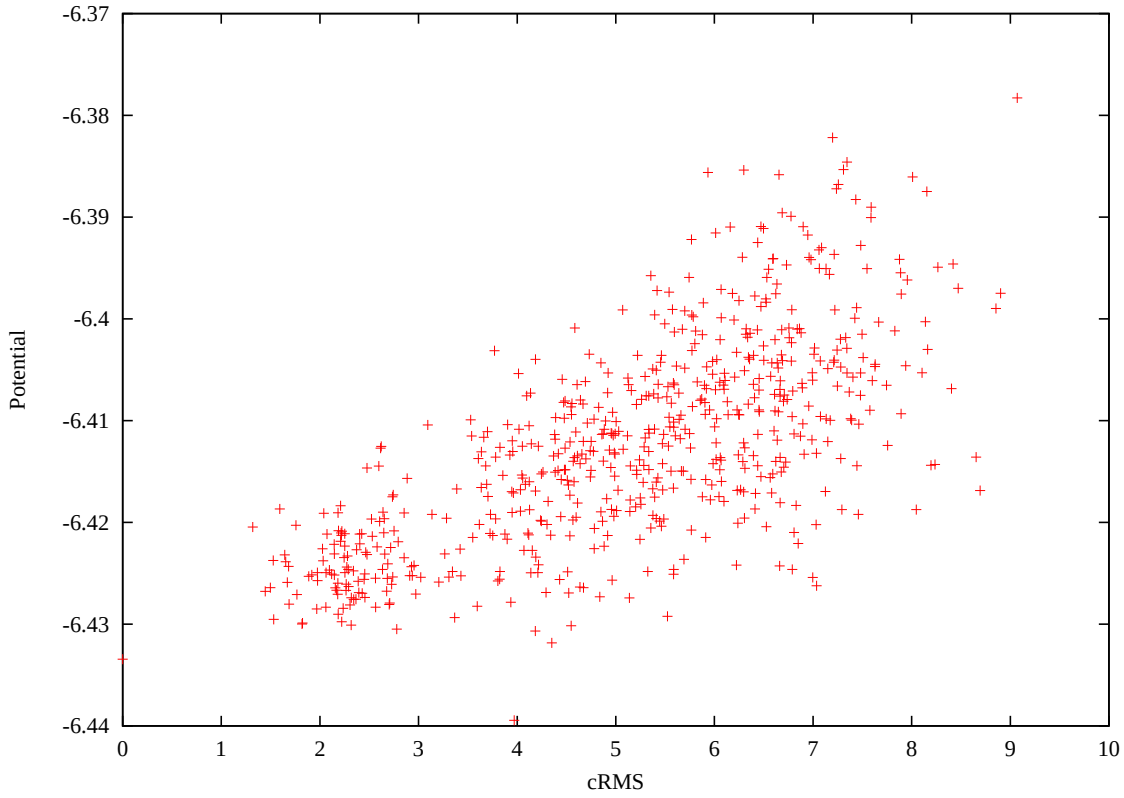
- 653 decoys, correlation 0.66

3icb – Alpha Database



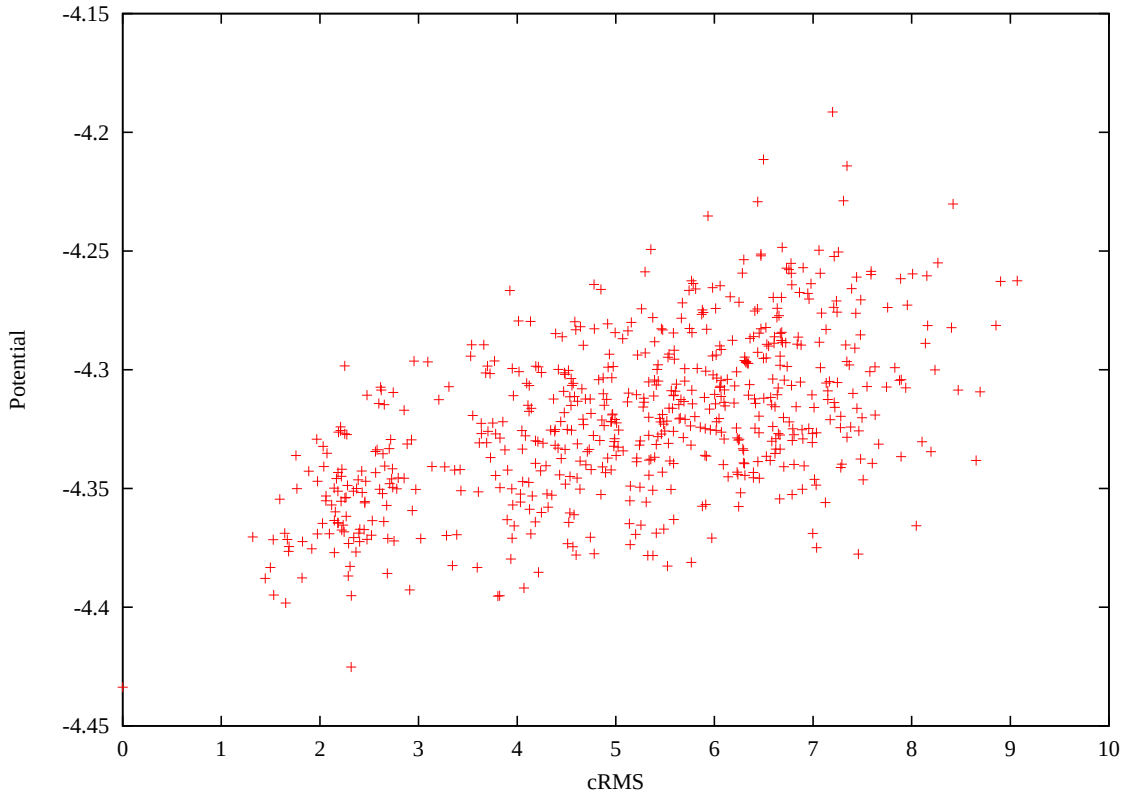
- 653 decoys, correlation 0.79

1ctf – Full Database



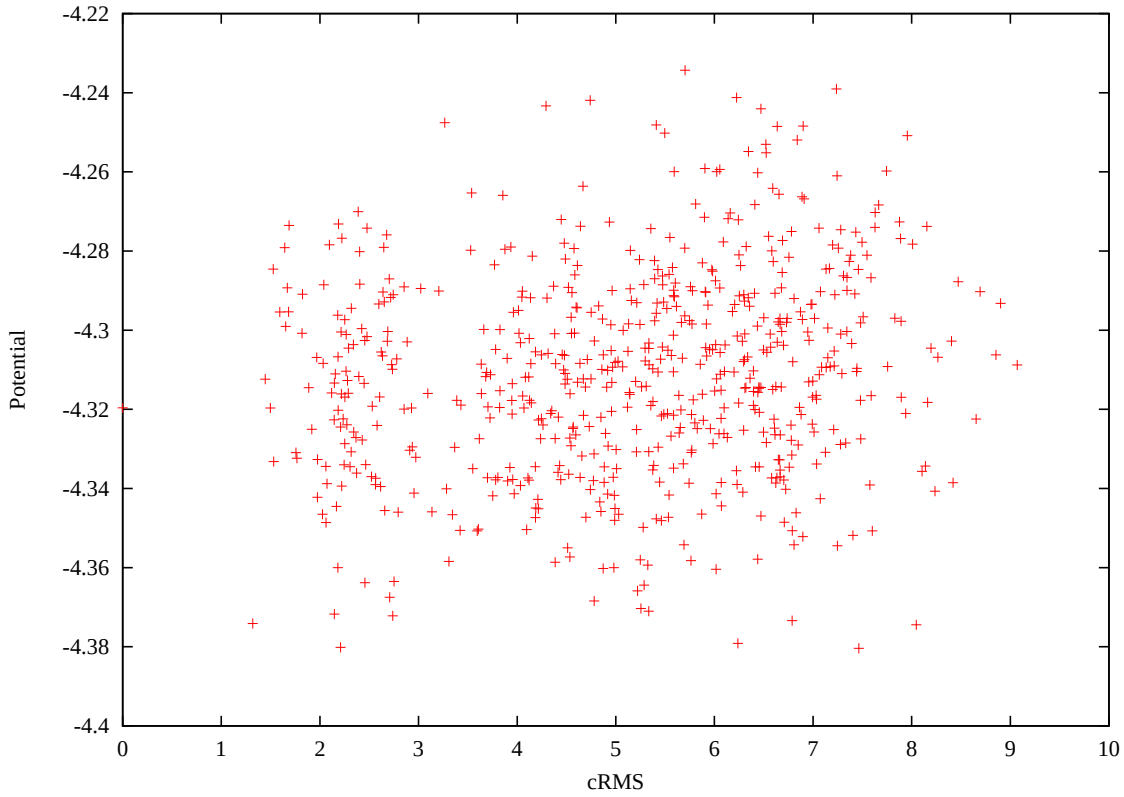
- 630 decoys, correlation 0.70

1ctf – Alpha Database



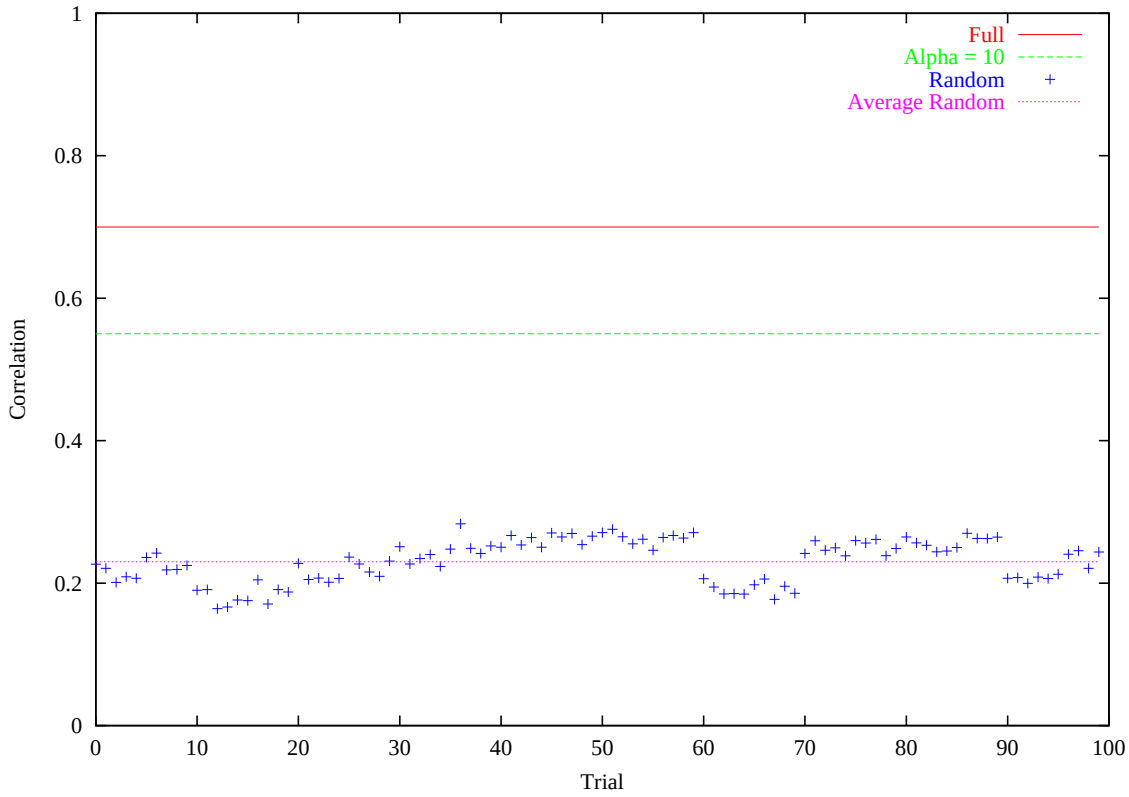
- 630 decoys, correlation 0.56

1ctf – Random 12.3% Database



- 630 decoys, correlation 0.16

Random Trials



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

Correlation

protein	type	# residues	# decoys	correlation	
				ap	α
1ctf	74	a+b	630	0.70	0.56
1r69	69	a	675	0.31	0.43
1sn3	65	a	660	-0.04	0.002
2cro	75	a	674	0.32	0.52
3icb	75	a	653	0.66	0.79
4pti	58	small	687	0.18	0.04
4rxn	54	small	677	-0.08	-0.22

Improved Selection

protein	best cRMS	selected (rank)		native	
		ap	α	ap	α
1ctf	1.32	3.97 (156)	2.32 (57)	2	1
1r69	0.88	5.35 (410)	4.47 (268)	368	217
1sn3	1.31	6.56 (428)	6.36 (383)	77	106
2cro	0.81	4.19 (246)	1.87 (24)	365	604
3icb	0.94	6.32 (387)	1.45 (16)	16	21
4pti	1.41	6.28 (424)	4.75 (169)	2	2
4rxn	1.36	3.91 (140)	7.01 (606)	148	332

Experiments

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

Discussion

- No significant progress in 20 years: Geometry?
- Future work:
  - ⋆ Restrict to domains
  - ⋆ Alternate functions and Multivariate regression
  - ⋆ No distance dependence
- Why cRMS?
- Induced Questions: rigidity, local computation