Protein-DNA Specificity Prediction

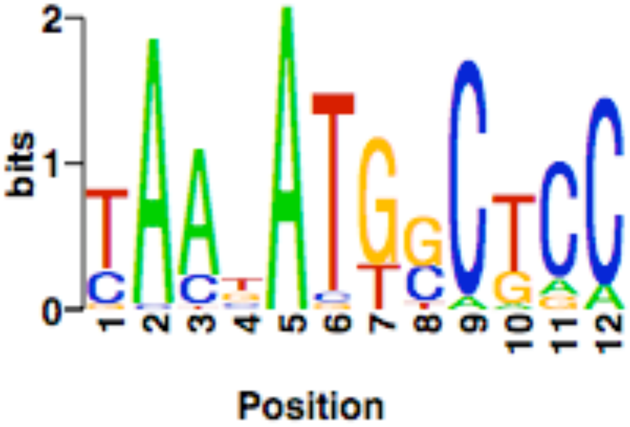
Chen Yanover
Bradley lab
Program in Computational Biology

FRED HUTCHINSON
CANCER RESEARCH CENTER
A LIFE OF SCIENCE

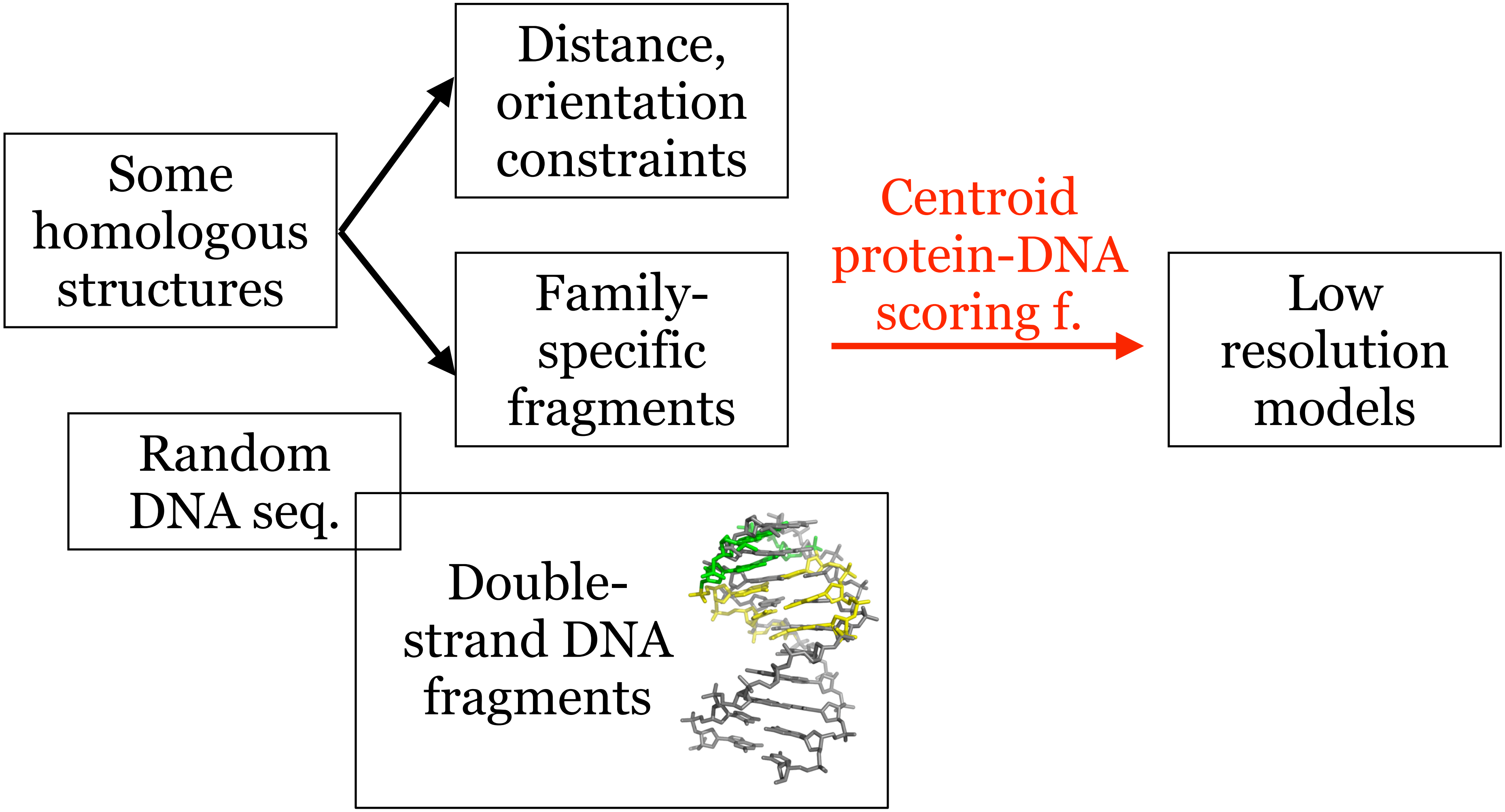
ROSETTACON 2009

Goal: sequence→structure→specificity

```
>1ubd_C mol:protein length:124  PROTEIN (YY1 ZINC FINGER DOMAIN)
MEPRTIACPHKGCTKMFRDNSAMRKHLHTHGPRVHVCAECGKAFVESSKLKRHQLVHTGE
KPFQCTFEGCGKRFSLDFNLRTHVRIHTGDRPYVCPFDGCNKKFAQSTNLKSHILTHAKA
KNNQ
```



Low resolution phase



Some homologous structures

Distance, orientation constraints

Family-specific fragments

Centroid protein-DNA scoring f.

Low resolution models

Random DNA seq.

Double-strand DNA fragments

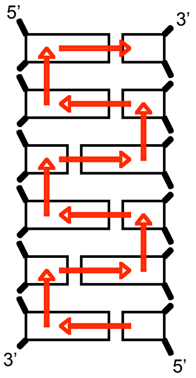
High resolution phase

Low resolution models

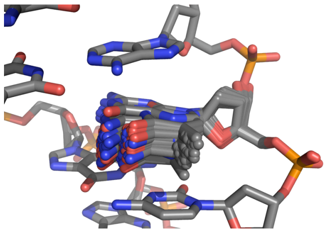
All atom refinement, w/ simultaneous sampling of DNA sequence

High resolution models

DNA base-centric kinematics



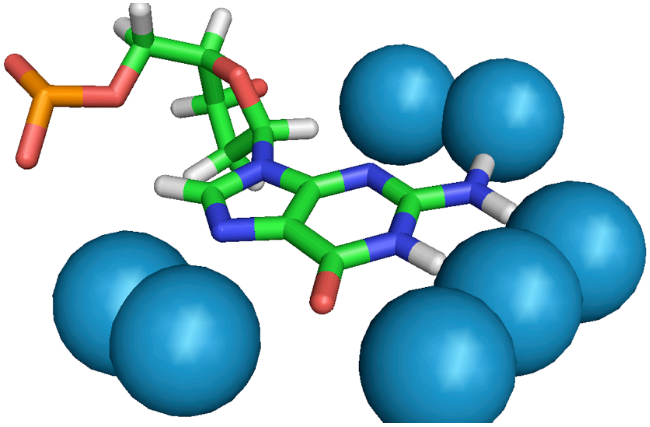
DNA “rotamers”



Full-atom score function

- Explicit electrostatic (hack_elec)

- LK orientation-dependent



- Base-step ref energy

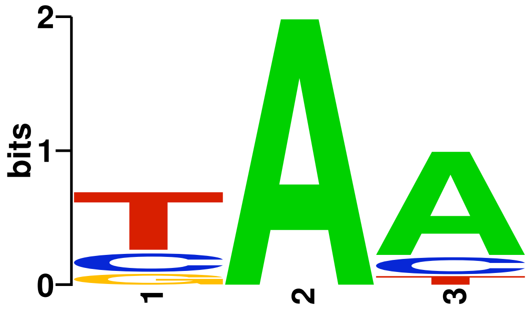
Decoys to specificity

energy	DNA
-212.866	aca
-211.624	gaa
-211.259	taa
-209.915	tac
-209.865	taa
-209.282	taa
-209.112	gaa
-208.631	ccc
-208.593	gac
-208.281	gta
-208.204	caa
-208.058	caa

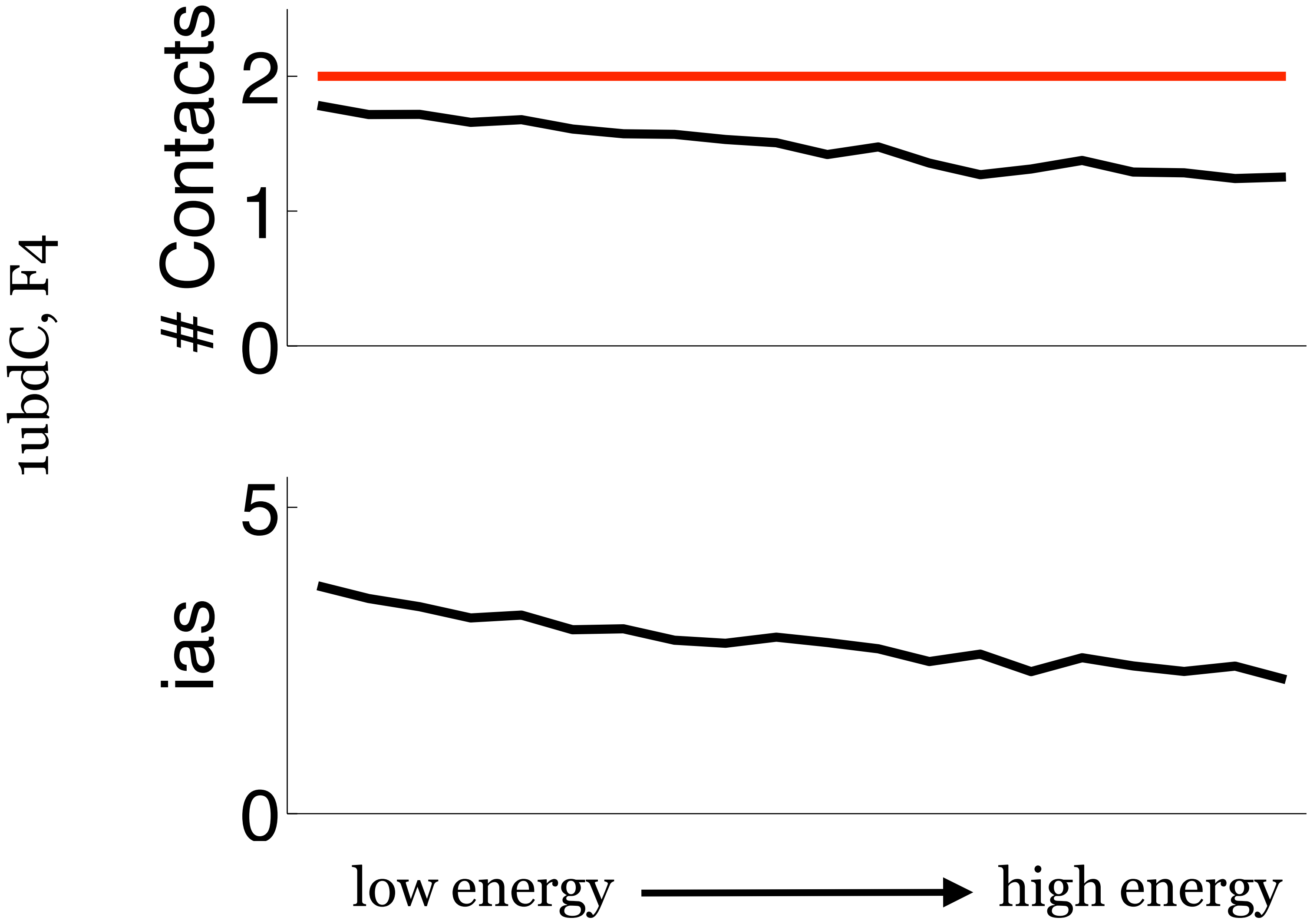
[1ubdC, F4]

DNA	#
taa	259
caa	139
tac	90
gaa	89
tat	51

A	0,	628,	487
C	139,	0,	90
G	89,	0,	0
T	400,	0,	51



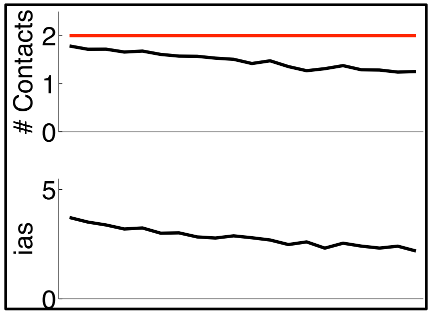
Results: structure prediction

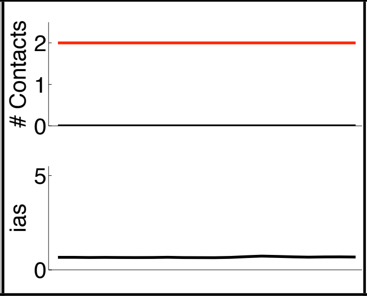
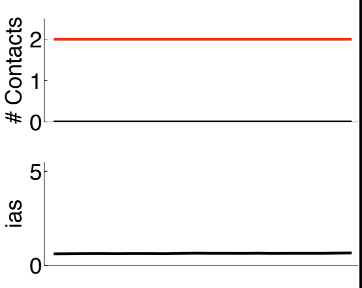
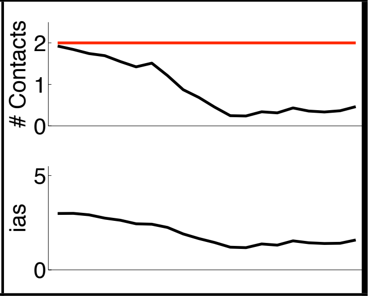
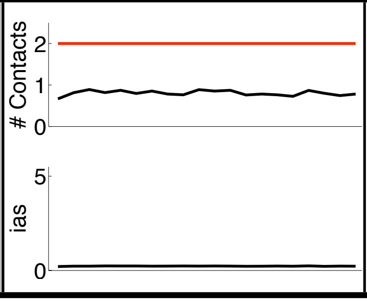
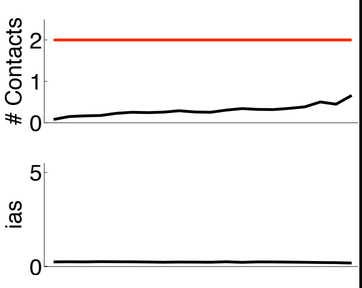
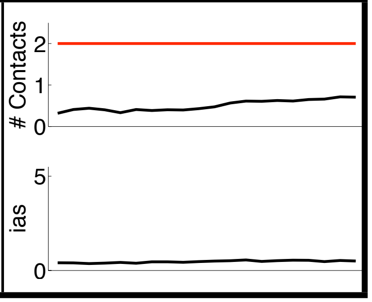


Results: structure prediction

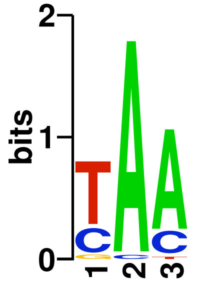
1ubdC, F4

Full protocol

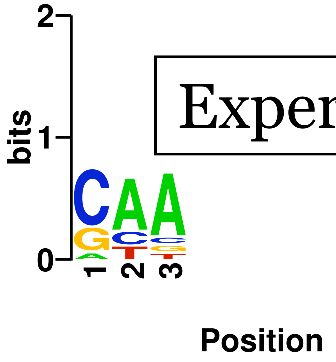


	repack	sc-mcm	high res.
1aayA, F1			
2drpA, F1			

Results: single ZF specificity prediction

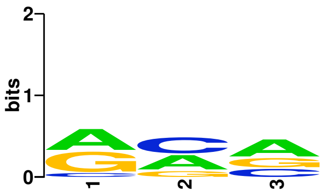
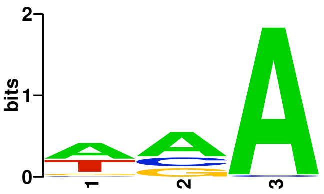
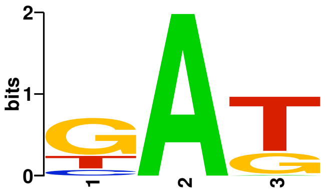


Predicted

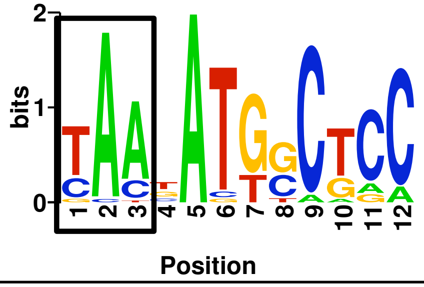
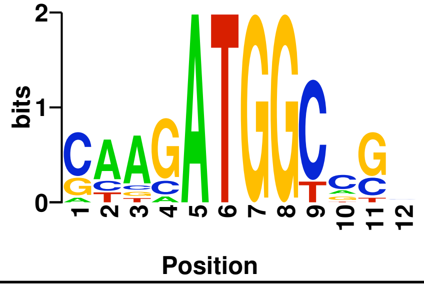
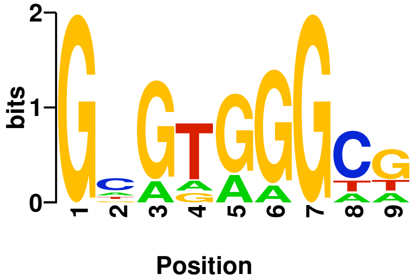
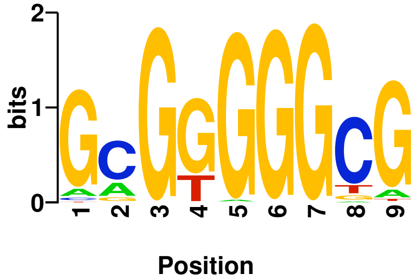
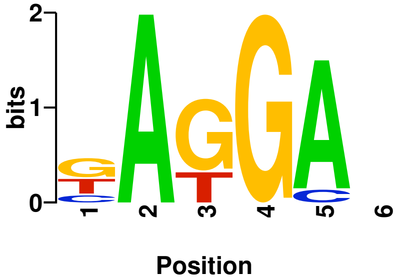
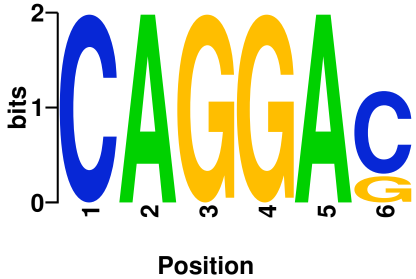


Experimental

1ubdC, F4

	repack	sc-mcm	high res.
1aayA, F1			
2drpA, F1			

Results: multi-ZF specificity prediction

	Predicted	Experimental
YY1 [1ubdC]		
Egr1 [1aayA]		
TTK [2drpA]		

Summary

- Structure-based prediction of DNA binding specificity, using family info
- Works well for C2H2 ZFs
  - more families to follow

Acknowledgements

- Phil Bradley, Bradley lab
- FHCRC

FRED HUTCHINSON
CANCER RESEARCH CENTER

A LIFE OF SCIENCE