Using Symmetry to Predict Protein-DNA Interactions

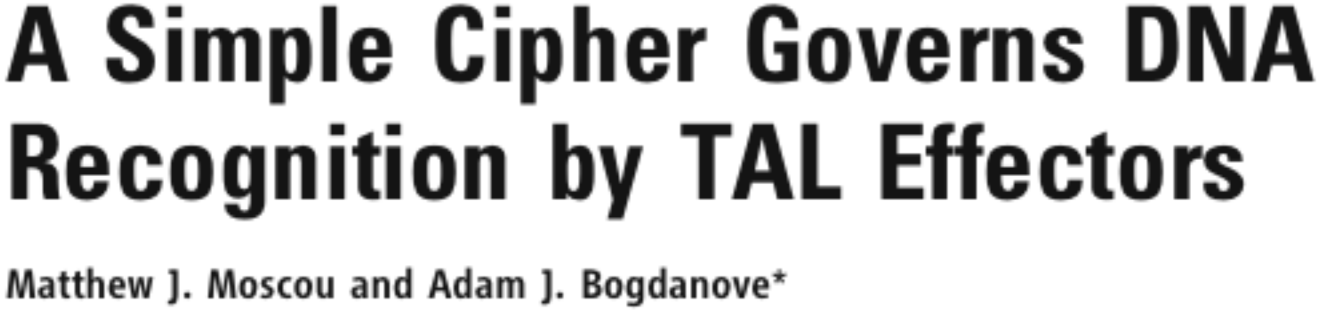
Phil Bradley

Computational Biology Program

FRED HUTCHINSON
CANCER RESEARCH CENTER

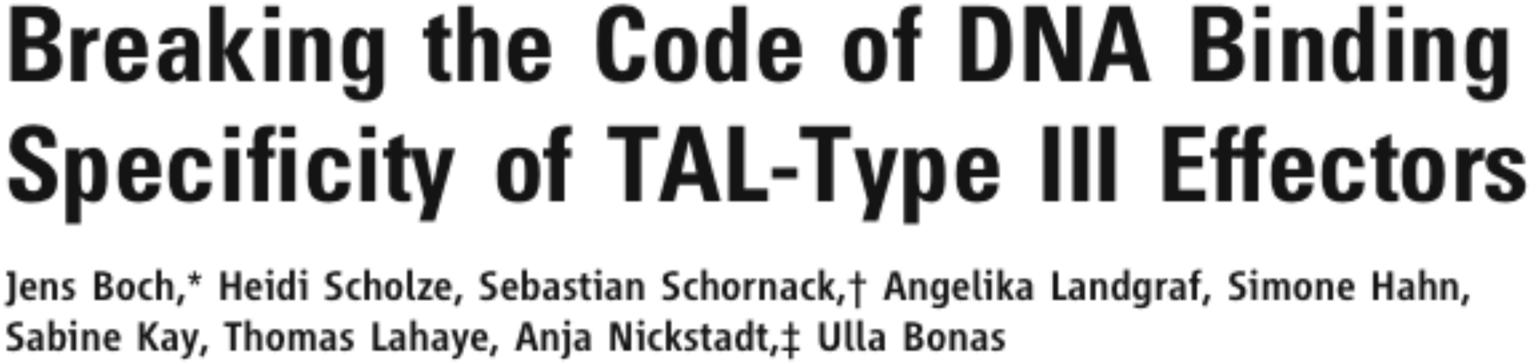
A LIFE OF SCIENCE

TAL Effectors: A new and versatile DNA recognition mode



A Simple Cipher Governs DNA Recognition by TAL Effectors

Matthew J. Moscou and Adam J. Bogdanove*

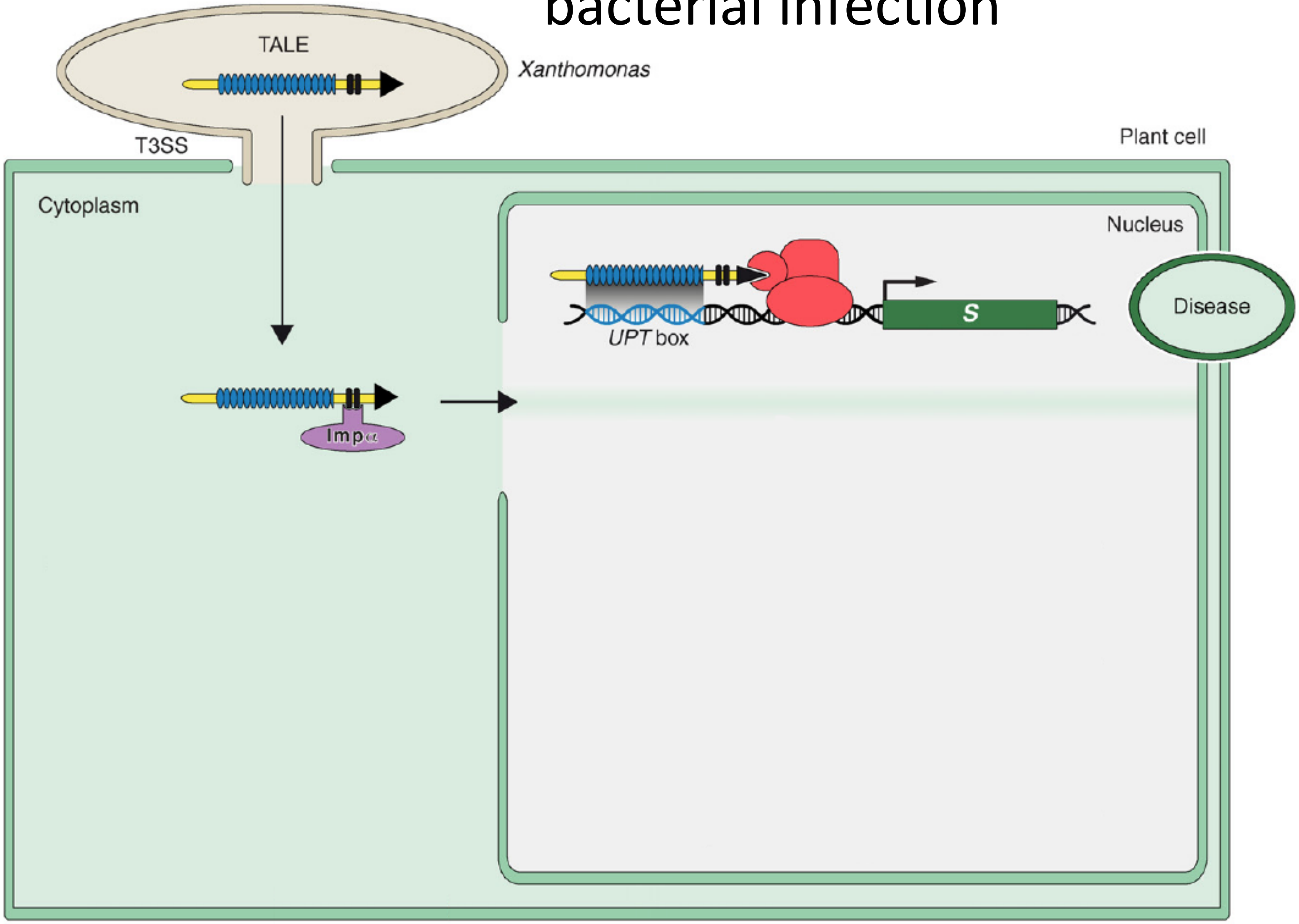


Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors

Jens Boch,* Heidi Scholze, Sebastian Schornack,† Angelika Landgraf, Simone Hahn, Sabine Kay, Thomas Lahaye, Anja Nickstadt,‡ Ulla Bonas

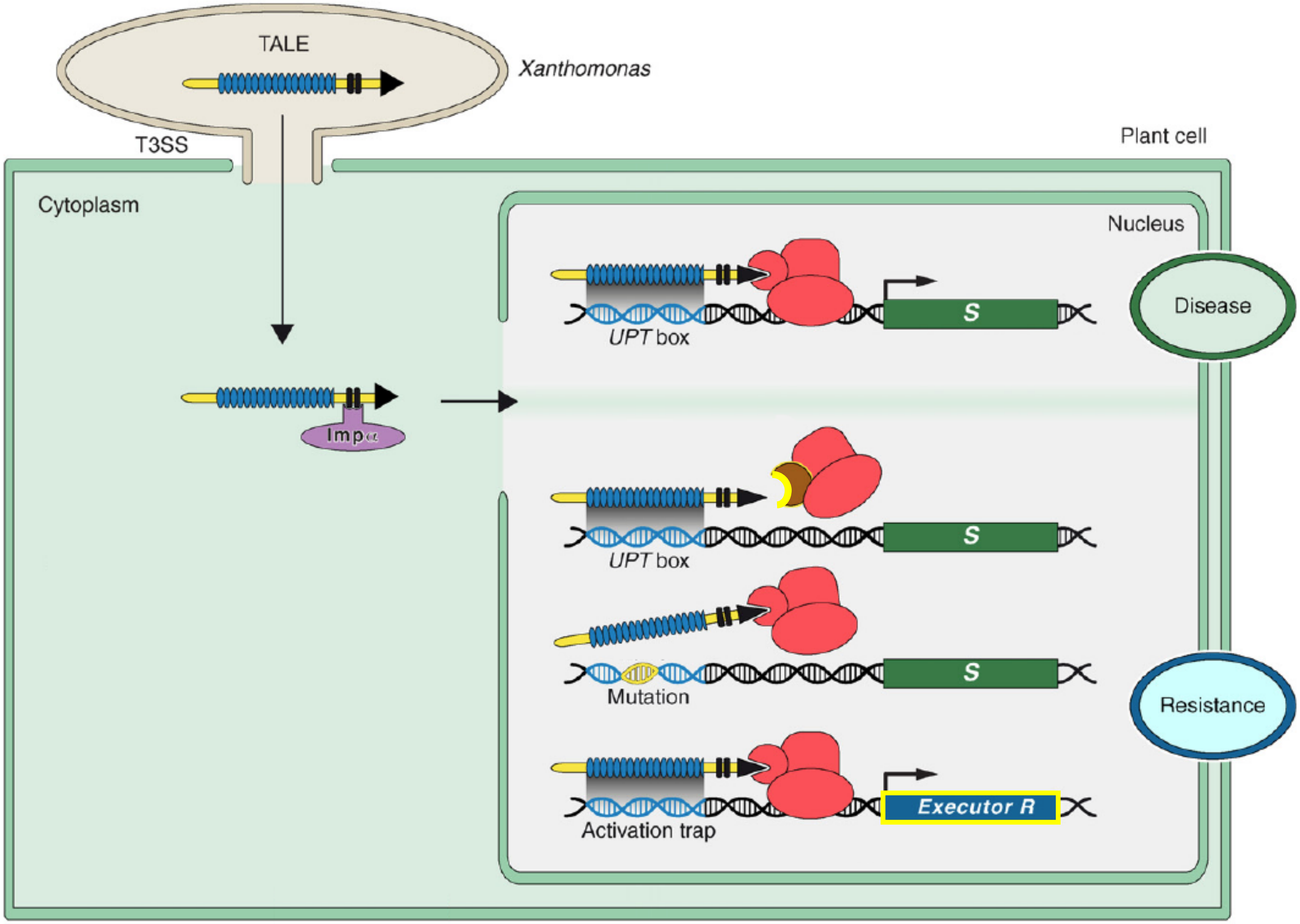
December 2009, Science

TAL Effectors (TALEs) are trans-kingdom transcription factors that activate plant genes which promote susceptibility to bacterial infection



Bogdanove *et al* Curr. Opinion Plant Biol. (2010) 13 (4): 394 - 401

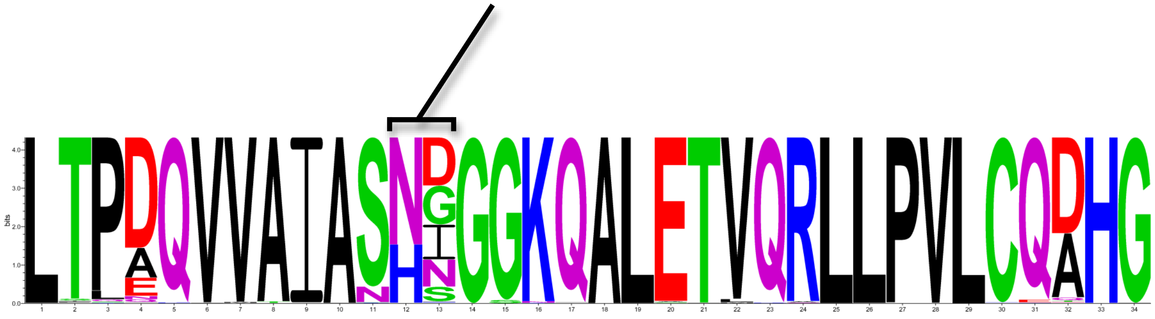
Plants can evolve evasion mechanisms that lead to resistance

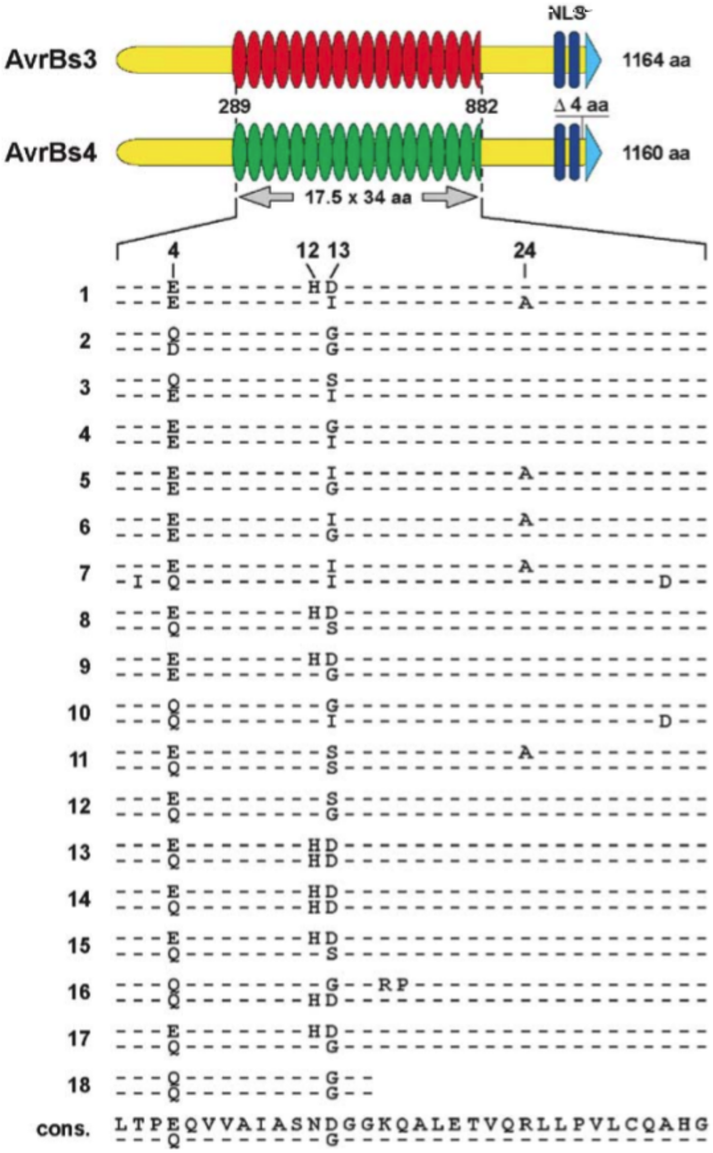


Bogdanove *et al* Curr. Opinion Plant Biol. (2010) 13 (4): 394 - 401

TAL effectors have a central repeat region that forms the DNA binding domain

The tandem, 34 amino acid repeats show very high sequence similarity, with most changes restricted to positions 12 and 13, termed the “Repeat Variable Diresidue” (**RVD**)





A simple cipher governs RVD-DNA associations, allows prediction and re-engineering of target sites.

AvrXa27 - *Xa27*
```
NI NN N* NG NS NN NN NN NI NN NI N* HD HD NI NG NG
 A  G  A  A  G  A  A  G  A  G  A  C  C  C  A  T  A
```

AvrBs3 - *Bs3*
```
HD NG NS NG NI NI NI HD HD NG NS NS HD HD HD NG HD NG
 A  T  A  T  A  A  A  C  C  T  A  A  C  C  A  T  C  C
```

AvrBs3 - *UPA20*
```
HD NG NS NG NI NI NI HD HD NG NS NS HD HD HD NG HD NG
 A  T  A  T  A  A  A  C  C  T  G  A  C  C  C  T  T  T
```

AvrBs3Δrep16 - *Bs3-E*
```
HD NG NS NG NI NI NI HD HD NG HD NG HD NG
 A  T  A  T  A  A  A  C  C  T  C  T  C  T
```

AvrBs3Δrep109 - *Bs3*
```
HD NG NS NG NI NI NI HD HD NG NS NS NG HD NG
 A  T  A  T  A  A  A  C  C  T  A  A  C  C  A
```

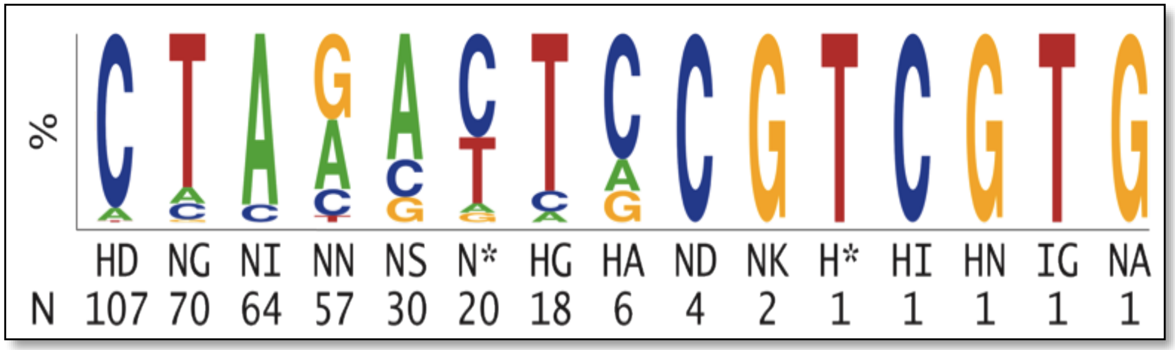
AvrHah1 - *Bs3*
```
NN IG NI NI NI HD HD NG NN NI HD HD HD NG
 A  T  A  A  A  C  C  T  A  A  C  C  A  T
```

PthXo1 - *Os8N3*
```
NN HD NI HG HD NG N* HD HD NI NG NG NI HD NG NN NG NI NI NI NI N* NS N*
 G  C  A  T  C  T  C  C  C  C  C  T  A  C  T  G  T  A  C  A  C  C  A  C
```

PthXo6 - *OsTFX1*
```
NI H* NI NN NN NN NN NN HD NI HD HG HD NI N* NS NI NI HG HD NS NS NG
 A  T  A  A  A  A  G  G  C  C  C  T  C  A  C  C  A  A  C  C  C  A  T
```

PthXo7 - *OsTFIIAγ*
```
NI NG NI NI N* NN HD HD N* NI NI NI NG HD HG NN NS NN HD HD NG NG
 A  T  A  A  T  C  C  C  C  A  A  A  T  C  C  C  C  T  C  C  T  C
```

Tal1c - *OsHEN1*
```
HD HD HD HD HD NG HD NN HD NG HG NN HD N* NG NG
 C  C  C  C  C  T  C  G  C  T  T  C  C  C  T  T
```



	HD	NG	NI	NN	NS	N*	HG	HA	ND	NK	H*	HI	HN	IG	NA
N	107	70	64	57	30	20	18	6	4	2	1	1	1	1	1

Moscou and Bogdanove Science (2009) 326: 1501.
Boch et al. Science (2009) 326: 1509 – 1512.

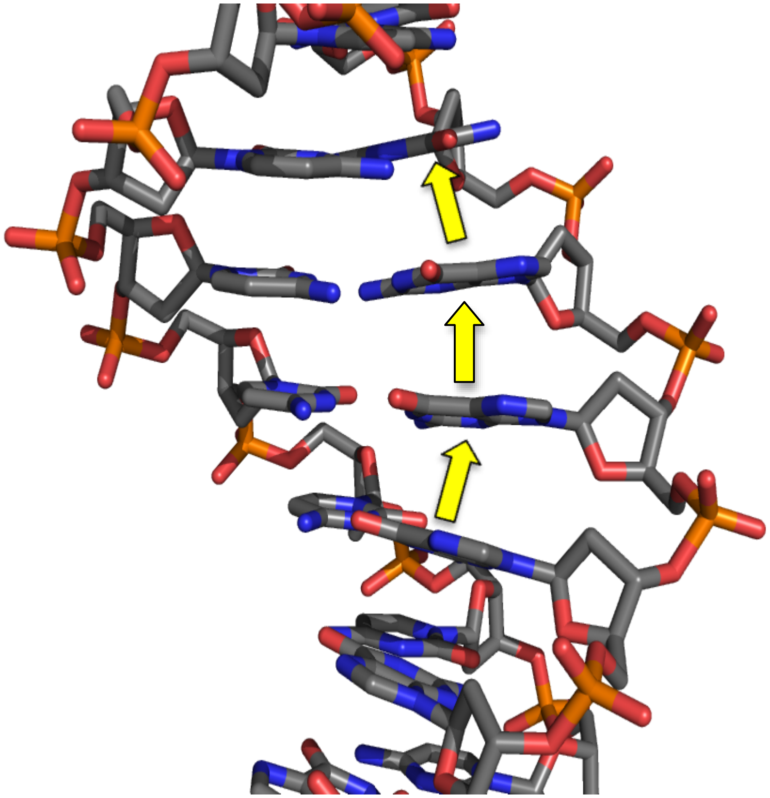
Questions for model building

- How do you fit 34 amino acid repeats next to the DNA in 1-1 mapping with basepairs? (Is it even possible, sterically?)
- Is the DNA B-form?
- Is the TAL repeat structure similar to TPR repeats (another 34aa repeat family)?
- How do the RVDs recognize specific base pairs?

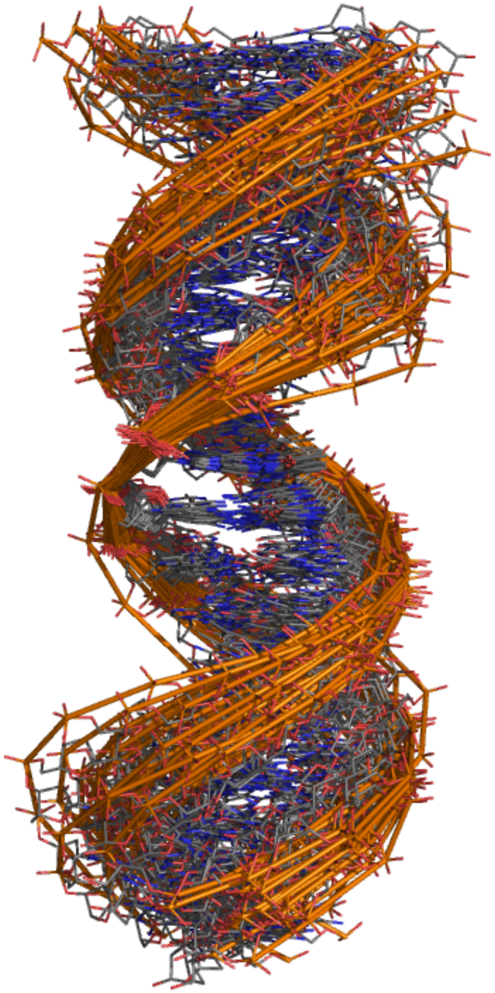
Model-building assumptions: Symmetry and RVD-DNA contact

- DNA is structurally symmetric across the target site
- One or both of the RVD positions contacts DNA
- Repeats of the same RVD:base (e.g. NI:A) association have the same structure
- Repeats with different RVDs have similar structures

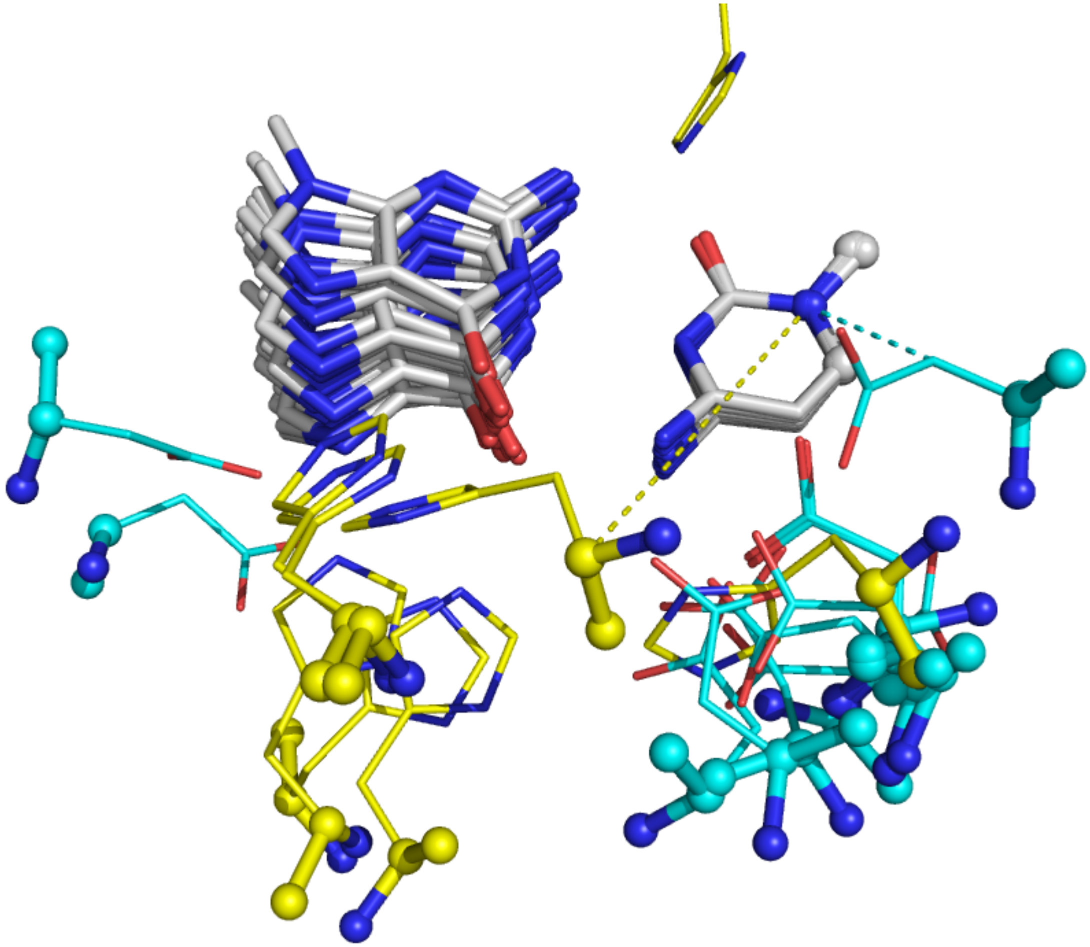
A library of symmetrical DNA structures



The same base-step transform (yellow arrow) is repeated multiple times to generate a symmetrical double-helix

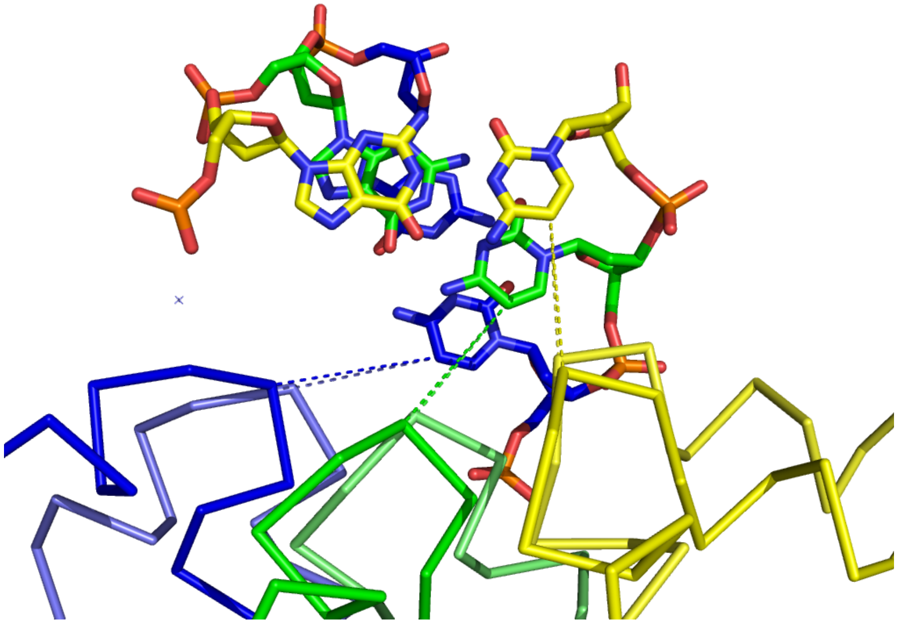


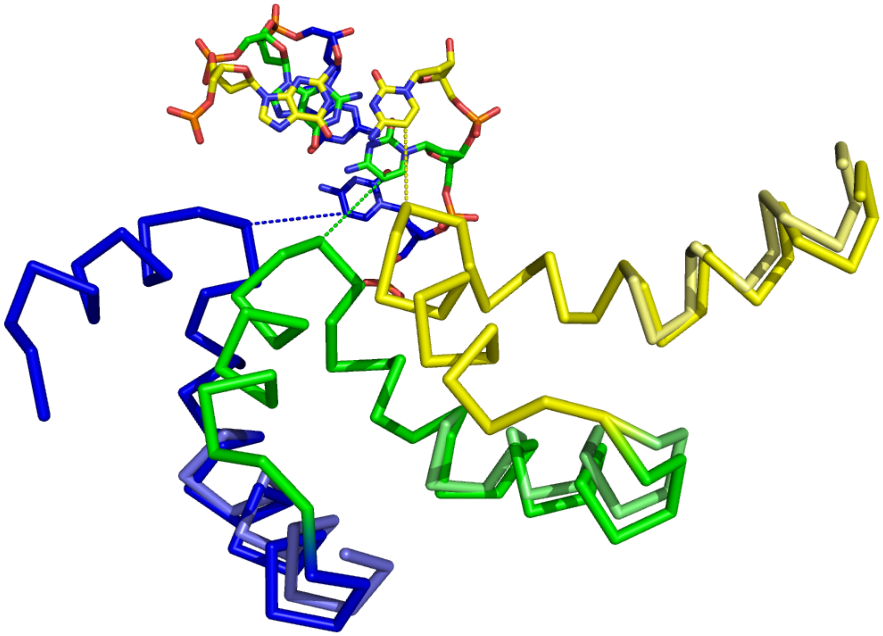
A library of RVD-base contacts observed in protein-DNA structures



Examples of Histidine and Aspartate residues contacting G:C base pairs

Symmetrical fragment-replacement moves for protein-DNA interfaces





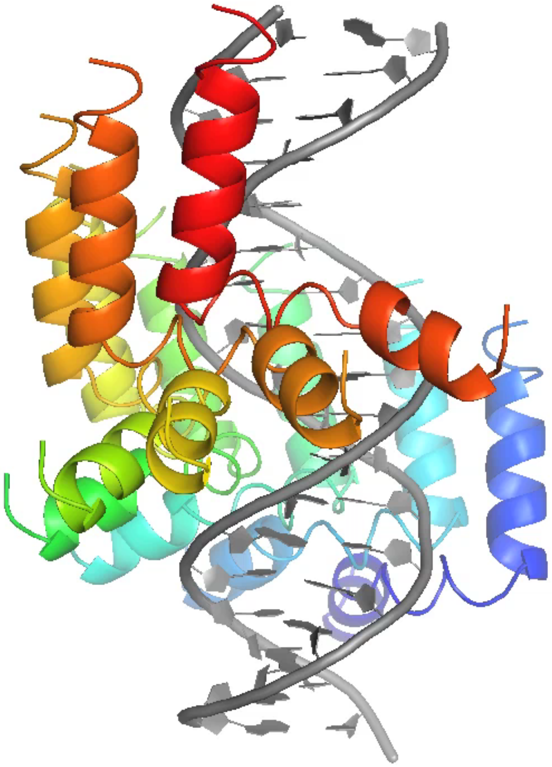
Update the RVD:DNA contact geometry for all repeats

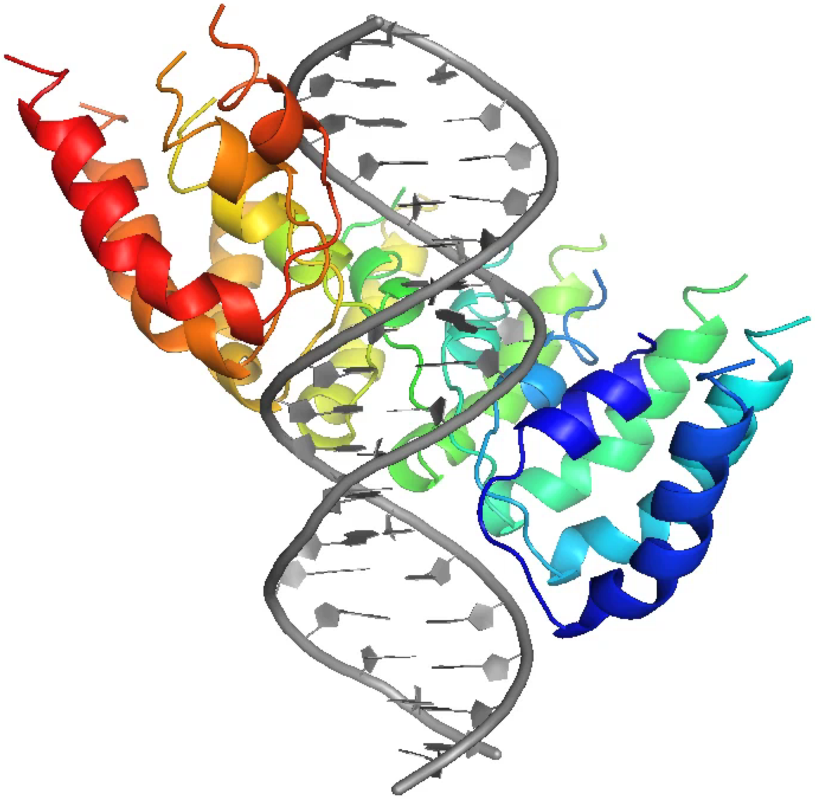
Update the protein backbone conformation of all repeats

Several thousand independent folding/docking simulations generate a population of TAL-DNA models

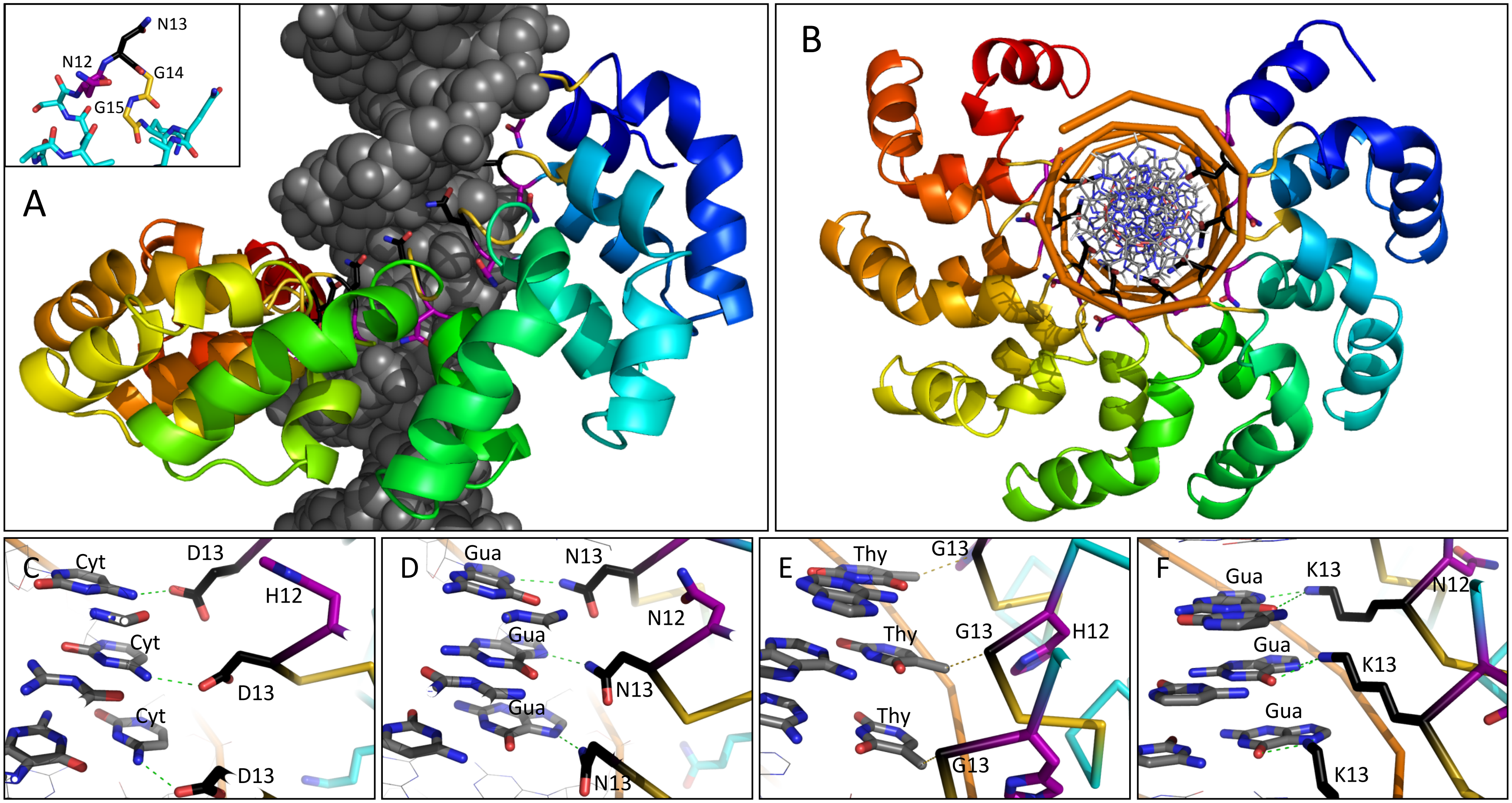
Each simulation models a single RVD-DNA association repeated multiple times with perfect symmetry

RVD-DNA contact is guaranteed by construction: the 3D structure of each repeat unit is built outward from the RVD loop, which is anchored to its cognate base-pair by a flexible linker



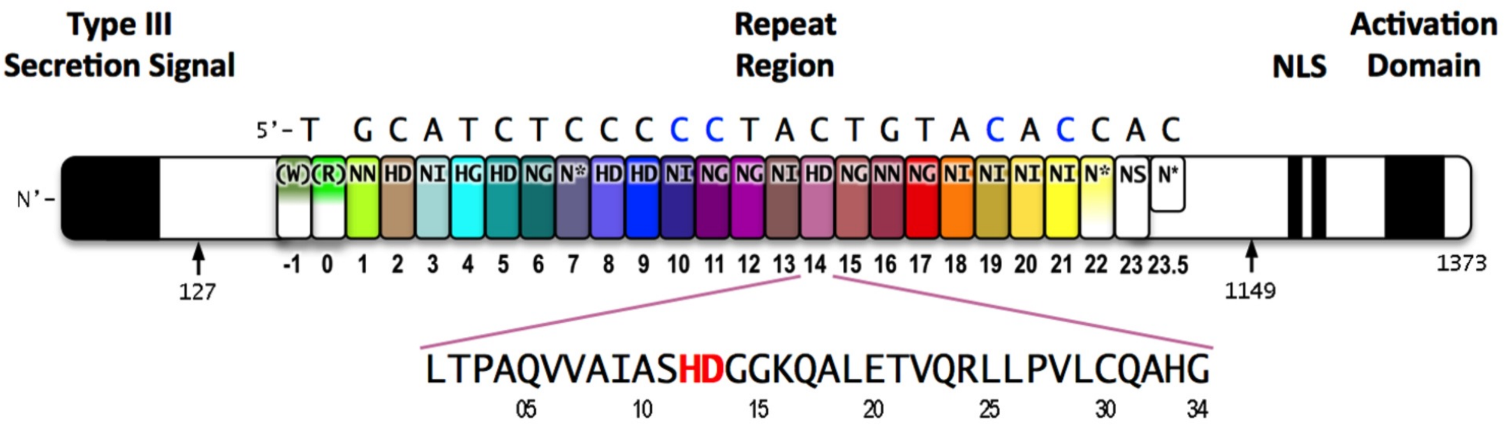


Final models were clustered to yield a predicted structure with good geometry and favorable protein-DNA interaction energies. Structural model provides explanation for observed RVD-DNA associations.

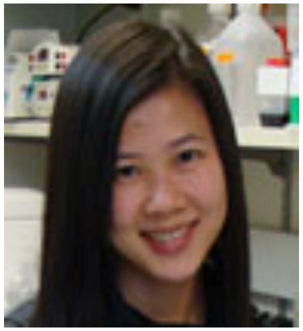


HD:C NN:G HG:T NK:G

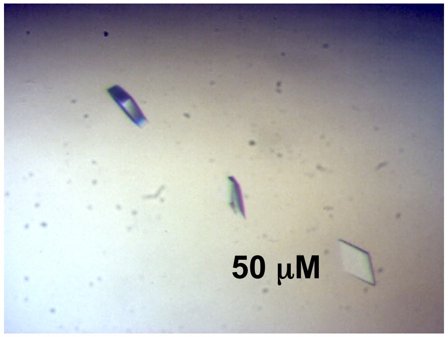
Experimental validation:
Crystallization of PthXo1 from *Xanthomonas Oryzae*


Type III
Secretion Signal
Repeat
Region
NLS
Activation
Domain
5'-T G C A T C T C C C C C T A C T G T A C A C C A C
N'-
(W) (R) NN HD NI HG HD NG N* HD HD NI NG NG NI HD NG NN NG NI NI NI NI N* NS N*
127
-1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 23.5
1149
1373
LTPAQVVAIASHDGGKQALETVQRLLPVLCQAHG
05 10 15 20 25 30 34

TAGATATGCATCT CCCCC TACTGTACACCACCAAAAGT
CAATCTATACGTAGAGGGGGATGACATGTGGTGGTTTT



Amanda Mak
Post-doc
Stoddard Lab



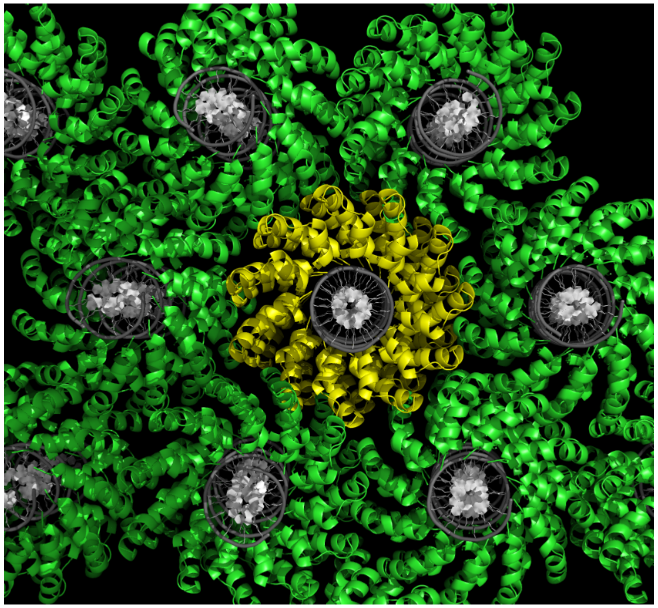
$P2_12_12_1$
a = 95.6 Å b = 248.5Å c = 54.6 Å
d_{min} = 3 Å

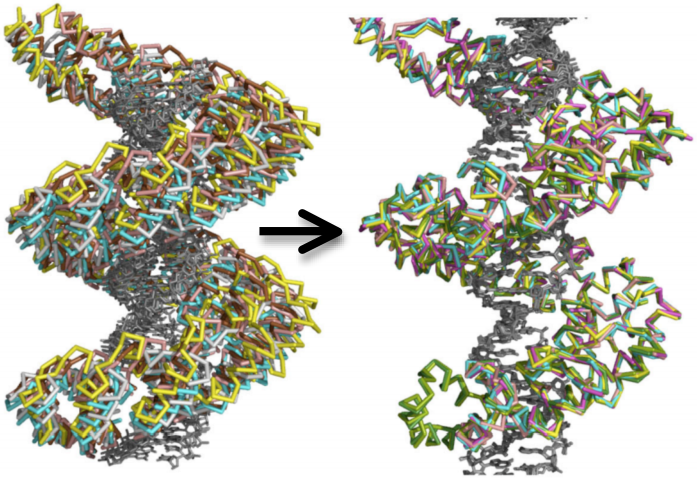
0.1 M MES pH 6
0.25 M Sodium Acetate
14% v/v PEG 400

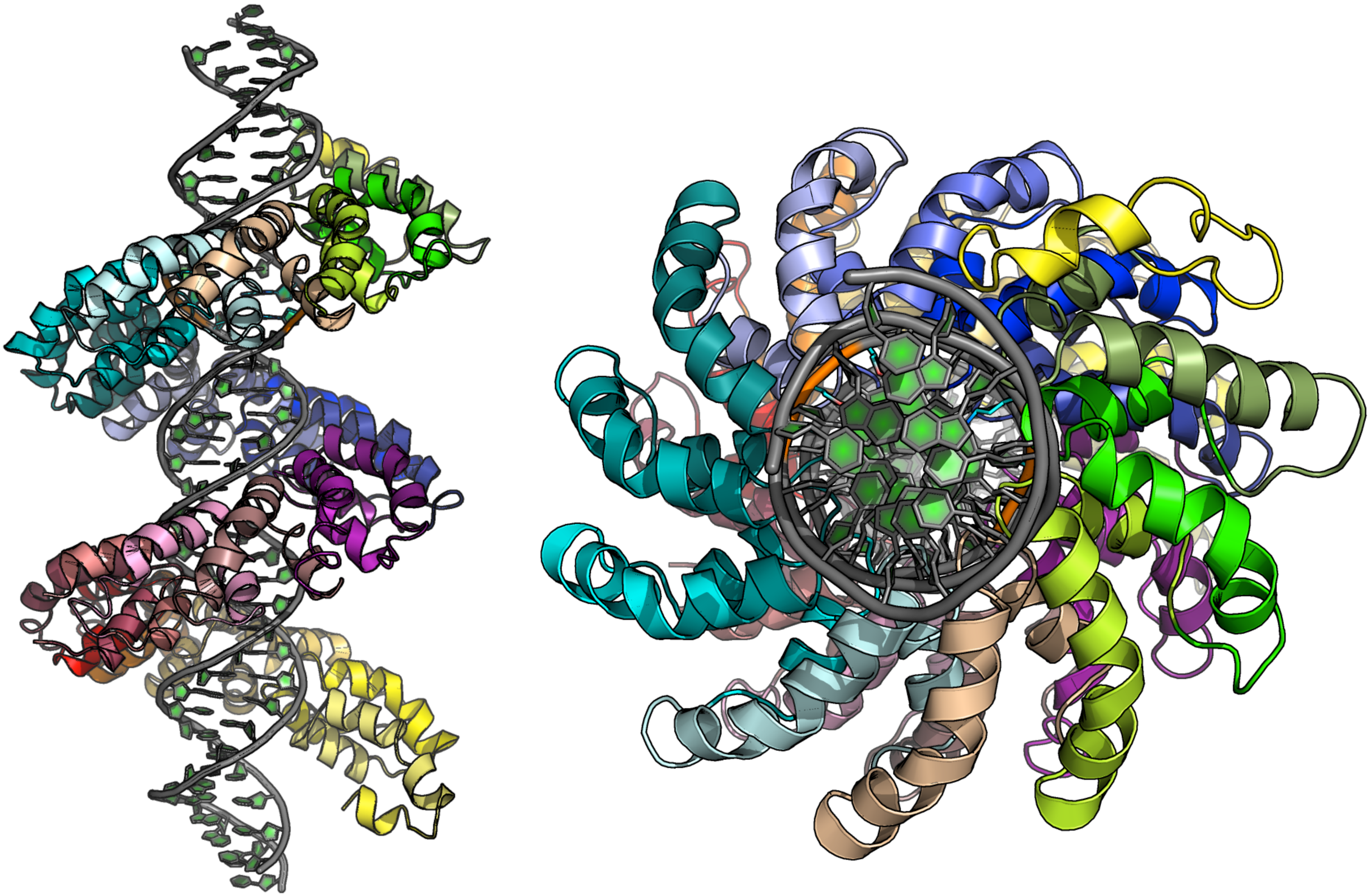
Barry Stoddard
Basic Sciences
FHCRC

Model-based structure determination

- Heroic efforts by Amanda Mak to identify constructs/conditions that would yield good crystals
- Initial attempts at experimental phasing were unsuccessful
- Molecular replacement searches with *de novo* models gave good scores, reasonable crystal packing
- Large-scale model-building and iterative refinement led to high-resolution structure







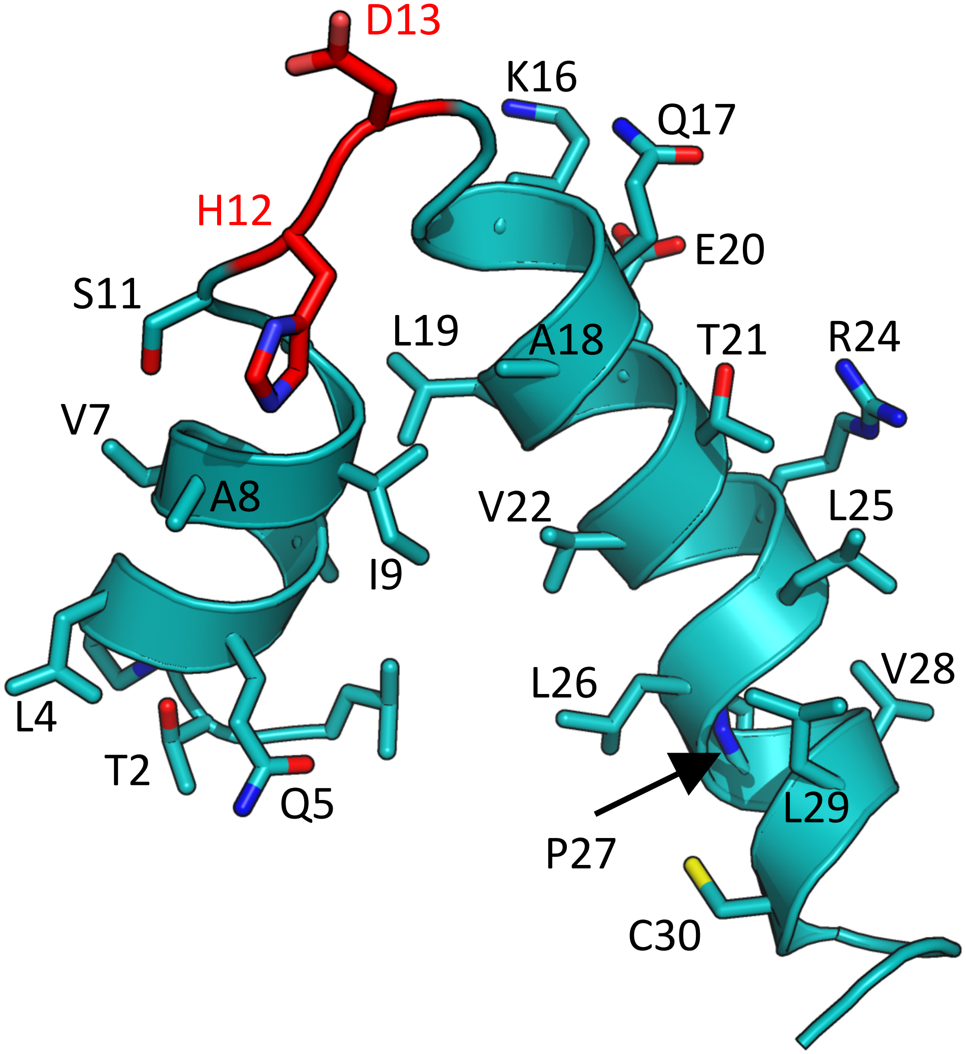
3.0 Å resolution (96.6% completeness; 5.6x redundancy)

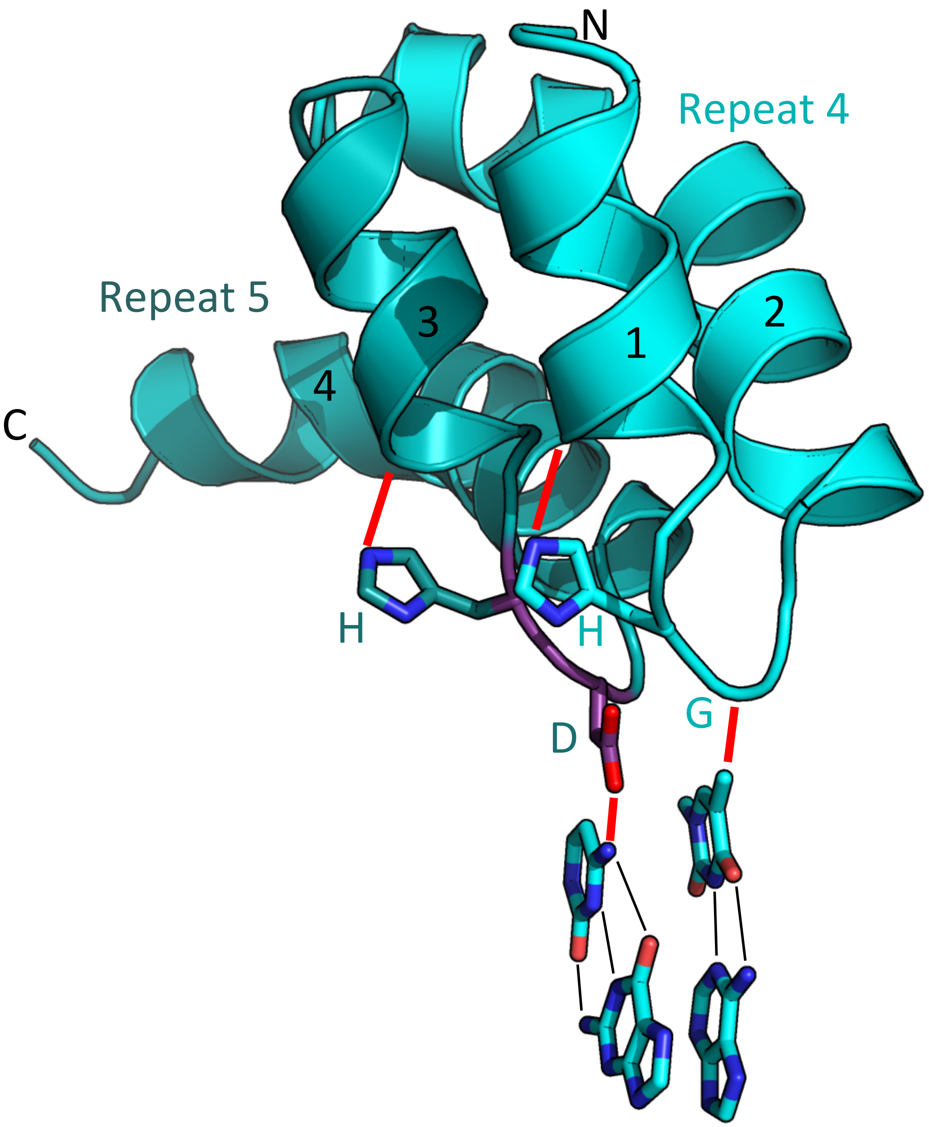
$R_{work/}R_{free}$ = 0.264 / 0.296   (2086 protein atoms, 1552 DNA atoms, 216 solvent molecules)

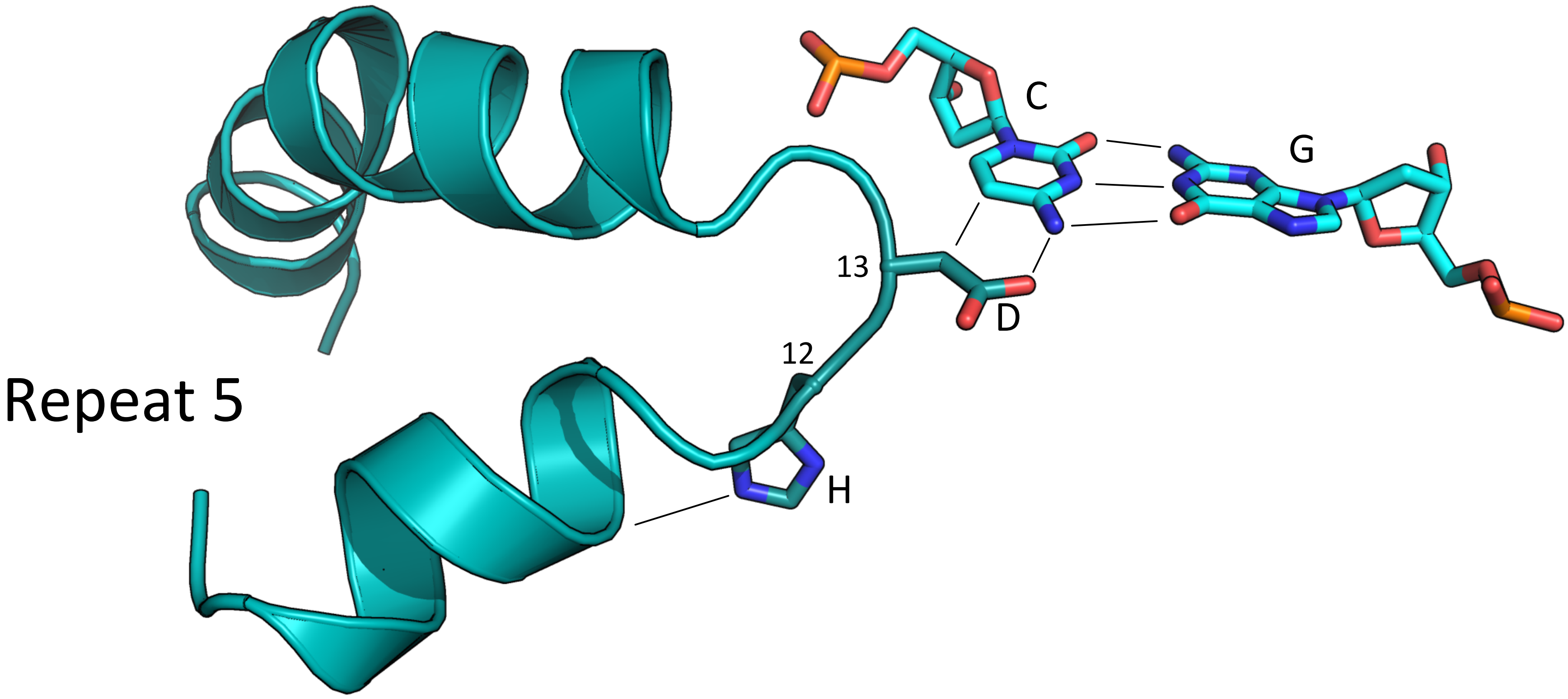
Ramachandran Distribution: 73.6 % core, 26.4% allowed; 0.0% in generous/disallowed

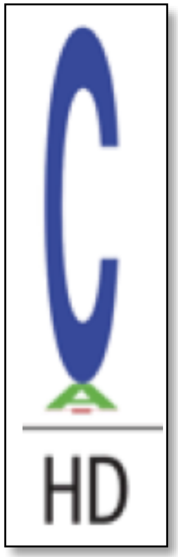
Individual TAL repeats form left-handed helical bundles that self-associate to interact with sequential bases of the DNA target sense strand.

The RVDs occupy a loop that connects the repeat helices, penetrates the DNA major groove and interacts with DNA bases





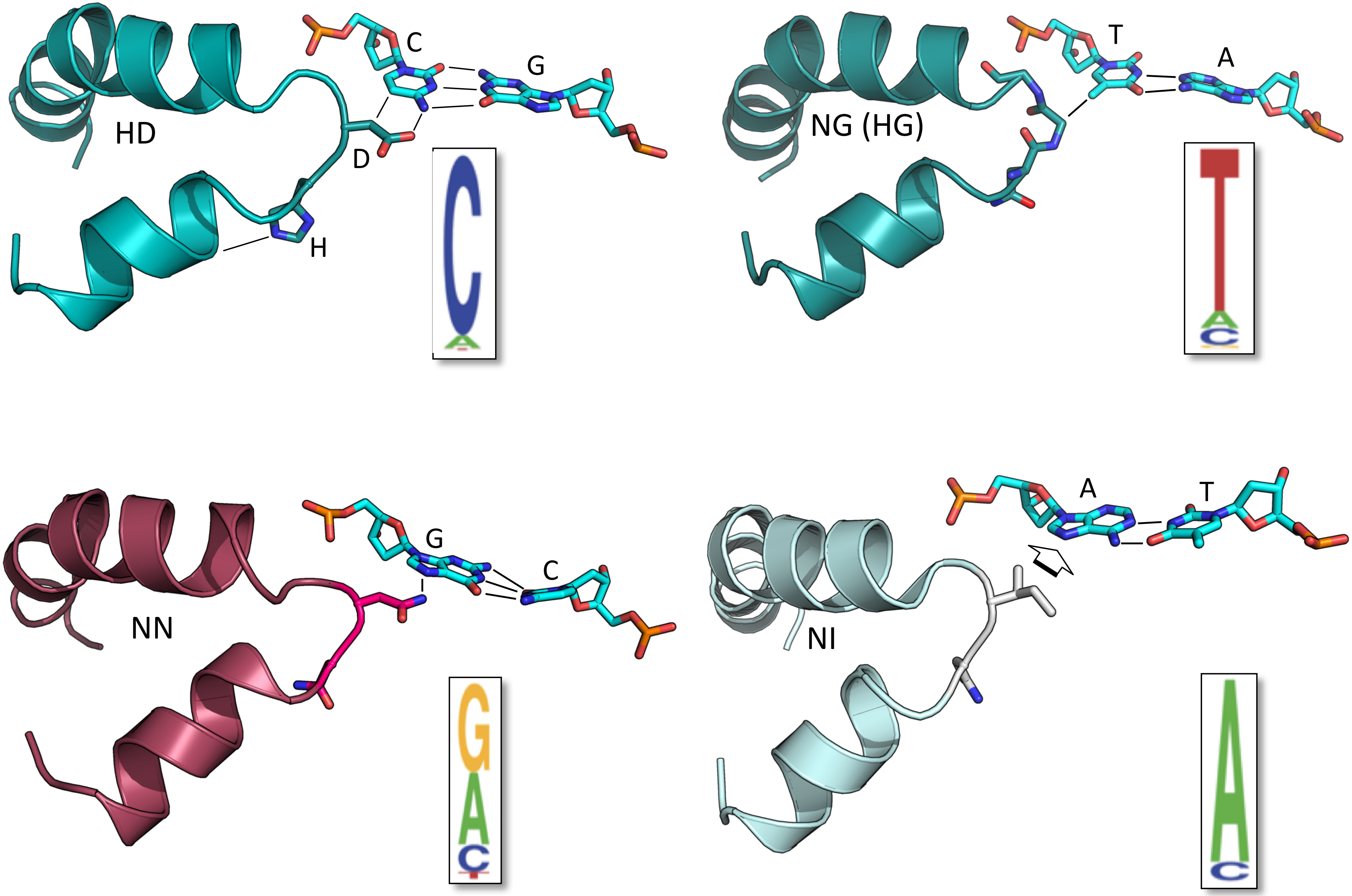




Residue 12 (usually N or H) of each RVD makes a structural interaction with preceding protein backbone carbonyl to stabilize loop conformation.

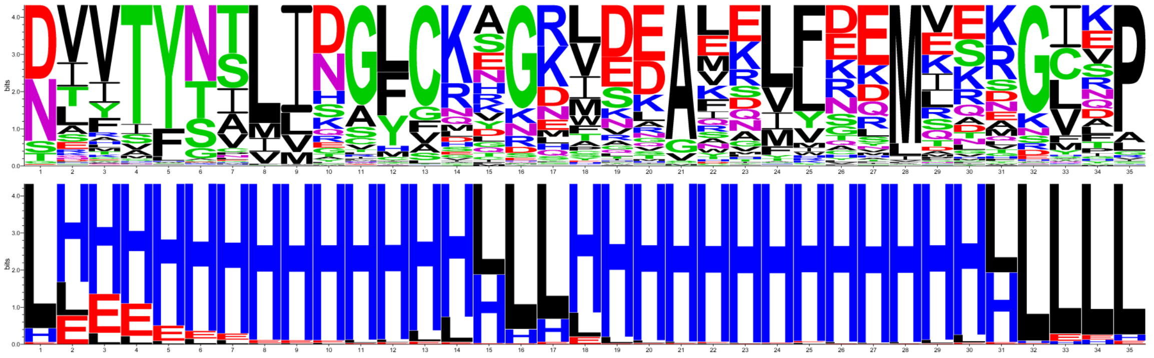
Residue 13 makes individual base-specific contacts

The most common sequence-specific RVDs (HD, NG, HG, NN, NI)



Pentatricopeptide repeats (PPR)

- First identified in *A. thaliana*
- a large family of mitochondrial and plastid proteins thought to bind RNA and regulate processing, editing, and translation
- greatly expanded in land plants (~450 in *A. thaliana*)
- tandem, degenerate ~35 amino acid repeats
- suggested to bind RNA in a modular, 1-1 fashion
- some experimental evidence on residues important for specificity (Ian Small; Alice Barkan)

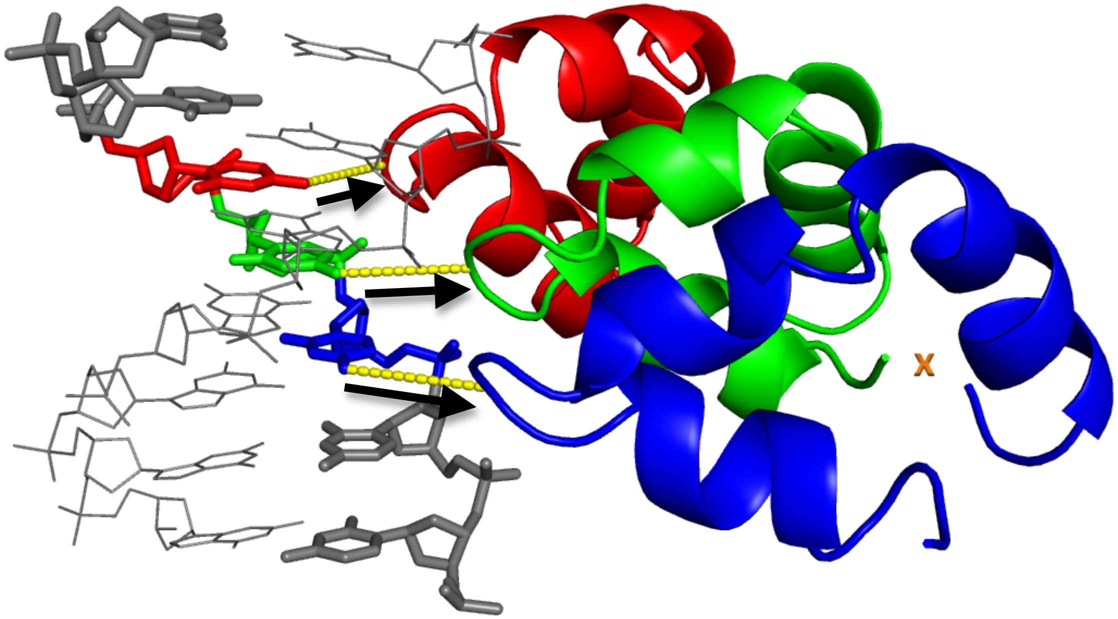


Predicting PPR:RNA interactions

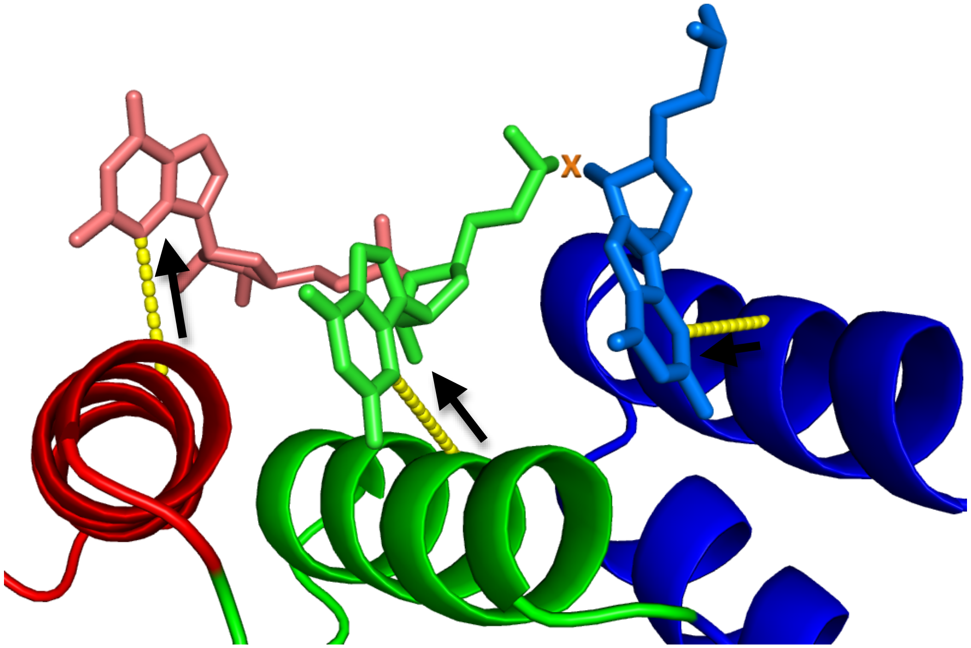
- Assume structural symmetry of protein repeats and RNA partner
- Build protein as connected, symmetric peptide chain
- Anchor RNA bases to protein repeats using flexible linkers, preserve symmetry of linkers and RNA conformation
- Generate ~20,000 models, cluster low-energy models

Kinematics

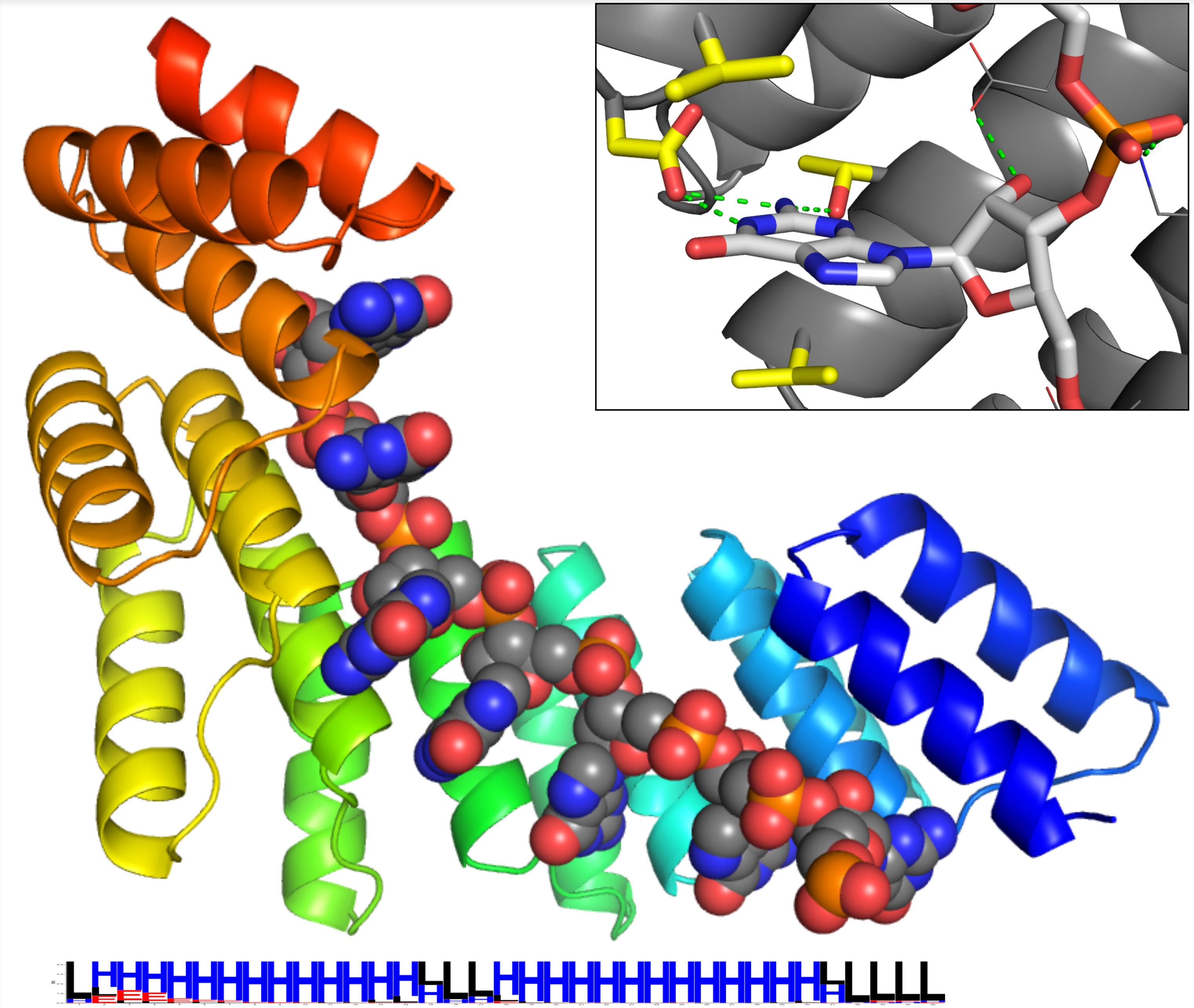
In the TAL:dsDNA simulations (top), the repeat units were built outward from their target base pairs.



In modeling the more flexible single-stranded RNA ligand (bottom), protein chain connectivity is maintained, RNA is built outward from protein



x = chainbreak



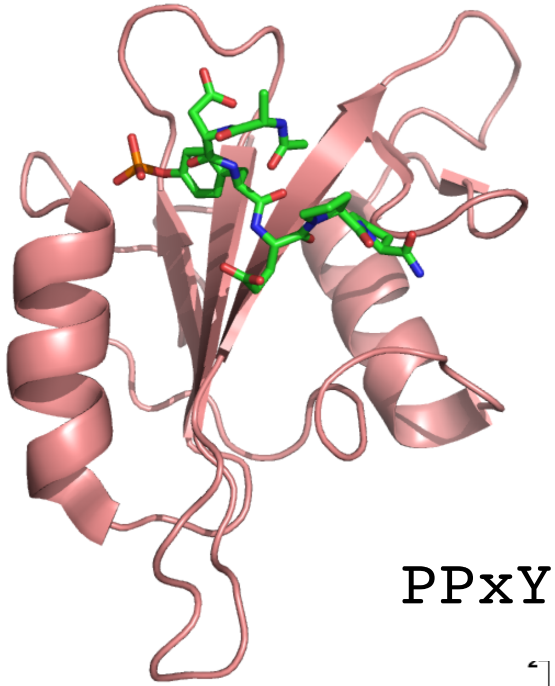
8***D**V**V**TY**T**TLIDGLCKAGRLDEALELFDEMVEKGIKP -> 7*Gua

12345678901234567890123456789012345

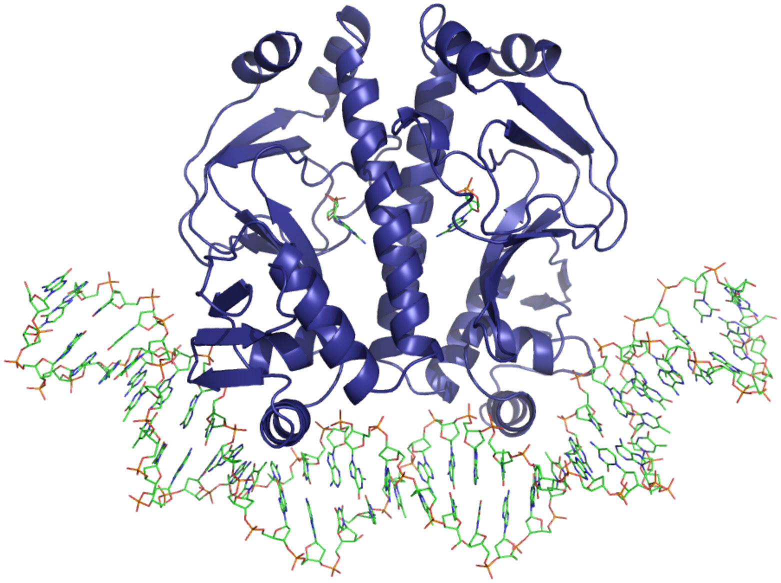
Thank you

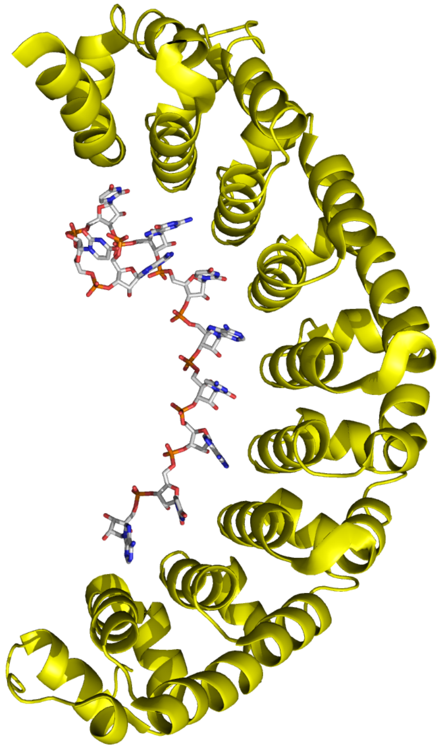
- Lab members:
  - Chen Yanover
  - Angela Liu
  - Chris King
  - Shen Li
  - Cecile Morales
- Barry Stoddard and Amanda Mak
- Adam Bogdanove
- Funding:
  - NIH
  - Searle Scholars
  - Sloan Foundation
  - FHCRC

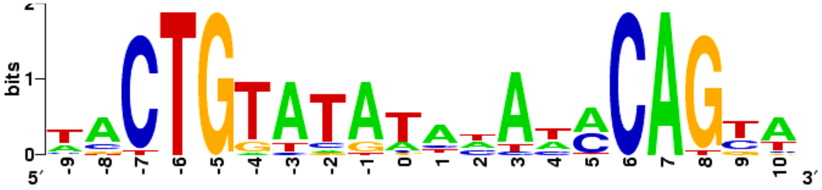
Many proteins recognize linear motifs in the sequences of their polymer partners



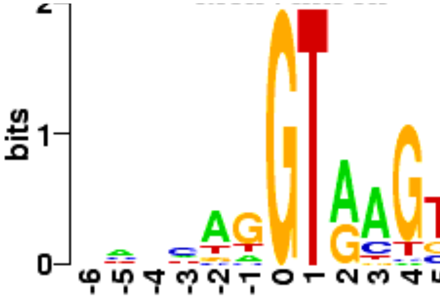
PPxY

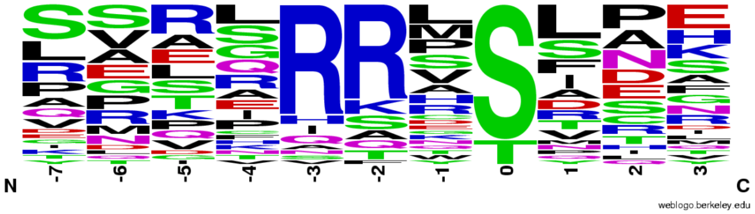




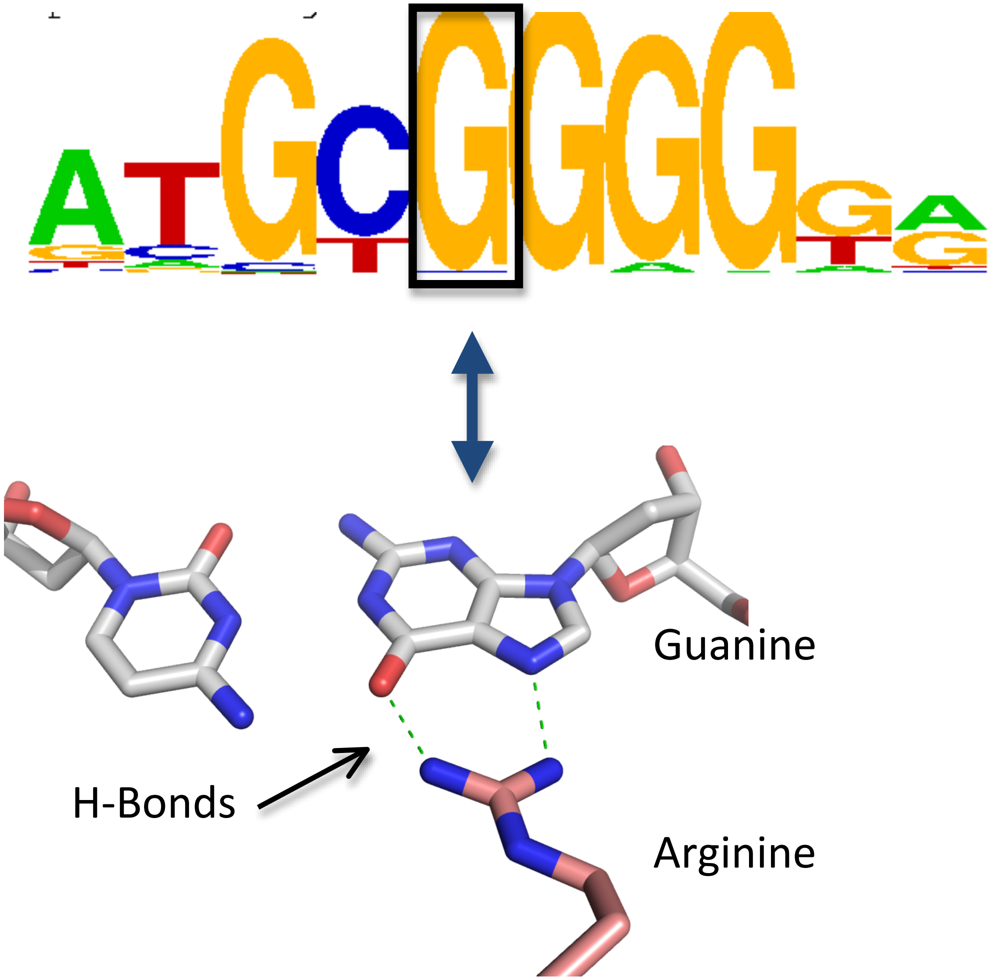


P[TS]x[LVI]P





Examination of three-dimensional structures suggests that structural modeling might be used to predict these interactions



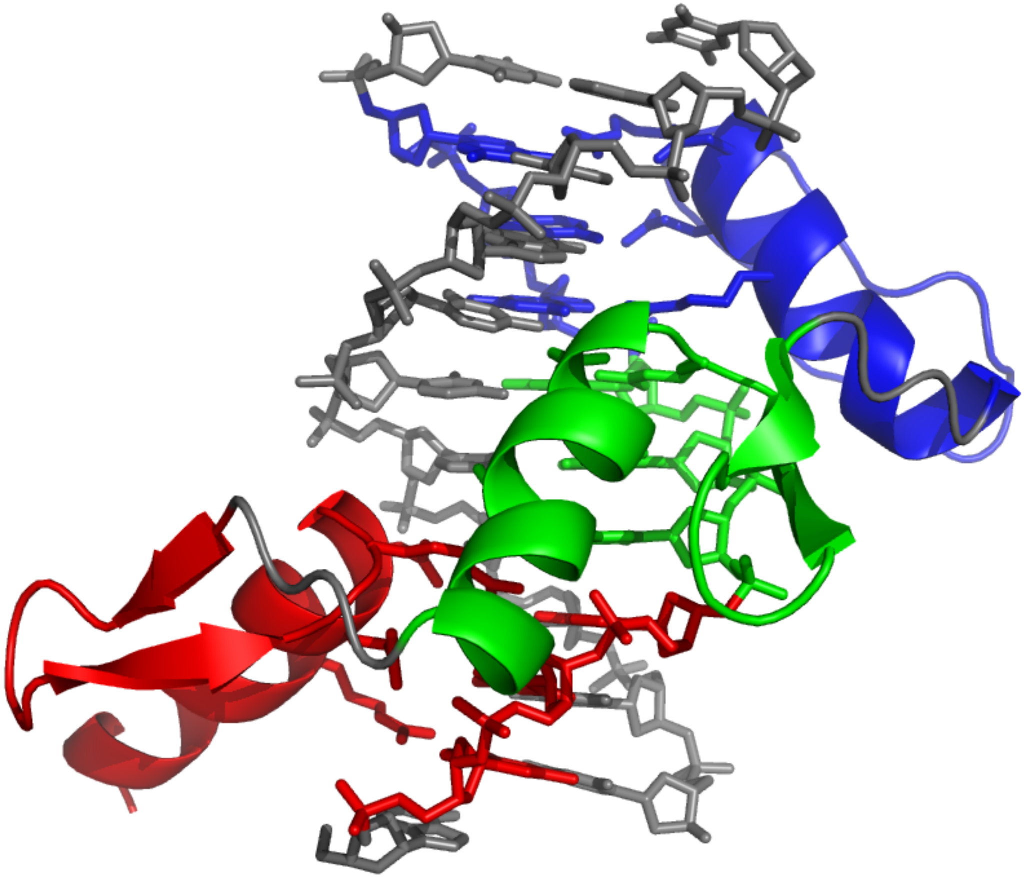
Model system: C_2H_2 Zinc Fingers

The C_2H_2 zinc finger family accounts for roughly half of all human transcription factors.

Each finger recognizes ~3 base pairs of DNA

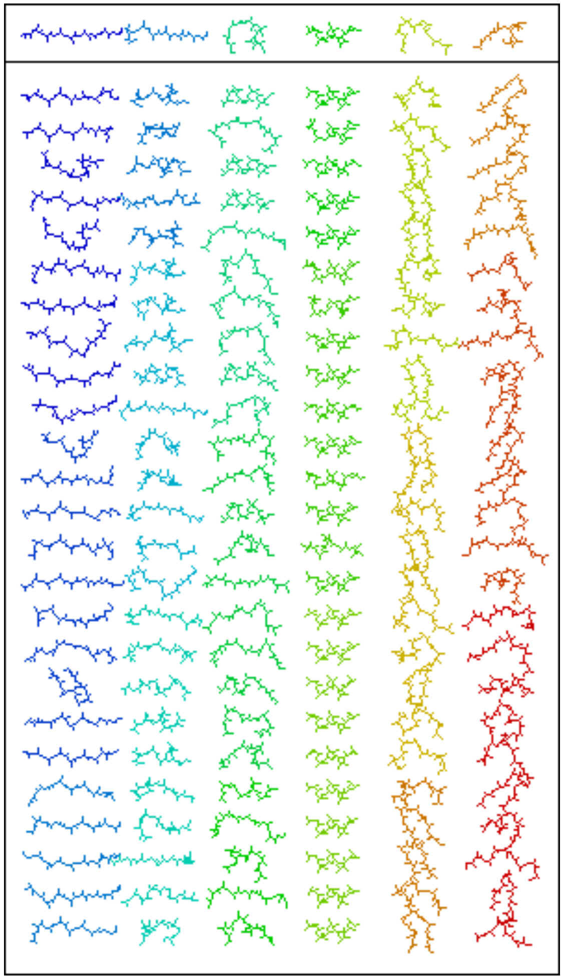
ZF proteins often have tandem arrays of 2 to 20 fingers.

Have been engineered to bind to new target sites.

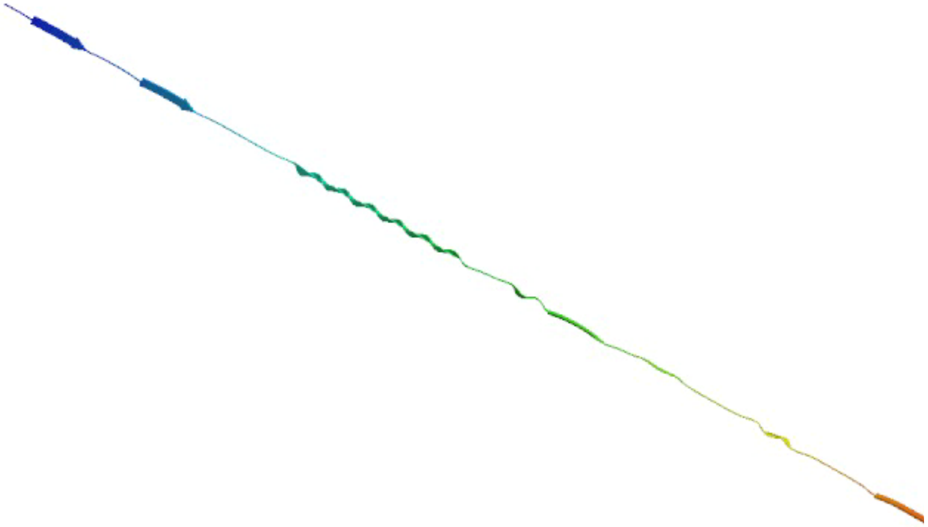


Multiple experimental structures available, but diversity in binding site sequence and structure makes template-based predictions challenging

Sampling tools from *de novo* structure prediction: fragment assembly

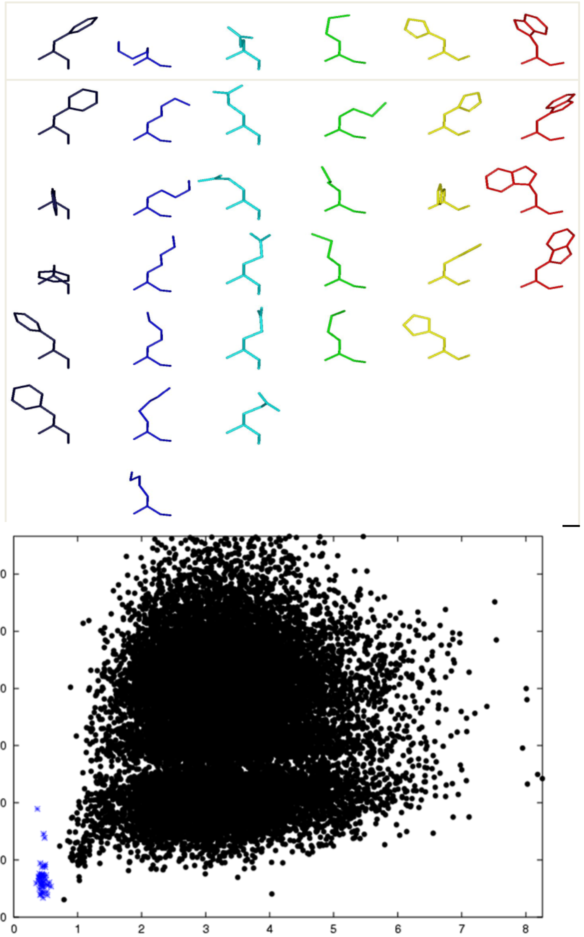


Fragment libraries use local sequence to identify candidate backbone conformations

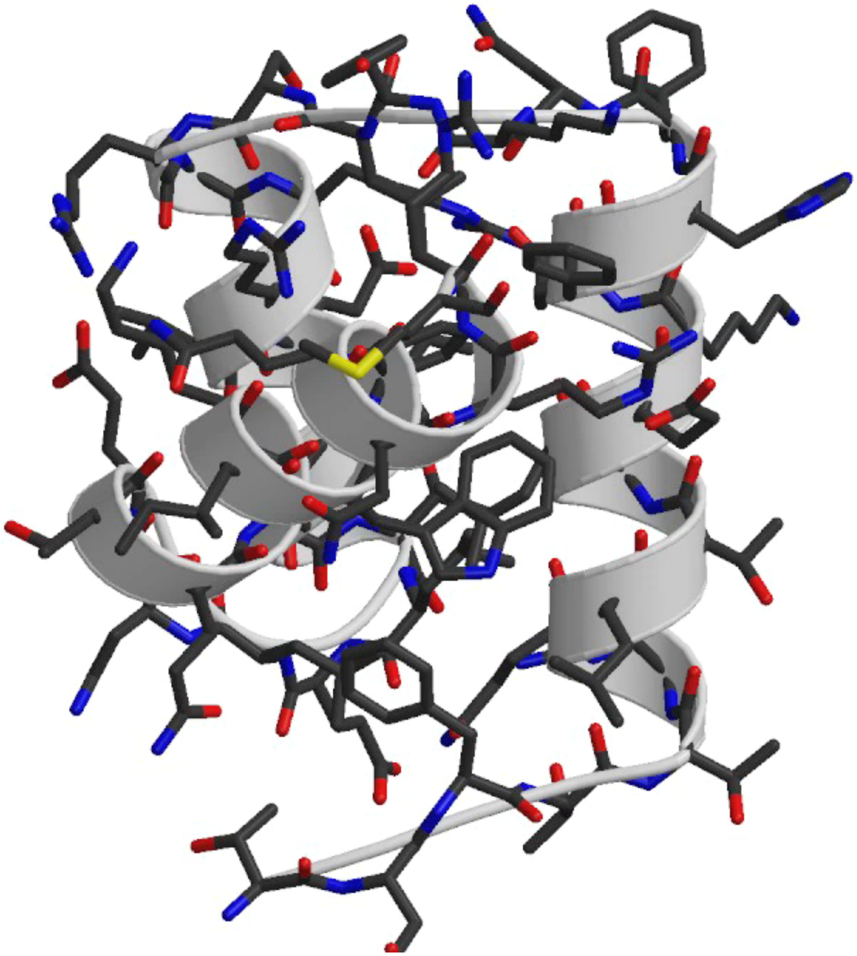


Fragment assembly simulations can efficiently explore conformational space

Sampling tools from *de novo* structure prediction: all-atom refinement

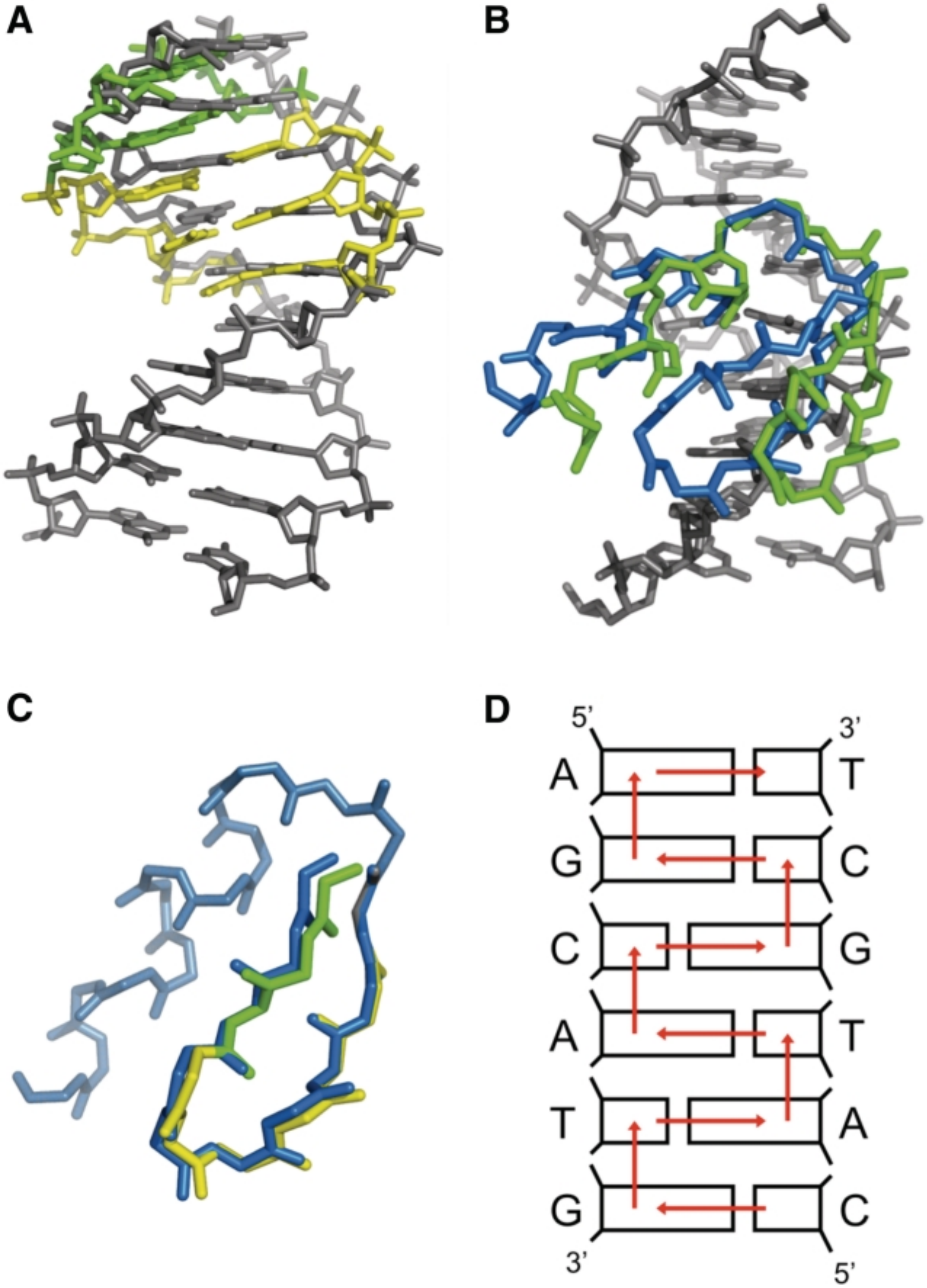


Rotamer libraries and physically realistic potential energy functions model sidechains



All-atom refinement simulations can pick out native-like conformations

Protein-DNA interfaces require new sampling moves

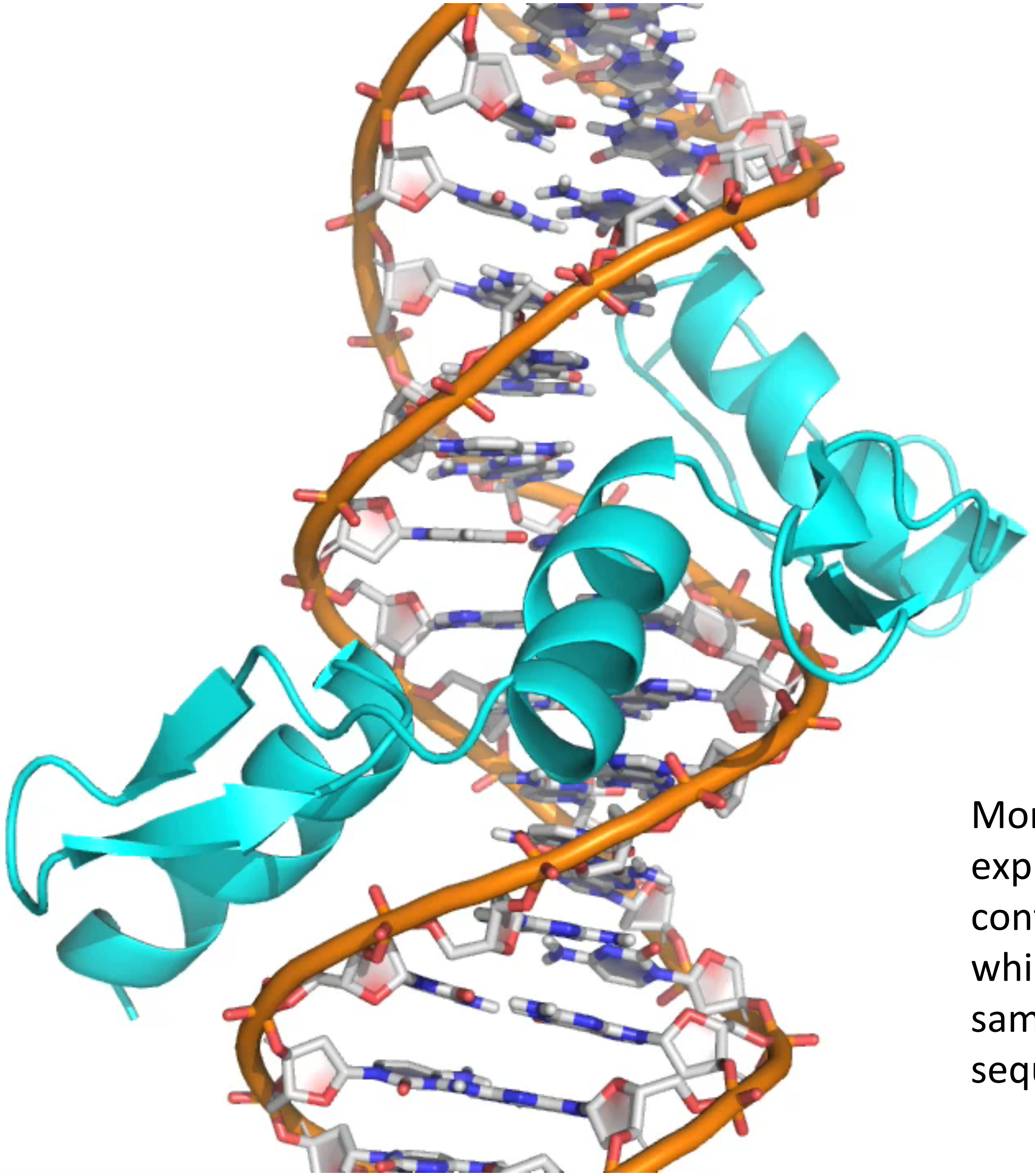


Double-helical DNA fragment insertions preserve base-pairing outside the region of fragment insertion

Interface moves sample the protein-DNA rigid body orientation using homologous structures as templates

Protein fragment insertions sample backbone conformation without perturbing DNA or binding mode

Kinematic structure for DNA allows torsion-space (internal coordinate) sampling while maintaining the DNA duplex

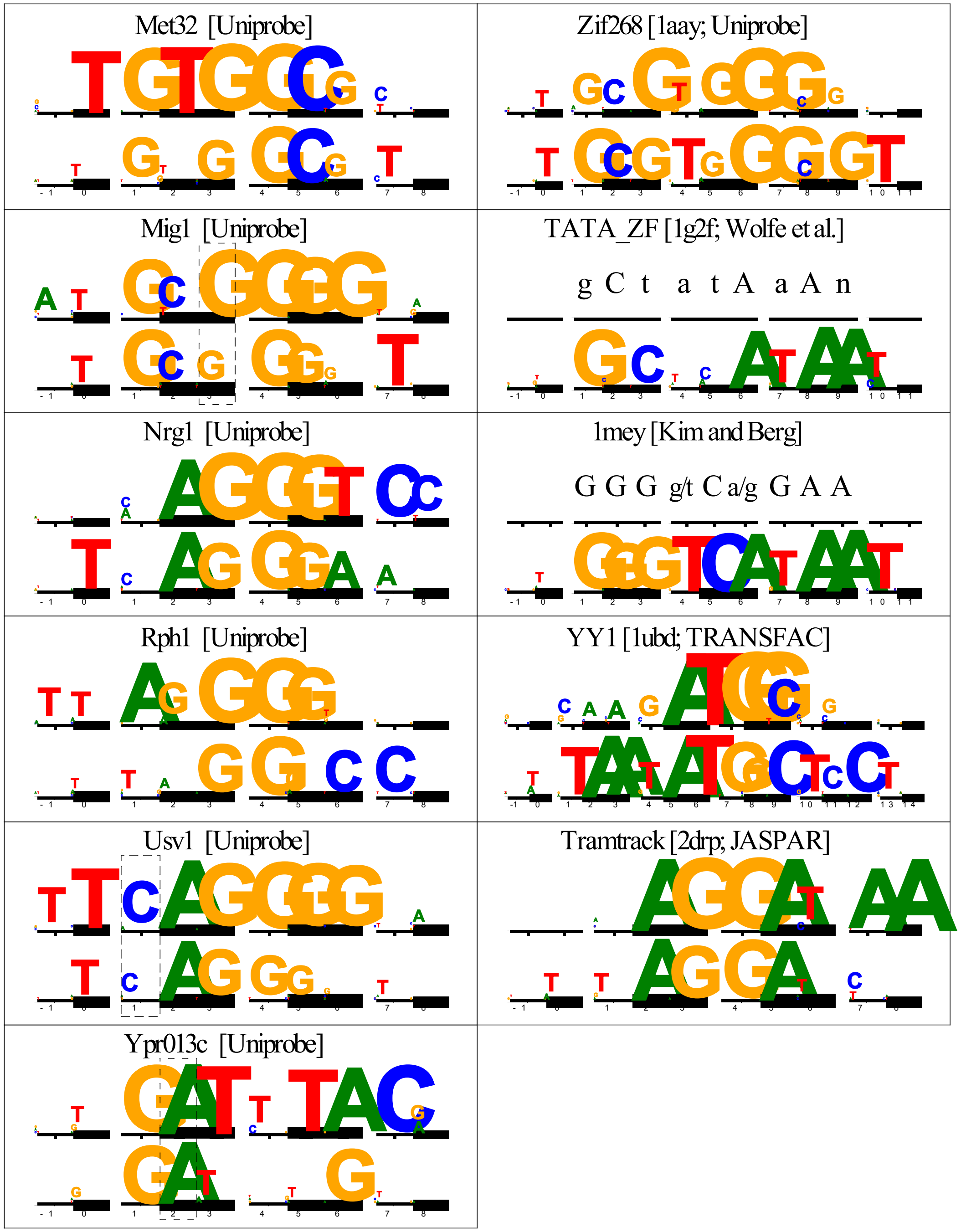


Monte Carlo simulation explores protein-DNA conformational space while simultaneously sampling DNA target site sequences

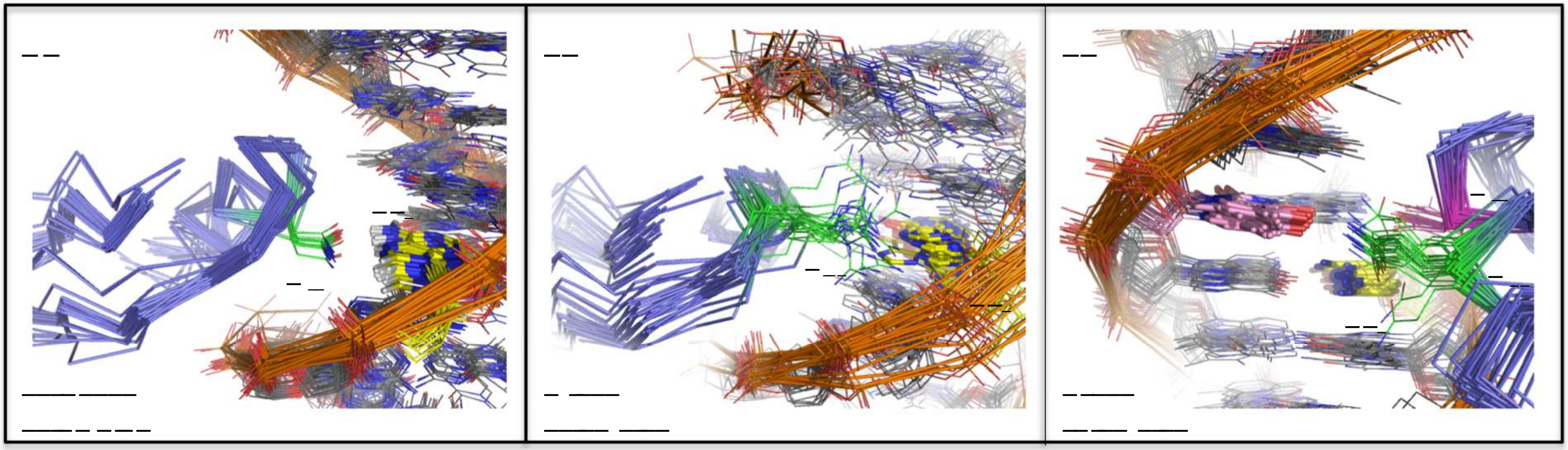
Predictions for a benchmark set of ZFs with 2-4 fingers

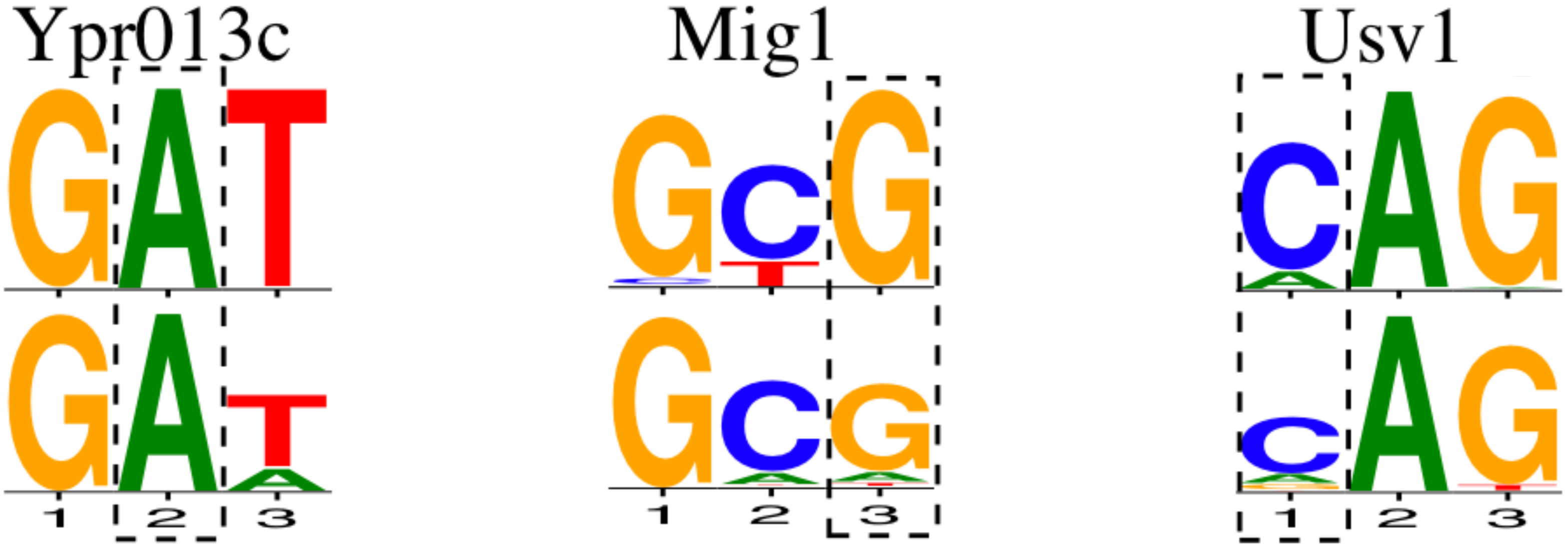
75-80% prediction accuracy

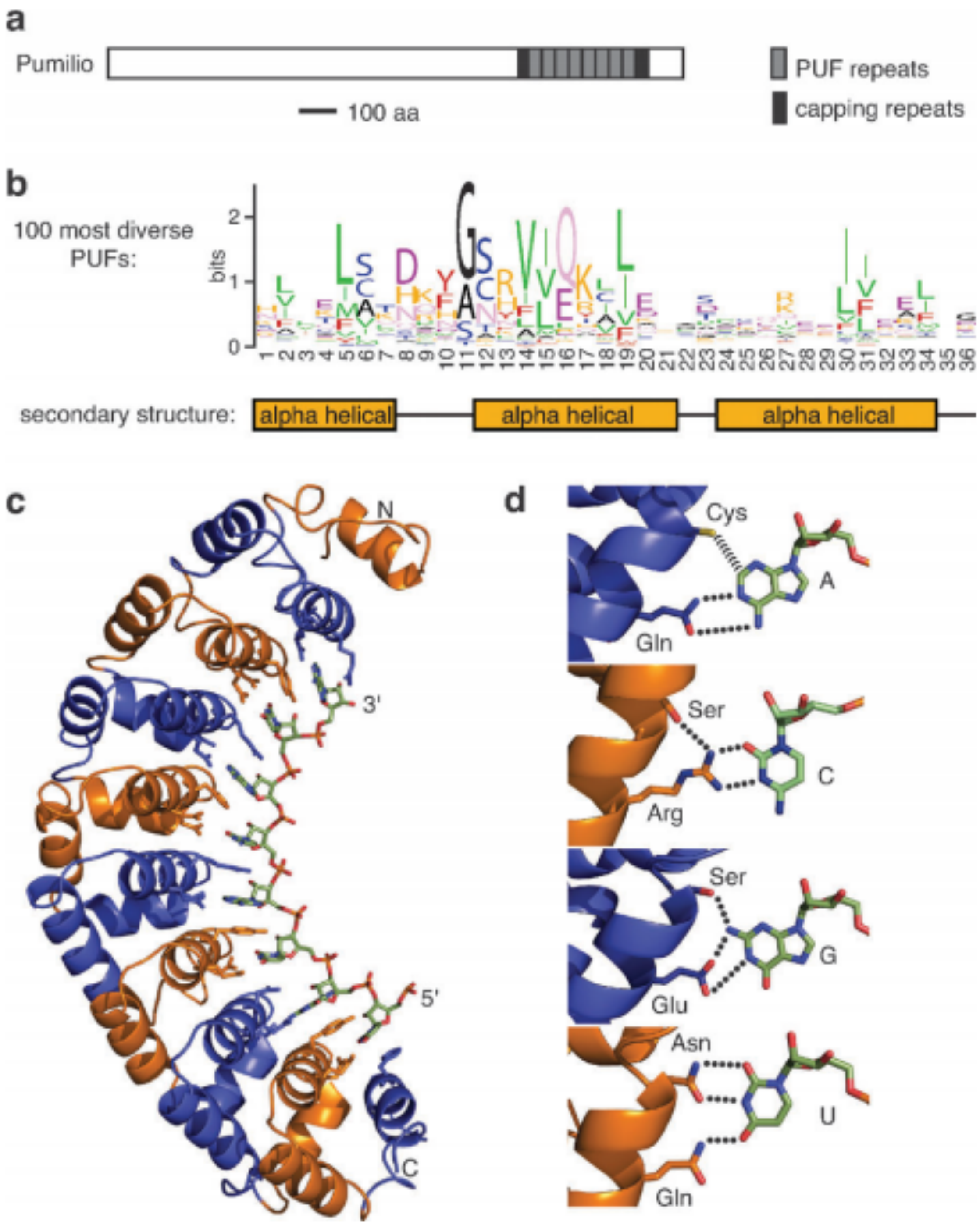
Similar performance (80%) on larger set of ~400 engineered zinc fingers



Simulations suggest structural basis of binding specificity







Perhaps related to PUF repeat proteins, which recognize ssRNA in a modular, 1-1 fashion, or tetratricopeptide repeat (TPR) proteins, which are involved in a wide range of protein interactions

(look at some helical repeat proteins in PyMOL)

Filipovska & Rackham
Mol. Biosyst 2012