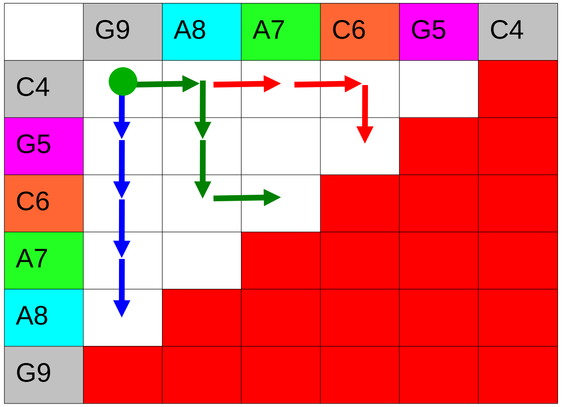
An Enumerative Ansatz for RNA and Protein Modeling

Rhiju Das

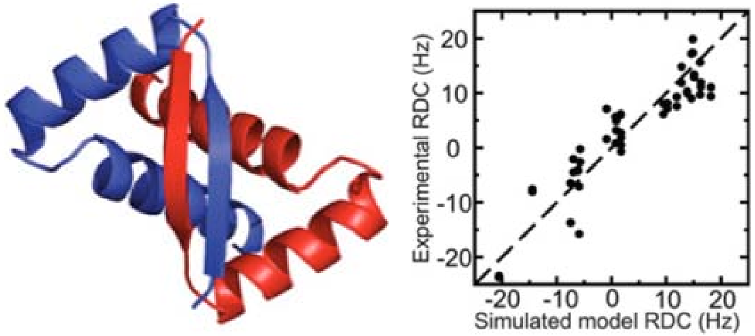
Aug. 4, 2010

RosettaCon!

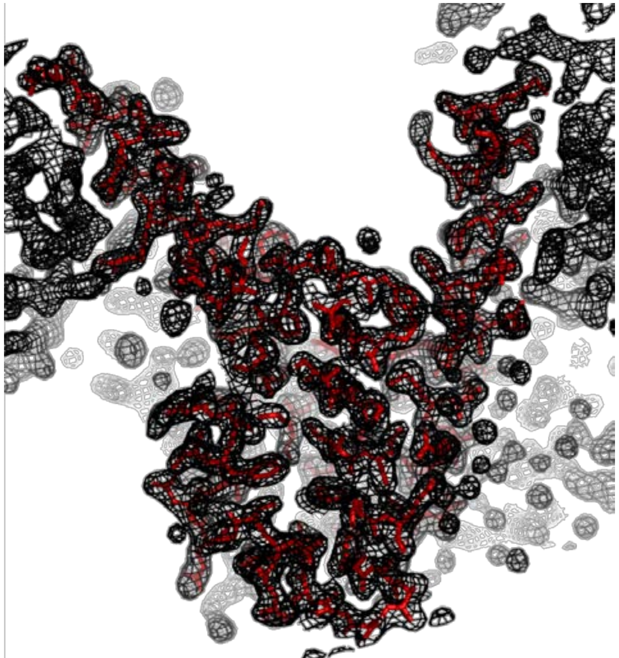


	G9	A8	A7	C6	G5	C4
C4	●					🟥
G5					🟥	🟥
C6				🟥	🟥	🟥
A7			🟥	🟥	🟥	🟥
A8		🟥	🟥	🟥	🟥	🟥
G9	🟥	🟥	🟥	🟥	🟥	🟥

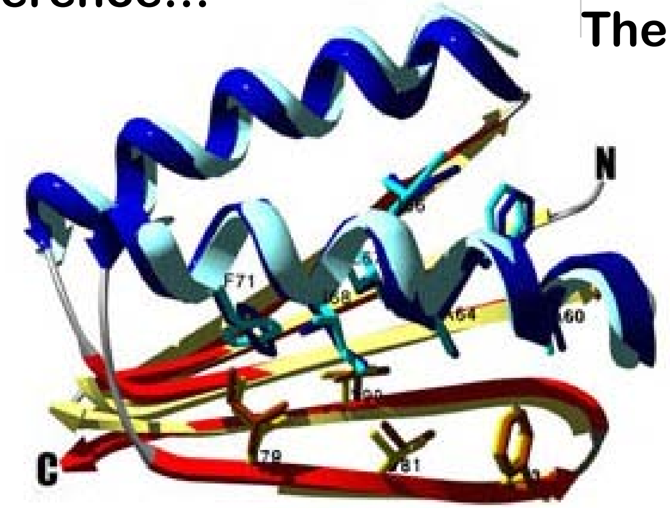
De novo modeling: connections to the real world



Accelerating & enabling NMR structural inference…



The crystallographic phase problem



Engineering new protein folds and new enzymes

This stuff doesn't always work

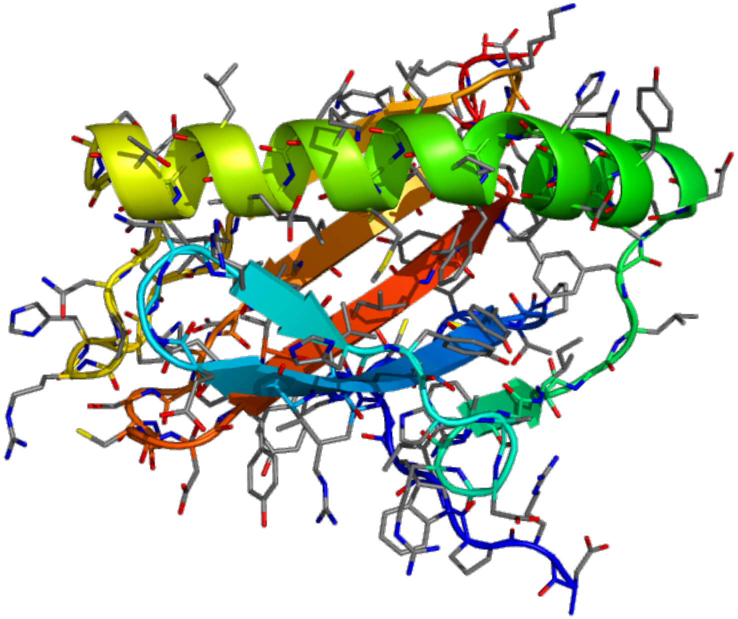
Macromolecule structure at atomic resolution

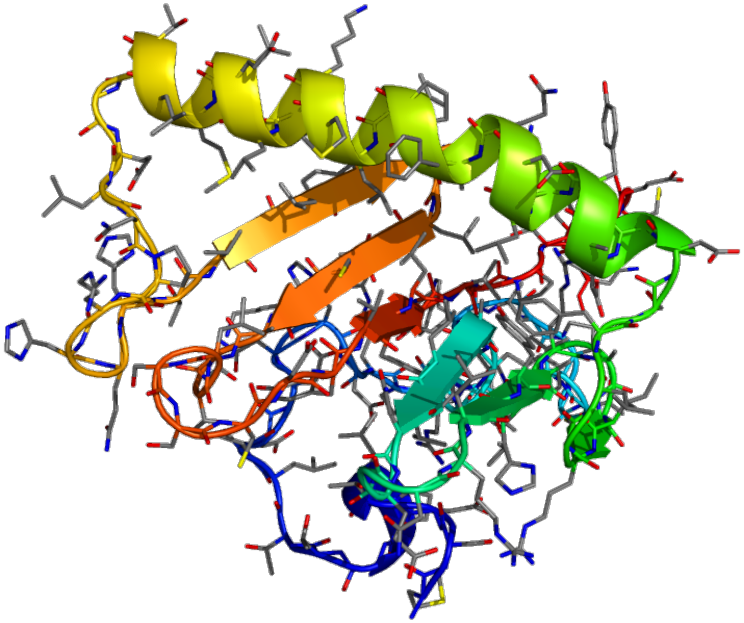
1. Three flaws in our sampling approaches

1. Little RNA puzzles

3. Little protein puzzles

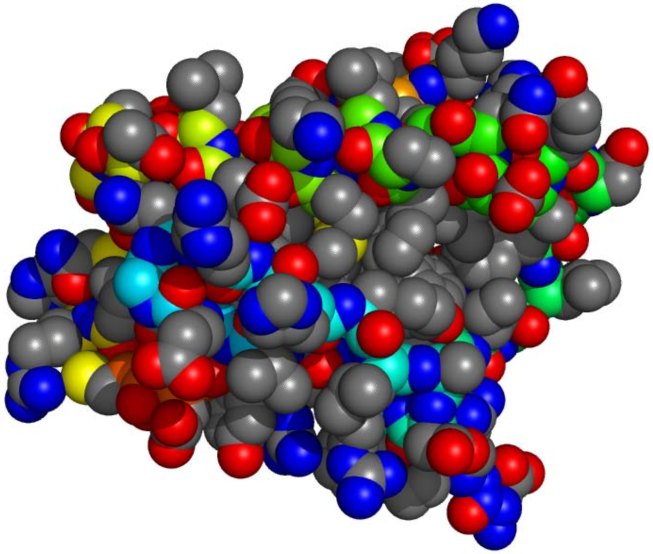
Can you pick out the right one?



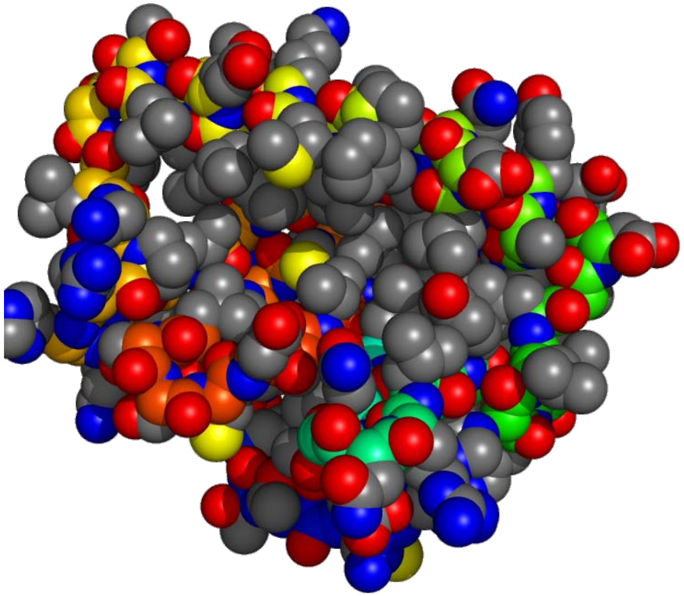


T304 (CASP7)

Can you pick out the right one?



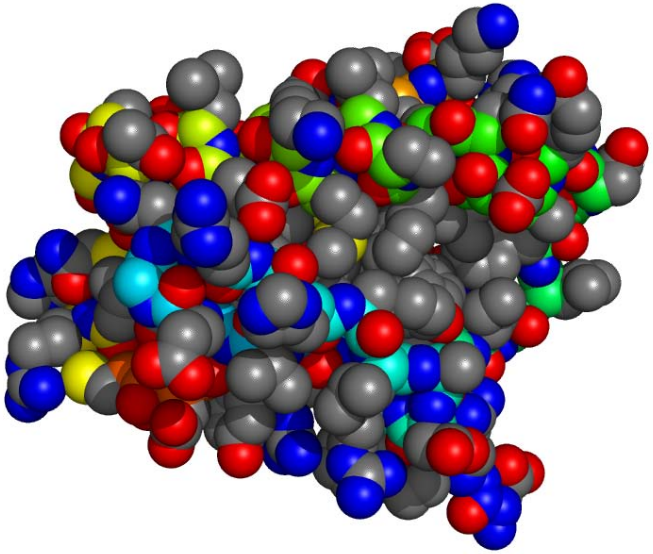
Crystallographic model



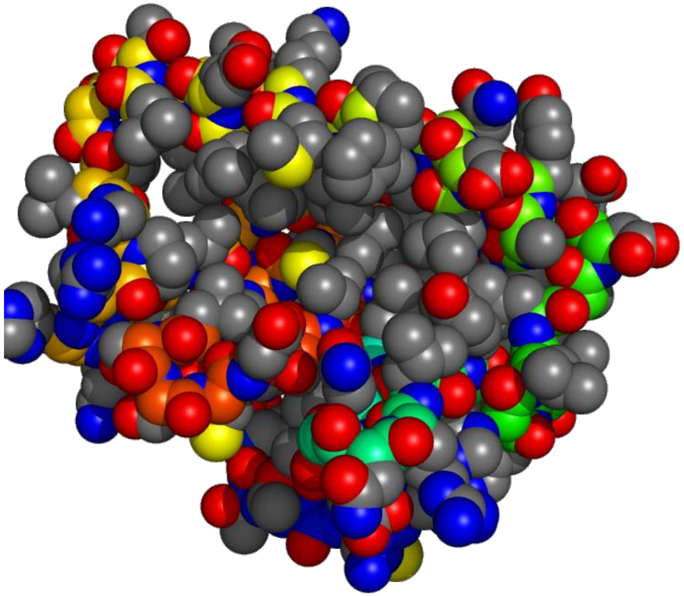
Best CASP model

T304 (CASP7)

Can you pick out the right one?



Crystallographic model

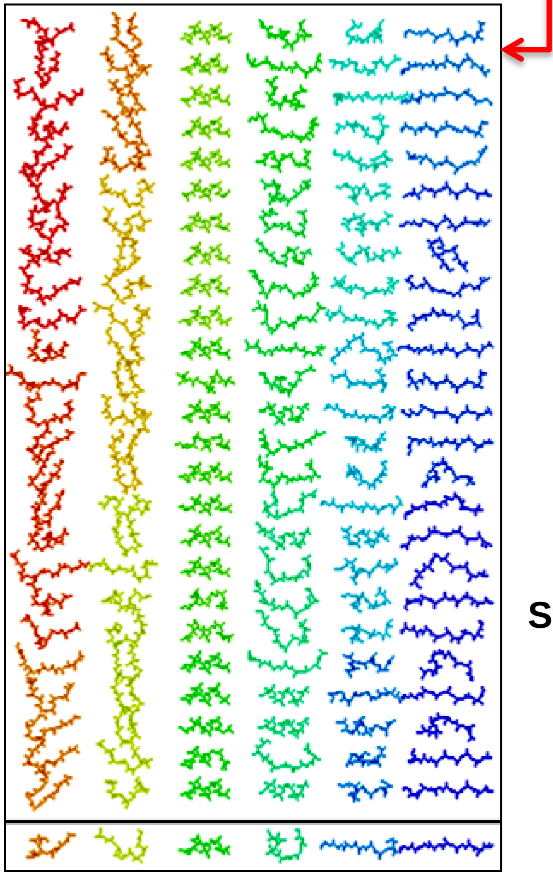


Best CASP model

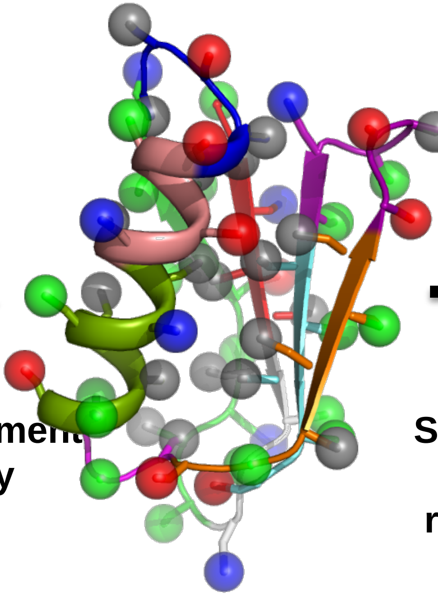
T304 (CASP7)

The state of *de novo* structure prediction

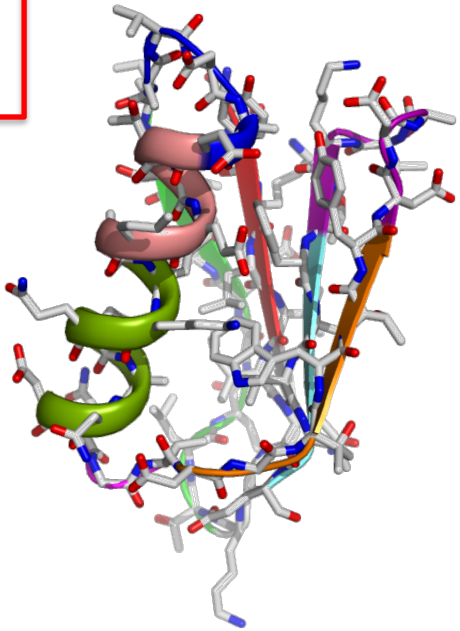
Issue #1. Relying on the existing database of structures.



Issue #3. Not directly searching all-atom energy function



Stage I. Fragment Assembly

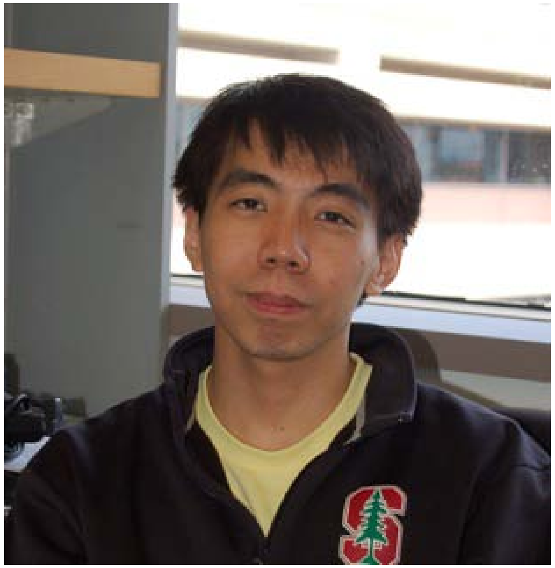


Stage II. All-atom refinement

Issue #2. Randomness – no guarantee of enumeration

The standard ROSETTA routine. SEE ALSO: Work by David Jones, Skolnick & Zhang (TASSER), others

A StepWise Ansatz for 3D modeling



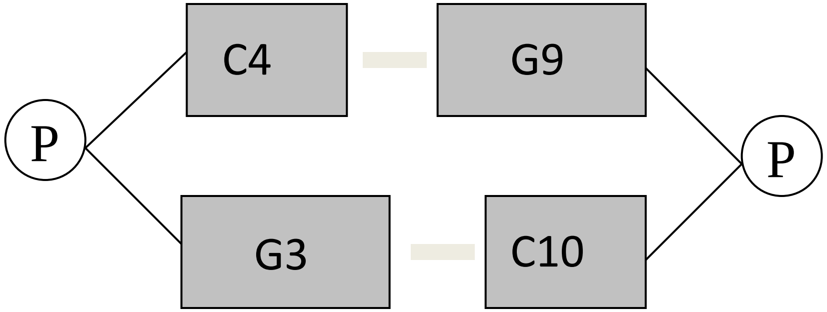
Parin Sripakdeevong

Step-by-step sampling

This sequence forms a ***highly stereotyped fold**** *. What is it?*

C A

G A



P

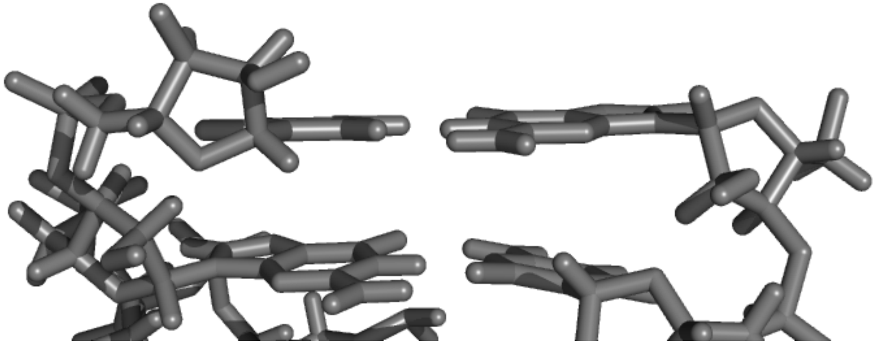
C4

G9

P

G3

C10

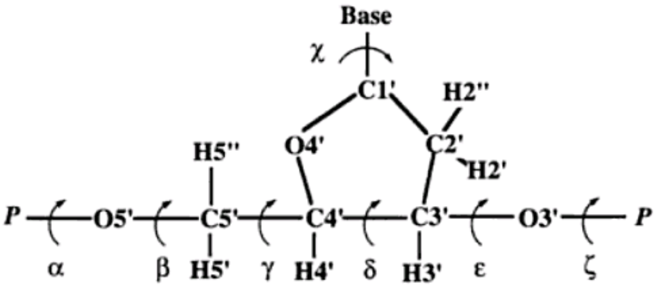


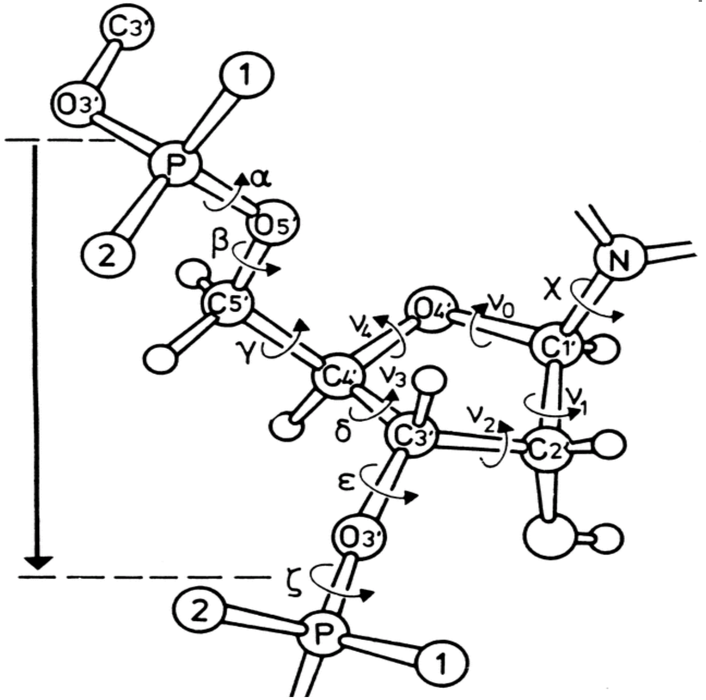
*NMR characterization, multiple crystal models in different helical contexts.

Conformation of a single nucleotide

- **Assume ideal bond length and bond angles**
- **7 torsional degree of freedom**
  - (α, β, γ, δ, ε, ζ, χ)

Q: How many unique conformations?

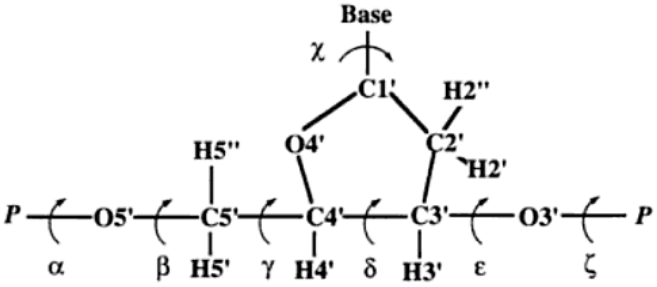




Conformation of a single nucleotide

- **Assume ideal bond length and bond angles**

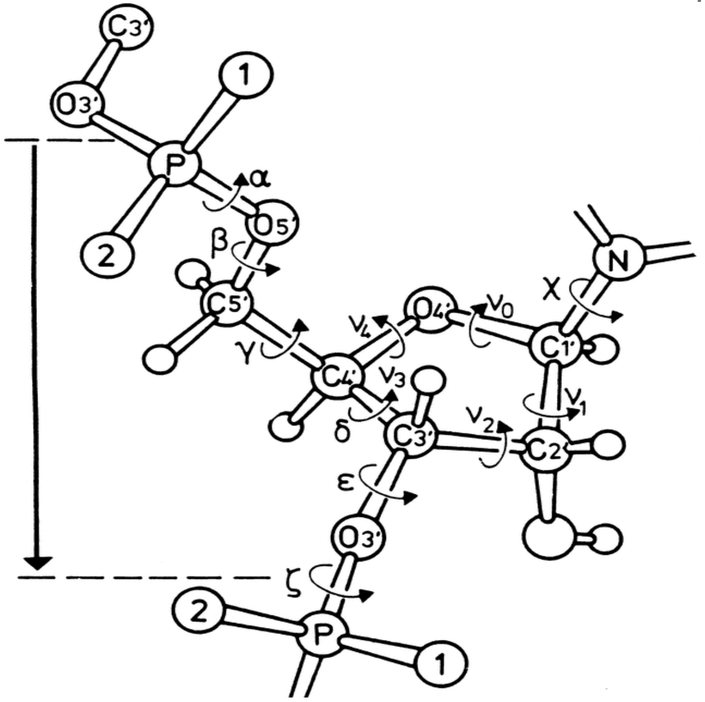
- **7 torsional degree of freedom**
  - (α, β, γ, δ, ε, ζ, χ)



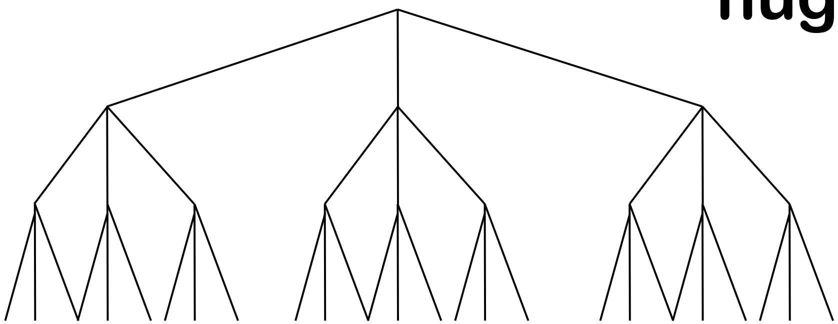
Q: How many unique conformations?

A: Depends on how fine you cluster:

all-atom rmsd cluster size (Å)	# Unique Conformations
3.0	~100
2.0	~1000
1.5	~10,000
1.0	~100,000



Levinthal-style: The conformational space is huge!



# Nucleotides	# Unique Conformations through exhaustive enumerations
1	~10^{5}
2	~10^{10}
4	~10^{20}
10	~10^{50}
20	~10^{100}

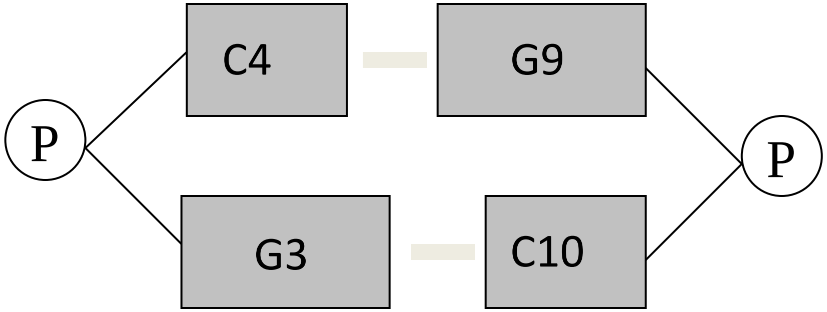
Typical RNA motif length → 10

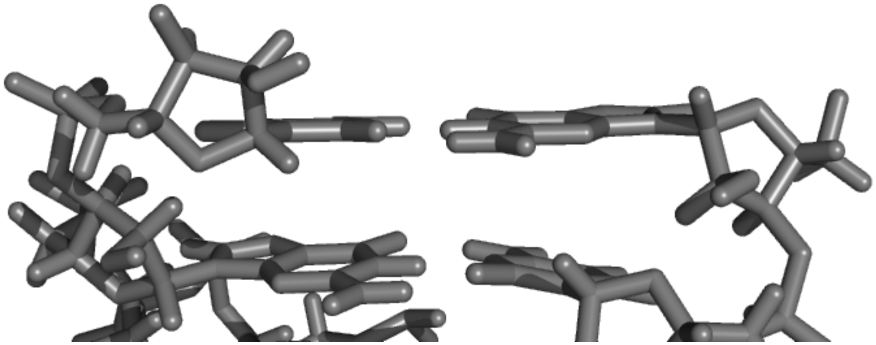
A billion years to sample a tetraloop → ~10^{20}

Step-by-step sampling

C A

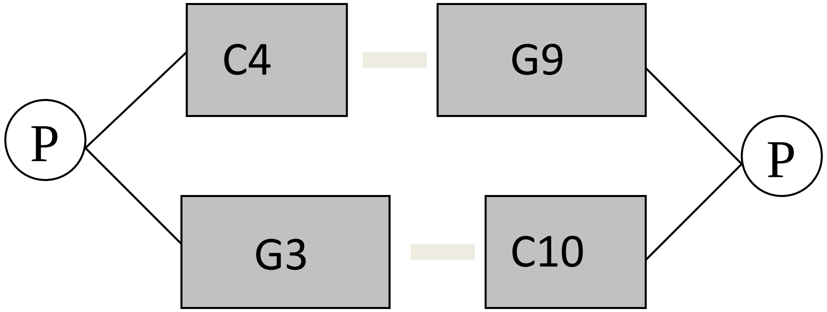
G A

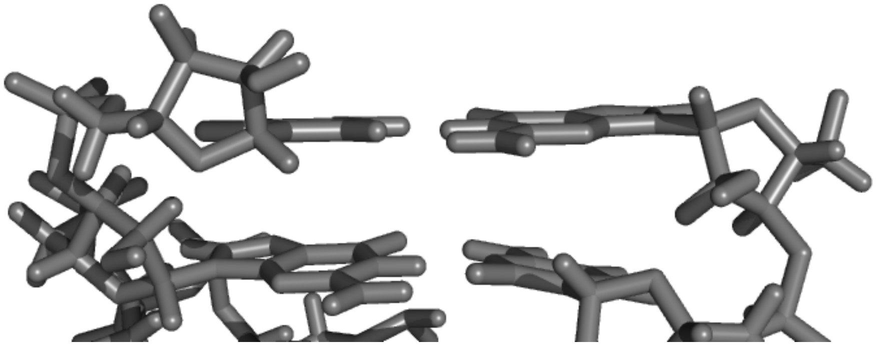




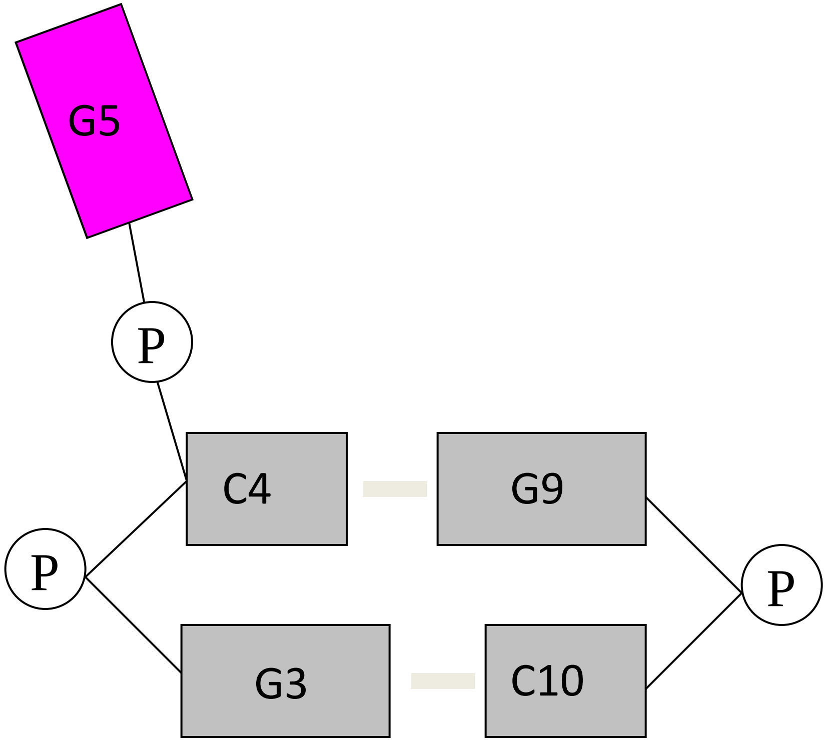
Step-by-step sampling

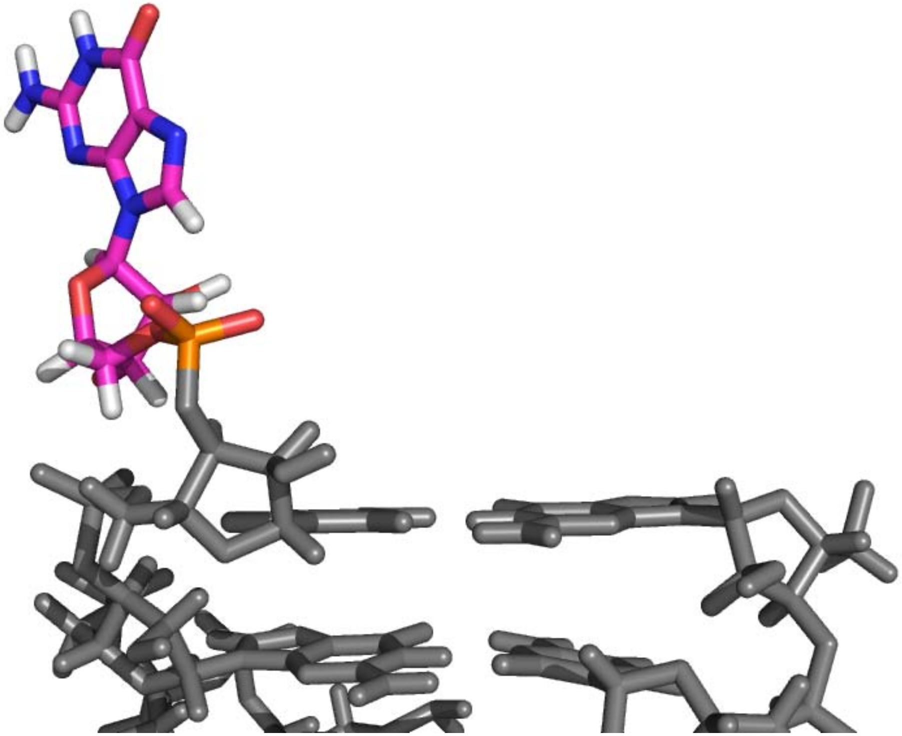
G



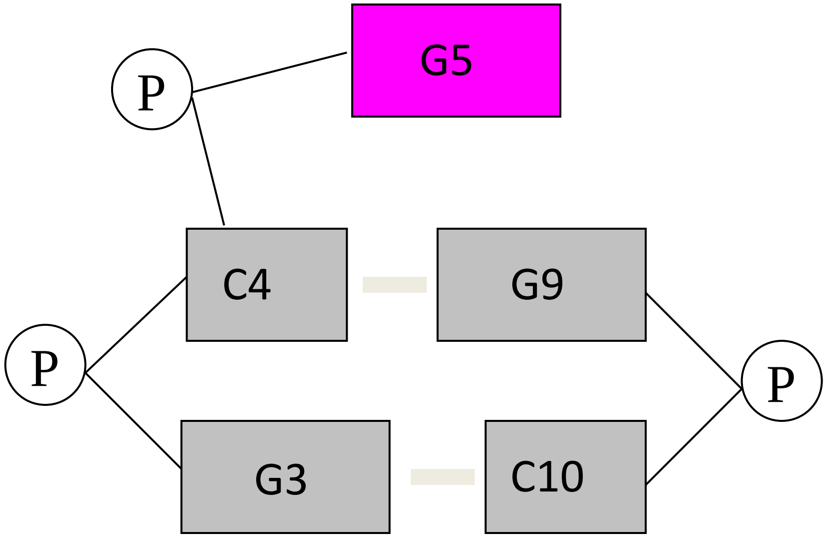


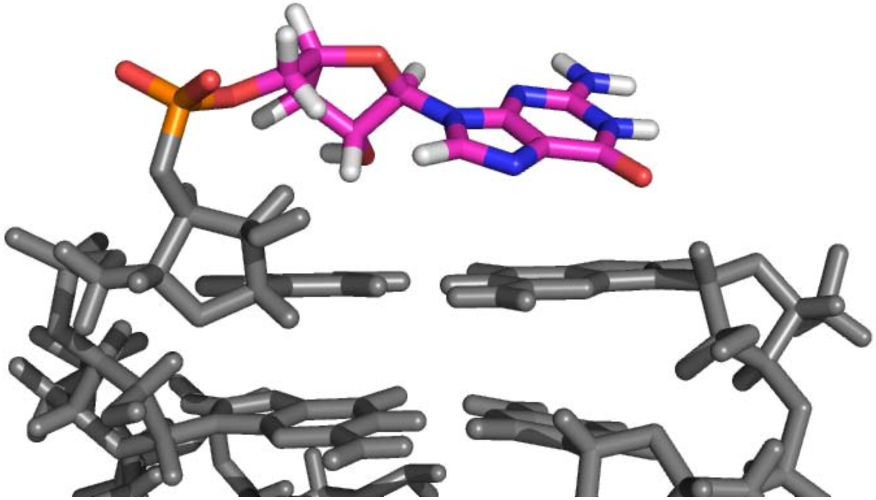
Step-by-step sampling



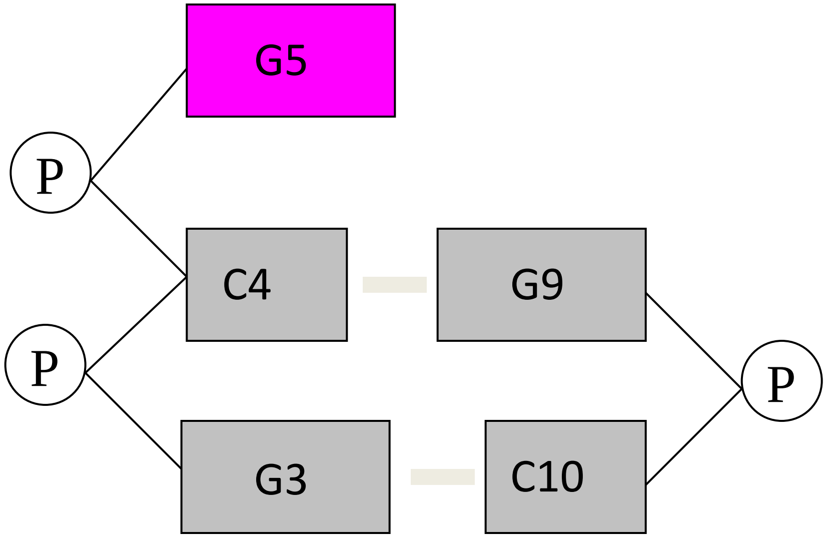


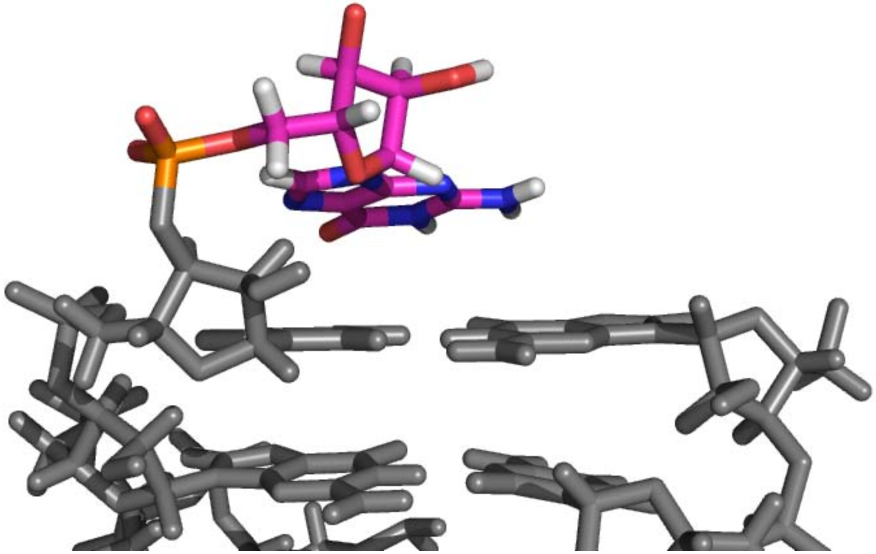
Step-by-step sampling



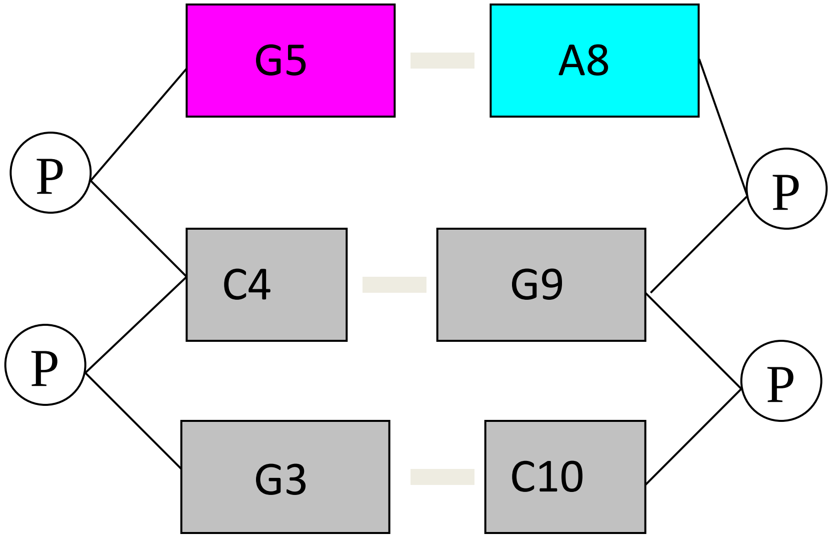


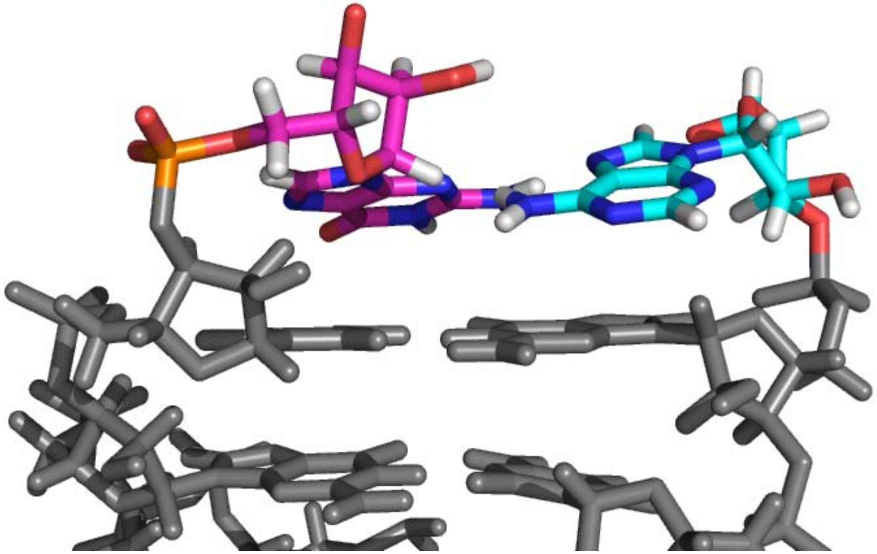
Step-by-step sampling



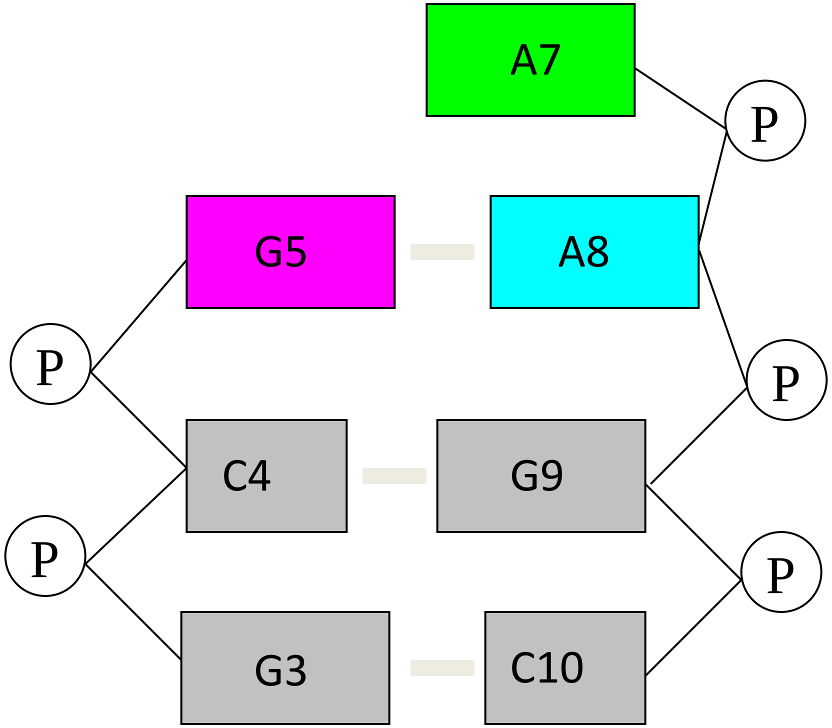


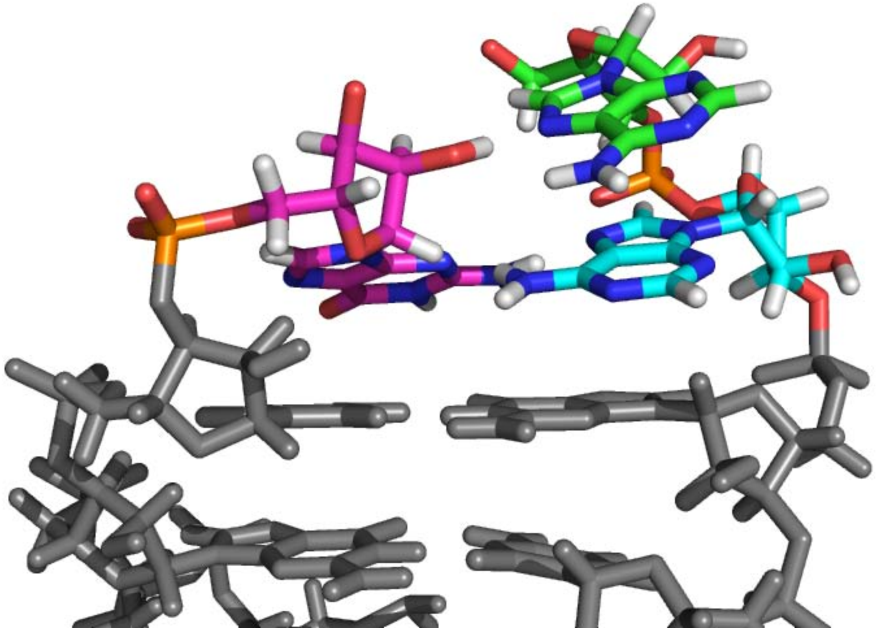
Step-by-step sampling



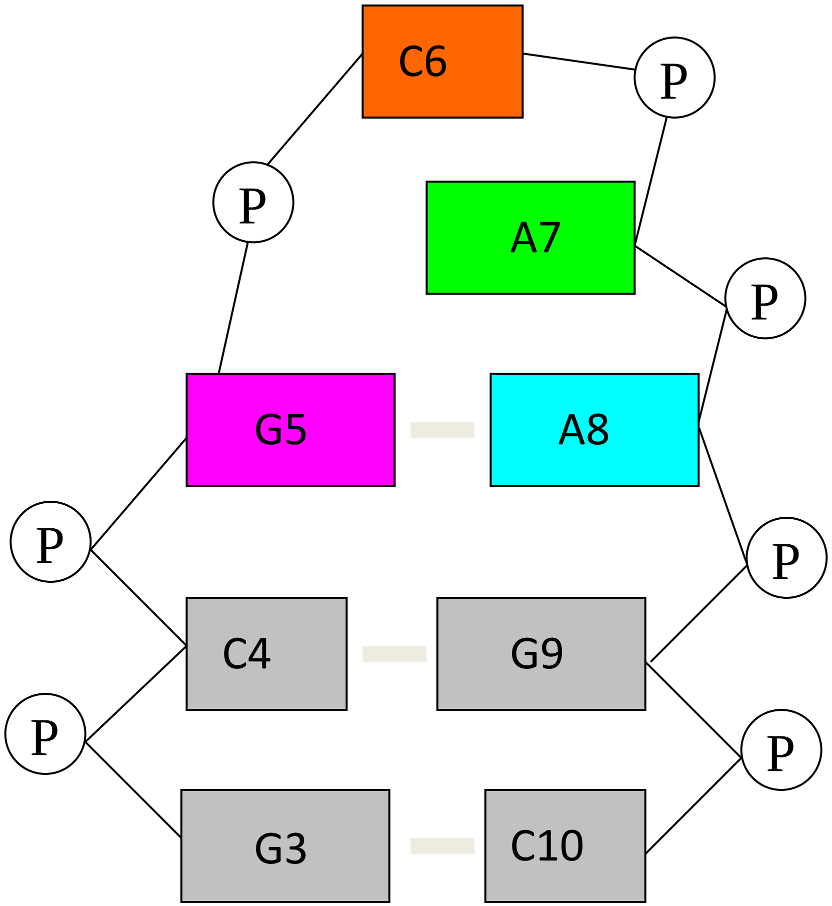


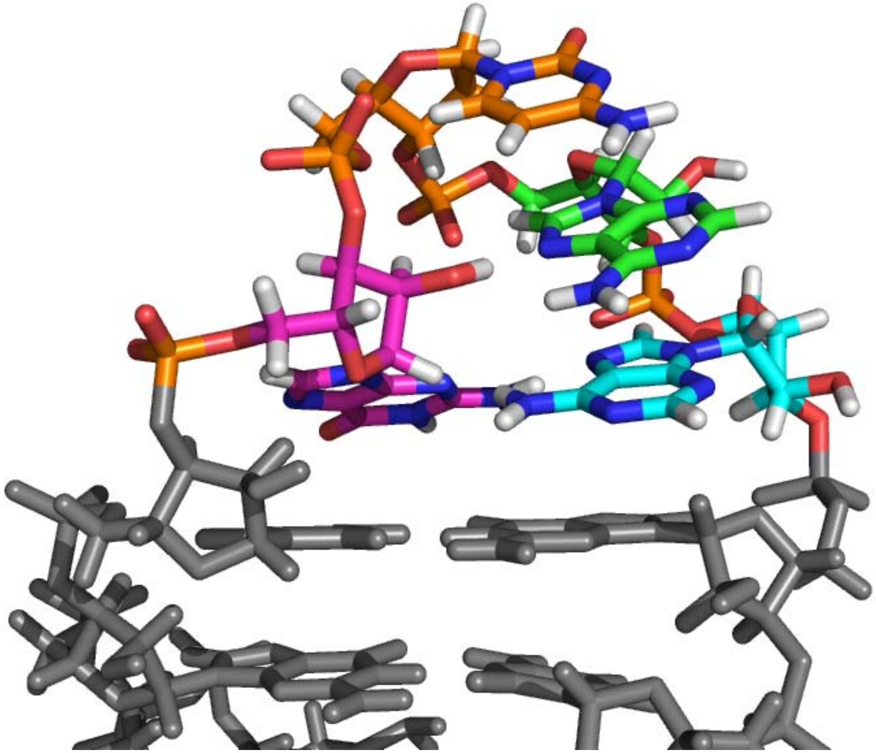
Step-by-step sampling



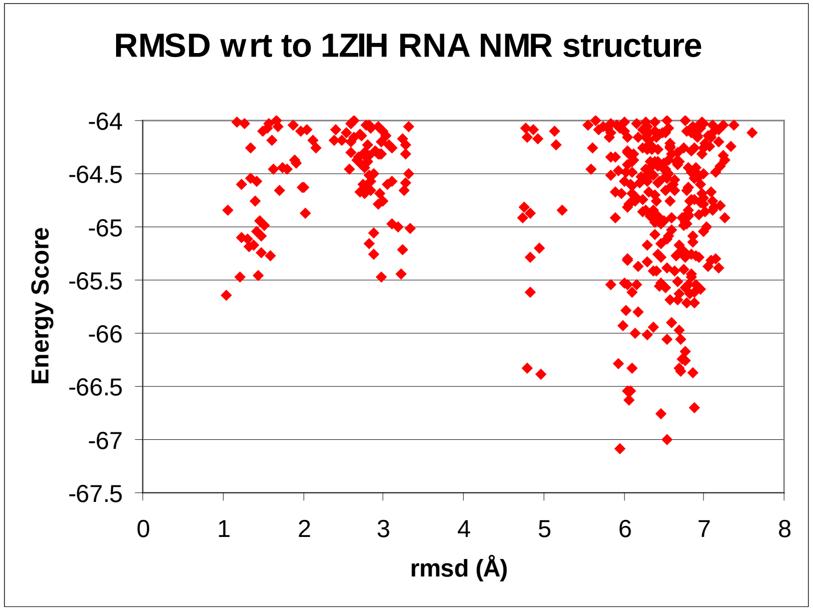


Step-by-step sampling

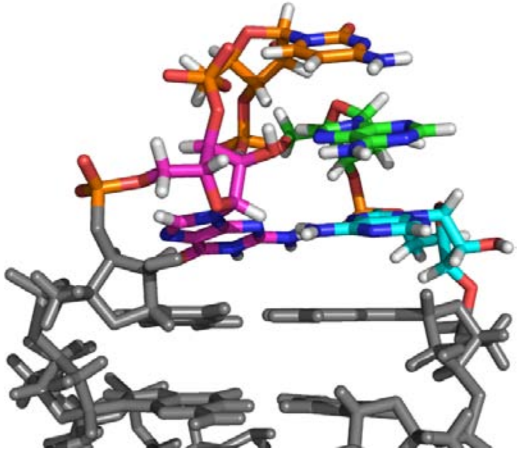




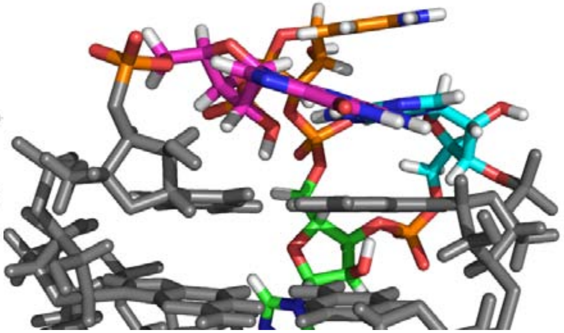
Step-by-step sampling



1ZIH NMR



Lowest Energy

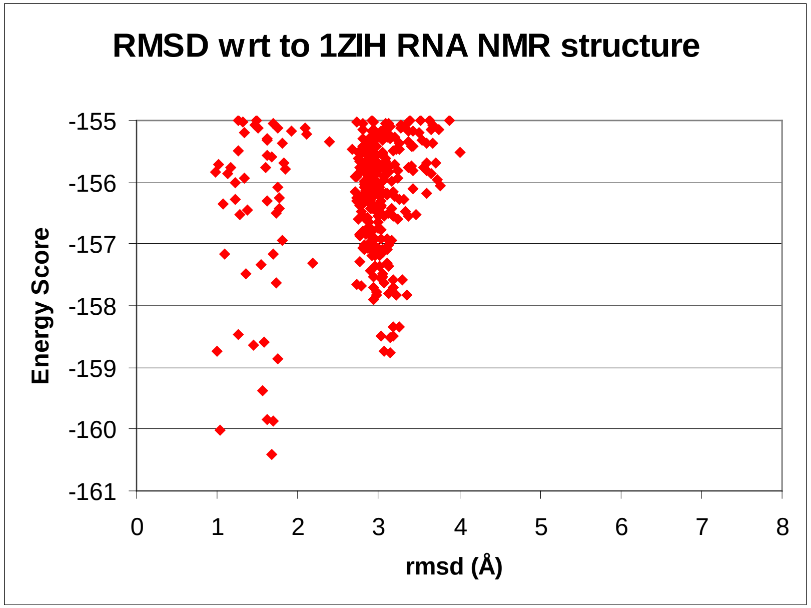


Aha – terms for:

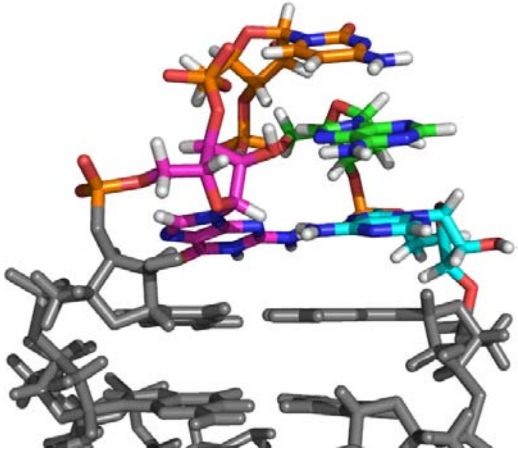
- base stacking
- RNA torsional potential

Had been dialed down to zero. (*A legacy of fragment assembly*)

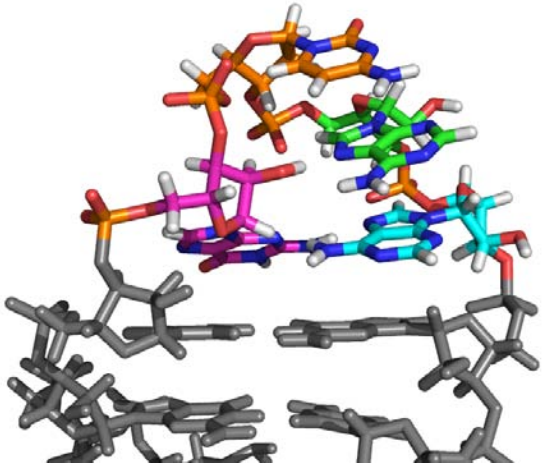
Step-by-step sampling



1ZIH NMR



Lowest Energy

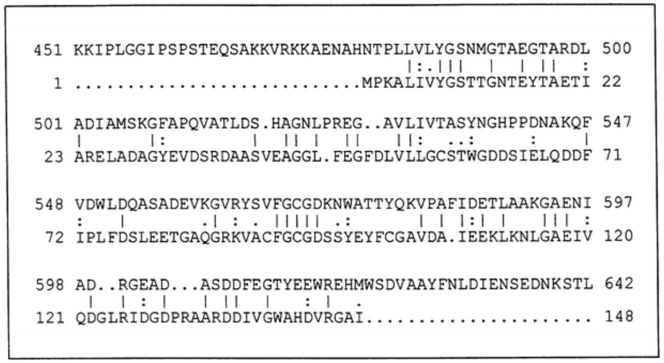


Wait, there’s still a cheat!

There are other pathways (2^N total)

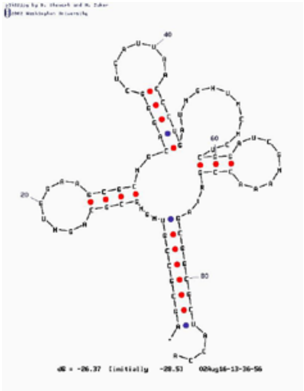
How to sample all paths?

Sequence alignment

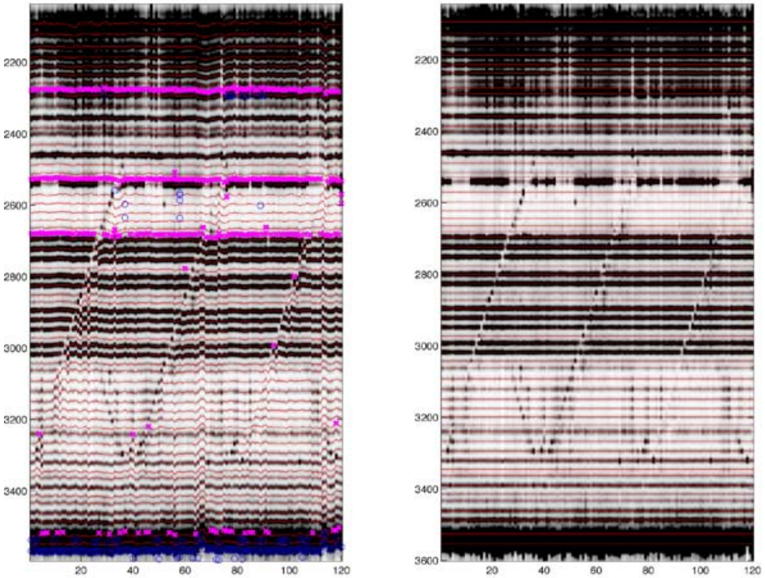


```
451 KKIPLGGIPSPSTEQSAKKVRKKAENAHNTPLLVLYGSNMGTAEGTARDL 500
                                    |:.|||  |  | ||  :
  1 ............................MPKALIVYGSTTGNTEYTAETI 22

501 ADIAMSKGFAPQVATLDS.HAGNLPREG..AVLIVTASYNGHPPDNAKQF 547
    |      |:        |  || |  ||   ||:  ..:     :    |
 23 ARELADAGYEVDSRDAASVEAGGL.FEGFDLVLLGCSTWGDDSIELQDDF 71

548 VDWLDQASADEVKGVRYSVFGCGDKNWATTYQKVPAFIDETLAAKGAENI 597
    :   |       .| : . ||||| .:      | | |:| |   ||| :
 72 IPLFDSLEETGAQGRKVACFGCGDSSYEYFCGAVDA.IEEKLKNLGAEIV 120

598 AD..RGEAD...ASDDFEGTYEEWREHMWSDVAAYFNLDIENSEDNKSTL 642
     |  | : |   |  ||  |   : |  .
121 QDGLRIDGDPRAARDDIVGWAHDVRGAI...................... 148
```

Nucleic acid 2° structure



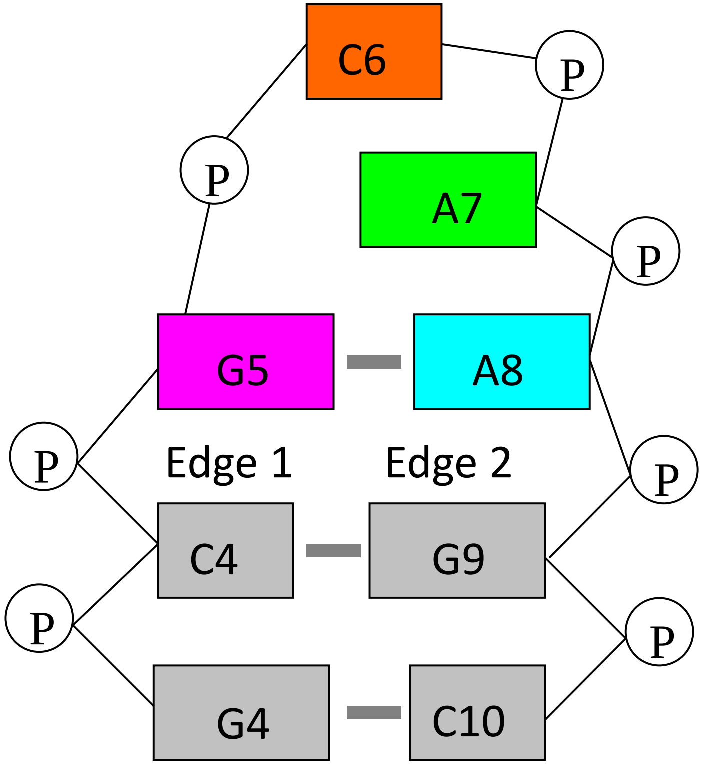
Electrophoretic trace alignment

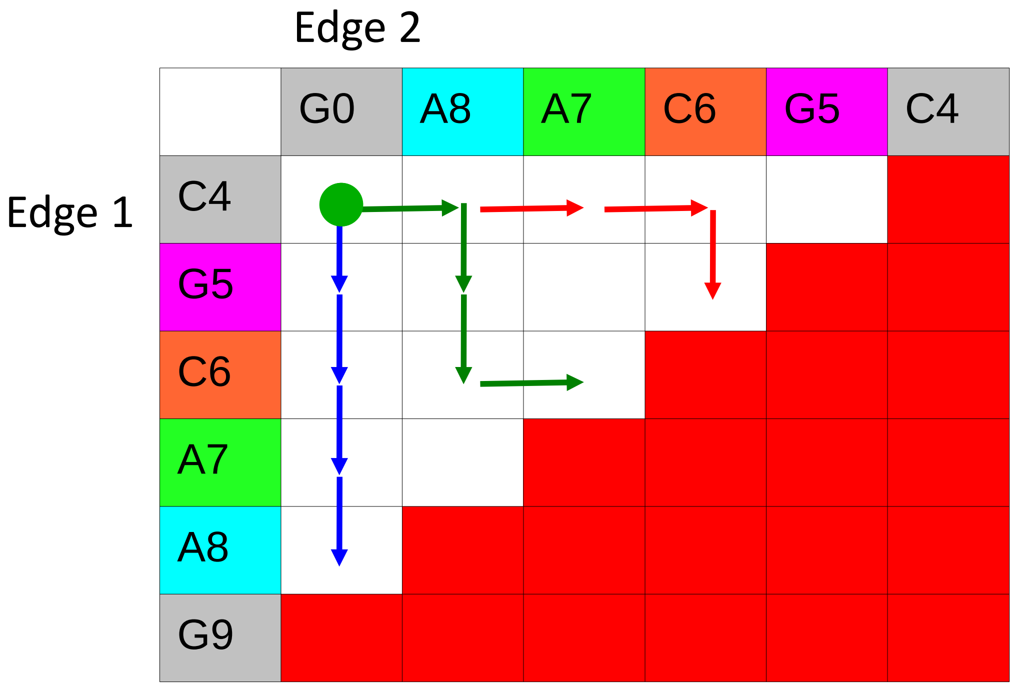


Ordering primers for PCR assembly for the least $$$.

```
TTCTAATACGACTCACTATAGGCCAAAACAACGGAATTGCGGGAAAGGGGTCAACAGCCG->1
                                       ||||||||||||||||||||| 71.8
                                    2<-GCCCTTTCCCCAGTTGTCGGCAAGTCATGGTTCAGAGTCCCCTTTGAAACTCTACCG
                                                                              |||||||||||||||||| 58.1
                                                                              GGGAAACTTTGAGATGGCCTTGCAAAGGGTATGGTAATAAGCTGACGGACATG->3
                                                                                                               |||||||||||||||||||| 58.3
                                                                                                            4<-CATTATTCGACTGCCTGTACCAGGATTGGTGCGTCGGTT
                                                                                                                                               |||||||
                                                                                                                                               CAGCCAA


CAGGATTCAGTTG
||||||||||||| 60.5
GTCCTAAGTCAACAGATCTTCTGTTGATATGGATGC->5
              |||||||||||||||||||||| 58.5
           6<-CTAGAAGACAACTATACCTACGTCAAGTTTTGGTTTGGTTTCTTTGTTGTTGTTGTTG
```

Dynamic programming: all pathways

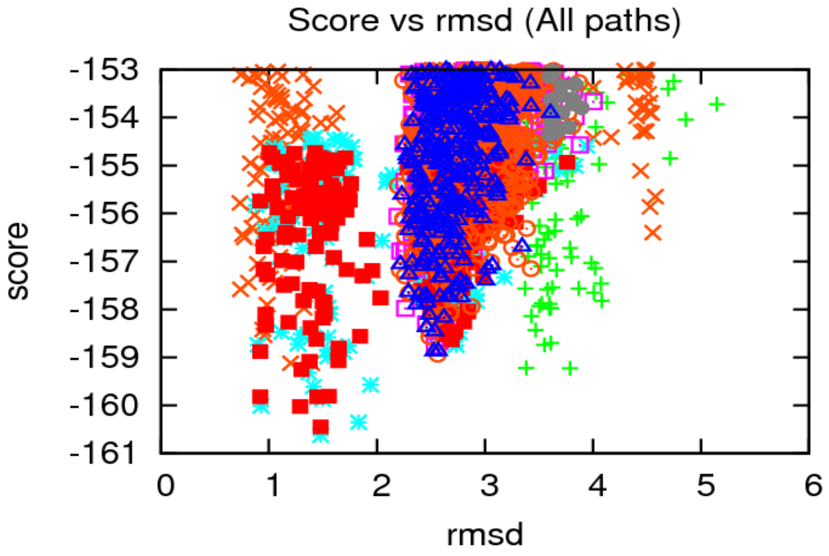


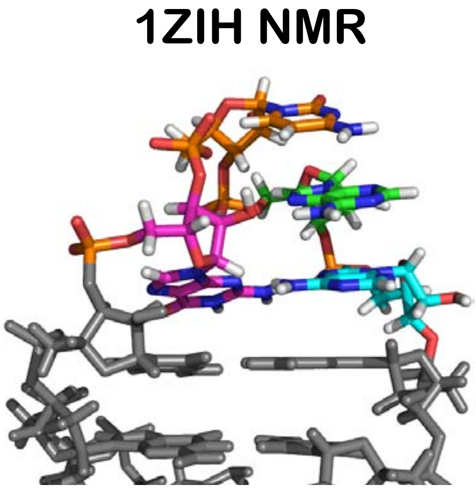

Edge 2

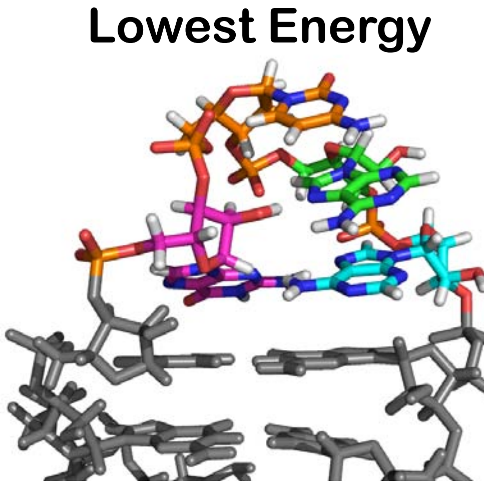
Edge 1

	G0	A8	A7	C6	G5	C4
C4	●					🟥
G5					🟥	🟥
C6				🟥	🟥	🟥
A7			🟥	🟥	🟥	🟥
A8		🟥	🟥	🟥	🟥	🟥
G9	🟥	🟥	🟥	🟥	🟥	🟥

Dynamic programming: all pathways

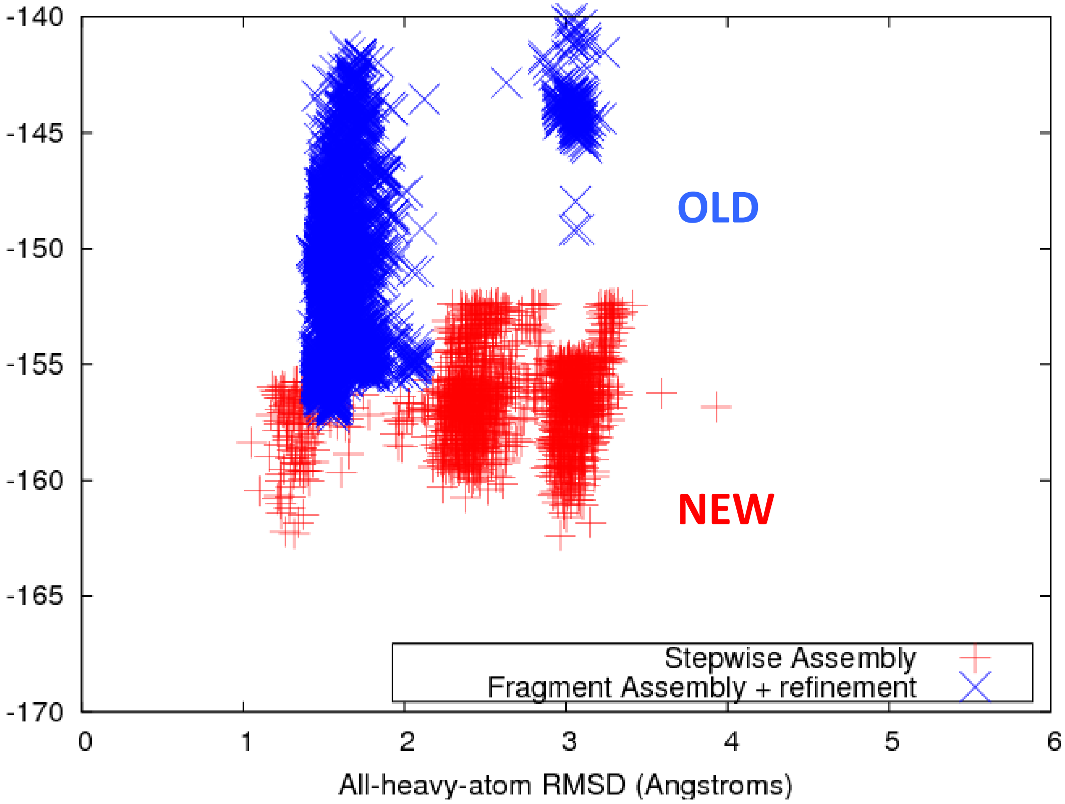






Each point style represents a rebuild path

What have we gained?



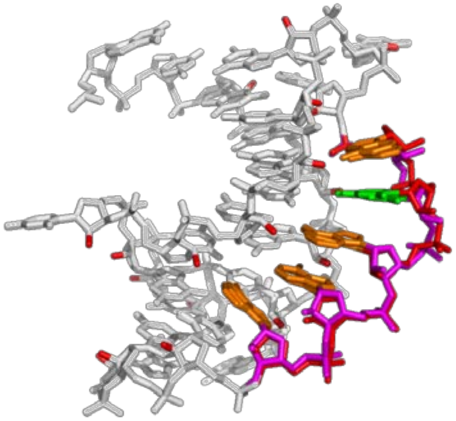
1. Does not use pieces of existing structures
2. Enumerative [$O(N^2)$]
3. Directly searches the all-atom representation.

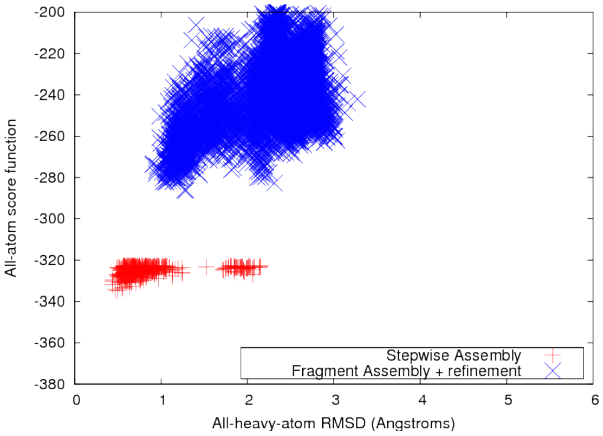
But we only search conformations reachable in a stepwise manner – this is the *Ansatz*.

Overall results

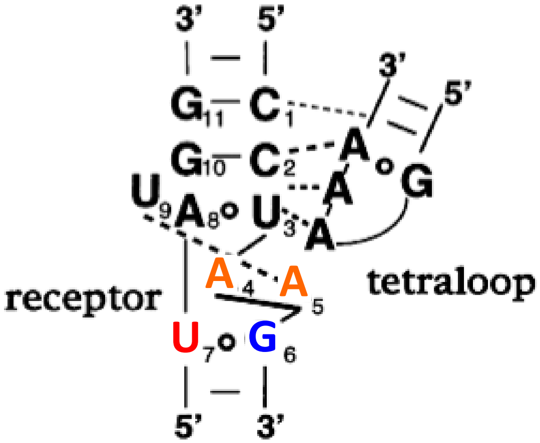
PDB	Length (# non-canonical nucleotides)	Motif Description	All-atom rmsd wrt to exp. structure (Å)*	
			Best RMSD Model	Lowest Energy Score Model
1ZIH	4	GCAA tetraloop	0.9	1.5
1F7Y	4	UUCG tetraloop	1.0	3.4
2PN3	4	5'UU3'/5'UC3' mismatch in HCV IRES	1.0	1.2
1L2X	7	Loop region of a Viral RNA Pseudoknot	0.7	4.6
2R8S	7	Tetraloop Receptor (build receptor only)	0.9	1.0
1Q9A	9	Bulged G-motif from the sarcin/ricin loop	1.1	5.3
1LNT	10	Highly Conserved Internal Loop of SRP RNA	1.2	1.7
354D	10	Purine rich region in the 5S rRNA Loop E motif	0.8	1.1

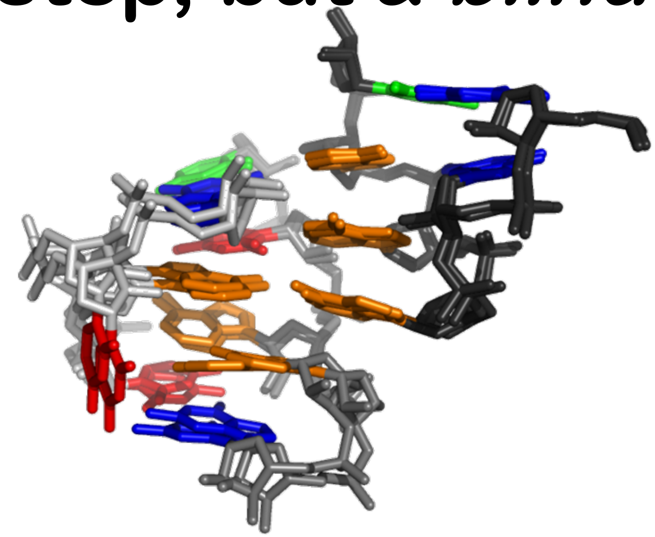
*All-atom RMSD, excluding bulge nucleotides



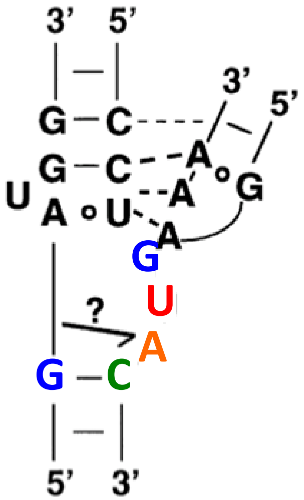


A baby step, but a *blind* one.

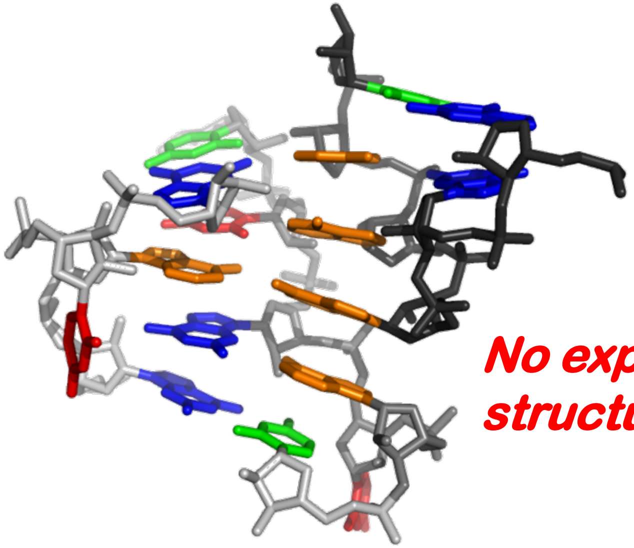




Just rebuilding the colored residues



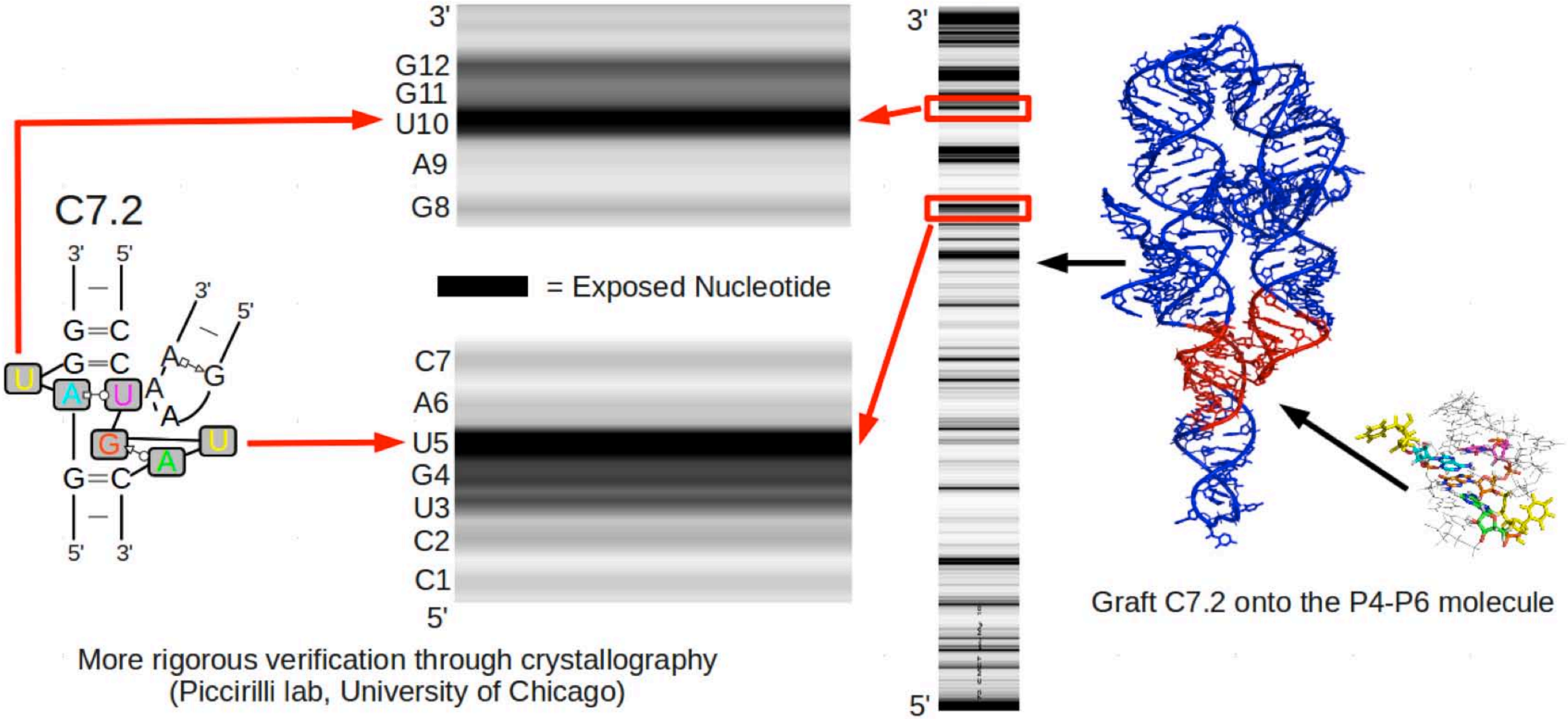
C7.2



No experimental structure yet.

Initial validation

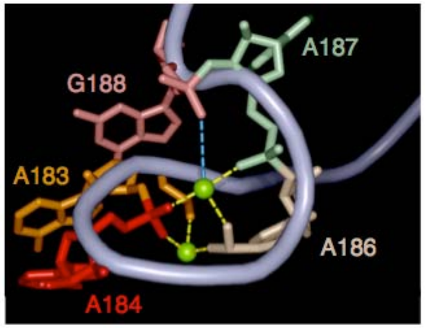
CMCT chemical accessibility data (G,U)



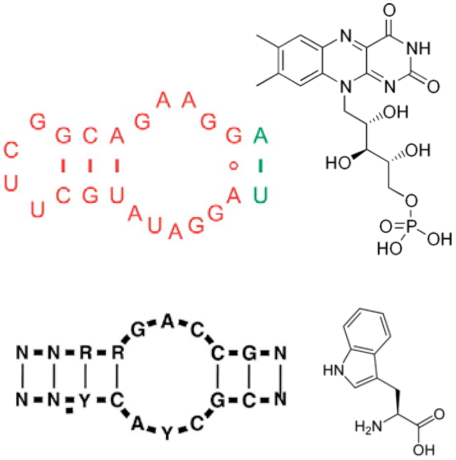
Graft C7.2 onto the P4-P6 molecule

More rigorous verification through crystallography
(Piccirilli lab, University of Chicago)

A stepwise enumerative ansatz: next.

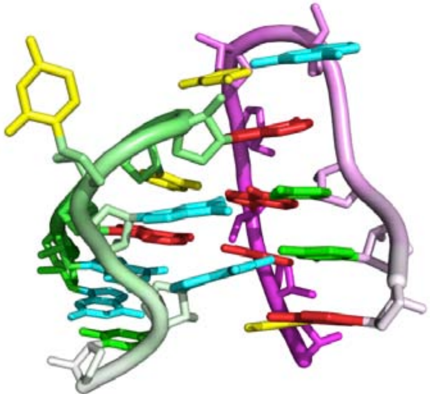


Metal ions, solvation, all that – fixing the energy function



A plethora of RNA aptamers.

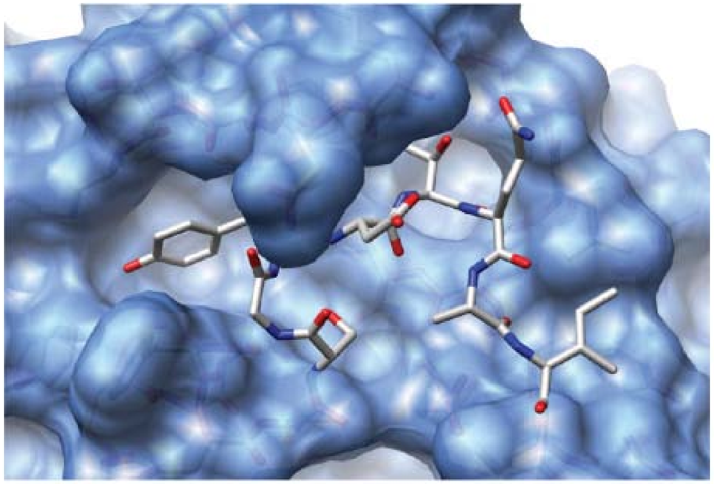
What about proteins?



O(N^4)

More complex motifs/RNAs

Small protein puzzles

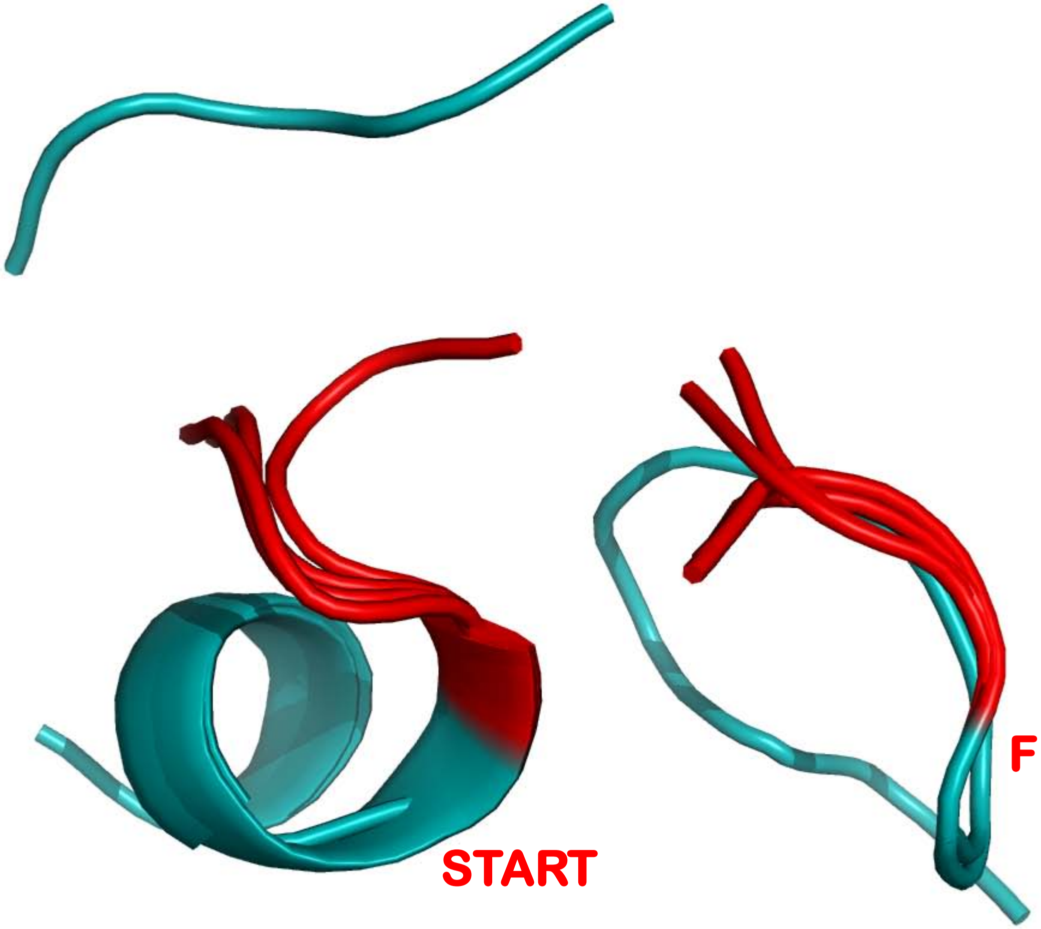


Sellers, Zhu, Zhao, Friesner, & Jacobson 2008.

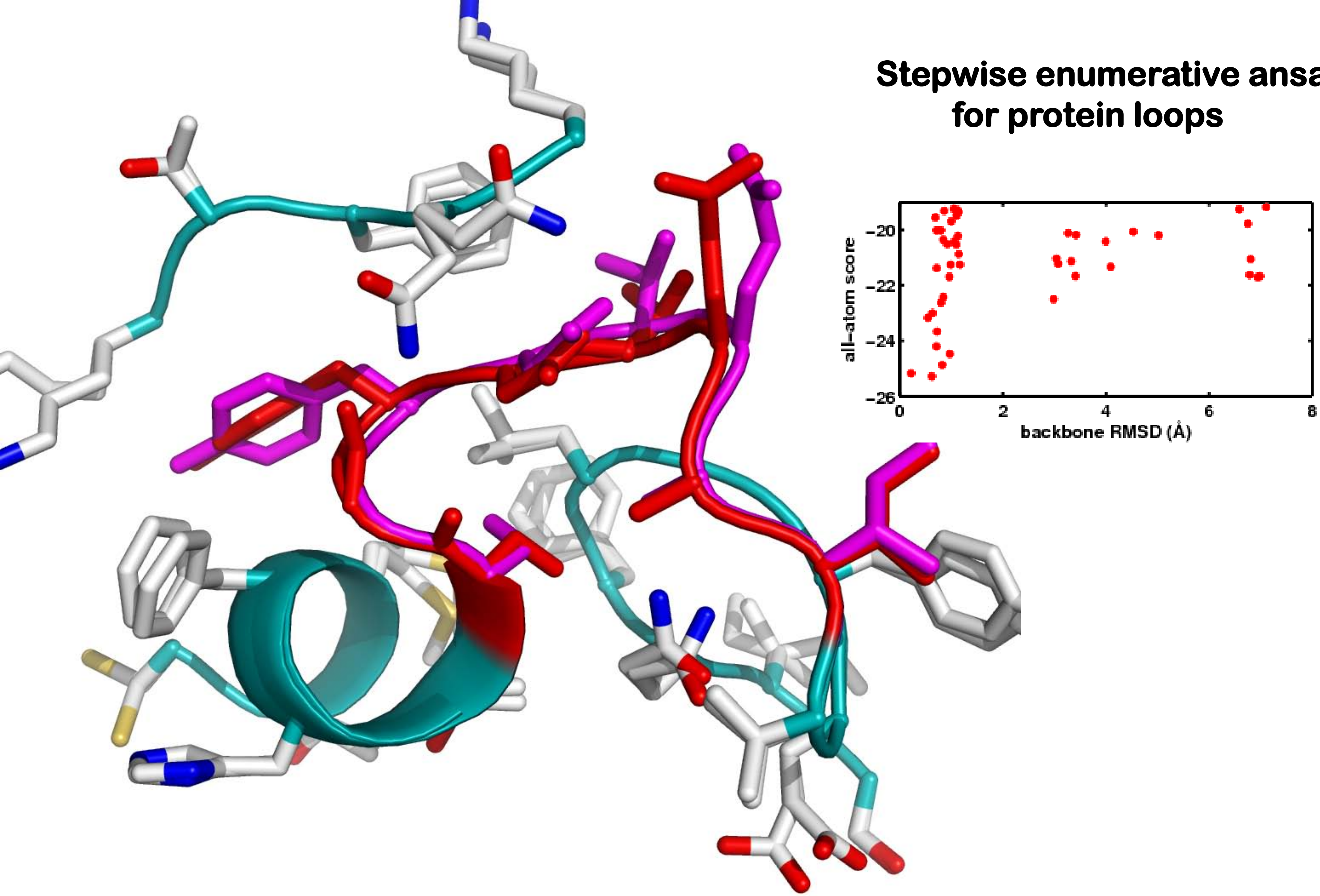
1ALC 34–41

See also: Rosetta fragment-based modeling (Rohl), with CCD (Wang), Monte Carlo Minimization with kinematic loop closure (Mandell et al.)

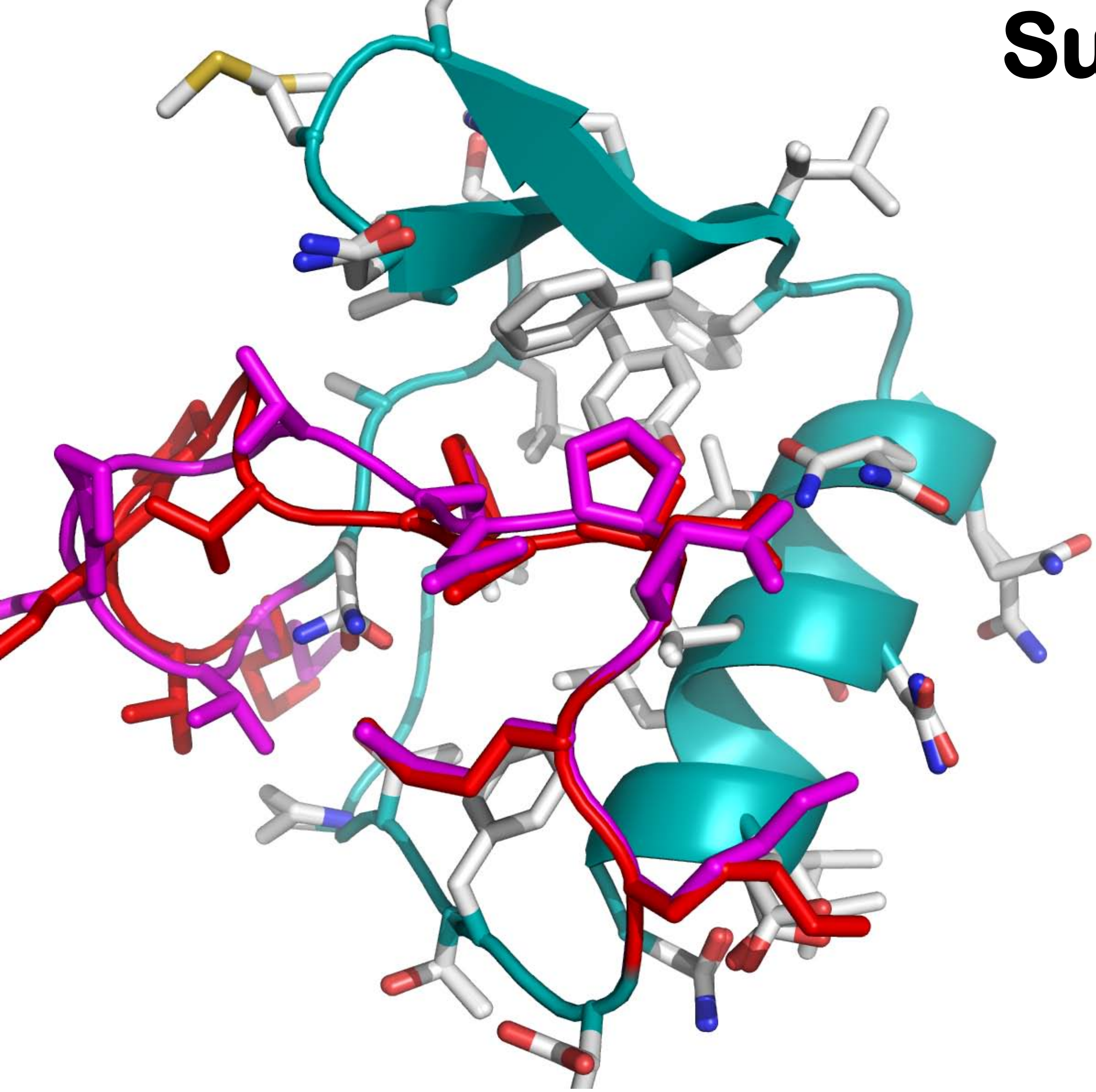
Stepwise enumerative ansatz for protein loops



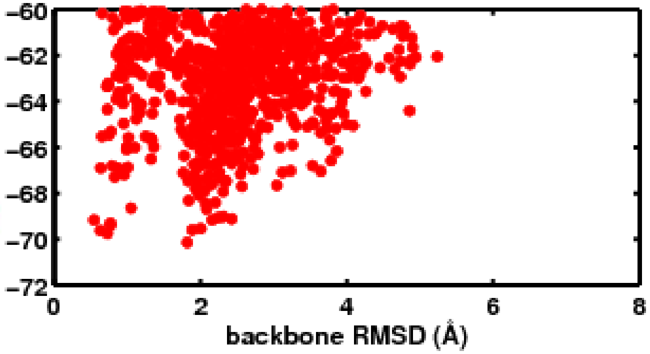
Stepwise enumerative ansatz for protein loops

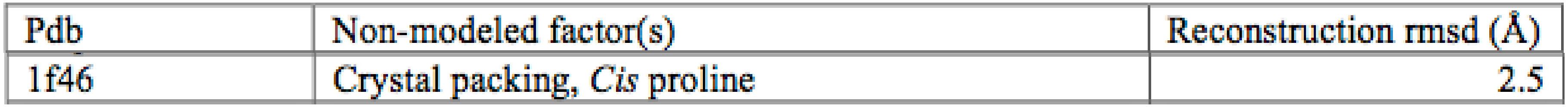


Success on prior hard cases



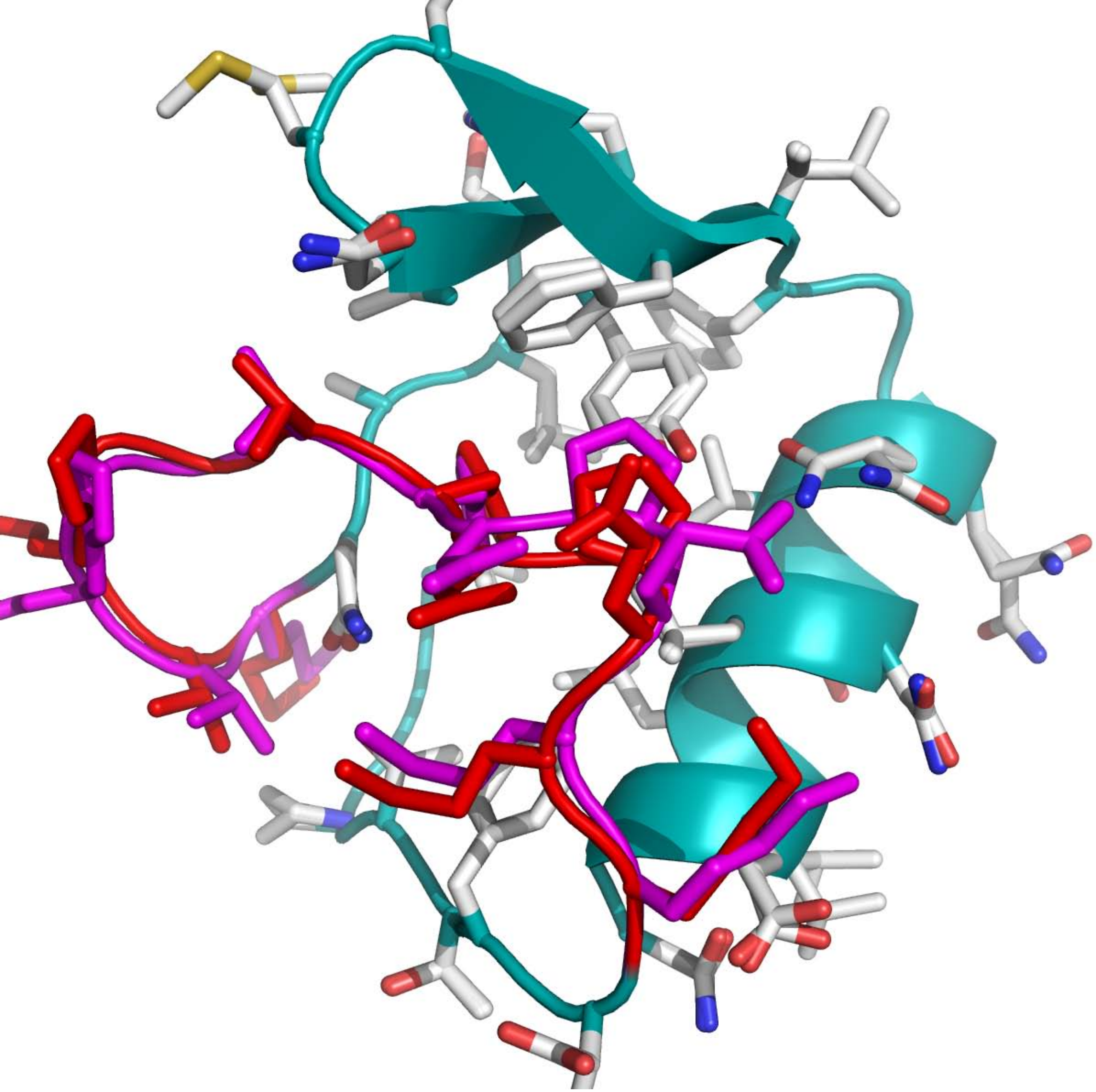
1F46 64–75





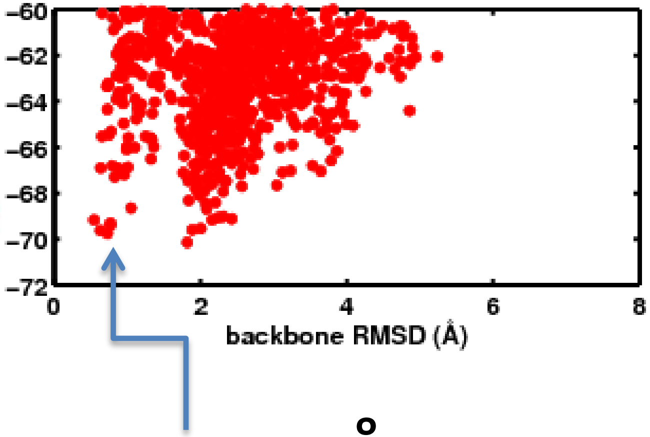
Pdb	Non-modeled factor(s)	Reconstruction rmsd (Å)
1f46	Crystal packing, *Cis* proline	2.5

Mandell, Coutsias, Kortemme 2009



ccess on prior hard cases

1F46 64–75



0.6–0.7 Å

Pdb	Non-modeled factor(s)	Reconstruction rmsd (Å)
1f46	Crystal packing, *Cis* proline	2.5

Mandell, Coutsias, Kortemme 2009

Loop modeling made easy?

Loop	Accuracy
1ALC 34–41	0.5 Å
1CLC 313–320	0.5 Å
1F46 64–75	0.6, 1.9 Å (equal score)
3TGL 82–87	0.5 Å
2CI2 34–46	1–3 Å
T0308 21–31	1.0 Å
T0308 56–64	0.6 Å
T0308 65–75	0.7 Å
T0308 99–107	1 Å
T0311 38–43	0.3 Å
T0453 32–45	0.5–1.5 Å
T0488 10-17	1 Å

Stepwise enumerative assembly

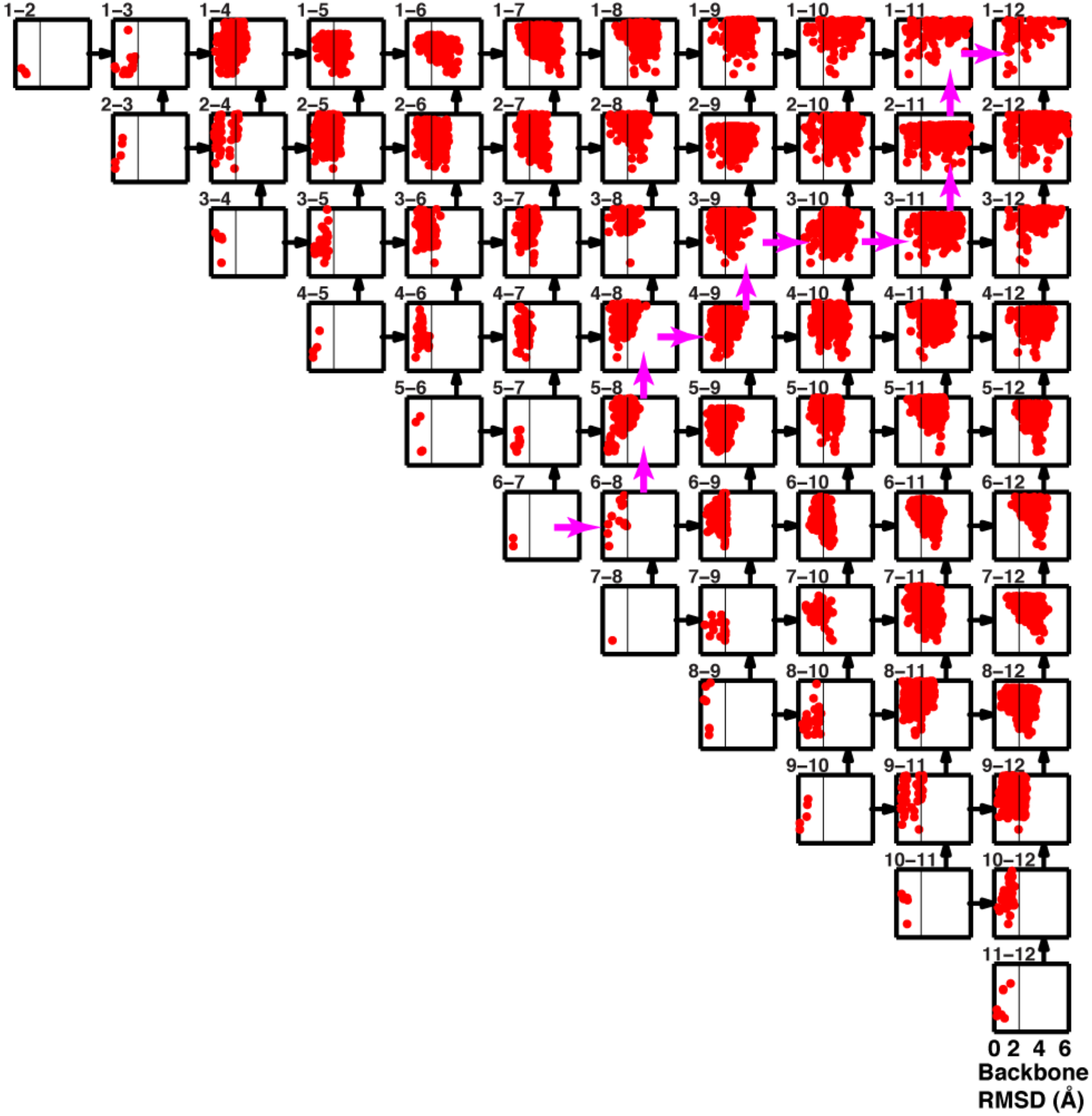
• extremely good at picking up "memory" imprinted outside loop

• extremely sensitive to any *errors*, e.g. as occurs in homology modeling – testing now in CASP9!

• Need "self-contained" de novo tests: mini-proteins?

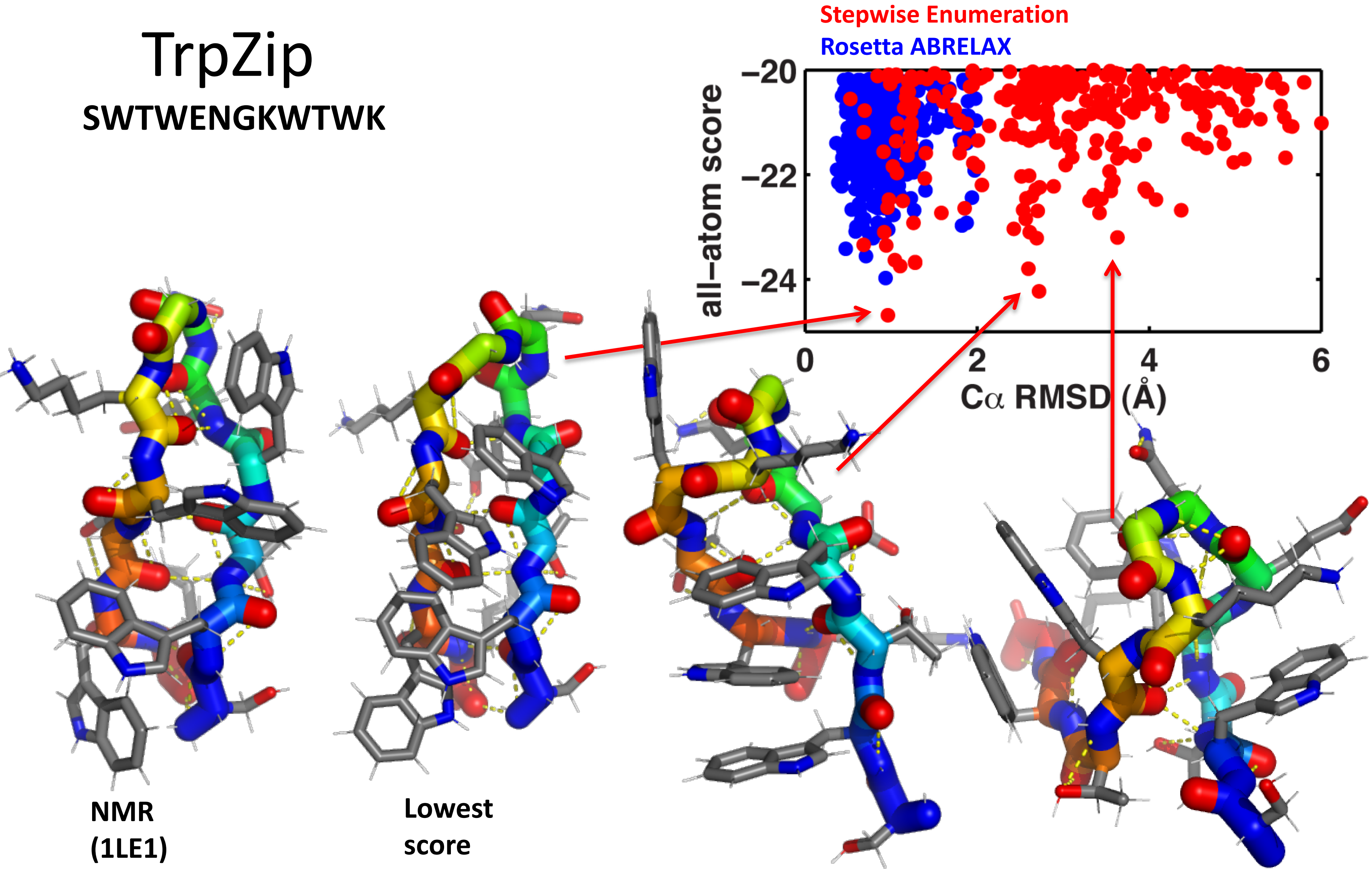
TrpZip

SWTWENGKWTWK



TrpZip

SWTWENGKWTWK

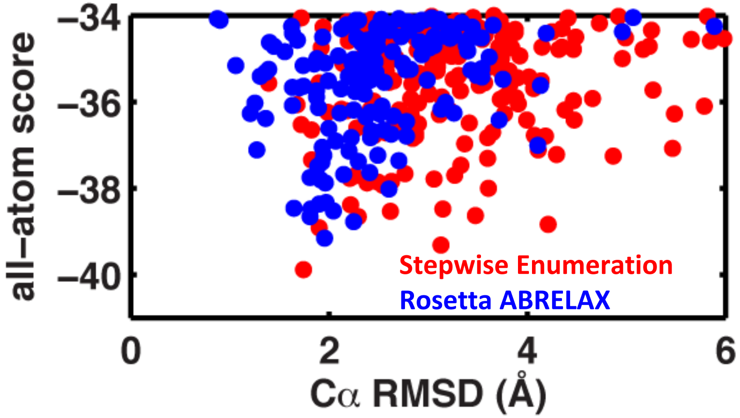


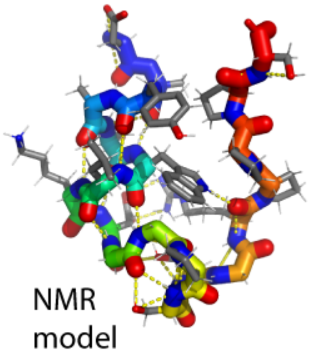
NMR (1LE1)

Lowest score

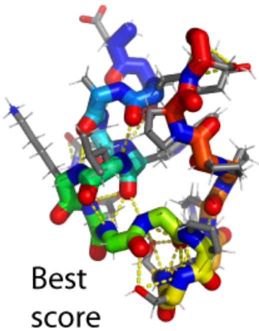
Mini-proteins: discrimination disaster

Trp cage: DAYAQWLKDGGPSSGRPPPS



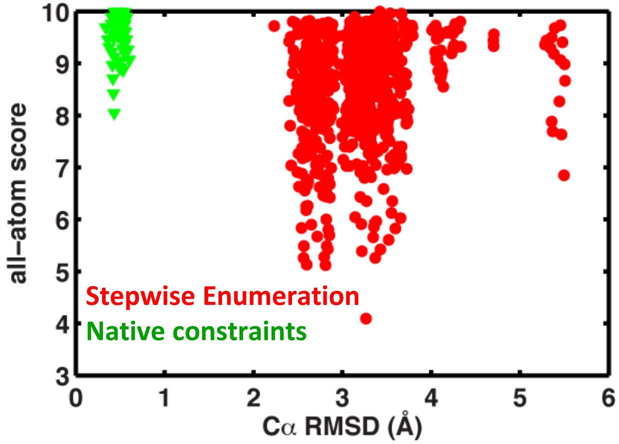


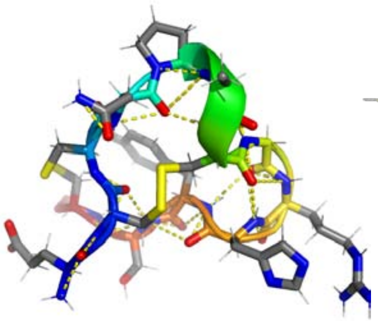
NMR model (2JOF)



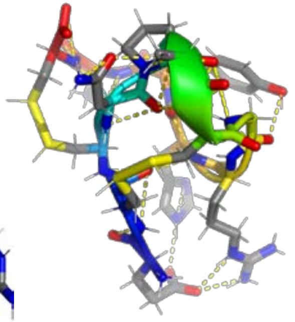
Best score (step-wise assembly)

A marine snail venom toxin: ECCNPACGRHYSC



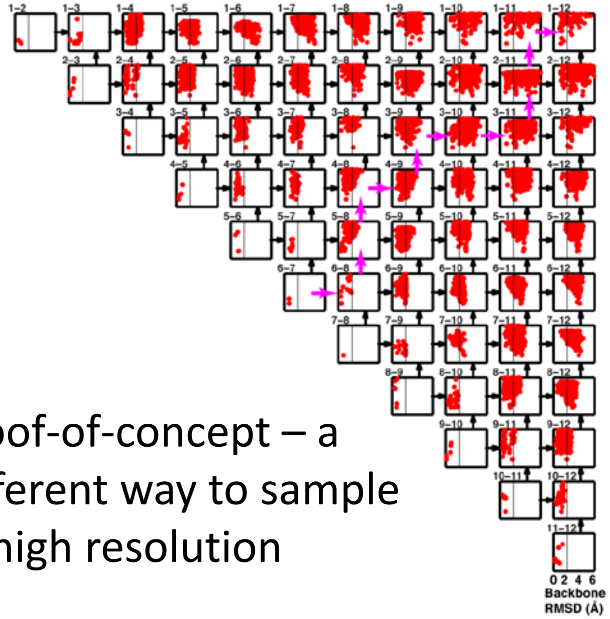


Crystallographic model (1NOT)

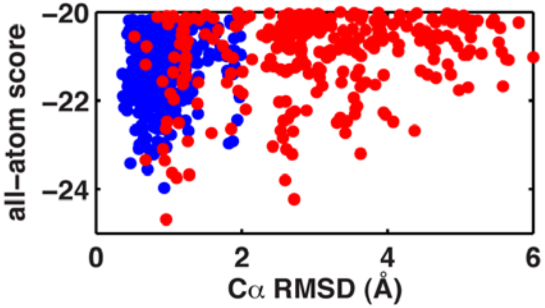


Best score (step-wise assembly)

A stepwise enumerative ansatz for macromolecules

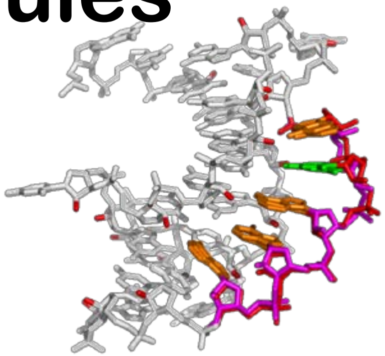


Proof-of-concept – a different way to sample at high resolution

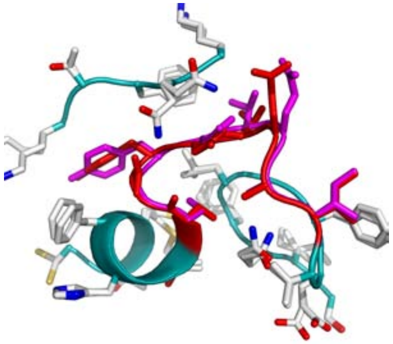


Lower energies & more parts of conformational space than fragment-assembly/refinement

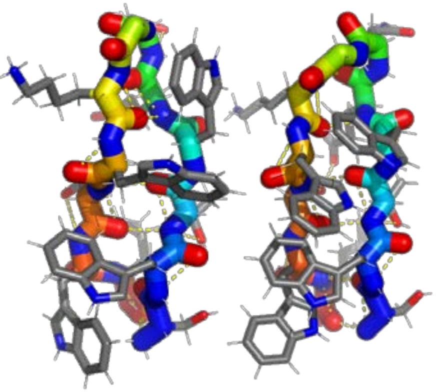
RNA motifs



Protein loops

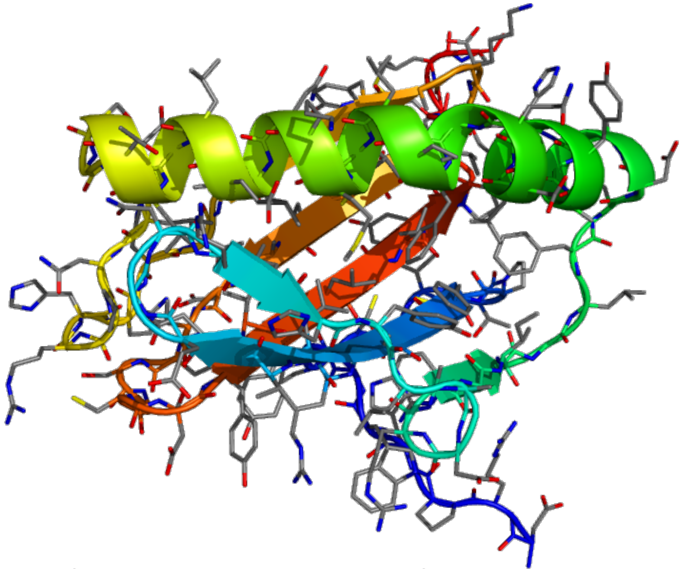


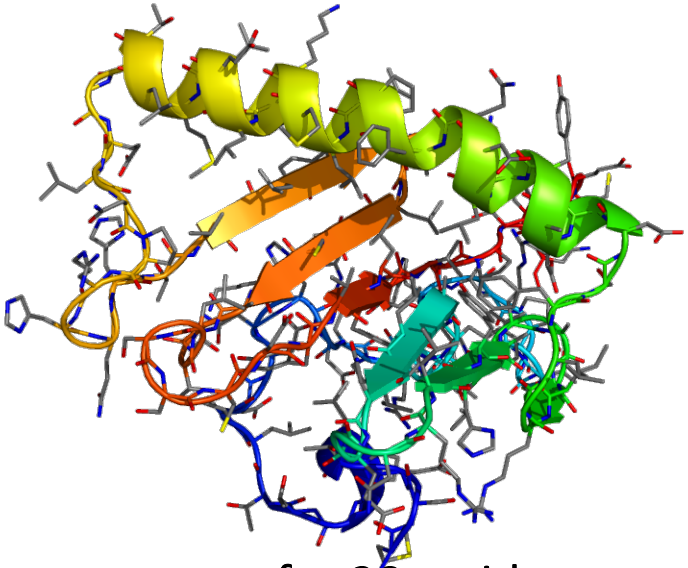
Mini-proteins



Ongoing: blind tests

How about a 150 residue protein?





- Currently, takes 10,000 CPU-hours [400 cores, 1 master, 24 hours] for 20 residues.

- Assuming:

  O(N^2) [no. steps]

  x O(N) [minimize takes longer] x O(N) [more poses],

  150 residue protein will require **100 million CPU-hours**.

- Caveats:

(a) “single-residue steps” may not be appropriate.

(b) No. of poses in “thermal ensemble” may increase with N.

(c) Energy function issues…

Thanks to:

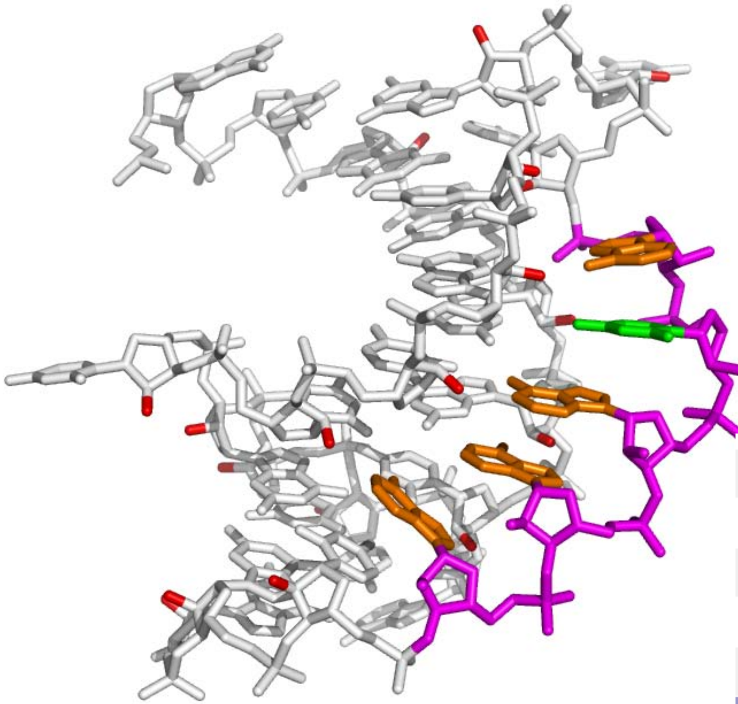
- Parin Sripakdeevong [all the RNA stuff!]

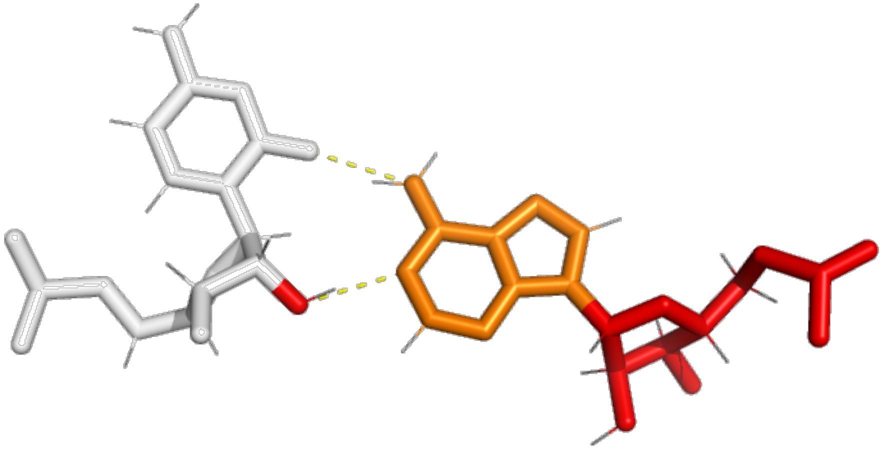
- Ann Kladwang [tetraloop/receptor data]

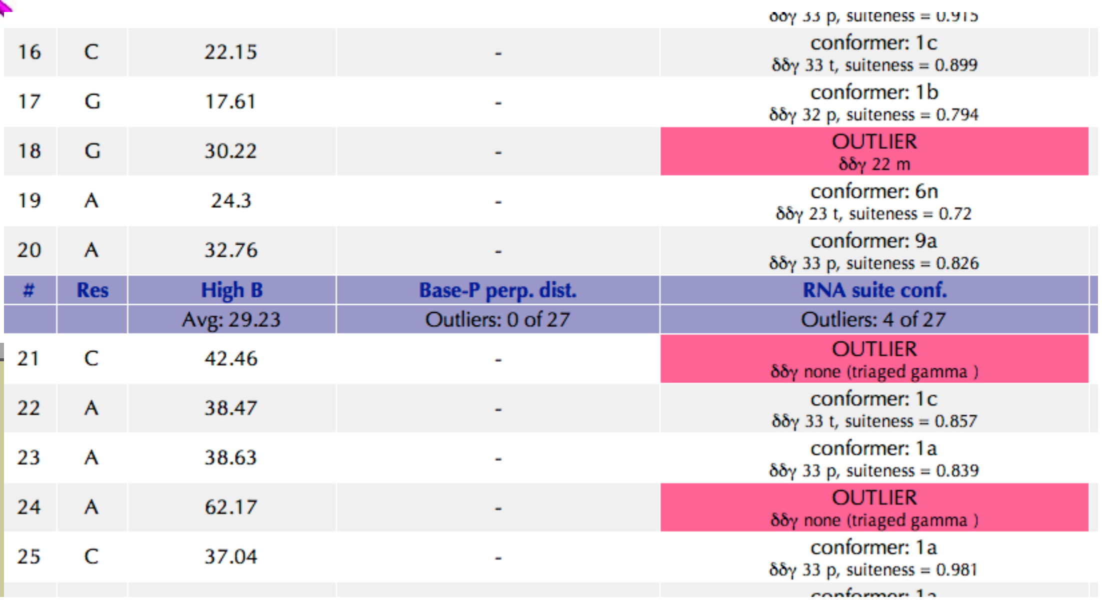
- NSF BioX^2 cluster at Stanford; Burroughs-Wellcome foundation

- Rosetta community

A previously impossible toy problem

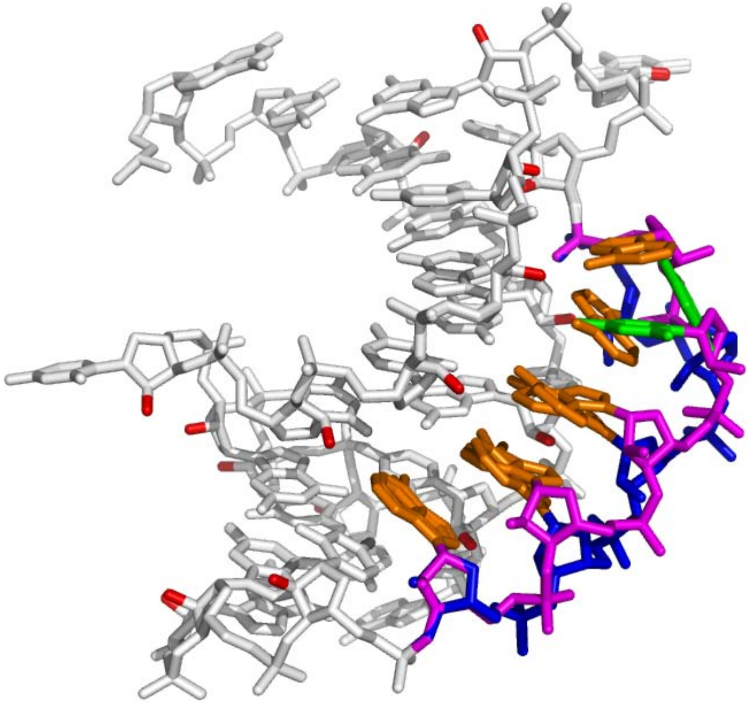


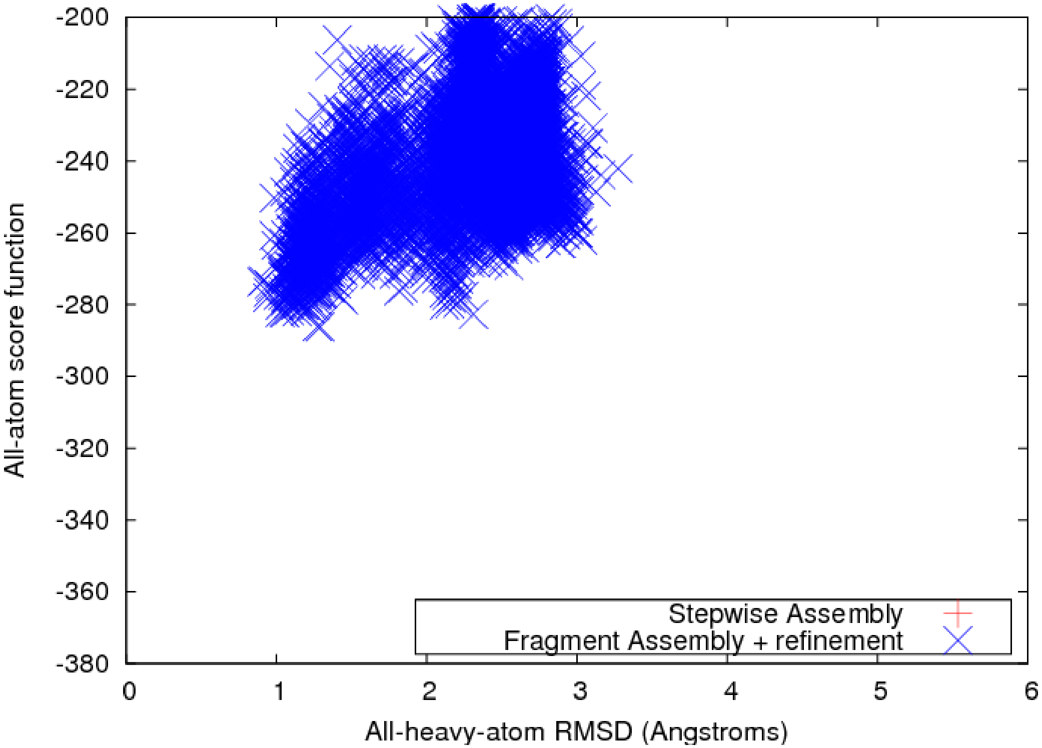




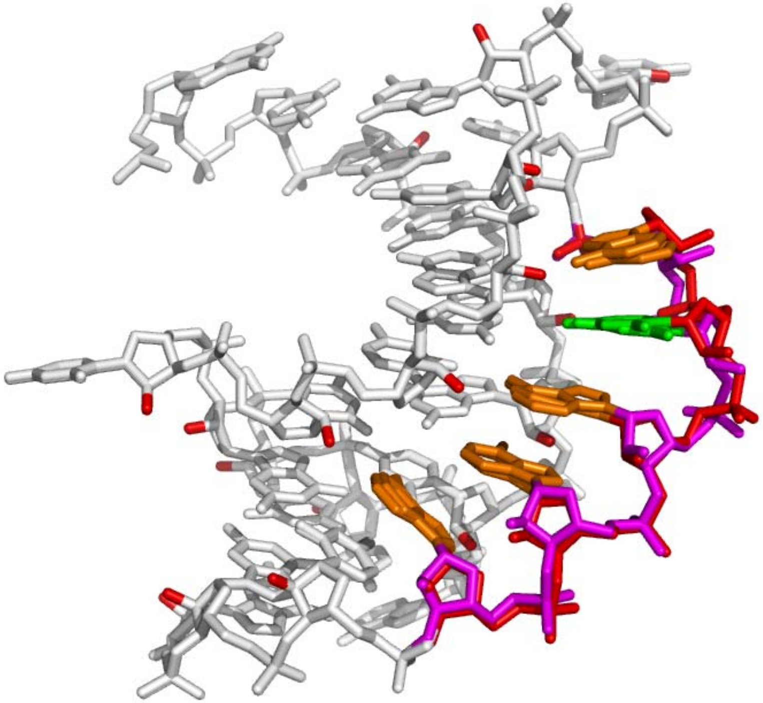
				δδγ 33 p, suiteness = 0.915
16	C	22.15	-	conformer: 1c δδγ 33 t, suiteness = 0.899
17	G	17.61	-	conformer: 1b δδγ 32 p, suiteness = 0.794
18	G	30.22	-	OUTLIER δδγ 22 m
19	A	24.3	-	conformer: 6n δδγ 23 t, suiteness = 0.72
20	A	32.76	-	conformer: 9a δδγ 33 p, suiteness = 0.826
#	Res	High B	Base-P perp. dist.	RNA suite conf.
		Avg: 29.23	Outliers: 0 of 27	Outliers: 4 of 27
21	C	42.46	-	OUTLIER δδγ none (triaged gamma)
22	A	38.47	-	conformer: 1c δδγ 33 t, suiteness = 0.857
23	A	38.63	-	conformer: 1a δδγ 33 p, suiteness = 0.839
24	A	62.17	-	OUTLIER δδγ none (triaged gamma)
25	C	37.04	-	conformer: 1a δδγ 33 p, suiteness = 0.981
				conformer: 1a

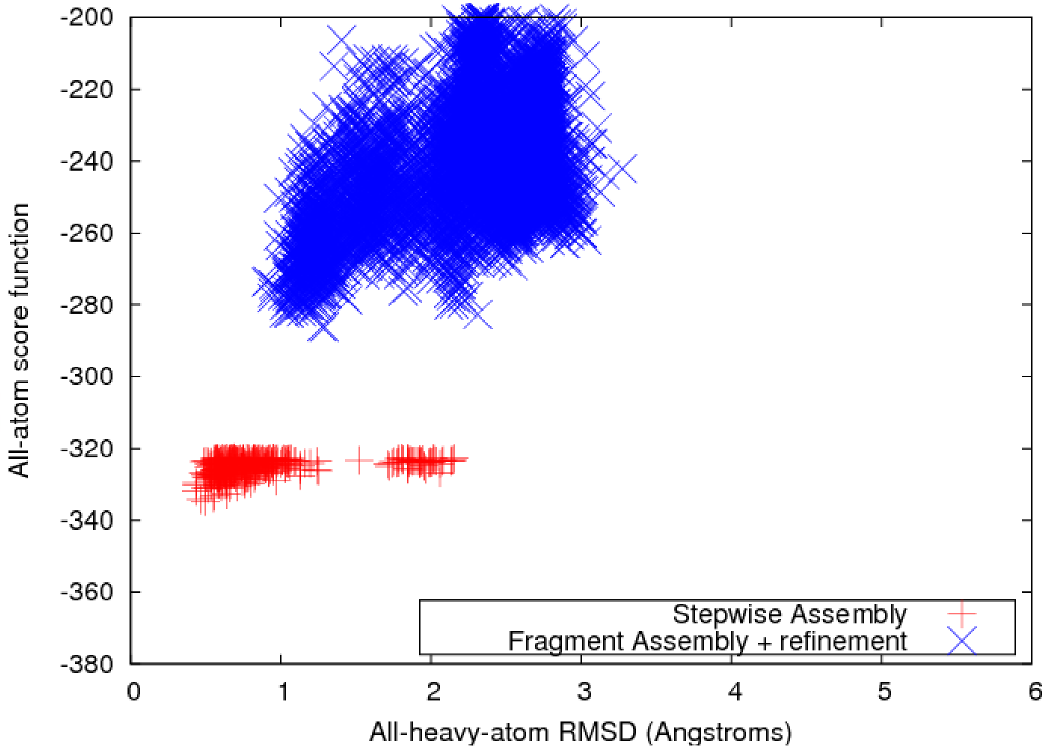
A previously impossible toy problem



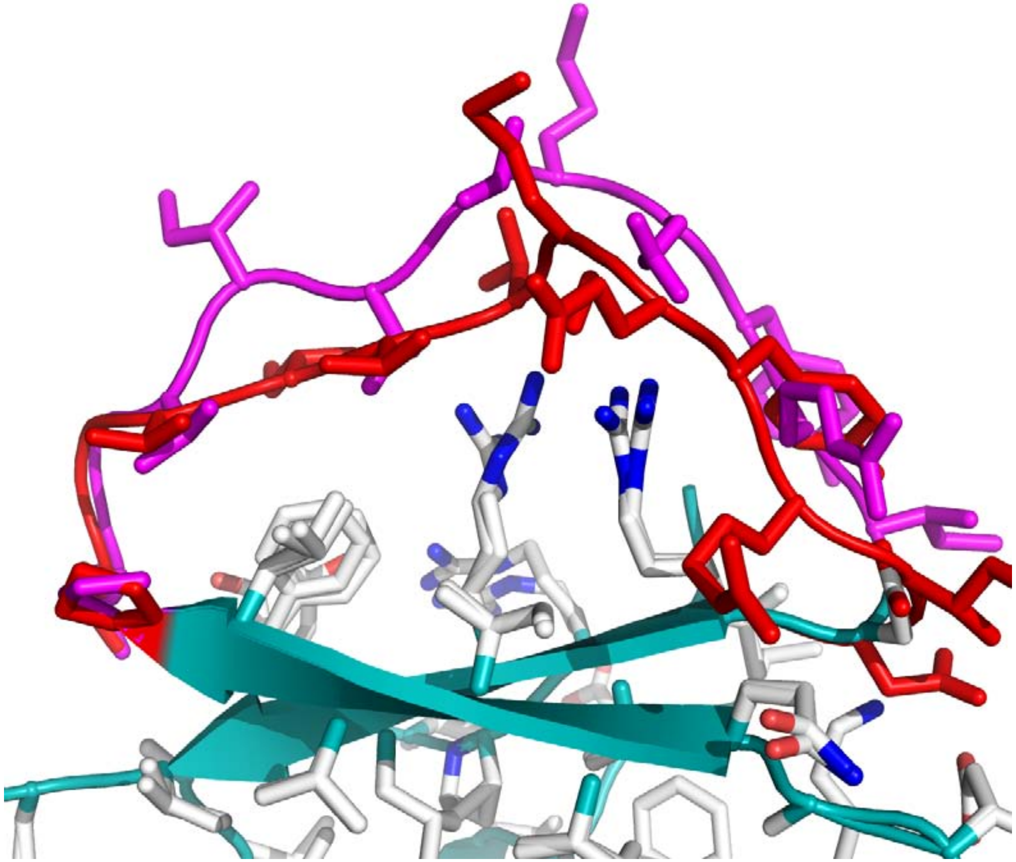


A previously impossible toy problem

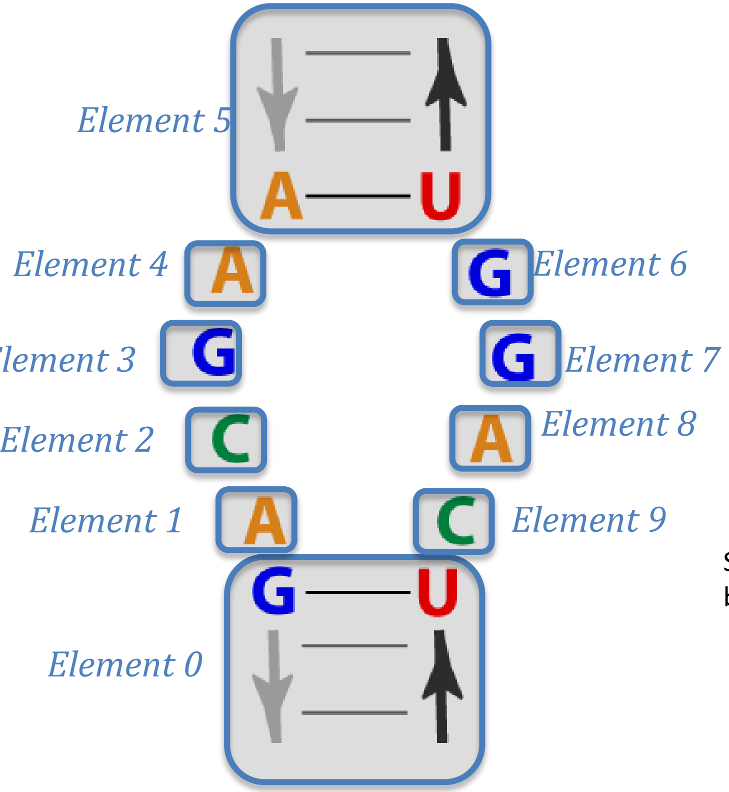


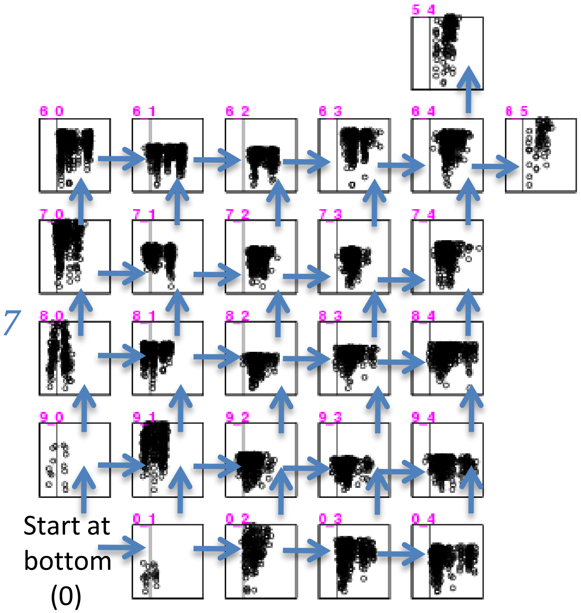


Chymotrypsin inhibitor (2ci2)

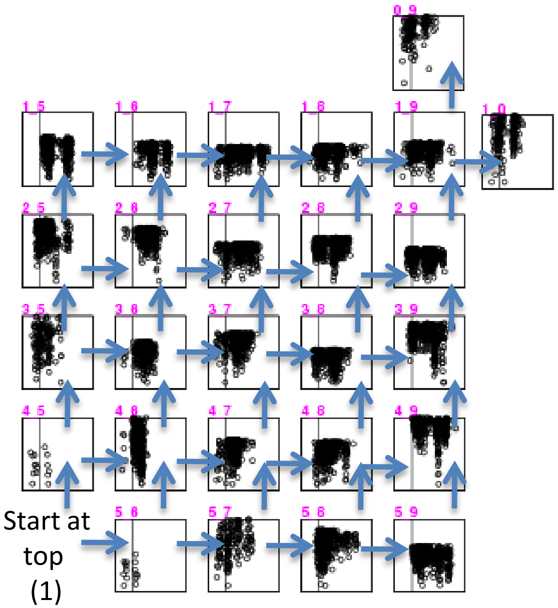


A more complex motif



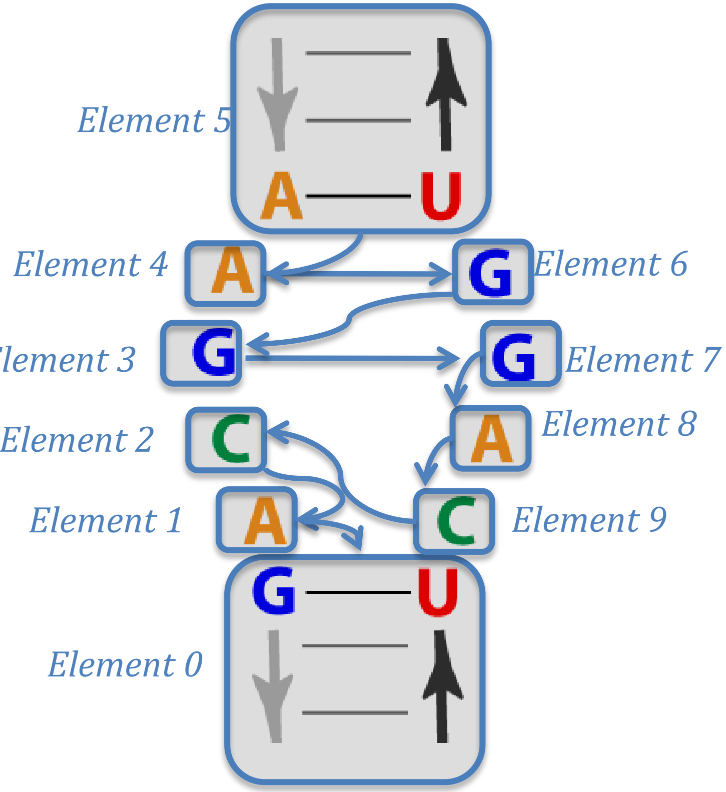


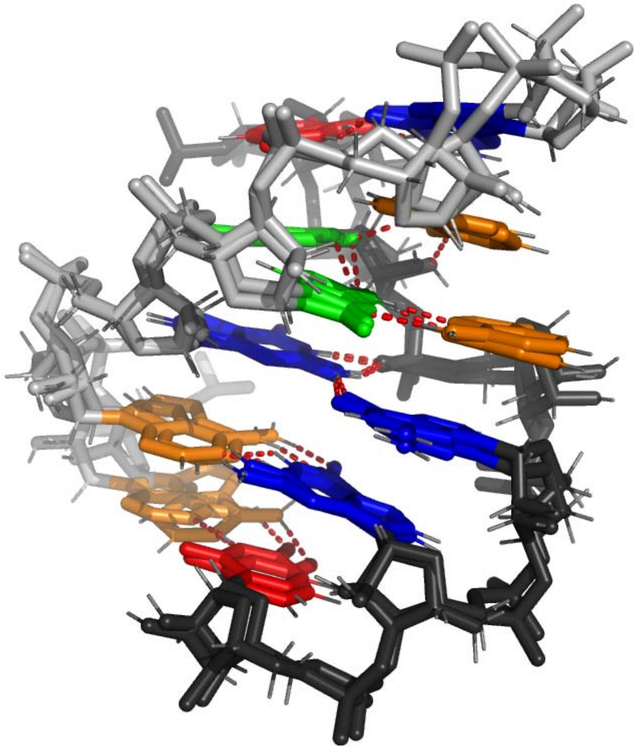
Start at bottom (0)



Start at top (1)

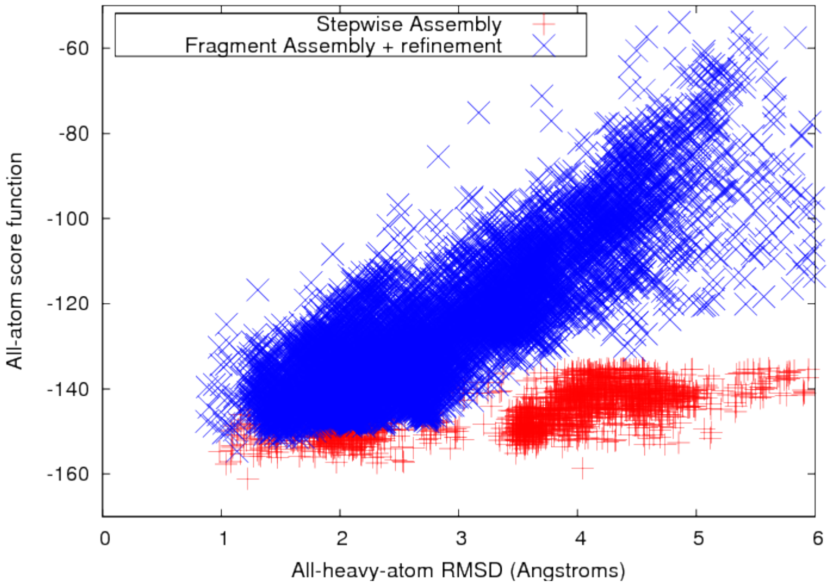
A more complex motif

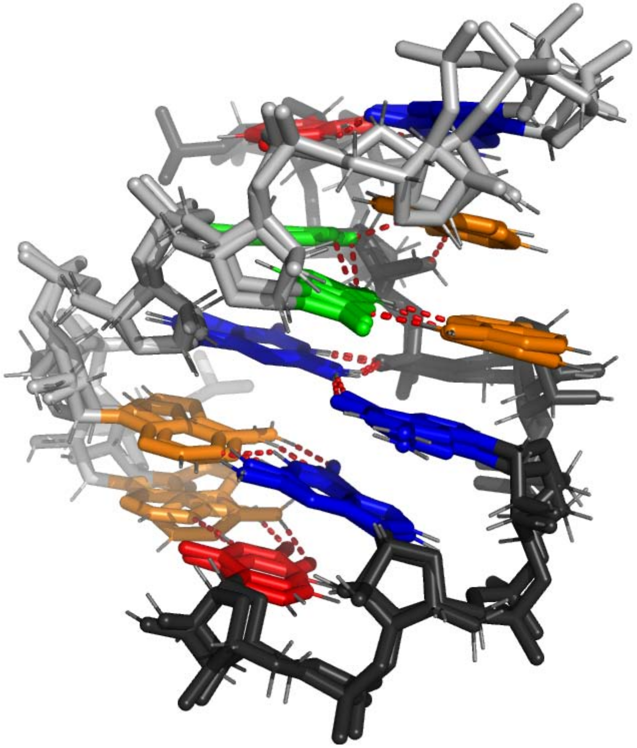




1.09 Å heavy-atom RMSD from crystallographic model

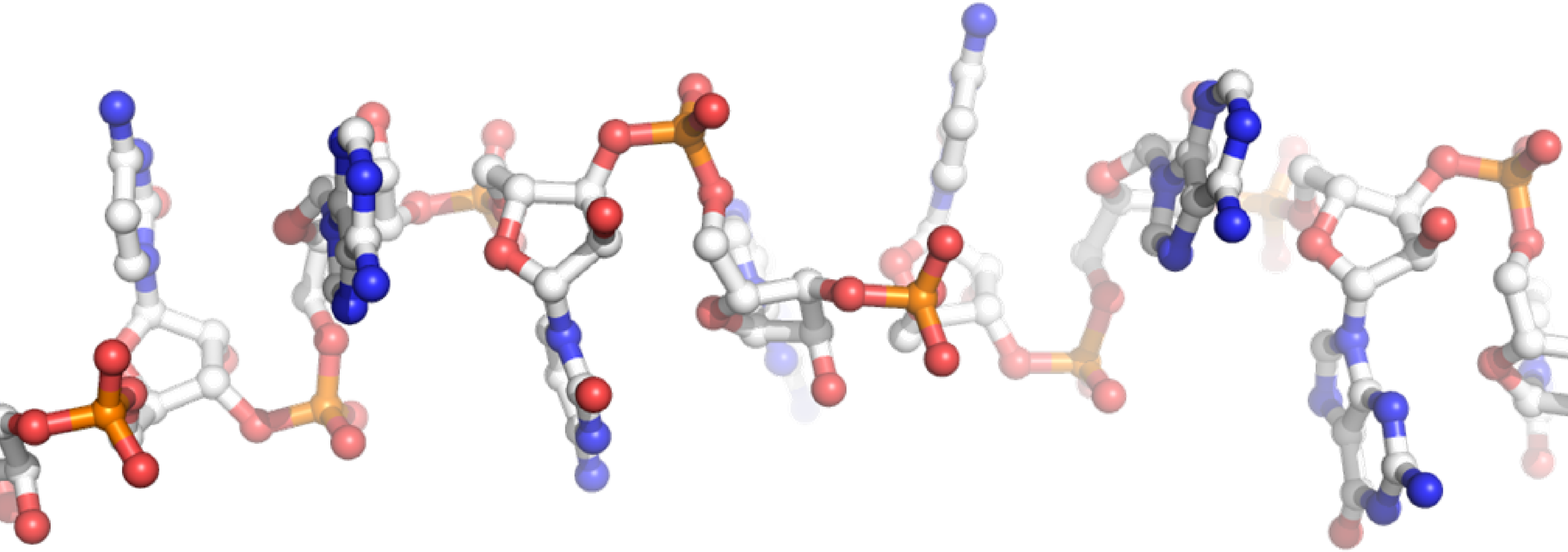
A more complex motif





1.09 Å heavy-atom RMSD from crystallographic model

A simple recipe – find the optimum



A simple recipe – find the optimum

$$U=\sum \tfrac{1}{2}K_b(b-b_o)^2+\sum \tfrac{1}{2}K_\theta(\theta-\theta_o)^2$$

All Bonds All Angles

$$+\sum K_\phi[1-\cos(n\phi+\delta)]$$

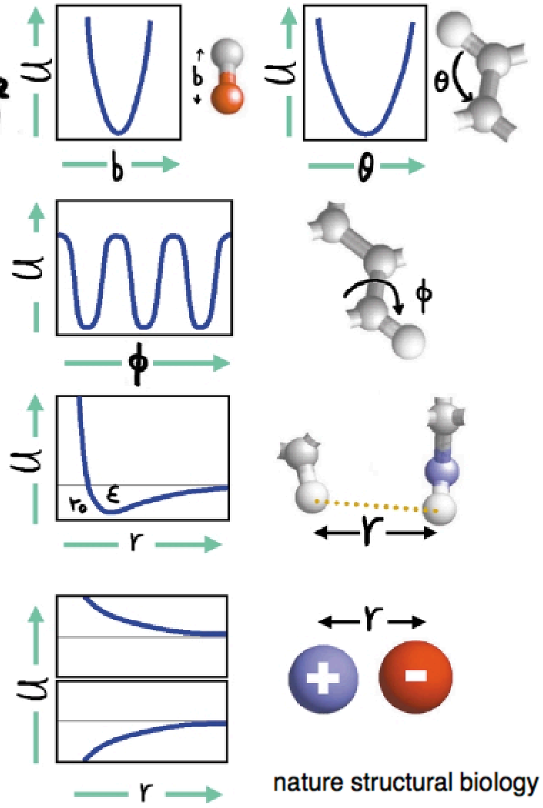
All Torsion Angles

$$+\sum \epsilon\left[\left(\frac{r_o}{r}\right)^{12}-2\left(\frac{r_o}{r}\right)^{6}\right]$$

All nonbonded pairs

$$+\sum 332 q_i q_j/r$$

All partial charges

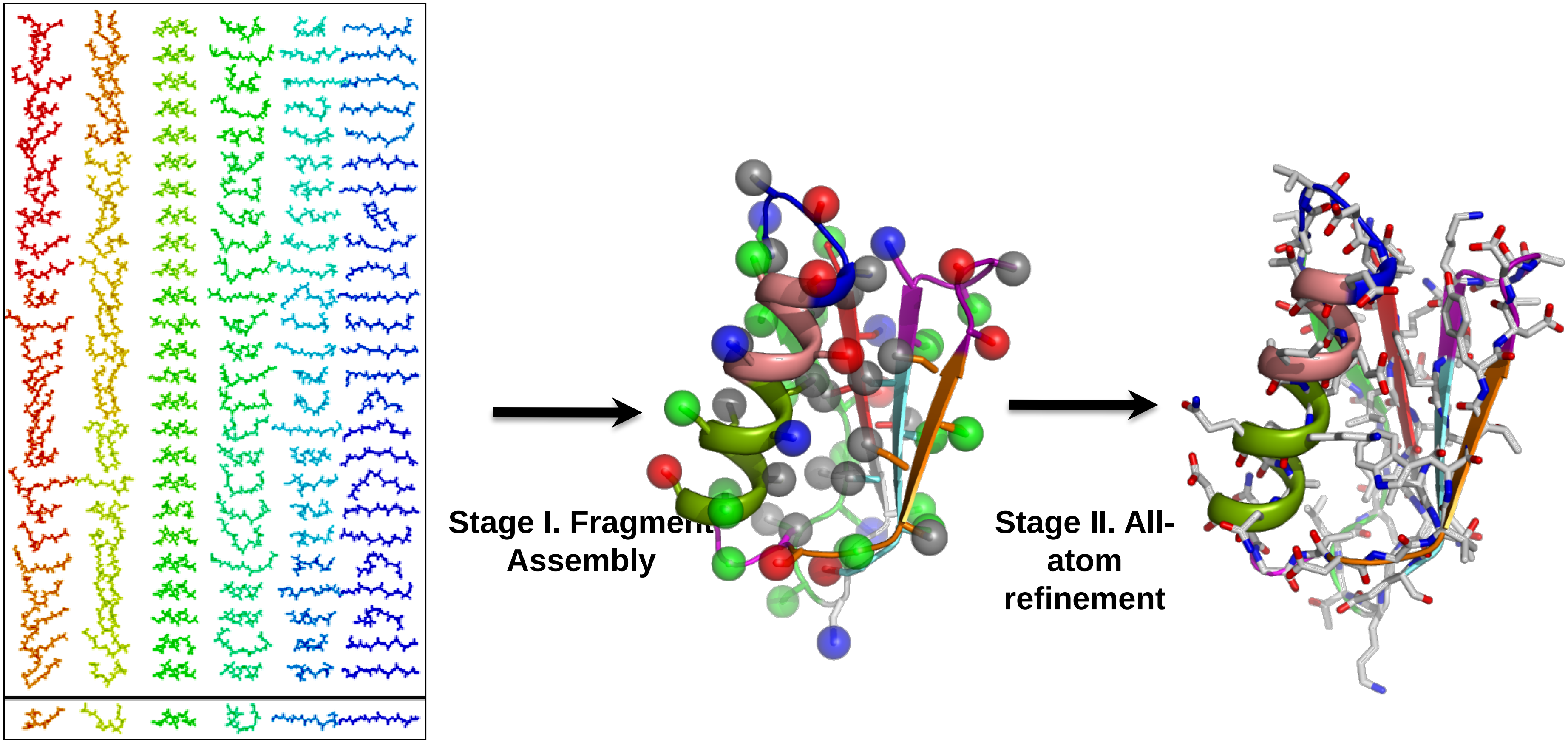


nature structural biology • volume 8 number 5 • may 2001

The birth of computational structural biology

Michael Levitt

The state of *de novo* structure prediction



Stage I. Fragment Assembly

Stage II. All-atom refinement

The standard ROSETTA routine. SEE ALSO: Work by David Jones, Skolnick & Zhang (TASSER), others