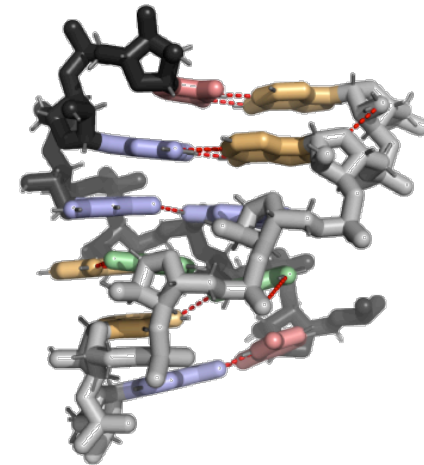
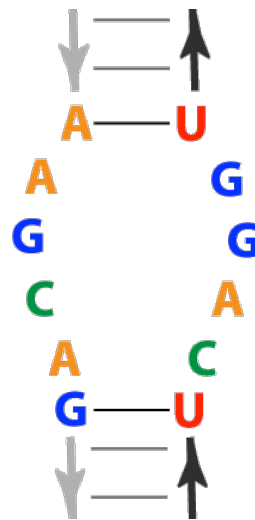
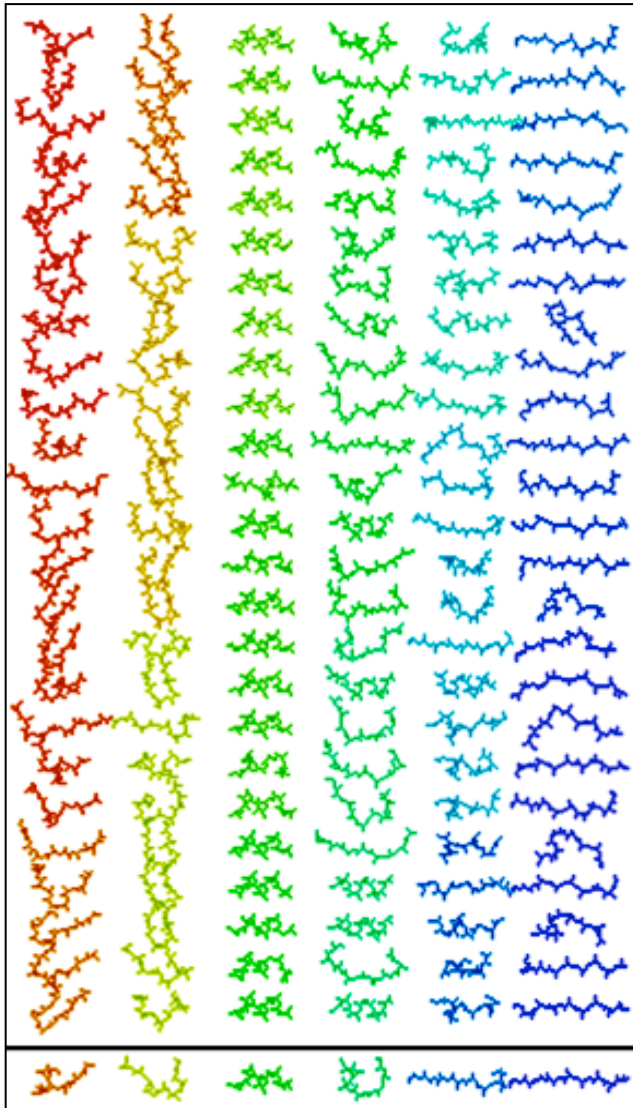


Structures from scratch

Rhiju Das, Departments of
Biochemistry & Physics

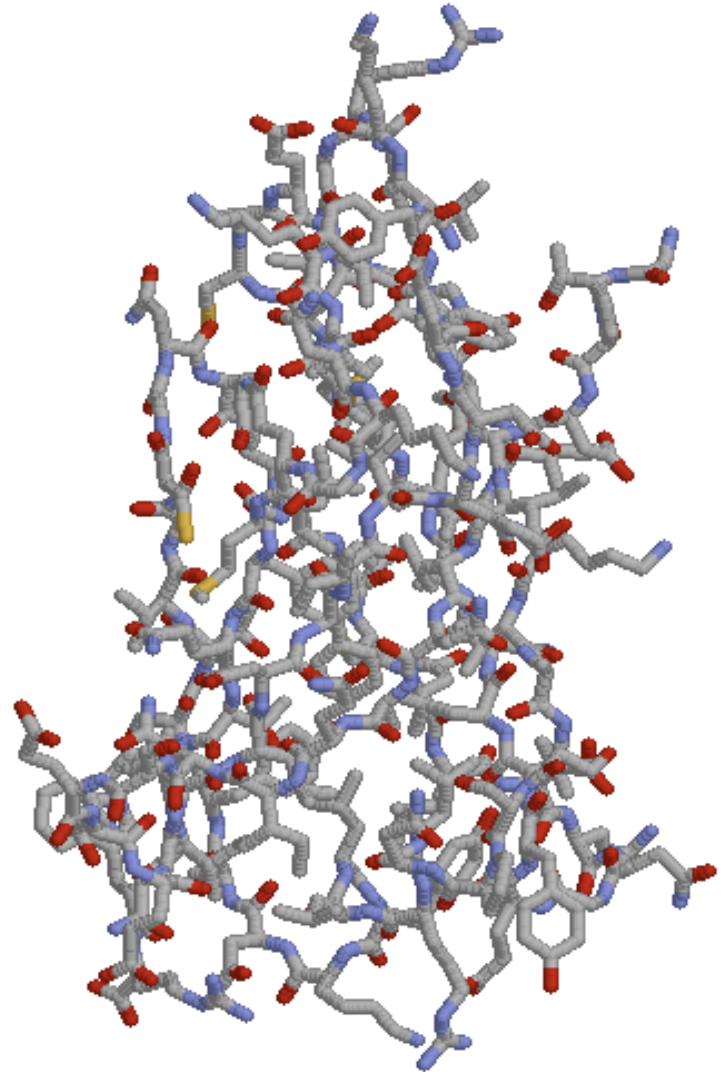
BIOC 218

Feb 2010



Predicting protein structure

GTPDIIVNAQINS
EDENVLDFIIEDEY
YLKKRGVGAHIK
VASSPQLRLLYKN
AYSTVSCGNYGVL
CNLVQNGEYDLN
AIMFNCAEIKLNK
GQMLFQTKIWR



Automatic Protein Fold Prediction Results

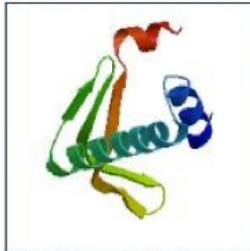
<http://swissmodel.expasy.org/>

Workunit: P000002 Title: HU E. coli



Go to: [Template Selection] [Alignment] [Modelling Log] [Evaluation]

Model Details: ? Segment 1



Model info:

modelled residue range: 1 to 90
based on template **1mula** (2.30 Å)
Sequence Identity [%]: 84.444
Evalue: 3.25e-22

display model: as pdb - as DeepView project

download model: as pdb - as Deepview project - as text

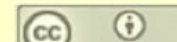
Alignment ? [top]

TARGET	1	MNKTQLID VIAEKAELSK TQAKAALEST LAAITESLKE GDAVQLVGFG
1mula	1	mnktqlid viaekaelsk tqakaalest laaiteslke gdavqlvgfg

TARGET		hhhhh hhhhh h hhhhhhhhh hhhhhhhhh sss ss
1mula		hhhhh hhhhh h hhhhhhhhh hhhhhhhhh sss ss

TARGET	49	TFKVNRAER TGRNPQTGKE IKIAAANVPA FVSGKALKDA VK
1mula	49	tfkvnhrae- ----- ---aaanvpa fvsgkalkda vk-

TARGET		ssssss sssss sss ss sssss sssshhhhh h
1mula		ssssss sssss sssss sssshhhhh



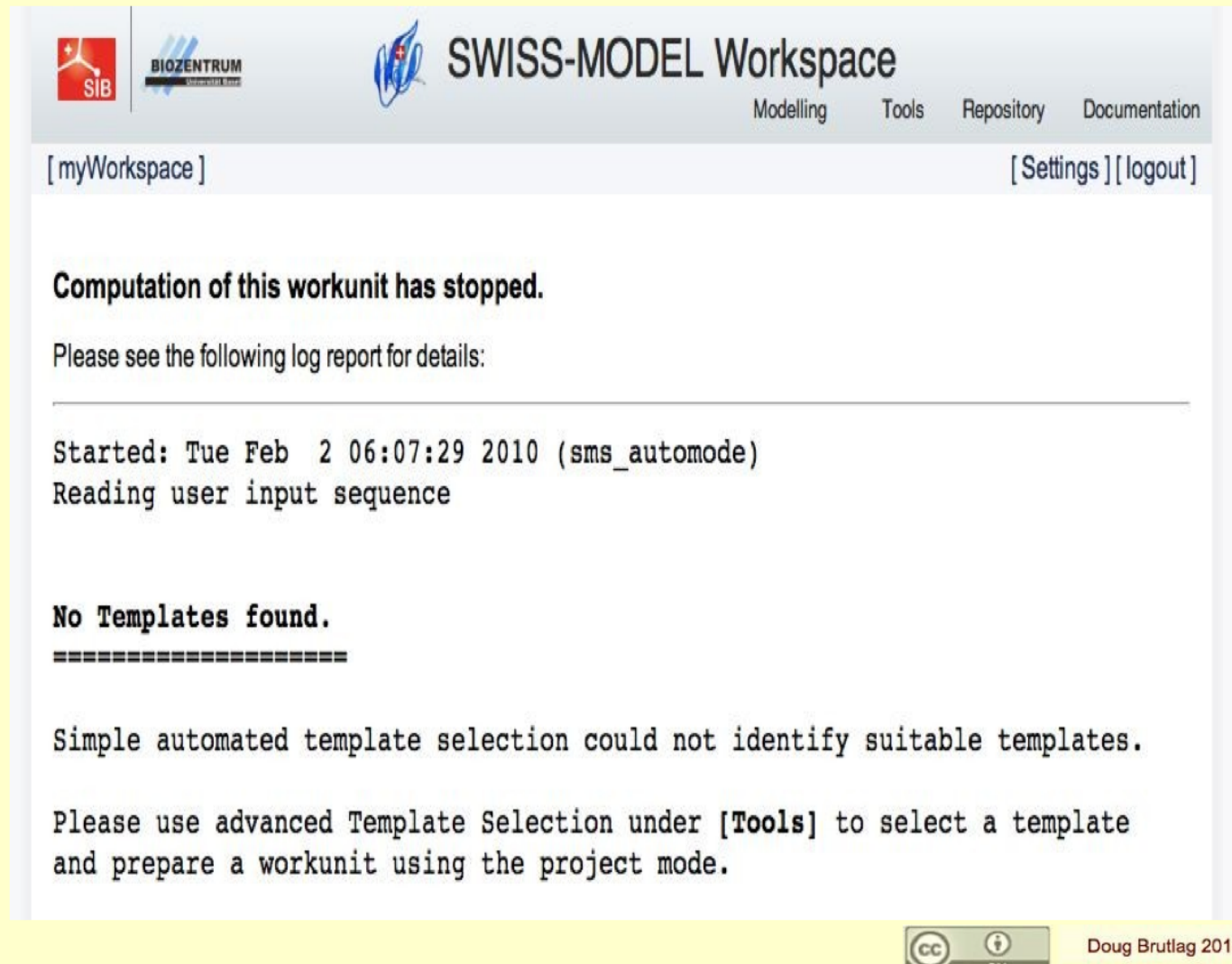
Doug Brutlag 2010

**This will
happen
to you a
lot.**

ODEL

Automatic Protein Fold Prediction Results

<http://swissmodel.expasy.org/>



The screenshot shows the SWISS-MODEL Workspace interface. At the top, there are logos for SIB, BIOCENTRUM, and SWISS-MODEL. The main header reads "SWISS-MODEL Workspace" with navigation links for "Modelling", "Tools", "Repository", and "Documentation". Below the header, there are links for "[myWorkspace]" and "[Settings] [logout]". The main content area displays a message: "Computation of this workunit has stopped." followed by "Please see the following log report for details:". The log report text is as follows:

```
Started: Tue Feb  2 06:07:29 2010 (sms_automode)
Reading user input sequence

No Templates found.
=====

Simple automated template selection could not identify suitable templates.

Please use advanced Template Selection under [Tools] to select a template
and prepare a workunit using the project mode.
```

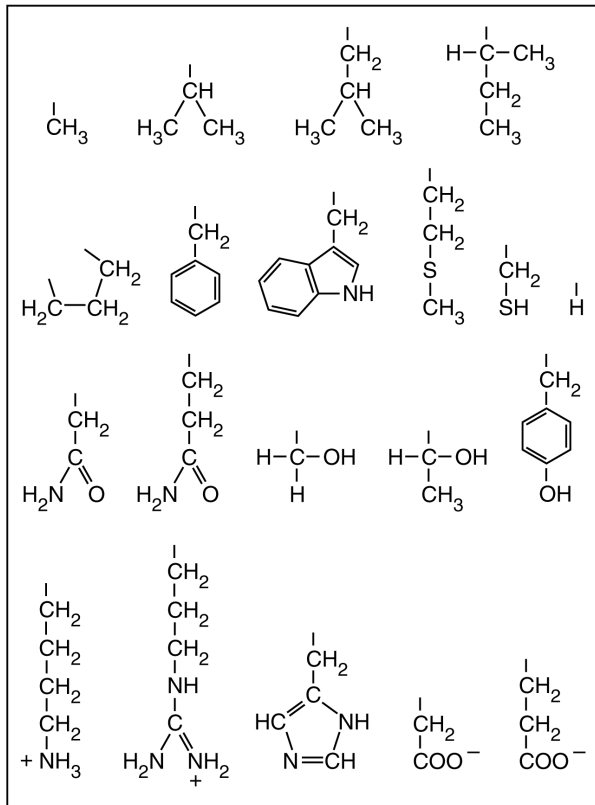
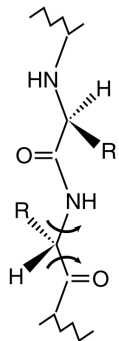
At the bottom right of the interface, there is a Creative Commons license icon (CC BY) and the text "Doug Brutlag 2010".

Proteins

Backbone

Side Chains

R ≡

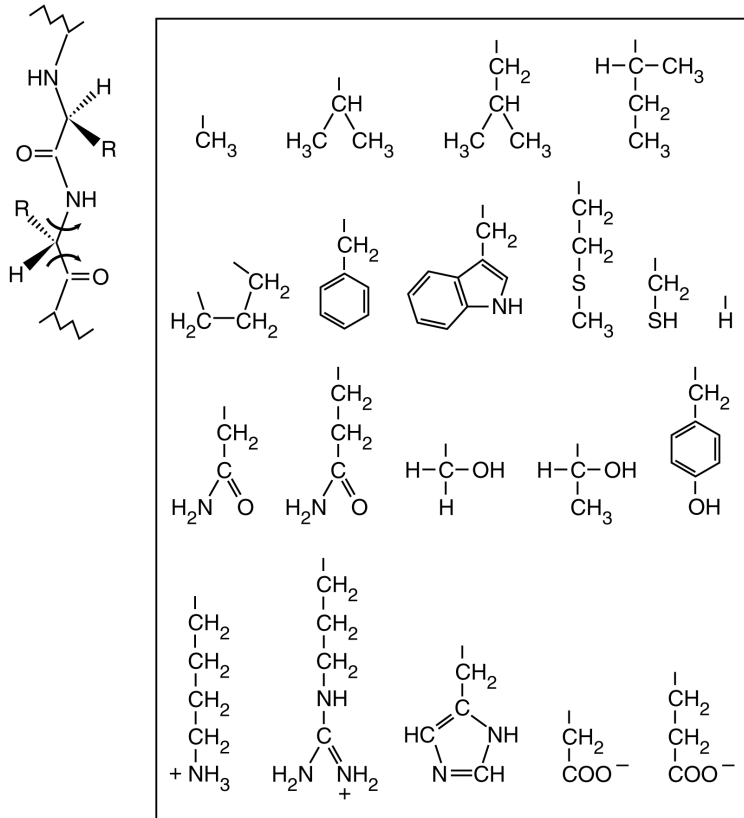


Proteins

Backbone

Side Chains

$R \equiv$



THE STRUCTURE OF PROTEINS: TWO HYDROGEN-BONDED HELICAL CONFIGURATIONS OF THE POLYPEPTIDE CHAIN

BY LINUS PAULING, ROBERT B. COREY, AND H. R. BRANSON*

GATES AND CRELLIN LABORATORIES OF CHEMISTRY,
CALIFORNIA INSTITUTE OF TECHNOLOGY, PASADENA, CALIFORNIA†

Communicated February 28, 1951

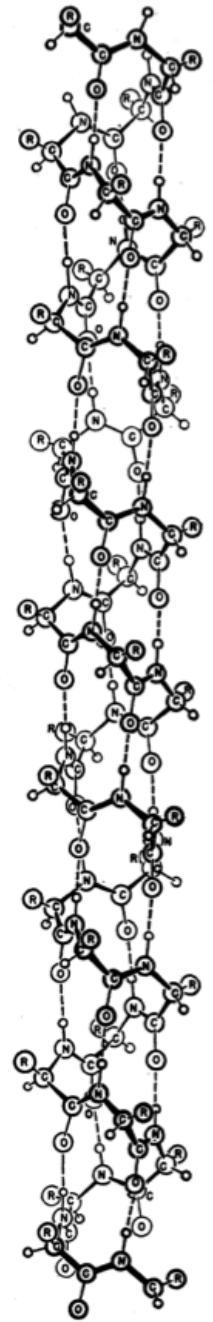
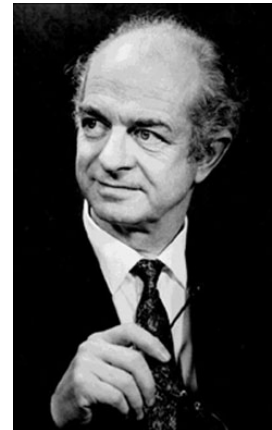
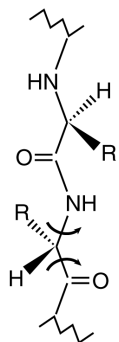


FIGURE 2
The helix with 3.7 residues per turn.

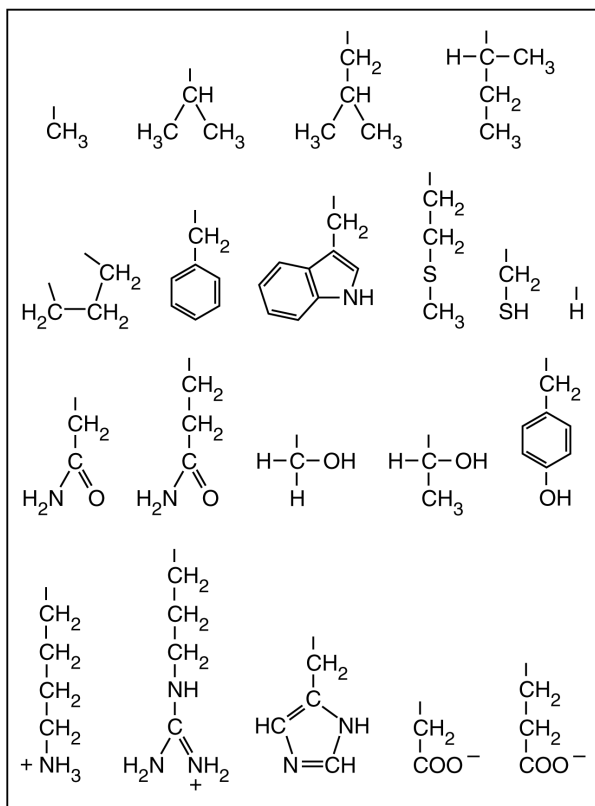
Proteins

Backbone

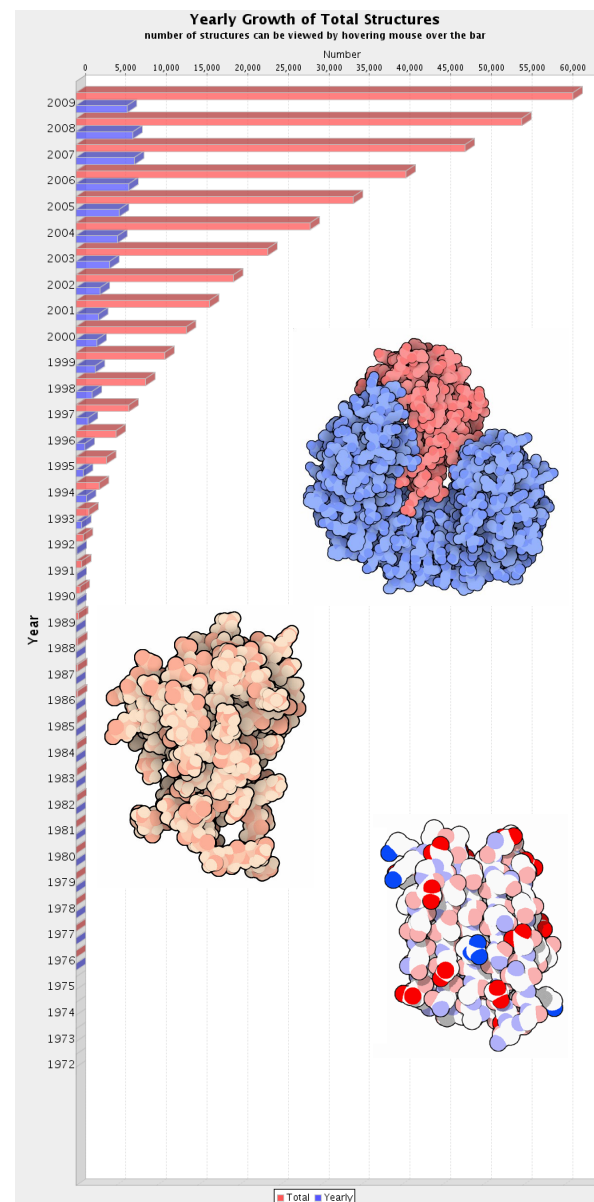


Side Chains

R ≡



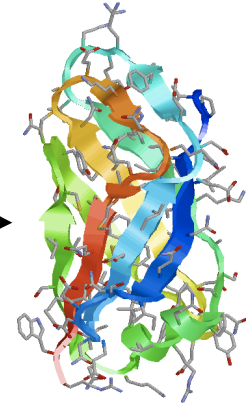
RCSB **PDB**
PROTEIN DATA BANK



Two fundamental problems

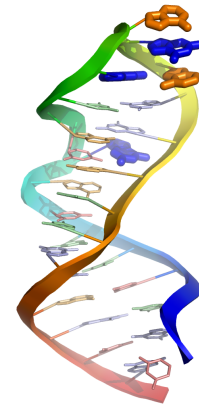
1. Predicting protein structure

GTPDIIVNAQINSEENVLDF
IIEDEYYLKRGVGAHIKVAS
SPQLRLLYKNAYSTVSCGNYG
VLCNLVQNGEYDLNAIMFNC
AEIKLNKGQMLFQTKIWR



2. Predicting RNA structure

ugcuccuaguacgag
aggaccggagug



On the formation of protein tertiary structure on a computer

(protein folding/computer simulation/protein evolution/role of glycines)

ARNOLD T. HAGLER* AND BARRY HONIG†

* Department of Chemical Physics, Weizmann Institute of Science, Rehovot, Israel; and † Department of Physical Chemistry, The Hebrew University, Jerusalem, Israel

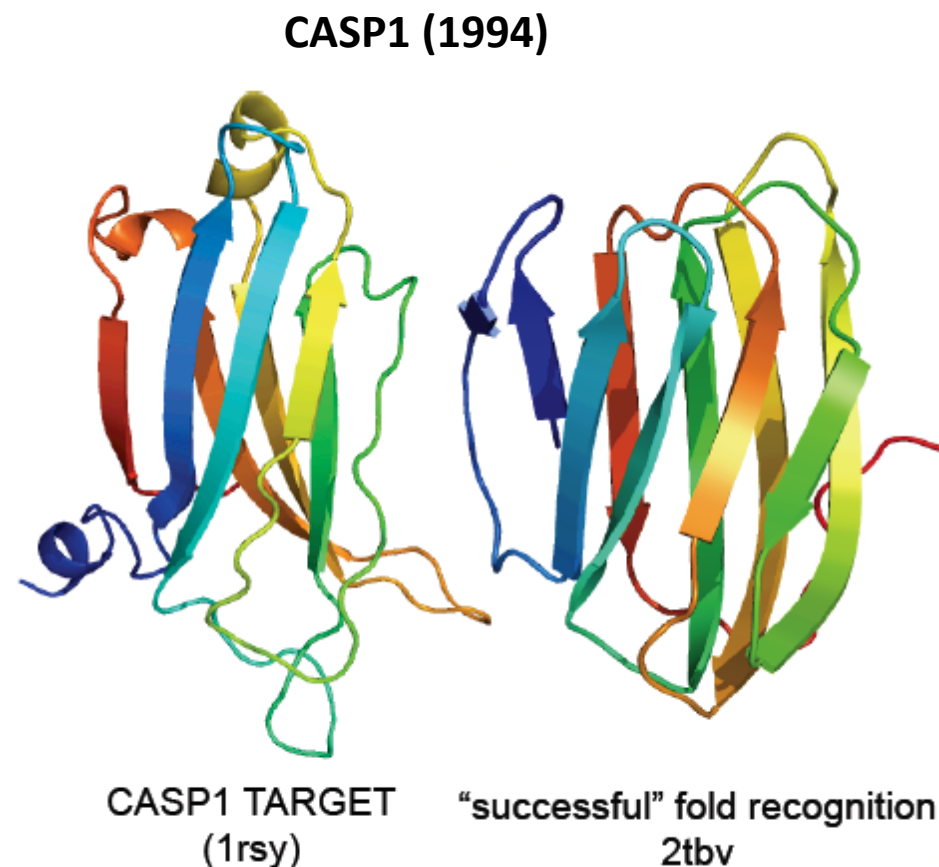
Communicated by Cyrus Levinthal, November 17, 1977

The impression generated by these various simulations is that major progress has been made towards predicting the tertiary structure of a protein from its amino-acid sequence; i.e., the folding problem may be far more tractable than has generally been considered (9).

Driving innovation in protein structure prediction: “CASP”

Critical Assessment of
Structure Prediction

**Five *blind*
predictions per
target**



RMDS: 16.0 Å

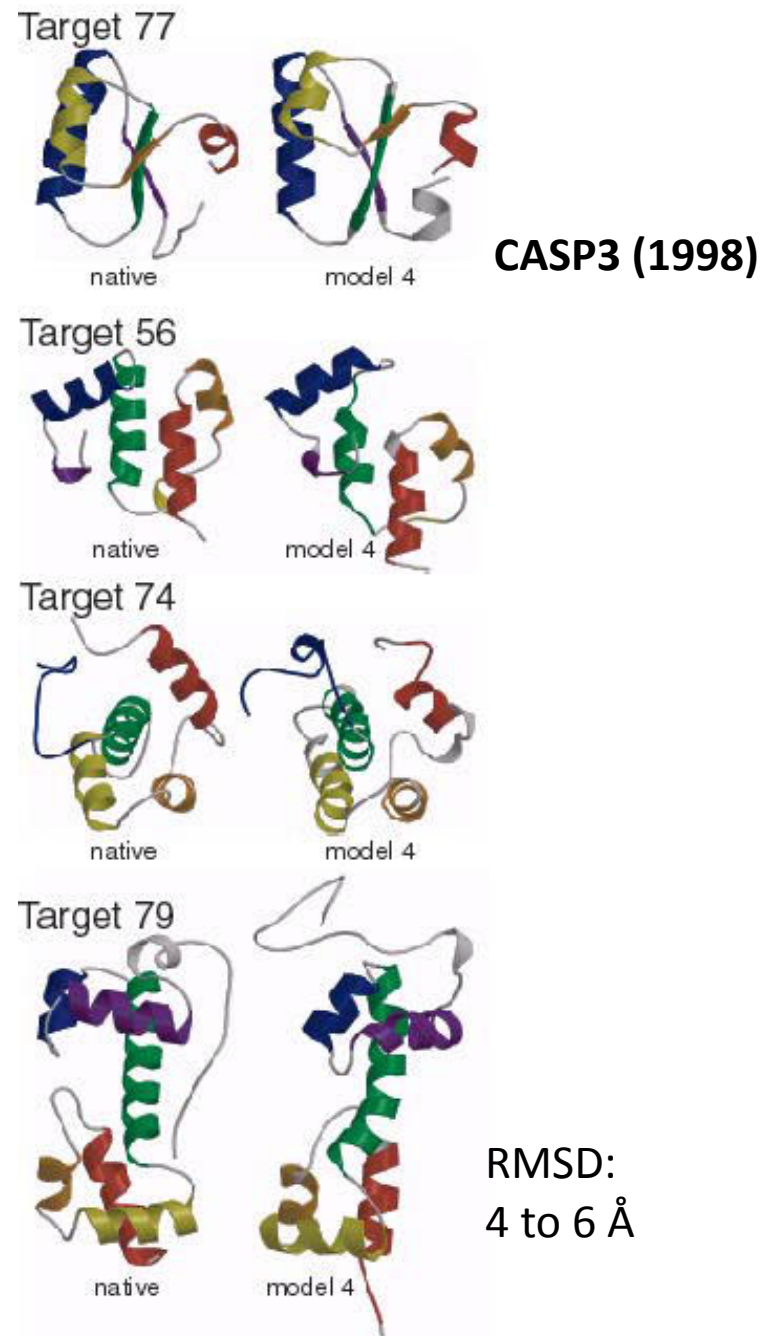
From Neil Clarke, CASP7 assessor’s talk

Driving innovation in protein structure prediction: “CASP”

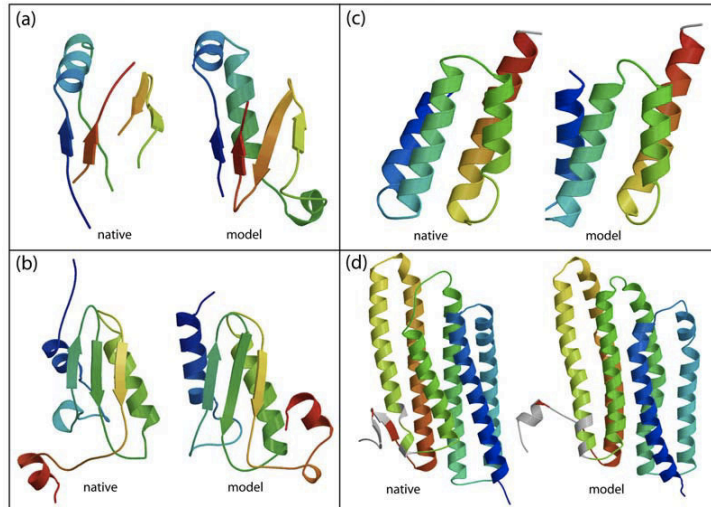
Critical Assessment of
Structure Prediction

**Five *blind*
predictions per
target**

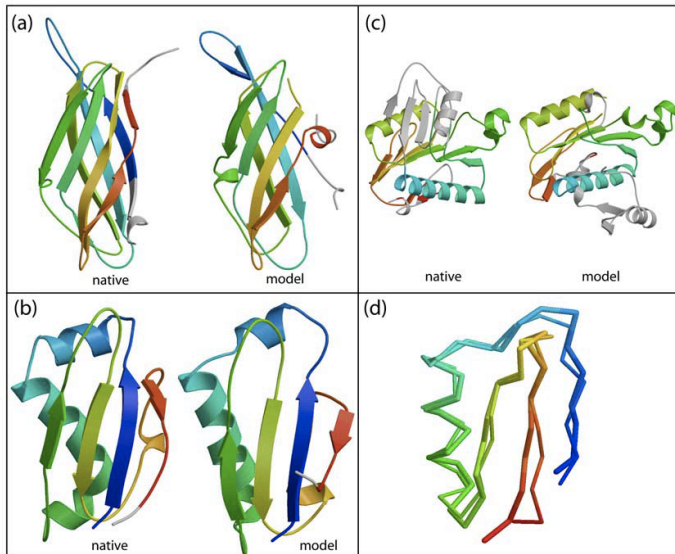
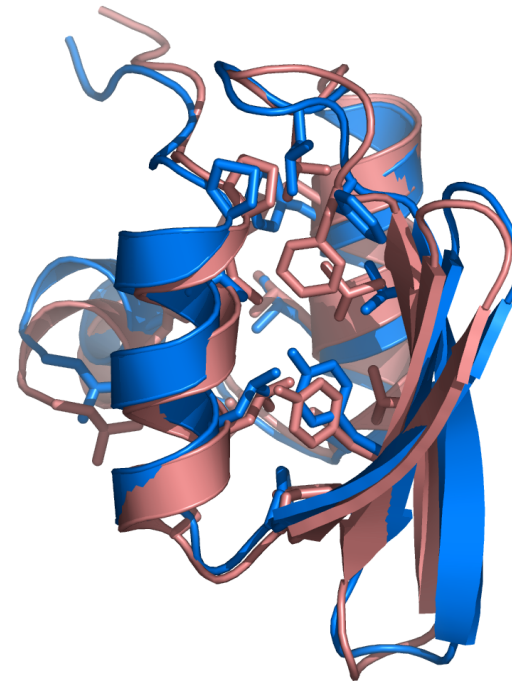
DAVID BAKER & colleagues



CASP6 (2004)



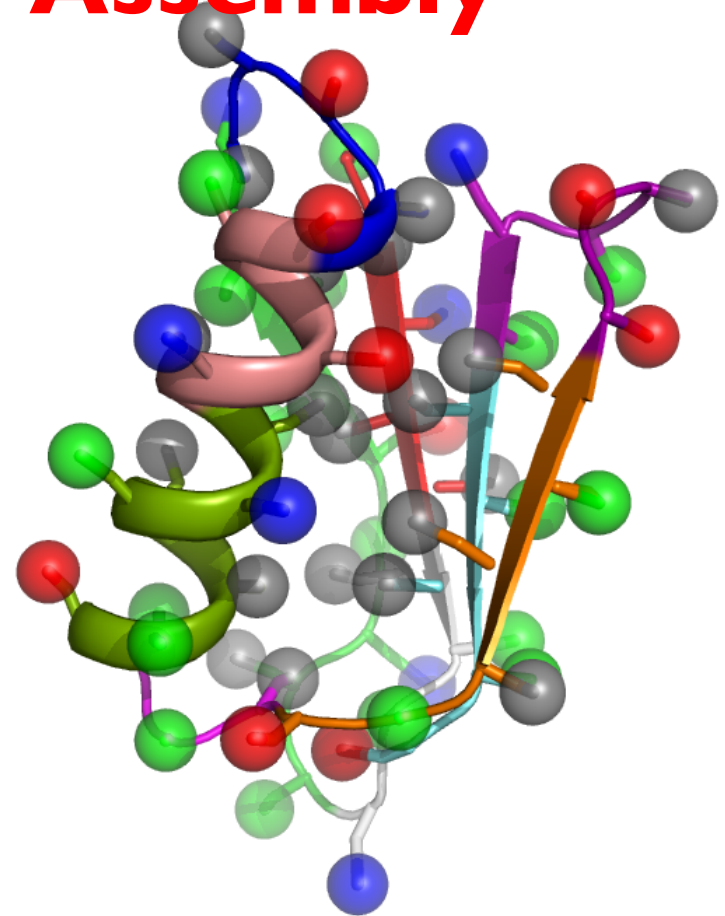
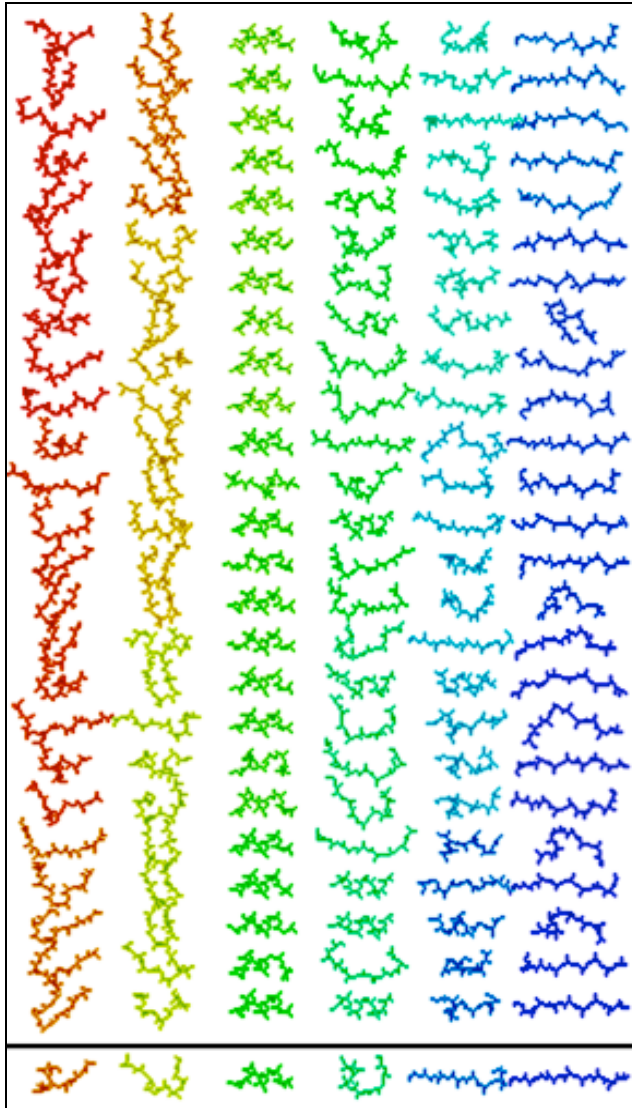
T0281 (1.6 Å over 70 residues)



DAVID BAKER & colleagues

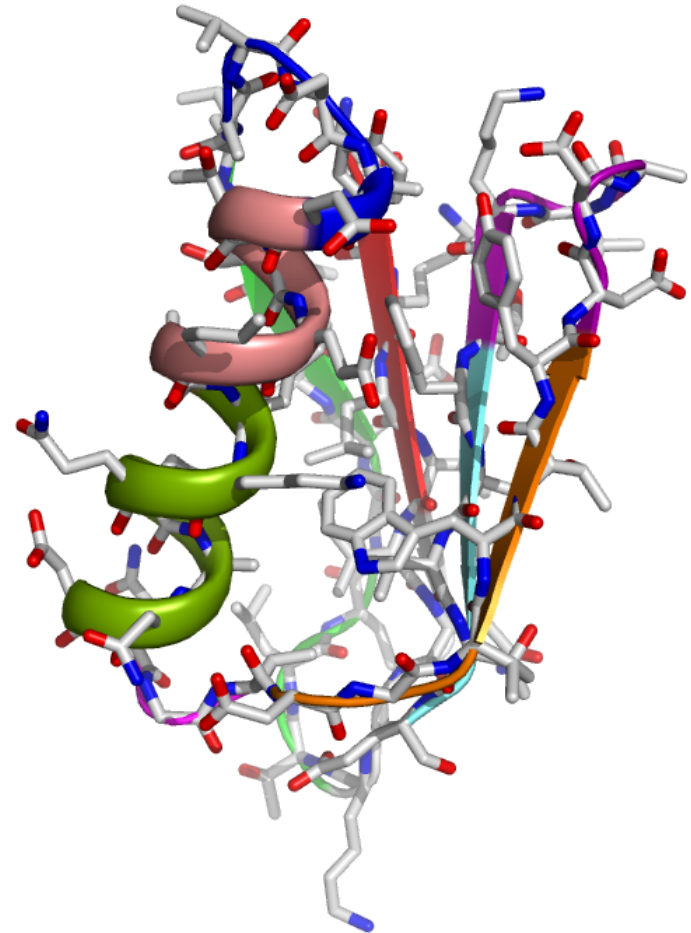
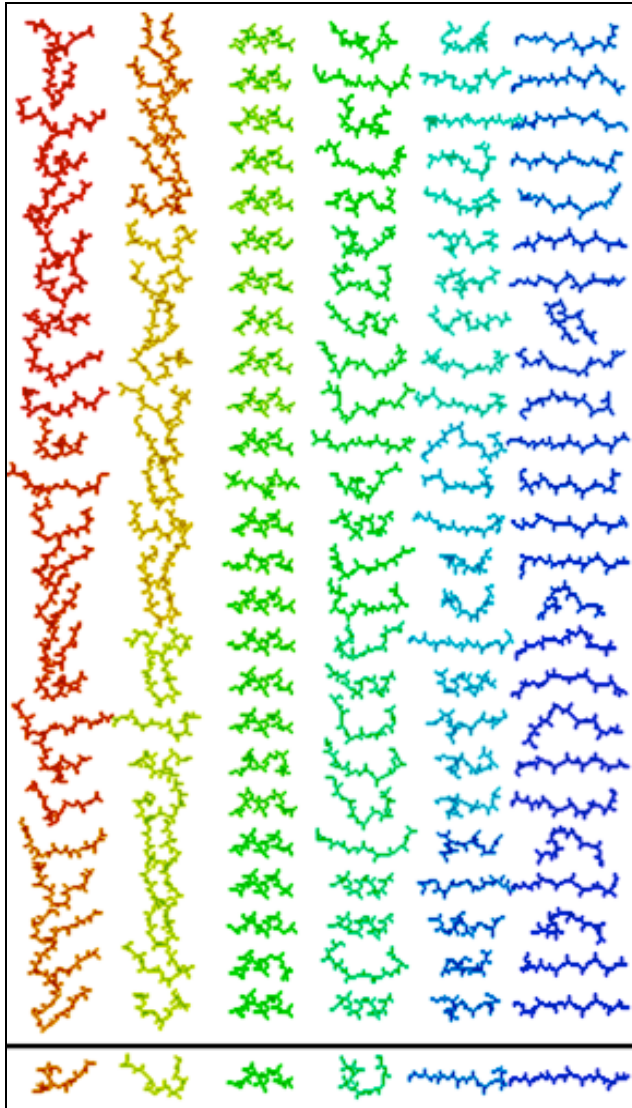
De novo Modeling with Rosetta

Stage I. Fragment Assembly



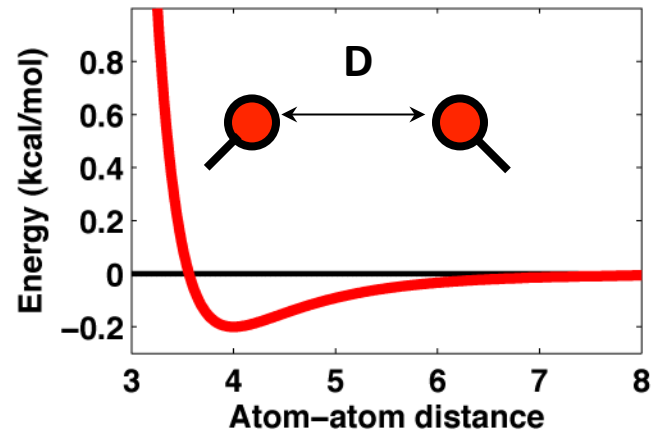
De novo Modeling with Rosetta

Stage II. All-atom refinement

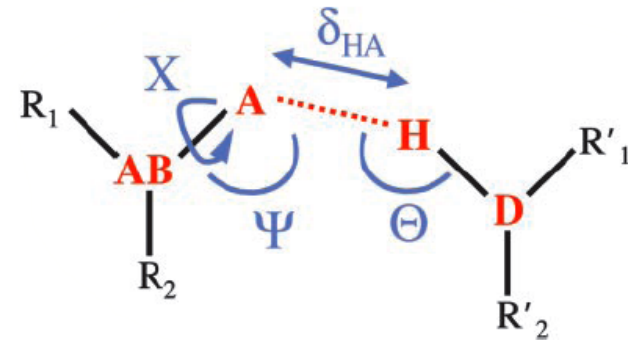


Ingredients of a high resolution potential

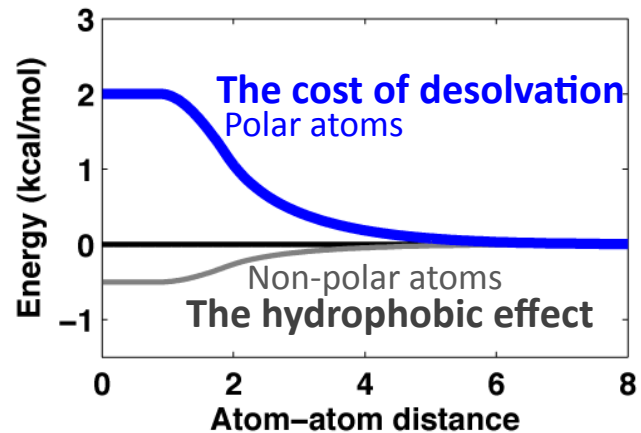
1. Van der waals packing



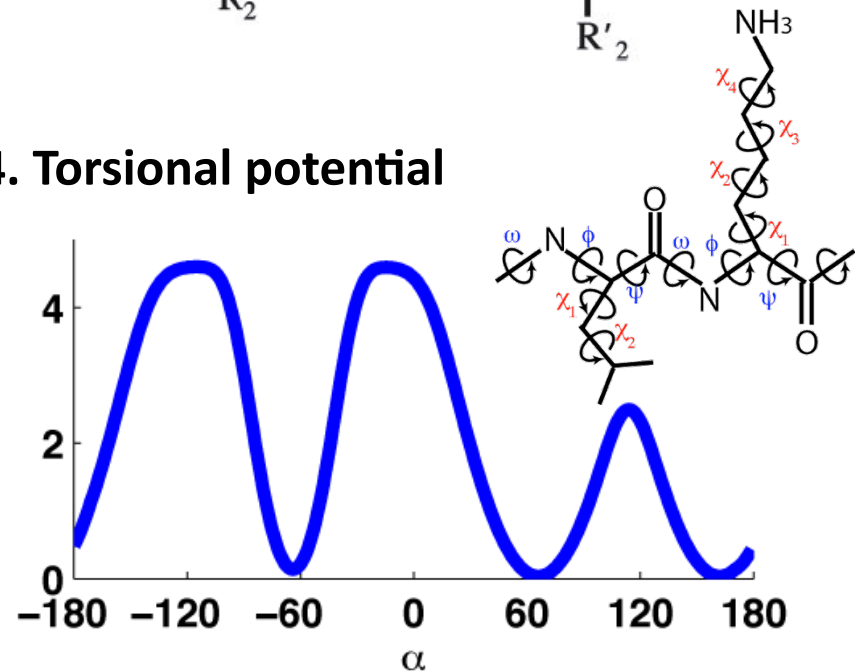
2. Hydrogen bonds



3. Manifestations of water

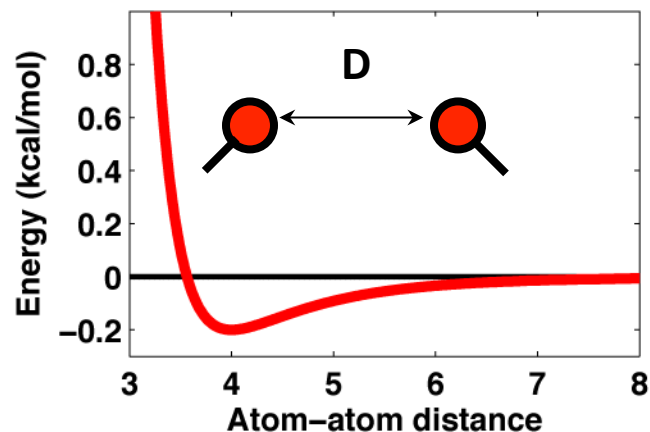


4. Torsional potential

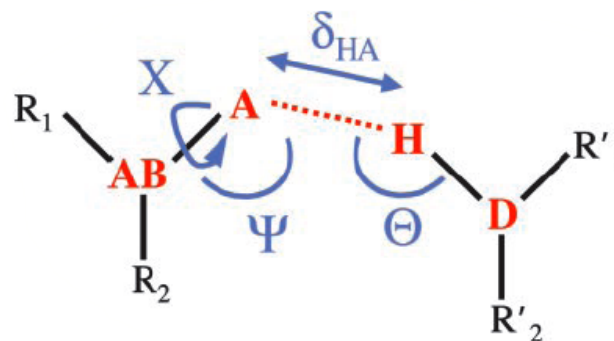


Ingredients of a high resolution potential

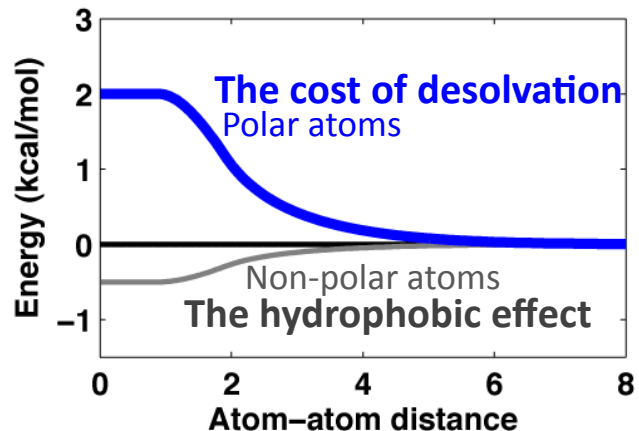
1. Van der waals packing



2. Hydrogen bonds



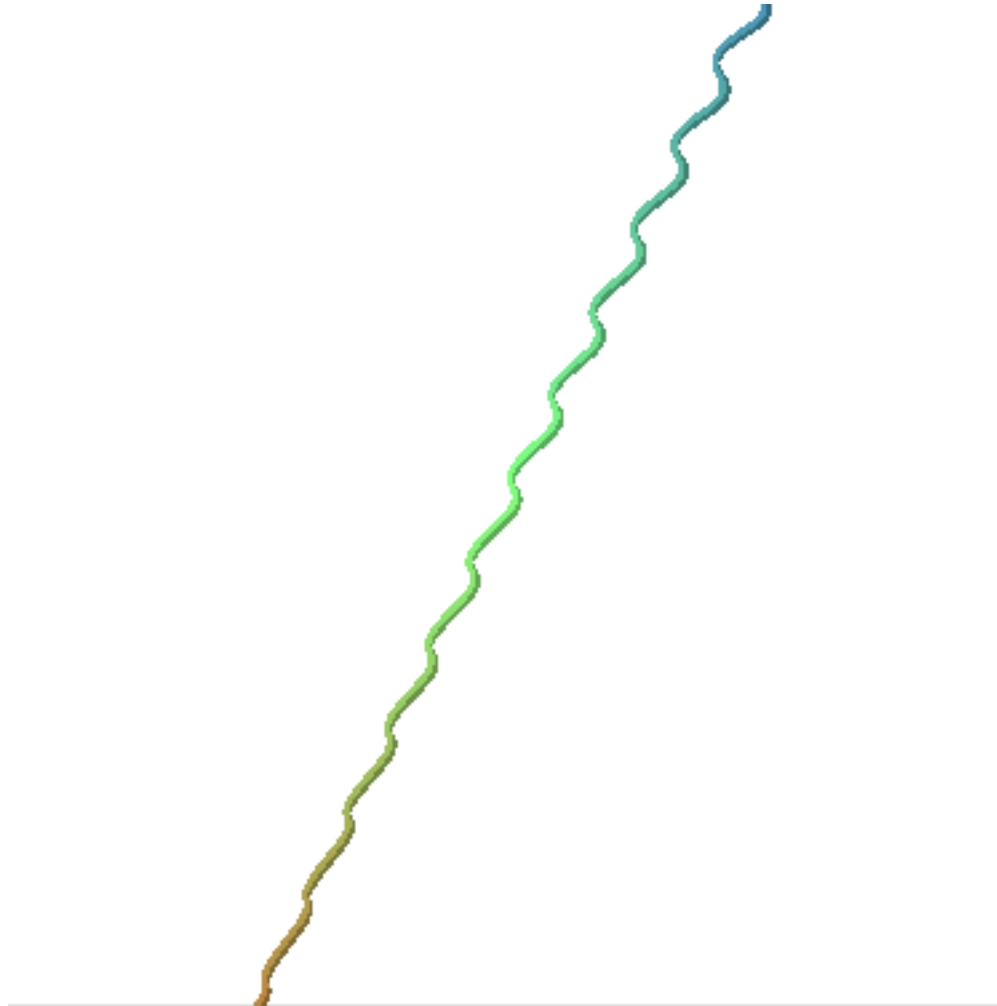
3. Manifestations of water



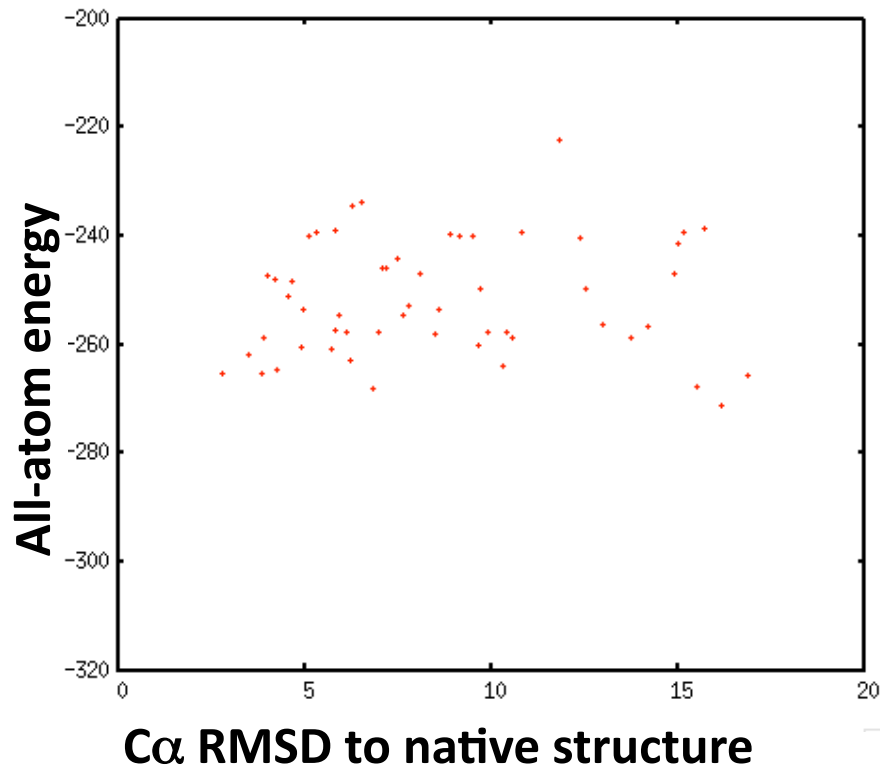
the clustering of non-polar groups, the enhanced hydrogen bond attraction in a hydrophobic environment and the access of water molecules to those polar groups which are not involved in hydrogen bonds.

– Michael Levitt, 1969

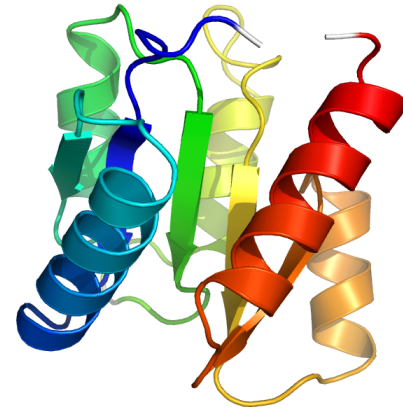
Rosetta in action



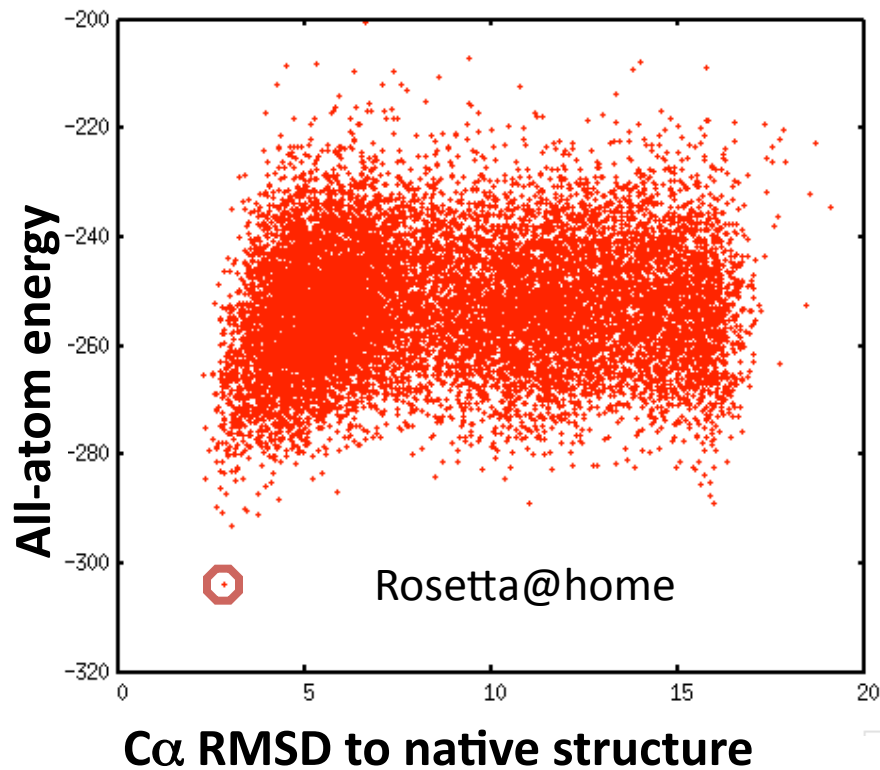
A ~1000-fold increase in computational power



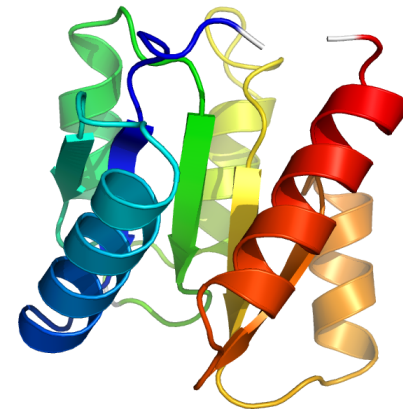
Native (CheY)



A ~1000-fold increase in computational power



Native (CheY)

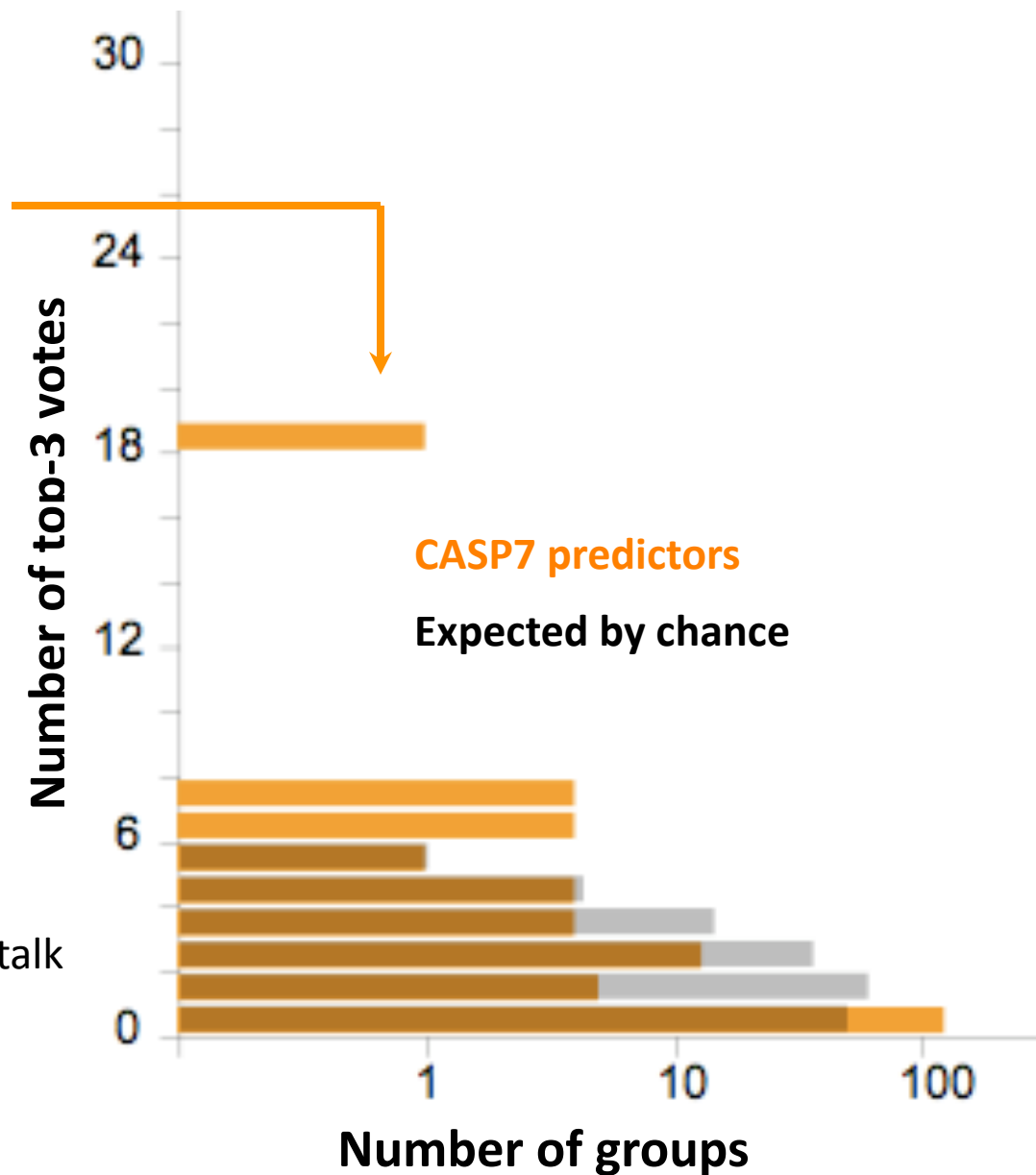


Lowest energy
Rosetta structure



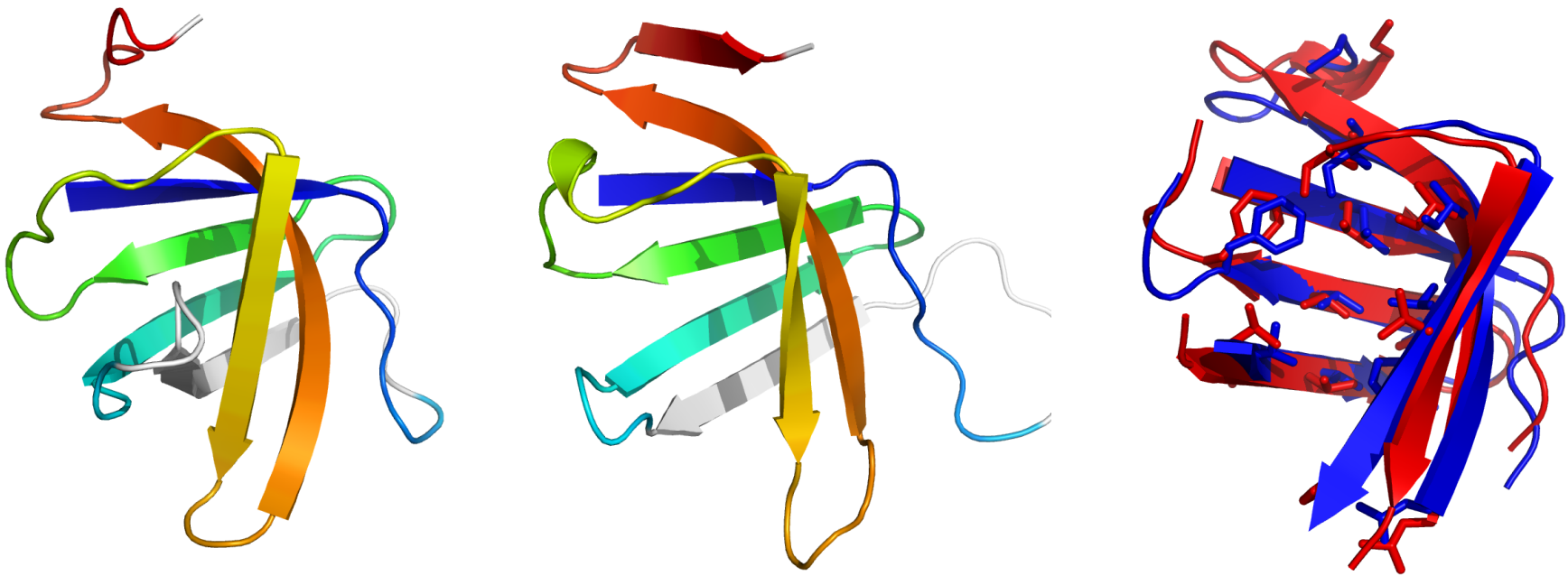
Rosetta@home in CASP7

From Neil Clarke, CASP7 assessor's talk
on "free modeling"



De novo successes: all- β

CASP7 target T0316 (domain 3)



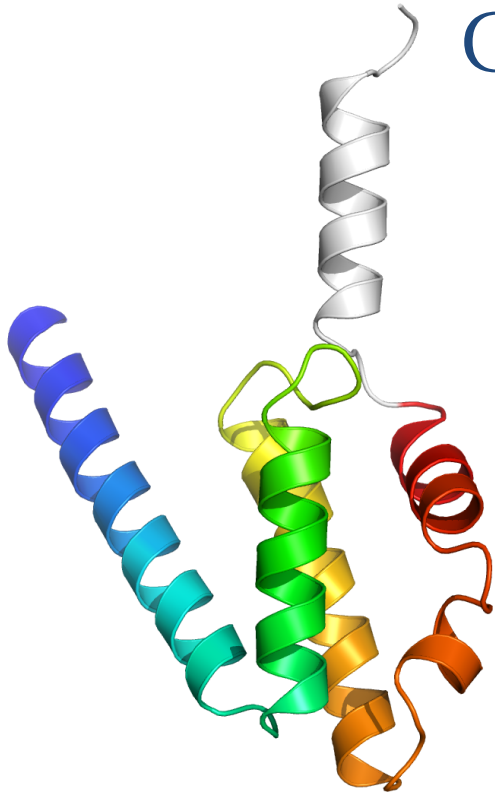
Native

Model

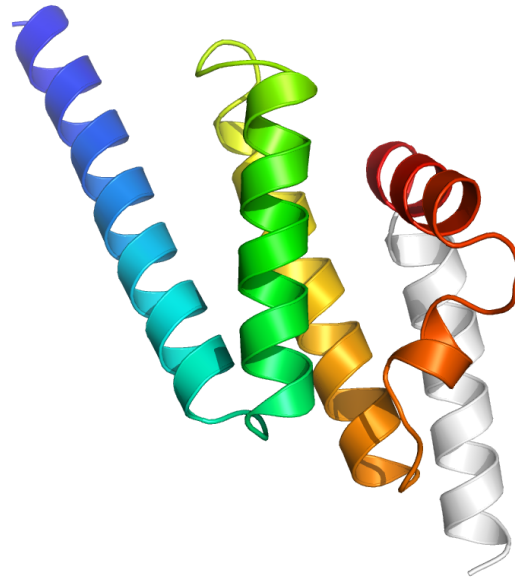
2.0 Å over 61 residues

De novo successes: all- α

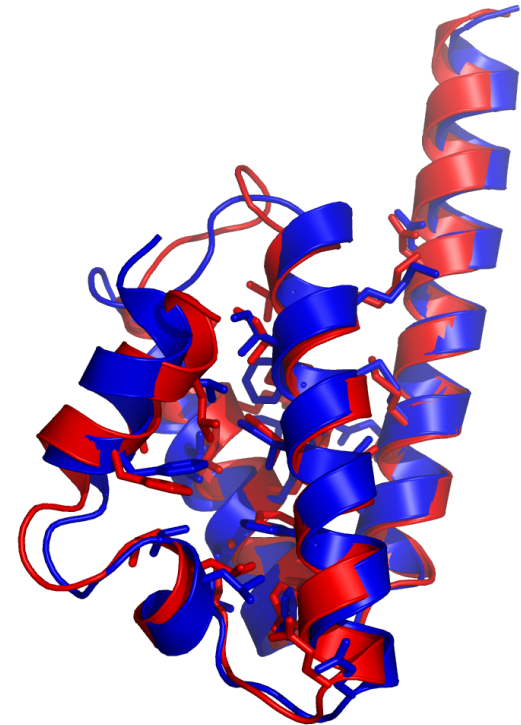
CASP7 target T0283 (112 residues)



Native

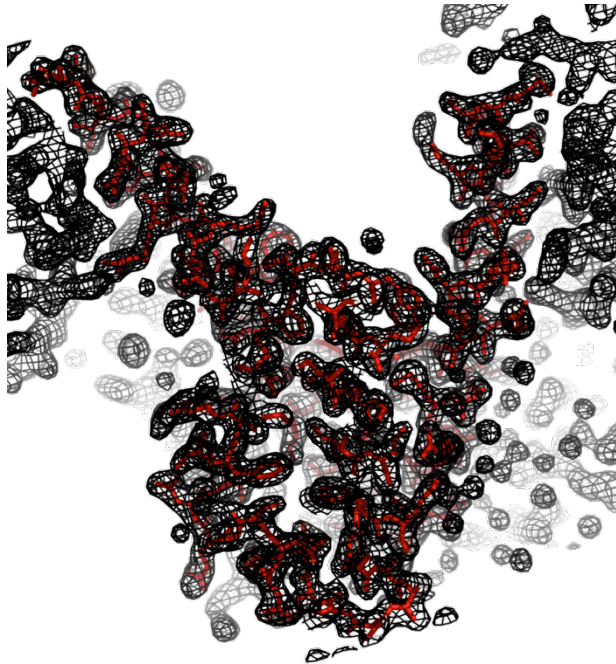


Model

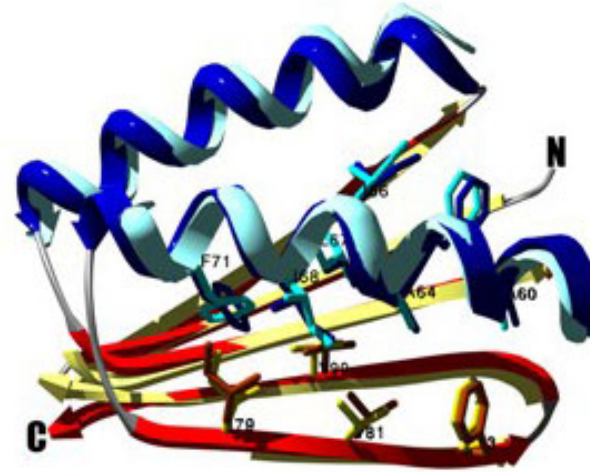


1.4 Å over 90 residues

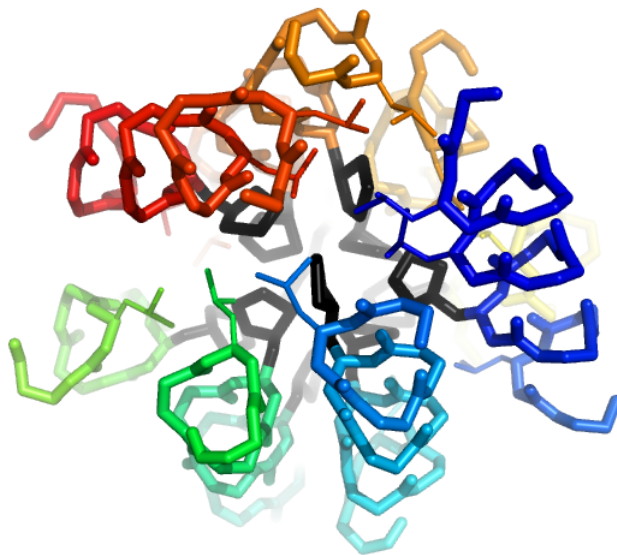
De novo modeling: connections to the real world



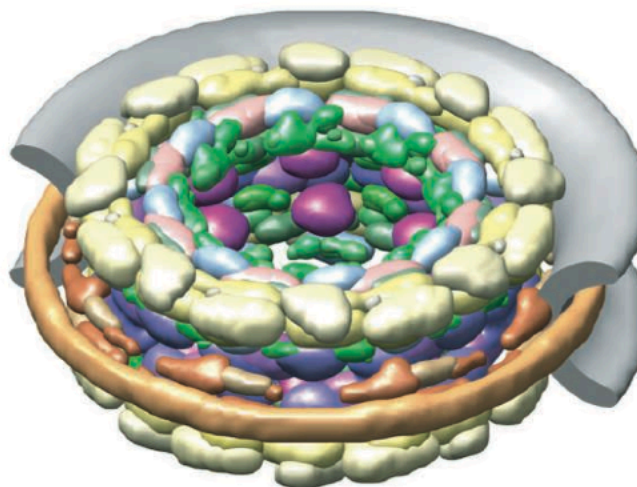
The crystallographic phase problem



Engineering new protein folds and new enzymes



Non-biological polymers: beta proteins



NEWS

Problem Solved*

(*sort of)

Researchers have toiled for decades to understand how floppy chains of amino acids fold into functional proteins. Learning many of those rules has brought them to the verge of being able to make predictions about proteins they haven't even discovered

IN 1961, CHRISTIAN ANFINSEN, A BIOCHEMIST at the U.S. National Institutes of Health, saw something that continues to perplex and inspire researchers to this day. Anfinsen was studying an RNA-chewing protein called ribonuclease (RNase). Like all proteins, RNase is made from a long string of building blocks called amino acids that fold up into a particular three-dimensional (3D) shape to give RNase its chemical abilities.

Anfinsen raised the temperature of his protein, causing it to unravel into a spaghetti-like string. When he cooled it back down again, the protein automatically refolded itself into its normal 3D shape. The implication: Proteins aren't folded by some external cellular machine. Rather, the subtle chemical push and pull between amino acids tugs proteins into their 3D shapes. But how? Anfinsen's insights helped earn him a share of the 1972 Nobel Prize in chem-

istry—and laid the foundation for one of biology's grand challenges.

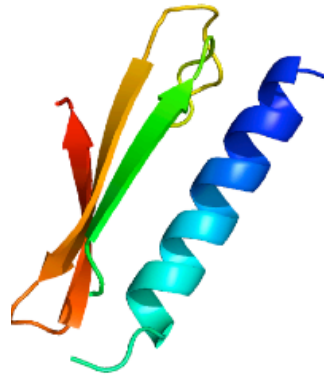
With an astronomical number of ways those chains of amino acids can potentially fold up, solving that challenge has long seemed beyond hope. But now many experts agree that key questions have been answered. Some even assert that the most daunting part of the problem—predicting the structure of unknown proteins—is now within reach, thanks to the inexorable improvements in computers and computer networks. “What was called the protein-folding problem 20 years ago is solved,” says Peter Wolynes, a chemist and protein-folding expert at the University of California, San Diego.

Most researchers won't go quite that far. David Baker of the University of Washington, Seattle, believes that such notions are “dangerous” and could undermine interest in the field. But all agree

(Hype)

Reality

target



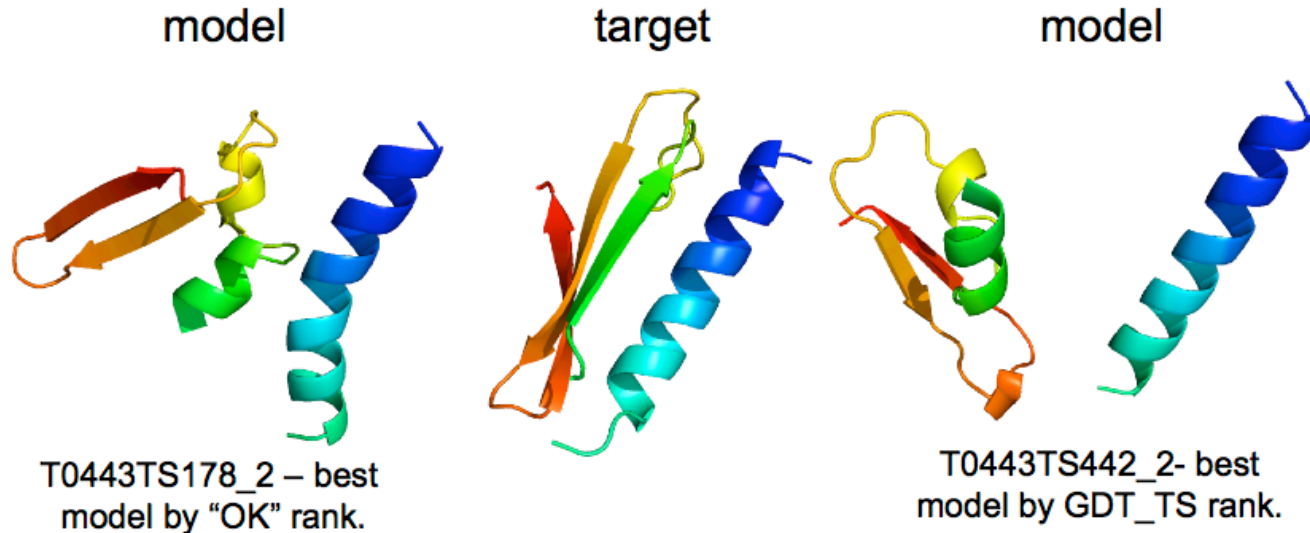
5-Dec-2008

CASP8

“Free Modeling”

Joel L. Sussman
Weizmann Institute of Science

Reality



**Each of the 3 visual assessors independently said:
“no good model for this target”**

- **No real progress from CASP7**

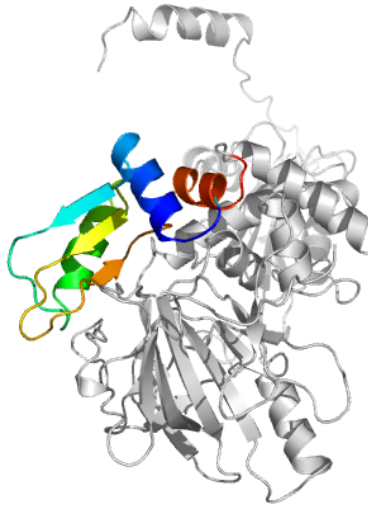


CASP8

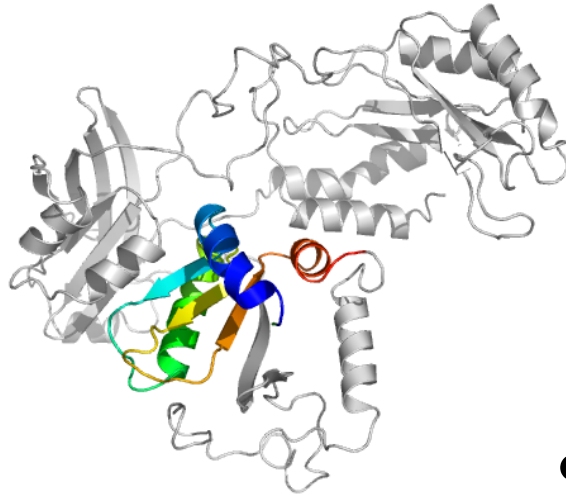
“Free Modeling”

5-Dec-2008

Is protein folding *solved*?



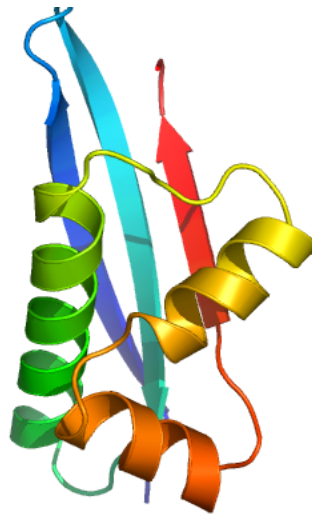
Native



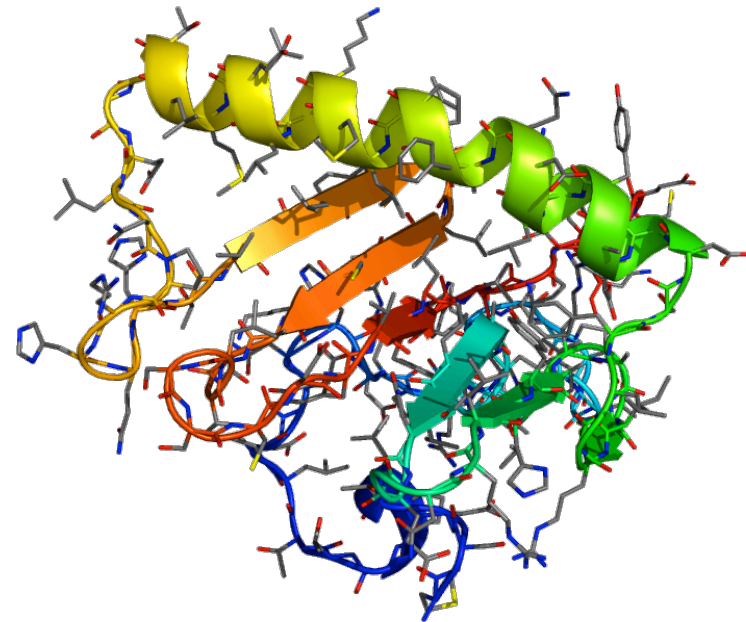
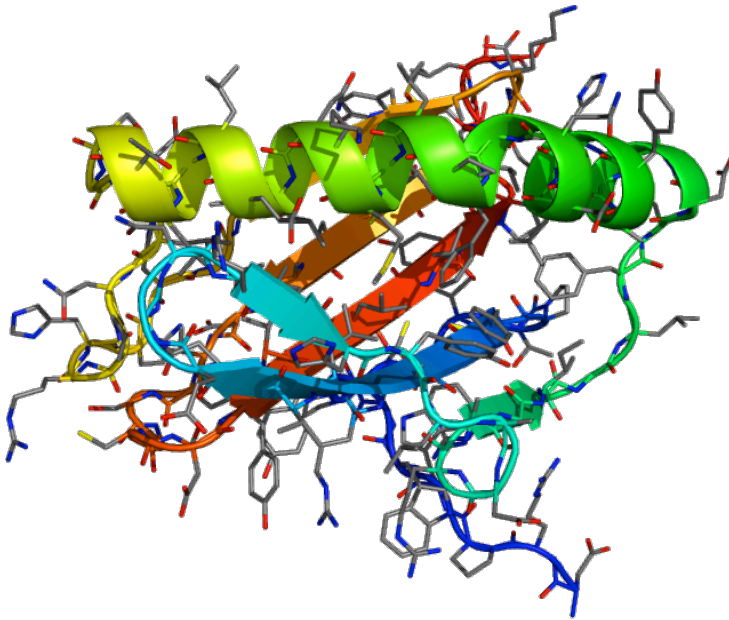
Model

NO!

- Success in $<1/3$ of cases.
- Conformational sampling still a huge issue

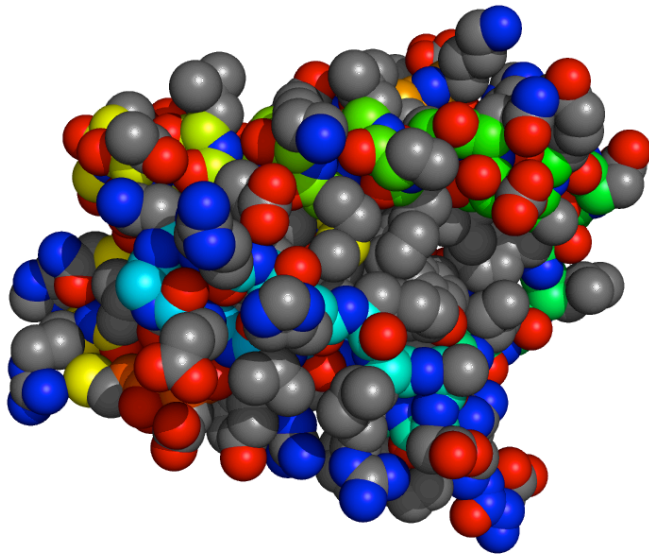


Can you pick out the right one?

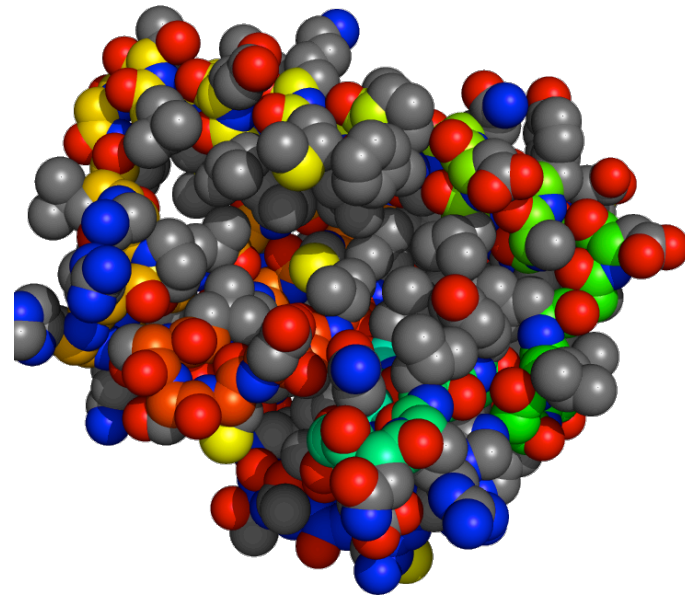


T304 (CASP7)

Can you pick out the right one?



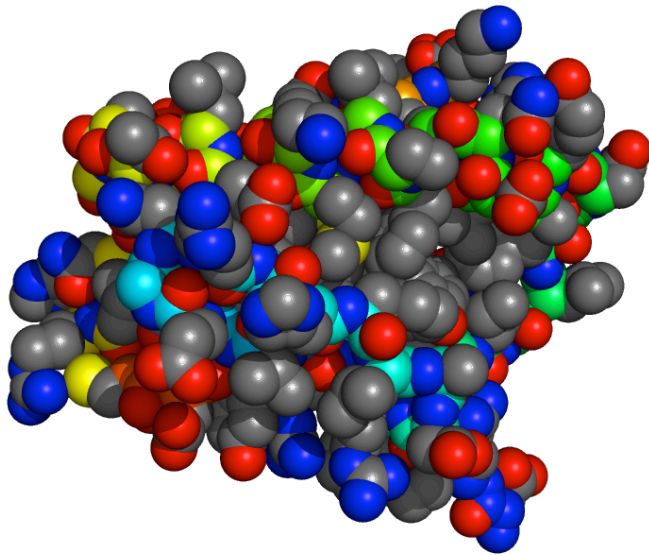
Crystallographic model



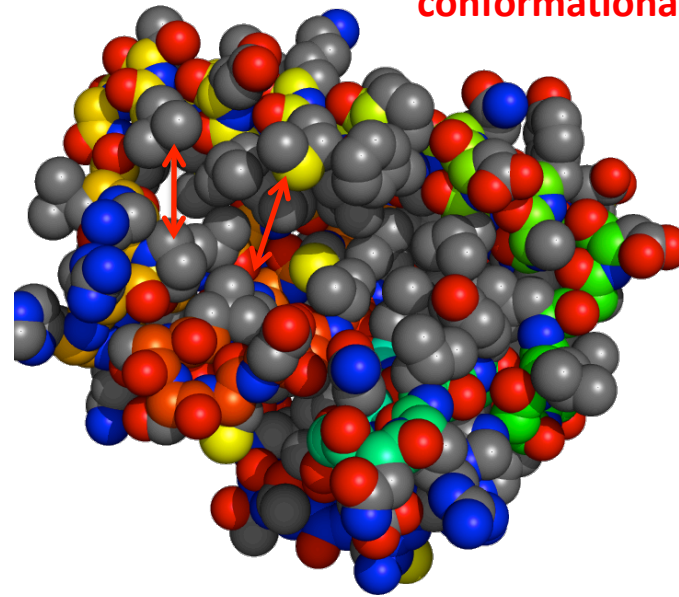
Best CASP model

T304 (CASP7)

Can you pick out the right one?



Crystallographic model



A symptom of poor
conformational sampling

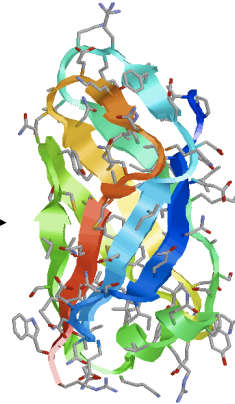
Best CASP model

T304 (CASP7)

Two fundamental problems

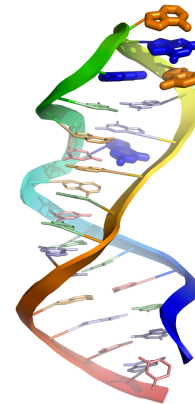
1. Predicting protein structure

GTPDIIVNAQINSEENVLDF
IIEDEYYLKKRGVGAHIKVAS
SPQLRLLYKNAYSTVSCGNYG
VLCNLVQNGEYDLNAIMFNC
AEIKLNKGQMLFQTKIWR



2. Predicting RNA structure

ugcuccuaguacgag
aggaccggagug

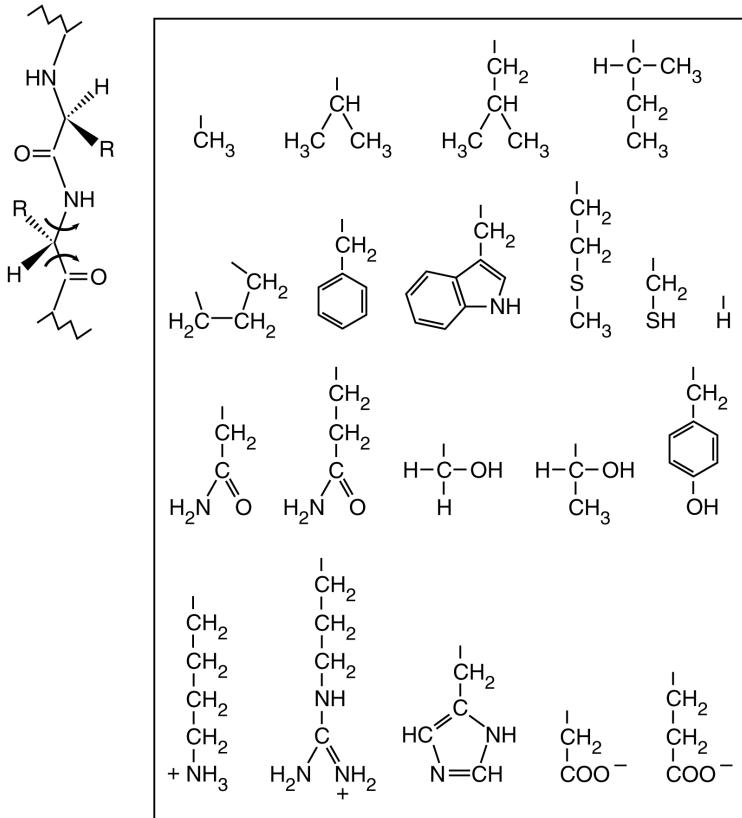


Proteins

RNA

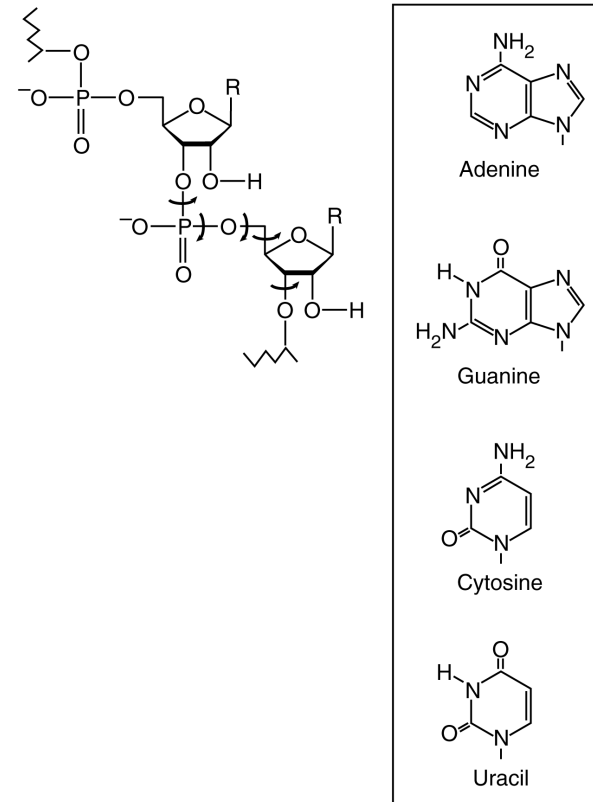
Backbone

Side Chains

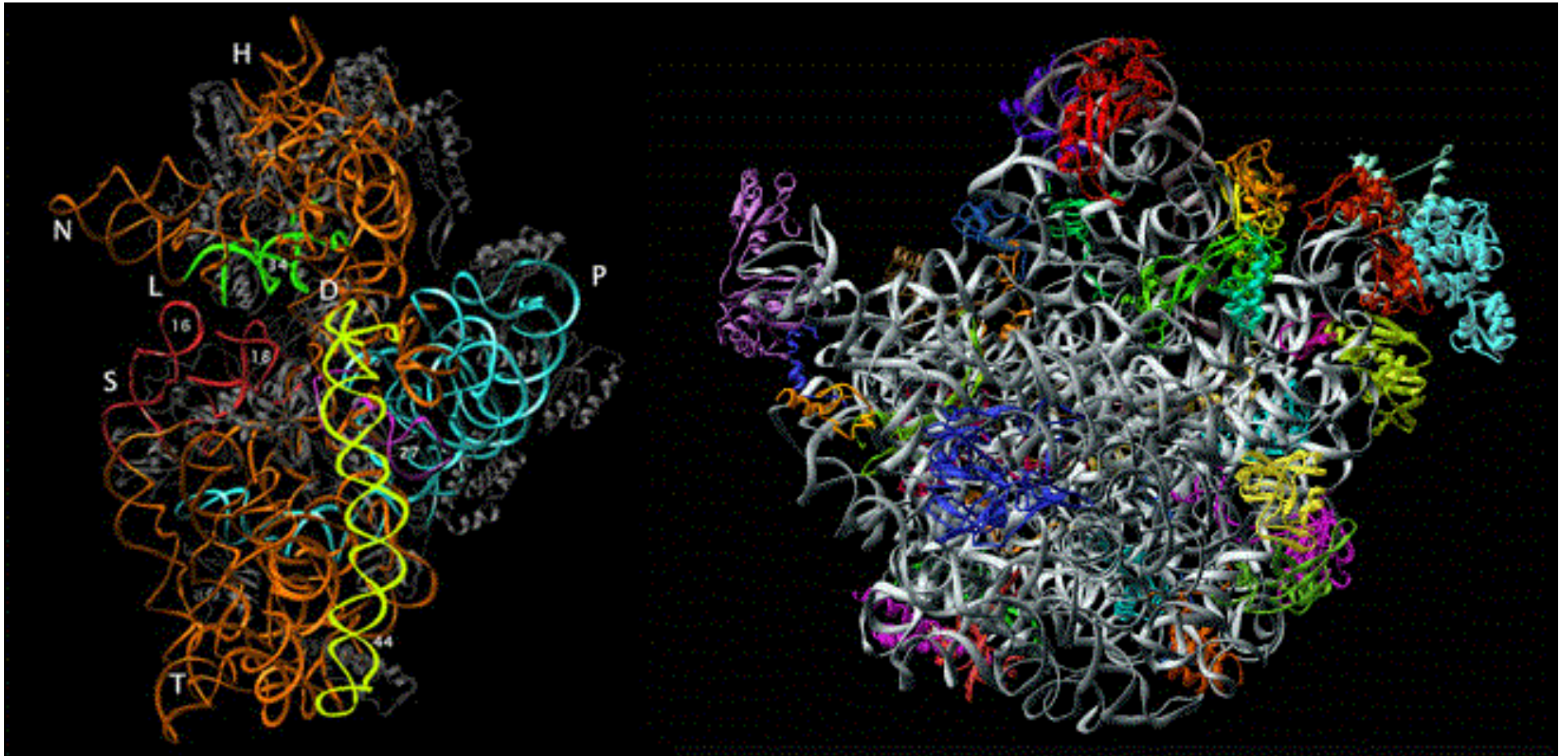
$$R \equiv$$


Backbone

Side Chains

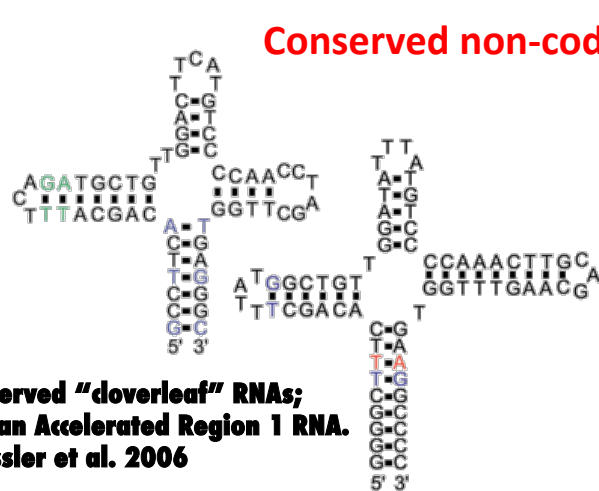
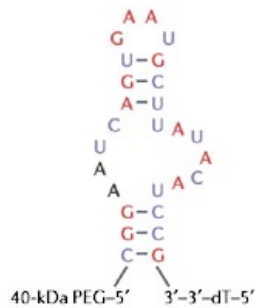
 $R \equiv$ 

How a physicist got into biochemistry (2000)

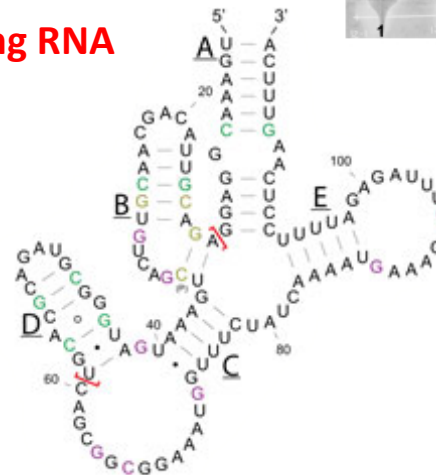


A flourishing RNA world

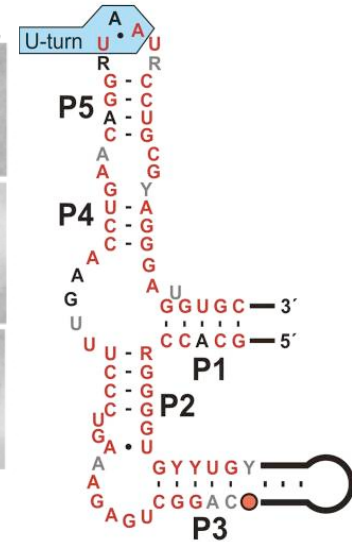
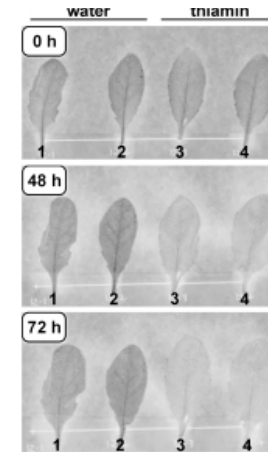
Engineered ribozymes and aptamers



Conserved non-coding RNA



"Riboswitches"

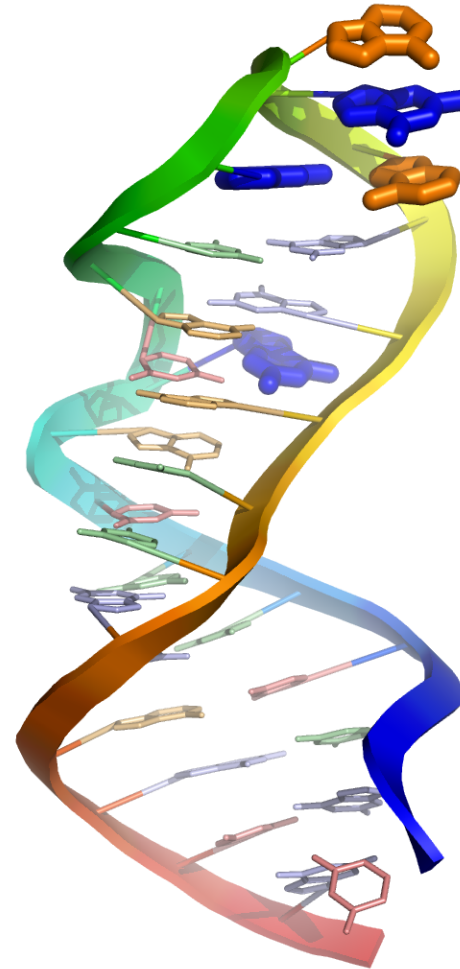


Breaker & colleagues, 2007

The Das Lab

Goal: Nucleic Acid Structures You Can Trust

ugcuccu
aguacga
gaggacc
ggagug

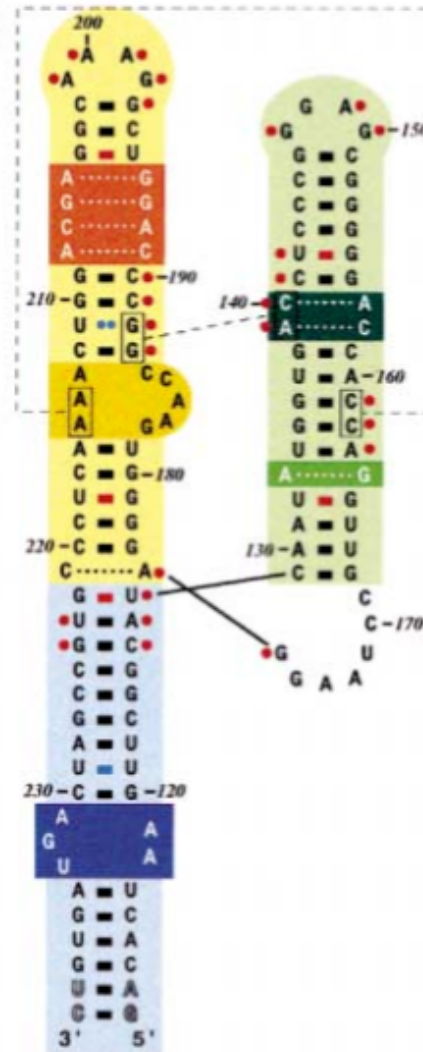


With *de novo* protein
structure modeling as an
inspiration, how far can we
get with computers?

Words and grammar for RNA?

GACACUAAGUUCGGCA
UCAUAUGGUGACCUC
CCGGGAGCGGGGGGACC
ACCAGGUUGCCUAGAG
GGGUGAACCGGCCCAG
GUCGGAAACGGAGCAG
GUCAAAACUCCCGUGC
UGAUCAGUAGUGU

Signal Recognition Particle RNA
Oubridge et al., 2002

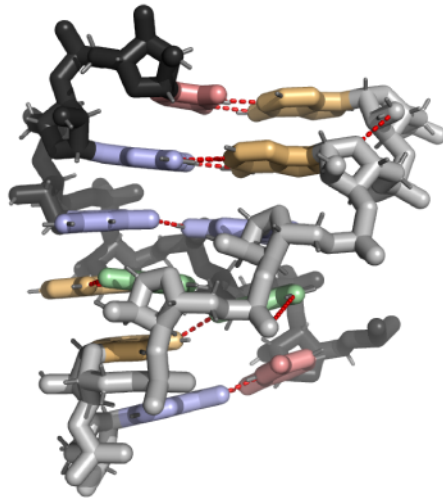
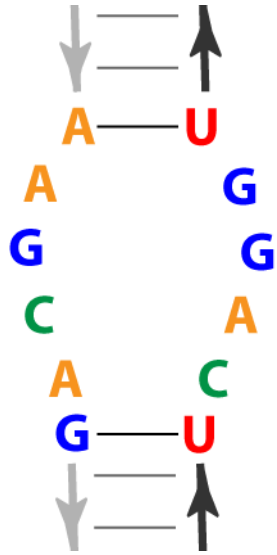


Words and grammar for RNA?

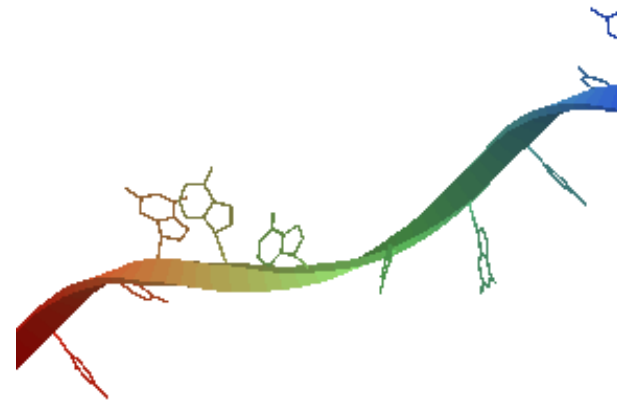
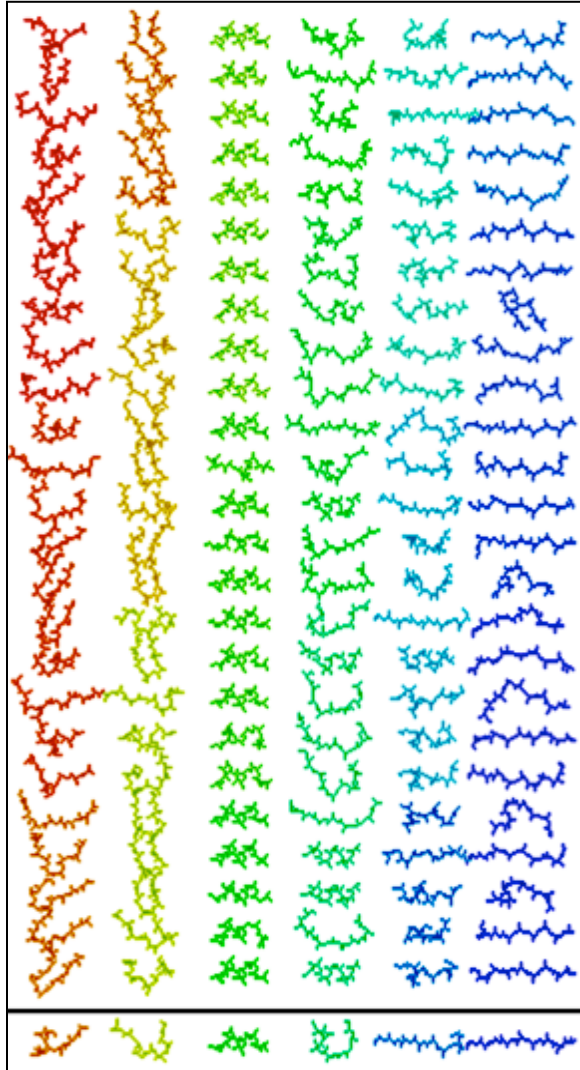
Canonical double helices
Non-canonical regions



Words and grammar for RNA?

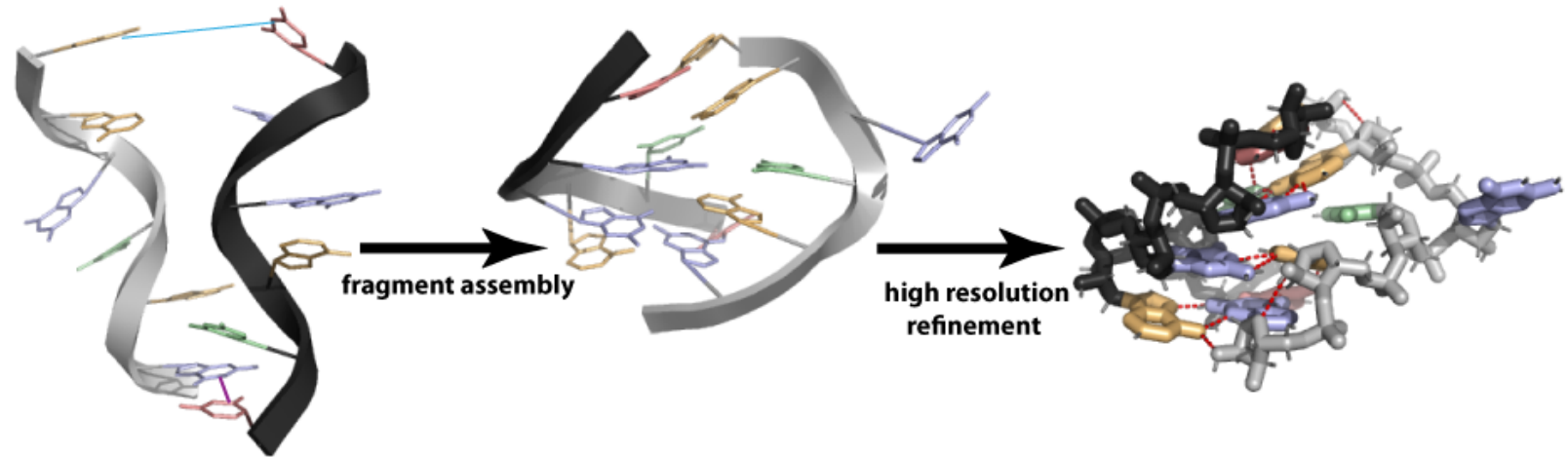


De novo modeling



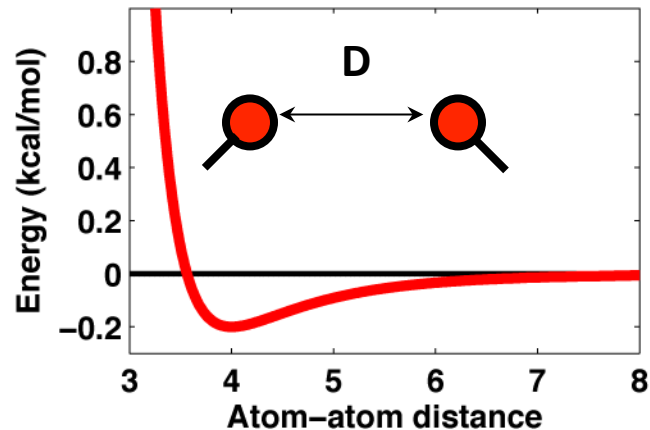
Fragment Assembly of RNA (FARNA)

De novo modeling

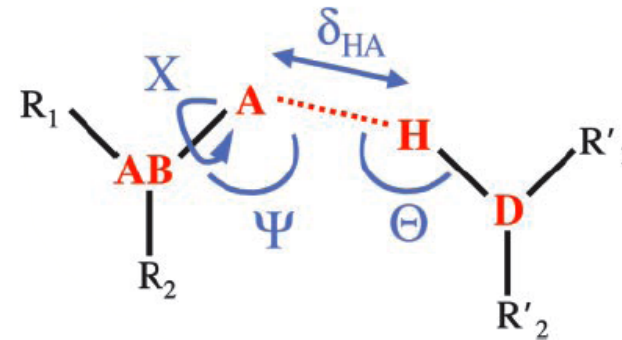


Ingredients of a high resolution potential

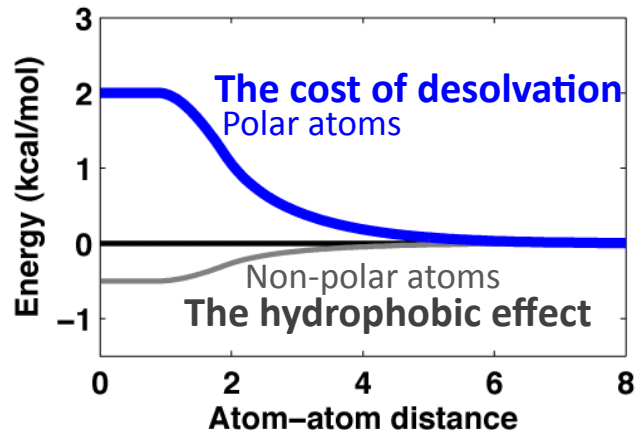
1. Van der waals packing



2. Hydrogen bonds



3. Manifestations of water



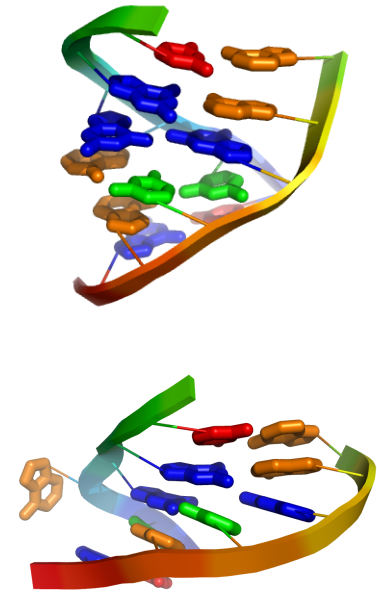
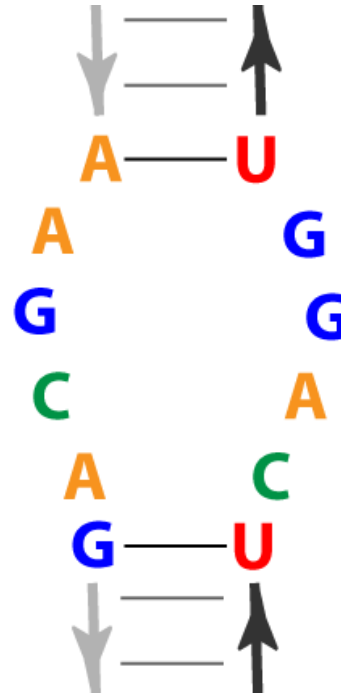
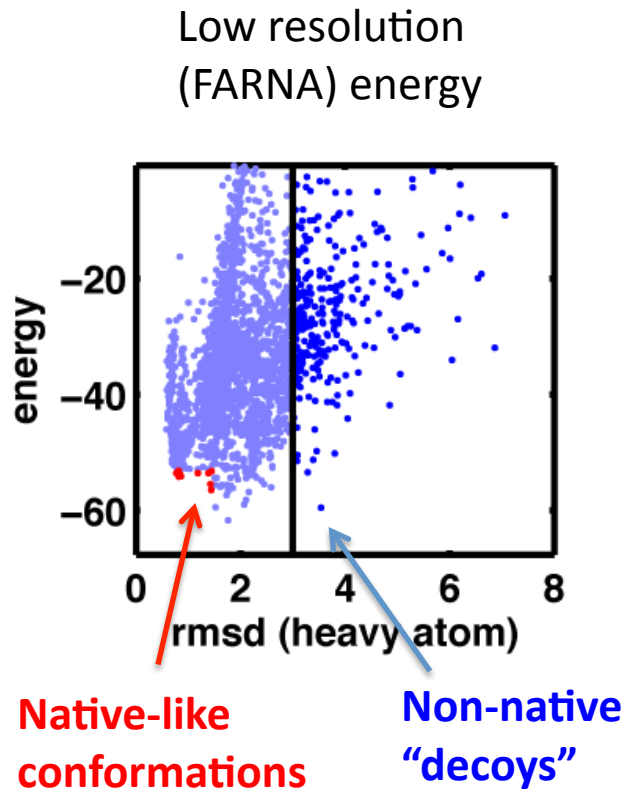
the clustering of non-polar groups, the enhanced hydrogen bond attraction in a hydrophobic environment and the access of water molecules to those polar groups which are not involved in hydrogen bonds.

– Michael Levitt
“Detailed molecular
model of transfer
RNA”, *Nature* 1969.

Does it work?

Native-state discrimination

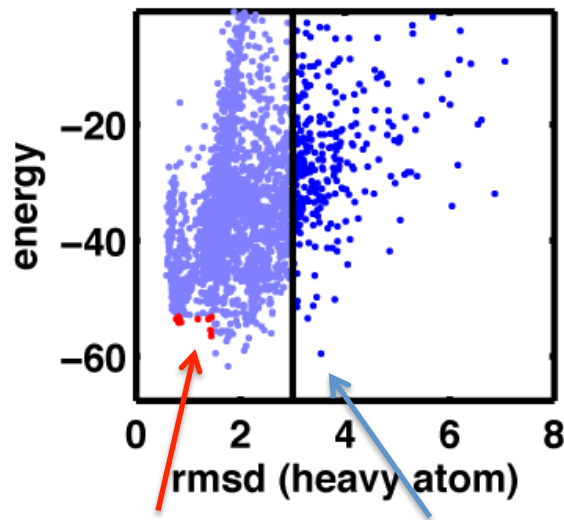
The most conserved region of the
signal recognition particle



Native-state discrimination

The most conserved region of the
signal recognition particle

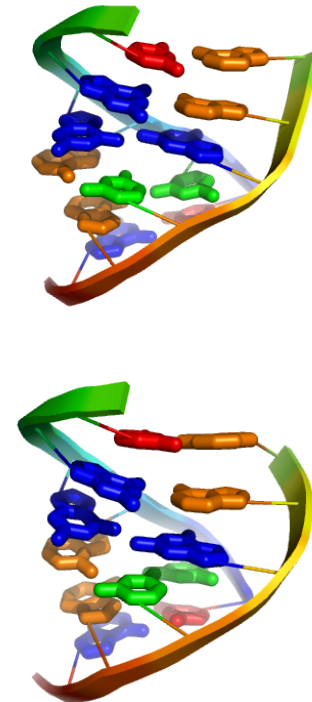
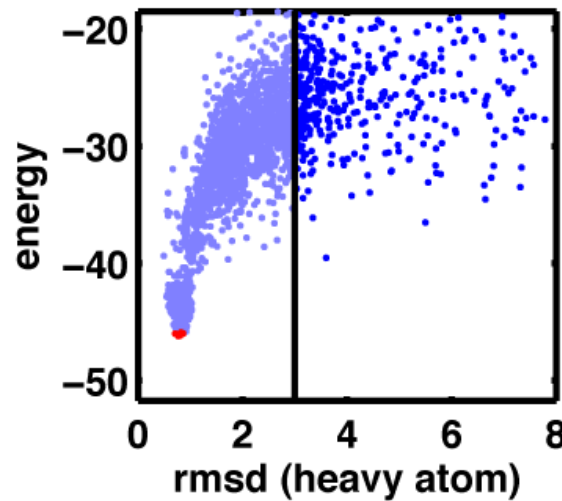
Low resolution
(FARNA) energy



Native-like
conformations

Non-native
"decoys"

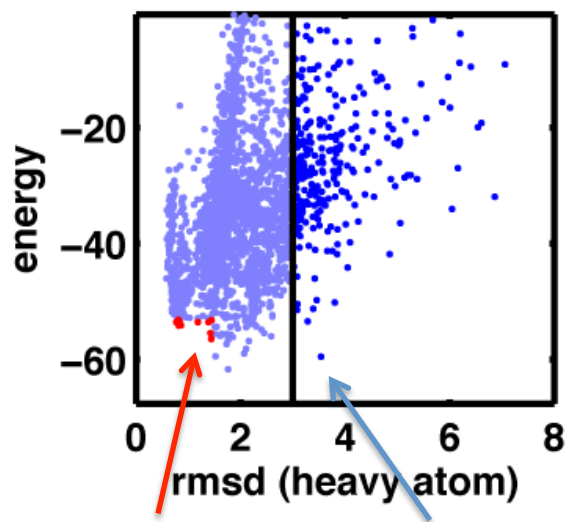
High resolution
energy



Native-state discrimination

The most conserved region of the
signal recognition particle

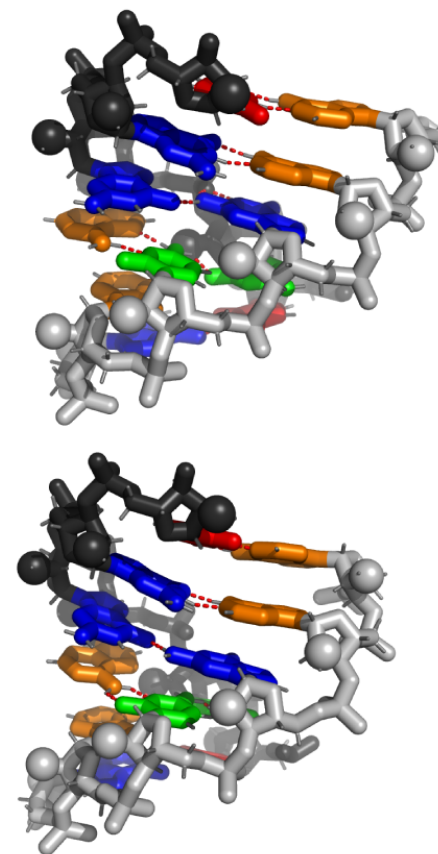
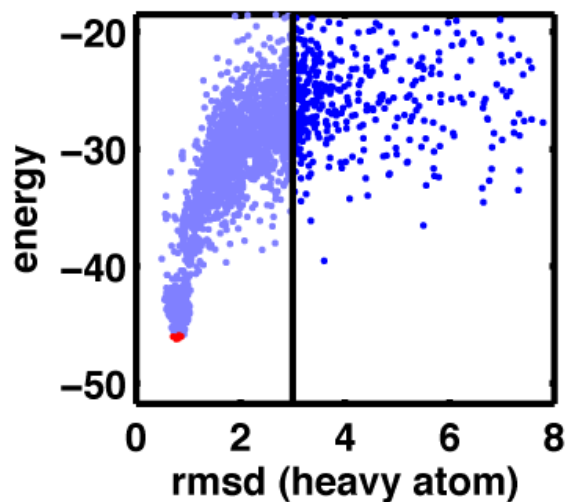
Low resolution
(FARNA) energy



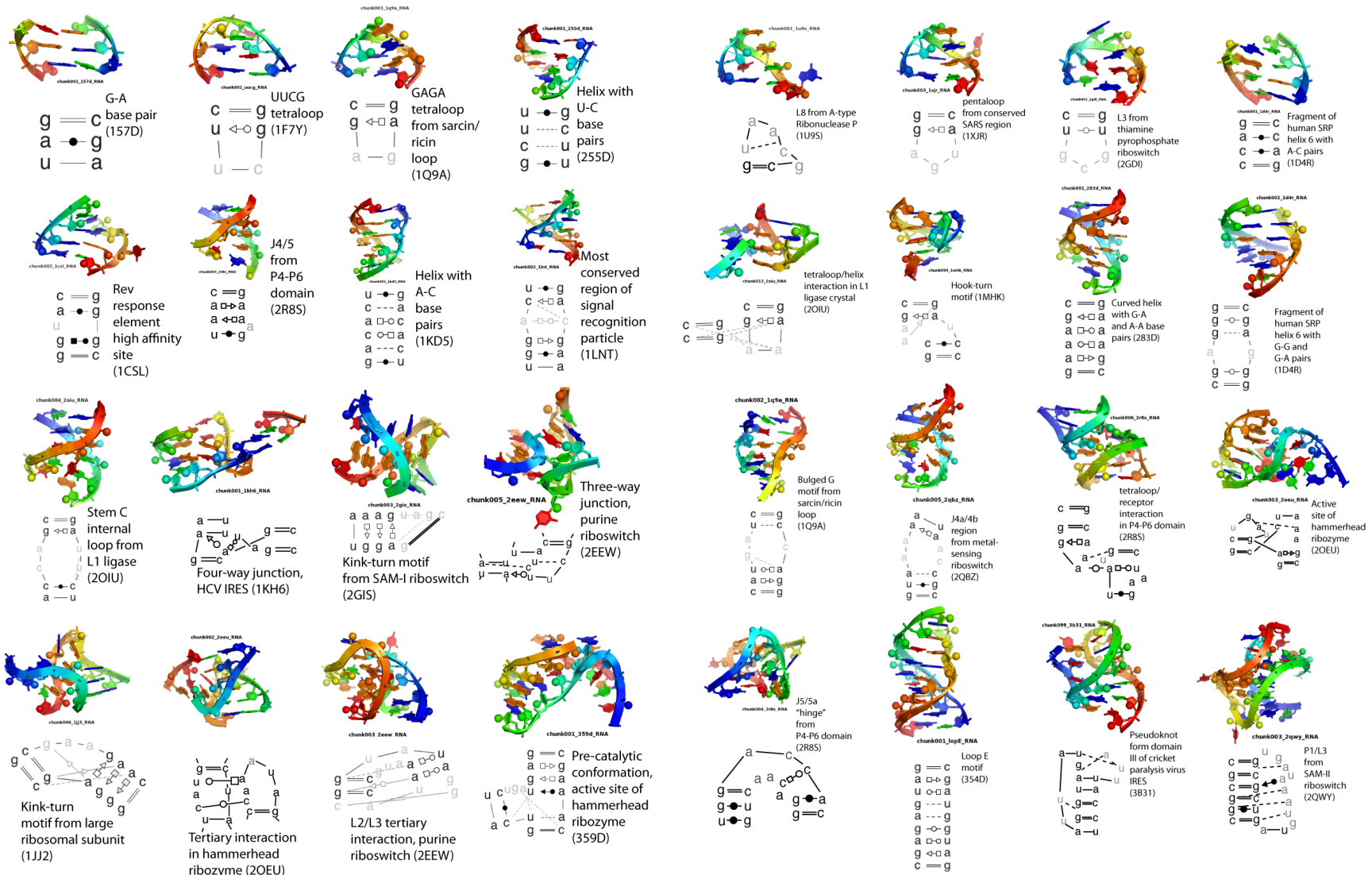
Native-like
conformations

Non-native
“decoys”

High resolution
energy

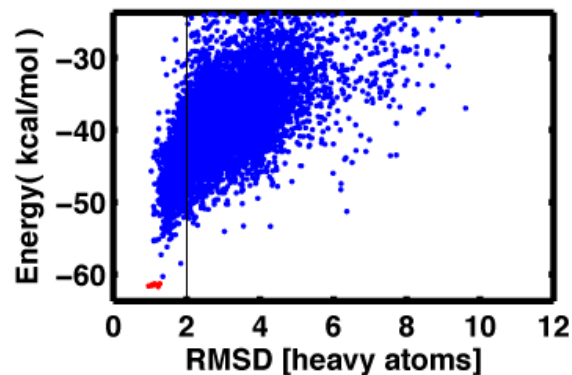
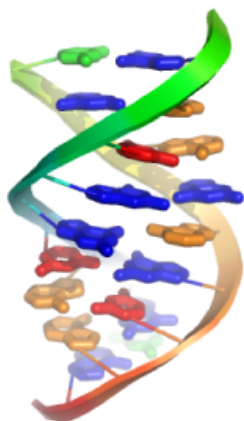
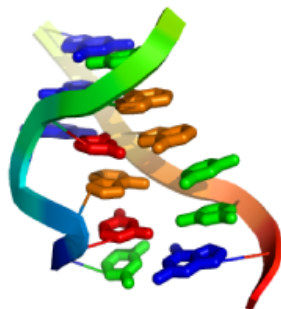
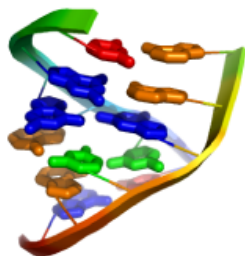


Can we decipher all the known “words”?

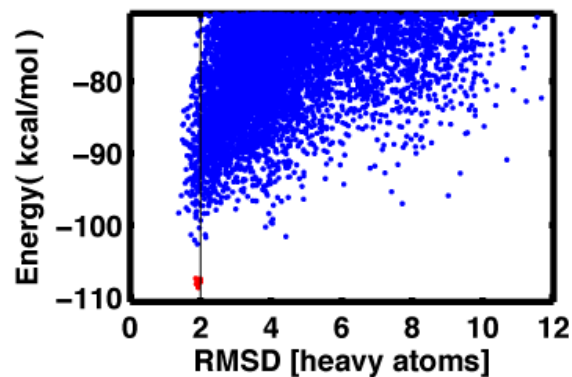
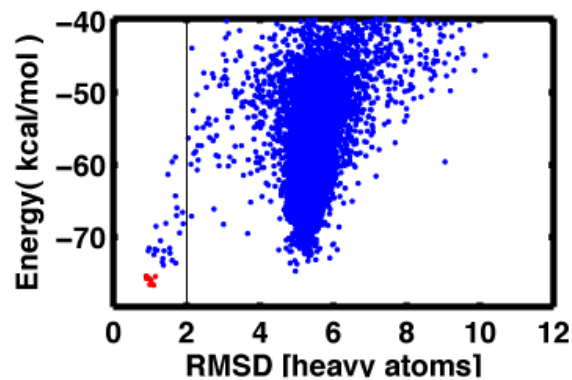


De novo modeling

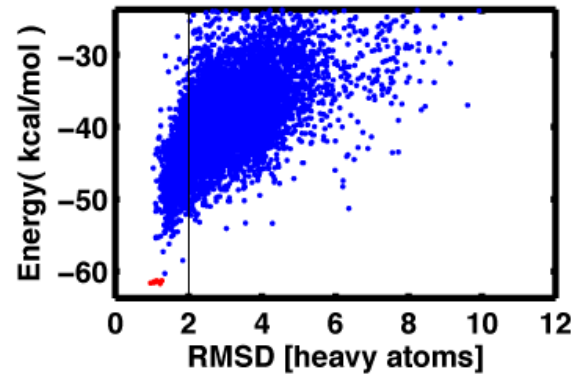
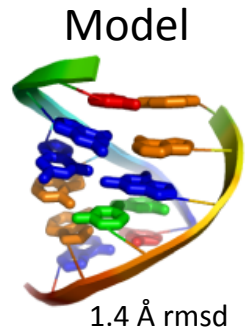
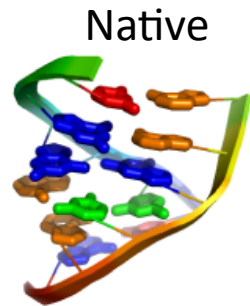
Native



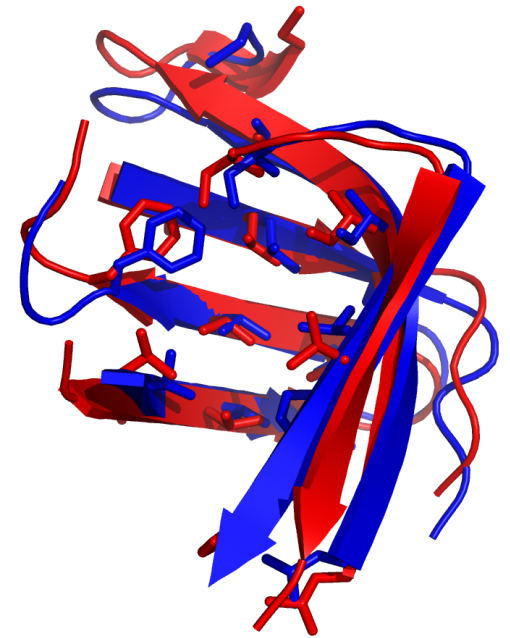
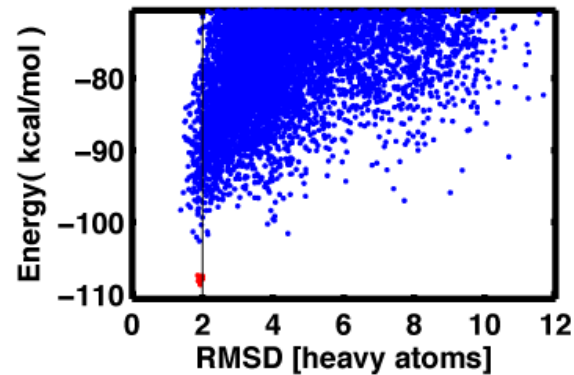
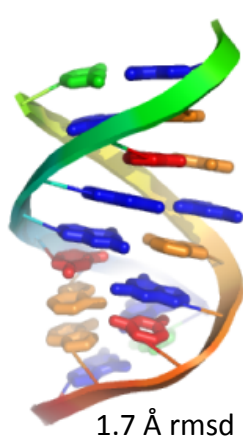
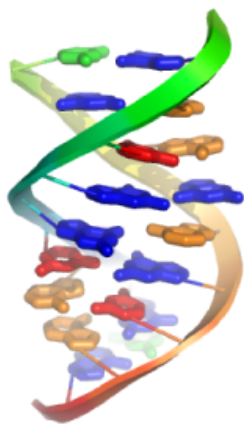
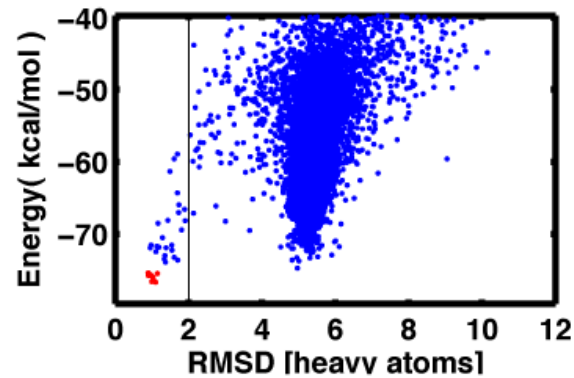
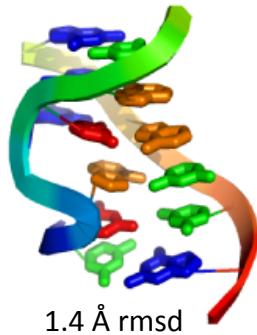
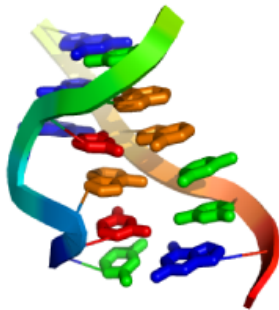
In half the cases, de novo modeling achieves $< 2.0 \text{ \AA}$ structures, and selects them.



De novo modeling

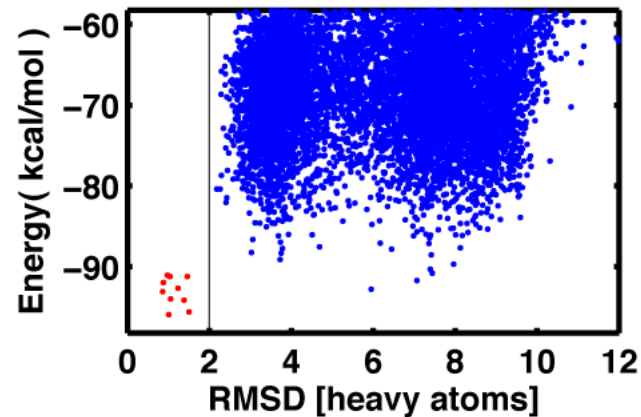
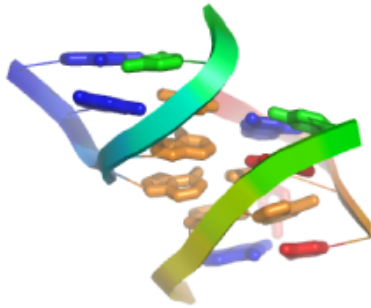


In half the cases, de novo modeling achieves < 2.0 Å structures, and selects them.



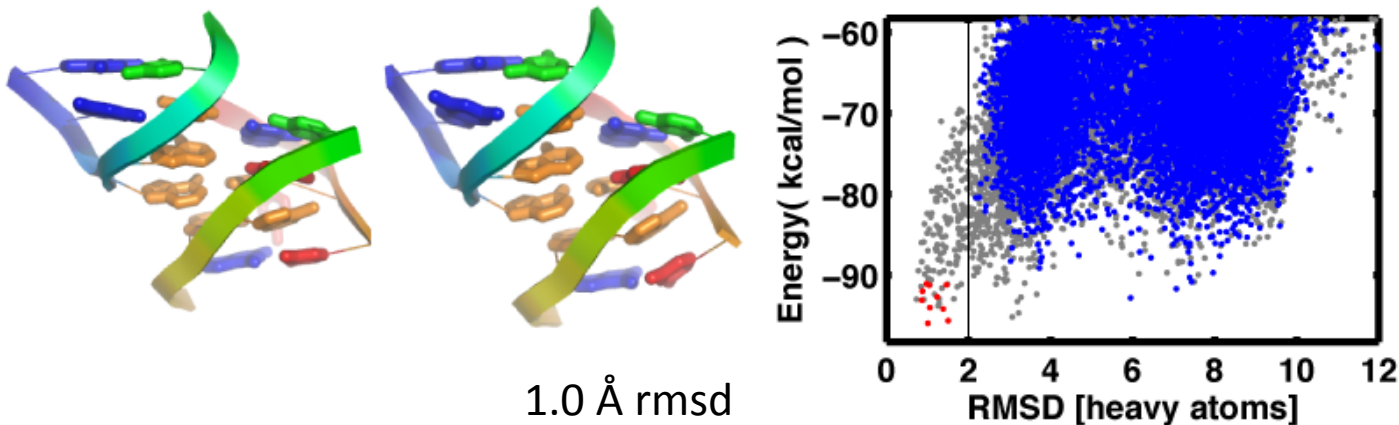
De novo modeling

The biggest bottleneck: conformational sampling



De novo modeling

The biggest bottleneck: conformational sampling



We know the rules of the game
(but we have to play it better)

A universal obsession

domains (<85 residues). The primary bottleneck to consistent high-resolution prediction appears to be **conformational sampling**.

Toward High-Resolution de Novo Structure Prediction for Small Proteins

Philip Bradley, Kira M. S. Misura, David Baker*

generated during CASP8. Developing more effective **conformational sampling algorithms** and protocols is a critical area for current research in protein structure prediction.

Structure prediction for CASP8 with all-atom refinement using Rosetta

Srivatsan Raman,¹ Robert Vernon,¹ James Thompson,² Michael Tyka,¹ Ruslan Sadreyev,³ Jimin Pei,³ David Kim,¹ Elizabeth Kellogg,¹ Frank DiMaio,¹ Oliver Lange,¹ Lisa Kinch,³ Will Sheffler,² Bong-Hyun Kim,⁴ Rhiju Das,¹ Nick V. Grishin,^{3,4} and David Baker^{1,2,3*}

to the functional loop regions. However, despite progress in loop prediction methods^{1,2}, design applications are limited by the difficulty in modeling purely local conformational moves and by the **need for advances in sampling** and evaluating loop conformations.

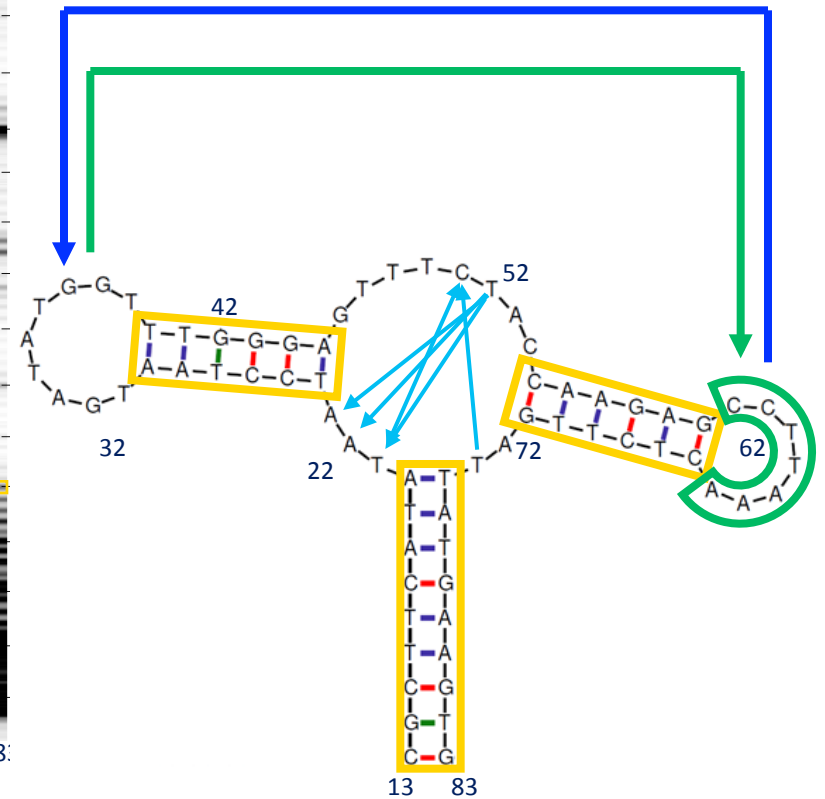
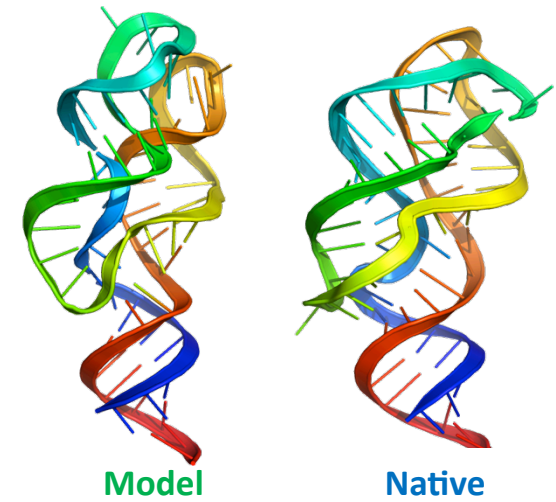
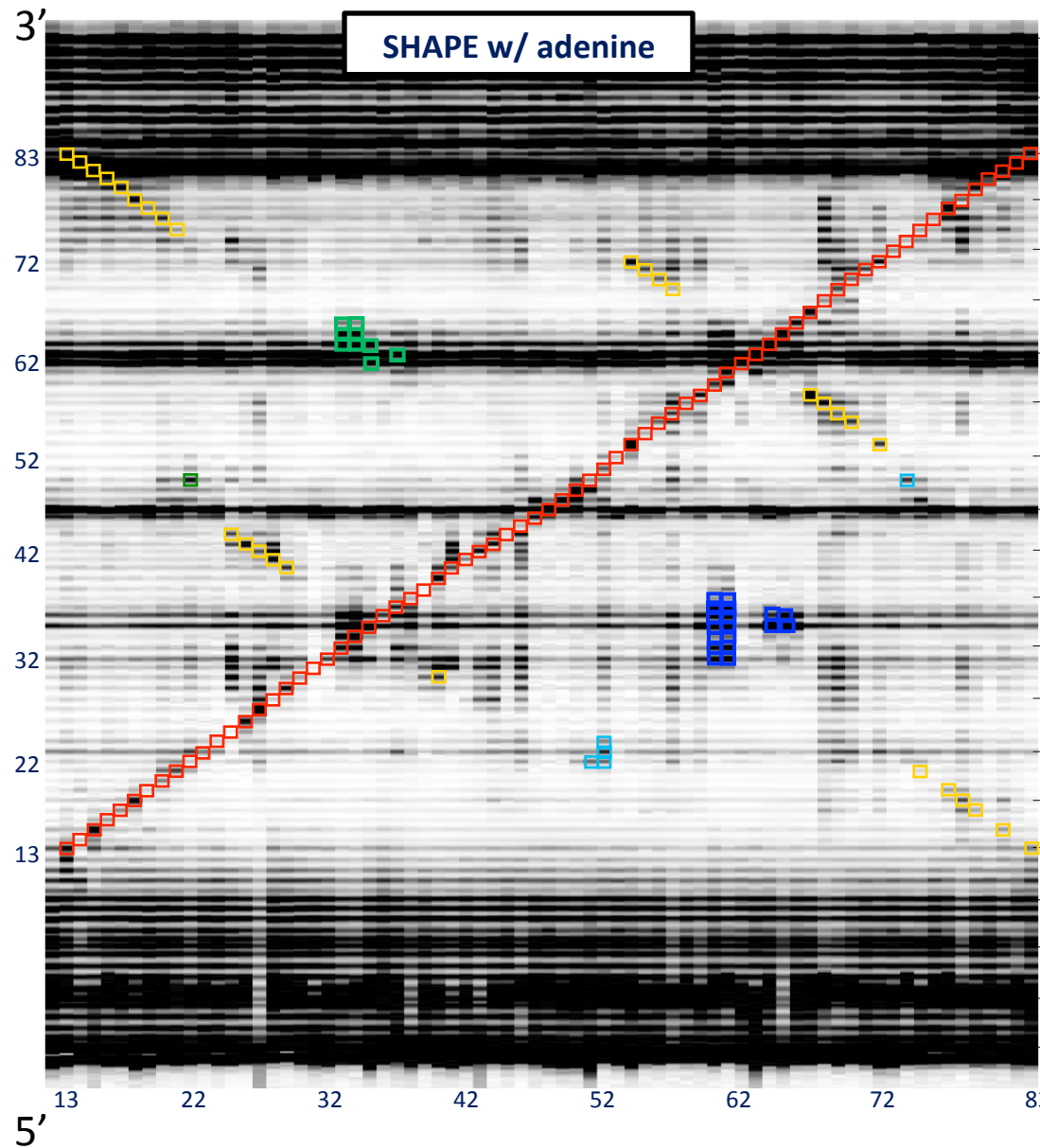
Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling

Daniel J Mandell^{1,2}, Evangelos A Coutsiyas³ & Tanja Kortemme^{1,2,4}

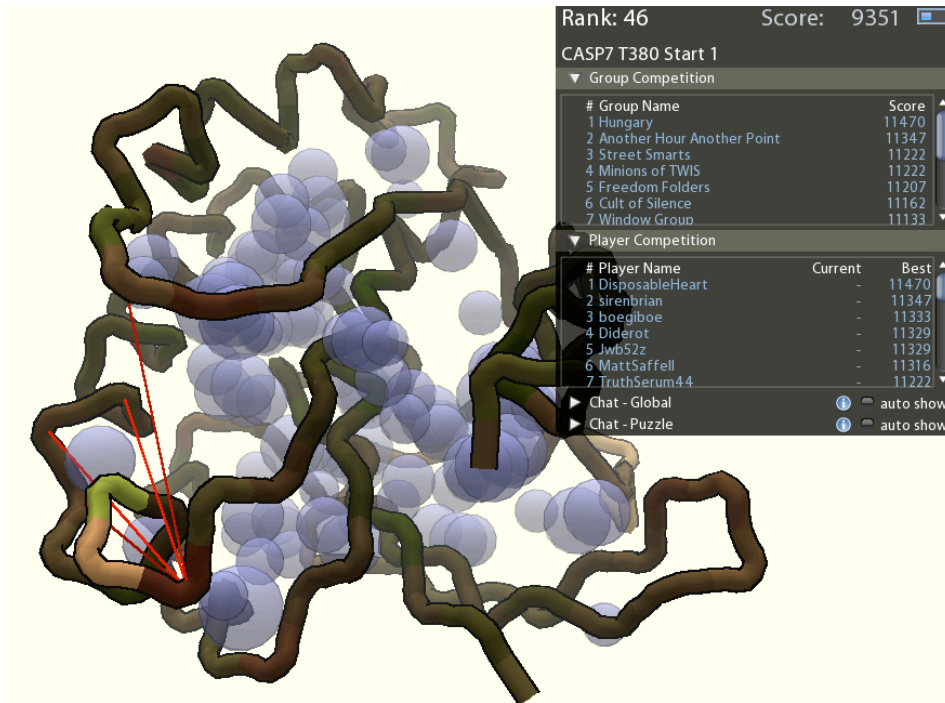


**Beating the “astronomical”
conformational sampling
problem**

Solution 1? Data



Solution 2? Humans



FOLD.IT

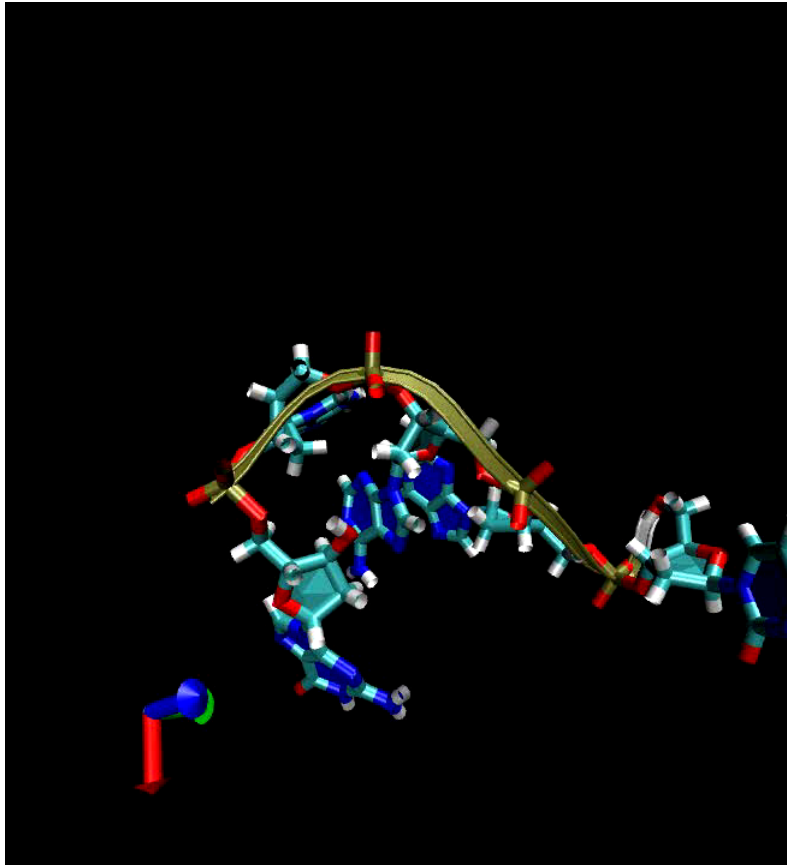
Baker lab With UW Comp. Sci. (Adrien Treuille, Seth Cooper, Zoran Popovic, David Salesin, others...)



ETERNA

With Adrien Treuille (now at Carnegie-Mellon) and Jeehyung Lee

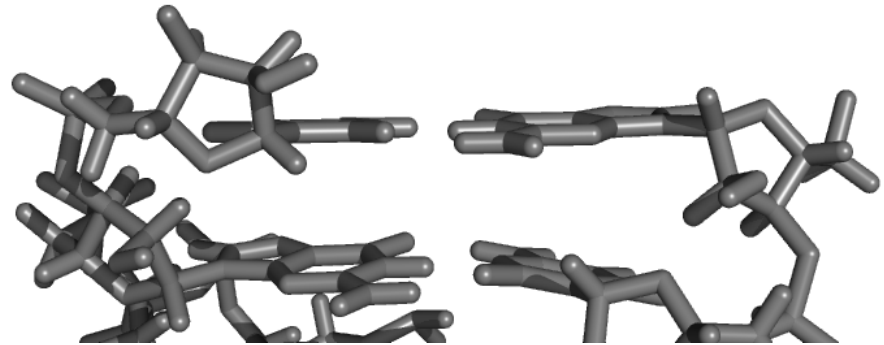
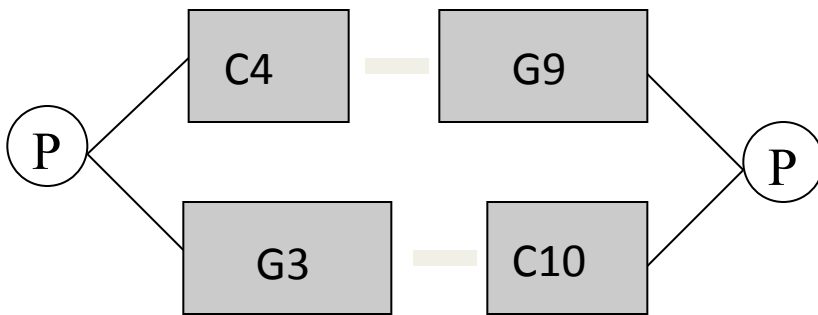
Solution 3? Physics



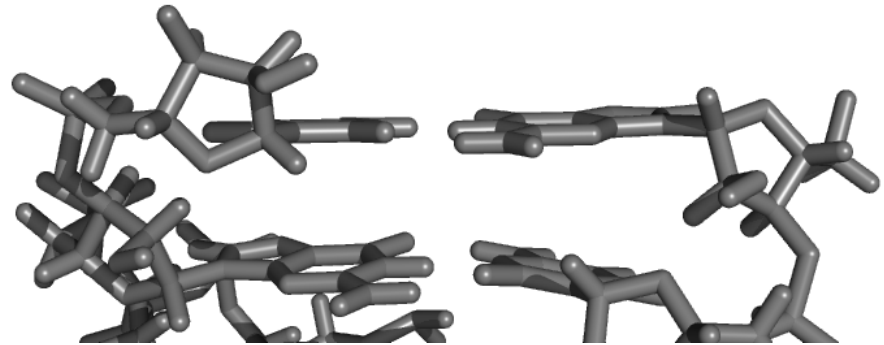
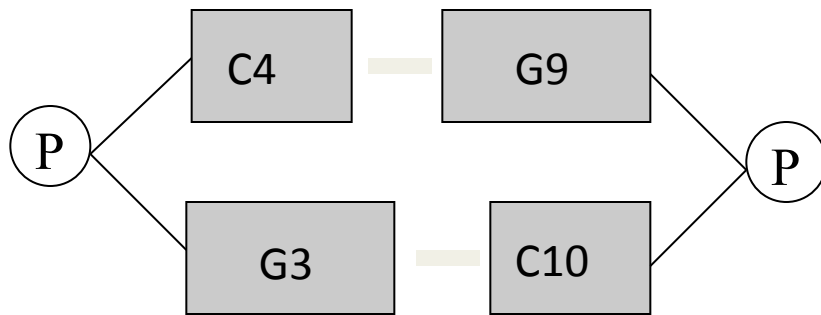
Computationally expensive, but getting faster

Still no case of blind predictions of structure

Solution 4. Better algorithms

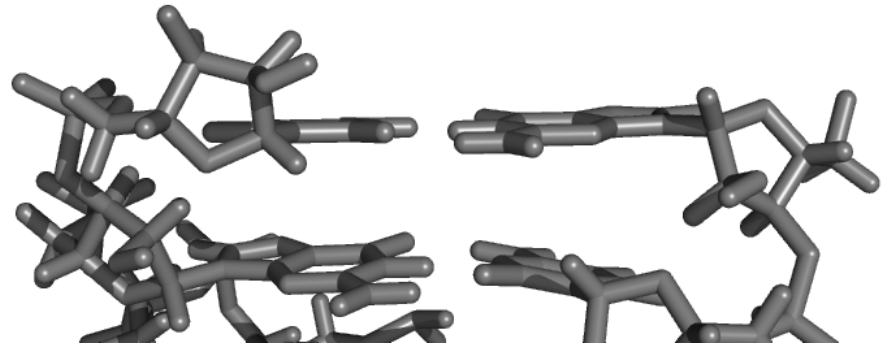
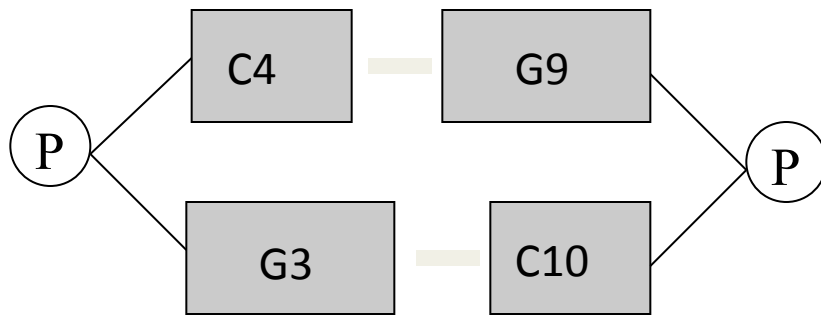


Step-by-step sampling

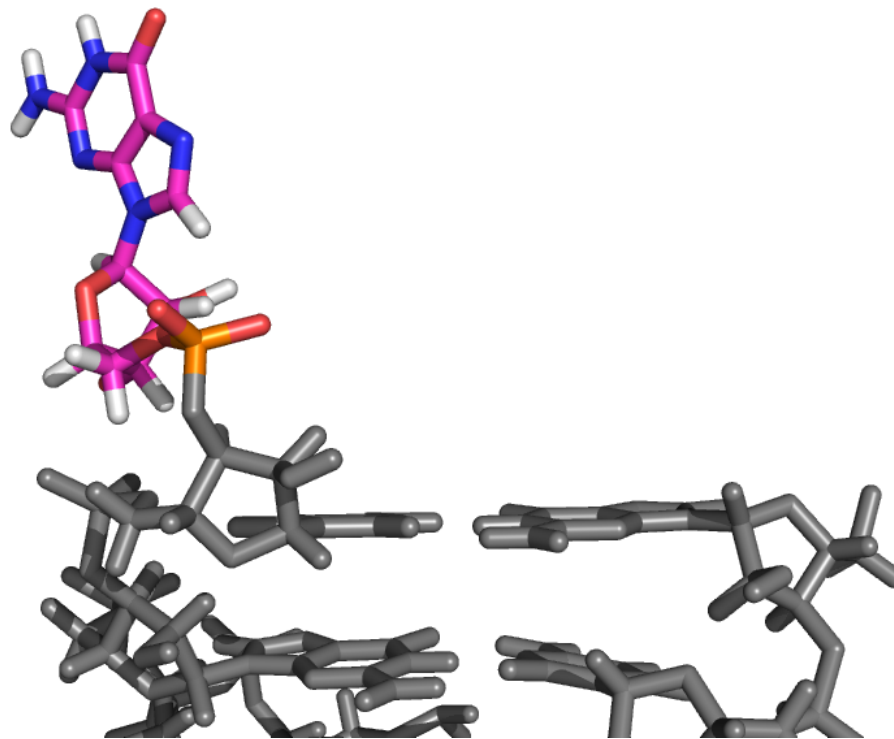
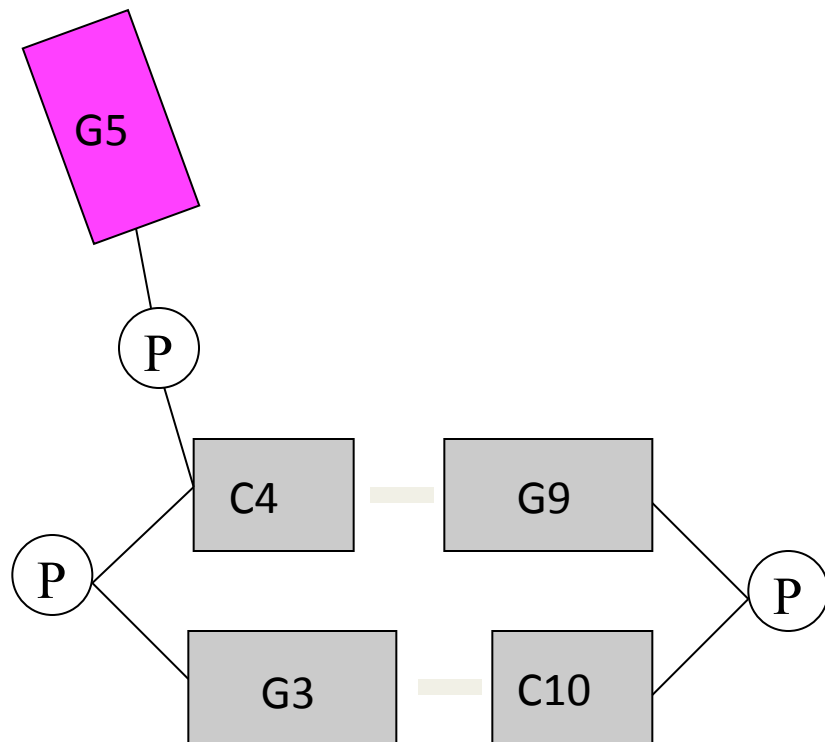


Step-by-step sampling

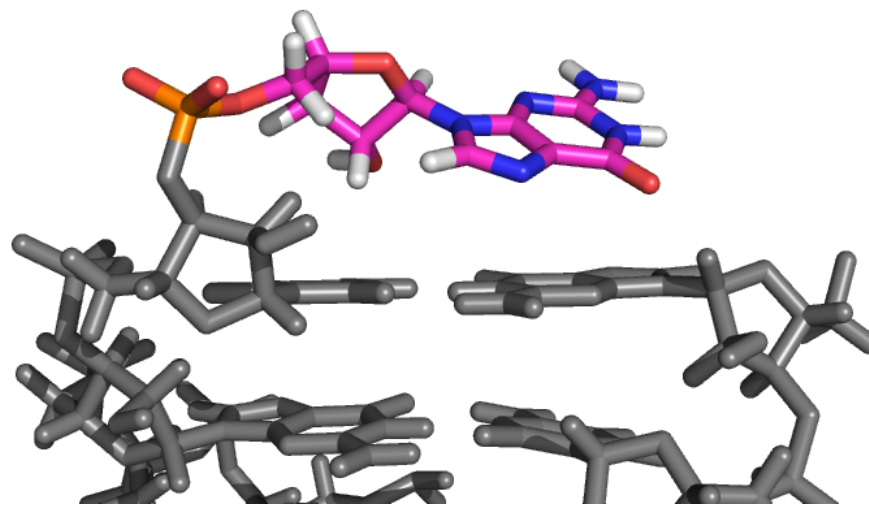
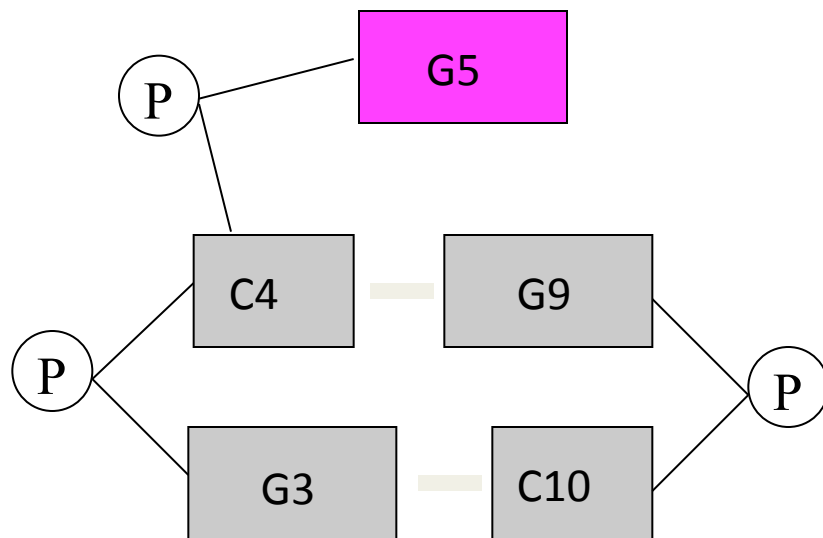
G



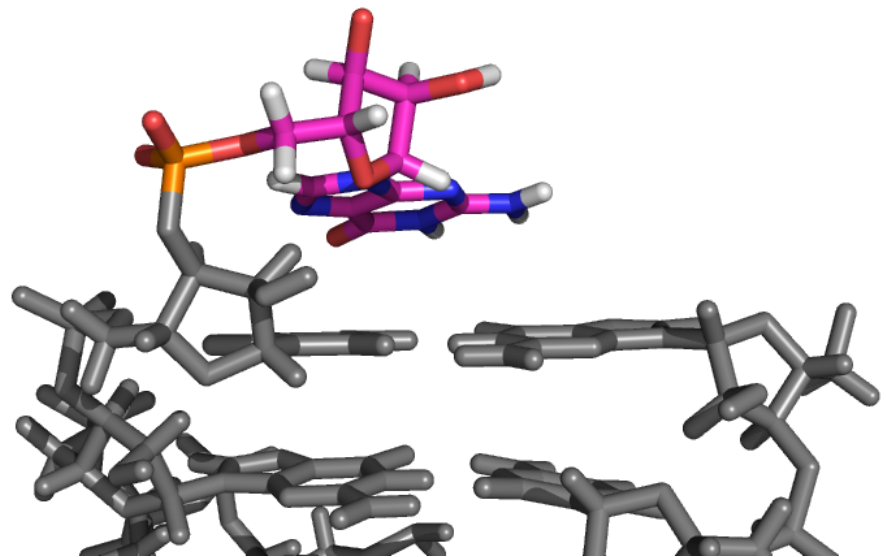
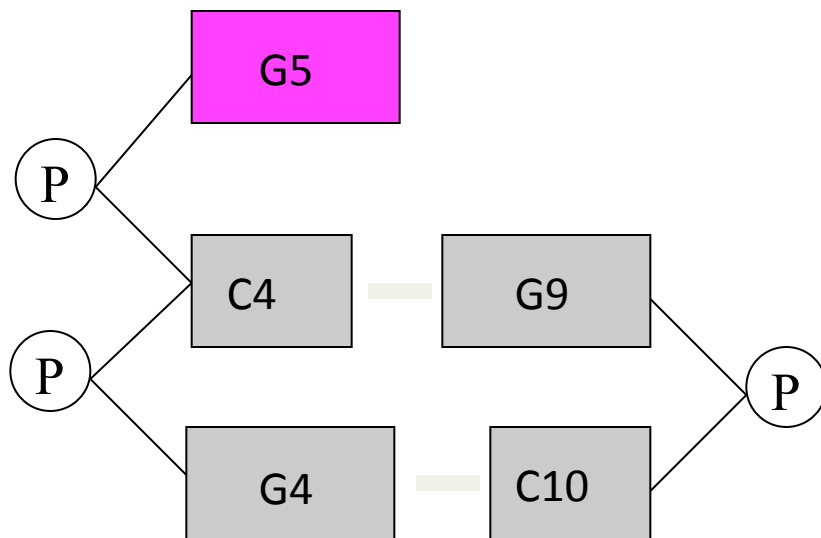
Step-by-step sampling



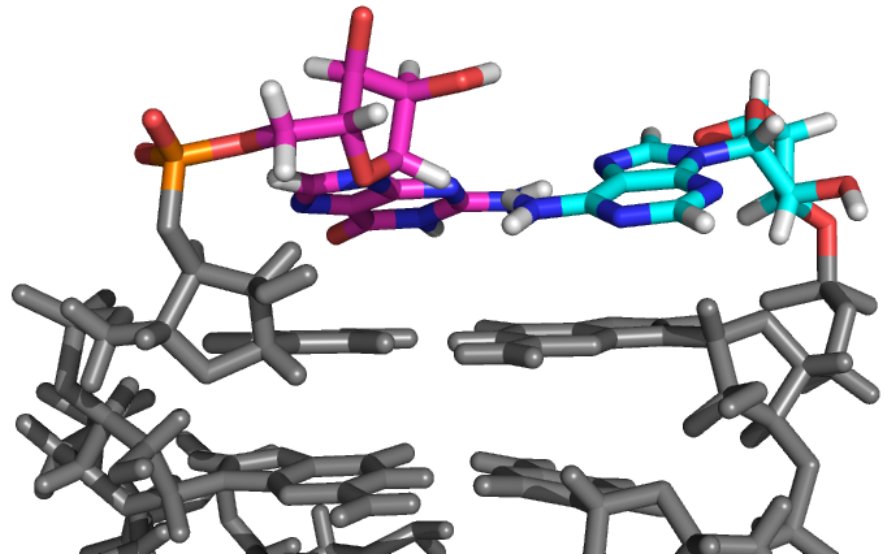
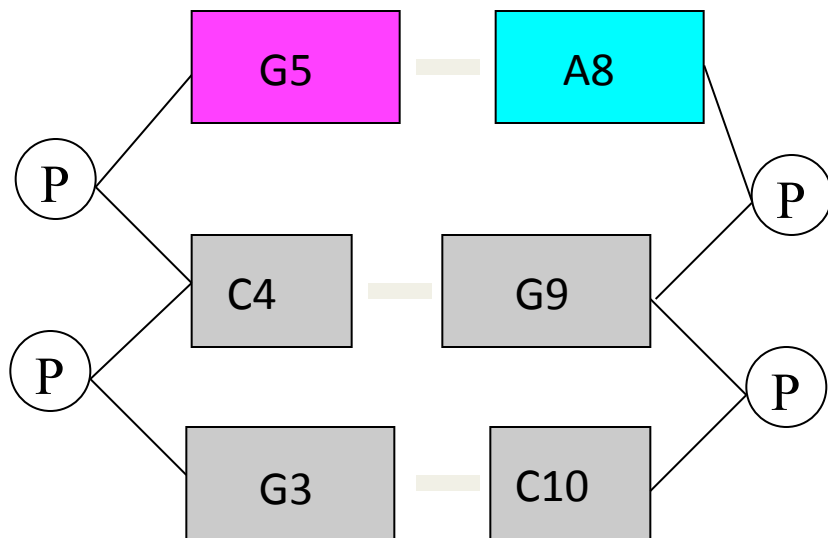
Step-by-step sampling



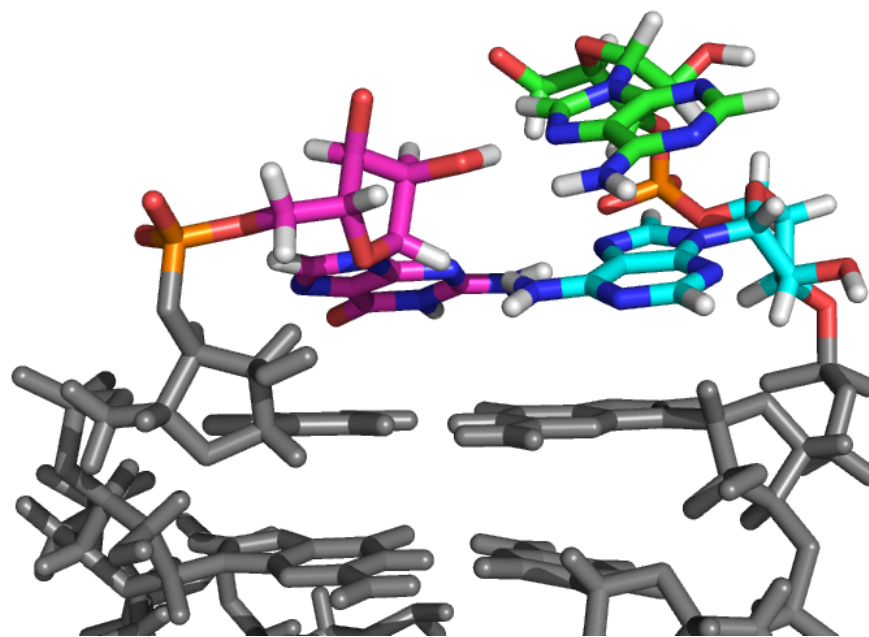
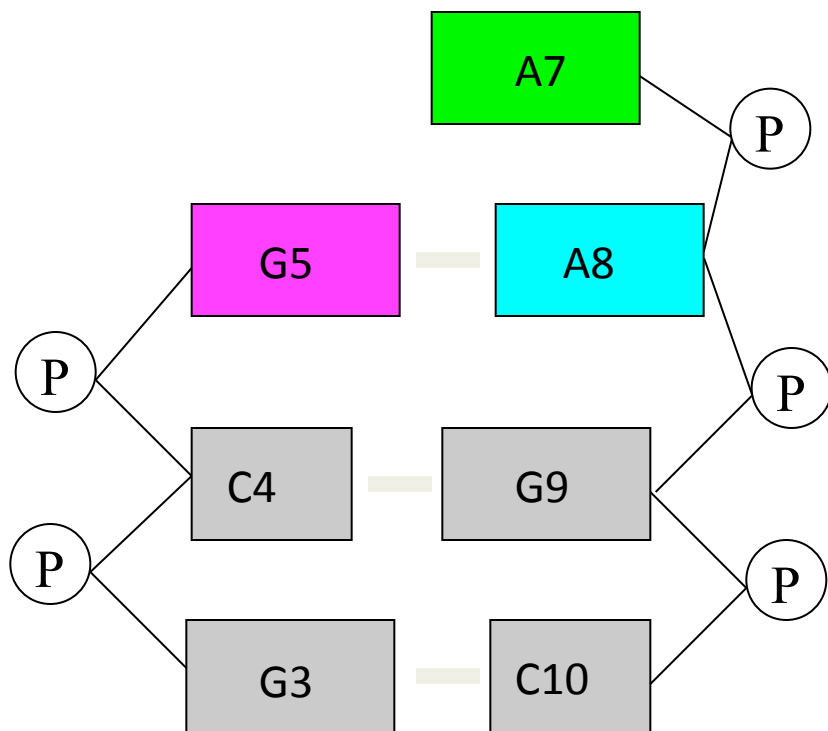
Step-by-step sampling



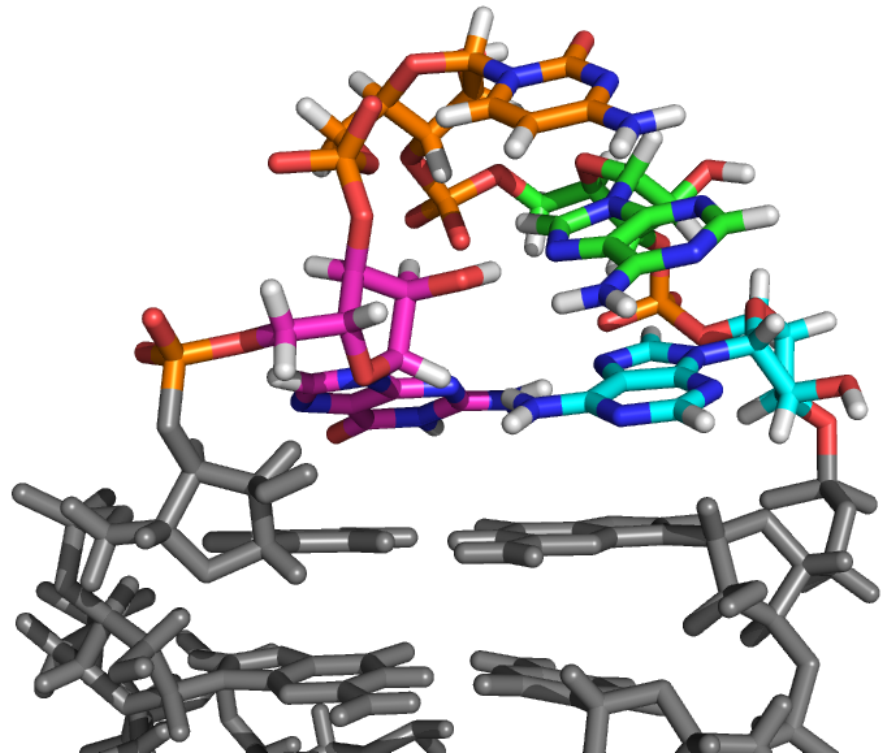
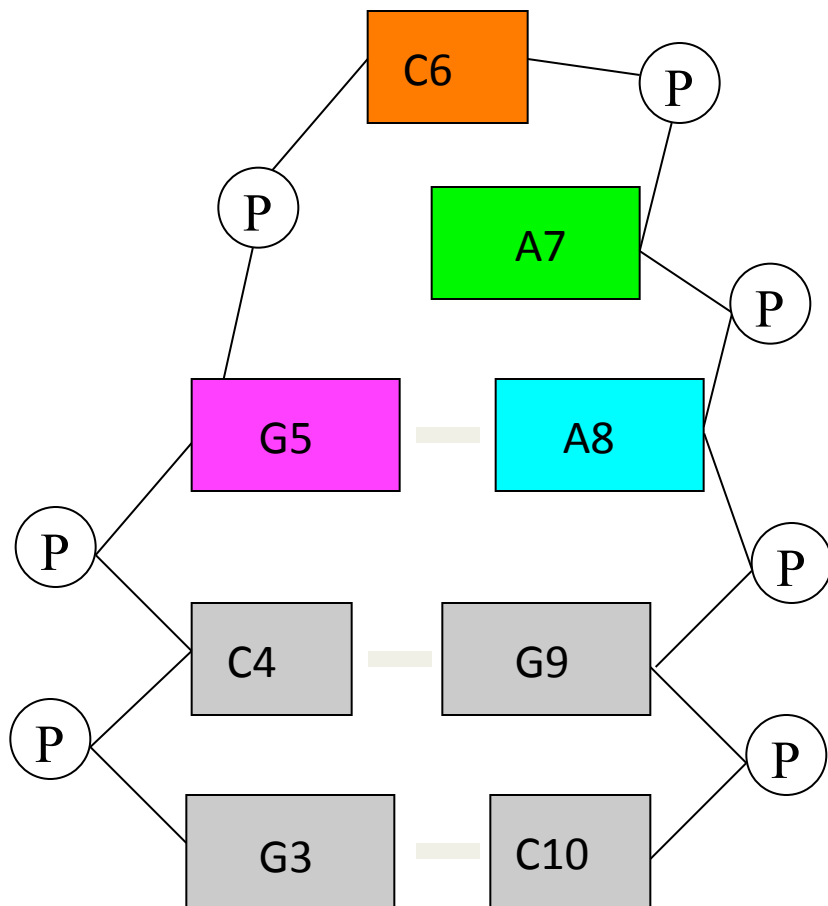
Step-by-step sampling



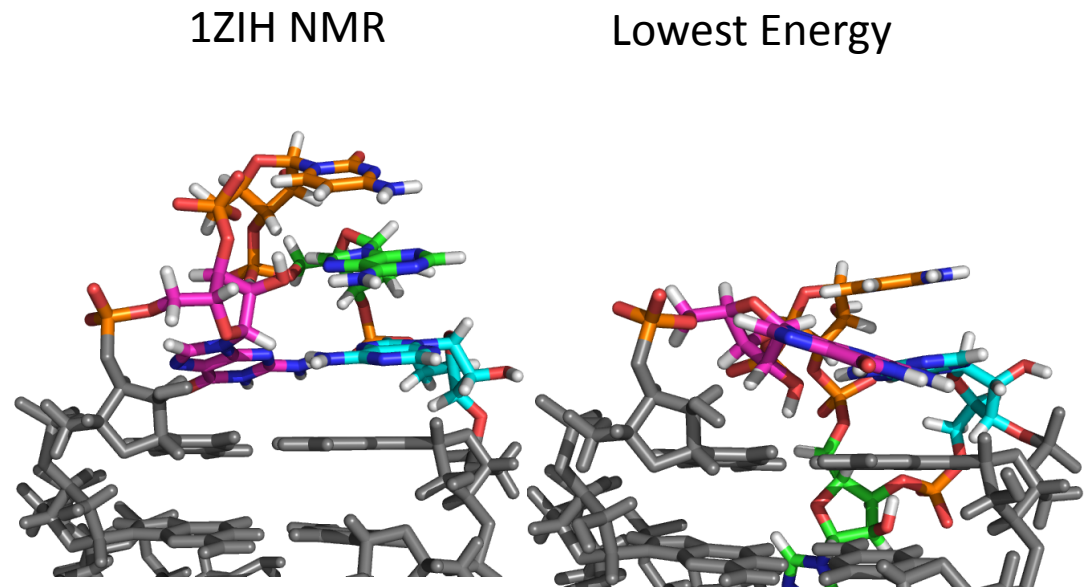
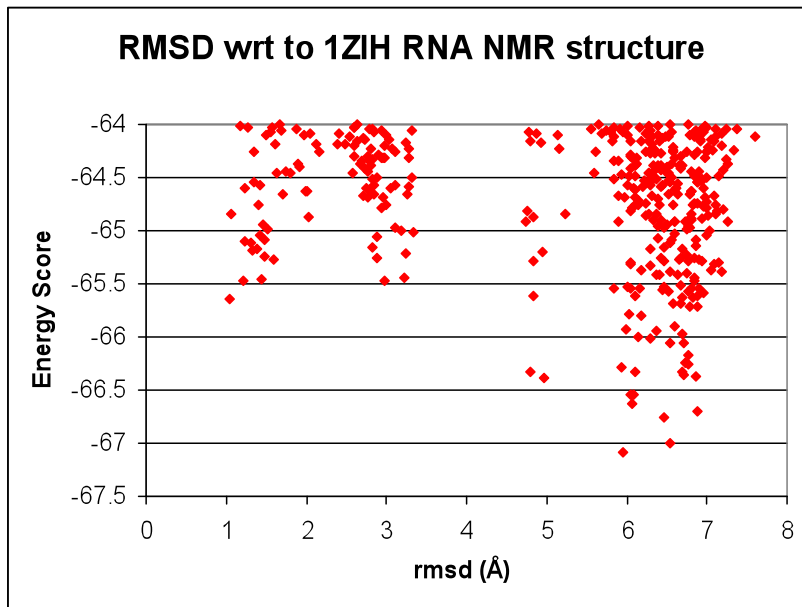
Step-by-step sampling



Step-by-step sampling



Step-by-step sampling

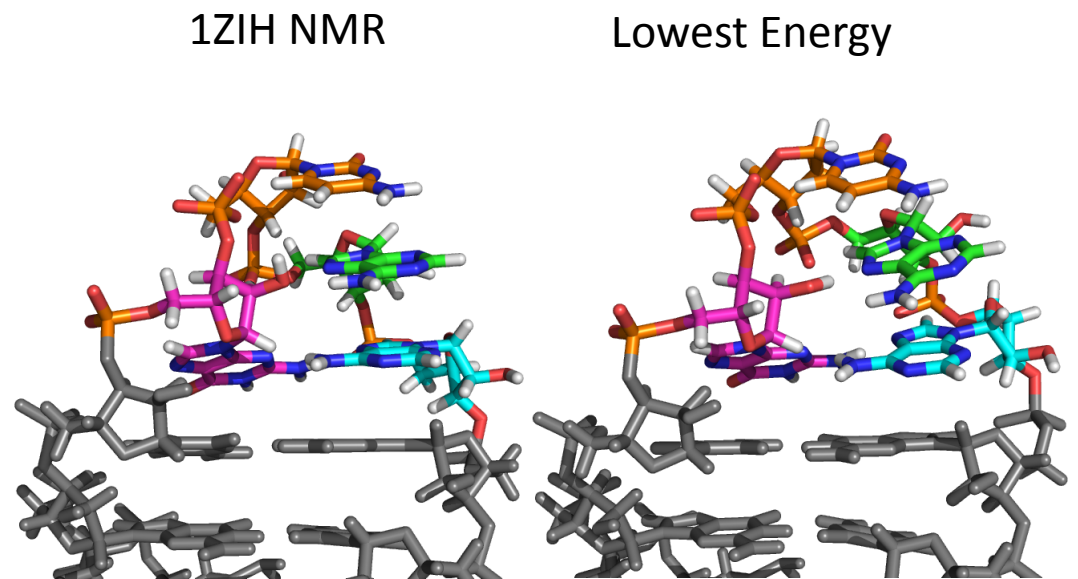
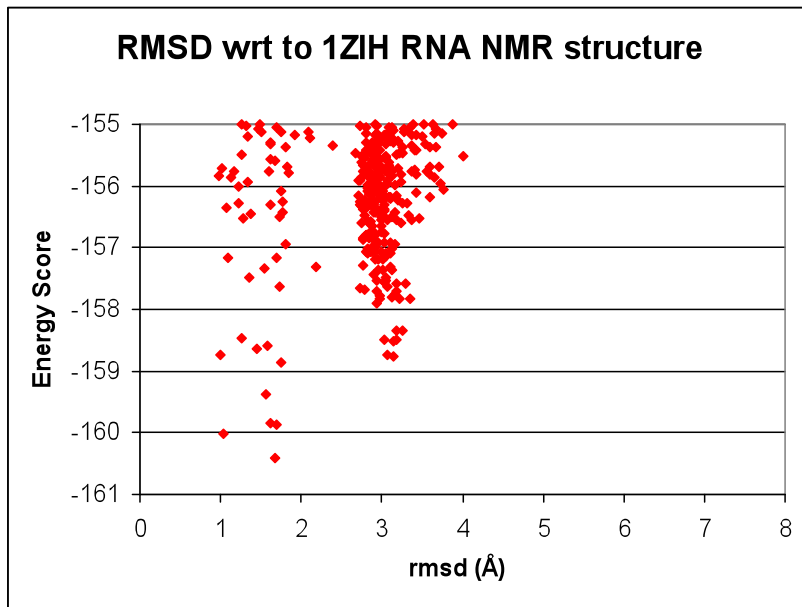


Aha – terms for:

- base stacking
- RNA torsional potential

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

Step-by-step sampling



Wait, there's still a cheat!

There are other pathways (2^N total)

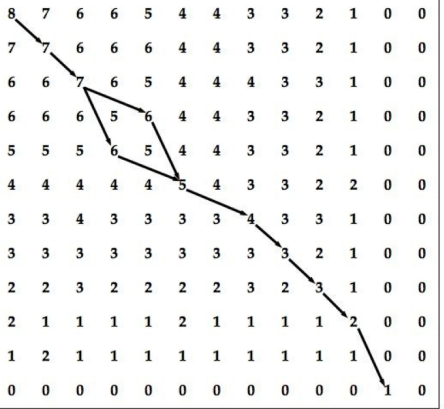
How to sample all paths?

Steal a trick from “dynamic programming”, i.e. recursion.

Sequence alignment

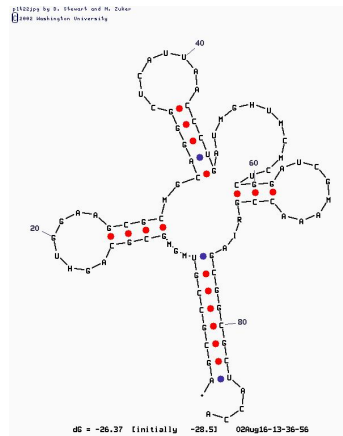
Needleman Wunsch Alignment Algorithm

	A	D	C	N	Y	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	7	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
D	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

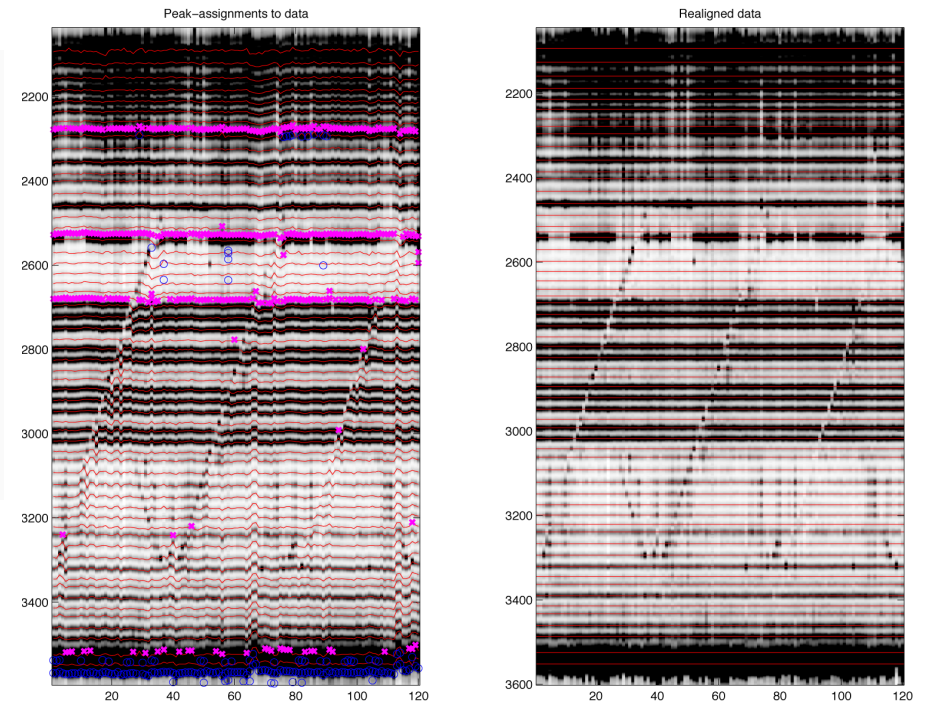


 Doug Brutlag

Nucleic acid 2° structure



Electrophoretic trace alignment



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

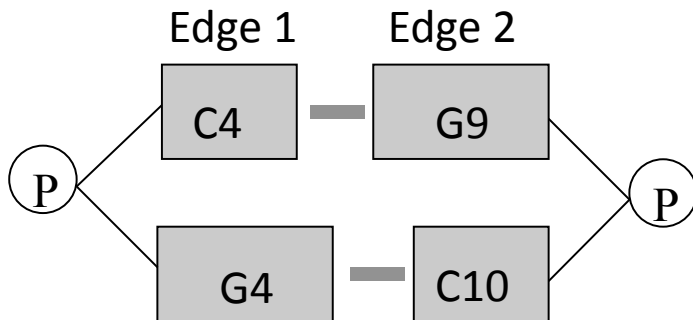
```

TTCTAATACGACTCACTATAGGCCAAAAACAACGGAATTGCGGGAAGGGTCAACAGCCG->1
||||| 71.8
2<-GCCCTTTCCTCCAGTTGTCGGCAAGTCATGGTTCAGAGTCCCTTTGAACTCTACCG
||||| 58.1
GGGAACTTTGAGATGGCCTTGCAAAGGGTATGGTAATAAGCTGACGGACATG->3
||||| 58.3
4<-CATTATTCGACTGCCTGTACCAGGATTGGTGCCTCGGTT
|||||
CAGCCAA

CAGGATTCAGTTG
||||| 60.5
GTCCCTAAGTCAACAGATCTTCTGTTGATATGGATGC->5
||||| 58.5
6<-CTAGAAGACAACACTATACCTACGTCAAGTTTGGTTTGGTTTCTTTGTTGTTGTTGTTG

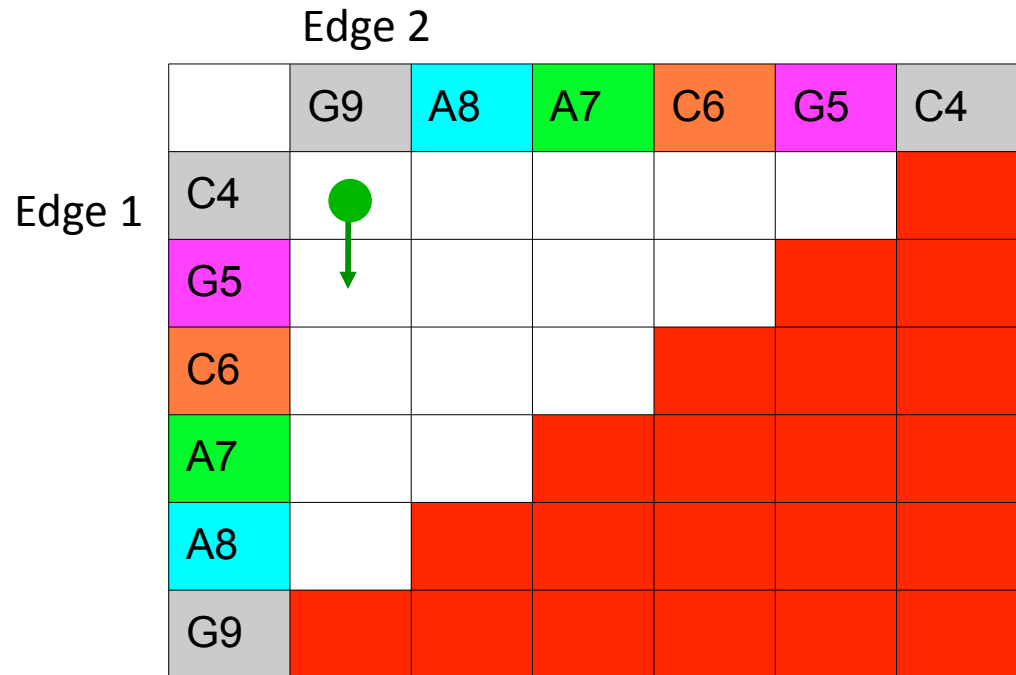
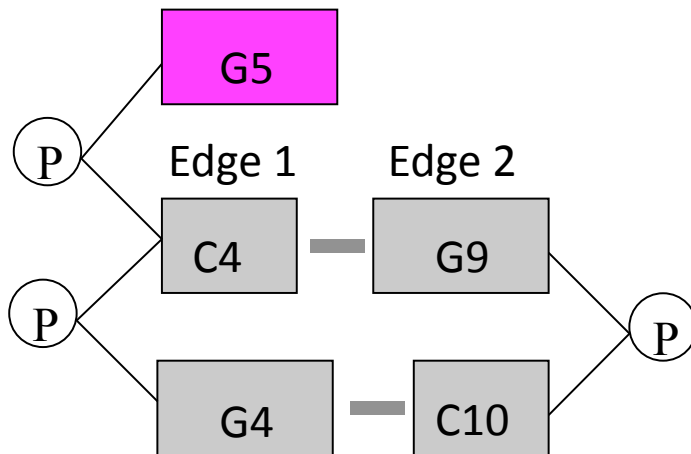
```

Rebuild Pathways: Tetraloop (1zih)



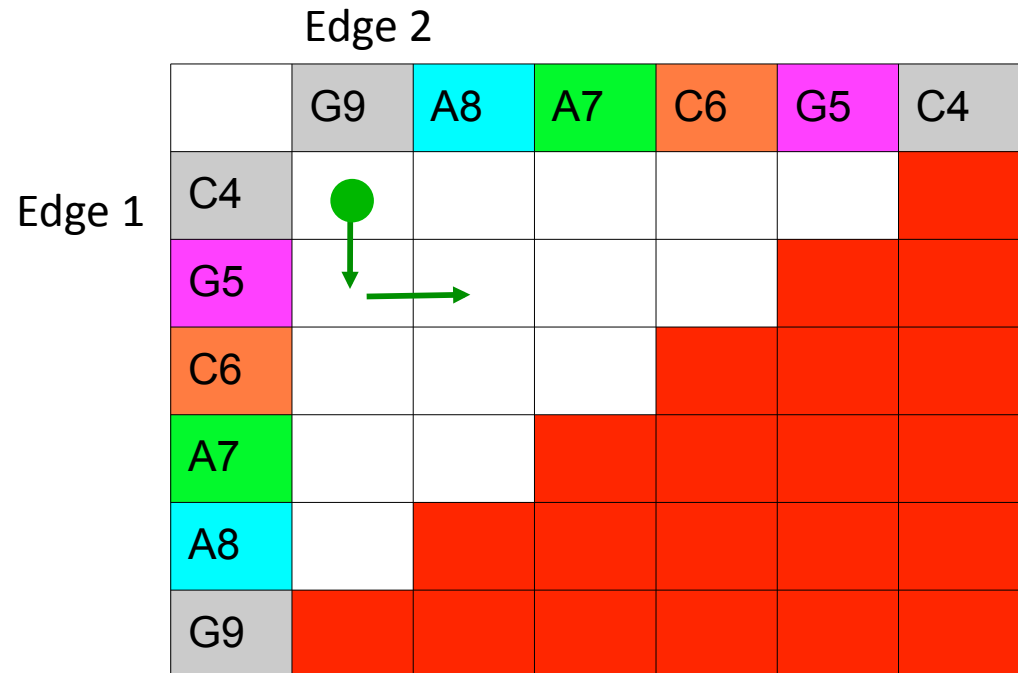
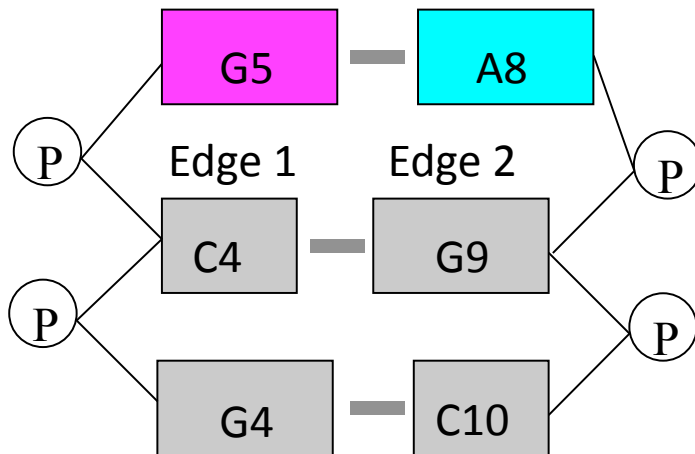
Rebuild Pathway: Tetraloop (1zih)

Build G5 from Edge 1



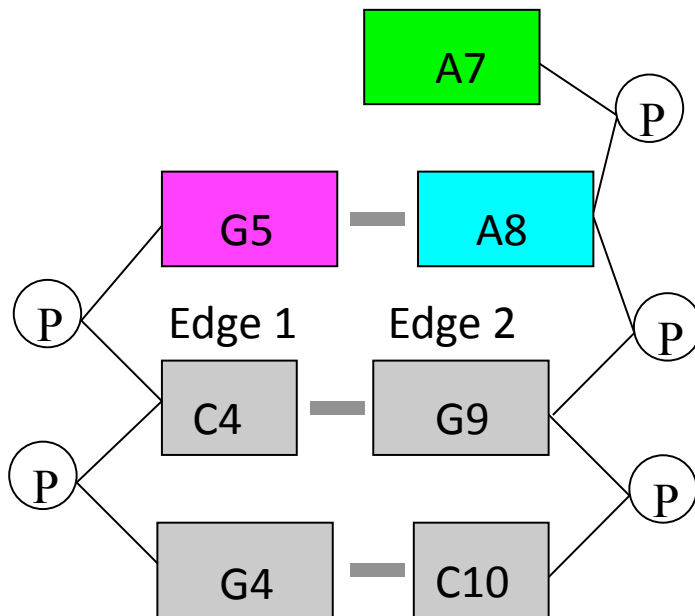
Rebuild Pathway: Tetraloop (1zih)

Build A8 from Edge 2



Rebuild Pathway: Tetraloop (1zih)

Build A7 from Edge 2

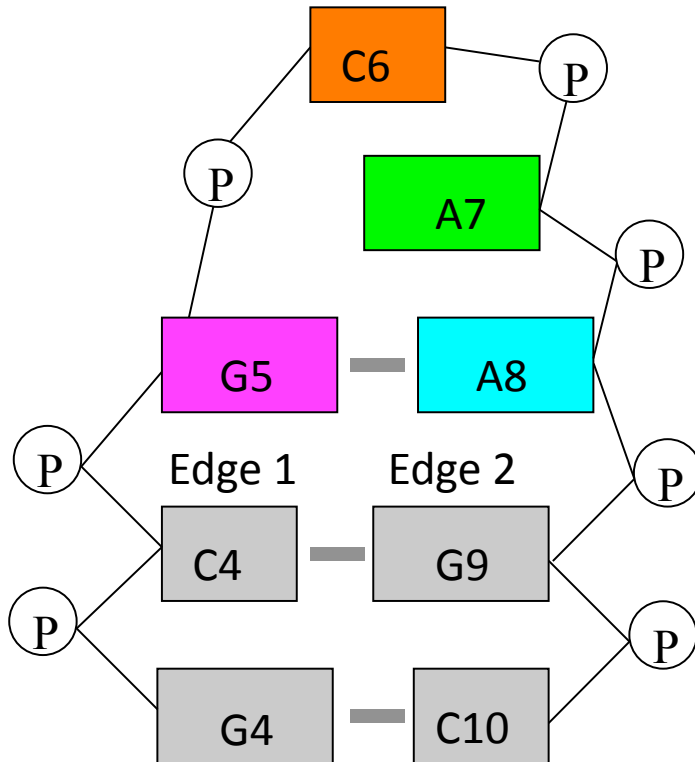


Edge 1

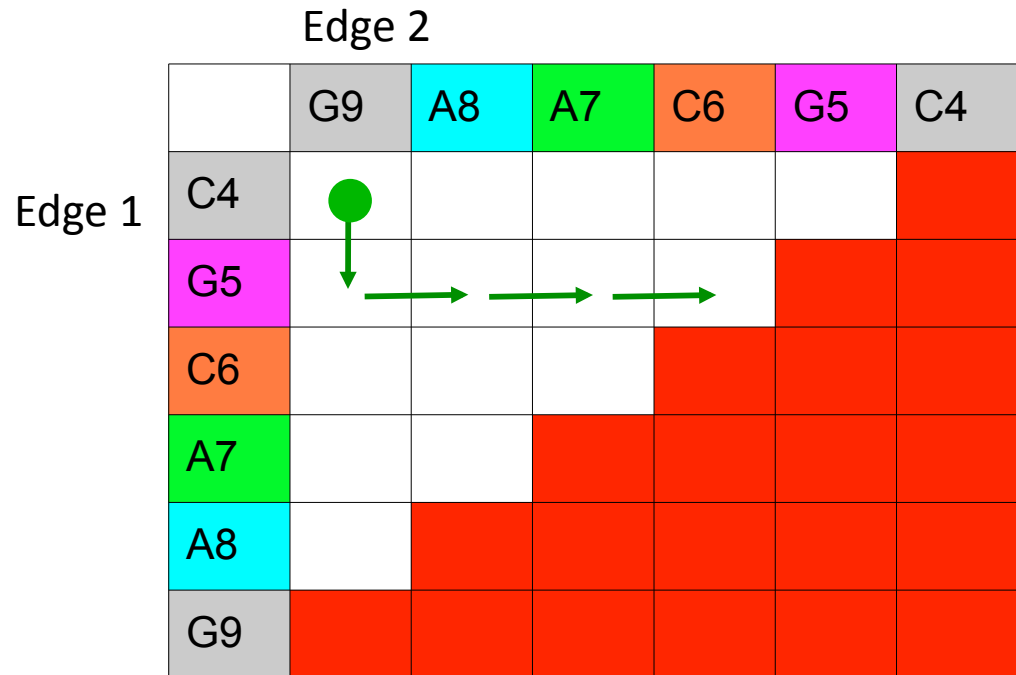
Edge 2

	G9	A8	A7	C6	G5	C4
C4						
G5						
C6						
A7						
A8						
G9						

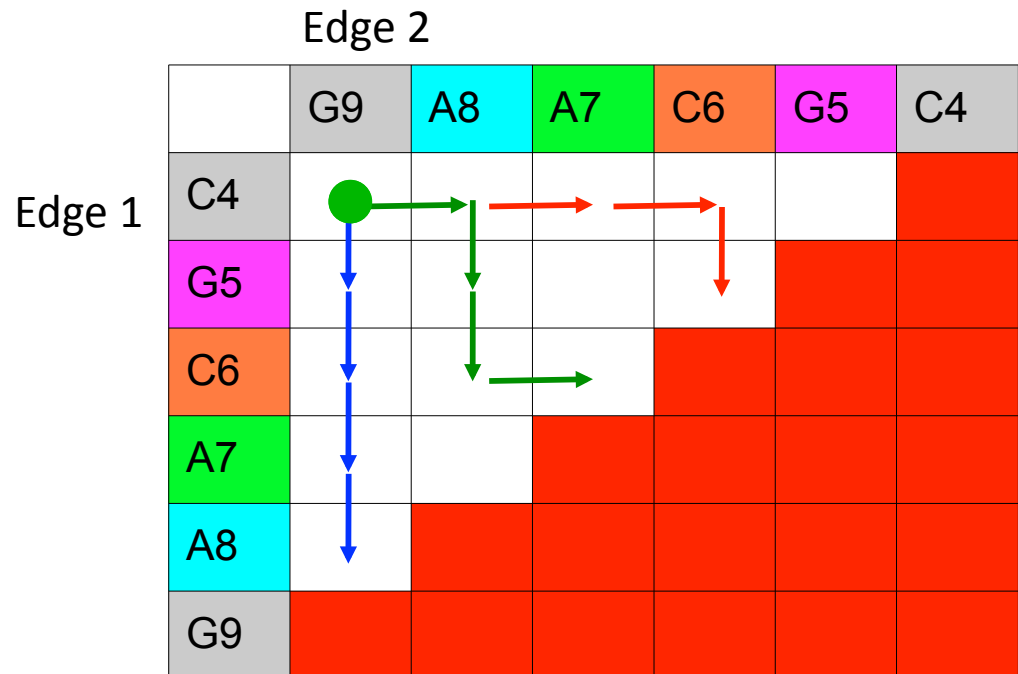
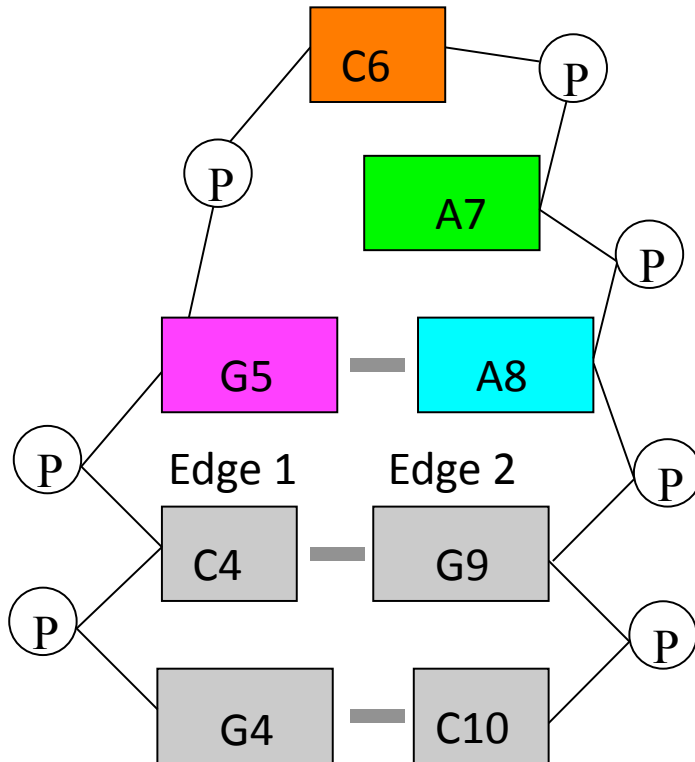
Rebuild Pathway: Tetraloop (1zih)



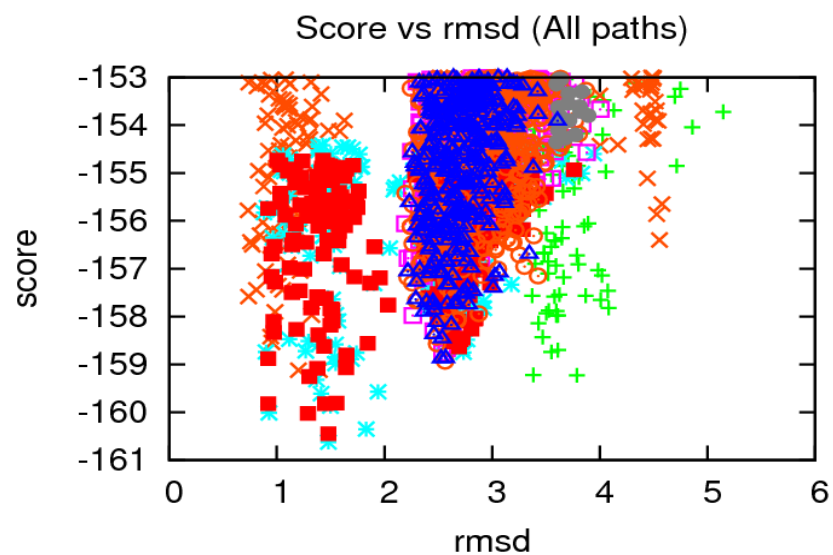
Build C6, close chain



Can we rebuild using other pathways?



All pathways

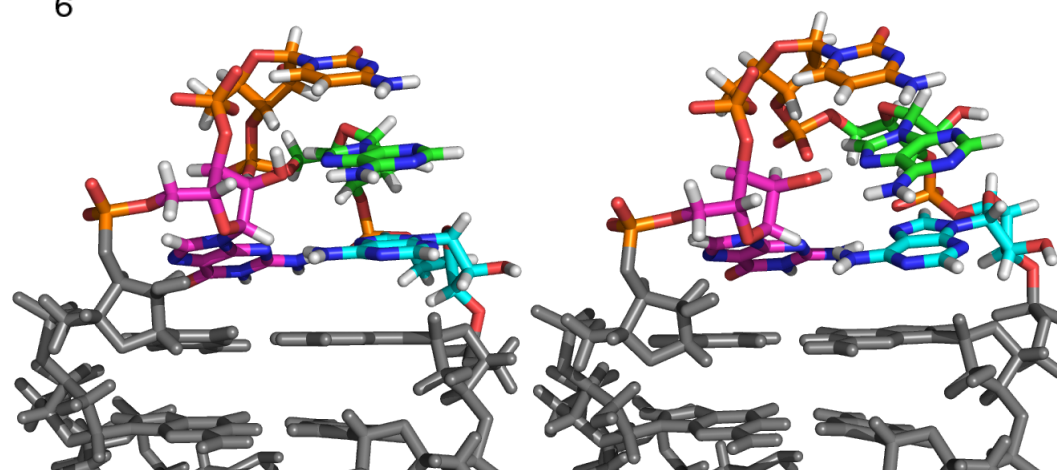


Each point style represents a rebuild path

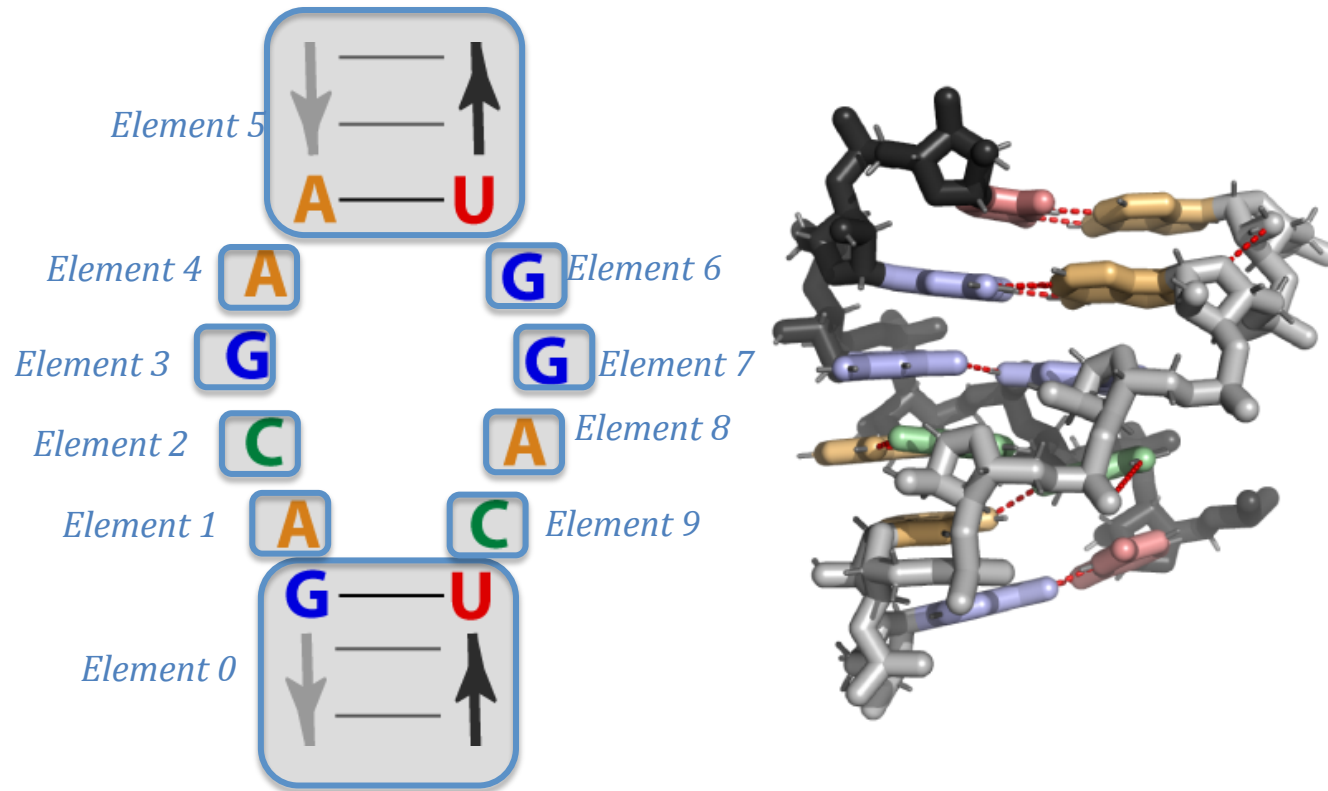
1ZIH NMR

Lowest Energy

Best energy score decoy:
~1.5 Å from NMR structure



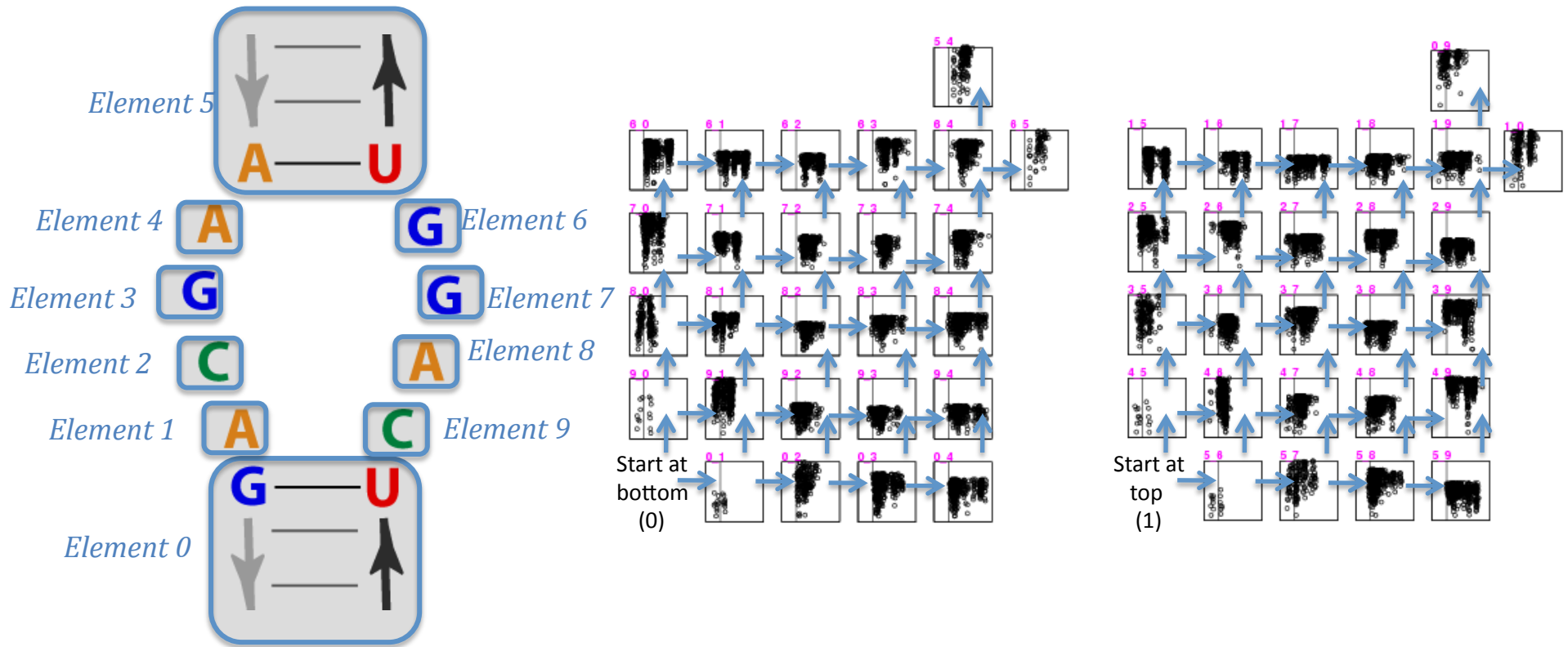
A more complex motif



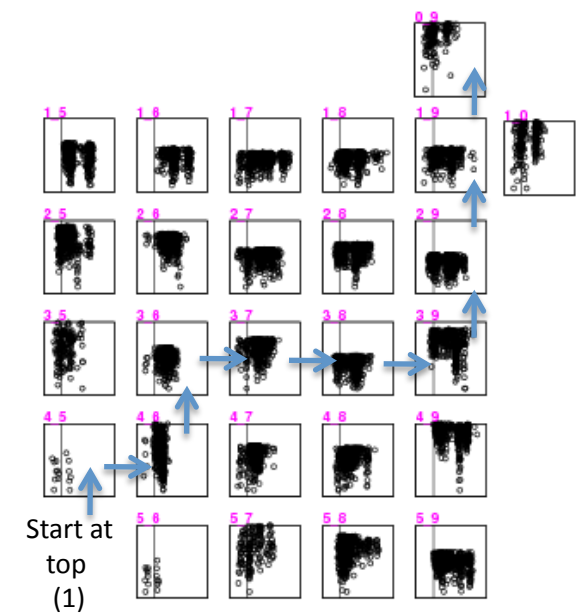
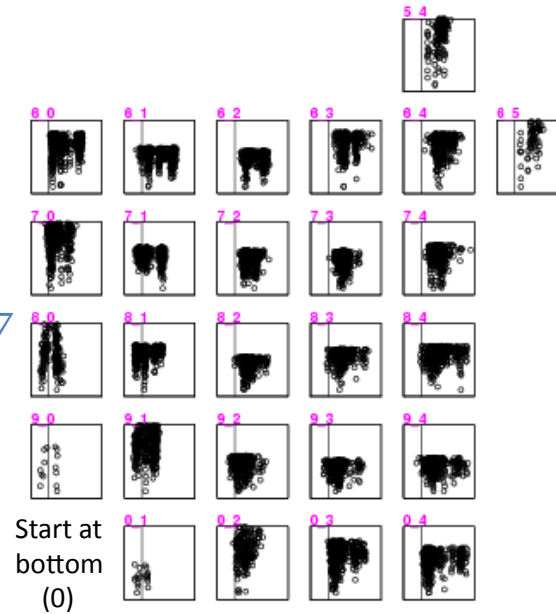
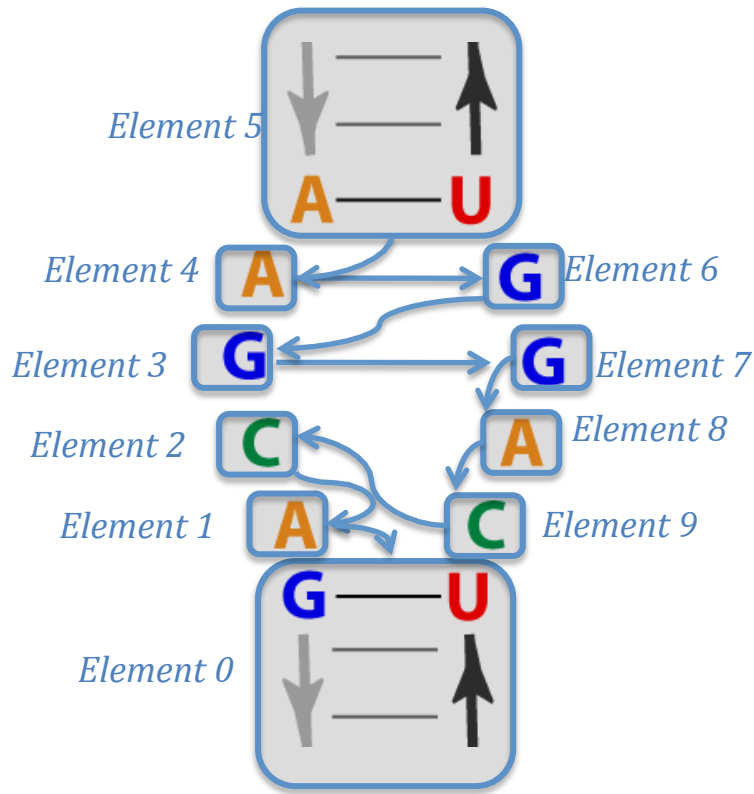
The most conserved domain of the signal recognition particle

(highly stereotyped fold – four crystal structures w/ and w/o protein)

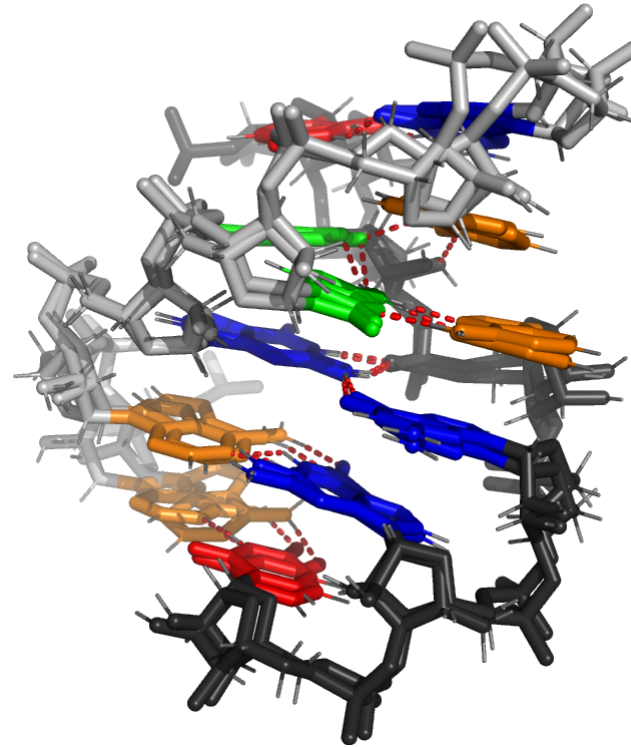
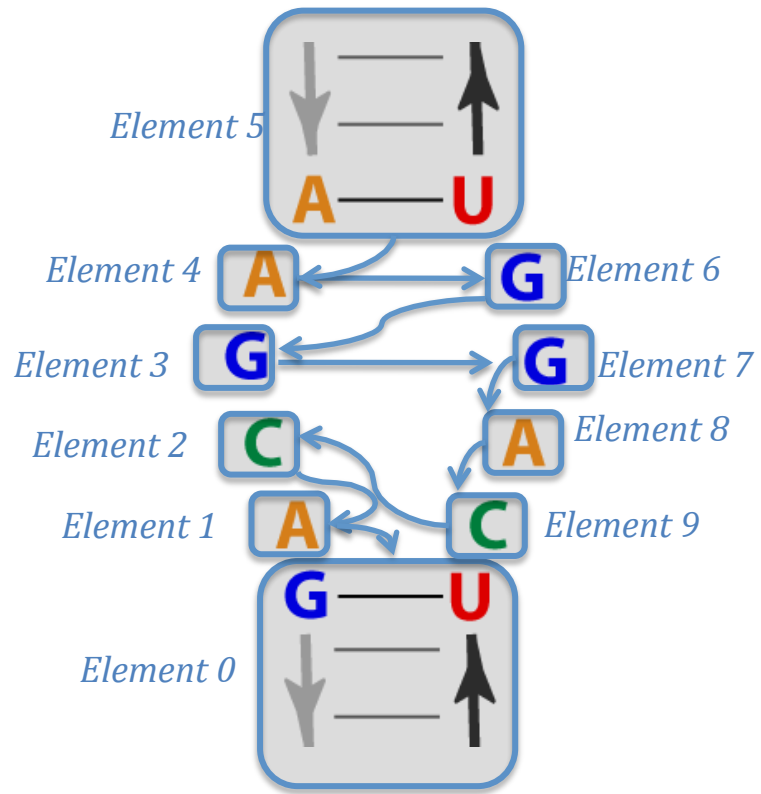
A more complex motif



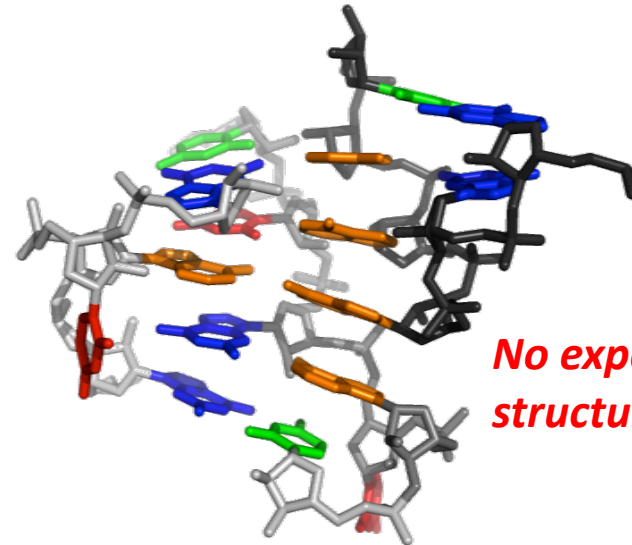
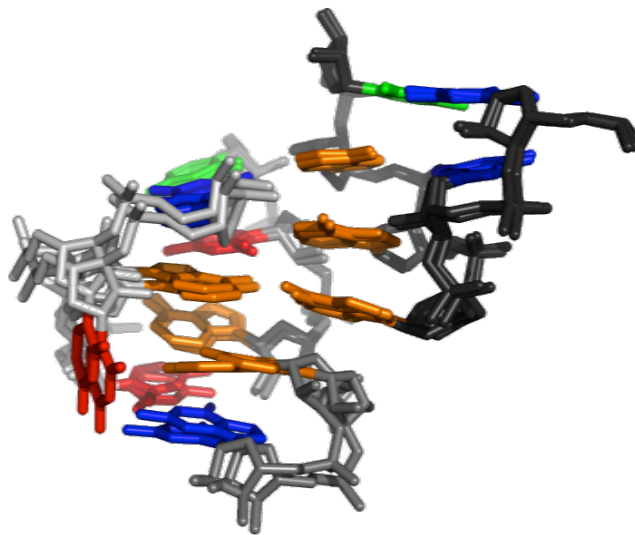
A more complex motif



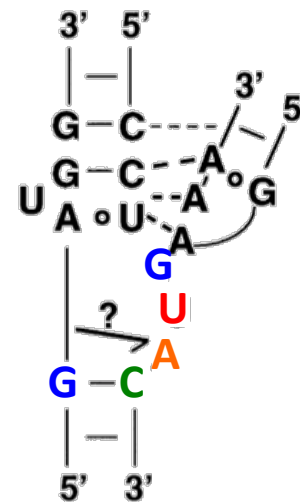
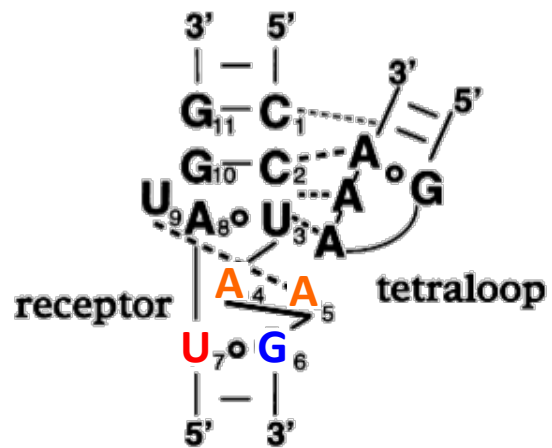
A more complex motif



A baby step, but a *blind* one.



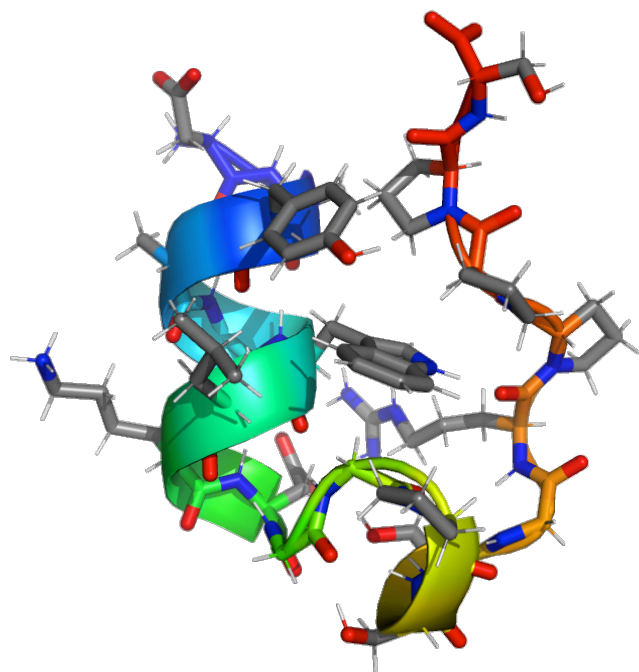
No experimental structure yet.



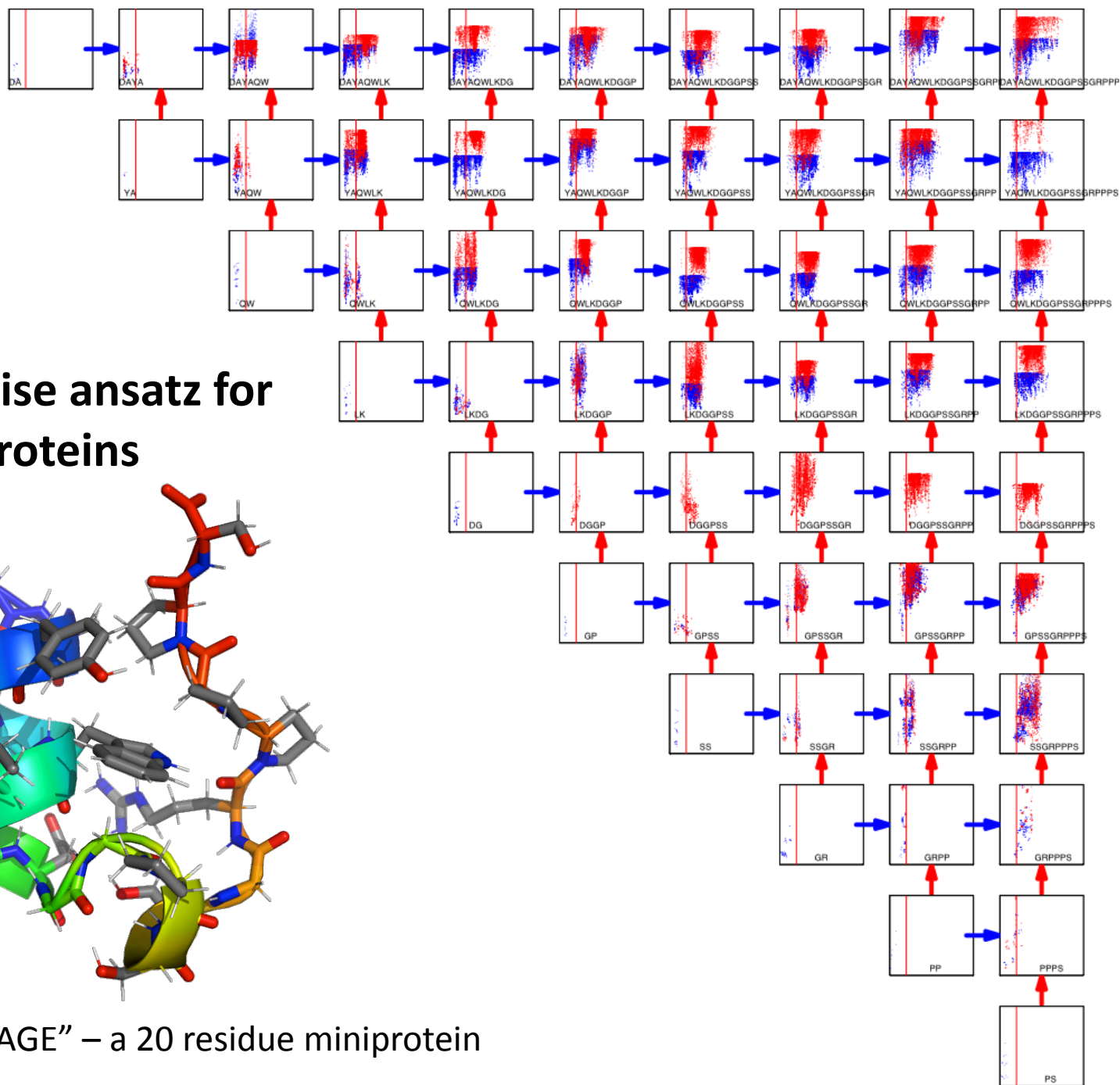
Just rebuilding the *colored* residues

C7.2

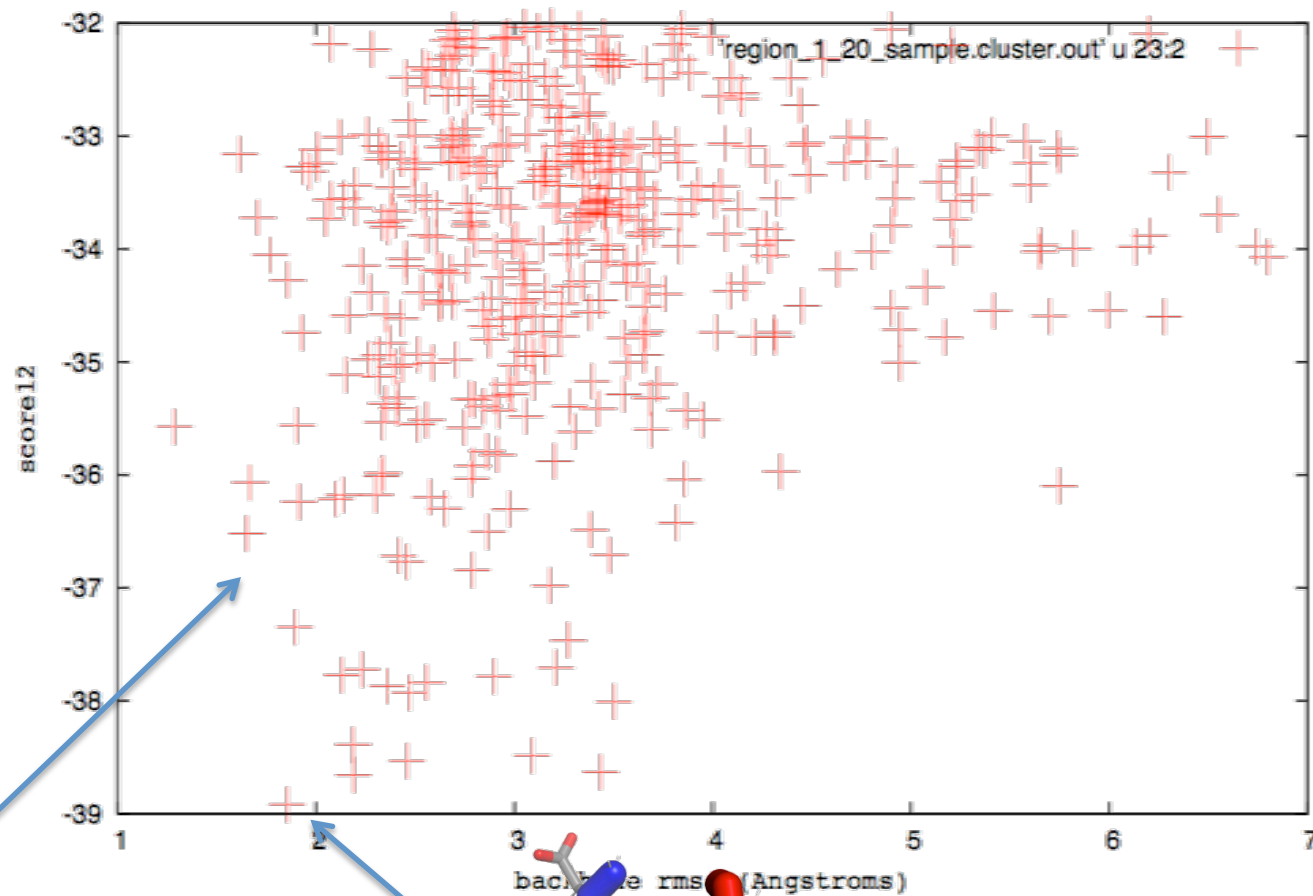
A stepwise ansatz for (mini) proteins



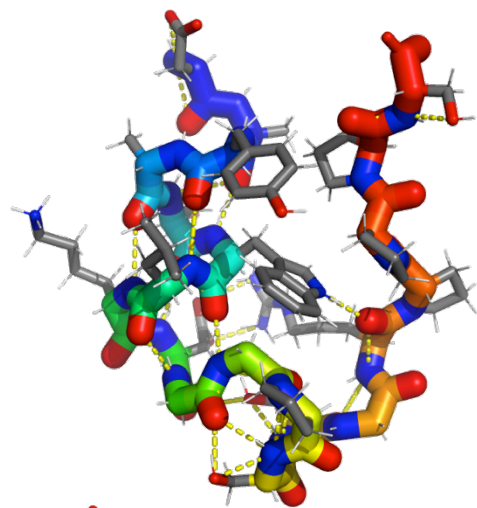
“Trp CAGE” – a 20 residue miniprotein



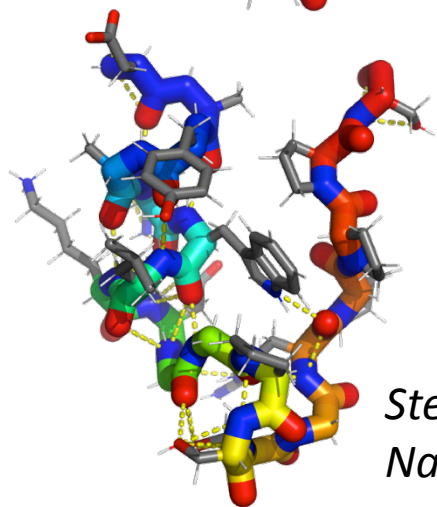
trp cage, 2 by 2, ALL PATHS



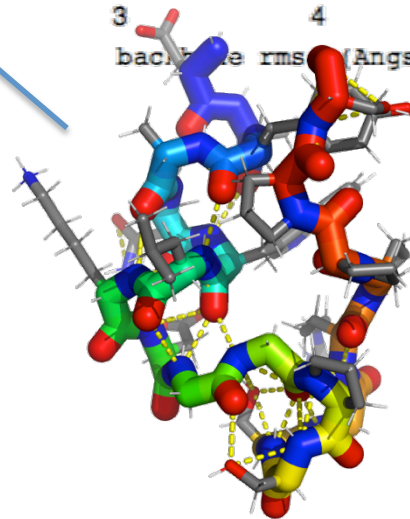
NMR model

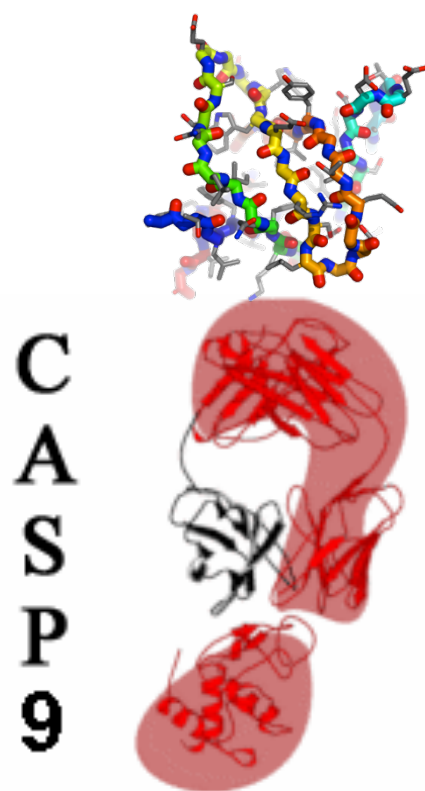


*Stepwise assembly –
Native-like model*



*Stepwise assembly –
Best all-atom score*





May-July 2010.

Stepwise building: a new idea?

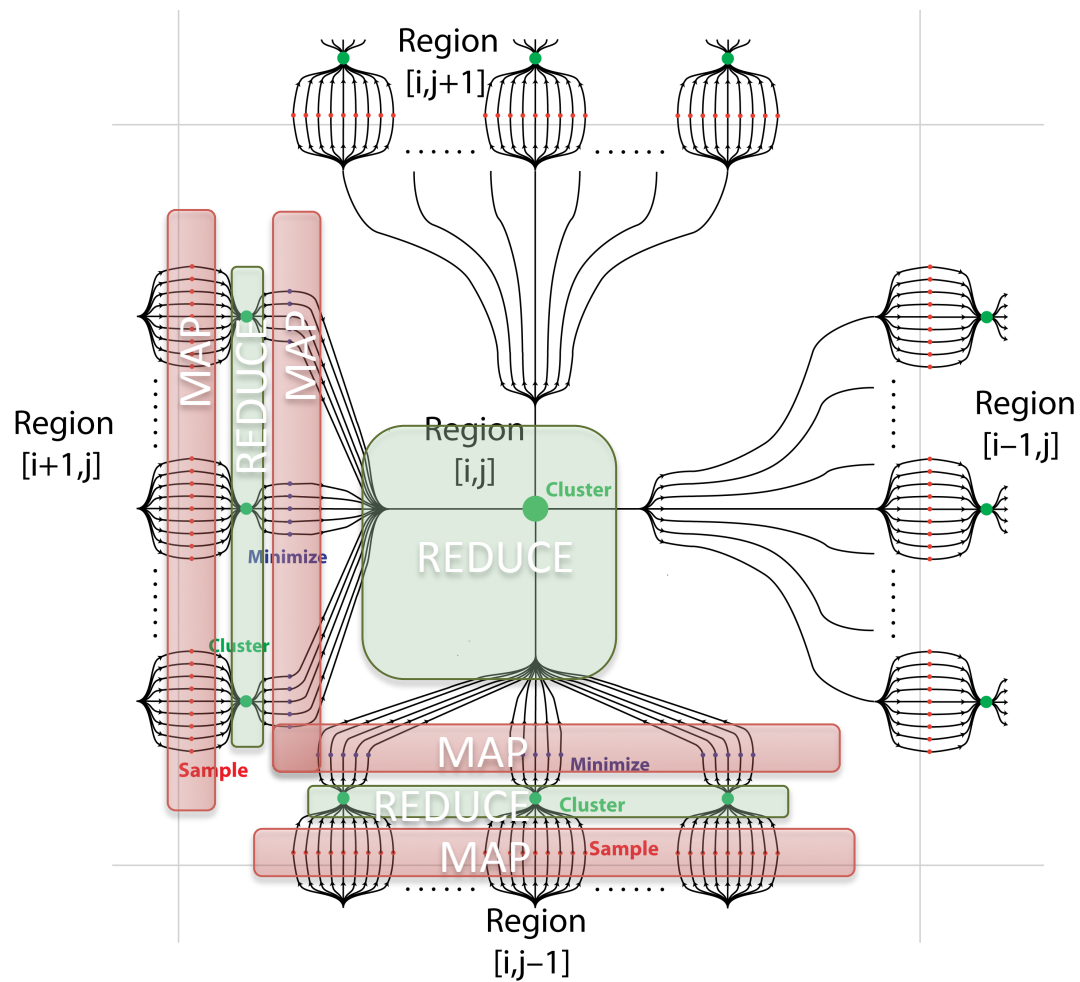
ARE THERE PATHWAYS FOR PROTEIN FOLDING ?

by CYRUS LEVINthal

[Massachusetts Institute of Technology, Department of Biology Cambridge, Massachusetts.]

Thus, the computer-aided model building is not designed to find the configuration of minimum energy rather, it is designed as an aid to the investigator as various sequentially folding steps are tried.

Finally, the computer system has been used in attempts to deduce plausible folding pathways for myoglobin and lysozyme.

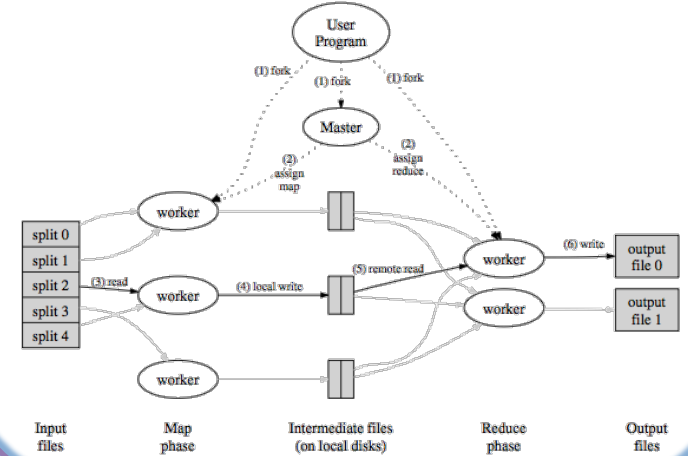


MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.



Prooogole

|MFACNNFAGAPQRSTLNYIIALSDDWVQPPWP

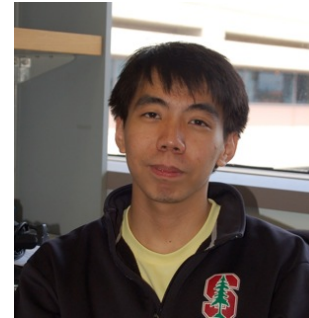
[Advanced Search](#)
[Language Tools](#)

Google Search

I'm Feeling Lucky

BIOC 218 (Feb, 2014)

Acknowledgments



Parin
Sripakdeevong



Kyle Beauchamp

- David Baker & lab
- Vijay Pande and group (MD expertise)
- Adrien Treuille, Jee Lee + colleagues
- Andrew Leaver-Fay, Sergey Lyskov, **Rosetta community**



Jane Coffin
Childs
Foundation

