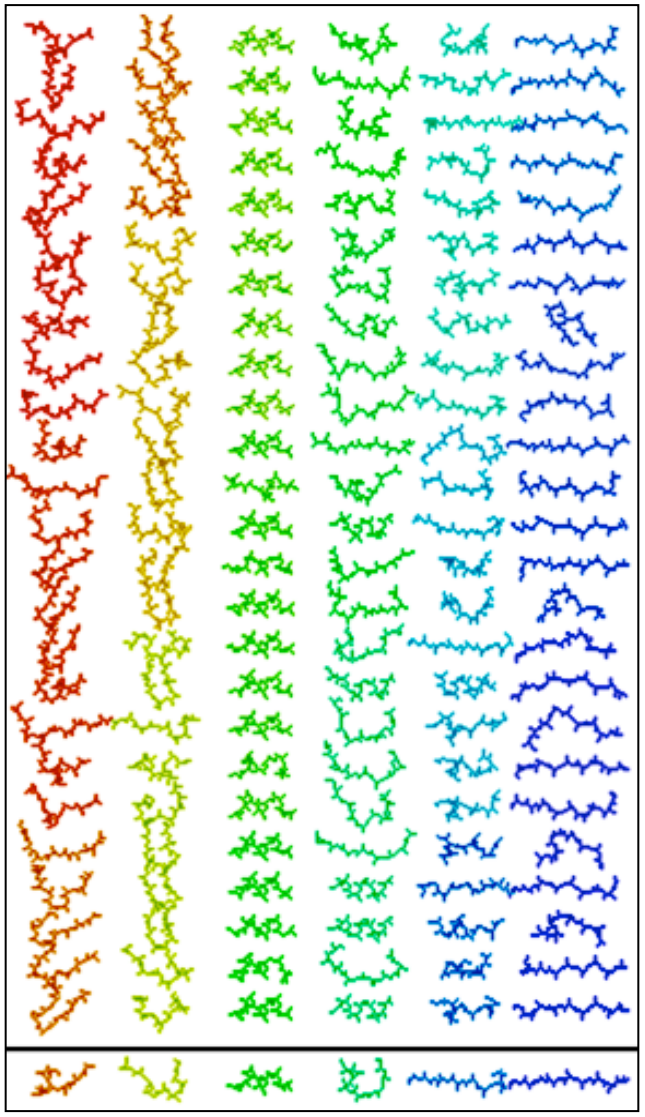
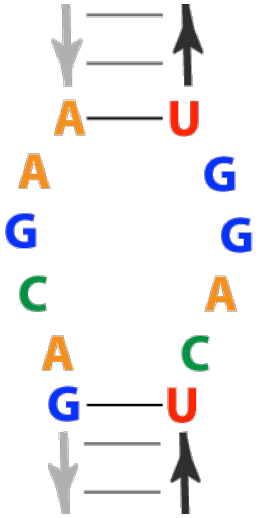


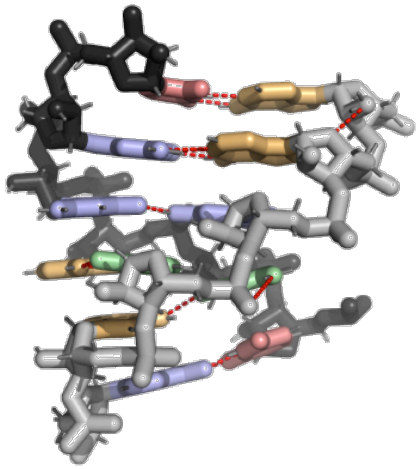
Structures from scratch

Rhiju Das, Departments of Biochemistry & Physics

BIOC 218

Feb 2010

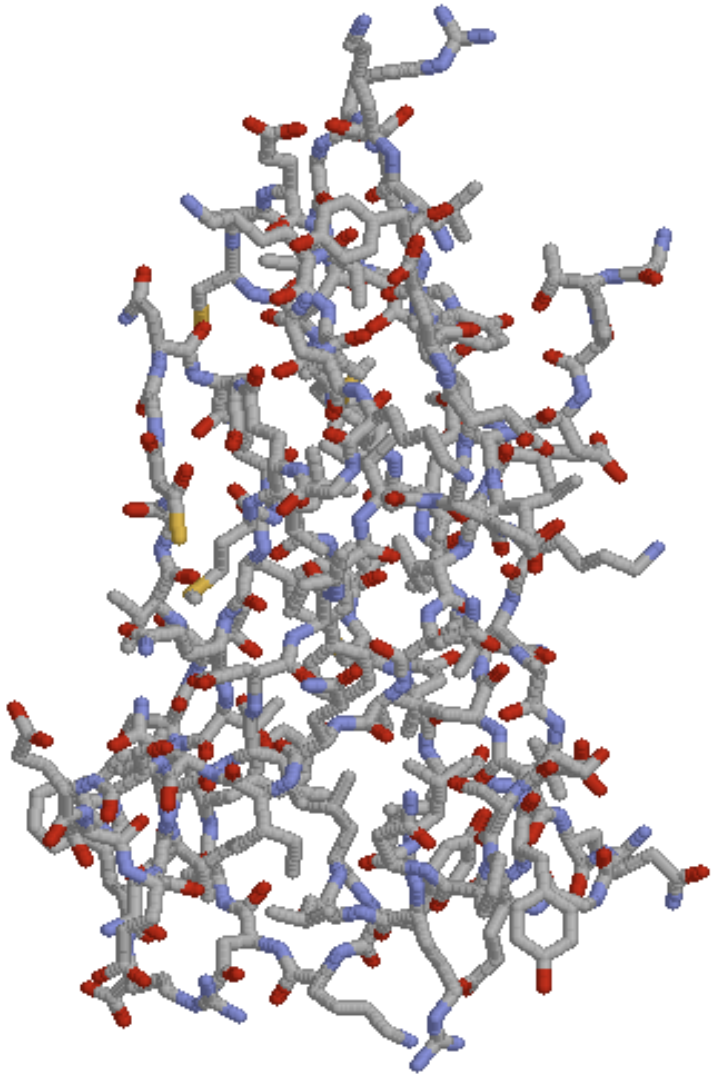




Predicting protein structure

GTPDIIVNAQINS
EDENVLDFIIEDEY
YLKKRGVGAHIIK
VASSPQLRLLYKN
AYSTVSCGNYGVL
CNLVQNGEYDLN
AIMFNCAEIKLNK
GQMLFQTKIWR

?



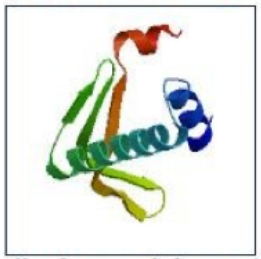
Automatic Protein Fold Prediction Results

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             hhhhhh hhhhhh   h hhhhhhhhhh hhhhhhhhhh    sss  ss
1mulA              hhhhhh hhhhhh   h hhhhhhhhhh hhhhhhhhhh    sss  ss


TARGET    49    TFKVNHRAER TGRNPQTGKE IKIAAANVPA FVSGKALKDA VK
1mulA     49    tfkvnhrae- ---------- ---aaanvpa fvsgkalkda vk-

TARGET          sssssss    sssss  sss ss   sssss sssshhhhhh h
1mulA           sssssss                     sssss sssshhhhhh
```

This will happen to you a lot.

Automatic Protein Fold Prediction Results

http://swissmodel.expasy.org/

SWISS-MODEL Workspace

Modelling Tools Repository Documentation

[myWorkspace] [Settings] [logout]

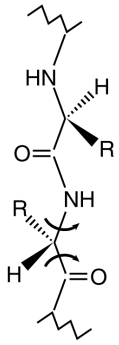
Computation of this workunit has stopped.

Please see the following log report for details:

```
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.
```

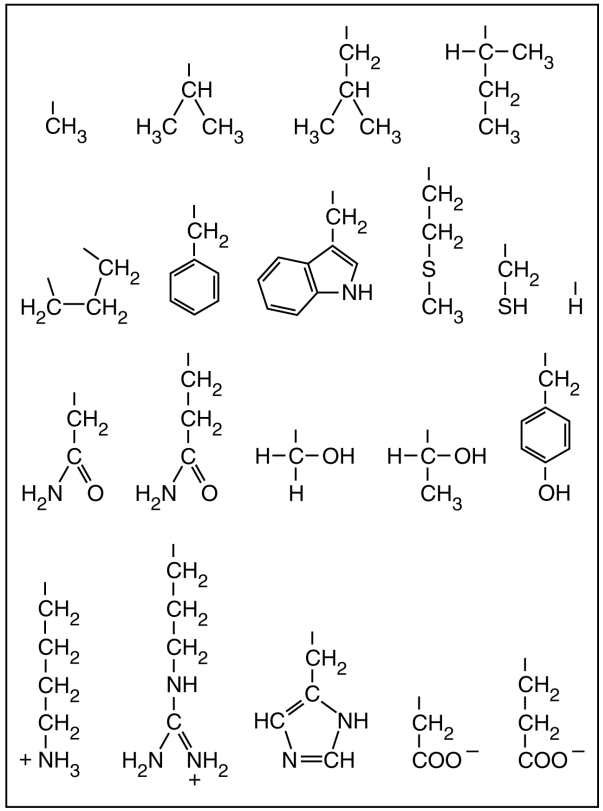
Proteins

Backbone



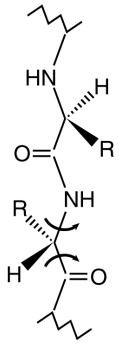
Side Chains

R ≡



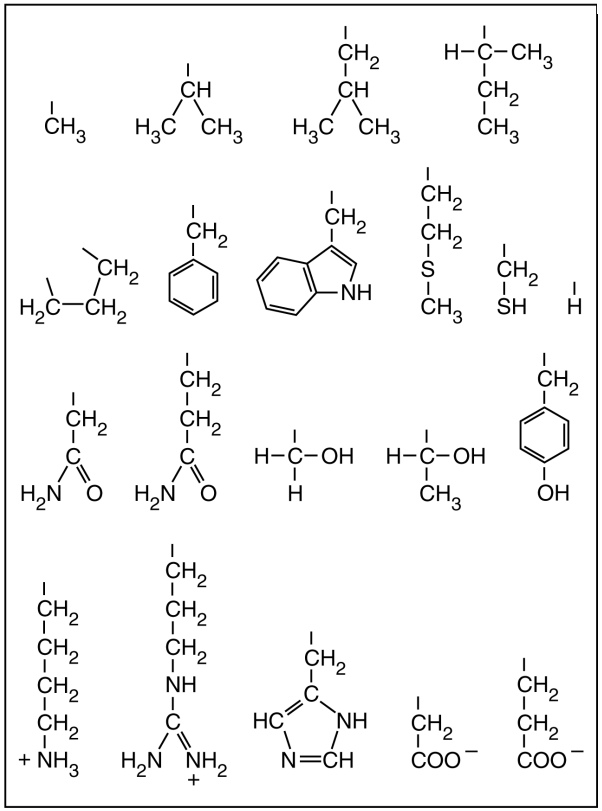
Proteins

Backbone



Side Chains

R ≡

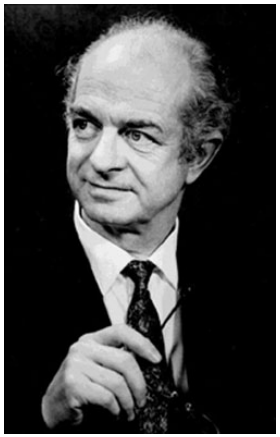


THE STRUCTURE OF PROTEINS: TWO HYDROGEN-BONDED HELICAL CONFIGURATIONS OF THE POLYPEPTIDE CHAIN

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



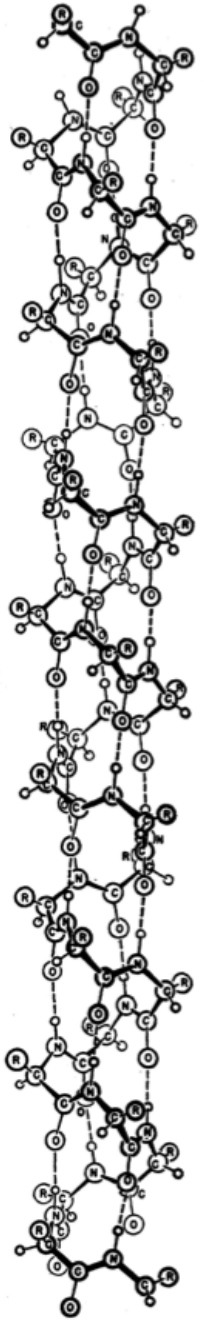
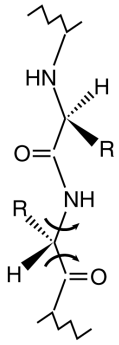


FIGURE 2
The helix with 3.7 residues per turn.

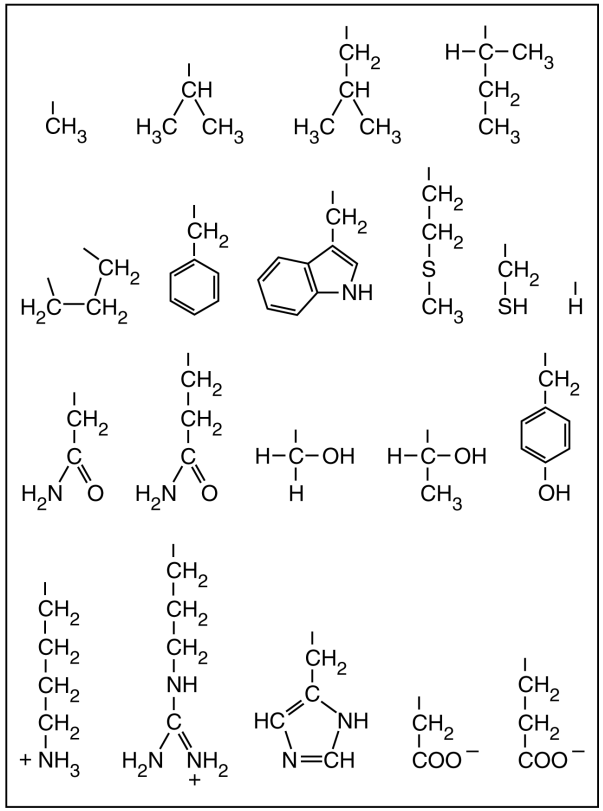
Proteins

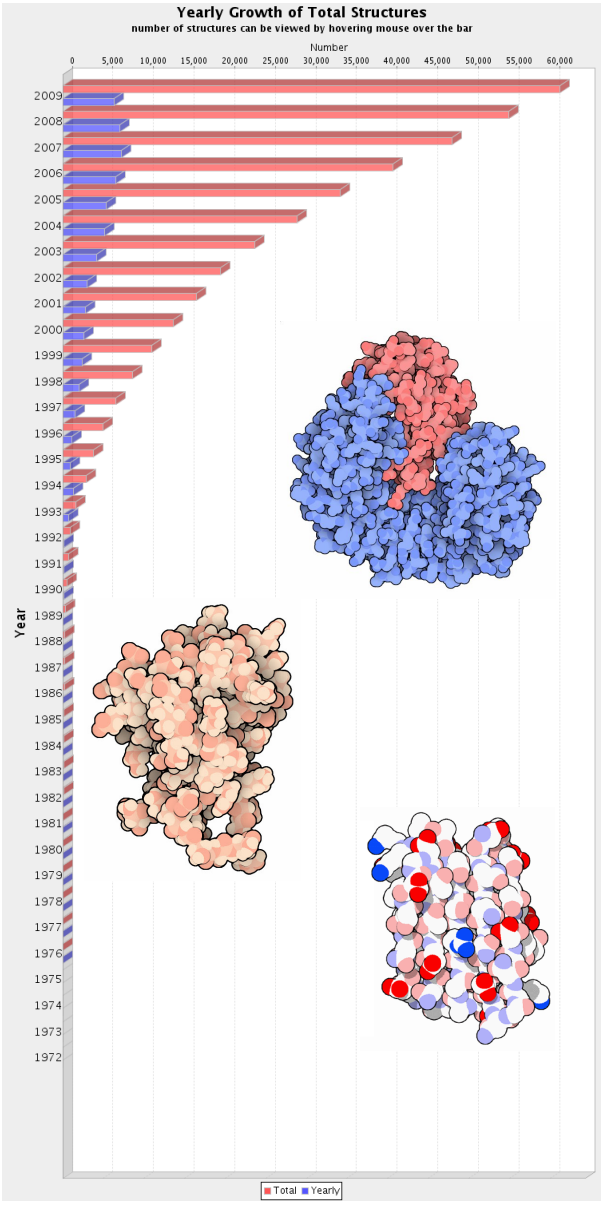
Backbone



Side Chains

R ≡





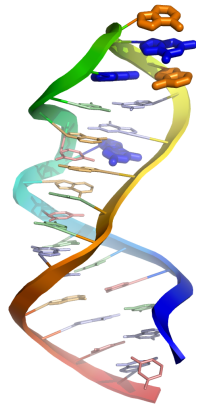
Two fundamental problems

1. Predicting protein structure

GTPDIIVNAQINSEDENVLDF
IIEDEYYLKKRGVGAHIIKVAS
SPQLRLLYKNAYSTVSCGNYG
VLCNLVQNGEYDLNAIMFNC
AEIKLNKGQMLFQTKIWR

2. Predicting RNA structure

ugcuccuaguacgag
aggaccggagug



Proc. Natl. Acad. Sci. USA
Vol. 75, No. 2, pp. 554–558, February 1978
Chemistry

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

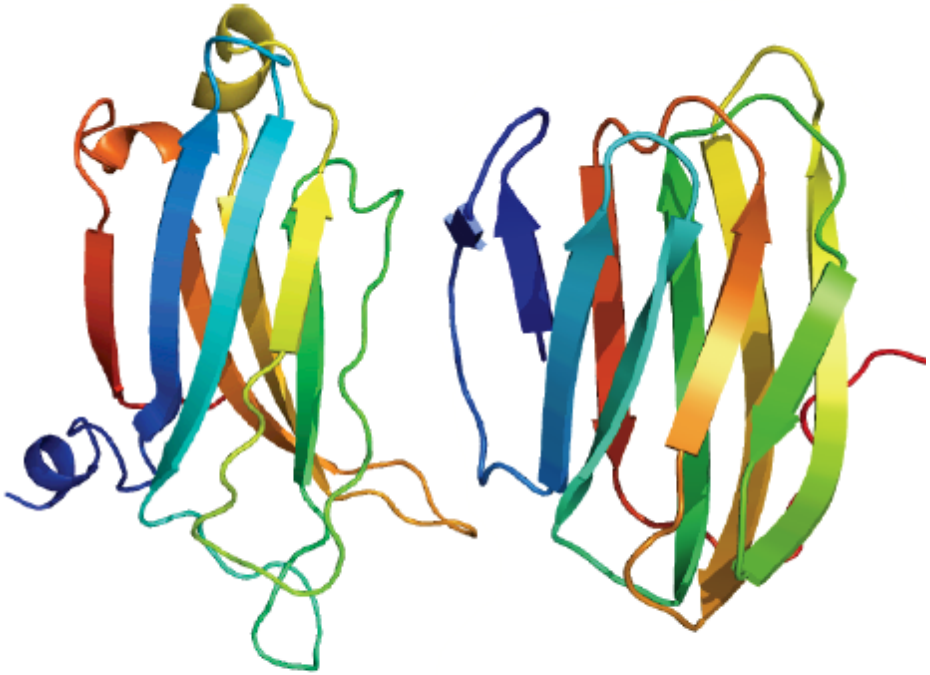
The impression generated by these various simulations is that major progress has been made towards predicting the tertiary structure of a protein from its amino-acid sequence; i.e., the folding problem may be far more tractable than has generally been considered (9).

Driving innovation in protein structure prediction: "CASP"

Critical Assessment of Structure Prediction

Five *blind* predictions per target

CASP1 (1994)



CASP1 TARGET (1rsy)

"successful" fold recognition 2tbv

RMSD: 16.0 Å

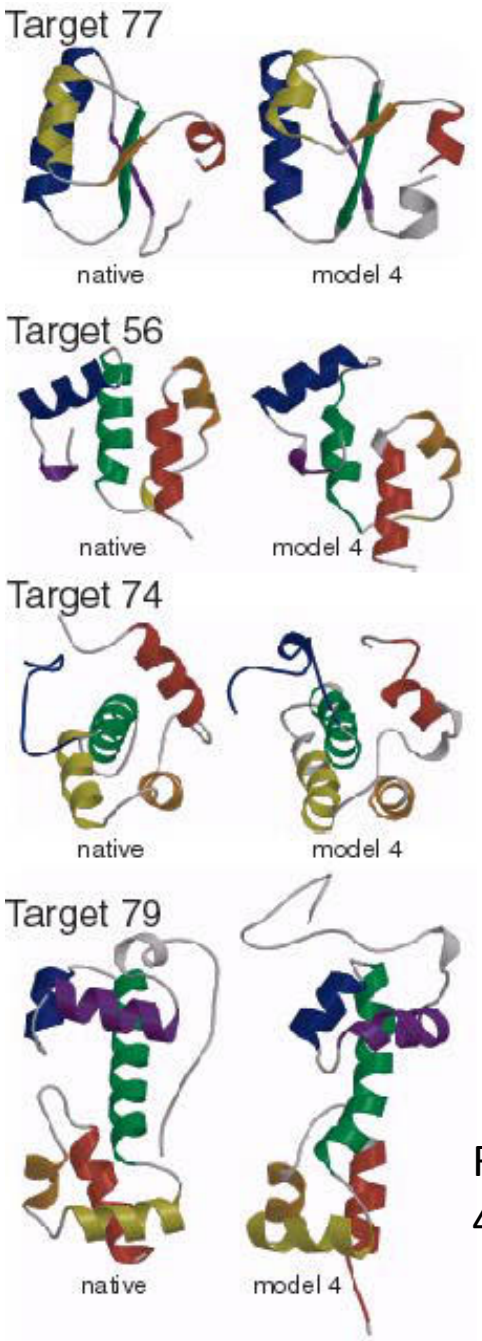
From Neil Clarke, CASP7 assessor's talk

Driving innovation in protein structure prediction: "CASP"

Critical Assessment of Structure Prediction

Five *blind* predictions per target

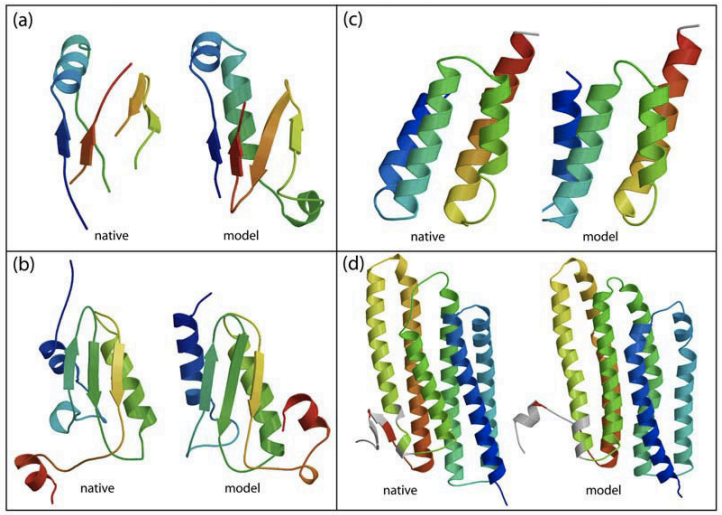
DAVID BAKER & colleagues

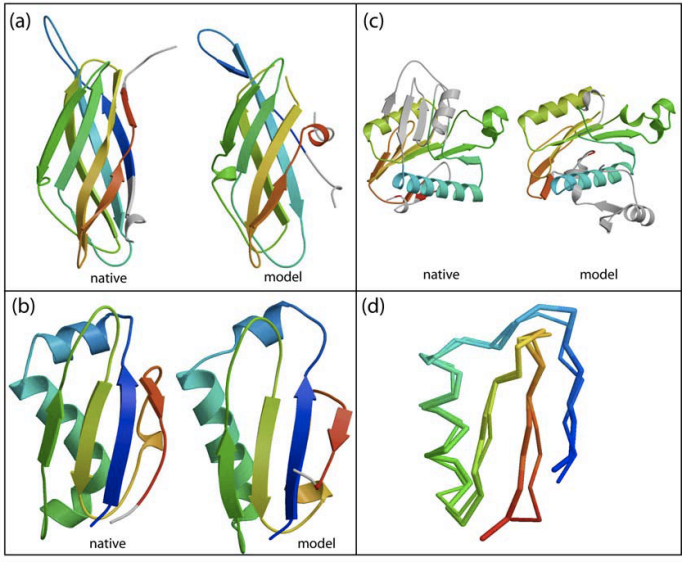


CASP3 (1998)

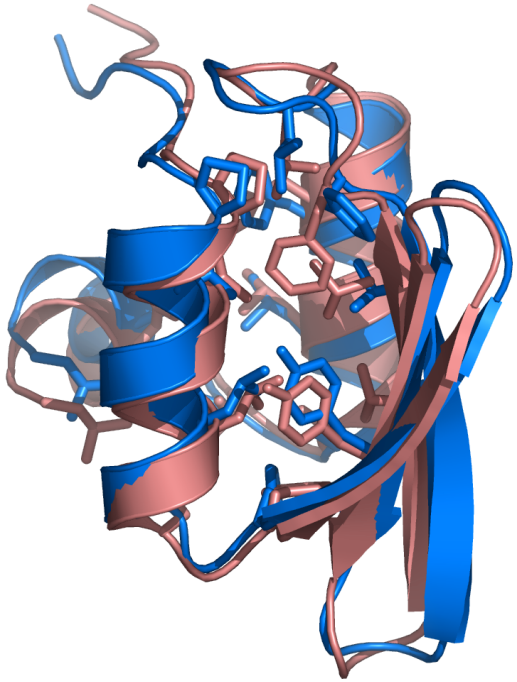
RMSD:
4 to 6 Å

CASP6 (2004)



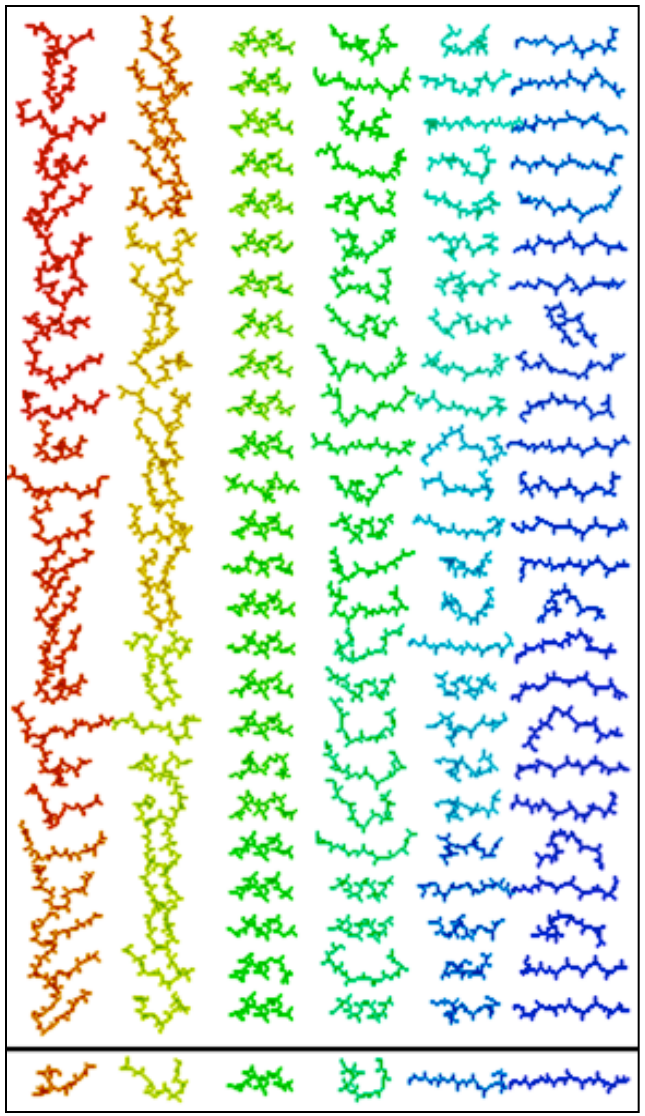


T0281 (1.6 Å over 70 residues)

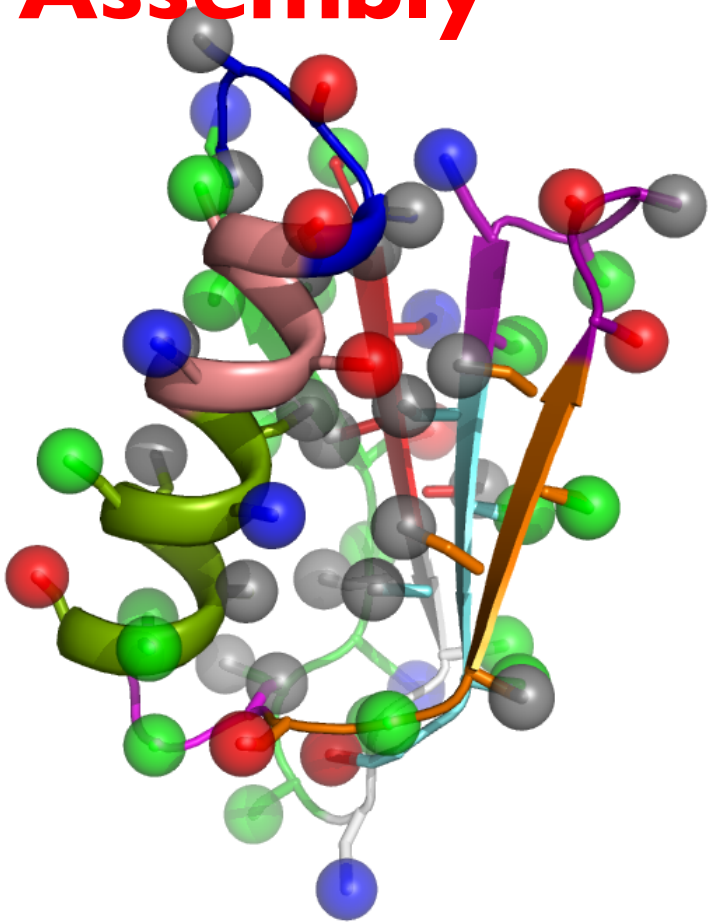


DAVID BAKER & colleagues

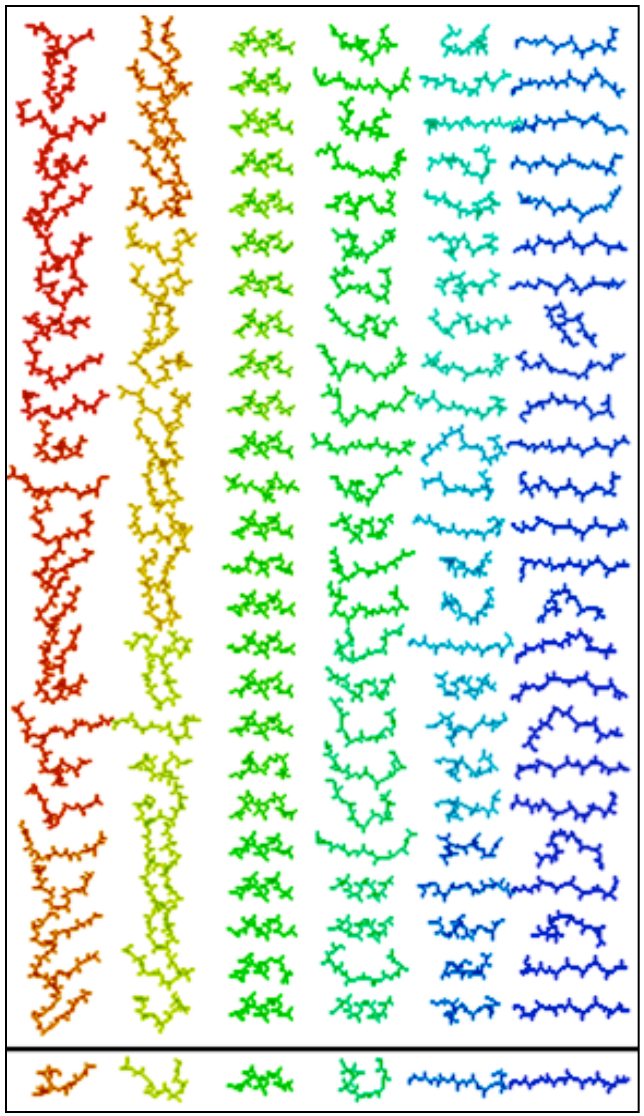
De novo Modeling with Rosetta



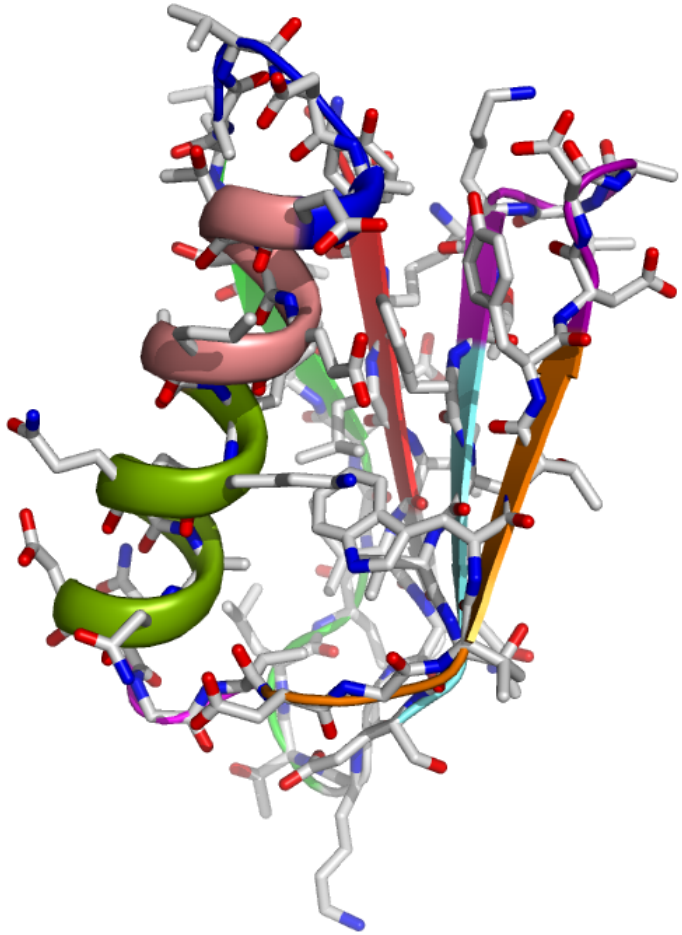
Stage I. Fragment Assembly



De novo Modeling with Rosetta

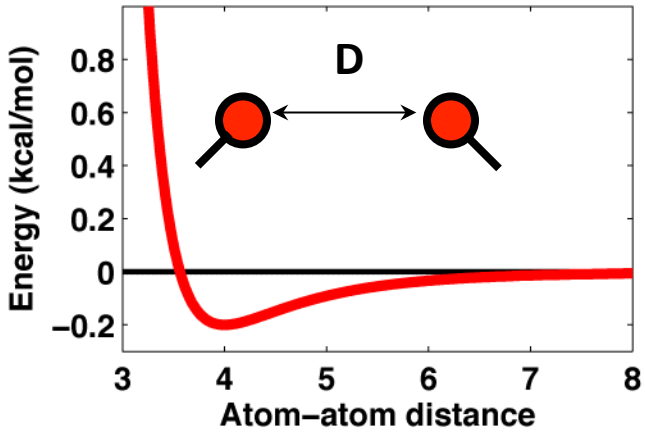


Stage II. All-atom refinement

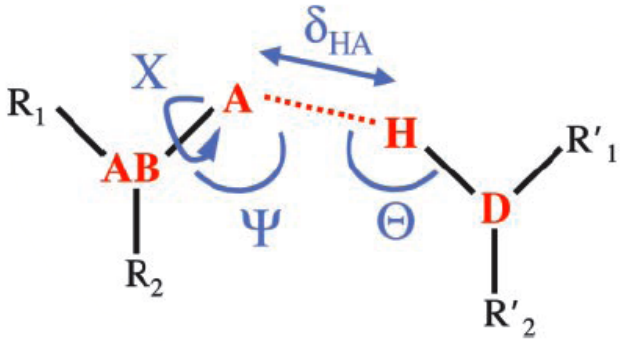


Ingredients of a high resolution potential

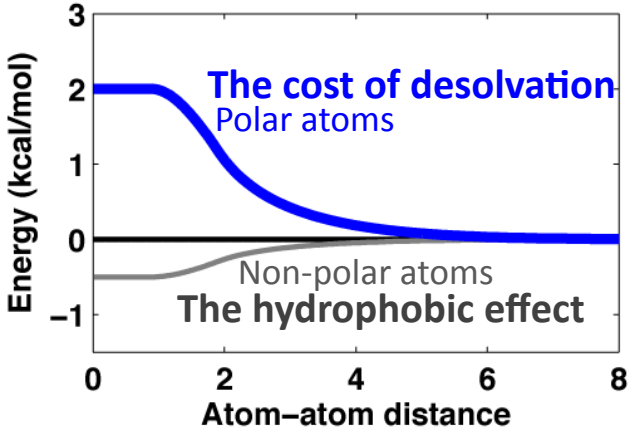
1. Van der waals packing



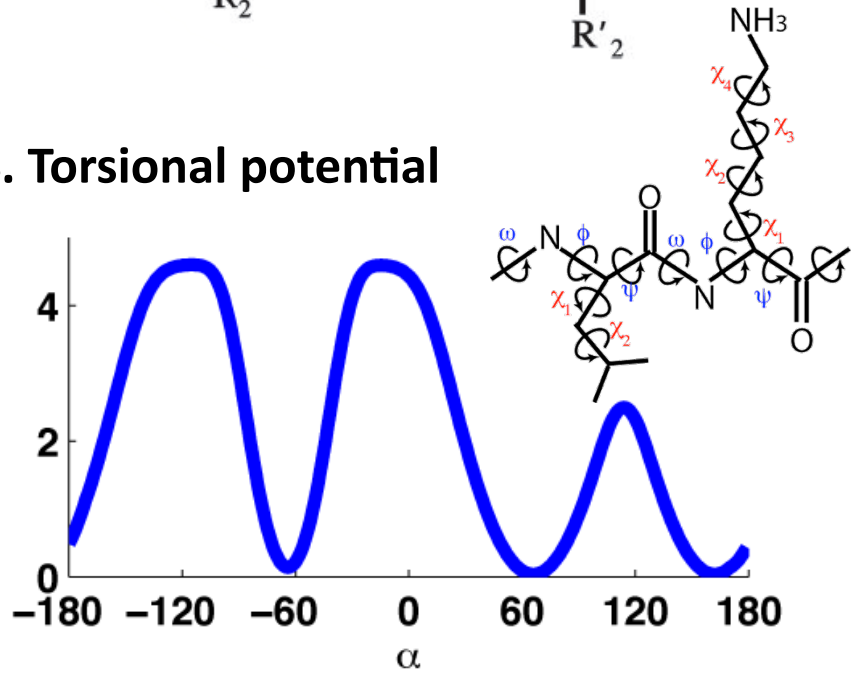
2. Hydrogen bonds



3. Manifestations of water

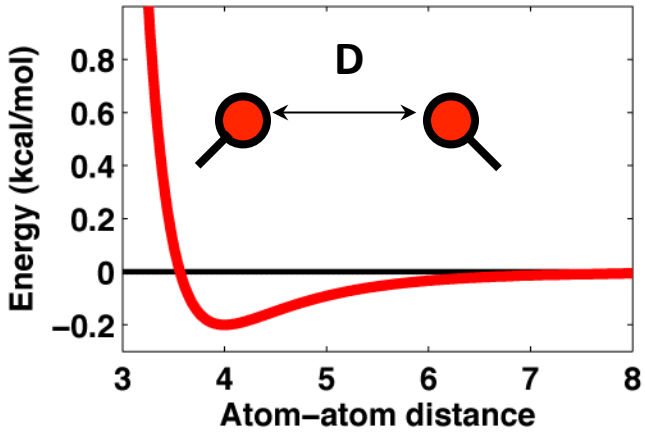


4. Torsional potential

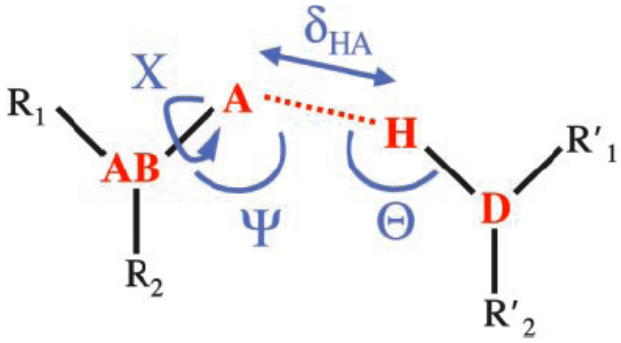


Ingredients of a high resolution potential

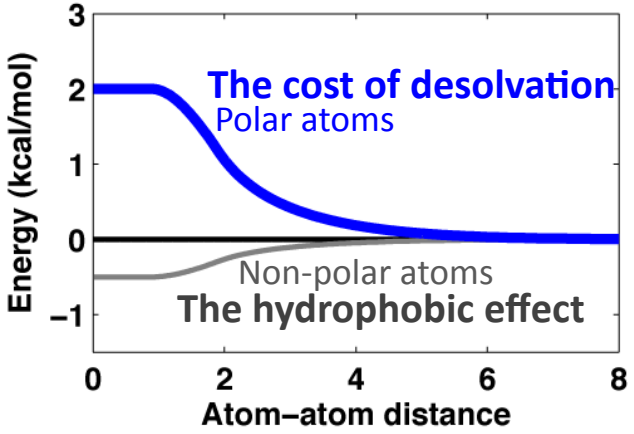
1. Van der waals packing



2. Hydrogen bonds



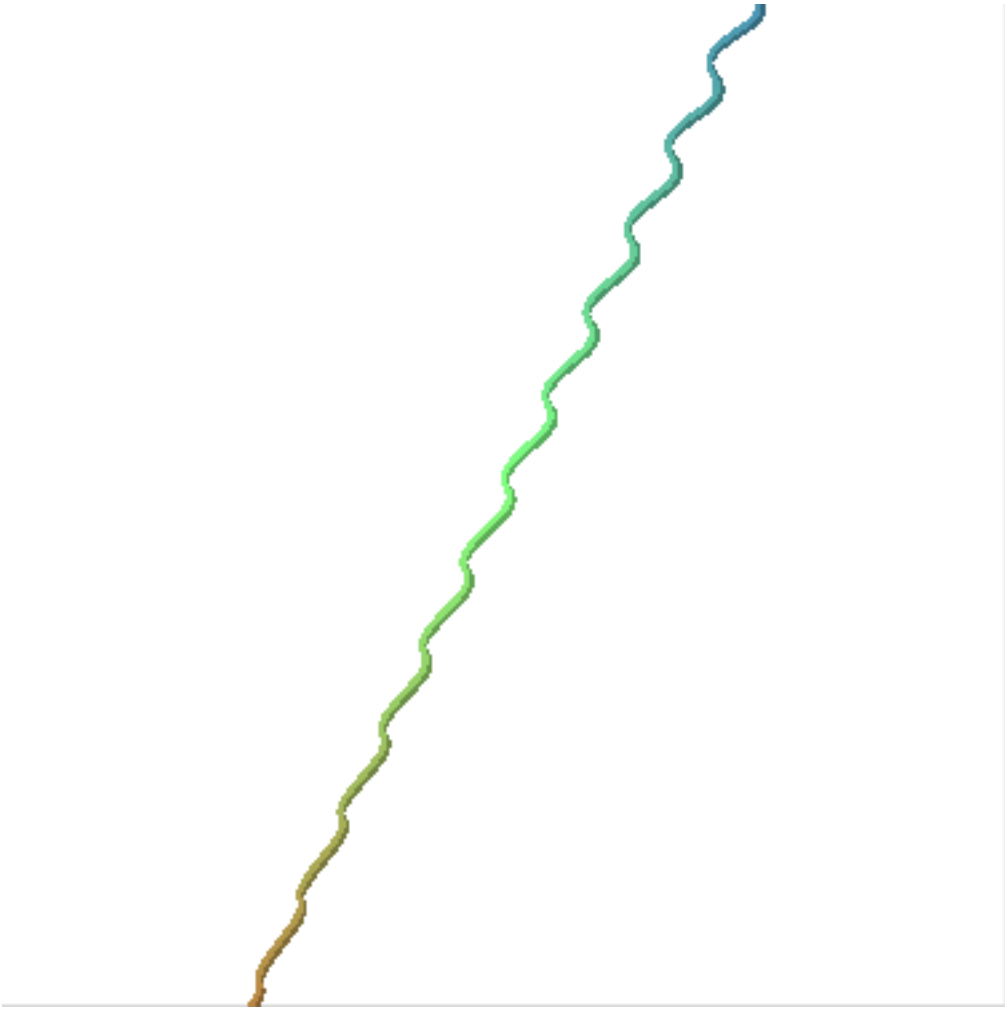
3. Manifestations of water



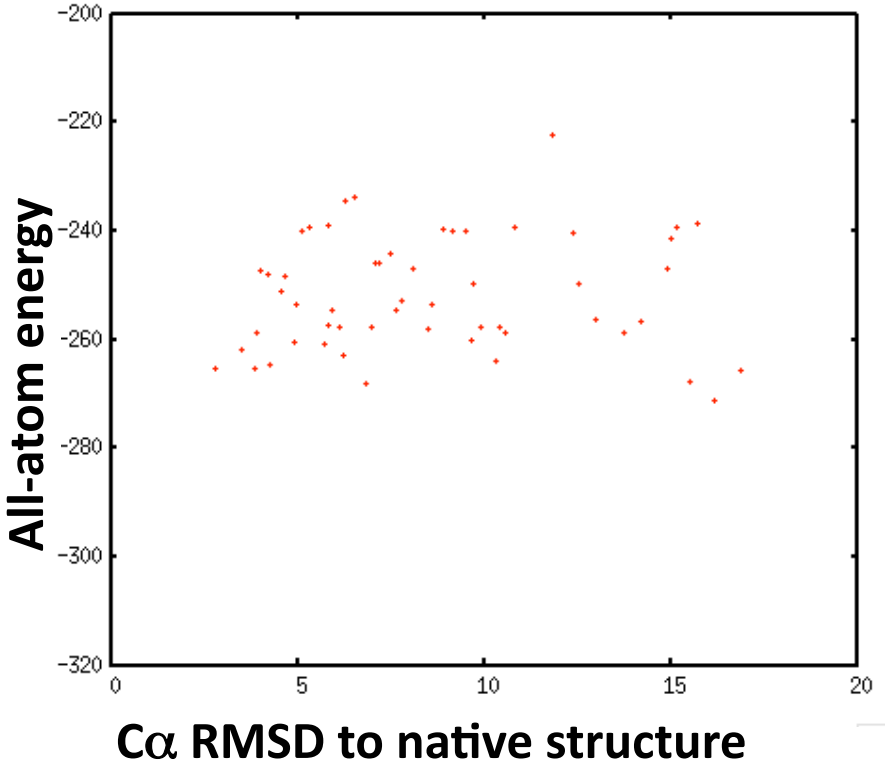
the clustering of non-polar groups, the enhanced hydrogen bond attraction in a hydrophobic environment and the access of water molecules to those polar groups which are not involved in hydrogen bonds.

– Michael Levitt, 1969

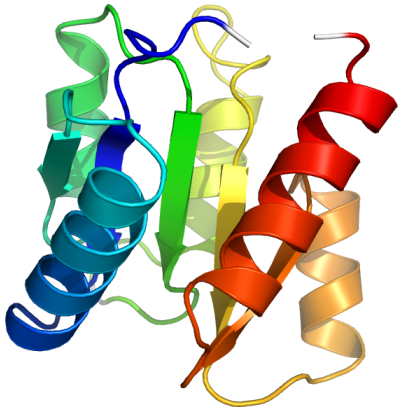
Rosetta in action



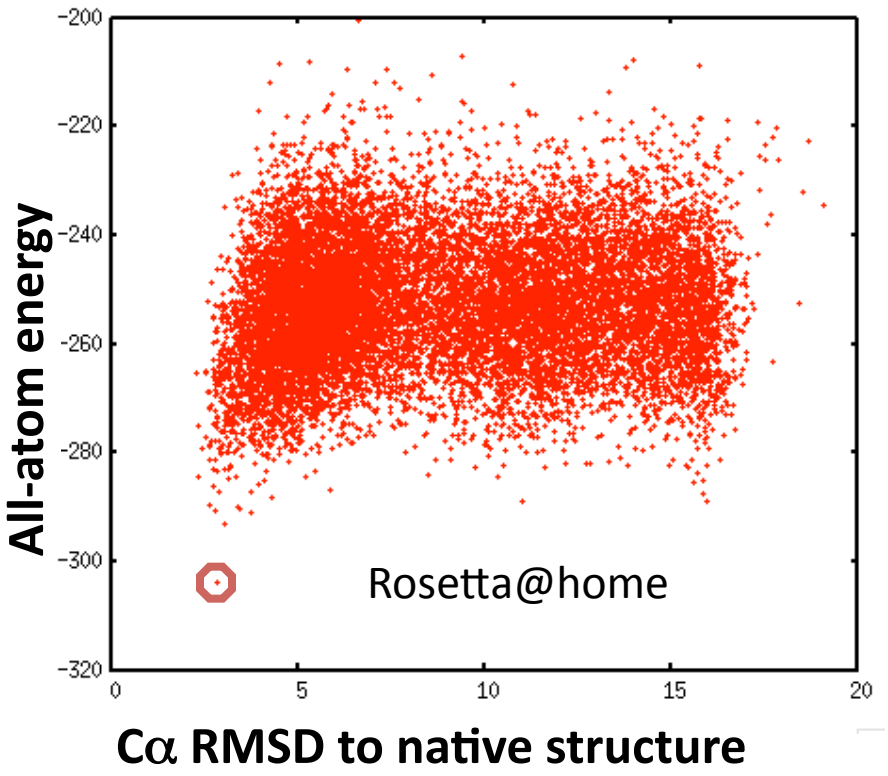
A ~1000-fold increase in computational power



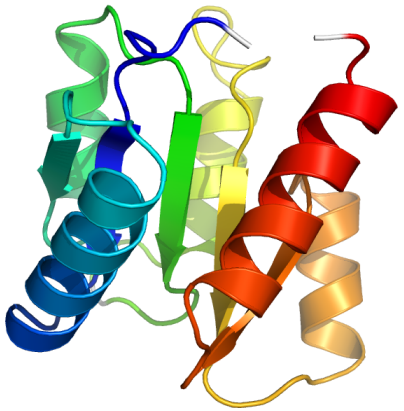
Native (CheY)



A ~1000-fold increase in computational power

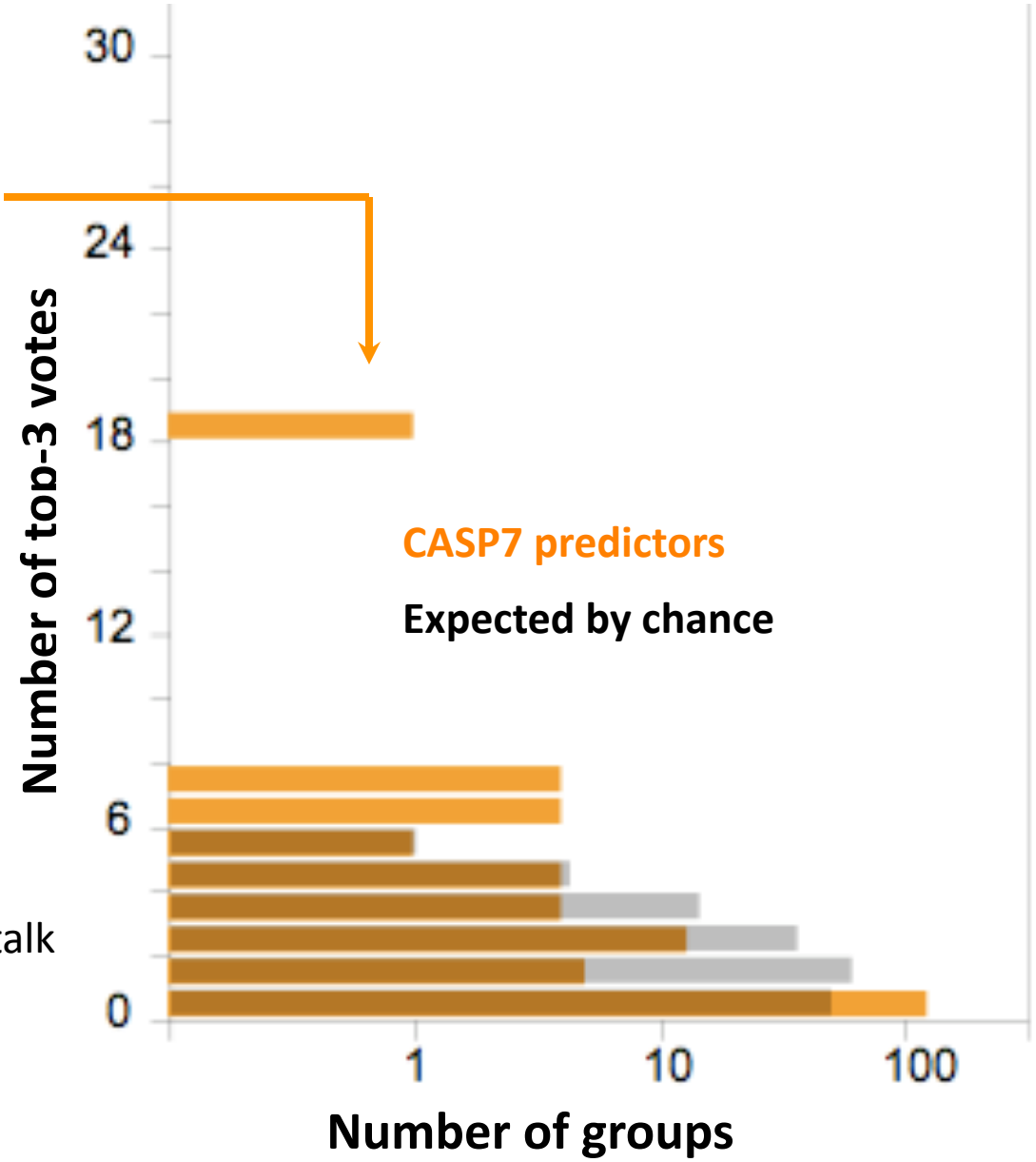


Native (CheY)



Lowest energy
Rosetta structure

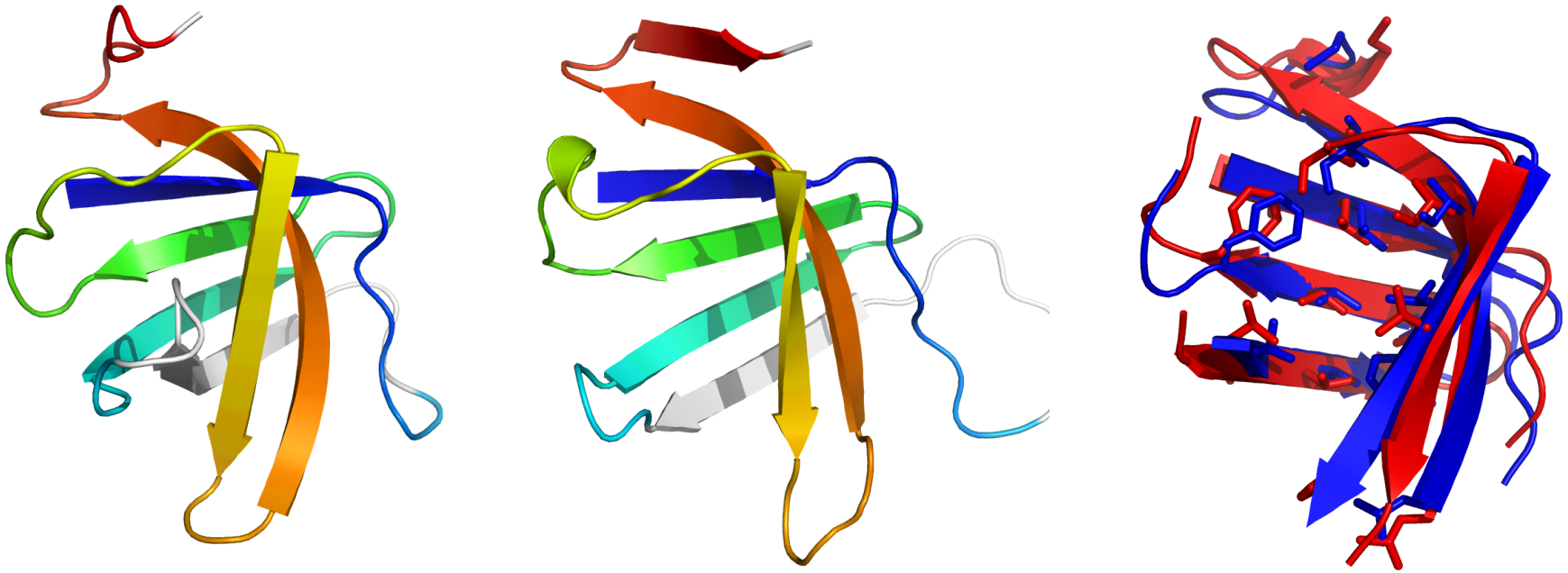
Rosetta@home in CASP7



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

De novo successes: all-β

CASP7 target T0316 (domain 3)

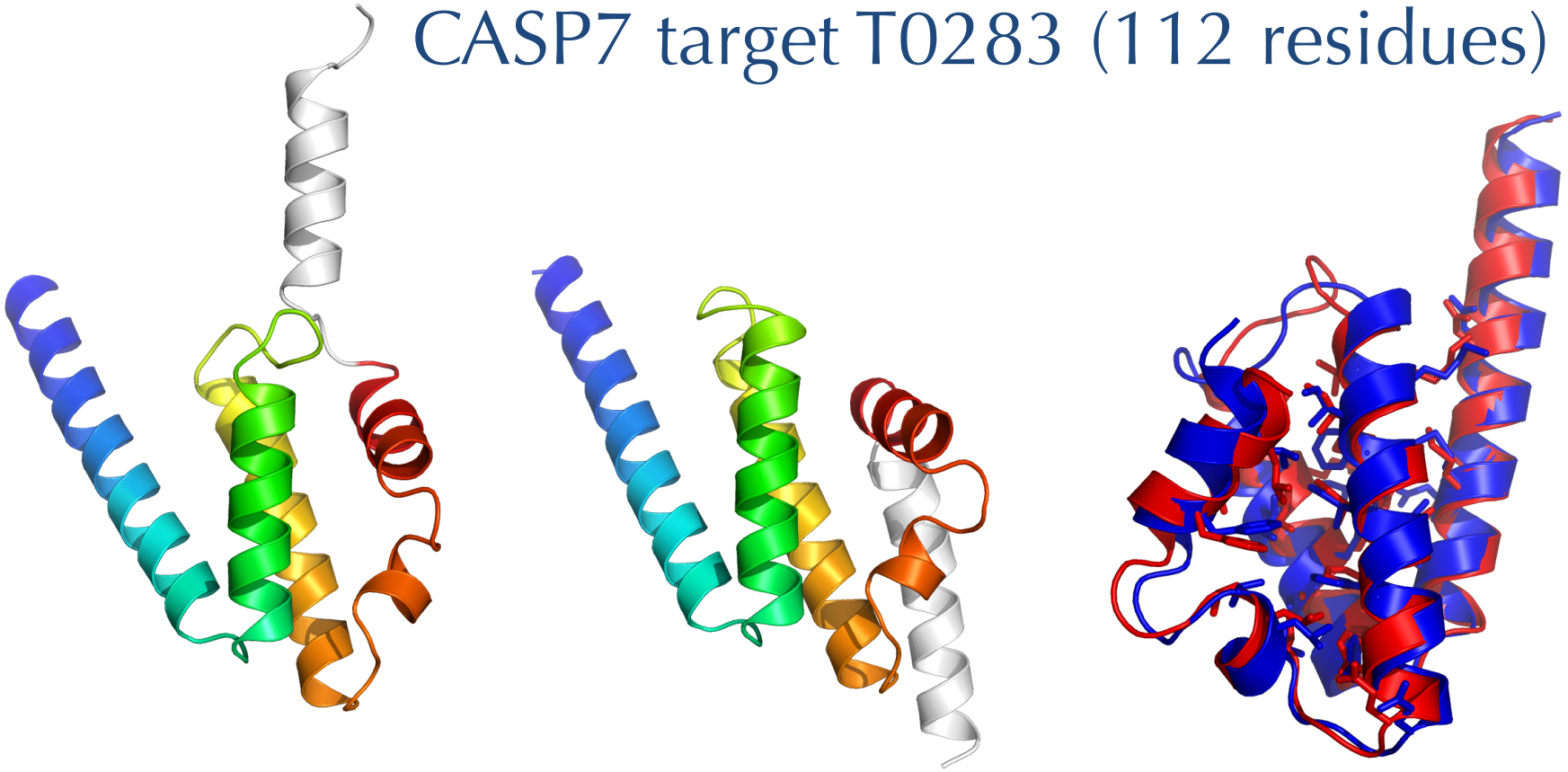


Native Model

2.0 Å over 61 residues

De novo successes: all-α

CASP7 target T0283 (112 residues)

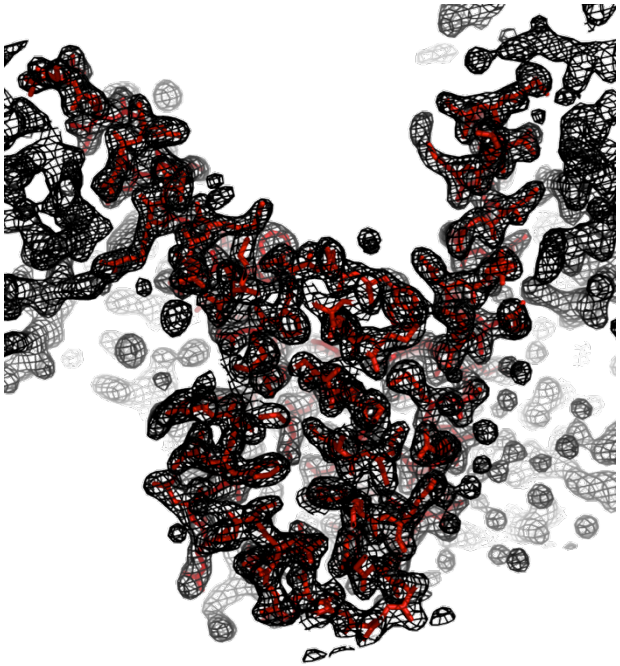


Native

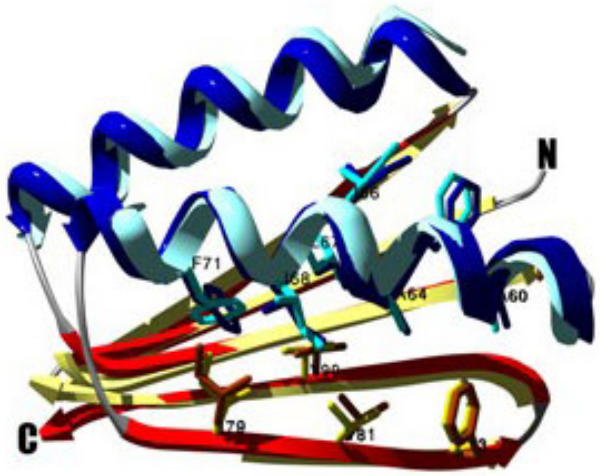
Model

1.4 Å over 90 residues

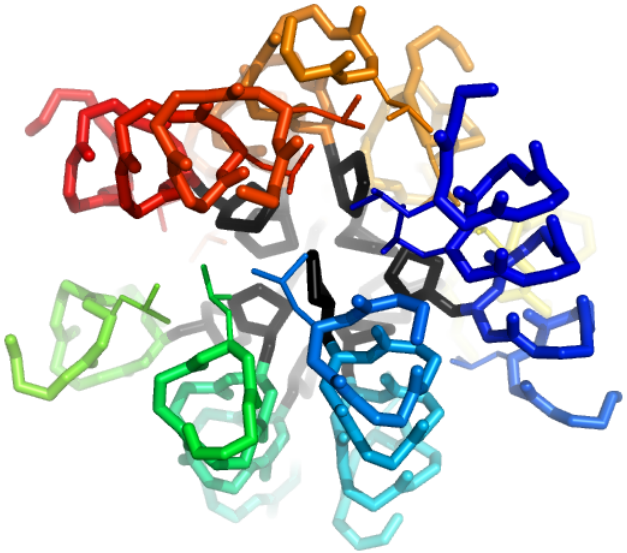
De novo modeling: **connections to the real world**



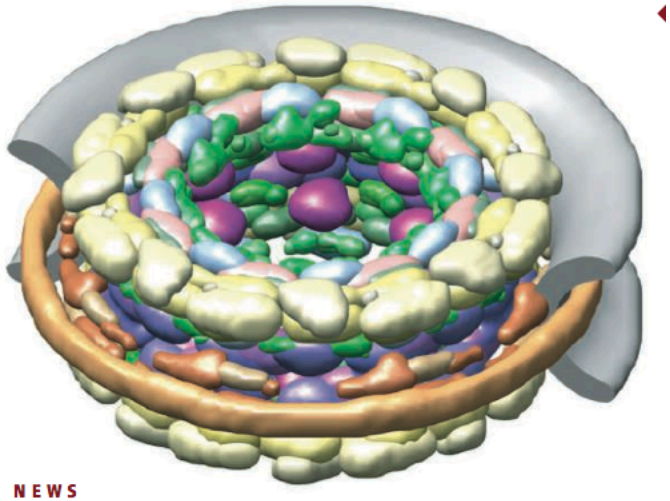
The crystallographic phase problem



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

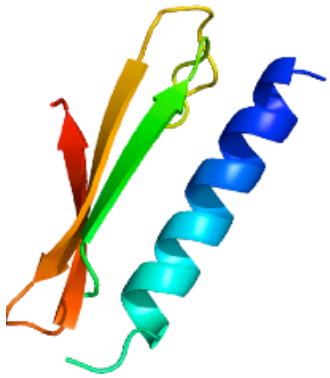
With an astronomical number of ways those chains of amino acids can potentially fold up, solving that challenge has long seemed beyond hope. But now many experts agree that key questions have been answered. Some even assert that the most daunting part of the problem—predicting the structure of unknown 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



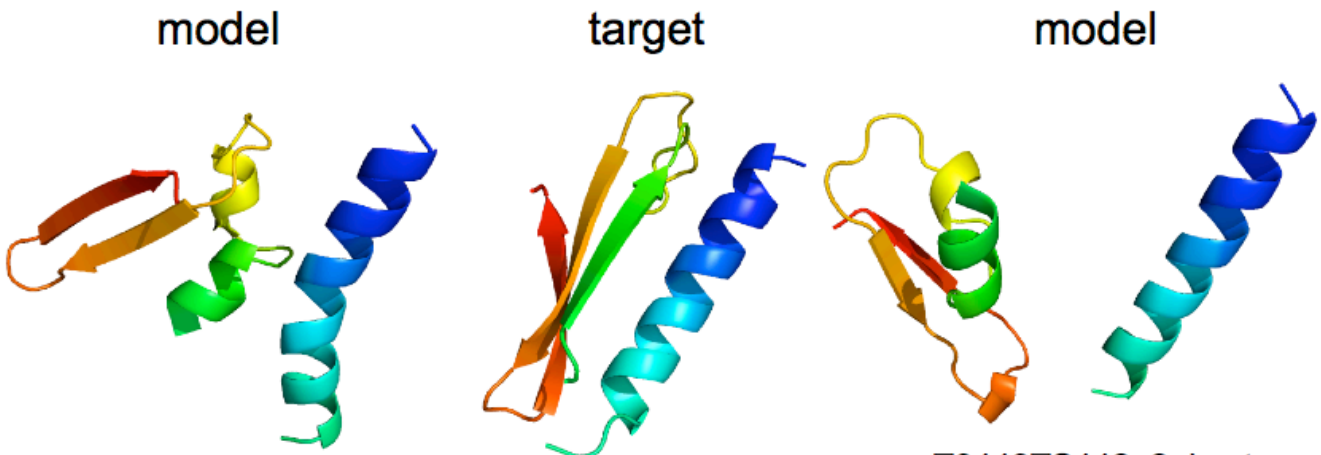
5-Dec-2008

CASP8

"Free Modeling"

Joel L. Sussman
Weizmann Institute of Science

Reality

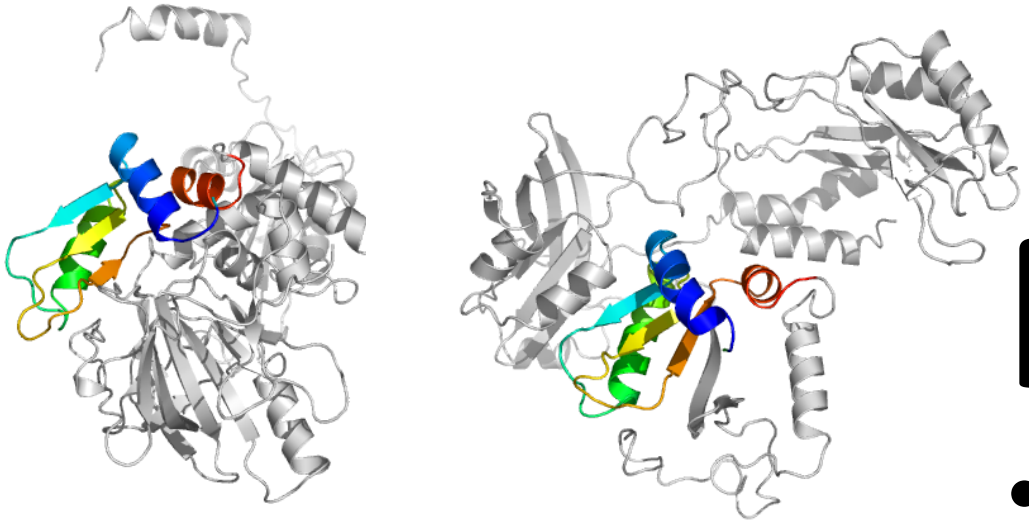


Each of the 3 visual assessors independently said: "no good model for this target"

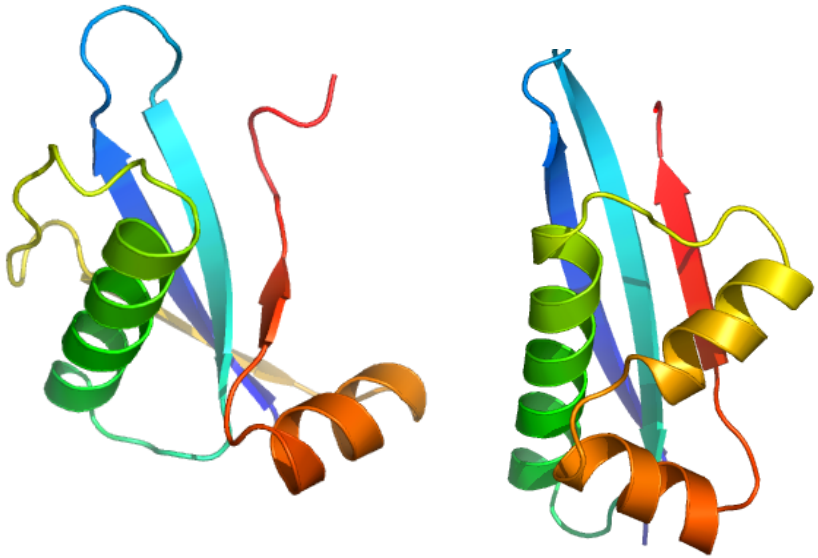
- **No real progress from CASP7**

Joel L. Sussman
Weizmann Institute of Science

Is protein folding *solved?*



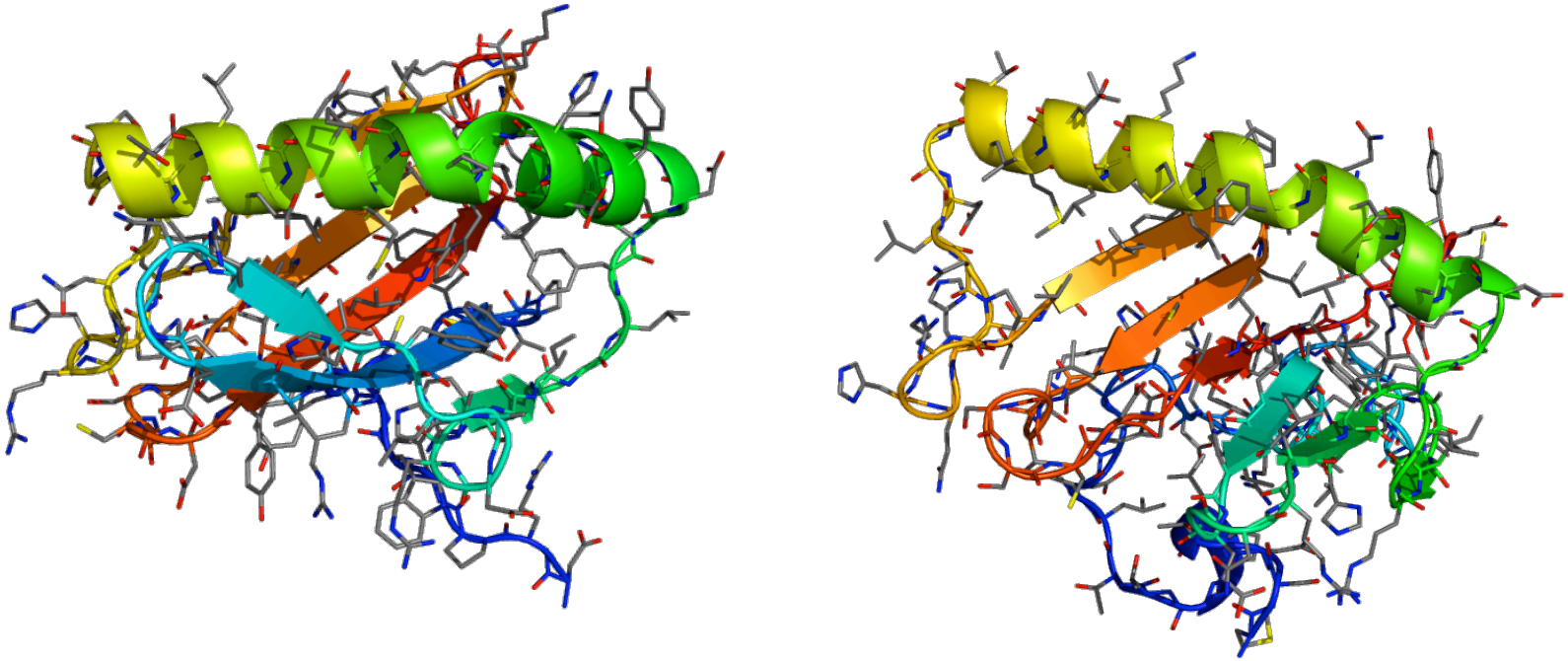
Native Model



NO!

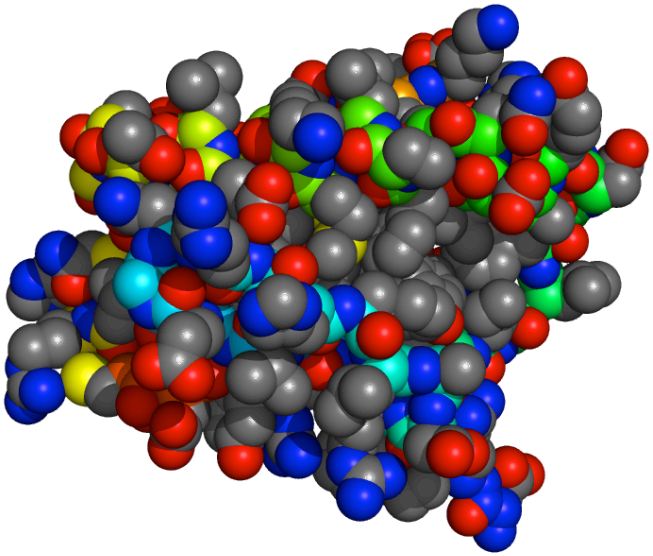
- Success in <1/3 of cases.
- Conformational sampling still a huge issue

Can you pick out the right one?

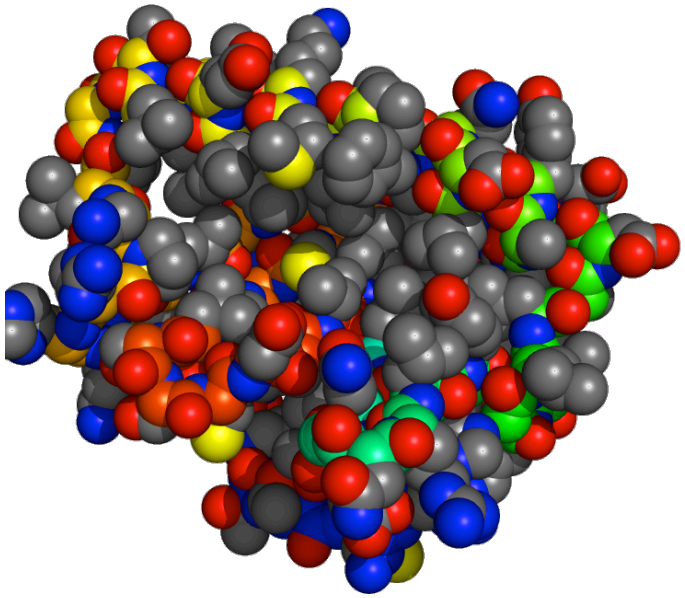


T304 (CASP7)

Can you pick out the right one?



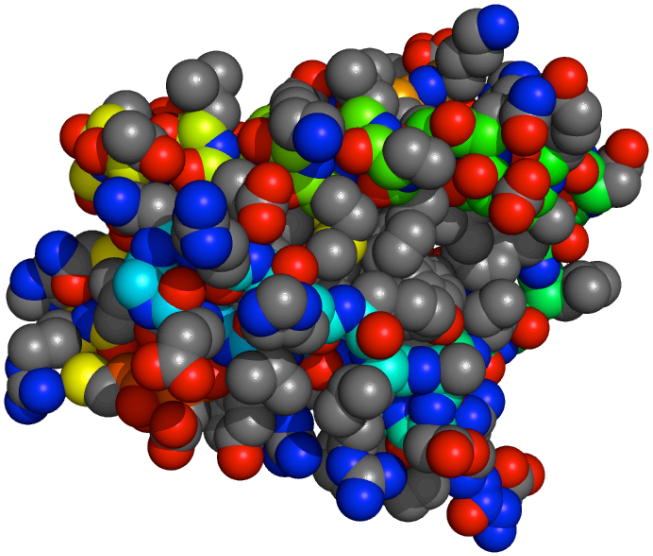
Crystallographic model



Best CASP model

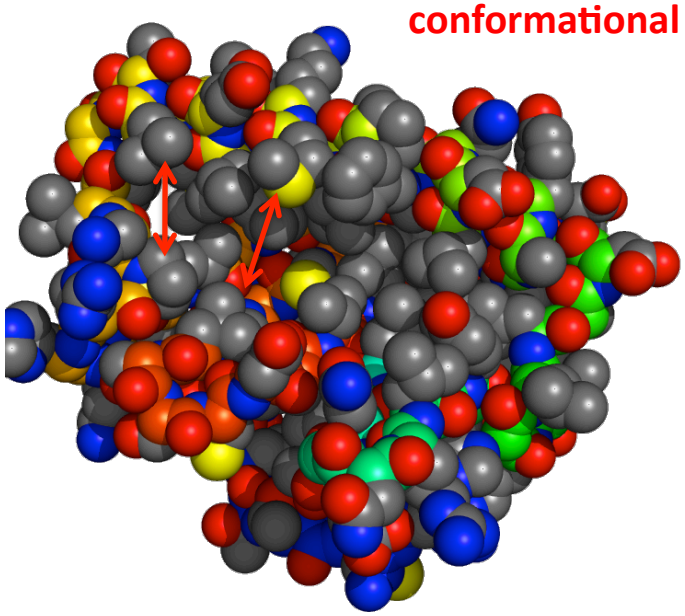
T304 (CASP7)

Can you pick out the right one?



Crystallographic model

A symptom of poor conformational sampling



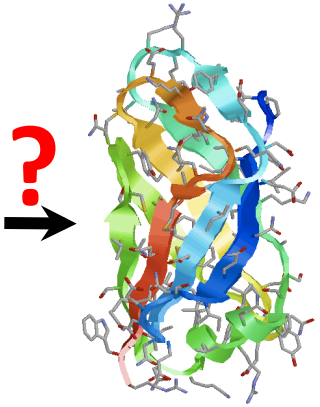
Best CASP model

T304 (CASP7)

Two fundamental problems

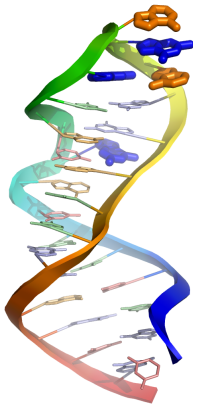
1. Predicting protein structure

GTPDIIVNAQINSEDENVLDF
IIEDEYYLKKRGVGAHIIKVAS
SPQLRLLYKNAYSTVSCGNYG
VLCNLVQNGEYDLNAIMFNC
AEIKLNKGQMLFQTKIWR



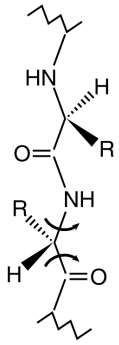
2. Predicting RNA structure

ugcuccuaguacgag
aggaccggagug



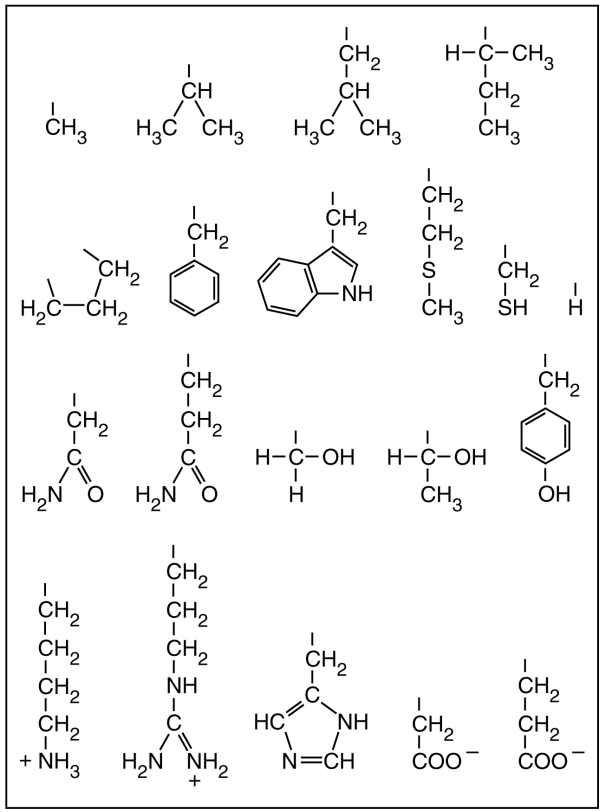
Proteins

Backbone



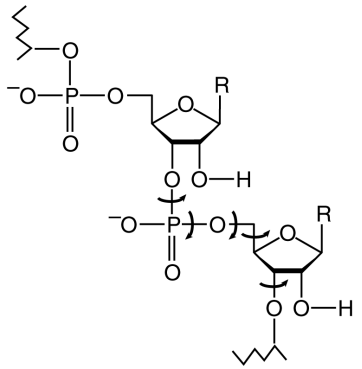
Side Chains

R ≡



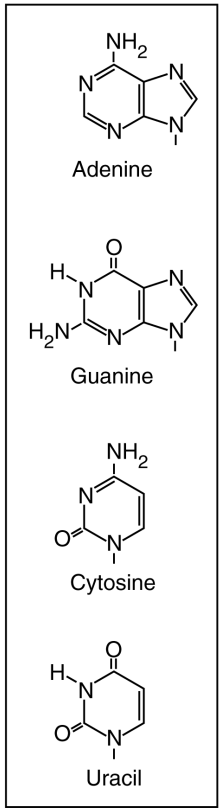
RNA

Backbone

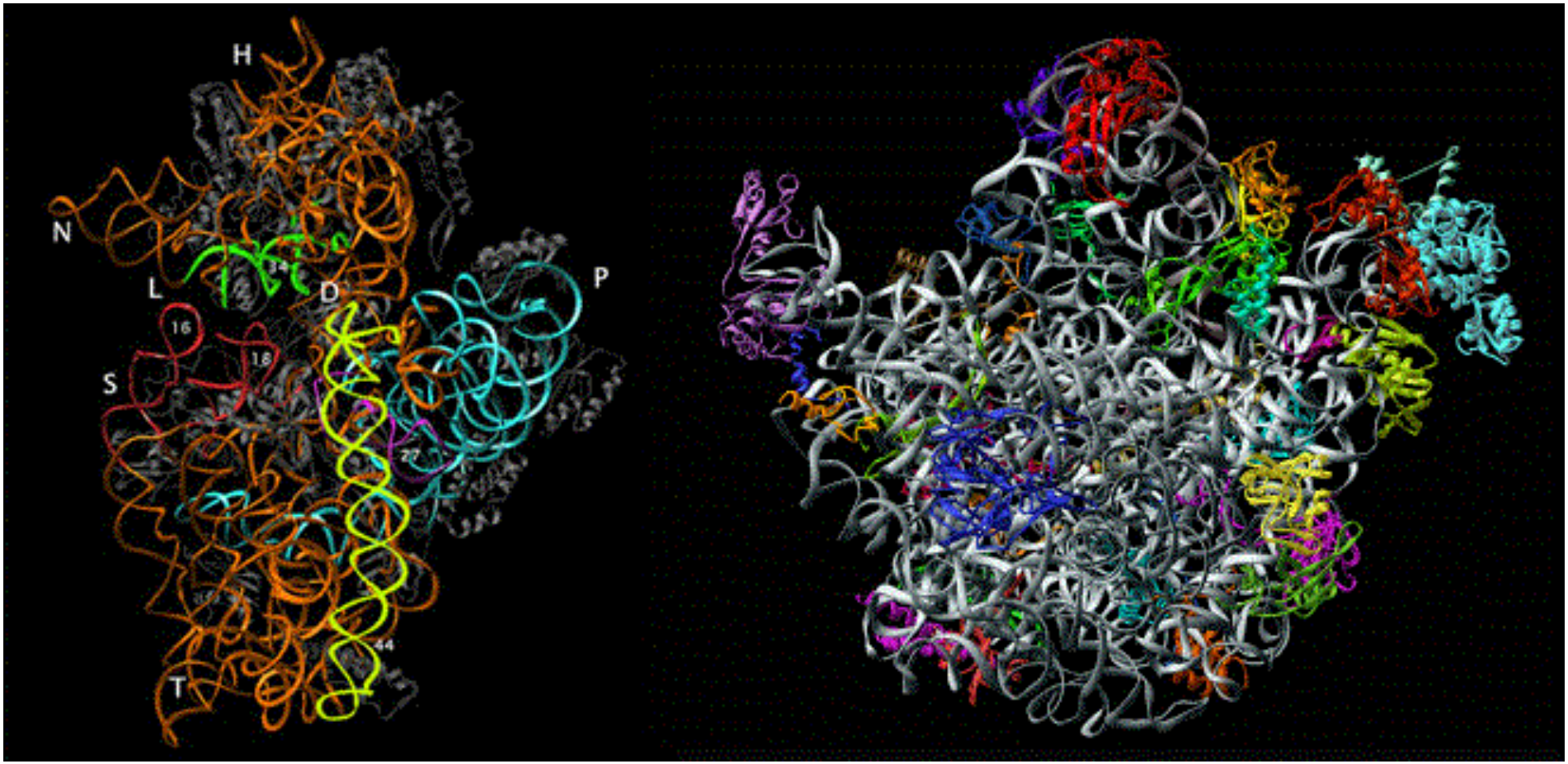


Side Chains

R ≡

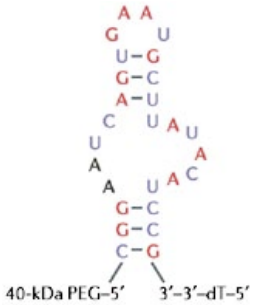


How a physicist got into biochemistry (2000)

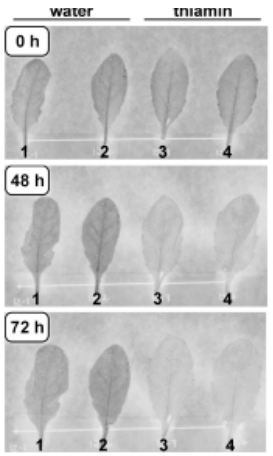


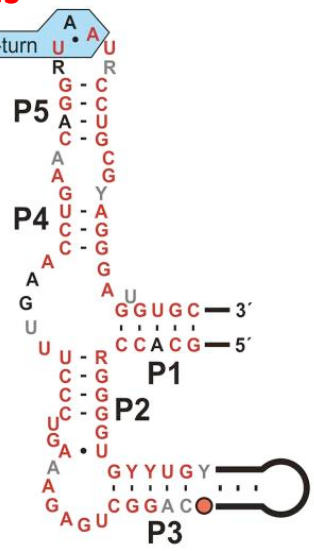
A flourishing RNA world

Engineered ribozymes and aptamers



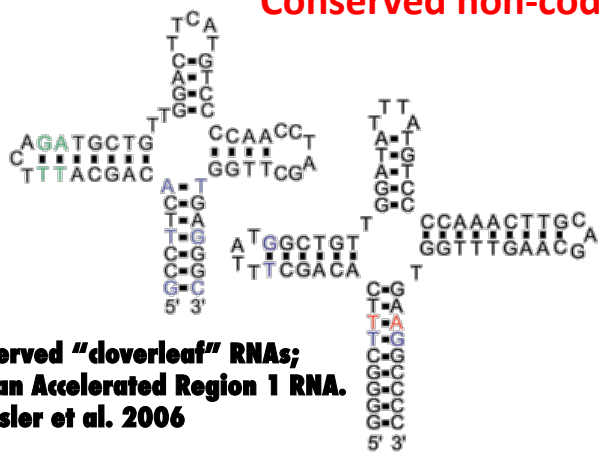
"Riboswitches"



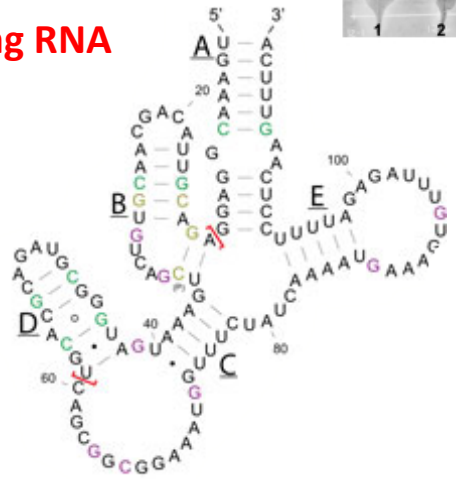


Breaker & colleagues, 2007

Conserved non-coding RNA



Conserved "cloverleaf" RNAs; Human Accelerated Region 1 RNA. Haussler et al. 2006

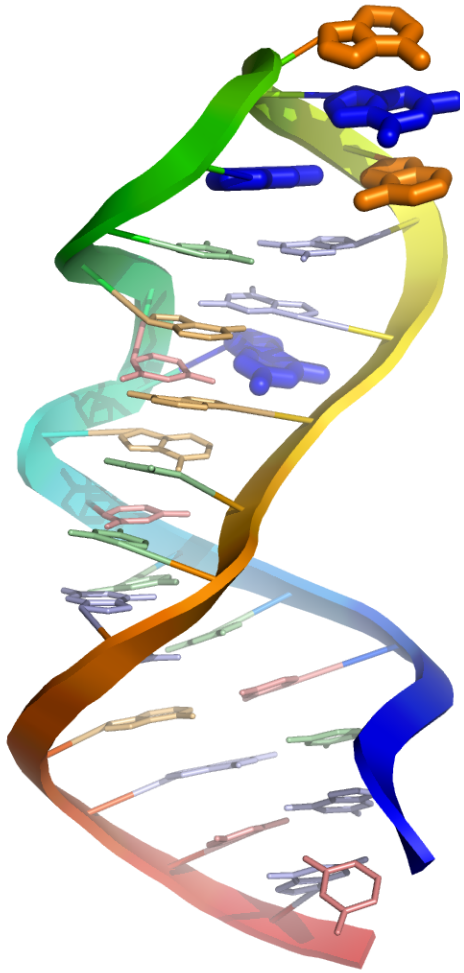


The Das Lab

Goal: Nucleic Acid Structures You Can Trust

ugcuccu
aguacga
gaggacc
ggagug

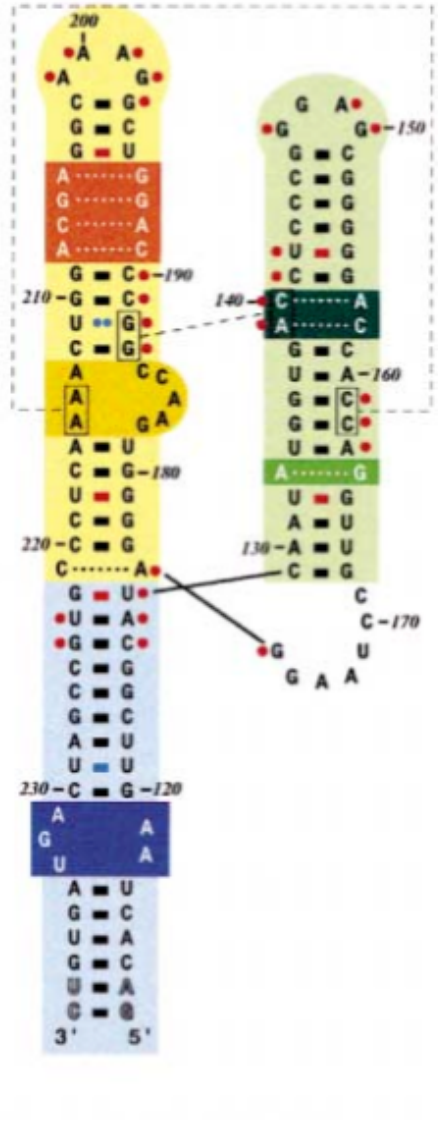
?



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

Words and grammar for RNA?

GACACUAAGUUCGGCA
UCAAUAUGGUGACCUC
CCGGGAGCGGGGGACC
ACCAGGUUGCCUAGAG
GGGUGAACCGGCCCAG
GUCGGAAACGGAGCAG
GUCAAAACUCCCGUGC
UGAUCAGUAGUGU

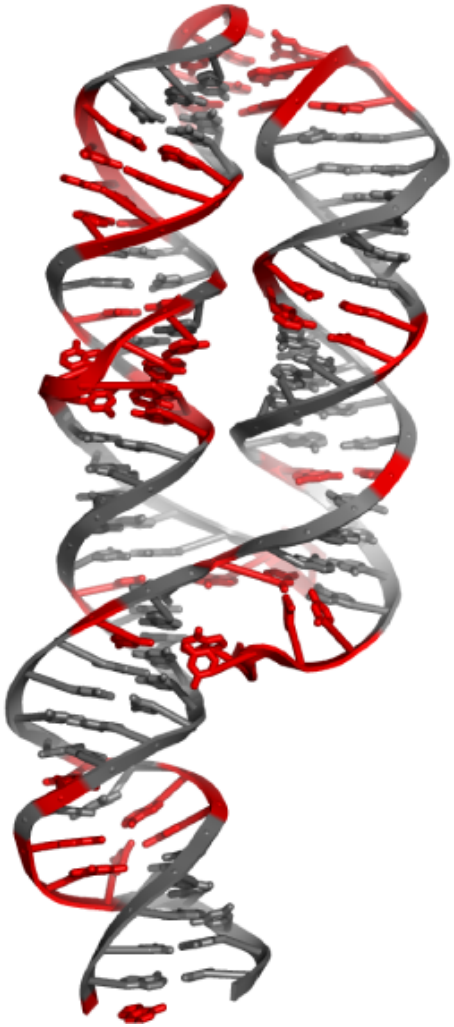


Signal Recognition Particle RNA
Oubridge et al., 2002

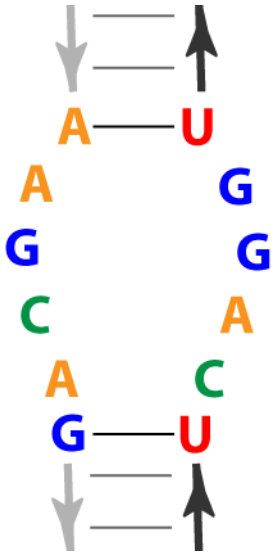
Words and grammar for RNA?

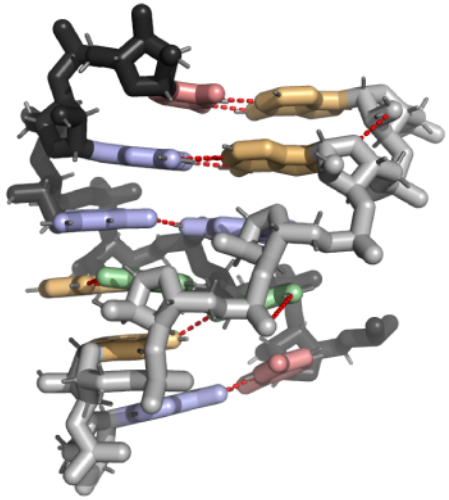
Canonical double helices

Non-canonical regions

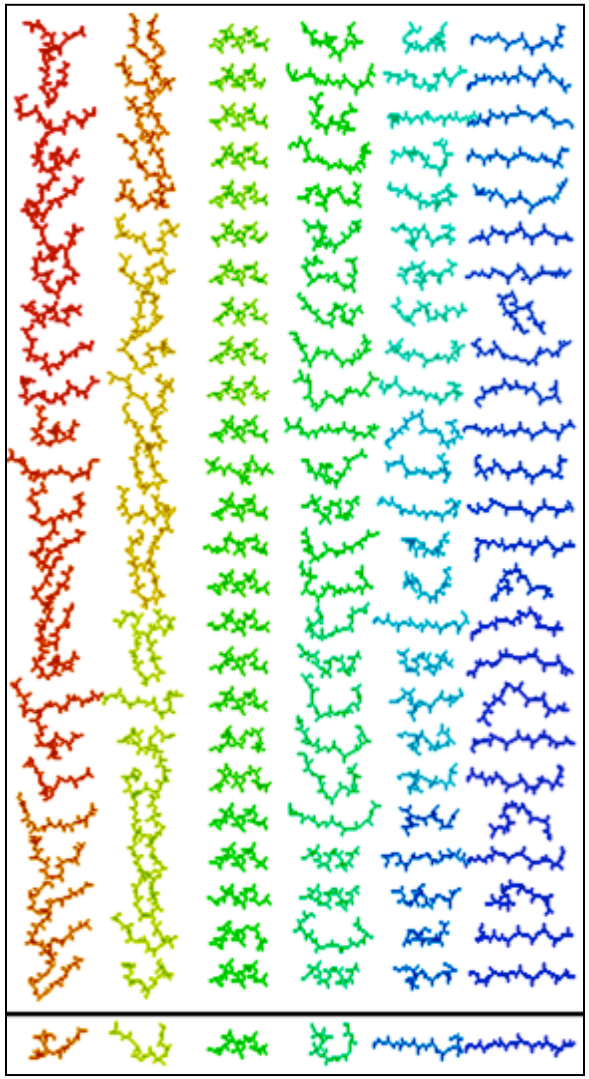


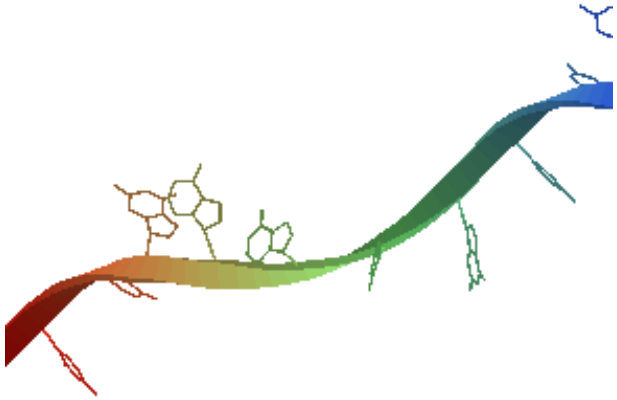
Words and grammar for RNA?





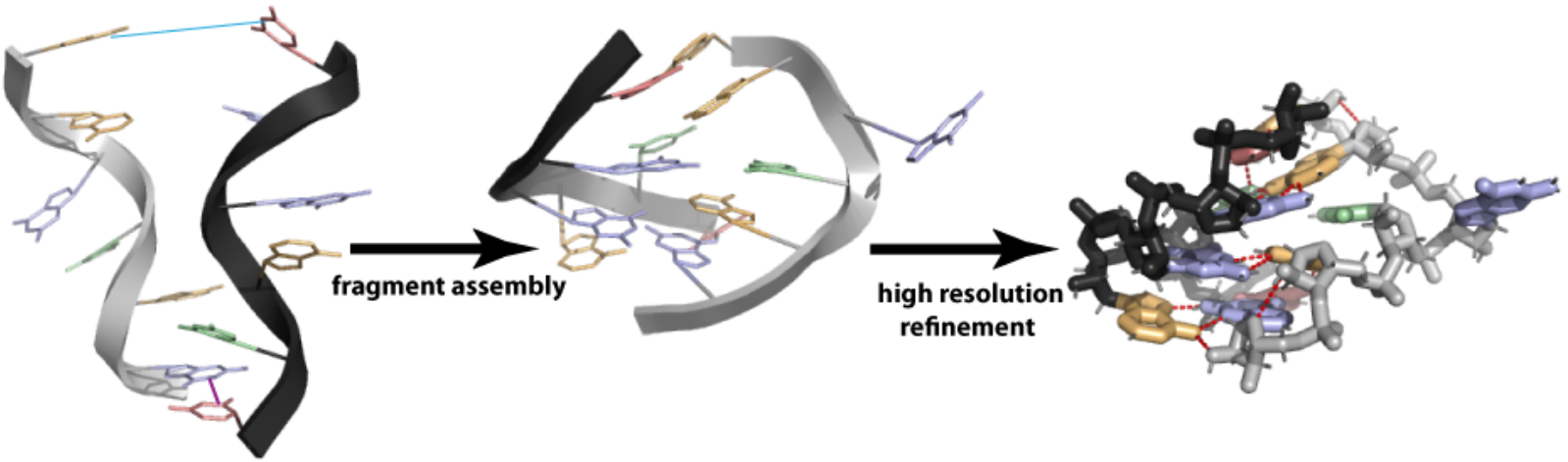
De novo modeling





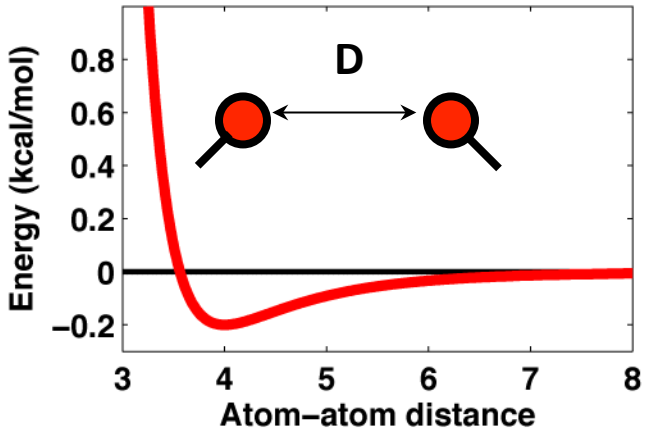
Fragment **A**ssembly of **RNA** (**FARNA**)

De novo modeling

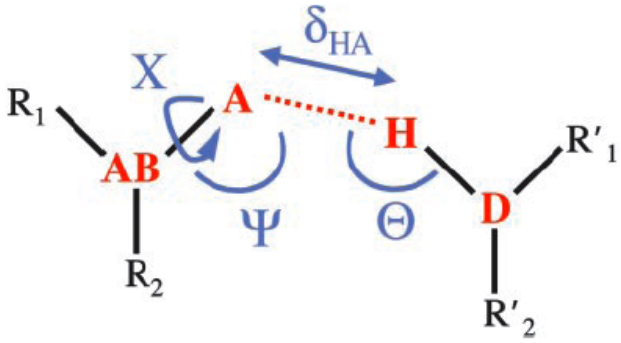


Ingredients of a high resolution potential

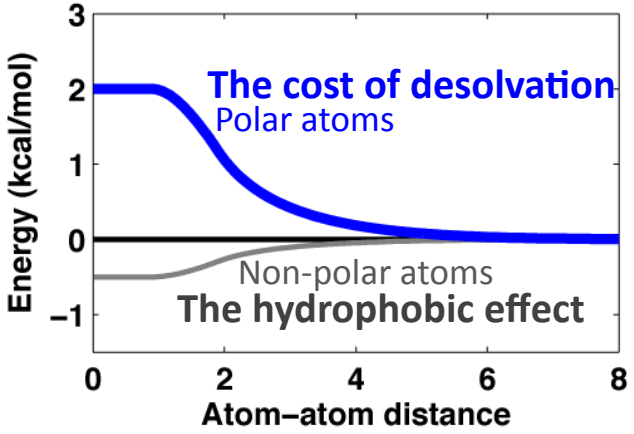
1. Van der waals packing



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.

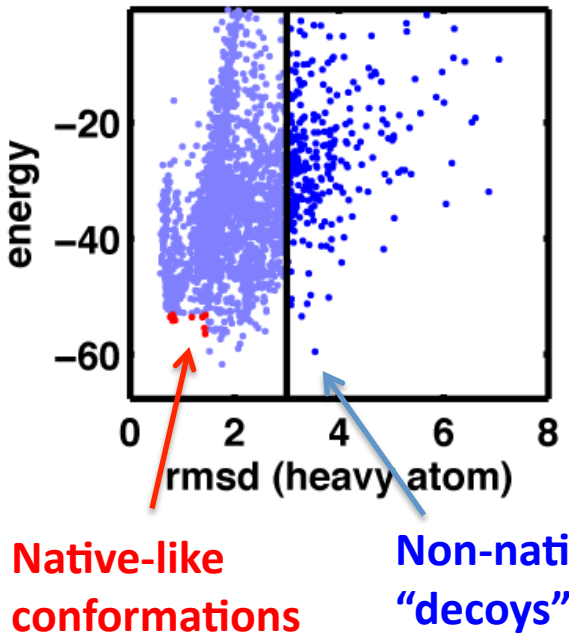
– Michael Levitt

"Detailed molecular model of transfer RNA", *Nature* 1969.

Does it work?

Native-state discrimination

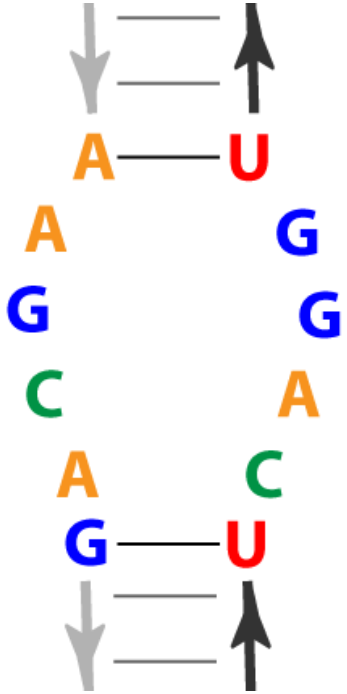
Low resolution
(FARNA) energy

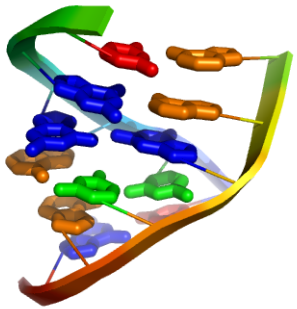


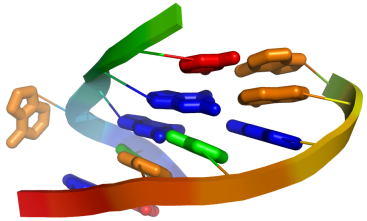
Native-like conformations

Non-native "decoys"

The most conserved region of the signal recognition particle



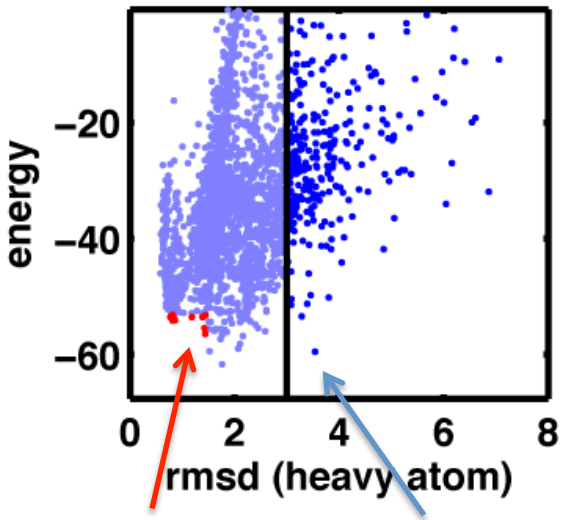




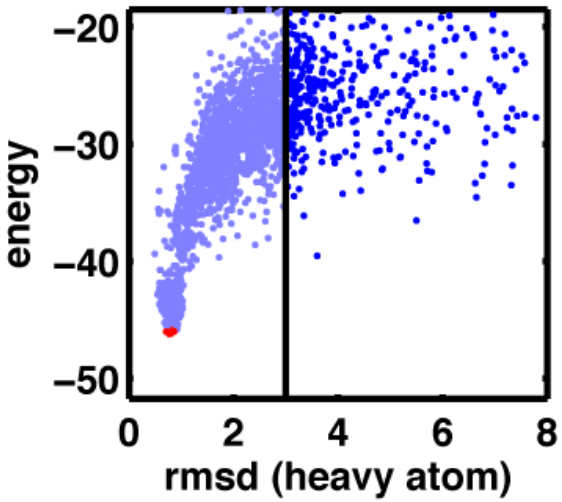
Native-state discrimination

The most conserved region of the signal recognition particle

Low resolution (FARNA) energy

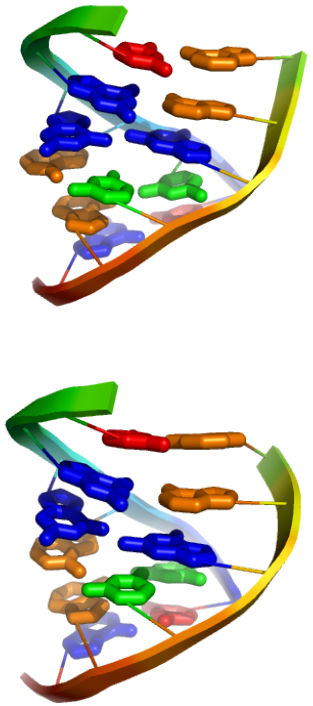


High resolution energy



Native-like conformations

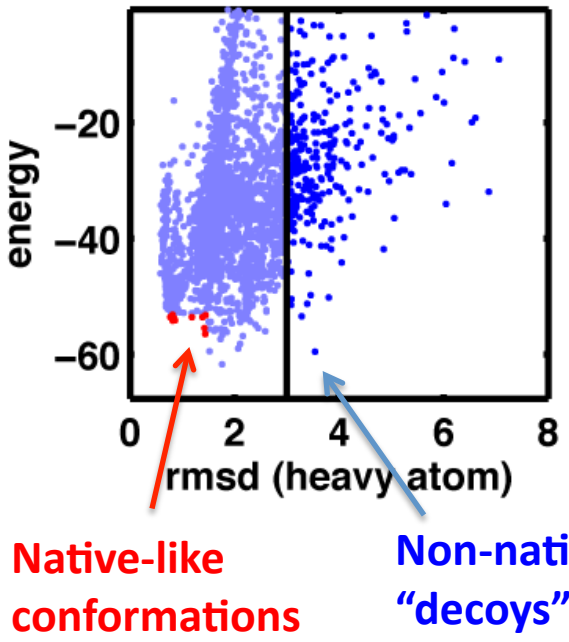
Non-native "decoys"



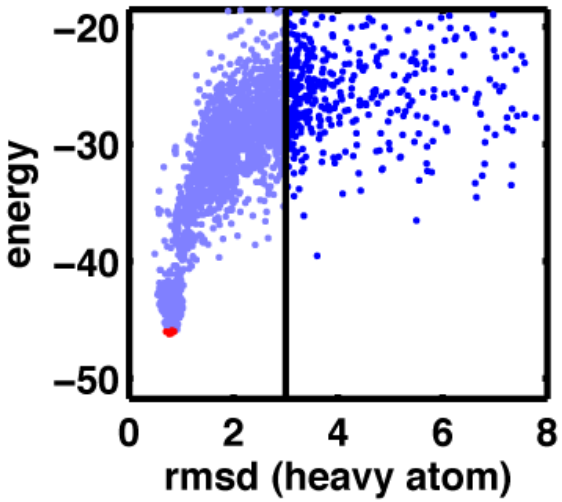
Native-state discrimination

The most conserved region of the signal recognition particle

Low resolution (FARNA) energy

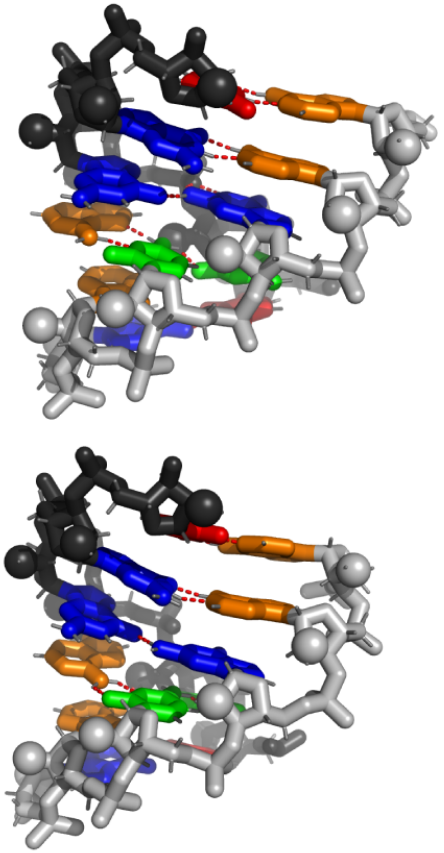


High resolution energy

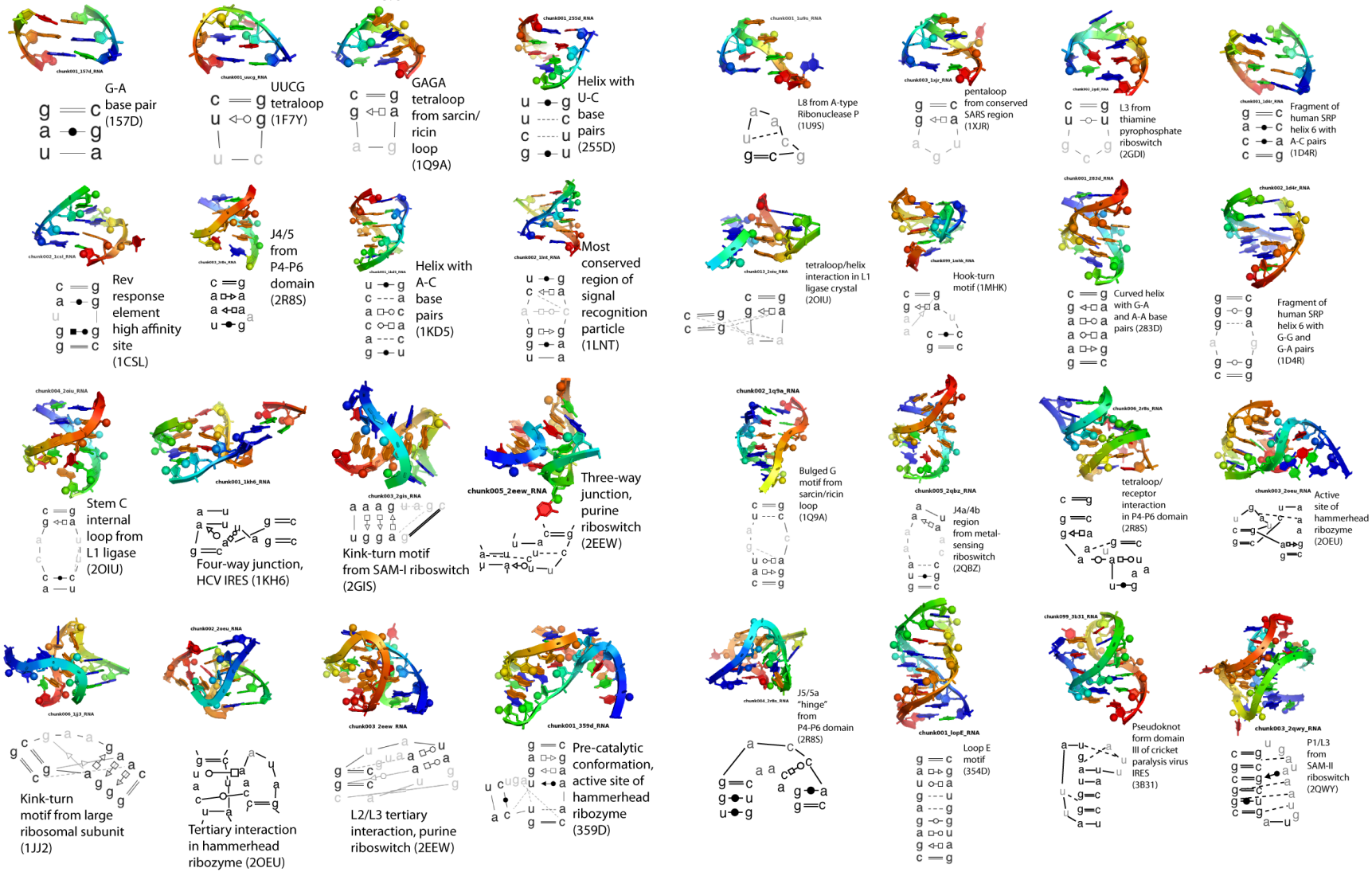


Native-like conformations

Non-native "decoys"

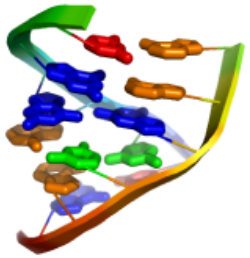


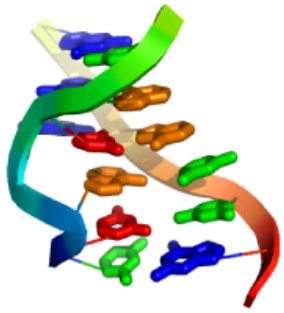
Can we decipher all the known “words”?

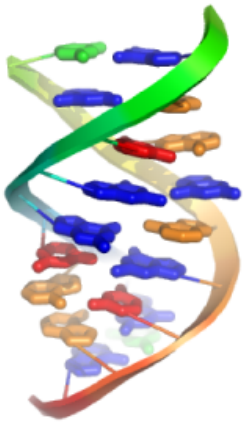


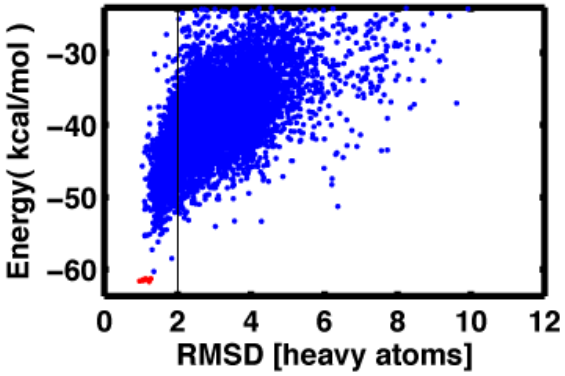
De novo modeling

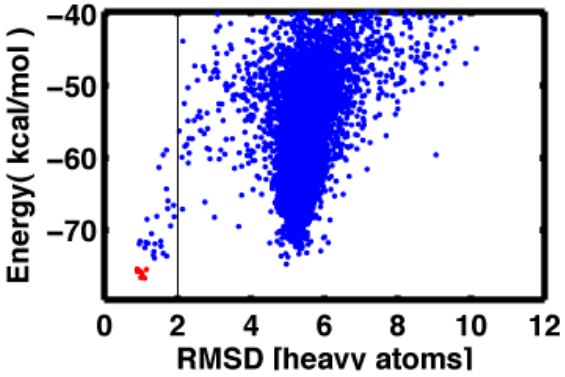
Native

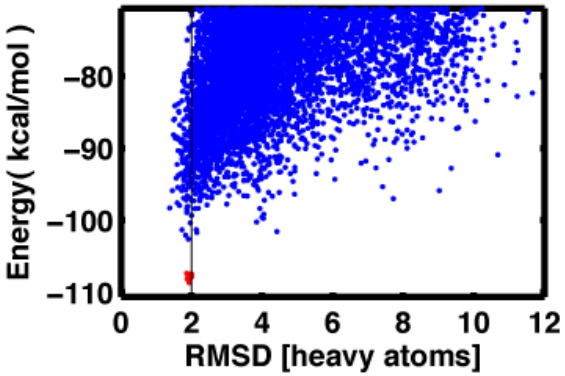






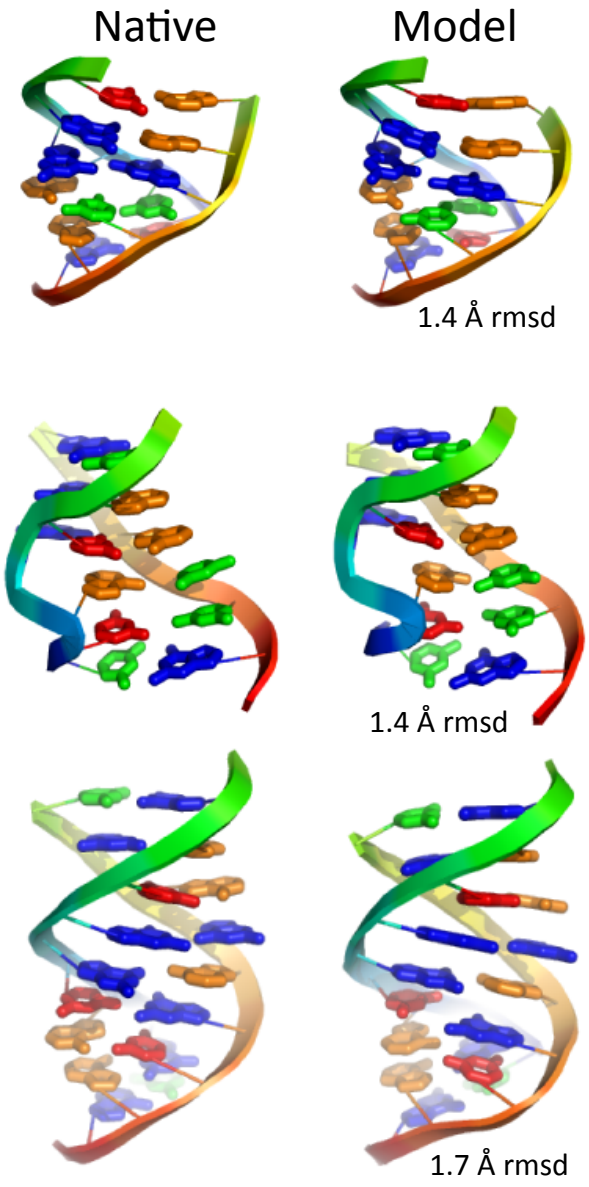


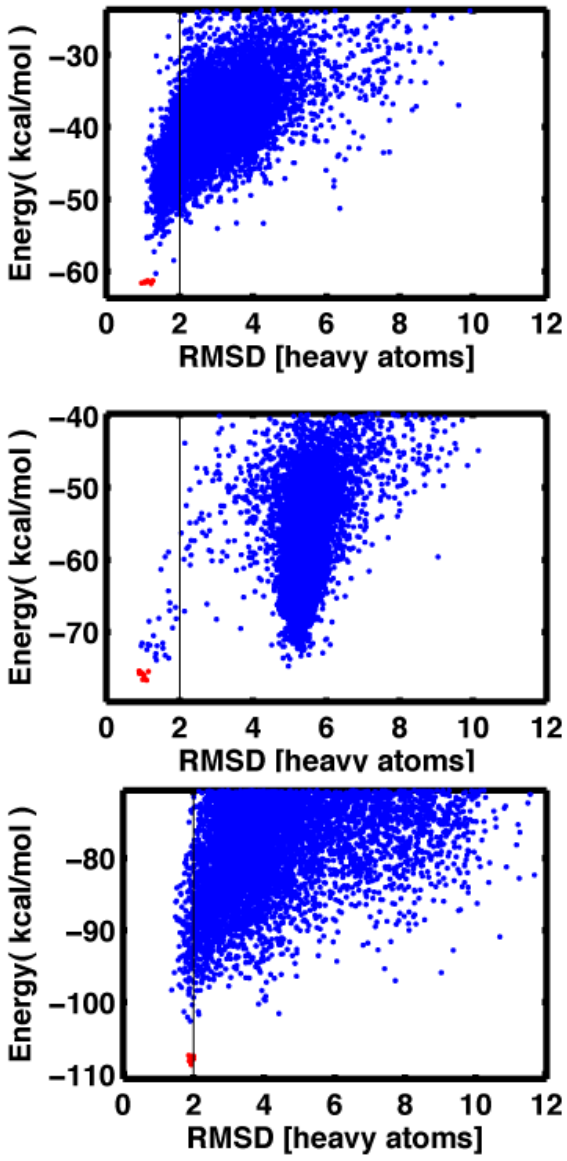




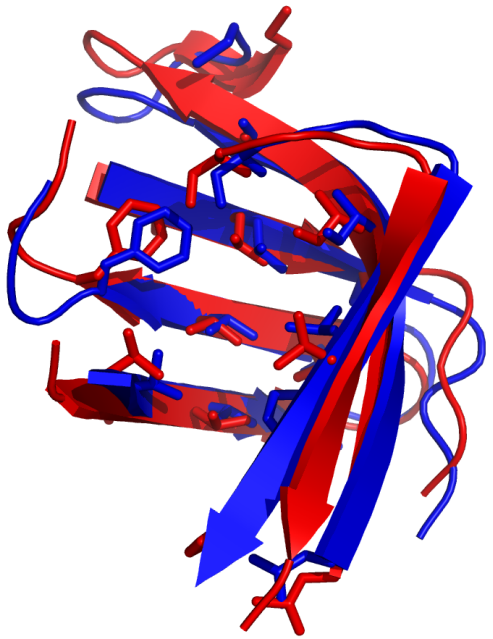
In half the cases, de novo modeling achieves < 2.0 Å structures, and selects them.

De novo modeling



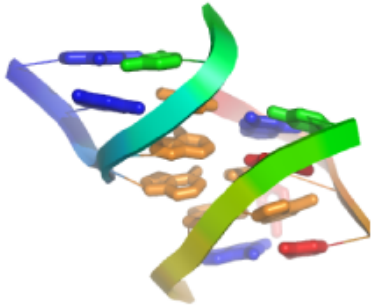


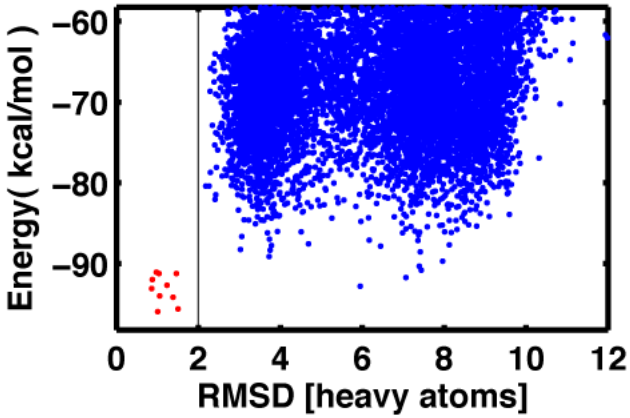
In half the cases, de novo modeling achieves < 2.0 Å structures, and selects them.



De novo modeling

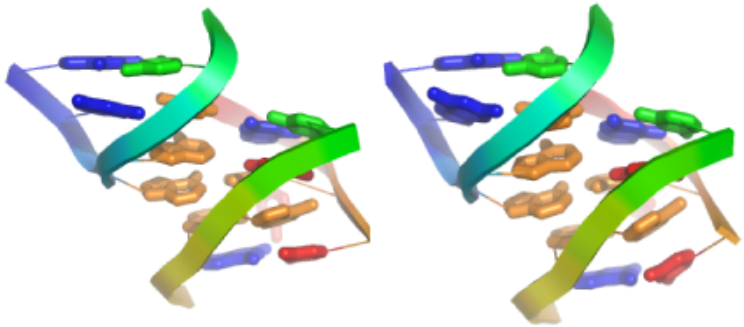
The biggest bottleneck: conformational sampling



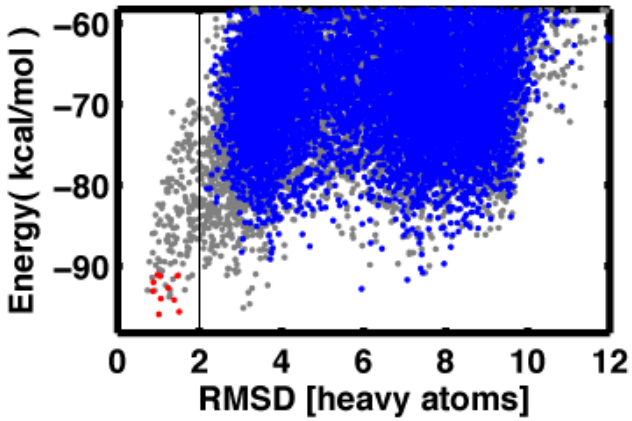


De novo modeling

The biggest bottleneck: conformational sampling



1.0 Å rmsd



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.

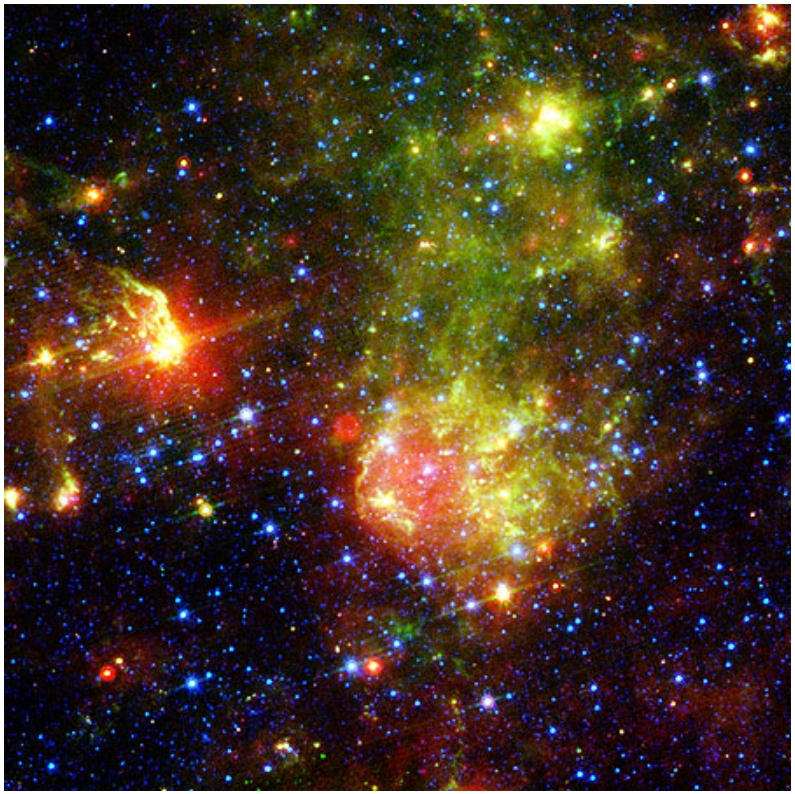
Structure prediction for CASP8 with all-atom refinement using Rosetta

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.

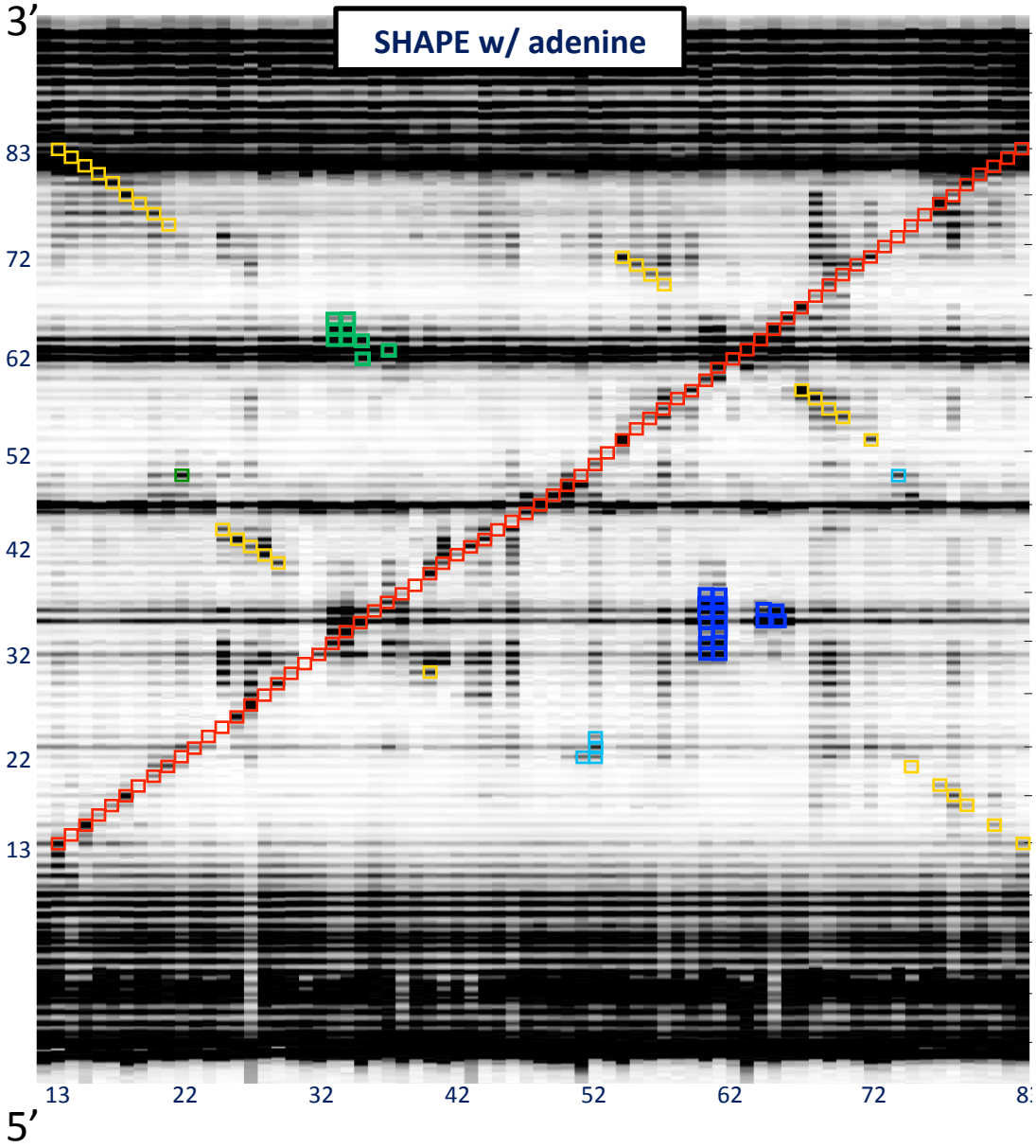
Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling

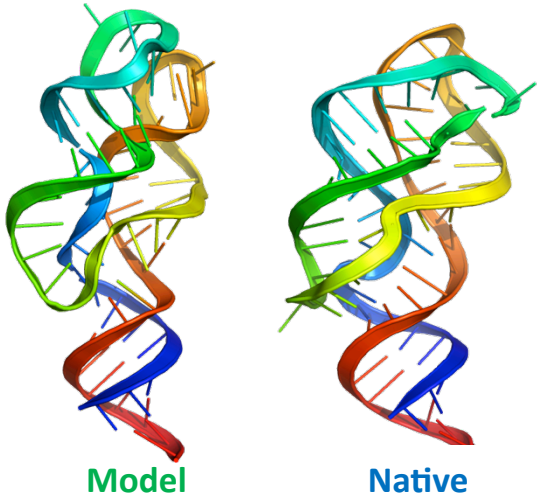
Daniel J Mandell[1,2], Evangelos A Coutsias[3] & Tanja Kortemme[1,2,4]

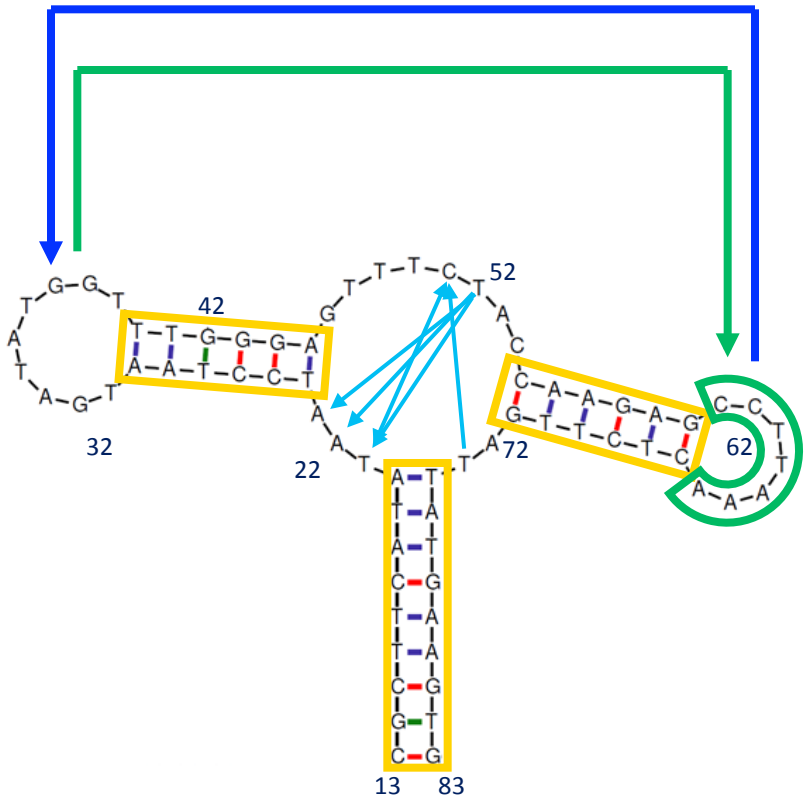


Beating the “astronomical” conformational sampling problem

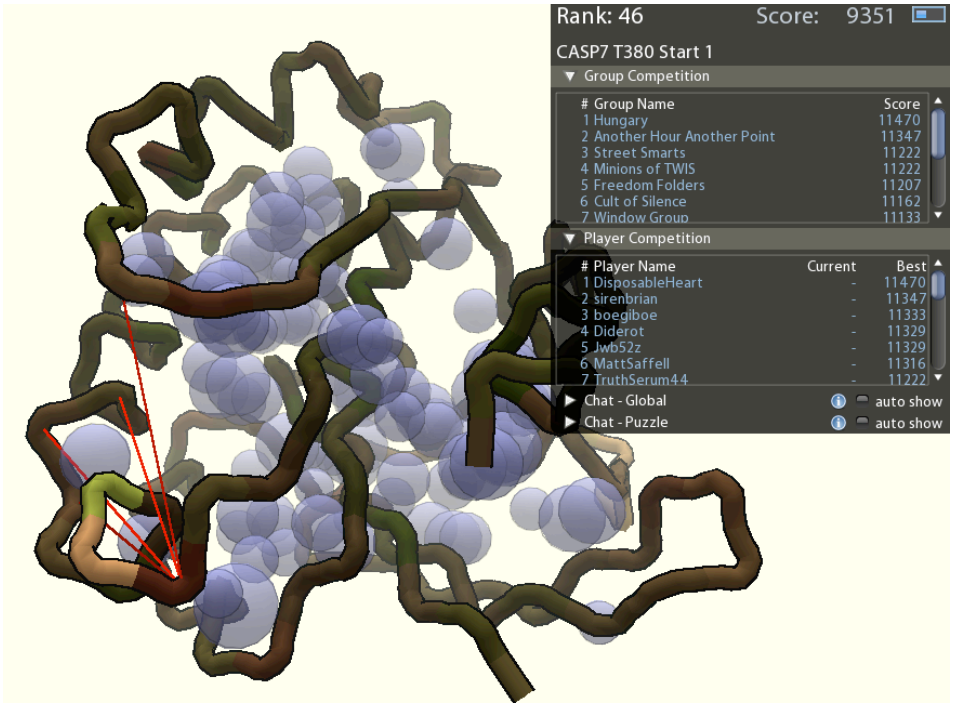
Solution 1? Data







Solution 2? Humans



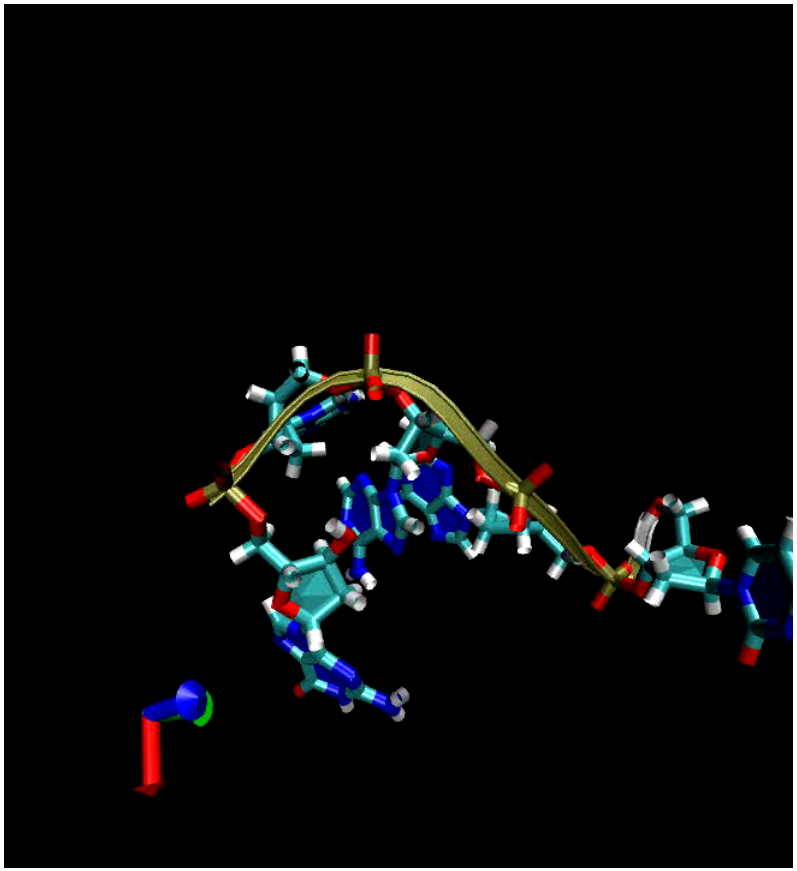
FOLD.IT

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

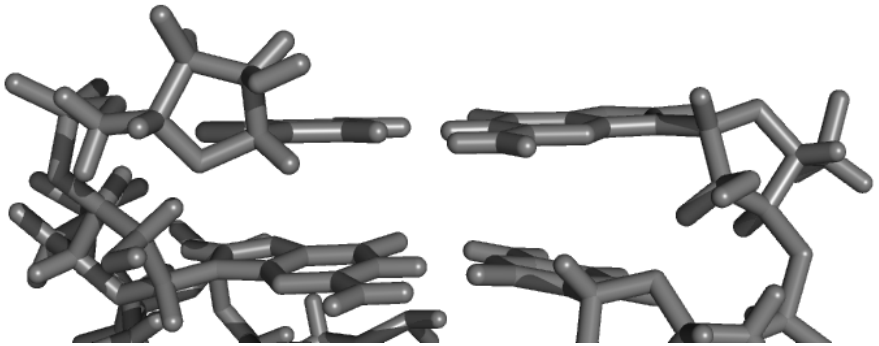
Solution 3? Physics



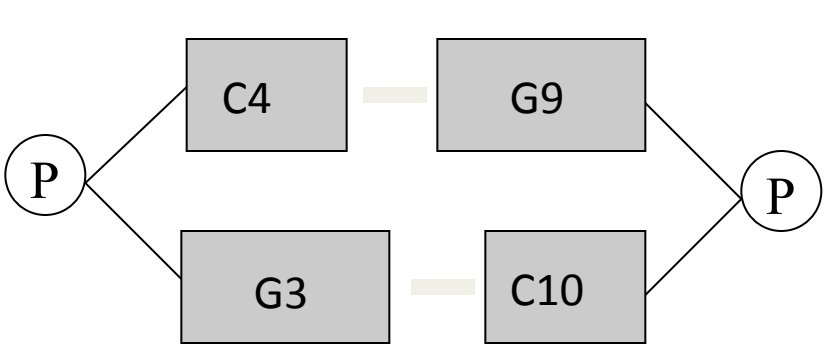
Computationally expensive, but getting faster

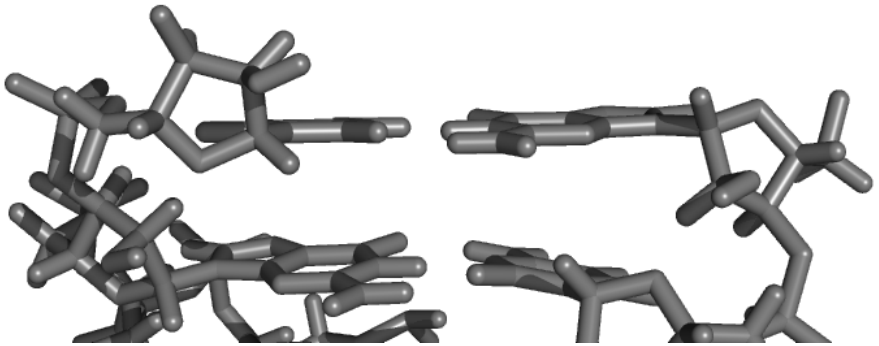
Still no case of blind predictions of structure

Solution 4. Better algorithms

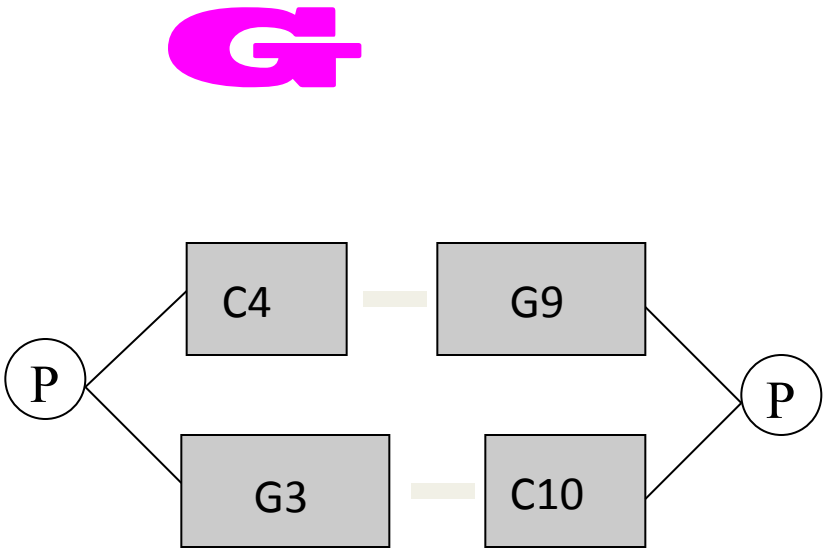


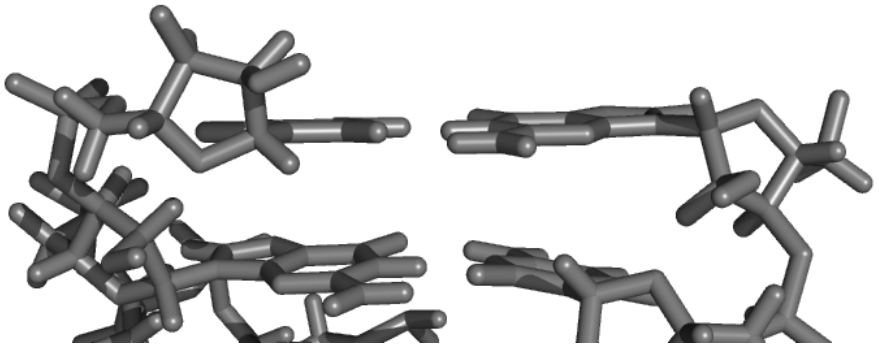
Step-by-step sampling



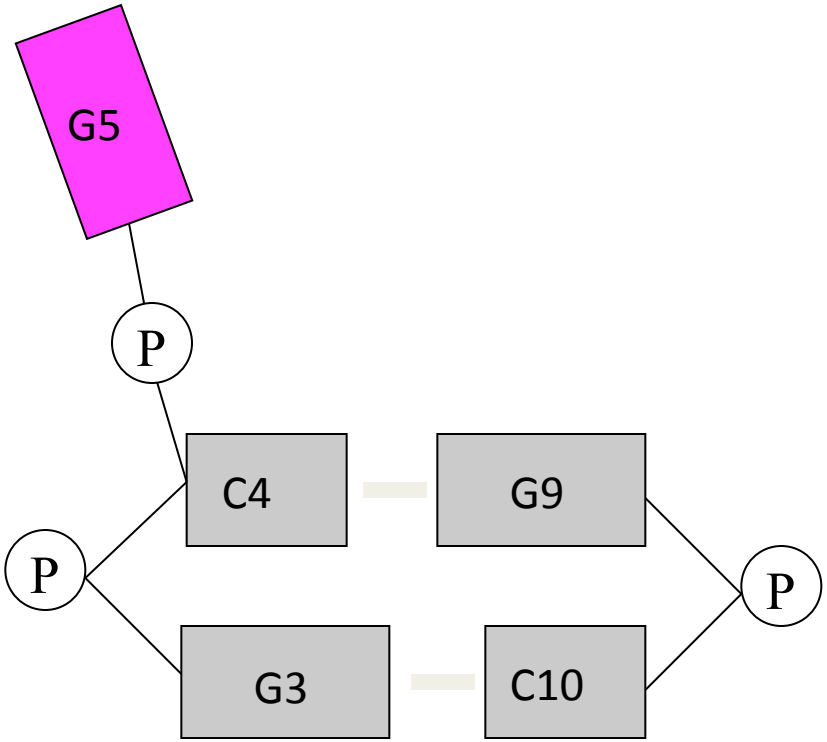


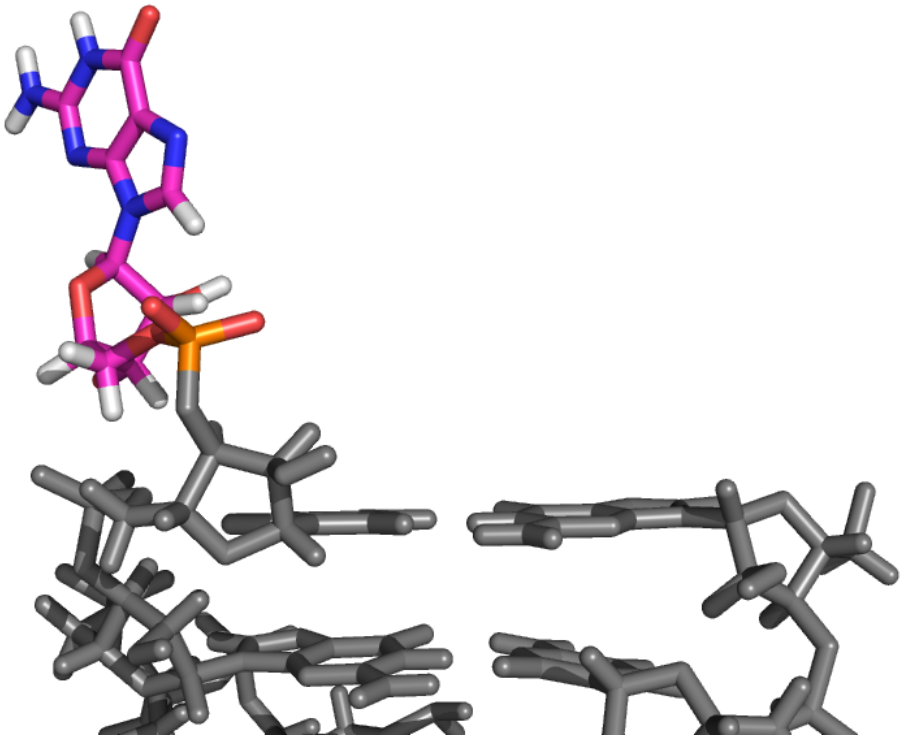
Step-by-step sampling



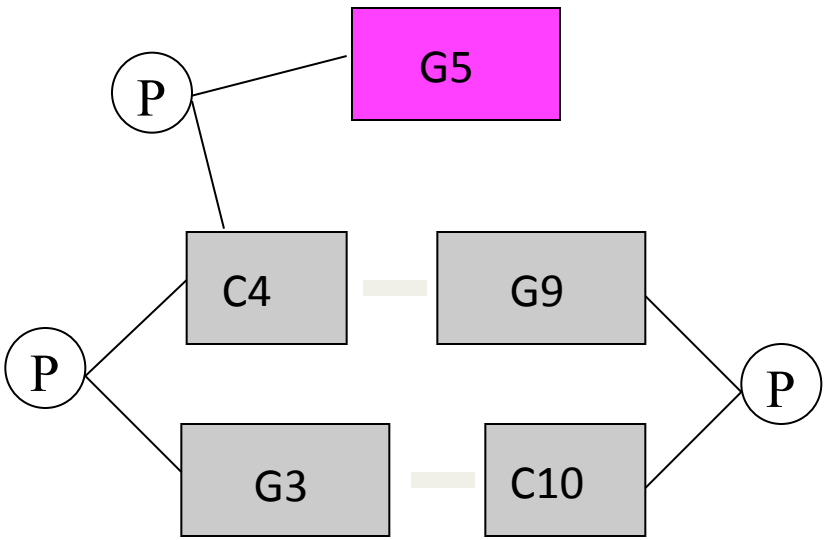


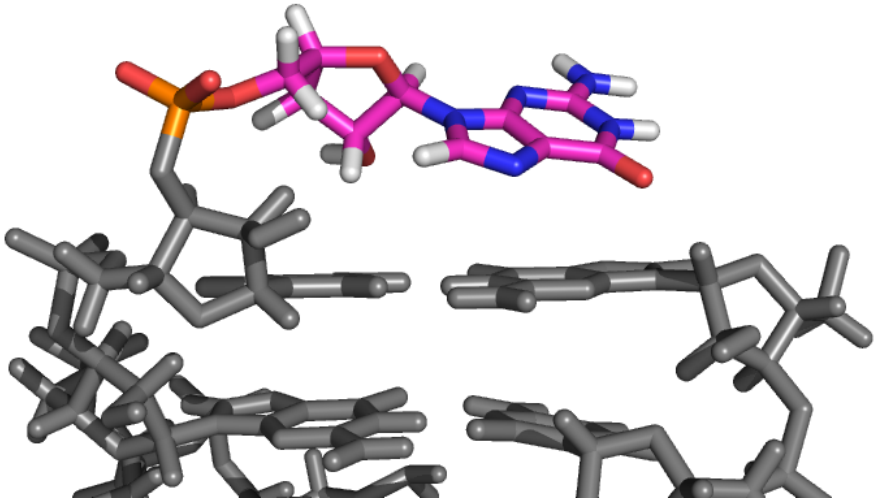
Step-by-step sampling



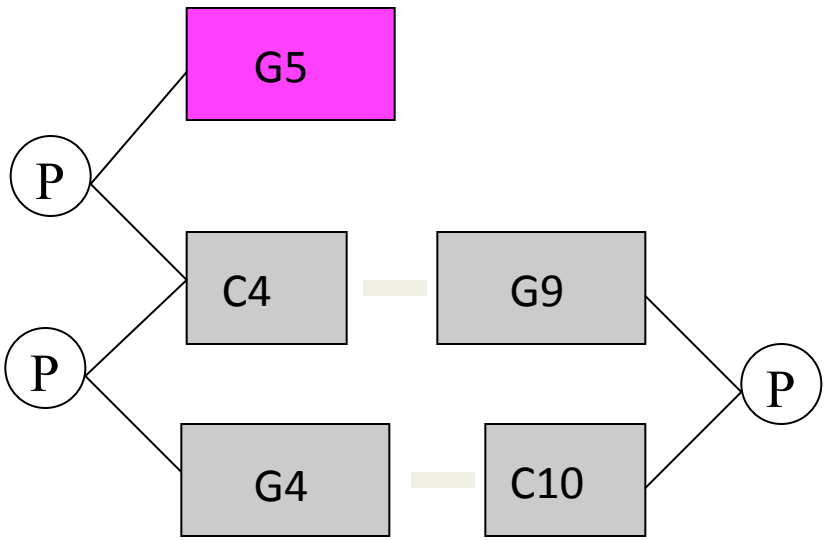


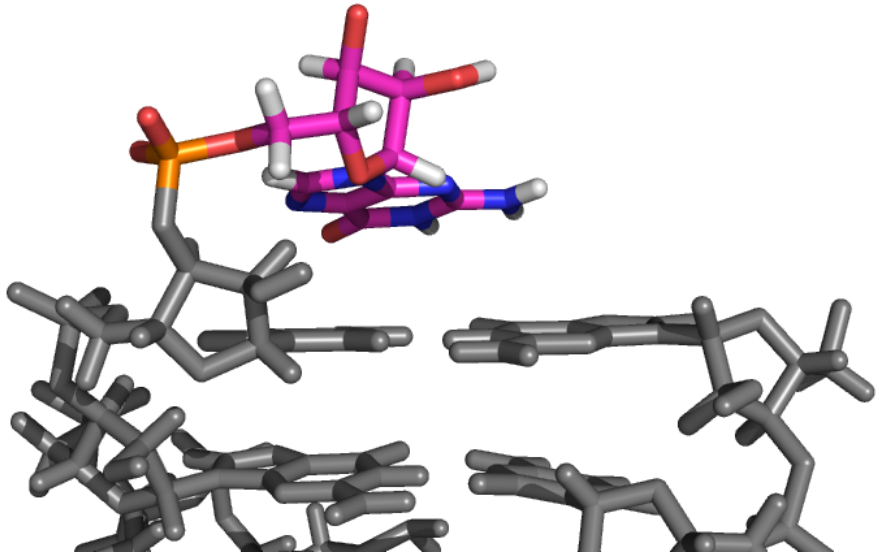
Step-by-step sampling



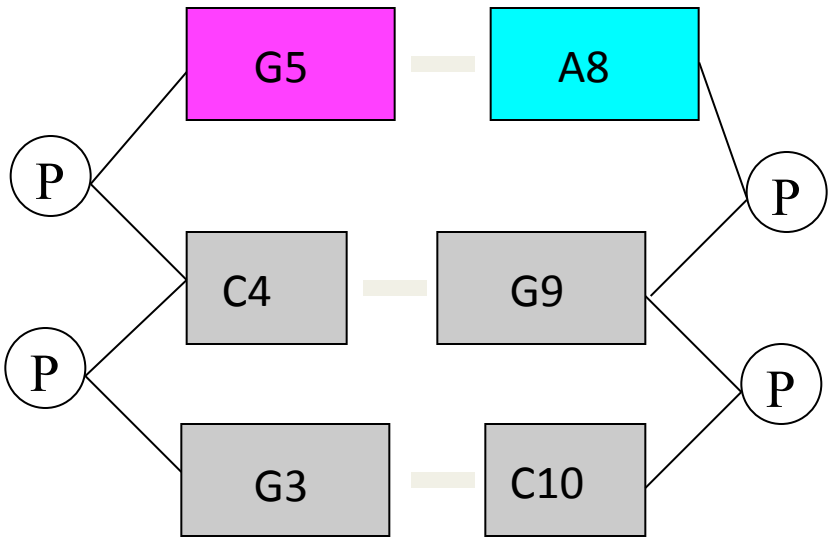


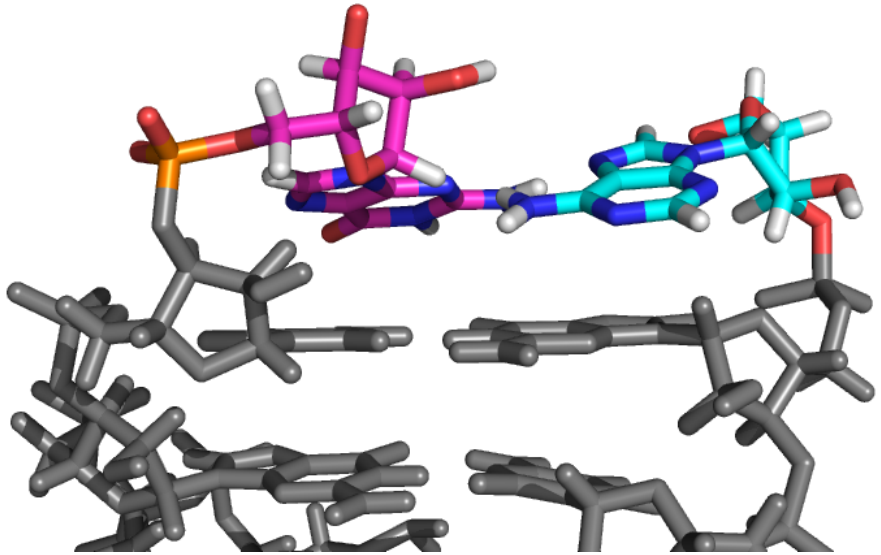
Step-by-step sampling



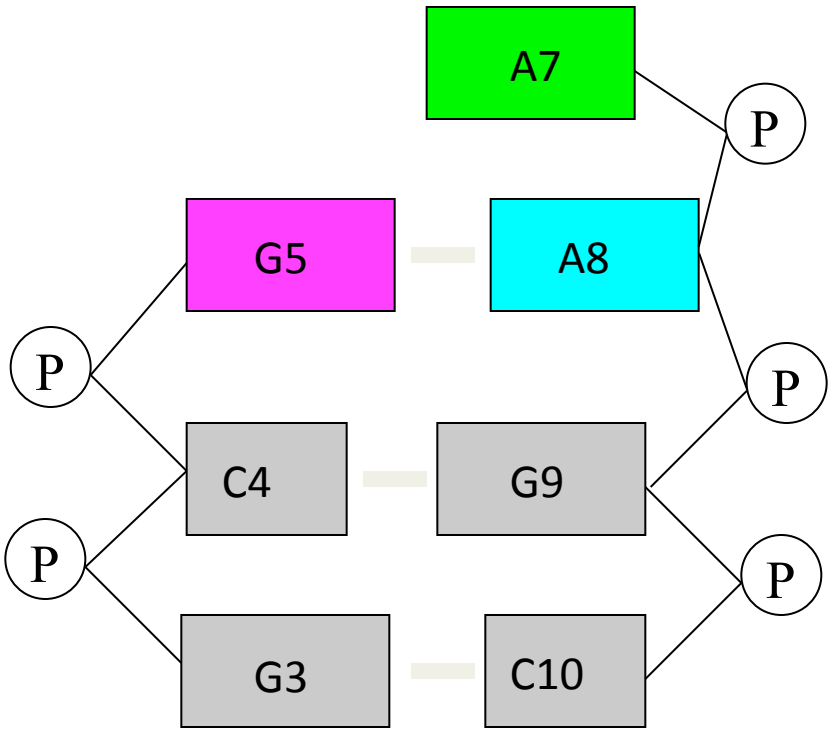


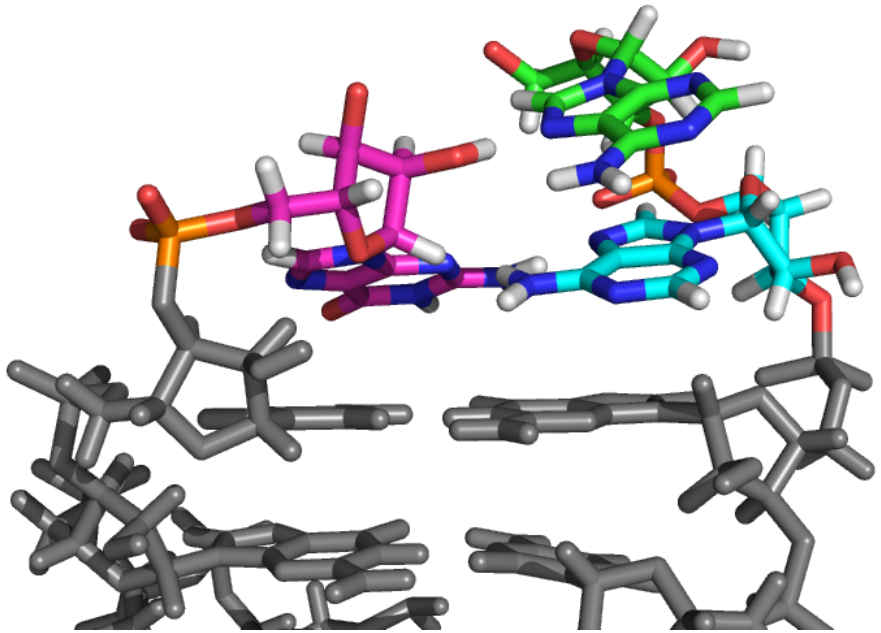
Step-by-step sampling



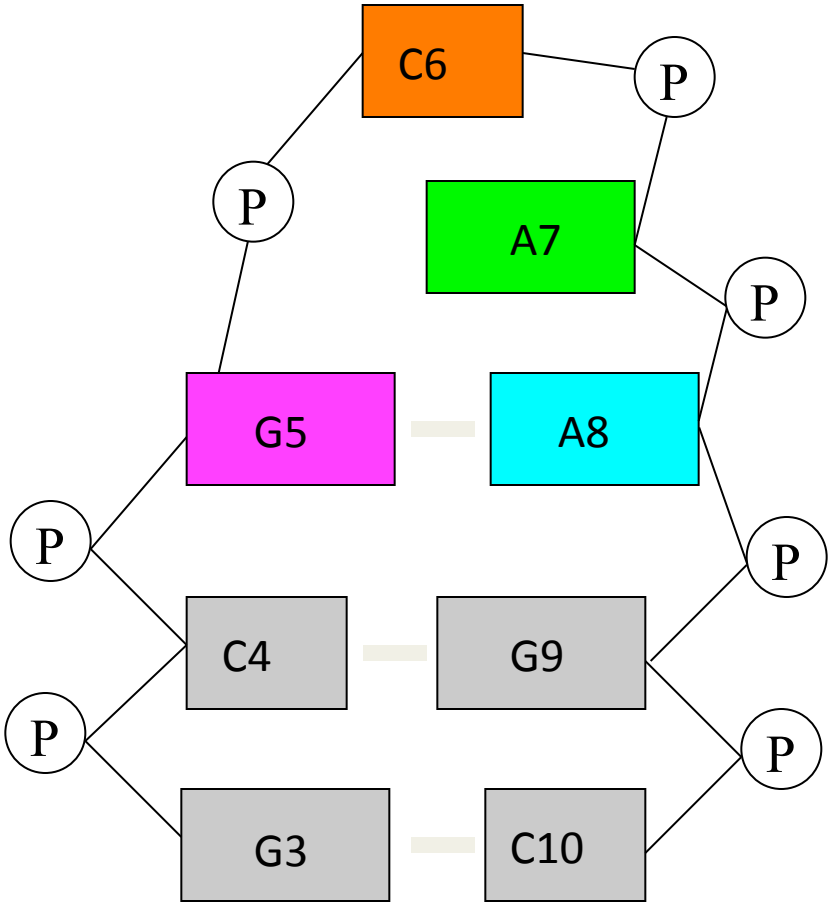


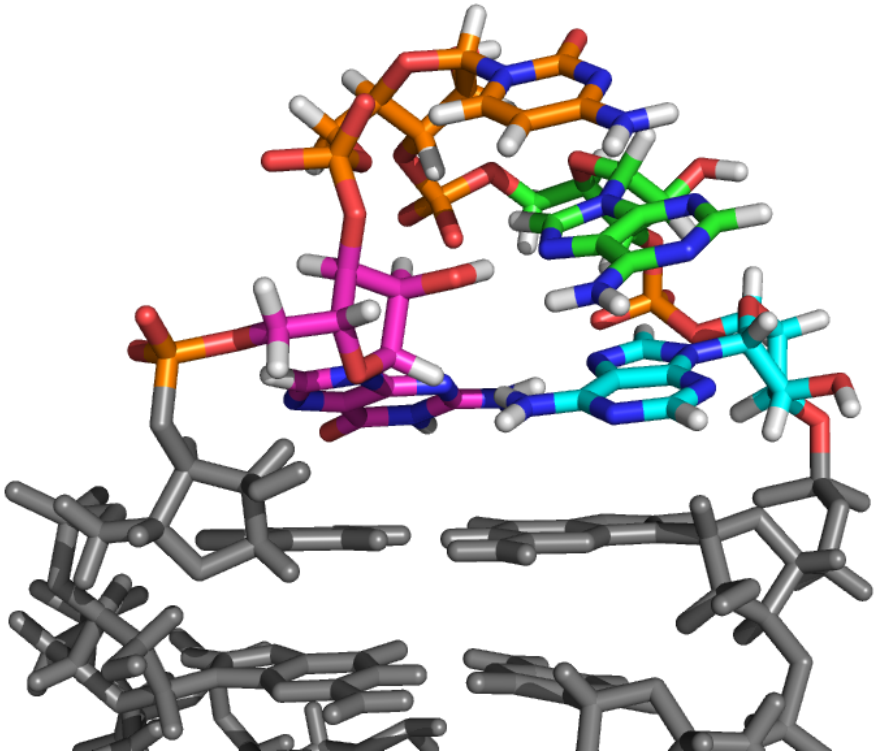
Step-by-step sampling



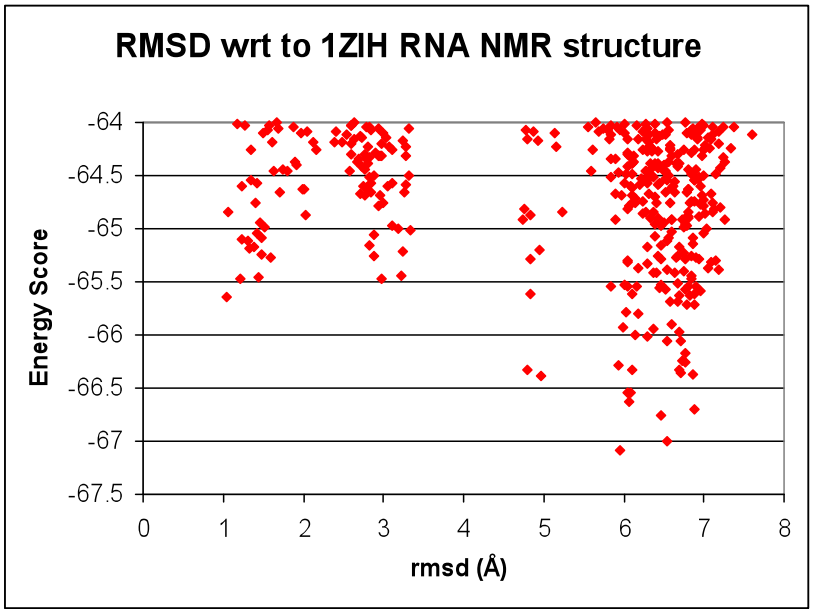


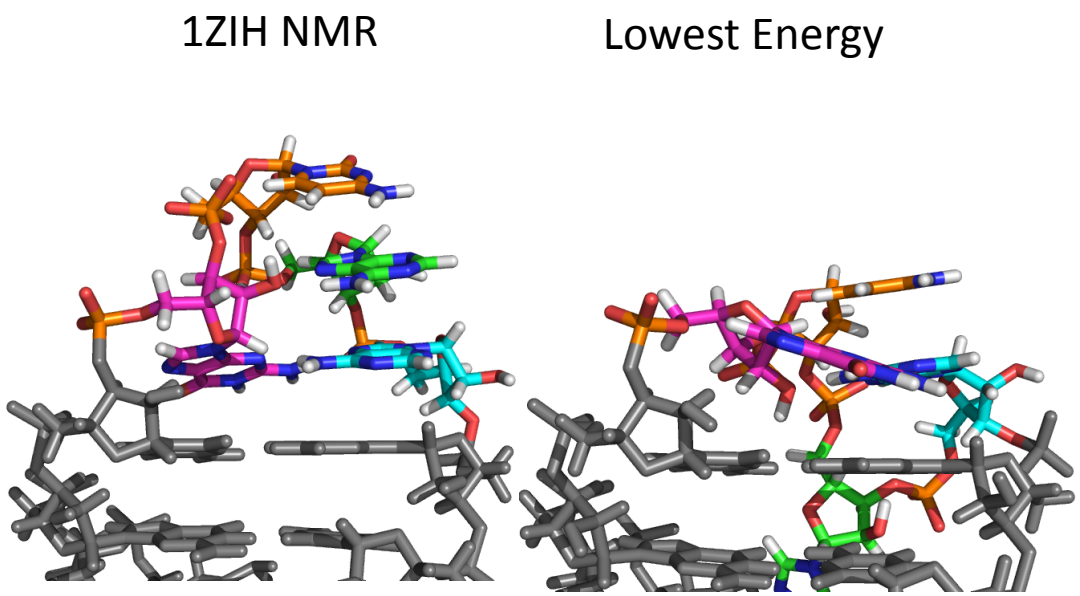
Step-by-step sampling





Step-by-step sampling



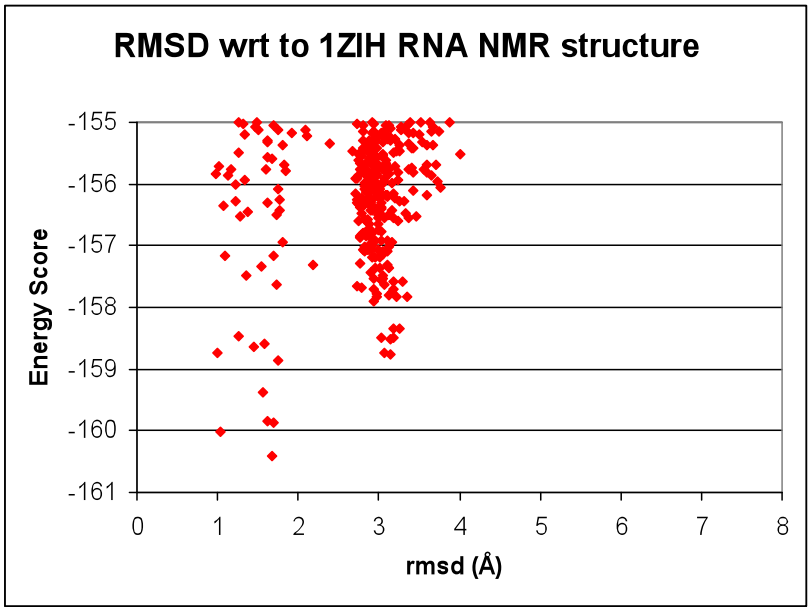


Aha – terms for:

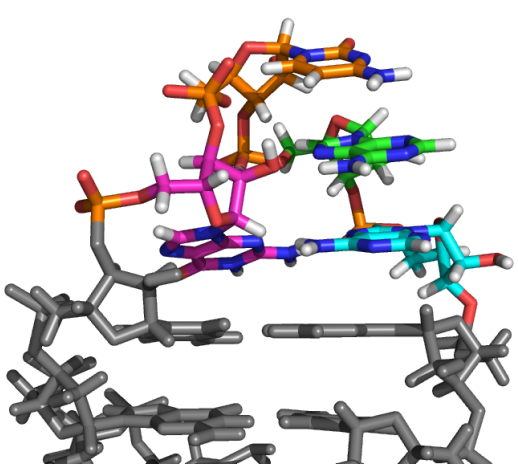
- base stacking
- RNA torsional potential

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

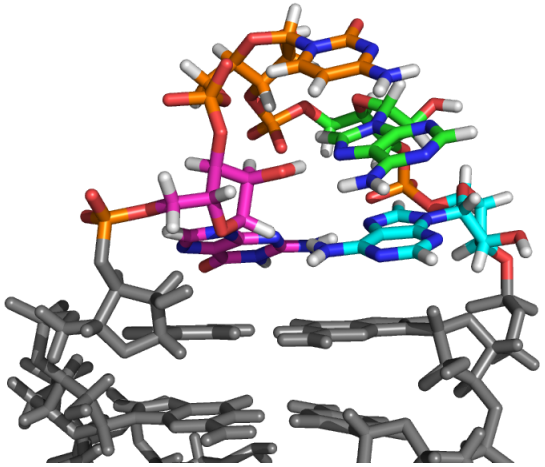
Step-by-step sampling



1ZIH NMR



Lowest Energy



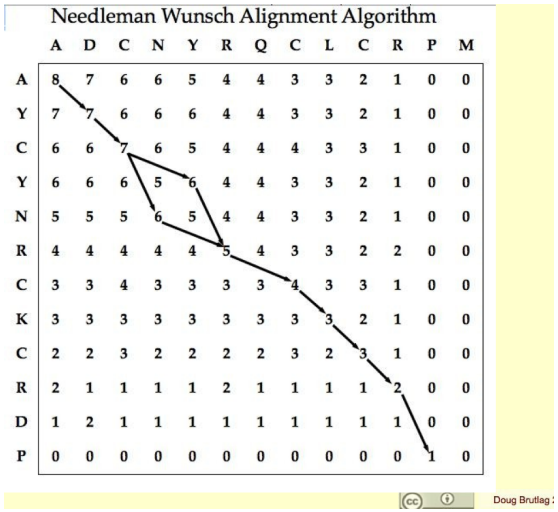
Wait, there's still a cheat!

There are other pathways (2^N total)

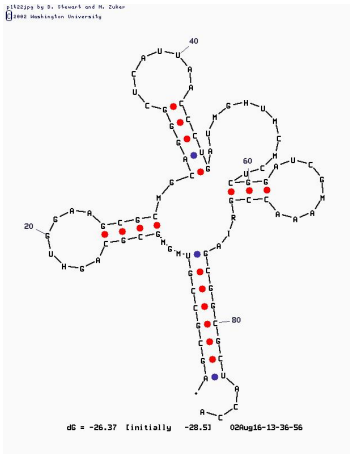
How to sample all paths?

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

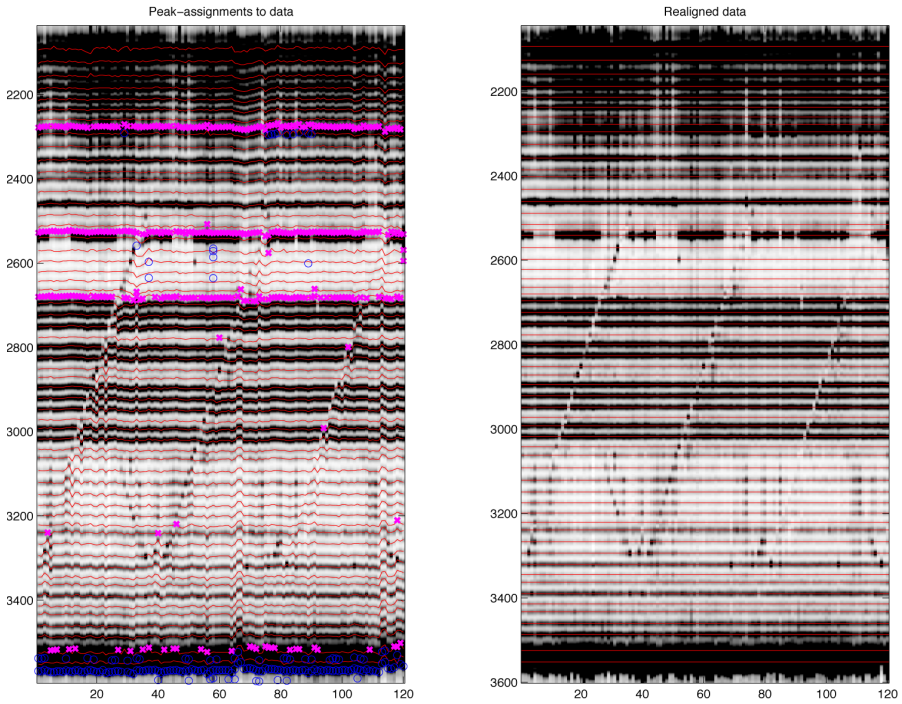
Sequence alignment



Nucleic acid 2° structure



Electrophoretic trace alignment

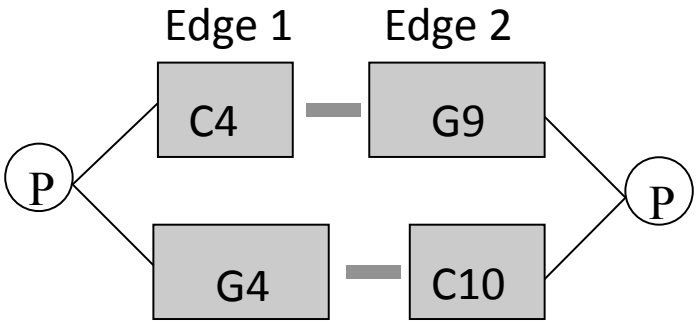


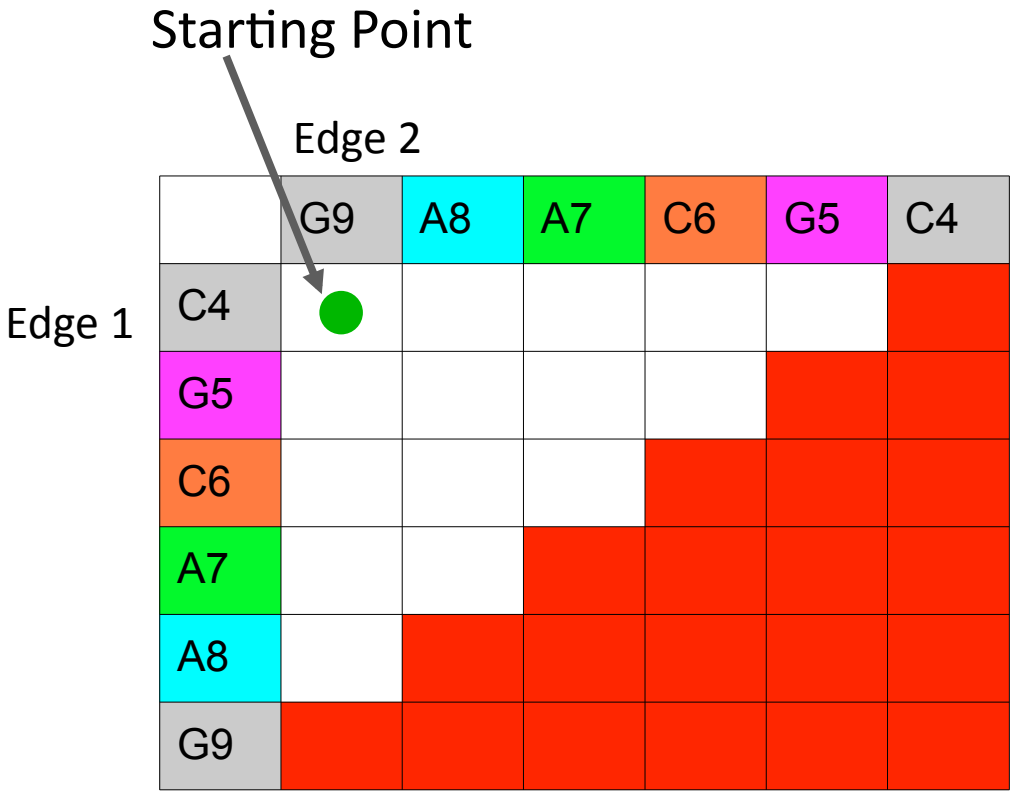
Ordering primers for PCR assembly for the least $$$.

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

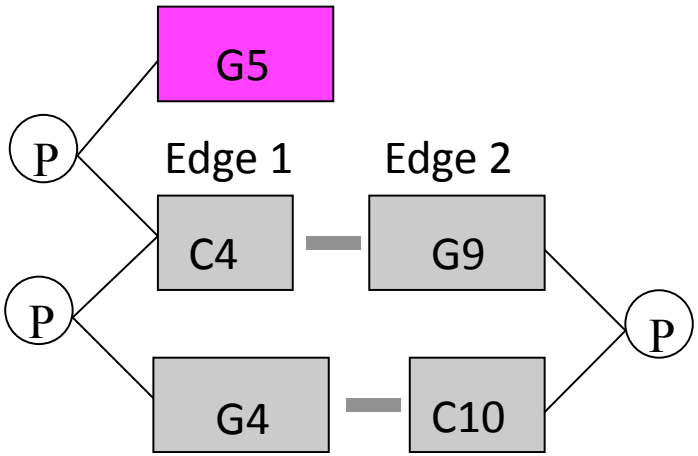
```
CAGGATTCAGTTG
||||||||||||| 60.5
GTCCTAAGTCAACAGATCTTCTGTTGATATGGATGC->5
            ||||||||||||||||||||||| 58.5
          6<-CTAGAAGACAACTATACCTACGTCAAGTTTTGGTTTGGTTTCTTTGTTGTTGTTGTTG
```

Rebuild Pathways: Tetraloop (1zih)

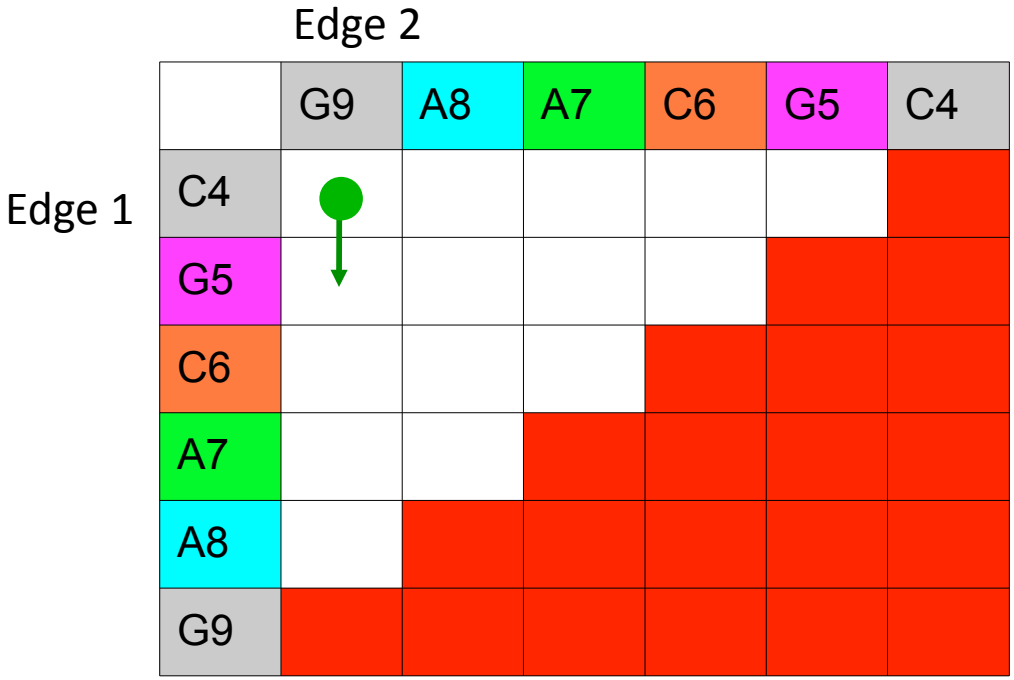




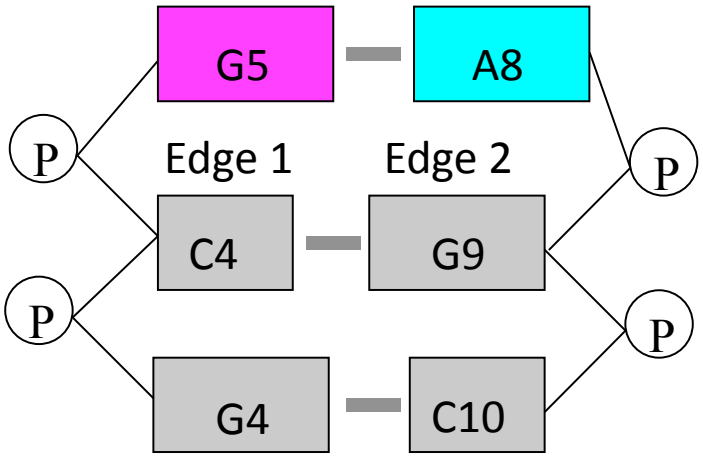
Rebuild Pathway: Tetraloop (1zih)



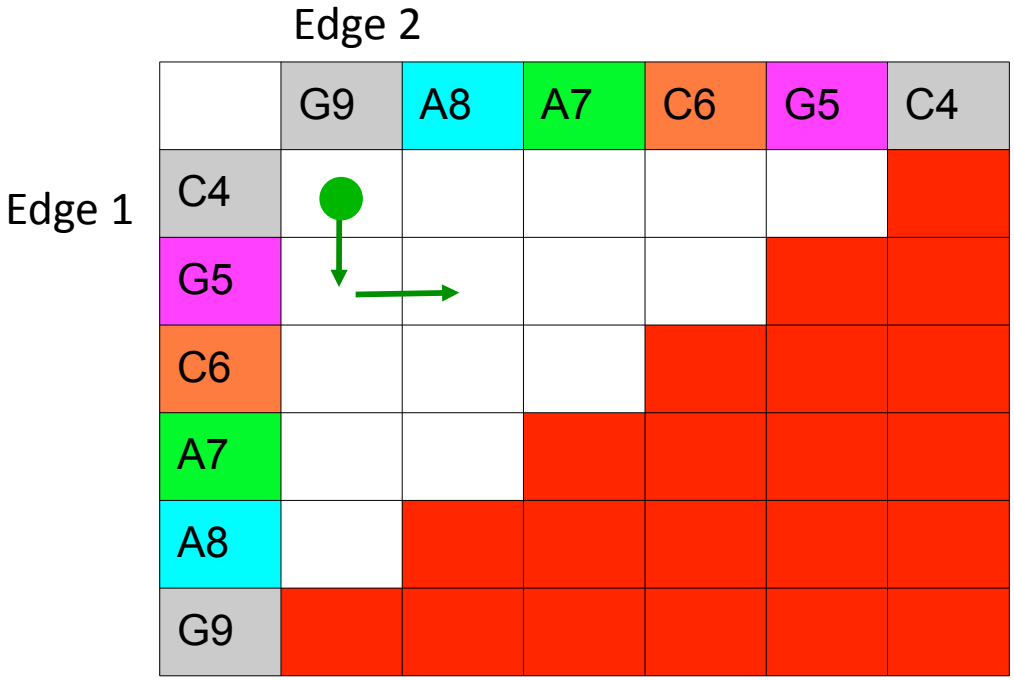
Build G5 from Edge 1



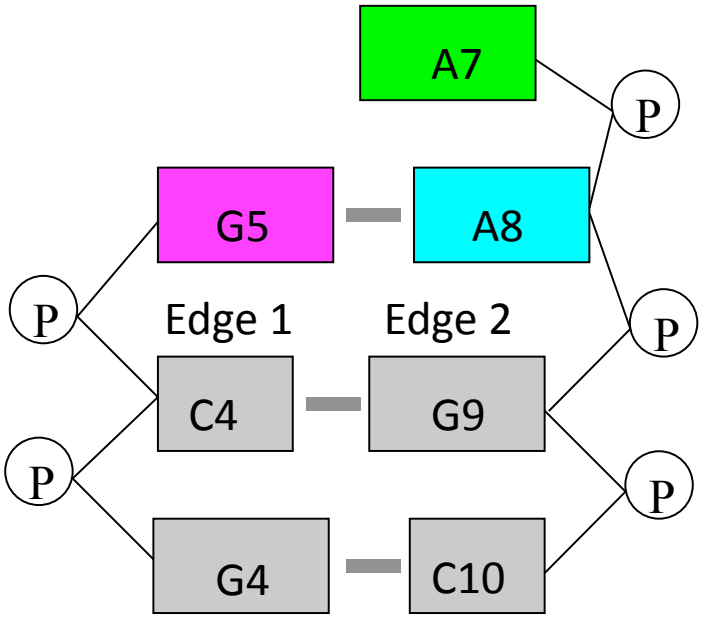
Rebuild Pathway: Tetraloop (1zih)



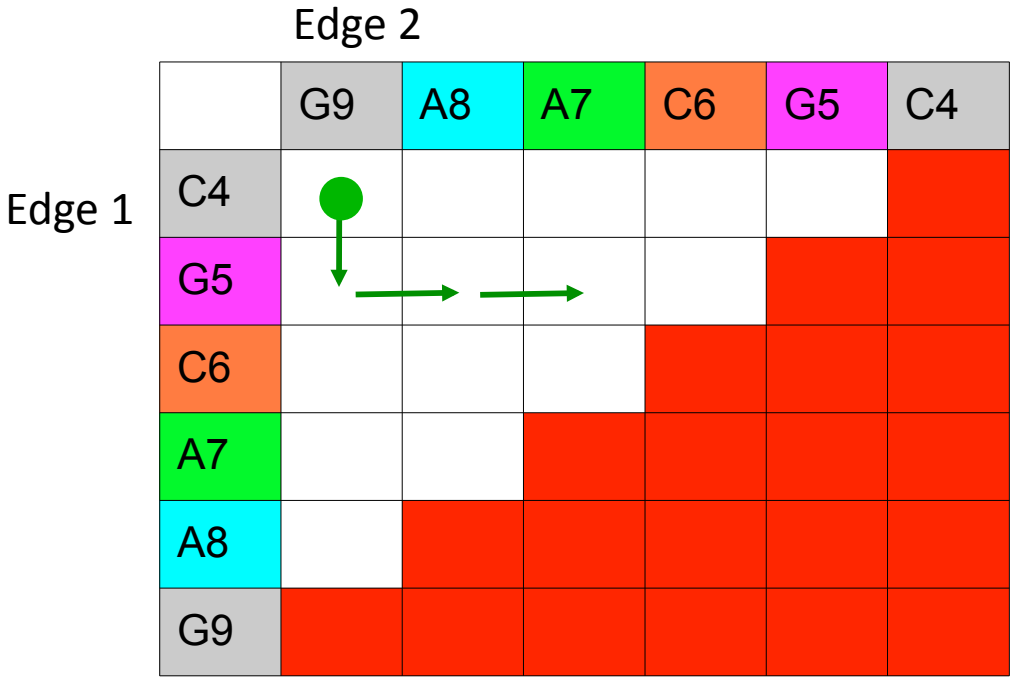
Build A8 from Edge 2



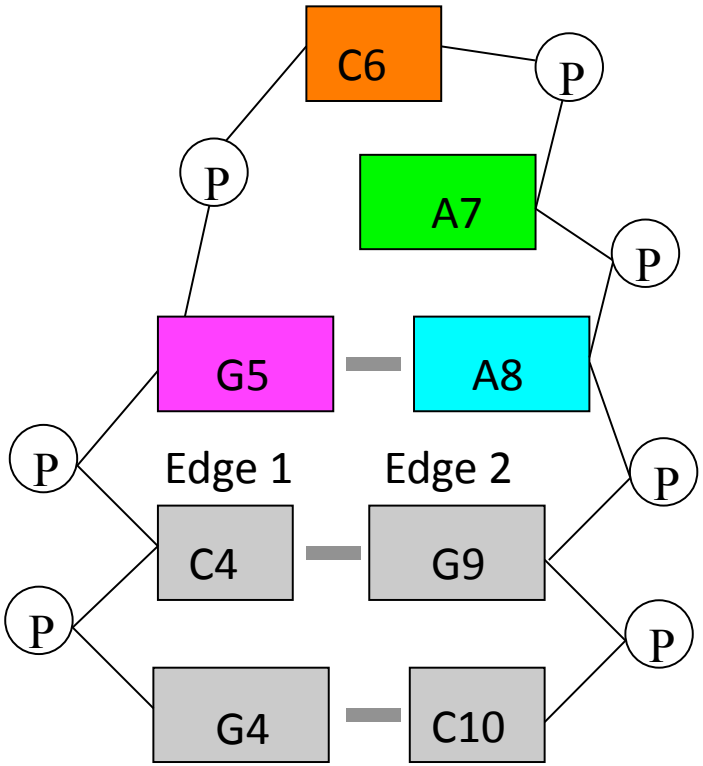
Rebuild Pathway: Tetraloop (1zih)



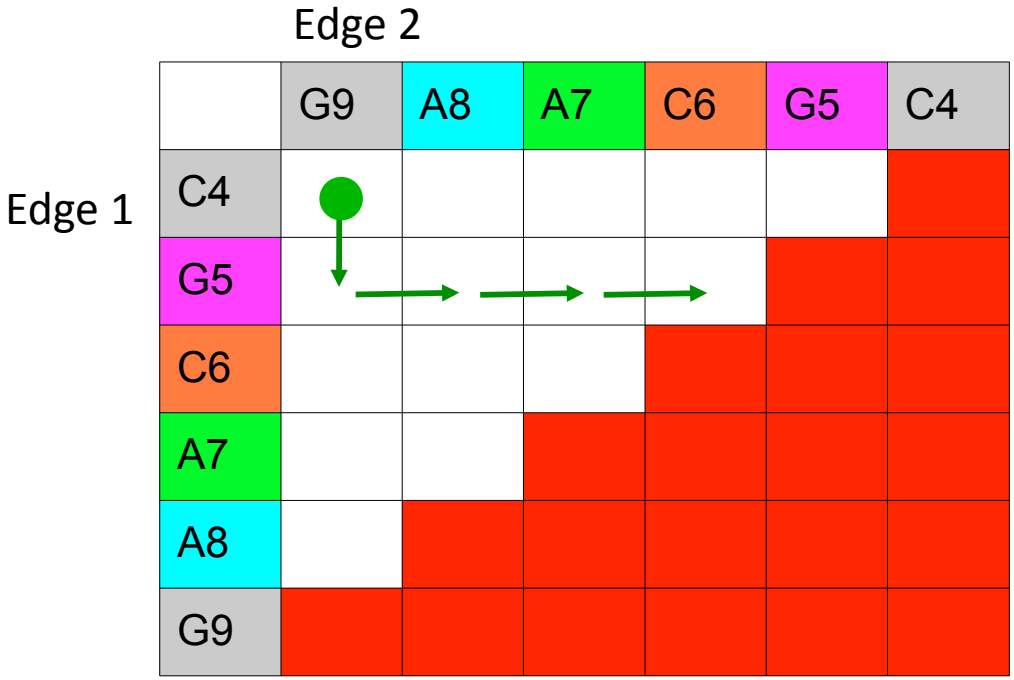
Build A7 from Edge 2



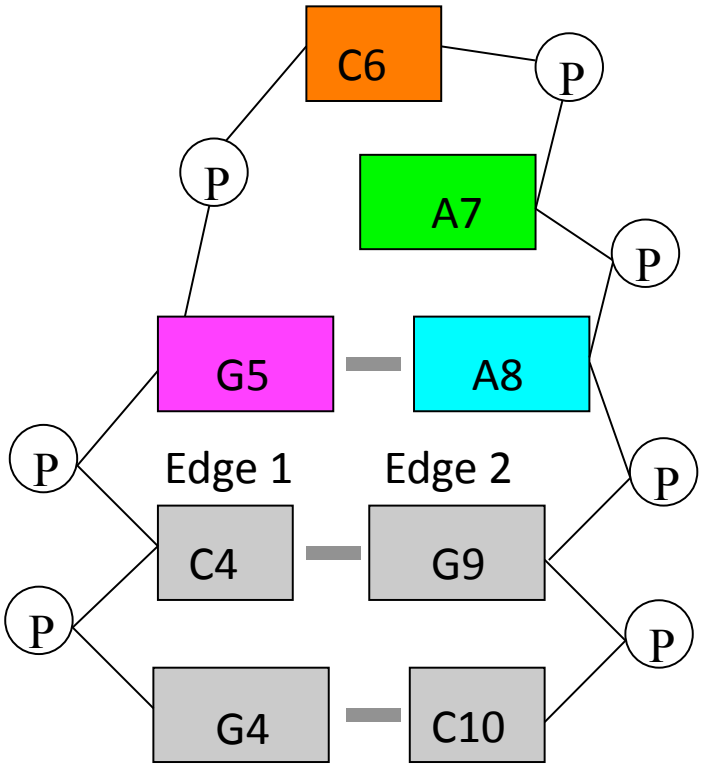
Rebuild Pathway: Tetraloop (1zih)

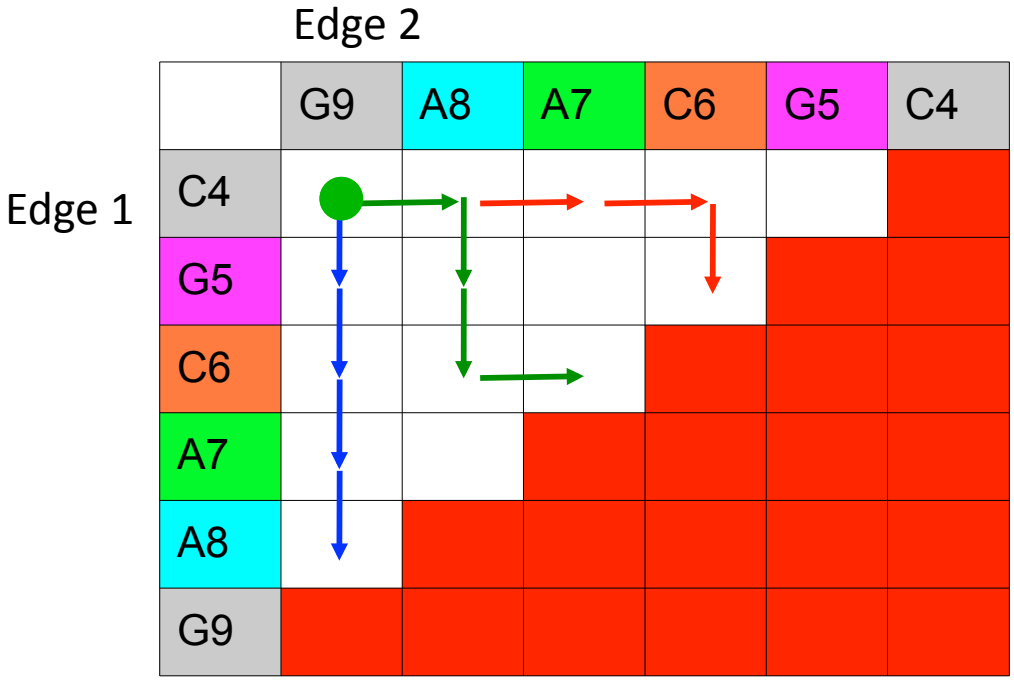


Build C6, close chain

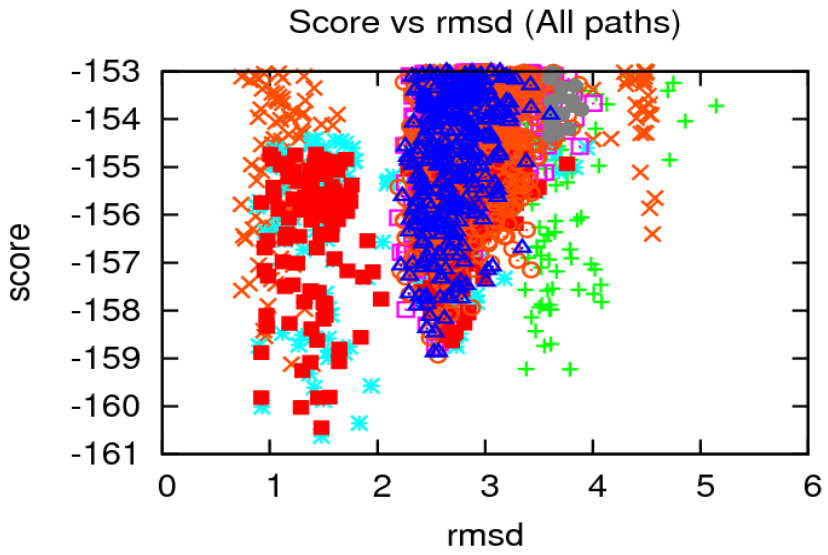


Can we rebuild using other pathways?





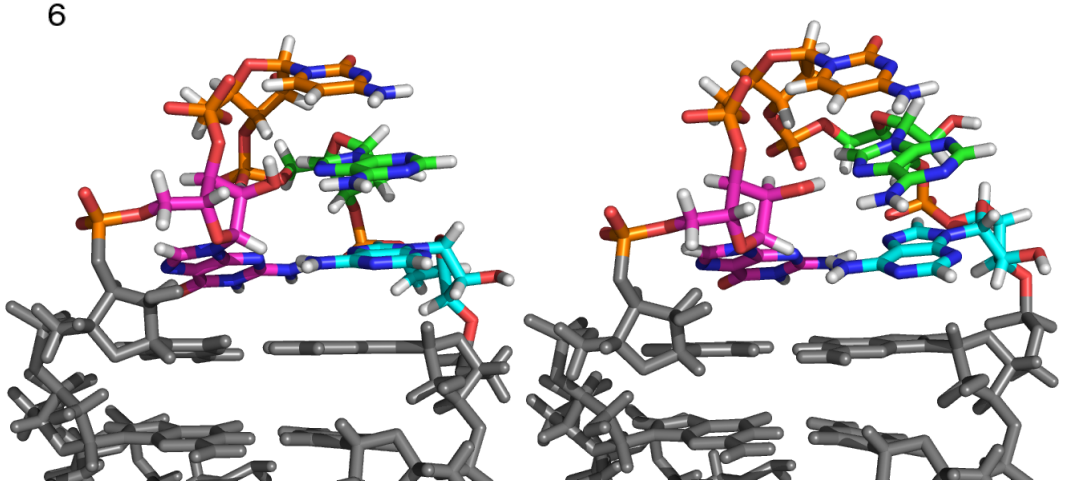
All pathways



Each point style represents a rebuild path

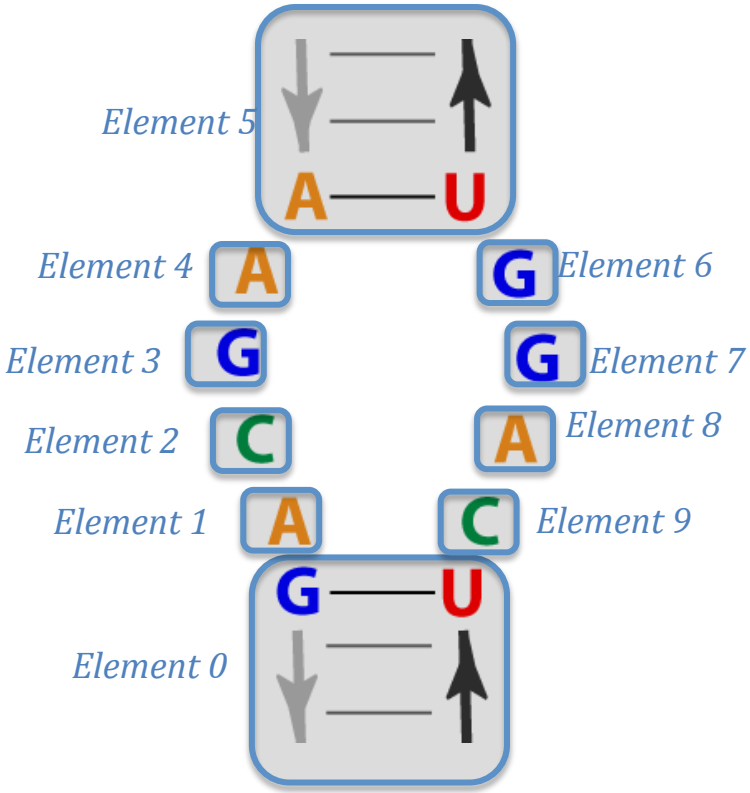
1ZIH NMR

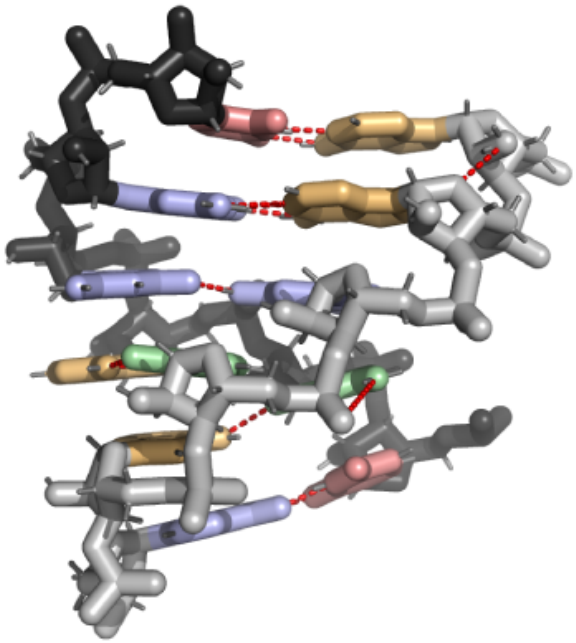
Lowest Energy



Best energy score decoy:
~1.5 Å from NMR structure

A more complex motif

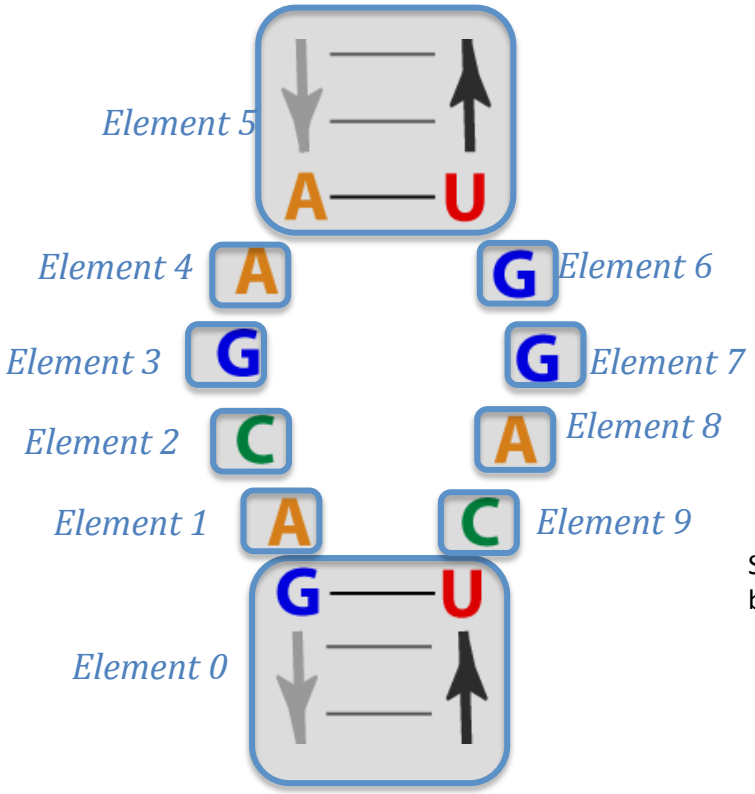


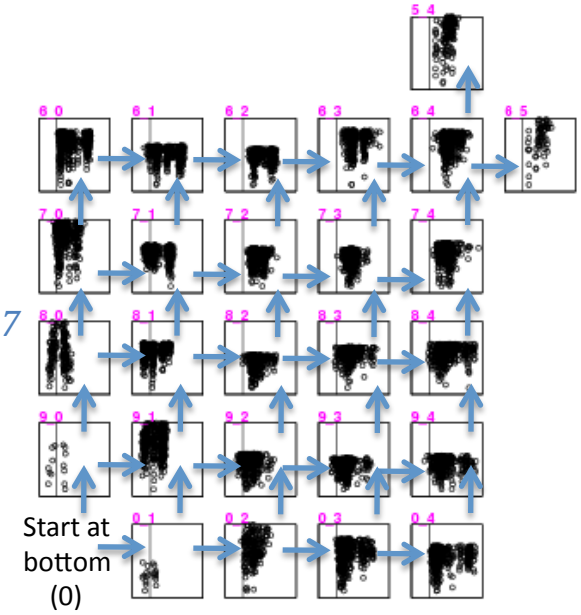


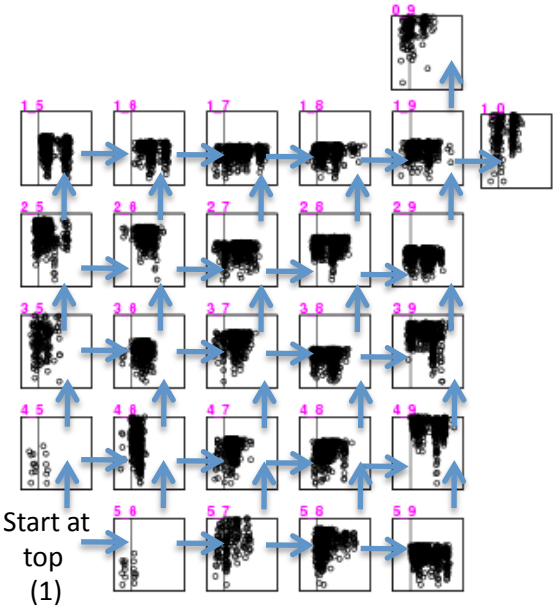
The most conserved domain of the signal recognition particle

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

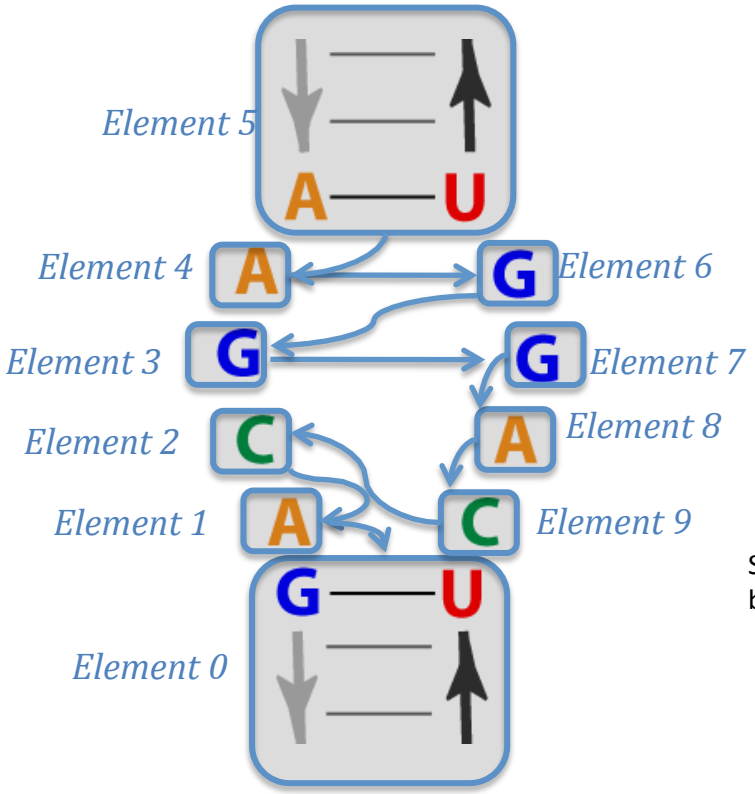
A more complex motif

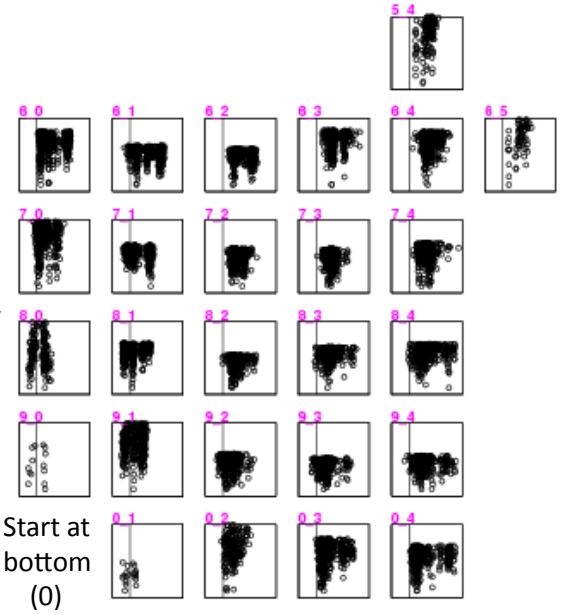


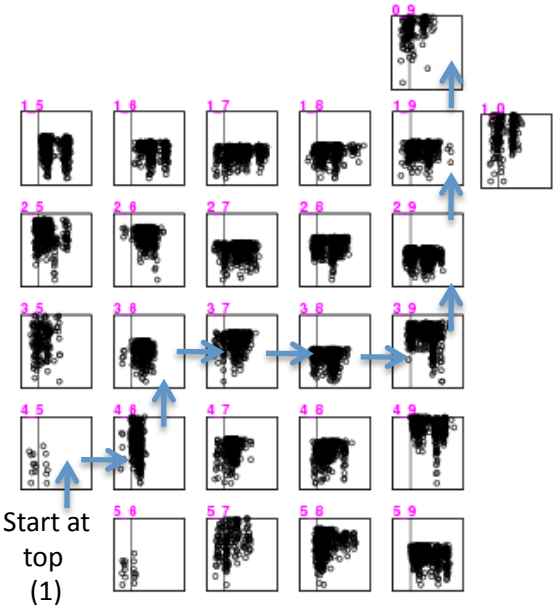




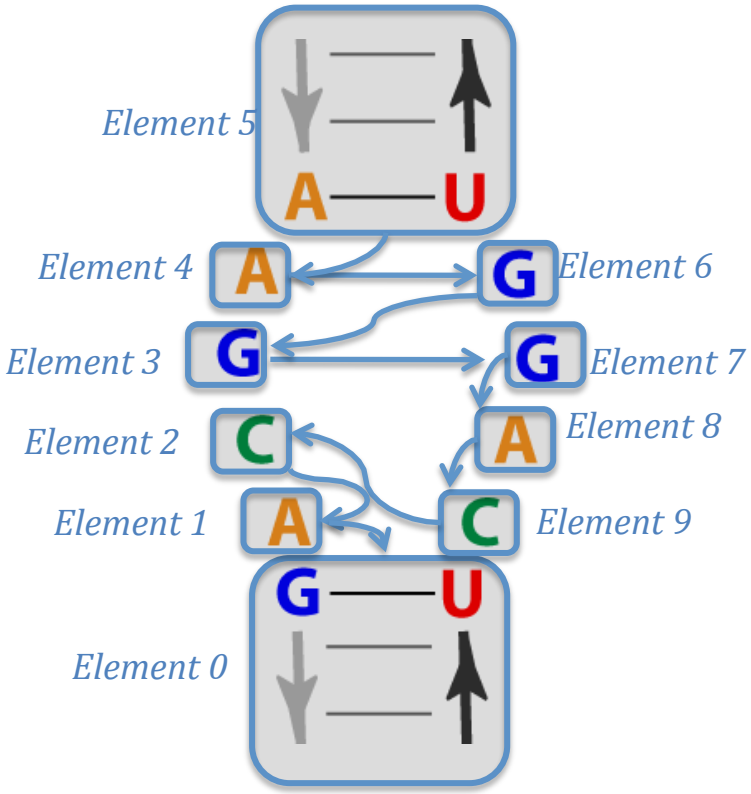
A more complex motif

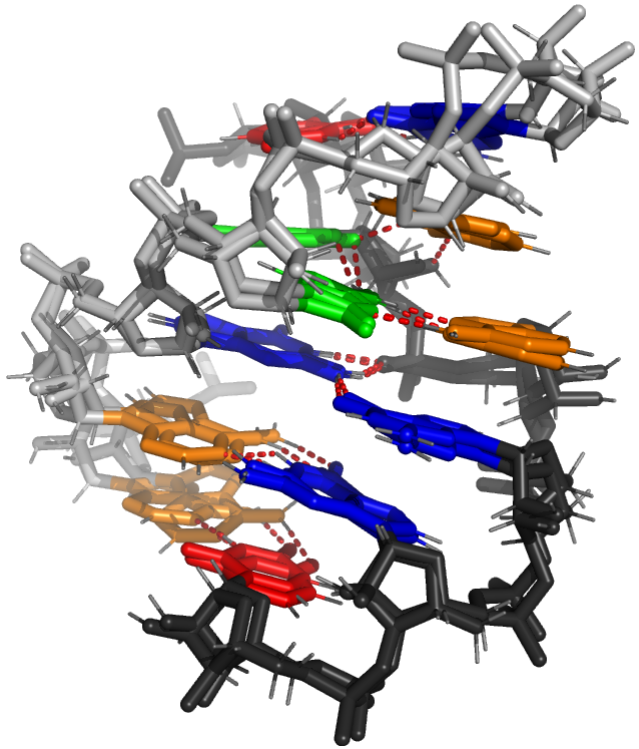




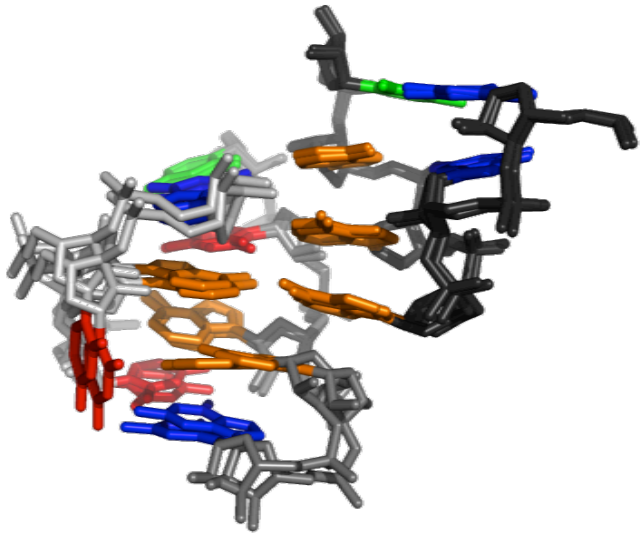


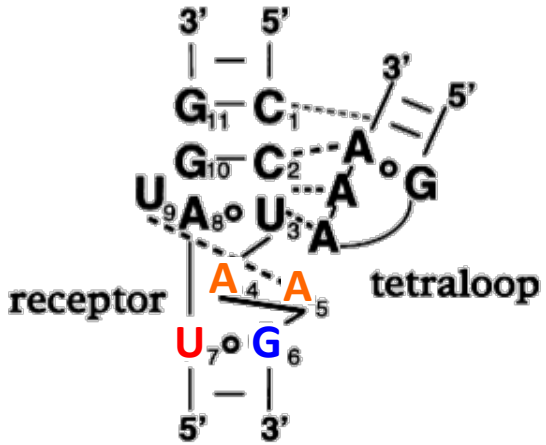
A more complex motif



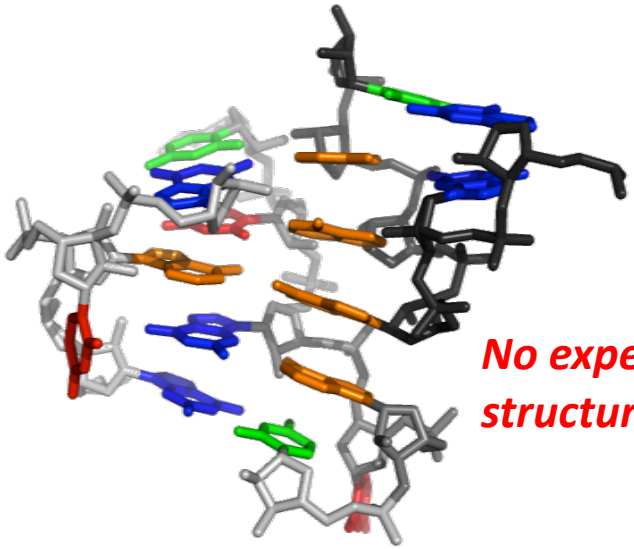


A baby step, but a *blind* one.

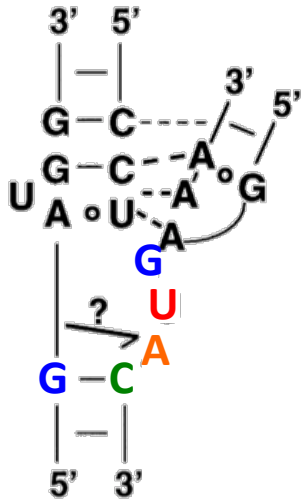




Just rebuilding the colored residues

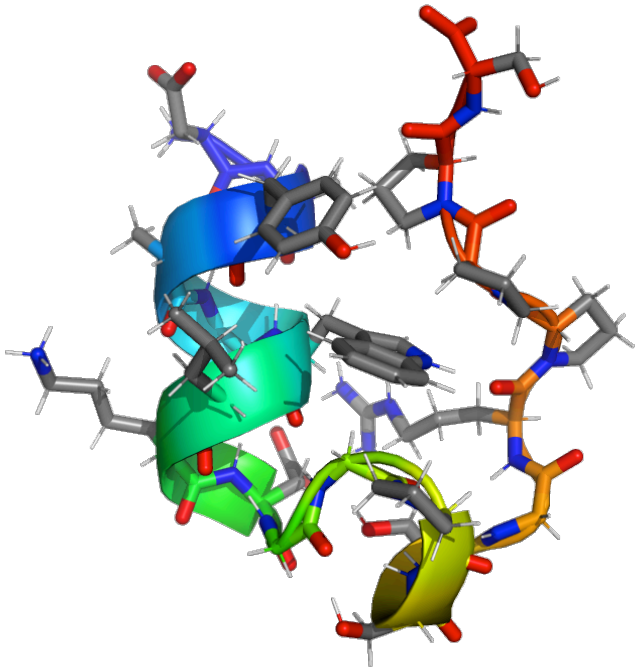


No experimental structure yet.

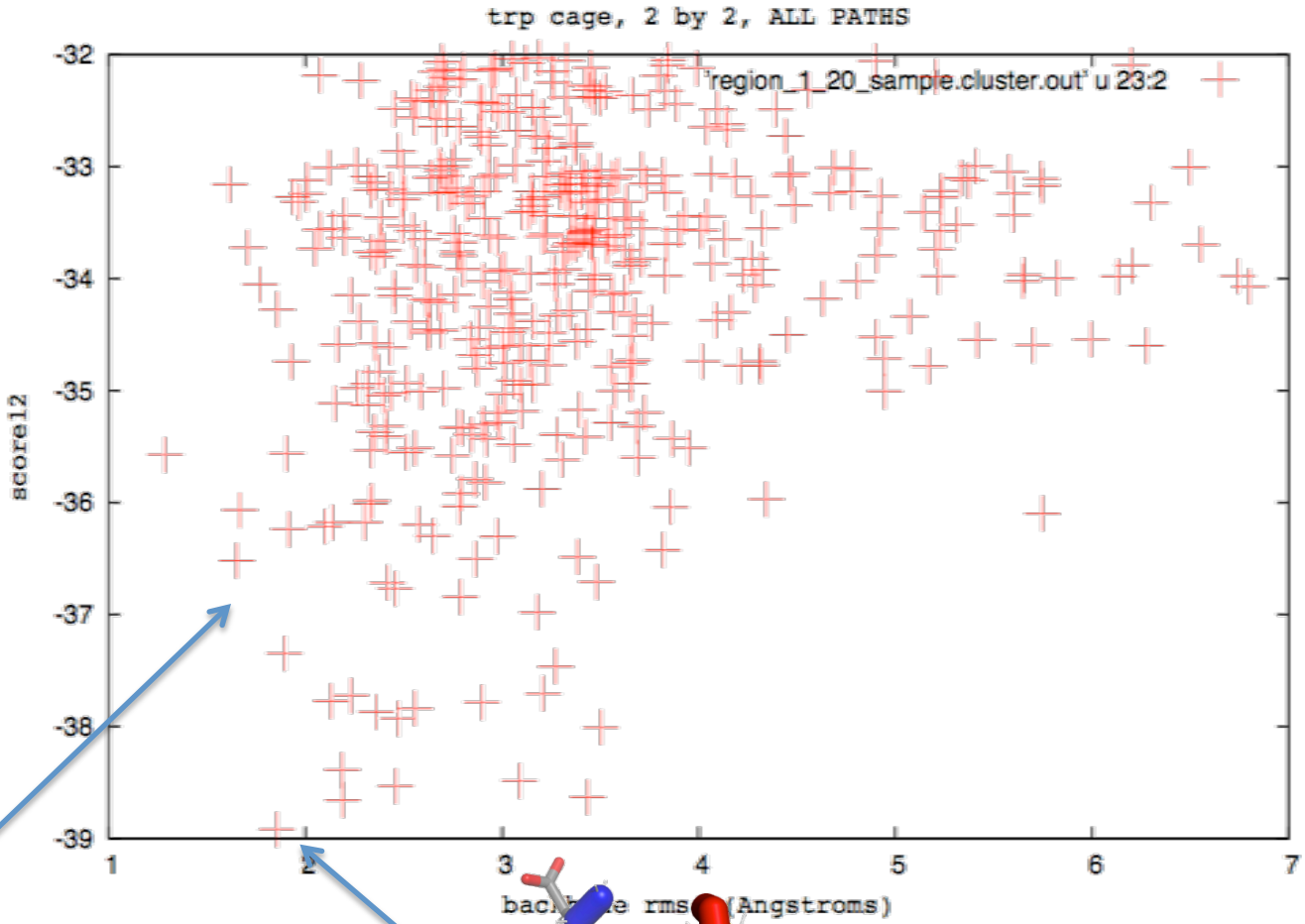


C7.2

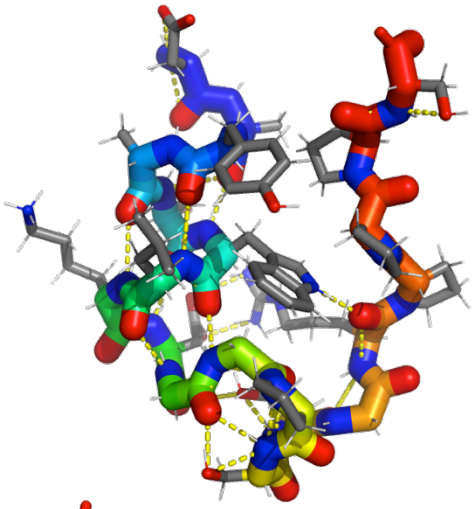
A stepwise ansatz for (mini) proteins

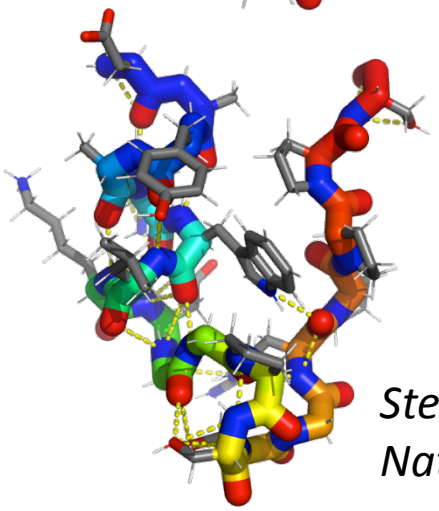


“Trp CAGE” – a 20 residue miniprotein

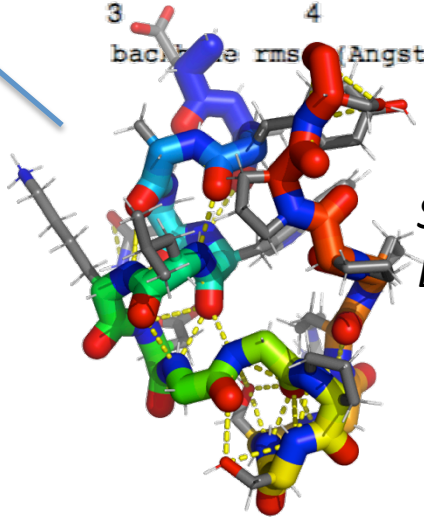


NMR model

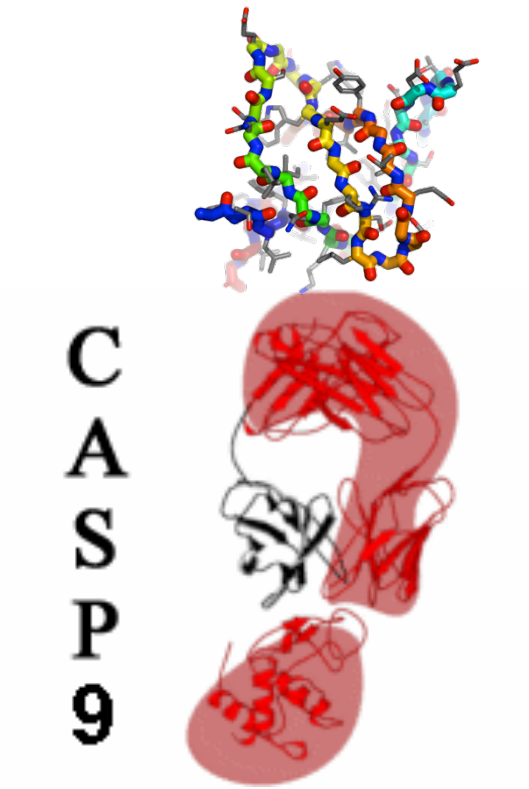




Stepwise assembly –
Native-like model



Stepwise assembly –
Best all-atom score



May-July 2010.

Extrait du Journal de Chimie Physique, 1968, **65** n° 1, p. 44.

Stepwise building: a new idea?

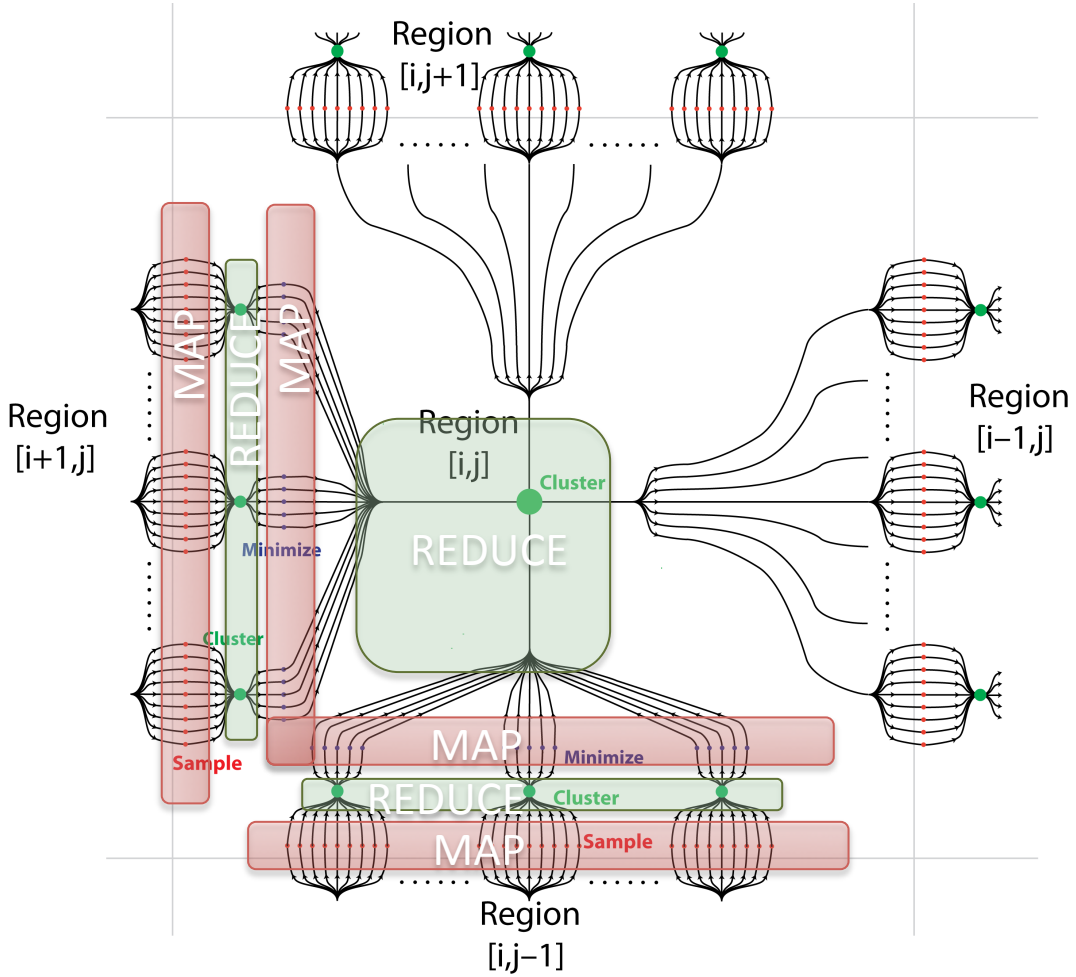
ARE THERE PATHWAYS FOR PROTEIN FOLDING ?

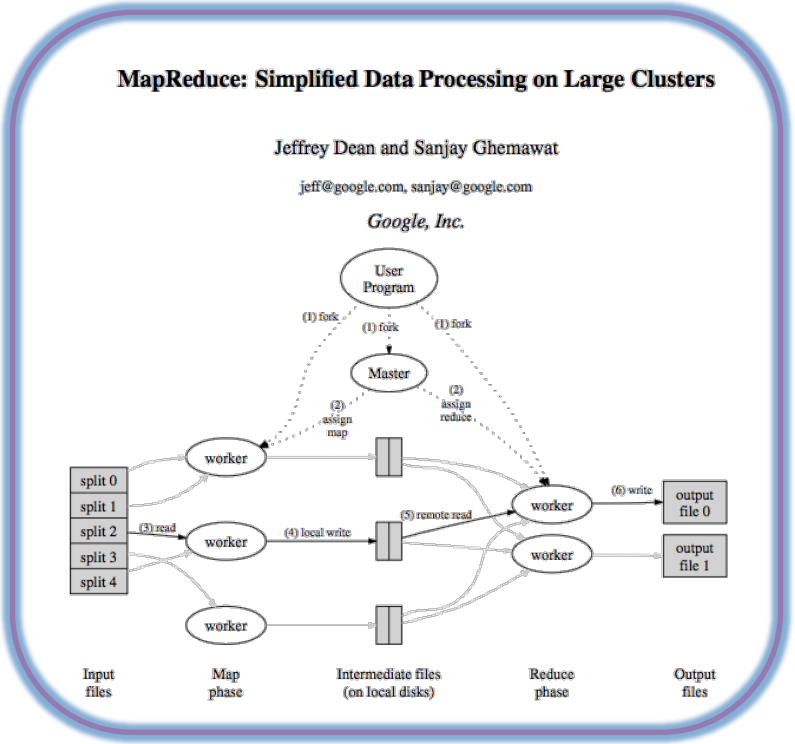
by CYRUS LEVINTHAL

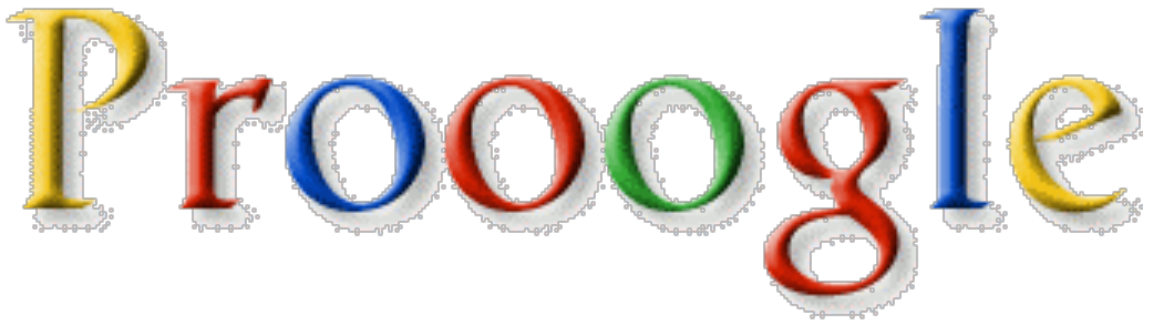
[*Massachusetts Institute of Technology, Department of Biology Cambridge, Massachusetts*.]

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.


Region
[i,j+1]
Region
[i+1,j]
Region
[i,j]
Region
[i−1,j]
Region
[i,j−1]
MAP
REDUCE
MAP
Cluster
REDUCE
Minimize
Cluster
Sample
MAP
Minimize
REDUCE
Cluster
MAP
Sample


MapReduce: Simplified Data Processing on Large Clusters
Jeffrey Dean and Sanjay Ghemawat
jeff@google.com, sanjay@google.com
Google, Inc.
User Program
(1) fork
(1) fork
(1) fork
Master
(2) assign map
(2) assign reduce
worker
split 0
split 1
split 2
split 3
split 4
(3) read
worker
(4) local write
worker
worker
(5) remote read
worker
(6) write
output file 0
output file 1
Input files
Map phase
Intermediate files (on local disks)
Reduce phase
Output files



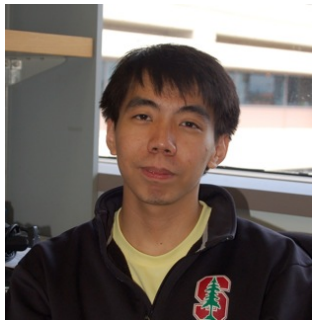
MFACNNFAGAPQRSTLNYIIALSDDWVQPPWP

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**

NATIONAL INSTITUTES OF HEALTH

NSF

Jane Coffin Childs Foundation

BURROUGHS WELLCOME FUND