Protein Folding: From Sequence to Structure

Karl F. Freed

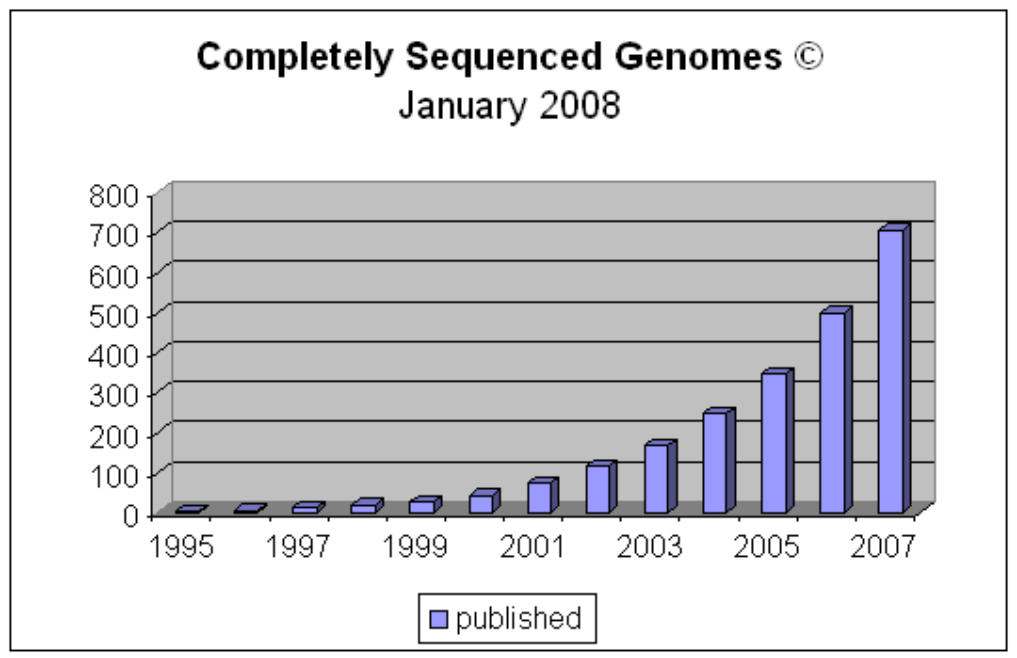
James Franck Institute
and Department of Chemistry
University of Chicago
UC/ANL Computation Institute

Argonne, Oct. 12, 2010

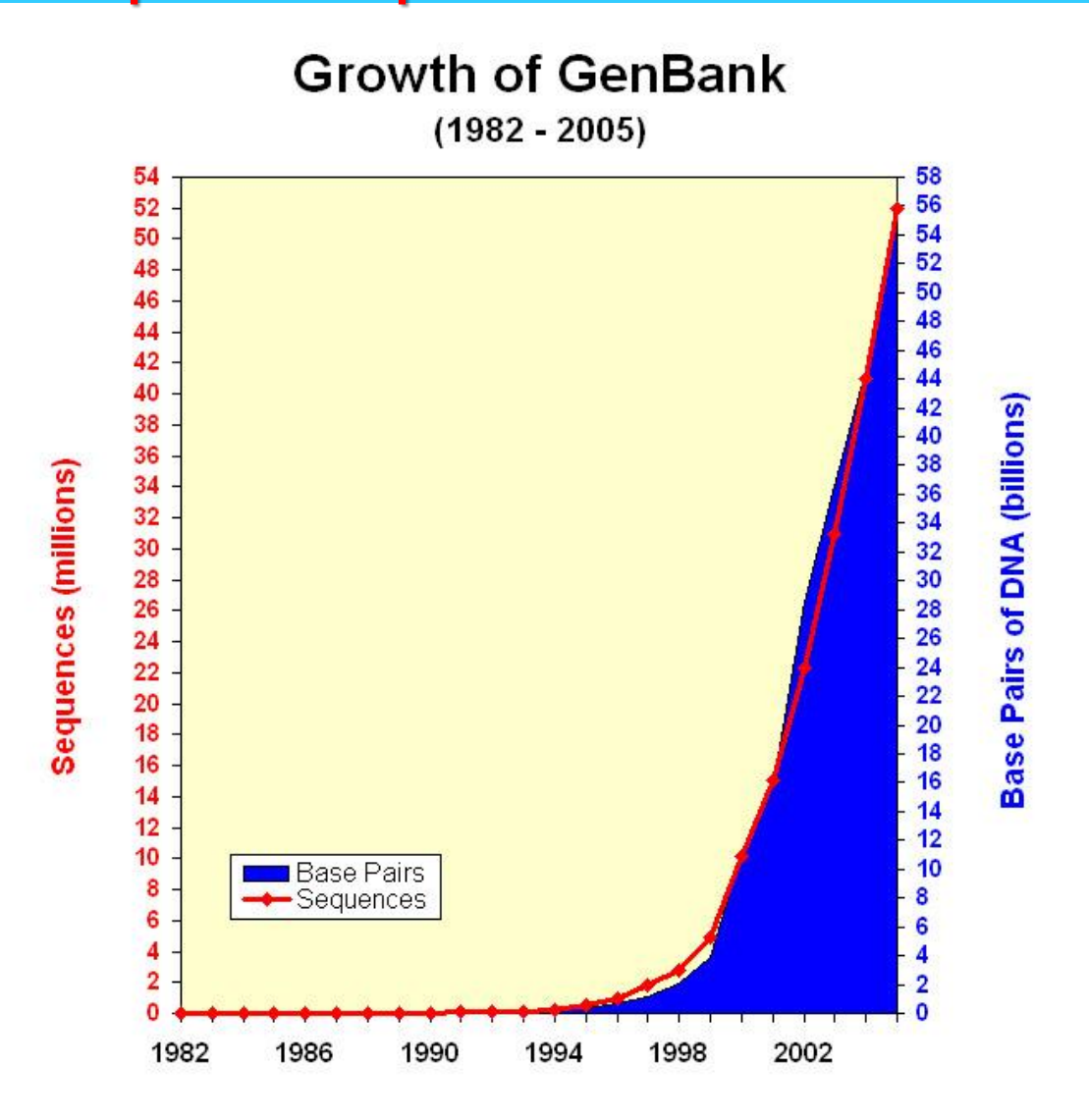
Human, Dog, Rat, Worm Genome Projects:

Obtain genes which code for protein sequences

Midwest Center for Structural Genomics, ANL, SWAXS, APS

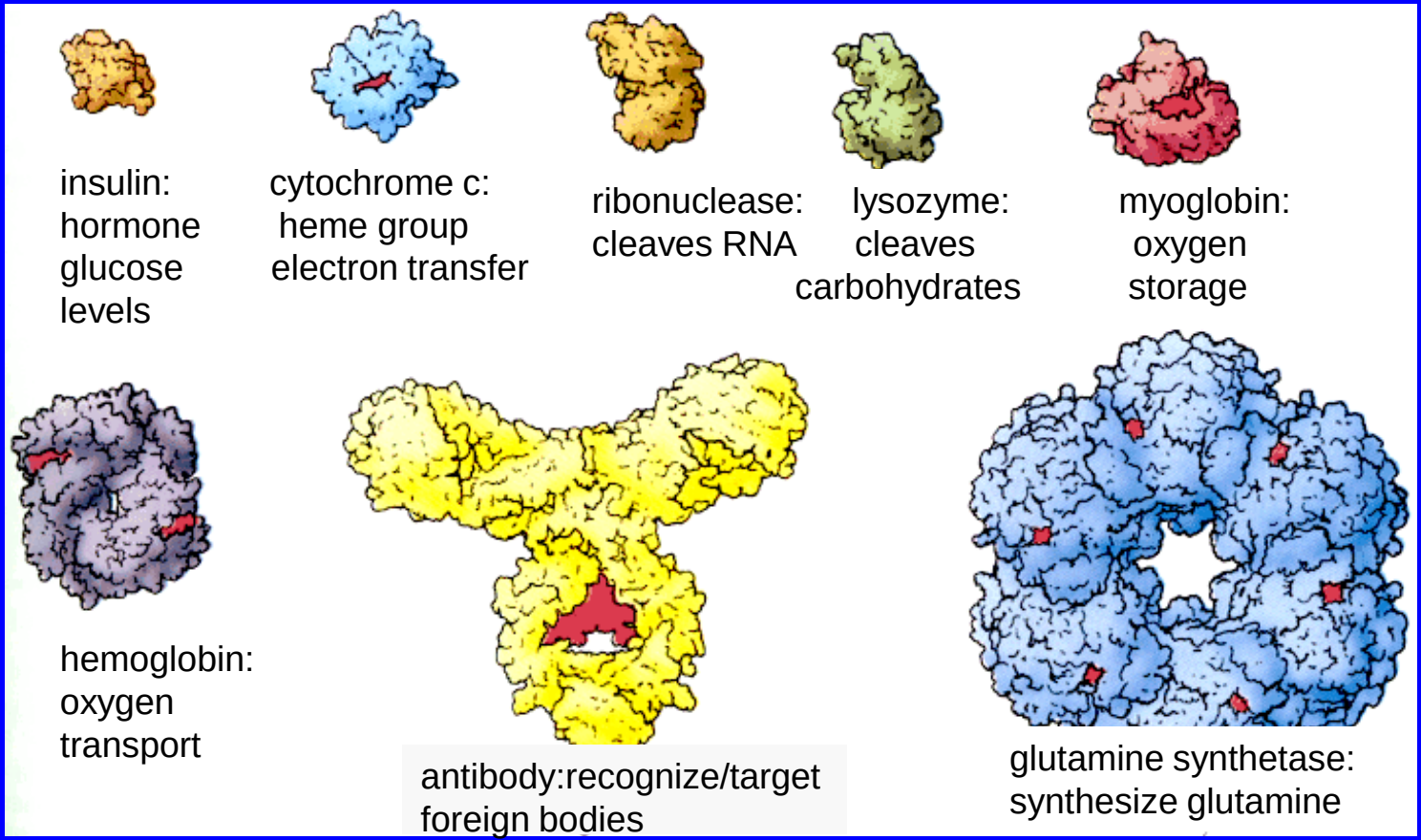


www.genomesonline.org/images/gold_s1.gif



www.ncbi.nlm.nih.gov/Genbank/genbankgrowth.jpg

Proteins: The primary functional biomolecules



Function follows from structure

Our goal: determine folded structures from sequence

Protein Folding Problem

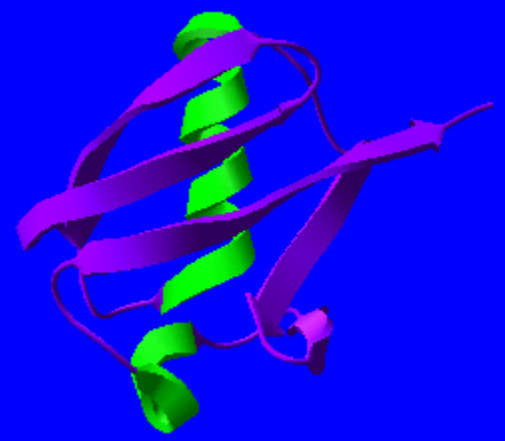
SEQUENCE

"Unfolded State"
Many conformations

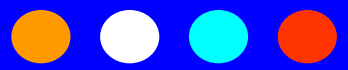
Folding

0.00001 – 10+ sec

3D STRUCTURE



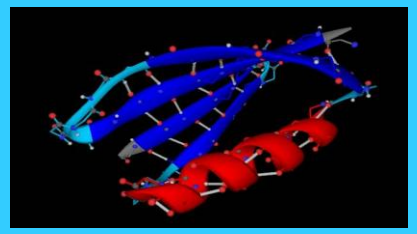
"Native State"
Responsible for function



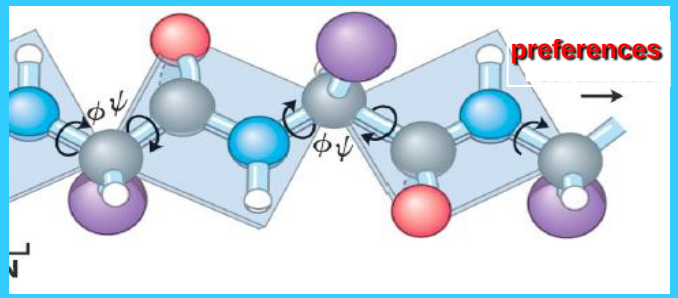
Amino acids - polar, neutral, charged, or hydrophobic

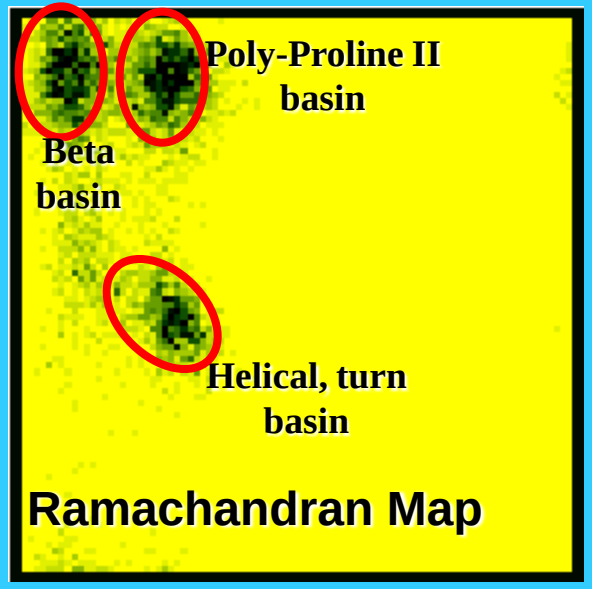
Vast Conformational Search: Levinthal Paradox

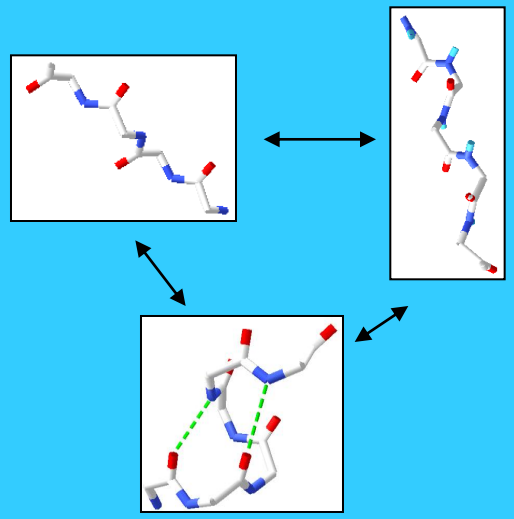
How does a protein find the time to fold?



Polypeptide backbone is flexible, adopting specific conformations







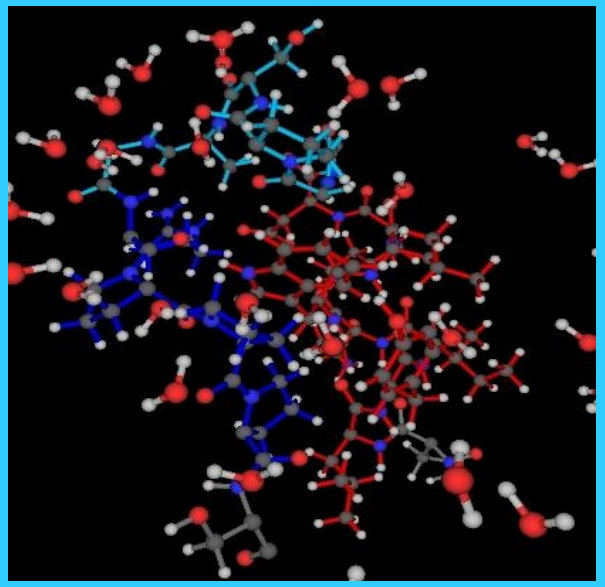
Each ϕ,ψ pair: 3 conformations ⇒ ~3^n ~$10^{0.5n}$ conformations/amino acid

e.g., 10^{50} states for a 100 a.a. protein

Random, unbiased search would take 10^{35} s even with fsec sampling

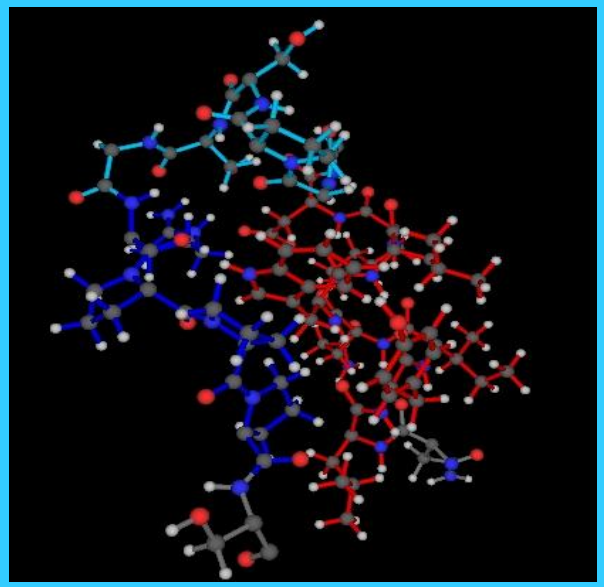
And we have to search too!

What info is needed to fold proteins? all-atom ⇒ dihedral angles



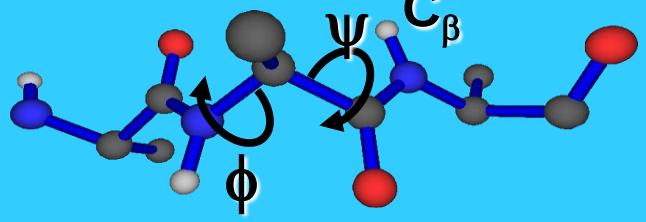
All-atom protein + solvent
Simulation would take decades

Step 1: Remove solvent –
(hydrophobic, screening, etc terms)

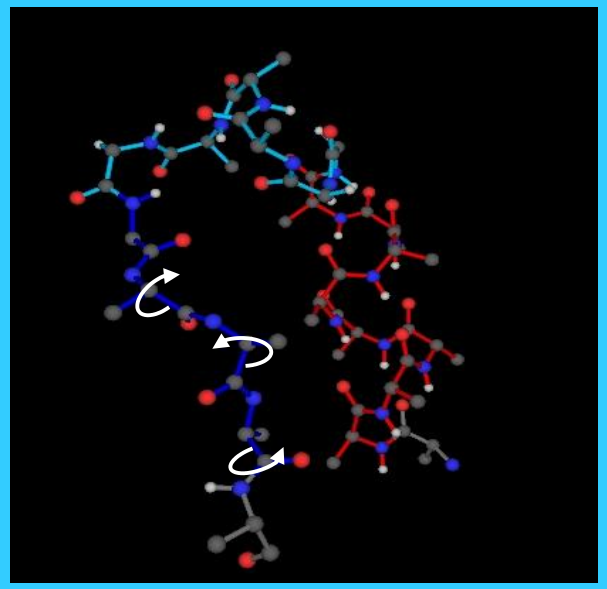


All-atom protein

Step 2: Replace side chains by C_β atoms.

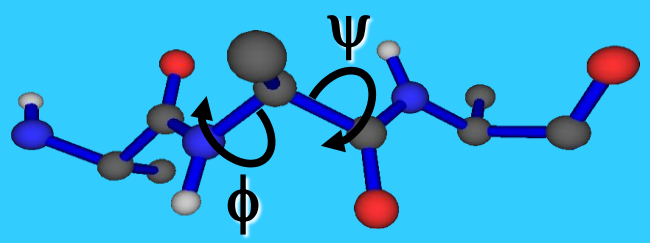


Only main chain heavy atoms and C_β atoms.
Backbone ϕ,ψ angles are the only coordinates.
2N total for N aa's (need ~10x3N for x,y,z of heavy atoms)



1) Proteins bend in certain ways

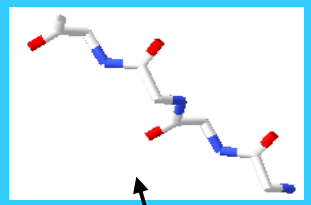
Sampling in dihedral space: ϕ-ψ angles and Ramachandran BASINS



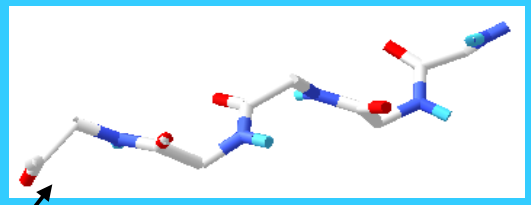
ϕ & ψ show very strong preferences for certain regions of the ϕ-ψ map, called **Ramachandran basins**.

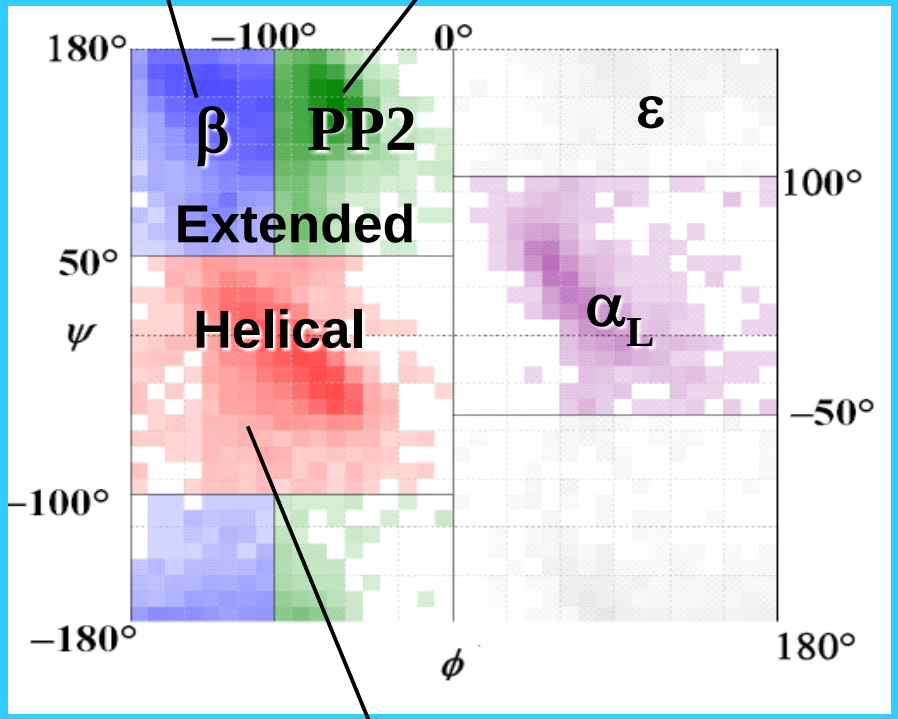
(due to steric & electrostatic interactions)

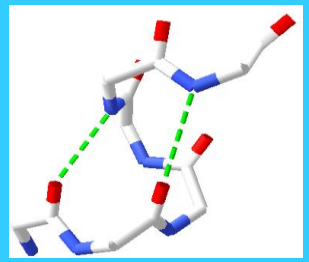
β-basin



PPII-basin



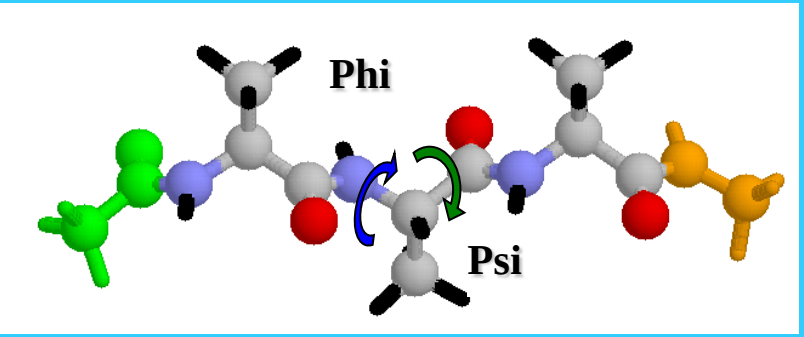




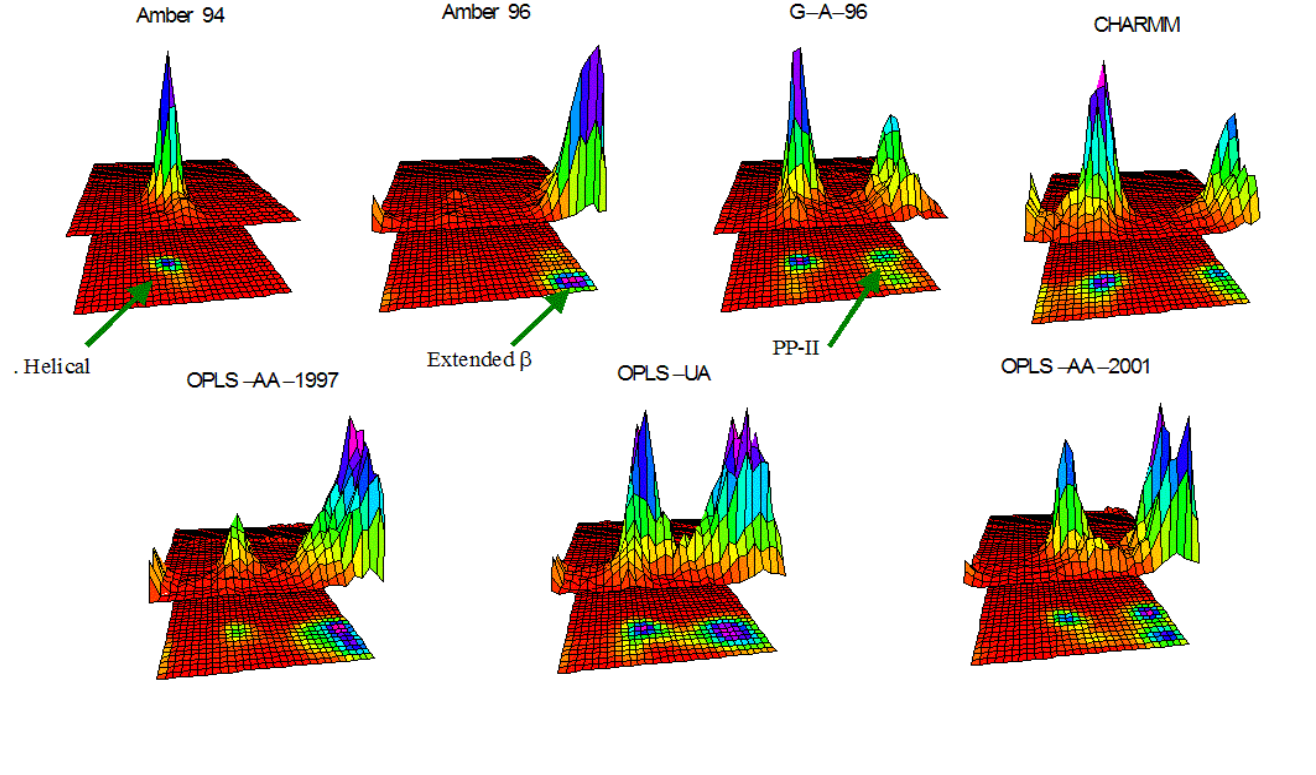
α-basin

Where to get this information?

From computer simulations?



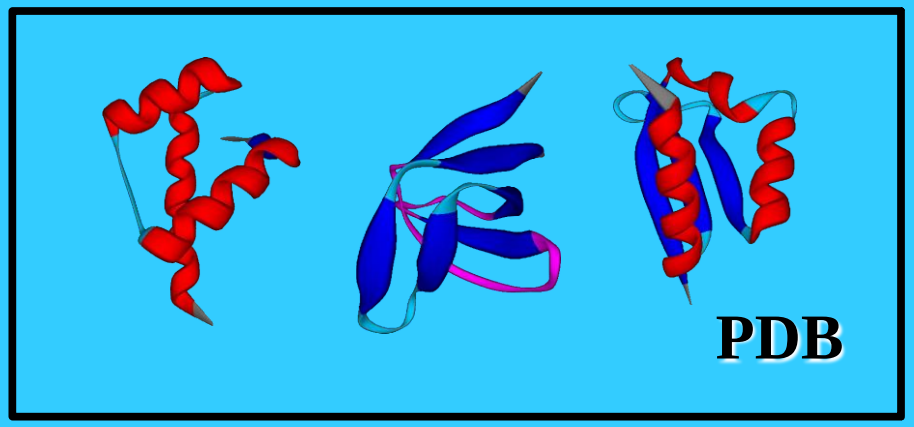
But, force fields can vary widely



Zaman et al.
JMB 2003

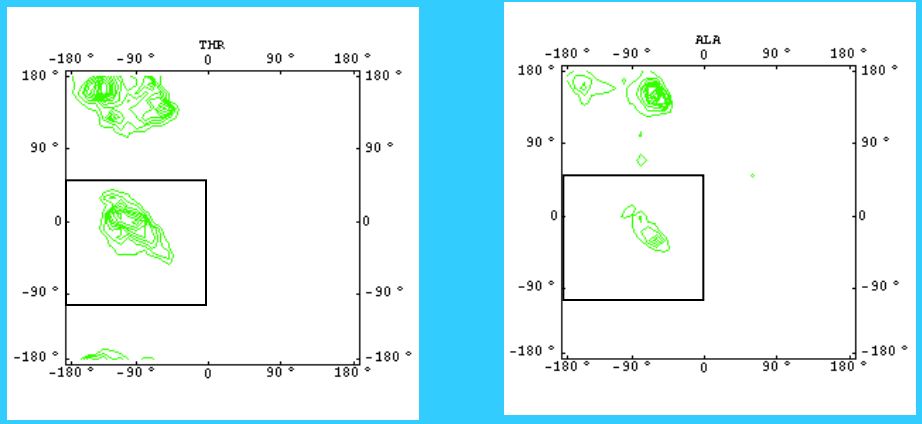
Data mine protein data base (PDB) of crystal structures:

Extract the distributions for each type of amino acid



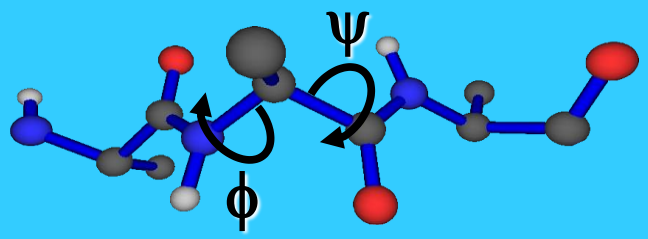
THR

ALA



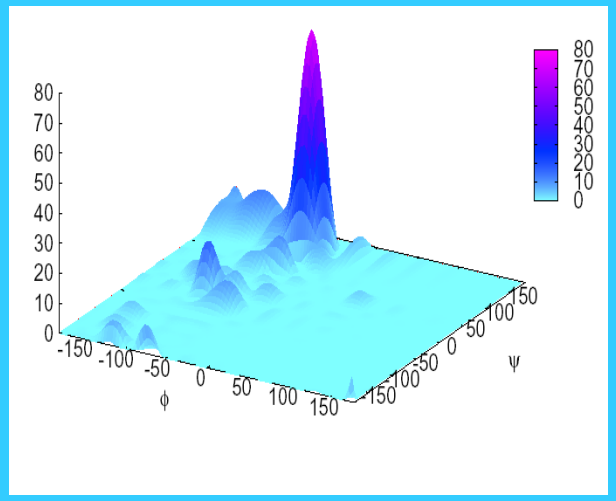
ψ

ϕ

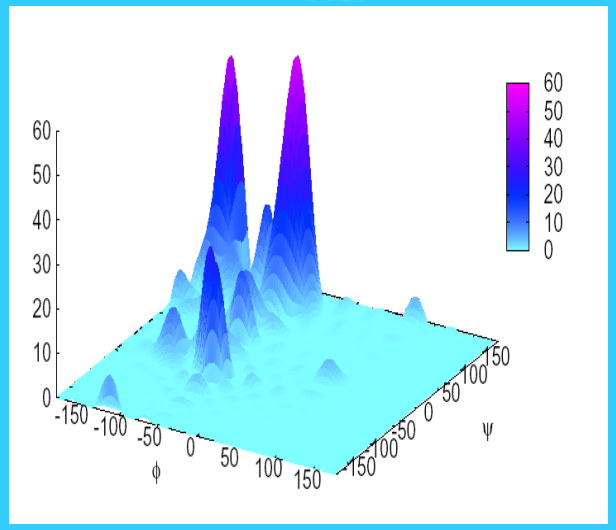


Ramachandran Map of ALA with neighbors=ALA, ASP

ALA_{ALA}



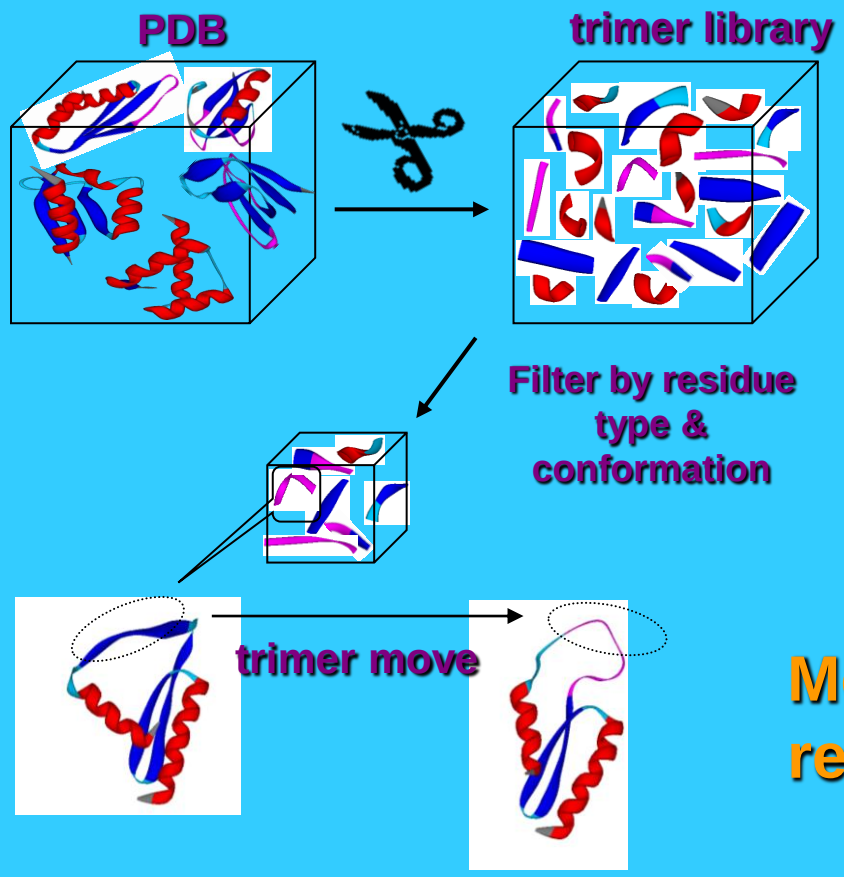
ALA_{ASP}



Map depends on neighbor type and conformation

Strong correlations in sequence

Move set uses highly selected trimers from PDB

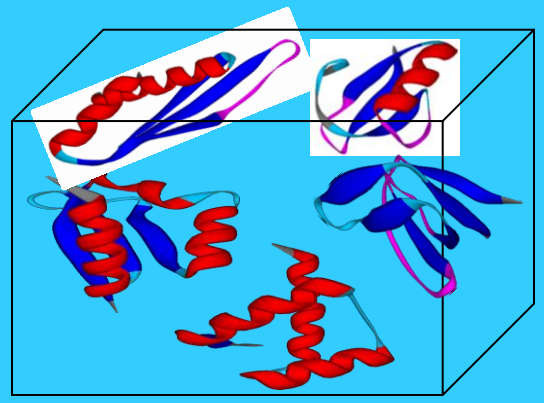


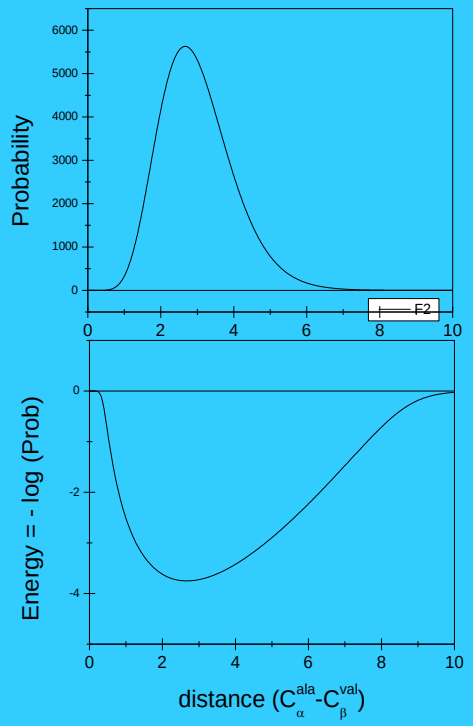
Move ϕ & ψ of central residue of trimer

Specifying side chain type and backbone geometry implicitly includes all-atom side chain information

Knowledge-based energy function: Assign interaction energies according to the observed distances in the PDB

PDB





$Prob_{PDB}(r_{ij})$

Probability of finding 2 atoms some distance apart

e.g. $Dist(C_\alpha^{ala} – C_\alpha^{val})$.

$Energy_{PDB}(r_{ij}) = -ln(Prob_{PDB}(r_{ij}))$

Each stage of folding is a Monte Carlo simulated annealing minimization

Secondary structure prediction methods…

Statistics

Single residue and nearest neighbors
Chou and Fasman (1974)

QVD

Secondary structure probability conditional on identity of neighbors

Machine Learning

Neural Network
Jones (1999)
Dor & Zhou (2007)

Support Vector Machines
Kim (2003)
Hu et al. (2004)

Multiple sequence alignments

+ SASA, …

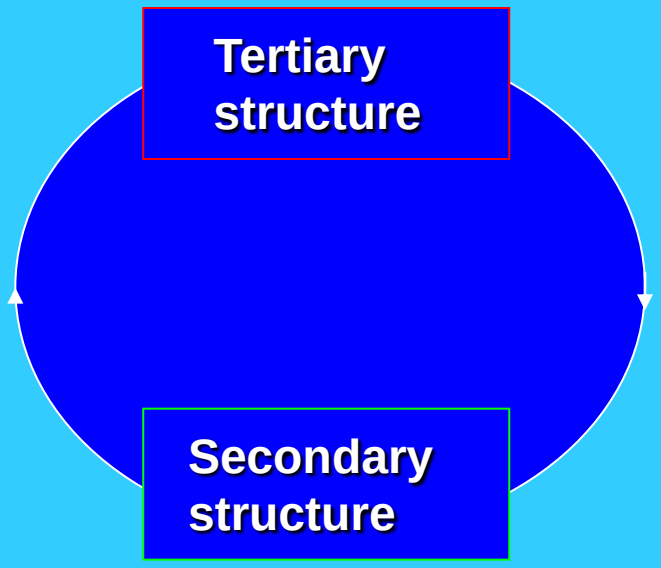
Current Accuracy of Best Methods is 75 – 80 %:
Good for α, but poorer for β and α/β

McGruffin and Jones (2003): PSIPRED
Dor and Zhou (2007)

What we do differently:

Couple secondary and tertiary structure *during the folding process*

Restrict possible secondary structure as the chain folds



"Eliminate all other factors, and the one which remains must be the truth."

A. C. Doyle, *The Sign of the Four* (1890)

ItFix Algorithm

Starting configuration (no SecStr restriction)

Round 1

Stage 1. Fold 100 times with trimers from entire PDB

Stage 2. Refine each Stage 1 end structure

Round 2

Stage 1. Fold 100 times with restricted library

Stage 2. Refine each Stage 1 end structure

Round 3

Stage 1. Fold 100 times with restricted library

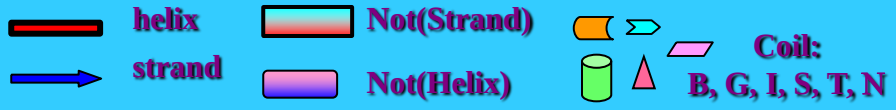
Stage 2. Refine each Stage 1 end structure

Repeat until no further fixing is possible…

Final Round

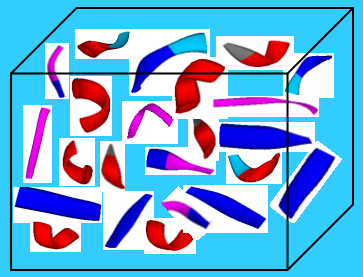
Stage 1. Fold 100 times with restricted library

Stage 2. Refine each Stage 1 end structure



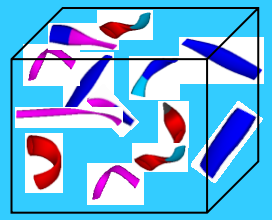
Predicted secondary structure and 3-D model

Trimer library: Full PDB



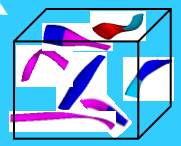
Keep track of where the trimers came from (Helix, extended, coil)

Round 1: Filter by residue identity only

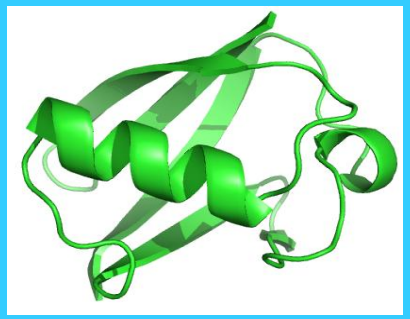


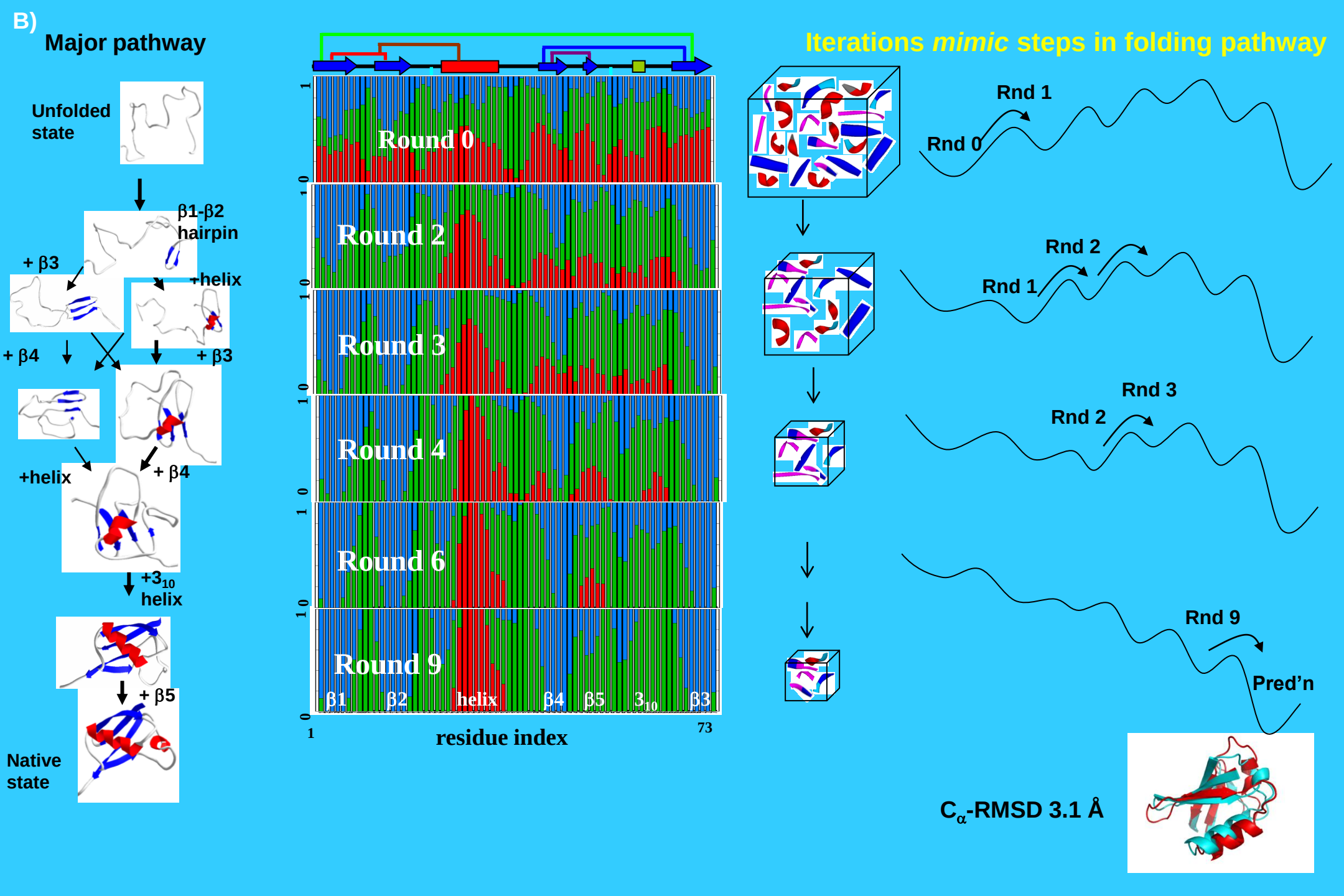
Eliminate an option (H,E,C) if P_{Helix}, P_{Coil} < 10%, P_{Ext} < 5%

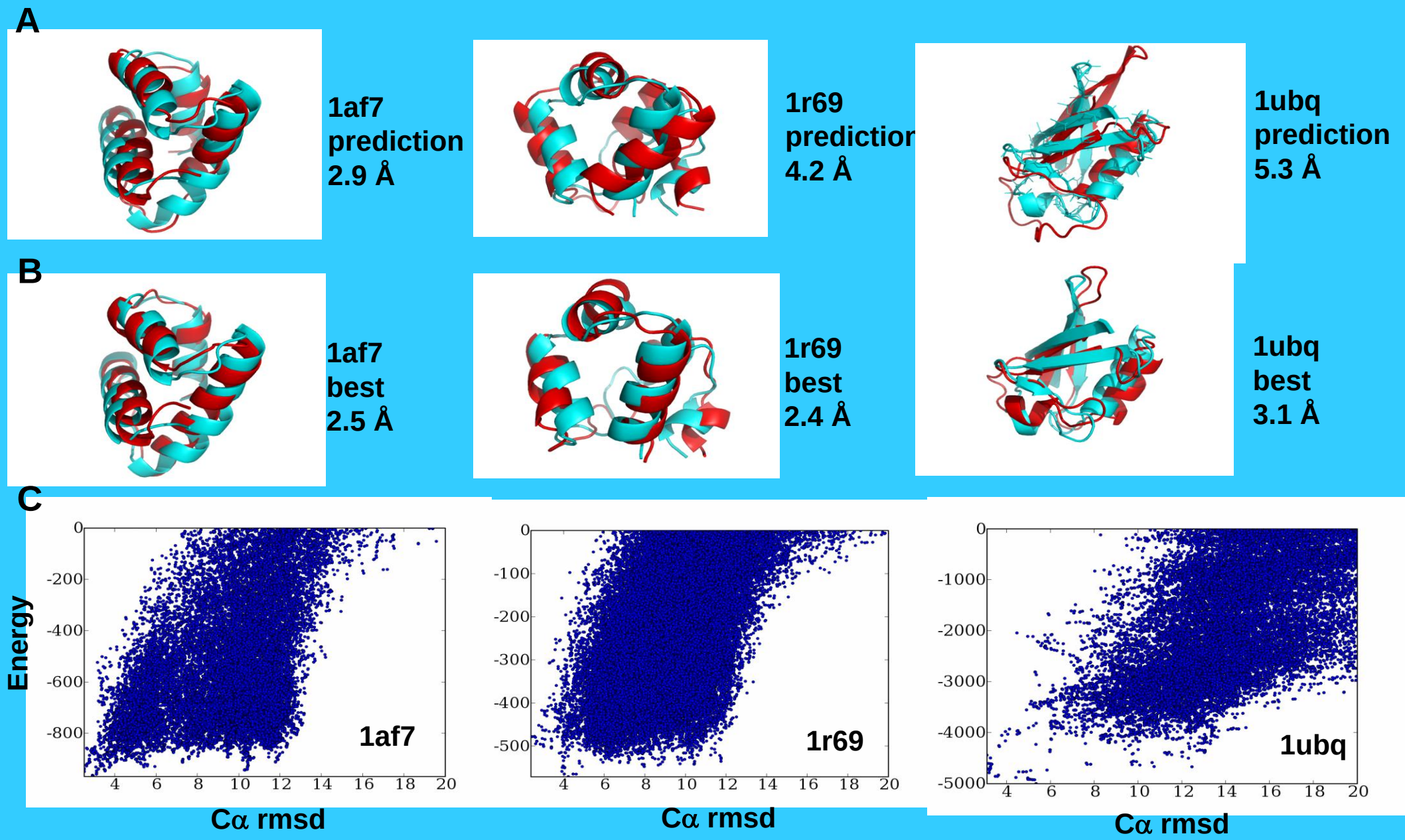
Round 2



Rounds 3,4 …



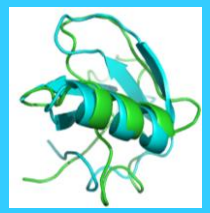

B)
Major pathway
Unfolded state
β1-β2 hairpin
+ β3
+helix
+ β4
+ β3
+helix
+ β4
+3₁₀ helix
+ β5
Native state
Round 0
Round 2
Round 3
Round 4
Round 6
Round 9
β1
β2
helix
β4
β5
3₁₀
β3
1
73
residue index
Iterations mimic steps in folding pathway
Rnd 0
Rnd 1
Rnd 2
Rnd 3
Rnd 9
Pred'n
Cα-RMSD 3.1 Å


A
1af7
prediction
2.9 Å
1r69
prediction
4.2 Å
1ubq
prediction
5.3 Å
B
1af7
best
2.5 Å
1r69
best
2.4 Å
1ubq
best
3.1 Å
C
Energy
0
-200
-400
-600
-800
4 6 8 10 12 14 16 18 20
1af7
Cα rmsd
0
-100
-200
-300
-400
-500
4 6 8 10 12 14 16 18 20
1r69
Cα rmsd
0
-1000
-2000
-3000
-4000
-5000
4 6 8 10 12 14 16 18 20
1ubq
Cα rmsd

1AF7 3.4 Å RMSD

```
Native   ---HHHHHHHHHHHHHHH-----GGGHHHHHHHHHHHHHHHT---HHHHHHHHHH-TT-THHHHHHHH-
ItFix    ---THHHHHHHHHHHHHHT---------HHHHHHHHHHHTT----THHHHHHHH--T-HHHHHHHHH-
SSPro    ---HHHHHHHHHHHHHHHHHHE-TTHHHHHHHHHHHHHHHHT--HHHHHHHHHHT-TTHHHHHHHHHH-
PSIPRED  ---HHHHHHHHHHHHHHH-----HHHHHHHHHHHHHHHHHH----HHHHHHHHH----HHHHHHHHH--
```

1TIF 5.4 Å RMSD

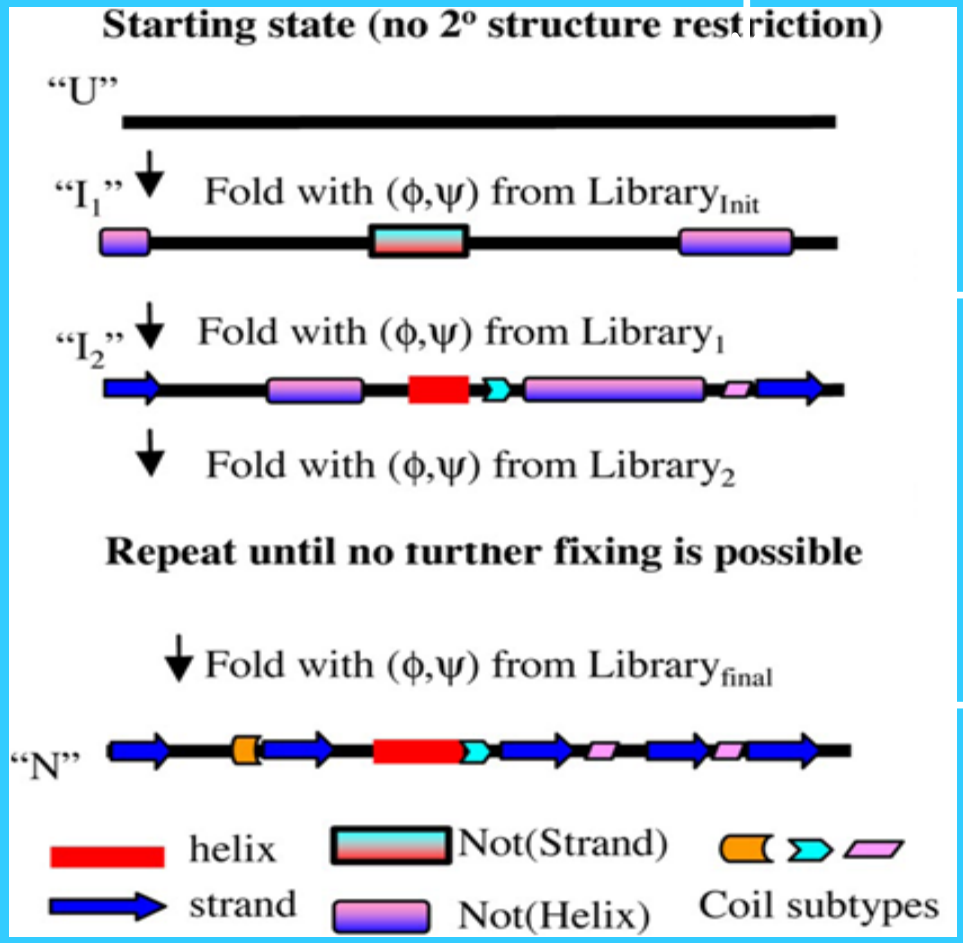


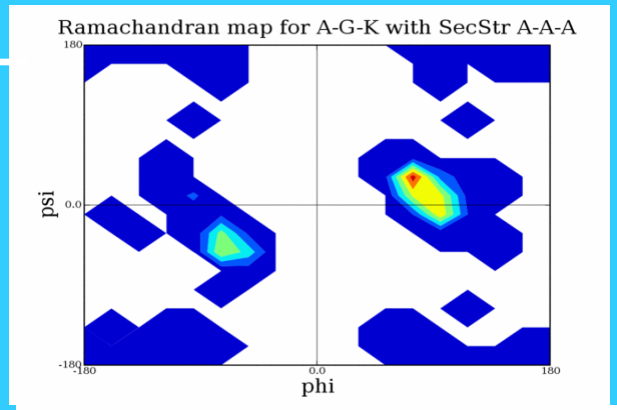
```
Native   --BGGG---SEEEEE-TTS-EEEEEEHHHHHHHHHHTT-EEEEEETTSSS-EEEEE-
ItFix    -BB-T------EEEE-TTS-EEEEE-HHHHHHHHHHTT--EEEE-TT----EEEEE-
SSPro    --BBTEEE-EEEEEEETTT-EEEEE-HHHHHHHHHHHT--EEEE-TT----EEEE--
PSIPRED  ----------EEEEE-----EEEEE-HHHHHHHHHH----EEEE-------EEEE--
```

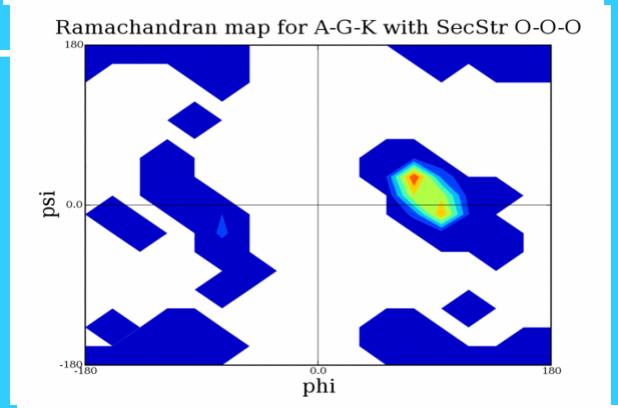
1SAP 7.8 Å RMSD

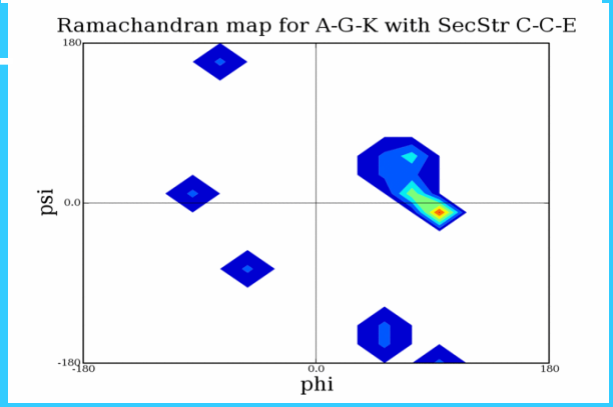
```
Native     --SEEEEETTEEEEE-GGGEEEEEEETTEEEEEE-BTTB--EEEE-GGG--HHHHHHHHHHHHHH-
ItFix      -EEEEEETT-EEEEET-EEEEEEEET-EEEEEE-TTT------E--T---HHHHHHHHHHHHHH-
SSPro      -EEEEEEE--EEEE-HHHEHEEEEHEEEEEEEE-TT-----EEE--T--HHHHHHHHHHHHH---
PSIPRED    -EEEEEEE---HHH--HHHHHHHHHH--EEEEEE----------------HHHHHHHHHHHHHH--
```

ItFix (Iterative Fixing): An algorithm for homology free protein structure prediction



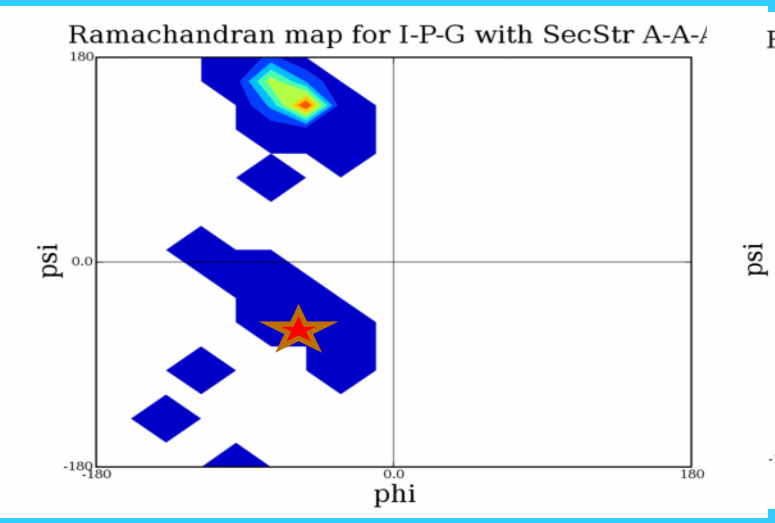


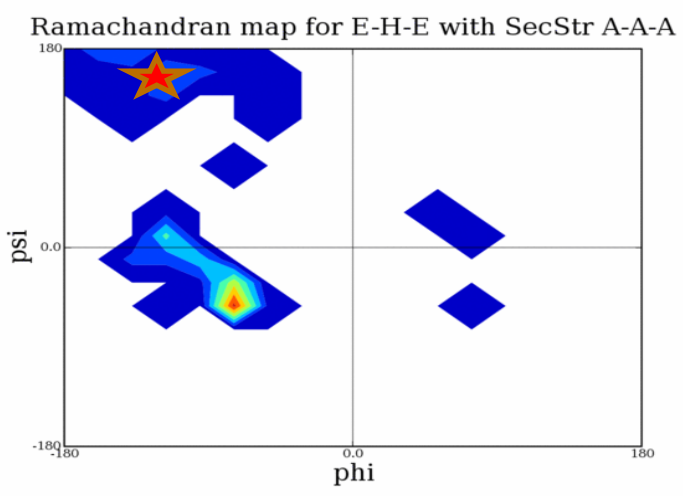


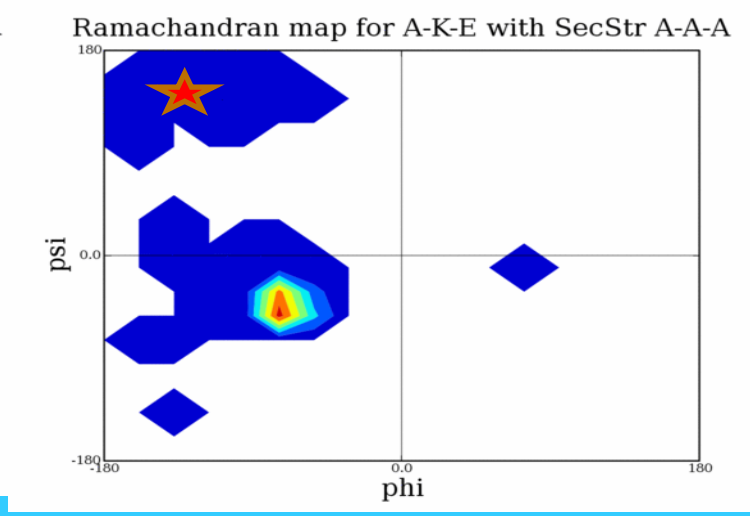


1dcj: a big homology-free failure

```
1dcj sequence DHTLDALGLRCPEPVMMVRKTVRNMQPGETLLIIADDPATTRDIPGFCTFMEHELVAKETDGLPYRYLIRKG
1dcj SecStr   NEEENNTTNNTTHHHHHHHHHHHHNNTTNNEEEEENSTTHHHHHHHHHHTTNEEEEEENSSSSEEEEEENN
```







★ **Native angles**

ItFix Enhanced with Evolutionary Diversity

target sequence

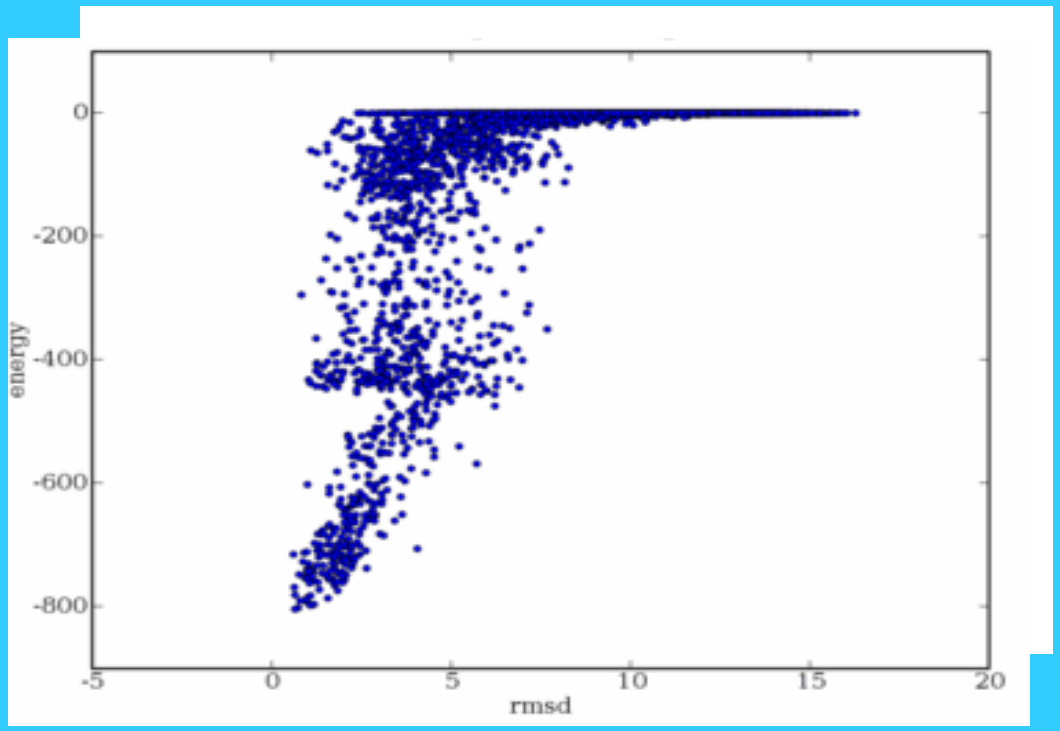
MQIFVKTLTGKTITLEV

EEEEEEETTSNEEEEEE

sequence database

IEIKIRDIYSKTYKFMA
IEITCNDRLGKKVRVKC
MRLFIRSHLHDQVVISA
MKLSVKSPNGRIEIFNE
LQFFVRLLDGKSVTLTF
IEITLNDRLGKKIRVKC
IEIWVNDHLSHRERIKC
MDVFLMIRRQKTTIFDA
IIVTVNDRLGTKAQIPA
MRISVIKLDSTSFDVAV
MNVNFRTILGKTYTITV
MLLTVRDRSELTFSLQV
MQIFVTTPSENVFGLEV
MSLTIKF-GAKSIALSL
MKYRIRTISNDEAVIEL
INVVVKH-QGKKYDVDV
MLVEFRVLTGERFRLNL
MQVLIKTLSGQRVPFDF
MKLTVKTLKGTQFEIRV
ITLIVKAPNQQIETIKC
MQLFVRGT--KTHTLQI
MQLFVC--AQELHTLEV
MKVTIRSSTEISYSVTI
IQLKVRSPDGSEVYFKI
IRIYVEN-AGRTCTLDT
MQIFVKTLTGKTITLEV
MLITVRDLSEVAFSLQV
MQIFEKTLTDKTIILDV

… ~1000 sequences

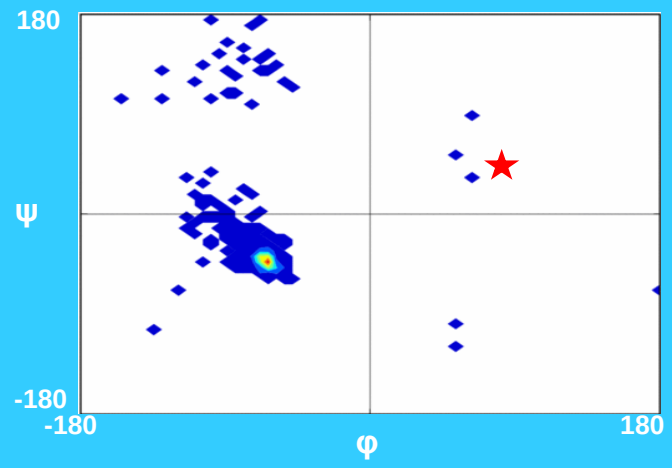


1af7 3.4 Å prediction from simulation of final round

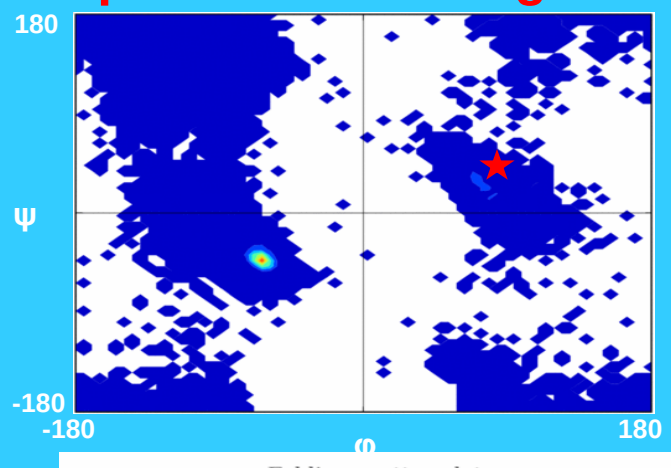
SPEED can enhance the Ramachandran distribution sampled during folding

1b72 AA seq TTRQLTELEKEFHFNKYLSRARRVEIAATLELNETQVKIWFQNRRMKQKK
1b72 SecStr NNHHHHHHHHHHHTTNSSNNHHHHHHHHHHHTNNHHHHHHHHHHHHHHHHHN

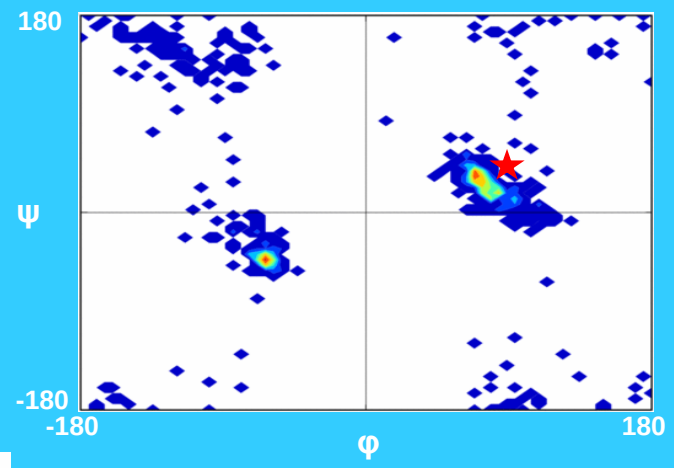
L-E-L

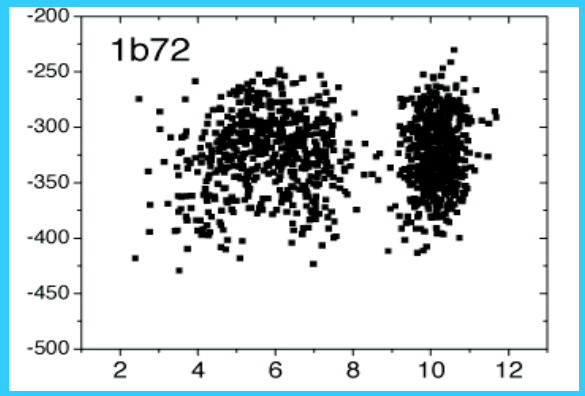


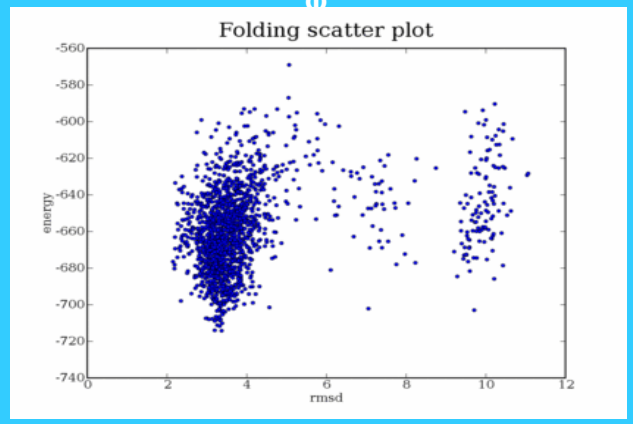
position 31 in alignment



L-G-L







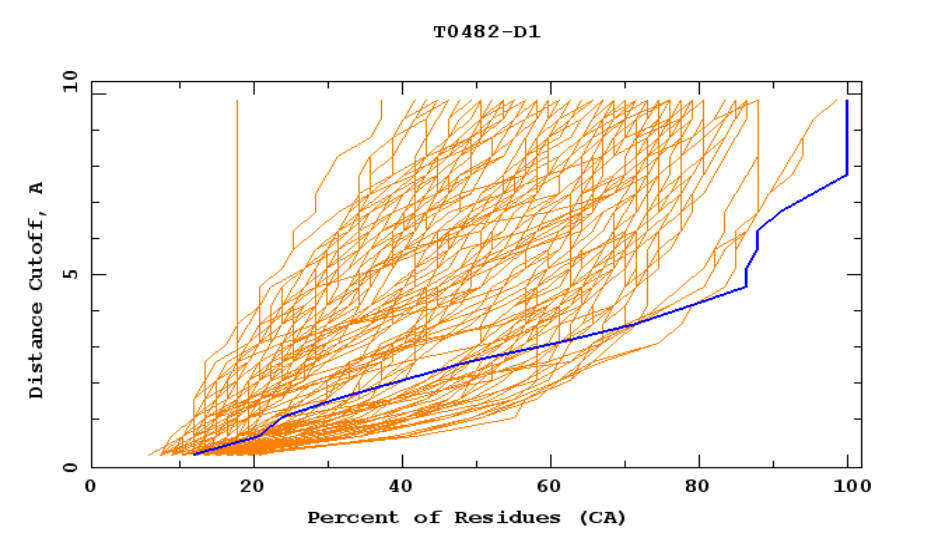
★ Native

ItFix Homology-free

ItFix SPEED

Protein			Q3 2° structure accuracy Q3 (Q8) % accuracy				3° structure accuracy Lowest (Best), Å	
PDB code	Length	Fold	ItFix	SPEED	SSPro	PSIPRED	ItFix	ItFix SPEED
1af7	69	α	97 (86)	94 (87)	86 (81)	90	2.9 (2.5)	4.0 (2.0)
1b72	50	α	88 (84)	98 (92)	68 (72)	84	3.5 (1.6)	3.3 (1.3)
1csp	67	β	79 (67)	87 (70)	75 (67)	88	10.5 (6.0)	9.8 (4.7)
1dcj	72	αβ	45 (29)	83 (65)	65 (56)	89	13.3 (7.6)	5.5 (4.3)
1mky	77	αβ	86 (70)	83 (65)	87 (71)	90	6.9 (6.1)	7.8 (4.3)
1r69	61	α	93 (89)	98 (89)	74 (72)	92	4.2 (2.4)	4.0 (2.3)
1shfA	59	B	76 (56)	71 (51)	85 (69)	80	12.2 (6.7)	6.5 (5.5)
1tif	57	αβ	89 (79)	91 (81)	76 (70)	93	11.3 (4.2)	5.2 (3.5)
1tig	86	αβ	83 (70)	88 (81)	69 (67)	83	6.4 (5.3)	6.4 (4.5)
1ubq	73	αβ	92 (69)	94 (68)	88 (67)	90	5.3 (3.1)	2.6 (1.9)

ItFix with SPEED successfully applied in CASP8



T0482

- **No PDB homologues**
- **Good sequence alignment**
- **4.8 Å prediction**

Conclusions

We couple the formation of tertiary structure and the identification of secondary structure to improve both types of structure prediction

Goal: Predict secondary structures – Exceed the 90% barrier… without homology

Future applications:
Improved *ab initio* structure prediction
Assist in identifying and ranking of threading targets, large insertions, etc.

Future improvements: Bigger proteins, multi-domain, docking, design
Better energy functions = better folding and predictions
Use SWAXS data from APS/ANL to aid theory

Acknowledgements

Prof. Tobin Sosnick

Joe DeBartolo - Ab initio folding, secondary structure prediction

Dr. Andres Colubri – Folding simulations, software

Dr. Abhishek Jha (Agios) – coil library, unfolded state, α, β propensities, structure refinement, electrostatics

James Fitzgerald (Stanford) – Stat. potentials, torsional dynamics

Dr. Michael Wilde (ANL) – High performance computing

Glen Hocky (Columbia) – Folding & high performance computing

All-atom statistical potential

Dr. Min-yi Shen (Applied Biosystems), **Prof. A. Sali, UCSF**

Funding: NIH, NSF, Burroughs Wellcome Fund