Introduction into the Rosetta Monte Carlo simulation program.

Jens Meiler/Shannon Smith

CPBP 8330

17 August 2020

shannon.t.smith.1@vanderbilt.edu

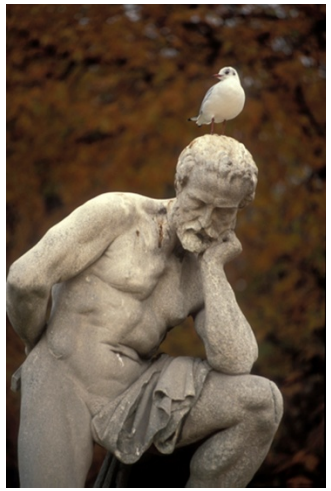
Goals for this talk

1. **Intro to general topics:**
   a) Computational structural biology
   b) Sampling: MC/simulated annealing
   c) Scoring: statistical potentials, physics-based, empirical, machine learning
2. Ab initio structure prediction
   a) Low-resolution centroid predictions
   b) High-resolution full-atom refinement
   c) CASP (Critical Assessment of protein Structure Prediction)
3. Comparative modeling
   a) Why we do comparative modeling? MPs/GPCR example
   b) Sequence alignments + threading
4. Loop modeling
   a) Why we "mind the gap"? HCDR3 example
   b) Cyclic Coordinate Descent (CCD)
   c) Kinematic closure (KIC)

Central Dogma of Structural Biology

DNA

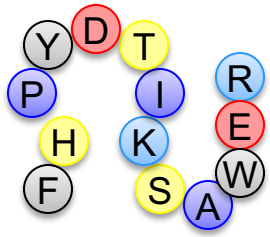
Function



Transcription

RNA

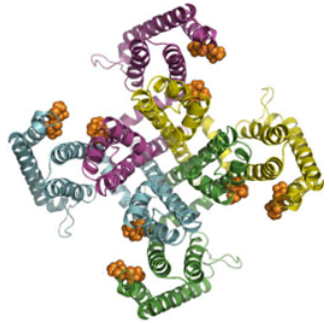
Translation



Folding

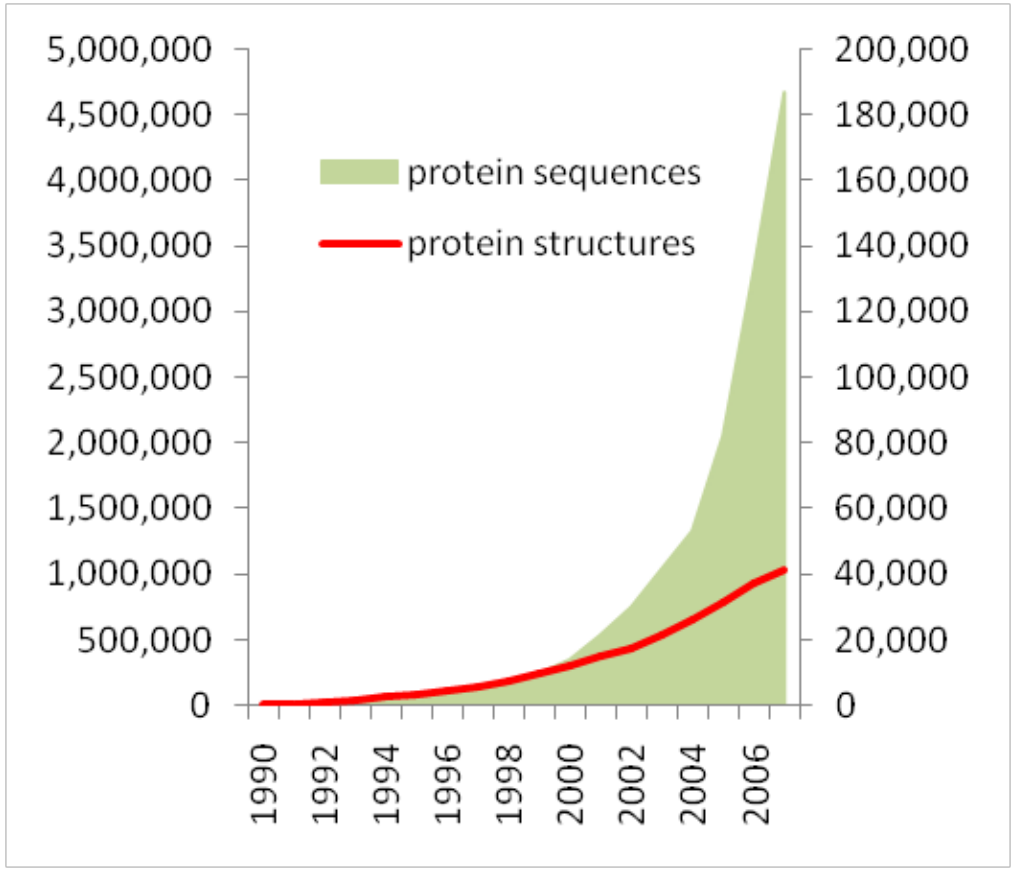
Sequence

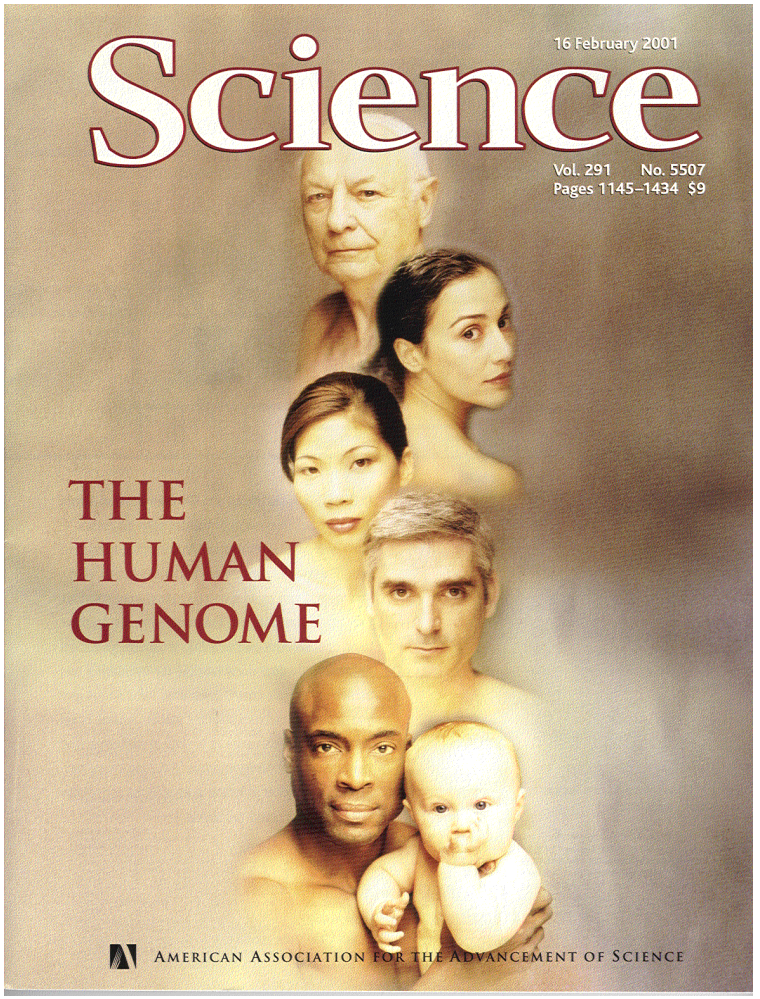
Structure



Structural Biology After the Human Genome Project

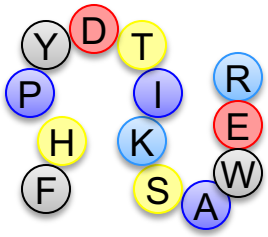
- Sequence versus Structure



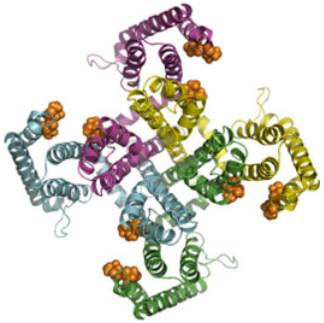


Protein Sequence and Structure Data

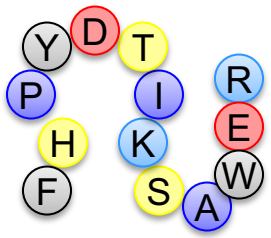
- Genbank
  - ~20,000,000 sequences
- Protein Databank
  - ~200,000 structures



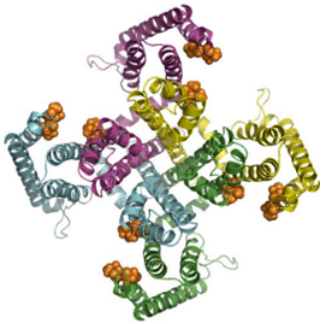
Sequence ⇄ Structure (Folding)



Protein Folding Problem

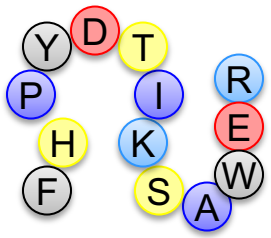


Sequence → Structure

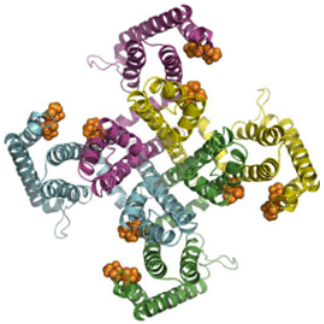


- Given a protein's AA sequence, what is its 3-dimensional fold, and how does it get there?
- Assume 100 conformations for each amino acid in a 100 amino acid protein ⇨ 10^{200} possible conformations!
- Cyrus Levinthal's paradox of protein folding, 1968.

Inverse Protein Folding Problem (also known as protein design)

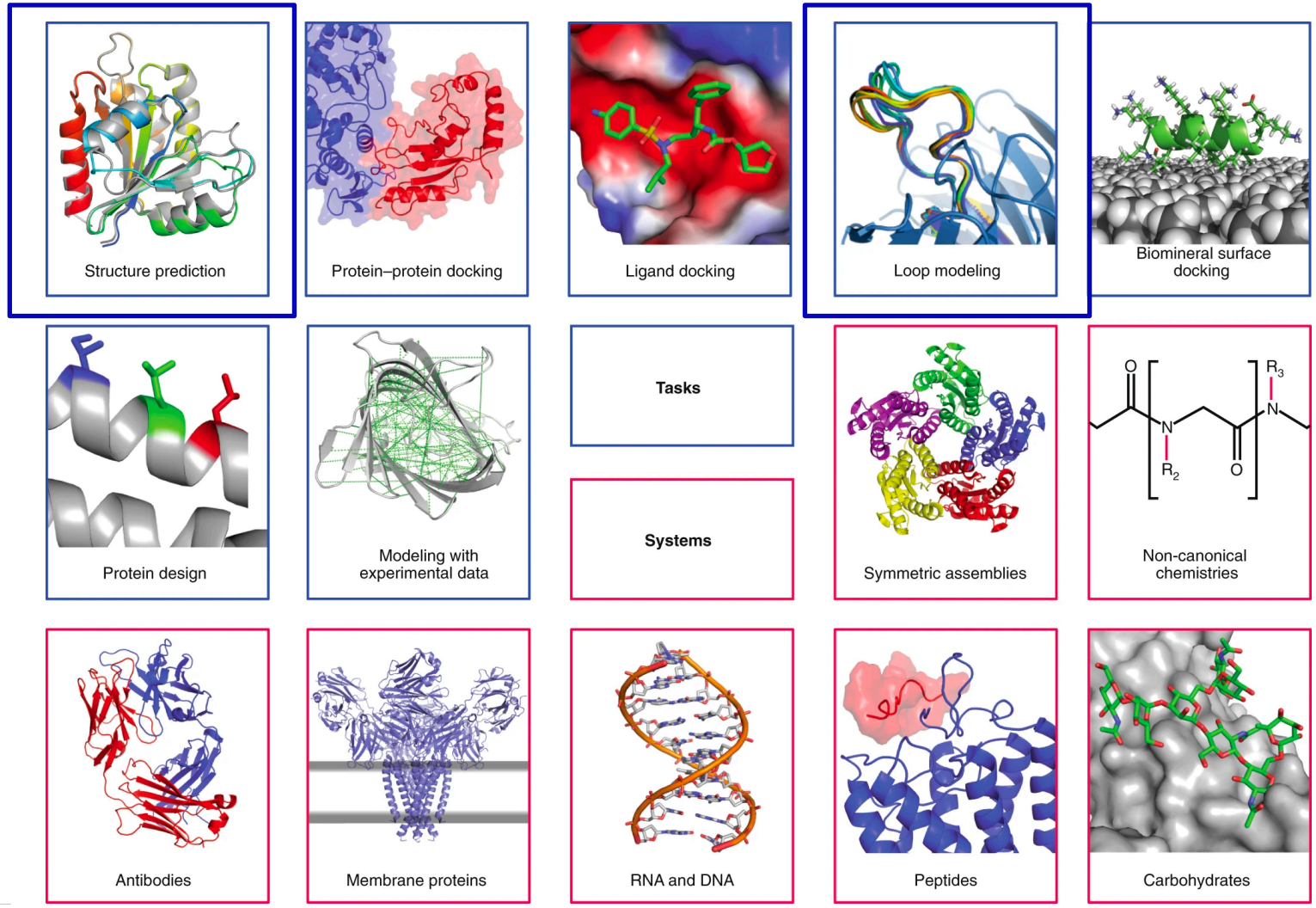


Sequence ⇦ Structure



- Given a protein fold, which primary sequence(s) fold into it?
- Assume a total of 100 conformations for all 20 natural occurring amino acids side chains in a 100 amino acid protein ⇨ 10^{200} possible conformations!
- Earth is less than 10^{10} years old.

Rosetta: A Unified Framework for Tackling Molecular Modeling



Koehler et al. "Macromolecular modeling and design with Rosetta: recent methods and frameworks" *Nature Methods (2020)*.

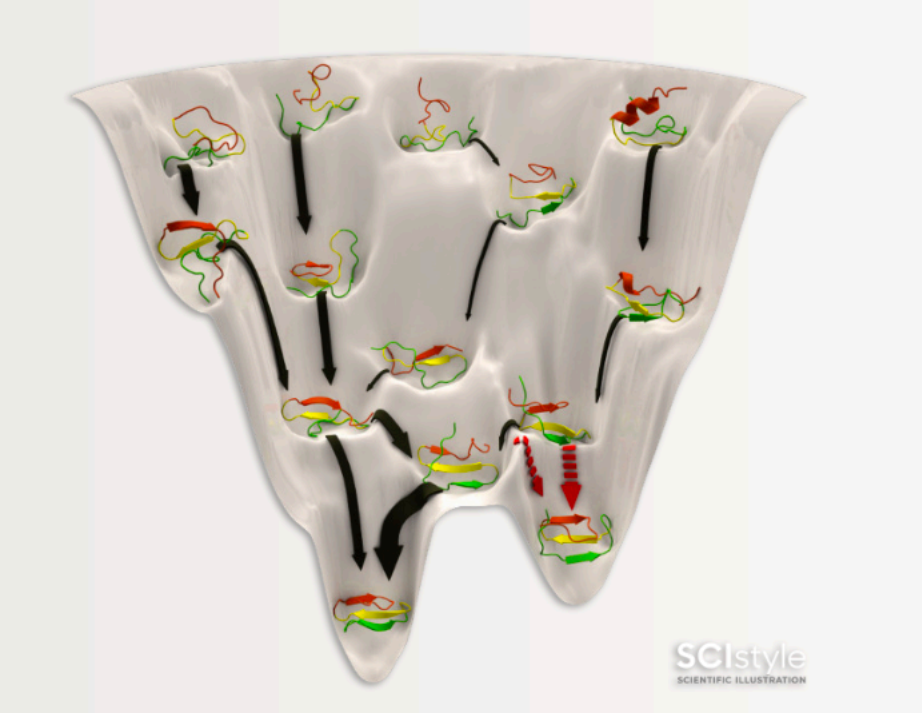
It all comes down to 2 questions: Sampling vs. scoring problems

Sampling problem
How do we find a "global minimum" or physiologically-relevant structure?

- Molecular dynamics (MD) methods: time-dependent, based on forces of individual atoms
- Monte Carlo (MC) methods: time-independent, larger conformational changes

Scoring problem
Given a sampled pose, how do we distinguish good from bad structures?

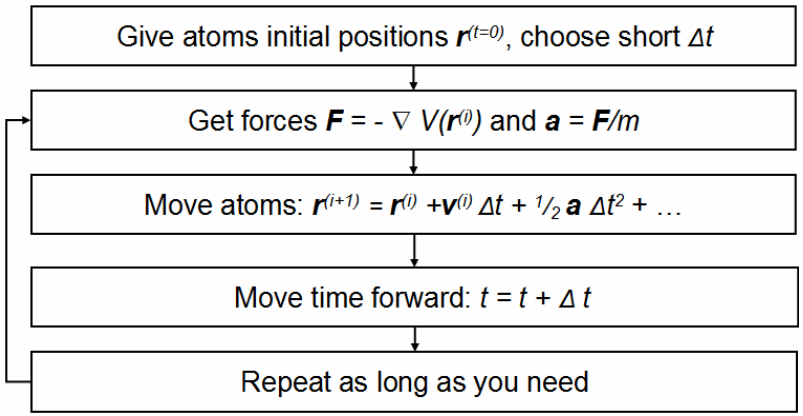
- Assign each structure a score that correlates with the free energy of that pose
- Types of score functions: physics-based, empirical, knowledge-based, machine-learning



https://medium.com/proteinqure/welcome-into-the-fold-bbd3f3b19fdd

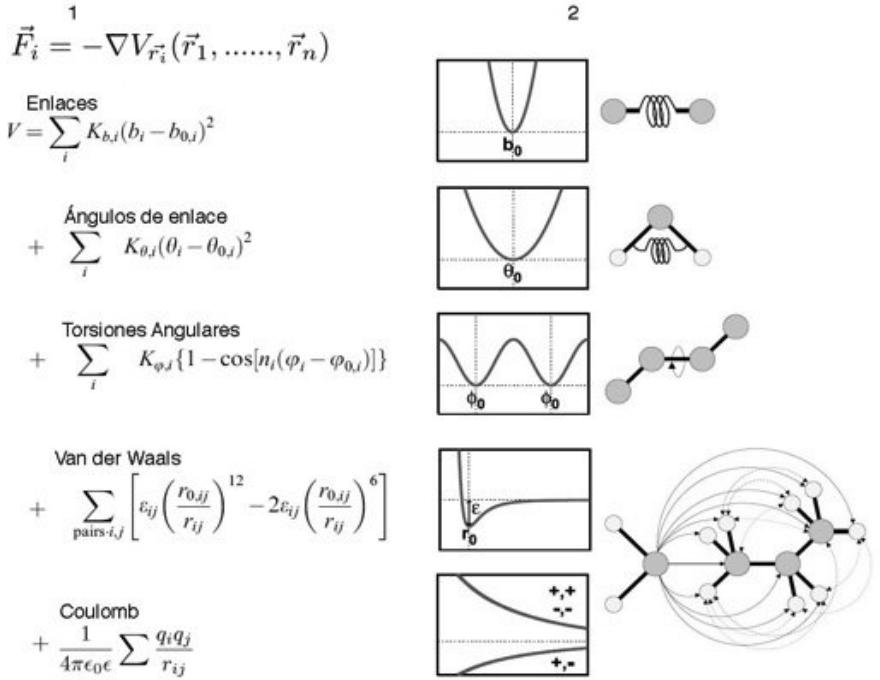
Molecular dynamics (MD) is a computer simulation of physical movements of atoms and molecules in the context of N-body simulation

- trajectories of atoms and molecules are determined by numerically solving the Newton's equations of motion for a system of interacting particles



- Δt needs to be small to hold cumulative errors from numerical integration small <=> only recently and only small, fast-folding peptides and proteins can be studied

- Molecular mechanics force field constraints distances, angles, torsions, van-der waals interactions, and coulomb interactions



Almost everything in Rosetta is done using MCM with Simulated Annealing

- Staring from a protein model M_1 a conformational change is applied to arrive at protein model M_2. The score/energy of both models is determined as E_1 and E_2.

- The new conformation is accepted according to the Metropolis criteria:

$$\begin{cases} E_2 \leq E_1: & accept \\ E_2 > E_1: & accept\ with\ P = e^{-(E_2 - E_1)/T} \end{cases}$$

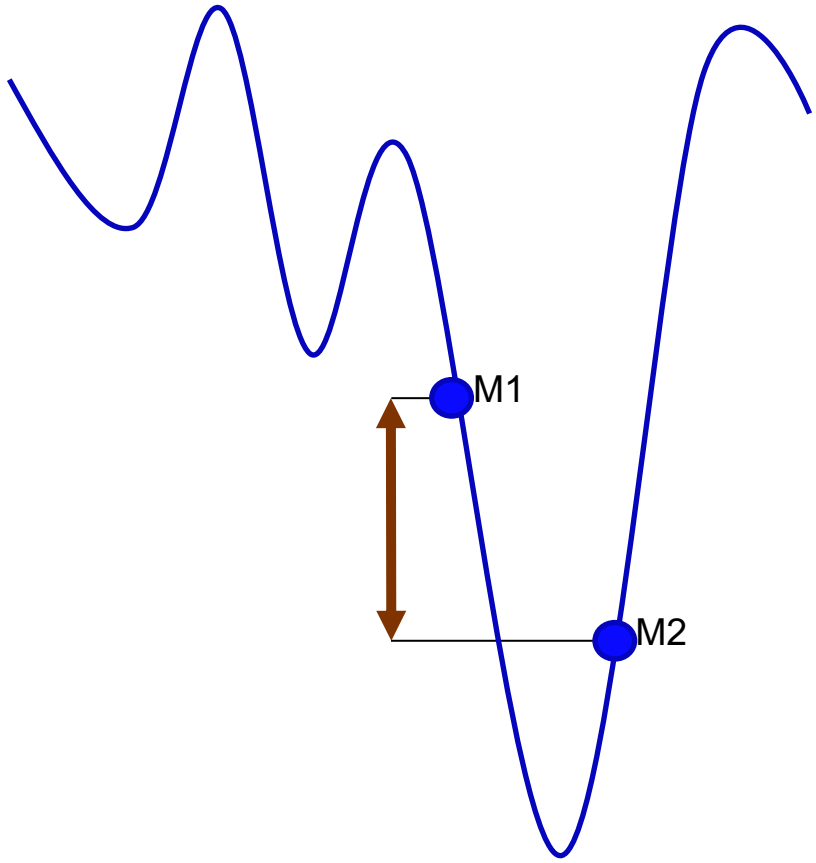
- Note that $P = 1$ if $E_2 = E_1$. As E_2 becomes larger than E_1 (unfavorable change in energy) P approaches quickly 0.

- T is a temperature parameter. The higher T, the more likely an unfavorable energy increase is accepted. In a "simulated annealing" simulation T is stepwise reduced to approach zero. I.e. the system is cooled down from a high to a low energy.

Monte Carlo + Metropolis criteria

Apply conformational change to protein model M1 to new structure model M2

$$\begin{cases} E_2 \leq E_1: & accept \\ E_2 > E_1: & accept\ with\ P = e^{-(E_2 - E_1)/T} \end{cases}$$



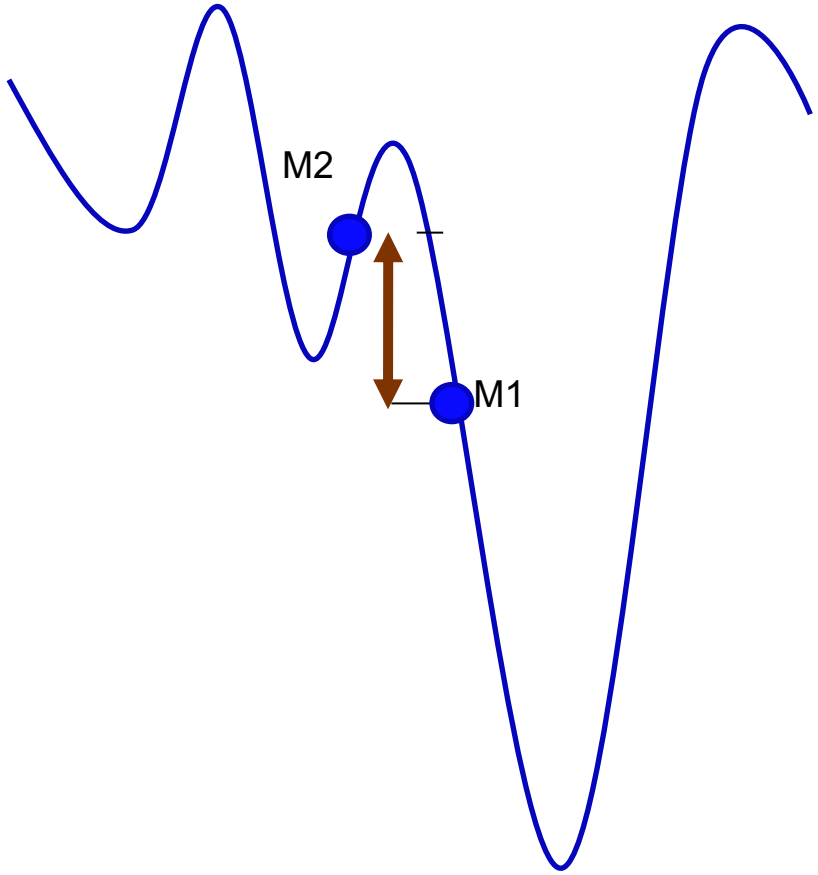
Monte Carlo + Metropolis criteria

Apply conformational change to protein model M1 to new structure model M2

$$\begin{cases} E_2 \leq E_1: & accept \\ E_2 > E_1: & accept\ with\ P = e^{-(E_2-E_1)/T} \end{cases}$$

Choose random number R [0,1)

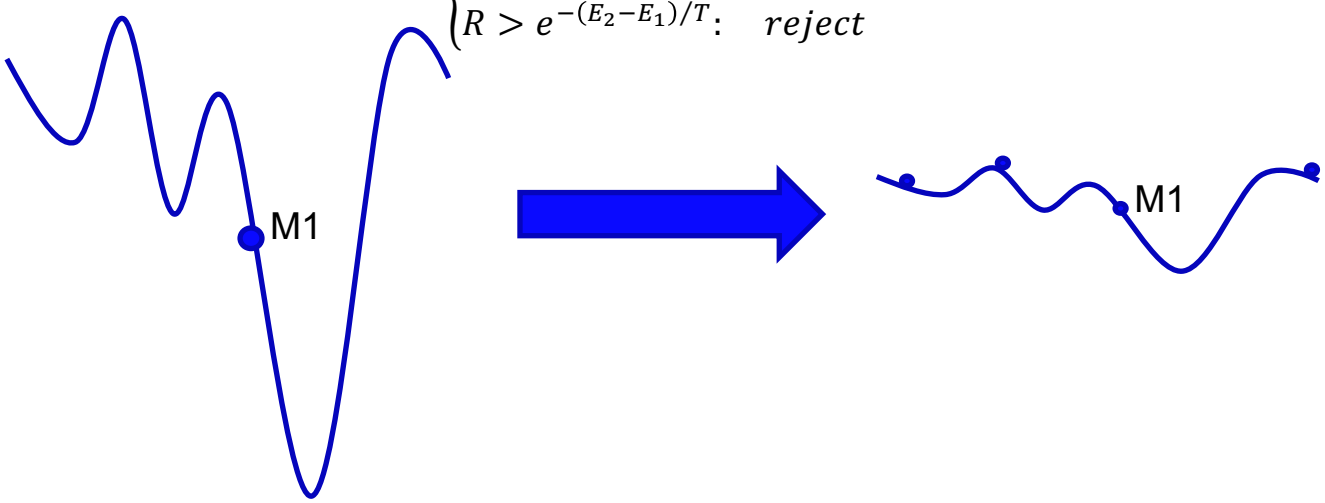
$$\begin{cases} R \leq e^{-(E_2-E_1)/T}: & accept \\ R > e^{-(E_2-E_1)/T}: & reject \end{cases}$$



Monte Carlo + Metropolis criteria with simulated annealing

What happens when the temperature is <u>increased</u>?

$$\begin{cases} R \leq e^{-(E_2-E_1)/T}: & accept \\ R > e^{-(E_2-E_1)/T}: & reject \end{cases}$$

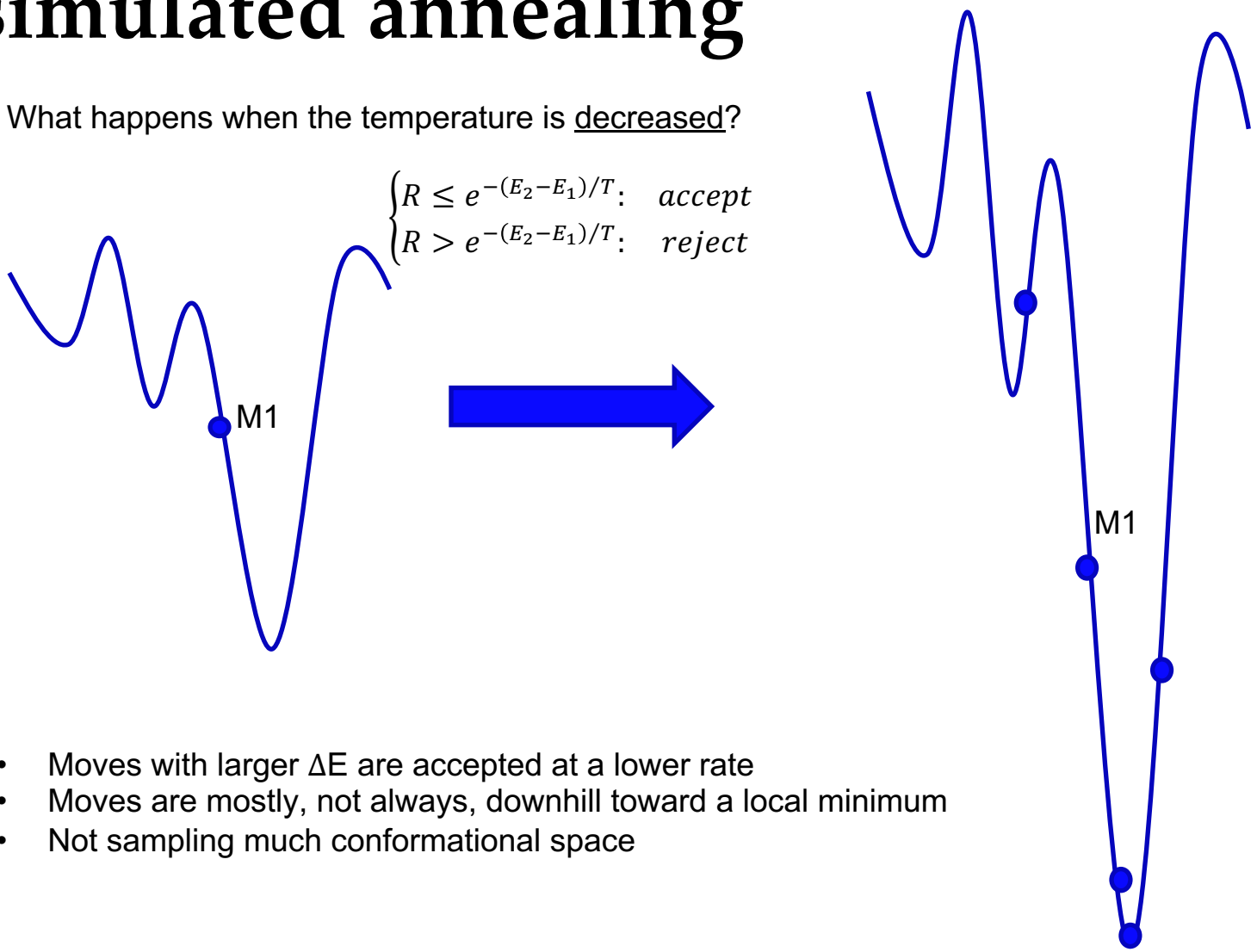


- Moves with larger ΔE are accepted at a higher rate, effectively flattening the energy landscape
- Possible to sample larger conformational space

Monte Carlo + Metropolis criteria with simulated annealing

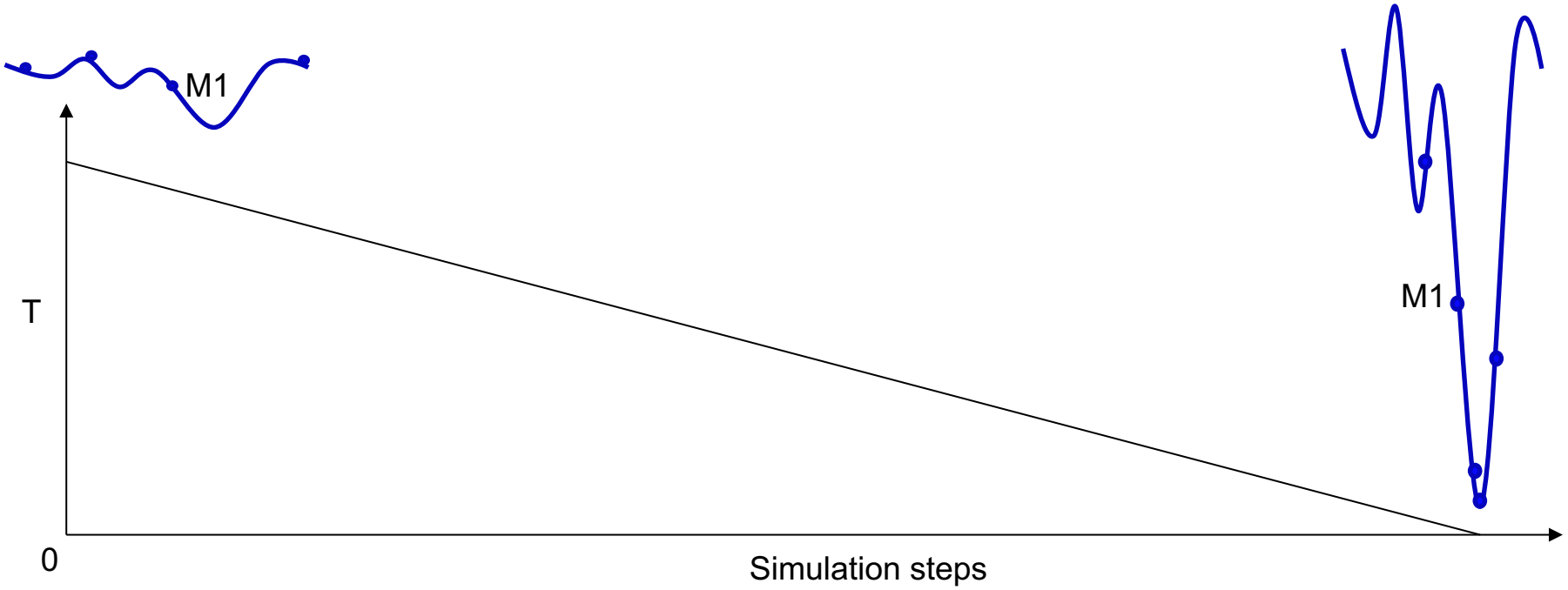
What happens when the temperature is decreased?

$$\begin{cases} R \leq e^{-(E_2-E_1)/T}: & accept \\ R > e^{-(E_2-E_1)/T}: & reject \end{cases}$$



- Moves with larger ΔE are accepted at a lower rate
- Moves are mostly, not always, downhill toward a local minimum
- Not sampling much conformational space

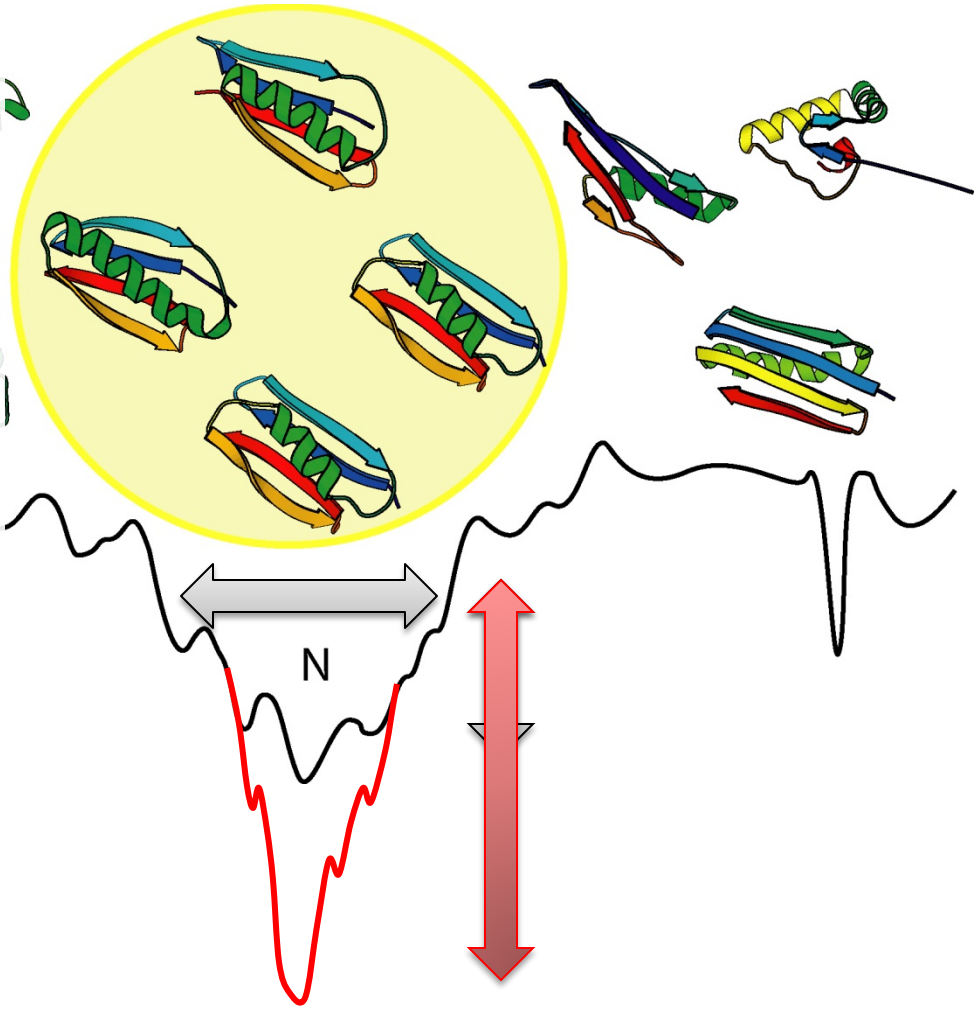
Simulated annealing gradually lowers the T to converge towards an energy minimum



Do this many times to get different outputs to get ensemble of structures that we assume are in a local minima

Native-like Protein Models Form Large Clusters

- The free energy minimum corresponds (usually) to the native protein fold
- Its depth is obscured because of the simplified energy approximation
- However, the width of the funnel leading to the free energy minimum of the native protein fold is well preserved

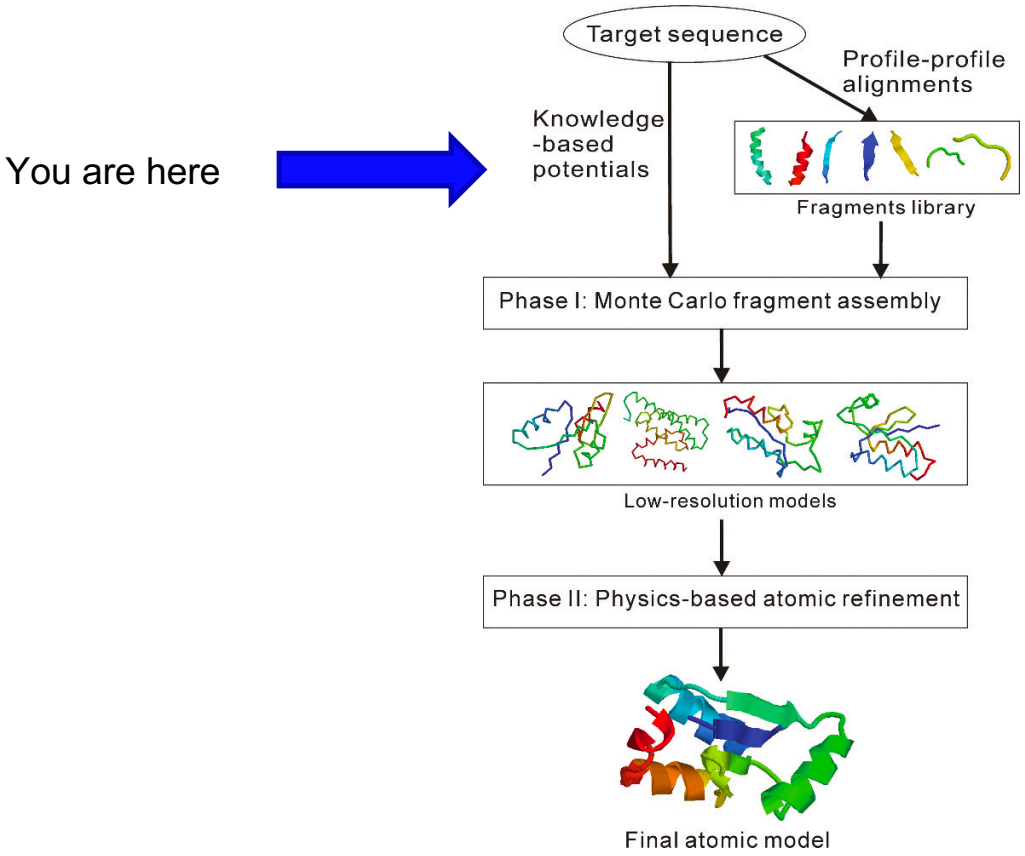


Goals for this talk

1. Intro to general topics:
   a) Computational structural biology
   b) Sampling: MC/simulated annealing
   c) Scoring: statistical potentials, physics-based, empirical, machine learning
2. **Ab initio structure prediction**
   a) Low-resolution centroid predictions
   b) High-resolution full-atom refinement
   c) CASP (Critical Assessment of protein Structure Prediction)
3. Comparative modeling
   a) Why we do comparative modeling? MPs/GPCR example
   b) Sequence alignments + threading
4. Loop modeling
   a) Why we "mind the gap"?
   b) Cyclic Coordinate Descent (CCD)
   c) Kinematic closure (KIC)

General overview of de novo structure prediction

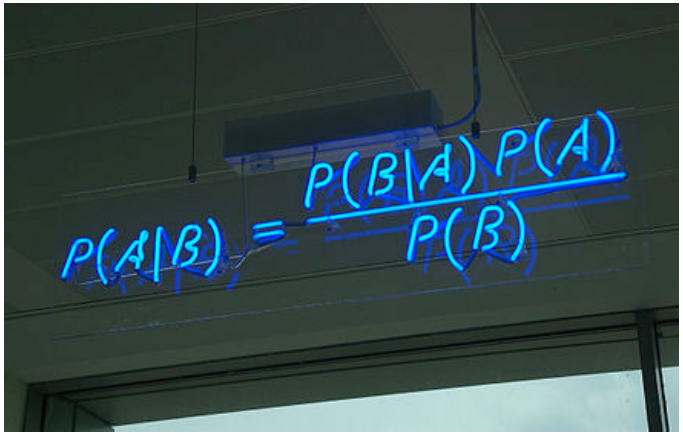
How do we go from amino acid sequence to 3D structure without other information?



Statistical / Bayesian / Knowledge-Based scoring

We assume that the more common a conformation is observed, it is because it is energetically more favorable

$P(A|B)$ =“Given B is true, what is the probability of A being true.”



Structure prediction:

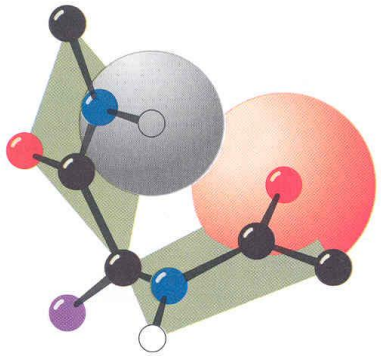
$P(str|seq)$ = Given a sequence *seq*, what is the probability of observing structure *str*?

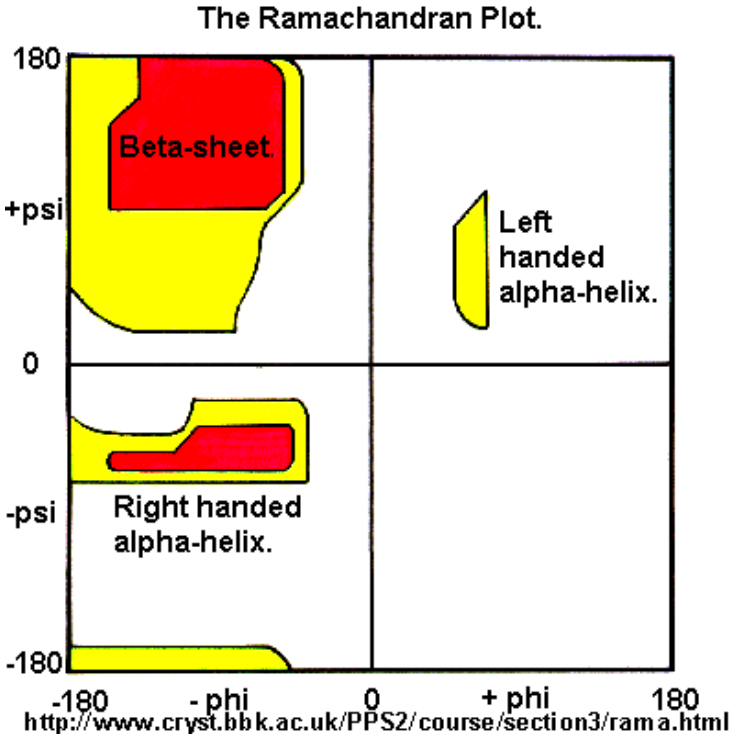
Design:

$P(seq|str)$ = Given a structure *str*, what is the probability of observing sequence *seq*?

K. T. Simons, C. Kooperberg, E. Huang and D. Baker; "Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions"; *J. Mol. Biol.*; **1997**; Vol. 268 p. 209-225; K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff and D. Baker; "Improved Recognition of Native-Like Protein Structures Using a Combination of Sequence-Dependent and Sequence-Independent Features of Proteins"; *Proteins: Structure, Function, and Genetics*; **1999**; Vol. 34 p. 82-95.

The Ramachandran Plot is a common example of using knowledge-based methods

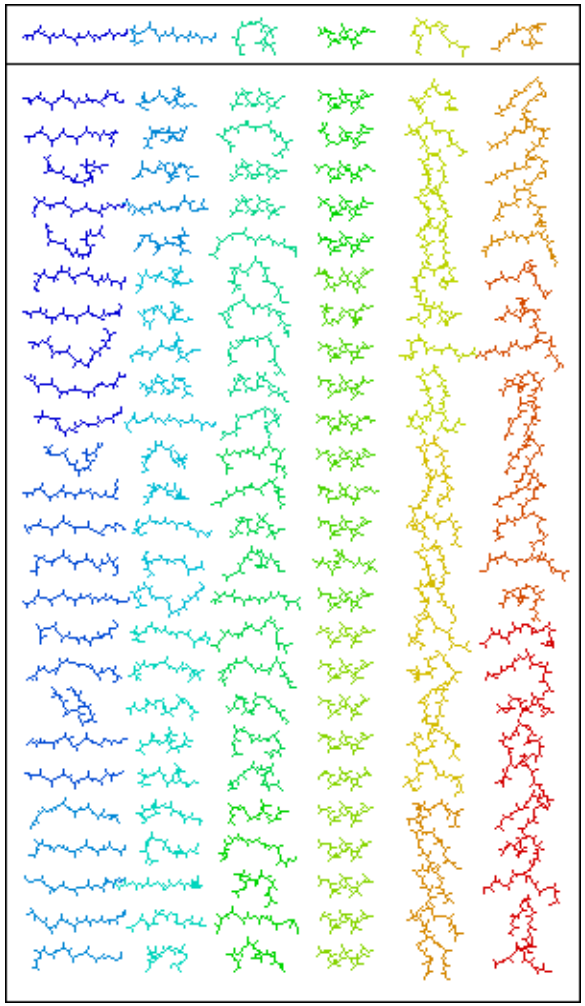




$$E_{aa,SecStr} = -\ln(P(SecStr|aa))$$

- Features with greater probability are assigned a more negative score (aka favorable score)
- Features with less probability are assigned a less negative or positive score (not favorable)

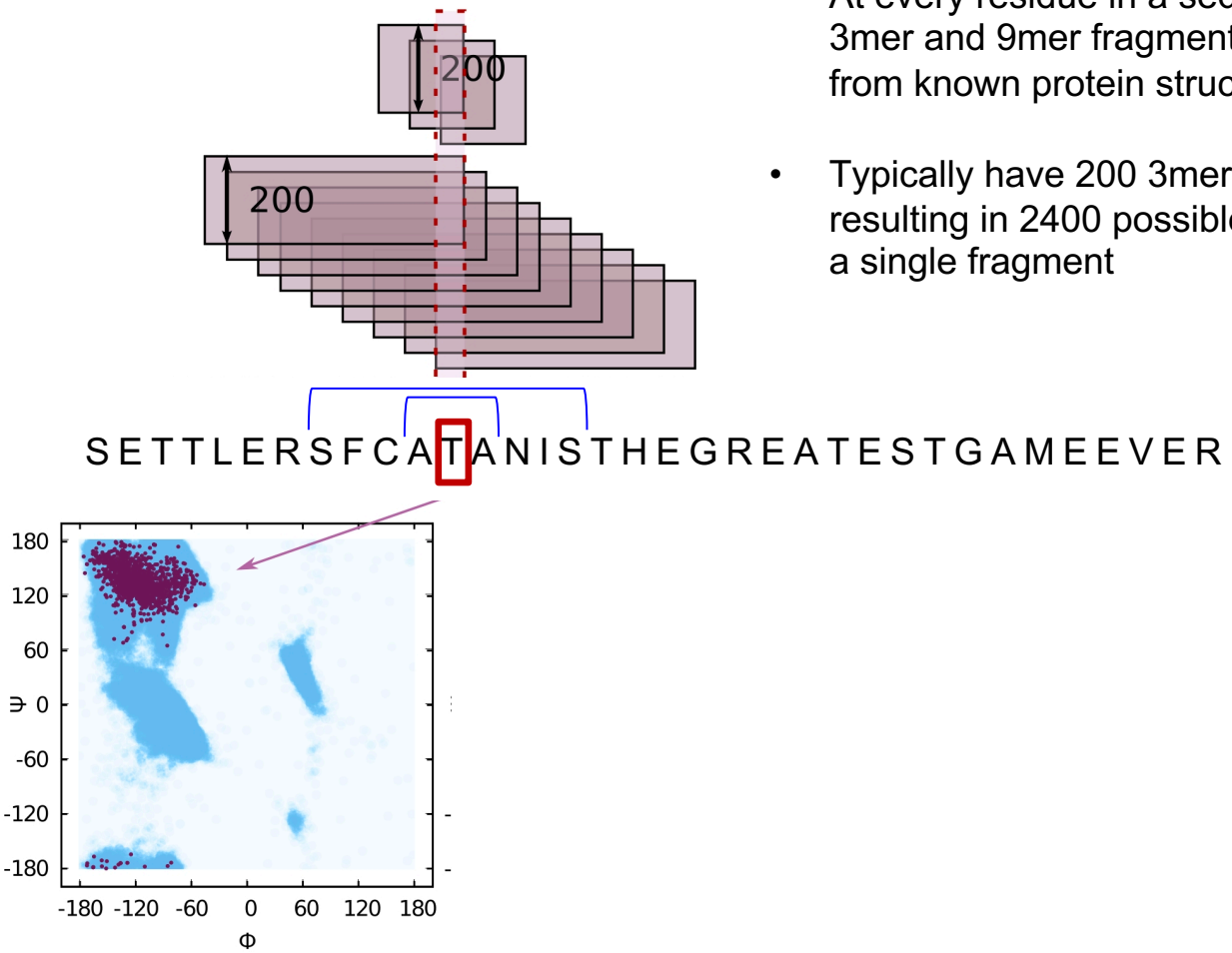
Local Sequence Bias – Rapid Approximation of Local Interactions



- **While not every protein fold is present in the protein databank, all possible conformations of small peptides fragments are!**

- Approximate local interactions using the distribution of conformations seen for similar sequences in known protein structures
- For each sequence window, select fragments that represent the conformations sampled during folding

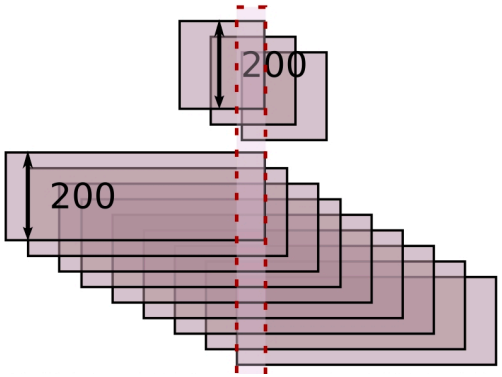
Fragment sampling



- At every residue in a sequence, generate 3mer and 9mer fragments defined by Φ/Ψ from known protein structures
- Typically have 200 3mers and 200 9mers, resulting in 2400 possible combinations at a single fragment

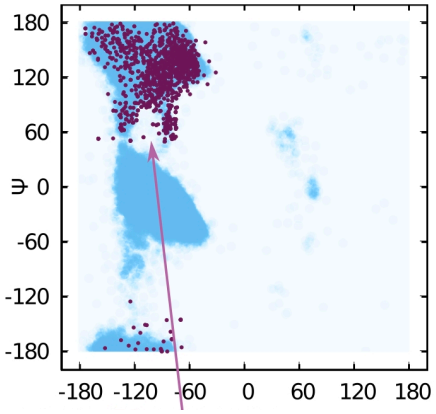
Gront D, Kulp DW, Vernon RM, Strauss CE, Baker D. Generalized fragment picking in Rosetta: design, protocols and applications. Plos one. 2011 ;6(8):e23294. DOI: 10.1371/journal.pone.0023294.

Fragment sampling



- At every residue in a sequence, generate 3mer and 9mer fragments defined by Φ/Ψ from known protein structures
- Typically have 200 3mers and 200 9mers, resulting in 2400 possible combinations at a single fragment

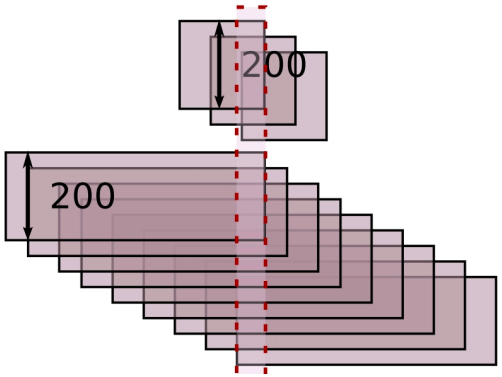
SETTLERSFCATANISTHEGREATESTGAMEEVER



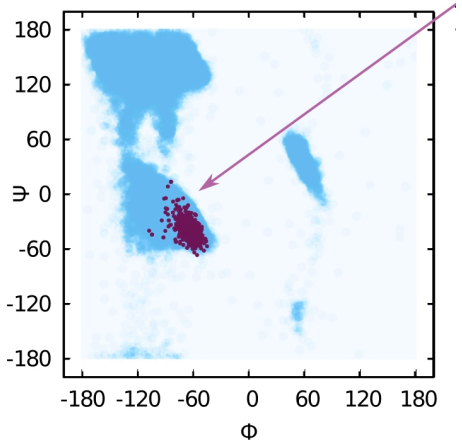
Gront D, Kulp DW, Vernon RM, Strauss CE, Baker D. Generalized fragment picking in Rosetta: design, protocols and applications. Plos one. 2011 ;6(8):e23294. DOI: 10.1371/journal.pone.0023294.

Fragment sampling

- At every residue in a sequence, generate 3mer and 9mer fragments defined by Φ/Ψ from known protein structures
- Typically have 200 3mers and 200 9mers, resulting in 2400 possible combinations at a single fragment



S E T T L E R S F C A T A N I S T H E G R E A T E S T G A M E E V E R

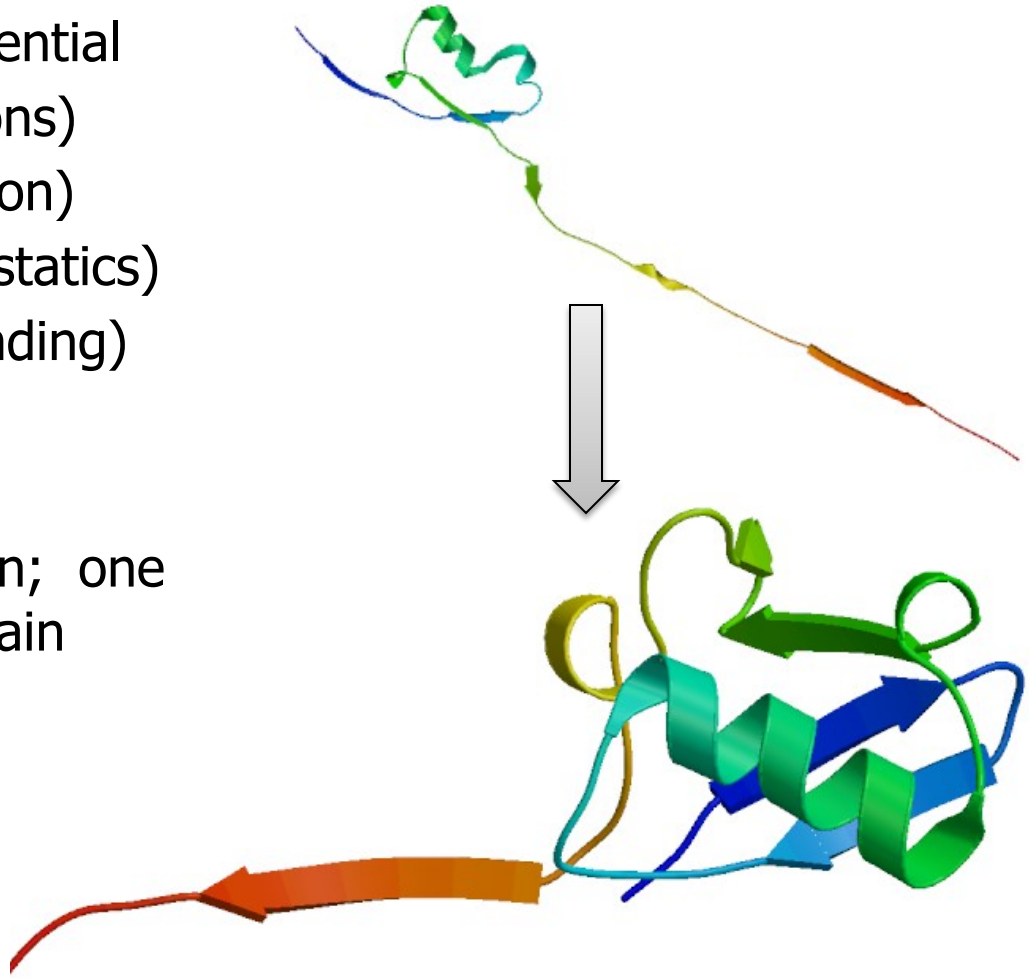


Gront D, Kulp DW, Vernon RM, Strauss CE, Baker D. Generalized fragment picking in Rosetta: design, protocols and applications. Plos one. 2011 ;6(8):e23294. DOI: 10.1371/journal.pone.0023294.

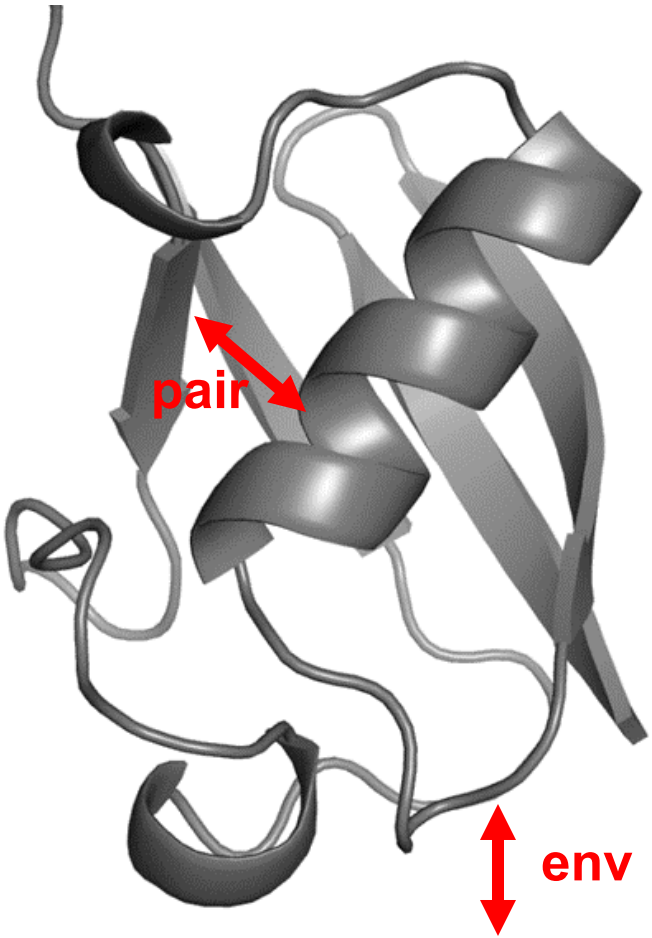
Non-local Interactions Govern Protein Folding Process

- Statistically-derived scoring potential
  - Steric overlap (vdw interactions)
  - Residue environment (solvation)
  - Pairwise interactions (electrostatics)
  - Strand pairing (hydrogen bonding)
  - Compactness (solvation)

- Simplified protein representation; one centroid per amino acid side chain

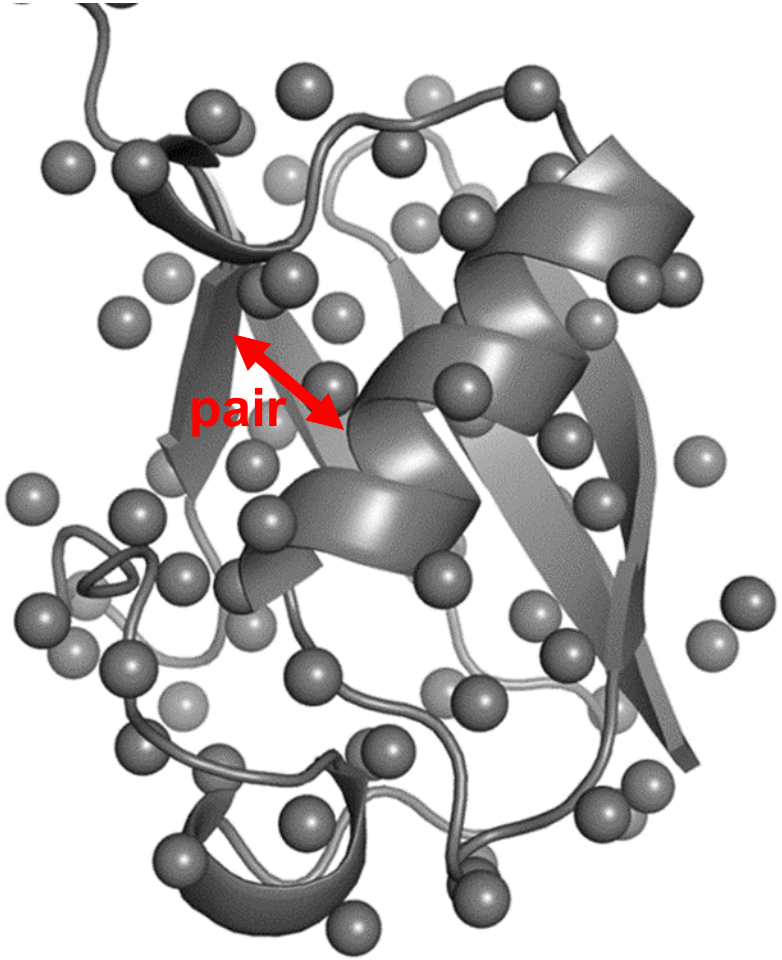


Scoring Function Terms



- Explicit: Interaction within the protein
  - Covalent interactions: disulfides
  - Electrostatic interactions
  - Hydrogen Bonds
  - Van der Waals interactions
- Implicit: Interactions with environment
  - Solvation
  - Radius of Gyration

Low Resolution Scoring Terms: Explicit Pairwise Interactions



Residue pair interactions (electrostatics, disulfides)

$$\sum_i \sum_{j>i} -\ln\left[\frac{P(\mathrm{aa}_i, \mathrm{aa}_j | s_{ij} d_{ij})}{P(\mathrm{aa}_i | s_{ij} d_{ij}) P(\mathrm{aa}_i | s_{ij} d_{ij})}\right]$$

Strand pairing (hydrogen bonding)

$$\text{SchemeA} : \mathrm{SS}_{\phi,\theta} + \mathrm{SS}_{hb} + \mathrm{SS}_d$$

$$\text{SchemeB} : \mathrm{SS}_{\phi,\theta} + \mathrm{SS}_{hb} + \mathrm{SS}_{d\sigma}$$

where

$$\mathrm{SS}_{\phi,\theta} = \sum_m \sum_{n>m} -\ln\left[P(\phi_{mn}, \theta_{mn} | d_{mn}, \mathrm{sp}_{mn}, s_{mn})\right]$$

$$\mathrm{SS}_{hb} = \sum_m \sum_{n>m} -\ln\left[P(\mathrm{hb}_{mn} | d_{mn}, s_{mn})\right]$$

$$\mathrm{SS}_d = \sum_m \sum_{n>m} -\ln\left[P(d_{mn} | s_{mn})\right]$$

$$\mathrm{SS}_{d\sigma} = \sum_m \sum_{n>m} -\ln\left[\mathrm{P}(d_{mn}\sigma_{mn} | \rho_m, \rho_n)\right]$$

Strand arrangement into sheets

$$-\ln\left[P(n_{\mathrm{sheets}} n_{\mathrm{lonestrands}} | n_{\mathrm{strands}})\right]$$

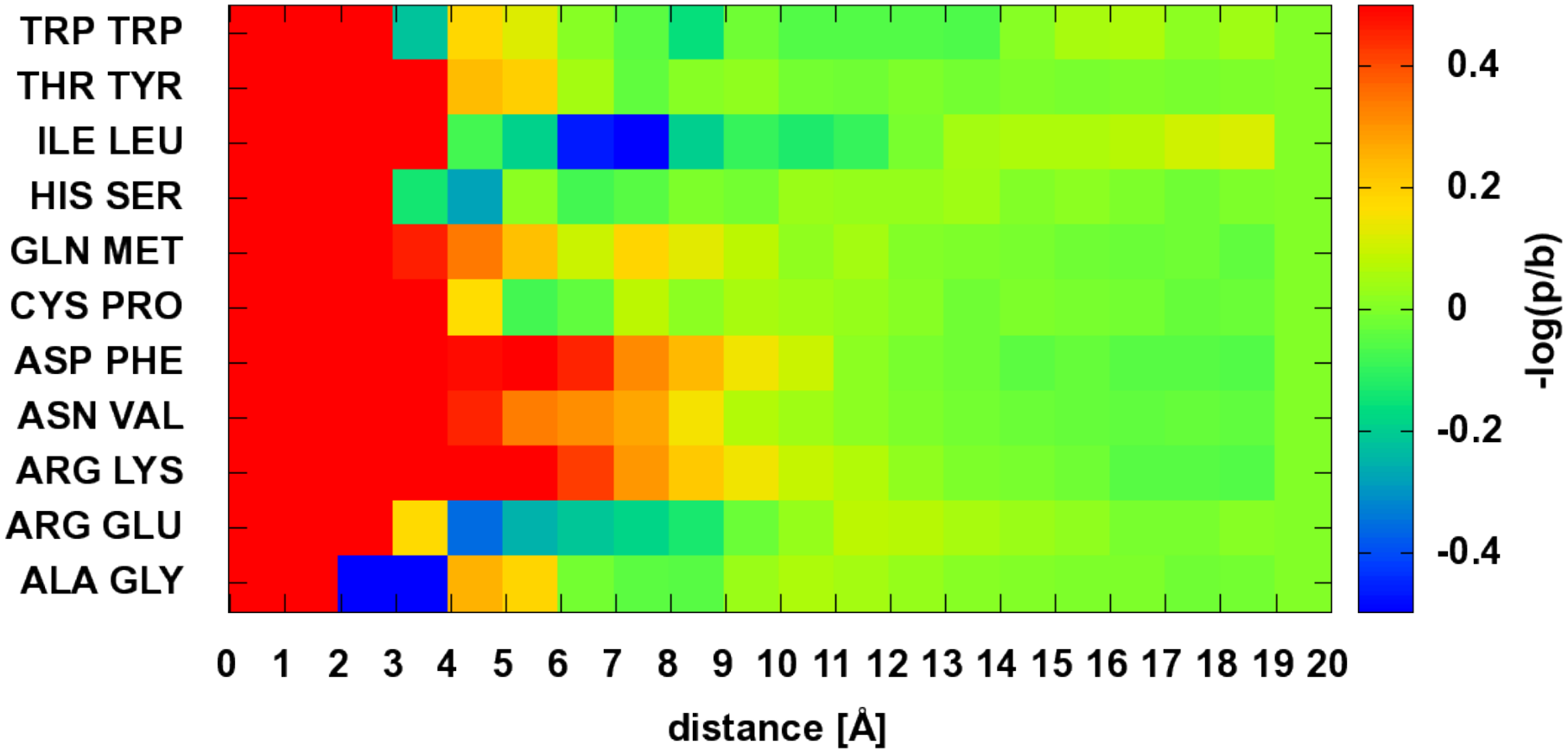
Helix–strand packing

$$\sum_m \sum_n -\ln\left[P(\phi_{mn}, \psi_{mn} | \mathrm{sp}_{mn} d_{mn})\right]$$

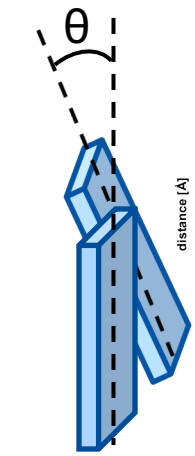
Steric repulsion

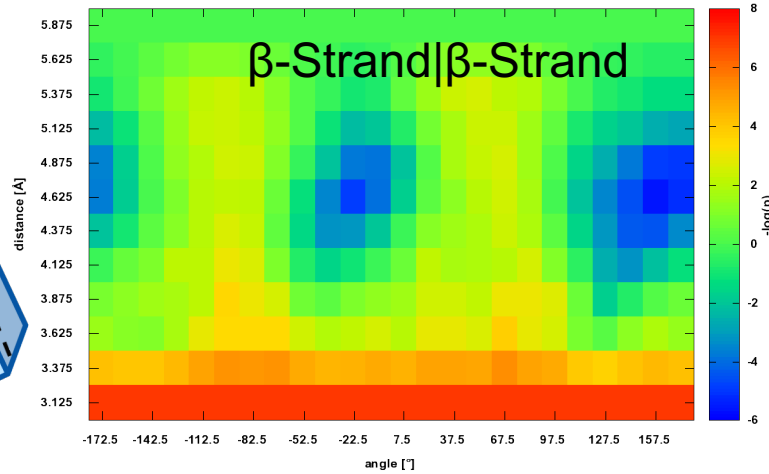
$$\sum_i \sum_{j>i} \frac{\left(r_{ij}^2 - d_{ij}^2\right)^2}{r_{ij}}; \; d_{ij} < r_{ij}$$

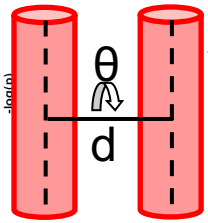
How likely are two amino acids in this distance? $P(aa_i, aa_j | d_{ij})$

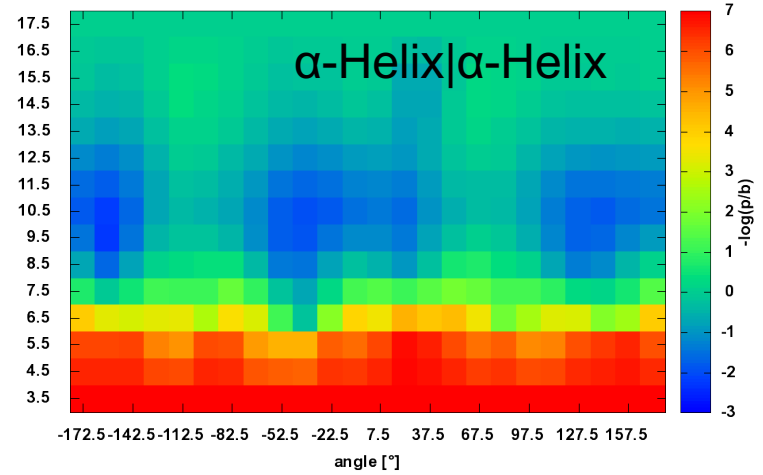


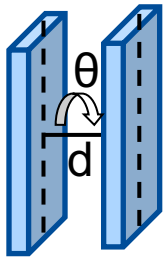
How likely are two SSEs in a certain arrangement? $P(d_{ij}, \theta_{ij} | SSE_{i,j})$

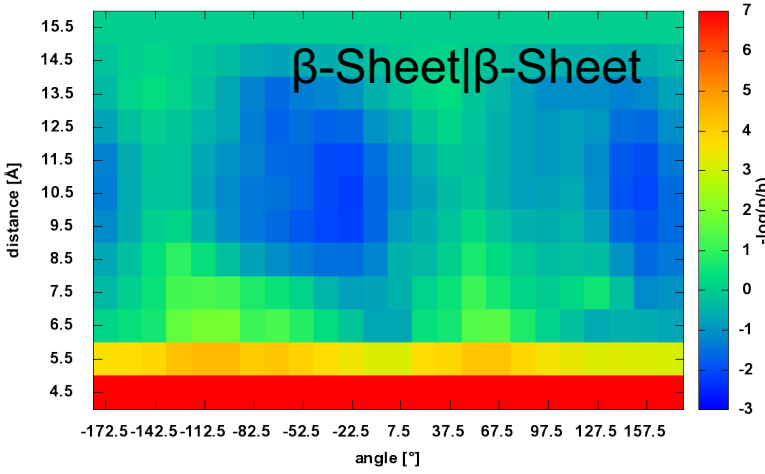


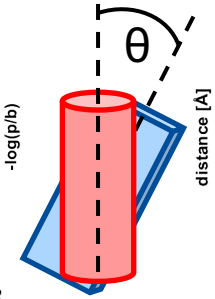


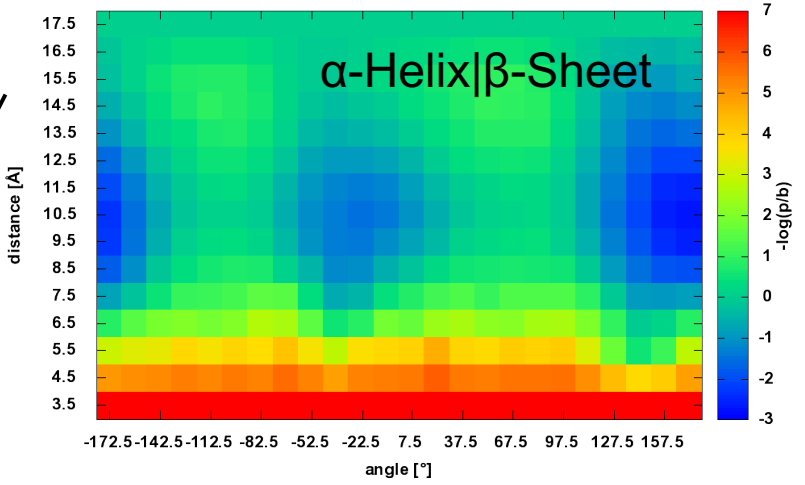




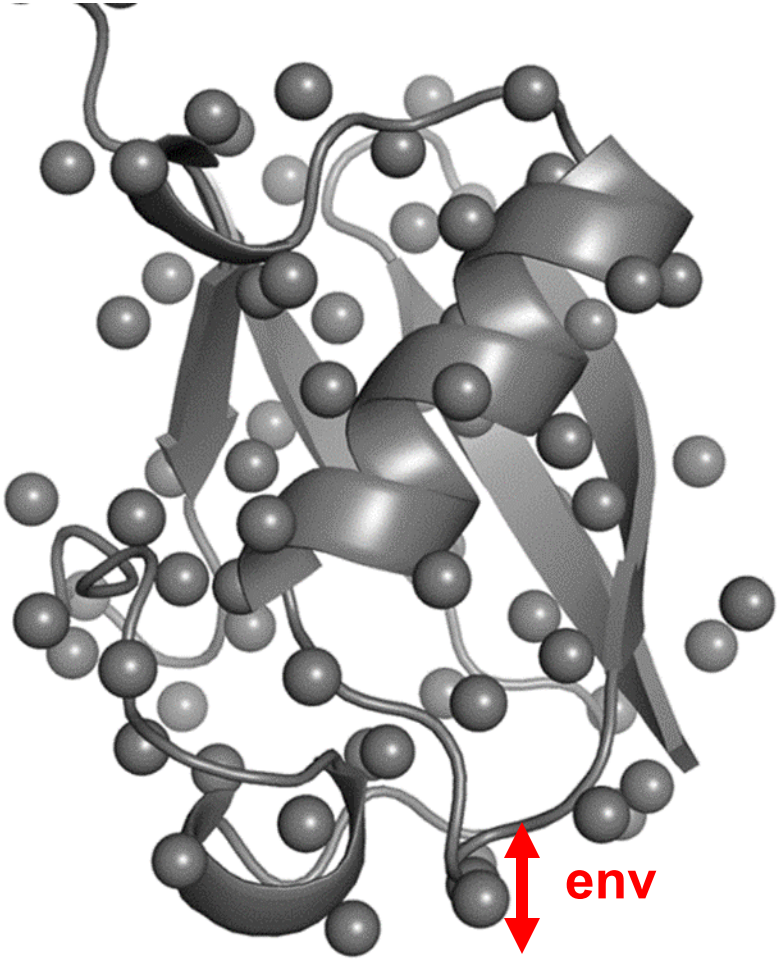






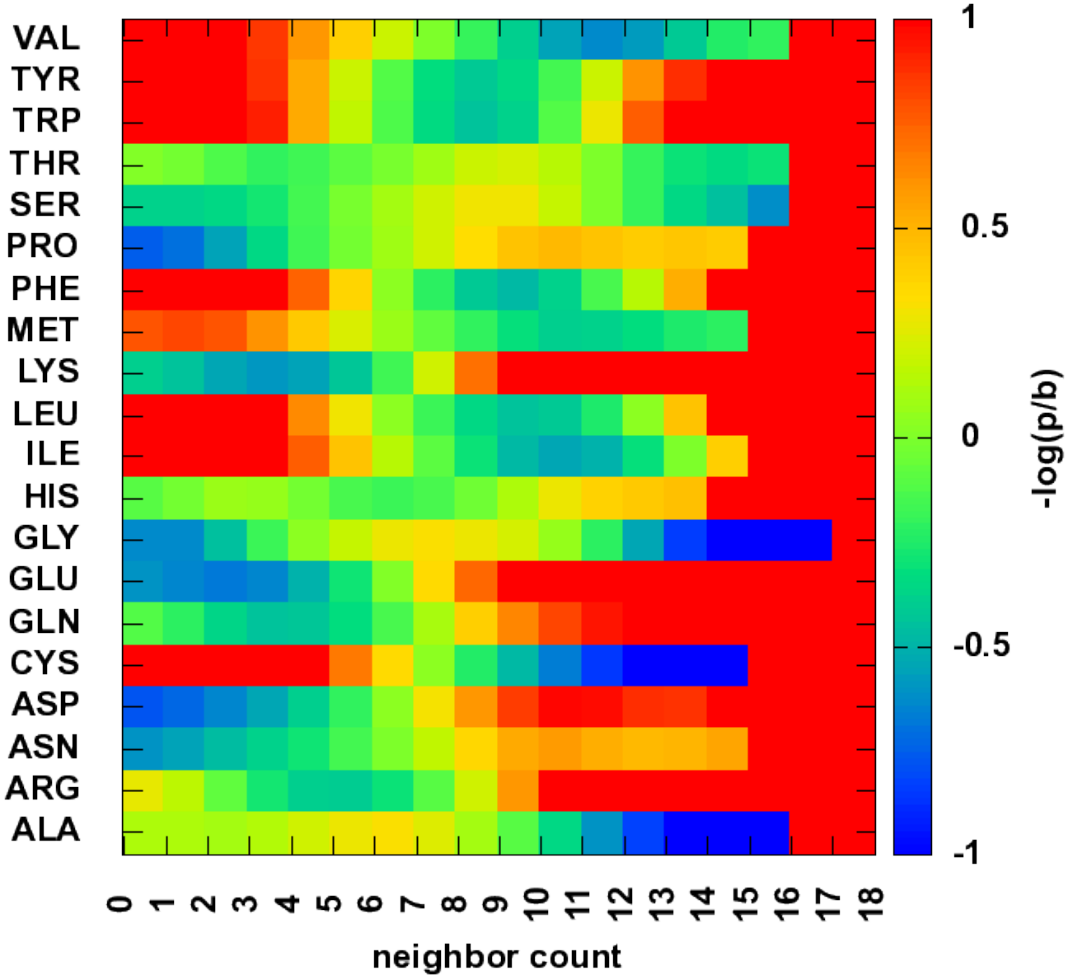


Low Resolution Scoring Terms: Implicit Interactions with Solvent



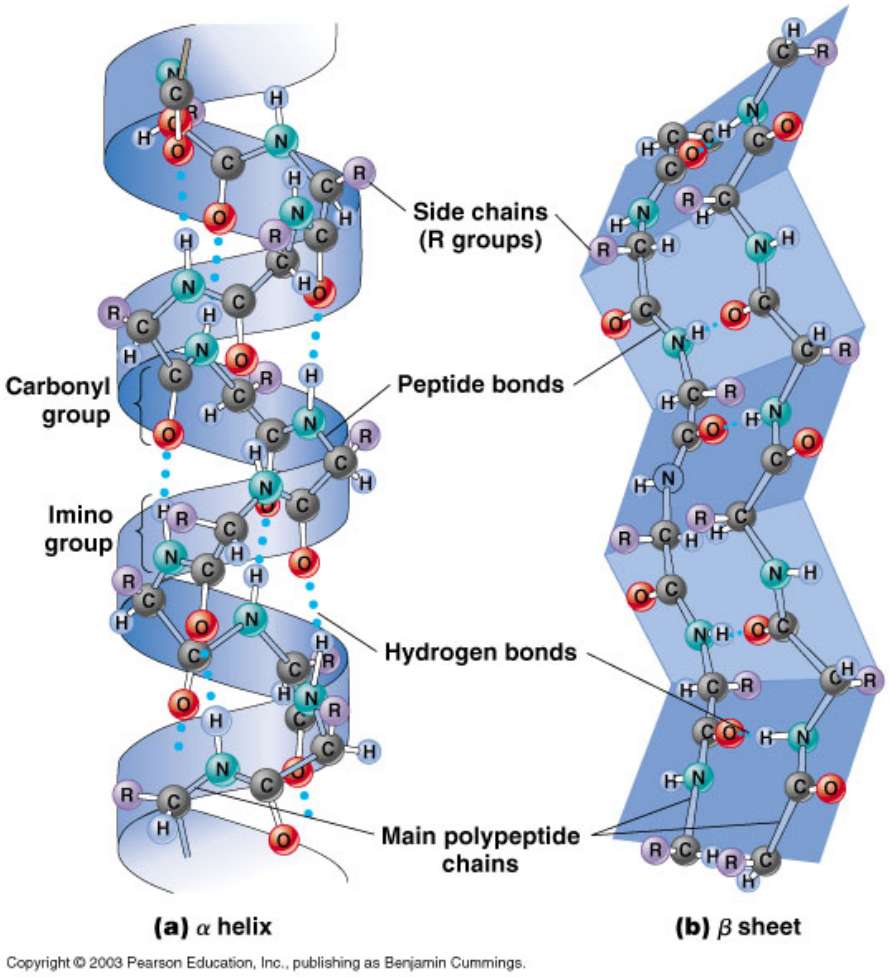
Residue environment (solvation)	$\sum_i -\ln\left[P(\mathrm{aa}_i \mid \mathrm{nb}_i)\right]$
Radius of gyration (vdw attraction; solvation)	$\sqrt{\langle d_{ij}^2 \rangle}$
$C\beta$ density (solvation; correction for excluded volume effect introduced by simulation)	$\sum_i \sum_{sh} -\ln\left[\frac{P_{\mathrm{compact}}(\mathrm{nb}_{i,sh})}{P_{\mathrm{random}}(\mathrm{nb}_{i,sh})}\right]$

How likely is that amino acid in this exposure state? $P(aa_i \mid env)$



Secondary Structure: Build from Backbone Hydrogen Bonds

- *α-Helix:*
  - *Periodicity = 3.6*
  - *Rise = 1.5Å*
  - *Pitch = 5.4Å*
- *β-Sheet:*
  - *Periodicity = 2*
  - *Translation = 3.4Å*
  - *Distance = 5.4Å*



Secondary Structure Propensities of Amino Acids in Numbers

Table 1 Amino acid parameter sets

Name	Ξ[a]	α[b]	υ_v [c]	π[d]	I[e]	α[f]	β[g]
ALA	1.28	0.05	1.00	0.31	6.11	0.42	0.23
GLY	0.00	0.00	0.00	0.00	6.07	0.13	0.15
VAL	3.67	0.14	3.00	1.22	6.02	0.27	0.49
LEU	2.59	0.19	4.00	1.70	6.04	0.39	0.31
ILE	4.19	0.19	4.00	1.80	6.04	0.30	0.45
PHE	2.94	0.29	5.89	1.79	5.67	0.30	0.38
TYR	2.94	0.30	6.47	0.96	5.66	0.25	0.41
TRP	3.21	0.41	8.08	2.25	5.94	0.32	0.42
THR	3.03	0.11	2.60	0.26	5.60	0.21	0.36
SER	1.31	0.06	1.60	-0.04	5.70	0.20	0.28
ARG	2.34	0.29	6.13	-1.01	10.74	0.36	0.25
LYS	1.89	0.22	4.77	-0.99	9.99	0.32	0.27
HIS	2.99	0.23	4.66	0.13	7.69	0.27	0.30
ASP	1.60	0.11	2.78	-0.77	2.95	0.25	0.20
GLU	1.56	0.15	3.78	-0.64	3.09	0.42	0.21
ASN	1.60	0.13	2.95	-0.60	6.52	0.21	0.22
GLN	1.56	0.18	3.95	-0.22	5.65	0.36	0.25
MET	2.35	0.22	4.43	1.23	5.71	0.38	0.32
PRO	2.67	0.00	2.72	0.72	6.80	0.13	0.34
CYS	1.77	0.13	2.43	1.54	6.35	0.17	0.41

[a] Steric parameter (graph shape index)
[b] Polarizability
[c] Volume (normalized van der Waals volume)
[d] Hydrophobicity
[e] Isoelectric point
[f] Helix probability
[g] Sheet probability

Secondary Structure Propensities of Amino Acids in Numbers

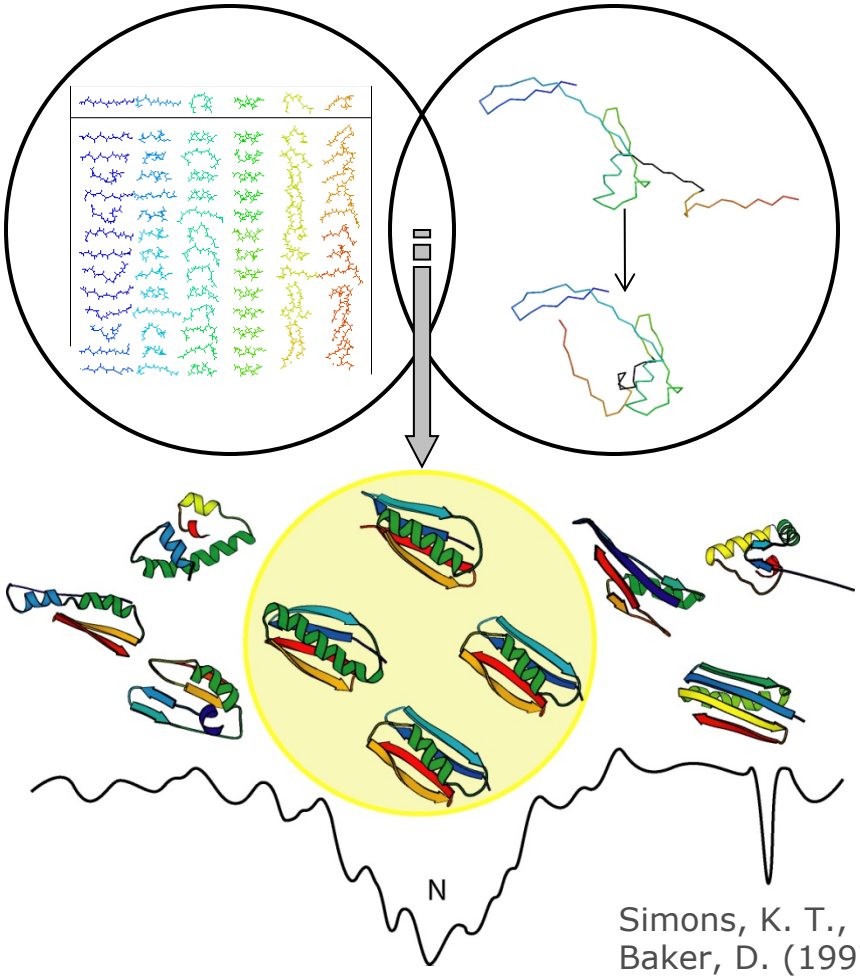
Table 1 Amino acid parameter sets

Name	Ξ[a]	α[b]	υ_v [c]	π[d]	I[e]	α[f]	β[g]
ALA	1.28	0.05	1.00	0.31	6.11	0.42	0.23
GLY	0.00	0.00	0.00	0.00	6.07	0.13	0.15
VAL	3.67	0.14	3.00	1.22	6.02	0.27	0.49
LEU	2.59	0.19	4.00	1.70	6.04	0.39	0.31
ILE	4.19	0.19	4.00	1.80	6.04	0.30	0.45
PHE	2.94	0.29	5.89	1.79	5.67	0.30	0.38
TYR	2.94	0.30	6.47	0.96	5.66	0.25	0.41
TRP	3.21	0.41	8.08	2.25	5.94	0.32	0.42
THR	3.03	0.11	2.60	0.26	5.60	0.21	0.36
SER	1.31	0.06	1.60	-0.04	5.70	0.20	0.28
ARG	2.34	0.29	6.13	-1.01	10.74	0.36	0.25
LYS	1.89	0.22	4.77	-0.99	9.99	0.32	0.27
HIS	2.99	0.23	4.66	0.13	7.69	0.27	0.30
ASP	1.60	0.11	2.78	-0.77	2.95	0.25	0.20
GLU	1.56	0.15	3.78	-0.64	3.09	0.42	0.21
ASN	1.60	0.13	2.95	-0.60	6.52	0.21	0.22
GLN	1.56	0.18	3.95	-0.22	5.65	0.36	0.25
MET	2.35	0.22	4.43	1.23	5.71	0.38	0.32
PRO	2.67	0.00	2.72	0.72	6.80	0.13	0.34
CYS	1.77	0.13	2.43	1.54	6.35	0.17	0.41

[a] Steric parameter (graph shape index)
[b] Polarizability
[c] Volume (normalized van der Waals volume)
[d] Hydrophobicity
[e] Isoelectric point
[f] Helix probability
[g] Sheet probability

Sampling and Scoring for Protein Folding Simulation

- Local Sequence Bias
  - Approximate local interactions using the distribution of conformations seen for similar sequences in known protein structures
- Monte Carlo simulations
  - Select broadest minima using cluster analysis

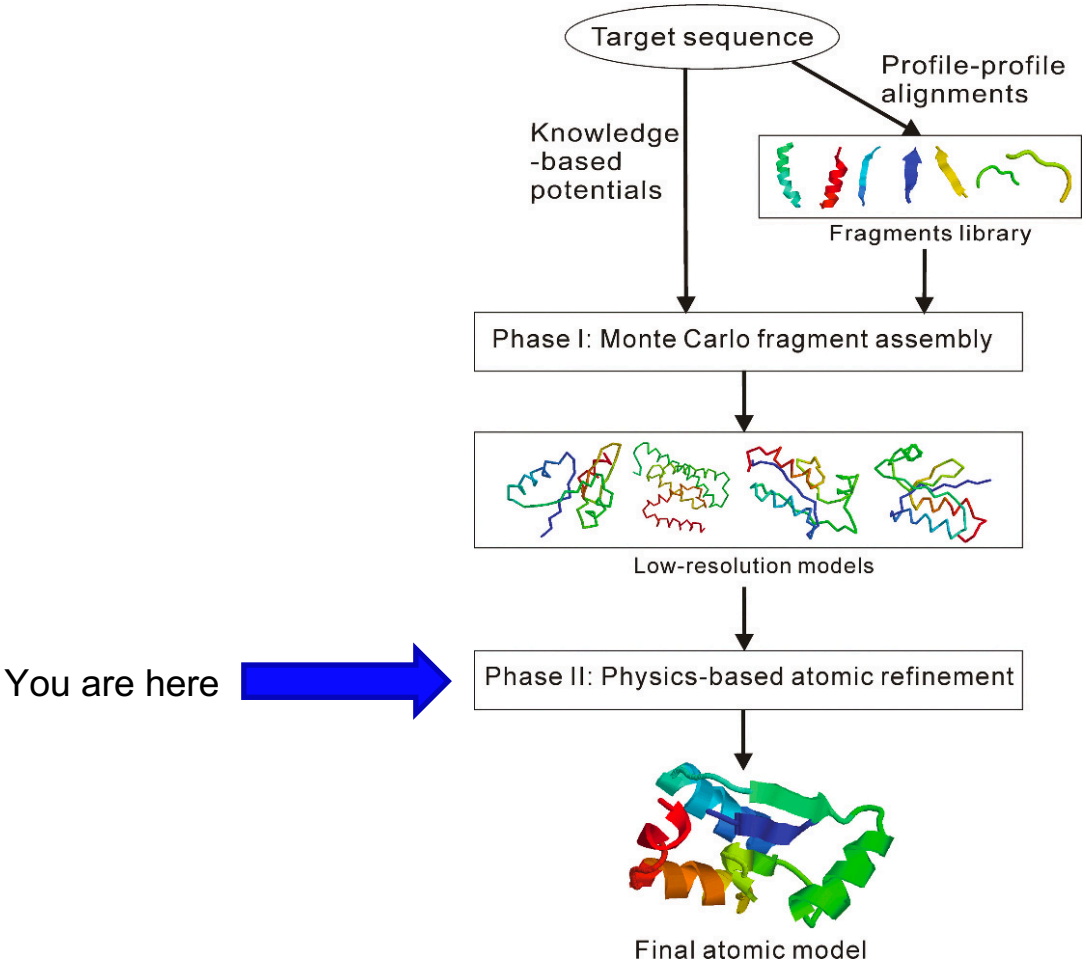


- Energy evaluation of non-local interactions using knowledge-based energy function
  - Steric overlap
  - Residue environment
  - Pair wise interactions
  - Strand pairing
  - Compactness
  - Secondary Structure Packing

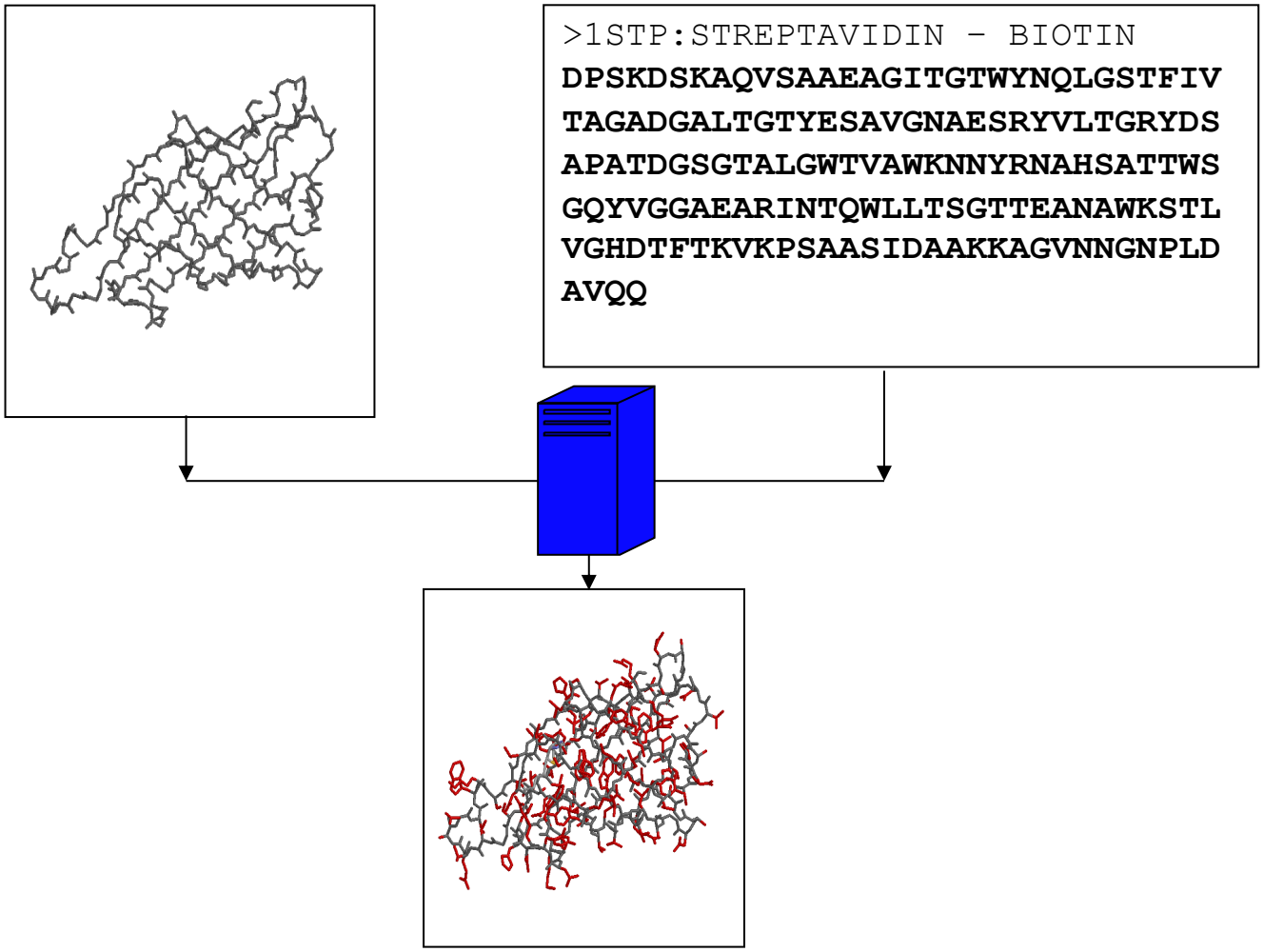
Simons, K. T., Kooperberg, C., Huang, E. and Baker, D. (1997) *J. Mol. Biol.*, 268, 209-225.

Folding 76AA Protein Ubiquitin *de novo* to 1.93Å

General overview of de novo structure prediction



Given Protein Backbone and Sequence, Model all Protein Atoms



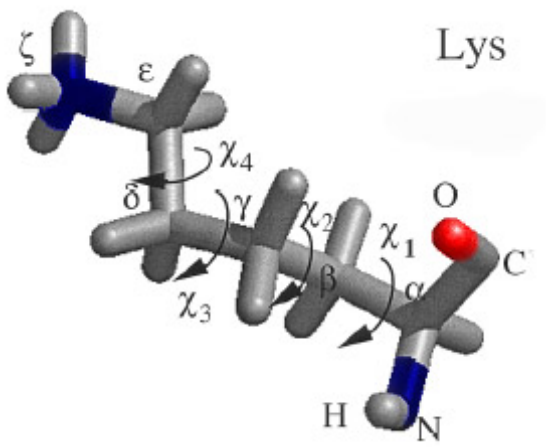
Why model side chain conformations?

- Necessary for constructing a full 3D model of proteins (homology modeling, fold recognition and ab-initio methods).
- The amino acid side chains determine the global fold of the protein. Identity and structure of individual side chains determine the protein stability and its interactions with other molecules.
- Complete the structural information not resolved by experimental procedures.
- Incorporate protein flexibility in docking algorithms

Components and complexity of the problem

- The side chain modeling problem can be divided into searching procedure and scoring function.
  - The searching procedure should sample the search space (in our case usually the torsion angle space) and create conformations.
  - The scoring function evaluates each conformation created by the searching procedure. The evaluation scores are used to rank the conformations and pick the best one to be the final model.
- Consider a short protein with 50 amino acids. Each side chain has on average two dihedral angles (χ angles). Assuming that we will sample every $40°$ in the dihedral angle space, the number of conformations to search becomes $N = (360°/40°)^{50\times2} \approx 10^{95}$!

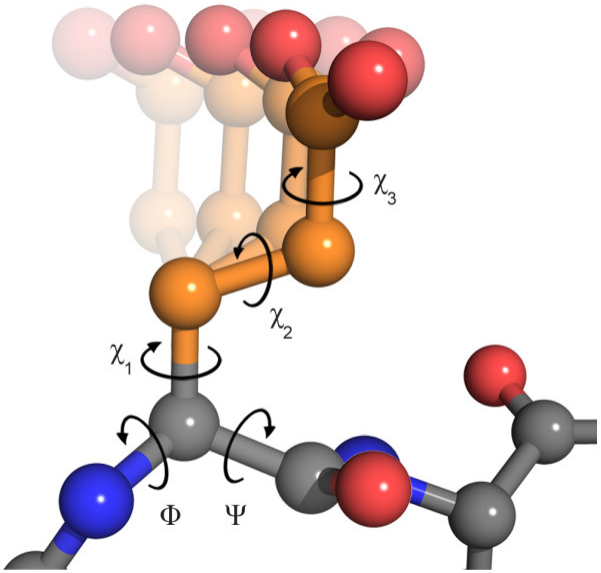
Algorithms are needed that find good solutions by screening only parts of the search space are needed



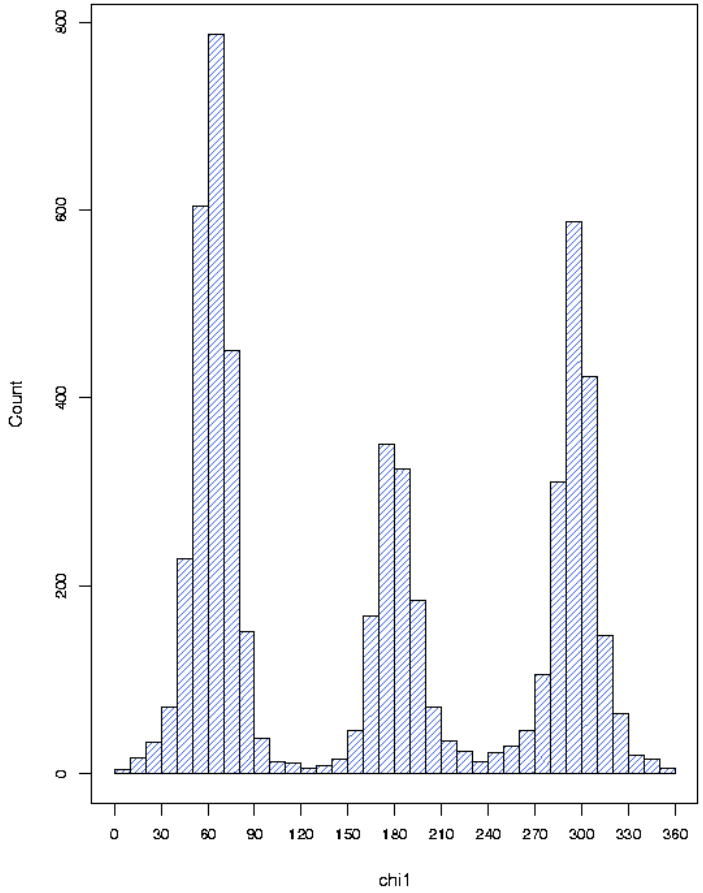
Rotamer terminology

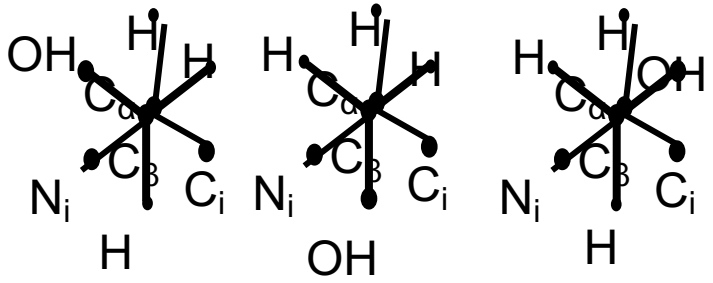
- *Rotamer:* short for "rotational isomer", is a single side-chain conformation represented as a set of values, one for each dihedral-angle degree of freedom.
- *Rotamer library:* collection of rotamers for each residue type with frequency information
- Rotamer libraries can be *backbone-independent*, *secondary-structure-dependent,* or *backbone-dependent.*
  - *Backbone-independent* rotamer libraries make no reference to backbone conformation
  - *Backbone-dependent* rotamer libraries dependent on the local backbone conformation as defined by the backbone dihedral angles ϕ and ψ, regardless of secondary structure.
  - *Secondary-structure-dependent* libraries based on α-helix, β-sheet, or coil secondary structures.

Computational prediction of amino acid side chain conformations



Not all rotamers are created equal



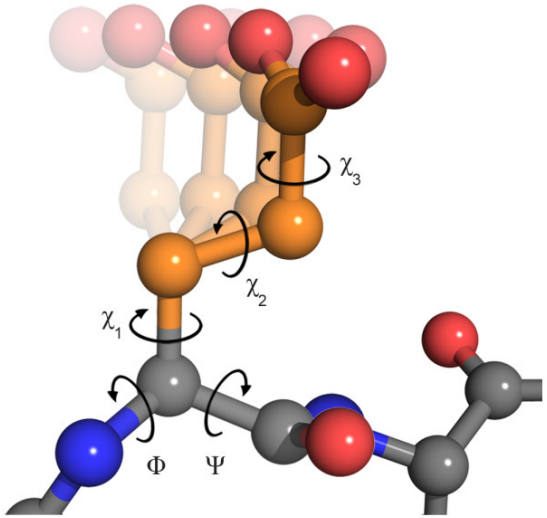


http://www.biomedsearch.com/nih/Beyond-rotamers-generative-probabilistic-model/20525384.html

Rotamer libraries in the 21st century

Authors	Year	Type	Number of proteins	Resolution
Chandrasekaran [2]	1970	BBIND	3	
Janin [4]	1978	BBIND, SSDEP	19	2.5
Bhat [3]	1979	BBIND	23	
James and Sielecki [5]	1983	BBIND	5	1.8, R-factor < 0.15
Benedetti [6]	1983	BBIND	238 peptides	R-factor < 0.10
Ponder & Richards[7]	1987	BBIND	19	2.0
McGregor[8]	1987	SSDEP	61	2.0
Tuffery[9]	1991	BBIND	53	2.0
Dunbrack & Karplus[10]	1993	BBIND, BBDEP	132	2.0
Schrauber et al.[11]	1993	BBIND, SSDEP	70	2.0
Kono & Doi [12]	1996	BBIND	103	
Lasters, DeMaeyer[13]	1995	BBIND	19	2.0
Dunbrack & Cohen[14]	1997-2002	BBIND, BBDEP	850*	1.7
Lovell et al. [15••]	2000	BBIND, SSDEP	240	1.7

* Latest update, May 2002.



"Complete" Rotamer Library introduced in 1987

J. W. Ponder and F. M. Richards; "Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes"; *J Mol Biol*; **1987**; Vol. 193 (4): p. 775-91.

res	rotamer	chi1	chi2		Frequency
ARG	-t	-67.6	176.9		0.463
ARG	tt	-174.1	-178.6		0.232
ARG	+t	80.0	175.6		0.098
ARG	--	-67.0	-71.7		0.085
ARG	t+	178.2	69.5		0.049
ARG	++	57.1	82.8		0.024
ARG	-+	-76.9	54.2		0.024
ASN	--	-68.3	-36.8		0.303
ASN	t0	-177.1	1.3		0.213
ASN	-+	-67.2	128.8		0.131
ASN	+0	63.9	-6.8		0.115
ASN	tt	-174.9	-156.8		0.115
ASN	++	63.6	53.8		0.066
ASP	-0	-68.3	-25.7		0.477
ASP	t0	-169.1	3.9		0.336
ASP	+0	63.7	2.4		0.159
CYS	-	-65.2			0.606
CYS	t	-179.6			0.245
CYS	+	63.5			0.138
GLN	-t	-66.7	-178.5		0.367
GLN	tt	-174.6	-177.7		0.211
GLN	--0	-58.7	-63.8	-46.3	0.144
GLN	t+0	-179.4	67.3	26.8	0.089
GLN	+t	70.8	-165.6		0.067
GLN	--t	-51.3	-90.4	165.0	0.044
GLN	t+t	167.5	70.9	174.2	0.022
GLU	-t0	-69.6	-177.2	-11.4	0.272
GLU	tt0	-176.2	175.4	-6.7	0.259
GLU	--0	-64.6	-69.1	-33.4	0.111
GLU	-+0	-55.6	77.0	25.3	0.086
GLU	+t0	69.8	-179.0	6.6	0.086
GLU	t+0	-173.6	70.6	14.0	0.062
GLU	+-0	63.0	-80.4	16.3	0.049
HIS	--	-62.8	-74.3		0.341
HIS	t+	-175.2	-87.7		0.250
HIS	-+	-69.8	96.1		0.159
HIS	+-	67.9	-80.5		0.136
HIS	t-	-177.3	100.5		0.091
HIS	++	48.0	85.9		0.023
ILE	-t	-60.9	168.7		0.452
ILE	--	-59.6	-64.1		0.183
ILE	+t	61.7	163.8		0.161
ILE	tt	-166.6	166.0		0.129
ILE	t+	-174.8	72.1		0.032
LEU	-t	-64.9	176.0		0.639
LEU	t+	-176.4	63.1		0.245
LEU	tt	-165.3	168.2		0.048
LEU	++	44.3	60.4		0.020
LYS	-t	-68.9	-178.4		0.409
LYS	tt	-172.1	175.3		0.236
LYS	--	-58.1	-74.9		0.164
LYS	t+	173.4	83.4		0.082
LYS	+t	71.5	-174.3		0.036
LYS	t-	-175.8	-63.9		0.027
LYS	-+	-104.0	74.6		0.018
MET	---	-64.5	-68.5	-75.6	0.375
MET	-t	-78.3	-174.7		0.250
MET	tt	178.9	179.0		0.188
PHE	-9	-66.3	94.3		0.463
PHE	t9	-179.2	78.9		0.250
PHE	+9	66.0	90.7		0.213
PHE	-0	-71.9	-0.4		0.063
PRO	+	26.9	-29.4		0.394
PRO	-	-21.8	31.2		0.349
PRO	0	0.3	-0.8		0.234
SER	+	64.7			0.480
SER	-	-69.7			0.286
SER	t	-176.1			0.235
THR	+	62.7			0.479
THR	-	-59.7			0.450
THR	t	-169.5			0.047
TRP	-+	-70.4	100.5		0.379
TRP	+-	64.8	-88.9		0.207
TRP	t-	-177.3	-95.1		0.138
TRP	t+	-179.5	87.5		0.103
TRP	--	-73.3	-87.7		0.069
TRP	++	62.2	112.5		0.034
TYR	-9	-66.5	96.6		0.486
TYR	t9	-179.7	71.9		0.327
TYR	+9	63.3	89.1		0.150
TYR	-0	-67.2	-1.0		0.037
VAL	t	173.5			0.571
VAL	-	-63.4			0.262
VAL	+	69.3			0.054

Backbone independent rotamer library

- Dunbrack & Cohen, 1997

R.L. Dunbrack Jr, F.E. Cohen,. "Bayesian statistical analysis of protein sidechain rotamer preferences" Prot. Science, 6 (1997), pp. 1661-1681

	No. χ_1	No.	p	σ	p|χ_1	σ	χ_1	σ	χ_2	σ
SER 1 0 0 0	4125	4125	46.61	0.43	100.00	0.00	65.0	10.7		
SER 2 0 0 0	2059	2059	23.27	0.37	100.00	0.00	179.6	11.7		
SER 3 0 0 0	2665	2665	30.12	0.40	100.00	0.00	-64.2	11.0		
THR 1 0 0 0	4165	4165	48.38	0.44	100.00	0.00	61.1	8.8		
THR 2 0 0 0	686	686	7.98	0.24	100.00	0.00	-173.3	12.8		
THR 3 0 0 0	3757	3757	43.64	0.44	100.00	0.00	-60.4	8.2		
TRP 1 1 0 0	337	215	9.56	0.51	63.62	2.13	61.7	9.7	-90.9	9.4
TRP 1 2 0 0	337	16	0.74	0.15	4.92	0.96	65.6	7.5	-16.7	40.9
TRP 1 3 0 0	337	106	4.73	0.36	31.47	2.06	59.4	12.0	88.2	10.1
TRP 2 1 0 0	786	359	15.94	0.63	45.64	1.45	-178.4	12.5	-104.1	15.1
TRP 2 2 0 0	786	139	6.19	0.41	17.72	1.11	-175.5	12.4	18.2	31.0
TRP 2 3 0 0	786	288	12.80	0.57	36.63	1.40	179.8	8.8	84.8	9.7
TRP 3 1 0 0	1127	106	4.73	0.36	9.45	0.71	-70.4	13.2	-91.4	15.4
TRP 3 2 0 0	1127	303	13.46	0.59	26.90	1.08	-68.5	9.9	-2.5	26.8
TRP 3 3 0 0	1127	718	31.86	0.80	63.66	1.17	-67.4	11.3	99.8	16.4

Backbone dependent rotamer library

- Dunbrack & Cohen, 2002

R.L. Dunbrack Jr, F.E. Cohen,. “Bayesian statistical analysis of protein sidechain rotamer preferences” Prot. Science, 6 (1997), pp. 1661-1681

	ϕ	φ			p	χ_1	χ_2	χ_3	χ_4
ARG	-180	-100	0	2 2 3 1	0.025418	-173.4	176.6	-62.2	107.1
ARG	-180	-100	0	1 2 2 1	0.024425	57.1	179.0	178.6	87.1
ARG	-180	-100	0	3 2 2 3	0.024220	-69.4	-178.8	-175.1	-85.2
ARG	-180	-100	0	1 2 2 3	0.022793	57.8	-172.1	-178.6	-84.3
ARG	-180	-100	0	3 2 3 3	0.022259	-70.4	-171.4	-65.5	-86.7
ARG	-180	-100	0	3 2 2 1	0.021848	-69.2	179.3	178.7	88.0

Rosetta currently uses an updated smoothed potential of the BB-dependent rotamer library (2011)

M.V. Shapovalov, R. L. Dunbrack Jr. “A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions.” Structure. 2011 Jun 8; 19(6): 844–858.

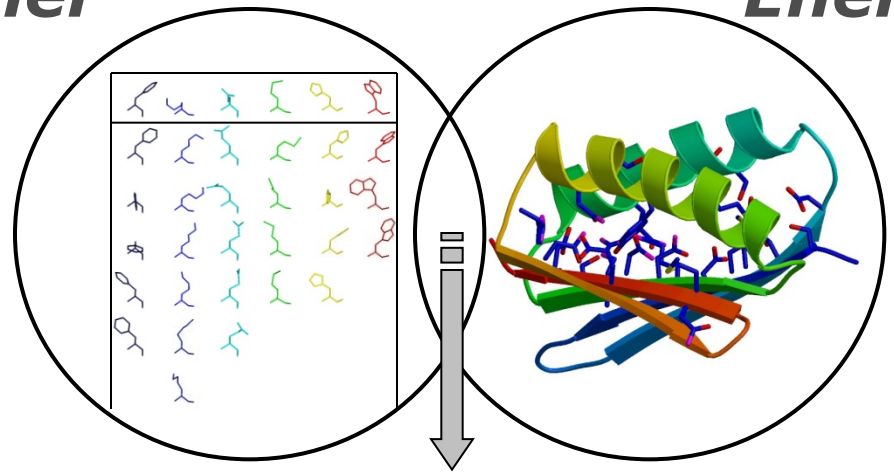
What do rotamer libraries provide?

- Rotamer libraries reduce significantly the number of conformations that need to be evaluated during the search.
- This is done with almost no risk of missing the real conformations. Even small libraries of about 100-150 rotamers cover about 96-97% of the conformations actually found in protein structures.
- The probabilities of each rotamer in the library can be applied to estimate the potential energy due to interactions within the side chain and with the local backbone atoms, using the Boltzmann relation: $E \cong -ln(P)$

Sampling and Scoring for Side Chain Repacking and Design

Local Rotamer Bias

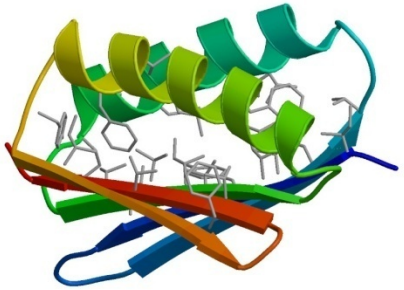
Approximate interactions within sidechain using the distribution of sidechain confor-mations (rotamers) seen in known protein structures



Energy function

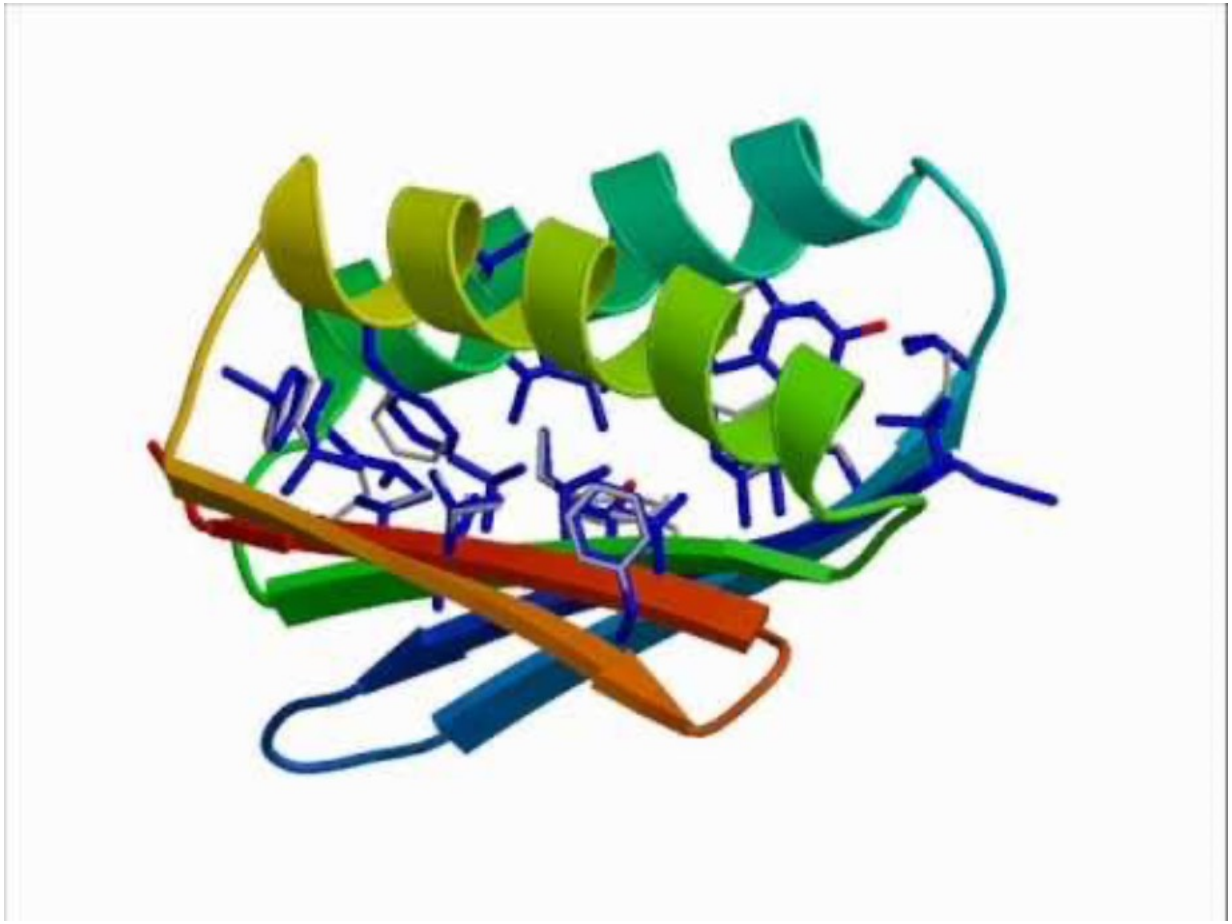
Statistically derived potential function
- ***VDW interaction***
- ***solvation***
- ***hydrogen bonding potential***
- ***pair wise interactions***
- ***rotamer probability***

Simulated Annealing Monte Carlo energy minimization



Dahiyat, B. I. and Mayo, S. L. (1997) *Science,* 278, 82-7
Dunbrack, R. L., Jr. and Karplus, M. (1993) *J Mol Biol,* 230, 543-74.
Kuhlman, B., et. al. (2003) *Science,* 302, 1364-1368.

Repacking 76AA Protein Ubiquitin *de novo* to 1.93Å



Rosetta full-atom energy function is a combination of different kinds of scoring methods

1. VDW: 6-12 Lennard-Jones potential
2. Lazaridis-Karplus implicit solvation model (penalizes buried polar atoms)
3. Orientation-dependent hydrogen bonding
4. Electrostatic pair potential
5. Knowledge-based rotamer preferences
6. Knowledge-based backbone angle preferences
7. Amino acid reference energies for unfolded state (used in protein design only)
8. Disulfide terms

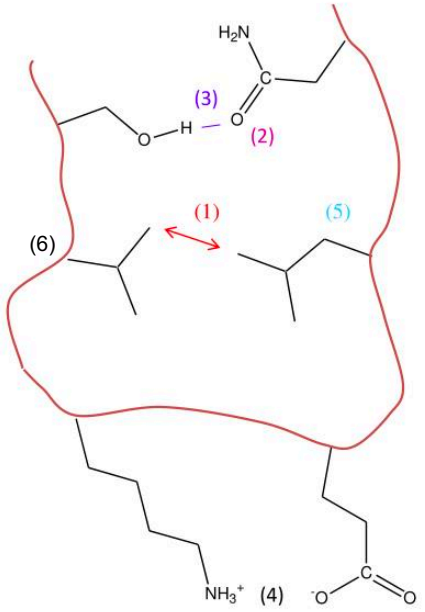
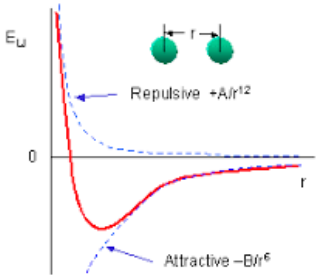


Image adapted from Dr. Brian Kuhlman, UNC

Physics-based scoring similar to molecular mechanics

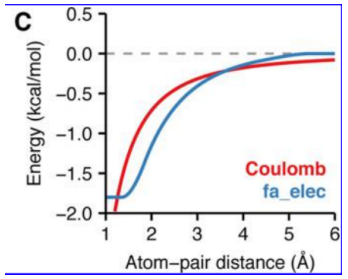
Van der Waals:
Lennard Jones 6-12 potential

$$E_{\text{vdw}}(i,j) = \varepsilon_{i,j}\left[\left(\frac{\sigma_{i,j}}{d_{i,j}}\right)^{12} - 2\left(\frac{\sigma_{i,j}}{d_{i,j}}\right)^{6}\right]$$



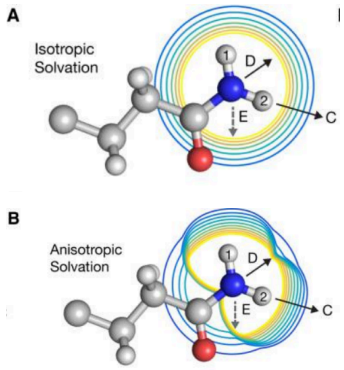
Electrostatics:
Coulomb's Law

$$E_{\text{Coulomb}}(i,j) = \frac{C_0 q_i q_j}{\epsilon}\frac{1}{d_{i,j}}$$



Solvation:
Lazaridis-Karplus
implicit solvent model

$$f_{\text{desolv}} = -V_j\frac{\Delta G_i^{\text{free}}}{2\pi^{3/2}\lambda_i{\sigma_i}^2}\exp\left[-\left(\frac{d_{i,j}-\sigma_{i,j}}{\lambda_i}\right)^2\right]$$



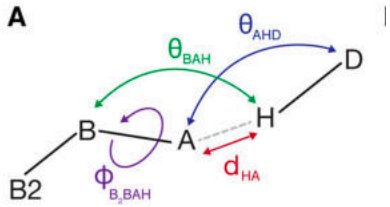
f_{desolv} represents the energy to desolvate atom i when approached by atom j

($d_{i,j}$ is atom-pair distance, ΔG_i^{free} is the experimentally-determined vapor-to-water transfer free energy, $\sigma_{i,j}$ is the sum of atomic radii, λ_i is correlation length and V_j is atomic volume)

Alford et al. 2016.

Hydrogen bonding terms

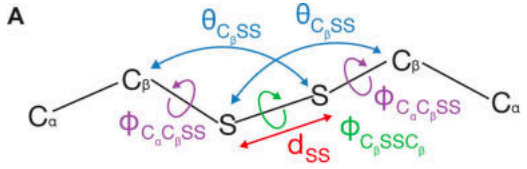
Define angles in
H-bond network



A = Acceptor H = Hydrogen
D = Donor B = Base B2 = Parent

Structural information
from Hbonds in PDB

$$E_{\text{hbond}} = \sum_{\text{H,A}} w_{\text{H}} w_{\text{A}} f(E^{\text{HA}}_{\text{hbond}}(d_{\text{HA}}) + E^{\text{AHD}}_{\text{hbond}}(\theta_{\text{AHD}}) + E^{\text{BAH}}_{\text{hbond}}(\theta_{\text{BAH}}) + E^{\text{B}_2\text{BAH}}_{\text{hbond}}(\rho, \phi_{\text{B}_2\text{BAH}}, \theta_{\text{BAH}}))$$



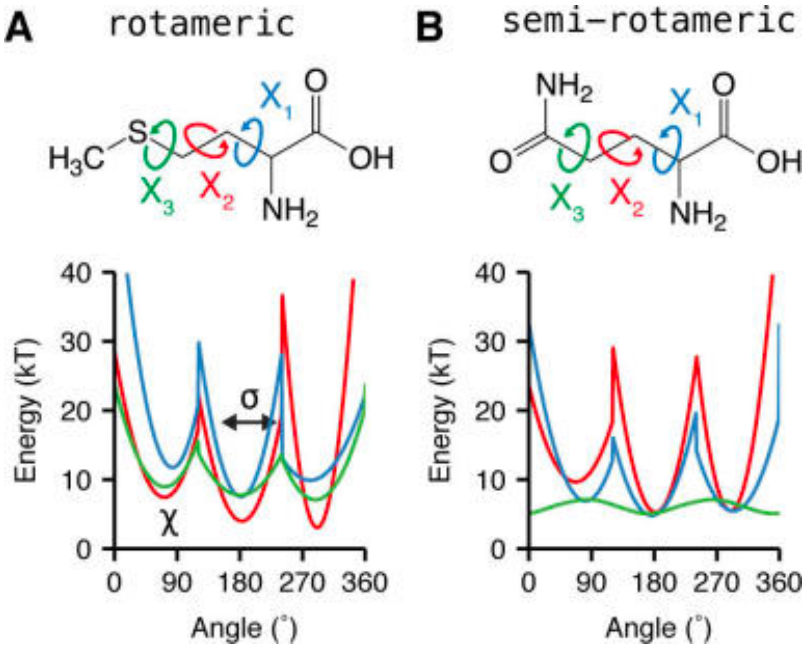
Structural information
from disulfides in PDB

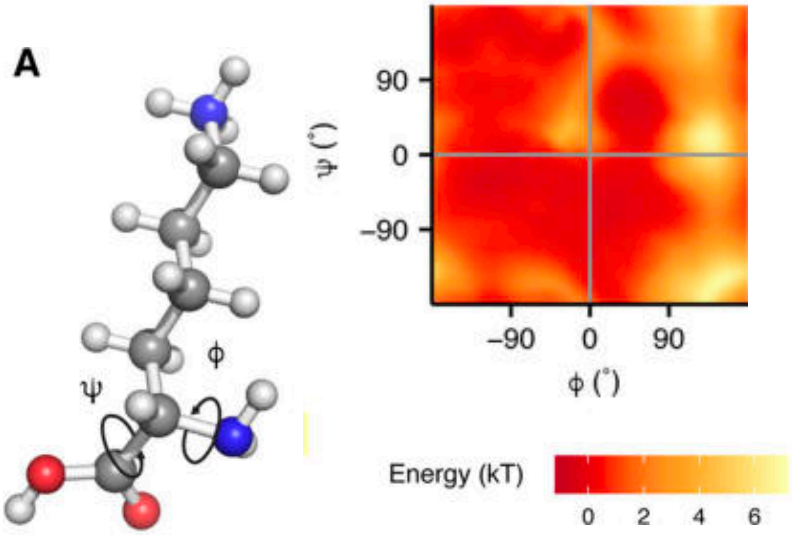
$$E_{\text{dslf\_fa13}} = \sum_{\text{S}_1,\text{S}_2} E^{\text{SS}}_{\text{dslf}}(d_{\text{SS}}) + E^{\text{CSS}}_{\text{dslf}}(\theta_{\text{C}_{\beta 1}\text{SS}}) + E^{\text{CSS}}_{\text{dslf}}(\theta_{\text{C}_{\beta 2}\text{SS}}) + E^{\text{C}_\alpha\text{C}_\beta\text{SS}}_{\text{dslf}}(\phi_{\text{C}_{\alpha 1}\text{C}_{\beta 1}\text{SS}}) + E^{\text{C}_\alpha\text{C}_\beta\text{SS}}_{\text{dslf}}(\phi_{\text{C}_{\alpha 2}\text{C}_{\beta 2}\text{SS}}) + E^{\text{C}_\beta\text{SSC}_\beta}_{\text{dslf}}(\phi_{\text{C}_{\beta 1}\text{SSC}_{\beta 2}})$$

Alford et al. 2016.

Sidechain and backbone torsion angles are knowledge-based potentials

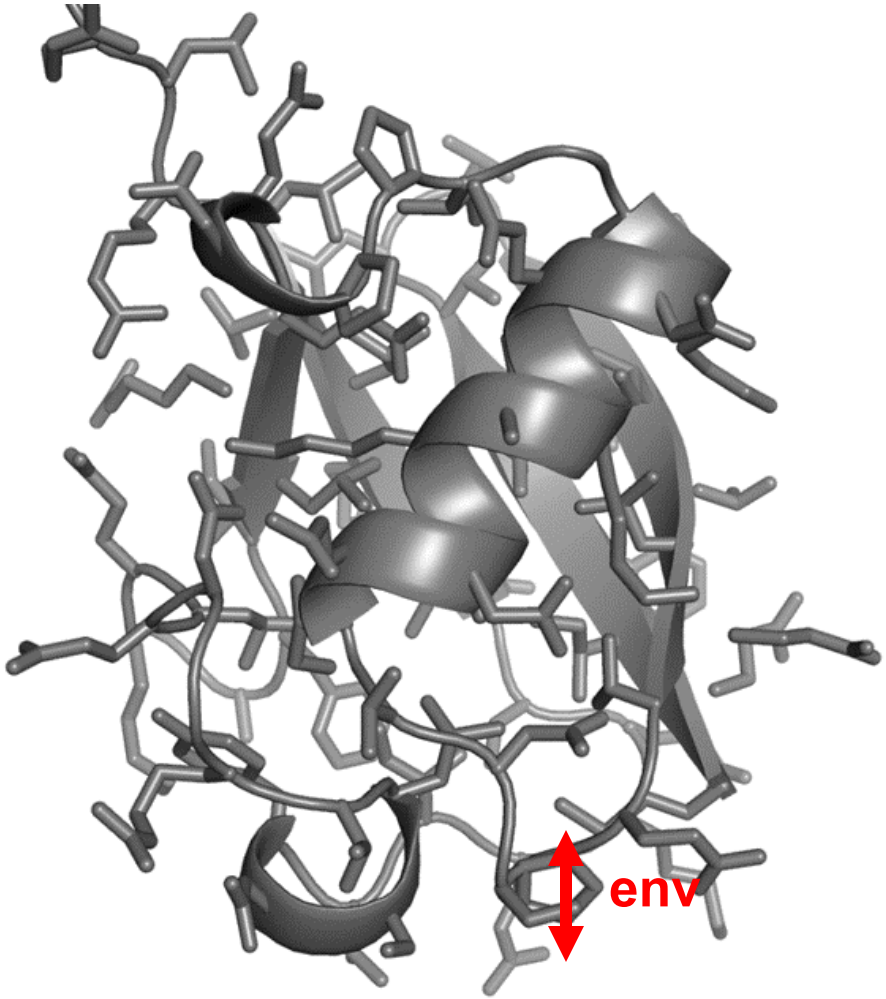
We assume that the more common a conformation is observed, it is because it is energetically more favorable





Alford et al. 2016.

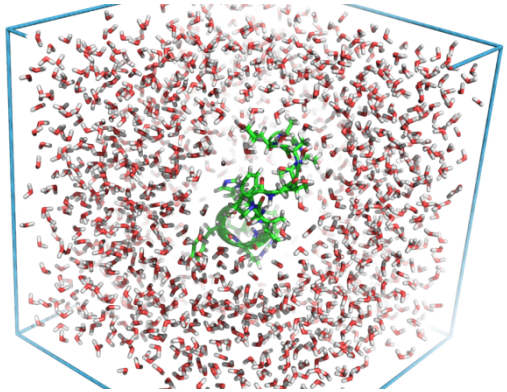
High Resolution Scoring Terms: Implicit Interactions with Solvent



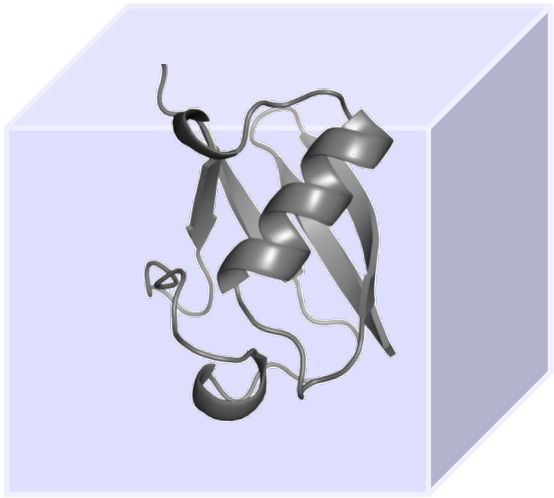
Solvation	$\sum_i \left[\Delta G_i^{ref} - \sum_j \left(\frac{2\Delta G_i^{free}}{4\pi^{3/2}\lambda_i r_{ij}^2} e^{-d_{ij}^2} V_j + \frac{2\Delta G_i^{free}}{4\pi^{3/2}\lambda_j r_{ij}^2} e^{-d_{ij}^2} V_i \right) \right]$
Ramachandran torsion preferences	$\sum_i -\ln [P(\phi_i, \psi_i \mid aa_i, ss_i)]$
Rotamer self-energy	$\sum_i -\ln \left[\frac{P(rot_i \mid \phi_i, \psi_i) P(aa_i \mid \phi_i, \psi_i)}{P(aa_i)} \right]$
Unfolded state reference energy	$\sum_{aa} n_{aa}$

Implicit Solvation Potentials

- Implicit Solvation Models can be:
  - knowledge-based (RosettaMembrane low-res)
  - physics-based (IMM1 in CHARMM)



- SASA model
  - $\Delta G_{\text{solv}} = \sum_i \sigma_i \ ASA_i$
- Poisson-Boltzmann
  - $\vec{\nabla} \cdot \left[\epsilon(\vec{r})\vec{\nabla}\Psi(\vec{r})\right] = -4\pi\rho^f(\vec{r}) - 4\pi \sum_i c_i^\infty z_i q \lambda(\vec{r}) e^{\frac{-z_i q \Psi(\vec{r})}{kT}}$
- Generalized Born Approximation
  - $G_s = \frac{1}{8\pi}\left(\frac{1}{\epsilon_0} - \frac{1}{\epsilon}\right)\sum_{i,j}^{N} \frac{q_i q_j}{f_{GB}}$
  - $f_{GB} = \sqrt{r_{ij}^2 + a_{ij}^2 e^{-D}}$



The Lazaridis – Karplus Model EEF1: A Model of Desolvation

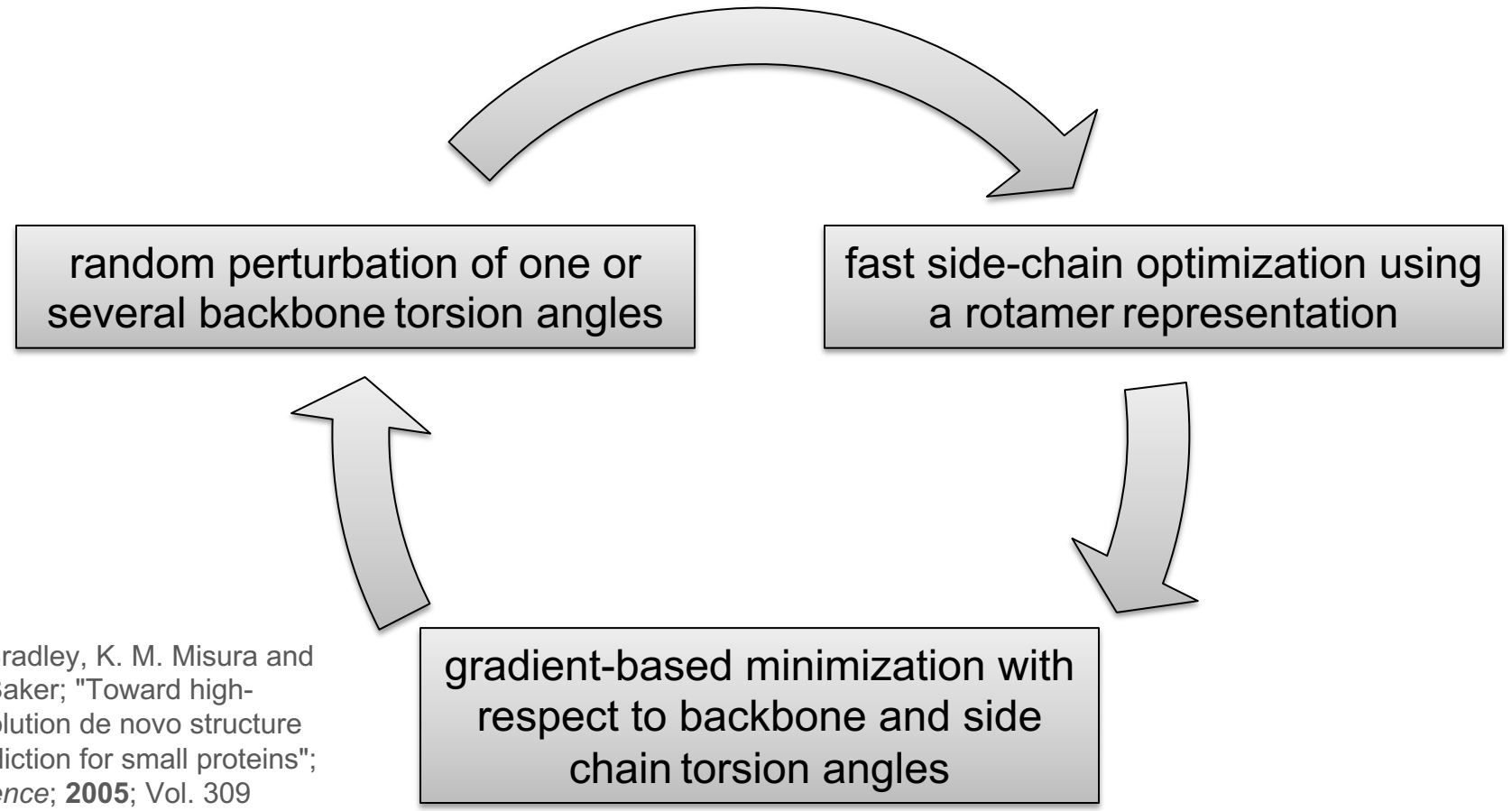
- EEF1: Gaussian-shaped solvent exclusion

$$\Delta G_i^{solv} = \Delta G_i^{ref} - \sum_j \int_{Vj} f_i(r) dr$$

- unaccounted effects:
  - hydrophobic effect
  - viscosity of the solvent
  - H-bonds with solvent
- limitations:
  - choice of solvent is important
  - models are only tested on small proteins!!!
  - ionization of charged groups neglected

T. Lazaridis and M. Karplus; "Effective energy function for proteins in solution"; *Proteins*; **1999**; Vol. 35 (2): p. 133-152.

Refinement Cycle with Side Chain Repacking and All Atom Minimization

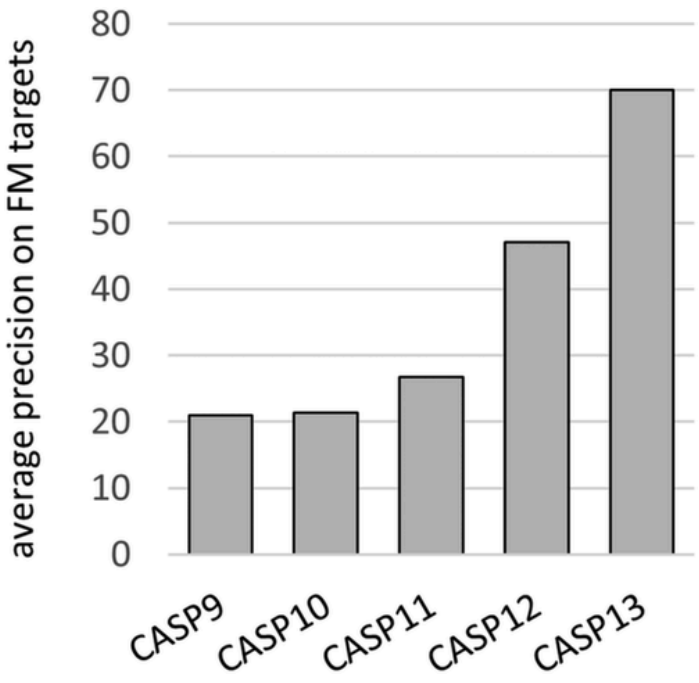


P. Bradley, K. M. Misura and D. Baker; "Toward high-resolution de novo structure prediction for small proteins"; *Science*; **2005**; Vol. 309 (5742): p. 1868-71.

How do we test protocols? CASP (Critical Assessment of protein Structure Prediction)

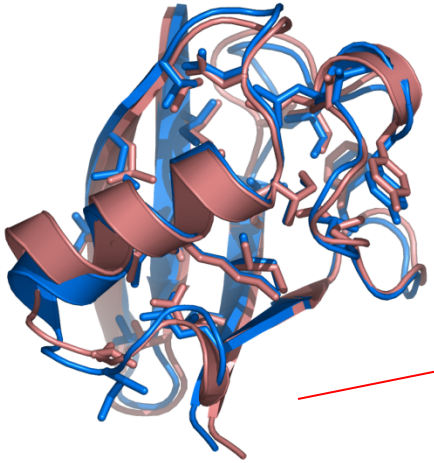
- Benchmarking is important to understand the strengths and limits of our methods
- CASP is a double-blind experiment

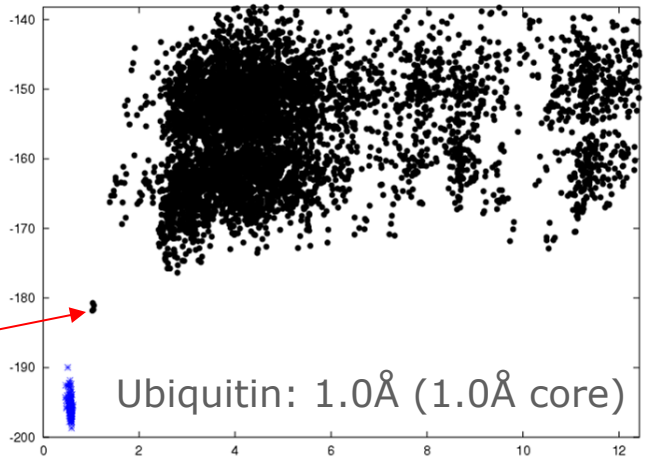
- Ab initio
- Template-based models
- Model refinement
- Accuracy estimation
- Protein assemblies

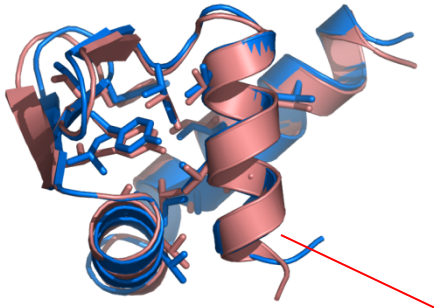


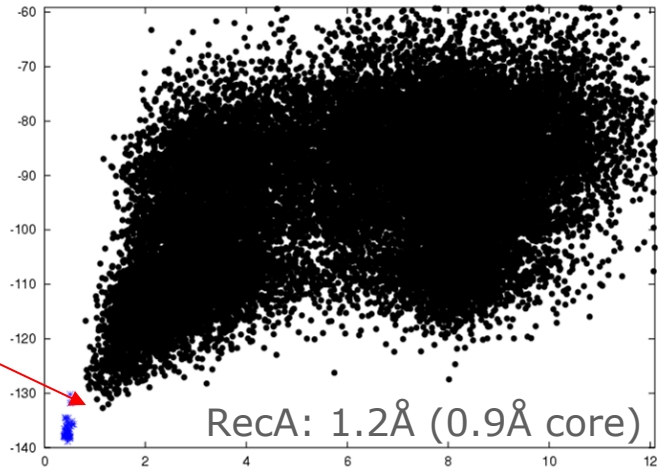
Benchmarking CASP predictions

1.6Å Cα-RMSD blind structure prediction for CASP6 target T0281, hypothetical protein from Thermus thermophilus Hb8. Superposition of our submitted model for this target in CASP6 (blue) with the crystal structure (red; PDB code 1whz)

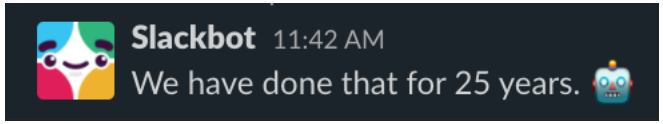






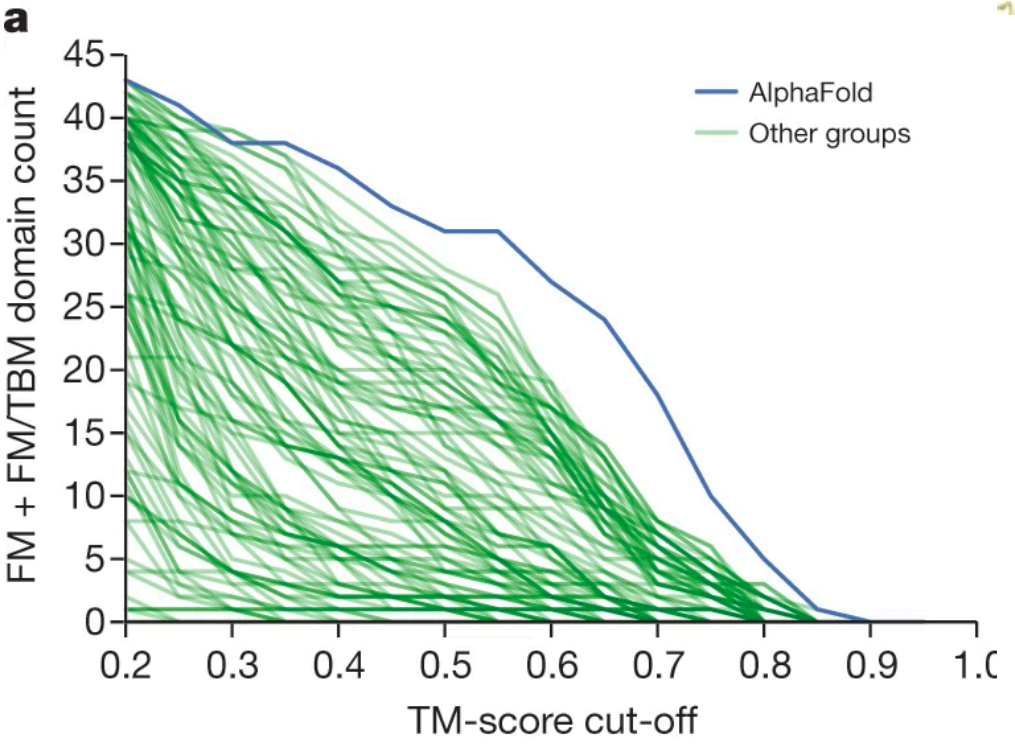


CASP13: Machine learning



- ML is excellent at picking out patterns in large datasets
- AlphaFold trained ANNs on residue pair distance information

Rosetta potentials used



Senior, A.W., Evans, R., Jumper, J. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577,** 706–710 (2020).

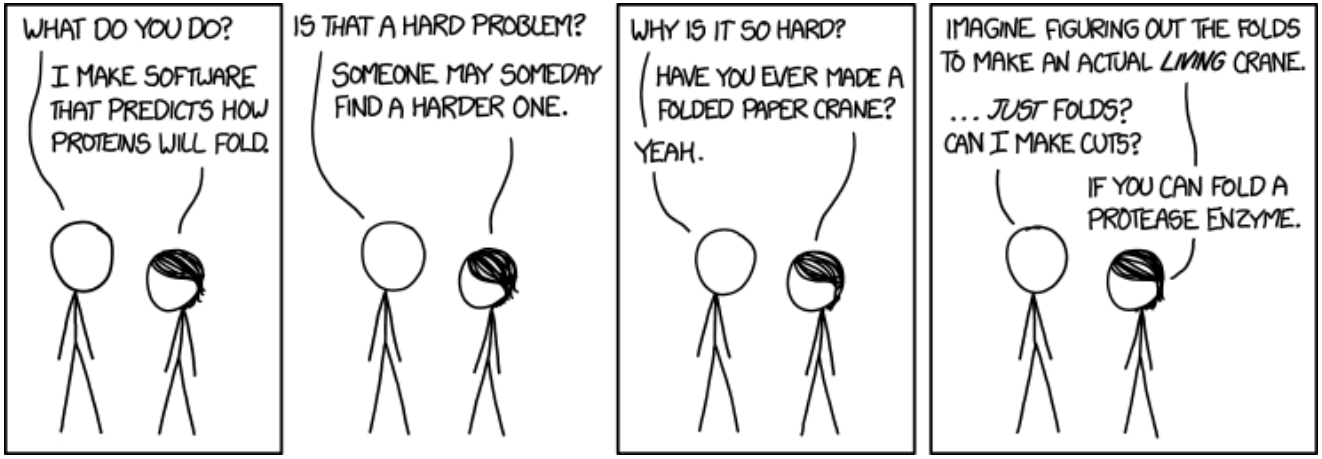
Takeaways from the *de novo* section

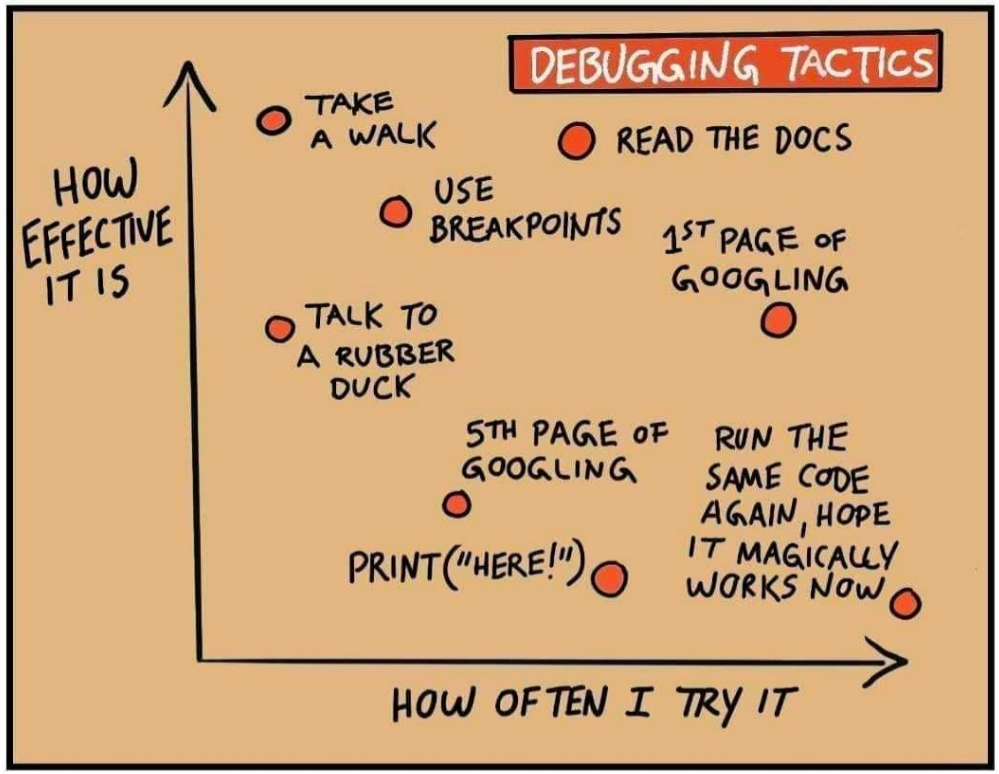
Low-resolution stage:

- Centroid atom representation
- Sampling using fragments to obtain possible global folds
- Scoring using statistics-based metrics

High-resolution stage:

- All-atom refinement
- Sampling sidechain rotamers via rotamer libraries
- Scoring using physics-based, empirical-based and statistical-based methods

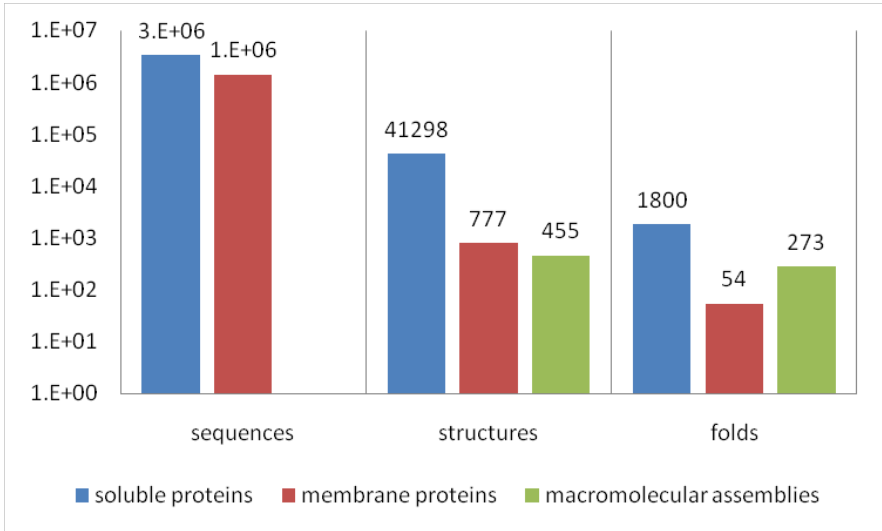

WHAT DO YOU DO?
I MAKE SOFTWARE THAT PREDICTS HOW PROTEINS WILL FOLD.
IS THAT A HARD PROBLEM?
SOMEONE MAY SOMEDAY FIND A HARDER ONE.
WHY IS IT SO HARD?
HAVE YOU EVER MADE A FOLDED PAPER CRANE?
YEAH.
IMAGINE FIGURING OUT THE FOLDS TO MAKE AN ACTUAL LIVING CRANE.
... JUST FOLDS? CAN I MAKE CUTS?
IF YOU CAN FOLD A PROTEASE ENZYME.


DEBUGGING TACTICS
HOW EFFECTIVE IT IS
HOW OFTEN I TRY IT
TAKE A WALK
READ THE DOCS
USE BREAKPOINTS
1ST PAGE OF GOOGLING
TALK TO A RUBBER DUCK
5TH PAGE OF GOOGLING
RUN THE SAME CODE AGAIN, HOPE IT MAGICALLY WORKS NOW
PRINT("HERE!")

Goals for this talk

1. Intro to general topics:
   a) Computational structural biology
   b) Sampling: MC/simulated annealing
   c) Scoring: statistical potentials, physics-based, empirical, machine learning
2. Ab initio structure prediction
   a) Low-resolution centroid predictions
   b) High-resolution full-atom refinement
   c) CASP (Critical Assessment of protein Structure Prediction)
3. **Comparative modeling**
   a) Why we do comparative modeling? MPs/GPCR example
   b) Sequence alignments + threading
4. Loop modeling
   a) Why we "mind the gap"?
   b) Cyclic Coordinate Descent (CCD)
   c) Kinematic closure (KIC)

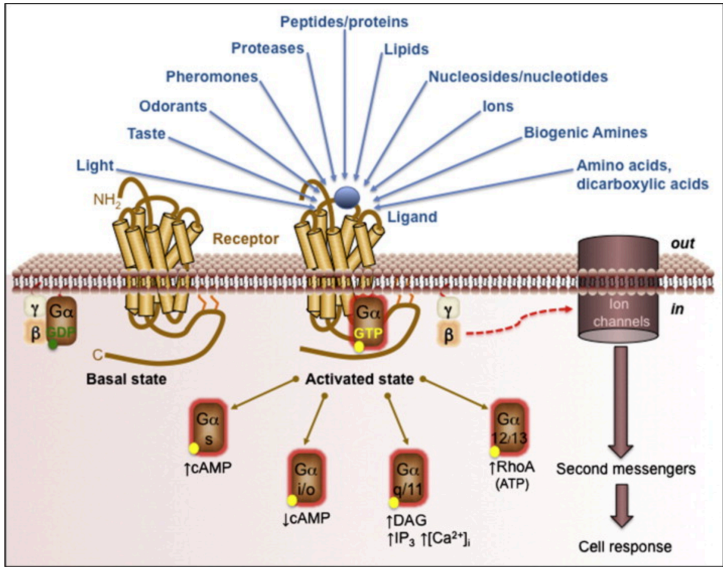
Homology modeling is commonly used in GPCR structural biology



- ~50% of available therapeutics target MPs
- Difficult to structurally characterize

- Humans have ~800 unique GPCRs and are targets for many therapeutics

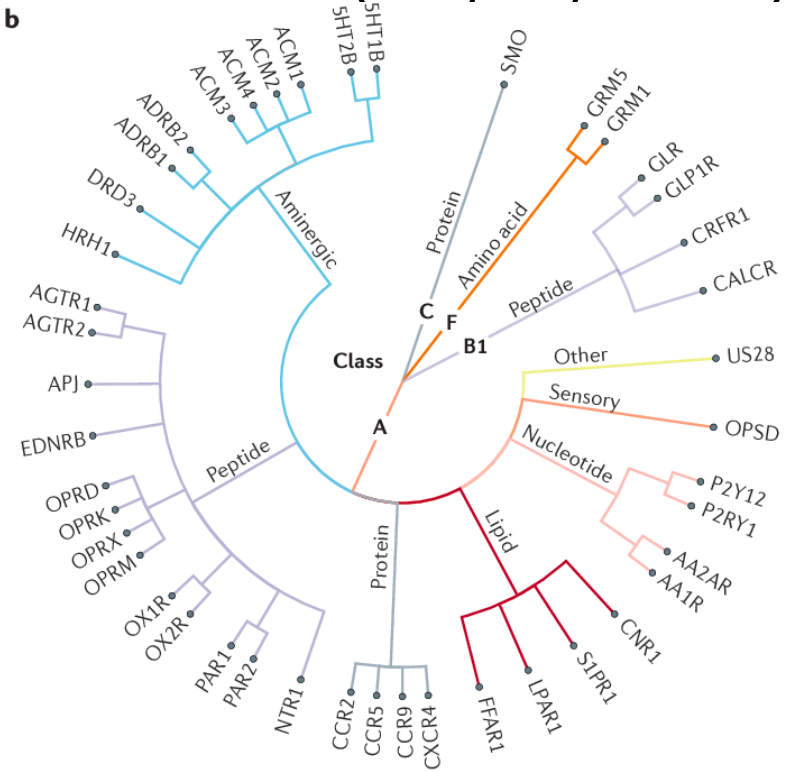
Structural and sequence conservation of GPCRs:
- 7 transmembrane helices
- 3 intracellular and 3 extracellular loops
- Highly conserved residues



Munk *et al.* An Online Resource for GPCR Structure Determination and Analysis. *Nat. Methods.* **2019.**

GPCR structural biology: where are we?

- What you need to do to crystalize GPCR's:
  - Cut-off N-terminus and C-terminus
  - Insert a stabilizing protein at the position of ICL3 (T4-lysozyme etc.)
  - nanobodies
  - Thermostabilizing mutations
  - A solubilization method
  - Sampling, luck and patience

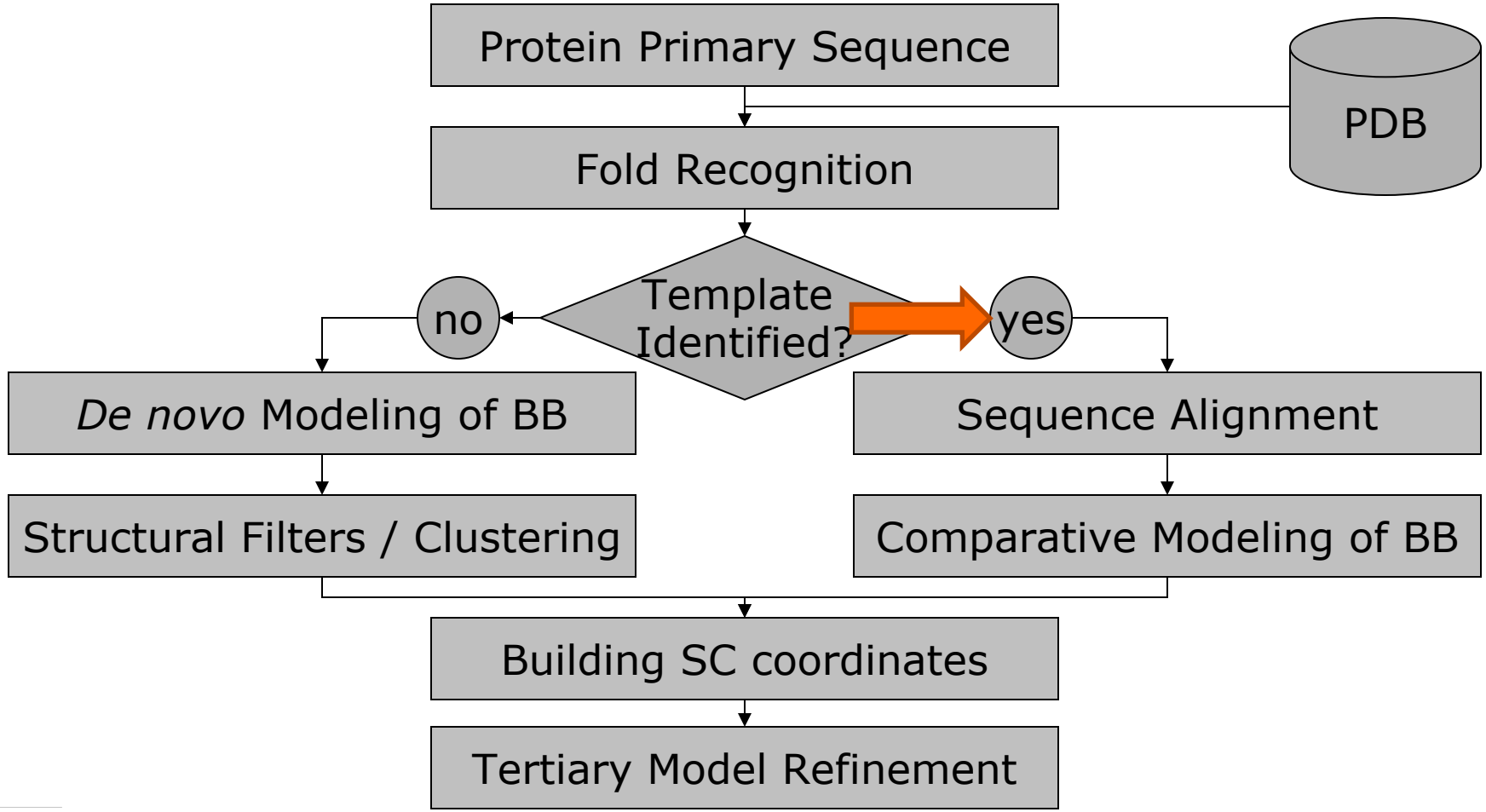


→ We are modeling GPCR crystal structures, we are not modeling GPCRs in their natural environment!

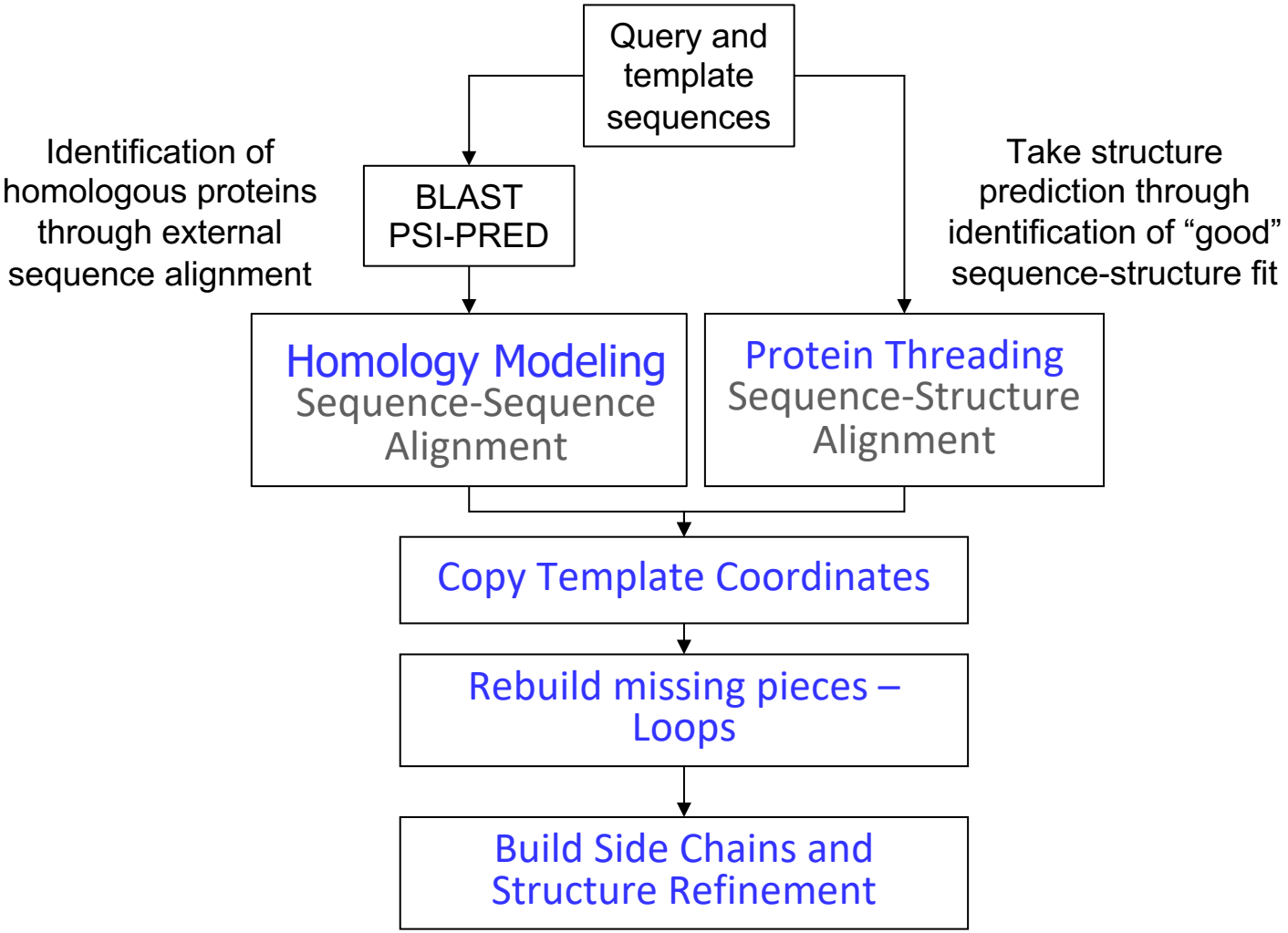
Hauser, A., Attwood, M., Rask-Andersen, M. *et al.* Trends in GPCR drug discovery: new agents, targets and indications. *Nat Rev Drug Discov* **16,** 829–842 (2017). https://doi.org/10.1038/nrd.2017.178

Comparative modeling

Using known structures with similar sequences to guide structure prediction



Comparative modeling: build model based on an existing structure(s)



How is sequence alignment useful

- Many sequences with unknown structure or function
  - "Good" alignment may imply that the sequences are homologous
  - If one sequence has known structure/function, the alignment may yield insights about the other
- Typical questions answered by sequence alignment:
  - Is my newly sequenced gene already known?
  - Two proteins have the same function – are there sequence features that they share?
  - Protein A has known structure, protein B is similar in terms of sequence – can we build a rough 3D model of the latter protein?
  - Given a set of good pairwise sequence alignments, can we construct a *better* alignment involving *all* of them (multiple sequence alignment)

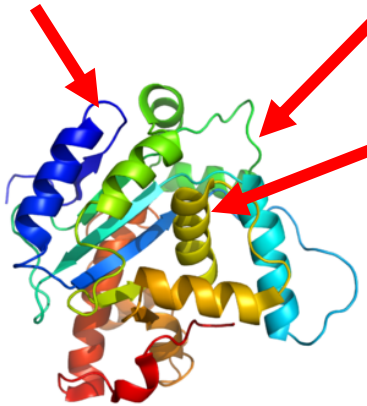
Alignment Algorithms

- There is no "best" alignment method
  - Visual
  - Brute Force
  - Word-Based/k tuple (BLAST/PSI-PRED algorithm)
  - Many other optimization methods, but not discussed here
- "Good" alignment is based on biological knowledge
- Alignment algorithms find the best possible alignment(s) for a given scoring regime
- Alignment of two biological sequences is **only a model!**

Threading – Energy Function measures Likelihood $P(seq|str)$

- Unlike sequence-sequence alignment where amino acids are aligned, a sequence-structure alignment aligns amino acids with structural environments
- A simple definition of structural environment E_s could include
  - secondary structure: alpha-helix, beta-strand, loop
  - solvent accessibility: 0, 10, 20, …, 100% of accessibility
  - amino acid pair distances

MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE



How likely is that amino acid in this secondary str.

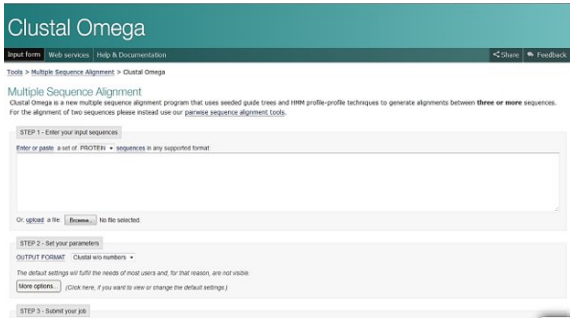
How likely is that amino acid in this exposure state

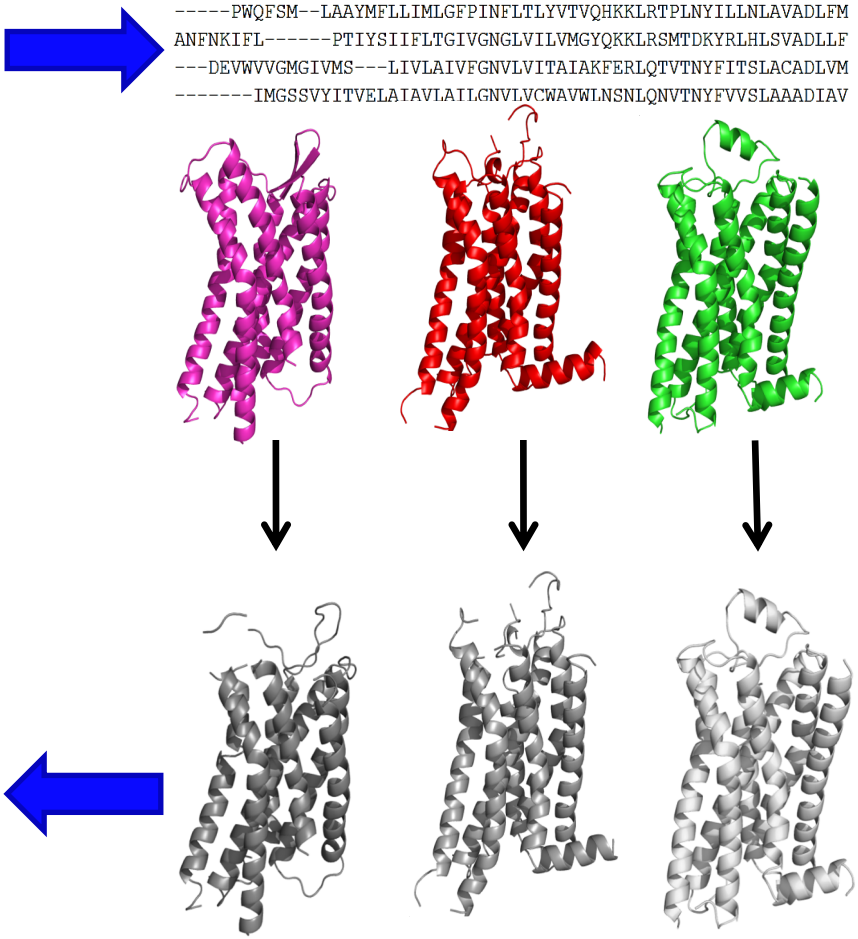
How likely is are two amino acids in this distance

$$P(seq|str) = \prod_i P(aa_i|sse) \times \prod_i P(aa_i|env) \times \prod_{i,j} P(aa_i, aa_j|d_{ij})$$

$$E(seq|str) = E_{gap} + \sum_i E(aa_i|sse) + \sum_i E(aa_i|env) + \sum_{i,j} E(aa_i, aa_j|d_{ij})$$

RosettaCM protocol typically relies on pre-determined alignment



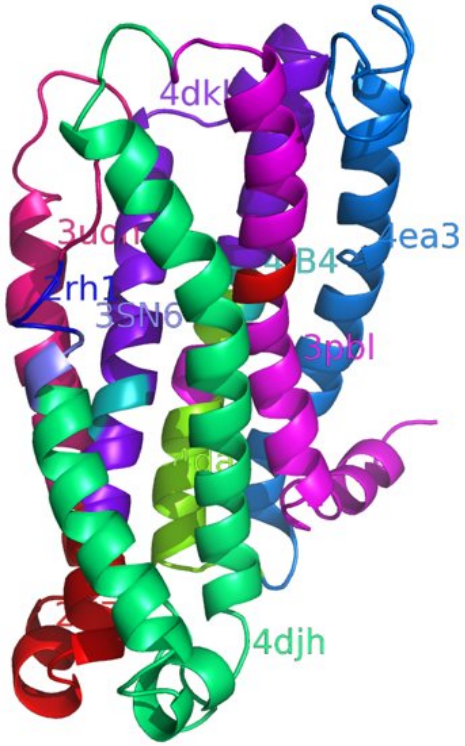


- Model loops
- All-atom refinement

Using multiple templates improves model accuracy

- Templates are ideally >30% sequence identity to target
- It is advisable to use multiple templates due to the low sequence identity in available templates

Template	PDB ID	% Seq id
β2-adrenoceptor	3SN6	36
5-HT1B receptor	4IAR	32
β2-adrenoceptor	3D4S	34
5-HT2B receptor	5TVN	32
M1 receptor	5CXV	32
H1 receptor	3RZE	31
M4 receptor	5DSG	29
A2A receptor	2YDO	28
A1 receptor	5N2S	27

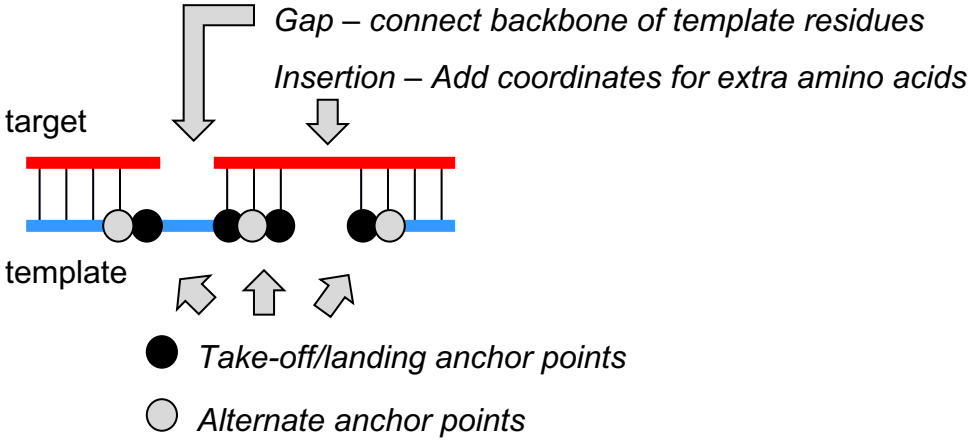


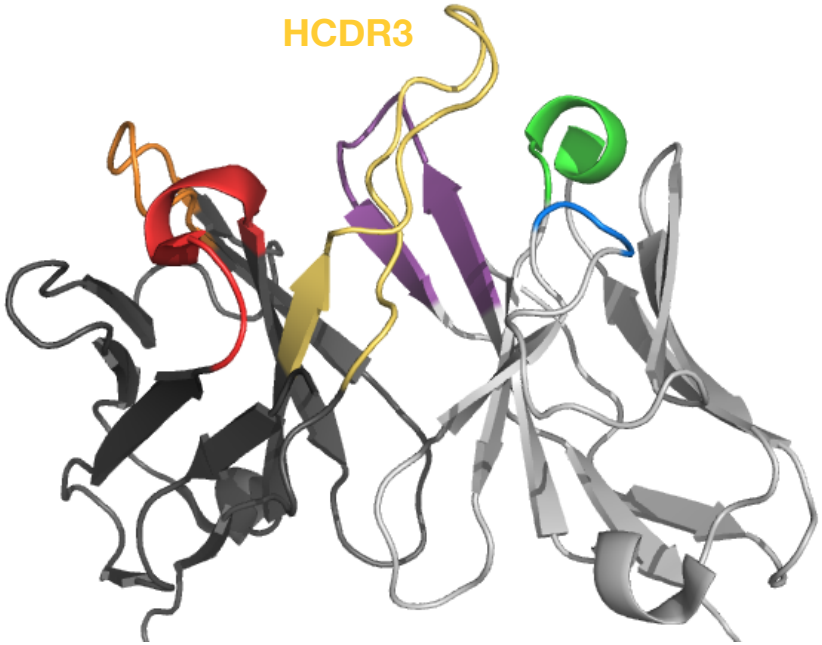
Goals for this talk

1. Intro to general topics:
   a) Computational structural biology
   b) Sampling: MC/simulated annealing
   c) Scoring: statistical potentials, physics-based, empirical, machine learning
2. Ab initio structure prediction
   a) Low-resolution centroid predictions
   b) High-resolution full-atom refinement
   c) CASP (Critical Assessment of protein Structure Prediction)
3. Comparative modeling
   a) Why we do comparative modeling? MPs/GPCR example
   b) Sequence alignments + threading
4. **Loop modeling**
   a) Why we "mind the gap"?
   b) Loop closure methods (CCD/KIC)

Why do we model loops?

- Fixing chain breaks (homology modeling)
- Protein-protein recognition and binding (E.g. HCDR3 loop is highly variable and important for binding specificity)
- Often times not resolved in experimental structures
- Success <12 residues





Loop Closure Problem – Definition

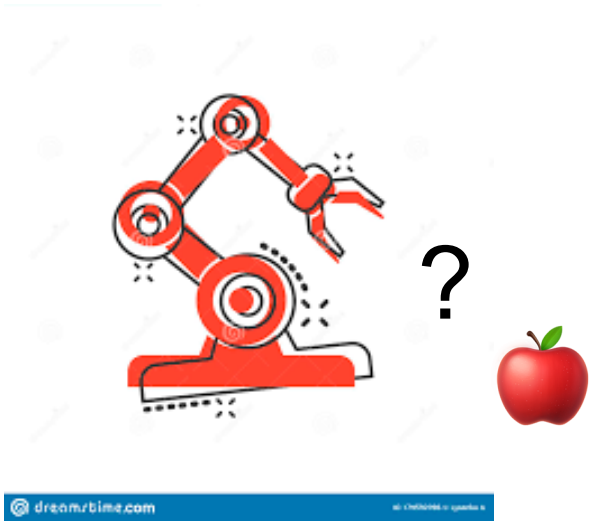
Input

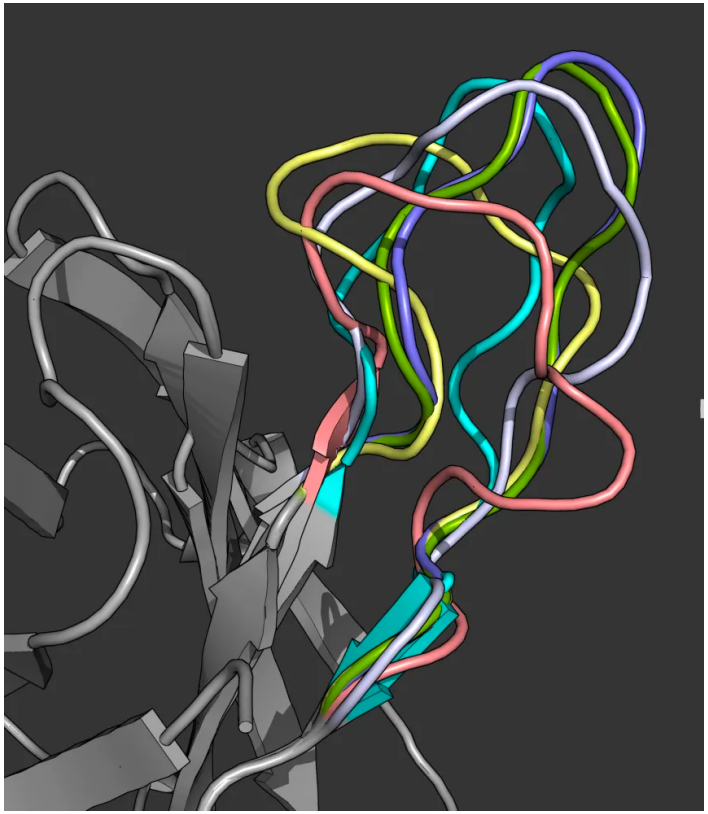
- 2 Anchor residues
- Length of missing fragment

Output

- A small number of candidate structures for missing fragment

The relative placement of loop take-off and landing points can be described with six parameters in 3D space – one translation $T(x, y, z)$ and one rotation $R(\alpha, \beta, \gamma)$



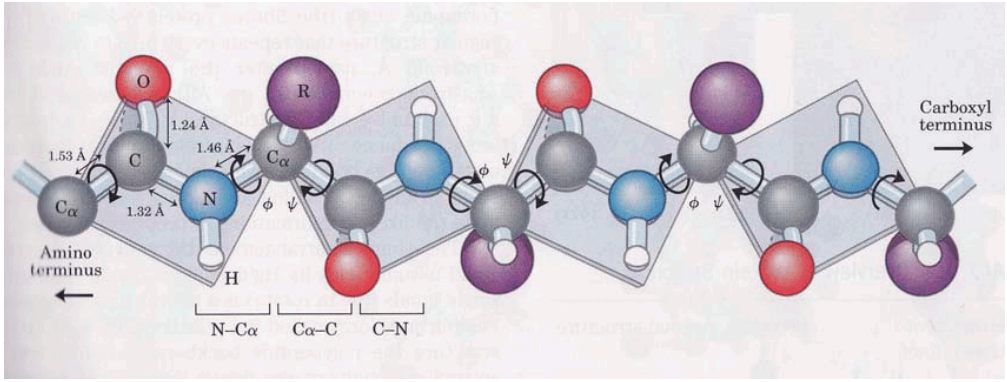


https://www.blopig.com/blog/2016/01/loop-model-selection/

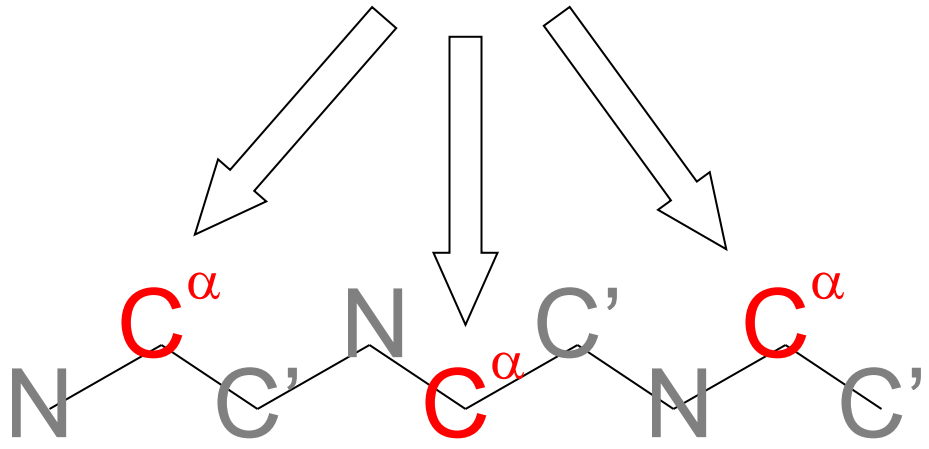
The Tripeptide Problem has six DOFs

3 residues, each with Φ and Ψ
= 6 degrees of freedom

“joints” are the angles centered at the Cα



Spherical Joints

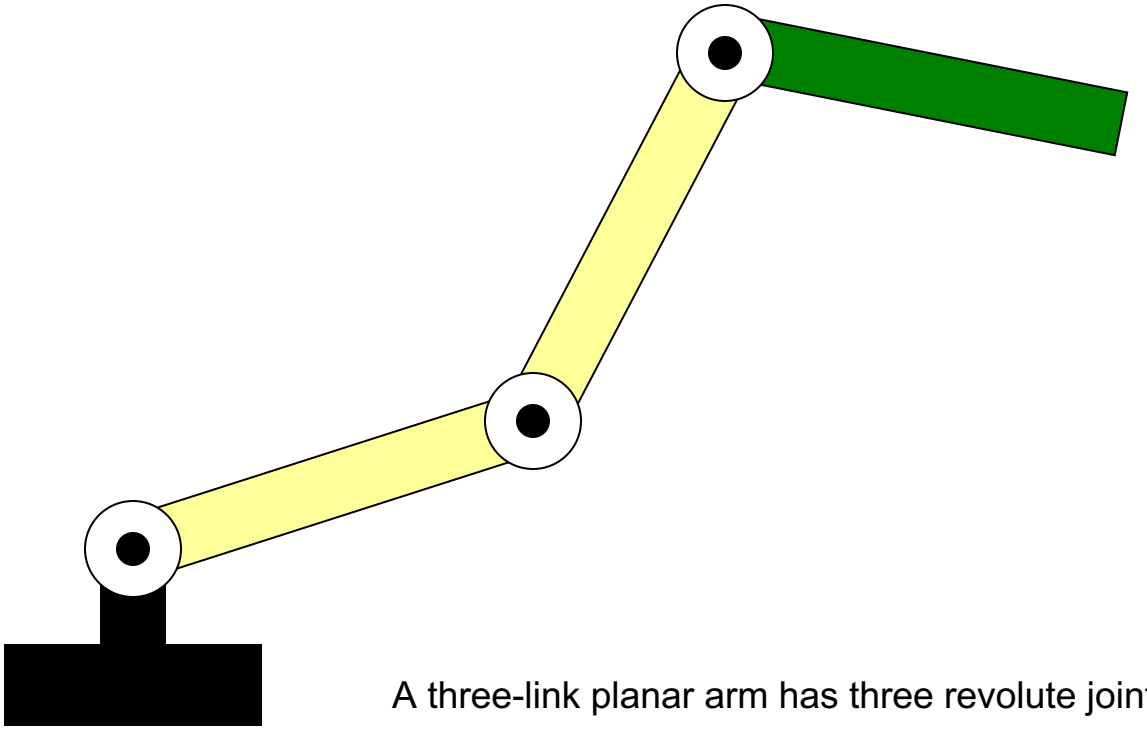


The Molecular Loop Closure Problem and the Degrees of Freedom (DOF)

- The relative placement of loop take-off and landing points can be described with six parameters in 3D space – one translation $T(x, y, z)$ and one rotation $R(\alpha, \beta, \gamma)$
- < 3 residues: fewer than 6 DOFs, might have no solutions
- 3 residues: < 16 possible solutions, possibly none
- > 3 residues: > 6 DOFs, might have an infinite number of solutions.

The Robotic Loop Closure Problem: Which Φ/Ψ angles connect the anchor points?

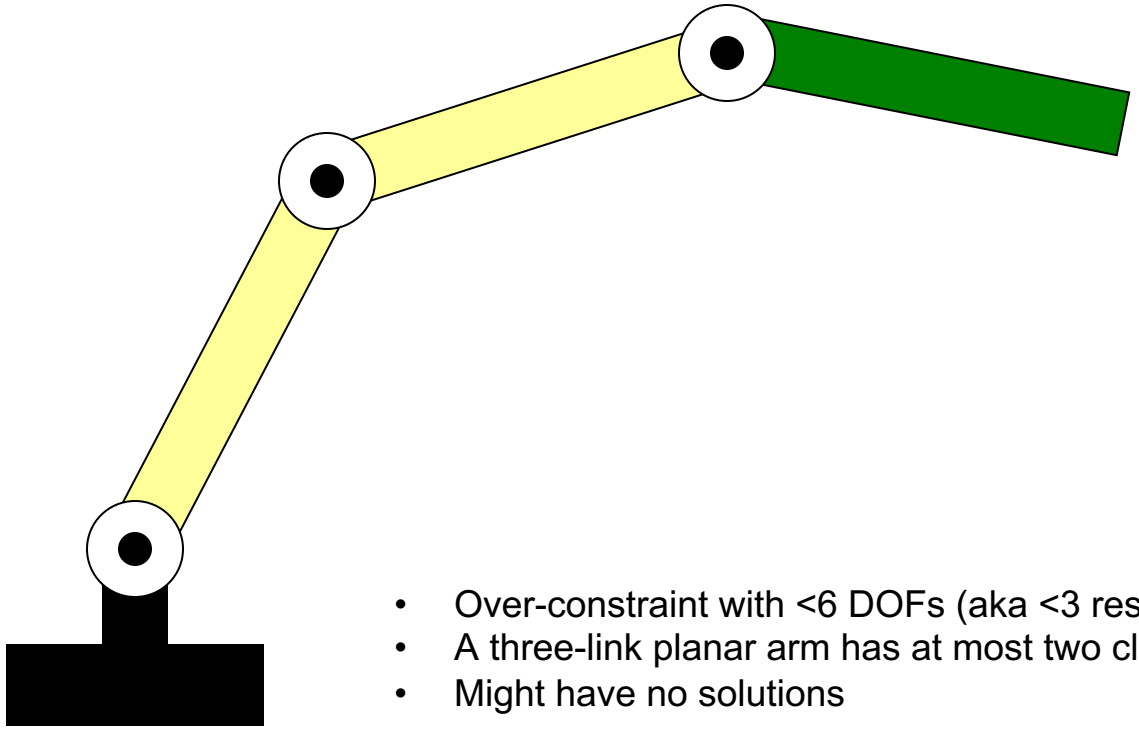
Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.



A three-link planar arm has three revolute joints.

The Robotic Loop Closure Problem: Which Φ/Ψ angles connect the anchor points?

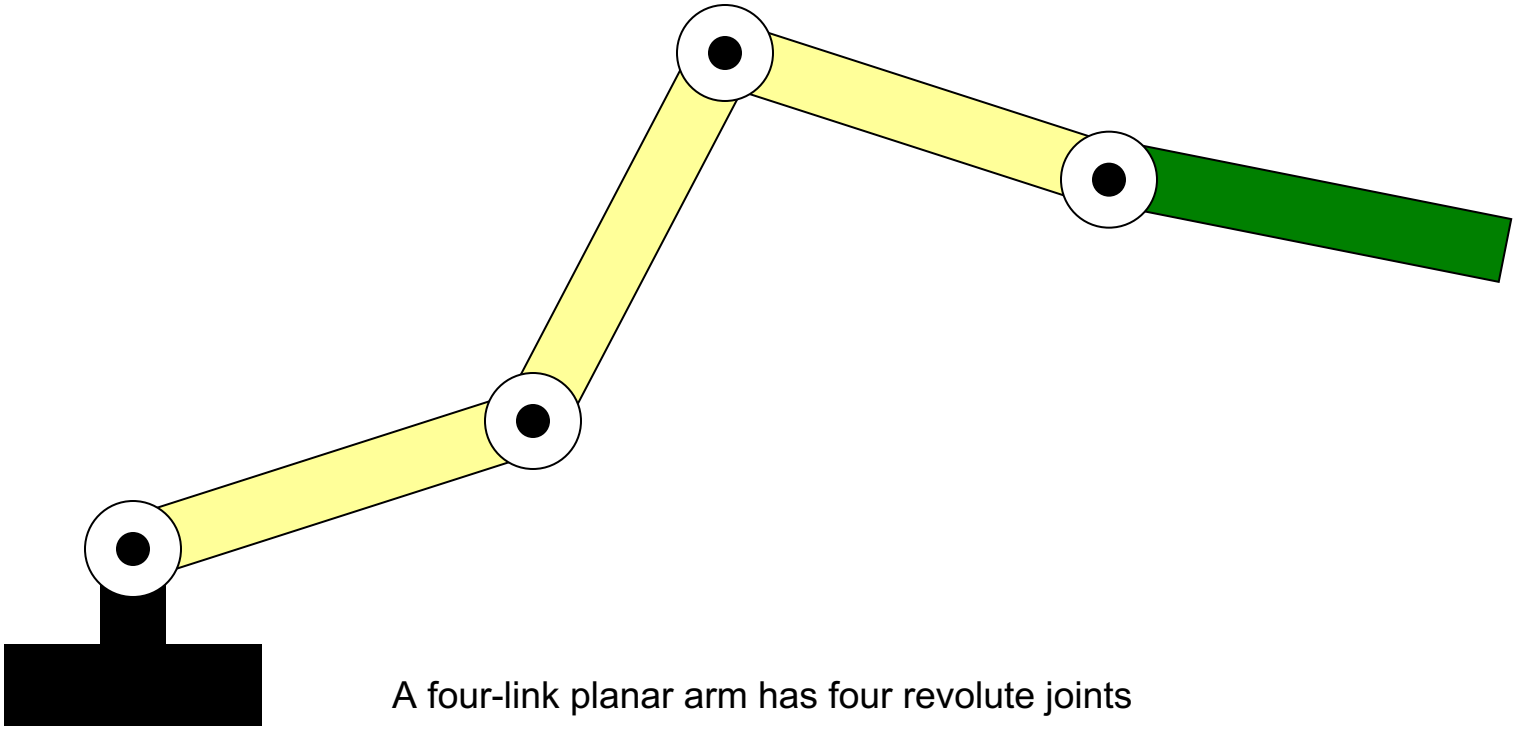
Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.



- Over-constraint with <6 DOFs (aka <3 residues)
- A three-link planar arm has at most two closed loop conformations
- Might have no solutions

The Robotic Loop Closure Problem: Which Φ/Ψ angles connect the anchor points?

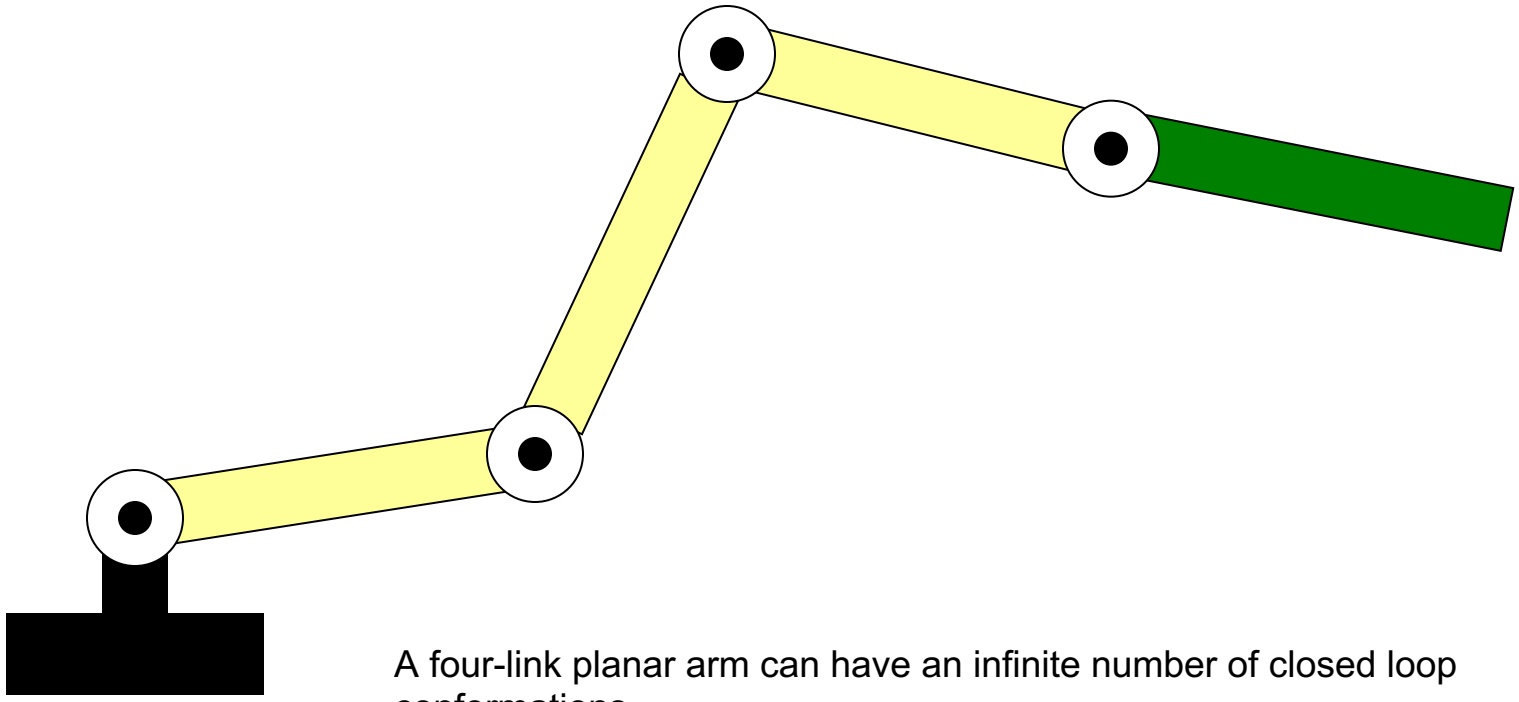
Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.



A four-link planar arm has four revolute joints

The Robotic Loop Closure Problem: Which Φ/Ψ angles connect the anchor points?

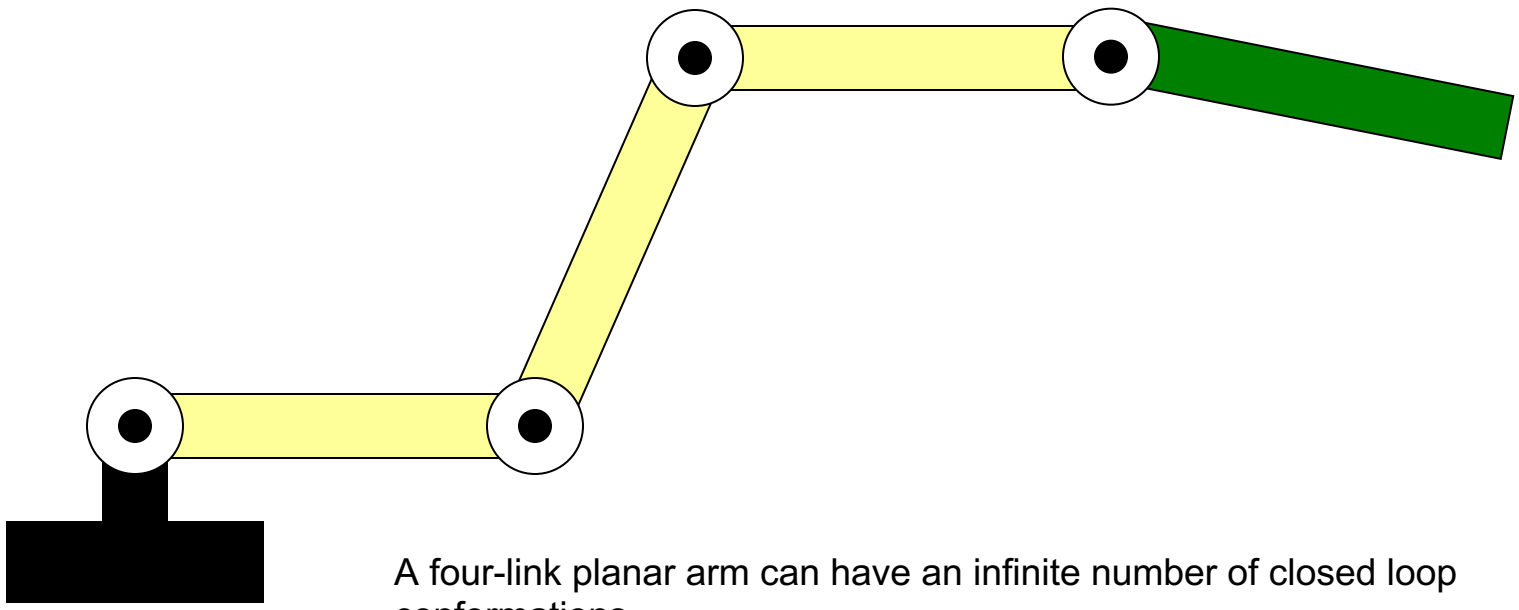
Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.



A four-link planar arm can have an infinite number of closed loop conformations.

The Robotic Loop Closure Problem: Which Φ/Ψ angles connect the anchor points?

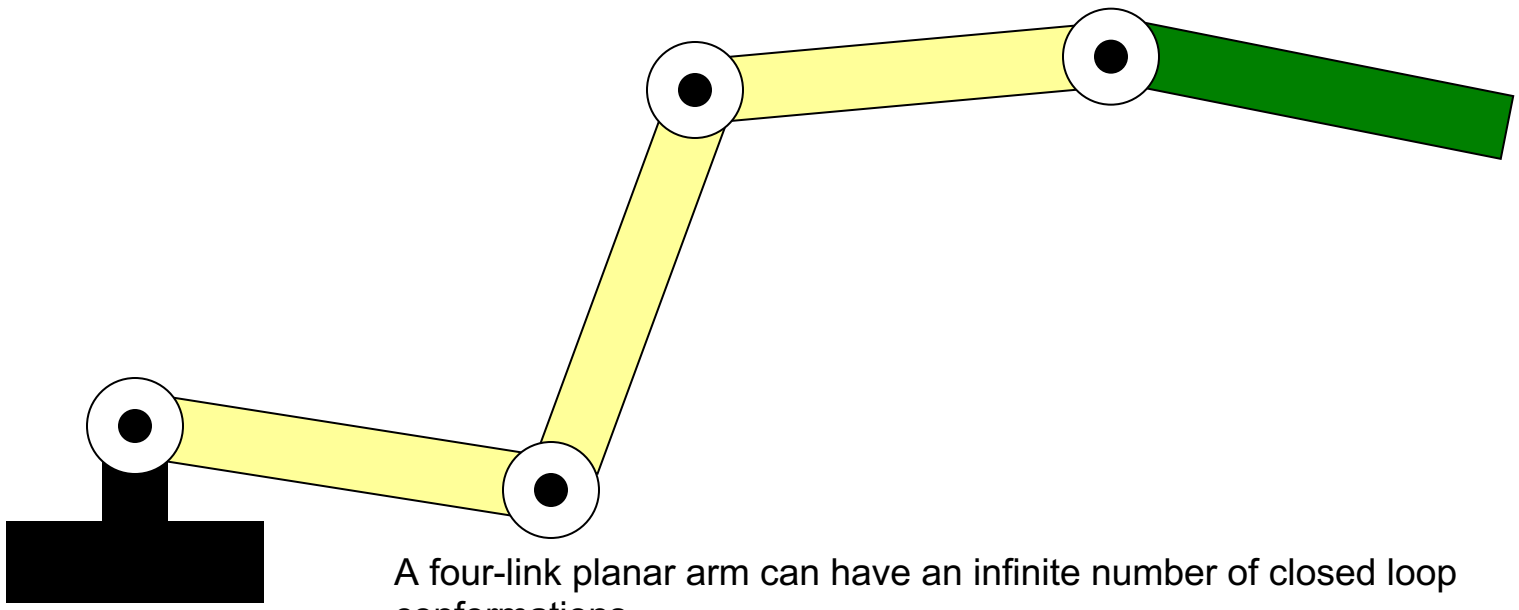
Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.



A four-link planar arm can have an infinite number of closed loop conformations.

The Robotic Loop Closure Problem: Which Φ/Ψ angles connect the anchor points?

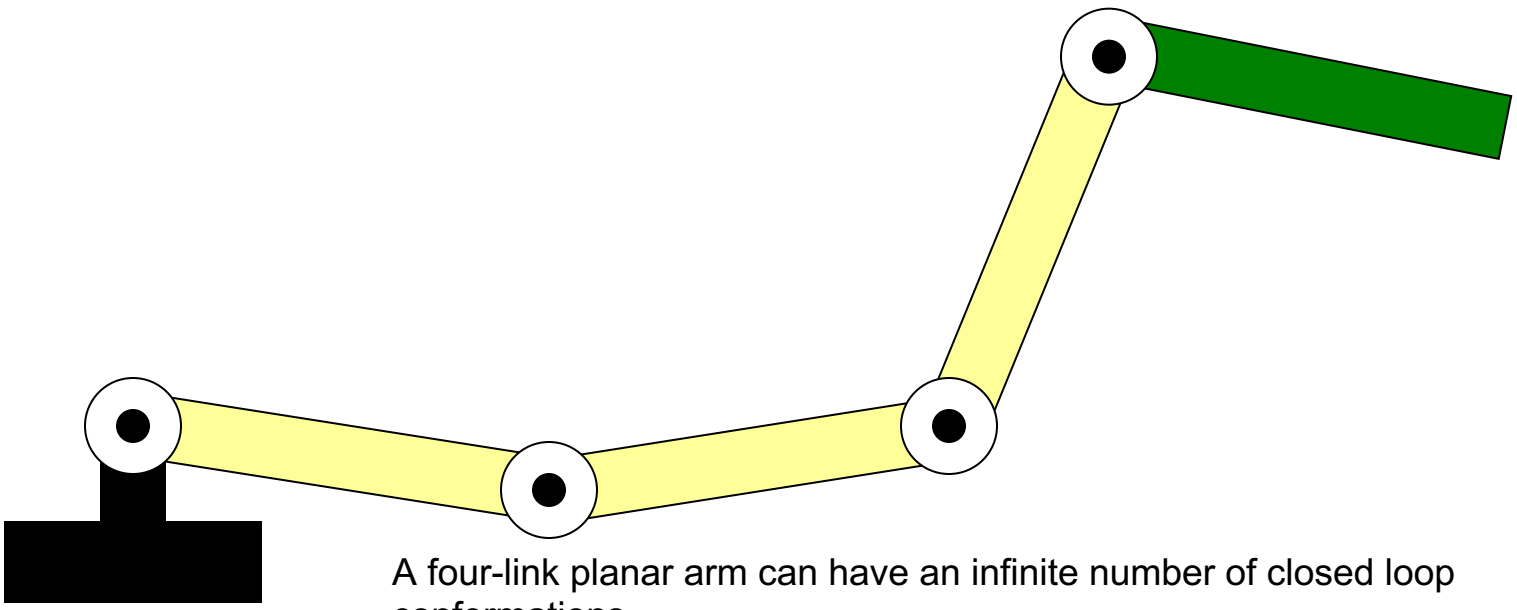
Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.



A four-link planar arm can have an infinite number of closed loop conformations.

The Robotic Loop Closure Problem: Which Φ/Ψ angles connect the anchor points?

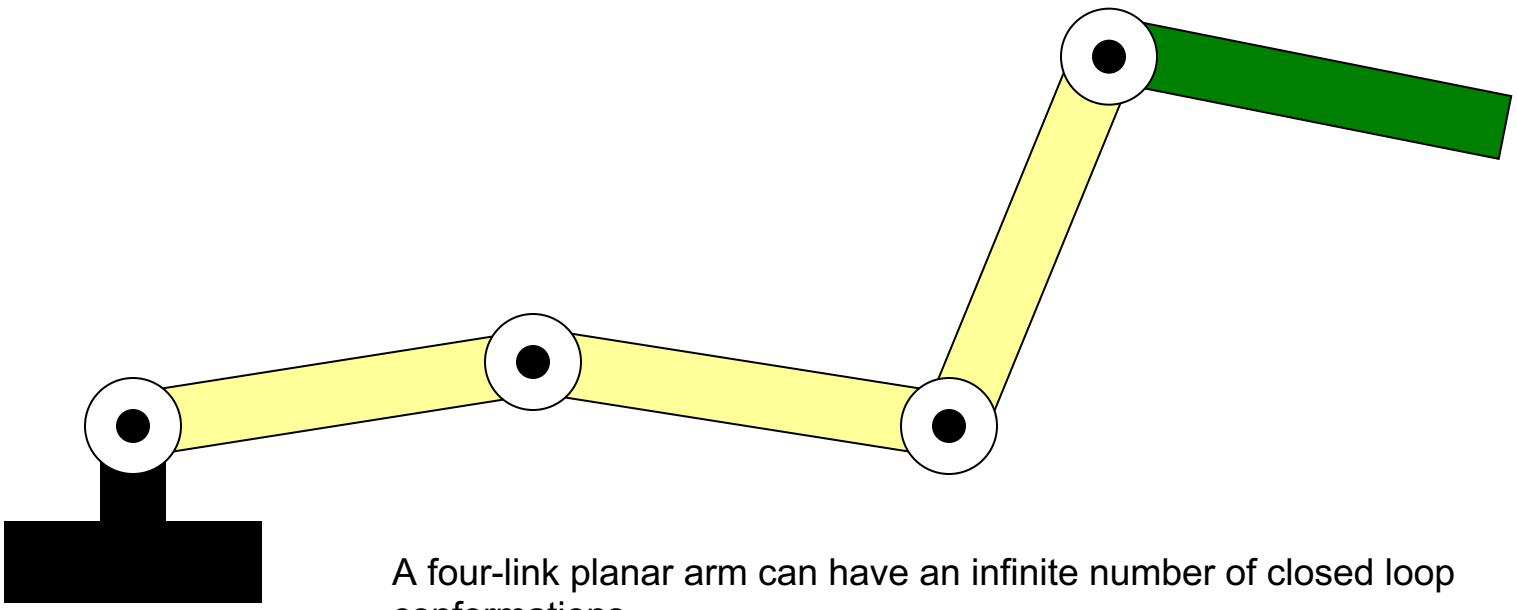
Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.



A four-link planar arm can have an infinite number of closed loop conformations.

The Robotic Loop Closure Problem: Which Φ/Ψ angles connect the anchor points?

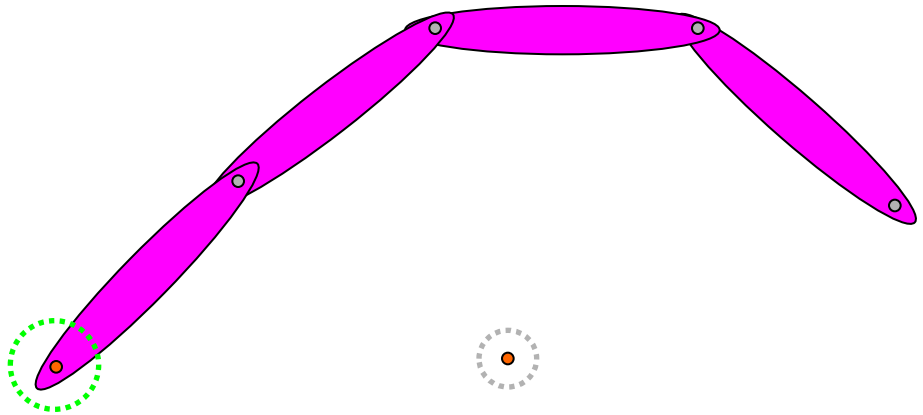
Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.



A four-link planar arm can have an infinite number of closed loop conformations.

Loop Closure: Which Φ/Ψ angles connect the anchor points?

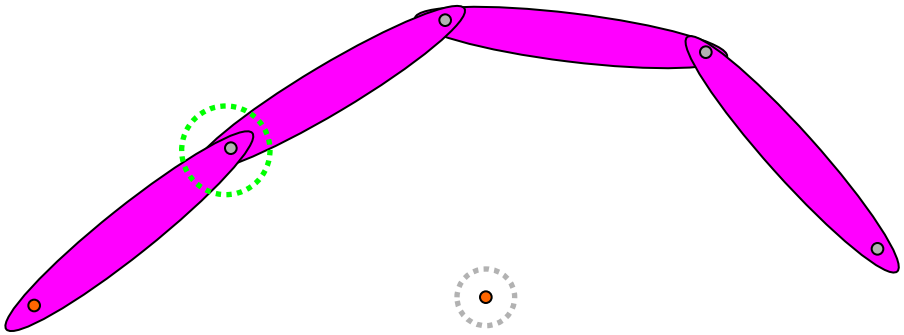
- Generate random conformation
- Close using Cyclic Coordinate Descent (CCD)



- Canutescu, A. A.; Dunbrack, R. L., Cyclic coordinate descent: Protein Sci 2003, 12, 963-972.

Loop Closure: Which Φ/Ψ angles connect the anchor points?

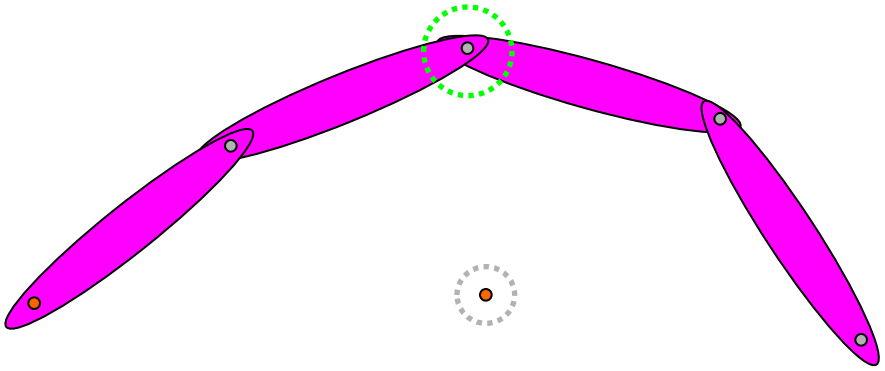
- Generate random conformation
- Close using Cyclic Coordinate Descent (CCD)



- Canutescu, A. A.; Dunbrack, R. L., Cyclic coordinate descent: Protein Sci 2003, 12, 963-972.

Loop Closure: Which Φ/Ψ angles connect the anchor points?

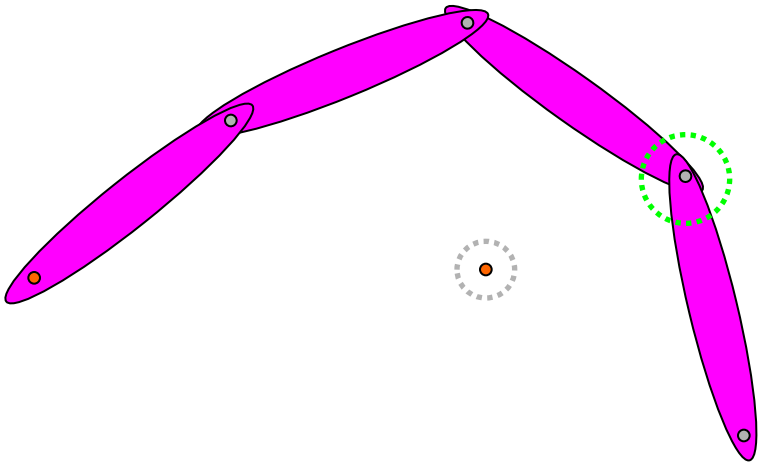
- Generate random conformation
- Close using Cyclic Coordinate Descent (CCD)



- Canutescu, A. A.; Dunbrack, R. L., Cyclic coordinate descent: Protein Sci 2003, 12, 963-972.

Loop Closure: Which Φ/Ψ angles connect the anchor points?

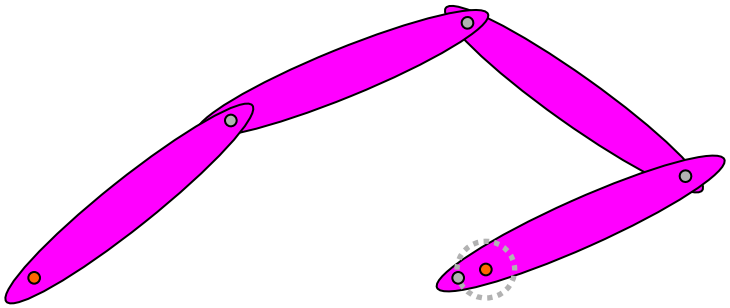
- Generate random conformation
- Close using Cyclic Coordinate Descent (CCD)



- Canutescu, A. A.; Dunbrack, R. L., Cyclic coordinate descent: Protein Sci 2003, 12, 963-972.

Loop Closure: Which Φ/Ψ angles connect the anchor points?

- Generate random conformation
- Close using Cyclic Coordinate Descent (CCD)



- Canutescu, A. A.; Dunbrack, R. L., Cyclic coordinate descent: Protein Sci 2003, 12, 963-972.

Cyclic Coordinate Descent: A Robotics Algorithm for Loop Closure

One angle at a time, simply minimizing the angle between final dihedral and anchor residue

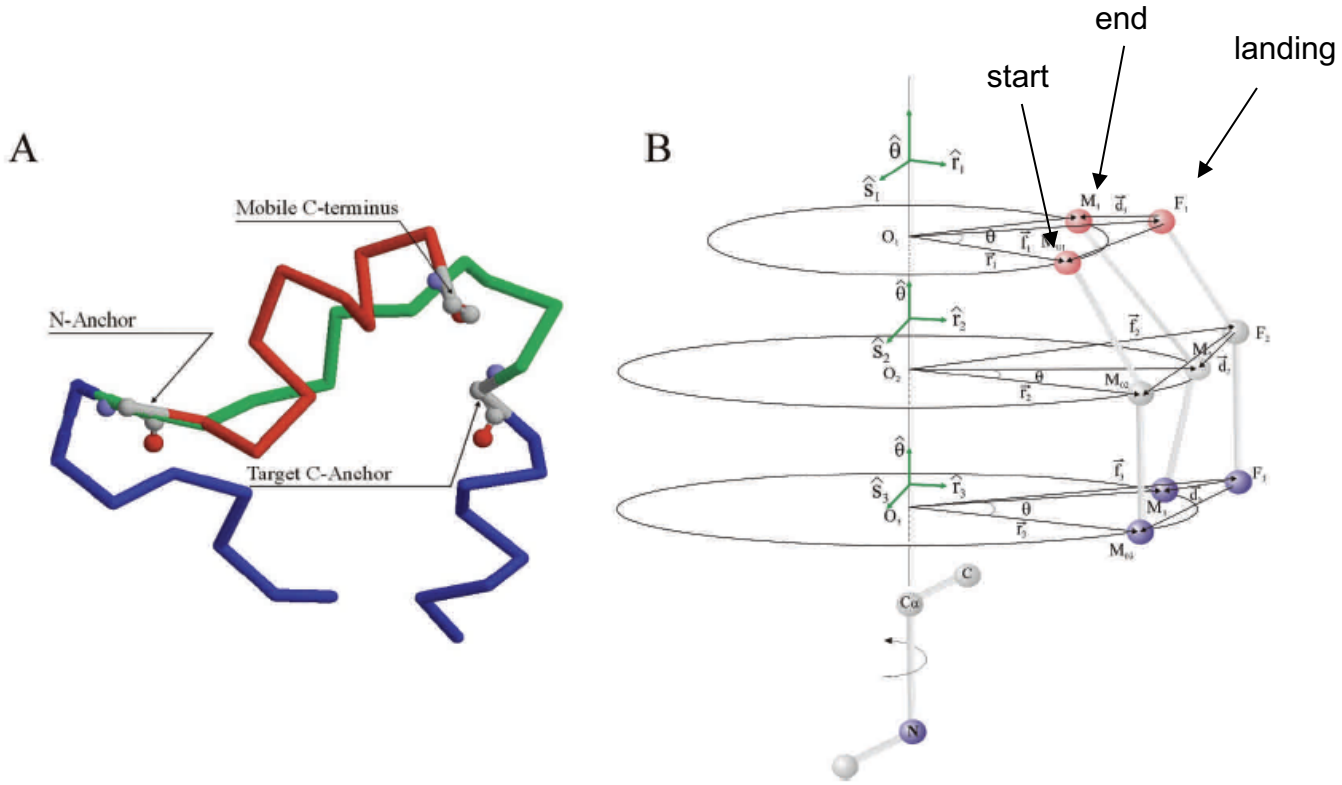
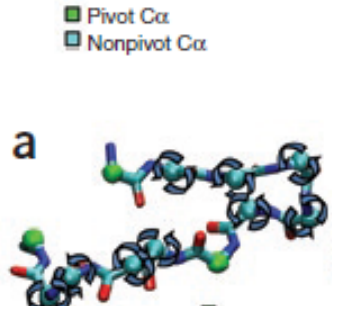


Figure 1. (*A*) C_α trace of a loop before (red) and after (green) closure with the flanking secondary structures (blue). The moving C-terminal anchor and the fixed C-terminal anchor are indicated. The loop closure problem is to adjust the dihedral angle degrees of freedom of the loop so that the moving C-terminal anchor is superimposed on the fixed C-terminal anchor. (*B*) Schematic of the CCD algorithm. Variables are defined in the text.

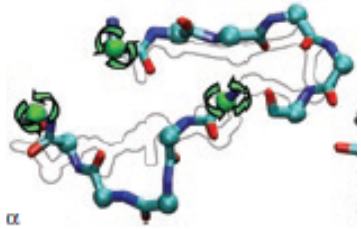
- Canutescu, A. A.; Dunbrack, R. L., Cyclic coordinate descent: Protein Sci 2003, 12, 963-972.

Kinematic Loop Closure (KIC)

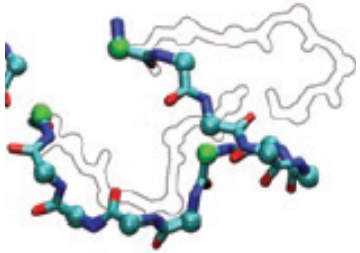
- 3 Cα atoms defined in N length chain as pivot atoms
- N-3 Cα atoms are non-pivot atoms

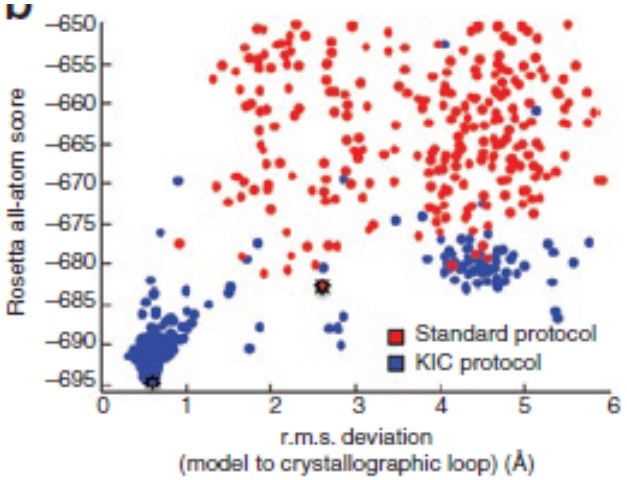


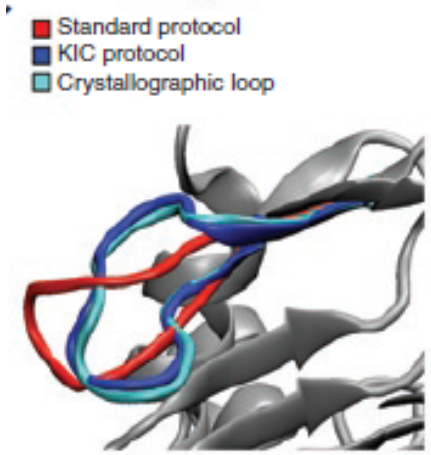
- Non-pivot atoms sampled based on Ramachandran-favored torsion angles (notice the chainbreak)



- Sample pivot angles until loop is closed







D. J. Mandell, E. A. Coutsias and T. Kortemme; "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling"; *Nat Methods*; 2009; Vol. 6 (8): p. 551-2.

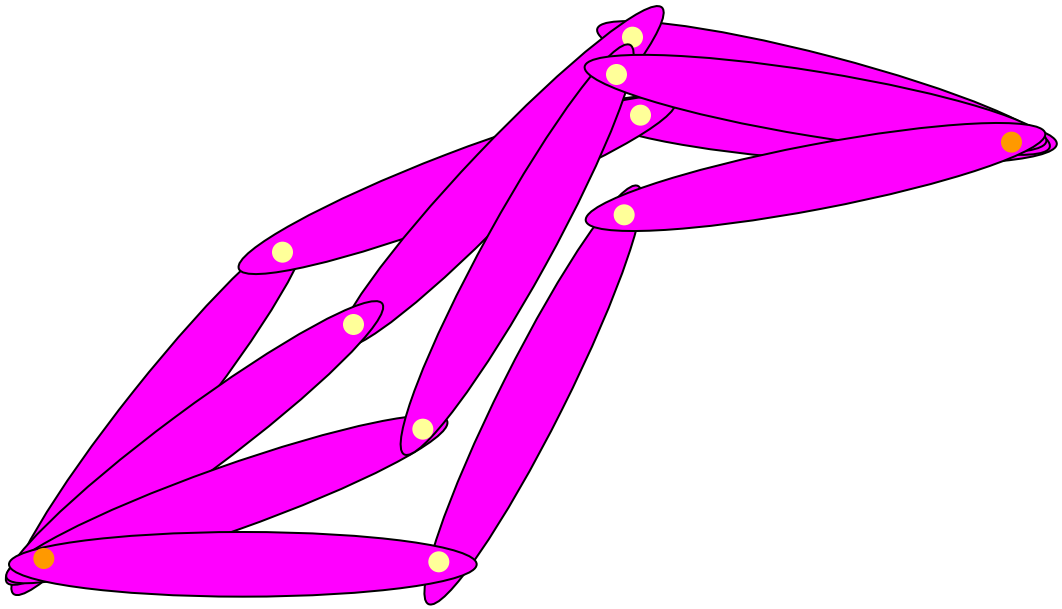
CCD vs. KIC

- CCD uses fragment insertion, originally KIC did not.
- FWIW: most people use KIC and newer implementations include GenKIC and KIC with fragments
- KIC has shown success in short loops (homology modeling gaps) and longer loops < 12 residues

These were strictly written as loop closure solutions, loop refinement is done with all-atom Rosetta energy function using KIC protocols.

Loop Refinement

- Minimize energy while retaining closure using all-atom energy refinement and smaller conformational sampling moves



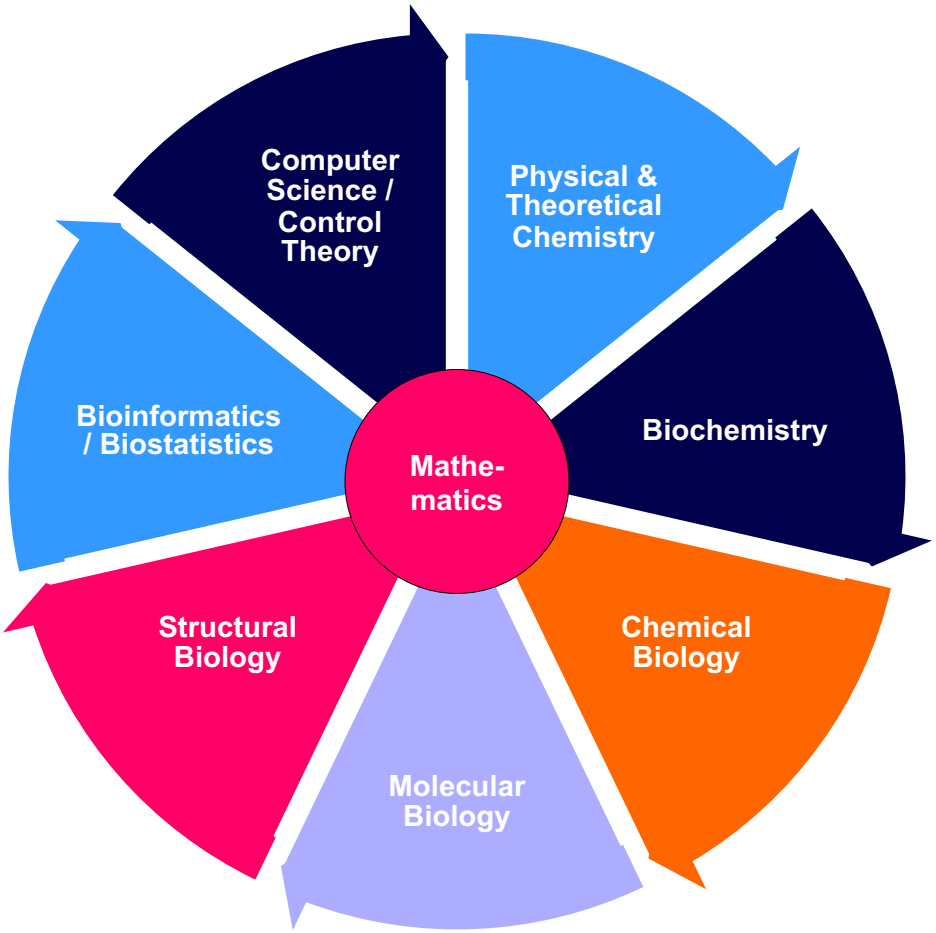
Loop modeling Recap

- Used in homology modeling—usually shorter loops where sequence lengths differ
- Used for longer loops important in protein-protein interactions

- CCD and KIC are both methods for loop closure

- Loop refinement done by all-atom scoring scheme

Computational Structural Biology



Questions?