

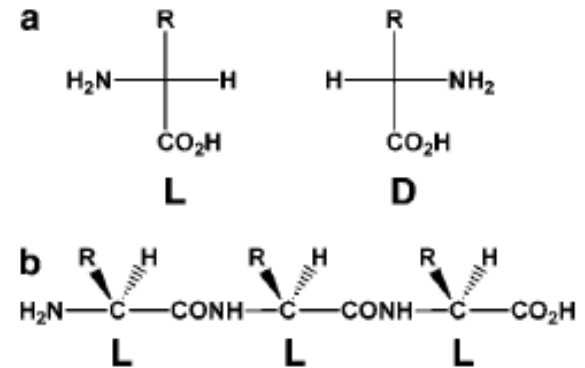
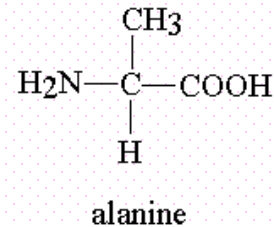
Computational Protein Design

Vibin Ramakrishnan

IIT Guwahati

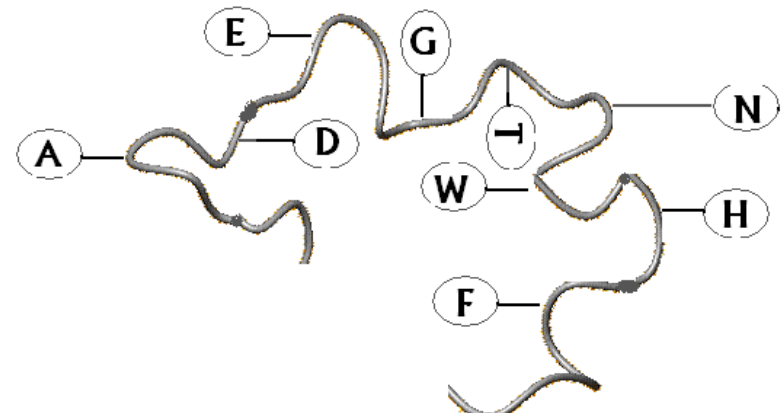
Protein is a (Hetero) polymer

Typical monomer

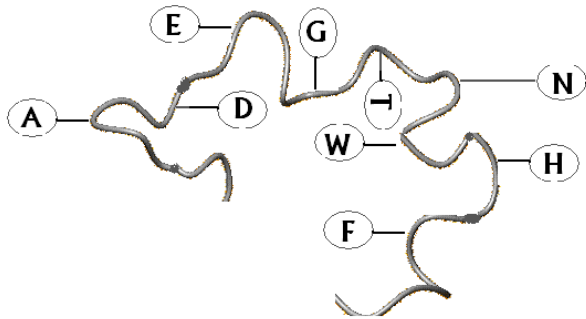


Amino acid sequence:

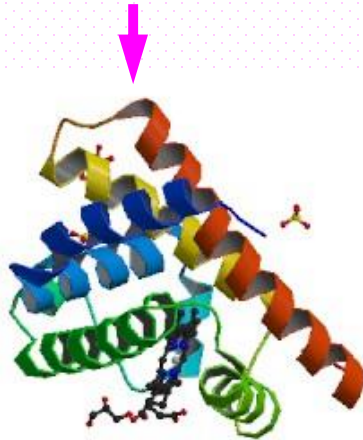
GLSDGEWQQVLNVWGKVEADIAGHGQEVLI
RLFTGHPETLEKFDKFKHLKTEAEMKASEDL
KKHGTVVLTALGGILKKKGHHEAELKPLAQS
HATKHKIPIKYLEFISDAIIHVLHSHKHPGDFGA
DAQGAMTKALELFRNDIAAKYKELGFQG



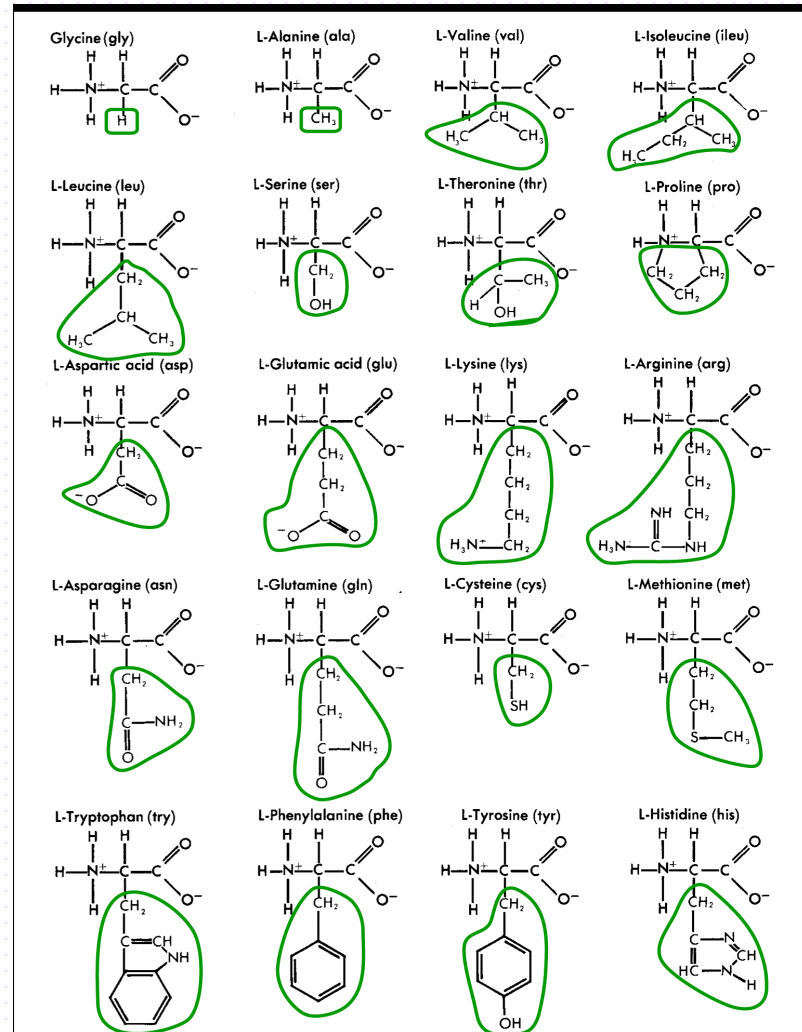
Folding of Protein to its prescribed structure



Non - Functional form



Functional form



Protein Design: Inverse of Protein Folding Problem

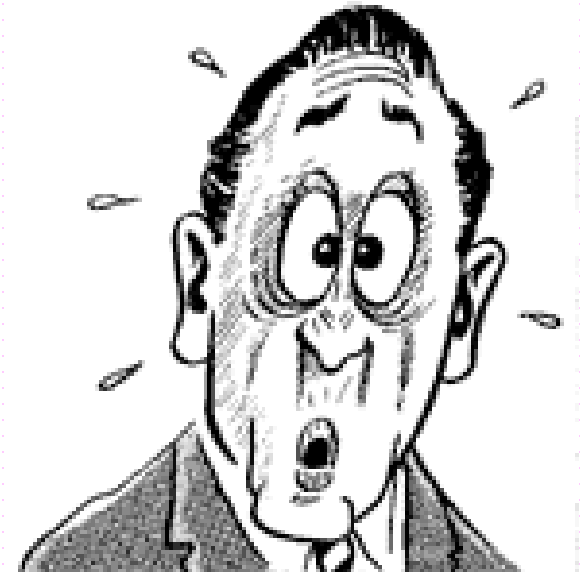
- Predicting Folds from Sequences → Protein-folding problem
- Predicting Sequences from Folds → 'Inverse Folding' Protein Design problem

Pre-requisites for protein design

- Main-chain Fold
- Side-chain Rotamer Library
- Energy Functions
- Search Algorithms

The Designers challenge

- Total no. of amino acids is 20
- For a 100 amino acid long protein 20^{100} sequence variants are theoretically possible.
- Total search space of 20^{100} variants
- For a Library of 500 rotamers, search space becomes 20^{500}



The Designers Advantage

- ❖ Protein Chain can have astronomical number of possible conformations.

-----Cyrus Levinthal (1969)

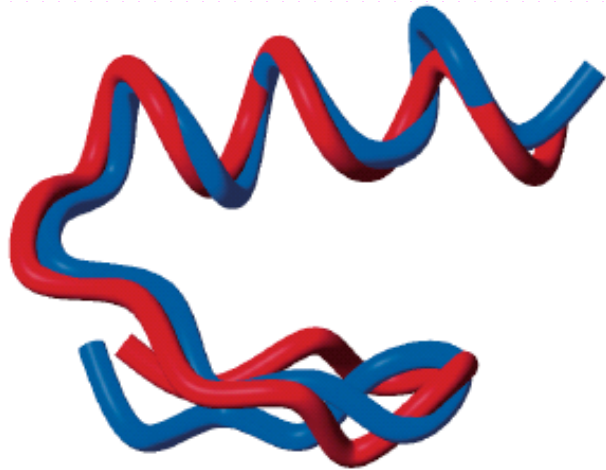
For any given fold there may be large number of sequences that are compatible with that fold.

- ❖ But number of protein folds discovered is about 10^3

Therefore, '*A given structure can accommodate many sequence variants*'.

- ❖ This is highly advantageous to a protein designer and is precisely the reason why inverse protein design problem is more tractable than the folding problem.

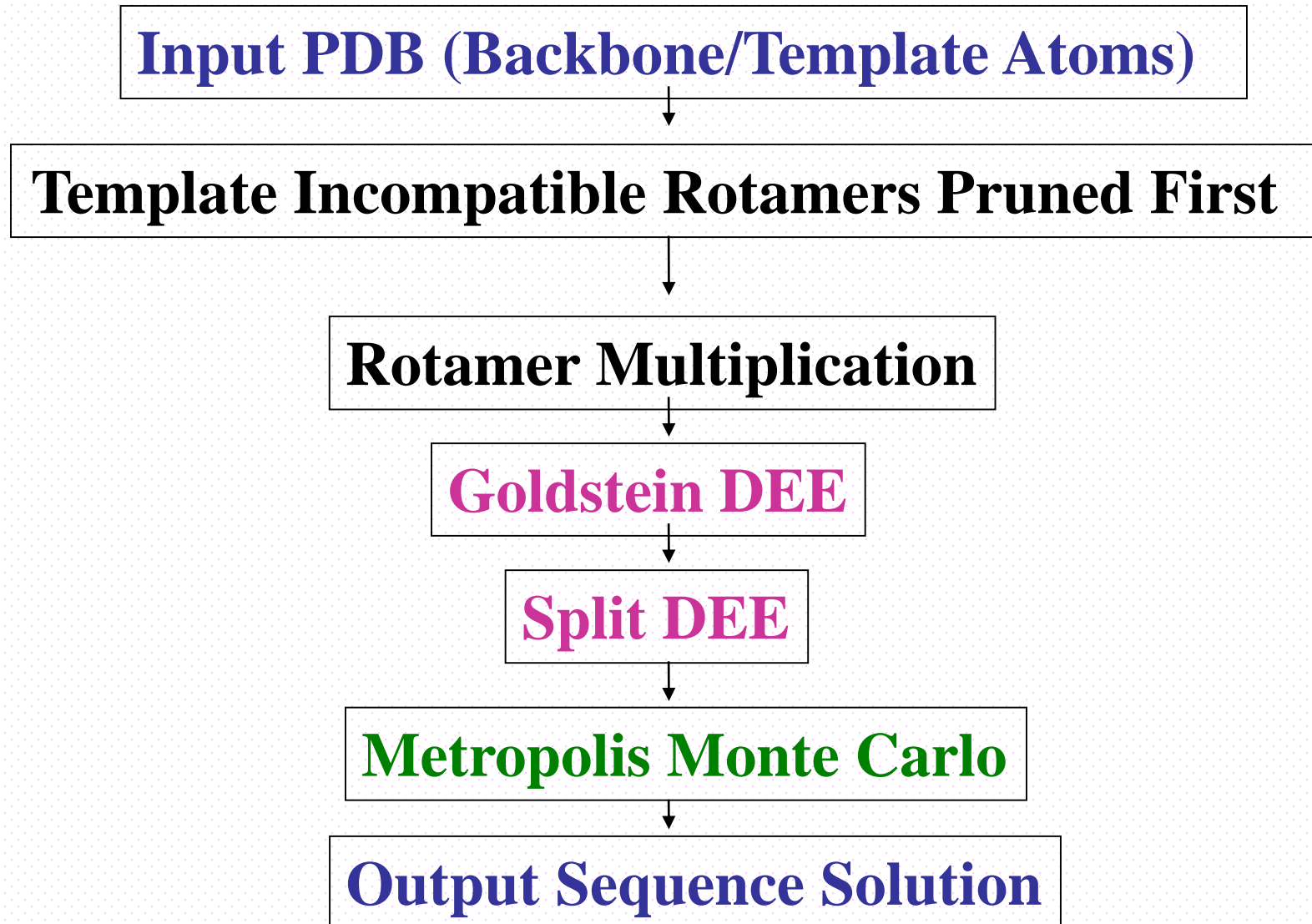
Main Chain Fold or Backbone



Steps for Protein Design on a chosen fold.

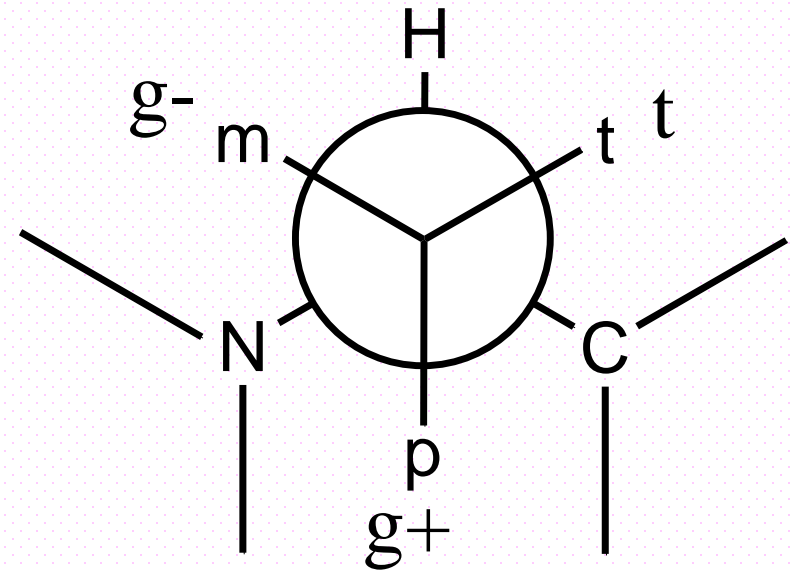
- Mutation of chosen/all sites
- Estimation of Energy
- Pruning by using an optimization algorithm

Sequence of events



Side-chain Rotamer Library

- Library of statistically significant conformers
- Backbone Dependent
- Backbone Independent
- Secondary Structure Dependent



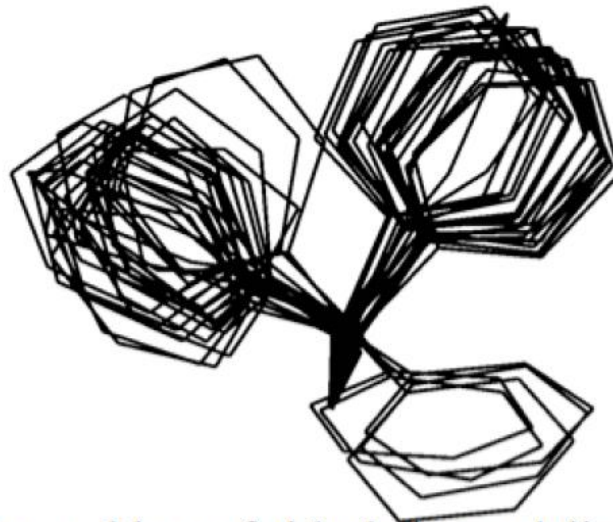
Rotamer

Dunbrack, R.L., Jr., Karplus, *J. Mol. Biol.* 230:543-571, 1993.

Lovell et. *al Proteins* 40: 389-408. (2000)

Sidechain angle space -- rotamers

A random sampling of Phenylalanine sidechains, when superimposed, fall into three classes: **rotamers**.



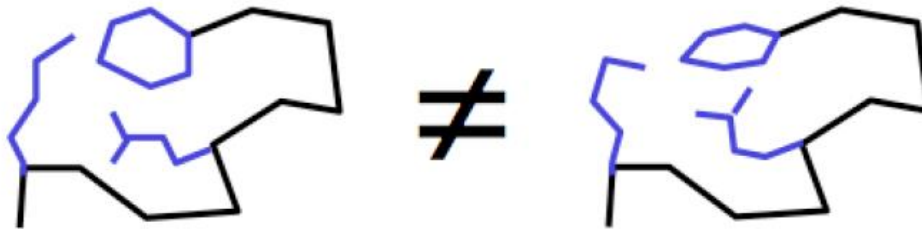
This simplifies the problem of sidechain modeling.

Dunbrack, R.L., Jr., Karplus, *J. Mol. Biol.* 230:543-571, 1993.

Lovell et. *al Proteins* 40: 389-408. (2000)

Sidechain modeling

Given a backbone conformation and the sequence, can we predict the sidechain conformations?



Energy calculations are sensitive to small changes. So the wrong sidechain conformation will give the wrong energy.


Dunbrack, R.L., Jr., Karplus, *J. Mol. Biol.* 230:543-571, 1993.

Lovell et. *al Proteins* 40: 389-408. (2000)

Energy Functions for Protein Design

- **Non-bonded interaction energy terms**

- van der Waals
- Electrostatics
- Solvation
- Hydrogen bonding
- Entropy


$$E_{vdW} = D_0 \left[\left(\frac{R_0}{R} \right)^{12} - 2 \left(\frac{R_0}{R} \right)^6 \right]$$


$$E_{elec} = 322.0637 \left(\frac{Q_i Q_j}{\epsilon R} \right)$$

Energy functions for protein design

D Benjamin Gordon*, Shannon A Marshall* and Stephen L Mayo†

Current Opinion in Structural Biology 1999, 9:509–513

Energy Functions for Protein Design

Solvation Energy

$$V_{\text{solvation}}(\mathbf{r}) = \sum_{i=1}^M \phi_i A_i(\mathbf{r})$$

$$\Delta E_{\text{Solvation}} = E_{\text{folded}} - E_{\text{reference}}$$

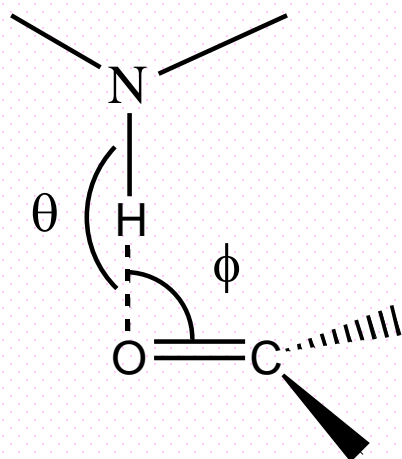
Hydrophobic burial (40% and more) → Reward 1.08784 kJ/mol

Aromatic hydrophobic amino acid exposure (40% or more) → Penalty 8.368 kJ/mol

Polar amino acid burial (more than 60%) → Penalty 4.184 kJ/mol

Energy Functions for Protein Design

Hydrogen bonding



$$E_{HB} = D_0 \left[5 \left(\frac{R_0}{R} \right)^{12} - 6 \left(\frac{R_0}{R} \right)^{10} \right] F(\theta)$$

$$\begin{array}{ll} sp^3 \text{ donor} - sp^3 \text{ acceptor} & F = \cos^2 \theta \cos^2 (\phi - 109.5) \\ & \theta > 90^\circ, \phi - 109.5^\circ < 90^\circ \end{array} \quad (3)$$

$$\begin{array}{ll} sp^3 \text{ donor} - sp^2 \text{ acceptor} & F = \cos^2 \theta \cos^2 \phi \\ & \phi > 90^\circ \end{array} \quad (4)$$

$$sp^2 \text{ donor} - sp^3 \text{ acceptor} \quad F = \cos^4 \theta \quad (5)$$

$$sp^2 \text{ donor} - sp^2 \text{ acceptor} \quad F = \cos^2 \theta \cos^2 (\max[\phi, \varphi]) \quad (6)$$

Energy functions for protein design

D Benjamin Gordon*, Shannon A Marshall* and Stephen L Mayo†

Current Opinion in Structural Biology 1999, 9:509–513

Energy Functions for Protein Design

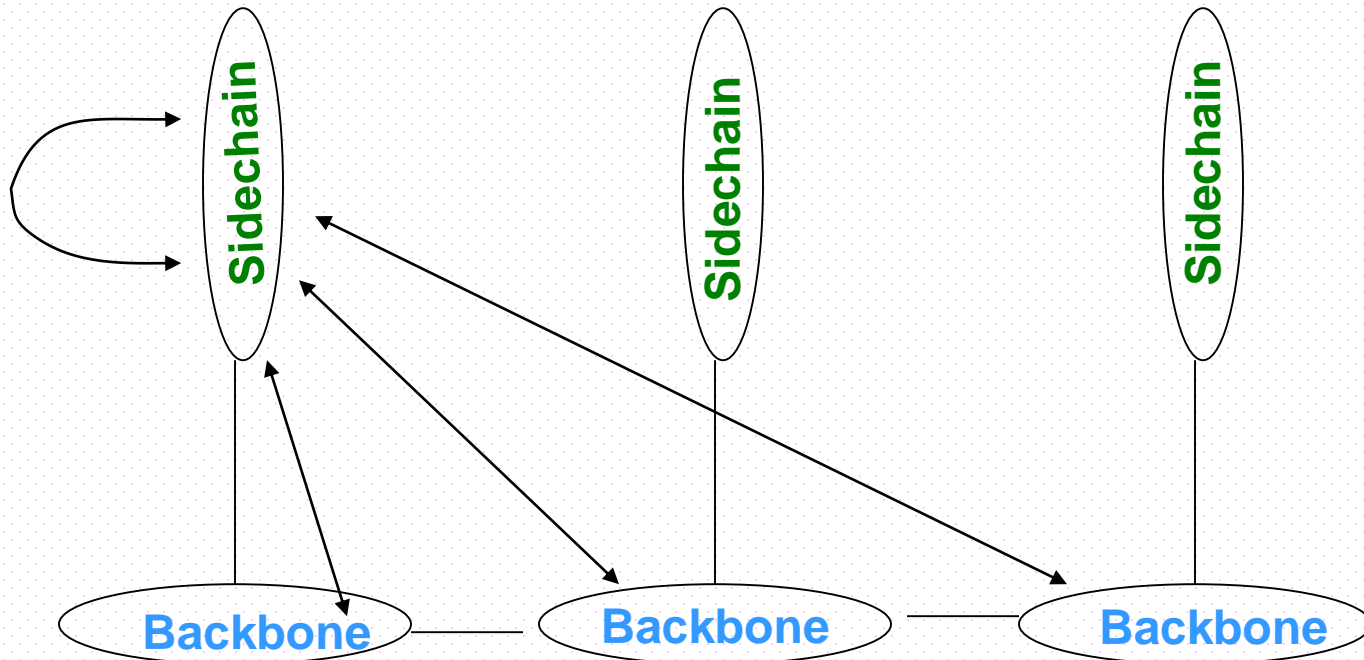
Entropy

$$\Delta G_s = RT \ln \Omega$$

Amino Acid	Omega (kcal/mol)	Amino Acid	Omega (kcal/mol)
Ala	0.2	Glu	2.2
Arg	2.6	His	1.9
Asn	1.9	Ile	1.7
Asp	1.5	Leu	1.7
Cys	0.7	Lys	3.3
Gln	2.6	Met	5.0
Phe	0.9	Ser	0.7
Thr	0.9	Trp	1.3
Tyr	0.9	Val	1.1

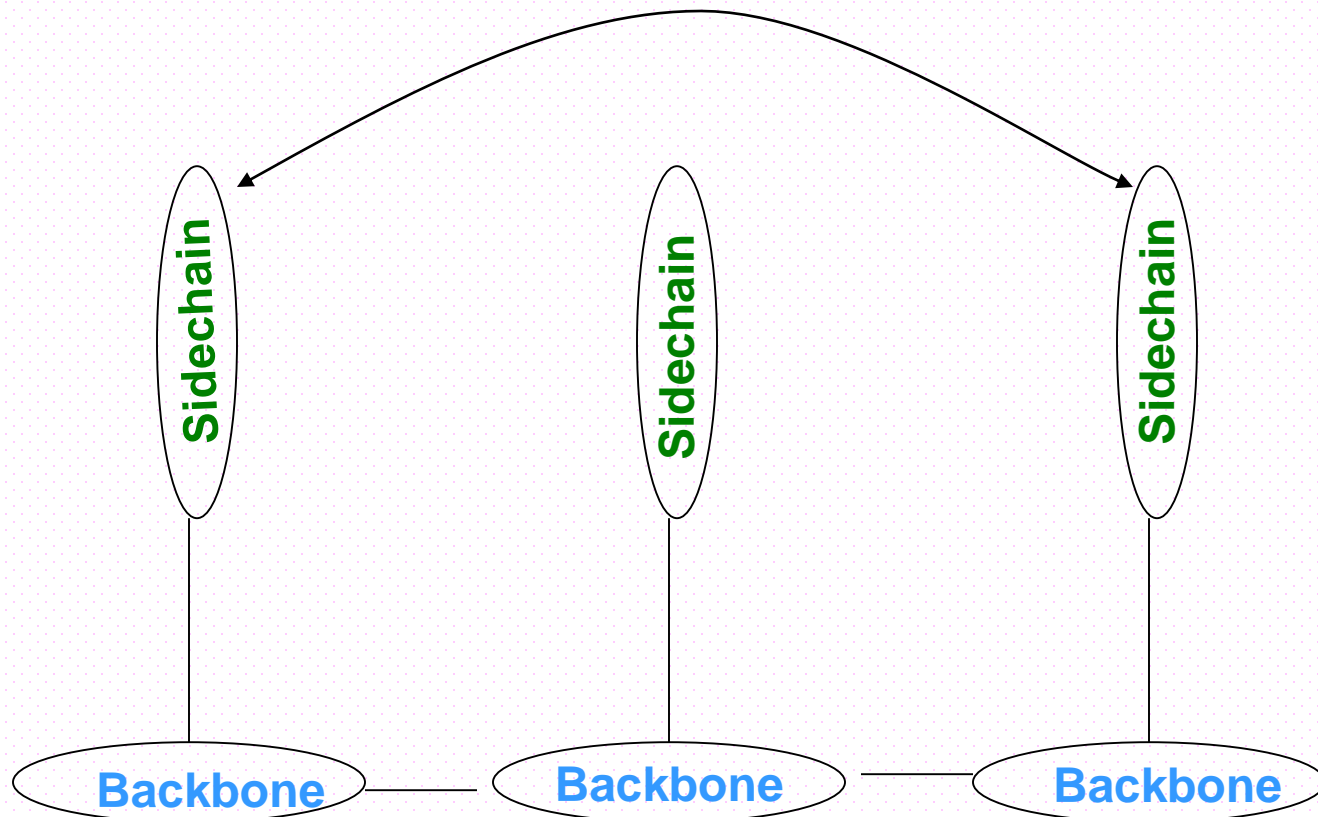
Energy Decomposition

$E(i_r)$ **Self Energy:** Interaction energies between the atoms of rotamer r of side chain i , plus the interaction energies of the atoms of rotamer r with all the backbone atoms.



Energy Decomposition

$E(i_r, j_u)$ **Interaction Energy**: Interaction energy between atoms of rotamer r of side chain i and atoms of rotamer u of side chain j



Energy Decomposition

If a rotamer has been chosen for each side chain, then the potential energy of the protein is

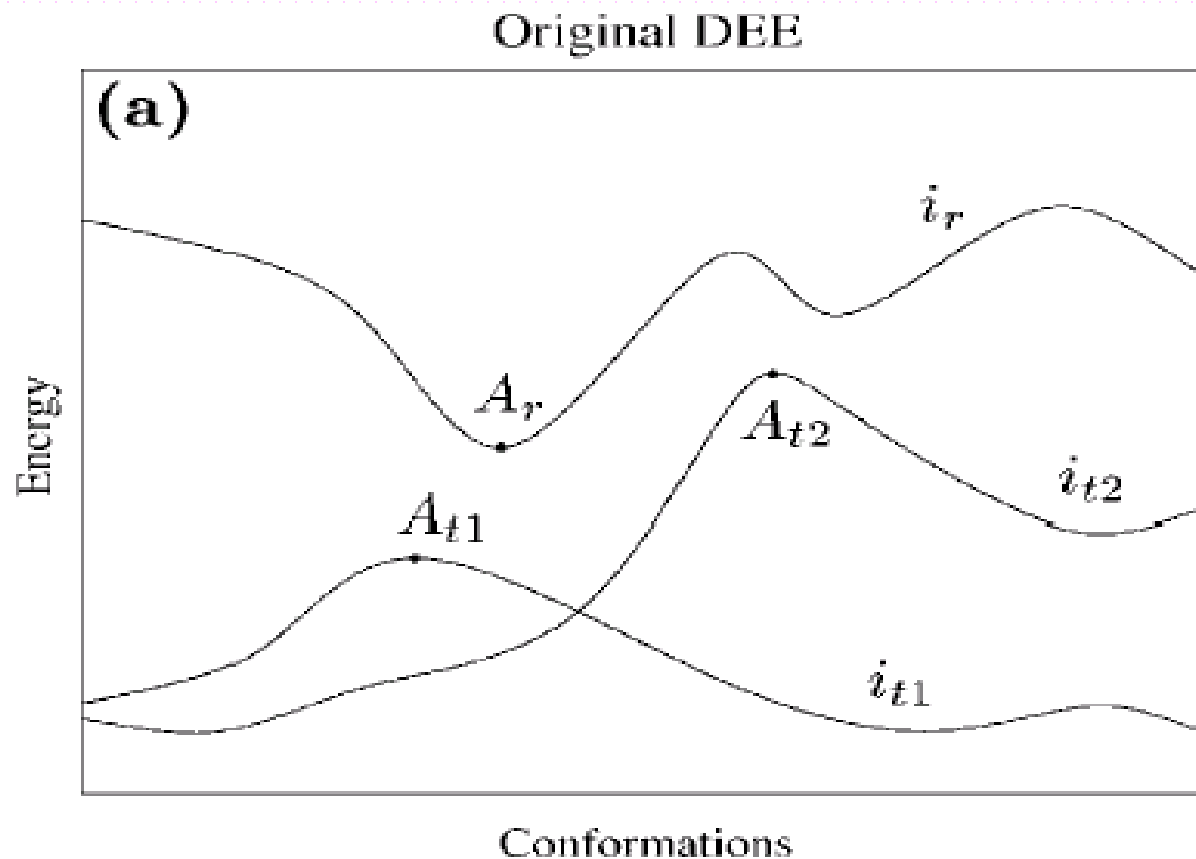
$$E = E_{bb} + \sum_{i=1}^{N_{sch}} \overbrace{E(i_r)}^{\text{self energies}} + \sum_{i=1}^{N_{sch}-1} \sum_{j=i+1}^{N_{sch}} \overbrace{E(i_r, j_u)}^{\text{Interaction energies}}$$

N_{sch} = number of side chains

Optimization Algorithms

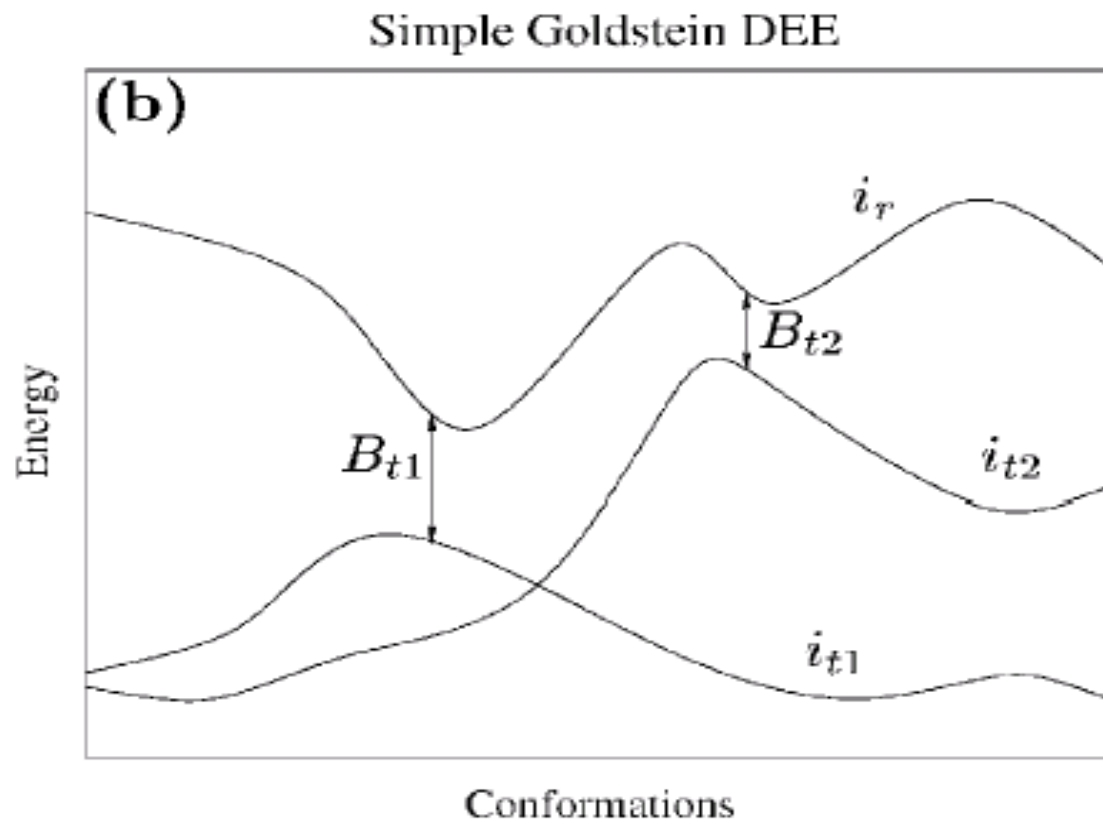
- **Large variable space** (Combinatorial Explosion)
For a small protein of 100 amino acids, number of sequences that are possible when 20 naturally occurring amino acids are considered is 20^{100} ($\sim 10^{130}$)
- ✓ **Systematic (deterministic) search**
 - Dead End Elimination, Branch and Bound
 - **Random search**
 - Genetic Algorithms, Monte Carlo, Simulated Annealing
- Each method has its own strengths and weaknesses.

Optimization by Dead End Elimination (DEE)



$$E(i_r) + \sum_{j, j \neq i} \min_u E(i_r, j_u) > E(i_t) + \sum_{j, j \neq i} \max_u E(i_t, j_u)$$

Goldstein DEE

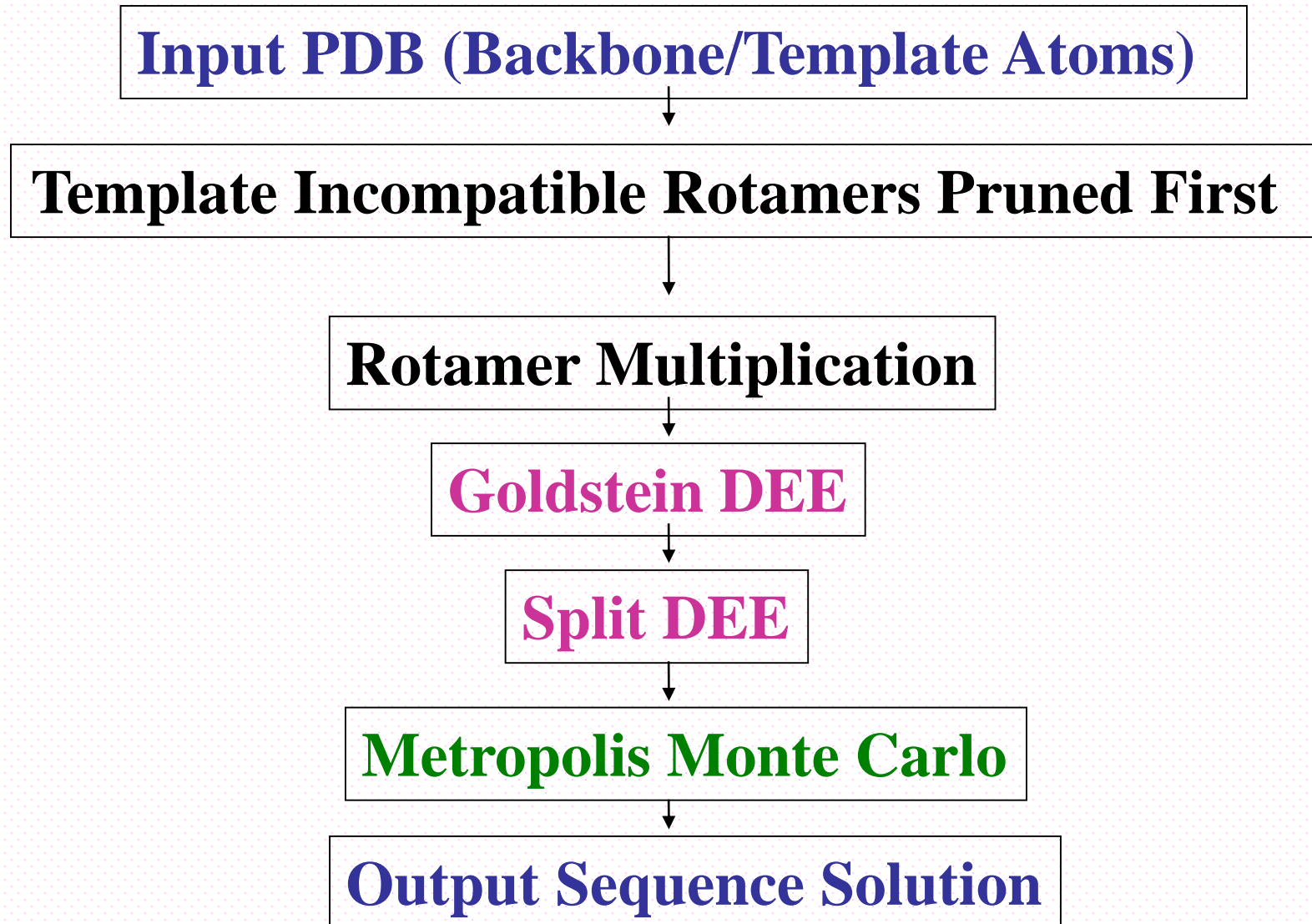


$$E(i_r) - E(i_t) + \sum_{j, j \neq i} \left\{ \min_u [E(i_r j_u) - E(i_t j_u)] \right\} > 0$$

Metropolis Monte carlo

$\epsilon_1 \leq \epsilon_0$	$\epsilon_1 > \epsilon_0$ $e^{-(\epsilon_1 - \epsilon_0)/kT} < R$	$\epsilon_1 > \epsilon_0$ $e^{-(\epsilon_1 - \epsilon_0)/kT} > R$
The energy of State-1 is less than or equal to State-0. This means that the new state is accepted and becomes the new State-0.	The energy of State-1 is greater than State-0, but the energy difference is small enough that it is probabilistically accepted. This means that the new state is accepted and becomes the new State-0.	The energy of State-1 is greater than State-0, and the energy difference is large enough that it is probabilistically rejected. This means that the system stays in State-0.

Sequence of events

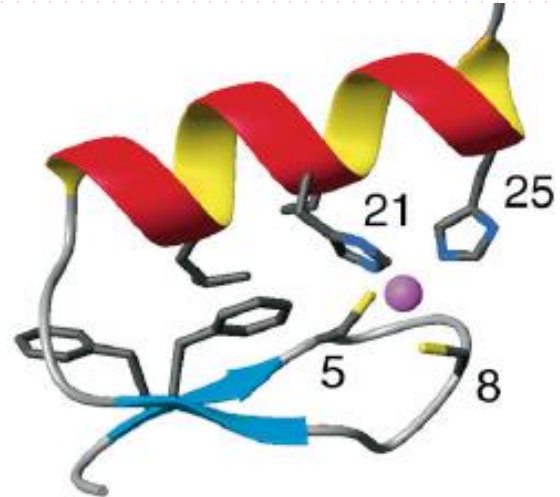


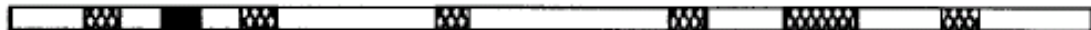
Case Study 1: Full Protein Design FSD -1

BBA motif of Zinc Finger Domain

Search space of 1.9×10^{27} possible sequences

Sequence Solution



	5	10	15	20	25																				
																									
FSD-1	Q Q Y T A K I K G R T F R N E K E L R D F I E K F K G R																								
Zif268	K P F Q C R I C M R N F S R S D H L T T H I R T H T G E																								

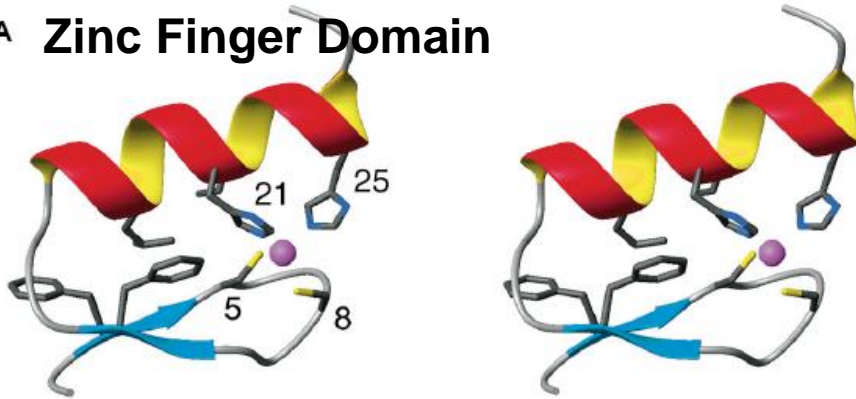
De Novo Protein Design: Fully Automated Sequence Selection

Bassil I. Dahiyat† and Stephen L. Mayo*

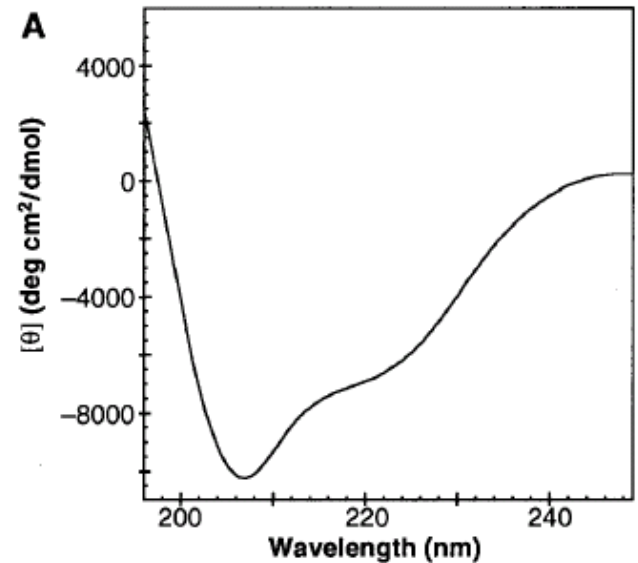
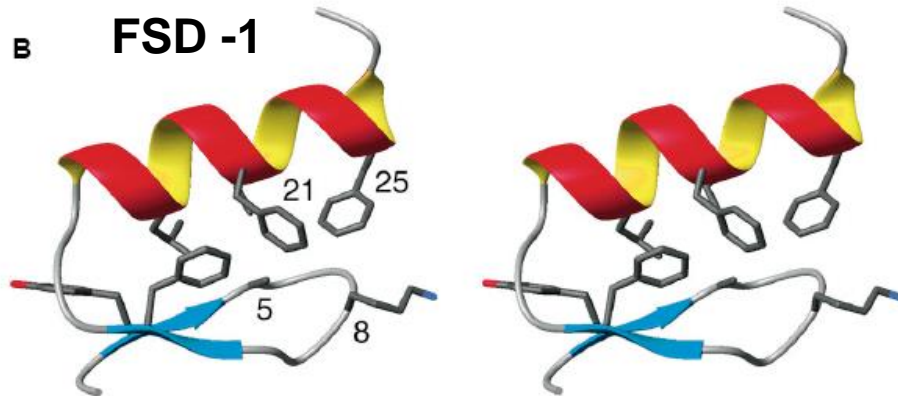
SCIENCE • VOL. 278 • 3 OCTOBER 1997

Case Study 1: Full Protein Design FSD -1

A Zinc Finger Domain



B FSD -1



CD Spectrum of the designed protein

De Novo Protein Design: Fully Automated Sequence Selection

Bassil I. Dahiya† and Stephen L. Mayo*

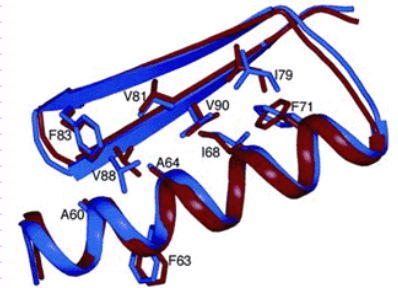
SCIENCE • VOL. 278 • 3 OCTOBER 1997

Case Study 2

Top7: De novo designed Artificial Protein

(Top7 is a 93 residue α/β protein)

Top7 Design Protocol



The target topology was selected first, totally new artificial structure
A starting sequence was designed for each structure by searching an amino acid rotamer library

- All amino acids except cysteine were allowed at 71 sites. The remaining 22 positions are on the sheet surfaces and were restricted to polar amino acids

Kuhlman B., G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker.
"Design of a novel globular protein fold with atomic-level accuracy."
Science 302, no. 5649 (Nov 21, 2003): 1364-8.

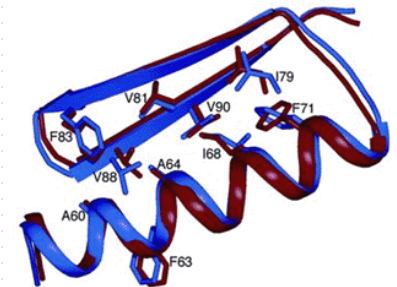
Case Study 2

Top7: De novo designed Artificial Protein

Top7 Design Protocol

Backbone optimization: Monte Carlo minimization to alter the structure so as to accommodate existing sequence

- Sequence optimization: Rotamers of new amino acids explored for low-energy side-chain packing
- 15 alternating cycles of each gives a final energy-minimized structure



Kuhlman B., G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker.
"Design of a novel globular protein fold with atomic-level accuracy."
Science 302, no. 5649 (Nov 21, 2003): 1364-8.

Case Study 2

Top7: De novo designed Artificial Protein

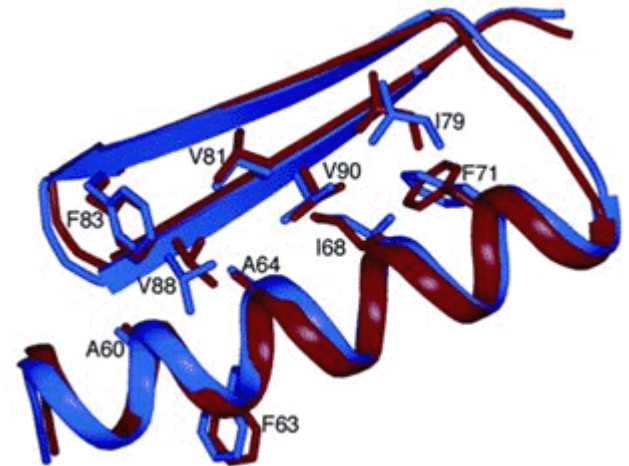
Top7 Design Protocol

Characteristics of Final structure :

Final backbone model is only 1.1Å different from starting model

Only 31% of residues are retained from initial to final design

Synthesized protein is thermally stable upto 98°C



Kuhlman B., G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker.
"Design of a novel globular protein fold with atomic-level accuracy."
Science 302, no. 5649 (Nov 21, 2003): 1364-8.