

---

# Ab Initio Methods for Protein Structure Prediction

Slides from Shuai C., Li at University of  
Waterloo

# Motivation

---

- homology modeling
  - No knowledge about the physical nature of the protein folding and stability.
- ab-initio methods can
  - augment fold-recognition and homology (refinement, large loops, side chains).
- it can ease experimental structure determination.
- It can find new folds

# Ab Initio Methods

---

- Ab initio: “From the beginning”.
- Assumption
  - All the information about the structure of a protein is contained in its sequence of amino acids.
  - The structure that a (globular) protein folds into is the structure with the lowest free energy.
  - The native structure is contained in the search space
- Finding native-like conformations require
  - A scoring function (potential).
  - A search strategy.

# ab-initio protein structure prediction

---

## ■ **Optimization problem**

- Define some initial model.
- Define a function mapping structures to numerical values (the lower the better).
- Solve the computational problem of finding the global minimum.

## ■ **Simulation of the actual folding process**

- Build an accurate initial model (including energy and forces).
- Accurately simulate the dynamics of the system.
- The native structure will emerge.
- **No hope due to large search space**

# Energy Minimization (Theory)

---

- Treat Protein molecule as a set of balls (with mass) connected by rigid rods and springs
- Rods and springs have empirically determined force constants
- Allows one to treat atomic-scale motions in proteins as classical physics problems

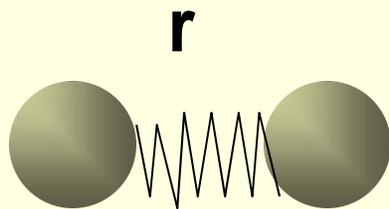
# Standard Energy Function

---

$$\begin{aligned} E = & K_r(r_i - r_j)^2 + && \text{Bond length} \\ & K_\theta(\theta_i - \theta_j)^2 + && \text{Bond bending} \\ & K_\phi(1 - \cos(n\phi_j))^2 + && \text{Bond torsion} \\ & q_i q_j / 4\pi\epsilon r_{ij} + && \text{Coulomb} \\ & A_{ij}/r^6 - B_{ij}/r^{12} + && \text{van der Waals} \\ & C_{ij}/r^{10} - D_{ij}/r^{12} && \text{H-bond} \end{aligned}$$

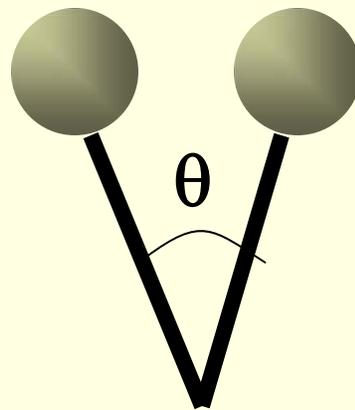
# Energy Terms

---



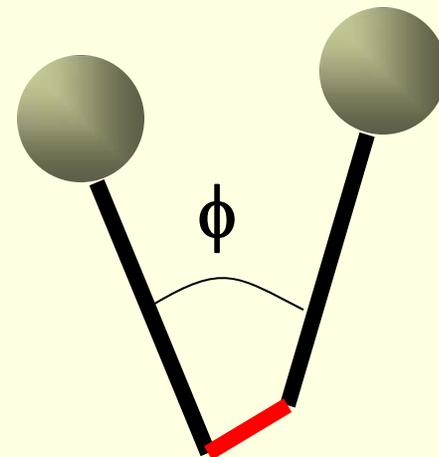
$$K_r(r_i - r_j)^2$$

**Stretching**



$$K_\theta(\theta_i - \theta_j)^2$$

**Bending**

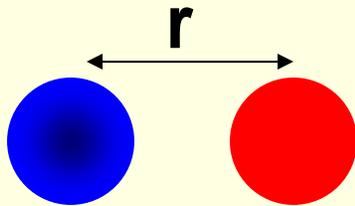


$$K_\phi(1 - \cos(n\phi_j))^2$$

**Torsional**

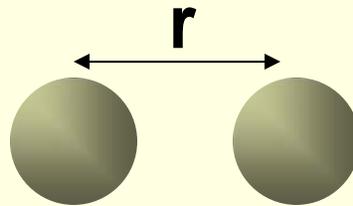
# Energy Terms

---



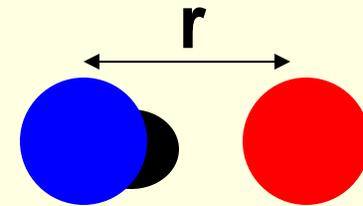
$$q_i q_j / 4\pi\epsilon r_{ij}$$

**Coulomb**



$$A_{ij}/r^6 - B_{ij}/r^{12}$$

**van der Waals**



$$C_{ij}/r^{10} - D_{ij}/r^{12}$$

**H-bond**

# Reduced complexity models

---

- No side chains
  - sometimes no main chain atoms either
  - Or represent the side chain with  $C_{\beta}$
- Reduced degrees of freedom
- On-or off-lattice
- Generally have an environment -based score and a knowledge-based residue-residue interaction term
- Sometimes used as first step to prune the enormous conformational space, then resolution is increased for later fine-tuning

# Basic element

electrons &  
protons

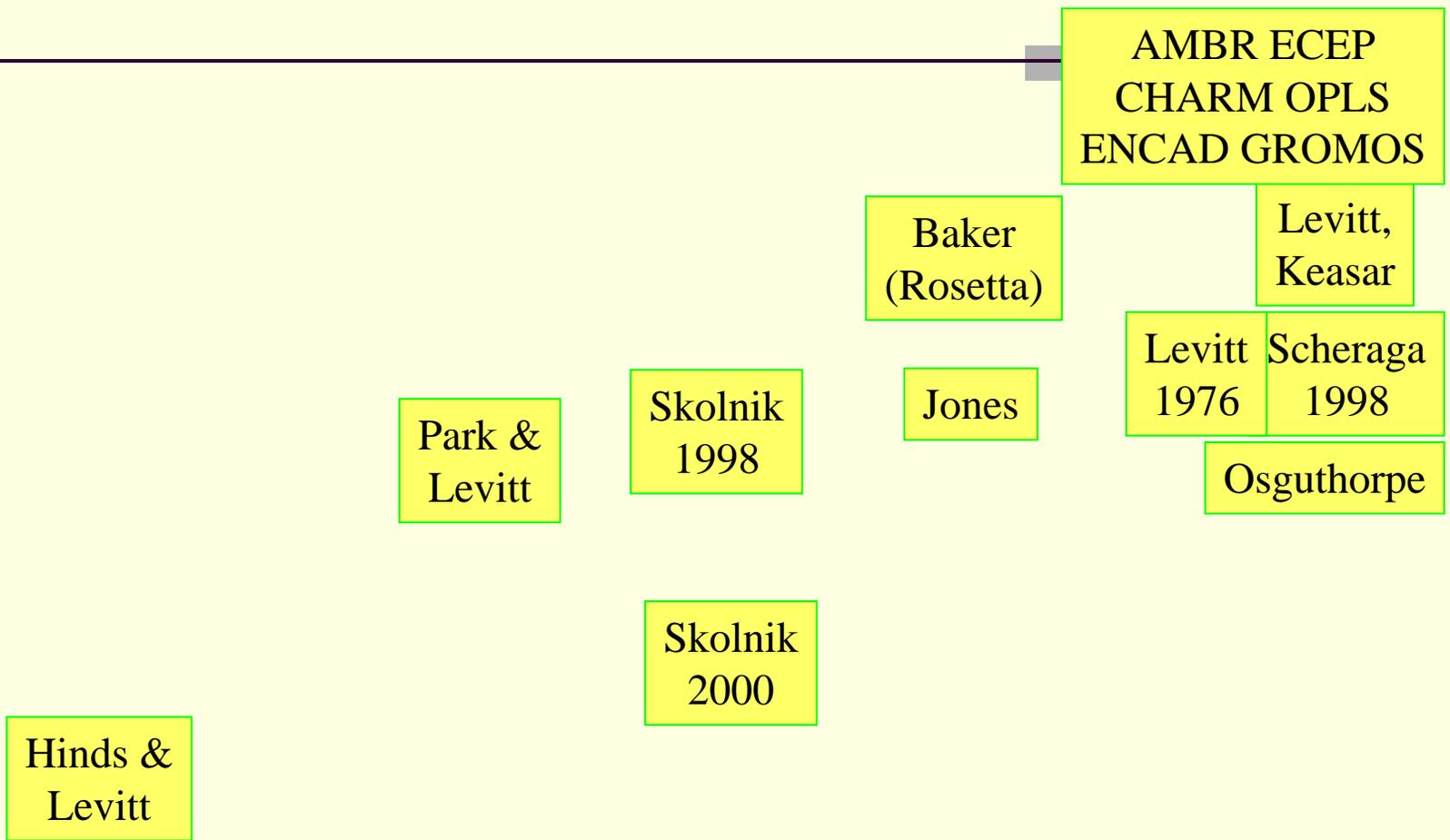
atom

extended  
atom

half a  
residue

residue

Some  
residues



diamond  
lattice

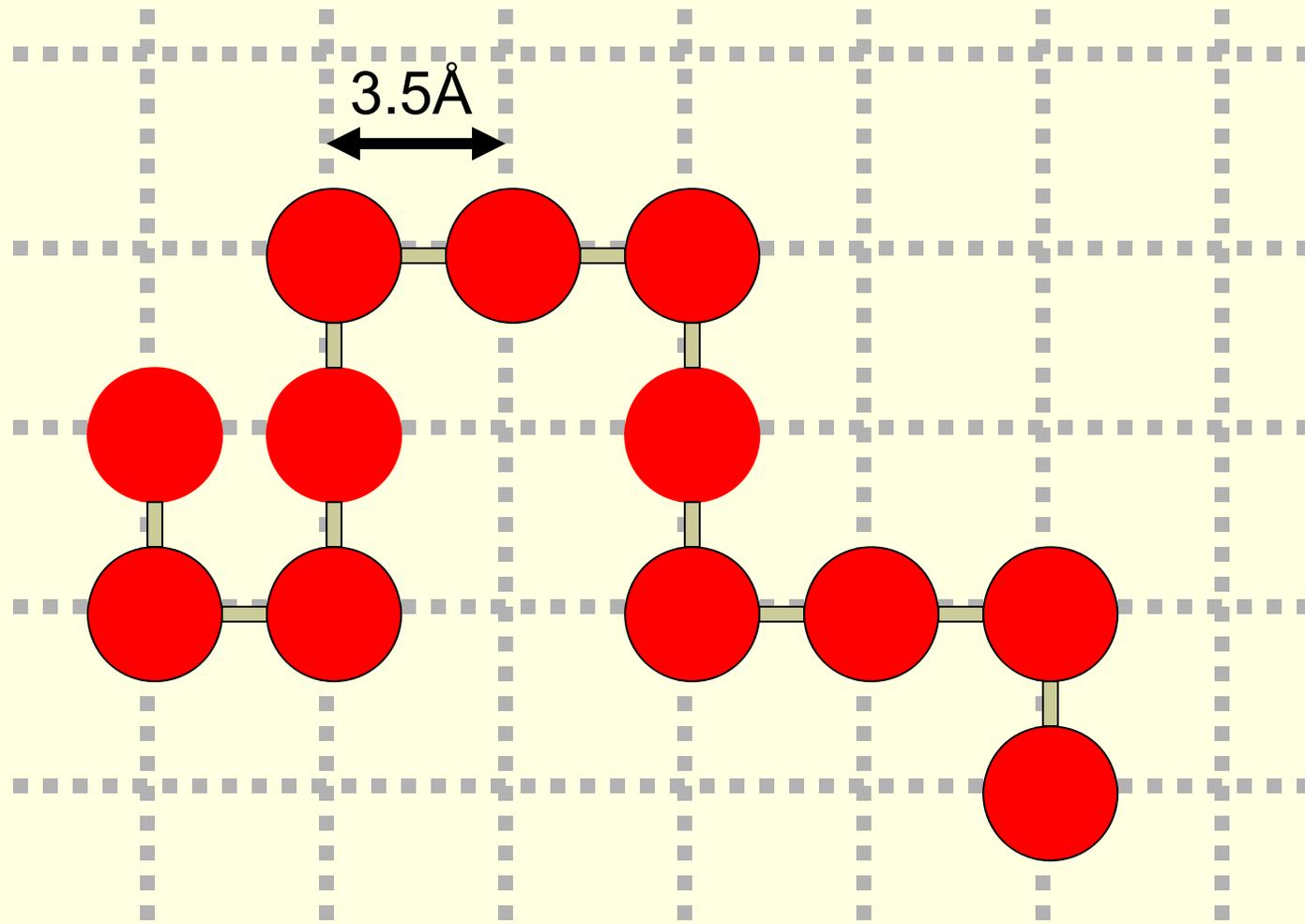
torsion  
angle lattice

fine square  
lattice

fragments

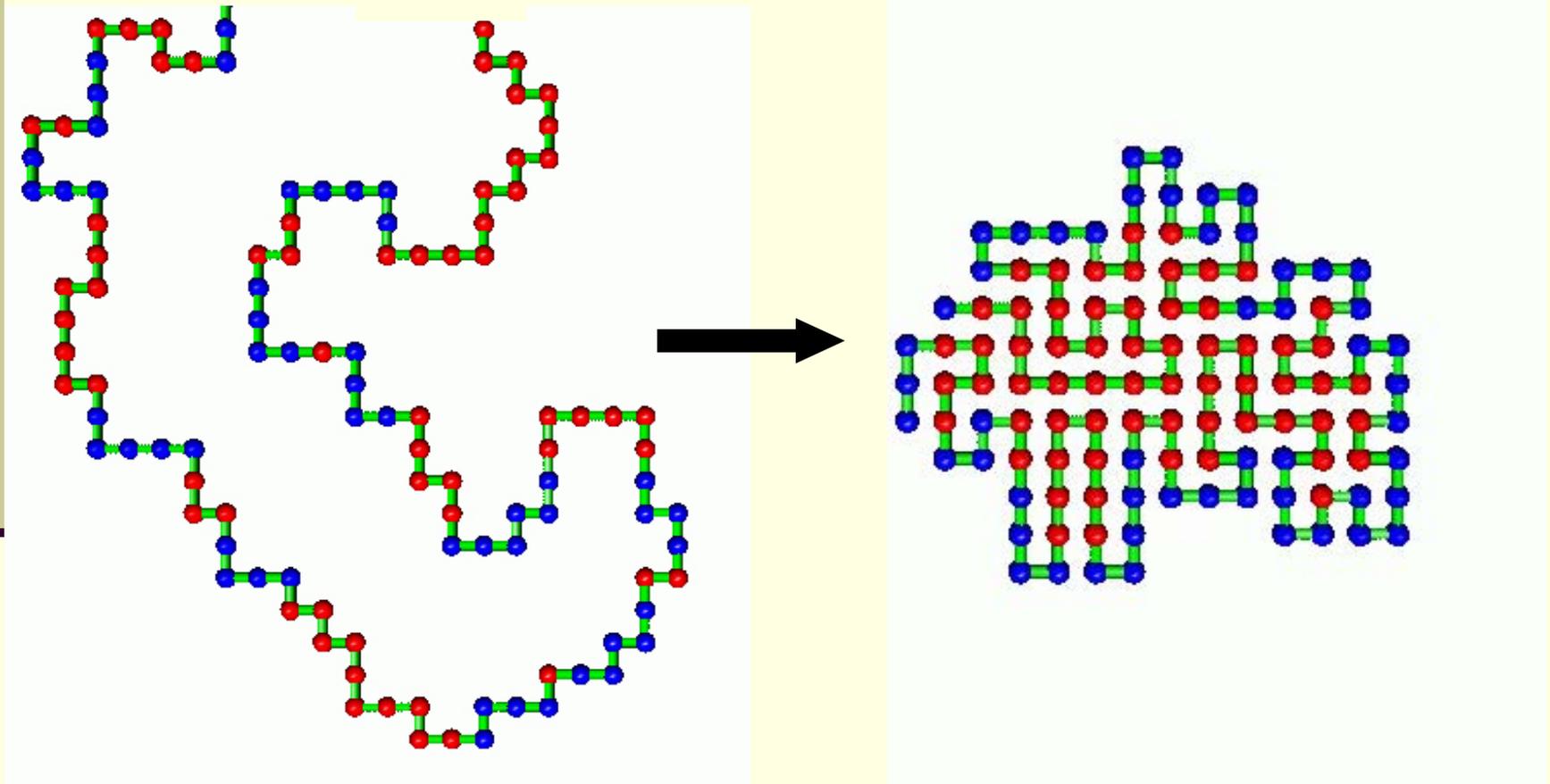
continuous

# A Simple 2D Lattice



# Lattice Folding

---



# Lattice Algorithm

---

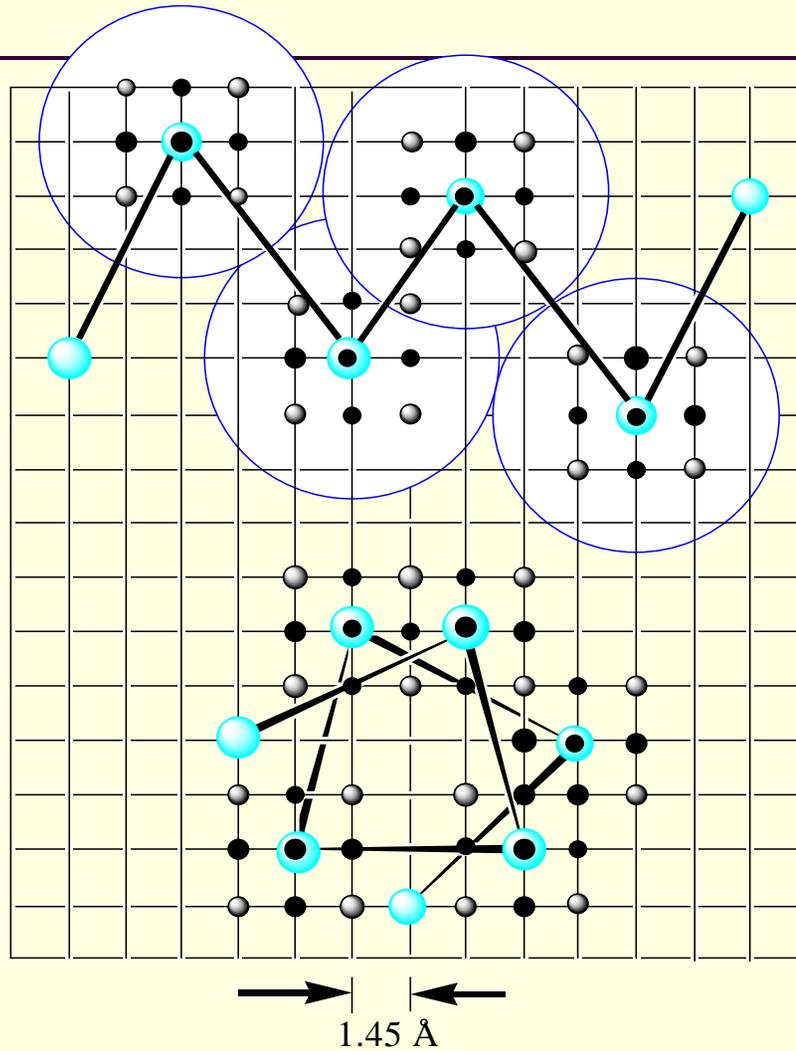
- *Build a “n x m” matrix (a 2D array)*
- *Choose an arbitrary point as your N terminal residue (start residue)*
- *Add or subtract “1” from the x or y position of the start residue*
- *Check to see if the new point (residue) is off the lattice or is already occupied*
- *Evaluate the energy*
- *Go to step 3) and repeat until done*

# Lattice Energy Algorithm

---

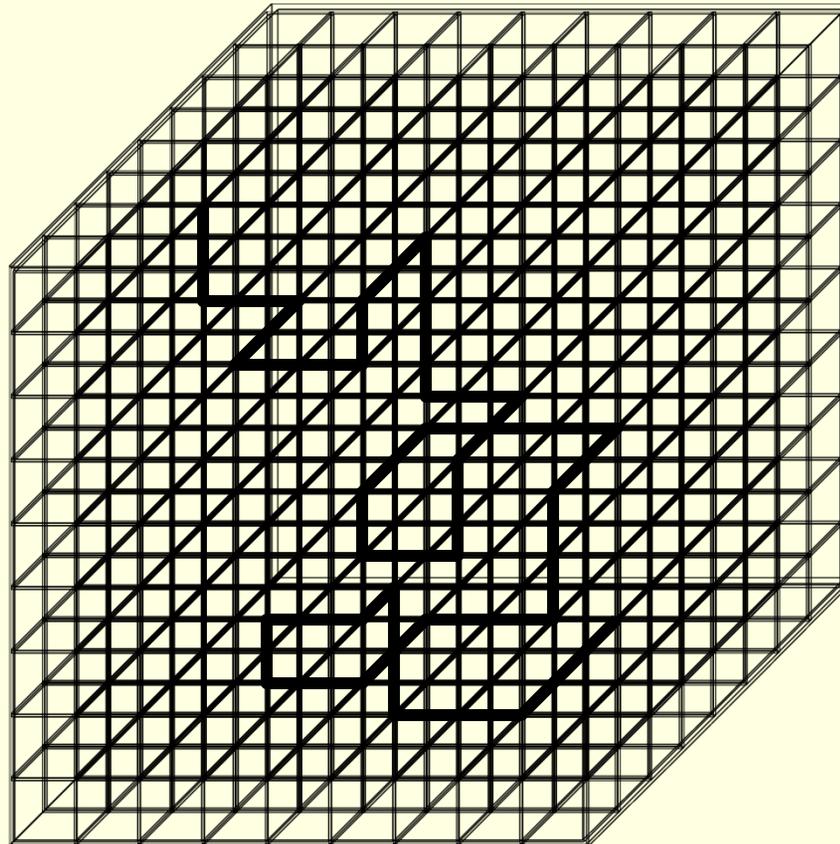
- *Red = hydrophobic, Blue = hydrophilic*
- *If Red is near empty space  $E = E+1$*
- *If Blue is near empty space  $E = E-1$*
- *If Red is near another Red  $E = E-1$*
- *If Blue is near another Blue  $E = E+0$*
- *If Blue is near Red  $E = E+0$*

# More Complex Lattices



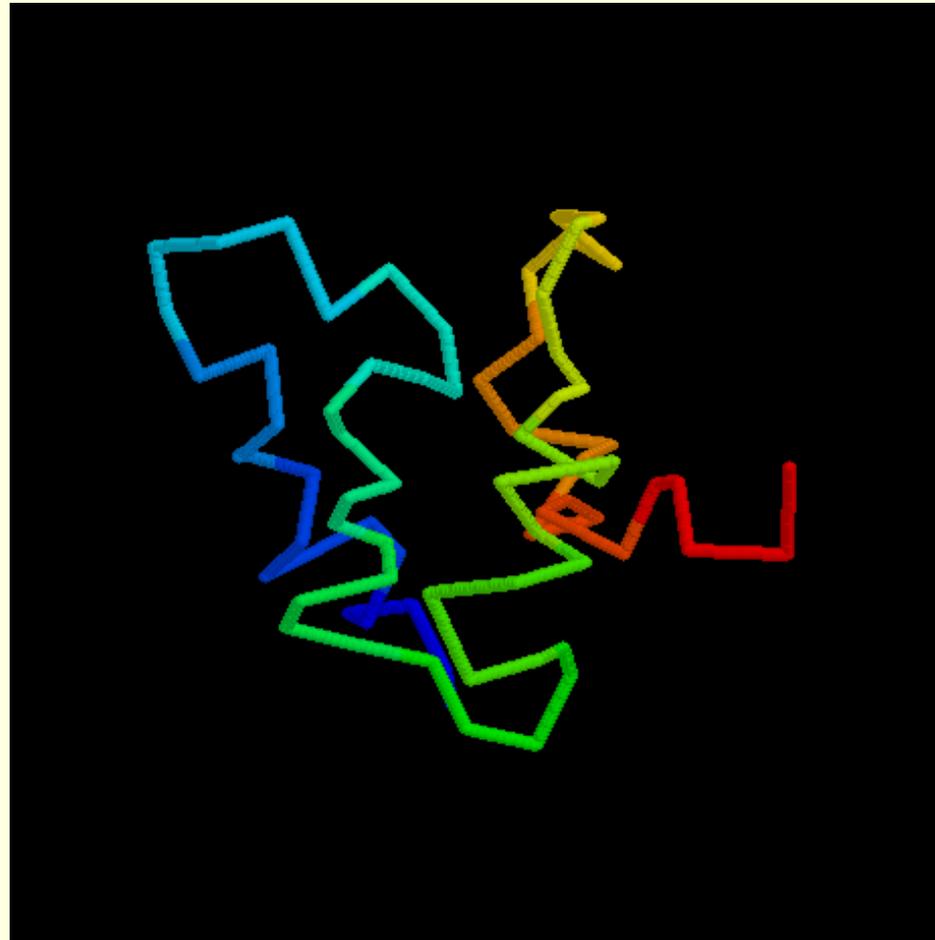
# 3D Lattices

---



# Really Complex 3D Lattices

---



J. Skolnick

# Lattice Methods

---

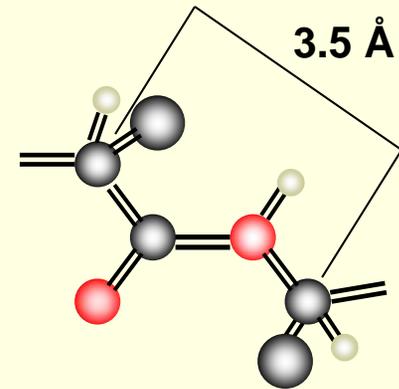
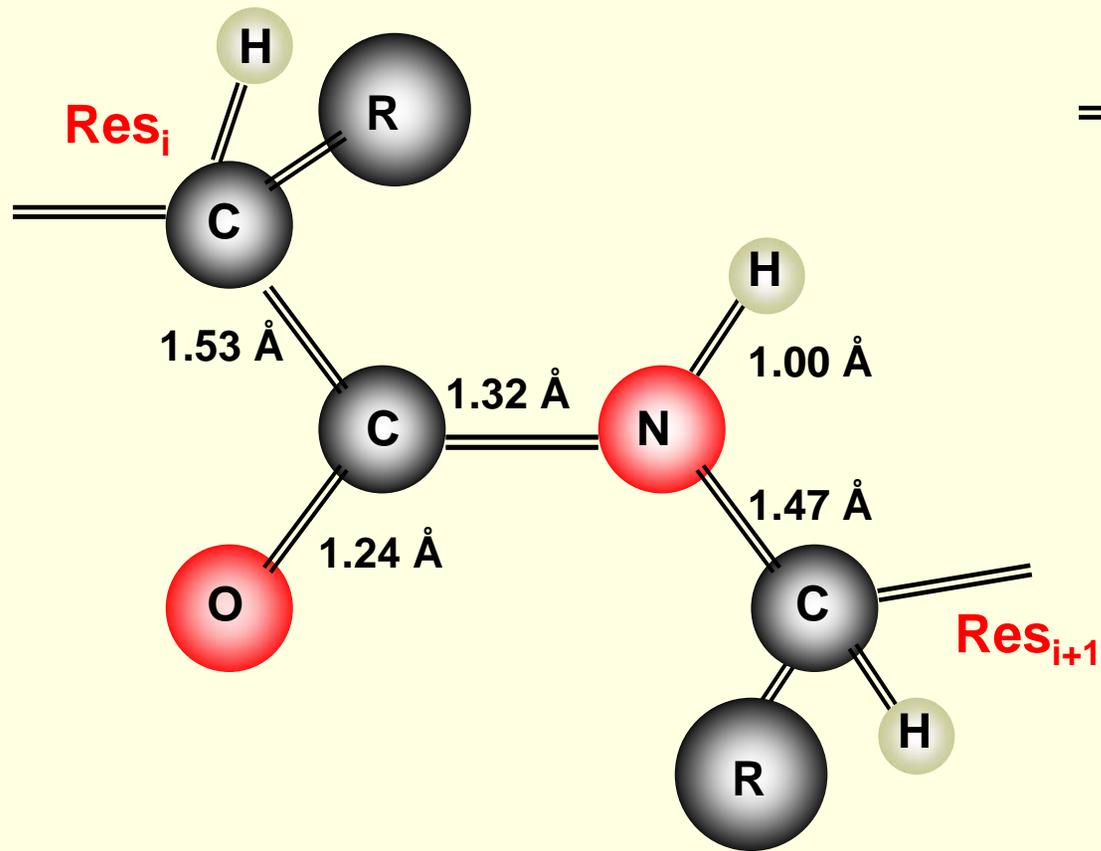
## Advantages

- Easiest and quickest way to build a polypeptide
- More complex lattices allow reasonably accurate representation

## Disadvantages

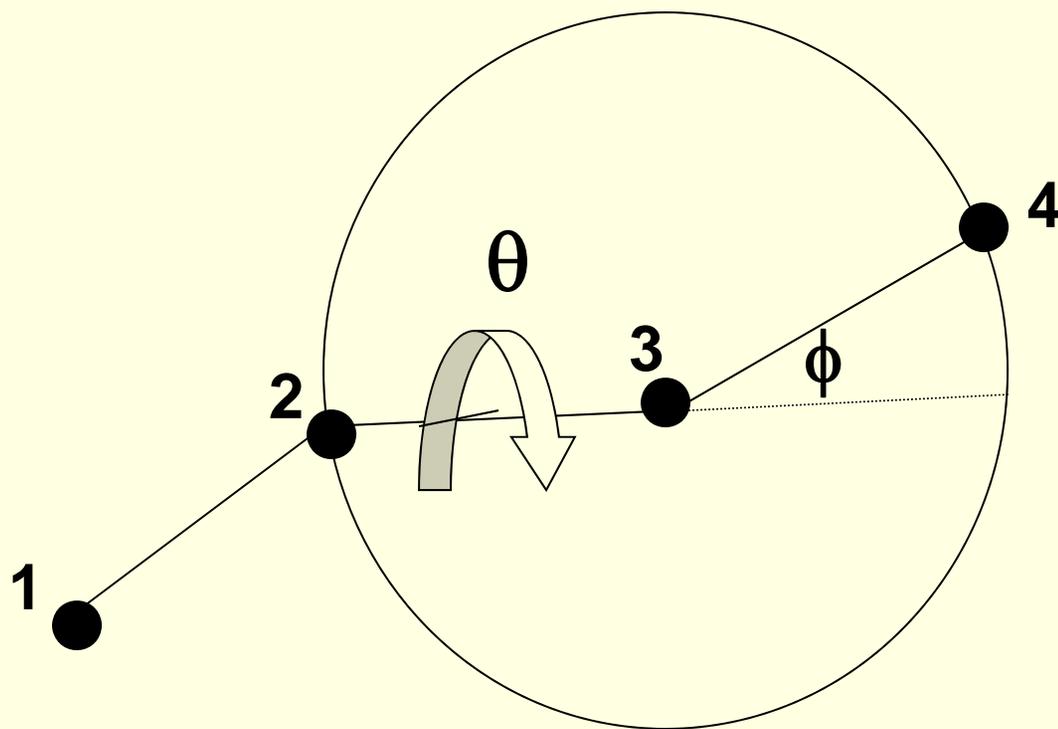
- At best, only an approximation to the real thing
- Does not allow accurate constructs
- Complex lattices are as “costly” as the real thing

# Non-Lattice Models



# Simplified Chain Representation

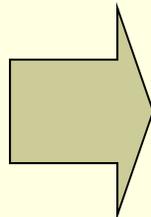
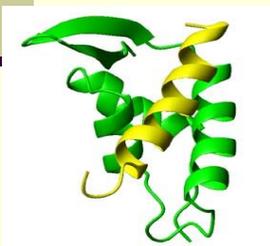
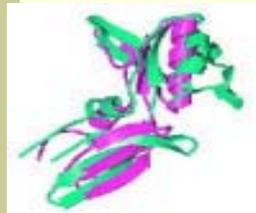
---



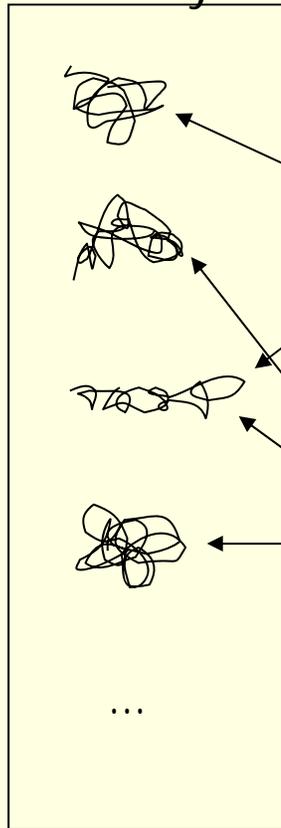
**Spherical Coordinates**

# Assembly of sub-structural units

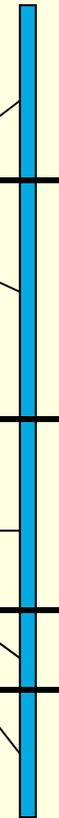
known structures



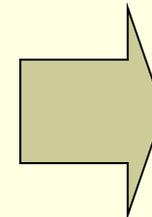
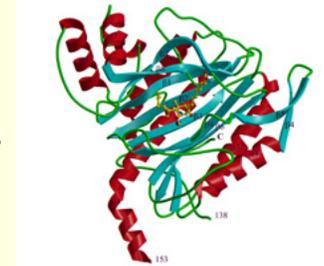
fragment library



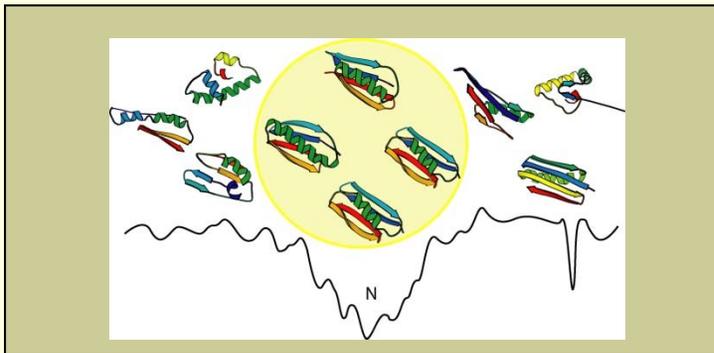
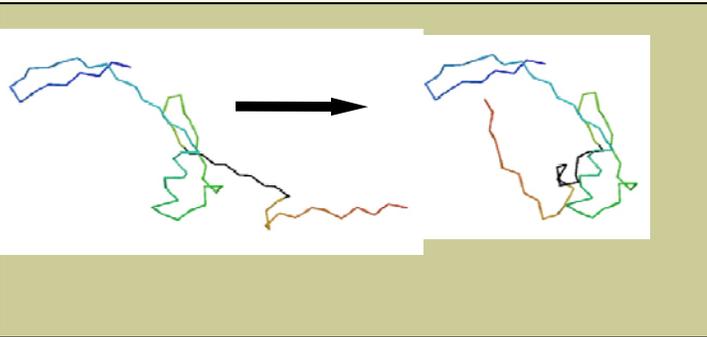
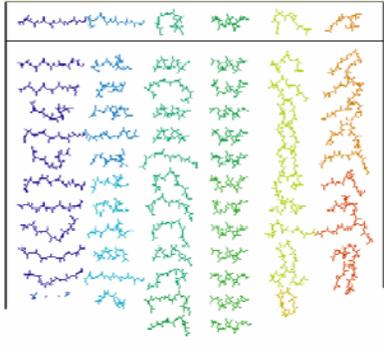
protein sequence



predicted structure



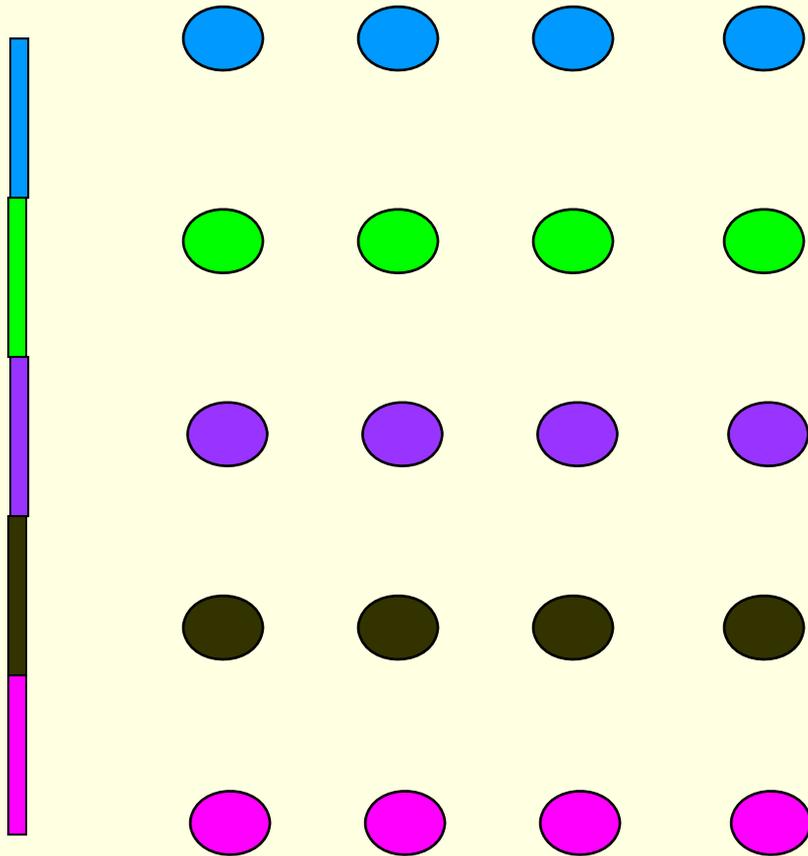
# *Structure Prediction with Rosetta*



- Select fragments consistent with local sequence preferences
- Assemble fragments into models with native-like global properties
- Identify the best model from the population of decoys

# Modelling

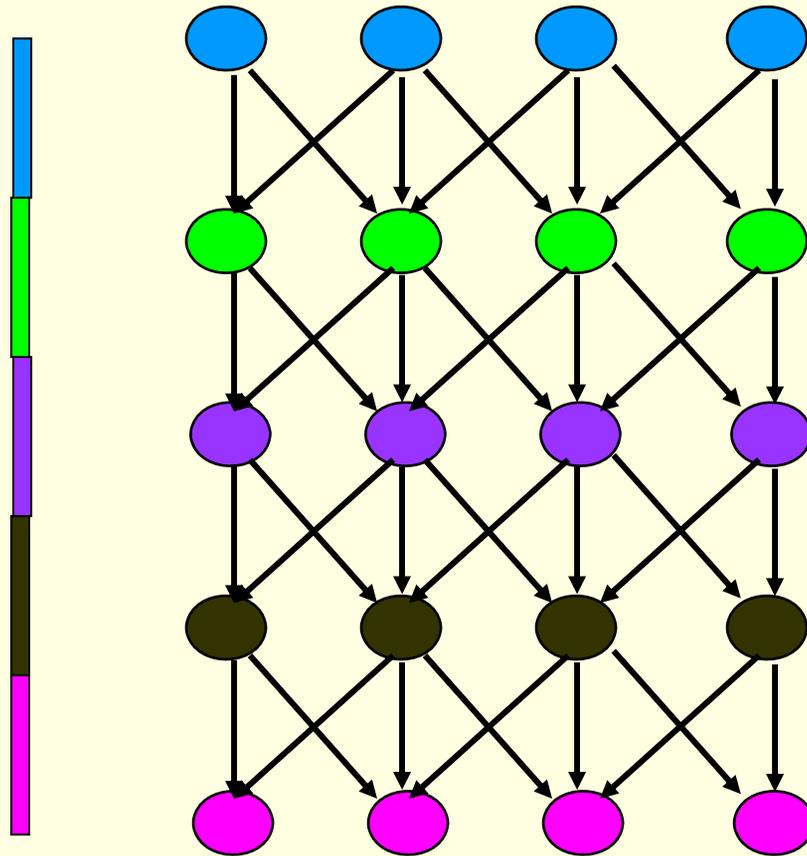
Protein sequence



- Model each candidate local structure as a node

# Modelling

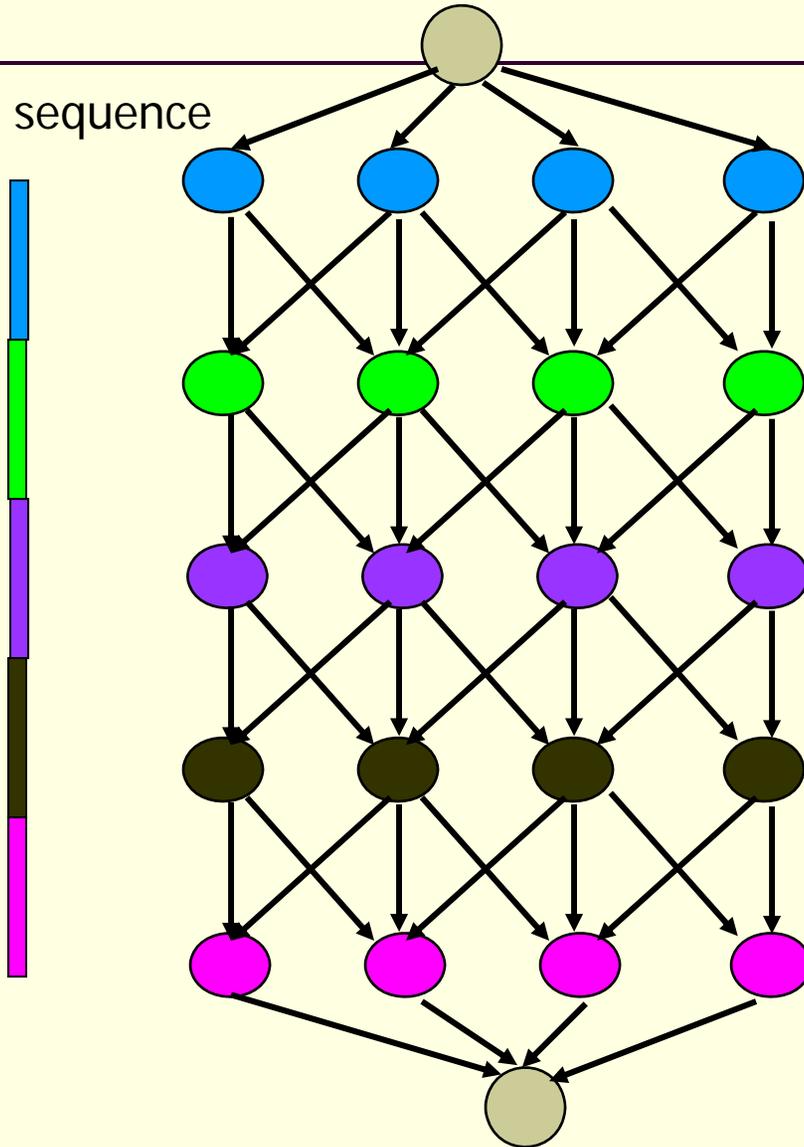
Protein sequence



- Model each candidate local structure as a node
- If two consecutive local structure are compatible, an edge joins them

# Modelling

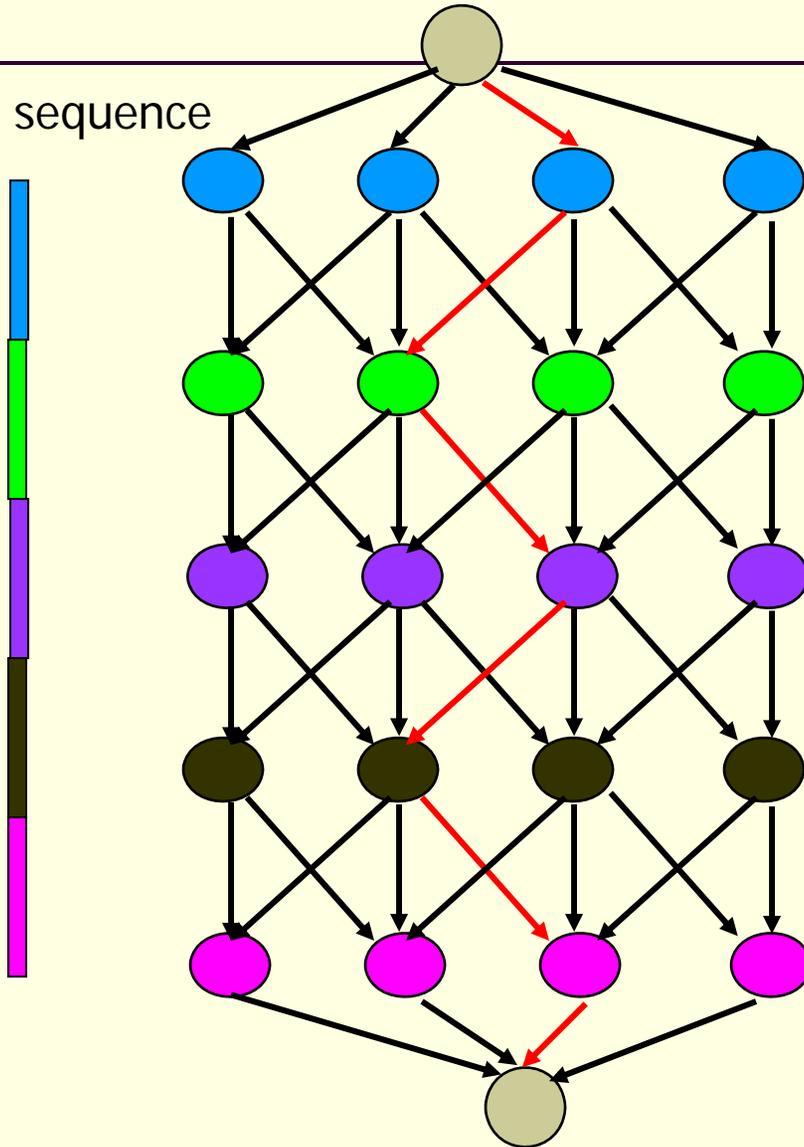
Protein sequence



- Model each candidate local structure as a node
- If two consecutive local structure are compatible, an edge joins them
- Add a source  $s$  and sink  $t$  to the graph

# Modelling

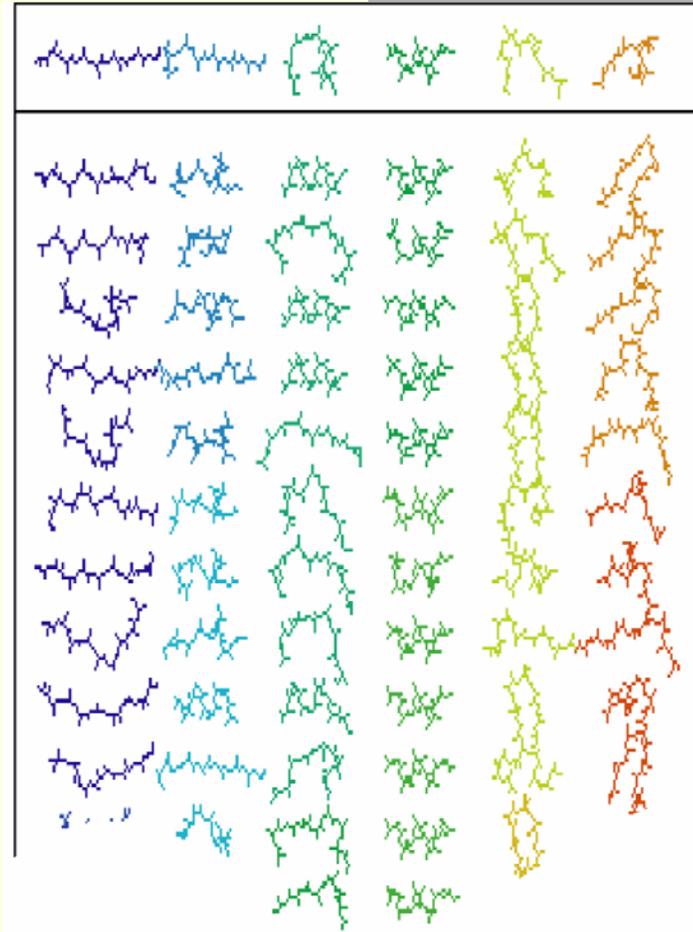
Protein sequence



- Each path from s to t forms a candidate structure
  - At least one of the s-t paths is native-like structure
  - A good search strategy should pick up this path with less time consuming
  - A good model should reduce the search space

# Build the Fragment Library-Rosetta

- Extract possible local structures from PDB



# Generate the Fragment Library

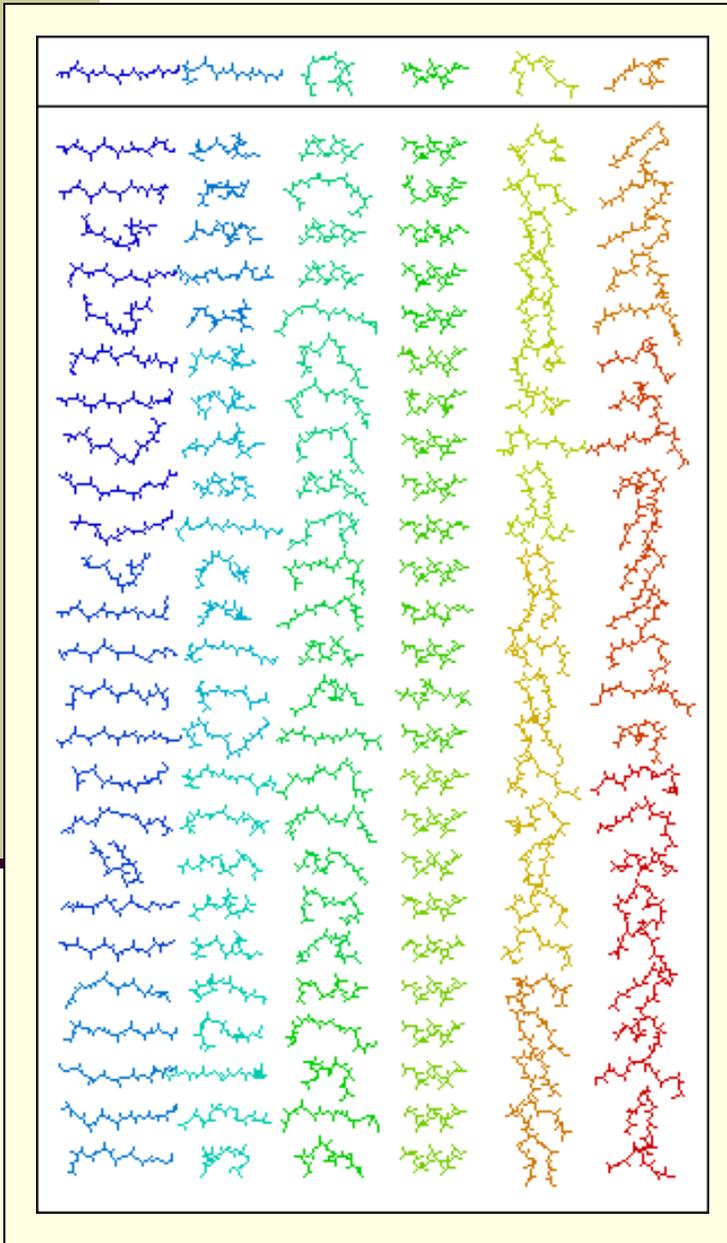
---

- Select PDB template
  - Select Sequence Families
  - Each Family has a single known structure (family)
  - Has no more than 25% sequence identity between any two sequence
- Clustering the fragments
  - Generate all the fragments from the selected families
  -

# Find Local Structures

- Given a subsequence, a local structure to be identified
  - Represent each subsequence with a vector
    - $V = \{v_1, v_2, \dots, v_k\}$
    - eg:  $V$  as a  $20 \times l$  matrix, with the  $(i, j)$ -th entry represent the frequency of amino acid  $j$  occurs at position  $i$
  - Represent each substructure with a vector
    - $V' = \{v_1', v_2', \dots, v_k'\}$
    - eg:  $V'$  as a  $20 \times l$  matrix, with the  $(i, j)$ -th entry represent the frequency of amino acid  $j$  occurs at position  $i$
  - Rank the structure according to:
    - $\sum_i |v_i - v_i'|$
    - This implies that the entries of the vectors are independent.

# *Rosetta Fragment Libraries*



- 25-200 fragments for each 3 and 9 residue sequence window
- Selected from database of known structures  
> 2.5Å resolution  
< 50% sequence identity
- Ranked by sequence similarity and similarity of predicted and known secondary structure

# Search Strategy

---

- Reduce the Search Space
- Design Better Search Strategies

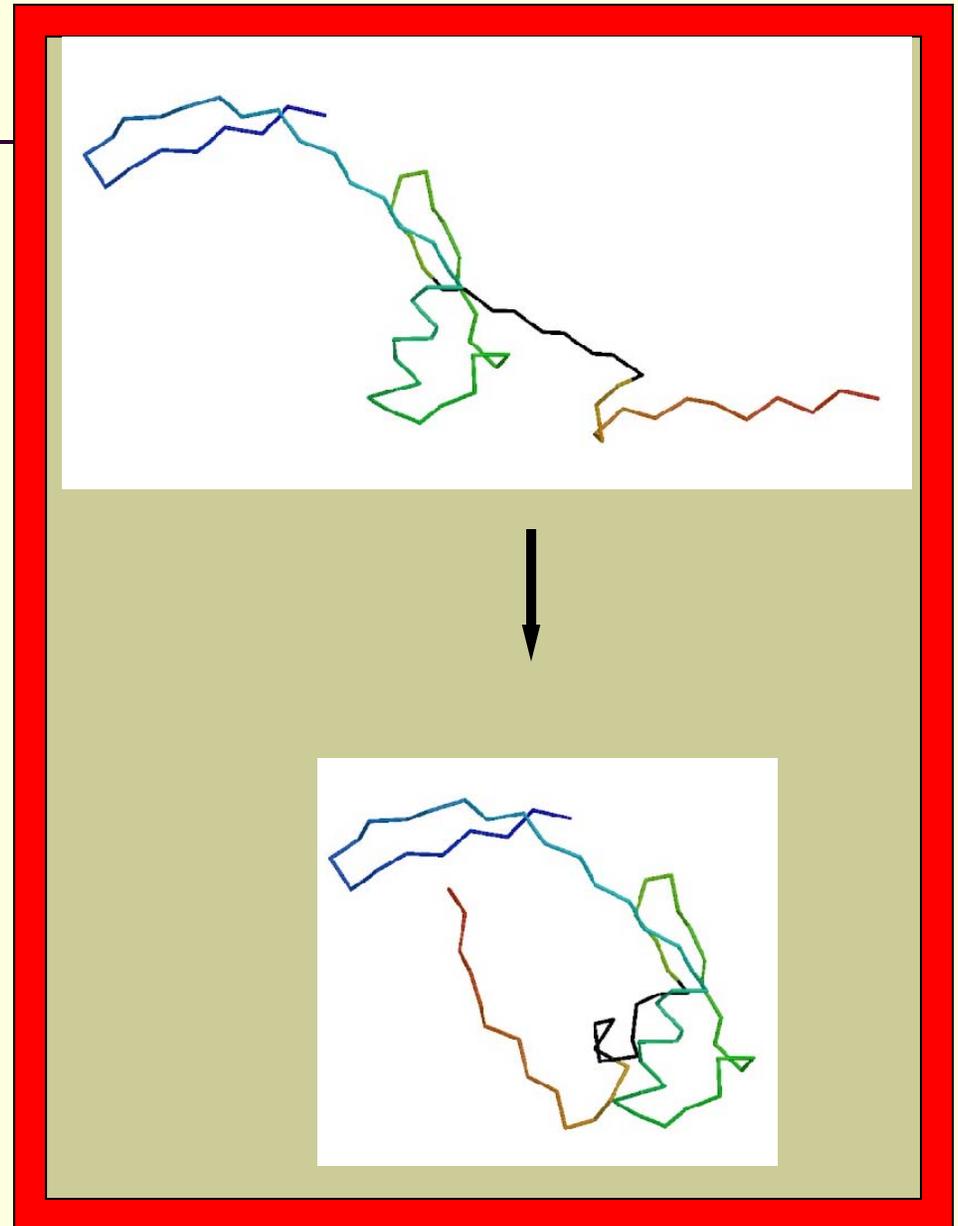
# Scoring Function

---

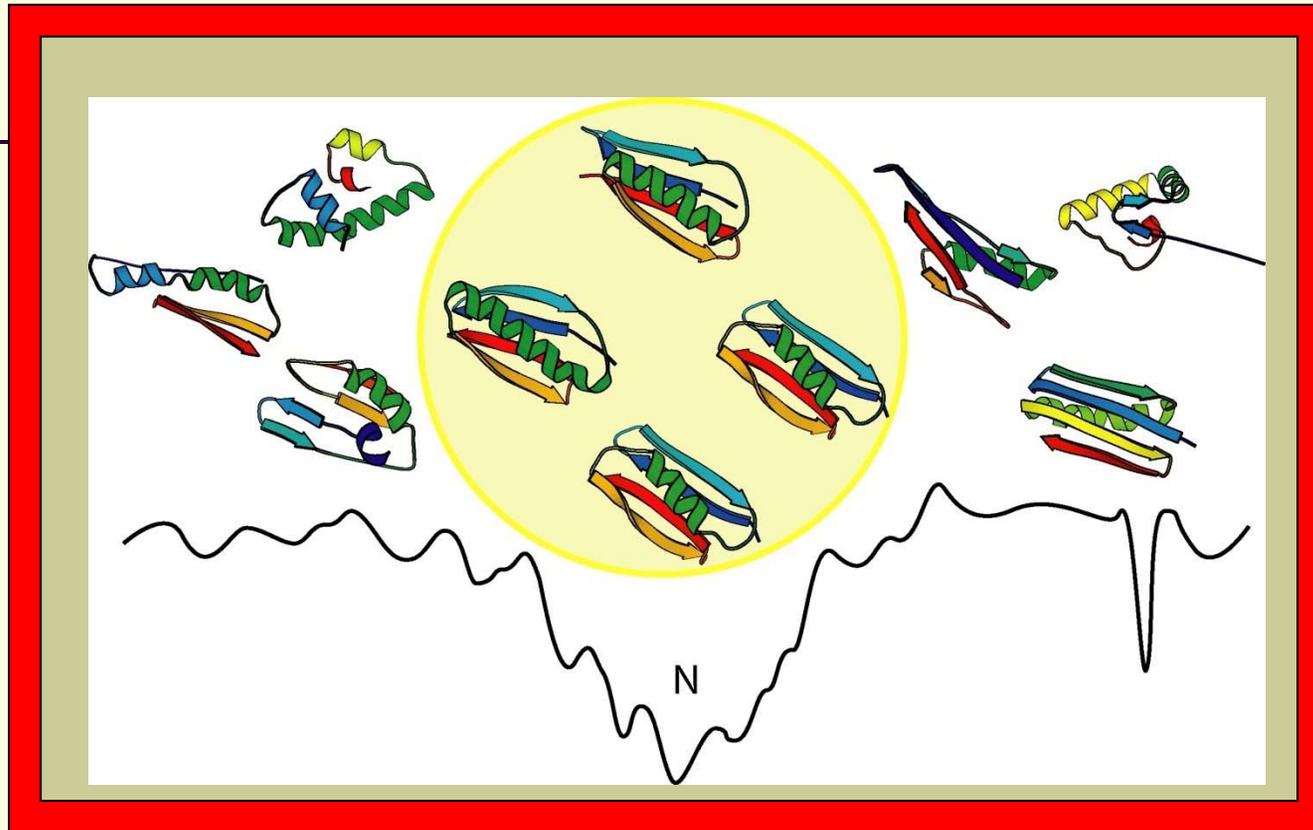
- Ideal energy function
  - Has a clear minimum in the native structure.
  - Has a clear path towards the minimum.
  - Global optimization algorithm should find the native structure.

# *Rosetta Potential Function*

- Derived from Bayesian treatment of residue distributions in known protein structures
- Reduced representation of protein used; one centroid per sidechain
- Potential Terms:
  - environment (solvation)
  - pairwise interactions (electrostatics)
  - strand pairing
  - radius of gyration
  - C $\beta$  density
  - steric overlap



# *Decoy Discrimination: Identifying the Best Structure*



- 1000-100,000 short simulations to generate a population of 'decoys'
- Filter population to correct systematic biases
- Full atom potential functions to select the deepest energy minimum
- Cluster analysis to select the broadest minimum
- Structure-structure matches to database of known structures

# The Rosetta Scoring Function

---

$$P(\text{structure}|\text{sequence}) \propto P(\text{sequence}|\text{structure}) \times P(\text{structure})$$

Sequence dependent:

- hydrophobic burial
- residue pair interaction

Sequence independent:

- helix-strand packing
- strand-strand packing
- sheet configurations
- vdW interactions

# The Sequence Dependent Term

---

$$P(aa_1, \dots, aa_n | X) =$$

$$\prod_i P(aa_i | X) \times$$

$$\prod_{i < j} \frac{P(aa_i, aa_j | X)}{P(aa_i | X)P(aa_j | X)} \times$$

$$\prod_{i < j < k} \frac{P(aa_i, aa_j, aa_k | X)P(aa_i | X)P(aa_j | X)P(aa_k | X)}{P(aa_i, aa_j | X)P(aa_i, aa_k | X)P(aa_j, aa_k | X)} \times$$

...

# The Sequence Dependent Term

---

$$P(\text{sequence}|\text{structure}) \approx P_{\text{env}} \times P_{\text{pair}}$$

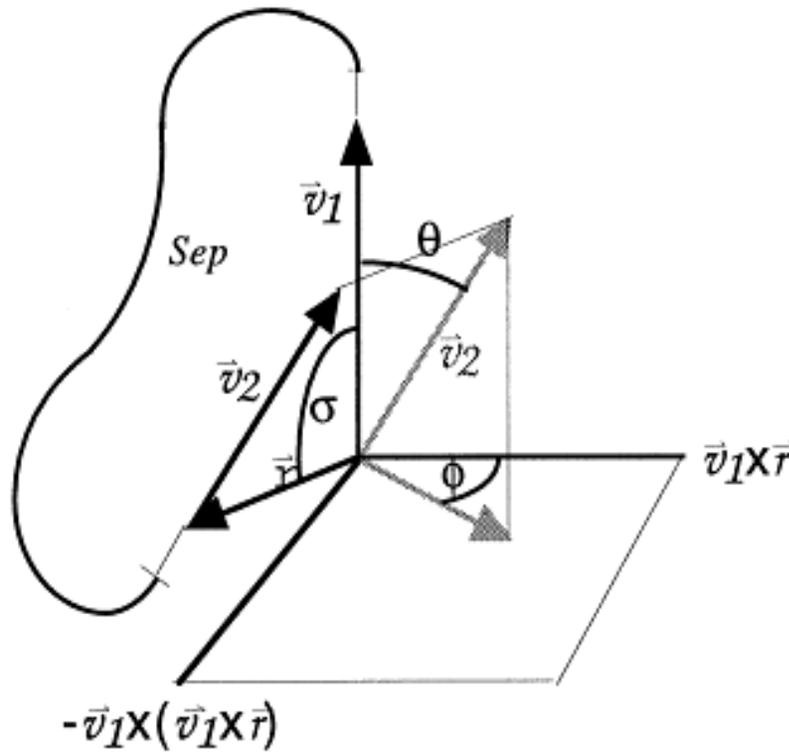
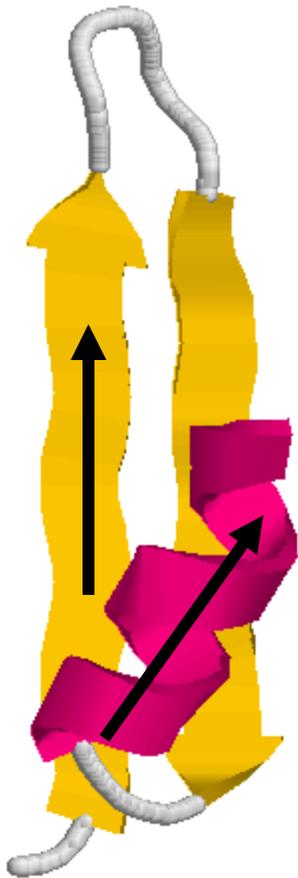
$$P_{\text{env}} = \prod_i P(\text{aa}_i | E_i)$$

$$P_{\text{pair}} = \prod_{i < j} \frac{P(\text{aa}_i, \text{aa}_j | E_i, E_j, r_{ij})}{P(\text{aa}_i | E_i, r_{ij}) P(\text{aa}_j | E_j, r_{ij})}$$

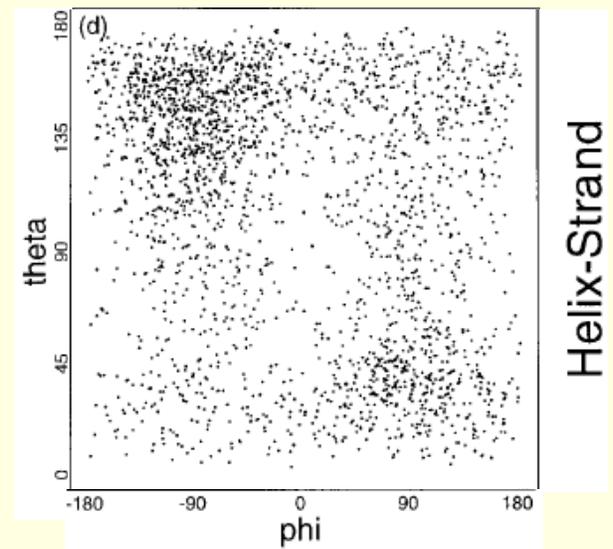
# The Sequence Independent Term

$$P(r, \phi, \theta, \sigma, hb|sep) \approx$$

$$P(\phi, \theta|r, sep) \times P(hb|r, sep) \times P(\sigma|r, sep) \times P(r|sep)$$



vector representation



# The Model

---

$$P(\text{structure}) = P_A^{w_A} P_B^{w_B} P_C^{w_C}, \quad w_X > 0.$$

$$-\log P(\text{structure}|\text{sequence}) \propto$$

$$-\log P(\text{sequence}|\text{structure}) - \log P(\text{structure})$$

$$g(\text{rmsd}) = w_{\text{protein}} + w_{\text{HS}} \log P_{\text{HS}} + w_{\text{SS}} \log P_{\text{SS}} + w_{\text{vdW}} \text{VdW} + \\ w_{\text{sheet}} \log P_{\text{sheet}} + w_{\text{seq}} (\log P_{\text{env}} + \log P_{\text{pair}})$$

# Search Strategy

---

## ■ Requirement

- Identify the native structure easily
  - Filter out those non-native ones
- Eliminate the non-native candidates as early as possible
- Jumping out from the local minimum
- No repetitions
- ...

## ■ Search Strategies

- Taboo search, simulated annealing, genetic algorithms, multi-agent, ...
-

# ROSETTA search algorithm

## Monte Carlo/Simulated Annealing

---

- Structures are assembled from fragments by:
  - Begin with a fully extended chain
  - Randomly replace the conformation of one 9 residue segment with the conformation of one of its neighbors in the library
  - Evaluate the move: Accept or reject based on an energy function
  - Make another random move, tabu list is built to forbidden some local minimums
  - After a prescribed number of cycles, switch to 3-residue fragment moves

# A Filter for Bad $\beta$ -Sheets

Many decoys do not have proper sheets. Filtering those out seems to enhance the rmsd distribution in the decoy set. Bad features we see in decoys include:

- No strands,
- Single strands,
- Too many neighbors,
- Single strand in sheets,
- Bad dot-product,
- False sheet type (barrel),

# ROSETTA Obstacles & Enhancements

---

- generate lots of unrealistic decoys
  - Filter based on contact order
  - quality of  $\beta$ -sheets
  - poor packing
- large search space
  - Bias fragment picking by predicted secondary structure, faster computational algorithms
- low confidence in the result
  - – Fold many homologs of the target, cluster the answers, report the cluster with highest occupancy