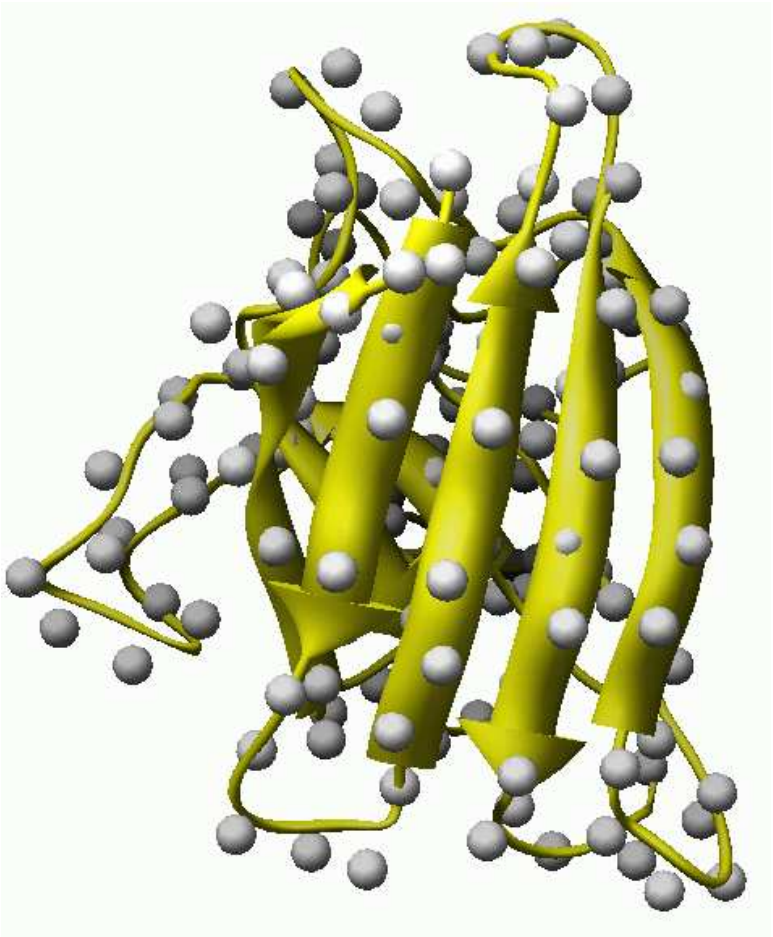
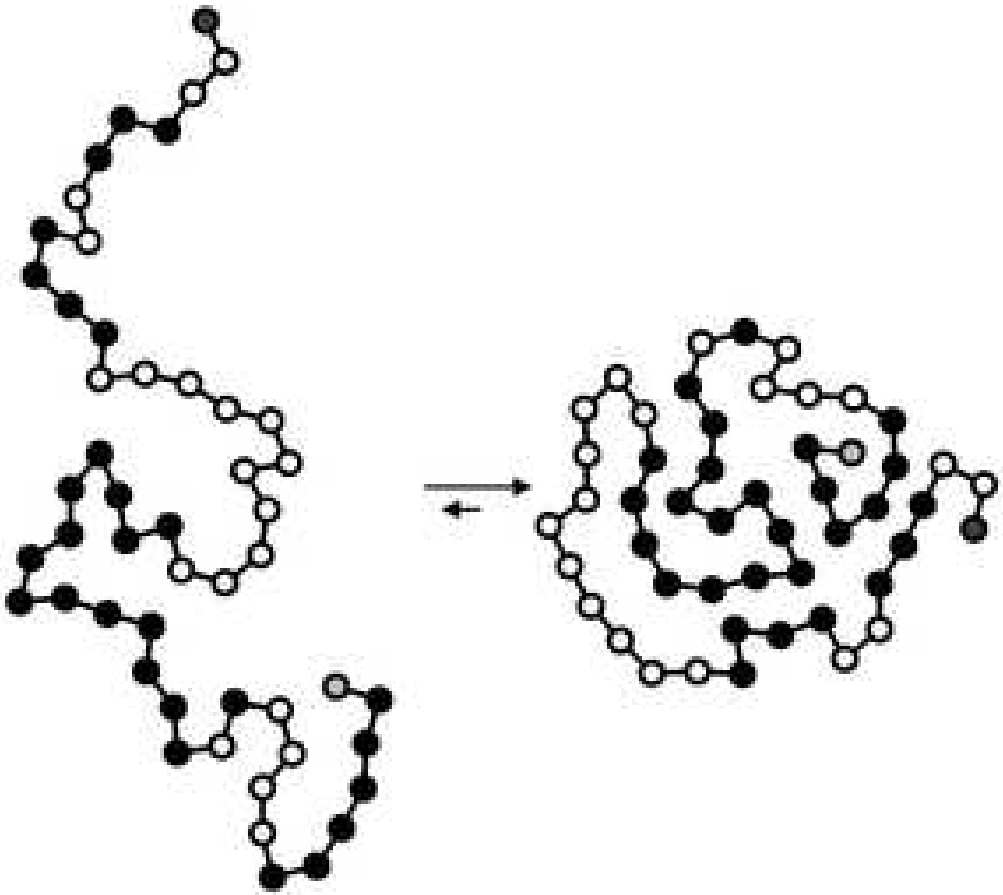
How does this sequence fold?

VQAVAVLKGDAGVSGVVKFEQASESEPTTVSYEIAGNSPNAERGFHIHEFGDATNGCVSA
GPHFNPFKKTHGAPTDEVRHVGDMGNVKTDENGVAKGSFKDSLIKLIGPTSVVGRSVVIH
AGQDDLGKGDTEESLKTGNAGPRPACGVIGLTN



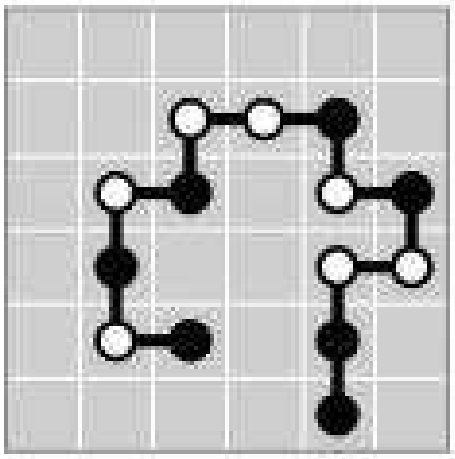
Protein folding: schematic



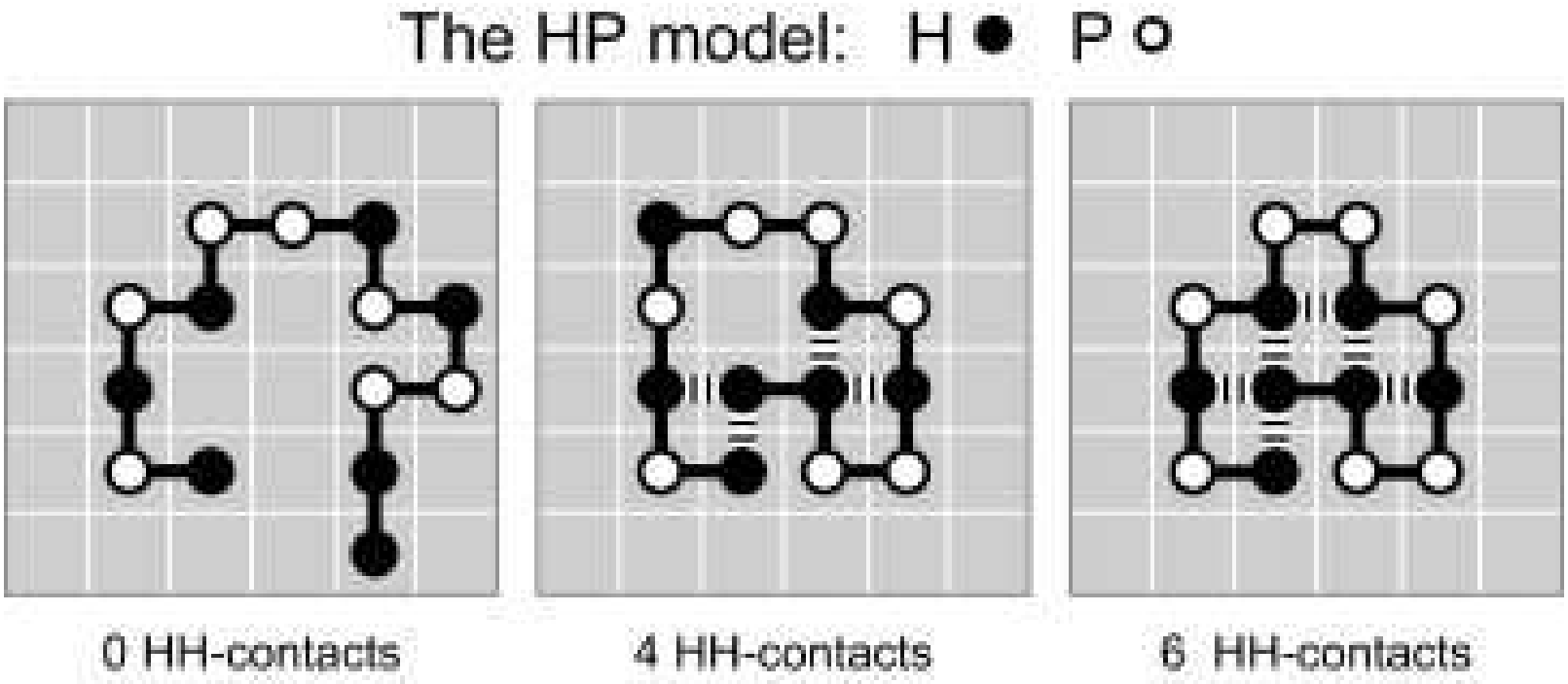
Lattice model

- model a protein as a chain of hydrophobic (H) and polar (P) residues

- a conformation is a self-avoiding walk on a 2D square lattice



Conformations



Ab initio structure prediction

Kim T. Simons, Charles Kooperberg, Enoch Huang and David Baker
“Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions”
J. Mol. Biol., vol. 268, 209-225 (1997).

A simulated annealing procedure needs:

- method for generating structures
- scoring function

Generating structures

Three-dimensional structures are generated by splicing together fragments of proteins of known structure with similar local sequences.

Earlier studies showed a strong correlation between local sequence and local structure of nine residue fragments.

For each segment of length 9 in the sequence being folded, the 25 nearest sequence neighbours in the structure database were identified.

The conformation of each of these segments was adjusted to give ideal bond lengths and angles.

The percentage of neighbours structurally similar to the true structure is greater when multiple sequence information is available.

Estimating P(structure)

In fold recognition, we can assume that each known fold (a finite set) is equally probable.

However, when considering a vast number of synthesised conformations, many of which are highly improbable, we need some way of assessing the feasibility of each conformation.

Simons et al. (1997) suggest a simple approach in which P(structure) is zero if atoms overlap, and otherwise P(structure) is related to the compactness of the structure, measured by the “radius of gyration”.

The radius of gyration is defined as the square root of the mass average of r_i^2 for all of the mass elements.

Estimating P(sequence|structure)

Similar to scoring a sequence-fold match when threading.

Profiles:

$$\prod_i P(aa_i \mid E_i)$$

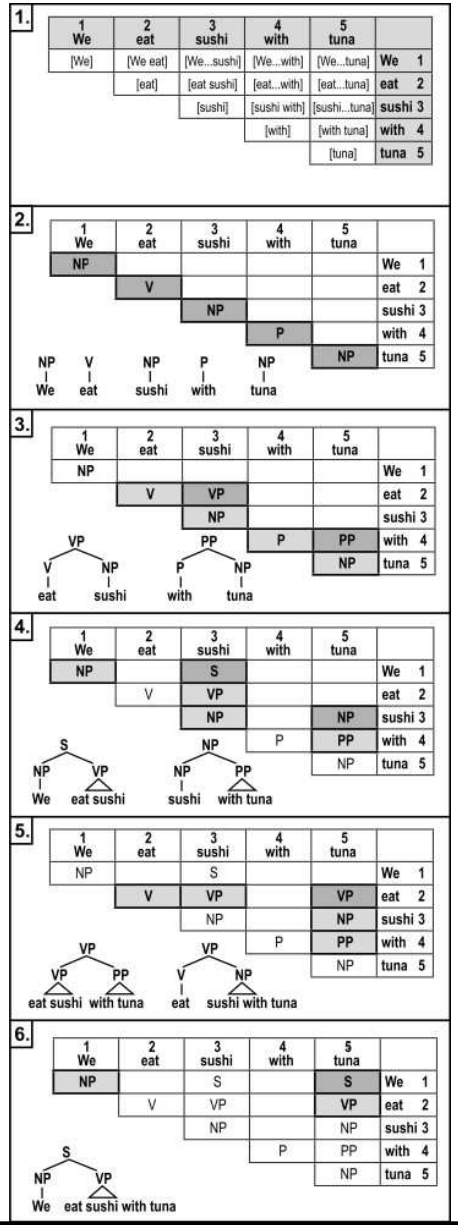
Pairwise potentials:

$$\prod_{i<j} P(aa_{i,}aa_j \mid r_{ij})$$

Simons et al. (1997):

$$\prod_i P(aa_i \mid E_i) \times \prod_{i<j} \frac{P(aa_{i,}aa_j \mid r_{ij,}E_{i,}E_j)}{P(aa_i \mid r_{ij,}E_{i,}E_j)P(aa_j \mid r_{ij,}E_{i,}E_j)}$$

The CKY algorithm — natural language



Cocke-Kasami-Younger algorithm

- bottom-up parsing
- dynamic programming

Parsing natural language vs. folding a protein

Parsing natural language:

a) start with one-dimensional string of words;

b) consider all possible topologies representing possible relationships among words and phrases;

c) chooses the one that conveys the correct single meaning of the sentence.

Folding a protein:

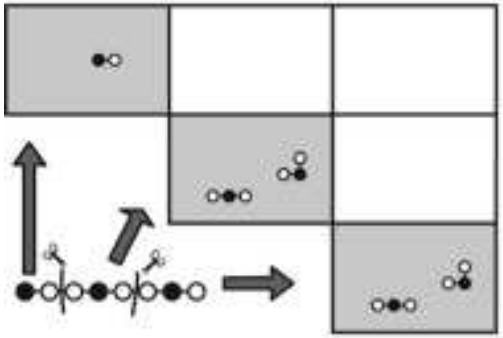
a) start with one-dimensional string of amino acid residues;

b) consider all possible topologies representing possible native substructures of a protein;

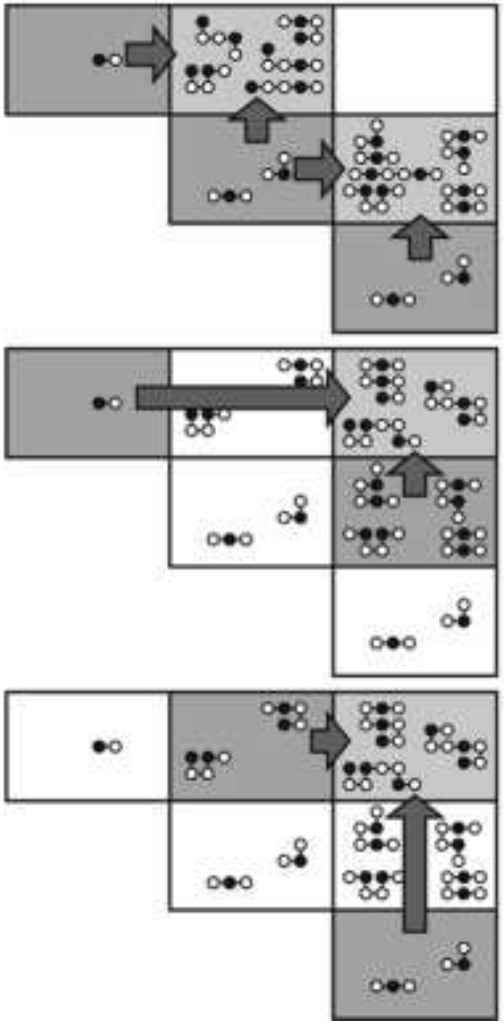
c) chooses the one that has the global minimum free energy.

The CKY algorithm — protein structure

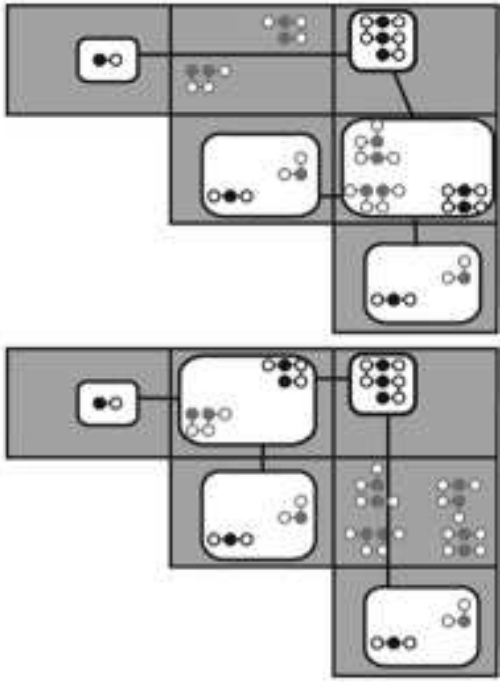
1. Initialize the chart



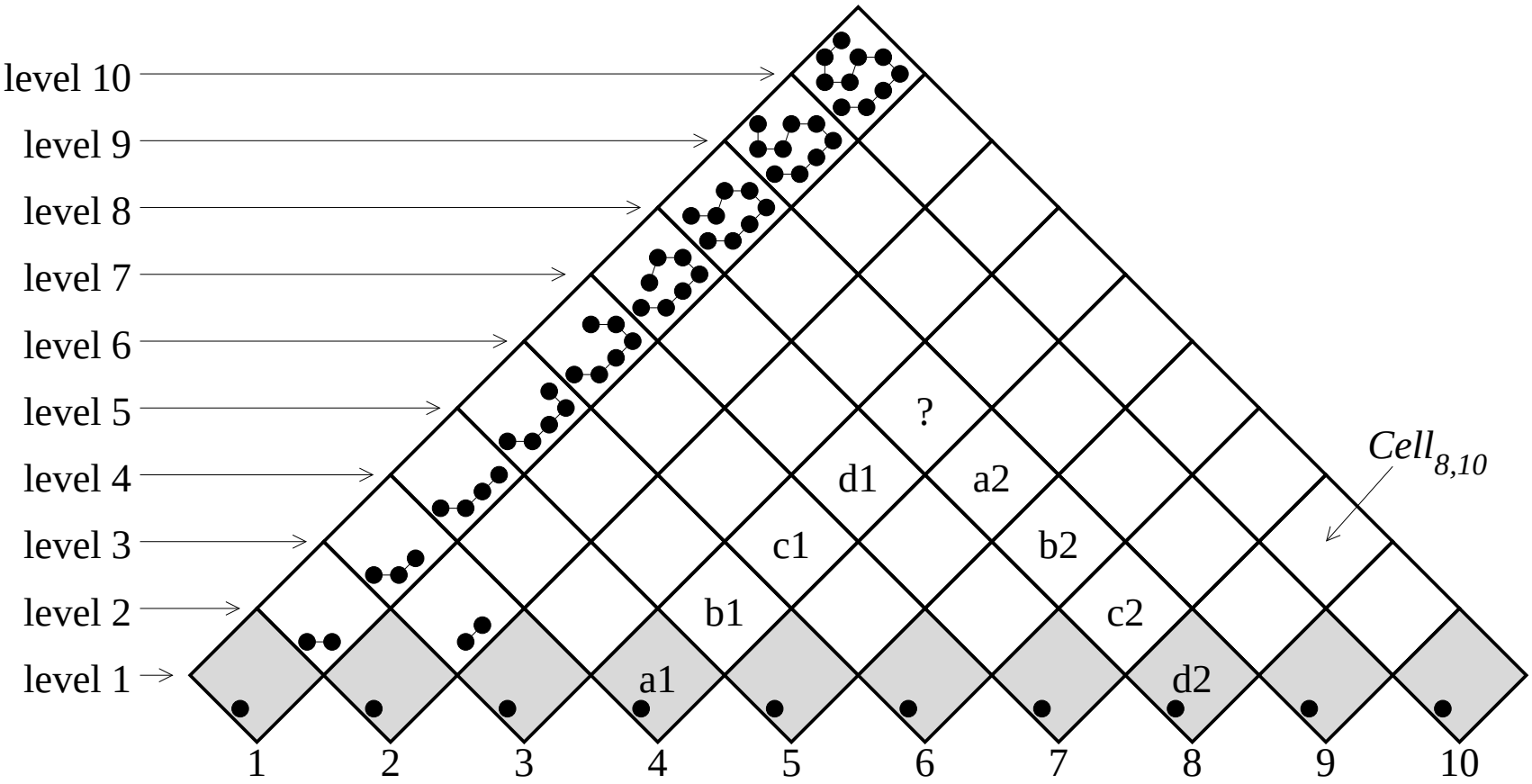
2. Fill the chart



3. Extract the trees



Zipping and assembly



Zipping and assembly with constraints: information used

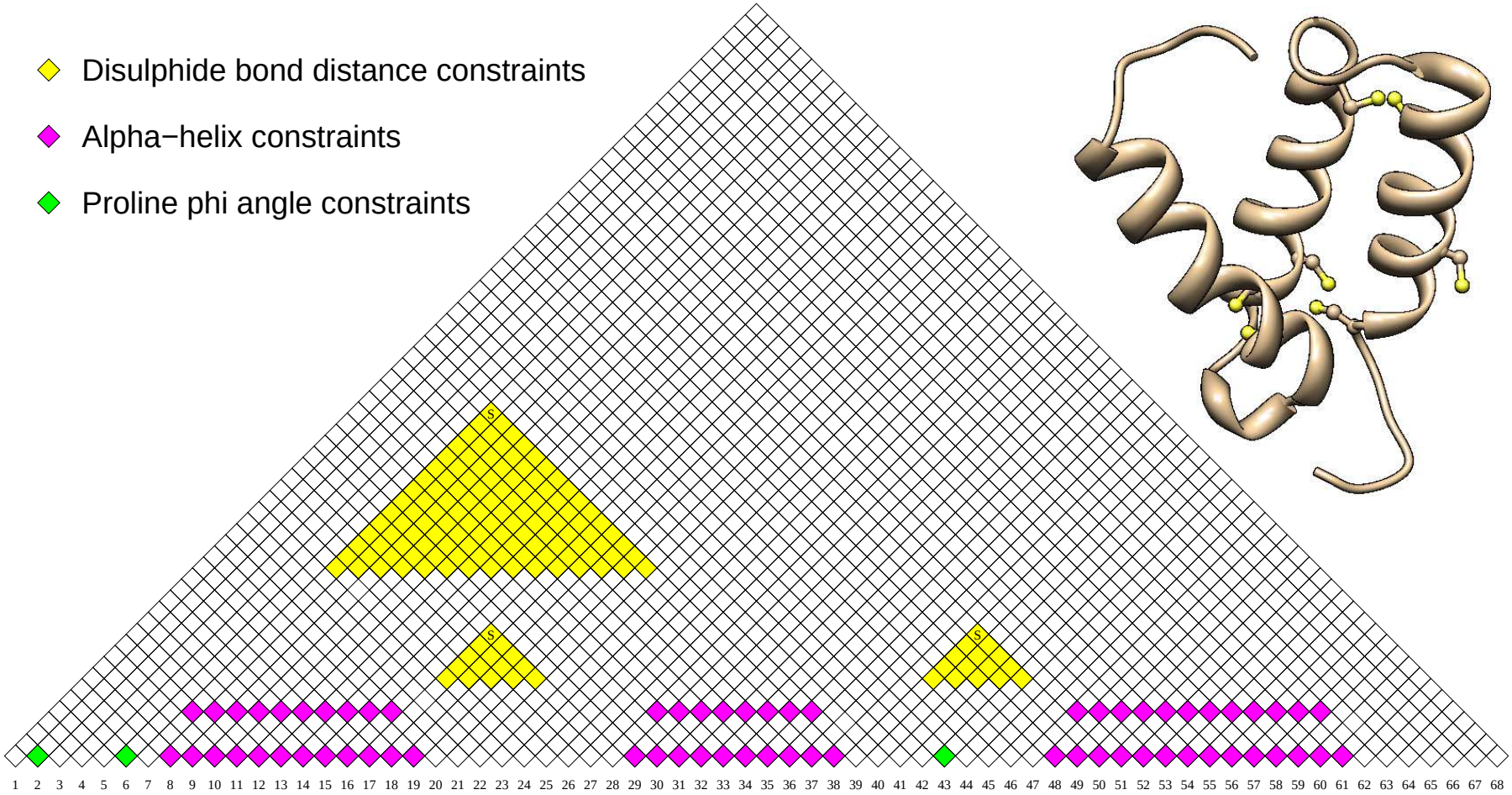
Protein amino acid residue sequence

Constraints

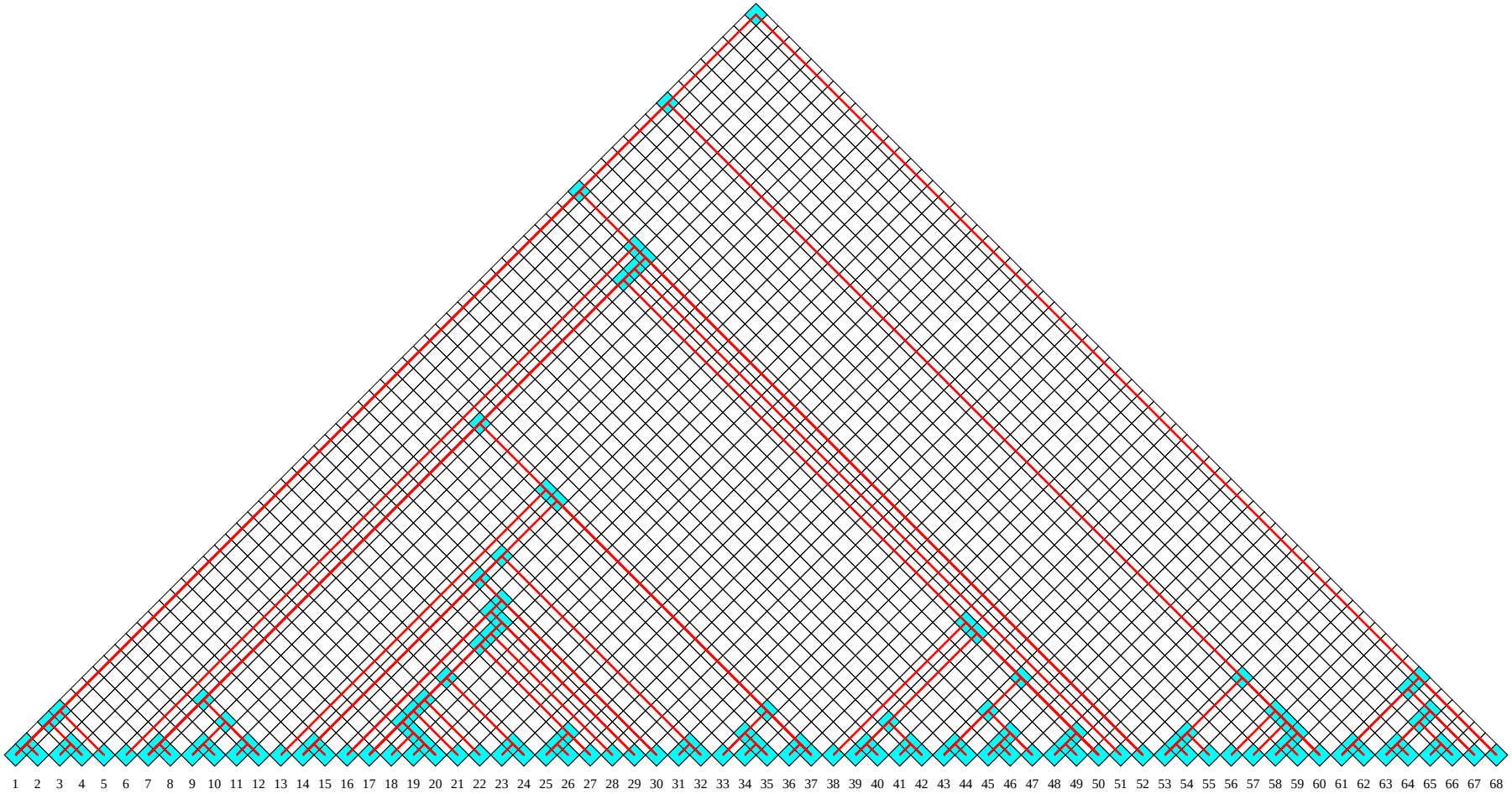
- Angle constraints:
  - torsion angle ranges predicted from chemical shifts
- Distance constraints:
  - main chain N and O involved in hydrogen bonds in secondary structures
    - HN-HN NOEs from 4D NMR experiments
    - from predicted secondary structure
  - disulphide bridges
  - no steric overlaps
  - ...

Constraints used in modelling human p8MTCP

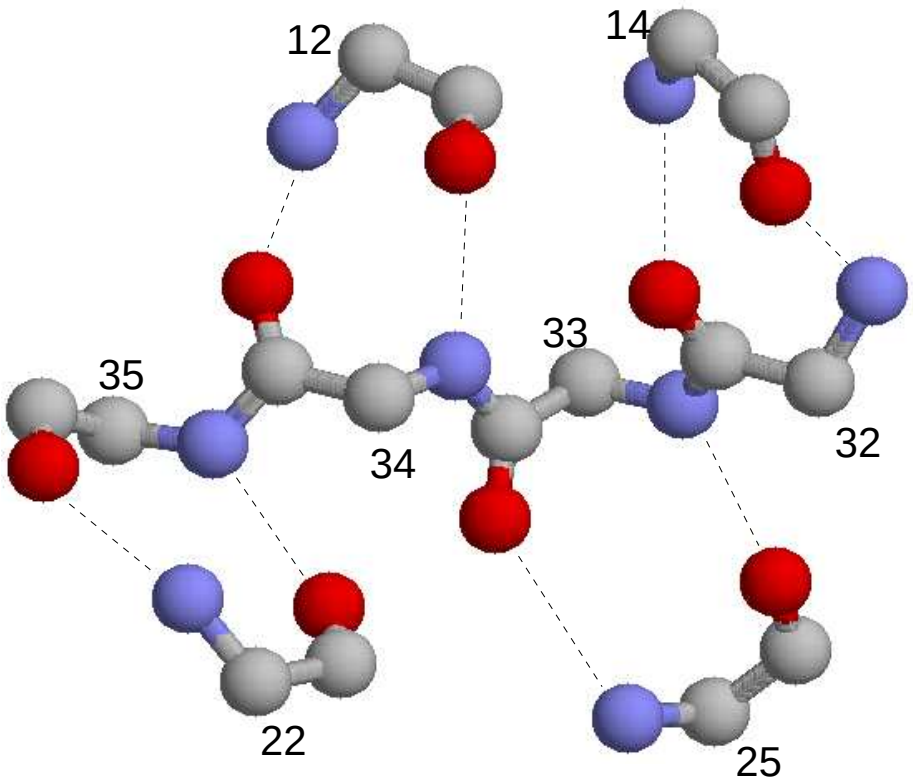
- Disulphide bond distance constraints
- Alpha−helix constraints
- Proline phi angle constraints



Actual cells used in constructing one model



Human β-defensin 6: antiparallel bridges

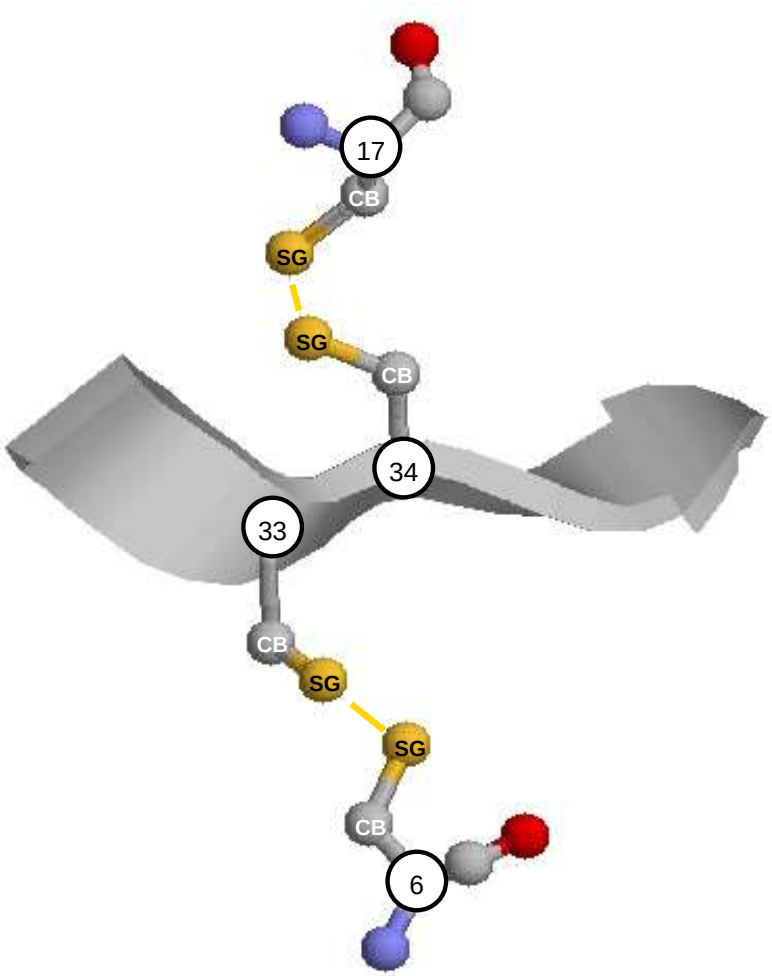



```
residue(1,'PHE').
residue(2,'PHE').
residue(3,'ASP').
residue(4,'GLU').
residue(5,'LYS').  % etc.

disulphide_bond(6,33).
disulphide_bond(13,27).
disulphide_bond(17,34).

alpha_helix(4,8).
```

```
antiparallel_bridge(12,34).
antiparallel_bridge(14,32).
antiparallel_bridge(22,35).
antiparallel_bridge(25,33).
```

Prolog facts

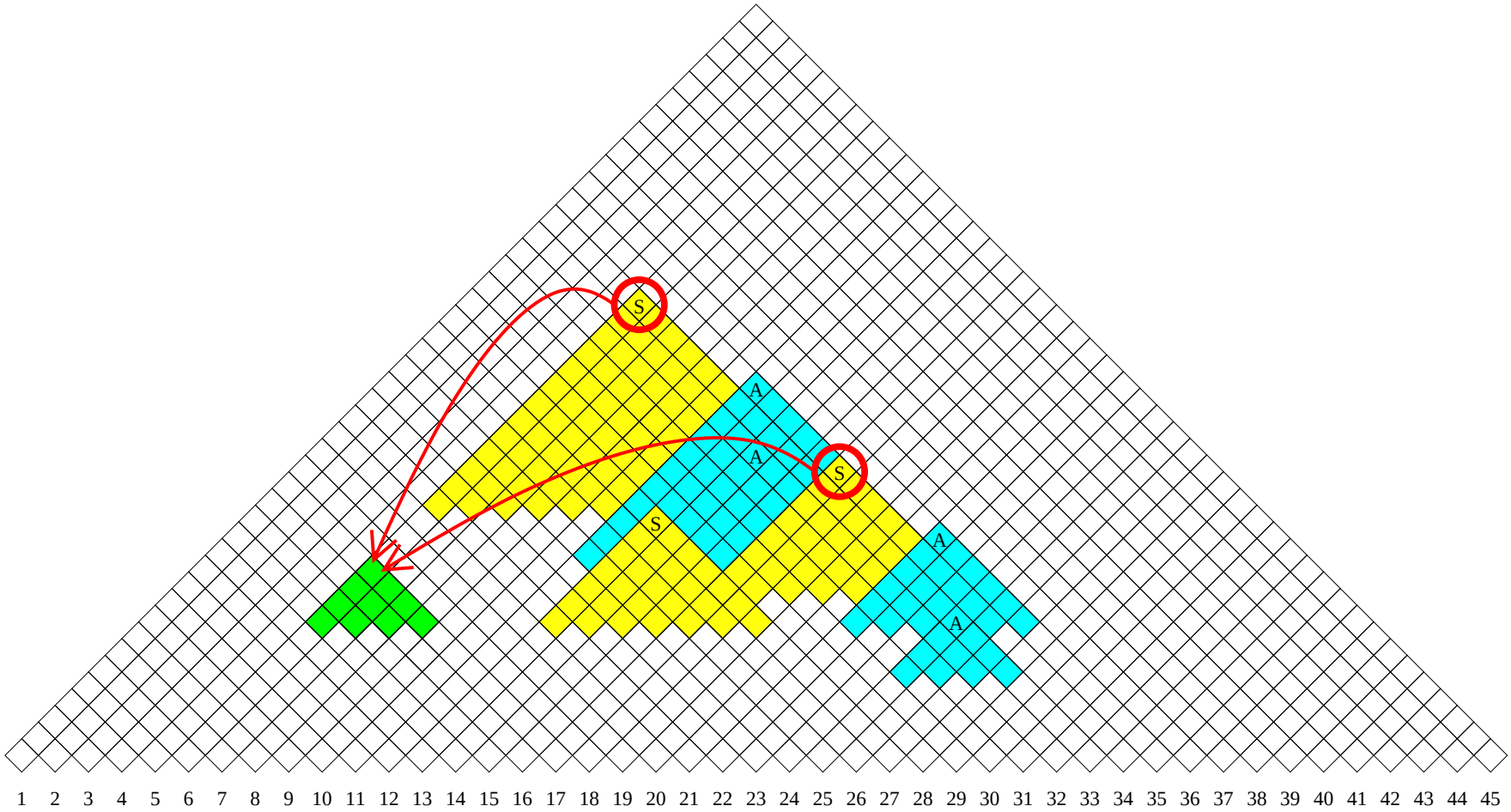
Adjacent residues in a strand



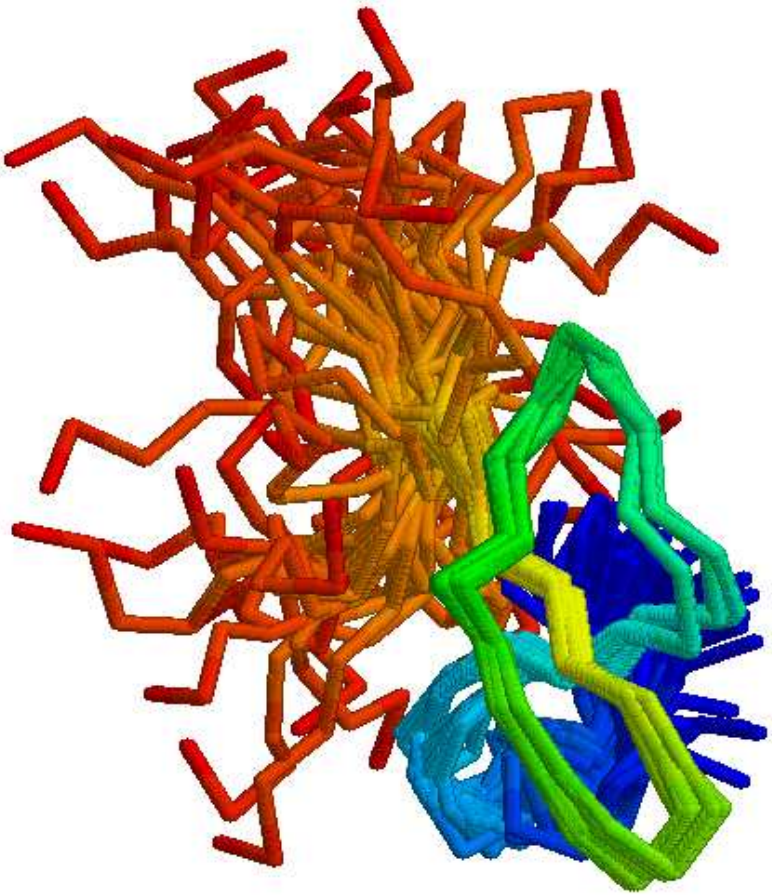
Additional rule:

```
disulphide(A,B) :- disulphide_bond(A,B).
disulphide(A,B) :- disulphide_bond(B,A).

disulphide_distance_constraints :-
  disulphide(A,B),
  disulphide(C,D),
  1 is C-B,
  strand(StrandStart,StrandEnd),
  B >= StrandStart,
  C =< StrandEnd,
  assert(lower_distance_bound(
            (A,'CA'),(D,'CA'),13.0)),
  assert(upper_distance_bound(
            (A,'CA'),(D,'CA'),15.0)),
  fail.
```

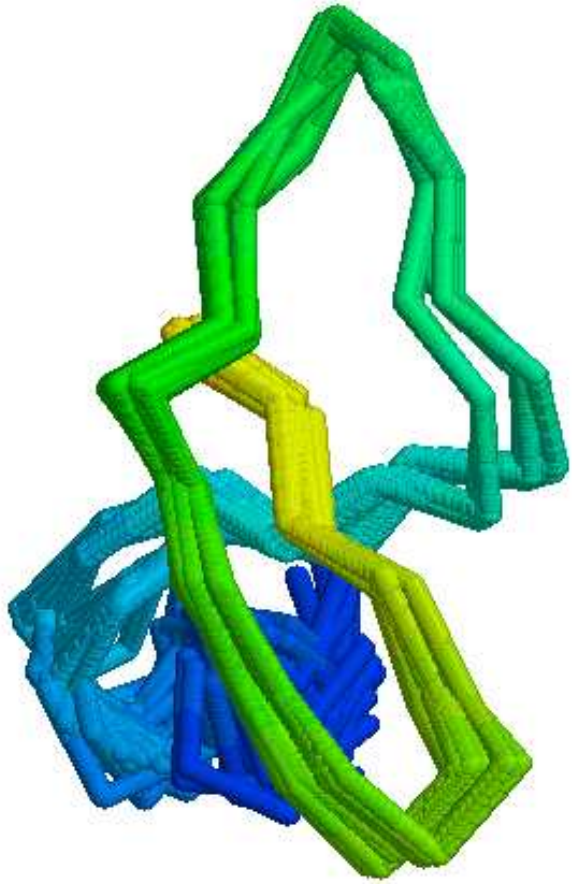
Human β-defensin 6: distance constraints



Human β-defensin 6: 50 best models



All residues



Core residues: 4-35

Claims made for ZAMDP method

- local-first-global-later explains quick folding, and avoidance of vast stretches of conformational space
- reflects parallel nature of physical kinetics
- captures relationship between contact order (whether contacts are mainly local or mainly non-local) and folding rate
- identifies slow- and fast-folding proteins, and slow- and fast-folding routes