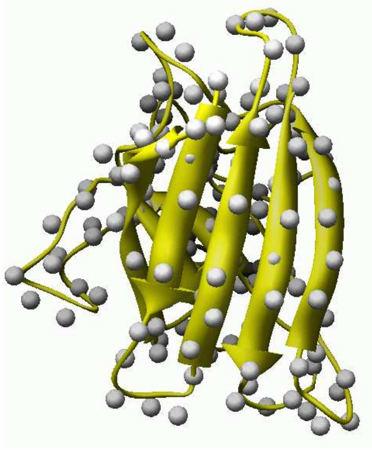
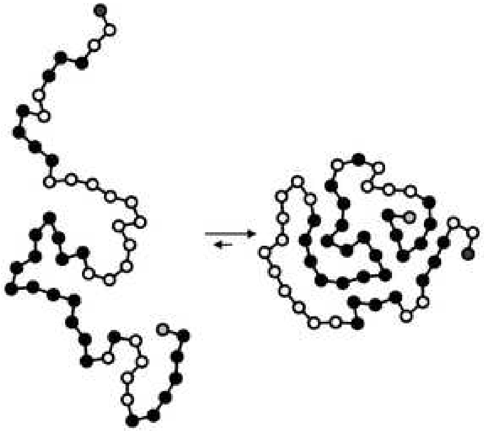
How does this sequence fold?

```
VQAVAVLKGDAGVSGVVKFEQASESEPTTVSYEIAGNSPNAERGFHIHEFGDATNGCVSA
GPHFNPFKKTHGAPTDEVRHVGDMGNVKTDENGVAKGSFKDSLIKLIGPTSVVGRSVVIH
AGQDDLGKGDTEESLKTGNAGPRPACGVIGLTN
```

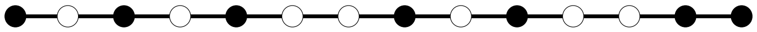


Protein folding: schematic

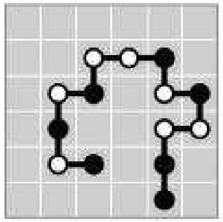


Lattice model

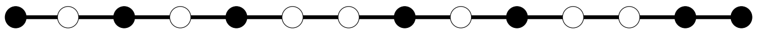
- model a protein as a chain of hydrophobic (H) and polar (P) residues

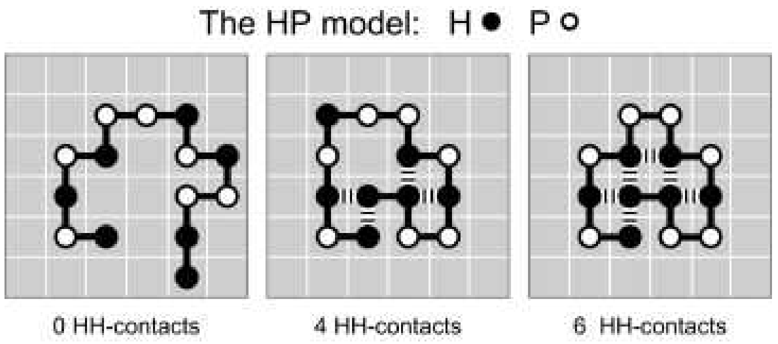


- a conformation is a self-avoiding walk on a 2D square lattice



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:

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

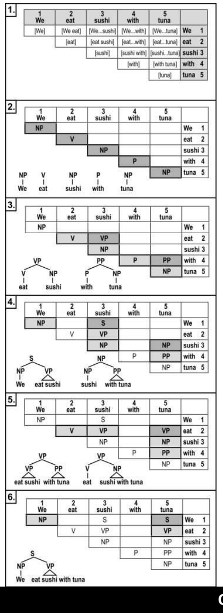
Pairwise potentials:

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

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.

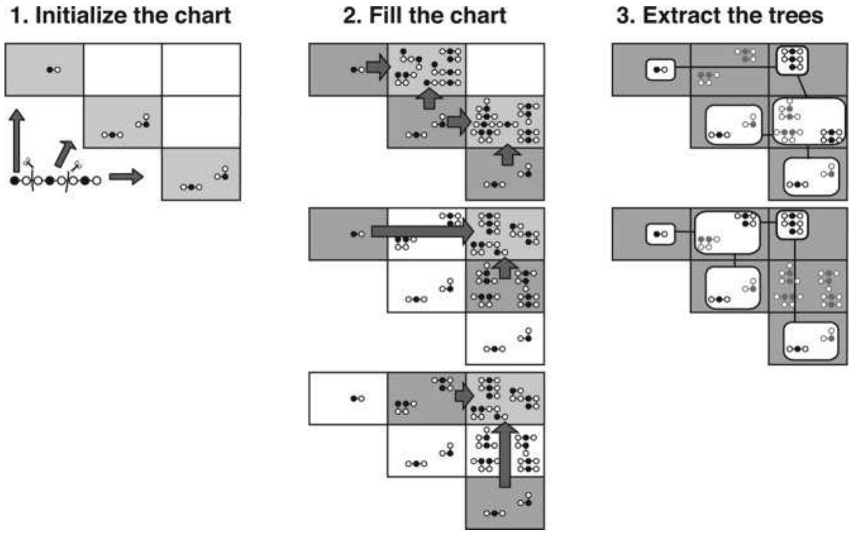
Folding a protein:

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

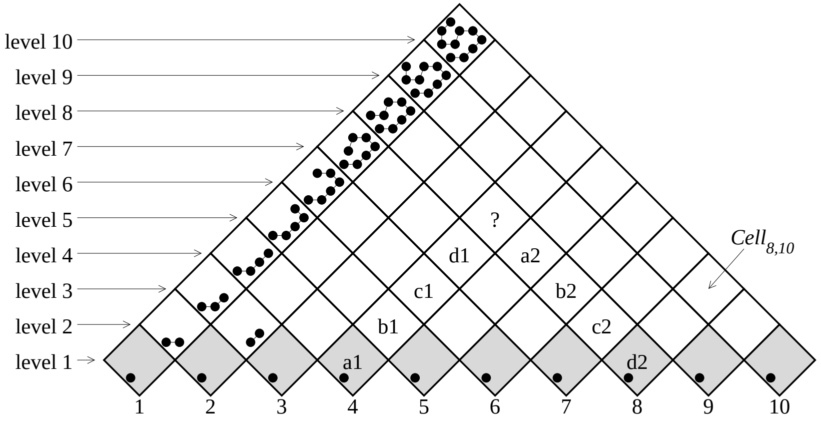
b) consider all possible topologies representing possible native substructures of a protein;

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

The CKY algorithm — protein structure



Zipping and assembly



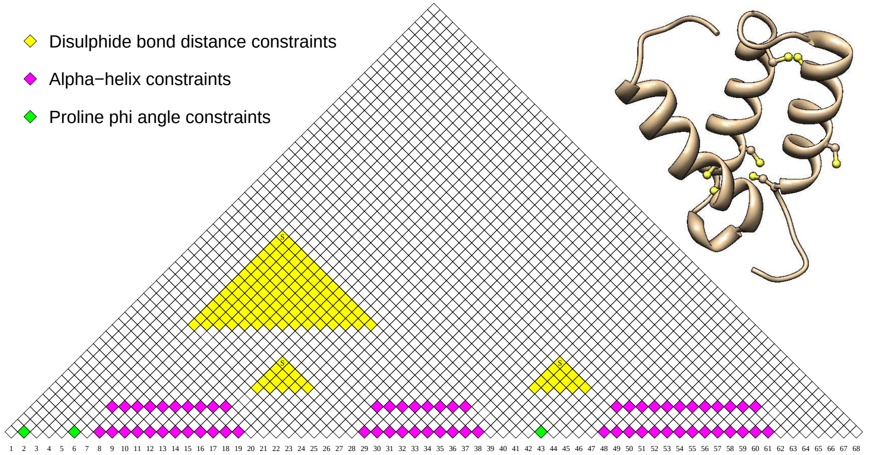
Zipping and assembly with constraints: information used

Protein amino acid residue sequence

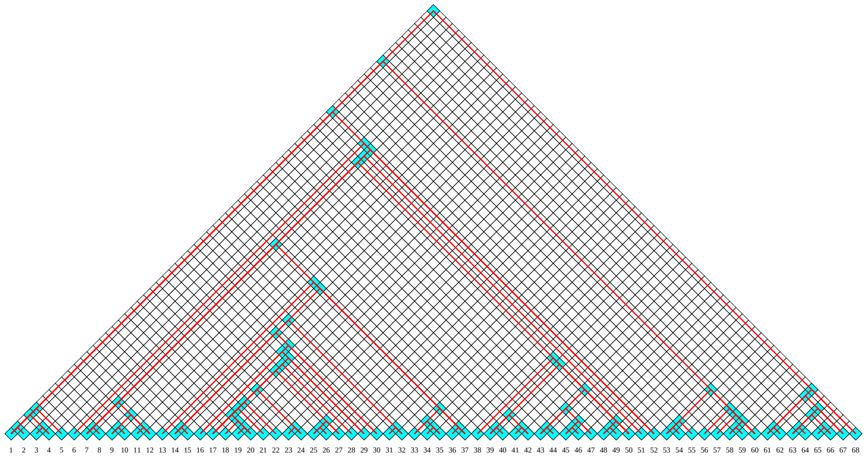
Constraints

- 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
  - ...

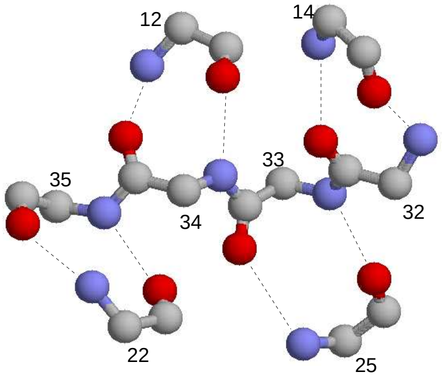
Constraints used in modelling human p8MTCP



Actual cells used in constructing one model



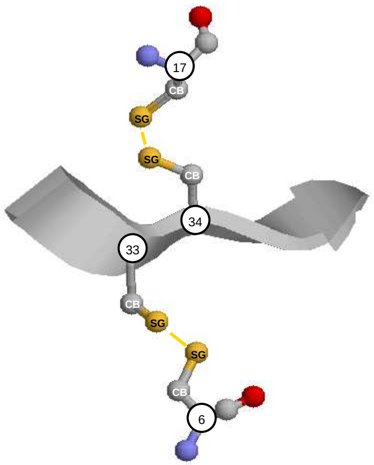
Human β-defensin 6: antiparallel bridges



```
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

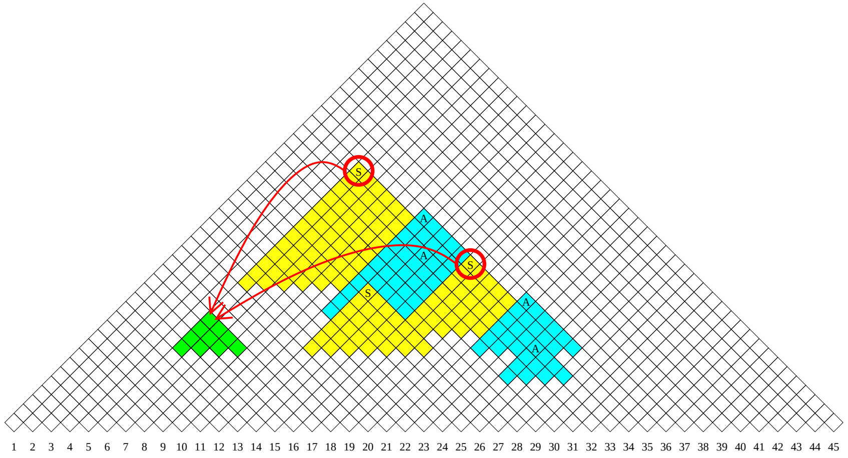
Adjacent residues in a strand



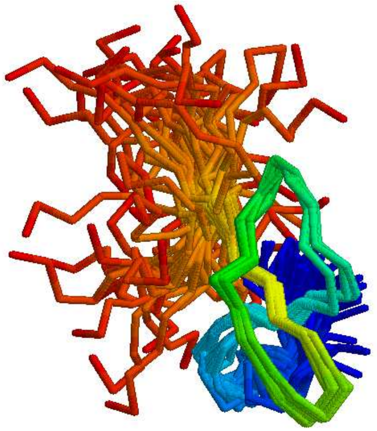
Additional rule:

```
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.
```

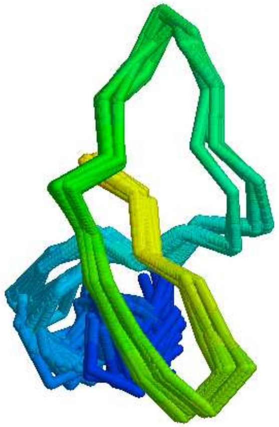
Human β-defensin 6: distance constraints



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

DALI: Distance-matrix ALIgnment

Holm, L. and Sander, C. (1996)
Mapping the Protein Universe
Science vol. 273, 595-602.

The objective of shape comparison in DALI is to assign a one-to-one equivalence between the residues, where non-matching residues can be skipped in either chain.

This is done by finding similar patterns in distance matrices.

Constructing distance matrices (or “contact maps”) is easy;
finding maximal matching sub-matrices is hard.

Two algorithms in DALI

Scan for obvious similarities using a fast (but, in general, less accurate) algorithm, then rescan for more subtle similarities using more sophisticated (but slower) algorithms.

A) Fast heuristic 3D lookup (“hashing”)

Catches easy-to-find structural similarities.

Represent secondary structure elements by 3D line segments; match vector relationships from the query protein with a stored list; when enough matches are found with a database protein, sample a limited set of superpositions.

B) Branch-and-bound algorithm

Guaranteed to find the global optimum, but slower
(worst case: exponential number of steps).

Find the best matching sub-matrices for proteins A and B;
then recursively split the solution sub-space.

Shape comparison in DALI

(i) a suitable representation:

list of Cα atoms described by their x, y and z coordinates.

(ii) an objective function to be optimised:

accommodate the largest possible number of equivalent points within small deviations in position (typically less than 2 to 3 $angstrom$).

(iii) a comparison algorithm:

find matching sub-matrices and merge these into larger consistent blocks of agreement by removing intervening rows and columns.

(iv) appropriate decision rules:

statistical significance of comparison score (Z-score);
equivalent sets of residues (structural alignment);
3D view of the matched parts superimposed.

Problems when searching a protein structure database

(Want to perform all-against-all comparison)

Unequal representation of protein families.

Some redundancy can be eliminated by removing proteins with mutual sequence identity greater than 25%.
But many structurally similar proteins remain.

The problem of domains.

Similar sub-structures recur between several proteins.

Today we can identify sets of domains with distinct folds from resources like CATH and SCOP.