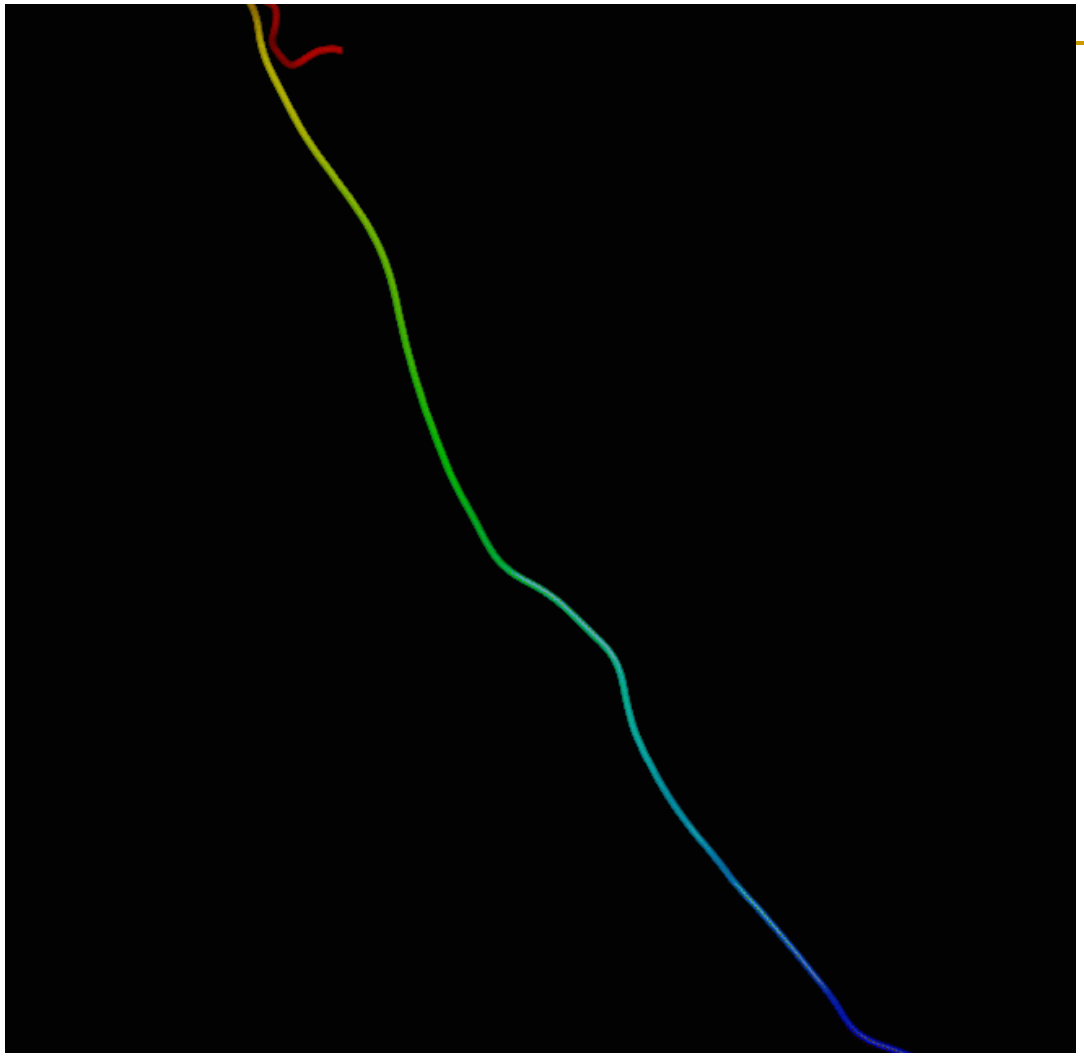
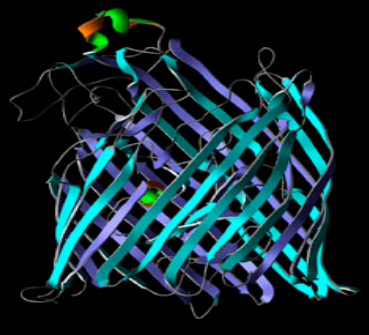


The Protein Folding Problem

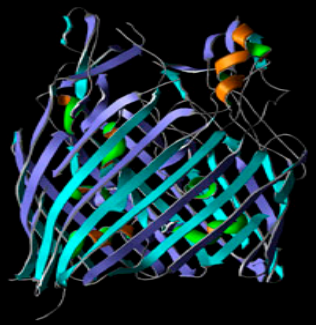
Statistical Mechanics models

Mixed character of the problem :

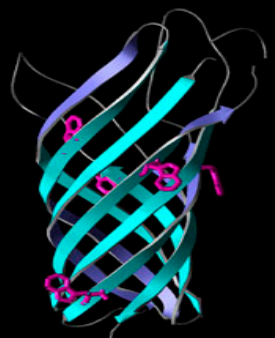
continuous mathematics -- geometry of surfaces &
discrete mathematics -- combinatorics of folds



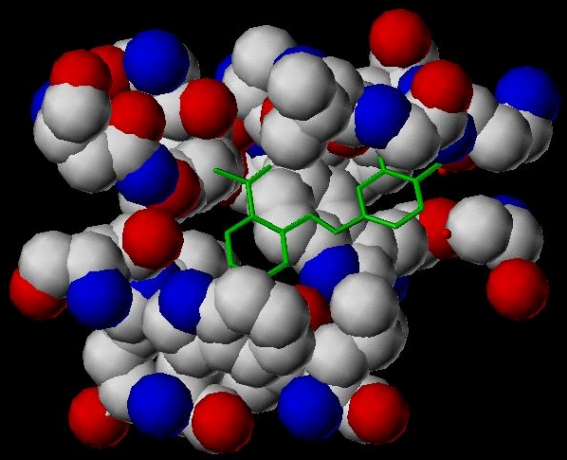
FhuA
(Ferguson et al., 1998)

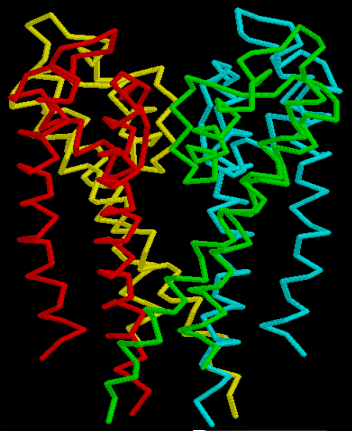


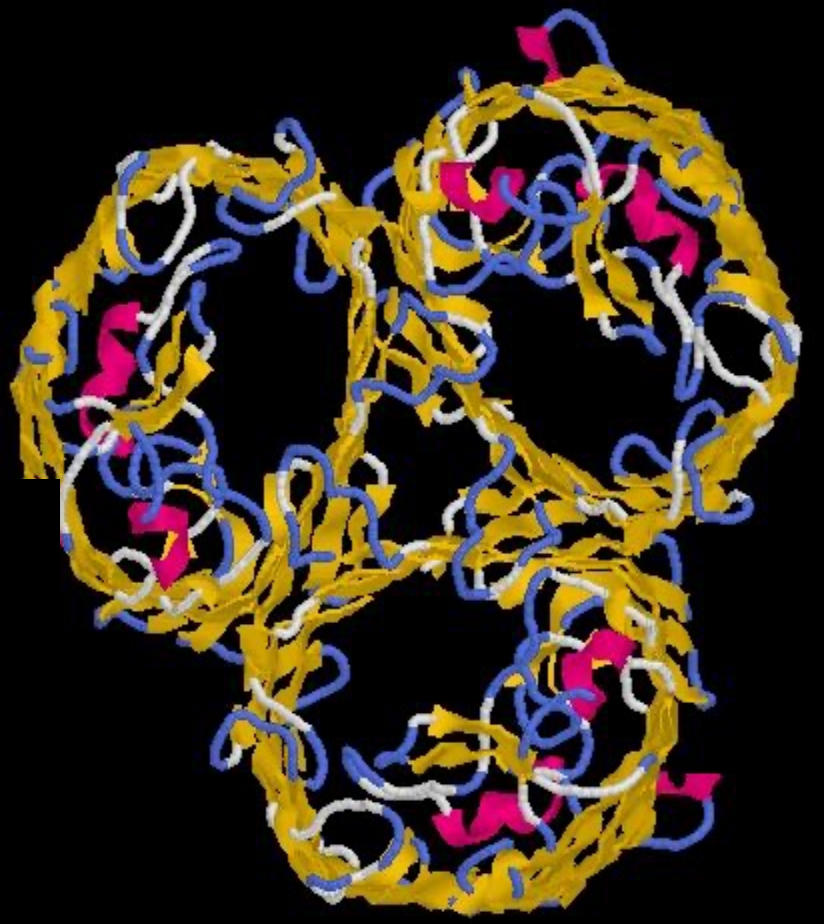
FepA
(Buchanan et al., 1999)

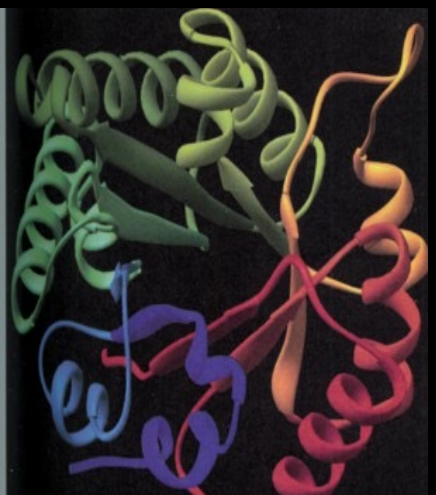


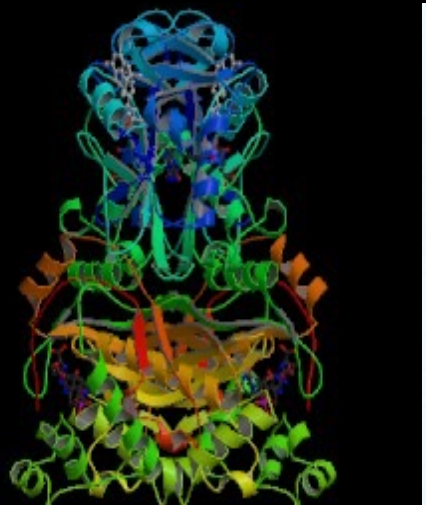
OmpA
(Pautsch & Schulz, 1998)

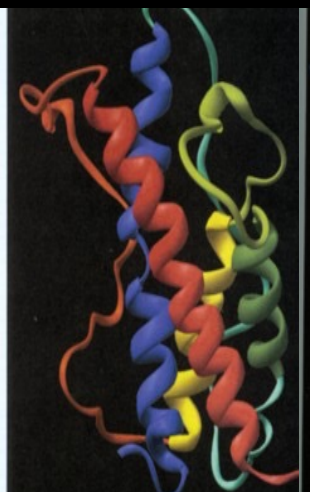




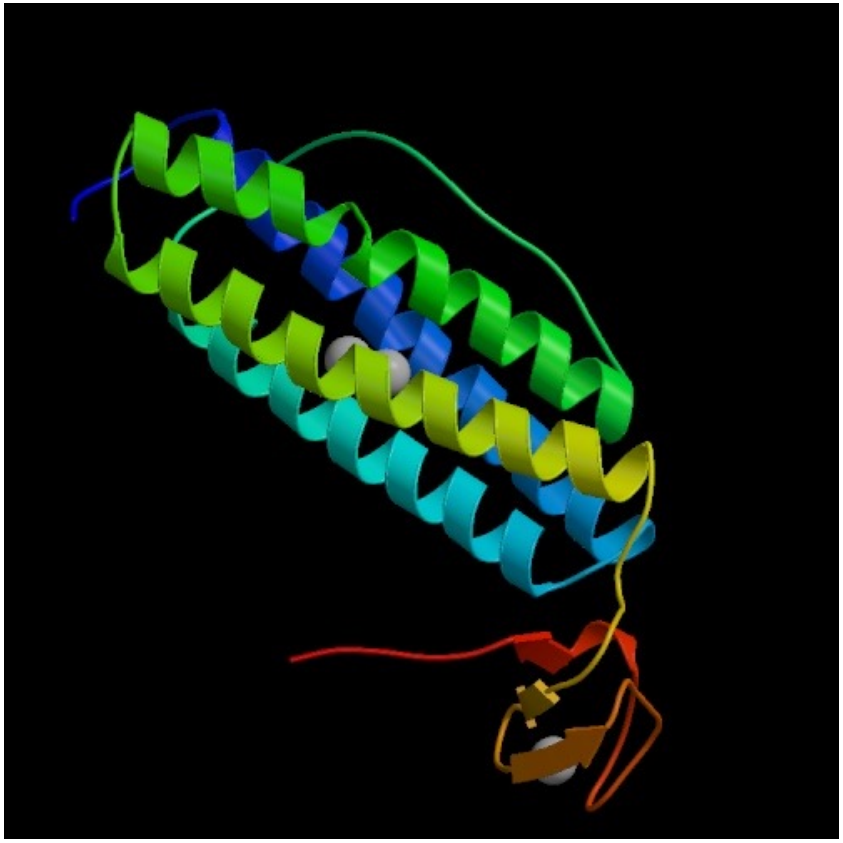




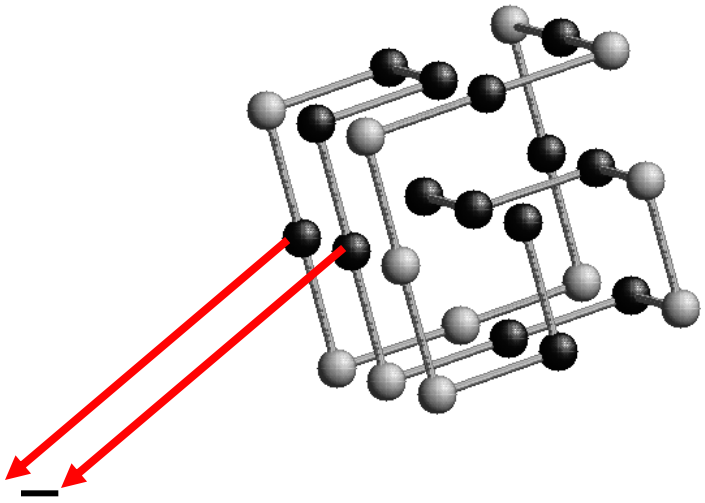


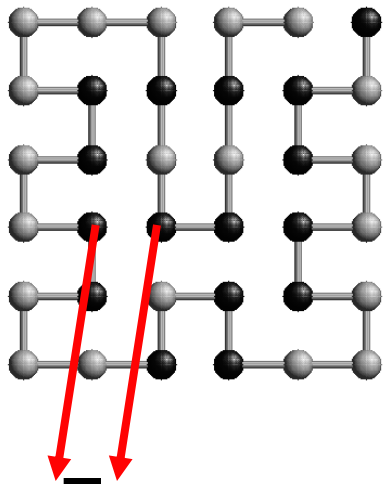


SELF-AVOIDING WALK



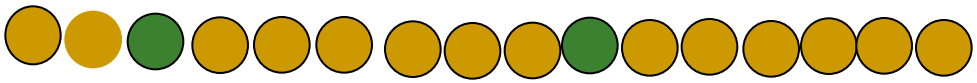
CONTACT





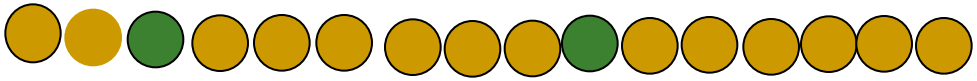
<= 4 Amstrongs

CONTACT MAPS

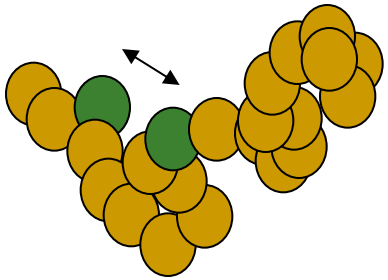


Unfolded protein

CONTACT MAPS

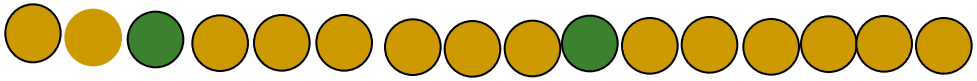


Unfolded protein

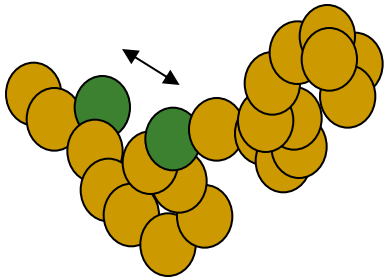


Folded protein = contacts

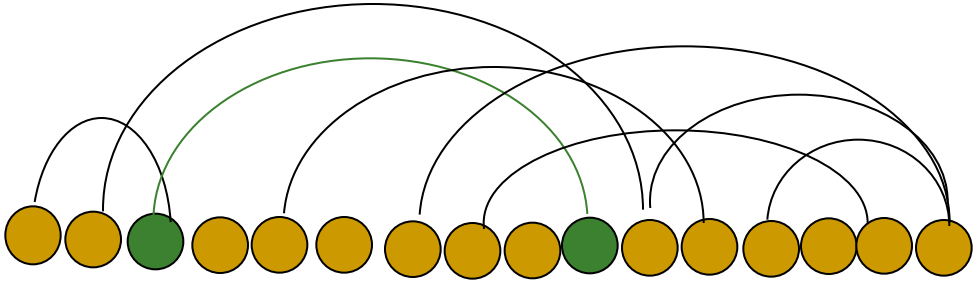
CONTACT MAPS



Unfolded protein

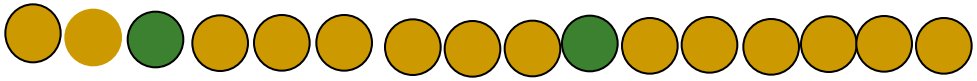


Folded protein = contacts

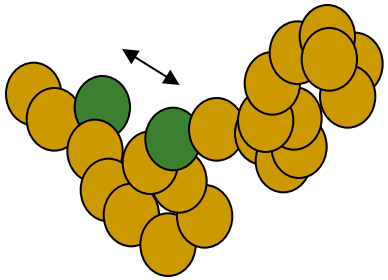


Contact map = graph

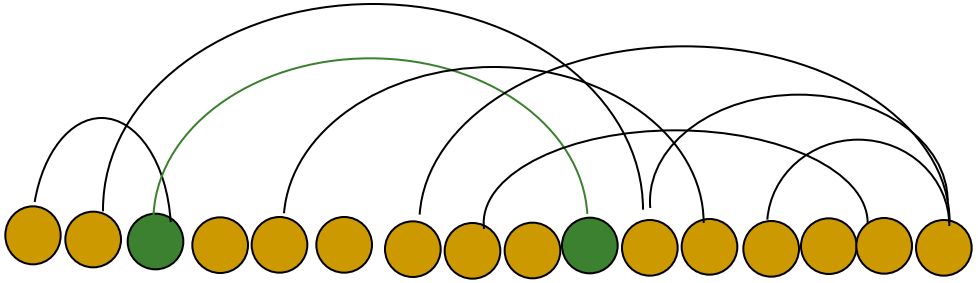
CONTACT MAPS



Unfolded protein



Folded protein = contacts

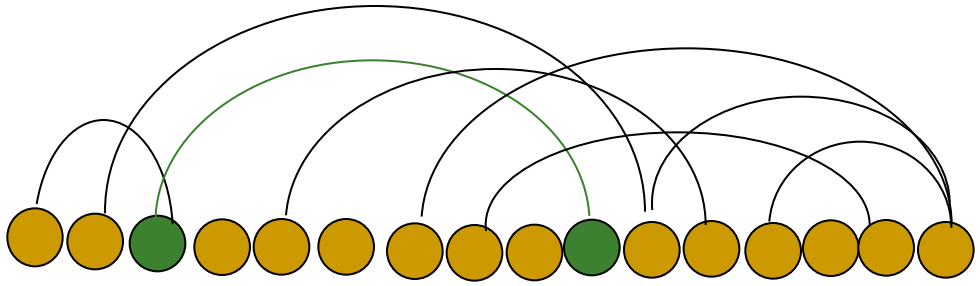


Contact map = graph

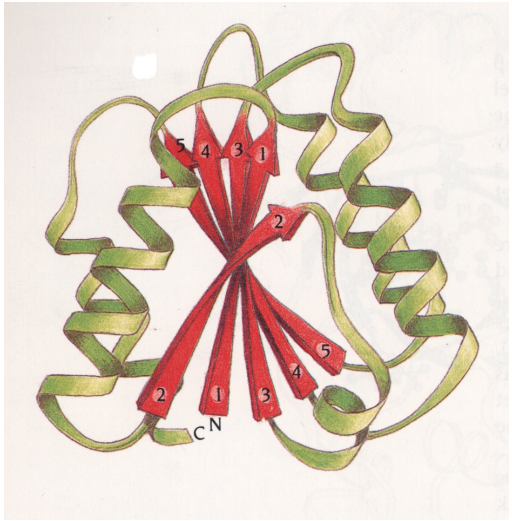
OBJECTIVE: align 3d folds of proteins
= align contact maps

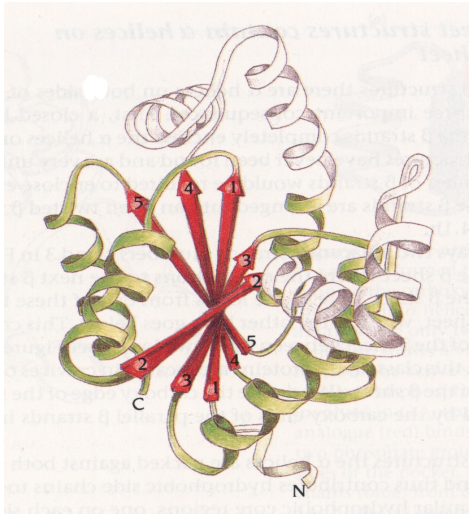
Contact Map

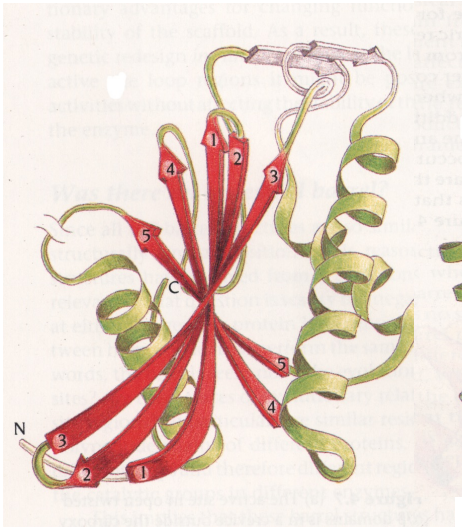
A *contact map* (n, E) is an undirected graph $G = (V, E)$ such that the set of vertices $V = \{1, 2, ..., n\}$ is linearly ordered.

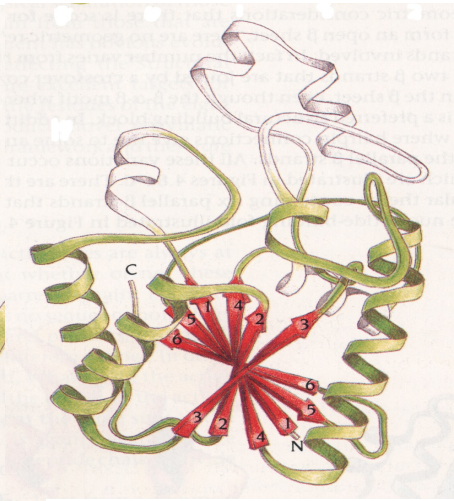


ASILOMAR: Protein Folding -- A Paradigm-Shift









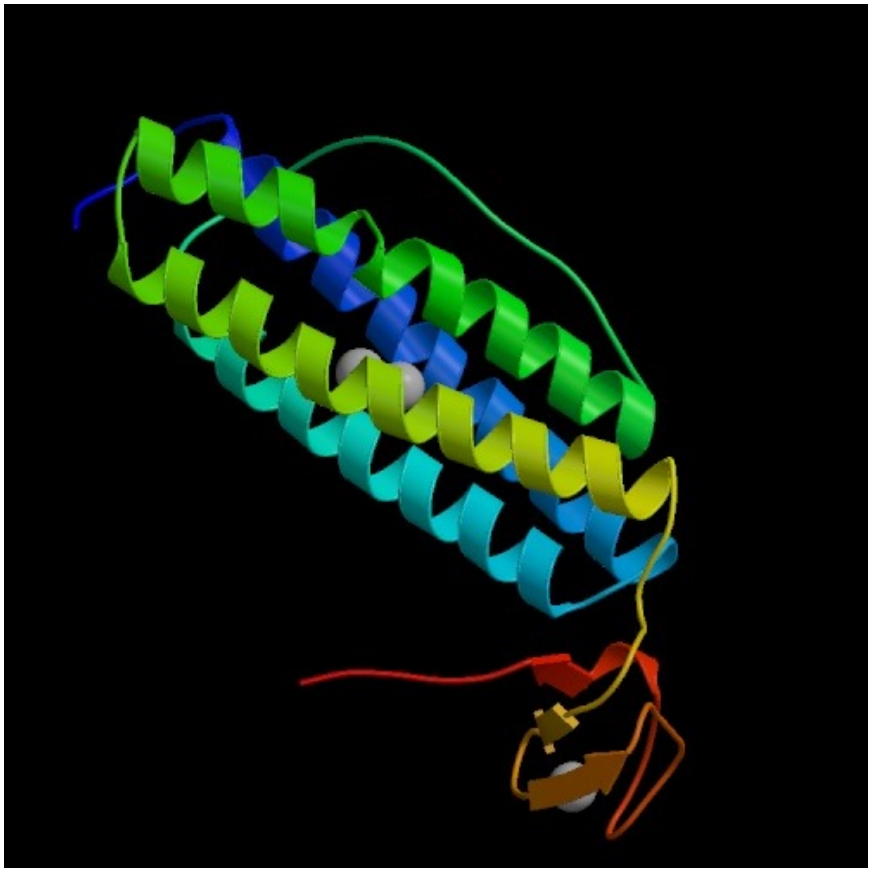
Structure Similarity

Structure Alignment

Fold Recognition

Fold Assignment

FOLD RECOGNITION



Measure of Protein Structure Similarity

- Root-mean-square distance (RMSD)
- Difference of the distance matrices (DDM)
- Contact Map Overlap (CMO)
- Various more or less ad hoc scoring schemes based on local secondary structure, hydrogen bonding pattern, burial status, interaction pattern

The Contact Map Overlap

- No other measure comes even close to satisfying the above list of desiderata

Measure of Protein Structure Similarity

- Root-mean-square distance (RMSD)
- Difference of the distance matrices (DDM)
- Contact Map Overlap (CMO)
- Various more or less ad hoc scoring schemes based on local secondary structure, hydrogen bonding pattern, burial status, interaction pattern

ROOT MEAN SQARE DISTANCE (minimization of distance)

$$RMSD(AB) = \sqrt{\frac{1}{Naligned} \sum_{i=1}^{N} [\bar{r}^A(i) - \bar{R}^B(AB(i))]^2}$$

where

- amino acid i in protein A is aligned (''equivalenced'') with amino acid $AB(i)$ in protein B
- $\bar{r}^A(i)$ is the position of the C_α atom in protein A
- $\bar{R}^B(AB(i))$ is the position of the C_α atom in protein B *after optimal superposition*.
- the set of equivalenced positions $[i, AB(i)]$ is defined before hand in the form an alignment between the two protein sequences.

DIFFERENCE OF DISTANCES

(minimization of distance)

$$DDM(AB) = \frac{1}{Npairs} \sum_{i>j} \mid d_{ij}^{A} - d_{AB(i)AB(j)}^{B} \mid$$

where

- d_{ij}^{A} is the distance between C_α atoms i and j of protein A

- $d_{AB(i)AB(j)}^{B}$ is the distance between C_α atoms $AB(i)$ and $AB(j)$ of protein B

- the equivalenced pairs $[i, AB(i)]$ need to be specified in advence as an alignment between the two protein sequences

CONTACT MAP OVERLAP

(maximization of similarity)

$$CMO(AB) = \frac{1}{Ncontacts} \sum_{i>j} C^A_{ij} C^B_{AB(i)AB(j)}$$

where

- C^A is the contact map (matrix) of protein structure A; $C^A_{ij} = 1$ if and only if i and j are in contact

- Similarly C^B for protein B

Conceptual Difficulties

- Some notorious non-robust
- Very little relationship between edit distance of two proteins and their 3D similarity
- Hydrophobic-hydrophilic character of residues is often not reflected in distance calculations
- Most fail to to account for the “excluded volume” aspect, I.e., the protein backbone is a self-avoiding walk
- Computation of similarity of measures require solutions to intractable optimization problems
- The optimization draw their complexity from the non-locality of their scoring function and the handling of insertions and deletions
- All existing structural alignment algorithms use ad hoc simplifications either of their scoring function or search procedure

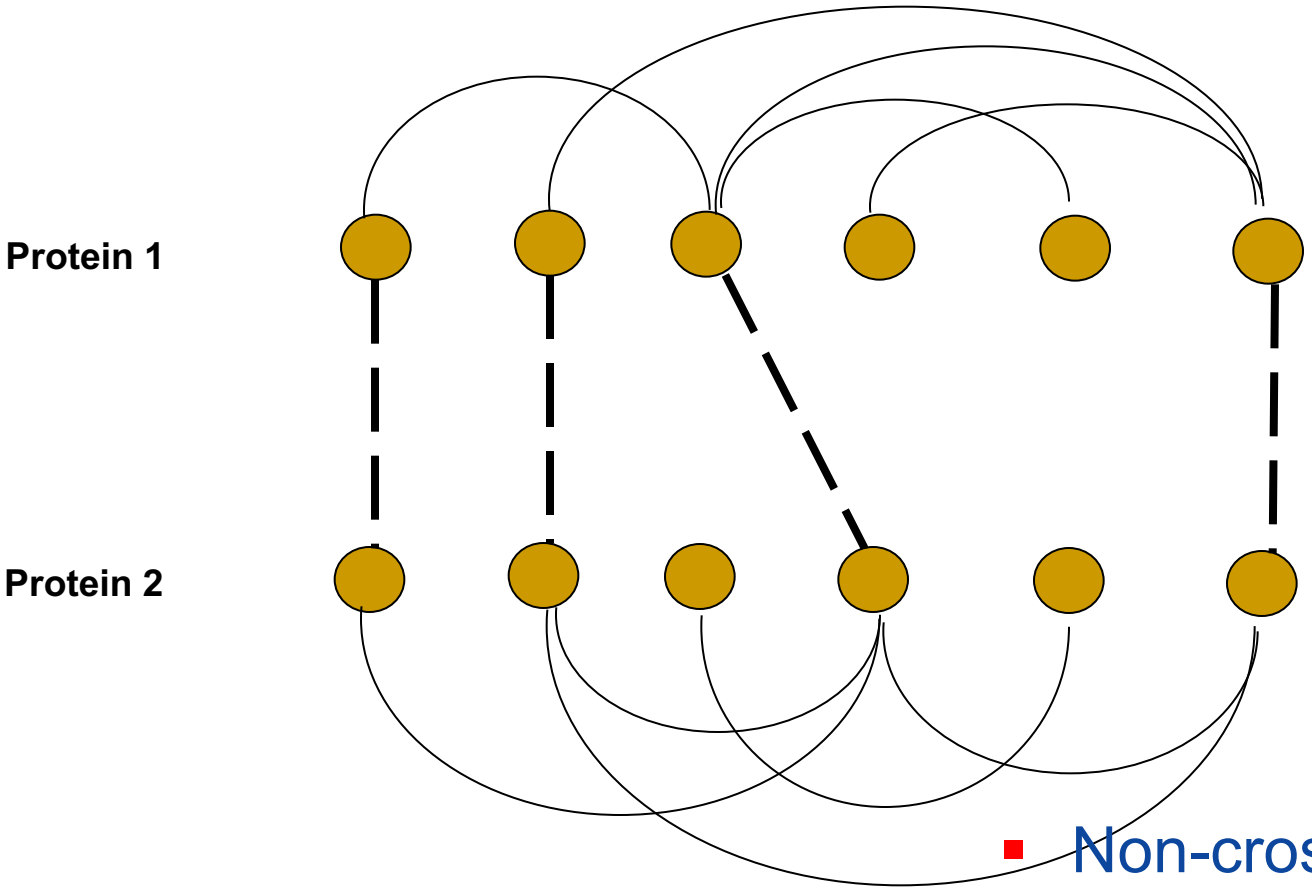
AXIOMS/Desiderata for a structural similarity measure

1. **Not penalize to heavy indels**
2. **Reasonably robust, I.e., small perturbations of the definition should not make too much difference in the measure**
3. **It should be easy to compute, or at least rigorously approximated**
4. **It should be able to discover both local and global alignments**
5. **It should take into accounts the self-avoiding walk nature of the backbone**
6. **It should be subject to empirical studies on the Protein Data Bank (PDB) data**
7. **Even for a "perfect" measure it will difficult to replace entrenched measures used for years by protein scientists. Acceptance in the field is thus a further desideratum**

The Contact Map Overlap

- No other measure comes even close to satisfying the above list of desiderata

Contact Map Overlap Alignment

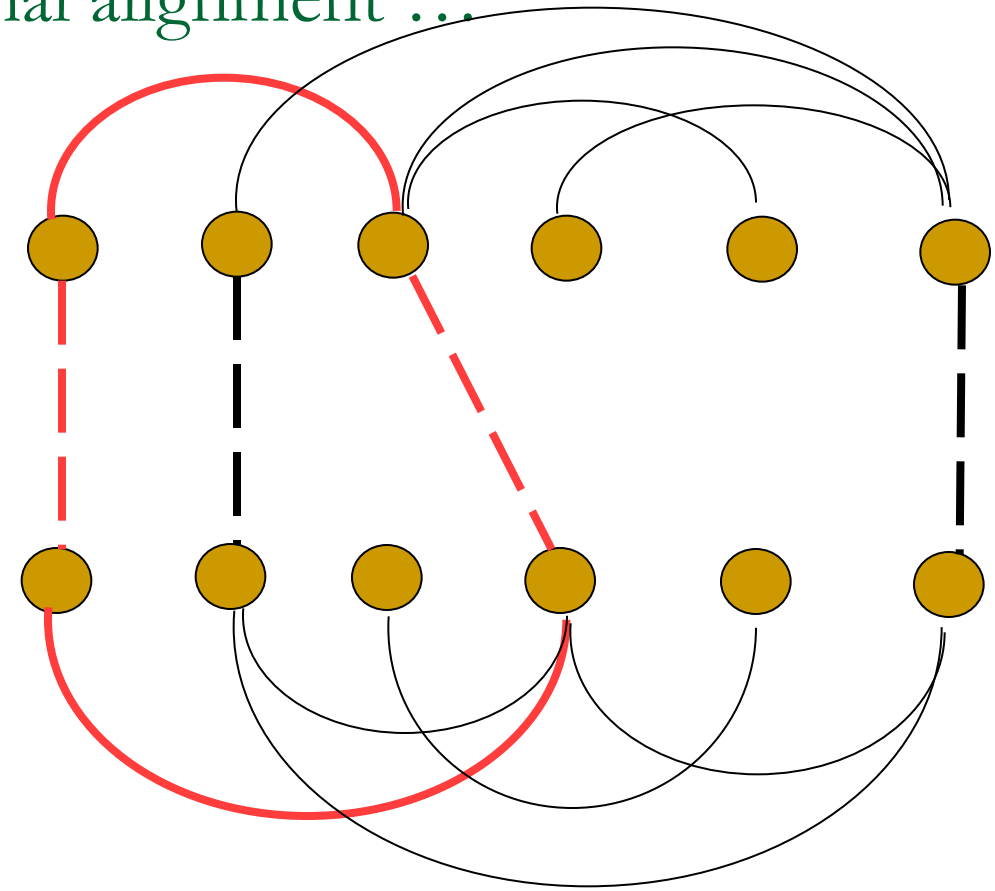


- Non-crossing alignments

non-crossing map between residues in protein 1 and protein 2

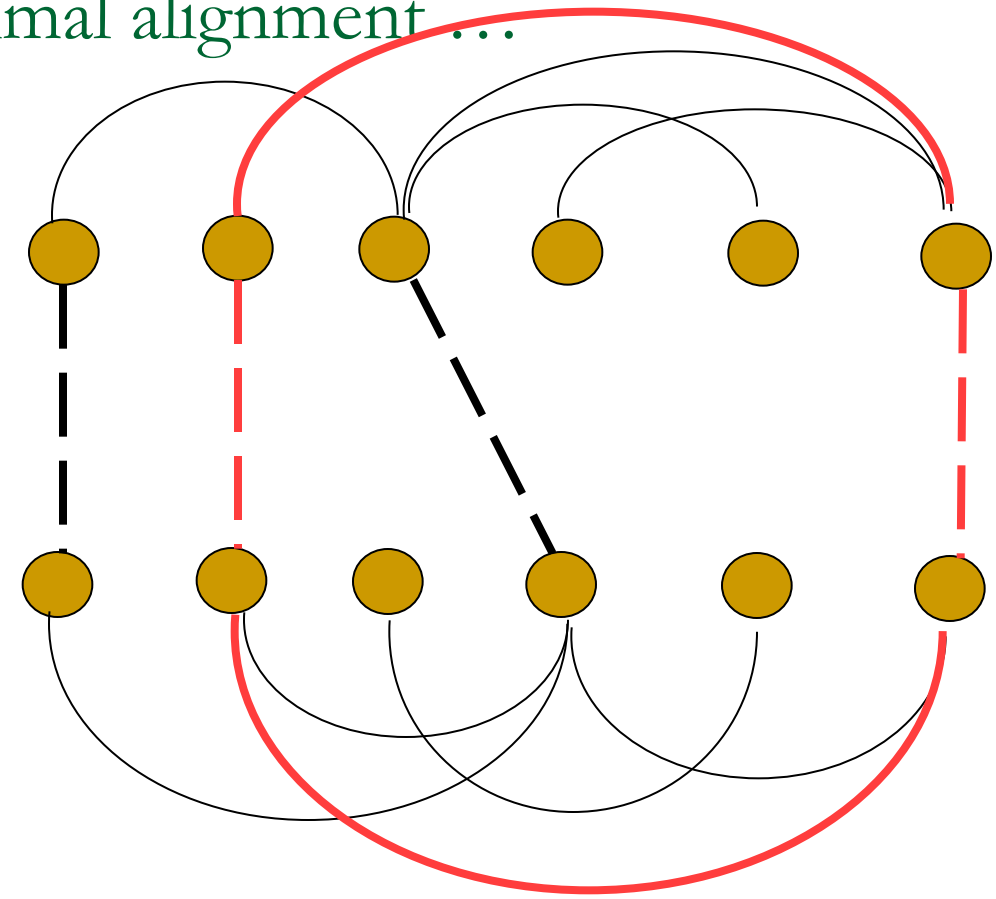
Contact Map Overlap Alignment

The optimal alignment …



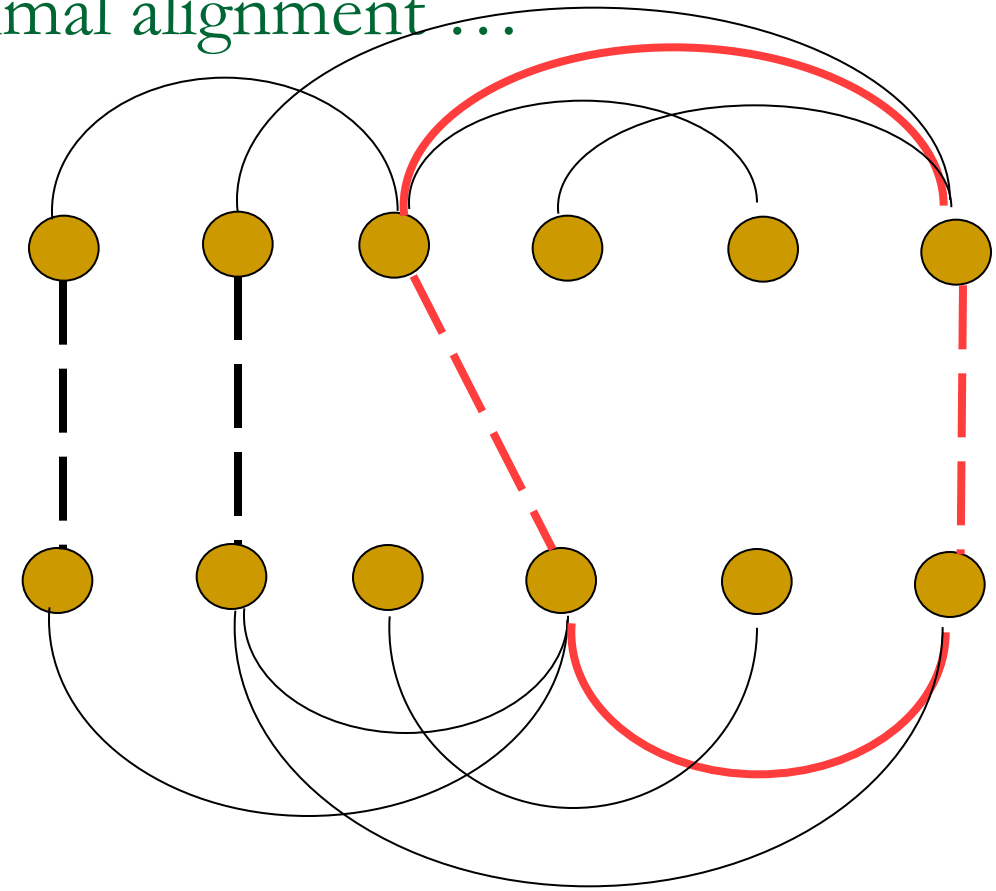
Contact Map Overlap Alignment

The optimal alignment …



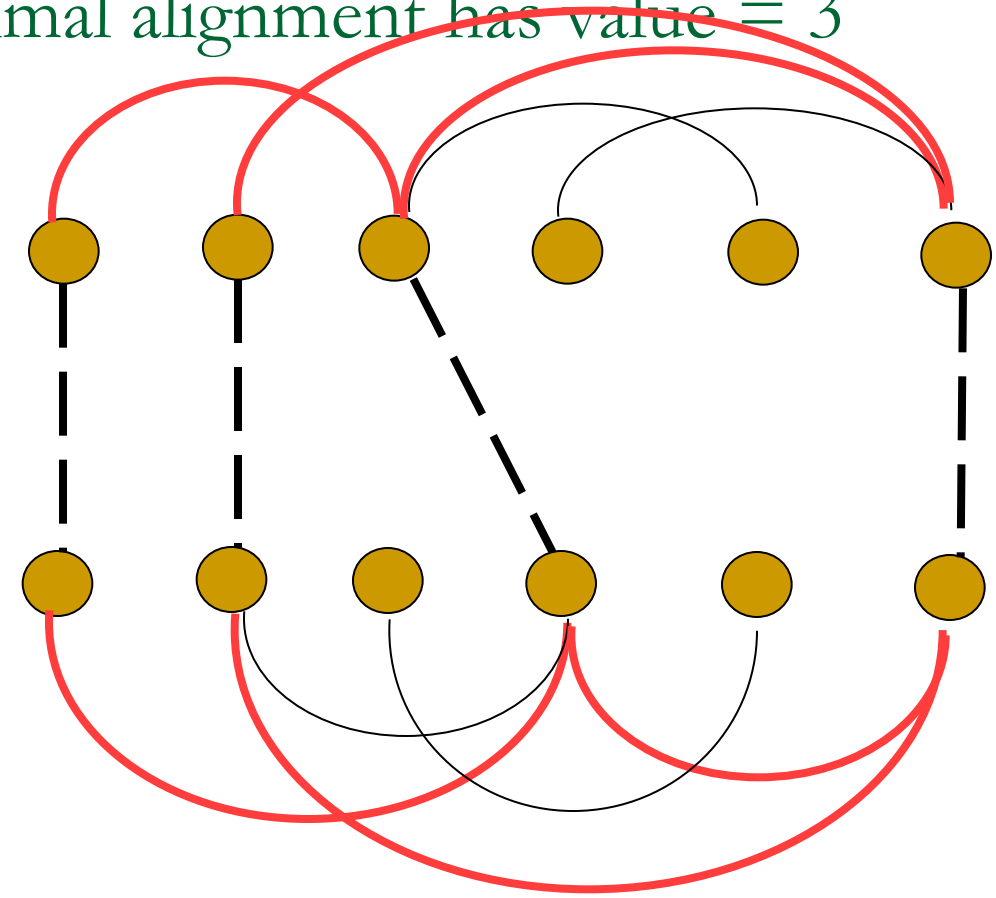
Contact Map Overlap Alignment

The optimal alignment …



Contact Map Overlap Alignment

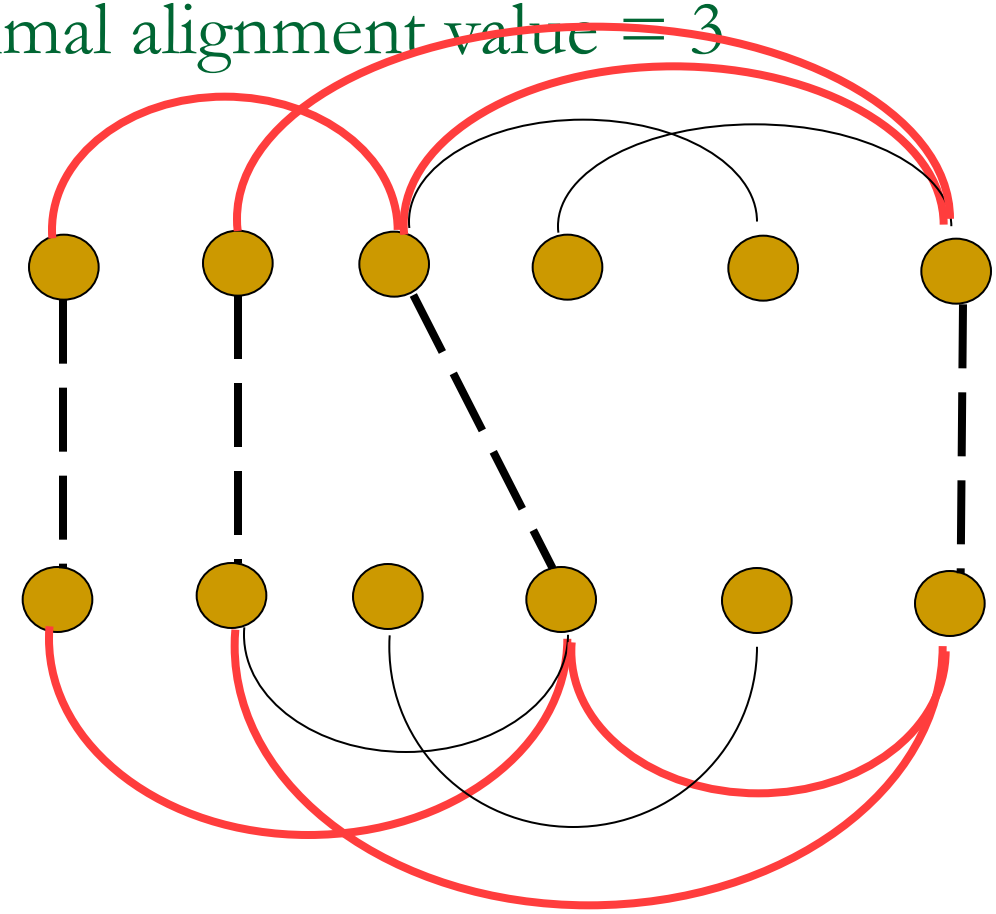
The optimal alignment has value = 3



Value = 3

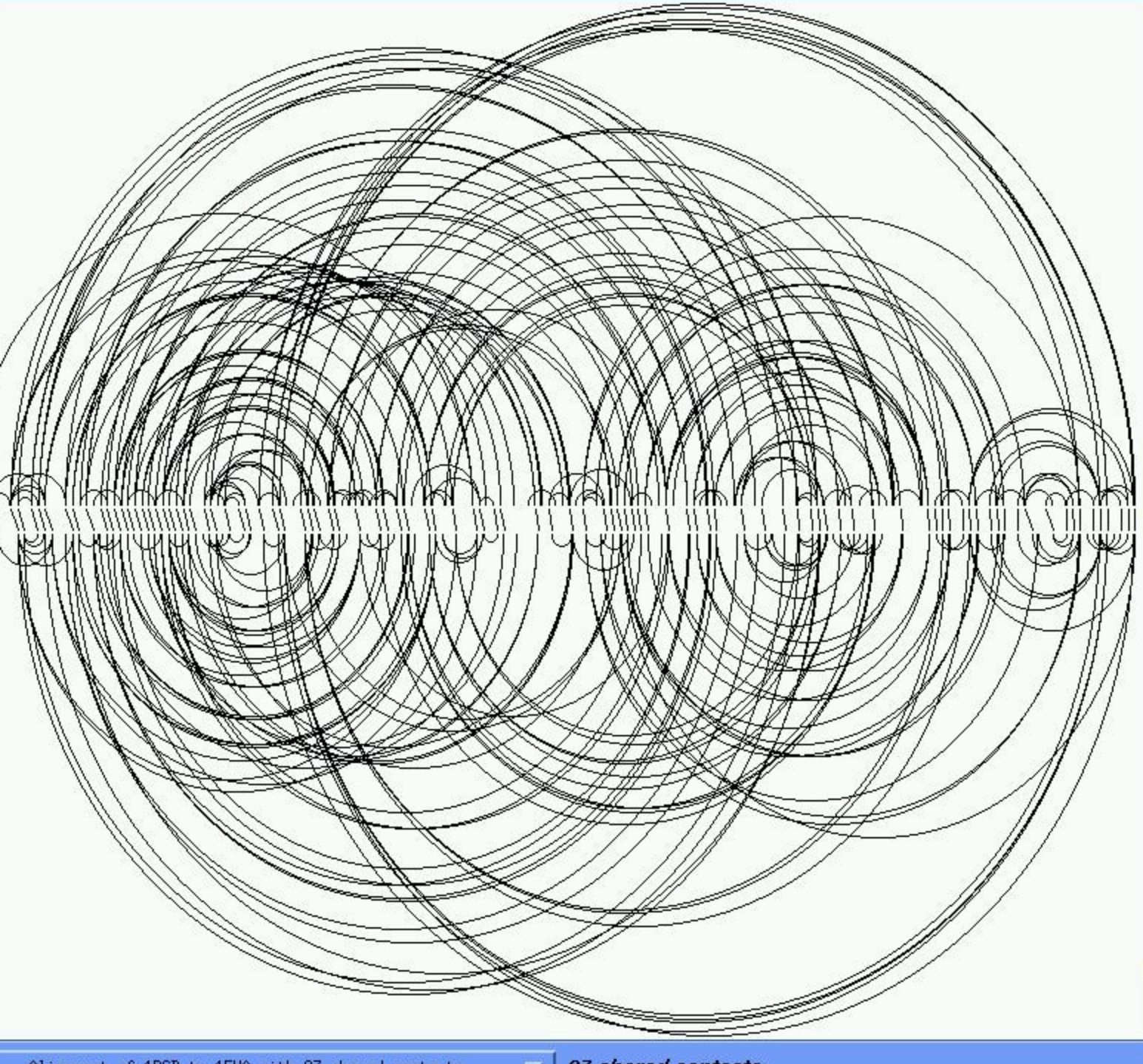
Contact Map Overlap Alignment

The optimal alignment value = 3



Value = 3

We want to maximize the value



{1RCD} IRON STORAGE

{1FHA} IRON STORAGE

The Contact Map Overlap

- No other measure comes even close to satisfying the above list of desiderata

Skolnick Clustering Test

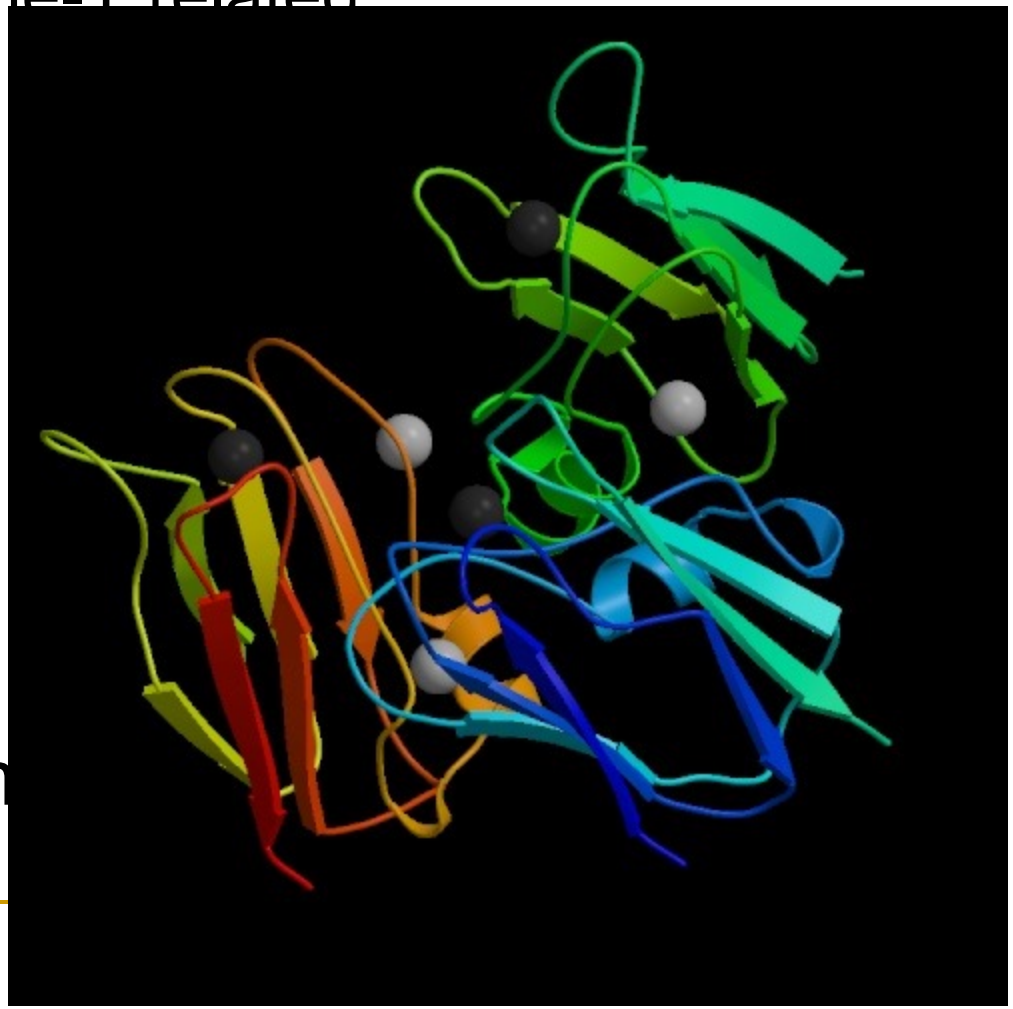
Skolnick Test

- Four Families
  1. Flavodoxin-like fold Che-Y related
  2. Plastocyanin
  3. TIM Barrel
  4. Ferratin

- alpha-beta
- 8 structures
- up to 124 residues
- 15-30% sequence sim
- < 3Å RMSD

Skolnick Test

- Four Families
  1. Flavodoxin-like fold Che-Y related
  2. Plastocyanin
  3. TIM Barrel
  4. Ferratin



- beta
- 8 structures
- up to 99 residues
- 35-90% sequence sim
- < 2Å RMSD

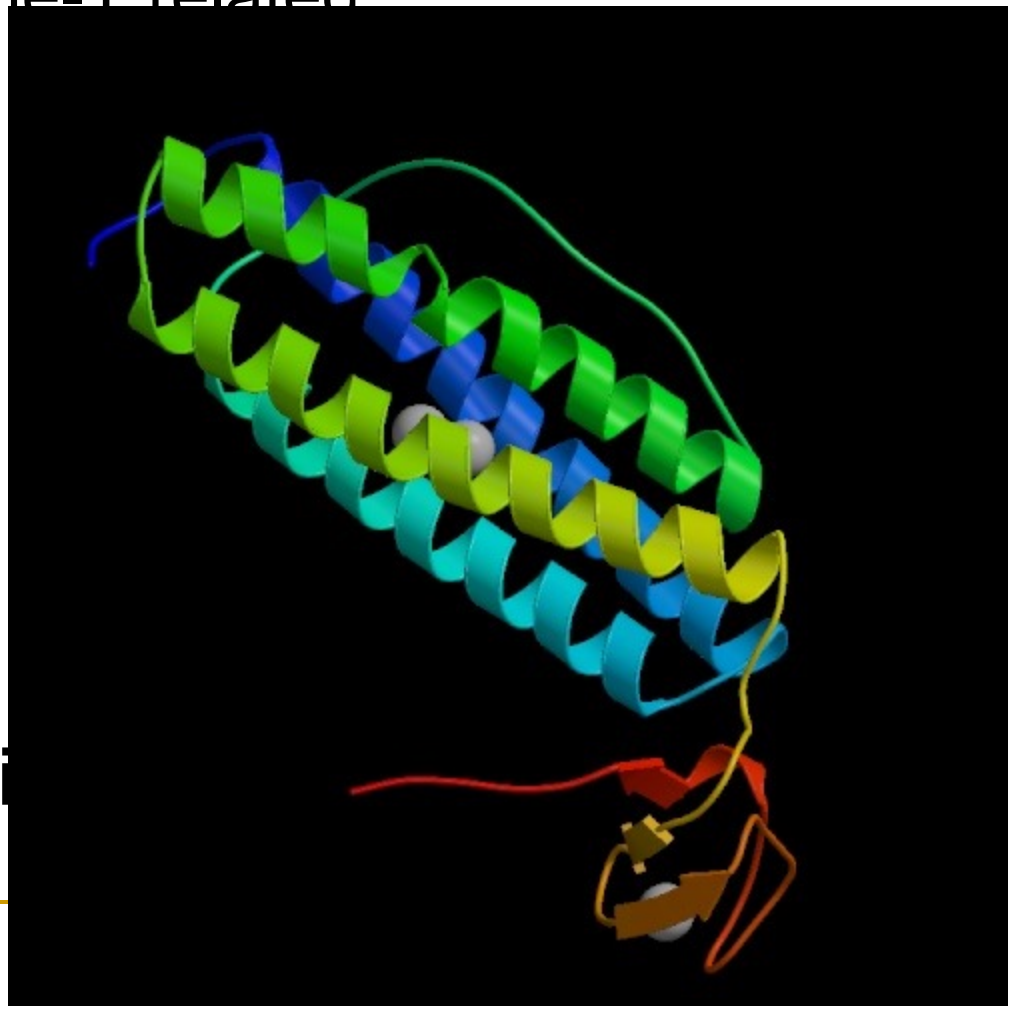
Skolnick Test

- Four Families
  1. Flavodoxin-like fold Che-Y related
  2. Plastocyanin
  3. TIM Barrel
  4. Ferratin

- alpha-beta
- 11 structures
- up to 250 residues
- 30-90% sequence sim
- < 2Å RMSD

Skolnick Test

- Four Families
  1. Flavodoxin-like fold Che-Y related
  2. Plastocyanin
  3. TIM Barrel
  4. Ferratin



- alpha
- 6 structures
- up to 170 residues
- 7-70% sequence simi
- < 4Å RMSD

Clustering

Define score(P_1, P_2) as

$$0 <= \frac{\text{\# shared contacts}}{\text{Min \# of contacts of } P_1, P_2} <= 1$$

Put P1, P2 in same family if score(P1, P2) >= *threshold*

If P1, P2 too big, use G.A. and local search to compute score

L.P. gives then bounds:

HEURISTICS score <= OPT score <= LP bound

and we know how far off OPT we are

Skolnick Test Results

- Performance
  - 528 alignments
  - 1.3% false negative
  - 0.0% false positive