

# Five hierarchical levels of sequence-structure correlations in proteins

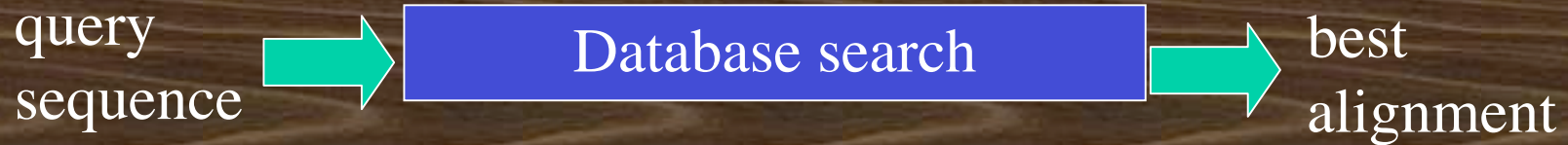
**Chris Bystroff**  
**Rensselaer Polytechnic**  
**Institute**  
**Troy, New York, USA**

# What does structure prediction tell us about the physics of folding?

Check one:

- A.** If we can predict protein structures, then we know how proteins fold.
- B.** If we know how proteins fold, then we can predict protein structures.

# Two ways to predict protein structure...



(statistics)



(physics)



...two very different *Underlying principles*

query  
sequence



*Darwin:*

Proteins with a common  
ancestor have the same  
fold.



best  
alignment

millions of years

query  
sequence



*Boltzmann:*

Proteins adopt a minimum  
the free energy  
conformation.

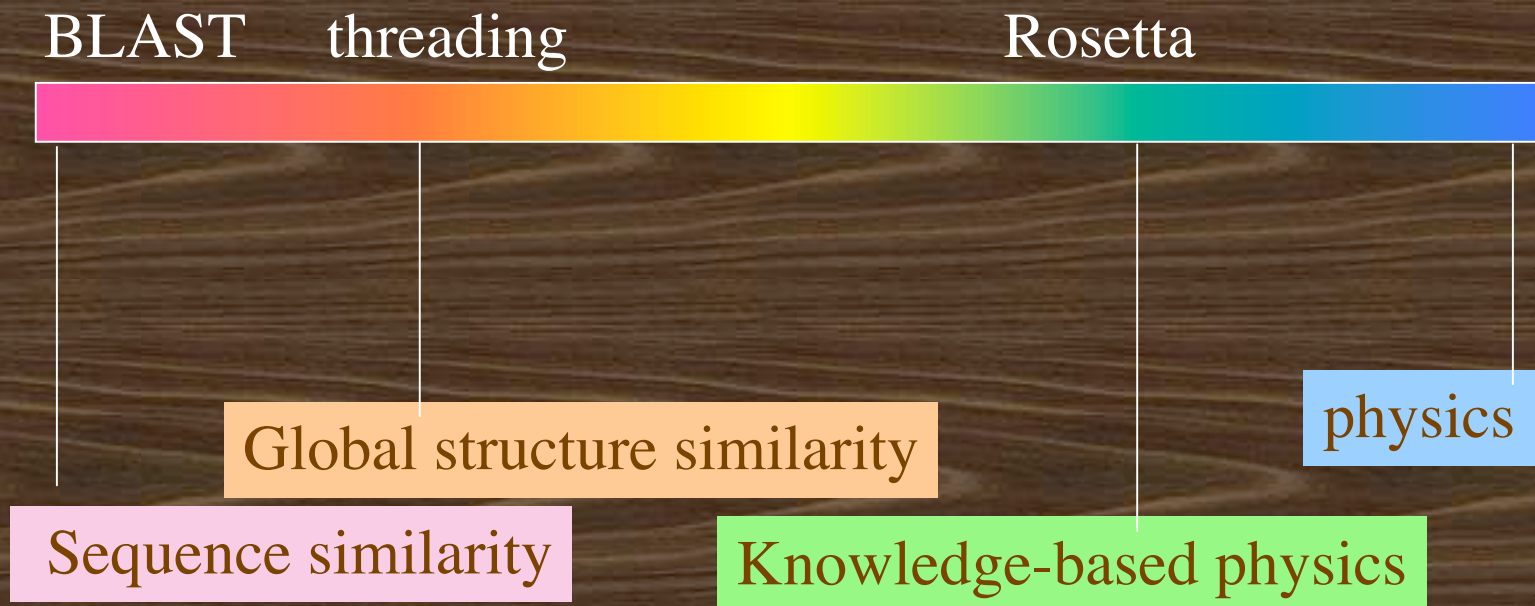


lowest  
energy

microseconds to seconds



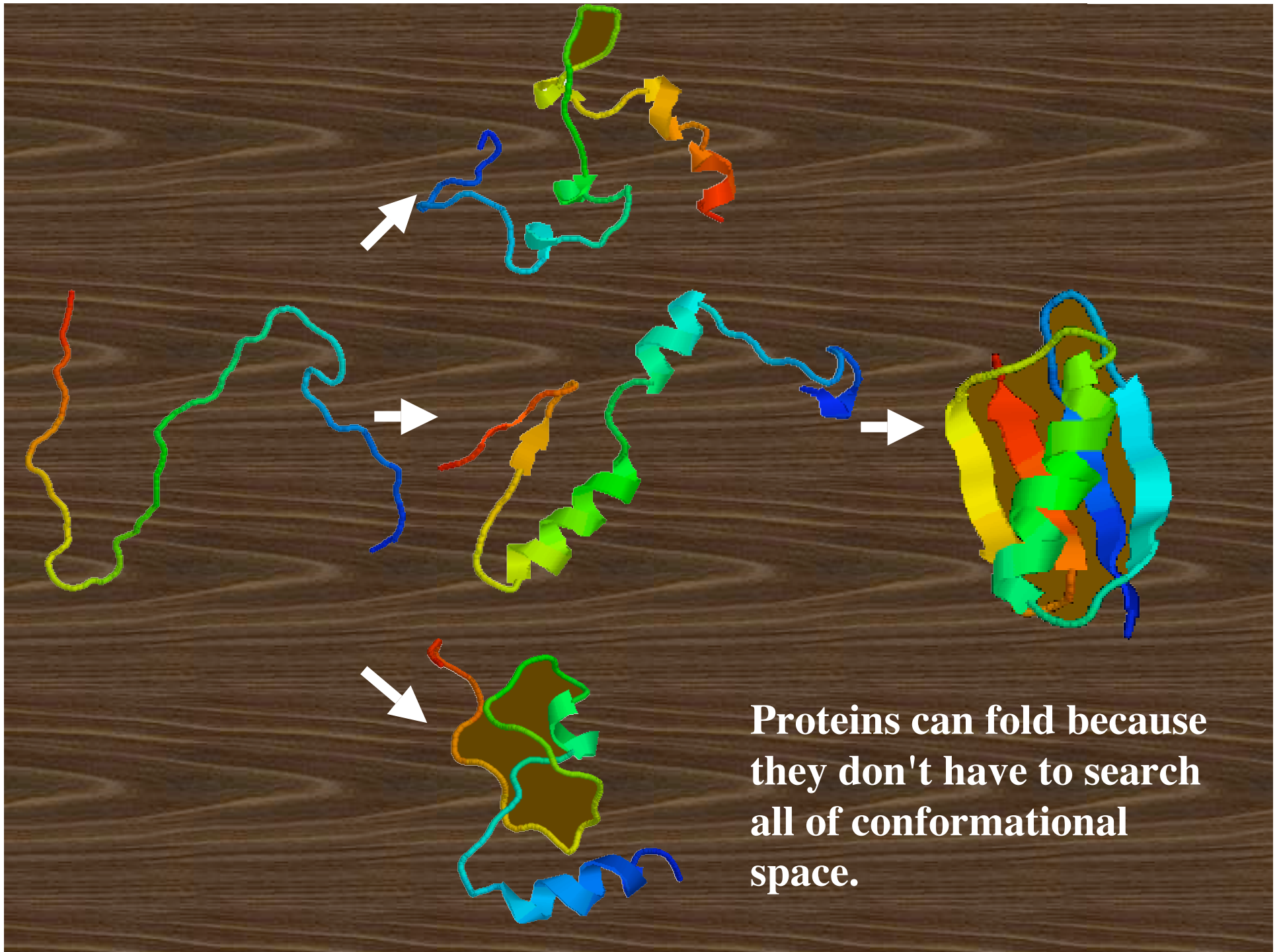
# Darwin versus Boltzmann. Do hybrid models make sense?



We know proteins fold via pathways.



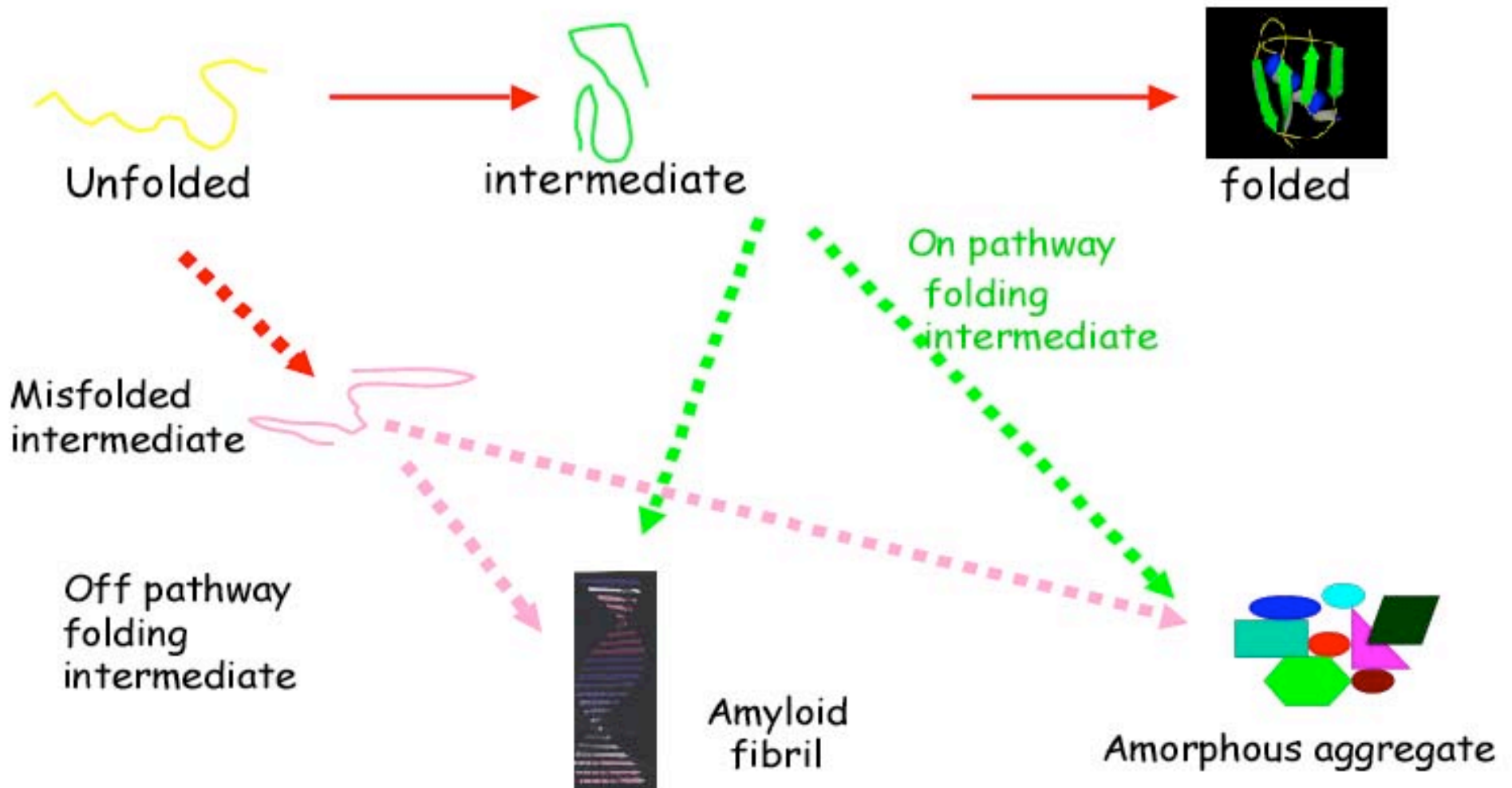
**local structure first, eliminating  
alternate pathways, then global**



**Proteins can fold because they don't have to search all of conformational space.**



# Protein Misfolding



# Protein Misfolding diseases

Alzheimer's disease

Creutzfeldt-Jakob Disease (CJD)\*

Scrapie\*

Kuru\*

Huntington's Disease

Parkinson's Disease

Type-2 diabetes

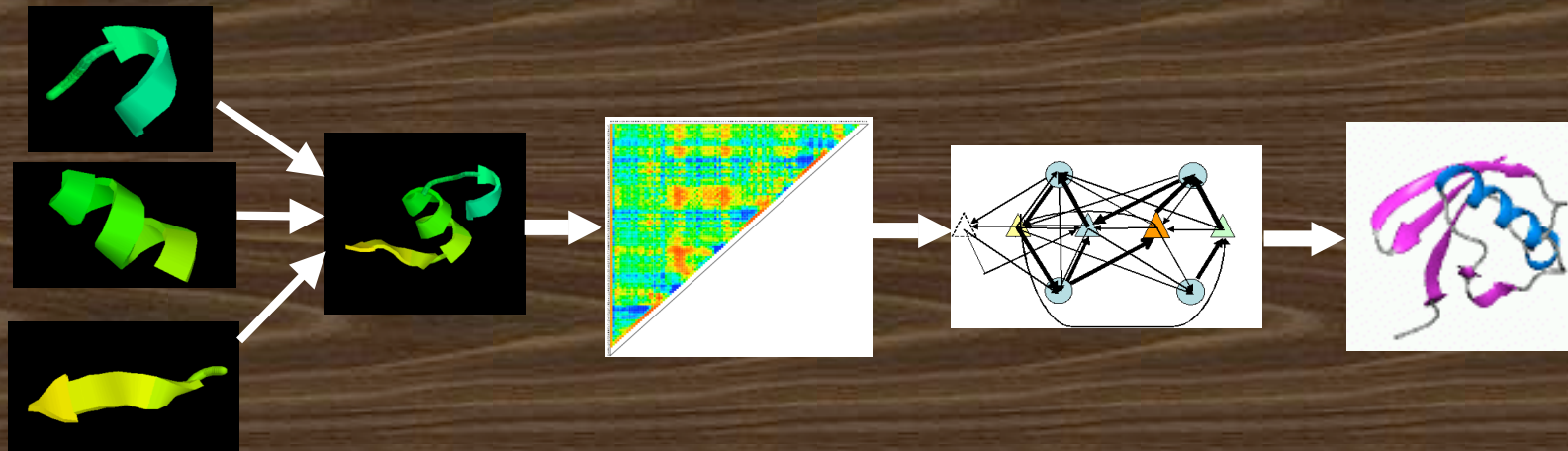
Familial Amyloid Polyneuropathy (FAP)

\*Prion-linked

# The goal: understand protein folding pathways

By modeling the ways proteins fold we can:

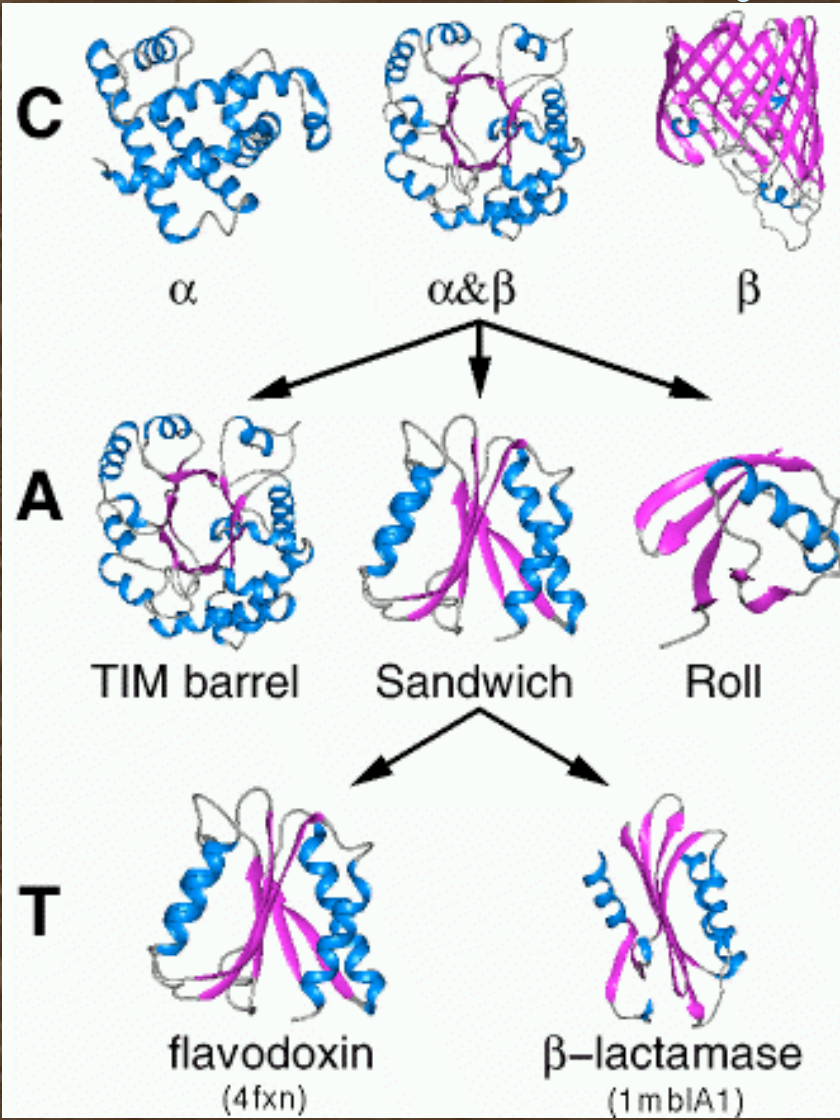
- (1) Predict the structure from the sequence
- (2) Predict the effects of any mutation
- (3) Design a new structure with a new function





# We know that proteins have a hierarchy of structural similarity...

Class



conserves...

2° content

Architecture

packing of 2°

Topology\*

chain connectivity

\*Fold recognition algorithms work at this level

Image borrowed from CATH database

# Can we use the database to make models for folding pathways?

early



late

## Steps along the folding pathway:

- (1) Initiation
- (2) propagation
- (3) condensation
- (4) molten globule
- (5) native state motifs

## Steps in data mining:

- local motifs
- extended local motifs
- pairs of motifs
- multiple motifs
- aligned multiple

# Heirarchical level 1: Folding initiation site motifs

Non-homologous sequences

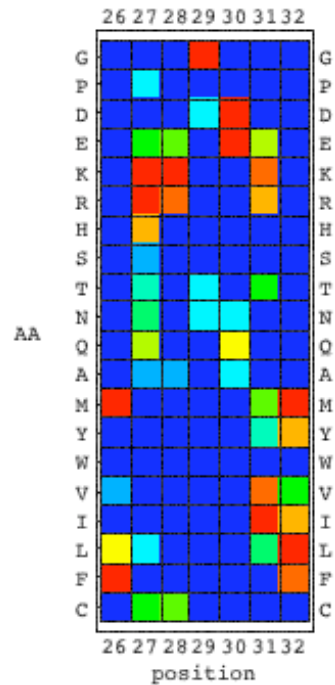
recurrent  
sequence

HDFPIEGGDSPMQTIFFWSNANAKLSHGY  
CPYDNIWMQTIFFNQSAAVYSVLHLIFLT  
IDMNPQGSIEMQTIFFGYAESA  
ELSPVVNFLEEMQTIFFISGFTQTANS  
INWGSMQTIFFEEWQLMNVMDKIPS  
IFNESKKGIA MQTIFFILSGR  
PPPMQTIFFVIVVNYNESKHALWCSVD  
PMMWNLMQTIFFISQQVIEIPS  
MQTIFFVFSHDEQMCLKGLKGA

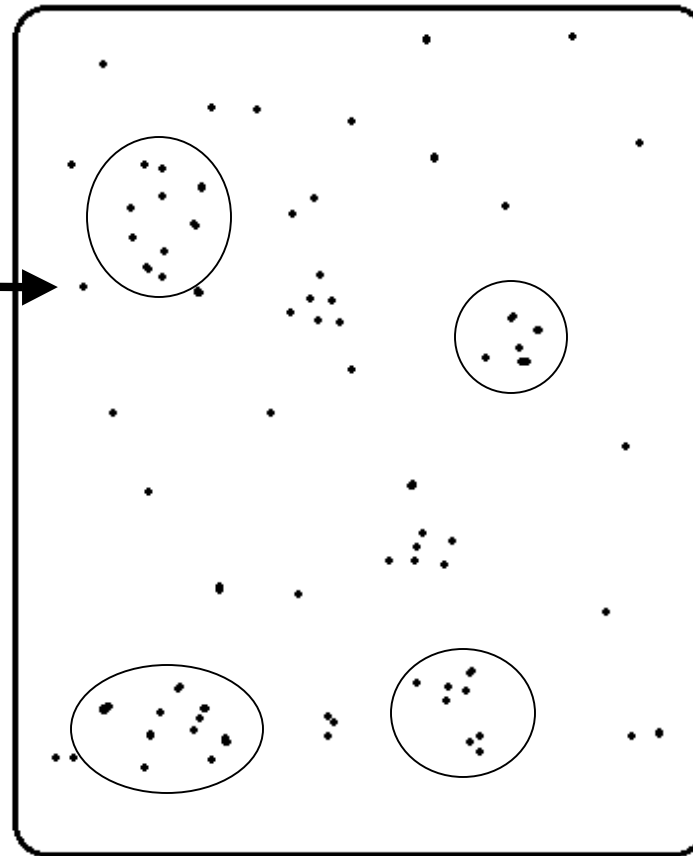
Is it a recurrent structure?



# Clustering sequence profiles to find recurrent patterns



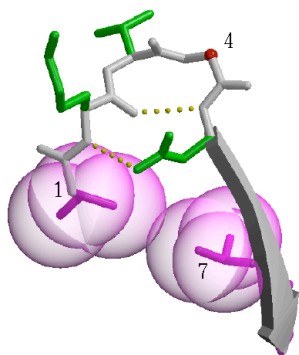
Each dot  
represents a  
short profile



similarity metric (product of log-likelihood ratios)

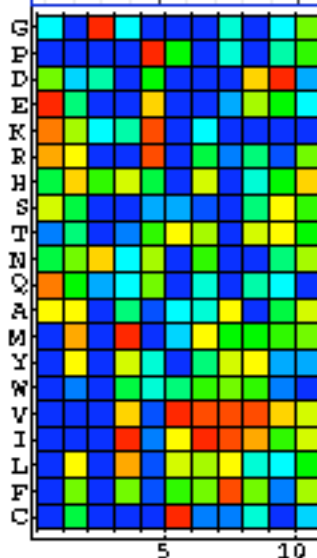
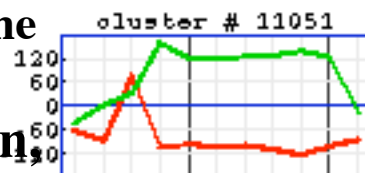
$$D(p, q) = \sum_j \sum_i LLR(p_{ij}) LLR(q_{ij})$$

# The I-sites Library



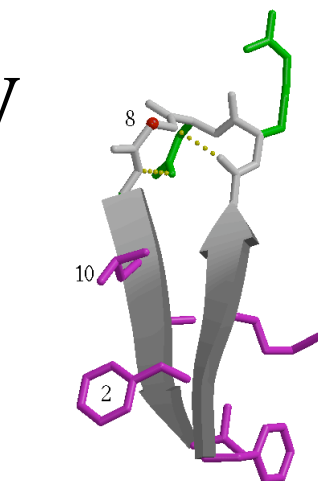
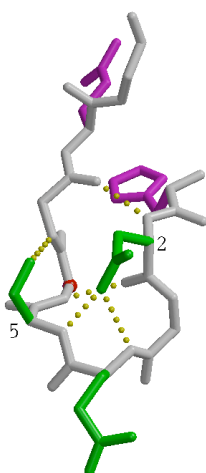
**diverging type-2  
turn**

Backbone  
angles:  
 $\psi$ =green  
 $\phi$ =red

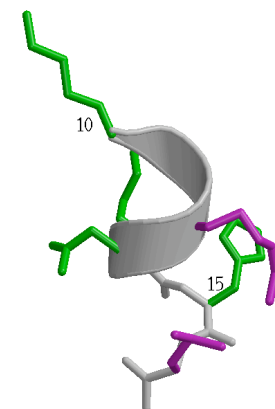


Amino acids  
arranged  
from non-  
polar to polar

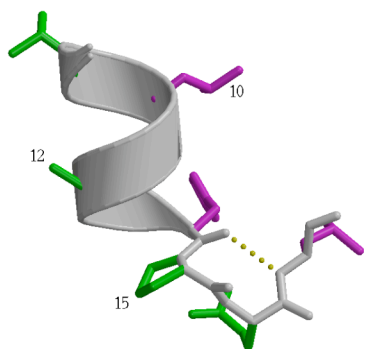
**Serine  
hairpin**



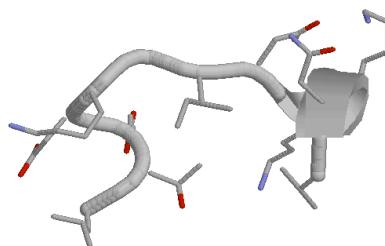
**Type-I  
hairpin**



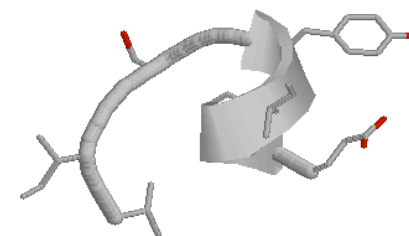
**Frayed  
helix**



**Proline helix C-cap**

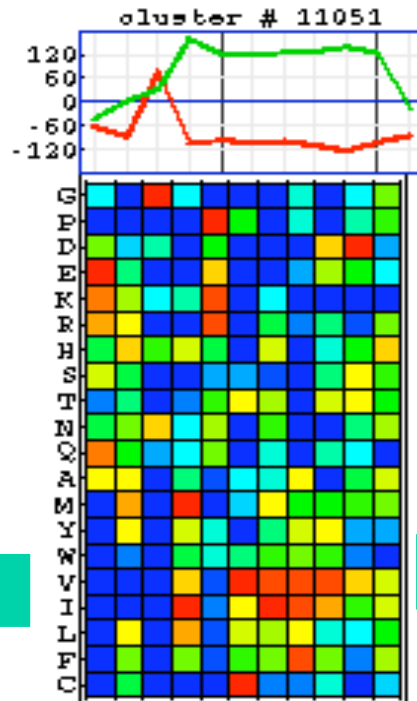


**alpha-alpha corner**

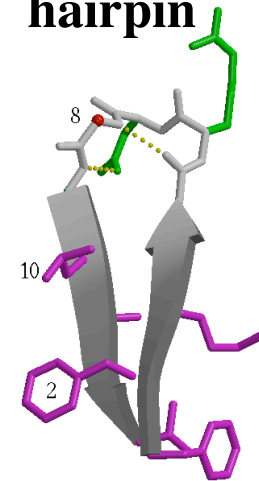


**glycine helix N-cap**

# Finding I-sites



Type-I hairpin



ELLEHSLYTQPENHSSEIAIWKEDLEYGIPVVFVDAGALQNELLEHSLYTQPENHSCELLEHSLYTQ

# Are I-sites really folding initiation sites?

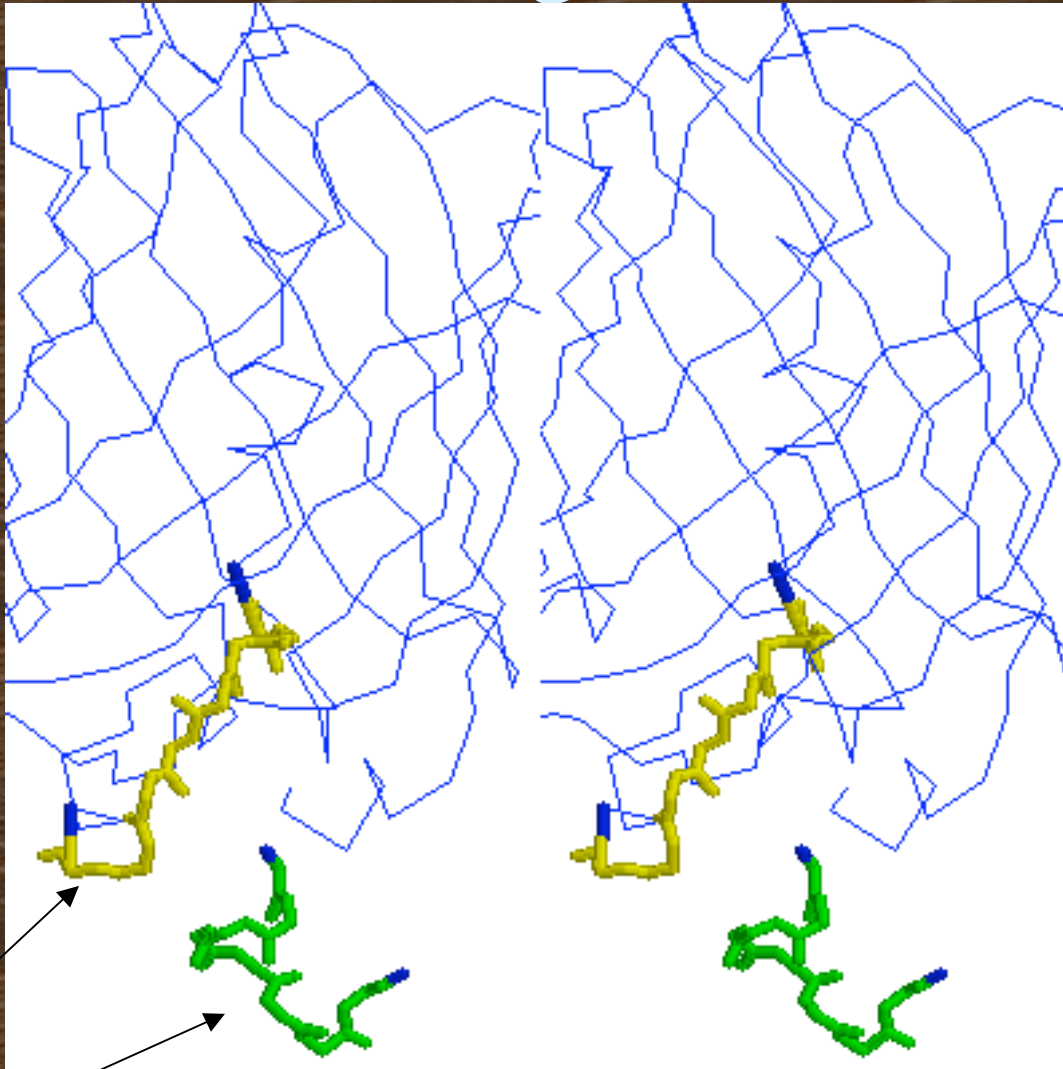
Prediction experiments (Bystroff & Baker, Proteins, 1997)

NMR data on peptides (Yi *et al*, J.Mol.Biol., 1998)

Molecular dynamics simulations (Bystroff & Garde, Proteins, 2002)



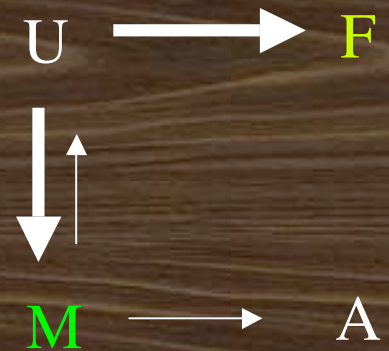
# Misfolding initiation sites?



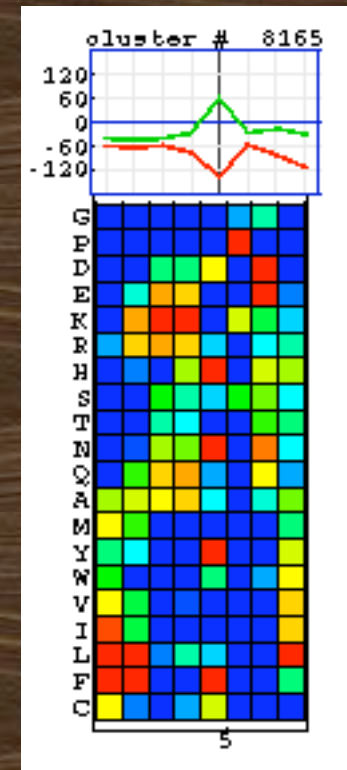
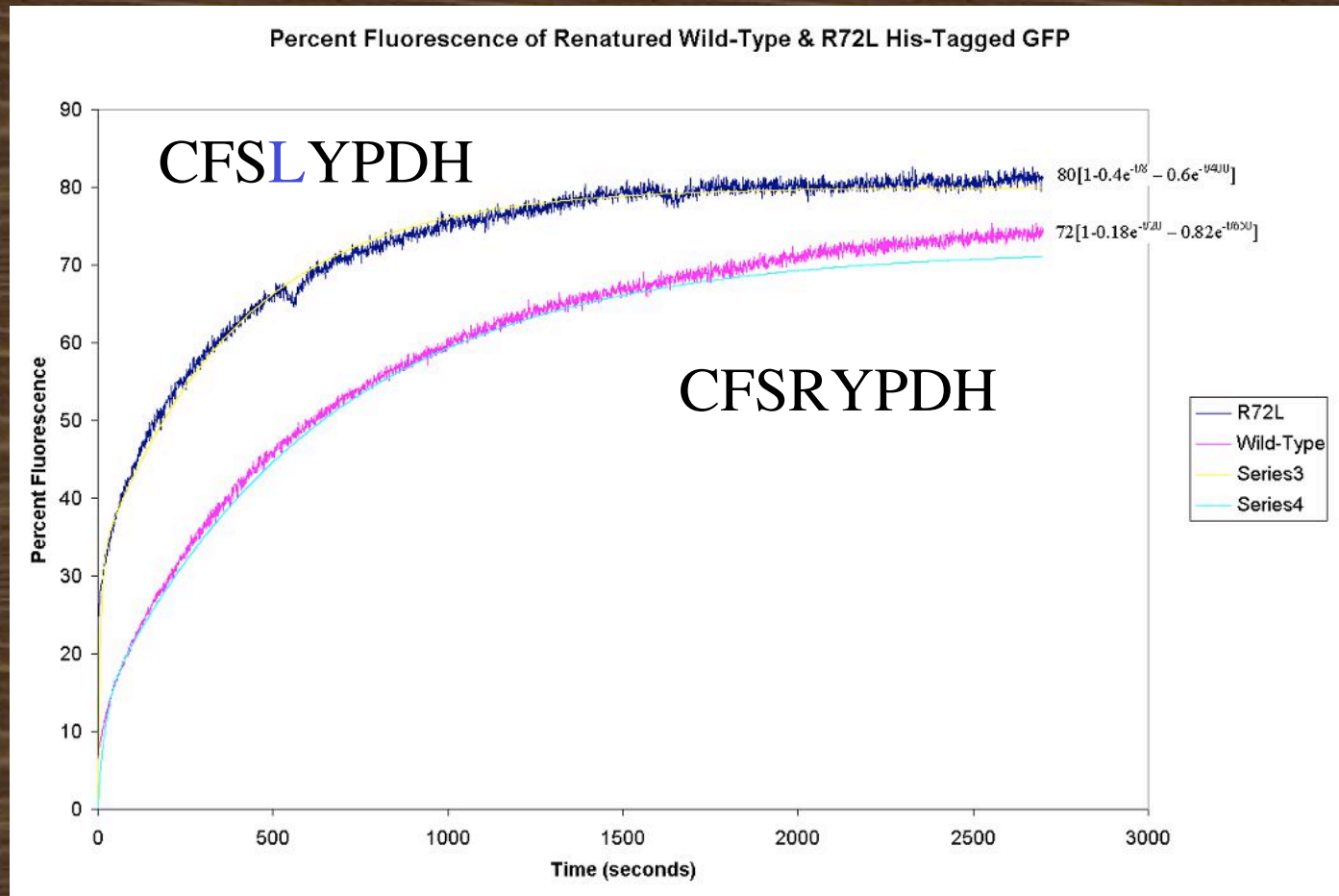
correctly  
folded

misfolded piece

U=unfolded  
M=misfolded  
F=folded  
A=aggregated

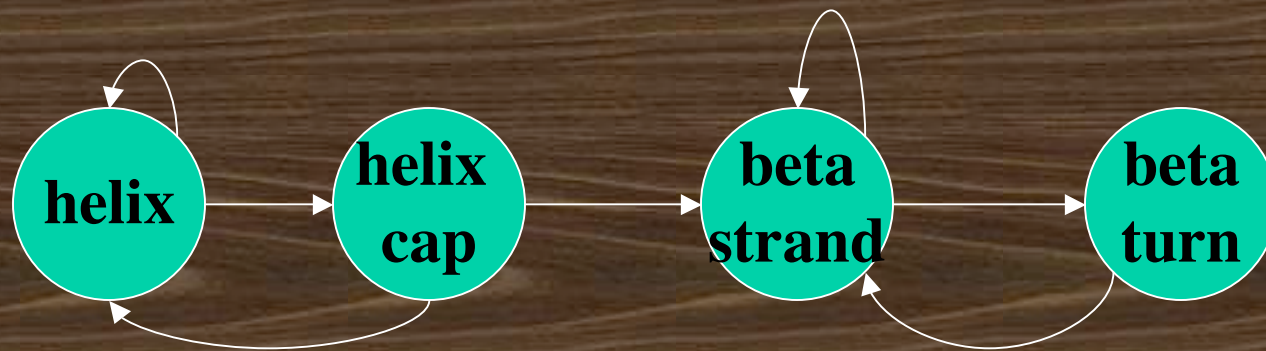


# Destabilizing the misfolding initiation site



## Level 2. Motif grammar

Arrangement of I-sites motifs in proteins is highly non-random

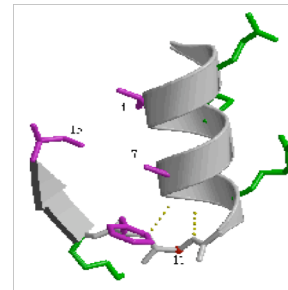
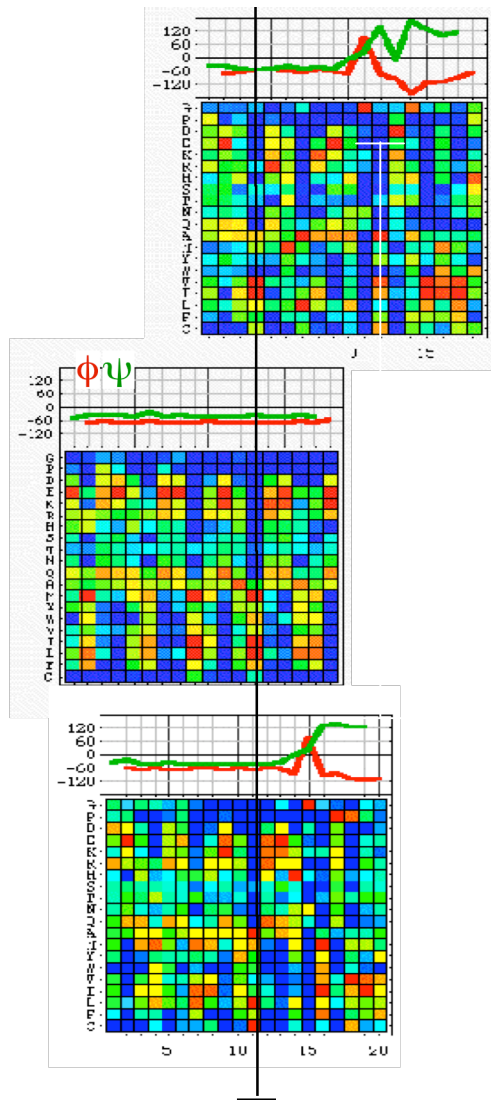


Adjacencies can be modeled as a Markov chain

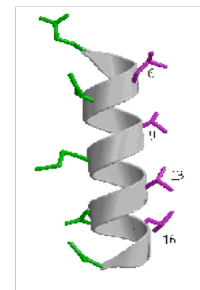


# Aligned motifs become a Markov chain

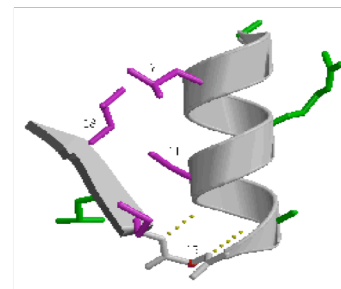
aligned profiles



Type-1  
G  $\alpha$  C-cap



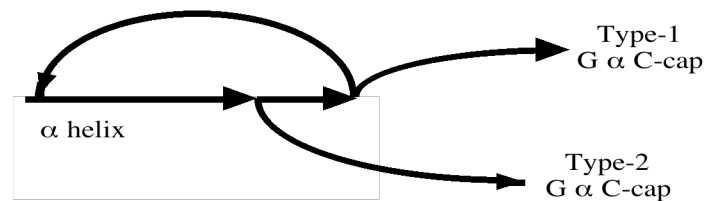
$\alpha$  helix

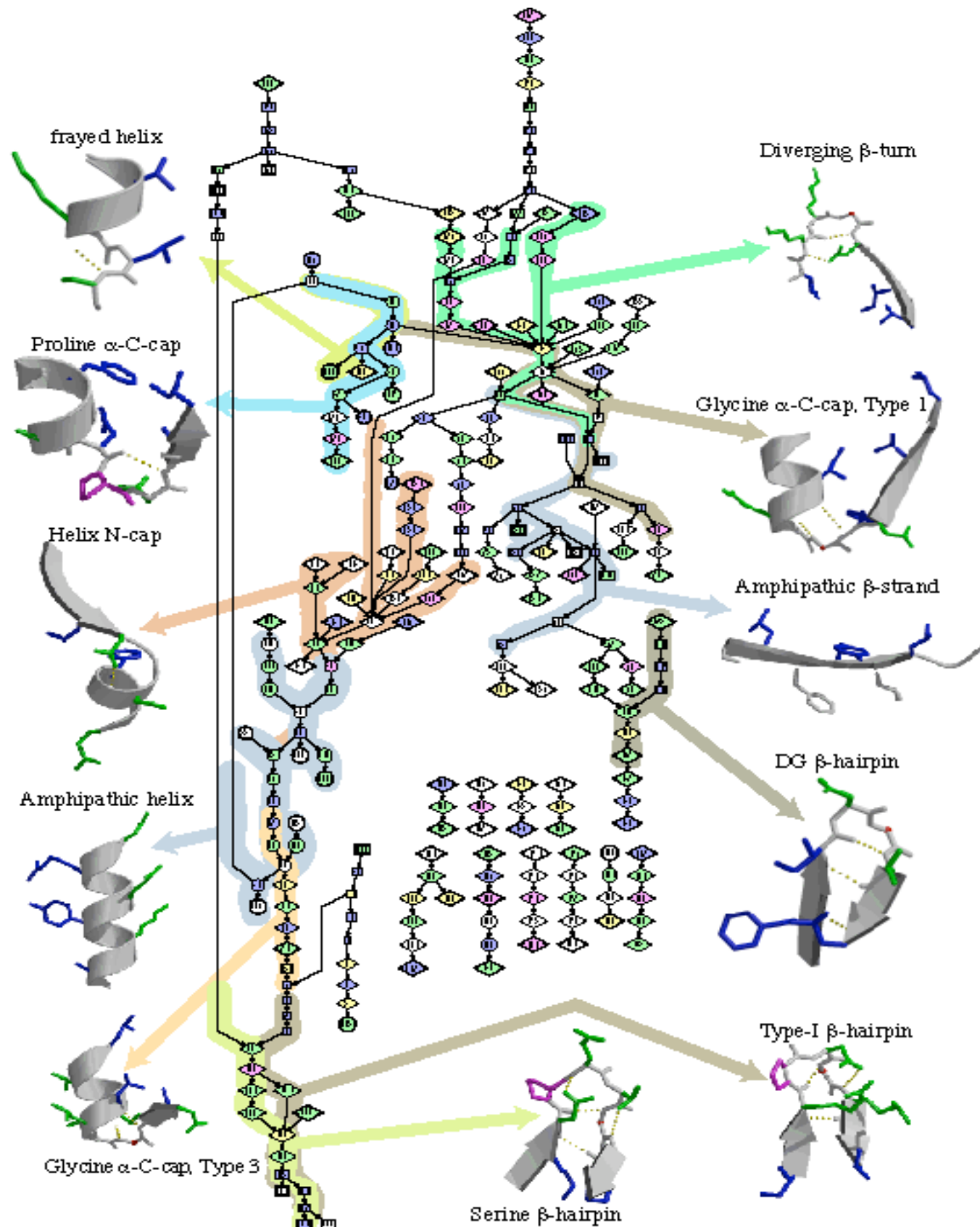


Type-2  
G  $\alpha$  C-cap

aligned structures

state topology:





# HMMSTR

Hidden  
Markov  
Model for  
local protein  
STRucture

282 nodes

317 transitions

Unified model  
for 31 distinct  
sequence-  
structure motifs

(Bystroff & Baker, J.  
Mol. Biol., 2000)

# How an HMM works

We have  $S$  (the sequence).  
We want  $Q$  (the state sequence),  
 $P(Q|S)$  is the probability of  $Q$  given  $S$

$$P(Q | S) = \pi_{q_1}(s_1) \prod_{t=2, \dots, N} a_{q_{t-1}q_t} B_{q_t}(s_t)$$

starting states

arrows

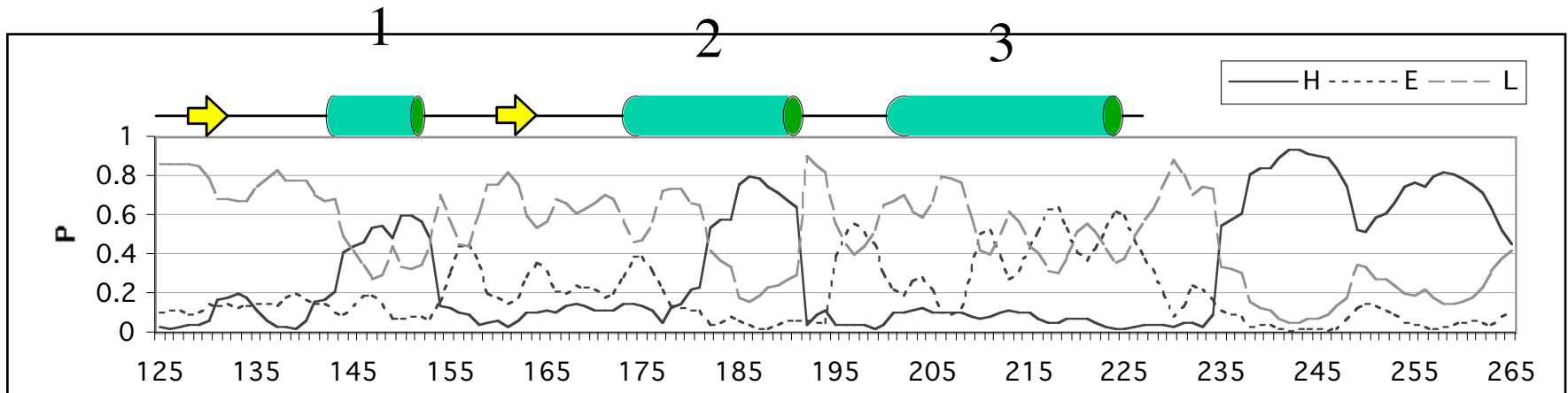
amino acid profiles

$$B_i(s_t) = \begin{pmatrix} d_i(D_t) \\ r_i(R_t) \\ c_i(C_t) \end{pmatrix} b_{q_i}(O_t)$$

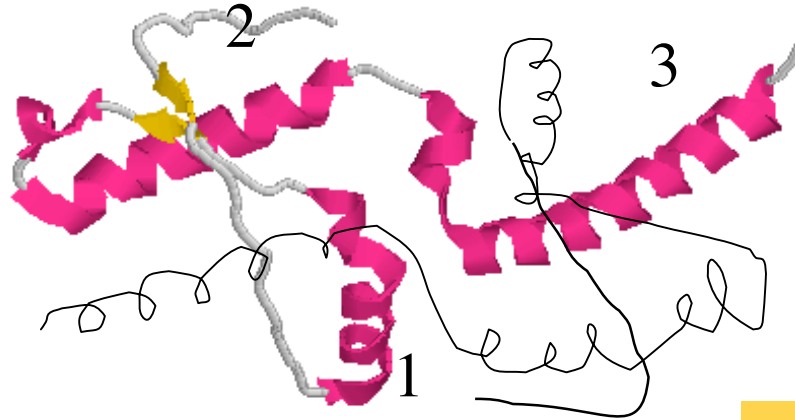


# HMMSTR can predict which parts of a structure might misfold.

HMMSTR secondary structure prediction



Human prion protein fragment.  
(X-ray structure solved in 2002)

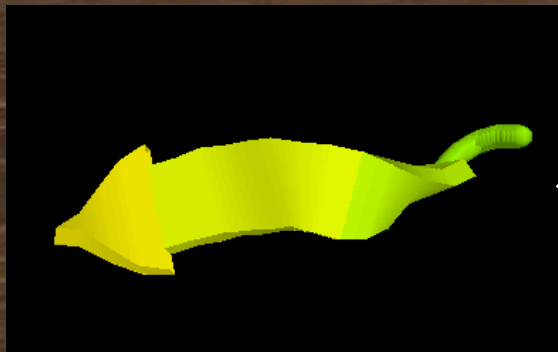
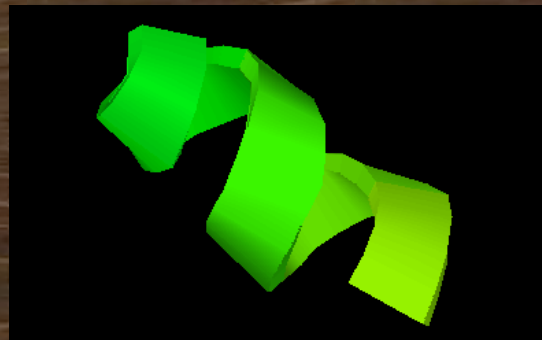
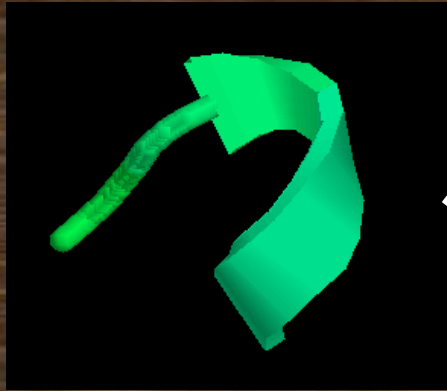


Helix 3 is known to be the location of familial prion disease

Knaus et al, NSB 8:770-4, 2001

Level 1: I-sites

Level 2: HMMSTR



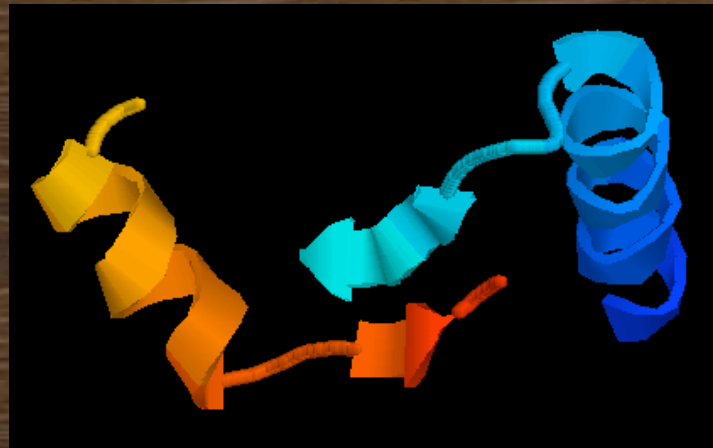
initiation



propagation

## Level 3: Pairwise Motif-Motif Contact Potentials

- $G(p, q, s)$  represents the free energy of a motif-motif contact.

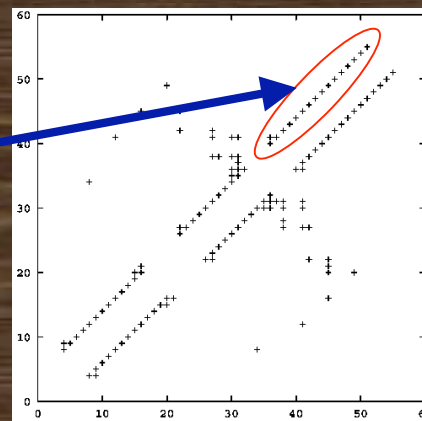
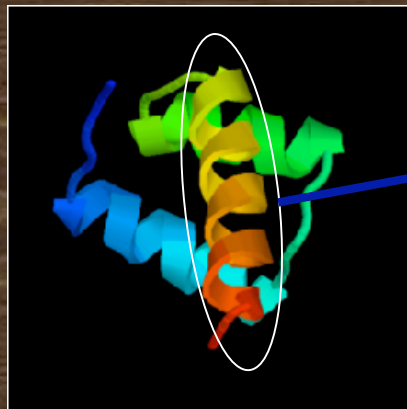
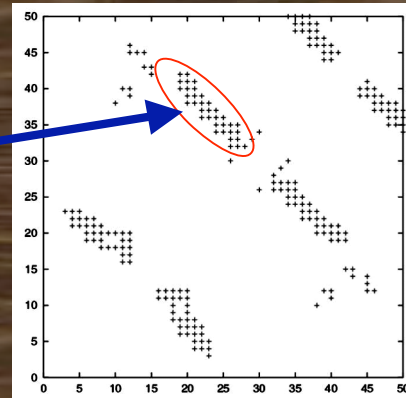
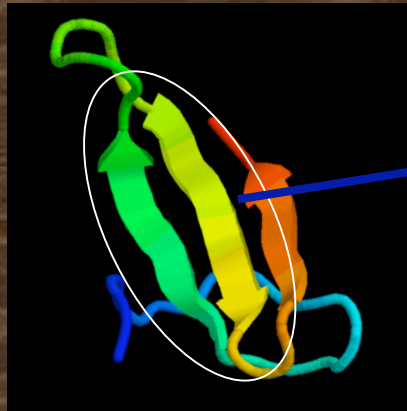


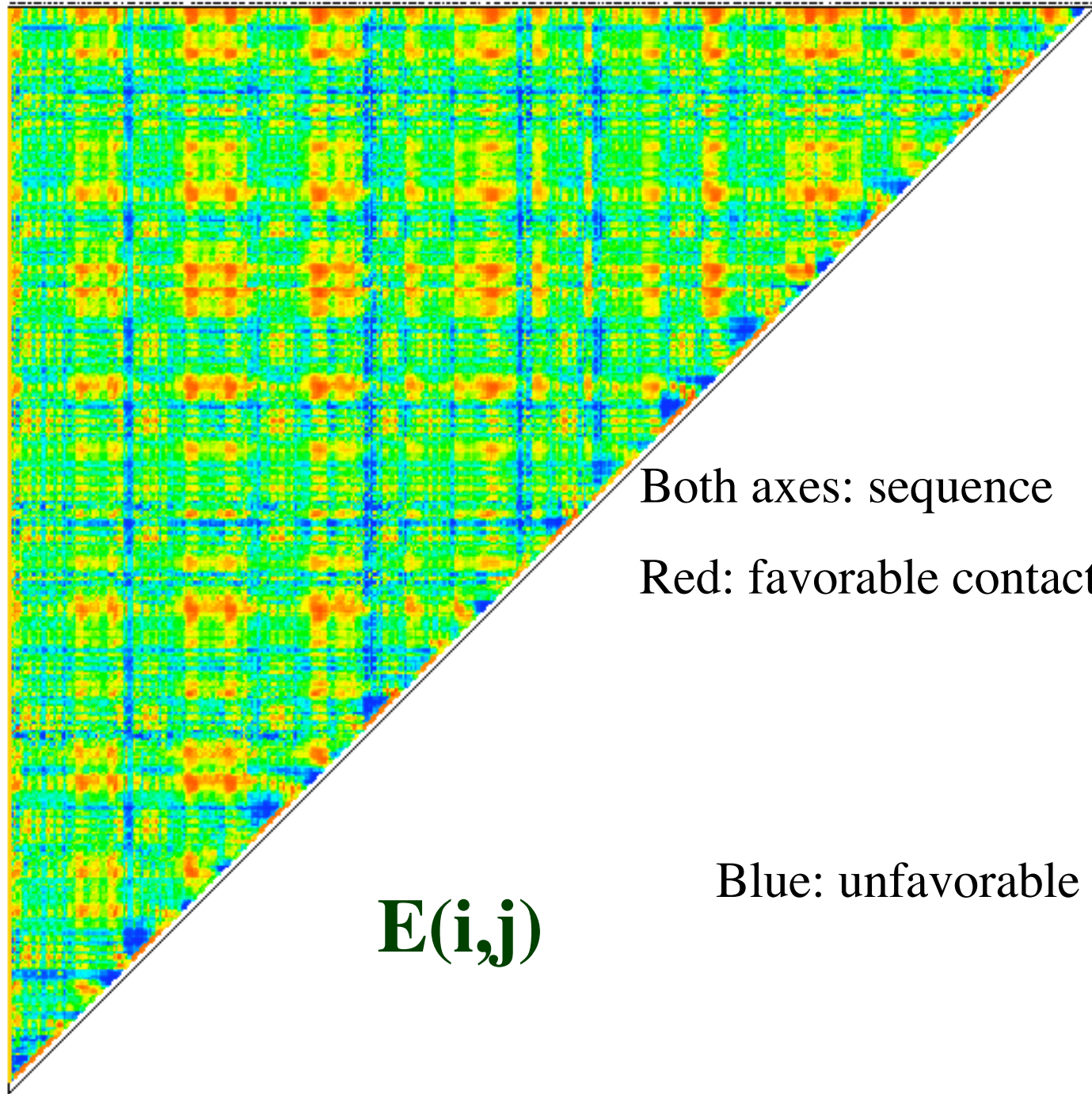
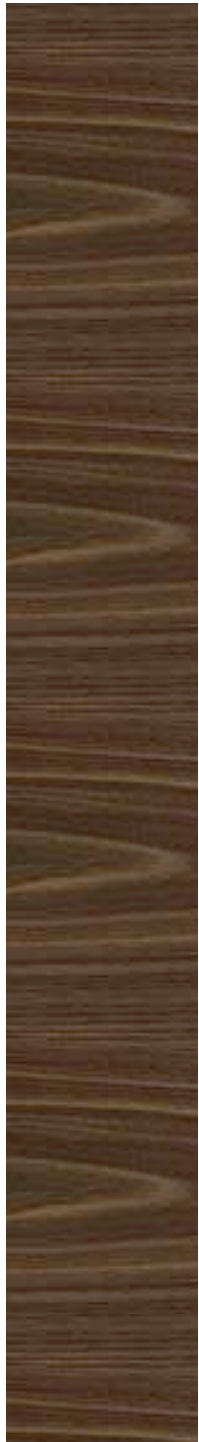
$$G(p, q, s) = -\log \frac{\sum_{PDBselect} \sum_{i \ni D_{i,i+s} < 8\text{\AA}} \Gamma(i, p) \Gamma(i + s, q)}{\sum_{PDBselect} \sum_i \Gamma(i, p) \Gamma(i + s, q)}$$



# What is a contact map?

**Definition:**  $S(I, J) = \begin{cases} 1 & \text{if } d(i, j) \leq D \\ 0 & \text{if } d(i, j) > D \end{cases}$





Both axes: sequence

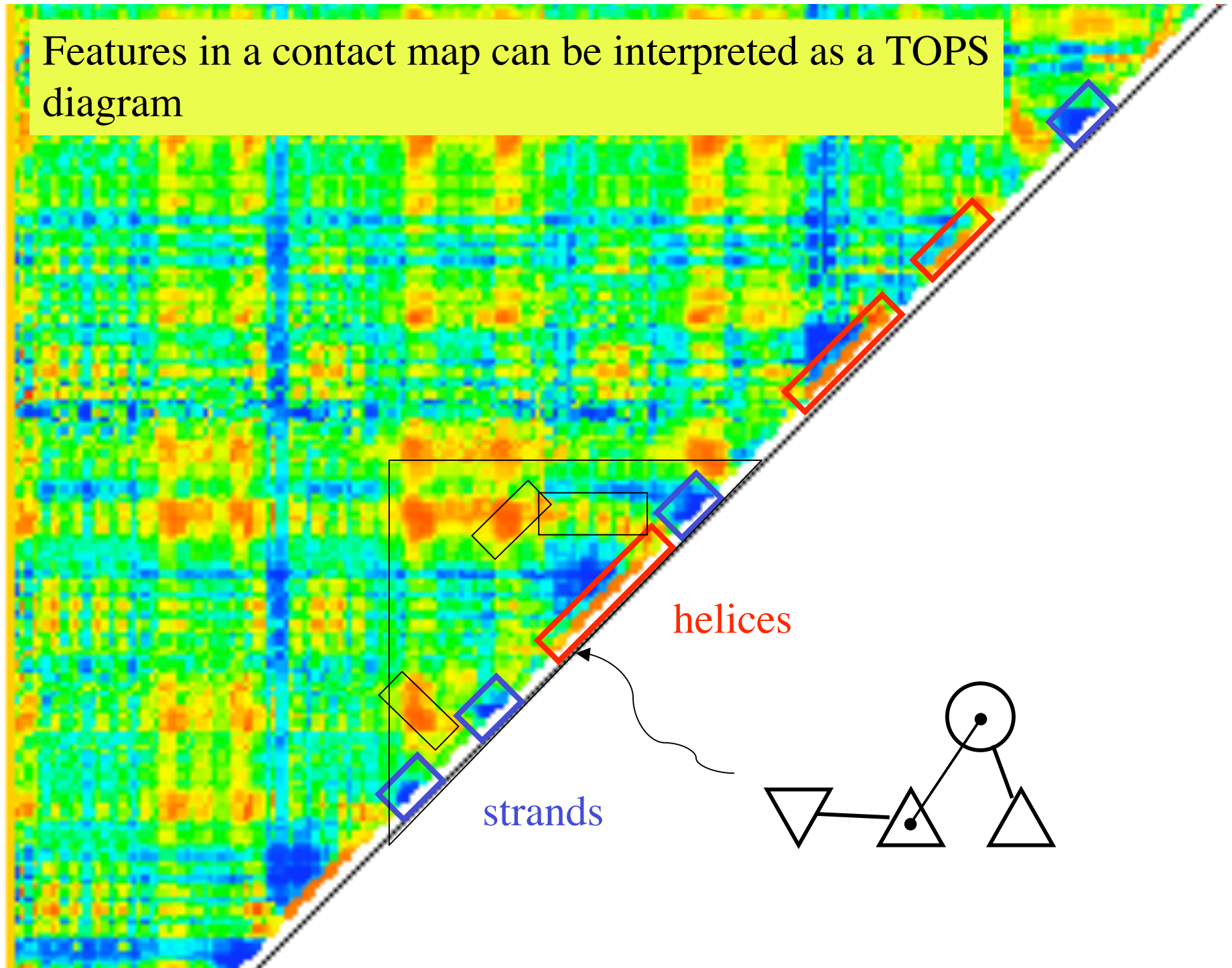
Red: favorable contact

Blue: unfavorable

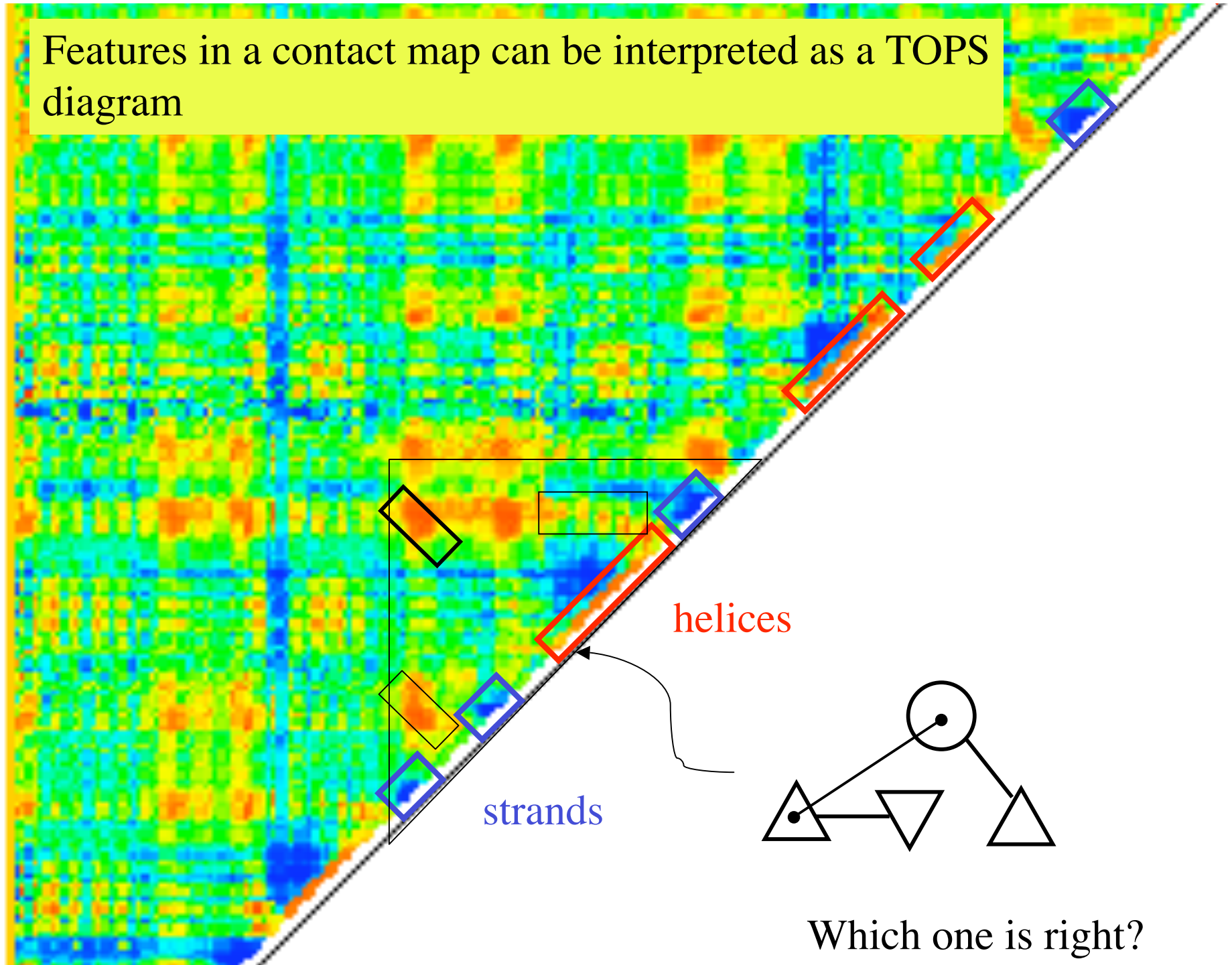




Features in a contact map can be interpreted as a TOPS diagram



Features in a contact map can be interpreted as a TOPS diagram

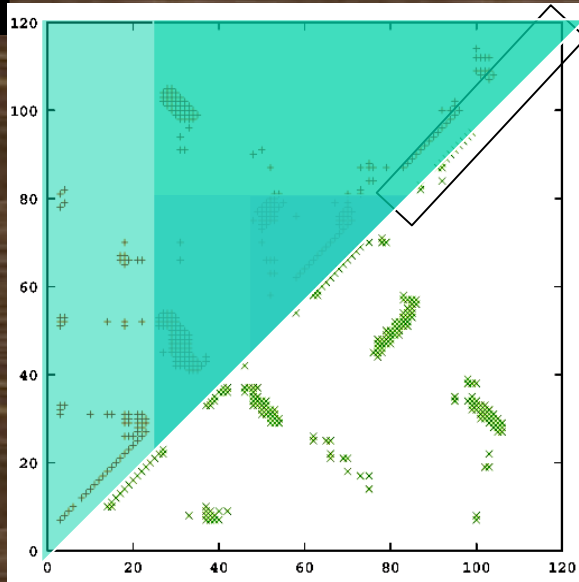


Which one is right?



CASP5

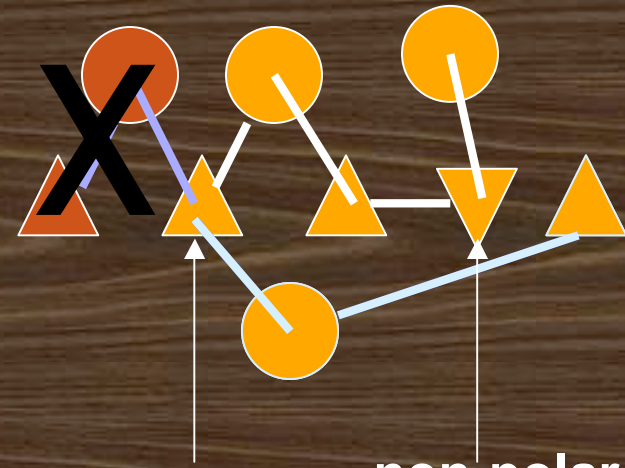
*ab initio* Prediction



True Contact Map  
T0130

True contact map

A rule-based simulation procedure.

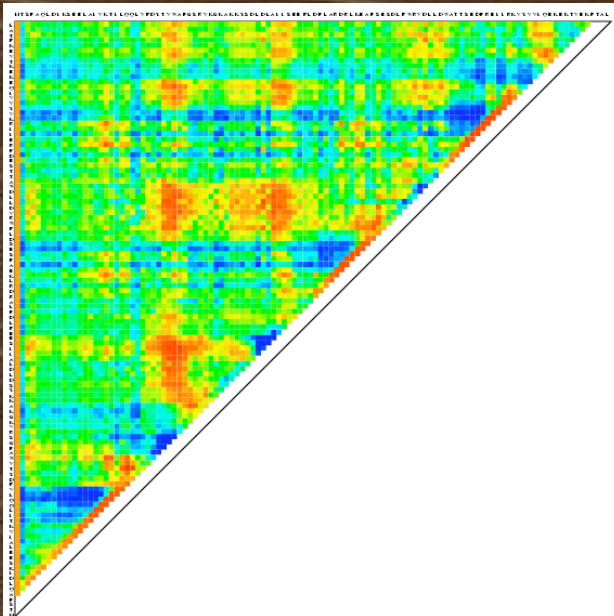


amphipathic non-polar



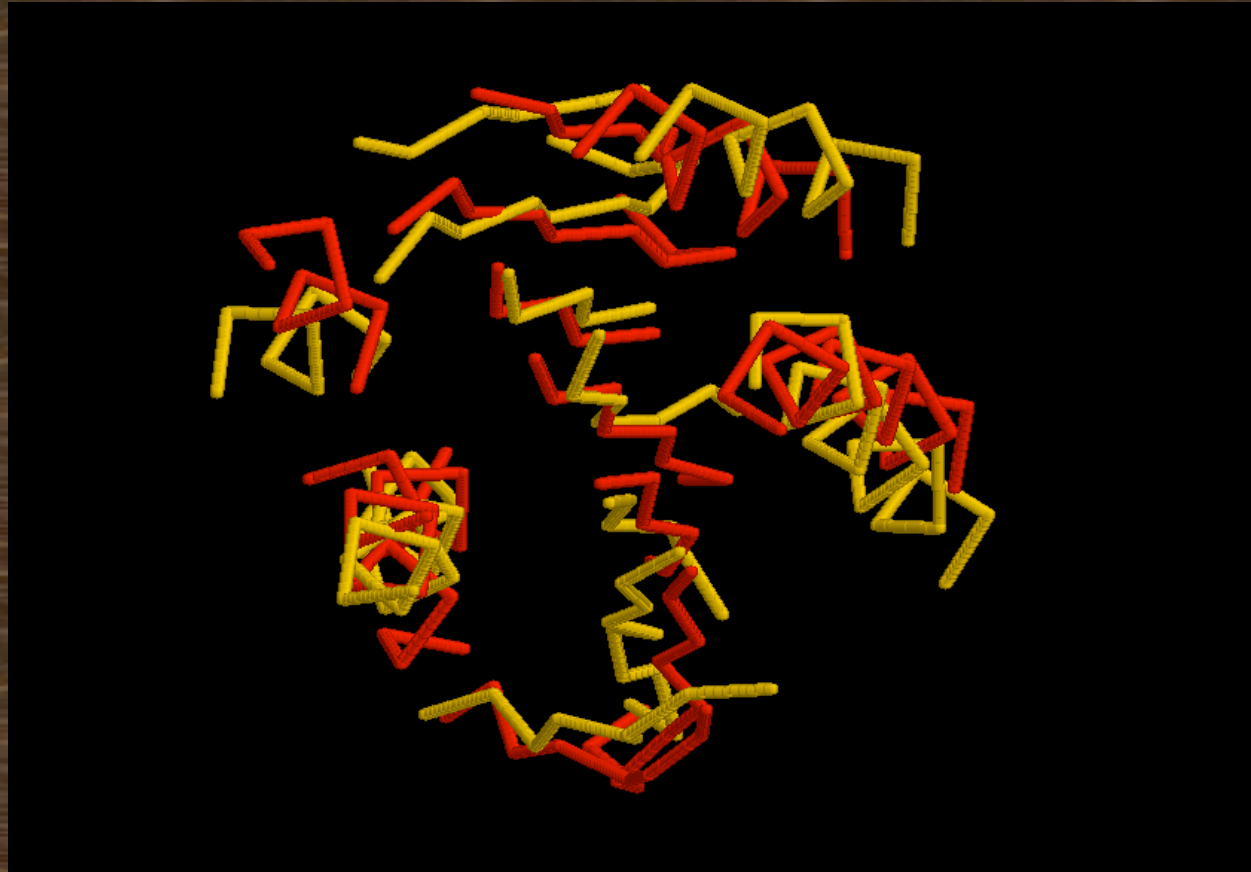
T0130

Contact energies



## Level 4: Multibody arrangements of local motifs

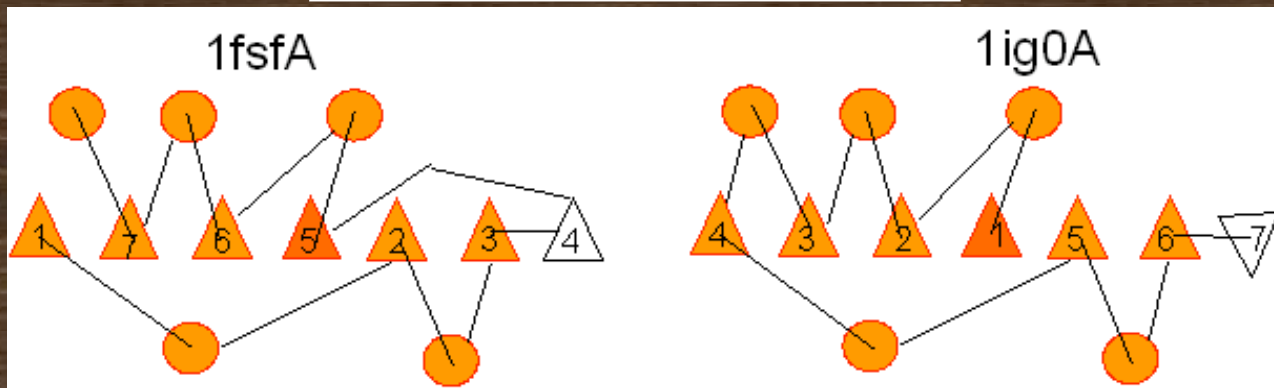
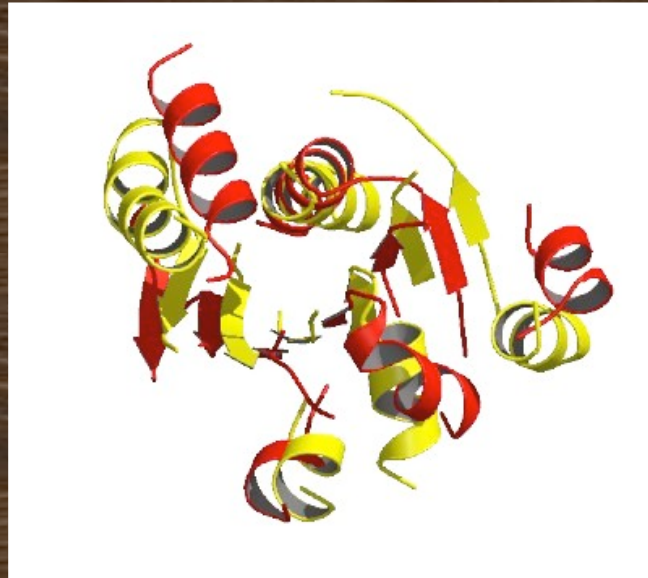
It is difficult to see similarities between these two proteins, but...







# SCALI : Structural Core ALignment

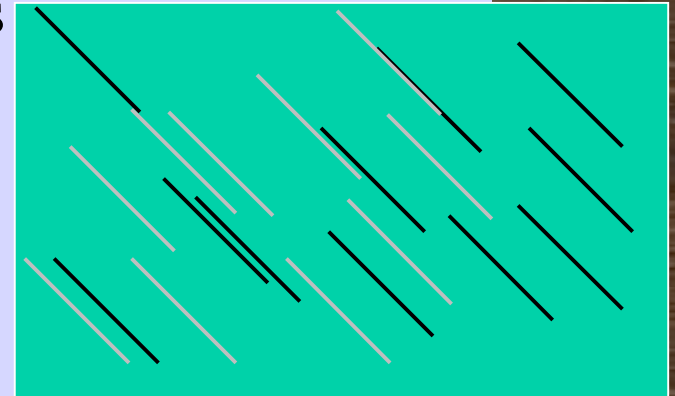




# How SCALI works

(1) Gapless alignment of HMMSTR states

(2) Initialize tree search w/ one gapless fragment.

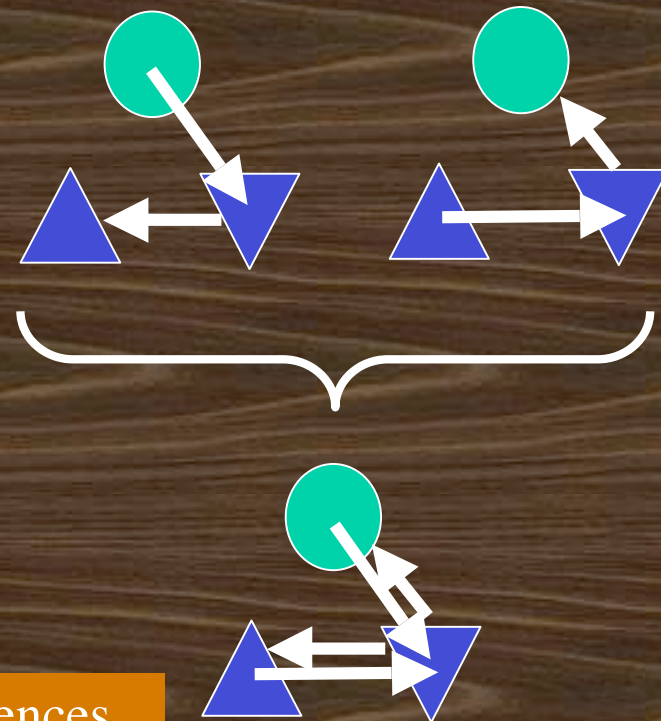
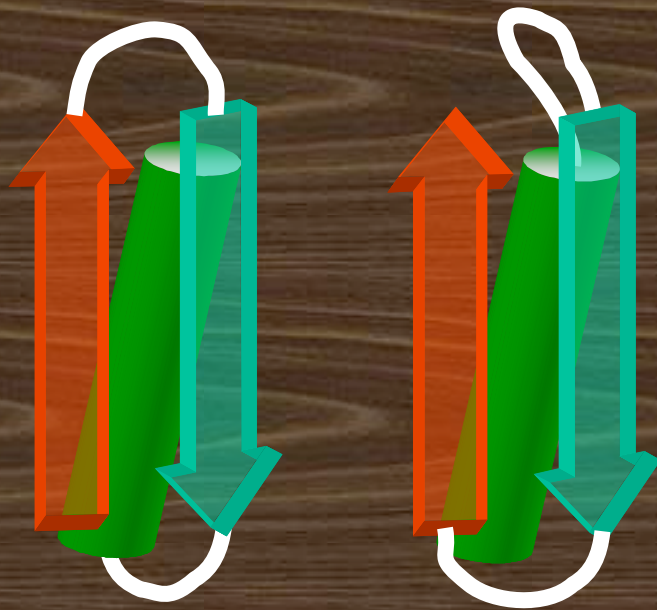


(3) Add a new fragment *iff* it is compatible and has a high score .

(4) Tree leaves when no fragments can be added.

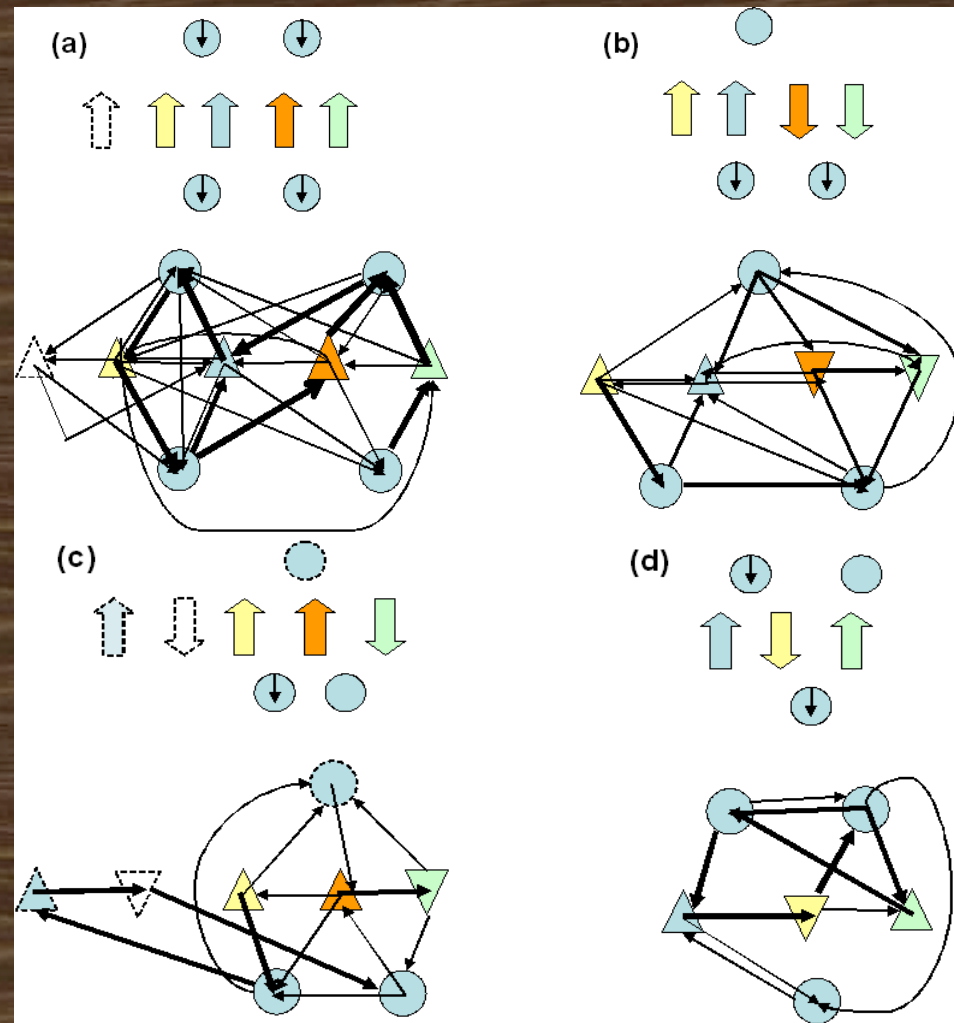
Score of leaves = aligned contacts + permutation penalty.

# HMMs may be built based on non-sequential alignments



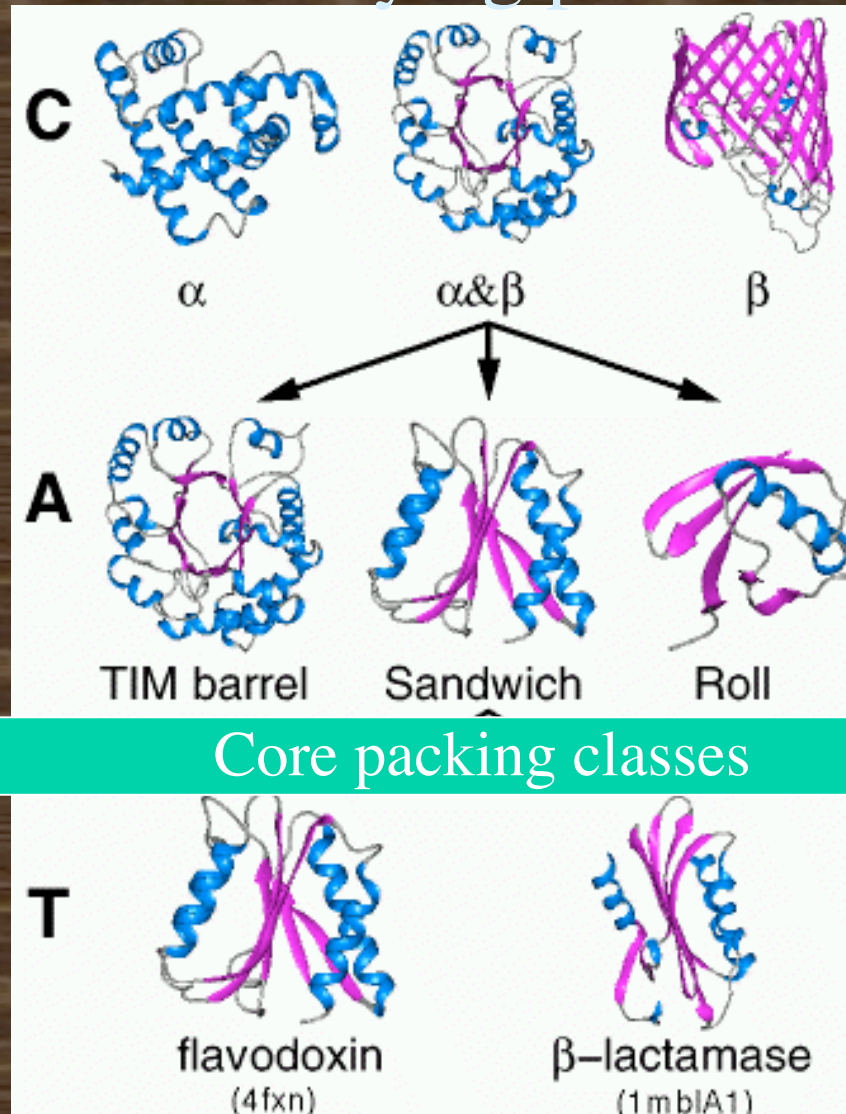
Markov states represent amino acid sequences and positions in space. Connections between them represent loops.

# Hidden Markov models for a/b/a proteins





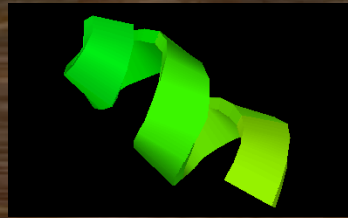
# Non-sequential clusters may be a useful for classifying proteins



Multiple non-sequential alignments are more specific than “architecture” but not as specific as “topology”.



Level 1: I-sites



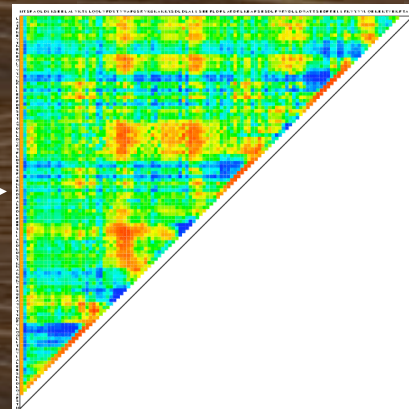
initiation

Level 2: HMMSTR



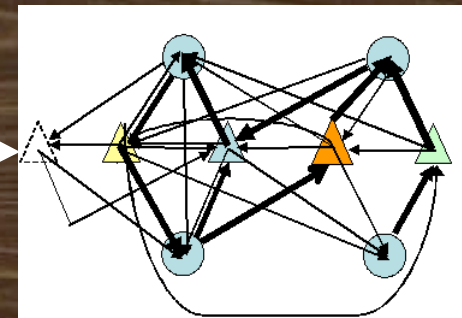
propagation

Level 3: HMMSTR-CM



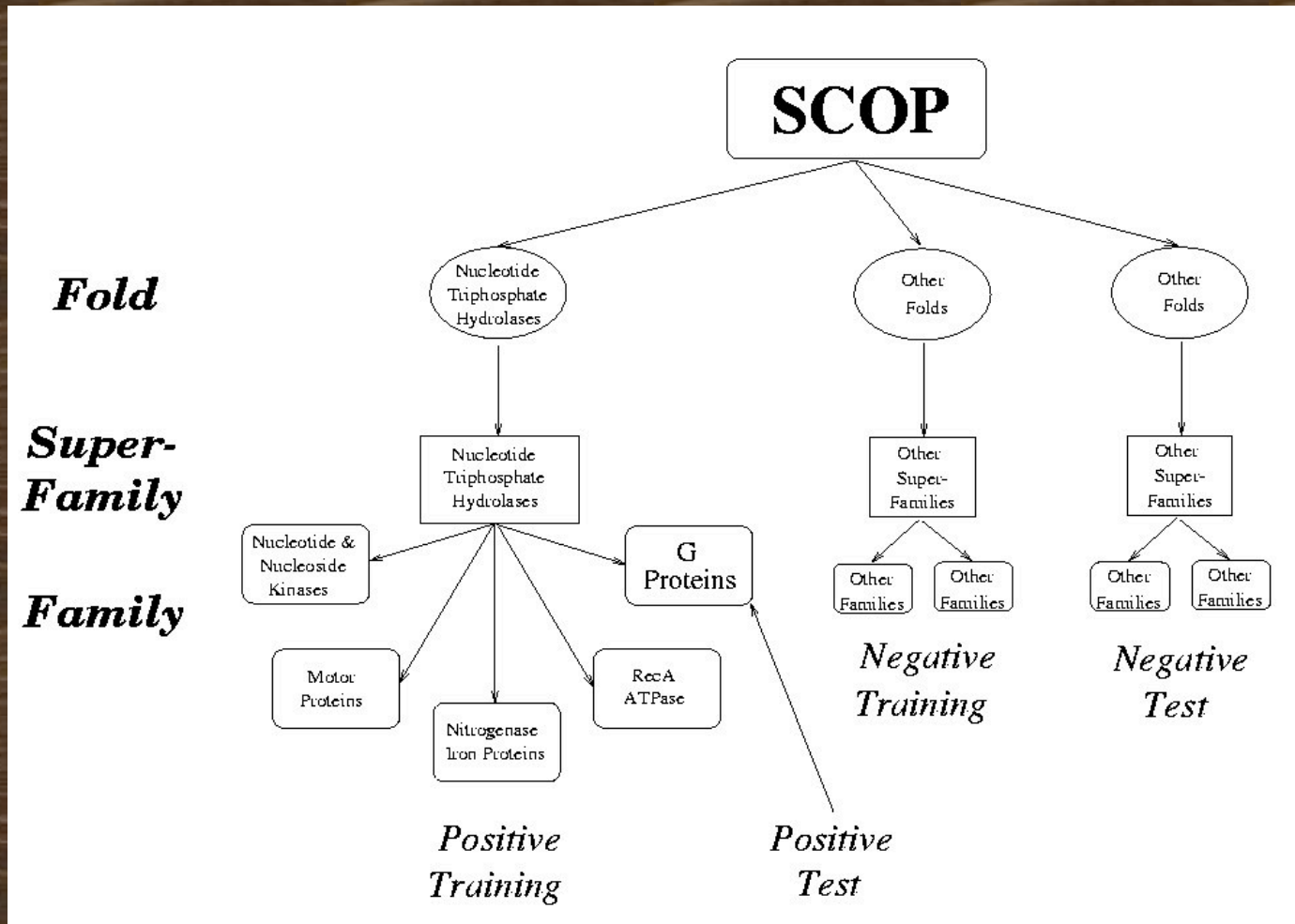
condensation

Level 4: SCALI



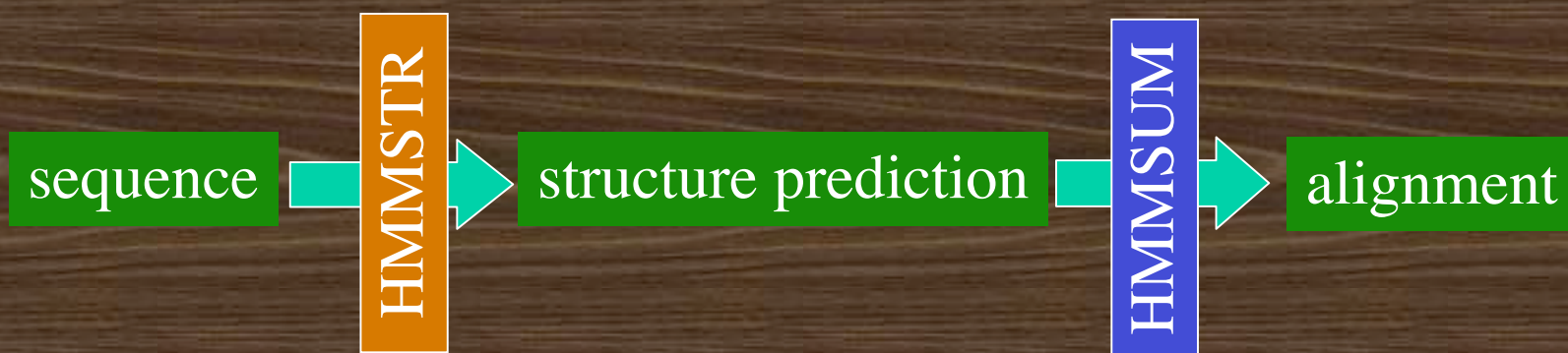
molten  
globule

# Level 5: Global topology



Separation of the SCOP 1.53 database into training and test sequences, shown for the G proteins test family

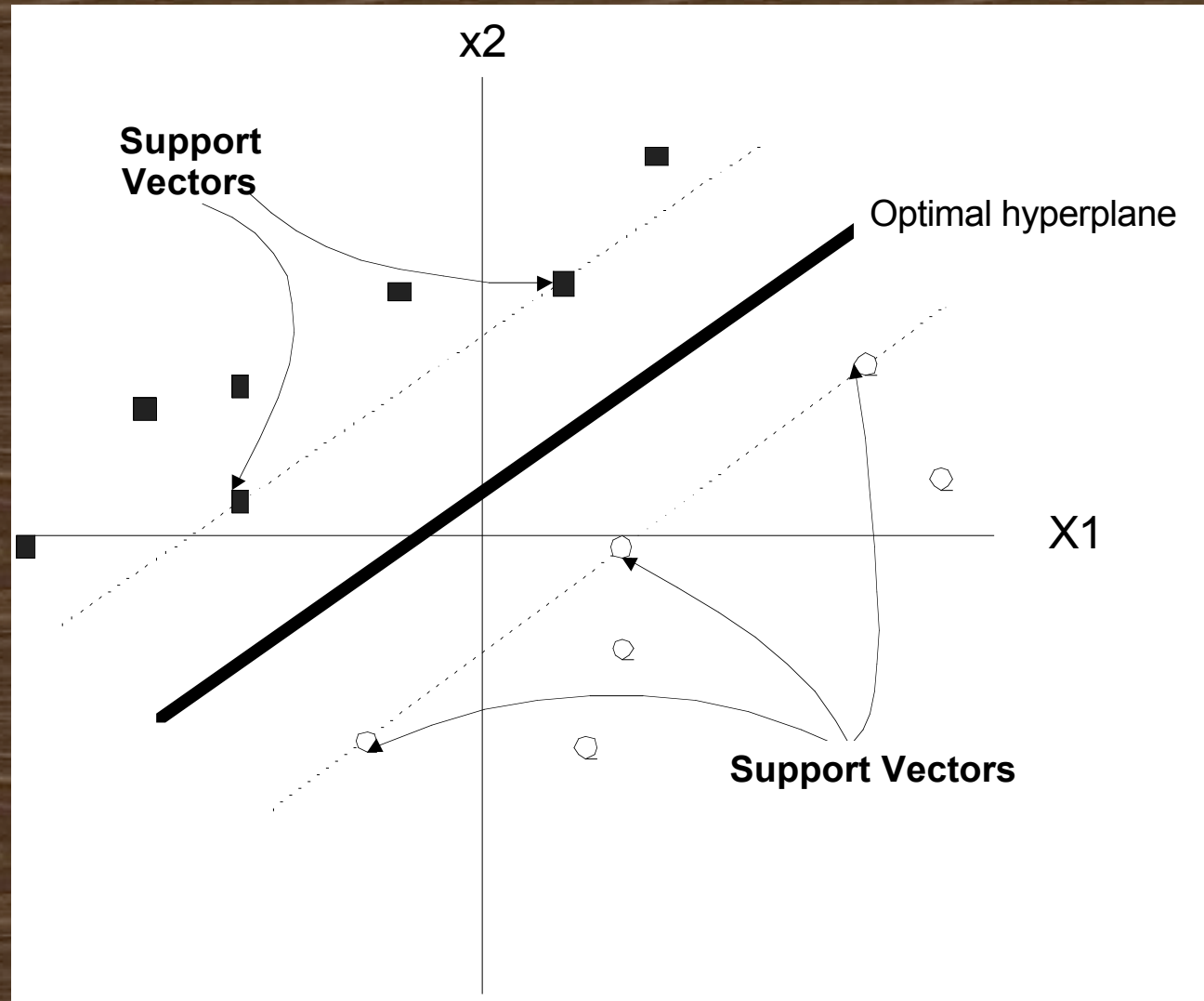
# Aligning Twilight Zone sequences using HMMSUM



Matrix	Gap penalty		Correct matches		Accuracy		Coverage (%)
	Opening	Extension	Counts	P value	%	P value	
BLOSUM50	8	2.3	12,211	-	41.8	-	35.9
HMMSUM-M	15	0.9	13,850	<0.001	42.4	0.316	40.8
HMMSUM-L	12	1.2	13,551	<0.001	43.8	0.110	39.9
HMMSUM-D	21	0.5	15,927	<0.001	46.0	<0.001	46.9

# Support Vector Machine

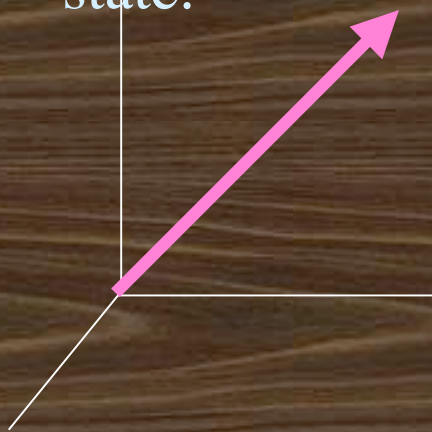
4052 proteins -->  
54-dimensional  
vector. Each  
dimension is the  
order of  
appearance  
HMMSTR states  
for one family.



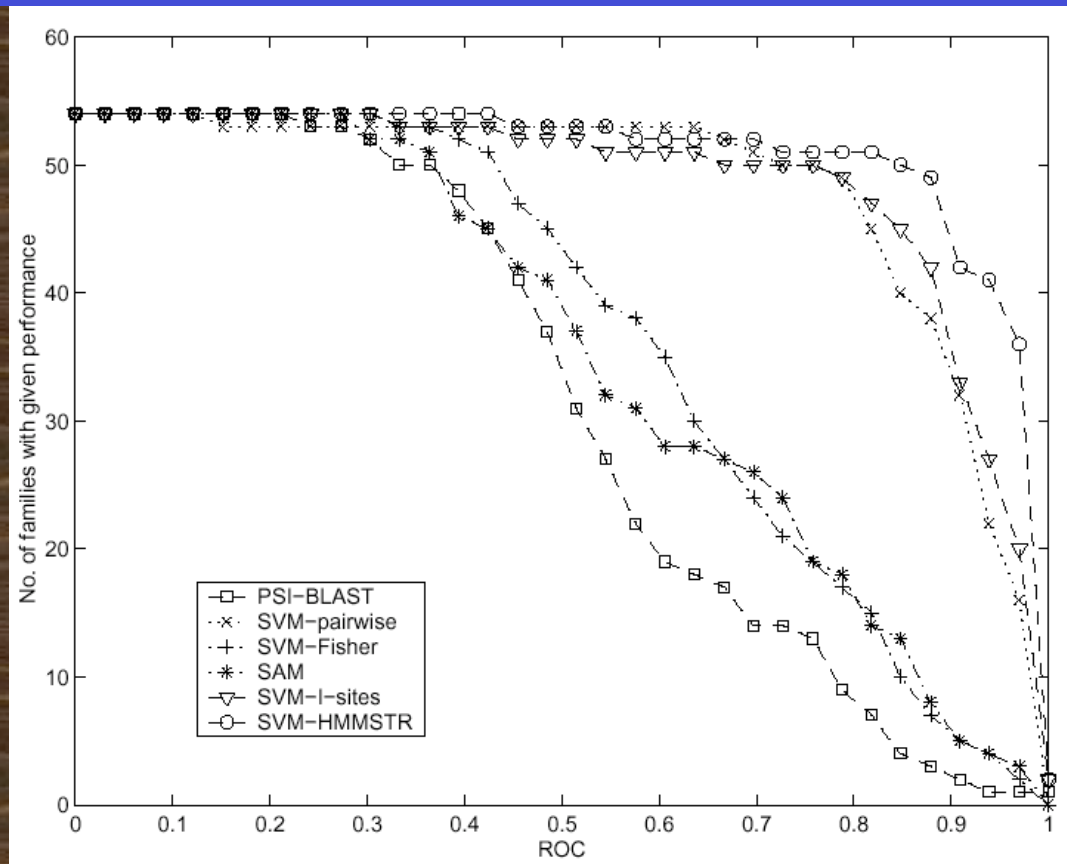


# HMMSTR as the basis for a Support Vector Machine

4052 proteins,  
represented as  
282-  
dimensional  
vector = Prob of  
each HMMSTR  
state.




SCOP benchmark of 54 sequence families



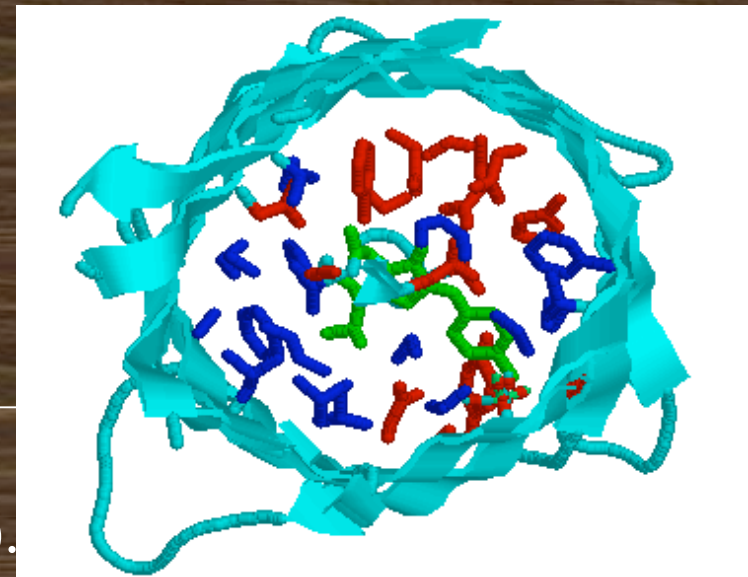
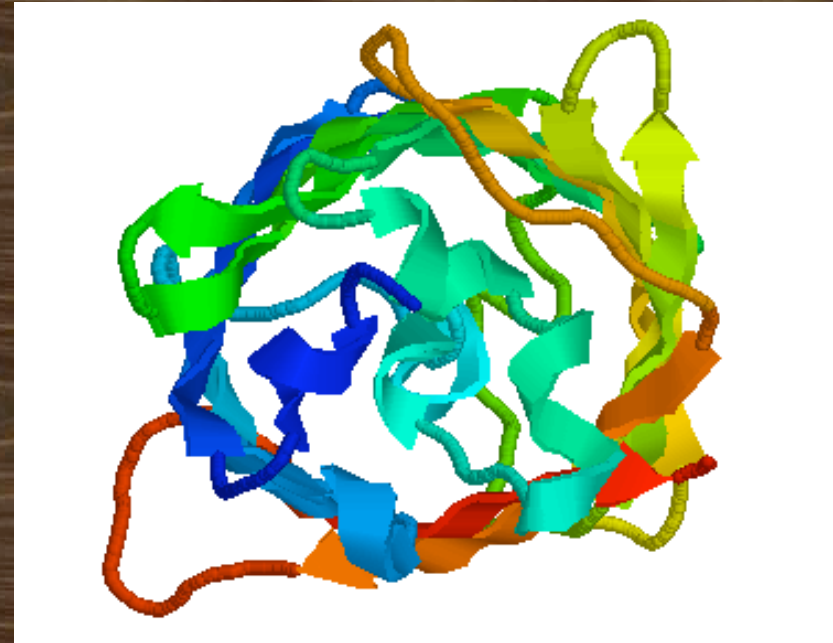
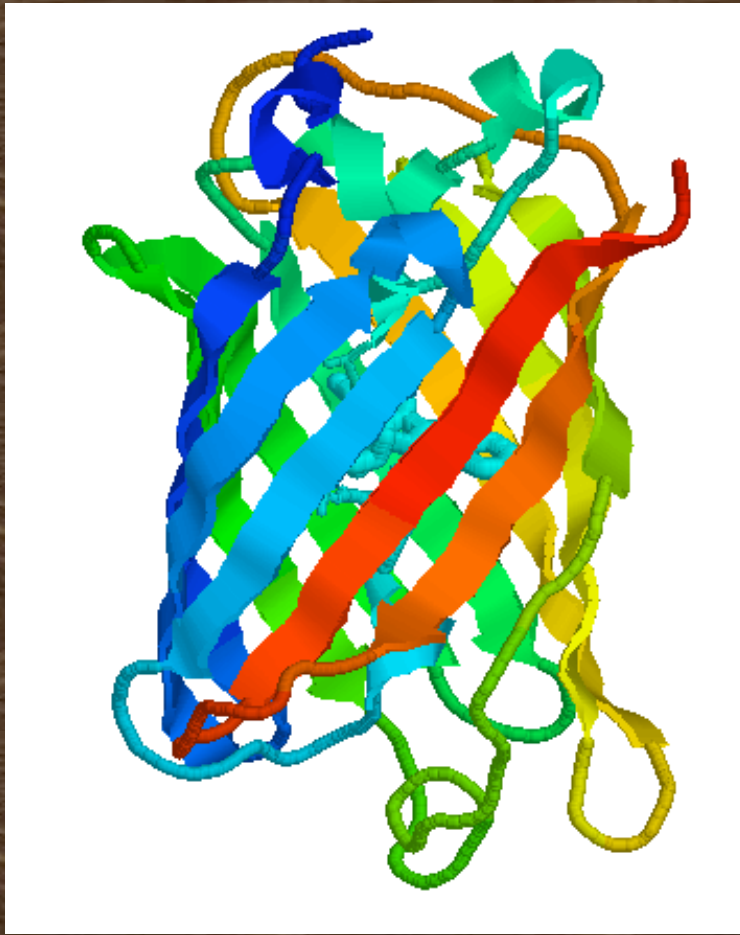
(Hou, Y *et al*, Bioinformatics, 2003; Proteins, 2004)

# No sparse data problem as we mine longer and longer patterns! Why?



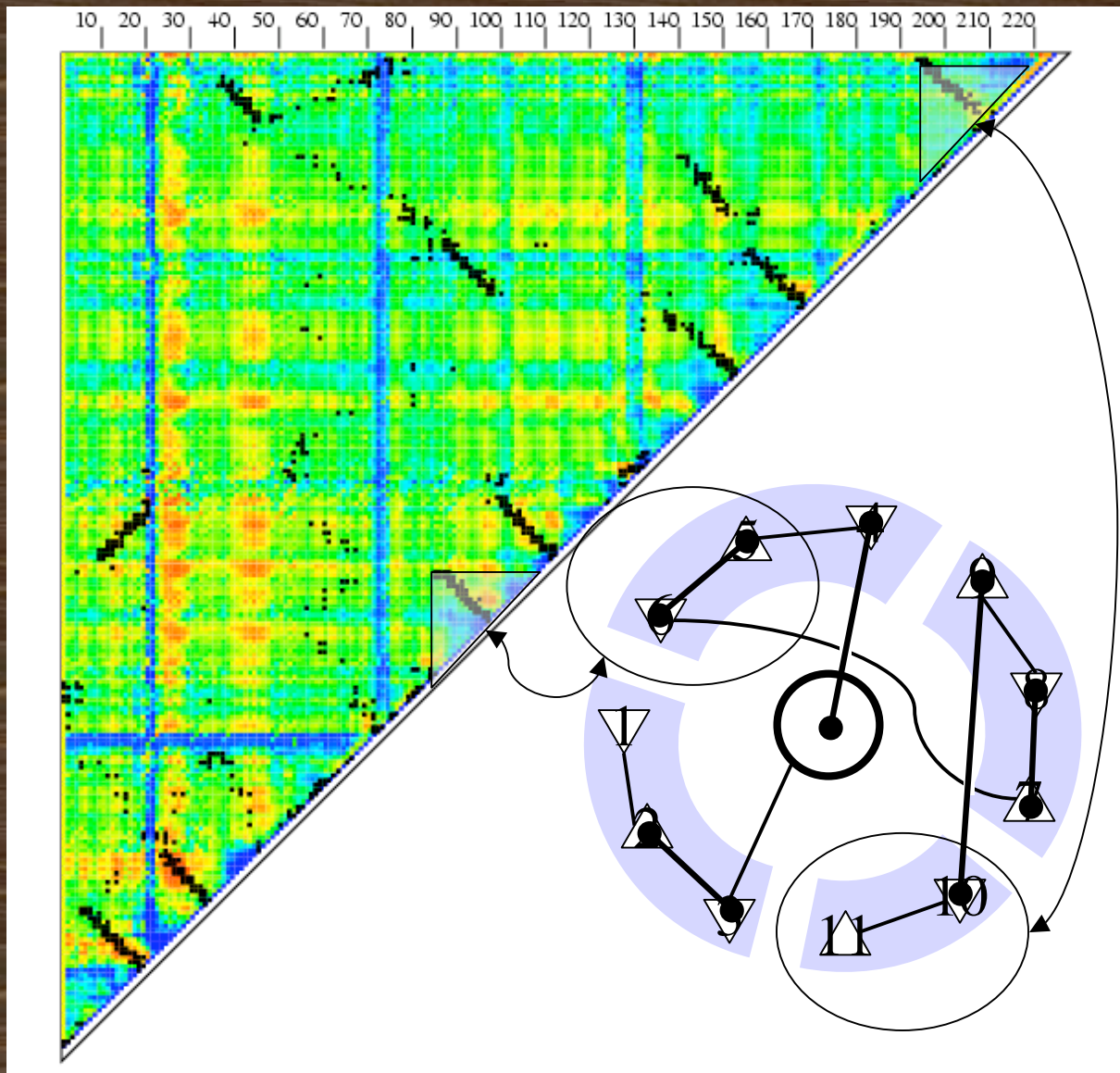
<u>Steps along the folding pathway:</u>	<u>Model</u>	<u>Complexity</u>
(1) Initiation	I-sites	~40 motifs
(2) propagation	HMMSTR	1.1 transitions/node
(3) condensation	HMMSTR-CM	~1% of pairs occur
(4) molten globule	SCALI	only self-avoiding paths
(5) native state folds	SVM-HMMSTR	~1000-2000

GFP is an 11-stranded anti-parallel beta barrel.

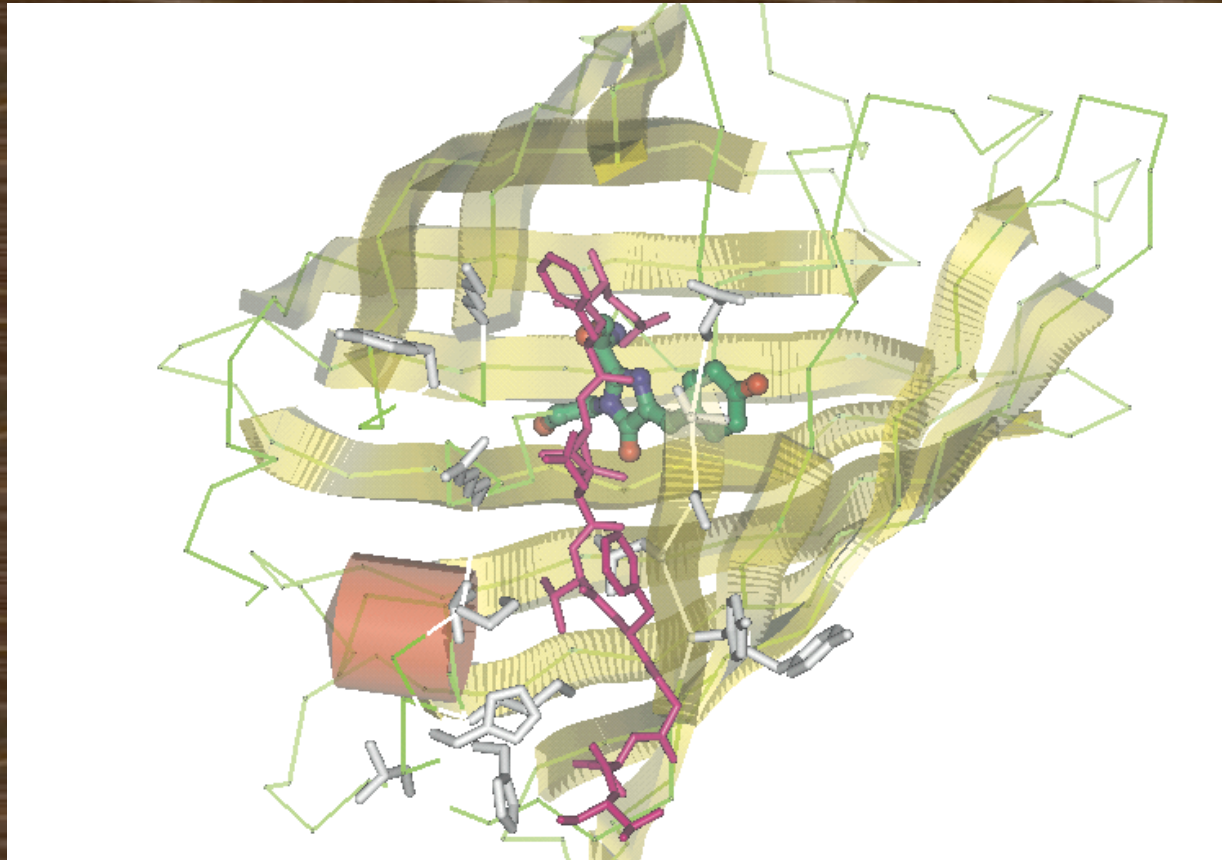


The core of the protein contains non-polar (blue) and polar sidechains (red).



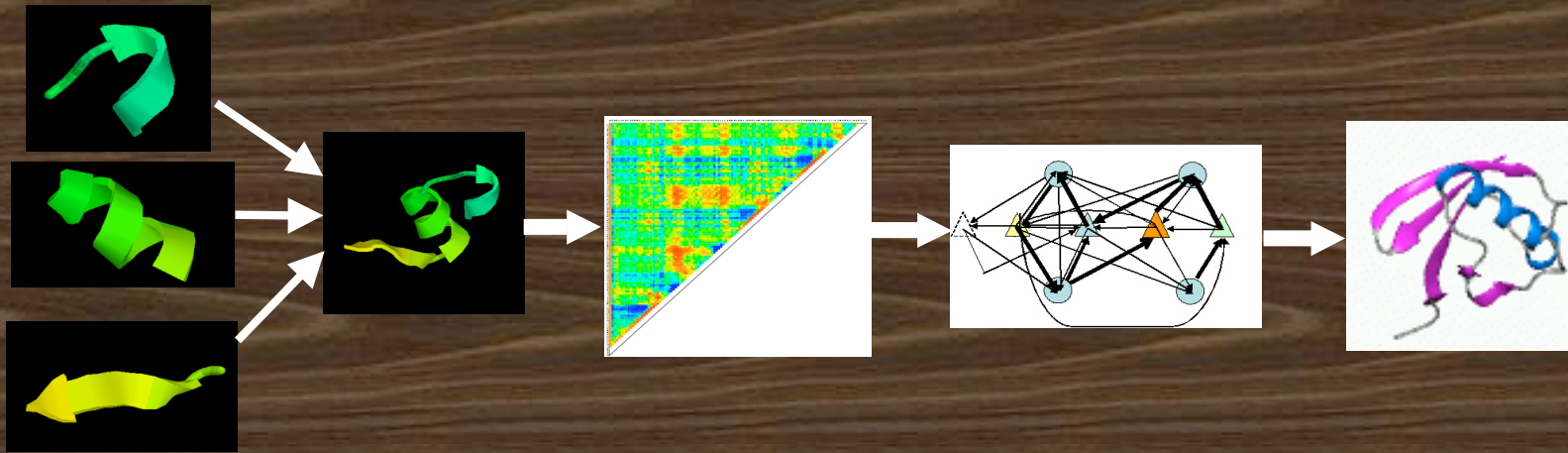


# A peptide biosensor based on GFP

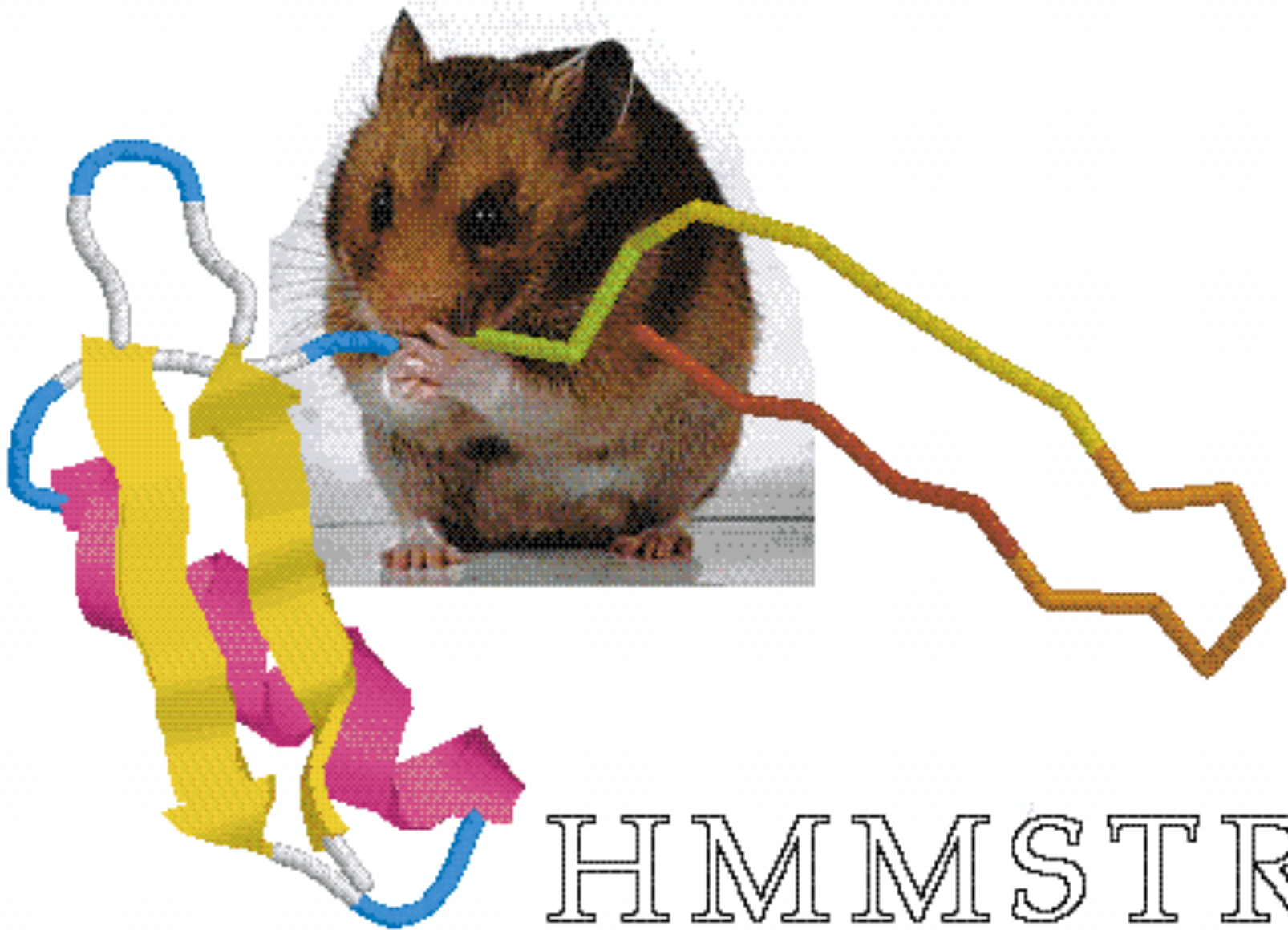


# Are there any conclusions?

We assumed that proteins fold in a certain, hierarchical manner, mined the data accordingly and found recurrence at every level, from short motifs to global structure.







HMMSTR

HMMSTR says: *Think Globally, Act Locally.*