

Five hierarchical levels of sequence-structure correlations in proteins

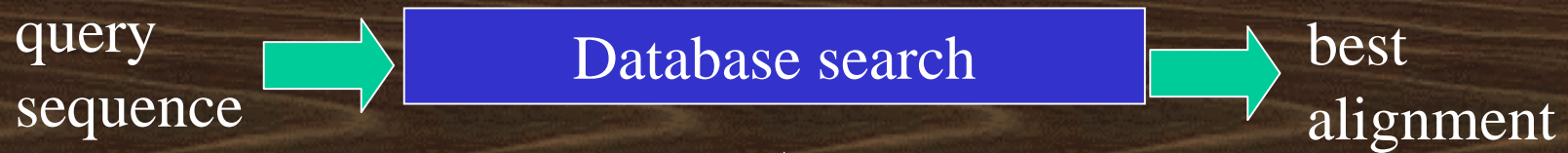
Chris Bystroff
Rensselaer Polytechnic Institute
Troy, New York, USA

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...

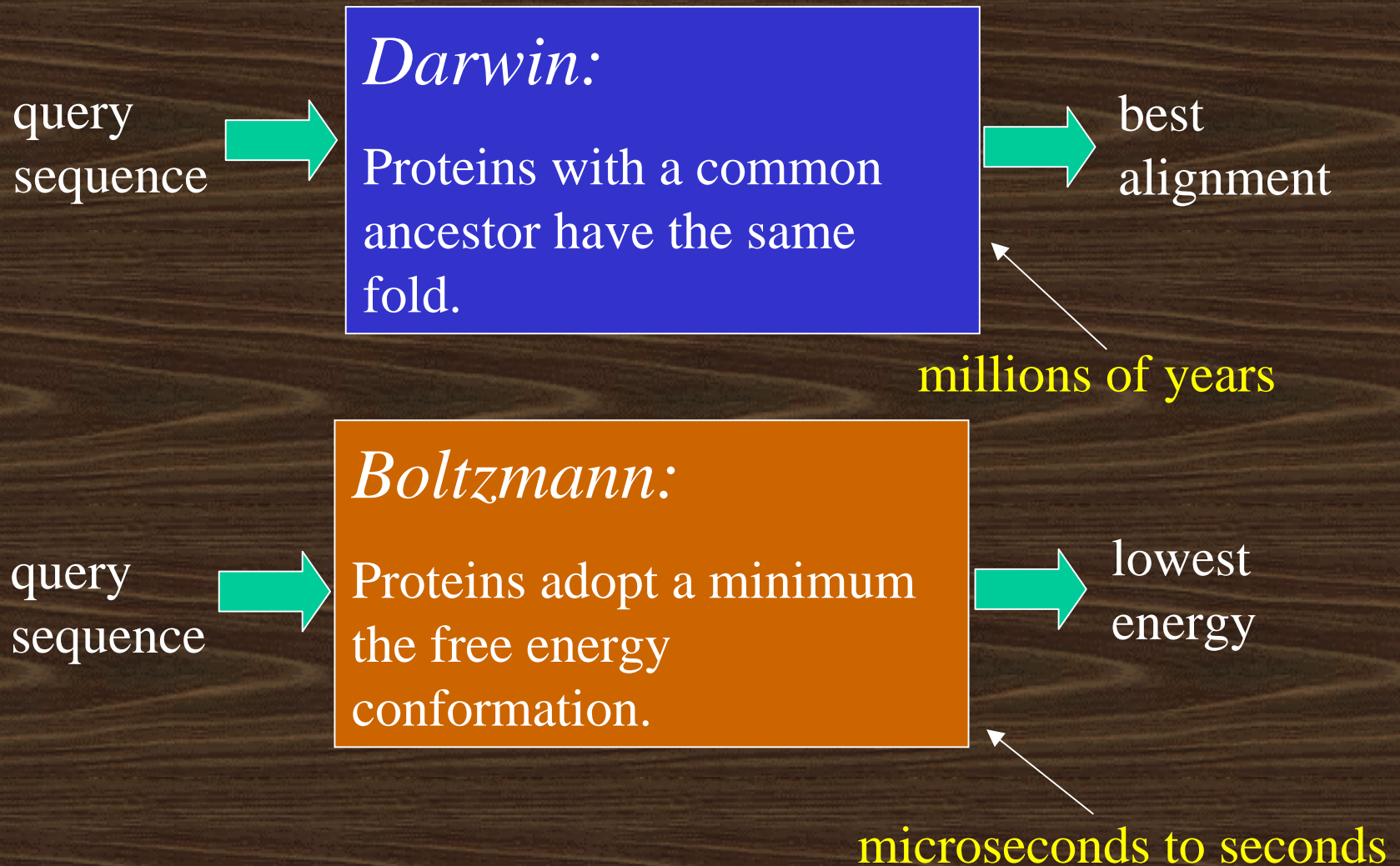


(statistics)

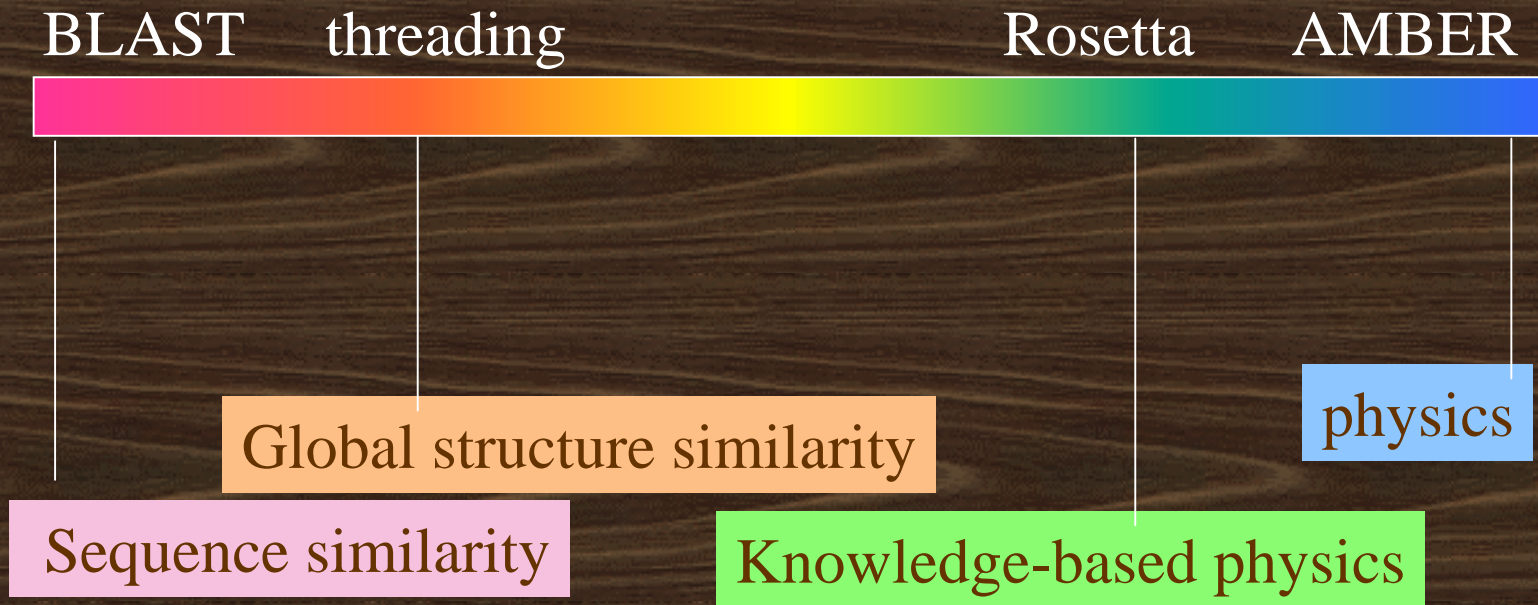


(physics)

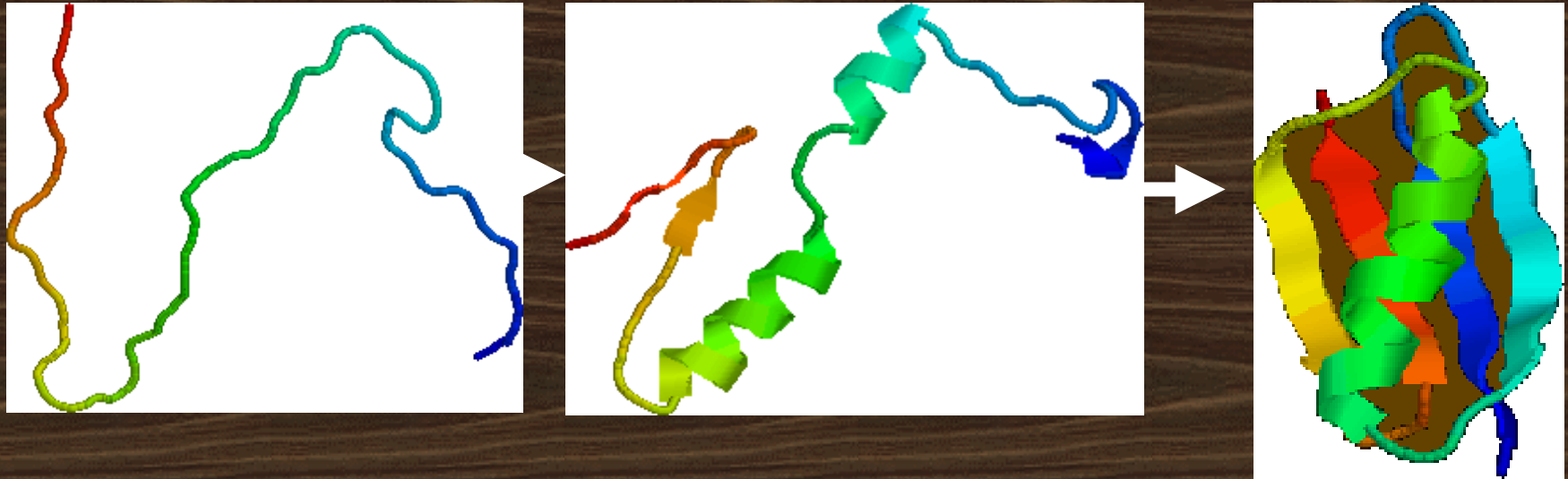
...two very different *Underlying principles*



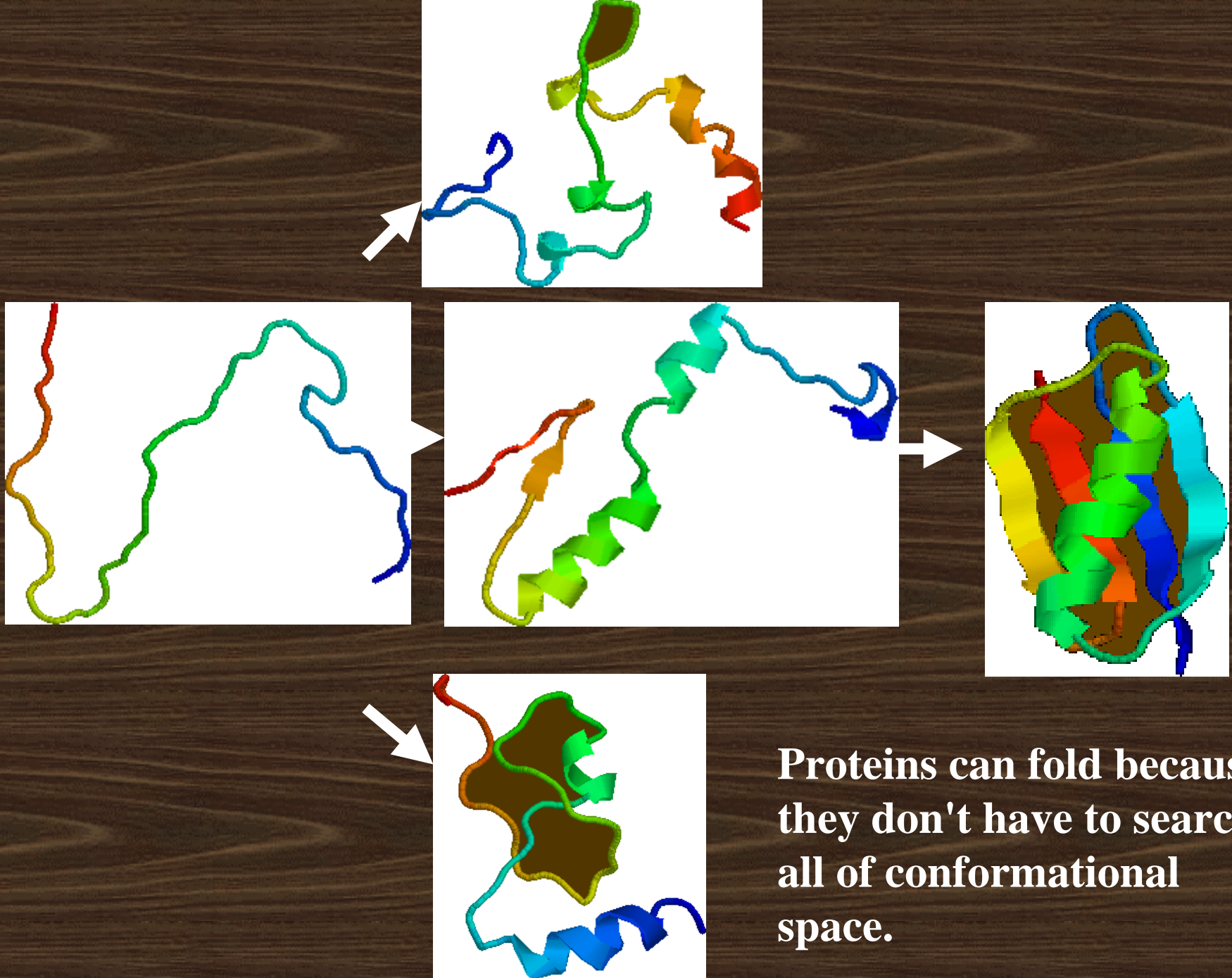
Darwin versus Boltzmann. Do hybrid models make sense?



We know proteins fold via pathways.



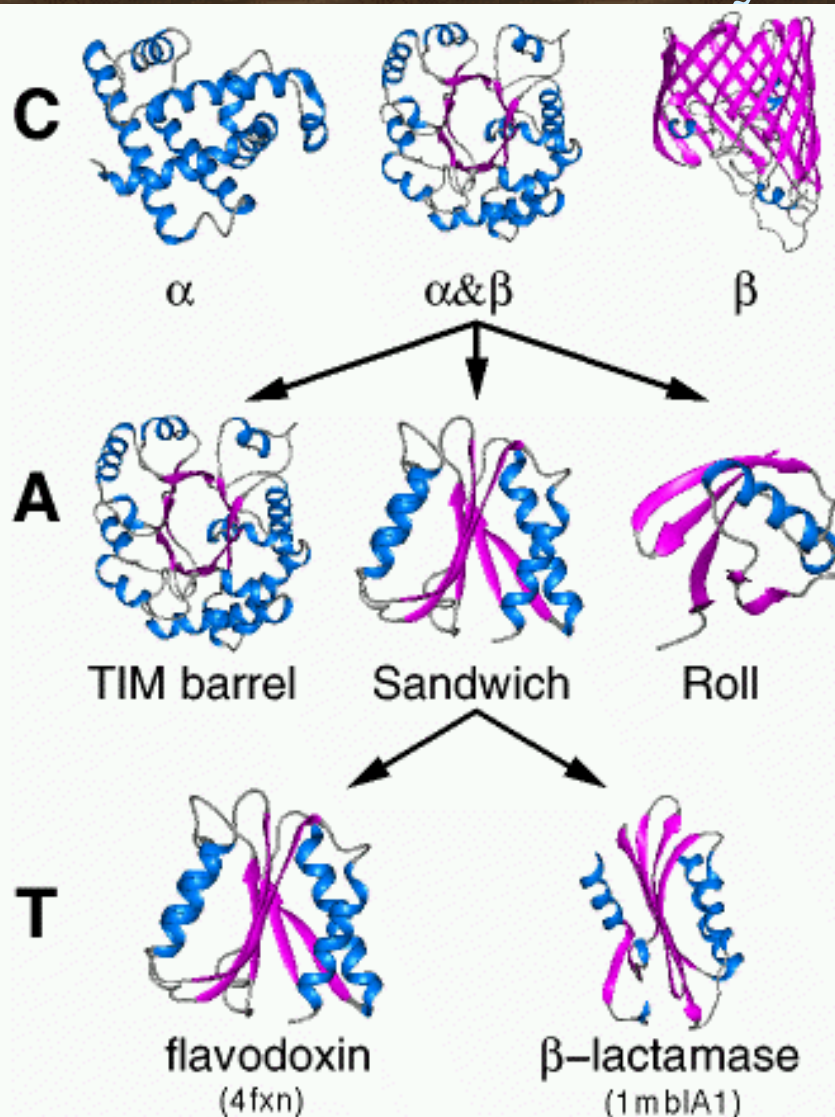
**local structure first, eliminating
alternate pathways, then global**



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

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?



<u>Steps along the folding pathway:</u>	<u>Steps in data mining:</u>
(1) Initiation	local motifs
(2) propagation	extended local motifs
(3) condensation	pairs of motifs
(4) molten globule	multiple motifs
(5) native state	aligned multiple motifs

Heirarchical level 1: Folding initiation site motifs

Non-homologous sequences

recurrent
sequence

HDFPIEGGDS **P M Q T I F F** W S N A N A K L S H G Y
CPYDNIW **M Q T I F F** N Q S A A V Y S V L H L I F L T
IDMNPQGSI **E M Q T I F F** G Y A E S A
ELSPVVNFLE **E M Q T I F F** I S G F T Q T A N S D
INWGS **M Q T I F F** E E W Q L M N V M D K I P S
IFNESKKKGI **A M Q T I F F** I L S G R
PPPM **Q T I F F** V I V N Y N E S K H A L W C S V D
PW M W N L **M Q T I F F** I S Q Q V I E I P S
M Q T I F F V F S H D E Q M K L K G L K G A

Is it a recurrent structure?

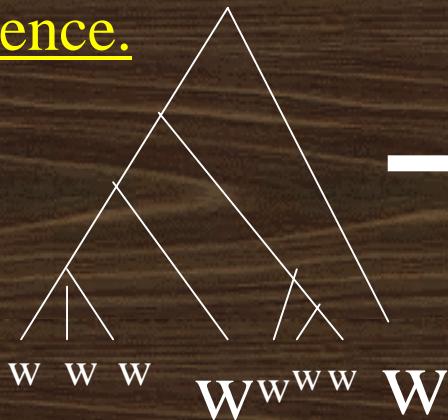
Removing database redundancy

(1): Cluster sequences into phylogenetic trees.



One family, one count.

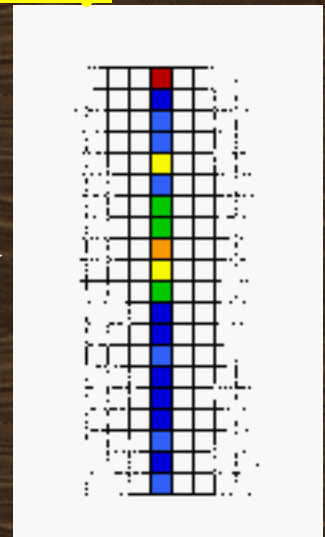
(2): apply a tree weight to each sequence.



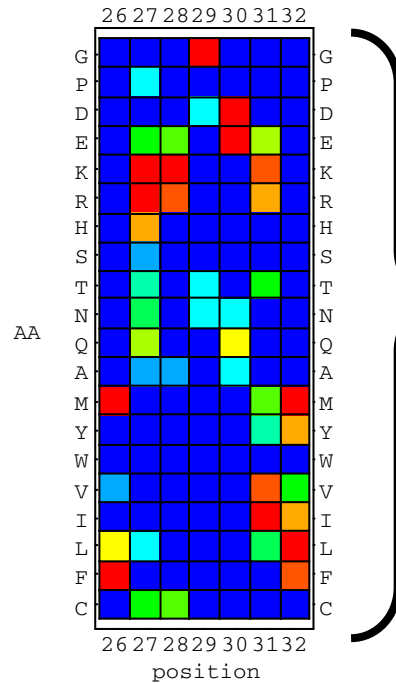
(3): Convert each position to a probability distribution.

$$P_{ij} = \frac{\sum_{k=seqs} w_k \delta(s_{kj} = a_i)}{\sum_{k=seqs} w_k}$$

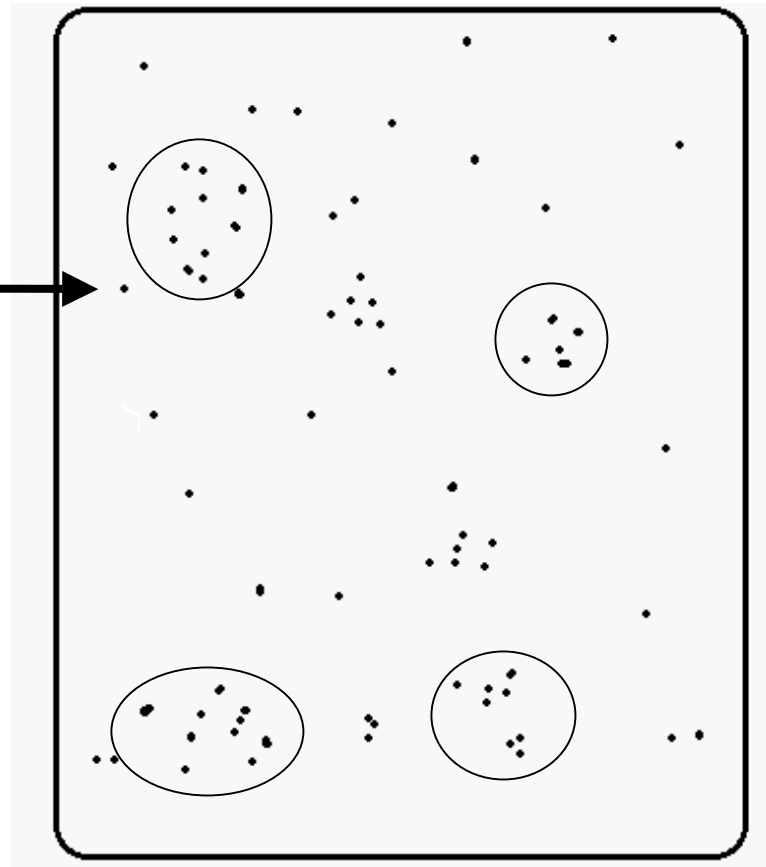
"sequence profile"



Clustering sequence profiles to find recurrent patterns



Each dot
represents a short
profile

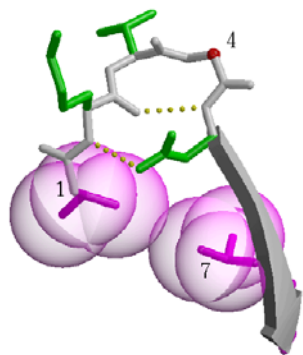


similarity metric (product of log-likelihood ratios)

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

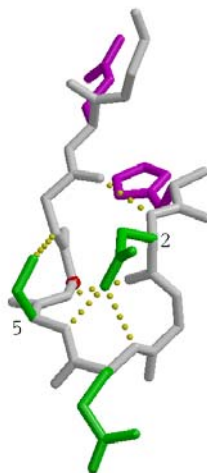
j positions
 i amino acids

The I-sites Library

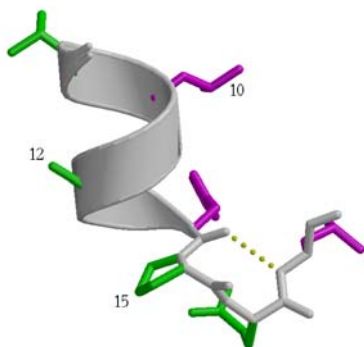


**diverging type-2
turn**

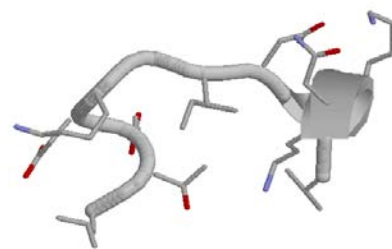
**Serine
hairpin**



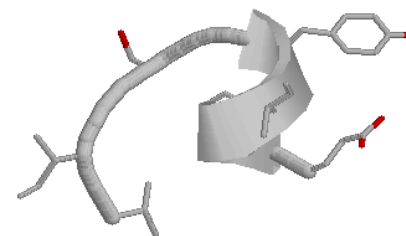
**Amino
acids
arranged
from non-
polar to
polar**



Proline helix C-cap

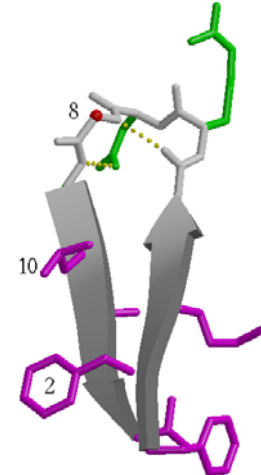
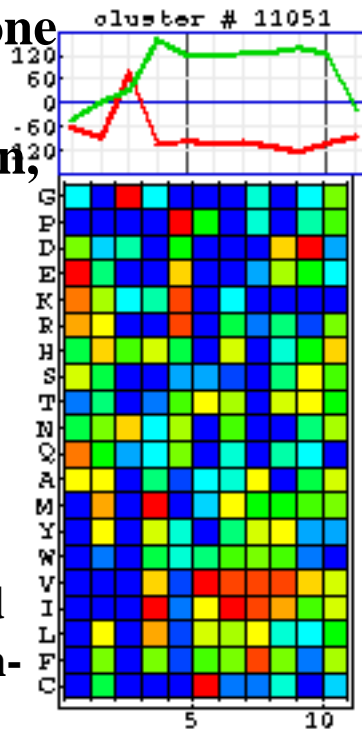


alpha-alpha corner



glycine helix N-cap

**Backbone
angles:
 ψ =green,
 ϕ =red**



**Type-I
hairpin**



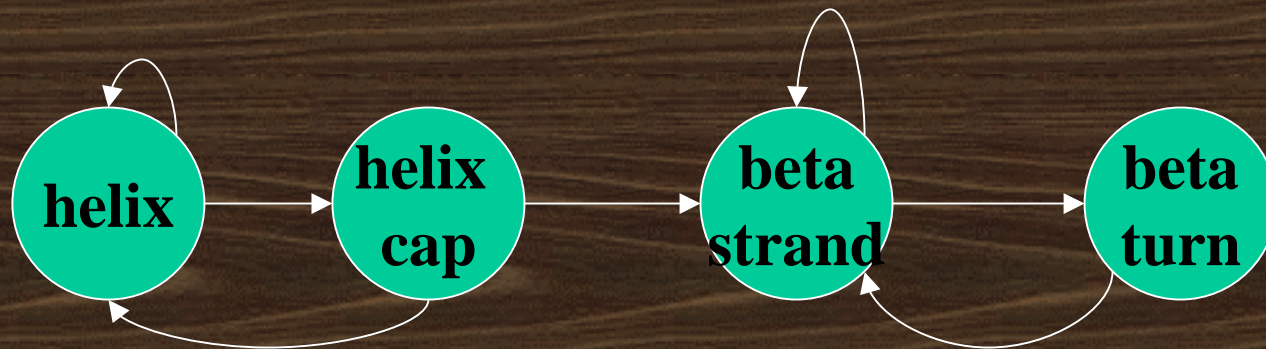
**Frayed
helix**

Are I-sites really folding initiation sites?

Prediction experiments	(Bystroff & Baker, Proteins, 1997)
NMR data on peptides	(Yi <i>et al</i> , J.Mol.Biol., 1998)
Molecular dynamics simulations	(Bystroff & Garde, Proteins, 2002)

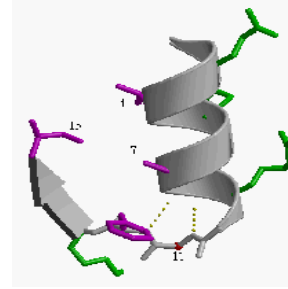
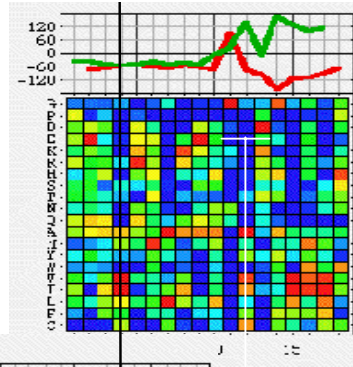
Level 2. Motif grammar

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

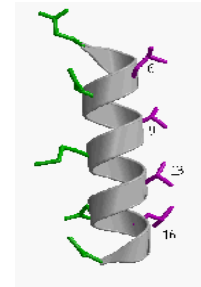
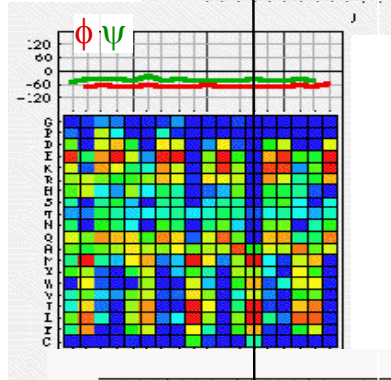


Adjacencies can be modeled as a Markov chain

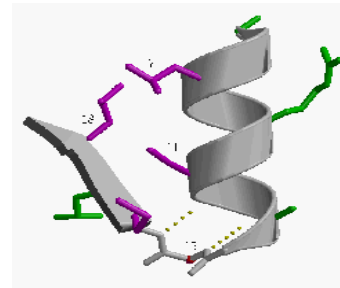
Aligned motifs become a Markov chain



Type-1
G α C-cap



α helix

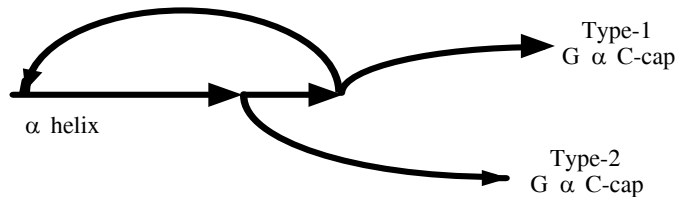


Type-2
G α C-cap

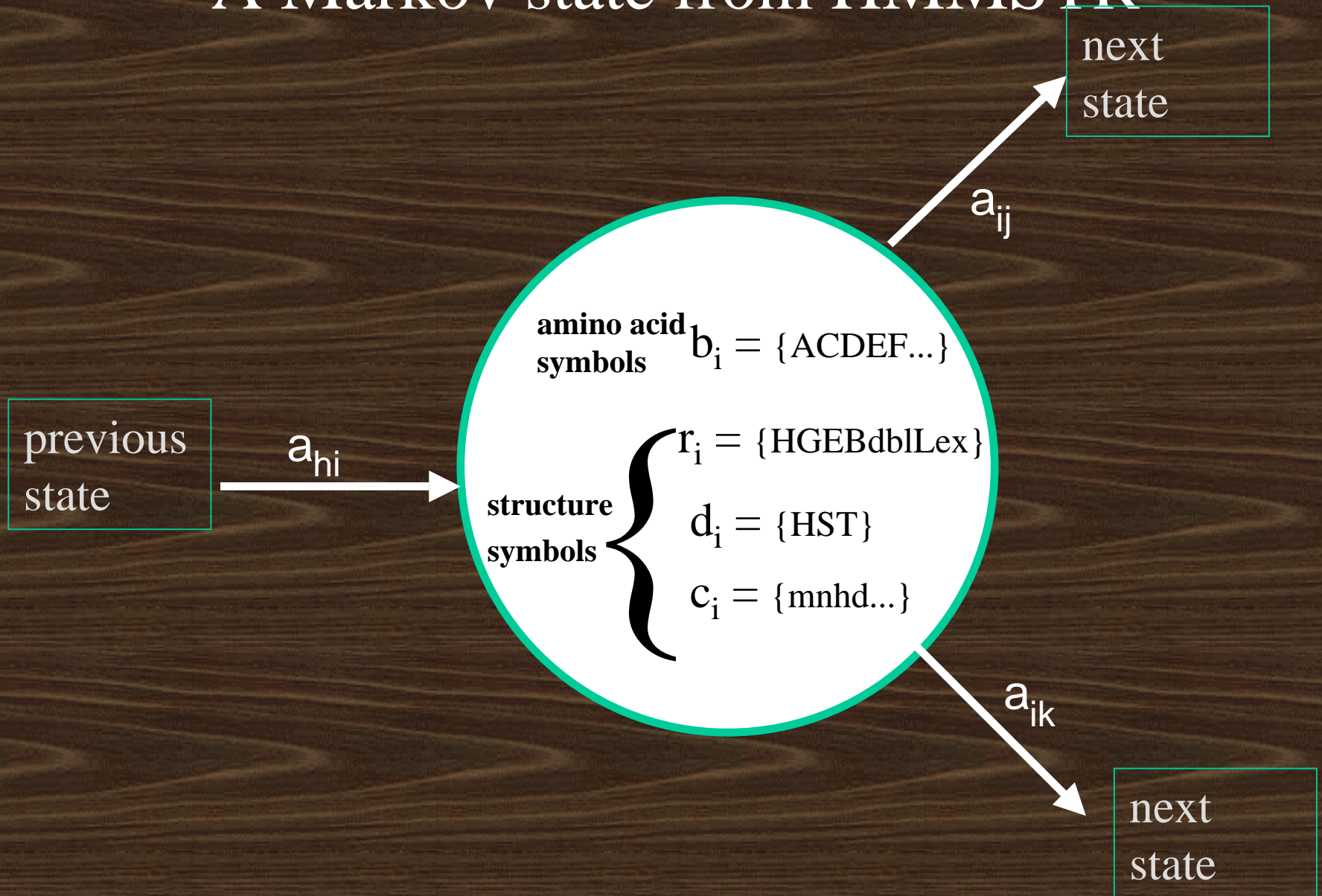
aligned
profiles

aligned
structures

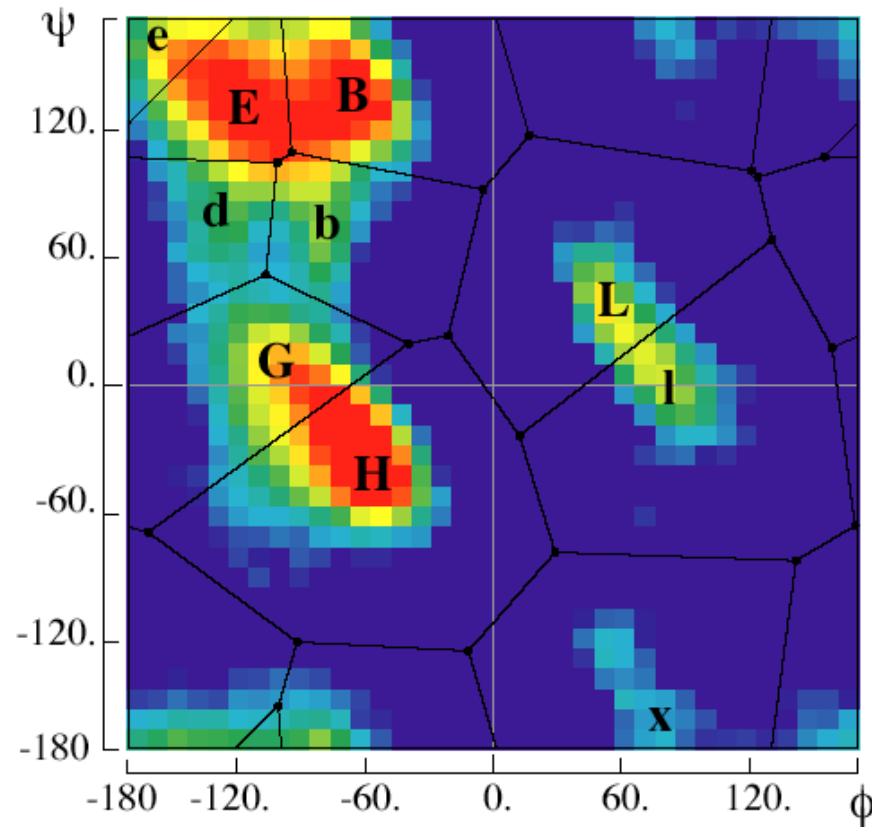
state
topology:



A Markov state from HMMSTR



Discretized structure states: backbone angle regions (r_i)



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

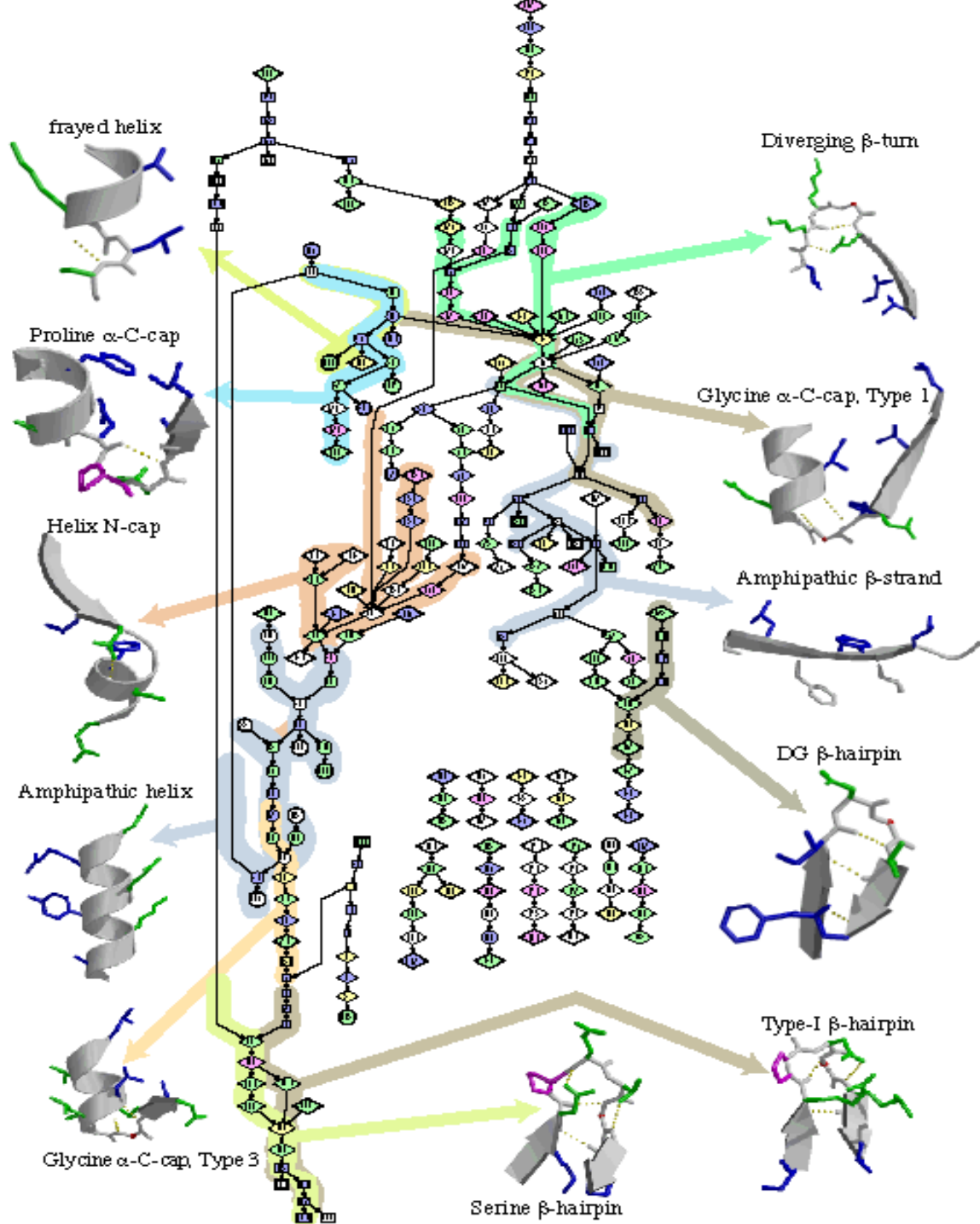
$$P(Q | S) = \pi_{q_1}(s_1) \prod_{t=2, N} a_{q_{t-1}q_t} B_{q_t}(s_t)$$

starting states

arrows

amino acid profiles

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



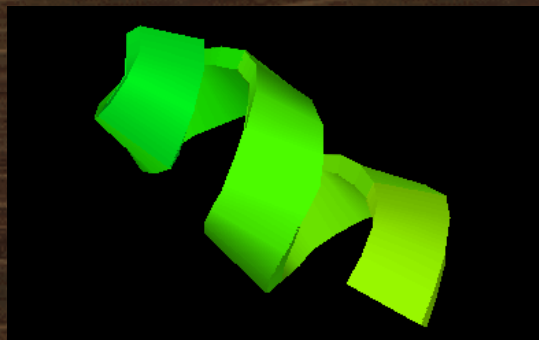
HMMSTR
Hidden Markov Model for local protein STRucture

282 nodes
 317 transitions
 Unified model for 31 distinct sequence-structure motifs

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

Level 1: I-sites

Level 2: HMMSTR



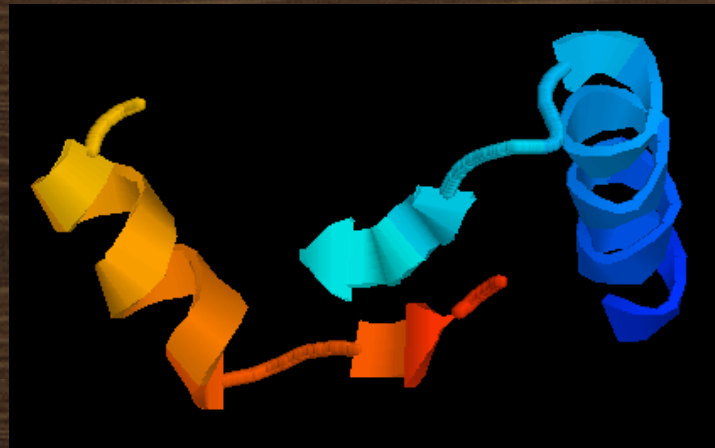
initiation



propagation

Level 3: Pairwise Motif-Motif Contact Potentials

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

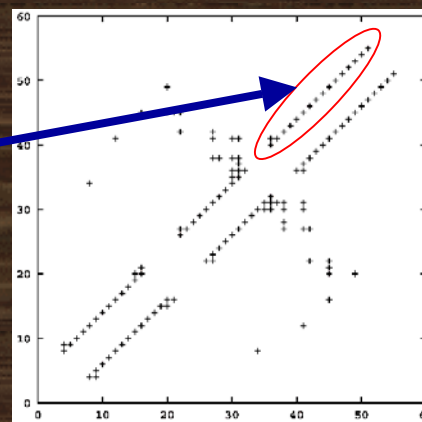
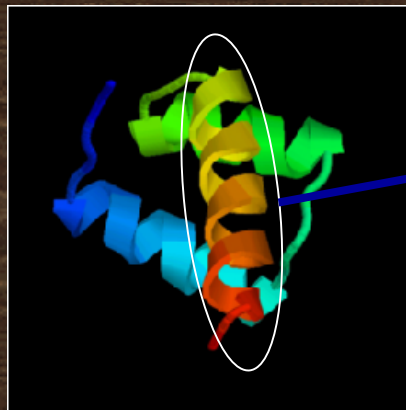
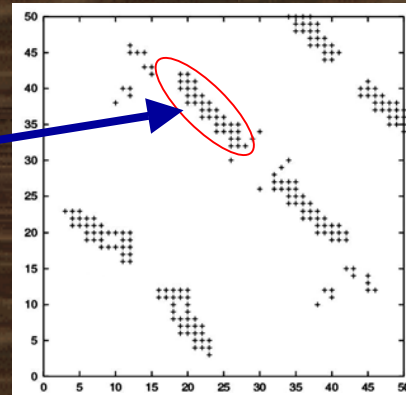
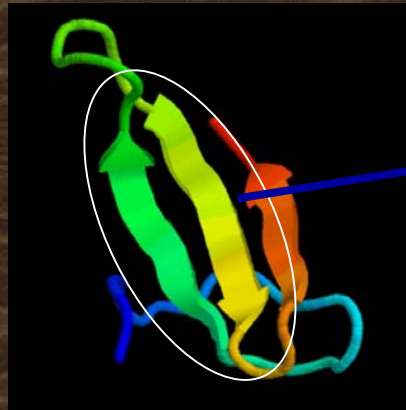


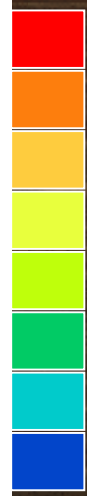
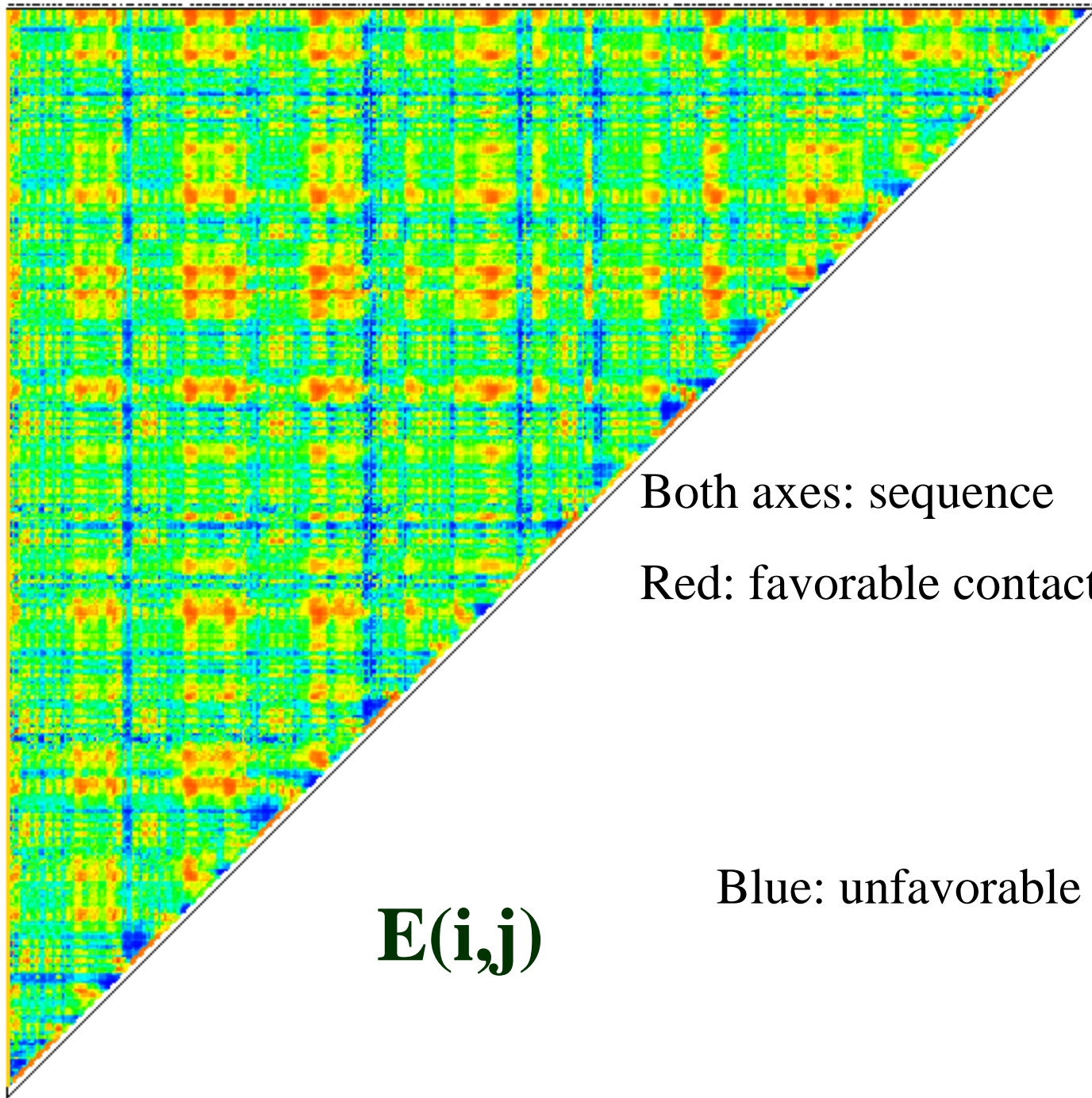
$$G(p, q, s) = -\log \frac{\sum_{PDBselect} \sum_{i \ni D_{i,i+s} < 8\text{\AA}} \Gamma(i, p) \Gamma(i + s, q)}{\sum_{PDBselect} \sum_i \Gamma(i, p) \Gamma(i + s, q)}$$

What is a contact map?

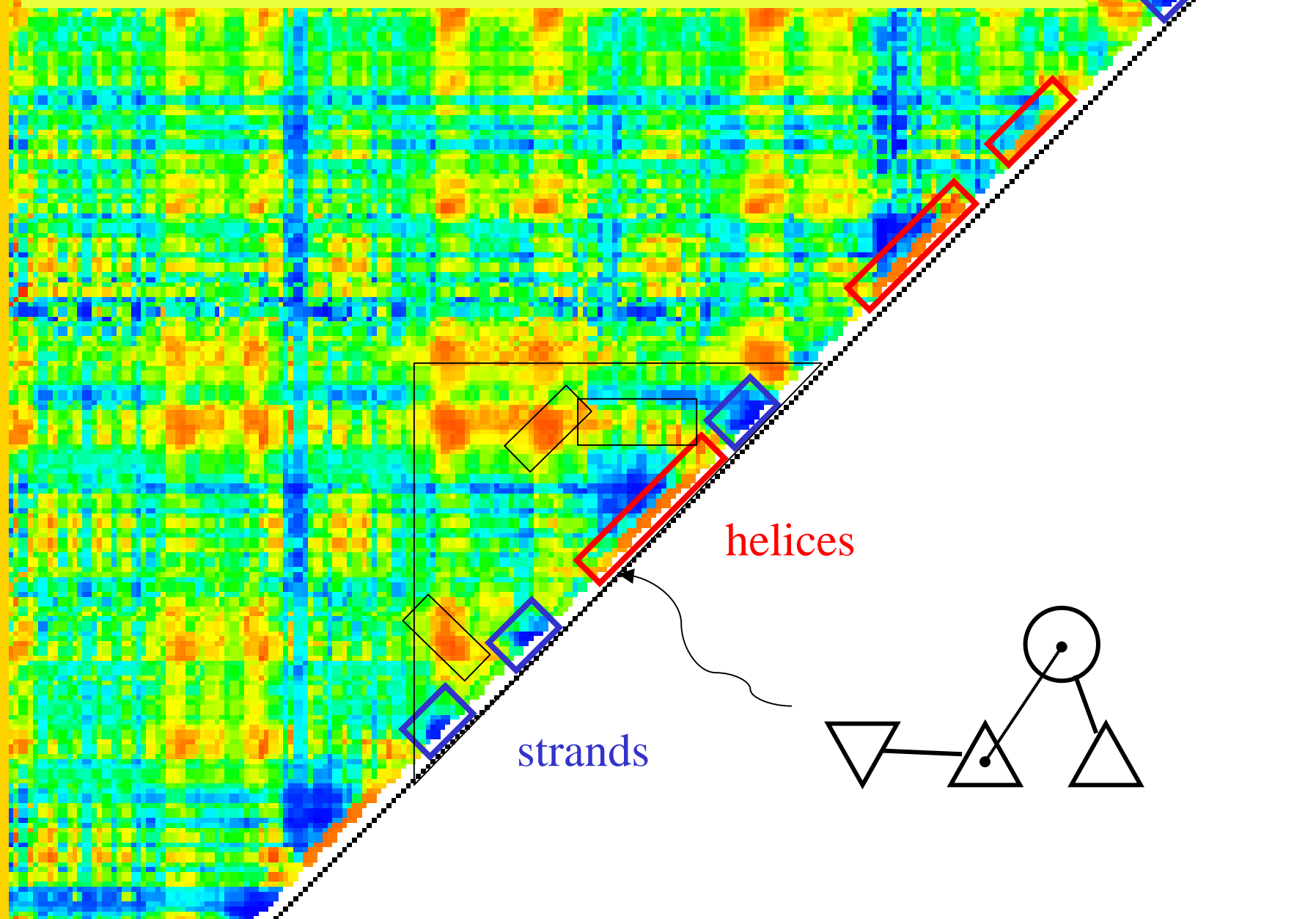
Definition:

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

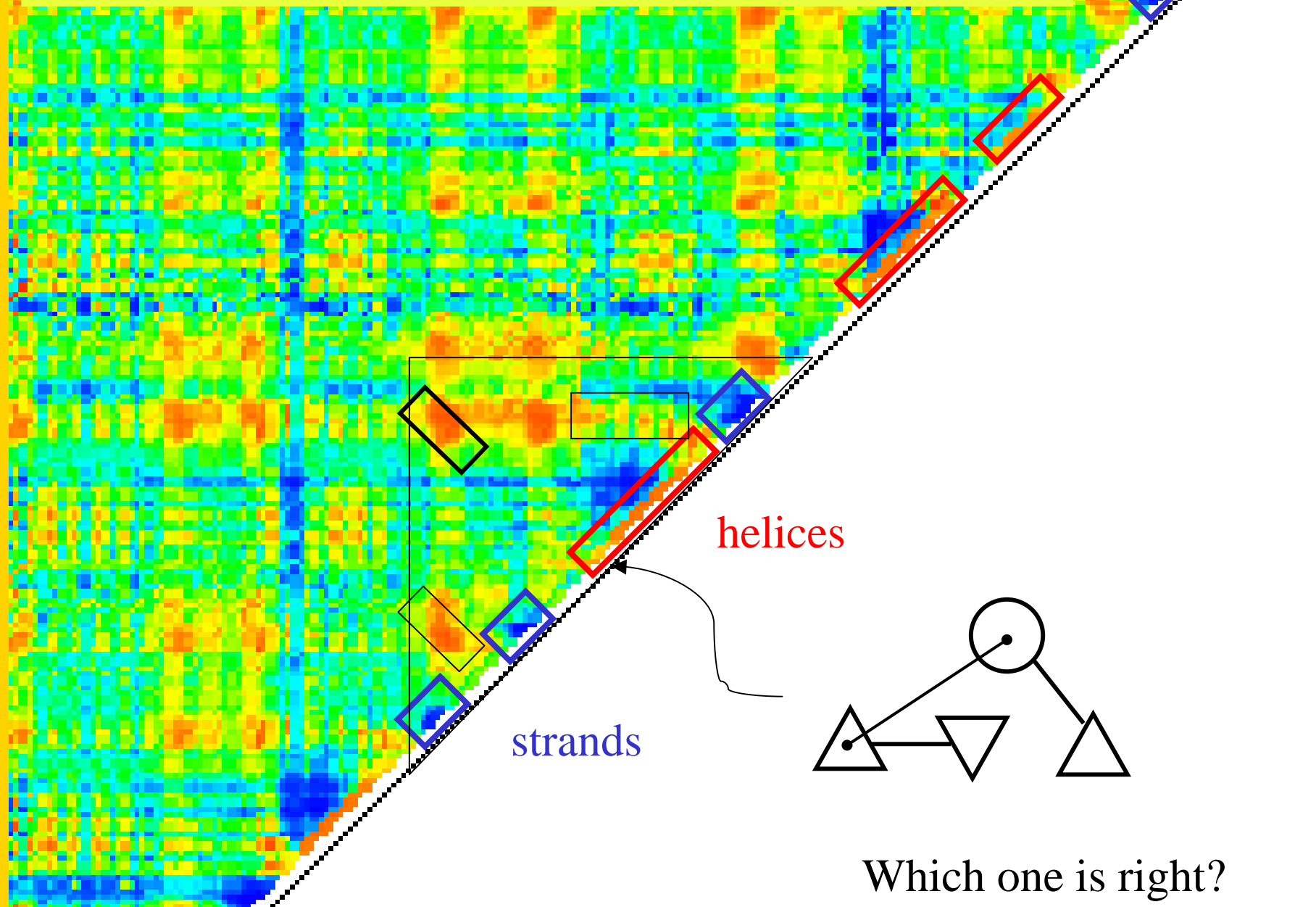




Features in a contact map can be interpreted as a TOPS diagram

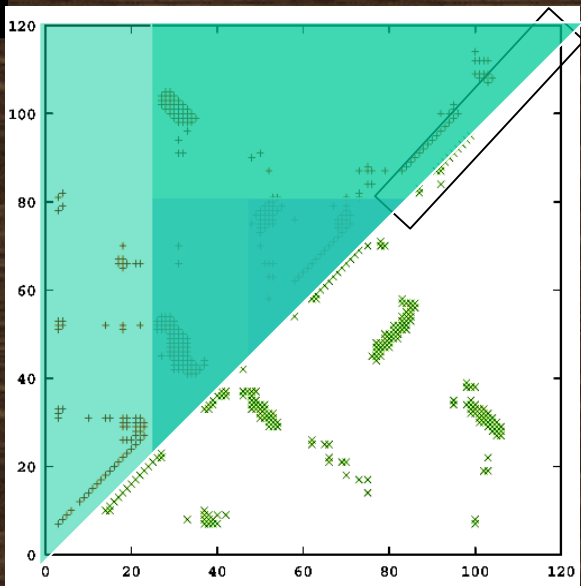


Features in a contact map can be interpreted as a TOPS diagram



Which one is right?

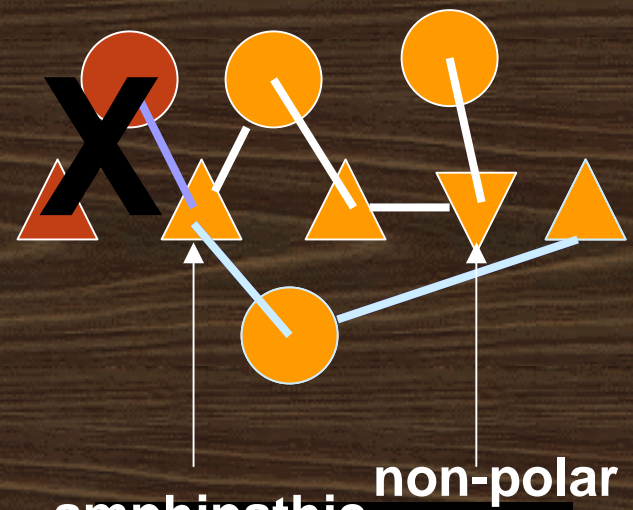
ab initio Prediction



True Contact Map
T0130

True contact map

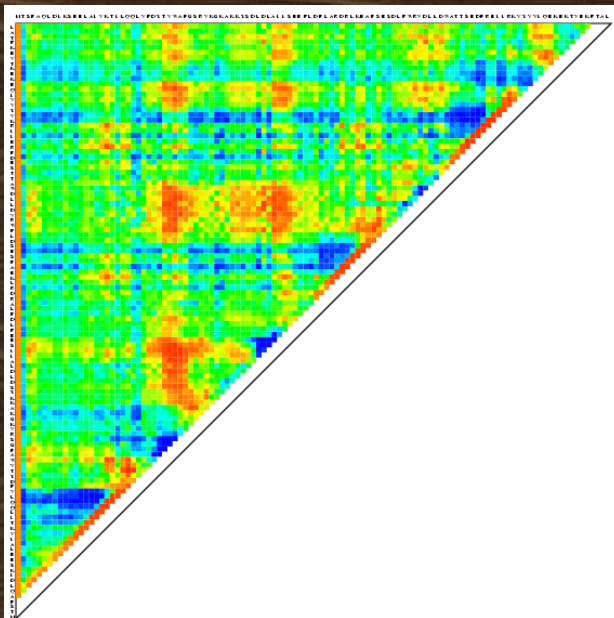
A rule-based simulation procedure.



amphipathic

non-polar

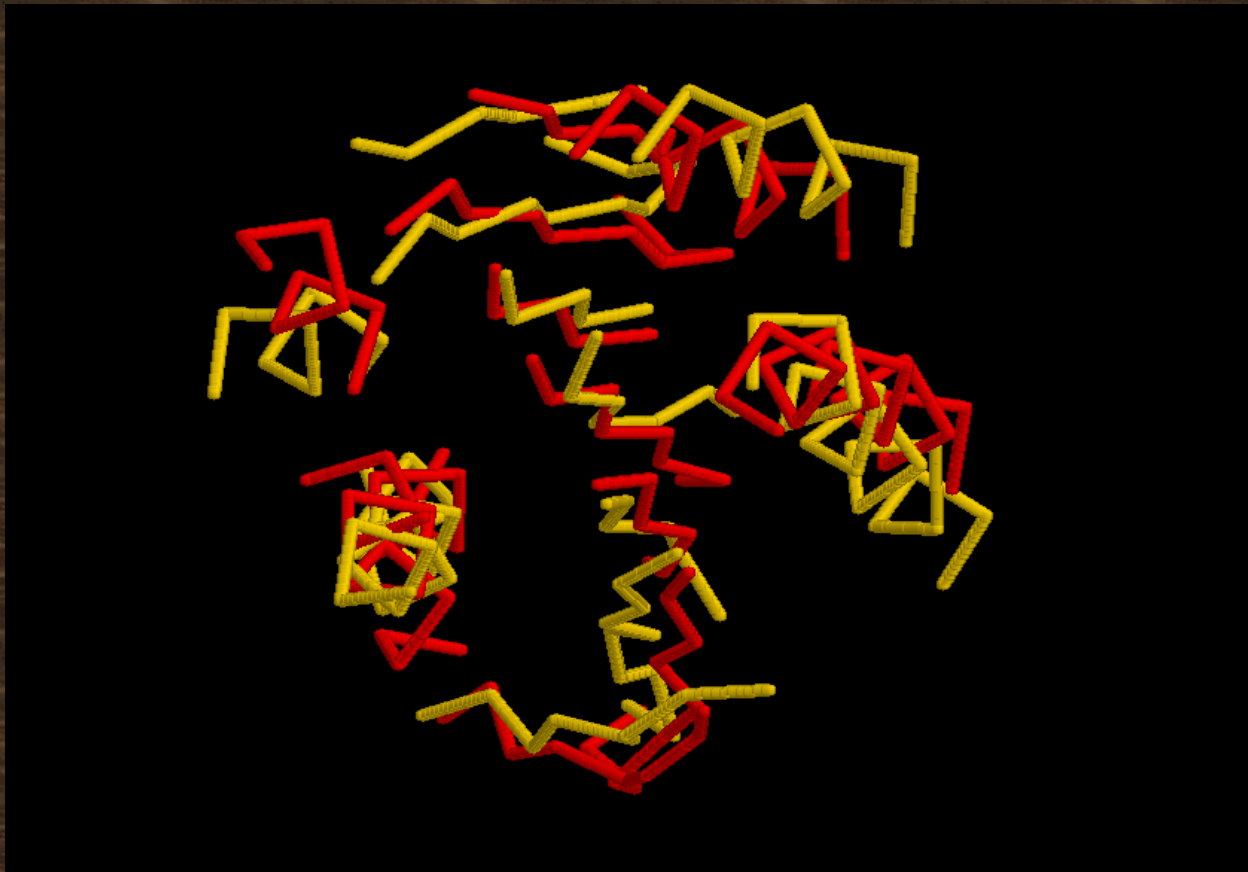
Contact energies



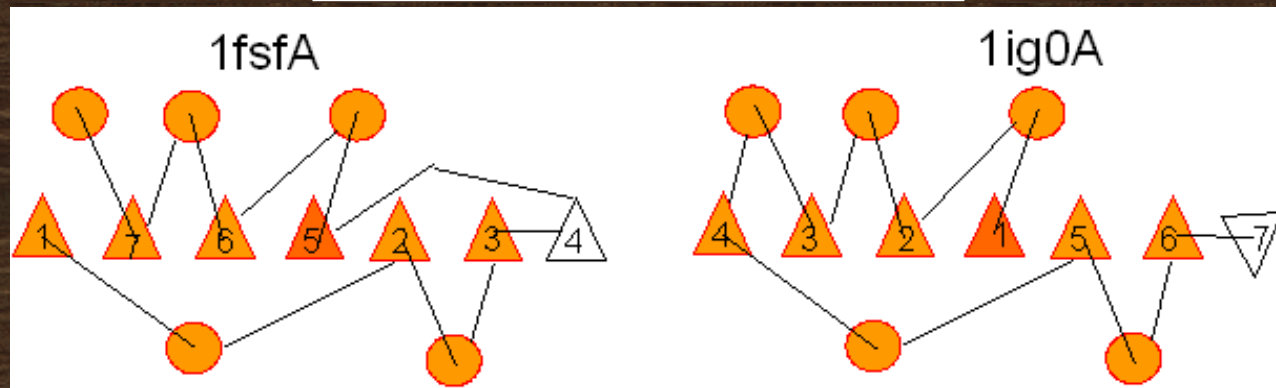
T0130

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

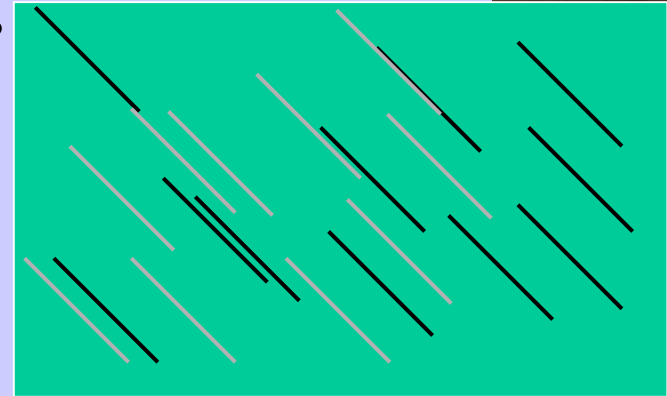
(1) Gapless alignment of HMMSTR states

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

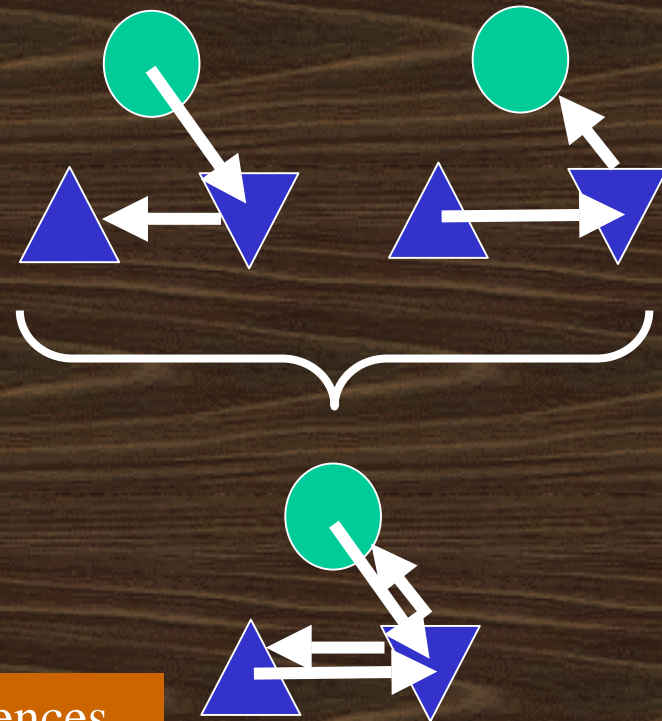
(3) Add a new fragment *iff* it is compatible and has a high score .

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

Score of leaves = aligned contacts + permutation penalty.

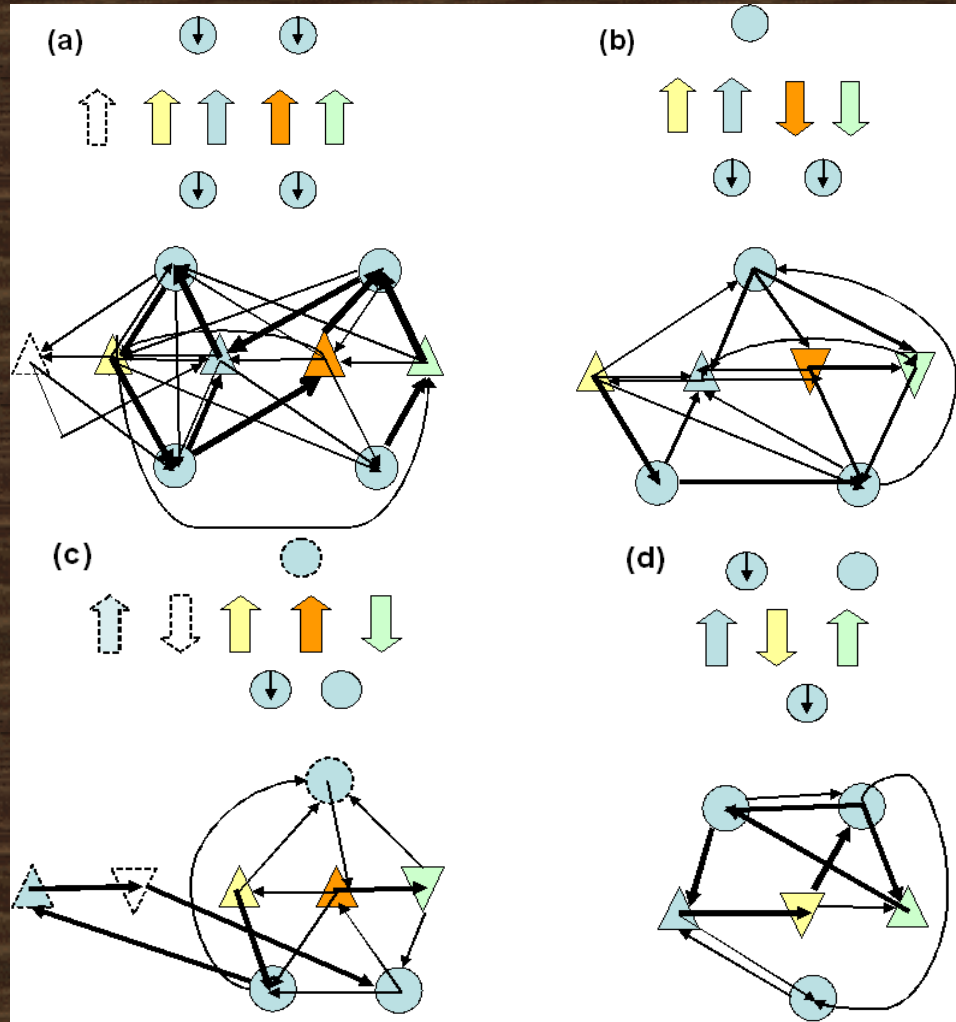


HMMs may be built based on non-sequential alignments

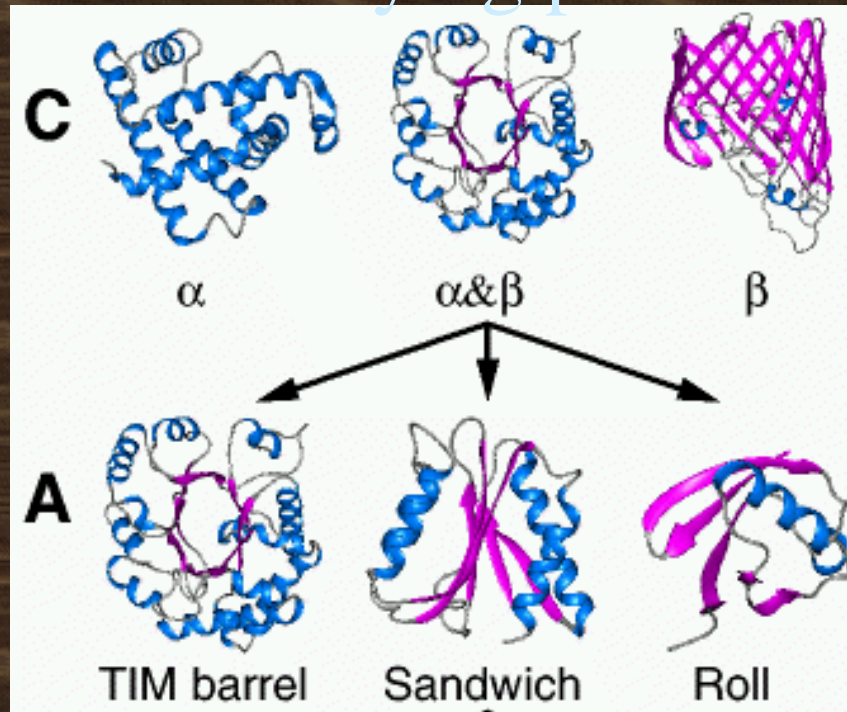


Markov states represent amino acid sequences and positions in space. Connections between them represent loops.

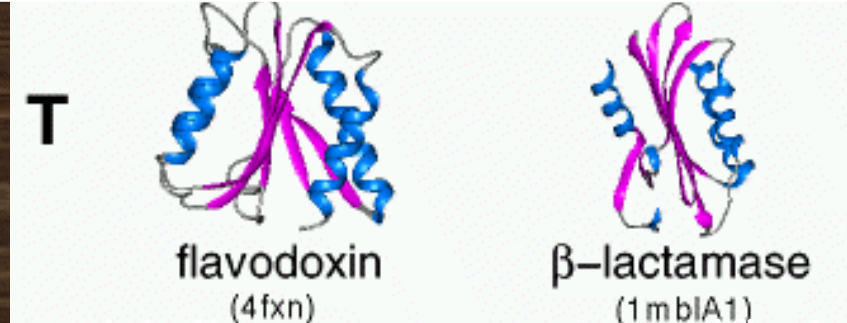
Hidden Markov models for $\alpha/\beta/\alpha$ proteins



Non-sequential clusters may be a useful for classifying proteins



Core packing classes



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

Level 1: I-sites

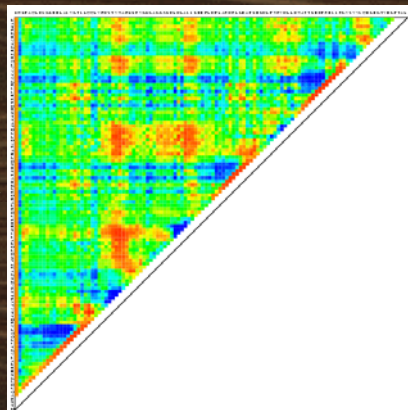


Level 2: HMMSTR



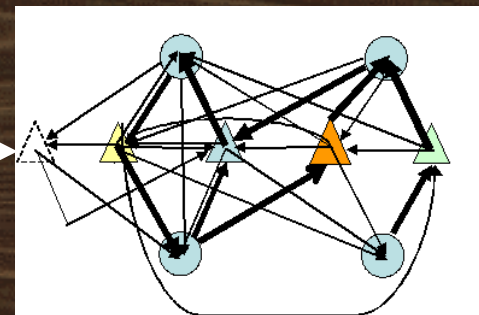
propagation

Level 3: HMMSTR-CM



condensation

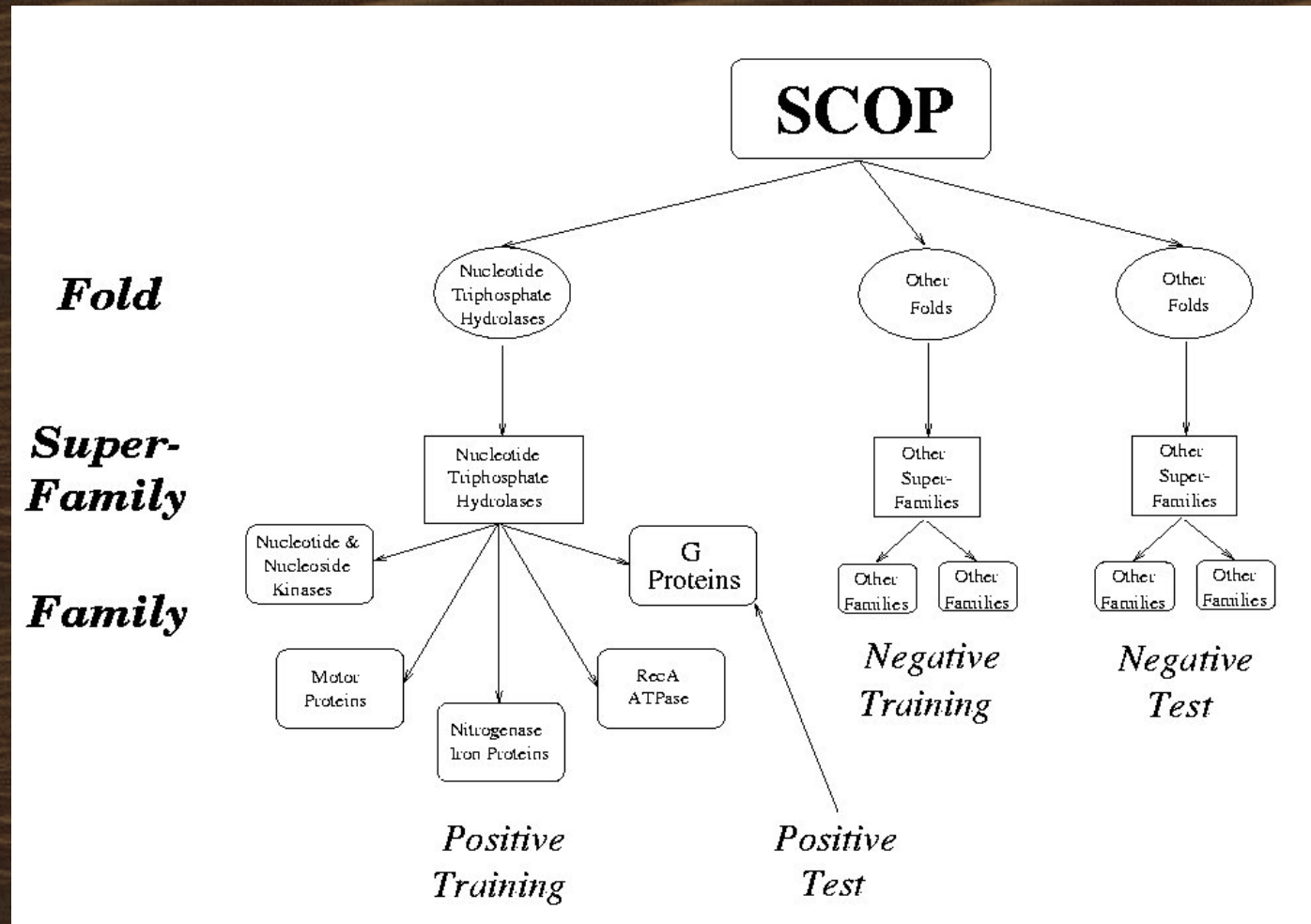
Level 4: SCALI



molten
globule

initiation

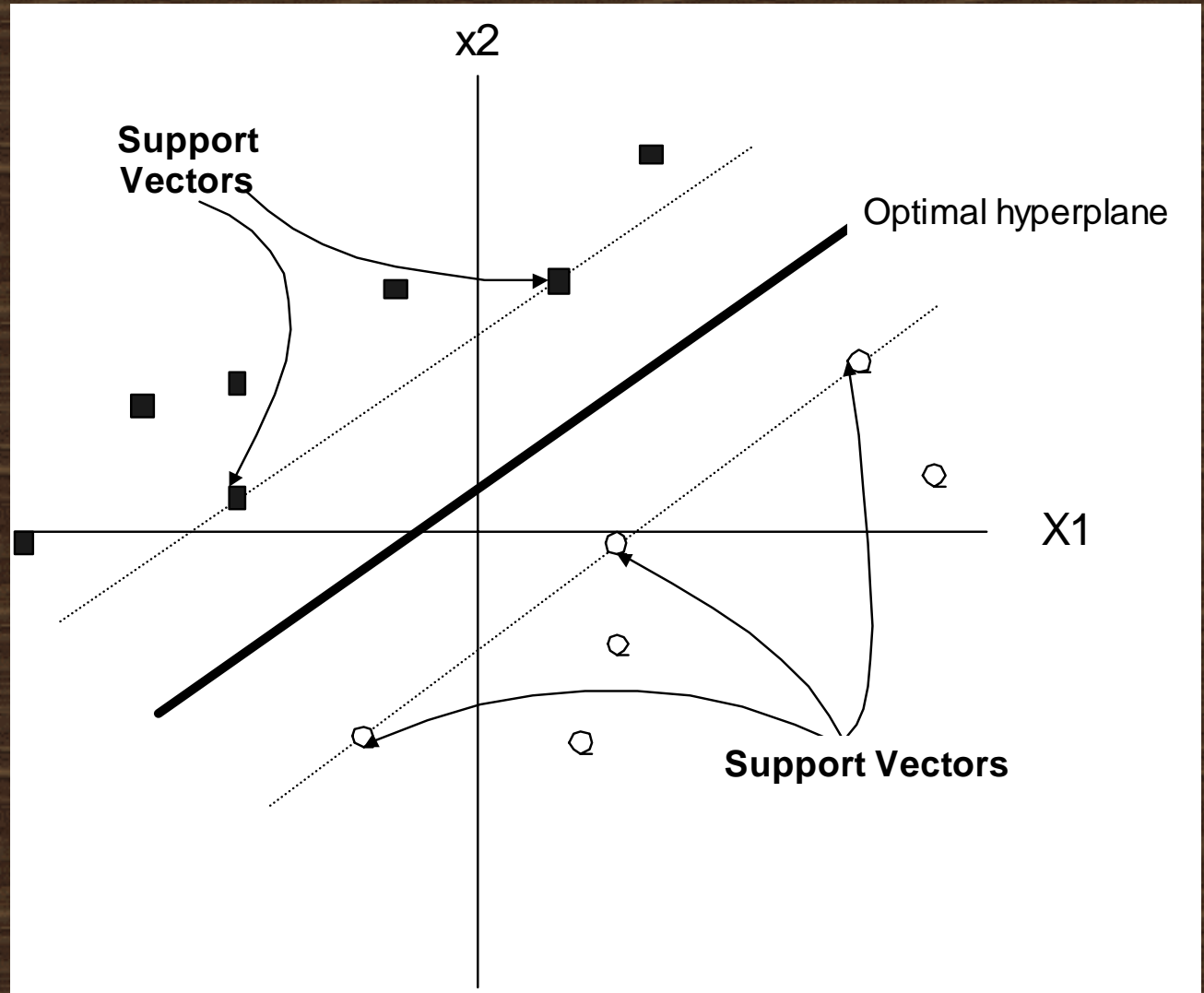
Level 5: Global topology



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

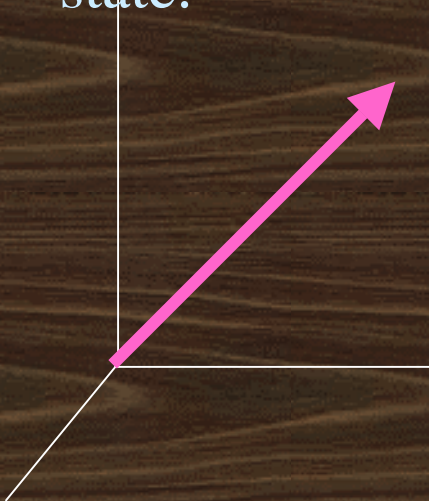
Support Vector Machine

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

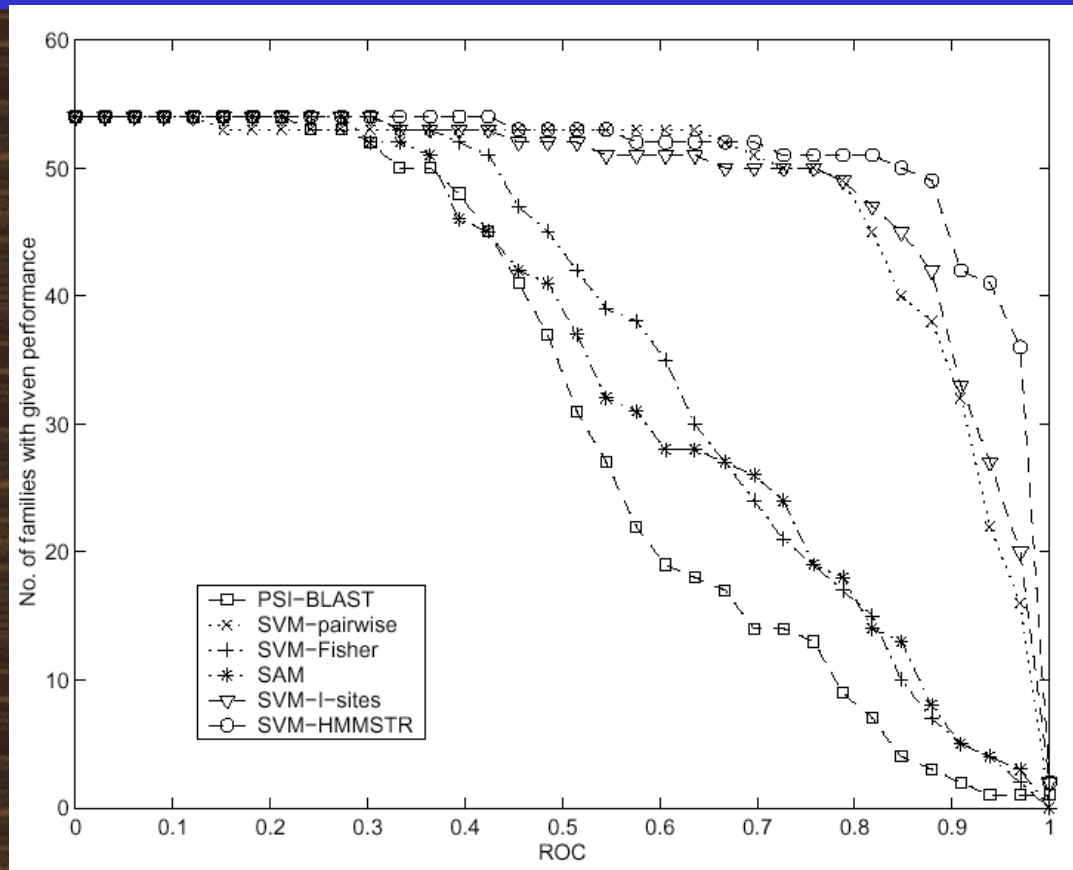


HMMSTR as the basis for a Support Vector Machine

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?

Steps along the

folding pathway:

Model

Complexity

(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

SVM-HMMSTR

~1000-2000 folds

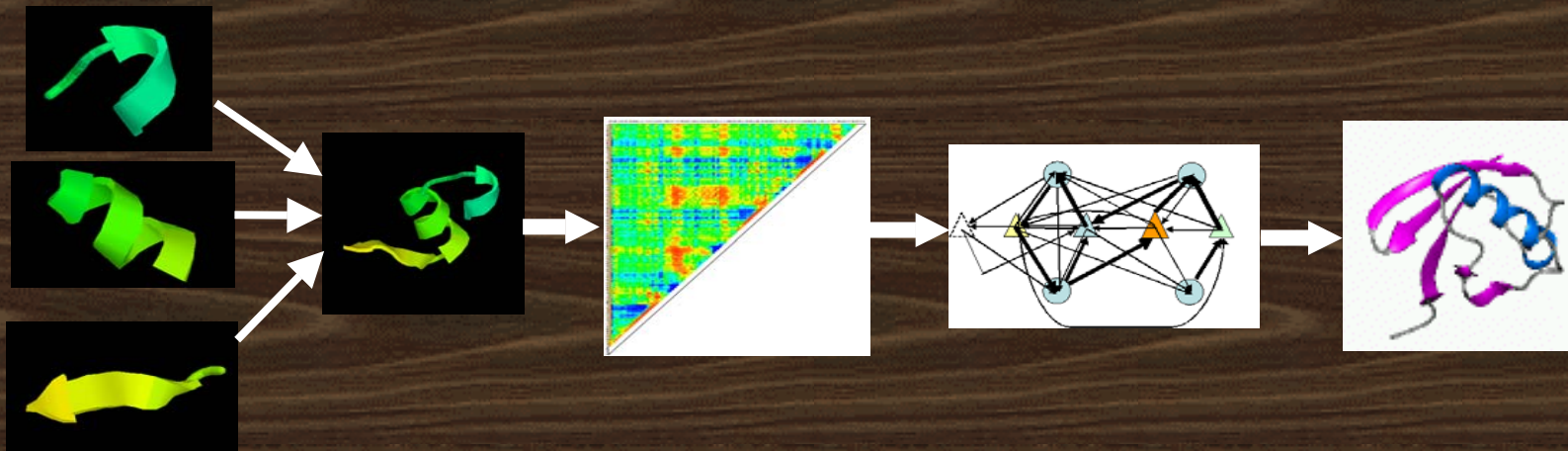
early



late

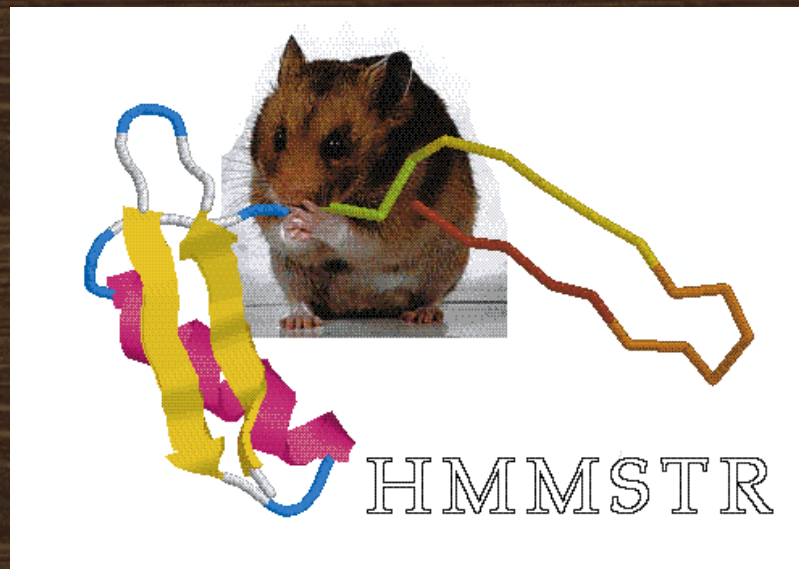
Are there any conclusions?

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:
Chris Bystroff
Vesteinn Thorsson
David Baker



Funding from:
NSF-CISE

Bystroff Lab

Yu Shao
Xin Yuan
Kwang Kim

Yaoming Huang
Donna Crone
Rachel van Duyne
Ben Cole

SVM-HMMSTR
(Nat.Univ.Singapore)

Yuna Hou
Mong-Li Lee
Wynne Hsu

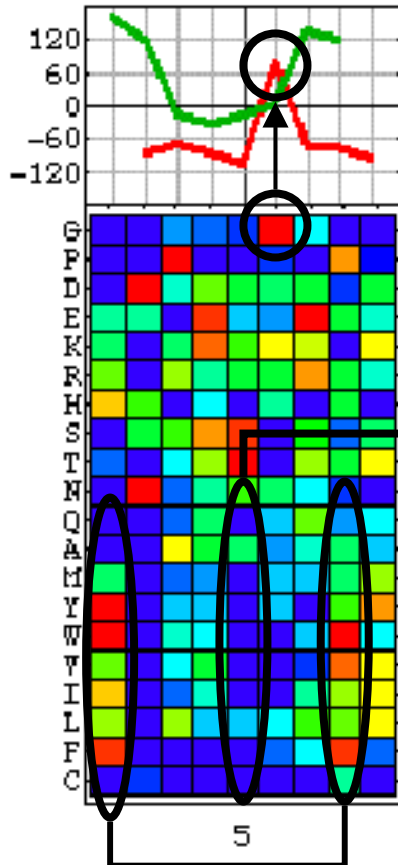
www.bioinfo.rpi.edu/~bystrc/

HMMSTR says: Think Globally, Act Locally.

Are I-sites folding initiation sites?

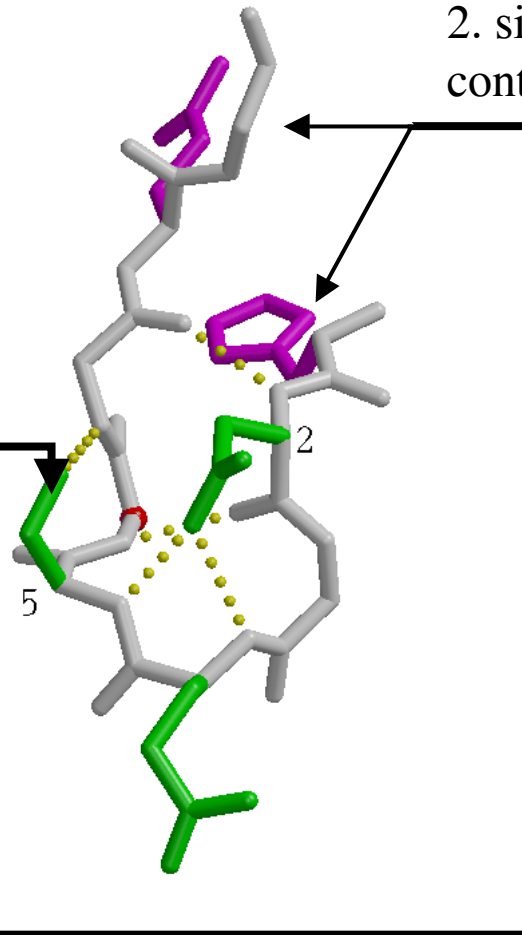
Patterns of conservation suggest energetic motive

1. backbone angle constraints

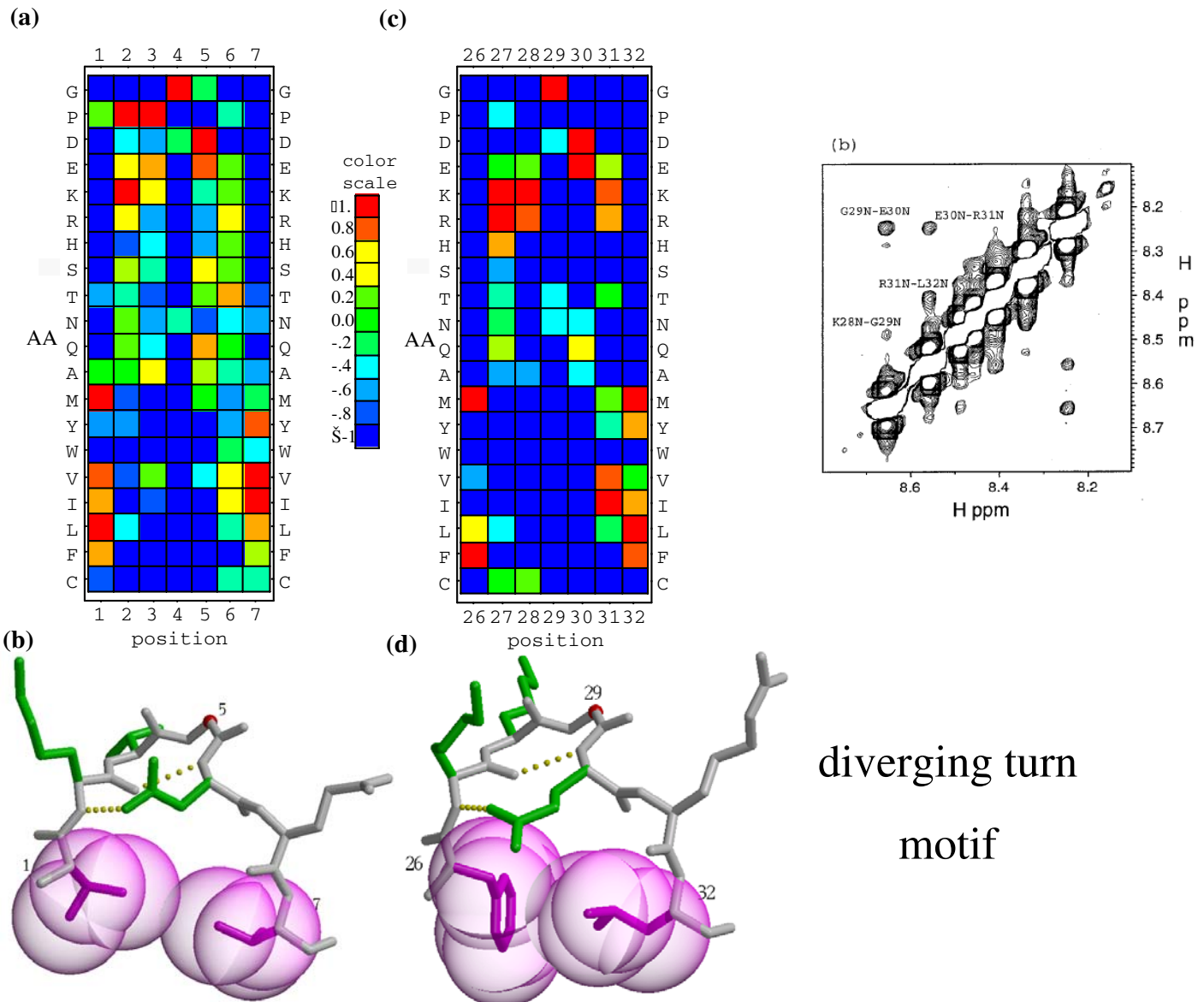


3. negative design

2. sidechain contacts



NMR structures confirm independent folding



NMR structure of a 7-residue I-sites motif in isolation (Yi *et al*, J. Mol. Biol, 1998)

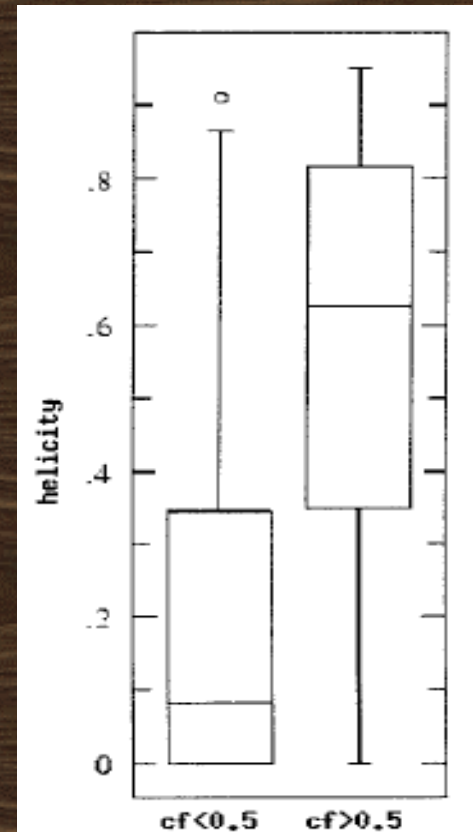
Peptide simulations show a correlation between sequence and stability

- AMBER 6.0
- 800-900 waters.
- Ion balance(Na, Cl)
- 340°K
- ≥ 5 ns

QuickTime™ and a decompressor are needed to see this picture.

Sequences simulated

AAALDRMR	FHMYFMLR	PRDANTSH
AALEALLR	FSVMNDAS	QDDARKLM
AANRSHMP	FYSSYVYL	QGIIDKLD
AARYKFIE	GQLMALKQ	QKMKTYFN
ADFKAAVA	HNLIEAFE	QTLAQLSV
AFDGETEI	IEHTLNEK	RDFEERMN
AKELVVVY	IQNGD W TF	RIILDRHR
AKGVETAD	KAAIAQLR	RLLLKAYR
ARFTKRLG	KKYRPETD	RPIARMLS
ATLEEKLN	KNPDNVVG	RVLGRDLF
CNGGHWIA	KPMGPLLV	SCDVKFPI
DAVTRYWP	KQAHPLDK	TEVMKRLV
DEAIDAYI	KQDKHYGY	TLNEKRIL
DELTRHIR	KSYLRSLR	YASLRSLV
DYVRSKIA	LDLHQTYL	YESHVGCR
EDLVERLK	NAVWAAIK	
EELKQALR	NETHSGRK	
EEMVSKLK	NFLEVGEY	
EKLLESLE	NPVKESRH	
EKPFGTSY	PAIISAAE	
EQIKA AVK	PLQHHNLL	



3-state secondary structure prediction

Training Set	H^{pred}	S^{pred}	T^{pred}	Total
H^{obs}	35943	1794	8983	46720
S^{obs}	1484	18154	10454	30092
T^{obs}	7115	6406	54306	67827
Total	44542	26354	73743	144639

74.9%
correct

Test Set	H^{pred}	S^{pred}	T^{pred}	Total
H^{obs}	4919	318	1342	6579
S^{obs}	196	2286	1361	3843
T^{obs}	793	717	6708	8218
Total	5908	3321	9411	18640

74.6%
correct

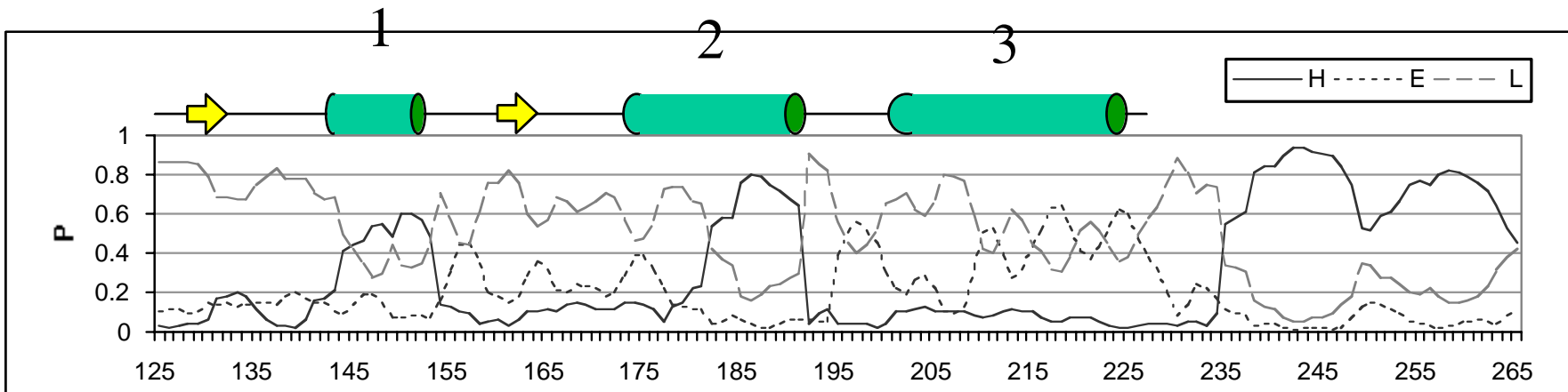
Predicting super-secondary context

$P_d/(P_h+P_d)$	Hairpin	Diverging	$P_n/(P_m+P_n)$	Middle	End
0.0-0.1	2	3	0.0-0.1	88	7
0.1-0.2	18	4	0.1-0.2	307	77
0.2-0.3	17	4	0.2-0.3	279	106
0.3-0.4	23	22	0.3-0.4	240	141
0.4-0.5	43	50	0.4-0.5	196	132
0.5-0.6	23	36	0.5-0.6	98	81
0.6-0.7	17	22	0.6-0.7	68	83
0.7-0.8	5	29	0.7-0.8	11	31
0.8-0.9	4	18	0.8-0.9	8	25
0.9-1.0	0	34	0.9-1.0	1	4
Total	152	222	Total	1296	687

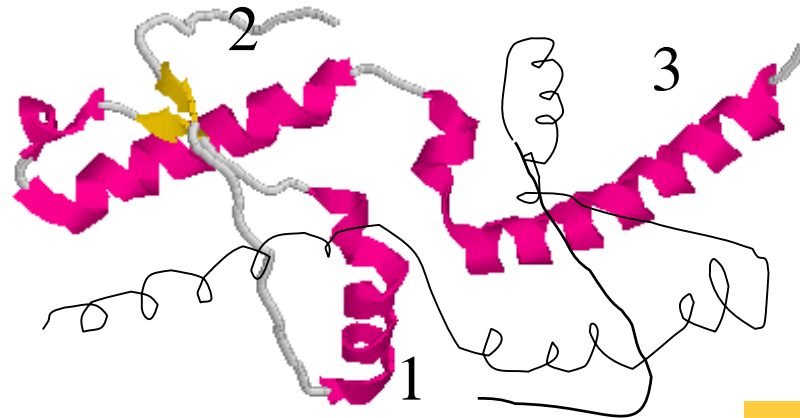
Results are for the independent test set.

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 mutations.

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