CMPS 6630: Introduction to Computational Biology and Bioinformatics

Secondary Structure Prediction

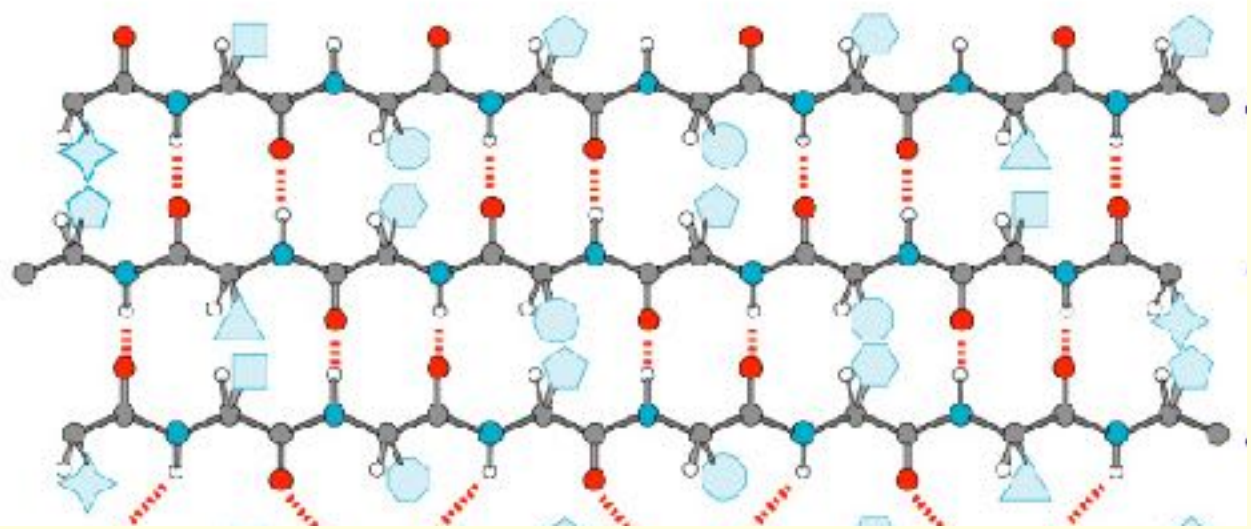
Secondary Structure Annotation

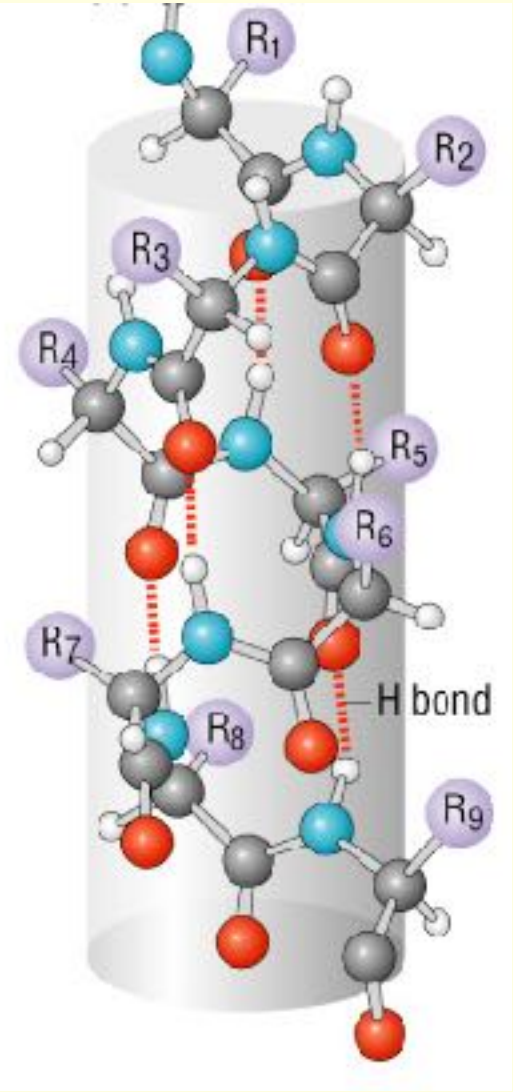
- Given a macromolecular structure
- Identify the regions of secondary structure

Assumptions:

High resolution structure

There is a reasonable and consistent way to define Secondary Structure





Petsko, Ringe, Prot Struct and Function, 2004

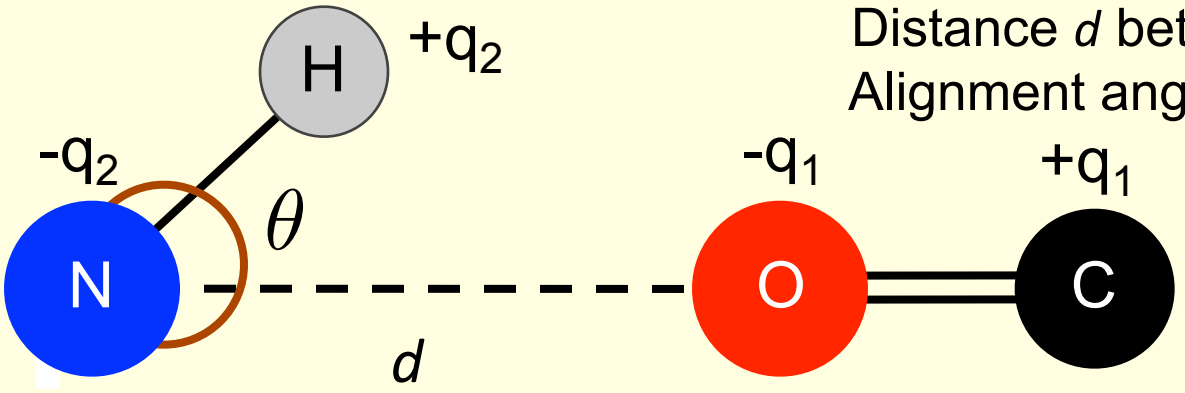
DSSP (Defined Secondary Structure of Proteins)

- A gold standard for secondary structure identification (NOT prediction)
- Used by the PDB database for secondary structure annotation
- Based on **Hydrogen-Bonding Patterns**

Quality of H-bond is function of:

Distance *d* between donor and acceptor

Alignment angle θ



$$E = f\frac{q_x q_y}{r_x y}$$

$$E = q_1 q_2 \left(\frac{1}{r_{\mathrm{ON}}} + \frac{1}{r_{\mathrm{CH}}} - \frac{1}{r_{\mathrm{OH}}} - \frac{1}{r_{\mathrm{CN}}} \right) f$$

DSSP

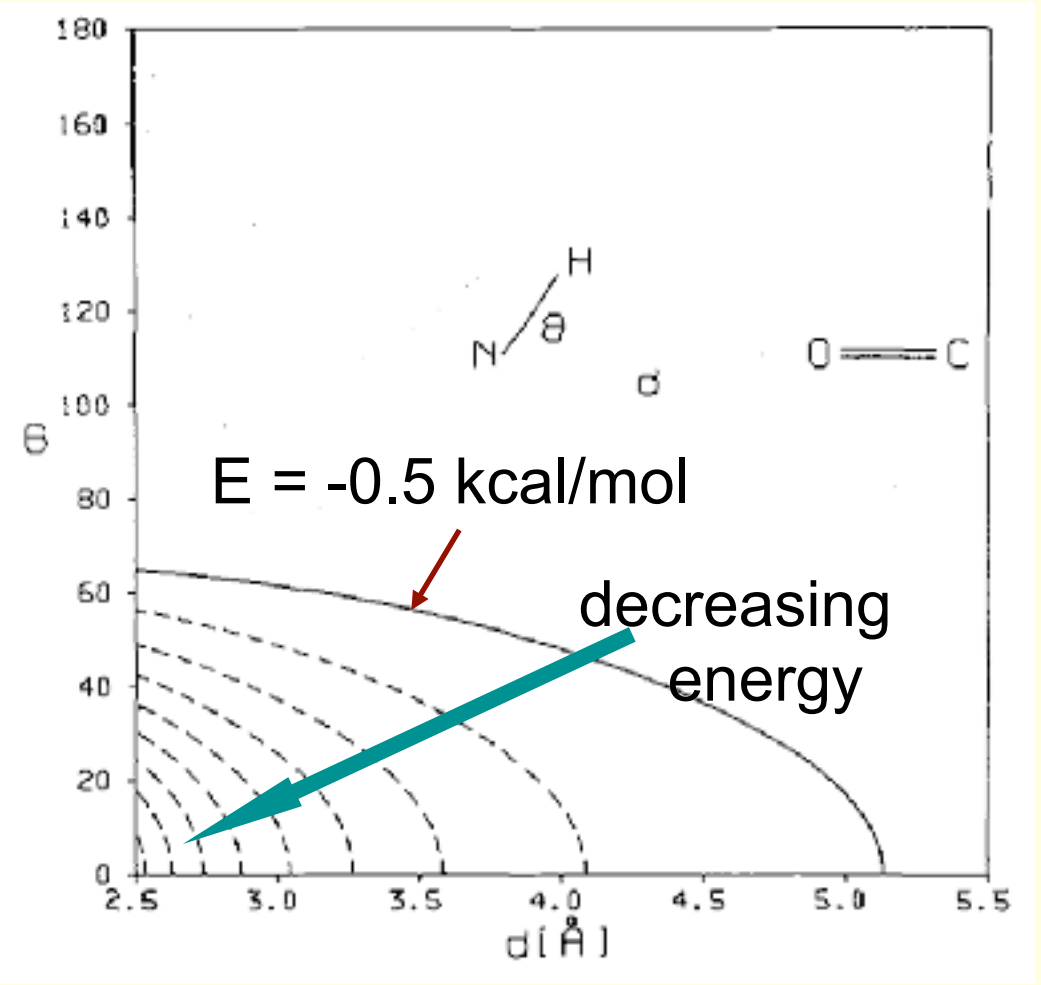
Defines a **Polar Interaction** as a H-bond with computed energy < -0.5 kcal/mol

An ideal H-bond has

d = 2.9 Å

θ = 0

E = -3.0 kcal/mol



DSSP

n-Turn

An *n*-turn is present at residue *i* if there is an H-bond from CO(*i*) to NH(*i*+*n*). Two or more form a helix.

***alpha*-turn**(*i*) = Hbond(*i*,*i*+*4*) (most common)

***3*-turn**(*i*) = Hbond(*i*,*i*+*3*) (3_{10} -helix)

***pi*-turn**(*i*) = Hbond(*i*,*i*+*5*) (very rare)

Bridge

A bridge is formed from two non-overlapping stretches of 3 residues if one of the following H-bond patterns is seen:

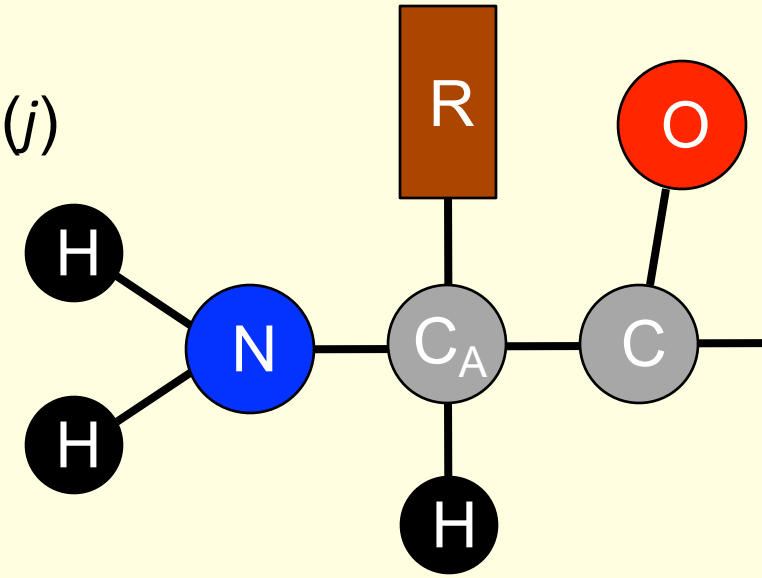
Parallel Bridge (*i*, *j*) = [Hbond(*i*-1,*j*) and Hbond(j,*i*+1)] OR
[Hbond(*j*-1,*i*) and Hbond(*i*,*j*+1)]

AntiParallel Bridge (*i*, *j*) = [Hbond(*i*,*j*) and Hbond(j,*i*)] OR
[Hbond(*i*-1,*j*+1) and Hbond(*j*-1,*i*+1)]

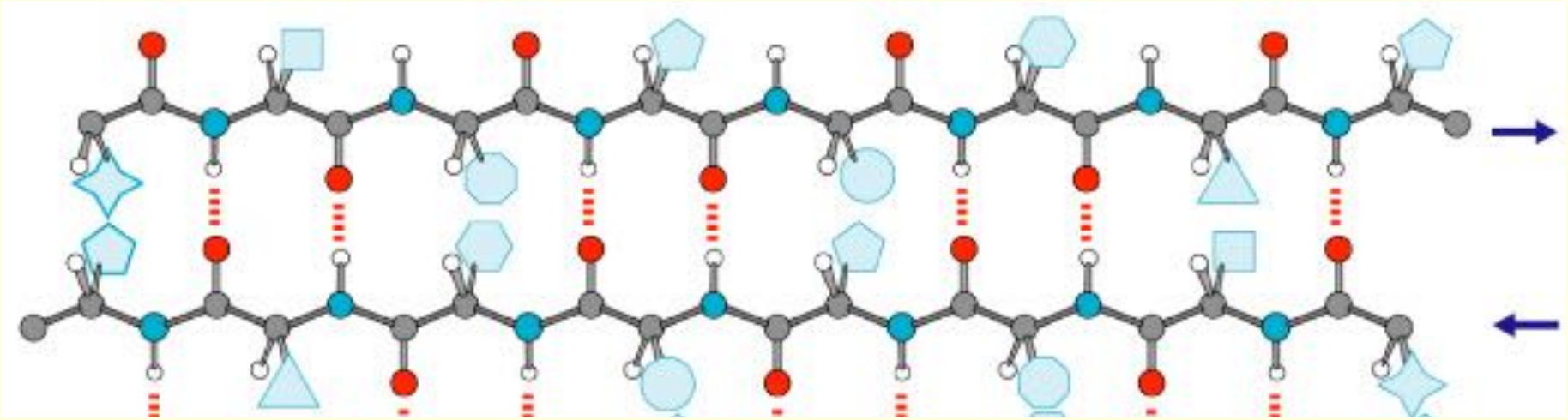
Directionality

Hbond(*i*,*j*) = H-bond from CO(*i*) to NH(*j*)

Proteins written from
N-terminus to C-terminus



i

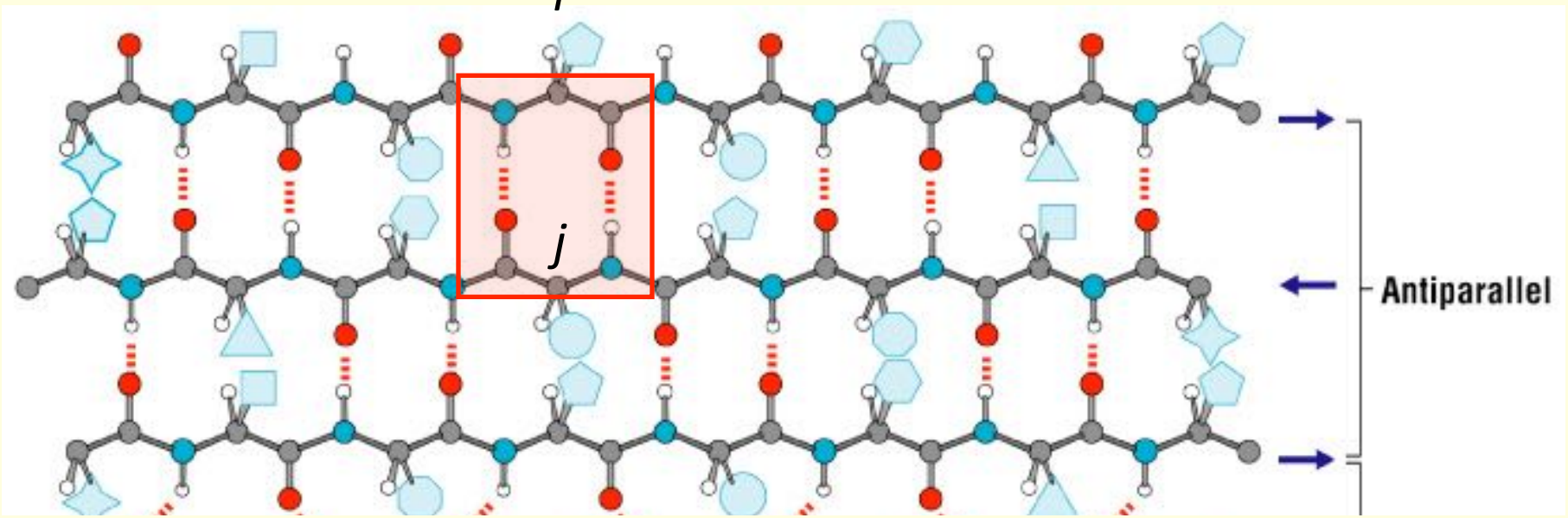


j

DSSP

AntiParallel Bridge (i, j) = [Hbond(i,j) and Hbond(j,i)] OR
[Hbond(i-1,j+1) and Hbond(j-1,i+1)]

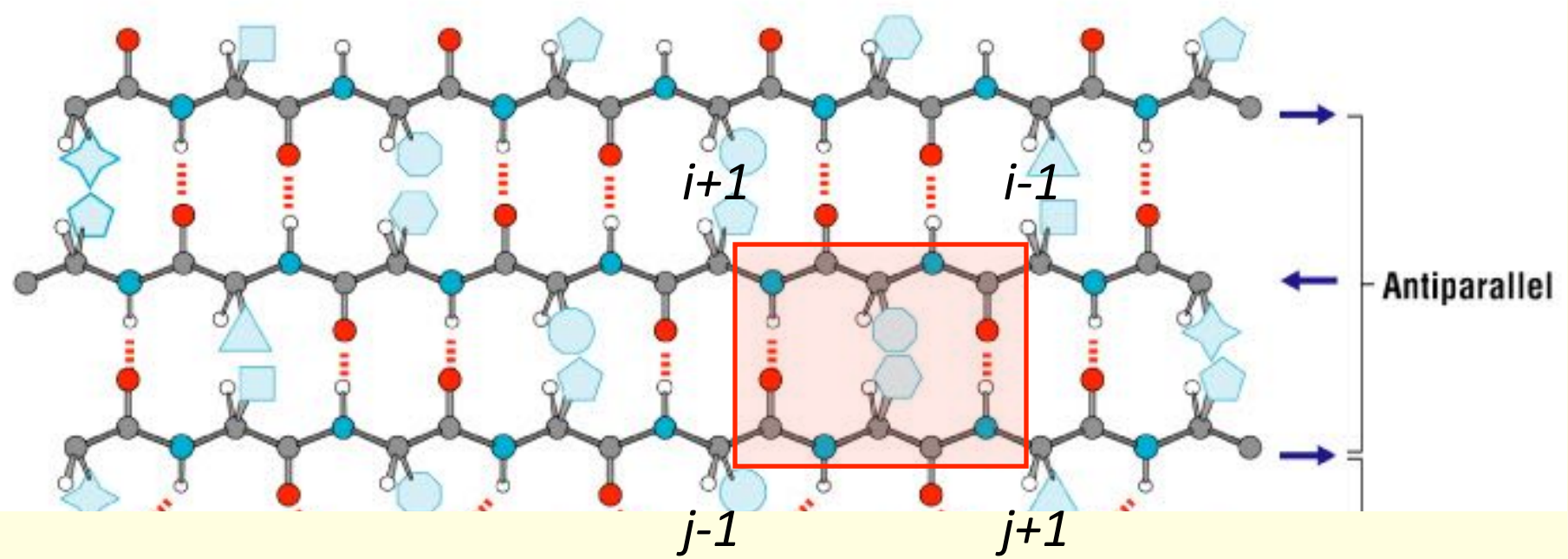
Hbonds $i \rightarrow j$, $j \rightarrow i$



DSSP

AntiParallel Bridge (*i*, *j*) = [Hbond(*i*,*j*) and Hbond(j,*i*)] OR
[Hbond(*i*-1,*j*+1) and Hbond(*j*-1,*i*+1)]

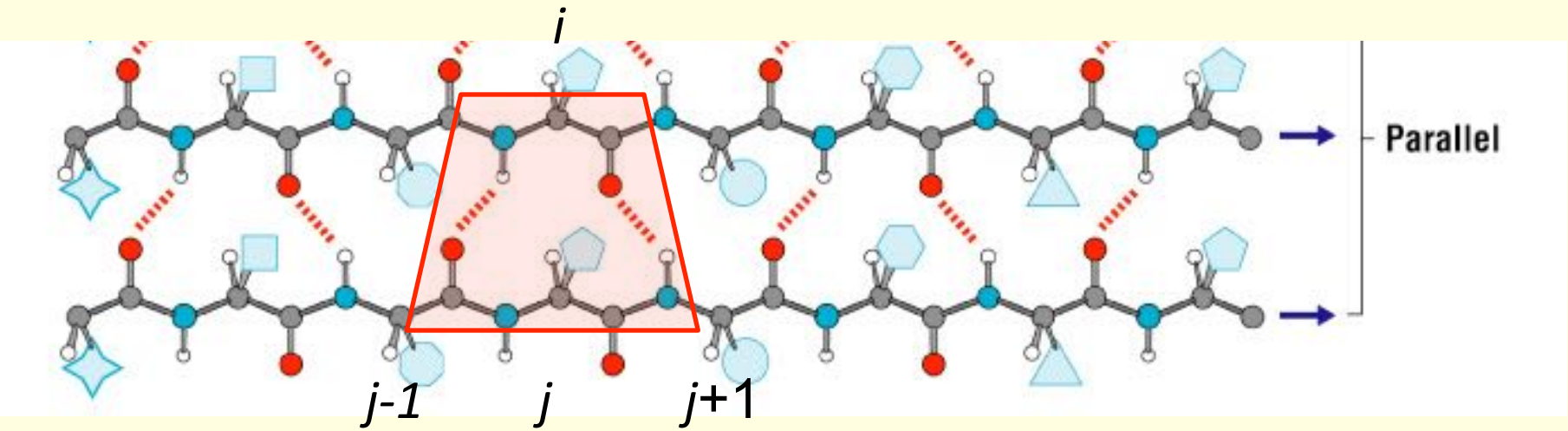
Hbonds *i-1*➡*j+1*, *j-1*➡*i+1*



DSSP

Parallel Bridge (i, j) = [Hbond(i-1,j) and Hbond(j,i+1)] OR
[Hbond(j-1,i) and Hbond(i,j+1)]

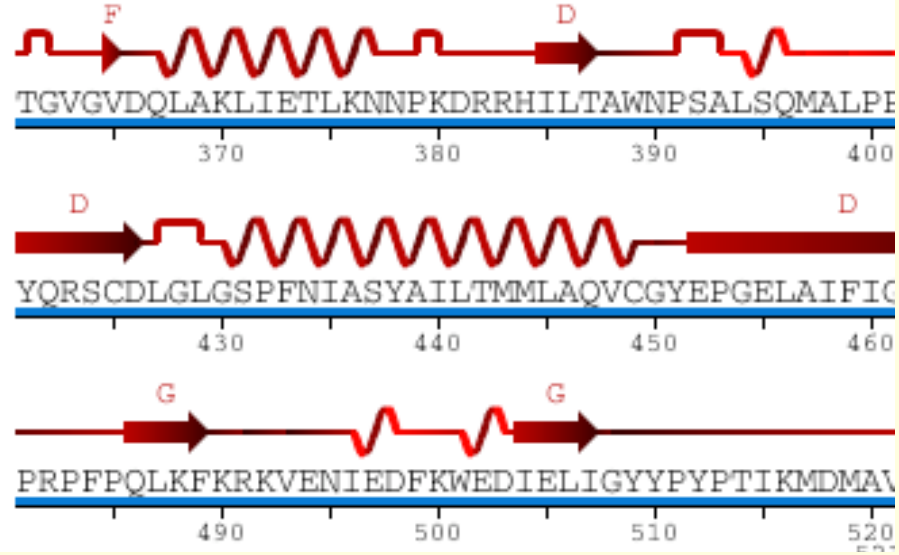
Hbonds *j-1*➡*i*, *i*➡*j*+1



Secondary Structure Prediction

Input: Primary Sequence only

Output: Annotation of alpha-helices, beta-strands, loops



Amino acid	Preference α-helix	β-strand	Reverse turn
Glu	**1.59**	0.52	1.01
Ala	**1.41**	0.72	0.82
Leu	**1.34**	1.22	0.57
Met	**1.30**	1.14	0.52
Gln	**1.27**	0.98	0.84
Lys	**1.23**	0.69	1.07
Arg	**1.21**	0.84	0.90
His	**1.05**	0.80	0.81
Val	0.90	**1.87**	0.41
Ile	1.09	**1.67**	0.47
Tyr	0.74	**1.45**	0.76
Cys	0.66	**1.40**	0.54
Trp	1.02	**1.35**	0.65
Phe	1.16	**1.33**	0.59
Thr	0.76	**1.17**	0.90
Gly	0.43	0.58	**1.77**
Asn	0.76	0.48	**1.34**
Pro	0.34	0.31	**1.32**
Ser	0.57	0.96	**1.22**
Asp	0.99	0.39	**1.24**

Petsko, Ringe, Prot Struct and Function, 2004

Chou-Fasman (First-Generation)

Compute propensities for each amino acid, a_i in structural conformation $s_j \in \{\alpha, \beta, \rho\}$

$$\frac{\Pr[A = a_i \mid S = s_j]}{\Pr[A = a_i]}$$

Categorize each AA as *helix-former*, *helix-breaker*, *helix-indifferent* and *sheet-former*, *sheet-breaker*, *sheet-indifferent*.

Keep chaining residues to SS element while average propensity is above some threshold.

Accuracy?

3-State Accuracy: Percent of residues for which a method's predicted secondary structure (Helix (α), Strand (β), Neither (ρ)) is correct.

Chou-Fasman: 50-60% 3-State Accuracy

Priors:

30% Helices
20% Strands
50% Neither

If you always predicted 'Neither' you would have 50% 3-State Accuracy

Need to incorporate additional information!

GOR Method

What if we look at “blocks” of sequence to determine secondary structure? (Garnier-Osguthorpe-Robson, late 70s)

$$s_j = f(r_{j-k}, \ldots, r_{j+k})$$

GAVLIFYWMVLLAGIFFST

GOR Method

Main Idea: Look at the mutual information between nearby amino acids to assess whether they influence secondary structure at a particular position.

$$
\begin{aligned}
I(x;y) &= \sum_x \sum_y \Pr[x,y] \log \frac{\Pr[x,y]}{\Pr[x]\Pr[y]} \\
&= \sum_x \sum_y \Pr[x,y] \log \frac{\Pr[x \mid y]\Pr[y]}{\Pr[x]\Pr[y]} \\
&= \sum_x \sum_y \Pr[x,y] \log \frac{\Pr[x \mid y]}{\Pr[x]}
\end{aligned}
$$

GOR Method

Consider mutual information between structural class and sequence:

$$I(s_j; r_{j-8}, \ldots, r_{j+8}) = \sum_{s_j} \sum_{\langle r_{j-8}, \ldots, r_{j+8} \rangle} \Pr[s_j, r_{j-8}, \ldots, r_{j+8}] \log \frac{\Pr[s_j \mid r_{j-8}, \ldots, r_{j+8}]}{\Pr[s_j]}$$

Select the structural class which maximizes:

$$I(s_j = x; r_{j-8}, \ldots, r_{j+8}) - I(s_j \neq x; r_{j-8}, \ldots, r_{j+8})$$

Problem: Each conditional probability needs to be computed.

$\Pr[s_j \mid r_{j-8}, \ldots, r_{j+8}]$ table size: 3×20^{17}

Assume 'independence':

$$I(s_j; r_{j-8}, \ldots, r_{j+8}) = \sum_{k=-8}^{8} I(s_j = x; r_{j+k})$$

GOR Method Extension

Larger training set (513 vs 267 sequences)
Because 'Neither' regions were over predicted, they added a margin by which Neither must be preferred over Helix or Coil
Use of doublet and triplet statistics
Optimized window length - 13 better than 17...

Previously:

$$I(s_j; r_{j-8}, \ldots, r_{j+8}) = \sum_{k=-8}^{8} I(s_j = x; r_{j+k})$$

Incorporating conditional information on residue r_j :

$$I(s_j; r_{j-8}, \ldots, r_{j+8}) = I(s_j = x; r_j) + \sum_{k=-8, k\neq 0}^{8} I(s_j = x; r_{j+k} \mid r_j)$$

Other “Learning” Methods

Neural Networks

17 Residue Window

Each Residue represented with 21-bits

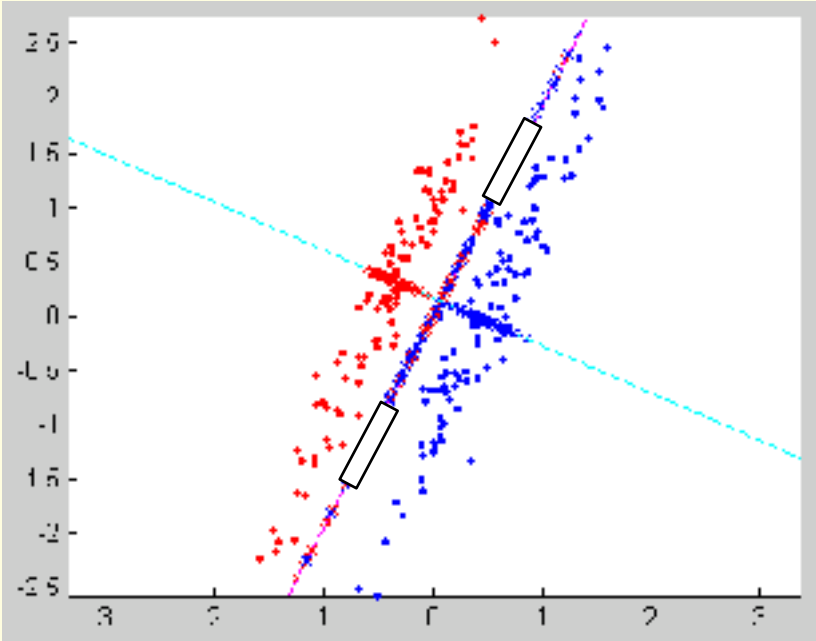
One for each AA, one for beginning/end of sequence

Each sample has 17x21 bit feature vector with 17 non-zero entries

Support Vector Machines (SVM)

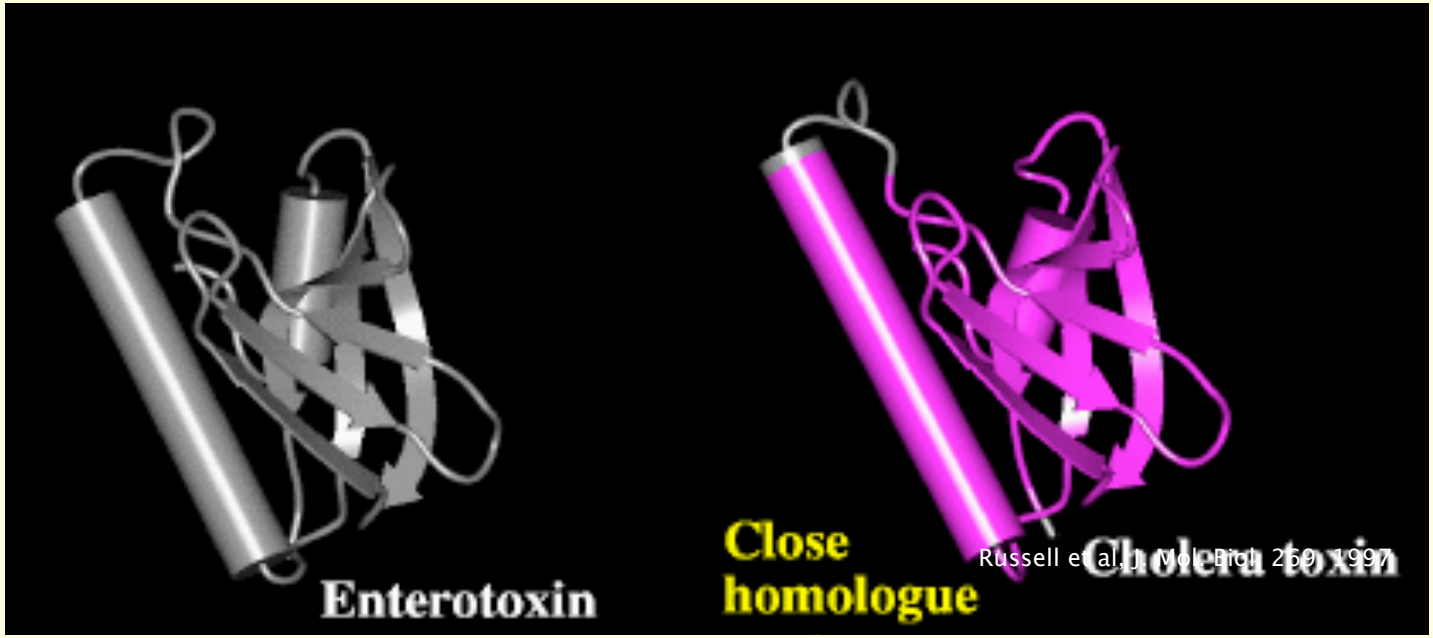
Linear Discriminant Analysis (LDA)

- Returns H, E, or C with confidence or probability.
- Requires k sequential residues with confidence above some threshold.



Third-Generation

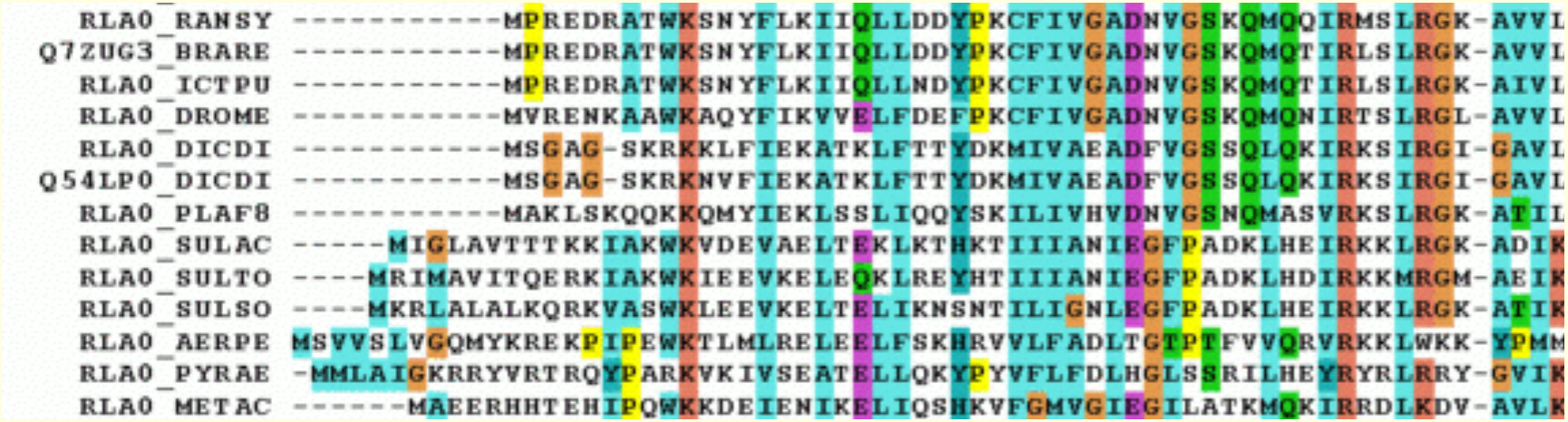
Observation: Protein structure is more conserved than protein sequence.



Two proteins with >30% sequence identity are likely to have similar structures.

Multiple Sequence Alignment

We saw that multiple primary sequences of DNA, RNA, or protein identify similarity that may be a consequence of functional, structural, or evolutionary relationships.



Idea:

Predict secondary structure type by using information from homologs, then combine predictions.

Multiple Sequence Alignment

Have:

List of similar (but different) sequences

Assumption:

These aligned sequences fold into the same structure

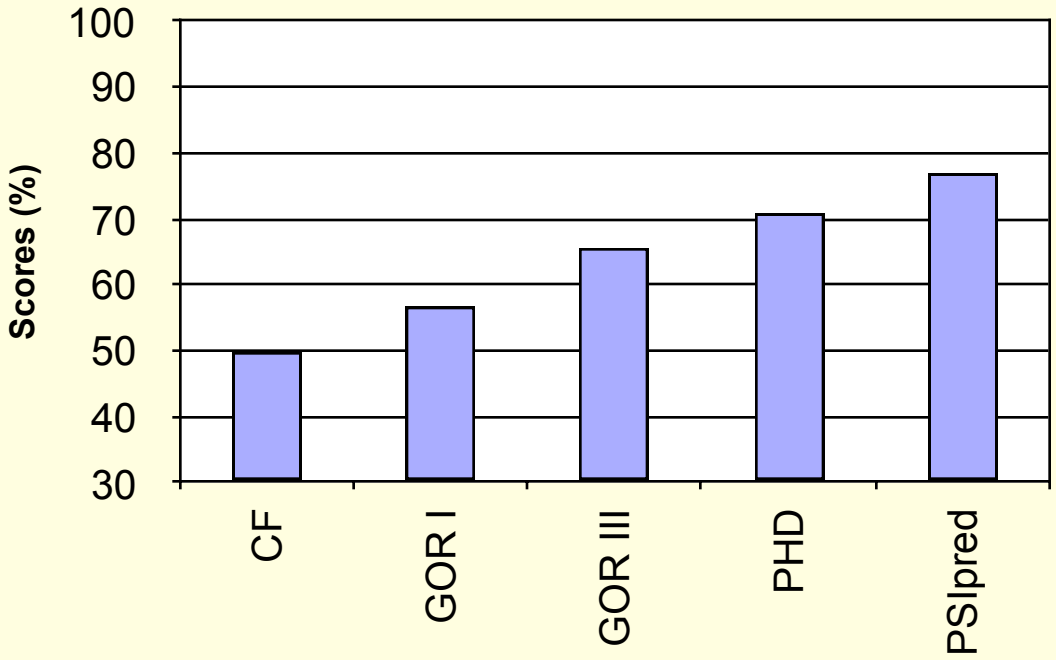
Consider the probability distribution over all amino acids at residue i, a 20-dimensional vector

Examine 13-residue window

Each residue represented by the 20-dimensional probability vector

Use favorite classification technique (NN, LDA, SVM)

Accuracy Limits



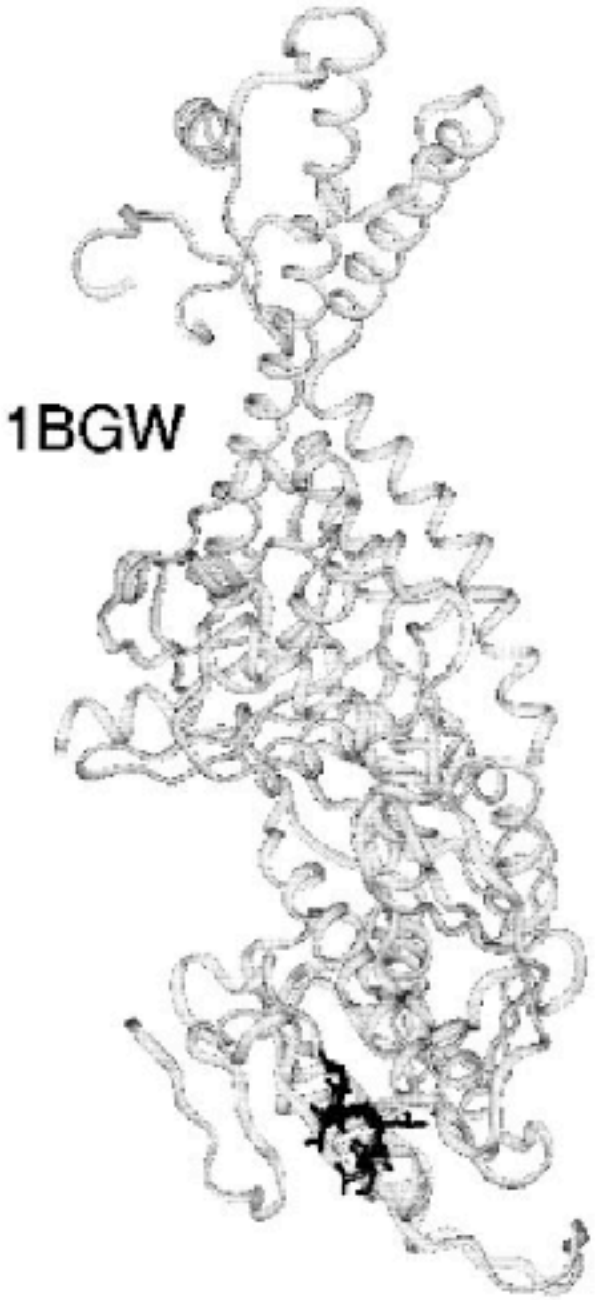
What is the ***gold standard***? (~12% variability)

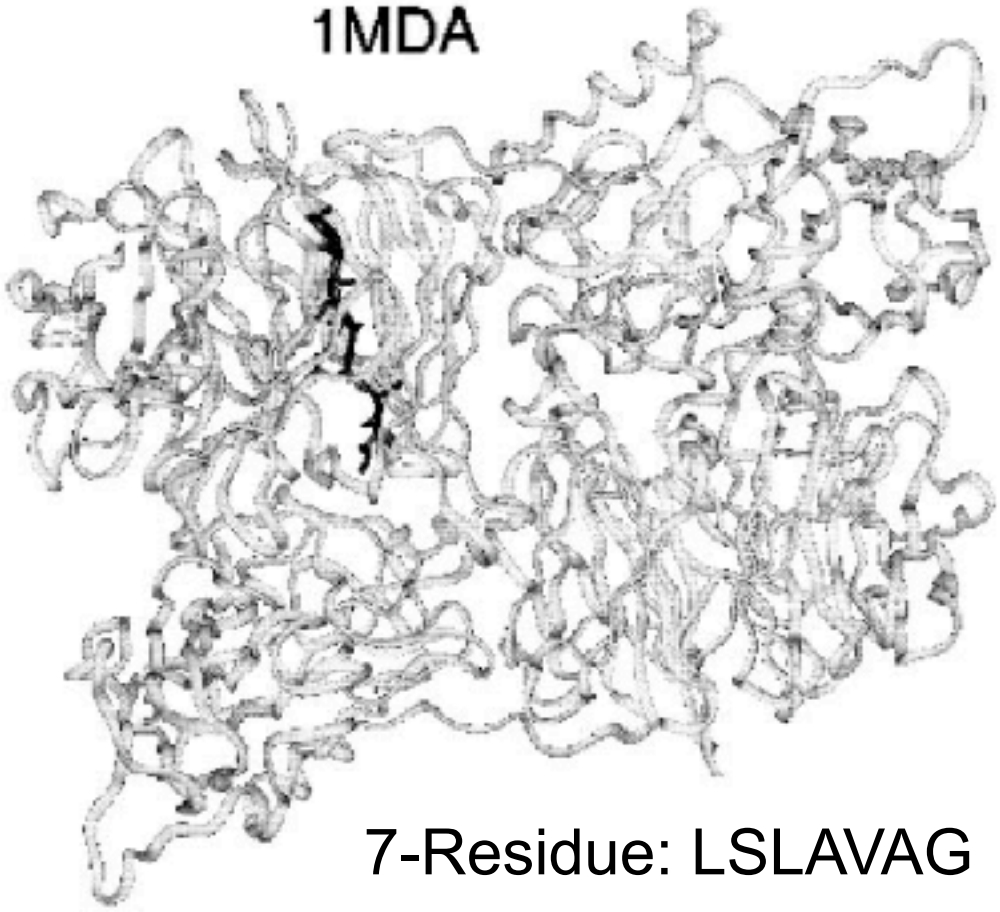
Puts upper bound at 88% accuracy

Sequence ➡ Structure Degeneracy

Some 11-long AA sequences can fold into both an alpha-helix and a beta-strand

Chameleon Sequences



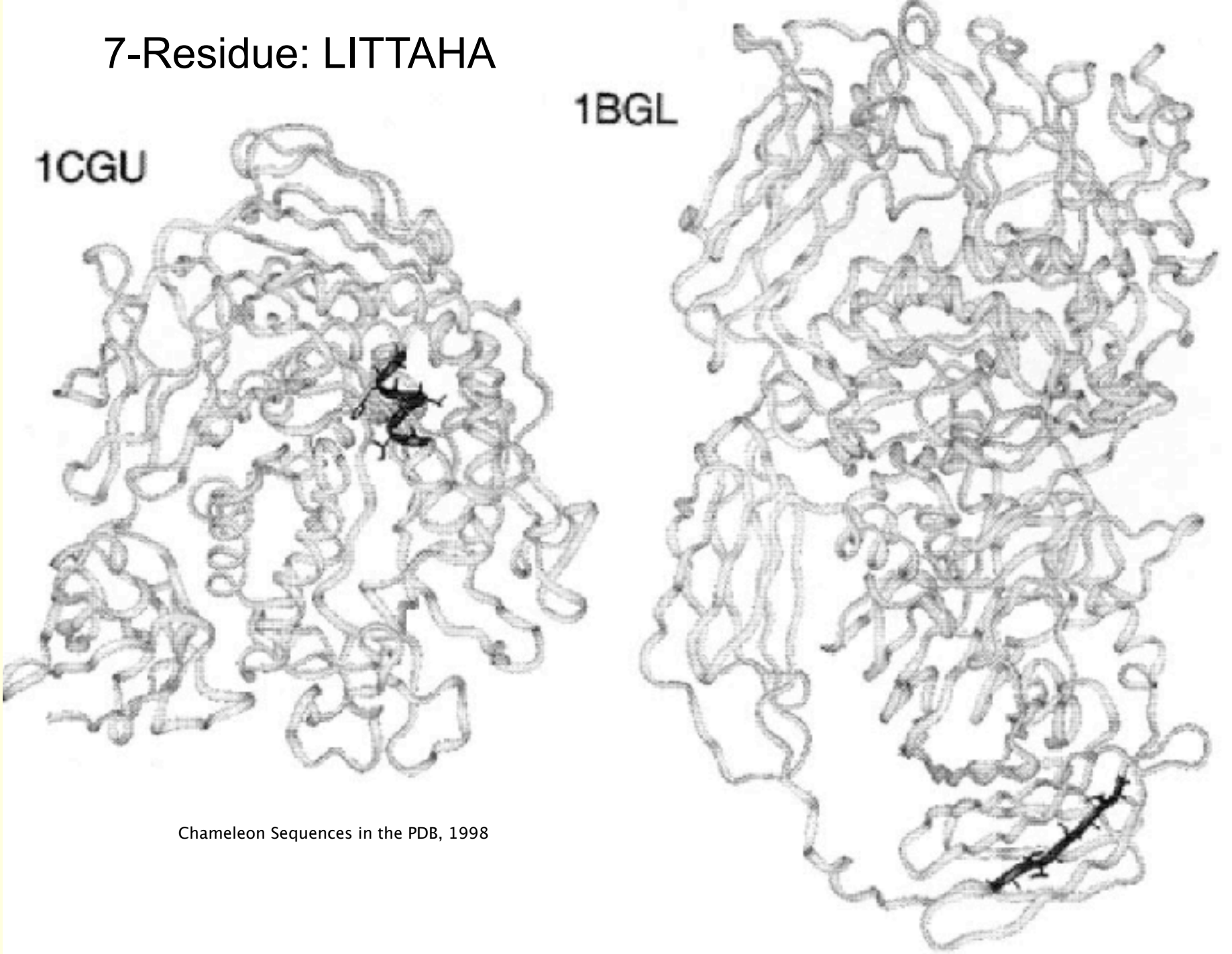


7-Residue: LSLAVAG

Many 5-residue chameleons
Fewer 6-residue chameleons
Few 7-residue chameleons

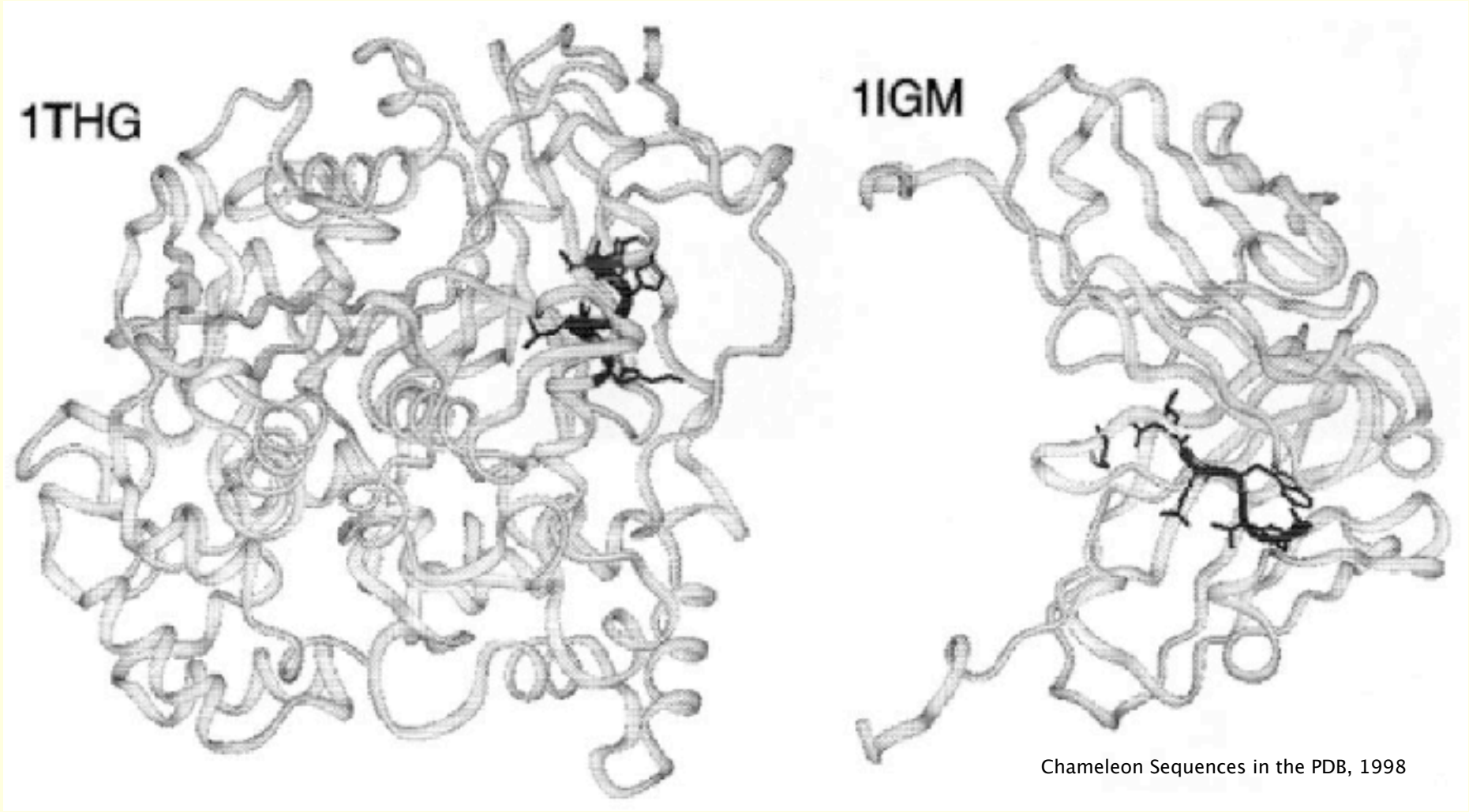
Chameleon Sequences in the PDB, 1998

7-Residue: LITTAHA



Chameleon Sequences in the PDB, 1998

7-Residue: KGLEWVS



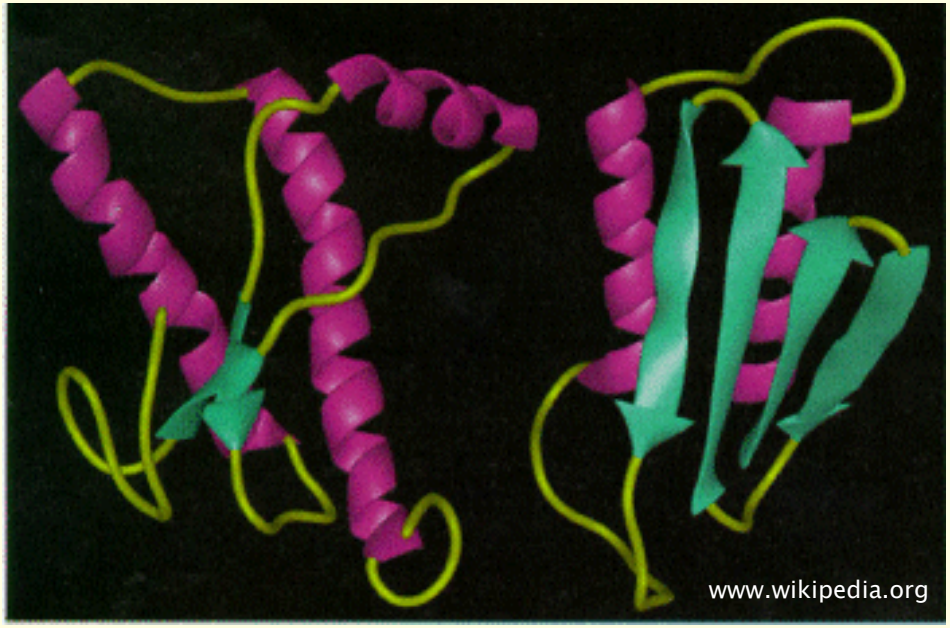
Chameleon Sequences in the PDB, 1998

Prions

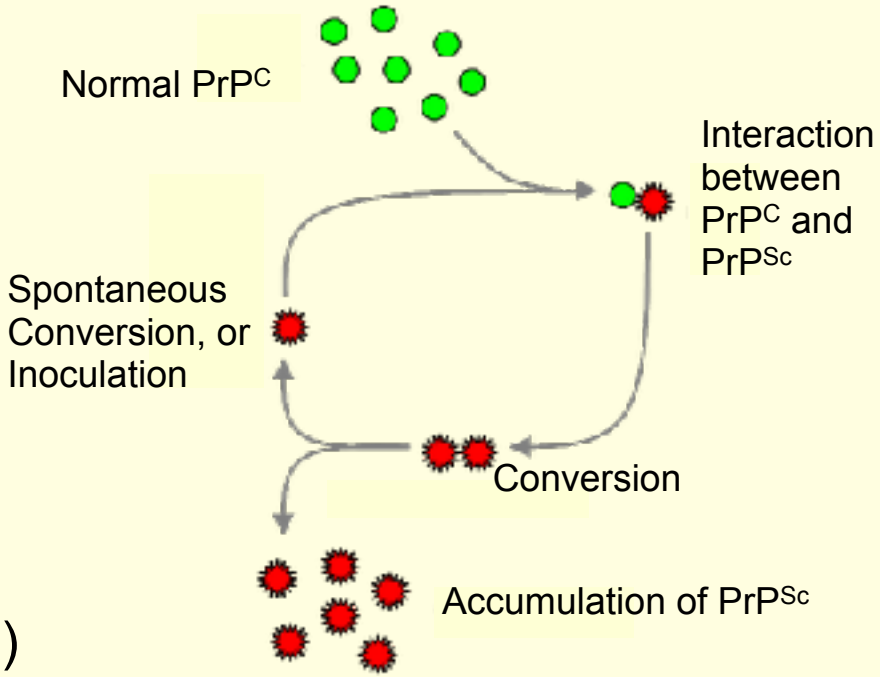
Prion: from “proteinaceous infectious”
The prion protein is naturally found in the body.
The infectious agent is the same protein but with a different fold.

Disease fold induces normal copies of the protein to fold into disease form.

Accumulation of disease folds forms cytotoxic aggregates.



PrP^{C} (NMR) PrP^{Sc} (Proposed)



Prions

Prion: from “proteinaceous infectious”
The prion protein is naturally found in the body.
The infectious agent is the same protein but with a different fold.

Disease fold induces normal copies of the protein to fold into disease form.

Accumulation of disease folds forms cytotoxic aggregates.

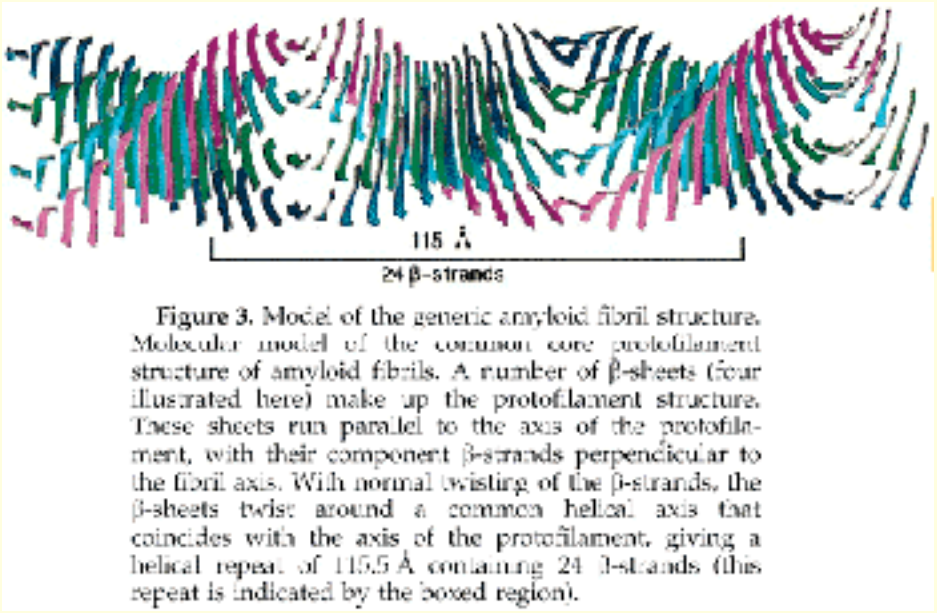
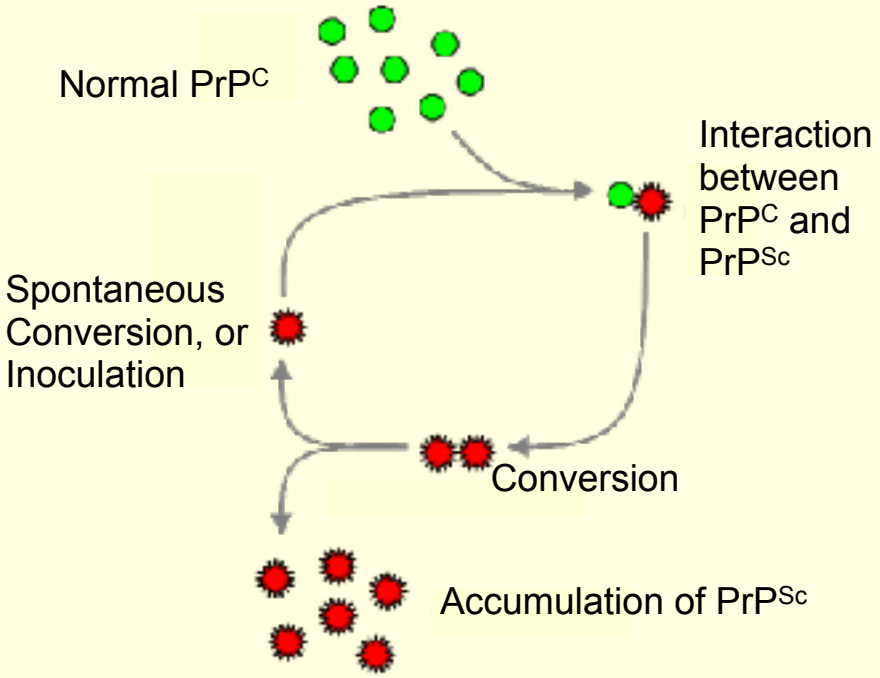


Figure 3. Model of the generic amyloid fibril structure. Molecular model of the common core protofilament structure of amyloid fibrils. A number of β-sheets (four illustrated here) make up the protofilament structure. These sheets run parallel to the axis of the protofilament, with their component β-strands perpendicular to the fibril axis. With normal twisting of the β-strands, the β-sheets twist around a common helical axis that coincides with the axis of the protofilament, giving a helical repeat of 115.5 Å containing 24 β-strands (this repeat is indicated by the boxed region).

Sunde et al, J Mol Biol, 273(3), 1997.



“Amyloidosis” leads to tissue damage.