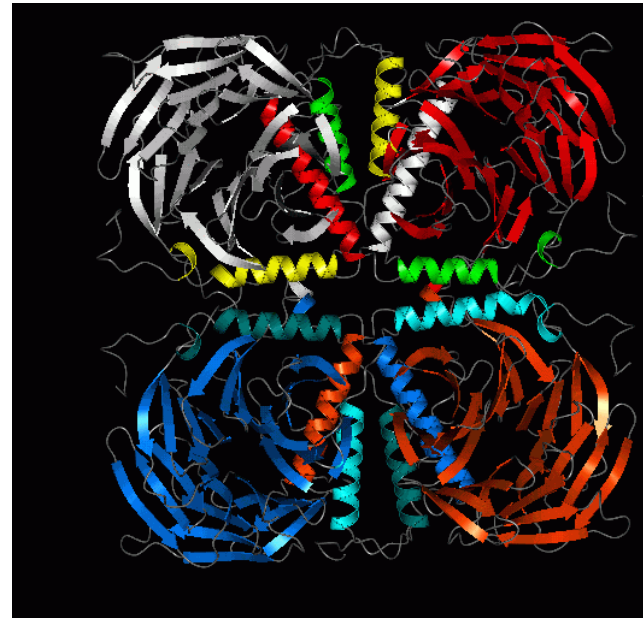
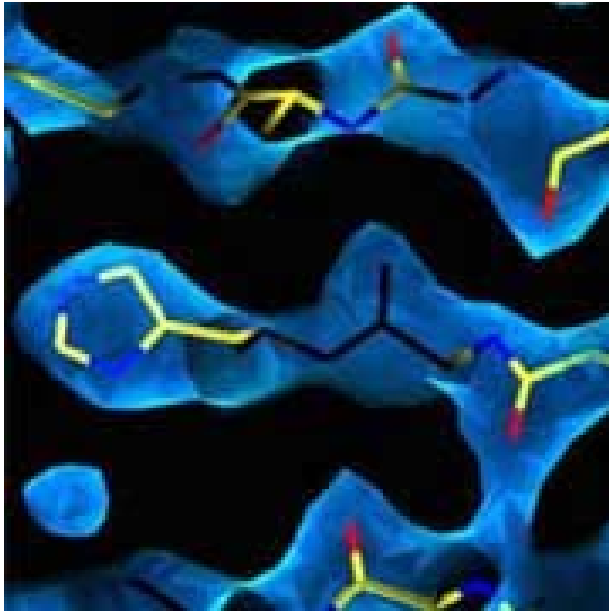


# 7.91 – Lecture #6 Michael Yaffe

## Protein Secondary Structure Prediction



Scansite Menu - Netscape


File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

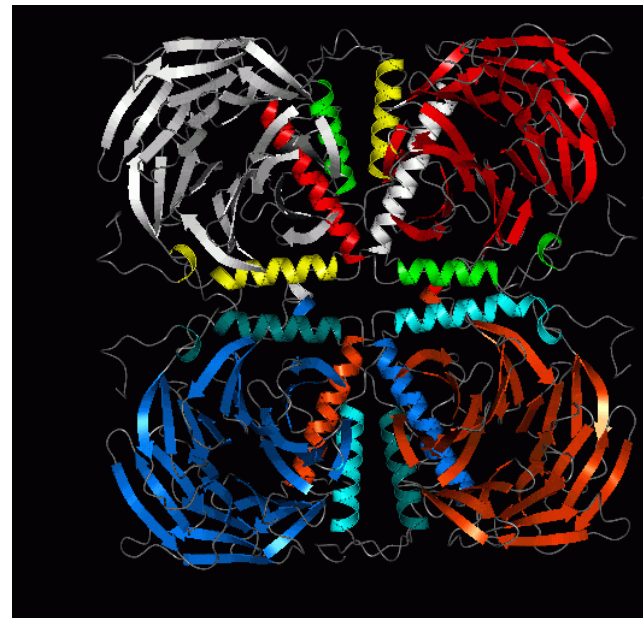
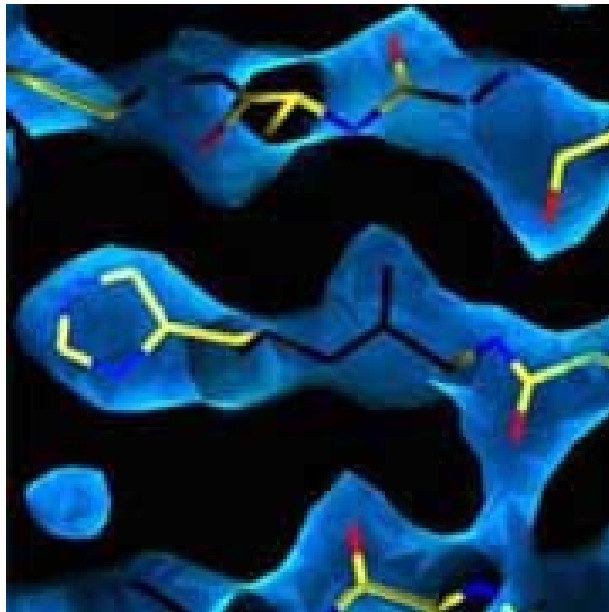
Bookmarks Location: <http://scansite.mit.edu/>

MIT Home Page MIT Certificate MIT Directory Yellow Pages MIT Search Channels

# Scansite

 **News:** Nov. 5, 2001 - Genpept, SwissProt, and TrEMBL databases have been updated in Scansite's Database Search.

Massachusetts Institute of Technology [▶ About SCANSITE](#)



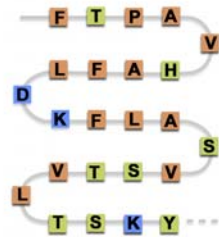
# Outline

- Brief review of protein structure
- Chou-Fasman predictions
- Garnier, Osguthorpe and Robson
- Helical wheels and hydrophobic moments
- Neural networks
- Nearest neighbor methods
- Consensus prediction approaches

# Hierarchy of protein structure

## Hierarchy of Protein Structure

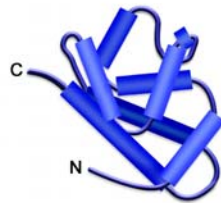
Primary



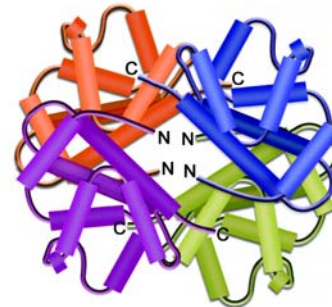
Secondary



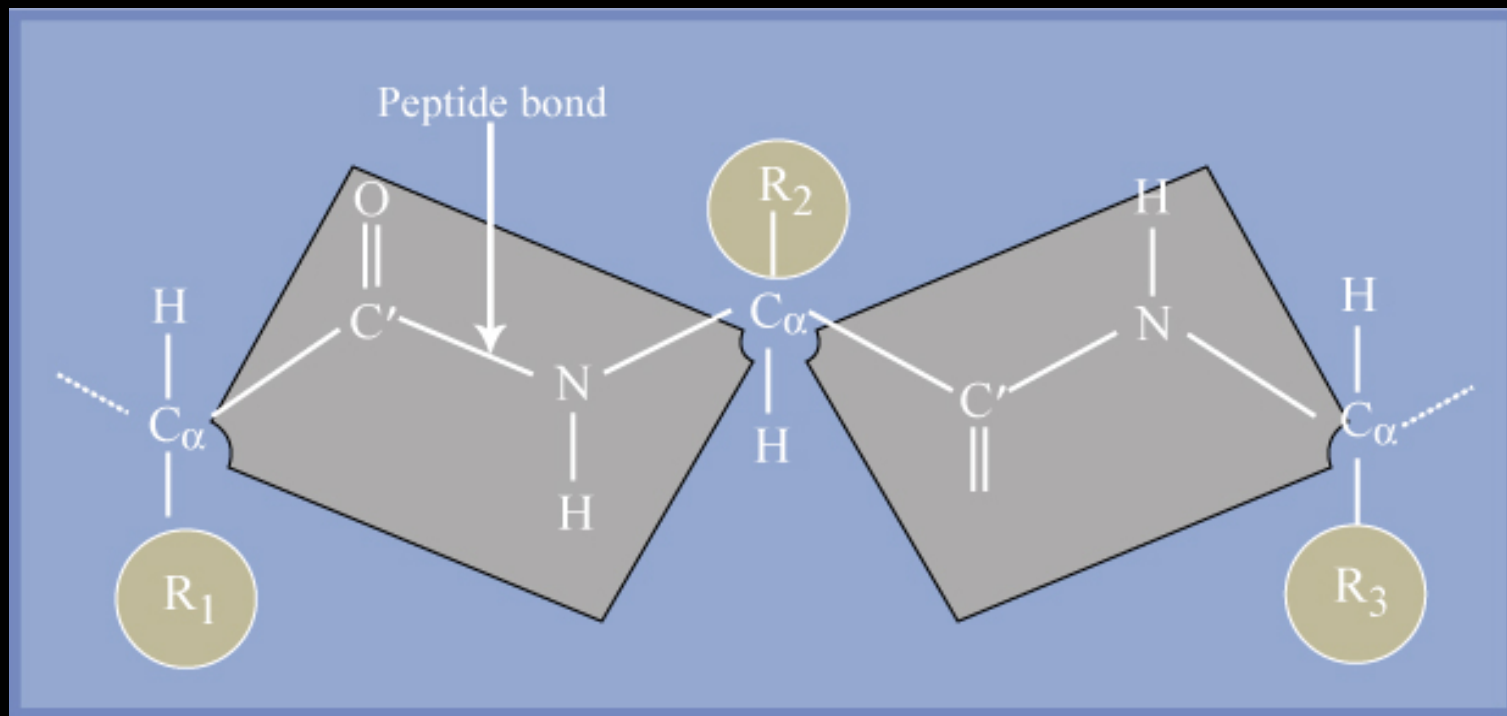
Tertiary



Quaternary



# Resonance of peptide bond implies planarity



# Dihedral angles define secondary structure

Please refer to Branden, Carl, and John Tooze. *Introduction to Protein Structure*.  
2nd ed. Garland Publishing, Inc., 1999. ISBN: 0815323042.

# Structure of $\alpha$ -helices

Please refer to Branden, Carl, and John Tooze. *Introduction to Protein Structure*. 2nd ed. Garland Publishing, Inc., 1999. ISBN: 0815323042.

# $\alpha$ -helix dipole moment

Please refer to Branden, Carl, and John Tooze. *Introduction to Protein Structure*. 2nd ed. Garland Publishing, Inc., 1999. ISBN: 0815323042.



# Anti-parallel $\beta$ -sheets

Please refer to Branden, Carl, and John Tooze. *Introduction to Protein Structure*. 2nd ed. Garland Publishing, Inc., 1999. ISBN: 0815323042.

# The “pleat” - a function of the tetrahedral C $\alpha$ carbon

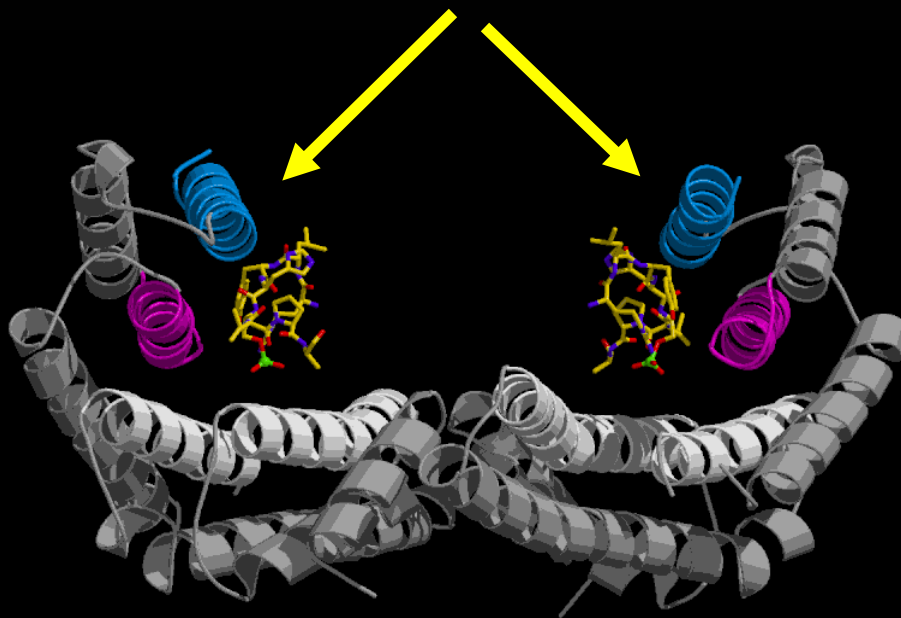
Please refer to Branden, Carl, and John Tooze. *Introduction to Protein Structure*. 2nd ed. Garland Publishing, Inc., 1999. ISBN: 0815323042.

# The parallel $\beta$ -sheet

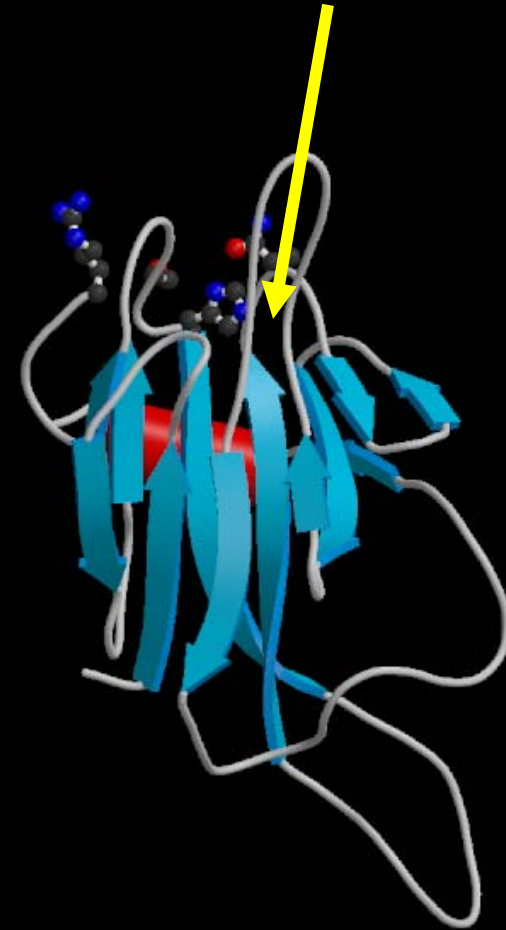
Please refer to Branden, Carl, and John Tooze. *Introduction to Protein Structure*. 2nd ed. Garland Publishing, Inc., 1999. ISBN: 0815323042.

# Protein Classes – defined by secondary structural elements

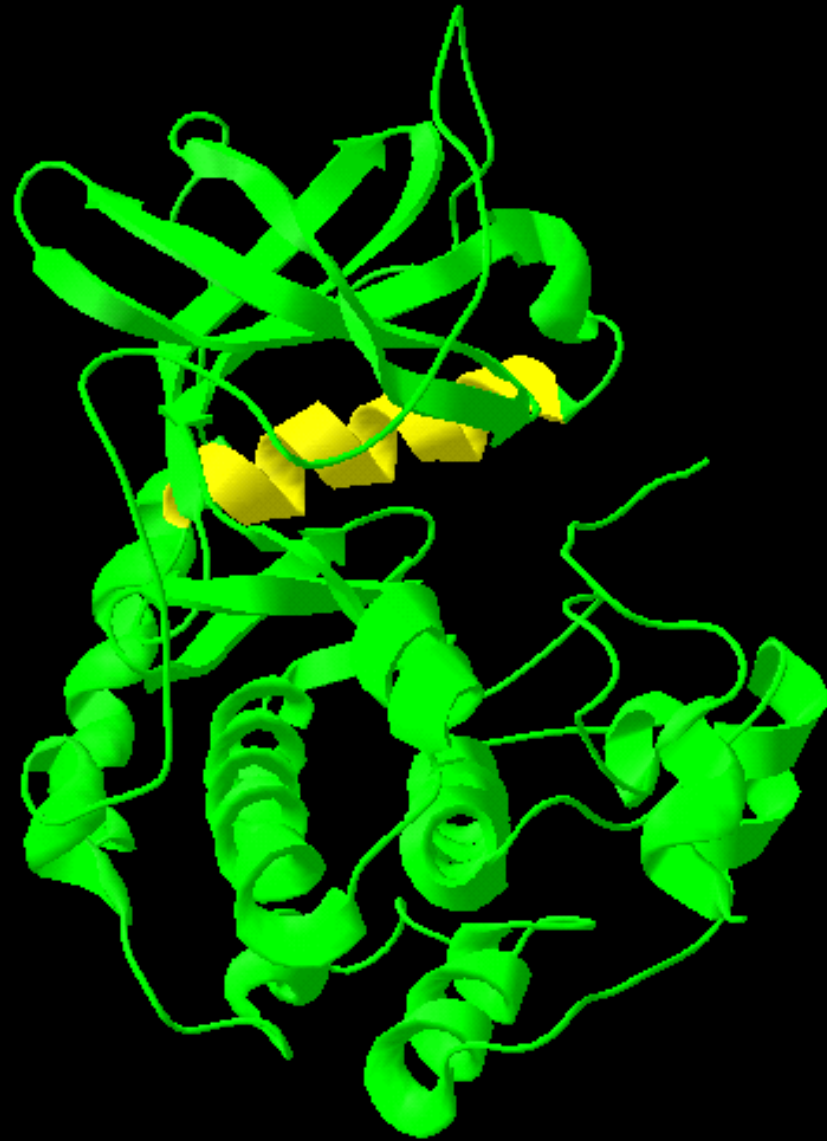
**All  $\alpha$ -helical**



**All  $\beta$ -sheet**



$\alpha/\beta$ -protein



# Chou-Fasman

Biochemistry, 13: 222-245, 1974

---

- **Statistical Method**

- **Based on 15 proteins of known conformation, 2473 total amino acids**

- **Determined “protein conformational parameters”**  
 $P_{\alpha}$ ,  $P_{\beta}$ , based on  $f_i^s / (\sum f_j^s / 20) \rightarrow 0.5-1.5$

**Helical residues**

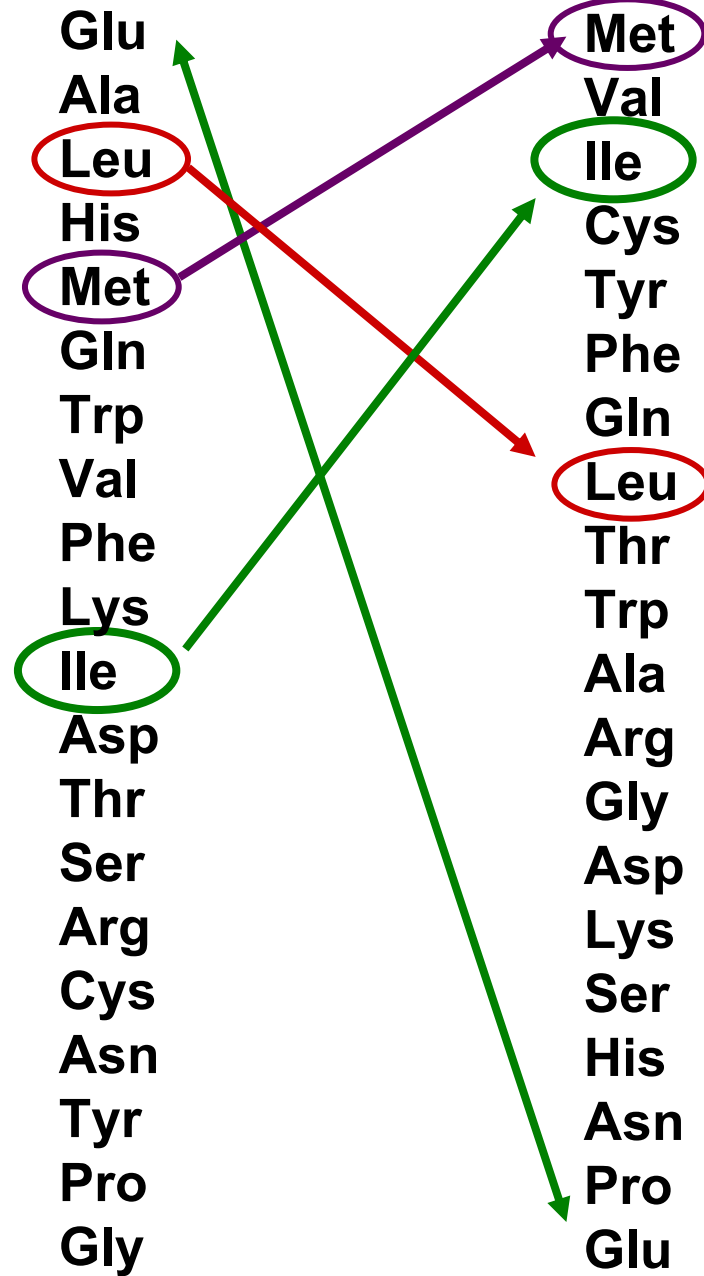
	$P_{\alpha}$		
<b>Glu</b>	<b>1.53</b>	}	<b>H<math>\alpha</math></b> <b>Strong helix former</b>
<b>Ala</b>	<b>1.45</b>		
<b>Leu</b>	<b>1.34</b>		
<b>His</b>	<b>1.24</b>	}	<b>h<math>\alpha</math></b> <b>Helix former</b>
<b>Met</b>	<b>1.20</b>		
<b>Gln</b>	<b>1.17</b>		
<b>Trp</b>	<b>1.14</b>		
<b>Val</b>	<b>1.14</b>		
<b>Phe</b>	<b>1.12</b>	}	<b>l<math>\alpha</math></b> <b>Weak helix former</b>
<b>Lys</b>	<b>1.07</b>		
<b>Ile</b>	<b>1.00</b>	}	<b>i<math>\alpha</math></b> <b>Helix indifferent</b>
<b>Asp</b>	<b>0.98</b>		
<b>Thr</b>	<b>0.82</b>		
<b>Ser</b>	<b>0.79</b>		
<b>Arg</b>	<b>0.79</b>	}	<b>b<math>\alpha</math></b> <b>Helix breaker</b>
<b>Cys</b>	<b>0.77</b>		
<b>Asn</b>	<b>0.73</b>	}	<b>B<math>\alpha</math></b> <b>Strong helix breaker</b>
<b>Tyr</b>	<b>0.61</b>		
<b>Pro</b>	<b>0.59</b>		
<b>Gly</b>	<b>0.53</b>		

<b><math>\beta</math>-Sheet residues</b>		$P_{\beta}$		
Met	1.67	}	$H_{\beta}$	Strong sheet former
Val	1.65			
Ile	1.60			
Cys	1.30	}	$h_{\beta}$	Sheet former
Tyr	1.29			
Phe	1.28			
Gln	1.23			
Leu	1.22			
Thr	1.20			
Trp	1.19	}	$l_{\beta}$	Weak sheet former
Ala	0.97			
Arg	0.90	}	$i_{\alpha}$	Sheet indifferent
Gly	0.81			
Asp	0.80			
Lys	0.74	}	$b_{\beta}$	Sheet breaker
Ser	0.73			
His	0.71			
Asn	0.65			
Pro	0.62	}	$B_{\beta}$	Strong sheet breaker
Glu	0.26			



$\alpha$ -helical

$\beta$ -sheet



# Chou-Fasman

## *Empirical rule set for secondary structure nucleation using $\langle P_\alpha \rangle$ , $\langle P_\beta \rangle$*

- Search for helical nuclei: locate clusters of 4 ( $H_\alpha$  or  $h_\alpha$ ) out of 6 residues. Unfavorable if  $> 1/3$  ( $b_\alpha$  or  $B_\alpha$ ).
- Extend helical segments in both directions until terminated by tetrapeptides with  $\langle P_\alpha \rangle < 1.0$ . Helix breakers include  $b_4$ ,  $b_{3i}$ , etc. Some of the tetrapeptide residues can be in the helical ends (except Pro).
- Refine boundaries: Pro, Asp, Glu prefer N-terminal end, His Lys, Arg prefer C-terminal end.
- **Rule #1 – Any segment  $\geq 6$  residues with  $\langle P_\alpha \rangle \geq 1.03$  and  $\langle P_\alpha \rangle > \langle P_\beta \rangle$ , satisfying above conditions is predicted as helical.**

# Chou-Fasman

## *Empirical rule set for secondary structure*

### *nucleation using $\langle P_\alpha \rangle$ , $\langle P_\beta \rangle$*

- Search for  $\beta$ -sheet nuclei: locate clusters of 3  $\beta$  residues ( $H_\beta$  or  $h_\beta$ ) out of 5 residues. Unfavorable if  $> 1/3$   $\beta$  breakers ( $b_\beta$  or  $B_\beta$ ).
- Extend  $\beta$ -sheet segments in both directions until terminated by tetrapeptides with  $\langle P_\beta \rangle < 1.0$ .  $\beta$ -sheet breakers include  $b_4$ ,  $b_{3i}$ , etc.
- Refine boundaries: Glu occurs rarely in  $\beta$ -region and Pro equally uncommon within inner  $\beta$ -sheets. Charged residues rare at either end. Trp most frequently at N-terminal end
- **Rule #2 – Any segment  $\geq 5$  residues with  $\langle P_\beta \rangle \geq 1.05$  and  $\langle P_\beta \rangle > \langle P_\alpha \rangle$ , satisfying above conditions is predicted as  $\beta$ -sheet.**

# Chou-Fasman

## *Results*

- ~50-60% accurate in reality, though paper claimed much higher results (limited data set)
- Seemed to be particularly less accurate for  $\beta$ -sheets.

# Chou-Fasman

## *$\beta$ -Turn potentials*

- Typical  $\beta$ -turn is 4 amino acids

	$f_i$	$f_{i+1}$	$f_{i+2}$	$f_{i+3}$
Arg	0.051	0.127	0.025	0.101
Asn	0.101	0.086	0.216	0.065
Asp	0.137	0.088	0.069	0.059
Pro	0.074	0.272	0.012	0.062
Trp	0.045	0.000	0.045	0.205

$$\langle f_j \rangle = \sum j/N = 65/2343 = 0.07$$

# Chou-Fasman

## *$\beta$ -Turn potentials*

- Typical  $\beta$ -turn is 4 amino acids

	$f_i$	$f_{i+1}$	$f_{i+2}$	$f_{i+3}$
Arg	0.051	<b>0.127</b>	0.025	<b>0.101</b>
Asn	<b>0.101</b>	<b>0.086</b>	<b>0.216</b>	0.065
Asp	<b>0.137</b>	<b>0.088</b>	0.069	0.059
Pro	<b>0.074</b>	<b>0.272</b>	0.012	0.062
Trp	0.045	0.000	0.045	<b>0.205</b>

$$\langle f_j \rangle = \sum j/N = 165/2343 = 0.07$$

$$P(t) = f_i f_{i+1} f_{i+2} f_{i+3} \quad P(t) > 7.5 \times 10^{-5} \rightarrow \text{turn}$$

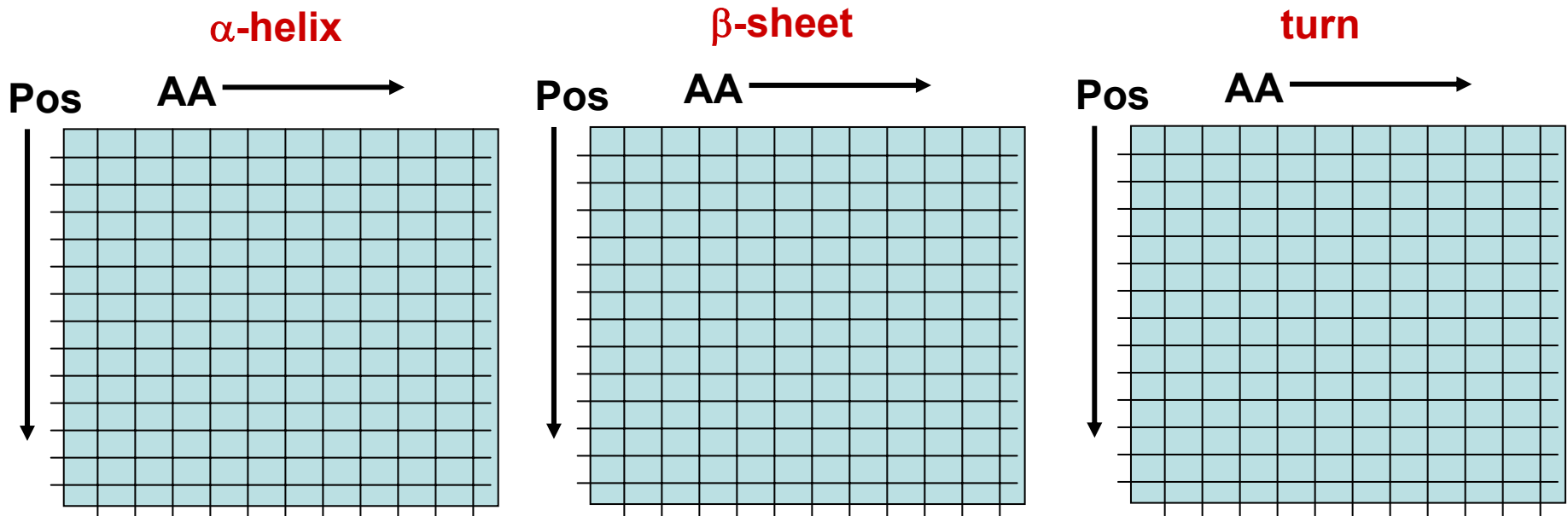
# Garnier, Osguthorpe, Robson

- Alternative approach to Chou-Fasman.
- Original version called “GOR”. Now up to GOR-3. Uses a scanning window of 17 amino acids centered on residue being examined.
- Based on assumption that each amino acid individually influences the propensity of the central residue to adopt a particular secondary structure.
- Each flanking position evaluated independently ...like a PSSM!

# GOR Scoring Tables (original)

3 states –  $\alpha$ -helix,  $\beta$ -sheet, turn

-8	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7	+8
W	R	Q	I	C	T	V	N	A	F	L	C	E	H	S	Y	K



Note – each table is INDEPENDENT of the central amino acid!



# GOR Scoring Tables

- Add the scores – assign secondary structure based on highest score.
- Problems: Limited data set for scoring table. 17 amino acids –  $20^{17}$  possibilities =  $1.3 \times 10^{22}$  possible sequences, yet based on only 200-300 proteins!
- What do the scoring numbers mean? We are treating them as log-odds ratios, representing units of structural information.

# GOR Scoring Tables

- Based on information theory approach of Robson and Pain.
- Step 1- Consider the joint probability of amino acid R being in conformation S. The information function is  $I(S,R)=\text{Log}(P(S,R)/P(R)P(S))$  - this is Chou-Fasman
- Step 2 – For Garnier, in each conformation, calculate the difference of information functions,  
 $I(\Delta S,R)=\text{Log}(P(S,R)/P(S',R))+\text{Log}(P(S')/P(S))$  where S' = all other conformations except S. These terms are the values in the lookup tables.
- Probability terms calculated based on observed frequencies in the database of known structures as on 1978. Can actually use the net probability sum to calculate absolute probability ratios – so can estimate likelihoods.

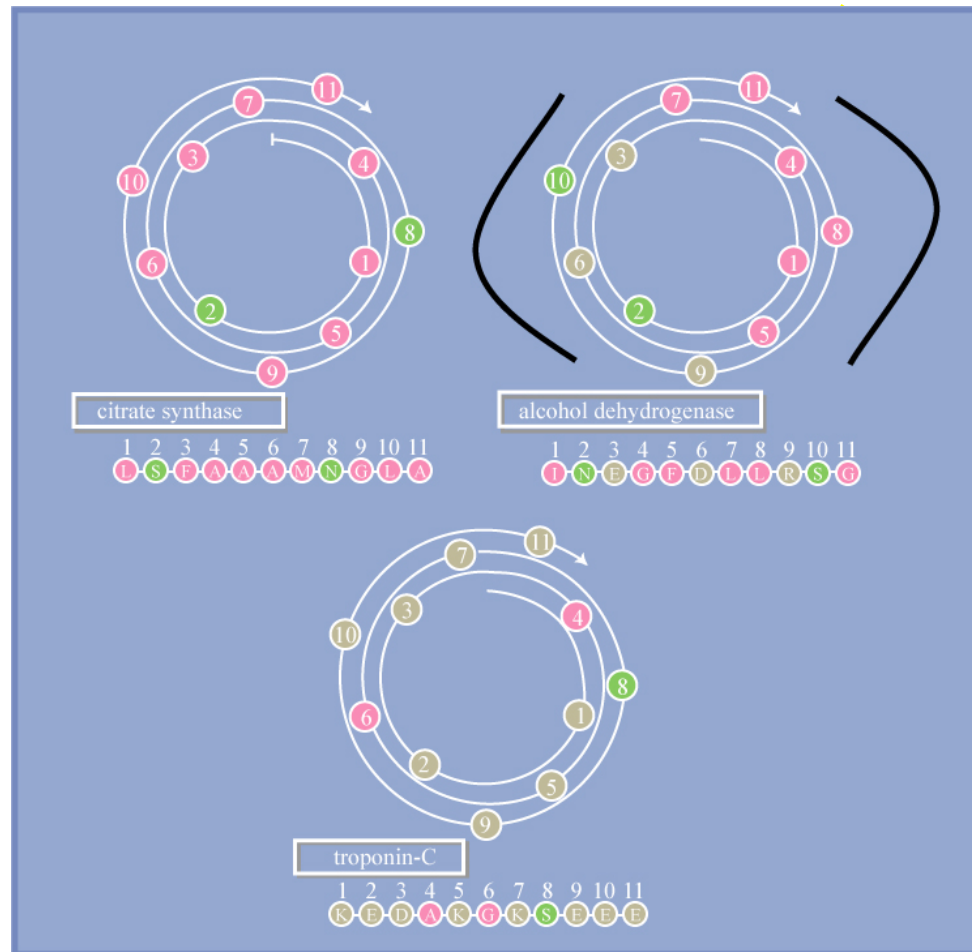
# GOR Results

- ~ 65% accurate.
- Can use information from experiments (circular dichroism) to improve accuracy of predictions.
- Later versions allowed pairwise combinations of amino acids in flanking regions + central amino acid (GOR-2), or combinations of two amino acids in the flanking region (GOR-3) influence the final conformation of the central amino acid.

# Fred Cohen's Approach-1989

- Both Garnier and Chou-Fasman work well for globular proteins
- Cohen: Turns demarcate elements of secondary structure
- Therefore, start by predicting turns first.
- Fill in helices, strands after that.
- Use pattern recognition algorithms (forerunner of neural networks).
- In  $\alpha/\beta$  proteins -  $\sim 85\%$  accurate. But how do you know you have an  $\alpha/\beta$  protein to begin with?

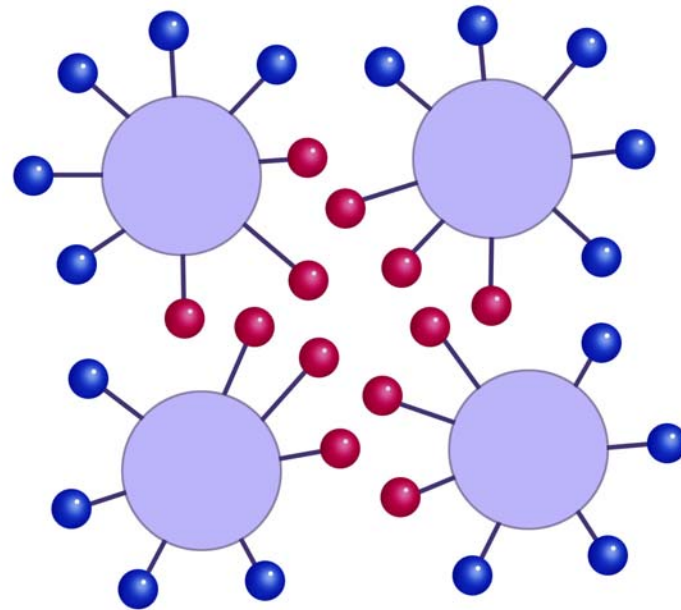
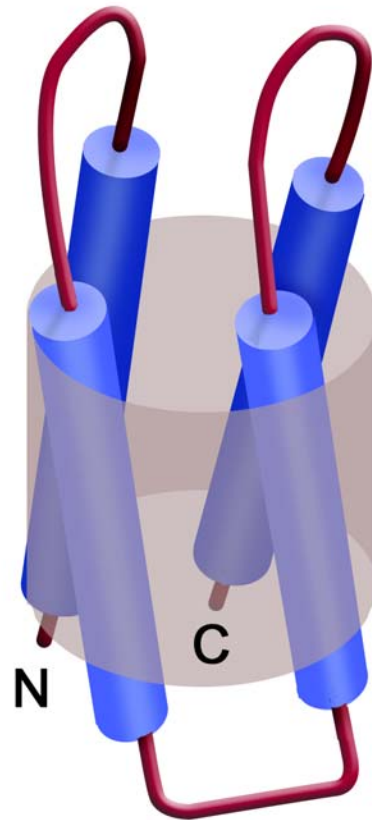
# Helical wheels and hydrophobic moments



hydrophobic

# Amphipathic helices

## Amphipathic helices



# Alternating hydrophobic and hydrophilic positions in $\beta$ -sheets

Please refer to Branden, Carl, and John Tooze. *Introduction to Protein Structure*. 2nd ed. Garland Publishing, Inc., 1999. ISBN: 0815323042.

# Eisenberg-Hydrophobic moments

- Standard approach – Kyte and Doolittle – calculate hydrophobicity using a running window and typical scale of hydrophobicity based on oil-water partition coefficients of free amino acid side chains.
- Eisenberg's idea – Plot hydrophobicity as function of sequence # - look for periodic repeats by fourier transform:
  - Period = 2 amino acids –  $\beta$  sheet
  - Period = 3-4 amino acids –  $\alpha$  helix



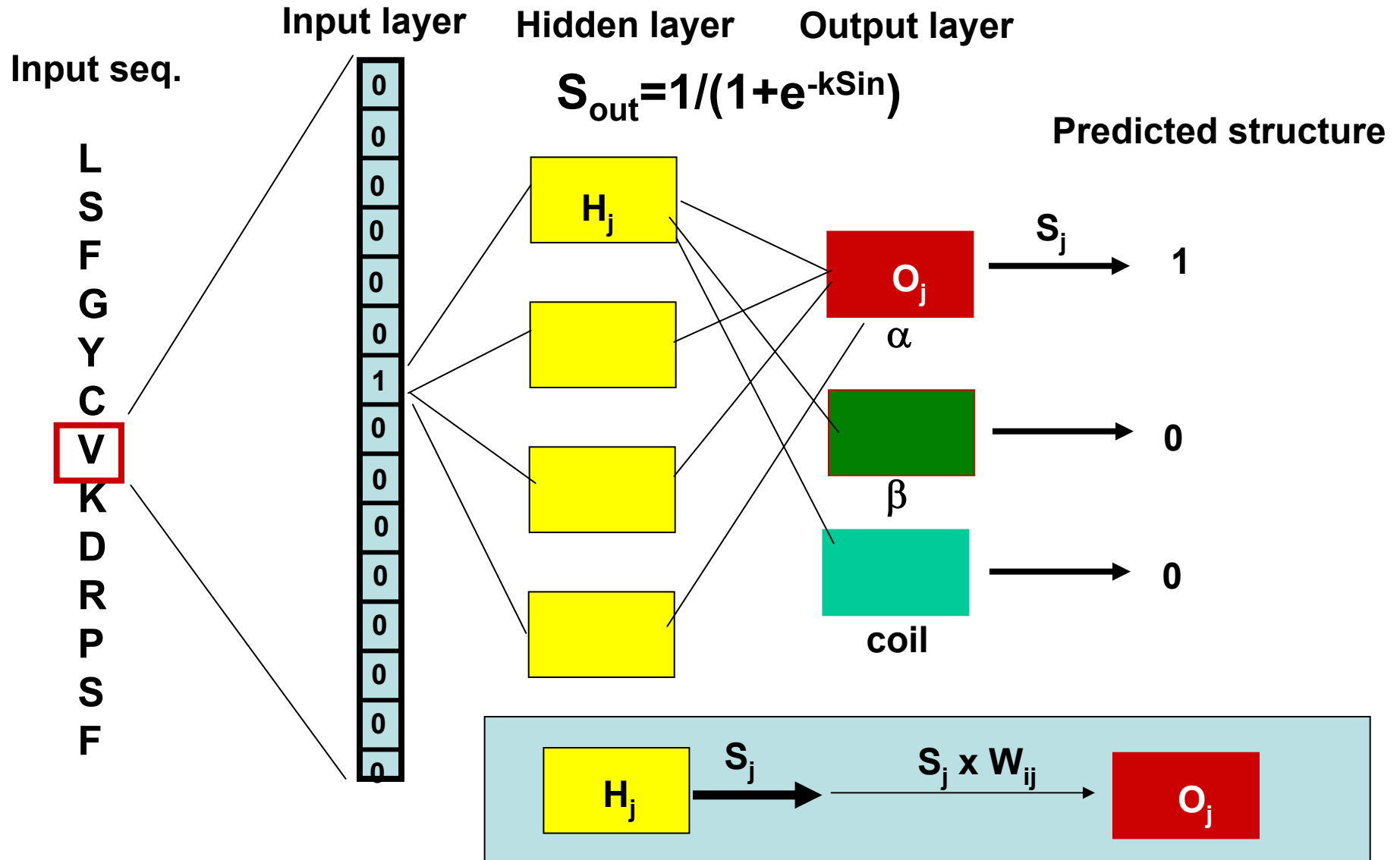
# Neural network approach

- Look for amino acid patterns that patterns in a protein sequence that coincide with known secondary structures.
- Use machine learning approaches and a test set of proteins to decipher the best pattern recognition algorithm.
- Simulate the operation of the brain, where complex synaptic connections underlie function. Some neurons collect data, some process data, some deliver output.

# Neural network approach

- Use sliding window of 13-17 amino acids.
- 3 processing layers in feed-forward multilayer network: input layer → hidden layer → output layer
- Each input modified by a weighting factor and many inputs are fed into the hidden layer. The hidden layer integrates the inputs and outputs a number close to 0 or 1 by feeding inputs into a sigmoid trigger function that mimics neuronal firing.
- Signals from hidden units sent to the each of three output units (one for helix, sheet or other), weighed again, and all the inputs integrated again. Final output from each output unit is a 1 (predicts that particular secondary structure) or a 0 (not predicted).

# Neural network approach



# Neural network approach

- Train network on training set to optimize the weighting factors  $W_{ij}$  using feedback.
- Usually done by Jack-knife testing.
- Can use multiple different network architectures and select final secondary structure by jury decision.
- Increases predictive accuracy to  $\sim 70-72\%$ .
- Best example: PHD (Profile network from HeiDelberg).
- Gives reliability indices for each predicted portion of the protein based on differences between output signals from the network.

# Nearest-neighbors Methods

- Also machine learning-based
- Identify sequences similar to the query in known structures. The known structures in the training set are divided into ~16 amino acid sequence fragments and secondary structure of central amino acid is recorded.
- Take similar window in the query sequence, match to best ~50 sequences in the training set. Use frequency of secondary structure of central amino acid in training data to infer structure in the query.
- Feed these structural predictions as input into a neural network to obtain the final prediction.
- Very accurate algorithms >72% correct prediction

# Nearest-neighbors Methods

- PREDATOR – another NN method that also considers amino acid patterns that can form H-bonds between adjacent  $\beta$ -strands and between  $n$  and  $n+4$  in  $\alpha$ -helices.
- Also considers substitutions found in sequence alignments, and gaps as likely to be “coils”
- Accuracy is  $\sim 75\%$  - most accurate prediction algorithm to date.

# Best overall strategy

- JPRED <http://jura.ebi.ac.uk:8888/>
- Developed by Geoffrey Barton
- A consensus approach to predicting secondary structure. Utilizes 6 different methods for prediction – PHD, linear discrimination (DSC), NNSSP, PREDATOR, ZPRED (conservative number weighted prediction), MULPRED (consensus single sequence method combination).
- \*\*\*\*Looks in pdb for homologues\*\*\*
- Available over the web, Q3=72.9%