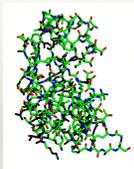


## Protein Structure Prediction

## Protein Structure Representation



CPK: hard sphere model



Ball-and-stick



Cartoon

## Degrees of Freedom in Proteins

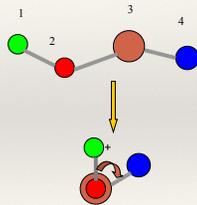
*Bond length*



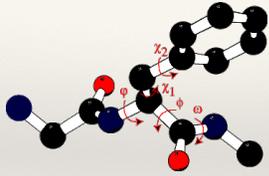
*Bond angle*



*Dihedral angle*



### Protein Structure Representation



**Backbone:** 3 angles per residue :  $\phi$ ,  $\psi$  and  $\omega$

**Sidechain:** 1 to 7 angles,  $\chi_i$ ; each  $\chi_i$  has 3 favored values:  $60^\circ$ ,  $-60^\circ$ ,  $180^\circ$ .

---

---

---

---

---

---

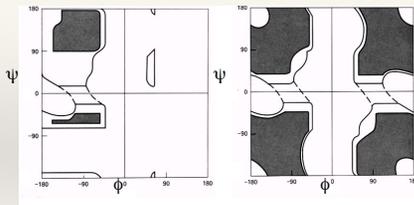
---

---

---

---

### Ramachandran Plots



All residues, but glycine

Glycine

*Acta Cryst.* (2002). D58, 768-776

---

---

---

---

---

---

---

---

---

---

### Root Mean Square Distance (RMSD)

To compare two sets of points (atoms)  $A=\{a_1, a_2, \dots, a_N\}$  and  $B=\{b_1, b_2, \dots, b_N\}$ :

-Define a 1-to-1 correspondence between A and B

for example,  $a_i$  corresponds to  $b_i$  for all  $i$  in  $[1, N]$

-Compute RMS as:

$$RMS(A, B) = \sqrt{\frac{1}{N} \sum_{i=1}^N d(a_i, b_i)^2}$$

$d(A_i, B_i)$  is the Euclidian distance between  $a_i$  and  $b_i$ .

---

---

---

---

---

---

---

---

---

---

## Structural Bioinformatics: Proteins

Proteins: Sources of Structure Information

Proteins: Homology Modeling

Proteins: Secondary structure prediction



---

---

---

---

---

---

---

---

## Structural Bioinformatics: Proteins

Proteins: Sources of Structure Information

Proteins: Homology Modeling

Proteins: Secondary structure prediction

Proteins: Ab initio prediction



---

---

---

---

---

---

---

---

## Proteins: Finding the Primary Structure

*Methods for finding the sequence of a protein:*

**-Translating gene sequence**

- For proteins from prokaryotes, direct translation
- For proteins from eukaryotes, we need the sequence of mRNA or cDNA

**-Edman degradation**

limited to "small" proteins, up to 50 amino acids for automated sequencer

**-Mass spectrometry**

---

---

---

---

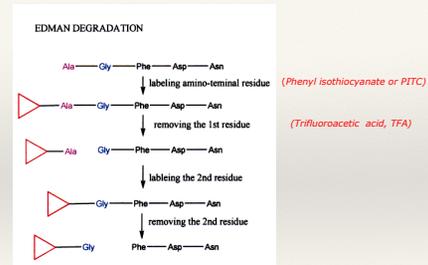
---

---

---

---

## Proteins: Finding the Primary Structure



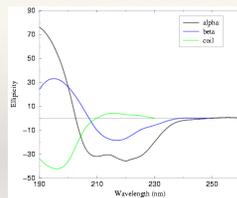
([http://en.wikibooks.org/wiki/Structural\\_Biochemistry/Protein/Protein\\_sequence\\_determination\\_techniques](http://en.wikibooks.org/wiki/Structural_Biochemistry/Protein/Protein_sequence_determination_techniques))

## Proteins: Finding the Tertiary Structure

Methods for finding the 3D structure of a protein:

- Circular Dichroism  
(low resolution; provides information on secondary structure)
- X-ray crystallography  
(high resolution; finds structure of a protein in a crystal)
- NMR spectroscopy  
(high resolution; finds structure of a protein in solution)

## Proteins: Circular Dichroism



Circular dichroism (CD) spectroscopy measures differences in the absorption of left-handed polarized light versus right-handed polarized light which arise due to structural asymmetry.

Different secondary structures in proteins have different CD spectra as they have different asymmetry.

CD therefore can detect secondary structures in protein

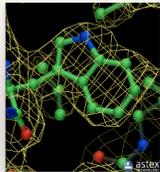


## From Diffraction to Electron Density Map

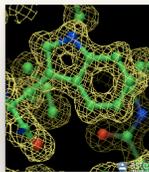


**One hidden problem:** diffraction patterns provide intensities; for Fourier transform, need intensity and phase. A significant step in X-ray crystallography is the solve the "phase problem".

## Fitting the structure: Influence of the resolution



2.6 Å resolution



1.2 Å resolution

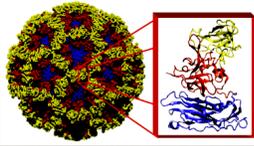
## Resolution of X-ray structures

Resolution (Å)	Meaning
>4.0	Individual coordinates meaningless
3.0 - 4.0	Fold possibly correct, but errors are very likely.
2.5 - 3.0	Fold likely correct except that some surface loops might be modelled.
2.0 - 2.5	Many small errors can normally be detected. Fold normally correct and number of errors in surface loops is small.
1.5 - 2.0	Water molecules and small ligands become visible. Many small errors can normally be detected. Folds are extremely rarely incorrect, even in surface loops.
0.5 - 1.5	In general, structures have almost no errors at this resolution. geometry studies are made from these structures.

([http://en.wikipedia.org/wiki/Resolution\\_\(electron\\_density\)](http://en.wikipedia.org/wiki/Resolution_(electron_density)))

Large molecular assemblies:  
X-ray crystallography and Cryo-EM

**X-ray structure**  
(180 copies of  
the same protein)



(Norwalk virus: <http://www.bcm.edu/molvir/norovirus>)

---

---

---

---

---

---

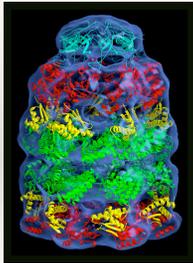
---

---

---

---

Large molecular assemblies:  
X-ray crystallography and Cryo-EM



**Cryo-EM:**

-Microscopy technique;  
as such, do not need crystal  
(closer to physiological  
conditions)

-Not high-resolution enough  
to provide atomic details;  
used in combination with  
modeling

---

---

---

---

---

---

---

---

---

---

Structural Bioinformatics: Proteins

Proteins: Sources of Structure Information

Proteins: Homology Modeling

Proteins: Secondary structure prediction



---

---

---

---

---

---

---

---

---

---

### Why we need Homology Modeling?

- ◊ Aim to solve the structure of all proteins: this is too much work experimentally!
- ◊ Solve enough structures so that the remaining structures can be inferred from those experimental structures
- ◊ The number of experimental structures needed depend on our abilities to generate a model.

---

---

---

---

---

---

---

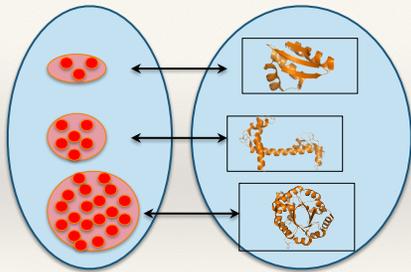
---

---

---

Sequence Space

Structure Space



---

---

---

---

---

---

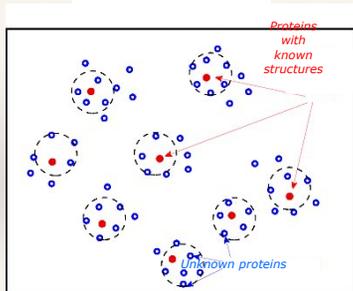
---

---

---

---

### Why we need Homology Modeling?



---

---

---

---

---

---

---

---

---

---



## Homology Modeling: Input

The query, also called target sequence

The template structure and sequence  
(need high quality structure)

The sequence alignment between query and template sequence  
(Probably the most important input!)

## Homology Modeling: Which program to use?

### 1) **Web service: SwissModel**

<http://swissmodel.expasy.org/>

3 modes:

- fully automatic
- "Alignment mode": you provide your own target-template alignment
- "Project mode": provides an environment to edit alignment

### 2) **Software: Modeller**

<http://www.sallab.org/modeller/>

Probably the best maintained software the homology modeling

## Do not start a homology modeling project before checking...

### - **Swiss-Model** repository

(<http://swissmodel.expasy.org/repository/>)

- Companion to the Swiss-Model tools – over 3.8 million models of protein domains

### - **ModBase**

(<http://modbase.compbio.ucsf.edu/>)

- Companion to **Modeller**

ModBase Contents	
Number of Models	38,312,698
Number of Unique Sequences modeled	6,501,740
Unique sequences attempted	8,082,334
Number of PDB chains	97,771
Statistics updated at {Fri Jan 30 10:03:05 PM PST 2026}	

## Structural Bioinformatics: Proteins

Proteins: Sources of Structure Information

Proteins: Homology Modeling

Proteins: Secondary structure prediction



---

---

---

---

---

---

---

---

## Secondary Structure Prediction

◊ Given a protein sequence  $a_1a_2\dots a_N$ , secondary structure prediction aims at defining the state of each amino acid  $a_i$  as being either H (helix), E (extended=strand), or O (other) (Some methods have 4 states: H, E, T for turns, and O for other).

◊ The quality of secondary structure prediction is measured with a "3-state accuracy" score, or  $Q_3$ .  $Q_3$  is the percent of residues that match "reality" (X-ray structure).

---

---

---

---

---

---

---

---

## Secondary Structure Assignment

Determine Secondary Structure positions in known protein structures using DSSP or STRIDE:

1. Kabesh and Sander. Dictionary of Secondary Structure in Proteins: pattern recognition of hydrogen-bonded and geometrical features. Biopolymer 22: 2571-2637 (1983) (DSSP)
2. Frischman and Argos. Knowledge-based secondary structure assignments. Proteins, 23:566-571 (1995) (STRIDE)

---

---

---

---

---

---

---

---

## Early methods for Secondary Structure Prediction

### ◊ *Chou and Fasman*

(Chou and Fasman. Prediction of protein conformation. Biochemistry, 13: 211-245, 1974)

### ◊ *GOR*

(Gamier, Osguthorpe and Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol., 120:97- 120, 1978)

## Chou and Fasman

◊ Start by computing amino acids propensities to belong to a given type of secondary structure:

$$\frac{P(i / \text{Helix})}{P(i)} \quad \frac{P(i / \text{Beta})}{P(i)} \quad \frac{P(i / \text{Turn})}{P(i)}$$

Propensities > 1 mean that the residue type i is likely to be found in the corresponding secondary structure type.

## Chou and Fasman

Amino Acid	$\alpha$ -Helix	$\beta$ -Sheet	Turn	
Ala	1.23	0.80	0.76	
Cys	1.11	0.74	0.80	
Leu	1.30	1.02	0.59	
Met	1.47	0.97	0.39	
Glu	1.44	0.75	1.00	
Gln	1.27	0.80	0.97	
His	1.22	1.05	0.69	
Lys	1.23	0.77	0.96	
Val	0.91	1.49	0.47	
Ile	0.97	1.45	0.51	
Phe	1.07	1.32	0.68	Favors
Tyr	0.72	1.25	1.05	$\beta$ -strand
Trp	0.59	1.14	0.75	
Thr	0.82	1.21	1.03	
Gly	0.56	0.92	1.64	
Ser	0.82	0.95	1.33	Favors
Asp	1.04	0.72	1.41	turn
Asn	0.90	0.76	1.23	
Pro	0.52	0.64	1.81	
Arg	0.96	0.99	0.88	

## Chou and Fasman

### Predicting helices:

- find nucleation site: 4 out of 6 contiguous residues with  $P(\alpha) > 1$
- extension: extend helix in both directions until a set of 4 contiguous residues has an average  $P(\alpha) < 1$  (breaker)
- if average  $P(\alpha)$  over whole region is  $> 1$ , it is predicted to be helical

### Predicting strands:

- find nucleation site: 3 out of 5 contiguous residues with  $P(\beta) > 1$
- extension: extend strand in both directions until a set of 4 contiguous residues has an average  $P(\beta) < 1$  (breaker)
- if average  $P(\beta)$  over whole region is  $> 1$ , it is predicted to be a strand

## Chou and Fasman

### Position-specific parameters

#### for turn:

Each position has distinct amino acid preferences.

#### Examples:

-At position 2, Pro is highly preferred; Trp is disfavored

-At position 3, Asp, Asn and Gly are preferred

-At position 4, Trp, Gly and Cys preferred

f(i)	f(i+1)	f(i+2)	f(i+3)	i	i+1	i+2	i+3
Ala	0.060	0.076	0.035	0.058			
Arg	0.070	0.106	0.029	0.085			
Asp	0.147	0.110	0.173	0.081			
Asn	0.161	0.063	0.193	0.091			
Cys	0.149	0.050	0.117	0.128			
Glu	0.056	0.060	0.077	0.064			
Gln	0.074	0.098	0.037	0.098			
Gly	0.102	0.085	0.190	0.152			
His	0.140	0.047	0.093	0.054			
Ile	0.043	0.034	0.013	0.056			
Leu	0.061	0.025	0.036	0.070			
Lys	0.055	0.115	0.072	0.095			
Met	0.068	0.082	0.014	0.055			
Phe	0.059	0.041	0.065	0.065			
Pro	0.102	0.301	0.034	0.068			
Ser	0.120	0.139	0.125	0.106			
Thr	0.086	0.108	0.065	0.079			
Trp	0.077	0.013	0.064	0.167			
Tyr	0.082	0.065	0.114	0.125			
Val	0.062	0.048	0.028	0.053			

## Chou and Fasman

### Predicting turns:

- for each tetrapeptide starting at residue  $i$ , compute:
  - $P_{turn}$  (average propensity over all 4 residues)
  - $F = f(i) * f(i+1) * f(i+2) * f(i+3)$

- if  $P_{turn} > P_{\alpha}$  and  $P_{turn} > P_{\beta}$  and  $P_{turn} > 1$  and  $F > 0.000075$  tetrapeptide is considered a turn.

### Chou and Fasman prediction:

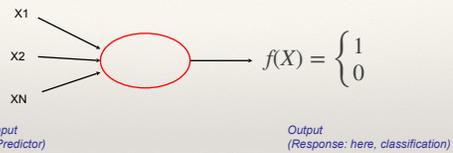
[http://fasta.bioch.virginia.edu/fasta\\_www/chofas.htm](http://fasta.bioch.virginia.edu/fasta_www/chofas.htm)

## Neural Networks

The most successful methods for predicting secondary structure are based on neural networks. The overall idea is that neural networks can be trained to recognize amino acid patterns in known secondary structure units, and to use these patterns to distinguish between the different types of secondary structure.

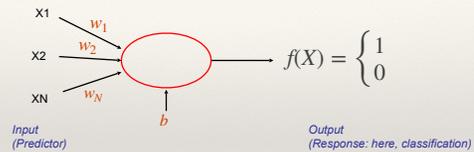
Neural networks classify "input vectors" or "examples" into categories (2 or more). They are loosely based on biological neurons.

## The perceptron



The perceptron classifies the input vector  $X$  into two categories.

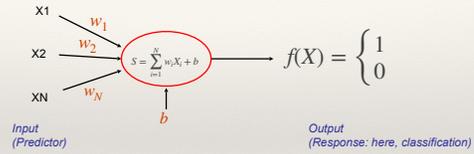
## The perceptron



The perceptron classifies the input vector  $X$  into two categories.

Add weights  $w$  to all predictors, and a bias  $b$

## The perceptron

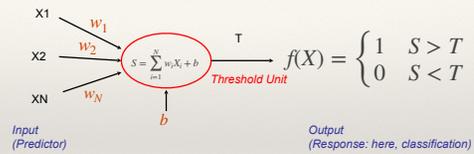


The **perceptron** classifies the input vector  $X$  into two categories.

Add weights  $w$  to all predictors, and a bias  $b$

Define simple "function"  $S$  as weighted sum of the predictors plus a bias

## The perceptron



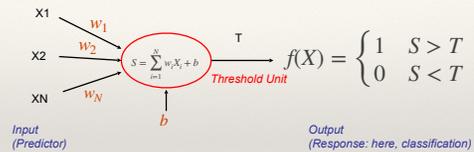
The **perceptron** classifies the input vector  $X$  into two categories.

Add weights  $w$  to all predictors, and a bias  $b$

Define simple "function"  $S$  as weighted sum of the predictors plus a bias

Add threshold to the function  $S$

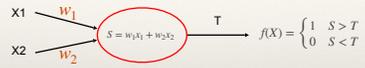
## The perceptron



If the weights, bias, and threshold  $T$  are not known in advance, the perceptron must be **trained**. Ideally, the perceptron must be trained to return the correct answer on all training examples, and perform well on examples it has never seen.

The training set must contain both type of data (i.e. with "1" and "0" output).

### Example: the OR gate



x1	x2	Output	S
1	1	1	$w_1 + w_2$
1	0	1	$w_1$
0	1	1	$w_2$
0	0	0	0

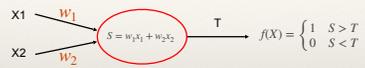
### Example: the OR gate



x1	x2	Output	S
1	1	1	$w_1 + w_2$
1	0	1	$w_1$
0	1	1	$w_2$
0	0	0	0

*Possible solution:*  
 $w_1 = 1$   
 $w_2 = 1$   
 $T = 0.5$

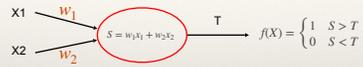
### (Counter) Example: the XOR gate



x1	x2	Output	S
1	1	0	$w_1 + w_2$
1	0	1	$w_1$
0	1	1	$w_2$
0	0	0	0

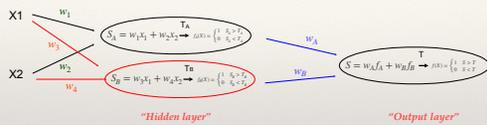
*Cannot find a solution with a "simple" perceptron*

### (Counter) Example: the XOR gate



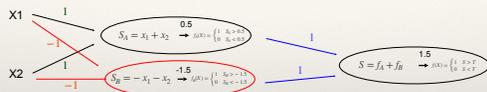
X1	X2	Output	S
1	1	0	$w_1 + w_2$
1	0	1	$w_1$
0	1	1	$w_2$
0	0	0	0

### A more complex network for the XOR gate



X1	X2	Output	S <sub>1</sub>	S <sub>2</sub>	S
1	1	0	$w_1 + w_2$	$w_1 + w_2$	$w_5w_1 + w_6w_2$
1	0	1	$w_1$	$w_1$	$w_5w_1 + w_6w_2$
0	1	1	$w_2$	$w_2$	$w_5w_1 + w_6w_2$
0	0	0	0	0	$w_5w_1 + w_6w_2$

### A more complex network for the XOR gate



X1	X2	Output
1	1	0
1	0	1
0	1	1
0	0	0

## The perceptron

Training a perceptron:

Find the weights  $W$  that minimizes the error function:

$$E = \sum_{i=1}^P (F(X^i \cdot W) - t(X^i))^2$$

$P$ : number of training data  
 $X^i$ : training vectors  
 $F(W \cdot X^i)$ : output of the perceptron  
 $t(X^i)$ : target value for  $X^i$

Use steepest descent:

- compute gradient:

- update weight vector:

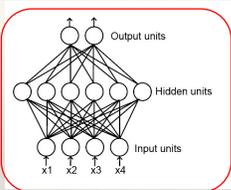
- iterate

$$\nabla E = \left( \frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \frac{\partial E}{\partial w_3}, \dots, \frac{\partial E}{\partial w_N} \right)$$

$$W_{new} = W_{old} - \epsilon \nabla E$$

( $\epsilon$ : learning rate)

## Neural Network

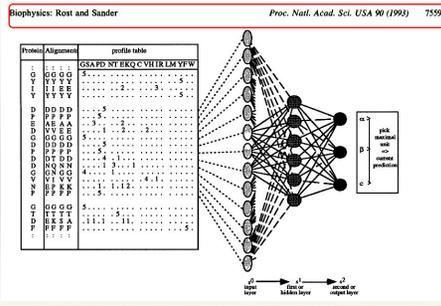


*A complete neural network is a set of perceptrons interconnected such that the outputs of some units becomes the inputs of other units. Many topologies are possible!*

Neural networks are trained just like perceptron, by minimizing an error function:

$$E = \sum_{i=1}^{Ndata} (NN(X^i) - t(X^i))^2$$

## PHD: Secondary structure prediction using NN





### Protein Structure Prediction

One popular model for protein folding assumes a sequence of events:

- ◊ Hydrophobic collapse
- ◊ Local interactions stabilize secondary structures
- ◊ Secondary structures interact to form motifs
- ◊ Motifs aggregate to form tertiary structure

---

---

---

---

---

---

---

---

### Protein Structure Prediction

A physics-based approach:

- ◊ - find conformation of protein corresponding to a thermodynamics minimum (free energy minimum)
- ◊ - cannot minimize internal energy alone! Needs to include solvent
- ◊ - simulate folding...a very long process!
- ◊ Folding time are in the ms to second time range; however, Folding simulations at best run 1 ns in one day...

---

---

---

---

---

---

---

---

### PHD: Secondary structure prediction using NN

- **Sequence-Structure network:** for each amino acid  $a_j$ , a window of 13 residues  $a_{j-6} \dots a_j \dots a_{j+6}$  is considered. The corresponding rows of the sequence profile are fed into the neural network, and the output is 3 probabilities for  $a_j$ :  $P(a_j, \text{alpha})$ ,  $P(a_j, \text{beta})$  and  $P(a_j, \text{other})$
- **Structure-Structure network:** For each  $a_j$ , PHD considers now a window of 17 residues: the probabilities  $P(a_k, \text{alpha})$ ,  $P(a_k, \text{beta})$  and  $P(a_k, \text{other})$  for  $k$  in  $[j-8, j+8]$  are fed into the second layer neural network, which again produces probabilities that residue  $a_j$  is in each of the 3 possible conformations
- **Jury system:** PHD has trained several neural networks with different training sets; all neural networks are applied to the test sequence, and results are averaged
- **Prediction:** For each position, the secondary structure with the highest average score is output as the prediction

---

---

---

---

---

---

---

---