CISC 436/636 Computational Biology & Bioinformatics (Fall 2016)

Protein Structure Prediction

Protein Secondary Structure

Protein structure

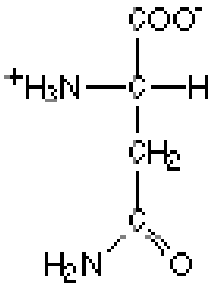
- Primary: amino acid sequence of the protein
- Secondary: characteristic structure units in 3-D.
- Tertiary: the 3-dimensional fold of a protein subunit
- Quaternary: the arrange of subunits in oligomers

Experimental Methods

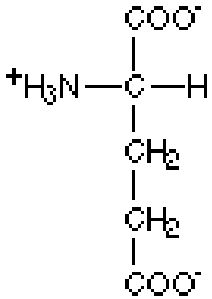
- X-ray crystallography
- NMR spectroscopy
- Neutron diffraction
- Electron microscopy
- Atomic force microscopy

- Computational Methods for secondary structures
  - Artificial neural networks
  - SVMs
  - …
- Computational Methods for 3-D structures
  - Comparative (find homologous proteins)
  - Threading
  - *Ab initio* (Molecular dynamics)

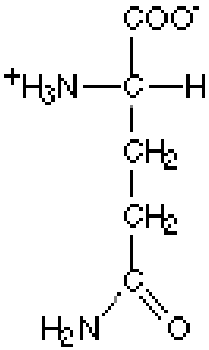
Amino acids with hydrophilic side groups



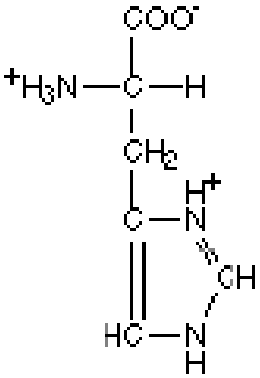
Asparagine
(asn)



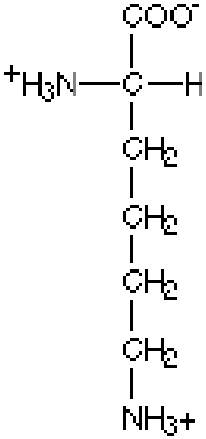
Glutamic acid
(glu)



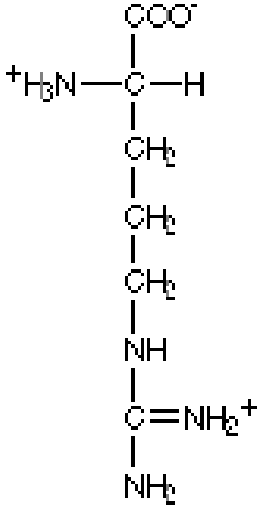
Glutamine
(gln)



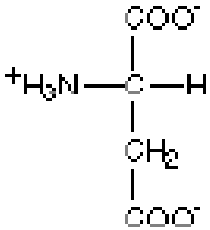
Histidine
(his)



Lysine
(lys)

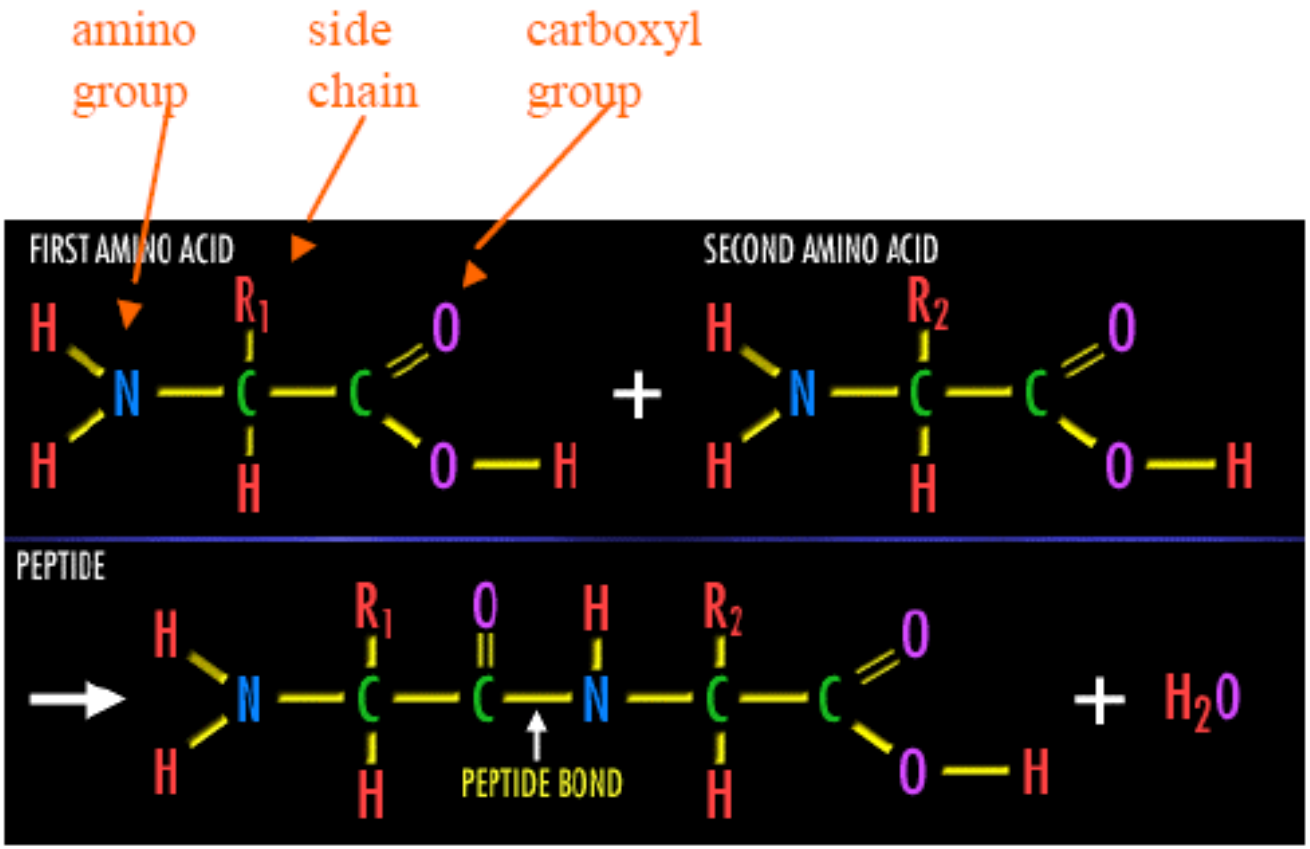


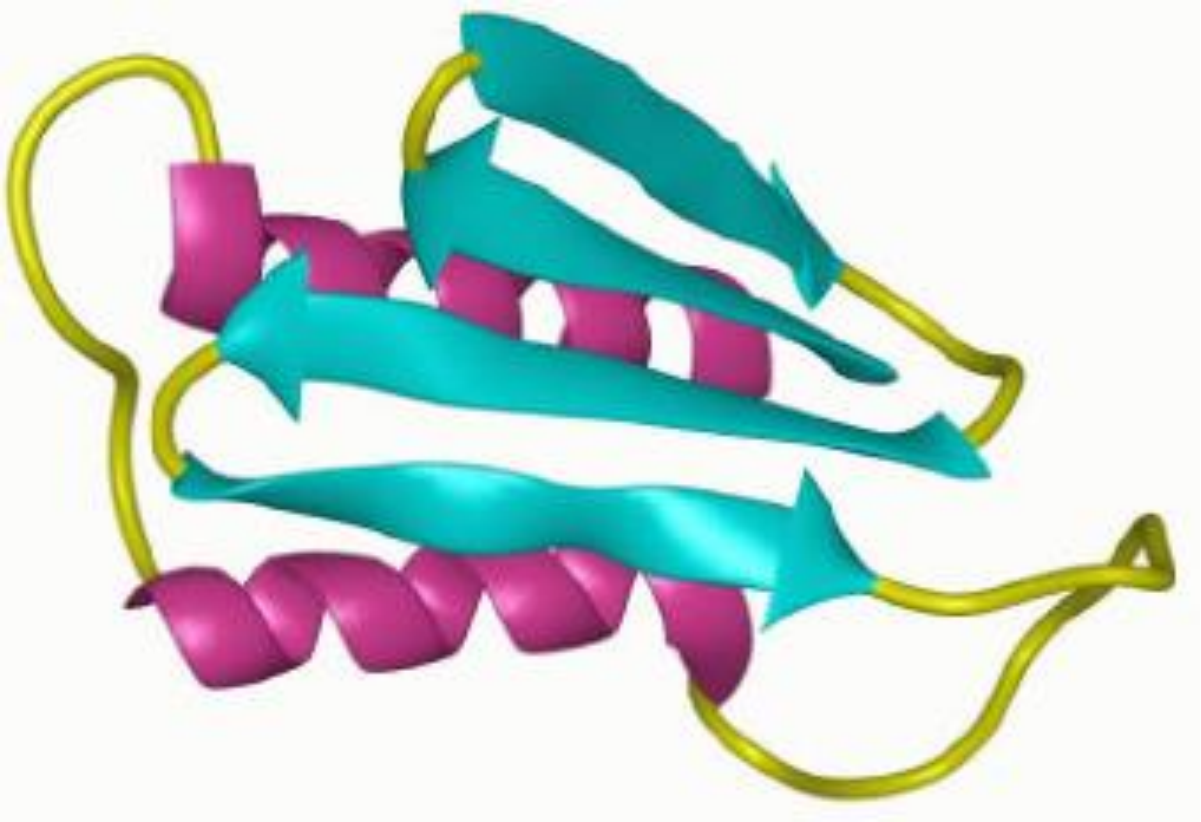
Arginine
(arg)



Aspartic acid
(asp)

Peptide Bonds





Scop Classification Statistics

SCOP: Structural Classification of Proteins. **1.65** release
20619 PDB Entries (1 August 2003). 54745 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

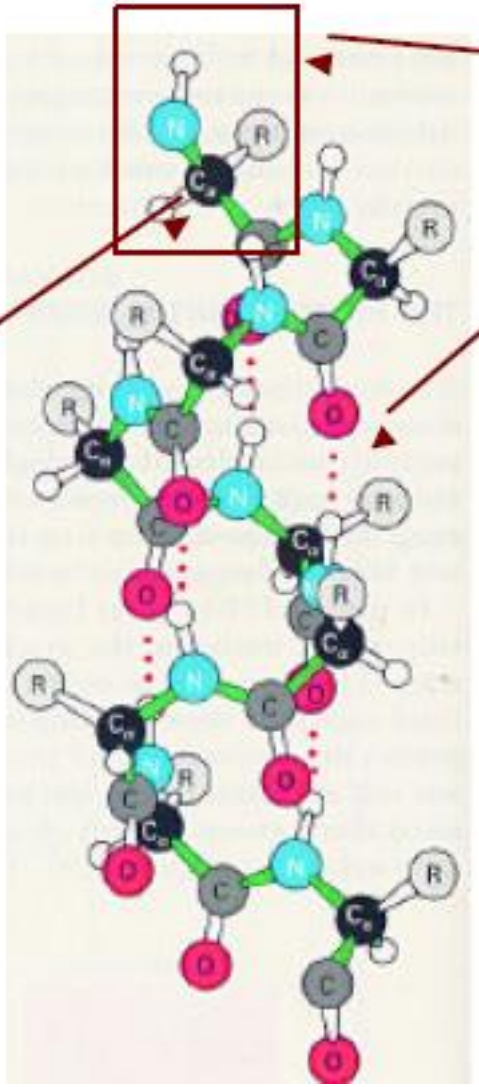
Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	179	299	480
All beta proteins	126	248	462
Alpha and beta proteins (a/b)	121	199	542
Alpha and beta proteins (a+b)	234	349	567
Multi-domain proteins	38	38	53
Membrane and cell surface proteins	36	66	73
Small proteins	66	95	150
Total	800	1294	2327

Root: scop

Classes:

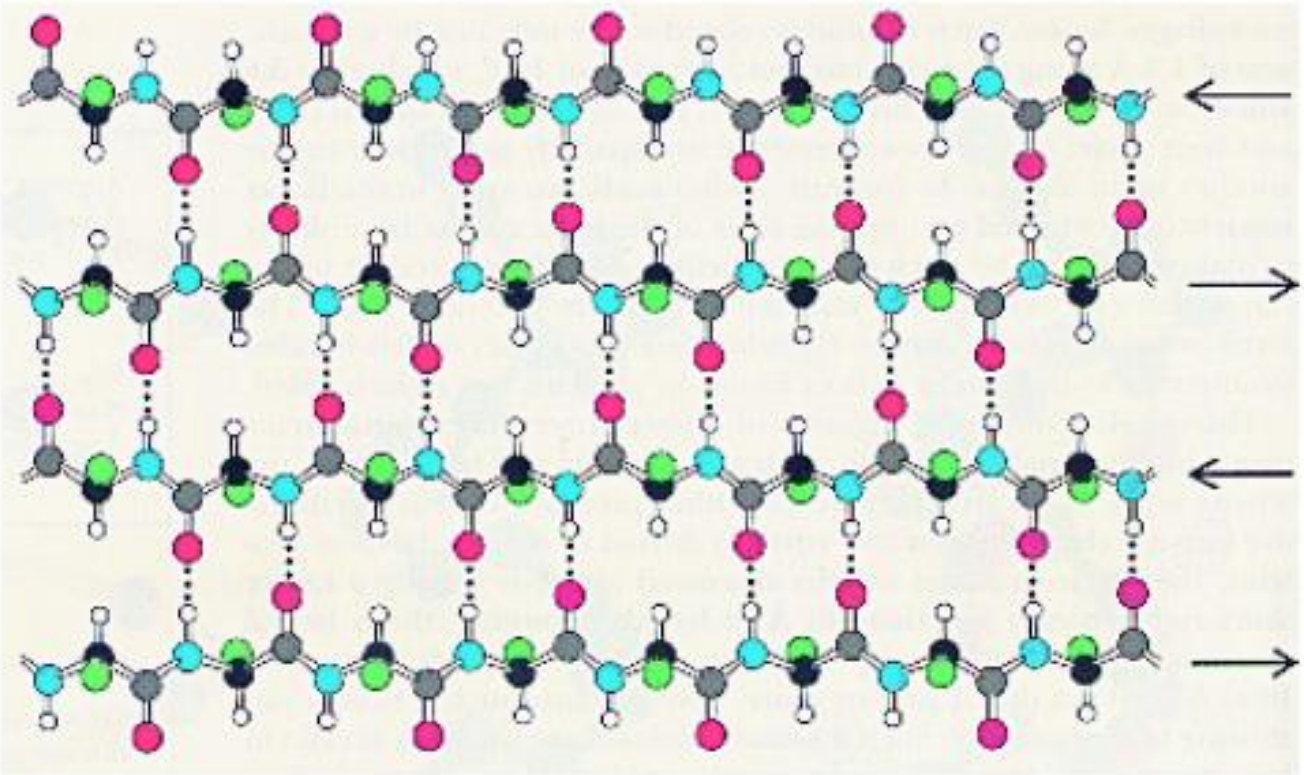
1. All alpha proteins [46456] (226)
2. All beta proteins [48724] (149)
3. Alpha and beta proteins (a/b) [51349] (134)
   Mainly parallel beta sheets (beta-alpha-beta units)
4. Alpha and beta proteins (a+b) [53931] (286)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) [56572] (48)
   Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides [56835] (49)
   Does not include proteins in the immune system
7. Small proteins [56992] (79)
   Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Coiled coil proteins [57942] (7)
   Not a true class
9. Low resolution protein structures [58117] (24)
   Not a true class
10. Peptides [58231] (116)
    Peptides and fragments. Not a true class
11. Designed proteins [58788] (42)
    Experimental structures of proteins with essentially non-natural sequences. Not a true cla

α Helices

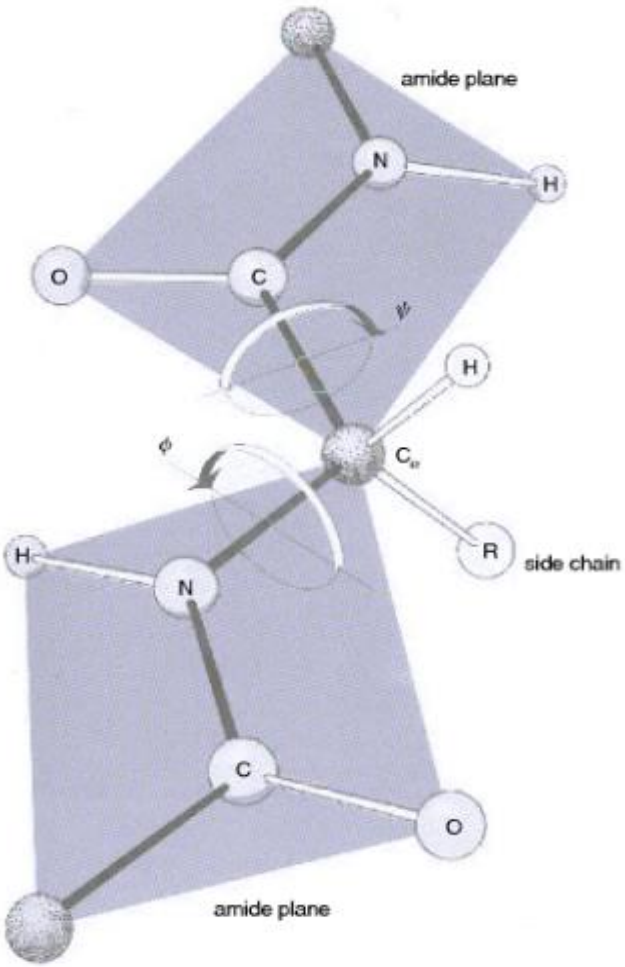


- Helix complete turn every 3.6 AAs
- Hydrogen bond between (-C=O) of one AA and (-N-H) of its 4th neighboring AA

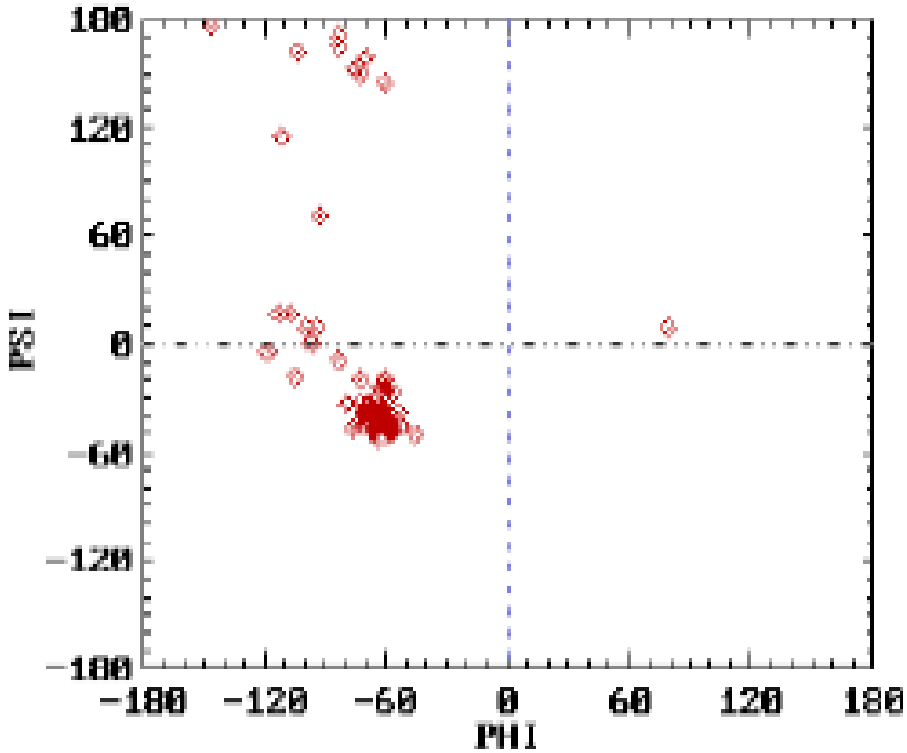
β Strands



Hydrogen bond b/w carbonyl oxygen atom on one chain and NH group on the adjacent chain

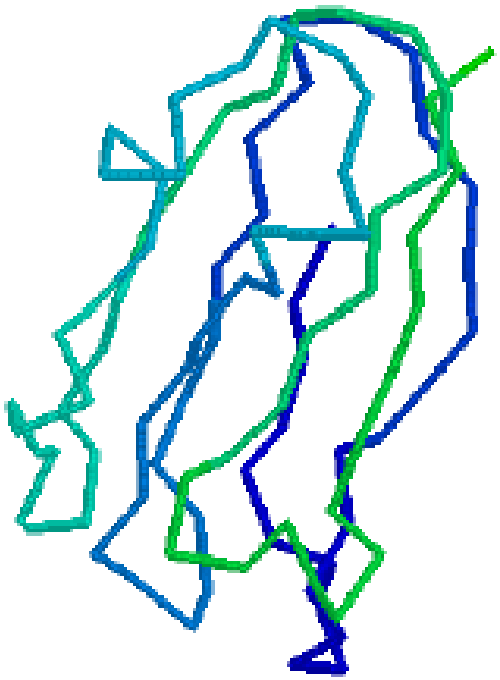

amide plane
N
H
O
C
ψ
H
φ
Cα
R
side chain
H
N
C
O
amide plane

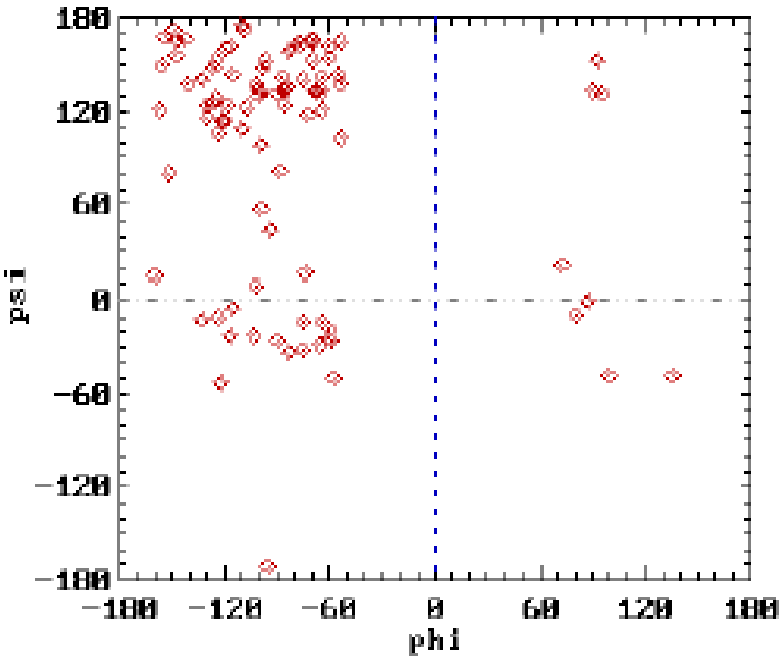
Ramachandran Plot



PHI: -57; PSI -47

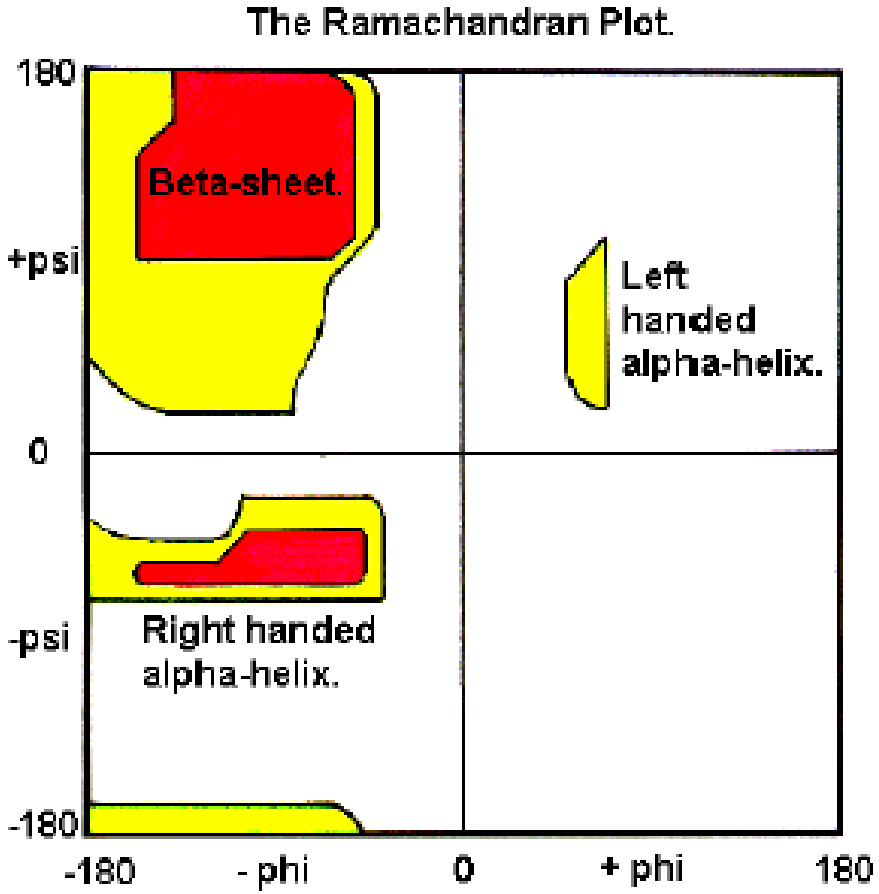
Ramachandran Plot





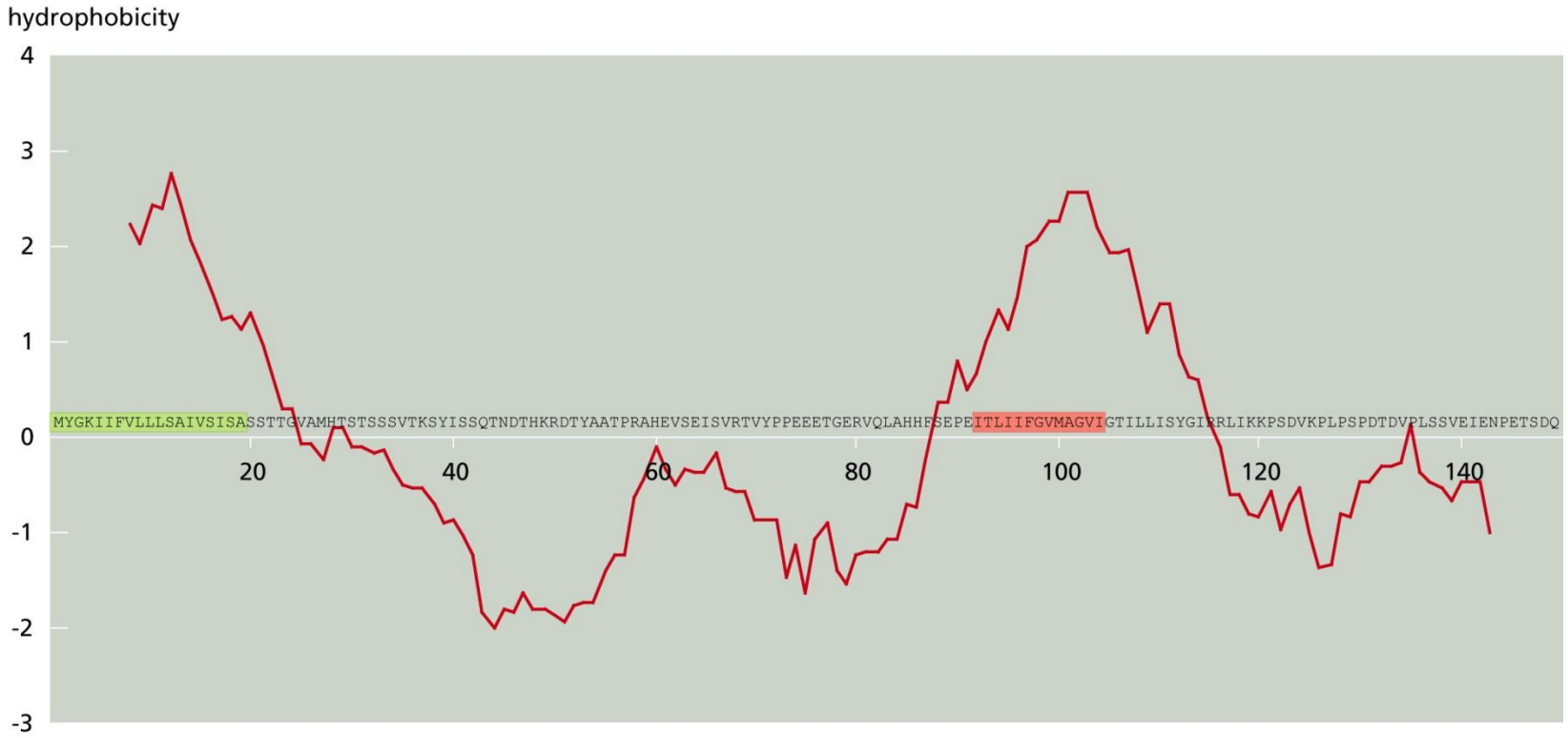
Parallel: PHI: -119; PSI: 113

Anti-parallel: PHI: -139; PSI: 135


The Ramachandran Plot.
180
+psi
0
-psi
-180
-180
- phi
0
+ phi
180
Beta-sheet.
Left handed alpha-helix.
Right handed alpha-helix.

Hydrophobicity Scales

	Kyte-Doolittle	Hopp-Woods
Alanine	1.8	-0.5
Arginine	-4.5	3.0
Asparagine	-3.5	0.2
Aspartic acid	-3.5	3.0
Cysteine	2.5	-1.0
Glutamine	-3.5	0.2
Glutamic acid	-3.5	3.0
Glycine	-0.4	0.0
Histidine	-3.2	-0.5
Isoleucine	4.5	-1.8
Leucine	3.8	-1.8
Lysine	-3.9	3.0
Methionine	1.9	-1.3
Phenylalanine	2.8	-2.5
Proline	-1.6	0.0
Serine	-0.8	0.3
Threonine	-0.7	-0.4
Tryptophan	-0.9	-3.4
Tyrosine	-1.3	-2.3
Valine	4.2	-1.5


hydrophobicity
4
3
2
1
0
-1
-2
-3
MYGKIIFVLLLSAIVSISASSTTGVAMHTSTSSSVTKSYISSQTNDTHKRDTYAATPRAHEVSEISVRTVYPPEEETGERVQLAHHFSEPEITLIIFGVMAGVIGTILLISYGIRRLIKKPSDVKPLPSPDTDVPLSSVEIENPETSDQ
20
40
60
80
100
120
140

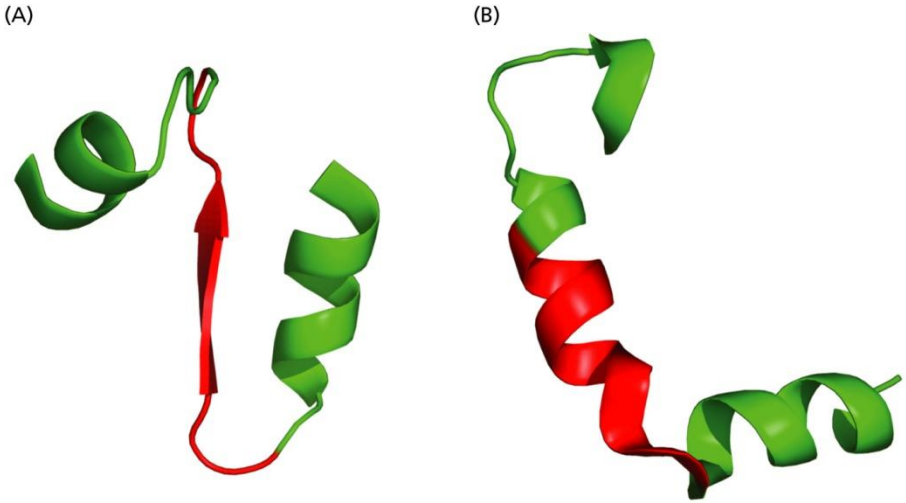
Residue conformation preferences

Helix: A, E, K, L, M, R

Sheet: C, I, F, T, V, W, Y

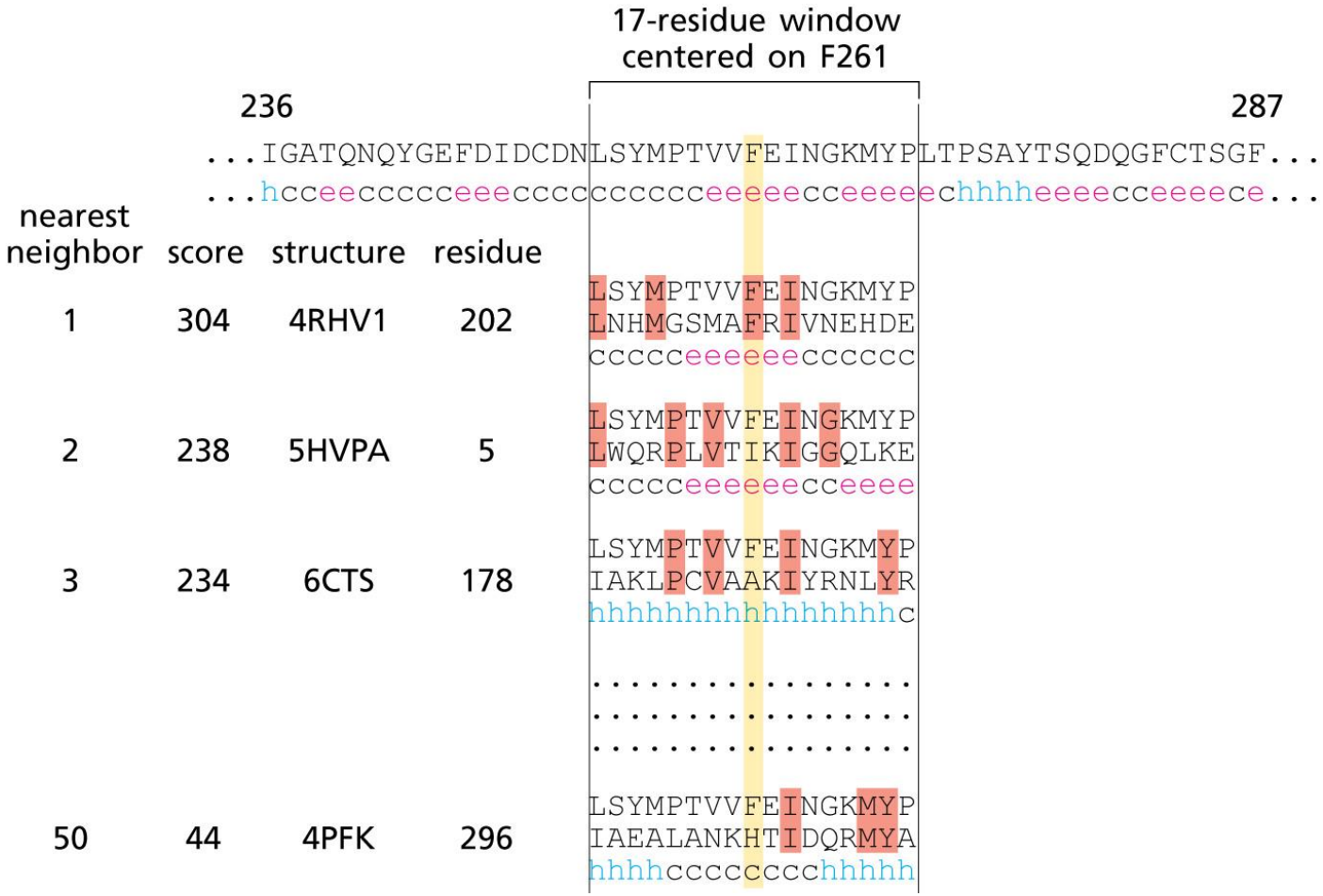
Coil: D, G, N, P, S

Structures are modulated by nearby sequence



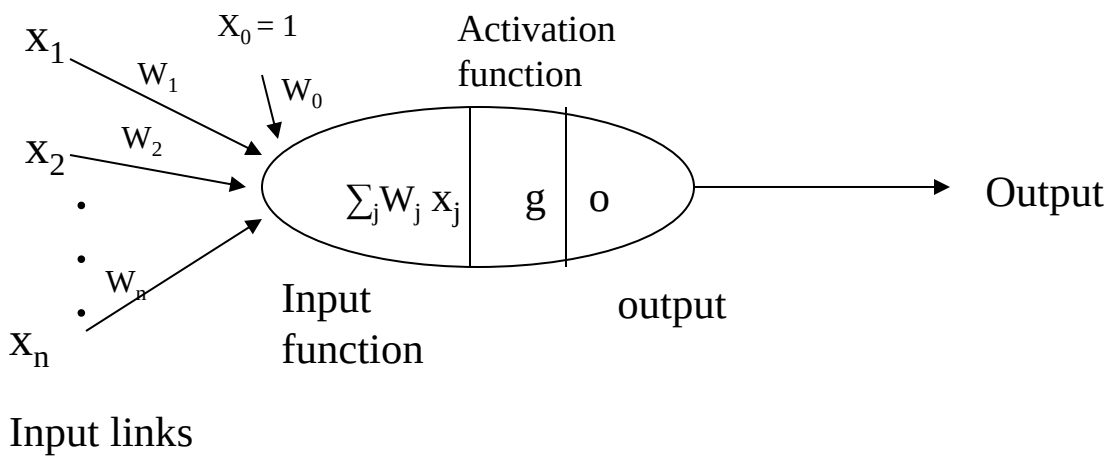
The nine-residue sequence KGVVPQLVK (in red) occurs in two proteins (1IAL and 1PKY) but with completely different structures. (Fig. 12.12)

The Nearest-neighbor methods based on segment similarity

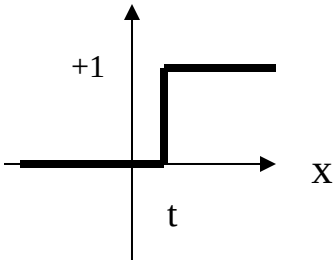


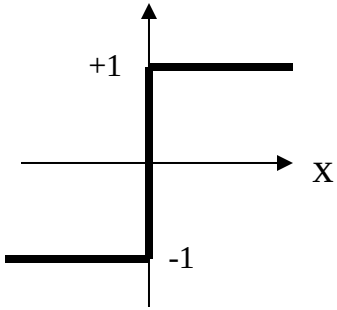
Artificial neural networks

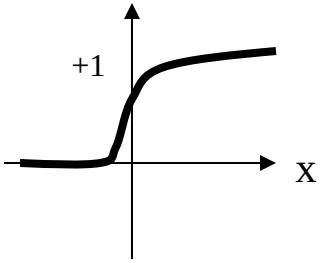
- Perceptron $o(x_1, \ldots, x_n) = g(\sum_j W_j x_j)$



- Activation functions







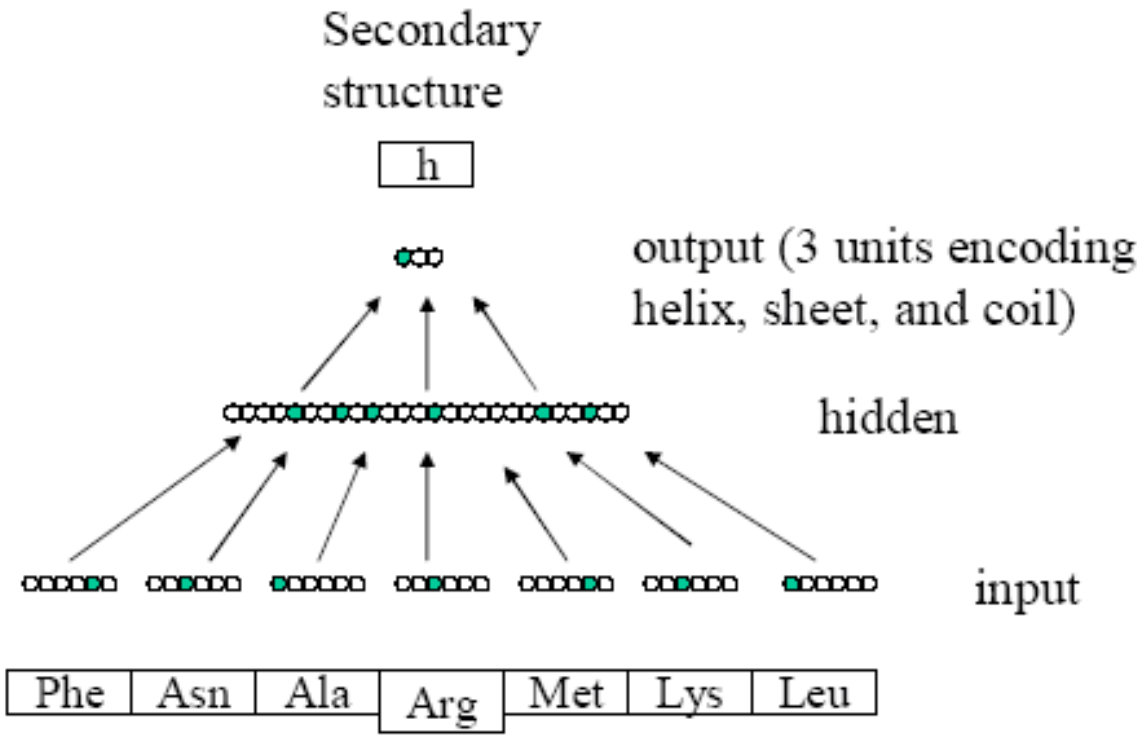
$$\text{Step}(x) = \begin{cases} 1 \text{ if } x \geq t \\ 0 \text{ otherwise} \end{cases}$$

$$\text{Sign}(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ -1 \text{ otherwise} \end{cases}$$

$$\text{Sigmoid}(x) = 1/(1+e^{-x})$$

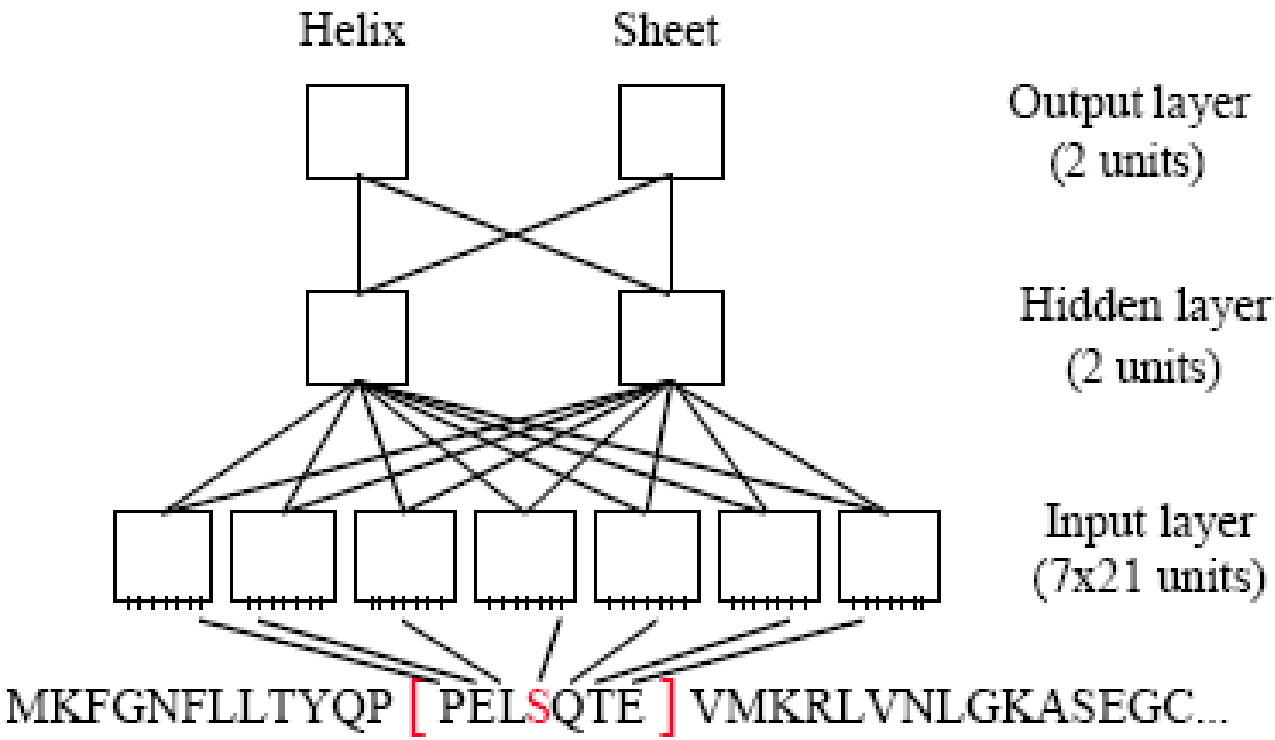
Artificial Neural Networks

Qian & Sejnowski, JMB 202(1988)865-884



Sequence of amino acid processed as sliding windows of fixed-length (7 to 17 aa) segments. The central residues are then classified by a three-state (helix, sheet, or coil) prediction.

2-unit output

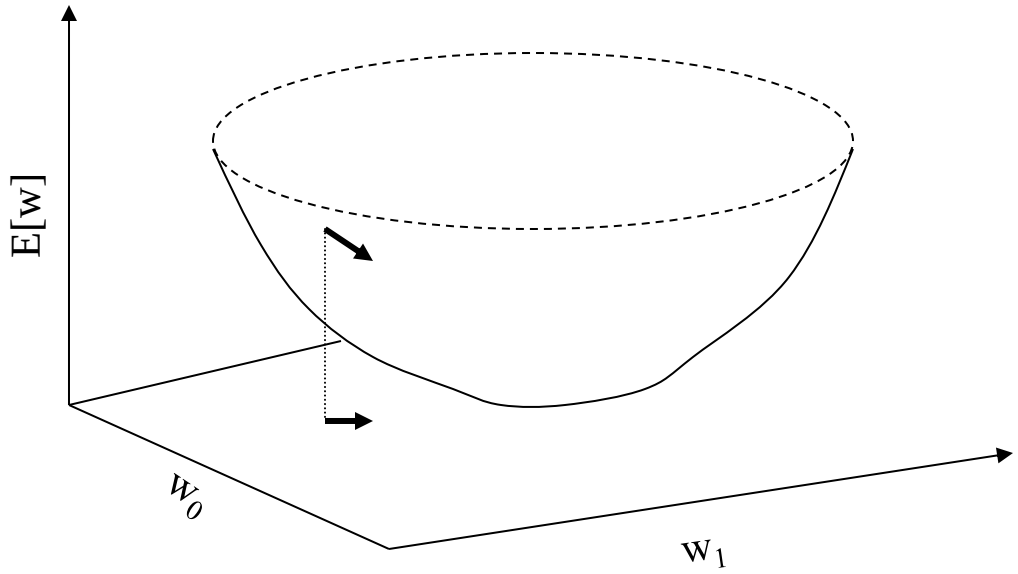


- Learning: to determine weights and thresholds for all nodes (neurons) so that the net can approximate the training data within error range.
  - Back-propagation algorithm
    - Feedforward from Input to output
    - Calculate and back-propagate the error (which is the difference between the network output and the target output)
    - Adjust weights (by *gradient descent*) to decrease the error.

Gradient descent

$$\mathbf{w}_{new} = \mathbf{w}_{old} - r\,[\partial E/\partial \mathbf{w}]$$

where r is a positive constant called learning rate, which determines the step size for the weights to be altered in the steepest descent direction along the error surface.



Data representation

- Direct sequence encoding
  - BIN4:

    A →1000; T → 0100; G → 0010; C → 0001; - → 0000
  - BIN2:

    A →00; T → 01; G → 10; C → 11
  - For amino acids: each amino acid → a vector of 21 bits (This is called BIN21
  - Other properties of amino acids, such as hydrophobicity.
- Indirect sequence encoding

  Sequence features and information content can be extracted by various scoring mechanisms.
  - Residue frequency
- Input trimming

  Reduce dimensions and condense information content
  - Decision trees
  - Singular value decomposition (SVD)
  - Principle component analysis (PCA)

- Issues with ANNs
  - Network architecture
    - FeedForward (fully connected vs sparsely connected)
    - Recurrent
    - Number of hidden layers, number of hidden units within a layer
  - Network parameters
    - Learning rate
    - Momentum term
  - Input/output encoding
    - One of the most significant factors for good performance
    - Extract maximal info
    - Similar instances are encoded to “closer” vectors

An on-line service

Address http://www.cmpharm.ucsf.edu/cgi-bin/nnpredict.pl

Results of nnpredict query

Tertiary structure class: none

Sequence:

```
MANLGYWLLALFVTMWTDVGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGTWGQPH
GGGWGQPHGGSWGQPHGGSWGQPHGGGWGQGGGTHNQWNKPSKPKTNLKHVAGAAAAGAV
VGGLGGYMLGSAMSRPMIHFGNDWEDRYYRENMYRYPNQVYYRPVDQYSNQNNFVHDCVN
ITIKQHTVTTTTKGENFTETDVKMMERVVEQMCVTQYQKESQAYYDGRRSSSTVLFSSPP
VILLISFLIFLIVG
```

Secondary structure prediction ***(H = helix, E = strand, - = no prediction)*:**

```
---HHHHHHHHHHH-------------------------------------------
---------------------------------------------HHHHHHHHHHHE
E------EE-------EEE-------HH-------------------------------
EEE--E-E-E--------HHHHHHHHHHH-HHH----------EE------EEEE-----
-EEEEHHEEEEE--
```

- Performance
  - ceiling at about 65% for direct encoding
    - Local encoding schemes present limited correlation information between residues
    - Little or no improvement using multiple hidden layers.
  - Surpassing 70% by
    - Including evolutionary information (contained in multiple alignment)
    - Using cascaded neural networks
    - Incorporating global information (e.g., position specific conservation weights)

Cathy Wu, Computers Chem. 21(1997)237-256

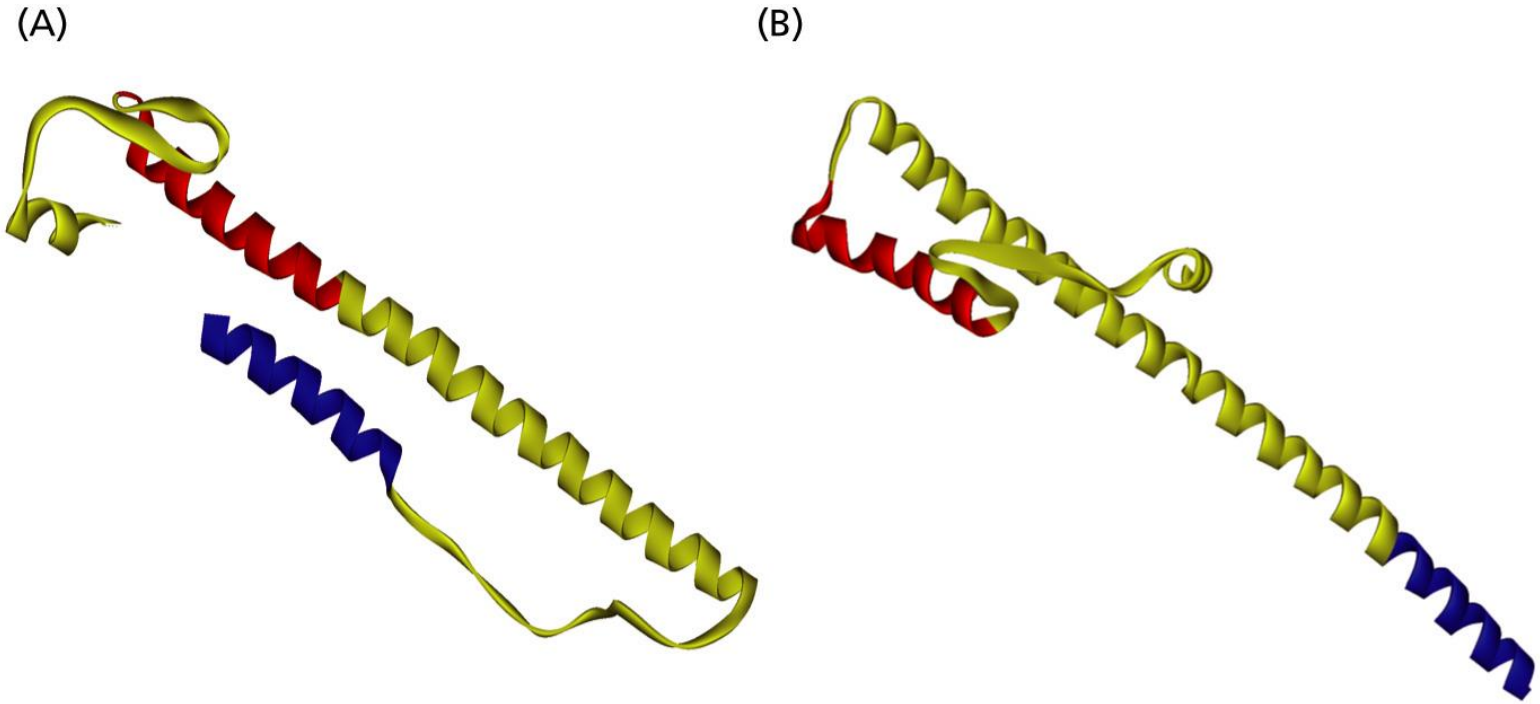
Table 1. Neural network applications for DNA/RNA sequence analysis

Reference	Application	Neural network*	I/O encoding†
Intron/Exon (I/E) Discrimination and Gene Identification			
Uberbacher and Mural, 1991	Coding region recognition	4L/FF/BP	FEAT7/1(Y,N)
Uberbacher *et al.*, 1996	Coding region recognition	3L/FF/BP	FEAT13/1(Y,N)
Snyder and Stormo, 1993	I/E feature weighting	2L/FF/Delta	FEAT6/1(Inequality)
Snyder and Stormo, 1995	I/E feature weighting	2,3L/FF/Delta,BP	FEAT6/1(Inequality)
Brunak *et al.*, 1991	Splicing donor/acceptor site prediction	3L/FF/BP	BIN4/1(Y,N)
Farber *et al.*, 1992	I/E discrimination	2L/FF/BP	BIN4,FREQ/1(Y,N)
Granjeon and Tarroux, 1995	I/E compositional constraints	3L/FF/BP	BIN4/3(I,E,O)
Reczko *et al.*, 1995	Parallel implementation for I/E discrimination	3L/FF/BP,QP,RP	BIN4/1(I,E)
Prediction and Analysis of Ribosome-binding Sites, Promoters and Other Sites			
Stormo *et al.*, 1982a	Ribosome-binding site prediction	Perceptron	BIN4/1(Y,N)
Bisant and Maizel, 1995	Ribosome-binding site prediction	3L/FF/BP	BIN4/1(Y,N)
Abremski *et al.*, 1993	*E. coli* promoter prediction	3L/FF/BP	BIN4/1(Y,N)
Demeler and Zhou, 1991	*E. coli* promoter prediction	3L/FF/BP	BIN2,BIN4/1(Y,N)
O'Neill, 1991, 1992	*E. coli* promoter prediction	3L/FF/BP	BIN4/1(Y,N)
Horton and Kanehisa, 1992	*E. coli* promoter prediction	2L/FF/BP	BIN4 + 3 + FREQ/1(Y,N)
Mahadevan and Ghosh, 1994	*E. coli* promoter prediction	2 × 3L/FF/BP	BIN4/1(Y,N)
Pedersen and Engelbrecht, 1995	Transcription start site and feature detection	3L/FF/BP	BIN4/1(Y,N)
Larsen *et al.*, 1995	Eukaryotic promoter prediction	3L/FF/BP	BIN4/1(Y,N)
Matis *et al.*, 1996	RNA polymerase II binding site prediction	4L/FF/BP	FEAT13/1(Y,N)
Nair *et al.*, 1994	Prediction of transcriptional terminator	3L/FF/BP	BIN4,REAL1/1(Y,N)
Nair *et al.*, 1995	Prediction of transcription control signal	3L/FF/BP	BIN4/1(RTL)
DNA/RNA Sequence Analysis, Phylogenetic Classification and Code Mapping			
Arrigo *et al.*, 1991	Clustering and functional region identification	2L/Kohonen	REAL1/Map(30)
Giuliano *et al.*, 1993	Clustering and functional region identification	2L/Kohonen	REAL1/Map
Leblanc *et al.*, 1994	Phylogenetic classification	2L/ART	BIN4/19(Class)
Wu and Shivakumar, 1994	Ribosomal RNA classification	2 × 3L/FF/BP,CP	FREQ,SVD/220,15(Class)
Sun *et al.*, 1995	Transfer RNA gene recognition	3L/FF/BP	BIN4/10(Class)
Tolstrup *et al.*, 1994	Genetic code mapping	3L/FF/BP	BIN4/20(Class)

*Neural network architectures: 2L/FF = two-layer, feedforward network (i.e. perceptron); 3L or 4L/FF = three- or four-layer, feedforward network (i.e. multi-layer perceptron).

Neural network learning algorithms: BP = Back-propagation; Delta = Delta rule; QP = Quick-propagation; RP = Rprop; ART = Adaptive resonance theory; CP = Counter-propagation.

Environmental effects



3HMG

1HTM

Credit: Fig. 12.15

Resources

- Protein Structure Classification
  - CATH: http://www.biochem.ucl.ac.uk/bsm/cath/
  - SCOP: http://scop.mrc-lmb.cam.ac.uk/scop/
  - FSSP:
- PDB: http://www.rcsb.org/pdb/