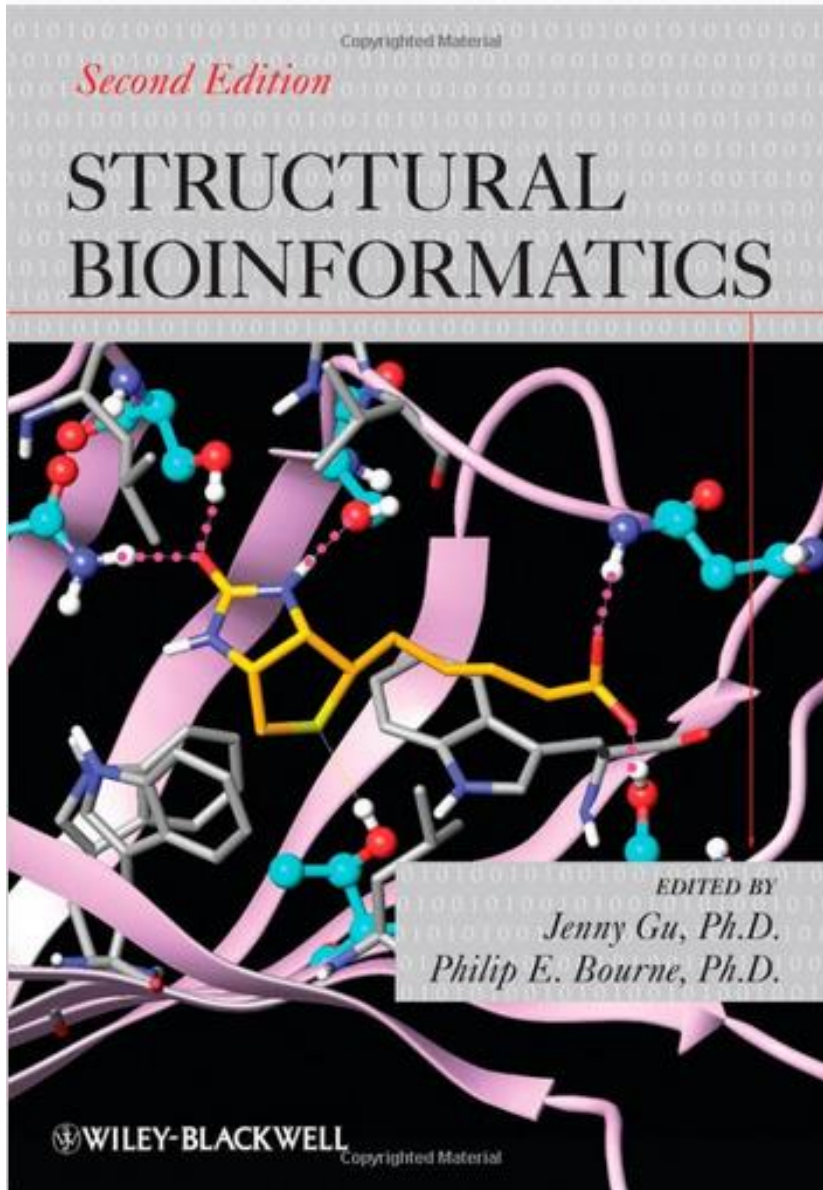


GLOBEX Bioinformatics (Summer 2015)

Protein Structure Prediction

- Basic concepts
- Secondary structure
- Lattice model



A big volume:
40 chapters
1096 pages

Protein structure

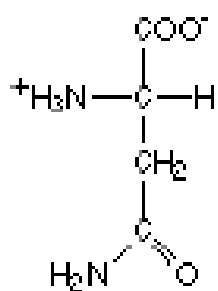
- Primary: amino acid sequence of the protein
- Secondary: characteristic structure units in 3-D.
- Tertiary: the 3-dimensional fold of a protein subunit
- Quaternary: the arrange of subunits in oligomers

Experimental Methods

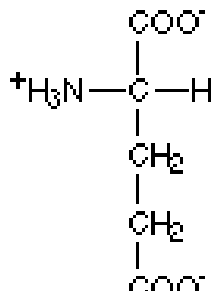
- X-ray crystallography
- NMR spectroscopy
- Neutron diffraction
- Electron microscopy
- Atomic force microscopy

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

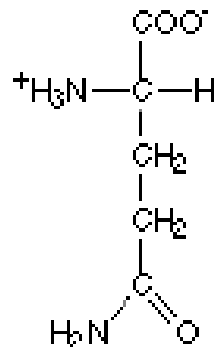
Amino acids with hydrophilic side groups



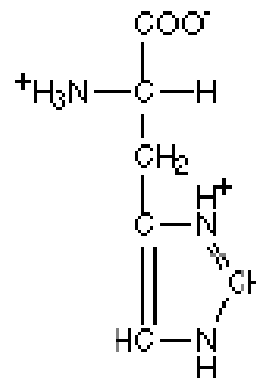
Asparagine
(asn)



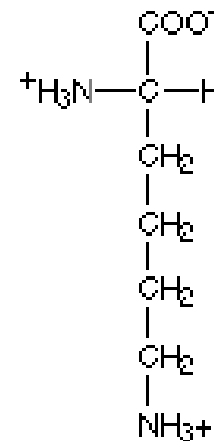
Glutamic acid
(glu)



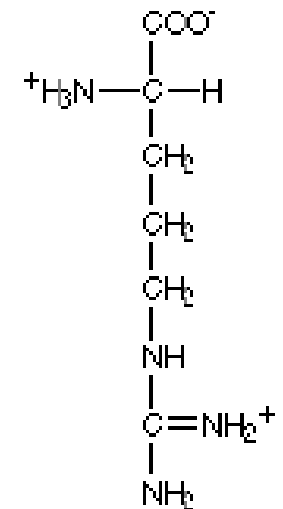
Glutamine
(gln)



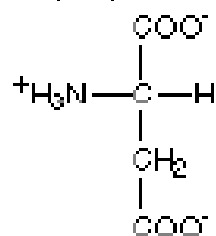
Histidine
(his)



Lysine
(lys)

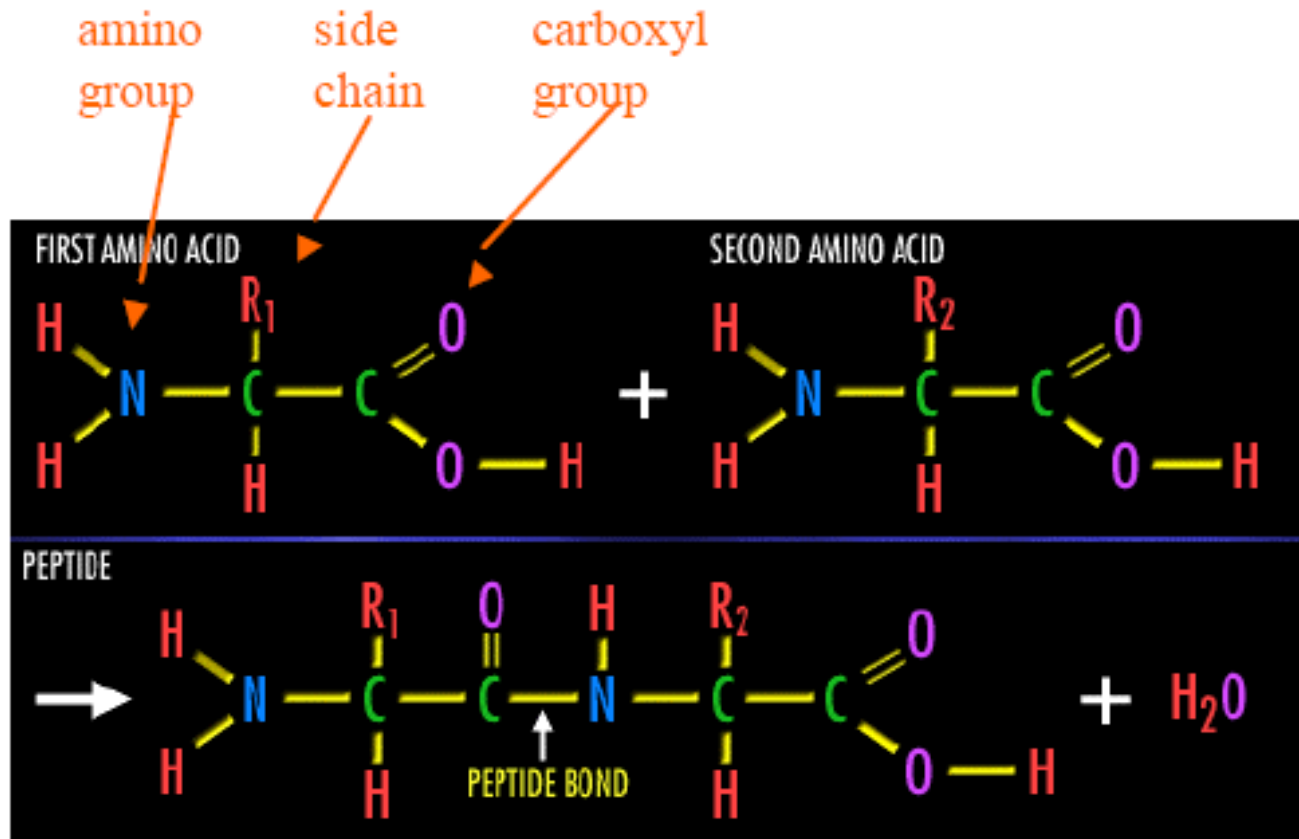


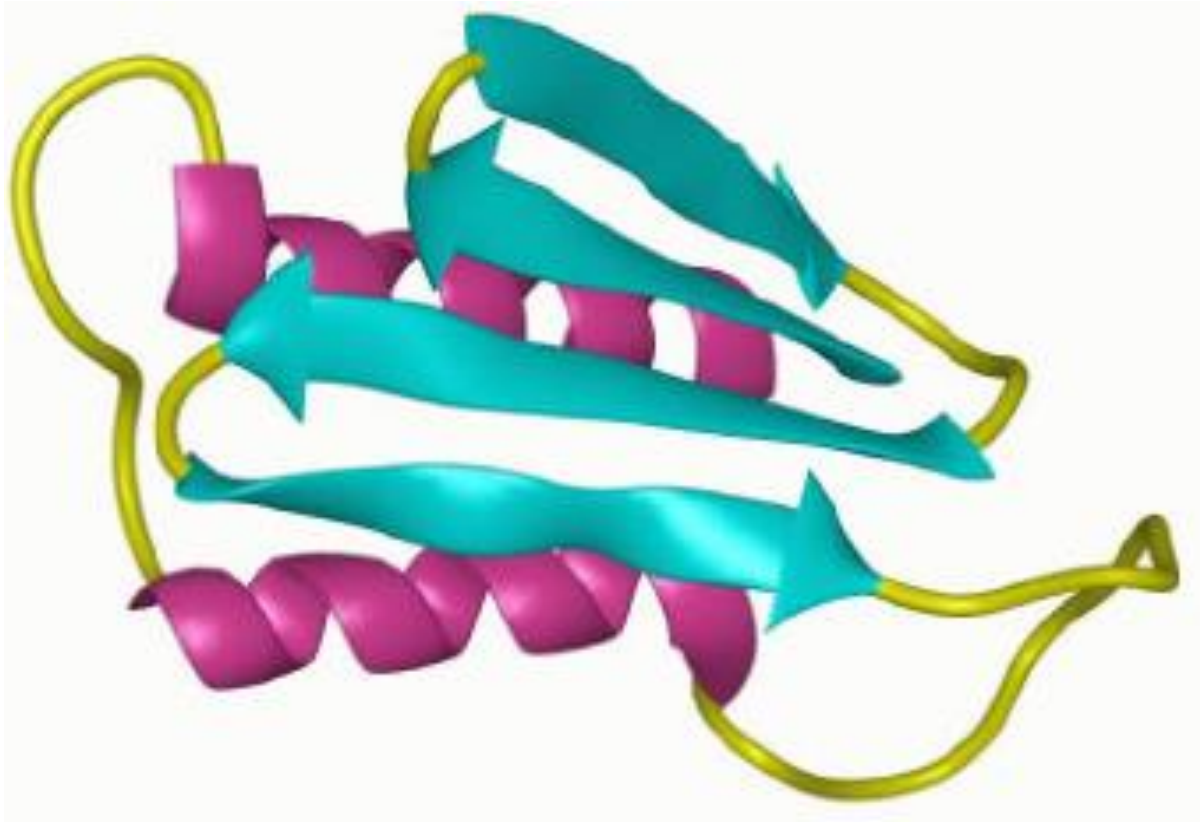
Arginine
(arg)



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:

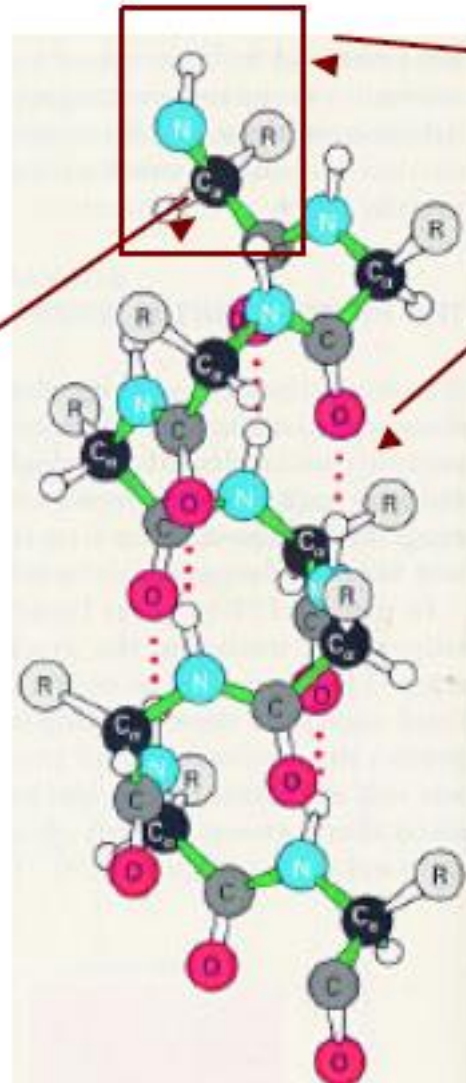
1. [All alpha proteins](#) [46456] (226)  
2. [All beta proteins](#) [48724] (149)  
3. [Alpha and beta proteins \(a/b\)](#) [51349] (134)  
Mainly parallel beta sheets (beta-alpha-beta units)
4. [Alpha and beta proteins \(a+b\)](#) [53931] (286)  
Mainly antiparallel beta sheets (segregated alpha and 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 class

α Helices

α carbon

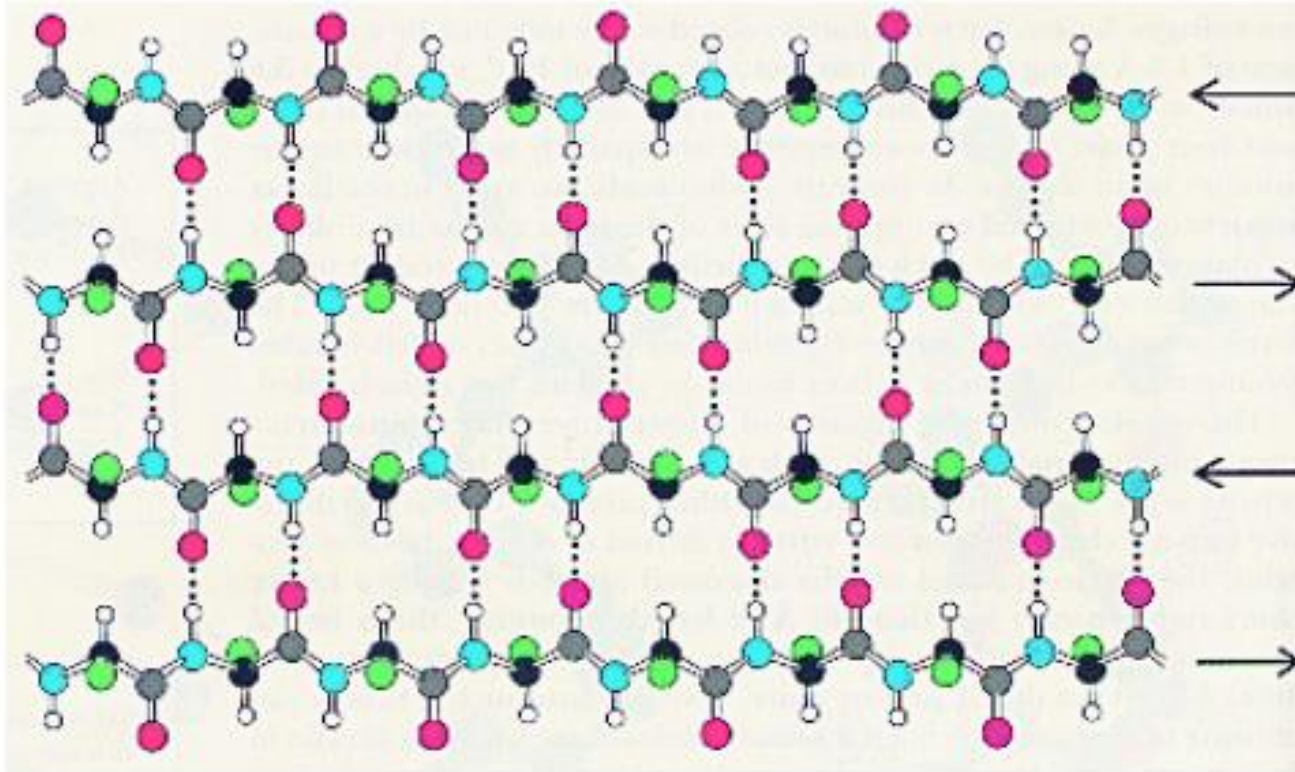
individual amino acid

hydrogen bond

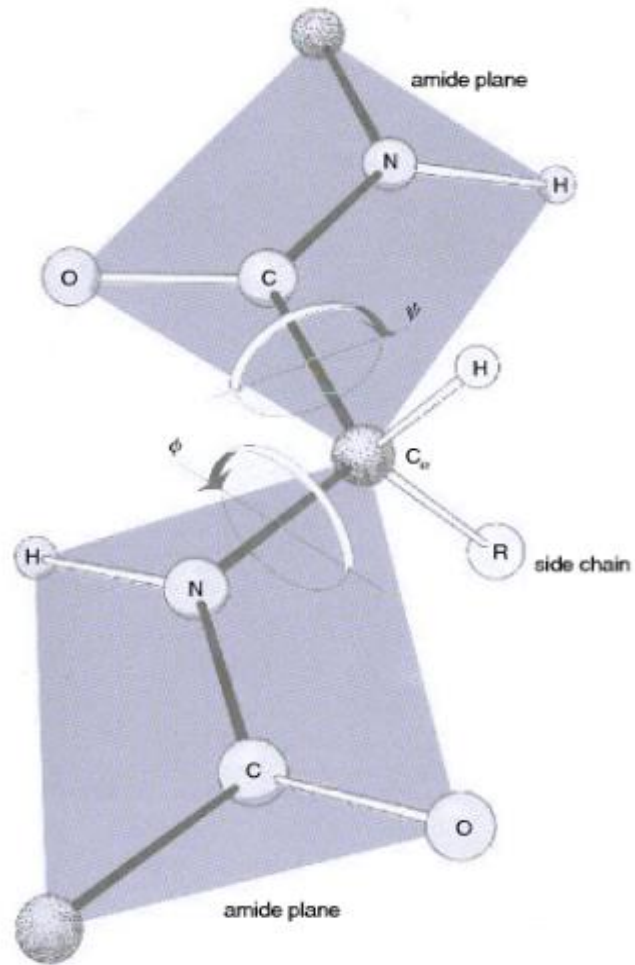


- Helix complete turn every 3.6 AAs
- Hydrogen bond between (-C=O) of one AA and (-N-H) of its 4th neighboring AA

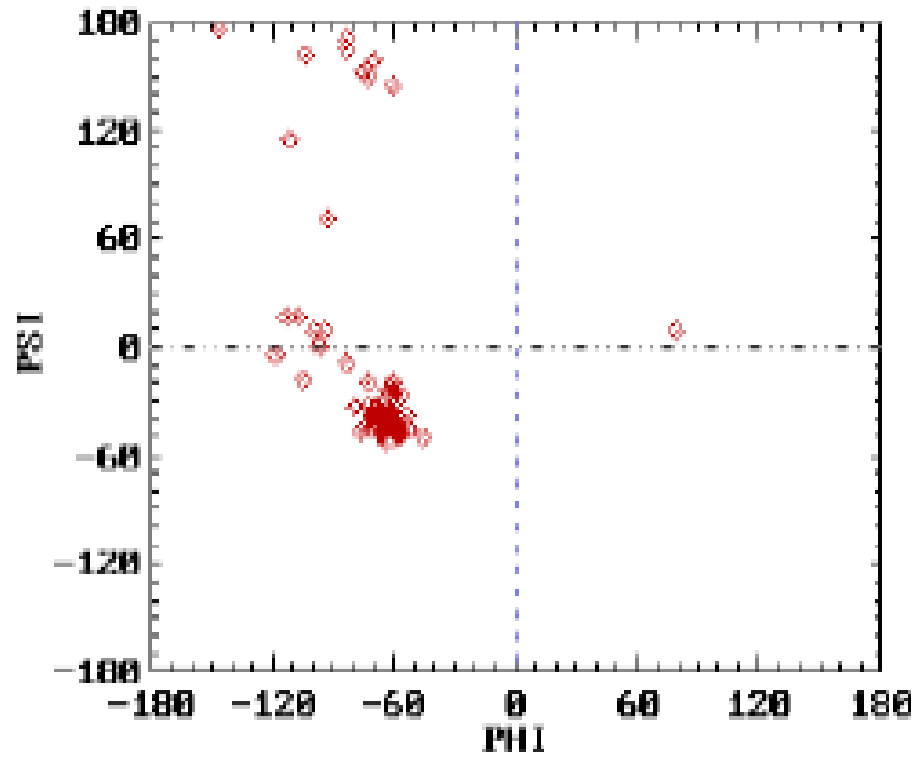
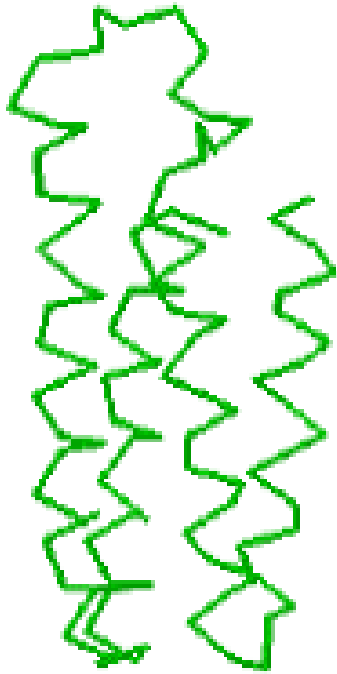
β Strands



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

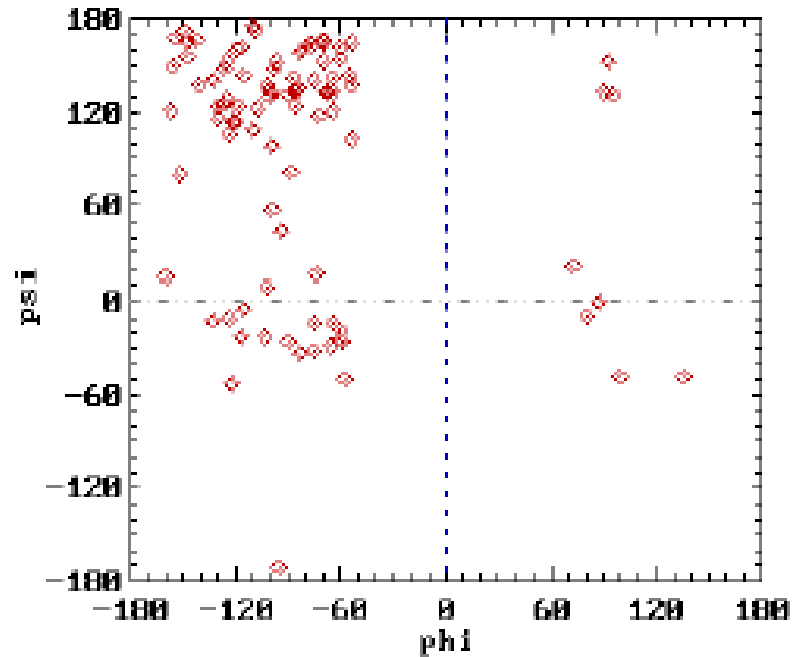
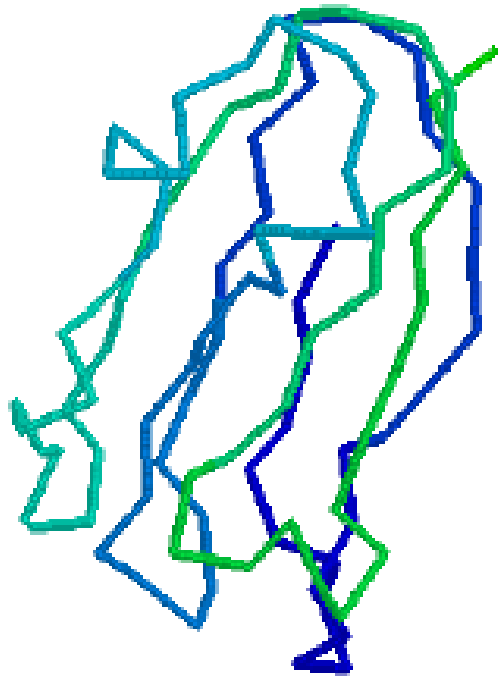


Ramachandran Plot



PHI: -57; PSI -47

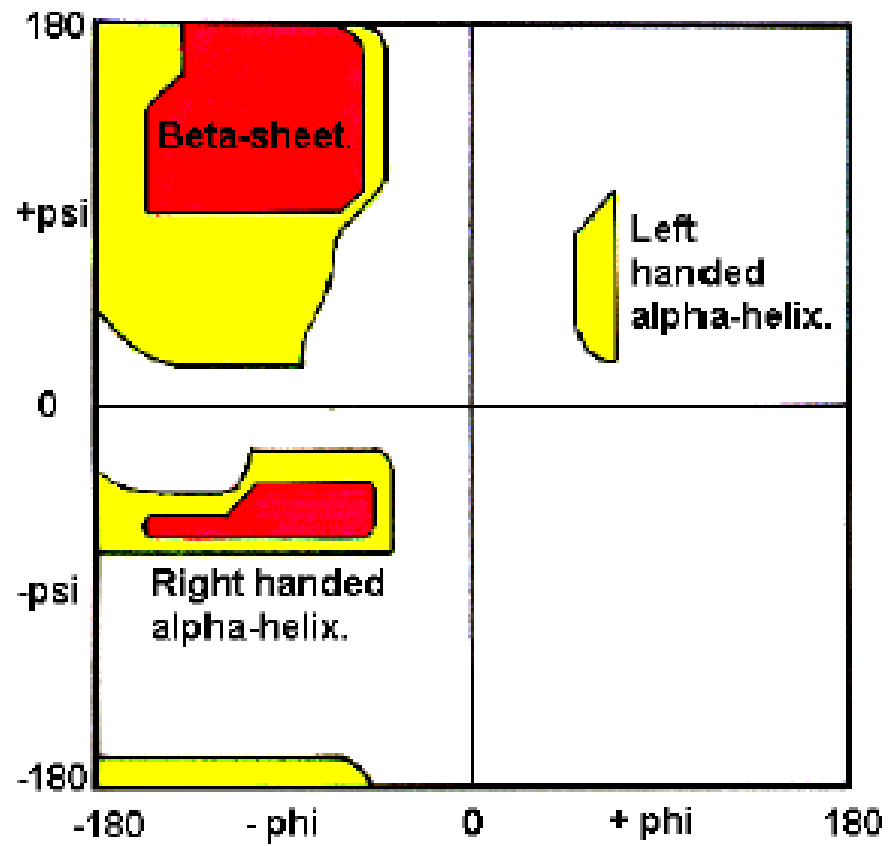
Ramachandran Plot



Parallel: PHI: -119; PSI: 113

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

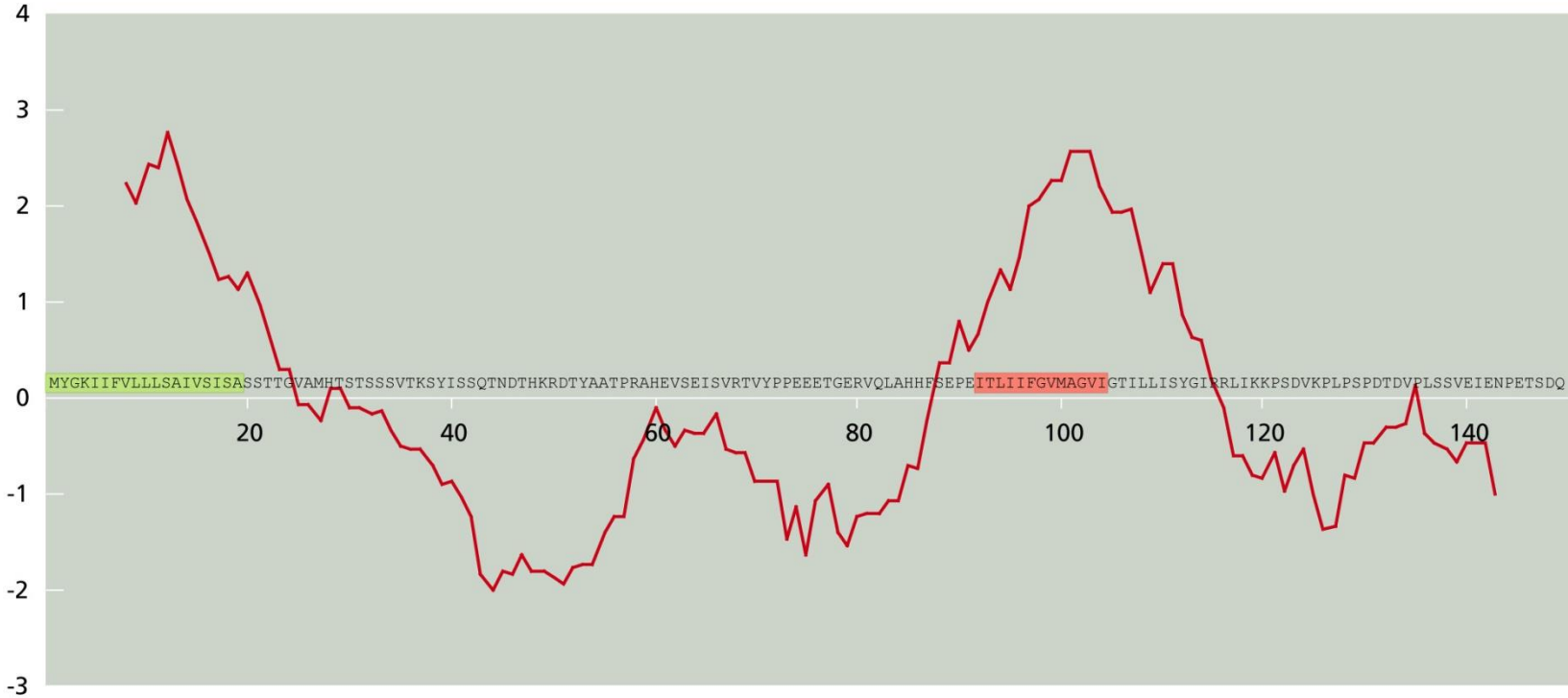
The Ramachandran Plot.



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



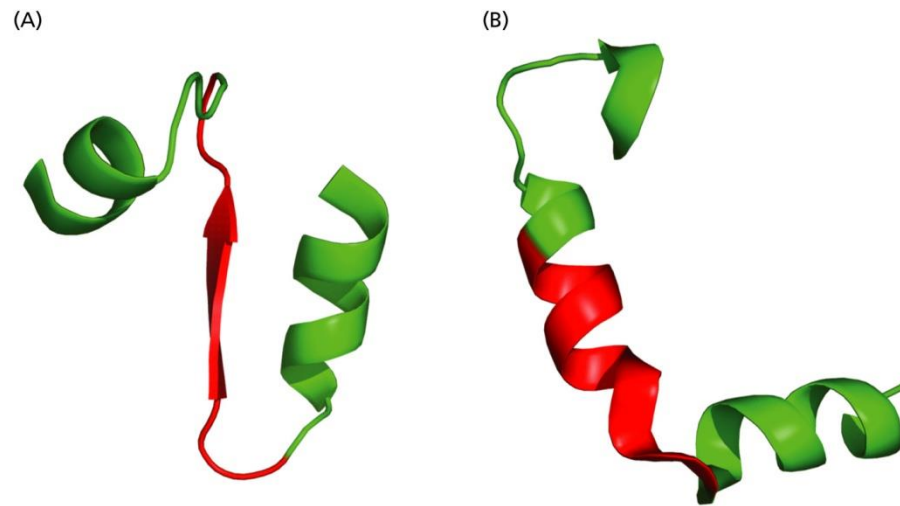
Residue conformation preferences

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

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

Coil: D, G, N, P, S

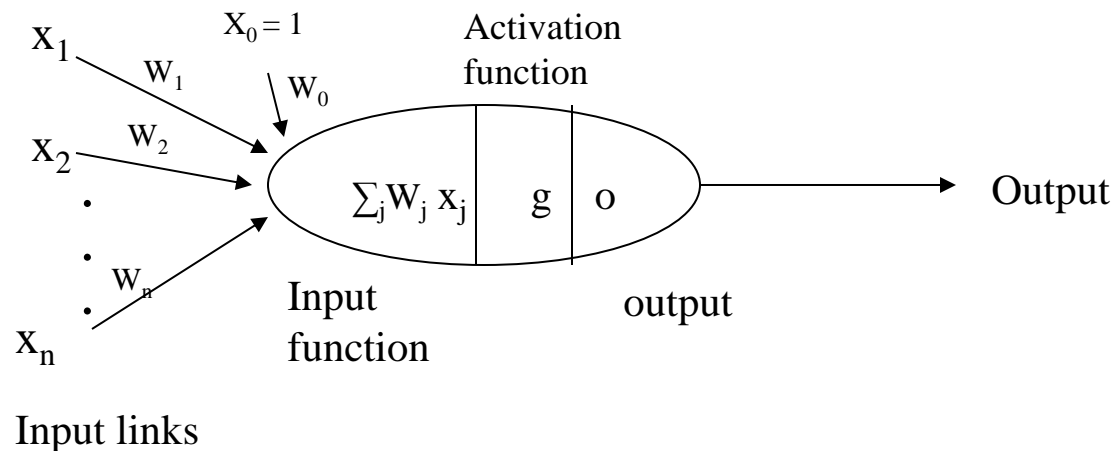
Structures are modulated by nearby sequence



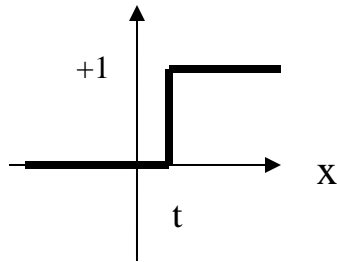
The nine-residue sequence KGVVPQLVK (in red) occurs in two proteins (1IAL and 1PKY) but with completely different structures. (Fig. 12.12)

Artificial neural networks

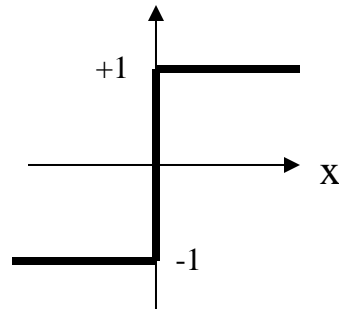
- Perceptron $o(x_1, \dots, x_n) = g(\sum_j W_j x_j)$



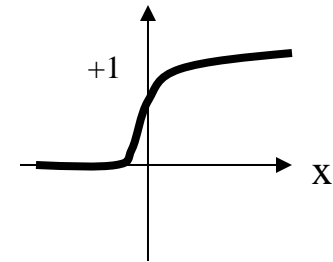
- Activation functions



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



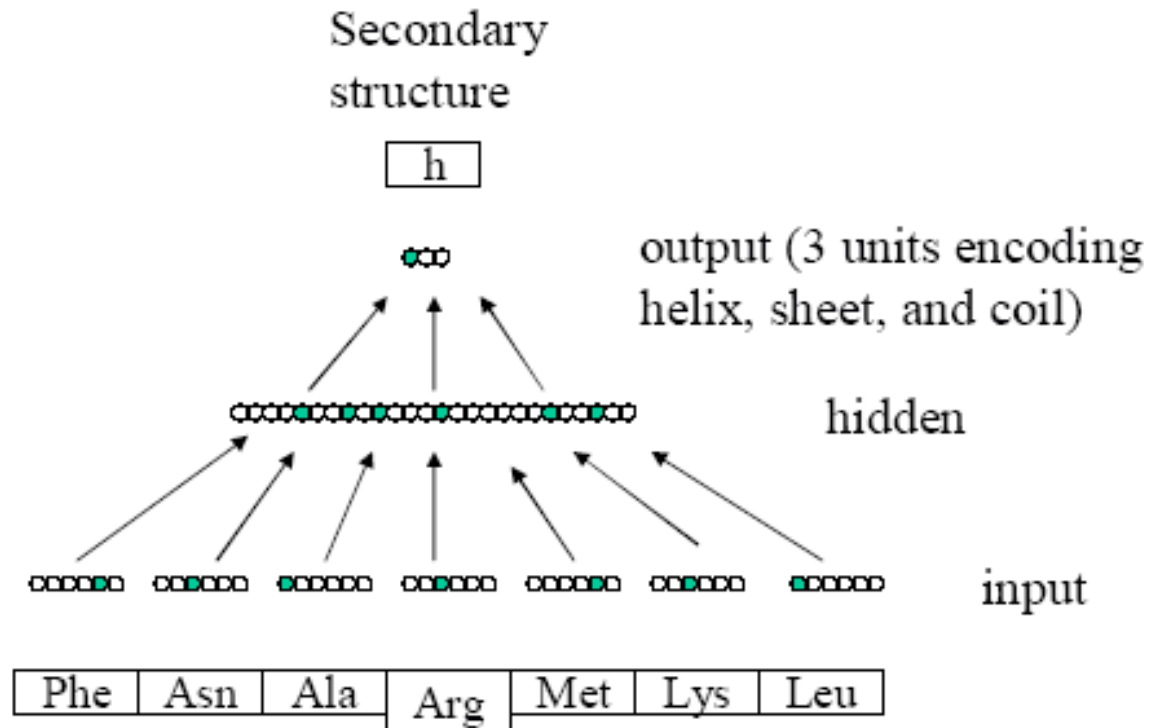
$$\text{Sign}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{otherwise} \end{cases}$$



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

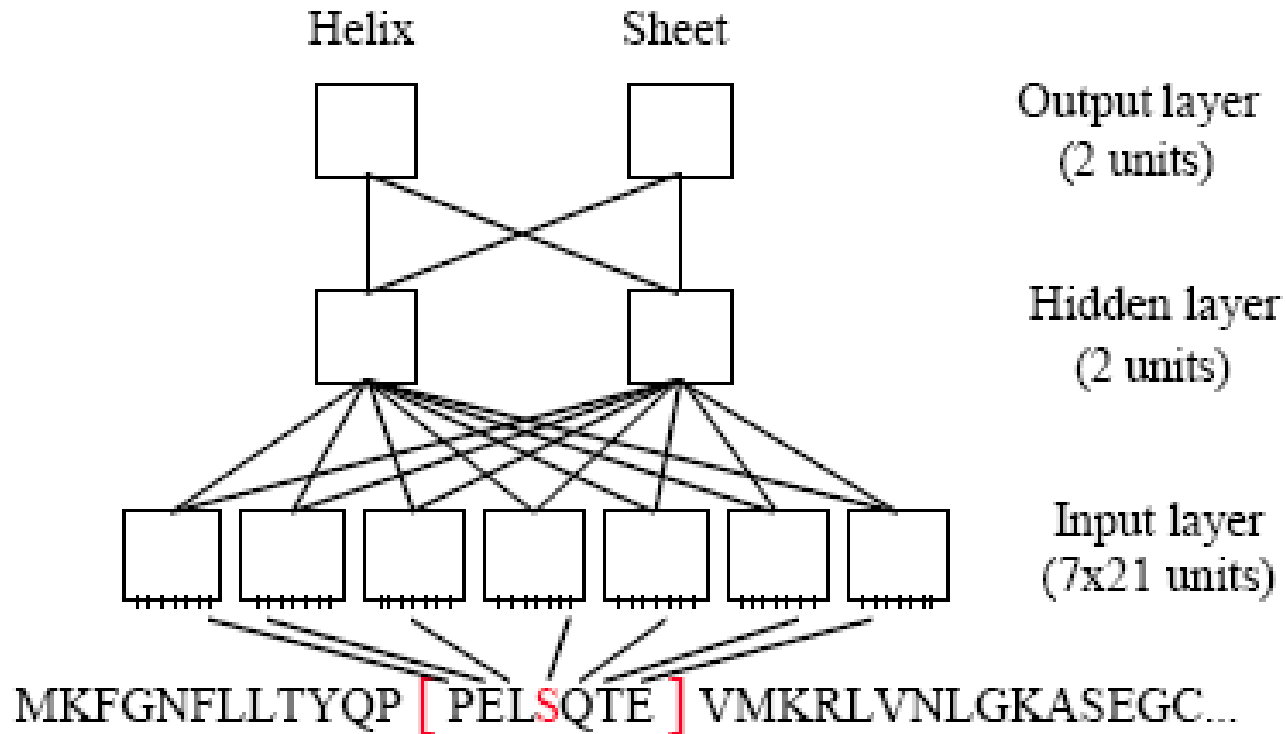
Artificial Neural Networks

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

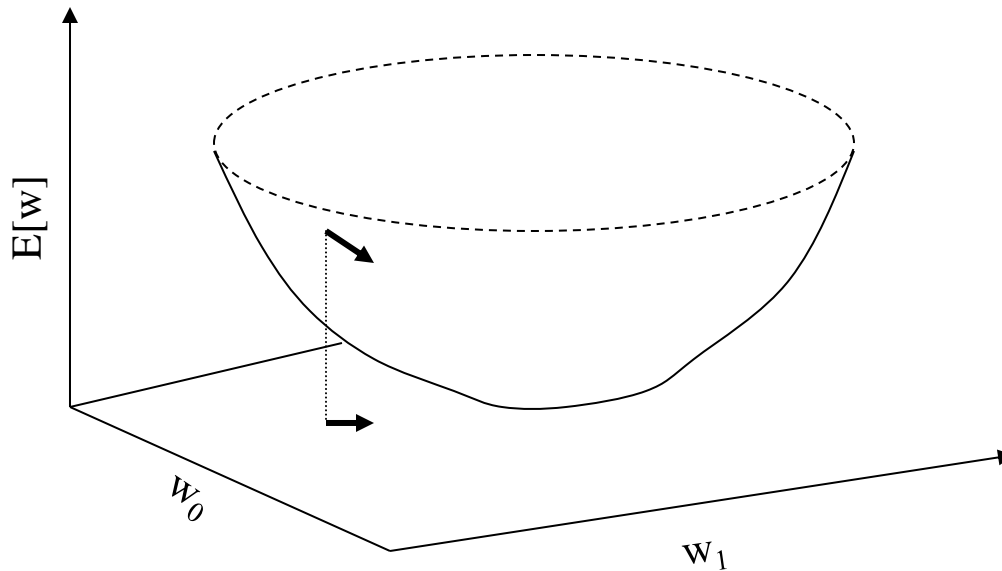


- Learning: to determine weights and thresholds for all nodes (neurons) so that the net can approximate the training data within error range.
 - Back-propagation algorithm
 - Feedforward from Input to output
 - Calculate and back-propagate the error (which is the difference between the network output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

Gradient descent

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

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
GGGWGQPHGGSWGQPHGGSWGQPHGGGGWGQGGGTHNQWNKPSKPKTNLKHVAGAAAAGAV
VGGLGGYMLGSAMSRPMIHFGNDWEDRYRENMYRYPNQVYYPVDQYSNQNNFVHDCVN
ITIKQHTVTTTTKGENFTETDVKMMERVVEQMCVTQYQKESQAYYDGRRSSTVLFSSPP
VILLISFLIFLIVG
```

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

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

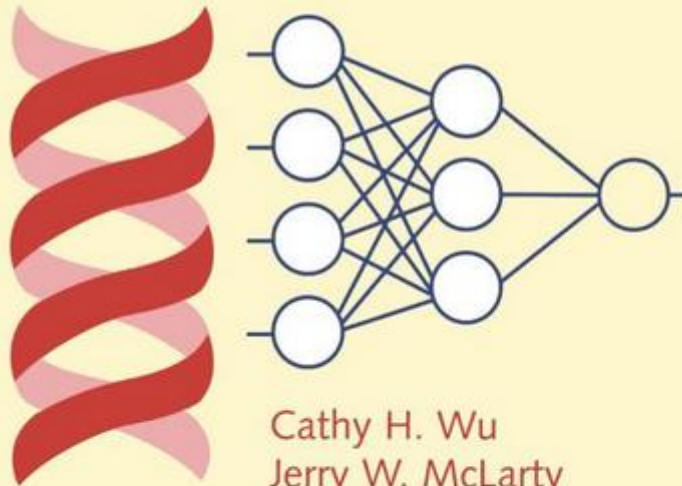
- Performance
 - ceiling at about 65% for direct encoding
 - Local encoding schemes present limited correlation information between residues
 - Little or no improvement using multiple hidden layers.
 - Surpassing 70% by
 - Including evolutionary information (contained in multiple alignment)
 - Using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

Copyrighted Material
METHODS IN COMPUTATIONAL
BIOLOGY AND BIOCHEMISTRY

Volume
1

SERIES EDITOR: A.K. KONOPKA

Neural Networks and Genome Informatics



ELSEVIER

Copyrighted Material

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

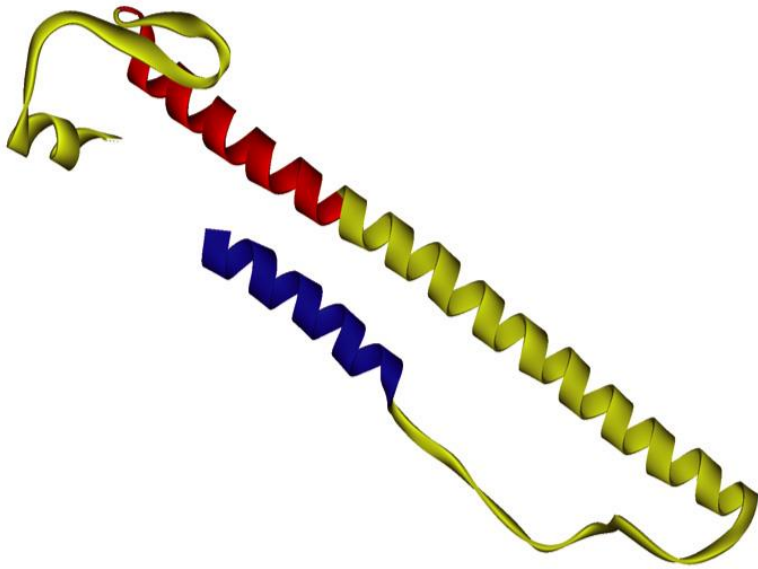
Reference	Application	Neural network*	I/O encoding†
Intron/Exon (I/E) Discrimination and Gene Identification			
Uberbacher and Mural, 1991	Coding region recognition	4L/FF/BP	FEAT7/1(Y,N)
Uberbacher <i>et al.</i> , 1996	Coding region recognition	3L/FF/BP	FEAT13/1(Y,N)
Snyder and Stormo, 1993	I/E feature weighting	2L/FF/Delta	FEAT6/1(inequality)
Snyder and Stormo, 1995	I/E feature weighting	2,3L/FF/Delta,BP	FEAT6/1(inequality)
Brunak <i>et al.</i> , 1991	Splicing donor/acceptor site prediction	3L/FF/BP	BIN4/1(Y,N)
Farber <i>et al.</i> , 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 <i>et al.</i> , 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 <i>et al.</i> , 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 <i>et al.</i> , 1993	<i>E. coli</i> promoter prediction	3L/FF/BP	BIN4/1(Y,N)
Demeler and Zhou, 1991	<i>E. coli</i> promoter prediction	3L/FF/BP	BIN2,BIN4/1(Y,N)
O'Neill, 1991, 1992	<i>E. coli</i> promoter prediction	3L/FF/BP	BIN4/1(Y,N)
Horton and Kanehisa, 1992	<i>E. coli</i> promoter prediction	2L/FF/BP	BIN4 + 3 + FREQ/1(Y,N)
Mahadevan and Ghosh, 1994	<i>E. coli</i> 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 <i>et al.</i> , 1995	Eukaryotic promoter prediction	3L/FF/BP	BIN4/1(Y,N)
Matis <i>et al.</i> , 1996	RNA polymerase II binding site prediction	4L/FF/BP	FEAT13/1(Y,N)
Nair <i>et al.</i> , 1994	Prediction of transcriptional terminator	3L/FF/BP	BIN4,REAL1/1(Y,N)
Nair <i>et al.</i> , 1995	Prediction of transcription control signal	3L/FF/BP	BIN4/1(RTL)
DNA/RNA Sequence Analysis, Phylogenetic Classification and Code Mapping			
Arrigo <i>et al.</i> , 1991	Clustering and functional region identification	2L/Kohonen	REAL1/Map(30)
Giuliano <i>et al.</i> , 1993	Clustering and functional region identification	2L/Kohonen	REAL1/Map
Leblanc <i>et al.</i> , 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 <i>et al.</i> , 1995	Transfer RNA gene recognition	3L/FF/BP	BIN4/10(Class)
Tolstrup <i>et al.</i> , 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.

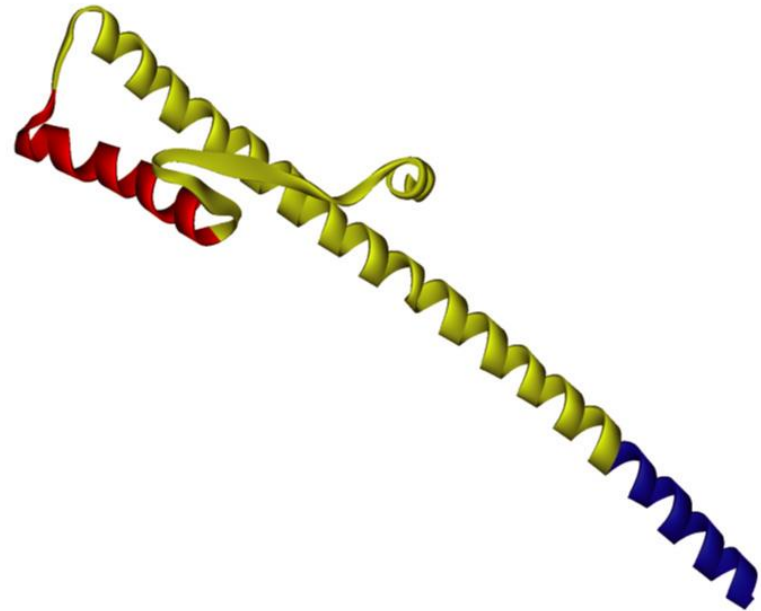
Environmental effects

(A)



3HMG

(B)



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/>

Foundation

- Anfinsen Hypothesis: A native state of protein corresponds to a free energy minimum. (This hypothesis was by Anfinsen based on some experimental findings for smaller molecules, 1973)
- Levinthal paradox: how can a protein fold to a native conformation so rapidly when the number of possible conformations is so huge?
 - Experimental observation: proteins fold into native conformations in a few seconds
 - If a local move by diffusion takes 10^{-11} seconds, it would take 10^{57} seconds to try $\sim 5^{100}$ possible conformations for a protein of 100 AAs.
- NP-completeness

Computational Methods for 3-D structures

- Comparative (find homologous proteins)
- Threading (recognize folds)
- *Ab initio* (Molecular dynamics)

Root Mean Square Deviation (RMSD)

$$\sqrt{\frac{\sum_{i=1, N} (\bar{x}_i - \bar{y}_i)^2}{N}}$$

Where x_i are the coordinates from molecule 1 and y_i are the *equivalent** coordinates from molecule 2.

*Which atoms are *equivalent* is based on an **alignment**.

Credit: Chris Bystroff @ RPI

Comparative (homology based) modeling

- Identification of structurally conserved regions (using multiple alignment)
- Backbone construction
- Loop construction
- Side-chain restoration
- Structure verification and evaluation
- Structure refinement (energy minimization)

Least squares superposition

Problem: find the rotation matrix, \underline{M} , and a vector, \underline{v} , that minimize the following quantity:

$$\sum_i \left| \underline{M} \vec{x}_i + \underline{v} - \vec{y}_i \right|^2$$

Where x_i are the coordinates from one molecule and y_i are the *equivalent** coordinates from another molecule.

**equivalent* based on alignment

Credit: Chris Bystroff @ RPI

Mapping structural equivalence = aligning the sequence

Any position that is aligned is included in the sum of squares.

```
4DFR:A ISLIAALAVDRVIGMENAMPWNLPA DLAWFKRNTLDKPVIMGRHTWESIG-RPLPGRKNI
1DFR:_ TAFLWAQNRNGLIGKDGHL PWHL PDDLHYFRAQTVGKIMVVGRRTYESFPKRPLPERTNV

4DFR:A ILSSQ-PGTDDRVTWVKSVD EAI AAC--GDVPEIMVIGGGRVYEQFLPKAQKLYLTHIDA
1DFR:_ VLTHQEDYQAQGAVVVHDVA AVFAYAKQHLDQELVIAGGAQIFTAFKDDVDTL LVTRLAG

4DFR:A EVEGDTHFPDYEPDDWESV FSEPHDADAQNS--HSYCFKILERR
1DFR:_ SFEGDTKMIPLNWDDPTKVSSRTVEDT---NPALHTTYEVWQKK
```

Unaligned positions are not.

Credit: Chris Bystroff @ RPI

least-squares superimposed molecules



Credit: Chris Bystroff @ RPI

Lattice Models

Simplifications

- All residues have the same size
- Bond length is uniform
- Positions of residues are restricted to positions in a regular lattice (or grid).

HP model

- Energy function $B_{i,j}$ for a pair of residues w_i and w_j
 - = 0 when the two residues
 - do not have contact (i.e. not topological neighbors), or
 - one residue is polar/hydrophilic and the other is hydrophobic, or
 - both are polar/hydrophilic
 - = -1 when two residues
 - are topological neighbors, and
 - both are hydrophobic.

Note: Despite these simplifications, it is still NP complete to find an optimal conformation

- 2-dimensional and 3-dimensional lattice
- Two residues w_i and w_j are *connected neighbors* when $j = i+1$ or $i-1$.
- Two residues w_i and w_j are *topological neighbors* when $j \neq i+1$ or $i-1$, and

$$\| w_i - w_j \| = 1$$

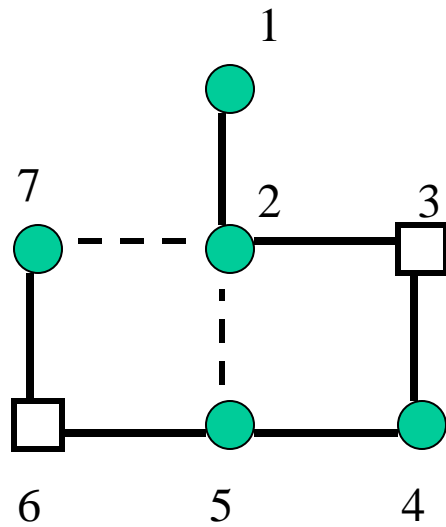
- A native state is a conformation that has the minimum contact energy

$$E = \sum_{1 \leq i+1 < j \leq n} B_{i,j} \delta(w_i, w_j)$$

where $\delta(w_i, w_j) = 1$ when w_i and w_j are *topological neighbor*, and $=0$ otherwise.

- The HP model approximates the hydrophobic force, which is not really a force, but rather an aggregate tendency for nonpolar residues to minimize their contact with the solvent.

- Example HP model



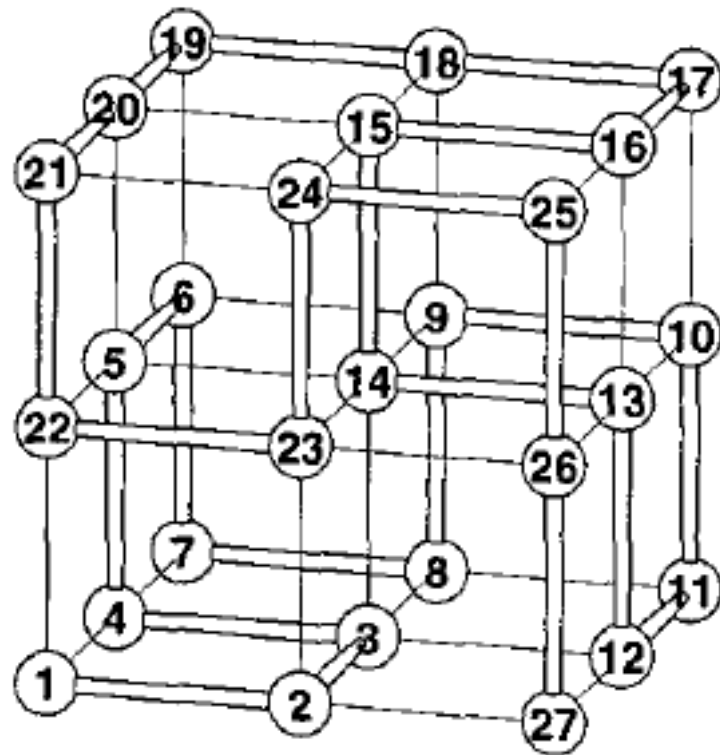
□ P

● H

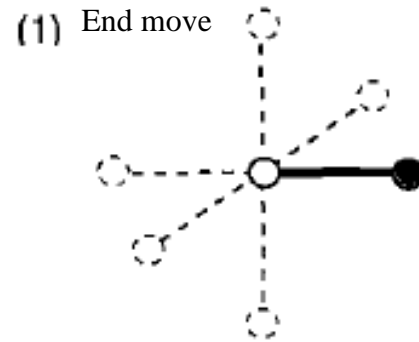
— Peptide bond

- - Topological contact

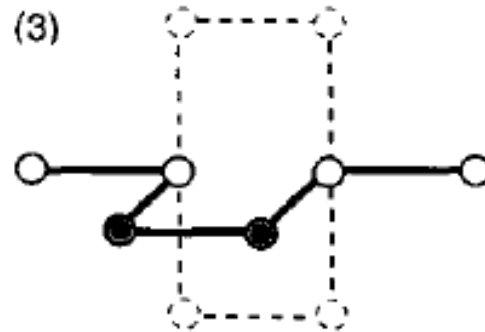
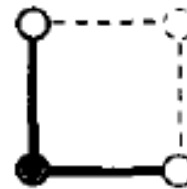
- The HP model by SSK (Sali, Shakhnovich, and Karplus, 1994)
 - A 27-bead heteropolymer in a 3-d lattice.
 - Contact energy normally distributed
 - Possible conformation: $5^{26} \sim 10^{18}$
 - Compact conformation: in a 3x3x3 cube there are 103346 distinct conformations
 - Folding time t_0 is short if energy gap is large
 - Solved Levinthal paradox



Local moves



(2) Corner move



Crankshaft move

(4)

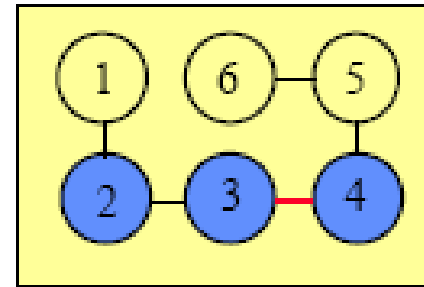
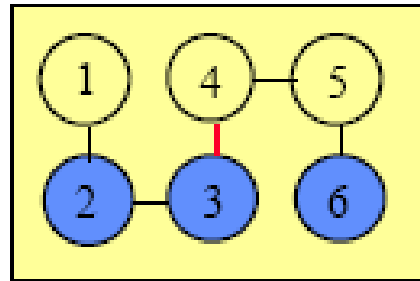


Impossible move

Crossover to create new conformations

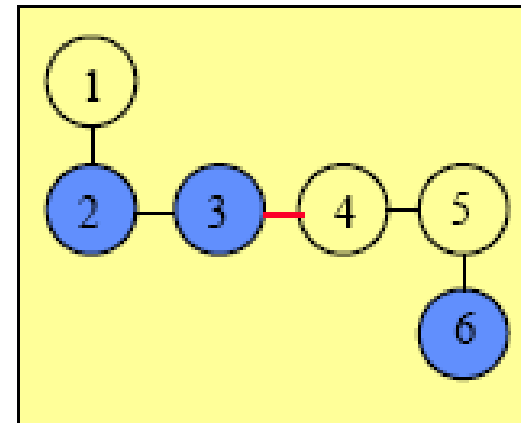
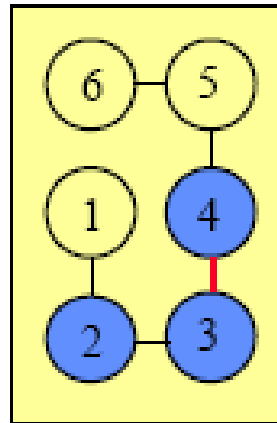
Parents

10 00 01 00 10
 10 00 00 01 11

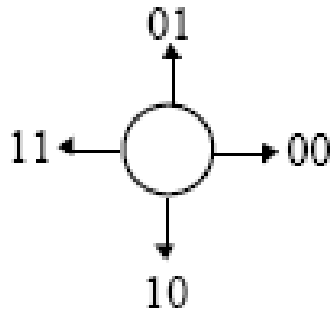


Children

10 00 01 01 11
 10 00 00 00 10



Encoding:



Credit: Iosif Vaisman @ GAU

Genetic algorithm

Input

- P, the population,
- r: the fraction of population to be replaced,
- f, a fitness,
- ft, the fitness_threshold,
- m: the rate for mutation.

Initialize population (randomly)

Evaluate: for each h in P, compute Fitness(h)

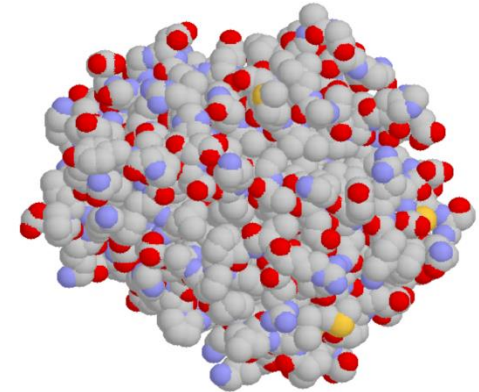
While [$\text{Max}_h f(h)$] < ft

do

1. Select
2. Crossover
3. Mutate
4. Update P with the new generation Ps
5. Evaluate: f(h) for all h \in P

Return the h in P that has the best fitness

Ab initio approaches



Molecular Dynamics

$$F_i = m_i a_i$$

$$a_i = dv_i / dt$$

$$v_i = dr_i / dt$$

$$- dE / dr_i = F_i$$

$$- dE / dr_i = m_i d^2r_i / dt^2$$

Force fields

$$E(r^N) = \sum_{\text{bonds}} k_b(l - l_0)^2 + \sum_{\text{angles}} k_a(\theta - \theta_0)^2$$
$$+ \sum_{\text{torsions}} \sum_{N-1}^n \frac{1}{2} V_n [1 + \cos(n\omega - \gamma)]$$
$$+ \sum_{j=1} \sum_{i=j+1}^N f_{ij} \left\{ \epsilon_{ij} \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$$

Popular force fields: AMBER, CHARMM, CVFF, GROMOS



The Nobel Prize in Chemistry 2013

Martin Karplus, Michael Levitt, Arieh Warshel

Share this: [f](#) [g+](#) [t](#) [+](#) 883 [e](#)

The Nobel Prize in Chemistry 2013



Photo: A. Mahmoud

Martin Karplus

Prize share: 1/3



Photo: A. Mahmoud

Michael Levitt

Prize share: 1/3



Photo: A. Mahmoud

Arieh Warshel

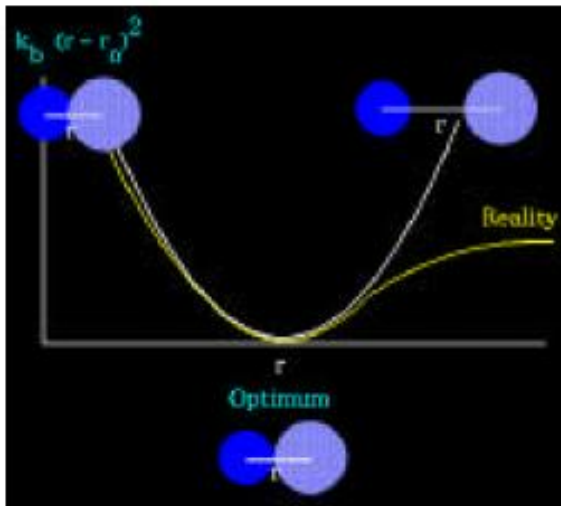
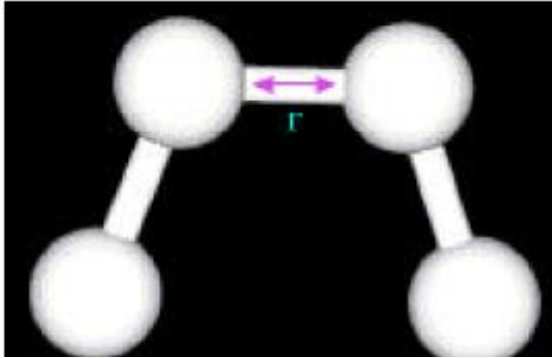
Prize share: 1/3

The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel *"for the development of multiscale models for complex chemical systems"*.

Photos: Copyright © The Nobel Foundation

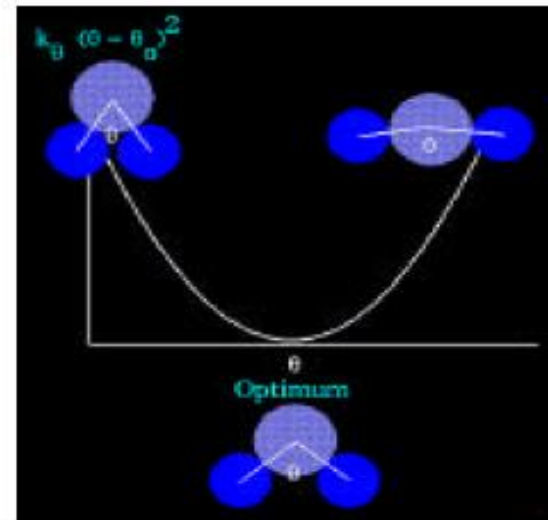
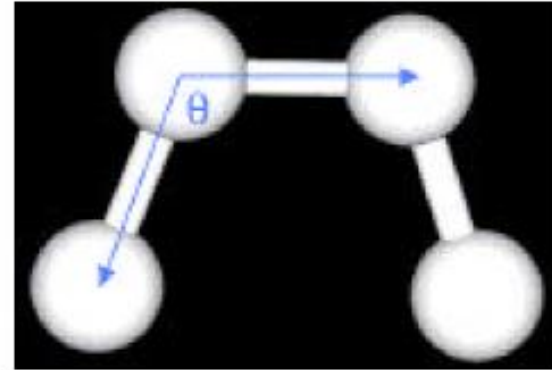
Bond length

$$E = \sum_{\text{bonds}} k_b (r - r_0)^2$$



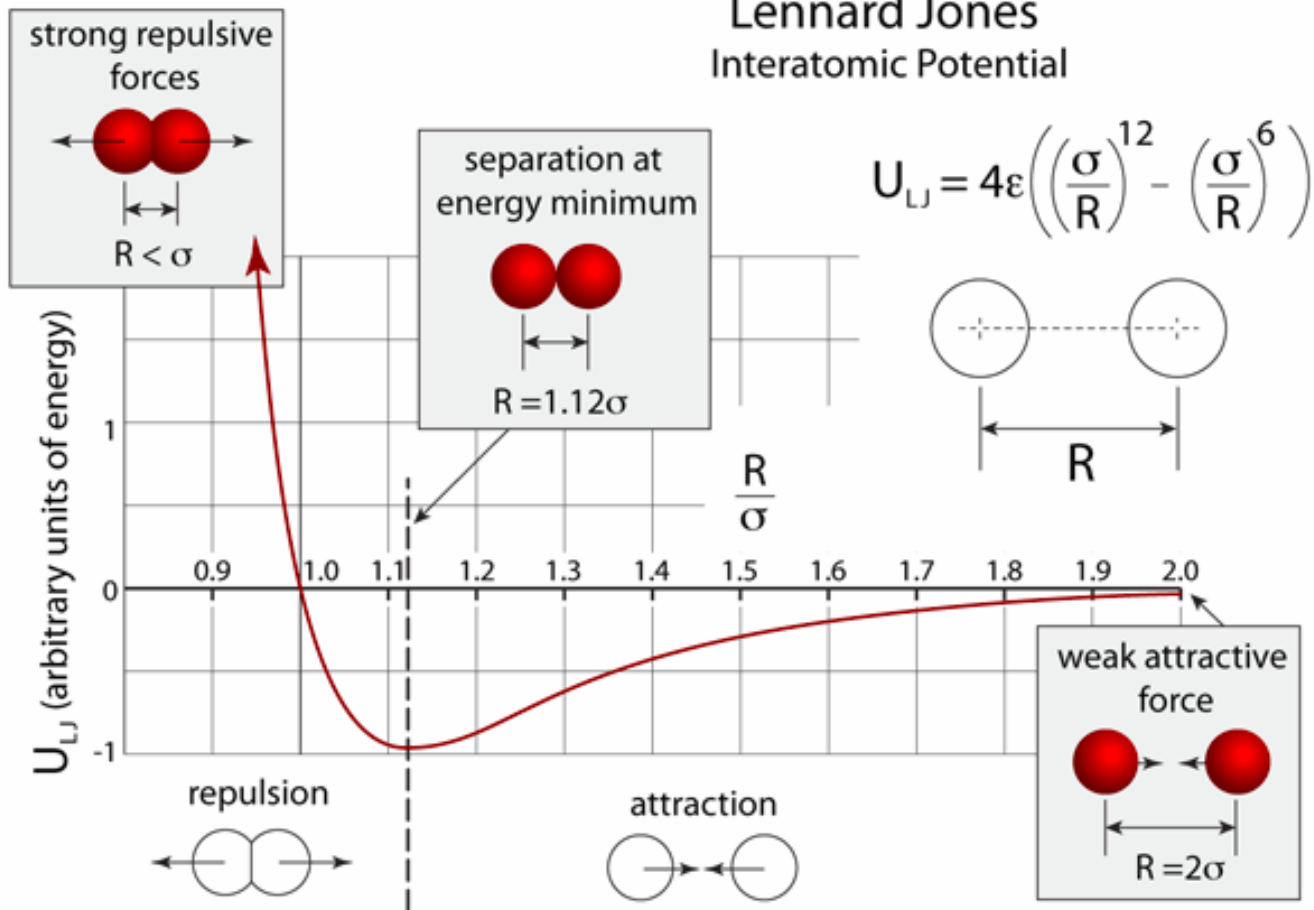
Bond angle

$$E = \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2$$



Credit: Iosif Vaisman @ GMU

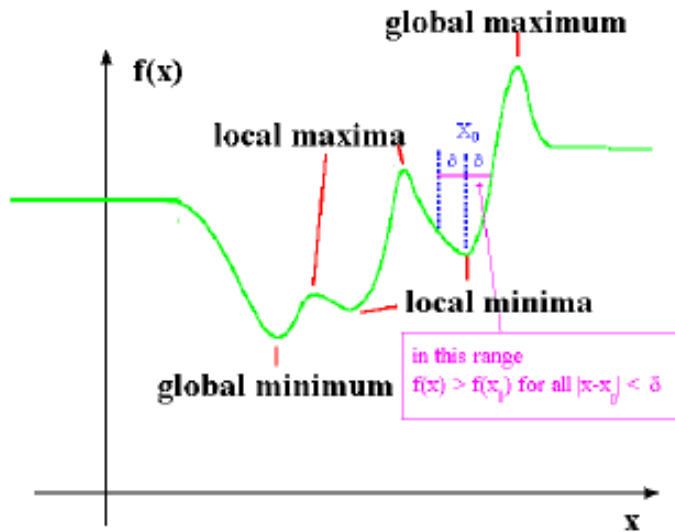
Lennard Jones Interatomic Potential



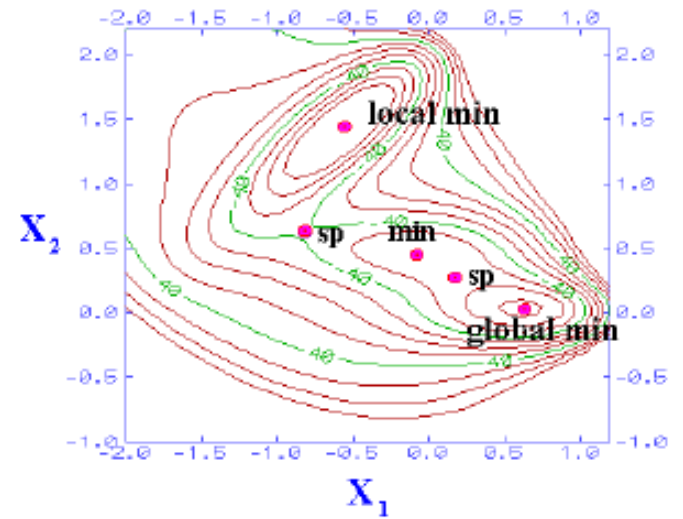
Lennard Jones Potential. The graph above plots the Lennard–Jones potential function, and indicates regions of attraction and repulsion. Atoms try to minimize their potential energy and at the lowest temperatures are sitting at the bottom of the potential curve. When the atomic separations are to the left of the minimum the atoms repel, otherwise they attract one another.

Credit: atomsinmotion.com

Energy Minimization

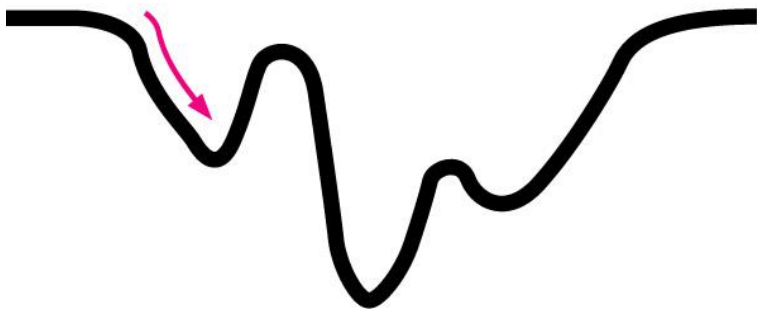


Energy Minimization

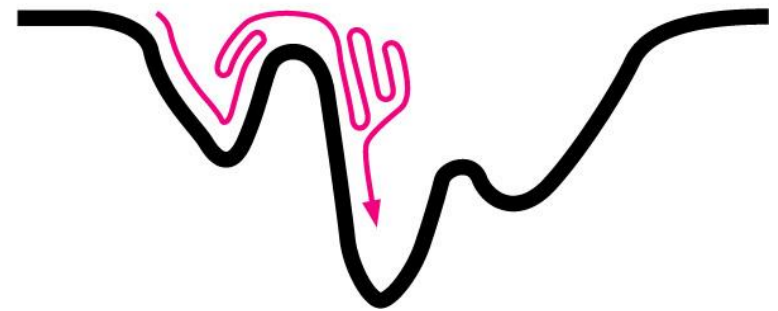


Credit: Iosif Vaisman @ GMU

(A)



(B)



Molecular Dynamics

$$F_i = m_i a_i$$

$$a_i = dv_i / dt$$

$$v_i = dr_i / dt$$

$$- dE / dr_i = F_i$$

$$- dE / dr_i = m_i d^2r_i / dt^2$$

Credit: Iosif Vaisman @ GMU

Resources

Homology modeling programs

<http://www.expasy.ch/swissmod>

Threading:

<http://www.ncbi.nlm.nih.gov/Structure/RESEARCH/threading.shtml>

Ab initio

<https://www.rosettacommons.org/>