Statistical Machine Learning Methods for Bioinformatics

IV. Neural Network Applications in Bioinformatics

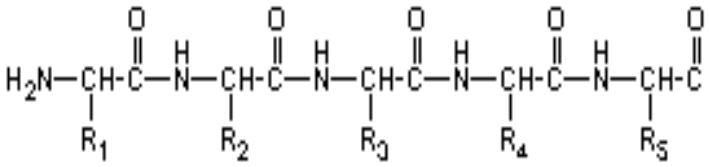
Jianlin Cheng, PhD
Department of Computer Science
University of Missouri, Columbia
2012

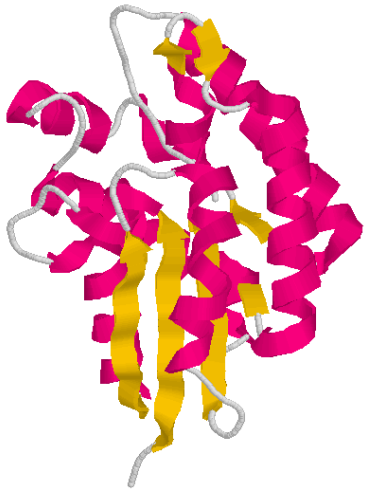
Neural Network Application in Bioinformatics

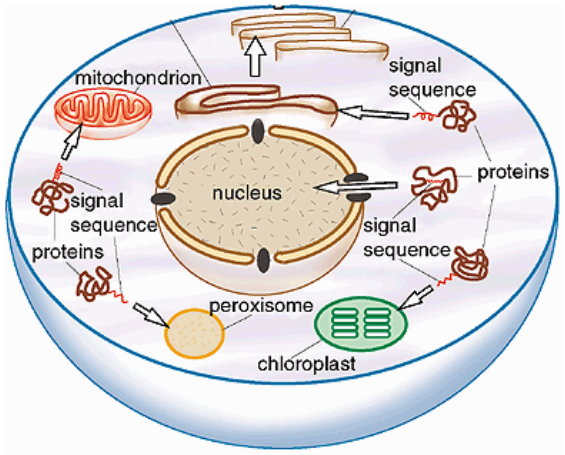
- Neural network is one of the most widely used methods in bioinformatics.
- It is used in gene structure prediction, protein structure prediction, gene expression data analysis, … Almost anywhere when you need to do classification.
- Here we specifically focus on applying neural networks to protein structure prediction (**secondary structure**, solvent accessibility, disorder region, contact map).

Sequence, Structure and Function

AGCWY……



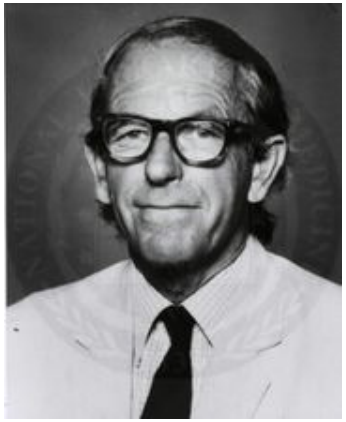


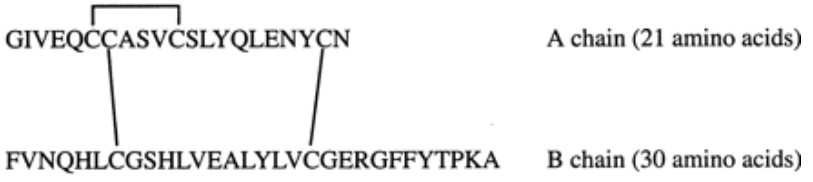


Cell

Protein Sequence – Primary Structure

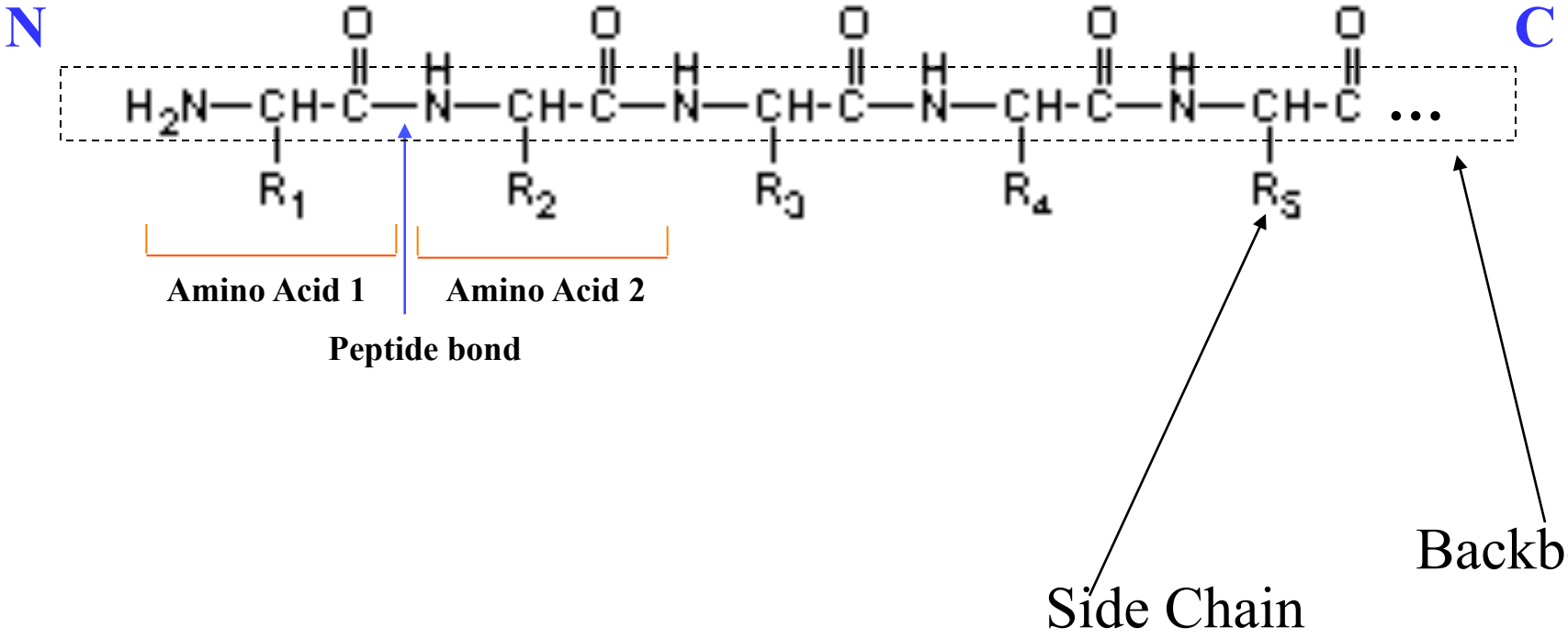
- The first protein was sequenced by Frederick Sanger in 1953.
- Twice Nobel Laureate (1958, 1980) (other: Curie, Pauling, Bardeen).
- Determined the amino acid sequence of insulin and proved proteins have specific primary structure.



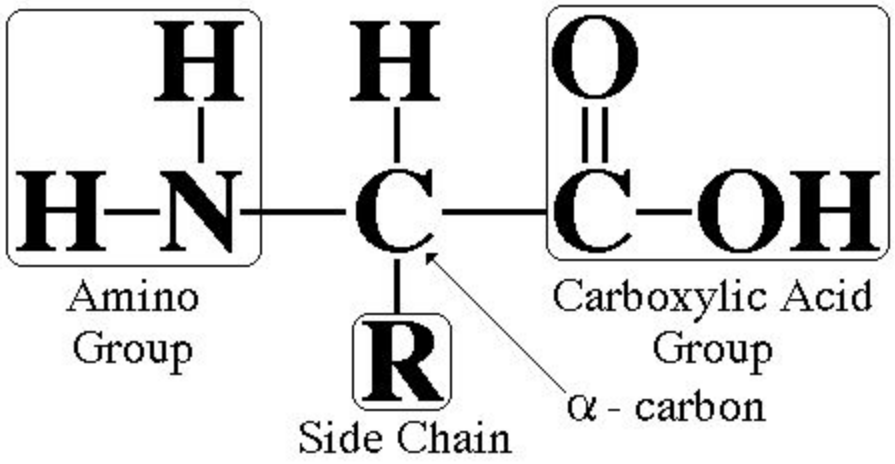


Protein Sequence

A directional sequence of amino acids/residues



Amino Acid Structure



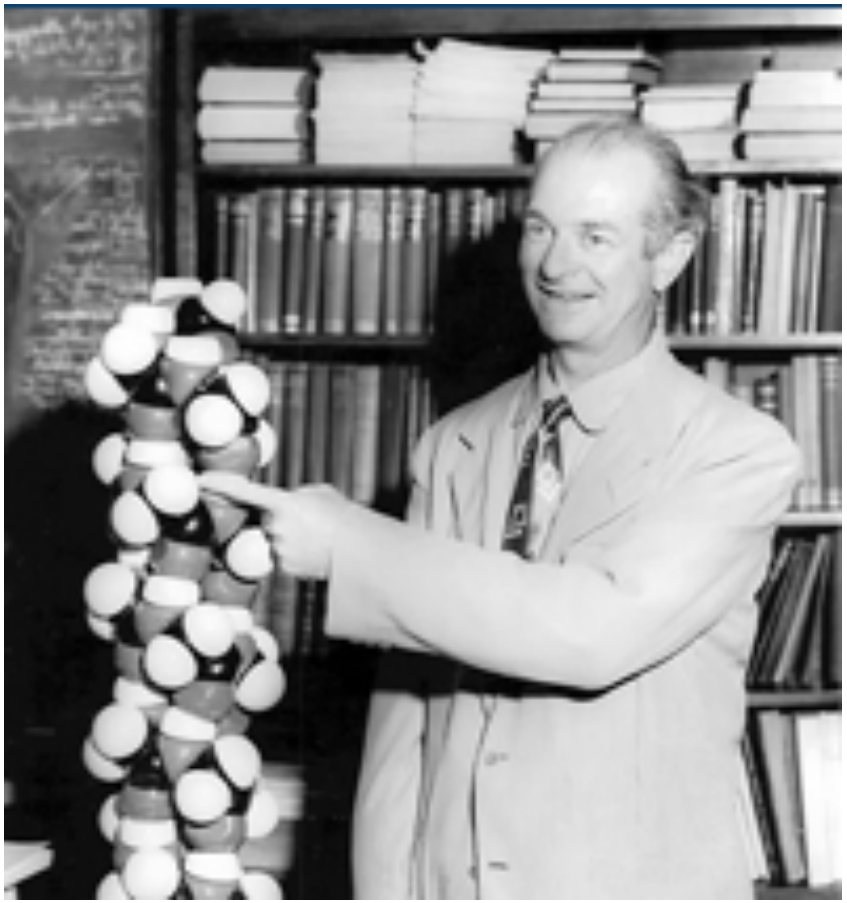
Amino Acids

Amino acid	Abbrev.	Side chain	Hydro-phobic	Polar	Charged	Small	Tiny	Aromatic or Aliphatic	van der Waals volume	Codon	Occurrence in proteins (%)
Alanine	Ala, A	$-CH_3$	X	-	-	X	X	-	67	GCU, GCC, GCA, GCG	7.8
Cysteine	Cys, C	$-CH_2SH$	X	-	-	X	-	-	86	UGU, UGC	1.9
Aspartate	Asp, D	$-CH_2COOH$	-	X	negative	X	-	-	91	GAU, GAC	5.3
Glutamate	Glu, E	$-CH_2CH_2COOH$	-	X	negative	-	-	-	109	GAA, GAG	6.3
Phenylalanine	Phe, F	$-CH_2C_6H_5$	X	-	-	-	-	Aromatic	135	UUU, UUC	3.9
Glycine	Gly, G	-H	X	-	-	X	X	-	48	GGU, GGC, GGA, GGG	7.2
Histidine	His, H	$-CH_2-C_3H_3N_2$	-	X	positive	-	-	Aromatic	118	CAU, CAC	2.3
Isoleucine	Ile, I	$-CH(CH_3)CH_2CH_3$	X	-	-	-	-	Aliphatic	124	AUU, AUC, AUA	5.3
Lysine	Lys, K	$-(CH_2)_4NH_2$	-	X	positive	-	-	-	135	AAA, AAG	5.9
Leucine	Leu, L	$-CH_2CH(CH_3)_2$	X	-	-	-	-	Aliphatic	124	UUA, UUG, CUU, CUC, CUA, CUG	9.1
Methionine	Met, M	$-CH_2CH_2SCH_3$	X	-	-	-	-	-	124	AUG	2.3
Asparagine	Asn, N	$-CH_2CONH_2$	-	X	-	X	-	-	96	AAU, AAC	4.3
Proline	Pro, P	$-CH_2CH_2CH_2-$	X	-	-	X	-	-	90	CCU, CCC, CCA, CCG	5.2
Glutamine	Gln, Q	$-CH_2CH_2CONH_2$	-	X	-	-	-	-	114	CAA, CAG	4.2
Arginine	Arg, R	$-(CH_2)_3NH-C(NH)NH_2$	-	X	positive	-	-	-	148	CGU, CGC, CGA, CGG, AGA, AGG	5.1
Serine	Ser, S	$-CH_2OH$	-	X	-	X	X	-	73	UCU, UCC, UCA, UCG, AGU,AGC	6.8
Threonine	Thr, T	$-CH(OH)CH_3$	X	X	-	X	-	-	93	ACU, ACC, ACA, ACG	5.9
Valine	Val, V	$-CH(CH_3)_2$	X	-	-	X	-	Aliphatic	105	GUU, GUC, GUA, GUG	6.6
Tryptophan	Trp, W	$-CH_2C_8H_6N$	X	-	-	-	-	Aromatic	163	UGG	1.4
Tyrosine	Tyr, Y	$-CH_2-C_6H_4OH$	X	X	-	-	-	Aromatic	141	UAU, UAC	3.2

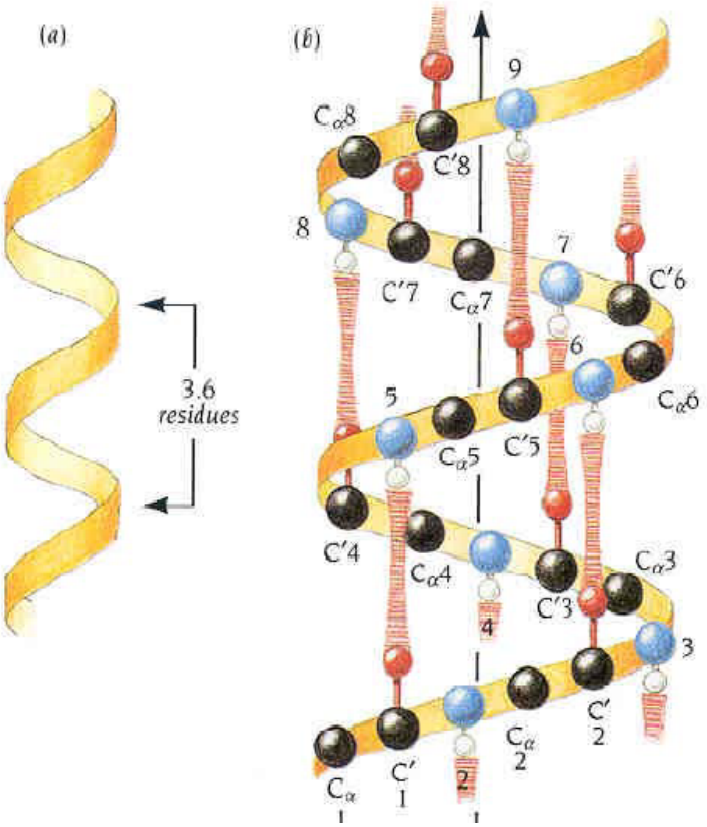
Hydrophilic

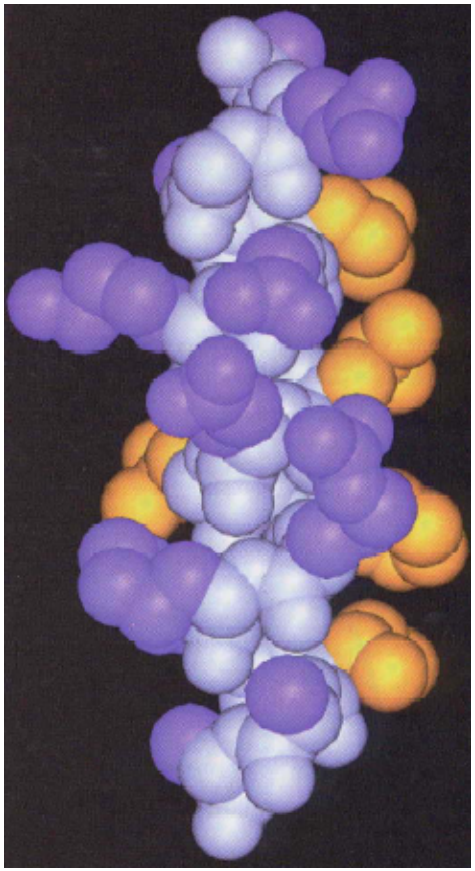
Protein Secondary Structure

- Determined by hydrogen bond patterns
- 3-Class categories: alpha-helix, beta-sheet, loop (or coil)
- First deduced by Linus Pauling et al.



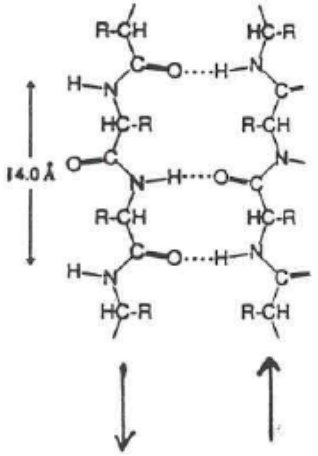
Alpha-Helix

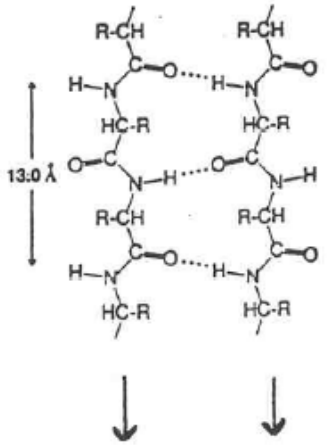


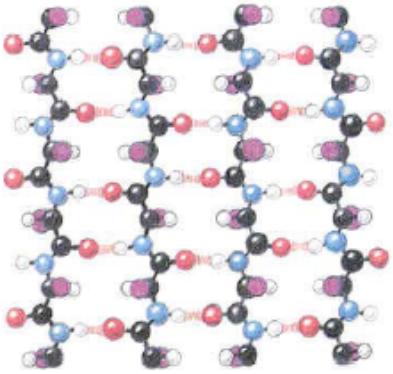


Jurnak, 2003

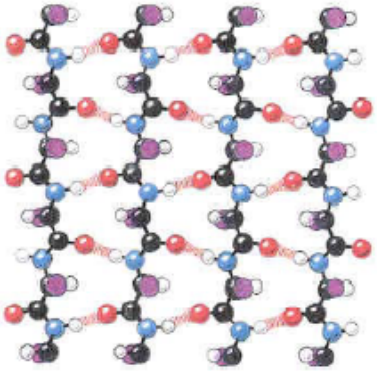
Beta-Sheet







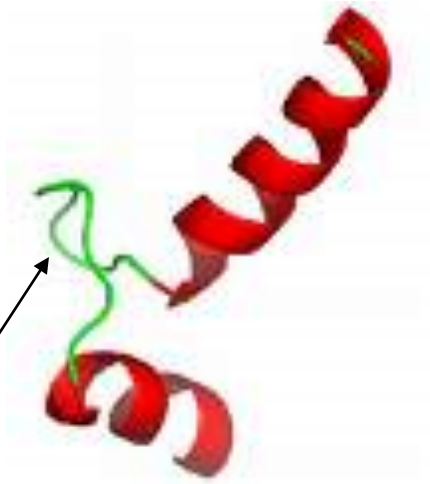
Anti-Parallel



Parallel

Non-Repetitive Secondary Structure

Beta-Turn



Loop

Tertiary Structure

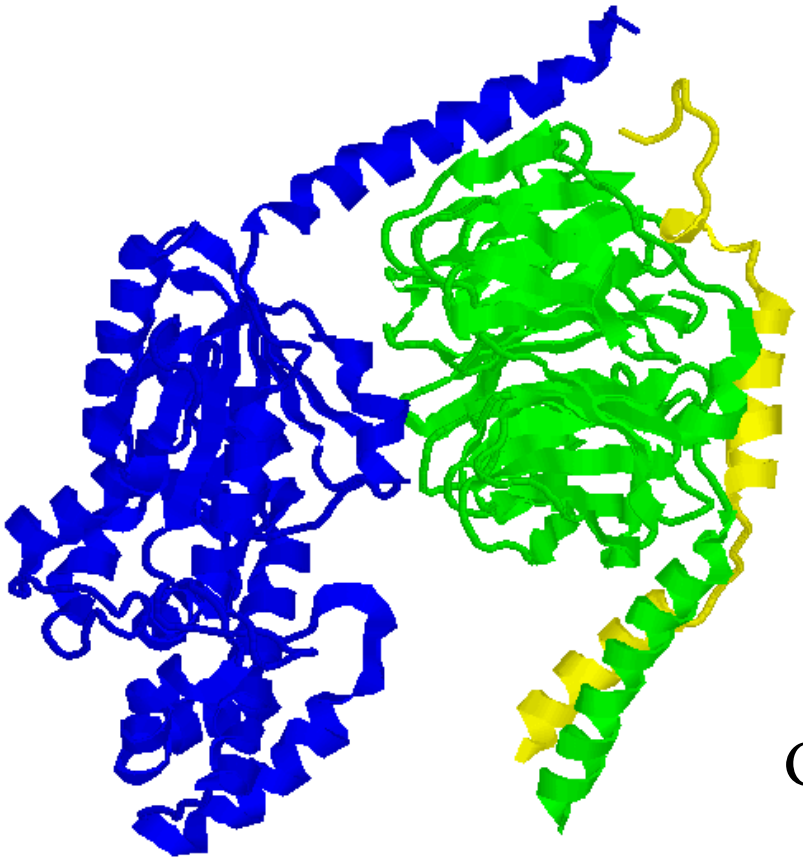
- John Kendrew et al., Myoglobin
- Max Perutz et al., Haemoglobin
- 1962 Nobel Prize in Chemistry

Perutz Kendrew

myoglobin

haemoglobin

Quaternary Structure: Complex



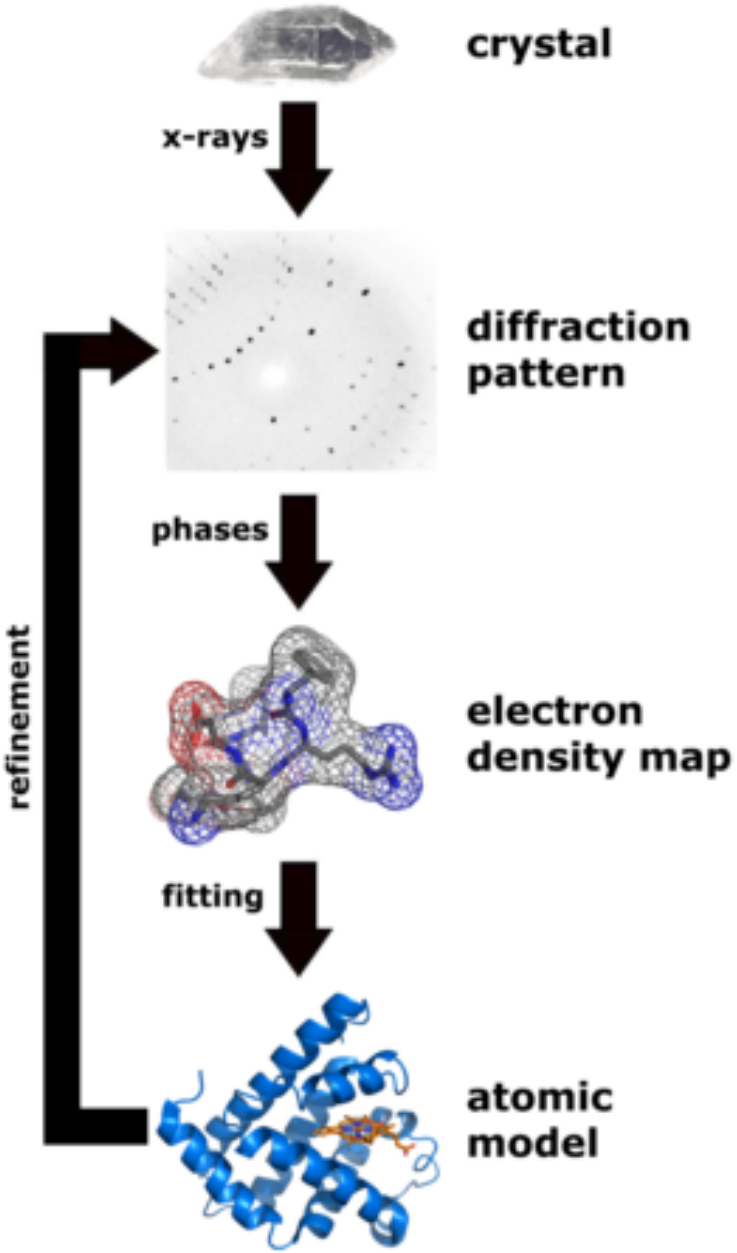
G-Protein Complex

Anfinsen's Folding Experiment

- Structure is uniquely determined by protein sequence
- Protein function is determined by protein structure

Protein Structure Determination

- X-ray crystallography
- Nuclear Magnetic Resonance (NMR) Spectroscopy
- X-ray: any size, accurate (1-3 Angstrom (10^{-10} m)), sometime hard to grow crystal
- NMR: small to medium size, moderate accuracy, structure in solution

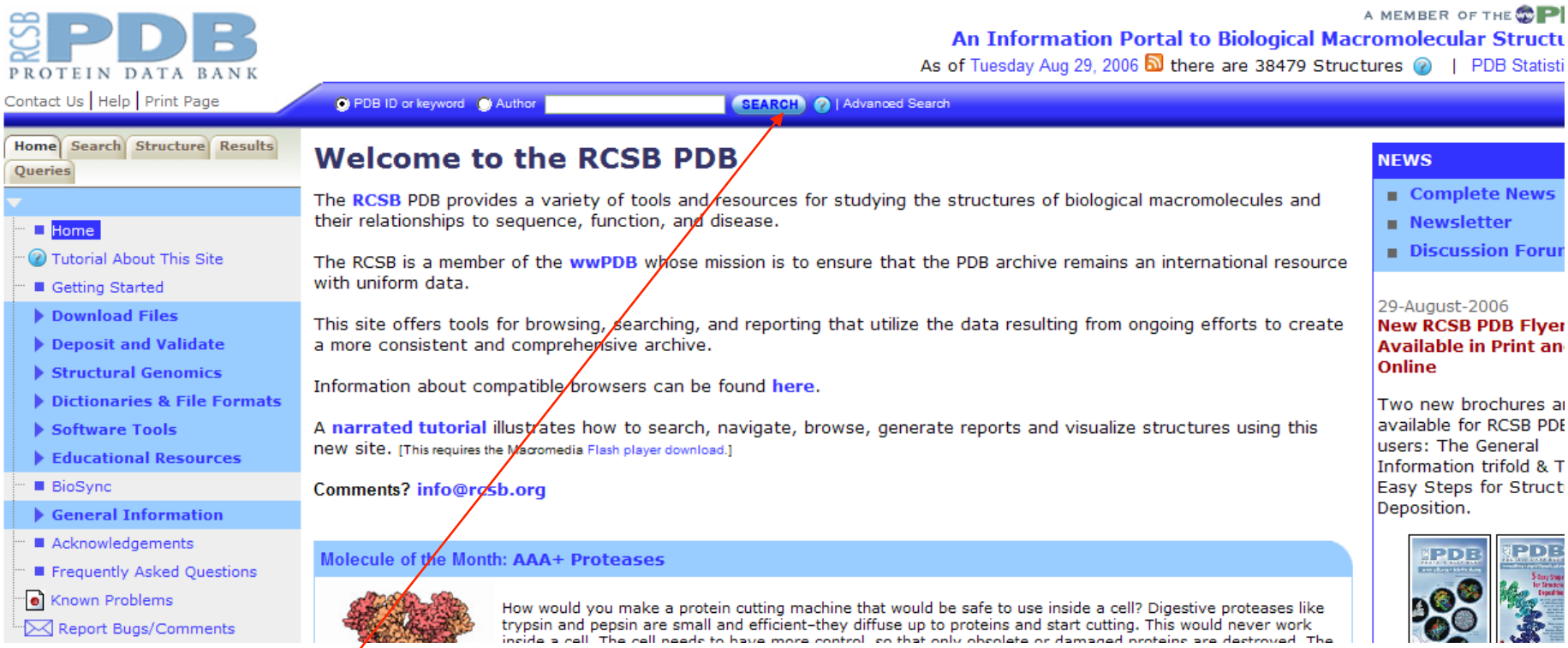


Wikipedia, the free encyclopedia

Pacific Northwest National Laboratory's high magnetic field (800 MHz, 18.8 T) NMR spectrometer being loaded with a sample.

Wikipedia, the free encyclopedia

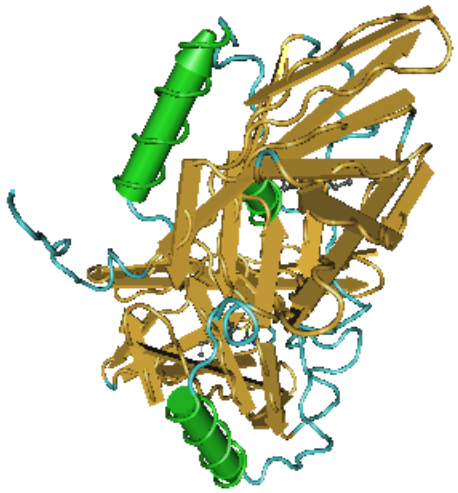
Storage in Protein Data Bank

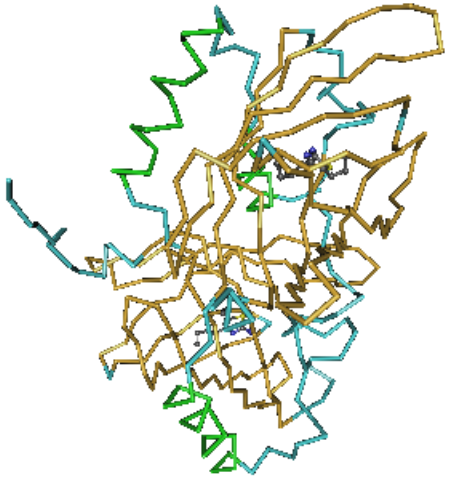


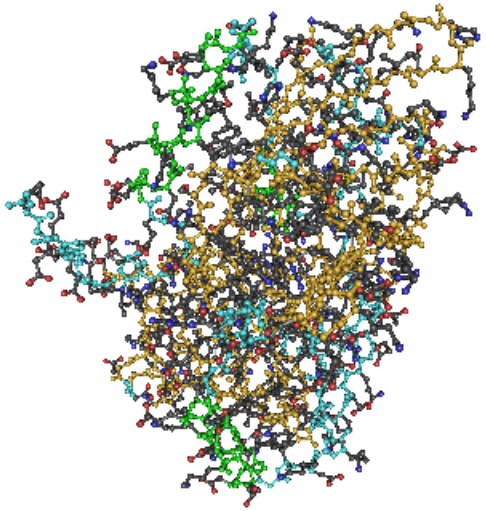
Search database

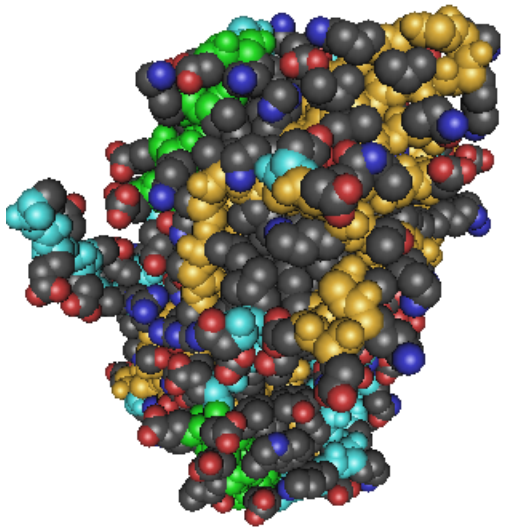
Search protein 1VJG

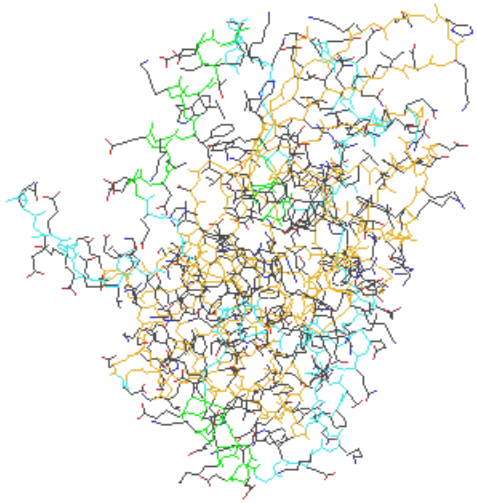
```
SEQRES   1 A   21  GLY ILE VAL GLU GLN CYS CYS THR SER ILE CYS SER LEU
SEQRES   2 A   21  TYR GLN LEU GLU ASN TYR CYS ASN
SEQRES   1 B   29  PHE VAL ASN GLN HIS LEU CYS GLY SER HIS LEU VAL GLU
SEQRES   2 B   29  ALA LEU TYR LEU VAL CYS GLY GLU ARG GLY PHE PHE TYR
SEQRES   3 B   29  THR PRO LYS
FORMUL   3  HOH   *31(H2 O1)
HELIX    1   1 GLY A    1  CYS A    7  1                                   7
HELIX    2   2 SER A   12  ASN A   18  1                                   7
HELIX    3   3 GLY B    8  GLY B   20  1                                  13
HELIX    4   4 GLU B   21  GLY B   23  5                                   3
SSBOND   1 CYS A    6    CYS A   11                          1555   1555
SSBOND   2 CYS A    7    CYS B    7                          1555   1555
SSBOND   3 CYS A   20    CYS B   19                          1555   1555
CRYST1   78.608   78.608   78.608  90.00  90.00  90.00 I 21 3        24
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.012721  0.000000  0.000000        0.00000
SCALE2      0.000000  0.012721  0.000000        0.00000
SCALE3      0.000000  0.000000  0.012721        0.00000
ATOM      1  N   GLY A   1      45.324  26.807  11.863  1.00 24.82           N
ATOM      2  CA  GLY A   1      45.123  27.787  12.967  1.00 24.93           C
ATOM      3  C   GLY A   1      43.756  27.627  13.605  1.00 25.16           C
ATOM      4  O   GLY A   1      43.107  26.591  13.438  1.00 25.00           O
ATOM      5  N   ILE A   2      43.313  28.661  14.323  1.00 25.21           N
ATOM      6  CA  ILE A   2      42.050  28.622  15.065  1.00 25.39           C
ATOM      7  C   ILE A   2      40.818  28.303  14.200  1.00 25.69           C
ATOM      8  O   ILE A   2      39.935  27.565  14.635  1.00 25.56           O
ATOM      9  CB  ILE A   2      41.816  29.917  15.917  1.00 25.39           C
```





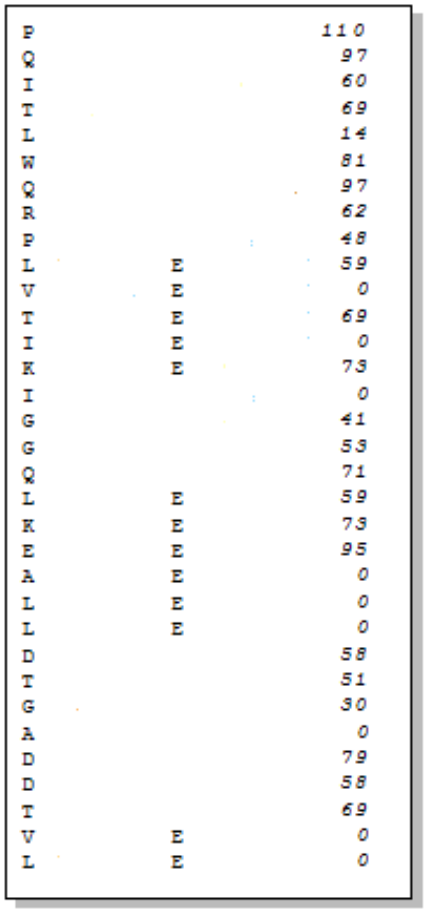




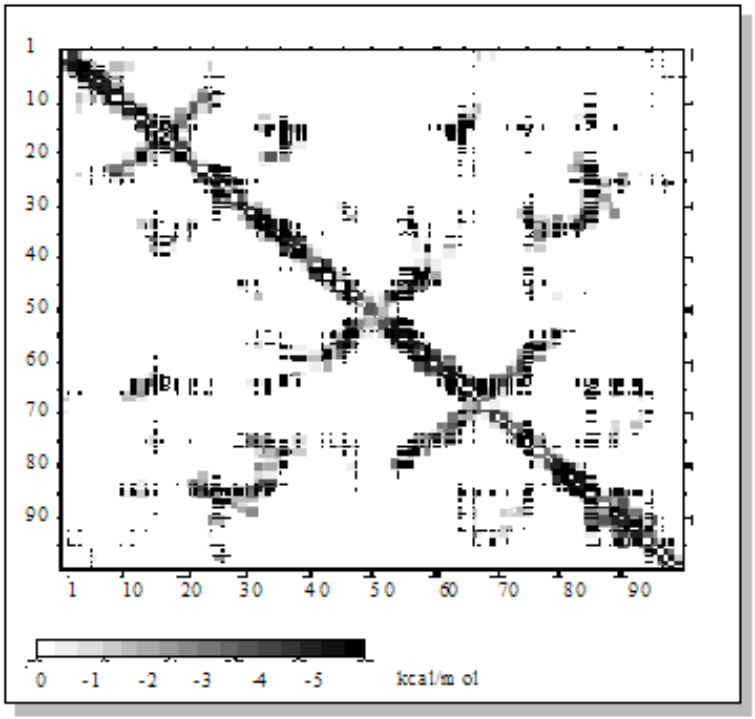


J. Pevsner, 2005

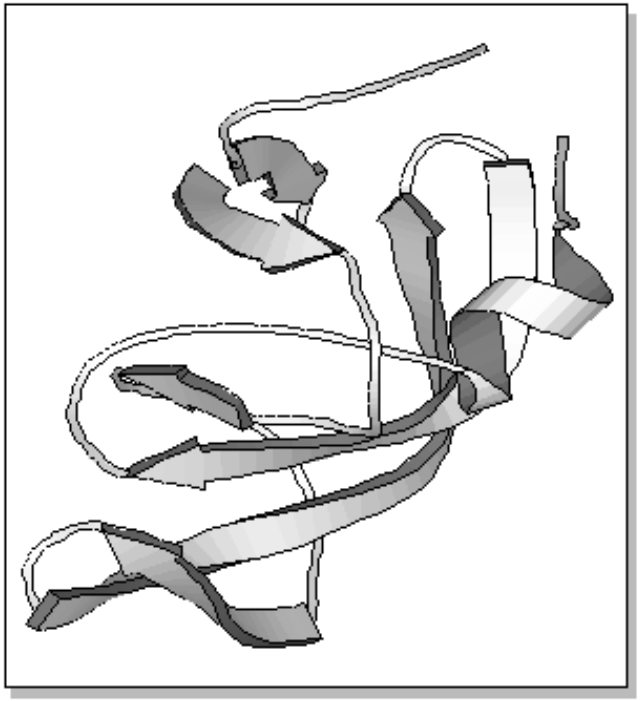
1D, 2D, 3D Structure Prediction



1D



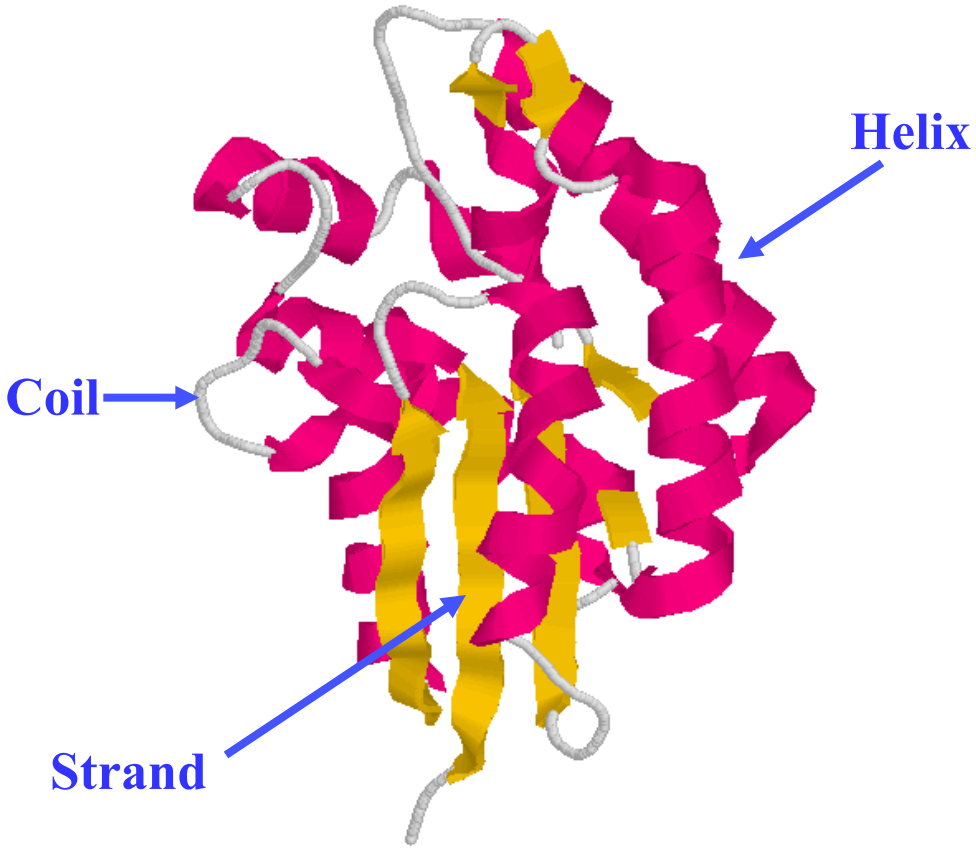
2D



3D

B. Rost, 2005

1D: Secondary Structure Prediction



MWLKKFGINLLIGQSV…

↓

Neural Networks + Alignments

↓

CCCCHHHHHCCCSSSSS…

Accuracy: 78%

Pollastri et al. *Proteins*, 2002. **Cheng et al.. *NAR*, 2005**

How to Use Neural Network to Predict Secondary Structure

- Create a data set with input sequences (x) and output labels (secondary structures)
- Encode the input and output to neural network
- Train neural network on the dataset (training dataset)
- Test on the unseen data (test dataset) to estimate the generalization performance.

Create a Data Set

- Download proteins from Protein Data Bank
- Select high-resolution protein structures (<2.5 Angstrom, determined by X-ray crystallography)
- Remove proteins with chain-break (Ca-Ca distance > 4 angstrom)
- Remove redundancy (filter out very similar sequences using BLAST)
- Use DSSP program (Kabsch and Sander, 1983) to assign secondary structure to each residue.

Train and Test

- Use one data set as training dataset to build neural network model
- Use another data set as test dataset to evaluate the generalization performance of the model
- Sequence similarity any two sequences in test and training dataset is less than 25%.

Create Inputs and Outputs for Feed-Forward NN for a Single Sequence

Protein Sequence:

MWLKKFGINLLIGQSVQTRSWYYCKRA

How to encode the input for each position?
How to encode the output for each position?

SS Sequence:

CCCCHHHHHHEEEEEHHHHEEEEEECC

Create Inputs and Outputs for Feed-Forward NN for a Single Sequence

Protein Sequence:

MWLKKFGINLLIGQSVQTRSWYYCKRA

Use 20 inputs of 0s and 1s for each amino acid
Use 3 inputs to encode the SS alphabet

SS Sequence:

CCCCHHHHHHEEEEEHHHHEEEEEECC

100: Helix, 010: Extended strand, 001: Coil
Similarly for 20 different amino acids

Use a Window to Account for Context

Protein Sequence:

MWLKKFGINLLIGQSVQTRSWYYCKRA

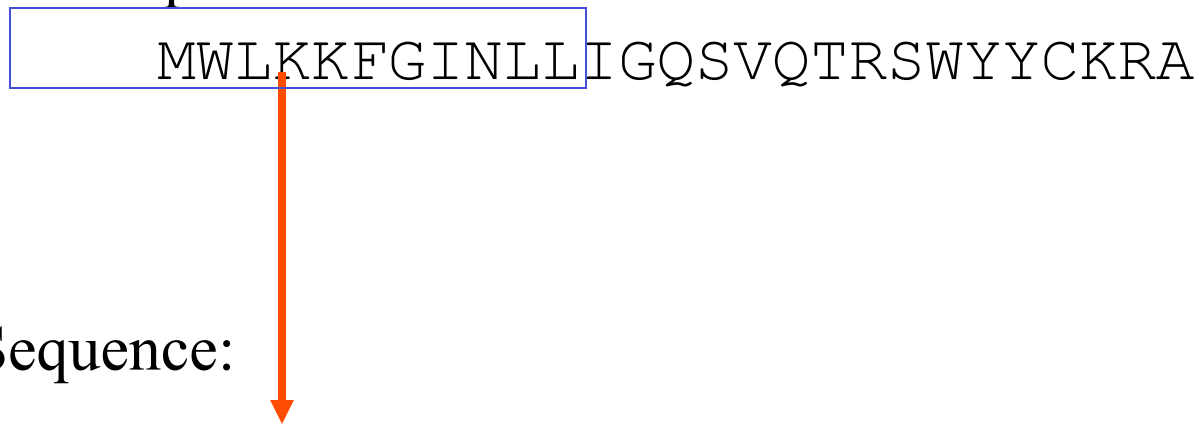
SS Sequence:

CCCCHHHHHHEEEEEHHHHEEEEEECC

Total number of inputs is window size (*l*) * 20. *l* is a parameter to tune.

Use an Extra Input to Account for N- and C- Terminal Boundary

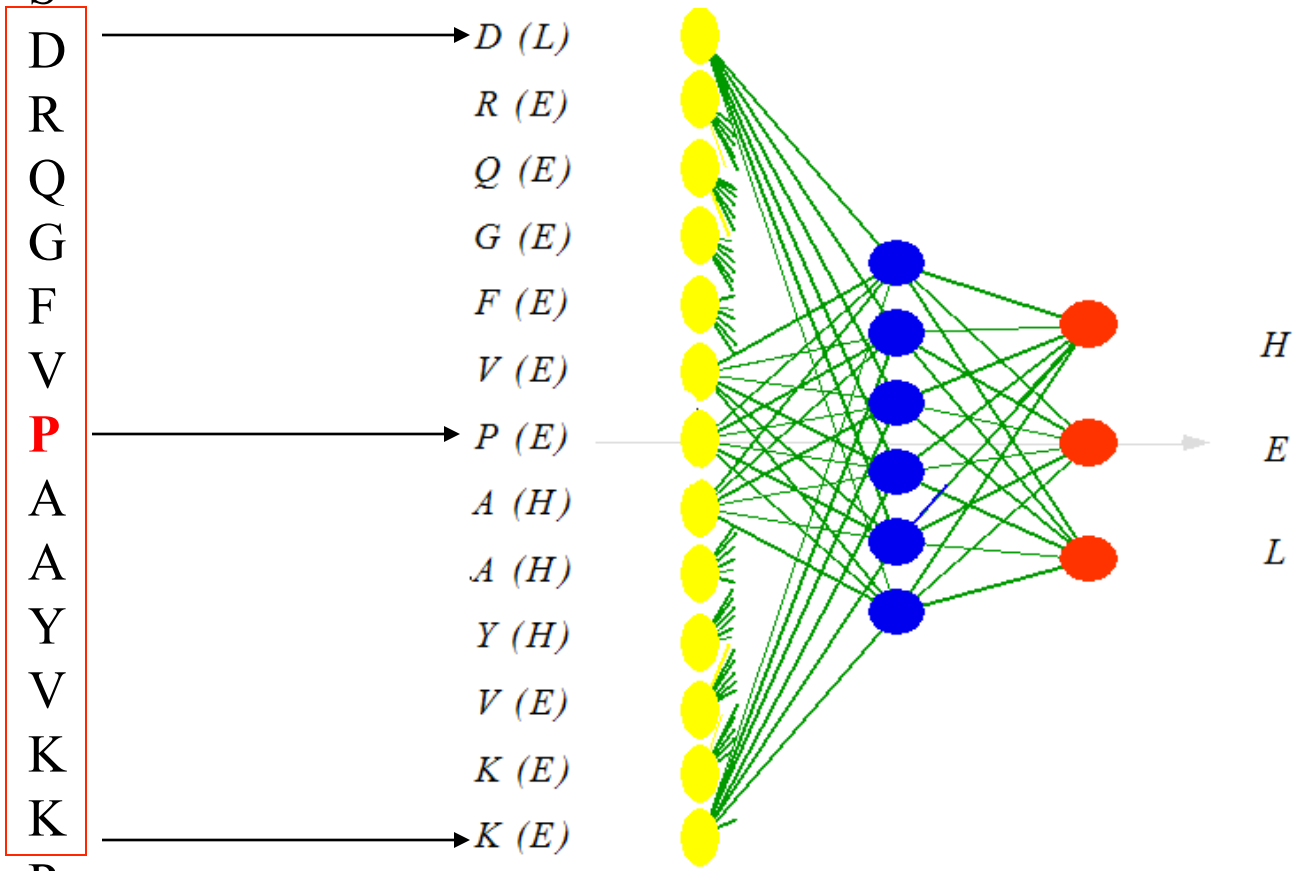
Protein Sequence:



Add an extra input for each position to indicate if it is out of the boundary of the sequence.

Total number of inputs is window size (l) * 21. l is a parameter to tune.

Secondary Structure Prediction (Generation III – Neural Network)



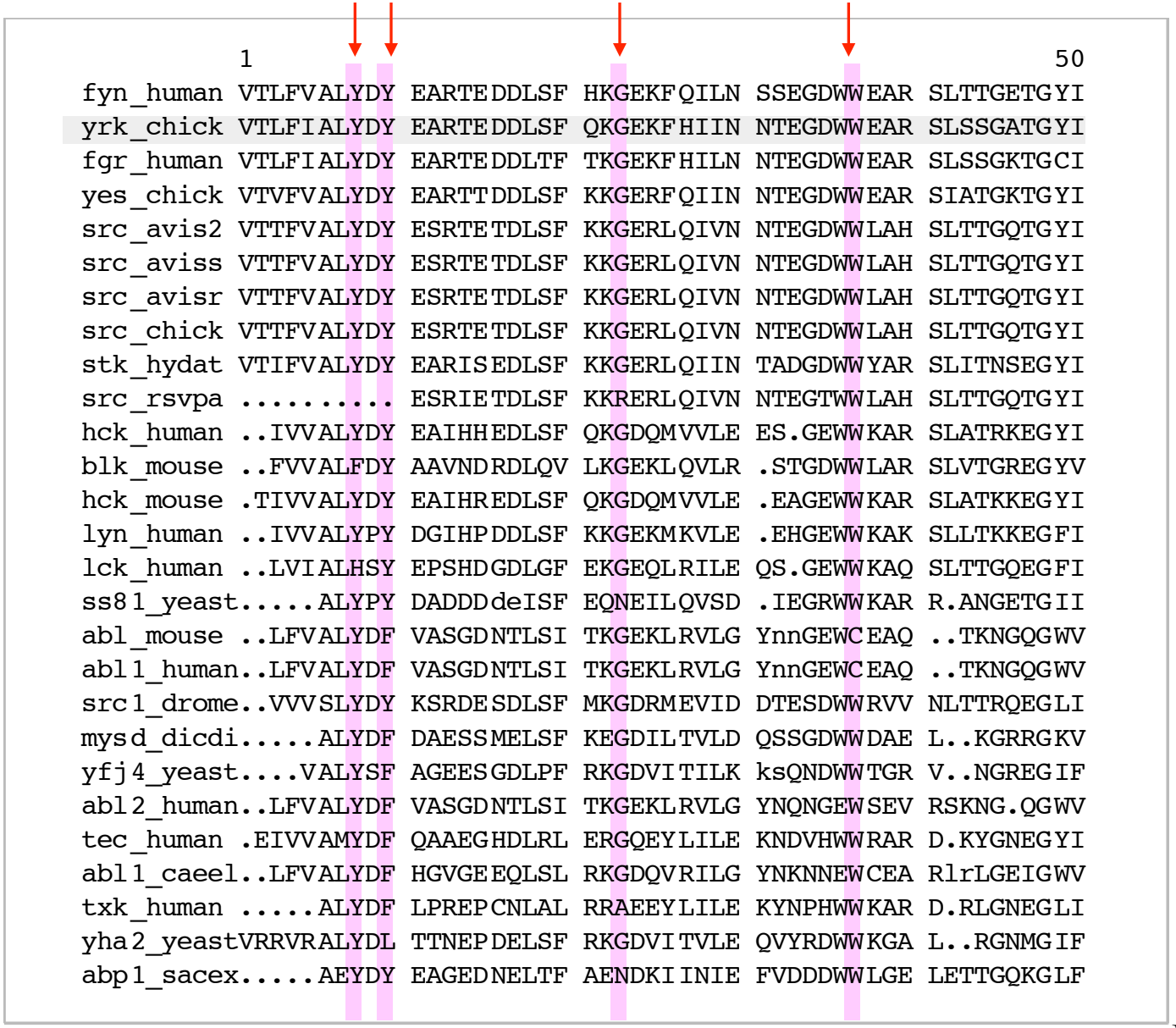
Output:
Prob (H)
Prob (E)
Prob (L or C)

B. Rost, 2005

Evolutionary Information is Important

- Single sequence yields accuracy below 70%.
- Use all the sequences in the family of a query sequence can improve accuracy to 78%.
- Structure is more conserved than sequence during evolution. The conservation and variation provides key information for secondary structure prediction.

Second Breakthrough: **Evolutionary Information - Profile**

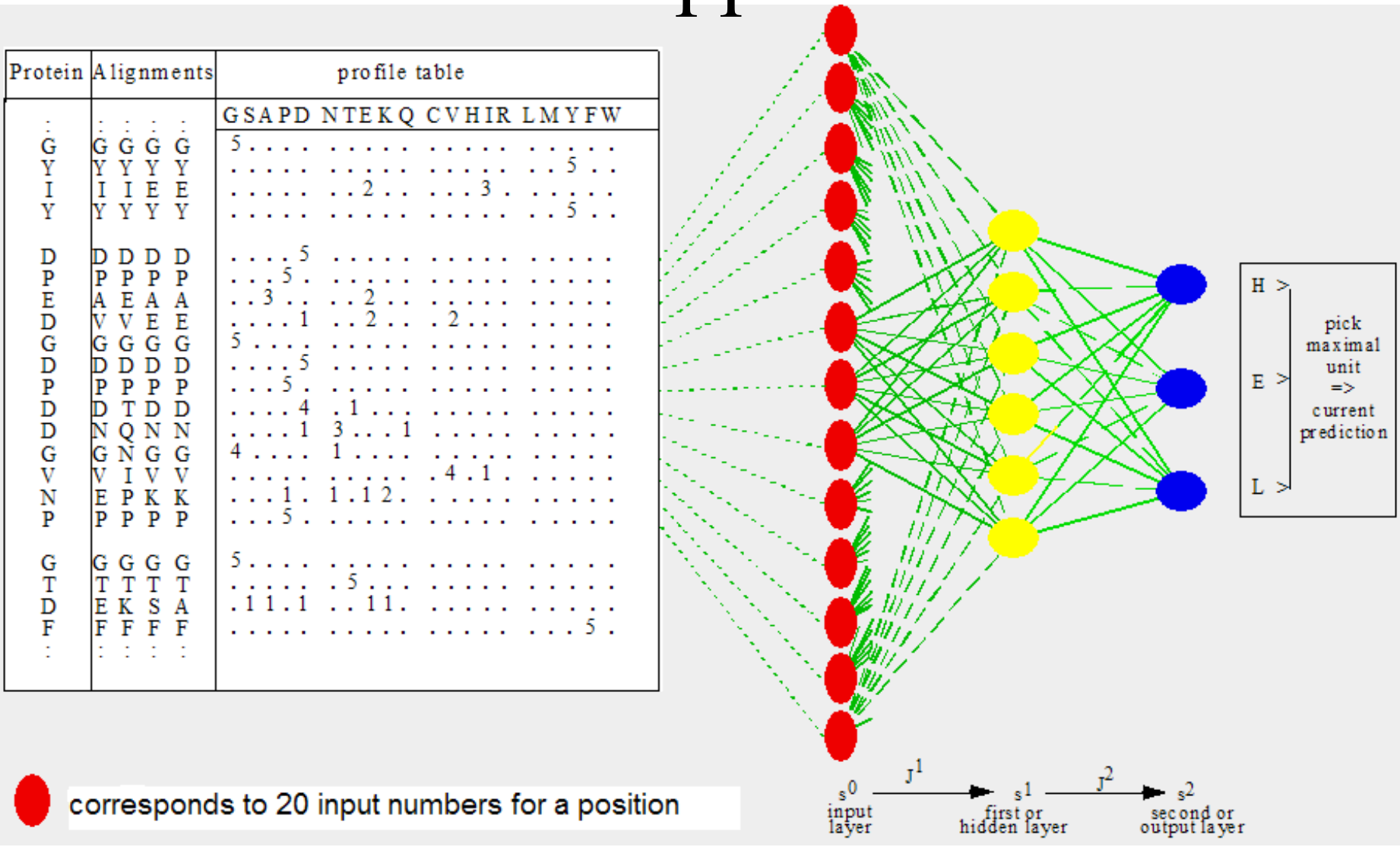


B. Rost, 2005

How to Find Homologous Sequences and Generate Alignments

- Use PSI-BLAST to search a query sequence against the very large non-redundant protein sequence database (NR database, compiled at NCBI)
- Combine the pairwise alignment between the query sequence and other sequences into a multiple sequence alignment using the query sequence as the center.
-

PhD Approach



Comments: frequency is normalized into probability and sequence needs to be weighted.

Reference: Rost and Sander. Proteins, 1994.

B. Rost, 2005

PSI-PRED Approach

- PSI-PRED does not use probability matrix instead it uses the another kind of profile: Position Specific Scoring Matrix generated by PSI-BLAST during sequence search.
- The weighting of the sequences is done implicitly by PSI-BLAST.
- The raw PSSM is transformed into values within [0,1] using sigmoid function.

Reference: Jones, Journal of Molecular Biology, 1999.

PSI-PRED Input

Position-based scoring matrix used

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

Window of 15 rows

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
0.4	0.3	0.3	0.3	0.2	0.9	0.3	0.3	0.4	0.4	0.4	0.3	0.4	0.9	0.1	0.4	0.4	0.5	0.7	0.4
0.3	0.2	0.3	0.8	0.4	0.3	0.7	0.1	0.6	0.2	0.4	0.3	0.5	0.2	0.1	0.4	0.8	0.2	0.3	0.2
0.1	0.1	0.4	0.3	0.5	0.1	0.1	0.3	0.1	0.1	0.4	0.2	0.4	0.9	0.3	0.4	0.4	0.9	0.3	0.6
0.6	0.3	0.3	0.1	0.3	0.5	0.5	0.2	0.1	0.4	0.4	0.3	0.6	0.9	0.1	0.5	0.1	0.5	0.7	0.4

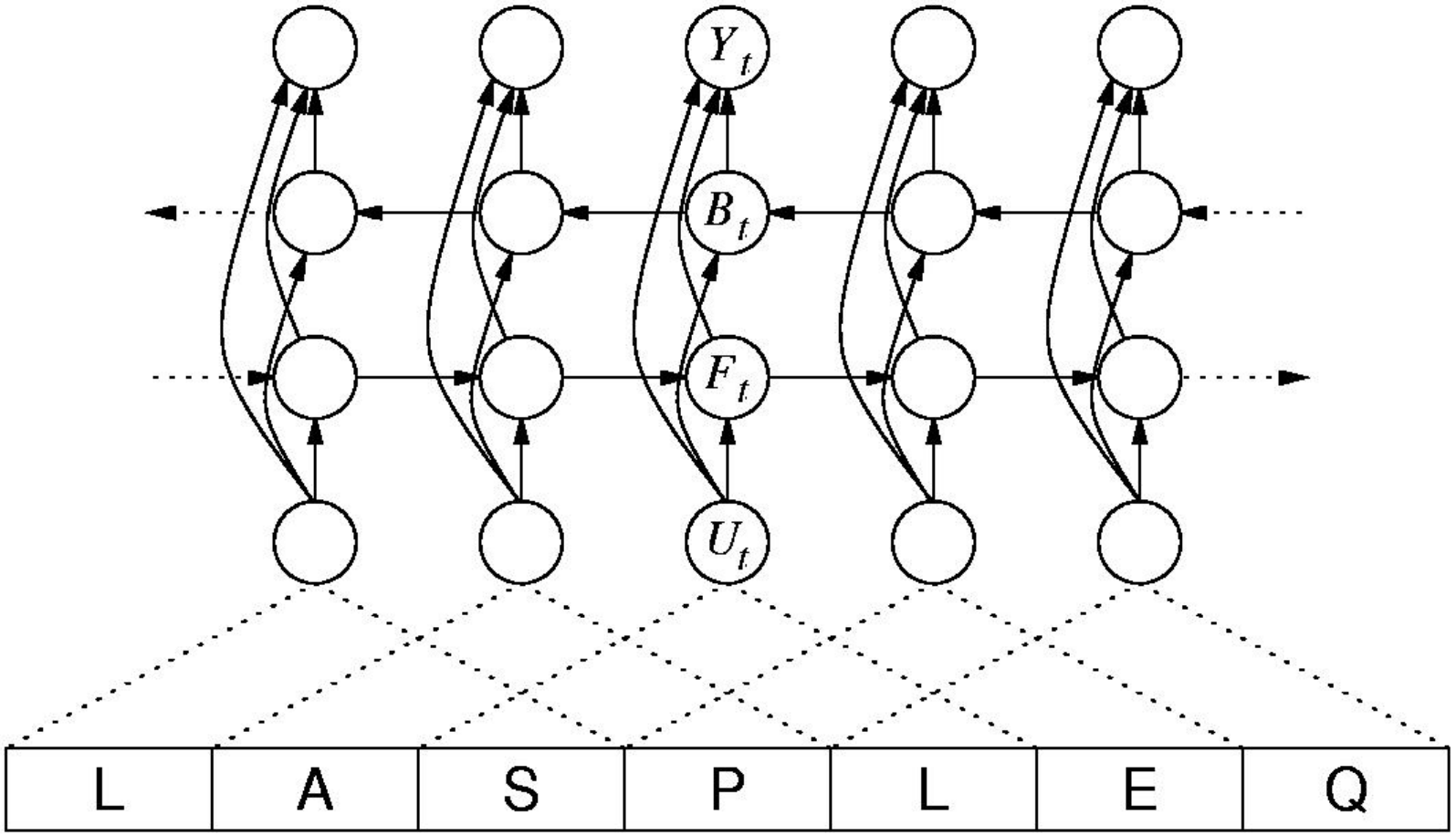
Reference: Jones, Journal of Molecular Biology, 1999.

SSpro Approach

- SSpro uses probability matrix as inputs
- SSpro uses an information theory approach to weight sequences
- The main novelty of SSpro is to use 1-Dimensional Recurrent Neural Network (1D-RNN)

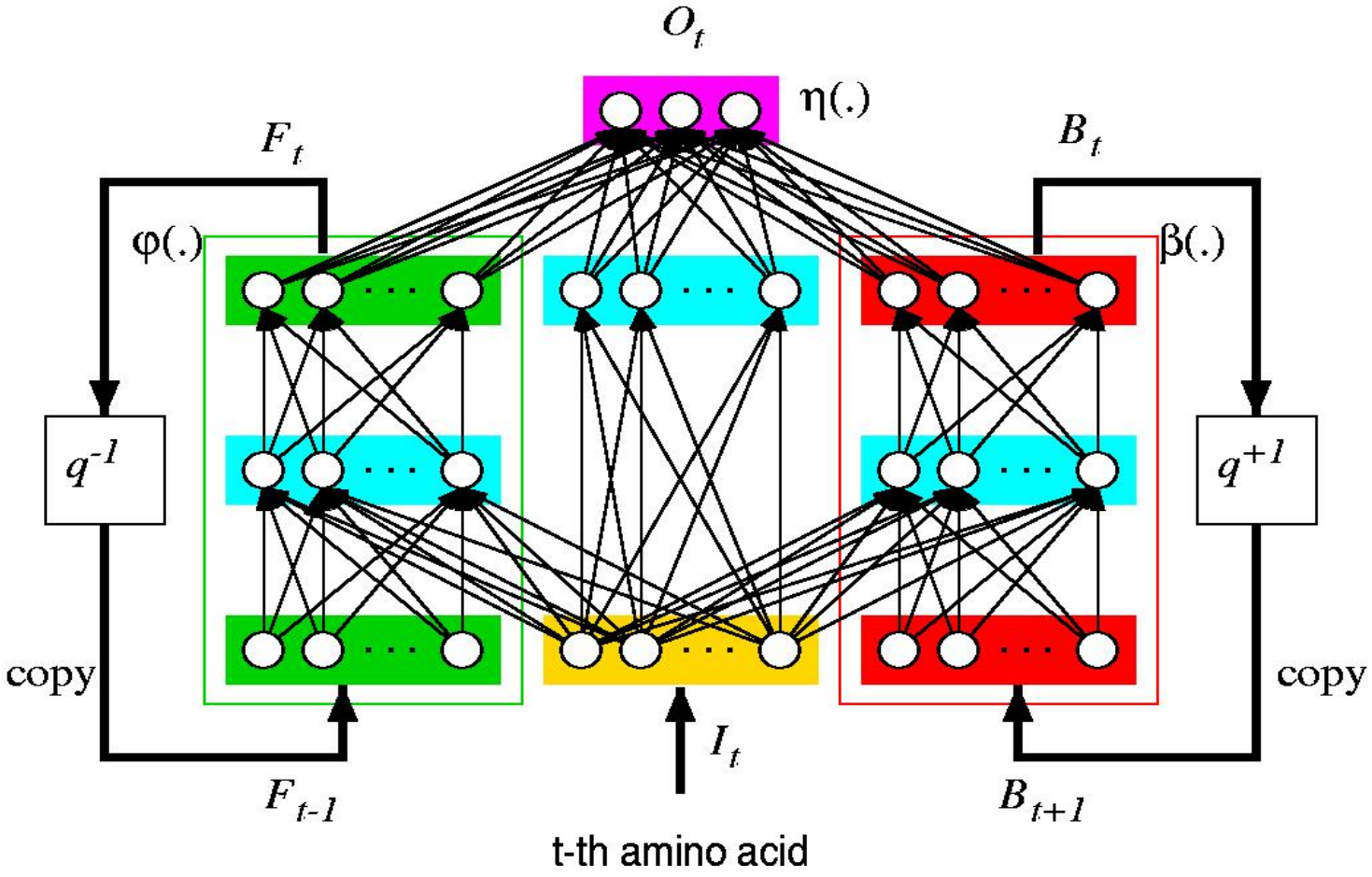
Pollastri et al.. Proteins, 2002.

Bi-directional Input Output Hidden Markov Model for SS Prediction



Baldi, 2004

1D-Recursive Neural Network



Baldi, 2004

Advantage and Disavantages of SSpro

- Directly take a sequence with variable length as inputs.
- Hopefully can utilize more information than a fixed-window approach
- More complex, thus harder to implement than feed-forward neural network.

Second Neural Network to Smooth Output Predictions

- Raw output from one neural network may contain weird predictions such as helix of length 1. But minimum length is 2.
- So use another neural network to smooth output. The inputs are a window of predicted secondary structure. The outputs are the true secondary structures.
- The second neural network makes the predictions more protein-like.

Secondary Structure Prediction Project (4th project)

- Training dataset with sequences and secondary structures (1180 sequences) and test dataset (126 sequences). (training data was created by Pollastri et al. and test data was created by Rost and Sander.) (www.cs.missouri.edu/~chengji/mlbioinfo/ss_trian.txt (and ss_test.txt)
- Generate multiple alignments using generate_flatblast.sh in SSpro 4 package (http://casp.rnet.missouri.edu/download/)

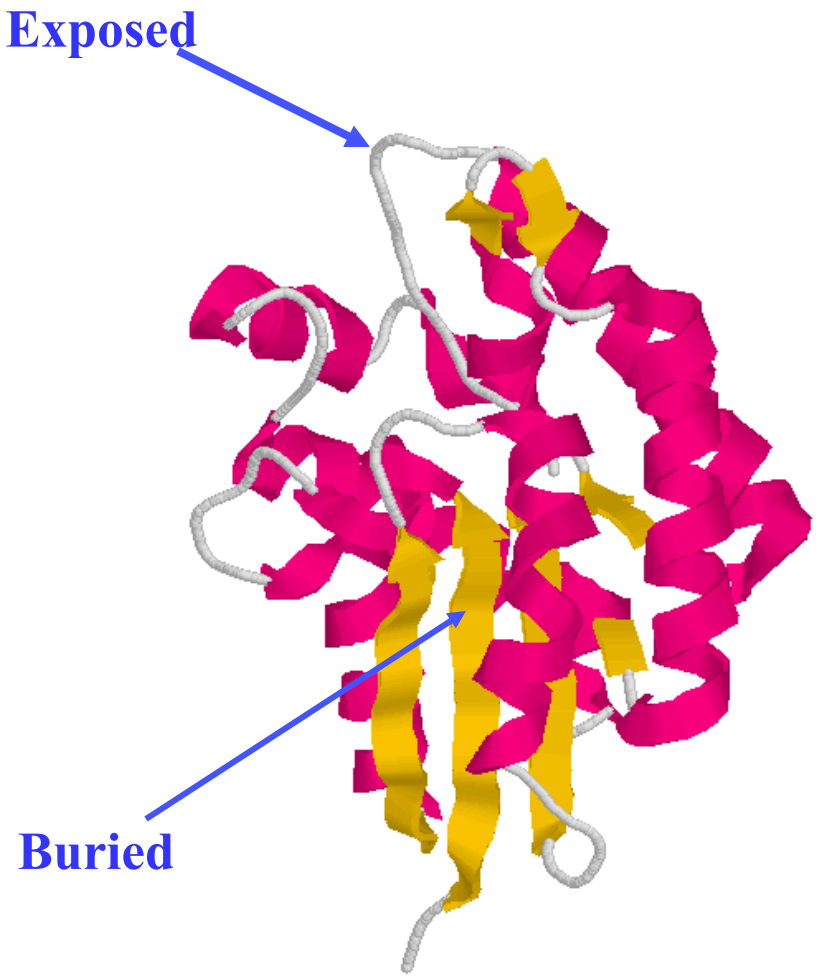
Secondary Structure Prediction Project (continued)

- Generate inputs and outputs (probability matrix or Position Specific Scoring Matrix)
- Train neural network using NNClass or Weka
- Test neural network on test dataset

References for the Project:

- B. Rost and C. Sander. Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure. *Proteins*. 1994.
- D.T. Jones. Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. *JMB*. 1999.
- G. Pollastri, D. Przybylski, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*. 2002
- J. Cheng, A. Randall, M. Sweredoski, and P. Baldi. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Research. 2005.

1D: Solvent Accessibility Prediction



MWLKKFGINLLIGQSV…

↓

Neural Networks + Alignments

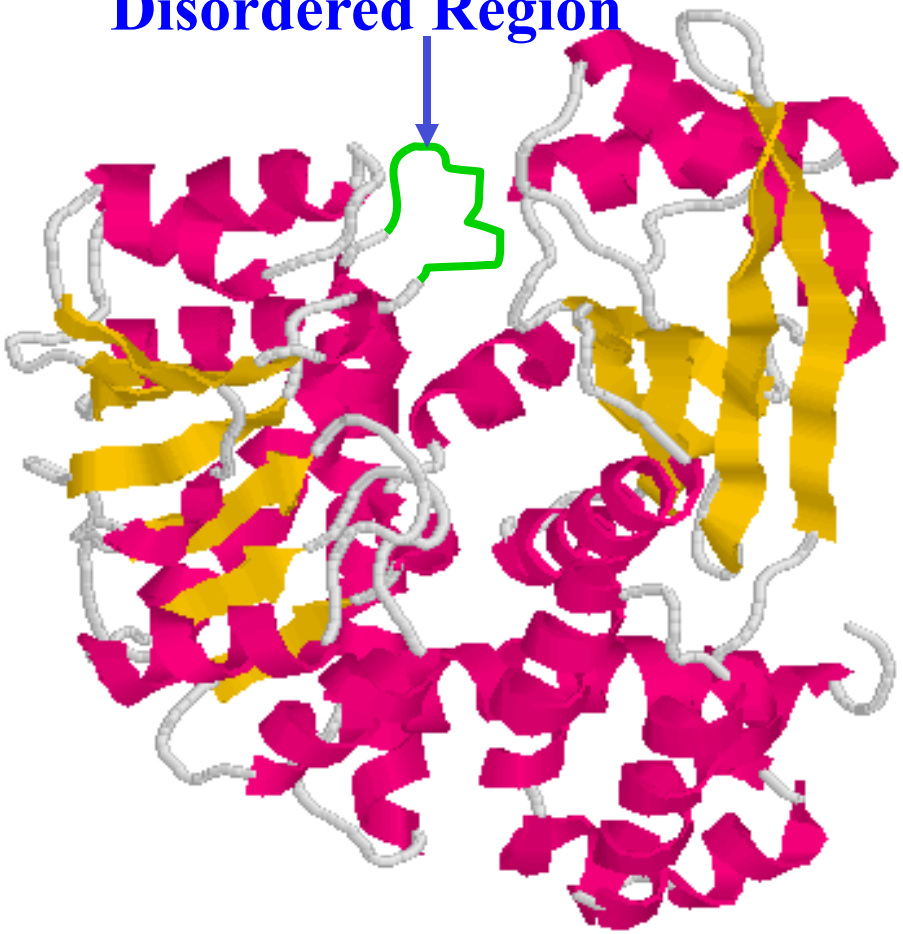
↓

eeeeeeebbbbbbbbeeeebbb…

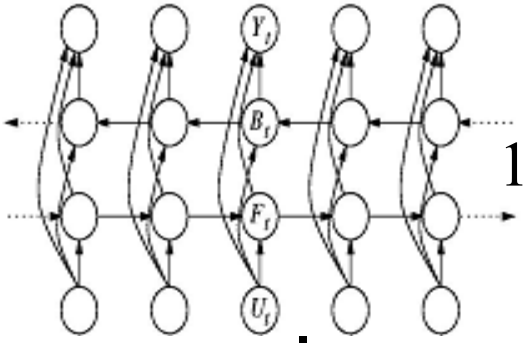
Accuracy: 79%

Pollastri et al. *Proteins*, 2002.
Cheng et al. *Nucleic Acid Research*, 2005

1D: Disordered Region Prediction Using Neural Networks



MWLKKFGINLLIGQSV…



1D-RNN

OOOOODDDDOOOOO…

93% TP at 5% FP

Cheng, Sweredoski, Baldi. *Data Mining and Knowledge Discovery*, 2005

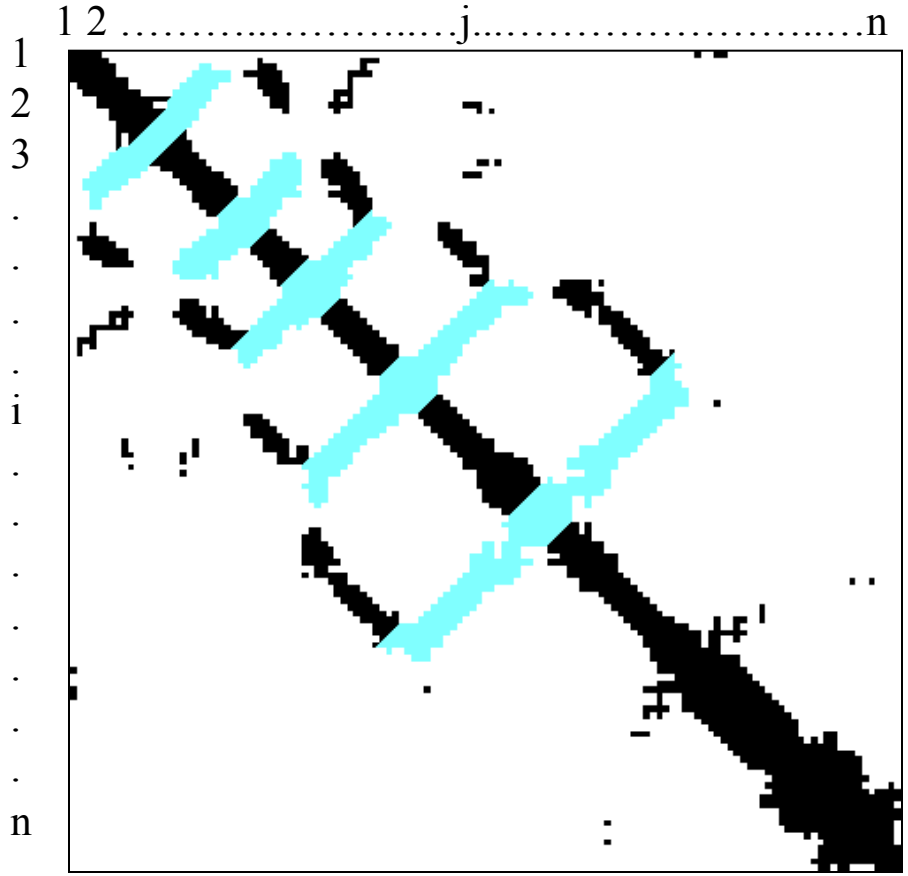
2D: Contact Map Prediction

3D Structure

2D-Recursive Neural Network

Support Vector Machine

2D Contact Map



Distance Threshold = 8A°

Cheng, Randall, Sweredoski, Baldi. *Nucleic Acid Research*, 2005
Cheng and Baldi. *BMC Bioinformatics*, 2007.