

Statistical Machine Learning Methods for Bioinformatics

IV. Neural Network Applications in Bioinformatics

Jianlin Cheng, PhD

Department of Computer Science

University of Missouri, Columbia

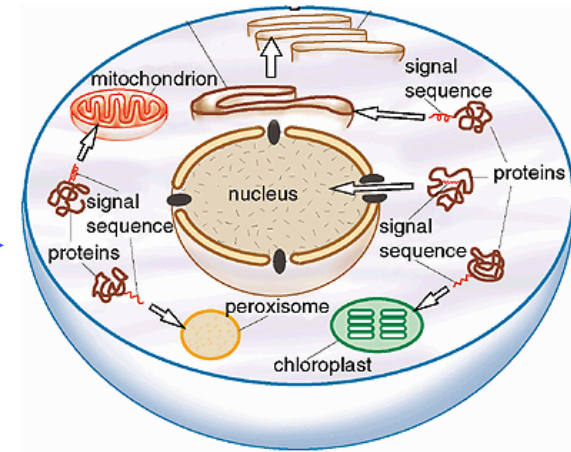
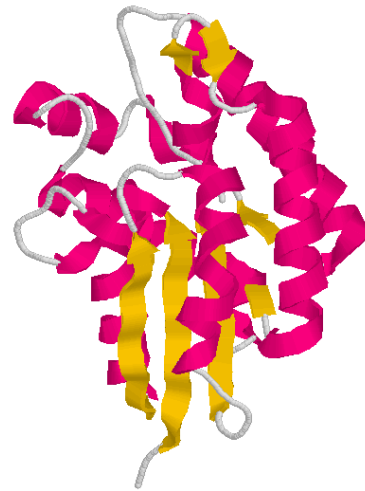
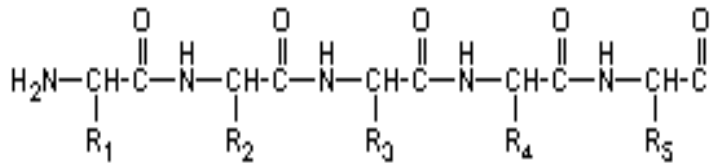
2012

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

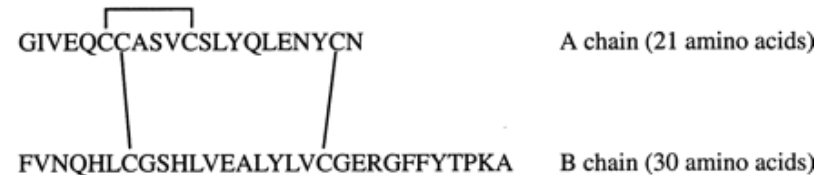
AGCWY.....



Cell

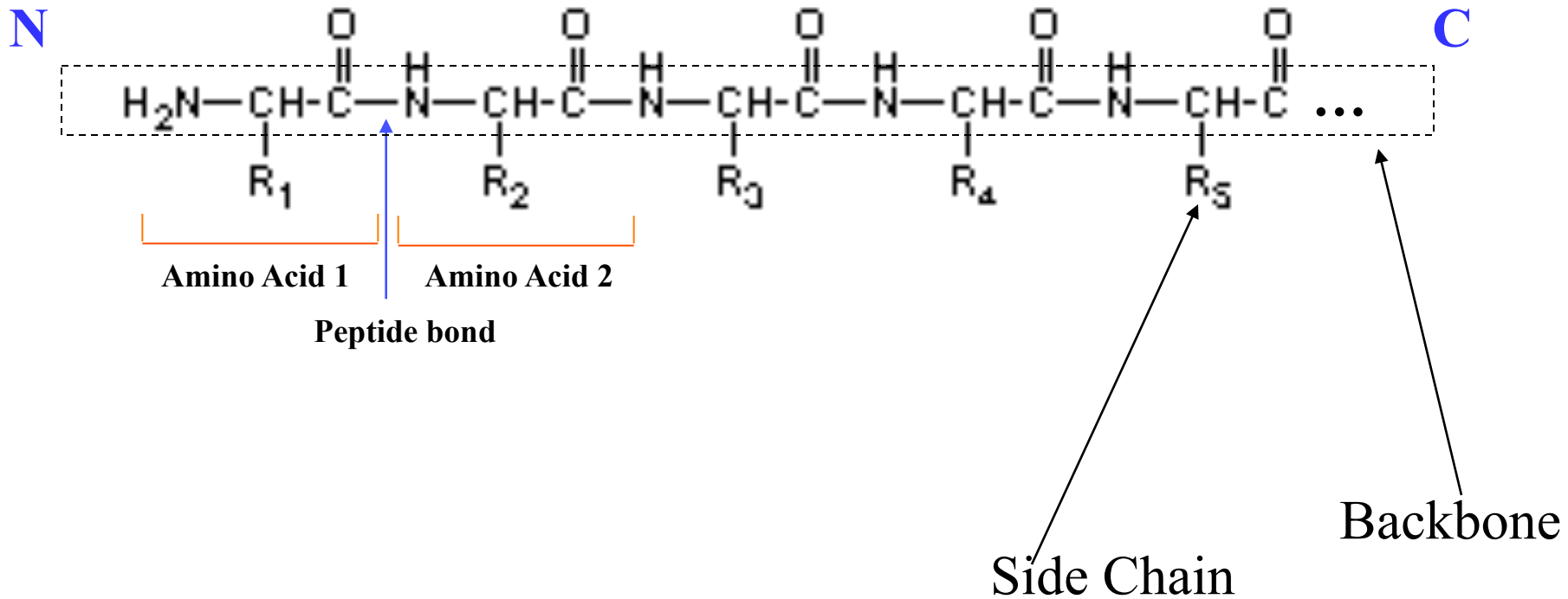
Protein Sequence – Primary Structure

- 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.

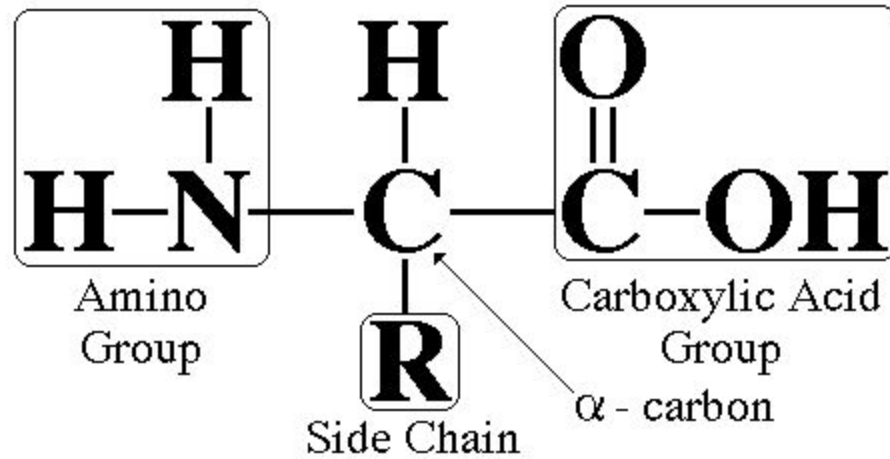


Protein Sequence

A directional sequence of amino acids/residues



Amino Acid Structure



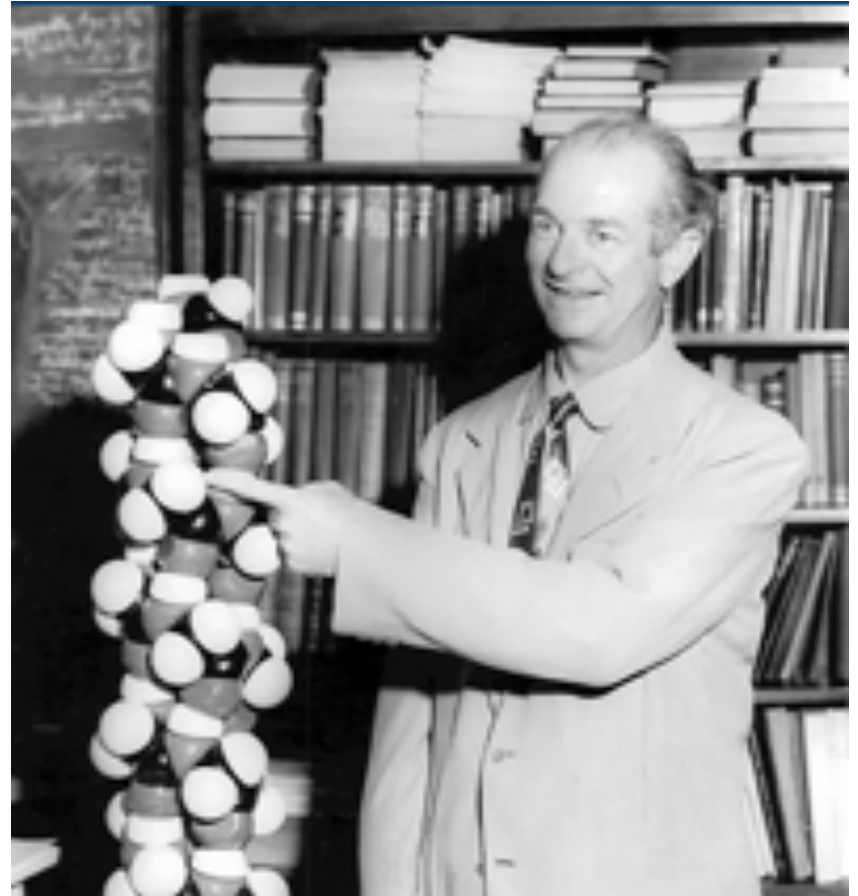
Amino Acids

Amino acid	Abbrev.	Side chain	Hydrophobic	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 ₂ SH	X	-	-	X	-	-	86	UGU, UGC	1.9
Aspartate	Asp, D	-CH ₂ COOH	-	X	negative	X	-	-	91	GAU, GAC	5.3
Glutamate	Glu, E	-CH ₂ CH ₂ COOH	-	X	negative	-	-	-	109	GAA, GAG	6.3
Phenylalanine	Phe, F	-CH ₂ C ₆ H 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 ₃ H ₃ N ₂	-	X	positive	-	-	Aromatic	118	CAU, CAC	2.3
Isoleucine	Ile, I	-CH(CH ₃)CH ₂ CH ₃	X	-	-	-	-	Aliphatic	124	AUU, AUC, AUA	5.3
Lysine	Lys, K	-(CH ₂) ₄ NH ₂	-	X	positive	-	-	-	135	AAA, AAG	5.9
Leucine	Leu, L	-CH ₂ CH(CH ₃) ₂	X	-	-	-	-	Aliphatic	124	UUA, UUG, CUU, CUC, CUA, CUG	9.1
Methionine	Met, M	-CH ₂ CH ₂ SCH ₃	X	-	-	-	-	-	124	AUG	2.3
Asparagine	Asn, N	-CH ₂ CONH ₂	-	X	-	X	-	-	96	AAU, AAC	4.3
Proline	Pro, P	-CH ₂ CH ₂ CH ₂ -	X	-	-	X	-	-	90	CCU, CCC, CCA, CCG	5.2
Glutamine	Gln, Q	-CH ₂ CH ₂ CONH ₂	-	X	-	-	-	-	114	CAA, CAG	4.2
Arginine	Arg, R	-(CH ₂) ₃ NH-C(NH) NH ₂	-	X	positive	-	-	-	148	CGU, CGC, CGA, CGG, AGA, AGG	5.1
Serine	Ser, S	-CH ₂ OH	-	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 ₂ C ₈ H ₆ N	X	-	-	-	-	Aromatic	163	UGG	1.4
Tyrosine	Tyr, Y	-CH ₂ -C ₆ H ₄ OH	X	X	-	-	-	Aromatic	141	UAU, UAC	3.2

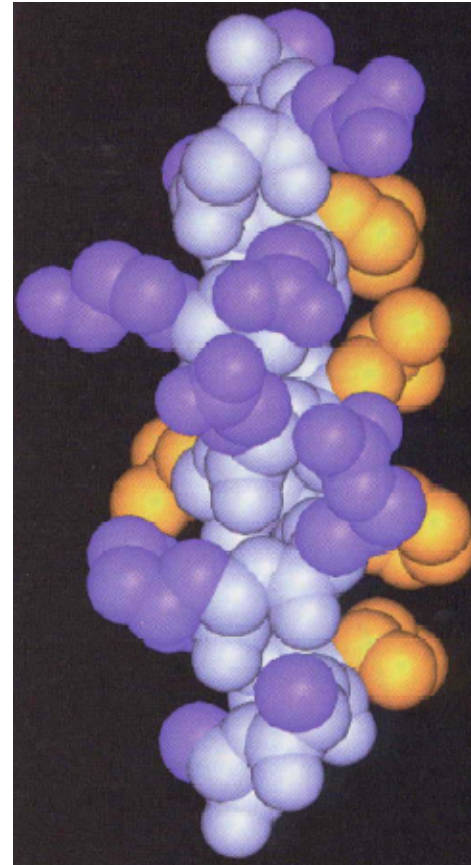
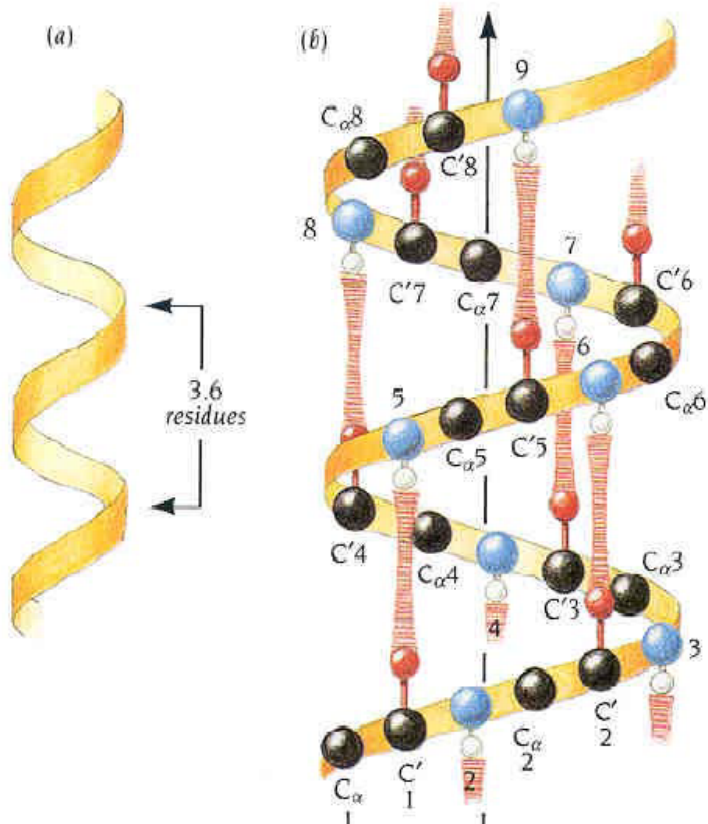
↑
Hydrophilic

Protein Secondary Structure

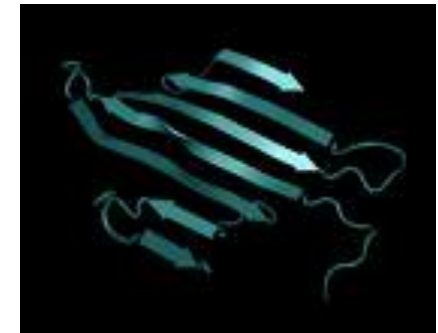
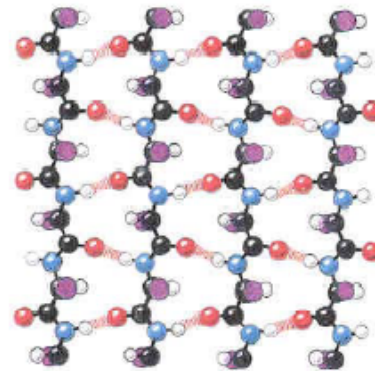
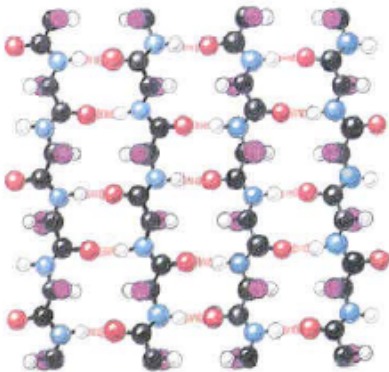
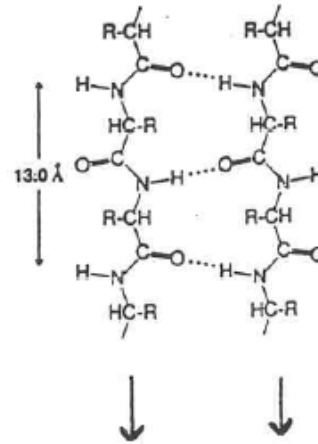
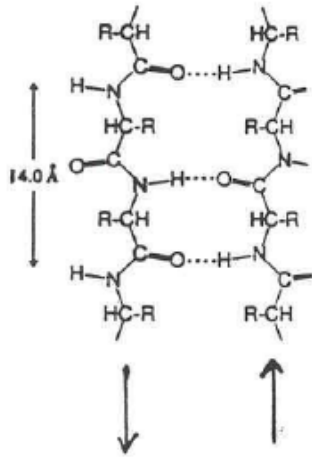
- Determined by hydrogen bond patterns
- 3-Class categories: alpha-helix, beta-sheet, loop (or coil)
- First deduced by Linus Pauling et al.



Alpha-Helix



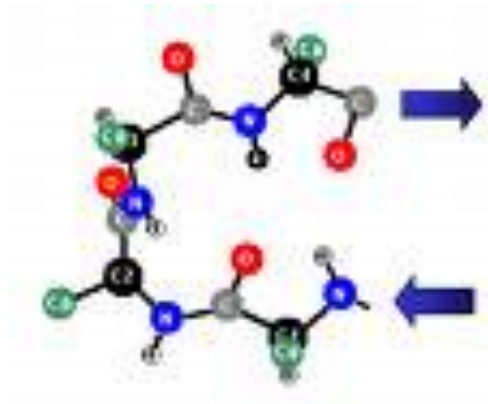
Beta-Sheet



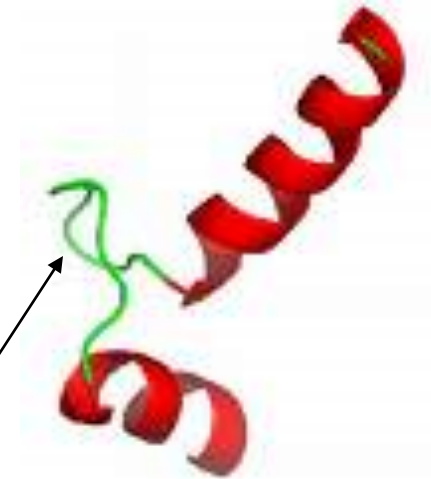
Anti-Parallel

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

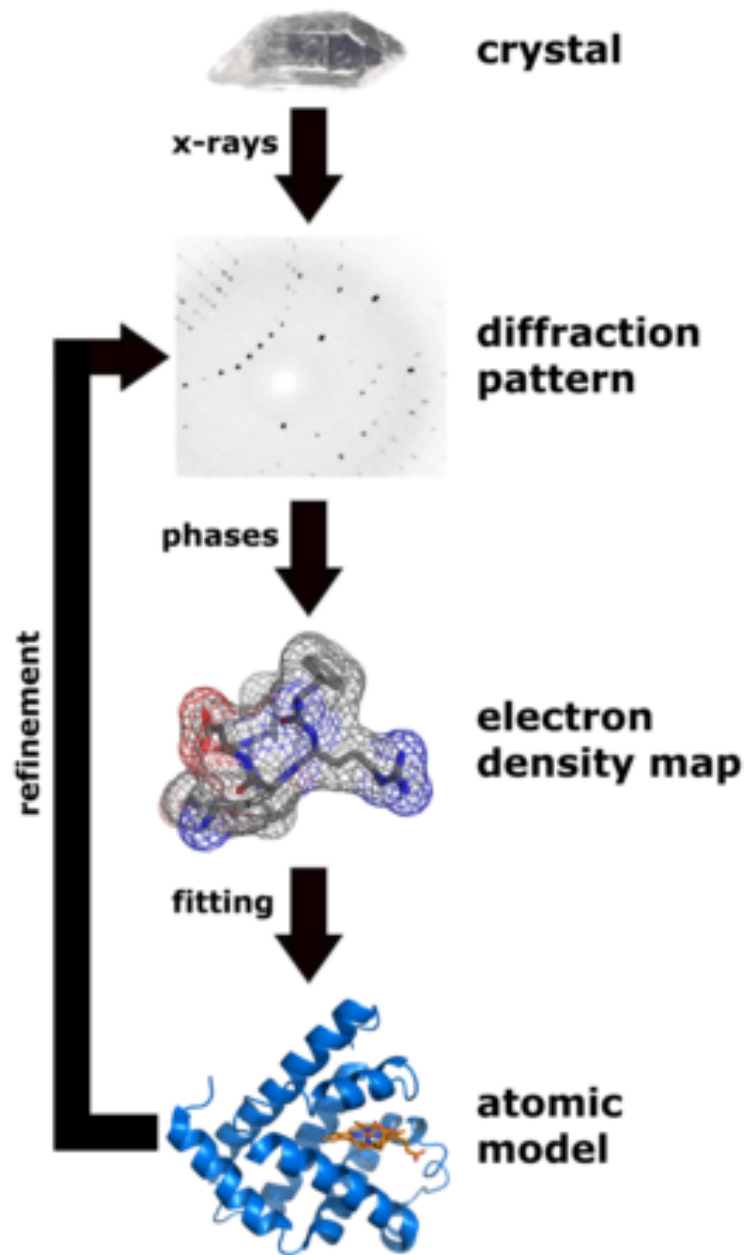
Anfinsen's Folding Experiment

- 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





[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

- Home
- Tutorial About This Site
- Getting Started
- Download Files
- Deposit and Validate
- Structural Genomics
- Dictionaries & File Formats
- Software Tools
- Educational Resources
- BioSync
- General Information
- Acknowledgements
- Frequently Asked Questions
- Known Problems
- Report Bugs/Comments

Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the **wwPDB** whose mission is to ensure that the PDB archive remains an international resource with uniform data.

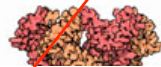
This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found [here](#).

A **narrated tutorial** illustrates how to search, navigate, browse, generate reports and visualize structures using this **new site**. [This requires the Macromedia Flash player download.]

Comments? info@rcsb.org

Molecule of the Month: AAA+ Proteases



How would you make a protein cutting machine that would be safe to use inside a cell? Digestive proteases like trypsin and pepsin are small and efficient—they diffuse up to proteins and start cutting. This would never work inside a cell. The cell needs to have more control, so that only obsolete or damaged proteins are destroyed. The

NEWS

- Complete News
- Newsletter
- Discussion Forum

29-August-2006
New RCSB PDB Flyer Available in Print and Online

Two new brochures are available for RCSB PDB users: The General Information trifold & The Easy Steps for Structure Deposition.



Search database

RCSB PDB : Structure Explorer - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.rcsb.org/pdb/navbsearch.do?newSearch=yes&isAuthorSearch=no&radioSet=All&inputQuickSearch=1vjg&image.x=0&image.y=0&image=Search

Google pdb

RCSB PDB
PROTEIN DATA BANK

A MEMBER OF THE **PDDB**

An Information Portal to Biological Macromolecular Structures

As of Tuesday Oct 10, 2006 there are 39323 Structures | PDB Statistics

Contact Us | Help | Print Page

PDB ID or keyword Author **SEARCH** | Advanced Search

Home Search **Structure** Queries Structure Summary Biology & Chemistry Materials & Methods Sequence Details Geometry

1VJG

- Download Files
- FASTA Sequence
- Display Files
- Display Molecule
- Structural Reports
- Structure Analysis
- Help

1VJG Images and Visualization

Biological Molecule



Display Options

- KING
- Jmol
- WebMol
- Protein Workshop
- QuickPDB
- All Images

Title Crystal structure of putative lipase from the G-D-S-L family from Nostoc sp. at 2.01 Å resolution

Authors Joint Center for Structural Genomics (JCSG)

Primary Citation Joint Center for Structural Genomics (JCSG) Crystal structure of putative lipase from the G-D-S-L family from Nostoc sp. at 2.01 Å resolution. *To be published*

History Deposition 2004-02-19 Release 2004-03-16

Experimental Method Type X-RAY DIFFRACTION Data [EDS]

Parameters

Resolution [Å]	R-Value	R-Free	Space Group
2.01	0.175 (obs.)	0.218	P 3 ₂ 2 1

Unit Cell

Length [Å]	a	b	c
56.19	56.19	129.32	129.32
Angles [°]	alpha	beta	gamma
90.00	90.00	120.00	120.00

Molecular Description Asymmetric Unit Polymer: 1 Molecule: putative lipase from the G-D-S-L family Chains: A

Functional Class Structural Genomics Unknown Function

Source Polymer: 1 Scientific Name: **Nostoc sp. pcc 7120** Common Name: **Bacteria** Expression system: **Nostoc sp. pcc 7120**

Done

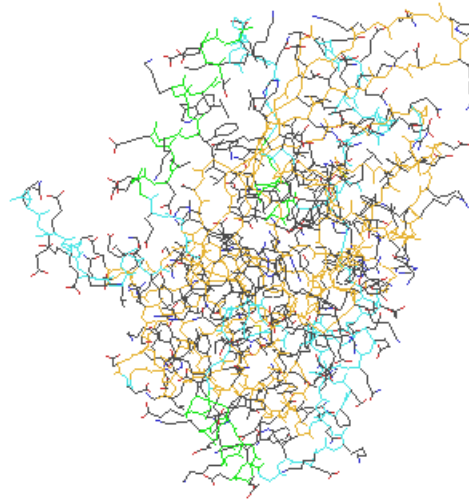
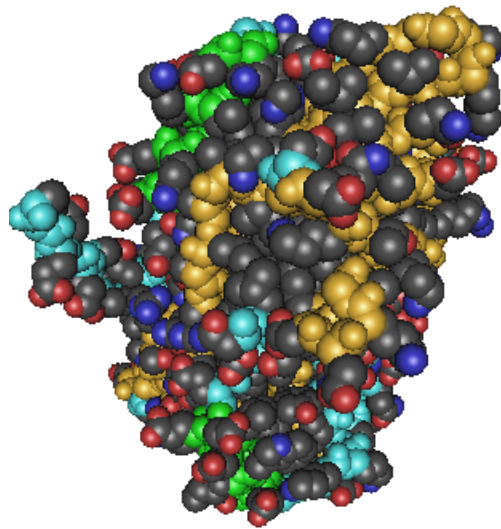
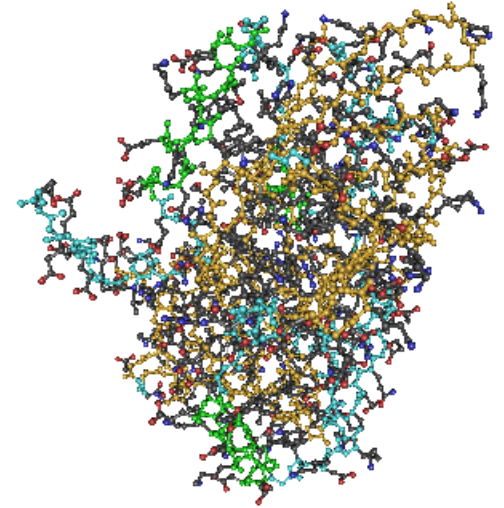
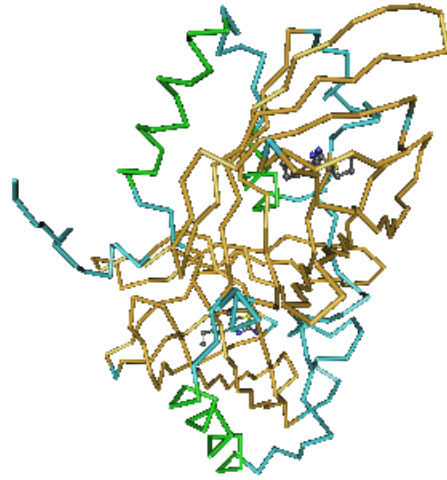
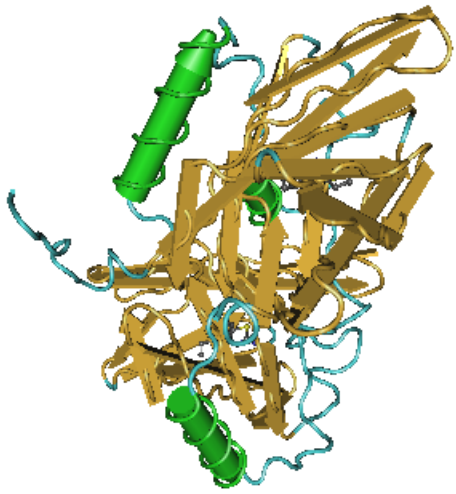
Start

Inbox - Outlook Express CAP5937 slides13 slides1 RCSB PDB : Structure ...

Entrez cross-database s... untitle - Paint

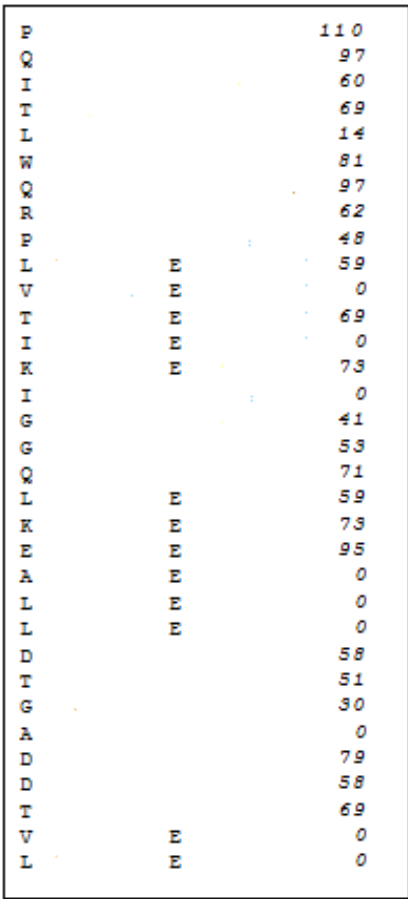
10:42 AM Monday

Search protein 1VJG

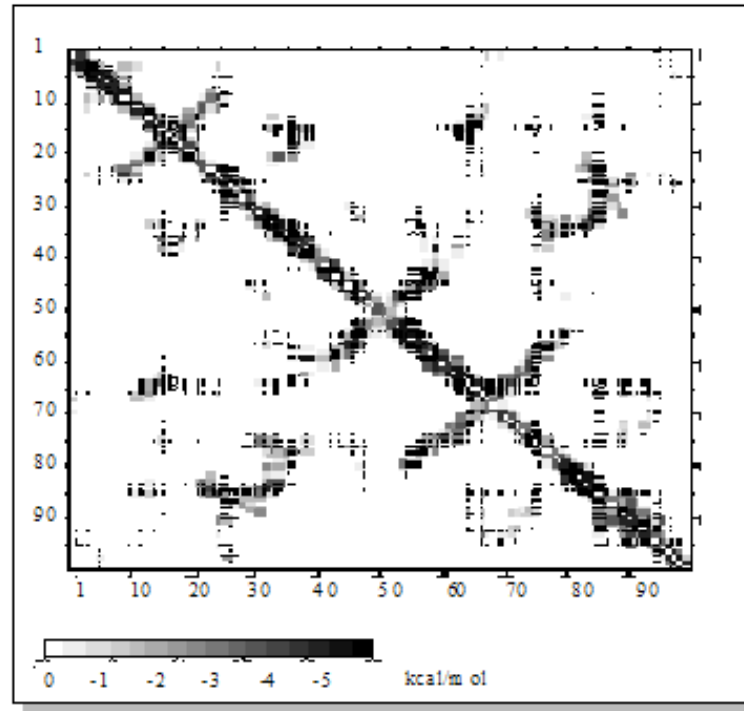


J. Pevsner, 2005

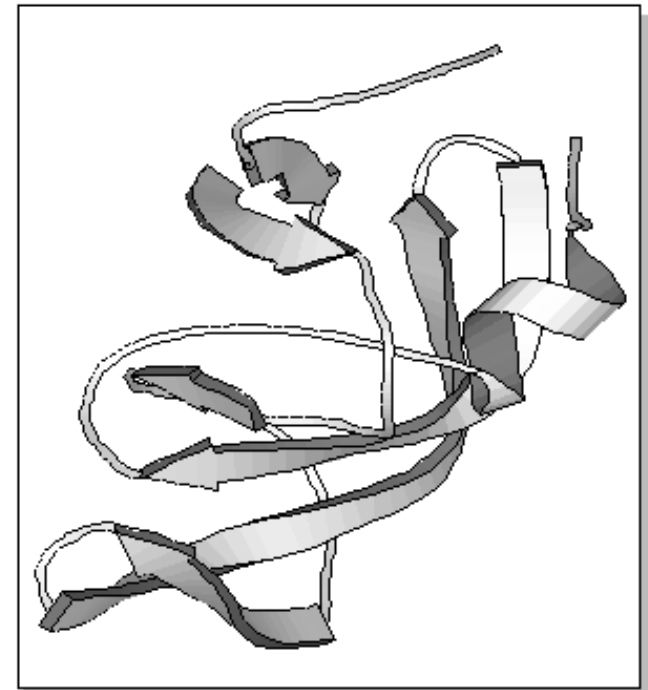
1D, 2D, 3D Structure Prediction



1D

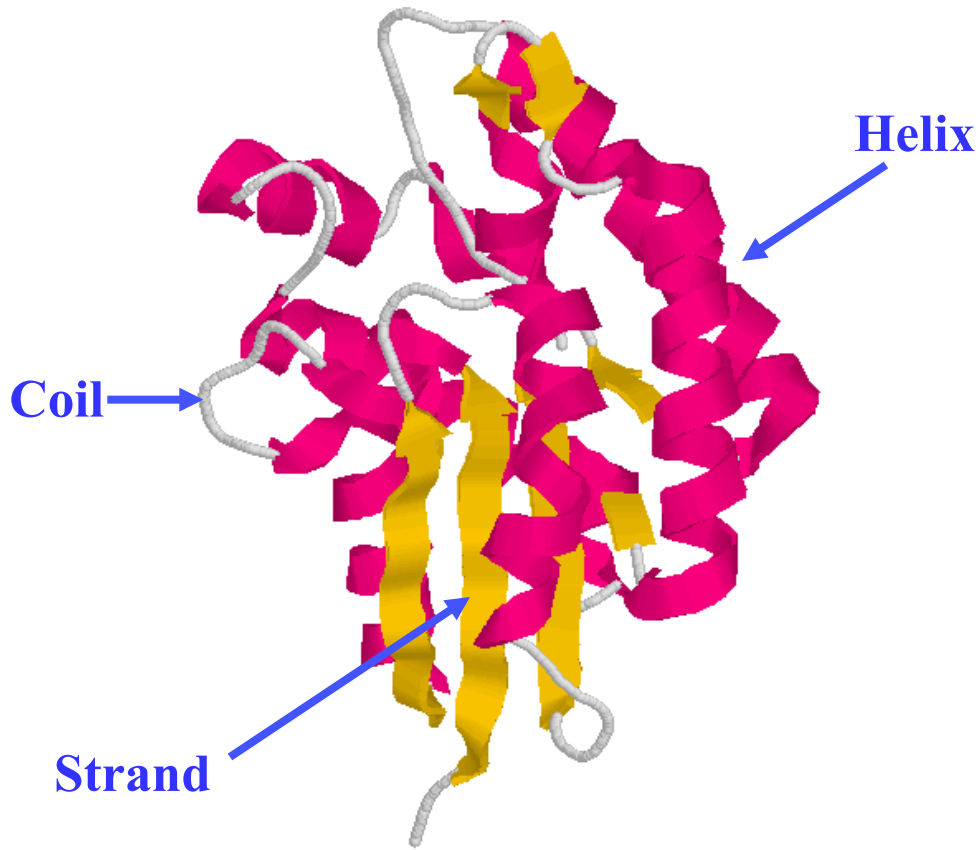


2D



3D

1D: Secondary Structure Prediction



MWLKKFGINLLIGQSV...

↓
**Neural Networks
+ Alignments**
↓

CCCCHHHHHCCCSSSSS...

Accuracy: 78%

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:

MWLKKKFGINLLIGQSVQTRSWYYCKRA

SS Sequence:

CCCCHHHHHHEEEEEHHHHEEEEECC



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

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

Protein Sequence:

MWLKKKFGINLLIGQSVQTRSWYYCKRA

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

SS Sequence:

CCCCHHHHHHEEEEEHHEEEEECC

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:

CCCCHHHHHHEEEEEHHHHEEEEECC

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:

MWLKKFGINLLIGQSVQTRSWYYCKRA



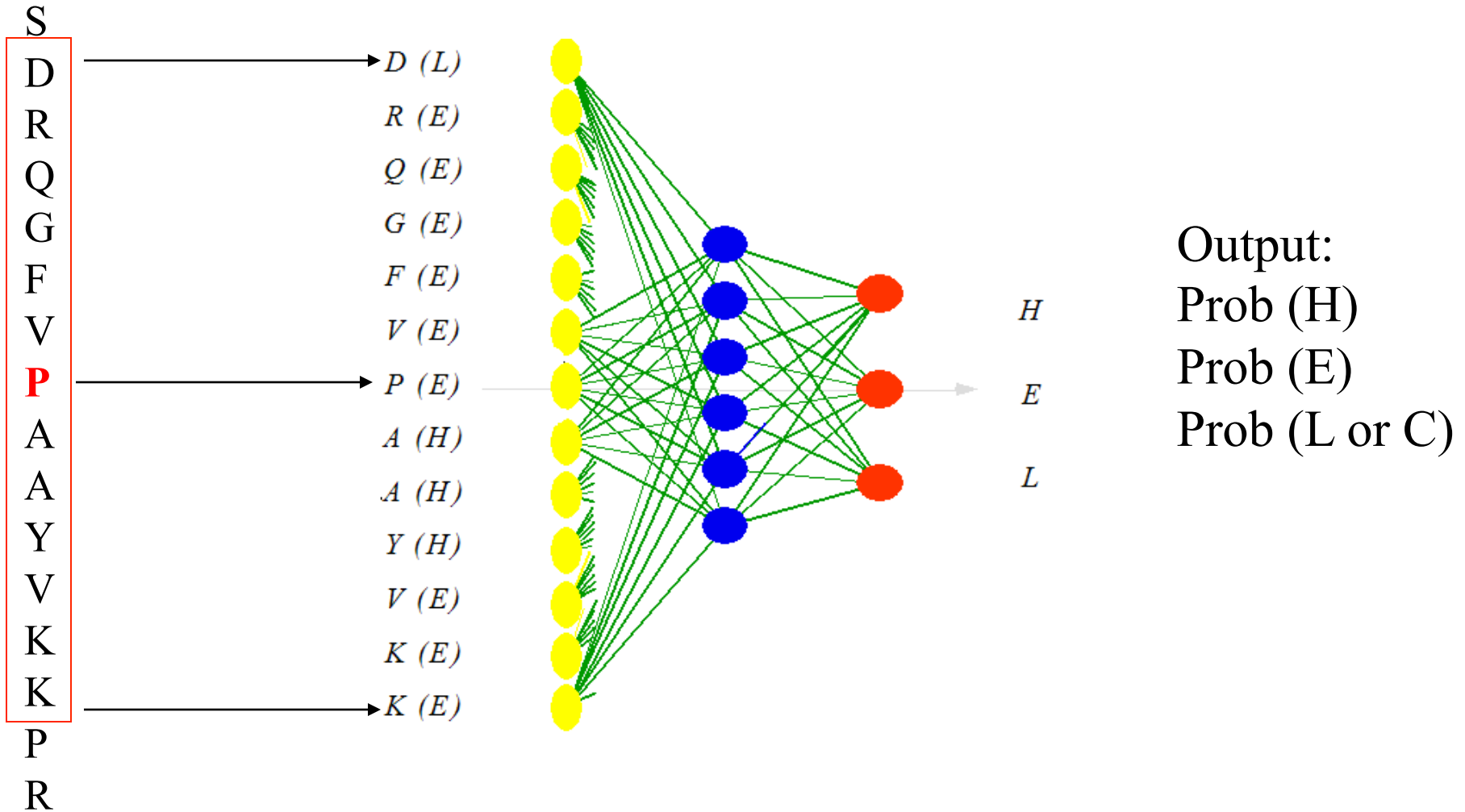
SS Sequence:

CCCCHHHHHHEEEEEHHHHEEEEECC

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)



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

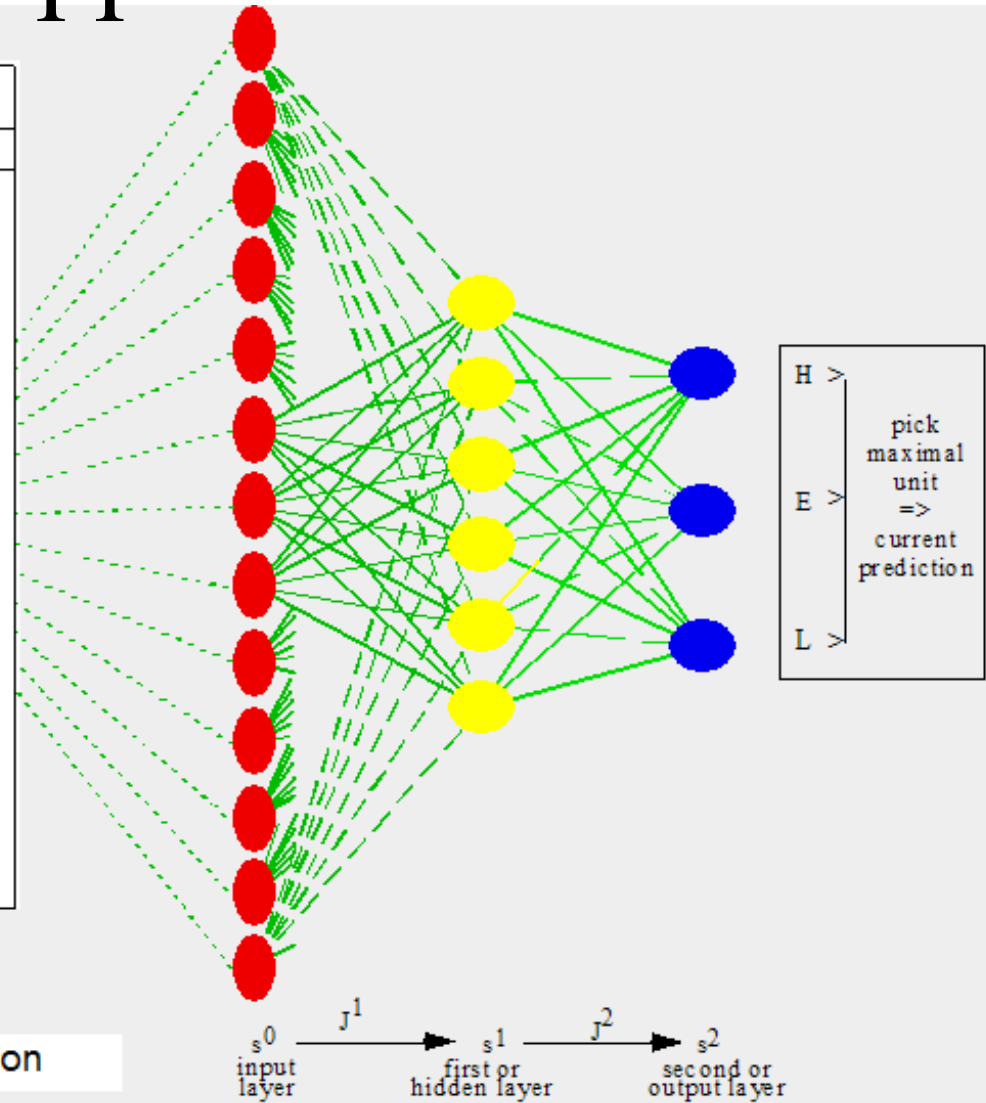
	1				50		
fyn_human	VTLFVALYDY	EARTEDDL	HKGEK	FQILN	SSEGDWWEAR	SLTTGETGYI	
yrk_chick	VTLFIALYDY	EARTEDDL	QKGEK	FHIIN	NTEGDWWEAR	SLSSGATGYI	
fgr_human	VTLFIALYDY	EARTEDDL	TKGEK	FHILN	NTEGDWWEAR	SLSSGKTGCI	
yes_chick	VTVFVALYDY	EARTTDDL	KKGER	FQIIN	NTEGDWWEAR	SIATGKTGYI	
src_avis2	VTTFVALYDY	ESRTE	TDLSF	KKGER	LQIVN	NTEGDWWLAH	SLTTGQTGYI
src_avis	VTTFVALYDY	ESRTE	TDLSF	KKGER	LQIVN	NTEGDWWLAH	SLTTGQTGYI
src_avisr	VTTFVALYDY	ESRTE	TDLSF	KKGER	LQIVN	NTEGDWWLAH	SLTTGQTGYI
src_chick	VTTFVALYDY	ESRTE	TDLSF	KKGER	LQIVN	NTEGDWWLAH	SLTTGQTGYI
stk_hydat	VTIFVALYDY	EARISE	EDLSF	KKGER	LQIIN	TADGDWWYAR	SLITNSEGYI
src_rsvpa	ESRIE	TDLSF	KKRER	LQIVN	NTEGTWWLAH	SLTTGQTGYI
hck_human	..IVVALYDY	EAIHHE	EDLSF	QKGDQ	MVVLE	ES.GEWVKAR	SLATRKEGYI
blk_mouse	..FVVALFDY	AAVND	RDLQV	LKGEK	LQVLR	.STGDWWLAR	SLVTGREGYV
hck_mouse	.TIVVALYDY	EAIHRE	EDLSF	QKGDQ	MVVLE	.EAGEWVKAR	SLATKKEGYI
lyn_human	..IVVALYPY	DGIHP	DDL	SS		.EHGEWVKAK	SLLTKKEGFI
lck_human	..LVIALHSY	EP	SHDGLGF	EKGEQ	LRI	QS.GEWVKAQ	SLTTGQEGFI
ss81_yeastALYPY	DADDD	deISF	EQNEI	LQVSD	.IEGRWVKAR	R.ANGETGII
abl_mouse	..LFVALYDF	VASGD	N	TKGEK	L	YnnGEWCEAQ	..TKNGQGW
abl1_human	..LFVALYDF	VASGD	N	TKGEK	L	YnnGEWCEAQ	..TKNGQGW
src1_drome	..VVVSLYDY	KSRDE	SDLSF	MKGDR	MEVID	DTESDWWRVV	NLTTRQEGLI
mysd_dicdiALYDF	DAESS	MEL	KEGD	L	QSSGDWWDAE	L..KGRRGKV
yfj4_yeastVALYSF	AGEES	GDL	RKGD	V	ksQNDWWTGR	V..NGREGIF
abl2_human	..LFVALYDF	VASGD	N	TKGEK	L	YNQNGEWSEV	RSKNG.QGW
tec_human	.EIVVAMYDF	QAAEG	HDLRL	ERGOE	YLILE	KNDVHWWRAR	D.KYGNEGYI
abl1_caee	..LFVALYDF	HGVGE	EQLSL	RKGD	Q	YNKNNEWCEA	RlrLGEIGW
txk_humanALYDF	LPREP	CNLAL	RRAEE	YLILE	KYNPHWWKAR	D.RLGNEGLI
yha2_yeast	VRRVRALYDL	TTNEP	DELSF	RKGD	V	QVYRDWWKGA	L..RGNMGI
abp1_sacexAEYDY	EAGED	NELTF	AENDK	I	FVDDDDWWLGE	LETTGQKGLF

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

Protein	Alignments	profile table
		GSAPD NTEKQ CVHIR LMYFW
:	:	:
G	G G G G	5
Y	Y Y Y Y 5 . .
I	I I E E 2 . . . 3
Y	Y Y Y Y 5 . .
D	D D D D 5
P	P P P P 5
E	A E A A	. . 3 2
D	V V E E 1 . . 2
G	G G G G	5
D	D D D D 5
P	P P P P 5
D	D T D D 4 . . 1
D	N Q N N 1 3 . . . 1
G	G N G G	4 1
V	V I V V 4 . 1
N	E P K K 1 . 1 . 1 2
P	P P P P 5
G	G G G G	5
T	T T T T 5
D	E K S A	. 1 1 . 1 . . 1 1
F	F F F F 5 . .
:	:	:



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

Reference: Rost and Sander. Proteins, 1994.

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.

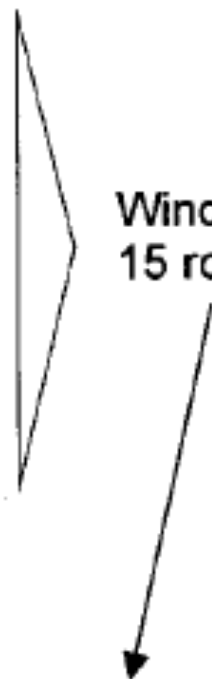
Reference: Jones, Journal of Molecular Biology, 1999.

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

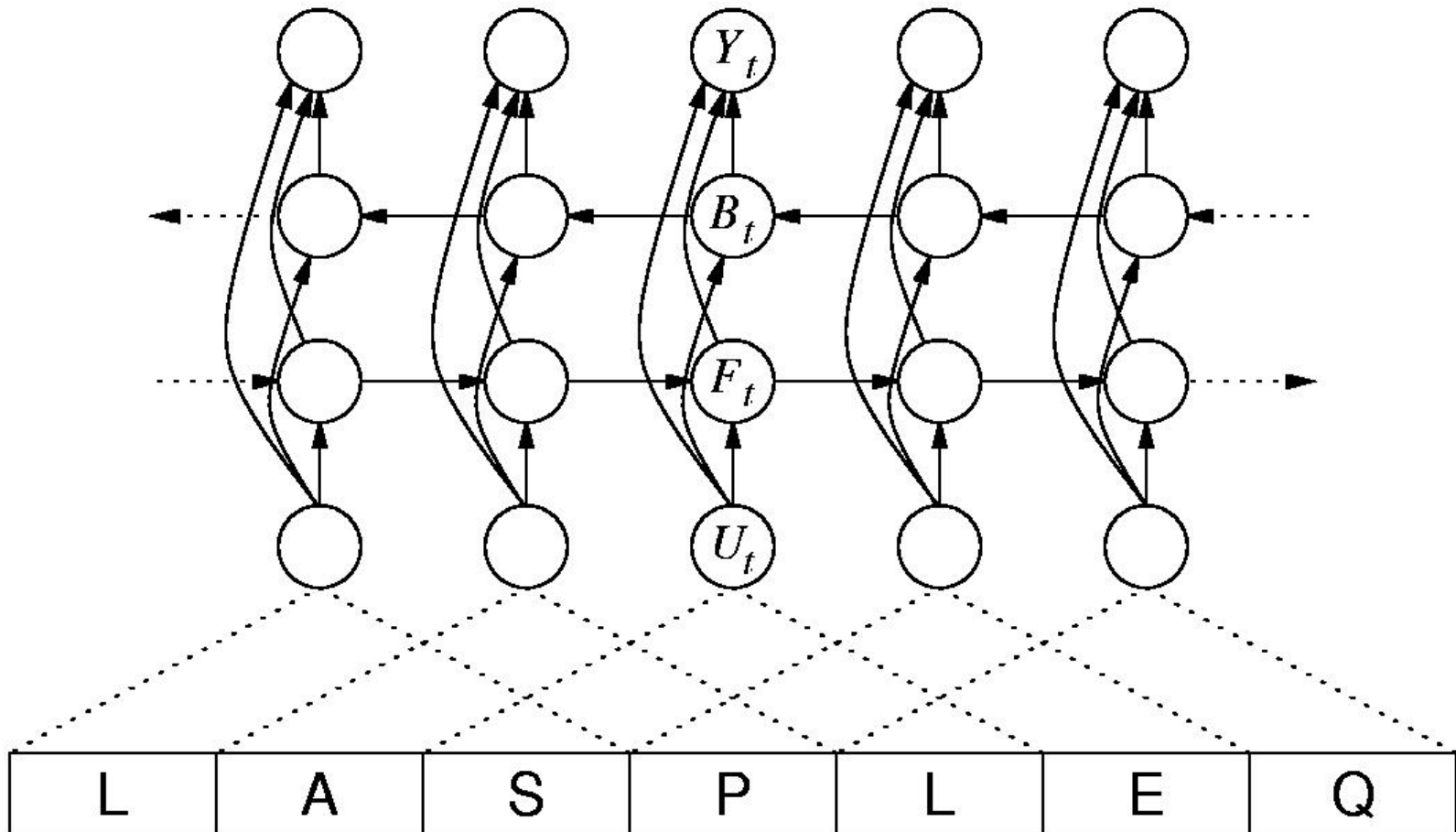
Reference: Jones, Journal of Molecular Biology, 1999.

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)

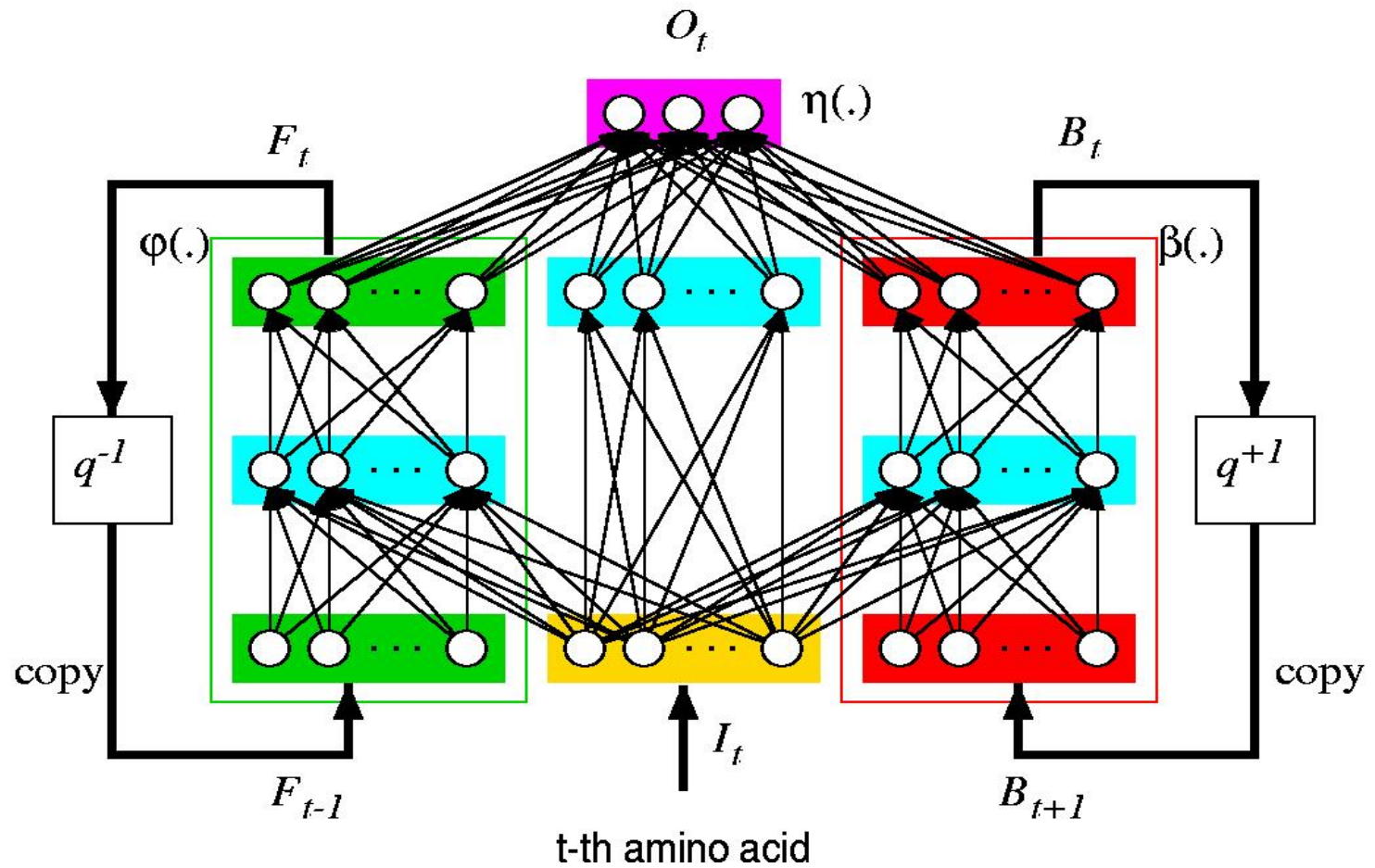
Pollastri et al.. Proteins, 2002.

Bi-directional Input Output Hidden Markov Model for SS Prediction



Baldi, 2004

1D-Recursive Neural Network



Baldi, 2004

Advantage and Disadvantages 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](http://www.cs.missouri.edu/~chengji/mlbioinfo/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)

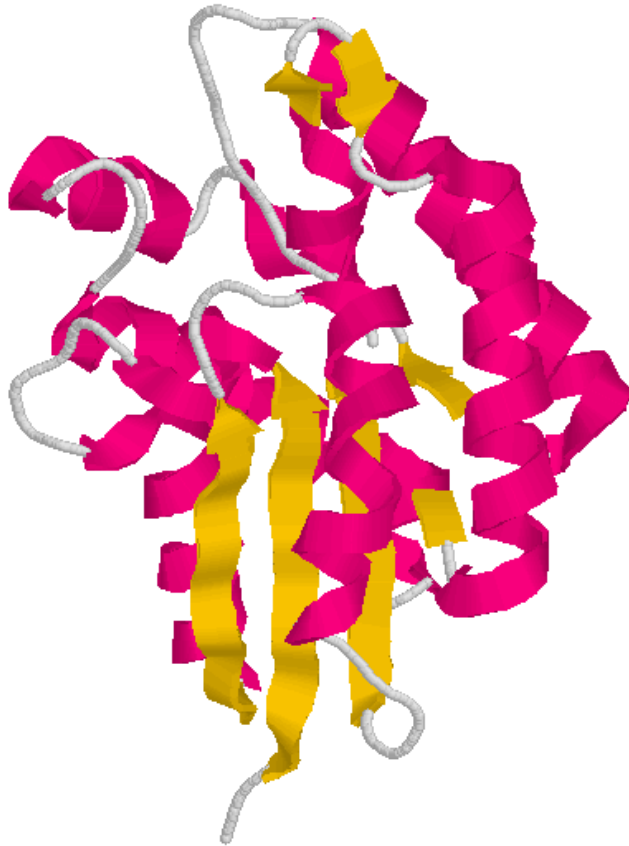
- Generate inputs and outputs (probability matrix or Position Specific Scoring Matrix)
- Train neural network using NNClass or Weka
- Test neural network on test dataset

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.

2D: Contact Map Prediction

3D Structure

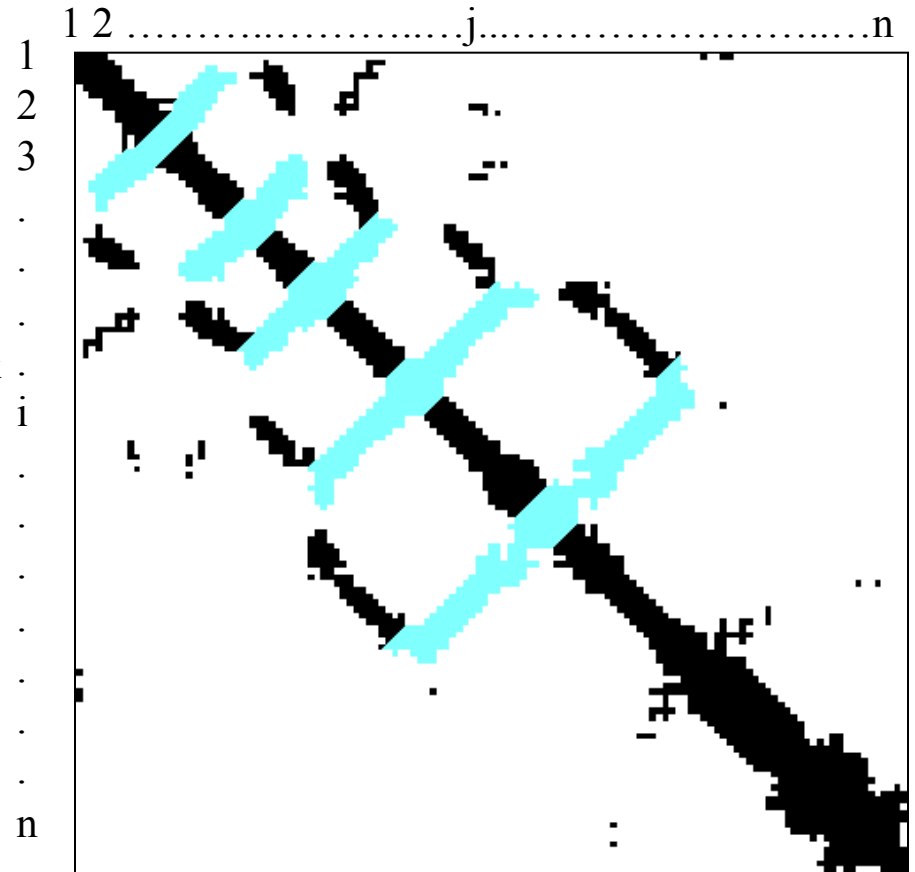


2D-Recursive
Neural Network



Support Vector
Machine

2D Contact Map



Distance Threshold = 8\AA

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