MACHINE LEARNING IN BIOINFORMATICS

Part 9: Neural Network Applications in Bioinformatics

František Mráz
KSVI MFF UK

**(Based on slides of Jianlin Cheng, PhD
Department of Computer Science
University of Missouri, Columbia)**

Neural Network Application in Bioinformatics

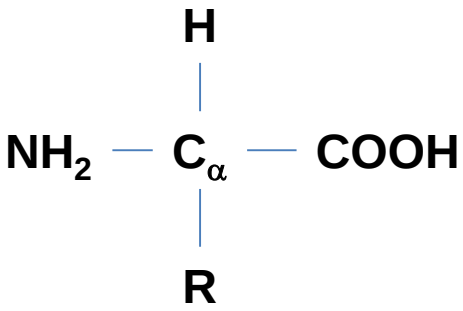
- Neural network have numerous applications in bioinformatics
- They are used in gene structure prediction, protein structure prediction, gene expression data analysis, … Almost anywhere when you need to do classification.
- Here we specifically focus on applying neural networks to protein structure prediction (**secondary structure, solvent accessibility,** disorder region, contact map).

Outline

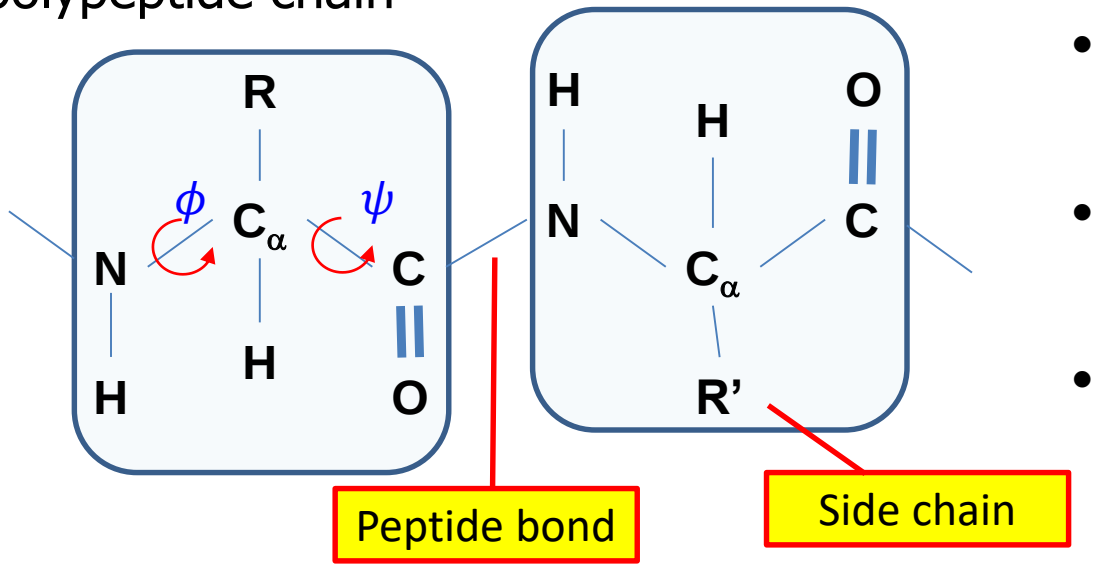
1. **Proteins**
2. Secondary structure
3. Protein structure determination
4. Using neural networks for protein structure prediction
5. Predicting solvent accessibility, disordered region, contact map,

Proteins

- A protein is a chain of amino acids joined by peptide bonds
- The structure of an amino acid



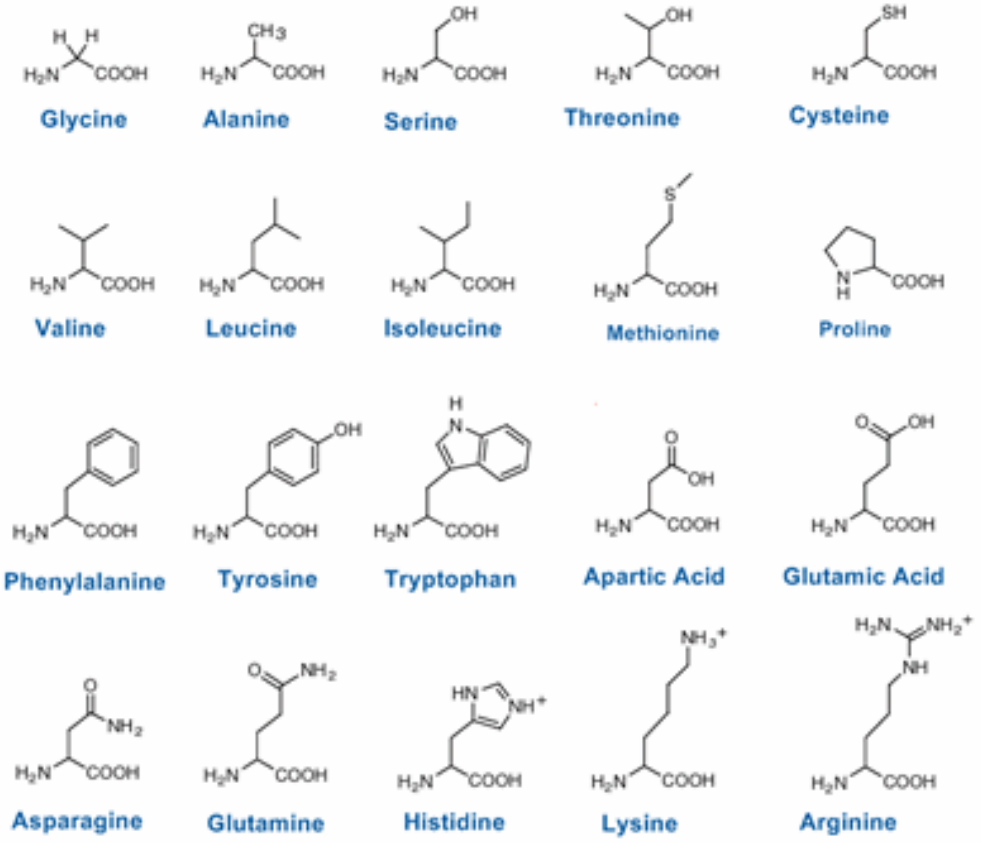
- amino acids are composed into a polypeptide chain



- N – C_α – C make up the backbone of the protein.
- Each amino acid has two rotational degrees of freedom ϕ and ψ
- The angle between C=O and N–H is always approximately 180°

Amino Acids

- http://groundupstrength.wdfiles.com/local--files/amino-acids/amino-acids.gif



Hierarchy of protein structure

- The **primary structure** is the chemical structure of the polypeptide chain(s) in a given protein, i.e. its sequence of amino acid residues that are linked by peptide bonds.
- The **secondary structure** is folding of the molecule that arises by linking the C=O and NH groups of the backbone together by means of hydrogen bonds.
- The **tertiary structure** is the 3D structure of the molecule consisting of secondary structures linked by "looser segments" of the polypeptide chain stabilized (primarily) by sidechain interactions.
  - Protein shape determines most of its function
  - Experimental determination of protein structure via x-ray crystallography is hard and time consuming
    - We would like to determine the structure of a protein from its sequence
- The **quaternary structure** is the aggregation of separate polypeptide chains into the functional protein.

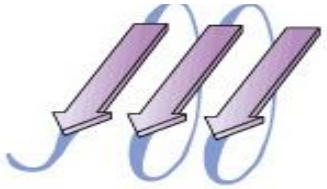
Four Levels of Protein Structure

a) Primary structure

–Ala–Glu–Val–Thr–Asp–Pro–Gly–

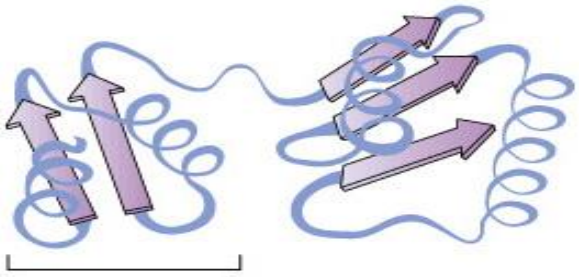
b) Secondary structure

α helix



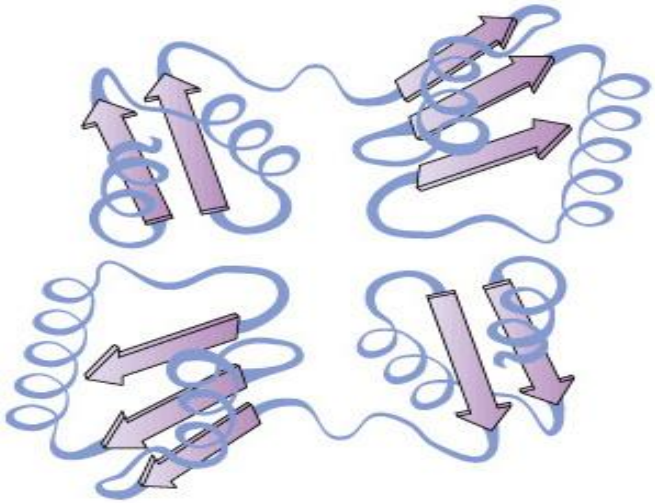
β sheet

c) Tertiary structure



domain

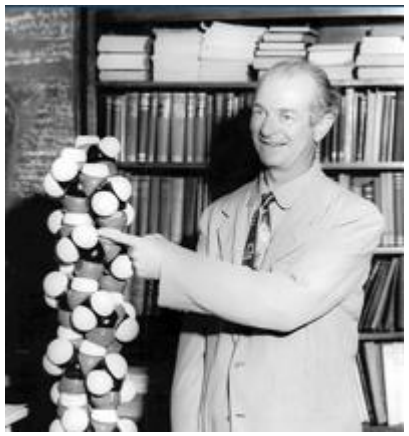
d) Quaternary structure



Outline

1. Proteins
2. **Secondary structure**
3. Protein structure determination
4. Using neural networks for protein structure prediction
5. Predicting solvent accessibility, disordered region, contact map,

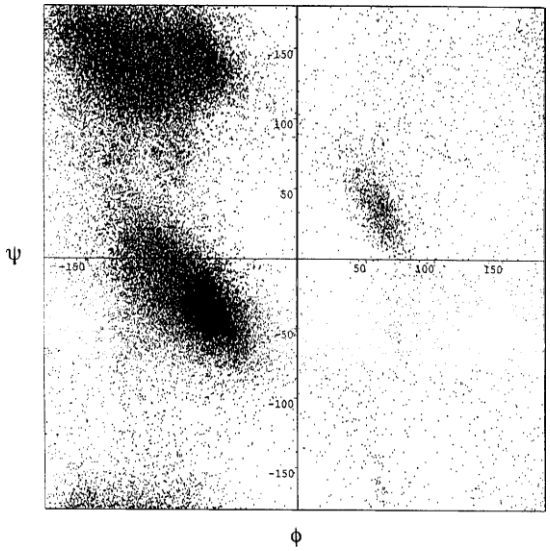
Secondary structure

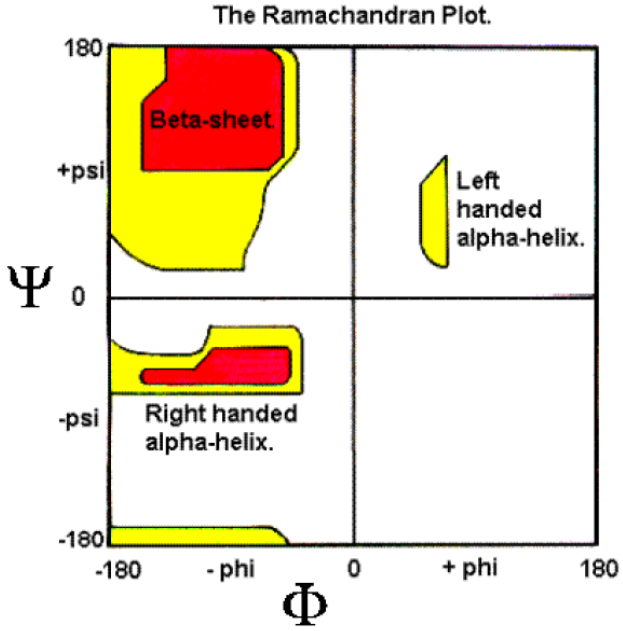


- Determined by hydrogen bond patterns
- 3-Class categories:
  - α helix
  - β sheet,
  - loop (or coil)
- First deduced by Linus Pauling et al.
- α helix and β sheet correspond to specific choices of the ϕ and ψ angles along the chain

Secondary structure Ramachandran plot

- plot of observed pairs of the ϕ and ψ angles in a collection of known protein structures

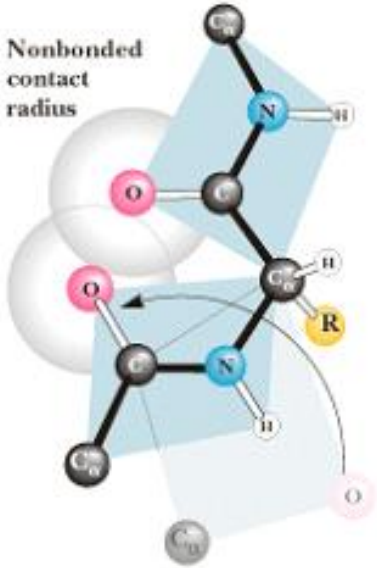




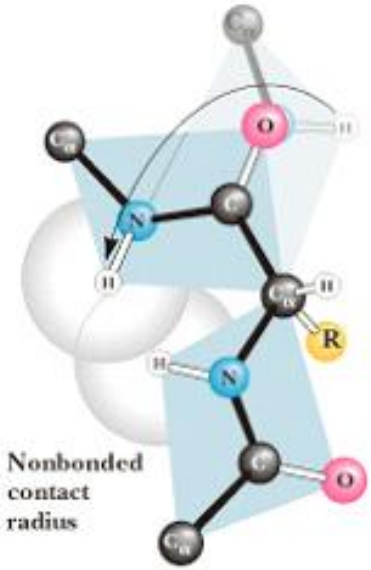
- Describes acceptable ϕ/ψ angles for individual AA's in a polypeptide chain.
- Helps determine what types of secondary structure are present

- The pairs near $\phi = -60°$ and $\psi = -60°$ correspond to helices.
- The pairs near $(-90°, 120°)$ correspond to strands.

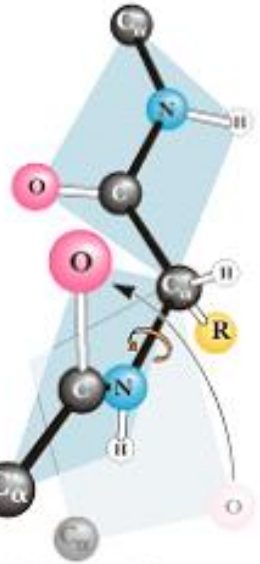
Not all ϕ/ψ angles are possible



$\phi = 0°, \psi = 180°$

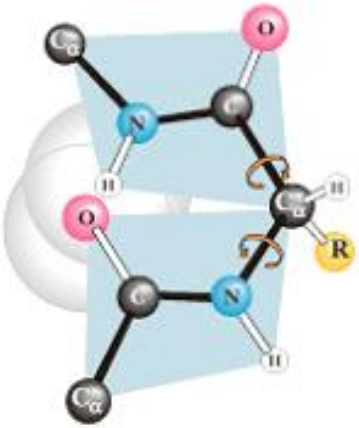


$\phi = 180°, \psi = 0°$



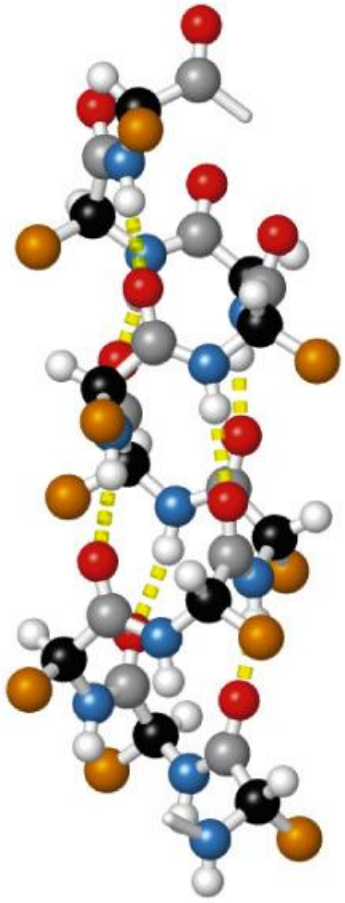
$\phi = -60°, \psi = 180°$

A further ϕ rotation of 120° removes the bulky carbonyl group as far as possible from the side chain

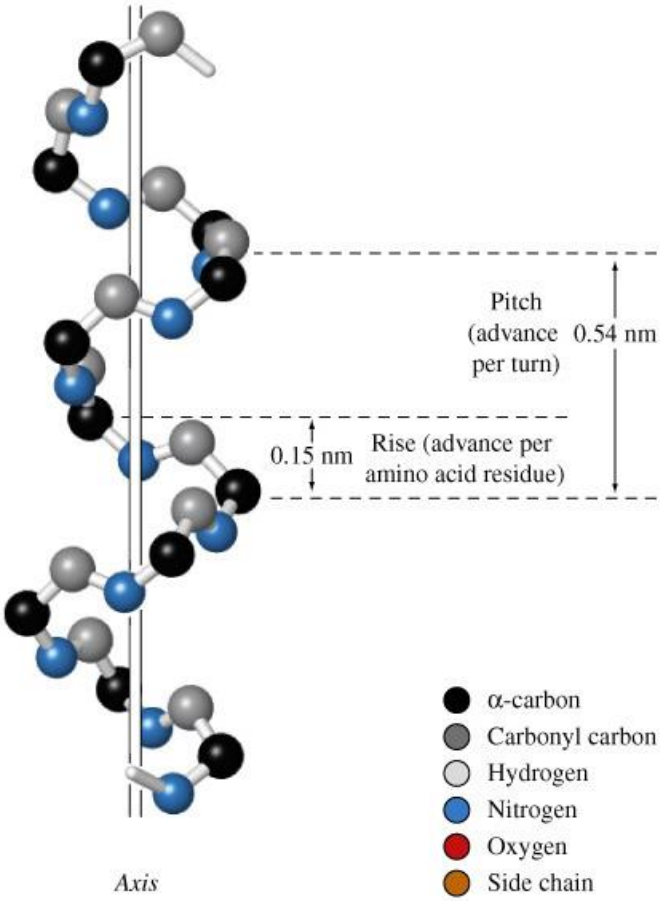


$\phi = 0°, \psi = 0°$

α Helix



Right-handed α-helix

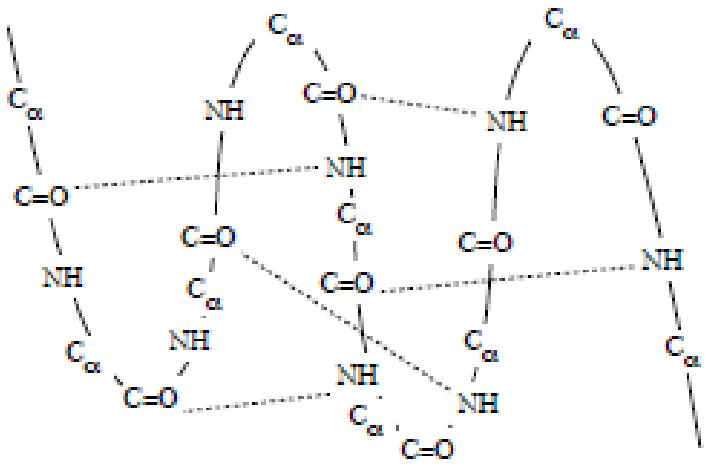


Axis

- Residues per turn: 3.6
- Rise per residue: 1.5 Angstroms
- Rise per turn (pitch): 3.6 x 1.5A = 5.4 Angstroms
- amino hydrogen H-bonds with carbonyl oxygen located 4 AA's away forms 13 atom loop

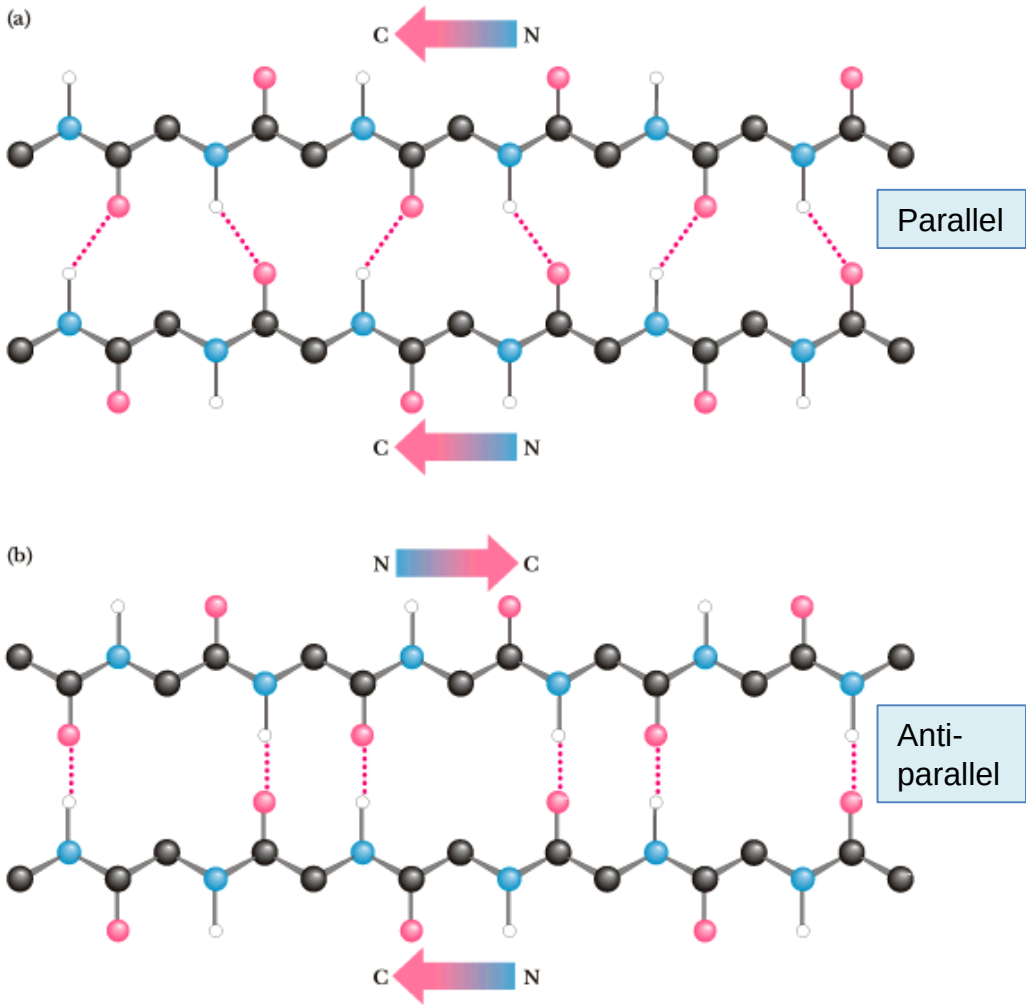
Helices

- Helices arise when hydrogen bonds occur between (the C=O group of) the amino acid at position i and (the NH group of) the amino acid at position $i + k$ (with $k = 3, 4 \text{ or } 5$), for a run of consecutive values of i.
- Most often, $k = 4 \text{ or } 5$ and the resulting structure is called an α helix, whereas $k = 3$ gives rise to a 3_{10} helix
- α helix:



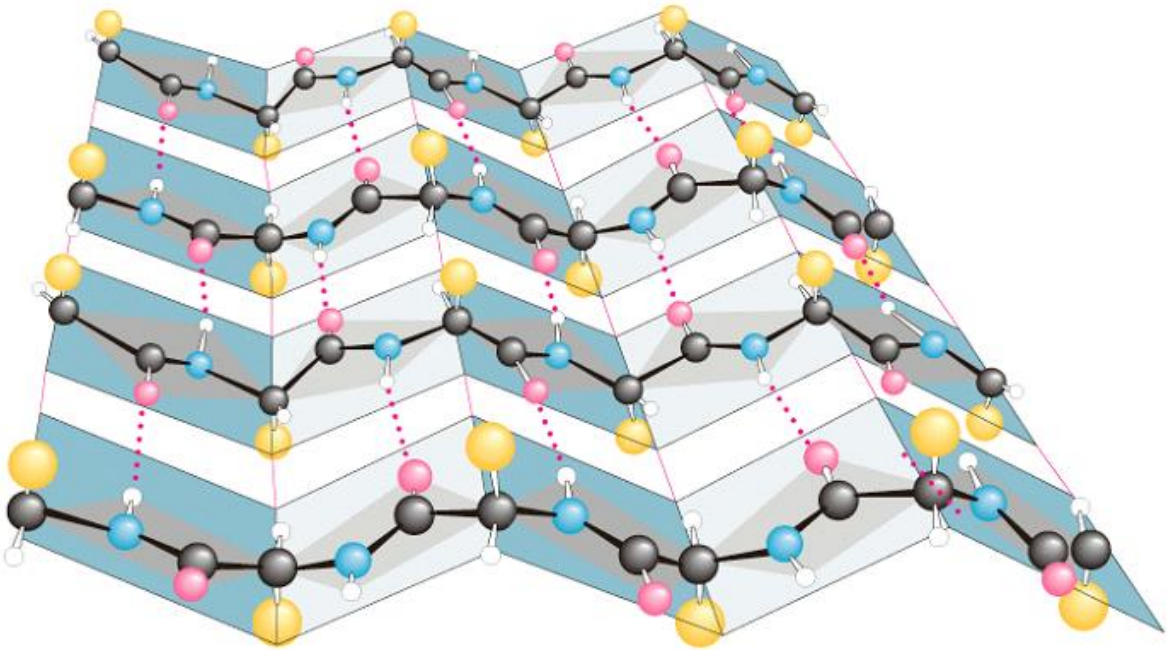
β Sheets

- β-sheets formed from multiple side-by-side beta-strands.
- Can be in parallel or anti-parallel configuration
- Anti-parallel beta-sheets more stable



β Sheets

- Side chains point alternately above and below the plane of the β sheet
- 2- to 15 β-strands/β-sheet
- Each strand made of ~ 6 amino acids

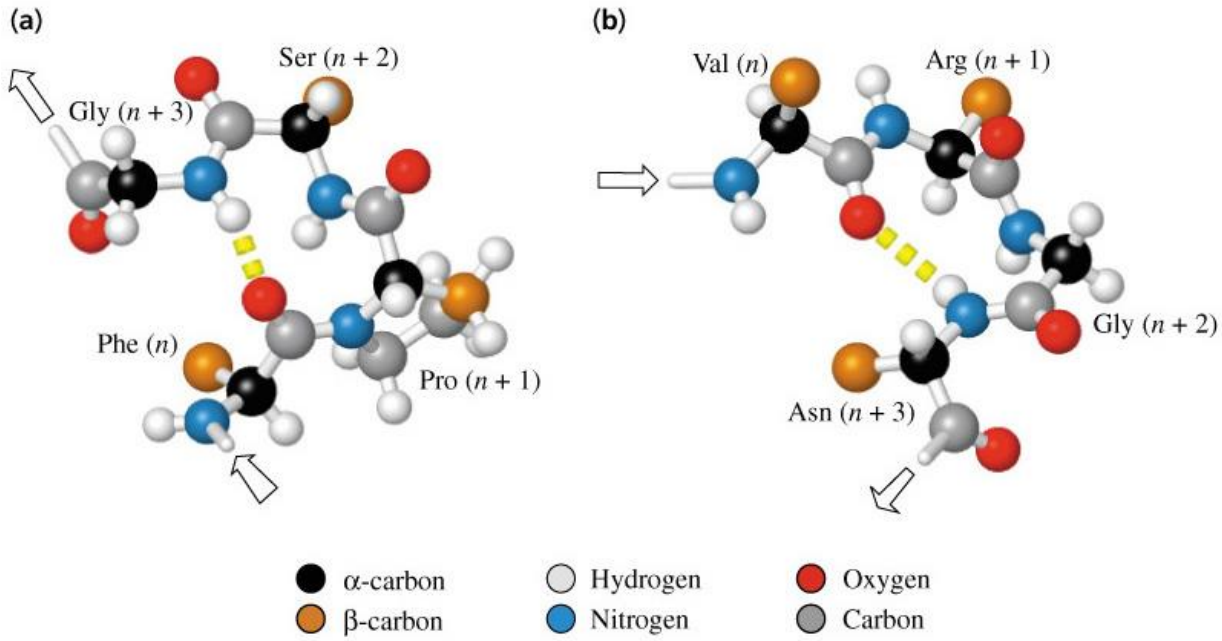


Loops and turns

- Loops
  - Loops usually contain hydrophilic residues.
  - Found on surfaces of proteins
  - Connect α helices and β sheets
- Turns
  - Loops with < 5 AA's are called turns
  - β turns are common

β Turns

- allows the peptide chain to reverse direction
- carbonyl C of one residue is H-bonded to the amide proton of a residue three residues away
- proline and glycine are prevalent in beta turns



*A region of secondary structure that is not a helix, a sheet, or recognizable turn is also called a **coil**.*

Outline

1. Proteins
2. Secondary structure
3. **Protein structure determination**
4. Using neural networks for protein structure prediction
5. Predicting solvent accessibility, disordered region, contact map,

3. Protein Structure Determination

- X-ray crystallography
  - X-ray: any size, accurate (1-3 Ångström (10^{-10} m)), sometimes hard to grow crystal

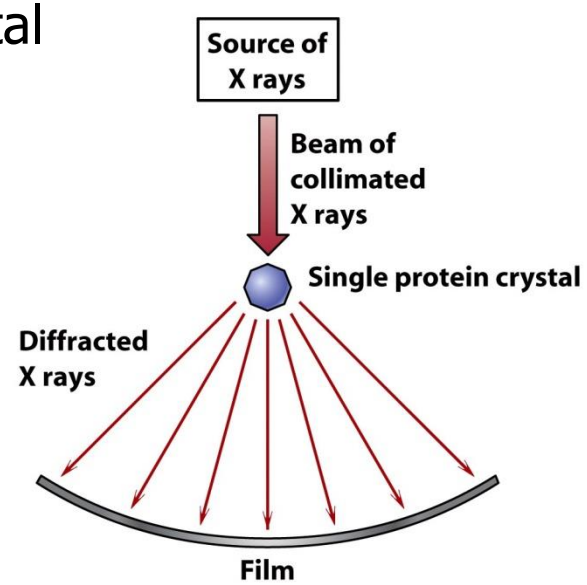


Figure 4-2a Principles of Biochemistry, 4/e
© 2006 Pearson Prentice Hall, Inc.

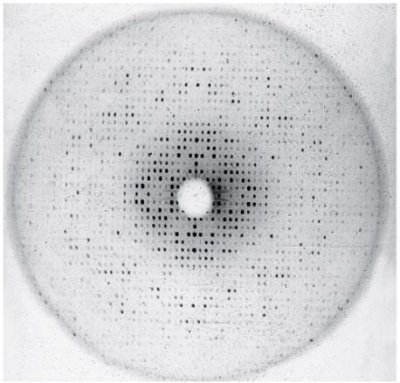
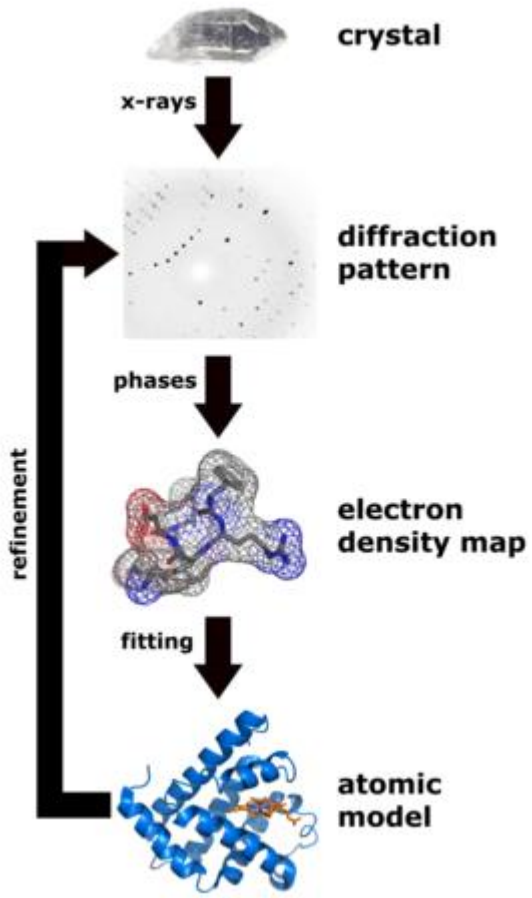


Figure 4-2b Principles of Biochemistry, 4/e
© 2006 Pearson Prentice Hall, Inc.

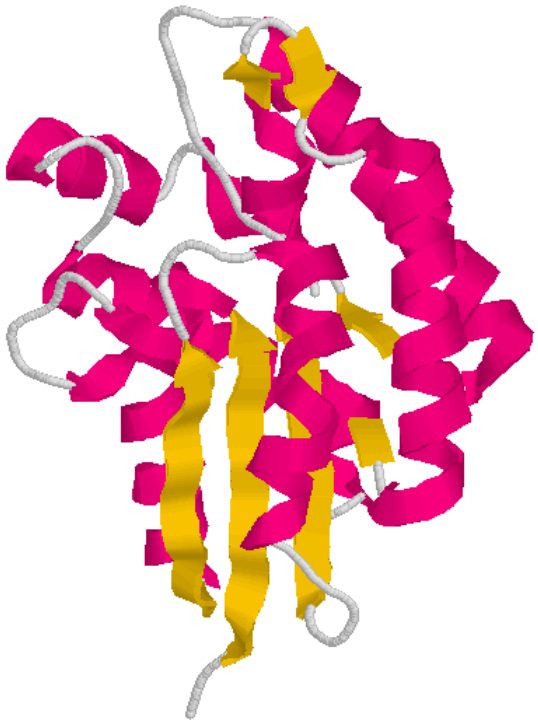
- Nuclear Magnetic Resonance (NMR) Spectroscopy
  - small to medium size, moderate accuracy, structure in solution

X-ray crystallography



Wikipedia, the free encyclopedia

1D: Secondary Structure Prediction



MWLKKFGINLLIGQSV…

↓

Neural Networks + Alignments

↓

H – α helix
S – β sheet
L – loop (or coil)

↓

LLLLHHHHHLLLSSSSS…

Accuracy: 78%

Outline

1. Proteins
2. Secondary structure
3. Protein structure determination
4. **Using neural networks for protein structure prediction**
5. Predicting solvent accessibility, disordered region, contact map,

How to Use Neural Network to Predict Secondary Structure

1. Create a data set with input sequences (X) and output labels (secondary structures)
2. Encode the input and output to neural network
3. Train neural network on the dataset (training dataset)
4. 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 Ångström, determined by X-ray crystallography)
- Remove proteins with chain-break ($C_\alpha - C_\alpha$ distance > 4 Ångström)
- Remove redundancy (filter out very similar sequences using BLAST)
- Use DSSP program (Kabsch and Sander, 1983) to assign secondary structure to each residue.
  - DSSP is a database of secondary structure assignments (and much more) for all protein entries in the Protein Data Bank (PDB). DSSP is also the program that calculates DSSP entries from PDB entries. DSSP does **not** predict secondary structure.

    [*A series of PDB related databases for everyday needs.*
    Wouter G Touw, Coos Baakman, Jon Black, Tim AH te Beek, E Krieger, Robbie P Joosten, Gert Vriend.
    Nucleic Acids Research 2015 January; 43(Database issue): D364-D368.]

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 should be 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:

LLLLHHHHHHEEEEEHHHHEEEEEELL

H – α helix
E – β sheet – extended strand
C – loop (or coil)

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

Protein Sequence:

MWLKKFGINLLIGQSVQTRSWYYCKRA

One-hot-encoding

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

SS Sequence:

LLLLHHHHHHEEEEEHHHHEEEEEELL

100: Helix
010: Extended strand
001: Loop (or coil)

Similarly for 20 different amino acids

Use a Window to Account for Context

Protein Sequence:

MWLKKFGINLLIGQSVQTRSWYYCKRA

SS Sequence:

CCCCHHHHHHEEEEEEHHHHEEEEEEECC

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

Use an Extra Input to Account for Nand C- Terminal Boundary

Protein Sequence:

MWLKKFGINLLIGQSVQTRSWYYCKRA

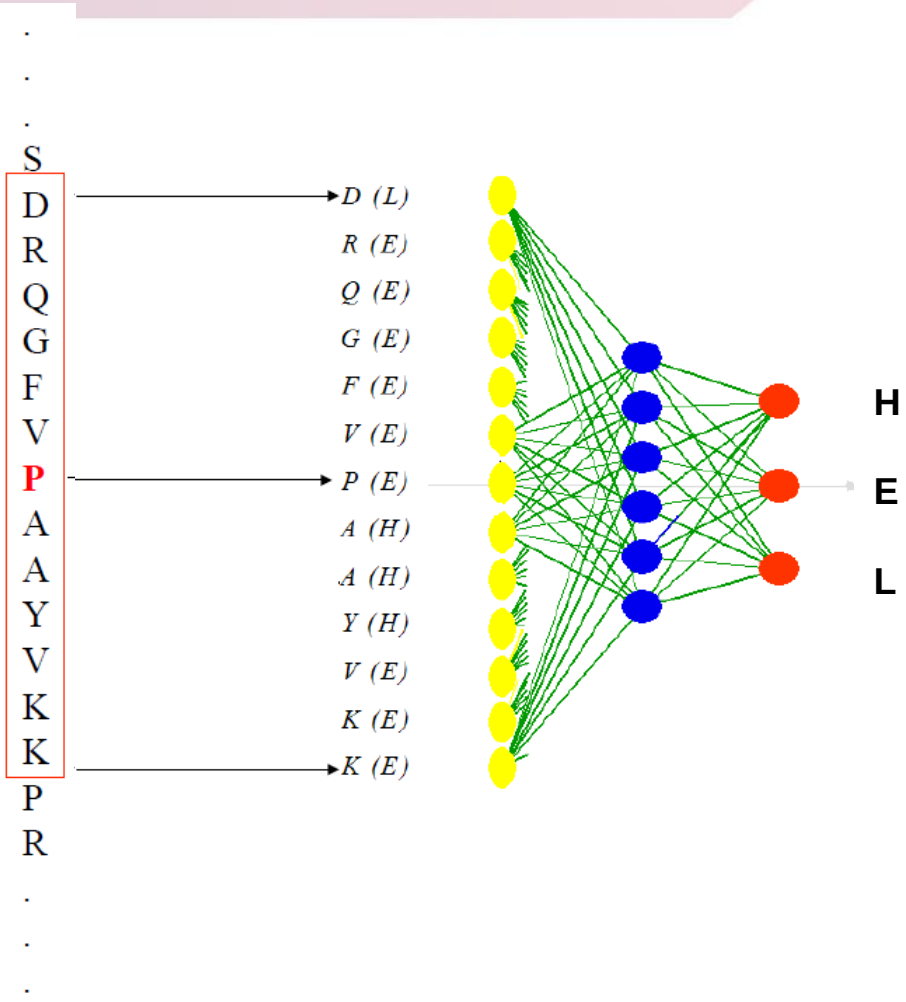
SS Sequence:

CCCCHHHHHHEEEEEHHHHEEEEEECC

Add an extra input for each position to indicate if it is out of the boundary of the sequence (‘spacer’).

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)

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 EARTEDDLSF HKGEKFQILN SSEGDWWEAR SLTTGETGYI
yrk_chick VTLFIALYDY EARTEDDLSF QKGEKFHIIN NTEGDWWEAR SLSSGATGYI
fgr_human VTLFIALYDY EARTEDDLTF TKGEKFHILN NTEGDWWEAR SLSSGKTGCI
yes_chick VTVFVALYDY EARTTDDLSF KKGERFQIIN NTEGDWWEAR SIATGKTGYI
src_avis2 VTTFVALYDY ESRTETDLSF KKGERLQIVN NTEGDWWLAH SLTTGQTGYI
src_aviss VTTFVALYDY ESRTETDLSF KKGERLQIVN NTEGDWWLAH SLTTGQTGYI
src_avisr VTTFVALYDY ESRTETDLSF KKGERLQIVN NTEGDWWLAH SLTTGQTGYI
src_chick VTTFVALYDY ESRTETDLSF KKGERLQIVN NTEGDWWLAH SLTTGQTGYI
stk_hydat VTIFVALYDY EARISEDLSF KKGERLQIIN TADGDWWYAR SLITNSEGYI
src_rsvpa .......... ESRIETDLSF KKRERLQIVN NTEGTWWLAH SLTTGQTGYI
hck_human ..IVVALYDY EAIHHEDLSF QKGDQMVVLE ES.GEWWKAR SLATRKEGYI
blk_mouse ..FVVALFDY AAVNDRDLQV LKGEKLQVLR .STGDWWLAR SLVTGREGYV
hck_mouse .TIVVALYDY EAIHREDLSF QKGDQMVVLE .EAGEWWKAR SLATKKEGYI
lyn_human ..IVVALYPY DGIHPDDLSF KKGEKMKVLE .EHGEWWKAK SLLTKKEGFI
lck_human ..LVIALHSY EPSHDGDLGF EKGEQLRILE QS.GEWWKAQ SLTTGQEGFI
ss81_yeast.....ALYPY DADDDdeISF EQNEILQVSD .IEGRWWKAR R.ANGETGII
abl_mouse ..LFVALYDF VASGDNTLSI TKGEKLRVLG YnnGEWCEAQ ..TKNGQGWV
abl1_human..LFVALYDF VASGDNTLSI TKGEKLRVLG YnnGEWCEAQ ..TKNGQGWV
src1_drome..VVVSLYDY KSRDESDLSF MKGDRMEVID DTESDWWRVV NLTTRQEGLI
mysd_dicdi.....ALYDF DAESSMELSF KEGDILTVLD QSSGDWWDAE L..KGRRGKV
yfj4_yeast....VALYSF AGEESGDLPF RKGDVITILK ksQNDWWTGR V..NGREGIF
abl2_human..LFVALYDF VASGDNTLSI TKGEKLRVLG YNQNGEWSEV RSKNG.QGWV
tec_human .EIVVAMYDF QAAEGHDLRL ERGQEYLILE KNDVHWWRAR D.KYGNEGYI
abl1_caeel..LFVALYDF HGVGEEQLSL RKGDQVRILG YNKNNEWCEA RlrLGEIGWV
txk_human .....ALYDF LPREPCNLAL RRAEEYLILE KYNPHWWKAR D.RLGNEGLI
yha2_yeastVRRVRALYDL TTNEPDELSF RKGDVITVLE QVYRDWWKGA L..RGNMGIF
abp1_sacex.....AEYDY EAGEDNELTF AENDKIINIE FVDDDWWLGE LETTGQKGLF
```

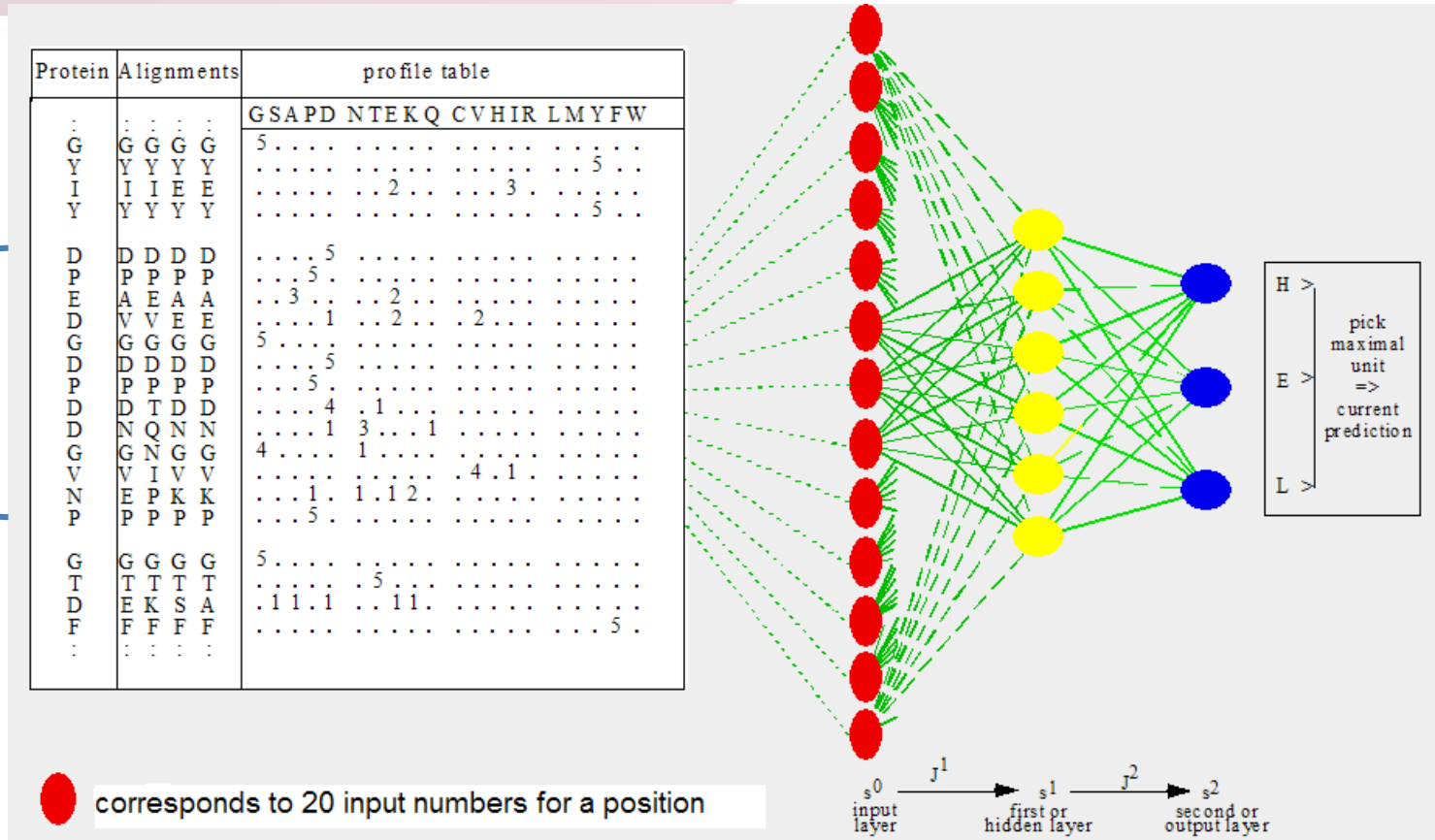
B. Rost, 2005

How to Find Homologous Sequences and Generate Alignments

Position Specific Iterated BLAST

- 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.

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.

PHD Approach

Local alignment 13 adjacent positions

Global statistics on whole protein

Input local in the sequence – for each residue (13) use

- 20 values from the profile
- 1 'spacer' – yes/no position out of the sequence
- 2 number of insertions and deletions in the alignment
- 1 'cons' conservation weight

Global statistics

- 20 amino acid composition
- 4 protein length ($\leq 60, \leq 120, \leq 240, > 240$)
- 8 distances of the window from the protein end ($\leq 40, \leq 30, \leq 20, \leq 10$)

First level NN:

- Hidden layer and
- 3 outputs (helix, strand, other) for the central residue

Second level NN:

- Input for each residue (13)
  - 3 output of the first level
  - 'spacer'
  - 'cons'
- Hidden layer and output similar to the first level

PhD Approach

- The second level introduces a correlation between adjacent residues, otherwise, e.g., too short helices are outputted
- The distribution of examples is uneven
  - 32% of the residues in helices,
  - 21% in strand, and
  - 47% in loop
- A balanced training is used – it improved results for less frequent states but not decreased accuracy for high frequency residues ⇒ lower overall accuracy
- Final decision – a **jury**: an arithmetic average over 4 differently trained networks: all combinations of the first level NN with balanced/unbalanced training and second level NN with balanced/unbalanced training
- Final prediction = unit with the maximal value

PSI-PRED Approach

- PSI-PRED does not use probability matrix instead it uses the another kind of profile: Position Specific Scoring Matrix (PSSM) 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.

What is the difference between probability matrix and PSSM?

Reference: Jones, Journal of Molecular Biology, 1999.

PSI-PRED Input

Position-based scoring matrix used

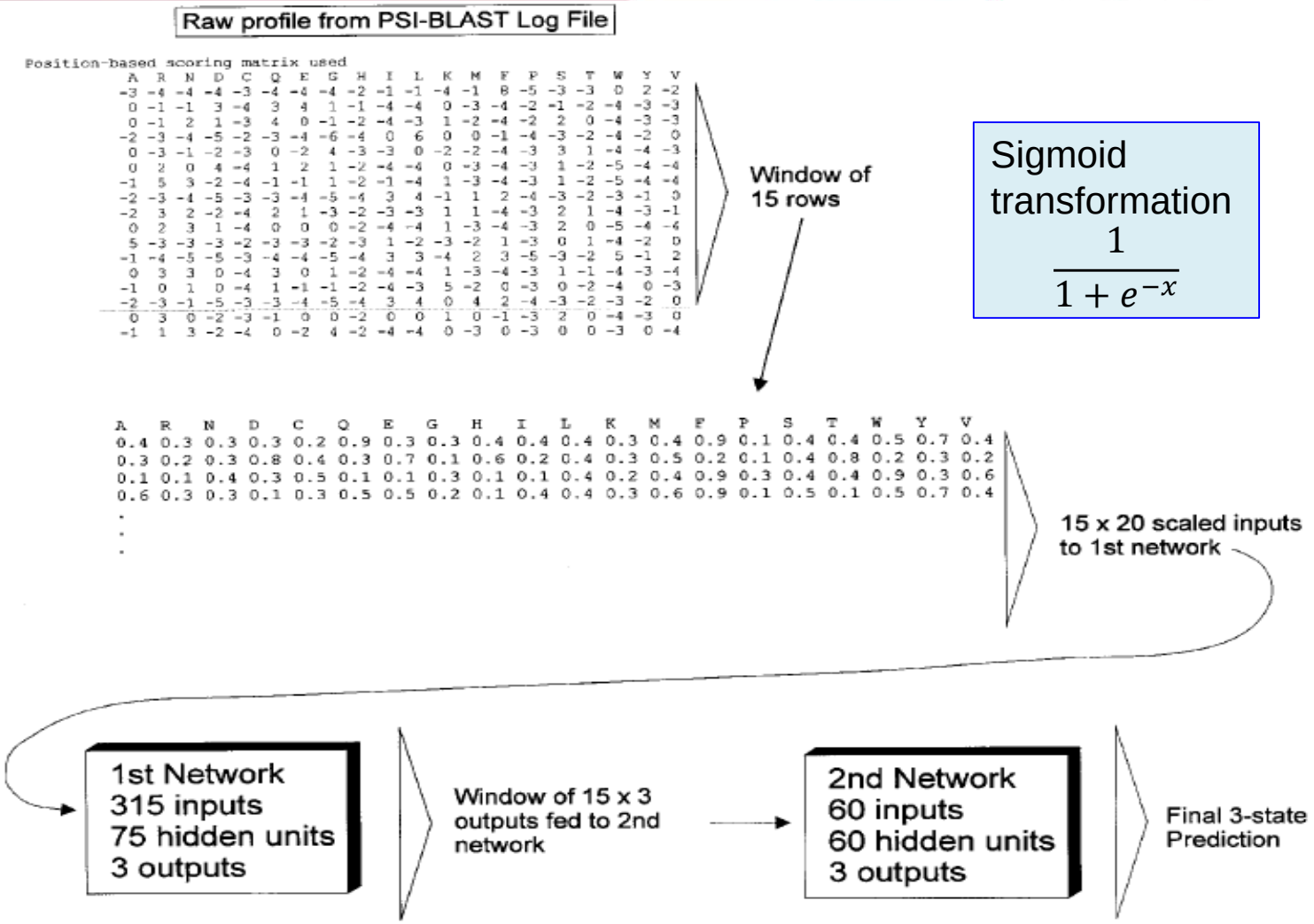
A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
-3	-4	-4	-4	-3	-4	-4	-4	-2	-1	-1	-4	-1	8	-5	-3	-3	0	2	-2
0	-1	-1	3	-4	3	4	1	-1	-4	-4	0	-3	-4	-2	-1	-2	-4	-3	-3
0	-1	2	1	-3	4	0	-1	-2	-4	-3	1	-2	-4	-2	2	0	-4	-3	-3
-2	-3	-4	-5	-2	-3	-4	-6	-4	0	6	0	0	-1	-4	-3	-2	-4	-2	0
0	-3	-1	-2	-3	0	-2	4	-3	-3	0	-2	-2	-4	-3	3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

PSI-PRED

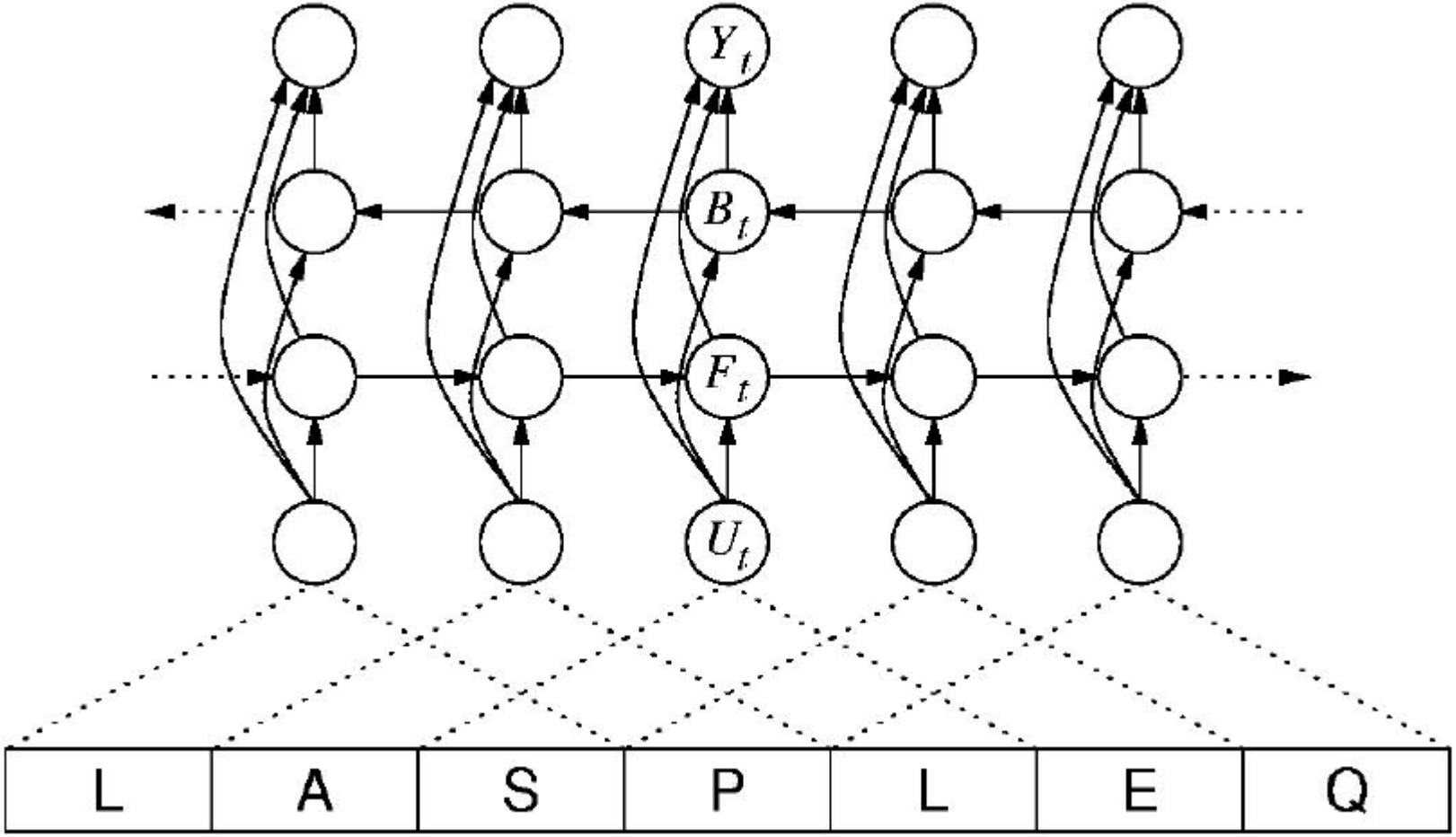


SSpro Approach

- 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



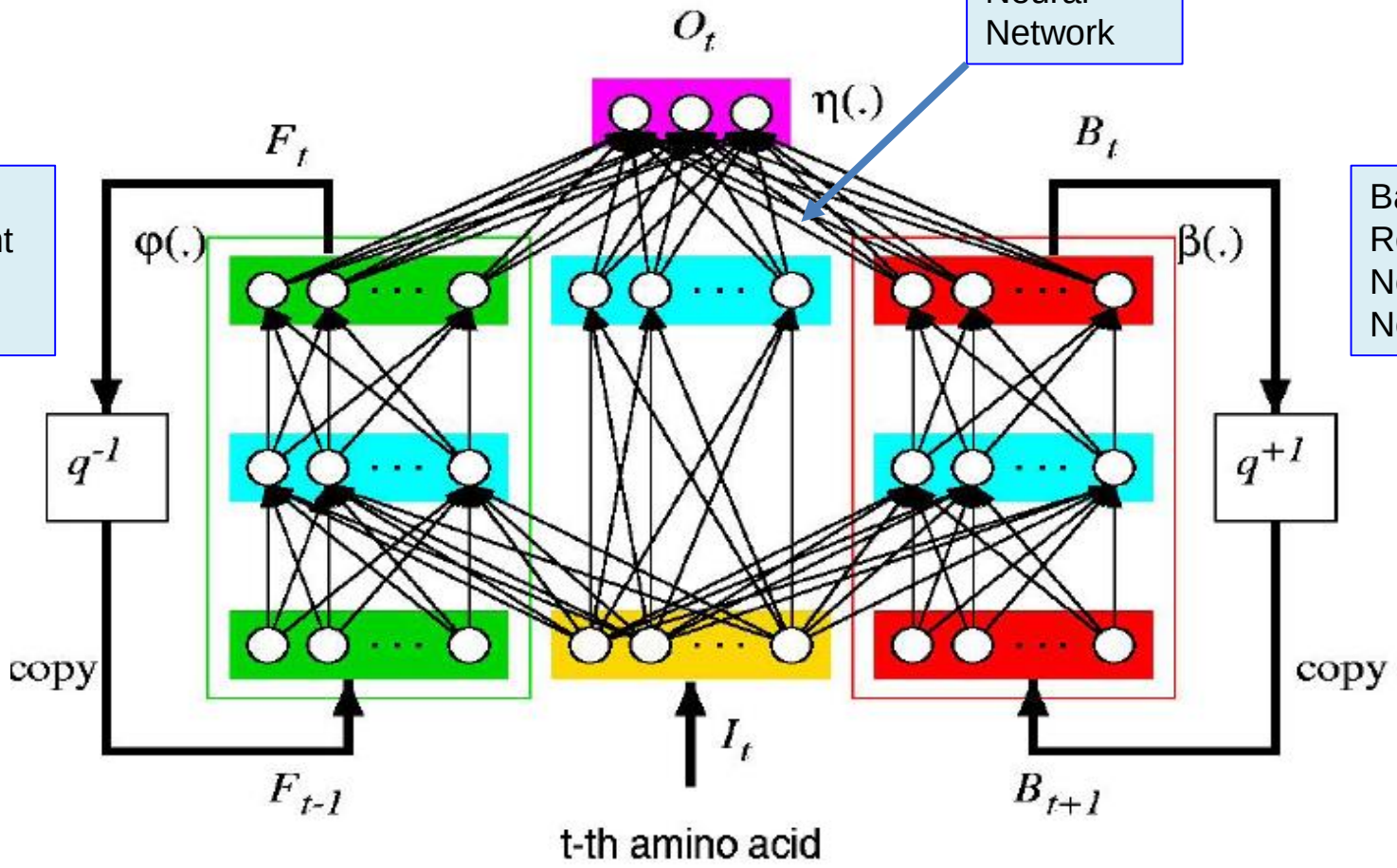
Baldi, 2004

1D-Recursive Neural Network

Feed Forward Neural Network

Forward Recurrent Neural Network

Backward Recurrent Neural Network



Baldi, 2004

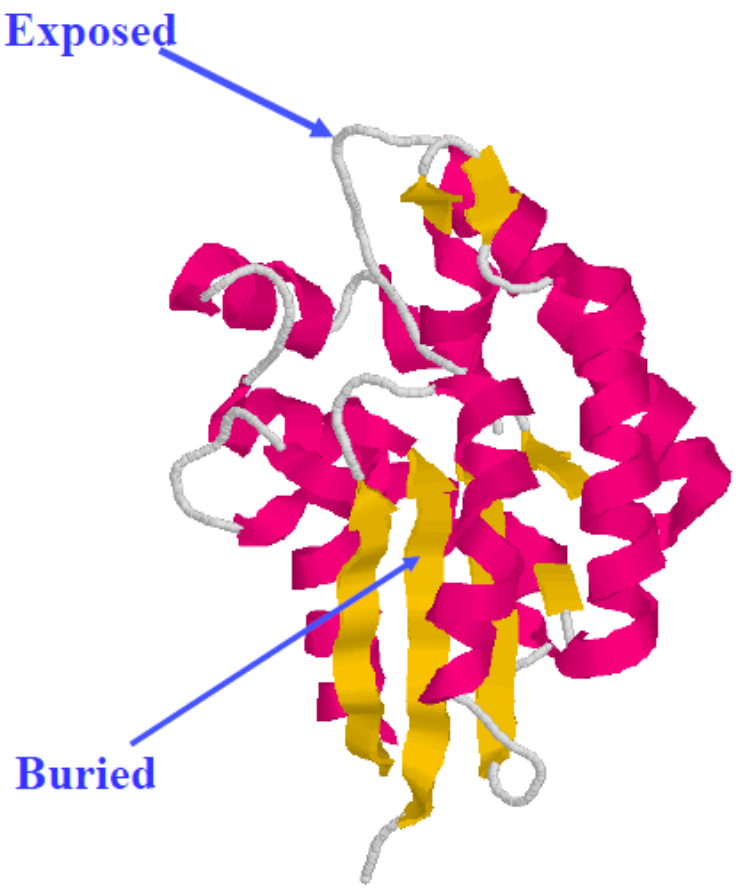
Advantage and Disavantages of SSpro

- 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.

Outline

1. Proteins
2. Secondary structure
3. Protein structure determination
4. Using neural networks for protein structure prediction
5. **Predicting solvent accessibility, disordered region, contact map**

1D: Solvent Accessibility Prediction



MWLKKFGINLLGAQVSBG…

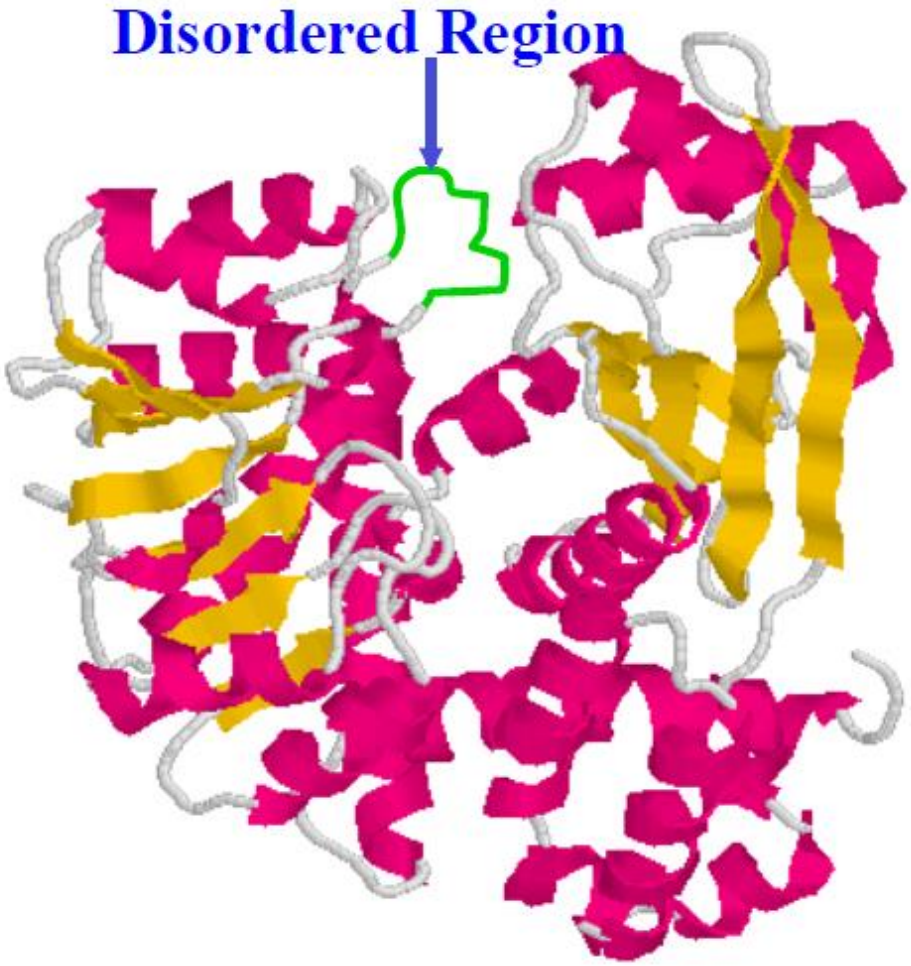
Neural networks+alignments

eeeeeebbbbbbbbeeeeebbb…

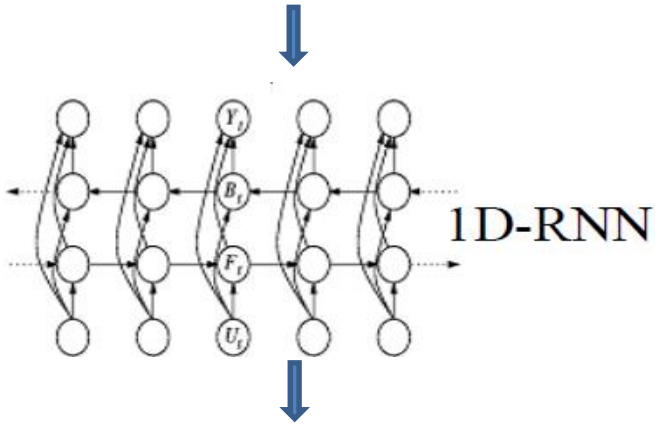
Accuracy: 79%

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

1D: Disordered Region Prediction Using Neural Networks



MWLKKFGINLLGAQVSBG…



oooooddddddooooooo…

93% TP at 5% FP

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

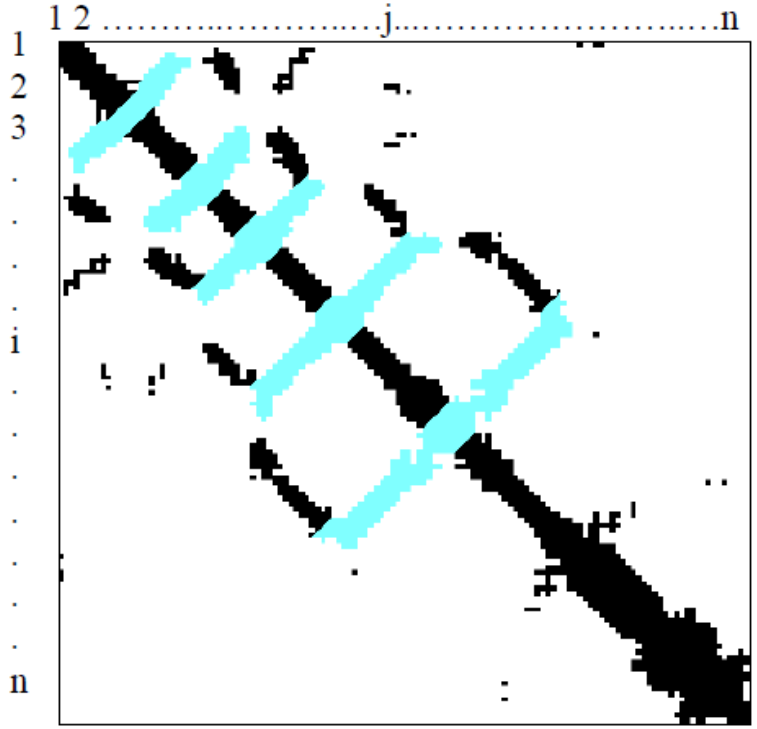
2D: Contact Map Prediction

3D Structure

2D-Recursive Neural Network

Support Vector Machine

2D Contact Map



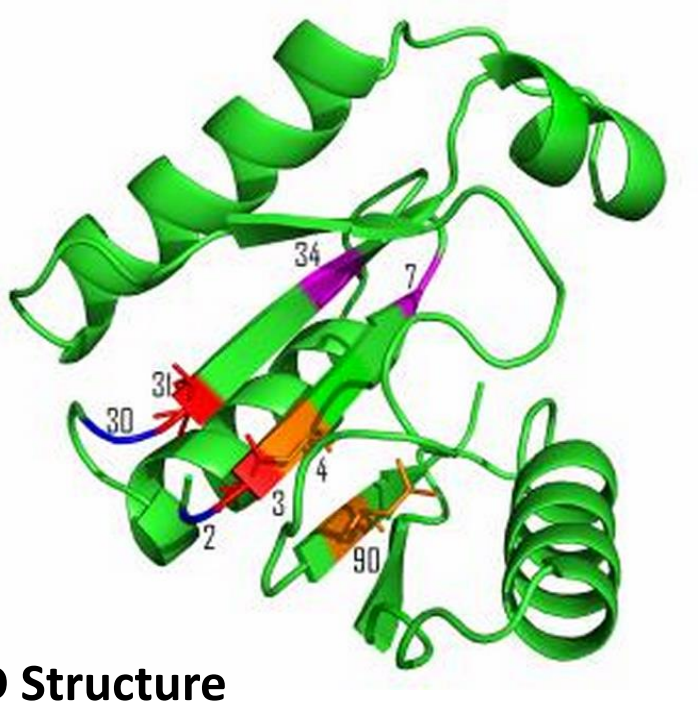
Distance Threshold = 8A°

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

Residue-Residue Contact Prediction

1D Sequence

SDDEVYQYIVSQVKQYGIEPAELLSRKYGDKAKYHLSQRW



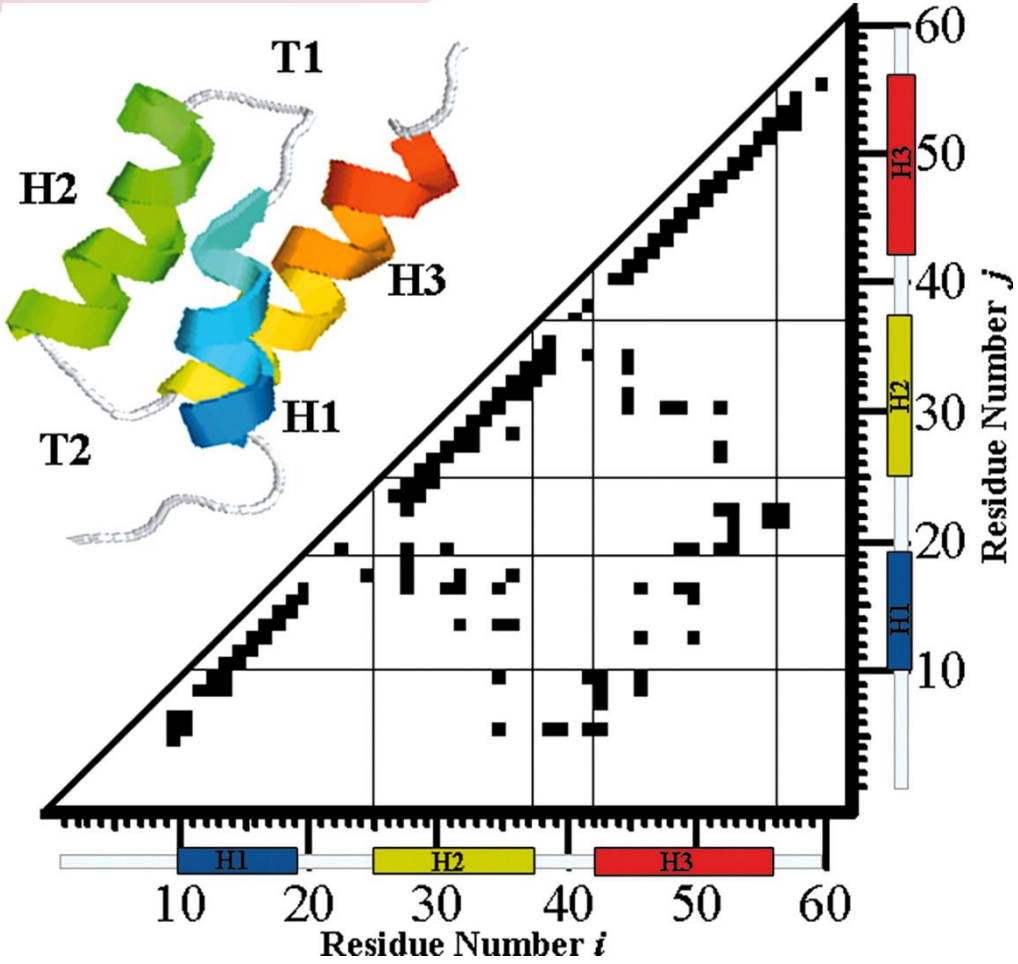
3D Structure

Objective:

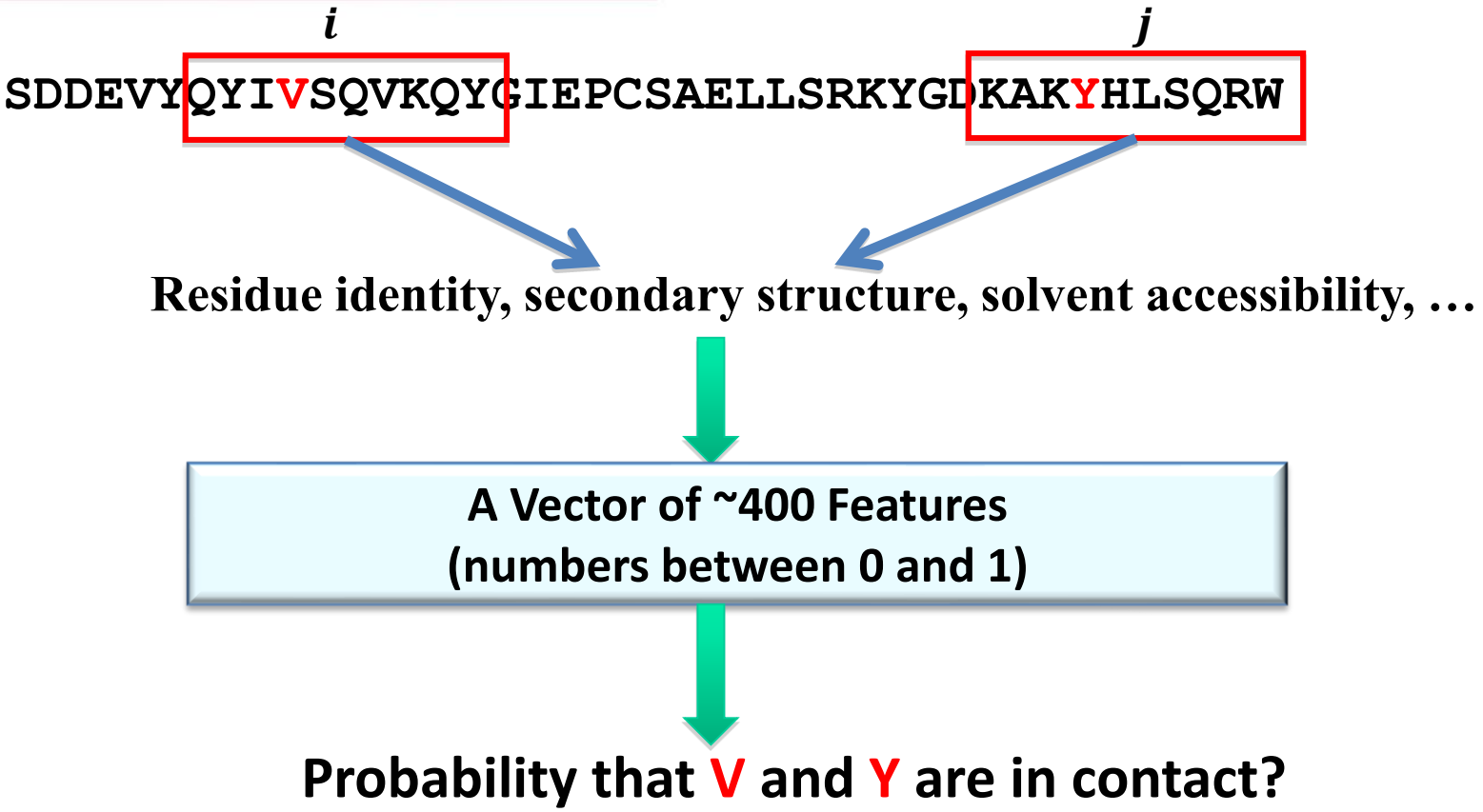
Predict if two residues (i, j) are in contact (spatially close), i.e. $distance(i\ j) < 8$ Ångstrom

Eickholt & Cheng, 2012

Visualization of a Contact Map



A Binary Classification Problem



Cheng & Baldi, 2007; Tegge et al., 2009; Eickholt & Cheng, 2012

Solvent Accessibility Input Features

SDDEVYQYIVSQVKQYGIEPCSAELLSRKYGDKAKYHLSQRW

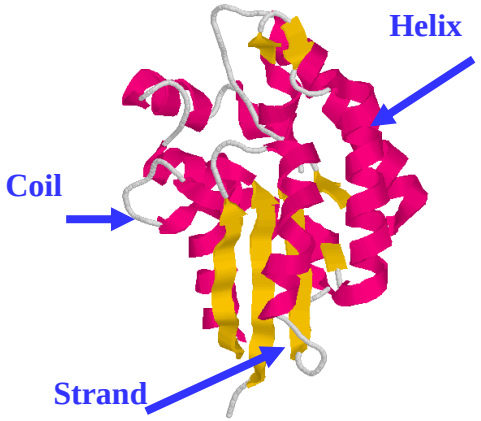
20 binary numbers

A 10000000000000000000
C 01000000000000000000
D 00100000000000000000

.
.
.
.
.
.
.
.

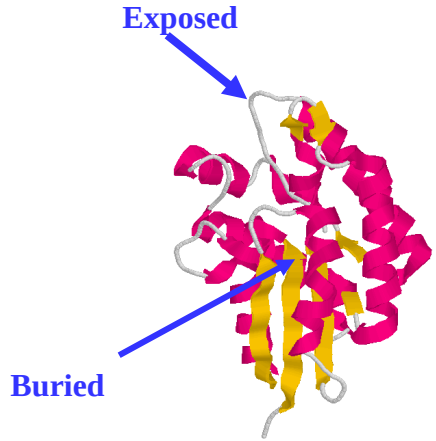
Y 00000000000000000001

3 numbers



Helix 100
Strand 010
Coil 001

2 numbers



Exposed 10
Buried 01

$25 \cdot 18 = 400$ features for a pair (i, j)