Prediction of Protein Secondary Structure using Neural Networks at Better than 70% Accuracy

Burkhard Rost and Chris Sander

By
Kalyan C. Gopavarapu

1

Presentation Outline

- Major Terminology
- Problem
- Method
- Results
- References

2

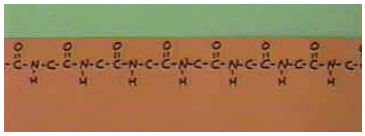
Major Terminology

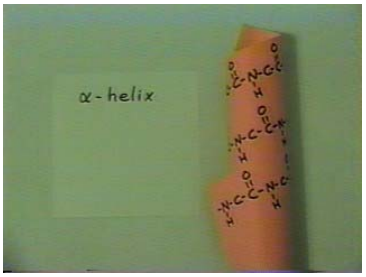
Secondary structure:

The alpha-helical or beta-sheet configuration of the polypeptide backbone of a protein.

Alpha-helix:

One of the most common types of secondary structure is referred to as an **alpha helix**. Essentially what is involved in an alpha helix is that the primary structure of the protein twists around upon itself in such a way, somewhat like rolling up this paper model, to form a recurring pattern that is very important.



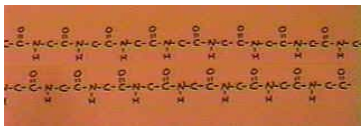


3

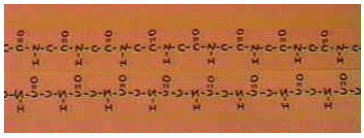
Contd…

- Beta Sheet

It is also possible for the protein's strands to line up side by side with one another and form hydrogen bonds in this way. When that happens with the strands both running in the **same direction** it is called **parallel**.



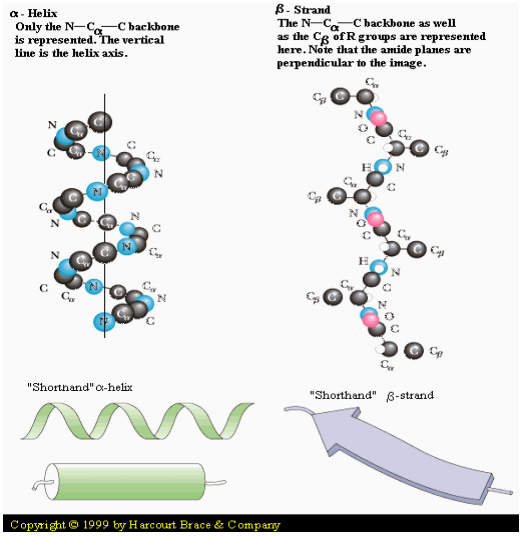
When that happens with the strands running in the **opposite direction** it's called **antiparallel**.



Parallel or antiparallel, this particular kind of secondary structure is referred to as **beta sheet** structure.

4

Alpha-Helix and Beta Strand



5

Contd…

Sequence Profiles

Sequence profiles are essentially patterns where each position in the sequence of the segment has been assigned a probability value for each possible amino-acid residue type.

Multiple Sequence Alignment

The alignments are taken from HSSP data bank

Local Interactions

The integrations between the residues within the helix of an amino acids.

Non-local Interactions

The integrations between the residues in the different helices.

Neural Network

6

Problem

To predict the secondary structure of a protein at better than 70% accuracy.

Input: K E L N D L E K K Y… (protein sequence)

Output: α α β α β L L β L α … (secondary structure)

7

Prediction Approach

- A combination of three levels of network with sequence profiles generated from multiple sequence alignments as input (instead of single sequences).
- Reliability index is used for winner-take-all decision.
- Performance accuracy of a prediction tool is verified by a sevenfold cross validation test.
- Non-local interactions of amino acids are considered.

8

Method

- 7-fold Cross-validation technique

  130 protein chains are taken from PDB

  111 protein chains are used for training

  19 for testing

- Repeated 7 times with different sets of 19 until all proteins have been used for testing exactly once.

- How can the data bank of known structures be used to estimate the performance on new proteins?
  - In training the network

9

Measures of Protein secondary structure prediction accuracy

Compute ratios that reflect the number of properly predicted residues.

Coefficients are derived from 3X3 accuracy table *A*.

A_{ij}= number of residues predicted to be in structure type j and observed to be in type i

The sums over the columns of *A* give the number of residues predicted to be in structure i:

$$a_i = \sum_{j=1}^{3} A_{ji}, \quad \text{for } i = \alpha, \beta, \text{L}.$$

The sums over the rows give number of residues observed to be in structure i.

$$b_i = \sum_{j=1}^{3} A_{ij}, \quad \text{for } i = \alpha, \beta, \text{L}.$$

10

Contd…

The sum of overall elements of A is the number of residues in the data bank used,

$$b = \sum_{j=1}^{3} b_j = \sum_{j=1}^{3} a_j.$$

For class i the percentage of residues correctly predicted to be in class i

-- relative to those observed to be in class i are given by

$$Q_i = Q_i^{\%\text{obs}} = \frac{A_{ii}}{b_i} \times 100, \quad \text{for } i = \alpha, \beta, \text{L},$$

-- from all residues predicted to be in i are given by

$$Q_i^{\%\text{pred}} = \frac{A_{ii}}{a_i} \times 100, \quad \text{for } i = \alpha, \beta, \text{L}.$$

11

Contd…

The overall three state accuracy is given by

$$Q_{\text{total}} = \frac{\sum_{i=1}^{3} A_{ii}}{b} \times 100$$

The above percentages describe performance accuracy for a prediction tool.

12

Example:

A C D E F G H I I L

Pred: L L α α α α L β β β

Obs: L α α α L L β β β L

Pred

Obs	α	β	L
α	2	0	1
β	0	2	1
L	2	1	1

Predicted to be in α and observed in L = 2

What is the accuracy if we have multiple chains?

$$\langle Q \rangle_{\text{chain}} = \frac{1}{N^{\text{chain}}} \sum_{c=1}^{N^{\text{chain}}} Q^{c}_{\text{total}},$$

N^{chain} – number of all chains in the data bank.

Q^{c}_{total} – accuracy defined by above Q_{total} for chain c.

13

Contd…

A more complicated measure of accuracy is given by correlation coefficient

$$C_i = \frac{p_i n_i - u_i o_i}{\sqrt{(p_i + u_i)(p_i + o_i)(n_i + u_i)(n_i + o_i)}},$$

for $i = \alpha, \beta, \text{L},$

p_i = number of properly predicted residues in conformation i.

n_i = number of those correctly not assigned to structure i.

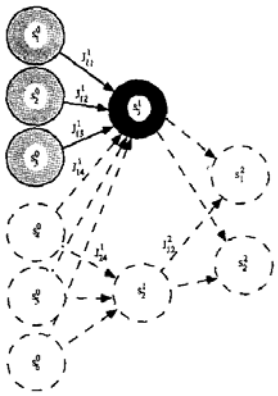
u_i = number of underestimated

o_i = number of overestimated conformations.

14

Classifications by a layered network

The simple two layered network:



Perceptron:

Here j represents the junctions or weights

the input to the first layer neurons is

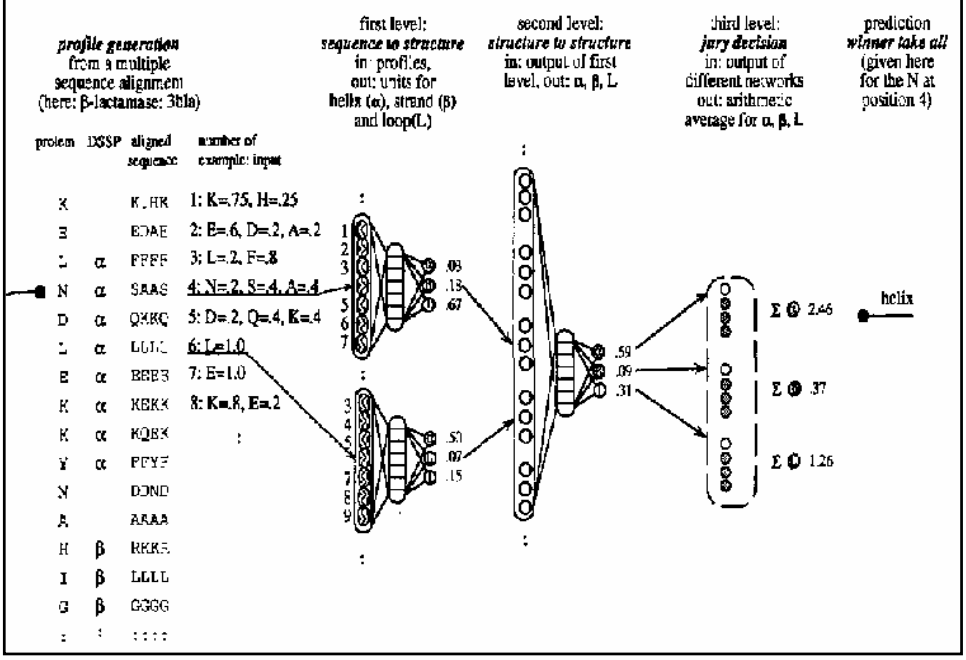
$$h_i^1 = \sum_{j=1}^{N^0+1} J_{ij} s_j^0 \quad (\text{here, } i = 1).$$

The output is computed by sigmoid trigger function:

$$s_i^1 = \frac{1}{1+\exp(-h_i^1)}.$$

15

Network System:



First Level: sequence-to-structure net

Input: Profile sequence generated from MSA taken from HSSP data bank.

Window size (w) = 7

How many basic input units are the input for a particular pattern?

The input for a particular pattern = 20w input units.

What is the input for the first residue?

Input for the particular pattern is (20+1)w input units.

Output: Secondary structure class of the central residue.

Coding concept: binary or real values.

0000 for frequency	$f < 0{\cdot}02$
0001 for	$0{\cdot}02 \leq f < 0{\cdot}33$
0011 for	$0{\cdot}33 \leq f < 0{\cdot}66$
0111 for	$0{\cdot}66 \leq f < 0{\cdot}98$
1111 for	$f \geq 0{\cdot}98.$

17

Contd…

The output unit i of the network for sample v is:

$$s_i^{2,\nu} = f\left\{\sum_{j=1}^{N^1+1} J_{ij}^2 f\left\{\sum_{k=1}^{N^0+1} J_{jk}^1 s_k^{0,\nu}\right\}\right\},$$

The sigmoid trigger function is

$$f(x) = \frac{1}{1+e^{-\beta x}}.$$

The error E for pattern v is

$$E\left(\{J^1\}, \{J^2\}\right) = \sum_{i=1}^{3} (s_i^{2,\nu} - d_i^{\nu})^2,$$

The gradient is given by

$$\Delta J(t+1) = J(t) - \varepsilon \frac{\partial E}{\partial J}(t) + \eta \Delta J(t-1),$$

18

Second level: structure-to-structure net

Input: Output of first level
Window size = 17

Output: Secondary structure class of the central residue.

Coding: same as first level.

19

Third level: jury decision

Input: Output of different network architectures.

Output: Arithmetic average of secondary structure classes (α, β, L)

$$\langle s_i \rangle_{\text{jury}} = \frac{1}{\chi} \sum_{a=1}^{\chi} s_i^a, \quad \text{for } i = \alpha, \beta, \text{L}.$$

How to generate different Architectures?

- By training
  - balanced and unbalanced fashion
  - real coding and binary coding

20

Reliability index for the prediction

The following formula separates the input vectors into secondary structure classes.

$$s_i^{2,\nu} = f\left\{\sum_{j=1}^{N^1+1} J_{ij}^2 f\left\{\sum_{k=1}^{N^0+1} J_{jk}^1 s_k^{0,\nu}\right\}\right\},$$

Finally winner-take-all decision:

the highest output value is chosen as the prediction.

Reliability index : increases the difference between the output values.

$$RI = \mathrm{INTEGER}(10 \times (out_{\max} - out_{\mathrm{next}})),$$

out_{max} – Output of the unit with highest value.

out_{next} – Output of the unit with next highest value.

21

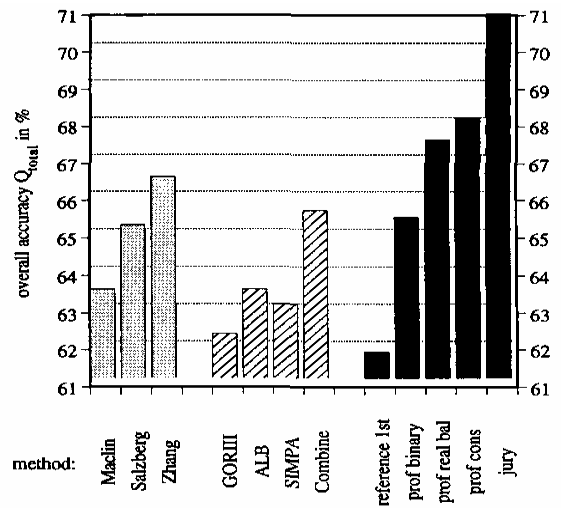
Results

- More than six percentage points gained by use of sequence profiles.
- Two percentage points increased by the jury decision.
- Reliability index helps to evaluate the prediction.
- Balanced prediction by balanced training
- Substantial improvement in predicting segment lengths.
- Secondary structure content predicted successfully.

22

Contd…

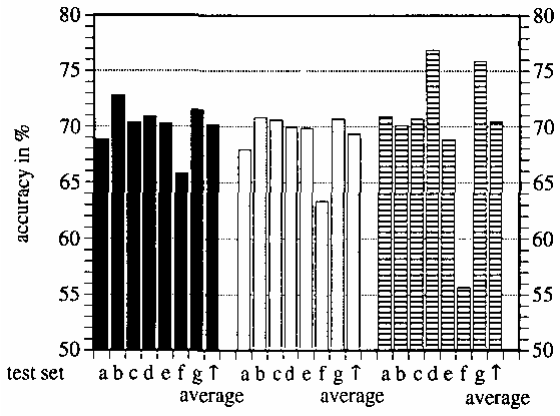
Overall accuracy of various methods



23

Results:

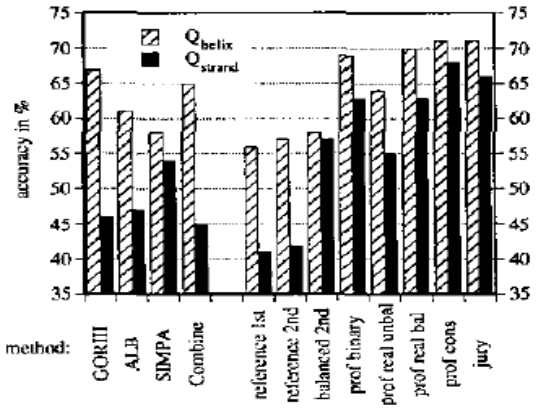
Variation of prediction accuracy with choice of test.



- Filled bars Q_{total}
- Stippled bars Q_{helix}
- Hatched bars Q_{strand}

24

Comparison of Helix/Strand accuracies for various methods.



25

References:

- Zhang, X., Mesirov, J. P. & Waltz, D. L. (1992). Hybrid system for proteim secondary structure prediction.
- Taylor, W. R. (1988). Pattern matching methods in protein sequence comparison and structure prediction.
- Maclin, R. & Shavlik, J. W. (1983). Using knowledge based neural networks to improve prediction algorithms.
- Garnier, J. (1993). Prediction of protein structure. In Biological Sequences: Finding Structure and Function by Neural Networks.
- http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Retrieve&dopt=Citation&list_uids=8345525

26

Thank you.

Questions???