Predicting Protein Secondary Structure using Artificial Neural Networks

A Short Tutorial

Sara Silva

May 2005

Contents

List of Figures

List of Tables

Foreword

This report is based on my M.Sc. thesis on predicting protein secondary structure using artificial neural networks. It was completed more than five years ago at FCUL[1], Portugal, under the supervision of J. Felix Costa and Pedro J.N. Silva. No publication ever came out of this thesis, mainly because of lack of computational power to perform the numerous additional validations that would have been needed to ensure reliable and publishable results. A few years ago, this thesis was given new life and began serving as an application example to the bioinformatics students of the FCUL-IGC post-graduate programme[2], in the 'Biologically Inspired Algorithms' course coordinated by Carlos Lourenço.

The prediction system described here was inspired in the best system available then, and the results presented appear to be slightly better. This was probably due to the general increasing quality of the publicly available data, or to the specific removal of any dubious elements from our data, resulting in high quality data sets.

However, this report is *not* about results. Those are outdated, as may be other information herein contained. This work should not be regarded as anything more than a mere example of the application of artificial neural networks to the secondary structure prediction task. This prediction system could have been developed in many different ways, of which only a few are mentioned along the text. Many details were left out, and many options left unjustified. Still, after completing this condensed translation of my thesis to serve as additional course material, I hope you will find it useful, too.

I would like to thank the Fundação para a Ciência e a Tecnologia, Portugal, for my M.Sc. grant (PRAXIS XXI BM/15046/98), as well as both my M.Sc. advisors in the University of Lisboa, especially Pedro for once again reviewing my writing. Great thanks also go to Ernesto Costa, currently my Ph.D. advisor in the University of Coimbra, for his interest and suggestions regarding this work. Last but not least, I thank Carlos Lourenço for taking my thesis out of the shelf and making it see the light of day again.

Sara Silva

May 10th, 2005

[1] http://www.fc.ul.pt/

[2] http://bioinformatics.fc.ul.pt/

Chapter 1

Introduction

The idea of using neural networks in the prediction of protein secondary structure originated on a curious episode. The NETtalk system, developed by Sejnowski and Rosenberg [1], consists of a neural network that learns how to pronounce English written text. A 7-letter window moves along the text while the network is trained to pronounce the phoneme corresponding to the central letter. Following a presentation about NETtalk, someone in the audience suggested Sejnowski that, using amino acids instead of letters, the system could learn how to predict protein secondary structure [2]. The work then published by Qian and Sejnowski [3] proved that neural networks could achieve better results than any other existing secondary structure prediction method.

A long series of similar prediction methods followed, leading to the PHD system [4], the first method to reach the 70% accuracy barrier, and certainly the most successful for many years afterwards. Our prediction system was inspired in PHD.

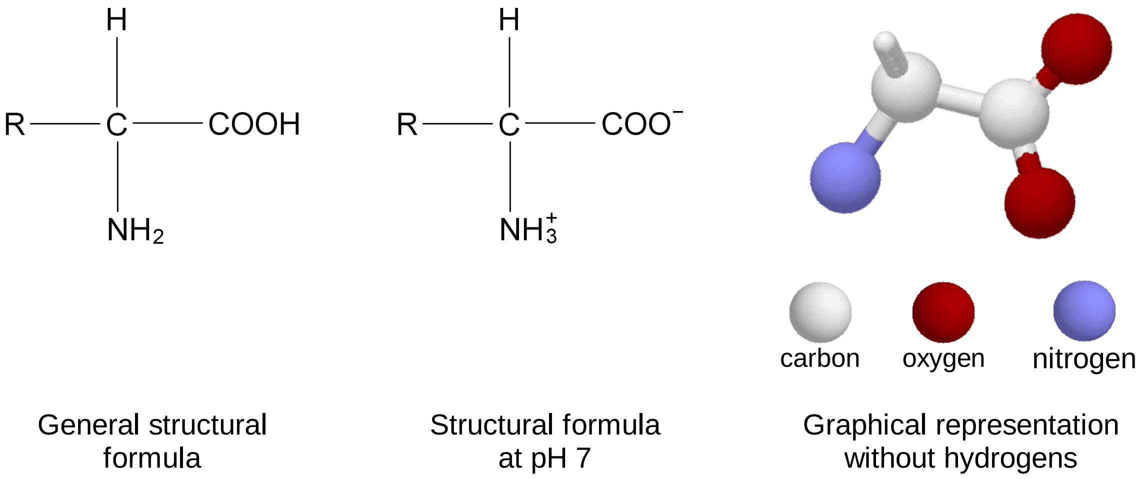
The next chapter of this report describes proteins: the constitution of the several levels of their structure, the structural classes in which they can be divided, and some notions about homology and alignments. Chapter 3 addresses the reasons for predicting secondary structure, and briefly describes some important prediction methods, including the PHD. Chapter 4 explains the workings of our system: how data was obtained and pre-processed, how the different components of the system interact with each other, and how to interpret the output provided. Chapter 5 describes how the results can be presented, and how to make them more informative and reliable. Finally, Chapter 6 contains some considerations regarding the difficulties of predicting protein secondary structure.

Chapter 2

Proteins

Proteins are macromolecules coded by DNA, such as enzymes, antibodies, and many hormones. They are built from several molecules called *amino acids*, connected to each other in a linear sequence called *polypeptide chain*. Each protein is made from one or more of these chains.

Amino acids are formed by a central carbon bonded to an hydrogen, a *carboxyl group* (COOH), an *amino group* (NH_2), and a *side chain* whose structure determines the distinctive physical and chemical properties of each amino acid. Figure 2.1 shows a generic amino acid, and table 2.1 lists the names of the 20 standard amino acids found in proteins, along with the three-letter and one-letter symbols used to designate them.



General structural formula

Structural formula at pH 7

Graphical representation without hydrogens

Figure 2.1: Generic amino acid.

Table 2.1: The 20 standard amino acids found in proteins.

Name	3-letter symbol	1-letter symbol
Alanine	Ala	A
Cysteine	Cys	C
Aspartic acid	Asp	D
Glutamic acid	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	Thr	T
Valine	Val	V
Tryptophan	Trp	W
Tyrosine	Tyr	Y

2.1 Structure

When two amino acids connect to each other, they release a water molecule, forming a *peptide bond*. What is left of each amino acid is then called *residue*, although both terms are used interchangeably. Each polypeptide chain may contain from a few dozen to several hundred residues. The *primary structure* of a protein is the sequence of residues of its polypeptide chain(s). Figure 2.2 shows the primary structure of a protein designated as 'protein G'. The highlighted residues constitute what is called the 'B1 domain' (see figures 2.5 and 2.6 for other structural levels).

```
              10         20         30         40         50
  1 MEKEKKVKYF LRKSAFGLAS VSAAFLVGST VFAVDSPIED TPIIRNGGEL
 51 TNLLGNSETT LALRNEESAT ADLTAAAVAD TVAAAAAENA GAAAWEAAAA
101 ADALAKAKAD ALKEFNKYGV SDYYKNLINN AKTVEGIKDL QAQVVESAKK
151 ARISEATDGL SDFLKSQTPA EDTVKSIELA EAKVLANREL DKYGVSDYHK
201 NLINNAKTVE GVKELIDEIL AALPKTDTYK LILNGKTLKG ETTTEAVDAA
251 TAEKVFKQYA NDNGVDGEWT YDDATKTFTV TEKPEVIDAS ELTPAVTTYK
301 LVINGKTLKG ETTTKAVDAE TAEKAFKQYA NDNGVDGVWT YDDATKTFTV
351 TEMVTEVPGD APTEPEKPEA SIPLVPLTPA TPIAKDDAKK DDTKKEDAKK
401 PEAKKDDAKK AETLPTTGEG SNPFFTAAAL AVMAGAGALA VASKRKED
```

Figure 2.2: Primary structure of 'protein G', with the residues of the 'B1 domain' highlighted.

Polypeptide chains are not unidirectional structures. Several chemical interactions among residues and between residues and the solvent fold the chain in different directions. The hydrophobic properties of some amino acids also play a part in its final conformation. Several motifs repeatedly occur along the polypeptide chain(s). Their identification is called the *secondary structure* of the protein.

Two of the most common structural motifs are the *α-helix* and the *β-sheet*. In a α-helix the backbone of the polypeptide chain forms a helicoidal structure with 3.6 residues per turn, stabilized by hydrogen bonds between every 4 residues, with all side chains turned outwards. Figure 2.3 shows three different representations of a α-helix[1]. Other types of helix occur less frequently, like the *π-helix* and the *3_{10}-helix*.

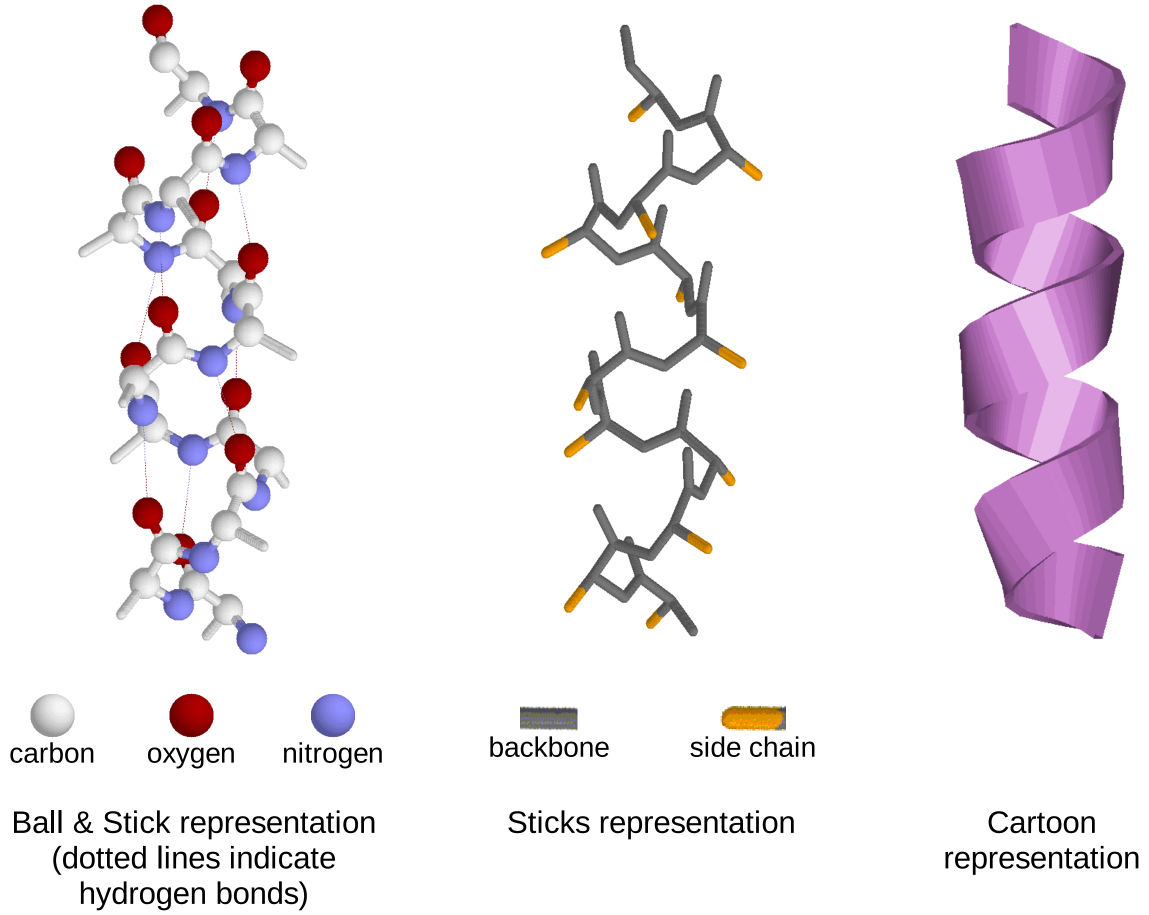
In a β-sheet different segments of the polypeptide chain, or even of different chains, are connected by hydrogen bonds between all the residues, forming a planar structure with the side chains turned upwards and downwards, never interacting with each other. β-sheets can be *parallel* (all the segments oriented in the same direction), *antiparallel* (adjacent segments oriented in opposite directions), or *mixed*. Figure 2.4 shows two different representations of a mixed β-sheet.

The *tertiary structure* of a protein is the arrangement of all its atoms in tridimensional space. Figure 2.5 shows the tertiary structure of the B1 domain of protein G (see figure 2.2 for primary structure), in a stereoscopic pair[2]. It is often useful to visualize both secondary and tertiary structures at the same time, in which case the representation of the side chains is omitted and the most common structural motifs are represented pictorially (see figures 2.3 and 2.4). Figure 2.6 shows both secondary and tertiary structures of the B1 domain of protein G.

Quaternary structure exists only when the protein is formed by more than one polypeptide chain, and it describes their relative positions and interactions. Figure 2.7 (page 11) shows the quaternary structure of a human hemoglobin, with each of its four chains represented pictorially in a different color.

[1] These illustrations were produced by one of the molecular visualization programs available at `http://www.umass.edu/microbio/rasmol/`

[2] Stereoscopic pairs should be visualized using the cross-viewing method.

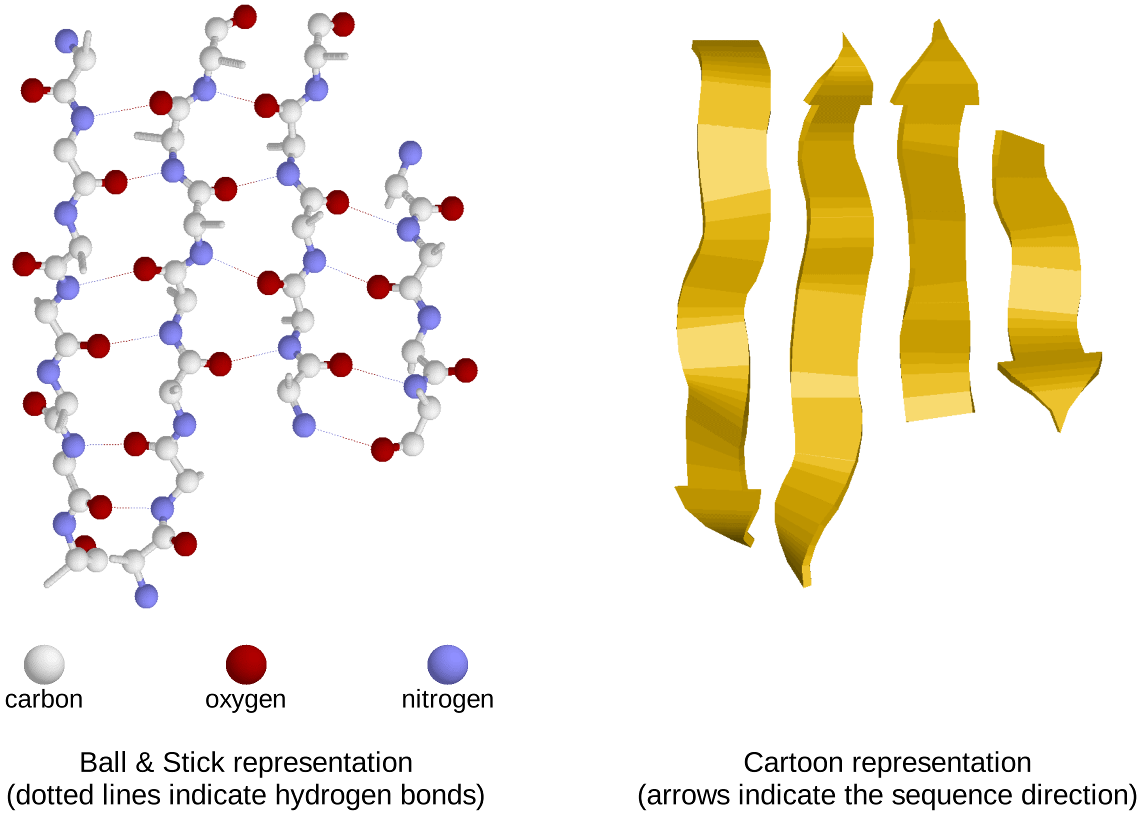


Ball & Stick representation (dotted lines indicate hydrogen bonds)

Sticks representation

Cartoon representation

Figure 2.3: Different representations of a α-helix.



Ball & Stick representation (dotted lines indicate hydrogen bonds)

Cartoon representation (arrows indicate the sequence direction)

Figure 2.4: Different representations of a β-sheet.

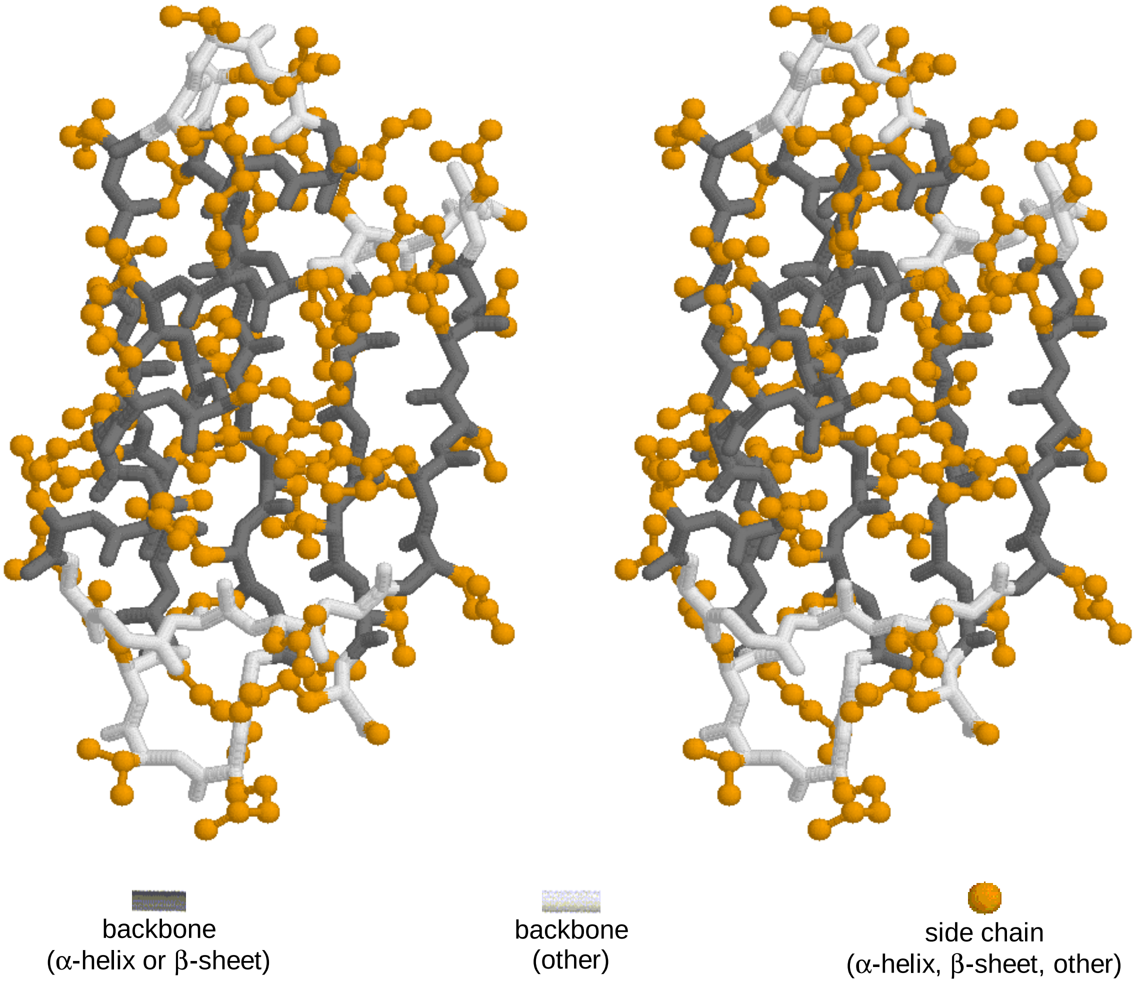


Figure 2.5: Tertiary structure of ‘protein G’, ‘B1 domain’, in stereoscopy.

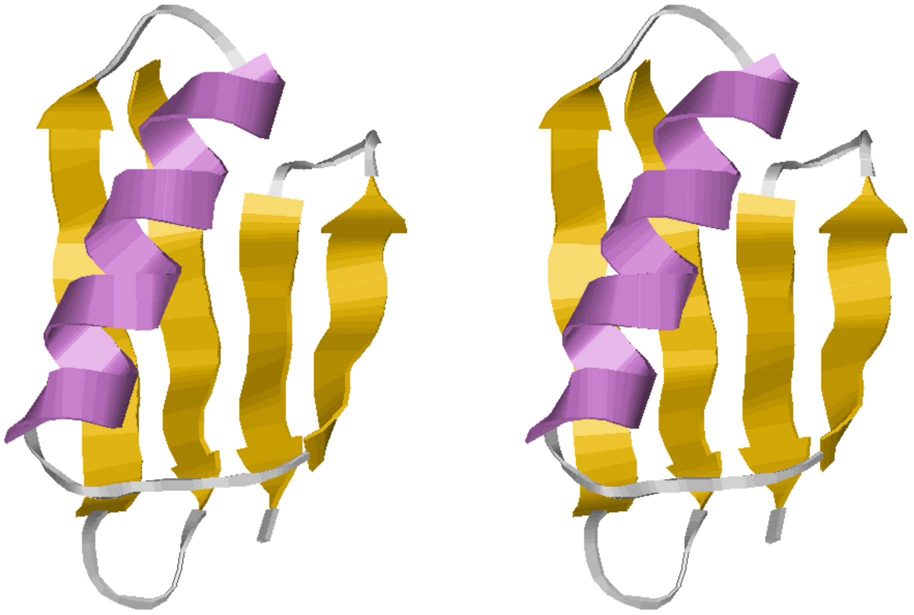


Figure 2.6: Secondary and tertiary structure of ‘protein G’, ‘B1 domain’, in stereoscopy.

2.2 Structural class

Depending on their spatial structure, proteins can be grouped in one of four classes, represented by α/α, β/β, α/β, and $\alpha + \beta$. α/α proteins are formed almost exclusively by α-helices, with any β-sheets located in the periphery. Human hemoglobin (figure 2.7, page 11) is an example of a α/α protein. β/β proteins are formed almost exclusively by β-sheets, mostly antiparallel, with any α-helices located in the periphery. Figure 2.8 (left) represents a β/β protein. In proteins belonging to the α/β structural class, α-helices and β-sheets alternate in such a way that the sheets (typically parallel) form a central agglomerate surrounded by helices. Figure 2.8 (right) represents a α/β protein. The $\alpha + \beta$ class includes the proteins that are not dominated by any of the secondary structure motifs, nor show the alternateness typical of the α/β class. The B1 domain of protein G (figure 2.6, page 9) belongs to the $\alpha + \beta$ class.

Different domains of the same protein frequently belong to different structural classes. Some proteins cannot be classified in any class, either because its sequence is too short or because it shows practically no secondary structure motifs. Knowing the structural class to which a protein belongs may be helpful in predicting its entire secondary structure because it allows the usage of methods specialized in each class [5]. However, achieving a reliable prediction of the structural class based on the primary structure alone may prove too hard to be advantageous [4].

2.3 Homology

When genes suffer mutations, the proteins they code may be affected, most commonly by substitutions, insertions, and deletions of single amino acids of the sequence. Some proteins contain a group of amino acids essential to its structure and function, called *functional center* (or *active site*, in enzymes). When a mutation affects the functional center of a protein, it usually impairs or even disables its function, so the mutation is rapidly lost. On the other hand, substitutions between similar amino acids rarely affect the conformation of the protein, and are very common. Conformation is generally more important than sequence, therefore more evolutionarily conserved.

Two proteins are said to be *homologous* when they share a common ancestor. In practical terms, homology is considered when the primary sequences are at least $n\%$ identical, with n usually 20, 25, or 30. It is indeed highly improbable for two sequences evolving independently to arrive at such high similarity.

An *alignment* of proteins is an arrangement of their sequences such that the aligned residues correspond to the same residue in a common ancestor. Although an alignment may use only two sequences, a multiple alignment is more reliable from the biological point of view, and may contain more than evolutionary information. Namely, it can reveal the functional centers of homologous proteins, identified by one or more groups of extremely conserved consecutive residues.

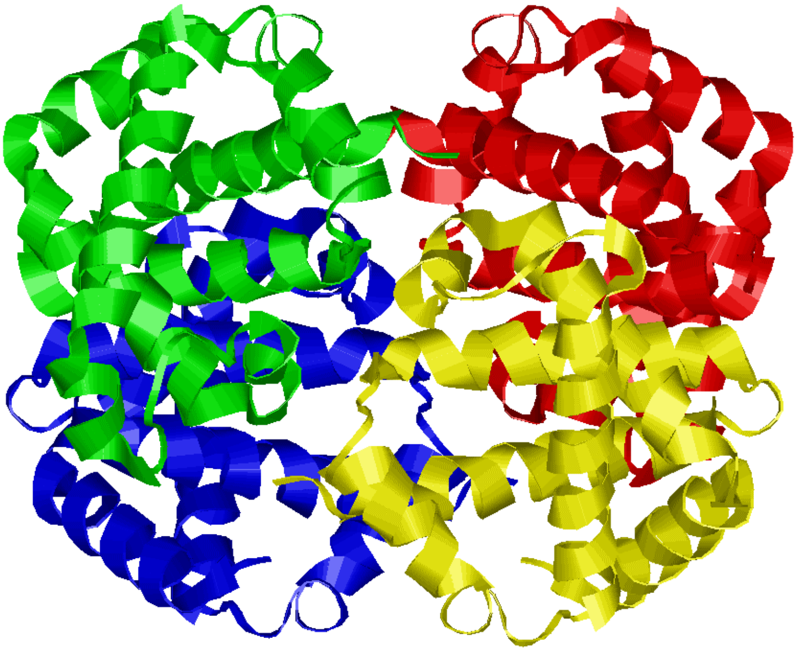


Figure 2.7: Quaternary structure of a human hemoglobin.

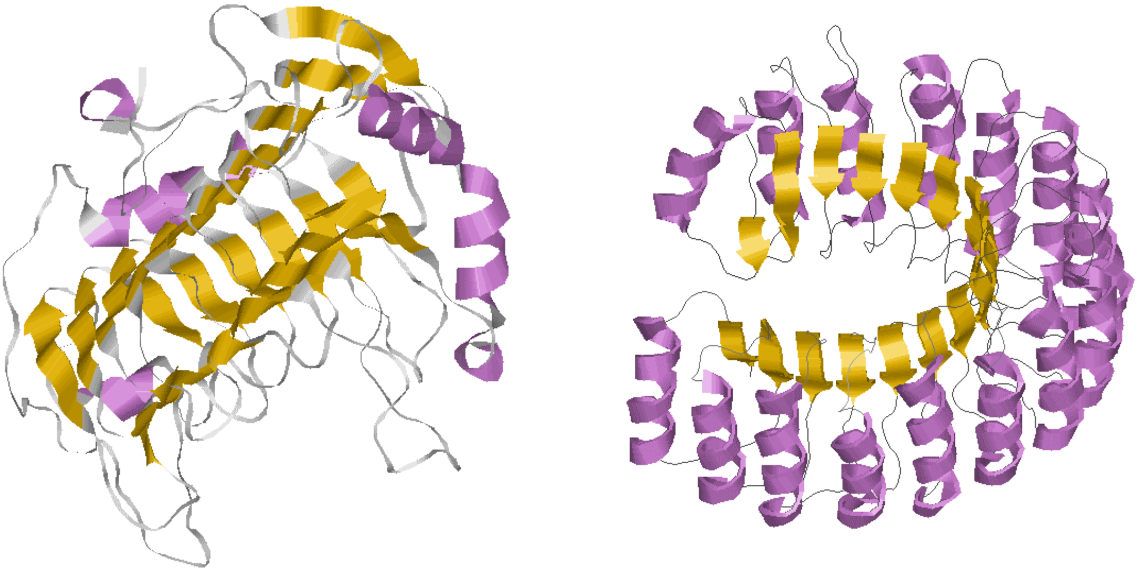


Figure 2.8: β/β protein (left) and α/β protein (right).

Chapter 3

Predicting secondary structure

Experimental methods, like X-ray crystallography and multidimensional Nuclear Magnetic Resonance (NMR) spectroscopy can be used to determine protein structure. However, for more than a decade they have not been able to keep pace with the rapidly growing number of known sequences, and the last few years have dramatically increased that difference. Figure 3.1 plots the growth of the number of annotated sequences available in Swiss-Prot[1] and the number of structures available in PDB (Protein Data Bank)[2], between 1986 and now. On May 10th of 2005, the number of structures was 31237 and the number of sequences was 168297, plus more than 1.5 million (precisely 1589670) sequences stored in the auxiliary TrEMBL database, awaiting annotation before being also admitted in Swiss-Prot.

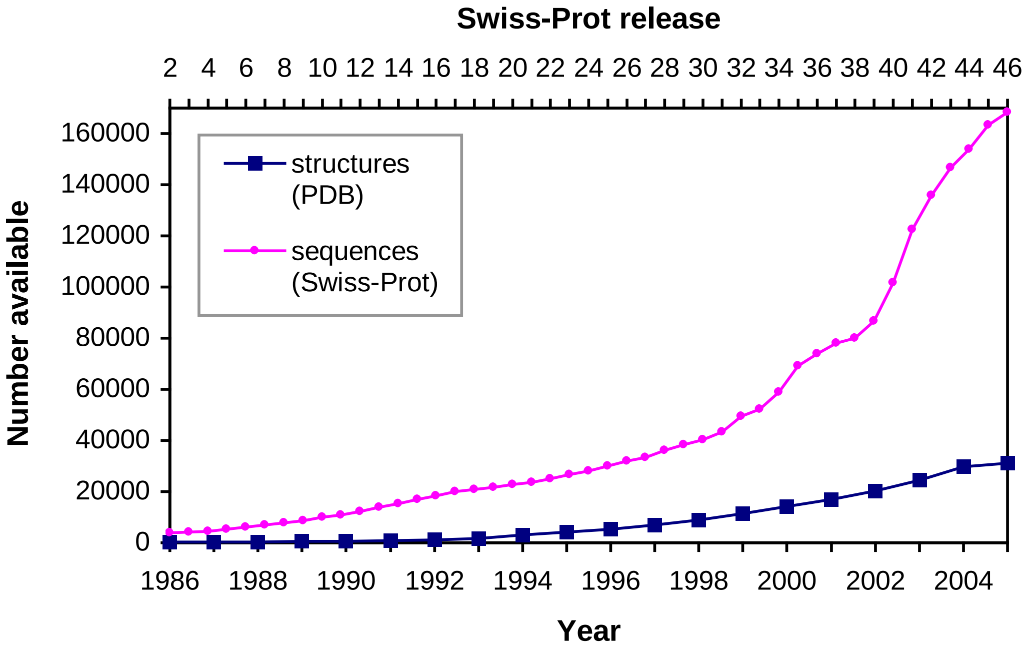


Figure 3.1: Growth of the number of sequences and structures available.

Given the difficulty in experimentally determining protein structure, several attempts have been made to *predict* it, based on the fundamental assumption that sequence determines conformation. Many of the methods developed are aimed at what seems to be an easier task: predicting secondary structure[3] (see [6, sect. 6.3]). The importance of this strategy should not be underestimated. Besides providing invaluable information about a protein, a sufficiently reliable secondary structure prediction may be used as a stepping stone for attempting the far more difficult task of predicting its complete tridimensional conformation.

[1] `http://www.expasy.org/sprot/`

[2] `http://www.rcsb.org/pdb/`

[3] For a collection of protein secondary structure analysis and information sites, see `http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-2-struct.html`

Table 3.1: Three generations of secondary structure prediction methods.

Name	Year	Type	Accuracy
Chou-Fasman	1974	1st generation: ☐ local interactions ☐ alignments	50%
GOR III	1987	2nd generation: ☒ local interactions ☐ alignments	60%
PHD	1993	3rd generation: ☒ local interactions ☒ alignments	70%

Table 3.1 summarizes three different secondary structure prediction methods: Chou-Fasman [7, 8], GOR III (Garnier-Osguthorpe-Robson) [9], and PHD (Profile network from HeiDelberg) [4, 10]. These methods are important because they have introduced what is usually considered to be three different *generations* of secondary structure prediction methods. Each generation is characterized by the usage - or not - of information regarding local interactions between amino acids, and alignments (see section 2.3).

3.1 Chou-Fasman

The Chou-Fasman method was developed in 1974, and the first one to be widely used in the prediction of protein secondary structure. Based solely on the probability of each amino acid to belong to a helix or sheet, its accuracy did not reach more than 50% of correctly classified residues.

3.2 GOR III

GOR III, developed in 1987 as an improvement to GOR [11], was the first method to use information about local interactions between amino acids. This means that, to predict the secondary structure of a given amino acid, GOR III also uses the information about which amino acids are following and preceding it in the sequence. This method reports an accuracy of 60%.

3.3 PHD

Possibly still the best and most used secondary structure prediction method, PHD was the first to use the evolutionary information contained in alignments. When receiving a sequence to classify, the first priority of PHD is to obtain a multiple alignment from homologous sequences stored in Swiss-Prot, a task performed by the auxiliary program MaxHom [12].

Based on artificial neural networks (ANNs), PHD contains four processing levels, the first two being multilayer perceptrons previously trained with proteins of known structure. The first level receives vectors concerning sequences of 13 consecutive amino acids in the alignment, and returns the likelihood of the central residue being in a helix, sheet, or other motif. The second level receives the values from the first level, along with some global information about the protein (like its length), and returns new likelihood values with the same meaning as the ones returned by the first level. The highest value determines the classification of the central residue, and the difference between the two highest values is used as a reliability index indicating the confidence the program has in the prediction made. Several neural networks, trained independently from each other, perform the classification of all the residues of the sequence, and the third computational level of PHD consists in choosing the classifications with the highest reliability indices. The fourth and last computational level consists in submitting the classification obtained to a filter that removes obvious mistakes (like helices less than three residues long, an impossible occurrence).

The PHD method was the first to surpass the 70% accuracy barrier. If only the half of residues with higher reliability index are considered, this accuracy rises above 80%. PHD can be freely used online through the server PredictProtein[4] [13].

[4] http://www.embl-heidelberg.de/predictprotein/

Chapter 4

ANN-based prediction system

Although inspired in PHD, the prediction system described here is considerably simpler, containing only two processing levels, both implemented as multilayer perceptrons. Figure 4.1 shows a diagram of this system.

4.1 Outline

As in the PHD system, inputs to the first processing level consist of windows of 13 consecutive positions[1] in the alignment, and the values returned represent the likelihood of the central position being in a helix, sheet, or other motif. The second level receives these likelihood values and, using 17 × 3-value wide windows, returns updated likelihood values for the central position. Although implemented in a different manner, this second processing level is similar to the forth processing level of PHD, filtering out the most obvious prediction errors. The prediction for each residue is the motif with highest likelihood value, accompanied by a reliability index somewhat less optimistic than the PHD index.

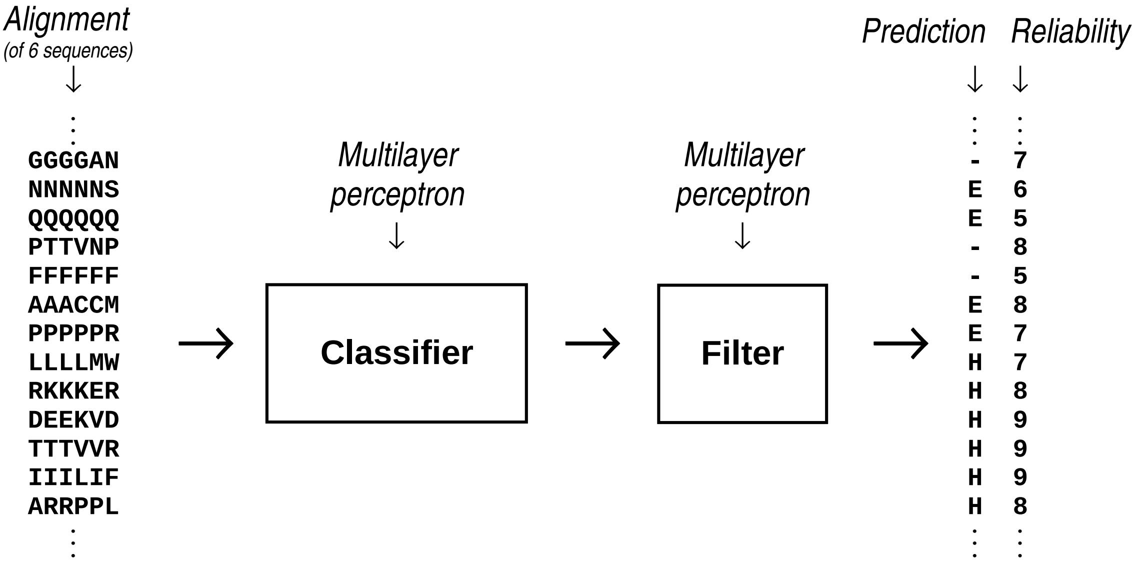


Figure 4.1: ANN-based prediction system. H → helix, E → sheet, – → other motifs.

[1] Any other odd number is acceptable. The higher the number, the larger amount of information is available to the neural network, but the longer computational time is needed for learning. Some authors have used larger windows, of size 17 [14] or even 51 [15].

4.2 Data

All the data used to train our prediction system was obtained in the HSSP (Homology-derived Secondary Structure of Proteins) [12] public database[2]. Frequently updated, on November 23rd of 2004 this database contained 26213 files referring to proteins whose structure is available in PDB and for which the secondary structure was determined using the DSSP (Database of Secondary Structure in Proteins) [16] program[3]. The names of the HSSP files are identical to the names of the corresponding PDB files, and their internal format is exemplified in figure 4.2, abbreviated for better visualization.

HSSP files

The file header contains information about the protein, like its origin (not shown in the figure) and identification, sequence length, number of chains, and number of sequences used in the alignment. The file shown in figure 4.2 concerns a protein identified as *2fiv* in PDB. It is made of four chains, of which only two are used in this file, with total length 118. As much as 16 sequences are used in the alignment.

Next follows a section initiated by the string "`## ALIGNMENTS`", containing the sequence, secondary structure, and alignments. The position of each residue in the sequence is identified by a number in columns 2-6; columns 7-11 identify the corresponding positions in the PDB file.

The residues of the sequence are then identified by their one-letter symbols (see table 2.1, page 6) in column 15, preceded by an identifier of the chain to which they belong, in column 13. When there are doubts concerning the true identity of the residues, symbols different from the ones presented in the table are used. The symbol "`!`" generally indicates the end of a chain and beginning of another, but it may also indicate a gap inside the same chain, caused by an error or omission in the corresponding PDB file.

The secondary structure of the protein, as determined by the program DSSP, is then specified in column 18. Seven symbols are used to identify different motifs (only some appear in the figure), listed in table 4.1. The isolated β-sheet is a β-sheet only one residue long, therefore usually not considered to be a regular β-sheet; the loop with hydrogen bond usually consists of a fraction of a 3_{10}-helix or a π-helix, too small to be considered a true helix; the absence of any symbol indicates that the residue is not in any recognizable structural motif, nor in a zone of sufficient curvature to be considered a loop. When there is superposition of motifs, priority is given by the order of appearance in the table.

Table 4.1: Symbols used to identify secondary structure motifs.

Symbol	**Motif**
H	α-helix
G	3_{10}-helix
I	π-helix
E	β-sheet
B	isolated β-sheet
S	loop
T	loop with hydrogen bond
–	other

After additional information regarding structure and alignment, all the sequences participating in the alignment (including the sequence targeted by this file) are then presented, from column 52 onwards. Dots indicate deletions; pairs of lowercase symbols indicate insertions that took place between the two residues (not shown in the figure).

A new section follows in the HSSP file, identified by the string "`## SEQUENCE PROFILE AND ENTROPY`". It contains a matrix where each row indicates the percentages of each residue in that position in the sequence, calculated from the alignment. Because of the possible existence of unidentified residues, some rows may have all percentages null. This section also contains additional information, like the number of insertions and deletions that occurred in each position of the sequence.

[2] `http://www.sander.ebi.ac.uk/hssp/`

[3] `http://www.sander.ebi.ac.uk/dssp/`

```
HSSP       HOMOLOGY DERIVED SECONDARY STRUCTURE OF PROTEINS , VERSION 1.0 1991
PDBID      2fiv
DATE       file generated on 14-Aug-98
SEQBASE    RELEASE 36.0 OF EMBL/SWISS-PROT WITH  74019 SEQUENCES
...
SEQLENGTH   118
NCHAIN        4 chain(s) in 2fiv data set
KCHAIN        2 chain(s) used here ; chain(s) :  A,I
NALIGN       16
...
## ALIGNMENTS    1 -   16
 SeqNo  PDBNo AA STRUCTURE BP1 BP2  ACC NOCC  VAR
....:....1....:....2....:....3....:....4....:....5....:....6....:....7
     1    4 A V              0   0  117    7    4  VVI       VVV
     2    5 A G        +     0   0   75    7    0  GGG       GGG
     3    6 A T        -     0   0   32    7   46  TTT       VVV
     4    7 A T  E     -A  226   0A  79   10   29  TTT     E TTTE E
     5    8 A T  E     -A  225   0A  25   10   53  TTT     Y YYYL L
...
    96   99 A Q  S    S-     0   0   27    4    0  QQQ.............
    97  100 A P        -     0   0   18   11   39  PPPEEEEEDE......
    98  101 A L  E     -fH  25  34B   0   17   17  LLLVVVVVIVIIIIII
    99  102 A L  E     -f   26   0B   0   17   12  LLLLLLLLILLLLIII
   100  103 A G    >>  -     0   0    0   17    0  GGGGGGGGGGGGGGGG
...
   113  116 A M              0   0   17   17    7  MMMMMMMMLMLLLMFM
   114         !  !          0   0    0    0    0
   115  202 I X              0   0   51    0    0
   116  203 I V  E     -KO  30 238C    0    1    0
   117  204 I X  E     - O   0 237C    0    0    0
...
## SEQUENCE PROFILE AND ENTROPY
 SeqNo PDBNo   V   L   I   M   F  ...  A   P   D  NOCC NDEL NINS  ... WEIGHT
    1    4 A  86   0  14   0   0  ...  0   0   0     7    0    0  ...  1.46
    2    5 A   0   0   0   0   0  ...  0   0   0     7    0    0  ...  1.54
    3    6 A  43   0   0   0   0  ...  0   0   0     7    0    0  ...  0.66
    4    7 A   0   0   0   0   0  ... 30   0   0    10    0    0  ...  1.44
    5    8 A   0  20   0   0   0  ...  0   0   0    10    0    0  ...  0.67
...
   96   99 A   0   0   0   0   0  ...  0   0   0     4   13    0  ...  0.67
   97  100 A   0   0   0   0   0  ...  5   0   9    11    6    0  ...  0.71
   98  101 A  35  24  41   0   0  ...  0   0   0    17    0    0  ...  1.11
   99  102 A   0  76  24   0   0  ...  0   0   0    17    0    0  ...  1.36
  100  103 A   0   0   0   0   0  ...  0   0   0    17    0    0  ...  1.57
...
  113  116 A   0  24   0  71   6  ...  0   0   0    17    0    0  ...  1.41
  114           0   0   0   0   0  ...  0   0   0     0    0    0  ...  1.00
  115  202 I   0   0   0   0   0  ...  0   0   0     0    0    0  ...  1.00
  116  203 I 100   0   0   0   0  ...  0   0   0     1    0    0  ...  1.00
  117  204 I   0   0   0   0   0  ...  0   0   0     0    0    0  ...  1.00
...
```

Figure 4.2: HSSP file (partial view).

From sequence profiles to input vectors

The input vectors to our prediction system are based on the matrix of sequence profiles of the HSSP files[4]. A sliding window of odd size n transforms each segment of n consecutive residues into a vector of $n \times 20$ elements, containing the n matrix rows one after another. Every time the window moves, a new input vector is generated. Although both PHD and our system use windows of size 13, figure 4.3 illustrates this process with a window of size 3, for easy visualization. Because each vector refers only to the central residue of the window, the first $(n-1)/2$ vectors - referring to the first $(n-1)/2$ residues of the sequence - contain a high amount of null values that correspond to the window areas outside of the sequence (red in the figure). In the PHD system, vectors have $n \times 21$ elements, where the 21st element indicates the areas outside the sequence.

Normalizing the input vectors is a normal practice that usually improves the learning ability of the multilayer perceptron. In our prediction system, normalization is a two-step process. Although each input vector has $n \times 20$ elements, it can also be treated as n vectors of 20 elements each. As a first step, each of these vectors is normalized independently, and only on a second step the entire vector is normalized as a whole. This ensures that each residue has the same weight as the others in the final vector. Figure 4.4 illustrates this process.

From structural motifs to output vectors

The output vectors used to train our prediction system are obtained from the seven-symbol classification found in the HSSP files (see table 4.1, page 15). In secondary structure prediction it is common practice to reduce the number of motifs to only three: helix, sheet, and other. The three types of helix are reduced to simply *helix*; the β-sheet becomes simply *sheet*; the remaining motifs become *other*[5]. So the seven symbols are reduced to only two: H (helix) and E (sheet), with the absence of symbol indicating other. Each of these three cases is identified by a different binary[6] output vector, as specified in table 4.2.

Table 4.2: From structural motifs to output vectors.

Structural motif in HSSP files	**Structural motif in prediction system**	**Output vector**
H, G, I	helix (H)	`[1,0,0]`
E, B	sheet (E)	`[0,1,0]`
S, T, –	other (–)	`[0,0,1]`

Data quality and usage

Several prediction systems contemporary with PHD claimed to achieve accuracy rates much higher than PHD, in spite of it being considered the best secondary structure prediction tool ever developed. The contradiction arises from the fact that many authors test their systems in proteins homologous to the ones used for training. Because homologous proteins have highly similar sequences, the generalization ability of the system appears to be (misleadingly) high.

Our system was trained and tested with two non-homologous protein data sets. One of them was obtained from the PDB_SELECT[7] [19, 20], a regularly updated list of polypeptide chains available in PDB, with less than 25% similarity between them. On November 23rd of 2004 PDB_SELECT listed 2485 chains. The other data set contained most of the 240 chains used by Michie *et al.* [21].

In spite of all the care taken, errors and omissions are usually found in PDB files, inevitably propagated to HSSP files and others. New releases of these databases usually contain many new entries, along with some entries that substitute previous faulty or incomplete ones. To train and test our prediction system, we have discarded all the chains containing discontinuities or omissions, and in some cases also the ones with few sequences in the alignment[8], thus obtaining relatively

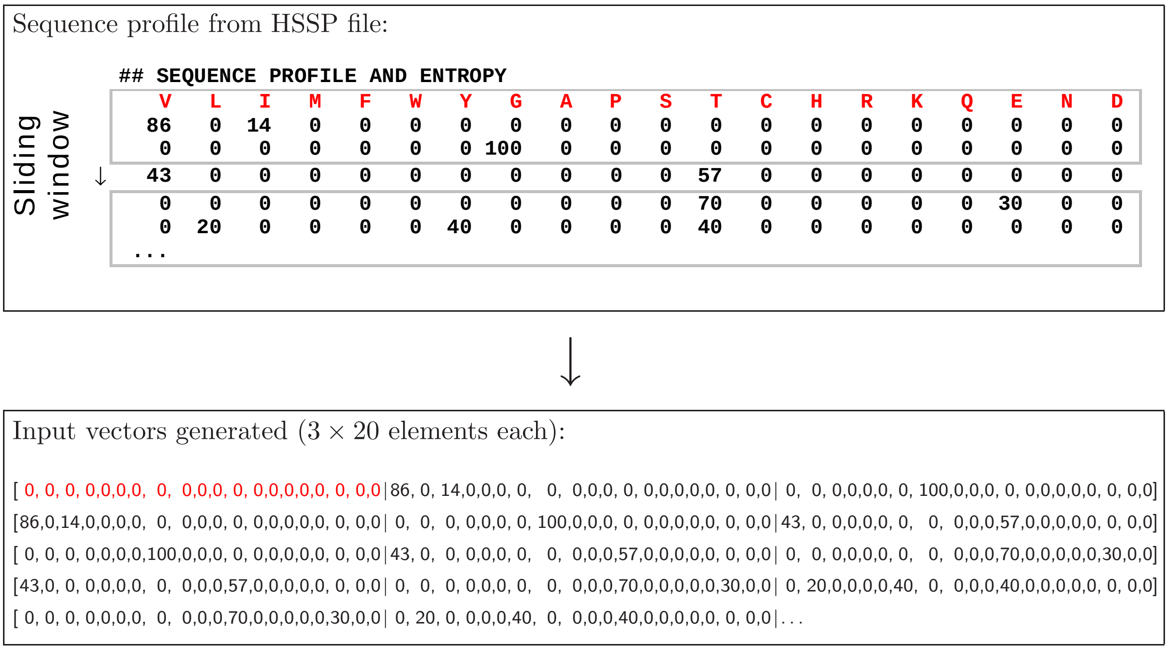
[4] In this context, whenever we mention a *residue*, we are no longer referring to a residue in a single sequence, but to a position in the alignment of several sequences.

[5] Some authors classify the isolated β-sheet as *sheet* [4] while others regard it as *other* [17]. We have chosen the second option.

[6] Bipolar vectors could be used instead.

[7] `http://homepages.fh-giessen.de/~hg12640/pdbselect/`

[8] The prediction accuracy seems to be dependent on the number of sequences used in the alignment [4].



Sequence profile from HSSP file:

```
## SEQUENCE PROFILE AND ENTROPY
      V    L    I    M    F    W    Y    G    A    P    S    T    C    H    R    K    Q    E    N    D
     86    0   14    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
      0    0    0    0    0    0    0  100    0    0    0    0    0    0    0    0    0    0    0    0
     43    0    0    0    0    0    0    0    0    0    0   57    0    0    0    0    0    0    0    0
      0    0    0    0    0    0    0    0    0    0    0   70    0    0    0    0    0   30    0    0
      0   20    0    0    0    0   40    0    0    0    0   40    0    0    0    0    0    0    0    0
    ...
```

Input vectors generated (3 × 20 elements each):

```
[ 0, 0, 0, 0,0,0,0,  0,  0,0,0, 0, 0,0,0,0,0, 0, 0,0|86, 0, 14,0,0,0, 0,   0,  0,0,0, 0, 0,0,0,0,0, 0, 0,0| 0,  0, 0,0,0,0, 0, 100,0,0,0, 0, 0,0,0,0,0, 0, 0,0]
[86,0,14,0,0,0,0,  0,  0,0,0, 0, 0,0,0,0,0, 0, 0,0| 0,  0,  0, 0,0,0, 0, 100,0,0,0, 0, 0,0,0,0,0, 0, 0,0|43, 0, 0,0,0,0, 0,   0,  0,0,0,57,0,0,0,0,0, 0, 0,0]
[ 0, 0, 0, 0,0,0,0,100,0,0,0, 0, 0,0,0,0,0, 0, 0,0|43, 0,  0, 0,0,0, 0,   0,  0,0,0,57,0,0,0,0,0, 0, 0,0| 0,  0, 0,0,0,0, 0,   0,  0,0,0,70,0,0,0,0,0,30,0,0]
[43,0, 0, 0,0,0,0,  0,  0,0,0,57,0,0,0,0,0, 0, 0,0| 0,  0,  0, 0,0,0, 0,   0,  0,0,0,70,0,0,0,0,0,30,0,0| 0, 20,0,0,0,0,40,  0,  0,0,0,40,0,0,0,0,0, 0, 0,0]
[ 0, 0, 0, 0,0,0,0,  0,  0,0,0,70,0,0,0,0,0,30,0,0| 0, 20, 0, 0,0,0,40,  0,  0,0,0,40,0,0,0,0,0, 0, 0,0|. . .
```

Figure 4.3: From sequence profiles to input vectors. Red indicates areas outside of the sequence.

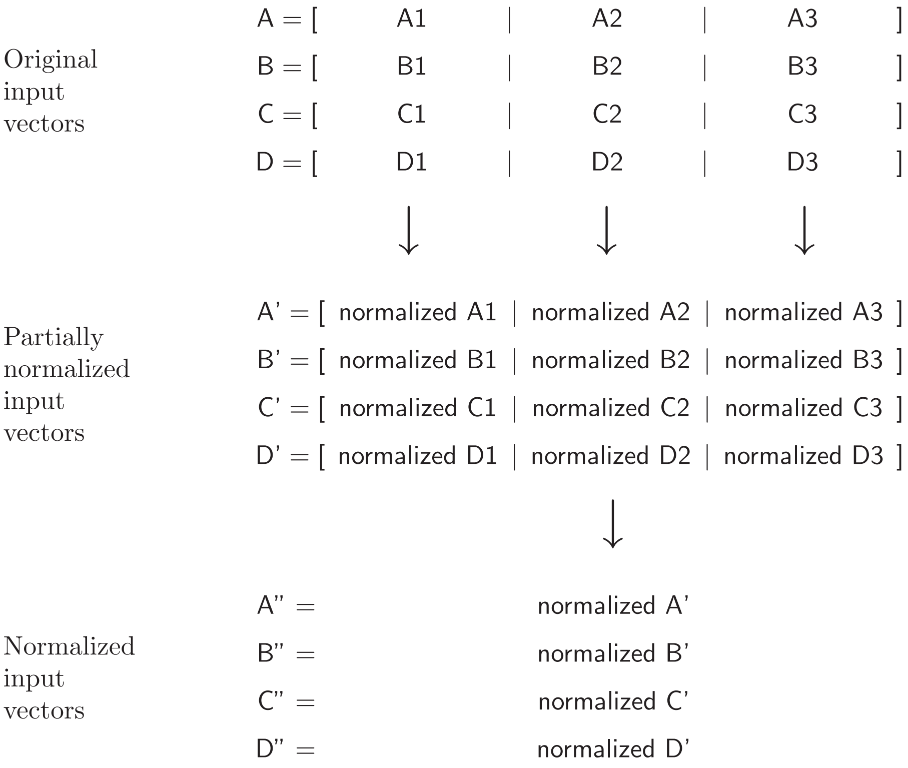


Figure 4.4: Two-step normalization of the input vectors.

high quality data sets. In most of the experiments performed, the data set used was divided in three parts: 10% for validation, 20% of the remaining for testing, and the remaining for training. In some cases with fewer data the division was only in two parts: 80% for training and 20% for testing.

4.3 Classifier

The *classifier* - the first processing level of our prediction system - is a multilayer perceptron with $13 \times 20 = 260$ input neurons, 35 hidden neurons, and 3 output neurons. During training, it receives the input vectors along with the expected output vectors, described in section 4.2. When making predictions, it returns output vectors representing the likelihood of each residue being in a helix, sheet, or other motif. Figure 4.5 illustrates a classifier receiving several input vectors and returning the predicted output vectors, comparing it with what could be the correct (expected) classification.

The classifier alone is able to perform a valid prediction, by choosing the motif with higher likelihood value. Figure 4.7 (top) shows what could be the prediction made by this classifier in a short sequence. But, most likely, this prediction can be easily improved by removing several obvious mistakes (signaled with arrows in the figure), like broken helices or helices that are impossibly short. Because of this, the output vectors returned by the classifier are not used for prediction, but instead given to the second processing level of the system, described next.

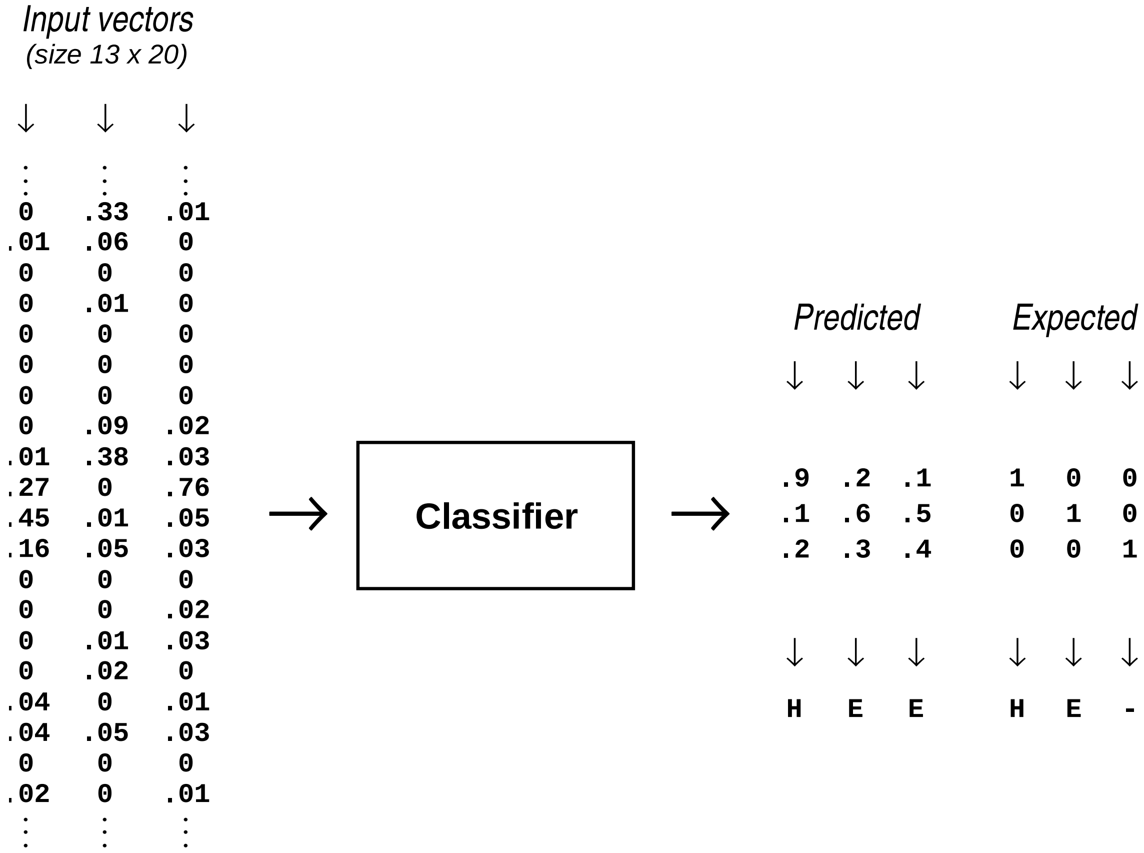


Figure 4.5: Classifier. Example of prediction of three input vectors.

4.4 Filter

The *filter* - the second and last processing level of our prediction system - is also a multilayer perceptron. The input vectors it receives are built from 17-residue windows over the prediction returned by the classifier. Because this prediction consists of a triplet of likelihood values for each residue, each input vector has size 17 × 3 = 51, containing the 17 triplets positioned size by side. The filter has 51 input neurons, 17 hidden neurons, and 3 output neurons. The output vectors used for training it are the same used for the classifier. Figure 4.6 exemplifies a filter receiving several input vectors and returning the predicted output vectors, once again comparing it with what could be the correct (expected) classification. Figure 4.7 (middle) shows what could be the prediction made by the filter, an improvement over the classifier prediction by the removal of several noticeable errors (see top of figure).

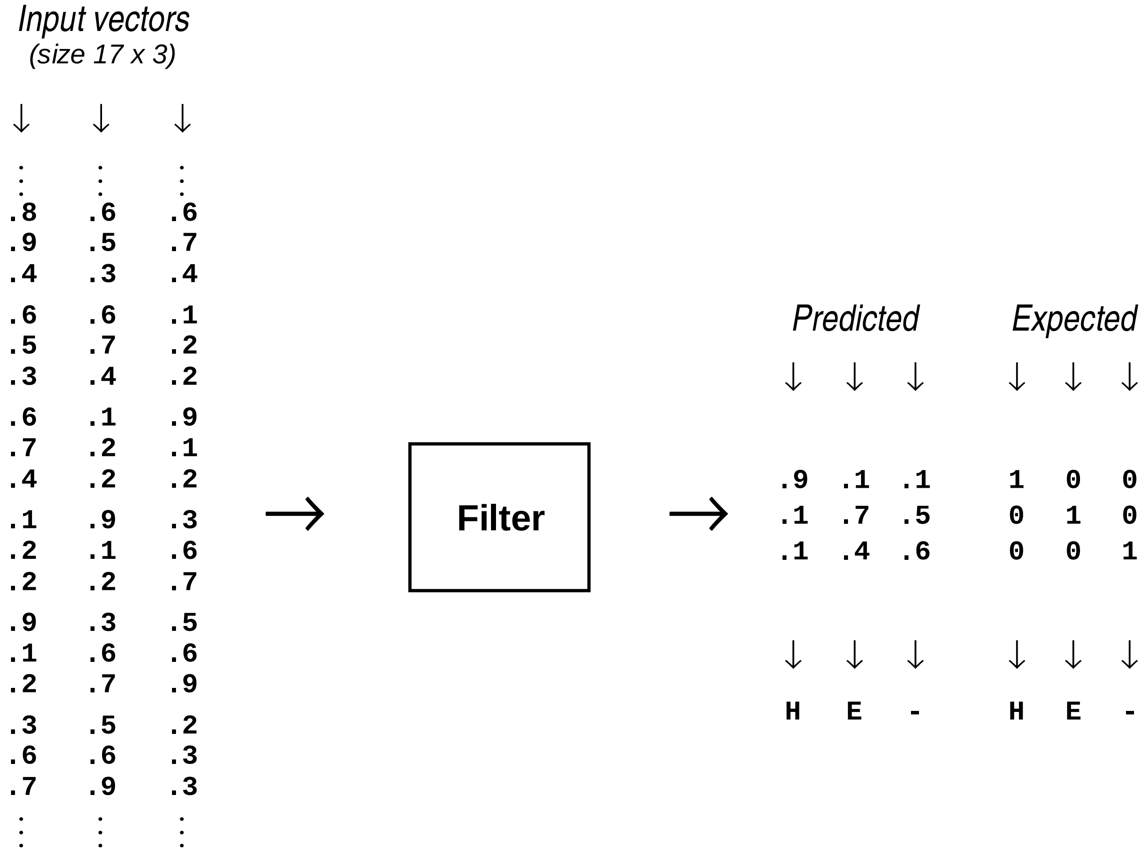


Figure 4.6: Filter. Example of prediction of three (consecutive) input vectors made from the output vectors of the classifier.

Prediction made by the classifier:

```
-HHH-HH-H-HHHH-EEEE--EEEHEE--HHHHH----H---EEEE--
    ↑  ↑ ↑              ↑             ↑
```

Prediction with obvious errors removed by the filter:

```
-HHHHHHHHHHHHH-EEEE--EEEEEE--HHHHH--------EEEE--
```

Prediction with reliability index:

```
-HHHHHHHHHHHHH-EEEE--EEEEEE--HHHHH--------EEEE--
788997678999987567776545666233788976312244675662
```

Figure 4.7: Improvement of prediction quality by the filter and reliability index (example). Top: prediction made by the classifier alone, with several obvious errors identified with arrows; Middle: prediction after being filtered, with the obvious errors removed; Bottom: prediction along with reliability index.

4.5 Reliability index

Along with the secondary structure prediction for each residue, the PHD system returns a value between 0 and 9 indicating the confidence the system has in the prediction made (9 represents the highest confidence). This reliability index is based on the difference between the highest and the lowest value of the output vector giving the prediction.

Given the extreme usefulness of this extra output (see chapter 5), our prediction system also provides a reliability index, based on the PHD index multiplied by the highest likelihood value. While the PHD system only uses the difference between the highest and lowest likelihood values returned, our system also considers the magnitude of the highest value. This results in lower values of the reliability index, making our system a bit less optimistic than PHD.

Table 4.3 illustrates the calculation of the reliability index used in our system for two different residue predictions. Note the low values achieved for such apparently reliable predictions. Figure 4.7 (bottom) shows what could be the reliability index values accompanying the prediction made by the filter.

A linear correlation between reliability and accuracy is a desirable property of any reliability index, shared by both the PHD and our index (although slightly higher with our index). Figure 4.8 shows, for a given data set, the percentage of residues receiving each reliability value, along with their prediction accuracy, for the PHD index (left) and our index (right). As expected, our system returns lower reliability values than PHD (for example, 47% of residues are classified with reliability higher than 5 with our index, against 59% with the PHD index).

Table 4.3: Calculation of the reliability index.

Output vector	**[0.9,0.5,0.2]**	**[0.5,0.1,0.1]**
Reliability index		
difference between the two strongest outputs...	$0.9 - 0.5 = 0.4$	$0.5 - 0.1 = 0.4$
...multiplied by the strongest...	$0.4 \times 0.9 = 0.36$	$0.4 \times 0.5 = 0.2$
...scaled between 0 and 9...	$0.36 \times 9 = 3.24$	$0.2 \times 9 = 1.8$
...and rounded	**3**	**2**

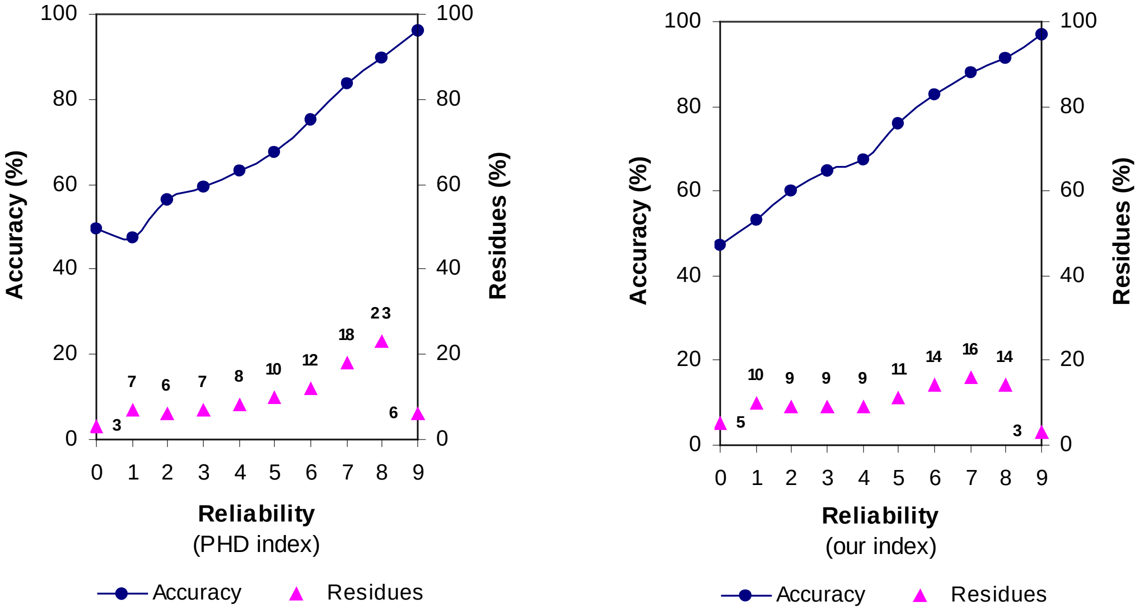


Figure 4.8: Reliability *versus* residues *versus* accuracy (both indices).

Chapter 5

Results

The results presented next are only an example of what can be achieved with a system like the one described, and a good quality data set. More than that, they exemplify how secondary structure prediction results can be presented, how much more information is gained by using the reliability index, and finally how accuracy measures alone can represent a dangerous pitfall.

5.1 Without reliability index

The most common way to present accuracy measures for a given classification is by giving the percentage of correctly classified elements, along with the omission and commission errors obtained for each class. The omission error represents the percentage of elements that *should* have been classified as belonging to a certain class, and were not; the commission error represents the percentage of elements that should *not* have been classified as belonging to a certain class, and were. Figure 5.1 (bottom left) shows the omission and commission errors obtained with our prediction system in one of the data sets used, for all the elements (residues) of all the proteins in the set. Errors in the classification of sheets are typically higher than in the other classes; helices are the easiest motif to predict.

However, when presenting accuracy measures regarding a set of proteins, it is much more useful to consider each protein separately. For each protein, the percentage of correctly classified residues is calculated. The results are then given as the mean and standard deviation of these percentages, as shown also in figure 5.1 (top left). Along with it, an histogram of the accuracy values is also very useful (same figure, right). Users of such a prediction system know that most of the proteins are classified with accuracy between 70% and 80%, that some are classified with less than 50% accuracy, that very few are classified with accuracy higher than 90%, and so on. Still, these results do not hint whether new proteins will be classified with as low as 40% or as high as 100% accuracy - users cannot rely on the predictions made by this system, unless they get a little more information. Hence the importance of the reliability index.

5.2 With reliability index

Figure 4.8 (page 22) shows a clear correlation between the values of the reliability index and the accuracy of classification. Knowing that, it is safe to trust predictions with high reliability, while disregarding the ones with low reliability. Even along the same polypeptide chain reliability is variable, so the user can obtain a partial but highly accurate prediction by retaining only the residues classified with high confidence.

Figure 5.2 illustrates another way of presenting the information contained in figure 4.8 (our index, right, page 22). The plot shows the percentage of residues classified with a certain minimum reliability (equal or higher than a certain value), and the accuracy obtained in their classification. This tells us, for example, that almost half of the residues were classified with reliability 6 or higher, and the accuracy obtained on these residues was higher than 85% - certainly much more informative than the results provided without the reliability index .

Accuracy (%):

Mean	74
Std Dev	9

Errors (%):	helix	sheet	other
Omission	24	31	25
Commission	19	36	26

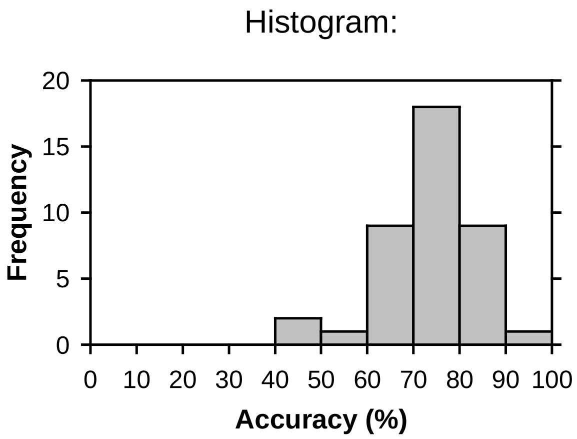


Figure 5.1: Results without reliability index: mean and standard deviation accuracy for all proteins (top left); omission and commission errors for all residues (bottom left); histogram of accuracy for all proteins (right).

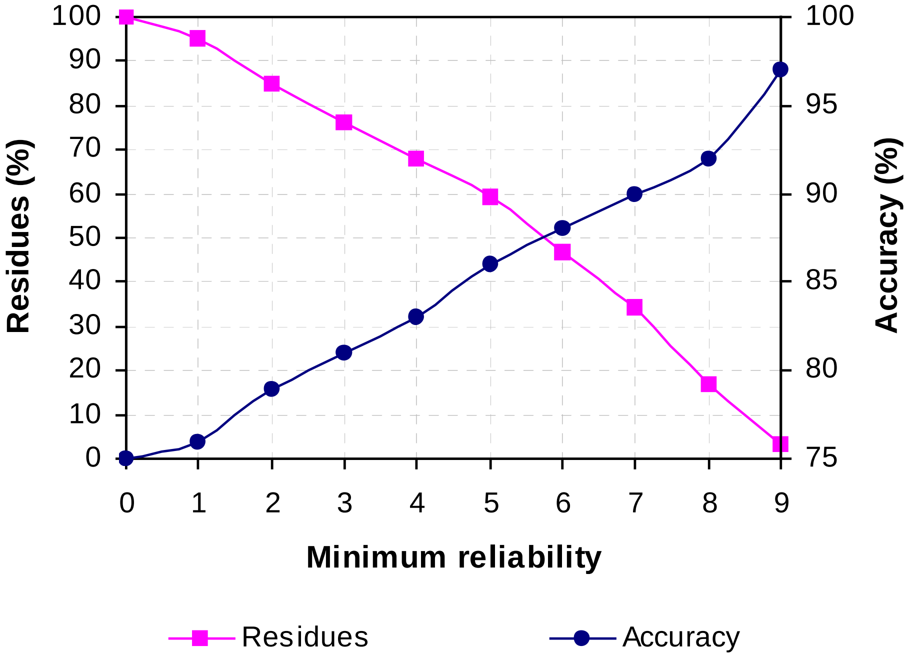


Figure 5.2: Results with reliability index: minimum reliability *versus* residues *versus* accuracy.

5.3 A note of caution on accuracy

Presenting accuracy as the percentage of correctly classified residues may be misleading in the context of protein secondary structure. Figure 5.3 shows an example. *C* is the true (correct) secondary structure of a small sequence; *P1* and *P2* are two different predictions for the same sequence. P1 correctly identifies the three motifs, although not with the right length or placement - its accuracy is 57%. P2 provides a truly messy prediction, where helices and sheets are impossibly short and interleaved - its accuracy is 79%!

It is much more important to provide a *realistic* prediction, where motifs are clearly identified, even if slightly misplaced, than a prediction which is clearly impossible, even if most of the residues are matched with the right motifs. P2 is useless, and still its accuracy is very high. Therefore, the accuracy measures presented here should be regarded with caution. The performance of a prediction system should also be assessed with other techniques, more appropriate in secondary structure prediction (see [6, sect. 6.7]).

```
 C:  --HHHHHH---EEEE----HHHHHH---
P1:  -HHHHH---EEEEEEEE-----HHHHH-    → 57% accuracy
P2:  --HEHEHH--HEEHE----HEHH-H---    → 79% accuracy!
```

Figure 5.3: A note of caution on accuracy measures. C: the correct secondary structure; P1 and P2: two different predictions, and respective accuracy values. The unrealistic prediction (P2) is the one with highest accuracy!

Chapter 6

Final considerations

Predicting protein secondary structure, based only on its sequence, is an apparently simple task that has been challenging several generations of prediction methods for already 30 years.

Using the evolutionary information contained in sequence alignments has allowed third generation methods to improve the prediction accuracy dramatically. However, if there are no available homologues to the protein whose structure we want to predict, the absence of the alignments results in impaired performance. Fortunately, this limitation tends to gradually disappear as the number of known sequences continues to rise.

Using a limited size window for the prediction of the central residue also poses a difficulty, as the exact same segment may fold differently when found in different proteins. In fact, even an entire sequence may adopt different conformations, depending on the properties of the solvent, so increasing the window size cannot guarantee success. Another problem is that the conformation of homologous proteins can vary more than 10% (although mostly in the extremities of the sequence), immediately imposing an upper limit of less than 90% accuracy for any third generation prediction method [18].

Predicting protein secondary structure with total accuracy is now accepted to be an impossible task. The best prediction methods have instead managed to achieve another difficult task: providing reliable and useful predictions, in spite of their limitations.

Bibliography

[1] Sejnowski, T.J., Rosenberg, C.R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, 1: 145–168

[2] Anderson, J.A., Rosenfeld, E., coords. (1998). *Talking nets: an oral history of neural networks.* MIT Press.

[3] Qian, N., Sejnowski, T.J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, 202: 865–884

[4] Rost, B., Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, 232: 584–599

[5] Cohen, B.I., Cohen, F.E. (1994). Predictions of protein secondary and terciary structure. In Douglas W. Smith, coord., *Biocomputing - Informatics and Genome Projects*. Academic Press, 203–232

[6] Baldi, P., Brunak, B. (2001). *Bioinformatics: The Machine Learning Approach*, second edition. Bradford Books.

[7] Chou, P.Y., Fasman, G. (1974). Conformational parameters for amino acids in helical, b-sheet, and random coil regions calculated from proteins. *Biochemistry*, 13: 211–222

[8] Chou, P.Y., Fasman, G. (1974). Prediction of protein conformation. *Biochemistry*, 13: 222–245

[9] Gibrat, J.F., Robson, B., Garnier, J. (1987). Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.*, 198: 425–443

[10] Rost, B., Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19: 55–72

[11] Garnier, J., Osguthorpe, D.J., Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, 120: 97–120

[12] Sander, C., Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9: 56–68

[13] Rost, B., Sander, C., Schneider, R. (1994). PHD - an automatic mail server for protein secondary structure prediction. *CABIOS*, 10: 53–60

[14] Holley, L.H., Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. USA*, 86: 152–156

[15] Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Lautrup, B., Nrskov, L., Olsen, O.H., Petersen, S.B. (1988). Protein secondary structure and homology by neural networks. The a-helices in rhodopsin. *FEBS Lett.*, 241: 223–228

[16] Kabsch, W., Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22: 2577–2637

[17] Riis, S.K., Krogh, A. (1996). Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comp. Biol.*, 3: 163–183

[18] Rost, B., Sander, C., Schneider, R. (1994). Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, 235: 13–26

[19] Hobohm, U., Scharf, M., Schneider, R., Sander, C. (1992). Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Sci.*, 1: 409–417

[20] Hobohm, U., Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.*, 3: 522–524

[21] Michie, A.D., Orengo, C.A., Thorton, J.M. (1996). Analysis of domain structural class using an automated class assignment protocol. *J. Mol. Biol.*, 262: 168–185