17 Secondary Structure of Proteins

learning objectives:

- fundamentals of description of protein structure in terms of amino acid subunits, and of protein secondary structure: α-helix, β-sheet and coil
- how to code the amino acid sequence for input into a network
- use of the moving window scan to process the protein chain as a series of overlapping neighborhoods
- comparison of results from network with results of a traditional structure prediction method

17.1 The Problem

Polypeptides and proteins are made up of elementary building blocks, the amino acids (a polypeptide is a short chain of amino acids; a protein is a long chain); apart from some special cases, only 20 of the many different amino acids occur in proteins. Figure 17-1 shows two of these amino acids, along with their abbreviations (a three- and a one-letter code).

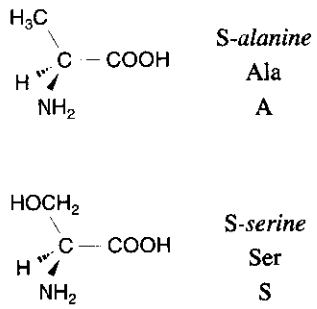

Figure 17-1: Two naturally occurring amino acids, their structures and their three- and one-letter abbreviations.

These amino acids are arranged sequentially in a protein; the exact sequence is called the *primary structure*. Figure 17-2 shows the sequence of amino acids in a segment of a protein. (Amino acids in a protein are generically called "residues".)

This linear sequence folds and turns into a unique **three**-dimensional structure, which contains global features that are referred to as the *secondary structure*. There are three types of secondary structures: α-helix, β-sheet, and random coil.

In an α-helix structure, the protein chain turns continuously in the same direction to form a "spiral"; in a β-sheet, two or more parts of the same chain are aligned parallel in space; the term "coil" collects all the other more or less irregular three-dimensional arrangements of

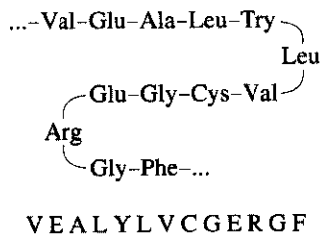

Figure 17-2: Part of the primary structure of the protein *insulin*.

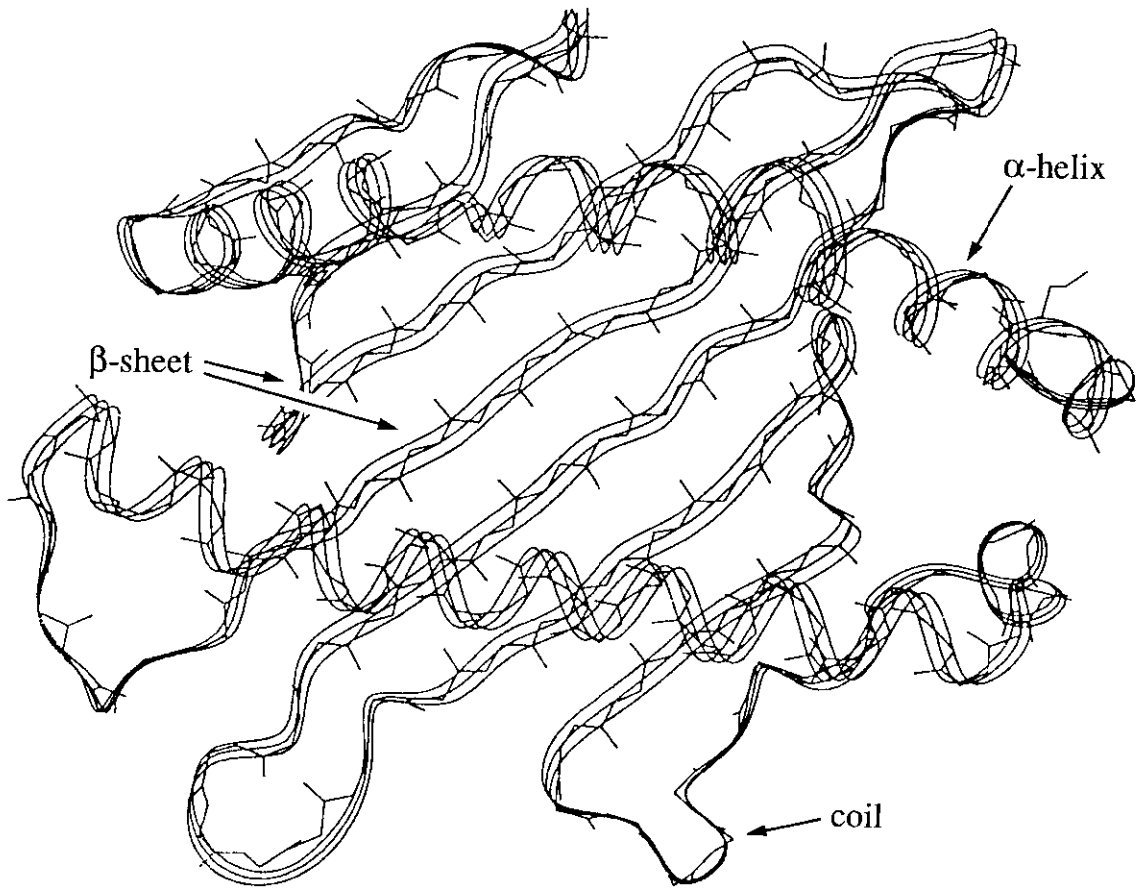

Figure 17-3: Residues 1 to 180 of human lymphocyte antigen A2. The three secondary structural features of a protein chain (α-helix, β-sheet, and coil) can be clearly seen. H_1, H_2, H_3 are three α-helices; the arrows indicate five "pleats" of a β-sheet; all the remaining parts of the molecule comprise a meandering "random coil". (Picture courtesy by Gerhard Müller and Horst Kessler, Org. Chem. Institut, TU München).

amino acids. All of these can be found in one protein. See Figure 17-3.

The secondary structure of a protein is of utmost importance to its biological activity.

Hence, there is much interest in predicting the secondary structures of proteins from their primary structures. The most widely used traditional approach is the method of Chou and Fasman (see References 17-8 and 17-9), which allows one to predict from the amino acid sequence whether a certain amino acid is part of an α-helix, a β-sheet, or a coil structure with about 50 – 53% correctness.

In recent years, numerous papers have been published on the use of neural networks to predict secondary structures of polypeptides from their amino acid sequences. The pioneers in this field were Qian

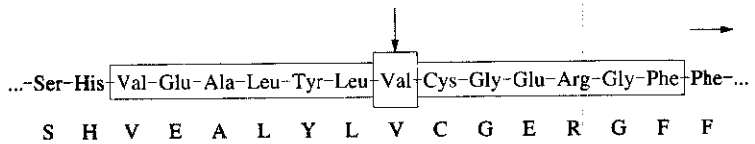

Figure 17-4: Segment (window) of the amino acid sequence thought to influence the secondary structure of the protein at a site centered on the middle amino acid (here, *valine*, Val).

and Sejnowski. Since quite a few other research groups adopted the essentials of their input representation, we will look at their work in some detail (see Reference 17-1).

The basic assumption in the work of Chou and Fasman and of Qian and Sejnowski is that the identities of an amino acid and its neighbors determine the secondary structure of that neighborhood. A sort of "window scan" over a whole polypeptide segment might in principle give the secondary structure of the whole chain (Figure 17-4).

17.2 Representation of Amino Acids as Input Data

In order to determine the dependence of secondary structure on the amino acid sequence, we must input the amino acid under consideration and a certain number (in this example six) of amino acids preceding and following it, a total of 13 amino acids. The sequence of 13 one-letter amino acid symbols, x_i^{orig} (original input variable), will be referred as to the **original input vector**, $\boldsymbol{X}^{orig}$.

Each of the 20 naturally occurring amino acids is coded as a 21-bit string with **one** specific bit turned on and the others all zero; for example, *proline* is represented by a 1 in position 14 of this string. The 21st position is special, and will be explained in the next section. Because of its discrete character, each original variable, x_i^{orig} (in this case, each amino acid label) must be represented by 21 binary or bipolar (see Section 4.2, Equation (4.1)) variables. Such a coding scheme is called a *distributed representation* (Figure 17-5).

We discussed in Section 16.2 (Equation (16.3)) the (very good) reasons for substituting a discrete variable by a bit string containing as many bits as the variable has discrete values. Representing an amino acid by a sequential number running from 1 to 21 would imply that the numbering of amino acids is a quantitative measure for the similarity between them; that is, if two amino acids have numbers that

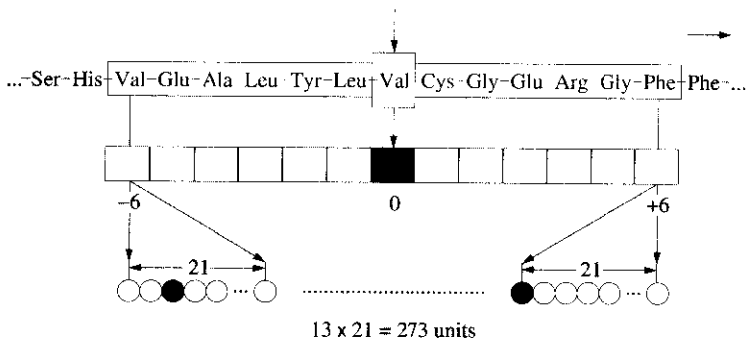

Figure 17-5: Coding scheme for the amino acid sequence.

differ by only one (e.g. 6 and 7) this would imply that these two amino acids are more closely related than, say, numbers 6 and 20.

However, what is important in determining secondary structure is not similarities in the structures of the amino acids, but similarities in their sequence order, that is, the relative position of a given amino acid from the center of the window plays the crucial role in the decision making, and not its structure.

(We are speaking here from an **information science** point of view, not a chemical one. Of course, the substituent groups ("R-groups") that distinguish one amino acid from another are primarily responsible for protein structure – but, as a first approximation, we **do not care** why the amino acids do what they do.)

Therefore, in this example the 13 original input variables x_i^{orig} are replaced by a 13 x 21 = 273-element **binary** input vector x_i. This means that the same number, 273, of units is required for input in the network, each input unit receiving one binary value, (0 or 1).

> The i-th original input variable, x_i^{orig}, tells us which of the twenty amino acids is presently at the $(i - 7)$-th position relative to the central amino acid in the 2 x 6 + 1 = 13-residue-long window.

The concept of a window bracketing a certain neighborhood, whether of a topological, sequential, or time-dependent nature, was used in the process control examples. See Sections 9.5 and 16.2.

Input of the entire sequence of amino acids of a protein (primary structure) is achieved by moving this window of 13 amino acids along the entire sequence in steps of one amino acid at a time. At each of

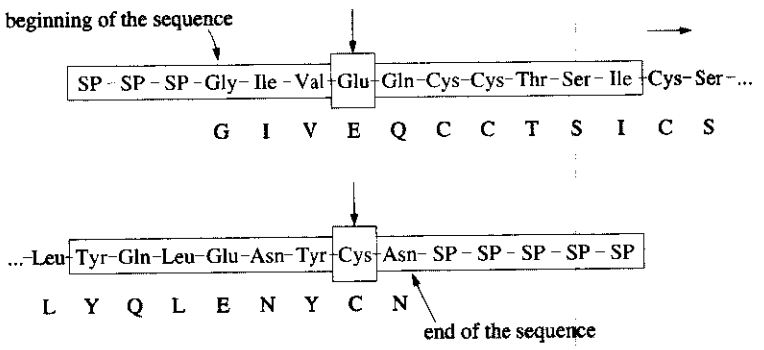

Figure 17-6: When the moving window is at the end of the chain, it has to be filled out with spacers.

these steps the corresponding window of amino acids is input into the network.

Special provision has to be made when the *moving window* is at the beginning or end of an amino acid chain; if it extends beyond the end, it will span fewer that the usual 13 residues (Figure 17-6).

So as not to complicate the algorithm, we will assume that there are always 13 residues in the window; at the beginning or end of the chain, any empty spaces will be filled with a special code called a *spacer*, coded using the 21st position of the bit vector.

(This coding scheme was inspired by Sejnowski's (very successful) work in training a neural network, NetTalk, to derive the pronunciation of a letter in an English word from the letters surrounding it.)

17.3 Architecture of the Network

In the network for predicting the secondary structure of proteins, three output neurons were used, one each for α-helix, β-sheet, and random coil. Qian and Sejnowski tried different numbers of hidden neurons (0 – 80) and decided that the optimum number is forty. Thus, the two-layer neural network has an architecture of (273 x 40 x 3), amounting to (273 + 1) x 40 + (40 + 1) x 3 = 11,083 weights, including those to the bias (Figure 17-7).

The training set consists of 106 proteins, having altogether 18,105 amino acid residues. Each of these is accompanied by a specification of the kind of secondary structure it is embedded in. (A given amino acid can be in different types of structures in different proteins, or even in different parts of the same protein.) Another 15 proteins with a

total of 3,520 amino acids and their known participation in a secondary structure are taken as the test set. The training is a supervised learning process performed with the back-propagation algorithm.

17.4 Learning and Prediction

A network consisting of more than 11,000 weights is quite a large one. Ten epochs of training with 18,000 amino acid sequences requires about 2 billion ($11 \times 10^3 \times 18 \times 10^3 \times 10$) weight corrections. Obviously, this is a major undertaking. Commercial neural-network software is now available, and, for problems of this size, special accelerator boards for plugging into your PC. Some of these can accelerate the calculation by factors of 50 or more compared to a 33 MHz 486 computer.

The network gave 62.7% right answers on the test set. This is a remarkable improvement over the method of Chou and Fasman, which has a predictive ability of only 50 – 53%.

The publication of Qian and Sejnowski stirred quite some interest among protein-structure chemists. Since then (October, 1988), a number of papers on this subject have been published, and a number of improvements and suggestions have been made, from enlarging the window to learning a larger number of amino acid sequences.

In fact, the tide of scientific work has continued to rise since publication of the first edition of this book. The work up to 1996 has been summarized by B. Rost. An excellent overview of all aspects of protein structure prediction has appeared by the same author in the Encyclopedia of Computational Chemistry.

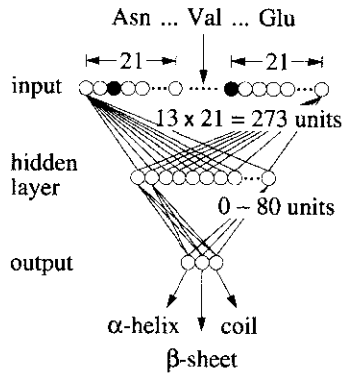

Figure 17-7: Architecture of the neural network for deriving the secondary structures of proteins from their amino acid sequences.

17.5 References and Suggested Readings

17-1. N. Qian and T. J. Sejnowski, "Predicting the Secondary Structure of Globular Proteins Using Neural Network Models", *J. Mol. Biol.* **202** (1988) 865 – 884.

17-2. L. H. Holley and M. Karplus, "Protein Secondary Structure Prediction with a Neural Network", *Proc. Natl. Acad. Sci. USA* **86** (1989) 152 – 156.

17-3. D. G. Kneller, F. E. Cohen and R. Langridge, "Improvements in Protein Secondary Structure Prediction by An Enhanced Neural Network", *J. Mol. Biol.* **214** (1990) 171 – 182.

17-4. H. Bohr, J. Bohr, S. Brunak, R. M. J. Cotterill, B. Lautrup, L. Norskov, O. H. Olsen and S. B. Petersen, "Protein Secondary Structure and Homology by Neural Networks. The α-Helices in Rhodopsin", *FEBS Lett.* **24** (1988) 223 – 228.

17-5. H. Bohr, J. Bohr, S. Brunak, R. M. J. Cotterill, F. Fredholm, B. Lautrup, O. H. Olsen and S. B. Petersen, "A Novel Approach to Prediction of the 3-Dimensional Structures of Protein Backbones by Neural Networks", *FEBS Lett.* **261** (1990) 43 – 46.

17-6. H. Andreassen, H. Bohr, J. Bohr, S. Brunak, T. Bugge, R. M. J. Cotterill, C. Jacobsen, P. Kusk and B. Lautrup, "Analysis of Secondary Structure of the Human Immunodeficiency Virus (HIV) Proteins p17, gp120, and gp41 by Computer Modeling Based on Neural Network Methods", *J. Acquired Immune Defic. Syndr.* **3** (1990) 615 – 622.

17-7. S. Brunak, J. Engelbrecht and S. Knudsen, "Neural Network Detects Errors in the Assignment of MRNA Splice Sites", *Nucleic Acid Res.* **18** (1990) 4797 – 4801.

17-8. P. Y. Chou and G. D. Fasman, "Conformational Parameters for Amino Acids in Helical, β-Sheet, and Random Coil Regions Calculated from Proteins", *Biochemistry* **13** (1974) 211 – 222.

17-9. P. Y. Chou and G. D. Fasman, "Prediction of Protein Conformation", *Biochemistry* **13** (1974) 222 – 241.

17-10. B. Rost, "PHD: Predicting one-dimensional protein structure by profile based neural networks", *Meth. Enzymol.* **266** (1996) 525 – 539.

17-11. B. Rost, "Protein Structure: Prediction in 1D, 2D, and 3D", in *Encyclopedia of Computational Chemistry*, Eds.: P. v. R.

Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III and P. R. Schreiner, Wiley, Chichester, UK, 1998, pp. 2242 – 2255.