Elements of
ARTIFICIAL NEURAL NETWORKS
with selected applications in Chemical Engineering, and Chemical & Biological Sciences

Sanjeev S. Tambe, Bhaskar D. Kulkarni
National Chemical Laboratory
Pune 411 008, India

and

Pradeep B. Deshpande
University of Louisville
Louisville, KY 40292, USA.

Contributors

Chapter 8

Sriram Ramani, Nithya Srinivasan & Raul Miranda
University of Louisville

Chapter 9

T. Murlidharan Nair
National Chemical Laboratory
Pune, India.

Simulation & Advanced Controls, Inc.
Suite 307, The Landmark Building, 304 West Liberty Street,
Louisville, KY 40202, USA.

Chapter 9. Artificial Neural Networks in Biological Sciences

The quest to understand biological phenomena and answer questions related to them requires a sound understanding of both structure and function. A variety of techniques have been used in answering such questions. Toni Kazic in her article " Reasoning about biochemical compounds and processes" (1992) has very neatly broken problems related to biology by posing two questions:

What is it ?
What does it do ?

While it is important to understand and explicitly delineate the relationship between structure, the 'it', and function, the 'do', it is equally important to identify the 'it' and find out the 'do' that is related to it. Biological systems being very complex, it is often difficult to identify individual components of the systems and establish the way in which they interact with each other. The inherent complexity of biological systems makes it very difficult to understand them as well as to model them phenomenologically. Further, owing to the difficulties associated in obtaining complete information of the interacting components of a biological system, they appear to be very complex in nature. Thus, it is important to use an alternative approach that can be applied to systems about which only partial information is known. The approach should further help in providing some clue to the identification of the 'it' and the 'do' it is related to.

Starting from the issue of protein folding to that of the non-coding regions in DNA, many questions need to be answered. There have been a number of

approaches, both experimental and theoretical, that have been directed towards understanding these and other complex biological problems. Theoretical approaches include statistical analysis, spectral analysis, linguistic analysis, Monte-Carlo methods and molecular dynamics approaches. Further, there is also a growing need to develop faster and newer methods in understanding biological processes. This need mainly stems from the explosive generation of sequence information as a result of the different genome projects. These projects would have no meaning if the valuable biological information hidden in the sequences is not extracted. Thus, it is of paramount importance to develop techniques to unscramble the words in the sequence and read the hidden message. The encryption of messages in biological sequences is complex. It is now being established that sequences no longer carry a single message (e.g., the triplet code which are instructions for protein synthesis) but, in fact, carry overlapping messages like the DNA shape code and the chromatin code. Other signals which are responsible for vital cell activities like transcription are also encoded in different regions. While the aforementioned methods are very useful in understanding a variety of biological phenomena, there still remain certain processes which cannot be analyzed using these techniques.

ANNs appear to be one of the most suited alternative tools. Artificial neural networks are mathematical approximations of the biological synapses and were initially developed as models for understanding the brain mechanisms involved in perception. The ability of the ANNs to perform nonlinear mapping and their powerful internal representation capability has led neural networks to be used as a tool for modeling rather than understanding the brain functions per se. Recent advances in neural network theory and technology have made them a powerful tool that helps to identify complex processes in the presence of noisy or incomplete information, colinearity of data, and time delays. It can also be used on incomplete data without assumed models or postulated formulas. Further, there are several features of neural networks that have encouraged their application to the analysis of protein and nucleic acid sequences. ANNs can incorporate both positive and negative information, that is, both sequences with the feature of interest and without the feature are used to impart knowledge to the network. They are also able to detect second- and higher-order correlations in patterns, and thus, are more useful in determining complex correlations than the conventional methods based simply on the frequency of occurrence of residues at certain positions. An artificial neural network based on the knowledge it acquires at the time of training makes its own internal representation of the system being modeled and then automatically determines which residues and which positions are important. Neural networks are thus ideally suited for parallel sequence processing and are increasingly applied to the study of biological macro-molecules. They aim at mapping nucleic acid/protein sequences on to spatial structure/functionality. These properties of ANNs make them applicable to solve a number of complex biological problems. Their biological application is not limited to only sequence analysis; they have been employed to solve problems in medicine and many other areas in biology as well. There has been an astounding increase in the application of ANNs in biology in the past few years

(see reviews: Hirst and Sternberg, 1992; Burns and Whitesides, 1993; Sumpter et al., 1994; Rawlings and Fox, 1994), and the present chapter aims at providing a comprehensive overview of the applications of ANNs in biological sciences. Further, the chapter aims to bring forth to the reader, areas in biology that are amenable to black-box modeling, from which potential problems may be identified and tackled using the techniques detailed in Chapters 3 to 5. The chapter will not be detailing the ANN algorithms as these have been discussed in the first few chapters.

9.1. APPLICATION OF ANNs TO NUCLEIC ACID SEQUENCE PREDICTION.

Neural networks have been extensively used in the analysis of nucleic acid sequences ; the main reason being the explosion of data which provides the learning space for the network. ANNs have found numerous applications in the areas of promoter recognition, terminator recognition, non-coding region of DNA, capturing transcription control signals, phylogenetic analysis, etc.

9.1.1. DNA Promoter Recognition. A promoter is a start site at the beginning of a gene or a gene cluster which harbors the signal that directs the enzyme RNA polymerase to initiate RNA synthesis. Since RNA polymerase recognizes a particular region upstream to the start site of the gene, this region, in principle, should contain signals which may be a function of the structure associated with the region or a result of the sequence distribution. Earlier analysis of promoters (most of them from *E. coli*) has revealed the presence of two conserved regions situated 10 and 35 bp upstream from the transcription starting point, viz. the -10 and the -35 boxes. The coding strand conserved sequences are TATAAT and TTGACA at the -10 and the -35 regions, respectively. Identification of promoter regions becomes important to delineate crucial sequence elements in the process of recognition that harbor signals to initiate the RNA synthesis.

Most of the work on DNA promoters has been done from the sequences from *E. coli*. This is mainly due to the fact that there is enough data available on *E. coli* as a result of the identifiction, sequencing and characterization of a large number of promoters. Nakata et al. (1988) used discriminant analysis to predict promoter regions in *E.coli*. The criteria for discrimination were based on the accuracy of consensus sequence patterns measured by the perceptron algorithm, the thermal stability map and the base composition. The discriminant analysis also takes into account the Calladine-Dickerson rules. These rules were formulated by Calladine (1982) to explain the departure of the DNA dodecamer C-G-C-G-A-A-T-T-C-G-C-G, from the ideal regular helical structure of B-DNA. Calladine used principles of elastic beam mechanics to analyze this clash and ways of relieving it. He proposed that the DNA chain may ameliorate these van der Waals clashes in four ways:

(i) Flatten the propeller twist in one or both pairs.
(ii) Open up the roll angle between base pairs on the side where clash is found.
(iii) Shift its backbone sideways towards the pyrimidines.
(iv) Decrease the local helix twist angle at the step at which the clash occurs.

For more details the reader is advised to refer to the original work by Calladine (1982) and Dickerson (1983). While the unaided perceptron predicted promoter regions in *E. coli* with 67% accuracy, the inclusion of other information in the prediction algorithm increased the predictive ability to 75%. Lukashin et al. (1989), Demeler and Zhou (1991), and Mahadevan and Gosh (1994) have tackled the same problem with networks of increasing complexities, and obtained accuracies of 94%-99%. Two three-layer neural networks were used by Lukashin et al. (1989) in capturing the conserved regions present in the -10 and the -35 region. In their procedure for training the network, a small part of the total set of promoter sequences was used to develop a system of distinctive features, which was later used as a reference in identifying promoters against the background of random sequences. A recognition accuracy of 80% was achieved by O'Neill (1991) in predicting *E. coli* promoters of the 17 base spacer class with false positive rates below 0.1%. The training set used by O'Neill (1991) comprised 5148 58-base sequences drawn from 39 promoters and 4000 random sequences which were 60% AT- rich. Further, the true set of sequences was expanded by permuting all possible single base changes in positions other than those known to harbor promoter point mutations.

In the analysis carried out by Demeler and Zhou (1991), an optimized error-back-propagation (EBP) network was developed by using two different coding schemes (two-bit and four-bit) and by training the network using various ratios of promoter sequences to non-promoter sequences (1:1 to 1:20). In the two-bit coding scheme, the nucleotides were coded in a linearly dependent set of dimension 2 called CODE-2 (00=A; 01=T; 10=G; 11=C), while in the orthonormal set of dimension 4, called CODE-4, the nucleotides were coded in 4-bit binary (0001=C; 0010=G; 0100=A; 1000=T). The data for training were taken from the *E. coli* RNA polymerase promoter sequences compiled by Hawley and McClure (1983). From these, 80 bacterial and phage promoters were used for the training, and the remaining 30 plasmid and transposon promoter sequences and the promoters generated by mutation were used in a test set. The promoter sequences were arranged into the following three independent training and corresponding test sets :

(i) 20 bases centered on -10 region containing the TATAAT consensus sequence wherein the first T was placed at the 12th position,

(ii) 20 bases centered on the -35 region containing the TTGACA consensus sequence with the first T at the 10th position, and

(iii) 44 bases as aligned in (i) and containing the -35 region without a gap between the conserved -10 and -35 regions.

The non-promoter or random sequences with an equal composition of A, T, G and C were generated using a pseudo-random number generator. The network was trained to an error level of 10^{-4}. The effect of the extent of training was determined using the network in the prediction mode for error levels between 0.001-10. The number of neurons in the hidden layer were also varied to determine the effect of

network complexity on the prediction capability. The simulation results indicated that CODE-4 was a better choice for DNA data representation. CODE-4 is a unitary coding matrix with an identical Hamming distance between vectors. Hamming distance is a measure of the difference between vectors arrived at by counting the number of different bit entries; for example, a vector (1,1) has a Hamming distance 1 compared to the vector (0,1) while it is 2 when compared to the vector (0,0). In CODE-4 encoding, all the vectors have a Hamming distance of 2 between each other. The number of neurons in the hidden layer did not seem to influence the results obtained in a predictable fashion and differences were minor. It was further seen that increasing the number of neurons in the hidden layer increases the representational power of the network but decreases the generalization ability of the network. As regards the error levels during the training, lower network error levels improve promoter prediction accuracy, while larger *r:p* ratios improve the network's ability to filter out false positives. The paper presents a thorough work on the DNA promoter recognition. In another approach, the EBP networks have been trained to recognize *E. coli* promoters of the 17 base spacing class (O'Neill, 1991). The entire promoter sequence length was 58 bases i.e., -50 to +8. The output of the net was either 0 or 1, depending on whether the input is a non-promoter or a promoter. The network with 15 hidden neurons gave an optimal performance. An optimal efficiency of 80% was achieved on the test data set with a false positive rate below 0.1%.

Another interesting work on the analysis of *E. coli* promoters using a module approach was addressed by Mahadevan and Gosh (1994). A back-propagation neural network was trained to identify *E. coli* promoters using a three-module structure. In the first neural net module, the consensus boxes were identified; in the second module promoters were aligned to a length of 65 bases, and in the third the entire sequence of 65 bases was recognized. Module-I comprised two neural networks which learnt the -35 and the -10 boxes separately. A sequence of 6 bases from each consensus region was coded using the CODE-4 strategy and presented to the net, thus constituting 24 neurons in the input layer. The output layer housed one neuron and the hidden layer, 2 neurons. The output corresponding to the promoter was designated 1 and for non-promoters it was 0. The training set contained 106 promoters. The non-promoters were random hexamers and did not match with more than two bases of the conserved -10 and -35 boxes. After elimination of the duplicate boxes from the set of 106 promoters, a total of 58 unique boxes of the -10 region, and 72 of the -35 region were used for training. The net was presented with promoters and non-promoters in the ratio 1:1. Module-II of the network aligned the sequences with respect to the boxes and spacers identified by module-I. The boxes with an output greater than 0.8 with a spacer of 15 to 21 bases between them were considered potential promoters and aligned by introducing a gap. Module-III of the network learnt the aligned sequences of length 65 from the 106 promoters. The network had 260 neurons in the input layer and 1 neuron in the output layer. The number of neurons in the hidden layer was varied between 2 and 12 and the prediction capability of the net ascertained. Not much variation in the prediction

was observed with varying number of neurons in the hidden layer. Seven neurons were found to be optimal and used for training purposes. The module-III of the network was tested on 60 bacterial, 26 bacteriophage and 40 mutant promoters. The output values greater than 0.8 were considered promoters and those below 0.8 non-promoters. Further, the trained net was used in the identification of single point mutations in p22ant promoter. Mutations in p22ant promoter and their effect on promoter activity has been well studied experimentally (Youderian et al., 1982; Moyle et al., 1991). The network was further trained with 11 mutated non-promoters and the 8 mutated promoters of the p22ant promoter. The network learned all the 11 mutated non-promoters (RU369, RU454, RE167, RU204, RU267, RU523, RU541, RU428, R-34TG, R-31CG and R-7TG) and 8 mutated promoters (RU1150, RU287, RU1002, RU1156, RU1012, R1173, RU1197 and R30AG) along with the p22ant promoters. The results of the prediction show that all the promoters except the two (RU1041 and R-9AC) and the non-promoters were predicted correctly by the network. The network was further used in the identification of the promoters of the pBR322 sequence. The network successfully identified the locations of P1, P2, and P3 promoters.

In a very different approach, Shavlik and Towell used problem-specific knowledge in understanding *E. coli* promoters (Shavlik and Towell, 1992; Towell and Shavlik, 1993; 1994). They used knowledge-based neural networks (KBANN) in the prediction of *E. coli* promoters and demonstrated the superiority of the KBANN to alternative techniques. Figure 9.1 summarizes the KBANN approach. It uses inference rules about the current biological problems which need to be only approximately correct to initially configure a neural network. The analytical techniques used combine neural network learning with the rule-based approach of the expert system. As mentioned earlier, the inference need not be perfectly correct and may be considered heuristic guesses. The algorithm maps this information, called *domain theory*, into a neural network and then uses a collection of examples from the

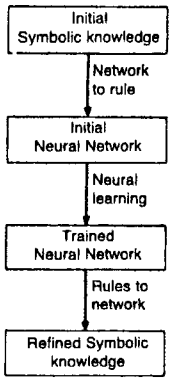

Figure 9.1. KBANN approach to sequence analysis. Reproduced with permission from Towell and Shavlik (1993).

training space to refine the initial knowledge. The refined information which is in the form of a large collection of numeric weights is incomprehensible. The algorithm then maps this incomprehensible information into more easily interpretable symbolic terms. The simulation results reveal that KBANN performs better as compared to any other approach. Shavlik and Towell (1992) have made a comparison of the different approaches. The KBANN algorithm can be used as a general method that provides a means for improving the existing roughly correct knowledge by the use of machine learning technique.

9.1.2. Transcription Terminator Recognition. ANN application in the recognition of terminators, which are sequences that primarily regulate gene expression by providing stop signals at the end of transcription units, was carried out by Nair et al. (1994). Earlier studies (Rosenberg and Court, 1979; Brendel and Trifonov, 1984) have shown that factor-independent terminators shared features like G-C rich dyad symmetry followed by a stretch of 4-8 adjacent thymine residues immediately upstream of the last nucleotide incorporated into the RNA chain. It should also be noted that there exist many independent terminators that do not comply with the consensus pattern of dyad symmetry and T-stretch (Brendel and Trifonov, 1984).

Learning space for the network developed by Nair and coworkers was from the terminator sequence compilation by Brendel et al. (1986). From a total of 128 terminators of length 51, 88 were chosen for training the network. The remaining 40 terminators were used as the test data set. A pseudo-random number generator was used for constructing random sequences with an equal composition of A, T, G and C. These random sequences were combined with the terminator sequences in the ratio 1:3 (one terminator followed by three random sequences). Nair et al. (1994) have used two different coding strategies in presenting the input data. In one case, the network was presented with data coded in binary, similar to CODE-4 used by Demeler and Zhou (1991). The target to each input sequence was coded '1' for a terminator sequence and '0' for the random sequence. The second type of coding is based on the *Informational Spectra Method* (ISM) (Veljkovic and Metlas, 1987; Veljkovic and Cosic, 1987; Veljkovic et al., 1985) which is a mathematical and a physical method for the analysis of the information content of DNA and protein sequences (Veljkovic and Metlas, 1988; Cosic et al., 1986). In this form of coding, the electron ion interaction potential (EIIP) associated with each nucleotide is calculated using the following equation.

$$W = 0.252\, Z^* \sin(1.04\,\pi\, Z^*) / 2\,\pi \tag{9.1}$$

where Z^*, the quasi-valence number is determined as

$$Z^* = \sum_{i=1}^{m} n_i\ Z_i \,/\, N \tag{9.2}$$

In this equation, Z_i denotes the valence number of the ith atomic component, n_i the number of atoms of the ith component, m the number of atomic components in the molecule, and N the total number of atoms.

The informational spectra method that uses the EIIP values is a tool for the analysis of the informational content of proteins and nucleotide sequences. It has also been used to obtain consensus spectra for different sequences (Cosic et al., 1986) and is aimed at establishing a relation between a sequence and its biological activity. Biological processes that take place in nature are highly specific, and are due to the selective interactions and recognition between macromolecules which take place at a relatively larger separation. The basis of molecular recognition has been attributed to the electric forces determined by the electrostatic potential around a molecule (Harrison, 1970). The electrostatic potential depends upon the distribution and the energy state of the valence electrons and the EIIP values are the physical parameters that influence the delocalized electrons. Studies have established a correlation between the EIIP values of organic molecules and their biological activity (Cosic and Nesic, 1987). The EIIP values for the nucleotides obtained using eqs. (9.1) and (9.2) are : A, 0.1260; T, 0.1335; G, 0.0806 and C, 0.1340. Thus, each nucleotide, irrespective of its position, is represented by a definite number, and the numerical series so obtained are finite length deterministic discrete signals. The normalized signals represent the input to the network. The targets are represented in a manner analogous to CODE-4. This representation will henceforth be referred to as the *EIIP code*.

Two separate neural nets were trained by Nair and coworkers using the aforementioned coding strategies. Since the CODE-4 method did not reflect any intrinsic property of the bases, the EIIP coding strategy which reflects a physical property of the system under study was used. Hence, the network was presented with a numerical series of the EIIP values which are finite length deterministic discrete signals corresponding to the terminator sequence. The network architecture for CODE-4 representation consisted of 204 (sequence length x 4) neurons in the input layer, a single hidden layer with 7 neurons and an output layer with 1 neuron. The network architecture for the sequences coded with their EIIP values consisted of an input layer with 51 neurons, a single hidden layer with 7 neurons and an output layer with 1 neuron. The number of neurons in the hidden layer was optimized to 7. The net was presented with 352 patterns consisting of 88 terminators and 264 random sequences for training. After every epoch, which corresponds to the presentation of all the 352 training patterns once to the net, the weights were extracted and used for predicting the test data set comprising 160 patterns (40 terminators and 120 random sequences). A network output lying between 0.5 and 1 indicates that the input pattern is a terminator and an output less than 0.5 corresponds to a random sequence. The network using the CODE-4 and EIIP strategies, correctly predicted 157 (98.125%) and 153 (95.625%) patterns, respectively, from the test set comprising 160 patterns. Neither coding strategy predicted any false positives, that is, none of the random sequences was predicted as terminators. Further, it has also been shown that a network trained by coding the four nucleotides with properly spaced arbitrary numbers (A, 0.25; T ,0.50; G, 0.75 and C, 1.0) and possessing the same architecture was inferior to the one using the EIIP coding strategy. Only 147 patterns out of 160 were predicted correctly and the

net also took a longer time to converge. The marginally lower prediction capability of the EIIP-code has been attributed to the smaller network size which means a lesser parameter space as compared to CODE-4. However, coding sequences by their EIIP values has the advantage that it reduces the network size to one fourth as compared to CODE-4; concequently, it reduces the training time.

Nair et al. (1994) also presented a very simple approach to delineate the sequences that play a crucial role in the process of recognition. In this approach, a fixed calliper of sequences were randomized and presented to the net for prediction. The analysis revealed that the sequences between 30 and 51 were most important in the recognition process. Figure 9.2 shows the error profiles of the network when different regions in the sequence were randomized. It is noteworthy that the region corresponding to the maximum error also corresponds to the region in the sequence containing the dyad symmetry and the T-run which are the well-known features of the terminators. This approach can be used as an alternative strategy to sequence alignment in arriving at the consensus of a sequence responsible for a particular biological function.

9.1.3 Translation Initiation Region. Translation initiation sites or ribosome binding sites are the protein synthesis initiation regions. They are defined by several groups performing nuclease protection experiments as the 20-40 base-long segments of mRNA that remain protected by the ribosome attachment when complexes formed of phage mRNA, ribosomes and fMet-tRNA$_{fmet}$ are exposed to ribonuclease (Gupta et al., 1970; Hindley et al., 1969; Steitz, 1969).

The following factors have been identified in the definition and function of a ribosome binding site (Bisant and Maizel, 1995) :

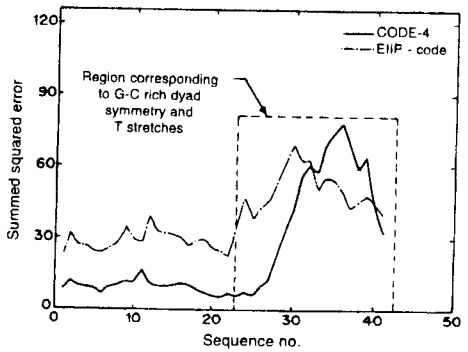

Figure 9.2. Calliper randomization approach to arrive at the consensus of a region responsible for a particular biological function. Reproduced with permission from Nair et al. (1994).

(a) the Shine-Dalgarno Element (SDE) (Gillam, et al., 1984),

(b) the start codon and the bases immediately surrounding it (Munson et al.,1984),

(c) the presence of multiple competing initiation codon (Matteucci and Heyneker, 1983),

(d) the distance between the start codon and the SDE (Hui et al., 1984),

(e) the sequence of the bases between the start codon and SDE (Matteucci and Heyneker, 1983; Ringquist et al.,1992),

(f) the composition of the bases far downstream (Petersen, 1987),

(g) the composition of the bases far upstream (Coleman et al., 1985),

(h) the secondary structure (Munson et al., 1984; Davis et al., 1985), and

(i) the presence of translation enhancers (Thanaraj and Pandit, 1989; Ivanov et al., 1992).

Recognition of these sites based on sequence data is difficult due to the multiple determinants that define them. The earliest approach of the neural network model to sequence analysis was by Stormo et al. (1982a), in which a perceptron with no hidden layer was used to predict the translation initiation sites in *E. coli*. The main objective was to investigate the features of mRNA that allow recognition by ribosomes, that is, to identify nucleotides which may play an important role in the selection of initiation codons by ribosomes of *E.coli*. They analyzed the sequences with the objective of obtaining a weighting function, which, when applied to any sequence would give a value, and depending on the magnitude of which it would be possible to determine whether that particular sequence is a ribosome binding site. The sequences were encoded in a $4 \times N$ matrix (N being the length of the sequence). The elements were encoded in binary (0 or 1), where 1 represented the presence of a base at a position and 0 its absence. The weighting function was determined using the perceptron algorithm. The perceptron convergence theorem guarantees that, if a solution exists, it will be found in a finite number of steps; the number of steps, however, may be large. The perceptron was trained with 124 known gene beginnings and 167 false beginnings. In the predictive mode, the perceptron identified six gene beginnings out of ten correctly and five false beginnings incorrectly. The rule-based approach (Stormo et al., 1982a) had predicted only five true beginnings and twelve false ones.

Recently, Bisant and Maizel (1995) explored the multilayered ANNs in the identification of *E. coli* ribosome binding sites. For their study, a new set of *E. coli* ribosome binding sites was compiled. Performance of all the neural networks and perceptron was evaluated by *receiver-operating-characteristic* (ROC) analysis (Green and Swets, 1988; Meistrell, 1990). The best neural network used an input window of 101 nucleotides and a single hidden layer of 9 units. Further, the result also indicated that the neural networks performed significantly better than the

perceptron of Stormo and coworkers (1982). The neural networks and perceptrons trained on the new compilation also performed better than the original perceptron (Stormo et al., 1982a).

9.1.4. mRNA Splice Site Recognition. In eukaryotic mRNA processing, splicing is a major event. It is characterized by the removal of the non-coding pieces of RNA molecules (introns) and the joining together of the remaining protein-coding fragments (exons) to form a continuous mRNA molecule. There are specific features that characterize the exons, introns and their junctions. The coding region harbors the specific three-base periodicity of their sequence, which can be used to locate them and the junctions between introns and exons. Further, the nucleotide sequence at the junction exhibits some degree of sequence conservation, like the GU sequence at the beginning of every intron and AG at the end (Breathnach and Chambon, 1981). Detailed analysis has revealed the presence of some semiconserved elements for the intron-exon junction, 5'GUAAGU, and the exon-intron junction: 5'-(U)nNCAG (Mount, 1982). Owing to a diffused sequence pattern distribution, these regions were investigated using a variety of approaches. Trifonov (1985) developed an algorithm for locating the splice junctions. Nakata et al. (1985) predicted mRNA sequences by discrimiant analysis of information, including consensus sequence patterns around splice junctions, free energy of snRNA and mRNA base pairing, and base composition and periodicity. In an interesting analysis, Brunak et al. (1990) used ANNs to predict the splice site location in human pre-mRNA. The analysis was similar to the promoter recognition studies. Separate networks were constructed to recognize donor and acceptor sites in the DNA. The data set contained 544 exons with a total of 133,372 nucleotides, the average length of the exons being 245bp. The data set comprised a total 95 genes (Brunak et al., 1991). The data set that was used for training the network contained the 1st 65 entries and the remaining 30 entries were used as the test data set. The total set of 95 genes contained 449 donor sites (and 449 acceptor sites), of which 331 donor/acceptor sites were part of the training set and the remaining of the test set. For further details of the data used, the reader may refer to the original work. The sequence similarity analysis, dinucleotide propensity analysis and the Shanon information analysis of the data set have also been carried out.

The network was a feedforward type, which received input from the windows scanning the DNA sequence. Each window configuration was represented numerically as a binary string. The sequences were classified as one of the following catagories, splicing donor site or non-splicing donor site, or splicing acceptor site or non-splicing acceptor site, or coding or noncoding. The performance of the network with widely different architectures was studied. Multilayer perceptrons with 0, 5, 10, 15, 20, 40 and 50 hidden units were used, all with 11, 13, 15, 17, 21, 31, 41 and 51 nucleotides visible in the input layer. Even though the results obtained were better than those by the weight-matrix methods of Staden (1984), the false positive rate of the donor was about five times the number true positives. In the case of the exon/intron boundary recognition, the net correctly detected 94% of boundaries not presented to the net earlier with 0.1% false

identification. The intron/exon detection was predicted with an 87% accuracy, with 0.2% false identification. The combined method detected the intron/exon and exon/intron boundaries with 95% accuracy. The false identification in the case of intron/exon was 0.4%, while in the case of exon/intron it was 0.1%. In comparison to this, the weight matrix method of Staden (1984) gave more false positives, that is, 0.7% for the same level of detection of true boundaries.

In yet another interesting application, neural networks were used by Brunak and coworkers (1990) to detect errors in the assignment of mRNA splice sites. They trained a network to recognize mRNA splicing signals in 33 human genes. During the training they noticed that some sequences appeared to disturb the learning process as the network weights did not stabilize on a specific signal assignment. Analysis of the data used for training revealed discrepancies from the original papers for three genes due to misprints and other errors of interpretation. A possible use of neural network as proofreaders to detect errors before accepting the data into a database has been suggested.

9.1.5. Coding and Non-coding Regions. Another very important application of neural networks in molecular biology is in finding protein coding regions in DNA sequences. The emerging importance of using neural networks in the identification of coding region is mainly due to the ongoing genome projects. The final goal of the genome project is to identify an estimated 100,000 genes which lie buried in the 3×10^9 bases of the human genome. While sequencing is progressing on at an alarming pace, data analysis will certainly become a rate limiting step. The succeeding phases of the project would then depend largely on interpreting nucleotide sequences by *in computo* experiments with a view to providing some insight into the location structure and functional class of protein-coding genes. It is needless to emphasize the importance of the problem and the consequent increase in the number of approaches, algorithms, and software to solve the problem is self-evident. For an overall information on the gene identification problem the reader is referred to the excellent review by Fickette (1995). Neural network analysis in the determination of protein-coding regions in eukaryotic DNA sequences has met with considerable success. The earliest attempts in this direction were by Lapedes et al. (1989) wherein the data representation consisted of isolated codon information. They however did not consider the Markov dependence of the codon. The intron-exon sensitivity that they obtained was 98.4%. The false positive rates were also high. In another detailed analysis, Farber et al. (1992) used dicodon frequencies and obtained a prediction accuracy of 99%.

A radically different approach was used by Uberbacher and Mural (1991), wherein a comprehensive system to analyze and characterize the genetic structure of DNA sequences was developed. The system is known as GRAIL (Gene Recognition and Analysis Internet Link). It uses a multi-sensor/neural network, expert system and parallel search tool combination to recognize and interpret genes in DNA sequences (Mural et al., 1992). Experimental methods used to locate genes are labor intensive, time consuming, and are subject to interpretation problems,

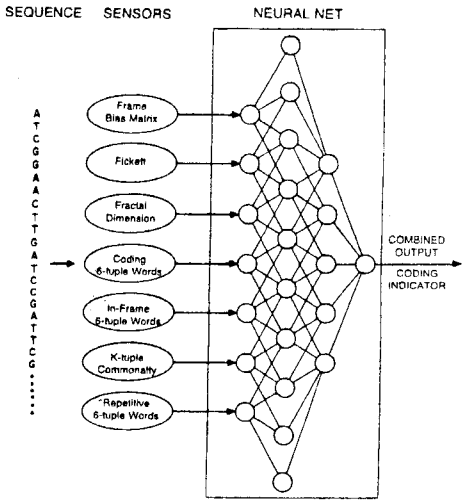

Figure 9.3. Schematic diagram of the coding recognition module. Reproduced with permission from Uberbacher and Mural (1991).

while the traditional statistical and consensus methods for gene feature recognition are often complicated by the problem of large number of false positive signals. The input to the net was the result of a series of statistical tests (sensors) that had been previously used to differentiate between the coding and non-coding regions (see Figure 9.3). The analysis was based on the concept suggested by robotic environmental sensing, wherein the perception of the robot's surrounding occurs via an integration of information that it receives from different sensors. While the sensors supply partially redundant information with varying degrees of accuracies, optimal integration, however, can be achieved with the help of machine learning. Such an estimate is better than considering the result from any of the output of any of the sensors individually. This approach has been applied to recognize the location of exons which encode protein. The environment of DNA consisting of strings of bases is sensed using different algorithms. For example, the coding recognition module (CRM) incorporates a group of seven sensor algorithms each of which gives the extent to which a given sequence is in the coding region. These sensor algorithms are based on statistical, mathematical, and linguistic principles. The sequences are analyzed over a calliper range of 100 bases (see Figure 9.4). All the seven tests are applied to each calliper and thus form one input node to the neural net. The net, while undergoing iterative training, learns to recognize the relative importance of each test and adjusts the weighting to each input sensor and those

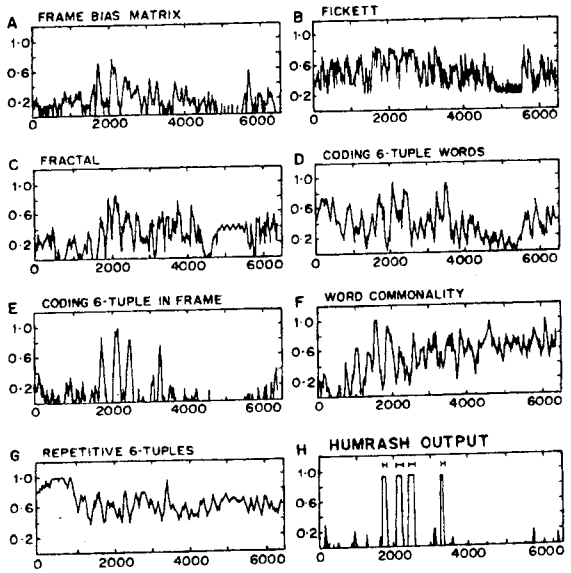

Figure 9.4. (A-G) Outputs of the seven sensor algorithm in the case of the human HRA S1 protooncogene. Panel H is the neural network output showing the predicted and actual coding regions (horizontal bars). Reproduced with permission from Uberbacher and Mural (1991).

between its internal nodes so as to recognize the relative importance of each test. The converged network when provided with sensor data from a test sequence can distinguish non-coding and coding regions of DNA with a very high degree of accuracy.

An overview of the sensors is given below (also see Uberbacher and Mural, 1991).

1. *Frame bias matrix* : It is based on the non-random frequency with which each of the four bases occupies each of the three positions within codons. This is mainly due to codon bias and the unequal usage of amino acids. This bias represented in the form of a matrix is used to probe potential coding regions and the preferred reading frame. The propensity of a region to code for a protein is based on the extent of correlation between the bias matrix and the other two reading frames. The correlation coefficient between the matrix and each reading frame is used as an indicator of the coding potential.

2. *Fickette algorithm* : This was developed by Fickette (1982) which accounts for the different properties of the coding sequences. For a window of sequence, it examines the 3-periodicity of each of the four bases and compares them to the periodic properties of the coding DNA. It further compares the overall base composition of the test DNA with that of the coding and non-coding DNA.

3. *Dinucleotide fractal dimension* : The non-randomness in the distribution of dinucleotides is indicated by the frequent occurrence of AA and TC as compared to CG. The DNA sequence may be viewed as a dynamic function based on the transition of sequential dinucleotides. In the Boltzmann sense it may be considered as a change in energy using the energy scale $E = -\ln(p)$, where p is each dinucleotide probability. These fluctuations are characterized by their fractal dimension and has been found to be lower for the coding region than for the non-coding region (Hsu and Hsu, 1990).

4. *Coding 6-tuple word preferences* : This is a statistical estimate of a nucleotide "word" of a given length in the DNA sequence (Claverie et al., 1990). Different regions have varying distribution of word occurrences. There are certain 6-base 'words' that occur much more frequently in coding regions. In the Uberbacher's words "An analogy is, if you looked at a page of an engineering text and a page of a romance novel, you could tell by looking at a few dozen random words which was which." The preference value for a word is calculated as the logarithmic ratio of its normalized frequency of occurrence in coding vs. noncoding human DNA, and the sum of preference values in the window provides a coding indicator. The 6-tuple frequencies for protein-coding DNA were compiled from known protein-coding portion of 122 cDNA sequences (210,000 nucleotides) and the information on the non-coding DNA from a data set of about 175,000 bases of sequence from human introns.

5. *In-frame 6-tuple words* : In this, the observed 6-tuples in the test DNA are compared with the preference values of in-frame 6-tuples compiled from coding DNA. The total preference is computed once for each reading frame. The predicted reading frame is taken to be the one that provides the best 6-tuple in-frame coding vs. non-coding preference, and the sensor value corresponds to the total preference for this frame.

6. *K-tuple commonalty* : The overall frequency of occurrence of a given 6-tuple in bulk DNA is related to its context. Introns use extremely common words and exons relatively rare words. The number of distinct k-tuples that can be formed are 4^k and thus there would be 4096 nucleotide 6-tuples. Further, the 6-tuple commonalty is calculated by summing all (overlapping) 6-tuple commonalties contained completely in the analysis window.

7. *Repetitive 6-tuple word preference* : The sequence under study is compared with 6-tuple statistics for several classes of repetitive DNAs. The largest total preference in the window (the best similarity to a repetitive type) is used as the sensor.

Furthering their efforts in this direction, Xu et al. (1994) have upgraded the GRAIL system to make more accurate predictions of coding region. The new version of GRAIL, GRAIL-II is responsible for the recognition of coding regions. The network architecture of GRAIL-II is shown in Figure 9.5. The factor that distinguishes the two versions of GRAIL from other approaches is that their analysis of coding region is based on longer DNA words, such as 6-mers, or dicodons, rather than on the triplets. The distinction between the coding and non-coding regions based on longer words was demonstrated by Claverie et al. (1990). The existence of a correlation between adjacent codons in exonic DNA was demonstrated in a study by Farber et al. (1992). The other distinguishing feature is the combining of multiple types of information in the exon discrimination process. The GRAIL-II system uses eleven different information elements in the exon recognition process. These elements do not provide a direct coding measure but they include the splice junction quality and properties of adjacent non-coding regions. Similar to the previous version, GRAIL-II makes use of neural networks for integrating information. The earlier version used a fixed sequence window of 100 bases in length to evaluate the coding potential of regions in DNA sequence, and a prediction efficiency of 90% was achieved. However, the difficulty was in

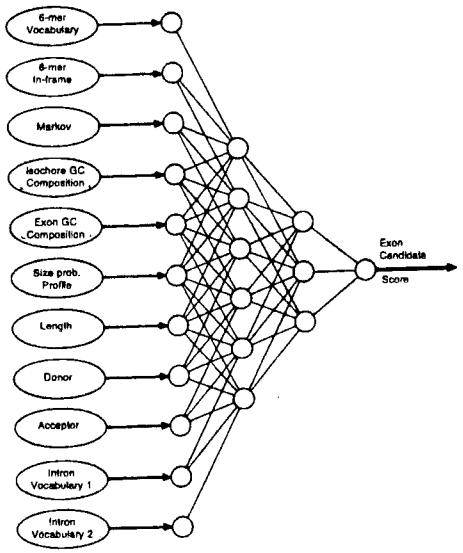

Figure 9.5. Neural network architecture of GRAIL II. Reproduced with permission from Xu et al. (1994).

locating exons that are much shorter than the window size due to the inclusion of introns. The feature of GRAIL-II is the improved performance in predicting short exons. The use of variable length windows makes it possible to consider all possible exons. GRAIL-II predicts discrete coding regions in a DNA sequence instead of a continuous coding probability function. Further, it has also achieved error-free reading frame prediction. GRAIL-II has four steps in exon recognition:

(i) generation of an initial coding pool,

(ii) eliminating improbable candidates using heuristic methods,

(iii) evaluation of exon candidates with neural network, and

(iv) clustering scored candidates.

The reader may refer to the original work for the details of the steps given above and used by GRAIL-II. The performance of GRAIL-II was markedly improved by incorporating these new strategies. It recognized 93% of all exons as compared to GRAIL-I. While GRAIL-I predicted only 50% of exons less than 100 bases in length, GRAIL-II located about 81% of them. Further, the edge error for GRAIL-II was only 14 bases in comparison to 30 bases for GRAIL-I. These predictions are further improved in the output of GRAIL-II's gene assembly program (GAP III) (Xu et. al., 1995, 1995a). GAP III is used for constructing gene models from a set of accurately-predicted exons. The program takes the set of clusters of exon candidates generated by GRAIL-II as input, and uses a dynamic programming algorithm to construct a gene model, which may be complete or partial by optimizing a predefined objective function. The gene models obtained correspond well with the structures of genes which have been determined experimentally and reported in the genome sequence databases. Discussing the details of GAP III is beyond the scope of the present chapter. GRAIL-II, however, like other methods, is sensitive to insertions and deletions. This is so because an insertion or a deletion may change the open reading frame and, hence, interrupt the reading frame of the coding region. Thus, a true exon may not be predicted as an exon due to sequencing errors that cause insertions or deletions. GRAIL-II has also been used in combination with dynamic programming approaches in correcting sequencing errors (Xu et al., 1995b).

In another interesting approach, Snyder and Stormo (1993) have also used dynamic programming in combination with neural networks to identify coding regions. Their GeneParser program scores the sequence for splice sites, codon usage, local compositional complexity, 6-tuple frequency, length distribution and periodic asymmetry. GeneParser employs the dynamic programming algorithm to find the highest scoring combination of introns and exons subject to the constraint that introns and exons must be adjacent and non-overlapping. The GeneParser correctly identified 75% of the exons and it correctly predicted 86% of coding nucleotides as coding with 13% false positives (non-exon bps being predicted as coding). The correlation coefficient for exon prediction was 0.85. Extending their work in this direction, Snyder and Stormo (1995) used the dynamic programming

algorithm to parse a sequence into an arbitrary number of classes subject to a number of grammatical constraints. In view of the fact that biological sequences consist of intrinsically discrete elements, the use of dynamic programming is very much justifiable. The dynamic programming algorithm was then applied to find the combination of introns and exons that maximizes the likelihood function. Snyder and Stormo (1995) have used this method to generate ranked suboptimal solutions, each of which is the optimum solution containing a given intron-exon function. The system performed satisfactorily on a large collection of human genes. A correlation coefficient of 0.89 for exon nucleotide prediction was achieved. For a subset of G+C-rich genes a correlation coefficient of 0.94 was achieved.

From the above discussion it would be clear that in understanding the sequence distribution associated with the coding and non-coding regions there is still a long way to go.

9.1.6. Phylogenetic Classification of Ribosomal RNA Sequence. Neural networks have been successfully employed by Wu and Shivakumar (1994) in classifying ribosomal RNA sequences according to phylogenetic relationship. The molecular sequences were encoded into neural input vectors using an *n-gram* extraction method. This was originally proposed by Cherkassky and Vassilas (1989) for associative database retrieval. The concept is similar to the k-tuple method. The algorithm extracts from the various patterns of n-consecutive residues of a sequence string, the number of occurrences of all possible letter pairs, triplets, etc. The method has the advantage that it is : (1) sequence length invariant; (2) residue insertion and deletion invariant; and (3) independent of the *a priori* recognition of certain specific patterns. Further, the singular value decomposition method has been used to reduce the size of the input vector. A three-layered feedforward neural network that has been employed uses supervised learning paradigms, involving back-propagation and a modified counterpropagation algorithm. The neural net was trained to classify new sequences into predefined classes with the data from RDP (Ribosomal Database Project) database. Each rRNA sequence entering in RDP was fully aligned and the percentage similarity was converted to an evolutionary distance which was then used to place the entry on a phylogenetic tree. The prediction accuracy of the neural network was expressed as : (a) the total number of correct patterns (true positives), (b) the total number of incorrect patterns (false positives), and (c) the total number of unidentified patterns (negatives). After training, the network was able to classify query sequences into more than one hundred phylogenetic classes with 100% accuracy, at the rate of less than 0.3 CPU seconds per sequence on a workstation. The performance of the network was an order of magnitude faster than other methods like Similarity Rank, Blast and Fasta.

9.1.7. Analysis of Transcription Control Signals of a Eukaryotic Protein Coding Gene. Nair et al. (1995) have made use of the mutation data available from wet labs in building a neural network capable of predicting the transcriptional levels based on the sequence of the upstream region. The neural net has been trained using

mutation data taken from the studies carried out by Myers et al. (1986), wherein saturation mutagenesis (Myers et al., 1985) was used to introduce random single base substitutions into the mouse β-globin promoter region. The effects of single base substitutions in the β-globin promoter were determined by comparing the levels of correctly initiated RNA derived from the test and reference plasmids co-transfected into HeLa cells and expressed as the relative transcription level (RTL) of each mutant. From a total of 129 mutants obtained, with mutations between -101 and +20, 117 were used as the network training data set and the remaining 12 as test data set. A network with 484 neurons in the input layer, 8 neurons in the hidden layer, and 1 neuron in the output layer gave optimal results. Using the model as a heuristic device, Nair and coworkers have simulated all possible single base mutations associated with the upstream region of the globin gene. The simulation results have helped in identifying sequence elements within the conserved region which when mutated did not seem to affect the transcriptional levels (Nair et al., 1995). However, it should be borne in mind that the simulation results could not be validated because of the non-availability of mutation data. Nevertheless, the results can be used as a guide in designing mutation experiments since an *a priori* estimate of the possible outcome of a mutation can be obtained.

Since most of the controls of gene expression are effected at the level of transcription, an insight into the process of transcription would help in understanding the process of gene expression as a whole. Further, the analysis of transcription control signals in predicting the rate of mRNA synthesized also underscores the need to build a mutational database (Nair and Kulkarni, 1996). The database would serve as a reservoir of information which would assist a theoretical analyst to build black-box/empirical models. Pooling up of information will also help in validating simulation results. Moreover, development of more realistic models will be possible only if sufficient data is made available. These highly developed models could then be applied to important systems such as the globin gene in understanding hitherto unknown genetic disorders caused by a loss of transcription control signals.

In yet another approach, Larsen and coworkers (1995) have analyzed eukaryotic promoter sequences using neural networks in combination with a variety of other tools like Shannon information (Shannon and Weaver, 1971), multiple alignment and weight matrix methods, and revealed the presence of a systematically occurring C-T signal. Neural networks were used to predict the exact location of the transcription initiation site in a data set consisting of 481 promoter regions from mammals. The input layer was presented with 71nt. The region was set to [-250; 250], [-150; 150], [-100; 100], and [-60;60]. The number of neurons in the hidden layer was varied from 0 to 100. The best result was 27.2 and 44.6% recognition of true sites with 1.0 and 4.9% false positive rate. Further analysis of the result using networks and the Shannon information measure showed three important regions, viz. the TATA box region [-30;-24], the cap signal [-1;0] and a new region [+5; +10]. Using the multiple alignment method, the consensus of this

new region (CT signal) obtained was CTNCNG. The results show how important information can be captured by using neural networks in combination with other analytical techniques.

9.2. PREDICTION OF PROTEIN STRUCTURAL FEATURES.

Another very important area where ANNs have found widespread applications is the prediction of protein structural features. Proteins are composed of linear chains of polypeptides with amino acids as their monomers. Accurate prediction of the secondary and tertiary structures from their amino acid sequences which constitute the primary structure has been one of the dreams of structural biologists. The secondary structure is the local spatial organization of the polypeptide backbone without taking into consideration the conformation of its side chains. Thus α-helix, β-sheet, and β-turn belong to this class. The tertiary structure is the arrangement of all the atoms in space, which includes the disulfide bridges and the position of the side chains, taking into consideration all short- and long-range interactions. All the attempts at predicting protein structure are based on the tacit assumption that there exists adequate information in the amino acid which would finally enable the determination of the 3-D structure. The assumption owes its basis to the ability of a sequence to return to its native conformation in an environment devoid of other protein synthesizing components.

With the advent of newer approaches in the elucidation of structure, there has been an exponential growth in the database of atomic resolution protein structures which were earlier observed by X-ray crystallography and recently by NMR. Along with the growth of the database, there has been an increase in the variety of theoretical methods for analyzing the growing wealth of information. Studies have ranged from the correlation of local, secondary structural features to long-range, tertiary structural interactions. The importance of predicting the three-dimensional structure of a protein stems from the fact that the function of an enzyme *in vivo* and *in vitro* is a direct consequence of its folded structure and its physical properties. Further, with the amount of effort going into the prediction of genes from the giga bases of DNA sequences, generated by the genome projects, it is also important to understand the structural type of the protein that is encoded by these genes and find out whether the protein is related to one of the known functions or structures. Thus, understanding the relationship between structure and function in these biological macromolecules would be of paramount importance in modern medical applications, including therapeutics and diagnostics. The most interesting practical use has been the prediction of antigenic oligopeptides as potential vaccines.

9.2.1. Prediction of Protein Secondary Structure. Earlier approaches towards predicting the secondary structure were made empirically with the use of secondary structures observed in proteins as a basis to formulate statistical and rule-based algorithms. Most of the work on protein structure prediction using artificial neural network started after the work by Qian and Sejnowski (1988). Holley and Karplus (1989) and Bohr et al. (1988) have also used similar networks but different architectures in analyzing protein structural features. In the case of the Qian and

Sejnowski network, the input to the net was an amino acid sequence window of length 13, and the output of the net was the corresponding secondary structure of the amino acid at the center of the window. The definition of the secondary structure was as defined by Kabsch and Sander (1983). They obtained a prediction accuracy of 64.3% on the three types of secondary structures: α-helix, β-sheet, and coil, with correlation coefficients, $C_\alpha = 0.41$, $C_\beta = 0.31$, and $C_{coil} = 0.41$. The Qian and Sejnowski net has been used to predict the secondary structure of the catalytic domain of C-H-ras oncogene protein. The network results have been compared with the prediction by Chou and Fasman (Holbrook, 1993). The Holley and Karplus (1989) network gave a prediction accuracy of 63% for the three states: helix, sheet and coil. In this case also the secondary structure assignments used were based on the method of Kabsch and Sander (1983). The input layer was a moving window of size 17 in the amino acid sequence and the predictions were made for the central amino acid sequence. A three-layer neural network was developed by Bohr et al. (1988) which predicted the transmembrane α-helices in rhodopsin with an accuracy of 73%. A comparative evaluation of the prediction with Argos et al. (1982) using the Chou and Fasman method (Chou and Fasman, 1974; 1978) has also been carried out. The prediction accuracy for the amino-terminal region was more accurate probably due to the lack of influence of the rest of the structure on its folding.

Andreassen et al. (1990) have used neural networks to predict the secondary structure of HIV proteins p17, gp41, and gp120. They have also compared their results with the traditional approaches. Stolorz et al. (1992) give a comparative performance of neural network methods and Bayesian statistical methods in the prediction of the secondary structure of proteins. The Bayesian method makes the assumption that the probability of an amino acid occurring in each position in the protein is independent of the amino acid occurring elsewhere. The neural network architecture was of the perceptron type (no hidden layers). While the predictive accuracy of the networks increases with the addition of hidden units, only a marginal improvement was seen for the secondary structure problem (Qian and Sejnowski, 1988). The use of hidden units improved the overall prediction accuracy by only 0.3%, at the expense of greater computational time. Their results essentially point to the fact that the predictive accuracy of the Bayesian method is only marginally lower than other sophisticated methods. Muskal and Kim (1992) have used two neural networks placed in tandem to predict the secondary structure content of water-soluble globular proteins. The first network predicts the helix and strand contents of the protein based on the amino acid composition, molecular weight and heme presence. In order to overcome the problem of generalization encountered by the first net, a second network was designed which determined the generalization state of the first one. The combined networks gave predictions with errors as low as 5 and 5.6% for helix and strand content, respectively. A comparative study of performance of other methods as against that of the tandem network revealed that the latter performed better than the other methods. This network was also applied to the prediction of secondary structure composition of the c-H-ras catalytic domain from its amino acid composition, which was shown to

be helix, 39%; strand, 21%; and coil, 40%. These are in good agreement with the observed values of helix, 25%; strand, 25%; and coil, 40%.

Kneller et al., have used separate networks on different protein classes to improve the prediction capabilities. They have compiled a database of all α, all β, and α / β proteins (Kneller et al., 1990). A general database slightly different from the Qian and Sejnowski (1988) was compiled and the performance evaluated. An average accuracy of 63% was obtained on the testing subset of this database. The obtained correlation coefficients, $C_\alpha = 0.35$, $C_\beta = 0.30$, and $C_{coil} = 0.41$, are similar to those in other studies. The 105 proteins in the database were divided into 22 all α, 24 all β, and 20 α / β proteins and were trained and tested by a jackknife (leave-one-out) procedure. Prediction accuracies of 79, 70 and 64% were obtained for all α, all β, and α / β proteins, respectively. In comparison to the Qian and Sejnowski's results, the results for all α and all β proteins were much better (Kneller et al., 1990). With a view to improving the prediction accuracy, the authors incorporated additional spatial nodes. These nodes described the helix and strand hydrophobicity moment, charge pairing due to oppositely charged residues at positions i and $i+4$, and average hydrophobicity over a window. The results did not offer any substantial improvements in the prediction capabilities of the network. However, based on the slight improvements of helix prediction, these nodes were retained in their all α -predictions.

The β-turns form a third type of secondary structural element apart from the α-helix and β-strands. β-turns (also known as *reverse turns, tight turns, β-bends, hairpin bends, 3_{10} kinks, wedgets,* etc.) are the most prevalent type of the non-repetitive structures that have been recognized. These were first recognized from a theoretical conformational analysis by Venkatachalam (1968). β-turns are localized over four residues, and, based on the hydrogen bonding, they are of 7 to 8 different classes. Mostly, all β-turns are either Type I or Type II, which close the turn by a hydrogen bond between the carbonyl oxygen of residue i and the amino of residue $i+3$. Type I turns have approximately, $\phi_2 = -60°, \psi_2 = -30°, \phi_3 = -90°, \psi_3 = 0°$; and for Type II, $\phi_2 = -60°, \psi_2 = -120°, \phi_3 = -90°, \psi_3 = 0°$. For further details the reader may refer the extensive review on protein conformation by Richardson (1981). A neural network has been developed to recognize the Type I, Type II, non-specific β-turn and a non-turn (McGregor et al., 1989; 1990). The input to the net was the four residue amino acid sequence involved in various types of turns as well as the four residue sequence in non-turn regions. The network architecture was 80 input neurons, 8 hidden neurons and 4 output neurons. The network predictions were better than the predictions of turns by the Chou-Fasman technique. The perceptron network gave a prediction accuracy of 26.1%, while the addition of 8 hidden neurons resulted in a prediction accuracy of 28.7%. The performance of the perceptron was better (56.2%) in comparison to the network with eight hidden nodes (53.8%) when a leverage of one or two residues is given while predicting the turns. On applying this method to the p21-ras protein a single β-turn was predicted.

Sasagawa and Tajima (1993, 1993a) have used neural networks with a modular architecture to predict the secondary structure of proteins. Globular proteins whose secondary structures were determined by Kabsch and Sander's method were used for training the network. The input to the network comprised twenty amino acids, B, X, Z and chain breaks. Here, B means that an amino acid is aspartic acid or asparagine, X is the undetermined amino acid, and Z refers to glutamic acid or glutamine. These were represented as a 24-bit binary input with only one bit put 'on' in each case while the other bits are 0. A neural network with modular architecture has n-1 modules if the secondary structure has n states (n = 3, 4, 8). Each module is a three-layered neural network and has one output unit. The network learns whether the secondary structure is the ith state of secondary structure or not ($1 \le i \le n - 1$). The final results of learning are obtained by combining the outputs of all modules with certain rules in the unification unit (Figure 9.6). The prediction accuracies of the network with a modular architecture were higher than that of the ordinary neural network (Sasagawa, 1993).

Dalmas et al. (1994) used ANNs to predict the secondary structure from circular dichroism (CD) spectra. ANNs have been used as an alternative to the statistical linear techniques for secondary structure prediction from CD spectra. Dalmas and coworkers used two networks, a back-propagation network for optimizing parameters and a combination of self-organizing map (SOM) and a back-propagation network (SOMBPN). SOMBPN is a hybrid network that uses principles of both the unsupervised and supervised learning paradigms, where the first layer of network is allowed to mature through a self-organization process. This enables the processing elements in the second layer to be sensitive to only a small region in the input space. The training data set consisted of CD spectra of 21 proteins and one polypeptide measured in units of differential absorption coefficient for left and right circularly polarized light, in the wavelength range 178 to 260 nm.

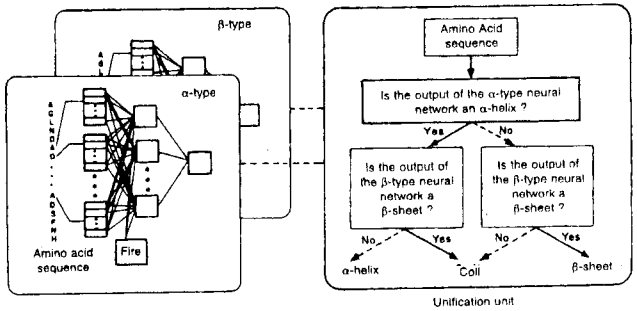

Figure 9.6. Neural network with modular architecture and the unification unit. Reproduced with permission from Sasagawa and Tajima (1993a).

digitized at 0.5 nm intervals. The simulation results point to the fact that proteins with high α -helical content fared better than the others. They obtained the best results for haemoglobin, myoglobin and glyceraldehyde-3-phosphate dehydrogenase.

9.2.2. Prediction of Protein Tertiary Structure. Bohr et al. (1990) have used neural networks for predicting the 3D-structure of protein backbones. The network was trained on a class of functionally but not structurally homologous proteins. The training data set consisted of sequences and distance matrices of 13 proteases, including trypsin and subtilisin. The distance matrix was obtained by plotting protein sequence along both the vertical and horizontal axes and points placed on the graph to indicate where two C_α positions are within a specified cluster in the three-dimensional structure. Tertiary structural information generated by the network, in the form of binary distance constraints for the C_α atoms in the protein backbone, was used to obtain a folded conformation of the protein backbone.

Wilcox et al. (1990) have used neural nets to predict secondary and tertiary structures from amino acid sequences. The training space consisted of 15 proteins of 140 residues which included some homologous proteins as well. The network input was in the form of normalized (range -1 to +1) hydrophobicity values as given in Liebman et al. (1985). The network with 140 input units, one hidden layer containing 15-240 units and an output layer representing the window of distance matrices of 19600 (140 x 140) units performed well on the training set. The generalization capability of the network was poor, which may be due to the heterogeneity of the training space.

9.2.3. Other Applications in Identification of Protein Structural Features. Neural networks have also been used in understanding many other aspects of protein structural features. Dubchack et al. (1993) have used feedforward networks in discriminating protein folding patterns. The input to the network was the amino acid percent composition of the protein. The network architectures with one to four output neurons, each representing a protein folding pattern were tested. The network with a single output distinguished a particular fold from among the others, while the network with two, three or four outputs was useful in distinguishing the fold of interest from among the three folds. The four classes tested and their prediction accuracies are given below.

4 – α helical bundle	81%
Eight-stranded parallel α / β barrel	85%
Nucleotide binding or Rossmann fold	91%
Immunoglobulin fold	78%

It is important to underline the fact that, although statistical analysis did not reveal differences in the amino acid percentage, the networks were able to distinguish between folding types. The network weights were analyzed to determine

the amino acids important in the formation of the specific folding patterns. The main difficulty in understanding protein folding problems was the lack of known 3-D structure of proteins. This problem was circumvented by assigning similar folding pattern to proteins of the same family showing high homology with proteins of known structure. The network trained on these assumptions correctly predicted the folding class of C-H ras oncogene.

The role of cysteine in imparting stability to proteins by forming disulfide bonds has lead to attempts to predict the disulfide bonding patterns using an artificial neural network. Muskal et al. (1990) have developed a neural net which was trained to differentiate between a cysteine involved in a disulfide bond and that existing as a free sulfhydryl. The input to the network was the amino acid sequence surrounding (but not including) the cysteine of interest. This was based on the assumption that the propensity or aversion of cysteine to form disulfide bonds depends on the neighboring sequences. The network predicted the presence of a disulfide bond with an accuracy of 81.4% and the free cysteines with an accuracy of 80%. The three cysteines of the C-H-ras p21 protein not involved in disulfide bonding were correctly predicted by this method.

Exposure of different residues on the surface of a protein is an important aspect determining the interaction with other macromolecules and ligands, antigenic determinants, etc. These regions also form potential mutation sites. It is possible to determine surface accessibility from the three-dimensional structure of a protein. However, in the absence of a known structure, the intrinsic hydrophobicity of the amino acid is used to draw conclusions about surface accessibility and exposure to solvent. Holbrook et al. (1990) have developed a neural network by using high resolution protein crystal structure during the training process, to predict (i) whether a residue is buried or exposed; (ii) whether a residue is completely buried, partially exposed or completely exposed; and (iii) the actual exposure of the residues. The network input consisted of the amino acids whose exposure was in question and its flanking sequence was on both sides. The results of the predictions show that the influence of the surrounding residues is quite small since 72% correct prediction was obtained using a window size of 9 residues as against 70% prediction obtained using only the central residue. This approach has been used to calculate the surface accessibility for the residues of the C-H-ras protein and the predictions have been compared with the fractional accessibilities determined from the crystal structure.

Networks have also been used in clustering proteins into families. Ferran and Ferrara (1992) have used a Kohonen network, trained by an unsupervised learning method, to cluster proteins into families. The network was composed of 7x7 neurons and it used as input matrix patterns derived from the bipeptide composition of 447 proteins belonging to 13 different families. During training, the network organized the activation of its neurons into topologically ordered maps, into which the proteins were correctly clustered into the corresponding families. A prediction

accuracy of 96.7% was achieved. Furthering their efforts, Ferran and Pflugfelder (1993) have proposed a new hybrid method which used both statistical and ANN approaches. The statistical methods are used to cluster a set of bipeptidic matrices into families in three main stages: (i) principal component analysis ; (ii) determination of the optimal number of M clusters ; and (iii) classification of the bipeptidic matrices into M clusters. In this kind of analysis, the results of the statistical method are used to choose the number of neurons and inputs of the network. Such networks which are fed with a few principal components of the bipeptidic matrices are easy to train and faster convergence is achieved.

In yet another important application, Wu et al. (1992) and Wu (1993) used neural networks for protein classification. The protein classification artificial neural system (ProCANS) was developed by them for rapid superfamily classification of unknown proteins. Its sequence encoding scheme involved an n-gram hashing function. There were several advantages of using this method since it is invariant of sequence length as well as insertion and deletion of a residue. It is also independent of *a priori* recognition of certain specific patterns. The network architecture used in developing ProCANS was a set of independent multiple modules of a three-layered feedforward neural network that employs the back-propagation learning algorithm (Figure 9.7). The architecture involves two levels of modularization, viz. database modularization and encoding modularization. The training set derived from the PIR (protein identification resource) protein sequence database was broken down into multiple sets according to functional groups. Each set was trained by a separate

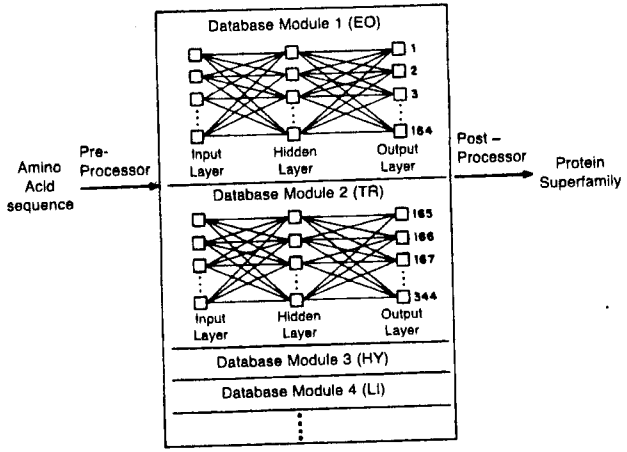

Figure 9.7. ProCANS network architecture. Reproduced with permission from Wu (1993).

module called *database module*. In all, the current ProCANS had four neural network modules developed with seven protein functional groups, which consisted of 690 superfamilies and 2724 entries of the annotated PIR database (PIR1, Release 32.0, 31 March 1992). For details of the PIR protein entries used to develop ProCANS, the reader may see Wu (1993). The configuration of the four networks was made up of 462 input units, 200 hidden units and 164, 180, 192 and 154 output units. The outputs represented the number of superfamilies represented in each network module.

Xin et al. (1992) used back-propagation neural networks for prediction of protein folding patterns. A very large network was employed which was made to learn the entire distance matrices that encode tertiary as well as secondary structure elements. The distance matrix consisted of a square array whose side was the number of residues in a protein and whose elements contained the distance in Euclidean space from the ith to the jth residue in the protein. Their approach was mainly directed towards associating selected physical and chemical properties of amino acids (mostly amino acid hydrophobicity) with the three-dimensional structure of protein observed by crystallography. A set of 20 proteins representing seven families were used in the training phase, and the generalization capability of the net was tested by a set of homologous proteins representing 4 of these families. The network (Bignet) was able to capture the 3-D representation of a large variety of structural features. Recreation of the embedded three-dimensional Euclidean structure (Wilcox et al., 1990; Liebman et al., 1985) was performed on the basis of distance matrices predicted by the Bignet. The analyses presented are a pointer to the fact that residue hydrophobicity is a good approach to describe sequence data, and further that distance matrices can be learned to an accuracy of > 99%.

Chen (1992) has developed a geometric learning system (GLS) with a view to providing a basis for representing, characterizing and learning the essential geometric structures of proteins in the human genome. Using the existing database of proteins structures (the Brookhaven protein data bank) as a sample space for learning, a mapping from amino acid sequence to protein tertiary structures was implemented for learning and predicting new protein structures. During the learning phase, GLS determined a set of invariant geometric feature vectors that characterize the tertiary structure. The GLS was implemented on three levels :

(i) learning of the invariant geometric feature vectors from amino acid sequence using neural networks,

(ii) using the clustering algorithm to determine the secondary structures from invariant geometric feature vectors, and

(iii) using the clustering algorithm to determine the domain structure from the secondary structure.

Their prediction result pointed to the inadequacy of the existing database.

Lohmann et al. (1994) used neural networks in the prediction of membrane-spanning amino acid sequences. Earlier approaches to predict membrane spanning regions in amino acids were based on hydrophobicity profiles (Kyte and Doolittle, 1982; Vogel et al., 1985). This approach has met with limited success (Esposti et al., 1990) with respect to the prediction accuracy. The investigators have used a set of 7 property scales of amino acids normalized to (-1,1) intervals in representing amino acid sequence data as follows:

(i) Hydrophobicity (Engelman et al., 1986)
(ii) Hydrophilicity (Hopp and Woods, 1981)
(iii) Polarity (Jones, 1975)
(iv) Volume (Zamyatnin, 1972)
(v) Surface area (Chothia, 1976)
(vi) Bulkiness (Jones, 1975)
(vii) Refractivity (Jones, 1975).

Out of the total 47 human integral membrane proteins, 36 were used as training data set and the remaining 11 as test sequences. The training and test data sets also contained negative examples (non-transition regions). The ratio of positive (transition regions) to negative examples was 1:4. The prediction accuracy of the network was about 92.3%. The prediction power of the network was attributed to the coding of amino acids in terms of physico-chemical properties and the inductive training techniques which were part of the structure evolution. Lohmann and coworkers conclude that the features of membrane transition regions extracted during the network development seem to be mainly based on the characteristic distribution of charges, potentials and dipoles, and the geometric properties (volume, surface area, bulkiness) are less important.

There are also a variety of other applications where neural networks have been employed in understanding different aspects of protein structural features. Schneider and Wrede (1994) developed artificial neural filters for pattern recognition in protein sequences. The networks had a feedforward architecture and provide adaptive neural filter systems for pattern recognition in primary structures and sequence classification. The prediction accuracy of the network on an independent test set of sequences was 97%. The network architecture appears to be most suited for the analysis of *E. coli* signal peptidase cleavage sites. Neural networks have also been used, in combination with simulated molecular evolution, for sequence-oriented protein design (Schneider et al., 1994; 1995). In another interesting application, ANNs were used to analyze and predict the human immunodeficiency virus type I reverse transcriptase inhibitor. The training and control sets included 44 molecules (Tetko et al., 1994). The trained network was used to predict the activities of 20 new molecules. The prognosis of new molecules revealed one molecule as possibly very active and this was further confirmed by biological tests. Further, ANNs have also found application in peptide sequencing

from high energy CID spectra (Scarberry and Zhang, 1996) and prediction of the zinc finger DNA binding protein (Nakata, 1995).

Knowledge-based neural networks have been used to improve algorithms (Maclin and Shavlik, 1993). This approach has been applied to improve the Chou and Fasman algorithm (Chou and Fasman, 1978). The multistrategy approach of KBANN has lead to a statistically-significant and more accurate solution than both the original Chou-Fasman algorithm and a neural network trained by the standard approach.

9.3. APPLICATIONS OF FEEDFORWARD NETWORKS TO OTHER AREAS IN BIOLOGY.

Apart from the use of feedforward nets in analyzing protein and nucleic acid sequences, there has been innumerable applications of networks in other branches of biology where conventional modeling is not possible. This section would take stock of these applications. We have tried to cover as many applications as possible; however, some of the applications might have been inadvertently missed out.

Back-propagation networks have been used to simulate the response properties of posterior parietal neurons. In an interesting application, Zipser and Andersen (1988) have developed a network model to decode the spatial information from area 7a neuron (which contains the visual and eye position neurons) of the posterior parietal cortex of monkeys and to account for their observed response properties. The neural network model was able to show how eye position independent location can be extracted from a population of 7a neurons. The model was also able to reproduce the nonlinear interactions of eye position and retinal position information seen in actual area 7a neurons. The data for training the neural net was collected from the earlier studies, carried out on awake, unanaesthetized monkeys. A three-layered network was trained to map visual targets to head-centered coordinates, given any arbitrary pair of eye and retinal positions. The visual input consisted of 64 Gaussian-shaped receptive fields, with 1/*e* widths of 15 degrees with each peak separated by 10 degrees in an 8 x 8 array. The eye position input consisted of four sets of 8 units with single sets for positive and negative slope for horizontal and vertical eye positions. Two representations of location in head-centered coordinates at the output layer were used.

1. *A Gaussian format* : In this, each unit had a Gaussian receptive field similar to the representation of the retinal input, but the coding location is in head-centered rather than retinal co-ordinates.

2. *The monotonic format* : In this case, the activity of each neuron is a linear function of the location of the stimulus in head-centered coordinates.

The network was trained using randomly selected pairs of input eye positions and retinal position. The output of the network was the true spatial location in head-

centered coordinates implied by the inputs. The output was represented in either the monotonic or Gaussian format. The converged network was able to generate retinal receptor fields remarkably similar to the experimentally observed fields.

Neural networks have also found widespread applications in the area of medicine. One of them is the application of neural computing in cancer drug development. Weinstein et al. (1992), have developed a neural network model capable of predicting the mechanism of action of a drug from its pattern of activity against a panel of 60 malignant cell lines in the U.S. National Cancer Institute's drug screening program. The input to the net was drug's pattern of activity in the cell line and the output was the category to which the drug belonged. The simulation results reveal that in assigning six possible classes of mechanism, the network missed only 12 out of 141 agents (8.5%). The network performance has been reported to be better than linear discriminant analysis, which missed 20 out of 141 (14.2%). Apart from this, neural networks have also been used in the early detection and diagnosis of cancer (Rogers et al., 1994). Snow et al. (1994) have used neural networks in the diagnosis and prognosis of prostrate cancer. The neural network predicted the biopsy results with 87% overall accuracy. Further, it also predicted tumor recurrence with 90% overall accuracy. Other applications in the area of medicine include the estimation and prediction of myocardial infarction size (Mandal et al., 1994). Baxt (1990) used ANN for the diagnosis of myocardial infarction. The converged network had a diagnostic sensitivity of 97% and a specificity of 96.2%. Neural networks have also been used in developing predictive models for growth (Yee et al., 1993). For a short review of the applications of EBP networks to medicine and related areas, the reader may refer Erb (1995).

Neumann et al. (1992) have developed a self-learning system to analyze and classify courtship songs of *Drosophila* males. Courtship songs produced by wildtype *Drosophila* males as well as the cacophony and dissonance behavioral traits were studied for adaptive acoustic analysis and classification. The self-learning system developed by them used several techniques involving feedforward networks, learning vector quantization of signals and nonlinear adaptation of data analysis. A three-layered feedforward network was used to discriminate among the song types. The network was used to map song-pulse NAVs (node activation vectors) which formed the input space on the correct genotype (output). A genotype was ideally represented by an activation value of 1 for the corresponding output unit and a value of 0 for the remaining output units. A comparative performance of genotype prediction using ANN and multivariate analysis technique showed ANNs to be a much better tool.

Goodacre et al. (1994) have compared the extent of production of recombinant cytochrome b5 in pλ-ncyt (obtained spectrophotometrically) with that assessed by pyrolysis mass spectrometry (Pyms) using artificial neural networks. The training data for the network consisted of normalized triplicate pyrolysis mass spectra derived from *E. coli* pλ-Ocyt, pλ-2cyt, and pλ-5cyt, and the outputs were

the actual (true) amount of cytochrome b5 as determined spectrophotometrically and expressed as a percentage of the total protein. The result of the predictions show that ANNs can be used to quantify the amount of mammalian cytochrome b5 expressed in *E. coli*.

Neural nets were also used in the classification of species by training networks using digitized images. Simpson et al. (1991 ; 1992 ; 1993) and Williams et al. (1993) used networks for recognizing phytoplankton. Following their work Culverhouse et al. (1994) used neural networks in the categorization of five species of *Cymatocylis* (protozoa, Tintinnida). Photomycrographs of five species of *Cymatocylis* were digitized, converted into binary and edited by hand to remove large debris that distorted the image. This data was further preprocessed by Fourier transformation before using it for training. The network was trained to categorize 201 of these specimens. The simulation results revealed 70% correct categorization of the data used in the training set. The optimal network performed with an error rate of 11%.

9.4. CONCLUSION.

In this chapter, an attempt has been made to give an overview of the different applications of neural networks in biology. The utility of neural network approach can be assessed from the studies that rigorously compare the neural network performance with other statistical methods. However, it becomes difficult to draw a very fine conclusion from these studies as there are many empirical parameters like window size, number of hidden units, learning space, and convergence definition that need to be optimized. Neural networks have been more widely applied in nucleic acid sequencing than in protein sequencing. This may be due to the availability of information on a large number of nucleic acid sequences, and the fact that bases of DNA can be encoded in a 4-bit binary while amino acids require a 20-bit binary.

The KBANN approach seems to be an attractive proposition. This is mainly because the inference rules about the problem need to be only approximately correct. Further, it has also been used in extracting rules and in refining algorithms. The approach of Uberbacher and Mural (1991) suggests an indirect analysis which may be more successful than the direct one. The calliper randomizarion approach used by Nair et al. (1994) presents an alternative methodology to identify regions in DNA that are crucial to the process of recognition. The method can be used to arrive at the consensus of a region involved in common biological function. Putting an ANN to work on mutation data to capture the transcription control signals (Nair et al., 1995) is one of the important ANN applications. The connotation of this work is its direct application to understand the unknown genetic disorders.

The take-home message is that data representation plays a very important role in network analysis. Extracting information from data using different analytical techniques helps to analyze different faces of the same data. Using outputs of these as inputs to the network, greatly helps to reduce the network-size (especially in case

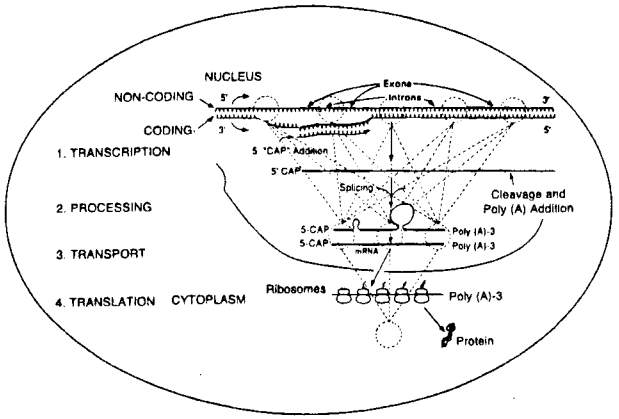

Figure 9.8. A possible perspective of capturing the whole process of gene expression in a neural network.

of sequence analysis). What the future holds in store for artificial neural networks is yet to be seen and would greatly depend on how we extend our imagination. The ultimate aim is to pick out genes from sequences, predict their expression levels and finally the structure of the proteins that they encode. A possible perspective of this is given in Figure 9.8. It goes without saying that all this must be backed up with adequate experimental evidence.

"Models are to be used but not to be believed"
Henry Theil

References

Andreassen, H., H. Bohr, J. Bohr, S. Brunak, T. Bugge, R. Cotterill, C. Jacobsen, P. Kusk, B. Lautrup, S.B. Petersen, T. Saermark and K. Ulrich (1990). Analysis of the secondary structure of the human immunodeficiency virus (HIV) proteins p17, gp120, and gp41 by computer modeling based on neural network methods. *J. AIDS*, **3**, 615-622.

Argos, P., J. Rao, and P. Hargrave, (1982). Structural prediction of membrane bound proteins. *Eur. J. Biochem.*, **128**, 565-575.

Baxt, W. (1990). Use of an artificial neural network for data analysis in clinical decision-making : The diagnosis of acute coronary occlusion. *Neural Computation*, **2**, 480-489.

Baxt, W. (1991). Use of an artificial neural network for the diagnosis of myocardial infraction. *Annals of Int. Med.*, **115**, 843-848.

Bisant, D. and J. Maizel (1995). Identification of ribosome binding sites in Escherichia coli using neural network models. *Nucl. Acids Res.*, **23**, 1632-1639.

Bohr, H., J. Bohr, S. Brunak, R. Cotterill, B. Lautrup, L. Norskov, O. Olsen and S. Petersen (1988). Protein secondary structure and homology by neural networks. *FEBS Lett.*, **241**, 223-228.

Bohr, H., J. Bohr, S. Brunak, R. Cotterill, H. Fredholm, B. Lautrup, and S. Petersen (1990). A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS. Lett.*, **261**, 43-46.

Breathnach, R. and P. Chambon (1981). Organization and expression of eukaryotic split genes coding for proteins. *Ann. Rev. Biochem.*, **50**, 349-383.

Brendel, V. and E. Trifonov (1984). A computer algorithm for testing potential prokaryotic terminators. *Nucl. Acids Res.*, **12**, 4411-4427.

Brendel, V., H. Hamm, and E. Trifonov (1986). Terminators of transcription with RNA polymerase from *E. coli* : What they look like and how to find them. *J. Biomol. Struct. Dyn.*,**3**, 705-723.

Brunak, S., J. Engelbrecht, and S. Knudsen (1990). Neural network detects errors in the assignment of mRNA splice sites. *Nucl. Acids Res.*, **18**, 4797-4801.

Brunak, S., J. Engelbrecht, and S. Knudsen (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, **220**, 49-65.

Burns, J. and G. Whitesides (1993). Feedforward neural networks in chemistry : Mathematical systems for classification and pattern recognition. *Chem. Rev.*, **93**, 2583-2601

Calladine, C. (1982). Mechanics of sequence-dependent stacking of bases in B-DNA. *J. Mol. Biol.*, **161**, 343-352.

Chen, S. (1992). Characterizing and learning of protein conformation. In *Proceedings of the Second International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*, [Eds.] H. Lim, J. Fickette, C. Cantor and R. Robbins, pp. 391-399.

Cherkassky, V. and N. Vassilas (1989). Performance of back-propagation network in associative database retrieval. *Proc. Int. Joint Conf. Neural Networks*, **1**, 77-83.

Chothia, C. (1976). The nature of accessible and buried surfaces in proteins. *J. Mol. Biol.*, **105**, 1-14.

Chou, P. and G. Fasman (1974). Conformational parameters for amino acids in helical, β-sheet and random coil regions calculated from proteins. *Biochemistry*, **13**, 211-222.

Chou, P. and G. Fasman (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology*, **47**, 45-148.

Claverie, J-M., I. Sauvaget, and L. Bougueleret (1990). k-Tuple frequency analysis from intron/exon discrimination to T-cell epitope mapping. *Methods in Enzymology*, **183**, 237-252.

Coleman, J., M. Inouye, and K Nakamura (1985). Mutations upstream of the ribosome-binding site affect translational efficiency. *J. Mol. Biol.*, **181**, 139-143.

Cosic, I., D. Nesic, M. Pavlovic, and R. Williams (1986). Enhancer binding proteins predicted by informational spectrum method. *Biochem. Biophys. Res. Comm.*, **141**, 831-834.

Cosic, I. and D. Nesic (1987). Prediction of hot spots in SV40 enhancer and relation with experimental data. *Eur. J. Biochem.*, **170**, 247-252.

Culverhouse, P., R. Ellis, R. Simpson, R. Williams, R. Pierce, and J. Turner (1994). Automatic categorization of five species of cymatocylis (Protozoa, Tintinnida) by artificial neural network. *Marine Ecol. Prog. Series*, **107**, 273-280.

Dalmas, B., G. Hunter, and W. Bannister (1994). Prediction of protein secondary structure from circular dichroism spectra using artificial neural network techniques. *Biochem. Mol. Biol. Int.*, **34**, 17-26.

Davis, M., R. Simons, and N. Kleckner (1985). Tn 10 protects itself at two levels from fortuitous activation by external promoters. *Cell*, **43**, 379-387.

Demeler, B. and G. Zhou (1991). Neural network optimization for *E. coli* promoter recognition. *Nucl. Acids Res.*, **19**, 1593-1599.

Dickerson, R. (1983). Base sequence and helix structure variation in B and A DNA. *J. Mol. Biol.*, **166**, 419-441.

Dubchack, I., S. Holbrook, and S-H. Kim (1993). Prediction of protein folding class from amino acid composition. *Proteins : Struct. Funct. Genet.*, **16**, 79-91.

Engelman, D., T. Steitz, A. Goldman (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophyschem.*, **15**, 321-353.

Erb, R. (1995). The backpropagation neural network-A Bayesian classifier : Introduction and applicability to pharmacokinetics. *Clin. Pharmacokinet.*, **29**, 69-79.

Esposti, D., M. Crimi, G. Venturoli (1990). A critical evaluation of the hydropathy profile of membrane proteins. *Eur J. Biochem.*, **190**, 207-219.

Farber, R., A. Lapedes, and K. Sirotkin (1992). Determination of eukaryotic protein coding regions using neural networks and information theory. *J. Mol. Biol.*, **226**, 471-479.

Ferran, E. and P. Ferrara (1992). Clustering proteins into families using artificial neural networks. *CABIOS*, **8**, 39-44.

Ferran, E. and B. Pflugfelder (1993). A hybrid method to cluster protein sequences based on statistics and artificial neural networks. *CABIOS*, **9**, 671-680.

Fickette, J. (1982). Recognition of protein coding regions in DNA sequences. *Nucl. Acids Res.*, **10**, 5303-5318.

Fickette, J. (1995). The Gene identification problem : An overview for developers. in *Proceedings of the fourth International Workshop on Open Problems in Computational Biology*, [Ed.] A. Konopka, also appearing in *Comp. Chem.* **20**, 103-118.

Frishman, D. and P. Argos (1992). Recognition of distantly related protein sequences using conserved motifs and neural networks. *J. Mol. Biol.*, **228**, 951-962.

Gillam, S., C. Astell, P. Jahnke, C. Hutchison and M. Smith (1984). Construction and properties of a ribosome-binding site mutation in Gene E of ϕ-X174 bacteriophage. *J. Virol.*, **52**, 892-896.

Goodacre, R., A. Karim, M. Kaderbhai and D. Kell (1994). Rapid and quantitative analysis of recombinant protein expression using pyrolysis mass spectrometry and artificial neural networks : Application to mammalian cytochrome b5 in Escherichia coli. *J. Biotech.*, **34**, 185-193.

Green, D. and J. Swets (1988). *Signal Detection Theory and Psychophysics*. Peninsula Publishing, Los Altos, CA.

Gupta, S., J. Chen, L. Schaefer, P. Lengyel and S. Weissmann (1970). Nucleotide sequence of a ribosome attachment site of bacteriophage F2 RNA. *Biochem. Biophys. Res. Commun.*, **39**, 883-888.

Harrison, A. (1970). *Solid-State Theory*, McGraw-Hill, New York.

Hawley, D. and W. McClure (1983). Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucl. Acids Res.*, **11**, 2237-2255.

Hindley, J. and D. Staples (1969). Sequence of a ribosome binding site in bacteriophage Qβ-RNA. *Nature*, **224**, 964-967.

Hirst, J. and M. Sternberg (1992). Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochemistry*, **31**, 7211-7218.

Holbrook, S., S. Muskal, and S-H. Kim (1990). Predicting surface exposure of amino acids from protein sequence. *Protein Engg.*, **3**, 659-665.

Holbrook, S. (1993). Application of computational neural networks to the prediction of protein structural features. In *Genetic Engineering Principles and Methods*, Vol. 15, [Ed.] Jane Setlow.

Holley, L. and M. Karplus (1989). Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci.*, USA, **86**, 152-156.

Hopp, T. and K. Woods (1981). Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci.*, USA, **78**, 3824-3828.

Hsu, K. and A. Hsu (1990). Fractal geometry of music (physics of melody). *Proc. Natl. Acad. Sci.*, USA, **87**, 938-941.

Hui, A., J. Hayflick, K. Dinkelspiel and H. De Boer (1984). Mutagenesis of the three bases preceding the start codon of the β-galactosidase m-RNA and its effect on translation in Escherichia coli. *EMBO J.*, **3**, 623-629.

Ivanov, I., R. Alexandrova, B. Dragulev, D. Leclere, A. Saraffova, V. Maximova and M. Abouhaidar (1992). Efficiency of the 5' terminal sequences (omega) of tobacco mosaic virus RNA for the initiation of eukaryotic gene translation in *E. coli*, *Eur. J. Biochem.*, **209**, 151-156.

Jones, D. (1975). Amino acid properties and side-chain orientation in proteins : Cross-correlation approach. *J. Theor. Biol.*, **50**, 167-183.

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

Kazic, T. (1992). Reasoning about biochemical compounds and processes. In *Proceedings of the Second International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*, [Eds.] H. Lim, J. Fickette, C. Cantor and R. Robbins, pp. 35-49.

King, R., J. Hirst, and M. Sternberg (1993). New approaches to QSAR : Neural networks and machine learning. *Perspectives in Drug Discovery and Design*, **1**, 279-290.

Kneller, D., F. Cohen, and R. Langridge (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.*, **214**, 171-182.

Kocur, C., S. Rogers, M. Kabrisky, J. Hoffmeister, K. Baer and J. Steppe (1995). Neural network wavelet feature selection for breast cancer diagnosis (submitted to *IEEE Engng. Med. Biol.*).

Kyte, J. and R. Doolittle (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105-132.

Lapedes, A., C. Barnes., R. Farber, and K. Sirotkin (1989). Application of neural networks and other machine learning algorithms to DNA sequence analysis. In *Computers and DNA, SFI Studies in the Science of Complexity*, [Eds.] G. Bell, and T. Marr, Vol. 7, pp 157-182, Addison-Wesley, Reading, MA.

Larsen, N., J. Engelbrecht, and S. Brunak (1995). Analysis of eukaryotic promoter sequences reveal a systematically occurring CT-signal. *Nucl. Acids Res.*, **23**, 1223-1230.

Liebman, M., C. Venanzi, and H. Weinstein (1985). Structural analysis of carboxypeptidase A and its complexes with inhibitors as a basis for modeling enzyme recognition and specificity. *Biopolymers*, **24**, 1721-1758.

Liebman, M. (1992). Application of neural networks to the analysis of structure and function in biologically active macro-molecules. In *Proceedings of the Second International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*, [Eds.] H. Lim, J. Fickette, C. Cantor and R. Robbins, pp. 331-347.

Lohmann, R., G. Schneider, D. Behrens and P. Wrede (1994). A neural network model for the prediction of membrane-spanning amino acid sequences. *Protein Science*, **3**, 1597-1601.

Lukashin, A., V. Anshelevich, B. Amirikyan, A. Gragerov and M. Frank-Kamenetskii (1989). Neural network models for promoter recognition. *J. Biomol. Struct. Dyn.*, **6**, 1123-1133.

Maclin, R. and J. Shavlik (1993). Using knowledge-based neural network to improve algorithms : Refining the Chou-Fasman algorithm for protein folding. *Machine Learning*, **11**, 195-215.

Mahadevan, I. and I. Gosh (1994). Analysis of *E. coli* promoter structures using neural networks. *Nucl. Acids Res.*, **22**, 2158-2165.

Mandal, M., A. Mandal, and A. Nath (1994). On estimation and prediction of myocardial infarction size using artificial neural network. *J. Sci. Ind. Res.*, **53**, 831-839.

Marabini, R and J. Carazo (1994). Pattern recognition and classification of images of biological macromolecules using artificial neural networks. *Biophys. J.*, **66**, 1804-1814.

Matteucci, M. and H. Heyneker (1983). Targeted random mutagenesis : the use of ambiguously synthesized oligonucleotides to mutagenize sequences immediately 5' of an ATG initiation codon. *Nucl. Acids Res.*, **11**, 3113-3121.

McGregor, M., T. Flores, and M. Sternberg (1989). Prediction of β-turns in proteins using neural networks. *Protein Engng.*, **2**, 521-526. [published erratum in *Protein-Engng.*, 3(5), 459-460 (1990)].

Meistrell, M. (1990). Evaluation of neural network performance by receiver operation characteristic (ROC) analysis : Examples from the biotechnology domain. *Comput. Meth. Prog. Biomed.*, **32**, 73-80.

Mount, S. (1982). A catalogue of splice junction sequences. *Nucl. Acids Res.*, **10**, 459-472.

Moyle, H., C. Waldburger, and M. Susskind (1991). Hierarchies of base pair preferences in the P22 ant promoter. *J. Bact.*, **173**, 1944-1950.

Munson, L., G. Stormo, R. Niece and W. Reznikoff (1984). lac Z translation initiation mutations. *J. Mol. Biol.*, **177**, 663-683.

Mural, R., J. Einstein, X. Guan, R. Mann and E. Uberbacher (1992). An artificial intelligence approach to DNA sequence feature recognition. *Trends in Biotech.*, **10**, 66-69.

Muskal, S., S. Holbrook, and S-H. Kim (1990). Prediction of the disulfide bonding state of cysteine in proteins. *Protein Engng.*, **3**, 667-672.

Muskal, S. and S-H. Kim (1992). Predicting protein secondary structure content : A tandem neural network approach. *J. Mol. Biol.*, **225**, 713-727.

Myers, R., S. Lerman, and T. Maniatis (1985). A general method for saturation mutagenesis of cloned DNA fragments. *Science*, **229**, 242-247.

Myers, R., K. Tilly, and T. Maniatis (1986). Fine structure genetic analysis of a β-Globin promoter. *Science*, **232**, 613-618

Nair, T. M., S. Tambe, and B. Kulkarni (1994). Application of artificial neural networks for prokaryotic transcription terminator prediction. *FEBS Lett.*, **346**, 273-277.

Nair, T. M. , S. Tambe, and B. Kulkarni (1995). Analysis of transcription control signals using artificial neural networks. *CABIOS*, **3**, 293-300.

Nair, T. M. and B. Kulkarni (1996). Prediction of transcription levels using neural networks : A pointer towards building a mutational database. *CABIOS* (submitted).

Nakata, K. (1995). Prediction of zinc finger DNA binding protein. *CABIOS* , **11**, 125-131.

Nakata, K., M. Kanehisa, and C. DeLisi (1985). Prediction of splice junctions in mRNA sequences. *Nucl. Acids Res.*, **13**, 5327-5340.

Nakata, K., M. Kanehisa, and J. Maizel (1988). Discriminant analysis of promoter regions in *E. coli* sequences. *CABIOS*, **4**, 367-371.

Neumann, E., D. Wheeler, A. Bernstein, J. Burnside, and J. Hall (1992). Artificial neural network classification of Drosophila courtship song mutants. *Biol. Cybern.*, **66**, 485-496.

O' Neill, M. (1991). Training back-propagation neural networks to define and detect DNA-binding sites. *Nucl. Acids Res.*, **19**, 313-318

Petersen, C. (1987). The functional stability of the lacZ transcript is sensitive towards sequence alterations immediately downstream of the ribosome binding site. *Mol. Gen. Genet.*, **209**, 179-187.

Qian, N. and T. Sejnowski (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, **202**, 865-884.

Rawlings, C. and J. Fox (1994). Artifical intelligence in molecular biology : a review and assessment. *Phil. Trans. R. Soc. Lond. B*, **344**, 353-363.

Richardson, J. (1981). The anatomy and taxonomy of protein structure in : *Advances in Protein Chemistry*, [Eds.] C. Anfinsen, J. Edsall, F. Richards, Vol. 34, pp. 167-339.

Ringquist, S., S. Shinedling, D. Barrick, L. Green, J. Binkley, G. Stormo and L. Gold (1992). Translation initiation in *E. Coli* : Sequences within the ribosome binding site. *Mol. Microbiology*, **6**, 1219-1229.

Rogers, S., D. Ruck, and M. Kabrisky (1994). Artificial neural networks for early detection and diagnosis of cancer. *Cancer Lett.*, **77**, 79-83.

Rosenberg, M. and D. Court (1979). Regulatory sequences involved in the promotion and termination of transcription. *Ann. Rev. Genet*, **13**, 319-353.

Sasagawa, F. (1993). Application of neural network with a modular architecture to protein secondary structure prediction. *Fujitsu Scientific and Technical Journal*, **29**, 250-256.

Sasagawa, F. and K. Tajima (1993). Prediction of protein secondary structures by a neural network with a modular architecture and super-computer. In *Computer-Aided Innovation of New Materials* II, [Eds.] M. Doyama, J. Kihara, M. Tanaka and R. Yamamoto, Elsevier Science B.V.

Sasagawa, F. and K. Tajima (1993a). Prediction of protein secondary structures by a neural network. *CABIOS*, **9**, 147-152.

Scarberry, R. and Z. Zhang (1996). Peptide sequence determination from high-energy CID spectra using artificial neural networks. *JASMS* (In press).

Schneider, G and P. Wrede (1994). The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution : De Novo design of an idealized leader peptidase cleavage site. *Biophys. J.*, **66**, 335-344.

Schneider, H., J. Schuchhardt and P. Wrede (1994). Artificial neural networks and simulated molecular evolution are potential tools for sequence-oriented protein design. *CABIOS*, **10**, 635-645.

Schneider, H., J. Schuchhardt and P. Wrede (1995). Peptide design in Machina : Development of artificial mitochondrial protein precursor cleavage sites by simulated molecular evolution. *Biophy. J.*, **68**, 434-447.

Schneider, H. and P. Wrede (1993). Development of artificial neural filters for pattern recognition in protein sequences. *J. Mol. Evol.*, **36**, 586-595.

Shannon, C. and W. Weaver (1971). *A Mathematical Theory of Communication*. University of Illinois press, Urbana, Illinois.

Shavlik, J. and G. Towell (1992). Using neural networks to refine existing biological knowledge. *Int. J. Genome Res.*, **1**, 81-107.

Simpson, R., P. Culverhouse, R. William, and R. Ellis (1991). Classification of Dinophyceae by artificial neural networks. In *IEEE Conf. Neural Networks and Ocean Engineering*, Washington DC, USA, August 1991. IEEE, Piscataway, NJ, pp., 223-229. Also see *Toxic Phytoplankton Blooms in the Sea* (1993). {Eds.} T. Smayda, Y. Shimizu, and Elsevier Science, New York, pp. 183-190.

Simpson, R., P. Culverhouse, R. William, and R. Ellis (1992). Biological pattern recognition by neural networks. *Mar. Ecol. Prog. Ser.*, **79**, 303-308.

Snow, P., D. Smith, and W. Catalona (1994). Artificial neural networks in the diagnosis and prognosis of prostrate cancer : A pilot study. *J. Urol.*, **152**, 1923-1926.

Snyder, E. and G. Stormo (1993). Identification of coding regions in genomic DNA sequences : an application of dynamic programming and neural networks. *Nucl. Acids Res.*, **21**, 607-613.

Snyder, E. and G. Stormo (1995). Identification of protein coding regions in genomic DNA. *J. Mol. Biol.*, **248**, 1-18.

Staden R. (1984). Computer methods to locate signals in nucleic acids sequences. *Nucl. Acids Res.*, **12**, 505-519.

Steitz, J.A. (1969). Polypeptide chain initiation : Nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. *Nature*, **224**, 957-964.

Stolorz, P., A. Lapedes, and Y. Xia (1992). Predicting protein secondary structure using neural net and statistical methods. *J. Mol. Biol.*, **225**, 363-377.

Stormo, G., T. Schneider, and L. Gold (1982). Characterization of translational initiation sites in *E. Coli*. *Nucl. Acids Res.*, **10**, 2971-2996.

Stormo, G., T. Schneider, L. Gold, and A. Ehrenfeucht (1982a). Use of perceptron algorithm to distinguish translational initiation sites in *E. coli*, *Nucl. Acids Res.*, **10**, 2997-3011.

Sumpter, B., C. Getino, and D. Noid (1994). Theory and applications of neural computing in chemical science. *Ann. Rev. Phys. Chem.*, **45**, 439-481.

Tetko, I., V. Tanchuk, N. Chentsova, S. Antonenko, G. Poda, V. Kukhar and A. Luik (1994). HIV-1 reverse transcriptase inhibitor design using artificial neural network. *J. Med. Chem.*, **37**, 2520-2526.

Thanaraj, T. and M. Pandit (1989). An additional ribosome binding site on mRNA of highly expressed genes and a bifunctional site on the colicin fragment of 16S rRNA from Escherichia coli : important determinants of the efficiency of translation-initiation. *Nucl. Acids Res.*, **17**, 2973-2985.

Towell, G. and J. Shavlik (1993). Extracting refined rules from knowledgbe-based neural networks. *Machine Learning*, **13**, 71-107.

Towell, G. and J. Shavlik (1994). Knowledge-based artificial neural networks. *Artificial Intelligence*, **70**, 119-165.

Trifonov, E. (1985). Construction of an algorithm for locating splicing junctions. *CODATA*, pp. 119-124.

Uberbacher, E. and R. Mural (1991). Locating protein-coding regions in human DNA sequences by a multiple sensors-neural network approach. *Proc. Natl. Acad. Sci.*, USA, **88**, 11261-11265.

Veljkovic, V. and I. Cosic (1987). A novel method of protein analysis for prediction of biological function : Application to tumor toxins. *Cancer Biochem. Biophys.*, **9**, 139-148.

Veljkovic, V. and I. Slavic (1972). Simple general-model pseudopotential. *Phys. Rev. Lett.*, **29**, 105-107.

Veljkovic, V. and R. Metlas (1987). Theoretical prediction of region in IL-2 primary structure important for its activity. In *Proc. Protein Engineering*, Oxford, p. 102.

Veljkovic, V. and R. Metlas (1988). Identification of nanopeptide from HTLV-III, ARV-2 and LAV envelope gp120 determining binding to T4 cell surface protein. *Cancer Biochem. Biophys.*, **10**, 191-206.

Veljkovic, V., I. Cosic, B. Dimitrijevic and D. Lalovic (1985). Is it possible to analyze DNA and protein sequences by the method of digital signal processing? *IEEE Trans. Biomed. Engng.*, **32**, 337-341.

Venkatachalam, C. (1968). Stereochemical criteria for polypeptides and proteins.V. Conformation of a system of three linked peptide units. *Biopolymers*, **6**, 1426-1436.

Vogel, H., J. Wright, and F. Jaehnig (1985). The structure of the lactose permease derived from Raman spectroscopy and prediction methods. *EMBO J.*, **4**, 3625-3631.

Weinstein, J., K. Kohn, M. Grever, V. Viswanadhan, L. Rubinstein, A. Monks, D.A. Scudiero, L. Welch, A. Koutsoukos, A. Chiausa and K. Paul (1992). Neural computing in cancer drug development : Predicting mechanism of action. *Science*, **258**, 447-451.

Wilcox, G., M. Poliac, and M. Liebman (1990). Neural network analysis of protein tertiary structure. *Tetrahed. Comput. Meth.*, **3**, 191-211.

Williams, R., P. Culverhouse, R. Simpson, R. Ellis, J. Lindley, H. McCall, B. Requera, I. Bravo, and T. Parasini (1993). Identification of species of Ceratium and Dinophysis by artificial neural networks. In *Proc 6th Int. Conf. Toxic Marine Phytoplankton*. Nantes, Oct. 18-22.

Wu, C. and S. Shivakumar (1994). Backpropagation and counter-propagation neural networks for phylogenetic classification of ribosomal RNA sequences. *Nucl. Acids Res.*, **22**, 4291-4299.

Wu, C. (1993). Classification neural networks for rapid sequence annotation and automated database organization. *Comput. Chem.*, **17**, 219-227.

Wu. C., C. Wang, and I. Yazdanpanahi (1992). Protein classification artificial neural system : A filter program for database search. In *Proceedings of the Second International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*, [Eds.] H. Lim, J. Fickette, C. Cantor, R. Robbins, pp. 348-358.

Xin,Y, T. Carmeli, M. Liebman, and G. Wilcox (1992). Use of the back-propagation network algorithm for prediction of protein folding patterns. In *Proceedings of the Second International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*, [Eds.] H. Lim, J. Fickette, C. Cantor, R. Robbins, pp. 358-375.

Xu, Y., R. Einstein, R. Mural, M. Shah and E. Uberbacher (1994). An improved system for exon recognition and gene modeling in human DNA sequences. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, Stanford University, San Franscisco, CA, Aug. 14-17.

Xu, Y., R. Mural, M. Shah and E. Uberbacher (1995). Recognizing exons in genomic sequence using GRAIL-II. In *Genetic Engineering, Principles and Methods*. Plenum Press.

Xu, Y., R. Mural, and E. Uberbacher (1995a). Constructing gene models from accurately-predicted exons : An application of dynamic programming. *CABIOS* (In press)

Xu, Y, R. Mural, and E. Uberbacher (1995b). Correcting sequencing errors in DNA coding regions using a dynamic programming approach. *CABIOS* (In press).

Yee, D., M. Prior, and L. Florence (1993). Development of predictive models of laboratory animal growth using artificial neural networks. *CABIOS*, **9**, 517-522.

Youderian, P., S. Bouvier, and M. Susskind (1982). Sequence determinants of promoter activity. *Cell*, **30**, 843-853.

Zamyatnin, A. (1972). Protein volume in solution. *Prog. Biophys. Mol. Biol.*, **24**, 107-123 ; Also see *Biofizika*, **16**, 163-171 (1971).

Zipser, D. and R. Andersen (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Science*, **331**, 679-684.

■■■