Chapter 15
Neural Networks in Bioinformatics

Masood Zamani and Stefan C. Kremer

1 Introduction

Bioinformatics or computational biology is a multidisciplinary research area that combines molecular biology, computer science, and mathematics. Its aims are to organize, utilize and explore the vast amount of information obtained from biological experiments for understanding the relationships and useful patterns in data. Bioinformatics problems, such as protein structure prediction and sequence alignments, are commonly categorized as non-deterministic polynomial problems, and require sophisticated algorithms and powerful computational resources. Artificial Intelligence (AI) techniques have a proven track record in the development of many research areas in the applied sciences. Among the AI techniques, artificial neural networks (ANNs) and their variations have proven to be one of the more powerful tools in terms of their generalization and pattern recognition capabilities. In this chapter, we review a number of bioinformatics problems solved by different artificial neural network architectures.

In a field as young and diverse as bioinformatics, it is always a challenge to try to organize the scope of problems and their respective solutions in a sensible way. In text, this organization is further constrained to a mostly linear narrative. If we view biological systems as information processing devices, then we can trace a typical flow of information from DNA sequences, to RNA, and then to protein sequences (following the path of the "Central Dogma of Molecular Biology" [18]). From there, the information can be viewed to move on to protein structure, functionality and higher level of biological phenomena. We can use this flow to guide our narrative in this chapter.

Many problems in bioinformatics involve predicting later stages in the information flow from earlier ones. Bioinformatics methods capable of such predictions

Masood Zamani · Stefan C. Kremer
The School of Computer Science at the University of Guelph,
Guelph, Ontario
e-mail: {mzamani,skremer}@uoguelph.ca

M. Bianchini et al. (Eds.): *Handbook on Neural Information Processing*, ISRL 49, pp. 505–525.
DOI: 10.1007/978-3-642-36657-4_15

can often eliminate costly, difficult, or time-consuming tasks in important biological research. For example, predicting protein structure and function based on amino acid sequence, is an essential component of modern drug design, and can replace expensive wet-lab work.

In our narrative we will situate problems, first, along their input data and secondly along their outputs as shown in Tables 1, 2 and 3. For each problem type we will proceed to describe, section by section, the nature of the data, its representation and any special considerations when using this data with artificial neural networks. We also consider the nature of the computational problem to be solved and discuss how to effectively apply a neurally inspired solution to it.

More specifically, this chapter is organized as follows. In Section 2 we discuss problems involving the analysis of DNA, including the detection of promoter regions, RNA coding regions, rare events, new motifs and DNA barcoding. Next, in Section 3 we turn our attention to peptide, or amino acid sequences. This section covers many problems related to the elucidation of structure and function of proteins, including: identifying secondary structure components, structural domains, disulphide bonds, contact points, solvent accessibility and protein binding sites, motifs, protein stability and protein interactions. Finally, in Section 4 we discuss the highest level of bioinformatic analysis including the diagnosis of cancers using spectrometry and microarray data.

This chapter is intended to provide an introduction to the predominant research areas and some of the approaches used within bioinformatics.

2 Analyzing DNA Sequences

In this section, we examine approaches that involve analyzing DNA sequences. DNA is a class of molecules that consist of a helical pair of polymers. The polymers are complementary and encode identical information. Each polymer is composed of many nucleotides that are joined in sequential fashion along a *backbone*. The information encoded in DNA can be viewed as a very long sequence of 4-base symbols since there are only four standard nucleic acids in DNA. These long strings of information are then transcribed into shorter segments by a process known as transcription. The shorter strings are composed of a similar molecule called RNA that employs the same type of 4-base representation; and, each such RNA string represents a code for a specific molecule. In many cases the RNA molecules are not themselves end products, but merely an encoding of a different type of molecule called a protein. Proteins are also polymers composed of simpler components joined in sequence, but the building blocks of proteins are amino acids (instead of nucleic acids). As there are 20 different types of standard amino acids, it takes at 3 symbols in the 4-base RNA code to uniquely identify a single symbol in the 20-base protein code. In fact, there is a redundant encoding from the 64 possible, 4-base triples to the 20-base amino acids.

Since DNA is the carrier of heritability, this is a reasonable place to start our discussion. It is relatively easy to build a neural system that processes DNA. Typically,

Table 1 Nonlinear Model Results (pt.1)

Input data	Output	Method
DNA sequence	Promoter regions	Promoter region identification [10]
DNA sequence	RNA gene	Non-coding RNA gene finder [56]
DNA sequence	Functional RNA genes	Detection of functional RNA genes using feed-forward neural networks [15]
DNA sequence	Classifying rare events in human genome	Detection of rare event in unbalanced data using neural networks [16]
DNA sequence	Clustered gene expression patterns	Analyzing correlated gene expression patterns using unsupervised neural networks [31]
DNA sequence	DNA motifs	Identifying unknown DNA motifs on DNA sequences using unsupervised neural networks [4]
DNA sequence	Classification of DNA barcoding genes	Inferring species membership via DNA barcoding with back-propagation neural networks [68]
DNA sequence	mRNA's donor and acceptor sites	Predicting donor and acceptor location on human pre-mRNA with feed-forward neural networks [12]
AA sequence	Sequence classifications	Protein Sequence Classification using Bayesian neural networks [62]
AA sequence	Clustered sequences	Unsupervised Kohonen learning technique [26]
AA sequence	Coil locations	Coil prediction [30]
AA sequence	β-sheet locations	Predicting protein β-sheets using alignment, neural networks and graph algorithm [13]
AA sequence	β-turn locations	Prediction of protein β-turn structure using evolutionary information and neural networks [36]
AA sequence	Protein Structural domains	Decomposition of protein structures into structural domains using profile and ANN [28]
AA sequence	Protein domain boundaries	Predicting protein domain using bidirectional recurrent neural networks [60]
AA sequence	Disulphide bonds	Disulphide bond prediction with a 2D-recurrent network [59]
AA sequence	Prediction of residue contacts	2D-recurrent neural networks for Protein contact map prediction [58]

Table 2 Nonlinear Model Results (pt.2)

Input data	Output	Method
AA sequence	Secondary structure	Predicting the secondary structure of globular proteins using MLP [52]
AA sequence	Secondary structure	Prediction of protein secondary structure using sequence profiles and neural networks [53]
AA sequence	Secondary structure	Prediction of protein secondary structure using evolutionary information and neural networks [54]
AA sequence	Secondary structure	Prediction of protein secondary structure using Position Specific Scoring Matrix(PSSM) and neural networks [34]
AA sequence	Secondary structure	Prediction of protein secondary structure using hidden neural networks [47]
AA sequence	Secondary structure	Prediction of protein secondary structure using bidirectional recurrent neural networks [7]
AA sequence	Real values of the solvent accessibility	Feed-forward neural networks for predicting the real values of solvent accessibility of amino acid [2]
AA sequence	Real values of the solvent accessibility	Approximating the real-value relative solvent accessibility (RSA) of AA residues [1]
AA sequence	Protein binding sites	Binding site prediction with neural network [37]
AA sequence	Secondary structure, solvent accessibility, backbone structural motifs, and contact density	Predicting 1D structural properties using structural alignment method (SAMD) and recursive neural networks [50]
AA sequence	Signal peptides	Detection of signal peptides in proteins [51]
AA sequence	Detection of protein stability	Prediction of protein stability changes using statistical potentials and multi-layer feed-forward neural networks [20]
AA sequence	Detection of protein disorders	Predicting protein disorder for N-, C- and internal regions [46]
AA sequence	Detection of motifs	Predicting proteasome cleavage motifs using artificial neural networks [38]
AA sequence	Detection of drug resistant factor	Predicting HIV drug resistance with neural networks [21]
AA sequence	Protein superfamilies	Classifications of protein sequences based on superfamily classes [66]

Table 3 Nonlinear Model Results (pt.3)

Input data	Output	Method
Mass spectrometry data	Diagnosis of tumours	Classifying human tumour and identification of biomarkers [8]
DNA microarrays	Diagnosis of cancers	Classification and prediction of cancers using gene expression profiling and artificial neural networks [39]
DNA microarrays	Diagnosis of breast cancers	Detecting breast cancer using artificial neural networks [45]
DNA microarrays	Classification of diseases	Classification of gene expression data using ensemble neural networks [48]

a sliding window of fixed length is applied to the sequence, and the nucleic acids that fall within the window are encoded in a one-hot fashion. That is, four input units are used to represent each nucleotide and exactly one of these units (corresponding to one of the four different nucleotides) is activated each time. In this section, we consider four different goals in analyzing DNA: (i) identifying RNA coding regions in the DNA (arbitrary and specific fRNA), (ii) identifying promoter regions in the DNA, (iii) detecting disease carriers, and (iv) DNA barcoding.

While the central dogma of molecular biology encompasses how DNA is transcribed into RNA and then translated into protein sequences, most DNA does not code for proteins. Originally, called "Junk DNA" these parts of the genome are beginning to be better understood. In some cases, DNA is transcribed into functional RNA (fRNA) that is never translated into a protein but rather performs a directly useful biological function. Such RNA can be referred to as "non-coding" and the DNA regions that prescribe it are called "non-coding genes". Non-coding RNA genes have been explored for their hidden and important roles in cells. A challenging task is the identification of non-coding RNA genes due to the diversity and the lack of consensus patterns for their genes. One avenue is to identify transcription factor binding sites: locations in the DNA where special molecules attach and begin the process of transcribing the DNA into RNA. A novel approach using fuzzy neural networks for non-coding RNA gene prediction was proposed in [56]. The hybrid approach has the advantages that give the nodes and parameters in the neural network physical meanings and provide a means to incorporate the qualitative prior knowledge by fuzzy set theory.

Another research area related to RNA is the detection of the gene encoding functional RNA (fRNA). In brief, fRNAs are the set of RNA genes which generate functional RNA products such as transfer RNA(tRNA) and microRNA(miRNA) without translation to protein. For instances, tRNA is involved in translation of the three-letter code in messenger RNA into the amino acids of proteins. In [15], a feed-forward neural network is employed for fRNA gene detection. Evolutionary computation is used to optimize the architecture of the neural networks. In other words, the neural network is evolved and optimized by deletions and insertions of nodes and connections and also adjusting the weights associated between two nodes.

Another type of pattern that can be found in DNA is the promoter region. These regions provide convenient places for the RNA polymerase proteins to attach to a DNA strand and begin the transcription process. In this fashion, these regions serve a regulatory role. Identifying promoter regions using artificial neural networks has been also studied in [10]. The traditional promoter prediction methods mainly search for motifs. However, recent studies in [35], [42] and [61] indicate that DNA structural features such as curvature, and stress-induced duplex destabilization (SIDD) also provide valuable information. In [10], SIDD profile data obtained from *E. coli* is used as the training data for the neural network.

One challenge faced by bioinformaticians is an usual sparsity of data. While there are often many long genetic sequences available, the most interesting phenomena are sometimes extremely rare. Therefore, a rare event leads to a variety of needle-in-a-haystack problems which have to be modelled and understood. Rare events are log normally distributed, so methods based on statistics that assume Gaussian distributions (e.g. arithmetic means) fail. However, sample stratification is a useful technique for rare event detection in unbalanced data especially in molecular biology. The technique makes each class in a sample data have equal weight in decision making. Using a neural network for sample stratification and detection of rare events was examined in [16]. The experiment was carried out on human genome DNA, and it showed significant improvement for rare event detection.

A common task with regard to the voluminous data in molecular biology is the detection of unique features from DNA sequences. In [4], an unsupervised learning class of ANNs, known as self-organizing map (SOM) [41], was studied in order to detect new motifs (domains) in DNA sequences. It was used to detect the signal peptide coding region on a dataset of human insulin receptor genes. SOMs are useful in pattern clustering and feature detection since this class of neural networks form internal representations that model the underlaying structures of input data. In the study, no prior knowledge, such as sequence alignment analysis, was embedded in the neural network. Yet, after the neural network training, the existence of minimal similarity patterns (MSPs) among the trained data was found by a statistical measure called “Tanimoto similarity” which is proportional to the difference between the input and weight vectors. The proposed method may potentially facilitate the identification of other DNA domains such as functional DNA patterns by performing further analysis on MSP clusters.

The final problem that we will discuss in this section stems from the field of taxonomy. Traditional taxonometric methods identify species by painstaking observation of morphological features—the physical characteristics of an organism. While this method has served scientists since before the days of Aristotle, it can be problematic. Many organisms are so small that observation of physical differences even using microscopy is difficult. In other cases, organisms have multiple life stages with very different forms that need to be individually identified, or significant differences among sexes. Sometimes the physical form of an organism is affected by its environment (including diet, habitat, etc.). In these cases, relying on the observation of physical traits is problematic. With the advent of genetic sequencing another approach is possible. By directly comparing the DNA of organisms it is possible to make species identifications [29]. Ideally, this is done by focusing on specific genetic traits that vary among species but not within species. A first approach might be to identify a specific gene with this property and then to measure differences among instances of this gene across organisms using a classical genetic distance measure (such as alignment scores). Current distance-based methods for species identifications using DNA barcoding sequences are frequently criticized for treating the nearest neighbour as the closest relative using a number of raw similarity scores. In [68], a feed-forward neural network is employed for the classification of DNA barcoding sequences. The results indicate a better performance compared to the previous methods such as basic local alignment search tool(BLAST) [3] which is a simple genetic distance-based method.

2.1 Example Application

In the following, we briefly explain an application of ANNs for the identification of donor and acceptor sites on messenger RNA (mRNA). In eukaryotic organisms an important stage before the translation of a messenger RNA molecule to a correct protein is the remove the introns (non-coding regions) and joining exons (coding regions) in a process known as RNA splicing. In other words, the DNA coding for a particular protein will often be discontinuous and interrupted by these introns. The splicing mechanism removes these introns and concatenates the exons to form a correct RNA molecule for the protein to be assembled.

An mRNA is a molecule that is copied from DNA during a process called transcription. An mRNA molecule carries a "blueprint" of the genetic information for synthesizing a protein. In vertebrates, small ribonucleoproteins recognize the splice sites by detecting the sequences around exon-intron transitions. In [12], a feed-forward neural network has been applied to predict splicing sites in human pre-mRNA. In this study, a joint prediction scheme for exon and intron regions was developed since the transitions between exon and intron regions control cutoff positions for the splicing process, and can therefore lead to the prediction of splicing sites. The dataset used for the training and test was obtained from GenBank Release 62.0.

A subset of the dataset was eliminated based on sequences with only one intron, no complete RNA transcript, or more than one gene. In total, 95 genes remained for training and testing after the eliminations based on the afore mentioned criteria. The dataset was divided into two parts in which 65 genes were used for the training of neural networks and the remaining 30 genes for testing. Since many genetic datasets that are collected in this way tend to have strong sister sequences that are nearly identical to each other and thus trivialize the problem, it is important to keep such sister sequences together (in either the training or testing datasets), rather than separating them (into training and testing). To avoid such high similarities among genes, the genes were alphabetically order prior to dividing the dataset.

Another challenge to measuring the predictive performance of neural networks arises since the non-donor/non-acceptor sites outnumber the donor/acceptor sites, resulting in largely imbalanced classes. To overcome the imbalanced classes, the correlation efficient method [49] was applied for the evaluation of the neural networks. The neural network inputs were prepared based on a sliding window that scanned the DNA sequence. The window length was the number of nucleotides within it. The nucleotides *A*, *G*, *C*, *T* and unknown nucleotides were represented using a one-hot encoding in 4-digit binary numbers as 1000, 0100, 0010, 0001 and 0000 respectively. A single output of the neural network predicted whether or not the nucleotide represented in the middle of the input window was a donor or acceptor site. The neural network had a single hidden layer, and its performance was evaluated with different numbers of neurons in that hidden layer and varying inputs neurons (equivalent to window size).

The experimental results indicated the optimal prediction of neural networks were achieved whit a window size of 15 nucleotides and 40 neurons in the hidden layer for donor sites as compared to a window size of 41 nucleotides and 20 neurons in hidden layer for acceptor sites. With the selected architectures for the neural networks, the percentage of positive prediction of donor and acceptor sites were 95%, whereas the percentage of false predictions for donor and acceptor sites were only 0.1% and 0.4% respectively.

2.2 Conclusion

In this section, we have surveyed some neural approaches to DNA analysis. The representation of a DNA sequence as an input vector to an artificial neural network via a sliding-window approach is straightforward and has been used to great effect in the methods described above.

3 Peptide Sequence Analysis

All biological functions are based on the interactions of proteins. Proteins serve as catalysts in many biochemical reactions and play critical roles in the structure and behaviour of all cells. Protein's chemical properties are determined by their amino acid constituents, as well as their shapes since the shape of a protein affects the accessibility of the amino acids. Since amino acid sequence generally determines the shape of a protein (prions are a notable exception), it should be possible to predict a protein's shape based on its amino acids. This quest has long been viewed as a "holy grail" of bioinformatics.

In order to tackle such an important and challenging problem, a number of subtasks to the problem of determining a protein's three-dimensional shape have been identified. Proteins exhibit regions of structural patterns whose shapes are well defined. They can be held together by bonds between non-neighbouring amino acids (in the protein's chain) called disulphide bonds. Moreover, there is a fairly strong structural homology (similar shapes) among homologous protein sequences (similar sequences).

Coding amino acid sequences as neural network inputs can be accomplished by a sliding window and a representation scheme for individual amino acids. Since there are twenty amino acids, it is possible to encode them by twenty input units with a one-hot encoding. This tends to result in a very large input vector which in turn leads to a large number of connections, and trainable parameters. Having too many trainable parameters can result in over fitting (especially if the sample space is small). Of course, twenty distinct values can be encoded in a 5-bit vector using a binary representation. This type of representation may not be ideal however, since certain patterns (00000 and 00001) are much closer in Euclidean or Hamming space than others (01100 and 10011). Thus, a binary representation can implement a bias in the network favouring output mappings which treat the similar input patterns similarly. By contrast, the 20-input one-hot approach places every amino acid at a point equidistant to every other amino acid using any metric (Euclidean, Hamming, etc.). A number of alternative amino acid encodings using physicochemical properties of amino acids have been proposed and employed in [44], [67].

Once an encoding of the input has been developed, the next task is to determine what the neural network should output or predict. The first stage in understanding a protein's structure (and thereby gleaning insight into its function) is often to recognize particular sequence patterns. In [62], a Bayesian neural network approach is used to classify protein sequences. The features selected from the protein sequences for the input of the neural networks are based on both the global and local similarities.

Organizing and searching for homologous sequences in DNA and protein databases are essential tasks. An interesting clustering result with an accuracy of 96.7% for protein sequences into families using ANNs was accomplished in [26]. The unsupervised Kohonen learning method [41] has been used to train the network and cluster protein sequences since the number of composition and protein families were not known in advance. The neural network clustering approach is different than the non-hierarchical statistical methods for clustering data that usually require the number of expected classes be defined in advance [5].

Fortunately, there are some constraints on the structures of proteins. Because of weak covalent bonds between the hydrogen atoms in the amino acids, the amino acids themselves are often drawn into tight, stable arrangements. Two common arrangements are helices (where covalent bonds form between nearby amino acids coiling the polymer), and sheets (where covalent bonds form between two or more long polymer strands that run parallel or anti-parallel to each other). A third structure called a *coil* is more of a catch-all category that covers irregular regions of proteins. These three structure categories are referred to as secondary structure—the primary structure being the linear sequence of amino acids, while the tertiary structure is the detailed three-dimensional shape of the protein.

ANNs have also been used to predict which parts of a protein form each type of secondary structure and the detailed arrangements within the secondary structures. The most challenging parts of the prediction are perhaps coils which are irregular structural patterns. In [30], neural networks were used to predict dihedral angle probability distributions of coils from protein sequences. The network inputs are organized in a predefined window of residues. The dihedral angle probability distribution is predicted for the middle residue. The results indicate improvements compared to those using statistical methods [43], [70].

In addition to the covalent hydrogen bonds between amino acids, there are also di-suplide bonds. These form between pairs of one particular type of amino acid, cysteine. These bonds are also known as bridges as they form connection between amino acids that would otherwise be separated larger regions of space. Having prior knowledge of disulphide bridge locations in a protein structure is very valuable for the prediction of the protein backbone conformation. A recurrent artificial neural network has been successfully applied for disulphide bonding prediction [59]. This approach creates a two-dimensional matrix of potential disulphide bridges. Each dimension represents one amino acid in the potential disulphide bridge. This matrix representation can then be used both to formulate the input and the output of a neural network. Additionally, recurrent connections can then be used to propagate information about amino acids that surround the potential disulphide bond. An important challenge that many bioinformatics approaches face is that the number of exemplars (actual proteins that exist in nature) tends to be relatively small compared to the input space (all possible amino acid chains). This can result in a high dimensional parameter space with only few points, and consequent overfitting (as noted previously). The work in this area uses alignment profiles and homologies to expand the input examples and thus mitigate this issue. In addition, the study shows that using multiple alignment profiles improves the prediction accuracy which emphasizes on the importance of evolutionary information.

A protein's secondary structure is defined based on its common 3D structural patterns. Protein secondary structures are grouped into three structural classes: the α-helix (a spiral conformation), the β-sheet (a twisted, pleated sheet) and the coil (the most irregular structural pattern). Most proteins are composed of sections of all three of these classes. It is possible to assign each amino acid in a protein to one of these categories based on the protein shape at that amino acid's location. Thus, an amino acid sequence can be converted into a structural class sequence called a secondary structure. Predicting secondary structure, in turn, can be used as initial information by methods using free energy minimization to study protein pathways leading ultimately to 3D structure predictions. The pioneering work of using neural networks for predicting the secondary structure of globular proteins was proposed in [52]. In the study, an MLP with one hidden layer was used. The main drawback of the method's architecture was the over-fitting problem. Several techniques were introduced to address the problem, such as ensemble averages by training independently different neural networks, different input information [53, 54] and alternate learning procedures [34]. Using multi-sequence alignments for the network input instead of raw amino acid sequence data has significantly improved the result because secondary structures tend to be preserved across homologous proteins.

In [54], evolutionary information obtained from protein sequence alignments is used as the neural network inputs, and in turn it increased the classification accuracy. The effectiveness of incorporating protein evolutionary information with a neural network has been also studied for predicting β-turn patterns [36]. A fast protein secondary structure prediction using MLP was also proposed in [47], where the outputs of the neural network are passed to a Hidden Markov Model (HMM) to optimize the predictions.

In addition, a bidirectional recurrent neural network was used for the protein secondary structure problem in [7] to overcome the limitations of the past methods that used fixed-size input windows. While a conventional network's input is limited by its fixed-size input layer, a recurrent network uses feedback to process information over time (this distinction is similar to that between combinatorial and sequential logic circuits). By processing information over time, the input is not limited to a fixed size. This results in an architectural parsimony whereby a network with fewer adaptable parameters is able to process large input patterns. This, in part, helps to overcome the overfitting problem. In a three-stage method proposed in [13] for predicting protein β-sheets, a recurrent neural networks has been used as a primary step to obtain the residue pairing probabilities of all pairs in β-sheets. The inputs of the neural network are generated from profile, secondary structure and solvent accessibility information.

In a protein, structural domains have important applications in protein folding, evolution, function and design. They are common structures that occur in multiple proteins. These native-like structures are independent of the rest of the protein in the sense that they remain folded if separated from the rest of the protein. Methods for protein domain decomposition, such as using graph theory, are not accurate when the number of structural domains in a protein is not known prior to partitioning. To effectively assess the quality of a partition, in [28] the structural information of a protein including the hydrophobic moment profile and a neural network were used to evaluate the quality of identified domains. Using neural networks contributed significantly to an increase in prediction accuracy from 74.5% to 81.9%. Bidirectional recurrent neural networks have been utilized to predict protein domain boundaries [60]. The performance of these neural networks relies on protein sequences, evolutionary information and protein structural features.

Detecting signal peptides (SP) in proteins using ANNs was studied in [51], and the Swiss-Prot protein database was used to evaluate the performance of the proposed method. Signal peptides are the short fragments of proteins that lead newly synthesized proteins to find their destinations. One advantage of the proposed method is its computational speed compared to those of other approaches in [22], [25].

Another method for describing the shape of a protein uses a similar technique described earlier for the disulphide bond prediction. Specifically, it aims to identify neighbouring amino acids (without the presence of covalent bonds, like those between sulphur atoms). These neighbouring amino acids are considered “contacts”, and a complete 2D matrix showing the degree of proximity between all pairs of amino acids is called a “contact map”. Protein contact maps have important applications in proteins such as inferring protein folding rates, evaluating protein models and improving protein 3D structure predictions. Protein contact maps are encoded as a matrices of residue-residue contacts within a distance threshold. The NNcon method proposed in [58] is a protein contact map prediction technique based on 2D-recurrent neural networks. It maps 2D input information into 2D output targets. NNcon has been ranked among the best contact map prediction methods in CASP8. NNcon can be used to predict both general residue-residue contacts and specific beta contacts in β-sheets.

In [2], a feed-forward neural network was used for predicting the real values of solvent accessibility of amino acid residues. Solvent accessibility identifies which parts of a protein are accessible on the surface of the three dimensional structure as opposed to lying in the interior hydrophobic core. Understanding accessibility sheds light on the chemical properties of the protein. The method in [2] predicts the real values of accessible surface area from the sequence of information without a prior classification of exposure states unlike the past techniques classifying residues into buried and exposed states with varying thresholds. The categorical nature of such methods reduces the amount of information. Surface accessibility values are regarded as features used to improve the techniques applied for protein structure prediction. In [1], a neural network-based regression method was used to approximate the real-value relative solvent accessibility (RSA) of amino acid residues. Unlike other methods, the approach is not based on a classification problem which needed arbitrary boundaries among the classes. Instead, the method employs several feed-forward and recurrent neural networks and eventually combines them into one consensus predictor.

Using the 1D structural properties of a protein is an alternate way of exploring the correlation between a protein sequence and its 3D structure. Predicting the structural properties is valuable for protein structure and function prediction. The automated structural alignment method, SAMD, proposed in [50] for protein 1D structure prediction employs a recursive neural network and uses remote homology information unlike most 1D prediction methods that do not incorporate the homology information into the prediction process. The method is able to predict four structural properties which are: secondary structure, solvent accessibility, backbone structural motifs and contact density. The structural information is coded into the templates of structural frequency profiles and used as additional inputs to the recurrent neural networks to predict 1D-structural properties of query sequences. The systems is capable of making predictions by relying on data of only remotely homologous sequences whose structures are known in the Protein Data Bank (PDB) [9].

Predicting protein function is important to understand protein folding mechanisms. Protein function information is also correlated to its 3D structure. There are a number of neural network applications in the protein function prediction area, in particular, the identification of protein binding sites such as in [37]. Also, designing proteins that are able to function robustly in unusual environments such as in extreme pH and temperatures is very important. An interesting exploration also would be to change protein properties with a number of substitutions, and then predicting whether the mutations affect the stabilities of the proteins. In [20], a fast and accurate method was proposed for predicting protein stability changes when amino acids are mutated in a protein sequence. Statistical potentials and a multilayer feed-forward neural network are the two main components used in this approach to predict protein stability changes.

Artificial neural networks have also been used to predict protein disorder for N-, C-termini and internal regions [46]. A polypeptide chain has two unbounded ends which are a carboxyl group (*-COOH*) called the C-terminus and an amino group (*-NH2*) called the N-terminus. The translation of a protein from a messenger RNA (mRNA) starts from the N-terminus and ends with the C-terminus. A protein's function depends on its 3D structure in native state when it is known to be completely folded. By contrast, it has been observed, e.g. in [27], [11], that some proteins are partially or completely unfolded in their native states. The so-called natively unfolded or disordered proteins were investigated and it was postulated that the disordered regions due to the lack of a fixed 3D structure could be the integral parts of a novel protein function [63], [65]. The experiment in [46] indicates a higher prediction accuracy for disordered regions compared to those of discriminant analysis and logistic regression methods [17], [24].

Predicting proteosome cleavage motifs has also been examined using artificial neural networks in [38]. The motif prediction is a crucial step to understanding a cellular process such as metabolic adaptation and regulation of immune responses. The artificial neural network application has been also examined for the prediction of HIV protease mutants [21]. Predicting the resistant factor helps current HIV therapies in developing more effective treatments.

3.1 Example Application

In the following, an ANN application for protein "superfamily" [19] classification is explained in summary. Although the term superfamily refers to a group of evolutionarily related proteins, it is also applied to a group of structurally or functionally related proteins. A common task with regards to amino acid sequences is the classification of these protein sequences into superfamilies which often possess a common origin, structure and function. An example of protein classification techniques at the superfamily level that employs artificial neural networks is ProCANS (Protein Classification Artificial Neural Networks System) [66]. ProCANS has been implemented on a Cray supercomputer and is used for classifying unknown proteins at the superfamily level by embedding the information of the Protein Identification Resource (PIR) database [55] which is organized according to the superfamily concept. The two main steps of the ProCANS system are the sequence encoding scheme, to extract information from sequences without knowing the importance of its features in the classification model, and modular network architecture, a collection of small feed-forward neural networks instead of one large neural network to increase the processing of large and complex databases [40].

The important part of the sequence encoding scheme used in the study is a hashing function called the *n*-gram extraction method [14]. Using the *n*-gram extraction method, all patterns of possible *n* consecutive residues (or an alphabetic set) in a sequence are extracted, and the total number of each pattern's occurrences is recorded. Then, the sequence is represented as an m-dimensional vector called a "count vector" where each element of the vector corresponds to each pattern's total count. For instances, with twenty amino acids there are four hundreds possible residue pairs (patterns) using a 2-gram (bigram) extraction method. In ProCANS, ten sequence encodings were used according to two alphabet sets: amino acids and exchange group sets consisting alphabets of size twenty and six respectively. The first five encodings are based on count vectors which combine the various patterns of amino acids and exchange group patterns. For example, the encodings "a2" and "a2e2" are the bigram amino acids and the concatenation of the bigram amino acids and exchange group patterns respectively. The rest of the five encodings are based on a "position vector". Position vectors are generated according to the n-gram method, however each element of the vector is the average of each pattern's position (or order) in the sequence.

The second step in the ProCANS system is database modularization. At this step, the PIR database is divided to multiple sets according to protein functional groups such as electron transfer proteins, growth factors, etc.. The described encoding schemes are applied on the sets, and each set is used for the training of a feedforward neural network called a "database module". In the study, the PIR database is divided into four sets and four neural networks called database modules are trained for each of the ten encoding schemes. All trained neural networks have a single hidden layer fixed with 200 neurons, but the number of inputs for each module depends on the selected encoding scheme. The total number of outputs for all four neural networks is 620 corresponding to 620 protein superfamilies. Moreover, each neural network is trained to classify a range of superfamilies. The number of superfamilies classified by the four database modules are 148, 157, 178 and 137. Each input neuron is fed with a real value, computed by the product of the n-gram count and the corresponding residue's frequency in nature. Then, the product is mapped to the range [0,1].

To evaluate classification accuracy, a protein sequence is classified by all four database modules. Therefore, among 620 classification scores from 0 (no match) to 1.0 (perfect match), the superfamilies of the three highest scores are selected as the predicted superfamilies of the protein. The classification accuracy is expressed by the total number of correctly identified patterns (superfamilies), the total number of incorrectly identified patterns and the total number of unidentified patterns. A protein's superfamily is considered correctly classified if one of the best three scores is above a defined threshold value and matches the known superfamily number of the protein. The predictive accuracy is examined by comparing threshold values ranging from 0.01 to 0.9. By choosing a lower threshold value, more superfamilies (patterns) are identified which results a higher sensitivity (more true positives) and a lower specificity (more false positives). In contrast, a higher specificity and a lower sensitivity are achieved if a higher threshold value is chosen. The experimental results reached a 90% classification accuracy with 9% false positives.

At a threshold of 0.9 the classification accuracy decreases to 68% with zero percent false positive. Also, the results indicate that the best encoding scheme is the concatenation of 1-gram (single letter) amino acids and 2-gram protein exchange groups.

3.2 Conclusion

In this section, we have examined approaches that begin with amino acid sequences and aim to predicting protein structure and function. We have seen that there are many intermediate goals along the way to full 3D structure prediction. We have also noted the danger of overfitting, which is caused by a sparsity of exemplars in a high dimensional space and various methods that are effective at mitigating this problem.

4 Diagnostic Predictions

Even if we knew the functions of proteins or had a tool to predict them, we would still be at a loss to explain many biological processes. This is because most of these processes involve protein-protein interactions and the presence or absence of proteins are varied by specific and complex regulatory mechanisms. Many biological processes can be turned on or off by causing particular genes to be variably expressed under different conditions. The study of gene regulation seeks to understand this variable expression. Variable expression, in turn, can then be used to shed light on the processes going on in an organism. This can provide valuable diagnostic tools for medicine and other applications.

Modern microarray technology uses 2D arrays of short DNA or RNA sequences called “probes” that bind to specific complementary RNA strands found in cells. A microarray may contain several thousands such probes (or even millions with the newest technology). By providing florescently died RNA materials, it is possible to copy the RNA produced in a functioning cell, photograph such an array, and based on the brightness of specific grid points, measure the expression of specific RNA patterns. By building custom designed microarrays for specific genes or interesting gene segments, it is thus possible to capture a snap shot of the proteins being produced in a cell at a given point in time.

An example using this datatype is disease classification and the identification of indicative biomarkers for detecting the early onset of diseases. Deciphering the gene expression signatures for classification of cancerous diseases is challenging for such a high dimensional and complex dataset. In the seminal paper [39], a neural network was used to classify cancers into a number of diagnostic categories based on their gene expression signatures obtained by cDNA microarray analysis. In genetic engineering, a complementary DNA (cDNA) is a type of DNA synthesized from a messenger RNA template in which the introns are removed in order to be able to clone eukaryotic genes in prokaryotes. Selecting an optimal subset of gene markers is an extremely difficult task since there are a high number of gene markers, and these markers can reach over one million with new chip technology.

In [45], using ANNs resulted in reducing the gene signatures from 70 to 9 which accurately together predicted breast cancer from microarray data.

In [31], an unsupervised artificial neural network was used to analyze gene expression data obtained from DNA array experiments for correlated gene expression patterns. Clustering techniques are common tools applied to gene expression data. However, the proposed method uses a growing neural network which adopts the topology of a binary tree. The outcome is a hierarchical clustering that also has the robustness of a neural network. With regards to the rapidly growing DNA array technology and the huge amount of information, the proposed method is claimed to be faster and more accurate compared to the other hierarchical clustering techniques [64], [23], and [32].

In proteomics one of the pioneering applications of artificial neural networks was in the mining of the mass spectrometry data for the protein screening of cancer patients [8]. In the study, the ANN's weights were analyzed and the ions that had the highest contribution for the classification were identified. By further analyses, it was discovered that two ions in combination are able to predict the tumour grade with the highest accuracy.

4.1 Example Application

An application of ANNs for classification of gene expression data for disease diagnosis is explained in the following. DNA microarrays can be used to simultaneously measure the expression levels of large numbers of genes. As a result, gene expression data has become an effective tool in clinical purposes such as disease diagnosis. In order to develop a reliable technique to diagnose a disease, the bodies response can be measured via gene expression patterns in individual cells. Machine learning can then be used to identify and classify specific expression patterns and thereby label healthy and sick cells. Such patterns may include finding co-regulated genes. In this regard, the greatest challenge of this work is to build accurate classification models that are capable of processing the large amount of gene expression data consisting of thousands features (genes) and a few number of samples.

An essential component in mining such high dimensional data for key features is a robust feature selection technique. In [48], a combinatorial feature selection method and an ensemble neural network are used to classify gene expression data. As mentioned earlier, due to the limited samples of the gene expression data, the bootstrap method is used to resample the data 100 times. This increase of training data is necessary to ensure the accuracy, robustness and generalization of a classifier.

It has been verified that different feature selection techniques applied to a dataset result in different profiles for the dataset. Therefore, the combinatorial feature selection method provides more information for a classifier by employing three feature selection methods: ranksum test [69], principle component analysis(PCA) [57] and masked out clustering [33]. The ranksum test extracts and selects 30 top genes; the PCA method selects 15 principle components; and the masked out clustering clusters the data into 50 groups and then, using the t-test, selects the 30 top genes.

The proposed neural network ensemble consists of three feed-forward neural networks. Each competitive and cooperative neural network of the ensemble network has a single hidden layer with 10 units. The selected features produced by the ranksum, PCA and masked out clustering methods are separately fed as the inputs to each neural network in the ensemble network. Each neural network learns to classify based on the information extracted from the training data. The partial classification accuracies generated by all networks are aggregated according to a soft-voting scheme where the confidence of each network is considered as a voting value instead of the crisp values of 0 and 1.

An advantage of the proposed classifier is that its overall classification accuracies were higher or comparable to those of the other standard techniques on seven different types of gene expression datasets (Lung Cancer, DLBCL, etc.) due to combining the information extracted by the three different feature selection techniques.

4.2 Conclusion

In this section we have explored even higher levels of analysis, relating to the function and expression of genes. This analysis enables new diagnostic tools for diseases. Neural networks have been successfully used to identify markers and patterns in these cases, as well as feature selection.

5 Conclusion

In this chapter we have pursued the flow of information from the basic hereditary patterns in DNA to RNA to protein to functionality, and finally biological processes. We have reviewed a large number of neural network approaches to processing the data and making useful predictions. In completing any review such as this we have had to purposely omit some works. In general we have tried to give a broad overview of the scope of the field rather than focusing on variations on specific approaches. At the same time, there are constantly new developments in dealing with the rich and voluminous data being produced by modern molecular techniques, and these provide a great opportunity for future work. While we have tried our best to tie together in a logical way the diverse work covered in this survey using an information flow approach, there is one additional theme that resurfaces in our observations on the work. That is the challenge of working in high dimensional spaces, where the number of example patterns (although large and growing) is still fairly small compared to the dimensionality of the spaces to be considered. This creates a problem for any adaptive technique in the form of a danger of overfitting parameters to the data [6]. Regularization methods must be employed to manage this issue. Throughout this chapter we have pointed out some such methods which rely on input encoding, recurrent networks, alignment profiles, homologies and ensemble averages. We expect this to remain a challenge as datasets continue to become richer and be a prevalent theme of work in this area in the next decade.

Acknowledgement. Dr. Stefan C. Kremer is funded by the NSERC of Canada and uses that funding to support his students, including Mr. Masood Zamani.

References

1. Adamczak, R., Porollo, A., Meller, J.: Accurate prediction of solvent accessibility using neural networks-based regression. Proteins: Structure, Function, and Bioinformatics 56(4), 753–767 (2004)
2. Ahmad, S., Gromiha, M.M., Sarai, A.: Real value prediction of solvent accessibility from amino acid sequence. Proteins: Structure, Function, and Bioinformatics 50(4), 629–635 (2003)
3. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology 215(3), 403–410 (1990)
4. Arrigo, P., Giuliano, F., Scalia, F., Rapallo, A., Damiani, G.: Identification of a new motif on nucleic acid sequence data using Kohonen's self-organizing map. Computer Applications in the Biosciences: CABIOS 7(3), 353 (1991)
5. Auray, J.P., Duru, G., Zighed, D.A.: Analyse des données multidimensionnelles: Les Méthodes de structuration. A. Lacassagne (1990)
6. Bai, B., Kremer, S.C.: In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops (2011) (to appear)
7. Baldi, P., Brunak, S., Frasconi, P., Pollastri, G., Soda, G.: Bidirectional dynamics for protein secondary structure prediction. Sequence Learning, 80–104 (2001)
8. Ball, G., Mian, S., Holding, F., Allibone, R.O., Lowe, J., Ali, S., Li, G., McCardle, S., Ellis, I.O., Creaser, C., et al.: An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. Bioinformatics 18(3), 395–404 (2002)
9. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.: et al. The protein data bank: A computer-based archival file for macromolecular structures. Journal of Molecular Biology 112(3), 535–542 (1977)
10. Bland, C., Newsome, A., Markovets, A.: Promoter prediction in e. coli based on SIDD profiles and artificial neural networks. BMC Bioinformatics 11(suppl. 6), S17 (2010)
11. Bloomer, A.C., Champness, J.N., Bricogne, G., Staden, R., Klug, A.: Protein disk of tobacco mosaic virus at 2.8 a resolution showing the interactions within and between subunits. Nature 276(5686), 362 (1978)
12. Brunak, S., Engelbrecht, J., Knudsen, S.: Prediction of human mRNA donor and acceptor sites from the DNA sequence. Journal of Molecular Biology 220(1), 49–65 (1991)
13. Cheng, J., Baldi, P.: Three-stage prediction of protein-sheets by neural networks, alignments and graph algorithms. Bioinformatics 21(suppl. 1), i75–i84 (2005)
14. Cherkassky, V., Vassilas, N.: Performance of back propagation networks for associative database retrieval. In: International Joint Conference on Neural Networks, IJCNN, pp. 77–84. IEEE (1989)
15. Cheung, M., Fogel, G.B.: Identification of functional RNA genes using evolved neural networks. In: Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2005, pp. 1–7. IEEE (2005)
16. Choe, W., Ersoy, O.K., Bina, M.: Neural network schemes for detecting rare events in human genomic DNA. Bioinformatics 16(12), 1062–1072 (2000)
17. Cox, D.R., Snell, E.J.: Analysis of binary data, vol. 32. Chapman & Hall/CRC (1989)
18. Crick, F.H.: On protein synthesis. In: Symposia of the Society for Experimental Biology, vol. 12, p. 138 (1958)

19. Dayhoff, M.O., McLaughlin, P.J., Barker, W.C., Hunt, L.T.: Evolution of sequences within protein superfamilies. Naturwissenschaften 62(4), 154–161 (1975)
20. Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P., Rooman, M.: Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks. Bioinformatics 25(19), 2537–2543 (2009)
21. Draghici, S., Potter, R.B.: Predicting HIV drug resistance with neural networks. Bioinformatics 19(1), 98–107 (2003)
22. Dyrløv Bendtsen, J., Nielsen, H., von Heijne, G., Brunak, S.: Improved prediction of signal peptides: Signalp 3.0. Journal of Molecular Biology 340(4), 783–795 (2004)
23. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95(25), 14863 (1998)
24. Eisenbeis, R.A., Avery, R.B.: Discriminant analysis and classification procedures: theory and applications. Lexington Books (1972)
25. Fariselli, P., Finocchiaro, G., Casadio, R.: Speplip: the detection of signal peptide and lipoprotein cleavage sites. Bioinformatics 19(18), 2498 (2003)
26. Ferrán, E.A., Ferrara, P.: Clustering proteins into families using artificial neural networks. Computer Applications in the Biosciences: CABIOS 8(1), 39–44 (1992)
27. Fletcher, C.M., Wagner, G.: The interaction of eif4e with 4e-bp1 is an induced fit to a completely disordered protein. Protein Science 7(7), 1639–1642 (1998)
28. Guo, J., Xu, D., Kim, D., Xu, Y.: Improving the performance of domainparser for structural domain partition using neural network. Nucleic Acids Research 31(3), 944–952 (2003)
29. Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H., Hallwachs, W.: Ten species in one: Dna barcoding reveals cryptic species in the neotropical skipper butterfly astraptes fulgerator. Proceedings of the National Academy of Sciences of the United States of America 101(41), 14812 (2004)
30. Helles, G., Fonseca, R.: Predicting dihedral angle probability distributions for protein coil residues from primary sequence using neural networks. BMC Bioinformatics 10(1), 338 (2009)
31. Herrero, J., Valencia, A., Dopazo, J.: A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics 17(2), 126–136 (2001)
32. Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson, J., Boguski, M.S., et al.: The transcriptional program in the response of human fibroblasts to serum. Science 283(5398), 83 (1999)
33. Jager, J., Sengupta, R., Ruzzo, W.L.: Improved gene selection for classification of microarrays. In: Pacific Symposium on Biocomputing 2003, Kauai, Hawaii, January 3-7, p. 53. World Scientific Pub. Co. Inc. (2002)
34. Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 292(2), 195–202 (1999)
35. Kanhere, A., Bansal, M.: Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. Nucleic Acids Research 33(10), 3165 (2005)
36. Kaur, H., Raghava, G.P.S.: A neural network method for prediction of β-turn types in proteins using evolutionary information. Bioinformatics 20(16), 2751–2758 (2004)
37. Keil, M., Exner, T.E., Brickmann, J.: Pattern recognition strategies for molecular surfaces: III. binding site prediction with a neural network. Journal of Computational Chemistry 25(6), 779–789 (2004)
38. Keşmir, C., Nussbaum, A.K., Schild, H., Detours, V., Brunak, S.: Prediction of proteasome cleavage motifs by neural networks. Protein Engineering 15(4), 287–296 (2002)

39. Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., et al.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 7(6), 673–679 (2001)
40. Kimoto, T., Asakawa, K., Yoda, M., Takeoka, M.: Stock market prediction system with modular neural networks. In: International Joint Conference on Neural Networks, vol. 1, pp. 1–6. IEEE (1990)
41. Kohonen, T.: Self-organization and associative memory. In: Self-Organization and Associative Memory, 100 figs. XV, 312 pages. Springer Series in Information Sciences, vol. 8, p. 1. Springer, Heidelberg (1988)
42. Kozobay-Avraham, L., Hosid, S., Bolshoy, A.: Involvement of DNA curvature in intergenic regions of prokaryotes. Nucleic Acids Research 34(8), 2316 (2006)
43. Kuang, R., Leslie, C.S., Yang, A.S.: Protein backbone angle prediction with machine learning approaches. Bioinformatics 20(10), 1612 (2004)
44. Lac, H., Kremer, S.: Inducing fold dynamics from known protein structures using machine learning. PhD thesis, CIS, University of Guelph (April 2009)
45. Lancashire, L.J., Powe, D.G., Reis-Filho, J.S., Rakha, E., Lemetre, C., Weigelt, B., Abdel-Fatah, T.M., Green, A.R., Mukta, R., Blamey, R., et al.: A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks. Breast Cancer Research and Treatment 120(1), 83–93 (2010)
46. Li, X., Romero, P., Rani, M., Dunker, A.K., Obradovic, Z.: Predicting protein disorder for N-, C-, and internal regions. Genome Informatics Series, 30–40 (1999)
47. Lin, K., Simossis, V.A., Taylor, W.R., Heringa, J.: A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics 21(2), 152–159 (2005)
48. Liu, B., Cui, Q., Jiang, T., Ma, S.: A combinational feature selection and ensemble neural network method for classification of gene expression data. BMC Bioinformatics 5(1), 136 (2004)
49. Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure 405(2), 442–451 (1975)
50. Mooney, C., Pollastri, G.: Beyond the twilight zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information. Proteins: Structure, Function, and Bioinformatics 77(1), 181–190 (2009)
51. Plewczynski, D., Slabinski, L., Ginalski, K., Rychlewski, L.: Prediction of signal peptides in protein sequences by neural networks. Acta Biochimica Polonica 55(2), 261–267 (2008)
52. Qian, N., Sejnowski, T.J.: Predicting the secondary structure of globular proteins using neural network models. Journal of Molecular Biology 202(4), 865–884 (1988)
53. Rost, B., Sander, C.: Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proceedings of the National Academy of Sciences of the United States of America 90(16), 7558–7562 (1993)
54. Rost, B., Sander, C.: Combining evolutionary information and neural networks to predict protein secondary structure. Proteins-Structure Function and Genetics 19(1), 55–72 (1994)
55. Sidman, K.E., George, D.G., Barker, W.C., Hunt, L.T.: The protein identification resource (PIR). Nucleic Acids Research 16(5), 1869 (1988)
56. Song, D., Deng, Z.: A novel ncRNA gene prediction approach based on fuzzy neural networks with structure learning. In: 2010 4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE), pp. 1–5. IEEE (2010)
57. Speed, T.P.: Statistical analysis of gene expression microarray data. CRC Press (2003)

58. Tegge, A.N., Wang, Z., Eickholt, J., Cheng, J.: NNcon: improved protein contact map prediction using 2d-recursive neural networks. Nucleic Acids Research 37, w515–w518 (2009)
59. Vullo, A., Frasconi, P.: Disulfide connectivity prediction using recursive neural networks and evolutionary information. Bioinformatics 20(5), 653–659 (2004)
60. Walsh, I., Martin, A.J.M., Mooney, C., Rubagotti, E., Vullo, A., Pollastri, G.: Ab initio and homology based prediction of protein domains by recursive neural networks. BMC Bioinformatics 10(1), 195–214 (2009)
61. Wang, H., Noordewier, M., Benham, C.J.: Stress-induced DNA duplex destabilization (SIDD) in the e. coli genome: Sidd sites are closely associated with promoters. Genome Research 14(8), 1575 (2004)
62. Wang, J.T.L., Ma, Q., Shasha, D., Wu, C.H.: Application of neural networks to biological data mining: a case study in protein sequence classification. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 305–309. ACM (2000)
63. Weinreb, P.H., Zhen, W., Poon, A.W., Conway, K.A., Lansbury Jr., P.T.: NACP, a protein implicated in alzheimer's disease and learning, is natively unfolded. Biochemistry 35(43), 13709–13715 (1996)
64. Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L., Somogyi, R.: Large-scale temporal gene expression mapping of central nervous system development. Proceedings of the National Academy of Sciences 95(1), 334 (1998)
65. Wright, P.E., Dyson, H.J.: Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. Journal of Molecular Biology 293(2), 321–331 (1999)
66. Wu, C., Whitson, G., Mclarty, J., Ermongkonchai, A., Chang, T.C.: Protein classification artificial neural system. Protein Science: A Publication of the Protein Society 1(5), 667 (1992)
67. Zamani, M., Chiu, D.: An evaluation of DNA barcoding using genetic programming-based process. Life System Modeling and Intelligent Computing, 298–306 (2010)
68. Zhang, A.B., Sikes, D.S., Muster, C., Li, S.Q.: Inferring species membership using DNA sequences with back-propagation neural networks. Systematic Biology 57(2), 202–215 (2008)
69. Ziegel, E.R.: Probability and statistics for engineering and the sciences. Technometrics 46(4), 497–498 (2004)
70. Zimmermann, O., Hansmann, U.H.E.: Support vector machines for prediction of dihedral angle regions. Bioinformatics 22(24), 3009 (2006)