Spectroscopy 16 (2002) 53–69
IOS Press

An alternative method for rapid quantification of protein secondary structure from FTIR spectra using neural networks

Joachim A. Hering [a], Peter R. Innocent [b] and Parvez I. Haris [a,*]

[a] *Department of Biological Sciences, De Montfort University, The Gateway, Leicester, LE1 9BH, UK*
[b] *Department of Computer Science, De Montfort University, The Gateway, Leicester, LE1 9BH, UK*

Abstract. Lack of reliable methods for accurate estimation of protein secondary structure from infrared spectra of proteins is a major barrier in its widespread use in protein secondary structure characterisation. Here we report a method for protein secondary structure estimation, from FTIR spectra of proteins, based on a multi-layer feed-forward neural network approach using an enhanced "resilient backpropagation" learning algorithm. The method utilises a database consisting of infrared spectra of 18 proteins, with known X-ray structure, as the reference set. Our study revealed that providing the neural network analysis with only part of the amide I region from empirically determined structure sensitive regions in combination with appropriate pre-processing of the spectral data produced the best overall results. This lead to a standard error of prediction (SEP) of 4.47% for α-helix, an SEP of 6.16% for β-sheet, and an SEP of 4.61% for turns. Compared to a previous factor analysis study by Lee et al., using the same set of 18 FTIR spectra of proteins, the error in prediction of α-helix and β-sheet was improved by 3.33% and 3.54% respectively, with minor increase for turns by 0.31%. Generally, our neural network analysis achieved comparable, in most cases even better prediction accuracy than most of the alternative pattern recognition based methods that were previously reported indicating the significant potential of this approach.

Keywords: Protein secondary structure prediction, FTIR spectroscopy, neural networks, resilient backpropagation, boxcar averaging

1. Introduction

Fourier transform infrared (FTIR) spectroscopy has been demonstrated to be a very useful technique for rapid characterisation of protein secondary structure [1–6]. In matter of minutes, measurements on small quantities of proteins can be carried out in solution, or in other environments. However, for reliable predictions to be made from FTIR spectra of proteins, further improvements in the analysis and interpretation of FTIR data is necessary. This includes development and improvement of methods that can be used for accurately quantifying protein secondary structure. Development of such methods is particularly important and timely since the sequencing of the human genome has just been completed, and the focus is now on protein structure–function relationships. The work done in quantitative estimation of protein secondary structure based on FTIR spectra along with its potentials and pitfalls has been thoroughly reviewed [7–12]. To date, the existing methods to estimate the fractions of protein secondary structural conformations from FTIR spectral data fall into two main categories: Those based on band narrowing and decomposition of mainly the amide I band shape into its underlying components often referred to as frequency-based or curve fitting approaches and those based on the principle of pattern recognition.

*Corresponding author. Tel.: +44 116 250 6306; Fax: +44 116 257 7287; E-mail: pharis@dmu.ac.uk.

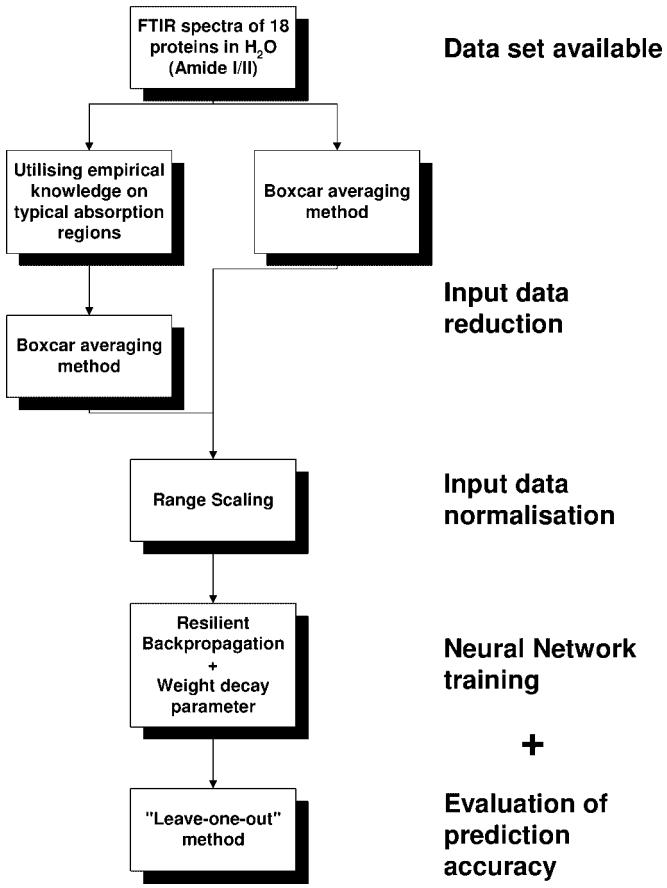

Fig. 1. An overview of the techniques used in the present study. The main aim was to introduce methods to reduce the weight connections in the neural networks to facilitate generalisation and hence prediction accuracy.

Curve fitting is probably the most commonly applied method for protein secondary structure quantification based on FTIR spectral data [1,3,10,13–26]. Alternative methods for quantitative analysis of protein secondary structure have been introduced that require fewer assumptions and remove much of the subjectivity inherent in the curve fitting method. These methods can be generally referred to as "pattern recognition based" approaches [27–38] where multivariate data analysis techniques are by far the most widely applied techniques of this category.

All of the currently used methods for protein secondary structure prediction from FTIR spectra have varying degrees of advantages and disadvantages and neither curve fitting methods nor pattern recognition based methods are free of problems and further improvements in this field are necessary to achieve increased quality of prediction accuracy [8,39].

More recently, neural network approaches have been explored to allow predictions about protein secondary structure conformations. However, they have been most widely used for secondary structure prediction from amino acid sequences [40–43] as well as from CD spectral data [38,44,45]. Keiderling, Pancoska and co-workers have used neural network analysis for obtaining protein structural information from infrared spectra of proteins [36–38]. More recently, one of us has reported a different neural network based approach for predicting protein secondary structure from FTIR spectra [27].

A schematic overview of the techniques involved in our neural network analysis is shown in Fig. 1. Neural networks are also often referred to as "classifier of inputs" or "black box approaches" since they can learn to map certain input values to certain output values. Here, a neural network is trained to learn a mapping from FTIR spectral data of proteins to their secondary structure fractions. The present study is based on the same reference set of 18 FTIR spectra as in a previous factor analysis study by Lee et al. [35]. This allows direct comparison of the results achieved. Finally, representative work done in the field of protein secondary structure prediction based on FTIR spectra and that achieved by our neural network approach is compared.

2. Materials and methods

2.1. The spectral data set

We have used the same set of protein infrared spectra as in Lee et al.'s paper [35] where details regarding sample preparation and FTIR measurements can be found: alcohol dehydrogenase (equine liver), trypsin inhibitor (bovine pancreas), carbonic anhydrase (bovine erythrocyte), concanavalin A (Jack Bean), chymotrypsinogen A (bovine pancreas), chymotrypsin (bovine pancreas), cytochrome c (equine heart), elastase (porcine pancreas), hemoglobin (bovine erythrocyte), insulin (porcine pancreas), lysozyme (chicken egg white), myoglobin (sperm whale skeletal muscle), nuclease (*Staphylococcus aureus*), prealbumin (human plasma), papain (*Papaya latex*), alkaline serine protease B (*Streptomyces griseus*), ribonuclease A (bovine pancreas), and ribonuclease S (bovine pancreas). We have taken the same protein structure quantities based on the work from X-ray crystallography studies by Levitt and Greer [46]. Table 1 shows a simple statistical characterisation of the distribution for each of these secondary structure conformations across these 18 proteins.

2.2. The software tools used

The present study makes use of the Stuttgart Neural Network Simulator (SNNS), Version 4.2, which is a complex simulator for neural networks developed since 1989 at the Institute for Parallel and Distributed High Performance Systems ("Institut für Parallele und Verteilte Höchstleistungsrechner", IPVR) at the University of Stuttgart, Germany.

2.3. The "leave-one-out" method

Given the limited number of 18 FTIR spectra from proteins in H_2O used in our study, the "leave-one-out" method was employed for validation where each protein, in turn, is eliminated from the analysis

Table 1
Distribution of the secondary structural conformations as determined from X-ray crystallography studies by Levitt and Greer [46] for the18 proteins used (in % structure)

Secondary structure	α-helix	β-sheet	Turns
Minimum	3	0	7
Maximum	88	65	28
Mean	31	37	19
Standard deviation	26	20	6

and a training set generated from the remainder is used to predict its secondary structure [47]. This method has also been used by Lee et al. [35] and others in the area of FTIR protein secondary structure prediction [30–33,48].

2.4. Resilient backpropagation: A locally adaptive neural network learning scheme

Since the introduction of the backpropagation algorithm [49] there have been several suggestions to improve weight training in feed-forward neural networks based on the concept of supervised learning in multi-layer perceptrons using the technique of gradient-descent.

The present study employs an enhanced resilient backpropagation learning algorithm where a weight decay parameter has been added to the error function for improved prediction accuracy [50]. Resilient backpropagation makes use of a local adaptation strategy to improve the conventional backpropagation learning technique where a harmful influence of the size of the partial derivative on the weight step can be observed. In order to eliminate this potential problem, in resilient backpropagation only the sign of the derivative is considered to indicate the direction of the weight update. The exact mathematical formulation of the resilient backpropagation learning algorithm employed is given elsewhere [50]. In all our experiments, the resilient backpropagation learning algorithm is configured with the following parameter settings: The increase and decrease factors are set to fixed values of 1.2 and 0.5, respectively [51]. The weight decay exponent is set to 4.0.

The danger of possible overfitting during neural network training was considerably diminished by using the resilient backpropagation learning algorithm. Therefore, a stop criterion could be safely defined merely on the basis of a predefined number of epochs and a predefined minimum limit for the sum of squared errors (SSE) of the training patterns. In our studies, neural network training is stopped when either the number of epochs exceeds 2000 or the SSE falls below 0.07. A relatively high SSE of 0.07 has been chosen as stop criterion to facilitate generalisation.

2.5. The characteristics of the neural network

We used multi-layer feed-forward neural network topologies, which are trained with the resilient backpropagation learning algorithm described above. We investigated using the entire spectral data of the amide I/amide II regions for neural network training where the number of input units of the neural network corresponds directly to the data points of the respective amide region(s). In this case, the absorbance values for each wavenumber were used as inputs to the neural network. In subsequent experiments, data compression was applied in order to reduce the number of inputs to the neural network (see Section 2.7).

Each of the three outputs of the network corresponds to the fraction of α-helix, β-sheet, and turns respectively. A sigmoidal activation function is used. Hence, the output values fall in the region between 0.0 and 1.0, where 0.10 for example means a fraction of secondary structure of 10 percent.

In order to keep the number of connection weights as low as possible while still maintaining the neural network's ability to approximate secondary structural fractions from the shapes of the spectral data, a neural network topology with one hidden layer containing two hidden units is used.

2.6. Data pre-processing

Before presenting the spectral data to the neural network, it is important to pre-process it in a way suited to the problem at hand in order to facilitate neural network training. Normalisation of the spectral region under investigation is commonly used in protein secondary structure prediction techniques based

on FTIR spectra [27,29,30,33–35]. In our studies we decided to employ range scaling normalisation. Range scaling transforms the spectral data to fall between 0 and 1. Using range scaling on a data set which contains outliers or an uneven spread of values results in a large proportion of the data values being squashed into a small part of the input range, leaving most of the rest unused. Being aware of this problem, the data set was tested for the presence of outliers. Since our data set contains no outliers, range scaling could be safely applied.

2.7. Input data reduction

One important factor for good prediction accuracy in neural networks is the number of inputs and hence the number of weight connections. The neural network should have a considerably low number of inputs to facilitate generalisation [52]. In the current study this was achieved by employing a simple but efficient input data reduction technique, namely, the boxcar averaging method. Recently, improved predictions were reported using boxcar averaging applied to the optimisation of functional group predictions from infrared spectra using neural networks [53]. The boxcar averaging method takes a predefined number of data points at a time and replaces them by their average. The main aim of employing the boxcar averaging technique is to reduce the number of input data points while preserving the overall features of the FTIR spectra. Boxcar averaging has the additional effect to smooth possible noise present in the spectral data.

Apart from a variation in the number of input units, the same neural network settings as described above were used. The spectral data of the 18 protein FTIR spectra was processed in the following way: The absorbance values of the spectral region under investigation were compressed using the boxcar averaging method. Subsequently, the resulting data was normalised using range scaling normalisation as described above. The number of data points averaged at a time was varied from 2 to 10. However, only the number of data points averaged at a time achieving the best results are reported.

3. Results and discussion

By far the most widely used spectral region for protein secondary structure quantification from FTIR spectral data is the amide I band although the amide II and amide III bands are also used [1–3,5,13,23, 30,31,35,54]. Our analysis presented here concentrates primarily on the amide I region and to a lesser extent on the amide II region. Figure 1 gives an overview of the main steps involved in our neural network analysis.

3.1. Prediction using the amide I region alone

Initially, we focused on providing our neural network analysis with data from the amide I region only (1700–1600 cm^{-1}). Different pre-processing techniques were applied to facilitate better neural network training and hence prediction accuracy (see Table 2).

3.1.1. Effect of amide I band normalisation on prediction accuracy

When using infrared spectral data for neural network training, possible variation in path length and absorbance amongst the protein spectra recorded should be taken into account. Therefore, some form of normalisation needs to be applied to the raw spectral data prior to analysis. Lee et al. set the area under the amide I band and the ordinate at 1700 cm^{-1} to constant values by appropriate multiplication and offsetting of the spectral data [35].

Table 2
Standard error of prediction (in % structure) using the original data of the amide I band, range scaling, and boxcar averaging. BOXCAR_5 means that every 5 data input points were replaced by their average. The number of inputs presented to the neural network for each FTIR spectrum as well as the resulting number of weight connections of the neural network are also given

Method used	No. of input units	No. of weight connections	α-helix	β-sheet	Turns	Average
Factor analysis [35]	N/A	N/A	7.8	9.7	4.3	7.27
Orig. data	101	208	14.86	12.28	5.22	10.79
Range scaling	101	208	7.85	9.15	4.68	7.23
BOXCAR_5	21	48	6.78	8.67	4.34	6.59

Table 2 shows the SEPs for the factor analysis approach presented by Lee et al. [35] and the SEPs of our neural network approach using the spectral data of the amide I region with and without the application of range scaling. The results clearly demonstrate that the application of range scaling normalisation to the spectral data of the amide I region prior to training does have an advantageous effect on the resulting prediction accuracy. The average of the SEPs was already improved compared to Lee et al.'s factor analysis approach [35].

Prior to this study, we have explored various normalisation methods which have been thoroughly investigated by Klawun and Wilkins [53] who were looking at the optimisation of functional group predictions from infrared spectra using neural networks. However, we chose to employ range scaling normalisation mainly because it transforms the spectral data (absorbances) to fall in the region between 0 and 1. This is particularly important for our neural network analysis where some input units cover the range of relatively high values and others cover the range of relatively low values. Clearly, errors due to higher value inputs would have a greater effect during neural network training than those errors due to lower value inputs as their magnitude would be greater. Ensuring that every input unit covers the same range also ensures that errors on each input unit contribute the same proportion to the change in network weights. Since range scaling preserves the relative positions of each data point along the range, constant molar absorptivities for each type of structure need not be assumed. Additionally, the overall bandshapes of the FTIR spectra are preserved. The relatively poor prediction accuracy using the original data for neural network training underlines the need for the application of a normalisation method.

3.1.2. Effect of spectral data reduction on prediction accuracy

Various techniques to compress infrared spectral data prior to neural network analysis have been explored recently [53]. The main aim of these techniques is to improve prediction accuracy and facilitate faster neural network training. In the present study, boxcar averaging proved to be the most suitable technique for our study. Table 2 shows that the best results were achieved by replacing every 5 data inputs of the amide I region by their average (BOXCAR_5). Hence, with only 21 data points per FTIR spectrum the lowest average of SEPs (6.59%) was achieved, a better result than that obtained by using the entire spectral data of the amide I region with range scaling alone. It was observed that reducing the amide I region (1700–1600 cm^{-1}) from 101 absorbance values down to only 11 data points per FTIR spectrum using the boxcar averaging technique produced a lower average of SEPs than Lee et al.'s factor analysis approach [35]. The results of subsequent experiments (see Tables 2–4) suggest that the optimal number of data input points to be replaced, by their average, may vary depending on the spectral region under investigation. However, in our study, averaging a number around three data points has shown to be most appropriate. Clearly, further investigation on possibly larger protein FTIR spectral reference sets would be helpful to determine the most appropriate parameter for the boxcar averaging procedure.

Generally, the boxcar averaging method proved to be very useful for significantly reducing the amount of input data while still maintaining its important spectral features for the neural network to make good predictions about the secondary structure of FTIR spectra from unknown proteins. Due to the resulting reduction of weight connections, the neural network is more likely to be forced to generalise despite possible loss of information contained in the input data while still maintaining good performance on the training set. There is a huge variation in the literature suggesting the optimal number of inputs to the neural network and hence the number of weight connections to achieve good generalisation [52,55,56]. Lange and Männer claim that a critical training set size N exists:

$$N \approx \frac{3}{8}w, \tag{1}$$

where $w =$ number of connection weights, with overfitting and bad generalisation below N and no overfitting and good generalisation above N. In our case, for a neural network with 101 input units, 2 hidden units, and 3 output units, the critical training set size would be 78. Since the training set size in the present study using the "leave-one-out" method is limited to 17, a number of weight connections less or equal to 45 would fulfil the criterion in Eq. (1) for good generalisation. This criterion was met for BOXCAR_6 to BOXCAR_10. In our case, even a number of weight connections above 45 resulted in good generalisation. The results in Table 2 show that the boxcar averaging method is highly appropriate to significantly reduce the number of data points of the protein FTIR spectra while still maintaining the overall "features" required to make reasonable predictions about their secondary structure. Particularly with regards to other pattern recognition based approaches, further studies would be helpful to investigate if the application of the boxcar averaging technique will generally lead to an improved prediction accuracy. Since in our neural network analysis the boxcar averaging method did have a beneficial effect on neural network training in terms of generalisation, it is also used in the following.

3.2. Prediction using the amide II region alone

Here, we looked at providing our neural network analysis with the amide II region alone (1600–1500 cm^{-1}) to see how prediction accuracy compares to that achieved by using the amide I region. Table 3 shows the best results achieved by replacing every ten absorbance values by their average as well as the results obtained without boxcar averaging applied. The results clearly indicate, that our neural network approach is in agreement with the empirical findings, that the amide I region is more useful for secondary structure prediction from FTIR spectra than the amide II region. The best average of SEPs

Table 3

Standard error of prediction (in % structure) using the boxcar averaging method based on the amide II region alone as well as those based on the combined amide I and amide II regions. BOXCAR_N means that every N data input points were replaced by their average. The number of inputs presented to the neural network for each FTIR spectrum as well as the resulting number of weight connections of the neural network are also given

Method used	No. of input units	No. of weight connections	α-helix	β-sheet	Turns	Average
Amide II						
Orig. data	101	208	16.7	15.88	5.26	12.61
BOXCAR_10	11	28	17.18	13.88	5.68	12.25
Amide I and II combined						
Orig. data	201	408	9.64	10.38	4.05	8.02
BOXCAR_3	67	140	8.92	9.34	4.78	7.68

that was reached using the amide II region is 12.25% which is almost twice as high as the SEP achieved using the amide I region (6.59%). We can therefore confirm, that the amide II region does in fact contain less features relating to secondary structure that can be picked up by our neural network than the amide I region.

3.3. Prediction using the amide I and amide II regions together

Table 3 shows the results when both the amide I and amide II region (1700–1500 cm^{-1}) are used for neural network analysis. The results achieved when replacing every three absorbance values by their average as well as the results obtained without boxcar averaging are shown in Table 3. Although the best average of SEPs of 7.68% is significantly better than the average of SEPs that was achieved using the amide II region alone, it is still not better than the results obtained by merely using the amide I region. Using our neural network approach, we could therefore not confirm the observation made by others using multivariate data analysis techniques, where combination of the amide I and amide II bands did increase the accuracy of secondary structure prediction [24,29,32,34]. Other than the methodological difference, a possible explanation for this discrepancy may be that different sets of protein infrared spectra have been used. Different proteins will have different secondary and primary structures and this will influence the prediction quality of different algorithms.

3.4. Prediction utilising empirical knowledge about the amide I region

Empirical studies have identified certain frequencies within the amide I band, which are particularly sensitive to secondary structural conformation [1–6]. We have directed our focus on structure sensitive regions only within the amide I band because this region has been explored more thoroughly.

Table 4 shows the best results obtained using structure sensitive regions within the amide I band, namely the intervals 1625–1635 cm^{-1}, 1649–1659 cm^{-1}, and 1675–1685 cm^{-1} in combination with the boxcar averaging method replacing every 3 data points by their average. An SEP of 4.47% for α-helix, an SEP of 6.16% for β-sheet, and an SEP of 4.61% for turns was achieved. The resulting average of SEPs is 5.08%.

The majority of proteins in H_2O of known structure have been found to display infrared absorption in the range from 1650 to 1658 cm^{-1} for α-helical conformation, from 1620 to 1640 cm^{-1} for β-sheet structure. For turns, the spectral region is less well determined. However, absorption around 1680 cm^{-1} can be assigned to turns [1–6,57].

Table 4

Standard error of prediction (in % structure) using the boxcar averaging method based on empirically determined structure sensitive regions (1625–1635 cm^{-1}, 1649–1659 cm^{-1}, 1675–1685 cm^{-1}) and insensitive regions (1600–1609 cm^{-1}, 1696–1700 cm^{-1}). BOXCAR_N means that every N data input points were replaced by their average. The number of input data units presented to the neural network for each FTIR spectrum as well as the resulting number of weight connections of the neural network are also given

Method used	No. of input units	No. of weight connections	α-helix	β-sheet	Turns	Average
Structure sensitive regions						
Orig. data	33	72	5.46	7.37	4.62	5.82
BOXCAR_3	11	28	4.47	6.16	4.61	5.08
Structure insensitive regions						
Orig. data	15	36	33.92	25.04	6.87	21.94
BOXCAR_3	5	16	29.07	21.94	6.07	19.03

In this "specific amide I frequency based analysis", we were interested in whether knowledge about structure sensitive spectral regions leads to further improvement of protein secondary structure prediction accuracy. Thus, based on these empirical findings, we determined one characteristic wavenumber for each secondary structural conformation under investigation. For the α-helix and β-sheet structures, we took the central band frequencies of the given wavenumber ranges resulting in 1654 cm^{-1} and 1630 cm^{-1}, respectively. For turns, we took the wavenumber 1680 cm^{-1}. Subsequently, we performed a series of "leave-one-out" runs with varying interval sizes around the three determined characteristic wavenumber positions. Intervals ranging from ± 1 wavenumbers up to ± 10 wavenumbers around the three determined characteristic wavenumber positions were extracted from the original FTIR spectra. These regions were subsequently used for neural network analysis both with and without boxcar averaging applied prior to analysis. Best results were obtained by using an interval of ± 5 data points around the three characteristic wavenumbers as determined above (Table 4). Clearly, further investigations based on FTIR spectra from a larger number of proteins would be necessary to determine the optimal positions and widths of structure sensitive regions within the amide I band. However, based on the 18 protein FTIR spectra, the current study successfully demonstrates the power of incorporating empirical knowledge about structure sensitive regions of protein infrared spectra for quantifying their secondary structure. By utilising empirical knowledge on structure sensitive regions in combination with the boxcar averaging data compression technique, significantly fewer data points (i.e., only 11 data points per infrared spectrum) had to be provided for our neural network analysis even leading to an improved prediction accuracy. As a result, neural network training is considerably faster which will become increasingly important with a growing number of infrared spectra in the reference set.

Table 5 shows the prediction accuracy, i.e., the average of absolute differences (in % structure) between target output as determined by X-ray crystallography and predicted output over the secondary structure classes of interest for each protein in the reference set. Based on the same set of protein spectra, prediction accuracy is shown for Lee et al.'s factor analysis method [35], Severcan et al.'s neural network method [27] and our current neural network method. In our study, best prediction was observed for ribonuclease A (1.29%) with poorest prediction for lysozyme (8.68%). Severcan et al.'s neural network approach is in good agreement with our results where the best prediction was observed for both ribonuclease A and papain (2%) with poorest prediction for lysozyme (11%). However, for Lee et al.'s factor analysis approach [35], best prediction was achieved for alcohol dehydrogenase (1%) with poorest prediction for trypsin inhibitor (14%). This discrepancy may reflect the difference between the relatively similar neural network approaches and the more mathematical factor analysis approach. Interestingly, good correlation of prediction accuracy between our study and Severcan et al.'s study can be observed (0.8) whereas prediction accuracy between our study and Lee et al.'s study shows very weak correlation (0.21).

Another interesting observation could be made. One would expect good predictions to be made about the real protein secondary structure contents for a new protein, if it is similar to the overall secondary structure composition (i.e., the average of known secondary structure contents) of proteins in the reference set. However, in all of the latter studies, no correlation between the prediction accuracy for each protein left out and its deviation of average secondary structure contents from the mean for all proteins of the reference set (see Table 1) was observed. The calculated correlation coefficients were 0.34, 0.09, and -0.14 for Lee et al.'s study, Severcan et al.'s study, and our study, respectively. E.g., very good predictions were made for chymotrypsin (see Table 5). However, relatively high absolute differences of secondary structure contents (in % structure) of 19.78, 12.89, and 5.89 for α-helix, β-sheet, and turns respectively, were observed from the average of secondary structure contents of all proteins (see Table 1).

Table 5

Averages of absolute differences (in % structure) between known X-ray data and predictions made for secondary structures are listed for each protein left out. Results are shown from Lee et al.'s factor analysis study [35], Severcan et al.'s neural network study [27] and our specific amide I frequency based analysis

Protein	Average of absolute differences[c] (% structure)		
	Lee et al.	Severcan et al.	Our study
Alcohol dehydrogenase	1	3.33	2
Trypsin inhibitor	14	9	5.01
Carbonic anhydrase	7	7	7.28
Concanavalin A	8	4.33	1.87
Chymotrypsinogen A	4.33	4.33	3.41
Chymotrypsin	1.67	2.67	2.2
Cytochrome c	5	7	6.43
Elastase	4.33	5.67	5.49
Hemoglobin	10	7	3.46
Insulin	5	6.33	4.42
Lysozyme	8	11	8.68
Myoglobin	6.67	3.67	2.36
Nuclease	2.67	5.33	3.99
Prealbumin	10.67	4	2.93
Papain	4.67	2	4.21
Alkaline serine protease B	5.33	5.33	5.74
Ribonuclease A	4	2	1.29
Ribonuclease S	5	6.67	6
Standard deviation	3.26	2.38	2.05
Correlation coefficients	0.21[a]	0.8[b]	

[a]Correlation between Lee et al.'s predictions and those from our study.
[b]Correlation between Severcan et al.'s predictions and those from our study.
[c] Only the average of absolute differences over the structural properties of interest is shown.

Table 5 also shows the standard deviation of average absolute differences between target and predicted outputs over all proteins in the reference set. The relatively low standard deviation of our method (2.05) suggests better reliability of prediction compared to Lee et al.'s factor analysis approach (3.26).

To further confirm the findings of the empirical studies on structure sensitive regions within the amide I band, we subsequently used data points, which are not recognised as being particularly sensitive to protein secondary structure, i.e., the regions 1600–1609 cm^{-1} and 1696–1700 cm^{-1} [8]. As expected, the resulting SEPs were significantly higher than those in previous experiments. The best average of SEPs achieved was 19.03% with very poor SEPs for α-helix (29.07%) and β-sheet (21.94%) and a surprisingly good SEP of 6.07% for turns (see Table 4). The good SEP for turns may be related to its relatively weak distribution of target output proportions shown in Table 1. It should be noted that the number of data input points used in this experiment should have been sufficient to produce good results. The best result achieved by our neural network approach is based on merely 11 data input points (see Table 4).

3.5. Possible impact of amino acid side chain absorption within the amide I band

Side chain absorption from certain amino acids has been found to occur within the amide I and amide II regions for proteins in D_2O and proteins in H_2O [58–60]. It has been shown that significant absorption

Table 6
Amino acid proportions (in %) in relation to the overall amino acid composition along with the mean and standard deviation for amino acids that are known to absorb within the amide I band, namely tyrosine, phenylalanine, glutamine, arginine, and lysine

Protein	Tyrosine	Phenylalanine	Glutamine	Arginine	Lysine	Sum
Alcohol dehydrogenase	1.1	4.8	5.6	3.2	8	22.7
Trypsin inhibitor	6.9	6.9	1.7	10.3	6.9	32.7
Carbonic anhydrase	3.1	4.6	4.2	2.7	9.3	23.9
Concanavalin A	3	4.6	2.1	2.5	5.1	17.3
Chymotrypsinogen A	1.6	2.4	4.1	1.6	5.7	15.4
Chymotrypsin	1.7	2.5	4.1	1.2	5.8	15.3
Cytochrome c	3.8	3.8	2.9	1.9	18.1	30.5
Elastase	4.6	1.2	6.2	5	1.2	18.2
Hemoglobin	2.1	5	0.7	2.1	7.8	17.7
Insulin	7.8	5.9	5.9	2	2	23.6
Lysozyme	2.3	2.3	2.3	8.5	4.7	20.1
Myoglobin	2	3.9	3.3	2.6	12.4	24.2
Nuclease	4.7	2	4	3.4	15.4	29.5
Prealbumin	3.9	3.9	0	3.1	6.3	17.2
Papain	9	1.9	6.1	5.7	4.7	27.4
Alkaline serine protease B	4.9	2.7	1.1	4.3	0.5	13.5
Ribonuclease A	4.9	2.4	4.9	3.3	7.3	22.8
Ribonuclease S	4.8	2.4	5.6	3.2	8.8	24.8
Mean	4.01	3.51	3.6	3.7	7.22	22.04
Standard deviation	2.21	1.55	1.95	2.39	4.54	5.61

may arise in the amide I region [58–60]. Regarding our reference set of infrared spectra of 18 proteins, we were interested in possible impact of amino acid side chain absorption with regards to our neural network analysis. Hence, for the amino acid sequence of each protein in our reference set we calculated the percentage of amino acids that have been reported to display significant absorbance in the amide I region, namely tyrosine, phenylalanine, glutamine, arginine, and lysine [58,59]. Table 6 shows these percentages along with the mean and standard deviation. Despite a relatively high average proportion of these interfering amino acids in relation to the average number of amino acids present in the 18 proteins of our reference set (22.04%), good prediction accuracy of our neural network approach was achieved. This may be explained by one important property of pattern recognition approaches, i.e., they base their predictions mainly on band shape variation. As a result, our neural network analysis should be at least partially immune to amino acid side chain contributions when they are similar for all proteins of the reference set. However, when amino acid side chain variation is high across the spectral data set, deterioration of prediction accuracy may be expected. This could also be confirmed by others with regards to multivariate data analysis [60]. In our case, the relatively low standard deviation for each interfering amino acid (see Table 6) seemed to be sufficiently low for the neural network not to get confused by its interference within the amide I band. Hence we believe that with the pattern recognition based approaches, contribution from amino acid side chain absorption may generally not have such high impact on prediction accuracy if its variation is considerably low amongst the proteins of the reference set. However, good prediction accuracy may also be explained by the fact that absorbance for each of these amino acid side chains individually is rather weak (see Table 6) and that they absorb in different regions within the amide I band. Possibly, side chain absorption from these amino acids is spread in a way such that it does not have high impact on the overall shape of the amide I band. In our neural

network analysis, using only specific structure sensitive regions within the amide I band in combination with compression of the resulting regions using boxcar averaging has the additional effect to cancel out respectively to smooth possible noise in the data arising from amino acid side chain absorption. Clearly, further work is necessary to get a better understanding of the impact of amino acid side chain absorption on prediction accuracy with respect to our neural network analysis in particular and pattern recognition based approaches in general. Although the average of SEPs using our specific amide I frequency based analysis is considerably low (see Table 4), we should certainly not disregard that further improvement may be achieved by appropriate subtraction of side chain absorbance from the amide I region [59,60].

3.6. Comparison of prediction accuracy of our method with other currently established prediction methods

Although the fact that information on protein secondary structure can be derived from FTIR spectra has already been established about 50 years ago [61–63] and a considerable amount of work has been done in that field since, there is still no single method which is commonly agreed upon as the best method for secondary structure prediction from FTIR spectra. Two approaches have mainly evolved over the years: Curve fitting methods and pattern recognition based approaches.

Table 7 shows that our neural network approach competes well with the widely used multivariate data analysis methods employed for protein secondary structure prediction from FTIR spectra. Employing our specific amide I frequency based analysis, the SEP for helix is better than that of any of the listed multivariate data analysis methods and even better than some of the curve fitting approaches. With the exception of Dousseau et al.'s work [34], this is also true for β-sheet structure. Although not the best SEP was achieved for turns, it is only 0.07% above average of the multivariate data analysis techniques listed (4.54%).

In our study the same set of 18 FTIR spectra as in a previous factor analysis study by Lee et al. [35] is used. This allows us to directly compare the two sets of results. Overall, the average of SEPs for α-helix, β-sheet, and turns was reduced by 2.19% by employing our specific amide I frequency based analysis. By removing one FTIR spectrum from the training set, reducing the training set size to 16, Lee et al. [35] achieved an average of SEPs of 6.27%, which is still higher than that achieved by our neural network approach without removing a spectrum from the data set. In our study it was important to us to keep all protein infrared spectra available in the reference set to demonstrate the potential of our approach in dealing with problematic spectra in the reference set.

In addition to the results achieved by our specific amide I frequency based analysis, Table 7 summarises representative contributions made along with the reported prediction accuracy. With the exception of the work presented by Goormaghtigh et al. [20] and Baello et al. [33], the results are based on experiments with FTIR spectra recorded from proteins in H_2O. Only those authors are listed, where we were able to re-calculate the quality of prediction in terms of SEPs from the information provided in the respective publications. Note that in Table 7 results for curve fitting studies are shown irrespective of how the individual bands have been identified (FSD or second derivative spectra) and to what spectrum the actual curve fitting procedure has been applied (original spectrum, FSD spectrum, or derivative spectrum). The information provided in Table 7 refers to the configuration resulting in the best predictions when based on the entire infrared spectral data set available.

The curve fitting procedure has been reported to provide a good estimate of protein secondary structure (see Table 7). However, no general procedure for determining the parameters for both deconvolution and derivation exists, which may be consistently applied for band narrowing prior to curve fitting in order

Table 7
Comparison of various secondary structure prediction methods from FTIR spectra in terms of the SEP

Ref.	Year	Data set size	FTIR spectra combined with CD data	Spectral region used (FTIR) for best results[a]	Method used for calculating target secondary structure fractions[b]	Prediction method used[c]	SEP for helix[d] (%)	SEP for sheet[d] (%)	SEP for turns (%)	SEP for other[g] (%)	Average of SEPs (%)
[13]	1986	11	No	I	LG	C	2.17	2.76	NP[e]	NP	2.47
[1]	1986	6	No	I	LG	C	2.24	2.55	NP	NP	2.4
[20]	1990	14	No	I'	LG	C	10.31	6.87	NP	NP	8.59
[23]	1990	12	No	I	LG	C	5.76	6.82	8.07	6.02	6.67
[24]	1994	14	No	$I + II$	R	C	5.95	2.56	3.99	3.27	3.94
[35]	1990	18	No	I	LG	M	7.8	9.7	4.3	NP	7.27
[34]	1990	13	No	$I + II$	LG/KS	M	5.11	3.71	NP	5.14	4.65
[31]	1991	17	No	I	KS	M	9.8	11.22	6.61	9.18	9.20
[30]	1993	21	Yes	I	KS	M	7	9.5	7	10	8.38
[29]	1996	39	No	$I + II$	KS	M	12.14	9.08	NP	NP	10.61
[32]	1997	23	No	$I + II$	KS	M	8.6	7.34	1.39	3.79	5.28
[33]	2000	23	No	$I + II + I' + II'$	KS	M	5.34	6.33	3.39	5.37	5.11
[27]	2001	18	No	I	LG	N	7.7	6.4	4.8	NP	6.3
NN[f]	2001	18	No	I	LG	N	4.47	6.16	4.61	NP	5.08
Average							6.74	6.5	4.91	6.11	

[a] I: amide I; II: amide II; $I + II$: Both amide I and amide II region; $I + II + I' + II'$: amide I, amide II, amide I', and amide II' regions were used.
[b] LG: Levitt and Greer [46]; KS: Kabsch and Sander's DSSP [65]; R: Ramachandran plots.
[c] C: Curve fitting; M: Multivariate data analysis; N: Neural network analysis.
[d] If the secondary structure has been further divided (e.g., parallel, anti-parallel β-sheet), the average has been taken.
[e] NP: This type of secondary structure class has not been predicted.
[f] Our neural network approach presented in this paper.
[g] This structural class is also often referred to as "unordered", "random coil", "random", "irregular", and "undefined".

to arrive at a generalised procedure for protein secondary structure estimation. In contrast, the choice of parameters is rather a process of trial and error. Additionally, in a last step of the curve fitting procedure, spectral bands need to be manually assigned to secondary structure conformations to calculate the overall secondary structure fractions. Hence, a high degree of subjectivity is involved in curve fitting. In a critical assessment of the determination of protein secondary structure by FTIR spectroscopy, Surewicz et al. have given a few notes of caution regarding the general validity of the curve fitting approach [39]. They claim that due to a lack of uniqueness in band assignment, the assumption of equal molar absorptivities across different conformers within the amide I band, and the large number of adjustable parameters in the mathematical procedure of curve fitting, the characterisation of this procedure as a generally valid method to assess quantitatively the absolute content of protein secondary structure may be questioned.

It was observed, that curve fitting methods were generally applied to a very small number of proteins. Better results have been achieved when the analysis has been based on only few protein infrared spectra. E.g., the best average of SEPs (2.4%) was obtained for the curve fitting study based on merely 6 protein infrared spectra. The poorest average of SEPs (8.59%) was obtained based on the largest set of 14 protein infrared spectra.

It has been shown that pattern recognition based techniques as an alternative approach to curve fitting can be applied with a reduced amount of subjectivity. Most work has been done regarding multivariate data analysis methods. In the current study, we suggest an alternative pattern recognition approach based on artificial neural networks. One important choice to be made in neural network analysis is which training algorithm to employ. Much of the work in the application of neural networks to problems in analytical biochemistry has been performed using feed-forward multi-layer perceptrons trained with the conventional backpropagation algorithm [49]. However, we chose to employ resilient backpropagation over conventional backpropagation since it has several advantages over the conventional algorithm. The resilient backpropagation algorithm modifies the size of the weight-step directly by introducing the concept of resilient update values, resulting in an adaptation effort, which is not susceptible to unpredictable gradient behaviour. Additionally, in a study comparing backpropagation to resilient backpropagation and two other adaptive learning methods it has been demonstrated on a couple of representative benchmark problems that local adaptive algorithms, in particular resilient backpropagation, converge considerably faster than the ordinary backpropagation (gradient-descent) algorithm [51]. Furthermore, the choice of initial parameters does not have such an important impact on the outcome of the neural network training as with the conventional backpropagation and most other pattern recognition based approaches. This certainly is an important factor with respect to reliability of our neural network approach.

Neural network training using standard resilient backpropagation has already been demonstrated, in a study involving one of us [27], to be well suited in the area of protein secondary structure prediction from FTIR spectra. Since the latter study has been based on the same set of 18 protein FTIR spectra as in our study, our results can be directly compared to the results achieved in that study. Although Severcan et al. [27] made use of a more complex version of the "leave-one-out" method where a form of cross-validation is used to train the neural networks, poorer results were achieved than in our specific amide I frequency based analysis. Using our current approach, a major improvement of the SEP was achieved for α-helix (3.23%). For β-sheet and turns the SEP was slightly improved by 0.24% and by 0.19%, respectively. Severcan et al. [27] employed the standard resilient backpropagation learning algorithm. Improvement in prediction accuracy in the current study was achieved by successfully utilising empirical knowledge on structure sensitive frequencies within the amide I region. Additionally, we employed an extension to the standard resilient backpropagation learning algorithm where a weight decay parameter has been added to the error function to further improve generalisation and hence prediction accuracy [50]. Riedmiller demonstrated that the introduction of a weight decay parameter in combination with a relatively low maximum step size lead to an improved generalisation [64]. Overfitting did not occur even with long training times (i.e., a large number of epochs). Overfitting refers to the case where the neural network has begun to "memorise" each individual training pattern rather than settling for weights that generally describe the mapping for all cases.

4. Summary

The results in Table 7 show that neural network analysis offers great potential for predicting protein secondary structure from FTIR spectra of proteins. In particular, the results of our study demonstrate that neural networks are a valid alternative to existing methods and are worthwhile further exploration. In fact, our specific amide I frequency based neural network approach achieved better prediction accuracy than most of the pattern recognition based approaches and even better results than some of the curve fitting approaches. At the bottom of Table 7, the average of SEPs for each secondary structure across all

methods listed are given. For each secondary structure, the SEP achieved by our neural network approach is generally better than the average taken for that secondary structure across all methods listed. Clearly, the effect of using different sets of protein FTIR spectra on the resulting prediction accuracy requires further investigation. However, where the same reference set has been used, best SEPs were achieved by our method.

References

[1] H. Susi and D.M. Byler, Resolution-enhanced Fourier-transform infrared-spectroscopy of enzymes, *Methods in Enzymology* **130** (1986), 290–311.

[2] P.I. Haris and D. Chapman, Does Fourier-transform infrared-spectroscopy provide useful information on protein structures, *Trends in Biochemical Sciences* **17**(9) (1992), 328–333.

[3] W.K. Surewicz and H.H. Mantsch, New insight into protein secondary structure from resolution-enhanced infrared spectra, *Biochimica et Biophysica Acta* **952** (1988), 115–130.

[4] S. Krimm and J. Bandekar, Vibrational spectroscopy and conformation of peptides, polypeptides and proteins, *Adv. Protein Chem.* **38** (1986), 181–364.

[5] L.K. Tamm and S.A. Tatulian, Infrared spectroscopy of proteins and peptides in lipid bilayers, *Quarterly Reviews of Biophysics* **30**(4) (1997), 365–429.

[6] E. Goormaghtigh and J.M. Ruysschaert, *Molecular Description of Biological Components by Computer Aided Conformational Analysis*, R. Brasseur, ed., CRC Press, 1998, pp. 285–329.

[7] P.I. Haris, Characterization of protein structure and stability using Fourier transform infrared spectroscopy, *Pharmacy and Pharmacology Communications* **5**(1) (1999), 15–25.

[8] P.I. Haris, *Fourier Transform Infrared Spectroscopic Studies of Peptides: Potentials and Pitfalls*, B.R. Singh, ed., ACS Symposium series, American Chemical Society, 2000, pp. 54–95.

[9] P.I. Haris and D. Chapman, Analysis of polypeptide and protein structures using Fourier transform infrared spectroscopy, microscopy, optical spectroscopy, and macroscopic techniques, in: *Methods in Molecular Biology*, C. Jones, B. Mulloy, and A.H. Thomas, eds, 1994, pp. 183–202.

[10] B.R. Singh, *Basic Aspects of the Technique and Applications of Infrared Spectroscopy of Peptides and Proteins*, B.R. Singh, ed., ACS Symposium series, American Chemical Society, Washington, DC, USA, 1999, 2000, pp. 2–37.

[11] S. Krimm, *Interpreting Infrared Spectra of Peptides and Proteins*, B.R. Singh, ed., ACS Symposium series, American Chemical Society, Washington, DC, USA, 1999, 2000, pp. 38–53.

[12] M. Jackson and H.H. Mantsch, The use and misuse of FTIR spectroscopy in the determination of protein-structure, *Critical Reviews in Biochemistry and Molecular Biology* **30**(2) (1995), 95–120.

[13] D.M. Byler and H. Susi, Examination of the secondary structure of proteins by deconvolved FTIR spectra, *Biopolymers* **25**(3) (1986), 469–487.

[14] D.G. Cameron and D.J. Moffatt, A generalized approach to derivative spectroscopy, *Applied Spectroscopy* **41** (1987), 539–544.

[15] D.G. Cameron and D.J. Moffatt, Deconvolution, derivation and smoothing of spectra using Fourier transforms, *Journal of Testing and Evaluation* **12** (1984), 78–85.

[16] J.K. Kauppinen, D.J. Moffatt, H.H. Mantsch and D.G. Cameron, Fourier self-deconvolution: a method for resolving intrinsically overlapped bands, *Applied Spectroscopy* **35** (1981), 271–276.

[17] J.K. Kauppinen, D.J. Moffatt, H.H. Mantsch and D.G. Cameron, Fourier transforms in the computation of self-deconvoluted and first-order derivative spectra of overlapped band contours, *Analytical Chemistry* **53** (1981), 1454–1457.

[18] M. Ruegg, V. Metzger and H. Susi, Computer analysis of characteristic infrared bands of globular proteins, *Biopolymers* **14** (1975), 1465–1471.

[19] J. Villalain, J.C. Gomez-Fernandez, M. Jackson and D. Chapman, Fourier transform infrared spectroscopic studies on the secondary structure of the Ca^{2+}-ATPase of sarcoplasmic reticulum, *Biochimica et Biophysica Acta* **978** (1989), 305–312.

[20] E. Goormaghtigh, V. Cabiaux and J.M. Ruysschaert, Secondary structure and dosage of soluble and membrane proteins by attenuated total reflection Fourier-transform infrared spectroscopy on hydrated films, *Eur. J. Biochem.* **193** (1990), 409–420.

[21] B.R. Singh, D.B. DeOliveira, F. Fu and M.P. Fuller, Fourier transform infrared analysis of amide III bands of proteins for the secondary-structure estimation [1890-11], Biomolecular spectroscopy III, in: *Proceedings – SPIE the International Society for Optical Engineering*, L.A. Nafie and H.H. Mantsch, eds, SPIE, 1993, pp. 47–55.

[22] B.R. Singh, M.P. Fuller and B.R. DasGupta, Botulinum neurotoxin type A: structure and interaction with the micellar concentration of SDS determined by FT-IR spectroscopy, *J. Protein Chem.* **10** (1991), 637–649.

[23] A. Dong, P. Huang and W.S. Caughey, Protein secondary structure from second derivative amide I infrared spectra, *Biochemistry* **29** (1990), 3303–3308.
[24] T.F. Kumosinski and J.J. Unruh, Global-secondary-structure analysis of proteins in solution – Resolution-enhanced deconvolution Fourier-transform infrared-spectroscopy in water, molecular modeling, in: *ACS Symposium Series*, Anonymous, ed., Amer. Chemical Soc., 1994, pp. 71–98.
[25] H.H. Mantsch and D.J. Moffatt, Computer-aided methods for the resolution enhancement of spectral data with special emphasis on infrared spectra, in: *NATO ASI Series C, Mathematical and Physical Sciences*, R. Fausto, ed., Kluwer Academic Publishers, USA, 1993, pp. 113–124.
[26] P.R. Griffiths and G.L. Pariente, Introduction to spectral deconvolution, *Trends Anal. Chem.* **5** (1986), 209–215.
[27] M. Severcan, F. Severcan and P.I. Haris, Estimation of protein secondary structure from FTIR spectra using neural networks, *Journal of Molecular Structure* **565** (2001), 383–387.
[28] V. Baumruk, P. Pancoska and T.A. Keiderling, Predictions of secondary structure using statistical analyses of electronic and vibrational circular dichroism and fourier transform infrared spectra of proteins in H_2O, *Journal of Molecular Biology* **259**(4) (1996), 774–791.
[29] K. Rahmelow and W. Huebner, Secondary structure determination of proteins in aqueous solution by infrared spectroscopy: A comparison of multivariate data analysis methods, *Analytical Biochemistry* **241**(1) (1996), 5–13.
[30] R. Pribic, I.H.M. Van Stokkum, D. Chapman, P.I. Haris and M. Bloemendal, Protein secondary structure from Fourier transform infrared and/or circular dichroism spectra, *Analytical Biochemistry* **214**(2) (1993), 366–378.
[31] R.W. Sarver and W.C. Krueger, Protein secondary structure from Fourier transform infrared spectroscopy: A database analysis, *Analytical Biochemistry* **194** (1991), 89–100.
[32] S. Wi, P. Pancoska and T.A. Keiderling, Predictions of protein secondary structures using factor analysis on Fourier transform infrared spectra: Effect of Fourier self-deconvolution of the amide I and amide II bands, *Biospectroscopy* **4**(2) (1997), 93–106.
[33] B.I. Baello, P. Pancoska and T.A.a. Keiderling, Enhanced prediction accuracy of protein secondary structure using hydrogen exchange Fourier transform infrared spectroscopy, *Analytical Biochemistry* **280**(1) (2000), 46–57.
[34] F. Dousseau and M. Pezolet, Determination of the secondary structure content of proteins in aqueous solutions from their amide I and amide II infrared bands. Comparison between classical and partial least-squares methods, *Biochemistry* **29** (1990), 8771–8779.
[35] D.C. Lee, P.I. Haris, D. Chapman and R.C. Mitchell, Determination of protein secondary structure using factor – Analysis of infrared-spectra, *Biochemistry* **29**(39) (1990), 9185–9193.
[36] P. Pancoska, H. Fabian, G. Yoder, V. Baumruk and T.A. Keiderling, Protein structural segments and their interconnections derived from optical spectra. Thermal unfolding of ribonuclease T-1 as an example, *Biochemistry* **35**(40) (1996), 13 094–13 106.
[37] V. Cabiaux, K.A. Oberg, P. Pancoska, T. Walz, P. Agre and A. Engel, Secondary structures comparison of aquaporin-1 and bacteriorhodopsin: A Fourier transform infrared spectroscopy study of two-dimensional membrane crystals, *Biophysical Journal* **73**(1) (1997), 406–417.
[38] P. Pancoska, V. Janota and T.A. Keiderling, Novel matrix descriptor for secondary structure segments in proteins: Demonstration of predictability from circular dichroism spectra, *Analytical Biochemistry* **267**(1) (1999), 72–83.
[39] W.K. Surewicz, H.H. Mantsch and D. Chapman, Determination of protein secondary structure by Fourier transform infrared spectroscopy: A critical assessment **32**(2) (1993), 389–394.
[40] J. Moult, The current state of the art in protein structure prediction, *Current Opinion in Biotechnology* **7**(4) (1996), 422–427.
[41] B. Rost and S. O'Donoghue, Sisyphus and prediction of protein structure, *CABIOS* **13**(4) (1997), 345–356.
[42] P. Baldi, S. Brunak, P. Frasconi, G. Soda and G. Pollastri, Exploiting the past and the future in protein secondary structure prediction, *Bioinformatics* (Oxford) **15**(11) (1999), 937–946.
[43] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen and H. Nielsen, Assessing the accuracy of prediction algorithms for classification: An overview, *Bioinformatics* (Oxford) **16**(5) (2000), 412–424.
[44] M.A. Andrade, P. Chacon, J.J. Merelo and F. Moran, Evaluation of secondary structure of proteins from UV circular-dichroism spectra using an unsupervised learning neural-network, *Protein Engineering* **6**(4) (1993), 383–390.
[45] B. Dalmas, G.J. Hunter and W.H. Bannister, Prediction of protein secondary structure from circular dichroism spectra using artificial neural network techniques, *Biochemistry and Molecular Biology International* **34**(1) (1994), 17–26.
[46] M. Levitt and J. Greer, Automatic identification of secondary structure in globular proteins, *Journal of Molecular Biology* **114** (1977), 181–293.
[47] R. Hecht-Neilson, *Neurocomputing*, Addison Wesley, 1990.
[48] P. Pancoska, V. Janota and T.A. Keiderling, Interconvertibility of electronic and vibrational circular dichroism spectra of proteins: a test of principle using neural network mapping, *Applied Spectroscopy* **50**(5) (1996), 658–668.
[49] J.L. McClelland and D.E. Rumelhart, Parallel distributed processing: explorations in the microstructure of cognition, in: *Psychological and Biological Models*, Vol. 2, MIT Press, 1987.

[50] M. Riedmiller and H. Braun, A direct adaptive method for faster backpropagation learning: The RPROP algorithm, in: *IEEE International Conference on Neural Networks (ICNN-93)*, H. Ruspini, ed., 1993, pp. 586–591.
[51] M. Riedmiller, Advanced supervised learning in multi-layer perceptrons – from backpropagation to adaptive learning algorithms, *International Journal of Computer Standards and Interfaces*, Special Issue on Neural Networks **5** (1994), 8.
[52] S. Haykin, *Neural Networks, A Comprehensive Foundation*, Macmillan College Publishing Company, Inc., 1994, pp. 176–177.
[53] C. Klawun and C.L. Wilkins, Optimization of functional group prediction from infrared spectra using neural networks, *Journal of Chemical Information and Computer Sciences* **36**(1) (1996), 69–81.
[54] M.S. Braiman and K.J. Rothschild, Fourier-transform infrared techniques for probing membrane – protein structure, *Annual Review of Biophysics and Biophysical Chemistry* **17** (1988), 541–570.
[55] E.B. Baum and D. Haussler, What size net gives valid generalisation?, *Neural Computing* **1** (1989), 151–160.
[56] R. Lange and R. Männer, Quantifying a critical training set size for generalisation and overfitting using teacher neural networks, in: *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, P. Morasso and M. Marinaro, eds., 1994, pp. 497–500.
[57] J.L.R. Arrondo and F.M. Goni, in: *Protein–Lipid Interactions*, A. Watts, ed., Elsevier, 1993, pp. 321–349.
[58] Y.N. Chirgadze, O.V. Fedorov and N.P. Trushina, Estimation of amino acid residue side chain absorptions in infrared spectra of protein solutions in heavy water, *Biopolymers* **14** (1975), 679–694.
[59] S.Y. Venyaminov and N.N. Kalnin, Quantitative IR spectrophotometry of peptide compounds in water (H_2O) solutions. I. Spectral parameters of amino acid residue absorption bands, *Biopolymers* **30** (1990), 1243–1257.
[60] K. Rahmelow, W. Huebner and T. Ackermann, Infrared absorbances of protein side chains, *Analytical Biochemistry* **257**(1) (1998), 1–11.
[61] A. Elliot and E.J. Ambrose, Structure of synthetic polypeptides, *Nature* **165** (1950), 921–922.
[62] E.J. Ambrose and A. Elliot, Infrared spectroscopic studies of globular protein structure, *Proc. R. Soc. London Ser. A* **208** (1951), 75–90.
[63] A. Elliot, Infrared spectra of polypeptides with small side chains, *Proc. R. Soc. London Ser. A* **226** (1954), 408–421.
[64] M. Riedmiller, Untersuchungen zur Konvergenz und Generalisierungsfähigkeit überwachter Lernverfahren mit dem SNNS, Workshop SNNS-93: Simulation Neuronaler Netze mit SNNS (Anonymous), Universität Stuttgart, Fakultät Informatik, 1993, pp. 107–116.
[65] W. Kabsch and C. Sander, Dictionary of protein secondary structure – pattern-recognition of hydrogen-bonded and geometrical features, *Biopolymers* **22**(12) (1983), 2577–2637.