Empirical Knowledge and Genetic Algorithms for Selection of Amide I Frequencies in Protein Secondary Structure Prediction

Joachim A. Hering[#], Peter R. Innocent[§], Parvez I. Haris[*,#]

[#] Leicester School of Pharmacy, De Montfort University, The Gateway, Leicester, LE1 9BH, UK

[§] Department of Computer Science, De Montfort University, The Gateway, Leicester, LE1 9BH, UK

pharis@dmu.ac.uk

Abstract

Here we investigate an extension of a previously suggested "automatic amide I frequency selection procedure" where we introduce an additional criterion utilizing empirical knowledge on regions within the amide I band (1600-1700 cm^{-1}) found to be particularly sensitive to protein secondary structure. We show that the genetic algorithm provides a solution with good protein secondary structure prediction accuracy.

Based on an evaluation set of 13 protein infrared spectra from proteins not contained in the reference set, it is demonstrated that our method is capable of making good predictions for proteins it has never seen before during training. In the present study, where the genetic algorithm is guided towards a solution with a higher number of empirically determined, structure sensitive amide I frequencies selected, minor improvement in prediction accuracy for α-helix and β-sheet structure could be achieved compared to our previous study, where no such knowledge has been provided. Despite the very limited number of protein spectra in the reference set (18), the neural networks were able to generalize with an overall average of standard errors of prediction of 4.36 % based on the evaluation set of protein spectra, which is even better than that achieved during the analysis based on the reference set of protein spectra (4.8 %). This clearly indicates the potential of our approach once more protein infrared spectra are available to base the analysis on.

Keywords: Protein secondary structure prediction; FTIR spectroscopy; Neural networks; Genetic algorithms; Empirical knowledge

1 Introduction

Several different techniques are available for quantification of protein secondary structure from FTIR spectra of proteins (Haris, 2000). Recently, we have introduced an alternative method based on the combination of a genetic algorithm and neural networks (Hering et al. 2002). The main aim of this study was to let the genetic algorithm search for an optimal composition of amide I frequencies producing good predictions across protein secondary structures under investigation. Prediction accuracy was evaluated in terms of the average of the standard error of prediction (SEP). In that study, the genetic algorithm was merely guided by the average of SEPs and the number of frequencies selected.

In the present study, we investigated whether the genetic algorithm could be further guided by additionally providing information on empirically determined structure sensitive regions (Haris, 1999; Haris, 2000). Additionally, we were interested in whether the found solution would allow for good predictions to be made based on an evaluation set of protein spectra never seen during training. The general principle of genetic algorithms has been described elsewhere, e. g. (Goldberg, 1989). In the present paper we make use of the same terminology used in our previous paper (Hering et al. 2002) originally introduced by John Holland (Holland, 1975).

2 Materials and Methods

2.1 Protein reference and evaluation set

Two independent protein sets are used, a reference set to find an optimal composition of amide I frequencies and an evaluation set to test the found solution on an independent set of proteins not contained in the reference set.

As a reference set we used the same set of 18 FTIR spectra from proteins in aqueous solution as used in our previous study (Hering et al. 2002). Target fractions of secondary structure (in %) as determined by Kabsch and Sander's "Database of Secondary Structure in Proteins" (DSSP) method (Kabsch and Sander, 1983) as well as the corresponding Protein Data Bank (PDB) (Berman et al. 2000) codes for the proteins of the reference set can be found there. A detailed description regarding sample preparation and FTIR measurements is given in Lee et al.'s paper (Lee et al. 1990).

For validation of the best solutions found by the genetic algorithm based on the reference set, we employed an independent set of 13 protein infrared spectra (see Table 1). These protein spectra were chosen from a set of protein spectra kindly provided by Dong et al. (Dong, 2002) and Keiderling et al. (Keiderling, 2002). Details on sample preparation and FTIR measurements can be found at the respective internet pages (Dong, 2002; Keiderling, 2002).

Protein	Species	PDB Code[a]	Spectra provided by[b]	Kabsch and Sander[c]				
				Helix[d]	Sheet[e]	Turn	Bend	Other[f]
Holo-Lactoferrin	Human milk	1LFG	D	34.3	18.23	18.38	10.71	18.38
Albumin	Human serum	1BM0	D	72.6	0	9.42	5.53	12.45
Beta-Lactoglobulin B	Bovine milk	1BSQ	D	16.05	37.65	9.88	16.67	19.75
Alpha-Lactalbumin (Ca-depleted)	Bovine milk	1F6R	D	44.67	9.84	17.9	10.38	17.21
Ferritin (Apo)	Horse spleen	1IES	D	73.4	0.19	7.44	3.53	15.44
Rhodanese	Bovine liver	1RHD	K	29.69	13.31	16.38	10.92	29.69
Subtilisin Carlsberg	Bacillus licheniformis	1SCD	D	29.2	19.71	16.42	10.22	24.45
Triose phosphate isomerase	Yeast	1YPI	K	44.24	16.36	9.49	7.27	22.63
Deoxyribonuclease I	Bovine pancreas	3DNI	D	28.96	28.57	8.11	10.04	24.32
Gluthathione reductase	Wheat germ	3GRS	K	34.27	25.16	12.15	9.11	19.31
Trypsin	Bovine pancreas	3PTN	K	9.87	34.98	14.8	15.25	25.11
Lactate dehydrogenase	Rabbit	6LDH	K	43.77	17.02	8.51	9.73	20.97
Thermolysin	Bacterial	8TLN	K	41.07	16.93	10.34	13.48	18.18

[a] Protein Data Bank code for reference structure (Berman et al. 2000).

[b] D: Dong et al. (Dong, 2002); K: Keiderling et al. (Keiderling, 2002)

[c] Target fractions of secondary structure (in %) as determined by Kabsch and Sander's DSSP method (Kabsch and Sander, 1983).

[d] Including α-helix, 3_{10}-helix, and π-helix.

[e] Including isolated beta bridge and extended strand.

[f] Including all structure not explicitly given.

Table 1 Secondary structure for the evaluation set of proteins as determined by X-ray studies

Secondary Structure	Protein training set (18 proteins)[a]					Evaluation set (13 proteins)[b]				
	Helix	Sheet	Turn	Bend	Other	Helix	Sheet	Turn	Bend	Other
Minimum	3.8	0	5.88	1.96	11.76	9.87	0	7.44	3.53	12.45
Maximum	80.39	51.53	20.93	17.24	31.73	73.4	37.65	18.38	16.67	29.69
Mean	27.75	26.62	13.4	9.33	22.9	38.62	18.3	12.25	10.22	20.61
Standard deviation	21.77	16.84	3.92	4.16	6.21	18.49	11.45	3.98	3.6	4.6

[a] Set of protein FTIR spectra from Lee et al. (Lee et al. 1990)

[b] Subset of protein FTIR spectra from Dong et al. (Dong, 2002) and Keiderling et al. (Keiderling, 2002)

Table 2 Distribution of secondary structural conformation (in % structure) as determined by X-ray crystallography studies for the18 proteins of the reference set as well as for the 13 proteins of the evaluation set.

We have selected a subset of 13 spectra as the evaluation set such that the set of proteins from the reference set and the set of proteins from the evaluation set do not contain same proteins. Additionally, only those protein spectra were selected for validation where target fractions of secondary structure fall within the range covered by the reference set of proteins (see Table 2). Target fractions of secondary structure (in %) were calculated using the same method as used for the reference set of proteins. Details on the proteins used for the evaluation set are given in Table 1.

A statistical characterization of the distribution for each secondary structure as calculated by Kabsch and Sander's DSSP method (Kabsch and Sander, 1983) for both the reference set and the evaluation is given in Table 2.

2.2 The software used

The Stuttgart Neural Network Simulator (SNNS, Version 4.2) is used for neural network analysis. SNNS is a complex simulator for neural networks developed at the Institute for Parallel and Distributed High Performance Systems at the University of Stuttgart, Germany.

On top of the "batchman" interface provided by SNNS, we implemented a simple genetic algorithm using Java (JDK 1.3).

2.3 Prediction accuracy evaluation

As in our previous study (Hering et al. 2002), the "leave-one-out" method is used for prediction accuracy evaluation during neural network training. It is worth noting that for each "leave-one-out" run, the protein left out from the analysis for prediction accuracy evaluation is not seen by the neural network during training. However, since the genetic algorithm on top of the neural network training favors "leave-one-out" runs producing good SEPs, it guides the neural network analysis towards finding a set of weight connections optimal for the protein spectra within the reference set. To get a better understanding about the true generalization capabilities of the best solution found by the genetic algorithm, it was evaluated using a separate evaluation set of 13 protein infrared spectra not known during training. At the end of the genetic algorithm, the best individual found produced 18 neural networks - one for each protein in the reference set left out from the training for evaluation. The absorbance values at the frequencies determined by the best individual found by the genetic algorithm were extracted from each protein spectrum of the evaluation set (see Table 1), normalized, and fed through each of the 18 neural networks. The final prediction for each secondary structure was then calculated as the average of predictions produced by the 18 neural networks.

Prediction accuracy is generally measured in terms of the standard error of prediction (SEP) which has also been employed previously (Lee et al. 1990; Hering et al. 2002):

$$SEP = \sqrt{\frac{\sum_{j=1}^{n}\left(p_{cj} - p_{xj}\right)^2}{n}} \quad (1)$$

where p_{cj} = the proportion of structure predicted for protein j by the respective method, p_{xj} = the proportion of structure calculated from the original X-ray data for protein j, and n = the number of proteins in the reference set respectively the proteins of the evaluation set.

2.4 Neural network topology and neural network training algorithm

In the present study, the same basic neural network topology trained by the same neural network algorithm is used as in our previous study (Hering et al. 2002). The number of inputs to the neural network is varied based on the number of frequencies selected by the genetic algorithm.

2.5 The frequency selection procedure

In the present study, we employed a combination of a genetic algorithm and neural network analysis to automatically select a set of amide I frequencies. This set of frequencies determines a set of absorbance values for each protein infrared spectrum to be used for quantification of its protein secondary structure by providing those values to a neural network analysis. The genetic algorithm used in the present study is an extension of the genetic algorithm used in our previous study where details can be found (Hering et al. 2002). In the present study, an additional term has been introduced to the fitness function to account for studies on empirically determined structure sensitive regions (Haris, 1999; Haris, 2000). According to those studies, we have determined three frequencies, namely, 1654 cm^{-1}, 1630 cm^{-1}, and 1672 cm^{-1}, as mid points of those structure sensitive regions for helix, sheet, and turn structure, respectively. Here, we have defined structure sensitive regions as intervals +/- 5 frequencies around those mid points resulting in 1649-1659 cm^{-1}, 1625-1635 cm^{-1}, and 1667-1677 cm^{-1}. The fitness of each individual in the population is calculated as follows:

$$Fitness_{Individual} = \underbrace{Round(100 - AverageOfSEPsInPercent_{Individual}, 3) \cdot 10^9}_{\text{The lower the average of SEPs, the better the fitness}} + \underbrace{NoFrequenciesSelectedFromEmpirical Regions_{Individual} \cdot 10^3}_{\text{The more frequencies from empirical regions selected, the better the fitness}} + \underbrace{ChromosomeLength - NoFrequenciesSelected_{Individual}}_{\text{The fewer frequencies selected, the better the fitness}} \quad (2)$$

The "number of frequencies selected from empirical regions" was then determined for each individual as the number of frequencies that fall into those regions. Overall, evaluation of an individual was performed as follows: Lower average of SEPs always results in better fitness (larger fitness value). If two individuals have the same average of SEPs, but their "number of frequencies selected from empirical regions" is different, then the individual with higher "number of frequencies selected from empirical regions" will have better fitness. If two individuals have both the same average of SEPs and the same "number of frequencies selected from empirical regions", but their overall number of frequencies selected for neural network training is different, then the individual with the lower number of frequencies selected will have higher fitness. This way, the genetic algorithm is mainly guided by the achieved prediction accuracy in terms of the average of SEPs. When prediction accuracy amongst individuals is the same, the genetic algorithm is encouraged to pursue solutions with as few frequencies selected as possible and as many of those frequencies stemming from empirically determined structure sensitive regions. Since we believe that a selected number of amide I frequencies below 10 may not be sensible, we have restricted the minimum number of 1's in a chromosome (i. e., the number of frequencies selected) to be greater than or equal to 10.

3 Results and Discussion

3.1 Empirical knowledge to guide the genetic algorithm towards a better solution

Based on the fitness function described above, the genetic algorithm converged after 1115 generations to its best solution where the following 35 frequencies were selected (cm^{-1}): 1601, 1603, 1604, 1606, 1608, 1609, 1611, 1612, 1621, 1624, 1625, 1630, 1631, 1632, 1633, 1634, 1636, 1638, 1640, 1645, 1646, 1647, 1648, 1652, 1653, 1655, 1656, 1659, 1663, 1670, 1671, 1676, 1678, 1680, and 1681.

Table 3 shows both the results of our previous "automatic amide I frequency selection" procedure (Hering et al. 2002) and the results of the present study based on the reference set of protein spectra used during the analysis.

Results are only shown for the best individuals found. Based on a frequency selection pattern of 35 absorbance values, minor improvement of the average of SEPs (0.18 %) was achieved compared to our previous study (Hering et al. 2002).

Reference	Helix	Sheet	Turn	Bend	Other	Avg
(Hering et al. 2002)	4.58	5.72	4.42	3.95	6.12	4.96
Present study	4.42	5.8	4.26	3.71	5.8	4.8

Table 3 Standard errors of prediction (in % structure) are given for our previous study (Hering et al. 2002) as well as for the present study.

Our main interest in the current study was to investigate whether providing the genetic algorithm with empirical knowledge on structure sensitive regions within the amide I band would guide the search towards a better solution. Table 3 shows the SEPs achieved in the present study as well as the SEPs achieved by our previous study where no empirical knowledge on structure sensitive regions was embedded in the genetic algorithm's fitness function (Hering et al. 2002). Based on our reference set of 18 protein infrared spectra, no significant improvement could be achieved by additionally providing the genetic algorithm with empirical knowledge on structure sensitive regions. The average of SEPs could be improved by merely 0.18 %. Comparing the set of frequencies selected by our previous study (Hering et al. 2002) with the set of frequencies selected in the present study, only 12 frequencies were the same. In the present study, a clear shift towards lower frequencies selected within the amide I band was observed, where the highest frequency selected was 1681 cm^{-1}. In our previous study 9 frequencies above that value were selected including 1700 cm^{-1}. For both our previous study (Hering et al. 2002) and the present study, Table 4 shows the number of frequencies that fall in each empirically determined structure sensitive region within the amide I band.

Amide I region (cm^{-1})	Secondary structure	No. of frequencies[a, b] previous study (Hering et al. 2002))	No. of frequencies[a, b] present study
1648-1660	α-helix	4 (30.77 %)	6 (46.15 %)
1620-1640 1670-1695	β-sheet	14 (29,79 %)	17 (36.17 %)
1620-1640 1650-1695	turn	19 (28,36 %)	23 (34.33 %)
1640-1657 1660-1670	other	8 (27,59 %)	11 (37.93 %)
1600-1619 1696-1700	none of the above	6 (24 %)	8 (32 %)
No. of frequencies selected (best individual)		29	35

[a] Since the regions given are overlapping, same frequencies may be assigned to multiple regions.

[b] The number in brackets gives the percentage of the number of frequencies in relation to the overall number of frequencies of the respective structure sensitive region within the amide I region.

Table 4 Empirically determined structure sensitive regions. For both our previous "automatic amide I frequency selection procedure" and the present study the number of frequencies are given that fall in the respective regions. The number of frequencies selected which do not fall in any of these empirically determined regions is also shown.

The number of frequencies selected outside these empirically determined regions are given at the bottom of Table 4. From this table, it can be seen that the genetic algorithm of the present study did in fact favor a solution with more frequencies within structure sensitive regions. However, despite the fact, that the solution found in the present study is in better agreement with empirical studies on structure sensitive regions(Haris, 1999; Haris, 2000), no significant improvement in prediction accuracy was achieved based on our reference set of 18 protein spectra. Table 5 shows the averages of SEPs for the protein spectra of the evaluation set based on the evolved set of amide I frequencies and neural networks from both the present study and our previous study (Hering et al. 2002).

The overall average of SEPs achieved by our previous study is 4.54 %, which is comparable to that achieved by the present study (4.36 %). However, by guiding the genetic algorithm towards a set of frequencies within empirically determined structure sensitive regions, improved predictions were made in the present study for helix and sheet structure. For turn structure, no improvement in prediction accuracy was observed. This is possibly due to the fact that for turn structure, assignments are less well defined resulting in a wide range of frequencies where turn structure has been found to absorb (1620 cm^{-1} to 1640 cm^{-1} and 1650 cm^{-1} to 1695 cm^{-1}) within the amide I band (Haris, 1999; Haris, 2000). However, since overall only minor improvements were

observed, further studies are required to investigate if providing empirical knowledge on structure sensitive regions has the potential to significantly improve overall prediction accuracy.

Protein	Species	Helix[c]	Sheet[d]	Turn	Bend	Other[e]	Avg
Holo-Lactoferrin	Human milk	1.9	3.97	4	1.06	6.03	3.39
Albumin	Human serum	10.46	1.74	1.8	0.9	4.99	3.98
Beta-Lactoglobulin B	Bovine milk	2.66	2.46	3.86	6.31	4.07	3.87
Alpha-Lactalbumin (Ca-depleted)	Bovine milk	6.53	1.01	4.01	1.83	5.99	3.87
Ferritin (Apo)	Horse spleen	6.16	1.29	4.33	1.17	3	3.19
Rhodanese	Bovine liver	3.64	4.09	2.1	1.69	5.24	3.35
Subtilisin Carlsberg	Bacillus licheniformis	6.04	8.1	2.2	1.2	1.78	3.87
Triose phosphate isomerase	Yeast	12.83	1.96	5.41	3.27	2.78	5.25
Deoxyribonuclease I	Bovine pancreas	9.88	2.97	6.03	0.86	0.35	4.02
Gluthathione reductase	Wheat germ	4.92	16.18	2.47	1.86	5.04	6.09
Trypsin	Bovine pancreas	6.63	1.91	0.9	4.18	0.83	2.89
Lactate dehydrogenase	Rabbit	2.1	8.99	5.78	1.07	2.74	4.13
Thermolysin	Bacterial	15.11	13.89	2.15	8.48	4.38	8.8
Average		6.84	5.27	3.46	2.61	3.63	4.36
Average previous study (Hering et al. 2002)		7.97	5.85	3.3	2.84	2.72	4.54

[a] Protein Data Bank code for reference structure (Berman et al. 2000).

[b] Target fractions of secondary structure (in %) as determined by Kabsch and Sander's DSSP method (Kabsch and Sander, 1983).

[c] Including α-helix, 3_{10}-helix, and π-helix.

[d] Including isolated beta bridge and extended strand.

[e] Including all structure not explicitly given.

Table 5 Standard errors of prediction (in % structure) for each protein of the evaluation set. For each spectrum, the frequencies of the best individual found by the GA were extracted, normalized, and fed through each of the 18 neural networks generated by the "leave-one-out" method. The averages of errors produced by the 18 neural networks are given here.

3.2 Generalization

Merely employing a genetic algorithm in combination with neural networks to search for an optimal set of amide I frequencies to produce good predictions based on a reference set of protein spectra should not be viewed as our main goal, but rather an intermediate step. Our procedure can only be said to be successful, if it also achieves good prediction accuracy based on infrared data from any other protein outside the reference set. Hence, the best solution found by the present study was evaluated using a set of 13 infrared spectra from proteins outside the reference set (see Table 1). Standard errors of prediction are shown in Table 5, where SEPs are given for each protein of the evaluation set along with the averages. An overall average of SEPs of 4.36 % was achieved. Additionally, averages of SEPs for each secondary structure are shown where the spectral data of the evaluation set has been presented to the best solution found by our previous study (Hering et al. 2002). Here, an overall average of SEPs of 4.54 % was achieved. Despite the very limited number of protein spectra in the reference set (18), good predictions were made resulting in average errors (in % structure) of 6.84 %, 5.27 %, 3.46 %, 2.61 %, and 3.63 %, for helix, sheet, turn, bend, and other structure, respectively. Overall, an average error of prediction of 4.36 % was achieved. Predictions made for sheet, turn, bend, and other structure are even better than those made based on the reference set using the "leave-one-out" method (see Table 3 and Table 5). For helix structure, the average of SEPs based on the evaluation set was 2.42 % higher. Bearing in mind that the reference set and the evaluation set of protein spectra were recorded by different groups in different laboratories, the results demonstrate that our neural network approach is very well capable of dealing with protein spectra recorded in different laboratories.

When looking at the prediction errors made for the proteins of the evaluation set individually, high variation in SEPs is observed. Prediction errors for helix structure range from 1.9 % to 15.11 %, for sheet structure from 1.01 % to 16.18 %, for turn structure from 0.9 % to 8.85 %, for bend structure from 0.86% to 8.48 %, and for other structure from 0.35 % to 6.03 %. This clearly indicates that the reference set of protein spectra does not contain a sufficiently large number of spectra required to represent all spectral features to be able to make consistently good predictions for proteins outside the reference set. Table 2 shows the range of quantities (in % structure) covered by the proteins both of the reference set and the evaluation set for each secondary structure. The broad ranges for the more complex helix and sheet structures underline the need for a larger reference set to sufficiently represent all spectral variation for the neural network to be able to predict any quantity within that range reliably. For the less complex turn and bend structures of the reference set, however, the ranges of quantities covered are less broad (see Table 2). Here, the number of protein spectra in the reference set seemed to be sufficient to allow good predictions to be made for protein spectra outside the reference set.

4 Summary

In the present study, we have extended our previously reported "automatic amide I frequency selection procedure" (Hering et al. 2002) by additionally embedding empirical knowledge on structure sensitive regions within the amide I band into the fitness function to guide the genetic algorithm towards a better solution.

The fact that both genetic algorithm studies found solutions with comparable prediction accuracy but substantially different sets of frequencies selected from the same set of 18 FTIR spectra of proteins leads to the conclusion, that only sub-optimal solutions were found. However, this is not surprising bearing in mind the vast number of possible solutions of over $2.5*10^{30}$ where in both our studies only 36030 possible solutions were explored. However, in both our previous study (Hering et al. 2002) and the present study, the genetic algorithm could be guided very rapidly towards solutions based on a set of merely 18 FTIR spectra of proteins capable of making good predictions about the secondary structure of proteins not known during the analysis. Good prediction accuracy was demonstrated based on an evaluation set of 13 protein infrared spectra outside the reference set (see Table 5). This clearly demonstrates the potential of our approach, once an optimal solution is found. For such an optimal solution to be found, there is a need for a sufficiently large and representative reference set of protein spectra to base our genetic algorithm on, better hardware to allow for broader exploration of the search space, and possibly further enhancements in the fitness function to guide the genetic algorithm towards a better solution more rapidly. Once all these criteria are met, our genetic algorithm approach will provide us with a powerful tool in proteomics research where protein secondary structure predictions from FTIR spectra of proteins can be generally made with good prediction accuracy.

5 Acknowledgements

We would like to thank Dr. A. Dong et al. and Dr. T. A. Keiderling et al. for kindly providing us with infrared spectral data from their studies.

6 References

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research* **28** (1):235-242.

Dong, A.: Protein Infrared Database, http://www.unco.edu/chemist/aichun/irdata.htm. Accessed: 25 Oct 2002.

Goldberg, D.E. (1989) Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley.

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

Haris, P. I. (2000): Fourier Transform Infrared Spectroscopic Studies of Peptides: Potentials and Pitfalls. Proc. ACS Symposium series, USA : Washington, DC, **750**:54-95, American Chemical Society.

Hering, J.A., Innocent, P.R. and Haris, P.I. (2002) Automatic Amide I frequency selection for rapid quantification of protein secondary structure from FTIR spectra of proteins. *Proteomics* **2** (7):839-849.

Holland, J.H. (1975) Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor, Michigan.

Kabsch, W. and Sander, C. (1983) Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **22** (12):2577-2637.

Keiderling, T. A.: The Keiderling Group - Download Page, http://www.uic.edu/labs/takgroup/FARMR/farmr.html. Accessed: 25 Oct 2002.

Lee, D.C., Haris, P.I., Chapman, D. and Mitchell, R.C. (1990) Determination of Protein Secondary Structure Using Factor- Analysis of Infrared-Spectra. *Biochemistry* **29** (39):9185-9193.