

**Original citation:**

Hall, Vincent Austin, Sklepari, Meropi and Rodger, Alison. (2014) Protein secondary structure prediction from circular dichroism spectra using a self-organizing map with concentration correction. *Chirality*, 26 (9). pp. 471-482.

**Permanent WRAP url:**

<http://wrap.warwick.ac.uk/75651>

**Copyright and reuse:**

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**Publisher's statement:**

"This is the peer reviewed version of the following article: Hall V., Sklepari M. and Rodger A. (2014), Protein Secondary Structure Prediction from Circular Dichroism Spectra Using a Self-Organizing Map with Concentration Correction, *Chirality*, 471-482, DOI: 10.1002/chir.22338, which has been published in final form at <http://dx.doi.org/10.1002/chir.22338>. This article may be used for non-commercial purposes in accordance with [Wiley Terms and Conditions for Self-Archiving](#)."

**A note on versions:**

The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP url' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: [publications@warwick.ac.uk](mailto:publications@warwick.ac.uk)

warwick**publications**wrap  
highlight your research

<http://wrap.warwick.ac.uk>

# Protein Secondary Structure Prediction from Circular Dichroism Spectra Using a Self-organising Map with Concentration Correction

VINCENT HALL,<sup>1</sup> MEROPI SKLEPARI,<sup>2</sup> AND ALISON RODGER<sup>2\*</sup>

1. MOAC, Department of Chemistry and School of Engineering, University of Warwick, Coventry CV4 7AL, UK.
2. Warwick Centre for Analytical Science and Department of Chemistry, University of Warwick, Coventry, CV4 7AL, UK. Phone: +44 2476574696. Fax: +44 2476575795. Email: A.Rodger@warwick.ac.uk

Short title: Secondary structure from CD using a SOM

*KEY WORDS:* Artificial Neural Network, Kohonen map, “Secondary Structure Neural Network”, SSNN, peptides, MATLAB, CDPro, Dichroweb

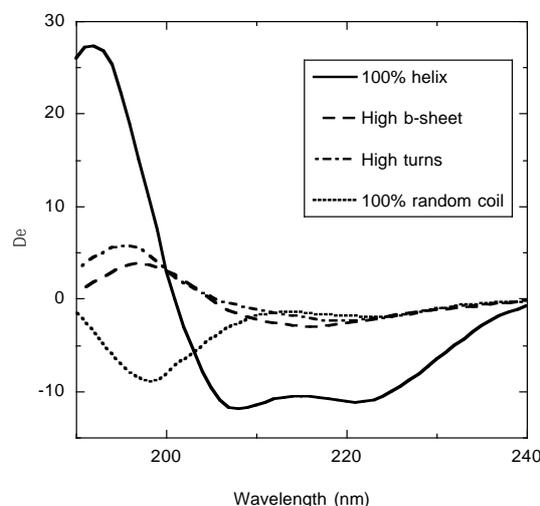
Contract grant sponsor: EPSRC; Contract grant number: EP/F500378/1.

**ABSTRACT** Collecting circular dichroism (CD) spectra for protein solutions is a simple experiment, yet reliable extraction of secondary structure content is dependent on knowledge of the concentration of the protein—which is not always available with accuracy. We previously developed a self-organising map (SOM), called Secondary Structure Neural Network (SSNN), to cluster a database of CD spectra and use that map to assign the secondary structure content of new proteins from CD spectra. The performance of SSNN is at least as good as other available protein CD structure fitting algorithms. In this work we apply SSNN to a collection of spectra of experimental samples where there was suspicion that the nominal protein concentration was incorrect. We show that by plotting the normalized root mean square deviation of the SSNN predicted spectrum from the experimental one versus a concentration scaling-factor it is possible to improve the estimate of the protein concentration while providing an estimate of the secondary structure. For our implementation (51 data points 240 – 190 nm in nm increments) good fits and structure estimates are obtained if the NRMSD (normalised root mean square displacement, RMSE/data range) is < 0.03; reasonable for NRMSD < 0.05; and variable above this. We have also augmented the reference database with 100% helical spectra and truly random coil spectra.

## INTRODUCTION

To extract secondary structure information for globular proteins from circular dichroism (CD) spectra, expert opinion must be sought; this is usually from either a person who has worked in the field for a long time, or a software methodology. A number of such methodologies are available in Dichroweb<sup>1,2</sup> and at the CDPro website.<sup>3,4</sup> However, all such available methodologies are dependent on the accuracy of the protein concentration. In other papers<sup>5,6</sup> we reported the development of SSNN, “Secondary Structure Neural Network”, which is a software package to assign secondary structures using a self-organising map (SOM) methodology. Our approach is similar in intent to the family of k2d programs<sup>7</sup> but more flexible in terms of reference data set and wavelength range. It has also been validated by testing it in a leave-one-out methodology using the CDDATA.48 reference set from CDPro<sup>3,4</sup> as a 47-member training set with one test protein, repeating the test 48 times and comparing with CDSSTR, SELCON3, and k2d.<sup>5,6</sup> CDDATA.48 has structure vectors associated with it and may be found on the CDPro website (<http://lamar.colostate.edu/~sreeram/CDPro/main.html>), which is maintained by Sreerama *et al.* at Colorado State University.<sup>8</sup> The structure labels used are written as a vector throughout this work and refer to: ( $\alpha$ -helix, distorted  $\alpha$ -helix,  $\beta$ -sheet, distorted  $\beta$ -sheet, turn, other) as in references.<sup>8,9,10</sup> The assignments come from DSSP annotation with the 2 residues at each end of helices and 1 residue at each end of  $\beta$ -strands being taken

as distorted.<sup>11</sup> In this work the structure vectors are quoted as fractions of 1 in this order. Total  $\alpha$ -helix content is thus the sum of the first two components and total  $\beta$ -sheet content is the sum of the third and fourth components. Pure CD structure types produce spectra similar to those seen in Figure 1.



**Figure 1:** The CD spectra of the proteins of reference set CDDATA.(48+5) with the most extreme structure types: an extrapolation of the CD of peptide Aurein 2.5 to model 100%  $\alpha$ -helix; rat intestinal fatty acid binding protein protein is the highest  $\beta$ -sheet content in the reference set (58.4%); Azurin has the largest turn content in the reference set (31.2 %); N-formyl acetic acid is 100% random coil protein.

SSNN proceeds in three distinct units. In the first unit, SSNN1, the spectra of a reference set of known proteins (with known structures) is organised on a map of chosen size so that similar spectral shapes are neighbours. All nodes on the map are given spectra interpolated between those of the original reference set. In SSNN2, secondary structure vectors corresponding to the spectrum of that node are assigned to all nodes. In SSNN3, the best position on the map for the spectrum of an unknown protein is found and its structure vector determined from its position. For a given reference set, SSNN1 and SSNN2 need only be run once.

SSNN performed at least as well as other available fitting methodologies in our earlier work.<sup>5,6</sup> Its worst structure suggestions were where unknown proteins were on the edge of the structures map or where the intensity in the 208–222 nm region gave a relatively small spectral NMRSD (normalised root mean square displacement, see below) but the shape of the experimental spectrum was reminiscent of an  $\alpha$ -helix and that of the predicted model spectrum  $\beta$ -sheet (or conversely). Our motivation in developing a new structure fitting method was to have an approach that could be used in a wide variety of situations. Our first attempts to apply SSNN ‘in the real world’ were not entirely successful for two completely different reasons. The first reason was that some proteins and peptides (particularly the latter) ended up on the very edge of the spectral map because our reference set did not include very high helical or completely unfolded spectra. The second was that despite our best efforts and those of our colleagues, estimates of protein concentration were never as accurate as we thought. This has the automatic consequence for any fitting methodology of introducing an error into the estimates of secondary structure.

In this work we have therefore augmented the SSNN reference set with 5 spectra created to represent 100%  $\alpha$ -helical proteins and 100% ‘random coil’, bringing the reference set up to 53 proteins. Here random coil refers to the spectrum observed for an unfolded peptide. Its structure vector is 100% ‘other’, *i.e.* (0,0,0,0,0,1). We have also developed a way of using SSNN to improve the estimates of protein concentration.

## METHODS

A SOM is a type of unsupervised neural network that takes high-dimensional data (in our case CD spectra) and clusters them, then visualises it in a manner that is much easier to understand than the high-dimensional data set. SSNN is described in reference <sup>5</sup> and the details of how to implement the code are given in <sup>6</sup>. The process of training the SOM with a reference data set (SSNN1) is first to make a matrix of

$n \times n$  (in this case  $40 \times 40$ ) vectors containing pseudo-random numbers. Then one selects, at random, a protein spectrum from the reference set and compares it with all of the random spectra in the matrix. The random vector in the ‘spectra map’ that has the smallest Euclidean distance from the selected protein spectrum is called the best matching unit (BMU) and is the ‘winner’. Next this BMU is made more similar to the selected spectrum using a learning rule, which makes the numbers in the random vector more similar to the protein spectrum. At the same time the vectors in the map near the BMU, the neighbourhood, are made more similar to the protein spectrum as well, but to a lesser degree in a distance (map coordinate)-dependent way. This is done for thousands of iterations, 28000 in our case. Once this is finished, the spectra map is trained, and ready for the next stage. Due to the random selection, the physical appearance of the maps change each time SSNN1 is run, though the clustering will be the same and thus ultimate fits should not change. One layer of a trained SOM can be seen in Figure 2a for our augmented version of CDDATA.48 which we denote CDDATA.(48+5). The figure shows the clustered CD spectra in a  $40 \times 40$  spectra map for the 222 nm data point. This only shows one of the stack of 51 data points for each spectrum, so there are 50 other nm points for each spectrum, and thus a stack of 50 other maps could be produced. The map shows the 53 reference set spectra, and also virtual interpolated spectra filling in the gaps to make up the 1600 spectra.

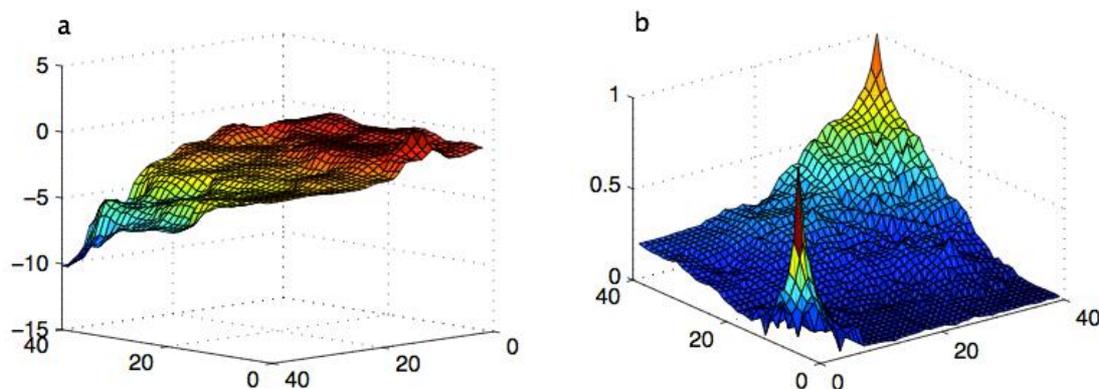
The second module, SSNN2 takes the clustered spectra and constructs a structures map by finding the coordinates of the BMUs of the 53 proteins in the spectra map and placing their structures at the same coordinates in the structures map. For the virtual structures, SSNN2 takes a distance-weighted sum of 5 of the structures of neighbouring spectra from the reference set. A typical result is shown in Figure 2b for the  $\alpha$ -helix. There is a structures map for each of the 6 structure types used in this work. The two peaks in Figure 2b in this map show that not all  $\alpha$ -helix-rich proteins have the same spectra and hence structure vectors.

In SSNN3, a model of a test spectrum is determined (*e.g.* Figure 5a(ii)) as a weighted sum of its 5 BMUs (positions of BMUs on the SOM are also given in the output files). The structures then follow from the same weighted sum as in the structures map. The model or predicted (fitted) spectrum has an NRMSD (normalised root mean squared deviation) associated with it, and this is used to indicate how much the structures prediction can be trusted. We use these definitions of RMSD and NRMSD,

$$RMSD = \sqrt{\frac{\sum_{i=1}^N (S_i - M_i)^2}{N}} \quad (1)$$

$$NRMSD = \frac{RMSD}{M_{\max} - M_{\min}} \quad (2)$$

where  $S_i$  are the elements of the real spectrum, and  $M_i$  are elements of the model spectrum,  $N$  is the number of data points in a spectrum (51 in this case).  $M_{\max}$  and  $M_{\min}$  are the largest and smallest observed values, in this case the largest and smallest values in the model spectrum being evaluated.



**Figure 2:** (a) Spectral intensity map for 222 nm CD signals. (b) Structures map for  $\alpha$ -helix structure vector component after optimisation starting from a reference set of 53 proteins which includes CDDATA.48 from CDPPro and an additional 5 spectra (see text).

In this implementation of SSNN we have chosen to represent CD spectra as vectors of 57 numbers, the first 51 being the  $\Delta\epsilon$  intensities (where the concentration is that of amino acid residues not protein molecules) for 240–190 nm in 1 nm steps and the last 6 being the structure vector components. In our previous work<sup>5</sup> we showed that with a reference set of 48 spectra, the SOM size of 40×40 nodes or spectra was optimal. The same will be true for 53 spectra, though a large increase or decrease would require a larger or smaller map. A version of SSNN (SSNNGUI) is available pre-trained with CDDATA.(48+5) at [http://www2.warwick.ac.uk/fac/sci/chemistry/research/arodger/arodgergroup/research\\_intro/instrumentation/ssnn/](http://www2.warwick.ac.uk/fac/sci/chemistry/research/arodger/arodgergroup/research_intro/instrumentation/ssnn/) and the instructions for its use are given in details in reference<sup>6</sup>. The test spectra must then be matching 51-number column vectors. Alternatively SSNN1\_2.app may be trained on reference sets for any wavelength range or set of proteins, as also described in reference<sup>6</sup>. The key innovation of this work is to have made both SSNNGUI and SSNN1\_2.app useable when one only has an estimate of the protein concentration. In this case a concentration scaling factor range should be entered on the GUI (graphical user interface) and also a step size. Bearing in mind that more calculations take longer to perform, it is preferable to do a coarse-grained calculations first then refine the step size for a smaller range. A spectral NRMSD against concentration scaling factor plot will be additional output if these parameters are entered on the GUIs.

#### *Insulin and polylysine sample preparation and data collection*

Insulin and polylysine were obtained from Sigma–Aldrich (insulin from bovine pancreas I6634, polylysine P 4707 MW 70000–150000). For the pH 2.3 insulin, 0.44 mg of insulin was dissolved in NaOH (0.1 M). Sodium phosphate buffer (to final concentration 4 mM) was added, resulting in an insulin solution of ~0.3 mg/ml (concentration calculated by measuring the UV absorbance at 278 nm and using the Sigma–Aldrich extinction coefficient of 6080 mol<sup>-1</sup>cm<sup>-1</sup>dm<sup>3</sup>). The pH was adjusted to 2.3 with HClO<sub>4</sub> 0.1M and it was diluted by factor of 4. For the zinc-free pH 7.3 sample, a similar procedure was followed, but EDTA (equal molarity to insulin) was added with the phosphate buffer and no HClO<sub>4</sub> was required. Polylysine was made to nominal 0.10 mg/mL in water and adjusted to pH 11.2 with NaOH resulting in a nominal 0.094 mg/mL solution.

For the UV absorbance measurements, a Jasco V-660 spectrophotometer was used. All the CD spectra were taken in a Jasco J-815 of J-715 CD spectropolarimeter. The balance used was a Mettler Toledo XP2U and the pH meter Mettler Toledo Seven Compact pH/Ion S220 InLab Nano Sensors.

## **RESULTS AND DISCUSSION**

### *Applying SSNN3 to real data*

Following our successful leave-one-out validation of SSNN using spectra from the CDPro website,<sup>12,13</sup> we embarked on applying SSNN to real data. This project had mixed success until we considered the fact that although the proteins all had a nominal value of 0.1 mg/mL, this concentration was unlikely to be correct. We therefore wrote a version of SSNN3 that processes an input spectrum through SSNN several times, each time multiplying the spectrum vector by different factors, which we call the ‘concentration scaling factors’. As illustrated in the examples below, we found that the plot of NRMSD versus scaling factor often has a single minimum. If the value of the minimum NRMSD is small (<0.03) we are confident this scaling factor gives a reasonable estimate of the true concentration and secondary structure estimate. It is advisable to view the output for a number of scaling factors near the minimum if the structure estimates differ significantly—this is particularly true if *e.g.* the low wavelength data quality is poor. For larger NRMSDs, visual inspection of the overlay of model and experimental spectra is advisable as discussed below.

For ease of use we have made a single GUI option for using SSNN3 pre-trained with reference set CDDATA.(48+5). We previously used it with one single concentration but have added the option to scale the concentration automatically. This application is denoted SSNNGUI.app. More advanced users can use a re-trainable version, which includes SSNN1 and SSNN2 as well as SSNN3, called SSNN1\_2.app. SSNN is written in MATLAB, and for the GUI, MATLAB’s GUIDE (graphical user interface development environment) was used. SSNN3-single, runs SSNN once for each unknown protein in the test set to determine the secondary structures of the proteins in question at the stated concentration. SSNN3-multiple allows the user to select a range of concentration scaling factors and gives secondary structure analyses and NRMSD values for each scaling factor each as part of the output. The GUI of SSNN3 takes less than 2 seconds to run (per spectrum) on a 2009 MacBook Pro laptop with 8 GB RAM, and a 2.26 GHz processor once the MATLAB Compiler Runtime is installed (for details see reference<sup>6</sup>). The training process of SSNN1 and SSNN2 running for 28,000 iterations takes about 20 minutes to run. SSNNGUI is available for

Mac and Windows at  
[http://www2.warwick.ac.uk/fac/sci/chemistry/research/arodger/arodgergroup/research\\_intro/instrumentation/ssnn/](http://www2.warwick.ac.uk/fac/sci/chemistry/research/arodger/arodgergroup/research_intro/instrumentation/ssnn/).

Guide lines for the formatting of the input files can be found in the instructions text file on the SSNN website. The output is plots to show where the BMUs making the model spectrum are relative to the reference set members, an overlay of the model (or predicted) spectrum and the original experimental spectrum, along with the spectral NRMSD value (see *e.g.* Figures S12(a-d) in SI A, and 5a(ii) below). When SSNNGUI is run in multiple mode, a plot of NRMSD versus scaling factor is also produced. The results reported in this paper have been determined using the SSNNGUI.

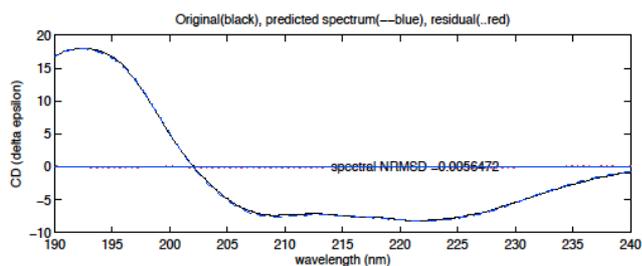
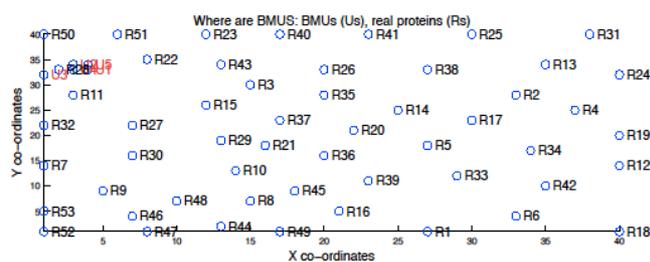
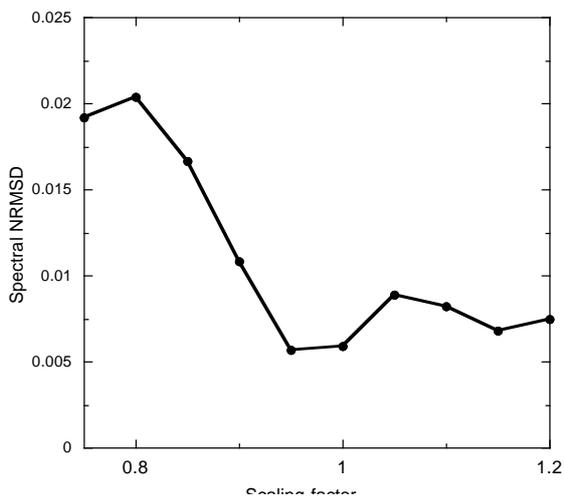
### *Applications*

The remainder of this paper is structured around particular examples that illustrate aspects of the performance of SSNN. The first example is a selection of the results of an extensive study of CD spectra of insulin at different pH and different concentrations to illustrate the strengths and weaknesses of applications of SSNNGUI to solution-phase protein structure and concentration determination. In order to undertake a systematic study to indicate the extent to which we could rely on SSNN to determine protein secondary structures and concentrations we chose to focus on insulin.

#### *Membrane peptide: $\alpha$ -helix*

The CD of an  $\alpha$ -helical peptide whose concentration was known fairly accurately is shown in Figure ?? The plot of spectral NRMSD vs scaling factor suggests 0.95 (83% helix) and 1.0 (84% helix) are best fits, with 0.9 (81%) and 1.05 (85%) still having very good fits. All have 0%  $\beta$ -sheet content. We therefore conclude this peptide has 83 $\pm$ 2% helix, 0% sheet, 5 $\pm$ 1% turn and 12 $\pm$ 1% Other. As discussed below for mixed structure systems it may be appropriate to declare a bigger uncertainty.

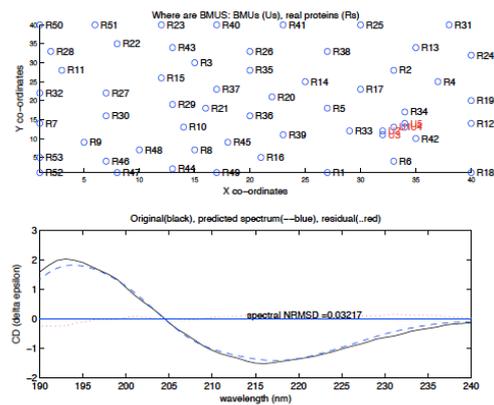
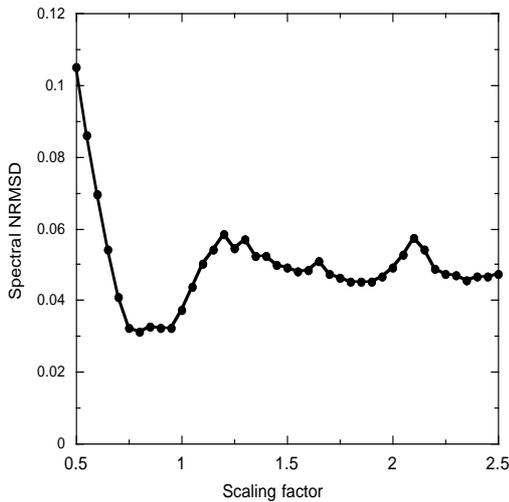
Concentration scaling factor	$\alpha$ -Regular	$\alpha$ -Distorted	$\beta$ -Regular	$\beta$ -Distorted	Turn	Other
0.9	0.624	0.184	0.003	0.002	0.052	0.136
0.95	0.655	0.176	0	0	0.048	0.121
1	0.67312	0.16736	0	0	0.045	0.114
1.05	0.69919	0.15312	0	0	0.043	0.105



*Polylysine:  $\beta$ -sheet*

Polylysine (~1.1 mg/mL) was dissolved in H<sub>2</sub>O and the pH was adjusted to 11.4 with NaOH then heated at 55°C for 30 min to produce a  $\beta$ -sheet.<sup>14,15</sup> As shown in Figure ??, the  $\beta$ -sheet structure has minimum spectral NRMDS with a scaling factor of 0.9 and structure vector (0.020,0.052,0.29,0.14,0.20,0.29). With this highly sheet protein (and others we tested, data not shown), the accuracy of the concentration is not a great concern as the  $\alpha$ -helical percentage was 7% and  $\beta$ -sheet percentage was 44% for concentration scaling factors ranging from 0.65–1.2. The local minima in the NRMDS plots at higher scaling factors are clearly bad fits as they look helical.

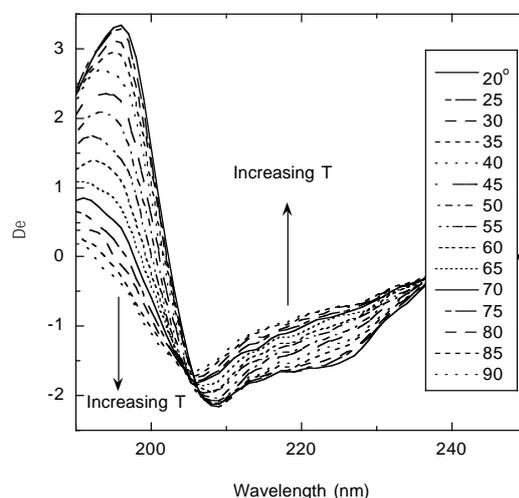
We chose to analyse polylysine because it is a simple system which is deemed to give the archetypical  $\alpha$ -helix,  $\beta$ -sheet, and random coil spectra under different conditions. After extensive work (data not shown) we remain unconvinced that the pH 11.2 room temperature polylysine structure is a pure  $\alpha$ -helix, at least in our laboratory.



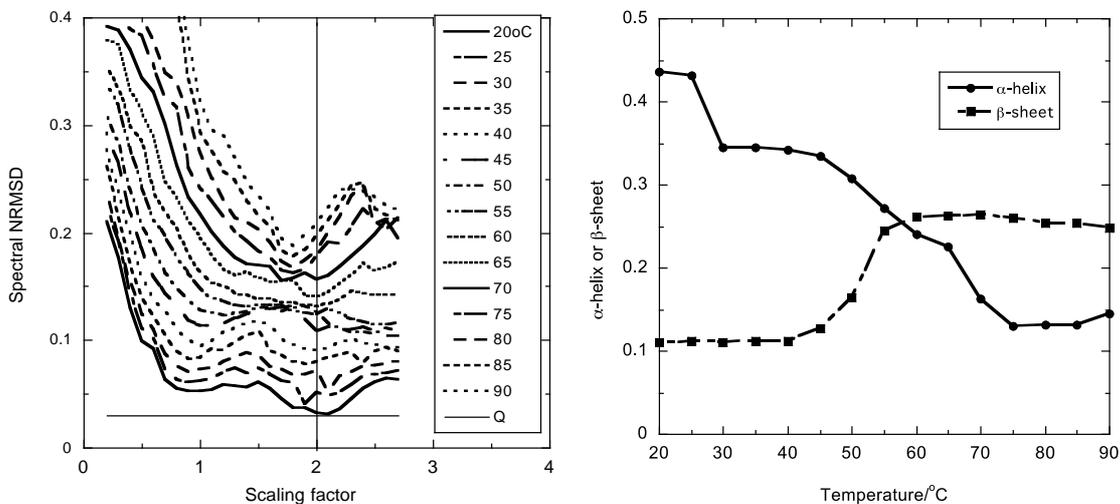
### Insulin

Insulin is a small 2-peptide protein whose crystal structure for the neutral pH zinc containing protein (PDB entry 4INS) was annotated to have structure vector (0.29, 0.23, 0.02, 0.04, 0.05, 0.36) for Woody and Sreerama's data base.<sup>4</sup> Insulin is a challenging protein to get and keep in solution, and its structure varies with pH and whether it has zinc present or not. Despite being extensively studied, its structural details remain unclear in some environments, so resulting data will be useful. In addition, in leave-one-out testing with SSNN and SELCON3 this fairly helical protein was the third worst structure analysis.<sup>5</sup> The results shown below are for insulin at pH 2.4, which is a zinc-free structure as it illustrates the limitations of SSNN effectively.

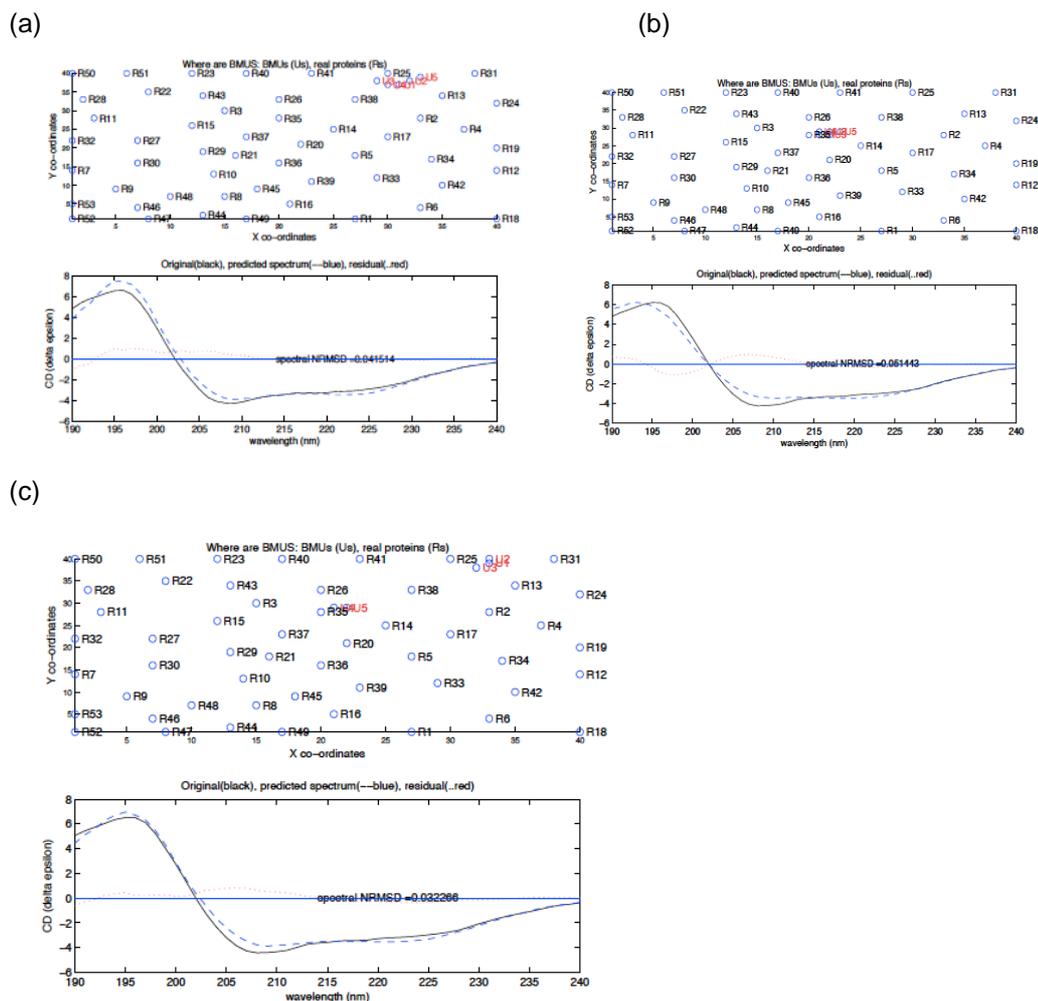
CD spectra for pH 2.4 insulin are shown in Figure 3 as a function of temperature (assuming nominal 0.1 mg/mL concentration). The SSNN spectral NRMSDs are plotted in Figure 4a on a displaced vertical scale and the predicted helical and sheet content are plotted as a function of temperature using scaling factor 2.0 for the first 11 spectra and 1.8 for the last 4 (some evaporation occurred during the experiment as seen by the absorbance signals). The room temperature structure predictions are ~9% less helical than both the crystal and the zinc-free pH 7.3 solution data (not shown) which is in accord with the available low pH NMR data (PDB 2HIU,<sup>16</sup> for insulin with the B chain carboxy-terminus native alanine mutated to threonine) which is 39% helix (according to DSSP annotation). The peculiar step in Figure 3b between 25°C and 30°C provides an interesting illustration of another value of the variable concentration methodology as a means of estimating reliability of structure predictions. As shown in Figure 4a, none of the spectra NRMSDs are below the somewhat arbitrary quality threshold of 0.03. The factor 2.0 BMUs and fits are illustrated in Figures 5a and b for 25°C and 30°C and the factor 2.1 for 30°C. The BMUs for the two 2.0 fits are in slightly different parts of the map and that for 2.1 spans both parts. This instability reflects the fact that the reference set does not contain the 'right' type of helical spectra which also raises difficulty for the peptide in trifluoro ethanol discussed below.



**Figure 3:** (a) CD spectra of insulin (pH 2.4) as a function of temperature (assuming nominal concentration 0.1 mg/mL in 1 mm path length cuvette).



**Figure 4:** (a) NRMSD for SSNN output for data of Figure 3. Plots are vertically displaced by 0.01 for each temperature increment, with 20°C at the bottom. (b) SSNN  $\alpha$ -helix and  $\beta$ -sheet as a function of temperature, using scaling factor 2 for 20–70° and 1.8 for the remainder.

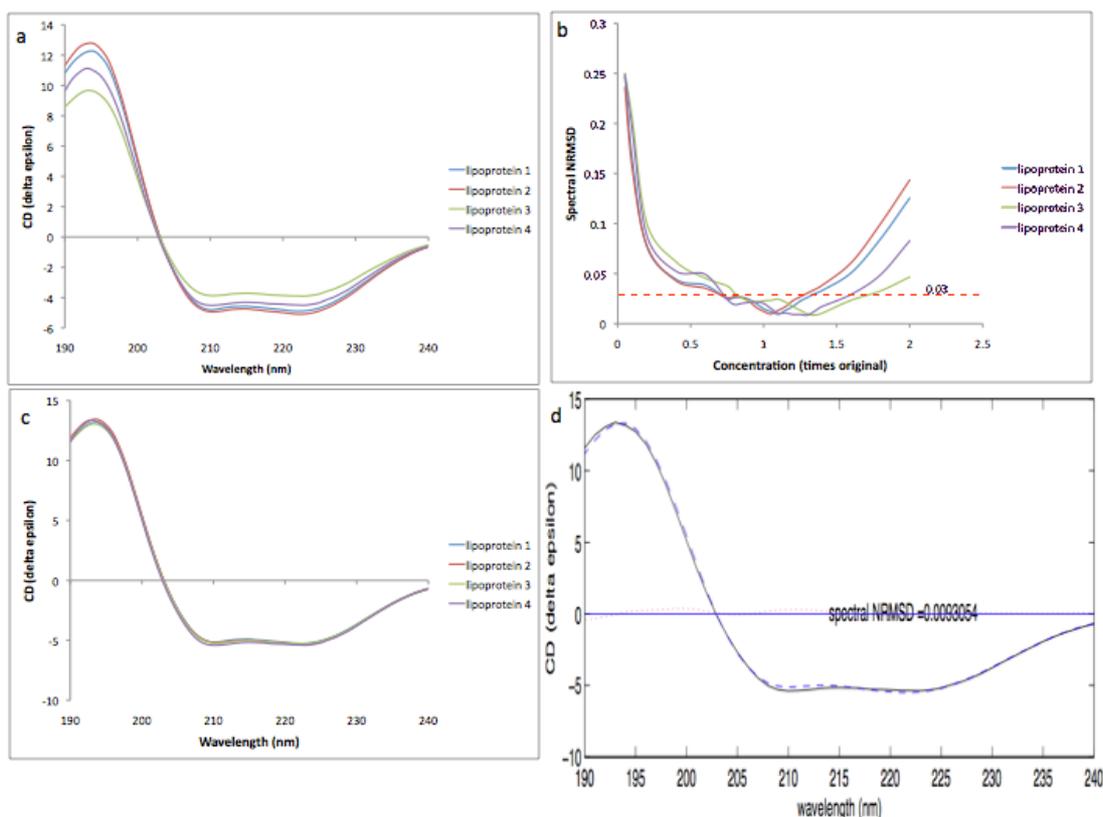


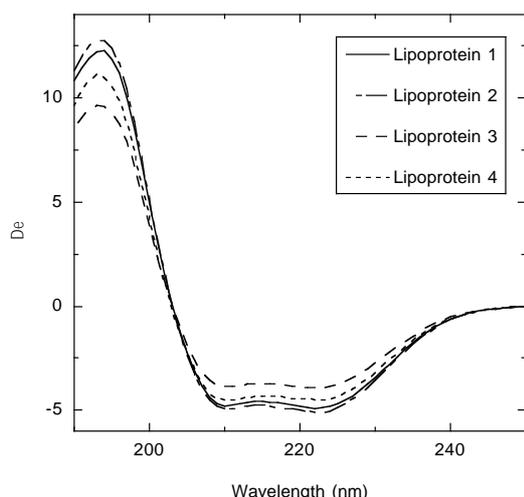
**Figure 5:** SSNN output for the pH 2.4 insulin data of Figure 3. (a) 25° C with concentration scaling factor 2.0, (b) 30° C with concentration scaling factor 2.0 (34% helix), (c) 30° C with concentration scaling factor 2.1 (41% helix).

### Lipoproteins

CD spectroscopists frequently use protein concentrations in units of mg/mL since, for a pure protein, 0.1 mg/mL corresponds to ~ 910–950  $\mu$ M amino acid residue concentration which gives a good far UV (*i.e.* amide chromophore) CD spectrum in a 1 mm cuvette (as long as the buffer does not absorb light significantly). However, by definition, lipoproteins include lipids which are invisible in the spectrum but contribute to the mass and presumably also affect protein concentration determination methods. Figure 6a shows the overlay of the CD spectra of 4 high density lipoproteins (L1–L4), which were all thought to be 0.1 mg/mL in concentration. Literature (*e.g.* <sup>10</sup>) led us to expect helical content when folded of between 50% and 80%. Some of our spectra are almost identical in spectral shape, but differ in magnitude. These spectra had been collected in our laboratory and abandoned, as the results could not be interpreted with available structure fitting methodologies.

Plots of spectral NRMSDs versus concentration scaling factor for L1–L4 are shown in Figure 6b. The NRMSD minima are for scaling factors: 1.05, 1.05, 1.35, and 1.2 for L1 to L4 respectively. The different predictions for neighbouring scaling factors give an indication of the percentage errors in the fits. Thus we conclude that the original protein concentrations were respectively 0.093, 0.095, 0.074, 0.083 mg/mL (within  $\pm 5\%$ ). Figure 6c shows an overlay of the spectra rescaled with these scaling factors and Figure 6d illustrates the quality of the model spectrum overlaid with the original for L4 (the worst fit of the 4 proteins) indicating we can trust the secondary structure vectors to within 1–2% (difference between vector for minimum NRMSD and the neighbouring points, see SI Tables SI and SII). L1 and L3 are 63% helix, 6–7 % sheet and L2 and L4 are 65% helix, 6% sheet with errors (as determined by nearest scaling factor fits) of ~3% for the helices and 2% for the sheets.



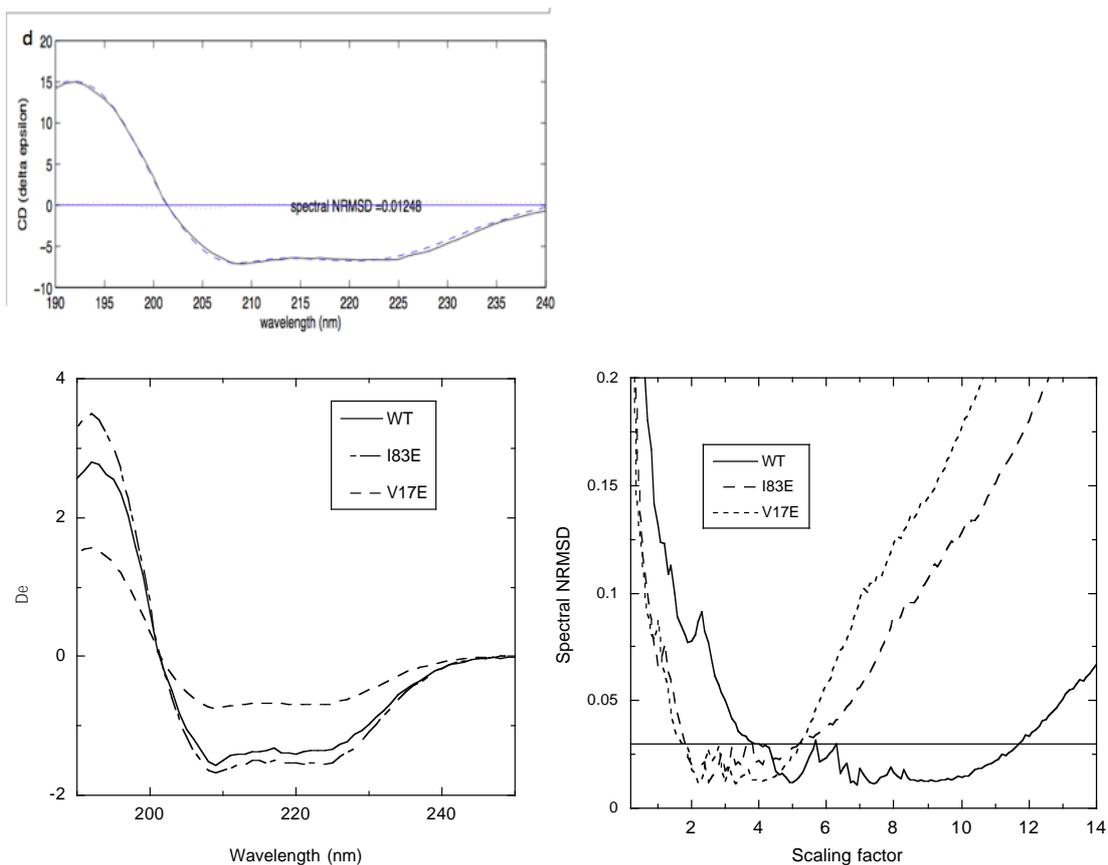


**Figure 6:** (a) CD spectra of 4 lipoproteins, L1–L4, converted to  $\Delta\epsilon$  assuming original protein concentration was 0.1 mg/mL in a 1 mm path length cuvette and the average amino acid molecular mass was 105  $u$ . (b) SSNN3-multiple NRMSD for the 4 proteins versus concentration scaling factor. (c) SSNN3 output for L4 with concentration scaling factor 1.2. Please send me data to plot b – I can't find it

#### *ZapA: wild type and mutants*

*Escherichia Coli* ZapA is a 104-residue protein which binds to the cell division protein FtsZ.<sup>17,18,19</sup> We had been interested in whether the protein's structure changed when key residues were mutated. Although we had collected CD data (Figure 7a), our analysis had been hindered by very inaccurate concentration determinations. We therefore implemented SSNNGUI and plotted the NRMSD versus concentration scaling factor. The answer is less clear than in the previous example with the NRMSD curves oscillating significantly and having a large fairly flat region. Despite the oscillations, I83E is a minimum in the region of scaling factor 2.5 with a good fit between scaled experiment and model spectra (Figure S4b). Inspection of the model spectra for WT and V17E overlaid with the different scaling factors of the experimental data (see *e.g.* Figure 7d for WT), show that the different NRMSD minima optimize different parts of the spectrum. We chose the 8.8 and 4.10 minima for respectively the WT and V17E mutant as the best estimates because these fits do not emphasise one part of the spectrum at the expense of others and because they are in flat parts of the NRMSD curves. (Note: we would expect a flatter NRMSD plot in the 8.8 region of the spectrum as a 0.1 step in the 1.0 region equates to a 0.9 step at 8.8). The corresponding best-fit structure vectors are then similar with WT: (0.40, 0.20, 0.02, 0.03, 0.13, 0.22); and V17E: (0.43, 0.20, 0.02, 0.02, 0.11, 0.22). I83E has a fairly flat NRMSD curve for scaling factors between 2.5 and 4 (whose BMUs are on the edge of the map). The two extrema structure vectors are (0.26, 0.17, 0.08, 0.06, 0.16, 0.27) and (0.29, 0.18, 0.045, 0.047, 0.17, 0.25).

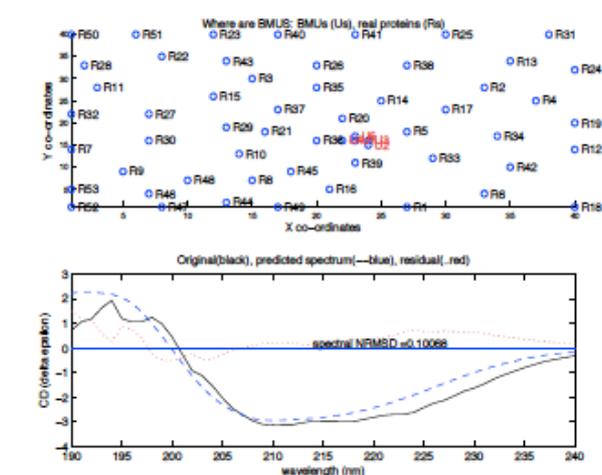
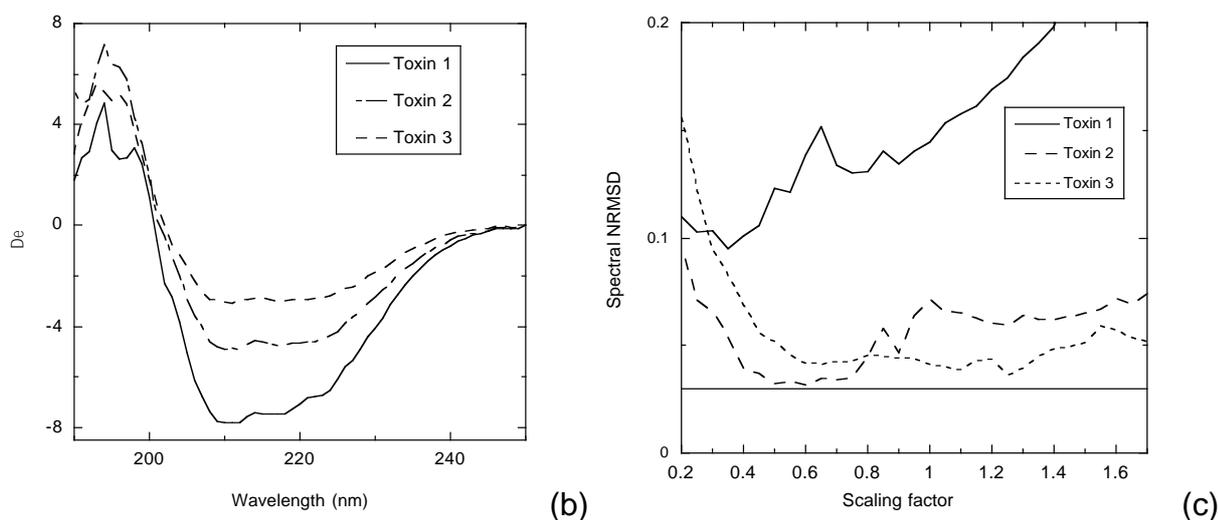
Even allowing for the ranges suggested in the predictions, I83E is significantly less helical than the other two proteins. According to the *Pseudomonas aeruginosa* crystal structure<sup>19</sup> we would expect ~ 64% helix and ~10% sheet for the WT ZapA. These percentages compared well with total helix (~60%) and total sheet (~5%) components of the above structure vectors. However, I83E is predicted to be ~45%  $\alpha$ -helix and ~12%  $\beta$ -sheet. Even allowing for errors in the concentration estimates (and hence the structure vectors), we may conclude that the I83E mutant disrupts some of the helical character of ZapA but the V17E mutant does not. Consistent with this is the fact that position 83 is in the coiled-coil region of the protein whereas position 17 is at the end of strand 2.<sup>18, 19</sup>



**Figure 7:** CD spectra of ZapA. (a) WT, I83E and V17E with  $\Delta\epsilon$  determined assuming nominal concentration 0.1 mg/mL in 1 mm pathlength cuvette and the average amino acid molecular mass 116 u. (b) SSNN3-multiple NRMSD for the 3 proteins versus concentration scaling factor with 0.03 quality indicator shown. (c) Overlay of experiment and model spectra for scaling factors = 4.1 for V17E. *include BMU??*

### Toxins

CD data were collected for a series of related bacterial proteins (toxins) as shown in Figure 8a. All samples had been dialysed to remove high concentrations of salt that interfered with CD data collection resulting in unknown concentrations. SSNNGUI was implemented with a range of scaling factors. Toxins 2 and 3 have spectra of almost the same shape, but the poor quality of the spectra at low wavelength result in slightly different structure predictions, respectively 29% helix and 20% sheet for toxin 2 with scaling factor 0.6 and 32% helix and 12% sheet for toxin 3 with scaling factor 1.1. No reasonable fit emerged for Toxin 1 led us to re-examine the raw CD data files, which showed HT voltages above 600 V below 200 nm for Toxin 1 and below 196 nm for the others. So the short wavelength data were at fault not the fitting methodology. We therefore prepared new protein samples whose typical minimum NRMSD was 0.016 giving 35% helix and 12% sheet (the small increase reflection the previous attenuation of the low wavelength signal).



**Figure 8:** (a) Toxin CD spectra with  $\Delta\epsilon$  determined assuming nominal concentration 0.1 mg/mL in 1 mm pathlength cuvette and the average amino acid molecular mass 105 u. (b) SSNN3-multiple NRMSD for the 3 proteins versus concentration scaling factor with 0.03 quality indicator shown. (c) Overlay of experiment and “best fit” (see text for comment) model spectra for Toxin 1, scaling factor = 0.2.)

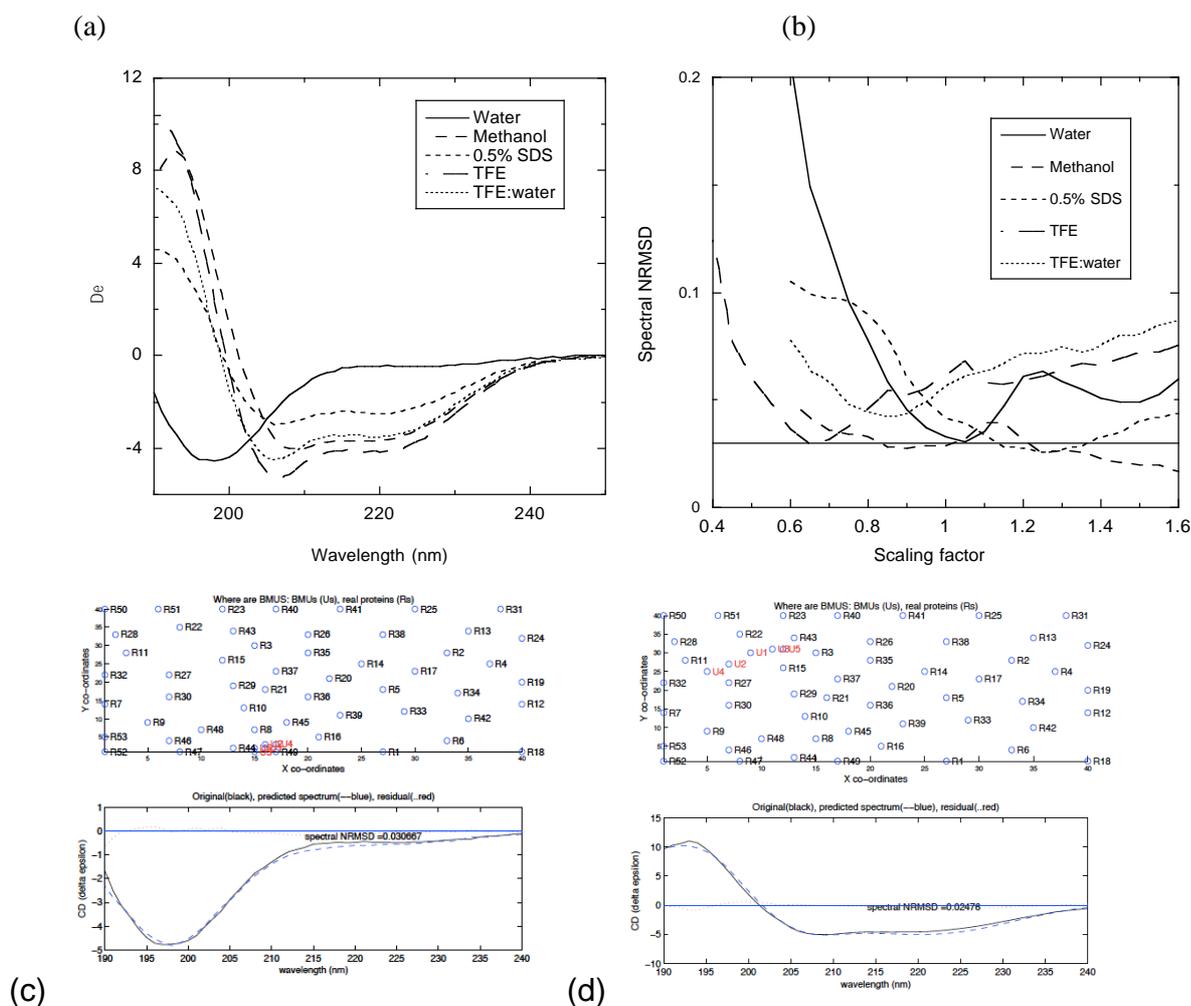
### Peptide structure fitting

Peptides are a challenge to structure fitting programs as they tend to adopt a single secondary structure motif with frayed ends. Further, any sample may be a mixture of populations. We wished to see whether SSNN could at least be used to rank relative amounts of folding for peptides. Initial attempts to use SSNN with CDDATA.48 resulted in BMUs at the very edge of the maps and poor fits (see *e.g.* Figures SI7a and b?? delete) for unfolded peptides and for peptides suspected to be well-folded helices. With hindsight this should not have been surprising given CDDATA.48 contained data only from globular proteins which all have a mix of secondary structure motifs. To produce better peptide structure estimates we enhanced CDDATA.48 with 5 more spectra. Three spectra to mimic unfolded peptides were included: MSLSRRQAAQASGIALCAGAVPLKASA in water taken from reference,<sup>20</sup> and the spectra for N-formyl acetic acid and N-acetyl valine,<sup>21</sup> which have 100% ‘random coil’ structure. Two more reference spectra were constructed by taking the spectrum for myoglobin (from CDDATA.48) and for a helical aurein peptide<sup>22</sup> and scaling them to have the accepted maximum magnitude value of  $-13 \text{ mol}^{-1}\text{cm}^{-1}\text{dm}^3$ <sup>23</sup> at 222 or 208 nm. Scaled-myoglobin and the aurein were taken to have 100 %  $\alpha$ -helix structure.

To test the usefulness of SSNN with the enhanced reference set for assessing secondary structures of peptides we used the data from a systematic study of a 27-mer SufI signal peptide from the *Escherichia coli* Tat system<sup>17</sup> with sequence: MSLSRRQFIQASGIALCAGAVPLKASA. Here we consider the peptide structure in water, methanol, 0.05% SDS, TFE, and TFE:water (Figure 9a). The overlay of the NRMSD plots versus concentration factor in Figure 9b suggest that the nominal concentration of 0.1 mg/mL is fairly close to the true value for all solvents except TFE. The water model spectrum for scaling factor 1.05 (Figure 9c), gives a good fit to experiment letting us conclude that the prediction of only a small amount of secondary

structure (5% helix, 9% sheet) is correct. One of the BMUs for this fit is number 49, one of our additions to the training set. The structure predictions for the neighbouring scaling factors suggest a possible 2% error.

The methanol NRMSD plot (Figure 9b) is fairly flat, with multiple minima. The factor 1.25 fit looks best (Figure 9d) suggesting methanol induces 51% helix and 10% sheet. The factor 1.0 vector (0.30,0.18,0.070,0.055,0.12,0.29) leads us to conclude that in this case error is of the order of 5%. The other three solvents all have reasonable but not good fits at their NRMSD minima that suggest partial folding in each case with similar helix/sheet percentages being 32%/17% for SDS factor 1.2, 36%/16% for TFE factor 0.65 model spectra, and 33%/16% for TFE:water factor 0.85. To confirm these values we would suggest some titration experiments with different mixed solvents. One would expect TFE to induce more helical structure than methanol, so we suspect that the quality of the fit is suffering from a lack of reference spectra where 208 nm is significantly larger in magnitude than 222 nm.



**Figure 9:** (a) CD spectra (with conversion to  $\Delta\epsilon$  per amino acid performed by assuming a concentration of 0.1 mg/mL in a 1 mm path length cuvette) of Sufl peptide in different solvents. Data taken from reference <sup>20</sup>. (b) SSNN3-multiple NRMSD for Sufl in different solvents versus concentration scaling factor with 0.03 quality indicator shown. (c) Overlay of experiment and best fit model spectrum for Sufl in water [scaling factor = 1.05, structure vector (0.02,0.03,0.056,0.036,0.06,0.80)]. (d) Overlay of experiment and best fit model spectrum for Sufl in methanol [scaling factor = 1.25, structure vector: (0.32,0.19,0.051,0.046,0.15,0.25)].

On the basis of this peptide example, we can conclude that when used with care, inspecting the fits, and considering the location of the BMUs on the map, SSNN can be a useful tool for peptides. In the original analyses of reference <sup>20</sup> it was assumed that the peptide could only adopt a helical or a random coil structure and percentages were determined using the CD magnitude at 222 nm. The results of using SSNN indicate a more complicated structural landscape for this 27-mer, which is in accord with the small size (or lack of with TFE) of the dip in CD intensity at ~216 nm expected between the 208 nm and 222 nm negative maxima for a helical protein.

## Conclusion

**We previously recommended a greater variety of  $\beta$ -sheet spectra be added to the reference set due to the fact that magnitudes of  $\beta$ -structure CD is less related to concentration than is the case for  $\alpha$ -helices. While this remains the case, it is only a big problem for highly  $\beta$ -sheet proteins where the mixed structure members of the reference set do not cover the space. In this case it transpires that the structural fit is largely independent of concentration**

### Importance of high quality data for structure fitting

#### State errors

**Reference set needs to cover the space. In general this is indicated by poor spectra NRMSDs. We have also noted that SSNN-multiple has the added advantage of making it clear when this is a problem as neighbouring scaling factors will probably settle on different parts of the map.**

**. One would expect TFE to induce more helical structure than methanol, so we suspect that the quality of the fit is suffering from a lack of reference spectra where 208 nm is significantly larger in magnitude than 222 nm.**

In this paper we have shown how SSNN can be used to provide fairly reliable secondary structure estimates for proteins, even when the concentration of the sample is not known, as long as the spectral NRMSD versus concentration scaling factor plots have a clear minimum. Our experiences as summarized in the results section suggests that for the size of spectrum vector used in this work (51 data points spaced at 1 nm intervals), an NRMSD<0.03 gives a reasonable structure prediction. Where an oscillating NRMSD versus concentration scaling factor is observed, this is usually due to the set of BMUs changing as a function of concentration scaling factor. In such a situation an educated eye may be able to discern the best fit. Sometimes a poor fit reflects inadequacies in the reference set used to train SSNN, other times it reflects deficiencies in the experimental data such short wavelength data being significantly attenuated due to low photon count. In some cases the results may suggest one of a small number of possibilities and an alternative technique such as infrared absorbance, Raman or Raman optical activity may be required to select between them. A series of spectra as a function of temperature may help clarify the appropriate scaling factor.

Most of the work reported in this paper has been performed with a GUI for the final module SSNN3-multiple. It takes as input the spectral SOM from SSNN1 and the structure map from SSNN2 as well as the experimental spectrum as a vector of 51 data points from 240 nm to 190 nm in 1 nm intervals. If a different wavelength range is required or a change in the reference data set is proposed, SSNN1 and SSNN2 will need to be rerun using the SSNN1\_2.app. With the hindsight of the peptide work reported above, we have augmented CDDATA.48 reference set from CDPro to include 100% helical structures and 100% random coil structures thus extending the parameter space. The SSNNGUI that is now available at [http://www2.warwick.ac.uk/fac/sci/chemistry/research/arodger/arodgergroup/research\\_intro/instrumentation/ssnn/](http://www2.warwick.ac.uk/fac/sci/chemistry/research/arodger/arodgergroup/research_intro/instrumentation/ssnn/) was trained with this larger reference set of 53 protein spectra. The fitting could be improved further by *e.g.* adding TFE-induced structures to a reference set. The lack of appropriate spectra in a reference set is usually apparent because the BMUs for a protein are very near the edge of the SOM rather than surrounded by other BMUs.

In looking for examples on which to test this concentration-optimising version of SSNN, we had many old data sets that had never been analysed once it had become apparent that we did not know the concentration. We are now able to proceed further with such data sets. A key to further progress will be enhancement of the training sets with particular classes of proteins, *e.g.* membrane proteins. Collecting the data for new spectra is one aspect of this, however, equally important is determination of the structure vector.

## Acknowledgements

VH thanks EPSRC for a PhD studentship through the MOAC Doctoral Training Centre grant number EP/F500378/1 and MS acknowledges funding from the FP7 Marie Curie Innovative Doctoral Training Programme.

## References

1. Whitmore L, Wallace BA. DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Res* 2004;32(Web Server issue):W668-73.
2. Whitmore L, Wallace, B.A. Protein secondary structure analysis from circular dichroism spectroscopy: methods and reference databases. *Biopolymers* 2008;89:392-400.
3. Sreerama N, Woody RW. A self-consistent method for the analysis of protein secondary structure from Circular dichroism. *Analyt. Biochem.* 1993;209:32-44.
4. Sreerama N, Woody RW. Estimation of Protein Secondary Structure from Circular Dichroism Spectra: Comparison of CONTIN, SELCON, and CDSSTR Methods with an Expanded Reference Set. *Analyt. Biochem.* 2000;287:252–260.
5. Hall V, Nash, A., Hines, E., Rodger, A. Elucidating protein secondary structure with circular dichroism and a neural network. *Journal of Computational Chemistry* 2013;34(32):2774-2786.
6. Hall V, Nash, A., Rodger, A. SSNN, a method for neural network protein secondary structure fitting using circular dichroism data. Submitted to *Analytical Methods*.
7. Andrade MA, Chacon P, Merelo JJ, Moran F. Evaluation of Secondary Structure of Proteins from Uv Circular-Dichroism Spectra Using an Unsupervised Learning Neural-Network. *Protein Engineering* 1993;6(4):383-390.
8. Sreerama N, Venyaminov S.Y., Woody, R.W. Estimation of protein secondary structure from CD spectra: Inclusion of denatured proteins with native protein in the analysis. *Anal. Biochem.* 2000;287:243-251.
9. Lees JG, Miles AJ, Wien F, Wallace BA. A reference database for circular dichroism spectroscopy covering fold and secondary structure space. *Bioinformatics* 2006;22(16):1955-1962.
10. Sevugan Chetty P, Mayne L, Kan Z-Y, Lund-Katz S, Englander SW, Phillips MC. Apolipoprotein A-I helical structure and stability in discoidal high-density lipoprotein (HDL) particles by hydrogen exchange and mass spectrometry. *Proceedings of the National Academy of Sciences* 2012;109(29):11687-11692.
11. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577–2637.
12. Hall V, Nash A, Hines E, Rodger A. Elucidating protein secondary structure with circular dichroism and a neural network. *J. Comp. Chem.* 2013.
13. Hall V, Nash A, Rodger A. SSNN, a method for neural network protein secondary structure fitting using circular dichroism data. Submitted to *Analytical Methods* 2013.

14. Shibata A, Yamamoto M, Yamashita T, Chiou J-S, Kamaya H, Ueda I. Biphasic Effects of Alcohols on the Phase Transition of Poly( L-lysine) between  $\alpha$ -Helix and  $\beta$ -Sheet Conformations. *Biochemistry* 1992;31:5728-5732.
15. Greenfield N, Fasman GD. Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry* 1969;8:4108-4116.
16. Hua QX, Gozani SN, Chance RE, Hoffmann JA, Frank BH, Weiss MA. Structure of a protein in a kinetic trap. *Nat. Struct. Biol.* 1995;2:129-138.
17. Pacheco-Gomez R, Cheng X, Hicks MR, Smith CJ, Roper DI, Addinall S, Rodger A, Dafforn TR. Tetramerization of ZapA is required for FtsZ bundling. *Biochem J*;449(3):795-802.
18. Small E, Marrington, R., Roder, A, Dafforn, T.R. FtsZ polymer-bundling by the *Escherichia coli* ZapA orthologue, YgfE involves a conformational change in bound GTP. *J. Mol. Biol.* 2007;369:211-221.
19. Low H, Moncrieffe M, Lowe J. The Crystal Structure of ZapA and its Modulation of FtsZ Polymerisation. *J. Mol. Biol.* 2004;341:839-52.
20. Miguel MS, Marrington R, Rodger PM, Rodger A, Robinson C. An *Escherichia coli* twin-arginine signal peptide switches between helical and unstructured conformations depending on hydrophobicity of the environment. *Eur. J. Biochem.* 2003;270:3345-3352.
21. Gokce I, Woody RW, Anderluh G, Lakey JH. Single peptide bonds exhibit poly(pro)II ("random coil") circular dichroism spectra. *J. Am. Chem. Soc.* 2005;127:9700-9701.
22. Nordén B, Rodger A, Dafforn TR. Linear dichroism and circular dichroism: a textbook on polarized spectroscopy. Cambridge: Royal Society of Chemistry; 2010. 304 p.
23. Berova N, Nakanishi K, Woody RW, editors. Circular dichroism principles and applications. 2nd ed. New York: Wiley-VCH; 2000.