Analytical Methods

RSCPublishing

ARTICLE

SSNN, a method for neural network protein secondary structure fitting using circular dichroism data

Cite this: DOI: 10.1039/x0xx00000x VINCENT HALL,[a,b] ANTHONY NASH,[a,b,c] ALISON RODGER*[b,d]

Received (in XXX, XXX) Xth XXXXXXXXX 2013, Accepted Xth XXXXXXXXX 20XX

Circular dichroism (CD) spectroscopy is a quick method for measuring data that can be used to determine the average secondary structures of proteins, probe their interactions with their environment, and aid in drug discovery. This paper describes the operation and testing of a self-organising map (SOM) structure-fitting methodology named Secondary Structure Neural Network (SSNN), which is a methodology for estimating protein secondary structure from CD spectra of unknown proteins using CD spectra of proteins with known X-ray structures. SSNN comes in two standalone MATLAB applications for estimating unknown proteins' structures, one that uses a pre-trained map and one that begins by training the SOM with a reference set of the user's choice. These are available at http://www2.warwick.ac.uk/fac/sci/chemistry/research/arodger/arodgergroup/research_intro/instrumentation/ssnn/ as SSNNGUI and SSNN1_2 respectively. They are available for both Macintosh and Windows formats with two reference sets: one obtained from the CDPro website, referred to as CDDATA.48 which has 48 protein spectra and structures, and one with 53 proteins (CDDATA.48 with 5 additional spectra). Here we compare SSNN with CDSSTR, a widely-used secondary structure methodology, and describe how to use the standalone SSNN applications. Current input format is $\Delta\varepsilon$ per amino acid residue from 240 nm to 190 nm in 1 nm steps for the known and unknown proteins and a vector summarising the secondary structure elements of the known proteins. The format is readily modified to include input data with *e.g.* extended wavelength ranges or different assignment of secondary structures.

Received 00th October 2013,
Accepted 00th January 2013

DOI: 10.1039/x0xx00000x
www.rsc.org

Introduction

Circular dichroism spectroscopy is the difference in absorbance of left and right circularly polarized light. It is probably most often used to estimate the percentages of different secondary structures that are present in proteins. Assuming that high quality data have been collected for the sample of interest and the concentration is accurately known, then the question arises as to what method is best used to assign secondary structure motifs quantitatively. It is now generally recognised that the best approach is to use spectral data and secondary structures for an extensive reference set of proteins and then implement a process to estimate the secondary structure content of the unknown protein.[1-6]

Some CD spectral features are readily apparent, for example, an α-helix is characterised by a large positive band at 190 nm (part of a $\pi\rightarrow\pi^*$ exciton couplet), and two smaller negative bands at 208 nm (the other $\pi\rightarrow\pi^*$ component) and 222 nm ($n\rightarrow\pi^*$). β-sheets usually show a positive peak between 195 nm and 202 nm, and a negative signal between 215 nm and 220 nm, though sometimes resemble the 'random coil' and poly-proline II spectra which have a negative signal at 200 nm.[4] It is now widely accepted that this 200 nm negative band is dominated by contributions from residues with poly(proline) type-II conformations.[7] β-turns have a large negative band at 180 nm–190 nm, a positive signal in the 200 nm–205 nm range ($\pi\rightarrow\pi^*$), and a negative signal at 225 nm ($n\rightarrow\pi^*$).

A number of secondary structure analysis programs, based either in statistical methods or intelligent systems, exist that make quantitative assignments of percentages of structure type.[*e.g.* 1, 8-16] It is possible to make use of some of these on Dichroweb, an online server hosted at Birkbeck, University of London.[17] The commonly used statistical packages which are available on Dichroweb include: CONTIN which is a ridge regression technique; CDSSTR (an update of 'VARSLC') which is a variable selection, or feature selection, method; and SELCON (now SELCON3) which is a self-consistent method together with a singular value decomposition, SVD, algorithm. Dichroweb includes one intelligent system approach, called K2d,[10] which is a self-organising map (SOM) neural network approach. Although the intelligent systems approaches appears to have many advantages, K2d and its successors including K2D2/3 [13, 15, 18] and SOMCD,[14] have not been widely adopted by the CD community.

As these methods are only available as pre-trained SOMs where the reference set and structural categories have been defined by the original authors, we chose to develop our own CD structure fitting SOM: Secondary Structure Neural Network (SSNN). This was done so we could test it back-to-back with statistical methods and enable any user to train it with new spectral reference sets and different structure assignment methodologies if the researcher wished to do so. This is timely as new data bases are currently being developed—greatly facilitated by the recently established Protein Circular Dichroism Data Bank.[19]

In summary, SSNN has three independent modules that operate in sequence.

SSNN1: takes spectra for a set of proteins of known secondary structure content (the reference set) and trains (organises) them so that related spectra are put near each other on the map. The map has many more nodes to put spectra in than there are spectra, so the gaps are filled-in with intermediate, virtual spectra. These virtual spectra are made by taking weighted sums of the nearby experimentally-obtained spectra.

SSNN2: puts vectors of the secondary structure contents of the reference proteins onto a structures map that matches the output of SSNN1, with structure vectors created for the virtual spectra by using the same weighting for the virtual nodes as used for the spectra.

SSNN3: takes as input a CD spectrum of a structurally unknown protein (currently in units of $\Delta\varepsilon/(\mathrm{mol}^{-1}\mathrm{dm}^{3}\mathrm{cm}^{-1})$, where the concentration is that of amino acid residues rather than molecules of protein) and produces as output an estimate of its secondary structure, a model spectrum, and the spectral NRMSD (normalised root mean squared deviation) defined as

$$NRMSD = \frac{\sqrt{\left(\frac{\sum_{i}\left(x_{i,experiment} - x_{i,model}\right)^{2}}{N}\right)}}{M - m}$$

where x_i is the value at each wavelength (or structure), N is the number of data points, M is the largest intensity, and m is the smallest, so $(M–m)$ is the range. SSNN1 and SSNN2 need only be performed once for a given reference set.

In a previous paper [11] we determined the parameters required to optimize the performance of SSNN and showed that it compared well with SELCON3, K2d, and SOMCD in a leave-one-out comparison using CDDATA.48 from the CDPro web site (http://lamar.colostate.edu/~sreeram/CDPro/). At this time we made SSNN3 available pre-trained but without detailed instructions for use. The overall performance of SSNN-47 and SELCON3-47 was similar: SELCON3-47 "won" for 23 out of 48 spectra, being slightly better on average for α-helix and Other structure estimates. Whereas SSNN-47 "won" for 25 out of 48 spectra and on average was slightly better for mixed α-helix/β-sheets, and β-sheets and turns. Comparisons between SSNN and K2d and SOMCD were hampered by lack of re-trainable executable versions of K2D and SOMCD. So we worked with what was available in Dichroweb [4] and the literature [14] creating a comparison methodology that dis-favoured SSNN. In summary, even when given a handicap, SSNN out-performed K2d and performed better than SOMCD for 22 of the 33 proteins for which results were available in reference [14].

The aim of this paper is to provide other CD users, including those in the biopharmaceutical industry, with the tools required to use all three modules of SSNN, thus enabling them to work with new reference sets and structure definitions should they so desire. We also show that SSNN has a firm place in the tool-kit for CD structure analysis by showing that it compares very well with CDSSTR, which Woody and Sreerama previously showed was better than SELCON3 for β-distorted structures.[3]

Methods

Reference data set

The reference data sets used by us to train SSNN in this work are CDDATA.48 (with data from 240 nm–190 nm taken from the CDPro website: http://lamar.colostate.edu/~sreeram/CDPro/) and CDDATA.48+5 which is CDDATA.48 augmented with 5 additional spectra, 2 representing 100% α-helix and 3 representing 100% Other structures (some of these are extrapolations). The spectra are given in per residue molar absorbance units ($\Delta\varepsilon$= mol^{-1} dm^{3} cm^{-1}) and the structures are assigned to 6 structure categories. This is the largest available CD reference set that has been consistently annotated with secondary structures. Each member of the input reference set has 57 numbers. In order they are: 51 for the spectral intensity at each wavelength from 240 nm to 190 nm, and 6 for the structure values in the order: (α-helix regular, α-helix distorted, β-sheet regular, β-sheet distorted, turns, Other structures). The CDPro website and its references explain the structure types which are summarised in reference [3]. In summary the 'regular' structures are the middles of helices and sheets and the 'distorted' structures the ends. In the leave-one-out tests reported in this paper, each spectrum was removed from the reference set in turn and used as the test spectrum. CDSSTR or SSNN1 and SSNN2 was/were run with the resulting 47-member reference set. This was repeated for each member of the reference set, in so-called leave-one-out cross-validation.

SSNNGUI

To run SSNN3, or "SSNNGUI", pretrained with CDDATA.48+5 file, the correct version for the computer being used should be downloaded from http://www2.warwick.ac.uk/fac/sci/chemistry/research/arodger/arodgergroup/research_intro/instrumentation/ssnn/.

The details of how to install and run SSNNGUI are in the supplementary information. What follows below is an outline of how to proceed and the results that will be obtained. Table 1 summarises the steps once the program has been installed on the user's computer. Figure 1 shows SSNNGUI_versA, the version of SSNN3 that does not need SSNN1 and SSNN2 to be trained before it is used as it already includes the results of training with CADATA.48+5.

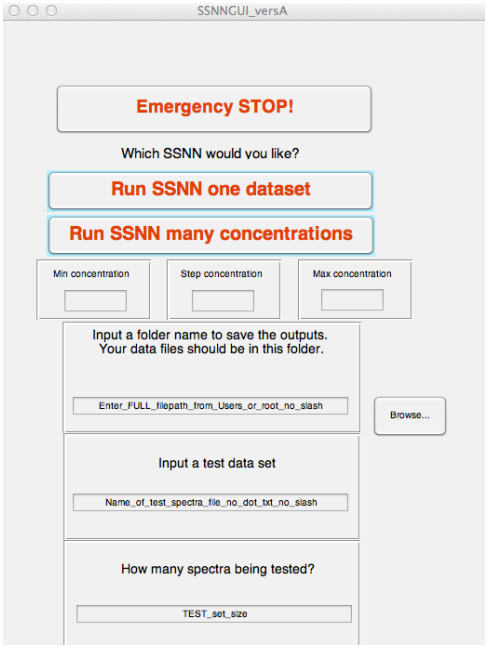


Figure 1: Screenshot of SSNNGUI 2013a.

Table 1. Protocol to run SSNN3 for test proteins of known concentration once the application has been installed (see Methods and Supplementary Information for details).

1. Navigate to the folder containing SSNNGUI_2013a.
2. Click the SSNNGUI.app icon or run the .sh file using Terminal as described in the Supplementary Information.
3. Click on the browse button beside the field that says <Input a folder name to save the outputs...>. Navigate to the folder containing < SSNNGUI_2013a>, and select *any* file in it (you may need to change the selection criterion to <All files>), as long as the folder is correct. The text in the SSNNGUI box will not change.
4. In the field called <Input a test data set> type the name of the test file, for example <toxins experiment 1 Friday>. The file type should be .txt, but do *not* include <.txt> in the file name. The test file should have columns of 51 data points representing the CD spectra from 240 nm to 190 nm. Multiple spectra can be in the same .txt file as tab or comma delimited columns. No other text should be present. See the example file included in the package.
5. In the field <How many spectra being tested?> put the number of spectra in that test file.
6. Hit the <Run SSNN one dataset> button ("one dataset" means one file with, perhaps, multiple spectra). In a few seconds one should be presented with the same number of windows as the number entered in step 5. These windows will contain two plots. The top plot will be the locations of all of the best matching units (BMUs), the bottom plot will be the experimental spectrum of a protein compared with the model of it that SSNN made, and a residual (the difference between the original spectrum and the model spectrum). The txt output files should also appear in an output folder labelled with the input data file name.

SSNN1_2

To make SSNN future compatible we have made SSNN available as a re-trainable package. The user will thus be able to use SSNN with different data sets as they become available. In the supplementary information is a guide on how to use the incarnation of SSNN that can be retrained.

RESULTS AND DISCUSSION

The main result of this paper is the production of a stand-alone pretrained GUI for SSNN, called SSNNGUI. We have also made versions of SSNN1, 2, and 3 all available in one application denoted SSNN1_2. They are available at http://www2.warwick.ac.uk/fac/sci/chemistry/research/arodger/arodgergroup/research_intro/instrumentation/ssnn/.

SSNN1_2 will allow CD practitioners to train the methodology with their own desired reference sets. They may then estimate CD secondary structures, with the confidence of having used their own reference spectra of test proteins. Preferred reference sets might include additional proteins, different methods of structural annotation (including the number of structure types chosen), different wavelength ranges *etc.* This is particularly attractive as more good spectra and structures become available from CD, X-ray crystallography, NMR and other sources.

The SSNN application has various parameters that can be varied to train the program in a way tailored to the reference set used. For example, practitioners might like to change the number of iterations, as SSNN may take longer to learn a larger reference set than the 28,000 iterations that we found was best for about 50 spectra.[11]

Care should be taken in changing parameters. For example, the initial learning rate follows a negative exponential curve throughout the training process. Changing this rate will have an impact on the whole training. Making the initial learning rate too low will cause the spectral map spectra to model the reference set (experimental) spectra too slowly. Making it too high will move spectra into position too quickly, which might not produce a broad enough region for each type of structure. Our experiments showed us learning too quickly can be a bad thing.[11] This may also depend on the size of the map.

The emergency stop button has been included in SSNNGUI, though in our experience it has seldom needed had to be used in in practice. The emergency stop shuts down the application completely, although it takes a little time to do so.

Among other functions, SSNN model spectra residuals highlight which wavelengths of the test protein spectra produce the most error in model spectra. SSNN treats all wavelengths equally, so a larger error in one region of a spectrum could throw off the estimation. If this correlates with a region of the spectrum where the data quality is poor, the user may want to ignore this contribution to the NRMSD. Users should also check that model spectra have appropriate intensities at certain wavelengths. For example β-sheet-rich protein models should have a single negative peak between 215 nm and 220 nm, whereas α-helix-rich proteins should have negative peaks at 208 nm *and* 222 nm.

Comparison of SSNN and CDSSTR

We previously ran SSNN1, SSNN2, and SSNN3 48 times in a leave-one-out methodology using 47 spectra of CDDATA.48 as the reference data set and the omitted one as the test spectrum

each time.[11] We refer to this as SSNN-47. In this way we ensured that the test spectra spanned structural space and also that the structure annotation was the same for the reference set and the test spectra. Table 2 and Table SI1 show the net results (final rows) of the previous work and details of an analogous leave-one-out test performed here for CDSSTR (CDSSTR-47) using the executable version available on the CDPro web site. The sum of the absolute values of the deviations from the fractions of real structures is also given for SELCON3-47 in the final lines of the Table SI1. SELCON3-47 data are from [11]. We also ran the leave-one-out test for SSNN-52 (*i.e.* using CDDATA.48+5 in leave-one-out mode) and have shown the detailed results in Table SI1.

In summary, for the CDDATA.48 reference set SELCON3 is best for high α-helix structures, SSNN is best for medium α-helix and 'Other' structures, and they are joint best for β-sheets. CDSSTR is best for distorted β-sheets. When reduced to 3 structure types α-helix, β-sheet and Turns-plus-Other (data not shown), SELCON3 remains best for α-helices and SSNN is best for β-sheets and Turns-plus-Other. The somewhat disconcerting result given in Table SI1 is that CDSSTR has the best *spectral* NRMSD (Table SI1) but this does not translate into the best structural predictions. It is satisfying to note that the augmented reference set, CDDATA.48+5, leads to the SSNN-52 results being better than all the CDDATA.48 results.

As noted previously,[11] it is important to inspect the model spectrum outputted by a fitting program. The metrics currently used to assess fit do not give extra significance to the 218 nm region of the spectrum, which is a negative maximum for β-sheet structures and a negative minimum between two negative maxima for α-helical structures. This is illustrated for carboxypeptidase A which has low spectral NRMSDs for CDSSTR-47 of 0.075 and for SSNN-47 of 0.067, but high structural NRMSDs of 0.68 and 0.34 respectively. The overlay of their plots is given in Figure 2a, showing the difference in spectral structure that is 'obvious' to the eye but not to an equally weighted numeric estimate of spectral NRMSD. Conversely Rat Intestinal Fatty Acid Binding Protein (Figure 2b) is modelled to be more helical than reality, due in this case to the fact that this highly β-sheet spectrum has unusually large magnitude which is not reflected in the reference set (Figure 2b).

Table 2 shows the performance of the programs on particular classes of proteins. Mixed α/β proteins remain the most challenging for all the programs and a critical human eye is an invaluable tool for assessing the output from any of the programs. The challenge in estimating the Other class is that it is an amalgam of various undefined structures that have different spectral features. Putting these into one class implies they have something in common—which is misleading.

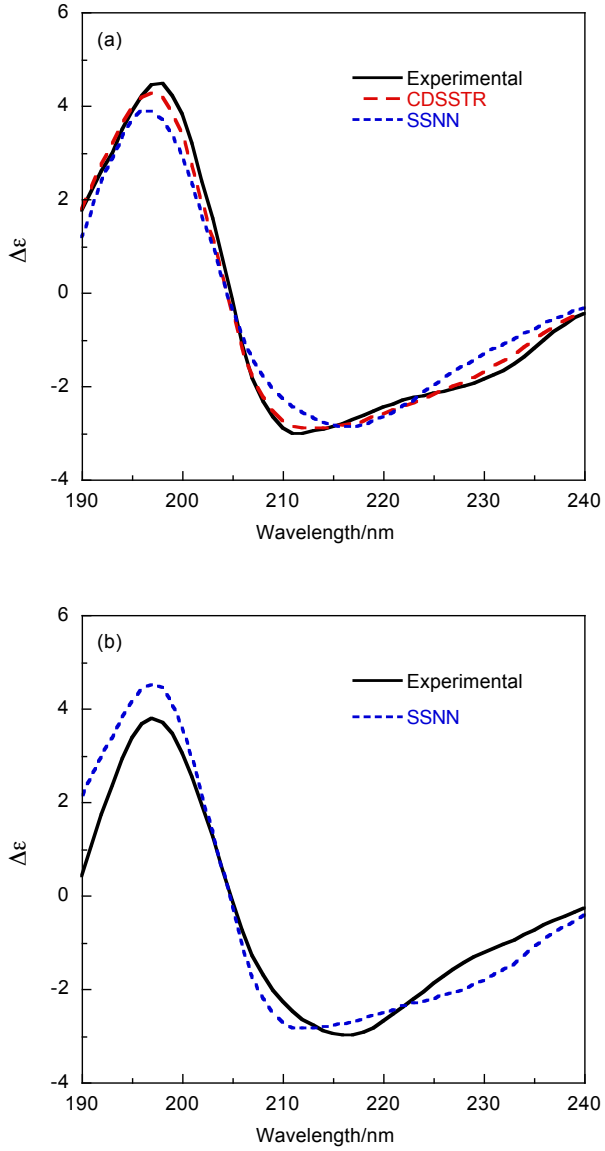


Figure 2. (a) Carboxypeptidase A experimental CD spectrum and best fit from CDSSTR-47 and SSNN-47. (b) Rat Intestinal Fatty Acid Binding Protein experimental spectrum and best fit from SSNN-47.

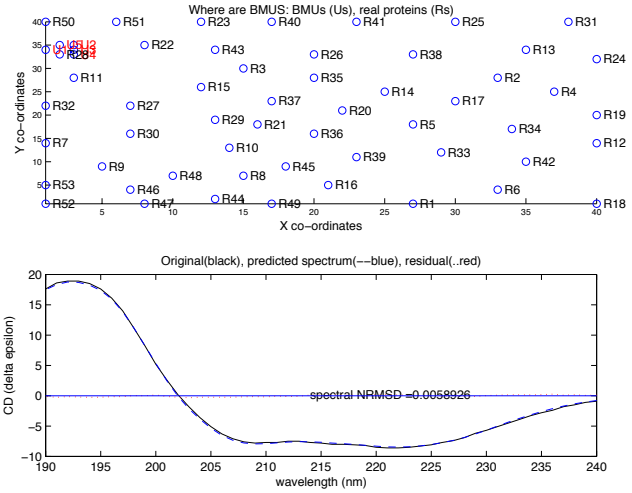
Table 2. Structural NRMSDs for SELCON3-47, SSNN-47, CDSSTR-47, and SSNN-52 from the data in Table SI1 for different structural classes of protein. "Errors" are one standard deviation of the variation between proteins in the class.

Program	Overall 6-column	> 50% α-helix	30%–50% α-helix	>30% β-sheet	>50% Other
SELCON3-47	0.2±0.2	0.1±0.1	0.3±0.3	0.3±0.2	0.2±0.2
CDSSTR-47	0.3±0.2	0.2±0.2	0.3±0.2	0.2±0.2	0.3±0.2
SSNN-47	0.2±0.2	0.2±0.1	0.2±0.2	0.2±0.2	0.2±0.1
SSNN-52	0.1 ± 0.1	0.10±0.07	0.14±0.08	0.2±0.2	0.10±0.06

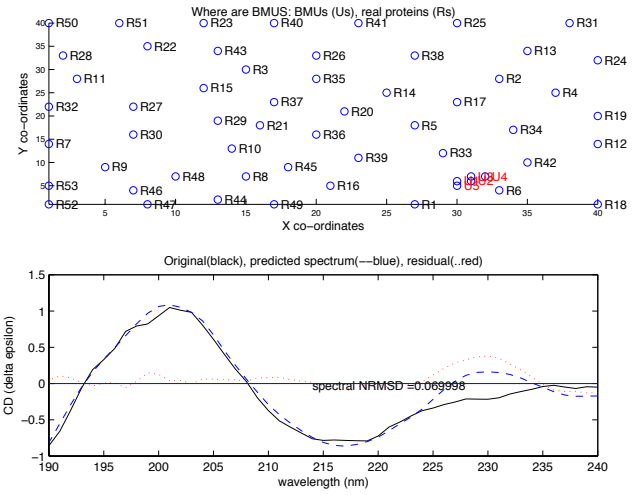
Some examples

To further illustrate SSNN and to give a new user some known examples to test their installation of the software, data sets for the biopharmaceutical product spectra of Figure 3 are included on the web site. The highly helical protein shows an exceedingly good fit with 84% helix and 0% sheet. The antibody spectral fit is not as good, but the key spectra features are reproduced with 5% helix and 38% sheet which is consistent with a slightly relaxed version of antibody crystal structures *e.g.* PDB 1IGT which is annotated by DSSP to be 6% helix and 47% sheet [20] and consistent with CDSSTR which suggests the spectrum is 3% helix and 35% sheet. The mixed-structure asparaginase spectrum of Figure 3c has a small NRMSD and indicates 31% α-helix and 16% β-sheet, which compares with the crystal structure values of 31% and 23%.[21] The lower estimates of sheet content from CD and SSNN on solution data compared with X-ray diffraction on crystals are not surprising given the dynamic nature of proteins in solution.

(a)



(b)



(c)

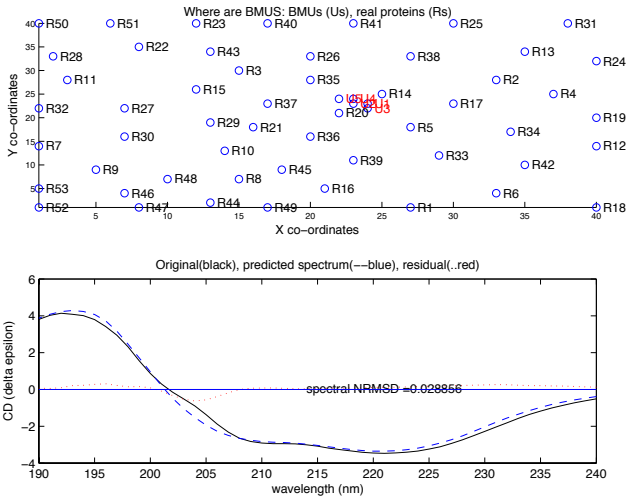


Figure 3. Test spectra (*i.e.* experimental data of unknown structure), SSNN model CD spectra, and BMU plots for some biopharmaceutical proteins. (a) A highly helical protein with an extremely good model spectrum. The SSNN structure vector (0.67, 0.17, 0.00, 0.00, 0.04, 0.11). (b) An antibody spectrum with moderately good fit. The SSNN structure vector (0.01, 0.04, 0.23, 0.15, 0.23, 0.34). (c) Asparaginase with an extremely good model spectrum. The SSNN structure vector (0.17, 0.14, 0.09, 0.07, 0.23, 0.31). Input data sets are available in SI.

Conclusions

We have made all parts of a neural network self organising map method of protein secondary structure determination from circular dichroism spectroscopy available as a stand-alone program. SSNNGUI.app is a version that has been pre-trained with a currently available reference set. If a different reference set is desired, the first and second modules, SSNN1 and SSNN2 need to be run once for every new reference set used. These, and SSNN3 for structure estimation, are available in the second of two GUIs, named SSNN1_2.app. The output from the first two modules is internally used as input to SSNN3 which gives a secondary structure estimate for unknown proteins. All spectral data (reference set and test proteins) are formatted in our implementation as columns of 51 intensity points at 1 nm resolution from 240 nm to 190 nm in units of $\Delta\varepsilon$ (where the concentration is that of amino acids). Any other resolution or units can be used as long as the reference and test spectra use the same set. This means that a SOM secondary structure fitting methodology is now available for use with new reference sets, *e.g.* for membrane proteins.

We have also shown that SSNN compares well with the statistical programs CDSSTR and SELCON3.[11] Although on average the methods can each be deemed to perform best for some secondary structures. For proteins known to have a high α-helix content, SELCON3 should be used to estimate structures. Both CDSSTR and SSNN are better at estimating β-sheet-rich proteins, and for intermediate and 'Other' proteins, SSNN is best. In practice by augmenting CDDATA.48 with 5 more reference spectra we have enhanced the performance of SSNN beyond that of the other methods in all but two cases. In practice, to obtain the most accurate picture of the structures of proteins it is advisable to use a few different secondary structure estimation methodologies. If they agree then there can be confidence in the results and if they differ or the NRMSDs are high then care must

be taken. Due to the simple metrics currently used by all fitting programs to assess goodness of spectral fit, we recommend complementing fitting programs with a visual inspection of the overlay of experimental data and model spectrum, particularly in the 215 nm region of the spectrum.

SSNN can be trained with pretty much any reference set for which structures are available. Thus SSNN could easily be adapted to a variety of data; examples could include CD data for membrane proteins, and infrared spectroscopy data.

Acknowledgements

We thank the Engineering and Physical Sciences Research Council for the funding for Vincent Hall and Anthony Nash through the MOAC Doctoral Training Centre (Grant number EP/F500378/1). Meropi Sklepari's help with the manuscript is gratefully acknowledged.

Notes and references

[a] *Molecular Organisation and Assembly in Cells Doctoral Training Centre, University of Warwick. Coventry, CV4 7AL, UK*

[b] *Department of Chemistry, University of Warwick. Coventry, CV4 7AL, UK. Fax: 44 24 76575795; Tel: 44 24 76574696; E-mail: a.rodger@warwick.ac.uk*

[c] *Now at Department of Chemistry, University College London, WC1H 0AJ, UK*

[d] *Warwick Centre for Analytical Science, University of Warwick. Coventry, CV4 7AL, UK.*

† Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/b000000x/

1. N. Sreerama and R. W. Woody, *Analyt. Biochem.*, 1993, 209, 32-44.
2. R. W. Woody, in *Circular dichroism principles and applications*, eds. K. Nakanishi, N. Berova and R. W. Woody, VCH, New York, 1994.
3. N. Sreerama and R. W. Woody, *Analyt. Biochem.*, 2000, 287, 252–260.
4. L. Whitmore and B. A. Wallace, *Nuc. Acids Res.*, 2004, 32, W668-673.
5. B. A. Wallace and R. Janes, eds., *Modern Techniques for Circular Dichroism Spectroscopy*, IOS Press, Amsterdam, 2009.
6. N. Berova, K. Nakanishi and R. W. Woody, eds., *Circular dichroism principles and applications*, Wiley-VCH, New York, 2000.
7. R. W. Woody, *J. Am. Chem. Soc.*, 2009, 131, 8234-8245.
8. S. W. Provencher, *Compter Phys. Comm.* , 1978, 27, 229-242.
9. W. C. Johnson, *Proteins Struct. Funct. Genet.*, 1999, 35, 307-312.
10. M. A. Andrade, P. Chacon, J. J. Merelo and F. Moran, *Prot. Eng.* , 1993, 6, 383-390.
11. V. Hall, A. Nash, E. Hines and A. Rodger, *J. Comp. Chem.*, 2013, 34, 2774–2786.
12. N. J. Greenfield, *Analytical Biochemistry*, 1996, 235, 1–10.
13. C. Louis-Jeune, M. A. Andrade-Navarro and C. Perez-Iratxeta, *Proteins: Struc. Func. Bioinf.* , 2012, 80, 374-381.
14. P. Unneberg, J. J. Merelo, P. Chaco and F. M. n, *PROTEINS: Structure, Function, and Genetics*, 2001, 42, 460-470.
15. C. Perez-Iratxeta and M. A. Andrade-Navarro, *BMC Structural Biology* 2008, 8:25.
16. V. Hall, M. Sklepari and A. Rodger, *Chirality*, 2014.
17. A. Lobley, L. Whitmore and B. A. Wallace, *Bioinformatics* 2001, 18, 211-212.
18. http://www.ogic.ca/projects/k2d3/, *accessed on 14th March 2012*.
19. B. A. Wallace, L. Whitmore and R. W. Janes, *Proteins-Structure Function and Bioinformatics*, 2006, 62, 1-3.
20. L. J. Harris, S. B. Larson, K. W. Hasel and A. McPherson, *Biochemistry*, 1997, 36, 1581-1597.
21. J. Lubkowski, M. Dauter, K. Aghaiypour, A. Wlodawer and Z. Dauter, *Acta Crystallogr.,Sect.D*, 2--3, 59, 84.