Protein secondary structure determined from independent and integrated infra-red absorbance and circular dichroism data using the algorithm SELCON

Søren Vrønning Hoffmann[a]*, Nykola C. Jones[a], Alison Rodger[b]

Protein circular dichroism (CD) and infrared absorbance (IR) spectra are widely used to estimate the secondary structure content of proteins in solution. A range of algorithms have been used for CD analysis (SELCON, CONTIN, CDsstr, SOMSpec) and some of these have been applied to IR data, though IR is more commonly analysed by bandfitting or statistical approaches. In this work we provide a Python version of SELCON3 and explore how to combine CD and IR data to best effect. We used CD data in $\Delta\varepsilon$/amino acid residue and scaled the IR spectra to similar magnitudes. Normalising the IR amide I spectra scaled to a maximum absorbance of 15 gives best general performance. Combining CD and IR improves predictions for both helix and sheet by ~2% and helps identify anomalously large errors for high helix proteins such as haemoglobin when using IR data alone and high sheet proteins when using CD data alone.

Introduction

Protein circular dichroism (CD) spectra, particularly with the high-quality data produced by synchrotron radiation sources,[1-3] are usually the first choice for determining the average secondary structure of proteins in aqueous solution. However, infra-red absorbance (IR) data are sometimes used in preference, particularly when high concentration samples or high concentrations of buffer components are present. There are a few studies where CD and IR data have been combined to good effect. In this paper we show the value of integrating both types of spectral data where possible.

When extracting secondary structure percentages from CD data alone we have found it valuable to use different reference sets and different algorithms. SELCON and SOMSpec seem to be the most reliable methods for CD analysis.[4] SELCON [5] for CD is currently available as Fortran code and also *via* the Dichroweb server with a selection of reference sets.[6] SOMSpec is available as a MATLab code.[7]

Infra-red (IR) absorbance spectra, particularly of the amide I band between 1600–1700 cm^{-1}, are also generally recognised to contain information about the protein's secondary structure. The vibrational contribution of the amide I band is dominated by the C=O stretching of the amide group coupled with the in-phase bending of N–H bonds and stretching of C–N bonds[8,9]. A great deal of work has been done on protein IR spectroscopy, but the best way to extract secondary structure information for, for example, regulatory or research purposes remains unclear.

A range of different curve fitting methods, often preceded by band-narrowing,[10,11-14] have been implemented as summarised and illustrated in reference[7]. The recent consensus, e.g.,[15,16] is that 1620–1640 cm^{-1} is attributed to β-sheet, 1640–1650 cm^{-1} to other structures, 1650–1656 cm^{-1} to α-helix, and 1670–1685 cm^{-1} to turns. Errors of between 10−20% were found with band-fitting and derivative-band fitting methods.[7]

Factor analysis methods[17,18] including the BioTools (Jupiter, US) program ProtaTM provides reasonably good structure estimates, but the details of the fittings cannot be interrogated by the user. Oberg et al.[19] have extensively explored the application

[a] ISA, Department of Physics and Astronomy, Aarhus University, 8000 Aarhus C, Denmark.

[b] School of Natural Sciences, Macquarie University, Sydney, NSW 2109, Australia.

† Electronic Supplementary Information (SI) is available at

ORCIDs: AR: 0000-0002-7111-3024, SVH: 0000-0002-8018-5433, NCJ: 0000-0002-4081-6405

of a partial least squares analysis (PLS) using reference sets and concluded that the most important issue is the quality of the reference set. They observed that larger reference sets usually do not perform better than smaller ones, as they may include more 'anomalous' spectra – so it is important to be able to interrogate results rather than simply accept a number. Goormaghtigh et al. had significant success with an approach which identifies three key wavenumbers for the three structural features that can be distinguished in the IR spectrum.[20,21] However, their preference for using a data point in the amide II band for films of proteins is in our experience not transferable to biopharmaceutical formulations, as we have observed that the magnitude of this band varies significantly with different solution components.

We have explored the application of our self-organising map circular dichroism structure fitting algorithm, SOMSpec,[7,22] to the analysis of protein infrared spectra and found it to be more accurate than most of the other methods used. The key feature of SOMSpec is that it organises a reference set onto a map by similarity of spectra shape then places the unknown spectra on the same map to extract secondary structure estimates.

There are limited examples in the literature where both CD and IR data have been used to give better estimates of secondary structure data. Most applications, such as in[14,17,18,23] involve independent consideration of CD and IR spectra, usually with some kind of band-fitting approach for the IR data. Oberg et al. [19] explored combining CD and IR (amide I and II) spectra. They measured CD and IR of 50 proteins and used Partial Least Squares (PLS) regression as their main analysis method with DSSP annotation for the spectra, and also considered principal component analysis followed by multiple regression and SELCON. Overall, they found that in general IR is better for β-sheet and turn estimates and CD for α-helix and 'the rest'. They also noted that when the α-helix content from either CD or IR is noticeably lower than from the other, then the larger one will be closer to reality. In general, using the combined data set gave better estimates, but noting when the independent estimates differed significantly was a good indication of failed analyses. The authors also concluded larger reference sets are better but warned against enhancing a reference set with anomalous spectra.

Although Oberg et al.[19] had success with using their PLS approach with CD and IR data, both independently and combined, and we have extensively used our self-organising map approach with both CD and IR data,[7,22] and could have used it for combined data sets, neither approach has been implemented for routine use. As the CD community has used SELCON successfully for decades and Oberg et al. [19] found it worked as well as their approach, the goal of this work was therefore to make SELCON available for CD, IR Amide I and CD+IR analysis and to test how well it works.

Methods

The SELCON3 routine used in this work is based on an implementation of the algorithm in the MatLab script SelMat,[24] re-written into Python. The new Python-based program and reference set package SSCalcPy includes this implementation of SELCON3 and is available on Zenodo [25] and GitHub.[26] The SSCalcPy package includes two reference sets for CD secondary structure calculations, namely SP175[24] and SMP180[27] obtained from the PCDDB.[28-30] Details of the origin of the original code and the reference datasets are given in the supplementary information as **S1**. The secondary structure assignments for the CD reference data are based on a DSSP[31] method, see **S1** for more details.

The reference set used for SELCON3 analysis of IR data is based on the RaSP50[20,32] data available in the Supplementary Material of the SOMSpec analysis publication.[7] A detailed description for the method of sample preparation and data collection can be found in the 2006 paper of Goormaghtigh et al.[20] A list of the 50 proteins, their SOMSpec annotation, and crystal secondary structure can be found in **S2** of the supplementary information. This reference set contains IR spectra collected on protein samples dried on an ATR crystal and has data in the wavenumber range 1600-1800 cm^{-1}. They scaled their IR data to a maximum of 1 and when combining CD and IR data they scaled the CD spectral intensities by 0.0015.[19] The IR data to be analysed must have a baseline spectrum subtracted making sure that the 2100 cm^{-1} region is flat and the data zeroed at 1718 cm^{-1} prior to normalization.

From the proteins included in RaSP50, SP175 and SMP180 we have identified 28 common proteins where both high-quality IR and CD spectra are available. We used these to produce a combined CD-IR reference set (CD-IR28). For a list of the 28 proteins, see **S3** in the supplemental information. To perform a SELCON3 Leave One Out Validation (LOOV) analysis of both the RaSP50 and the new CD-IR28 reference sets, each spectrum is removed from the reference set and subjected to SELCON3 analysis with the remaining 49(RaSP50)/27(CD-IR28) spectra used as reference sets. The LOOV Python script is included in the SSCalcPy package under the folder "Tools".

Since the IR reference data are scaled to a maximum absorbance of 1 and the CD spectra are in molar extinction ($\Delta\varepsilon$) units, the CD data magnitudes are typically much larger than the magnitude of the IR data. To optimize the SELCON3 analysis of the combined CD and IR data, the IR data have been scaled further to achieve more similar magnitudes between the CD and IR spectra. This scaling factor (IRscale) has been varied between 1 and 20 in the analysis, and we give suggestions for the optimum scaling of the IR spectra as part of the analysis in the Results and Discussion section. A scaling of zero results in analysis of the secondary structure based on the CD spectra alone. This approach is opposite from that taken by Oberg et al. [19] who scaled the CD spectra and used normalised IR spectra.

Results and discussion

The Python SELCON3 code in the SSCalcPy package was tested using CD data files and the reference set SP175 and satisfactorily compared with the results from the server Dichroweb[6,33-35] prior to its use for IR data. First, we validated the performance of SELCON3 for the 50 IR spectra in the RaSP50 reference dataset by performing LOOV analysis and calculating for each protein the difference between the SELCON3 secondary structure (SS_i) and the crystal secondary structure (cSS_i), $\Delta_{IR,i} = SS_i - cSS_i$ for protein i (see **S2** for a list of SS_i and cSS_i). The results are shown in Figure 1 and compared to the similar analysis using the SOMSpec method.[7] The reconstructed spectra generated by the SELCON3 algorithm displayed with the corresponding protein spectrum are shown in **S4** of the supplementary information for each of the members of the RaSP50 reference dataset.
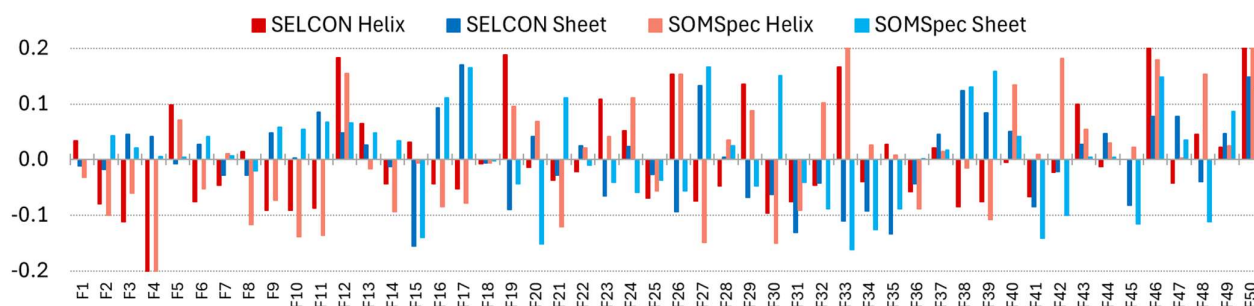

Figure 1: The fractional difference between the calculated helix and sheet content for the 50 proteins in the RaSP50 reference set using SELCON (this work) and SOMSpec (previous work [7]). Helix denotes combined α-helix and 3-10 helix and sheet denotes β-sheet. In order to improve visibility of the smallest values, the scale has been limited to +/- 0.2, so the absolute values of the largest differences are not shown, see the text for a discussion of these outliers

Visual inspection of Figure 1 indicates that SELCON and SOMSpec secondary structure predictions from IR data are of similar reliability. To quantify the performance of both the SELCON3 and the SOMSpec methods on the IR data, two metrics are calculated: the average of the absolute differences, $avg(\Delta_{abs,IR}) = \Sigma_i |\Delta_{IR,i}| / n$, and the standard deviations of the differences $\sigma(\Delta IR)$. For the IR SELCON3 analysis, n is 49 for RaSP50 as one protein is left out for the LOOV. Both are calculated individually for the helix and the sheet differences. A summary of this analysis is shown in Table 1.

Table 1: The overall performance of SOMSpec and SELCON3 for secondary structure predictions

Method	Metric	Helix	Sheet
SOMSpec	$avg(\Delta_{abs,IR})$	0.09	0.07
SELCON3	$avg(\Delta_{abs,IR})$	0.08	0.06
SOMSpec	$\sigma(\Delta_{IR})$	0.12	0.09
SELCON3	$\sigma(\Delta_{IR})$	0.10	0.07

For both of these metrics, SELCON3 performs slightly better than SOMSpec. However, careful inspection of Figure 1 does reveal that in a few cases SOMSpec outperforms SELCON3 (e.g. F47) and for other cases the opposite is true (e.g. F42). The average SELCON helix deviation is 8% and sheet deviation is 6% which are slightly better than the 9% and 7% of SOMSpec. The standard deviations of the errors are also slightly tighter for SELCON being 0.10 and 0.07 versus the 0.12 and 0.09 of SOMSpec.

Overall, there is no evidence for a general under or overprediction of secondary structure, although helical contents of highly helical proteins tend to be under-estimated. As noted previously, highly helical proteins with very similar spectra may have quite different amounts of helix. Haemoglobin (F4) is particularly problematic with the crystal structure having 77% helix and the IR-prediction being only 51%. Also, Metallothionein II (F50) and Soy Trypsin Inhibitor (F46) structures are not well predicted, essentially because they do not retain any helical structure, and for Metallothionein II not even sheet structure, in their crystal structure.

When combining IR and CD reference datasets for analysis, some care should be made in not emphasizing one set over the other, thus skewing the results. To this end, we reduced the IR dataset to include data in the 1600 − 1720 cm^{-1} wavenumber range in 2 cm^{-1} steps. The data outside this range is essentially zero after baseline subtraction and 2 cm^{-1} steps is sufficient to represent the spectral features in the IR data. This brings the number of data points (wavenumbers) for the IR data down to 61, very similar to the 66 data points (wavelengths) in the CD reference dataset.

The LOOV analysis of this combined CD and IR data set containing 28 proteins, CD-IR28 (see **S3** for the full list), was performed for a range of scaling factors (IRscale) of the IR data ranging from 0, i.e. pure CD data analysis, up to 20. We note that for the highest scaling factor, the IR spectrum is significantly larger than the CD spectrum for low helix content proteins. For each scaling factor, the difference between the SELCON3 results and the crystal structure was calculated for each protein, $\Delta_{CD-IR,i}$, and the

metrics $avg(\Delta_{abs,CD-IR})$ and $\sigma(\Delta_{CD-IR})$ derived. The Δ_{CD-IR}, for all proteins in CD-IR28 and for each scaling factor are shown in the Supplementary Information **S5**.
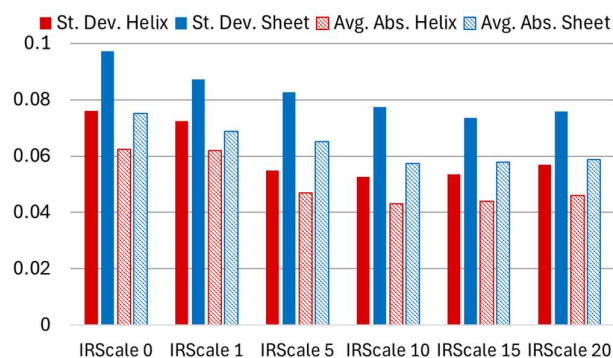

Figure 2: The standard deviations and the average absolute differences for helix and sheets for a range of scaling factors (IRscale) of the IR data in the combined CD and IR reference dataset.

In Figure 2, both the standard deviations and the average absolute differences for helix and sheets are shown for each scaling factor. We note that for a scaling factor of zero, i.e. pure CD analysis, both metrics show a better performance for the determination of helical content compared to the sheet content. In contrary, the metrics in Table 1 show that IR has a better performance in determining the sheet over the helical content. This is in line with the notion that CD is more sensitive to helical content, whereas IR is more sensitive to sheet content.

From Figure 2 it is clear that the performance of SELCON3 is improved when including IR data in the analysis (IRscale > 0), not only for sheet content, as expected due to the higher sensitivity to sheets, but also for the helix content. For an IRscale = 10 the metric $avg(\Delta_{abs,CD-IR})$ are 0.043/0.058 for helix/sheet to be compared to the pure IR SELCON3 fit values of 0.08 / 0.06, i.e. an improvement for helix without sacrificing the performance for sheets. From the analysis in Figure 2, taking both metrics into consideration, the best choice of IR scaling factor is 15, but 10 may also be considered a good choice.
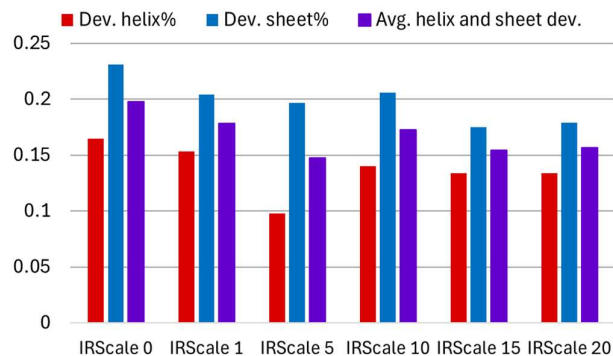

Figure 3: The maximum absolute difference between the SELCON3 calculated secondary structure and the crystal secondary structure, individually for helix and sheets, as well as their average.

To further analyse the performance of SELCON3 for the CD-IR28 data reference set, the maximum of the absolute difference between the SELCON3 results and the cSS, $max(|\Delta_{CD-IR,i}|)$, is shown in Figure 3. This metric is a measure of the worst performance of SELCON3, for both helix, sheet and their average, and using this metric a scaling factor of 5 minimizes the outliers.

In combination, these three metrics show that the best choice of IRscale is in the range 5-15. This range seems very reasonable when considering that an IRscale of 15 brings the IR spectrum magnitude close to that of the CD spectra for helix-rich proteins, and a scale of 5 brings the IR spectra close to the magnitude of low helix content proteins.

As the final metric considered to elucidate the optimum IRscale, we have calculated the root mean square deviation (RMSD) between the protein spectrum under analysis and the SELCON3 reconstructed spectrum. The average RMSD for all 28 proteins in CD-IR28 is shown in Figure 4 (top) for both the individual CD and IR parts and for the combined CD-IR spectrum.
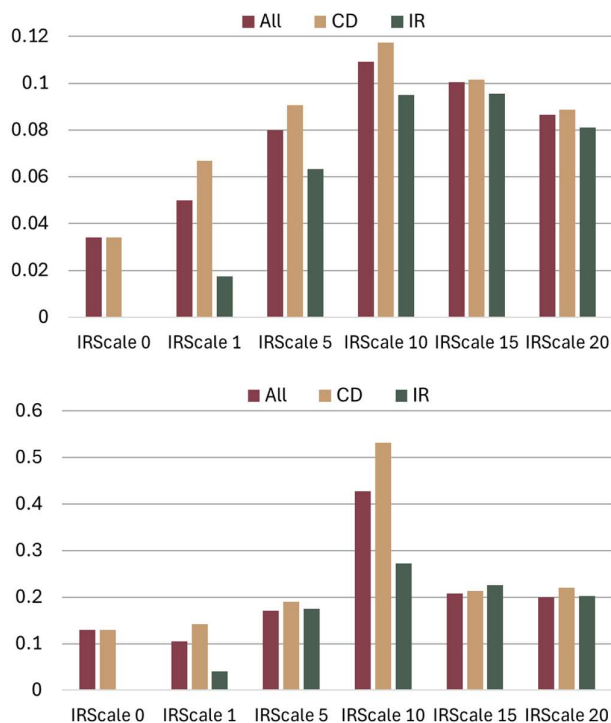
Figure 4: The average (top) and the maximum of the RMSD (bottom) between the protein spectrum under analysis in LOOV and the SELCON3 reconstructed spectrum. The RMSD is shown for both the individual CD and IR parts and for the combined CD-IR spectrum.

For scaling factors up to 10, the average RMSD for the IR part of the spectrum increases with increasing scaling factor. If we consider the simple case where the reconstructed spectrum has the same shape and is only scaled, then the RMSD would increase linearly with the scaling factor. Hence, the general increase is well understood for the IR RMSD. However, the increase in the RMSD of the CD part of the spectrum is not a direct result of the IR scaling factor. To understand this increase, we must consider how reference spectra are selected in the SELCON3 method. First the reference spectra are sorted according to their RMSD with respect to the protein query spectrum, and then an increased number of the reference spectra, most similar to the query spectrum, are included while searching for valid solutions.[36] When we concatenate the IR spectrum to the CD spectrum, other reference spectra might be more similar, i.e. have lower RMSD with respect to the query spectrum, than those for the CD spectrum only. This gives rise to reconstructed CD spectra that are no longer optimized for the CD part of the combined spectrum, but rather optimized for the both the CD and IR spectra. Therefore the overall RMSD increases with scaling factor, while still retaining a better prediction of the secondary structure as evidenced by the metrics $\text{avg}(\Delta_{abs,CD-IR})$ and $\sigma(\Delta_{CD-IR})$ in Figure 2.

To understand why the average RMSD is highest at a scaling factor of 10, the maximum RMSD among the proteins in CD-IR28 is shown in Figure 4 (bottom). The scaling factor of 10 is a clear outlier here, driven by a badly reconstructed CD spectrum for Chymotrypsinogen A. The combined CD and IR spectrum for Chymotrypsinogen A is shown in **S6** for scaling factors of 10 and 15. For IRscale = 10 the selected reference spectra in SELCON3 give rise to wavelength shifted peaks in the CD spectrum, resulting in a high RMSD, whereas the more dominating IR part of the spectrum at IRscale = 15 assists SELCON3 in selecting proteins that give results in a better fit between the CD part of the Chymotrypsinogen A spectrum and its reconstructed spectrum.

Overall, the analysis of all the considered metrics points to a scaling factor of 15 provides reliable results for the predictive power of SELCON3 using combinations of CD and IR reference spectra. The SSCalcPy software allows the user to select other scaling factors, and in particular for proteins with low helix content, i.e. with lower magnitude CD signal, we suggest that lower scaling factors are examined and compared to higher scaling factor predictions for consistency.

Conclusions

We have shown that the algorithm originally created by Sreerama and Woody [5] to extract secondary structure estimates from circular dichroism spectra can be used with amide I infrared protein absorbance data with slightly more average accuracy than any other method reported to date for analysis of IR spectra. The SELCON3 IR results are very similar to those we obtained previously using our self-organising map algorithm SOMSpec. Furthermore, the combination of CD and IR data was shown to give improved prediction accuracy in SELCON3 analysis compared to separate CD or IR analysis.

Author Contributions

All authors contributed equally to the formulation of the research goals and aims and different aspects of the data generation and analysis. SVH coded SELCON3 in Python.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Notes and references

1. Bruque MG, Rodger A, Hoffmann SV, et al. Analysis of the Structure of 14 Therapeutic Antibodies Using Circular Dichroism Spectroscopy. *Analytical Chemistry*. 2024/09/24 2024;96(38):15151-15159. doi:10.1021/acs.analchem.4c01882

2. Krokengen OC, Touma C, Mularski A, et al. The cytoplasmic tail of myelin protein zero induces morphological changes in lipid membranes. *Biochim Biophys Acta Biomembr*. Oct 2024;1866(7):184368. doi:10.1016/j.bbamem.2024.184368

3. Guerra JPL, Blanchet CE, Vieira BJC, et al. Controlled modulation of the dynamics of the Deinococcus grandis Dps N-terminal tails by divalent metals. *Protein Science*. 2023;32(2):e4567. doi:https://doi.org/10.1002/pro.4567

4. Hall V, Nash A, Hines E, Rodger A. Elucidating protein secondary structure with circular dichroism and a neural network. *J Comp Chem*. 2013;34:2774–2786. doi:10.1002/jcc.23456

5. 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. doi:doi:10.1006/abio.2000.4880

6. Whitmore L, Wallace BA. DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Res*. Jul 1 2004;32(Web Server issue):W668-73. doi:10.1093/nar/gkh371

7. Pinto Corujo M, Olamoyesan A, Tukova A, et al. SOMSpec as a General Purpose Validated Self-Organising Map Tool for Rapid Protein Secondary Structure Prediction From Infrared Absorbance Data. Original Research. *Frontiers in Chemistry*. 2022-January-27 2022;9doi:10.3389/fchem.2021.784625

8. Bandekar J. Amide modes and protein conformation. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*. 1992/04/08/ 1992;1120(2):123-143. doi:https://doi.org/10.1016/0167-4838(92)90261-B

9. Krimm S, Bandekar J. Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins. *Adv Protein Chem*. 1986;38:181-364.

10. Kauppinen JK, Moffatt DJ, Mantsch HH, Cameron DG. Fourier Self-Deconvolution: A Method for Resolving Intrinsically Overlapped Bands. *Applied Spectroscopy*. 1981/05/01 1981;35(3):271-276.

11. Maddams WF, Tooke PB. Quantitative Conformational Studies on Poly(vinyl Chloride). *Journal of Macromolecular Science: Part A - Chemistry*. 1982/04/01 1982;17(6):951-968. doi:10.1080/00222338208056495

12. Susi H, Byler MD. Protein structure by Fourier transform infrared spectroscopy: Second derivative spectra. *Biochemical and Biophysical Research Communications*. 1983/08/30/ 1983;115(1):391-397. doi:https://doi.org/10.1016/0006-291X(83)91016-1

13. Byler DM, Susi H. Examination of the secondary structure of proteins by deconvolved FTIR spectra. *Biopolymers*. Mar 1986;25(3):469-87. doi:10.1002/bip.360250307

14. Calero M, Gasset M. Fourier Transform Infrared and Circular Dichroism Spectroscopies for Amyloid Studies. In: Sigurdsson EM, ed. *Amyloid Proteins: Methods and Protocols*. Humana Press; 2005:129-151.

15. Kong J, Yu S. Fourier transform infrared spectroscopic analysis of protein secondary structures protein FTIR data analysis and band assignment. *Acta Biochimica et Biophysica Sinica*. 2007;39:549-559.

16. Yang H, Yang S, Kong J, Dong A, Yu S. Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. *Nature Protocols*. 02/05/online 2015;10:382. doi:10.1038/nprot.2015.024

17. Baumruk V, Pancoska P, Keiderling TA. Predictions of Secondary Structure using Statistical Analyses of Electronic and Vibrational Circular Dichroism and Fourier Transform Infrared Spectra of Proteins in H2O. *Journal of Molecular Biology*. 1996/06/21/ 1996;259(4):774-791. doi:https://doi.org/10.1006/jmbi.1996.0357

18. Pancoska P, Yasui SC, Keiderling TA. Statistical analyses of the vibrational circular dichroism of selected proteins and relationship to secondary structures. *Biochemistry*. 1991/05/01 1991;30(20):5089-5103. doi:10.1021/bi00234a036

19. Oberg KA, Ruysschaert J-M, Goormaghtigh E. The optimization of protein secondary structure determination with infrared and circular dichroism spectra. *European Journal of Biochemistry*. 2004;271:2937-2948.

20. Goormaghtigh E, Ruysschaert J-M, Raussens V. Evaluation of the Information Content in Infrared Spectra for Protein Secondary Structure Determination. *Biophysical Journal*. 2006/04/15/ 2006;90(8):2946-2957. doi:https://doi.org/10.1529/biophysj.105.072017

21. De Meutter J, Goormaghtigh E. FTIR Imaging of Protein Microarrays for High Throughput Secondary Structure Determination. *Analytical Chemistry*. 2021/03/02 2021;93(8):3733-3741. doi:10.1021/acs.analchem.0c03677

22. Hall V, Sklepari M, Rodger A. Protein secondary structure prediction from circular dichroism spectra using a self-organizing map with concentration correction. *Chirality*. 2014;26:471-482. doi:10.1002/chir.22338.

23. Ascoli GA, Pergami P, Luu KX, et al. Use of CD and FT-IR to determine the secondary structure of purified proteins in the low-microgram range. *Enantiomer*. 1998;3(4-5):371-81.

24. 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. doi:10.1093/bioinformatics/btl327

25. Hoffman S, Jones N. *Zendo* 2024;10.5281/zenodo.13995323

26. Hoffman S, Jones N. *GitHub*. 2024;github.com/AU-SRCD/SSCalcPy

27. Abdul-Gader A, Miles AJ, Wallace BA. A reference dataset for the analyses of membrane protein secondary structures and transmembrane residues using circular dichroism spectroscopy. *Bioinformatics*. 2011;27:1630-1636.

28. Whitmore L, Woollett B, Miles AJ, Klose D, Janes RW, Wallace BA. PCDDB: the protein circular dichroism data bank, a repository for circular dichroism spectral and metadata. *Nucleic acids research*. 2011;39:D480-D486.

29. Whitmore L, Miles AJ, Mavridis L, Janes RW, Wallace BA. PCDDB: new developments at the Protein Circular Dichroism Data Bank. *Nucleic Acids Research*. 2016;45(D1):D303-D307. doi:10.1093/nar/gkw796

30. Ramalli SG, Miles AJ, Janes RW, Wallace BA. The PCDDB (Protein Circular Dichroism Data Bank): A Bioinformatics Resource for Protein Characterisations and Methods Development. *Journal of Molecular Biology*. 2022/06/15/ 2022;434(11):167441. doi:https://doi.org/10.1016/j.jmb.2022.167441

31. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. . *Biopolymers*. 1983;22:2577–2637.

32. Oberg KA, Ruysschaert J-M, Goormaghtigh E. Rationally selected basis proteins: A new approach to selecting proteins for spectroscopic secondary structure analysis. *Protein Science*. 2003;12:2015-2031.

33. Lobley A, Whitmore L, Wallace BA. DICHROWEB: an interactive website for the analysis of protein secondary structure from circular dichroism spectra. *Bioinformatics*. 2002;18(1):211-212. doi:10.1093/bioinformatics/18.1.211

34. Whitmore L, Wallace BA. Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers*. May 2008;89(5):392-400. doi:10.1002/bip.20853

35. Miles AJ, Ramalli SG, Wallace BA. DichroWeb, a website for calculating protein secondary structure from circular dichroism spectroscopic data. *Protein Sci*. Jan 2022;31(1):37-46. doi:10.1002/pro.4153

36. Sreerama N, Woody RW. A self-consistent method for the analysis of protein secondary structure from Circular dichroism. *Analyt Biochem*. 1993;209:32-44.