Determination of protein fold class from Raman or Raman optical activity spectra using random forests

Myra Kinalwa, Ewan W. Blanch, and Andrew J. Doig*

Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess Street, Manchester M1 7DN, United Kingdom

Received 30 March 2011; Revised 8 June 2011; Accepted 23 June 2011

DOI: 10.1002/pro.695

Published online 15 July 2011 proteinscience.org

Abstract: Knowledge of the fold class of a protein is valuable because fold class gives an indication of protein function and evolution. Fold class can be accurately determined from a crystal structure or NMR structure, though these methods are expensive, time-consuming, and inapplicable to all proteins. In contrast, vibrational spectra [infra-red, Raman, or Raman optical activity (ROA)] are rapidly obtained for proteins under wide range of biological molecules under diverse experimental and physiological conditions. Here, we show that the fold class of a protein can be determined from Raman or ROA spectra by converting a spectrum into data of 10 cm^{-1} bin widths and applying the random forest machine learning algorithm. Spectral data from 605 and 1785 cm^{-1} were analyzed, as well as the amide I, II, and III regions in isolation and in combination. ROA amide II and III data gave the best performance, with 33 of 44 proteins assigned to one of the correct four top-level structural classification of proteins (SCOP) fold class (all α , all β , α and β , and disordered). The method also shows which spectral regions are most valuable in assigning fold class.

Keywords: protein structure; machine learning; vibrational spectroscopy; α -helix; β -sheet

Introduction

At the broadest level, protein structures can be classified into classes primarily based on secondary structure content. For example, the most abundant classes in SCOP are all α , all β , and α and β (α/β or $\alpha+\beta$),¹ whereas the main classification levels for CATH are mainly α , mainly β , mixed $\alpha-\beta$, and few secondary structures.² Knowledge of the fold class of a protein is valuable as fold class gives an indication of protein function and evolution.

The fold class of a protein can be definitively assigned by determining its three-dimensional structure by X-ray crystallography or NMR. However,

there are limitations to their applicability as many biomolecules are difficult to crystallize, while others are very large to be solved by current NMR methods. Both are always time consuming to implement. Alternatively, vibrational spectroscopies, such as infra-red, Raman, or Raman optical activity (ROA) can rapidly give data on protein structure. Although they do not usually provide information at atomic resolution, they can be routinely applied to a wide range of biological molecules under diverse experimental and physiological conditions providing information on structure and dynamics.^{3–5} Vibrational spectra of proteins are rich in structural information, with spectral features corresponding to secondary structure motifs, side chain tautomers, hydrogen bonding interactions, and other features. Protein secondary structure contents (typically, three-state: α , β , and coil) can be rapidly and accurately determined by circular dichroism, Raman or ROA,⁶ though this information is not the same as a class assignment. Here, we use the machine learning method of random forests (RF) to determine protein fold class from Raman and ROA spectra.

Abbreviations: ROA, Raman optical activity; RF, random forest; OOB, out-of-bag; MDS, multidimensional scaling.

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Andrew J. Doig, Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess Street, Manchester M1 7DN, United Kingdom. E-mail: andrew.doig@manchester.ac.uk

Figure 1. Metric MDS for whole Raman spectra. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Results

Whole spectra

RF dissimilarities were mapped to multidimensional scaling (MDS) using three different fitness functions: Sammon mapping,⁷ metric stress (sum of squares of the inter-point distances), and squared stress. For Raman spectra, the mean stresses were: 0.07 for Sammon mapping, 0.243 for metric stress, and 0.114 for squared stress; for ROA spectra, the mean stresses were: 0.12 for Sammon mapping, 0.313 for metric stress, and 0.169 for squared stress. Sammon mapping was, therefore, used for all our subsequent work, as it gives the closest mean fit for distances to dissimilarities for both Raman and ROA spectra. The correlations between the RF dissimilarities and the Euclidean distances after the mapping were 0.83 for the ROA data and 0.91 for Raman, showing that the mapping largely preserved the RF dissimilarity data (Supporting Information Fig. 1).

Figure 2. Metric MDS plot for whole ROA spectra. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table I. Percentage of Proteins Whose Classes Were Correctly Predicted

	Raman (%)	ROA (%)
Amide I	54 (13/24)	64 (28/44)
Amide II	8 (2/24)	64 (28/44)
Amide III	46 (11/24)	59 (26/44)
Amide I+II	50 (12/24)	73 (32/44)
Amide II+III	38 (9/24)	75 (33/44)
Amide I+III	50 (12/24)	61 (27/44)
Amide I+II+III	38 (9/24)	70 (31/44)
Whole	38 (9/24)	55 (22/44)

Numbers in parenthesis show the number of correctly predicted observations out of the total number of observations

The Raman and ROA spectra results are depicted in two-dimensional (2D) scaling plots (Fig. 1, Raman; Fig. 2, ROA). The data points are shown by their SCOP structural class: diamond α/β , circle α , triangle β , square disordered, or other protein structure. The Raman spectral data has α/β tightly clustered and α , β and disordered more dispersed. Hence, the algorithm has difficulty in distinguishing α from β or disordered proteins using whole spectrum Raman data. The ROA whole spectra plot, in contrast, shows that the α , β and α/β classes are roughly clustered, whereas the disordered proteins are dispersed. Disordered proteins are, thus, not readily distinguished from folded proteins using ROA data, though overall the ROA whole spectrum data are more useful in assigning fold class. Table I confirms that the RF for Raman has a poor performance for Raman data, though is better for ROA. Accuracies using whole spectrum data are 38% for Raman and 55% for ROA, compared to 25% for random assignment to one of four classes.

Variable analysis of the ROA and Raman data shown in Figure 3 and Figure 4 highlights the spectral bands that are important in identifying the fold classes. For Raman spectra, the most important bins are 905–925 cm^{-1} and 945–975 cm^{-1} and for ROA spectra 955–965 cm^{-1} , 1135–1145 cm^{-1} , 1165–1175 cm^{-1} , 1335–1355 cm^{-1} , 1685–1695 cm^{-1} . The strong correlation of the two ROA bins with protein fold is unsurprising as these two bins usually contain the most intense marker bands for α -helix.^{8,9} Therefore, our analysis verifies the band assignments previously used for the empirical analysis of protein structure using ROA spectra. However, the result for Raman spectra is unexpected as the body of protein Raman literature dating back decades has generally relied on analysis of the amide I region (1600–1700 cm^{-1}) with the amide II (1450–1500 cm^{-1}) and amide III (1200–1340 cm^{-1}) regions also being occasionally utilized. In fact, although there are known Raman marker bands in this vicinity, we are not aware of any previous research making use of the region of 965–975 cm^{-1} for protein structural analysis.

Figure 3. Raman whole variable importance (top) and Gini index (bottom) plots. The data were divided into bins of averages of 10 wavenumbers (cm^{-1}). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Spectral subdivisions

It is well established that bands within the amide I–III regions are sensitive to secondary structure content.¹⁰ We, therefore, investigated whether we could assign protein fold classes using just bands within these regions. The amide I–III regions were examined individually, and in combination, for both Raman and ROA (Fig. 5, Table I; Supporting Information Figs. 3 and 4). For Raman, the highest accuracy was for the amide I region alone, with only the amide II region having a performance worse than the whole spectrum. Similarly, using data from only amide regions was always better than using the whole spectra for ROA, with using amide II and III data in combination the most accurate. This suggests that fold class information is present largely in the amide I–III regions in vibrational spectra and that adding data from the rest of the spectra mostly adds noise, leading to a poorer algorithm performance. Even though Figures 2 and 3 show that there are bands outside the Amide regions (such as $965\text{--}975\text{ cm}^{-1}$) that have high variable importance, the value of adding them to the training data is outweighed by the addition of bins that contain little more than noise.

Variable importance and Gini plots show which bins are most important in fold class assignment for Raman amide region data (Supporting Information Fig. 5). Table II lists the most important bins and their corresponding positions in wavenumbers. Key regions for Raman fold prediction are $1255\text{--}1265\text{ cm}^{-1}$ and $1665\text{--}1685\text{ cm}^{-1}$. The first bin contains established marker bands for β -sheet, β -turn, and disordered structural motifs, while the other corresponds to the amide I region and contains marker bands for these motifs.^{10,11}

Variable importance and Gini plots for ROA amide data are shown in Supporting Information Figure 6 and summarized in Table II. The most important regions for ROA class assignment are $1335\text{--}1345\text{ cm}^{-1}$ and $1655\text{--}1665\text{ cm}^{-1}$. These two regions both contain strong marker bands for α -helix.⁹

Discussion

Our previous work used principal components analysis to separate ROA whole spectra.⁴ The first principal component separated spectra on helix versus sheet contents, while the second separated on folded versus disordered. While fold classes did fall into distinct regions using this method, assignments could only be made by eye from where a protein fell in the principal component analysis (PCA) plot, rather than definitively as from the random forest analysis. Second, our method shows which bins are of most value in class assignment, and this was not previously reported. Finally, we have used Raman spectra to assign fold class for the first time, though the results are noticeably poorer compared to ROA. To our knowledge, this is the first method that can assign protein fold class, other than crystallography or NMR.

The best performing method uses the ROA amide II and III regions. Adding amide I or whole spectrum data decreases accuracy, showing that any additional useful information in these regions is outweighed by noise. This is surprising, as the amide I region contains intense ROA bands that have proven useful for the analysis of protein structure.^{5,10,12} The dominant amide I ROA marker bands for α -helix, β -sheet, and disordered structures occur within a narrow range, of the order of 30 cm^{-1} , and it is possible that the optimum bin size used here, of

Table II. Most Important Spectral Regions for Assigning Fold Class

	Amide I	Amide II	Amide III	Amide I+II	Amide II+III	Amide I+III	Amide I+II+III	Whole
Raman (cm ⁻¹)	1665–1695	1525–1535	1255–1265	1665–1685	1235–1265	1665–1685	1665–1685	905–925, 945–975
ROA (cm ⁻¹)	1655–1665	1535–1545	1335–1345	1575–1585, 1655–1675	1335–1345	1335–1345, 1655–1665	1335–1345, 1655–1675	955–965, 1135–1145, 1165–1175, 1335–1355, 1685–1695

10 cm⁻¹, is very large to correctly discriminate between them. By comparison, the principal amide III ROA marker bands span a range of over 100 cm⁻¹, and so this bin size is still fine enough to discriminate between these. It is interesting to note that inclusion of the amide II region improves prediction accuracy as the ROA bands in this region are comparatively weak and so not often used for protein structure analysis. This result indicates that the amide II region of the ROA spectrum should receive greater attention. ROA is clearly superior to Raman spectroscopic analysis for distinguishing protein structure. ROA spectra contain clearer and more distinct marker bands for sheet, helix, loops, and turns than Raman spectra. These structures are obviously important to define the fold class of a protein. Although the accuracies for superfold prediction are not as high as our recently published predictions of secondary structure content from the same Raman and ROA spectra,⁶ they are still far above random chance (25%) for a four-class prediction.

In conclusion, ROA and Raman data can be used to accurately assign a protein into a fold class, rapidly giving potentially valuable information on protein structure and function. Although we have, so far, only investigated the ability of Raman and ROA spectra to identify the protein superfold class, these analyses should be equally applicable to recognizing the fold or even family class of a protein. A more comprehensive coverage of protein fold space will be required to achieve this, but the potential to determine the tertiary structure of a protein in solution from its vibrational spectra has been shown in this work. As these spectroscopic techniques are nondestructive and are relatively fast, there is clearly potential for these methods to complement atomic resolution structural techniques.

Materials and Methods

Random forests

A random forest (RF) is a combination of many decision trees where each tree depends on randomly selected vectors.^{13,14} The trees are grown using binary partitioning where each parent is split into two child nodes. Randomness is introduced by growing each tree on a random subset of the training dataset. After selection of the subset, the node is split using the best variable among the subset of variables. The best variable is that which produces children with the largest difference (Gini index). This process is repeated in its entirety, and at each node, a new subset is selected at random. As a result of this process, the most important variables will eventually be selected in the tree.

We used the RF implementation of Breiman and Cutler (http://www.stat.berkeley.edu/breiman/RandomForests/cc_home.htm).

Figure 4. ROA whole variable importance (top) and Gini index (bottom) plots. The data were divided into bins of averages of 10 wavenumbers (cm^{-1}). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Gini index

The criterion used in splitting the nodes in RF trees is the Gini index rule. The splitting decision is based on the decrease in the impurity at the node. The Gini value of a split at a node into subsets is defined as:

$$\Delta\text{imp}(\text{sp}, n) = \text{imp}(n) - P_L \text{imp}(n_L) - P_R \text{imp}(n_R)$$

where n is a given node; sp is a split in the node; P_R is the number of variables at the given node n that split into the right child node n_R ; P_L is the number of variables at the given node n that split into the left child node n_L ; $\text{imp}(n_R)$ is the impurity of the child node on the right n_R ; $\text{imp}(n_L)$ is the impurity of the child node on the left n_L .

The Gini impurity maximizes the average purity of the two children nodes. The selected splits are those that decrease the Gini index the most. The impurity value imp for the node n is: $\text{imp}(n) = 1 - \sum(p_1^2, p_2^2, \dots, p_n^2)$ where $p_1 \dots p_n$ are the frequency ratios of class 1... n in node n .

The split variable that generates the smallest Gini value $\Delta\text{imp}(\text{sp}, n)$ is chosen to split the node.

RF prediction

The trees are grown to their maximal size so that each terminal node contains a single vector. To combine the trees, the votes for each class are counted, and the class with the most votes is the overall winner. The number of votes is the RF score. The proportion of the votes is related to the probability of the class membership. RF allows setting of class weight to correct imbalances in the data and to boost the accuracy of the speci-

fied class. Class member scores were, therefore, weighted in inverse proportion to their class sizes, thus increasing the importance of smaller class members and avoiding the danger of the algorithm assigning every member to the largest class.

Out-of-bag data

Each tree is grown on two-third of the original data, randomly assigned, while the remaining one-third is used to test a single tree. The omitted data, referred to as out-of-bag (OOB), are used to assess the performance accuracy of the model at each tree. Model results are combined to give a single prediction through a voting system for classification problems. All training data is used in the training but only one subset at a time.

Figure 5. Metric MDS plot for ROA amide I and II spectra. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Variable importance

Classification trees assign each case to class only. Using the margin, the difference between the proportions of correctly predicted classes and the total number of cases predicted, the RF implementation allows the importance of each variable for the assignments to be made.¹⁵ Taking the left out OOB cases with known margins M_0 , the values of the x th variable on the k th tree are permuted and the new margin recalculated (M_1). The variable importance represents the lowering of the margin across all OOB cases. Mutating the variables decreases the proportion of true positives yielding a smaller margin of the permuted variable M_1 . The difference between the original margin and the permuted subset mutation represents the variable importance. Alternatively, variable importance can be determined using the Gini index, as above.

Distance metric

RF evolves trees to their entirety without pruning. The left out data are run down each tree. If observations i and j end up in the same terminal node the similarity between i and j is increased by one. The more similar data records are, the more likely they will land in the same terminal node of a tree. This distance metric can be used to construct a dissimilarity matrix for input into MDS. The dissimilarities matrix represents the pairwise relationships between the variables derived from the data features. Dissimilarity matrices reduce the high-dimensionality of large data, as the size of the matrices is directly proportional to the number of objects and is independent of the dimension of the data. MDS is a method used to analyze dissimilarity matrices to produce distances based on a specified number of dimensions. MDS analysis based on two dimensions, as used for the analysis presented here, produces a 2D representation of the data variables, with the RF distances represented in a 2D map. In MDS, the data variables are moved to find the configuration that best represents the observed distances. The MDS algorithm evaluating the different possible configurations uses a function minimization with the aim of maximizing the goodness-of-fit. The most common measure of goodness-of-fit of the approximation between the original dissimilarities and the interpoint distances produced by MDS is the stress measure, defined as:

$$\sqrt{\frac{\sum (p_{ij} - d_{ij})^2}{d_{ij}^2}}$$

where i and j are the original and mapped points, respectively, p_{ij} is the distance between i and j for the original dissimilarities and d_{ij} is the distance between i and j for the mapped points.

The axes on the MDS dimensional map are arbitrary meaning that the distances remain the same whichever way the map is looked at. MDS projects the data points into Euclidean space so that the Euclidean distances correlate approximately to the dissimilarity scores between the points. MDS chooses the lowest dimensional configuration possible to minimize the stress function. The MDS function was implemented with the MATLAB software (<http://www.mathworks.com/help/techdoc/>).

The Raman and ROA data were labeled classes 1 to 4: α -helix, β -sheet, α/β and disordered, respectively, according to the SCOP structural class classification. The aim of the RF was to differentiate the data into clusters of their respective classes. The data were pre-processed by binning, with widths of 10 cm^{-1} , starting from 605 cm^{-1} . Scaling was applied before RF to avoid attributes in greater numeric ranges dominating those in the smaller numeric ranges. For ROA spectra, which are bisignate, each attribute in the vector was linearly scaled to the $[-1, +1]$ range. Scaling to the range of $-1, +1$ means that the lowest value allowed in each vector is -1 , and the highest value allowed is $+1$. The rest of the values in the vector are adjusted to lie in between this range. The linear scaling method used is shown below.

$$z_i = \frac{(z_i - m_i)}{(M_i - m_i)}$$

where z_i is the attribute in the vector, M_i is the maximal value of the vector, m_i is the minimal value of the vector. ROA spectra have negative and positive peaks and therefore scaling between -1 and $+1$ reflects the nature of the data. Raman data are all positive, so, were scaled to be between 0 and $+1$. Data for both Raman and ROA spectra were analyzed between 605 and 1785 cm^{-1} .

Acknowledgment

The authors are grateful to Professor L.D. Barron and Dr L. Hecht at the Department of Chemistry of the University of Glasgow for provision of Raman and ROA spectra.

References

1. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536–540.
2. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH: a hierarchical classification of protein domain structures. *Structure* 5:1093–1108.
3. Ashton LA, Dusting J, Imomoh E, Balabani S, Blanch EW (2009) Shear induced unfolding of Lysozyme monitored in situ. *Biophys J* 96:4231–4236.
4. Blanch EW, McColl IH, Hecht L, Nielsen K, Barron LD (2004) Structural characterization of proteins and viruses using Raman Optical Activity. *Vibrational Spectroscopy* 35:87–92.

5. Barron LD, Hecht L, Blanch EW (2002) Unfolded proteins studied by Raman Optical Activity. *Adv Prot Chem* 62:51–90.
6. Kinalwa MN, Blanch EW, Doig AJ (2010) Accurate determination of protein secondary structure content from Raman and Raman Optical Activity spectra. *Anal Chem* 82:6347–6349.
7. Sammon JW (1969) A nonlinear mapping for data structure analysis. *IEEE Trans Comp* C18:401–409.
8. McColl IH, Blanch EW, Hecht L, Barron LD (2004) A study of α -helix hydration in polypeptides, proteins and viruses. *J Am Chem Soc* 126:8181–8188.
9. Blanch EW, Bell, A.F., Hecht, L., Day, L.A. and Barron, L.D. (1999) Raman Optical Activity of filamentous bacteriophages: Hydration of α -Helices. *J Mol Biol* 290:1–7.
10. Barron LD, Hecht L, Blanch EW, Bell AF (2000) Solution structure and dynamics of biomolecules from Raman optical activity. *Prog Biophys Mol Biol* 73:1–49.
11. Barron LD (2006) Structure and behaviour of biomolecules from Raman Optical Activity. *Curr Opin Struct Biol* 16:638–643.
12. Ashton LA, Czarnik-Matusiewicz B, Hecht L, Hyde J, Blanch EW (2006) Two-dimensional correlation analysis of Raman optical activity data on the α -helix-to β -sheet transition in poly(L-lysine) *Mol Phys* 104: 1429–1445.
13. Breiman L (2001) Random forests. *Machine learning* 45:5–32.
14. Shi T, Horvath S (2006) Unsupervised learning with random forest predictors. *J Comp Graph Stat* 15: 118–138.
15. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinf* 8: 25–46.