doi:10.1016/j.jmb.2006.08.038 J. Mol. Biol. (2006) 363, 19–26

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

COMMUNICATION

Delineation of Protein Structure Classes from Multivariate Analysis of Protein Raman Optical Activity Data

Fujiang Zhu[1], George E. Tranter[2], Neil W. Isaacs[1], Lutz Hecht[1] and Laurence D. Barron[1]*

[1]*WestCHEM, Department of Chemistry, University of Glasgow, Glasgow G12 8QQ UK*

[2]*Biological Chemistry, Division of Biomedical Sciences, Imperial College, London SW7 2AZ, UK*

**Corresponding author*

Vibrational Raman optical activity (ROA), measured as a small difference in the intensity of Raman scattering from chiral molecules in right and left-circularly polarized incident light, or as the intensity of a small circularly polarized component in the scattered light, is a powerful probe of the aqueous solution structure of proteins. On account of the large number of structure-sensitive bands in protein ROA spectra, multivariate analysis techniques such as non-linear mapping (NLM) are especially favourable for determining structural relationships between different proteins. Here NLM is used to map a dataset of 80 polypeptide, protein and virus ROA spectra, considered as points in a multidimensional space with axes representing the digitized wavenumbers, into readily visualizable two and three-dimensional spaces in which points close to or distant from each other, respectively, represent similar or dissimilar structures. Discrete clusters are observed which correspond to the seven structure classes all α, mainly α, αβ, mainly β, all β, mainly disordered/irregular and all disordered/irregular. The average standardised ROA spectra of the proteins falling within each structure class have distinct features characteristic of each class. A distinct cluster containing the wheat protein A-gliadin and the plant viruses potato virus X, narcissus mosaic virus, papaya mosaic virus and tobacco rattle virus, all of which appear in the mainly α cluster in the two-dimensional representation, becomes clearly separated in the direction of increasing disorder in the three-dimensional representation. This suggests that the corresponding five proteins, none of which to date has yielded high-resolution X-ray structures, consist mainly of α-helix and disordered structure with little or no β-sheet. This combination of structural elements may have functional significance, such as facilitating disorder-to-order transitions (and *vice versa*) and suppressing aggregation, in these proteins and also in sequences within other proteins. The use of ROA to identify proteins containing significant amounts of disordered structure will, *inter alia*, be valuable in structural genomics/proteomics since disordered regions often inhibit crystallization.

Keywords: raman optical activity; multivariate analysis; non-linear mapping; natively unfolded proteins; disordered structure

Abbreviations used: Agli, A-gliadin; NMV, narcissus mosaic virus; PMV, papaya mosic virus; PPII, poly(L-proline) II helix; PVX, potato virus X; ROA, Raman optical activity; TMV, tobacco mosaic virus; TRV, tobacco rattle virus; UVCD, ultraviolet circular dichroism; VCD, vibrational circular dichroism; NLM, non-linear mapping; SCP, scattered circular polarization; ICP, incident circular polarization.

E-mail address of the corresponding author: laurence@chem.gla.ac.uk

The determination of protein structure and function is a central aspect of biomolecular science in the post-genomic era. Although X-ray crystallography is the technique of choice in this enterprise, supplemented by multidimensional NMR for smaller structures in solution, these core techniques are not always applicable.[1,2] A major impediment to the application of X-ray crystallography is the lack of suitable crystals. This can be due to a number of factors, the most fundamental being that the protein lacks a compact tertiary fold in its native state. Such proteins are variously named "natively unfolded", "intrinsically unstructured" or "intrinsically disordered" and are now recognized as constituting an important structure class the members of which exhibit a variety of functions.[3–7] The DisProt database† provides an archive of information about structure and function of such proteins. There is an urgent need for the development of techniques that can provide structural information for the large number of proteins, be they folded, unfolded or partially unfolded, that are inaccessible to X-ray and NMR methods.

Spectroscopic techniques such as vibrational infrared[8] and Raman,[9] electronic circular dichroism measured in the ultraviolet (UVCD)[10] and extended into the vacuum ultraviolet using synchrotron radiation,[11] and vibrational circular dichroism (VCD),[12] have all been applied to this problem. Here we focus on the novel technique of Raman optical activity (ROA), which combines the additional sensitivity to three-dimensional structure of chiroptical methods such as UVCD with the advantages of vibrational spectroscopy and consequently reports on chirality associated with all $3N-6$ fundamental molecular vibrational transitions, where N is the number of atoms in the molecule.

ROA measures a small difference in the intensity of vibrational Raman scattering from chiral molecules in right and left-circularly polarized incident light or, equivalently, the intensity of a small circularly polarized component in the scattered light using incident light of fixed polarization.[13–16] The first and second experiments are called incident circular polarization (ICP) and scattered circular polarization (SCP) ROA, respectively. ROA has been developed to the point where it is an incisive probe of the structure and behaviour of biomolecules.[17] The ability to study aqueous solutions, with no restrictions on the size of the biomolecules, makes ROA ideal for the study of folded, unfolded and partially folded proteins,[18–22] together with intact glycoproteins[23,24] and viruses.[25]

We are currently building up a data set of the ROA spectra of polypeptides and proteins in aqueous solution that covers as much of fold space as possible. Some recent examples of typical protein ROA spectra may be found in the literature.[18–20] One advantage ROA has over other spectroscopies for protein structure analysis is that resolved signatures of loops and turns appear in addition to those of secondary structure, thereby also providing information about tertiary structure. One consequence is that, although some individual band assignments may be uncertain, or not even valid due to the extensive vibrational coupling often involved in the generation of large ROA signals, overall ROA band patterns can be characteristic of certain motifs and sometimes even folds.[20] ROA is also unique among spectroscopic methods in having the ability to distinguish between hydrated and unhydrated α-helix.[19]

In addition to folded proteins, ROA is also valuable in the study of unfolded proteins, especially for the identification of poly (L-proline II) (PPII) helix in such states. Although originally defined for the conformation adopted by polymers of L-proline, the PPII helix can be supported by amino acid sequences other than those based on L-proline and has been recognized as a common structural motif within the longer loops in the X-ray crystal structures of many proteins.[26] It consists of a left-handed helix with 3-fold rotational symmetry in which the ϕ,ψ angles of the constituent residues are restricted to values around −78°, +146°, corresponding to a region of the Ramachandran surface adjacent to the β-region. The extended nature of the PPII helix precludes intrachain hydrogen bonds, the structure being stabilized instead by main-chain hydrogen bonding with water molecules and side-chains.[27] The PPII helix currently attracts much interest as a major conformational element of disordered polypeptides and unfolded proteins in aqueous solution,[28,29] and may be important in the regulatory multiple weak interactions that are increasingly being recognized as associated with intrinsically disordered sequences within proteins.[6,7,30,31] It can be distinguished from random coil in polypeptides using UVCD,[28–31] VCD,[12,32] infrared and Raman[33] and ultraviolet resonance Raman,[34] but these techniques have difficulty in identifying PPII helix when other conformational elements are present, as in structured proteins. However, it is readily identified even in proteins using ROA,[35] which has proved valuable for studying PPII in unfolded and partially folded proteins[21,22] and its possible role in amyloid fibril formation in certain protein misfolding diseases.[21,22,36]

The large number of resolved structure-sensitive bands in protein ROA spectra makes them highly suitable for the application of multivariate analysis techniques to extract structural information. We have previously shown that useful structural relationships among proteins may be obtained by analyzing their ROA spectra using the well-known method of principal component analysis (PCA).[18,20] Here, we show that the application of the more advanced multivariate analysis method called nonlinear mapping (NLM)[37] yields even better results, including structural similarities within a subset of unfolded and partially folded proteins. Specifically, we used monotone non-metric multidimensional

† http://www.disprot.org