VIBRATIONAL SPECTROSCOPY TOOLBOX & APPLICATIONS: DETERMINATION OF THE SECONDARY STRUCTURE OF PROTEINS

Vladimír Kopecký Jr.[1, 3], *Jiří Bok*[1], *Kateřina Hofbauerová*[2, 3]

[1] Institute of Physics, Faculty of Mathematics and Physics, Charles University
[2] Department of Physical Chemistry and Macromolecular Chemistry, Faculty of Science, Charles University
[3] Institute of Physiology, Academy of Sciences of the Czech Republic

Abstract

The part of the Matlab toolbox of vibrational spectroscopy designated for a determination of the secondary structure of proteins is introduced. Functions and applications serve for analysis of Amide I and Amide III bands in Raman spectra and Amide I and II in infrared spectra of proteins. Mathematical procedures for the secondary structure determination are based on least squares analysis with reference spectra of proteins or with intensity profiles and some of them partially used factor analysis methods. Applications are designed for user-friendly analysis of a protein spectrum. They are based on Matlab functions mentioned above but they include automatic spectra treatments.

Introduction

More then half of dry weight of cells is made of proteins. Proteins serve as the main instruments of the molecular recognition, catalysis and also determine structure of the cell. The function of proteins is connected with their structure. Therefore if we would like to fully understand the enzymatic or a structure function of a protein, we will need to determine the protein structure. Probing the protein structure plays one of the most important roles in present molecular biology. Although very precise methods exist for determination of the whole molecular structure, *i.e.* X-ray or NMR structure determination, their practical usage is complicated. Crystallography of proteins is too much time consuming and needs excellently purified proteins in special physico-chemical conditions. NMR spectroscopy requires high concentrations of samples and proteins bigger than 30 kDa can be hardly determined in the structure. For that reasons other spectroscopic methods are used in simple way although they can not provide such complex information about the protein structure. (Basic overview on the spectroscopic methods for determining the protein structure can be found in [1].) One from the favourite is Raman and infrared (IR) spectroscopy – parts of the vibrational spectroscopy branch.

Vibration spectra of proteins exhibit important broad bands sensitive to the secondary structure conformations. These vibrations can be simply explained by a model of the peptide linkage (which links aminoacids in proteins) - N-methylacetamid (Fig. 1). In Raman spectra two bands can be used for the secondary structure diagnostics: Amide I (1630–1700 cm^{-1}) – C=O, C-N and C-N-H stretch vibrations and Amide III (1230–1330 cm^{-1}) – N-H/C-H deformation vibrations. IR spectra exhibit except Amide I band Amide II (1510–1570 cm^{-1}) – N-H/C-N deformational vibrations. The secondary structure, in terms α-helix, β-sheet, β-turn and other structure (see Fig. 2), can be determined by an analysis of Amide I–III bands in the spectra.

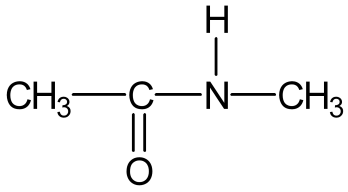

Figure 1. N-methylacetamid – the model compound used for a vibrational analysis of characteristics group frequencies for the peptide linkage.

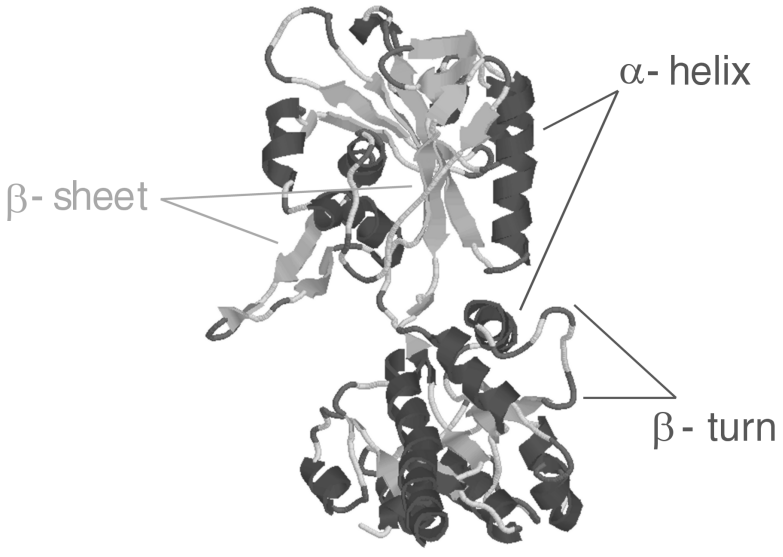

Figure 2. Main types of the secondary structures presented in proteins, *i.e.* helixes, sheets and turns.

Functions

Functions, mentioned here, are M-files with open Matlab® source code. They use matrices with reference data sets whose are in the MAT-format. The functions represent only part of the Vibrational Spectroscopy Toolbox & Applications project and not all mentioned functions are explained here. However brief overview of all used algorithms for a determination of the secondary structure content of proteins by means of Raman and IR spectroscopy is given in the next paragraphs.

`Amide1fast` – a very simple and fast method for the quantitative determination of the secondary structure contents of a protein by some parameters of Raman Amide I band. This method is based on statistical analysis, described in [2], between the structural data on one hand, obtained from X-ray crystallography, and the spectroscopic parameters of Raman Amide I band on the other. The basic principle used in this method lies in the expression of the percentage of a specific conformation of a given protein as a liner function of spectroscopic parameters, which reads $S\ [\%] = a_0 + a_1P_1 + a_2P_2 + \ldots + a_mP_m$ where S is type of the structure (α-helix, β-sheet, β-turn, unordered structure), a are a coefficients and P are the chosen parameters of Raman Amide I band (*e.g.* maximum of the band, left and right halfwidths). This method gives good results (better than ± 8 % in a structure content) from the knowledge of the frequency of the maximum of the Raman Amid I peak, however the core lies in statistics of the set of proteins not used furthermore in this method. Thus it can serve for basic and approximate determination of the secondary structure content only.

`Amide1lstsq` and `Amide3lstsq` – provides a least-squares analysis of Raman Amide I and Amide III bands, which is a standard method for determination of the secondary structure of proteins from the Amid bands. This method use a simple principle – it solve equation $Ax = b$, where A is matrix containing normalized reference spectra of proteins whose structures are known, column b contains the normalized spectrum of the protein being analyzed. The solution leads to equation $f = Fx$, where F is a matrix containing classes of the secondary structure of reference proteins, then f contains fractions of the secondary structure for the protein being analyzed. The functions were programmed according to [3]. Raman spectra mentioned in [3] were interpolated at resolution 1 cm^{-1} by a cubic spline interpolation (for Amide I analysis only). The spectra and secondary structure matrix (according to [3]) are stored in MAT-files `amide1lstsq` and `amide3lstsq` respectively. The arrangement of matrices is same as in [3], thus it gives chance to simply modify the set of reference proteins if necessary. Both of these functions gives a determination of the secondary structure content better than ± 5 % and the algorithm is mathematically simple and physically coherent.

`Amide1rip` – the most popular method in analysis of the Raman Amide I band. It is based on the same mathematical and software background as previous, *i.e.* least-squares analysis. In comparison with two functions mentioned above, it used reference intensity profiles (Fig. 3) determined by authors of this method [4]. The method enables to add to algorithm a profile of a solution or to divide a profile for α-helix to several separate bands whose improves the analysis. Dependent on the used profiles, this method is called 3 or 4-Reference Intensity Profiles Method [4]. Usage of additional profiles improves an estimation of the secondary structure content to be better than ± 4 %.

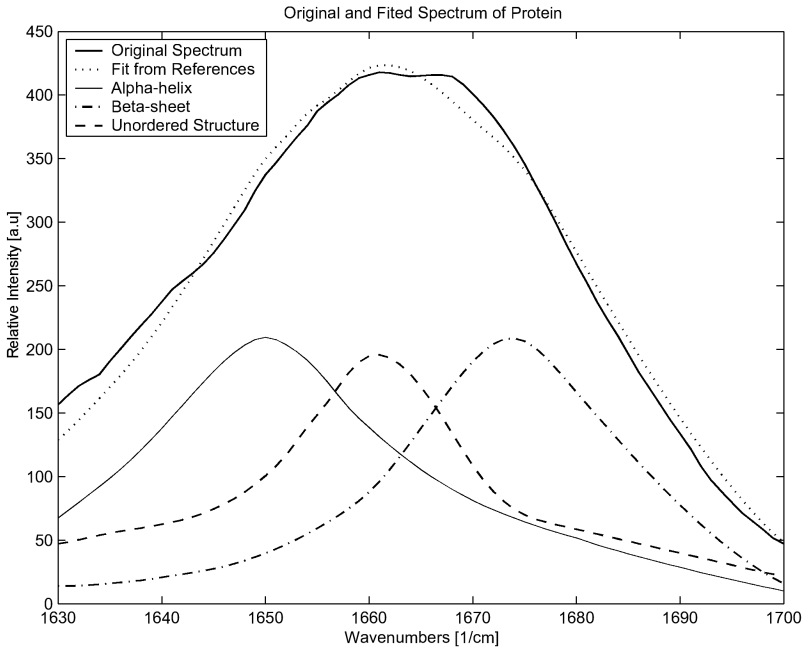

Figure 3. The result from the *3-Reference Intensity Profiles Method* of the Raman Amide I band analysis using *Amide* application. The original spectrum of a protein was fitted by reference profiles for each type of the secondary structure, *i.e.* α-helix, β-sheet and unordered structure. Thus relative representation of each of spectral profiles correspond to a relative amount of the secondary structure in the protein.

`Amide2lstsq` – The same mathematical principles as in Raman spectroscopy are used in an analysis of the IR Amide I and II bands region. However functions for IR spectra are our own construction. They use a protein spectra set taken from [5], mathematical methods were taken elsewhere. The least-squares analysis for IR spectra of proteins was reported later than for Raman spectra because of complicated data treatment due to a high contribution of the water band to the region of Amide I and II in IR spectra [6], nevertheless the mathematical algorithm is same as reported for Raman spectroscopy [3, 4].

`Amide2pls` – to simplify operations with matrices the Partial Least-Squares Analysis was developed [6]. This method use factor analysis to decompose calibration set A of IR protein spectra into $A = TB + E$, where B is an matrix which rows are the loading spectra, T is an matrix of intensities or scores in the new co-ordinate system and E represents spectral residuals. Because columns in matrix T are orthogonal, the simplest matrix operations can be used in a comparison with a classical least-square solution. This mathematical procedure can increase accuracy in the secondary structure estimation about 1 %.

`Amide2fa` – uses factor analysis methods for IR spectra according to [7]. The factor analysis using singular value decomposition algorithm is used to generate abstract factors of eigenspectra, which may be combined linearly to enable the original spectra to be reconstituted within experimental error. Factors that account only for noise in the original spectrum are discarded and the remaining factor loadings are analyzed for their contribution to each of the spectra forming the calibration set. Then a multiple linear regression is used to establish correlations between the factor loadings and the composition of the calibrations samples in terms of the properties of interest, those with little correlations are eliminated. When an unknown is analyzed, the factor loadings required to reproduce the spectrum are determined by using factors generated from the calibration set. The value for each property in the unknown is then obtained by substituting in the relevant regression equation. The error of this method in the secondary structure estimation is approximately same as in the previous method.

Applications

An application superstructure, called `amide` was programmed. This applications use all functions mentioned above. Thus it enables by modifications of the functions change spectral sets or algorithms. These changes are then used in the applications but graphical user interface (GUI) stay same. GUI plays the most important role in applications. Philosophy of the Vibrational Spectroscopy Toolbox & Applications project is to enable professional analysis of spectra without special knowledge of programming. Applications give users sophisticated professional graphical output but mathematical part of analysis still modifiable by usage of the simple Matlab® mathematical expressions. Application `amide` was designed only for the secondary structure estimation in protein spectra, thus it enables read data by usage of function `loadspec` in different data formats (ASCII, SPC, SPE *etc.*) used in spectroscopy. Results of Amide band analysis can be write to files in ASCII format. Moreover this application automatically treats data according to relevant procedures [2–7]. It tests presence of amide bands, fits data to appropriate density using the cubic spline function and then norms the tested spectrum for the correct analysis. However the full assistance for data treatment in `amide` can not be done. For this purposes exist a wide range application `asa` – the *Advance Spectra Analyzer*. This applications includes all functions from our toolbox and

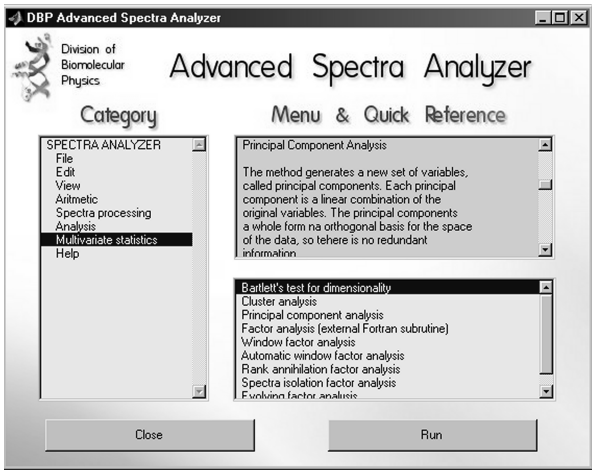

Figure 4. The screen of the *Advanced Spectra Analyzer* – the application for vibrational spectroscopy, which enables wide range of methods for the spectra analysis.

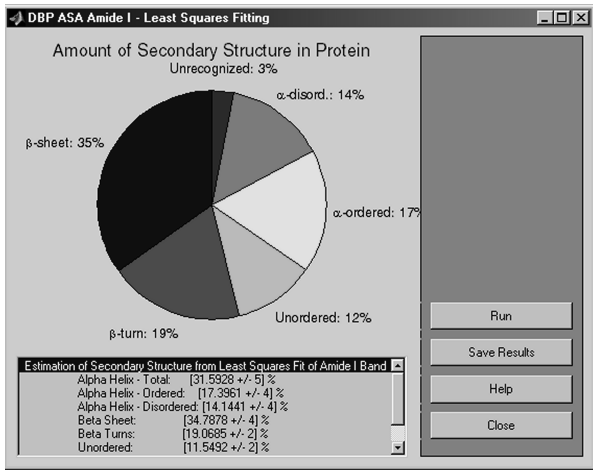

Figure 5. The screen of the *Least Squares Analysis* of the Raman Amide I band – the part of the *Advanced Spectra Analyzer*.

methods from other Matlab® toolboxes useful in a data treatment and analysis. It enables full manipulation with data including corrections of background, multivariate statistics, viewing *etc*. All functions share the same data sets which simplify operations with spectra.

Conclusion

Determination of the secondary structure of proteins in solutions plays an important role in molecular biology. The basic per cent content of α-helix, β-sheet, β-turn and undetermined structure can be simply done by analysis of Amide bands in Raman or IR spectra. Our Vibrational Spectroscopy Toolbox & Applications includes functions for analysis of Amide I, III and Amide II bands in Raman and IR spectra respectively. Connection of the open Matlab® code in functions, with graphical user interface in applications, enables advanced users to manage both reference sets and algorithms, which can be useful to improve the analysis. Primary users appreciate an automatic spectra treatment and a graphical intuitive interface in the applications. In the present time no freeware software for determination of the secondary structure from Raman or IR spectra can be found on web pages. Some commercial subroutines for program packages are available for IR spectroscopy, no one was found for Raman spectra. Thus our Toolbox & Applications represent the widest range of the methods for the secondary structure estimation from Raman and IR spectra freely available on Internet in the present time.

Acknowledgement

The authors are very grateful to *Dr. V. Baumruk* (Charles University, Czech Republic) for his fruitful suggestions and advises. The support of this study by the research plan of the Ministry of Education (MŠMT VS 97 113) and of the Faculty of Science (MSM 113100001), by the Volkswagen Foundation (I/74 679) and by grants No. 204/01/0254, No. 204/01/1001 of the GACR, is acknowledged.

References

[1] A. H. Havel (Ed.): Spectroscopic Methods for Determining Protein Structure in Solution. *VCH Publishers*, New York – Weinheim – Cambridge (1996).

[2] A. J. P. Alix, G. Pedanou and M. Berijot: Fast Determination of the Quantitative Secndary Structure of Proteins by Using Some Parameters of the Raman Amide I Band. *Journal of Molecular Structure* **174** (1988), 156–164.

[3] R. W. Williams: Protein Secondary Structure Analysis Using Raman Amide I and Amide III Spectra. *Methods in Enzymology* **130** (1986), 311–331.

[4] M. Berjot, J. Marx and A. J. P. Alix: Determination of the Secondary Structure of Proteins from the Raman Amide I Band: The Reference Intensity Profiles Method. *Journal of Raman Spectroscopy* **18** (1987), 289–300.

[5] V. Baumruk, P. Pancoska and T. A. Keiderling: Predictions of Secondary Structure using Statistical Analyses of Electronic and Vibrational Circular Dichroism and Fourier Transform Infrared Spectra of Proteins in H_2O. *Journal of Molecular Biology* **259** (1996), 774–791.

[6] F. Dousseau and M. Pézolet: Determination of the Secondary Structure Content of Proteins in Aqueous Solutions from Their Amide I and Amide II Infrared Band. Comparison between Classical and Partial Least-Squares Methods. *Biochemistry* **29** (1990), 8771–8779.

[7] D. C. Lee, P. I. Haris, D. Chapman and R. C. Mitchell: Determination of Protein Secondary Structure Using Factor Analysis of Infrared Spectra. *Biochemistry* **29** (1990), 9185–9193.

Contact Addresses

Vladimír Kopecký Jr. ✉, Jiří Bok
Institute of Physics, Faculty of Mathematics and Physics, Charles University, Ke Karlovu 5, CZ-121 16 Prague 2, Czech Republic,
Tel.: +420/2/2191 1472, Fax: +420/2/2492 2797, E-mail: kopecky@karlov.mff.cuni.cz, URL: http://atrey.karlin.mff.cuni.cz/~ofb

Kateřina Hofbauerová
Department of Physical Chemistry and Macromolecular Chemistry, Faculty of Science, Charles University, Albertov 2030, CZ-128 40 Prague 2, Czech Republic

Kateřina Hofbauerová, Vladimír Kopecký Jr.
Institute of Physiology, Academy of Sciences of the Czech Republic, Vídeňská 1083, CZ-142 20 Prague 4, Czech Republic