International Journal of Computational and Theoretical Chemistry
2024, Vol. 12, No. 1, pp. 18-23
https://doi.org/10.11648/j.ijctc.20241201.13

Methodology Article

Automation of the *Ad Hoc* Approach for Derandomization of Proteins: A Tutorial for Undergraduates in Molecular Sciences

Adewale Olamoyesan[1, 2, 3, *]

[1]Department of Molecular Sciences, Macquarie University, Sydney, Australia
[2]Department of Chemistry, University of Lagos, Lagos, Nigeria
[3]College of Science and Computing, Wigwe University, Isiokpo, Nigeria

Abstract

Data analysis and manipulation software are vulnerable to user error during data processing and computations take considerable time when handling huge data and multiple repetitive tasks. These problems are usually mitigated by creating an app to repeat any given task reproducibly any number of times. This paper discusses the development of app that systematically automates the *ad hoc* approach for derandomization of proteins and, or peptides. Thirty second-year undergraduates with little-to-no prior knowledge of computer programming are (were) asked to create this app with modules that sequentially convert spectra from original units to molar extinction and subtract baseline spectrum from the resultant spectra, derandomize the spectra by removing suspected significant unfolded domains from them, concatenate the generated files to a single file in an acceptable format for structural analysis, process our group structural algorithm output files into a user-friendly format to ease data analysis. In addition, they are (were) asked to prepare protein solution, determine its concentration spectroscopically, collect circular dichroism measurements of the protein, derandomize the protein spectra, and determine the secondary structure of the resultant protein spectra with our structure algorithm. The assessment results demonstrated that the students could prepare samples for CD analysis, collect spectra of proteins, and create an app to automate the *ad hoc* approach. The hands-on activities enable students to acquire knowledge in basic programming and circular dichroism, CD spectroscopy.

Keywords

MATLAB, Application, Second-Year Undergraduate, Circular Dichroism, Protein

1. Introduction

Software and apps such as Microsoft Excel, Origin, and Sigmaplot are frequently used to import, view, analyze, and interpret results obtained from analytical instruments in laboratories. These tools can visualize and present raw experimental data in different forms, calculate answers, and simulate processes. Their drawbacks are that they are very vulnerable to user error during

[*]Corresponding author: adewale.olamoyesan@hdr.mq.edu.au (Adewale Olamoyesan), adewale.olamoyesan@unilag.edu.ng (Adewale Olamoyesan)

Received: 24 May 2024; **Accepted:** 15 June 2024; **Published:** 27 June 2024

data processing and computations take considerable time, particularly when multiple repetitive tasks are required and huge data are being processed. A more reliable and efficient alternative for these tasks is to program an application to repeat any given task reproducibly any number of times. Examples of programming languages available to actualize this automation or simulation are MATLAB, Python, and C++. MATLAB [1-4] and Python are both scripting languages, the former has a limited range of technical applications compared with the latter, while C++ [5, 6] is a full-featured programming language with a range of applications and numerical precision that surpasses that of the aforementioned two languages. All the software has obtained wide applications in the domain of science and engineering, which prove the general acceptability of any of them without paying much emphasis on each language's strengths. In this study, the preference for MATLAB was based on its advantages, the author's strong familiarity with this particular software and its availability in the Information Technology (IT) services of the university for all students and staff. The advantages of which MATLAB was chosen for manipulation of CD spectra are (i) ease of learning and understanding of the code, (ii) its capability of developing an app or software via a platform (known as appdesigner), *etc* [1].

Some reports have demonstrated the use of MATLAB in teaching science concepts, as this exposes students to coding and helps reinforce relevant concepts [7-11]. Francisco and co-workers designed a nuclear magnetic resonance, NMR viewer by using MATLAB graphical user interface, GUI for visualizing 1H NMR spectra in a user-friendly way, this project provides the students with knowledge of computational programming and spectroscopy [7]. A MATLAB program for the simulation of dynamic infrared, IR spectroscopy was written by Zoerb and Harris [8] based on the optical Bloch equations, as these models successfully predicted the effects of known exchange rate on the IR spectrum and how the fast site exchange affects the input vibrational line shapes. Adian Fisher [9] reported an app (generated with MATLAB GUI) for the calculation of binding energies of cadmium, Cd, and selenium, Se, by using Slater's rule to predict their X-ray photoelectron spectroscopy, XPS spectra. Simulated spectrum outputs are quite similar to experimental spectra, as they are derived from uniting mathematical model and experimentally derived values. Of relevance to this work, Vincent *et al* [10] developed a self-organizing map, structure fitting methodology SSNN (secondary structure neural network) which worked in MATLAB. SSNN predicts the secondary structure of an unknown protein based on reference spectra from proteins with known X-ray structures. Seeing the potential of this approach for also analyzing infrared and Raman absorbance data, Ang and co-workers revised the SSSN to SOMSpec (Self-organizing map spectroscopy), by generalizing this app for predicting the secondary structures for three spectroscopic methods [11]. This program furnishes researchers and spectroscopists with a tool that enables them to estimate the secondary structure of proteins and, or peptides that might have or have not been derandomized. The proteins and, or peptides to which the derandomization approach was applied are typical biopharmaceutical products or drugs [12-15]. For example, insulin is used to manage diabetics in patients [16, 17].

MATLAB GUI proved to be the ideal tool to systematically automate the previous *ad hoc* approach for derandomizing spectra of proteins and, or peptides in solution, with which the effect of variables such as temperature and time of measurement of spectra after the addition of buffer can be identified. The proteins or peptides resulting from this process are said to have significant unfolded domains, which may have an impact on determining the right answer from the predicted spectra for each condition, *e.g*., lysozyme at 30 ℃, U3.4 R7A immediately after buffer addition at t = 0 h [14, 15]. Toward that goal, a teaching project on this was designed to provide second-year undergraduates of the Department of Molecular Sciences majoring in chemistry with hands-on experience in programming while solving problems in chemistry.

This paper reports the development of a series of modules built with MATLAB GUI (available as CD derand) that can derandomize protein spectra and eliminate bottlenecks in data analysis of SOMSpec output files. Each of these modules can either generate visualizable plots of the resultant data or write them into an Excel file (xlsx format) after the original data has been manipulated. The plots of the resultant spectra for the first module (delta epsilon) of the app are displayed in the figure window like other plots for scaled spectra and derandomized spectra, while distinct outputs (data files) in either Excel or txt format that can be used for data analysis and, or as input file for the self-organizing map, SOM analysis is generated for all the modules in the app.

2. Objectives

Undergraduates studying chemistry in the department have option of taking scientific computing course towards understanding programming and later applying the knowledge to solve problems in chemistry. For the structure and overview of the course, and longitudinal survey data (see S1 and S2 in supporting information). The teaching project described in this text is structured to achieve three objectives:

1) Develop the thinking abilities of students to solve computation problems in chemistry or relevant subjects.
2) Improve student attitudes on using computers to solve chemistry problems.
3) Reinforce chemistry and relevant artificial intelligence concepts taught during lectures through hands-on activities in the laboratory.

3. Implementation

3.1. Lectures

The author prepared teacher and lecture notes to introduce

the instructors to chemistry-related concepts and relevant artificial intelligence, AI concepts or tools like MATLAB and subsequently the students (see lecture and teacher notes). Instructors are to walk the students through chemistry-related and artificial intelligence concepts and, or tools, class problems *e.g.* creating of IR train/test app used to generate test and train files for alternate validation approach. Thirty second-year undergraduates batched together in fifteens that attended the lectures were grouped in threes for the computer-based and bench laboratories to follow scheduled for ten days in succession.

3.2. Development of CD Derand App

3.2.1. Module 1: Delta Epsilon

The task is to build a "delta epsilon" (alternatively known as molar extinction) module used for converting CD spectra of proteins and, or peptides in millidegrees to molar extinction (most generally used units by spectroscopists and acceptable by all structure analysis programs such as our SOMSpec and Dichroweb). First, students were asked to write the formulae for converting from millidegrees to molar extinction units …to molar ellipticity… to absorbance, and vice versa. After that, students were instructed to write code in the components' callbacks using the conversion formula and any MATLAB functions and operators found suitable for the purpose. These programming statements are to execute a series of events like accepting input parameters (path length (mm), concentration (mg/mol) and mean residue weight (Daltons), converting to units of interest (molar extinction) from millidegrees (original units), subtracting the baseline from the resultant spectra, plotting the baseline corrected spectra, and writing the processed results into Excel.

3.2.2. Module 2: Scaling Factor

This module is used to scale CD spectra of proteins or peptides to account for spectroscopists' concentration error and to convert certain spectra to molar extinction (*e.g.*, CD spectra of insulin with scaling factor, 2.69). The students were instructed to create an interface for this task and write lines of code in the components' callbacks to use the scaling factor, MATLAB functions and operators found suitable for executing the computations (somewhat similar to that in the delta epsilon module).

3.2.3. Module 3: Disordered

Students were instructed to create a module that can remove varying fractions of unfolded domains from baseline -corrected spectra (in molar extinction units). This module is used to automate the *ad hoc* process of derandomizing spectra of proteins or peptides. The requirement in terms of the mathematical model (can be found in Equation 6, in the lecture note) to implement this process, while they decide on their own the MATLAB functions and operators suitable to write programming statements in the components' callbacks of the interface.

3.2.4. Module 4: Concatenate

The concatenate module is used to merge derandomized spectra into a single file. This enables SOMSpec to analyse the data as a whole chunk saved in an acceptable format (txt). The instructor(s) asked students to create an interface with components and write lines of code in the components' callback to automate the computations.

3.2.5. Module 5: Convert

In addition to the other modules developed to prepare input data for SOMSpec, students were asked to create a module to extract SOMSpec output (txt) into a user-friendly format (xls) to ease data analysis. The experimental and predicted spectra written by SOMSpec as row vectors without including the wavelength data, are challenging to analyse in this particular form. Instructors walked students through the computational steps to process the data to ease data analysis and hinted to them about a few MATLAB functions and, or operators to use. Then, students follow the same approach as in other modules to create the required module (named convert) to produce xls files. The five modules were made available as a "CD derand" app created in a tab container as advised by the instructors.

3.3. Bench Laboratory

To access the student's ability to reproduce experiments in the laboratory with minimum supervision and little or no variability in the results benchmarked against the author or instructors' results. They were instructed to prepare buffer, protein solutions, estimate the concentration of the protein solutions using a UV spectrometer, and collect CD spectra for the protein solutions. For step-by-step procedure for the bench laboratory see S1 in lecture and teacher notes.

4. Assessment

For instance, CD measurements made on lysozyme by students every 10 ℃ from 20 ℃ to 100 ℃ were checked for their reliability by instructors, by inspecting the total absorbance and the ratio of the peaks' intensity at 208 to 222 nm. Each spectrum with total absorbance between -0.2 to 1.1 and a 208 nm to 222 nm peak intensity ratio of 1.25 ±0.05 scored 2.5 marks, while the spectrum whose total absorbance and, or the intensity ratio of the peaks deviated greatly from the aforementioned range was scored 0.5 marks or 1.5 marks. A perfect score received for all spectra will add up to a total score of 25 marks for all. The modules that made up a "CD derand" app were graded on the use of appropriate components for the interfaces and execution of computation steps and results obtained. A module with fewer than required components that were not aligned in logical order and failed to run was scored

3 marks out of 10 marks. Thus, an app that satisfies all the criteria will receive a perfect score (50 marks). In the case of SOMSpec outputs (spectral and structure predictions), the reliability of spectra collected with a spectrophotometer and the accuracy of the computations by the "CD derand" app were used to extrapolate the scores the students received (the perfect score that can be awarded for the results here is 25 marks). The reason is that when the processes are repeated with the same data set, SOMSpec prediction outcome and normalized root mean square deviation, NRMSD seem not to vary more than 1% and 10% respectively [13, 18]. It was found that nearly half of the students scored below 50 marks for all the activities, however, a significant percentage scored about 18 marks for the bench laboratory. This suggests the students are adept at preparing samples for CD analysis and collecting results and develop good understanding of computer programming. Noteworthy, both bench and computer-based laboratories are aimed to reinforce the student's learning and provide hands-on training to understand concepts taught during the lectures.

5. Hazard and Safety

5.1. Ultra Violet (UV) Radiation and Precautions

5.1.1. Ultra Violet (UV) Radiation

The light source (UV lamp) of both CD spectrophotometers and UV-vis spectrometer emits UV radiation that can cause serious effects to the eyes and other hazards. This depends on the intensity, duration and wavelength of the radiation.

Figure 1. *UV radiation symbol*

5.1.2. Precautions

1) Wear safety goggle or glasses approved to protect from UV radiation.
2) Avoid looking at the UV lamp when turned on.
3) Ensure the spectrophotometer or spectrometer lid is closed when collecting spectra.

5.2. High Pressure and Precautions

5.2.1. High Pressure

The high pressure built up in the hot lamps can cause it to explode when it is being replaced or removed while hot.

5.2.2. Precautions

1) Ensure the lamp cools down completely before being removed or replaced from its holder, to protect eyes and skin from flying shards.
2) Avoid touching the lamp with bare skin, as oil from it on a hot lamp will etch the quartz causing local overheating and strains on the lamp that may lead to premature, catastrophic failure.
3) Wipe the lamp with alcohol before installing it.

5.3. Ozone and Precautions

5.3.1. Ozone

Allotrope of oxygen (ozone), which is more reactive than its diatomic counterpart is produced by the absorption of UV light by oxygen. Ozone can damage the optics of the spectrophotometer.

5.3.2. Precautions

1) Make sure the space where this equipment is adequately ventilated to ensure that the build-up of nitrogen does not lead to asphyxiation.
2) Purge instrument with nitrogen for 15 min before turning on the lamp usage.

6. Conclusions

A teaching project to help second-year undergraduates boost their understanding and confidence in basic programming and chemistry related concepts, and how to automate the *ad hoc* approach of derandomizing CD spectra of proteins was designed and implemented. A series of modules have been built by the students to prepare input CD spectra for SOMSpec analysis and to extract the SOMSpec output into a user-friendly format. This study provides support for the conclusion reached by Erik [19] that solving problems in chemistry using computers enhances students' engagement and learning. An important question for future studies is whether to streamline this suite of modules into a single unit.

Abbreviations

CD	Circular Dichroism
GUI	Graphical User Interface
NMR	Nuclear Magnetic Resonance
MATLAB	Matrix Laboratory

RC	Random Coil
IR	Infrared
SOM	Self Organizing Map
SOMSpec	Self Organizing Map Spectroscopy
SS	Secondary Structure
SSNN	Secondary Structure Neural Network
UI	User Interface
UV	Ultra Violet
XPS	X-ray Photoelectron Spectroscopy

Supplementary Material

The supplementary material can be accessed at https://doi.org/10.11648/j.ijctc.20241201.13

Acknowledgments

I immensely appreciate the invaluable support and guidance provided by the project manager (Alison Rodger) and those researchers who had implemented the *ad hoc* approach for derandomization of protein spectra via Excel, the instructors involved and the students who took part in the teaching project.

Author Contributions

Adewale Olamoyesan is the sole author. The author read and approved the final manuscript.

Funding

The support from the international Macquarie University Research Excellence Scholarship (iMQRES) is gracefully acknowledged.

Data Availability Statement

The data supporting the outcome of this research work has been reported in this manuscript.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] MATLAB. Available from: https://www.mathworks.com/products/matlab.html (accessed January 2024).

[2] Van Loan, C. F. Introduction to Scientific Computing, 3. ed; London: Pearson Education, Limited; 2005, pp 17-50.

[3] Chapman, S. J. MATLAB Programming for Engineers. Stamford: Thomson; 2004, pp 2-8.

[4] Moler, C. B. Numerical Computing with MATLAB. Philadelphia: Siam; 2004, pp 1-55. https://doi.org/10.1137/1.9780898717952

[5] Python vs. C++: Key differences and uses. Available from: https://www.ionos.com/digitalguide/websites/web-development/python-vsc/#:~:text=C%2B%2B%20duel%20lacks%20a%20clear,requires%20more%20time%20to%20master (accessed January 2024).

[6] Python vs C++: Which One Should You Use? Available from: https://blog.udemy.com/python-vs-c-plus-plus/ (accessed January 2024).

[7] Arrabal-Campos, F. M, Cortés-Villena, A., Fernández, I. Building "My First NMRviewer": A Project Incorporating Coding and Programming Tasks in the Undergraduate Chemistry Curricula. Journal of Chemical Education. 2017, 94(9), 1372-1376. https://doi.org/10.1021/acs.jchemed.7b00304

[8] Zoerb, M. C., Harris, C. B. A Simulation Program for Dynamic Infrared (IR) Spectra. Journal of Chemical. Education. 2013, 90, 4, 506–507. https://doi.org/10.1021/ed3006852

[9] Fisher, A. A., An Introduction to Coding with Matlab: Simulation of X-ray Photoelectron Spectroscopy by Employing Slater's Rules. Journal of Chemical Education. 2019, 96, 1502-1505. https://doi.org/10.1021/acs.jchemed.9b00236

[10] Hall, V., Nash, A., Rodger, A. SSNN, A Method for Neural Network Protein Secondary Structure Fitting Using Circular Dichroism Data. Analytical Methods. 2014, 6(17), 6721-6726. https://doi.org/10.1039/c3ay41831f

[11] Ang, L. D. Biophysical and Computational Studies of Biomolecular System. Ph. D. Dissertation, Western Sydney University, Sydney, 2019.

[12] A Pinto Corujo M., Olamoyesan A., Tukova A, Ang D, Goormaghtigh E., Peterson J., Sharov V., Chmel N. Rodger A. SOMSpec as a General Purpose Validated Self-Organising Map Tool for Rapid Protein Secondary Structure Prediction from Infrared Absorbance Data. Frontier Chemistry. 2022, 9, 784625. https://doi.org/10.3389/fchem.2021.784625

[13] Bansal, R., Elgundi, Z., Goodchild, S. C., Care, A., Lord, M. S., Rodger, A., Sunna, A. The Effect of Oligomerization on a Solid-binding Peptide Binding to Silica-based Materials. Nanomaterials 2020, 10 (6), 1070. https://doi.org/10.3390/nano10061070

[14] Olamoyesan, A., Ang, D., Rodger, A. Circular Dichroism for Secondary Structure Determination of Proteins with Unfolded Domains Using a Self-organising Map Algorithm SOMSpec. RSC Advances 2021, 11 (39), 23985-23991. https://doi.org/10.1039/ d1ra02898g

[15] Olamoyesan, A., Rodger, A. Application of Derandomisation to Peptide Circular Dichroism Spectra to Determine their Secondary Structure Content. South African. Journal Chemistry. 2024, 78, 52–60. https://doi.org/10.17159/0379-4350/2024/v78a10

[16] Sklepari, M., Rodger, A., Reason, A., Jamshidi, S., Prokesa, I., Blindauera, C. A. Biophysical Characterization of a Protein for Structure Comparison: Methods for Identifying Insulin Structural Changes. Analytical Methods. 2016, 8, 7460-7471. https://doi.org/10.1039/c6ay01573e

[17] Vecchio, I., Tornali, C., Bragazzi, N., Martini, M. The Discovery of Insulin: An Important Milestone in the History of Medicine. Frontiers Endocrinology. 2018, 613 (9). 1-8. https://doi.org/10.3389/fendo.2018.00613

[18] Hall, V. A. Self-organising Map Machine Learning Approach to Pattern Recognition for Protein Secondary Structures and Robotic Limb Control, Ph.D. Dissertation, University of Warwick, 2014.

[19] Erik J. M. Series of Jupyter Notebooks Using Python for an Analytical Chemistry Course. Journal of Chemical. Education. 2020, 97, 3899-390. https://doi.org/10.1021/acs.jchemed.9b01131.3

Biography

Adewale Olamoyesan received his PhD in Chemistry and Biomolecular Sciences from Macquarie University. Olamoyesan has been at University of Lagos for a couple of years as an honorarium lecturer before taking up another role at Wigwe University. At University of Lagos, Adewale teaches Physical Chemistry and coordinates laboratory sessions for this course.

Research Field

Adewale Olamoyesan: Biophysics, Computational Chemistry, Biophysical Chemistry, Physical Chemistry, Material Science.