Research Article

iMedPub Journals
http://www.imedpub.com

American Journal of Computer Science and Information Technology

2021
Vol. 9 No. 2: 76

Design a Simple Computational Protocol to Estimate the Secondary Structure of Proteins Using the Amino Acid Propensities and the Graphic Programming Language LabVIEW

Glomen Tovar*

Institute of Biomedical Research (BIOMED), University of Carabobo, Núcleo Aragua, Las Delicias Maracay, Venezuela

***Corresponding author:**
Glomen Tovar, Institute of Biomedical Research (BIOMED), University of Carabobo, Núcleo Aragua, Las Delicias Maracay, Venezuela

tglomen@gmail.com

Citation: Tovar G (2021) Design a Simple Computational Protocol to Estimate the Secondary Structure of Proteins Using the Amino Acid Propensities and the Graphic Programming Language LabVIEW. Am J Compt Sci Inform Technol Vol.9 No.2: 76.

Abstract

A simple computational protocol to estimate the secondary structure of a protein is developed, using the amino acid propensities and the graphic programming language LabVIEW. The protocol estimates the number of residues of the structures α-helix, β-sheet and turns; their validation can be checked by the method of least-squares in the different established relationships. 31 proteins were analyzed using the NADH dehydrogenase as a standard to evaluate the performance of protocol, the results obtained when comparing the values of the programmed structures with those calculated, showed that the Pearson's correlation coefficients (r) obtained were statistically significant (P<0.001). This result shows that the number of α-helix, β-sheet and turns are a linear function of the number of amino acid residues in the proteins. It can be concluded that this procedure can be useful for novice researchers with limited technological resources and poor laboratory infrastructure, to estimate the secondary structure of a variety of proteins, suggesting some consequences in protein folding.

Keywords: Protocol; Estimate; Least squares; Propensities; Secondary structure; LabVIEW

Received: February 08, 2021; **Accepted:** February 22, 2021; **Published:** March 01, 2021

Introduction

Estimate the secondary structure of proteins is a fundamental part in the analytical study of their structure and function, whose sequences can help to assess their superior structures. Establishing a computational protocol to estimate the secondary structure of proteins, it can be knows the composition in terms of its amino acid sequence in order to reach its presumption [1]. In first instance, it can be visualized in a UniProt Swiss-Prot text-based computer file format FASTA the amino acid sequence, transpose it and generate a file. In the past decade's technique of predicting the secondary structure of proteins has been developed in several studies. The Chou-Fasman method [2], was one of the first methods tested for this purpose, which is based on a successful statistical procedure on the potential conformation or propensities for all amino acid residues [3]. However, this method presented limitations due to the low precision in the results. These values have been utilized to offer a simple procedure without complex computer calculations to predict the secondary structure of proteins from their known amino acid sequence. Other prediction methods were developed, based on different algorithms, such as: the advent of X-ray crystallography [4], involving compose, to create the amino acids of the secondary structures of the proteins [5], exhaustive statistical analyses of the amino acid sequence as a function of (C) and (N) terminal [6, 7], improvement of multiple linear regression methods based on the frequency of different amino acids and the length of the protein [8,9], improvement of the Chou-Fasman method thanks to the recent development in the folding of proteins, through improvements in propensities and the wavelet transform [10], choice of the best data set of proteins to check their propensities [11]. Other methods develop algorithms based on, such as, non-linear models of neural networks [12,13], methods of the nearest neighbor [14,15], using multiple alignments [16], production of FASTA files from complete α-helix, β-sheet, and turns sequences using their propensities in computational applications or by their hydrophobic characteristics [17] among others. In the present work, the goal is to develop a simple protocol to estimate the secondary structure of proteins using LabVIEW programming language, the propensities of amino acids are finally evaluated and validated by comparing the results obtained in the structure programmed with the calculated structure. The protocol can be

useful for novice researcher with limited technological resources and laboratory infrastructure.

Materials and Methods

Propensity of amino acids in different types of secondary structures

The Chou-Fasman method is an empirical technique for predict of secondary structures of proteins, originally developed. The method is based on the analysis of the relative frequencies of each amino acid in α-helix, β-sheets and turns on structures of known proteins solved with X-ray crystallography. From these frequencies, a set of probability parameters was derived for the appearance of each amino acid in each type of secondary structure, and these parameters are used to predict the probability that a given sequence of amino acids will form α-helix, a β-chain or a turn in a protein.

Residual propensity values (P (ES)) in different types of secondary structures, were determined from the ratio of the frequency of occurrence of the residue in α-helix(Pα), β-chain(Pβ) and turns(Pturn) versus their frequency of appearance in the protein by the following expression:

P (ES)=fraction of the residue/fraction of total residue (1)

In this work the propensities of Chou Fasman were used. The values can be seen, with the following considerations:

1. Any segment of six residues or more in a native protein with <Pα> ≥ 1.03 and <Pα >><Pβ>, is predicted as helical.

2. Any segment of three residues or more in a native protein with <Pβ> ≥ 1.05 and <Pβ >><Pα>, is predicted as β-sheet.

3. Any segment of four or more residues in a native protein with <Pturn> ≥ 1.00 and P (α-helix)<P (turn)>P (β-sheet), is predicted as a turn sheet.

The values shown in (1) were applied to get the graphs of the secondary structures (α, β and turns), considering the values of <Pα>, <Pβ> and <Pturn> ≥ 1.00 both for non-superimposed and superimposed on the front panel.

However, there are other versions of propensity estimates for the amino acids of other authors but they do not differ appreciably.

Method of least-squares: The least-squares is a numerical analysis procedure, in which we try to find the continuous function that best approximates the data by providing a visual demonstration of the relationship, between the points of the data and is expressed in the equation of the straight line, expressed as follows:

$$y=mx+b \quad (2)$$

Where m is the slope and b the interception with the "x" axe.

The development and application of the equation, allows to search for a line of better fit that explains the possible relationship between the independent variable (in this case is the number of amino acid residues of the NADH dehydrogenase protein) and the dependent variable (number of residues of the structures secondary α-helix, β-sheet and turns). From these designations is obtained the equation for the line of best fit, determined from the method of least-squares validating the values gotten in the protocol designed [18].

Software used: The software LabVIEW (a trademark of National Instruments of Austin TX brand) is used to create the computational protocol, this software is a programming language that uses icons instead of lines of text. The user interface (known as the front panel) is constructed from graphic codes. The block diagram has the lines connected by guiding the flow of information from one code to the next [19].

Protocol: A set of proteins obtained from the UniProt database were used to build and validate the protocol. This protocol can be used to estimate about the secondary structure of the proteins, by means of a programmatic procedure and validated using the method of the least-squares [20].

The DATASET or amino acid sequences of the proteins were obtained through the UniProt FASTA. A total of 31 proteins randomly chosen with avary of the number of residues from 54 to 476 represent a significant sample to relate them to the number of secondary structures (α-helix, β-sheet and turns).This relationship was made based on its conformational parameters or propensities, which represents an intrinsic property of the amino acids.

Once the DATASET has been selected, it is necessary to create a SAR (Amino acid Residue String), that is amino acid residues of the protein coming from the FASTA format, which must be aligned through program TAAS (Transposition Algorithm Amino acid Sequences), to generate a xls file and converter a csv file for transposition or GFTS (Generate File Transpose Sequences). These files are introduced to the program or SSPA (Secondary Structure Predictive Algorithm), whose development will generate the -xls file of the estimated protein or GFPSS (Generate File Protein Secondary Structure), whose operability is explained in the programs following (**Figure 1**).

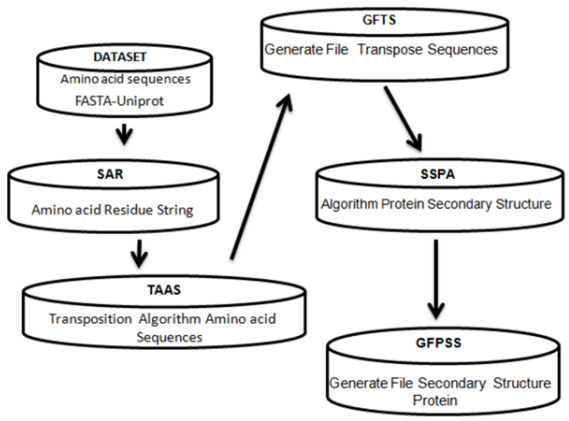

Figure 1: Protocol to predict secondary structure of proteins: DATASET (Amino Acid Sequences FASTA-Uniprot) Abbreviations: SAR: Amino acid Residues String; TAAS: Transposition Algorithm Amino Acid Sequences; GTFS Generate File Transpose Sequences; SSPA: Secondary Structure Predictive Algorithm; GFPSS: Generate File Secondary Structure Protein.

Operability of the programs

TAAS (Transposition Algorithm Amino acid Sequences): The algorithm generated in the so-called block diagram allows to concatenate amino acid sequences (FASTA) through the concatenate string function, which are connected to the programming structures: While loop that has the function of stopping program execution, the programmatic structure event structure that executes the action commands of the program and the for Loop structure that allows iteration of the data supplied from the concatenated sequences [21]. These structures generate data that is quantified using the string length function and once the program is activated, it generates the information; it transmits and connects to the reshape array function, which changes the dimensions of an array according to the size values of the dimension. When the previous actions are completed, an output array is generated that connects to the Insert into array function to which a control or column name is added and whose result is finally connected to a write to spreadsheet file, where it is generated the information that designs the route from the -xls file (**Figure 2**).

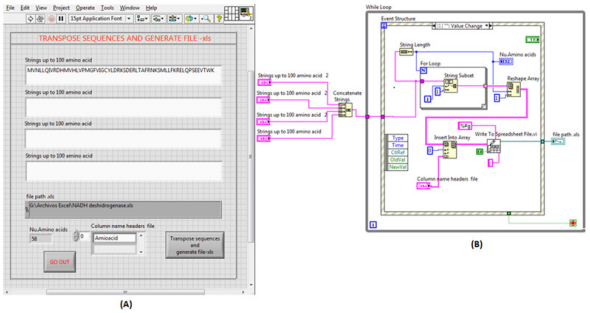

Figure 2: Block diagram to visualize programming structure (while loop, even structure and for loop) to transpose sequences in NADH dehydrogenase and generate-xls file.

Front panel: When a virtual instrument (vi) is opened in LabVIEW, the front panel window representing the user interface appears. In the case of the TAAS program, it represents the virtual instrument of the user interface, which is made up of 4 string control where the "strings" that represent the amino acid sequences in the FASTA format and a file path control where the name and path of the generated file [22]. This file is generated by controls numerical where the size of the analyzed sequence appears, once the program has been activated with the button to generate the file, an array control to name the columns of the parameters of the data to be analyzed, and a program stop of exit button (**Figure 3**).

SSPA (Secondary Structure Predictive Algorithm): This algorithm originated in the so-called block diagram allows producing XY graphs, control tables and -xls file that offer data to distinguish the secondary structure of the protein. These elements are produced by beginning to explore through the programming functions Read from Spreadsheet File, the -csv files of the protein sample, the patterns of the secondary structures and the overlapping and not overlapping structures; which will then be connected through a programming function delete from array with a for loop structure to iterate with a Search 1D array and an index array that are associated with a ship register and a built array, to produce an Array containing the information required to complement the action of the program structures (while loop, event structure, for loop and case structure). These structures will operate to originate the expected results when associated with other programming functions such as fract/exp, string to number and transpose 2D array to get the XY graphs of the secondary structures α-helix, β-sheet, turns and overlay graphics. At the same time the array produced that has the information required to complement the action of the programmatic structures, is associated with another functions as transpose array and a delete from array to link with the programming structures for loop and case structure to discriminate among them the α, β and turn forms that when associated in a built array generate a table of overlapping secondary structures and a -xls file using a function write to spreadsheet file (**Figure 4**).

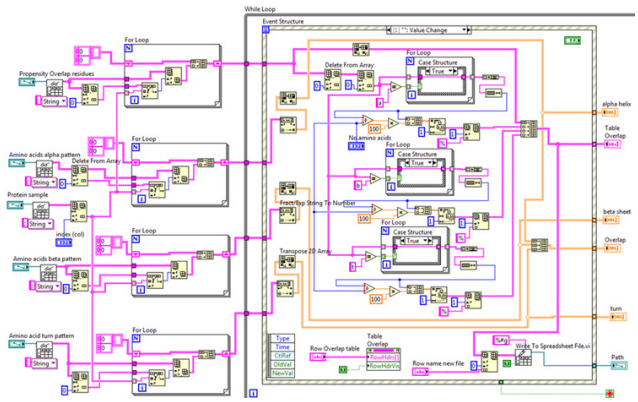

Figure 3: Front panel to visualize transpose sequences secondary structure in NADH dehydrogenase and generated path-xls file.

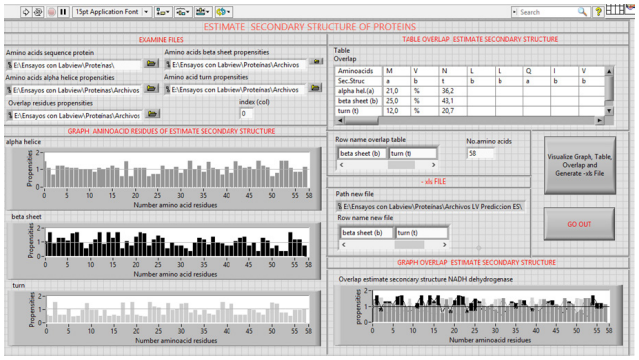

Figure 4: Block diagram for programming structures (for loop, while loop, case structure and event structure) in the estimation of the secondary structure in NADH dehydrogenase.

Front panel: When the virtual instrument for estimate the secondary structure of proteins is open in LabVIEW, the user interface appears, in this case, the estimate algorithm of the secondary structure of the protein is constituted by a set of express control that are associated with the functions and controls of the user interface. In this way, can be found text

control as file path control, for proteins, for amino acid patterns (α-helix, β-sheet and turns) and superimposed residuals, XY graph for α-helix, β-sheet, turns and overlays; list, table and tree of table for table of superimposed residues, string and path for generation of -xls file, numeric controls for name of rows of table and –xls file; button and switches to visualize graphics, tables, graphics of superposition, generate files and of exit of the program (**Figure 5**).

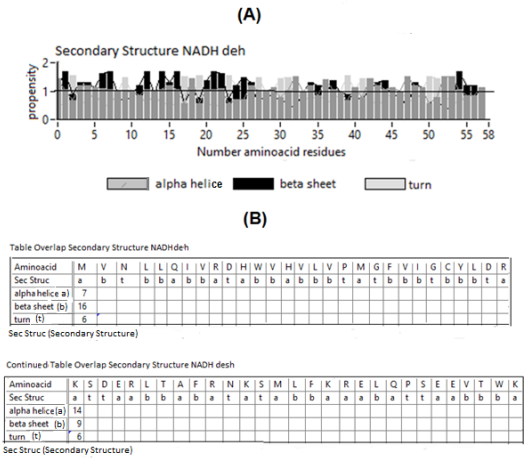

Figure 5: (A) Graph overlap estimate secondary structure in NADH dehydrogenase and (B) Table overlap estimate secondary structure in NADH dehydrogenase.

The programs are activated in the status toolbar with the program execution buttons.

Particle swarm-based scheduling agent

In this section, the scheduling factor is an imitation-based method to create an optimal scheduling of the PSO algorithm. In this algorithm, the population is equal to the number of particles in the problem space. The particles are randomly initialized. Each particle will have a compatibility value and will be evaluated by a compatibility function that must be optimized in each generation [23,24]. In other words, this solution is equivalent to a bird in the pattern of collective movement of birds. Each particle has a merit value calculated by a merit function. The closer the particle is to the target in the search space (food in the bird movement model), the greater the merit is. Each particle also has a velocity that directs the particle's motion. Each particle continues to move in the problem space by following the optimal particles in the current state (**Figure 6**).

Click on → and activate → to run the program

With the program activated the button is pressed and the user must introduce the information required by the program, in order to generate the -xls file and visualize the table and the graphics of amino acid residues, without superimposing and superimposed, that have been related to the propensities of the amino acids of secondary structure α-helix, β-sheet and turns (**Figure 7**) (**Tables 1 and 2**).

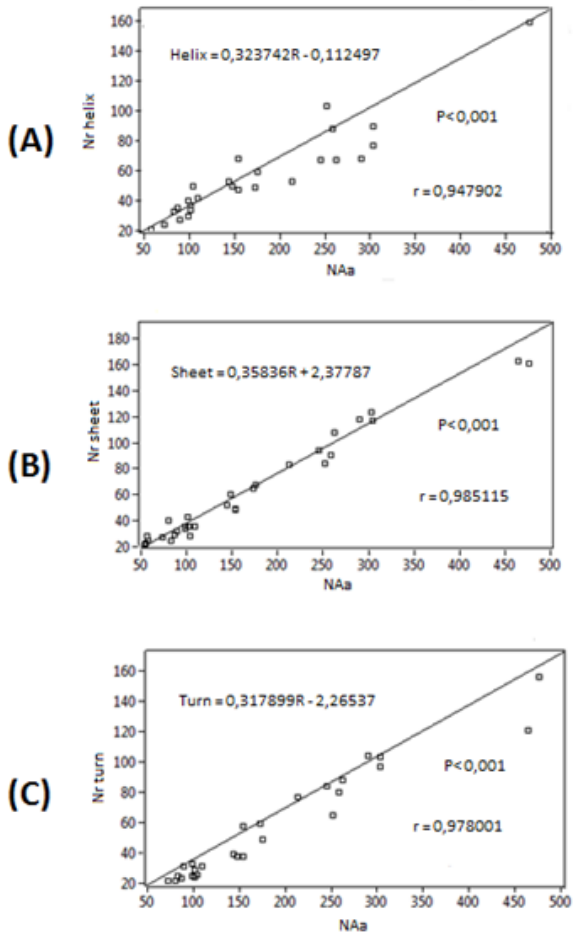

Figure 6: Least squares analysis as linear function between the number of residues (Nr) and the number of amino acids (Na) in NADH dehydrogenase for (A) alpha helice (B) beta sheet and (C) turns.

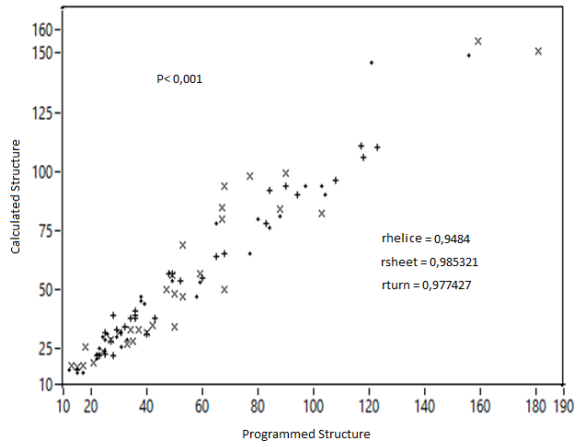

Figure 7: Pearson correlation between the secondary structures (xxx) alpha helice, (+++) beta sheet and (...) turns programmed and calculated of the 31 proteins analysed.

Amino acid	Propensity		
	Pα	Pβ	Pt
A-Ala	1,4	0,8	0,7
R-Arg	1,0	0,9	1,0
D-Asp	1,0	0,5	1,5
N-Asn	0,7	0,9	1,6
C-Cys	0,7	1,2	1,2
E-Glu	1,5	0,4	0,7
Q-Gln	1,1	1,1	1,0
G-Gly	0,6	0,8	1,6
H-His	1,0	0,9	1,0
I-Ile	1,1	1,6	0,5
L-Leu	1,2	1,3	0,6
K-Lys	1,1	0,7	1,0
M-Met	1,5	1,1	0,6
F-Phe	1,1	1,4	0,6
P-Pro	0,6	0,6	1,5
S-Ser	0,8	0,8	1,4
T-Thr	0,8	1,2	1,0
W-Trp	0,8	1,2	1,0
Y-Tyr	0,7	1,5	1,1
V-Val	1,1	1,7	0,5
Pα (Propensity alpha-helices)			
Pβ (Propensity beta-sheet)			
Pt (Propensity turn)			

Table 1: Relative amino acids propensity value for secondary structure used in the Chou-fasman methods.

Cod. Uniprot	Proteins	NAs	Programmed structure			Calculated structure		
			Helix	sheet	Turn	Helix	Sheet	Turn
			Nr	Nr	Nr	Nr	Nr	Nr
P18669	Phosphoglycerate mutase 1	525	103	84	65	81	93	78
P00168	Cytochrome b5	87	35	29	23	27	94	26
Q3E846	Cytochrome c oxidase	104	50	28	26	32	40	32
Q9HTK7	Rubredoxin 3	54	15	22	17	16	22	16
Q4HWU5	Trypsin inhibitor	56	13	28	15	17	23	17
P80176	High potencial iron	83	33	25	25	26	32	25
P08905	Lysosome C -2	148	50	60	38	47	55	46
P83443	Macrodontain-1	213	53	83	77	68	79	66
P40571	Ribonuclease P	144	53	52	39	45	54	44
P00272	Rubredoxin 2b	173	49	65	59	55	64	53
Q9HTK8	Rubredoxin 2a	55	17	23	15	16	22	16
P04069	Carboxypeptidase B	303	77	123	103	97	111	94
P00766	Chymotrypsinogen A	245	59	103	83	78	90	76
P07630	Carbonic anhydrase	253	88	90	80	82	95	80
P168870	Carboxypeptidase E	476	159	161	156	153	173	149
P17538	Chymotrypsinogen B	263	67	108	88	84	97	82
P02866	Chonconavalin A	290	68	118	104	93	106	90
P61949	Ravadoxin 1	176	59	68	49	56	65	54
P35557	Glucokinase	465	181	163	121	150	169	146
O75438	NADH dehydrogenase	58	21	25	12	17	23	17
P00441	Superoxide dismutase	154	47	49	58	49	58	47
P83748	Serine Protease	304	90	117	97	97	111	95
P00044	Cytochrome c iso-1	109	42	36	31	34	42	33
A0A1 K0FU49	Myoglobin	154	68	48	38	49	58	47
P022655	Apolipoprotein C2	101	34	43	24	31	39	31
F7VJQ1	Alternative P nonprotein	73	24	27	22	22	29	22
PO2656	Apolipoprotein C3	99	40	34	25	25	38	30
QOVDE8	Adipogenin	80	18	40	22	31	31	24
Q13015	Protein A F1 q	90	27	32	31	35	35	27
Q9NZD4	Alpha hemoglobin	102	37	36	29	39	39	31
P14621	Acyphosphatase-2	99	30	36	33	38	38	30
Na(Amino Acids Number) Nr(Residues Number)								

Table 2: Secondary structure (alpha-helices, beta-sheet and turn) programmed and calculated of protein Uniprot.

Results and Discussion

The operability of the program is evidenced with the use of a protein sample extracted from the DATASET FASTA of UniProt from a total of 31 proteins randomly chosen; the NADH dehydrogenase of 58 amino acid residues was selected. When the program is activated in the bar tools, the front panel asks, for the place and identification of the corresponding parameters, then the file name is generated, the representative graphs of the amino acid residues of the secondary structure will be displayed (α-helix, β-sheet and turns) on the front panel. The graph of the amino acid residues of the secondary structure superimposed (define the protein secondary structure estimated), the table of the superimposed residues that show the quantified state of the same and the site path of the -xls file can observed. The result of the process described in the graph and the table of the quantification of the superimposed amino acid residues of the NADH dehydrogenase were extracted.

The procedure of the activation of the program previously described, was applied completely in the set of 31 selected proteins (DATASET FASTA of UniProt), obtaining the results quantified programmatically for each of them.

The values obtained with the program were validated using the method of the least-squares.

Both results were tabulated for the 31 proteins, reflecting the number of amino acids (Na), and the number of amino acid residues (Nr) of the estimated secondary structures (α-helix, β-sheet and turns).

It was constructed a X Y graph to relate the values obtained by the method of least-squares, and the values of the programmed residues (Nr) with the number of amino acids (Na) of the protein samples, obtaining the expressions of the straight lines for each structure programmed with their pearson correlation coefficients (r), which shown statistically significant results for $P<0.001$.

With the expressions of the straight line of the secondary structures (α-helices, β-sheet, and turns), the values of the calculated residuals (Nr) were obtained. All the values for the amino acid residues (Nr) of the programmed and calculated secondary structure in the 31 proteins set and their Pearson correlation coefficients (r), which showed statistically significant results for a $P<0.001$. This result shows that the number of α-helix, β-sheet and turns are a linear function of the number of amino acid residues in the protein, it can be suggested this fact as a result for protein folding due fundamentally to the linearity of the turns in the peptide chain as indicated.

Conclusion

A simple computational protocol to estimate the secondary structure of a protein is developed in this research, using the propensities of the amino acids and the graphic programming language LabVIEW.

The NADH dehydrogenase protein is utilized to evaluates the performance of protocol, the results shown that Pearson's correlation coefficients (r) obtained is statistically significant ($P<0.001$).

Estimate the secondary structure of proteins using the LabVIEW programming language, provides a tool to be used in some daily laboratory routines where the resources of a laboratory and the technology are insufficient to fulfill the purposes suitable for an investigation. That is because this protocol could solve these deficiency and help in the routine tasks of novice researchers, and can be used to estimate the properties of a variety of proteins.

The program would allow the user to manage a data from an Excel file in the -csv format for amino acid sequence of a protein, using residues of the 20 amino acids with propensity for structures (α, β or turns) without overlapping and overlapping.

In this way, the user can manipulate any protein in the previously proposed scheme and get all the necessary information to graph, tabulate and store information generated in -xls files that will be useful for his research work.

Acknowledgements

I want to thank to my daughter, Carolina Tovar PhD, Professor of the Central University of Venezuela for her valuable advice and continue support in writing the manuscript. I also want to thank to my son, Engineer Raimundo Tovar for his insightful comments on the manuscript.

The world is trying to become completely wireless, demanding uninterrupted access to information anytime and anywhere with better quality, high speed, increased bandwidth and reduction in cost.

References

1. Anfinsen CB (1973) Principles that govern the folding of proteins chains. Science181: 223-230.
2. Chou PY, Fasman GD (1974) Prediction of protein conformation. Biochemistry 13: 222-244.
3. Chou PY, Fasman GD (1978) Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 47: 45-147.
4. Wlodawer A, Minor W, Dauter Z, Jaskolski M (2013) Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. FEBS Journal 280: 5705-5736.
5. Argos P, Palau J (1982) Amino acid distribution in protein secondary structure. Int J Peptide Protein Res 19: 380-393.
6. Brendel V, Bucher P, Nourbakhsh IL, Blaisdell BE, Karlin S (1992) Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Scie 89: 2002-2006.
7. Berezovsky IN, Kilozanidze GT, Tumanian VG, Kisselev LL (1999) Amini acid composition of protein termini are biased in different manners. Protein Engineering 12: 23-30.
8. Krigbaum WR, Knutton SP (1973) Prediction of the amount of secondary structure in a globular protein from its amino acid composition. Proc Nat Acad Sc 70: 2809-2013.

9. Zhang Chun-Tim, Zhang Z, He Z Prediction of the secondary structure content of globular protein based on three structural classes. J Protein Chem 17: 261-272.

10. Chen H, Gu F, Huang Z (2006) Improved Chou-Fasman method for protein secondary structure prediction. BMC Bioinformatics 7: 1-14.

11. Costantini S, Colonna G, Facchiano AM (2006) Amino acid propensities for secondary structure are influenced by the protein structural class. Bioch Byoph Res Comm 342: 441-451.

12. Qian N, Sejnowski TJ (1988) Predicting the secondary structure of globular proteins using neural network models. J Mol Biol 202: 865-884.

13. Rashid S, Saraswathi S, Kloczkowski A, Sundaram S, Kolinski A (2016) Protein secondary structure prediction using a small training set (compact model) combined with a Complex-valued neural network approach. BMC Bioinformatics 17: 1-18.

14. Yi TM, Lander ES (1993) Protein structure prediction using nearest neighbor methods. J Mol Biol 232: 1117-1129.

15. Akcesme FB (2015) Protein secondary structure prediction based on physicochemical features and PSSM by KNN. Southeast Eur J Soft Comput 4: 37-42.

16. Levin JM, Pascarella S, Argos P, Garnier J (1993) Quantification of secondary structure prediction improvement using multiple alignments. Protein Eng 6: 849–854.

17. Gromiha MM, Ponnuswamy PK (1995) Prediction of protein secondary structures from their hydrophobic characteristics. Int J Pept Protein Res 45: 225-240.

18. Miller S (2006) The method of least squares. 114: 1-7.

19. Kalkman CJ (1955) LabVIEW A software system for data acquisition, data analysis and instrument control. J Clin Monit 11: 51-58.

20. Rose GD, Wetlaufer DB (1977) The number of turns in globular protein. Nature 268: 769-770.

21. Levitt M (1978) Conformational preferences of amino acids in globular proteins. Biochemistry 17: 4277-4284.

22. Nakashima H, Nishikawa K, Ooi T (1986) The folding type of a protein is relevant to the amino acid composition. J Biochem 99: 153–162.

23. Gao J, Wu Z, Hu G, Wang Kui, Son J, et al. (2018) Survey of predictors of propensity production and crystallization with application to predict resolution of crystal structures. Curr Protein Pept Sci 19: 1-11.

24. Nawaz R, Islam UL, Roy Ch, Gupta PSS, Banerjee Sh, et al. (2018) PROPAB: Computation of propensities and other properties from segments of 3D structure of proteins. Bioinformation 14: 190-193.