PARS2: An R package for Protein Assignment Regarding Secondary Structure

Emran Heshmati*
Department of Biology, Faculty of science, University of Zanjan, Zanjan-Iran.
heshmati@znu.ac.ir

Amirhosein Fathinavid
Department of computer Science, College of engineering, Hamedan Branch
Islamic Azad University
Hamadan-Iran
a.fathi.navid@gmail.com

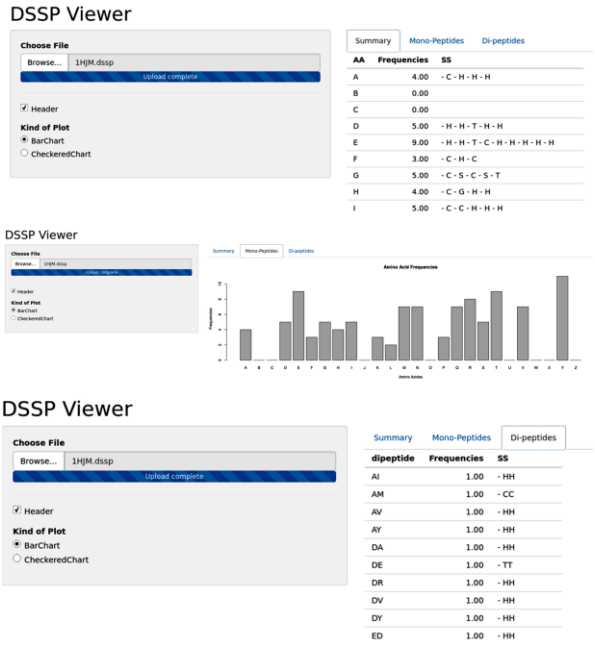
Khosrow Khalifeh
Department of Biology, Faculty of science, University of Zanjan, Zanjan-Iran.
khalifeh@znu.ac.ir

ABSTRACT
We developed a package, namely PARS2 (Protein Assignment Regarding Secondary Structure) to calculate the tendency of different mono, di and tripeptides to localize in different secondary structural elements of proteins according to the DSSP secondary structure nomenclature. The application may be used for constructing a dataset of di or tripeptides of selected protein groups. It is implemented in R programming language, using the dynamic programming method. Using this program it is now possible, for example, to calculate the probability of inserting a residue in the first or second position of a dipeptide according to the context of secondary structural element. The output of this program can not only be used to construct a detailed structural dataset for different fields of studies including machine learning methods, but it can also provide information essential for protein designing and site-directed mutagenesis in enzyme biotechnology.

Keywords
Protein, bioinformatics, R programming, monopeptide, dipeptide, secondary structure.

1. INTRODUCTION
One the one hand, twentieth centuries witnessed a huge progress in the field of molecular biology which results in the implementation of different biological databases[1]. On the other hand, there have been major progress in computer sciences which covered the ground between the raw biological data and their quantitative manipulation. However, not only will many fundamental questions in molecular biology remain unsolved, but also many more will arise in the future. Protein folding or the relationship between the linear sequence and the 3D structure of proteins, is one of the most important questions in structural biology[2]. The linear sequence of the amino acids within a protein gains the local conformation known as secondary structure. Different types of secondary structures join to each other via long range interactions and constitute the tertiary structure of a protein. It is now accepted that all the information for a protein to gain its tertiary structure is coded in its primary structure. Decoding this information is of great interest for scientist. It appears that advances in prediction algorithms accompanied by manipulation of biological data using these achievements can help the scientists to pave the way to unravel the protein folding problem. Initial works of Chou and Fasman[3,4] which implemented according to the probability of finding individual amino acids in each secondary structural element, has been used as the basis for the prediction of the protein structures. Thanks to the new developed PARS 1 software, the probability of finding di- and tripeptides in each secondary structure was investigated by our research group[5,6]. In recent work we present the new version of PARS known as PARS 2 which was set up in the R-language program and has new capabilities.

2. METHOD
For the implementation of the method, we have been used R/Shiny package. Shiny is an R package, making it easy to build interactive web applications (apps) directly from R (https://shiny.rstudio.com). To install the Shiny package, R session should be opened followed by running the > install.packages("shiny" command .
The main reasons for using this package are due to the fact that this programming language uses different data structures to perform statistical analysis. Furthermore, web development with this tool does not require extensive information. All the PDB files can be read through a dialog box and the files then would be read in client side. All process has been done in server side and the results could be sent to the client. After reading a PDB file, the program convert it into DSSP file format, extracts the amino acid contents, saves the result into data frame structure and lists them in final output.



Figur1: The results of calculation performed on monopeptides and dipeptides which are shown in distinct tabs.

Moreover, the application does the following important tasks: Through some functions, the list of monopeptides with the related frequencies of each secondary structure is calculated and the total frequencies regarding to individual

amino acids could also be shown in a diagram. This is useful for the researchers that want to find the most repetitions in a polypeptide chain. This application does the same work for dipeptides. Dynamic programming method is used for all calculations. The results are shown in distinct tabs (Fig. 1) and can be saved in both graphic and text formats (Figs. 2a and 2b).

In fact, the application can save the content of input for further processing. In the Fig.1 there are three columns. In the first one, the name of individual amino acids are shown. The second column shows their frequencies and in the third column the secondary structure of each amino acid is shown based on its frequencies.

	H	E	C	total
A	49	89	715	853
C	12	19	159	190
D	44	80	1257	1381
E	37	46	835	918
F	52	93	336	481
G	31	61	1878	1970
H	14	40	348	402
I	39	149	394	582
K	50	84	796	930
L	64	142	699	905
M	12	36	127	175
N	31	64	802	897
P	2	89	749	840
Q	23	35	394	452
R	35	79	654	768
S	32	115	959	1106
T	45	113	914	1072
V	69	154	507	730
W	22	24	123	169
Y	41	80	334	455
total	704	1592	12980	15276

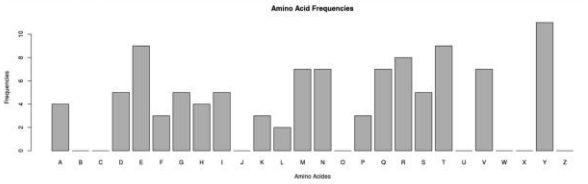


Figure2a: saved results in spreadsheet (up) and graphic (down) format for monopeptides.

3. RESULTS AND DISCUSSION

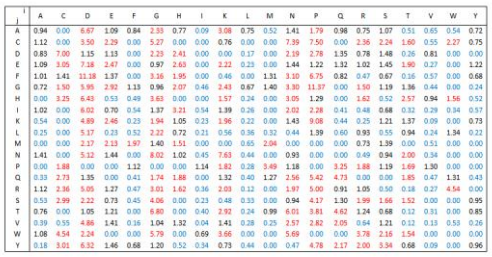
There are some advantages of using PARS 2 in comparison with previous version which include:

1. Accepting the PDB file rather than DSSP which allows data to be processed more simple and fast.
2. The output is categorized for individual combinations of the secondary structures including αα, αβ, αc, ββ, βα, βc, cα, cβ and cc.
3. The output for mono and dipeptides is presented graphically as well as .xls file format.

Using this software, it is possible to form different structural datasets of proteins for creation. The output of this software provides a data source containing valuable information for designing mutation in site-directed mutagenesis strategies. Also, it is possible now to develop software to automated mutation design.

4. ACKNOWLEDGEMENT

We thank the research council of the University of Zanjan for providing us this research opportunity.



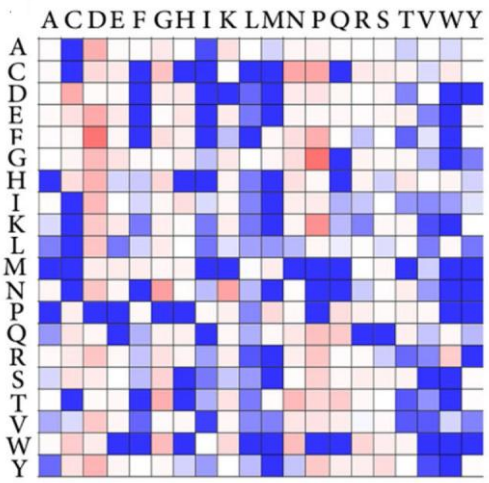


Figure2b: saved results in spreadsheet (up) and graphic (down) format for dipeptides in ββ conformation as an example.

REFERENCES

[1] E.C. Meng, E.F. Pettersen, G.S. Couch, C.C. Huang, T.E. Ferrin, Tools for integrated sequence-structure analysis with UCSF Chimera., BMC Bioinformatics. 7 (2006) 339. doi:10.1186/1471-2105-7-339.

[2] D. Baker, A. Sali, Protein structure prediction and structural genomics, Science (80-.). 294 (2001) 93–96. doi:10.1126/science.1065659.

[3] P.Y. Chou, G.D. Fasman, Prediction of protein conformation., Biochemistry. 13 (1974) 222–245. doi:10.1002/pro.5560021016.

[4] J. Kyngäs, J. Valjakka, Unreliability of the Chou-Fasman parameters in predicting protein secondary structure., Protein Eng. 11 (1998) 345–348. doi:061/10.

[5] M. Ghadimi, E. Heshmati, K. Khalifeh, Distribution of dipeptides in different protein structural classes: An effort to find new similarities, Eur. Biophys. J. (2017). doi:10.1007/s00249-017-1226-6.

[6] M. Ghadimi, K. Khalifeh, Emran Heshmati, Neighbor Effect and Local Conformation in Protein Structures, Amino Acids. Accepted A (2017). doi:10.1007/s00726-017-2463-9.