MANUSCRIPT

A PROTOTYPE WEB APPLICATION PACKAGE FOR BASIC DNA AND PROTEIN ANALYSIS USING R LANGUAGE

MR SWAMINATHAN VENKATARAMANAN[1]

SIVA KUMAR CHANDRAN[1]

PROF.DATO DR. MD GAPAR MD. JOHAR[2]

DEPARTMENT OF DIAGNOSTIC AND ALLIED HEALTH SCIENCE[1]

INFORMATION TECHNOLOGY AND INNOVATION CENTER[2]

Email: s_venkataramanan@msu.edu.my

ABSTRACT

Analysis of DNA and protein has become a very important aspect in the field of research, especially for Bioinformatics. This is important as the basic analysis of these protein and DNA can lead to further advanced analysis of the sequence, which may lead to new discoveries. Basic analysis of sequences is done in the industry, research as well as education. R language is a statistical program that is used in the analysis of DNA and protein sequences, through the application of packages in the Comprehensive R Archive Network. This analysis package helps to analyze sequences, but in a command prompt analysis. However, the process is slow as the researcher has to enter several lines of codes to obtain the result for the analysis. The research is to develop a prototype web application package with an interactive new interface for the DNA and protein analysis. The prototype is fully coded in R with options to download the results as well as providing information about the codes being used for the analysis and the package reference. This application is made to assist in the sequence analysis of DNA and protein without having to write the codes.

KEYWORDS: R, Bioinformatics, sequence analysis, web application, statistics

INTRODUCTION

Bioinformatics is a hybrid field consisting of different field such as biology, statistics, chemistry and genetics with the addition of information technology in the analysis and the interpretation of biological data. Some of the fields in Bioinformatics includes sequence and structure analysis of sequences of DNA and protein. Manipulation of sequences can be done via sequence analysis in Bioinformatics which includes statistical outputs as a theoretical value and result. R language is a GNU language(Kim, 2007) that emphasizes in statistical analysis with simple data analysis and data visualization. The packages in R are in alphanumeric form that helps programmers to script codes using the packages (Jinlong, 2011) There are packages in the R archives which contains codes for statistical analysis, including sequence manipulation. R language is an open source language, thus it is freely distributed among people in the Internet. R is also integrative as it allows the language to be implemented with other languages such as C++ (Dirk and Romain, 2011) and Java. In the field of Bioinformatics, R language is used in the analysis of sequences to produce statistical data using analysis packages which is accessed using a R terminal that requires writing long lines of codes and sometimes in a strange arrangement. These R codes can be used in the analysis of DNA sequences ("What is DNA", 2014) such as length, GC count (Zheng and Wu, 2010, Oliver and Marin, 1996, Henke et al, 1997 , base count (Lobry and Lobry, 1999) as well as reverse and complement of the sequences. It can also be used in the protein analysis in the study of evolutionary analysis (Mehmet *et al*, 2006), length determination (Kingshuk and Ken, 2009, Luciano and Samuel, 2005), the isoelectric point of the protein (Kawashima *et al*, 1999, Widmann *et al*, 2010), translation from DNA to protein, amino acid statistics (Kawashima *et al*, 1999) as well as the Dot Plot analysis of both sequences (Gibbs and Mcintyre, 1970). The R codes can be used to generate graphical plots such as box-plots, histograms and charts for better analysis of data (Tina, 2014)

PROBLEM STATEMENT

Previous analysis of DNA and protein sequences using R has been done using the command line analysis, which is redundant and time consuming. Besides that, the R language is deemed difficult due to the strange code structure, making it difficult for users to learn.

METHODOLOGY

Agile Unified Process

Agile Unified Process has four stages. The Inception stage where initial analysis for R and Bioinformatics is being done, Elaboration is done to see the compatibility of the R language to the current analysis packages in the repository. Construction is where analysis packages is combined with the documentation package under one large web application, using a special web application package as a framework. Then, the prototype is uploaded to the online server and deployed for testing.

Packages used for the analysis:

a) Shiny : web application package for the R codes. Combination of the analysis package as well as the documentation package is combined within the framework.
b) knitR : Dynamic report generating package using R language. The package acts as a secondary R terminal to include the input, codes and output into a report
c) seqinR: analysis package created for sequence extraction and analysis from the databases or from random input from the user.
d) rBase : the base package in R itself which is installed with the R language. Provides the structure and syntax for the R codes

RESULTS

Several comparison analysis has been conducted in terms of the methods used for the DNA and protein sequencing and the prototype produces an accurate analysis of the data, compared to the current online web application tool as well as the command line analysis. The prototype also has a good response from the users in terms of knowledge input and provides a good documentation which includes the input, codes for the analysis and the output as well. Based on the user acceptance survey and prototype evaluation survey, the prototype web application attains a good response in terms of user interface, system analysis as well as the result production. In terms of Bioinformatics, the users agree that statistical analysis is a very important aspect in the sequence analysis in bioinformatics. They also agree that providing the codes for the analysis in the interface as well as the documentation provides new knowledge in terms of analysis method and the codes being used for the analysis. The new documentation format is also preferred by the respondents as it shows the input, codes and the output of the analysis in the same report.

The deployed application is also tested with a random number of sequences for Protein and DNA analysis. The prototype web application is compared with another online web application tool as well as the R command line. Comparison are made for all the analysis present in the prototype web application, numeric and graphical. The prototype web application has shown to produce a good and accurate analysis of the sequences, similar to the R command line analysis as well as the online web application. Certain analysis such as the Dot Plot analysis and the Amino Acid Statistics which produces a graphical output is produced similarly in the R command line, but not in the online web application tool.

DISCUSSION

The prototype passes the user acceptance test in terms of user interface, system analysis as well as result production. This is because compared to the command line analysis which is R language's main access, the users don't have to write long lines of codes to access the analysis packages. Only a click of a button and the results will be shown in the interface as well as the report. There are two spaces for the input, which allows the user to do a rough comparison of the two sequences as well in the report produced. Reports are separated between the numerical analysis and the graphical analysis Figure 2 to prevent clutter in the report production as well as helping the users to understand the results better. Implementation of R language codes in the help section provides the users with a new knowledge because the prototype not only provides a simple explanation about the analysis method being used on the

sequences but also the codes that is being used for the manipulation of the sequences. This provides a new knowledge in users as they could perform the sequence analysis, learn about the methods being used in the sequence analysis as well as the codes that is being used for the analysis.

The prototype is also easy to use as the users only has to enter the input once to obtain all the outputs of the analysis such as in Figure 3. This method prevents the redundancy of users to enter the same sequence every time they want to perform sequence analysis. This also reduces the overall time taken for sequence analysis to be done, as the users do not have to take their time in writing long lines of codes redundantly to obtain the results of the sequence analysis. Presence of graphical outputs provides an easier way to interpret the sequence analysis results compared to the textual output such as the dot plot analysis and the amino acid statistics function in the prototype web application. The presence of the user interface such as in Figure 1 and Figure 4 helps in providing a better access to the sequence analysis because the packages in R are commonly accessed in command line prompt. The user interface helps to access the analysis and documentation packages in R without the users having to write the codes to access the packages and still obtain the results as accurate as the command line analysis. The user interface provides a good medium to link the users with the R sequence analysis packages without having to access the command line prompt.

CONCLUSION

The prototype web application for the basic DNA and protein analysis using R language provides a simple basic analysis of sequences with the addition of information about the methods of the analysis as well as the codes that is being used for the analysis as well. The prototype has several tabs for the information and analysis of the sequences. Users shows preferences for the prototype in terms of user interface, system analysis as well as result production. Accuracy of the prototype can also be proved by the accurate result comparisons with the command line and the online web application tool. Redundancy and time can be reduced as the users do not have to write long lines of codes for the analysis.

ACKNOWLEDGEMENT

Firstly, I would like to thank my student Sivakumar chandran and Prof Dato Gapar for helping this work in a great manner.

REFERENCES

Gibbs, A.J, McIntyre, A.G (1970), The Diagram, a Method for Comparing Sequences : Its Use with Amino Acids and Nucleotide Sequences, European Journal Of Biochemistry, Retreived from http://onlinelibrary.wiley.com/doi/10.1111/j.1432-1033.1970.tb01046.x/pdf

Henke.W, Herdel.K, Jung.K, Schnorr D, Loening, S.A, (1997), Betaine improves the PCR Amplification of GC-rich DNA sequences, Nucleic Acid Research, Vol 25, doi: 10.1093/nar/25.19.3957 Retreived from http://nar.oxfordjournals.org/content/25/19/3957.full.pdf+html

Jinlong,Z.(2011 November 25th), Ice Break of R: An Introduction to R Programming Language, Kadoori Farm and Botanical Garden. Retrived from http://phylodiversity.net/jinlongzhang/uploads/7/5/3/6/7536331/ice_break_r.pdf.

Kawshima.S,Ogata.H, Kanehisa,M (1998), Aaindex: Amino Acid Index Database, Nucleic Acid Research, volume 27, Retreived from http://nar.oxfordjournals.org/content/27/1/368.full.pdf+html

Kim,S., Earnest,L., (2007) , Statistics With R Using Biological Examples (pg 6). Retrived from http://cran.r-project.org/doc/contrib/Seefeld_StatsRBio.pdf

Lobry, J.R, Lobry, C, (1999), Evolution of DNA Base Composition Under No Strand Bias Condition When the Subtitution Rates Are Not Constant, Molecular Biology and Evolution. Retrived from http://mbe.oxfordjournals.org/content/16/6/719.long

Luciano.B, Samuel, K, (2005), Protein lengths in eukaryotic and prokaryotic proteomes, Nucleic Acid Research, doi:10.1093/nar/gki615, Retreived from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1150220/pdf/gki615.pdf

Mehmet,K.,Yohan,K., Umut,T., Shankar,S.,Wojciech,S., Ananth,G. (2006), Pairwise Alignment of Protein Interaction Network, Journal of Computational Biology, volume_13, page 182-199. Retrived from http://engr.case.edu/koyuturk_mehmet/publications/ppi_alignment_jcb.pdf.

Oliver,J.L, Marin, A (1996), A relationship between GC content and coding sequence length, JournalOf Molecular Evolution, Retreived from https://www.google.com.my/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0CCgQFjAB&url=http%3A%2F%2Fwww.researchgate.net%2Fpublication%2F14499424_A_relationship_between_GC_content_and_coding-sequence_length%2Ffile%2F79e4150bb1dd6c5e51.pdf&ei=Mdj3U6-4BNWRuATMo4C4Bw&usg=AFQjCNHw1slEWAe5An8lwGEkfI5vVkMjGg&sig2=9PUabbs8UFbbL5N8p588ew&bvm=bv.73612305,d.c2E

Tina,A.(2014 May 5th), Why the R programming language is good for business,[Weblog

post]. retrived from http://www.fastcolabs.com/3030063/why-the-r-programming-language-is-good-for-business.

Virginia,P.H, Harris,J.O, (1952), The isoelectric point of bacterial cells, Journal Of Bacteriology , 1953, 65(2):198, Retreived from http://jb.asm.org

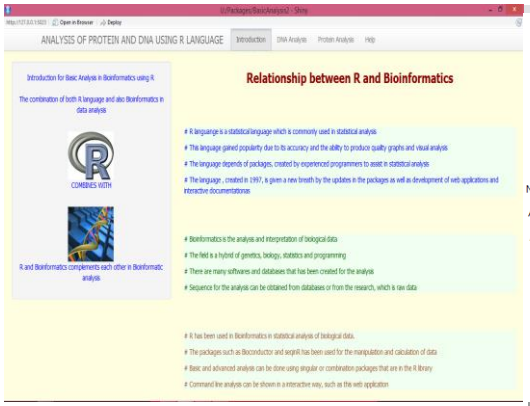
"What are protein and what do they do" (2014), Retrived from http://ghr.nlm.nih.gov/handbook/howgeneswork/protein.

"What is DNA" (2014), Retrived from http://ghr.nlm.nih.gov/handbook/basics/dna.

Widmann.M, Trodler.P, Pleiss,J, (2010), The isoelectric Regions of Proteins: A Systemic Approach, PLOSONE online journal, DOI: 10.1371/journal.pone.0010546,Retreived from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0010546

Zheng H, Wu H, (2010),Gene-centric association analysis for the correlation between the guanine-cytosine content levels and tempreture range conditions of prokaryotic species, BMC Bioinformatics, doi: 10.1186/1471-2105-11-S11-S7, Retreived from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3024870

FIGURES



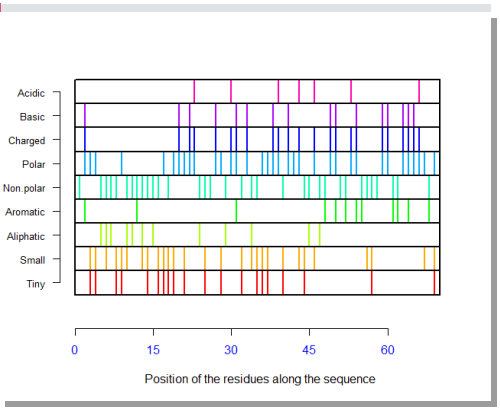
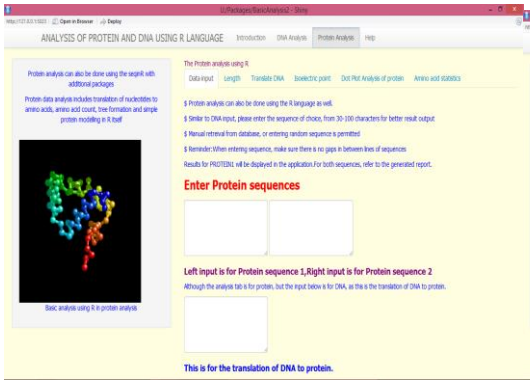


FIGURE 1: INTRODUCTION INTERFACE

FIGURE 2 : AMINO ACID STATISTICS



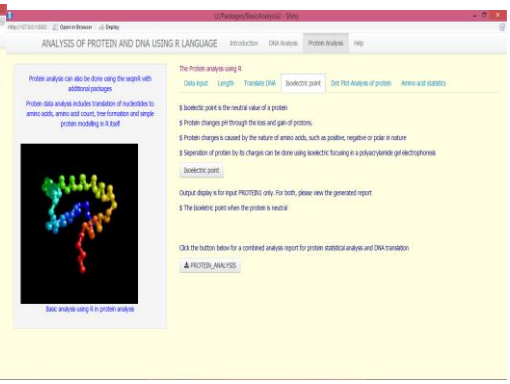


FIGURE 3: DATA INPUT INTERFACE

FIGURE 4 : ANALYSIS METHOD