Introducing 'BioAnalyticaV1.0' : A New Perl Based Software for Biological Sequence Analysis

Pankaj Chetia*, Soumitra Nath, Abhijit Mitra, Khairujjaman Mazumder, Monjur Ahmed Laskar, Manabendra Dutta Choudhury and Gauri Dutta Sharma
Bioinformatics Centre, Assam University, Silchar - 788011, India
*Correspondence; Email: pankaj.chetia@bioinfoaus.ac.in

Abstract

'BioAnalytica Version1.0', (BioAnalyticaV1.0.exe) developed and released in the website www.bioinfoaus.ac.in, is an open source, easy, Perl-based, biological sequence editing software tool, that will accept the input of a Nucleic acid or a Protein sequence, selecting the target file by file browsing and can produce a comprehensive report of various sequence analysis data. It will efficiently execute in silico Transcription, Reverse Transcription, Reverse Complement conversion, Joining of sequences, Total Count of nucleotides, Translation & search for a particular motif that is wide spread in organisms' genome in the target query sequence. And for Protein sequences, it will perform Total Amino Acid Count & search for a particular motif in the target query sequence.

Keywords: BioAnalytica, *In silico*, Perl, Motif

Introduction

With the rapid advancement of Molecular Biological, Genomics & Proteomics research, designing of more and more easier and user friendly sequence analysis softwares has become a very essential part of hardcore Bioinformatics. Such computer-based tools now play an immense critical role in the ongoing biological research all over the world. Designing of such software tools generally needs the logical synchronization of lots of individual programs executing their respective functions. Various programming languages, such as Perl, Java, C, C++, C#, FORTRAN, Java etc. are very extensively being used for developing softwares for Bioinformatics (Gibas *et al.*, 2001). Networks and integration of databases are keys to success in BioInformatics. The technology now exists for global access to databases and softwares (Kolatkar *et al.*, 1998).

Now-a-days, analysis of biological sequences, i.e., DNA, RNA as well as protein sequences has become a routine task for the researchers working in the field of Molecular Evolution that deals with the phylogenetic interactions between species to species, primer designing, vector designing, gene prediction, protein structure prediction etc. Sequencce editing has also got outstanding applications regarding mutation detection in DNA barcoding researches, prognosis of dangerous mutation oriented diseases especially cancer.

The Perl language is used to develop web applications and is frequently used by computational biologists.

Perl programs don't have to be compiled in order to run. Instead, each command in a Perl script is sent to a program called the Perl interpreter, which executes the commands (Schwartz *et al.*, 1997). Perl is an interpreted language; in other words, Perl programs don't have to be compiled in order to run. The language provides powerful text processing facilities without the arbitrary data length limits of many contemporary Unix tools, facilitating easy manipulation of text files.

Perl gained widespread popularity in the late 1990s as a CGI scripting language, in part due to its parsing abilities (Wall *et al.*, 2000). Perl has been used since the early days of the web to write CGI scripts, and is now a component of the popular LAMP (Linux/Apache/MySQL/Perl) platform for web development. Perl has been called "the glue that holds the web together". Large systems written in Perl include Slashdot, and early implementations of Wikipedia and PHP Perl finds many applications as a glue language, tying together systems and interfaces that were not specifically designed to interoperate. Systems administrators use Perl as an all purpose tool; short Perl programs can be entered and run on a single command line (Binny, 2008).

'Perl', that stands for 'Practical Extraction and Reporting Language' is a high-level, general-purpose, interpreted, dynamic programming language. 'Perl' was originally named as "Pearl". It was first developed by Larry Wall in 1987 as a general-purpose Unix scripting language to make report processing easier (Sheppard, 2000). Perl is a popular programming language that's extensively used in areas such as bioinformatics and web programming. Perl has become very popular amongst biologists, because it's so well-suited to several bioinformatics tasks (Tisdall, 2001). It runs on all the operating system such as Windows, Linux, Unix, Macintosh etc. The executable files of any software, developed using Perl language don't need installation in the computer like other softwares.

Perl/Tk (also known as **pTk or Perl Toolkit**) is a collection of modules and code that attempts to wed the easily configured Tk 8 widget toolkit to the powerful lexigraphic, dynamic memory, I/O, and object-oriented capabilities of Perl 5(Binny, 2008). In other words, it is an interpreted scripting language for making widgets and programs with Graphical User Interfaces (GUI).

The sequence editor, **'BioAnalytica Version1.0'** efficiently executes the following programs:

a. Sequence editing by loading the file by file browsing.

b. Transcription, i.e., conversion of a DNA sequence to its corresponding RNA sequence. Here, all the 'T' are replaced by 'U'.

c. Reverse Transcription, i.e., conversion of a RNA sequence back to its corresponding DNA sequence. Here, all the 'U's are replaced by 'T'

d. Conversion of a DNA/RNA sequence into its corresponding Reverse Complement form, i.e., the reverse form of the complementary strand.

e. Concatenating (i.e.,joining) two DNA or RNA sequences which is required in case of vector designing.

f. Total ATGC/AUGC count of the given DNA/RNA sequence.

g. Translation, i.e., conversion of a DNA/RNA sequence to its corresponding PROTEIN sequence.

h. Searching for a particular MOTIF sequence in the query DNA/RNA sequence.

i. Total AMINO ACID/RESIDUE count of the given PROTEIN sequence.

j. Searching for a particular MOTIF sequence in the query PROTEIN sequence.

Methodology

BioAnalyticaV1.0 is a Perl-based, biological sequence editing software tool that was developed and designed in Bioinformatics Centre, Assam University, Silchar, basically to be used by the Molecular Biologists for macromolecular sequence analysis purpose. It was developed in Windows7 Operating System using ActivePerl and Perl Tk modules. Initially, individual programs were written in Active Perl and then the Graphical User Interface (GUI) for the software was developed using Perl Tk. All the modules, that contain supportive files for programs to run, were downloaded from the Internet.

For developing **BioAnalytica V1.0** in Perl Programming Language, the Software 'ActivePerl-5.10.1.1007' was first downloaded from www.activestate.com/activeperl/downloads and was then installed. The packages of all the **Perl/Tk** modules for designing the GUI were then

downloaded from http://www.cpan.org/ and was then installed(Lidie *et al.*, 2002). **MS DOS** was used as command line editor and all the individual perl programs were input and synchronized in Notepad.

Using BioAnalytica

BioAnalyticaV1.0 accepts an input nucleic acid (DNA/ RNA) or protein sequence, which can be easily loaded on the editing screen by browsing the target file saved in certain location in the computer. The sequence can be picked up from a local file or retrieved from NCBI as a GenBank file via its accession number. Two main buttons are available on the software interface; 'NUCLEIC ACIDS' button for working with DNA/ RNA sequence and 'PROTEINS' button for working with amino acid or protein sequences. The 'NUCLEIC ACIDS' button contains 7 sub

buttons viz, **Transcription, Reverse Transcription, Reverse Complement, Join Sequences, Count, Translation & Motif Search** and the '**PROTEINS**' button contains 2 sub buttons, viz, **Count of Residues & Motif Search**.

For every conversion, the result will be displayed in a new Toplevel window and the result sequence can be easily saved in the word format. Two sequences need to be input for joining two sequences or finding the existence of a particular motif in the query nucleic acid or protein sequences.

All the necessary instructions for working with this software are clearly displayed on the screen below the main buttons and for further details, another button is there at the menubar called 'HELP'.

Figure 1: **The Main GUI for the Sequence Editor, 'BioAnalytica Version1.0'**

Figure 2: **The GUI for NUCLEIC ACID TOOL of 'BioAnalytica V1.0'**

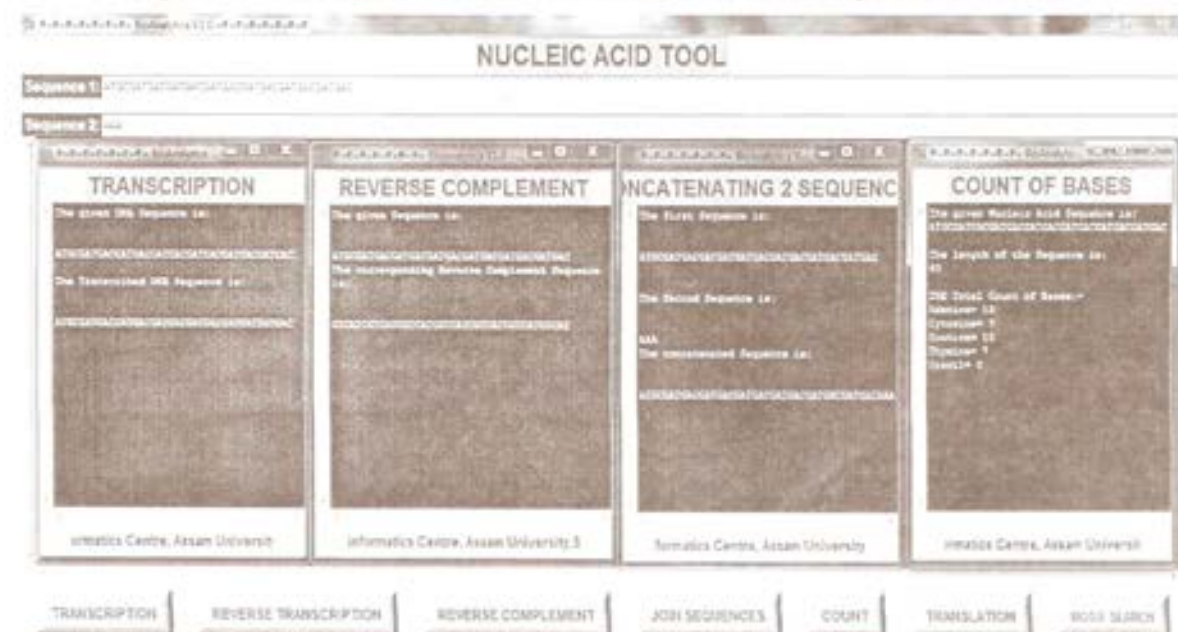Figure 3: The GUI for the PROTEIN TOOL of 'BioAnalytica Version1.0'

Figure 4 : Results of Various Programs Under Nucleic Acid Tool of 'BioAnalytica Version1.0'

Discussion

With respect to the recent trends in Molecular Biological, Genomics and Proteomics researches, development of high throughput macromolecular sequence analysis softwares has become a very

important part on the part of Bioinformaticians. Various open access or commercially available software tools play the key role in performing biological sequence analysis. The newly designed sequence editor **'BioAnalyticaV1.0'** is an attempt to meet the growing demand of easy and user friendly sequence analysis softwares. This open source software has been made available for the users via the website www.bioinfoaus.ac.in in the internet. It can be quite easily downloadable in the computers and being an executable file, it is platform independent. In the upcoming versions of **'BioAnalytica'**, programs for online nucleotide as well as protein BLAST, ORF search etc. would be incorporated.

References

Binny, V. A. (2008). Perl Tk Tutorial.http://www.binco.com/perl/perl_tk_tutorial/perl_tk_tutorial.pdf

Gibas, C. and Jambeck, P.(2001) Developing Bioinformatics Computer Skills. 1st edition. O'Reilly.

Kolatkar,P.R., Sakharkar,K., Tse,C.R., Kiong,B.K., Tan, T.W.(1998) Development of software tools at Bioinformatics Centre (BIC) at the National University of Singapore (NUS).

Lidie, S and Walsh, N.(2002) Mastering Perl/Tk. 1st edition. O'Reilly.

Schwartz, R. and Christiansen, T.(1997)Learning Perl. Introduction to Perl but assumes prior experience with another programming language. 2nd edition. O'Reilly & Associates.

Sheppard, D. (2000-10-16). "Beginner's Introduction to Perl". dev.perl.org. Retrieved 2011-01-08.

Tisdall, J.(2001)Beginning Perl for Bioinformatics. 1st edition. O'Reilly.

Wall,L.,Christiansen,T. and Orwant,J., (July 2000). Programming Perl, Third Edition. O'Reilly. ISBN 0-596-00027-8.