Short Communication

iMedPub Journals
www.imedpub.com

Annals of Clinical and Laboratory Research
ISSN 2386-5180

2018
Vol.6 No.4:265

DOI: 10.21767/2386-5180.100265

Bio-Stamp: Simple Tool for Routine Biological Data Analysis

Peter Tharwat Habib[1,3], Alsamman Mahmoud Alsamman[2] and Aladdin Hamwieh[3*]

[1]College of Biotechnology, Misr University for Science and Technology, 6th of October City, Egypt

[2]Department of Genetics and Breeding, Agricultural Genetic Engineering Research Institute, Giza, Egypt

[3]Depatment of Genetics and Breeding, International Center for Agricultural Research in the Dry Areas (ICARDA), Cairo, Egypt

[*]**Corresponding author:** Aladdin Hamwieh, Researcher, Department of Genetic and Breeding, International Center for Agricultural Research in the Dry Areas (ICARDA), Cairo, Egypt, Tel: + 2001026009094; E-mail: a.hamwieh@cgiar.org

Received Date: November 8, 2018; **Accepted Date:** November 20, 2018; **Published Date:** November 22, 2018

Citation: Habib PT, Alasamma AM, Hamwieh A (2018) Bio-Stamp: Simple Tool for Routine Biological Data Analysis. Ann Clin Lab Res Vol.6 No.4: 265.

Abstract

The massive extension in biological data induced a need for user-friendly bioinformatics tools could be used for routine biological data manipulation. Bio-Stamp is a simple analytical software implements variety of tools to perform common data analysis on different biological data types and databases. Bio-Stamp provide general aspects of data analysis such as handling nucleotide data, fetching different data formats information, NGS quality control, data visualization, performing multiple sequence alignment and sequence BLAST. These tools accept common biological data formats and produce human-readable output files could be stored on local computer machines. Bio-Stamp has a user-friendly graphical user interface to simplify massive biological data analysis and consume less memory and processing power. Bio-Stamp source code was written through Python programming language which provides less memory usage and initial start-up time. Bio-Stamp is free and open source software, where its code could be modified, extended or integrated in different bioinformatics pipelines.

Keywords: Blast; Visualization; Computational biology; Multiple sequence alignment; Dotplot; Restriction site; FASTA; Genbank; Data extraction; Quality control

Introduction

Bioinformatics has evolved and expanded continuously over the past years and has become very important basic demand in life science research. There is an enormous growth of biological data on network and databases due to the massive amount of research done daily. The public databases growth rate is increasing exponentially over years, for example: NCBI Gene database and Protein database, nucleotide database reached 24, 300 and 210 million records in 2016 and have 13.8%, 37.7% and 5.2% annually growth rate, respectively [1].

The biological data analysis and interpretation is getting a major bottleneck in Bioinformatics [2]. In order to extract the target information from different biological data, there are a plethora publicly available analysis tools, which could be used extract, analyse and visualize data. Some of the main differences between these software are availability, GUI user friendliness, visualization methods and performance. Each one of these software requires specific parameters in order to perform analysis or extract information about genes or gene clusters through simple and routine procedure. Many of the available bioinformatics open source tools uses command lines to perform different analysis, others have a Graphical User Interface (GUI) could simplify complex analytical procedures and provide a simple way to enter different parameters.

Tenth of different general-use bioinformatics softwares are publicly available. DNASTAR (Laser gene) is a commercial bioinformatics software that compresses different applications such as gene discovery, genomic visualization, NGS assembly with Sanger validation, primer design, Sanger sequence assembly, sequence alignment and others [3]. CLC workbench is another bioinformatics pipeline provided by QIAGEN company (www.qiagenbioinformatics.com), which provides different data analysis tools such as NGS read mapping, *de novo* assembly, variant analysis.

Assemble of DNA sequence data, multiple alignment sequence and reverse complement. EMBOSS is the European molecular biology open software suite, it integrates existing analytical programming packages and databases more effectively with over 100 applications and has the capability to be run with advanced graphical user interfaces [4].

In this study, we are introducing Bio-Stamp software, which is a bioinformatics tool that compresses simple and common data analysis applications with a user-friendly GUI. Bio-Stamp source code and freely available, where its code could be modified, extended or integrated in different bioinformatics pipelines. Bio-Stamp is a simple analytical software implements variety of tools to performing common data analysis on different biological data types and databases.

Implementation

Bio-Stamp was written using Python programming language (version 3.4+) that provides set of functions and tools to

implement a data analysis, extraction and visualization. Bio-python was used to implement some common applications [5]. An additional python codes were written to provide new other tools. Source code, installer and manual are publicly available at (https://github.com/peterhabib/biostamp).

Features and Applications

Bio-Stamp provide general aspects of data analysis such as handling nucleotide data, fetching different data formats information, NGS quality control, data visualization, performing multiple sequence alignment and sequence BLAST.

Nucleotide tools accept nucleotide sequence(s) or NCBI accessions as an input. These tools provide DNA translation, GC%, reverse complement, transcription, back transcription and open reading frame (ORF) finding. GC% content could be used in transcriptome mapping (HTM) in gene-dense domains with high GC content [6]. DNA translation could be used in protein sequence classification or finding statistically significant functional associations in genomic experimental [7]. Additionally the tool also has options to choose between different translation tables and stopping translation at first stop codon.

Data Extraction can be used to extract specific targeted information from Genbank sequence(s). This tool creates folder that contains text files holds specific user-defined information extracted separately from genomic data. This tool can be used for the exploration biological database depends on Genbank file data of drug discovery [8].

Database tools could be useful in handling specific NCBI accessions in different databases for sequence retrieval in FASTA or Genbank formats. This tool in discovering new mutations is responsible for diseases by comparing different database records for the same gene in specific gene family [9]. On the other hand, BLAST tool could be used to align specific sequence or ID to public NCBI database, in order to discover similar published sequence(s). This option could be helpful the characterization of novel genes belong to the different gene families [10].

Alignment is most daily used tools in bioinformatics to do local, global, Needleman or water nucleotide sequence alignment. Bio-Stamp offer different options to change alignment matrices according BLOSUM for either global or local alignments. Sequence alignment illustrates how different aligned sequences are related to each other, discovering genes with common ancestor or to improve protein secondary structure prediction [11].

Visualization tools draw the massive nucleotide sequence such as chromosome files, illustrating genes/CDS positions. This tool export illustrations in PDF file formats. This tool could be used in positioning genes on chromosome and depicting their rearrangement in different chromosomes [12]. Phylogenetic tree tool can reconstruct phylogenetic tree(s) produced by using different nucleotide sequences, in order to screen there genetic diversity [13]. Dot-plot tool draws graphs between two sequences to show the sequence similarity, which could be used to compare complete genome sequencing data [14]. GC% in visualization tools creates chart represent the GC% content of two FASTA file records, depicting the recombination drives the evolution of GC-content in different genomes [15]. Restriction site tool can build a circular and a linear representation for the position of restriction sites in DNA sequences, which could be helpful in rapid polymorphism identification and genotyping using restriction site associated markers [16]. The weblogo tool illustrates the consensus sequence in given record(s) which reflect the presence of the functional domains in protein such as: active site of or ligand binding site.

Quality control tools deal with FASTAQ files in order to do post-sequencing processing such as primer and adopters trim to prepare the reads for different analysis such genome assembly, mapping or any other application [17]. Also, convert FASTAQ format to FASTA format to allow user to do different analysis such as *de novo* transcript sequence reconstruction from NGS data [18].

References

1. Coordinators NR (2016) Database resources of the national center for biotechnology information. Nucleic Acids Res p: D7.
2. Pavlopoulos GA, Wegener AL, Schneider R (2008) A survey of visualization tools for biological network analysis. Bio Data Min 1(1): 12.
3. Burland TG (2000) DNASTAR's Lasergene sequence analysis software. In Bioinformatics methods and protocols 2000. Humana Press, Totowa, NJ pp: 71-91.
4. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European molecular biology open software suite. Trends Genet 16(6): 276-277.
5. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, et al. (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25(11): 1422-1423.
6. Versteeg R, Van Schaik BD, Van Batenburg MF, Roos M, Monajemi R, et al. (2004) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res 13(9): 1998-2004.
7. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) Panther: A library of protein families and subfamilies indexed by function. Genome Res 13(9): 2129-2141.
8. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, et al. (2006) Drug bank: A comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 34(1): 668-672.
9. Levy GG, Nichols WC, Lian EC, Foroud T, McClintick JN, et al. (2001) Mutations in a member of the ADAMTS gene family cause thrombotic thrombocytopenic purpura. Nature 413(6855): 488.
10. Qu X, Zhai Y, Wei H, Zhang C, Xing G, et al. (2002) Characterization and expression of three novel differentiation-related genes belong to the human NDRG gene family. Mol Cell Biochem 229(1-2): 35-44.

11. Cuff JA, Barton GJ (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 40(3): 502511.
12. Gandhi MS, Stringer JR, Nikiforova MN, Medvedovic M, Nikiforov YE (2009) Gene position within chromosome territories correlates with their involvement in distinct rearrangement types in thyroid cancer cells. Genes, Genes Chromosom Cancer 48(3): 222-228.
13. Van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30(2): 386-394.
14. Loman NJ, Quick J, Simpson JT (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12(8): 733.
15. Meunier J, Duret L (2004) Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol 21(6): 984-990.
16. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17(2): 240-248.
17. Nobuta K, McCormick K, Nakano M, Meyers BC (2010) Bioinformatics analysis of small RNAs in plants using next generation sequencing technologies. In Plant MicroRNAs, Humana Press pp: 89-106.
18. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8(8): 1494.