Advances in Bioscience and Biotechnology, 2019, 10, 33-41
http://www.scirp.org/journal/abb
ISSN Online: 2156-8502
ISSN Print: 2156-8456

BioAnalyzer: Bioinformatic Software of Routinely Used Tools for Analysis of Genomic Data

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

[1]Department of Bioinformatics, College of Biotechnology, Misr University for Science and Technology, 6th of October City, Egypt
[2]Department of Biotechnology, International Center for Agricultural Research in the Dry Areas (ICARDA), Cairo, Egypt
[3]Department of Genome Mapping, Molecular Genetics and Genome Mapping Laboratory, Agricultural Genetic Engineering Research Institute, Giza, Egypt
Email: *a.hamwieh@cgiar.org

How to cite this paper: Habib, P.T., Alsamman, A.M. and Hamwieh, A. (2019) BioAnalyzer: Bioinformatic Software of Routinely Used Tools for Analysis of Genomic Data. *Advances in Bioscience and Biotechnology*, **10**, 33-41.
https://doi.org/10.4236/abb.2019.103003

Received: January 22, 2019
Accepted: March 18, 2019
Published: March 21, 2019

Abstract

The massive extension in biological data induced a need for user-friendly bioinformatics tools could be used for routine biological data manipulation. Bioanalyzer is a simple analytical software implements a variety of tools to perform common data analysis on different biological data types and databases. Bioanalyzer provides 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. Bioanalyzer has a user-friendly graphical user interface to simplify massive biological data analysis and consume less memory and processing power. Bioanalyzer source code was written through Python programming language which provides less memory usage and initial startup time. Bioanalyzer is a free and open source software, where its code could be modified, extended or integrated in different bioinformatics pipelines. Bioinformatics Produce huge data in FASTA and Genbank format which can be used to produce a lot of annotation information which can be done with Python programming language that open the door form bioinformatics tool due to their elasticity in data analysis and simplicity which inspire us to develop new multiple tool software able to manipulate FASTA and Genbank files. The goal Develop new software uses Genomic data files to produce annotated data. Software was written using python programming language and biopython packages.

Keywords

Biopython, Bioinformatics, FASTA, Genbank, Multiple Sequence Alignment, BLAST, NCBI, Data Visualization

1. Introduction

1.1. Bioinformatics Data Set

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].

1.2. User-Friendly Bioinformatics Tools

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 plethora publicly available analysis tools, which could be used extract, analyze and visualize data. Some of the main differences between these softwares are availability, GUI user friendliness, visualization methods and performance. Each one of these softwares 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.

1.3. Similar Multiple Tools Softwares

Tenth of different general-use bioinformatics softwares are publicly available. DNASTAR (Lasergene) is a commercial bioinformatics software that compresses different applications such 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 assembly 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].

1.4. Bioanalyzer

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

2. Materials and Method

Bioanalyzer was developed using python libraries for perform data manipulation and using of Tkinter package to design the interface. About forty module and function from Biopython with integration from open source scripts and our self-wrote scripts.

2.1. Biopython

Biopython is python library for Genomic data analysis and annotation provides plethora on scripts such as: data reading and extracting from FASTA and Genbank files, Multiple Sequence Alignment, BLAST searching against NCBI database and even accessing to the NCBI database itself [5]. We used biopython scripts to create environment for data mining and annotation using scripts to read and manipulate FASTA and Genbank files using SeqIO module to produce annotated data in text format such as Multiple sequence Alignment, open reading frames, BLAST searching or NCBI query search, or in illustrated figures such as in chromosome genes, mRNA or tRNA visualization, enzymes restriction site using matplotlib and network module with integration of biopython modules.

2.2. Matrices and Algorithms for Proteins and Nucleotide Alignment

To maximize the accuracy of protein alignment, PAM and BLOSUM matrices is used for score the accepted mutation and find functional domains.

2.3. Packages Used in Data Visualization

Matplotlib is most sufficient and accurate for data visualization. Matplotlib used in the software draw and visualization of chromosome, restriction site, dotplot graph.

2.4. Tkinter Designing Graphical User Interface

Tkinter library is used to build the GUI that consist of frames, buttons, text boxes etc. tkinter provides availability to link scripts and functions with press of buttons and display the result text on text viewer [6].

2.5. Converting Python to Stand-Alone Executable Application

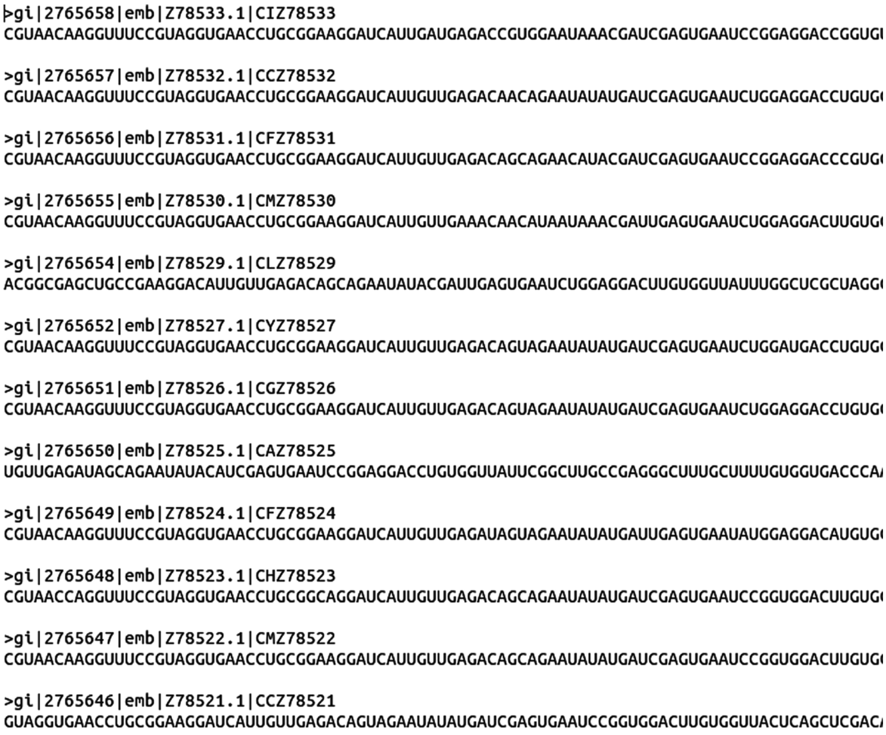
We used pyinstaller (http://www.pyinstaller.org/) to convert python file to standalone executable application. Pyinstaller collect the packages used in the python software and converting them locally installed packages in the directory of the software where the software can retrieve any function from this packages on this directory instead of calling the packages and function on system.

3. Results and Discussion

Bioanalyzer provides 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. The following description of each section of software with sample of results.

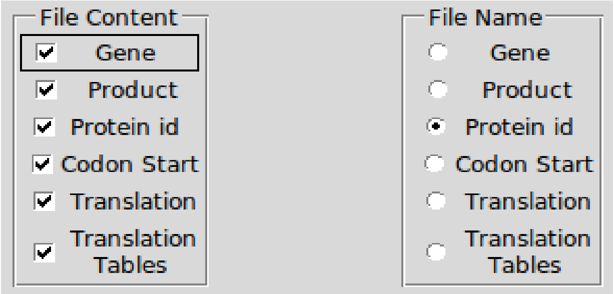
Nucleotide tools accepts nucleotide sequence(s) or NCBI accessions as an input. These tools provide DNA translation, GC%, reverse complement, transcription (**Figure 1**), back transcription and open reading frame (ORF) finding. GC% content could be used in transcriptome mapping (HTM) in gene-dense domains with high GC content [7]. DNA translation could be used in protein sequence classification or finding statistically significant functional associations in genomic experimental [8]. 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 genebank sequence(s) with option of choosing file content and name (**Figure 2**). 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 genebank file data of drug discovery



```
>gi|2765658|emb|Z78533.1|CIZ78533
CGUAACAAGGUUUCCGUAGGUGAACCUGCGGAAGGAUCAUUGAUGAGACCGUGGAAUAAACGAUCGAGUGAAUCCGGAGGACCGGUG

>gi|2765657|emb|Z78532.1|CCZ78532
CGUAACAAGGUUUCCGUAGGUGAACCUGCGGAAGGAUCAUUGUUGAGACAACAGAAUAUAUGAUCGAGUGAAUCUGGAGGACCUGUG

>gi|2765656|emb|Z78531.1|CFZ78531
CGUAACAAGGUUUCCGUAGGUGAACCUGCGGAAGGAUCAUUGUUGAGACAGCAGAACAUACGAUCGAGUGAAUCCGGAGGACCCGUG

>gi|2765655|emb|Z78530.1|CMZ78530
CGUAACAAGGUUUCCGUAGGUGAACCUGCGGAAGGAUCAUUGUUGAAACAACAUAAUAAACGAUUGAGUGAAUCUGGAGGACUUGUG

>gi|2765654|emb|Z78529.1|CLZ78529
ACGGCGAGCUGCCGAAGGACAUUGUUGAGACAGCAGAAUAUACGAUUGAGUGAAUCUGGAGGACUUGUGGUUAUUUGGCUCGCUAGG

>gi|2765652|emb|Z78527.1|CYZ78527
CGUAACAAGGUUUCCGUAGGUGAACCUGCGGAAGGAUCAUUGUUGAGACAGUAGAAUAUAUGAUCGAGUGAAUCUGGAUGACCUGUG

>gi|2765651|emb|Z78526.1|CGZ78526
CGUAACAAGGUUUCCGUAGGUGAACCUGCGGAAGGAUCAUUGUUGAGACAGUAGAAUAUAUGAUCGAGUGAAUCUGGAGGACCUGUG

>gi|2765650|emb|Z78525.1|CAZ78525
UGUUGAGAUAGCAGAAUAUACAUCGAGUGAAUCCGGAGGACCUGUGGUUAUUCGGCUUGCCGAGGGCUUUGCUUUUGUGGUGACCCA

>gi|2765649|emb|Z78524.1|CFZ78524
CGUAACAAGGUUUCCGUAGGUGAACCUGCGGAAGGAUCAUUGUUGAGAUAGUAGAAUAUAUGAUUGAGUGAAUAUGGAGGACAUGUG

>gi|2765648|emb|Z78523.1|CHZ78523
CGUAACCAGGUUUCCGUAGGUGAACCUGCGGCAGGAUCAUUGUUGAGACAGCAGAAUAUAUGAUCGAGUGAAUCCGGUGGACUUGUG

>gi|2765647|emb|Z78522.1|CMZ78522
CGUAACAAGGUUUCCGUAGGUGAACCUGCGGAAGGAUCAUUGUUGAGACAGCAGAAUAUAUGAUCGAGUGAAUCCGGUGGACUUGUG

>gi|2765646|emb|Z78521.1|CCZ78521
GUAGGUGAACCUGCGGAAGGAUCAUUGUUGAGACAGUAGAAUAUAUGAUCGAGUGAAUCCGGUGGACUUGUGGUUACUCAGCUCGAC
```

Figure 1. Multiple FASTA record file transcription generated by transcription tool.



File Content

☑ Gene
☑ Product
☑ Protein id
☑ Codon Start
☑ Translation
☑ Translation Tables

File Name

○ Gene
○ Product
● Protein id
○ Codon Start
○ Translation
○ Translation Tables

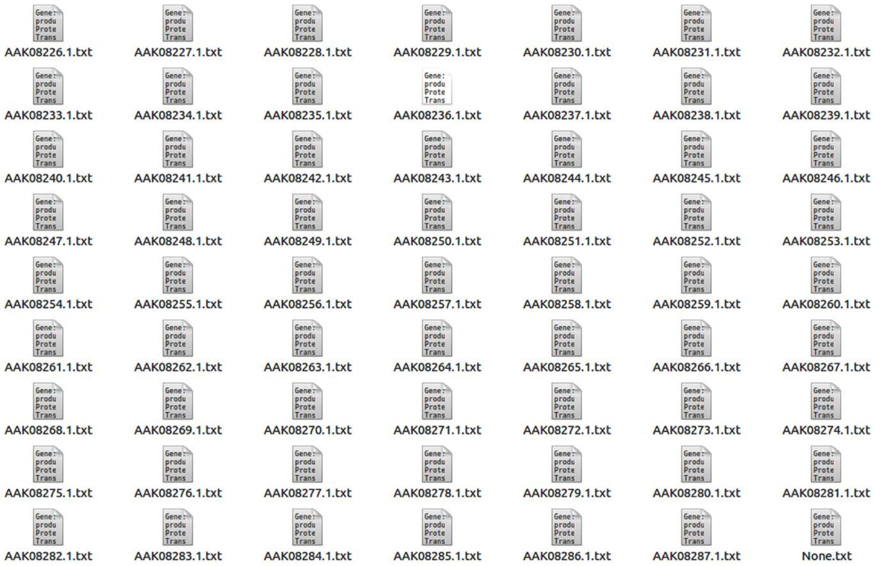
Figure 2. Data extraction tool is used to extract certain information from large genbank file. the left section from the figure is the content of each file, and the right section is the name of each file.

[9]. the resulted file is text format file contain the information according the user choices (**Figure 3**).

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 responsible for diseases by comparing different database records for the same gene in specific gene family [10]. On the other hand, BLAST tool (**Figure 4**) 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 [11].

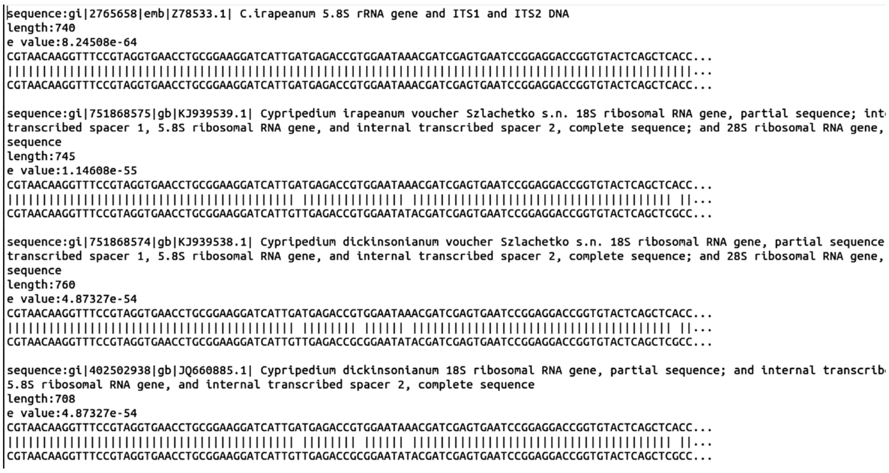
Alignment is most daily used tools in bioinformatics to do local, global, needleman or water nucleotide sequence or protein (**Figure 5**) alignment. Bioanalyzer offer different options to change alignment matrices according BLOSUM for either global or local alignments and the yielded score indicate how far those aligned sequences are similar to each other by giving score to every match, mismatch and gap. Sequence alignment illustrate how different aligned sequences are related to each other, discovering genes with common ancestor or to improve protein secondary structure prediction [12].

Visualization tools draw the massive nucleotide sequence such as chromosome files, illustrating genes/CDS positions (**Figure 6** and **Figure 7**). 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 [13]. Phylogenetic tree tool can reconstruct phylogenetic tree(s) produced by using different nucleotide sequences, in order to screen there genetic diversity [14]. Dotplot tool draws graphs between two sequences to show the sequence similarity, which could be used to compare complete genome sequencing data [15]. 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 [16]. 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 [17].



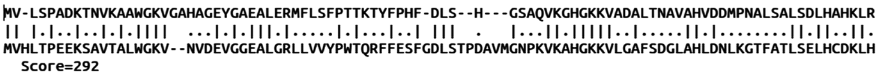
AAK08226.1.txt AAK08227.1.txt AAK08228.1.txt AAK08229.1.txt AAK08230.1.txt AAK08231.1.txt AAK08232.1.txt
AAK08233.1.txt AAK08234.1.txt AAK08235.1.txt AAK08236.1.txt AAK08237.1.txt AAK08238.1.txt AAK08239.1.txt
AAK08240.1.txt AAK08241.1.txt AAK08242.1.txt AAK08243.1.txt AAK08244.1.txt AAK08245.1.txt AAK08246.1.txt
AAK08247.1.txt AAK08248.1.txt AAK08249.1.txt AAK08250.1.txt AAK08251.1.txt AAK08252.1.txt AAK08253.1.txt
AAK08254.1.txt AAK08255.1.txt AAK08256.1.txt AAK08257.1.txt AAK08258.1.txt AAK08259.1.txt AAK08260.1.txt
AAK08261.1.txt AAK08262.1.txt AAK08263.1.txt AAK08264.1.txt AAK08265.1.txt AAK08266.1.txt AAK08267.1.txt
AAK08268.1.txt AAK08269.1.txt AAK08270.1.txt AAK08271.1.txt AAK08272.1.txt AAK08273.1.txt AAK08274.1.txt
AAK08275.1.txt AAK08276.1.txt AAK08277.1.txt AAK08278.1.txt AAK08279.1.txt AAK08280.1.txt AAK08281.1.txt
AAK08282.1.txt AAK08283.1.txt AAK08284.1.txt AAK08285.1.txt AAK08286.1.txt AAK08287.1.txt None.txt

Figure 3. The output files contain the informations after extraction using extraction tool.



```
sequence:gi|2765658|emb|Z78533.1| C.irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA
length:740
e value:8.24508e-64
CGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGATGAGACCGTGGAATAAACGATCGAGTGAATCCGGAGGACCGGTGTACTCAGCTCACC...
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||...
CGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGATGAGACCGTGGAATAAACGATCGAGTGAATCCGGAGGACCGGTGTACTCAGCTCACC...

sequence:gi|751868575|gb|KJ939539.1| Cypripedium irapeanum voucher Szlachetko s.n. 18S ribosomal RNA gene, partial sequence; int
transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene,
sequence
length:745
e value:1.14608e-55
CGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGATGAGACCGTGGAATAAACGATCGAGTGAATCCGGAGGACCGGTGTACTCAGCTCACC...
|||||||||||||||||||||||||||||||||||||||||| ||||||||||||||| |||||||||||||||||||||||||||||||||||||| ||...
CGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGTTGAGACCGTGGAATATACGATCGAGTGAATCCGGAGGACCGGTGTACTCAGCTCGCC...

sequence:gi|751868574|gb|KJ939538.1| Cypripedium dickinsonianum voucher Szlachetko s.n. 18S ribosomal RNA gene, partial sequence
transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene,
sequence
length:760
e value:4.87327e-54
CGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGATGAGACCGTGGAATAAACGATCGAGTGAATCCGGAGGACCGGTGTACTCAGCTCACC...
|||||||||||||||||||||||||||||||||||||||||| |||||||| |||||| |||||||||||||||||||||||||||||||||||||| ||...
CGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGTTGAGACCGCGGAATATACGATCGAGTGAATCCGGAGGACCGGTGTACTCAGCTCGCC...

sequence:gi|402502938|gb|JQ660885.1| Cypripedium dickinsonianum 18S ribosomal RNA gene, partial sequence; and internal transcrib
5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence
length:708
e value:4.87327e-54
CGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGATGAGACCGTGGAATAAACGATCGAGTGAATCCGGAGGACCGGTGTACTCAGCTCACC...
|||||||||||||||||||||||||||||||||||||||||| |||||||| |||||| |||||||||||||||||||||||||||||||||||||| ||...
CGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGTTGAGACCGCGGAATATACGATCGAGTGAATCCGGAGGACCGGTGTACTCAGCTCGCC...
```

Figure 4. BLAST result contain the sequence ID, description, length, e-value and sequence with alignment.



```
MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS--H---GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR
|| |.|..|..|.|.||||  ...|.|.|||.|.....|.|...|..| |||  .   |...||.|||||..|.....||.|........||.||..||.
MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH
  Score=292
```

Figure 5. The alignment two different protein can be produced by PAM or BLOSUM tools.

The weblogo tool illustrates the 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 (**Figure 8**).

Quality control tools deal with FASTAQ files in order to do post-sequencing processing such as primer and adopters trimming to prepare the reads for different analysis such genome assembly, mapping or any other application [18]. Also, convert FASTAQ format to FASTA format to allow user to do different analysis such as *de novo* transcript sequence reconstruction from NGS data [19].

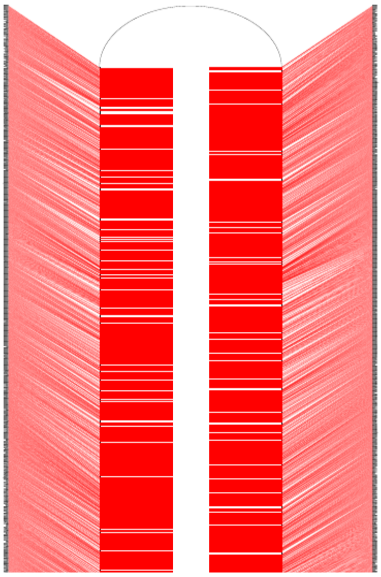


Figure 6. The form of chromosome produced by chromosome tool illustrating each position of gene on chromosome (zoom out).

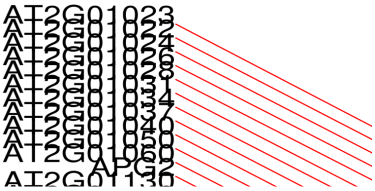


Figure 7. Genes ID generated by chromosome visualization tool showing each gene ID position on chromosome (magnification may be different due to the density of genes on chromosome, in this figure 500%).

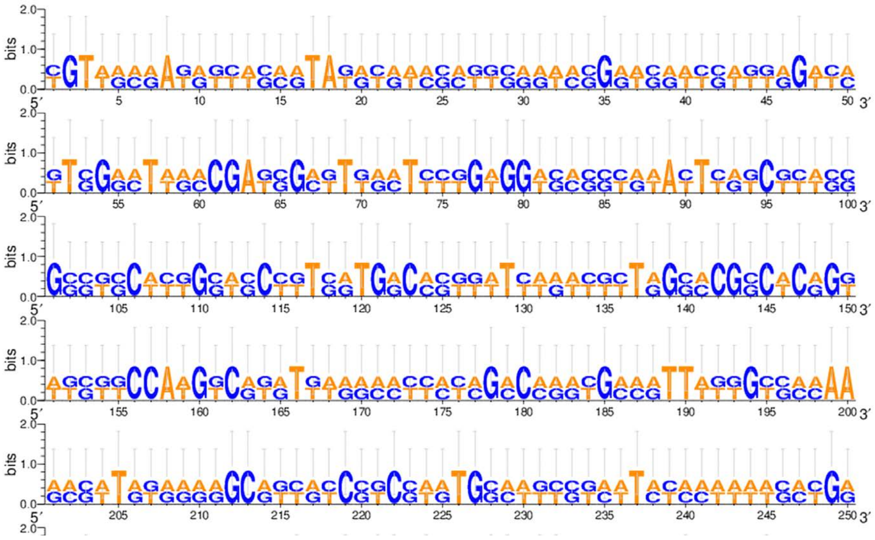


Figure 8. Weblogo tool construct graph represent the consensus between sequences in given multiple fasta file.

4. Conclusion

Bioanalyzer was written using Python programming language (version 3.4+) that provides set of new functions, new tools and already available tools with minor edition in order to improve its functionality and presenting the output in more ordered way to implement a data analysis, extraction and visualization all gathered in one software.

Data Availability

Bioanalyzer was written using Python programming language (version 3.4+) that provides set of functions and tools to implement a data analysis, extraction and visualization. An additional python codes were written to provide new other tools. source code, installer and manual are publicly available at (http://www.ageri.sci.eg/index.php/facilities-services/ageri-softwares/bioanalyzer or https://github.peterhabib/com/bioanalyzer).

Funding Statement

Research is funded by the corresponding author Aladdin Hamweih, senior scientist, Department of Biotechnology at International Center for Agricultural Research in the Dry Areas (ICARDA).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] Coordinators, N.R. (2017) Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **45**, D12. https://doi.org/10.1093/nar/gkw1071

[2] Pavlopoulos, G.A., Wegener, A.L. and Schneider, R. (2008) A Survey of Visualization Tools for Biological Network Analysis. *Biodata Mining*, **1**, 12. https://doi.org/10.1186/1756-0381-1-12

[3] Burland, T.G. (2000) DNASTAR's Lasergene Sequence Analysis Software. *Bioinformatics Methods and Protocols*, Humana Press, Totowa, 71-91.

[4] Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 276-277. https://doi.org/10.1016/S0168-9525(00)02024-2

[5] Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. and De Hoon, M.J. (2009) Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics*, **25**, 1422-1423. https://doi.org/10.1093/bioinformatics/btp163

[6] Shipman, J.W. (2013) Tkinter 8.4 Reference: A GUI for Python. New Mexico Tech Computer Center.

[7] Versteeg, R., van Schaik, B.D., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J. and van Kampen, A.H. (2003) 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 Research*, **13**, 1998-2004. https://doi.org/10.1101/gr.1649303

[8] Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. and Narechania, A. (2003) PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Research*, **13**, 2129-2141. https://doi.org/10.1101/gr.772403

[9] Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z. and Woolsey, J. (2006) DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Research*, **34**, D668-D672. https://doi.org/10.1093/nar/gkj067

[10] Levy, G.G., Nichols, W.C., Lian, E.C., Foroud, T., McClintick, J.N., McGee, B.M., Yang, A.Y., Siemieniak, D.R., Stark, K.R., Gruppo, R. and Sarode, R. (2001) Mutations in a Member of the ADAMTS Gene Family Cause Thrombotic Thrombocytopenic Purpura. *Nature*, **413**, 488. https://doi.org/10.1038/35097008

[11] Qu, X., Zhai, Y., Wei, H., Zhang, C., Xing, G., Yu, Y. and He, F. (2002) Characterization and Expression of Three Novel Differentiation-Related Genes Belong to the Human NDRG Gene Family. *Molecular and Cellular Biochemistry*, **229**, 35-44. https://doi.org/10.1023/A:1017934810825

[12] Cuff, J.A. and Barton, G.J. (2000) Application of Multiple Sequence Alignment Profiles to Improve Protein Secondary Structure Prediction. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **40**, 502-511. https://doi.org/10.1002/1097-0134(20000815)40:3<502::AID-PROT170>3.0.CO;2-Q

[13] Gandhi, M.S., Stringer, J.R., Nikiforova, M.N., Medvedovic, M. and Nikiforov, Y.E. (2009) Gene Position within Chromosome Territories Correlates with Their Involvement in Distinct Rearrangement Types in Thyroid Cancer Cells. *Genes*, *Chromosomes and Cancer*, **48**, 222-228. https://doi.org/10.1002/gcc.20639

[14] Van Oven, M. and Kayser, M. (2009) Updated Comprehensive Phylogenetic Tree of Global Human Mitochondrial DNA Variation. *Human Mutation*, **30**, E386-E394. https://doi.org/10.1002/humu.20921

[15] Loman, N.J., Quick, J. and Simpson, J.T. (2015) A Complete Bacterial Genome Assembled *de Novo* Using Only Nanopore Sequencing Data. *Nature Methods*, **12**, 733. https://doi.org/10.1038/nmeth.3444

[16] Meunier, J. and Duret, L. (2004) Recombination Drives the Evolution of GC-Content in the Human Genome. *Molecular Biology and Evolution*, **21**, 984-990. https://doi.org/10.1093/molbev/msh070

[17] Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A. and Johnson, E.A. (2007) Rapid and Cost-Effective Polymorphism Identification and Genotyping Using Restriction Site Associated DNA (RAD) Markers. *Genome Research*, **17**, 240-248. https://doi.org/10.1101/gr.5681207

[18] Nobuta, K., McCormick, K., Nakano, M. and Meyers, B.C. (2010) Bioinformatics Analysis of Small RNAs in Plants Using Next Generation Sequencing Technologies. *In Plant MicroRNAs*, Humana Press, 89-106. https://doi.org/10.1007/978-1-60327-005-2_7

[19] Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M. and MacManes, M.D. (2013) *De Novo* Transcript Sequence Reconstruction from RNA-seq Using the Trinity Platform for Reference Generation and Analysis. *Nature Protocols*, **8**, 1494. https://doi.org/10.1038/nprot.2013.084