Bioanalyzer: An Efficient Tool for Sequence Retrieval, Analysis and Manipulation

Hassan Tariq[1*], Shahid Nadeem[1], Sobia Idrees[1], Aisha Yousaf[1], Muhammad Sohail Raza[1], Tariq Niaz[2], Muhammad Ibrahim Rajoka[1]

[1]Department of Bioinformatics and Biotechnology
Government College University, Faisalabad, Pakistan
E-mail: hassantariq9@hotmail.com

[2]Directorate of Entomology
Ayub Agriculture Research Institute, Faisalabad, Pakistan

[*]Corresponding author

Received: October 16, 2010 **Accepted: January 5, 2011**

Published: January 31, 2011

Abstract: *Bioanalyzer provides combination of tools that are never assembled together. Software has list of tools that can be important for different researchers. The aim to develop this kind of software is to provide unique set of tools at one platform in a more efficient and better way than the software or web tools available. It is stand-alone application so it can save time and effort to locate individual tools on net. Flexible design has made it easy to expand it in future. We will make it available publicly soon.*

Keywords: *Analysis, Retrieval, Software.*

Introduction

One of the underpinning aims of the Human Genome Project is to provide the resources to support genetic analysis of human conditions and disorders. This aim is incomplete in part because neither the finished DNA sequence of the whole of the human genome has yet been established, nor has the task of identifying all of the genes within even the available DNA sequence been completed [5]. The rapid increase in large-scale sequencing and the challenges of the post-genomic era lead to a need for the rapid development of new applications, or the enhancement of existing software. This has been limited by the need for a general purpose framework for the academic development of sequence analysis software.

Computational analysis of biological sequences has become an extremely rich field of modern science and a highly interdisciplinary area, where statistical and algorithmic methods play a key role. In particular, sequence alignment tools have been at the hearth of this field for nearly 50 years [3].

Bioinformatics has developed into a full-blown discipline far beyond the scope of a review such as this. Bioinformatics is central to the interpretation of sequence data and to the generation of testable hypotheses arising from such data [4].

Such evidence indicates that bioinformatics resources are now fundamental tools in basic research and are increasingly being used in the training of biologists and in clinical medicine. However, given the burgeoning array of molecular biology databases as well as data retrieval and analysis tools, users are challenged daily to identify the resources that best fit their needs

and to use them effectively. This raises questions about the demographics of bioinformatics users, their needs, and libraries' roles in meeting those needs [2].

To deal with sequence analysis tasks in Bioinformatics, a sequence analysis package is to be introduced. The main objectives of our project are:

1. User and Researcher can retrieve and perform analysis at one platform;
2. to provide maximum sequence analysis tools regarding bioinformatics in one package;
3. to provide user friendly interface so a novice user can also use the tools and can have benefits;
4. to provide platform independent application so can be accessed on any operating system.

Material and method

We used the latest versions of Java and BioJava for our work. We developed this using the object oriented mode of programming. The modules used in the current project are regarding to sequence access, analysis, manipulation and proteomics.

Results

The detailed architecture of Bioanalyzer is very simple to use and easy to get started with as shown in Fig 1.

Database Access Lab

Accessing sequence from NCBI: Besides building your very own database-driven sequence repository, most users will need to fetch sequences from public data sources. A primary source of sequence information is NCBI. From its very beginning, Biojava is able to get sequences from NCBI with wrapper objects and methods. Most recently, the implementation of the Biojavax extension brought forth some changes (for example, namespaces) and the corresponding objects and methods were modified accordingly. To use this feature internet connection is needed as it is accessing sequences from NCBI.

Accessing PDB file: BioJava provides a PDB file parser that reads the content of a PDB file into a flexible data model for managing protein structural data. It is possible to parse individual PDB files from PDB. User would click the button Search and the retrieved file will be shown in the output Text Area. To use this feature internet connection is also a pre-requisite.

Sequence Analysis Lab

Replication, transcription and translation: The software provides the feature of replication, transcription and translation of the input sequences. Thus it covers central dogma of molecular biology along with other features.

Complement and reverse complement: In molecular biology, complementarity is a property of double-stranded nucleic acids such as DNA and RNA as well as DNA: RNA duplexes. Each strand is complementary to the other in that the base pairs between them are non-covalently connected via two or three hydrogen bonds. The software also provides the facility to do this.

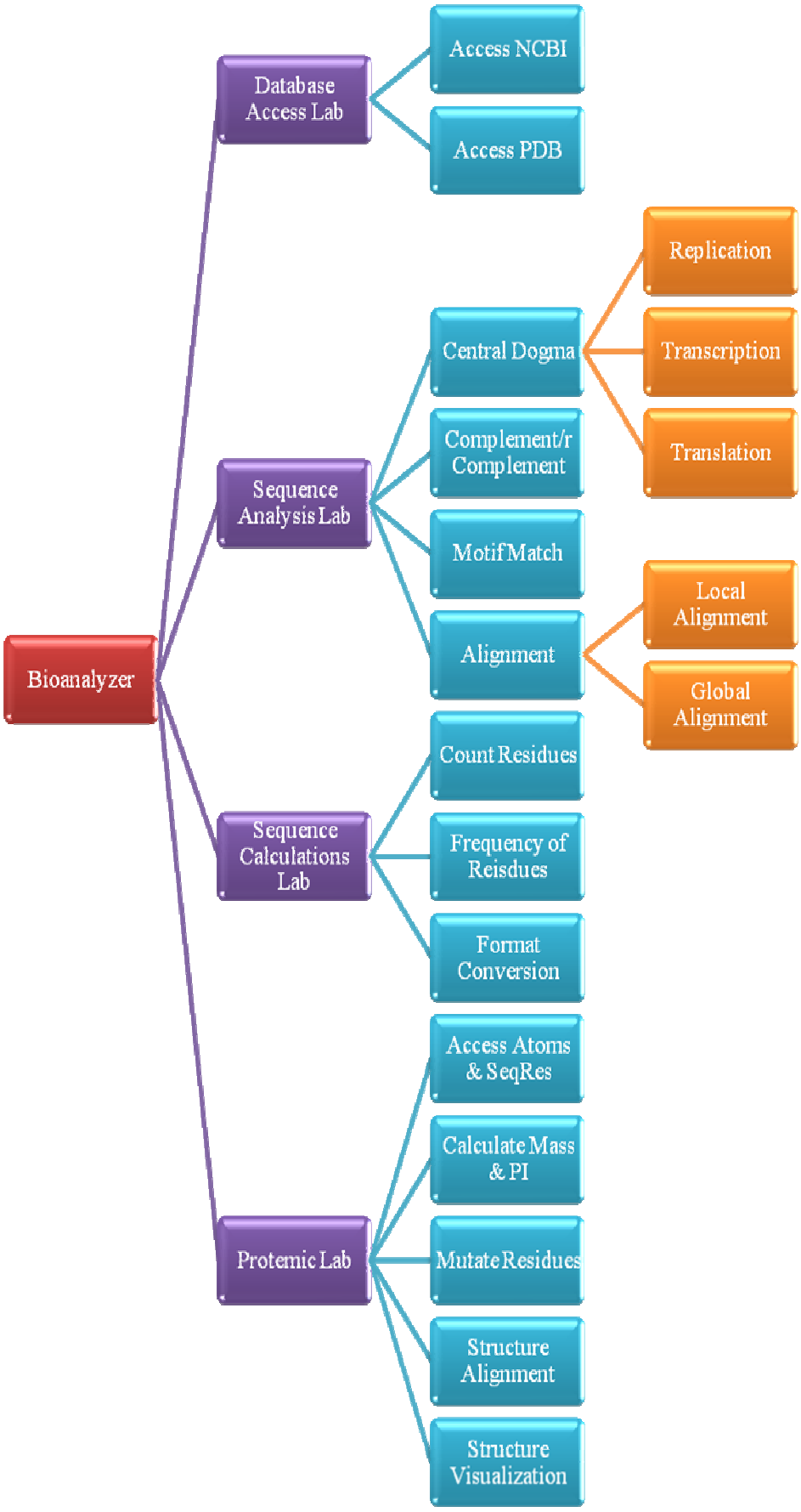

Fig. 1 The detail architecture of Bioanalyzer

Alignment: A very common task in bioinformatics is the alignment of two sequences also known as a "pair-wise alignment". Two common algorithms to generate pairwise alignments are the Needleman-Wunsch and Smith-Waterman algorithms which generate global and local alignments respectively. This feature would align in both Local and Global alignment. User would click the Alignment button and then can choose between Local and Global alignment option. The resulted alignment will be shown in the output Text Area.

Motif match: A sequence motif is a nucleotide or amino-acid sequence pattern that is widespread and has, or is conjectured to have, a biological significance. For proteins, a sequence motif is distinguished from a structural motif, a motif formed by the three dimensional arrangement of amino acids, which may not be adjacent. The software also provides a very good feature of finding motifs as well.

Proteomic Lab

Calculate mass and pl of peptide: In a proteomics project it is important to know what the approximate mass and pI of any putative gene is. You can give any sequence of you choice and calculations will be made with a blink of eye.

Accessing atoms in a structure: This feature can access the atoms in a structure.

PDB aligner: This feature is used for calculating a distance matrix based rigid body protein structure superimposition.

Mutate residues: This feature is used for mutating PDB structure file.

Structure visualization: Jmol is a popular open source 3D viewer written in Java. This feature can be used e.g. to visualize a protein structure.

Sequence Calculations Lab

Count residues: The software counting the residues in a Sequence in a fairly standard bioinformatics task.

Frequency of the symbol: This feature will calculate the frequency of symbol in a sequence.

Format conversion: Format conversion tools is used to convert all sequence file formats into other e.g., FASTA->EMBL or vice versa.

Discussion

Our main objective in developing Bioanalyzer was to provide a local tool for genetic analysis, but clearly, such studies can contribute to whole-genome annotation efforts. Bioanalyzer shares many features with other systems for the automated annotation of genomic sequence.

The main distinguishing features of Bioanalyzer as we see them are as follows:

1. Sequence Analysis Lab provides various analysis tools like basic central dogma analysis tools and local and global alignment;
2. Sequence Manipulation Lab provides various tools like format conversion that can convert any format into other;
3. Proteomic Lab provides protein structure simulator tool to perform analysis for large set of protein structures;

4. Bioanalyzer has a simple Graphic User Interface (GUI) in which user can provide information and presses submit button and resulted information can be visualized in output area. It provides detail description of the tools available to enhance the knowledge of the user about the usage of the system.

The GenBank sequence database incorporates DNA sequences from all available public sources, primarily through the direct submission of sequence data from individual laboratories and from large-scale sequencing projects [1]. Database Access Lab has feature of retrieving sequence and its information directly from NCBI server and it can access PDB file from PDB database.

Acknowledgments

The development of Bioanalyzer is supported by Department of Bioinformatics and Biotechnology, Government College University, Faisalabad, Pakistan. We are thankful to Dr. Shahid Nadeem who not only helped in laying the foundations of Database and Software Engineering Research Group in the Department of Bioinformatics and Biotechnology but also actively leading the Group.

References

1. Dennis A. B., S. B. Mark, J. L. David, O. James, O. Francis, A. R. Barbara, L. W. David (1998). GenBank, Nucl Acids Res, 27(1), 12-17.
2. Geer R. C. (2006). Broad Issues to Consider for Library Involvement in Bioinformatics, J Med Libr Assoc, 94(3), 286-298.
3. Giancarlo R., A. Siragusa, E. Siragusa, F. Utro (2007). A Basic Analysis Toolkit for Biological Sequences, Algo Mol Biol, 2(10), 404-406.
4. Hutchison A. C. (2007). DNA Sequencing: Bench to Bedside and Beyond, Nucl Acids Res, 35(18), 6227-6237.
5. Lander E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh (2001). Initial Sequencing and Analysis of the Human Genome, Nature, 409, 860-921.

Hassan Tariq, M.Sc. in Bioimformatics
E-mail: hassantariq9@yahoo.com

Hassan Tariq graduated from the Government College University – Faisalabad in 2008 with the distinction of securing second highest percentage in the session. Since 2009 he is a working as a Research Officer in the Department of Bioinformatics and Biotechnology. He is the founder of Database and Software Engineering Research Group in the Department of Bioinformatics and Biotechnology. He is also an active member of teaching Faculty of the department, teaching some of the interesting subjects like Bioinformatics Software Development, also publishing textbooks. Scientific interests: Biological Databases, Data Mining, Software Engineering in Bioinformatics, Web Engineering.

Shahid Nadeem, Ph.D.
E-mail:snadeem63@yahoo.com

Dr. Shahid Nadeem got his Ph.D. from the Punjab University, Lahore in 2002. He is working as a senior scientist in NIAB, Faisalabad. In 2009 he joined the Department of Bioinformatics and Biotechnology as a Chairman. He is among the founders of Department of Bioinformatics and Biotechnology in GC University Faisalabad. He is also the founder of Database and Software Engineering Research Group in the Department of Bioinformatics and Biotechnology. He is also an active member of teaching Faculty of the department, teaching some of the interesting subjects like Biotechnology and its applications, Social, Ethical aspects of Biotechnology also publishing textbooks. Scientific interests: Biotechnology, Plant breeding and Genetics, Microbiology.

Sobia Idrees
E-mail: sobia_binm@live.com

Sobia Idrees graduated from the Government College University Faisalabad in 2010. She is working as a dedicated member of Database and Software Engineering Research Group at Government College University Faisalabad. She holds the position of an enthusiastic software engineer in research group and working on many ongoing research projects held by the Group. Based on her research project she obtained 1st position in the software competition at national level. She has worked as active student member of IEEE-GCUF for the year 2008 and organized and participated in many events. She has interests in Software Engineering, Web and Database Development, Bioinformatics Sequence Analysis and in various programming languages like C++, VB.Net, Java, BioJava, Perl, Bioperl and PHP.

Aisha Yousaf
E-mail: aishayousaf@live.com

Aisha Yousaf is an active member of Database and Software Engineering Research Group at Government College University Faisalabad Pakistan. She graduated from the Government College University Faisalabad in 2010. In research group she is working on ongoing research projects. She has command over many Bioinformatics data analysis and structure prediction tools. Her interests are software development, drug designing, phylogenetic analysis and protein structure prediction.

Muhammad Sohail Raza
E-mail: muhammadsohailraza@live.com

Muhammad Sohail Raza is a dedicated member of Database and Software Engineering Research Group at Government College University Faisalabad Pakistan. He graduated from Government College University in 2010 and has the distinction of being the student with second highest percentage in the session. In Research group, he works as a database Developer and is working as an active participant in ongoing research projects. He has command over many Bioinformatics tools and software. He also works on various programming languages like java, Perl, BioPerl and Python. His core interests are Biological Database development, Bioinformatics software Engineering, Data mining, Drug discovery and Computer Aided Drug Designing.

Tariq Niaz, M.Sc. in Entomology
E-mail: tniaz08@hotmail.com

Tariq Niaz graduated from the University of Agriculture – Faisalabad in 1979 with the distinction of securing third highest percentage in the session. He is serving Ayub Agriculture Research Institute, Faisalabad since 1980 as an entomologist. He has also supervised several thesis of MSc and M.Phil student. He has published 40 research papers and review articles in International and national peer review journals and proceedings of conferences/workshops. Scientific interests: Plant protection, Bioassay of pesticides, Biological databases, Bioinformatics, Software development, Web engineering.

Muhammad Ibrahim Rajoka, Ph.D. in Biochemistry
E-mail: mibrahimrajoka47@gmail.com

M. I. Rajoka graduated from the University of New South Wales – Australia in 1981 with the distinction of securing second highest percentage in the session. He did Ph.D. in Biochemistry from University of the Punjab in 1990. After serving 41 years in Nuclear Institute for Agriculture and Biology – Faisalabad (19 years) and National Institute for Biotechnology and Genetic Engineering – Faisalabad (22 years), he joined as Professor in the Department of Bioinformatics and Biotechnology. He has supervised 25, 21 and 7 M.Sc., M.Phil. and Ph.D. students respectively and has published 140 research papers and review articles in International and national peer review journals and proceedings of conferences/workshops. He is also an active member of teaching Faculty of the department, teaching bioprocess technology, research methods in biological sciences, technical writing, communication skills and guiding faculty in writing research proposals. Scientific interests: Industrial strain development through conventional breeding and rDNA technology, bioprocess engineering.