Research Article

# Sequencia: The Primary Sequence Analysis Offline Tool

**Pratyoosh Tripathi*[1,2], K. Ganesan[1], Vaibhav Khandale[2], Sapana Mehendale[2]**
[1]SRM University, SRM Nagar, Kattankulathur, Kancheepuram District, Tamilnadu, India.
[2]RASA Life Science Informatics, 301, Dhanashree Apartments, opp Chitaranjan Vatika, Model colony, Shivaji Nagar, Pune, MH, India.
***Corresponding author's E-mail:** pratyooshshankar@gmail.com

**Accepted on:** 14-04-2015; **Finalized on:** 31-08-2015.

**ABSTRACT**

The current work is focused on the Software Development of an Offline Tool for "PRIMARY-SEQUENCE ANALYSIS" with JAVA and open source resources. SEQUENCIA Tool is an offline Tool of Primary Sequence Analysis, which is quite prevalent Topic for Researchers all over the world. Sequence Name, Sequence Length, Absorbance, Net charge, Iso electric charge, Amino acid composition, Amino acid classification, Aliphatic Index, Instability Index, Average Hydropathy etc are the Primary Sequence Analysis related Attributes for which this Tool exist. This work includes all the Physiochemical Properties- related to Primary Sequence Analysis under a common platform. JAVA, BIOJAVA were used under Platform Independent architecture. The Tool includes Properties of being offline where result can be stored in Text Format, here we can paste more than one sequence and also upload FASTA file to compute Parameters. SEQUENCIA tool may help researchers in carrying out their research and may lead to some useful outcome. For the Future Prospective I will design the SEQUENCIA Primary Sequence Analysis Offline tool for Android Application term under Updated version. The basic aim for this Android Application is because we should not always depend on computers for our tool to run.

**Keywords:** Biojava, Corejava, JAVA swing, Extinction Coefficient, Average Hydropathy, Instability Index.

## INTRODUCTION

The Primary Sequence Analysis term is always a fundamental topic of Research in the Bioinformatics World. The Primary Sequence Analysis originated from the Primary Structure of Protein concept. The Protein is the basic entity of Research in bioinformatics world. Protein sequence to structure and then to function and then to Homology Modeling and then to Drug Design is the dogma of Bioinformatics between which all types of research and novel Discussions take place that how to get specific result at all levels so to get specific drugs and to eventually eradicate specific diseases and to serve Mankind. The Protein structure is bimolecular. They are Polymers. They are specifically polypeptide sequences made of L-α AMINO ACIDS. Each unit of Protein can be called as Amino acid residue. The reason behind the term "Residue" because each amino acid residue form protein by losing its Water Molecule. Peptide and Protein are two different aspects. A chain of 40 residues is called Peptide.[1-4]

For the Biological Functions Protein is folded into one or more specific conformations. The Non covalent interactions which help in Protein conformations are Van der wall interaction, hydrophobic packaging, Hydrogen bonding, and Ionic interactions. To determine the functions of Protein at Molecular Level, the 3D Structure of protein is studied.[5]

The Tool "SEQUENCIA" is offline tool for Primary Structure Analysis of Protein .When the Primary Structure of Protein is discussed the main factor to know that the Primary Structure of Protein refers to linear sequence of Amino acid in the polypeptide chain. The Primary Structure is held together by the Covalent bonds as peptide bonds. Peptide bonds are formed during the process of Protein biosynthesis or translation. The ends of polypeptide chains are termed as carboxyl terminus and the other end as amino terminus. That is based on the end of free group at each extremity. The Primary Structure of the Protein is determined by the Gene corresponding to Protein. The sequence of Protein is unique to that Protein and that specific sequence of particular protein define the structure and function of that protein.

The Tool "SEQUENCIA" is offline Tool for the Primary Structure Analysis. SEQUENCIA will compute all the Physico-chemical Properties which is essential in Analysis of Primary Structure of Protein. The "SEQUENCIA" is first offline tool for Primary Sequence Analysis of Protein.

"SEQUENCIA" is tool which is part of Biosoftware Development using Biojava and the Core java. The Advantage of "SEQUENCIA" is that here we can paste more than one sequence and also we can give the FASTA file to compute the Parameters.

The Result of the SEQUENCIA can be stored in the workspace area as Text format. The Parameters are the Molecular Weight, Absorbance, Instability Index , Average Hydropathy, Sequence Name, Sequence Length, Amino Acid Composition, Amino Acid classification, Extinction Coefficient, Aliphatic Index, Net charge, Isoelectric point, Chemical nature of Protein, The nature of charge on Protein is also mentioned in the Tool. The Tool will be saved on Google Drive or central repository. Any

Researcher can download and use it. Thus under one roof researcher can get all the Properties computed and Analysis work can be more Organized and Easy.

**Description of Physio-Chemical Property**

***Absorbance***

Absorbance is proteins measured in water at 280 nm wavelength. Absorption at the 280nm is used for detection and quantification of purified proteins. The absorbance of each protein depends on the number and positions of its amino acid residues.

***Molecular Weight***

The Average molecular weight (a dimensionless quantity defined as the ratio of the particle mass to 1/12 of the mass of a 12C-atom and symbolized M) of a standard amino acid is nearer to 128. When an amino acid participates in the formation of a polypeptide one molecule of water (MW=18) is removed during peptide bond formation.[1]

***Iso electric point***

Because Amino acids contain ionisable groups, the predominant ionic form of these molecules in solution depends on the pH. Titration of the amino acid illustrates the effect of pH on amino acid structure. Consider alanine, a simple amino acid, which has two titrable groups(α-amino and α-carboxyl group). During titration with a strong base such as NaOH, alanine loses two protons in a stepwise fashion. In a strongly acidic solution, alanine is present mainly in the form in which the carboxyl group is uncharged. Under this condition the molecule net charge is +1, since the ammonium group is protonated. However, an increase in the pH results in the deprotonation of α-carboxyl group. At the point, alanine has no net charge and is electrically neutral. The pH at which this occurs is called the isoelectric point (pI).[6]

***Net Charge***

"Net charge is defined most generally as the sum over all charged groups that are covalently or tightly associated with a protein"[7]. Amino acid with ionizable side chains has more complex titration curves. Glutamic acid, for example, has a carboxyl side chain group. At the low pH glutamic acid has net charge +1. As the base is added (pH increases), the α-carboxyl group loses a proton to become a carboxylate group (in a polyprotic acid, the protons are first lost from the group with the lowest pKa). Glutamate now has no net charge. As still more base is added, the second carboxyl group loses a proton, and the molecule has a -1charge. Adding base further results in loss of protons from ammonium ions. At this glutamate has a net charge of -2. Thus net charge is the overall charge on the amino acid.

***Instability Index***

The instability index provides an estimate of the stability of your protein in a test tube. Statistical analysis of 12 unstable and 32 stable proteins has revealed that there are certain di peptides, the occurrence of which is significantly different in the unstable proteins compared with those in the stable ones.

The authors of this method have assigned a weight value of instability to each of the 400 different di peptides (DIWV). Using these weight values it is possible to compute an instability index (II) which is defined as:

A protein whose instability index is smaller than 40 is predicted as stable, a value above 40 predicts that the protein may be unstable.[5]

***Aliphatic Index***

The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular proteins.[5]

**Amino Acid Composition**

Amino acid composition is defined by the amino acid and their residue count and the calculated concentration of the respective amino acid.

**Amino Acid Classification**

***Uncharged Polar Amino Acid***

Six amino acids contain uncharged polar side chain - serine, threonine, cysteine, aspargine, glutamine, and tyrosine. Three amino acids serine, threonine, tyrosine contain hydroxyl groups attached to side chain. Cysteine is structurally similar to serine but contains a sulfhydryl or thiol group (-SH) in place of the hydroxyl group.

**Charged Polar Amino Acid**

***Positively charged group***

Lysine and arginine have side chains that contain positively charged groups at neutral pH or physiological pH. Lysine has amino group whereas arginine contains a guanidium group. Histidine contains an imidazole group, an aromatic ring. The imidazole group can be uncharged or positively charged near neutral pH, depending on its local environment.

***Negatively charged group***

Amino acid aspartate and glutamate contain acidic side chains that contain negatively charged groups at physiological pH.

**Amino acids with nonpolar side chain**

Among standard amino acids, nine amino acids contain nonpolar side chain or R group. These are glycine, alanine, valine, leucine, isoleucine, proline, methionine, phenylalanine; tryptophan.

Proline differs from other members in having its side chain bonded to both the nitrogen and the α-carbon atoms. Phenylalanine and tryptophan have aromatic side

chains. The side chains of phenylalanine contain a phenyl ring whereas tryptophan has an indole ring.[2]

### Extinction Coefficient

The extinction coefficient indicates how much light a protein absorbs at a certain wavelength. It is useful to have an estimation of this coefficient for following a protein which a spectrophotometer when purifying it.

It has been shown that it is possible to estimate the molar extinction coefficient of a protein from knowledge of its amino acid composition. From the molar extinction coefficient of tyrosine, tryptophan and cystine (cysteine does not absorb appreciably at wavelengths >260 nm, while cystine does) at a given wavelength.[5]

### Average Hydropathy

The Average Hydropathy value for a peptide or protein is calculated as the sum of Hydropathy values of all the amino acids, divided by the number of residues in the sequence.[5]

## MATERIALS AND METHODS

The "Sequencia" Primary Structure Analysis Tool came into Existence when while working over the Analysis of one Primary Structure of Protein; I came to realized that the online tool Available for it are limited to few Properties and also we can paste only one sequence at a time; even we cannot give any FASTA file for computation. The Result also cannot be stored somewhere for future proceedings thus delimiting all such limitations "SEQUENCIA" came into existence. It is part of BioSoftware Development. The "SEQUENCIA" is made in JAVA, BIOJAVA, and COREJAVA. The GUI of the Tool is made using JAVA SWING. The codes of the attributes are written using Corejava and Biojava all under the JAVA Architecture. To decide the attributes I gone through various Research Papers and the demand of the Properties which are mandatory while analyzing the Primary Sequence Analysis. The Physicochemical Properties which were discussed everywhere are the Molecular Weight, Instability Index, Extinction Coefficient, Sequence Length, Sequence Name, Aliphatic Index, Amino acid composition, Amino acid classification, Average Hydropathy, Isoelectric point, Net charge, chemical nature of protein, Absorbance.

In order to accomplish the desired aim collection of important attributes and uses them to build up a Tool certain methods and tools are being used. In this section introduction about tools used and the standard workflow information is provided. All these information is divided into following subcategories:

- Front-end tool
- Back-end tool
- Architecture of Tool
- Workflow of Software Development

### Front-end tool

#### *Java Swing*

Software Development in Java is impossible without Swing. SWING two components which make it portable and handy are its light weight and pluggable nature. The Lightweight component of swing make it non platform specific and to be written entirely into java. Thus this makes it more efficient and more flexible; the each component is derived by its swing not by the specific operating system. The other property of Java Swing is the Pluggable look which means each component code is in java only thus eventually feel will be off swing. With this Pluggable property we can modify one component without affecting others. Swing GUI has container and component. A component is independent and the container is having a group of components. J component class gives the swing components. Package is javax.swing. Class name begin with J.Class name for the label is Jlabel, class for push button is JButton, and class for scroll bar is jScrollbar. In containers Jframe, Japplet, JWindow, JDialog. Jtextfield allow us to edit one line of text. With the help of JAVA swing the GUI of the SEQUENCIA is designed where the home page consist of drop down box and we can choose the options of either the sequence or the FASTA file and according to the choice the Browser or the "INPUT SEQUENCE" page is displayed.[8]

### Back-end tools

#### *Core Java*

Java is set of software used for developing Applications. It is revolutionary Language. Encapsulation, Polymorphism, Inheritance, Abstraction are the Key features of the Core Java. The java is platform independent language. Exception Handling and using object as parameters are key features while writing the core java code for SEQUENCIA. Packages and Interfaces are special significance of the Core java. Multi threading concept is also unique feature of core java. Core java is used in writing the basic code of calculating the Attributes of SEQUENCIA. The Control Statement and the Interface, Exception Handling are the concept used in the Development of SEQUENCIA.[8]

It is used to write the code for the back end core program for the tool SEQUENCIA. The code for calculating the physicochemical properties of the Primary sequence is done through the core java.

### BioJava

Bio java is open source package for Biological Application. In the Biojava the java tools are used for the biological data. Biojava is where the java programming is used for the Sequences, Protein structures and also for Genome study. Biojava is having a group of library functions. Thus Biojava is used for many scientific applications. In SEQUENCIA to calculate the codes of some specific attributes like Extinction Coefficient packages of Biojava is imported.

### Operating System

JAVA is platform independent so no specification for the operating system.

### Architecture of the Tool

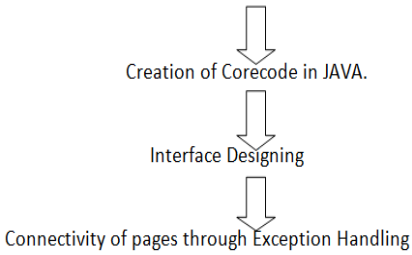
The Tool use JAVA SWING Architecture.

1. JAVA SWING
2. CORE JAVA
3. BIOJAVA

Main feature of this architecture is to make use of open source available component. All freely available elements make it easily reachable to general people community and also performance wise it is very efficient.

### Work Flow of Software Development

Collection of attributes related to Primary-sequence Analysis




## RESULTS AND DISCUSSION

### Collection of attributes related to Primary-sequence Analysis

For development of the proposed Tool the prime importance was to collect the meaningful attributes from reliable resources. For that we have chosen major research journals and their review articles related to ongoing researches and clues to Primary-Sequence Analysis. Majority of the journals were from Primary Structure of Protein background. After that attributes was collected and classified into various categories like physicochemical properties related attributes and other general information categories.

### Creation of Tool in JAVA

To create Primary Sequence Analysis Tool we used the Java platform. Firstly I studied about the online tools related to Primary Sequence Analysis and then I came to know the problem faced by common researchers regarding this aspect.

The Online Tool for Primary Sequence Analysis Protparam is limited to an extent as it take only one sequence at a time and it never take files for computation; the Result cannot be stored as Text File, The attributes are also limited. I wrote the source code named as SEQIO to calculate the physicochemical properties of the Sequencia. In the source code I also imported the packages of open source Biojava to calculate properties such as Extinction coefficient.

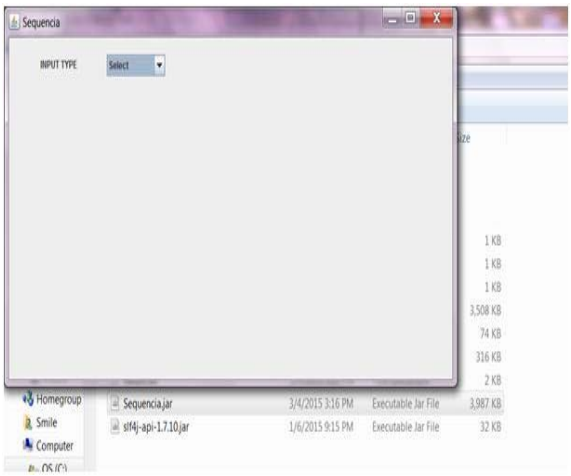



**Figure 1:** Home page of SEQUENCIA

### Interface Designing

Interface Designing is done using JAVA SWING. This GUI design is platform independent. Here there are two options for the interface design either the drag or drop box or can write the code for the expected GUI. I tried to make the GUI as such that any researcher can go and check the drop down box; there will be two options either can paste sequence or can paste or browse FASTA files. The GUI is design as such that while calculating parameters; a timer is visible on front end.

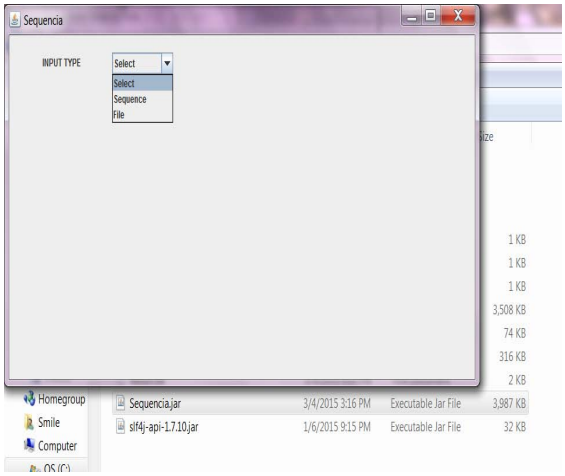



**Figure 2:** Dropdown box

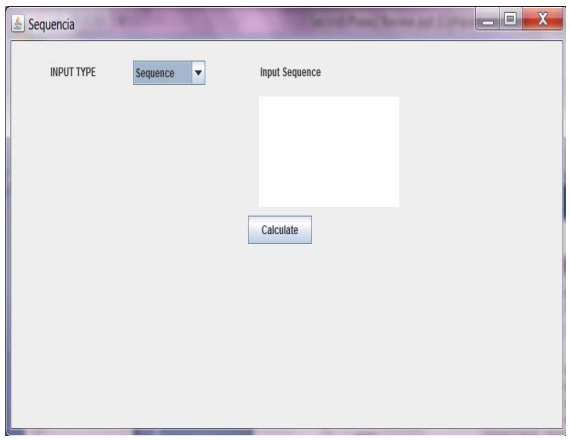



**Figure 3:** Depict the sequence page of SEQUENCIA

### Connectivity through Exception Handling

In interface we have lot of option to browse the needed data along with the keyword search using search boxes. These queries by user can only be answered by fetching information from Tool. For that these pages should get connected through Tool and that was done with help of Exception Handling which works with Java swing and send query to Tool and again print the output result for user. After the establishment of connection from Tool these search pages were actively worked.

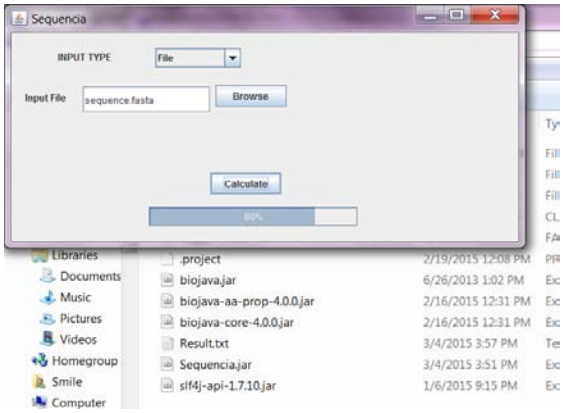



**Figure 4:** Browser page of SEQUENCIA

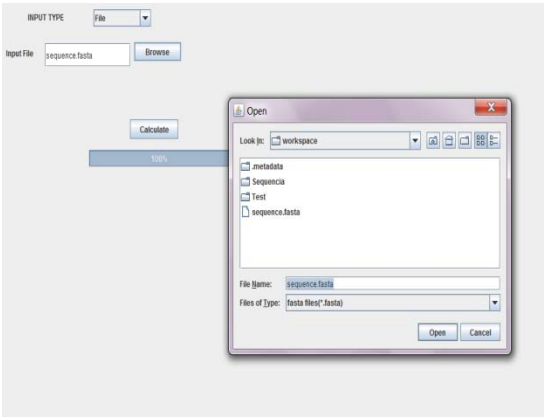



**Figure 5:** depict the File Page of SEQUENCIA and how the GUI Is linked to Source code.

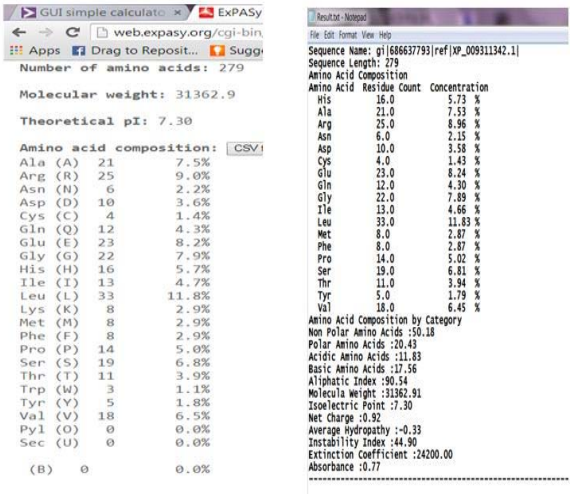



**Figure 6:** Comparisons of the Protparam and SEQUENCIA Result

## CONCLUSION

The current work aimed at developing a Primary Sequence Analysis Tool. The tool SEQUENCIA is offline tool which is part of Software Development; it is built in Java framework where the source code for calculating the physicochemical properties is written using core java and open source biojava. The GUI is design with the JAVASWING .SEQUENCIA is the tool for researchers where the attributes of primary sequence analysis all are grouped under one common platform. The SEQUENCIA is designed in such a way that a drop down box is there for sequence and files. Either can paste sequence or more than one sequence; meanwhile can give FASTA file also; SEQUENCIA will calculate the parameters and the result will be automatically stored in the workspace area.

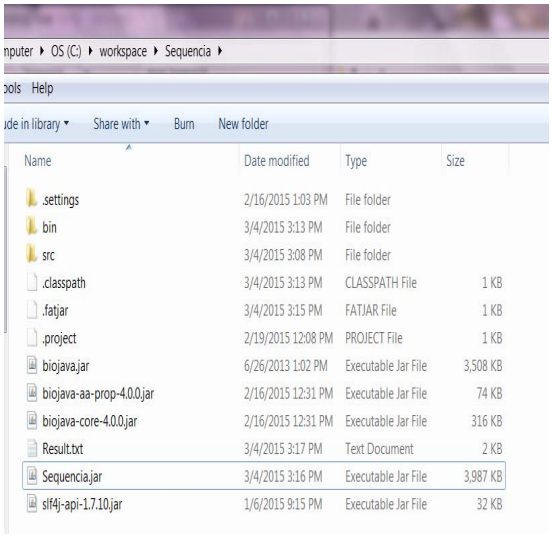



**Figure 7:** Depict the page where the result of the SEQUENCIA can be stored in text format.

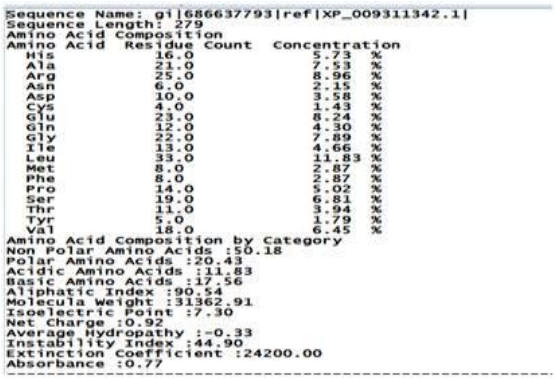



**Figure 8:** Depict the Result of the SEQUENCIA showing the calculated Parameters of Primary Sequence Analysis.

For the Future Prospective I will design the SEQUENCIA Primary Sequence Analysis Offline tool for Android Application term under Updated version. The basic aim for this Android Application is because we should not always depend on computers for our tool to run.

## REFERENCES


1. Kumar P, mina U, Life sciences fundamentals and practices part-1, 2014, Path finder publications, New Delhi, 2014.
2. Nelson DL, Cox MM, *Lehninger's Principles of Biochemistry,4th edition,* W. H. Freeman and Company, New York, 2005, 75-11.
3. Meister A, *Biochemistry of the Amino Acids, Vol 1&2,* 2nd edition, Academic Press Inc, New York, 1965, 75-109.
4. Cavasotto LN, Phatak SS, Homology modeling in drug discovery: current trends and applications, Drug Discovery Today, 14, 2009, 676-83.
5. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins M.R, Appel R.D, Bairoch A, *Protein Identification and Analysis Tools on the ExPASy Server,* 2005, 571-607.
6. Xia, Xuhua, Bioinformatics and the Cell, Springer, US, 2007, 207-219, http://dx.doi.org/10.1007/978-0-387-71337-3_10.
7. Gitlin I, Carbeck JD, Whitesides GM, Why Are Proteins Charged? Networks of Charge– Charge Interactions in Proteins Measured by Charge Ladders and Capillary Electrophoresis, Angew. Chem. Int. Ed, 45, 2006, 3022–3060.
8. Schildt H, Java: The Complete Reference, 7th Edition, Mc Graw Hill, New York, 2007, 15-864.


**Source of Support:** Nil, **Conflict of Interest:** None.