RESEARCH ARTICLE OPEN ACCESS

Protein sequence analysis for breast cancer disease

P.Sasikala*, K.Sathiyakumari **
**(Department of Computer Science, Bharathiar University, Coimbatore-641046*
Email: sasisethu96@gmail.com)
*** (Department of Information Technology, Bharathiar University, INDIA*
Email: sathiyakumari@psgrkc.ac.in)

ABSTRACT

The furthermost challenge facing the molecular biology community today is to make sense of the wealth of data that has been produced through the genome sequencing projects. The cells cover a central core called nucleus, which is warehouse of an important molecule known as DNA. These are packaged in small elements know as chromosomes. They are collectively known as the genome. While the computerized applications are used all around the world, there come to mind that the collection of a vast amount of data are accessed by peoples. The significant information hidden in vast data is attracting the researchers of multiple regulations to make study in developing effective approaches to gain the hidden knowledge within them. In protein and DNA analysis, the sequence mining techniques are used for sequence alignments, sequence searching and sequence classifications. The researchers are showing their interest on protein sequence analysis, in the field of protein sequence classifications. It has the capability to discover the persistent structures that exist in the protein sequences. This work explains various techniques methods to analyze protein sequence data and also provides an overview of different protein sequence analysis methods.

Keywords – protein sequences, DNA, amino acids, nucleiotide, peptide

Date of Submission: 26-08-2017 Date of acceptance: 13-09-2017

I. INTRODUTION

The bioinformatics is a field of discipline, in which various fields like information technology, biology, and computer science merge to form a single discipline. It is the promising field that deals with the biological problems on the molecular level based on the application of computers to the collection, organization, analysis, manipulation, presentation, and sharing of biologic data [1]. According to the author, bioinformatics is the mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information of proteins.

The major three important sub-disciplines within bioinformatics are, the growth of new algorithms and statistics to assess relationships among members of large data sets; finally the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information about proteins.

Over the past few years vast amount of developments in genomic and other molecular research technologies have combined to produce a wonderful amount of information related to molecular biology [2]. At the starting of the genomic mutiny, the main concern of bioinformatics was the creation and maintenance of a database to score biological information such as nucleotide and amino sequences. The information's are inclusive picture of normal cellular activities so that researchers may study how these activities are altered in different disease conditions [3]. For these reason, the field of bioinformatics has develop such that the most very important task to analyze and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures.

The reaming paper is organized as follows:
Section II gives the brief introduction about protein sequence. Section III demonstrates the breast cancer disease. Section IV describes the experiments and results; finally Section V gives the conclusion of this work.

II. PROTEIN SEQUENCE

Proteins are the large molecules that are formed with one or more chains of amino acids in a specific order. The normal size of a protein molecule may be hundred amino acids, while the large proteins can have a thousand amino acids. The 20 amino acids are {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, make the various array of proteins found in living things [4]. Protein Sequence is measured as a very important part of biological sequence data where the analysis and study have become an important research trend and content in bioinformatics domain.

The determination of amino acid sequence, that makes up the protein sequence. Some research can be performed on protein sequence pattern mining is an important task in the field of protein sequence or a protein family sequence making. One of the fundamental tools for the analysis of proteins is similarity searching [5]. The process can be executed in the amino acid sequence or in the spatial structure of the protein.

Searching for relationship among the proteins may have different applications. Based on the area, proteins can be analyzed at the level of amino acid sequence or with respect to different features of their structures [6]. Comparative analysis of the protein sequences may be helpful for the detection of proteins, identification of their functions and determination of their fundamental physical and chemical properties. The comparative analysis of the protein sequences carry much more information and is tremendously important in the processes such as the prediction of the properties of the newly discovered proteins that are difficult to be identified on the basis of amino acid sequences.

2.1 Protein Sequencing Strategy

The usual strategy for determining the amino acid sequence of a protein involves eight basic steps:

- Splitting polypeptide chain

Protein molecular should be separated and purified. Some of the polypeptides are combined together by non-covalent bonds; this is known as oligomeric protein. For example. 8 mol/L urea or 6mol / L guanidine hydrochloride can be used to deal with tetramer—Hb and dimer—Enolase [7].

- Detecting the number of polypeptide in protein moleculars

The number of polypeptides can be determined by detecting the relationship between the number of moles of amino acid residues and protein molecular weight.

- Breaking disulfide bonds

Several polypeptides chains are linked by disulfide bonds. Disulfide bonds will be reduced to thiol with excessive & [beta]- mercaptoethanol under the condition of 8mol / L urea or 6mol / L guanidine hydrochloride. And then it should be protected by alkyl reagents from re-oxidation.

- Detecting the amino acid composition of polypeptide chains and calculating the molecular ratio of amino acid composition.
- Sequencing N-terminal and C-terminal of polypeptide chains

Amino acid of polypeptides is divided into two categories: amino-terminal and carboxyl-terminal. The N-terminal is much more important in the analysis of amino acid sequence of peptide chains than C-terminal.

- Polypeptide can be cleaved into several small peptides. More than two methods can be used to break peptide samples into two or more sets of peptides or peptide fragments and then separate them.
- Determining the amino acid sequencing of each peptide
- Determining the sequence of peptide fragments in polypeptide chains
- Determining the position of disulfide bonds in the original polypeptide chains

Generally, pepsin will be used to deal with those polypeptide chains with disulfide bonds. And then 2D-electrophoresis technology will be used to separate each peptide fragment. Analyzing the composition and sequence of peptide fragments, which may contain disulfide bonds [7].Some methods are used to analyze peptide fragments to determine the position of disulfide bonds.

2.2 Protein sequence databases

Researchers have found out that the sequences of various proteins and store their details in databases. Some well-known databases are Protein Data Bank (PDB), UniProt, Swiss-Prot, SCOP, etc. The protein sequences can be extracted from above databases.

An example protein sequence of Hemoglobin beta chain is given below:

```
VHLTPEEKSAVTALWGKVNVDEVGGEALGRL
LVVYPWTQ
RFFESFGDLSTPDAVMGNPKVKAHGKKVLGA
FSDGLAHLD
NLKGTFATLSELHCDKLHVDPENFRLLGNVLV
CVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

III. BREAST CANCER

Cancer is a set of diseases that affect cells in the body to change and grow out of control. Most types of cancer cells are ultimately form a piece or bunch called a tumor, and is named after the part of the body where the tumor instigates [8]. Breast cancer starts in the breast tissue that is made up of glands for milk production, called lobules, and the ducts that connect the lobules to the nipple. The rest of the breast is made up of fatty, connective, and lymphatic tissues.

Breast cancer is normally detected in two ways either during a screening examination, before symptoms have developed, or after symptoms have been developed, when a woman feels a lump. Most masses observed on a mammogram and most breast lumps turn out to be benign; that is, they are not cancerous, do not mature uncontrollably or spread, so it is not life-threatening issue [9]. When cancer is assumed based on clinical breast exam or breast imaging, microscopic analysis of breast tissue is necessary for a perfect diagnosis and to determine the level of spread and differentiate the pattern of the disease. The tissue for microscopic analysis can be obtained via a needle or surgical biopsy method [10]. Choice of the type of biopsy is based on individual patient clinical factors, availability of particular biopsy devices, and resources for diagnose.

3.1 Staging

In easy terms the stage of a cancer illustrates the size of the tumor and determines whether it has spread and how far it has spread. The stage is important because it helps cancer specialists to decide on the best treatment option. Adjacent is a simplified description of a staging system for breast cancer. There are three main stages of breast cancer

- Early stage: refers to cancer that is confined to the fatty tissue of the breast.
- Locally advanced: which has spread to underlying tissue of the chest wall.
- Advanced or metastatic: where the tumour has spread to other parts of the body. The following figure shows the stages of breast cancer.

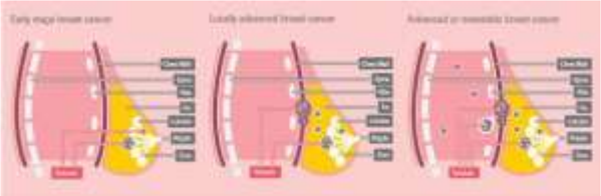

Early stage breast cancer
Locally advanced breast cancer
Advanced or metastatic breast cancer

Fig. 1. Stages of breast cancer

IV. EXPERIMENTS AND RESULTS

The experiment was carried out using Matlab 2014a platform. In this work protein sequence is analyzed using bioinformatics toolbox. Methods used for analyzing protein sequence are explained in detail as follows

4.1 Data collection

The breast cancer protein sequence was collected from neXtPort (https://www.nextport.org) database. This was a major milestone as this collection of data is already quite rich in information pertinent to modern biomolecular medical research. But there remains a large gap in our knowledge of human proteins in terms of functional information as well as protein characterization. At the SIB, neXtProt is developed in the CALIPHO (Computer Analysis and Laboratory Investigation of Proteins of Human Origin) group that was created jointly with the University of Geneva. Headed by Amos Bairoch and Lydie Lane, CALIPHO is an interdisciplinary team which aims to use a variety of methodologies to help uncover the function of uncharacterized human proteins.

- Fasta format

The downloaded dataset is fasta format. In bioinformatics, FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The format allows for sequence names and comments to herald the sequences. The format originates from the FASTA software package, but has now become a standard in the field of bioinformatics.

4.2 Results

- Amino Acids Structure

Protein sequencing is the convenient process of determining the amino acid sequence of all or part of protein or peptides. It is often desirable to know the unordered amino acid composition of a protein prior to attempting to find the ordered sequence, as this knowledge can be used to facilitate the discovery of errors in the sequencing process or to distinguish between ambiguous results [11]. An indiscriminate method often known as amino acid analysis [12] for determining amino acid frequency is as follows: i) Hydrolyse a known quantity of protein into its constituent amino acids. ii) Separate and quantify the amino acids in some way. The figure shows the amino acid structure.

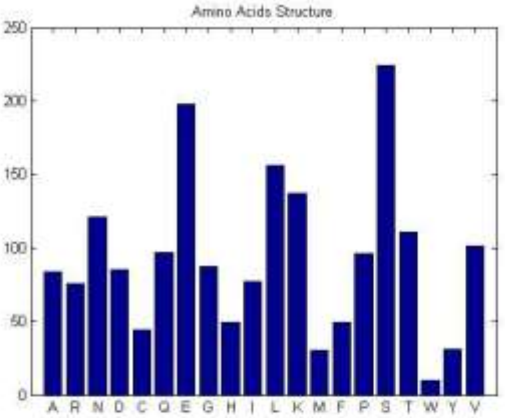


Fig. 2. Amino acid structure

- Isoelectric point

The isoelectronic point or isoionic point is the pH at which the amino acid does not migrate in an electric field. This means it is the pH at which the amino acid is neutral. The following figure illustrates the isoelectric point.

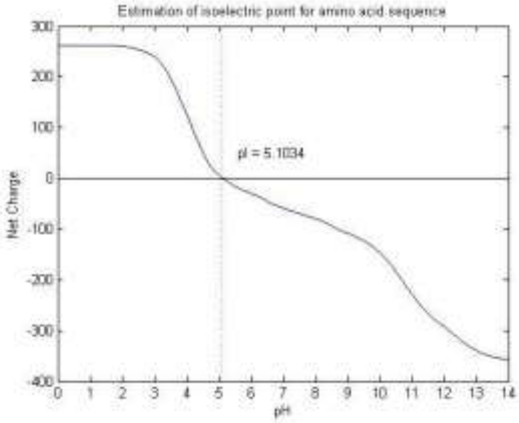


Fig.3. Iso-electric point chart

- Molecular weight

The matlab function *molweight* is used to find molecular weight of the sequence. Molecular weight for given sequence is as follows

Molweight= 2.0772e+05

- High-resolution isotope mass distribution and density function

The majority of elements occur in nature as a combination of isotopes. These are atom class of the same chemical element that has different masses. Isotopes have the same number of protons and electrons, but a different number of neutrons. The main elements occurring in proteins are CHNOPS. The following figure gives the high-resolution isotope mass distribution and density of the protein sequence.

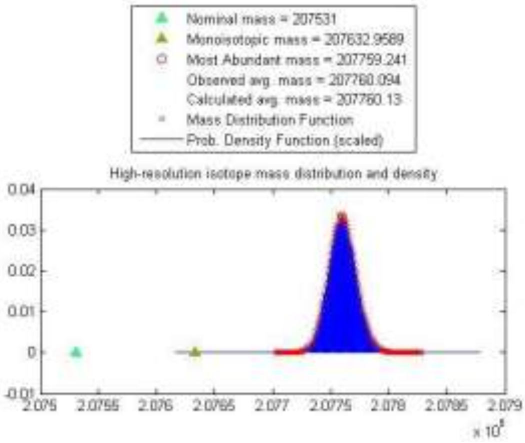


Fig.4. Isotope mass distribution

- Nucleotide density

The function ntdensity used to plots the density of nucleotides A, C, G, and T in sequence. The following figure displays the nucleotide density and counts.

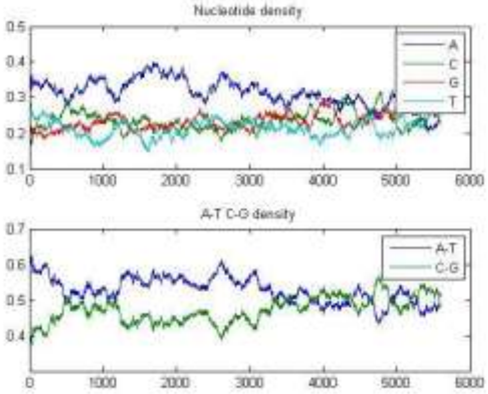


Fig.5.(a) Nucleotide density

A: 1751	A: 1200
C: 1321	C: 1317
G: 1317	G: 1321
T: 1200	T: 1751
Base count	Nucleotide counts in the reverse complement of a sequence

Fig.5.(b) Nucleotide counts

- Codon frames

In molecular biology, a reading frame is a way of dividing the sequence of nucleotides in a nucleic acid (DNA or RNA) molecule into a set of consecutive, non-overlapping triplets. Where these triplets associated to amino acids or the stop signals throughout translation, they are called codons. Example of Trinucleotides (codon) code for an amino acid is given below figure.

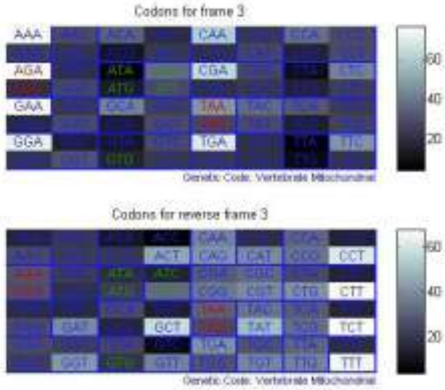


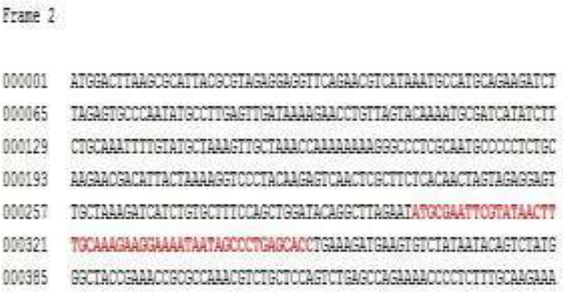
Fig.6. (a)Codons frame

AAA - 69 AAC - 59 AAG - 68 AAT - 62
ACA - 26 ACC - 30 ACG - 26 ACT - 29
AGA - 9 AGC - 35 AGG - 11 AGT - 35
ATA - 26 ATC - 33 ATG - 30 ATT - 18
CAA - 41 CAC - 26 CAG - 56 CAT - 23
CCA - 28 CCC - 22 CCG - 26 CCT - 20

Fig.6.(b) Trinucleotides (codon) code

- Open Reading Frames (ORFs)

In the molecular genetics technique, ORF is the element of a reading frame that has the potential to be translated. An ORF is a continuous stretch of codons that do not contain a stop codon (usually UAA, UAG or UGA). Sample of open reading frame if given below.



```
Frame 2

000001   ATGGACTTAAGCGCATTACGCGTAGAGGAGGTTCAGAACGTCATAAATGCCATGCAGAAGATCT
000065   TAGAGTGCCCAATATGCCTTGAGTTGATAAAAGAACCTGTTAGTACAAAATGCGATCATATCTT
000129   CTGCAAATTTTGTATGCTAAAGTTGCTAAACCAAAAAAAAGGGCCCTCGCAATGCCCCCTCTGC
000193   AAGAACGACATTACTAAAAGGTCCCTACAAGAGTCAACTCGCTTCTCACAACTAGTAGAGGAGT
000257   TGCTAAAGATCATCTGTGCTTTCCAGCTGGATACAGGCTTAGAATATGCGAATTCGTATAACTT
000321   TGCAAAGAAGGAAAATAATAGCCCTGAGCACCTGAAAGATGAAGTGTCTATAATACAGTCTATG
000385   GGCTACCGAAACCGCGCCAAACGTCTGCTCCAGTCTGAGCCAGAAAACCCCTCTTTGCAAGAAA
```

Fig.7. Open Reading Frames (ORFs)

- Sequence viewer

The sequence viewer is used to view the given protein sequence along with amino acid count and average percentage. The following figure shows the result of sequence viewer.

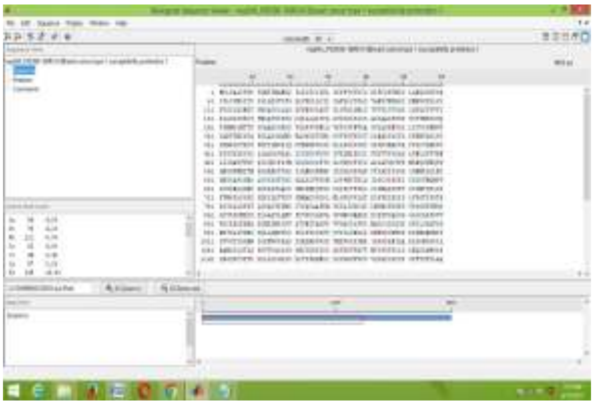


Fig. 8. Protein sequence viewer

V. DISCUSSION AND CONCLUSION

In recent trends, analysis of large amount of biological data like protein sequences is very difficult using traditional database system. Data mining techniques can be used to classify the unknown protein sequence. This work provided an overview protein sequence analysis using various methods. The analysis of breast cancer protein sequence discussed below

- Protein sequence of Amino acid structure is displayed and also iso-electric point of the sequence is determined and plotted.
- Molecular weight and isotope mass distribution are calculated. There are various types of mass distributions were calculated such as normal mass, mono-isotopic mass, most abundant mass, observed average mass, calculated average mass and finally probability of density function is calculated and plotted.
- Nucleotide density and counts are calculated; both base count and reverse complements of the sequence are also calculated.
- Codon frams and codes are evaluated, there are three repetitive frames are formed for given sequence and open reading frames (ORFs) are evaluated. Finally the sequence viewer is used to display the sequence with amino acid count and its average percentage.

REFERENCES

[1]. D. M. Mount. *Bioinformatics: Sequence and Genome Analysis 2nd ed.*(Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, ISBN 0-87969-608-7, 2004).

[2]. Hui-Huang Hsu. *Introduction to Data Mining in Bioinformatics* (Idea Group Publishing, ITB12936).

[3]. Jacques Cohen. *Computer science and bioinformatics* (Communications of the ACM, v.48 n.3, p.72-78, March ,2005)

[4]. Wang Z, Moult J. SNPs, *protein structure and disease* Human mutation 2001, 17(4):263-270.

[5]. Rabie Said, Mondher Maddouri , Engelbert Mephu Nguifo. *Protein sequences classification by means of feature extraction with substitution matrices*.

[6]. Padro Gabriel Ferreira, Paulo J.Azevedo. *Protein sequence classification through Relevant sequence Mining and Bayes Classifiers*

[7]. Rost, B., Sander. C. *Prediction of protein secondary structure at better than 70% accuracy* (J. Mol. Biol. 232:584– 599, 1993).

[8]. Sainsbury JR, Anderson TJ and Morgan DA. ABC of breast diseases: breast cancer. BMJ 2000; 321: 745-50.

[9]. Coleman MP, Quaresma M, Berrino F. et al. Cancer survival in five continents: a worldwide population-based study (CONCORD). Lancet Oncol 2008; 9: 730-56.

[10]. Cancer Help.org. Triple Negative Breast Cancer. Last accessed May 2011 at

http://www.cancerhelp.org.uk/about-cancer/cancer-questions/triple-negative-breast-cancer

[11]. Bogosian G, Violand BN, Dorward-King EJ, Workman WE, Jung PE, Kane JF (January 1989). *Biosynthesis and incorporation into protein of norleucine by Escherichia coli* The Journal of Biological Chemistry. 264 (1): 531–9. PMID 2642478.

[12]. Jump up Michail A. Alterman; Peter Hunziker (2 December 2011). *Amino Acid Analysis: Methods and Protocols* (Humana Press. ISBN 978-1-61779-444-5).

International Journal of Engineering Research and Applications (IJERA) is **UGC approved** Journal with Sl. No. 4525, Journal no. 47088. Indexed in Cross Ref, Index Copernicus (ICV 80.82), NASA, Ads, Researcher Id Thomson Reuters, DOAJ.

P.Sasikala. “ Protein sequence analysis for breast cancer disease.” International Journal of Engineering Research and Applications (IJERA) , vol. 7, no. 9, 2017, pp. 90–95.