# Comparative Study of Different Computational Systems for Analysis of DNA and Protein Sequences

Fakiha Shamsi, Zulfiqar A. Memon, Abdul Rehman Soomrani, and Qamar Udin Khand

***Abstract*—Comparison between DNA's and protein structures of human genome is the subject of this paper. We will discuss some basic models static and dynamic for modeling biological systems, different alignment algorithms and issues regarding each. For the rest of paper we will also discuss several issues relating with each approach there advantages and disadvantages and in the end we also will discuss future advancements that may be considered for further development in this area.**

***Index Terms*—DNA's and protein structure, biological systems.**

## I. Introduction

Managing biological data is one of crucial issue in research and also it is one of the most advance areas of research this is due to rapid growth of biological information. The elegant combination of information with technology have been produced a tremendous amount of information in molecular biology. Several projects have been succeeded in this area which provides a clear cut for monitoring the behavior of human genome and patterns it provides. Bioinformatics includes number of activities some of very much important are mapping and analyzing DNA and protein sequences form the collected information and on the basis of that information one can align different DNA and protein sequences and also can place the comparison between them to capture the relationships and after this one can generate 3-D model for viewing protein structures. Bioinformatics is the science of organizing, analyzing, mining, integrating, managing, retrieving and interpreting information form biological data at the genomic, metabalomlic, proteomic, psylogenetic, cellular or whole organism echelon. Genomic biology has much increased from few years and several tools have been developed to expertise genomic engineering which results in exponential growth in complete or partial databases. It deals with different algorithms, databases and information retrieval systems, web technologies, artificial intelligence and, structural biology, software engineering, data mining, image processing, modeling and simulation and several others. This emerging importance is due to the accumulation, transmission and growth of information in biological systems [1].

The field of bioinformatics is parallel to the biophysics and biochemistry this term "Bioinformatics" is actually refers to managing, analyzing, and retrieving biological information and store that integrated an manipulate information in the database that are created and highly designed to handle such type of biological data, primary purpose for maintaining such type of databases is not only to map or design issue relating to particular genomic information but also to develop and provides an interface specially for researchers for accessing of existing data as well as submission of new data [1]. Bioinformatics provides entailment for emerging, organizing and updating the databases which are warehouse for biological data. It also provides support for making tools that can further be utilized for analyzing the obtained information. Gathered information can then be used for discovering and developing gene-based drugs [2]. Major objective to compare biological structures such as DNA and proteins and to analyze sequence patterns is to find out the genetic patterns in a genome [2]. DNA and proteins are fundamental parts of every living organism; DNA is a unit of heredity it get transferred from one generation to another all of our characteristics like hair, skin, eyes are covered up in DNA, while proteins are the molecules that provides structure and function of a body like enzymes that are capable to break down food are proteins, hairs are almost proteins, skin cell are crammed full of proteins actually DNA molecule is cover or code for molecule of a protein they both make us what we are [3]. In each cell of individual living organism there is more than 75 special kind of proteins that with RNA molecules form a single protein according to the instructions provided by DNA. Proteins are most important part in every living organism they worked as small machines in our body and this will be all done with the help of DNA so DNA and protein both are important not only one. So the comparison between DNA and proteins is nothing but it is all about finding the relationship between them [4].

Manuscript received September 19, 2013; revised November 22, 2013.

Fakiha Shamsi, Zulfiqar A. Memon, Abdul Rehman Soomrani, and Qamar ud din Khand are with department of Computer Science at Sukkur Institute of Business Administration (Sukkur IBA), Sindh, Pakistan (e-mail: shamsi.fakiha@gmail.com, zulfiqar@iba-suk.edu.pk, rehman@iba-suk.edu.pk, qamar@iba-suk.edu.pk).

## II. Methodology

After alignment the sequence generated by protein and DNA may be used either to find out regions of similarity that defines a preserved consensus pattern of characters (nucleotides or amino acids) in all sequence combinations. These generated sequences may be used to evolutes a relationships between the sequences for this to be done he alignments should be well-formed and capable to infer relationships Commonly, the problem is: let $\Sigma$ be an alphabet and $S = \{S1, \ldots, Sk\}$ be a set of string defined over $\Sigma$. A multiple sequence alignment of S is a set $S\_ = \{S\_1 \ldots S\_k\}$ such that• $S\_ \in (\Sigma \cup \{-\})$ for each $i = 1, \ldots, k$• $Si$ is obtained from $S\_i$ by reducing all gap symbols $\{-\}$• $|S\_1| = |S\_2| = \ldots$

DOI: 10.7763/IJBBB.2014.V4.316

= |*S_k*|. A scoring function defined on the alphabet Σ is a map σ : (Σ ∪ {−})*k* → *R*. It has the following properties:

1) Reflexivity: Reflexivity will provide the maximum score if all the sequences are similar.
2) Symmetry: Regardless of where the differences are made the resulting score is based on the evaluation of multiset of characters in the arguments;

$$\sigma(x1, ..., xi, a, xi+2, ..., xj, b, xj+2, ..., xk) = \sigma(x1, ..., xi, b, xi+2, ..., xj, a, xj+2, ..., xk)$$

3) Triangle Inequality: It should be noticed that similarity should be the opposite of distance:

*D*(|*S*1|, |*S*2|, . . . , |*Sk*|) is considered as the best score for aligning *k* sequences *S*1, *S*2, . . . , *Sk* with respect to *σ* is the one that maximizes the sum of the *σ*s across all positions:

*i* ∈1..*n* *σ*(*S*_ 1[*i*], *S*_ 2[*i*] . . . *S*_ *k*[*i*]). If |*S*1| = |*S*2| = |*Sk*| = *n*, then the space and the time complexity of the best currently known algorithm is:

*O*(*nk*) and *O*(2*knk*) × *O*(computation of the *σ* function), respectively.

## III. Models for Biological Systems

### A. Static Modeling

Static modeling includes the sequences of proteins, nucleic acids and peptides it provides an interaction of data among proteins, Nucleic acids and peptides including microarray and networks of proteins and metabolites.

### B. Dynamic Modeling

When dealing with dynamic modeling, system biology is most preferable which includes metabolites concentrations and reaction fluxes. Cellular events such as signaling, transcription and reaction dynamics are most preferably captured by Multi agent-based approaches. More than thousands of DNA sequences of different organisms have been stored in databases after decoding, the information gathered in these database are then can further be analyzed and updated, and will try to find out genes that are capable enough to encode polypeptides(proteins), RNA genes, structural motifs, repetitive sequences and regulatory sequences. A comparison between several species of different organisms that may be similar of different with genes can illustrate relation between species or similarities between protein functions. As long as the data is continue to grow it is not possible to analyze genetic codes and DNA sequences manually so computer programs are most preferable. With the huge volume of biological data the field of bioinformatics provides a platform for two contradictory disciplines to work together and support each other for common goal of analyzing and studying the genome completely but not partially. Several software programs have been being built to do the work with great speed and accuracy, regardless of the amount of data given to the database.

Many of the computer programs such as BLAST that is used to search sequences from more than 26,0000 organisms, may have greater than 190 billion nucleotides. The program can then be reimbursed for certain operations such as mutations (exchanged, deleted or inserted bases) in the DNA sequence, just before identify sequences that are related, but not indistinguishable. A variant of this sequence alignment may be used in the sequencing process itself. Most preferable sequence alignment technique is shot-gun alignment which is not capable enough to produce complete set of chromosomes instead it can produce large number of small DNA fragments ranging from 35 to 900 nucleotides, again it depends on the type of sequence technology we are using with respect to the requirement of application.

Shot-gun sequencing technique is complicated for larger genomes, but it is good enough to produce sequence data very fast. For larger genome like human genome which requires large number of CPU cycles to perform computations. Shot-gun sequencing technique provides a simplest and fastest way to assemble genomes [5].

### C. Disadvantage

With traditional sequence alignment techniques, when dealing with larger genomes, such as for human genomes significant number of CPU time will be utilized, it also requires large-memory and large time to perform computations. With many sequence analysis techniques we may found gaps in the result which has to be full filled later on. To reduce gaps between sequences shot-gun technique is most preferable [5].

## IV. Computer Generated 3-D Modeling of Biological Data

### A. 3-D Models of DNA

Fig. 1, Fig. 2 and Fig. 3, shows 3-D image of DNA packing tight into a structure. The structure is inimitable in that the genome is entirely unknotted, meaning, that no matter how dense it is but you can draw it, easily get to the region you want to set down, read it off, and put it back when you're done [6].

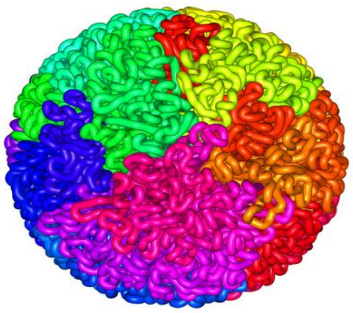

Fig. 1. 3-D model of DNA.

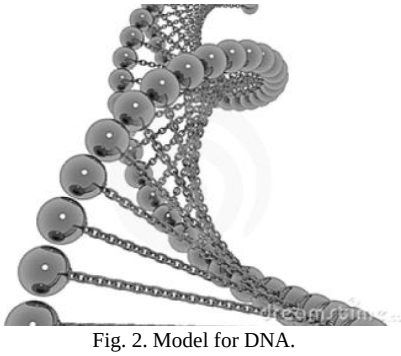

Fig. 2. Model for DNA.

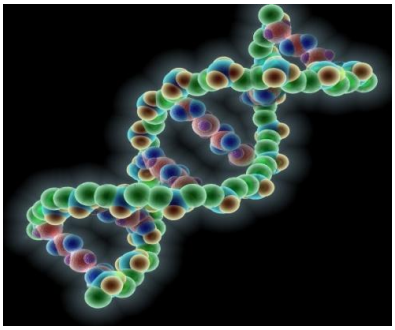

Fig. 3. 3-D model for molecule of DNA.

### B. 3-D Models of Proteins

A 3-D model of protein, shown in Fig. 4, Fig. 5 and Fig. 6, depicts the structure packing of protein that will be compared with DNA structures to identify the relationship between both of them.

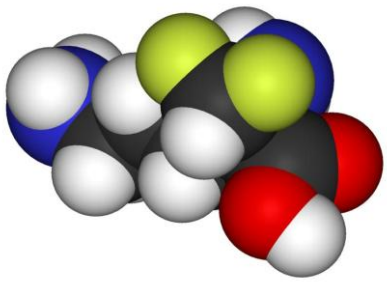

Fig. 4. A computer generated 3D image of a protein model.

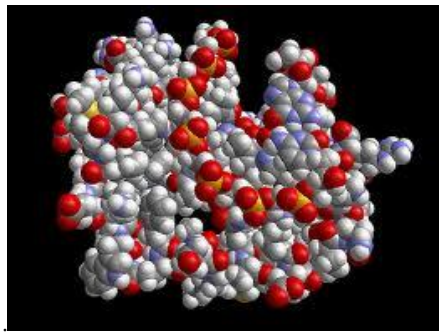

Fig. 5. Structure of protein.

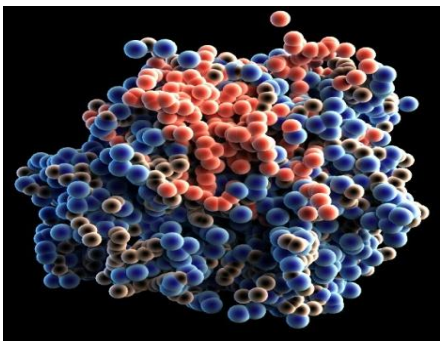

Fig. 6. 3-D model for molecules of protein.

## V. Comparison between Sequences of Proteins and DNA

For the comparisons of DNA and protein sequences several packages have been introduced such as FAST which has been extended to FASTX and FASTY which are capable to compare DNA sequences to a protein sequence database, It will translate DNA sequences in more than two frames and then for performing comparisons alignment will be performed. We have discussed in this paper several alignment algorithms such as BLAST or Dotplots with their mechanism with the comparison of FASTX and FASTY TFASTX and TFASTY have also been introduced that will compare protein sequence to a DNA sequence [6]. FASTX, FASTY, TFASTX, and TFASTY perform the calculation in four steps 1: By using a lookup table that will find the identical regions of DNA and protein 2: Perform rescanning of regions with the help of matrices (BLOSUM50 scoring matrix) 3: Merge regions of similarities after full filing the gaps 4: Compute an optimal solution [6].

For translation and comparison of DNA to protein several algorithms have been introduced. Sequence between protein and DNA is early known in analysis [7], protein sequences present more information then DNA and when we needs to align both, there are various algorithm for doing this. The comparison between DNA and protein sequences can be made by two means first we could compare on similarity bases by considering physical similarities between sequences and by putting them align the major objective of alignment is to measure the identities that could be global that attempt to span full length of sequences or local that will try to find sub-sequences with good score. Alignment could be done through number of algorithms i-e through Dotplots that can reveal local alignments and BLAST as we have discussed BLAST in this paper before that could find optimal local alignments from large databases of proteins.

Another way to compare DNA and protein is “Homology” that will measure the similarity due to evolutionary relationship between DNA and protein to find either they are homologous or not [7].

Global alignments are good statistical scoring while local alignments are difficult to score on the basis of complexity both protein and DNA have low complexity regions and DNA sequences contain repetitive elements these regions may cause confusions in database search to remove such confusions a process of filtering may be used we may remove the confusions or complexities by analyzing query sequences first or by filtering the results [7].

We may perform another way for analyzing sequences through annotations there are various software’s for performing annotations. These annotations are capable to mark the genes and other biological features in DNA sequences.

## VI. Conclusion

Main objective to compare genomes or other biological data is to provide the establishment to the correspondence between genes and other biological features of different organisms. We have discussed some of most important points to compare DNA and protein sequences with some models may be static or dynamic that describes the DNA and protein sequencing. The main objective of doing this is to catch the patterns of sequences, later on these sequences will help to identify the relationship among biological data, and will help in various field or researches in bioinformatics such as mining biological data, disease recognition from human genomes and so on. We have also discussed different types of

comparisons with their complexities. We have discussed "Homology" that is one of the key idea in bioinformatics, which is used to predict the function of a biological data. We have discussed different alignment algorithms such as Dotplots that is used to plot the similarity between sequences in the form of a dot it will use one straight diagonal across for similar sequences, and BLAST that will perform the sequence comparison on the basis of similarities. In the comparison section we have discussed about two methods of comparison FASTX and FASTY that can be further extended to TFASTX and TFASTY they both will be used to compare the sequence for getting the relationship after aligning DNA and protein.

This study is a footstep toward widespread rationalization of DNA and protein sequences that will help to recognize a comprehensive relationship among biological data.

## REFERENCES


[1] eBook Info list ENIN-11645-Computer Science. [Online]. Available: http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-1-85233-671-4

[2] Christine. (March 2011). Bioinformatics–The application of the different technologies for gathering and analyzing information. [Online]. Available: http://www.ismb2008.org/bioinformatics-the-application-of-the-different-technologies-for-gathering-and-analyzing-information

[3] P. E. Gary, W. A. Reynolds, and R. Reynolds, "DNA: The key to life," *Programmed Biology Series, Educational Methods*, Chicago, 1977.

[4] Genetic Science Learning Center. (August 6, 2012). Tour of the Basics. *Learn. Genetics*. [Online]. Available: http://learn.genetics.utah.edu/content/begin/tour

[5] P. H. Searls and B. David, "The roots of bioinformatics in theoretical biology," *PLoS Computational Biology*, vol. 7, no. 3, 2011.

[6] W. R. Pearson, T. Wood, Z. Zhang, and W. Miller, "Comparison of DNA sequences with protein sequences," *Genomics*, vol. 46, issue 1, pp. 24-36, 1997.

[7] A. Elumalai, "An overview of bioinformatics," *Journal of Science – Biological Science*, vol. 2, Issue 2, pp. 71-80, 2012.


**Fakiha Shamsi** was born and raised in Sukkur Pakistan. As a child growing up in Pakistan she was fortunate to attend the well regarded educational institutes. She has done her FSC from Hira Public Higher Secondary School, a well reputed institute and BS from Shah Abdlu LAtif University Khairpur Sindh. The year 2011 was a best year of her life, in that year, she have joined Sukkur IBA and started her MS in computer Science. It was the actual place where she got many research ideas and started her research career.

October 2013 is one of the best months in her life. In this month she got her first research paper accepted for presentation and this was an honor for her. This year it is her hope to secure a good GP for her MS degree and continue with her dream of PhD with the wonderful support of her husband and love of her parents. She wants to become a good philosopher and will serve her country Pakistan.

**Zulfiqar A. Memon** is an associate professor in Department of Computer Science at Sukkur Institute of Business Administration (Sukkur IBA), Sindh, Pakistan. He obtained his doctorate degree in 2010, in Computer Science specialization from VU University, Amsterdam, The Netherlands. His thesis title was "Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective". He joined Sukkur IBA as Lecturer in 2002. Dr. Memon has co-authored many international publications, including workshops, conference papers and journal articles. He has also served as a reviewer for international conferences. Apart from international publications, Dr. Memon has worked in various national and international research projects funded by well-reputed international donor agencies, like ADB, World Bank, etc.

**Abdul Rehman Soomrani** is an associate professor in Department of Computer Science at Sukkur Institute of Business Administration (Sukkur IBA), Sindh, Pakistan.

**Qamar Udin Khand** is an associate professor in Department of Computer Science at Sukkur Institute of Business Administration (Sukkur IBA), Sindh, Pakistan.