Volume 3, No. 3, May-June 2012

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Data Mining of Biological Data in Bioinformatics using Transcription, Translation Algorithm and Pattern Matching of Protein Sequences

Vivek Gangwar*[1], Yogendra Singh[2] and Dr. Udayan Ghose[3]
M.Tech IT, USIT(GGSIPU)[*1], M.Tech CSE, USIT(GGSIPU)[2], Assistant Professor, USIT(GGSIPU)[3],
Dwarka,New Delhi, India
gangwarvivek86@gmail.com*[1], yogendrasng1@gmail.com[2], g_udayan@lycos.com[3]

Abstract: Data mining of biological data in Bioinformatics is an emerging area of research. In this paper algorithms of Transcription(conversion of DNA to RNA) and Translation (RNA to Protein conversion) will be described. Since whenever human body is affected by any type of bacteria or virus ,to over come this our body produces protein for fighting against them. So by analyzing the protein sequence we can easily find out the disease by which human body is affected. For this we also describe a method for finding these type of diseases by pattern matching and percentage of matching algorithm of protein sequences.
So for this we provide a platform, in which different types of algorithms will work together and fetch the useful data with database in which we stored protein sequences of different types of diseases.

Keywords: Bioinformatics, Biological Data, DNA, RNA, Protein, Transcription, Translation, Start, Stop.

I. INTRODUCTION

Bioinformatics and data mining provide exciting and challenging research and application areas for computational science. Bioinformatics is the science of managing, mining, and interpreting information from biological sequences and structures.

Biological research is becoming increasingly database driven, motivated, in part, by the advent of large-scale functional genomics and proteomics experiments such as those comprehensively measuring gene expression.

Consequently, a challenge in bioinformatics is integrating databases to connect this disparate information as well as performing large-scale studies to collectively analyze many different data sets. This approach represents a paradigm shift away from traditional single-gene biology, and it often involves statistical analyses focusing on the occurrence of particular features (e.g., folds, functions, interactions, pseudogenes , or localization) in a large population of proteins. Moreover, the explicit application of machine learning techniques can be used to discover trends and patterns in the underlying data.

II. BIOLOGICAL DATA

In bioinformatics biological data mainly related to DNA, RNA and protein[10]. In many experiments and research these biological data are considered. These are main chemicals of living body.

A. DNA Sequence:

DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms.

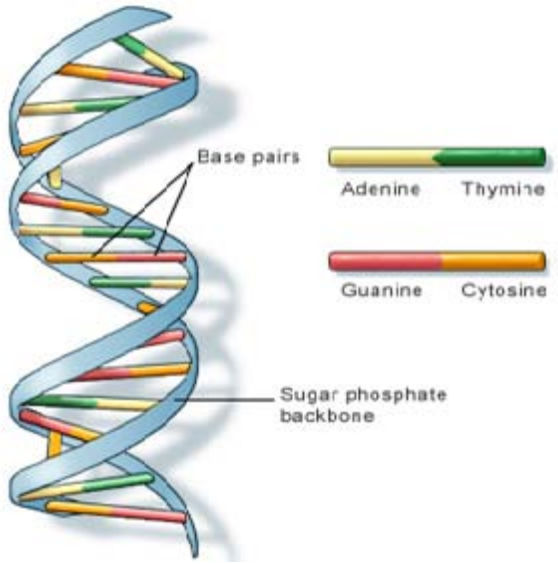

Figure. 1 DNA structure [4]

The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T).

DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule.

B. RNA sequence:

Ribonucleic acid or RNA, is one of the three major macromolecules (alongwith DNA and proteins) that are essential for all known forms of life.

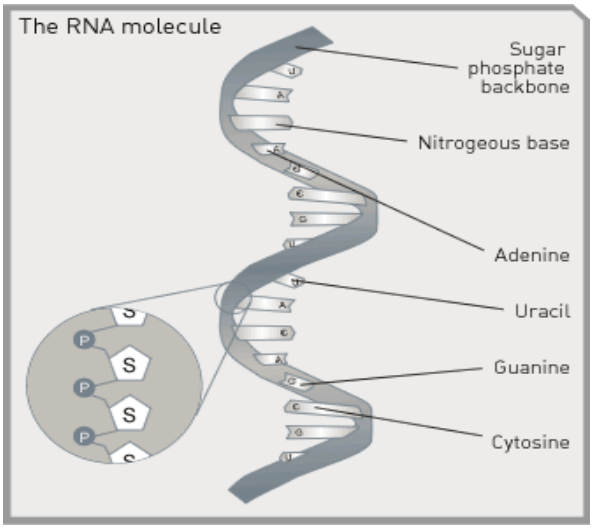

Figure.2 RNA structure [4]

Each nucleotide in RNA contains adenine (A), cytosine (C), guanine (G), or uracil (U). Adenine and guanine are purines, cytosine, and uracil are pyrimidines.

C. Protein Sequence:

The protein as read off from the mRNA may not be in the final form that will be used in the cell. Some proteins contains signal Peptide , this signal peptide is used to guide the protein out of the nucleus towards it´s final cellular localization[7]. This signal peptide is cleaved-out at the cleavage site once the protein has reach (or is near) it´s final destination.Various Post-Translational modifications .The final protein is called the "mature peptide".

III. CENTRAL DOGMA OF MOLECULAR BIOLOGY

In the central dogma[6] of molecular biology there are two processes one is transcription and other one is translation. After these two processes DNA has been changed to equivalent protein.

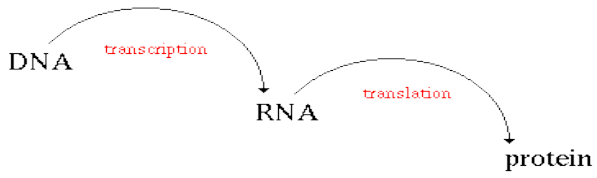

Figure.3 Central dogma [4]

A. Transcription:

Transcription[6] is the process by which the information contained in a section of DNA is transferred to a newly assembled piece of messenger RNA . In this conversion of DNA to RNA by replacing compliment of 'A' to 'U' ,'C' to 'G'.'T' to 'A' and 'G' to 'C'.

B. Translation:

Translation is the process of decoding a mRNA molecule into a polypeptide chain or protein. Each combination of 3 nucleotides on mRNA is called a codon or three-letter code word. Each codon specifies a particular amino acid that is to be placed in the polypeptide chain (protein)[9].

Table 1. Conversion of codens to protein [6]page no 280

Protein	Code triplets	Amino Acid Name
A	AAA,AAG	Lysine
B	AAU,AAC	Asparagine
C	ACU,ACC,ACA,ACG	Threonine
D	CGU,CGC,CGA, CGG,AGA,AGG	Arginine
E(start)	AUG	Methionine
F	AUU,AUC,AUA	Isoleucine
G	CAA,CAG	Glutamine
H	CAU,CAC	Histidine
I	CCU,CCC,CCA,CCG	Proline
K	GAA,GAG	Glutamic acid
L	GAU,GAC	Aspartic acid
M	GCU,GCC,GCA,GCG	Alanine
N	GGU,GGC,GGA,GGG	Glycine
O	GUU,GUC,GUA,GUG	Valine
P	UAU,UAC	Tyrosine
R	UCU,UCC,UCA, UCG,AGU,AGC	Serine
S	UGG	Tryptophane
T	UGU,UGC	Cysreine
U	UUA,UUG,CUU, CUC,CUA,CUG	Leucine
W	UUU,UUC	Phenylalmine
Stop	UAA,UAG,UGA	None

C. Transcription and Translation Algorithms:

Since DNA is the sequence made of four elements those are ACTG, and conversion of RNA to protein[11] will be done in the form of triplets .Therefore the sequence of DNA will be also in the form of triplets .So there will be 64 (4*4*4)combinations of four elements ACTG taking three at a time .

Some important points while making the algorithms of Transcription and Translation, those are as under-

a. DNA sequence should be in the multiple of 3, means should be in the form of triplets and consist of only ACTG characters.
b. Valid DNA sequence always consists of a Start triplet (TAC) and a Stop triplet (ATT, ATC, ACT).
c. Conversion of DNA to RNA will start from Start triplet and end before stop triplet. Before start the DNA sequence will not be converted into RNA , similarly applicable after stop.

Algorithm –

Input DNA( )

i. DNA ← DNA sequence entered by user
ii. (DNA.length)%3==0 //sequence is multiple of 3 or not.
//check sequence contain ACTG character only.
iii. for i ← 1 to DNA. length
iv. do if (DNA[i]!='A')
v. if(DNA[i]!='C')
vi. if(DNA[i]!='T')
vii. if(DNA[i]!='G')
viii. Call InputDNA( )
ix. else
x. call Check Start Stop(DNA)

CheckStartStop(DNA)

// To check whether sequence contain Start and Stop.
i. for i ← 1 to DNA. length
ii. Sub ← take subsequences of length 3 from DNA
iii. if (Sub=="TAC") // check sequence contain Start.
iv. Start ← i
// check if sequence contain Stop.
v. for j ← i to DNA. length
vi. if ((Sub=="ATT")||(Sub=="ATC")|| (Sub=="ACT"))
vii. Stop ← j
viii. call Convert RNA (DNA, Start, Stop)
ix. else
xi. call Input DNA( )

Convert RNA (DNA, Start, Stop)

i. RNA ←null
ii. For i ← Start to Stop
iii. if (DNA[i]=='A')
iv. RNA ← RNA + "U"
v. if (DNA[i]=='T')
vi. RNA ← RNA + "A"
vii. if (DNA[i]=='C')
viii. RNA ← RNA + "G"
ix. if (DNA[i]=='G')
x. RNA ← RNA + "C"
xi. Call Convert Protein(RNA)

Convert Protein (RNA)

i. Protein ← null
ii. for i ← 1 to RNA. length
iii. Sub ← take subsequences of 3 from RNA
iv. if (Sub=="AUG")
v. Protein ← Protein + "E"
vi. if ((Sub=="UUU")||(Sub=="UUC"))
vii. Protein ← Protein + "W"
viii. if((Sub=="UUA")||(Sub=="UUG")||(Sub=="CUU")||(Sub=="CUC")|| (Sub=="CUA")||(Sub=="CUG"))
ix. Protein ← Protein + "U"
x. if((Sub=="AUU")||(Sub=="AUC")||(Sub=="AUA"))
xi. Protein ← Protein +"F"
xii. if((Sub=="GUU")||(Sub=="GUC")||(Sub=="GUA")||(Sub=="GUG"))
xiii. Protein ← Protein +"O"
xiv. if((Sub=="UCU")||(Sub=="UCC")||(Sub=="UCA")||(Sub=="UCG")|| (Sub=="AGU")||(Sub=="AGC"))
xv. Protein ← Protein +"R"
xvi. if((Sub=="CCU")||(Sub=="CCC")||(Sub=="CCA")||(Sub=="CCG"))
xvii. Protein ← Protein +"I"
xviii. if((Sub=="ACU")||(Sub=="ACC")||(Sub=="ACA")||(Sub=="ACG"))
xix. Protein ← Protein +"C"
xx. if((Sub=="GCU")||(Sub=="GCC")||(Sub=="GCA")||(Sub=="GCG"))
xxi. Protein ← Protein +"M"
xxii. if((Sub=="UAU")||(Sub=="UAC"))
xxiii. Protein ← Protein + "P"
xxiv. if ((Sub=="CAU")||(Sub=="CAC"))
xxv. Protein ← Protein + "H"
xxvi. if ((Sub=="CAA")||(Sub=="CAG"))
xxvii. Protein ← Protein + "G"
xxviii. if ((Sub=="AAU")||(Sub=="AAC"))
xxix. Protein ← Protein + "B"
xxx. if ((Sub=="AAA")||(Sub=="AAG"))
xxxi. Protein ← Protein + "A"
xxxii. if ((Sub=="GAU")||(Sub=="GAC"))
xxxiii. Protein ← Protein + "L"
xxxiv. if ((Sub=="GAA")||(Sub=="GAG"))
xxxv. Protein ← Protein + "K"
xxxvi. if ((Sub=="UGU")||(Sub=="UGC"))
xxxvii. Protein ← Protein + "T"
xxxviii. if ((Sub=="UGG"))
xxxix. Protein ← Protein + "S"
xl. if((Sub=="CGU")||(Sub=="CGC")||(Sub=="CGA")||(Sub=="CGG")|| (Sub=="AGA")||(Sub=="AGG"))
xli. Protein ← Protein + "D"
xlii. if((Sub=="GGU")||(Sub=="GGC")||(Sub=="GGA")||(Sub=="GGG"))
xliii. Protein ← Protein + "N"
xliv. return RNA, Protein sequences

So by applying above algorithm[3] user can get RNA, Protein sequences of any DNA sequences[8].

We can easily understand by taking an example-

Suppose a DNA sequence is as under

ATGTAGGATTACAAAAAGAATAACGAAGAGGATG
ACTAATAGTATCAACAGCATCACAGAAGGAGTAGC
GGAGGGGGTGGCTGATGGTGTTGCCGACGGCGTCG
CATAATGGTAGTGGTTGTCTTATTGTTTTTCCTACT
GCTTCTCACAACGACCGCAGCGGCTGCCTCATCGT
CTTCCCCACCGCCTCCCATCAAAAAGACT

Equivalent RNA sequence will be as under

AUGUUUUUCUUAUUGCUUCUCCUACUGAUUAUCA
UAGUUGUCGUAGUGUCUUCCUCAUCGCCUCCCCC
ACCGACUACCACAACGGCUGCCGCAGCGUAUUAC
CAUCACCAACAGAAUAACAAAAAGGAUGACGAAG
AGUGUUGCUGGCGUCGCCGACGGAGUAGCAGAAG
GGGUGGCGGAGGG

Equivalent Protein sequence is as under

EWWUUUUUUFFFOOOORRRRIIIICCCCMMMMPPHH
GGBBAALLKKTTSDDDDRRDDNNNN

IV. BIOLOGICAL DATA PROCESSING SYSTEM

Whenever a human body is suffered from a particular disease, it is due to attack of virus or bacteria. Because of this Protein sequence gets little bit changed .To counter attack of virus or bacteria our human body produces protein. By analyzing these protein sequences we can easily find out the disease by which the body is suffered. In the database we have stored predefined protein sequences of different diseases. So by this system we can easily analyze [1]the different biological sequences and find out whether the

human body is suffered or not, if suffered then by which disease.

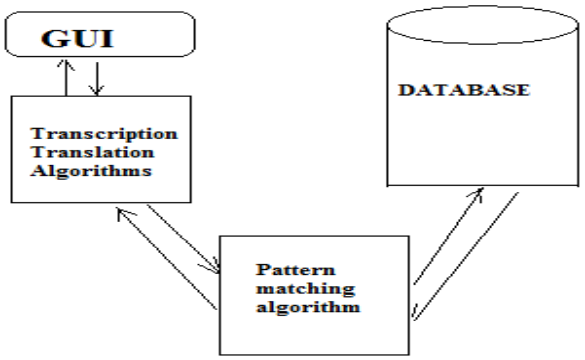

Figure.5 Processing system[11]

In this system different types of algorithms are used for fetching the data from database and also performing the processes of transcription and translation[2]**.** When any DNA sequence will be entered then automatically it will be converted into equivalent RNA and then further converted into equivalent protein. After that , the protein sequence will be matched with stored protein sequences of different diseases[5] in the database. If it will be matched then we can easily say that a human body is affected by that disease ,whom we took DNA for examine

V. CONCLUSION

Bioinformatics and data mining are developing as interdisciplinary science. Data mining approaches seem ideally suited for bioinformatics, since bioinformatics is data-rich but lacks a comprehensive theory of life's organization at the molecular level. However, data mining in bioinformatics is hampered by many facets of biological databases, including their size, number, diversity and the lack of a standard ontology to aid the querying of them as well as the heterogeneous data of the quality and provenance information they contain. Another problem is the range of levels the domains of expertise present amongst potential users, so it can be difficult for the database curators to provide access mechanism appropriate to all. The integration of biological databases is also a problem. However by applying different approaches we can find diseases in human body.

VI. REFERENCES

[1]. DNA sequence matching using Boolean algebra MATCHING USING BOOLEAN ALGEBRA, V. Anitha, Sr Lecturer MCA, Panimalar Engineering College Chennai, India., B. Poorna , Professor MCA, Easwari Enginnering College Chennai, India. 2010 International Conference on Advances in Computer Engineering

[2]. Compressed Pattern Matching in DNA Sequences, Lei Chen, Shiyong, and Jeffrey Ram ,Wayne State University. 2004 IEEE Computational Systems Bioinformatics Conference

[3]. Intrduction to Algorithms By Thomas H. Cormen, Charles E. Leiserson, Ronald L Rivest, Clifford Stein.

[4]. http//www.ncbi.nlm.nih.gov

[5]. http://www.mayoclinic.com

[6]. Principles of Genetics by Gardner/Simmons/Snustad.

[7]. Integrative Data Mining: The New Direction in Bioinformatics by Paul Bertone1 and Mark Gerstein21Dept. of Molecular, Cellular, and Developmental Biology2Dept. of Molecular Biophysics and Biochemistry and Dept. of Computer Science, Yale University, July/August 2001

[8]. R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley and Sons, 2001.

[9]. Data Mining in Bioinformatics , Jinyan Li, Institute for Infocomm Research, Limsoon Wong, National University of Singapore, Qiang Yang, Hong Kong University of Science and Technology, 2005 IEEE.

[10]. H. Liu, J. Li, and L. Wong, "Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data," Bioinformatics, vol. 21, no. 16, 2005, pp. 3377–3384.

[11]. S. Ray and M. Craven, "Learning Statistical Models for Annotating Proteins with Function Information Using Biomedical Text," BMC Bioinformatics, vol. 6, suppl. 1, 2005, p. S18.