Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Detection of disease and Prediction of Post Risk Level from DNA Sequence Using Pattern Matching and GA, A Proposed System

Das Sucharita[1*]

Dept. of Computer Science and Engineering, Siliguri Institute of Technology, Siliguri WB India
Email: sdsucharita@rediffmail.com Mob. +919434458486

ABSTRACT:

Pattern matching algorithm finds the occurrences of a small sequence called pattern in a larger sequence text, takes an essential responsibility in bioinformatics and parallel in medical science. In this paper I have proposed a system that will find the occurrences of infected genes in any sample of human gene sequence. KMP string matching algorithm is being used as sequence matcher. Depending upon the frequency of occurrence of the pattern the disease will be leveled. Along with detecting the stage of the infection the proposed system will be played an important role in detection of upcoming growth for expected cell mutation.

KEYWORDS: DNA sequence, String Matching, KMP Algorithm, Genetic Algorithm.

***Corresponding Author**

Mrs. Sucharita Das

Department of Computer Science and Engineering

Siliguri Institute of Technology, Sukna, Siliguri

Darjeeling-734009, WB, India

Email: sdsucharita@rediffmail.com

INTRODUCTION:

String matching is a technique of[1] searching a particular pattern from a massive volume of stored data. String matching algorithm that concern with finding one or all occurrences of a specific pattern from a given large text dataset. Pattern matching algorithm is an[2] element of bioinformatics that deals with:

--------"the collection, classification, storing and analyzing of biochemical and biological information using computers especially as applied to molecular and genomics."

We can predict future disease that may affect human and essential precautions can be taken to avoid them, from the DNA sequence database of human. By matching[3] exact and approximate pattern of a patient DNA sequence with a database DNA sequence, we can recognize the level of infected DNA in patient. Very few number of[4] DNA sequences of human vary from person to person frequently, but generally gene sequences are quite similar. Because of the varying region of DNA sequence the sample can be identified for a particular human being called genetic code.

The genetic[5] matter of cells that bears biological data (information) in an encoded form of base spring cell to offspring cell genetically, is called as Deoxyribonucleic acid(DNA).DNA is made up of four similar type of bases: Adenine(A), Guanine(G), Cytosine(C), and Thymine(T). Human genome has about three billion pairs of bases. The Specific ordering of As, Gs, Cs and Ts is very essential to detecting the particular species.

The Watson-Crick model states that the DNA molecule is a structure of two strands twisted together, called double helix. Two strands are strongly held together with A=T and G≡C bound only. The model also states that, two strands are complementary of one another i.e. if one strand is ATGAC then the other would have be TACTG –complementary combination of bases in each other[5] in reverse direction.

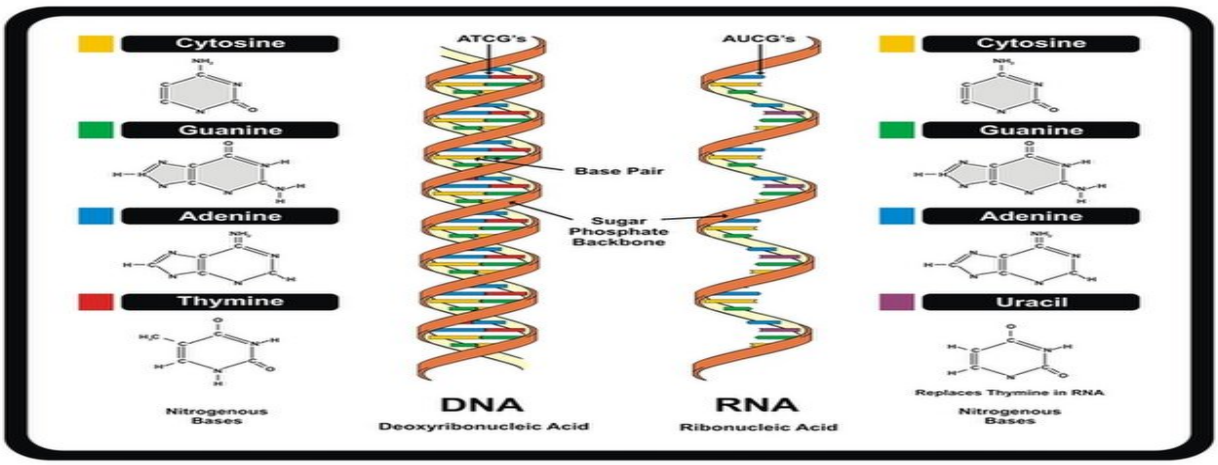

Figure 1: The chemical structure of DNA and RNA.

Ribonucleic acid (RNA) uses the information to enable the cell to synthesize the particular protein. DNA ⇒ mRNA ⇒ Protein

RNA [5] is a combination of four nucleotides: Adenine(A), Guanine(G), Cytosine(C) and Uracil(U). RNA is single-stranded structurally.

- RNAs are of various types-
- mRNA(messenger RNA)
- tRNA (transfer RNA)
- rRNA (ribosomal RNA)
- snRNA (small nuclear RNA)

Among them messenger RNA is a photocopy of gene. DNA is broken down into coils called chromosomes; human beings have 23 pairs of chromosomes, which are further broken into genes (total 70,000 genes in 23 pairs of chromosomes and every gene has its own function.

Proteins are building of 20 different amino acids[5]. The chains of amino acids are linked by peptide bounds called polypeptide and long, complex polypeptides forms proteins.

The basic material and operational units of heredity in gene, bring knowledge for building all proteins needed by organisms. Chain of nucleotides in some particular manner from genes and placed at a specific location of chromosome. Information of DNA is converted into a functional product, proteins, through the gene expression process. Expressed genes produce two types of RNA that are:

- DNA in a gene is copied to produce mRNA (messenger RNA) called transcription and then the transcribed "message" of DNA is translated into proteins of cell.
- Transcription into tRNA and rRNA are not able to translate into proteins.

The complete set of genetic instructions of an organism that contains all the information to built and grow[3] the organism is genome. To detect[4] probable in accuracy or irregularity in any DNA sequence analysis can be used, by comparing a specific gene with other analogous gene from same or different organisms. As a result if any mismatched DNA sequence is encountered, then its functionality can be predict based on its similarity with any sample DNA sequence.

Genetic Algorithm (GA)

A well known probabilistic and adaptive searching algorithm is Genetic Algorithm. GA is a search-based optimization technique based on the principles of Genetics and Natural Selection[6]. It is frequently used to find optimal or near-optimal solutions. GA has three major applications, namely--intelligent search, optimization and machine learning[7]. Each cycle of GA gives new generation.

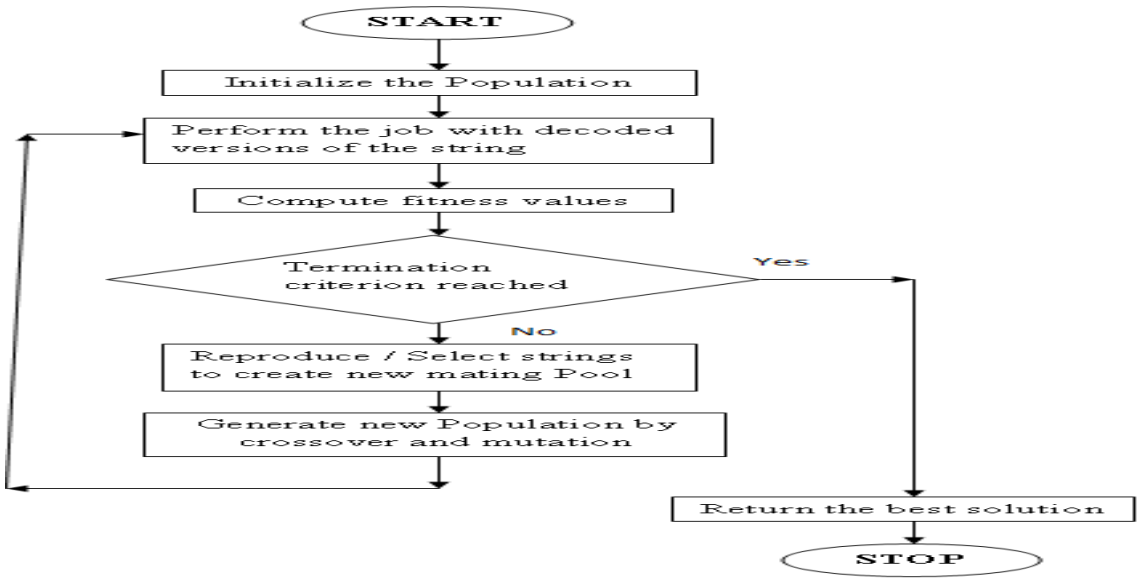

Figure 2: Flowchart of Genetic Algorithm to select Population

Initially a population is created to begin the search process by encoding them into bit-strings called chromosomes. The performances of the chromosomes are then estimated through fitness functions and mating pools are selected for genetic operations. Apply crossover and mutation operators on the mating pool to generate new population. The previous population by the new and estimate again.

Mutation operator helps to bring[8] about a rear but random and unpredictable change in the chromosomes.

String Matching Algorithm (KMP)

Knuth, Morris and Pratt(KMP) is a linear-time string matching algorithm, which has a precompiled auxiliary function π from the pattern string in time $\Theta(m)$ and store in an array $\pi[1\ldots m]$.

Algorithm want to determine exactly how many steps pattern P can be shifted to the right until there is some hope of another match. Determine this number by looking for a maximum suffix that is equal to a prefix of pattern P. Find the largest suffix of P[i-1] that is equal to a prefix of P[i-1]. If the length of this suffix is j, then the mismatched character in A can be matched against P[j+1] directly, without going through all the other redundant matches.

Compute_Next(P, m):
Input: Pattern (a string of size n) .
Output: next (an string of size m) .
1.Next(1) ←- 1.
2.Next(2)← 0.
3.for i←3 to m do
4. set j← next[i-1] +1
5. while(Pattern [i-1]≠ Pattern [j] and j>0)
6. set j← next[j] +1
7. Endwhile
8.Endfor

String_March(Text, Pattern, n, m)
1. set shift←0, i←1, j←1
2. while(i≤n)then
3. if(Pattern[j]==Text[i])then
4. set i←i+1
5. set j←j+1
6. Else
7. set j←next[j]+1
8. Endif
9. if(j=m+1)
10. Print :"Match occurs at (i-m) position of the text ".
11. set j←1and i←i+1
12. Endif
Endwhile

RELATED WORKS

Bioinformatics collects, accumulates, merges and then analyzes the biological data by using computer technology and expose the prosperity of biological information hidden in the large volume of data and discovers a patent into the basic biology of organisms[9].

The string matching algorithm can obtain the occurrences of a particular DNA sequence which cause a specific disease from a large sample of DNA sequence. By pattern matching unknown sequence present in the gene database that transformed from previous generation to offspring[10] can be recognize and also state its impact factors.

The accuracy of the pattern matching algorithm is depends on the capability to find the occurrence of match characters and the ability to remove any mismatch characters[10].

Pattern matching provides a way in diagnosing the disease by identifying the presence[3] of diseased DNA sequence in DNA sample. Compares[3] similar values with threshold value and stores specific output of diseased one and finally can identify optimal result using voting on all possible outputs.

To locate all possible occurrences of any pattern in a text several algorithms exist. A divide–and–conquer technique[11] can divide the large value of text in smaller independent sub-texts and then search for the pattern in sub-texts parallel and point of division have been chosen by applying genetic algorithm (GA) approach. This algorithm gain better performance because of parallel execution and can use a pool of processes to simulate the parallelism[11]. The drawback of this algorithm is that it will take more time to complete execution if pattern does not occur in the text, because of time required in genetic algorithm,

The healthier and stronger individual[12] are selected by nature from a particular generation. The process of natural selection is depends on probability. Genetic algorithm can suggest particular sub-section of the text, are the probability of matching the pattern is maximum[12].

In content of retrieval and pattern searching string matching algorithms play the most

important[1] role. Most string matching algorithm for detecting and extracting the required part of the text were designed for specific purpose and the efficiency differs based on the used dataset. To design an efficient algorithm for string matching for gene dataset, to reduce the execution time and to increase the accuracy, basic working principles of all string matching algorithms are vital.

A small[4] quantity of DNA sequence varies from human to human rapidly but generally human share very similar DNA sequences. Genetic code of an organism is represents by DNA sequence, which define the combination of proteins in the organism. The actual efficiency of any pattern matching algorithm is depends on the accuracy and performance on different dataset[4].

PROPOSED SYSTEM

Description of the Proposed System

This paper is about finding a particular pattern of DNA sequence present in a sample sequence of a person. This algorithm is able to detect whether the sample sequence has any affected DNA or not, by matching the sample with a DNA of the patient who has that specific disease by applying KMP string matching algorithm.

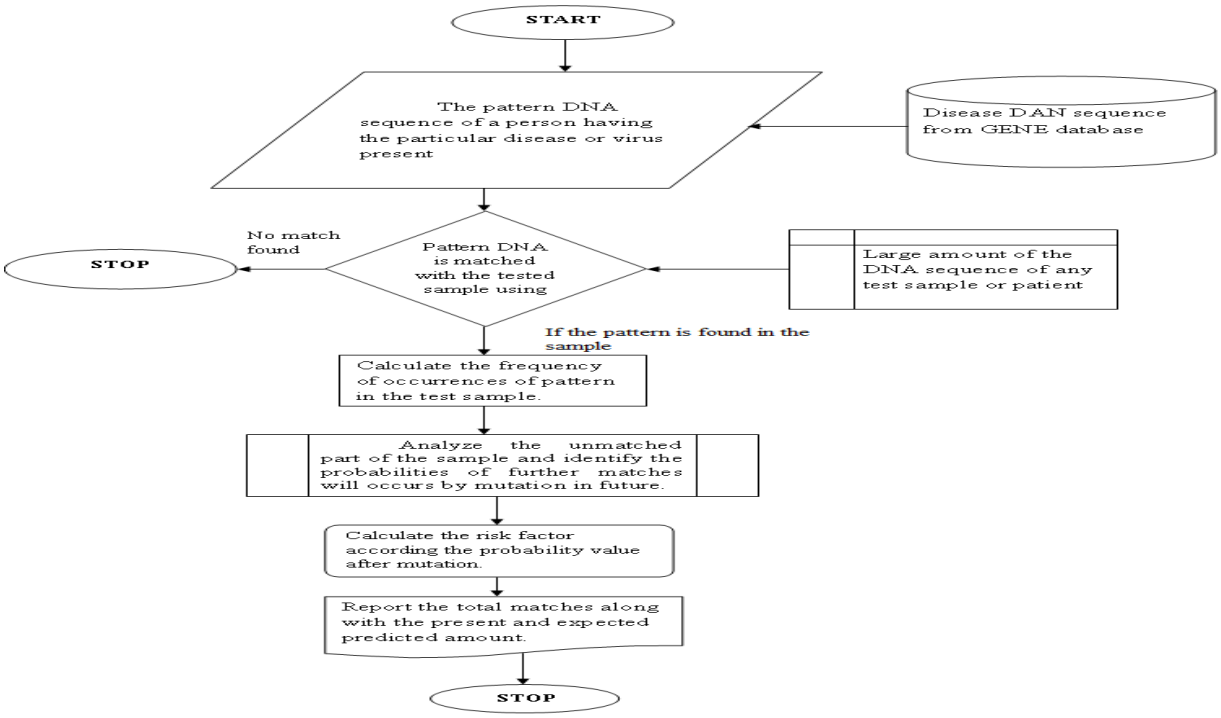

Figure 3: Flowchart of the Proposed System

If KMP finds any match then next target is to searching for the number of occurrences i.e. the frequency of affected DNA present in the sample. Then from that value we can recognize the intensity factor of the disease on that human. How much that disease spread on that person cells.

After recognizing the presence of a particular sequence in a sample sequence the disease is

being identified and the numbers of occurrences of the pattern sequence defines the intensity of the specific disease and also percentage of damaged cells by that disease. Furthermore it is also essential to predict the chances or probabilities of influencing the unaffected DNA by the specific disease caused due to mutation or alteration.

The disease causes due to the occurrence of unusual[13] or mutated sequences. Individual disease has its own format of sequence and the intensity of that disease depends on the rate of occurrence of the mutated gene sequence in the particular DNA. Cancer is caused due to the unrestrained growth of cells through mutations of the genetic material.

If by certain mutation any changes will occur in DNA sequence then what will be the increasing rate of the particular disease in future. In this way we can prefigure the disease at early stage and necessary preventive measure can be taken if possible.

By this proposed method we able not only to detect the disease but also protect the person from further infection by detecting the risk probabilities, which facilitate the treatment and the follow-up process.

CONCLUSION

In this proposed system the DNA sequences of test sample will be mapped with the reference sample of the particular disease using KMP string matching algorithm. If any match will occur then the number of occurrences or the frequency of the present sample pattern will be calculated and intensity of the particular disease will be measured. After the detection of the specific disease it also important to prescribe the future probable chances of further expansion of the infected disease along with the preventive measure. The sequence of any DNA may change by mutation. The proposed system will also have able to measure the approx possibilities of future infection due to mutation occur in cells. The Genetic Algorithm (GA) will be used by this part of system phase to prescribe the probabilities. The proposed system will able to play a huge role in medical science to identify, scale and predict the possibilities of deadly diseases.

REFERENCES

1. Jiji. N and Dr. T Mahalakshmi, *"Survey of Extract String Matching Algorithm for Detecting Patterns in Protein Sequence"*, Advances in Computational Sciences and Technology, 2018;10: 2707-2720, ISSN 0973-6107.
2. M.R. Pooja, M.B. Chandak, "Comparative Study of String Matching Algorithms for DNA dataset", International Journal of Computer Sciences and Engineering, May 2018; 6(5): 1067-1074, May-2018, E-ISSN: 2347-2693.

3. Nyo Me Tun, Thin Mya Mya Swe, *"Comparison of Three Pattern Matching Algorithms using DNA Sequences"*, International Journal of Scientific Engineering and Technology Research, November-2014; 03(35): 6951-6955, ISSN 2319-8885.
4. A. Izzat and N. Maryam ,*"String Matching Evaluation Methods for DNA Comparison"*, International journal of Advanced Science and Technology, October, 2012; 47.
5. C. Kun-Mao, *"Basic Concepts of DNA, Proteins, Genes and Genomes"*, supported in part by NSC grants 94-2213-E-002-018 and95-2221-E-002-126-MY3 from the National Science Council, October 2, 2006, Taiwan.
6. https://www.tutorialspoint.com/genetic_algorithms/genetic_algorithms_introduction.htm.
7. K. Amit, *"Artificial Intelligence and soft Computing Behavioral and Cognitive Modeling of the Human Brain"*, CRC Press, 2000, Boca Raton, Florida, ISBN 0-8493-1385-6.
8. 8. R. Samir, C. Udit, *" Introduction to Soft computing Nuro-Fuzzy and Genetic Algorithms"*, First Impression, Pearson, India, 2013, ISBN 978-81-317-9246-9.
9. S. Kuhu, B. Samarjeet, P. Sunil, *"An Analysis of Influential DNA Sequencing Algorithms*", International Journal of Application or Innovation in Engineering & Management, November 2012; 1(3) ISSN- 2319-4847.
10. S. Rajesh, S. Prathima, Dr. L.S.S. Reddy, *"Unusual Pattern Detection in DNA Database using KMP Algorithm"*, International Journal of Computer Applications, February 2010, 1(22):1–5, Published By Foundation of Computer Science.
11. B. Sagnik, C. Tamal, S. Devadatta , *"Finding all Occurrences of a Pattern by a Genetic Algorithm based Divide-and-Conquer Method"*, International Journal of Computer Applications, February 2013; 64-No.18, (0975-8887).
12. B. Sagnik, C.Tamal, S. Devadatta, *"A Genetic Algorithm Based Pattern Matcher"*, International Journal of Scientific & Engineering Research, November-2012; 3(11).
13. S. Rajesh, S. Prathima, Dr. L. S. S. Reddy, *"Unusual Pattern Detection in DNA Database Using KMP Algorithm"*, International Journal of Computer Application, 1(22).
14. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein, *"Introduction to Algorithms"*, Third Edition, The MIT Press Cambridge, Massachusetts London, England , ISBN 978-0-262-53305-8.
15. K. Michael, K. Leung, D. Andrew, A. Babak, and B.J. Frey *"Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets"*, Proceedings of the IEEE, Jan. 2016;104(1)