DNA Sequence Alignment Using Matching Algorithm to Identify the Rare Genetic Mutation in Various Proteins

Bipin Nair B J [#1]

[#] Department of Computer Sciences
Amrita School of Arts and Sciences, Mysuru Campus
Amrita Vishwa Vidyapeetham, Amrita University, India
bipin.bj.nair@gmail.com

***Abstract*—DNA sequence equivalent identification by implementing string matching algorithm to identify the rare genetic mutation intentions at ascertaining the intricacies involved in decisive the modification emerged in human DNA sequence. The string matching algorithm is chosen from specific areas like mapping. The algorithms are grouped in such way that it can be able to process DNA SEQUENCE. The outset provided for the mapped outputs are epitomized .The available DNA SEQUENCE analyzer tools require expert users to carry out experiments. Sequence matching helps to rectify variations in genotype-phenotype. The scope of this relies in the area of outset of the mapping DNA sequence. The consequence can be made interactive such that the outputs can be with ease interpreted by the user.**

Keyword- DNA sequence, Mutation, Genotype, Phenotype, Alignment

I. INTRODUCTION

Identifying the rare genetic disorder by mapping the mutation rendering in the DNA sequence. There are some substitute customs and tackles are supporting to identify mutation occurred in the DNA sequence .But it carries some of the downsides with respect to processing time and put away more memory space to compute mutation in DNA sequence. In this paper we are going to transform knuth morris pratt string matching algorithm to identify the mutation occurred in DNA sequence. It shapes the partial table of DNA pattern in the initial phase .Advantage of structuring the partial table is suppose if a particular nucleobase is not matching in that move, instead of moving to next nucleobase .Directly it will move to next nucleobase which is already calculated and stored in the partial table. In this way of mapping DNA sequence helps to identify the exact mutated position in the DNA sequence with less memory and processing time. Input will be the DNA sequence which is generated by the machine learning algorithm for the particular symptom. This algorithm avoids multiple scans of DNA sequence because of the partial table generated in the initial stage of matching the DNA sequence. Result will be the exact position of mutation occurred in the sequence. Running time of this algorithm is very fast (O(m+n)).

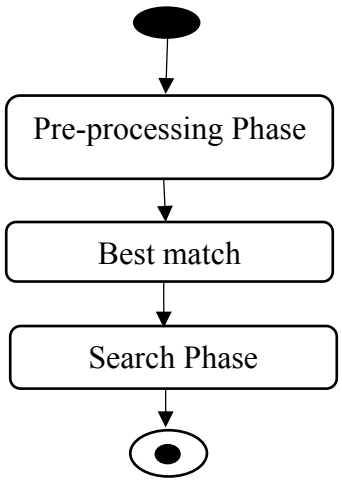

Figure 1: A sample flow diagram for working of the algorithm

II. OVERVIEW

This paper deals with the identification of the alteration in the nucleotide sequence of the genome of an organism, virus, or extra chromosomal genetic element position. Changes in DNA sequence results from unrepaired damage to DNA sequence or to RNA genomes; Alteration can result in several different types of change in sequences. Manually scrutinizing and mapping the genomic variation is very problematic. Since many genetics shares akin symptoms, it is very difficult to identify the alteration position in the DNA sequence accurately. To analyze DNA sequence in medical labs it requires amino acid analyzer which is very expensive. So, implementing the string matching algorithm in such a way that it can handle the DNA sequence data sets and to envision the result precisely in comprehensible manner. Implementing the string matching algorithm gives the advantage of providing the result with the partial sequence matching table. If the DNA sequence is not matching at some position at the time it is not necessary to start the iteration from the beginning. DNA sequence Partial table gives information about next move if the nucleotide sequence of the genome is not matching.

III. PROBLEM STATEMENT

DNA sequence equivalent identification by implementing string matching algorithm to identify the rare genetic mutation intentions at ascertaining the intricacies involved in decisive the modification emerged in human DNA sequence. There are excellent tools that are available. But it has got some snags. They are,

- It cannot handle the DNA sequence without preprocess.
- On juncture training and support is indispensable to use these tools.
- Normally sequence matching algorithm will not redirect directly to the next DNA matching sequence if the current element does not match exactly.
- Number of scan will be more in the other sequence matching algorithm which takes lots of time for processing.
- They provide limited user interaction.
- By hand analyzing the DNA SEQUENCE disparity is very challenging.
- It required lot of memory for storing the sequence.
- It might require to move backward for re scanning the DNA sequence. Which will become less efficient for large file.

IV. PROBLEM FORMULATION

The proposed paper helps the user to work with the DNA sequence matching algorithms in a relaxed way, and the user is delivered with an interactive and effective visualization that benefits to make a useful decision. It mainly overwhelms the confines of the existing work in terms of user participation and visualization.

The Proposed work performs,

- The DNA sequence algorithm handles text data.
- Unindustrialized an understanding of the solicitation domain, relevant prior knowledge and the goals of the end-user.
- Selecting the data set which contains the DNA sequence generated by the machine learning algorithms records which consists the genomic symptoms for particular disease.
- The system will be user friendly; a user will know what is happening to their input data.
- Partial table will be generated for the input pattern of the DNA sequence.
- Based on the partial table DNA sequence matching will happen.
- If the DNA sequence is not matching in the pattern, it will check for the next move in the partial table until the sequence matches.
- Because of the partial table DNA sequence matching can be performed in an effective approach.

V. RELATED WORK

ALGORITHM:

n = Length of the input DNA sequence

m = Length of the Pattern of DNA sequence.

u = Prefix −function of pattern (p).

q = Number of nucleobase matched.

Define the variable:

q=0, the beginning of the match.

Compare the first nucloebase of the pat tern with first nucleobase of input DNA sequence

DNA sequence
I f match is not found, substitute the value of u[q] to q.
If match is found, then increment the value of q by 1.
Check whether all the pat tern DNA sequence are matched with input DNA sequence
If not, repeat the search process.
I f yes, print the number of shift s taken by the Pat tern of DNA sequence.
Look f o r the next match.

```
n   length[S]
m   length[p]
a   Compute Prefix function
q   0
  for i   1 to n do
     while q > 0 and p[q + 1] , S[i] do
          q   a[q]
          if p[q + 1] = S[i] then
                q   q + 1
          end if
          if q == m then
                q   a[q]
          end if
     end while
  end for
```

Here a = u

VI. EXPERIMENT RESULT

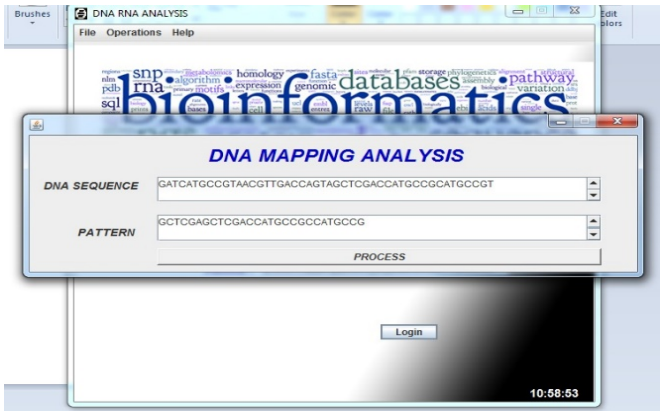

Figure 2 Taking DNA sequence input

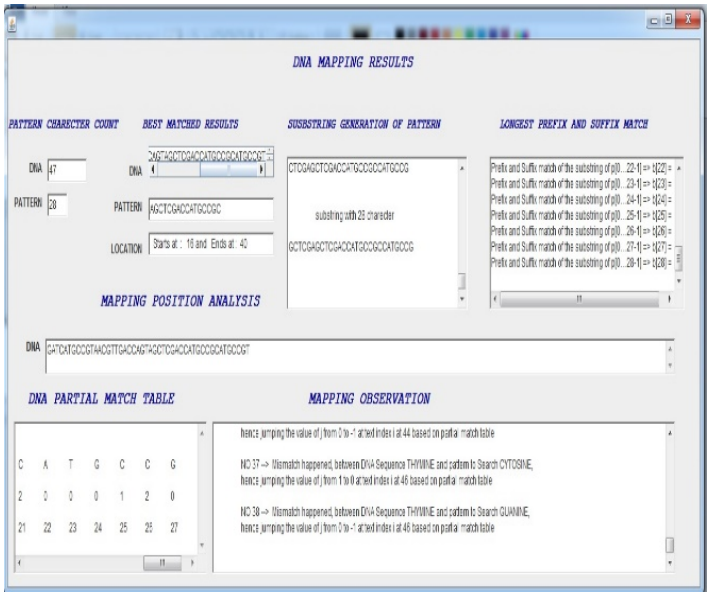

Figure 3 DNA sequence mapping result

VII. LITERATURE SURVEY

The method of Belief networks was functional to 20 SNPs in the apolipoprotein (apo) E gene and plasma apoE levels in a sample of 702 personages from Jackson, MS. Plasma apoE level was the primary target mutable. These scrutinizes indicate that the brink between SNP 4075, coding for the well-known ε2 allele and plasma apoE level was strong. Belief networks can effectually describe complex ambiguous processes and can both learn from data and combine prior knowledge [19].

Phylogeny and statistical modelling have typically been used to examine molecular and phenotypic evolutionary distance separately. We have used this background to develop phylogenetic substitution models to test for associations between evolutionary rate distance of genotype and phenotype. Here they introduce a new method for rating mixture rate matrices between genotype and phenotype [20].

A novel knowledge-driven systems biology method that exploits qualitative knowledge to build Dynamic Bayesian network (DBN) to represent the biological network underneath a specific phenotype [21]

VIII. CONCLUSION

The paper Analysis, The advantage of the proposed system is to progress the efficacy of the prevailing system by the sinking the complexities intricate in dominant the mutation happened in mortal DNA sequence. This result has a profound impact in medical sciences. This work will help to detect DNA sequence miss match in same category of genetic disorder in a resourceful way so that it will lessen the through put and in improve efficiency. The projected idea will aid physicians to spot the region of the genetic mutation and cultivate strategies for its therapy.

This work is part of a research paper, so there were time constraints in the implementation of the paper. As the technologies being used for this were changed in order to come up with the best result, all that has been achieved within this time is the final paper.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their very helpful comments and suggestions and also we extend our gratitude to Dr.Vikas Modi and family Amrita krupa Hospital Mysore, for helpful discussion.

REFERENCES

[1] Data Mining- Introductory and Advanced Topics
[2] Data Mining: Concepts and Techniques by Jiawei Han
[3] A Comparative Analysis of Data Mining Tools in Agent Based Systems by Sharon Christa, K. Lakshmi Madhuri, V. Suma
[4] CSE5230 Tutorial The Na ve ayes lassifier
[5] The Apriori Algorithm – a Tutorial by Markus Hegland
[6] DATA MINING: Theory and Practices by Dr.Shyam Divakar and Dr.K.P.Soman.
[7] Naive Bayes Classifier example - Eric Meisner lecture05-NaiveBayes-2up.pdfAn Improved k-Nearest Neighbor Classification Using Genetic Algorithm ,N. Suguna1, and Dr. K. Thanushkodi2
[8] International Journal of Computer Trends and Technology (IJCTT) - volume4Issue4 –April 2013 ENHANCED DBSCAN ALGORITHM by Priyamvada Paliwal#1, Meghna Sharma.
[9] An Introduction to Neural Networks by Vincent Cheung and Kevin Cannons.

[10] Jiawei Han and Micheline Kamber, University of lllinois at urbana-champaign, Concepts and Techniques, Data mining, second edition(2006), [3]Barandela, R., Sánchez, J.S., García, V., Rangel, E. Strategies for Learning in Class Imbalance Problems. Pattern Recognition 2003, 36(3), pp.849-851
[11] K.P Soman, Shyam Diwakar, V.Ajay, data mining theory and practice, Amrita Vishwa Vidyapeetham, (2010).
[12] Data Mining: Theory and Practices by Dr.Shyam Divakar and Dr.K.P.Soman.
[13] Apriori Algorithm Review for Finals Presentation by SE 157B, Spring Semester 2007 Professor Lee By Gaurang Negandhi.
[14] PDF-"Using TF-IDF to Determine Word Relevance in Document Queries" by Juan Ramos, Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855
[15] Figure 1 – Schematic drawing made through Rational Rose.
[16] Figure 2 to 5- Resulting Snapshots obtained by implementing algorithms Naïve Bayes and KNN.
[17] David Posada,Taylor J. Maxwelland Alan R. TempletonTreeScan: a bioinformatic application to search forgeno-type/phenotype associations using haplotype trees, Variagenics, Inc., 60 Hampshire Street, Cambridge, MA 02139, USA and Department of Biology, WashingtonUniversity, St Louis, MO 63130-4899, USAReceived on November 5, 2004; revised on January 10, 2005; accepted on January 25, 2005Advance Access publication January 28, 2005, Vol. 21 no. 9 2005, pages 2130–2132 doi:10.1093/bioinformatics/bti293
[18] AdrienCoulet, MalikaSmaïl-Tabbone, Pascale Benlian, AmedeoNapoliand Marie-Dominique Devignes ,‖ Ontolo-gy-guided data preparation for discovering genotype-phenotype relationships, Address:KIKA Medical, Paris, F-75012, France,LORIA (UMR 7503 CNRS-INPL-INRIA-Nancy2-UHP),Vandoeuvrelès-Nancy, F- 54506, France andUniversi-té Pierre et Marie Curie - Paris6, INSERM UMRS 538 Biochimie-BiologieMoléculaire,Paris, F-75571, FranceEmail: AdrienCoulet* - adrien.coulet@loria.fr; MalikaSmaïl-Tabbone - malika.smail@loria.fr; Pascale Benlian - pascale.benlian@sat.ap-hop-paris.fr; Amedeo Napoli - amedeo.napoli@loria.fr; Marie-Dominique Devignes – marie-dominique.devignes@loria.frCorresponding author
[19] Andrei S. Rodin and Eric Boerwinkle," Mining genetic epidemiology data with Bayesian networks I: Bayesian networks and example application(plasma apoE levels)", Human Genetics Center, School of Public Health and Institute of Molecu-lar Medicine, University of Texas Health Science Center, Houston, TX 77030, USAReceived on September 24, 2004; revised on May 3, 2005; accepted on May 17, 2005Advance Access publication May 24, 2005, Vol. 21 no. 15 2005, pages 3273–3278 doi:10.1093/bioinformatics/bti505.
[20] Timothy D. O'Connor and Nicholas I. Mundy," Genotype–phenotype associations: substitution models to detect evolutionary associations between phenotypic variables andgenotypic evolutionary rate", Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK, Vol. 25 ISMB 2009, pages i94–i100doi:10.1093/bioinformatics/btp231
[21] Rui Chang and Wei Wang, "A Novel Knowledge-driven Systems Biology Approach for Phenotype Prediction Upon Genetic Intervention", Department of Chemistry and Biochemistry, University of California, San Diego, CA, USA.

AUTHOR PROFILE

Bipin Nair B J has completed his MCA from Amrita Vishwa Vidyapeetham, Amrita University, B.Sc. Industrial Chemistry and PG diploma in Medical Biochemistry from Kerala University. Now he is working as Lecturer, Department of Computer Science, Amrita Vishwa Vidyapeetham, Amrita University, Mysuru Campus.