International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 02 | Feb-2018 www.irjet.net p-ISSN: 2395-0072

Next Generation Sequences Analysis Using Pattern Matching Algorithm

M Mohamed Divan Masood[1], D Manjula[2]

[1] Research Scholar, Computer Science and Engineering, Anna University, Chennai.
[2]Head of the Department, Computer Science and Engineering, Anna University, Chennai.

Abstract - *The last decades Next Generation Sequences (NGS) analysis are interesting research topics in computation biology. The DNA gene sequences create huge volume of data so we need novel methodology for analyzing the sequences. In this paper suggested a Karp Pattern Matching algorithm (KPM) for analyzing the DNA sequences. First stage have solved the base version of application by using dynamic programming. Essentially the dynamic programming process based on window based process and before that our methodology have extend to identification of unidentified region from DNA sequences and K-mer algorithm have been used for preprocess stage. Our method have used to identify the most common gene sequences without lab experimental and has exposed the good classification accuracy.*

Key Words: Next Generation Sequences, M Karp Pattern Matching algorithm, Global Alignment Algorithm, FASTA, Pattern Matching Algorithm.

1. INTRODUCTION

Genetic sequence alignment is used to align sequences for inferring useful information from the arranged sequence [1]. Ancestral sequence shows more or less the same sequence repeats and using this technique many applications implemented in forensics and other fields. There are many ways to align sequences first to align sequence by making use of the scoring matrices by giving the gap penalty for them. Identical sequences may sometime represent characterization towards a particular disease or identification of a disease in earlier stage. Many alignment algorithms have been introduced to align the given sequence because aligning the sequence may yield regular repetitive patterns called motifs. Motifs are said to have a biological function and identifying them are necessary to know more about the sequence.

Motifs also has databases to store the obtained data for further reference because storing these data helps to identify diseases or other things easily and in the earlier possible stages. FASTA is used to find global optimal solutions and used along with global optimization algorithm-Needleman Wunsch [2], which is used to align sequences after searching sequences necessary to align them to obtain the best score. Multiple sequence alignments are useful to sort numerous sequences at a given time instead of sorting them one by one and the time can be reduced considerably. Dynamic Programming used along with this algorithm gives the best optimal solution for which without this the optimal solution is not possible.

Sequences present in the DNA provides various information about characterization of disease. Here first sequences are searched from the data through a quick process and then the optimal solution. Dynamic Programming methodology is also used to divide the problem into simpler sub problems and solve them which yields the result. The aligned sequences can be stored into databases for later use and for medical purposes because the sequence does not change much in case of diseases, while finding relationships between different members of the family is easy [3]. Database consists of possible sequence for diseases such as Down syndrome, Alzheimer disease etc., the obtained sequence may not be full accurate but to a certain extent because the sequence arrangement may not the same for all organisms. gens causing genes for blood, breast, leukemia etc., have been found and the development of drugs becomes easy once the sequence is found because once the sequence is found the drug are developed according to make the recovery speedy and act appropriately to where the sequence can be present [4].

Ongoing studies prove that no drugs have been introduced to directly change the sequence because it may lead to side effects where the given drug may act in any other part rather than acting on the particular diseased gene sequence. Drugs developed so far focuses on changing the protein formation or provide special antibodies which the proteins fail to produce due to changes in them. DNA provides proteins and if the proteins are not formed as expected then abnormal changes occurs within the organisms [15]. In this scenario drug development is most challenging process in computational biology. So we need novel method to analysis the gene sequences datasets. Deep Learning helps to identify the drug target.

In our paper consists as follows. Section2 define literature reviews. Section 3 defines system design of my proposed work. Section4 define experimental result and finally section 5 define conclusion.

2. Literature Survey

Blaisdell [5] introduced new method for analyzing the DNA sequence, mainly this method was first alignment methodology in computational biology. In this study have been used for calculated correlation among the two sequences. Initially the proposed methodology resolved the fast access from database to searching the sequences. Previously most of researcher are works on identifying the relationship among the DNA sequences [6]. The k-mer

frequencies methodology have been used for build the phylogenetic trees by traditionally at the same time their approaches worked based on multiple sequence alignment methodology [7]. In this approaches are also used for identified the relationship between RNA, DNA and proteins. However, the result was not satisfied.

A few important participations fall outside these broad categories. For example, DME composes a discriminative model by ordering list of perturbation of a PWM; ANN-Spec [8] learns a neural network; and DEME learns a discriminative PWM model with a conjugate gradient algorithm. None of these approaches gives statistical hypothesis testing or error estimation by taking test or cross-validation. Seeder method that reduces a suitable distance between a seed and a larger sequence and bears some similarity to the LR algorithm method introduced here [10]. Whereas hypothesis testing and false discovery rates are clearly says, there is no consideration of holdout-validation or cross-validation.

Dynamic Programming Algorithm is used to solve complex problems by breaking into simple sub-problems which is used in pattern searching. Various techniques have been used to find similar patterns. The root element is chosen and the algorithm is chosen, which gives the best performance using the matching problem so that the pattern is chosen with minimum matching cost. Root-element Best-Matching problem(R-BMP) along with dynamic programming DP-R_{BMP} divides the problem into simpler sub problems by constructing a matrix displaying the cell cost and the list of indices of matched elements. DP-R_{BMP} follows the criteria of early abandoning when the minimum cost is obtained and it is compared with other cells in the rows, thus filling the cells can be stopped when the minimum cell cost is obtained [11].

Windowed DP algorithm is used to improve the complexity of DP-R_{BMP} by increasing the number of columns strictly more than the number of rows and the binary search method is use to search for patterns which gives perfect fit and it can be used for large problems. The windowed (openness) size increases due to which the resolution of conflicts also increases. The optimal matching for patterns are low, so exact matching is found which in turn increases the cost. The base model is extended under four criteria used for various applications and in genomics. The first criteria includes the presence of genomic regions in the form of intervals which can be reduced using centroid. They can also be used in dissimilar regions without making changes in windowed algorithm, secondly multi-track patterns are used when a given pattern can be identified in more than one track and the next criteria uses partial and negative matching [12] [13]. In negative matching the pattern is not present or in other words the result is always negative it consists of a buffer to store the distance of the successive matches found in positive matching. A collection of these negative matching is called as valid area, partial matching is used to find missing elements and finally region attributes are used to find region distance between a couple of matched regions.

3. System Architecture

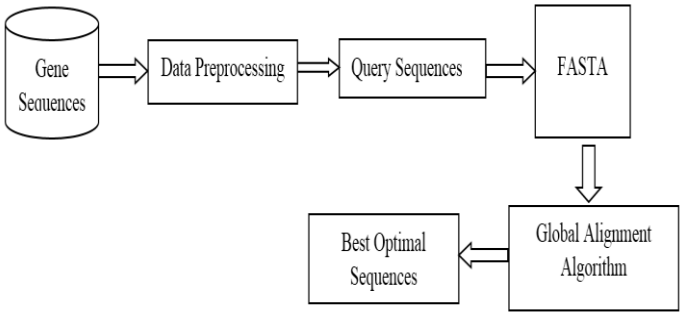

Figure 1 System Architecture for present work

The Fig 1 mentioned below shows the align sequence based on Global alignment algorithm. The genetic sequence obtained from the database is given as input and the query sequence is also given explicitly by the programmer which is based on the trial and error method because the given query may sometime obtain best results and vice versa. FASTA algorithm is used to search the given query in the database quickly and is useful for global alignment, once the query sequence is given it searches for the possible sequence in the database and returns all results which matches the query sequence. After obtaining the sequence the global alignment algorithm is used to align the sequence because it is use to match the sequence from end to end so this becomes the best when compared to local optimization where it matches only the particular sequence. Dynamic programming (DP) splits the problem into smaller sub problems and using it along with Needleman Wunsch algorithm leads to best optimal sequence because performing the algorithm along with DP though increases the time gives the best optimal result.

Algorithm 1 Karp Pattern Matching algorithm

KarpSet(s[1,2…….n], s means sequences

set ls:= blank Set

foreach s in ms

insert ls(su]ms[1..k]) into ls

LS: = function (s[1..k])

for i then 1 to n-k+1

if LS € ls & s [i..i+k-1] € ms

return i

LS := function(s[i+1..i+k])

Break

4. EXPERIMENTAL EVALUVATION

In our proposed method have been implemented on HP laptop with Intel i3, 1.33GHz processor with 8GB RAM. The Next Generation Sequences dataset has been stored in arff file format and it has taken from UCI machine learning repository. The first preprocessing stage, in this stage total 1GB dataset has been preprocessed. It has removed an irrelevant symbols and unmatched words except A, T, C and G. the data have been cleaned based on k-mer method. The assessed DNA sequences size was 625Mb, which is size calculated by given formula (1),

Sequences Size= K-mer /Peak Depth ……. (1)

Likewise, the positive replication frequency can basis of replication peak in total number of peak frequencies. Fig 2 is different between frequencies and depth of the total number of peak frequencies using the k-mer distribution methodology.

where, depth defines the sequences in the series and it has changed based on the size of the datasets. Present method has been used 1GB dataset. Our method has given good accuracy compared to previse method.

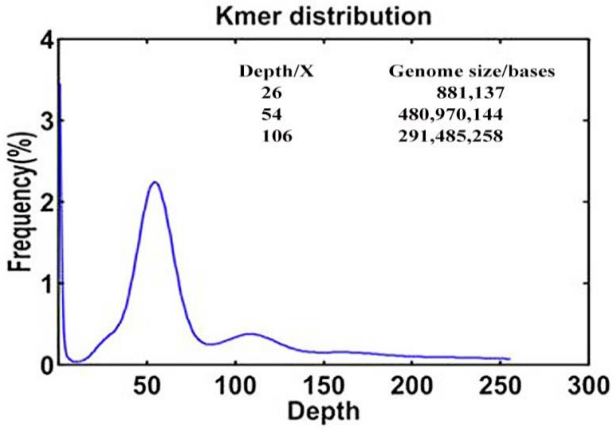

Figure 2 Analysis of result between Depth and Frequency

Sequence gathering and GC content analysis

Previously normal GC result have been satisfied of sequences in values of 38.64% but our method has been taken 40.25% sequences frequencies by using k-mer methodology. GC content may cause sequence bias on the Illumina sequencing platform, thus seriously affecting genome assembly [15]. Furthermore, the GC depth was somewhat congested, showed in fig 3. Maybe only one of the two sets of homologous chromosomes in the diploid was assembled, which resulted in the emergence of the lower layer and homologous chromosomes have assembled in lower layer [12].

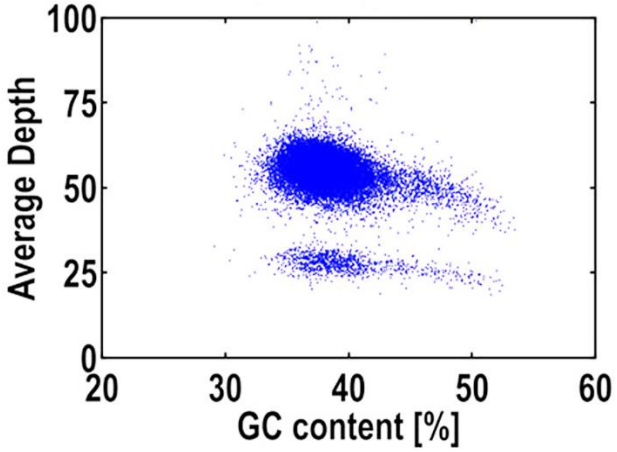

Figure 3. Shows the average of sequences depth based on GC content

5. CONCLUSION

The best optimal sequence is obtained which matches fully with the given sequence and from this the inference can be made as such it may belong to the same ancestor or the sequence found may be the relevant one. Giving the diseased query and searching them in the database may obtain the result sometime because the same diseased sequence or the relevant one can be found which may also detect the early stage of a disease, it's better to test with diseased sequence.

REFERENCES

[1] Durbin, Richard M.; Eddy, Sean R.; Krogh, Anders; Mitchison, Graeme (1998), Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (1st ed.), Cambridge, New York: Cambridge University Press, doi: 10.2277/0521629713, ISBN 0-521-62971-3, OCLC 593254083.

[2] Needleman, Saul B and Christian D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins." Journal of molecular biology 48.3 (1970): 443-453. H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," Bioinformatics, vol. 25, no. 14. pp. 1754–1760, 2009

[3] P. Chen, C. Wang, X. Li, and X. Zhou, "Hardware acceleration forthe banded Smith-Waterman algorithm with the cycled systolic array, in Proc. Int. Conf. Field-Programmable Technol., 2013, pp. 480–481.

[4] Lipinski, Christopher A., et al. "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings." Advanced drug delivery reviews 64 (2012): 4-17.

[5] B.E. Blaisdell, "A Measure of the Similarity of Sets of Sequences not Requiring Sequence Alignment," Proc. Nat'l Academy of Sciences USA, vol. 83, no. 14, pp. 5155-5159, 1986.\

[6] A. Globerson, G. Chechik, F. Pereira, and N. Tishby, "Euclidean embedding of co-occurrence data," J. Mach. Learn. Res., vol. 8, pp. 2265–2295, Oct. 2007.

[7] B. Harris, A. C. Jacob, J. M. Lancaster, J. Buhler, and R. D. Chamberlain, "A banded smith-waterman FPGA accelerator for mercury BLASTP," in Proc. Int. Conf. Field Programmable Logic Appl., 2007, pp. 765–769

[8] Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011.

[9] T. Wu, Y. Huang, and L. Li, "Optimal Word Sizes for Dissimilarity Measures and Estimation of the Degree of Dissimilarity between DNA Sequences," Bioinformatics, vol. 21, no. 22, pp. 4125-4132, 2005.

[10] A. Globerson, G. Chechik, F. Pereira, and N. Tishby, "Euclidean embedding of co-occurrence data," J. Mach. Learn. Res., vol. 8 pp. 2265–2295, Oct. 2007.

[11] A. L. Sullivan, C. Benner, S. Heinz, W. Huang, L. Xie, J. M. Miano, and C. K. Glass, "Serum response factor utilizes distinct Promoterand Enhancer-Based mechanisms to regulate cytoskeletal gene expression in macrophages," Mol. Cell. Biol., vol. 31, pp. 861–875, Feb. 15, 2011.

[12] Li Y, Hu Y, Bolund L, Wang J. State of the art de novo assembly of human genomes from massively parallel sequencing data. Hum Genet. 2010; 4: 271–277.

[13] C. T. Ong and V. G. Corces, "CTCF: an architectural protein bridging genome topology and function," Nat. Rev. Genet., vol. 15, no. 4, pp. 234–246, Apr., 2014.

[14] G. D. Erwin, N. Oksenberg, R. M. Truty, D. Kostka, K. K. Murphy, N. Ahituv, K. S. Pollard, and J. A. Capra, "Integrating Diverse Datasets Improves Developmental Enhancer Prediction," PLoS Comput. Biol., vol. 10, no. 6, p. e1003677, Jun., 2014.

[15] C. Cortes and M. Mohri, "AUC optimization vs. error rate minimization," in Proc. Adv. Neural Inf. Process. Syst., 2004, vol. 16, pp. 313–320.