ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 9, September 2016

Cancer Detection via Pattern Matching

[1]Dushyant, [2]Jaisalsingh Pawar, [3]Nilesh Suryawanshi, [4]Vivek Tigga, [5]Prof. Vaibhav Tumane
[1] Student, [2] Student, [3] Student, [4] Student, [5] Professor, Dept. of Computer Science and Engineering,
Nagpur Institute of Technology

***Abstract:* — Pattern matching is an important task of the pattern discovery process in today's world for finding the structural and functional behavior in proteins and genes. Cancer is mainly cause because of the genetic disorder. This disorder in genes affects the persons DNA which changes the pattern of its DNA sequence. We are using pattern matching technique in this project to detect whether the patient having cancer possibility or not. here We are using Knuth Morris Pratt (KMP) and Brute Force (BFS) Algorithm. By comparing the patient mutated DNA with the available infected cancer DNA pattern and after examining the match percentage we can determine the possibility of cancer. The latest and most effective algorithm is used in this project for pattern matching.**

***Index Terms*— Cancer, KMP Algorithm, Brute Force Algorithm, Pattern matching and genes**

I. INTRODUCTION

DNA are the particularly encoded genetic information which guide the functionality of different cell in body. They are the long molecule that looks like a twisted ladder. it is made of four type of letter which represent the range of DNA which are the repeating units in DNA. Thus latter are called nucleotides. Thus nucleotides are the repeating unites in DNA sequence. There are four types of nucleotides (A, T, G AND C). Based on thus sequences different information are carries. Another term is genes, a segment of DNA. Genes are like sentences made of the “letters” of the nucleotide alphabet, between them genes direct the physical development and behavior of an organism. Genes are like a information or instruction that are provided to organism so that depending upon that, the function of organism is depended. In general the DNA is the instruction or specifications which guide the development of cell within body. But DNA independently can’t directly involve in cell division or development. For that purpose another molecular structure which is RNA is user. The RNA are very much similar to DNA just the difference is between there functionality. As specified the DNA are gnarly provide specification for the cell development whereas the RNA are directly involved in cell development. For that purpose the RNA get the instruction from the DNA, based upon that instruction the RNA produce the different proteins. But still the RNA individually not involved in cell or protein development. Furthermore the RNA is again classified into various sub groups, all of them having different responsibility. Such as mRNA, tRNA, rRNA, microRNA. mRNA are sad to be messenger RNA which hold the genetic information that direct synthesis of specific proteins. Many viruses encode their genetic information using RNA

Genome Similarly there is different functionality for micro RNA. The micro RNA is a small non coding RNA molecule found in plants, animal and some viruses, that functions in RNA silencing and post transcriptional regulation of gene expression. When this co-ordination release with improper way then this may result into unwanted behavior. Cancer is somehow result of improper co-ordination between DNA and RNA. The micro RNA is an essential component which supports the RNA molecule. They are basically encoded RNA molecule which is use to neutralizing the RNA behavior in essential condition. When there is improper behavior of body take place, than they may result in DNA mutation. In the tumor of cancer tumor, the DNA of the affected area gets changed. This results in different DNA structures. Our main intention is to identify this irregular DNA structure to identifying the possibility of cancer. As the cancer affected tissue contain irregular DNA structure. These structures shows different pattern. This pattern could be taken as input to determining the possibility of the cancer in the affected tissue from which the mutated DNA. Here the technique is to provide the mechanism which matches the patient DNA with the mutated DNA patter in the terms to finding the similarity between them.

II. BIOPSY

As for pattern evaluation we need data which is provides as input to the system which is DNA string. For extracting DNA from patient we need the procedure which is called biopsy. In which the small portion of tissue taken from the affected area which is further been processed. As in the terms of cancer the affected area or tumor may reflect different structure in DNA string representation. This DNA required to be extracted from there. This tissue is further evaluated to get the DNA string which is further been used in this system.

The other part of system which process the mutated DNA. This DNA is collected from the tissue which reflects the cancer. The system contain different DNA string from cancer tissue so that the system having large array of reference to compare the possibility of cancer from the given DNA string.

III. OBJECTIVE

- ❖ Determinacy of the presence of cancer cell in the affected area by the means of matching DNA patterns is our AIM.
- ❖ Pattern Matching shall be taking place using single pattern matching algorithms.

Literature Review

1. A Survey of Human Cancer Classification,
Int. J. Comp. Tech. Appl., Vol 2 (5)

2. Micro RNAs and Cancer,
IEEE transactions on information technology in biomedicine, Vol. 13, No. 1

3. FSVML and GA-FSVML wrapper approaches for gene selection,
International Journal of Genetics, Vol. 4, Issue 2

IV. METHODOLOGY

The main intention behind this system is to find the similarity between these mutated DNA and input DNA string. we had a DNA sequence which are 96% mutated which result to cancer. This data is used as pattern for representing cancer affection. Our system will find match between thus mutated pattern go that of the input DNA of patients.

There are many different pattern matching algorithm which can possess and find the specific pattern from the given string. Previously there is brute force algorithm which is very simple to understand and to implement. Thus algorithm work sequentially by matching pattern to individual string character. Thus are very simple to implement and to understand. But the difficulty is within the operational phase. The brute force algorithm required much larger time as compared to other pattern matching algorithm. Thus the brute force algorithm having complexity which is O (m.n) which is very high.

KMP ALGORITHUM

```
algorithm kmp_search:
    input:
        an array of characters, S (the text to be searched)
        an array of characters, W (the word sought)
    output:
        an integer (the zero-based position in S at which W is found)

    define variables:
        an integer, m ← 0 (the beginning of the current match in S)
        an integer, i ← 0 (the position of the current character in W)
        an array of integers, T (the table, computed elsewhere)

    while m + i < length(S) do
        if W[i] = S[m + i] then
            if i = length(W) - 1 then
                return m
            let i ← i + 1
        else
            if T[i] > -1 then

                let m ← m + i - T[i], i ← T[i]
            else
                let m ← m + 1, i ← 0

    (if we reach here, we have searched all of S unsuccessfully)
    return the length of S
```

OUTPUT

```
             1         2
m: 01234567890123456789012
S: ABC ABCDAB ABCDABCDABDE
W:             ABCDABD
i:             0123456
```

BRUTE FORCE ALGORITHUM

```
c ← first(P)
while c ≠ Λ do
 if valid(P,c) then output(P, c)
 c ← next(P,c)
end while
Output
```

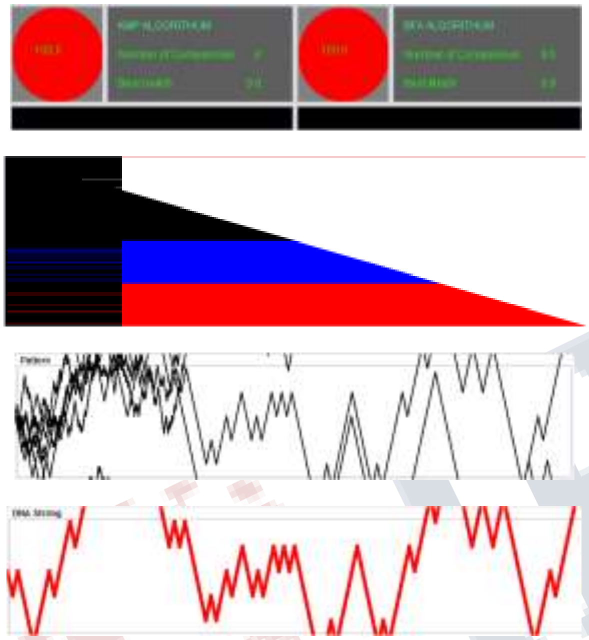


Figure 2.1 shows the graphical representation for the output of program

For preventing from that KMP algorithm is used. Which has advantages over the execution time and complexity. The Knuth Morris Pratt (KMP) algorithm having complexity which is O (m+n) which is very low as compared to brute force algorithm. There on such system KMP algorithm is going to implemented which result in reduced executions time. Here the main intention is to provide less execution time or complexity and to provide faster result. As KMP algorithm is used for pattern matching to find the substance of DNA which reflect the systems of cancer. The KMP algorithm works on two parts. Each for preparing the pattern as well to find the match of pattern go n given string. Tue two part of KMP are prefix operation and the other is pattern matching process.

A).Prefix operation

This is the first process which deal with the preparation of pattern for further evaluation. Here on to find the systems of cancer this phase will get the input from the mutated DNA string and prepare the pattern. These patterns are somehow logical consequences which reflect the mutated DNA string. As the patterns are prepared this pattern are been provided for further phases as the system contain data set of mutated DNA as provided which collected from the cancer tissues, which reflect the 96% of mutation in DNA when reflect the cancer possibility. Based upon this dataset further evaluation of patients DNA is take place. The prefix phase evaluates this DNA dataset to generate the pattern for further evaluation.

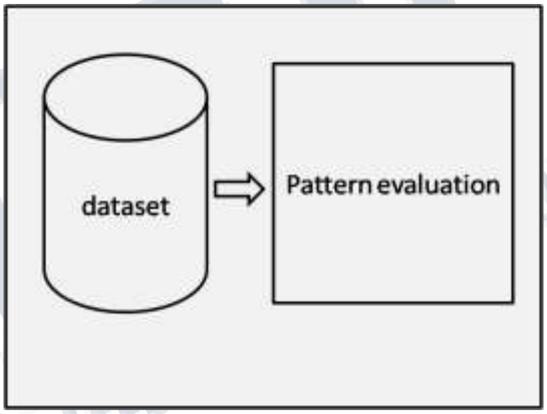
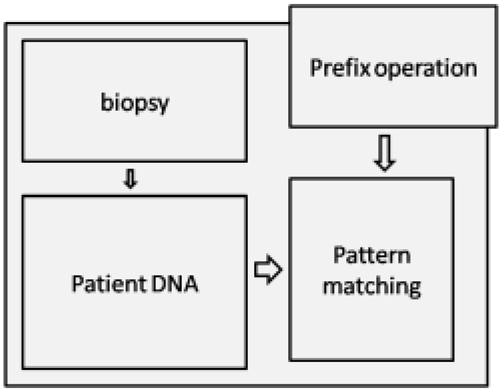


Figure 2.2 show the oprational structure of the system (Pattern evaluation)

B) Pattern matching phase

This phase work on real action where the persistent DNA are matched against the pattern which are evaluated at previous phase. There is simple approach for finding the pattern from the given string is just by comparing each substring to pattern. But this technique result in long processing time and increase complexity. But while we use KMP algorithm we actually increase the no of shift between each caparison. As the no of software reduced that the execution time also get reduced. As the data collected from prefix operation which represent the logical representation of sequences. These patterns are used to increase the no of shift between each comparison. As the no of shift are increased the required execution time gets reduced. Due to which the system can process much larger dataset for approximately finding the possibility of cancer from the reference of DNA which reflect the exact possibility of cancer.



Figur 2.3 Show the oprational structure fo the system(Pattern maching)

As the no of match percentage increase the possibility of cancer is increases respectively. As the system operates of the given data they may provide several results in percentage. But if the match percentage is greater than 85% then there is cancer absolutely. If the percentage drops below the certainly level that of the possibility of cancer reduces respectively. If the match percentage drops below the 50% then there is no cancer.

FUTURE SCOPE

1) The system could be real time pattern matching enhanced system

2) This system can be made hand held by developing the same for smart phones

3) Any other Algorithm better than KMP(if develop) can also be implemented for much better result

4) Advance system that can created the physical appearance of affected person for the after effects

V. CONCLUSION

The result of the KMP over BFS Algorithm is impressive and effective. DNA Pattern Matching using KMP Algorithm is less complex and gives better performing system compare to BFS Algorithm.

Hence matching DNA patterns for the possibility of CANCER disease using KMP Algorithm is scrutinized and effective.

REFERENCES

[1]. revethyvN. And Balasubramaniam," FSVML ad A-FSVML wrapper approaches for gene selection and classification using expression of very few genes",April 12,2012.

[2]. Rui xu,georgios c. Anognoslopaulasban Donald c. Wunsch,"Mukticlass Cancer classification using semi supervised ellipsoid ARTMAP and particle swarm optimization with gene expression data",val 4,no 1, January-march 2007

[3]. Anoslasis Oulas, Martin reczku, and Panayiata pairazi,"microRNA and cancer the search begins", val 13,no -1, January 2009.

[4]. Francisco azuqje,"making nenome expression data meaningful prediction and discovery of classes of cancer through a connectionsinst learning approach.

[5]. Thalia AFarazi, jessica , spirit zer, Pavel morozov and Thomas tuschi,"miRNA in human cancer",18 November 2010.