OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

"'X' Chromosomal Miner"

K. Sri Rama Reddy[1], Prof. G.V. Padma Raju[2]

[1]2/2 M.Tech(C.S.T) ,[2]Professor and Head of C.S.E, S.R.K.R Engineering College, Bhimavaram, Andhra Pradesh, India

ABSTRACT:- *Micro Satellites are helpful in identifying several diseases in early stages and disorders in the human based on his/her genome. Micro Satellites extraction became more crucial in the modern world. Several Micro Satellite Extractors exist and they fail to extract microsatellites on large data sets of giga bytes and tera bytes in size. This "'X' Chromosomal Miner" tool can extract both Perfect as well as Imperfect Microsatellites from large data sets of human genome 'X'.*

Keywords:- *Chromosome, DNA, Extraction, Micro Satellite, Nucleotide*

I. INTRODUCTION

Bio-Informatics is the term comprising of two words. Bio means life. Informatics refers to develop methods and software tools for understanding biological data. It mainly involves processing the genetical information of the target person to accomplish the task or working with the genetical information to explore new information.

It deals with the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data. Application Areas of Bio Informatics include Crime Investigation, DNA Testing and Syndrome Test by analyzing the Chromosomes, Unique Identification of a Person.

II. BACK GROUND OR HISTORY

Human body is made up of cells. Each cell contains DNA which is a helical Structure of bases namely A, T, C, G. DNA contains about 3 billion bases and more than 99% of bases are same in all people. The arrangements of these bases determine the maintenance of a person. 'A' pairs with 'T' and 'C' pairs with 'G'. Each base is attached to Sugar molecule and a phosphate molecule. Nucleotide is the combination of base and sugar molecule and a phosphate molecule. DNA has the tendency to replicate. Each strand of DNA in helical structure can serve as a pattern for duplicating sequence of bases. DNA Molecule is packaged into thread like structures called Chromosomes. Each cell contains 23 pair of Chromosomes. 22 pairs called autosomes look same in all people. 23 rd pair distinguishes the sex. A male has XY pair while the female has XX.

Genes are made of DNA. Genes contain information needed to make functional molecules called proteins. Genes play an important role in the production of proteins. The below are the two major steps in protein building.

Transcription:

The information stores in gene's DNA is transferred to similar molecule called RNA in cell nucleus. Now this Messenger RNA (m-RNA) that carries information for making a protein carries information from out of the nucleus into cytoplasm.

Translation:

It takes place in Cytoplasm. m-RNA interacts with specialized complex called ribosome that reads sequence of m-RNA bases. Now, the energy that the human takes through food are build into amino acids. These amino acids are grouped together to form a protein.

So, Genes play a vital role. Any mutation in the gene effect this process and leads to genetic disorders in the evolving species. The possible gene mutations are insertion, deletion and duplication of bases. These mutated genes are to be subjected to gene-therapy by replacing the gene or introducing new gene.

DNA contains coding and Non-coding regions. The region of DNA that is responsible for building of proteins is the Coding Region. Non-Coding Region is inherited from ancestors. Almost 98% of the DNA is Non-Coding.

DNA consists of simple sequence repeats (SSR) called Micro satellites generally size of 1-6 bp. These are abundantly found in coding region. SSR expansion or contraction may lead to loss or gain of gene function. The below Table I are the some of the disorders identified due to the expansion or contraction of the corresponding SSR in the 'X' Chromosome.

Micro Satellite Repeat	Disorder
CAG (Expansion)	Breast Cancer
AGT	Regulates gene Translation
CAAT	Mediating Phase Variation to adapt to host environmental Changes
$(A)_n$	Inactivates MMR genes and cause Human Cancer, Suppress tumour
GCAA, TTTA	Viral Genes
TATA	Hemophilia (Slows Blood Clotting)

Table I: Disorders caused by Micro Satellites

2.1 Classification of Micro Satellites:

Micro Satellites are further classified into two types.

Perfect: Continuous repetition of the bases without any substitutions

Examples:

ATATATAT, AAAAAAAA, CAGCAGCAGCAG

Im-Perfect: Performing Substitutions results in Continuous repetition of bases and more chance of gene mutation.

Examples:

1) ATAGACAG ('AT' repeat with substitutions =3)
2) AAATTTTA ('A' repeat with substitutions=4)
3) CAGTAGCATCAG ('CAG' repeat with substitutions=2)

III. MOTIVATION

This Perfect or Imperfect Micro Satellites extraction is the important task. Initially, Microsatellites are extracted. When the person faces any health problem, we can compare the newly extracted Micro Satellites with the previously extracted. It will be easy to identify which Micro Satellites were repeated and what would be the effect of their repetition. So, we can identify the disease in the earlier stages and subject the person to gene-therapy. We can take some precautions that helps in expanding the life time of the person.

IV. LITERATURE

Variations in SSR (expansions and contractions) in protein coding regions results in gain or loss of gene function [1]. The effect of SSR variations in un-translated (UTR) Regions is explained. Various effects are transcription slippage and produce expanded mRNA. Triplet SSRs located in the UTRs induce gene silencing. Various repeats and their effects on humans are explained. It provides a molecular basis for fast adaptation to environmental changes in both prokaryotes and eukaryotes. We came to know the importance of Micro Satellites [1].

The importance of Micro Satellites in 'X' Chromosome is explained [2]. Two individual daughters share the same micro satellites of the father in the non coding region. Sons inherit their 'X' chromosome from their mother. It tells the inner fact that to test Paternity, even if the father is no more, they can perform the test by matching daughters 'X' Chromosome with the Putative grandmother's 'X' Chromosome. So, Micro Satellites in 'X' Chromosome are also useful for Paternity testing. We came to know the importance of 'X' Chromosome [2].

They have developed a simple to use web software, called WebSat, for microsatellite molecular marker prediction and development. WebSat can be accessible through the Internet, requiring no program installation [3]. It makes use of Ajax techniques, providing a rich, responsive user interface. It allows the submission of sequences, visualization of microsatellites and the design of primers suitable for their amplification. This program allows full control of parameters and the easy export of the resulting data, thus facilitating the development of microsatellite markers [3].

Simple exact tandem repeats as well as non-simple repeats are found [4]. It has several advanced repeat search parameters/options compared to other repeat finder programs as it not only accepts GenBank, FASTA and expressed sequence tags (EST) sequence files. The minimum and maximum tandem repeat motif lengths that E-TRA finds vary from one to one thousand. It allows the researchers use different minimum motif repeats search criteria for varying motif lengths simultaneously. The results obtained indicated that 12.44% (679,800) of the human EST sequences contained simple and non-simple repeat string patterns varying from one to 126 nucleotides in length. The results also revealed that human organs, tissues, cell lines and different developmental stages differed in number of repeats as well as repeat composition, indicating that the distribution of expressed tandem repeats among tissues or organs are not random, thus differing from the un-transcribed repeats found in genomes [4].

Study on evaluation of 5 different Algorithms (TRF, Mreps, Sputnik, Star and Repeat Masker) [5]. It had shown that parameters change can influence micro satellite distributions in these Algorithms. These five algorithms were compared by fixing the parameter values. Now, Micro Satellites were observed [5].

A user-friendly Web application developed to minimize tedious manual operations and reduce errors [6]. This tool facilitates the integration, analysis and display of sequence data from SSR-enriched libraries [6].

V. METHODOLOGY

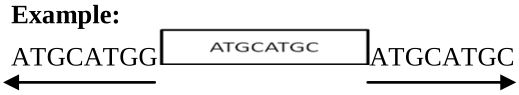
The Algorithm for "X Chromosomal Miner" takes the input file line by line and processes the line and extracts the Micro Satellites present in that line. When it takes input as the next line, it appends the last 6 characters of the previous line with the current input line. As a result, large no of Micro Satellites are extracted and it works even on the file of large size.

"**X Chromosomal Miner**" works on two step procedure.

STEP-1:

It searches for type-1 Satellites. Type 1 Satellites are those in which searching is done on both sides of the key string to find repeats. Necessary substitutions are made based on the user imperfection limit. It also extracts Perfect Micro Satellites.

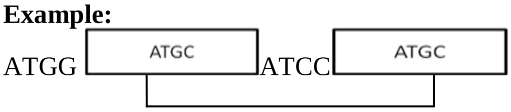
Example:



STEP-2:

It searches for type-2 satellites. Type 2 Satellites are those in which searching is done on skipped matching. The words that are skipped are verified for substitutions or imperfections. If there are perfect Micro Satellites, no need of skipping.

Example:



This tool stores the Micro Satellite information in the summary file that contains information about Micro Satellite name, iteration count, tract size, line number and percentage of A,T,C,G in the corresponding Micro Satellite.

As these input files differ in size for different people, this "X Chromosomal Miner" can be a useful tool that works on large size even gb's or tb's of input file and extract large number of Micro Satellites that are useful in identifying the neural diseases and cancers or genetic disorders in the pre-early stage. It also saves the memory space as it allows storing only one line of the input file at a time and over writes it the next time.

Central Algorithm Working Steps:

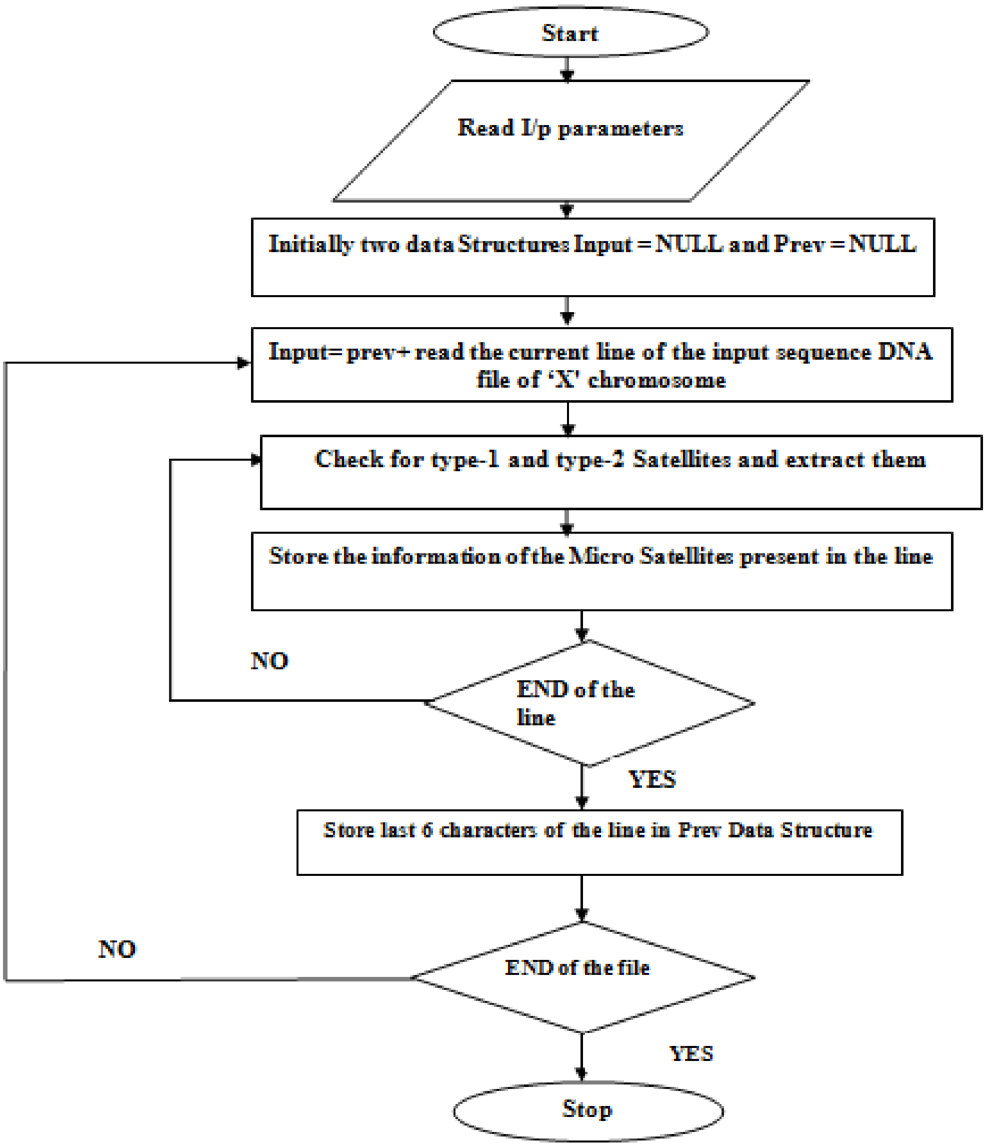
```
Step1: Start
Step2: Read input Parameters from the user
Step3: Initially 2 data Structures Input=NULL and prev=NULL
Step4: Input= prev+ (Read the line from 'X' Chromosomal i/p file)
Step5: Check for type-1 and type-2 Satellites in the i/p line
Step6: Extract Micro Satellite information and store the information
Step7: Check whether ending of line. If no, go to step5 else go to step8
Step8: Store the last 6 characters of the i/p line in the Prev Data Structure
```

Step9: Check whether ending of file. If yes, stop else go to step 4
Step10: Stop

Algorithm to Search for Type-1 and Type-2 Satellites:

Step 1: Start
Step2: Accept Minimum No of Substitutions, Minimum Repeat No from User, Imperfection Percentage
Step3: Check whether the motif of length=6 is a mono, di, tri nucleotide
Step4: If yes, check for next sequences whether it is such type of nucleotide and store that information. If not, go to step 5.
Step 5: Check Whether Nucleotide is tetra, penta or hexa.
Step6: Based on type of Nucleotide, compare with the next sequence. If matched, go on comparing. Store the Microsatellite Information.
Step7: Now expand on both sides of sequence
Step8: compare two regions to check whether they match or not. If match, store the Micro Satellite Information. If not, check for substitution and store the Micro Satellites that satisfy minimum no of substitutions.
Step 9: Stop

The flow chart for the Central Algorithm is explained in the Fig. 1



Start
Read I/p parameters
Initially two data Structures Input = NULL and Prev = NULL
Input= prev+ read the current line of the input sequence DNA file of 'X' chromosome
Check for type-1 and type-2 Satellites and extract them
Store the information of the Micro Satellites present in the line
NO
END of the line
YES
Store last 6 characters of the line in Prev Data Structure
NO
END of the file
YES
Stop

Figure 1: Flow Chart for the Algorithm

Algorithm to find whether a motif is mono or di or tri nucleotide:

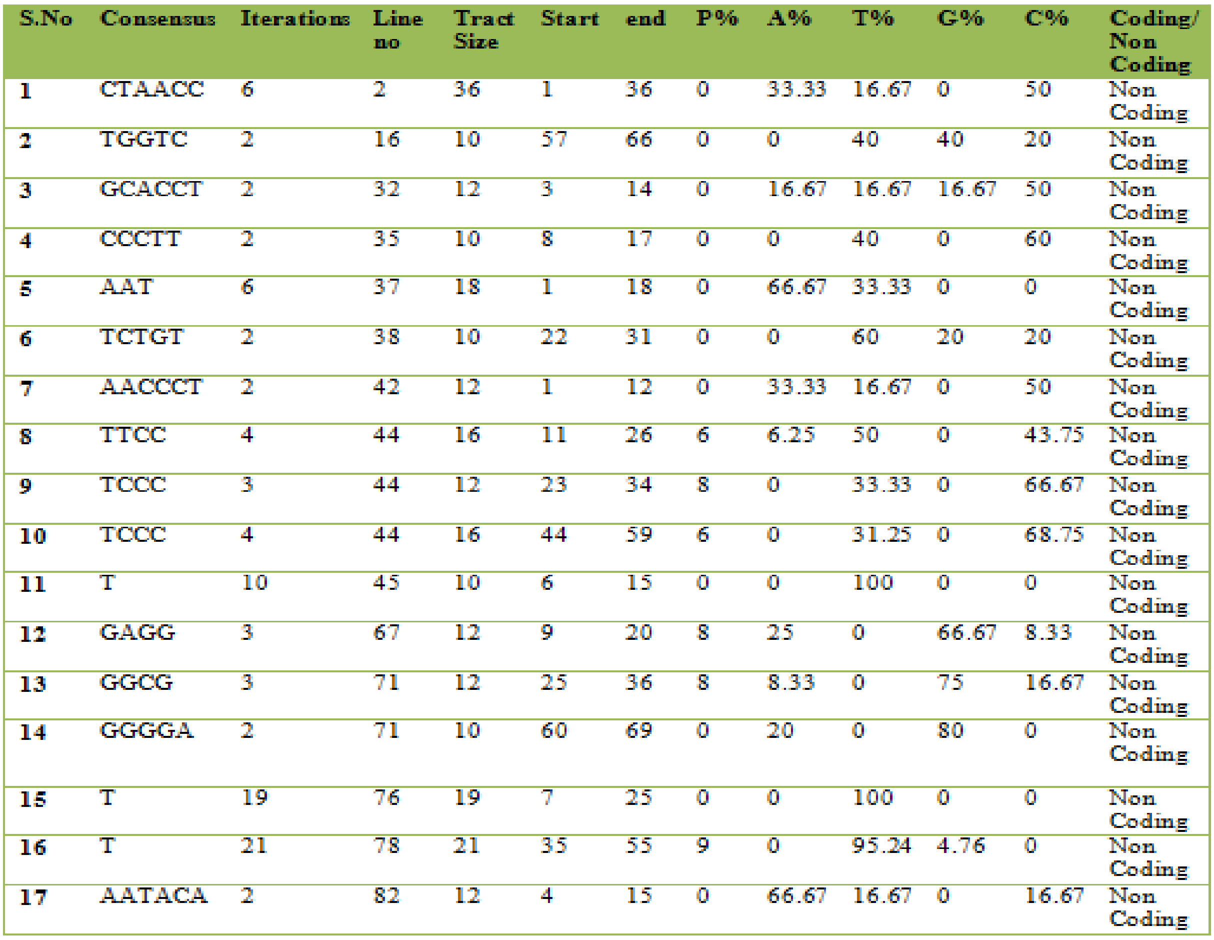
```
Step1: start
Step2: Get the length m of the satellite to check
Step3: if m==2,3,4,5,6 check whether it is mono. If so go to step7
Step4: if m==6, check whether it is di or tri nucleotide. If so go to step6
Step5: if m==4, check whether it is tri nucleotide. If so go to step6
Step6: Stop
```

Algorithm to compare 2 sequences:

```
Step1: Start
Step2: Accept the input parameter k (up to what length we have to compare)
Step3: Store the string up to length k in a string str1 every time incrementing k.
Step4: Store the string continuing from k in string str2 every time incrementing k.
Step5: compare str1 and str2
Step6: If matched, they are micro satellites. If not, continue.
Step7: stop
```

Algorithm to check for Substitution:

```
Step1: Start
Step2: Accept the input patterns and substitution limit given by the user as 'k'
Step3: Compare patterns based on index. If matched go to step4 else go to step5. If reached the end of the
        pattern go to step 6.
Step4: Increment their index and go to step 3
Step5: Increment number of mismatches 'r' and go to step4
Step6: if(r>k) i.e No of Mismatches substituted exceeds substitution limit, discard the satellite. Else, consider
        it as micro satellite.
Step7: Stop
```



S.No	Consensus	Iterations	Line no	Tract Size	Start	end	P%	A%	T%	G%	C%	Coding/ Non Coding
1	CTAACC	6	2	36	1	36	0	33.33	16.67	0	50	Non Coding
2	TGGTC	2	16	10	57	66	0	0	40	40	20	Non Coding
3	GCACCT	2	32	12	3	14	0	16.67	16.67	16.67	50	Non Coding
4	CCCTT	2	35	10	8	17	0	0	40	0	60	Non Coding
5	AAT	6	37	18	1	18	0	66.67	33.33	0	0	Non Coding
6	TCTGT	2	38	10	22	31	0	0	60	20	20	Non Coding
7	AACCCT	2	42	12	1	12	0	33.33	16.67	0	50	Non Coding
8	TTCC	4	44	16	11	26	6	6.25	50	0	43.75	Non Coding
9	TCCC	3	44	12	23	34	8	0	33.33	0	66.67	Non Coding
10	TCCC	4	44	16	44	59	6	0	31.25	0	68.75	Non Coding
11	T	10	45	10	6	15	0	0	100	0	0	Non Coding
12	GAGG	3	67	12	9	20	8	25	0	66.67	8.33	Non Coding
13	GGCG	3	71	12	25	36	8	8.33	0	75	16.67	Non Coding
14	GGGGA	2	71	10	60	69	0	20	0	80	0	Non Coding
15	T	19	76	19	7	25	0	0	100	0	0	Non Coding
16	T	21	78	21	35	55	9	0	95.24	4.76	0	Non Coding
17	AATACA	2	82	12	4	15	0	66.67	16.67	0	16.67	Non Coding

Table II: Micro Satellite Information

VI. IMPLEMENTATION

This tool has been developed in Java that is platform dependent. A java interface has been developed to interact with the user. The user can set the parameters like minimum number of repeats, minimum number of substitutions and imperfection percentage. The user can input the 'X' Chromosomal file of the Patient that can be of any size. The Parameters can be changed as per User requirements. The output is Summary file that contain information about Micro Satellites, their tract size, their type etc. The output file is available in html format.

VII. RESULTS AND DISCUSSION

Case Study:

We performed testing by taking input as the Human 'X' Chromosome. We however, performed testing on different input sizes.

Results:

This 'X' Chromosomal Miner can work on input of any size while the other tools fail to work. Most of the tools read the patient input file in an array. Now depending on the system RAM requirements, different systems accept different input sizes. This marker process line by line and to see that no satellites are missed at the intersection of the lines, it extracts at the intersecting part.

The above Table II is the output/summary file that gives the information about the microsatellites extracted. It is an HTML file.

This marker has the ability to work on any input file size ranging from gb to tb.

VIII. CONCLUSION

This "X Chromosomal Miner" extracts both Perfect and Imperfect Micro Satellites and can be used as a marker tool in Bio-informatics. It can extract Micro Satellites from the file of any size. The other tools that are available can work up to a fixed file size. It is flexible and user interactive. It provides a flexible environment for the user by allowing him/her to set the mutation limits. The future work includes storing the left and right flanking regions and standardization of compound micro satellites.

REFERENCES

Journal Papers:

[1]. You-Chun Li, Abraham B. Korol, Tzion Fahima and Eviatar Nevo, "Micro Satellites with in genes: Structure, function and evolution"

[2]. J.Edelmann, R.Lessig, M.Klintschar, R.Szibor, "Advantages of X-Chromosomal Micro Satellites in deficiency Paternity testing"

[3]. Wellington Santos Martins1, Divino Cesar, Kelligton Fabricio de Souza and David John Bertioli3, "WebSat - A web software for microsatellite marker development

[4]. Mehmet Karaca, Mehmet Bilgen , A . Naci Onus, Ayse Gul Ince and Safinaz Y. Elmasulu, "Exact Tandem Repeats Analyzer (E-TRA): A new program for DNA sequence mining"

[5]. Sebastien Leclercq, Eric Rivals and Philippe Jarne, "Detecting microsatellites within genomes: significant variation among algorithms".

[6]. Alexis Dereeper, Xavier Argout, Claire Billot, Jean-François Rami and Manuel Ruiz, "SAT, a flexible and optimized Web application for SSR marker development"