GENEXPLORER: AN INTERACTIVE TOOL TO STUDY REPEAT GENE SEQUENCE IN THE HUMAN GENOME

by

DEVANGANA KAR

(Under the Direction of Eileen T. Kraemer)

ABSTRACT

A large part of the human genome is made up of repeating elements. Out of these repeating elements segmental duplicates or repeating gene patterns constitute 5% of the total human genome. Researchers are investigating the relationship between the repeats in the genome to identify whether these repeats are identical by coincidence or whether some biological process duplicates them. The GeneXplorer tool is designed to aid the researchers to view the repeating patterns in the whole Human genome and to study the repeats in a chromosome or different chromosomes simultaneously. The genome can be magnified and the sequences viewed in detail. The details of the sequence in the gene level can be examined and area can be studied using UCSC Human Genome Browser. User can also select a sequence from various overlapping sequences present at a particular chromosome location.

INDEX WORDS: Chromosome, Gene, Human genome, Visualization, Repeat sequence

GENEXPLORER: AN INTERACTIVE TOOL TO STUDY REPEAT GENE SEQUENCE IN THE HUMAN GENOME

by

DEVANGANA KAR

BTech, Utkal University, India, 2005

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2008

GENEXPLORER: AN INTERACTIVE TOOL TO STUDY REPEAT GENE SEQUENCE IN THE HUMAN GENOME

by

DEVANGANA KAR

Major Professor: Eileen T. Kraemer

Committee: Maria Hybinette
Daniel M. Everett

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2008

DEDICATION

I would like to dedicate this thesis to my family for encouraging me to follow my dreams.

To Chandresh for his constant help and support.

To my friends for their constant support.

ACKNOWLEDGEMENTS

I would like to give my thanks to Dr. Eileen Kraemer for guiding me through the whole process of this thesis.

I would also like to thank Dr. Shaying Zhao for explaining the complexity of the human genome in simple terms so that I could understand and appreciate the subject.

I special thanks to Dr. Mucheng Zhang for taking the time out to explain the working of his algorithm and finding the time whenever I had a question.

Last but not the least I would like to acknowledge all my professors and colleagues for helping me and encouraging me throughout the period of my thesis.

TABLE OF CONTENTS

LIST OF FIGURES

Page

Chapter 1

Introduction

1.1 Purpose of Study

The purpose of this study is to aid researchers in obtaining a genome level view of the duplications that occurs in the mammalian (specifically Human) genome. Dr. Mucheng Zhang and Dr. Shaying Zhao at the University of Georgia have developed an algorithm that identifies duplications (at least 5 genes long) in a mammalian genome. Using the algorithm they identified 124,748 patterns, totaling to 467Mb and accounting for 23% of the human genome [Shaying *et al.*, unpublished]. By viewing these patterns researchers can study the mechanism of how the duplications of certain elements occur throughout the genome. These duplicated elements include transposons and retrotransposons. Transposons are segments of DNA that can move around to different positions in the genome of a single cell. Retrotransposons are a type of transposons that duplicate by first being transcribed into RNA and back to DNA by reverse transcriptase. Many retrotransposons have long terminal repeats (**LTRs**) at their ends that may contain over 1000 base pairs [Kimball, 1994] [Lanters et al, 2001]. Discovered by Dr. Shaying Zhao's group the presence of two composite LTR-retrotransposon-like elements in the human genome that has spread into 65 copies totaling to 5 MB, a hypothesis being researched is that these LTR-retrotransposon-like elements emerged and spread throughout via retrotransposon-like mechanism.

To aid in their research we have developed GeneXplorer, which is an interactive visualization tool for graphical display of repeating gene sequences in the entire human genome. It takes as input textual data information generated from the algorithm mentioned above, parses it and then displays the information in an interactive graphical manner. Interactive techniques are powerful tools for manipulating visualization to analyze, communicate and acquire information. GeneXplorer makes use of these techniques extensively. It incorporates a length-proportional representation of chromosomes and the sequences. GeneXplorer uses color but it also provides a tool for users to choose their own color, this feature is added keeping in mind that nearly 8% of male population and 5% of female population suffer from some form of color blindness [Hoffman, 1999]. It uses a technique of magnifying the chromosome to show the sequence in depth.

The tool is a very straightforward interface that has four different windows. The parent window, allows the user to choose the chromosome for which he would like to view the sequences in the magnification window. The Magnification window, displays the chromosome in a magnified view. The overlapping sequence window allows the user to select a chromosome from among different sequences that overlap at a particular location. The support window, which is very similar to the magnification window, is used to view the selected sequences on other chromosomes.

The user follows these steps to study the repeating sequence:

- Step1: The user chooses a chromosome that contains a repeat sequence of interest.

- Step2: When the chromosome is selected it pops up the Magnification window. The magnification window has three panels. The first panel contains the entire chromosome. The second panel shows the magnified view of the chromosome. From this panel the user selects the repeat sequence of interest. The third panel lists the gene information of the selected sequence. The user chooses a sequence of interest in the magnification panel. The user can view the beginning base-pair position and the ending base-pair position, the starting gene number and the ending gene number of the sequence.
    - On special cases a new window, the overlapping sequence window, will come up. This window represents all the sequences that overlap over the point on the chromosome that user clicked. The user may select one of these sequences and proceed in the usual manner.

  The magnification window also allows the user to view the sequence through the UCSC Human Genome Browser at the http://www.genome.ucsc.edu website [Kent *et al*, 2002].
- Step3: Once the user selects the sequence in the magnification window or in some cases the overlapping window, the repeating sequences are then represented in the Main window chromosome on which that sequence occurs. The user can now select the chromosome in which the selected sequence is present and this will pop up the support window where the user can study the repeat in detail.

These steps will be explained in detail in chapter 3.

1.2 Process cycle

The research can be divided into following steps:

Biological data retrieval

File consisting of chromosome information was retrieved from NCBI databank. Each file consisted of a list of gene with their beginning base-pair position, ending base-pair position, orientation and the gene name.

Statistical data processing

The files chromosome files retrieved from the NCBI databank is preprocessed and later analyzed for searching repeat gene sequences, of minimum 5-gene length, using an program that uses statistical algorithms and formulas. The output of this program is a file containing the repeating gene sequences information of the entire human genome [Watson, 1990].

Information visualization

The file contain the repeating gene sequence information is then preprocessed. The file is divided into 24 files, each containing the repeat sequence for one of the chromosomes of the human genome. These 24 sequence files are then used in the GeneXplorer visualization tool.

Biological Analysis

Using the GeneXplorer visualization tool the researchers can do biological analysis on the repeating gene sequences.

The complete life cycle of the process is shown in the following figure.

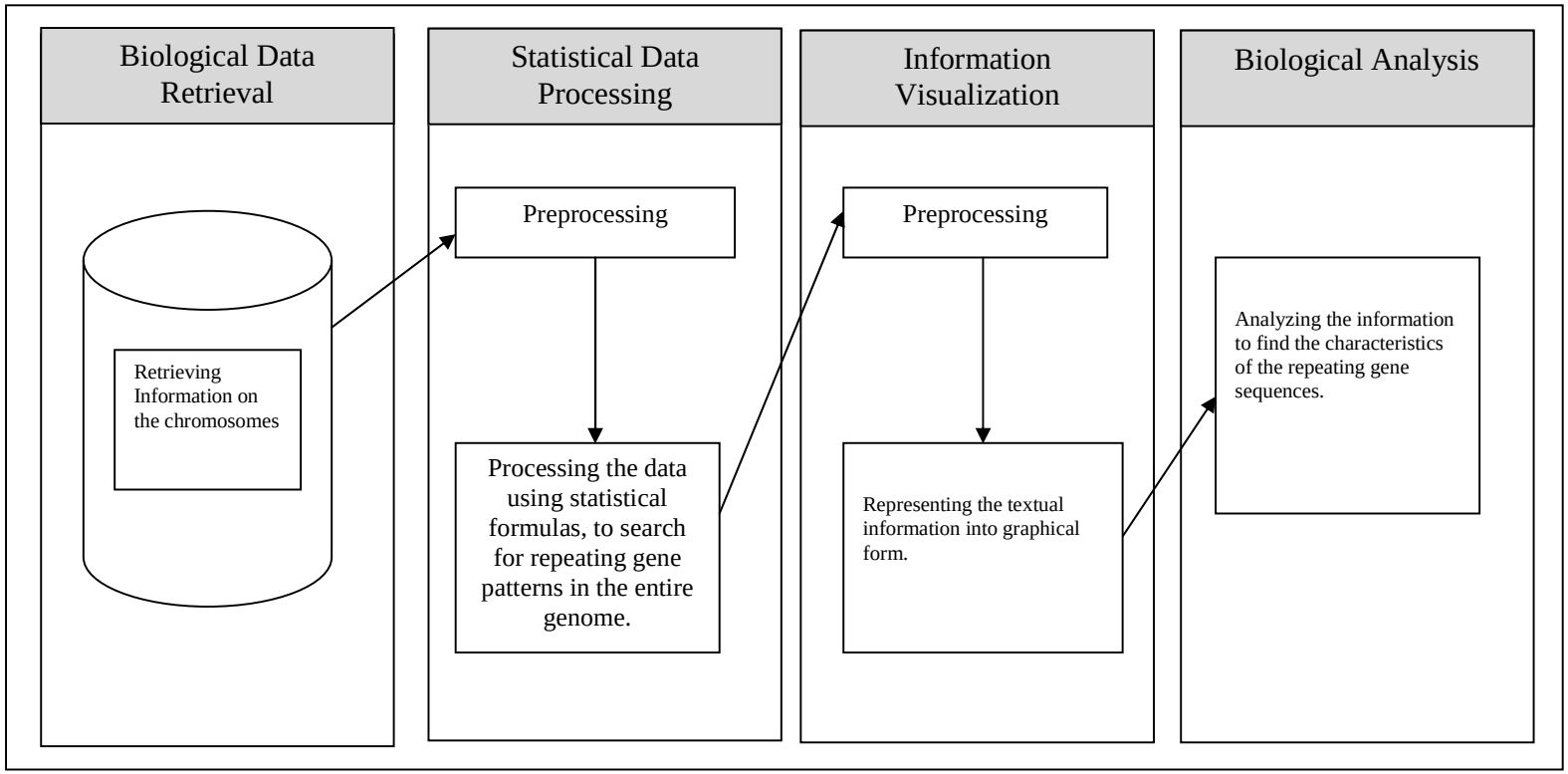

Fig1.1 Process cycle: This figure represents the complete cycle that the research goes through. The first step is retrieving biological information. The second step is to process the data using statistical formula and algorithm. A program is designed which implements these formulas and produces a file with the information regarding the repeating patterns in the genome. The third step preprocesses the sequences file and divides in to the 24 separate chromosome files and then represents this textual data into graphical form for the Biologists to biologically analyze the behavior of the repeating patterns in the fourth step.

CHAPTER 2

RELATED WORK

Advances in technology over the past decades have allowed scientists and researchers to produce vast quantities of biological sequence data. While this data may be viewed in purely textual formats, researchers can benefit greatly from well-designed graphical visualizations of this data, accompanied by well thought out interactions with the data and the display. A number of such tools have been developed that provide such displays of genomic data.

2.1 UCSC Human Genome Browser

The UCSC Human Genome Browser [Kent *et al*, 2002] is a tool that displays genomic information at various levels, aligned with annotation tracks. The browser can display assembly contigs and gaps, mRNA and expressed sequence tag alignments, multiple gene predictions, cross-species homologies, single nucleotide polymorphisms, sequence-tagged sites, radiation hybrid data, transposon repeats, and more as a stack of co-registered tracks. The user can search for a gene by its name, author, keyword or other gene information. The user can also view various sections of the chromosome by either specifying the chromosome band or by specifying the chromosome location by the base pair position. The UCSC Genome Browser itself is divided into three sections. The top most section contains controls that allow the users to search, zoom and scroll over a section of the chromosome. The middle section displays a graphical picture of

the particular section, along with the genome annotations. The bottom section has controls to fine-tune the display view.

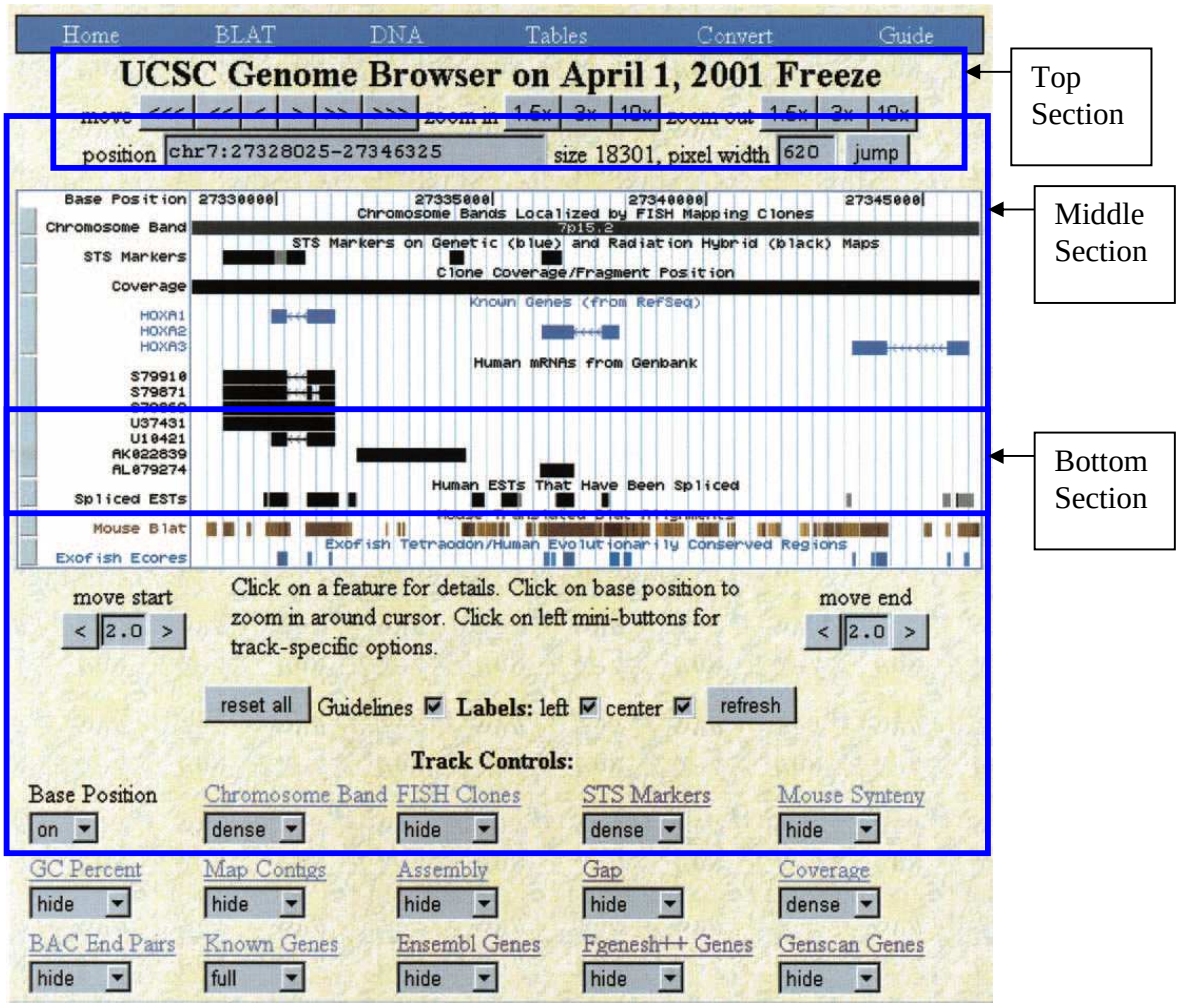

Fig 2.1: View of the UCSC Human Genome Browser depicting the three sections. The top section is for moving across the chromosome. The middle section is the graphical representation of the gene annotations and the bottom section is for fine-tuning the graphical view.

The browser represents the annotations in the form of tracks. The track annotations can be viewed as dense, full or hidden form. They can also be viewed from different levels of magnification starting from the whole chromosome to the splicing of a gene. This characteristic of the Genome Browser makes it very useful for studying sections of a chromosome. The major limitation with the browser is that we can only view the

chromosomal information one chromosome at a time. This falls short of fulfilling the requirements for searching the gene sequence repeats in the entire genome.

2.2 NCBI Map Viewer

NCBI (*National Center for Biotechnology Information)* Map Viewer [Dombrowski *et al,* 2003] presents genetic, radiation hybrid, cytogenetic, sequence-based, and clone maps for many genomes, where 'map' refers to a position of a particular type of object in a particular coordinate system. Some important features of Map Viewer that make it such a favored tool among researcher are the mechanism to compare maps in different coordinate systems; robust query interface; diverse options for configuring the display; multiple functions to report and download maps and annotated information; detailed descriptions of the objects displayed on the maps, etc. NCBI Map viewer can display genome specific BLAST (*Basic Local Alignment Search Tool)* searches [Altschul *et al., 1990*]. When a user clicks on the *Genome View* button they can see all the BLAST hits on the genome. Adding to that the flexibility of querying options like Boolean operators and wild cards makes it easier for users to search and navigate the chromosome. Other important features are zooming and ruler. The chromosome can be viewed as a whole or can be zoomed in to view specific locations and the ruler can be added to any map to show the size and location of a feature of interest. Users can get a genome level view for a gene search but as of now no attempts have been made to represents genome level view of a gene sequences. NCBI although a very useful tool for getting lower level repeats information in a genome falls short in producing high level repeat information.

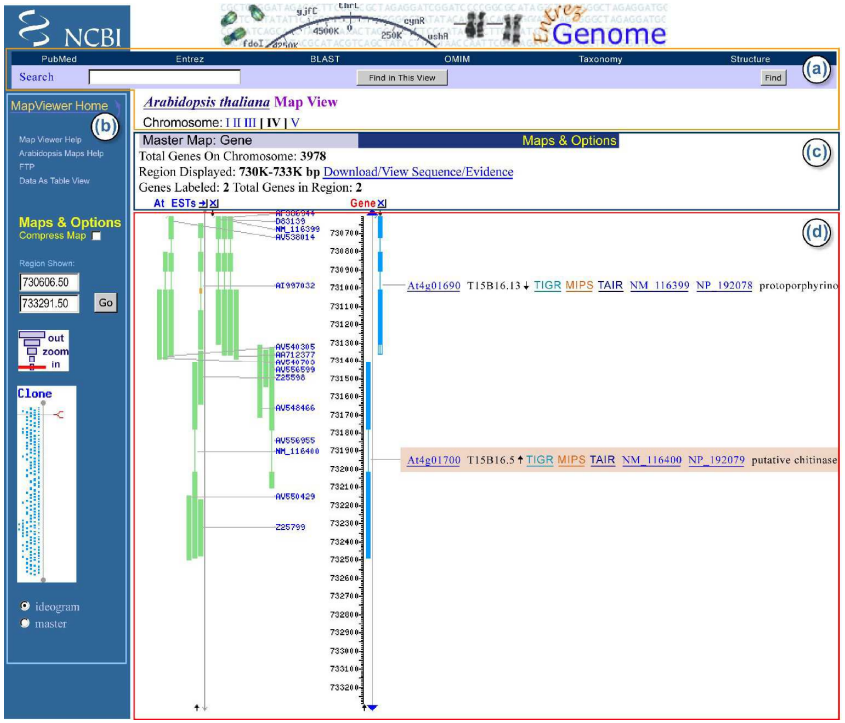

Fig 2.2a: NCBI Map Viewer can be used to see chromosomal information at different levels chosen by the user. There are many tools like the zoom-in zoom-out tool to view a particular section at various labels, graphical and textual chromosome navigator to navigate to any point of the chromosome, and a ruler to study the length of the genomic element. Different color-coding represents different types of data.

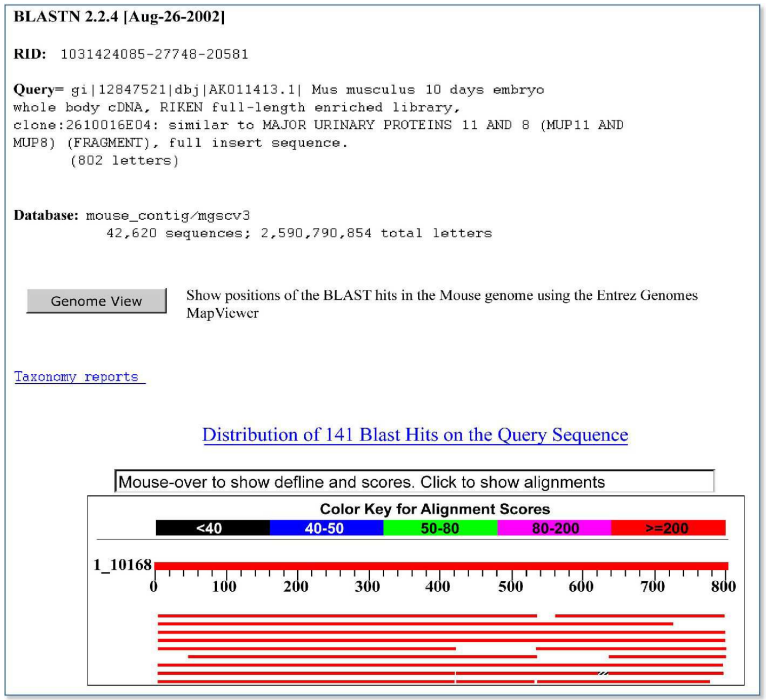

Fig 2.2b: Accessing the Map Viewer display from a genome-specific BLAST results page. Selecting the **Genome View** button shows all of the BLAST hits on the genome.

.

2.3 Ensembl

Ensembl's [Birney et al, 2004] strength lies in the versatility of the forms in which it can be used. It is available as an interactive website, a set of flat files and as a portable software system for handling genomes. As in NCBI Map Viewer, here too the genomic information can be viewed at various depths of magnification. It provides a platform for viewing different types of genomic information from various sources. Even with its ability to represent different types of genomic data, to the best of or knowledge no attempts have been made to represent gene sequences repeats.

Fig 2.3: DAS Visualizations [Flicek et al, 2008]. A 19 Mb region of human chromosome 11 showing identical data displayed with (from top to bottom) the color gradient, histogram and tiling array 'wiggle' format. The color gradient format transitions from yellow (low values) to blue (high value). The histogram display format supports merged data in bins across the genome; the display value is selectable to be either the average of the bin (shown here) or the maximum value in the bin to achieve greater data contrast. In the histogram format, the lowest value in the data set becomes the baseline. The tiling array format allows for the display of both positive and negative values with overlapping data points resulting in the maximum data point being displayed. All three-display formats support in-line data normalization. The ideal format will depend on the data to be displayed.

2.4 viewGene

viewGene [Kashuk et al, 2002] is a Java based visualization tool used for viewing sequence reference scaffold. viewGene is used to view genomic information at a local level, i.e. under a megabase of the sequence. Higher-level views can be viewed at a sub optimal level. It takes in input from various sequence format and analysis programs like GenBank [Benson *et al.,* 2007], FASTA [Lipman *et al.,* 1985], RepeatMasker [Smith *et al.,* 1997], CrossMatch [Green, 1993] [Green, 1994], BLAST [Altschul *et al., 1990*] and user-defined data. As mentioned viewGene does not give an optimal representation of higher level chromosomal views and hence falls short in fulfilling the requirements for researching gene sequences repeats.

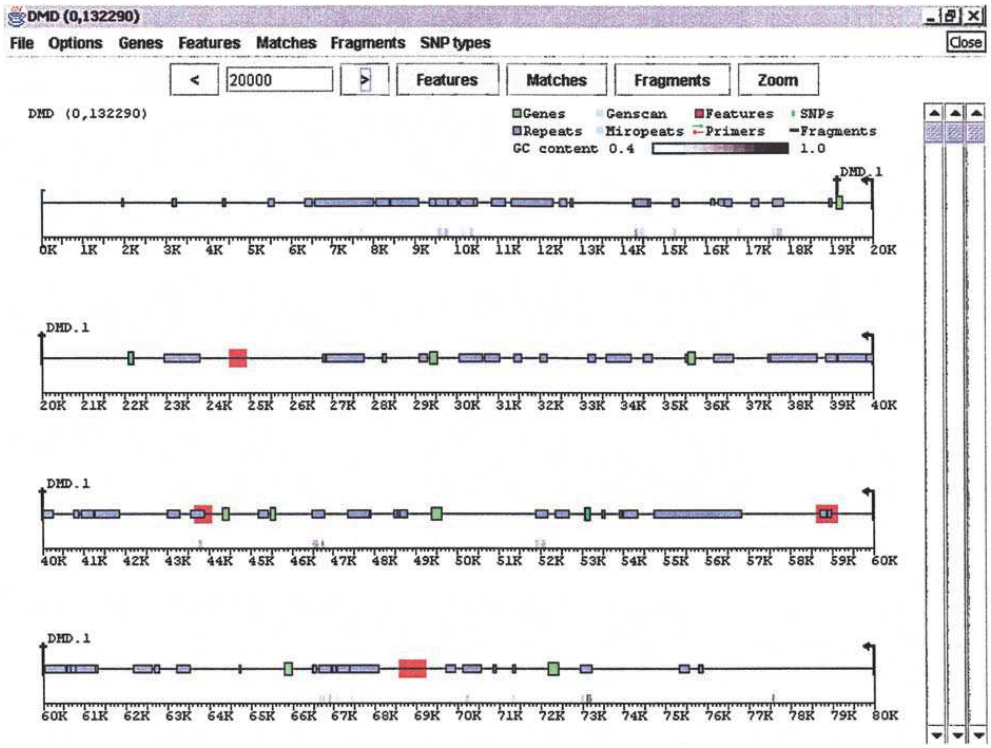

Fig 2.4: Data from GenBank record AL031542, containing 13 exons of the human dystrophin gene, is shown with green boxes representing the exons of "CDS" tags, blue boxes representing "repeat region" tags, and red areas defining "misc feature" tags. The user can click on any box and obtain details about the feature, control which types and classes of features are displayed, and label features [Kashuk et al, 2002].

2.5 REPvis

REPvis [Kurtz et al, 2001] is the visualization tool for the repeat searching program *REPuter*. It provides an interactive graphical display for examining repeat structures computed by *REPfind*. The repeats are represented in the following manner: First the most important or the longest repeat is shown. The smaller repeats are shown by using a color scale of 10 colors; with black being the least significant repeat. The user can move the slider to view the most significant to the least significant repeat. REPvis also has the feature of zooming in on the region of interest. REPuter like many of the other sequence repeats search tools also searches at the nucleotide level and not the higher levels like the gene levels, which is the requirement for this research.

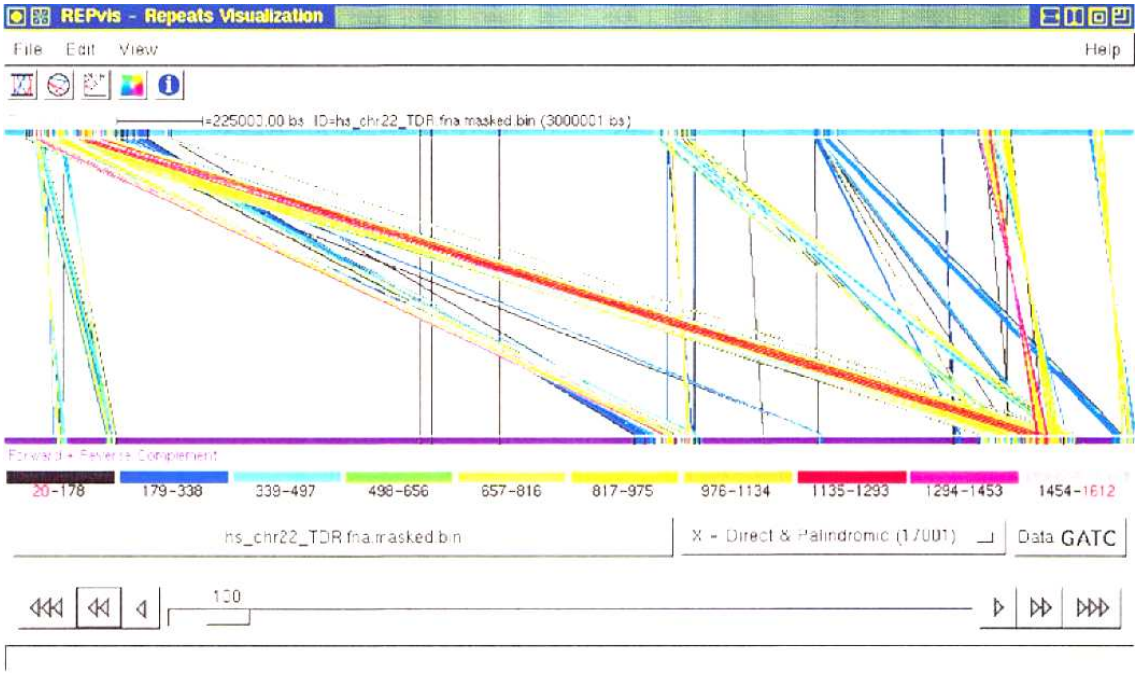

Fig 2.5: Assembly checking of human chromosome 22 [Kurtz et al, 2001]. *REPvis* display of exact direct and palindromic repeats with a minimum length of 300 bp. The chromosome (32 484 231 bp) consists of 11 concatenated contigs separated by vertical white lines (the separators are specified as an extra annotation and displayed by *REPvis*). The color-code for repeat length indicates that all other repeats are dwarfed by one long, exact repeat (light purple) of 190 014 bp.

2.6 GeneXplorer

GeneXplorer is visualization tool developed using Java. The tool is developed to study repeating gene sequences in the entire human genome. It takes in text-based files containing the information regarding the chromosomal repeats and produces a graphical visualization to aid researcher to further study the repeats from a genomic view. It has four windows: main window; magnification window; overlapping sequence widow; and support window. The main window represents the entire genome with all the 24 chromosomes displayed on the main panel. The magnification window displays the sequences in the main chromosome under investigation, which is called the primary chromosome. Only one magnification window can be opened as only one chromosome can be investigated at a time. As soon as the magnification window is loaded the main window gets refreshed with repeat sequences present in other chromosomes that correspond to sequences in the primary chromosome in the magnification widow. In the magnification window the user can select a sequence of interest and all the repeat sequences for the selected chromosomes in other chromosomes get highlighted in the main window. In order to get a detailed view of any of the repeat sequences in a particular chromosome, the user can open a support window for that chromosome. More than one support window can be opened at a time to investigate sequences that might repeat in multiple chromosomes, with the primary chromosome in the magnification window serving as the central point for investigating these repeat sequences. If the user selects a region in the magnification window, which falls inside multiple sequences, these

overlapping sequences will open in the overlapping sequence window. The user can then select one particular sequence from the list of overlapping sequences for that investigation.

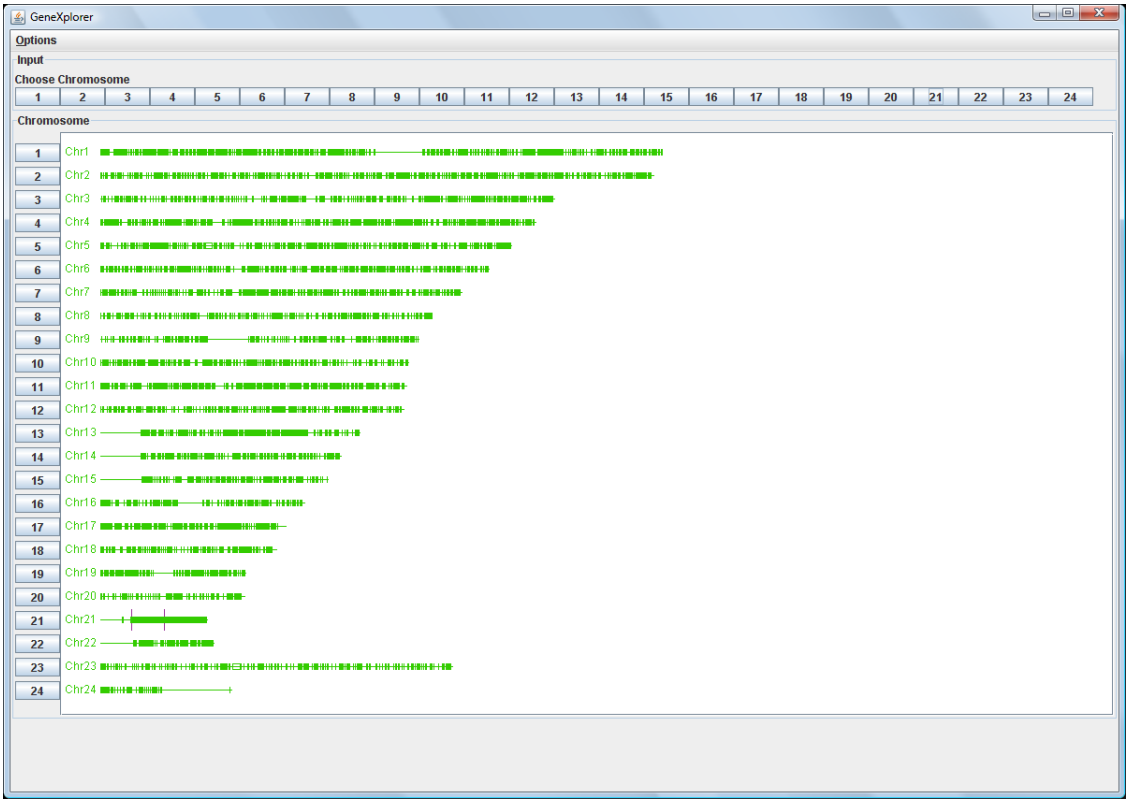

Fig 2.6: Genome level representation of the sequences repeats with Chromosome 21 being the primary chromosome.

This tool provides the following advantages over the other tools:

- The tool has been specifically designed to aid researchers in the study of repeating gene sequences in the human genome. Due to the specialized nature of the application, it provides an in-depth knowledge of the gene sequences repeats in the entire genome. The tool provides a genome level representation of the gene sequences repeats. No other tool to our knowledge does that.
- The application gives a high level, i.e., gene sequence level view of the

repeats where as other tools specialize in searching repeats in the base pair level of the genome.

- The application has been developed using Java Web Start, which provides the flexibility of deploying it as a web application but at the same time providing all the benefits of desktop application. Additionally, it does not require a constant internet connection as long as the user chooses not to the launch the UCSC Genome browser to get an in-depth view of the sequence at that particular chromosome coordinates.
- An added advantage of using Java and Java Web Start is that the application can run on any OS as it uses Java and simple text files as input. This environment can be simulated on any machine without having to recompile any part of the application.
- The application allows the users to customize the appearance of the sequences by choosing desired colors. This helps users who might have a disability in distinguishing / identifying certain colors.
- Users can download any chromosome data from the web and feed into the application as long as the input files follow certain guidelines as mentioned in Chapter 4.

2.7 Summary

In the above sections we have done a study of known visualization tools used in the study of genomic data. We have also done some comparative study of the pros and cons of each tool in relation to the tool developed for this thesis,

GeneXplorer. The comparison is based on each tool's ability to represent high level repeats, specifically gene sequence repeats in the entire Human genome.

The tools mentioned above have been used by researchers and have proven very useful in their own way but in the case of studying Gene sequence repeats in the entire genome as a whole all these tools fall short in one or more of the following functionality: representation of the information in the entire genome at one glance; representation of repeats at higher level like the gene level; etc.

The tool developed in this thesis, GeneXplorer, on the other hand is one such tool that can be used to view high level genomic information like gene sequences and their repeats in the Human genome. It is neither the first tool to do repeat analysis in the genome nor the first tool to display information in the whole genome at one glance. However, to the best of our knowledge it is the first tool to display gene sequence repeats in the complete genome. Other tools listed above have been comparing repeats in the genome for quite a while but most of the comparisons are done at the nucleotide level. GeneXplorer attempts to do the same at a higher level, the gene level, which is the primary focus of the research.

Chapter 3

User Level view

3.1 User Interface

The GeneXplorer user interface has been developed using Java and will be deployed using Java webstart. The GeneXplorer user interface comprises of the following screens.

- Main window
- Magnification window
- Overlapping sequence window
- Support window

3.1.1 Main window

The main screen is the first screen that the user will come across once the application is deployed. It has four main components:

- Option menu
- Input button panel
- Support button panel
- Genome view panel

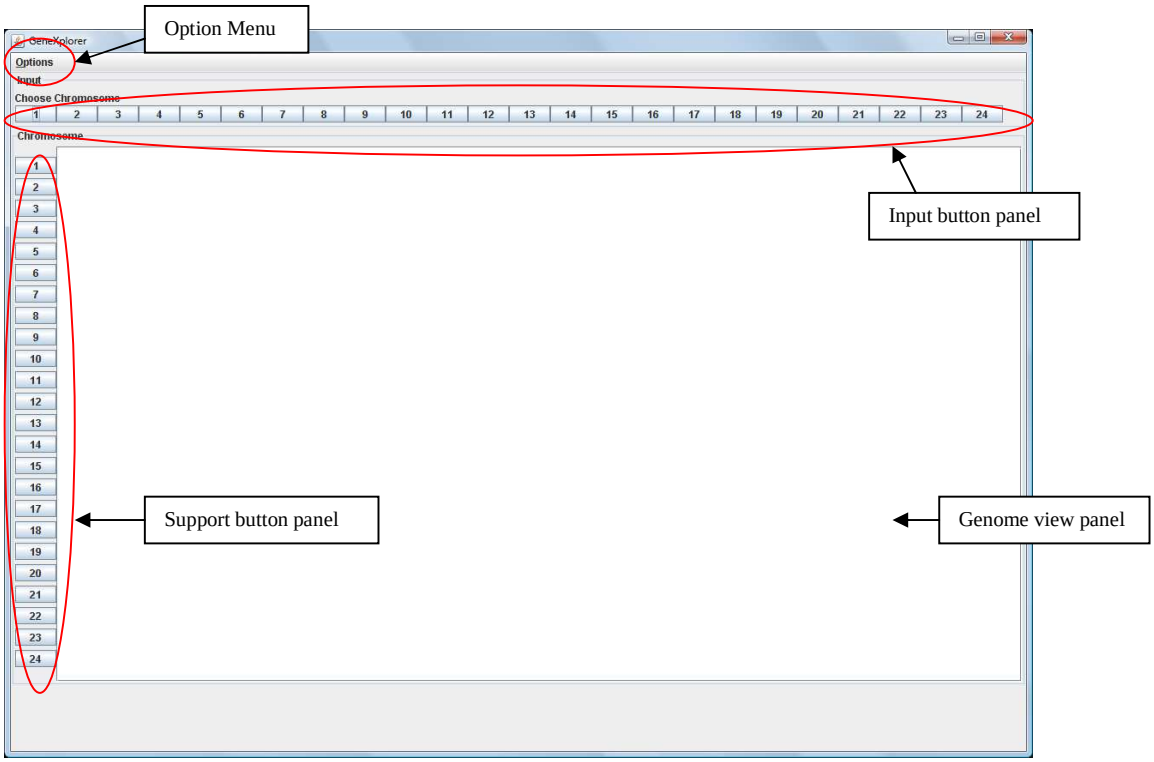

Fig 3.1: Main window when it was initially launched.

3.1.1a: The option menu

The option menu consists of the following options:

- Set Directory: This option is used to set the directory containing the sequences files and chromosome files. The sequences files are the files that are generated using the algorithm developed by Dr Mucheng Zhang and Dr Shaying Zhao at the University of Georgia. The chromosome files are the files containing the gene information of a particular chromosome. The user has the flexibility of choosing a location in the local machine to store the files required for running the application.
- Set Color: The user can choose the colors to represent the genomic information. This feature is added keeping in mind that nearly 8% percent

of the male population and 5% of female population have some form of color blindness. As color is one of the primary tools for representing genetic information, this feature will allow a wide range of users to use the application with ease.

3.1.1b: Input button panel

The input button panel consists of 24 buttons each representing different chromosomes in the Human genome. The user can click on any one of the input buttons and choose the chromosome of interest. Once the button is clicked the interface will load the relevant sequence file for that chromosome from the directory set by the ‘Set Directory’ option. The interface will then plot the genome level view of the sequences in the Genome view panel. It will also open a window, which will contain a magnified view of the chromosome along with its repeats. This window will be referred to as *Magnified window*.

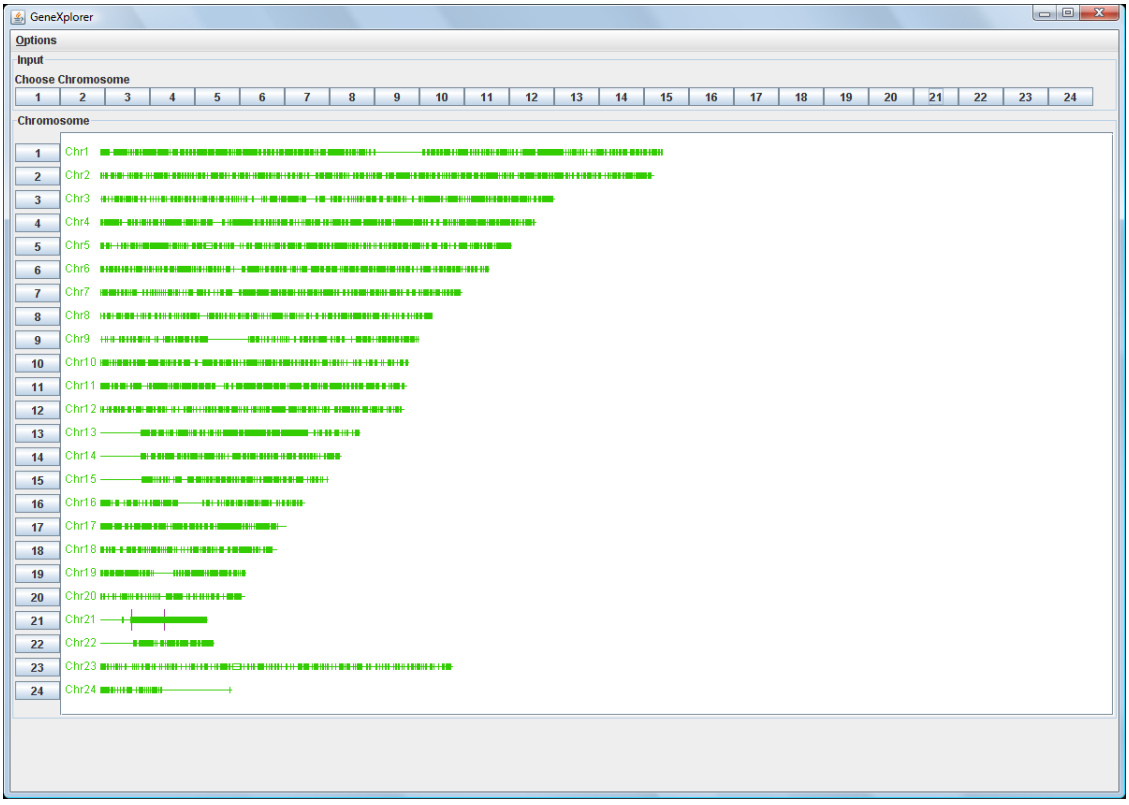

Fig 3.2: Main window once a chromosome from the input window is clicked. The sequence repeats on the other chromosomes corresponding to the sequences present in the selected chromosome is represented in the Genome view panel.

3.1.1c: Support button panel

The support button panel again consists of 24 buttons each representing different chromosomes in the human genome. Once the files are loaded from the default directory and the genome level view of the repeat sequences is plotted, the user can use one of these buttons to open a magnified view of the corresponding chromosome consisting of repeat sequences in that chromosome. This window will be referred to as *Support window*.

3.1.1d: Genome view panel

The genome view panel represents all the 24 chromosomes in the Human Genome. Once a chromosome button from the input button panel is clicked, the view panel displays the repeating sequences in other chromosomes that correspond to the sequences in the chosen chromosome. If a particular sequence is selected in the "Magnification Window" then all the repeats in other chromosomes are highlighted in the "Genome view panel". Users can then use the "Support button panel" buttons to study repeat information in other chromosomes in detail.

3.1.2 Magnification window

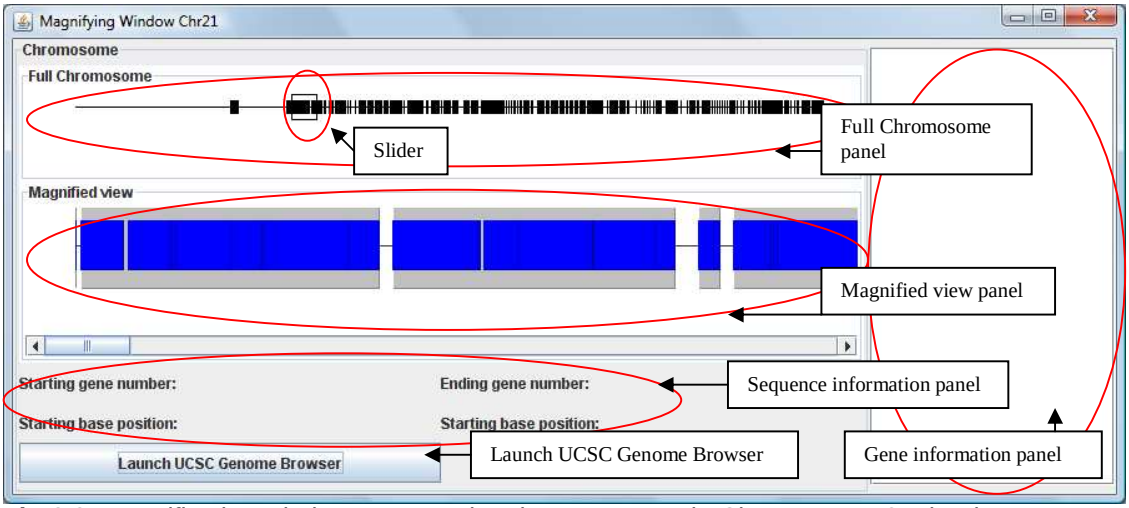

Fig 3.3: Magnification window representing the sequences in Chromosome 21 that have repeats in the same chromosome and the other chromosomes in the entire genome.

Once the chromosome button in the input panel of the Main window is clicked, the magnification window pops up representing the sequence information of the selected chromosome. The magnification window is divided into five major components:

- Full Chromosome panel
- Magnified view panel
- Gene information panel
- Sequence information panel
- Launch UCSC Genome Browser button

3.1.2a: Full chromosome panel

The full chromosome panel renders an overview of the entire chromosome with sequence information. It also features a slider, which can be used to identify a region that needs to be magnified for investigation.

3.1.2b: Magnified view panel

The magnified view panel renders the detailed sequence information of the selected chromosome. The slider in the “full chromosome panel” selects the view that is magnified and rendered in this view. It has two kinds of sequences. The taller blocks represent the complete sequences and the shorter blocks represent the subsequences. The user can select the sequence by clicking on it. Once the sequence is selected, the information about the list of genes in the selected sequence is shown in the ‘Gene information panel” and the sequence information is shown in the “Sequence information panel”.

3.1.2c: Gene information panel

The Gene information panel shows the information about the entire list of genes present in the sequence selected in the “Magnified view panel’. The information includes the starting base-pair position, ending base-pair position, orientation of the gene and name of the gene.

3.1.2d: Sequence information panel

This panel shows the sequence information of the sequence selected in the “Magnified view panel”. It contains the starting gene number, ending gene number, starting base-pair position and ending base-pair position.

3.1.2e: Launch UCSC Genome browser button

This button launches Internet Explorer and takes the user to the UCSC Genome browser website. The Genome browser is fed the information in the “Sequence information panel” as input and the user can study the selected region as needed (in detail).

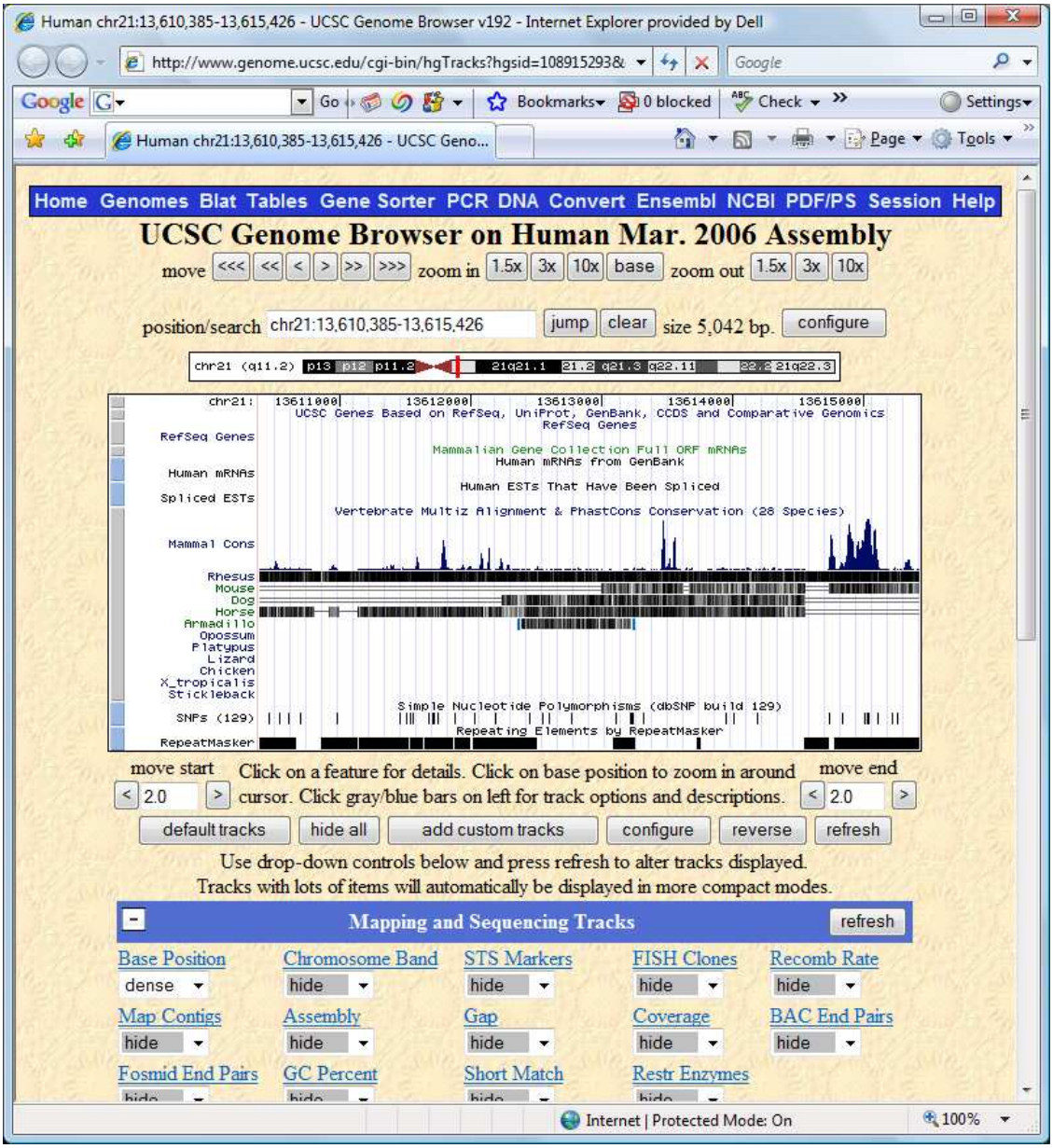

Fig 3.4: UCSC Genome Browser. Representing information about the chromosome area covered by the selected gene sequence.

3.1.3: Overlapping sequences window

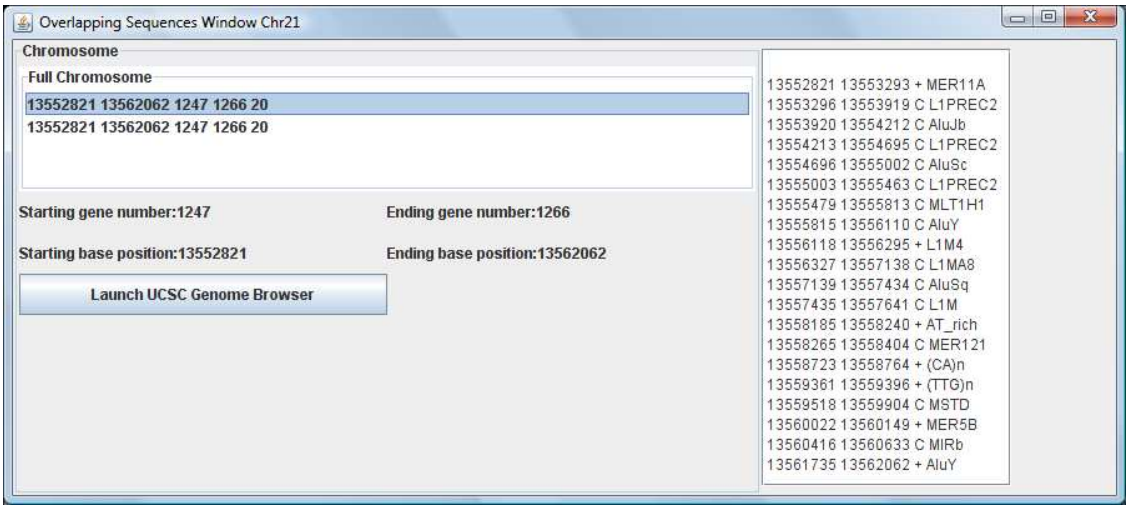

Fig 3.5: Overlapping sequence window representing the overlapping sequences in Chromosome 21 at a particular point selected by the user.

This window is launched through the “Magnification Window”. A user might select a sequence location in the “magnified view panel” of the “Magnification Window”. If only one sequence is located at that location then the information about that sequence is represented as described earlier. However, if multiple sequences are present at that location, then the interface launches a new window called “Overlapping sequences window”. This shows a textual list of all the sequences present at that location. It contains the same “Gene information panel”, “Sequence information panel” and the “Launch UCSC Genome Browser” button as described in the previous sections. The only difference is that the user will select a sequence from the textual list of sequences shown as compared to selecting a sequence shown graphically in the “Magnification window”.

3.1.4: Support window

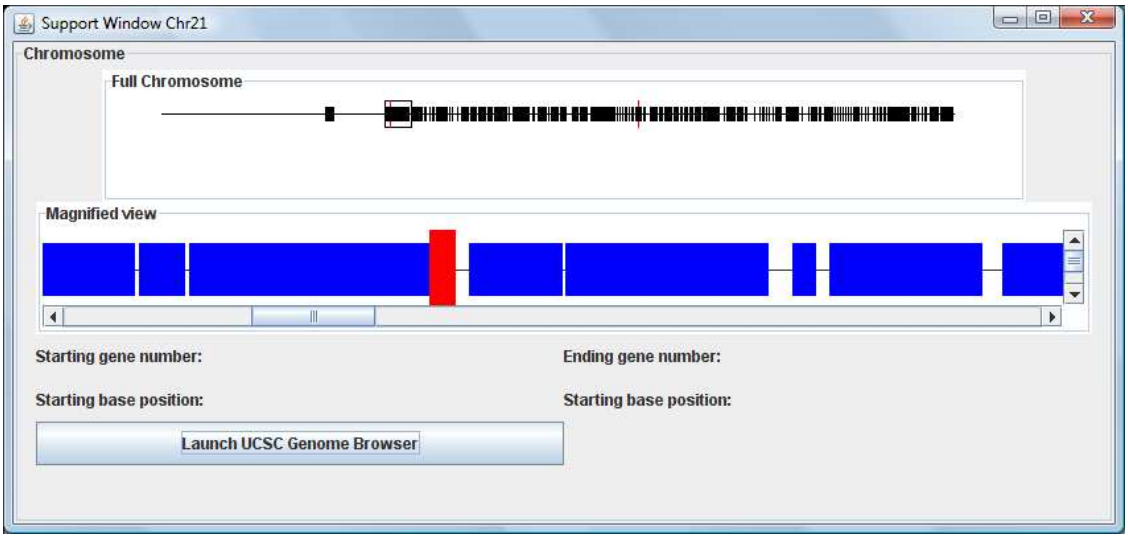

Fig 3.6: Support window representing the selected sequence in Chromosome 21 that is being repeated in the same chromosome.

The support window can be launched through the “Main window” by clicking on one of the buttons from the “Support button panel”. Once the user selects a chromosome to study by clicking on a chromosome button from the input button panel, the “Main window” renders the information of sequence repeats in other chromosomes that correspond to sequences in the selected chromosome. The user can then investigate those repeats in other chromosome in detail by clicking on one of the buttons in “Support button panel”. This will open a “Support window” for the corresponding chromosome. This window looks similar to the “Magnification window” except that it does not contain the “Gene information panel” as the genes in the sequence will remain almost the same as the sequence in the “Magnification window”. When a user selects a particular sequence in the Magnification window, if a repeat is present in the “Support window” then it is highlighted in a different color. The user can then move the

slider to magnify that section in the “Magnified view panel”. If the user clicks on that highlighted sequence, the “Sequence information panel” will represent the sequence information as described earlier. Clicking the “Launch UCSC Genome Browser” button will take the user to the UCSC Genome Browser with the sequence information pre-filled for further investigation.

3.2 Case Study

The GeneXplorer has been designed keeping in mind the user convenience. Certain extra features have been added to the application like ‘Color Chooser’ to involve a wider range of users specially people suffering some form of color blindness.

Color is a very important tool for classifying different data and has been extensively used in the field of Biology. GeneXplorer also uses color for classifying sequences, repeats, main sequences and subsequences. However, the choice of the color is left on the user. The users can go to the menu bar at the top left hand side and click on the menu item called ‘Options’. Under options there is sub menu item called ‘Select Color’. The user would click on that which would pop-up a dialog box with three divisions. The first section sets the color for the sequence window, the second for the main window and the third for the magnification window.

To make the application portable we have also added a feature where the user has a choice of choosing where he wants to store the files on the local system. After storing the files the user can then set the path to the directory containing the files from the application. The user can do so by clicking on Options then on

‘Set Directory’. This will bring up a directory chooser dialog box and the user can browse the local system and choose the directory.

When the user clicks on the chromosome button a new window ‘Magnification Window’ pops up. This window has three panels and a button. The first panel represents the full chromosomes and all the sequences in it. The second panel would represent the magnified view of a section of the chromosome. The third section shows the list of genes in the selected sequence and their information. The user needs to select the region of interest in the chromosome by clicking on the full chromosome.

By selecting a section of the chromosome the application will generate the graphical view of all the sequences in that region inside the magnification window.

The magnification window has two kinds of sequences. The taller blocks represent the complete sequences and the shorter blocks the subsequences. The user can select the sequence by clicking on it. It might happen that the point clicker by the user is covered by multiple sequences. In the case another window will pop-up with the list of sequences. The new window is similar to the magnification window except for the Full Chromosome and the Magnified chromosome panels. Here there is only one panel with the list of the sequences containing the point the user clicked in the Magnification Window. The user can now select the sequence of interest. Once the sequence of interest has been selected, the repeating sequences will be depicted in the genome view on the

Main Window. The list of gene and their information will be listed in the Gene List panel.

The user can also look at that particular sequence in the' UCSC Genome Browser' by clicking on the button at the bottom of the Magnification Window.

On the main window the repeats are highlighted on the entire genome. If the user wishes to study these repeats on the other chromosomes the user will have to click on the buttons next to the respective chromosomes. This will pop-up a window which looks like the Magnification window except that the gene list panel is not present. Here too the user has a choice of going to the UCSC Genome Browser using the button at the bottom.

Chapter 4

Developer level view

4.1 Objects

The software is modeled using the following objects

- Chromosome object:

  Each chromosome is manifested by its own object. The chromosome object representing the chromosome under study (for which the corresponding button was clicked on “Input button panel”) contains the main sequences and sub sequences which can be read from the “Sequence file” (explained in the next section). The object can also contain the list of genes for the chromosome. The chromosome object representing one of the other chromosomes, in which a repeat sequence might be found, will only contain information about those repeat sequences in the subsequence list.

- GeneSequence object

  This object represents any sequence (main sequence, sub sequence or repeat sequence). It contains the following information about sequence

  - A unique id identifying the sequence within the system
  - Begin gene position
  - End gene position
  - Begin base-pair position

- End base-pair position
- Length of the sequence

4.2 Input Files

The software takes two input files for each chromosome. The two different types of files are:

- Chromosome file – chr#.rmsk (where # represents the chromosome number)

  This file contains the list of genes present in the chromosome. Each gene is specified by the base-pair starting position, base-pair ending position, orientation of the gene and the name of the gene. The list is ordered by the position of the gene in the chromosome. A typical entry in the file will look as follows:

  1 38 + (CCCTAA)n

- Sequence file – seqChr#.txt (where # represents the chromosome number)

  The algorithm developed by Dr Mucheng Zhang and Dr Shaying Zhao at the University of Georgia generates this file. It contains the gene repeat sequence information for each chromosome. A typical entry in the file will look as follows:

  0 1 6353 11636 chrY 0 6352 1 2709520 chrX 0 6352 1 2709735

  Each entry in the file contains 14 elements. The first 9 elements relate to the chromosome for which this sequence file is generated and the remaining elements correspond to the chromosome where a similar

sequence can be found. The first element indicates whether this sequence is main sequence (0) or a sub sequence (1). The second element is the orientation of the sequence, the third element is the size of the sequence in terms of number of genes and the fourth is the number of times the sequence repeats within the same chromosome (at different locations) or in other chromosomes. The fifth element indicates the chromosome for which this sequence file is generated, sixth element indicates starting gene position, seventh element is the ending gene position, eighth is the starting base-pair position and ninth is the ending base-pair position. The tenth element indicates the chromosome where this sequence might repeat. The eleventh, twelfth, thirteenth and fourteenth element indicate the starting gene position, the ending gene position, the starting base-pair position the ending base-pair position respectively.

4.3 Input Processing

As soon as a user clicks on a button in the input button panel, the software loads the "Sequence File" for that chromosome from the directory set by the "Set Directory" option. For example, if the user clicks on "Chr1" the software will load the file –seqChr1.txt. The Sequence file is parsed line-by-line and each line is represented by at least two GeneSequence objects. The first sequence object is created with sequence information corresponding to the chromosome for which this sequence file has been generated. It is added to the chromosome's main

sequences list or sub sequences list depending on its type. The second sequence object is created with sequence information that corresponds to the second part of the entry as mentioned earlier for the chromosome in which this repeat occurs. This object is then added to that chromosome's sub sequence list (there will not be any main sequence list for other chromosomes). For example, in our example from the sequence file - the first sequence object will be added to chrY object's main sequences list and the second object will be added to the chrX object's sub sequences list. Since both the sequences are same, they are assigned the SAME sequence id. This id will later help us in identifying similar sequences.

4.4 Execution flow

As mentioned earlier, the user will first set the directory from which the input files can be read. Once a chromosome button from the "Input button panel" is clicked, the software reads the corresponding chromosome sequence file.

4.4.1 Reading Sequence file

It is parsed using the following algorithm.

1) Read the line

2) Split it into various elements based on space as the separator.

3) Create a new sequence object (GeneSequence).

4) Elements 1, 2,3,4,5,6,7,8 and 9 belong to the chromosome for which the input button is clicked.

a) Set the sequence object with the information from these elements. The 5th element should indicate the chromosome that contains this sequence. This should be the same as the chromosome the user intends to study. Please refer to the Input file description section for detailed information about what each element represents.

b) If the first element is 0, then it depicts that this is a main sequence. We check if the sequence has been added to the main sequence list in the Chromosome object for which the input button was clicked. If it has not been added, we increment the sequence id and assign it to the sequence. We then add it to the main sequence list.

c) If the first element is not 0, then it is a subsequence. We follow the same logic as b) but add the sequence to the sub sequence list of the chromosome.

5) Elements 10,11,12,13 and 14 belong to the chromosome that contains this repeat sequence of the sequence mentioned in Step 4. We then:

a) Create a new sequence object.

b) Set the related information about the sequence object using the description of the input file described earlier.

c) The sequence id should be the same as that calculated in Step 4. We assign this id to the sequence and add it to the respective

chromosome's sub sequence list (the chromosome in which it occurs).

6) Repeat until the end file.

Following the above logic helps us identify which sequences are main sequences and sub sequences when it comes to the chromosome under study. It also helps us identify which sequences are repeats and the chromosomes in which they occur along with their respective information (starting gene number, ending gene number, starting base-pair location and ending base-pair location). The repeats sequences or similar sequences, share a sequence id. In this way if one of the sequences instances is selected, we can easily identify the remaining instances of that sequence using the sequence id.

4.4.2 Main Window

After reading the input file and creating all chromosome objects, the chromosomes along with the corresponding sequences are drawn in the main window. Each chromosome object is passed a graphic object that it then uses to draw itself on the screen. The length of the longest chromosome is mapped to 600 pixels, i.e. Chromosome 1 is mapped to a length of 600 pixels. All other chromosomes have their relative mapping calculated based on the 600-pixel Chromosome 1 using the following formula:

(length of chromosome/ length of longest chromosome) * 600

For example, Chromosome 1 has a length of $2.47x10^8$ and Chromosome 2 has a length of $2.43\ x10^8$. So Chromosome 2 will be presented by a line, which is

$(2.43\ x10^8/2.47x10^8) * 600 = 591$ pixels in length.

A line can then represent each chromosome by calculating its relative length to chromosome one.

Once the chromosome is plotted using its mapped length, we can proceed to plot its subsequences. Each Chromosome object holds a list that includes all the subsequences. The chromosome under study will hold the list of main sequences that contain these sub sequences for that chromosome as well. However, for the purpose of visualization within the main window, we will ignore the main sequence list of that chromosome so that all chromosomes are treated uniformly. In order to plot a sequence over a chromosome, we calculate its relative location on the line using a formula similar to the one mentioned above. We can better explain this using the following example. We consider Chromosome 21 and say the sub-sequence:

chr21 0 258 9719769 9882589

We first get the length of this chromosome in pixels using the formula below.

$(4.70\ x10^7/2.47\ x10^8) * 600 = 114$ pixels

Then we map the starting base-pair position to a 114-pixels based location using the formula.

(base-pair position/ chromosome-length) * 114

Using the above formula on the example we have

$(9.71 \times 10^6 / 4.70 \times 10^7) * 114 = 23$

So the starting position of this sequence on the 600-pixel line will be 23 pixels. The end position will be

$(9.88 \times 10^6 / 4.70 \times 10^7) * 114 = 24$

We then assign a standard height of 6 pixels to this sequence and draw it in the overall view panel on the line of 600 pixels representing the chromosome with a width of 1 pixel. We repeat this process for all sub-sequences.

4.4.3 Magnification Window

Once the chromosome objects are created, the chromosome for which the button was clicked, i.e., the chromosome under study, is passed on to the "Magnification window". The window has two graphical sections - one to display an overview of chromosome and the other to display a magnified view.

4.4.3aOverview of Chromosome

Each chromosome object contains the overall length of the chromosome. This overall length is mapped to a default fixed length of a certain number of pixels. In our software, we assume a length of 600 pixels. The subsequences of the chromosome are then plotted on this line that is 600 pixels in length. We can better explain this using the following example. We consider Chromosome 21 and say the sub-sequence:

chr21 0 258 9719769 9882589

In order to plot this sequence we first map the entire chromosome length to 600. Then map the starting base-pair position to a 600 pixel based location using the formula.

(base-pair position/ chromosome-length) * 600

Using the above formula on the example we have

$(9.71 \times 10^6 / 4.70 \times 10^7) * 600 = 124$

So the starting position of this sequence on the 600-pixel line will be 124 pixels. The end position will be

$(9.88 \times 10^6 / 4.70 \times 10^7) * 600 = 126$

We then assign a standard height of 10 pixels to this sequence and draw it in the overall view panel on the line of 600 pixels representing the chromosome with a width of 2 pixels. We repeat this process for all sub-sequences. We do not plot main sequences in the overall view panel as we would like to show the user how the subsequences are concentrated. Plotting main sequences would make it difficult for the user to find areas of greater concentration of sub-sequences.

4.4.4b Magnified view

The overview contains a small slider, which the user can slide over the chromosome. The portion within the slider is then magnified within the magnified view. In order to get the portion of the chromosome that needs to be magnified, the user first clicks on an area of interest. Once we get the location based on the x and y coordinates, we draw the slider around the click. The slider is a square of 20 pixels. If the user clicks on the 600-pixel line representing the chromosome,

say at pixel 100, then the interface will draw the slider starting at pixel 100. The square will have a length of 20 pixels from the x location of the click and height of 10 pixels above below the line representing the chromosome.

Once this slider is plotted, we then calculate the portion of the chromosome that needs to be magnified. Going ahead with the previous example, if the user clicks on 100 then the magnified view will contain the chromosome represented between 100 and 120 pixels. The starting position can be calculated using the formula:

(pixel-location/ 600) * Chromosome-length

The starting location will be - (100/600) * 4.70 $x10^7$ = 7.83 $x10^6$.

The ending location will be - (120/600) * 4.70 $x10^7$ = 9.40 $x10^6$.

So we now know that the magnified view will contain the subsequences and the main sequences that fall between these two numbers.

We first plot the main sequences. We go through the list of all main sequences and find the sequences that start and end between these two numbers. We represent these main sequences by rectangles with a height of 70 pixels. We also include those sequences that start within the area of interest but do not end in it and also those that end within the area of interest but do not start in that area. We add the sequences that are visible in the magnified area to a separate list called "showingMainSequences". This list is necessary so that when the user clicks on a link we do not search all the sequences. We repeat the above process for the sub-sequences and store those sequences in the "showingSequences" list.

4.4.5 Support Window

The execution of support window, window displaying the repeats in other chromosomes, is very similar to the Magnification window. The support window is launched when a button from the "Support button panel" is clicked. It passes the corresponding chromosome object as input to the window. The views ; Overall and Magnified, are drawn in the same manner as discussed for the Magnified view. The only difference here is that there is no list of main sequences. They only contain the sequences that are repeats of sequences within the chromosome under study (i.e. the chromosome for which the button from "Input button panel" was clicked).

4.5 Interaction

The major interactive component within the software is a mouse click. The click event is prominent within the Magnification Window and Support Window. The difference is that the click event from the Magnification Window propagates to the Support Window and the Main Window. The click event within the Support Window is used to select a sequence and the event is contained within that window.

In order to build this framework, the first thing we need to go back to is the sequence id. All sequences that are related share a unique id. It does not matter whether the sequence is a main sequence or sub-sequence within the magnification window. It has a unique id. This id is then assigned to all the repeats, which can be found in other chromosomes. The idea is that when a

sequence is selected in the magnification window; all repeats within the current window or support window or main window should be highlighted.

When a mouse is clicked in the Magnification window within the "magnified view", we check the location of the click to see if it is within the area specified by any sub-sequence. In order to achieve this we associate a polygon object with each sequence (subsequence or main sequence) when the sequence is plotted in the magnified view.

On click, we then browse through the list of sub-sequences in “showingSequences” (sequences shown in the "magnified view") list to check which sequences contain this point. If there is only sequence, then its information is made visible in the "Sequence Information panel" and the "Gene information panel" shows the list of genes present in this sequence. In order to find this list of genes, we adopt the following algorithm.

4.6 Reading Chromosome file

This is the second file that is read which belongs to the chromosome for which the input button is clicked. As soon as user selects the unique sequence, we pass on the sequence's starting base-pair position, ending base-pair position and sequence length to the following algorithm.

It is parsed using the following algorithm.

1) Read the line.
2) Split into various elements based on space as the separator.

3) Check if the starting base-pair position within the line is same as the one in the selected sequence.

4) If not found, move on to the next line and repeat until we reach end of file or find the matching starting base-pair position.

5) Once we find the line, with matching starting base-pair position

- Create a new sequence object (OrgSeq)
- Add it to the list of genes.
- Keep reading next line and create sequence object until we find a line with matching ending base-pair position.
- Keep adding those objects to the list

6) Return the list.

At the end of the above algorithm we will have the gene sequence list for the chromosome that the user intends to study. The list is sorted by the starting base-pair position. This list is shown in the “Gene Information Panel”.

Once the sequence is set as the selected sequence, that information is passed onto the main window. The main window, which maintains a list of all chromosomes, passes this selected sequence id to all chromosome objects. The chromosome objects then refresh their view to highlight the selected sequence. They do so by checking if the sequence id of the sequence matches that of the selected sequence. If it does then the sequence height is increased to 22 pixels and the color is switched to highlight it.

If there is more than one sequence at the point of click, then the interface passes on the list to the "Overlapping window" The window shows the textual list of these sequences. On selecting one of the sequences from the list, the interface behaves in a similar manner as earlier. It sets the selected sequence id and passes onto the Main window. It then triggers the highlighting of selected sequences within the main window.

If no subsequences are found at the point of click, then we search the main sequences. The main sequences are plotted with a greater height giving the impression of them containing the subsequences but at the same time lying below them in terms of 3D space. We follow the same algorithm as earlier. We loop through showingMainSequence list to find the sequence that was highlighted. Usually, the main sequences will not overlap and hence, we break the loop as soon as we find the first sequence containing the point of click. This sequence will be set as selected sequence.

The Support window also renders the selected sequence in a different color in order to highlight it. Since the support window receives the same chromosome object that is available in the "Main window", it receives the selected sequence id as well. In the overall view, the interface goes through all the sequences to find the selected sequence and once it finds the sequence, it plots it in a different color with an increased height of 20 pixels. For the magnified view, the interface goes the list of sequences, which are showing within the magnified view only. It then plots the selected sequence with a height of 60 pixels and in a different color. Similar to the magnification window, on clicking on a sequence, its related

information comes up in the gene information panel and the sequence information panel. The sequence, which is clicked-on, is identified using the polygon object associated with each sequence, as discussed earlier.

4.7 Launching the Genome browser

At the time of sequence selection, we have the information about the sequence within the sequence object. We use the starting base-pair position and the ending base-pair position as inputs in a URL to launch the genome browser. We make an assumption that the software is running on a windows machine. We then launch the Internet Explorer from within the software and pass the URL with the sequence data as a parameter to the Internet explorer launch command. This command can be overridden to accommodate other browsers on other platforms by checking the Operating system. However, we reserve this implementation for future work.

4.8 Closing Magnification Window

Only one “Magnification window” may be opened at the same time. This is necessary because of the amount of memory involved in calculating the sequences and subsequences. If the “Magnification window” is closed, the “Main window” is informed of the closure all the remaining open support windows are closed. The “Main window” view is refreshed as well. All the chromosome objects have their internal sequence lists reset (main sequences and subsequences).

If a new “Magnification window” is opened while one is already opened, then first we close the window, which is already opened. It follows the same logic as described in previous paragraph. Once the windows are closed, and all sequences are reset in all chromosome objects, we go on to load the information about the new chromosome for which the input button was clicked.

This sums up the overview of internal execution of the GeneXplorer. It uses a simple object based approach with event propagation to simulate the visualization. As it is object based, it is easy to extend this visualization to build additional features as deemed necessary.

Chapter 5

Future work and Conclusion

5.1 Future work

The purpose of designing this application, GeneXplorer, was to graphically depict repeat gene sequences in the Human Genome to support researchers in performing biological analysis. The current version of the application is successful in accomplishing this. But we do acknowledge that this is just the beginning and there are still a lot of improvements that could be made to make this application better. Some of the changes that we plan on implementing in the future are:

- *Importing the data to a Database*

  In the present version of the application the data required for running the application is stored in files. Files were a chosen form of storage because the initial research was to be conducted in local machines and the portability of the files was conducive to this form of research. Now we would like the application to be available to a large group of researchers and this would require a centralized data access. The use of a properly indexed Database would also make the application run faster. Although the current application's performance is sufficient for the present research we would like to like to research into other data-types and storage

mechanisms, as the current method of data search is very memory intensive and requires sequential search.

- *Making the application a web application*

  The current version of the application is a desktop application, which is available online and can be downloaded using Java webstart. In the future we would like to build a web application where the user can run the application on the web server instead of loading the application to the local machine. This will encourage global resource sharing and as the server will handle most of the functionality it will not overtax the local machine.

- *Expanding the research scope to other organisms*

  Currently the application is specifically designed to study gene sequence repeats in the Human genome. We would like to expand out area of research to other organisms as well. In the future we would like to design an application that will dynamically generate the interface depending on the type of organism.

- *Building a 3 dimensional interface to represent overlapping sequences.*

  Presently we pull up a window that lists out the overlapping sequences in a textual format. The user has to select the sequence of interest from the list to view the repeats in genome. We would like to enhance this step graphically and bypass using a separate window by representing the repeats in a 3 dimensional manner.

5.2 Conclusion

GeneXplorer was designed to help researchers to view the repeating gene sequences in such a way that they could see the repeats in the complete human genome. This goal has been accomplished. The requirement of the project were to graphically depict textual information of the gene sequence repeats, to represent the information over the entire genome, to make it user interactive, to collaborate the information with the UCSC Human Genome Browser. To accomplish this we went through multiple paper prototypes and discussed it with our client. Once the client decided on a prototype, we went ahead with the actual coding. After the initial prototype was developed we went through multiple iterations to ensure that we met our clients' needs.

Once the final product was developed we asked some students from the department of Bioinformatics test the application. We gave them a brief overview of the application and asked them to test the application by looking at Chromosome 21 sequences repeats. We gave them the full freedom to use the application as they pleased and observed the way they used the application. Once the evaluation was over we got the some very positive feedbacks like that they found the application very intuitive, easy to understand, helpful in getting a genome level view of elements. Many of them said that they would like to use application in the future.

Based on our user testing we believe we have been successful in meeting with the requirements and have developed an application that can be easily used by the researchers to study repeat gene sequence in the entire human genome.

Reference

[Kent *et al*, 2002] Kent, W. James, Sugnet, Charles W., Furey, Terrence S., Roskin, Krishna M., Pringle, Tom H., Zahler, Alan M., Haussler, and David
The Human Genome Browser at UCSC
Genome Res. 2002 12: 996-1006

[Dombrowski *et al,* 2003] Susan M. Dombrowski and Donna Maglott
NCBI Handbook-Ch 20. Using the Map Viewer to Explore Genomes
Created: October 9, 2002, Updated: August 13, 2003

[Birney et al, 2004] Birney, Ewan, Andrews, T. Daniel, et al.
An Overview of Ensembl
Genome Res. 2004 14: 925-928

[Flicek et al, 2008] P. Flicek, B. L. Aken, et al.
Ensembl 2008
Nucleic Acids Res. 2008 January; 36(Database issue): D707–D714.

[Kashuk et al, 2002] Carl Kashuk, Sanghamitra SenGupta, Evan Eichler, *et al.*
viewGene: A Graphical Tool for Polymorphism Visualization and Characterization
Genome Res. 2002 12: 333-338

[Kurtz et al, 2001] Stefan Kurtz, Jomuna V. Choudhuri, Enno Ohlebusch, Chris Schleiermacher, Jens Stoye, and Robert Giegerich
REPuter: the manifold applications of repeat analysis on a genomic scale
Nucl. Acids Res. 29: 4633-4642.

[Shaying et al, unpublished] Shaying Zhao et al.
Search Repeat-patterns in a Genome
Unpublished

[Lanters et al, 2001] Lander ES et al.
Initial sequencing and analysis of the human genome.
Nature, 2001; 409(6822): 860-921

[Kimball, 1994] John W. Kimball
Kimball's Biology Pages
http://users.rcn.com/jkimball.ma.ultranet/BiologyPages

[Hoffman, 1999] Paul Hoffman, Cognetics Corporation
Usability Interface: Accommodating Color Blindness
Reprinted from Usability Interface, Vol 6, No. 2, October 1999

[Lipman *et al.,* 1985] DJ Lipman, WR Pearson
Rapid and sensitive protein similarity searches.
Science, Vol. 227, No. 4693. (22 March 1985), pp. 1435-1441.

[Altschul *et al., 1990*] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.
Basic local alignment search tool.
J Mol Biol. 1990 Oct 5;215(3):403-10.

[Benson *et al.,* 2007] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL.
GenBank
Nucleic Acids Res. 2008 Jan;36(Database issue):D25-30. Epub 2007 Dec 11.

[Smith *et al.,* 1997] A.F.A. Smit, R. Hubley & P. Green
RepeatMasker
http://repeatmasker.org

[Green, 1993] Green, P. 1993.
cross_match
(http://www.genome.washington.edu/UWGC/analysistools/phrap.htm).

[Green, 1994] Green, P. 1994.
phrap
(http://www.genome.washington.edu/UWGC/analysistools/phrap.htm).

[Watson, 1990] JD Watson (6 April 1990)
The human genome project: past, present, and future
Science **248** (4951), 44. [DOI: 10.1126/science.2181665]