An Integrated Tool for Investigating Genetic Disorder-Relevant Tandem Repeats in Human Genome

Jorng-Tzong Horng[1,2,*], Feng-Mao Lin[1], Ming-Yu Chen[1], Hsien Da Huang[2]

Department of Computer Science and Information Engineering[1],
Department of Life Science[2]
National Central University, Taiwan

Department of Biological Science and Technology & Institute of Bioinformatic[2]
National Chiao-Tung University, Hsin-Chu, Taiwan

horng@db.csie.ncu.edu.tw[*]

Abstract

Tandem repeats (TRs) are associated with human inherited diseases, play a role in evolution, and are important in regulatory processes. Because experts who are researching in genetic disorders may be interested in TRs with particular limits, and they will design primer sequence for experiment. Hence, the objectives of this study are to integrate the information about human genetic disorders, and to provide an efficient tool for observing the information about TRs and genetic disorders are the objectives of this study. To reach the first objective we need to establish a database which integrates gene information, TRs, and OMIM data. Then users can efficiently analyze genetic disorders with this tool. We also need to provide a primer design tool for identifying specific TRs. Then users can obtain interesting primers from specific disease features and can experiment the patient's sample for verifying the suspect. This tool is designed with a user-friendly interface and integrated information for experts to analyze genetic disorders, such as the primer design. In this work, we identify TRs in the complete human genome from the publicly available sequences and mapped to the genes located in. According to the relationship of genes and genetic disorders recorded by OMIM, the TRs, which potentially relevant to the genetic disorders, will be shown.

Keywords: Genetic Disorder, Tandem Repeats

Background

DNA tandem repeat is two or more contiguous occurrences of a specific pattern of nucleotides. Tandem repeats can arise from the tandem duplication, one of the less well understood mutational events in DNA molecules, which occurs to transform a segment of DNA sequence into two or more copies. For example, from DNA sequence TGCCA to TGCCGCCGCCA in which the single occurrence of a GCC pattern the tandem duplication is transformed into three identical, adjacent copies, and the sequence GCCGCCGCC is defined by three consecutive occurrences of a GCC pattern. Hence, tandem repeats, playing a variety of regulatory and evolutionary roles, are important laboratory and analytic tools.

Because of the instability of microsatellites, the amount of repeats can grow from on generation to the next generation [16]. Recently, several reports concerning hereditary diseases, and their association with tandem repeats have been published [5, 20]. The human triplet expansion diseases are predominantly neurological. Within the coding region of the gene or near their regulatory regions the human genetic disorders are caused by instability and expansion of trinucleotide repeats.

Trinucleotide repeats can be found in human genes which are associated with several neurological diseases. These diseases are separated into two categories: non-coding trinucleotide repeat disorders and polyglutamine diseases[5]. The first subclass has its repeats in non-coding sequences, including six diseases: Fragile X syndrome (FRAXA), Fragile XE MR (FRAXE), Friedreich's ataxia (FRDA), Myotonic dystrophy (MD), Spinocerebellar ataxia type 8 (SCA8), and Spinocerebellar ataxia type 12 (SCA12). In addition, FRAXA is caused by expansion of polymorphic $(CGG)_n$ repeat in 5'-untranslated region (UTR) of the gene FMR1 [23]. FRAXE is caused by expansion of $(GCC)_n$ repeat in promoter region of the gene FMR2 [10]. FRDA is caused by intronic GAA repeat expansion in the gene X25 [4], MD is caused by CTG expansion in 3'-UTR of gene DMPK [22]. SCA8 is caused by CTG

expansion in 3'-terminal exon of the gene SCA8 [11]. SCA12 is a disease caused by CAG repeats expansion in 5'-UTR of the PPP2R2B gene [8].

The polyglutamine diseases subclass characterized by exonic $(CAG)_n$ repeats that code for polyglutamine tracts, including: Spinobulbar muscular atrophy (SBMA) or Kennedy's disease [2], Dentatorubral-pallidoluysian atrophy (DRPLA) [9], Huntington's disease (HD) [6, 18], and several forms of spinocerebellar ataxia (SCA1, SCA2, SCA3, SCA6, SCA7) [18].

Besides, human triplet expansion diseases are caused by trinncleotide repeats. Researchers found that with specific length rather than trinucleotide pattern, tandem repeats may associate with genetic disorders. In recent years, one of the significant effort has shown that Spinocerebellar ataxia type 10 (SCA10) was caused by the expansion of ATTCT pentanucleotide repeat [15]. Molecular analysis, moreover, has proved that SCA10 is associated with expansion of the $(ATTCT)_n$ repeat from a normal range of ten to 22 to as many as 4500 copies in intron 9 of the SCA10 gene.

Because of the both instability and expansion of tandem repeats, tandem repeats expansion diseases have a predominantly hereditary component. By the expansion of existing repeats upon transmission from parents to offsprings, tandem repeats expansion diseases can be characterized. The large expansions observed for such diseases as fragile X syndrome, Friedreich's ataxia, and myotonic dystrophy may result from recombination or from aberrant processing of replication intermediates involving unusual DNA structures arising in certain repeat sequences [14, 24]. In afflicted individuals, the unusual number of copies in the specific repeat pattern resulting in the disease. Therefore, for clinicians and experts who research in the genetic disorders, an automatic, and integrated system is needed, and for investigating these genetic disorders and tandem repeats as user's wish has to be provided, an efficient tool.

Online *Mendelian Inheritance in Man* (OMIM) is a catalog of human genes and genetic disorders. OMIM focuses primarily on inherited, or heritable, genetic diseases [7]. OMIM is also a computerized database version of Victor McKusick's book, *Mendelian Inheritance in Man*, provided through the National Center for Biotechnology Information. The major difference between the two resources is that the online version is updated daily.

A distribution analysis of trinucleotide microsatellites in human, mouse, and rat suggested Table **1** shows the distribution of these genes associated tandem repeat sites.

Query interface

The system is a web-based tool for clinicians and researchers to observe the information about tandem repeats and genetic disorders. Users may click on the figure on the home page to start the accessing process of our system. The query interface provides users to define keywords for searching the genetic disorders, genes, and tandem repeats.

Graphical browsing interfaces

The system provides a graphical interface for users to

that the microsatellite motif CAG is one of the most abundant microsatellite motifs in human GenBank DNA sequences and is the most abundant microsatellite found in exons [20]. Stalling mentioned that this fact may explain why CAG repeats are thus far the predominant microsatellites expanded in human genetic diseases. Another analysis has discussed the distribution and association of tirnucleotide microsatellites with genes and other genomic regions [21]. This study revealed that AGC and CCG repeat were predominantly presented in the coding regions of the genome while UTRs and upstream sequences contained CCG repeats in relative abundance. Analysis of density of triplet repeats (bp/Mb) revealed that AAT and AAC were the abundant repeats whereas ACT and ACG were the rare repeats found in human genome. As Subramanian mentioned, their identification of 171 genes which contain a minimum of ten repeat copies will be of particular interest in future in correlating their association with any disease phenotype due to the expansion potential of repeats present in them.

For integrating the information about human genetic disorders, and providing an efficient tool for observing the information about tandem repeats and genetic disorders, we established a database based on genetic disorder in man. The database integrates not only genetic disorders from OMIM but also tandem repeats and gene information. Users can efficiently analyze genetic disorders with this tool. As well as a primer design tool for identifying specific tandem repeats has been provided, users can obtain interesting primers from specific disease features and experiment the patient's sample for verifying the suspect.

Results

The possible tandem repeats have been searched in the human genome, and the occurrence of repeats in the genomic regions has been identified based on the annotation of the human genome sequence in the GenBank database. The expansion of repeats in the coding region of the gene has been found that being associated with genetic disorders in the OMIM database. These data was stored in our database, and a web site was designed for accessing these information. There are 24 chromosomes, which contain 26,179 candidate genes, 1,246,831 distinct tandem repeat patterns and 34,263,072 tandem repeat sites. Most of these tandem repeat sites represented in non-genomic regions, only nearly 2% of them represented in such genomic regions as exons, introns, upstream and downstream of genes.

observe the information about tandem repeats, and we have referred to the accessing style of other public bioinformatics website such as Ensembl. An overview of the flowchart for accessing the web pages is shown in Figure 1. After users enter the introduction page of our system, the 'chromosome selection' provides users to browse the subclass of this system by choosing one of the human chromosome. 'Map view' provides the summary and distribution information of tandem repeats for the specified chromosome, chosen by the user in last step. Figure 2 shows an example of this function. The "info" field provides the summary about this chromosome which includes the length of the chromosome, the counting of tandem repeat sites according to length, copies, and location, the number of candidate genes, and the number of diseases from NCBI OMIM database (see A in Figure 2). By defining features at part B in Figure 2, users may make a keyword search in this chromosome, and by changing the box at part C in Figure 2, users may easily jump to browse another chromosome. The part of D in Figure 2 shows the distribution of tandem repeats and candidate genes. The diagram of bar charts shows the distribution of tandem repeat sites according to length and copies in different regions. It helps users to meet the suitable location in this chromosome, and users may browse their interest region by clicking on this graph.

'Contig view' provides the distribution of tandem repeat sites and occurrence of tandem repeats in coding regions: The mechanism of the refinement on this page supports user to meet the suitable, concerning tandem repeat sites. Figure 3 gives an example to illustrate this view. At part A in Figure 3 the figure denotes the region chosen by users. The "overview" field gives the summary information of cytogenetic-band, contigs, genes and diseases. The gene which locates at sense would be presented in dark blue, otherwise it would be presented in light blue (see B in Figure 3). The "detailed view" shows the distribution of tandem repeat sites and genes in specified arguments (see C in Figure 3), and it provides following functions.

Jumping and resizing. At the top of this field, there is a jumping and resizing block. Users can jump to a specified region of the genomic sequence (see D in Figure 3) or resize this view to show 1kb, 10kb, 50kb, 100kb, 200kb, 500kb, 1mb, or 2mb of the genomic sequence (see E in Figure 3).

Limiting the view. Users can define the limitation of tandem repeat sites to meet such the suitable tandem repeats as length (see F in Figure 3), copies (see G in Figure 3), pattern (see H in Figure 3) and location (see I in Figure 3). Therefore, the refinement mechanism would be activated to show this view in the new display parameters. If the tandem repeat site and gene locate at sense, the bars will be presented in blue. Otherwise, they will be presented in pink.

Getting summary information. Users can continue viewing the information of tandem repeats or genes in the 'TR list' or 'gene list' respectively. By clicking on the bar directly, the TR/gene which locates in the place where the user just clicked will be shown in TR/gene list. By clicking on the 'browse' button at the right-hand side of the bar chart (see J in Figure 3), a group of TRs/genes will be shown in TR/gene list respectively.

"TR list" provides the summary about the tandem repeats are chosen by the user, including the genes that contain this tandem repeat in its genomic regions and genetic disorders that associated to these genes. Figure 4 gives an example that describes this list. If the "Gene" field contains the graph, it means that the tandem repeat occurs in the genetic region. If the gene locates at sense, it will be presented in blue. Otherwise, it will be presented in orange. The occurrence of tandem repeats would be drawn under the graph of gene in red. By clicking on the checkbox at the left side of every tandem repeat, users can select the tandem repeats for the "TR selection for primer design" process later.

"TR selection for primer design" provides users to setup the parameters for primer design stage next. An example is shown in Figure 5. There is an introduction graph at the top of this page, which denotes some description about the parameters used in primer design. By clicking on the checkbox at the left of every record, users can select these tandem repeats continue to primer design.

"Primer result" offers the primer sequence for clinicians and researchers to experiment on the specified tandem repeats that they choose, and it is the final step of this tool. Figure 6 gives an example of the primer sequences designed for the 'atgca' repeat pattern chosen from the first entry in Figure 5. "5'Sequence" and "3' Sequence" in Figure 6(a) are the sequences for primer design. Complete sequence was printing in web user interface. At the bottom of Figure 6(a) is the constraints assigning in Figure 5 were showing. Figure 6(b) shows the candidate primers and restriction enzymes. In Figure 6(c) shows location of primer is showing.

Summary and future work

The existence of genetic disorders associated with the expansion of tandem repeats raises the interests of clinicians and experts who research in inherited diseases. As a tool, it may help clinicians and experts' studies in observation of the relationship between tandem repeats and genetic disorders. With a graphical user interface, it integrates not only the information about tandem repeats, genes, and genetic disorders but also primer design tool. Generations of the genomic regions and tandem repeats sites have been done and they have been mapped to genetic disorders in this study. Observing on these data as users' wish for retrieving tandem repeats that are associated with disorder-mapped genes can be reached via our tool, and then it provides primer sequences for experiment to verify the suspect.

Ensembl Human Disease View is a tool in Ensembl Human Genome Browser [3]. The disease view provides the disease name, Ensembl ID, HUGO synonyms, OMIM ID, OMIM web linkage, and chromosome cytolocation for each disease in OMIM. Ensembl ID is gene information related with the disease. Although Ensembl provides graphical interface for genes and repeats, it neither integrates the relationships among diseases, genes, and tandem repeats nor provides primer sequence for clinicians to experiment.

In the future, the information integration is an important issue for developing this tool. The information of homolog genes is important in mapping human genes in GenBank to the disorder-mapped genes recorded by OMIM. For example, users who browsing the cytogenetic-band 22q13.3 in chromosome 22 with our system would gain the tandem repeats ATTCT associated with gene E46L instead of the SCA10 gene, which was the disease SCA10 associated gene in OMIM. This should be done for making smooth-going of retrieving genetic disorders associated tandem repeats via this tool.

Furthermore, the information of single nucleotide polymorphisms (SNPs) can be integrated into our system for clinicians who research in human diseases. In recent years, SNPs are being extensively utilized as markers of choice in disease linkage studies [12, 13]. The data of SNPs can be obtained from the population SNP databases, such as dbSNP [19], which is a database stores genome variations in NCBI. These data integration will make our system as convenient as possible for providing comprehensive information about genetic disorders.

Materials and Methods

The present system contains two parts as being shown in Figure 7. They are data processing and result display. From GenBank the DNA sequences of *Homo sapiens* have been used for identifying tandem repeats, and the gene coding regions from GenBank have been used for retrieving gene location data. By combining these data, the occurrence of tandem repeats in such the genomic regions as exons, introns, upstream and downstream has been identified. As of Release 134.0 in February 2003, DNA sequences of Homo sapiens from GenBank contained 26,179 genes and nearly 2.86 billion nucleotide bases in a total of 24 chromosomes. The relationship between human genes and genetic disorders from OMIM has been retrieved, and there are 14,472 entries as of May 2003. From the above steps, the information of tandem repeats mapping genes and genetic genes mapping disorders has been used for generating the association of tandem repeats with disease phenotypes. The part of result representation has been developed with a user-friendly interface in PHP. Users can access the database via graphical web pages. The mechanism of query-refinement helps users for browsing in tandem repeats efficiently. To provide primer sequence for experiment on user-interesting genomic sequence regions, the primer design tool has been integrated into our system, and the primer3 has been applied to fit this purpose. Primer3 pick primers from a DNA sequence [17], and it can avoid choosing primers in transposable elements and can pick oligonucleotide for probe or primers.

The data processing phase

Tandem repeat identification

An overview of the data processing flowchart is shown in Figure 8. The gbk files of all chromosomes of *Homo sapiens* from GenBank[1] have been used for extracting the data of contigs, genes, and coding regions. The information retrieved from NCBI Map Viewer[2] has also used for generating the location of cytogenetic-bands and contigs in each chromosome. The contig sequences of GenBank are used for tandem repeats identification. The theoretically possible tandem repeats were searched by Tandem Repeat Finder, which the limit of pattern length we consider is up to 500. Tandem Repeat Finder is a popular program to locate tandem repeats in genome sequences [1]. It uses an algorithm for finding tandem repeats which works without the need to specify either the pattern or pattern size.

Marking TRs to genomic regions

The occurrence of the tandem repeats in genetic regions moreover was identified by combing the coding regions data from last step. The tandem repeats location information in Figure 8 denotes which tandem repeat locates in which genetic regions such as exons, introns, upstream, and downstream. The flag also would be marked if there is some mention about this tandem repeat in the annotation of any OMIM entry.

Mapping TRs to associated genetic disorders

In Figure 8, the step of mapping genes to OMIM entries has used to retrieve the genetic disorders and has related genes from OMIM. There is another procedure generated gene data by combining the output of processing data from OMIM and GenBank as described above. By combining these two relationships of tandem repeats to genes and genes to genetic disorders, the relationship between TRs and associated genetic disorders has been built.

The output files generated from these data processing procedures have been loaded into database via the tool named SQL*Loader, which is

[1] ftp://ftp.ncbi.nlm.nih.gov/genomes/h_sapiens/

[2] http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi

developed by Oracle Corporation and loads data from external files into tables of an Oracle database.

References

[1] Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Res, 1999. **27**(2): p. 573-80.

[2] Brooks, B.P. and K.H. Fischbeck, *Spinal and bulbar muscular atrophy: a trinucleotide-repeat expansion neurodegenerative disease.* Trends Neurosci, 1995. **18**(10): p. 459-61.

[3] Clamp, M., et al., *Ensembl 2002: accommodating comparative genomics.* Nucleic Acids Res, 2003. **31**(1): p. 38-42.

[4] Cossee, M., et al., *Frataxin fracas.* Nat Genet, 1997. **15**(4): p. 337-8.

[5] Cummings, C.J. and H.Y. Zoghbi, *Fourteen and counting: unraveling trinucleotide repeat diseases.* Hum Mol Genet, 2000. **9**(6): p. 909-16.

[6] Group, H.s.D.C.R., *A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group.* Cell, 1993. **72**(6): p. 971-83.

[7] Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2002. **30**(1): p. 52-5.

[8] Holmes, S.E., et al., *Expansion of a novel CAG trinucleotide repeat in the 5' region of PPP2R2B is associated with SCA12.* Nat Genet, 1999. **23**(4): p. 391-2.

[9] Ikeuchi, T., et al., *Dentatorubral-pallidoluysian atrophy: clinical features are closely related to unstable expansions of trinucleotide (CAG) repeat.* Ann Neurol, 1995. **37**(6): p. 769-75.

[10] Knight, S.J., et al., *Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation.* Cell, 1993. **74**(1): p. 127-34.

[11] Koob, M.D., et al., *An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8).* Nat Genet, 1999. **21**(4): p. 379-84.

[12] Kruglyak, L., *Prospects for whole-genome linkage disequilibrium mapping of common disease genes.* Nat Genet, 1999. **22**(2): p. 139-44.

[13] Martin, E.R., et al., *Analysis of association at single nucleotide polymorphisms in the APOE region.* Genomics, 2000. **63**(1): p. 7-12.

[14] McMurray, C.T., *DNA secondary structure: a common and causative factor for expansion in human disease.* Proc Natl Acad Sci U S A, 1999. **96**(5): p. 1823-5.

[15] Rasmussen, A., et al., *Clinical and genetic analysis of four Mexican families with spinocerebellar ataxia type 10.* Ann Neurol, 2001. **50**(2): p. 234-9.

[16] Richard, G.F. and F. Paques, *Mini- and microsatellite expansions: the recombination connection.* EMBO Rep, 2000. **1**(2): p. 122-6.

[17] Rozen, S. and H. Skaletsky, *Primer3 on the WWW for general users and for biologist programmers.* Methods Mol Biol, 2000. **132**: p. 365-86.

[18] Sermon, K., et al., *PGD in the lab for triplet repeat diseases - myotonic dystrophy, Huntington's disease and Fragile-X syndrome.* Mol Cell Endocrinol, 2001. **183 Suppl 1**: p. S77-85.

[19] Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.

[20] Stallings, R.L., *Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases.* Genomics, 1994. **21**(1): p. 116-21.

[21] Subramanian, S., et al., *Triplet repeats in human genome: distribution and their association with genes and other genomic regions.* Bioinformatics, 2003. **19**(5): p. 549-52.

[22] Timchenko, N.A., et al., *Molecular basis for impaired muscle differentiation in myotonic dystrophy.* Mol Cell Biol, 2001. **21**(20): p. 6927-38.

[23] Tolmacheva, E.N. and S.A. Nazarenko, *Polymorphism of trinucleotide repeats at loci FRAXA and FRAXE in the population of Tomsk.* Genetika, 2002. **38**(2): p. 268-73.

[24] Wilmot, G.a.W., ST, *Chapter 1, A new mutational basis for disease.*, in *Genetic Instabilities and Hereditary Neurological Diseases*, W.R.a.W. ST, Editor. 1998, Academic Press: San Diego, CA. p. 3-12.

Table 1. The distribution of tandem repeat sites which represented in various genomic regions.

Region	Occr.	Ratio
Exon	127,537	1.83%
Intron	6,574,552	94.37%
Upstream	127,519	1.83%
Downstream	137,212	1.97%
All	6,966,820	100.00%

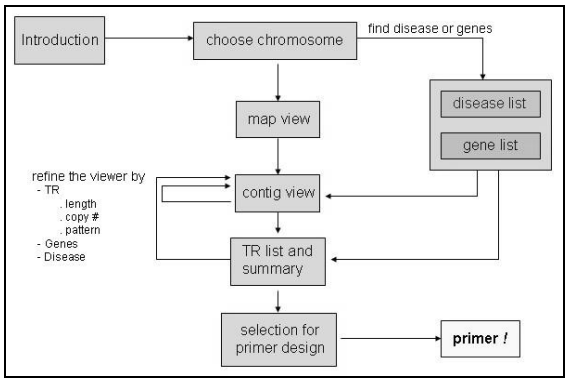


Figure 1. The flowchart of the web interfaces of the system.

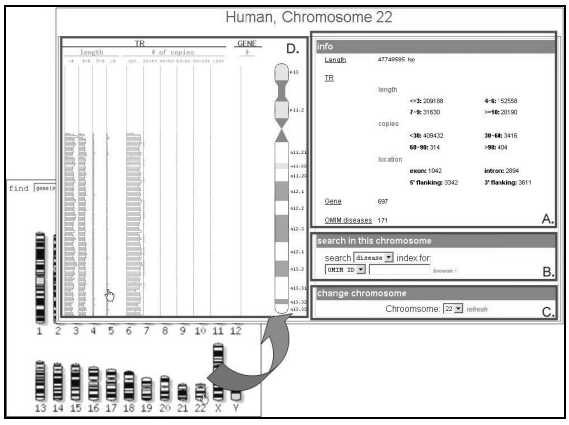


Figure 2. Chromosome summary information includes tandem repeats, genes and diseases. This is an example after selecting the chromosome 22 in 'chromosome selection' step.

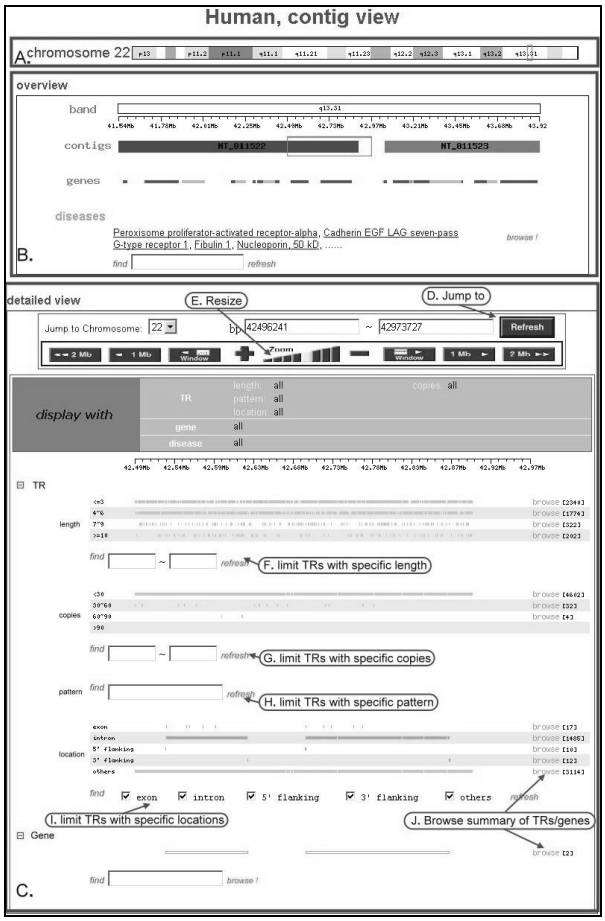


Figure 3. Detail information for a chromosome sub region.

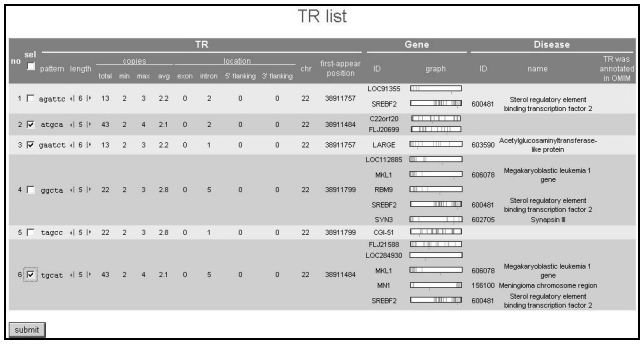


Figure 4. Detail information about the relationships among tandem repeats, genes and disorders.

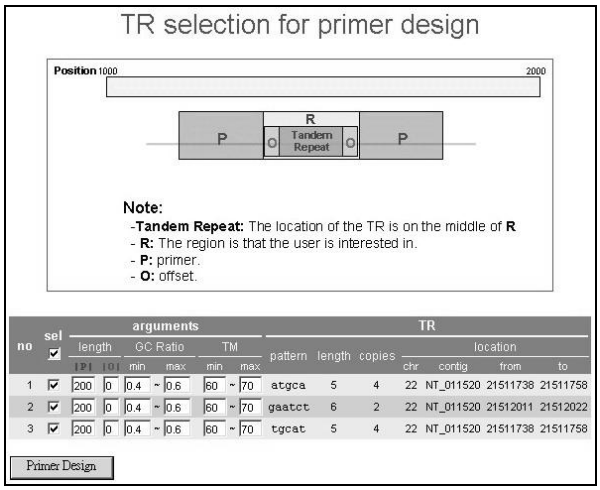
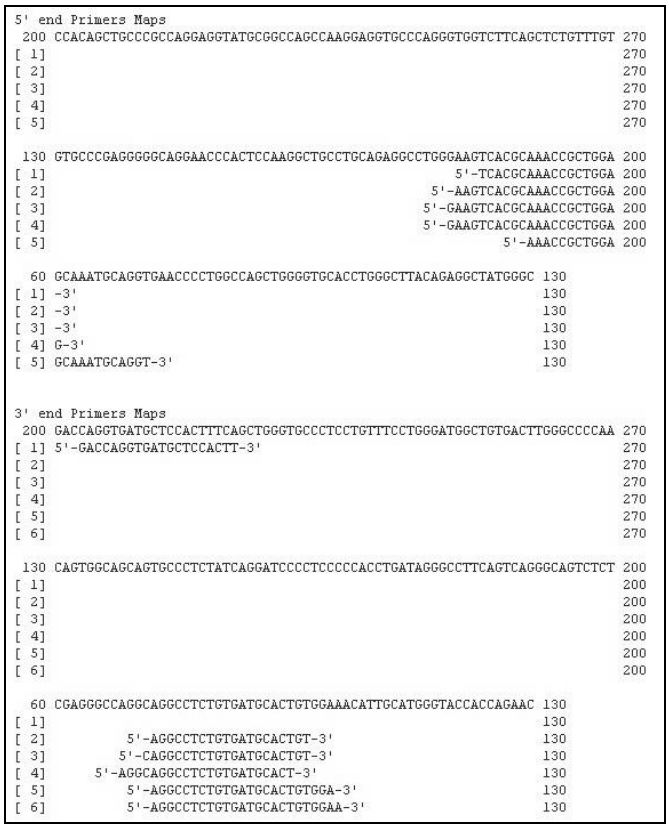


Figure 5. User interface for primer design with specific features.

(a)

(b)



(c)

Figure 6. Primer information for specific tandem repeats.

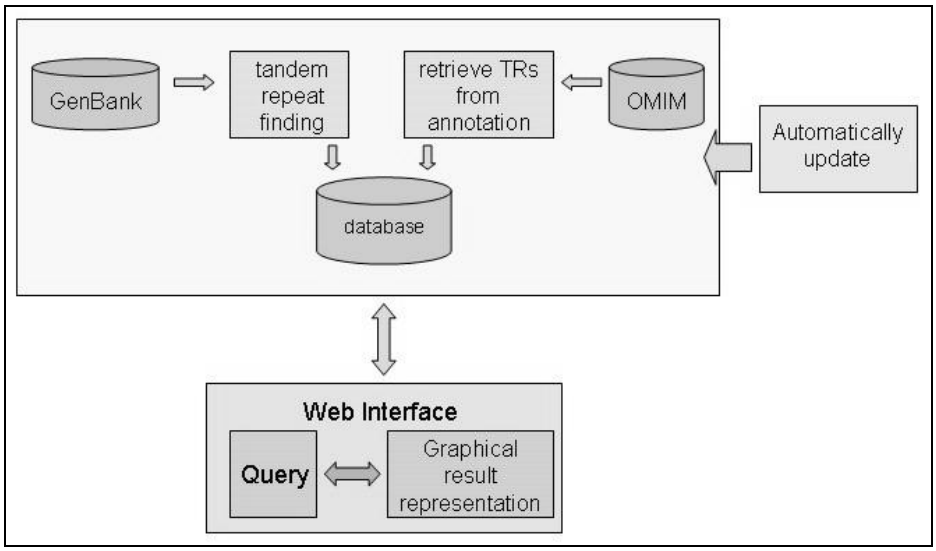


Figure 7. The system architecture contains two parts, data processing and web interface.

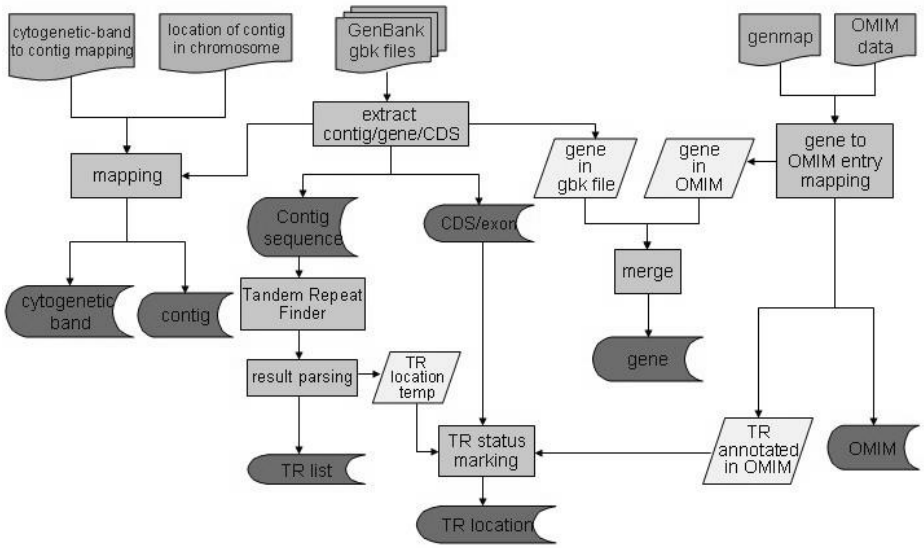


Figure 8. The data processing phase of this system.