

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

Feng-Mao Lin¹, Ming-Yu Chen², Hsien-Da Huang³, Jorng-Tzong Horng⁴

Keywords: genetic disorder, tandem repeats, single nucleotide polymorphisms

1 Introduction.

Tandem repeats (TRs) and single nucleotide polymorphisms (SNPs) are associated with human inherited diseases, play an important role in evolution and in regulatory processes [1-3]. Because the biologists, who are investigating in genetic disorders, are also interested in TRs and SNPs with particular limits, and they will design primer sequence for experiment. Hence, the goal in this work is to develop an effective tool for observing the information about TRs, SNPs and genetic disorders. We establish a database to store gene information, TRs, SNPs and OMIM [4] data. The system facilitates the analysis of genetic disorders. We develop a primer design tool for identifying specific TRs or SNPs when users interested in specific disease features. Web user interfaces and graphical interfaces are designed and implemented. Accordingly, the relationship of genes and genetic disorders recorded in OMIM are found. The main contribution of this work is to provide a user-friendly and effective tool for genetic disorders in the research of human inherited diseases.

2 Materials and methods.

The present system contains two parts. They are data processing and result display. From GenBank [5] 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. The relationship between human genes and genetic disorders from OMIM [6] has been retrieved. 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 [7], and it can avoid choosing primers in transposable elements and can pick oligonucleotide for probe or primers.

3 Results.

¹ Department of Computer Science and Information Engineering, National Central University, Taiwan. E-Mail: meta@db.csie.ncu.edu.tw

² Department of Computer Science and Information Engineering, National Central University, Taiwan. E-Mail: puenny@db.csie.ncu.edu.tw

³ Department of Biological Science and Technology & Institute of Bioinformatics, National Chiao-Tung University, Hsin-Chu, Taiwan. E-Mail: bryan@mail.nctu.edu.tw

⁴ Department of Computer Science and Information Engineering and Department of Life Science, National Central University, Taiwan. E-Mail: horng@db.csie.ncu.edu.tw

The possible tandem repeats have been searched in the human genome by Tandem Repeat Finder [8], 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. Table 1 shows the distribution of these genes associated tandem repeat sites.

| 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% |

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

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.

References

- [1] Cummings, C. J. and Zoghbi, H. Y. 2000. Fourteen and counting: unraveling trinucleotide repeat diseases. *Hum Mol Genet* 9:6 909-16.
- [2] Stallings, R. L. 1994. Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases. *Genomics* 21:1 116-21.
- [3] Subramanian, S., Madgula, V. M., George, R., Mishra, R. K., Pandit, M. W., Kumar, C. S., and Singh, L. 2003. Triplet repeats in human genome: distribution and their association with genes and other genomic regions. *Bioinformatics* 19:5 549-52.
- [4] Parton, M. J. 2003. Online Mendelian Inheritance in Man OMIM: www.ncbi.nlm.nih.gov/entrez. *J Neurol Neurosurg Psychiatry* 74:6 703.
- [5] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. 2003. GenBank. *Nucleic Acids Res* 31:1 23-7.
- [6] Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., and McKusick, V. A. 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 30:1 52-5.
- [7] Rozen, S. and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365-86.
- [8] Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27:2 573-80.