Linking Multiple Genes to Human Chromosomal Locations to Facilitate Knowledge Discovery

Tongsheng Wang[1, 2] Marc Ma[2, 3] Patricia Soteropoulos[1]

[1]Center for Applied Genomics, Public Health Research Institute, Newark, NJ 07103
[2]Department of Computer Science, and [3]Center of Applied Mathematics and Statistics, New Jersey Institute of Technology, Newark, NJ 07102

Abstract

A web-based tool, LinkGenes, has been developed for linking multiple genes to human chromosomal locations to facilitate knowledge discovery from DNA microarray experiments. LinkGenes consists of three functional elements: a MySQL database with gene annotation information, a serverlet container powered by Jakarta™ Tomcat 4.0, and web-user interface supported by HTML, JSP, and javascript. Upon a user's request, chromosomal mapping results are presented to the user in both a graphical viewer and tabular format. Using LinkGenes, we have been able to identify several human chromosomal regions that could be linked to either gene expression phenotypes or metabolic pathways. Genetic disease information associated with these genes is also obtainable through this tool.

Keywords: DNA microarray, chromosome mapping, multiple genes, gene expression, genetic disease.

1. Introduction

Microarray-based gene expression profiling has become a standard technique in many biological research laboratories. By allowing researchers to monitor thousands of mRNA levels simultaneously, DNA microarray technology provides opportunities to gather and analyze large amounts of data on gene function and regulation rapidly, increasing the speed of research [1]. However, RNA expression profiles provide insight into only the first level of gene regulation. The regulation of gene expression and gene products occurs at the transcriptional, translational and post-translational levels. Moreover, alternative splicing (AS) and/or alternative polyadenylation (AP) events most likely result in mRNA products that differ from those to which the oligonucleotide or cDNA probes on the array were designed [2,3]. In addition, for genes expressed at low levels, the signal values computed from pixel intensities are likely to be near the background of the chip and overwhelmed by the noise. Therefore some transcripts are below the level of detection of current microarray technology.

Over the past few years, signal extraction techniques and computational algorithms have been developed to facilitate the process of microarray analysis [4-6]. Each method has its advantages and disadvantages in addressing differential expression of genes on a large scale. However, there are interesting transcripts that due to AS, AP or low expression levels simply cannot be detected no matter how robust the statistics and the computation techniques. To address the issues associated with AS, AP and low expression levels, one must apply suitable biological knowledge and well designed bioinformatics techniques. Based on the observation that a large number of human expression phenotypes, genetic diseases and metabolic pathways show significant evidence of linkage to specific chromosomal locations [7, 8], we have made an initial effort to address gene discovery by linking genes with similar expression patterns to their genomic locations. We have designed a Web-based tool, LinkGenes to map gene expression data to specific chromosomal coordinates. This paves a way for finding new biomarkers and studying the expression and regulation of genes that are missed in expression microarray analysis, by simply searching though the defined chromosomal regions. LinkGenes for human chromosome mapping is our first step towards the more complex task of systematic large-scale microarray data mining.

Here we present the design of LinkGenes. A fully functional prototype is available to the public at http://siriusb.umdnj.edu:8081. Using LinkGenes, one can batch map human genes to their chromosomal locations. LinkGenes allows users to input GenBank, Unigene or Gene IDs to retrieve mapping results in both a tabular and graphical format directly on the client's web browser. We also strive to make genetic disease information about those genes retrievable through this tool. To test and validate the system, we have analyzed microarray experiments performed

using the Center for Applied Genomics (CAG) human 19K oligonucleotide array. Out of a total of approximately 19,000 genes, 2,177 genes exhibiting greater than 2-fold change in expression level between the two samples tested were processed by LinkGenes. Four chromosomal regions were found containing groups of genes which had similar expression patterns. We also showed that genetic disease information relating to the genes of interest can be easily retrieved.

2. Design of LinkGenes

2.1. Basic functionalities of LinkGenes

There are two basic functionalities available in LinkGenes:

- Genes and Locations: Maps a batch of genes to their chromosomal location and presents results in both tabular and graphical formats.
- MIM and Disease: Takes as input a batch of genes and retrieves relevant Mendelian Inheritance in Man (MIM) [9] disease information and presents it in a table.

2.2. Data sources

Human genomic annotations and human genetic disease information is publicly available. We have used four public data files to build the chromosome mapping database, which were downloaded from ftp://ftp.ncbi.nih.gov/gene/DATA/ and ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/maps/mapview/BUILD.35.1/. Another data file, mim.q, was also obtained from the latter website to set up genetic disease information database.

The chromosome mapping database was built from the following four data files: loc2acc, gene2unigene, gene.q, and seq_gene.md. File loc2acc has six columns. The first two columns contain Gene ID to GenBank ID mapping information. There are some entries in the GenBank ID column with "none". Those entries were not used because they are not yet well annotated. File gene2unigene only has two columns: the first column contains Gene ID and, the second column contains the corresponding Unigene ID. Only those Unigene IDs starting with "Hs." signifying they are human genes were used. The file gene.q contains 13 columns and the file seq_gene.md contains 15 columns. By using Gene ID as the key, corresponding gene description, gene symbol and cytogenetic information were obtained from the file gene.q. Chromosomal and contig coordinate information as well as strand information were obtained from the file seq_gene.md. Out of the ten columns of the file mim.q, Gene ID, MIM ID and description were used in our system.

2.3. Microarray data collection

The Human 19K microarray consists of approximately 19,000 oligonucleotide 65-mers (Compugen Human OligoLibrary™) each representing a single human gene. The arrays were spotted onto poly-l-lysine-coated glass microscope slides using an OmniGrid100™ microarraying robot. Labeling of two RNA samples and hybridization were performed using the Genisphere™ Array350 Kit following the protocol provided by the manufacturer. The arrays were scanned in a GenePix™ 4000B scanner and the data extracted using the GenePix™ Pro 5.0 software. Data filtering was performed by Microsoft Excel. Genes with a greater than 2-fold change in expression level were selected for further testing by LinkGenes.

2.4. Configuration and implementation of LinkGenes

The system configuration of LinkGenes is illustrated in Fig. 1. It is a three-tier system configured for human chromosome mapping and genetic disease information retrieval. Gid, Gbid Unigid, and Chr represent for Gene ID, GenBank ID, Unigene ID and chromosome respectively as used by NCBI for genome annotation. MIM is a genetic disease annotation system for human genes.

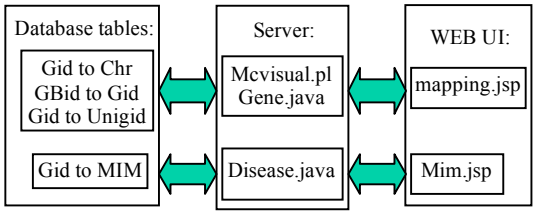


Fig. 1: A three-tier system configuration for human chromosome mapping.

The genomic mapping information is built into a relational database on the backend. A server was set up in the middle to accept user requests and gather information from the backend database. The front-end web user interface (Web UI) allows the user to perform chromosome mapping and disease information retrieval by accepting the user input and presenting the results back to the user.

The system is flexible in allowing users to input gene identifiers. Three types of gene identifiers are allowed, including Gene ID, GenBank ID and

Unigene ID. The system transforms the latter two into Gene ID prior to performing the chromosome mapping or disease information retrieval functions. Therefore the backend relational database is built by taking advantage of the fact that each Gene ID is unique in NCBI gene functional annotations. In our implementation we used Perl programs to parse downloaded data files with gene and genetic disease annotation information into tab-delimited flat files. The resulting flat files were used to populate the backend MySQL database, which was constructed for storage and query of data. The serverlet container is powered by Jakarta™ Tomcat 4.0 (http://jakarta.apache.org/tomcat/) and the server programming was performed using JAVA™. Graphical presentation was implemented using BioPerl [10] and Web UI constructed with JSP™, HTML and javascript.

2.5. How to use LinkGenes

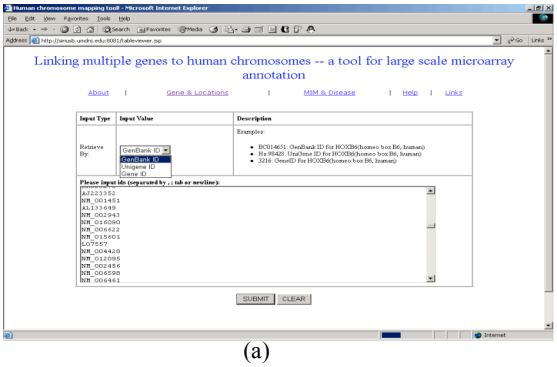
To use LinkGenes, the user first inputs a list of genes, defined by any of three types of identifiers, GenBank ID, Unigene ID or Gene ID. After the user submits a request, the Web UI JSP programs pass the request to server, and the server retrieves information from the database and organizes it into the appropriate format for client-end web presentation. When a user submits a large number of genes in one task, server response delay can be expected due to network traffic on the server. In case of submission of incorrect gene identifiers, all information about those genes will be returned as "NA".

3. Evaluation of LinkGenes

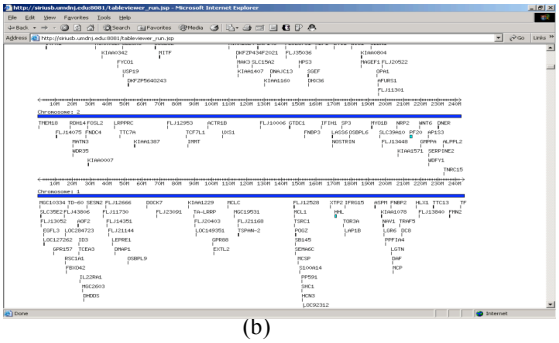
3.1. Chromosome mapping

Our gene mapping database was populated with 26,834 well-annotated genes along with their mapping information. To test and demonstrate the functionality of our system, we used our proprietary microarray data as an example. We chose to submit for mapping 2,177 genes exhibiting a greater than 2-fold change in expression between the two samples used for the experiment. Fig. 2 illustrates part of the Web UI for input and mapping results. We analyzed gene localization on chromosomes #11 and #12, and identified two regions on each chromosome with clusters of genes falling onto an approximately 5Mb window. Detailed numerical information about these chromosomal regions as is shown in Table 1.

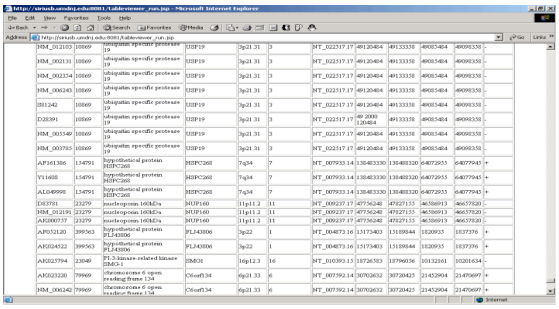
An important caveat to the definition of a clustered genomic location is the choice of window size. Since gene density is not uniformly distributed throughout the genome, rigorous statistics need to be assessed before any window size can be chosen. Currently we are working on defining such window sizes, which are highly likely to be genomic location specific.



(a)



(b)



(c)

Fig. 2: Mapping 2177 human genes to chromosome. (a) Web UI. (b) Graphic format of mapping results with chromosome coordinates indicated. (c) Tabular format of mapping results with genomic coordinates for each gene.

Table 1. Specific regions on chromosomes #11 and #12 defined by mapping human microarray data to genomic location. "From" and "To" are chromosomal coordinates in million bases (Mb).

Chromosome region		From	To	# of genes
#11	Region 1	59.92	64.37	7
	Region 2	68.57	74.03	9
#12	Region 1	26.16	30.80	5
	Region 2	47.95	52.00	8

3.2. Disease information retrieval

We also retrieved disease information about the same 2,177 genes used in chromosome mapping. Since only a small part of annotated genes are known to be associated with genetic diseases, our genetic disease information database only has a collection of 2,429 genes, which is less than 10% of genes (26,834) covered by the gene mapping database. Therefore, the disease information for most genes is unavailable, and the table entry appears as "NA" as shown in Fig 3. Nevertheless, the information retrieved can be of great use for studying genes associated with genetic diseases.

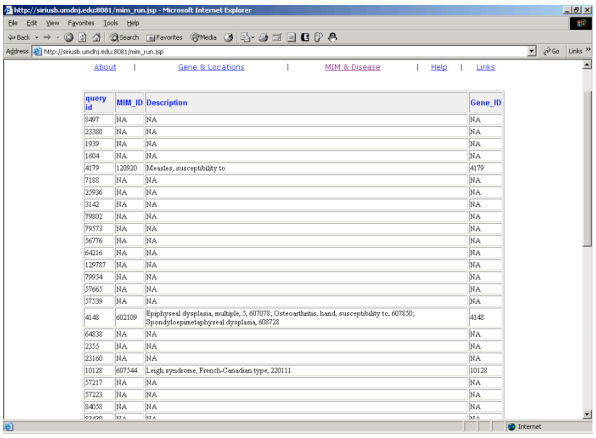


Fig. 3: Results of retrieving genetic disease information for 2177 genes. Not surprisingly, results for most genes are unavailable.

4. Summary and Future work

In this paper, we presented LinkGenes, a novel microarray data mining tool for linking multiple genes to human chromosomal locations to facilitate exploration of gene expression regulation and genetic disease association from microarray experiments. Since annotation of the human genome is still in progress and tremendous change may occur in the NCBI genome annotation data, periodic updating of our database will be necessary. In addition, mapping of genes to their genomic locations is only our first step toward gene discovery through linking gene loci to expression levels.

Statistical analysis of gene distribution across the human genome is underway in order to define a proper window size. Using properly defined window sizes and data for transcripts which we can detect, we hope to make accurate predictions about the expression of genes that are not detectable by conventional microarray analysis. Finally, we plan to develop algorithms for using gene expression data to predict genetic abnormalities and disease association.

5. References

[1] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95:14863-14868.

[2] Kim, P., Kim, N., Lee, Y., Kim, B., Shin, Y. and Lee, S. (2005) ECgene: genome annotation for alternative splicing. *Nucl. Acids Res.* 33:D75-D79.

[3] Zhang, H., Hu, J., Recce, M. and Tian, B. (2005) PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucl. Acids Res.* 33:D116-D120.

[4] Yang, Y.H., Dudoit, S., Luu, P., Li, D.M., Peng, V., Ngai, J. and Speed, T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucl. Acids Res.* 2002 Feb 15;30(4):e15.

[5] Li, C & Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* 98: 31-36.

[6] Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed. (2003) Summaries of Affymetrix GeneChip probe level data *Nucl. Acids Res*. 31:e15.

[7] Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S. and Cheung, V. G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* 430: 743-747.

[8] Korbel, J. O., Jensen, L. J., Mering, C. V. and Bork, P. (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nature Biotechnology* 22:911-917.

[9] Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucl. Acids Res*. 33:D514-7.

[10] Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H. et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12:1611-161.