Development of a co-regulated gene expression analysis tool (CREAT)

By

Min Wang

April, 2003

Project Documentation

Description of CREAT

CREAT (coordinated regulatory element analysis tool) are developed to assist in the prediction of regulatory elements for coordinated genes. Coordinated genes are those genes that exhibit same expression patterns under certain conditions. Most likely these genes are under the control of same regulatory elements under these conditions. The most important regulatory mechanism of transcription is the binding of a particular transcription factor to its corresponding regulatory element of its target gene. This binding will lead to an activation or inhibition effect on the expression of its target genes. CREAT provides a convenient way to find these regulatory elements, though further confirmation depends on experimental results.

With largely sequenced human genome, researchers are beginning to focus their interest on gene expression studies. Microarray and serial analysis of gene expression (SAGE) are two techniques that enable researchers to screen gene expression patterns of thousands of genes in parallel. They are very powerful techniques to identify co-regulated genes. CREAT was designed as a resource to assist in the analysis of gene regulation from large amount of microarray or SAGE data. It is intended to assist biological researches as well as computer analysis, in the study of human transcription signals.

CREAT is based on the information collected in promoter regulatory elements database. Promoter regulatory elements database contains information about promoter sequences and transcription factors that bind to the promoter sequences predicted by *MatInspector*. Large scale gene expession data such as results generated from microarry or SAGE experiments, can be analyzed using CREAT.

Initially the database was supported by the Center for Bioinformatics and Computational Biology (CBCB), University of Iowa. Currently, CBCB is using Sybase as its database server for microarray database constructed. The database is now supported by IDEA (Iowa Database and Emerging Applications) lab, Department of Computer Science at the University of Iowa. IDEA lab is using mySQL as the database server.

Description of promoter regulatory elements database

1. Data Processing

The structure and organization of HPD reflects the purpose of serving for microarray data mining and data analysis. In detail, upstream DNA sequences of 2004 human genes are downloaded from human promoter database website (http://zlab. bu. edu/~mfrith/HPD. html). All these promoter DNA sequences are further analyzed for potential transcription factor binding sites by using *MatInspector* from Germany (http://www.gsf.de/biodv/matinspector.html). The specific sequences and positions of transcription factor binding sites on each promoter region are indicated. All genes have been annotated by using local annotation tool: UI spider. All these information was organized into local promoter database with HPD ID entry as an associative key.

2. Database Installation and interface implementation

There are six tables in the database, containing promoter sequences, transcription factor binding information and annotation as shown in Figure 1.

The database is now supported by IDEA (Iowa Database and Emerging Applications) lab, Department of Computer Science at the University of Iowa. IDEA lab is using mySQL as the database server. The database system is hosted on an Intel Pentium workstation with a Linux operating system. The client system can be a machine running on various kinds of operating systems, such as a Windows 2000 professional operating system.

GeneSpring, a microarray analysis tool from Silicon Genetics, has been used to extract co-expressed genes' clusters from microarray data. A microarray database constructed by CBCB was used to accommodate microarray data.

3. Important Code and their descriptions

A C++ Code has been written to convert MatInspector output into suitable format for database input. It removes all unrelated information and save transcription factor binding information; concatenates HPD ID to binding position, and uses it as the primary key for transcription factor binding tables.

Several Perl scripts have been written to load data into database created with Sybase as the database server:

loadPromoterSequence.pl: load Promoter Sequence to HPD database

loadGeneAnnotation.pl: load gene annotation to HPD database

load TranscriptionFactorinformation.pl: load transcription factor binding sites information to HPD database

Several Perl CGI scripts have been written for database interface implementation:

creat.html: display of homepage

access.cgi: for database login

hpdQuery.cgi: for query initiation

queryOur.cgi: for output query result

Challenges of this project

Data integration is one of the most challenging tasks to handle large-scale biological data. Historically, biological data was produced in relatively small scale until the initiation of human genome project. Moreover, biological data is produced in a rather distribute research system, except for that produced in private companies, such as Celera. For example, human genomic sequence data is contributed by hundred of labs in at least 18 countries (http://www.ornl.gov/hgmis/). The fact that biological data is produced by different incompatible resources and stored in various systems leaded to various formatted data with different naming systems.

Recent years, with the production of large amount of DNA sequence, protein sequence and gene expression data from different labs, integration of those data stored across different systems is the first key step for further data mining.

Since the data stored in promoter regulatory elements database are collected from rather distributed systems and each system has different data formatting and object naming, a data integration process is taken so that information from heterogeneous resources is linked and uniformed. For example, potential promoter sequences are obtained from HPD (Human Promoter Database, http://zlab.bu.edu/~mfrith/HPD.html) and stored in promoterSequence table. The sequences are named with the HPD accession number (HPDID), which does not fit into the general definition of a gene.

Therefore, the sequence is used to blast search against MGC cDNA sequences and results are stored in promoterAnnotation table. The linkage between the two is the promoter sequences. Since the sequence is a very large field, HPDID is used as the linkage between the two tables.

Data processing is another challenge for this project. First of all, there are multiple bioinformatic centers around world available for selecting a data resource for data mining. To select an appropriate data resource, which provides a comprehensive and accurate data set, is a tough issue. A full understanding of the data resources and a comparison among several candidate data resource are extremely necessary for accomplish this task. For instance, there are several centers providing human promoter information, such as EPD(Eukaryotic Promoter Database) and human promoter database in Cold Spring harbor. Each of them are providing different kinds but overlapping information for human promoter. We need to decide how to make full use of those resources to obtain all the essential information. Besides choosing appropriate data resources, the data obtained from different resources contains unrelated information. Each analyzing result was scanned and parsed before the data was deposited into database.

Future Plans

1. Increase the Entries for the database

Today's trend in the study of transcription mechanism is trying to associate microarray data to transcription regulatory elements. Our long-term goal has been set to establish a gene expression regulatory network. However, the promoter entries in current known promoter databases are far less than the number of genes deposited on a microarray slide. Although a lot of efforts have been taken to predict promoters based on human genome sequence, it has been well known that current promoter prediction algorithms predict much more false positive promoters than real promoters. Even a combination of multiple prediction algorithms couldn't solve this problem. Potential promoter regions predicted from full-length cDNA has been demonstrated as a good resource for further transcriptional regulation study.

Several promoter and transcription factor databases have been developed for various kinds of purposes. With longest history, Eukaryotic Promoter Database (EPD) contains the most strictly selected promoter sequences with only 276 entries for human promoter. All the entries of EPD are associated with supporting experimental data. It has been used as a good resource for the development of promoter prediction algorithm. Recently, a database containing 2,004 entries for human promoters has been

developed based on the full-length cDNA sequences (http://zlab.bu.edu/~mfrith/HPD.html). However, with the progress of full-length cDNA project, more and more data are available for correctly prediction of human promoters. Mammalian gene collection (MGC) from National Center for Biotechnology Information (NCBI) currently has 9,949 human gene entries (http://mgc.nci.nih.gov/). Another attractive resource is Database of Human Transcriptional Start Sites (DBTSS) in Japan, which provides 7,889 gene entries. Compared to MGC, DBTSS are believed to be better for promoter identification, since it may provide longer 5'-end sequences and correct TSS, which are essential to determine promoter region for a particular gene. Although both resources provide more entries, they require extra work to determine the transcription start sites before corresponding promoter sequences can be extracted from human genomic sequence.

2. Statistical Analysis

The characterization of those predicted potential promoter regions are highly desired for result verification. For each of the potential promoter region, the existence information of promoter elements will be obtained, such as TATA box, GC box, Inr etc. On the other hand, the critical characterization of a promoter is essential for its function. For instance, the methylation status of CpG Island has been known as a suppression mechanism for a number of genes involved in cancer development. Whether a potential promoter is localized on CpG island or not will also be determined.

3. Database Implementation

The refurbishment of user interface is highly desired. Currently, CREAT provides a basic query for obtaining a list of shared transcription factors. A more complicate interface and query process should be designed for further study. How to automatically return a smaller set from a larger set of input genes, which share a similar gene expression pattern, is our next step goal. For instance, when a list of genes are loaded from the microarry database and input into the promoter database as a query, different lists of genes are returned. All genes in a list have the same property that they contain same consensus binding sequence for a particular transcription factor. Moreover, the security and maintenance of the database should also be considered.

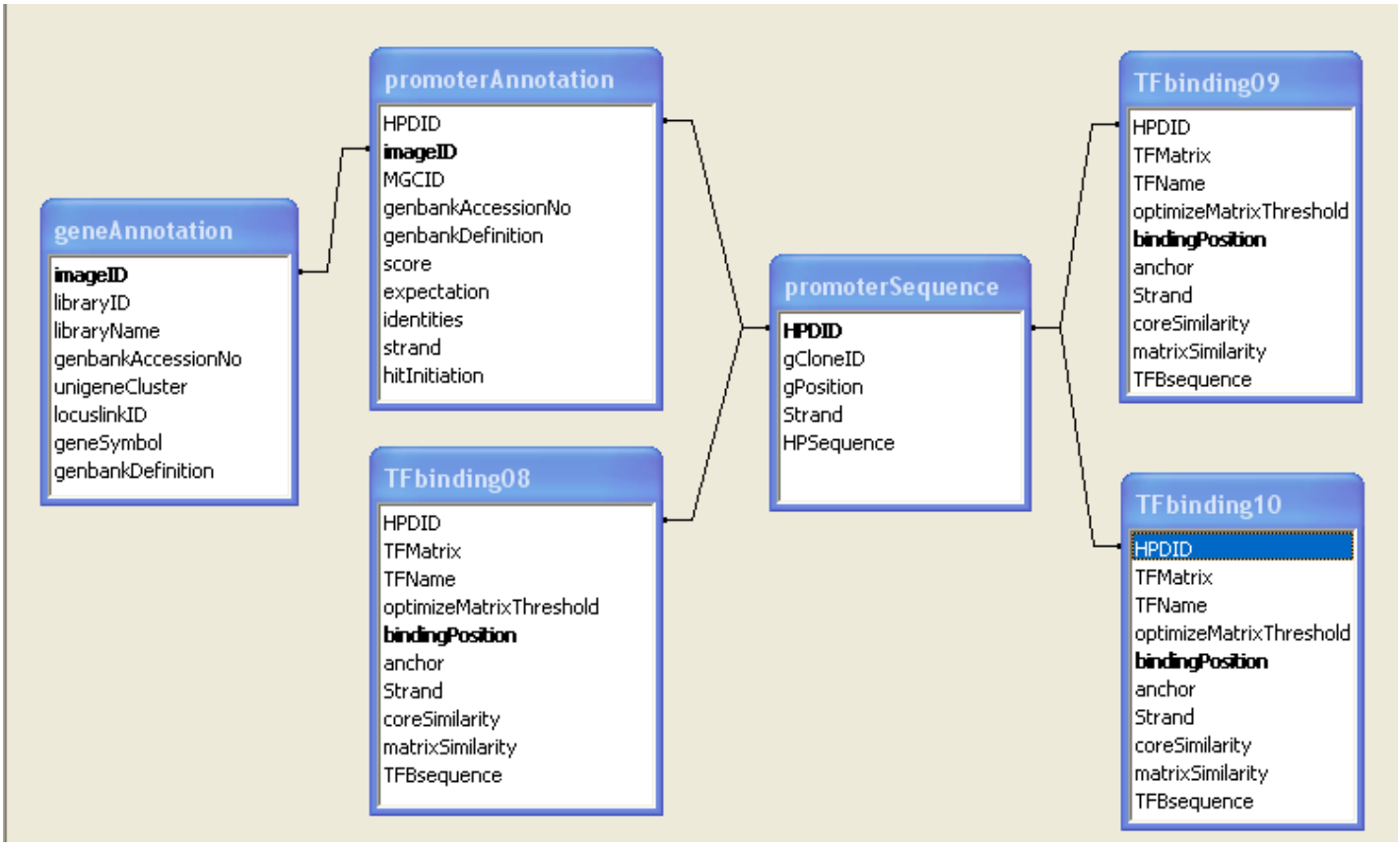

Figure 1. Schema for human potential promoter database.