International Journal of Engineering Inventions
ISSN: 2278-7461, www.ijeijournal.com
Volume 1, Issue 2 (September 2012) PP: 15-16

Feature Extraction from Gene Expression Data Files

Sayak Ganguli, Manoj Kumar Gupta, Sasti Gopal Das and Abhijit Datta
DBT Centre for Bioinformatics, Presidency University, Kolkata

Abstract—The Gene Expression Omnibus (GEO) maintained at the National Center for Biotechnology Information (NCBI) is an archive where microarray data is available freely to researchers. The database provides us with numerous data deposit options including web forms, spreadsheets, XML and Simple Omnibus Format in Text (SOFT). Along with the purpose of a repository a large collection of tools are available to help users effectively explore, and analyze gene expression patterns stored in GEO. The need for better analyses of gene expression data was explored using object oriented programming languages such as C and Practical extraction and reporting language (PERL). The test set that was used for this study were human transcription factors. The results obtained indicate that feature extraction was successful.

Keywords—Bioinformatics, Gene Expression Omnibus, NCBI, Object oriented programming.

I. INTRODUCTION

The GEO database (3) is one of the largest repositories of gene expression data on the web (1, 2) which has been generated using a large ensemble of high throughput measuring techniques. Such techniques include comparative genomic hybridization for analyzing gain or loss of genes and tiling arrays for the analyses and detection of transcribed regions or single-nucleotide polymorphism. Apart from these approaches data from ChIP-chip technology is also stored which provides information on the protein binding regions on nucleotide molecules. Data from serial analysis of gene expression (SAGE), massively parallel signature sequencing (MPSS), serial analysis of ribosomal sequence tags (SARST), and some peptide profiling techniques such as tandem mass spectrometry (MS/MS) are also accepted. Several tools such as GENE CHASER (5) exists which focus on the biological condition of expression of a particular data set, however, as microarray expression data comprise around 95% of the available data types in GEO it becomes necessary for providing a tool which shall enable the user to specifically analyze a particular data type from the expression data set (6).
The following are some existing tools which are used by the researchers for data analyses at GEO.

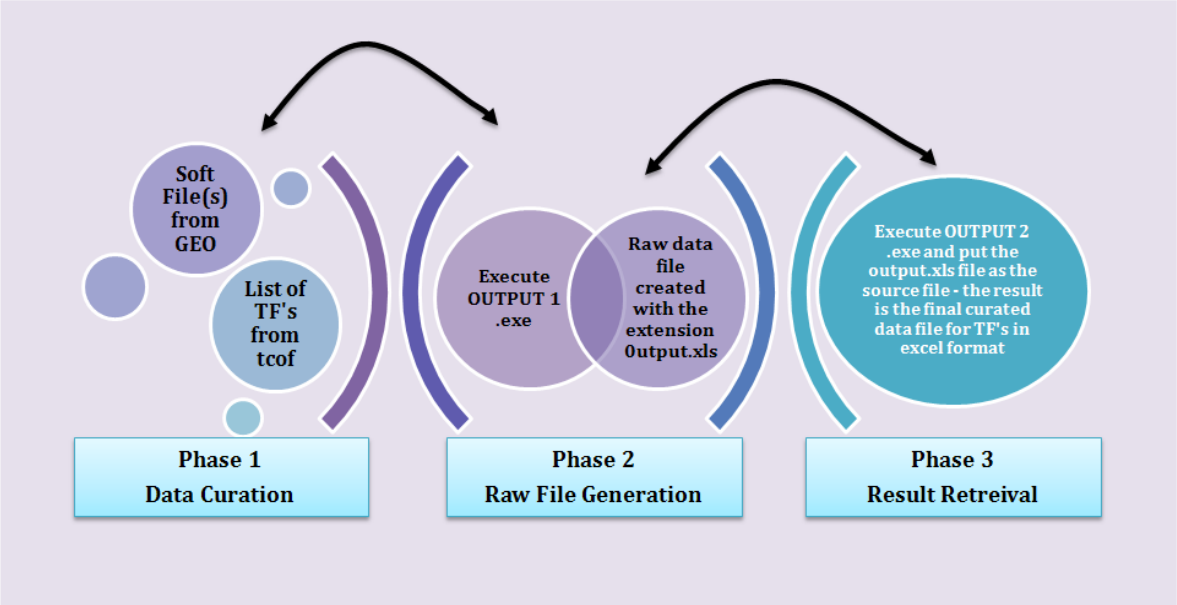
1. Profile neighbors retrieves gene connections that show a similar or reversed profile shape within a DataSet,
2. Sequence neighbors retrieves profiles which are related by nucleotide sequence similarity across all DataSets.
3. Homolog neighbors retrieve profiles of genes belonging to the same Homolo-Gene group.
4. Links menu enables the users to easily navigate from the GEO databases to associated records in other Entrez data domains including GenBank, PubMed, Gene, UniGene, OMIM, HomoloGene, Taxonomy, SAGEMap, and MapViewer.

However, analyses of specific gene sets from the files that are present in the GEO present a significantly labour intensive task. Here we present a feature extraction program written in C and PERL both belonging to the object oriented programming domain, a necessity when dealing with bioinformatics datasets as individual values are to be treated as objects and both these fundamental languages are easy to implement due to their platform independent property(4). The program allows the user to extract information regarding specific genes in a set of files. The program is currently tested on transcription factor expression data sets of human origin but would be extended to other datasets as well.

II. MATERIAL AND METHOD

This program is written in C and integrates PERL for feature extraction from the GEO data files.
Steps for Sorting Human Transcription Factors:

1. **Generating raw data**
2. **Generating transcription factors**
3. **Generating Gene Data File**
4. **Generating final transcription factors list**



III. CONCLUSIONS

With the increase in experimental techniques utilizing high throughput technologies, the need for automated data curation is paramount for successful time scale studies and for future implementation of robust analyses platforms. The feature extraction algorithm developed was tested on the most abundant class of genes that are present in general expression analyses – transcription factors. The results indicate that the algorithm can be extended in the development of an integrated webserver for the analyses and characterization of more such gene specific queries. Having sequence information together with expression information can help in the functional annotation and characterization of unknown genes or in finding novel roles for characterized genes. These data are also valuable to genome-wide studies, allowing biologists to review global gene expression in various cell types and states, to compare with orthologs in other species, and to search for repeated patterns of coregulated groups of transcripts that assist formulation of hypotheses on functional networks and pathways.

IV. ACKNOWLEDGEMENTS

The authors acknowledge the support of the Department of Biotechnology, Government of India BTBI scheme.

REFERENCES

1. Akerkar, R. A.; Lingras, P. An Intelligent Web: Theory and Practice, 1st edn. 2008, Johns and Bartlett, Boston.
2. Albert, R.; Jeong, H.; Barab´asi, A.-L. Diameter of the world-wide Web. Nature, 1999, 401, pp. 130–131.
3. Barrett T, Edgar Ron. Mining Microarray Data at NCBI's Gene Expression Omnibus (GEO) *Methods Mol Biol*. 2006; 338: 175–190.
4. Chakrabarti, S. Data mining for hypertext: A tutorial survey. SIGKDD explorations, 2000, 1(2), pp. 1–11.
5. Chen R., Mallelwar R., Thosar A., Venkatasubrahmanyam S., Butte A.J. (2008) GeneChaser: identifying all biological and clinical conditions in which genes of interest are differentially expressed, *BMC Bioinformatics* 9:548
6. Tarca AL, Romero R, Draghici S: Analysis of microarray experiments of gene expression profiling. Am J Obstet Gynecol, 2006 Aug;195(2):373-88