

Data Mining by Navigation – An Experience with Systems Biology

Amarnath Gupta¹, Michael Baitaluk¹, Animesh Ray², and Aditya Bagchi³

¹San Diego Supercomputer Center, Univ. of California San Diego, La Jolla, CA
92093, USA

{gupta, baitaluk}@sdsc.edu

²Keck Graduate Institute, 435 Watson Dr., Claremont, CA 91711, USA

Animesh.Ray@kgi.edu

³Indian Statistical Institute, Kolkata 700108, India

aditya@isical.ac.in

Abstract. This paper proposes a navigational method for mining by collecting evidences from diverse data sources. Since the representation method and even semantics of data elements differ widely from one data source to the other, consolidation of data under a single platform doesn't become cost effective. Instead, this paper has proposed a method of mining in steps where knowledge gathered in one step or from one data source is transferred to the next step or next data source exploiting a distributed environment. This incremental mining process ultimately helps in arriving at the desired result. The entire work has been done in the domain of systems biology. Indication has been given how this process can be followed in other application areas as well.

1 Introduction

In order to get insights of hypertension mechanisms, this paper ventures to discover the genes responsible for such hypertension. Traditionally the systems biologists depend on past experience, augmented by new laboratory experiments and statistical analysis. Results help in forming hypothesis substantiated by expert opinions. Though this approach enriches knowledge, it hardly utilizes past data obtained from different biological experiments. These past data are well documented in different databases and gene transcriptional/signaling networks, represented as graphs. This paper tries to make the knowledge discovery process sufficiently data driven avoiding dependence on hypotheses formed from experiences and expert opinions.

Since this type of knowledge discovery demands access to diverse data sources, it falls under distributed data mining paradigm. Moreover this study tries to associate related data items from different sources without deriving any rules. So the statistical processes applied here are different from the traditional *support* and *confidence* measures that use joint and conditional probabilities [7]. In addition, unlike considering only co-occurrence of items, combination of presence and absence of items may also be studied to discover additional interesting patterns. A new measure to this effect has already been developed [5]. Some applications demand collection of

data from different sources in order to find composite association of them. This notion of mining, popularly known as generalized association, tries to identify correlations among items considering both the absence and presence of them. A seminal work with this approach used chi-square test to measure significance of association. Authors demonstrated effectiveness of their method by testing it on census data and also by finding term dependence in a corpus of text documents [6]. Another recent work has proposed a method of compositional mining on multiple biological datasets. Authors have also proposed two new methods of mining; redescription and biclustering. However, here all the datasets are collected together to create a multirelational environment [9].

Though earlier research efforts explored the possibility of discovering association among items by extracting data from different sources, these efforts primarily handled same type of data representation and hence creation of a composite scenario was not a computationally challenging task. However, the application considered in this paper ventures to discover association among items with diverse representations distributed over different sites of a network. So this problem demands a new approach to mining different from the earlier efforts.

First, data representation, storage structure, access methods etc. are different in different data sources. So, porting data from different sites to one site and then bringing everything under a uniform representation may not be cost effective.

Secondly, semantic interpretation of even common items may be substantially different under different paradigms. For example, in a bioinformatics application, a gene may sometimes be viewed as a sequence and sometimes be viewed as a node in an interaction graph.

So, a mining effort is required which would choose to break the computational process to each data source and would transfer result of one to the next for further processing. In this navigational process a data source may be visited more than once depending on the application domain requirements. Thus the main motivation behind this paper is to develop a method of mining by navigation in a distributed environment.

While Section 2 of the paper covers the background knowledge for the data mining problem, Section 3 discusses about the diverse data sources utilized. Actual computational and navigational methodologies have been covered in Section 4. Section 5 discusses about the contribution made by the paper and concludes with possibility of other interesting applications.

2 Background Knowledge

Before going into the algorithmic detail, the following paragraph discusses about the problem domain for the appreciation of computer science community.

In general, Deoxyribonucleic Acid (DNA) is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms. DNA segments that carry this genetic information are called genes. DNA also contains the instructions or blueprints needed to construct other components of

cells like, proteins and RNA molecules. Ribonucleic acid (RNA) is transcribed from DNA by enzymes RNA polymerases. RNA is central to the synthesis of proteins. Here, a type of RNA called messenger RNA (mRNA) carries information from DNA and translates the information they carry into proteins. So in short, genes transcribe to mRNA in the process of protein production. Gene expression refers to the amount of mRNA produced. [8]. Two genes are considered to be co-expressed if, against a real life phenomenon, like a disease, both the genes are expressed simultaneously. Conversely, against a real life phenomenon, two genes are differentially expressed if while one gene is expressed other does not. Transcription of a gene is regulated (i.e. enabled and initiated) by a protein. While there can be other types of regulators, a regulator that helps in transcription is called a Transcription Factor (TF). The common expert opinion or hypothesis is - if two genes are co-expressed, they may have a common TF.

3 Data Sources

In this paper, the proposed mining effort uses three main data sources:

1. Gene Expression Omnibus (GEO), a well known repository that acts as curated, online resource for browsing, query and retrieval of gene expression data [4].
2. Promoter Analysis Pipeline (PAP) [2] is a Web-based workbench that provides the analysis of a set of co-expressed genes and the prediction of their transcriptional regulatory mechanisms.
 - PAP provides possible transcription factors that may regulate a set of co-expressed genes.
 - PAP can also predict other genes in the genome that are likely to be regulated by the same set of transcription factors.
 - It is also possible to get from PAP, the set of transcription factors that may regulate a particular gene.

All these information are statistical in nature and offered with a possible likelihood. The results are accepted when they cross a certain threshold of likelihood.

3. PathSys is a graph-based system for creating a combined database of biological pathways, gene regulatory networks and protein interaction maps. PathSys integrates over 14 curated and publicly contributed data sources along with Gene Ontology and is structured as an acyclic graph [1].

4 Computation and Navigation

Starting with an experiment to find the genes responsible for hypertension mechanism in mammalian species, this paper ventures to navigate over the different data sources mentioned in Section 4. Ultimately, a set of interacting genes with their transcription factors are discovered from the graphs of PathSys. This graph is

enriched by marking the connecting edges with their relative importance as obtained from computation with data from other sources. In the process, biologists acquire such knowledge about gene interactions that was not possible to get from laboratory experiments alone. The initial experiment was done on a hypertensive-hypotensive mice population to identify a set of differentially over-expressed genes apparently responsible for hypertension [3]. In the rest of the paper this would be referred as Friese experiment. Since the computational process has to navigate among diverse data sources it is necessary to justify the order of navigation so that the overall process of mining is well appreciated.

Genes identified by the Friese experiment, if placed on the PathSys graph directly as nodes, would give rise to a sub-graph as the molecular interaction graph that is responsible for hypertension in mammalian species. However, it does not reveal the importance of co-occurrence of all these genes. So it is necessary to identify which of these genes contribute significantly in the hypertension process when they co-express. This information is available in GEO database. So without pruning the gene set with co-expression data, directly accepting a sub-graph from PathSys may offer nodes and edges that may not have significant contribution to the study undertaken. Now since the co-expressed genes are likely to have common Transcription Factors and the genes that tend to co-express under hypertension have already been identified, the corresponding TFs can be obtained from the PAP workbench. Pruned set of genes as obtained from GEO database along with their TFs as obtained from PAP may now be placed on the PathSys graph. Experiment below would show that this process of identifying co-expressed genes and identification of sub-graph may be done more than once, ultimately to arrive at a final sub-graph.

Different steps of computation and navigation among data sources are detailed below:

Step 1. Correlation Phase:

Data Source: GEO Database.

Starting with the initial set of genes obtained from Friese experiment the GEO database is searched for experiments involving those genes, where in each such experiment at least one pair of co-expressed genes is identified from the initial set.

For each such pair of co-expressed genes X_i and Y_i , the expression values for all experiments are taken where they co-occur and Pearson correlation coefficient is computed as:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_X S_Y}$$

where, S_X and S_Y are the standard deviation of X and Y respectively (with 25 genes in the initial set, GEO database offered data about 630 experiments, causing a maximum data volume = $630 * ({}^{25}C_2)$).

Out of the correlation coefficients computed for all pair of genes in the initial set only those crossing a predefined threshold are taken for further study (considering a threshold of 0.7 for Pearson correlation coefficient, similar to Minimum-Support, 11 genes, as listed in Table 1, are retained for further processing).

Table1. Final set of co-expressed genes obtained from GEO

Gene Name
ACADVL (Acyl CoA dehydrogenase, very long chain)
CHGA (Chromogranin A)
CHGB (Chromogranin B)
CYP2E1 (Cytochrome P450 2E1)
EDN3 (Endothelin 3)
IGFBP4 (Insulin like growth factor binding protein-4)
IGFBP6 (Insulin like growth factor binding protein-6)
PNMT (Phenyl ethanolamine N-methyl transferase)
SCG2 (Chromogranin C)
SLC6A4 (Serotonin transporter)
TH (Tyrosine hydroxylase)

Step 2. Extraction of Transcription Factors (First Navigation):

Data Source: PAP Workbench.

Pairs of genes found to be highly co-expressed in Step 1 (correlation phase), are analyzed in the PAP workbench to find possible common Transcription Factors (TF). Considering human, rat and mouse data, only gene pairs present in all the three mammals are taken.

Since all TFs are identified with corresponding likelihood value, only TFs crossing a predefined threshold are listed in descending order and only a predefined number of members from the top of the list are taken for the subsequent step of analysis (the specific study considered 95% likelihood and only top 25 TFs are taken. These TFs are the regulator proteins for the genes obtained in Step 1).

Step 3. Identification of Protein-Protein Interaction (Second Navigation):

Data Source: PathSys Molecular Interaction Maps.

Since PathSys offers protein interaction maps, nodes corresponding to the original set of genes obtained from Step 1 and the nodes corresponding to the TFs obtained in Step 2 are identified along with their interconnecting edges. This study offers a sub-graph with the interconnected proteins and genes that is responsible for hypertension in mammalian species.

To extend the study further, one-neighborhood of these nodes is now considered on the PathSys graph thereby identifying even the regulator nodes connecting the TFs.

Step 4. Second Correlation phase (Third Navigation):

Data Source: GEO Database.

In order to ascertain the importance of the regulators in the original study of hypertension, the TF nodes as well as their regulators obtained from one-neighborhood are again taken for study of co-expression in GEO database. Repeating the process described in Step 1, the minimized set of highly co-expressed TFs and their regulators are identified.

Step 5. Graph Pruning (Fourth Navigation):

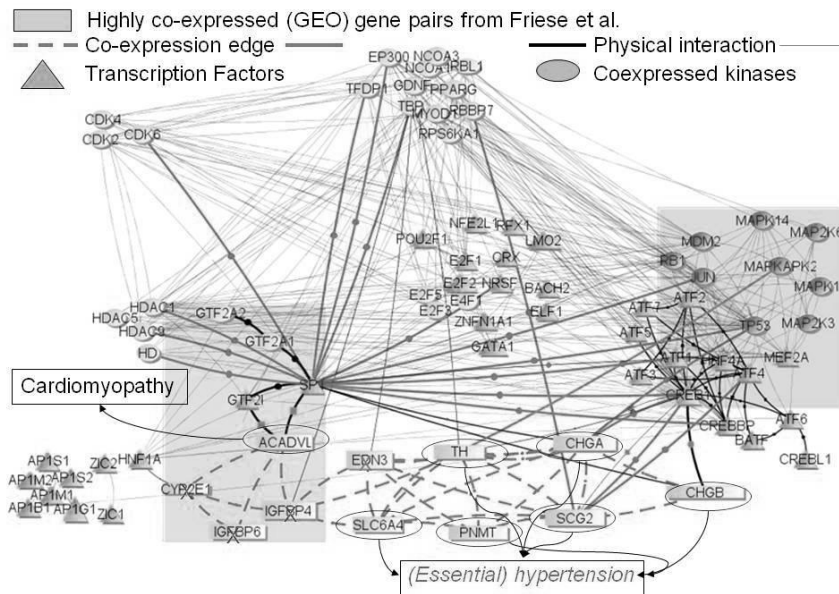
Data Source: PathSys Molecular Interaction Maps.

Sub-graph identified in Step 3 after considering one-neighborhood is now pruned with the result obtained from Step 4 and the extra nodes and edges obtained from one-neighborhood are deleted. Each co-expression edge (gene to gene and TF to TF) is now labeled with its co-expression value. This pruned graph with all its nodes as genes, proteins etc. and the interconnecting edges is functionally related to hypertension in mammalian species.

Step 6. Further Graph Pruning: (An additional study)

Hypothesis: If the network, as obtained from Step 5 is functionally significant, then co-expressed and interacting genes and their regulator proteins should not only show functional relation to hypertension but also to other related diseases like cardiovascular disorders and diabetics.

Accepting the hypothesis obtained from the domain experts, the graph pattern obtained from Step 5 is further pruned by using a fourth data source, the Online Mendelian Inheritance in Man or OMIM database. This database is a catalog of human genes and genetic disorders. Besides the original set of 11 genes, corresponding entries for other nodes are searched in OMIM database to ascertain whether they contribute to any or all of the diseases under scrutiny. For example, HNF1A a transcription factor for CYP2E1 gene is also a TF for other genes responsible for liver-specific disorders, and HDAC5 a regulator of SP1, a member of original set of TFs, is associated with cardiac problems.



CYP11B1 & B2, CYP17 (corticosteroid metabolism enzymes)

Fig. 1. PathSys Interaction Graph

Collecting these extra evidences from the fourth dataset used in this study, the set of selected genes and TFs is further pruned to get the final gene-protein interaction

graph that contributes to the study of hypertension in mammalian species. Figure 1 shows the final pruned graph as obtained from PathSys.

5 Conclusions

Studying a problem in systems biology, this paper has shown how composite mining strategy can be devised when related information are distributed in different data sources at different sites with substantial variation in data representation, data structures, storage strategies etc. Besides this navigational process, this paper has also proposed a new graph mining technique. Deviating from well known graph mining strategies, this process wants to identify the most interesting sub-graph from a large graph and augments it by evidences obtained from other problem related data sources. Interestingly, it may be possible to apply this graph pruning strategy in many other application areas like, traffic management, tour planning, social network etc. Authors are now trying to develop a generalized methodology for this type of graph mining.

References

1. Baitaluk, M., Qian, X., Godbole, S., Raval, A., Ray, A., Gupta, A.: PathSys: Integrating Molecular Interaction Graphs for Systems Biology, *BMC Bioinformatics* (2006) 7:55 doi:10.1186/1471-2105-7-55, <<http://www.biomedcentral.com/1471-2105/7/55>>
2. Chang, L. W., Fontaine, B. R., Stormo, G. D., Nagarajan, R.: PAP: a comprehensive workbench for mammalian transcriptional regulatory sequence analysis, *Nucleic Acids Research* (2007) 238-244 <<http://bioinformatics.wustl.edu/webTools/PromoterAnalysis.do>>
3. Friese, R.S., Mahboubi, P., Mahapatra, N. R., Mahata, S. K., Schork, N. J., Schmid-Schönbein, G. W., O'Connor, D. T.: Common genetic mechanisms of blood pressure elevation in two independent rodent models of human essential hypertension, *Am J Hypertension*, vol.18(5 Pt 1) (2005) 633-652
4. Gene Expression Omnibus (GEO) : <http://www.ncbi.nlm.nih.gov/geo/>
5. Pal, S., and Bagchi, A.: Association against Dissociation: some pragmatic considerations for Frequent Itemset generation under Fixed and Variable Thresholds, *ACM SIGKDD Explorations*, vol.7, issue 2 (2005) 151-159
6. Silverstein, C., Motwani, R., Brin, S.: Beyond Market Baskets: Generalizing Association Rules to Correlations, *Proceedings of the ACM SIGMOD International Conference on Management of Data* (1997) 265-276
7. Srikant, R.: Fast Algorithms for Mining Association Rules and Sequential Patterns, Ph.D. Thesis, University of Wisconsin, Madison, USA (1996)
8. Wikipedia, the free online encyclopedia : <http://www.wikipedia.org/>
9. Ying, J., Murali, T. M., and Ramakrishnan, N.: Compositional Mining of Multirelational Biological Datasets, *ACM TKDD*, vol.2, no.1 (2008) 1-35.