Formal Concept Analysis Applied to Transcriptomic Data

Mehwish Alam[2,3], Adrien Coulet[2,3], Amedeo Napoli[1,2], and Malika Smaïl-Tabbone[2,3]

[1] CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
[2] Inria, Villers-lès-Nancy, F-54600, France
[3] Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
{mehwish.alam,adrien.coulet,amedeo.napoli,malika.smail@inria.fr}

Abstract. Identifying functions shared by genes responsible for cancer is a challenging task. This paper describes the preparation work for applying Formal Concept Analysis (FCA) to complex biological data. We present here a preliminary experiment using these data on a core context with the addition of domain knowledge. The resulting concept lattices are explored and some interesting concepts are discussed. Our study shows how FCA can help the domain experts in the exploration of complex data.

Keywords: Formal Concept Analysis, Knowledge Discovery, Transcriptomic Data.

1 Introduction

Over past few years, large volumes of transcriptomic data were produced but their analysis remains a challenging task because of the complexity of the biological background. Some earlier studies aimed at retrieving sets of genes sharing the same transcriptional behavior with the help of Formal Concept Analysis [1, 2]. Further studies analyze gene expression data by using gene annotations to determine whether a set of differentially expressed genes is enriched with biological attributes [3, 4]. Several efforts have been made for integrating heterogeneous data [5]. For example, at the Broad Institute, biological data were recently gathered from multiple resources to get thousands of predefined genesets stored in the Molecular Signature DataBase (MSigDB) [6]. A predefined geneset is a set of genes known to have a specific property such as their position on the genome, their involvement in a molecular pathway etc.

This paper focuses on the preparation of biological data to data mining guided by domain knowledge. The objective is to apply knowledge discovery techniques for analyzing a list of differentially expressed genes and identifying functions or pathways shared by these genes assumed to be responsible for cancer. Section 2 explains the proposed approach for FCA-based analysis of biological data. Section 3 focuses on the conducted experiment. Section 4 discusses the results. Section 5 concludes the paper.

Local Proceedings in ISBN 978–84–695–5252–0,
Universidad de Málaga (Dept. Matemática Aplicada), Spain.

2 The Proposed Framework

We rely on the standard definition of FCA fully described in [7] and adapt it according to the current problem. Let G be the set of genes $\{g_1, g_2, g_3, ..., g_n\}$, and M be a set of attributes of MSigDB for describing genes. M will be considered as a partition of three points of view, $M = M_1 \cup M_2 \cup M_3$, with $M_i \cap M_j = \emptyset$ whenever $i \neq j$.

The first set of attributes M_1 refers to four types of attributes, "Location", "Pathway", "Transcription Factors" and "GO Terms" (see Table 1). For our convenience we have named MSigDB categories as types of attributes and used only C_1, C_2, C_3 and C_5. The category C_4 was not used as it keeps information on sets of genes related to a certain kind of cancer, which is not useful for the current problem. Thus we have a first context $K_1 = (G, M_1, I_1)$ where I_1 denotes the relation stating that gene g_i has an attribute m_j in M_1.

Types of Attributes	**Description**	**Data Provenance**
C1: Positional Gene Sets	Location of the gene on the chromosome.	Broad Institute
C2: Curated Gene Sets	Pathway	KEGG, REACTOME, BIOCARTA
C3: Motif Gene Sets	Transcription Factors	Broad Institute
C4: Computational Gene Sets	Cancer Modules	Broad Institute
C5: Gene Ontology (GO) Gene Sets	Biological Process, Cellular Components, Molecular Functions	AmiGO

Table 1. Types of attributes from MSigDB

The second set of attributes M_2 is related to the so-called "categories" where a category makes reference to a set of attributes with the "Pathway" type. For example, "Cell Growth and Death" is an example of category (see Figure 1). The categories in M_2 determine a second context, $K_2 = (G, M_2, I_2)$ where M_2 is the set of categories and I_2 denotes the relation between a gene and a category. It can be noticed that the categories are only related to the "Pathway" type and that they can be considered as domain knowledge.

Moreover, the third set of attributes, namely M_3, refers to the so-called "upper categories", which are defined as groupings of categories. Actually, we have for the type "Pathway" a hierarchy of categories with two levels, categories and upper categories (see Figure 1). The upper categories in M_3 define a third context $K_3 = (G, M_3, I_3)$ where M_3 is the set of upper categories and I_3 denotes the relation between gene g_i and an upper level category m_j. Upper categories are also related to the "Pathway" type and as categories, they can be considered as domain knowledge too.

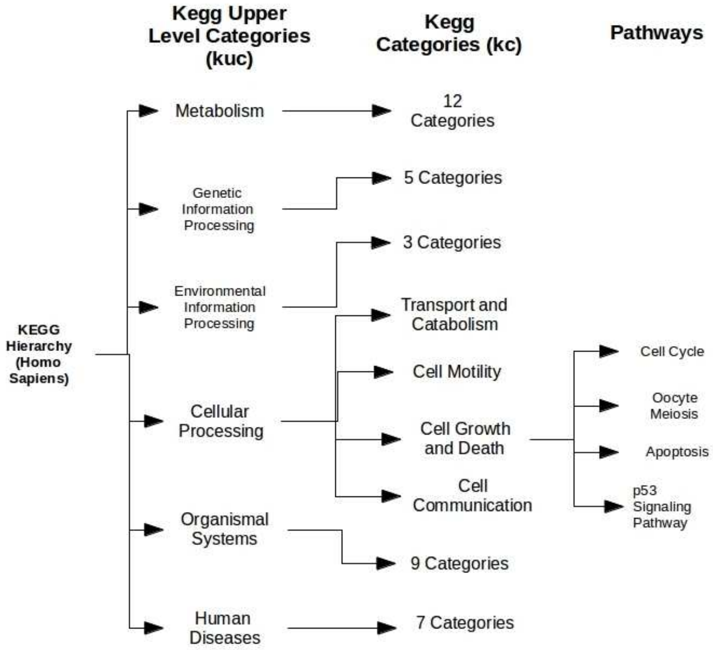

Fig. 1. Categories and upper categories in KEGG.

Now we consider the apposition of the three contexts K_1, K_2 and K_3, which yields the final context $K = (G, K_1 \cup K_2 \cup K_3, I_1 \cup I_2 \cup I_3)$. For example the context in Table 2 shows five genes described by attributes of M_1, M_2 and M_3.

3 Using FCA for Analyzing Genes

The framework described above was applied on three published sets of genes corresponding to Cancer Modules defined in [8]. Our test data are composed of three lists of genes corresponding to the so-called "Cancer Module 1" (Ovary Genes), "Cancer Module 2" (Dorsal Root Ganglia Genes), and "Cancer Module 5" (Lung Genes). For example, "PSPHL" is one gene with "Pathway" attribute as "PPAR Signaling" which belongs to category "kc:Endocrine System" and upper category "kuc:Organismal System". Considering the three lists of genes given by "Cancer Module 1", "Cancer Module 2" and "Cancer Module 5", we built three different contexts having the same form as the context in Table 2). Then we obtained three associated concept lattices with the help of the Coron Plate-form (`http://coron.loria.fr`). The concept lattice for Table 2 is given in Figure 2. The global characteristics of the three concept lattices are given in Table 3.

The exploration of a given concept lattice is carried out following the "Iceberg metaphor", i.e., the lattice is explored level by level according to the support of

Genes	ATP Binding (GO Term)	Serotonin Receptors (Pathway)	PPAR Signaling (Pathway)	V$POU3F2_02 (Transcription Factor)	Cellular Component Assembly(GO Term)	chr5q12 (Location)	kc:Endocrine System	kuc:Organismal Systems
BTB03	×				×	×		
PSPHL			×	×			×	×
CCT6A		×				×		
QNGPT1	×	×			×	×		
MYC	×			×				

Table 2. A toy example of formal context including domain knowledge.

each concept, where the the support of a concept is the cardinality of the extent. In addition, we also used stability for extracting interesting frequent and stable concepts [9].

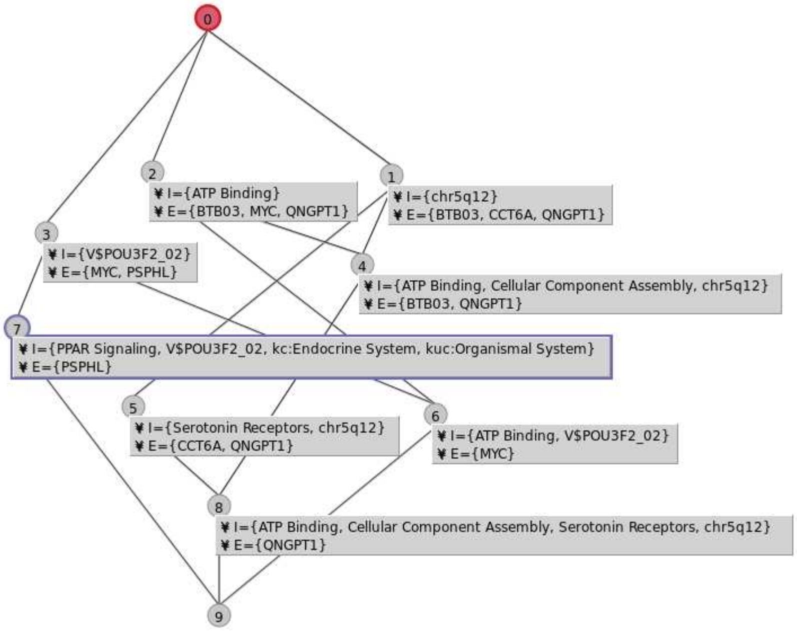
0
2
¥ I={ATP Binding}
¥ E={BTB03, MYC, QNGPT1}
1
¥ I={chr5q12}
¥ E={BTB03, CCT6A, QNGPT1}
3
¥ I={V$POU3F2_02}
¥ E={MYC, PSPHL}
4
¥ I={ATP Binding, Cellular Component Assembly, chr5q12}
¥ E={BTB03, QNGPT1}
7
¥ I={PPAR Signaling, V$POU3F2_02, kc:Endocrine System, kuc:Organismal System}
¥ E={PSPHL}
5
¥ I={Serotonin Receptors, chr5q12}
¥ E={CCT6A, QNGPT1}
6
¥ I={ATP Binding, V$POU3F2_02}
¥ E={MYC}
8
¥ I={ATP Binding, Cellular Component Assembly, Serotonin Receptors, chr5q12}
¥ E={QNGPT1}
9

Fig. 2. The concept lattice corresponding to table 2.

Data Sets	No. of Genes	No. of Attributes	No. of Concepts	Levels
Module 1	361	3496	9,588	12
Module 2	378	3496	6,508	11
Module 5	419	3496	5,004	12

Table 3. Concept lattice statistics for the cancer modules with domain knowledge.

4 Results

In this study, biologists are interested in links between the input genes in terms of pathways in which they participate, relationships between genes and their positions etc. We obtained concepts with shared transcription factors, pathways, locations of genes and GO terms. After the selection of concepts with a high support (≥ 10), we observed that there were some concepts with pathways either related to cell proliferation or apoptosis (expert interpretation). The addition of domain knowledge gives an opportunity to obtain the pathway categories shared by larger sets of genes (as categories and upper categories are there for maximizing the grouping of objects, see below).

Table 4 shows the top-ranked concepts found in each module. For example, in Table 4, we have the concept C_{4938}:*(KEGG Cytokine Cytokine Receptor Interaction, kc:Signaling Molecules and Interaction, kuc:Environmental Information Processing)* and the concept C_{4995}:*(kc:Signaling Molecules and Interaction, kuc:Environmental Information Processing)*. These two concepts are such as $C_{4938} \leq C_{4995}$, meaning that C_{4995} has greater support than C_{4938}. Moreover, we observed that the introduction of categories and upper categories in the global context allows us to consider concepts that otherwise would not be frequent. Actually, the role of categories and upper level categories is to facilitate the observation of sets of related genes.

This is a general way of obtaining larger sets of objects to interpret. When available, one can introduce a hierarchy of attributes –this is domain knowledge– and then insert the levels of each attribute in this hierarchy as a new attribute in the context. As a result, some classes of objects, that could not emerge before, will appear based on these hierarchical indications. Given the test data sets, the preliminary results obtained here constitute an interesting and positive control, and confirm that FCA-based analysis offers an efficient and practical procedure to explore complex and large sets of genes.

5 Conclusion

The preliminary study presented here shows how FCA can be applied to complex biological data and can give flexibility in using various types of attributes for analyzing a list of genes. In addition, domain knowledge can be introduced and guide the analysis.

Dataset	**Concept ID**	**Intents**	**Absolute Support**	**Stability**
Module 1	**9585**	GGGAGGRR _V$MAZ_Q6	51	0.99
	9571	GO Membrane Part	27	0.99
	9566	kc:Immune System, kuc:Organismal Systems	25	0.99
	9402	chr19q13	10	0.99
	9078	KEGG MAPK Signaling Pathway, kc:Signal Transduction, kuc:Environmental Information Processing	12	0.87
Module 2	**6502**	GGGAGGRR _V$MAZ_Q6	44	0.99
	6501	AACTTT _UNKNOWN	38	0.99
	6496	kc:Immune System, kuc:Organismal Systems	15	0.99
	6388	chr6p21	10	0.97
	6335	KEGG MAPK Signaling Pathway, kc:Signal Transduction, kuc:Environmental Information Processing	11	0.89
Module 5	**5002**	kuc:Cellular Processes	48	0.99
	5000	GGGAGGRR _V$MAZ_Q6	44	0.99
	4995	kc:Signaling Molecules and Interaction, kuc:Environmental Information Processing	26	0.99
	4933	chr19q13	11	0.99
	4985	kc:Immune System, kuc:Organismal Systems	11	0.99
	4938	KEGG Cytokine Cytokine Receptor Interaction, kc:Signaling Molecules and Interaction, kuc:Environmental Information Processing	11	0.87

Table 4. Top ranked concepts for each cancer module

As for future work, we plan to take into account relationships between genes and between terms (Gene Ontology relationships) and use the framework of relational concept analysis.

References

1. Kaytoue-Uberall, M., Duplessis, S., Kuznetsov, S.O., Napoli, A.: Two FCA-Based Methods for Mining Gene Expression Data. In Ferré, S., Rudolph, S., eds.: ICFCA. Volume 5548 of Lecture Notes in Computer Science., Springer (2009) 251–266
2. Rioult, F., Boulicaut, J.F., Crémilleux, B., Besson, J.: Using Transposition for Pattern Discovery from Microarray Data. In: DMKD. (2003) 73–79
3. Berriz, G.F., King, O.D., Bryant, B., Sander, C., Roth, F.P.: Characterizing gene sets with FuncAssociate. Bioinfo. **19**(18) (2003) 2502–2504
4. Doniger, S., Salomonis, N., Dahlquist, K., Vranizan, K., Lawlor, S., Conklin, B.: MAPPFinder: using Gene Ontology and GenMAPP to Create a Global Gene-expression Profile from Microarray Data. Genome Biology **4**(1) (2003) R7
5. Galperin, M.Y., Fernández-Suarez, X.M.: The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. Nucleic Acids Research **40** (2012) 1–8
6. Liberzon, A.: Molecular Signatures Database (MSigDB) 3.0. Bioinfo. **27**(12) (2011) 1739–1740
7. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin/Heidelberg (1999)
8. Segal, E., Friedman, N., Koller, D., Regev, A.: A Module Map Showing Conditional Activity of Expression Modules in Cancer. Nat.Genet. **36** (2004) 1090–8
9. Kuznetsov, S.O.: On stability of a Formal Concept. Ann. Math. Artif. Intell. **49**(1-4) (2007) 101–115