Formal Concept Analysis Applied to Transcriptomic Data

Mehwish Alam[2,3], Adrien Coulet[2,3], Amedeo Napoli[1,2], 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 or pathways shared by genes responsible for cancer is still a challenging task. This paper describes the preparation work for applying Formal Concept Analysis (FCA) to biological data. After gene transcription experiments, we integrate various annotations of selected genes in a database along with relevant domain knowledge. The database subsequently allows to build formal contexts in a flexible way. We present here a preliminary experiment using these data on a core context with the addition of domain knowledge by context apposition. The resulting concept lattices are pruned and we discuss some interesting concepts. Our study shows how data integration and FCA can help the domain expert in the exploration of complex data.

Keywords: Formal Concept Analysis, Knowledge Discovery, Data Integration, 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. In the field of transcriptomics, biologists analyze routinely the transcription or expression of genes in various situations (e.g., in tumor samples versus non-tumor samples).

Some earlier studies aimed at retrieving sets of genes sharing the same transcriptionl behaviour with the help of Formal Concept Analysis (see, e.g., [7, 10, 11]). Further studies analyze gene expression data by using gene annotations to determine whether a set of differentially expressed genes is enriched with biological attributes [1, 2, 13]. Many useful resources are available online and several efforts have been made for integrating heterogeneous data [5, 8]. A recent example is of the Broad Institute where biological data were gathered from multiple resources to get thousands of predefined gene sets stored in the Molecular Signature DataBase, MSigDB [4]. A predefined gene set is a set of genes known to have a specific property such as their position on the genome, their involvement in a

biological process (or a molecular pathway) etc. Subsequently, given an experimental gene list as input the GSEA (Gene Set Enrichment Analysis) program is used to asses whether each predefined gene set (in the MSigDB database) is significantly present in the input list by computing an enrichment score [3].

In this paper, we are interested in applying knowledge discovery techniques for analyzing a differentially expressed gene set and identifying functions or pathways shared by these genes assumed to be responsible for cancer. Knowledge discovery aims at extracting relevant and useful knowledge patterns from a large amount of data. It is an interactive and iterative process involving a human (analyst or domain expert) and data sources. We show how various gene annotations and domain knowledge are integrated in a database which is then queried for building in a flexible way formal contexts. We present here a preliminary experiments using these data. It was performed on a core context with the addition of domain knowledge (by context apposition). The considered domain knowledge are the hierarchical relationships between molecular pathways. Pruning the obtained lattices allows us to retrieve interesting concepts which we discuss. The results obtained from both experiments are also compared.

The plan of the paper is as follows: Section 2 introduces Formal Concept Analysis, Section 3 explains the data resources which are integrated, Section 4 focuses on the application of FCA, Section 5 discusses the results and Section 6 concludes the paper and presents future Work.

2 Formal Concept Analysis

We introduce here the basics of Formal Concept Analysis that are needed to understand what follows. Let G and M be the set of objects and set of attributes respectively and I be the relation between the objects and the attributes $I \subseteq G \times M$, where $g \in G$, $m \in M$, gIm is true iff the object g has the attribute m. The triple $K = (G, M, I)$ is called a formal context. If $A \subseteq G$, $B \subseteq M$ are arbitrary subsets, then a Galois connection denoted by $'$ is given by:

$$A' := \{m \in M \mid gIm \; \forall \; g \in A\} \tag{1}$$

$$B' := \{g \in G \mid g \; I \; m \; \forall \; m \in B\} \tag{2}$$

FCA framework is fully described in [6]. FCA helps in defining concepts which are composed of a maximal set of objects sharing a maximal set of attributes. However, given an input context, the resulting concept lattice can be very large leading to computational and interpretation problems. In order to have reduced and meaningful concepts, one can select concepts whose support is greater than a certain threshold, i.e., the iceberg lattice. For a concept (A,B), the support is the cardinality of the extent A. An alternative is to use the notion of stability that was proposed in [9, 12]. The stability index measures how much the concept intent depends on particular objects of the extent.

3 Complex Biological Data Integration

In this section, we introduce and describe the biological data on which we are working.

3.1 Molecular Signature Database (MSigDB)

Molecular Signature Database (MSigDB) is an up-to-date database which contains data from several resources such as KEGG, BIOCARTA, REACTOME, and Amigo [4]. It is a collection of 6769 predefined gene sets. A predefined gene set is a set of genes having a specific property such as their position on the genome (e.g., the genes at position chr5q12, i.e., band 12 on arm q of chromosome 5), their involvment in a biological process or a molecular pathway (e.g., the genes which are involved in the KEGG APOPTOSIS pathway)... A pathway is a series of actions among molecules in a cell that leads to a certain change in a cell. KEGG is a database storing hundreds of known pathways[4]. Besides, the MSigDB gene sets are grouped into five categories (Table 1). For instance, all the gene sets which are defined on the basis of gene position belong to the category C1. The category C5 groups the gene sets defined on Gene Ontology (GO) terms annotating the genes (with respect to their molecular function or their housing cellular component).

For our study, we used MSigDB Version 3.0. One entry, shown below in XML format, describes the gene set corresponding to the GO term 'RNA Polymerase II Transcription Factor Activity Enhancer Binding' (all the attribute names are underlined). The *Members* attribute contains the list of gene symbols belonging to the gene set. MSigDB was chosen as the main source for describing genes because it gathers up-to-date informations about many aspects of human genes.

<GENESET Standard Name ="RNA Polymerase II Transcription Factor Activity Enhancer Binding" Systematic Name = "M900" Historical Names ="" Organism ="Homo sapiens" Geneset Listing URL ="" Chip = "Human Gene Symbol" Category Code ="c5" Sub Category Code ="MF" Contributor ="Gene Ontology" Contributor Org ="GO" Description Brief ="Genes annotated by the GO term GO:0003705. Functions to initiate or regulate RNA polymerase II transcription by binding an enhancer region of DNA." Description Full ="" Members =" MYOD1, TFAP4, EPAS1, RELA, MYF5, MYEF2, NFIX, PURA, HIF1A" Members Symbolized = "MYOD1, TFAP4, EPAS1, RELA, MYF5, MYEF2, NFIX, PURA, HIF1A" Members EZID =" 7023, 2034, 5970, 3091" Members Mapping = " MYOD1, 4654-TFAP4, TFAP4, 7023-EPAS1, EPAS1, 2034-RELA, RELA, 5970-MYF5, MYF5, 4617-MYEF2, MYEF2, 50804-NFIX, NFIX, 4784-PURA, PURA, 5813-HIF1A" Status ="public" > </GENESET>

3.2 Domain Knowledge

Besides the gene annotations included in MSigDB, many types of domain knowledge are interesting to use when analyzing genes. The first type of such do-

[4] http://www.genome.jp/kegg/pathway.html

Table 1. Categories of MSigDB Gene Sets

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

main knowledge are the hierarchical relationships between GO terms or between KEGG pathways. Indeed, the KEGG hierarchy for human groups the KEGG pathways into 40 categories and 6 upper level categories. Figure 1 illustrates the KEGG hierarchy detailing on one upper-level category and one category.

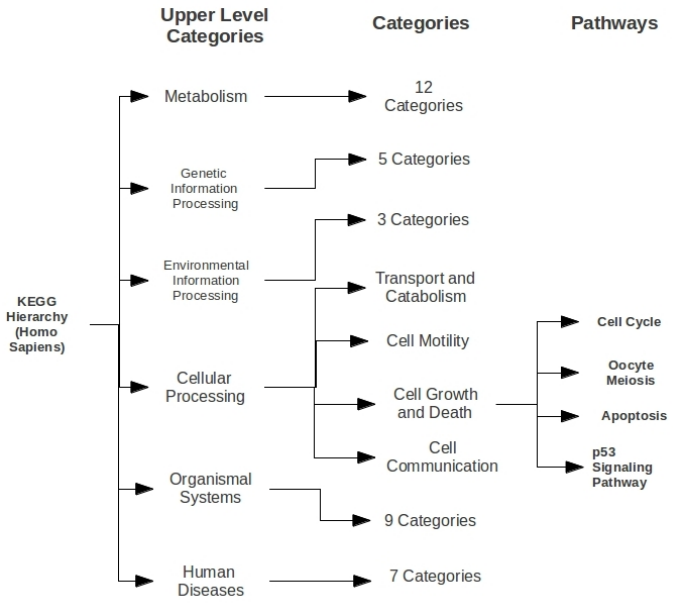

Fig. 1. Hierarchical Relationship in KEGG

In our study we have genes described by pathways involving them which may in turn be present in some category of pathways. For example, if a gene is involved in a pathway apoptosis it will also be in the category 'Cell Growth and Death'. In order to facilitate the knowledge discovery, it is important to

identify the relevant data sources, organize, and integrate the data at one single database. In our case, the relevant primary data sources are MSigDB, KEGG PATHWAYS database, and AmiGO database.

4 From Data to Knowledge

Once the data are integrated in our database the next step is to build formal contexts for applying FCA. Our experiment focuses on applying FCA to a core context describing genes by MSigDB-based attributes and shows its extension based on the addition of domain knowledge.

4.1 Test Data Sets

The experiments described here are based on three published sets of genes corresponding to Cancer Modules defined in [14]. The authors compiled gene sets from various resources and a large collection of micro-array data related to cancers. These modules correspond to gene sets whose expression significantly change in a variety of cancer conditions (they are also defined as MSigDB gene sets in the C4 category). Our test data are composed of three lists of genes corresponding to the Cancer Modules 1 (Ovary Genes), 2 (Dorsal Root Ganglia Genes), and 5 (Lung Genes).

4.2 Using FCA for Analyzing Genes

We apply FCA for analyzing a context describing genes of each Cancer Module with MSigDB-based attributes. Table 2 shows five genes (involved in Cancer Module 1) as a set of objects described by attributes which are the memberships to gene sets from MSigDB. For example, CCT6A is in the set of genes (gene set) whose standard_name is *Reactome Serotonin Receptors*. Interestingly, by querying our integrated database the analyst is able to select the predefined gene sets to include in the formal context.

In order to extend the analysis of a list of genes, we need to take into account the domain knowledge. Hence, the same experiment was conducted with the addition of the KEGG hierarchy knowledge to the core contexts resulting in three extended contexts. All KEGG categories and upper-level categories were added as a set of attributes. If a gene is member of a KEGG pathway which in turn belongs to a category and an upper level category then a cross '$\times$' is added in the corresponding cells in the extended context.

Table 2 shows five genes (from Cancer Module 1) with the addition of one KEGG category (kc) and one KEGG upper level category (kuc). In the given example *CCT6A* is involved in pathway *KEGG PPAR Signaling Pathway* which belongs to the category *kc:Endocrine System* and upper level category *kuc:Organismal Systems*. The lattices were generated and the statistics for each Cancer Module are given in Table 3. The concepts were filtered and ranked based on same criteria as in the first experiment.

Table 2. A Toy Example of Formal Context with Domain Knowledge

Genes	TTTGCAC,MIR-19A,MIR-19B	Reactome Serotonin Receptors	KEGG PPAR Signaling Pathway	V$POU3F2_02	GO Cellular Component Assembly	chr5q12	kc:Endocrine System	kuc:Organismal Systems
BTB03	×				×	×		
PSPHL			×	×			×	×
CCT6A		×				×		
QNGPT1	×	×			×	×		
MYC	×			×				

Table 3. Concept Lattice Statistics for the Cancer Modules with Domain Knowledge

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

5 Results

In this study, biologists are interested in links between the input genes in terms of pathways in which they participate, relationship between genes and microRNAs etc. We obtained concepts with shared transcription factors, pathways, positions of genes and some GO terms. After the selection of concepts with higher support, we observed that there were some concepts with pathways from KEGG and REACTOME as their intent. These pathways are either related to cell proliferation or apoptosis (cell death). The addition of domain knowledge effectively gives an opportunity to obtain the pathway categories shared by larger sets of of genes. Table 4 shows the top-ranked concepts found in each module. For example, in module 5, we have confirmation that *Cytokine Cytokine Receptor Interaction* pathway comes under the category *Signaling Molecules and Interaction* and upper level category *Environmental Information Processing* (Concept ID 4938). The absolute support and stability of the concept containing only the category *Signaling Molecules and Interaction* and upper level category *Environmental Information Processing* as its intent are higher (Concept ID 4995, Table 4) .

To sum up, we were able to discover interesting biological properties of subsets of genes in the three test data sets. As for example, the Focal Adhesion pathway was found to be associated to 17 genes in both modules 1 and 2; the

KEGG category Immune System was found to be shared by 11 to 25 genes in the three cancer modules (Table 4). Given the test data sets, these results are hopeful and constitute interesting positive control. This confirms that FCA-based analysis offers a powerful procedure to deeply explore sets of genes.

Table 4. Top-ranked Concepts with Domain Knowledge

Dataset	Concept ID	Intents	Absolute Support	Stability
Module 1	**9585**	M2192:GGGAGGRR _V$MAZ_Q6	51	0.99
	9571	M2598:GO Membrane Part	27	0.99
	9566	kc:Immune System, kuc:Organismal Systems	25	0.99
	9402	chr19q13	10	0.99
	9078	M10792:KEGG MAPK Signaling Pathway, kc:Signal Transduction, kuc:Environmental Information Processing	12	0.87
Module 2	**6502**	M2192:GGGAGGRR _V$MAZ_Q6	44	0.99
	6496	kc:Immune System, kuc:Organismal Systems	15	0.99
	6388	chr6p21	10	0.97
	6335	M10792:KEGG MAPK Signaling Pathway, kc:Signal Transduction, kuc:Environmental Information Processing	11	0.89
Module 5	**5002**	kuc:Cellular Processes	48	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	M9809:KEGG Cytokine Cytokine Receptor Interaction, kc:Signaling Molecules and Interaction, kuc:Environmental Information Processing	11	0.87

6 Conclusion and Future Work

Our study shows how Formal Concept Analysis can be applied to complex biological data. Data integration and FCA give the flexibility of using various types of attributes (pathways, GO terms, positions, microRNAs and Transcription Factor Targets) for analyzing a list of genes. Our approach gives an insight into how domain knowledge can be introduced in the analysis with the help of

FCA. As for future work, we plan to apply our approach to experimental gene lists and take into account gene-gene relationships (physical Protein Protein Interactions), term-term relationships (Gene Ontology relationships, namely *is-a*, *part-of*, and *regulates*) and relationships between gene positions. Moreover, in order to efficiently deal with the relationships present within the data we can use Relational Concept Analysis.

References

1. Gabriel F. Berriz, Oliver D. King, Barbara Bryant, Chris Sander, and Frederick P. Roth. Characterizing gene sets with FuncAssociate. *Bioinformatics*, 19(18):2502–2504, 2003.
2. Scott Doniger, Nathan Salomonis, Kam Dahlquist, Karen Vranizan, Steven Lawlor, and Bruce Conklin. MAPPFinder: using Gene Ontology and GenMAPP to Create a Global Gene-expression Profile from Microarray Data. *Genome Biology*, 4(1):R7, 2003.
3. Aravind Subramanian et al. Gene Set Enrichment Analysis: A Knowledge-based Approach for Interpreting Genome-wide Expression Profiles. *Proceedings of the National Academy of Sciences*, 102:15545–15550, 2005.
4. Arthur Liberzon et al. Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.
5. Michael Y. Galperin and Xosé M. Fernández-Suarez. The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Research*, 40(Database-Issue):1–8, 2012.
6. Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin/Heidelberg, 1999.
7. Mehdi Kaytoue-Uberall, Sébastien Duplessis, Sergei O. Kuznetsov, and Amedeo Napoli. Two FCA-Based Methods for Mining Gene Expression Data. In Sébastien Ferré and Sebastian Rudolph, editors, *ICFCA*, volume 5548 of *Lecture Notes in Computer Science*, pages 251–266. Springer, 2009.
8. Purvesh Khatri and Sorin Draghici. Ontological Analysis of Gene Expression Data: Current Tools, Limitations, and Open Problems. *Bioinformatics*, 21(18):3587–3595, 2005.
9. Sergei O. Kuznetsov. On stability of a Formal Concept. *Ann. Math. Artif. Intell.*, 49(1-4):101–115, 2007.
10. François Rioult, Jean-François Boulicaut, Bruno Crémilleux, and Jérémy Besson. Using Transposition for Pattern Discovery from Microarray Data. In *DMKD*, pages 73–79, 2003.
11. François Rioult, Céline Robardet, Sylvain Blachon, Bruno Crémilleux, Olivier G, and Jean-François Boulicaut. Mining Concepts from Large SAGE Gene Expression Matrices. In *In: Proceedings KDID03 co-located with ECML-PKDD 2003, Catvat-Dubrovnik (Croatia*, pages 107–118, 2003.
12. Camille Roth, Sergei A. Obiedkov, and Derrick G. Kourie. Towards Concise Representation for Taxonomies of Epistemic Communities. In *CLA*.
13. Zhong S, Storch F, Lipan O, Kao MJ, Weitz C, and Wong WH. GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. *Applied Bioinformatics*, 3(4):1–5, 2004.
14. Eran Segal, Nir Friedman, Daphne Koller, and Aviv Regev. A Module Map Showing Conditional Activity of Expression Modules in Cancer. *Nat.Genet.*, 36:1090–8, 2004.