Using formal concept analysis for mining hypomethylated genes among breast cancer tumors subtypes

Islam Ibrahim Amin¹, Samar K. Kassim², Aboul ella Hassanien³,∗
Hesham A. Hefny¹

¹Institute of Statistical Studies and Researches
Cairo University, Egypt
Email: eng.IsamAmin@gmail.com
²Faculty of Medicine, Ain Shams University, Cairo, Egypt
³Cairo University, Faculty of Computers and Information, Cairo, Egypt
∗Scientific Research Group in Egypt (SRGE) www.egyptscience.net

Abstract—Hypomethylation of DNA have been associated with cancer in several investigations. Hypomethylated of CPG islands associated with promoters can affect the expression of genes to be more expressed. The Illumina GoldenGate Methylation Cancer Panel I can measure DNA methylation at 1505 CpG loci of 806 cancer related genes. A powerful tools to analysis the DNA methylation data are needed. In this paper, formal concept analysis (FCA) is used as data mining tool for mining the hypomethylated genes among breast cancer subtypes, by building formal concepts with significant hypomethylated genes for each breast cancer subtypes. The concept lattice is constructed based on a formal context which is composed of formal concepts. This lattice reflects the biological relationships among breast cancer subtypes.

Keywords: Formal concept analysis, DNA methylation, hypomethylated genes, breast cancer, formal context, concept lattice.

I. INTRODUCTION

Cells are the basic unit of life. Each cell in human body performs a specific task by using many different types of proteins. A part of DNA (gene) is a blueprint of producing protein. Protein synthesis is described by the central dogma of molecular biology which is consisted of two steps: transcription and translation. In transcription, mRNA is produced from DNA. In translation, protein is synthesized from mRNA by using ribosomes. The process of synthesis protein from DNA is called gene expression. A specific protein is producing according to the biological state of the cell and its type. Analysing the complete set of mRNA in the cell is the best way to investigate gene expression by using the microarray. Microarray evaluated the expression of several thousands of genes by measuring the concentration of mRNA with one experiment under several conditions (e.g. normal and cancer tissues). Each gene is represented with a numerical value called gene expression value [1,2]. The output of the microarray is a two-dimensional array called a gene expression data (GED), rows refer to genes and columns indicate to situations [3,4]. The result of GED analysis is to identify differential genes which are responsible for causing cancer, therefore the analysis of GED is very interesting topic of research using data mining approaches (e.g. clustering, bi-clustering and FCA) [5,6]. DNA methylation can affect the gene expression regulation which causing many disease such as cancer. Genes are prevented from being expressed as a result of methylation in a specific region of genes (CPG sites in promoter). Thanks to the development of microarrays, the methylation level in a particular regions (CPG sites) can be measured. Illumine make adaptation in their GoldenGate to measure the methylation level of 1536 CPGs simultaneously in one sample [1,7]. In this paper, formal concept analysis (FCA) is used as data mining tool for mining the hypomethylated genes among breast cancer subtypes, by building formal concepts which representing each a set of significant hypomethylated genes of each breast cancer subtypes. These subtypes are called Basal-like, ERBB2+, luminal B, and luminal A [8]. The methylation data reported in the paper have been previously deposited in NCBI Gene Expression Omnibus (GEO) [9] and are accessible through GEO series accession number [GEO : GSE22135].

The rest of this paper is organized as follows: section II gives a brief background to DNA methylation, definitions of formal concept analysis and Statistical background. Section III shows the proposed method. Section IV shows experiment results. The proposed FCA for breast cancer subtypes approach is discussed in section V. Conclusion and future work discussed in section VI.

II. PRELIMINARIES

A. DNA Methylation

Gene promoter is a regulatory region of DNA that the RNA polymerase enzyme is attached to initiate the transcription process of a gene. Promoters are located upstream of the genes (refer to Figure 1) [10]. Methylation in a specific regions of genes (promoters) can inhibit transcription. Methylation located at specific genomic regions called CPG sites (CPGs). The ”p” in CPG indicate to phosphodiester bond between cytosine (c) and the guanine (G) as shown in Figure 2. Hypomethylated of CPG islands associated with genes can affect the expression of these genes to be more expressed, therefore these genes can be responsible for cancer. Searching for hypomethylated or hypermethylated become an active area of research. The biochemical process of DNA methylation is essential for normal development in the higher organisms.
DNA methylation occurs by adding a methyl group to the 5’ position of cytosine pyrimidine ring, cytosine is one of DNA four bases [11]. Hypomethylated indicating that regions are less methylated (genes be more expressed), but hypermethylated indicating that regions being more methylated (genes be less expressed).

Formal concept analysis was introduced as a mathematical theory modeling by Wille (1982). FCA provides a powerful framework to identify the groups of objects which sharing the common properties [13]. Discovering hidden dependencies by using FCA as a data analysis techniques based on formal context to build a lattice. This lattice is very important for knowledge representation and discovery therefore becoming more interesting for biologists. Showing the benefits of FCA in informations sciences in [14]. The standard definition of FCA in [15].

**Definition 1.** A formal context is a triple $K = (G,M,I)$ consist of a set of attributes $M$, a set of objects $G$ and the binary relation $I$ between $G$ and $M$. (i.e. $I \subseteq (G \times M)$. $(g,m) \in I$ which is read “object g has attribute m”. Formal concept analysis provides intention and extension, extension is a subset of objects and intension is a subset of attributes, which are defined as follows (definition 2).

**Definition 2.**
If a set of $A \subseteq G$,
$\bar{A} := \{ m \in M \mid \exists g \in A \text{ such that } gIm \}$
$\bar{B} := \{ g \in G \mid \exists m \in B \text{ such that } gIm \}$.

**Definition 3.**
Let $(G,M,I)$ is a formal context consists of a pair $(A,B)$ with $B \subseteq M$ called intension, $A \subseteq G$ called extension and $\bar{A} = B \land \bar{B} = A$.

**Definition 4.** Let $(A_1,B_2)$ and $(A_2,B_2)$ are concepts of a formal context $(G,M,I)$ and $A_1, A_2 \subseteq G \land B_1, B_2 \subseteq M$ (if $A_1 \subseteq A_2$ where $A_1 \subseteq A_2$ equivalent to $B_2 \subseteq B_1$), therefore $(A_1,B_1)$ is a subconcept of $(A_2,B_2)$. $(A_1,B_1) \subseteq (A_2,B_2)$ if the $(A_2,B_2)$ is a superconcept of $(A_1,B_1)$. The hierarchical order of the concepts is denoted by $\in \mathbb{B}(G,M,I)$ is called the concept lattice of the context $(G,M,I)$ which is the hierarchical order of the set of all concepts of $(G,M,I)$.  

**C. Statistical Background**

The main task of analysis DNA methylation data is to identify significant genes whose pattern demonstrate a differential change in methylation level under a certain experimental condition. In this paper DNA methylation data analysis implies two phases: non-specific filter and specific filter to identify the most significant hypomethylated genes.

1) **Non-Specific Filtering:** This phase aims to get ride of hypermethylated CPGs which are not candidate to be hypomethylation. By calculating $(\Delta \beta)$, is the difference between the mean methylation level of cancer samples with the mean of methylation level for the corresponding adjacent normal tissue [8], negative values indicate to hypomethylated, whereas positive values indicated to hypermethylation. The selection of hypomethylation markers was performed by applying a non-specific filter $\Delta \beta < \text{Zero}$.

2) **Specific Filtering:** This phase aims to determine the most differential hypomethylated CPGs using an appropriated statistical test after testing the normality of DNA methylation data.
III. THE PROPOSED MINING HYOMETHYLATED GENES

In this paper the DNA methylation level was analyzed in 1505 CPG of 806 cancer related genes in 28 breast cancer subtypes paired samples. The experiment representing the four major breast cancer subtypes performed in 30 paired breast tissues (normal and cancer). Two cancer samples has been ignored because their low methylations levels. Finally there are 28 paired breast tissues (normal and cancer) and four sample for peritumoral region. The methylation level expressed as a continuous values start from zero (completely unmethylated) to one (completely methylated) [8]. Figure (4) shows the main framework of the proposed approach. It contains two stages to identify the most significant hypomethylated genes: (1) non-specific filter and (2) specific filter.

A. Non-Specific Filtering

This step aims to determine the hypomethylated CPGs By calculating ($\Delta \beta$), the difference between the mean of methylation level for Cancer samples with the mean of methylation level for the corresponding adjacent normal tissue. Our experiment deals with 1,505 CPG loci. The final result of applying the non-specific filter ($\Delta \beta < 0$) to obtain the negative values are shown in Figure 3, which indicate to hypomethylated CPGs for each breast cancer subtypes (refer to Table-I).

B. Specific Filtering

According to the proposed model, the output of the non-specific filtering phase is used to be the input to the specific filtering phase. In [5], It is assumed that the distribution is normalized which is generally true for microarray data [17], but this paper takes into consideration the importance of
testing normality to determine which statistical test will be performed. One-sample Kolmogorov Smirnov test is used to test the normality for each breast cancer subtypes, if the data follows normality distribution then the paired t-test will be used otherwise Wilcoxon signed rank is the most appropriate test for a paired samples. After performing multiple testing, the additional filter is applied to reduce the false positives, also It seems logical to suppose that CPGs with a methylation value less than -0.2 are candidate to be significant hypomethylated CPGs [5].

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The non-specific phase has been performed by using the GeneFilter package in R [16]. The output of the first phase reducing CPGs to be 799, 1062, 958, 347 for basal-like, ERBB2+, luminal A and luminal B respectively, after that the normality tested with One-sample Kolmogorov Smirnov test by using SPSS statistical package. Figure (5) shows the results of the test of normality demonstrating that each of breast cancer subtypes methylation values does not follows normal distribution, therefore the most appropriate test is Wilcoxon signed rank (non-parametric test) for the paired samples. The specific filter (Wilcoxon signed rank) applied by using the GeneSelector package in R [18]. The output of this phase filter out CPGs to be 64, 508, 348 and 36 for basal-like, ERBB2+, luminal A and luminal B respectively. The additional filter is performed by using a specific difference between the cancer and normal samples \((Δβ < -0.2)\) for Basal-like and Luminal B. According to the large output of ERBB+2 and Luminal A, the additional filter is \((Δβ < -0.3)\). Table (I) shows the final results are 36, 69, 59 and 16 for basal-like, ERBB2+, luminal A and luminal B respectively which are considered as the most significant hypomethylated CPGs.

<table>
<thead>
<tr>
<th>Subtypes</th>
<th>(Δβ &lt; \text{Zero})</th>
<th>Wilcoxon signed rank</th>
<th>(Δβ &lt; -0.2)</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal-like</td>
<td>799 Cpgs</td>
<td>64 Cpgs</td>
<td>36 Cpgs</td>
<td>30</td>
</tr>
<tr>
<td>ERBB+2</td>
<td>1062 Cpgs</td>
<td>508 Cpgs</td>
<td>69 Cpgs</td>
<td>60</td>
</tr>
<tr>
<td>luminal A</td>
<td>958 Cpgs</td>
<td>348 Cpgs</td>
<td>59 Cpgs</td>
<td>50</td>
</tr>
<tr>
<td>luminal B</td>
<td>347 Cpgs</td>
<td>36 Cpgs</td>
<td>16 Cpgs</td>
<td>14</td>
</tr>
</tbody>
</table>

V. FCA FOR BREAST CANCER SUBTYPES

This section proposes FCA for mining the hypomethylated genes among breast cancer molecular subtypes. The formal context is first extracted, then we can construct the concept lattice based on a formal context. In this study the Java-based open source ConExp program was used as a formal concept analysis tool to generate the lattice diagram as shown in Figure 6 [19].

A. Formal context

The formal context represented as a table which is displaying the relations between cancer subtypes (objects) and hypomethylated genes (attributes) as shown in Table II.
B. Formal concept lattice

The formal context is used to construct a concept lattice. The concept lattice will construct every relationship between breast cancer subtypes and hypomethylated genes. According to the lattice in Figure 6, we can identify 15 concepts. The concept lattice makes these subtypes hierarchically grouped together according to their common hypomethylated genes. By using the lattice, concepts (1, 2, 3 and 4) identify the hypomethylated genes for each breast cancer subtype. The distinct common hypomethylated genes among breast cancer subtypes can be discovered from lattice as shown in Table III. This lattice helps us to answer many questions, for example if we need to know what are the common hypomethylated genes between Luminal A and Luminal B, the answer is, by using the red arrows we start navigating from concepts (3 and 4) to reach any higher connected concepts, finally Luminal A and Luminal B shared CXCL9, MAGEA1, VAMP8, PTPN6, PTK6, IL8, and MUC1. If we need to know what are subtypes shared genes TFF1 and SPl1, we need to find the concepts of these genes, by using the blue arrows we start navigating from concepts (11 and 12) to reach any lower connected concepts, finally the ERBB+2 and Luminal A are sharing these genes.

VI. CONCLUSION AND FUTURE WORK

The results of DNA methylation data analysis are encouraging, a powerful tool to discover hidden dependencies among breast cancer subtypes is needed. In this paper, formal concept analysis (FCA) is used to identify the distinct common hypomethylated genes in breast cancer subtypes, also answering many questions about the relationships between hypomethylated genes and breast cancer subtypes. Applying formal concept analysis begins with the construction of the formal context which represents the relations between cancer subtypes and hypomethylated genes. Based on formal context, the concept lattice is constructed which leads to discover a new relationship. This paper aims to analysis the DNA hypomethylated therefore, future research needs a combined analysis of hypomethylated and hypermethylated among breast cancer subtypes to discover more new relationships.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Subtypes</th>
<th>Hypomethylated Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>C(14)</td>
<td>Luminal A, Luminal B, Basal-like, ERBB+2</td>
<td>MUC1</td>
</tr>
<tr>
<td>C(13)</td>
<td>Luminal A, Luminal B, Basal-like</td>
<td>VAMP8, PTPN6, PTK6, IL8</td>
</tr>
<tr>
<td>C(12)</td>
<td>Luminal A, Luminal B, ERBB+2</td>
<td>NID1, TFF1</td>
</tr>
<tr>
<td>C(11)</td>
<td>Luminal A, Basal-like, ERBB+2</td>
<td>SPI1, PTPRH, PTHR1, IFNG, EMR3, HGH-52, BCAP31, FLJ20712</td>
</tr>
<tr>
<td>C(10)</td>
<td>Luminal A, Luminal B</td>
<td>CXCL9, MAGEA1</td>
</tr>
<tr>
<td>C(9)</td>
<td>Luminal B, ERBB+2</td>
<td>TRIP6</td>
</tr>
<tr>
<td>C(8)</td>
<td>Luminal A, ERBB+2</td>
<td>APOC1, DLCL1, EGF, CLDN4, GRB7, MOS, TDG, MCF2, MAGEC3, MAPK4, IL12B, NOS3, PADA4, PARA, PDGFB, PECAM1, PSCA, SNURF, PLA2G2A</td>
</tr>
<tr>
<td>C(7)</td>
<td>Luminal B, Basal-like</td>
<td>ITK</td>
</tr>
<tr>
<td>C(6)</td>
<td>Luminal A, Basal-like</td>
<td>TNFSF8, AIM2</td>
</tr>
<tr>
<td>C(5)</td>
<td>Basal-like, ERBB+2</td>
<td>CCL3, MMP10</td>
</tr>
<tr>
<td>C(4)</td>
<td>Luminal B</td>
<td>TPO, SEMA3C, ARG1GDH</td>
</tr>
<tr>
<td>C(3)</td>
<td>Luminal A</td>
<td>IL3, WNT16B, PESS1, MPL, MEST, MECP2, KRAS, KLK11, HLA-DPA1, CEACAM1, CD1A, CARD15</td>
</tr>
<tr>
<td>C(2)</td>
<td>Basal-like</td>
<td>BLK, CASP8, CCR5, CD2, ET52, EVL2A, FZ2, GABRG3, MT1A, RUNX3, SPP1, SRC</td>
</tr>
<tr>
<td>C(1)</td>
<td>ERBB+2</td>
<td>ARP10A, B3GALT5, BGN, BRCA1, C4B, CVPI2I, XIST, DSG1, EPHX1, GABRA5, H19, HDAC6, ILRN, IRAK1, MAPK10, MBD2, NIL1, PAP1, PROM1, S100A2, SNRPN, SDFEP, TDGFI, TIMP1, VMP1, CCKAR, DNASE1L1</td>
</tr>
</tbody>
</table>
Fig. 6. The concept lattice of hypomethylated genes in breast cancer subtypes.

REFERENCES


