Breast Cancer Combinatorial Biomarkers Identified Using FCA

1. Introduction

Complex disorders, such as cancer, have emerged as the most significant medical research problem of our time. A genetic component is known to exist in several of these disorders. As a result, DNA sequence analysis and gene expression measures are becoming increasingly important in both basic research and clinical therapy of disorders like cancer. The emphasis of this research is on combinatorial biomarkers that can be utilized for illness diagnosis or prognosis. A combinatorial biomarker is a collection of genes that may consistently discriminate between two classes (for example, healthy and malignant tissue, or metastasis and initial tumor) without the need of a single gene.

This means that a combinatorial biomarker can be made up of genes whose individual activity is unrelated to the classes of samples — the discriminative ability comes from the combination of genes. Many single genes have been found so far, each of which has a mutation or aberrant activity that promotes the spread of cancer. A mutation in one of the genes BRCA1 or BRCA2 (BRCA stands for BReast CAncer) has been linked to an increased chance of developing breast cancer, according to many studies. However, there are still numerous cases that aren't explained by single-gene anomalies. Combinatorial biomarkers are a natural extension of the single gene method and a way to learn more about complicated disorders.

2. Theoretical framework

This section explains the biological context and recalls the Formal Concept Analysis concepts that will be discussed throughout this paper.

2.1. Manifestation of genes

Chemical interactions between molecules underpin all of the operations in a live cell. Proteins are molecules that each cell produces from blueprints encoded on DNA. A *gene* is a fragment of DNA that contains the blueprint for a protein. The chemical processes that can occur can be regulated by modifying the number or composition of proteins present inside the cell. *Gene expression*, which can be observed in Figure 1, is the process through which a protein is produced from its gene and has two steps: transcription and translation.

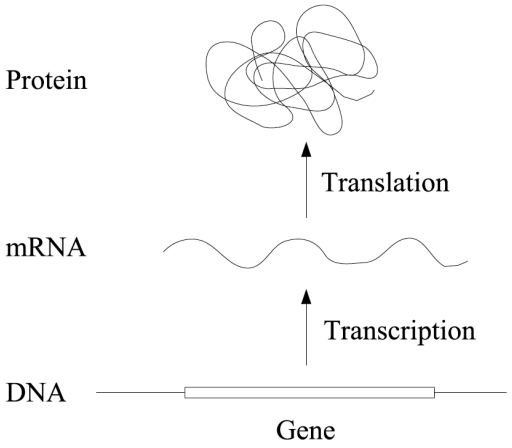

Figure 1 - The process of gene expression [1]

A copy of the gene is created on DNA in the first stage, transcription. This copy's "hardware" is an RNA molecule known as messenger RNA (or mRNA for short). In the translation phase, the protein is made from the gene's mRNA copy. The greater the number of mRNA copies of a gene present in a cell, the more proteins may be made from it. As a result, measures of mRNA abundance may be used to infer the condition of a cell. There are roughly 25.000 genes in human DNA.

The mRNA abundance of gene g in the sample is represented by an n-dimensional vector e_g , where $g \in \{1,..., n\}$ represents the mRNA abundance of gene g in the sample: The more mRNA transcripts there are, the greater the e_g.

The *expression value* of gene g is represented by the value e_g. A *gene expression data set* is a collection of such sample measurements summed in a $n \times m$ matrix E with each row corresponding to a gene and each column corresponding to a sample, so that egs is the expression value of gene g in sample s, $s \in \{1, ..., m\}$.

The purpose of this application is to uncover gene combinations that can distinguish between two types of samples. It is generally more convenient to evaluate gene expression variations between samples than their absolute expression values for this purpose. The *log ratio* is a standard metric for comparing the expression change of a gene g between two samples s and t.

$$l_g = log_2(\frac{e_{gs}}{e_{gt}})$$

If a gene has a log ratio of 1, its expression value in sample s is two times higher than in sample t. A log ratio of -1 indicates that the gene's expression value in sample s is two times lower than in sample t. Gene expression matrices typically contain many more genes than

samples, which is an unpleasant scenario. This imbalance makes statistical analysis of gene expression data extremely challenging.

2.2. Formal Concept Analysis (FCA)

The notation is quite similar to that of [2]. Let O denote a collection of objects, P denote a collection of qualities (also known as attributes), and $I \subseteq O \times P$ denote a binary relationship. To prevent confusion with the set of genes G that will appear later in the article, we use the letters O and P for the set of objects and characteristics. If $(o, p) \in I$, we may also write oIp and read "object o has the property p." A formal context is the triple $K = (O, P, I)$. We look at the most common derivation operators: Let $A \subseteq O$ and $B \subseteq P$, then:

$$A^{'} = \{p \in P : oIp \ \forall o \in A\}$$
$$B^{'} = \{o \in O : oIp \ \forall p \in B\}$$

We shall use the shorter notation o instead of o for sets containing only a single object, and similarly for attributes. If $A = B$ and $B = A$, the pair (A, B) is a notion of $K = (O, P, I)$. The set A is referred to as the extent, while the set B is referred to as the purpose (A, B) . A one-valued formal context is one in which an object either has the (one value of the) attribute or does not have the (one value of the) attribute. Because our attributes will have several values that might apply to an object, we'll require many-valued contexts for our discussion. Let W denote the range of values that attributes can have.

The quadruple $K = (O, P, W, I)$ is thus referred to as a many-valued context, with I now being a ternary relation ($I \subseteq O \times P \times W$). If $(o, p, w) \in I$, we may alternatively write $p(o) = w$, which means "the value of property p for the object o equals w." If o is an object and p is a property, $p(o) = w$ and $p(o) = v$ must imply that $w = v$. To locate ideas in a many-valued context, discretize the many-valued qualities and transform the context into a one-valued one. In Formal Concept Analysis, this method is known as *conceptual scaling*.

3. FCA for Combinatorial Biomarkers

The strategy for identifying combinatorial biomarkers that we propose may be used with both absolute gene expression data and log ratios as input data. In both circumstances, the input data are treated as a many-valued formal context $K = (S, G, W, I)$, where S is the set of samples, G is the set of genes and $(s, g, w) \in I$ (or $g(s) = w$) if gene g has an expression value or a log ratio w in sample s, respectively.

Furthermore, the samples are divided into two groups: a target class C_t (for example, sick tissue) and a background class C_b (healthy tissue). The goal is to discover a group of genes that separates these two classes without requiring a single member gene to do so. The class memberships of the samples in the gene expression data set are assumed to be known. The issue of discovering a combinatorial biomarker is thus approached as a traditional classification problem with certain extra limitations.

3.1. Context of Training and Validation

To begin, there is the context of various values. $K = (S, G, W, I)$ has two subcontexts: $KT = (ST, G, W, I)$ and $KV = (SV, G, W, I)$, with ST and SV being disjoint and $STSV = S$. The training and validation contexts are referred to as KT and KV, respectively. The training context will be used to identify combinatorial biomarkers using an FCA classification approach, while the validation context will be utilized to validate the results. ST must include examples from both classes, and SV should as well.

3.2. Scaling

Initially, we must scale the training environment in order to find combinatorial biomarkers for the target class. It's worth noting that the G gene set serves as the KT attribute set in this case. There are recommended two distinct scaling strategies depending on the kind of input data - which will be presented below:

- Absolute expression values
- Log ratios

Absolute expression values - regarding this application, the authors proposed that the scaling technique be guided by the general characteristics of gene expression data sets - which are:

1. the non-comparability of gene expression levels
2. a significant amount of noise

For each gene g, we wish to use a dichotomous scale of the form. This implies that for gene g, we set a tg threshold and substitute the many-valued characteristic g with the two one-valued attributes "expression value of $g \leq t_g$" and "expression value of $g > t_g$". Due to the incomparability of expression levels between genes, the threshold value t_g must be selected individually for each gene g.

$$\mathbb{S}_g := \begin{array}{|c||c|c|}\hline & \leqslant t_g & > t_g \\ \hline\hline \leqslant t_g & \text{X} & \\ \hline > t_g & & \text{X} \\ \hline \end{array} \quad (3)$$

The threshold should also be resistant to noise in gene expression data. If the data is supposed to have a noise level of l, we need to make sure that:

$$\begin{aligned} \max\{g(s) \ : \ g(s) \leqslant t_g, s \in S_T\} \leqslant t_g - \ell \\ \min\{g(s) \ : \ g(s) > t_g, s \in S_T\} > t_g + \ell. \end{aligned} \quad (4)$$

This condition ensures that even if the training context is disrupted by as much noise as l, using the thresholds t_g produces the same one-valued context as the original training context, and the subsequent combinatorial biomarker identification stays the same.

We performed the scaling technique by sorting the expression levels of gene g in the validation context so that $e_{g1} \leq e_{g2} \leq ... \leq e_{g|S_T|}$ and then looking for the greatest interval $[e_{gi}, e_{gi+1}]$, $i \in \{1,.., |S_T|\}$. Let's call this interval $[e_{gk}, e_{gk+1}]$ (if there are many greatest intervals, one is chosen at random). When $e_{gk+1} - e_{gk} \geq 2l$, condition (4) is met, and the threshold t_g is set to

$$t_g = \frac{e_{gk+1} + e_{gk}}{2}. \quad (5)$$

With the goal of improving the combinatorial biomarker's resilience against noise, those genes that do not fulfill the threshold tg requirement (4) are removed from the context. As a result, the scaling technique contains an implicit feature selection approach, whose strictness may be modified by the parameter.

Log ratios - when applying log ration all the properties for a single scale, given by

$$\mathbb{S} := \begin{array}{|c||c|c|}\hline & \leqslant -t & \geqslant t \\ \hline\hline \leqslant -t & \text{X} & \\ \hline \geqslant t & & \text{X} \\ \hline \end{array} \quad (6)$$

R^+'s threshold value t is used here. The two one-valued qualities "log ratio of $g \leq - t$" and "log ratio of $g \geq t$" replace each many-valued property g in this scaling. A common

threshold in gene expression data analysis is 1, which means that the gene's expression must vary by at least $2^1 = 2$. t can be either smaller or larger depending on the noise in the data set and the desired selectivity. The higher the t, the fewer samples will meet the scaled features, and the ensuing search for combinatorial biomarkers will be confined to genes with at least a $2^t - fold$ change in expression in one of the S_T samples.

3.3. The Methodology of Identification

$K_T = (S_T, G_S, J)$ signifies the scaled training context, where G_S is the collection of scaled attributes and J defines the obtained incidence relation. All FCA classification systems rely on the notion of positive and negative hypotheses, which is defined in terms of positive and negative hypotheses in [3]. We simplify notations and consider so-called homogeneous concept intentions to avoid separating KT into a positive and negative subcontext as in [3].

Assume that $B \subseteq G_S$ is a single notion purpose. B, obviously, indicates a subset of samples from the same class. Furthermore, if B_1 is a homogeneous concept intent and B_2 is a concept intent with $B_2 \supset B_1$, then B_2 is a homogeneous concept intent as well. We're particularly interested in the uniform idea intentions for the target class while looking for combinatorial biomarkers.

The smallest of these intentions, with the fewest features, are the most broad and apply to the biggest subgroups of target samples. As a result, we only produce the smallest homogenous concept intentions, i.e. those that are closest to the top element of the K_T concept lattice.

B comprises all combinatorial biomarkers for the target class if the extent corresponding to a homogenous idea intent B contains all samples of the target class in the training context. B is, in fact, the biggest combinatorial biomarker for the target class in K_T in this example. It's also possible that the target class has no one homogenous notion. The class is then specified by many homogenous concept intentions (with proper scaling), which might be read as a division into subclasses. In this situation, we would treat each subclass independently and create combinatorial biomarkers for each subclass as a result. The following formula can be used to calculate homogenous concept intents:

1. Check to see whether $S_T \cap C_t$ is a concept extent. Return $(S_T \cap C_t)'$ if this is the case
2. Alternatively, compute the subcontext's iceberg lattice $((S_T \cap C_t)'', G_S, J)$ and utilize the homogeneity of a concept intent as a halting condition (rather than the normal

support threshold used in iceberg lattices (see [14] for an introduction to iceberg lattices)). Return all concept intentions that were determined to be homogenous

3.4. Post-Processing and Validation

Due to the obvious imbalance in gene expression data (thousands of genes describing just a few samples), the homogenous concept intent B for the target class is likely to have many more genes than are required to properly separate the target and background classes. Combinatorial biomarkers for illness diagnosis and prognosis should be brief and reliable. When a biomarker has too many genes, evaluating it for a single patient takes a long time and is consequently costly. The marker, on the other hand, must be extremely resistant to noise and be able to accurately identify the target class.

With these conditions in mind, we select from the identified homogenous concept intent subsets of genes that are sufficient for identifying the target class, comprise only a few genes, and are relatively resilient in the training scenario. More exactly, we look for $M \subset G_s$ subsets:

1. $M' = B'$
2. $|M| \leq k$
3. $\forall s \in S_T \backslash C_T$ there are at least r genes $g_1,\ldots, g_r \in M$ with $\forall i \in \{1,\ldots, r\}\ g_i \notin s'$ (7)

The greatest number of genes permitted by the biomarker is k, and the lowest number of genes in a sample from the background class that do not meet the biomarker's criteria is r. The settings of these parameters must be chosen based on the application in question. We set for $s \in S_T \backslash C_t$ to simplify notations:

$$gap(S) = |\{g \in M : g \notin s'\}| \quad (8)$$

Due to its computing cost, an exhaustive search for all combinations of at most k genes is plainly prohibitive. As a result, we propose a simple heuristic, as shown in pseudo code in Figure 2.

1. set $M := \emptyset$, $L := S_T \setminus C_t$, `iter` := 0, `maxiter` := 1000
2. repeat
3. find a sample $s \in L$ with $|s' \cap B|$ maximal
4. $a := r - \text{gap}(s)$
5. $D := \{g \in B \ : \ g \notin s'\}$
6. randomly select genes $g_1, \ldots, g_a$ from D and set $M := M \cup \{g_1, \ldots, g_a\}$
7. $L := L \setminus \{s \in L \ : \ \text{gap}(s) \geqslant r\}$
8. if $|M| > k$, then $M := \emptyset$, $L := S_T \setminus C_t$, `iter` := `iter`+1
9. until $(M' = S_T \cap C_t)$ and $(L = \emptyset)$ or (`iter` = `maxiter`)

Figure 2 - Heuristic to extract candidate combinatorial biomarkers [1]

Starting with an empty set of genes M, we add genes from the homogenous concept intent B to M one by one until the conditions 1. and 2. from (7) are met. For this, we preserve a list L of the samples from the background class that haven't yet met criterion 3. In order for a sample $s \in S_T \setminus C_t$ to meet criterion 3., the biomarker M must contain r genes from the set $D = \{g \in B : g \in s'\}$. This is accomplished by gradually adding genes from D to M,, beginning with the most problematic samples, i.e. samples s for which $|s' \cap B|$ is the highest (lines 3. to 6.). When sample s is taken into account, M already includes $gap(s)$ genes whose threshold criteria are not met by s.

To acquire the r genes necessary by condition 3. for s, only genes must be introduced to M. (lines 4. and 6.). Samples from the background collection that meet criterion 3. are deleted from L after each addition of genes to$M$. (line 7.). If $|M|$ exceeds the pre-selected threshold k during this phase, the selection procedure is restarted (line 8.). The selection method is done at most maxiter times to avoid a possible non terminating loop. When the technique is repeated numerous times, the unpredictability in the gene selection creates a diversity of potential biomarker possibilities.

4. Breast Cancer Application of Combinatorial Biomarkers

Now the approach outlined in Section 3 is applied to a real-world dataset of gene expression data derived from breast cancer tumor biopsy samples. The data set includes 50 samples, 20 of which are metastases, 28 of which are initial tumors, and two of which are healthy tissue samples. The expression of 22.215 genes was assessed in each sample, resulting in a gene expression context $K = (S, G, W, I)$ with $|S|$ = 50 and $|G|$ = 22.215.

The goal is to discover biomarkers that may be used to detect metastases. The data set from this study involves systematic biases due to the biopsy samples being produced by separate laboratories, in addition to the noise problem that affects all gene expression data. On the one hand, this makes identifying a biomarker more difficult because classification procedures work best with homogeneous data sets, but on the other hand, this is a more

realistic scenario because a biomarker used in clinical practice must work regardless of the laboratory that provides the sample.

The training context K_T was chosen at random from the 50 samples, consisting of 16 primary tumors and 13 metastases. The remaining samples are in the validation context. The mean expression value of the 16 main tumors in K_T is used as a reference to translate the absolute expression values in both K_T and K_V into log ratios. The log ratio of gene g in sample s is calculated informally as

$$l_{gs} = log_2(\frac{e_{gs}}{\bar{e}}) \qquad (9)$$

where $\bar{e}$ denotes the average of gene g expression values among the 16 primary tumor samples used in the training. The log ratio scaling approach mentioned in Subsection 3.2 is used to scale K_T, with a threshold of $t = 1$. In K_T, the class of metastatic samples has just one homogenous concept intent B. There are 42 genes in this concept. We discover 8 combinatorial biomarkers that pass the validation using the heuristic from Figure 2 with k = 4 and r = 2.

The bulk of the genes in these indicators have extracellular activities, which is consistent with the idea that metastasis cells must modify their exterior structure and surroundings in order to travel across the body. Other genes are engaged in processes that are particular to breast and brain tissue, which makes sense given that the samples in the data set come from these areas (primary tumors from breast and metastases from brain). Three of the genes have already been linked to cancer progression and metastatic development in the literature.

In this article was utilized the scaled context of the breast cancer data set to infer decision trees with the MATLAB software's "treefit" function to gain an idea of how our FCA-based biomarker identification technique compares to other classification methods. To produce 2000 biomarkers, we employed both our selection heuristic based on homogenous concept intent and the treefit function. The selection method's parameters were k = 4 and r = 2. With default settings, the treefit function was utilized. We randomly arranged the columns of the training context to construct a number of distinct decision trees because the decision tree builder is deterministic. Table 1 summarizes the outcomes of the validation of the 2000 biomarkers. Three parameters are compared:

- **accuracy** - the number of samples properly categorized divided by the total number of samples (averaged over all produced biomarkers)
- **valid** - the amount of valid biomarkers that have been created

- **genes** - number of genes utilized in biomarkers on average

	accuracy	valid	genes
FCA	85.3 %	29	3.8
treefit	81.0 %	0	1

Table 1 - The FCA approach and the MATLAB decision tree method for biomarker identification are compared [1]

The average accuracy of both systems is not significantly different, as can be shown. However, the FCA technique generates a large number of genuine biomarkers (those with a 100% accuracy), whereas treefit generates none. The average number of genes employed in biomarkers, as seen in the third column of Table 2, is one explanation for this. There is just one gene in every decision tree formed. This is due to the fact that while a few single genes can discriminate between main tumor and metastasis in the training setting, none of them can do so in the validation situation.

5. Conclusion

The findings from the breast cancer study are promising. They demonstrate that our strategy may successfully identify valid combinatorial biomarkers in real-world gene expression data. The strategy gave satisfactory results even in the tough scenario of working with a heterogeneous data collection. The found combinatorial biomarkers not only exhibit the expected behavior in terms of their capacity to detect metastasis, but they also comprise genes that make biological sense. When we compare our method to a decision tree-based method, we can show that the FCA approach beats the decision tree algorithm.

Although there is still a long way from putting our combinatorial biomarkers into clinical practice due to the study's modest size, our findings may already be utilized to research the mechanisms involved in the formation of metastases. Rather than focusing on single genes whose expression changes between metastases and original tumors, our combinatorial biomarkers give insight on the gene-gene interaction that is distinctive of metastases. This addresses the concerns of a current line of medical study that suggests that the genesis of metastases is considerably more complex than single genes, requiring a certain constellation of expression changes to occur.

REFERENCES

[1] - Motameny, Susanne, Beatrix Versmold, and Rita Schmutzler. "Formal concept analysis for the identification of combinatorial biomarkers in breast cancer." *International Conference on Formal Concept Analysis*. Springer, Berlin, Heidelberg, 2008.

[2] - Ganter, Bernhard, and Rudolf Wille. *Formal concept analysis: mathematical foundations*. Springer Science & Business Media, 2012.

[3] - Kuznetsov, Sergei O. "Machine learning on the basis of formal concept analysis." *Automation and Remote Control* 62.10 (2001): 1543-1564.