# Inferring Microarray Relevance By Enrichment Of Chemotherapy Resistance-Based MicroRNA Sets

Koray Açıcı, Hasan Oğul
Computer Engineering Department
Baskent University
Ankara, Turkey
{korayacici,hogul}@baskent.edu.tr

***Abstract*—Inferring relevance between microarray experiments stored in a gene expression repository is a helpful practice for biological data mining and information retrieval studies. In this study, we propose a knowledge-based approach for representing microarray experiment content to be used in such studies. The representation scheme is specifically designed for inferring a disease-associated relevance of microRNA experiments. A group of annotated microRNA sets based on their chemotherapy resistance are used for a statistical enrichment analysis over observed expression data. A query experiment is then represented by a single dimensional vector of these enrichment statistics, instead of raw expression data. According to the results, new representation scheme can provide a better retrieval performance than traditional differential expression-based representation.**

***Keywords*—microRNA, information retrieval, content-based search, gene expression database.**

## I. INTRODUCTİON

Microarray technology has been resulted with a great amount of experimental results deposited in large public gene expression databases, such as GEO [1] and ArrayExpress [2]. Effective use of these data urges novel search strategies, which will go beyond simple meta-data based querying interfaces. A latest trend is to use content-based search as similar to its applications in multimedia information retrieval [3]. Implementing a content-based search interface requires the realization of a method for inferring relevance between two microarray entries. Since direct comparison of two gene expression matrices is not feasible due to several computational and biological constraints, we need a single dimensional representation to infer similarity between two entities. In previous studies, this problem has been approached using experiment fingerprints which are indexed by the corresponding genes involved in the experiment [4-9]. To quantify the behavior of genes, the fingerprint vector is usually filled out by their differential expression levels. Another approach is to use gene sets, instead of individual genes, to index the fingerprints. In this case, each entry in a fingerprint denotes its enrichment score in relevant to observed expression data [10, 11]. Although second approach is more promising in terms of the biological interpretability of results, it suffers from the lack of reliable sets for every context. For example, none of current set enrichment analysis implementations can offer predefined functional sets for microRNAs (miRNAs). Therefore, the concept of set enrichment analysis has not been used for miRNA microarray experiments so far. On the other hand, miRNAs have been recently shown to be important actors in gene expression [12]. Consequently, miRNA data have become a valuable resource in integrative analysis of gene regulation and its relations to diseases [13].

In this study, we particularly focus on retrieving microRNA microarray experiment where experiment relevance is defined based on their disease associations. To this end, we evaluate the ability of microRNA sets built according to their similarity on chemotherapy resistance. Hypothesizing that chemotherapy resistance is strongly relevant to the type of tumor under treatment; the knowledge-based representation that we propose here is anticipated to indentify relevance between different microarray samples with same disease conditions. Our experiments on a common information retrieval setup justify our argument.

## II. MATERIALS AND METHODS

### A. Microarray Experiment Retrieval

Given a microRNA expression matrix $E$, where $e_{ij}$ represents the expression of the $i^{th}$ microRNA in $j^{th}$ condition, the task is to find the matrix $M_k$ among the collection of matrices $\{M_1, M_2,\ldots,M_t\}$ in the microarray repository, where $k$ yields the highest similarity $s(E,M_k)$. The comparison of these matrices is usually realized through simpler one-dimensional vectors, called fingerprints. A fingerprint represents the experiment content in a retrieval framework. The information retrieval model is built on matching the fingerprints of overall microRNA expression profiles of two matrices, namely, the similarity $s(E,M_k)$ is defined over the fingerprints of $E$ and $M_k$ instead of their entire matrices. In this study, an experiment matrix refers to a microarray measurement of a number of miRNAs in two conditions; control and treatment.

A common way to represent a microarray experiment by a fingerprint is to design it as a vector of differential expressions of all microRNAs measured in the experiment. Having a differential expression $x_i$ for $i^{th}$ miRNA, an experiment $E$ is represented by the fingerprint vector $X=\{x_1,x_2,\ldots,x_N\}$ for a microarray that contains $N$ microRNA entries. Although it is easy to implement, this representation does not embody any biological knowledge. The fingerprint itself is usually not interpretable in defining the relevance between experiments. In

978-1-4673-7939-7/15/$31.00©2015 IEEE

this study, we offer a new knowledge-based and more interpretable way of representing miRNA microarray experiments in a targeted information retrieval task. Instead of indexing a fingerprint directly from involved miRNAs, we use an index of predefined miRNA sets based on chemotherapy resistance. In this case, the length of an experiment fingerprint becomes equal to the number of such miRNA sets in our repository (Fig. 1). The value of each miRNA sets contributing to the experiment fingerprint is calculated using GSEA (Gene Set Enrichment Analysis) algorithm.

### B. Gene Set Analysis

GSEA is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states [14]. Given a microRNA list with their differential expression value, the GSEA algorithm broadly follows these steps:

```
1. Rank microRNAs by differential expression
2. For each microRNA set:
    2.1. Compute cumulative sum for ranked microRNAs
    2.2. Report the maximum deviation from zero as the enrichment score (ES)
3. Report microRNA sets ranked by their ES values
```

The step 2.1 starts with initializing a sum, then goes over the ranked microRNA list as to increase the sum if the current microRNA is in the set, and decrease it otherwise. Here, the magnitude of increment depends on correlation of the microRNA with the phenotype under consideration.

Since current GSEA implementations cannot favor ready lists for functional miRNA sets, we create our own based on their chemotherapy resistance. These relations were obtained from CREAM database [15]. CREAM database provides an omnibus repository for discovering and depositing chemotherapy resistance-associated miRSNPs which were identified from multi-omics high-throughput data. Using this collection, one can infer the sets of miRNAs which are associated with same chemotherapy resistance. We created 276 miRNA sets from this collection, each of which responds to a different resistance type.

GSEA provides an enrichment score for each chemotherapy-based miRNA set. Instead of using this score as such, we select the miRNA sets which are significantly enriched in the experiment and set their fingerprint values to 1, indicating that this miRNA set is functional in this experiment. If the enrichment score is lower than a threshold value, corresponding fingerprint value for this miRNA set is determined as 0. The threshold is actively selected for each experiment. To realize an active selection, we choose a top K% percent of the sets which attain a p-value of lower than 0.05 in the ranked list of miRNA sets and determine these entries as significantly enriched. Optimal value of K is exhaustively searched over all collection of experiments so as to maximize the retrieval performance.

### C. Similarity Metric

Inferring the relevance of two experiments is done by computing a similarity score between their fingerprints. We use Tanimoto metric for this object. Given two fingerprint vectors X and Y with the length of N, their similarity is defined as follows:

$$s(X,Y)=\frac{\sum_i (x_i \cap y_i)}{\sum_i (x_i \cup y_i)}$$

In our context, this metric measures the overlap between the paired microRNA set categories, i.e. a binary indicator of being enriched or not in those experiments.

### D. Definition of Relevance

The relevance of two experiments is characterized by their disease associations. In this scheme, two experiments are said to be relevant if they both has a treatment sample labeled by the same disease. Our basic assumption here is that querying an experiment labeled with a specific disease should retrieve the other experiment associated with the same disease in the top of the retrieved list while scoring the other types of disease samples in the bottom of the list.

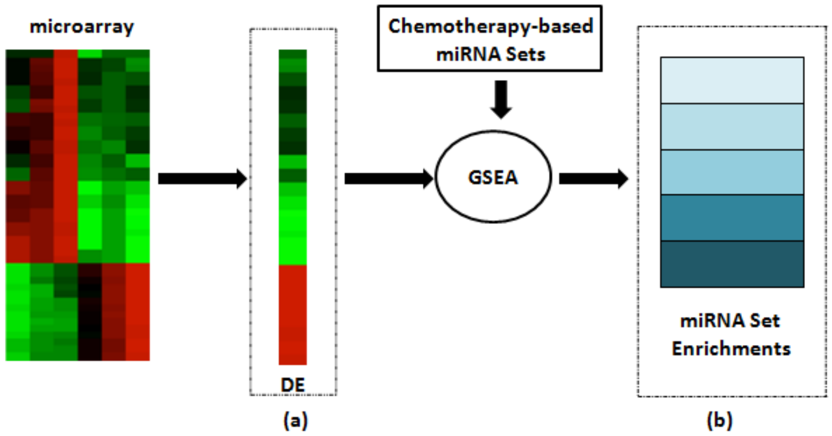

Fig. 1. Fingerprints for miRNA microarray experiments: (a) traditiional differential expression-based fingerprint, (b) new chemoterapy resistance-based fingerprint

### *E. Evaluation Criteria*

A content-based database search platform is simulated by taking all experiments in the collected dataset as a whole database and query each experiment in the database by leaving it out. The system is permitted to retrieve all experiments in the database in a decreasing order by their similarity scores. It is expected that the higher rank an experiment has, the more likely it is relevant to the query. A common way of evaluating the retrieval performance in such scenario is to use Receiver Operating Characteristic (ROC) curves. A ROC score for each positive sample (a relevant experiment) is calculated by the area under the ROC curve associated with it (See Algorihtm 1).

```
Sort the retrieved experiments by their similarity scores
Get a sorted list of relevance labels (1 or 0)
tp=0 /* Initialize true positive */
fp=0 /* Initialize false positive */
roc=0 /* Initialize ROC score */
for each of the sorted label do
    if (label=1)
        tp=tp+1
    else
        fp=fp+1
        roc=roc+tp
if (tp=0)
        roc=0
else if (fp=0)
        roc=1
else
        roc=roc/(tp*fp)
```

Algoirthm 1. The pseudo-code to compute the ROC score for a query experiment.

The overall performance is depicted by another curve plotted for the number of microarray experiments in the vertical axis having a higher ROC value than corresponding ROC value in horizontal axis. Average ROC scores are also reported for all alternative approaches. A higher value of average ROC score indicates a better retrieval performance where a ROC score of 1 represents the perfect case.

### *F. Dataset*

The dataset we used to assess retrieval performance involves 135 microarray entries, where each experiment is relevant to a specific disease, i.e. bladder cancer, brain cancer, breast cancer, colon cancer, ILD, kidney cancer, leukemia, lung cancer, pancreas cancer, prostate cancer, schwannoma and uterus cancer. All diseased samples are paired with at least one control microarray.

## III. RESULTS

The system's ability to retrieve relevant experiments is measured by ROC score computed for each experiment separately by the area under the corresponding ROC curve. A parameter to be adjusted for obtaining experiment fingerprints is the global *K* value which indicates the fraction of microRNA sets from GSEA output to be included in the fingerprint as significantly enriched. After an exhaustive search that is iteratively run, we find that the system performs best when *K* equals to 45 when we take the average ROC score of all experiments as the objective to be maximized (Fig. 2).

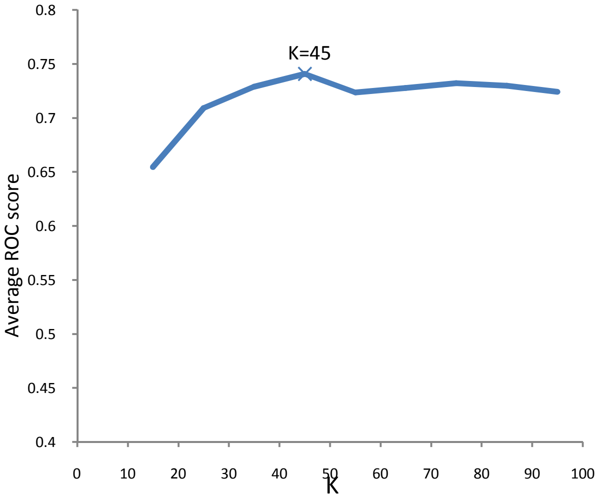

Fig. 2. Finding optimal K (the fraction of microRNA sets to be set as enriched in the experiment).

Figure 3 shows the plots of number of experiments for which the system performs better than given ROC score. A higher curve refers to a more effective retrieval performance. As shown, the new fingerprinting scheme can lead to more successful retrieval performance than differential expression-based fingerprints.

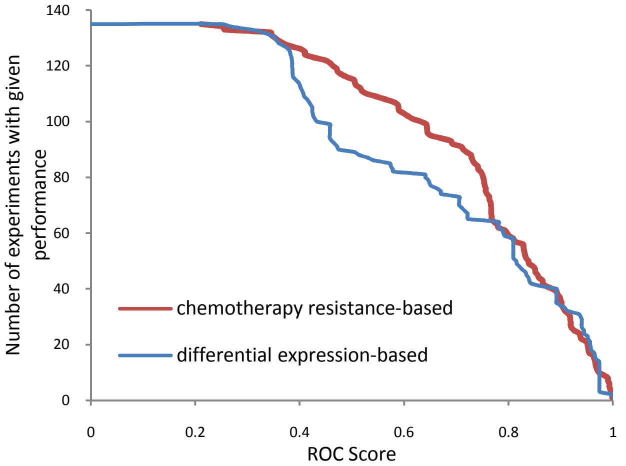

Fig. 3. Retrieval performance shown by the plots of number of experiments that exceeds a given threshold ROC with two fingerprinting scheme.

To justify our argument, we also present a scatter plot for ROC scores of two fingerprinting schemes (Fig. 4). The choice of similarity metric was validated by comparing Tanimoto metric with another common metric, Spearman Correlation Coefficient (CC). The Spearman CC directly uses the exact value of enrichment scores and computes a correlation between rank of two fingerprints. The scatter plot (Fig. 5) justifies that Tanimoto metric can provide a slightly better retrieval performance. Since the computational complexity of Tanimoto metric is much lower than that of Spearman metric, this result suggests the use of Tanimoto metric in this context.

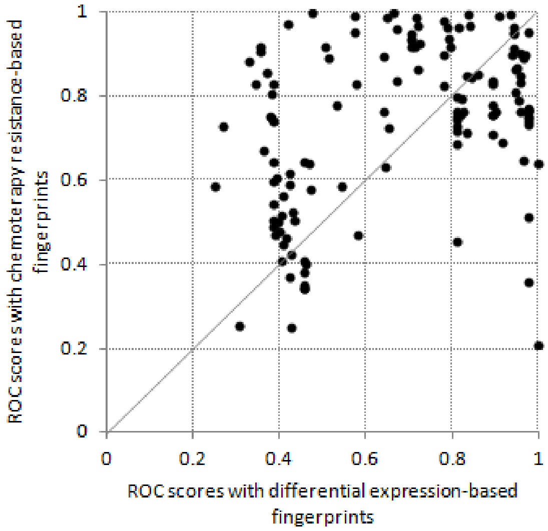

Fig. 4. Scatter plot for comparing chemotherapy resistance-based vs. differential expression-based representations.

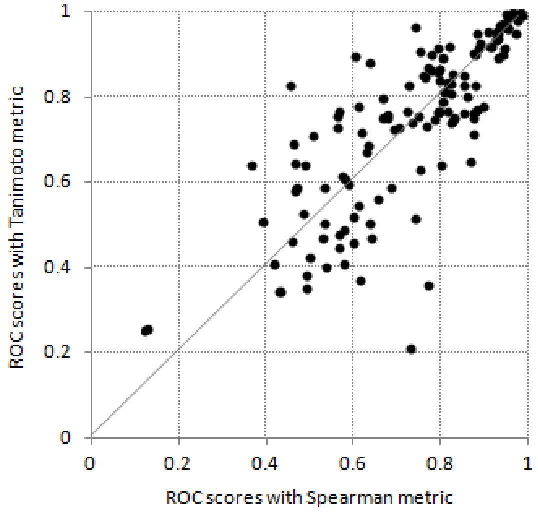

Fig. 5. Scatter plot for comparing Tanimoto vs. Spearman metrics.

## IV. CONCLUSION

This study offers a knowledge-based fingerprinting scheme for comparing miRNA microarray experiments. The fingerprints are built based on miRNA sets in relevant to their chemotherapy resistance. The empirical results obtained through an information retrieval setup shows that the new scheme can outperform traditional data-driven fingerprinting scheme based on differential expression. Furthermore, the representation is more interpretable from a biomedical point of view.

## ACKNOWLEDGMENT

This study was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under the Project 113E527.

## REFERENCES

[1] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets—update," Nucleic Acids Res., 41(Database issue):D991-5, 2013.

[2] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, and A. Brazma, "ArrayExpress—a public database of microarray experiments and gene expression profiles," Nucleic Acids Res., 35: D747–D750, 2007.

[3] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), v.2 n.1, p.1-19, 2006.

[4] P. B. Horton, L. Kiseleva, and W. Fujibuchi, "RaPiDS: an algorithm for rapid expression profile database search," International Conference on Genome Informatics, vol. 17, pp. 67-76, 2006.

[5] W. Fujibuchi, L. Kiseleva, T. Taniguchi, H. Harada, and P. Horton, "CellMontage: similar expression profile search server," Bioinformatics, vol. 23, pp. 3103-3104, 2007.

[6] R. Chen, R. Mallelwar, A. Thosar, S. Venkatasubrahmanyam, and A. J. Butte, "GeneChaser: Identifying all biological and clinical conditions in which genes of interest are differentially expressed," BMC Bioinformatics, vol. 9, p. 548, 2008.

[7] A. C. Gower, A. Spira, and M. E. Lenburg, "Discovering biological connections between experimental conditions based on common patterns of differential gene expression," BMC Bioinformatics, 12:381, 2011.

[8] J. M. Engreitz, A. A. Morgan, J. T. Dudley, R. Chen, R. Thathoo, R. B. Altman, and A. J. Butte, "Content-based microarray search using differential expression profiles," BMC Bioinformatics, 11:603, 2010.

[9] A. Hayran, H. Oğul, and E. E. Özkoç, "Content-based search in time-series microarray databases," 25th International Workshop on Database and Expert Systems Applications (DEXA), Munich, Germany (doi: 10.1109/DEXA.2014.33 ), 2014.

[10] J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, and S. Kaski, "Probabilistic retrieval and visualization of biologically relevant microarray experiments," Bioinformatics, 25:i145-153, 2009.

[11] E. Georgii, J. Salojärvi, M. Brosché, J. Kangasjärvi, and S. Kaski, "Targeted retrieval of gene expression measurements using regulatory models," Bioinformatics, 28(18):2349-2356, 2012.

[12] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," Cell, vol. 116, pp. 281–297, 2004.

[13] H. Oğul and M. S. Akkaya, "Data integration in functional analysis of microRNAs," Current Bioinformatics, vol. 6, 462-472, 2011.

[14] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gilette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," Proc. Natl. Acad. Sci. USA, vol. 102 no. 43, 15545-15550, 2005.

[15] E. Dai, Y. Lv, F. Meng, X. Yu, Y. Zhang, S. Wang, X. Liu, D. Liu, J. Wang, X. Li, and W. Jiang, "CREAM: a database for chemotherapy resistance-associated miRSNP," Cell Death and Disease, 5, e1272, 2014.