Combination Methods in Microarray Analysis

Han-Yu Chuang[1,*], Hongfang Liu[2], Fang-An Chen[3], Cheng-Yan Kao[1,*], and D. Frank Hsu[4,*]

1. *Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 106, Taiwan*
2. *Department of Information System, University of Maryland at Baltimore County, Baltimore, MD 21250, USA*
3. *Department of Otoloryngology, Division of Head and Neck Surgery, School of Medicine, New York University, New York, NY 10016, USA*
4. *Department of Computer and Information Science, Fordham University, New York, NY 10023, USA*

Email: r90002@csie.ntu.edu.tw, hsu@cis.fordham.edu

Abstract

Microarray technology and experiment can produce thousands or tens of thousands of gene expression measurement in a single cellular mRNA sample. Selecting a list of informative differential genes from these measurement data has been the central problem for microarray analysis. Many methods to identify informative genes have been proposed in the past. However, due to the complexity of biological systems, each proposed method seems to perform nicely in a particular data set or specific experiment. It remains a great challenge to come up with a selection method for a wider spectrum of experiments and a broader variety of data sets. In this paper, we take the approach of method combination using ***data fusion*** *and* ***rank-score graph*** *which have been used successfully in other application domains such as information retrieval, pattern recognition and tracking, and molecular similarity search. Our method combination is efficient and flexible and can be extended to become a general learning system for microarray gene expression analysis.*

1. Introduction

DNA microarrays are now capable of providing genome-wide measurements of gene expression across different conditions [1, 2]. One way to analyze these results is to determine observed significant differences and to select a list of most informative genes for further investigation [3]. Traditional statistical analyses such as pair t-test require replications of observation. However, since experiments using microarrays are costly and time consuming, most experiments are currently done without replications. Additionally, microarray data may contain a high level of noise due to its subtle nature. Therefore, the major challenge of microarray data analysis is to infer significant genes from the large number of measures given only a small number of samples.

Various approaches have been developed to solve this issue. However, it is often pointed out that there is no single method which is always the best in every study. The outcomes of different methods may differ substantially. This discordance causes difficulties in the interpretation of the data set. Moreover, it is unclear which method should be applied to new unknown data sets. However, the prevailing phenomenon implies that a gene is significantly worth further analyzing, if it is identified as an informative one in most commonly used methods. Thus, combining meaningful results from different methods seems to be a promising approach.

Recently, combination method and data fusion have been studied in a variety of different application domains such as information retrieval [9, 12, 18], pattern recognition [19], and molecular similarity searching [6]. There are basically at least two approaches to combine results from different methods. One is based on **result intersection**, where ones that are in results of all or most methods will be selected. The other is based on **result integration**, where score or rank of all methods integrate together using certain linear function or non-linear function to form a new score and ones with higher ranked measure will be selected. Method combination and data fusion fall under the second approach.

In this paper, we apply method combination and data fusion to analyze large data set from microarray experiments. We examine and combine 2 sets of 6 methods chosen from a total of 9 methods covering 5 parametric and 4 nonparametric gene selection methods for identifying informative genes: unpaired t-Test [8], paired t-Test, Fisher [16], Golub [7], SAM [17], TNoM [3], Wilcoxon rank sum test [15], Park [14] and WEPO

[*] Research supported by the project of Ministry of Economic Affairs, R.O.C., titled "Development of novel technology for diagnosis and treatment of angiogenesis-related disorders", under the contract "91-EC –17–A–19–S1–0016".

[4]. Each of the methods and combinations were evaluated by using precision measure on two publicly available dataset: Notterman *et al* [13] and Kuriakose *et al* [11] dataset. In each of the two sets of combination of 6 methods, we explore all the possible 63 = 2^6-1 combinations. We use the concept of a rank-score graph proposed in [9] to study the performance results of each of the two 63 combination experiments applied to the two data sets. Our experiments showed that method combination produces some very interesting patterns in the sense that certain combination does perform better than its individual part in microarray analyses. In many cases, genes selected by our combination approach are as informative as each of individual methods in the biological interpretation with the added advantages of efficiency and flexibility. Moreover, we observe that the combination of two heterogeneous and well-performing methods may achieve the best performance for microarray data. This phenomenon is in consistence with our previous research [5] and with those observed in other application domains (eg. [6, 12, 19]).

The rest of this paper is organized as follows: Section 2 discusses our combination methods. Section 3 describes the experiments (using 2 sets of combining 6 methods from the 9 gene selection methods and two datasets) for comparing individual and combination selection methods. Section 4 presents the result. Section 5 contains conclusions and suggests future research directions.

2. Method Combination

The proposed approach consists of two stages: the rank stage and the combination stage. In the first stage (rank stage), all genes in the interested dataset are ranked according to each of the selected methods. In this regard, each gene in the collection is assigned a score (which can be a measurement of variance, deviation, correlation, or probability) depending on a specific method. A ranking of the genes in the data set results from sorting them by their scores. The second stage (combination stage) is the process of combining the rank list obtained from the first stage.

When combining 2 or more ranked lists, we use the average rank combination. In other words, suppose we have N ranked lists R_i, $i = 1,2,...,N$ and M genes g_j, $j = 1,2,...,M$. For each gene g_j, we calculate a score which is the combined average ranking of that gene, $f(g_j) = \sum_{i=1}^{N} R_i^{-1}(g_j)/N$. Sorting $f(g_j)$ into ascending order gives rise to a new ranking $R*$ which is the resulting combined ranks. See [6, 18] for more details on combination method and data fusion.

3. Experiments

In the experiments, the combination method was applied to combine different gene selection methods. We then use a metric called **precision** to evaluate the performance on two different available datasets: Notterman *et al* [13] and Kuriakose *et al* [11] dataset. It measures how informative the selected genes are with respect to their established biological interpretations. We used the known informative genes in these datasets that have previously been proposed and confirmed to some extent by these authors to compute precision. In the following, we describe the datasets, precision calculation and gene selection methods in more detail.

3.1. Data Sets

In the following, we describe the two datasets used in the experiments: DS1: Notterman *et al* [13], and DS2: Kuriakose *et al* [11].

DS1: **Adenocarcinoma data set**;

The expression profile associated with this data set was collected by Notterman *et al* [13]. The Notterman team obtained 18 paired colon adenocarcinoma normal tissue samples from the Cooperative Human Tissue Network. The experiment was performed with the Human 6500 GeneChip Set (Affymetrix oligonucleotide array). The data set consists of 7457 genes and 18 paired samples, in which 18 are labeled "carcinoma" and 18 are labeled "normal". Additionally, Notterman *et al.* applied 4-fold relative expression to choose informative genes and 66 genes (1.78% of those detected) had been picked with significant difference between tumor tissue and the normal samples. 11 of them were confirmed by reverse transcription-PCR (RT-PCR), which were used to measure the precision of each gene selection method and combination methods.

DS2: **Head and Neck Squamous Cell Carcinoma (HNSCC) data set**;

In the study Kuriakose *et al* [11] of head and neck squamous cell carcinoma (HNSCC), RNA's extracted from 22 paired samples of HNSCC and normal tissue from the same clones were hybridized to the Affymetrics U95A chip. Forty two differentially expressed probe sets (18 up-regulated and 24 down-regulated) were selected for further validation by hierarchical clustering, multiple probe-set concordance, target-submit agreement, and RT-PCR analysis.

3.2. Precision of known informative genes

One reasonable way to evaluate a gene selection method is to measure the **precision** of known informative genes that were previously confirmed to be among the top selected genes. The precision we used is defined

1087-4089/04 $ 20.00 © 2004 IEEE

as $P(R,G)=\frac{\sum_{i}^{|G|}\frac{i}{|R_i|}}{|G|}$, where R is a ranking list, R_i is the top n elements in the ranking which include i genes in G, and G is a list of known informative genes.

3.3. Gene selection methods

In this paper, the following selection methods to identify information genes are considered: (A) unpaired t-test [8], (B) paired t-Test, (C) Fisher [16], (D) Golub *et al* [7], (E) SAM [17], (F) TNoM [3], (G) Wilcoxon rank sum test [15], (H) Park *et al* [14], and (I) WEPO [4]. The first 5 are of parametric nature while the last 4 are non-parametric.

In our study, we use two sets of 6 selection methods **M1**: (A), (C), (D), (F), (G), and (I), and **M2**: (A), (B), (E), (G), (H) and (I). Each of the M1 and M2 consists of both parametric and non-parametric selection methods.

4. Results

For each of the 2 data sets DS1 and DS2, we considered 2 sets M1 and M2 of 6 selection methods mentioned above to compute scores for each gene. Three experiments E1, E2, and E3 (**E1: DS1+M1, E2: DS2+M1, and E3: DS2+M2**) are conducted. In each experiment, all possible 63 combinations are considered and rank-score graphs are drawn.

Figure 1 demonstrates the precision of rankings for each individual and combination methods on the two datasets. Table 1 provides the detail of the corresponding rankings coded in Figure 1 for each experiment. The rankings are arranged with their number of methods combined and then the goodness of the corresponding precision. Figure 2 exhibits the rank-score graph of each individual method in these experiments.

In Figure 1, we see that the non-parametric methods almost outperform the parametric ones in each experiment. However, as we mentioned above, there is no “super star” methods. For example, Wilcoxon (G) is the best one for dataset DS1 when using precision measure and performs better than TNoM (F), but the reverse occurs for dataset DS2.

Nevertheless, Figure 1 displays a more exciting result that the performance of a combination of several methods is better than the worst case of each individual. As the number of methods combined increases, the minimum precision of rankings with the same number of methods combined increases, but the maximum drops when the number used is greater than two. The precision reaches the peak, which is higher than the one of any individual method, when combining two proper methods in each experiment. Moreover, the combination methods perform better if the individuals used are better in Table 2. A good instance can be found in the experiment shown in Figure 1 (c), where Wilcoxon (G), Park (H) and WEPO (I) are the top 1, 2, and 3 respectively when we use individual methods to rank. When combining two methods, the combination of any two of method G, H, and I are still the top ones. The same situation holds even enlarging the number of methods combined. The combination of G, H, and I is the best when combining three methods. Then, SAM (E) is included when combining four methods, and it is the top 4 on its own. Unpaired t-Test (A) is the next one to be added when combining five. Finally, paired t-Test (B) is added, since its performance is the worst. Similar phenomenon happens in the other two experiments.

Furthermore, by observing the rank-score graph in Figure 2, we found that the combination of two heterogeneous and well-performing methods will achieve better performance. Take dataset DS1 as an example. The combination of WEPO (I) and TNoM (F) resulted in the highest precision. The combination of WEPO (I) and Wilcoxon (G) is the next, and then follows the one of TNoM (F) and Wilcoxon (G). In Figure 2, the distance between the curve of I and of F/G is longer than that between F and G. In other words, method I is more different to F than F to G on the rank-score graph. The power of heterogeneous combination is also observed on each experiment.

5. Discussion and Conclusion

We have demonstrated that the combination method is a robust and efficient approach for microarray data analyses. From Figure 1, it is clear that no single gene selection method, at least until now, performs effectively across different data sets (and experiments) in different application domains. Results obtained in this paper using the combination methods shows that a combination approach almost always performs better than the less efficient individual, and in many cases, better than both in the cases of 2-combination. More significantly, the combination of two heterogeneous and well-performing methods achieves the best performance than any individual one in all tested datasets. All of this evidence indicates that method combination is highly likely to be a viable approach for microarray gene expression analysis on any dataset.

There are several other advantages of our combination methods for microarray data analyses. We mention only two characteristics: efficiency and flexibility. Sorting a list of n genes with assigned scores takes n*log$n$ steps. Moreover, combination of m rank lists should take no more than m*n*log$n$ steps. Calculation in the method combination becomes simple and easy to understand.

Selection of efficient and effective combinations would facilitate fast process and operation.

The proposed method is independent of each gene selection method. Individual selection method can be both parametric and nonparametric. Combination methods can use rank or score combination. In this paper we only use rank combination. Moreover, rank combination allows individuals the choice of using consensus building or voting, while score combination facilitates the options of using various linear, non-linear, or weighted combinations. Compared to other methods such as clustering association, or self-organized maps, method combination is more flexible as the outputs of both stages of the combination process are rank lists for the collection of genes.

The proposed method can be adapted to different application domains which may call for different combination algorithms. One of our long-term goals is to construct a system which can learn from the environments and phenomena in its application domain, and then evolve to become a more intelligent expert system in that particular domain.

In this paper, we described the method combination and have taken up our investigation using average linear combination of rankings to combine two or more gene ranking methods. Future work will explore other ways to combine different gene ranking methods. In other direction but closely related to the current study, Hsu and Palumbo [10] consider the rank space as a Cayley graph and examine how data fusion and method combination work in that graphical model.

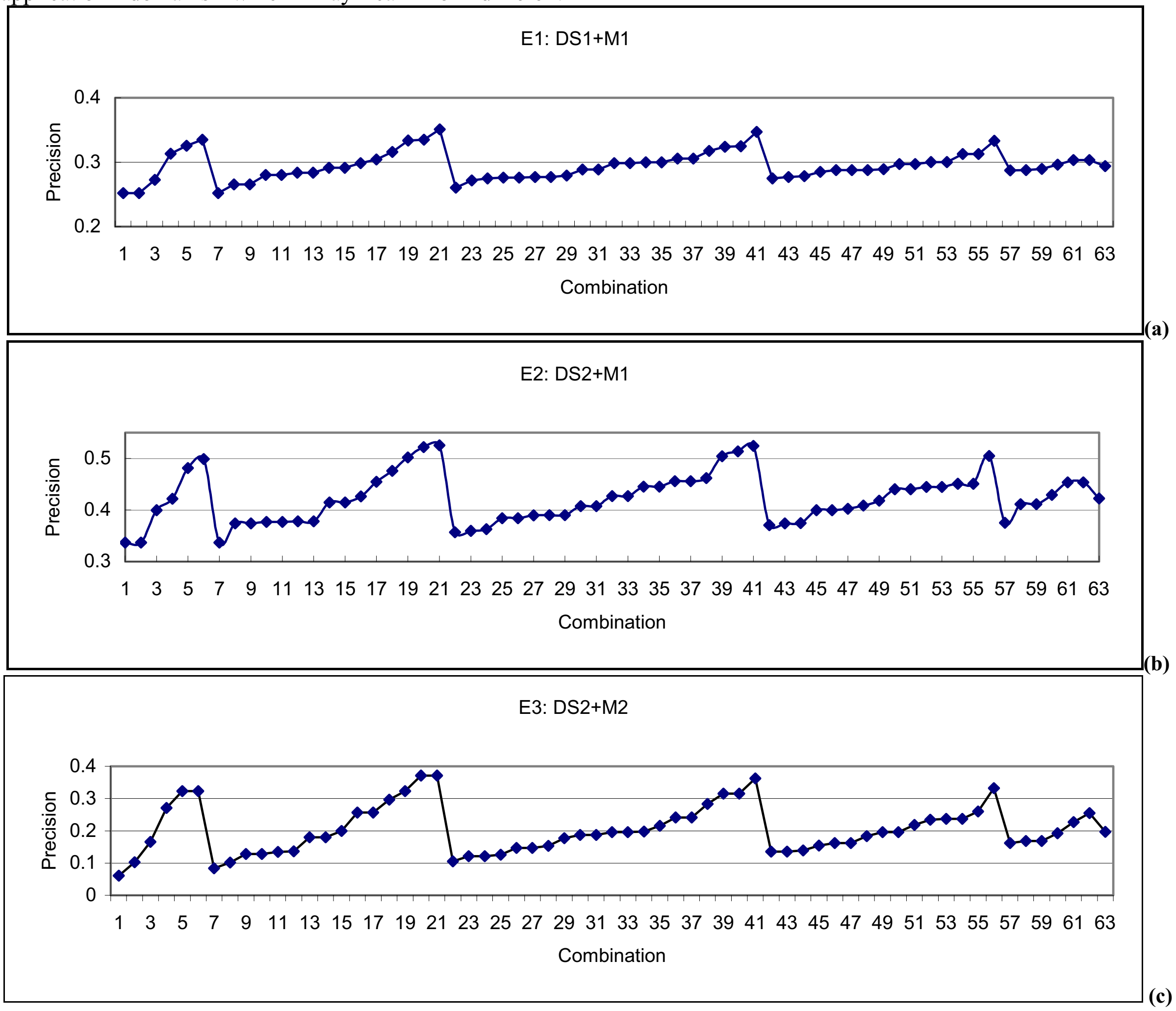
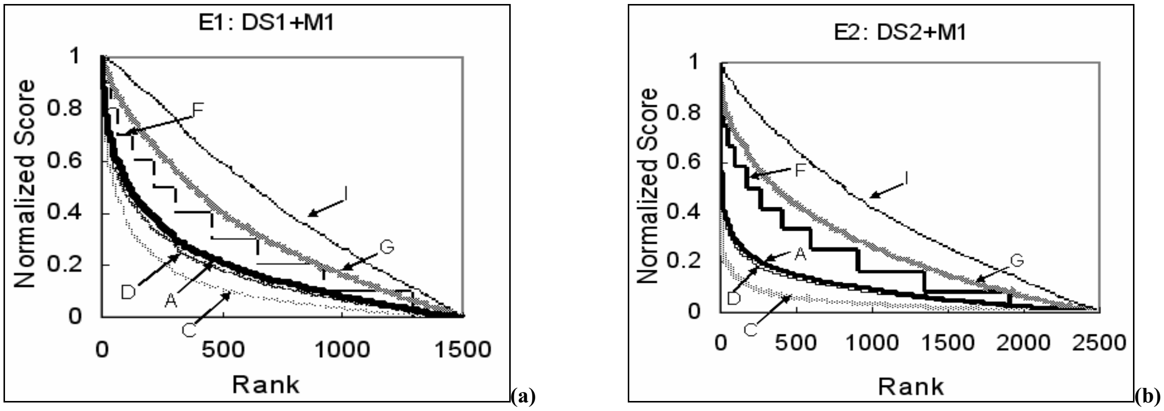


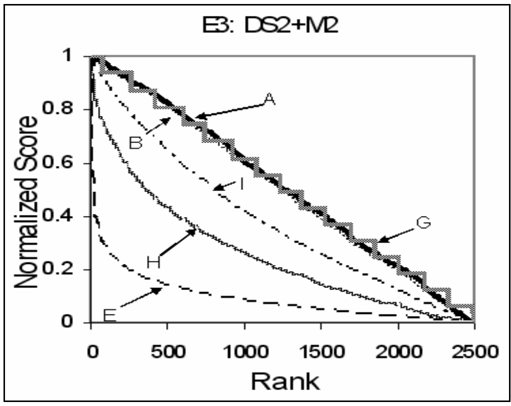
Figure 1. Precision v.s. combination in 3 experiments, E1, E2, and E3. **(a)** E1 = DS1 + M1, **(b)** E2 = DS2 + M1, and **(c)** E3 = DS2 + M2.

Table 1. The corresponding combination of methods coded in Figure 1. The columns record results from the experiment described in Figure 1 (a), (b), and (c), respectively.

Index	E1:DS1+M1	E2:DS2+M1	E3:DS2+M2	Index	E1:DS1+M1	E2:DS2+M1	E3:DS2+M2
1	A	A	B	**33**	CFG	FCD	IGB
2	C	C	A	**34**	AGI	GFA	IEA
3	D	I	E	**35**	CGI	GFC	GAH
4	F	D	I	**36**	AFI	IFA	IGA
5	I	G	H	**37**	CFI	IFC	IAH
6	G	F	G	**38**	DFG	IGD	GEH
7	AC	AC	BA	**39**	DFI	GFD	IEH
8	AD	IC	BE	**40**	DGI	IFD	IGE
9	CD	IA	BH	**41**	FGI	IGF	IGH
10	AF	CD	GB	**42**	ACDF	IGAC	BEAH
11	CF	AD	EA	**43**	ACDG	GACD	GBEA
12	AG	GC	IB	**44**	ACDI	IACD	IBEA
13	CG	GA	AH	**45**	ACFG	IGAD	GBAH
14	AI	FC	GA	**46**	ACFI	IGCD	IGBA
15	CI	FA	IA	**47**	ADFG	FACD	IBAH
16	DF	GD	EH	**48**	CDFG	GFAC	GBEH
17	DG	ID	GE	**49**	ACGI	IFAC	IBEH
18	DI	IG	IE	**50**	ADGI	GFAD	IGBE
19	FG	FD	GH	**51**	CDGI	GFCD	GEAH
20	GI	IF	IH	**52**	ADFI	IFAD	IGBH
21	FI	GF	IG	**53**	CDFI	IFCD	IGEA
22	ACD	GAC	BEA	**54**	AFGI	IGFA	IEAH
23	FAC	ACD	BAH	**55**	CFGI	IGFC	IGAH
24	ACG	IAC	GBA	**56**	DFGI	IGFD	IGEH
25	ADF	GAD	IBA	**57**	ACDFI	IGACD	GBEAH
26	CDF	GCD	BEH	**58**	ACDGI	IFACD	IBEAH
27	ADG	FAC	GBE	**59**	ACDFG	GFACD	IGBEA
28	CDG	IAD	IBE	**60**	ACFGI	IGFAC	IGBAH
29	ACI	ICD	GBH	**61**	ADFGI	IGFAD	IGBEH
30	ADI	IGA	EAH	**62**	CDFGI	IGFCD	IGEAH
31	CDI	IGC	GEA	**63**	ACDFGI	IGFACD	IGBEAH
32	ACG	FAD	IBH				



(A) unpaired t-test, (B) paired t-Test, (C) Fisher, (D) Golub *et al*, (E) SAM, (F) TNoM, (G) Wilcoxon rank sum test, (H) Park *et al*, and (I) WEPO.



(c)

Figure 3. rank-score graphs of each gene selection method in three experiments. The denotations of (a), (b) and (c) are the same as the ones of Figure 1.

References

[1] A.A. Alizadeh *et al.*, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", *Nature*, 2000, vol. **403**, pp.503-511.

[2] U. Alon *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proceedings of the National Academy of Sciences*, 1999, vol. **96**, pp. 6745-6750.

[3] A. Ben-Dor *et al.*, "Tissue Classification with Gene Expression Profiles", *Journal of Computational Biology*, 2000, vol. **7**, pp. 559-583.

[4] H.Y. Chuang, H.K. Tsai, Y.F. Tsai and C.Y. Kao, "Ranking genes for discriminability on microarray data", *Journal of Information Science and Engineering*, 2003, vol. **19**, pp. 953-966.

[5] H.Y. Chuang *et al.*, "Identifying significant genes from microarray data", *IEEE Bioinformatics and Bioengineering'04*, 2004.

[6] C.M.R. Ginn, P. Willett, and J. Bradshaw, "Combination of Molecular Similarity Measures Using Data Fusion", *Perspectives in Drug Discovery and Design*, 2000, vol. **20**, pp. 1-16.

[7] T. R. Golub *et al*., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", *Science*, 1999, vol. **286**, pp. 531-537.

[8] I. Hedenfalk *et al.*, "Gene-expression profiles in hereditary breast cancer", *New England J. Med.*, 2001, vol. **8**, pp. 344-539.

[9] D.F. Hsu, J. Shapiro, and I. Taksa, "Methods of Data Fusion in Information Retreival: Rank vs. Score Combination", *DIMACS Technical Report 58*, 2002.

[10] D. F. Hsu, and A. Palumbo, "A study of data fusion in Cayley graph $G(S_n, P_n)$", *Proceedings of I-SPAN'04, IEEE CS Press*, 2004.

[11] M.A. Kuriakose *et al.*, "Selection and Validation of differentially expressed genes in head and neck cancer", *manuscript*.

[12] K.B. Ng *et al.*, "Predicating the effectiveness of Naïve Data Fusion on the basis of system characteristics", *JASIS*, 2000, vol. **51**, pp. 1177-1189.

[13] D.A. Notterman, U. Alon, A.J. Sierk, and A.J. Levine, "Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays", *Cancer Research*, 2001, vol. **61**, pp. 3124–3130.

[14] P.J. Park, M. Pagano, and M. Bonetti, "A Nonparametric Scoring Algorithm for Identifying Informative Genes from Microarray Data", *Pacific Symposium on Biocomputing*, 2001, vol. **6**, pp. 52-63.

[15] R. Pollack *et al*., "Genome-wide analysis of DNA copy-number changes using cDNA microarrays", *Nature Genetics*, 1999, vol. **23**, pp. 41-46.

[16] S. F. Terrence *et al.*, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data", *Bioinformatics*, 2000, vol. **16**, pp. 906-914.

[17] V.G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response", *Proc Natl Acad Sci U S A*, 2001, vol. 98(9), pp. 5116-5121.

[18] C.C. Vogt, and G.W. Cotrell, "Fusion via a linear combination of scores", *Info. Ret.*, 1999, vol. **1**, pp. 151-172.

[19] L. Xu, A. Krzyzak, and C.Y. Suen, "Method of Combining Multiple Classifiers and their Application to Handwriting Recognition", *IEEE Trans SMC*, 1992, vol. **22**, pp. 418-435.