Extraction of Informative Genes from Integrated Microarray Data

Dongwan Hong[1], Jongkeun Lee[1], Sangkyoon Hong[1], Jeehee Yoon[1], and Sanghyun Park[2]

[1] Division of Information and Communication Engineering, Hallym University, Okcheon-Dong, Chuncheon, 200-702, Korea
{dwhong,jeikei,kyoons,jhyoon}@hallym.ac.kr
[2] Department of Computer Science, Yonsei University, Shinchon-Dong, Seodaemun-Gu, Seoul, 120-749, Korea
sanghyun@cs.yonsei.ac.kr

Abstract. We have recently proposed a rank-based approach as a new microarray data integration method. The rank-based approach, which converts the expression value of each sample into a rank value within the sample, enables us to directly integrate samples generated by different laboratories and microarray technologies. In this study, we show that a non-parametric scoring method can be efficiently employed for the rank-based data, and informative genes can be effectively extracted from the integrated rank-based data. To verify the statistical significance of the scoring results from the rank-based data, we compared the distribution of the score statistics to a set of distributions obtained from the randomly column-permuted data. We also validate our methods with experimental study using publicly available prostate microarray data. We compared the informative genes extracted from each individual data to the informative genes extracted from the integrated data. The results show that we can extract important prostate marker genes by directly integrating inter-study microarray data, which are missed in either single analysis.

Keywords: Informative genes selection, microarray data integration, prostate cancer, statistical significance verification.

1 Introduction

Microarray experiments enable scientists to obtain a tremendous amount of gene expression data at one time, so they are effectively used in identifying the phenotypes of diseases. In general, increasing sample size is quite desirable for more reliable and valid results. However, microarray experiments are still cost-expensive, so it is hard in reality to obtain experimental results based on a large number of samples. Thus, the experimental results from different investigations with the same research goals are somewhat different and usually contain many errors.

With the rapid accumulation of microarray data, it is of great interest and challenge to integrate inter-study microarray data to increase sample size, which leads to better experimental results. In our earlier work [1], we proposed a new

A. An et al. (Eds.): ISMIS 2008, LNAI 4994, pp. 618–627, 2008.

microarray integration method using a rank-based approach. The rank-based approach, which simply converts the expression value of each sample into a rank value within the sample, enables us directly integrate samples generated by different laboratories and microarray technologies.

In this study, we show that a non-parametric scoring method can be efficiently employed for the rank-based data, and informative genes can be effectively extracted from the integrated rank-based data. As a non-parametric scoring method, Park's method [2] is employed. However, as the scoring method compares the sample values of each gene to calculate a score, it may give slightly different score results when it is applied to the rank-based data and the actual expression value data, respectively. Here we verify the statistical significance of the scoring result from the rank-based data. We compared the distribution of the score statistics to a set of distributions which is obtained from the randomly column-permuted data. Golub's leukemia data [3] was tested, and its result was significant with the p-value of 0.0005 for the rank-based data. Then we compared the informative genes extracted from the rank-based data to the informative genes extracted from the actual expression value data. To exemplify the effectiveness of our integration method, we used three publicly available prostate microarray data. We compared the informative genes extracted from each individual data to the informative genes extracted from the integrated data. The results reveal that important marker genes are selected from the integrated data, which are missed from a single data.

2 Related Works

Experimental microarray data are organized as matrices where rows represent genes and columns represent samples. However, even when considering the microarray data with the same research goals, differences in platforms, protocols, set of genes, and scales of gene expression values lead to difficulties in integrating microarray data across experiments.

To integrate microarray data, the typical methods include *meta-analysis method* [4], *normalization and transformation method* [5, 6], and *rank-based approach* [1]. Instead of comparing microarray expression values from individual experiments, *meta-analysis method* combines the results of individual experiments by using statistical technique. However, there are many cases where the individual experimental results are not reliable due to the small sample size. So the integration of these results may bring an even worse analysis. *Normalization and transformation method* transforms the gene expression values of individual experimental data into a common scale, and then integrates inter-study data [5]. A classical method is the z-score transformation [6], which normalizes the expression values with the mean and standard deviation of each sample. Statistical tests, such as fold ratio, z ratio/test [6], and t statistical test, can be applied directly to the normalized data for predicting significant changes in gene expressions. However, there is still no consensus on the best method to perform data normalization [7]. *Rank-based approach* converts the expression value of each

sample into a rank value [1]. In statistical area, this method has been used as a noise reduction method [8]. Xu *et al.* [7] proposed a new classification method (top-scoring pair classifier) to select maker genes from the integrated rank-based data. However this method is only based on comparing relative expression values within each sample.

One of the difficulties in analyzing microarray data is the high dimensionality due to a large number of genes. However, only a small fraction of genes is informative for predicting significant changes in gene expressions. Currently, various methods are being presented to select informative genes precisely and effectively. Typically, informative genes are selected according to a test statistics. A *parametric method* assumes a statistical model representing the data, such as the t-statistics [9], Fisher [10], and Golub's method [3]. There are *non-parametric methods* such as TNom [11], Wilcoxon rank sum [12], and Park's method [2]. These methods define a minimum boundary and calculate the distance from the boundary as the score. On the other hand, when the gene is considered as a feature, the rank-based feature selection method [13] can be used. This method measures the significance of features and then ranks them. In this approach, the popular methods are Information Gain [13], Relief-F [14], and the method using Kendall's Correlation Coefficient [15]. However, all these methods use the gene expression values of each gene, and there is no consideration regarding the integration and normalization of the microarray data.

3 Methods

3.1 A Rank-Based Microarray Data Integration

The integration procedure of microarray data is shown as follows. First, only the experimental data of common genes are extracted from the individual microarray data, which has the same research goals. Then the expression value of each sample in each experiment is converted to a rank value within the sample. Once the expression values are changed to rank values, the integration of samples from different experiments becomes feasible. This method is simple and useful for integrating a large number of microarray samples without the need to perform any normalization. Hereafter, for simplicity, we call experimental data using the original expression values *raw data*, and experimental data using the rank values *rank data*. As the integrated data contains only the rank values rather than the actual expression values, there may be a slight loss of information. However, too big or too small expression values of each sample can be noises, which may give a negative effect on extracting informative genes. In return, we gain the robustness to external factors, such as noises.

3.2 Informative Genes Selection Method

Park's non-parametric scoring method [2] is extended and applied to the integrated microarray data. Park's method, which is proposed for a single microarray

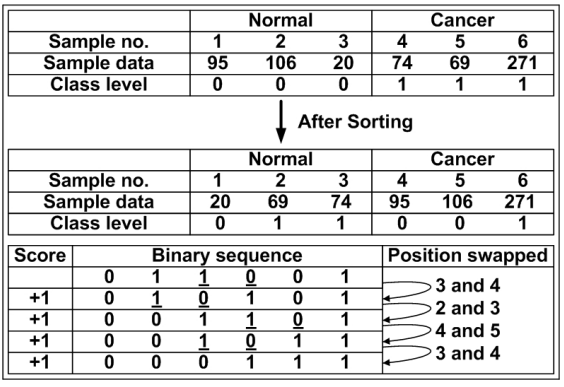


Fig. 1. An example of gene scoring

data, builds a binary sequence for a gene and calculates a score measuring how differently the genes are expressed in the two class groups, by using Kendall's Correlation Coefficient [15].

Let us explain the scoring method by using an example. Fig. 1 shows how to calculate the score of a gene with six sample data of 95, 106, 20, 74, 69, and 271. Here, we assume that each sample data represents the rank value. In this figure, samples 1, 2, and 3 represent normal class and samples 4, 5, and 6 represent cancer class. First, class label 0 is assigned the normal sample and class label 1 is assigned to the cancer sample, to obtain an initial binary sequence S = 000111, which represents the class labels of the gene data. Next, the sample data are sorted in ascending order along with the class labels. Thus, the sorted binary sequence T = 011001 is obtained, and it represents the class labels of the sorted gene data. A distance between S and T is used as the score of the gene. The distance is defined as the minimum number of swaps of neighboring 0 and 1 which is necessary to transform the sorted binary sequence into the initial binary sequence. Fig. 1 shows the process in which T = 011001 is transformed into S = 000111, and result score of 4. Suppose the number of normal samples is n_1 and the number of cancer samples is n_2, then the score ranges from 0 to $n_1 \times n_2$. Both low and high scores indicate differentially expressed genes, which are selected as informative genes.

3.3 Example

Next we illustrate data integration and informative gene selection procedures using an example. Let us consider two data, Data(A) and Data(B), which are generated independently but have the same research goals. As shown in Fig. 2, the scale of the expression values for each data is quite different and a direct integration is inappropriate. First we convert all expression values into ranks within each sample, and obtain Data(A)′ and Data(B)′ of *rank data*. As explained in Section 3.2, the score refers to the minimum number of swaps of neighboring

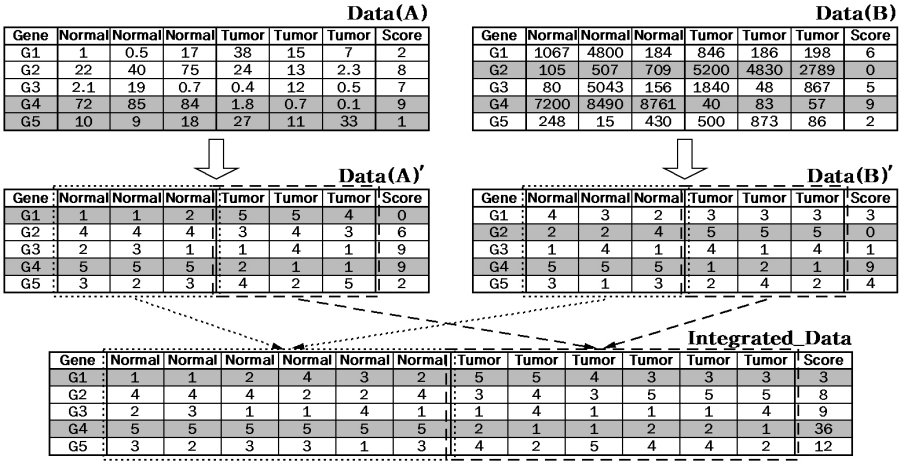


Fig. 2. An example of microarray data integration and informative gene selection

digits necessary to arrive at perfect splitting, with all the 0's on the left and all the 1's on the right. If only two genes are selected as informative genes from each data, the genes with the lowest and highest score are selected. For instance, "G5 and G4" and "G2 and G4" are selected from Data(A) and Data(B), respectively, and at the same time "G1 and G4" and "G2 and G4" are selected from Data(A)′ and Data(B)′, respectively. Notice that the extracted informative genes from *raw data* and *rank data* may be different. Next, Data(A)′ and Data(B)′ are merged and finally "G1 and G4" are selected from the Integrated_Data as informative genes.

3.4 Significance Test

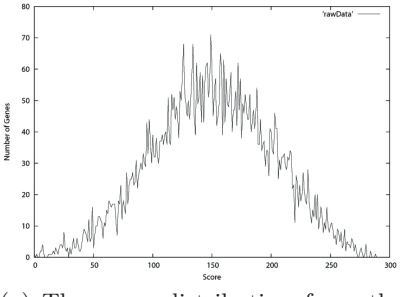
A permutation test is performed to test the significance of gene scoring result for the *rank data*. We generate a random permutation of entire columns, keeping all the rank values for each samples together. A p-value is then computed by comparing the distribution obtained from the original data to the set of distributions obtained from the randomly permuted data. To calculate a p-value, a cumulative function S_i of (Eq. 1) is used. For the comparison, we use the same function which is given in [2]. S_i is the measure of how much the i-th score distribution is different from the average of all the other score distributions. Here, f_i^* represents the average of all distributions except for the score distribution of the i-th column-permuted data, and M represents the number of column-permuted data. S_0 represents the difference between the score distribution of original data and the average of the score distributions of other column-permuted data. A significance probability $P(S_i \geq S_0)$ is now calculated. Here, the requirement of $i = 1, \cdots, M$ is met. If the p-value is smaller than the significance level, we assume that the gene scoring result is significant.

$$S_i = \sum_{j=0}^{n_1 n_2} (f_i(x_j) - f_i^*(x_j))^2, \quad i = 1, \cdots, M$$
$$f_i^*(x_j) = \frac{1}{M-1} \sum_{k=1, k \neq i}^{M} f_k(x_j) \tag{1}$$

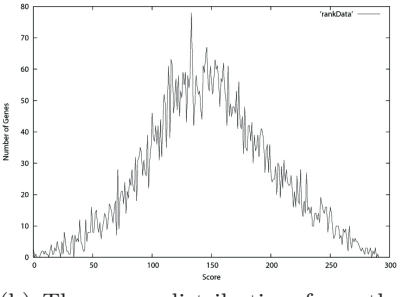
4 Results

4.1 Significance Test for Scoring Results

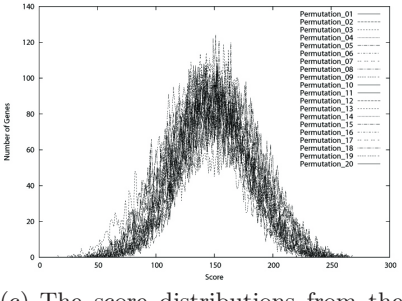
We applied the non-parametric scoring method described in Section 3.2 to the Golub's leukemia data [3]. Golub's data contains 38 bone marrow samples obtained from acute leukemia patients. 27 samples are from ALL class and 11 samples are from AML class. High-density oligonucleotide microarrays (produced by Affymetrix) containing 7129 probes for 6817 human genes are used. As stated in Section 3.4, a permutation test was performed for two data, *raw data* and



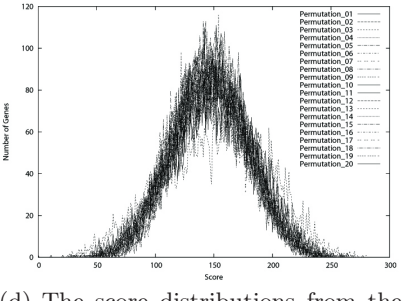
(a) The score distribution from the *raw data*



(b) The score distribution from the *rank data*



(c) The score distributions from the column-permuted *raw data*



(d) The score distributions from the column-permuted *rank data*

Fig. 3. Comparison of gene score distributions using *raw data* and *rank data* (Golub's data [3] is used)

rank data. We performed 10,000 permutations. Fig. 3 shows the score distributions from the original data and a set of randomly column-permuted data. Fig. 3-(a) shows the score distribution from the *raw data* where the score of each gene is calculated by using the gene expression values of samples. And Fig. 3-(b) shows the score distribution from the *rank data* where the score of each gene is calculated by using the rank values of samples. The results show that the two score distributions have very similar shapes and especially show heavier tails as expected, indicating many genes are differentially expressed in the two classes. Fig. 3-(c) shows a set of score distributions from the column-permuted *raw data*, and Fig. 3-(d) shows a set of score distributions from the column-permuted *rank data*. Here, only twenty score distributions are plotted for each case. The results show that the score distributions from the *raw/rank data* are more spread out with heavy tails, while the score distributions from the column-permuted *raw/rank data* are relatively concentrated with smaller variances. Based on the S_i values explained in Section 3.4, p=0.0005 was obtained from the *rank data*. Also, p=0.0053 was obtained from the *raw data*, which is consistent with the p-value reported by Park [2]. This result verifies our expectation that the scoring result from the *rank data* is statistically significant.

Next, we compared informative genes extracted from *rank data* to the informative genes from *raw data*. When top 1% of the genes are selected as informative genes, about 70% of the informative genes overlap each other. When top 5% of the genes are selected, about 76% of them overlap each other. Also, they include all 50 informative genes that were chosen by Golub's experiment [3].

4.2 Selection of Informative Genes from Integrated Data

To investigate whether more accurate informative genes can be selected from integrated data, the scoring method has been applied to the individual and integrated *rank data*. We used three prostate cancer microarray data which are publicly available. The platform of these data is Affymatrix HG_95Av2. Each data will be represented as an abbreviation of the first author of the paper, like as LaTulippe [16], Welsh [17], and Singh [18]. LaTulippe consists of 3 normal samples, 14 primary prostate cancer samples, and 9 metastatic prostate cancer samples. Welsh consists of 9 normal samples and 25 cancer samples, and Singh consists of 50 normal samples and 52 cancer samples.

As mentioned previously, we assume that larger sample size enables to extract more statistically significant genes. Also, we can expect a better statistical result when the number of test samples is almost equal to that of control samples. Singh's sample size is relatively larger than both LaTulippe's and Welsh's, and the number of its test samples is almost same as that of its control samples.

We merge LaTulippe and Welsh, using the 12600 common probe sets, to form an integrated data of increasing sample size. Here, (LaTulippe+Welsh) represents the integrated data resulting from the merging of LaTulippe and Welsh data. The scoring method is applied to the individual and integrated data, and top 1% of genes are selected as informative genes for each data. The selected informative genes are listed in Table 1.

Table 1. Comparison of informative genes extracted from LaTulippe, Welsh, Singh and (LaTulippe+Welsh)

Ranking	LaTulippe	Welsh	Singh	LaTulippe+Welsh	Ranking	LaTulippe	Welsh	Singh	LaTulippe+Welsh
1	FCGRT	MYL6	**HPN**	**ANGPT1**	64	KIAA0303	**CALM1**	AMACR	**HSPD1**
2	SOX5	CLU	**PTGDS**	**CALM1**	65	MEIS2	GJA1	**NME1**	MYL6
3	LCAT	PSIP2	**NELL2**	LPIN1	66	CDC5L	MYH11	CLDN3	TPM1
4	PNMT	**ANGPT1**	TRG@	SVIL	67	Rab11-FIP2	GSN	XBP1	SYP
5	IGF2	**DSCR1L1**	ANXA2P3	**COL4A6**	68	TRO	**GSTP1**	KIAA0977	ATP2A2
6	CYP3A5	FZD7	**HSPD1**	MEIS2	69	HSD11B1	RBPMS	SLC25A6	TACC1
7	CYP3A5	CBX7	ANXA2	CBX7	70	LDB2	TPM1	RPL12	KIAA0992
8	MDM1	KIAA0469	KLK3	LAPTM4A	71	NDN	MEIS2	GFPT1	JAM3
9	ELKS	FTO	PLA2G7	PRNP	72	ARHGEF4	CNN1	TNA	TRIP6
10	COL13A1	DMPK	PDLIM5	MYLK	73	FTO	TGFB1I1	TACSTD2	SH3GLB1
11	**ANGPT1**	RRAS	**STAC**	**GSTP1**	74	NRLN1	KANK	KIAA1157	STOM
12	CHRNA7	TRIP6	TMSNB	CLIPR-59	75	DOCK1	PMP22	AKR1B1	**HPN**
13	LDOC1	-	XBP1	GASP	76	SLK	ATP2A2	RBP1	LOC171220
14	RE2	PPP3CB	**DF**	NRLN1	77	DKFZP586A0522	CALD1	MYO6	CLU
15	GPR161	SVIL	**SPON1**	BART1	78	LAPTM4A	FLJ21174	MAP1GA1	**STAC**
16	CX3CR1	SRF	RGS10	SNX1	79	SRI	FLNA	MEG3	GNAZ
17	KIAA0888	DES	GUCY1A3	SPARCL1	80	MEIS1	CLIC4	PDIR	SLC2A5
18	LOC151584	PPP1R12B	**NME2**	**DAT1**	81	PGCP	GATM	**SC65**	FLNC
19	KIAA0534	KIAA0992	THBS4	**C7orf24**	82	TCF12	COL4A2	IL11RA	TBL1X
20	APEG1	OPTN	-	RBPMS	83	CDC42EP3	**DAT1**	**GSTP1**	ROR2
21	IGSF1	FLNA	RPL13A	GJA1	84	MAPRE1	RBPMS	ZNF146	WFS1
22	AIP1	BPAG1	SLC25A6	HOXC6	85	PPAP2B	NIFU	PHYHIP	DMPK
23	MKLN1	MYLK	**CALM1**	KIAA0725	86	STAT5B	ENO2	HOMER2	WFDC2
24	KIAA0980	**GSTP1**	SIM2	TCF8	87	SUSP1	ATP2B4	HSPA8	LDB3
25	TRO	TAZ	**DAT1**	**ANGPT1**	88	SLC2A5	LTBP1	MTHFD2	SEC23A
26	CHS1	PLS3	TSPAN-1	EDNRA	89	**SPON1**	GSTM5	**PYCR1**	SPG20
27	KPNA3	COL6A2	**C7orf24**	RBPMS	90	SYNGR1	-	ATP2C1	MXRA7
28	SNCG	FLNC	FBP1	TPM1	91	CLIPR-59	TPM2	LOC285843	ITPR1
29	D2LIC	BC008967	TACSTD1	TGFB1I1	92	VCL	TPM2	**NME1**	KRT18
30	TCF21	SDFR1	**COL4A6**	FNBP1	93	DHX38	TPM1	CYP1B1	**PYCR1**
31	ALDH1A2	CAV1	RPLP0	PPAP2B	94	BTBD3	TPM1	**PTGDS**	**NME2**
32	GSTM1	DPYSL3	ITSN1	CCND2	95	DKFZP434D1335	GASP	TRAF4	CNN1
33	ACTC	PRNP	P4HB	FEZ1	96	CETN2	ITGA8	SAT	RIMS3
34	MAP1LC3B	PLEKHC1	AGR2	DKFZP564M1416	97	MGC35048	DFNA5	ODC1	EMILIN1
35	PBX1	CCND2	-	MEIS3	98	GPRC5B	SMTN	EEF1G	FGFR1
36	SSX2IP	LMOD1	EPB41L3	CSRP1	99	**CALM1**	TEAD3	-	MYH11
37	C22orf2	RBPMS	TU3A	OPTN	100	C14orf132	BCMP1	S100A4	**NME2**
38	EFS	LPIN1	FOLH1	**GSTP1**	101	IGF1	ANXA6	TGFB3	SYNGR2
39	WFDC2	DSTN	G6PD	-	102	MADH6	CDC10	RPS10	TPM2
40	SSA2	TUBA3	MLP	MEIS1	103	RRAS	PTRF	MGC2650	DOCK1
41	FBXO7	JAM3	WSB2	RRAS	104	TGFB3	SDC4	CANX	DKFZP586A0522
42	DKFZP564M1416	FHL1	PLAB	KIAA1128	105	IGF1	CLIPR-59	ADCY3	RBPMS
43	DPT	**ANGPT1**	BUCS1	RRAS	106	RGN	SPARCL1	SH3YL1	ARMET
44	**PTGDS**	RBPMS	**GSTM4**	COX7A1	107	-	FEZ1	FASN	AKR1A1
45	**DF**	MYH11	RPL18A	ACTC	108	MADH4	MYL9	KIAA0934	RBPMS
46	SMARCD3	ITPR1	-	DMD	109	MADH4	ENIGMA	ERG	SMTN
47	**NELL2**	RBPMS	DKFZP586I2223	PLEKHC1	110	CASP9	FGF2	TM4SF2	ST5
48	BART1	**COL4A6**	RPS18	TPM2	111	PTGDS	**SC65**	U38A	DMN
49	FGFR2	SNX1	ATP6V1G1	DES	112	MLLT1	EDNRA	KIAA0746	CX3CL1
50	**ANGPT1**	COL6A1	RPS2	CDC42EP3	113	RBPMS	ACTG2	C2orf3	**NME1**
51	TGFB3	LAPTM4A	**DSCR1L1**	PTRF	114	TPS1	KIAA0467	GADD45G	PPP1R3C
52	RASA1	ACTN	DKFZP564B167	ACTG2	115	COL4A3	MEIS3	ANGPTL2	**CRYAB**
53	IGF2	TACC1	**ANGPT1**	FER1L3	116	TNFRSF4	-	CPD	EZH1
54	CX3CL1	ACTA2	CDC42BPA	SRF	117	MLLT7	KIAA0725	MPDH2	ATP1A2
55	PRSS11	TCF8	PENK	FLNA	118	FLJ32389	RIMS3	p100	CTF1
56	**GSTM4**	MBNL1	**CRYAB**	BC008967	119	GASP	KCNMB1	KIAA0342	TPM1
57	SMARCD3	RARRES2	UAP1	VCL	120	MEIS3	RNPC1	BMP5	GPM6B
58	SPOCK3	LDB3	**NME1**	MYL9	121	TIP120B	COX7A1	CLDN8	PPP3CB
59	COL4A3	EFEMP2	**NME2**	RARRES2	122	CYLD	FNBP1	RPL14	RIL
60	CTF1	GNAI2	EEF2	SDFR1	123	ASPA	STAT5B	RCL	PRKCB1
61	RAMP2	SYP	MGC5178	LMOD1	124	DBCCR1	NID	TFPI	RGN
62	ZNF288	ST5	ATP1A1	FGFR2	125	SPINK2	KIAA0193	ALCAM	COPE
63	COL13A1	MXRA7	FLRT2	GAS1	126	CXCL13	**SC65**	RPLP2	**DSCR1L1**

Genes that were selected in common are shown in bold and underlined.

We then compared the informative genes from independently conducted data to the informative genes from the integrated data. When we compare the informative genes of Singh and (LaTulippe+Welsh), 15 genes are found in common. In contrast, when we compare the common genes of Singh and (Latulippe+Welsh+Singh) with LaTulippe and Welsh, only 9 and 10 genes are found in common, respectively.

Furthermore, among the 15 informative genes we identified several tumor marker genes such as *HPN*, *C7orf24*, *NME1*, *NME2*, *CRYAB*, and *PYCR1*. *HPN* has been identified as a marker gene of prostate cancer in recent studies [19, 20]. *HPN* encodes hepsin, a cell surface transmembrane serine protease which plays an essential role in cell growth and the maintenance of cell morphology [7]. Also, *NME1* has been well known to be involved in the metastatic potential of several tumor cells, including prostate cancer cells [21]. Recently, Reference [22] reported that *C7orf24* may have an important role in cancer cell proliferation, and may be an appropriate therapeutic target molecule against cancer. However, these genes are not included in the list of LaTulippe or Welsh, either. These findings suggest that we can extract important marker genes which are missed in an individual data analysis by integrating several different microarray data.

5 Conclusion

In this paper, we showed the effectiveness of microarray integration and analysis method using *rank data*. To verify the statistical significance of the nonparametric scoring results, a random permutation test was performed for the *rank data*. With an experimental study using publicly available prostate microarray data, we also demonstrate that we can obtain more reliable and valid results from integrated data, based on a large number of samples.

Acknowledgment

This work was partially supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD, Basic Research Promotion Fund) (KRF-2007-531-D00019) and by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MOST) (No. R01-2006-000-11106-0).

References

1. Yoon, Y.M., Lee, J.C., Park, S.H.: Building a Classifier for Integrated Microarray Datasets through Two-Stage Approach. In: Proc. IEEE Symposium on Bioinformatics & Bioengineering, vol. 6, pp. 94–102 (2006)
2. Park, P.J., Pagano, M., Bonetti, M.: A nonparametric scoring algorithm for identifying informative genes from microarray data. In: Pacific Symposium on Biocomputing, pp. 52–63 (2001)

3. Golub, T.R., et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286, 531–537 (1999)
4. Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D., Chinnaiyan, M.: Meta-Analysis of Microarrays: Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer. Cancer Research 52, 4427–4433 (2002)
5. Jiang, H., Deng, Y., Chen, H.S., Tao, L., Sha, Q., Chen, J., Tsai, C.J., Zhang, S.: Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics 5, 81–93 (2004)
6. Cheadle, C., Vawter, M., Freed, W., Becker, K.: Analysis of Microarray Data Using Z Score Transformation. Journal of Molecular Diagnostics 5-2, 62–73 (2003)
7. Xu, L., Tan, A.C., Naiman, D.Q., Geman, D., Winslow, R.L.: Robust prostate cancer marker genes emerge from direct integration of inter–study microarray data. Bioinformatics Advance Access 21, 3905–3911 (2005)
8. Rosner, B.: Fundamentals of Biostatistics. Thompson 6, 540–544 (2003)
9. Shamir, B.A., Yakhini, R.Z.: Clustering gene expression patterns. J. Comput. Biol., 281–297 (1999)
10. Drăghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C., Krawetz, S.A.: Global functional profiling of gene expression. Genomics 81, 98–104 (2003)
11. Rogers, S., Williams, R.D., Campbell, C.: Class Prediction with Microarray Datasets. In: Bioinformatics using Computational Intelligence paradigms. Studies in Fuzziness and Soft Computing, vol. 176, pp. 119–141 (2005)
12. Deng, L., Pei, J., Ma, J., Lee, D.L.: A Rank Sum Test Method for Informative Gene Discovery. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), vol. 176, pp. 410–419 (2004)
13. Witten, I.H., Frank, E.: DATA MINING Practical Machine Learning Tools and Techniques, pp. 97–112. Morgan Kaufmann, San Francisco (2005)
14. Marko, R., Igor, K.: Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning Journal 53, 23–69 (2003)
15. Bailey, N.: Statistical methods in biology. Cambridge University Press, Cambridge (1995)
16. LaTulippe, E., Satagopan, J., Smith, A., Scher, H., Scardino, P., Reuter, V.: Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. Cancer Res. 62, 4499–4506 (2002)
17. Welsh, J.B., Sapinoso, L.M., Su, A.I., Kern, S.G., Wang-Rodriguez, J., Moskaluk, C.A.: Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Res. 61, 5974–5978 (2001)
18. Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209 (2002)
19. Hood, B., et al.: Proteomic Analysis of Formalin Fixed Prostate Cancer Tissue. Molecular & Cellular Proteomics 4, 1741–1753 (2005)
20. Pal, P., et al.: Variants in the HEPSIN gene are associated with prostate cancer in men of European origin. Hum. Genet. 210, 187–192 (2006)
21. Bemd, G., et al.: Mass spectrometric identification of human prostate cancer-derived proteins in serum of xenograft-bearing mice. Molecular & Cellular Proteomics 5, 1830–1839 (2006)
22. Iwaki, H., et al.: A novel tumor-related protein, C7orf24, identified by proteome differential display of bladder urothelial carcinoma. PROTEOMICS - Clinical Applications 1, 192–199 (2007)