The Korean Communications in Statistics **Vol. 14** No. 1, 2007, pp. 169–182

Statistical Method of Ranking Candidate Genes for the Biomarker*

Byung Soo Kim,[1)] Inyoung Kim,[2)] Sunho Lee[3)] and Sun Young Rha[4)]

Abstract

Receiver operating characteristic (ROC) approach can be employed to rank candidate genes from a microarray experiment, in particular, for the biomarker development with the purpose of population screening of a cancer. In the cancer microarray experiment based on n patients the researcher often wants to compare the tumor tissue with the normal tissue within the same individual using a common reference RNA. Ideally, this experiment produces n pairs of microarray data. However, it is often the case that there are missing values either in the normal or tumor tissue data. Practically, we have n_1 pairs of complete observations, n_2 "normal only" and n_3 "tumor only" data for the microarray. We refer to this data set as a mixed data set. We develop a ROC approach on the mixed data set to rank candidate genes for the biomarker development for the colorectal cancer screening. It turns out that the correlation between two ranks in terms of ROC and t statistics based on the top 50 genes of ROC rank is less than 0.6. This result indicates that employing a right approach of ranking candidate genes for the biomarker development is important for the allocation of resources.

Keywords: Biomarker; microarray; mixed data set; ranking genes; receiver operating characteristic (ROC) curve.

* B. S. Kim's work was supported by a grant No. (F01-2004-000-10044-0) from the Korea Science & Engineering Foundation. SH Lee was supported by a grant No. R04-2004-000-10145-0 from the Basic Research Program of the Korea Science and Engineering Foundation. S. Y. Rha was supported by a grant of the IMT-2000 project, Ministry of Health & Welfare, Republic of Korea (01-PJ11-PG9-01BT00A-0028).

1) Professor, Department of Applied Statistics, Yonsei University, Seoul 120–749, Korea. Correspondence : bskim@yonsei.ac.kr

2) Postdoctoral Fellow, Department of Epidemiology and Public Health, School of Medicine, Yale University, New Haven, CT 06520-8034, U.S.A.

3) Professor, Department of Applied Mathematics, Sejong University, Seoul 133-747, Korea.

4) Associate Professor, Cancer Metastasis Research Center, College of Medicine, Yonsei University, Seoul 120-752, Korea.

1. Introduction

Microarray technology holds the promise of becoming a new advance in modern cancer research and clinical diagnostics with its potential to quantitatively measure the expression levels of thousands of genes simultaneously. There are two types of a microarray; the single channel array, of which the Affymetrix system is the most predominant, and the dual channel array which includes a spotted cDNA microarray or a spotted oligonucleotide microarray. Here we focus on the cDNA microarray. One of the characterizing properties of cDNA microarray data is that it is subject to substantial variability, hence it is essential for the experimenter to carefully plan the experimental design driven by the study objective. For example, when using a cDNA array one must decide on a design for allocating specimens to labels and to array. The most commonly used design uses an aliquot of a reference RNA as one of the specimens for each array. This design is often referred to as a reference design or an indirect design. For most cancer microarray experiment, as Simon *et al.* (2002) indicates, there is not enough material available from one individual to create multiple arrays and thus the reference design would be the design of the choice. It is also a common practice in the cancer microarray experiment that a normal tissue is collected during the surgery from the same individual from which the tumor tissue was taken. The major reason of doing this is that it is not easy for the experimenter to collect normal tissues from healthy individuals. We observe that this practice of observing a matched pair adds a merit from a statistical aspect. For example, by observing a matched pair from a same individual one can reduce the inter-individual variability. It is often the case, however, that the experimenter can't extract enough RNA either from the tumor or the normal tissue to perform the microarray experiment due to poor quality of the tissue or other technical reasons. Therefore, collecting n cases does not necessarily end up with a matched pair sample of size n. (We use 'sample' to denote a random sample in statistics to distinguish it from a biological specimen.) Instead it usually consists of a matched pair sample of size n_1 and two independent samples of sizes n_2 and n_3, respectively for "reference versus normal only" and "reference versus tumor only" hybridizations ($n_1 + n_2 + n_3 = n$). Let X and Y denote the log fluorescent intensity ratios of reference versus normal and reference versus tumor hybridizations, respectively, for a given gene. Let U and V be independent copies of X and Y, respectively. Then we may observe three data types represented in Table 1.1. We refer to this data set as a mixed data set, as it contains a mix of fully observed and partially observed pair data.

Table 1.1: Three data types of the experiment. X and Y represent log intensity ratios for reference versus normal and reference versus tumor hybridizations, respectively. U and V are identically distributed with X and Y, respectively.

Hybridization		Number of cases
reference vs normal	reference vs tumor	
X	Y	n_1
U	missing	n_2
missing	V	n_3

The mixed data set of Table 1.1 occurred in clinical practice, particularly in the microarray experiment using human tissues (Kim *et al.*, 2005)
For handling the mixed data set of Table 1.1, standard procedures like the t statistic need to be modified properly. Kim *et al.* (2005) developed a $t-$based statistic as a means of combining all the data in the mixed data set when they detected differentially expressed (DE) genes between normal and tumor tissues.

The goal of the cancer screening is to detect tumors at an early stage so that the treatment is likely to be successful. Furthermore, it is essential that the screening tool is noninvasive and inexpensive to allow widespread application to the population. Another important aspect of the population screening of a cancer is that the screening tool maintains the low false positive rates. Even a small false positive rate translates into a large number of healthy people subject to diagnostic procedures that are unnecessary, costly and sometimes invasive. The aim of this note is to modify and extend the receiver operating characteristic (ROC) approach given the data set of Table 1.1. We then use the modified ROC approach to rank candidate genes that can be used for the biomarker development with the purpose of the population screening of a cancer.

2. Materials and Methods

2.1. Experiment and Data Pre-processing

Fresh specimens of cancer and normal tissues from each of 87 colorectal cancer patients were obtained during surgery at Severance Hospital, Yonsei Cancer Center, Yonsei University, College of Medicine, Seoul, Korea from May to December, 2002. These specimens were snap-frozen in liquid nitrogen right after the resection and stored at $-70\,^{\circ}\mathrm{C}$ until required. The median age of 87 patients was 65 with the range of 28–90. We had 46 and 41 for males and females. Other clinical

characteristics on location, carcinoembryonic antigen(CEA) level and stage were reported in Kim *et al.* (2005).

We attempted to extract total RNAs from tumor and normal tissues from each of 87 patients and wished to comprise a paired data set of size 87. From each of 36 patients we had RNA specimens both for tumor and normal tissues. However, from 19 patients RNA specimens for normal tissues only were available. From another 32 patients RNA specimens for tumor only were obtainable. Thus, we have a matched pair sample of size 36 and two independent samples of sizes 19 and 32. In terms of notations in Table 1.1, $n_1 = 36$, $n_2 = 19$ and $n_3 = 32$. We note that these tissues were taken by a single surgeon and there was no specific clinical or biological meanings on these three subgroups. Therefore, we assume that these three subgroups are independent samples from a population.

After total RNAs were extracted from fresh frozen tissues, the specimens were labeled and hybridized to cDNA microarrays based on the standard protocol established at Cancer Metastasis Research Center, Yonsei University College of Medicine (Park *et al.*, 2004)

We use $\mathrm{M} = \log_2(R/G)$ for the evaluation of relative intensity, where R and G represent the cy5 and cy3 fluorescent intensities, respectively. We first define no missing proportion (NMP) of a gene as the proportion of valid observation out of the total number of arrays. For example, if a gene has valid observation for 32 arrays out of 40, its NMP is 0.8.

We normalized M values using within-print tip group, intensity dependent normalization following Yang *et al.* (2002). We used 0.8 for the lower bound of the NMP, which deleted genes containing missing values for more than 20% of the total number of observations. We then imputed missing values employing $k-$nearest neighbor ($k = 10$) method. We averaged values for multiple spots. Finally, we ended up with a data set represented by 12850×123, where 12850 denotes the number of genes and 123 stands for the number of arrays.

2.2. Methods

We first briefly review the ROC approach when two independent random samples are given. We then extend the ROC approach to the matched pair sample and further to the mixed data set of Table 1.1. Comparison of ranks of genes in terms of the ROC approach and a $t-$based statistic, t_3 of Kim *et al.* (2005) is made based on 100 bootstrap samples. We perform small scale sensitivity analysis to investigate the stability of the ROC approach.

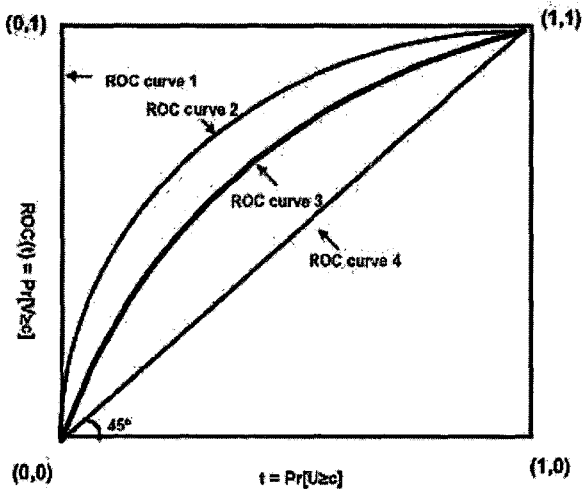

Figure 2.1: Examples of ROC curves. ROC curve 1 corresponds to the perfect gene, whereas ROC curve 4 stands for the uninformative gene. Most DE genes have their ROC curves between these two extreme cases. Better genes have ROC curves to the upper left corner.

We assume that $\{U_k\}_{k=1}^{n_2}$ and $\{V_l\}_{l=1}^{n_3}$ in Table 1.1 are independent random samples. Let $U = U_1$ and let $V = V_1$ just for simplifying the notation. The ROC curve is a plot of true positive versus false positive probabilities associated with varying thresholds for U and V. Just for the simplicity we present ROC curve for the up–regulated genes. However, the adaptation to the down–regulated gene is straightforward. For a given threshold c the false positive probability is given by $\Pr[U \geq c] \equiv t$, and the true positive probability is $\Pr[V \geq c] \equiv \mathrm{ROC}(t)$. Therefore, the ROC curve consists of $\{(t, \mathrm{ROC}(t)); 0 \leq t \leq 1\}$. Figure 2.1 shows four ROC curves corresponding to four hypothetical genes. The uninformative gene is one such that the probability distributions of expression levels are the same in the tumor and normal tissues, which results in $\Pr[U \geq c] = \Pr[V \geq c]$ for any threshold value c. The uninformative gene is represented by "ROC curve 4" in Figure 2.1. A perfect gene on the other hand completely separates tumor tissue from normal tissue. Its ROC curve is along the left and upper border of the positive quadrant, which is represented by "ROC curve 1" in Figure 2.1. Most differentially expressed (DE) genes have their ROC curves between these two extreme cases. Better genes have ROC curves to the upper left corner. We note that "ROC curve 2" is better than "ROC curve 3", because at any false

positive value t "ROC curve 2" has higher true positive probability than "ROC curve 3" and also at any true positive probability "ROC curve 2" maintains the smaller false positive probability than "ROC curve 3". Therefore, better genes have ROC curves closer to the upper left corner. Pepe *et al.* (2003) used empirical estimates of $\mathrm{ROC}(t_0)$ together with $p\mathrm{AUC} \equiv \int_0^{t_0} \mathrm{ROC}(t)dt$ for ranking genes of differential expression between normal and tumor tissues for a suitably chosen $t_0(\equiv \Pr[U \geq c_0])$ value which is also determined by the threshold c_0.

The ROC approach is now extended to the mixed data set of Table 1.1. We introduce some notations here. Let $\mathrm{ROC}_{pair}(t_0)$, $\mathrm{ROC}_{ind}(t_0)$ and $\mathrm{ROC}_{mix}(t_0)$ denote ROC values using the matched pair sample, two independent samples, and the mixed data set of Table 1.1, respectively, for a given threshold t_0, which, in turn, is determined by c_0 such that $t_0 = \Pr[U \geq c_0]$. Developing a ROC approach on the mixed data set consists of two steps. In the first step we devise a procedure for computing $\mathrm{ROC}(t_0)$, for a given threshold t_0, based on the paired data set of size n_1, which $\mathrm{ROC}_{pair}(t_0)$ denotes. Computation of $\mathrm{ROC}(t_0)$ based on two independent samples $\{U_k\}_{k=1}^{n_2}$ and $\{V_l\}_{l=1}^{n_3}$, denoted by $\mathrm{ROC}_{ind}(t_0)$, is straightforward along the line of Pepe *et al.* (2003). In the second step we properly average these two ROC values to derive $\mathrm{ROC}_{mix}(t_0)$ for the mixed data set.

Let $D = Y - X$ and let D_0 denote the hypothetical version of D under the null hypothesis of no differential expression. The distribution of D with a mean δ_a and the variance σ_a^2 is denoted by $D \sim (\delta_a, \sigma_a^2)$. The distribution of D_0 is represented by $D_0 \sim (\delta_0, \sigma_0^2)$. Let $\overline{D}$ denote the sample mean of D values based on n_1 paired observations. We augment the D notation by adding a superscript (g) to represent the g-th gene. Hence $D^{(g)}$ denote D for the g-th gene. We omit this superscript when the argument is based on each gene. We may note that we don't have any direct observations concerning the distribution of D_0. Let $|\overline{D}|_{(1)} \leq |\overline{D}|_{(2)} \leq \ldots \leq |\overline{D}|_{(p)}$ denote the order statistics of $\{|\overline{D^{(g)}}|\}_{g=1}^{p}$, where p is the number of genes spotted in a cDNA microarray. Let $g_{(j)}$ represent the index of a gene whose $|\overline{D}|$ value corresponds to $|\overline{D}|_{(j)}$ for $j = 1, \ldots, p$. We also refer to $S^2_{g(j)}(D)$ as the sample variance of D for the $g_{(i)}$-th gene based on n_1 paired observations.

The essence of computing $\mathrm{ROC}_{pair}(t_0)$ lies on estimating the baseline distribution of D under the null hypothesis of no differential expression. We assume, for simplicity, that D_0 has the same pattern of distribution with D except for the mean and variance. We expect that $\delta_0 \leq \delta_a$ and we don't necessarily assume that $\delta_0 = 0$. We further assume that $\sigma_0^2 \leq \sigma_a^2$. There are several ways of estimating the

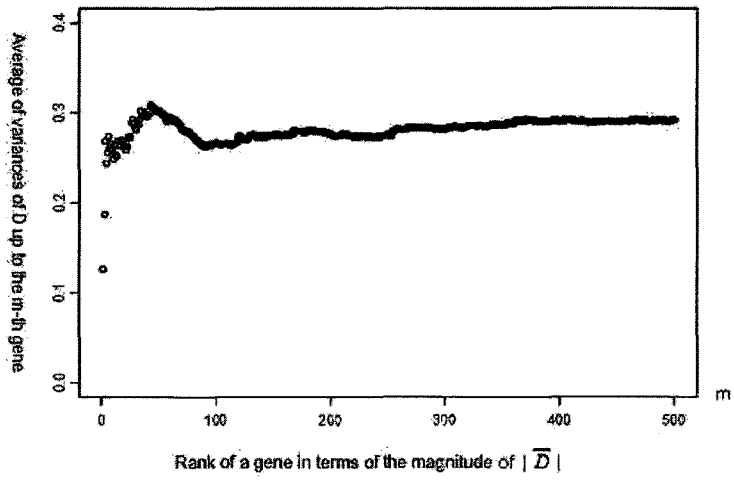

Figure 2.2: The plot of $(m, \frac{1}{m}\sum_{j=1}^{m} S^2_{g(j)}(D))$, for $m = 1, \ldots, 500$. X axis represents the rank (m) of $|\overline{D}|$ up to the 500^{th} and Y axis indicates the average of sample variances of D up to the m-th gene whose $|\overline{D}|$ value corresponds to the smallest m-th rank.

distribution of D_0 using the matched pair sample data. We choose a set of genes, denoted by $N_\epsilon = \{g; |\overline{D^{(g)}}| < \epsilon\}$ fot a small $\epsilon > 0$. The suitable choice of ϵ can be determined from the plot of $[\{m, (1/m)\sum_{j=1}^{m} S^2_{g(j)}(D)\}]_{m=1}^{p}$. We concluded from the plot of Figure 2.2 that the first 100 order statistics provide information on the non–DE genes. Based on these 100 genes in N_ϵ we can estimate δ_0 and σ_0^2. We derive the empirical distribution of D based on $\{D_i\}_{i=1}^{36}$, which the density of D in Figure 2.3 represents. We then shift the mean and adjust the scale of D by multiplying the factor σ_0/σ_a so that $D_0 = (D - \delta_a)(\sigma_0/\sigma_a) + \delta_0$ has mean δ_0 and variance σ_0^2. This results in a null hypothesis distribution of D represented by the density of D_0 in Figure 2.3. We finally proceed the calculation of $\mathrm{ROC}_{pair}(t_0)$ as is illustrated in Figure 2.3.

Once $\mathrm{ROC}_{pair}(t_0)$ and $\mathrm{ROC}_{ind}(t_0)$ are determined we may average these two ROC values to get $\mathrm{ROC}_{mix}(t_0)$. The initial idea was using the weighted mean of these two ROC's where the weights were proportional to the inverse of their variances. For calculating the variance of $\mathrm{ROC}_{ind}(t_0)$ we attempted using Result 5.1 of Pepe (2003). We noted that Result 5.1 of Pepe (2003) didn't work out for computing variance of $\mathrm{ROC}_{ind}(t_0)$, because for some genes which well

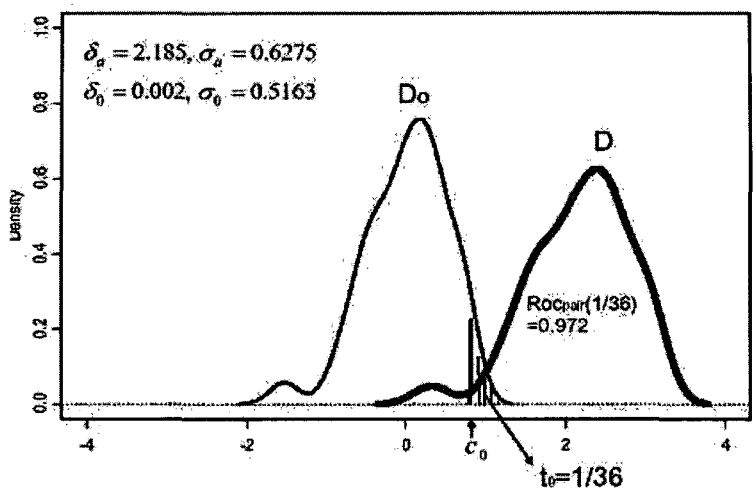

Figure 2.3: Schematic plot of calculating $\text{ROC}_{pair}(1/36)$ of a gene (transient receptor potential channel 3, AI655379, the 3^{rd} gene in terms of ROC_{mix} in Table 3.1) based on the paired data set of size 36. ROC_{mix} in Table 3.1 is a weighted mean of $\text{ROC}_{pair}(1/36)$ and $\text{ROC}_{ind}(1/36)$.

separated two distributions Equation (5.2) of Pepe (2003) involved division by zero. We attempted employing bootstrap method for calculating the variance of $\text{ROC}_{pair}(t_0)$. However, for some of genes which well separated distributions of normal and tumor tissues, we observed that zero variances occurred when we performed bootstrap procedure for calculating the variance of $\text{ROC}_{pair}(t_0)$. As an alternative we used the weighted average of Equation (2.1) to get $\text{ROC}_{mix}(t_0)$, where the weights were adopted from t_3 statistic of Equation (2.2).

Then we use the following weighted average to get $\text{ROC}_{mix}(t_0)$;

$$\text{ROC}_{mix}(t_0) = \frac{n_1 \text{ROC}_{pair}(t_0) + n_H \text{ROC}_{ind}(t_0)}{n_1 + n_H}, \tag{2.1}$$

where n_H is the harmonic mean of n_2 and n_3. We also use a newly developed $t-$based statistic, $t3$, of Equation (2.2) for detecting DE genes in a mixed data set of Table 1 (Kim *et al.*, 2005).

$$t_3 = \frac{n_1 \overline{D} + n_H(\overline{V} - \overline{U})}{\sqrt{n_1 S_D^2 + n_H^2(\frac{1}{n^2} S_U^2 + \frac{1}{n_3} S_V^2)}}, \tag{2.2}$$

where

$$\overline{D} = \frac{1}{n_1}\sum_{j=1}^{n_1} D_j \equiv \frac{1}{n_1}\sum_{j=1}^{n_1}(Y_j - X_j),$$

$$\overline{U} = \frac{1}{n_2}\sum_{k=1}^{n_2} U_k,$$

$$\overline{V} = \frac{1}{n_3}\sum_{l=1}^{n_3} V_l,$$

S_D^2, S_U^2 and S_V^2 are sample variances of D, U, and V, respectively, and n_h is the harmonic mean of n_2 and n_3. The central limit theorem can be invoked to approximate the null distribution of t_3 in Equation (2.2) by $N(0,1)$. We then compare the ranks of genes in terms of ROC_{mix} against the ranks in terms of t_3 statistic.

We generated 100 bootstrap samples from the mixed data set of Table 1.1 and computed ROC_{mix} and t_3 statistics for each sample. ROC_{mix_boot} and t_{3_boot} represent averages of these two statistics based on 100 bootstrap samples.

Actual sample sizes of the mixed data set of Table 1.1 are $n_1 = 36$, $n_2 = 19$ and $n_3 = 32$. These small sample sizes may raise a concern on the stability of ROC approach, in particular, for ranking candidate genes for the biomarker. We carried out a sensitivity analysis of the ROC approach based on aforementioned 100 bootstrap samples. Let $\mathrm{ROC}_{mix\_boot}(t_0)$ denote the average of $\mathrm{ROC}_{mix}(t_0)$ values based on 100 bootstrap samples for a given threshold t_0.

We report in Results section the top 50 genes ranked in terms of ROC approach based on a cDNA microarray experiment of 87 colorectal cancers.

3. Result

One can determine the false positive probability $t_0 = \Pr[U \geq c]$ in the baseline (nontumor) distribution very small, in particular, in the context of cancer screening. However, due to small sample sizes, the estimation of $\mathrm{ROC}(t_0)$ at very small t_0 is not possible. Thus, one needed to compromise in the real application, as Pepe *et al.* (2003) indicated, with the choice of t_0 such that it was small, but large enough to make $\mathrm{ROC}(t_0)$ reasonably precise. We chose t_0 to be $1/36$.

Table 3.1 shows the list of top 50 genes in terms of $\mathrm{ROC}_{mix}(1/36)$, another top 50 genes in terms of $\mathrm{ROC}_{mix\_boot}(1/36)$, their corresponding ranks in terms of t_3 and t_{3_boot} statistics. Two sets of top 50 genes in terms of ROC_{mix} and

Table 3.1: The top fifty genes in terms of $\mathrm{ROC}_{mix}(1/36)$ and their corresponding ranks and values in terms of other statistics including $\mathrm{ROC}_{mix\_boot}(1/36)$, t_3, and t_{3_boot} statistics. Two sets of top 50 genes in terms of ROC_{mix} and ROC_{mix_boot} procedures overlap 47 genes. The bottom three genes are included to list up the top fifty genes in terms of $\mathrm{ROC}_{mix\_boot}(1/36)$.

Rank in terms of		Gene Name	Gene Id	ROC	ROC	t_3	t_{3_boot}	Rank in terms of	
ROC $_{mix}$	ROC $_{mix\_boot}$			$_{mix}$	$_{mix\_boot}$			t_3	t_{3_boot}
1.5[a]	1	ATP–binding cassette, sub-family A (ABC1), member 8	AA634308	1.000	1.000	−23.930	−24.379	2	2
1.5	2	solute carrier family 4, sodium bicarbonate cotransporter, member 4	AA452278	1.000	0.993	−19.399	−19.842	14	12
4.0	3	transient receptor potential channel 3	AI655379	0.983	0.986	−25.108	−26.012	1	1
4.0	4	endothelial cell-spe cific molecule 1	W46577	0.983	0.981	21.124	21.803	5	5
4.0	6	stromal cell-derived factor 1	AI655374	0.983	0.972	−18.515	−19.017	25	25
6.5	8	tetraspan transmembrane 4 super family	AA046527	0.975	0.970	20.666	21.334	7	7
6.5	5	ets variant gene 4 (E1A enhancer-binding protein, E1AF)	AA010400	0.975	0.975	20.030	20.607	10	10
9.5	12	cadherin 3, type 1, P-cadherin (placental)	AA425217	0.971	0.960	21.304	22.172	4	4
9.5	9	chromogranin A (parathyroid secretory protein 1)	AA976699	0.971	0.969	−18.701	−19.157	23	23
9.5	11	transmembrane 4 superfamily member 2	N93505	0.971	0.962	−18.181	−18.650	28	28
9.5	7	carbonic anhydrase II	H23187	0.971	0.972	−17.341	−17.721	45	40
12.5	10	nuclear factor (erythroid –derived 2)–like 3	W76339	0.963	0.965	22.929	23.338	3	3
12.5	18	matrix metallo proteinase 11 (stromelysin 3)	AA954935	0.963	0.952	15.605	15.953	99	98
14.0	13	nhibin, beta A (activin A, activin AB alpha polypeptide)	AI925826	0.958	0.960	17.881	18.061	36	31
16.0	14	somatostatin	R51912	0.954	0.957	−20.647	−21.473	6	8
16.0	17	extracellular link domain-containing 1	AA704407	0.954	0.953	−20.220	−20.849	9	9
16.0	15	hypothetical protein FLJ21511	AI373245	0.954	0.956	−17.059	−17.393	52	50
18.5	29	B-cell CLL/lymphoma 2	W63749	0.950	0.922	−17.774	−18.241	33	36
18.5	20	carbonic anhydrase I	R93176	0.950	0.943	−16.088	−16.281	91	77
21.0	19	matrix metalloproteinase 7 (matrilysin, uterine)	AA031514	0.946	0.945	19.332	19.641	18	13
21.0	16	integral membrane protein 2A	AA775257	0.946	0.955	−18.661	−19.254	22	24
21.0	24	wingless-type MMTV integration site family, member 5A	W49672	0.946	0.937	16.909	17.726	44	53
23.5	22	Kruppel-like factor 4 (gut)	H45668	0.942	0.939	−19.118	−19.698	16	16
23.5	28	tryptophan hydroxylase (tryptophan 5-monooxygenase)	AA702193	0.942	0.925	−16.640	−17.465	49	61
25.0	25	CDC28 protein kinase 2	AA397813	0.933	0.931	18.958	19.113	24	19

26.5	44	UDP-glucose dehydrogenase	AA454086	0.929	0.882	−19.943	−20.336	11	11
26.5	21	ESTs	R31701	0.929	0.940	17.774	17.894	40	35
29.0	23	tryptophan hydroxylase (tryptophan 5-monooxygenase)	AA975820 AI733159	0.925	0.937	−19.105	−19.885	13	17
29.0	42	dyskeratosis congenita 1, dyskerin	AA052960	0.925	0.885	18.194	18.761	27	27
29.0	40	fucosidase, alpha −L−1, tissue	N95761	0.925	0.888	−18.073	−18.548	29	29
32.0	26	carbonic anhydrase XII	AA171613	0.921	0.928	−16.782	−17.339	53	57
32.0	27	glucagon	AI955772	0.921	0.926	−16.393	−16.884	66	69
32.0	31	serine/threonine kinase 15	R19158	0.921	0.913	15.613	15.967	98	97
34.0	30	No data	W80637	0.917	0.916	19.025	19.688	17	18
35.5	35	sushi−repeat− containing protein, X chromosome	AA449715	0.913	0.892	−17.558	−18.284	32	37
35.5	33	ESTs, Moderately similar to A40493 DNA topoisomerase [H.sapiens]	AI337434	0.913	0.902	16.071	16.745	71	79
37.0	43	procollagen (type III) N-endopeptidase	H98666	0.908	0.882	17.120	17.595	47	48
39.0	39	ectonucleoside triphosphate diphosphohydrolase 5	AI017442	0.904	0.890	−16.233	−16.864	67	73
39.0	38	proteolipid protein (Pelizaeus−Merzbacher disease, spastic paraplegia 2, uncomplicated)	R45264	0.904	0.890	−16.210	−16.565	78	74
39.0	32	peptide YY, 2 (seminalplasmin)	AI342688	0.904	0.907	−16.115	−17.002	62	76
41.0	37	chaperonin containing TCP1, subunit 5 (epsilon)	AA629692	0.900	0.891	20.908	21.278	8	6
42.0	59	cytoskeleton associated protein 2	AA504130	0.896	0.843	16.671	17.041	60	60
44.0	34	S100 calcium−binding protein A11 (calgizzarin)	AA464731	0.892	0.900	17.328	17.876	42	41
44.0	47	tryptophan hydroxylase (tryptophan 5-monooxygenase)	AI701018	0.892	0.879	−15.995	−16.648	76	84
44.0	41	carboxypeptidase M	AI367796	0.892	0.886	−15.650	−16.485	81	94
46.0	49	sorbitol dehydrogenase	AA700604	0.888	0.868	17.452	17.889	41	39
48.0	57	bone morphogenetic protein 2	AI569017	0.883	0.845	−16.511	−17.096	58	64
48.0	45	prostaglandin D2 synthase, hematopoietic	AI206447	0.883	0.880	−16.473	−16.759	70	65
48.0	36	matrilin 2	AA071473	0.883	0.892	−16.014	−16.097	94	83
50.0	52	minichromosome maintenance deficient (S.cerevisiae) 3	AI669374	0.879	0.860	17.516	18.062	35	38
51.5	46	ESTs	AA481059	0.875	0.879	18.746	19.312	21	21
61	48	ubiquitin carrier protein E2-C	AA430504	0.846	0.877	17.193	17.450	50	47
71.5	50	matrix metalloproteinase 3 (stromelysin 1, progelatinase)	W51794	0.808	0.866	17.792	18.404	31	34

*The average is taken for tied ranks

ROC_{mix_boot} have 47 genes in common. Thus, Table 3.1 contains 53 genes and these genes are sorted by ROC_{mix} values. We also note that 34 genes overlap between two sets of top 50 genes in terms of ROC_{mix} and t_3 statistics. This overlap proportion is in parallel with Pepe *et al.* (2003), which reports 8 over-

Table 3.2: Correlations among four ranks in terms of ROC_{mix}, ROC_{mix_boot}, t_3, and t_{3_boot} based on top 50 genes of ROC_{mix}.

	ROC_{mix}	ROC_{mix_boot}	t_3
ROC_{mix_boot}	0.923		
t_3	0.570	0.517	
t_{3_boot}	0.543	0.485	0.983

lapping genes between two sets of top 10 genes in terms of ROC and two sample t statistics.

Correlations among four sets of ranks in Table 3.1 are given in Table 3.2. We observed a correlation coefficient of 0.570 between two sets of ranks in terms of ROC_{mix} and t_3 based on the top 50 genes of ROC_{mix}. We note even smaller correlation (0.517) between these two sets of ranks based on the top 50 genes in terms of t_3 statistic (data not shown). We may note that t_3 statistics is less sensitive to the sampling variability than ROC_{mix}. We may recall that ROC approach ranks genes based on the true positive probability ($\text{ROC}(t_0)$) corresponding to the fixed small false positive probability t_0. Therefore, it was expected that ROC approach was more sensitive to the sampling variation than t_3 statistic, because ROC approach for ranking candidate genes ignored much of the information in the ROC curves, namely, ROC curves beyond $t > t_0$.

Among the selected genes, several genes were reported to relate with carcinogenesis and colorectal cancer. CXCR4_AI655374 (Ottaiano *et al.*, 2004) and CKS2_AA397813 (Li *et al.*, 2004) were significantly over–expressed in cancerous type (carcinomas and metastasis) compared to non–cancerous type. Several members of the S100 protein family of calcium-binding proteins (two isoforms of S100A9, S100A8, S100A11_AA464731 (Stulik *et al.*, 1999) and S100A6) were up-regulated in transformed colon mucosa. Meanwhile, SST_R51912 is known to induce cell apoptosis of large intestine cancer and inhibit cell proliferation (Mao *et al.*, 2005). Bone morphogenic protein 2(BMP2_AI569017) inhibits the colonic epithelial cell growth in vitro by promoting apoptosis and inhibiting proliferation (Hardwick *et al.*, 2004). ROC_{min}

4. Discussion and Conclusion

We have extended the ROC approach for ranking candidate genes for the biomarker development which is applicable to the mixed data set of microarray

experiment performed on human caner and normal tissues. The mixed data set of Table 1.1 occurs quite often in clinical practice when the tissue material is not large enough to yield the adequate amount of RNA for undergoing the DNA microarray experiment. There is a possibility that the ROC approach, in general, and the ranking the genes, in particular, might be sensitive due to the small sample size. We conducted a small sensitivity study of this ROC approach by random sampling. The high correlation coefficient of 0.923 between two sets of ranks in terms of ROC_{mix} and ROC_{mix_boot} in Table 3.2 indicated that even with this small sample size we could find a small set of DE genes which well separated tumor and nonturmor.

It was already shown that ROC approach was better than the classic measures of discrimination such as t statistic or Mann–Whitney U statistic in ranking candidate genes for the biomarker development for the purpose of the population screening of cancer (Pepe *et al.*, 2001, 2003). We observed that the correlation between two sets of ranks in terms of t statistic and ROC based on the top 50 ROC ranked genes was less than 0.6. Our result indicates that the proper method of ranking candidate genes, such as the ROC based approach, is quite important in allocating resources. Therefore, investigators should carefully choose the statistical measure for ranking the genes so that it fits the purpose of the experiment.

References

Hardwick, J. C., Van Den Brink, G. R., Bleuming, S. A., Ballester, I., Van Den Brande, J. M., Keller, J. J., Offerhaus, G. J., Van Deventer, S. J. and Peppelenbosch, M. P. (2004). Bone morphogenetic protein 2 is expressed by, and acts upon, mature epithelial cells in the colon. *Gastroenterology*, **126**, 111–121.

Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y. and Chung, H. C. (2005). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics*, **21**, 517–528.

Li, M., Lin, Y.M., Hasegawa, S., Shimokawa, T., Murata, K., Kameyama, M., Ishikawa, O., Katagiri, T., Tsunoda, T., Nakamura, Y. and Furukawa, Y. (2004). Genes associated with liver metastasis of colon cancer identified by genome-wide cDNA microarray. *International Journal of Oncology*, **24**, 305–312.

Mao, J. D., Wu, P., Xia, X. H., Hu, J. Q., Huang, W. B. and Xu, G. Q. (2005). Correlation between expression of gastrin, somatostatin and cell apoptosis regulation gene bcl-2/bax in large intestine carcinoma. *World Journal of Gastroenterology*, **11**, 721–725.

Ottaiano, A., Palma, A. di., Napolitano, M., Pisano, C., Pignata, S., Tatangelo, F., Botti, G., Acquaviva, A. M., Castello, G., Ascierto, P. A., Iaffaioli, R. V. and Scala, S. (2004). Inhibitory effects of anti-CXCR4 antibodies on human colon cancer cells. *Cancer Immunolgy, Immunotherapy*, **54**, 781–791.

Park, C. H., Jeong, H. J., Jung, J. J., Lee, G. Y., Kim, S. C., Kim, T. S., Yang, S. H., Chung, H. C. and Rha, S. Y. (2004). Fabrication of high quality cDNA microarray using a small amount of cDNA, *International Journal of Molecular Medicine*, **13**, 675–679.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, New York.

Pepe, M. S., Etzioni, R., Feng, Z., Potter, J., Thompson, M. L., Thornquist, M., Winget, M. and Yasui, Y. (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*, **93**, 1054–1061.

Pepe, M. S., Longton, G. M., Anderson, G. L. and Schummer, M. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics*, **59**, 133–142.

Stulik, J., Koupilova, K., Osterreicher, J., Knizek, J., Macela, A., Bures, J., Jandik, P., Langr, F., Dedic, K. and Jungblut, P.R. (1999). Protein abundance alterations in matched sets of macroscopically normal colon mucosa and colorectal carcinoma. *Electrophoresis*, **20**, 3638–3646.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research*, **30**, e15.

[Received September 2006, Accepted January 2007]