Prognostic index reflecting genetic alteration related to disease-free time for gastric cancer patient

MIJUNG KIM[1] and SUN YOUNG RHA[2,3,4]

[1]Institute for Mathematical Sciences, Yonsei University; [2]Cancer Metastasis Research Center, [3]Brain Korea 21 Project for Medical Sciences, [4]Department of Internal Medicine, Yonsei University College of Medicine, 134 Shinchon-Dong, Seodaemun-Gu, Seoul 120-752, Korea

Received March 3, 2009; Accepted May 5, 2009

DOI: 10.3892/or_00000454

Abstract. The study purpose was to develop a patient's prognostic index (PI) reflecting the genetic information in cDNA microarray-based CGH experiment data for estimating a gastric cancer patient's survival time. The developed methodology was fit to and validated using data from the Cancer Metastasis Research Center at Yonsei University; 30 pairs of gastric tumors and normal gastric tissues were used in the cDNA microarray-based CGH. The cDNA microarrays containing 17,000 sequence-verified human gene probes were directly compared. Genetic alteration score (GAS) was constructed based on the genes that had a high frequency of alteration among all the genes displaying small variations across the arrays. GAS was determined using a technique that finds linear combinations of the original variables that best account for the variability in the data. When classifying cancer patients with the PI predicted by the model incorporating GAS, the correct classification rate for recurrence was 83.33%. In conclusion, GAS allowed for providing an independent patient's PI that reflects the genetic information for prognosis on hazard rate of recurrence, which was capable of distinguishing a patient's recurrence status, survival status and cancer stage status. The predicted PI also provided each patient's estimated disease-free survival rate. In this study, 82 genes were selected for analysis based on a high frequency of alteration and small variations across the arrays. In addition, 13 genes displaying a possible relationship with disease-free survival time were identified. GAS was found to be associated with the recurrence status and survival status.

Correspondence to: Dr Mijung Kim, Institute for Mathematical Sciences, Yonsei University, 134 Shinchon-Dong, Seodaemun-Gu, Seoul 120-752, Korea
E-mail: mjkim@yonsei.ac.kr

Key words: cDNA microarray-based CGH, gene copy-number change, prognostic score, prognostic index, correct classification rate, factor analysis, Cox's proportional hazard model

Introduction

Gastric cancer is a major cause of human cancer-related mortalities (1). A genomic alteration detected by cDNA microarray-based CGH can easily be translated into both sequence and gene identification, which can provide additional information concerning the complex chromosomal rearrangements and imbalances (2). A gastric cancer related, cDNA microarray-based CGH experiment was performed for investigating genomic aberrations with a high resolution at the Cancer Metastasis Research Center at Yonsei University (3,4). Thirty pairs of gastric tumors and normal gastric tissues were used and the cDNA microarrays containing 17,000 sequence-verified human gene probes were directly compared. For analyzing cDNA microarray-based CGH data it was important to identify genes altered in gastric cancer since cDNA microarray-based CGH data include special features such as low-intensity spots. For the reason that analysis using mean values, such as the t-test, does not well identify 'altered gene' where its mean copy-number change should be over the criterion on alteration, frequency analysis was conducted in this study to detect subtle differences in copy-number change. For performing frequency analysis it was necessary to deal with variation of gene over the arrays because frequency analysis, such as a 1.5- or 2-fold change cut-off, does not consider variations of the gene over the arrays. In our previous study, a reproducible gene selection algorithm (RGSA) was developed for controlling variations of the genes across arrays with the same data as the cDNA microarray-based CGH at the Cancer Metastasis Research Center at Yonsei University (5). Kadota *et al* developed the pre-processing implementation for microarray (PRIM) for extracting reproducible data from the result of duplicate experiments (6). Rosner (7) and Dowdy and Wearden (8) introduced intra-class correlation coefficient and its application with random effect model. In RGSA, the variability of measurements across arrays was quantified via a random effect model and a measurement of reproducibility was incorporated using intra-class correlation coefficient. RGSA controls both reproducibility and the number of remaining

genes. The well filtered set of this article suggested had both reproducibility and number of remaining genes were maximized.

This study concerned the genes altered in gastric cancer and selected genes altered utilizing RGSA for filtering genes with small variations across arrays after taking the steps of within-print tip, intensity-dependent normalization on the data.

A genetic alteration score (GAS) was determined from genes displaying alterations in gastric cancer with the aim of finding characteristics that are related to the disease-free time. Using this scoring system, it was possible to search for specific genes that have a possible relationship to disease-free time of gastric cancer and to predict the prognostic index (PI) for the hazard rate of cancer recurrence, which reflects the genetic information of a gastric cancer patient. It was also possible to estimate a patient's disease-free survival rate for gastric cancer with this predicted PI. For this purpose, a genetic scoring system related to the hazard rate was established.

For scoring genetic information, Yang *et al* related the summation of changes in the number of gene copies of amplified genes (without considering deletions) to the recurrence of cancer (4) and Inoue *et al* assigned a weight of +1 or -1 to the gene depending on its characteristic for the five conventional pathological factors in relation to gastric cancer (9). Liu and Huang suggested a linear transformation method for cancer classification using rotation forest (10). Liebermeister applied independent component analysis to gene expression data for deriving a linear model based on hidden variables (11). Park *et al* developed a linear transformation method linking gene expression data with patient survival time using partial least squares method (12).

In the present study, not only was the gain considered but also the loss in the number of gene copies and a weight was assigned to each gene according to its contribution to the genetic score, which was related to the disease-free time of gastric cancer. In this process, the variability of the genes altered in gastric cancer was decomposed with several latent factor variabilities that represent common characteristics of the genes using factor analysis technique, where the latent factors were independent components and consisted of linear combinations of the genes. Among these common characteristics, the characteristic that was related to the disease-free time of gastric cancer was obtained to establish a score system that assesses the hazard rate with variable selection of the Cox's proportional hazard model. Cox (13) introduced regression model in life tables and proportional hazard model and Agresti (14) and Lee (15) also discussed regression model for lifetime. The selected common factor characterizing the relationship to the disease-free time provided the patient's genetic alteration level, which was based on changes in the gene copy-number. These genetic alteration levels were used as the GAS. GAS can be used to search for genes that have a possible relationship to the disease-free time and to predict a PI, which reflects the genetic information of a gastric cancer patient.

The strategy of this study was to: a) select genes that were altered in gastric cancer; b) construct the statistical model for the characteristics of the representative genes using a linear combination of common (latent) characteristics of the representative genes; c) find characteristics that were related to the disease-free time via variable selection in Cox's proportional hazard model; d) apply GAS for predicting the PI of a gastric cancer patient's hazard upon recurrence and thus disease-free survival rate and find genes that were related to disease-free time by investigating the loading of GAS for each gene.

The methods used to accomplishing this are discussed in the following sections of this study; 'Materials and methods' section explains data preparation for this study and describes how to establish GAS related to the disease-free time of gastric cancer. In 'Results' section genes related to the disease-free time are searched, the PI, which reflects the genetic information of each gastric cancer patient is predicted and cumulative disease-free survival rates on patient groups are also predicted and shown with figures. In addition, the recurrence status, survival status and cancer stage of gastric cancer patients was classified using the predicted PI and the correct classification rate was calculated.

Materials and methods

cDNA microarray-based CGH data were obtained from the Cancer Metastasis Research Center at Yonsei University and analyzed with an application of the RGSA, which was developed in our previous study. Patient and tissue samples were prepared as follows: 30 pairs of normal gastric mucosa and cancer tissues were obtained from gastric cancer patients who had undergone surgery at the Severance Hospital, Cancer Metastasis Research Center (CMRC), Yonsei University Health System, Seoul, Korea, from 1997 to 1999. The study followed the local ethical guidelines of the Institutional Review Board of the Yonsei University Medical Center, Seoul, Korea. The tissue samples were immediately frozen into liquid nitrogen at the time of resection and stored at -150˚C until further use. Clinical data description is shown in Table I.

Next, DNA extraction and cDNA microarray-based CGH was conducted as the following steps: genomic DNA extraction from the tissue was performed according to the conventional protocol using phenol/chloroform/isoamylalcohol method. The cDNA microarrays containing 17,000 human gene probes (CMRC-Genomictree, Korea) were used for CGH following the standard protocol of CMRC, Yonsei University (3,4). Briefly, 4 μg of the normal or cancer DNA from the same patient was fluorescently labeled with Cy3 or Cy5-dUTP (Amersham, USA), respectively, using a BioPrime DNA Labeling System (Invitrogen, USA). The labeling products were purified with a PCR purification kit (Qiagen, Germany) and combined with human Cot-1 DNA (30 μg; Gibco BRL, USA), yeast tRNA (100 μg; Gibco BRL) and poly(dA-dT) (20 μg; Sigma, USA). The hybridization mixture was then concentrated using Microcon 30 (Millipore, USA) and hybridized to the 17K microarray at 65˚C for 16-18 h. After washing, the microarray was scanned using GenePix 4000B (Axon Ins., USA). The experiment is done with direct comparison. In this experiment, normal and tumor genomic DNA samples are extracted from the same patient and hybridized on the same spotted array. cDNA (17K) microarray contained the 15,723 unique genes with 17,664 spots and

Table I. Clinical information of the patients.

Categorical variable	Class	Cases	Total cases
Survival status	Death	15	29
	Survival	14	
Stage	I, II	12	30
	III, IV	18	
Recurrence status	Recurrence	13	27
	Non-recurrence	14	
Gender	Female	3	30
	Male	27	
Continuous variable	**Range**	**Average (SD)**	**Total cases**
Age	Min 41, Max 78	63.830 (9.710)	30
Size	Min 9, Max 126	40.330 (29.680)	30
Lymph node metastasis	Min 0, Max 0.518	0.119 (0.158)	30

these unique genes were mapped for their chromosomal location using SOURCE (http://genome-www5.stanford.edu/cgi-bin/source/sourceSearch) and DAVID (http://apps1.niaid.nih.gov/david/).

Data preparation and establishing GAS

Data preparation. The transformation of the intensity signal to a ratio was carried out using the $\log_2$ red to green ratio, $\log_2(R/G)$, where R and G denote the fluorescent intensities of tumor and normal hybridizations, respectively. Pre-processing of the data was done with within-print tip, intensity-dependent normalization of Y following Yang *et al* (16). Genes showing missing values for >20% of the total number of observations were deleted and 10-nearest neighbor method was employed for imputation of missing values. Averaged values were used in case of the multiple spots. At this step, 10,514 genes were found from the 30 microarrays and the set of these data was the initial set for analysis; filtering genes with RGSA was performed on this set.

Park *et al* evaluated genome-wide measurement of copy-number of each gene in normal gastric cancer and placenta tissues for determining the criteria on a genomic alteration with the same data of cDNA microarray-based CGH; the range of genomic copy-number of normal tissues was found to be ±0.3 of the $\log_2$ fluorescence intensity ratio in the autosomal genes (3). This criterion was used for categorizing gene's copy-number change into alteration and non-alteration. The cDNA microarray-based CGH data for this study has been deposited into Array Express (http://www.ebi.ac.uk/arrayexpress/) Query:1283947172 E-TABM-171. Data analysis was performed with SAS V.9.1 (17).

Establishing GAS. Variations of genes across arrays when selecting genes that were altered in gastric cancer were controlled in the process of creating the GAS; genes with relatively large variations to total variations were removed so that the reproducibility for the set of remaining genes increased. To optimize both the reproducibility and number of remaining genes, the threshold for the screened set was determined when the product of the reproducibility and the number of remaining genes was maximal (the number of remaining genes decreases when the reproducibility of the remaining genes increases and *vice versa*). The threshold, k, was used to determine which genes would be removed, where genes of variations k times larger than the total variation were removed. This set was named the well-filtered set (denoted by S_{opt}) and was subjected to the RGSA. Genes from this set with a high frequency (at least 30% frequency) of alteration (gain or loss) across arrays were identified after the genes in each array were categorized into two categories, alteration and non-alteration.

The criterion used to define gain/loss was +0.3 and -0.3, that is, genes that had between a +0.3 and -0.3 log_2 copy-changes were categorized into the non-alteration group and those outside this regime were categorized into the alteration (gain or loss) group (3). Using this procedure, 101 genes were found to be altered in relation to gastric cancer and the data from 82 of the 101 genes were complete, while 152 genes were selected without considering variations in data across arrays. To compare the GAS obtained from the initial set, *INI*, and the well-filtered set, S_{opt}, two GAS scores (denoted by GAS_I and GAS, respectively) were calculated based on the change in the copy-number of the selected genes from the two sets, where the initial set, *INI*, consists of all the genes without consideration of the gene variations in the set and the well-filtered set, S_{opt}, consists of the screened genes where both the reproducibility and remaining genes were kept as large as possible. The estimated survival functions with significant prognostic factor and GAS_I or GAS (the significant prognostic factor incorporated in the survival model is discussed in Table IIIA) showed almost the same standard errors in each time interval even though GAS was created with a smaller number of genes than GAS_I. It is also worth noting that the genes used for GAS were selected from the set of genes with small variations compared to GAS_I. The average of the standard errors for the survival functions estimated with GAS_I and GAS were the same, 0.075.

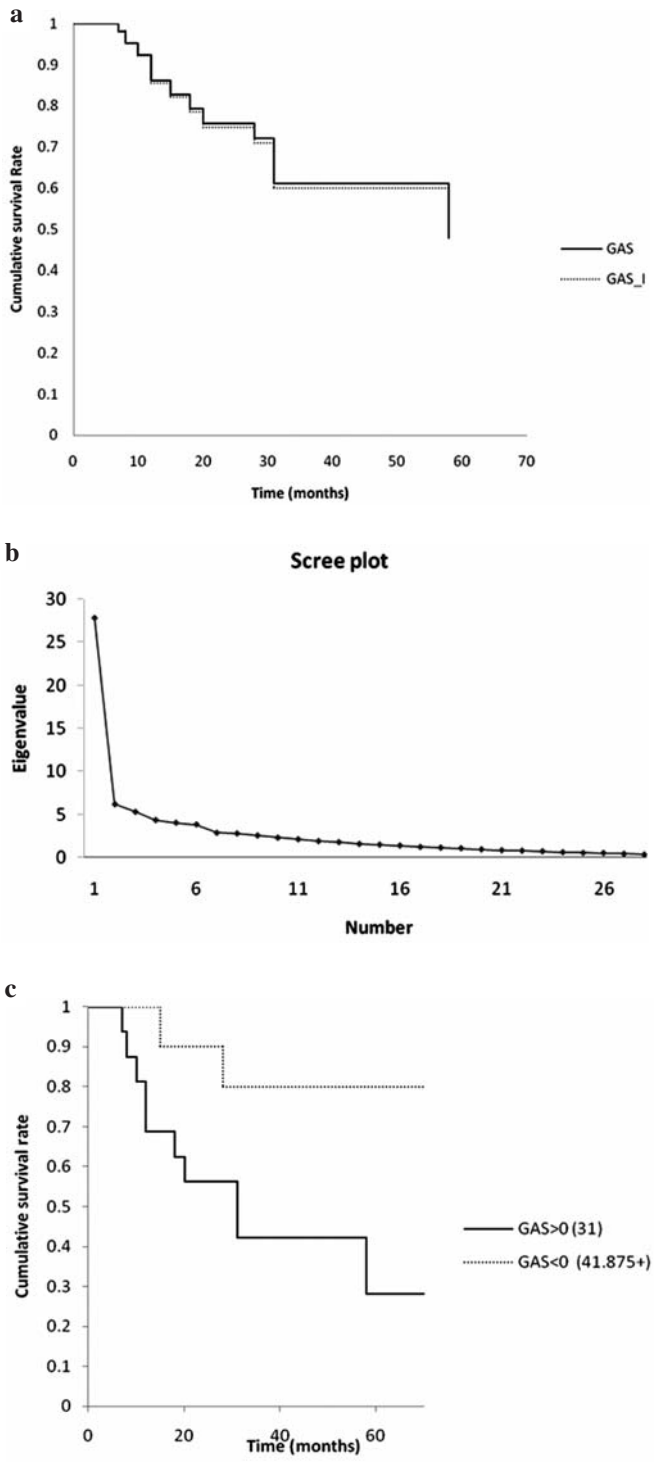

Figure 1. (a) Estimated disease-free survival curves using the score GAS_I and GAS. Estimated survival functions incorporating GAS_I and GAS, respectively where average prognostic factors incorporated in the Cox's model are 0.119 for lymph node metastasis and 0 for GAS_I and GAS. (b) Scree plot for the proportions of variability explained by the factors. x- and y-axis represent factor number and eigen value for the proportion of variability explained by the corresponding factor. (c) Kaplan-Meier survival curves on the two groups, GAS<0 and GAS>0. Value inside parenthesis is median survival time and a + indicates that the median survival time was limited to 65 months.

Fig. 1a shows the estimated disease-free survival curves on the recurrence of gastric cancer in patient with an average prognostic factor using the two score systems, GAS_I and GAS, where the x- and y-axis stand for the patient's survival time to recurrence and cumulative disease-free survival rate, respectively. There was no significant difference between the two survival curves and also no significant difference between the two prognostic indices estimated with GAS_I and GAS (p-value of 0.8623).

In the next step, common (latent) characteristics of the 82 representative genes were identified by decomposing the characteristic of the representative genes with several latent characteristics that were common to all the representative genes. This is done using factor analysis with the objective of finding characteristic that was related to the cancer patient hazard rates of recurrence of gastric cancer.

To determine the proper number of common factors, a scree plot was examined and nine factors were identified when the decrease in the eigen value became stable and the cumulative variability was at least 70% (the 73% variability was explained with the nine factors) (Fig. 1b).

Nine factors consisted of linear combinations of the selected 82 genes. The nine factors were used in a Cox's proportional hazard model, including the possible clinical prognostic factors.

Each gene had correlations with these nine common factors (characteristics), which are shown in Table IIA as factor loadings. Genes boldfaced showed a strong correlation with the corresponding factor. They in each factor explain the factor's characteristic that corresponded with the largest correlation when compared to the other factors.

The possible prognostic variables considered in the Cox's proportional hazard model were age, gender, lymph node metastasis (LN), size of tumor, cancer stage (early stage of I and II, late stage of III, IV).

Using stepwise variable selection with the nine factors including these clinical prognostic factors in the Cox's regression model, factor 6 and clinical prognostic factor, LN were identified as significant factors in relation to the hazard rate of recurrence of gastric cancer, with significance level of entry and effect to stay 0.05 (Table IIIA). Factor 6 was a common factor characterizing genes that were altered in gastric cancer patients, which was found to be related to the disease-free time of gastric cancer and was utilized as the GAS. This GAS was a linear combination of the changes in the copy-number of the selected 82 genes.

The rotated factor patterns of Table IIA show three genes, AA290624, AA421335 and AA278852, that were representative of the GAS characteristic and were distinguished from the characteristics of the other eight factors; these genes contribute the most among the nine factors to the GAS. It is noteworthy that of the three genes, AA290624 and AA421335 had a positive correlation with the GAS and displayed gains with a high frequency, and AA278852 had a negative correlation with the GAS and displayed a loss with a high frequency, which will be discussed with Table IIB in the section of 'Results'.

Table IIIA shows the result with the stepwise variable selection in the Cox's proportional hazard model. It indicates that the patients with a positive GAS had a higher risk of recurrence per unit time than a patient with an average GAS and LN* when adjusting LN*, where LN* is the lymph node

Table II. Characteristics for the 82 representative genes.

A, Rotated factor patterns for the 82 representative genes

Gene Bank Accession ID	Gene name	Loss	Gain	Mean	F1	F2	F3	F4	F5	F6	F7	F8	F9
AI094796	GIPC1	0	10	0.243	**0.822**	0.070	-0.056	0.051	0.295	0.009	0.082	-0.040	0.015
AI014388	TMEM165	0	17	0.297	**0.789**	0.169	-0.076	-0.093	0.084	0.233	0.058	-0.078	0.151
R22188	SKP2	0	17	0.316	**0.786**	-0.026	-0.256	-0.252	0.315	0.003	0.012	0.158	0.215
AA991514	POLK	0	15	0.293	**0.660**	0.242	-0.255	-0.131	0.036	-0.110	-0.022	-0.157	0.035
AI216112	ARHGEF11	0	14	0.315	**0.660**	0.106	-0.040	-0.244	-0.122	0.175	-0.019	0.021	-0.028
AI273225	LACTB	0	11	0.231	**0.650**	0.310	-0.300	-0.249	0.315	0.044	0.184	0.230	0.024
H54023	LILRB2	0	12	0.238	**0.640**	0.149	-0.032	-0.028	0.043	-0.133	-0.076	-0.111	-0.292
AI951840	SIX6	0	19	0.325	**0.630**	0.139	-0.328	-0.421	0.083	0.097	0.150	-0.094	0.126
AA932759	STS-1	0	12	0.268	**0.620**	0.274	-0.118	-0.193	-0.097	0.030	0.225	0.144	0.058
AA776891	SCARB2	0	15	0.275	**0.620**	0.294	-0.225	-0.333	0.137	-0.059	-0.331	0.072	0.139
N98412	LYZ	0	18	0.310	**0.609**	0.098	-0.466	-0.423	0.186	-0.084	0.035	-0.039	0.154
AI244751	SETBP1	0	11	0.243	**0.584**	0.347	0.002	-0.136	-0.245	0.245	-0.006	0.476	0.089
AI299228	-	0	12	0.268	**0.561**	0.230	-0.407	-0.027	0.408	0.120	0.013	-0.069	0.062
AA872372	-	0	9	0.260	**0.549**	0.361	-0.360	-0.099	0.250	0.147	-0.057	0.370	-0.119
AI312979	-	0	9	0.271	**0.520**	0.299	-0.177	-0.079	0.221	0.102	0.048	-0.373	0.207
AI369284	GTSE1	0	10	0.238	**0.484**	0.382	-0.325	-0.356	-0.142	0.294	0.000	-0.296	-0.022
AA279023	LRCH3	0	17	0.317	**0.477**	0.341	-0.043	-0.003	-0.034	0.195	0.063	0.287	-0.092
AI261377	-	0	12	0.280	**0.461**	0.377	-0.013	-0.404	-0.121	0.217	-0.069	-0.199	-0.084
AI339958		14	0	-0.244	**-0.520**	-0.267	0.182	0.282	0.160	-0.361	-0.239	0.062	-0.438
AA128162	MS4A4A	11	0	-0.255	**-0.598**	-0.357	0.008	-0.126	-0.234	-0.075	-0.302	0.300	-0.070
AI336859	ASXL2	12	0	-0.282	**-0.669**	0.018	0.380	0.223	0.166	-0.010	0.027	-0.007	0.024
AI348435	TNK2	0	20	0.332	0.183	**0.798**	-0.168	-0.304	0.184	0.172	-0.008	-0.184	-0.071
AI301753	EIF2S3	0	14	0.297	0.091	**0.771**	0.032	0.187	0.206	-0.007	0.042	-0.211	-0.070
AA983626	NOM1	0	19	0.374	0.138	**0.759**	-0.226	-0.258	0.020	-0.057	0.027	-0.043	-0.173
AA917892	YY1	0	11	0.237	0.252	**0.745**	-0.304	-0.116	-0.072	0.271	-0.071	-0.052	0.184
AI272002	CPM	0	9	0.199	0.275	**0.725**	-0.116	-0.036	0.070	0.019	-0.053	0.139	0.150
AI270875	CD55	0	15	0.327	0.353	**0.679**	-0.419	-0.074	0.029	-0.049	0.125	-0.120	0.121
AA917683	HERPUD1	0	9	0.221	0.242	**0.617**	-0.176	-0.318	0.208	0.024	-0.036	0.108	-0.228
AA988569	-	0	16	0.329	0.392	**0.598**	-0.477	-0.068	0.144	0.101	-0.065	-0.075	0.026
N70013	RP2	0	11	0.270	0.235	**0.578**	-0.138	-0.109	0.359	-0.472	0.066	0.187	0.212
AA278956	PHC3	0	9	0.249	0.142	**0.571**	-0.114	-0.205	-0.234	0.147	0.087	0.230	-0.251
AI421834	-	0	16	0.296	0.174	**0.559**	-0.063	-0.511	-0.102	0.084	0.177	-0.258	-0.096
AA933078	-	0	13	0.278	0.547	**0.556**	-0.115	-0.090	0.328	-0.100	-0.129	0.042	0.091
AA281064	FGFR1	0	9	0.262	0.372	**0.545**	0.295	0.129	-0.133	-0.144	-0.331	-0.055	0.208
AI262776	SLC12A3	0	14	0.312	0.304	**0.508**	-0.220	-0.253	0.226	0.356	0.042	0.096	0.150
AI630817	CDH22	0	9	0.203	-0.169	**0.457**	0.088	-0.187	-0.124	-0.062	0.245	0.006	0.252
R54968	COL16A1	0	12	0.251	0.428	**0.432**	-0.337	-0.170	0.249	0.407	-0.069	-0.189	0.173
AI611010	RPS10	0	9	0.248	0.421	**0.423**	-0.359	-0.412	0.308	0.265	0.001	-0.100	-0.131
R18845	ZNF559	9	0	-0.234	-0.341	**-0.471**	0.187	0.106	-0.138	-0.460	0.103	0.275	-0.337
T57875	PRKCI	0	11	0.240	0.015	**-0.574**	-0.112	-0.359	0.379	-0.153	-0.017	0.243	-0.161
H62387	ISLR	11	0	-0.227	0.016	-0.144	**0.777**	0.012	-0.320	-0.206	0.059	0.048	0.167
AI418042	RDH13	9	0	-0.255	-0.196	-0.049	**0.761**	0.383	0.032	0.090	-0.122	-0.192	0.154
AI564953	HAP1	18	0	-0.361	-0.207	-0.273	**0.720**	0.362	-0.132	-0.158	-0.028	-0.003	-0.140
AI669693	CD207	11	0	-0.291	-0.358	-0.051	**0.711**	0.421	-0.260	-0.083	0.082	-0.126	-0.064
AA995890	DKFZP564	16	0	-0.322	-0.091	-0.229	**0.703**	0.370	-0.255	-0.253	-0.156	0.114	-0.095
AA425008	CBLN1	9	0	-0.243	-0.174	-0.178	**0.647**	-0.079	0.069	0.254	0.080	-0.127	-0.034
AW009594	ARRB2	10	0	-0.263	-0.141	-0.392	**0.572**	0.417	-0.003	0.000	-0.190	-0.046	-0.275
AA873073	PEX5	11	0	-0.218	-0.065	-0.391	**0.566**	0.032	0.357	-0.343	0.002	-0.014	-0.082
AI681562	MTERFD2	9	0	-0.188	-0.340	-0.137	**0.539**	0.489	-0.311	-0.175	0.077	0.055	0.026
AA150402	COL4A1	9	0	-0.226	-0.523	-0.027	**0.534**	-0.099	0.183	0.190	-0.111	0.137	-0.026
AI768026	GTPBP1	12	0	-0.241	-0.144	0.122	**0.510**	-0.026	0.107	-0.252	-0.047	-0.061	0.089
AI207203	R3HDM1	10	0	-0.237	-0.029	-0.256	**0.461**	0.420	-0.293	-0.141	0.249	0.251	-0.130

Table IIA. Continued.

Gene Bank Accession ID	Gene name	Loss	Gain	Mean	F1	F2	F3	F4	F5	F6	F7	F8	F9
AA281426	SLC11A1	0	15	0.285	0.323	0.244	**-0.442**	0.195	0.334	0.157	-0.019	0.332	0.206
AA419268	PIK3R2	0	9	0.228	0.409	0.414	**-0.561**	-0.283	0.193	-0.034	-0.040	-0.180	0.208
AA987446	C3orf63	9	0	-0.210	0.269	-0.011	-0.168	**0.753**	-0.162	-0.286	-0.252	0.253	-0.032
AI018497	TCF3	14	0	-0.286	-0.305	0.033	0.106	**0.734**	0.120	0.064	-0.142	-0.149	-0.121
AA995875	AQP2	10	0	-0.218	-0.205	-0.337	0.329	**0.677**	-0.050	0.118	0.078	-0.057	-0.075
AI521736	BCL6B	11	0	-0.241	-0.058	-0.429	-0.042	**0.629**	-0.299	-0.050	-0.190	-0.123	-0.144
AA001403	ZMPSTE24	10	0	-0.220	-0.220	-0.176	0.179	**0.607**	-0.130	-0.190	-0.218	0.424	0.262
AI198289	ENAH	9	0	-0.191	-0.233	-0.524	0.308	**0.543**	0.032	-0.096	0.200	-0.160	0.064
H98218	HMGA2	10	0	-0.247	-0.255	-0.216	0.250	**0.373**	-0.060	-0.069	-0.138	0.063	-0.070
AI653069	DOCK10	0	10	0.231	0.314	0.028	-0.346	**-0.364**	-0.068	-0.032	0.046	-0.356	0.296
AW006471	CTTN	0	17	0.308	0.282	0.070	-0.261	**-0.547**	0.161	0.013	-0.040	-0.181	-0.069
AA236957	ARHGEF6	0	9	0.218	0.031	0.059	0.049	**-0.628**	0.168	-0.531	0.011	-0.097	-0.041
AA994976	SLC2A9	0	14	0.299	0.307	0.024	-0.364	**-0.698**	0.026	0.018	0.315	0.082	0.043
AI308989	ZNF594	0	12	0.261	0.277	0.148	-0.103	-0.120	**0.706**	0.155	-0.135	0.085	0.220
N71462	SCML2	0	10	0.271	0.329	-0.104	-0.307	-0.068	**0.694**	-0.201	-0.251	0.174	0.109
AI304790	PTGIS	0	10	0.234	-0.106	0.070	0.012	-0.218	**0.616**	0.107	0.510	-0.147	-0.219
AA278850	DCLRE1C	0	9	0.242	0.299	0.232	-0.118	-0.381	**-0.439**	-0.093	0.280	0.143	-0.107
AA281744	C18orf54	10	0	-0.242	-0.168	-0.349	0.108	0.337	**-0.629**	-0.033	-0.120	0.245	-0.004
AA290624	-	0	11	0.250	0.267	0.149	-0.142	0.038	0.075	**0.752**	0.114	0.099	0.048
AA421335	LDB1	9	0	-0.199	0.055	-0.299	0.139	0.144	0.023	**0.576**	-0.462	-0.124	0.036
AA278852	COPS2	9	0	-0.234	-0.033	-0.423	0.263	0.210	-0.154	**-0.645**	-0.170	0.054	0.035
AA886742	LOC44120	0	9	0.205	0.241	0.069	-0.028	-0.141	0.039	0.047	**0.805**	0.003	0.271
AA583574	S100A7	0	24	0.414	0.161	-0.048	-0.037	-0.079	-0.198	-0.093	**0.774**	-0.148	-0.105
AI630806	-	0	11	0.263	-0.256	-0.088	-0.026	-0.239	-0.040	0.487	**0.576**	0.175	-0.056
AA088258	LOC40098	9	0	-0.252	-0.142	-0.233	-0.114	0.160	0.018	-0.049	-0.056	**0.712**	0.028
AA923509	CPM	0	12	0.252	0.390	0.375	-0.104	-0.259	0.075	0.336	-0.200	**0.448**	-0.124
AI285331	STK4	0	12	0.285	0.366	0.060	0.064	0.137	0.106	0.011	-0.021	0.136	**0.772**
N65981	-	10	0	-0.232	0.023	-0.176	0.293	0.041	-0.088	-0.437	0.231	0.023	**-0.452**
AA490609	MRPL1	10	0	-0.245	0.017	0.032	-0.145	0.134	-0.302	-0.004	-0.468	0.210	**-0.511**
AA666234	PNMA2	0	18	0.335	0.368	0.252	0.230	0.396	0.006	-0.026	-0.120	0.126	**-0.538**

The third and the fourth columns are frequency of losses and gains observed from the 30 patients' arrays for each gene. The fifth column is the mean of copy-number changes over the 30 arrays for each gene. F1 through F9 denote Factor1 through Factor9, and entries are factor loadings of the nine factors for each gene. Bolded text, shows a strong correlation with the corresponding factor.

B, Genes having correlation of at least 0.3 with disease-free-time related score, GAS

Gene Bank Accession ID	Gene name	Loading of GAS	Frequency of gains	Frequency of losses	Mean
AA290624*	-	0.752	11	0	0.250
AA421335*	LDB1	0.576	0	9	-0.199
AI630806	-	0.487	11	0	0.263
R54968	COL16A1	0.407	12	0	0.251
AI262776	SLC12A3	0.356	14	0	0.312
AA923509	CPM	0.336	12	0	0.252
AA873073	PEX5	-0.343	0	11	-0.218
AI339958	-	-0.361	0	14	-0.244
N65981	-	-0.437	0	10	-0.232
R18845	ZNF559	-0.460	0	9	-0.234
N70013	RP2	-0.472	11	0	0.270
AA236957	ARHGEF6	-0.531	9	0	0.218
AA278852*	COPS2	-0.645	0	9	-0.234

Table III. Characteristics for GAS.

A, Selected factors with the Cox's proportional hazard model

Variable	df	Parameter estimate	Standard error	Chi-square	P-value	Hazard ratio
LN*	1	0.749	2.198	11.608	0.0007	2.114
GAS	1	0.753	0.349	4.661	0.0309	2.124

B, Associations of the GAS with cancer stage, recurrence and survival status

Patient status	Patient groups	GAS<0 (no. of patients)	GAS>0 (no. of patients)	P-value (Chi-square test)
Cancer stage	Early stage (I, II)	7	5	0.1758
	Late stage (III, IV)	6	12	
Recurrence status	Recurrence	2	11	0.0098[a]
	Non-recurrence	9	5	
Survival status	Survival	9	6	0.0654[a]
	Death	4	11	

[a]Denotes significance at 10% error rate.

C, Comparison of the prognostic indices between two groups

Category	Class	Cases	Median prognostic index	P-value (K-S test)
Survival status	Death	15	1.087	0.0025
	Survival	15	-0.852	
Cancer stage	Early stage (I, II)	12	-0.899	0.0033
	Late stage (III, IV)	18	0.471	
Recurrence status	Recurrence	13	1.479	0.0023
	Non-recurrence	14	-0.899	

D, Correct classification rates on patient status

Patient status	Sensitivity (%)	Specificity (%)	Correct classification rate (%)
Survival	83.33	83.33	83.33
Stage (I, II vs. III, IV)	88.89	77.77	79.17
Recurrence	83.33	83.33	83.33

Concerned event on sensitivity is death, late cancer stage and recurrence for survival status, stage status and recurrence status, respectively in part D shown above.

metastasis centralized about its mean 1.19 after being multiplied by 10. The hazard rate of the GAS is 2.124, which implies that the estimated risk of recurrence per unit time increases 2.124 times for patients that have a one unit increase of GAS, when adjusting LN*, relative to a patient with average values for lymph node metastasis and GAS. Since GAS was established as a variable following a normal standard distribution with mean 0 in factor analysis, two groups of patients with positive and negative GAS (denoted with GAS>0 and GAS<0) were investigated.

Kaplan-Meier survival curves, shown in Fig. 1c, indicate that the difference between the curves of the two groups, GAS>0 and GAS<0 was significant (p-value 0.036). The median survival time when GAS>0 was 31 months; when GAS<0, the median survival time was 41.875 months and this estimation was determined over a 65 month time period, since the estimated disease-free survival rate was large when GAS<0.

The associations of the genetic alteration score (GAS>0, GAS<0) with cancer stage (early/late stage), recurrence status, and survival status were analyzed using the Chi-square test.

From this analysis, the genetic alteration score was found to be associated with the recurrence status and survival status; however, it was not associated with cancer stage as shown in Table IIIB.

GAS reflects a patients' genetic alteration levels better than a simple summation of the gene genetic alteration levels in the sense that weights were assigned to genes according to their accountabilities for the score variable. This weight was the coefficient of the gene in the linear combination used in determining the GAS, while loading of the GAS for each gene explains the correlation between GAS and the corresponding gene. GAS had a significant relationship with the disease-free time and thus with the disease-free survival rate for the recurrence of gastric cancer. In addition, the GAS may provide insight into genes that affect the disease-free time and therefore, affect the disease-free survival rate. By investigating the loading of GAS for each gene, it is possible to find genes that are strongly related to GAS and therefore possibly related to the disease-free survival rate.

Results

Searching genes related to disease-free time in gastric cancer. Genes with large positive or negative loadings had a strong correlation with GAS and therefore these genes may be related to the disease-free time of gastric cancer. In Table IIB, the GAS loading for each gene were listed in ascending order for genes that showed a positive or negative correlation of at least 0.3 with GAS of the 82 selected gene in the S_{opt}. The frequency of gains and losses observed from the 30 arrays (patients) and the mean intensities (copy-number changes) are also provided. These were the genes that have a possible relationship with the disease-free time of gastric cancer since GAS is significantly related to the disease-free time of gastric cancer (Table IIIA).

As shown in Table IIB, genes that had a strong positive correlation with GAS also displayed a high frequency of gains, except AA421335; genes that had a strong negative correlation with GAS also had a high frequency of losses, except N70013 and AA236957. The three genes, AA290624, AA421335 and AA278852 described in Table IIA are highlighted with an asterisk.

Prognostic index and disease-free survival function reflecting genetic information. By incorporating the GAS in the Cox's regression model, each patient's PI, which reflects the genetic information for assessing the hazard rate of recurrence, can be obtained. By doing this it would be possible to predict an independent patient's disease-free survival rate.

Each patient's PI, $\log_e h_i(t, x)/h_0(t, x)$, which measures the patient's prognosis upon the recurrence of gastric cancer, was obtained using the following equation:

$$\log_e h_i(t, x)/h_0(t, x) = 0.749\mathrm{LN}^* + 0.753\mathrm{GAS} \quad (2)$$

where $h_i(t, x)/h_0(t, x)$ was the relative risk of recurrence per unit time, that is, the ratio of the risk for a patient with a given set of GAS and LN* to the risk for a patient with an average LN* and GAS value of 0; the coefficients for LN* and GAS are from Table IIIA obtained by variable selection with the Cox's proportional hazard model. This index was used to compare the prognosis or relative risk between patients with different GAS and LN* values.

The disease-free survival function for a patient with a prognostic index of PI, was estimated using

$$\widehat{S(t)_0}^{\exp(PI)} \quad \text{, where} \quad \widehat{S(t)_0}$$

was the estimated survival function at the baseline, which was obtained from Kaplan-Meier's survivor function and PI was obtained by the equation (2); this estimated disease-free survival function reflected not only the genetic information but also the significant clinical factor of each patient.

Based on changes in the copy-number of the 82 genes, each patient's GAS and estimated PI was determined and shown in Table IV.

The predicted PI was found to be very useful in determining a patient's prognosis. The predicted prognostic indices distinguished patients' cancer stage (early stage vs. late stage), recurrence status and survival status. More specifically, there was a significant difference in the predicted prognostic indices between the early cancer stage vs. late stage, non-recurrence vs. recurrence and survival vs. death. *Kolmogorov-Smirnov* test's p-values are shown in Table IIIC. Furthermore, the median prognostic indices were different between any of the two groups (Table IIIC).

The estimated disease-free survival curves with an average prognostic index PI for each of the two groups,

$$\widehat{S(t)_0}^{\exp(PI)}$$

indicate that there was a significant difference between the two groups; median survival time was denoted in parenthesis where a + indicates that the estimated median survival time was limited to 65 months (Fig. 2a1-3); a steep disease-free survival curve represents a low disease-free survival rate or short survival time to recurrence.

In addition, when classifying patients using the prognostic indices, the correct rate for classification upon their recurrence status was found to be 83.33%, which reached 95.83% when the same procedure for the three subsets of the 82 genes was applied to overcome the problems associated with a small sample size (shown in the next sub-section).

Classifying gastric cancer patients with the predicted prognostic index. The recurrence status of patients was classified based on the PI, where a threshold 0 was used to determine the classification of the patient since the ratio of the risk of recurrence per unit time for a patient with a prognostic index 0 to the risk for a patient whose prognostic factors were at their average values was expected to be 1.

Fig. 2b shows the classification of patients' recurrence status determined from the predicted prognostic index of the patient, where the patient with negative (positive) prognostic index was classified as non-recurrence (recurrence) shown on the left (on the right).

No information on the recurrence status or disease-free survival time of six patients out of the 30 patients was available; thus, they were not considered on classification. Twenty out of 24 patients were correctly classified whose correct classification rate was 83.33%.

It is noteworthy that GAS was constructed as a random variable that follows a standard normal distribution, which

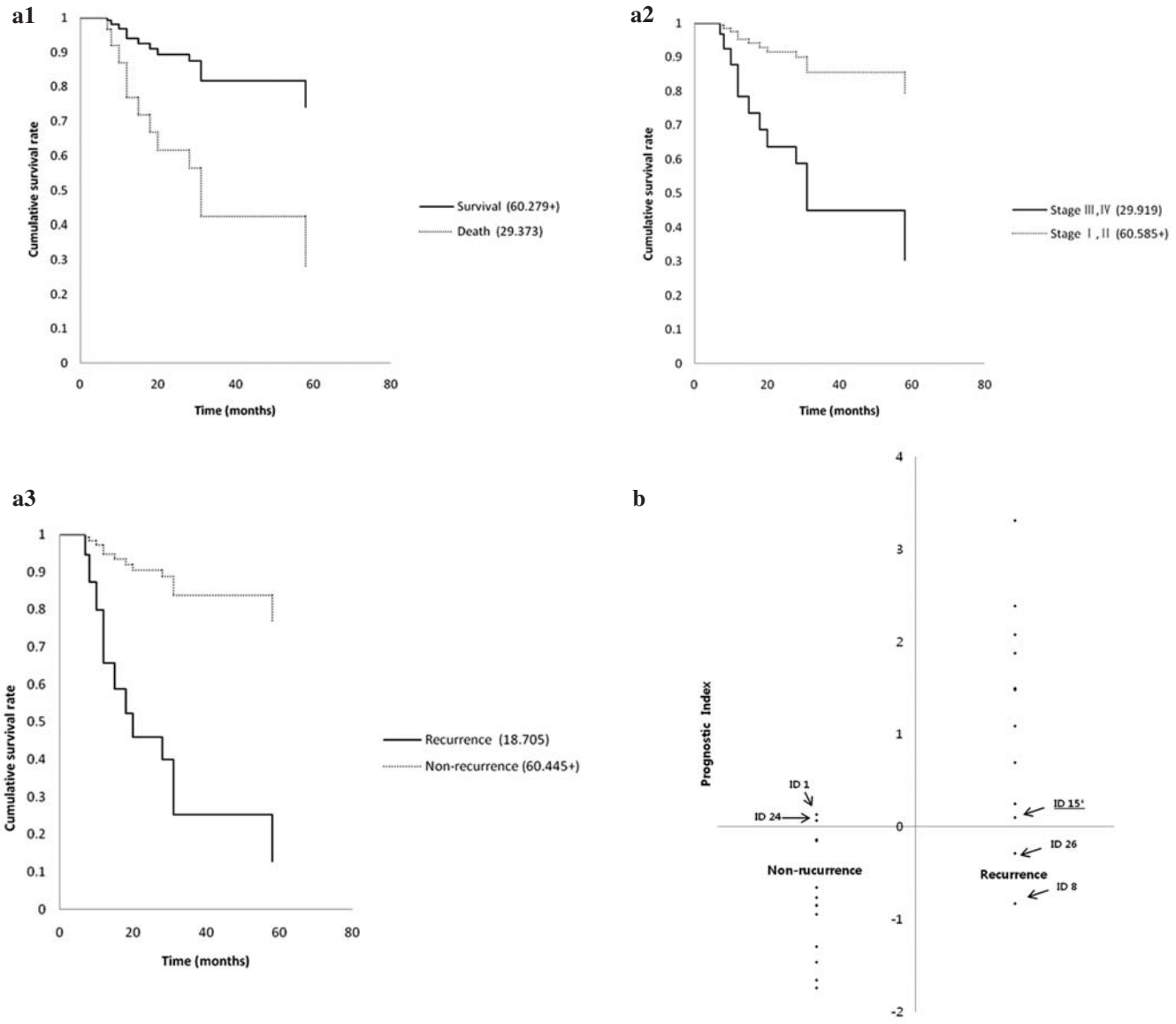

Figure 2. (a1-3) Estimated survival curves, $\widehat{S(t)_0}^{\exp(PI)}$ with an average prognostic index PI for each of the two groups. Value inside parenthesis is median survival time and a + indicates that the median survival time was limited to 65 months. (b) Classification on the 24 patients' recurrence status with prognostic index. Classification of patients' recurrence status was made with the predicted prognostic index of the patient; four patients, ID 1, ID 8, ID 24 and ID 26, were misclassified.

allows for finding the percentile of the patient's GAS score. For example, patient ID 15 had GAS score 1.069 which is ~86 percentile of GAS, thus had a high risk of recurrence and actually had recurrence. However, using only significant clinical factor, this patient was not correctly classified since this patient had a good prognostic factor, a small LN of 0.043 (LN^* of -0.751) while correctly classified by the model incorporating GAS.

Four patients, ID 1, ID 8, ID 24 and ID 26, were misclassified; a large GAS was found for the patients misclassified as recurrence (ID 1 and ID 24) and a small GAS was found for the patients misclassified as non-recurrence (ID 8 and ID 26). Since GAS was the one of the common characteristics for the 82 genes' variability and explained with a linear combination of the 82 genes, it could be better estimated when the sample size is at least the same as the number of genes coefficients that were being estimated. For this, the set of 82 genes was divided into three disjoint subsets such that the number of genes in each subset was not more than the sample size and the union of the three subsets included all 82 genes. Each subset was subjected to the same procedure that was performed on the set of 82 genes and the patient recurrence status was determined on the status with the larger frequency from the decisions obtained from the three Cox's proportional hazard models. Using this modified method, only one patient, ID 8, of the 24 patients, was misclassified, which correspond to a correct classification rate of 95.83%. This result indicates that it may be possible to achieve a 95.38% accuracy in classifying cancer patients using the prognostic index predicted by the method presented in this study when the sample size is large.

The classification rates when the predicted prognostic index with a threshold of 0 was used to assess the patient status on gastric cancer are shown in Table IIID.

Table IV. The patient GAS and predicted prognostic index in 30 patients.

Patient ID	LN*	GAS	Predicted prognostic index (PI)
1	-1.186	1.348	0.127
2	-1.186	0.190	-0.745
3	-1.186	0.229	-0.716
4	-0.873	-1.303	-1.636
5	-0.948	-0.159	-0.830
6	-1.186	-0.043	-0.921
7	-1.186	-1.067	-1.692
8	-1.186	0.109	-0.806
9	-1.186	-0.493	-1.260
10	-1.186	0.982	-0.148
11	-1.186	-0.959	-1.610
12	-1.186	-0.832	-1.515
13	-1.186	-0.457	-1.233
14	-0.835	-0.017	-0.638
15	-0.751	1.069	0.242
16	-1.186	-0.711	-1.424
17	-0.900	-0.825	-1.296
18	-0.853	0.668	-0.135
19	2.235	2.055	3.221
20	2.905	0.197	2.324
21	0.414	0.483	0.674
22	1.400	0.537	1.453
23	0.512	1.402	1.440
24	-0.054	0.144	0.068
25	3.993	-2.562	1.060
26	-0.721	0.347	-0.278
27	2.435	0.007	1.828
28	1.989	1.106	2.322
29	2.257	0.440	2.022
30	2.028	-1.883	0.100

LN* is the lymph node metastasis centralized about the mean after being multiplied by 10, that is, it is obtained by 10*LN-1.19.

The primary concern of this study was on the genes altered in gastric cancer and finding the characteristic related to disease-free time among the characteristics of those altered genes and therefore the model was built on the 82 genes which are showing alteration in gastric cancer rather than distinguishing recurrence status. It is noted that even though the classifications were made with prognostic index obtained with disease-free time related score, the correct classification rates were not low.

Discussion

In this study, a genetic alteration score that was related to the disease-free time of gastric cancer was established and genes with a possible relationship to disease-free survival rate were examined by investigating the loading of GAS. The use of a predicting PI based on the GAS score to assess the hazard rate and survival rate of recurrence was also investigated. This study was conducted with data from the Cancer Metastasis Research Center at Yonsei University, where 30 pairs of gastric tumor and normal gastric tissues were used in the cDNA microarray-based CGH.

The primary concern of this study was to investigate the characteristics of genes that were altered in gastric cancer and especially the characteristics that were related to the disease-free time and thus related to the disease-free survival rate of gastric cancer. To achieve this, a GAS that was related to the disease-free time was constructed and a PI that reflects the GAS was obtained with a model that was built on genes that displayed alteration with high frequency in relation to gastric cancer. The GAS was determined with linear combination of copy-number changes of the representative genes altered in gastric cancer. The copy-number of 82 genes were found to be altered (gain or loss) with at least a 30% frequency after genes with small variations across arrays were screened.

GAS has been found to be positively related to the risk of recurrence per unit time and thus patient with a positive GAS had a high risk of recurrence. The genetic alteration score (GAS>0 and GAS<0) was associated with both the recurrence status and survival status. It has also been shown that the estimated disease-free survival curves were statistically different between patients with positive GAS and negative GAS. GAS allowed for the identification of candidates for disease-free time related genes of gastric cancer, which was possible by examining the loading of GAS for each gene.

GAS was used to obtain a prognostic index that reflected the patient genetic information. The predicted PI provided each patient's estimated disease-free survival rate.

In regards to the characteristics of the genes altered in gastric cancer, another concern in relation to the GAS and prognostic index was the relationship of the PI with the recurrence status, survival status and cancer stage status. When this was examined, the patient prognostic indices were found to be statistically different between any of the two groups, early cancer stage versus late cancer stage, non-recurrence versus recurrence and survival vs. death. These results imply that the prognostic index predicted by the model that incorporated GAS can be used not only for comparing disease-free survival rate between patients with different LN and GAS but also for distinguishing patient cancer stage, recurrence status and survival status. When obtaining a patient's PI incorporating GAS into the Cox's regression model, it was possible to predict the recurrence status, survival status and cancer stage with a correct classification rate of 83.33, 83.33 and 79.17%, respectively, which could be increased at a larger sample size.

GAS was determined using a technique that finds linear combinations of the original variables that best account for the variability in the data and GAS was one of the characteristics that expressed copy-number changes using a linear combination of the 82 genes. Thus, GAS could be better estimated when the sample size is at least the same as the number of genes coefficients that were being estimated. A modified method was applied for investigating this fact. Three disjoint subsets were established such that the union of the three subsets resulted in a representative set of the 82 genes. The same procedure was used for each subset and the patient recurrence status was determined based on the outcomes from the three Cox's proportional hazard models.

Using this modified procedure, 23 of the 24 patients were correctly classified. Based on these results, it is expected that the correct rate for determining the recurrence of cancer would be improved when the sample size is large.

Acknowledgements

This study was supported by Korean Research Foundation Grant funded by the Korean Government (MOEHRD) (R03-2004-000-10048-0). The authors should like to express their gratitude to Cancer Metastasis Research Center at Yonsei University for permission of the data for this study and also thank Jin-Hyung Kim for his assistance in this study.

References

1. Pisani P, Parkin DM, Bray F and Ferlay J: Estimates of the worldwide mortality from 25 cancers in 1990. Int J Cancer 83: 18-29, 1999.
2. Squire JA, Pei J, Marrano P, *et al*: High-resolution mapping of amplifications and deletions in pediatric osteosarcoma by use of CGH analysis of cDNA microarrays. Genes Chromosomes Cancer 38: 215-225, 2003.
3. Park CH, Jeong HJ, Choi YH, *et al*: Systematic analysis of cDNA microarray-based CGH. Int J Mol Med 17: 261-267, 2006.
4. Yang SH, Seo MY, Jeong HA, *et al*: Gene copy number change events at chromosome 20 and their association with recurrence in gastric cancer patients. Clin Cancer Res 11: 612-620, 2005.
5. Kim M: Reproducible gene selection algorithm with random effect model in cDNA microarray-based CGH data: Expert Systems with Applications, DOI: 10.1016/j.eswa.2009.03.034'?.
6. Kadota K, Miki R, Bono H, *et al*: Preprocessing implementation for microarray (PRIM): an efficient method for processing cDNA microarray data. Physiol Genomics 4: 183-188, 2001.
7. Rosner B: Fundamentals of Biostatistics. 5th edition. Duxbury Publisher, Pacific Grove, pp557-567, 2000.
8. Dowdy S and Wearden S: Statistics for Research. 2nd edition. A Wiley-Interscience Publication, New Jersey, pp339-356, 1991.
9. Inoue H, Matsuyama A, Mimori K, Ueo H and Mori M: Prognostic score of gastric cancer determined by cDNA microarray. Clin Cancer Res 8: 3475-3479, 2002.
10. Liu KH and Huang DS: Cancer classification using Rotation Forest. Comput Biol Med 38: 601-610, 2008.
11. Liebermeister W: Linear modes of gene expression determined by independent component analysis. Bioinformatics 18: 51-60, 2002.
12. Park PJ, Tian L and Kohane IS: Linking gene expression data with patient survival times using partial least squares. Bioinformatics 18: 120-127, 2002.
13. Cox DR: Regression models in life tables. J Roy Stat Soc Ser B 34: 187-220, 1972.
14. Agresti A: Categorical Data Analysis. 2nd edition. Wiley-Interscience Publication, New Jersey, pp166-192, 2002.
15. Lee E: Statistical methods for survival data analysis. 2nd edition. Wiley-Interscience Publication, New Jersey, pp243-278, 1992.
16. Yang YH, Dudoit S, Luu P, *et al*: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30: e15, 2002.
17. SAS System for Windows V.9.1, SAS Institute Inc, 2000.