The Signature from Messenger RNA Expression Profiling Can Predict Lymph Node Metastasis with High Accuracy for Non-small Cell Lung Cancer

Naeyun Choi, MS,† Dae-Soon Son, MS,*† Jinseon Lee, PhD,† In-Seung Song, BS,† Kyung-Ah Kim, MS,‖ Sang-Ho Park, PhD,‖ Yoo-Sung Lim, BS,† Gil-Ju Seo, MS,† Jungho Han, MD,§ Hyejin Kim, MS,¶ Hye Won Lee, MS,¶ Jason Jong-ho Kang, PhD,‖ Jeong-Sun Seo, MD, PhD,‖ Ju Han Kim, MD, PhD,¶# and Jhingook Kim, MD‡*

Background: The extent of regional lymph node (LN) metastasis is the most important factor in the evaluation of resectability and prognosis of non-small cell lung cancer (NSCLC) to increase the chance of complete cure. The authors attempted to deduce a group of genes from the analysis of mRNA expression profiles of the tumor tissues of NSCLC patients with or without LN metastasis, and make a classification model for better prediction of LN metastasis.

Methods: The authors analyzed mRNA expression profiles of 79 NSCLC patients with or without LN metastasis, and deduced the gene signature for the predictive model of LN metastasis in lung cancer. The authors evaluated the predictive accuracy of each of four algorithms by applying them to another set of 33 NSCLC patients. Each algorithm's accuracy was calculated by 10-fold cross-validation, and a combined model showed a level of accuracy that was higher than any one of the better three component algorithms (i.e., ANN, DT, or NB). Avadis, SAS, ArrayXPath, and R-package were the statistical analysis software packages used.

Results: The authors selected 949 genes using a classical permutation *t* test ($p < 0.01$) and finally obtained a gene signature consisting of 31 genes by adjustment of multiple-hypothesis testing. The LN metastasis prediction model derived from the signature (31 genes) and their characteristic interactions provided a predictive accuracy of 84.85% when applied to a test set of 33 patients.

Conclusions: The authors have demonstrated that their gene signature developed by the expression profiling of mRNAs from the primary tissue could predict the LN metastasis status of NSCLC.

†Cancer Research Division, Center for Clinical Research, Samsung Biomedical Research Institute; Departments of ‡Thoracic Surgery and §Pathology, Samsung Medical Center, College of Medicine, Sungkyunkwan University School of Medicine; ‖Marcogen, Inc.; ¶Seoul National University Biomedical Informatics; and #Human Genome Research Institute, Seoul National University College of Medicine, Seoul, Korea.

*Contributed equally to the study.

Address for correspondence: Juhan Kim, MD, PhD, Seoul National University Biomedical Informatics (SNUBI), Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea. E-mail: juhan@snu.ac.kr

ISSN: 1556-0864/06/0107-0622

Key Words: Cancer, Lung cancer, Diagnosis and staging, Lymph node, Genetics.

(*J Thorac Oncol.* 2006;1: 622–628)

The optimal treatment of lung cancer relies on accurate disease staging, which is based on tumor size, regional nodal involvement, and the presence of metastasis. In particular, the extent of regional lymph node (LN) metastasis is the most important prognostic factor for those patients with resectable lung cancer. The cure rate following surgery is considerably lower if there is any LN involvement (<50% for stage II disease and 30% for stage IIIA disease).[1]

Although computed tomography (CT) has been widely used for the preoperative evaluation of tumor size and the invasion of adjacent structures, many studies have shown that CT has limitations for the staging of lung cancer because of its low reliability for LN staging.[2–5] The predictability of hilar or peribronchial LN metastasis (N1 disease) is even lower than that of mediastinal LN metastasis (N2 disease).[6] Recently, positron emission tomography (PET) using ^{18}F-fluorodeoxyglucose has been reported to increase the diagnostic accuracy for the differentiation of benign and malignant lesions and to improve the identification of nodal metastasis. However, it has been reported that ^{18}F-fluorodeoxyglucose PET is significantly less specific in depicting LN metastasis, especially in geographic areas with endemic granulomatous disease.[7]

Thus, to improve the cure rate, we need a more reliable tool with which to assess preoperative LN metastasis. The mechanisms by which malignant tumors leave the primary tumor site, invade the lymphatic system, and metastasize to the regional LNs are poorly understood. However, we are able to acquire information concerning which genes are differentially expressed according to such clinical parameters as LN metastasis, distant metastasis, recurrence, and survival.[8–14] Therefore, analysis of the differentially expressed genes between LN metastatic tumors and nonmetastatic tumors might allow researchers a better understanding of the pathogenesis of LN metastasis.

Obtaining mRNA expression profiles is an appropriate method for measuring the levels of expression for all genes in cancer. Along with clinical features, it may help to determine prognostic subgroups of cancer. It has been reported that cDNA microarray analysis could classify the subgroups of lung adenocarcinomas according to gene expression profiles and help create a risk index for survival.[12] Moreover, studies have used mRNA expression profiles as an adjunct to clinical decision-making. For example, several recent reports have suggested that gene expression profiles can predict recurrence of Dukes' stage B colon cancers,[10] diagnose LN metastasis from primary head and neck squamous cell carcinoma,[15] and predict the therapeutic response to docetaxel in patients with breast cancer.[16]

In this study, we attempted to deduce a group of genes from the analysis of mRNA expression profiles of the tumor tissues of NSCLC patients with or without LN metastasis, and create a classification model with the genes and an appropriate algorithm. Then, we investigated whether the model could predict the classification of regional LN metastasis involvement in another set of new NSCLC patients with reasonably high accuracy.

PATIENTS AND METHODS

Patients

Frozen tissues were collected from 302 patients among a total of 923 patients with primary NSCLC who underwent lung resection from January of 1995 to December of 2001 at the Samsung Medical Center, where the following patients were excluded: those with small cell or low-grade malignancy histology such as adenoid cystic carcinoma or atypical carcinoid, multiple pulmonary tumor nodules, preoperative neoadjuvant treatment for pathologically proven mediastinal LN metastases, chest or mediastinal invasion without LN metastasis (T3N0), or malignant pleural effusion, and patients who underwent an incomplete resection or who were without mediastinal node dissection. Frozen tissues were further selected by excluding the ones with less than 90% tumor cells in the microdissected specimen and the ones with degraded RNAs. As a result, frozen tissue specimens from 112 patients were used for the study. The clinical profiles of those patients consisting of 70 LN-negative patients and 42 LN-positive patients are shown in Table 1. Histologic types of samples were 48 adenocarcinomas, 58 squamous cell carcinomas, and six large cell carcinomas; 46 patients showed recurrence.

TABLE 1. Clinical Information of 112 Patients with Non-small Cell Lung Cancer

		TNM Stage I		TNM Stage II and III	
	Total	N0	N+	N0	N+
Histology					
Adenocarcinoma	48	34	0	14	0
Squamous cell carcinoma	58	32	0	26	0
Large cell carcinoma	6	4	0	2	0
Smoking history					
Nonsmoker	33	22	0	11	0
Smoker	77	46	0	31	0
Unknown	2	2	0	0	0
Sex					
Female	29	20	0	9	0
Male	83	50	0	33	0
Differentiation					
Well	19	15	0	4	0
Moderate	57	34	0	23	0
Poor	32	17	0	15	0
Unknown	4	4	0	0	0
Recurrence					
Recurrent	46	24	0	22	0
Nonrecurrent	66	46	0	20	0
Current status					
Alive	73	54	0	19	0
Dead	39	16	0	23	0
Operation Name					
Pneumonectomy	19	2	10	2	5
Bilobectomy	7	3	3	0	1
Lobectomy	82	44	14	20	4
Wedge resection	4	1	2	0	1

TNM, tumor, node, metastasis.

Tissue Preparation and RNA Extraction

Systematic LN dissection was performed for all cases by thoracotomy after routine mediastinoscopy. All the LNs encountered were removed from the American Thoracic Society LN map areas 11, 10, 9, 8, 7, 4, 3, and 2 for the tumors of the right lung; and from areas 11, 10, 9, 8, 7, 6, and 5 for the tumors of the left lung.[17] An average of six (range, five to eight) mediastinal stations were dissected. All the surgically removed tissues were immediately examined by experienced pathologists. One or two pieces (5×5 mm) of tumor tissue from the periphery of the tumor mass—to avoid the necrotic region—were immediately stored at −80°C until retrieved for the study.

The medical records were reviewed retrospectively, and the available hematoxylin and eosin slides of the LNs were reviewed by board-certified pathologists. A total of 2188 LNs were dissected from all the patients (the number of LNs per patient, 21.3 ± 10.6; range, 8–54), and 152 nodes were positive for metastasis.

Three hundred two frozen tissues were microdissected and lightly stained with hematoxylin to improve visualization. After staining, the dried frozen tissues and the tumor tissue containing normal cells that could not be microdissected were excluded. One hundred sixty-three tissues were used for RNA extraction, and the tissues having degraded RNA were removed. Finally, 112 tumor tissue specimens were used for the oligonucleotide microarray experiment. Each microdissected specimen for the study contained more than 90% tumor cells.

The microdissected tumor tissues were placed in 1 ml of Trizol reagent (Life Technologies, Rockville, MD) and immediately homogenized by vortexing. The total RNA was isolated according to the Trizol reagent protocol. The quality

of the total RNA was analyzed by electrophoresis using a 1% agarose gel containing 0.6 M formaldehyde and ethidium bromide. The quantity of the total RNA was analyzed using a Nanodrop spectrometer (Nanodrop Technologies, Rockland, DE).

Microarray Hybridization Analysis

Approximately 5 to 10 μg of the total RNA from the tumor tissues was used for the oligonucleotide microarray analysis (Macrogen, Seoul, Korea), and the remainder was used for reverse-transcriptase polymerase chain reaction (PCR) or real-time PCR. The Macrogen Oligo Human 10K microarray had been made by spotting 50-mer oligonucleotide probes that represented 10,416 human genes, 8032 known genes, and 2076 expressed sequence tags.[18] Each of the 112 total RNAs was hybridized with a universal human reference RNA (Stratagene, La Jolla, CA), and the oligonucleotide microarray hybridizations, scanning, and normalizations were carried out as described previously.[18]

Validation of Array Data with Real-Time Quantitative PCR

To validate the results of the oligonucleotide microarray analysis, two genes and 16 randomly selected samples were analyzed using real-time quantitative (RTQ) PCR (PCR). The same 16 poly-A RNAs used in the microarray analysis were reverse-transcribed with oligo(dT) priming (Superscript II; Life Technologies). RTQ-PCR was performed with specific primers for *BCL-3* and *GZMA*. The expression of the housekeeping gene *GAPDH* was used to normalize for variances in the input cDNA. A relative standard curve representing three 10-fold dilutions of cDNA from the mixed lung tissue samples (1:10:100:1000) was used for the linear regression analyses of the unknown samples. The cycling parameters used were 10 minutes at 95°C, 40 RTQ-PCR cycles of 15 seconds at 95°C and 60 seconds at 60°C for extension performed using an ABI Prism 7000 Sequence Detection System (Applied Biosystems, Foster City, CA) with a TaqMan Universal PCR Master Mix (Roche, Branchburg, NJ). Real-time monitoring of the PCR products was performed with the fluorescent dye FAM (Applied Biosystems). The expression levels of the specific genes were represented as their ratios against *GAPDH*. The PCR primer pairs (5′ to 3′) used for each gene were BCL-3 (catalogue No. Hs00180403_m1; Applied Biosystems) and GZMA (catalogue No. Hs00196206_m1; Applied Biosystems). Relatively high correlation between the results of microarray results and those of RTQ-PCR allowed us to proceed with the subsequent statistical analysis (data not shown).

Array Data Analysis and Classification Modeling

The microarray hybridization images were scanned and analyzed using ImaGene version 5.5 software (BioDiscovery, El Segundo, CA), and the data were normalized by the Lowess method, in which the value of the Cy-5/Cy-3 ratio was log transformed with base 2. Hierarchical clustering was performed to check whether the gene expression profiles of tumor tissues were significantly distinguishable between the presence and the absence of a patient's LN metastasis. Then, a permutation *t* test followed by multiple testing adjustment with the Westfall-Young method was conducted to select differentially expressed genes or signature.

Three algorithms (i.e., artificial neural network, decision tree, and naive Bayes) showed an acceptable level (>80%) of predictive accuracy in our case. However, to develop a classification model with a higher accuracy of LN metastasis prediction, we integrated a voting scheme in the combination of those three algorithms, where a positive or negative label for LN metastasis status was assigned when two or three individual algorithms agreed.

Validation of the Classification Model

The classification model obtained from the above training set of 79 patients was tested for its predictive accuracy by applying the test to a set of 33 new patients.

RESULTS

Location and Extent of LN Metastasis In NSCLC

Among 79 patients in our training samples, 49 (62%) were staged as pN0 and 30 (38%) showed lymphatic spread; 10 patients (33%) were staged as pN1 and 20 patients (67%) as pN2 according to the 1997 TNM classification. In 33 new samples for testing the developed model, 21 patients were pN0 and 12 patients showed lymphatic spread (N1, four patients; N2, eight patients) (Table 1). More than 2000 nodes ($n = 2349$), with an average of 21 LNs (range, 8–54) per patient, were clinicopathologically analyzed for metastatic infiltration. One hundred forty-eight LNs (6.3%) showed metastatic involvement, whereas 2201 LNs (93.7%) were free of tumor infiltration. LN involvement was present in 44% of the patients with squamous cell carcinoma and in 33% of the patients with large cell carcinoma. For patients with adenocarcinoma, the rate of LN involvement was 29%.

Distinctive Expression Profiles of Primary Tumors with LN Metastatic Involvement

We prepared mRNA expression profiles derived from primary tumors with or without LN metastatic involvement using oligonucleotide microarray and checked for the global discrepancy in gene expression between the two groups of tumors using hierarchical clustering. The results demonstrate that LN metastasis status was discriminated well, indicating the possibility of selecting a group of genes involved in LN metastasis (Figures 1 and 2).

Selection of LN Metastasis-Discriminating Genes

To identify the differentially expressed genes between patients with and without LN metastasis, we divided the 112 patients by stratified random sampling into two groups, 79 (49 without LN metastasis and 30 with LN metastasis) and 33 as the training set and the test set, respectively. Then, we analyzed the mRNA expression profiles from the training set and carried out the statistical analysis as described previously to generate a gene signature. We found that the signature

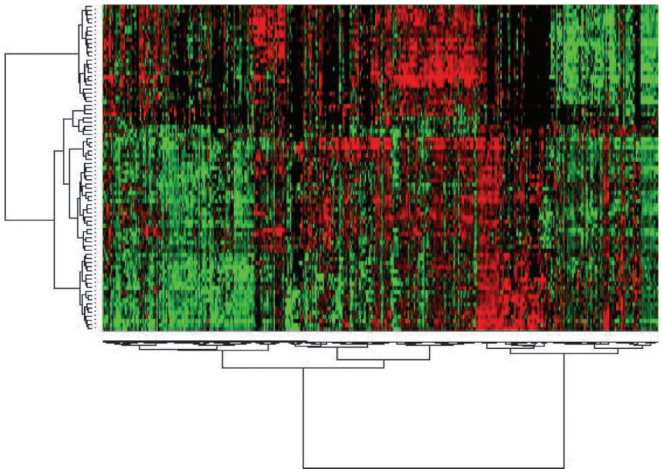

FIGURE 1. Supervised hierarchical clustering with average linkage method. Patients' samples are discriminated well by the presence or absence of LN metastasis. The *left* side in the heat map is LN metastasis-positive and the *right* side is LN metastasis-negative.

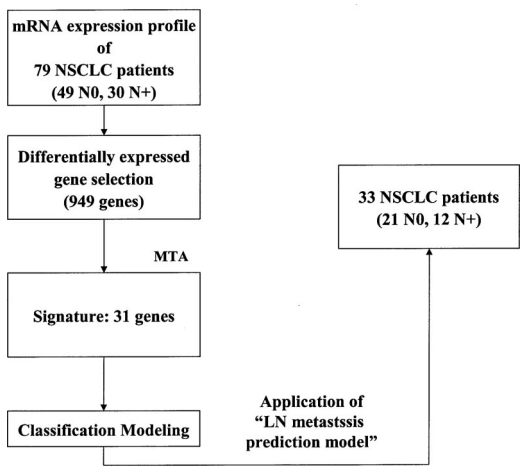

FIGURE 2. Prediction flow of LN metastasis. Gene select and construct model using 79 training samples measures model performance for the test set (33 patients). Thirty-one genes were selected as a signature from the training set, and the resulting LN metastasis prediction model was evaluated for its predictive accuracy by applying it to the test set.

consisted of 31 genes ($p < 0.05$) (i.e., 27 known genes and four expressed sequence tags), and their characteristics are summarized in Table 2.

Functional Classification of the Selected Discriminatory Genes

To obtain a brief overview of the general functional features of the LN metastasis-associated genes in NSCLC, we performed the classification and analysis of functions for the 949 genes ($p < 0.01$) that resulted from the *t* test before the multiple testing adjustment using the ArrayXPath. The results were deposited at the following website: http://biosvg.snubi.org:8080/ArrayXPath/ArrayXPath_SonDaeSoon_949.html. Specifically, 222 of 949 genes were identified in four public pathway databases (GenMAPP, PharmGKB, KEGG, and BioCarta) using ArrayXPath.[19] The results are useful for searching genes involved in cell proliferation, migration, and survival that are essential to LN metastasis in cancer cells.

Thirty-one genes as the signature for LN metastasis seem to be related to tumorigenesis and tumor suppression, which is supported by the following facts on the component genes: AATF, a recently identified human Rb-binding protein that inhibits the Rb growth suppressing function[20]; BAT3, related to apoptosis by caspase-3–mediated proteolytic activation[21]; and Cyb561, associated with senescence.[22] Further facts come from other genes: *GPC4* binds FGF2 and modulates fibroblast growth factor signal transduction[23]; *Mcm3* expression appeared to be higher in cancer cells than in normal proliferating cells of the uterine cervix and dysplastic cells[24]; *MUC5AC* is a target gene of epidermal growth factor receptor ligands in lung cancer cells, and up-regulation of this gene goes through concomitant activation of the epidermal growth factor receptor/Ras/Raf/extracellular signal-regulated kinase-signaling pathway and Sp1 binding to their promoters[25]; *SOX6* may be a potential diagnostic marker for gliomas[26]; and *TGIF2* may play an important role in the development and/or progression of some ovarian tumors through a mechanism of gene amplification.[27]

Development of the LN Metastasis Prediction Model and Its Predictive Accuracy

Next, we developed the LN metastasis prediction model using this gene signature and the modified algorithm described previously. When we tested this model for its predictive accuracy on the test set of 33 new patients, we achieved a sensitivity of 83.33% and a specificity of 85.71%, which is more accurate for the prediction of LN metastasis than obtained by any of the algorithms (Table 3). Our LN metastasis prediction model, which has been applied successfully to patients with their clinical information undisclosed, showed another practical example of being a useful prediction model as collectively structured parameters, not by one or a few individual genes involved.

DISCUSSION

Because preoperative chemotherapy plus surgery allows an increased median survival for patients with pathologically proven ipsilateral mediastinal lymph node metastasis (pathological N2 stage IIIA),[3,4] the neoadjuvant approach has been examined for earlier disease, such as stage IB, IIA, and IIB. However, contrary to the promising expectations, the outcomes of the previous studies were disappointing. In the phase II Bimodality Lung Oncology Team study performed by Pisters et al., although 53 of 94 (56%) patients had a major objective response after induction chemotherapy, only 81 patients (86%) had complete resection.[28] Considering that the candidates had T2N0, T1-2N1, or T3N0-1 staged NSCLC and more than 95% had R0 resection (i.e., complete resection with negative resection margins), this was too low a resection rate, and the unnecessary delay of surgery may have been attributable to unresectability. In addition, there was no remarkable survival benefit identified at 1 and 2 years after

TABLE 2. Differentially Expressed Genes from an Analysis of mRNA Expression Profiles of 79 Patients with Non-small Cell Lung Cancer as a Training Set (LN Metastasis Prediction Model Signature)

Gene Accession No.	Gene Description	Gene Symbol
AC005754	EST	
AF216972	EST	
AF309034	SRY (sex-determining region Y)-box 6	*SOX6*
ENSG00000084774	EST	
NM_000486	Aquaporin 2 (collecting duct)	*AQP2*
NM_000992	Ribosomal protein L29	*RPL29*
NM_001246	Ectonucleoside triphosphate diphosphohydrolase 2	*ENTPD2*
NM_001448	Glypican 4	*GPC4*
NM_001915	Cytochrome b-561	*CYB561*
NM_002262	Killer cell lectin-like receptor subfamily D, member 1	*KLRD1*
NM_002280	Keratin, hair, acidic, 5	*KRTHA5*
NM_002388	MCM3 minichromosome maintenance deficient 3 (*Saccharomyces cerevisiae*)	*MCM3*
NM_003093	Small nuclear ribonucleoprotein polypeptide C	*SNRPC*
NM_003365	Ubiquinol-cytochrome c reductase core protein I	*UQCRC1*
NM_003481	Ubiquitin specific protease 5 (isopeptidase T)	*USP5*
NM_003975	SH2 domain protein 2A	*SH2D2A*
NM_004639	HLA-B associated transcript 3	*BAT3*
NM_004711	Synaptogyrin 1	*SYNGR1*
NM_012138	Apoptosis antagonizing transcription factor	*AATF*
NM_014621	Homeo box D4	*HOXD4*
NM_014629	Rho guanine nucleotide exchange factor (GEF) 10	*ARHGEF10*
NM_015909	Neuroblastoma-amplified protein	*NAG*
NM_016153	LW-1	*LW-1*
NM_018695	erbb2 interacting protein	*ERBB2IP*
NM_020533	Mucolipin 1	*MCOLN1*
NM_021809	TGFB-induced factor 2 (TALE family homeobox)	*TGIF2*
NM_022338	Chromosome 11 open reading frame 24	*C11orf24*
NM_030673	SEC13-like 1 (*Saccharomyces cerevisiae*)	*SEC13L1*
NM_057090	Artemin	*ARTN*
XM_039877	Mucin 5, subtype B, tracheobronchial	*MUC5B*
XM_041585	EST	

EST, expressed sequence tag.

surgery, although the median survival had not yet been reached for their report on the year 2000. Most prominently, the authors mentioned that the interpretation of survival was hampered by the relative nonhomogeneity of the clinical/pathological stages of the treated patients.

In the Bimodality Lung Oncology Team study, despite performing modern CT imaging and mediastinoscopy, clinical staging was practically inaccurate, because only 16% of patients showed an equivalent clinical and pathological stage.[28] If induction chemotherapy might be effective only for patients with regional lymph node metastasis, surgery for patients with stage I NSCLC might have been unnecessarily delayed and the chance for survival was inadvertently taken away.

Therefore, the accurate preoperative assessment of LN involvement for patients with NSCLC is crucial for treatment planning. Imaging modalities such as CT, magnetic resonance imaging, and transesophageal endosonography provide limited predictive value in differentiation between benign and malignant nodes. McLoud et al. have shown that CT gives a sensitivity of 64% and a specificity of 62% when extensive sampling of nodes with a diameter of 1 cm as the upper limit is allowed.[2] PET and endoscopic ultrasound provide sensitivities of 84% and 78% and specificities of 89% and 71%, respectively, in diagnosis of LN metastasis in NSCLC.[29]

In the pretherapeutic LN staging of NSCLC, it has long been assumed that there exists a correlation between node size and metastatic infiltration. However, Prenzel et al. showed a lack of a statistically significant relationship between the size of the LNs and the likelihood of malignancy in a study of 2891 regional LNs from 256 patients.[6] Similar disappointing results were reported by Vogel et al. after investigation of 162 mediastinal LNs in 83 patients.[30] Despite these findings, CT still has been the method of choice for assessing LN metastasis in lung cancer.

In contrast, some researchers have tried to identify patients with LN metastasis by using biological methods. Analysis of serum vascular endothelial growth factor, microsatellite instability, or peanut agglutinin-binding carbohydrates in tumor suggested that these factors are related to

TABLE 3. Performance of the LN Metastasis Prediction Model in Four Classification Algorithms

	10-Fold Cross-Validation in Training Set					Prediction in Test Set of New Patients				
Classification Algorithm		**mN0**	**mN+**	**Total**	**Performance**		**mN0**	**mN+**	**Total**	**Performance**
Support vector machine	pN0	29	20	49	Sensitivity, 70.00%;					
	pN+	9	21	30	specificity, 59.18%					
					accuracy, 63.29%					
	Total	38	41	79						
Neural network										
	pN0	44	5	49	Sensitivity, 80.00%;	pN0	18	3	21	Sensitivity, 58.33%;
	pN+	6	24	30	specificity, 89.80%	pN+	5	7	12	specificity, 85.75%
					accuracy, 86.08%					accuracy, 75.76%
	Total	50	29	79		Total	23	10	33	
Decision tree										
	pN0	46	3	49	Sensitivity, 83.33%;	pN0	15	6	21	Sensitivity, 91.67%;
	pN+	5	25	30	specificity, 93.88%	pN+	1	11	12	specificity, 71.43%
					accuracy, 89.87%					accuracy, 78.79%
	Total	51	28	79		Total	16	17	33	
Naive Bayes										
	pN0	43	6	49	Sensitivity, 100%;	pN0	17	4	21	Sensitivity, 75.00%;
	pN+	0	30	30	specificity, 87.76%	pN+	3	9	12	specificity, 80.95%
					accuracy, 92.41%					accuracy, 78.79%
	Total	43	36	79		Total	20	13	33	
Combined model (neural network+ decision	pN0	44	5	49	Sensitivity, 93.33%;	pN0	18	3	21	Sensitivity, 83.33%;
tree + naive Bayes)	pN+	2	28	30	specificity, 89.80%	pN+	2	10	12	specificity, 85.71%
					accuracy, 91.14%					accuracy, 84.85%
	Total	46	33	79		Total	20	13	33	

nodal metastasis in lung cancer.[31–34] However, tumor metastasis is a complex process that involves many interactions between tumor cells and the extracellular matrix so that some individual genetic changes discovered cannot be generalized and are impractical to apply to clinical situations.

Extensive analysis of mRNA expression profiles or protein expression profiles on a scale and with the sensitivity unattainable by conventional techniques has been made possible by the improvement in high-throughput technologies. For lung cancer, these clinical genomic or proteomic approaches have identified subgroups of tumors that differ in terms of tumor type, histologic subclass, and patient survival, allowing prediction of regional LN metastasis with a small number of patients.[12]

The result of this study showed the possibility that gene signatures from mRNA expression profiling can predict LN metastasis with high accuracy for NSCLC. However, to be used in clinical practice such as in planning of preoperative chemotherapy, biopsy specimens should be collected from the primary tumor lesion and then analyzed for gene expression. Successful application of such a technique has already been reported by Borczuk et al.[35]

Although multigene signatures from mRNA expression profiling can predict LN metastasis, the extent of the metastasis, such as number of metastatic LNs or the delineation between N1 and N2, cannot be identified because the distinction between N1 and N2 is made by anatomical, not biological, criteria. Therefore, when preoperative analysis of gene expression of the biopsy specimen suggests possible LN metastasis, mediastinoscopy is imperative to discern N1 disease from N2 or N3 disease.

Even though we could select a group of genes from mRNA expression profiling, some individual genes in the group might not be useful for predicting LN metastasis as mentioned earlier because of the complexity of most types of cancer and the compound nature of gene functions. Thus, it is desirable that gene sets representing the characteristic gene expression profiles of LN metastasis are selected and their interactions interpreted as a whole in a reasonably acceptable manner. Furthermore, because our ultimate goal is prediction of LN metastasis for new patients with the model developed from the fixed data, the model's robustness is essential for the classification algorithm. For this reason, we decided to seek a modified model constructed from a series of known classification algorithms. We calculated the accuracy by 10-fold cross-validation for each algorithm and then selected and combined three independent algorithms (ANN, DT, and NB), showing a greater than 85% accuracy. We found the robustness in the resultant modified algorithm, which is good for accuracy in sound judgment.

CONCLUSIONS

In our study with the long-term goal of validating our gene set as a whole and not as individual genes for their effective prospective discrimination of LN metastasis, we

selected genes by a modified algorithm that had been derived by combining the last three of the above algorithms with a voting scheme. Our selected gene sets and their modeled interactions significantly improved the predictive accuracy of LN metastasis in the training set of patients compared with the other unmodified algorithms, and showed the prediction with high accuracy in the test set of patients. For ideal validation of this classification model, we should apply this model to another set of mRNA expression profiles from a new set of patients. This would provide us with an additional chance to refine the current model to the integrated version, allowing for even more accurate prediction.

ACKNOWLEDGMENTS

This work was partly funded by Korea Institute of Science & Technology Evaluation and Planning grant M6-0301-00-0017 and Samsung Biomedical Research grant C-A6-411-1.

REFERENCES

1. Vansteenkiste JF, De Leyn PR, Deneffe GJ, Lerut TE, Demedts MG. Clinical prognostic factors in surgically treated stage IIIA-N2 non-small cell lung cancer: Analysis of the literature. *Lung Cancer* 1998;19:3–13.
2. McLoud TC, Bourgouin PM, Greenberg RW, et al. Bronchogenic carcinoma: Analysis of staging in the mediastinum with CT by correlative lymph node mapping and sampling. *Radiology* 1992;182:319–323.
3. Townsend DW. A combined PET/CT scanner: The choices. *J Nucl Med* 2001;42:533–534.
4. Beyer T, Townsend DW, Blodgett TM. Dual-modality PET/CT tomography for clinical oncology. *Q J Nucl Med* 2002;46:24–34.
5. Scott WJ, Gobar LS, Terry JD, Dewan NA, Sunderland JJ. Mediastinal lymph node staging of non-small-cell lung cancer: A prospective comparison of computed tomography and positron emission tomography. *J Thorac Cardiovasc Surg* 1996;111:642–648.
6. Prenzel KL, Monig SP, Sinning JM, et al. Lymph node size and metastatic infiltration in non-small cell lung cancer. *Chest* 2003;123:463–467.
7. Yoon YC, Lee KS, Shim YM, Kim BT, Kim K, Kim TS. Metastasis to regional lymph nodes in patients with esophageal squamous cell carcinoma: CT versus FDG PET for presurgical detection prospective study. *Radiology* 2003;227:764–770.
8. Hao X, Sun B, Hu L, et al. Differential gene and protein expression in primary breast malignancies and their lymph node metastases as revealed by combined cDNA microarray and tissue microarray analysis. *Cancer* 2004;100:1110–1122.
9. Kikuchi T, Daigo Y, Katagiri T, et al. Expression profiles of non-small cell lung cancers on cDNA microarrays: Identification of genes for prediction of lymph-node metastasis and sensitivity to anti-cancer drugs. *Oncogene* 2003;22:2192–2205.
10. Wang Y, Jatkoe T, Zhang Y, et al. Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. *J Clin Oncol* 2004;22:1564–1571.
11. Bertucci F, Salas S, Eysteries S, et al. Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. *Oncogene* 2004;23:1377–1391.
12. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002;8:816–824.
13. Weiss MM, Kuipers EJ, Postma C, et al. Genomic profiling of gastric cancer predicts lymph node status and survival. *Oncogene* 2003;22:1872–1879.
14. Nagata M, Fujita H, Ida H, et al. Identification of potential biomarkers of lymph node metastasis in oral squamous cell carcinoma by cDNA microarray analysis. *Int J Cancer* 2003;106:683–689.
15. Roepman P, Wessels LF, Kettelarij N, et al. An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. *Nat Genet* 2006;37:182–186.
16. Chang JC, Wooten EC, Tsimelzon A, et al. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* 2003;362:362–369.
17. Murray JG, Breatnach E. The American Thoracic Society lymph node map: A CT demonstration. *Eur J Radiol* 1993;17:61–68.
18. Kono T, Obata Y, Wu Q, et al. Birth of parthenogenetic mice that can develop to adulthood. *Nature* 2004;428:860–864.
19. Chung HJ, Park CH, Han MR, et al. ArrayXPath II: Mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using Scalable Vector Graphics. *Nucleic Acids Res* 2006;33:W621–W626.
20. Burgdorf S, Leister P, Scheidtmann KH. TSG101 interacts with apoptosis-antagonizing transcription factor and enhances androgen receptor-mediated transcription by promoting its monoubiquitination. *J Biol Chem* 2004;279:17524–17534.
21. Wu YH, Shih SF, Lin JY. Ricin triggers apoptotic morphological changes through caspase-3 cleavage of BAT3. *J Biol Chem* 2004;279:19264–19275.
22. Kang MK, Kameta A, Shin KH, Baluda MA, Kim HR, Park NH. Senescence-associated genes in normal human oral keratinocytes. *Exp Cell Res* 2003;287:272–281.
23. Galli A, Roure A, Zeller R, Dono R. Glypican 4 modulates FGF signalling and regulates dorsoventral forebrain patterning in Xenopus embryos. *Development* 2003;130:4919–4929.
24. Ishimi Y, Okayasu I, Kato C, et al. Enhanced expression of Mcm proteins in cancer cells derived from uterine cervix. *Eur J Biochem* 2003;270:1089–1101.
25. Perrais M, Pigny P, Copin MC, Aubert JP, Van Seuningen I. Induction of MUC2 and MUC5AC mucins by factors of the epidermal growth factor (EGF) family is mediated by EGF receptor/Ras/Raf/extracellular signal-regulated kinase cascade and Sp1. *J Biol Chem* 2002;277:32258–32267.
26. Ueda R, Yoshida K, Kawakami Y, Kawase T, Toda M. Expression of a transcriptional factor, SOX6, in human gliomas. *Brain Tumor Pathol* 2004;21:35–38.
27. Starback P, Wraith A, Eriksson H, Larhammar D. Neuropeptide Y receptor gene y6: Multiple deaths or resurrections? *Biochem Biophys Res Commun* 2000;277:264–269.
28. Pisters KM, Ginsberg RJ, Giroux DJ, et al. Induction chemotherapy before surgery for early-stage lung cancer: A novel approach. Bimodality Lung Oncology Team. *J Thorac Cardiovasc Surg* 2000;119:429–439.
29. Toloza EM, Harpole L, McCrory DC. Noninvasive staging of non-small cell lung cancer: A review of the current evidence. *Chest* 2003;123(1 Suppl):137S–146S.
30. Vogel P, Daschner H, Lenz J, Schafer R. Correlation of lymph node size and metastatic involvement of lymph nodes in bronchial cancer. *Langenbecks Arch Chir* 1990;375:141–144.
31. Kalluri R. Basement membranes: Structure, assembly and role in tumour angiogenesis. *Nat Rev Cancer* 2003;3:422–433.
32. Ferrara N, Gerber HP, LeCouter J. The biology of VEGF and its receptors. *Nat Med* 2003;9:669–676.
33. Cavallaro U, Christofori G. Cell adhesion and signalling by cadherins and Ig-CAMs in cancer. *Nat Rev Cancer* 2004;4:118–132.
34. Stacker SA, Achen MG, Jussila L, Baldwin ME, Alitalo K. Lymphangiogenesis and cancer metastasis. *Nat Rev Cancer* 2002;2:573–583.
35. Borczuk AC, Shah L, Pearson GD, et al. Molecular signatures in biopsy specimens of lung cancer. *Am J Respir Crit Care Med* 2004;170:167–174.