Journal of Molecular Diagnostics, Vol. 6, No. 1, February 2004
Copyright © American Society for Investigative Pathology
and the Association for Molecular Pathology

Classification of Individual Lung Cancer Cell Lines Based on DNA Methylation Markers

Use of Linear Discriminant Analysis and Artificial Neural Networks

Alberto M. Marchevsky,[*] Jeffrey A. Tsou,[†] and Ite A. Laird-Offringa[†]

From the Department of Pathology and Laboratory Medicine,[*] Cedars-Sinai Medical Center, Los Angeles, California; and Departments of Surgery and of Biochemistry and Molecular Biology,[†] Norris Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, California

The classification of small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) can pose diagnostic problems due to inter-observer variability and other limitations of histopathology. There is an interest in developing classificatory models of lung neoplasms based on the analysis of multivariate molecular data with statistical methods and/or neural networks. DNA methylation levels at 20 loci were measured in 41 SCLC and 46 NSCLC cell lines with the quantitative real-time PCR method MethyLight. The data were analyzed with artificial neural networks (ANN) and linear discriminant analysis (LDA) to classify the cell lines into SCLC or into NSCLC. Models used either data from all 20 loci, or from five significant DNA methylation loci that were selected by a step-wise back-propagation procedure (*PTGS2*, *CALCA*, *MTHFR*, *ESR1*, and *CDKN2A*). The data were sorted randomly by cell line into 10 different data sets, each with training and testing subsets composed of 71 and 16 of the cases, respectively. Ten ANN models were trained using the 10 data sets: five using 20 variables, and five using the five variables selected by step-wise back-propagation. The ANN models with 20 input variables correctly classified 100% of the cell lines, while the models with only five variables correctly classified 87 to 100% of cases. For comparison, 10 different LDA models were trained and tested using the same data sets with either the original data or with logarithmically transformed data. Again, half of the models used all 20 variables while the others used only the five significant variables. LDA models provided correct classifications in 62.5% to 87.5% of cases. The classifications provided by all of the different models were compared with kappa statistics, yielding kappa values ranging from 0.25 to 1.0. We conclude that ANN models based on DNA methylation profiles can objectively classify SCLC and NSCLC cells lines with substantial to perfect concordance, while LDA models based on DNA methylation profiles provide poor to substantial concordance. Our work supports the promise of ANN analysis of DNA methylation data as a powerful approach for the development of automated methods for lung cancer classification. *(J Mol Diagn 2004, 6:28–36)*

There is an increasing interest in the use of molecular data from human neoplasms for diagnostic and prognostic purposes.[1] A prodigious amount of data regarding the molecular composition of a variety of neoplasms is being collected from tissue samples. To our knowledge, it is not yet clear how to analyze and interpret molecular data to yield diagnostic and prognostic information that will complement or replace the experience acquired over many years from the study of neoplasms with microscopy and other morphology-based methods. This problem is not entirely new to pathologists, as similar questions were faced in the past after the development of computerized image analysis systems that could process a large number of spatial and photometric features from neoplastic cells.[2–4] Various algorithms have been developed to accurately classify tumors based on morphometric and cytometric features.[2,3,5–10] Although most of these methods are not used in current pathology practice, as they are time consuming and difficult to standardize, they have led to the development of analytical instruments that are

Supported by a generous donation from Eri and Mary Lou Mettler, the Lung Cancer Methylation Fund, and the Mesothelioma Applied Research Foundation (MARF) (J.A.T. and I.A.L.O.)

Dr. Laird-Offringa is a shareholder of Epigenomics AG, which has a commercial interest in developing DNA methylation profiles for disease detection and diagnosis. None of the work performed in Dr. Laird-Offringa's laboratory is or has been supported by Epigenomics.

Accepted for publication October 15, 2003.

Address reprint requests to Ite A. Laird-Offringa, USC/Norris Cancer Center, 1441 Eastlake Avenue, Room NOR6420, Los Angeles, CA 90089-9176. E-mail: ilaird@usc.edu.

currently approved by the Food and Drug Administration for the routine screening of gynecologic smears.[11–13]

The World Health Organization and the International Association for the Study of Lung Cancer (World Health Organization/IASLC) recognize a large variety of malignant epithelial neoplasms of the lung.[14] The histopathological classification of some of these tumors is subject to considerable inter-observer variability. Indeed, inter-observer variability can be as high as 50% among expert pathologists classifying poorly differentiated lung neoplasms.[15] Malignant lung neoplasms are currently clustered in practice into two groups with distinct clinicopathological features: small cell carcinoma (SCLC) and non-small cell carcinoma (NSCLC). The classification of lung neoplasms into these two major groups is reproducible in approximately 90% of cases but the distinction between SCLC and NSCLC can be difficult when limited diagnostic material is available (eg, from a fine needle aspirate).[16] Diagnostic problems include the differential diagnosis between SCLC and the small cell variant of squamous cell carcinoma, atypical carcinoid tumors, and large cell neuroendocrine carcinoma.[17,18] Because of these problems, molecular markers specific for the various types of lung cancer would be very valuable. Molecular markers based on DNA have the advantage of allowing signal amplification by polymerase chain reaction (PCR), so that only limited material is required.[19] A very promising alteration of DNA that is commonly found in cancer is DNA methylation.[20–22]

DNA methylation is an epigenetic modification of DNA occurring in most living organisms.[23–25] In mammals, it consists of the addition of a methyl group to the carbon-5 position of cytosine and occurs primarily in palindromic CpG dinucleotides. DNA methylation is essential for mammalian development and plays a significant role in genomic imprinting, modulation of chromatin structure, and X-chromosome inactivation. CpG dinucleotides are frequently found in clusters called CpG islands in the promoter regions of genes, and these CpG islands are usually not methylated in normal cells. The unmethylated state of promoter CpG islands is associated with transcriptional activity. However, in cancer cells, certain promoter CpG islands exhibit increased methylation (hypermethylation). This local hypermethylation is associated with gene silencing, and can contribute to carcinogenesis when it occurs in the promoter regions of genes that regulate cell growth or other cellular functions associated with cell survival.[25–28] DNA methylation profiles can vary between cancers from different organs, suggesting that such profiles could be useful diagnostic tools.[24,29–31] Hypermethylation of a variety of genes in lung neoplasms has been reported (reviewed in[22]). *APC*, *CDKN2A*/ *p16*$^{\text{INK4A,}}$ *CDH13*, *RARB*, and *RASSF1A* have been found to be frequently hypermethylated in lung cancer by at least two independent studies (reviewed in[22]). In contrast, certain genes commonly methylated in other types of cancer, such as *ARF* and *CDKN2B*, are infrequently hypermethylated in lung neoplasms (reviewed in[22]).

Besides showing differences between cancers from different organs, methylation profiles are also thought to be distinct in different histological subtypes of cancer from the same organ. This is supported by our recent analysis of the methylation status of 23 loci in 47 NSCLC and 44 SCLC cell lines.[20] The methylation levels of 7 of 23 genes analyzed (*PTGS2*, *CALCA*, *MTHFR*, *ESR1*, *MGMT*, *MYOD1*, and *APC*) differed significantly between SCLC and NSCLC cell lines, supporting the idea that it may be possible to accurately distinguish SCLC from NSCLC using DNA methylation profiles.

To our knowledge, there is no current consensus on how to analyze multivariate molecular data for the identification of specific tumor cell types. Multiple studies done during the past four decades have developed objective diagnostic methods based on multivariate analysis of morphometric and densitometric data from neoplastic cells.[2–9,32–38] Such classificatory methods can be based on supervised or non-supervised paradigms.[4] “Unsupervised” methods attempt to explore or “mine” the data to detect relationships without any *a priori* information regarding the classification of the data. They identify groups of data elements that are highly correlated with each other and are more frequent in a particular group of a data set. These groups are subsequently compared with the diagnoses to explore whether a significant subset of variables correlates with a particular group of neoplasms. Studies of molecular data collected with high-throughput systems have mostly applied unsupervised methods such as hierarchical cluster analysis.[20] We previously used hierarchical clustering to analyze the data set used in the current study; this approach was able to correctly group SCLC with a specificity and sensitivity of 78%.[20] This percentage it is not sufficiently high to justify use of the tested DNA methylation markers in the clinical setting. However, the method used (hierarchical clustering) may not be optimal for the analysis of DNA methylation data, which is not normally distributed. We therefore undertook an exploration of alternative methods of classification.

In contrast to unsupervised methods, “supervised” methods are trained using data that is labeled with the correct answer. The trained models are thereafter tested with data that were not used during training. Most studies of morphometric and photometric data of neoplastic cells have used supervised methods, such as ad-hoc algorithms, linear discriminant analysis (LDA), logistic regression, artificial neural networks (ANN), rule-based expert systems, Bayesian belief networks, and others.[4] One of the important strengths of the “supervised” approach is that the trained function can be applied to classify unknown cases, as would occur when making a diagnosis of a new sample.[5–7,10,32] ANN trained with molecular data have been recently used by Khan and associates[39] in a study of classification of small round blue-cell tumors. These models classified correctly all their test cases into four diagnostic categories. Here we explore how supervised classification methods might be applied to lung cancer diagnosis based on DNA methylation profiles, using the previously obtained methylation data from 87 lung cancer cell lines as a model system. We compare the utility of linear discriminant analysis and artificial neural networks as classificatory tools of DNA methylation profiles, in an effort to develop diagnostic models that could distinguish SCLC from NSCLC.

Materials and Methods

Detection and Quantitation of CpG Island DNA Hypermethylation Profiles in SCLC and NSCLC Cell Lines

The tumor cell lines used in this study have been well characterized and were initiated by Gazdar and co-workers[20,40] at the National Cancer Institute and Hamon Cancer Center. DNA from 47 NSCLC and 44 SCLC cell lines was subjected to bisulfite conversion, which embeds methylation data into the DNA sequence by converting unmethylated Cs to U.[41] Methylation analysis was performed for 23 loci using the fluorescence-based, real-time PCR assay MethyLight, which utilizes primers and probes specifically designed to hybridize to fully methylated sequences.[42,43] The percentage methylated reference (PMR) for each locus was calculated by dividing the *GENE*:reference ratio of a sample by the *GENE*:reference ratio of highly methylated *Sss*I-treated human sperm DNA and multiplying by 100.[42,43] GENE methylation levels were normalized independently using each of the two reference reactions [one with β-actin (*ACTB*) and one with type 2 collagen (*COL2A1*)], and the mean of the resulting PMR was used as the final PMR value. The analysis resulted in over 2200 data points, representing DNA methylation levels at the 23 loci in the cell lines. These data, sorted into selected classes, was previously used for a study of hierarchical cluster analysis, which allowed classification of the cell lines into SCLC and NSCLC with 78% sensitivity and specificity.[20] For this study, loci that were uniformly negative (a PMR value of 0.00000) were removed from the data set, leaving 20 loci (*TYMS*, *TGFBR2*, *THBS1*, *CDKN2B*, *TIMP3*, *PTGS2*, *CALCA*, *MGMT E/I*, *MTHFR*, *ESR2*, *CDH1*, *HIC1*, *GSTP1*, *PGR*, *AR1*, *APC*, *MGMTPRO*, *MYOD1*, *CDKN2A*, and *ESR1*). Cell lines for which methylation information was incomplete were also excluded, leaving 87 cell lines (Table 1).

Selection of Significant Variables and Establishment of Data Sets

The PMR data from 20 loci was subjected to backward step-wise analysis to eliminate the variables that did not contribute to classification. Using automated step-wise backward elimination, a model with five independent variables (*ESR1*, *MTHFR*, *PTGS2*, *CDKN2A*, and *CALCA*) was selected. The data were next sorted randomly by cell line into 10 different data sets (numbered 1 to 10), each with training and testing subsets composed of 71 and 16 of the cases, respectively. SCLC and NSCLC cell lines were distributed in training and test sets at ~80% and ~20%, respectively.

Analysis with Artificial Neural Networks and Cross-Validation of Results

Artificial neural networks (ANN) are computerized mathematical models designed to emulate the architecture of the brain.[44] Data are divided into processing units or neurons. Neurons are organized in parallel layers: input, hidden (single or multiple), and output. Each neuron connects to all neurons of another layer but not to those in the same layer. Neurons process the data using a variety of mathematical functions; a probabilistic function with genetic algorithms was selected for this study, as previously reported.[5,32–34] All values are normalized to numbers ranging from 0 to 1. ANN were designed according to the number of variables being analyzed, with either 20 or five input neurons. Each input neuron processes the data from a DNA methylation locus. The hidden layer contained 89 neurons, the default value selected automatically by the software. The output layer consisted of two output neurons, one to classify the SCLC category and the other for the NSCLC category. The neurons in the input layer process the normalized values of each variable using a function. The values generated by each input neuron are “transmitted” to each hidden layer neuron. Each hidden layer neuron receives the numerical input from each of the input layer neurons and calculates two outputs. Each of the two output neurons “receives” one of the two outputs generated by each hidden layer neurons and calculates output values ranging from 0 to 1. If the value generated by an output neuron is larger than 0.5, the neuron is “activated” to yield its classification. Only one of the output neurons is “activated” during each training or testing cycle, resulting in the classification of a sample as NSCLC or SCLC.

The 10 training/testing sets were analyzed with ANN (NeuroShell2, Ward Systems, Frederick, MD). The ANN were trained using each of the 10 data sets, in an effort to emulate at a cross-validation procedure.[5,34,35] The test data were not used during ANN training. Five ANN models used all 20 variables as input neurons. The other five ANN used *PTGS2*, *CALCA*, *MTHFR*, *ESR1*, and *CDKN2A* as input neurons.

Analysis of the Data with Linear Discriminant Analysis

The data were analyzed with multivariate linear discriminant analysis (LDA) (Systat 10.0 SPSS, Chicago, IL) using the cell type (SCLC and NSCLC) as the dependent variable and the methylation levels as independent variables.[45] Prior probabilities were 0.50 for each group. The same 10 data sets used for ANN were used for the LDA. To address possible problems with non-linearity of the methylation data (loci can be very highly methylated in some cell lines, and unmethylated in others), the PMR values for the same 10 sets described above were also transformed logarithmically, using the formula log(1+variable), and analyzed with LDA. Linear discriminant functions were derived. A classification matrix in-

Table 1. PMR Values* for 20 Methylation Loci† in 87 Lung Cancer Cell Lines

Sample‡	Cell type	*TYMS*	*TGFBR2*	*THBS1*	*CDKN2B*	*TIMP3*	*PTGS2*	*CALCA*	*MGMT* E/I	*MTHFR*
H1770	NSCLC	0	0	0	4.66	0	2.80	0.62	10.96	71.16
H2170	NSCLC	0	0	0	0	0	0.01	0	11.71	106.74
H1755	NSCLC	0	0	0	0	0	71.13	0	0	156.72
H1693	NSCLC	0	0	0	0	0	0.22	2.76	0.02	2.83
H2087	NSCLC	0	0	0	0	0	1.20	2.17	0.47	30.48
H0522	NSCLC	0	0	1.41	0	0	5.97	10.73	5.19	10.27
H2347	NSCLC	0	0	0	0	77.40	13.55	65.38	1.76	0.36
H0023	NSCLC	0	0	0	0	0	0.01	20.98	9.10	115.43
HCC366	NSCLC	0	0	0	0	0.66	0.60	12.98	9.50	81.17
H1793	NSCLC	0	0	0	0	0	0.10	12.27	4.31	56.39
H1155	NSCLC	0	0	0	0	0	5.23	8.43	0	125.79
H0358	NSCLC	0	0	0	0	106.53	0.01	0	1.20	6.17
H0820	NSCLC	0	0	0	0	0	11.50	93.87	17.66	0
H1355	NSCLC	0	0	0	0	0.49	0.06	7.00	0.36	67.99
H0647	NSCLC	0	0	0	4.42	7.69	0.02	0.01	0.12	4.56
H0441	NSCLC	0	0	0	0	98.89	0	53.04	3.80	37.85
H0460	NSCLC	0	0	0	0	0	1.15	29.24	2.31	83.63
H2122	NSCLC	0	0	0	0	0	5.40	74.65	0	76.37
H1437	NSCLC	0	0	0	0	0.05	0.20	46.46	3.11	76.50
H1703	NSCLC	0	0	0	0	0	71.87	71.61	0.02	82.78
H2073	NSCLC	0	0	0	0	0.57	4.74	49.15	0.95	11.56
HCC044	NSCLC	0	0	0	0	0.28	0	0	0.05	35.12
H0322	NSCLC	0	0	0	0	108.43	6.53	87.90	16.98	90.29
H1435	NSCLC	0	0	0	0	0	4.07	5.19	2.31	39.73
H0838	NSCLC	0	0	0	0	0	1.81	50.47	0.21	15.59
H2077	NSCLC	0	0	0	0	0	2.37	169.41	7.48	157.47
H0125	NSCLC	0	0	0	0	0	0.34	17.97	1.37	34.04
H1792	NSCLC	0	0	0	2.29	0	0	5.78	3.50	107.85
H1395	NSCLC	0	0	0	0	59.23	0	3.23	34.95	2.25
H0720	NSCLC	0	0	0	57.47	167.04	120.10	150.61	38.91	178.23
H1993	NSCLC	0	0	0	0	0	0	29.95	22.57	110.01
H0920	NSCLC	0	0	0	0	0	0.32	6.56	0	103.74
HCC461	NSCLC	0	0	0	0	0	0	19.69	0	112.49
H2009	NSCLC	0	0	0	0	125.76	0.31	65.81	5.68	22.68
HCC015	NSCLC	0	0	0	0	0	1.88	11.40	0.05	75.30
H1944	NSCLC	0	0	0	0	0	0.01	11.08	3.55	12.48
H1264	NSCLC	0	0	0	0	0	0	0	0	31.52
H0157	NSCLC	0	0	0	0	0	0	0	6.69	114.68
HCC515	NSCLC	0	0	0	0	0	0	86.48	0	45.79
H1573	NSCLC	0	0	0	7.09	0	0.06	0	0	39.24
H2106	NSCLC	0	0	0	0	0	0	0	0.16	134.55
H1334	NSCLC	0	0	0	0	131.88	0.54	76.36	15.02	65.95
H0727	NSCLC	0	0	0	0	0	0.98	0	1.09	178.11
H0661	NSCLC	0	0	0	0	0	4.19	4.39	10.30	112.06
H1648	NSCLC	0	0	0	0	0	0.20	0	8.10	158.77
H1299	NSCLC	0	0	0	0	0	94.33	1.97	9.36	109.00
H0211	SCLC	0	0	0	0.01	0.18	0.23	3.27	5.99	46.09
H1618	SCLC	0	3.30	0	55.72	98.35	105.34	128.95	25.68	230.40
H1339	SCLC	0	0	0	1.72	12.57	113.11	166.74	55.89	260.75
H1284	SCLC	0	0	0	0	30.77	0	129.19	0	30.60
H0209	SCLC	0	0	0	0.62	0	152.96	0	8.37	62.63
H1994	SCLC	0	0	0	0.01	3.86	37.63	0	24.06	134.98
H0220	SCLC	0	0	0	0	19.09	23.66	19.33	0	70.63
H2107	SCLC	0	0	0	0	0	13.52	59.22	4.97	368.61
H1963	SCLC	0	0	0	0	67.04	136.70	87.56	19.67	107.52
H0146	SCLC	0	0	0	0	0	99.16	4.83	7.51	119.05
H0524	SCLC	0	0	0	0	0	0	0	19.24	169.26
H0510	SCLC	0	0	0	0	1.14	13.70	2.98	0.19	129.04
H1048	SCLC	0	0	0	0	0.02	44.74	154.24	8.25	119.77
H1417	SCLC	0	0	0	0	73.41	5.38	116.78	16.88	90.20
HCC970	SCLC	0	0	0	0	0	39.32	12.72	6.37	118.25
H0711	SCLC	0	0	0	0	80.96	38.25	111.65	26.22	96.68
H1607	SCLC	0	0	0	0	0	0.12	0	0.08	106.74
H1926	SCLC	0	0	0	0	45.44	106.79	104.60	2.63	153.47
H0069	SCLC	0	0	0	0	0	0.41	28.93	0.99	63.16
H1930	SCLC	0	0	0	42.49	114.33	4.11	181.11	5.57	101.06
H1284	SCLC	0	0	0	0	7.06	55.65	69.30	32.51	114.87
H0249	SCLC	0	0	0	0	0	0.86	70.31	3.98	82.94
H1514	SCLC	0.10	0	0	0	0	37.77	120.88	0.04	122.76

Table 1. *Continued*

ESR2	*CDH1*	*HIC1*	*GSTP1*	*PGR*	*AR1*	*APC*	*MGMT* PRO	*MYOD1*	*CDKN2A*	*ESR1*
0.1	5.76	0	0	18.11	0	0	0	0	0.10	0
0.02	0	68.47	0	140.30	0	0	0	0.99	0	0
0	0	23.91	0	37.08	0	0	0	0	0	0.12
0.07	0	68.65	0	72.74	0	73.45	0	0.18	30.46	86.47
1.72	0.44	124.43	17.50	19.02	36.34	153.78	0	63.89	0.65	48.89
20.62	0	53.50	0	4.67	0	87.83	0	0.56	0	1.24
0	0	49.53	4.88	2.17	101.70	0.25	0	135.70	38.14	2.45
20.23	1.57	117.10	0	9.39	0	109.70	0	21.60	32.10	24.54
0	0	108.95	89.49	178.72	3.64	91.98	0	83.81	44.07	29.15
0	0	100.91	0	0	0	72.54	0	0	0.12	0
0	5.90	101.07	180.06	0	0	129.30	0	0	0	61.14
132.50	0	43.11	169.04	85.04	7.44	67.22	0	67.14	93.57	137.93
0	0	265.06	0	0	8.36	238.82	0	0	0	0
6.75	0	130.11	291.83	0.64	0.42	76.11	0	0.38	48.52	2.72
0.12	0	161.95	17.37	0.56	0	148.23	0	0	0	1.16
69.07	0	87.97	112.26	83.06	48.58	73.46	20.03	70.78	46.06	95.45
15.77	80.69	89.77	0	14.82	0	90.13	0	0	0	45.77
2.62	0	98.22	0	35.05	0	84.23	0	0	0	98.74
1.74	16.28	89.13	82.01	14.61	1.31	83.20	0	8.85	23.79	78.18
59.09	1.74	143.12	0	1.19	0.04	0	0	20.02	0	94.23
0	1.65	30.37	0	0.91	78.72	0	0	10.06	0	104.12
0	0	11.74	0	1.76	0	18.42	0	0	0	0.51
109.47	0	111.71	388.92	123.42	67.23	121.83	135.26	69.95	0	56.52
13.43	8.24	36.69	0	0	1.16	68.51	2.47	0.46	36.70	0.59
0.02	0.55	162.24	0	12.31	0.01	70.29	0	0	0	35.15
0	11.69	62.52	0	0	0	86.86	0	71.31	102.00	0
1.40	0	144.69	0	5.32	0	99.46	0	6.15	0.07	21.49
0	0	114.95	6.19	31.41	0	0	0	5.14	0	28.49
0	0	54.27	240.43	14.52	0	4.02	0	0	0	0
0	13.49	100.58	196.99	169.67	0	0	0	0	99.37	3.98
19.21	0	89.60	215.02	13.54	47.05	135.01	247.22	21.43	0.07	79.29
0	0	76.76	0	4.69	0	0.32	0	14.27	0	0
4.81	0	117.15	0	0	0	0	0	53.32	77.01	93.44
0.04	0	51.56	76.53	1.05	0.01	15.62	0.91	79.25	0.02	0.63
2.49	0	87.12	0	5.23	0	3.07	0	7.97	0	21.78
4.32	0	125.78	0	3.10	0.16	0	0	0.12	0	0.02
0	0	43.35	0	0.20	0	145.50	0	0	0.44	0
0	0	178.39	0	102.44	0	60.01	95.77	0	0	0
0	0	36.00	0	0	49.38	72.04	0	9.83	0	0
0.04	0	223.52	86.48	0	0	65.08	0	0.74	26.41	48.91
1.56	0	88.51	254.08	0.26	0	0	0	0	0	0
30.98	0	101.81	90.42	104.56	24.39	96.85	10.15	27.52	0	50.27
0	0	108.55	241.67	12.00	0	0	43.13	0.13	40.30	52.97
19.35	0	89.94	0	7.44	0	0	0	0	0	0.18
0.13	0	66.18	0	2.71	0	0	0	0.17	0	0
3.46	42.08	89.53	0	193.71	1.26	291.97	0	51.77	96.95	265.16
19.11	35.63	111.98	0	42.91	0	97.55	0	0	0	25.65
0	85.58	83.33	267.30	167.77	0	1.15	0	1.48	66.32	16.09
91.80	9.63	137.29	13.41	8.38	3.67	217.04	0	0	8.21	0
0	0	74.01	37.24	0	0	35.71	0	0	65.58	88.13
0	0	57.21	0	0.10	0	0	0	0	0	0
0.41	0	78.05	31.23	0.19	0.03	35.96	0	0	0.02	0.07
0.08	0	209.35	0	10.75	49.44	103.71	0	4.03	3.84	48.47
0	0	104.34	0	21.95	2.43	2.81	0	6.89	0	0
0	0	116.24	198.31	0.27	40.11	68.96	0	0	0	60.16
0	0	99.24	0	0.72	0	0	0	0	0	0.73
0	0	82.15	10.72	0	0	0	0	0	0	0
0	0	71.09	2.74	1.40	0	0	3.65	0	0	0
23.30	0	112.08	0	2.90	21.90	113.55	0	6.37	0	0.02
0.20	0	78.04	0	0.32	1.94	0	0	0	0	1.29
1.97	0	108.96	0	61.35	56.75	0	0	8.21	0	5.15
58.51	0	42.01	239.14	0	0	0	0	0	0	0.53
0	0	181.80	0	0.39	0	44.17	0	0	0	0
0.40	0	41.33	0	4.77	0	0.97	0	0	0	0
0	0.94	53.18	10.74	2.33	0.04	4.06	3.11	25.71	1.14	5.55
0	0	101.45	0	46.79	0	101.86	0	0	0	1.08
0.41	0	89.70	31.42	1.00	14.15	0	0	0	0	0
0	15.17	94.25	191.69	0.01	0	110.25	0	0.48	0	77.05
0.35	0	67.07	0	143.71	0.14	0.33	0	0	0.60	0

(*Table continues*)

Table 1. *Continued*

Sample[‡]	Cell type	*TYMS*	*TGFBR2*	*THBS1*	*CDKN2B*	*TIMP3*	*PTGS2*	*CALCA*	*MGMT* E/I	*MTHFR*
HCC033	SCLC	0	0	0	0	9.57	9.07	67.51	6.76	44.84
H0748	SCLC	0	0	0	0	0	0.03	106.55	34.52	94.65
H1304	SCLC	0	0	0	0	0	0.01	31.32	21.42	119.37
H0889	SCLC	0	0	0	194.64	77.17	0.25	188.82	4.48	425.90
H2171	SCLC	0	0	20.29	0	197.82	0.43	48.71	8.01	151.86
H0740	SCLC	0	0	0	0	104.43	1.51	92.15	16.69	116.50
H1045	SCLC	0	0	0	0	616.64	17.49	146.67	0.60	276.87
H1184	SCLC	0	0	0	1.49	0	11.57	103.60	2.35	88.96
H2227	SCLC	0	0	0	0	0	33.71	86.55	2.01	83.75
H2196	SCLC	0	0	0	0	0	67.81	48.93	0	123.45
H2141	SCLC	0	0	0	0	11.99	0.07	107.91	12.39	107.82
H1105	SCLC	0	0.22	0	0	0	1.57	142.43	26.49	121.69
H0082	SCLC	0	0	0	0	0	11.41	0	17.00	173.18
H0526	SCLC	0	0	0	0	0	68.50	0	2.66	68.21
H1870	SCLC	0	0	0	0	0	111.91	365.60	0.31	138.99
H2029	SCLC	0	0	0	0	0	49.09	392.84	0	109.37
H2195	SCLC	0	0.03	0	0	0	49.49	1.56	0.62	58.69
H0060	SCLC	0	0	0	0	35.20	63.09	75.92	11.78	466.19

*PMR: Percentage Methylated Reference, values <0.02 and >0.00001 are listed as 0.01.

[†] Loci are designated by their Human Genome Organization (HUGO) name.

[‡] NCI-derived cell lines, designated as NCI-Hxxxx are listed without the NCI prefix. Cell lines derived from the Hamon Cancer Center are designated HCCxxx.

cluding the percentage of cases correctly classified by the models was generated.

Comparison of the Classifications Obtained by the 30 Classificatory Models: Kappa Statistics

The analysis of the test cases by the 10 ANN models and the 20 LDA models resulted in 30 test case classifications. The classifications provided by each model were compared with the true classification of cell type (known from the well-characterized cell lines) using kappa statistics.[46,47] This is a method that has been widely used for assessing intra-observer and inter-observer agreement for diagnoses provided by various observers. The kappa values allow for classification of agreements as: poor (< 0.41), moderate (0.41 to 0.6), substantial (0.6 to 0.8), and almost perfect (0.81 to 1.00).

Results and Discussion

Table 1 shows the DNA methylation analysis data, given as the percentage methylated reference (PMR) values for 20 loci in 87 cell lines (three uniformly negative loci and cell lines for which methylation data were incomplete were excluded from the original data set[20]). Using these 20 variables, a step-wise backward procedure selected a five significant-variable model with *PTGS2*, *CALCA*, *MTHFR*, *ESR1*, and *CDKN2A*. In the previously published hierarchical clustering study,[20] all 91 cell lines and all loci (even those with incomplete methylation data) were used to determine the significant variables. The four most significant loci found in the previous analysis ($P < 0.003$) were identical to four of the five loci selected here (*PTGS2*, *CALCA*, *MTHFR*, and *ESR1*). The difference in the other variable (*CDKN2A*) likely lies in the data set used, and in the fact that all of the loci were evaluated individually for statistically significant differences in methylation levels in the previous analysis, whereas they were evaluated *en bloc* here.

The data were next sorted randomly by cell line into 10 different data sets (numbered 1 to 10), each with training and testing subsets composed of 71 and 16 of the cases, respectively. SCLC and NSCLC cell lines were equally represented in training and test sets. The 10 training/testing models were analyzed with ANN, resulting in the classifications shown in Table 2. Models 1 to 5 used all 20 variables, while models 6 to 10 used the five variables selected by statistical analysis. In the latter five ANN models, two SCLC cell lines (NCI-H0069 and NCI-H0249) were classified incorrectly as NSCLC by some of the ANN. The results, summarized in Table 2, indicate that ANN can be trained to correctly classify up to 100% of the cell lines, based on the current methylation data set. Comparison of the 10 ANN models using kappa statistics indicates that the ANN using all variables was more accurate than the ANN using the five significant variables.

Next, LDA was used to analyze the same 10 data sets, using regular or log-transformed PMR values (see methods). The results of these analyses, with the kappa coefficients, are given in Table 2. Correct classification rates provided by the LDA models were variable and less accurate than those provided by ANN, ranging from 62.5% to 87.5%. Considerable variability based on the choice of training set was evident in the LDA using five genes. Indeed, four of the 10 LDA classificatory models yielded only poor kappa values. The comparison shown in Table 2 emphasizes the importance of performing cross-validation studies to analyze the accuracy of various supervised classificatory methods.

Our analyses demonstrate that it is possible to classify individual lung cancer cell lines into SCLC and NSCLC based on the analysis of DNA methylation markers using multivariate ANN. The ANN models using five genes correctly classified 87% or more of all test cases, while the ANN models using 20 input neurons were able to cor-

Table 1. *Continued*

ESR2	*CDH1*	*HIC1*	*GSTP1*	*PGR*	*AR1*	*APC*	*MGMT* PRO	*MYOD1*	*CDKN2A*	*ESR1*
0.12	0	100.55	4.24	0.31	0	0.43	0	0	0.01	0
0	0	171.31	233.74	0.10	0.01	0	0	0	0.26	0
0	0	128.46	13.19	0.12	53.95	0	0	2.39	0	0
0	0	114.62	99.18	25.67	0	4.09	0	0	0.03	0.03
18.91	0	163.11	131.76	0.07	0	0	0	0.67	0	6.71
1.13	0	65.45	106.85	0.55	0.01	30.28	0	0.30	0	0
155.13	0	310.47	0	0	0	255.28	84.17	0	3.94	0
0	0	101.16	16.19	2.48	0	0	52.28	0	0.11	6.91
124.55	6.34	79.68	0	73.21	0.01	117.82	0	3.49	92.37	20.93
67.79	2.62	99.25	11.03	20.64	5.66	0	5.05	0.50	0	0.07
0	0	80.04	0	47.49	0	0	0	0	0	0.16
0	0	92.49	222.94	8.70	0	0.14	136.79	0	0	0
0	0	119.62	0	0	0	192.82	0	124.35	0	0
0	0	63.26	0	1.02	0	0.57	0	0.21	0	0.40
39.40	0	97.22	187.56	51.52	0.18	0	1.68	0.29	0.11	1.40
0	0	79.20	55.85	0.81	0	88.47	0	0	0	0
1.10	0.17	47.09	0	193.27	6.78	0	66.64	33.52	0	0.62
44.09	0	53.19	112.63	1.13	0	0	0	0	0	0

rectly classify all test cases. This is a substantial improvement over the 78% correct classification observed using hierarchical clustering, and is also markedly better than the results obtained using LDA analysis (only 62 to 87% correct classifications, and substantial variability between models). The LDA models described in this study were developed using multivariate data from only 87 cell lines. When these data are randomly subdivided into training and testing sets, certain models are more likely to use cell lines that could not be classified with LDA as test subjects, while others may include them as part of the training set, resulting in variable correct classification rates for the test cases. Some of the variability may also be related to non-normality of the data in multivariate space, as the logarithmic transformation of the data slightly improved the accuracy of the LDA models using 20 variables. ANN models appear to handle these classification problems with less variability and misclassified only two cell lines. These two cell lines (H0069 and H0249) were also misclassified by most of the LDA models and by the previously carried out hierarchical clustering.[20]

Our results were attained with a modest set of methylation loci: 20 of the ~12,000 CpG islands present in the human genome. They compare favorably with studies of clinical inter-observer variability for the diagnosis of SCLC in biopsy materials and cytologic samples, where concordance rates of approximately 90% have been reported.[16] While it will be critical to verify our observations using human lung cancer tissue, the recent finding that methylation profiles in cell lines specifically resemble those found in tumors derived from the same organ suggests that analyses of lung tumor material will yield similar results.[48] Our recent comparison of lung adenocarcinoma and malignant mesothelioma cell lines and tumors indicates that methylation profiles in cell lines strongly resemble those in the corresponding tumors (Tsou JA, Shen LYC, Siegmund KP, Long TJ, Laird PW, Seneviratne CK, Koss MN, Pass HI, Laird-Offringa IA, manuscript submitted for publication). Future studies using a larger

Table 2. Classification of Test Cases (n = 16; 8 SCLC and 8 NSCLC) by Linear Discriminant Models and Artificial Neural Networks

Model training cell lines n = 71 (33 SCLC and 38 NSCLC)	Artificial neural network		Linear discriminant analysis		LDA after logarithmic transformation of the data	
	Number of correctly classified cell lines	Kappa coefficient	Number of correctly classified cell lines	Kappa coefficient	Number of correctly classified cell lines	Kappa coefficient
Models trained with all variables						
1	16 (100%)	1	12 (75%)	0.5	12 (75%)	0.5
2	16 (100%)	1	10 (62%)	0.25	12 (75%)	0.5
3	16 (100%)	1	12 (75%)	0.50	14 (87%)	0.75
4	16 (100%)	1	10 (62%)	0.25	11 (69%)	0.35
5	16 (100%)	1	10 (62%)	0.25	13 (81%)	0.65
Models trained with 5 variables (*PTGS2*, *CALCA*, *MTHFR*, *ESR1*, *CDKN2A*)						
6	16 (100%)	1	13 (81%)	0.62	13 (81%)	0.62
7	14 (87%)	0.75	10 (62%)	0.25	10 (62%)	0.25
8	14 (87%)	0.75	14 (87%)	0.75	13 (81%)	0.62
9	14 (87%)	0.75	13 (81%)	0.62	13 (81%)	0.62
10	15 (98%)	0.88	13 (81%)	0.62	13 (81%)	0.62

set of methylation markers, in combination with human lung cancer tissue samples, will be important to validate and extend our observations. The availability of high-throughput methods such as MethyLight, which allows the rapid processing of hundreds of samples,[42,43] will stimulate rapid progress in this area. Once optimal accuracy in diagnosis using tumor material is achieved, the same approaches can be used to develop non-invasive methods to detect lung cancer, such as the analysis of methylation profiles in the serum or sputum of subjects at risk for lung cancer.[22] The development of methods, such as those described here, to process complex molecular data and translate it into clinically meaningful information is crucial to realize the potential of DNA methylation analysis as a powerful molecular diagnostic tool.

References

1. Sullivan Pepe M, Etzioni R, Feng Z, Potter JD, Thompson MD, Thornquist M, Winget M, Yasui Y: Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001, 93:1054–1061
2. Marchevsky AM, Gil J, Jeanty H: Computerized interactive morphometry in pathology: current instrumentation and methods. Hum Pathol 1987, 18:320–331
3. Marchevsky AM, Hauptman E, Gil J, Watson C: Computerized interactive morphometry as an aid in the diagnosis of pleural effusions. Acta Cytol 1987, 31:131–136
4. Bartels PH, Thompson D, Weber JE: Image Analysis: A Primer for Pathologists. New York, Raven Press, 1994, pp 2
5. An CS, Petrovic LM, Reyter I, Tolmachoff T, Ferrell LD, Thung SN, Geller SA, Marchevsky AM: The application of image analysis and neural network technology to the study of large-cell liver-cell dysplasia and hepatocellular carcinoma. Hepatology 1997, 26:1224–1231
6. Walts AE, Morimoto R, Marchevsky AM: Computerized interactive morphometry and the diagnosis of lymphoid-rich effusions. Am J Clin Pathol 1993, 99:570–575
7. Walts AE, Marchevsky AM: Computerized interactive morphometry: an expert system for the diagnosis of lymphoid-rich effusions. Am J Clin Pathol 1989, 92:765–772
8. Marchevsky AM, Klapper E, Gil J: Computerized classification of nuclear profiles in non-Hodgkin's lymphomas. Am J Clin Pathol 1987, 87:561–568
9. Marchevsky AM, Gil J: Applications of computerized interactive morphometry in pathology: II. a model for computer-generated diagnosis. Lab Invest 1986, 54:708–716
10. Marchevsky A, Gil J, Silage D: Computerized interactive morphometry as a potentially useful tool for the classification of non-Hodgkin's lymphomas. Cancer 1986, 57:1544–1549
11. Cenci M, Nagar C, Vecchione A: PAPNET-assisted primary screening of conventional cervical smears. Anticancer Res 2000, 20:3887–3889
12. Mango LJ, Radensky PW: Re-screening of cervical Papanicolaou smears using PAPNET. JAMA 1998, 279:1786–1787
13. Troni GM, Cipparrone I, Cariaggi MP, Ciatto S, Miccinesi G, Zappa M, Confortini M: Detection of false-negative Pap smears using the PAPNET system. Tumori 2000, 86:455–457
14. Travis WD, Colby TV, Corrin B, Shimosato Y: Histological Typing of Lung and Pleural Tumours. Berlin, Springer Verlag 1999, pp 7–9
15. Marchevsky AM: Surgical Pathology of Lung Neoplasms. New York, Marcel Dekker, Inc, 1990, pp 77–211
16. Marchevsky AM, Gal AA, Shah S, Koss MN: Morphometry confirms the presence of considerable nuclear size overlap between "small cells" and "large cells" in high-grade pulmonary neuroendocrine neoplasms. Am J Clin Pathol 2001, 116:466–472
17. Travis WD, Rush W, Flieder DB, Falk R, Fleming MV, Gal AA, Koss MN: Survival analysis of 200 pulmonary neuroendocrine tumors with clarification of criteria for atypical carcinoid and its separation from typical carcinoid. Am J Surg Pathol 1998, 22:934–944
18. Travis WD, Gal AA, Colby TV, Klimstra DS, Falk R, Koss MN: Reproducibility of neuroendocrine lung tumor classification. Hum Pathol 1998, 29:272–279
19. Sozzi G: Molecular biology of lung cancer. Eur J Cancer 2001, 37(Suppl 7):S63–S73
20. Virmani AK, Tsou JA, Siegmund KD, Shen LY, Long TI, Laird PW, Gazdar AF, Laird-Offringa IA: Hierarchical clustering of lung cancer cell lines using DNA methylation markers. Cancer Epidemiol Biomarkers Prev 2002, 11:291–297
21. Holliday R: The significance of DNA methylation in cellular aging. Basic Life Sci 1985, 35:269–283
22. Tsou JA, Hagen JA, Carpenter CL, Laird-Offringa IA: DNA methylation analysis: a powerful new tool for lung cancer diagnosis. Oncogene, 2002, 21:5450–5461
23. Bird A: DNA methylation patterns and epigenetic memory. Genes Dev 2002, 16:6–21
24. Costello JC, Plass C: Methylation matters. J Med Genet 2001, 38: 285–303
25. Robertson KD: DNA methylation, methyltransferases, and cancer. Oncogene 2001, 20:3139–3155
26. Wade PA: Methyl CpG binding proteins: coupling chromatin architecture to gene regulation. Oncogene 2001, 20:3166–3173
27. Baylin SB, Esteller M, Rountree MR, Bachman KE:, Schuebel K, Herman JG: Aberrant patterns of DNA methylation, chromatin formation, and gene expression in cancer. Hum Mol Genet 2001, 10:687–692
28. Jones PA, Laird PW: Cancer epigenetics coming of age. Nat Genet 1999, 21:163–167
29. Esteller M, Sanchez-Cespedes M, Rosell R, Sidransky D, Baylin SB, Herman JG: Detection of aberrant promoter hypermethylation of tumor suppressor genes in serum DNA from non-small cell lung cancer patients. Cancer Res 1999, 59:67–70
30. Costello JC, Fruhwald MC, Smiraglia DJ, Rush LJ, Robertson GP, Gao X, Wright FA, Feramisco JD, Peltomaki P, Lang JC, Schuller DE, Yu L, Bloomfield CD, Caligiuri MA, Yates A, Nishikawa R, Huang JJS, Petrelli NJ, Zhang X, O'Dorisio MS, Held WA, Cavenee WK, Plass C: Aberrant CpG-island methylation has non-random and tumor-type-specific patterns. Nat Genet 2000, 25:132–138
31. Esteller M, Corn PG, Baylin SB, Herman JG: A gene hypermethylation profile of human cancer. Cancer Res 2001, 61:3225–3229
32. Marchevsky AM, Shah S, Patel S: Reasoning with uncertainty in pathology: artificial neural networks and logistic regression as tools for prediction of lymph node status in breast cancer patients. Mod Pathol 1999, 12:505–513
33. Singson RP, Alsabeh R, Geller SA, Marchevsky A: Estimation of tumor stage and lymph node status in patients with colorectal adenocarcinoma using probabilistic neural networks and logistic regression. Mod Pathol 1999, 12:479–484
34. Marchevsky AM, Patel S, Wiley KJ, Stephenson MA, Gondo M, Brown RW, Yi ES, Benedict WF, Anton RC, Cagle PT: Artificial neural networks and logistic regression as tools for prediction of survival in patients with stages I and II non-small cell lung cancer. Mod Pathol 1998, 11:618–625
35. Bellotti M, Elsner B, Paez DL, Esteva H, Marchevsky AM: Neural networks as a prognostic tool for patients with non-small cell carcinoma of the lung. Mod Pathol 1997, 10:1221–1227
36. Marchevsky AM, Truong H, Tolmachoff T: A rule-based expert system for the automatic classification of DNA "ploidy" histograms measured by the CAS 200 image analysis system. Cytometry 1997, 30:39–46
37. Marchevsky AM, Coons G: Expert systems as an aid for the pathologist's role of clinical consultant: CANCER-STAGE. Mod Pathol 1993, 6:265–269
38. Marchevsky AM: Expert systems for efficient handling of medical information: I. lung cancer. Anal Quant Cytol Histol 1991, 13:89–92
39. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001, 7:673–679
40. Phelps RM, Johnson BE, Ihde DC, Gazdar AF, Carbone DP, McClintock PR, Linnoila RI, Matthews MJ, Bunn Jr PA, Carney D, Minna JD, Mulshine JL: NCI-Navy Medical Oncology Branch cell line data base. J Cell Biochem Suppl 1996, 24:32–91
41. Olek A, Oswald J, Walter J: A modified and improved method for bisulphite-based cytosine methylation analysis. Nuclei Acids Res 1996, 24:5064–5066

42. Eads CA, Danenberg KD, Kawakami K, Saltz LB, Blake C, Shibata D, Danenberg PV, Laird PW: MethyLight: a high-throughput assay to measure DNA methylation. Nucleic Acids Res 2000, 28:E32
43. Trinh BN, Long TI, Laird PW: DNA methylation analysis by MethyLight technology. Methods 2001, 25:456–462
44. Rolston DW: Principles of Artificial Intelligence and Expert System Development. New York, McGraw-Hill Book Company, 1988
45. Afifi AA, Clark V: Computer-aided multivariate analysis. New York, Chapman & Hall/CRC, 1999, pp 243–280
46. Fleiss JL, Cuzick J: The reliability of dichotomous judgements: unequal numbers of judgements per subject: applied psychological measurement. Applied Psychol Meas 2003, 27:537–542
47. Marchevsky AM, Nelson V, Martin SE, Greaves TS, Raza AS, Zeineh J, Cobb CJ: Telecytology of fine needle aspiration biopsies of the pancreas: a study of well-differentiated adenocarcinoma and chronic pancreatitis with atypical epithelial repair changes. Diagn Cytopathol 2003, 28:147–152
48. Paz MF, Fraga MF, Avila S, Guo M, Pollan M, Herman JG, Esteller M: A systematic profile of DNA methylation in human cancer cell lines. Cancer Res 2003, 63:1114–1121