Genomics & Informatics- Vol. 4(4) 161-166, December 2006

Prediction Model for the Cellular Immortalization and Transformation Potentials of Cell Substrates

Min Su Lee[1], Clayton A. Matthews[2], Min-Ju Chae[2], Jung Yun Choi[3], Yeo Won Sohn[3], Min-Jung Kim[4], Su-Jae Lee[4] and Woong-Yang Park[2]*

[1]Department of Computer Science and Engineering, Ewha Womans University, Seoul 120-759, Korea, [2]Human Genome Research Institute and Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine, Seoul 110-799, Korea, [3]Biologics Headquater, Korea Food and Drug Administration, Seoul 122-704, [4]Laboratory of Radiation Experimental Therapeutics, Korea Institute of Radiological & Medical Sciences, Seoul 139-706, Korea

Abstract

The establishment of DNA microarray technology has enabled high-throughput analysis and molecular profiling of various types of cancers. By using the gene expression data from microarray analysis we are able to investigate diagnostic applications at the molecular level. The most important step in the application of microarray technology to cancer diagnostics is the selection of specific markers from gene expression profiles. In order to select markers of immortalization and transformation we used c-*myc* and H-*ras*^{V12} oncogene-transfected NIH3T3 cells as our model system. We have identified 8751 differentially expressed genes in the immortalization/transformation model by multivariate permutation F-test (95% confidence, FDR <0.01). Using the support vector machine algorithm, we selected 13 discriminative genes which could be used to predict immortalization and transformation with perfect accuracy. We assayed H-*ras*^{V12}-transfected "transformed" cells to validate our immortalization/transformation classification system. The selected molecular markers generated valuable additional information for tumor diagnosis, prognosis and therapy development.

Keywords: microarray, NIH3T3, immortalization, transformation, *c-myc*, H-*ras*^{V12}, prediction model, SVM

Introduction

Cell substrates can be used to produce recombinant vaccines or proteins for therapeutics. Although the number of residual cells, substrates and adventitious materials is usually minimal in the final products, we need to check whether these substances can lead to transformation and tumors in recipients. To check the tumorigenic potential of cell substrates, it is necessary to check the tumorigenicity *in vivo* using immunocomprimized animal models such as Balb/c-nu mice. However, these methods are time consuming and expensive making it impractical to test cell substrates candidates in this way. By using biomarkers for cellular transformation we can classify cell substrates according to their transforming potentials without the need for costly animal testing.

Normal diploid cells can divide a limited number of times before entering into an irreversible growth arrest termed replicative senescence (Hayflick, 1965). A small percentage of cells can acquire unlimited proliferation potential, which is a key step in tumorigenesis (Greider, 1999). Telomerase activity has been reported to be sufficient to immortalize human diploid fibroblasts (Bodnar *et al.*, 1998), while other studies have shown that the c-*myc* oncogene fulfils many of the criteria for a gene involved in the immortalization of human epithelial cells (Bouchard *et al.*, 1998).

Signaling from the small GTPase, Ras, has been under intense investigation over the past decade due to its involvement in mediating the pathways that control transcriptional activation. The pathways include mediators of critical and diverse cellular functions such as proliferation, development, differentiation, and apoptosis (Ayllon and Rebollo, 2000). Mutations conferring constitutive activation of Ras occur frequently in many types of human cancers (Bos, 1989). Signaling pathways downstream of Ras have been fairly well defined, and their contribution to Ras-mediated transformation has been studied extensively.

With the aim of reconstructing tumorigenesis, we have used mouse fibroblast NIH3T3 cells to establish models for immortalization and transformation by c-*myc* and/or H-*ras*^{V12}. The gene expression profiles of these cells were analyzed by Affymetrix microarray to find differentially expressed genes in each group. Finally, we selected 13 genes which could be used to predict whether the tested cells were normal, immortalized or transformed with one hundred percent accuracy. These genes and algorithm may be useful in the classification of cell substrates to predict their transforming potential without the need for *in vivo* tumorigenicity assays.

*Corresponding author: E-mail wypark@snu.ac.kr, Tel +82-2-740-8241, Fax +82-2-744-4534
Accepted 4 Dec 2006

Materials and Methods

Immortalization and transformation model

NIH3T3 mouse fibroblast cells were maintained in DMEM supplemented with 10% FBS and antibiotics of penicillin/streptomycin at 37 °C in a humidified incubator with a 5% CO_2 atmosphere. We established stable NIH3T3(c-*myc*) cell lines by transfecting NIH3T3 cells with pcDNAc-*myc* using Lipofectamine (Invitrogen, USA) and selecting stable transfectants with Geneticin (Invitrogen, USA). The other cell lines like NIH3T3(H-*ras*^{V12}) and NIH3T3(c-*myc*+H-*ras*^{V12}) were established using the same procedure. The expressions of transgenes was checked by Western blot analysis (data not shown), and the *in vitro* and *in vivo* transformation potential of cells was checked using soft agar assays and by tumor formation assays in Balb/c-nu mouse, respectively.

RNA extraction and microarray experiments

Total RNAs from cells was extracted using Trizol reagent (Invitrogen Inc., USA) and purified by column chromatography (RNeasy, Qiagen) according to the protocol provided by the manufacturer. The quality of RNA was checked by quantifying 260/280 and 260/230 ratios and by gel electrophoresis (data not shown). Following ethanol precipitation, RNA was stored at -80 °C after ethanol precipitation. Labeling and hybridization was performed as described in Affymetrix microarray protocol (Kim *et al.*, 2004). The array used in this experiment was Affymetrix mouse genome 430 2.0 array which includes 39,000 transcripts. Many experiments for control, c-*myc*, H-*ras*^{V12}, and c-*myc*+H-*ras*^{V12} microarray experiment were quadruplicated producing a total of 16 sets of microarray data.

Data analysis

Fluorescence intensity was processed and measured using GeneChip scanner 3000 and intensity data was imported to an in-house microarray database. In order to normalize data and remove systemic variance, RMA (Robust Multi-Array Average) normalization was applied to remove systematic variance (Irizarry *et al.*, 2003). The application of RMA allowed the raw intensity values to be background corrected, log2 transformed and then quantile normalized. A linear model was fitted to the normalized data to obtain an expression measurement for each probe set on each array.

To characterize the differences between immortalized and transformed cells, we identified genes that were differentially expressed among control, immortalized (c-*myc*), and transformed (c-*myc*+H-*ras*^{V12}) classes using the multivariate permutation F-test (Simon *et al.*, 2004; Korn *et al.*, 2004). The false discovery rate (FDR) was defined as the proportion of genes reported to be differentially expressed by our assay that were false positives. We used the multivariate permutation test to provide a 95% confidence level where the false discovery rate (FDR) was less than 1%.

Gene ontology (GO) (The Gene Ontology Consortium, 2000) groups of the selected genes were identified. This GO analysis was performed to provide information regarding whether the list of significant genes selected by the analysis was different to a randomly generated list selected from all genes in a given GO category. This type of analysis is different from a simple annotation of a gene list using GO categories. The data was expressed as the observed vs. the expected ratio where the observed was defined as the number of genes in the list of significant genes which fell into a GO category. The expected was defined as the average number of genes which would be expected to fall into that GO category in a subset of genes randomly selected from all genes in the analysis. A GO category consisted of not only the genes which were described by that GO term, but also any gene which was described by any members of that GO term.

We also developed models to predict the class for samples, where the gene ontology is unknown. Although the 95% confidence with FDR<0.01 was a strict constraint, the selected genes were too numerous to construct a prediction model. We performed another feature selection step to obtain compact gene sets for an optimal prediction model. Particularly, we used a method of gene selection utilizing support vector methods (SVM) (Vapnik, 1998) based on recursive feature elimination (RFE) (Guyon *et al.*, 2002). Gene selection with SVM based on RFE eliminates gene redundancy automatically and yields better and more compact gene subsets. We also developed a prediction model using the Sequential Minimal Optimization (SMO) algorithm with a logistic regression model and RBF kernel (Platt, 1998) incorporating the genes that were selected by SVM based on RFE. We estimated the prediction error of each model using leave-one-out cross-validation (LOOCV), which is a special case of *n*-fold cross-validation, where *n* is the number of samples in the dataset (Tan *et al.*, 2005). LOOCV has the advantage of utilizing as much data as possible for training. In addition, the test sets were mutually exclusive, and they effectively covered the entire dataset. LOOCV provided a method which maximized the data obtained from a small dataset and while generating as accurate an

estimate as possible (Tan *et al.*, 2005).

Results and Discussion

Models for immortalization and transformation

NIH3T3 mouse fibroblast cells can be maintained in monolayer cells, but it cannot be grown in soft agar. To establish models for immortalization and transformation, we introduced c-*myc* and H-*ras*^{V12} expression plasmid to NIH3T3 cells and selected stable cell colonies as reported previously (Wiehle et al., 1990). The tumorigenic potential of NIH3T3, NIH3T3(c-*myc*), NIH3T3(H-*ras*^{V12}) and NIH3T3(c-*myc*+H-*ras*^{V12}) cells were checked using the colony formation in soft agar and tumor formation *in vivo* in Balb/c-nu mice (Table 1). We compared the tumorigenicity using Vero and HeLa cells as negative and positive controls, respectively. Immortalized cells could form a colony in *in vitro* tumorigenicity assay, but not tumor mass in *in vivo* tumorigenicity assay, while NIH3T3(H-*ras*^{V12}) and NIH3T3(c-*myc*+H-*ras*^{V12}) cells could make massive tumor *in vivo* as well as *in vitro*.

Table 1. In vitro and in vivo tumorigenicity assay for NIH3T3, NIH3T3(c-myc), NIH3T3(H-rasV12) and NIH3T3(c-myc+H-rasV12) cells.

Groups	Species	Type	Tumorigenicity	
			in vitro	in vivo
HeLa	human	cervix cancer	+	+
Vero	monkey	fibroblast	-	-
NIH3T3	mouse	fibroblast	-	-
NIH3T3(c-myc)	mouse	fibroblast	+	-
NIH3T3(H-RasV12)	mouse	fibroblast	++	++
NIH3T3(c-myc+H-RasV12)	mouse	fibroblast	++	++

Identifying and characterizing differentially expressed genes

We identified 8,751 differentially expressed genes from control, immortalized, and transformed cells through multivariate permutation F-test with 95% confidence and 1% false discovery rate. Upon analysis of the distribution of those genes along the chromosomes, we did not find any preferential locus among 20 chromosomes of the mouse genome (Fig. 1). These results may indicate that the transcriptional regulations of immortalization and transformation do not ioverlab.

To identify the characteristics of the 8,751 genes, we performed gene ontology analysis based on observed vs. expected ratio. Our analysis showed statistically overrepresented GO terms within a group of genes (Table 2). Five categories of gene ontology were related to differentially expressed genes in the immortalization/transformation model. Oxidoreductase activity, especially related to superoxide radicals, was altered in our model. This is likely to be important in DNA damage and mutagenesis. In addition, the genes identified that have nucleotide kinase activity are likely to be important in

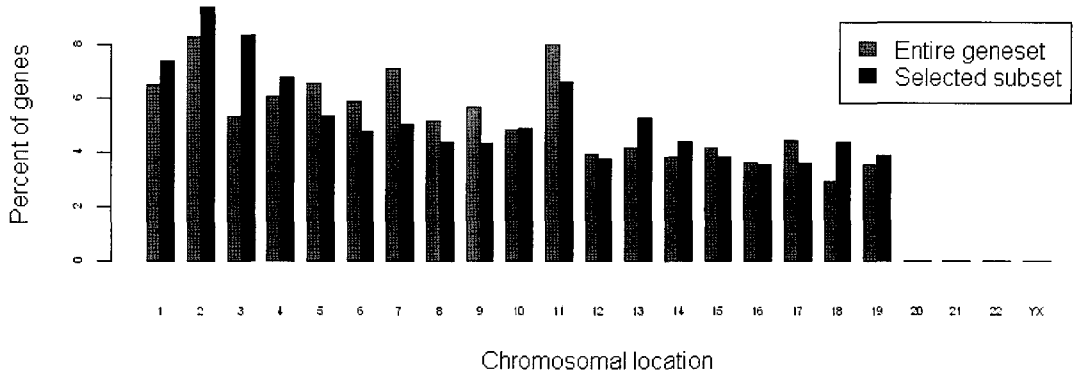

Fig. 1. Plotting of selected DEGs for immortalization/transformation and spotted genes for their distributions in mouse chromosome.

Table 2. Gene ontology analysis on molecular function of immortalization and transformation-related DEGs

GO No.	GO classification	Observed in selected subset	Expected in selected subset	Observed /Expected
16721	Oxidoreductase activity, acting on superoxide radicals as an acceptor	7	2.47	2.84
5173	Stem cell factor receptor binding	12	4.26	2.82
19201	Nucleotide kinase activity	20	7.85	2.55
30693	Caspase activity	13	5.16	2.52
30020	Extracellular matrix structural constituent conferring tensile strength	29	11.88	2.44

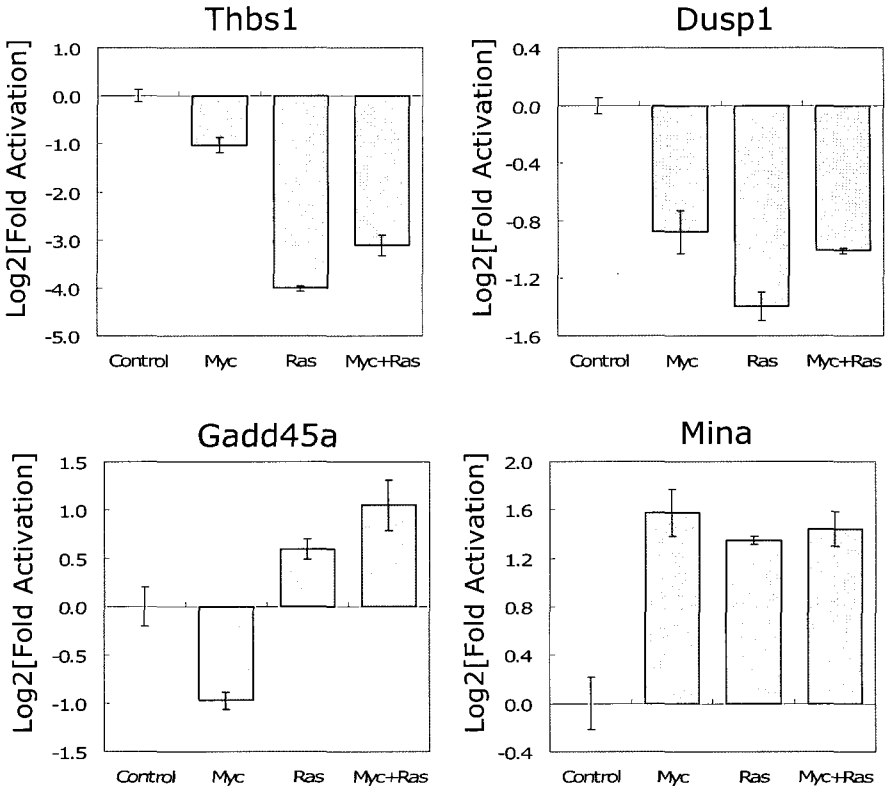

Fig. 2. Expression profiles of known c-*myc* target genes like Thbs1 (thrombospondin1), Dusp1 (dual specificity phosphatase 1), Gadd45a (growth arrest and DNA-damage-inducible 45 alpha) and Mina (*myc* inducible antigen)

aberrant DNA synthesis and metabolism predisposing cells to transformation. One of the most interesting categories deleted involved stem cell factor receptor binding which may be representative of cancer cell de-differentiation in our model.

We have picked several genes out of the DEG lists to explain the effects of each oncogenes. In particular, we tested four genes, which have been reported to be targets of c-*myc* transcriptional regulation. All of these genes were up- or down-regulated by the expression of c-*myc* (Fig. 2), however, the expression of all four genes was also affected by the ectopic expression of H-*ras*^{V12}. In the cases of Mina and Dusp1, H-*ras*^{V12} induced the expression of target genes at the level of c-*myc*-dependent regulation. The induction of Thbs1 by H-*ras* ^{V12} was higher than that observed for c-*myc*. More interestingly, the transcription of Gadd45a has been shown to be suppressed by c-*myc*, while H-*ras* ^{V12} could up-regulate Gadd45a expression. The effect of H-*ras* ^{V12} was sufficient to overcome the suppression by c-*myc* and up-regulate the expression of Gadd45a in cells expressing both c-*myc* and H-*ras* ^{V12}. Although we picked only four genes from the DEG list, the information on gene expression revealed many interesting regulatory relationships between different oncogenes.

Constructing prediction model

To obtain small sets of discriminative genes, we performed an additional feature subset selection process using linear SVM based on RFE and discovered 13 informative genes (Table 3). We constructed a prediction model using the SMO algorithm with a logistic regression model and the radial basis function (RBF)-kernel. The performance of the prediction model was assessed using LOOCV and was found to be one hundred percent accurate for the data analyzed. We also validated our prediction model using microarray data from four H-*ras*-transfected transformed cell lines and they were correctly classified as a transformed class.

In order to visualize selected biomarkers, we performed hierarchical clustering of 13 gene with average linkage and Manhattan distance measures for a total 16 sets of microarray data (Fig. 3). The heat map illustrated that data from H-*ras*^{V12}-transfected transformed cell (H-*ras*^{V12}) data were closely clustered with c-*myc*+H-*ras*^{V12} calss.

We have created composite profiles for an *in vitro* immortalization and transformation model, which was

Table 3. List of 13 biomarker genes selected for the prediction model

Gene Symbol	Gene Name	GeneBank No.
Pxn	paxillin	AF293883
B4galt7	xylosylprotein beta1,4-galactosyltransferase, polypeptide 7	BC027195
Glul	glutamate-ammonia ligase (glutamine synthase)	AI391218
Fert2	fer (fms/fps related) protein kinase, testis specific 2	AF286537
Mapre2	microtubule-associated protein, RP/EB family, member 2	BC027056
Cth	cystathionase (cystathionine gamma-lyase)	BC019483
Ifi27	interferon, alpha-inducible protein 27	AY090098
Bmp1	bone morphogenetic protein 1	L24755
Maged2	melanoma antigen family D, 2	AF319976
Stk24	serine/threonine kinase 24 (STE20 homolog, yeast)	BG060677
Ugt1a	UDP-glucuronosyltransferase 1A	D87867
Alcam	activated leukocyte cell adhesion molecule	U95030
Chmp4b	chromatin modifying protein 4B	BC011429

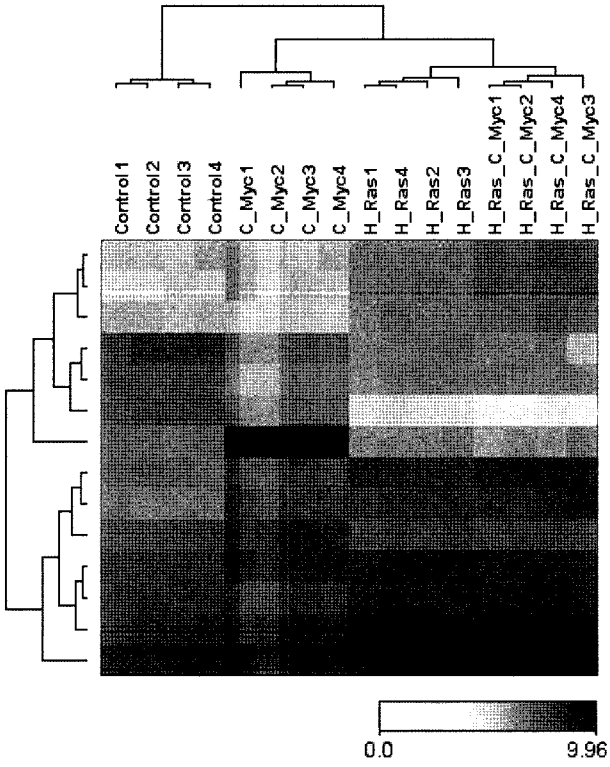

Fig. 3. Hierarchical clustering of NIH3T3, NIH3T3(c-*myc*), NIH3T3(H-*ras*^{V12}) and NIH3T3(c-*myc*+H-*ras*^{V12}) cells using 17 biomarker genes.

confirmed by *in vivo* tumorigenecity assays. Using the multivariate permutation F-test (95% confidence, FDR <0.01), we found 8,751 differentially expressed genes for our immortalization/transformation model. In c-*myc*-transfected "immortalized" cells, we found signatures of known immortalization markers related to telomere and transforming growth factor signaling. A comparison of the known targets for c-*myc* revealed that many were also regulated by H-*ras*^{V12}. In the case of Gadd45a, c-*myc* alone could suppress its expression while the co-expression of H-*ras*^{V12} caused a reversal of this suppression. To enhance the power of analysis of this sysem, more profiles on cell lines and cancer tissues need to be divulged. Based on the wrapper approach we were able to select 13 genes to establish a prediction model for immortalization and transformation, which gave perfect accuracy in calssifying

H-*ras*^{V12}-transfected cells as a transformed class.

Acknowledgements

This work was supported by research grant to W.-Y. Park from Korea Food and Drug Administration (06940-034, KFDA2006-7100).

References

Ayllon, V. and Rebollo, A. (2000). Ras-induced cellular events. *Mol Membr Biol.* 17, 65-73.

Bodnar, A.G., Ouellette, M., Frolkis, M., Holt, S.E., Chiu, C.P., Morin, G.B., Harley, C.B., Shay, J.W., Lichtsteiner, S., and Wright, W.E. (1998). Extension of life-span by introduction of telomerase into normal human cells. *Science* 279, 349–352.

Bos, J.L. (1989). Ras oncogenes in human cancer. *Cancer Res.* 49, 4682-4689.

Bouchard, C., Staller, P., and Eilers, M. (1998). Control of cell proliferation by Myc. *Trends Cell Biol.* 8, 202–206.

Greider, C.W. (1999). Telomerase activation. One step on the road to cancer? *Trends Genet.* 15, 109–112.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning* 46, 389-422.

Hayflick, L. (1965). The limited in vitro lifetime of human diploid cell strains. *Exp Cell Res.* 37, 614–636.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249-264

Kim, J.H., Ha, I.S., Hwang, C.I., Lee, Y.J., Kim, J., Yang, S.H., Kim, Y.S., Cao, Y.A., Choi, S., and Park, W.Y. (2004). Gene expression profiling of anti-GBM glomerulonephritis model: the role of NF-kappaB in immune complex kidney disease. *Kidney Int.* 66, 1826-1837.

Korn, E.L., Troendle, J.F., McShane, L.M., Simon, R. (2004) Controlling the number of false discoveries: Application to high-dimensional genomic data. *J Stat Plan Inf.* 124, 379-398.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization, advances in kernel methods - *Support Vector Learning.* MIT Press, Boston, MA.

Simon, R., Korn, E., McShane, L., Radmacher, M., Wright, G., Zhao, Y. (2004). Design and Analysis of DNA Microarray Investigations. Springer-Verlag New York, NY

Tan, P.N., Stenbach, M., and Kumar, V. (2005). *Introduction to data mining*, Addison Wesley, New York, NY.

The Gene Ontology Consortium. (2000). Gene Ontology: Tool for the unification of biology. *Nat. Genetics* 25, 25-29.

Vapnik, V.N. (1998). Statistical Learning Theory. Wiley, New York, NY.

Wiehle, R.D., Helftenbein, G., Land, H., Neumann, K., and Beato, M. (1990). Establishment of rat endometrial cell lines by retroviral mediated transfer of immortalizing and transforming oncogenes. *Oncogene* 5, 787-794.