Hiroshima J. Med. Sci.
Vol. 57, No. 1, 7~15, March, 2008
HIJM 57–2

Inference on Biological Mechanisms Using an Integrated Phenotype Prediction Model

Yumi ENOMOTO[1,*], Masaru USHIJIMA[2], Satoshi MIYATA[2], Masaaki MATSUURA [2,3] and Megu OHTAKI[4]

1) DYNACOM Co., Ltd. 643, Mobara, Mobara-shi, Chiba 297-0026, Japan
2) Genome Center, Japanese Foundation for Cancer Research, 3-10-6, Ariake, Koto-ku, Tokyo 135-8550, Japan
3) Department of Cancer Genomics, Cancer Institute, Japanese Foundation for Cancer Research, 3-10-6, Ariake, Koto-ku, Tokyo 135-8550, Japan
4) Department of Environmetrics and Biometrics, Research Institute for Radiation Biology and Medicine, Hiroshima University, 1-2-3, Kasumi, Minami-ku, Hiroshima 734-8551, Japan

ABSTRACT

We propose a methodology for constructing an integrated phenotype prediction model that accounts for multiple pathways regulating a targeted phenotype. The method uses multiple prediction models, each expressing a particular pattern of gene-to-gene interrelationship, such as epistasis. We also propose a methodology using Gene Ontology annotations to infer a biological mechanism from the integrated phenotype prediction model. To construct the integrated models, we employed multiple logistic regression models using a two-step learning approach to examine a number of patterns of gene-to-gene interrelationships. We first selected individual prediction models with acceptable goodness of fit, and then combined the models. The resulting integrated model predicts phenotype as a logical sum of predicted results from the individual models. We used published microarray data on neuroblastoma from Ohira et al (2005) for illustration, constructing an integrated model to predict prognosis and infer the biological mechanisms controlling prognosis. Although the resulting integrated model comprised a small number of genes compared to a previously reported analysis of these data, the model demonstrated excellent performance, with an error rate of 0.12 in a validation analysis. Gene Ontology analysis suggested that prognosis of patients with neuroblastoma may be influenced by biological processes such as cell growth, G-protein signaling, phosphoinositide-mediated signaling, alcohol metabolism, glycolysis, neurophysiological processes, and catecholamine catabolism.

Key words: *Biological mechanism, Gene-to-gene interrelationships, Epistasis, Multiple pathways*

A DNA microarray is a collection of microscopic DNA spots arrayed in high density on a solid surface. This technology enables us to obtain expression data on tens of thousands of genes in a single experiment. Gene expression analysis using the microarray has been performed for a variety of purposes, including classification of cancer type[8] and prediction of patient prognosis[19]. Numerous microarray data have recently been opened to the public, allowing researchers to obtain data from internet sites such as the NCBI Gene Expression Omnibus (GEO). However, it remains difficult to find important, relevant genes from microarray data for the following reasons: the reproducibility of microarray data is low, the number of genes measured is enormous but the number of samples is small, and biological interpretation of results is difficult. With microarray data analysis there is no gold standard method to employ in the search for relevant genes. Researchers have suggested various methods of microarray analysis depending on the purpose, such as hierarchical clustering[6], the self-organizing map (SOM)[24], principal component analysis (PCA)[18], the weighted voting method[8], support vector machines (SVM)[4] and AdaBoost[16].

In many cases it is thought that a given phenotype is regulated by multiple genes, and there are many possible patterns of gene-to-gene interrelationship affecting a phenotype. For example, models might have only the main effects of multiple genes or they might include interactions among genes. The latter expresses epistasis. Furthermore, it is possible that multiple pathways

* Address for correspondence: DYNACOM Co.,Ltd. 643, Mobara, Mobara-shi, Chiba 297-0026, Japan
Tel:+81-475-25-8282 e-mail address: enomoto@dynacom.co.jp

control a phenotype. Therefore, we undertook to develop a method to search for the combination of genes that affects a phenotype, the pattern of their interrelationship, and the multiple pathways regulating the phenotype, in order to understand the biological mechanisms involved.

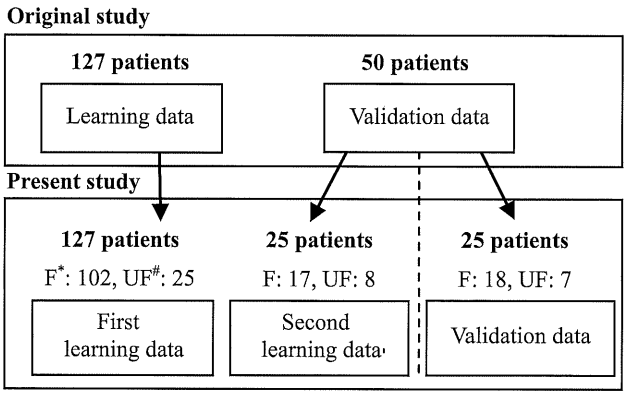
As described in Materials and Methods, we develop an integrated phenotype prediction model expressing multiple pathways by combining individual prediction models, each expressing a particular pattern of gene-to-gene interrelationship and fit using logistic regression. We also develop a methodology for using gene ontology annotations to draw inferences on the biological mechanisms controlling the phenotype based on our prediction models.

For illustration, we construct an integrated phenotype prediction model and infer the biological mechanisms controlling the prognosis of patients with neuroblastoma. For this we use the published data of Ohira et al[17], which are available from the GEO website, and validate our model by an independent subset of the data. We demonstrate that the model has excellent predictive performance despite comprising a small number of genes compared with the previously reported analysis of these data. Finally, we use Gene Ontology analysis to infer biological processes affecting the prognosis of neuroblastoma patients.

MATERIALS AND METHODS

Neuroblastoma Data

For illustration, we used published microarray data on neuroblastoma[17], available from the website of the NCBI Gene Expression Omnibus. The data include two sets of expression results: data on 5340 genes in 136 patients and data on 200 genes in 50 patients. Tumor samples were randomly selected from among specimens in the neuroblastoma tissue bank of the Division of Biochemistry, Chiba Cancer Center Research Institute, that were collected from a number of hospitals in Japan during the period 1996–2002. Informed consent was obtained at each hospital. The 136 patients in the first set comprised 41 with stage one, 22 with stage two, 33 with stage three, 28 with stage four, and 12 with stage 4s neuroblastoma. The 50 patients in the second set comprised 15 with stage one, 6 with stage two, 9 with stage three, 14 with stage four, and 6 with stage 4s neuroblastoma. All tumors were classified according to the International Neuroblastoma Staging System (INSS). Of the 136 patients, 89 were one year of age or younger at the time of diagnosis and 47 were more than one year old. Of the 50 patients, 30 were one year of age or younger at the time of diagnosis and 20 were more than one year old. In the original study, for the purpose of constructing a model predicting 2- or 5-year survival, the data on the 136 patients were used for model selection and the data on the 50 patients were used for validation. The 200 genes measured in the validation data were those selected according to the model for predicting prognosis in the original analysis (Ohira et al). We used the data on 127 out of the 136 patients in the first data set and the 50 patients in the second data set whose clinical outcomes (alive or deceased) at 2 years after diagnosis were known (Fig. 1). Among the 127 patients, 102 had a favorable outcome (defined as alive at 2 years) and 25 patients had an unfavorable outcome (defined as having died within 2 years). Among the 50 patients, 35 had a favorable outcome and 15 had an unfavorable outcome.



* F represents the group of patients with a favorable outcome (alive after 2 years).

UF represents the group of patients with an unfavorable outcome (died within 2 years).

Microarray experiments on the 127 and 50 patients were performed in separate laboratories. We constructed an integrated phenotype prediction model using a two-step learning approach employing the first and second learning data sets. We then validated the performance of the model using the validation data.

Fig. 1. Data used in this study

Data preprocessing

In the original study, the microarray experimental procedures-RNA preparation, probe labeling, and hybridization-were performed in separate laboratories for the 127 and 50 patients. To test whether the two datasets are comparable, we compared the distribution of expression values for the 200 genes available in the second dataset with that of the same 200 genes among the 127 patients in the first dataset. To make the comparison we used 95% confidence intervals (CI) for the average expression levels. Among favorable patients, only about 25% of the genes had CIs that overlapped between the two datasets; among the unfavorable patients, about 40% of the genes had overlapping CIs. However, for most of the genes, expression patterns (e.g., either a higher or lower level of expression among patients with a favorable outcome compared with patients having an unfavorable outcome; in other words, expression levels shifted up or down) were identical in the

two sets of data. As there was a low frequency of genes having overlapping CIs, we normalized the gene expression data using the Z score transformation. The Z score was calculated separately for the 127 and 50 patients as follows:

$$X_{in}=\frac{S_{in}-\bar{S}_i}{\sigma_i} \quad (i=1, \cdots,200).$$

where S_{in} is the log of sample signal divided by the reference signal for the i-th gene and n-th sample, $\bar{S}_i$ is the average of S_i, and σ_i is the standard deviation of S_i. Expression data in the set of 50 patients were replicated, so we used the average of the two expression values for each gene.

Model building

To build models for predicting phenotype, we used a two-step learning approach. Two learning data sets (one from each laboratory) and one validation data set were used as shown in Fig.1. We used the two learning data sets to fit prediction models by logistic regression and selected models with acceptable goodness of fit and predictive performance. Models confirmed to have excellent predictive performance in both the first and second learning steps were then combined to produce the integrated prediction model.

Prediction models based on logistic regression

In the first step, we selected models that predicted prognosis well, using the data on 127 patients (102 with a favorable outcome and 25 with an unfavorable outcome), as shown in Fig.1. In the second step, we validated the performance of the models selected in the first step, using data on 25 patients (17 with a favorable outcome and 8 with an unfavorable outcome, each selected at random from the patients with corresponding outcomes among the original 50 patients).

In the first step, we considered one-gene models for all genes and two-gene models for all combinations of two genes. In addition, we considered three-gene models by adding one additional gene to the two-gene models that fit well. Models were fit using logistic regression. We used a program written by ourselves to build logistic regression models. The program used Octave library (http://www.gnu.org/software/octave/).

The one-gene model is:

$$\text{logit}[Pr(Y=1)]=\alpha+\beta_j X_j$$

where logit is the log odds, namely, logit[Pr] = log[Pr / (1-Pr)]. For the value Y, we coded 1 for a favorable prognosis and 0 for an unfavorable prognosis. The X_j is the log of sample signal divided by the reference signal, normalized by the Z score transformation for gene j ($1 \leqq j \leqq 5340$). The α and β_j are unknown parameters; the former represents an overall average signal, the latter the effect of gene j.

We considered five forms of two-gene models:

$$\text{logit}[Pr(Y=1)]=\alpha+\beta_j X_j+\beta_k X_k \quad (1)$$

$$\text{logit}[Pr(Y=1)]=\alpha+\beta_j X_j+\beta_{jk} X_j X_k \quad (2)$$

$$\text{logit}[Pr(Y=1)]=\alpha+\beta_k X_k+\beta_{jk} X_j X_k \quad (3)$$

$$\text{logit}[Pr(Y=1)]=\alpha+\beta_j X_j+\beta_k X_k+\beta_{jk} X_j X_k \quad (4)$$

$$\text{logit}[Pr(Y=1)]=\alpha+\beta_{jk} X_j X_k \quad (5)$$

where β_{jk} is the effect of two-way interaction between genes j and k ($1 \leqq j < k \leqq 5340$). Equations (2), (3), (4) and (5) with interaction terms express epistasis. We examined about 71 million models corresponding to these five model forms and the combinations of 5340 genes selected two at a time. We performed the likelihood ratio test for the significance of each explanatory variable in each of the five model forms. We then selected as the best two-gene model the form with the smallest Akaike information criterion (AIC) for which all explanatory variables had p values below 0.01 (if such a model existed). We also considered for selection two-gene models having an AIC value of 70 or less.

We considered models for three genes derived from the following full model:

$$\begin{aligned}&\text{logit}[Pr(Y=1)]\\&=\alpha+\beta_j X_j+\beta_k X_k+\beta_l X_l+\beta_{jk} X_j X_k+\beta_{jl} X_j X_l\\&\quad+\beta_{kl} X_k X_l+\beta_{jkl} X_j X_k X_l \quad (6)\end{aligned}$$

where $1 \leqq j < k \leqq 5{,}340$ and l is different from j and k ($1 \leqq l \leqq 5340$, $l \neq j$, $l \neq k$). We considered sixteen forms of three-gene models derived from the full model (6) by taking subsets of the variables. We examined these sixteen model forms by adding one additional gene to the two-gene models that fit well. We performed the likelihood ratio test for the significance of each explanatory variable in the sixteen model forms and selected as the best three-gene model the form with the smallest AIC value for which all explanatory variables had p values below 0.01 (if such a model existed). We also considered for selection three-gene models having an AIC value of 70 or less.

In the second step, we checked the performances of the models selected in the first learning step by predicting the prognoses of the 25 patients in the second learning data set using predicted values $\hat{y}$ given by (7) and calculating the resulting error rate:

$$\hat{y}=\begin{cases}1, & \hat{p}>0.5\\0, & \hat{p}<0.5\end{cases},$$

$$\text{where } \hat{p}=\frac{1}{1+\exp(-\hat{z})} \text{ and } \hat{z}=\text{logit}[P_r(Y=1)] \quad (7)$$

where $\hat{p}$ is the predicted probability of favorable prognosis and $\hat{z}$ is the predicted log odds. For example, $\hat{z}=\alpha+\beta_i X_i$ for a one-gene model. For the predicted value $\hat{y}$, we coded 1 for a favorable

prognosis and 0 for an unfavorable prognosis. We selected models with error rates of 0.1 or less.

Integrated phenotype prediction model

It is possible that a phenotype is controlled by multiple pathways. Therefore, we constructed an integrated phenotype prediction model composed of multiple prediction models that were considered to reflect important pathways. The integrated model predicts prognosis by a logical sum of predicted unfavorable prognosis outcomes from individual prediction models. First, we selected the top ten prediction models according to AIC. To examine the characteristics of models and remove redundancy among the selected models, we performed a cluster analysis for those top ten models on the 25 patients using predicted prognosis. The clustering algorithm was set to average linkage clustering using the Euclidean distance. The models from different clusters performed differently in regards to predicting prognosis. Therefore, we selected the model with the smallest AIC in each cluster as representative of that cluster. We then defined the candidate integrated phenotype prediction model as the best combination of representative models (the combination with minimal error rate and fewest number of models) among the collection of all possible models constructed by logical sums of the individual models. The predicted value $\hat{y}_{imdl}$ is obtained from the candidate integrated phenotype prediction model and given by

$$\hat{y}_{imdl} = \begin{cases} 1, & \sum \hat{y}_a = m \\ 0, & \sum \hat{y}_a \neq m \end{cases} \quad (a = 1, \cdots, m). \tag{8}$$

For the predicted value $\hat{y}_{imdl}$, we coded 1 for a favorable prognosis and 0 for an unfavorable prognosis; m is the number of representative models used in the candidate integrated phenotype prediction model. To construct the integrated phenotype prediction model, we used the combination with the smallest error rate for $\hat{y}_{imdl}$ in the second learning data. We then used the data on the remaining 25 out of 50 patients for independent validation of the integrated model. The prognosis of patients in the validation data was predicted based on equation (8) and the error rate was calculated.

Inference on biological mechanism using Gene Ontology analysis

We performed Gene Ontology (GO) analysis to infer from the integrated model the biological mechanisms controlling the phenotype. The GO terms are controlled vocabularies to describe gene and gene product attributes, which are provided by Gene Ontology Consortium (http://www.geneontology.org/). We annotated individual genes in the neuroblastoma data with GO terms using Gene Compass software (DYNACOM), based on the tables "gene2go" and "gene2unigene" located on the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/). The "gene2go" table gives information on the GO terms from the Entrez Gene identifiers and the "gene2unigene" table gives corresponding UniGene cluster ID from the Entrez Gene ID. We used GO terms in the Biological Process category. We performed a statistical test to identify GO terms that were significantly over-represented in genes included in the top ten models. The null and alternative hypotheses for the test are:

$$H_0 : p_{go} = p_{go}^0 \text{ vs. } H_1 : p_{go} \neq p_{go}^0$$

Consider, for example, a certain GO term, XX. p_{go} is the proportion of genes appearing in the top ten models that have GO term XX. p_{go}^0 is the proportion of genes appearing in ten models randomly selected that have GO term XX. Define d as the number of genes appearing in the top ten models. We selected d genes randomly from among all genes in the neuroblastoma data, and repeated this e = 10,000 times. Denote by f the number of genes with GO term XX among genes appearing in the top ten models and denote by h the number of genes with GO term XX among the d genes randomly selected from the neuroblastoma data. An upper-tail Monte Carlo p value P_{mc} for GO term XX was calculated by

$$P_{mc} = \frac{\sum_{r=1}^{e} q_r}{e},$$

$$\text{where } q_r = \begin{cases} 1, & h_r \geqq f \\ 0, & h_r < f \end{cases} \quad (r = 1, \cdots, e).$$

We calculated P_{mc} for each GO term and selected GO terms with P_{mc} of 0.01 or less as significantly over-represented among genes appearing in the top ten models. To assess the stability of this approach based on the top ten models, we also performed the GO analysis using the top thirty models.

RESULTS

Individual prediction models fit by logistic regression

In the first learning step, no one-gene model met the selection criteria (all explanatory variables having likelihood ratio test p values below 0.01 and the model AIC value being 70 or less), whereas 72 two-gene models and 154,374 three-gene models were selected. In the second learning step, because the number of genes differs between the first and second learning data sets, only models using genes existing in both data sets could be analyzed; this restricted the analysis to 31 potential two-gene models and 7,051 potential three-gene models. Among these, one two-gene model and 739 three-gene models had error rates of 0.1

Table 1A. Two-gene model satisfying the selection criteria†

Covariate		Result of prediction using the second learning data			Model information	
			Number of errors / n			
X_1	X_2	Error rate	UF#	F*	Model form	AIC
MAOA	*MYCN*	0.08 (2/25)	2/8	0/17	X_1, X_2	62.2

† AIC 70 or less and error rate 0.1 or less.
# UF represents the group of patients with an unfavorable outcome.
* F represents the group of patients with a favorable outcome.

Table 1B. Top ten three-gene models satisfying the selection criteria†

Covariate			Result of prediction using the second learning data			Model information	
				Number of errors / n			
X_1	X_2	X_3	Error rate	UF#	F*	Model form	AIC
DDX1	*PRPH*	*TMEM66*	0.08 (2/25)	2/8	0/17	X_1, X_2, X_3	45.1
PRPH	*MYCN*	*POM121*	0.08 (2/25)	2/8	0/17	$X_1, X_2, X_3, X_{23}, X_{123}$	46.8
GNB1	*ENO1*	*HADHB*	0.08 (2/25)	2/8	0/17	X_1, X_2, X_3	46.9
MAOA	*MYCN*	*GAP43*	0.08 (2/25)	2/8	0/17	X_1, X_2, X_3	47.3
GNB1	*ENO1*	*TUBA3*	0.08 (2/25)	2/8	0/17	X_1, X_2, X_3	47.3
NCAM1	*MYCN*	*RPL4*	0.08 (2/25)	2/8	0/17	$X_1, X_2, X_3, X_{23}, X_{123}$	47.9
DDX1	*PRPH*	*VPS41*	0.08 (2/25)	2/8	0/17	X_1, X_2, X_3	47.9
CD44	*DDX1*	*GNB1*	0.08 (2/25)	2/8	0/17	X_1, X_2, X_3, X_{13}	48.2
MAOA	*MYCN*	*PRPH*	0.08 (2/25)	2/8	0/17	X_1, X_2, X_3	48.9
CD44	*DDX1*	*MORF4L2*	0.08 (2/25)	2/8	0/17	X_1, X_2, X_3	48.9

† AIC 70 or less and error rate 0.1 or less.
# UF represents the group of patients with an unfavorable outcome.
* F represents the group of patients with a favorable outcome.

or less.

The selected two-gene model is shown in Table 1A. This model incorporated *Monoamine oxidase A* [*MAOA*] and *V-myc myelocytomatosis viral related oncogene, neuroblastoma derived (avian)* [*MYCN*]. It did not have a significant interaction between *MAOA* and *MYCN*. Among the 739 three-gene models that satisfied our selection criteria, the top ten models are shown in Table 1B, ranked by AIC value. Genes appearing in the top ten models are shown in Table 2. Of 739 models, 116 had significant gene-gene interactions; thus, more than 80% of the selected models had no interactions. The gene most often included in these three-gene models was *MYCN*, appearing in 167 models. The second most frequently appearing gene was *CD44 molecule (Indian blood group)* [*CD44*] (118 models). The third most frequent was *Peripherin* [*PRPH*] (109 models). Moreover, among the models with interactions, *MYCN* was the gene most often included, appearing in about half (57) of these models. The second most frequently included gene was *Rho guanine nucleotide exchange factor (GEF) 7* [*ARHGEF7*], appearing in 20 models. The third most frequent was *CD44* (18 models). *MYCN, DEAD (Asp-Glu-Ala-Asp) box polypeptide 1* [*DDX1*], and *PRPH* appeared in four of the top ten models. *Guanine nucleotide binding protein (G protein) and beta polypeptide 1* [*GNB1*] appeared in three of the top ten models. *CD44*, *Enolase 1, (alpha)* [*ENO1*] and *MAOA* appeared in two of the top ten models. Three of the top ten three-gene models had significant interactions; *MYCN* interacted with another gene in two of these three models.

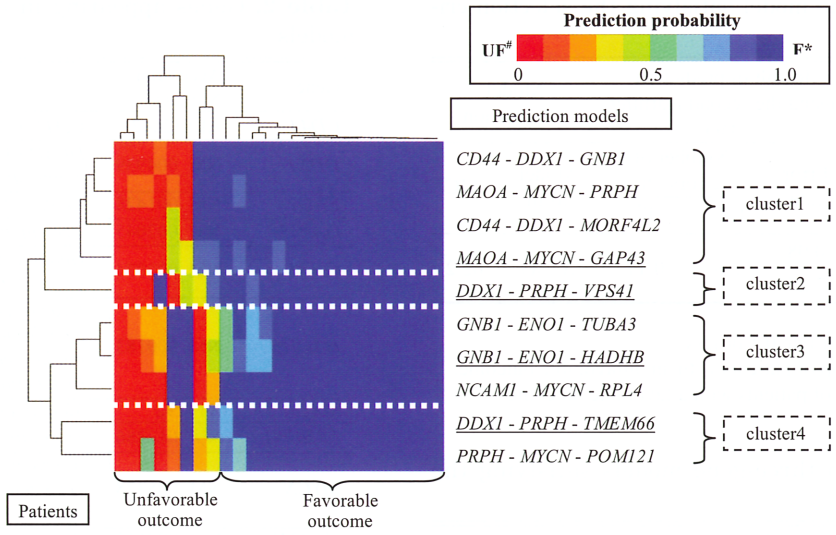
Table 2. Genes appearing in the top ten* three-gene models

Gene	Accession Number	Name
CD44	AL832642	*CD44 molecule (Indian blood group)*
DDX1	NM_004939	*DEAD (Asp-Glu-Ala-Asp) box polypeptide 1*
ENO1	NM_001428	*Enolase 1, (alpha)*
GAP43	NM_002045	*Growth associated protein 43*
GNB1	NM_002074	*Guanine nucleotide binding protein (G protein), beta polypeptide 1*
HADHB	NM_000183	*Hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), beta subunit*
MAOA	X17192	*Monoamine oxidase A*
MORF4L2	NM_012286	*Mortality factor 4 like 2*
MYCN	NM_005378	*V-myc myelocytomatosis viral related oncogene, neuroblastoma derived (avian)*
NCAM1	NM_000615	*Neural cell adhesion molecule 1*
POM121	AF036613	*POM121 membrane glycoprotein (rat)*
PRPH	NM_006262	*Peripherin*
RPL4	NM_000968	*Ribosomal protein L4*
TMEM66	AB028926	*Transmembrane protein 66*
TUBA3	NM_006082	*Tubulin, alpha 3*
VPS41	U87309	*Vacuolar protein sorting 41 (yeast)*

* The top ten three-gene models ranked according to AIC value.

Integrated phenotype prediction model

Cluster analysis using the top ten models and 25 patients of the second learning data set produced four clusters, as shown in Fig. 2. Many patients with an unfavorable outcome were located towards the left in the horizontal dimension. Taking the four representative models (models with the smallest AIC from each cluster), we searched for the best combination to construct the integrated phenotype prediction model. Combining two of the three-gene models led to an error rate of 0 in two instances: *GNB1-ENO1-HADHB* combined with *MAOA-MYCN-GAP43* and *GNB1-ENO1-HADHB* combined with *DDX1-*



* F represents patients with a favorable outcome (alive after 2 years).
# UF represents patients with an unfavorable outcome (died within 2 years).
Underlined models were chosen as representative models.

Fig. 2. Hierarchical clustering applied to the top ten models and 25 patients of the second learning data by predicted probability of prognosis

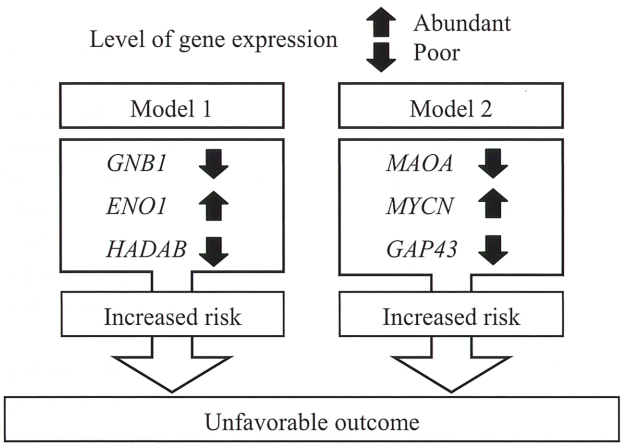


Fig. 3. The integrated prediction model composed of the logical sum of two selected models

Table 3. Integrated prediction model to predict prognosis in neuroblastoma patients

Model	Covariate			Result of prediction using the second learning data			Model information	
					Number of errors / n			
	X_1	X_2	X_3	Error rate	UF#	F*	Model form	AIC
Model 1	*GNB1*	*ENO1*	*HADHB*	0.12 (3/25)	2/7	1/18	X_1, X_2, X_3	46.9
Model 2	*MAOA*	*MYCN*	*GAP43*				X_1, X_2, X_3	47.3

UF represents the group of patients with an unfavorable outcome.
* F represents the group of patients with a favorable outcome.

PRPH-VPS41. Because the sum of AIC values for the combination *GNB1-ENO1-HADHB* and *MAOA-MYCN-GAP43* was smaller than that for *GNB1-ENO1-HADHB* and *DDX1-PRPH-VPS41*, we chose the former as the integrated phenotype prediction model (Fig. 3). To validate the performance of this model, we predicted the prognosis of the 25 patients in the independent validation data, obtaining an error rate of 0.12 (Table 3).

Biological differences between phenotype groups

A total of sixteen genes appeared among the top ten prediction models, as shown in Table 2. Among 215 GO terms that are involved with these 16 genes, those that were significantly over-represented (Monte Carlo p values of 0.01 or less) are listed in Table 4. Among the GO terms selected were: cell growth, G-protein signaling, alcohol metabolism, glycolysis, neurophysiological processes, and catecholamine catabolism. Among the genes in our integrated model, *GNB1* is involved in G-protein signaling and neurophysiological processes, *ENO1* in alcohol metabolism and glycolysis, *MAOA* in neurophysiological processes, and *GAP43* in cell growth and G-protein signaling. There were no GO terms involving *HADAB* or *MYCN* in the list of GO terms with a p value less than 0.01 among the top ten models. GO terms with Monte Carlo p values less than 0.01 for the 39 genes appearing in the top thirty models are similar to those selected for the top ten models, confirming the stability of our results based on the top ten models.

Table 4. GO terms significantly over-represented among genes appearing in the top ten models

GO term	P_{mc}	Genes	Number of genes	Mean number of genes by random selection*
regulation of cell growth	<0.001	*ENO1,GAP43, MORF4L2*	3	0.146
regulation of growth	<0.001	*ENO1,GAP43, MORF4L2*	3	0.188
cell growth	<0.001	*ENO1,GAP43, MORF4L2*	3	0.195
regulation of cell size	<0.001	*ENO1,GAP43, MORF4L2*	3	0.199
growth	0.001	*ENO1,GAP43, MORF4L2*	3	0.24
G-protein signaling, coupled to IP3 second messenger (phospholipase C activating)	0.001	*GAP43,GNB1*	2	0.046
alcohol metabolism	0.001	*DDX1,ENO1, MAOA*	3	0.254
phosphoinositide-mediated signaling	0.002	*GAP43,GNB1*	2	0.067
glycolysis	0.002	*DDX1,ENO1*	2	0.074
glucose catabolism	0.002	*DDX1,ENO1*	2	0.083
alcohol catabolism	0.002	*DDX1,ENO1*	2	0.087
hexose catabolism	0.002	*DDX1,ENO1*	2	0.087
monosaccharide catabolism	0.002	*DDX1,ENO1*	2	0.087
sensory perception of taste	0.004	*GNB1*	1	0.004
carbohydrate catabolism	0.004	*DDX1,ENO1*	2	0.104
cellular carbohydrate catabolism	0.004	*DDX1,ENO1*	2	0.104
neurophysiological process	0.004	*GNB1,MAOA, NCAM1*	3	0.332
glucose metabolism	0.004	*DDX1,ENO1*	2	0.115
glial cell differentiation	0.004	*GAP43*	1	0.004
main pathways of carbohydrate metabolism	0.005	*DDX1,ENO1*	2	0.117
hexose metabolism	0.007	*DDX1,ENO1*	2	0.147
cell organization and biogenesis	0.007	*DDX1,ENO1, GAP43,MORF4L2, POM121,VPS41*	6	1.904
morphogenesis	0.008	*CD44,ENO1, GAP43,MORF4L2*	4	0.848
monosaccharide metabolism	0.008	*DDX1,ENO1*	2	0.151
gliogenesis	0.008	*GAP43*	1	0.008
catecholamine catabolism	0.008	*MAOA*	1	0.008
dopamine catabolism	0.008	*MAOA*	1	0.008
peptidyl-proline modification	0.009	*POM121*	1	0.009
second-messenger-mediated signaling	0.009	*GAP43,GNB1*	2	0.144
generation of precursor metabolites and energy	0.009	*DDX1,ENO1, MAOA*	3	0.459

* Mean number of genes with a certain GO term included among 16 genes selected randomly (10000 repetitions).

DISCUSSION

Gene combinations and patterns of gene-to-gene interrelationships affecting phenotype

No one-gene model satisfied our selection criteria. However, one-gene models incorporating *MYCN*[3,21], *Neurotrophic tyrosine kinase receptor type 1* [*NTRK1*][13,14], *CD44*[7] or *FYN oncogene related to SRC, FGR, YES* [*FYN*][1] had likelihood ratio test p values less than 0.01. Furthermore, one-gene models incorporating *Cadherin 2 type 1 or N-cadherin (neuronal)* [*CDH2*][22] had p values less than 0.05. On the other hand, one-gene models incorporating *Pleiotrophin (heparin binding growth factor 8, neurite growth-promoting factor 1)* [*PTN*][15] or *Septin 7* [*SEPT7*][12] did not demonstrate significant likelihood ratio tests (data not shown). All seven of these genes have been reported to be markers of neuroblastoma prognosis. One two-gene model and 739 three-gene models satisfied our selection criteria. One of the known markers, *MYCN*, appeared in many of these models. The results suggest that the combination of *MYCN* and one or more other genes predicts well the prognosis of many patients. Thus, our analysis showed that expression of *MYCN* is an important factor for neuroblastoma prognosis, as was previously reported[17]. In addition, the present study suggests that *MYCN* tends to interact with other genes. It has been reported that elevated expression of *MYCN* induces the expression of many ribosomal proteins (RP) in neuroblastoma, which suggests that genes involved in the protein synthesis machinery are major targets of the *MYCN* protein[2,11]. *MYCN* might raise the efficiency of protein synthesis after transcription for several genes. Therefore, *MYCN* might demonstrate interactions with other genes. *MYCN* interacted with *NCAM1*, as seen in the top ten models. It has been reported that *MYCN* regulates the expression of NCAM cell-surface receptors; expression levels of both *NCAM1* and *MYCN* were elevated in SKNSH cells transfected with the vector containing *MYCN* cDNA[10].

It is possible that a phenotype is controlled by multiple pathways. Therefore, we constructed an integrated phenotype prediction model composed of multiple prediction models that were considered to reflect important pathways. In the present study, prognosis due to the integrated model is based on a logical sum of predicted unfavorable prognosis outcomes from individual prediction models. In future, other integrated models such as a logistic regression model using predicted probability from individual prediction models are worth examining, too.

Biological differences between patients with favorable and unfavorable outcomes

As mentioned above, the integrated phenotype prediction model is composed of two component models: *GNB1-ENO1-HADHB* (model 1) and *MAOA-MYCN-GAP43* (model 2). Each component model predicts prognosis based on characteristics of the cancer cells represented by expression of the three genes that appear in the model. Elucidation of the cancer-cell characteristics that influence prognosis is an important goal. Therefore, we examined the biological features of the sixteen genes appearing in the top ten models using GO terms. To assess the stability of the results, we also examined features of the 39 genes appearing in the top thirty models. Features of the genes in the top thirty models were similar to features of the genes in the top ten models.

Numerous genes involved in cell growth appeared in the top ten models. *ENO1* (model 1) is involved in cell growth, glucose catabolism, and alcohol catabolism. It has been reported that overexpression of *ENO1* is associated with survival outcome in lung cancer[5] and *ENO1* expression is elevated in hepatitis C virus-related hepatocellular carcinoma[23]. *ENO1* encodes the α subunit of enolase, which comprises three subunits: α, β, and γ. Enolase is an essential glycolytic enzyme and is a known molecular marker of advanced neuroblastoma. Cancer cells preferentially use anaerobic glycolysis for inefficient energy metabolism (the Warburg effect). A high level of *ENO1* expression increases the risk of unfavorable outcomes in model 1. The degree of activation of glycolysis might differ between patients with favorable and unfavorable outcomes. *GAP43* (model 2) is involved in cell growth and G-protein signaling; its expression is induced in the early stages of neuronal differentiation. In addition, it has been reported that expression of *GAP43* is induced by the cytokine TGF-β[25], which modulates neuroblastoma cell proliferation and differentiation *in vitro*[20]. A low level of *GAP43* expression increases the risk of unfavorable outcomes in model 2. Aberrant cell differentiation, or the lack of TGF-β signal to inhibit cell proliferation, might be a cause of low levels of *GAP43* expression. Hence, neuroblastoma cells of patients with favorable and unfavorable outcomes might differ in terms of cell growth.

Our GO analysis also suggested that patients with favorable and unfavorable outcomes might differ in terms of G-protein signaling and phosphoinositide-mediated signaling. *GNB1* (model 1) is involved in G-protein signaling and phosphoinositide-mediated signaling. *GNB1* encodes the β subunit of the G-protein, which is composed of α, β, and γ subunits. Deletion of the 1p36 region has been reported in neuroblastoma; the *GNB1* gene is located in this region[9]. A low level of *GNB1* expression increases the risk of unfavorable outcome in model 1. *GAP43* (model 2) is also involved in G-protein signaling and phosphoinositide-mediated signaling. This gene is involved in TGF-β signaling, which inhibits cell proliferation, as described above.

Our GO analysis also showed that there might be differences in alcohol metabolism and glucose catabolism between patients with favorable and unfavorable outcomes. *ENO1* (model 1) is involved in glycolysis, as described above. A difference in glycolytic energy supply may lead to a difference in cell growth, thereby affecting prognosis.

Many genes involved in neurophysiological processes appeared in the top ten models. *MAOA* (model 2) codes an enzyme that degrades amine neurotransmitters, such as dopamine and serotonin. A low level of *MAOA* expression increases the risk of unfavorable outcomes in model 2. Aberrant neurophysiological processes might affect patient outcome. Catecholamines, such as dopamine, are biomarkers of neuroblastoma.

Our GO analysis suggests that model 1 predicts patient outcome by the status of G-protein and phosphoinositide-mediated signaling mediated by *GNB1* and cell proliferation mediated by *ENO1*, whereas model 2 predicts patient outcome by the status of neurophysiological processes mediated by *MAOA* and G-protein and phosphoinositide-mediated signaling mediated by *GAP43*. There were no GO terms involving *HADAB* or *MYCN* in the list of GO terms with a p value less than 0.01 among the top ten models.

In summary, we developed a methodology for constructing an integrated phenotype prediction model by combining multiple prediction models, and we also developed a methodology for inferring biological mechanisms from the integrated phenotype prediction model using GO terms. Our results suggest that prognosis of neuroblastoma patients may be affected by biological processes such as cell growth, G-protein signaling, phosphoinositide-mediated signaling, alcohol metabolism, glycolysis, neurophysiological processes, and catecholamine catabolism. Although these results require confirmation by biological experiments, we think that our method is useful for inferring the biological mechanisms controlling a phenotype by combining information on multiple pathways, the combination of genes affecting the phenotype, and the patterns of gene-to-gene interrelationships. As information accumulates on GO, we will be better able to understand the biological mechanisms using our methods.

ACKNOWLEDGMENTS

We are very grateful to Shingo Dan for helpful discussion. We are also grateful to Miki Ohira for helpful comment. We also thank Hitoshi Fujimiya for providing the opportunity to conduct this study.

(Received November 5, 2007)
(Accepted December 6, 2007)

REFERENCES

1. **Berwanger, B., Hartmann, O., Bergmann, E., Bernard, S., Nielsen, D., Krause, M., Kartal, A., Flynn, D., Wiedemeyer, R., Schwab, M., Schafer, H., Christiansen, H. and Eilers, M.** 2002. Loss of a FYN-regulated differentiation and growth arrest pathway in advanced stage neuroblastoma. Cancer Cell **2**:377-386.
2. **Boon, K., Caron, H.N., van Asperen, R., Valentijn, L., Hermus, M.C., van Sluis, P., Roobeek, I., Weis, I., Voute, P.A., Schwab, M. and Versteeg, R.** 2001. N-myc enhances the expression of a large set of

genes functioning in ribosome biogenesis and protein synthesis. EMBO J. **20**:1383-1393.

3. **Brodeur, G.M., Seeger, R.C., Schwab, M., Varmus, H.E. and Bishop, J.M.** 1984. Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. Science **224**:1121-1124.
4. **Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M.,Jr. and Haussler, D.** 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc.Natl.Acad.Sci. USA. **97**:262-267.
5. **Chang, G.C., Liu, KJ., Hsieh, C.L., Hu, TS., Charoenfuprasert, S., Liu, HK., Luh, KT., Hsu, LH., Wu, CW., Ting, CC., Chen, CY., Chen, KC., Yang, TY., Chou, TY., Wang, WH., Whang-Peng, J. and Shih, NY.** 2006. Identification of alpha-enolase as an autoantigen in lung cancer: its overexpression is associated with clinical outcomes. Clin. Cancer Res. **12**:5746-5754.
6. **Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D.** 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl.Acad. Sci. USA. **95**:14863-14868.
7. **Favrot, M.C., Combaret, V. and Lasset, C.** 1993. CD44--a new prognostic marker for neuroblastoma. N.Engl.J.Med. **329**:1965.
8. **Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S.** 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**:531-537.
9. **Janoueix-Lerosey, I., Novikov, E., Monteiro, M., Gruel, N., Schleiermacher, G., Loriod, B., Nguyen, C. and Delattre, O.** 2004. Gene expression profiling of 1p35-36 genes in neuroblastoma. Oncogene **23**:5912-5922.
10. **Judware, R. and Culp, L.A.** 1995. Over-expression of transfected N-myc oncogene in human SKNSH neuroblastoma cells down-regulates expression of beta 1 integrin subunit. Oncogene **11**:2599-2607.
11. **Krasnoselsky, A.L., Whiteford, C.C., Wei, J.S., Bilke, S., Westermann, F., Chen, Q.R. and Khan, J.** 2005. Altered expression of cell cycle genes distinguishes aggressive neuroblastoma. Oncogene **24**:1533-1541.
12. **Nagata, T., Takahashi, Y., Asai, S., Ishii, Y., Mugishima, H., Suzuki, T., Chin, M., Harada, K., Koshinaga, S. and Ishikawa, K.** 2000. The high level of hCDC10 gene expression in neuroblastoma may be associated with favorable characteristics of the tumor. J.Surg.Res. **92**:267-275.
13. **Nakagawara, A., Arima, M., Azar, C.G., Scavarda, N.J. and Brodeur, G.M.** 1992. Inverse relationship between trk expression and N-myc amplification in human neuroblastomas. Cancer Res. **52**:1364-1368.
14. **Nakagawara, A., Arima-Nakagawara, M., Scavarda, N.J., Azar, C.G., Cantor, A.B. and Brodeur, G.M.** 1993. Association between high levels of expression of the TRK gene and favorable outcome in human neuroblastoma. N.Engl.J.Med. **328**:847-854.
15. **Nakagawara, A., Milbrandt, J., Muramatsu, T., Deuel, T.F., Zhao, H., Cnaan, A. and Brodeur, G.M.** 1995. Differential expression of pleiotrophin and midkine in advanced neuroblastomas. Cancer Res. **55**:1792-1797.
16. **Nguyen, S.T., Hasegawa, S., Tsuda, H., Tomioka, H., Ushijima, M., Noda, M., Omura, K. and Miki, Y.** 2007. Identification of a predictive gene expression signature of cervical lymph node metastasis in oral squamous cell carcinoma. Cancer Sci. **98**:740-746.
17. **Ohira, M., Oba, S., Nakamura, Y., Isogai, E., Kaneko, S., Nakagawa, A., Hirata, T., Kubo, H., Goto, T., Yamada, S., Yoshida, Y., Fuchioka, M., Ishii, S. and Nakagawara, A.** 2005. Expression profiling using a tumor-specific cDNA microarray predicts the prognosis of intermediate risk neuroblastomas. Cancer Cell **7**:337-350.
18. **Raychaudhuri, S., Stuart, J.M. and Altman, R.B.** 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac.Symp.Biocomput. 455-466.
19. **Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.I., Gascoyne, R.D., Muller-Hermelink, H.K., Smeland, E.B., Giltnane, J.M., Hurt, E.M., Zhao, H., Averett, L., Yang, L., Wilson, W.H., Jaffe, E.S., Simon, R., Klausner, R.D., Powell, J., Duffey, P.L., Longo, D.L., Greiner, T.C., Weisenburger, D.D., Sanger, W.G., Dave, B.J., Lynch, J.C., Vose, J., Armitage, J.O., Montserrat, E., Lopez Guillermo, A., Grogan, T.M., Miller, T.P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T. and Staudt, L.M.** 2002. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N.Engl.J.Med. **346**:1937-1947.
20. **Scarpa, S., Coppa, A., Ragano-Caracciolo, M., Mincione, G., Giuffrida, A., Modesti, A. and Colletta, G.** 1996. Transforming growth factor beta regulates differentiation and proliferation of human neuroblastoma. Exp.Cell Res. **229**:147-154.
21. **Schwab, M., Alitalo, K., Klempnauer, K.H., Varmus, H.E., Bishop, J.M., Gilbert, F., Brodeur, G., Goldstein, M. and Trent, J.** 1983. Amplified DNA with limited homology to myc cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour. Nature **305**:245-248.
22. **Shimono, R., Matsubara, S., Takamatsu, H., Fukushige, T. and Ozawa, M.** 2000. The expression of cadherins in human neuroblastoma cell lines and clinical tumors. Anticancer Res. **20**:917-923.
23. **Takashima, M., Kuramitsu, Y., Yokoyama, Y., Iizuka, N., Fujimoto, M., Nishisaka, T., Okita, K., Oka, M. and Nakamura, K.** 2005. Overexpression of alpha enolase in hepatitis C virus-related hepatocellular carcinoma: association with tumor progression as determined by proteomic analysis. Proteomics **5**:1686-1692.
24. **Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R.** 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc.Natl.Acad.Sci. USA **96**:2907-2912.
25. **Turco, A., Scarpa, S., Coppa, A., Baccheschi, G., Palumbo, C., Leonetti, C., Zupi, G. and Colletta, G.** 2000. Increased TGFbeta type II receptor expression suppresses the malignant phenotype and induces differentiation of human neuroblastoma cells. Exp. Cell Res. **255**:77-85.