Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 37 (2004) 260–268
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Discovering significant and interpretable patterns from multifactorial DNA microarray data with poor replication ☆

Ju Han Kim[a,b], Dooil Jeoung[c], Seongeun Lee[d], Hyeouneui Kim[e,*]

[a] Seoul National University Biomedical Informatics (SNUBI), Seoul 110-799, Republic of Korea
[b] Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Republic of Korea
[c] Department of Microbiology, Kangwon National University, Chuncheon 200-701, Republic of Korea
[d] In2Gen, 28 Yongon-dong Chongno-gu, Seoul 110-799, Republic of Korea
[e] Graduate Program in Health Informatics, University of Minnesota, Minneapolis, MN 55455, USA

Received 16 June 2004

Abstract

Motivation. Multivariate analyses are advantageous for the simultaneous testing of the separate and combined effects of many variables and of their interactions. In factorial designs with many factors and/or levels, however, sufficient replication is often prohibitively costly. Furthermore, complicated statements are often required for the biological interpretation of the higher-order interactions determined by standard statistical techniques like analysis of variance.

Results. Because we are usually interested in finding factor-specific effects or their interactions, we assumed that the observed expression profile of a gene is a manifestation of an underlying factor-specific generative pattern (FSGP) combined with noise. Thus, a genetic algorithm was created to find the nearest FSGP for each expression profile. We then measured the distance between each profile and the corresponding nearest FSGP. Permutation testing for the distance measures successfully identified those genes with statistically significant profiles, thus yielding straightforward biological interpretations. Association networks of genes, drugs, and cell lines were created as tripartite graphs, representing significant and interpretable relations, by using a microarray experiment of gastric-cancer cell lines with a factorial design and no replication. The proposed method may benefit the combined analysis of heterogeneous expression data from the growing public repositories.

Keywords: Gene expression; DNA microarray; Pattern recognition; Genetic algorithm; Gastric cancer

1. Introduction

Factorial designs have advantages in efficiency, power, and in the elegance of statistical testing [26]. Generalizations based on factorial experiments are broader than those obtained from single variable experiments, as the effect of treatment is studied across different conditions.

Replication is a cornerstone of scientific research. The importance of replication in microarray experiments has been highlighted as a means of increasing the precision of estimated quantities and of providing information about the uncertainties of estimates [16,19]. However, we often have to deal with poorly replicated experimental datasets because of unwanted limitations in resource, methodology, or knowledge.

DNA microarrays measure thousands of gene expression levels in a massively parallel way such that even a single classical two-dye technique microarray experiment may result in a reasonable estimation of statistical significance and experimental quality control.

☆ Availability: http://www.snubi.org/software/FSGP/.

* Corresponding author. Fax: +82 2 747 4830.
E-mail address: kimx0519@tc.umn.edu (H. Kim).

1532-0464/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2004.07.011

For example, the reliable identification of differentially expressed genes in a single-slide experiment has been demonstrated by a method choosing cut-offs in the distribution of ratios [7] and by a hierarchical Bayesian model based on posterior odds change [22]. Methods that determine differentially expressed genes using only a few replicates have also been introduced [1,34].

Previous methods relying on comparisons of two or more levels of a single factor, however, do not apply to multivariate cases. One needs a method flexible enough to allow for complex experimental design. Analysis of variance (ANOVA) is standard technique for analyzing such multivariate datasets.

One experimental dilemma, however, concerns the determination of the optimal number of replicates. In a factorial design with many factors and/or levels, replication may often be prohibitively costly. For example, in a 10 by 10, 2-factor 10-level experimental design, 100 microarray slides are needed with no replication. Even simple triplication of the experiment, which may not be sufficient, requires 200 more microarray slides. Clearly, the disadvantages of such replication, given limited resources, is to lose the opportunity to systematically explore the high-dimensional problem space enriched by many interesting factors that we want to measure. For example, a multivariate analysis technique requiring no replication may permit one to systematically explore 200 more factor levels (i.e., 10×20) using a 10 by 30 experimental design with no replication.

Biological interpretability, its clarity and relevance may be the most desired properties of a good microarray data analysis technique. However, biological interpretability of the results of standard ANOVA-type statistical methods may not be guaranteed in a factorial design with many factors and levels because of the large number of statistically significant higher-order interactions. As correctly pointed out by Pavlidis and Noble [24], when more than two levels are present for variables, ANOVA might indicate a significant effect of a factor on expression, but does not determine which factor levels show different expression from any others. Moreover, describing statistically significant higher-order interactions in a multifactorial experiment typically requires an extremely complicated statement.

With the growing number of microarray standards such as MIAME (Minimum Information About a Microarray Experiment, [4]) and MAGE-ML (Microarray Gene Expression Markup Language, [32]), and of public expression-data repositories such as ArrayExpress [5] and GEO [9], we clearly require more powerful analytical methods to discover the significant and interpretable patterns from large multifactorial microarray data with poor or insufficient replicates, to facilitate the mining of a huge collection of microarray data from heterogeneous sources.

In this paper, we propose a procedure for identifying genes that show both substantive and interpretable gene-expression patterns in multifactorial microarray experiments with no or poor replication. The proposed method identifies the optimal combination of biological factor(s) explaining the expression profile of the differentially expressed genes. In effect, we test all (biologically) interpretable patterns from a multifactorial design, select the nearest pattern for each expression profile, and evaluate statistical significance.

First, we define a factor-specific generative pattern (FSGP, see method), which is readily interpretable, and which represents all interpretable patterns enumerated from the particular design involved. We define the distance between a gene expression profile and FSGP, find the FSGP nearest each gene's expression profile, and measure the distance between each profile and the corresponding nearest FSGP. A Genetic Algorithm (GA) is created to determine the FSGP nearest a gene expression profile. Finally, we determine the FDR (false discovery rate)-corrected statistical significance of the distance by permutation testing.

The proposed procedure is illustrated using 54 cDNA microarray experiments with two factors (i.e., six chemotherapeutic agents and nine gastric cancer cell lines) where the cancer cell lines are labeled before and after a chemo-drug treatment. The procedure demonstrates how to reliably identify genes with drug and/or cancer-specific expression patterns, yielding straightforward biological interpretations, for multifactorial microarray data with poor replication.

This paper is organized as follows. In Section 2, we define FSGP and describe a GA implementation designed to find the FSGP nearest a given expression profile. Section 3.1 describes data preprocessing steps. Section 3.2 describes our permutation scheme for the strong control of type-I error. Genes identified by the proposed method are listed and investigated in relation to the associated drugs and cell lines. The association networks of the genes, drugs, and cell lines are reconstructed as tripartite graphs in Section 3.3, and this is followed by a discussion in Section 4.

2. Methods

2.1. Factor-specific generative patterns and pattern distance

Suppose that there are N factors denoted by n $(=1, \ldots, N)$ and K levels for each factor denoted by k_n $(=1_n, \ldots, K_n)$, a typical experimental design for the multifactorial analysis involves $\prod K_n$ microarrays.

An expression profile (or a pattern) of a gene can be represented by an N-dimensional matrix with $\prod K_n$ cells

denoted $C_{k_1},\ldots,k_N$. A simple two-dimensional case is illustrated in Fig. 1 ($N = 2$, $K_1 = 6$, and $K_2 = 9$).

Let f_n be a particular level of factor n such that $f_n \in \{1_n, 2_n, \ldots, K_n\}$. A pattern is defined to be specific to f_n if all $C_{k_1,\ldots,k_{n-1},f_n,k_{n+1},\ldots,k_N}$ are significantly changed.

A factor-specific generative pattern, $FSGP(k_i, k_j, \ldots)$, is defined as a pattern that is specific to all $k_i, k_j, \ldots$ and not to others. Therefore, there are in general $2^{\sum K_n}$ FSGPs for a microarray experiment with $\prod K_n$ slides.

As defined above, an expression pattern in a multifactorial design is generally defined as "specific" to one or more factors if the expression levels of all cells related to the factor(s) are all significantly changed. For example, the expression pattern in Fig. 1A is specific to the factors, t_1, t_2, t_3, and t_4, but not to others. Fig. 1C profile is specific to c_2 and 'nearly' specific to c_6 and t_3. Notice that we use the term, 'factor,' although 'level' may be a more precise term for multifactor design. We use the term for convenience and to prevent possible confusion between factor 'level' and gene-expression 'level.'

An FSGP is an expression pattern specific to a (combination of) factor(s) (or more precisely, factor level(s)). We view an observed expression profile as a manifestation of the underlying FSGP combined with noise. For example, we view that the expression profile in Fig. 1B is likely to be generated by the t_3-specific generative pattern, $FSGP(t_3)$, but has a pattern distance 1 from $FSGP(t_3)$ because of intervening noise. The FSGPs denoted by $FSGP(a_i, b_j, \ldots)$ in Fig. 1 can also be represented by a list of bit patterns, as such the pattern in Fig. 1C can be denoted as $FSGP((0,0,1,0,0,0),(0,1,0,0,0,1,0,0,0))$. Accordingly, there are in general $2^{\sum K_n}$ FSGPs for a microarray experiment with $\prod K_n$ slides.

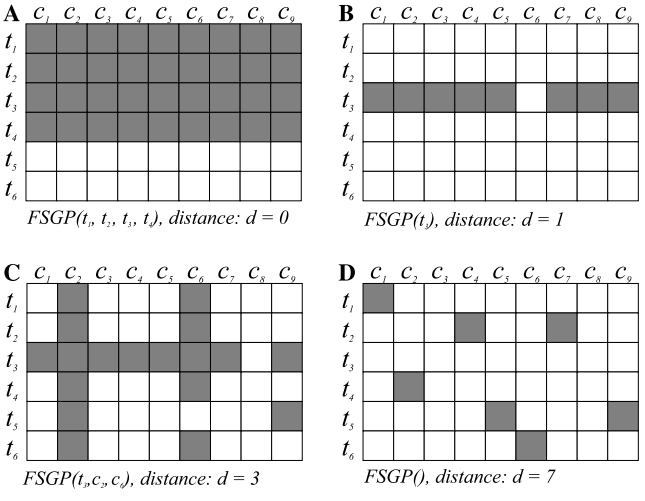

Fig. 1. Gene expression profiles with measured pattern distances with respect to the nearest factor-specific generative patterns. (A) t_1, t_2, t_3, t_4-specific generative pattern ($d = 0$), (B) t_3-specific generative pattern ($d = 1$), (C) t_3, c_2, c_6-specific generative pattern ($d = 3$), and (D) non-specific pattern or non-pattern ($d = 7$). Pattern distance in the binary space is simply the number of mismatches (i.e., the Hamming distance). c_i, the ith cell-line; t_i, the ith anti-cancer drug.

We define the pattern distance of an expression profile as the distance between the profile and the nearest FSGP. Thus, to measure the pattern distance of an expression profile, one first has to find its nearest FSGP. Section 2.2 demonstrates how to find the FSGP nearest a profile by implementing a GA. For the purpose of illustration, we first demonstrate the pattern distance in binary space and then generalize it.

The pattern distance is defined as (a kind of) Hamming distance between an expression profile and its nearest FSGP. In the binary space, where the expression level is dichotomized to zero (i.e., non-changed and depicted as blank cells) or one (i.e., significantly changed and depicted as dark cells), pattern distance simply equals the number of mismatches (Fig. 1). Therefore, the pattern distance between a two-factor profile C_{ij} and the nearest FSGP, C'_{ij}, equals $\sum |C_{ij} - C'_{ij}|$. More generally, the N-dimensional pattern distance of $C_{k_1,\ldots,k_N}$ equals $\sum |C_{k_1,\ldots,k_N} - C'_{k_1,\ldots,k_N}|$.

Fig. 1A demonstrates a case in which the nearest FSGP perfectly matches the observed expression profile, as such the pattern distance equals to zero. Fig. 1B exhibits an expression profile with one mismatch ($FSGP(t3), d = 1$) and (C) one with three mismatches ($FSGP(t3, c2, c6)$, $d = 3$). Moreover, a profile's nearest FSGP can be the null-pattern, $FSGP()$, as shown in Fig. 1D. Because we are interested in search for distinct patterns rather than a null-pattern, the distributions of pattern distances from the null-pattern is estimated separately and treated in the manner described in Section 3.2. The significance of the nearest FSGP obtained can be determined reliably by using the permutation test described in Section 3.2.

Generalization to non-binary space can easily be achieved by setting the range of expression levels in cells. For example, if the range of the expression level is .0–1.0 and the dark cell represents .9 and the blank cell .1 in Fig. 1B, then, because the $FSGP(t3)$ should ideally have ideal value of 1.0 for the cells in the third row ($t3$) and 0 for those in the extra rows, the pattern distance of Fig. 1B profile can be calculated as follows: $\{|1.0 - .9| * 8 + |1.0 - .1| * 1\} + \{|.0 - .1| * (9(\text{columns}) * 5(\text{rows}))\} = .8 + .9 + 4.5 = 6.2$. GAs for both binary and non-binary conditions are available at http://www.snubi.org/software/FSGP/.

2.2. Genetic algorithm used to find the nearest generative pattern

The GA is a non-deterministic optimization procedure based on a massively parallel search [14], where each potential solution to a problem is represented in the form of a string (or a chromosome) with encoded parameters (or attributes). A series of random strings

(or population) are initialized to represent data points in solution space and then evolve towards fitness by mutation, mating, and crossover. The selection process applies a fitness function to measure the goodness of fit to select the next generation population. The Boltzmann probability distribution may also be applied to affect acceptance or rejection based on an analogy to the stochastic free energy optimization.

To find the nearest FSGP of an observed gene expression profile, a simple GA was created using a classical binary fixed-string representation [11] for factor-specific effects. Factors (N) were represented as chromosomes whose lengths equaled the corresponding number of levels (K_n). For example, the nearest FSGP for the Fig. 1B profile is represented as $(0,0,1,0,0,0)$ in chromosome 1 and as $(0,0,0,0,0,0,0,0,0)$ in chromosome 2, representing drug- and cell-line specific effects, respectively. The nearest FSGP for the Fig. 1C profile is represented as $((0,0,1,0,0,0),(0,1,0,0,0,1,0,0,0))$. Therefore, the solution space has the size of 2^{6+9} FSGPs. The chromosomal representation can easily be scaled up to more than two factors and to many levels with $2^{\sum K_n}$ FSGPs. Both the binary and the non-binary (i.e., continuous-value) pattern distances described in Section 2.1 can be applied as measures of goodness of fit. Fairly standard implementation of GA with mutation, mating and crossover successfully identified the nearest FSGP for each expression profile used. A GA implementation for user evaluation with adjustable parameters written in the Python programming language is available at http://www.snubi.org/software/FSGP/.

3. Results

3.1. Data set and normalization

A data set from gastric-cancer research was studied. The primary goal of the study was to explore the potential gene-drug interactions in a search for novel drug targets. DNA microarray slides were prepared containing 2400 fully annotated genes. Six chemotherapeutic agents were administered to nine gastric-cancer cell lines, resulting in 54 experiments. A classical two-dye technique with Cy5 and Cy3 fluorescent dyes was applied before (Cy3) and after (Cy5) anti-cancer drug treatment.

Variance stabilizing normalization by Huber et al. [15] was applied with the 'vsn' package in Bioconductor using the *R* statistical package. After performing intensity-dependent global LOWESS regression, spatial and intensity-dependent effects were managed by pin-group LOWESS normalization, and this was followed by applying the approach described by Yang et al. [35].

For the purpose of illustration, we assigned dummy binary values to the data by applying six cut-off levels (i.e., 5, 10, 15, 20, 25, and 30%) and three-direction groups (i.e., 'up,' 'down,' or 'up-or-down' regulated groups). For example, we assigned 1's to the highest 5% and 0's to the others when we applied a 5% cut-off level. Overall, we created 18 dichotomized data sets (i.e., six cutoffs by three directions), which described the expression profiles of 2400 genes under 54 conditions.

3.2. Permutation test and false discovery rate

To perform formal statistical testing, we wanted to estimate the null distribution(s) for the proposed statistic, the pattern distance defined in Section 2.1. When scoring thousands of gene expression profiles simultaneously, we also had to deal with the problem of "multiple hypothesis testing." Two types of error measurements are commonly used in multiple-hypothesis testing: FWER (family wise error rate) and FDR (false discovery rate). FWER offers a very strict error measure of at least one false positive result among all significant hypotheses. FDR [2,33] is defined as the expected proportion of false positive results among all rejected hypotheses multiplied by the probability of making at least one rejection. FDR offers a much less strict criterion, which hence leads to an increase in statistical power. Genes with pattern distances greater than a threshold are considered potentially significant. The percentage of such genes identified by chance is the false discovery rate (FDR). We applied permutation test that does not require any distributional assumptions.

We ran ca. five million permutations of the 54 sample labels in the present study. It is possible to reduce the number of permutations by testing for all possible categories. For example, the permutation results are the same expression profiles having the same number of significant cells, i.e., the cells with 1s in the matrix. Thus, there are only 55 conditions (i.e., 0–54 1's in the six-by-nine matrix) requiring permutation. By permuting the 54 sample labels 10^5 times for the 55 groups and by applying GA to find the nearest FSGP for each gene, we ran the GA implementation over ca. 3 days using 18-nodes of a Linux cluster system. It should be noted that the illustration used in our study could be extended to more general cases with unbalanced factors and levels. Users may flexibly extend the proposed method for different data sets, and the number of testing permutations can be greatly reduced by excluding the null-pattern groups.

To estimate the FDR, the proportion of falsely significant expression profiles corresponding to the expression profile of each gene were computed by counting the number of permuted profiles showing equal or smaller pattern distances (to the corresponding nearest FSGP) than that of the observed (i.e., non-permuted) profile. The threshold can be adjusted to identify smaller or larger sets of profiles, and the FDRs are calculated for each set (Fig. 2).

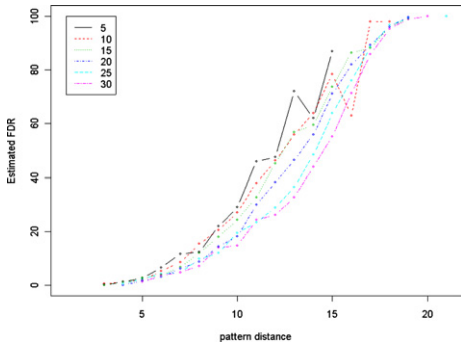

Fig. 2. Distribution of estimated FDR (false discovery rate). By permuting the sample labels, estimated distributions of FDRs were obtained across all levels of pattern distances and different six cut-off levels in the 'up-or-down' regulated group. The other two groups showed the same pattern (data not shown).

The FDR at each level of pattern distance was estimated by averaging the number of falsely significant expression profiles at each pattern distance level. Fig. 2 shows the estimated FDRs at all levels of pattern distance for six different cutoffs in the 'up-or-down' regulated group. The other two directional groups exhibited the same pattern (data not shown). For the purpose of illustration, the frequency and FDR (±SD) plots are overlapped with those of the 'up-or-down' regulated group at six cutoffs in Fig. 3. The other two directional groups showed the same pattern (data not shown). Fig. 4 demonstrates the distributions of the estimated FDRs at all pattern-distance levels across six cutoffs in the three directional groups. It seems that about 5% of the FDR can be obtained by applying a threshold of pattern distance, $d = 5$, for the six cutoffs.

3.3. Genes showing drug and cell-line specific patterns

For illustration purposes, the (two tailed) 'up-or-down' regulated group at the 10% cut-off level, which may be the most biologically relevant, was selected for further investigation. Thirty-seven genes were determined to be significant after applying strict control of the type-I error at FDR < 0.025, according to the proposed method (Table 1). We identified 17 more genes at FDR < 0.05.

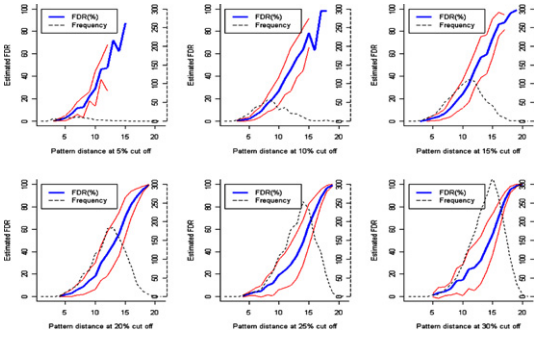

Fig. 3. Estimated FDRs (mean ± SD). Estimated FDRs of the 'up-or-down' regulated group are overlapped by the corresponding frequency histograms at all levels of pattern distance and at six cut-off levels, 5, 10, 15, 20, 25, and 30%.

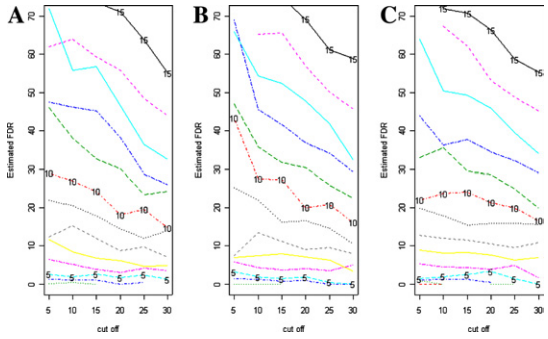

Fig. 4. Estimated FDRs across six cut-off and at all pattern-distance levels in the (A) 'up-or-down,' (B) 'up,' and (C) 'down' regulated groups.

It is worth noting that the proposed method reduces the effect of higher-order interactions from a multiplicative ($\prod K_n$, i.e., an effect of N factors on an expression level) down to an additive ($\sum K_n$, i.e., an effect of a factor on an expression profile) complexity, given the assumption that the factors independently affect gene expression level. By doing so, it greatly improves the interpretability of the analysis by permitting us to reconstruct an $(N + 1)$-partite-graph of the whole association network, which is equivalent to the set of dyadic associations identified as in Table 1.

k-Partite graph is a set of graph vertices decomposed into k disjoint sets such that no N graph vertices within the same set are adjacent. One can easily construct a tripartite graph by simply combining the dyadic associations among the three types of vertices, i.e., genes, drugs, and cell lines, described in Table 1.

Fig. 5 exhibits the tripartite-graph representation of the association network for the significant 54 genes, six drugs, and nine cell lines (FDR < 0.05) identified by the proposed method. Genes and associations having FDR < 0.025 are represented by bold characters and bold lines and those having FDR < 0.05 by italic characters and thin lines. Among the 37 genes at FDR < 0.05, 12 genes demonstrated a single drug-specific effect, 20 genes a single cell-line effect, three genes a dual cell-line effect, and two genes showed both drug and cell-line effects (Table 1 and Fig. 5).

Doxorubicin, an anthracycline antibiotic produced by the fungus *Streptomyces peucetius*, was found to specifically effect the MHC class I HLA-C-α-2 chain, the HLA class I locus C heavy chain, the immunoglobulin κ light chain and the interferon-inducible protein (IFI616, G1P3). Interestingly, all genes related to the anthracycline antibiotic were immune-response-related genes. Moreover, interferon α has been shown to modify the anti-tumor effect of doxorubicin and reduce bladder-

Table 1
List of genes showing cancer and/or drug-specific effects and the pattern distance, FDR (false discovery rate), and the number of cells significantly changed for each gene expression profile

Gene description	*d*	No[a] (=*n*)	Drug effect	Cell-line effect
Nuclear aconitase mRNA, encoding mitochondrial	3	(6)	CPT-11	
Transcription factor ZFM1 isoform B3. SF1: splicing factor 1	3	(3)		SNU601
α-2-Macroglobulin	4	(6)		SNU601
Immunoglobulin κ light chain	4	(7)	Doxorubicin	
Ferritin heavy chain	4	(6)		SNU1
Serine protease (Omi). PRSS25, protease, serine, 25	4	(6)		SNU620
Cytochrome *b*5. NQO1: NAD(P)H dehydrogenase, quinone 1	5	(7)		AGS
DNA polymerase epsilon, catalytic polype	5	(7)		SNU216
MAPK6: mitogen activated PK 6. (ERK3 protein kinase.)	5	(6)	Cisplatin	
Farnesyltransferase α-subunit	5	(7)		M1
RNA for *c*-fes. FES: felin sarcoma oncogene	5	(7)		M1
X-box binding protein-1 (XBP-1)	5	(7)		AGS
FABP5: Fatty acid binding protein homologue (psoriasis associated)	5	(7)		M74
GDNF family receptor α 2 (GFRA2)	5	(7)		M74
Zinc finger protein FPM315 (ZNF263)	5	(7)		SNU668
Putative src-like adapter protein (SLAP)	5	(7)		SNU620
Pyrroline 5-carboxylate reductase	6	(12)		AGS, SNU1
Rohu mRNA for rhodanese (HSROHU)	6	(9)	Cisplatin	
mRNA for APRIL protein. Acidic protein rich in leucines	6	(8)		AGS
Glutathione peroxidase (GPX1)	6	(8)		AGS
mRNA for P1cdc47. MCM7 minichromosome maintenance deficient 7	6	(8)		SNU601
(clone PWHCLC2-8) cardiac myosin light chain 2	6	(8)		SNU668
rhoGAP protein	6	(8)		M74
Human MHC class I HLA-C-α-2 chain and alternative mRNA	6	(7)	Doxorubicin	
UBE3A:ubiquitin prot. Ligase E3A (HPV E6-asso prot., Angelman synd)	6	(7)	Taxol	
KIAA0406	6	(7)	Taxol	
mRNA for Pr22 protein. STMN1: stethmin 1/onprotein 18	6	(8)		SNU601
LAMA2: laminin, α 2 (merosin, congenital muscular dystrophy) M chain	7	(13)		SNU620,SNU1
mRNA for KIAA0385 gene	7	(9)		SNU620
Human mRNA for HLA class I locus C heavy chain	7	(8)	Doxorubicin	
Ras-related protein (Krev-1). RAP1A: RAP1A, member of RAS oncogene family	7	(8)	Cisplatin	
mRNA for KIAA0288 gene	7	(7)		M74, SNU1
Myo-inositol monophosphatase 2	8	(11)	CPT-11	
SLC1A4: solute carrier family 1 (glutamate/neutral amino acid transporter), member 4	8	(12)	Taxol,	M1
SLC16A3: solute carrier family 16 (MCT3, Monocarboxylate transporters), member 3	9	(14)	TSA,	SNU601,SNU620
G1P3: interferon, α-inducible protein (cDNA, IFI616)	10	(15)	Doxorubicin	
CDC2: cell division cycle 2, G1 to S and G2 to M	10	(13)	5-FU	

[a] Number of the cells showing significantly changed expression levels.

cancer proliferation [23]. Doxorubicin damages DNA by intercalating the anthracycline portion, chelating metal ions, generating free radicals or by inhibiting DNA topoisomerase II.

MAPK6, Rohu, and Rap1 showed cisplatin-specific effects. The platinum drug, cisplatin, is a cell-cycle non-specific anti-cancer drug, which binds to DNA and causes the production of intrastrand cross-links and DNA adduct formation. Cisplatin treatment activates multiple signal transduction pathways, which can lead to several cellular responses, including cell cycle arrest, DNA repair, survival or apoptosis. Moreover, genotoxic stress induces multiple signal transduction pathways, which include the MAP kinase pathways [8,13,20,28]. The same pathways are also related to platinum drug resistance [25]. Rohu mRNA for rhodanese (mercaptopyruvate sulfurtransferase) catalyzes the transfer of a sulfur ion to cyanide during cyanide degradation. Platinum cyanide binds at the entrance of the active site pocket, involving Arg-186 and Lys-249 of rhodanese [21]. Although we were unable to find a report on the interaction between cisplatin and Rap1, the latter, a ras-related gene with transformation suppressor activity, is closely associated with the MAP kinase cascades [3,6,17,30,36].

Taxol (paclitaxel) binds to the tubulin heterodimer, hence preventing microtubules from disassembling and cells from dividing. A taxol-specific effect was found in ubiquitin protein ligase E3A, KIAA0406, and SLC1A4 (neutral amino acid transporter). An inhibitor or ubiquitin-dependent multicatalytic protease complex (proteasome) was shown to be cytotoxic to human myeloid leukemia cell lines, and pre-treating this inhibitor enhanced the cytotoxicities of taxol and cisplatin [31]. Parkin, a protein–ubiquitin E3 ligase, was found to be tightly bound to microtubules in

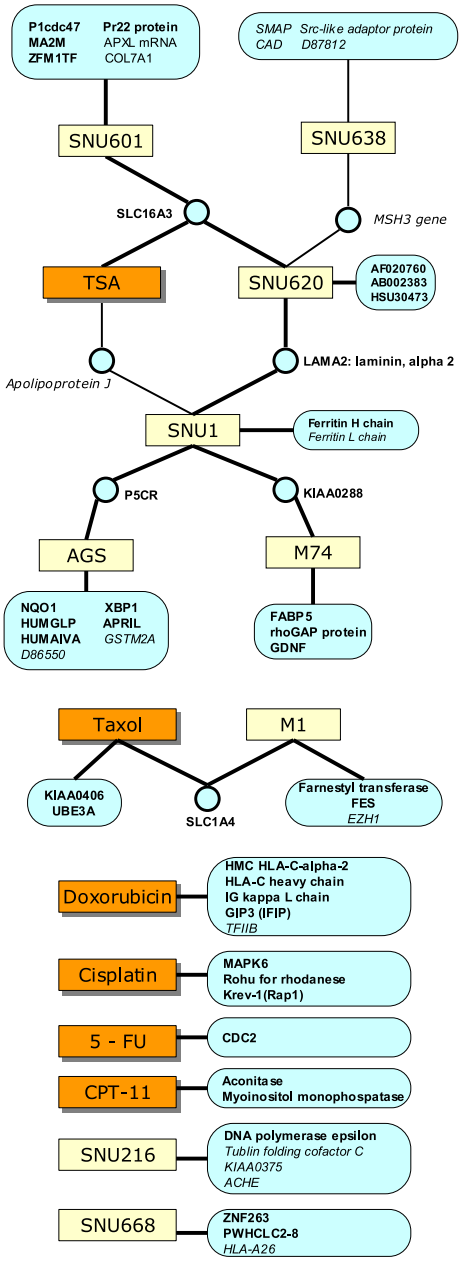

Fig. 5. Tripartite-graph representation of gene, drug, and cell-line association networks. Factor-specific generative patterns significantly associated to genes were determined and used for the reconstruction of the tripartite-graph networks.

taxol-mediated microtubule coassembly assays [27]. Although we could not find a report on the interaction between SLC1A4 and taxol, multidrug transporters have been suggested to be involved in chemotherapeutic-drug resistance [10,12,29]. Moreover, TSA (Trichostatin A) has been related to another transporter protein, SLC16A3. Interestingly, both of these transporters showed multiple drug-and-cancer-specific effects (i.e., Taxol and M1 vs. TSA and SNU601 and SNU620). KIAA0406 is a gene of unknown function.

5-Fluorouracil (5-FU) produced only one gene-specific effect on Cdc2 kinase. Cdc2 kinase forms a complex with B-type cyclins, which are central regulators of the progression from G2 to mitosis. Moreover, cyclin B levels following the treatment of a HepG2 hepatic cancer cell line with 5-FU or methotrexate were shown to be down regulated, and this cyclin B down-regulation was suggested as a means of regulating G2 arrest [18].

4. Discussion

We propose a method for identifying the optimal combination of biological factor(s) explaining the expression profile of the differentially expressed genes in a microarray experiment. The method outperforms when compared to typical multifactorial analysis such as ANOVA in two ways: reducing the higher-order interactions from a multiplicative to an additive complexity and, more importantly, no replicates are required.

Although sufficient replication is desirable whenever possible, the replication of microarray experiments may often be prohibitively costly. A balance must be found between the increased number of replicates and a reduction in the number of factors to be evaluated. Moreover, with increasing standardization [4,32] and growth of public repositories of expression data [5,9], it becomes important that powerful methods are developed to identify useful patterns from among the huge collections in heterogeneous expression databases.

In this study, genes specifically associated with drugs and/or cancers were successfully identified. We assumed that each observed expression profile was created by the underlying FSGP and noise, and devised a GA to determine the nearest FSGP to each expression profile and then measured the pattern distance of the profile with respect to the nearest FSGP. Finally, a statistical significance score was assigned using a permutation test.

One challenge presented by experiment with cancer cell-lines is that, because of cell-line heterogeneity, the in vitro effects observed do not accurately reflect the in vivo or clinical condition. The result of the proposed method, when it suggests a drug-specific effect, implies that the identified gene may consistently interact with the drug across many different cell lines. Thus, the technique suggests a more robust generalizability with respect to the cell-line heterogeneity such that it is more likely that the identified gene may also interact with real human cancer cells in vivo.

One property of the proposed method is that the nearest expression profile pattern provides a straightforward biological interpretation. High-biological interpretability becomes even more important when many factors and/or analysis levels are involved, when the

interpretation of higher-order interactions of factors can be extremely complicated. The proposed method can be viewed as a feature selection method augmented by statistical significance scoring. Extracted features may also be easily added to the analysis. Here we applied (a kind of) Hamming distance as a measure of pattern distance; however, one may flexibly choose an alternative appropriate distance metric.

In the present study, for the purpose of illustration, we applied a fairly standard GA with fixed chromosomal representation to explore the relatively small 2^{6+9} FSGP search space. The actual implementation or the performance of the GA itself is not the main issue in the present study. Rather, one can apply the proposed scheme of the present study of defining a pattern distance, devising a search algorithm to find the nearest FSGP, and evaluating the statistical significance. Other meta-heuristic methods like simulated annealing and Tabu search may also be successfully applied.

Association networks among genes, drugs, and cell lines were reconstructed from the identified FSGPs and showed significant associations with the observed gene expression profiles. The tripartite graphs in Fig. 5 can completely capture relevant multifactorial experimental information. In general, using the proposed method, the association networks of N factorial experiments can be represented as $(N + 1)$-partite graphs.

In contrast to traditional statistical test like ANOVA, which informs us of the statistical significance of a given null hypothesis, the nearest generative pattern of an expression profile informs us of the most likely underlying FSGP of the profile. However, the proposed method does not test all possible hypotheses. Rather it directly finds the best explanatory model and tests the statistical significance of the model using a permutation test. Therefore, it should be noted that our result does not necessarily exclude the statistical significances of remaining hypotheses. For example, the second or the third nearest generative pattern of an expression profile may also be statistically significant given the same FDR threshold. It is trivial to extend the method to test the significances of all competing hypotheses. When we search only for the nearest generative pattern, the proper interpretation of the result is that the expression profile of gene x in an experimental setting y can be 'best' explained by (the list of) factor(s), z_n.

A microarray experiment can be regarded as a data-driven method of massive hypothesis generation. For each hypothesis generated, whether positive or negative, we tried to find related publications. Although there are more than one million publications in PubMed, a microarray experiment with a multifactorial design generally addresses such a huge problem space (i.e., 157,286, $400 = 2^{6+9-1}\,(\text{factors}) \times 2400\,(\text{genes})$ in the present study) that it is literally impossible to find at least one report corresponding to the associations tested. It seemed that pre-genomic studies have explored the problem space only sparsely.

Acknowledgment

This study was supported by a grant from Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (03-PJ1-PG3-21000-0009).

References

[1] Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t test and statistical inferences of gene changes. Bioinformatics 2001;17(6):509–19.
[2] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 1995;57:289–300.
[3] Burney TL, Rockove S, Eiseman JL, Jacobs SC, Kyprianou N. Partial growth suppression of human prostate cancer cells by the Krev-1 suppressor gene. Prostate 1994;25(4):77–88.
[4] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 2001;29:365–71.
[5] Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, et al. ArrayExpress-a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2003;31:68–71.
[6] Caamano J, DiRado M, Iizasa T, Momiki S, Fernandes E, Ashendel C, et al. Partial suppression of tumorigenicity in a human lung cancer cell line transfected with Krev-1. Mol Carcinog 1992;6(4):252–9.
[7] Chen Y, Dougherty ER, Bittner ML. Ratio-based deisions and quantitative analysis of cDNA icroarray images. J Biomed Opt 1997;2:364–74.
[8] Eastman A. Activation of programmed cell death by anticancer agents: cisplatin as a model system. Cancer Cells 1990;2:275–80.
[9] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30(1):207–10.
[10] Endicott JA, Ling V. The biochemistry of P-glycoprotein-mediated multidrug resistance. Annu Rev Biochem 1989;58:137–71.
[11] Goldberger DE. Genetic algorithms in search, optimization and machine learning. Reading, MA: Addison-Wesley; 1989.
[12] Gottesman MM, Pastan I. Biochemistry of multidrug resistance mediated by the multidrug transporter. Annu Rev Biochem 1993;62:385–427.
[13] Hershberger PA, McGuire TF, Yu WD, Zuhowski EG, Schellens JH, Egorin MJ, et al. Cisplatin potentiates 1,25-dihydroxyvitamin D3-induced apoptosis in association with increased mitogen-activated protein kinase kinase kinase 1 (MEKK-1) expression. Mol Cancer Ther 2002;1(10):821–9.
[14] Holland JH. Adaptation in neural and artificial systems. Ann Arbor, MI: University of Michigan Press; 1975.
[15] Huber W, Von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 2002;18(Suppl. 1):S96–104.
[16] Kerr MK, Afshari CA, Bennett L, Bushel P, Martinez J, Walker N, Churchill GA. Statistical analysis of a gene expression microarray experiment with replication. Statist Sin 2002;12:203–17.

[17] Kitayama H, Sugimoto Y, Matsuzaki T, Ikawa Y, Noda M. A ras-related gene with transformation suppressor activity. Cell 1989;56(1):77–84.
[18] Krause K, Wasner M, Reinhard W, Haugwitz U, Dohna CL, Mossner J, et al. The tumour suppressor protein p53 can repress transcription of cyclin B. Nucleic Acids Res 2000;28(22):4410–8.
[19] Lee MT, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA 2000;97:9834–9.
[20] Levresse V, Marek L, Blumberg D, Heasley LE. Regulation of platinum-compound cytotoxicity by the c-Jun N-terminal kinase and c-Jun signaling pathway in small-cell lung cancer cells. Mol Pharmacol 2002;62(3):689–97.
[21] Lijk LJ, Kalk KH, Brandenburg NP, Hol WG. Binding of metal cyanide complexes to bovine liver rhodanese in the crystalline state. Biochemistry 1983;22(12):2952–7.
[22] Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol 2001;8:37–52.
[23] Okamoto E, Kinne RK, Sokeland J. Interferons modify in vitro proliferation of human bladder transitional cell carcinoma in the presence of doxorubicin and mitomycin C. J Urol 1996; 156(4):1492–5.
[24] Pavlidis P, Noble WS. Analysis of strain and regional variation in gene expression in mouse brain. Genome Biol 2001;2(10):RESEARCH0042.1–.15.
[25] Persons DL, Yazlovitskaya EM, Cui W, Pelling JC. Cisplatin-induced activation of mitogen-activated proteinkinases in ovarian carcinoma cells: inhibition of extracellular signal-regulated kinase activity increases sensitivity to cisplatin. Clin Cancer Res 1999;5(5):1007–14.
[26] Pedhazur EJ. Multiple regression in behavioral research: explanation and prediction. New York NY: Wadsworth; 1997.
[27] Ren Y, Zhao J, Feng J. Parkin binds to alpha/beta tubulin and increases their ubiquitination and degradation. J Neurosci 2003;23(8):3316–24.
[28] Sanchez-Perez I, Murguia JR, Perona R. Cisplatin induces a persistent activation of JNK that is related to cell death. Oncogene 1998;16(4):533–40.
[29] Sarkadi B, Muller M, Homolya L, Hollo Z, Seprodi J, Germann UA, et al. Interaction of bioactive hydrophobic peptides with the human multidrug transporter. FASEB J 1994;8(10):766–70.
[30] Sawada Y, Nakamura K, Doi K, Takeda K, Tobiume K, Saitoh M, et al. Rap1 is involved in cell stretching modulation of p38 but notERK or JNK MAP kinase. J Cell Sci 2001;114(Pt 6) :1221–7.
[31] Soligo D, Servida F, Delia D, Fontanella E, Lamorte G, Caneva L, et al. The apoptogenic response of human myeloid leukaemia cell lines and of normal and malignant haematopoietic progenitor cells to the proteasome inhibitor PSI. Br J Haematol 2001; 113(1):126–35.
[32] Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, et al. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 2002;3(9): research0046.1–.9.
[33] Storey JD, Tibshirani R. 2001. Estimating false discovery rates under dependence, with applications to DNA microarrays. Technical Report 2001-28. Department of Statistics Stanford University.
[34] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98:5116–21.
[35] Yang YH, Dudoit S, Luu P, Speed TP. 2000. Normalization for cDNA microarray data. Technical Report.
[36] York RD, Yao H, Dillon T, Ellig CL, Eckert SP, McCleskey EW, et al. Rap1 mediates sustained MAP kinase activation induced by nerve growth factor. Nature 1998;3926676: 622–626.