BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 24 2005, pages 4348–4355
doi:10.1093/bioinformatics/bti722

Gene expression

Differential coexpression analysis using microarray data and its application to human cancer

Jung Kyoon Choi[1], Ungsik Yu[1], Ook Joon Yoo[2] and Sangsoo Kim[3,*]

[1]National Genome Information Center, Korea Research Institute of Bioscience and Biotechnology, 52 Ueun-dong, Yuseong-gu, Daejeon, Korea, [2]Department of Biological Sciences, Korea Advanced Institute of Science and Technology, 373-1 Guseong-dong, Yuseong-gu, Daejeon, Korea and [3]Department of Bioinformatics and Life Science, Soongsil University, Seoul, Korea

Received on May 19, 2005; revised on September 2, 2005; accepted on October 16, 2005
Advance Access publication October 18, 2005

ABSTRACT

Motivation: Microarrays have been used to identify differential expression of individual genes or cluster genes that are coexpressed over various conditions. However, alteration in coexpression relationships has not been studied. Here we introduce a model for finding differential coexpression from microarrays and test its biological validity with respect to cancer.

Results: We collected 10 published gene expression datasets from cancers of 13 different tissues and constructed 2 distinct coexpression networks: a tumor network and normal network. Comparison of the two networks showed that cancer affected many coexpression relationships. Functional changes such as alteration in energy metabolism, promotion of cell growth and enhanced immune activity were accompanied with coexpression changes. Coregulation of collagen genes that may control invasion and metastatic spread of tumor cells was also found. Cluster analysis in the tumor network identified groups of highly interconnected genes related to ribosomal protein synthesis, the cell cycle and antigen presentation. Metallothionein expression was also found to be clustered, which may play a role in apoptosis control in tumor cells. Our results show that this model would serve as a novel method for analyzing microarrays beyond the specific implications for cancer.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: sskimb@ssu.ac.kr

1 INTRODUCTION

Gene expression profiling has been widely used for cancer research. Coupled with statistical techniques, gene expression patterns have been explored in many types of cancer. Most microarray analyses on cancer have focused on the comparison of tumor and normal tissues.

The identification of over- or underexpressed genes is one of the most widely used types of analysis in screens for potential tumor markers. Despite its great contribution to the field, however, differential expression analysis does not harness the full potential of microarrays in that information from only chosen genes rather than the entire set of transcripts is used. Moreover, genes are treated individually and interaction among them is not considered.

Other concerns with typical differential expression analysis include technical artifacts and limited applicability of results beyond the tissue studied. Individual genes are often studied in various cancer cells to draw a general conclusion about their behavior in more than one type of cancer. However, only a few studies have attempted such an approach on a genomic scale (Ramaswamy *et al*., 2001; Rhodes *et al*., 2004; Segal *et al*., 2004). As for technical artifacts, we have demonstrated that single microarray data are prone to false findings and the cross validation of datasets would significantly reduce those false findings and increase sensitivity (Choi *et al*., 2003, 2004).

With recent interest in biological networks, a gene coexpression network has emerged as a novel holistic approach for microarray analysis (Stuart *et al*., 2003; Bergmann *et al*., 2004; Lee *et al*., 2004; van Noort *et al*., 2004). van Noort *et al*. (2004) demonstrated the small-world and scale-free architecture of the yeast coexpression network. A human network was analyzed by Lee *et al*. (2004) with functional grouping and cluster analysis. Stuart *et al*. (2003) and Bergmann *et al*. (2004) separately constructed the gene coexpression network that connected genes whose expression profiles were similar across different organisms. They showed that functionally related genes are frequently coexpressed across organisms constituting conserved transcription modules. As for cancer study, Graeber and Eisenberg (2001) studied coexpression patterns of ligand–receptor pairs. They reported that some ligand–receptor pairs comprising an autocrine signaling loop had correlated mRNA expression in cancer, possibly contributing to cancer phenotypes. However, global coexpression patterns have not been determined for cancer.

To find common threads in different cancers while overcoming the inevitable artifacts of single-set analysis, we combined independent datasets on different types of cancer using the established meta-analytic methods (Choi *et al*., 2003). Most importantly, we wanted to explore transcriptional changes in terms of gene interactions rather than at the level of individual genes. To this end, we introduced a gene coexpression network and sought to

[*]To whom correspondence should be addressed.

Table 1. Datasets included in the analysis

Tissue origin	Author	Microarray platform	No. of normal samples	No. of tumor samples
Breast	Sørlie *et al.* (2001)	cDNA	13	13 (72)[a]
Colon	Notterman *et al.* (2001)	Hu6800[b]	22	22
Kidney	Boer *et al.* (2001)	membrane	81	81
Liver	Chen *et al.* (2002)	cDNA	76	76 (104)[a]
Lung	Bhattacharjee *et al.* (2001)	U95A[b]	17	17 (127)[a]
Lymphoma	Alizadeh *et al.* (2000)	cDNA	31	31 (77)[a]
Pancreas	Iacobuzio-Donahue *et al.* (2003)	cDNA	14	22
Prostate	Singh *et al.* (2002)	U95A[b]	50	52
Stomach	Chen *et al.* (2003)	cDNA	29	29 (103)[a]
Brain			8	20
+ Bladder	Ramaswamy *et al.* (2001)	Hu6800 +	7	11
+ Ovary		Hu35KSubA[b]	3	11
+ Uterus			6	10

All the cDNA microarray data were obtained from the Stanford Microarray Database (http://genome-www5.stanford.edu).

[a] Randomly selected from the total number in parentheses.

[b] GeneChip (Affymetrix, Santa Clara, CA).

find cancer-induced changes in the network. The identification of coexpressed pairs in tumor and normal tissues led to the construction of two distinct networks that represent tumor and normal states, respectively. We expected that biological changes would be reflected in transcriptional changes, which could be identified by comparing the two coexpression networks.

2 METHODS

2.1 Data collection, preprocessing and cross-platform gene mapping

To include tumors of diverse types in the analysis, we collected as many appropriate datasets as possible. Table 1 shows 10 collected datasets designed for the comparison of primary tumor and normal counterpart. The datasets were downloaded from the Stanford Microarray Database (http://genome-www5.stanford.edu) or the authors' websites. For the five datasets where the tumor sample size was much larger than the normal sample size (breast, liver, lung, lymphoma, stomach), we randomly selected the same number of tumor samples. Negative values from Affymetrix GeneChips were considered missing. Genes with >70 missing values or <4 observations were filtered out from each dataset. All of the expression values were base-two log-transformed. The integration of data generated on distinct microarray platforms required cross-platform gene mapping. Each clone was mapped to a UniGene accession based on UniGene build #162. For multiple clones matched with the same UniGene accession, we chose the one with the least missing values. Missing data were imputed using the mean of the gene vector. All the processed data can be found at http://centi.kribb.re.kr. Finally, we selected genes observed in at least five datasets to cover various cancer types and to increase reliability by means of interstudy validation.

2.2 Measuring correlation and differential expression for meta-analysis

Effect size, a standardized index measuring treatment or covariate effect, was employed for the meta-analysis (Choi *et al.*, 2003). Let μ be the average effect size and y_k be the observed effect size for independent datasets $k = 1, 2, \ldots, p$. The general model is given hierarchically as

$$\begin{aligned} y_k &= \theta_k + \varepsilon_k, \quad \varepsilon_k \sim N(0, s_k^2),\\ \theta_k &= \mu + \delta_k, \quad \delta_k \sim N(0, \tau^2), \end{aligned}$$

where between-study variance τ^2 represents the variability between datasets while within-study variance s_k^2 represents the sampling error conditioned on the i-th dataset. For each dataset k, the Pearson correlation r was calculated and converted into a standard normal metric using Fisher's r-to-z transformation as $z_r = 0.5 \log((1 + r)/(1 - r))$. z_r served as the observed effect size y_k while s_k^2 was given as $(n - 3)^{-1}$, where n is the sample size. For differential expression measure, the observed effect size y_k was given as $d = (\bar{X}_t - \bar{X}_n)/S_p$, where $\bar{X}_t$ and $\bar{X}_n$ represent the means of the tumor and normal groups, respectively and S_p indicates the pooled standard deviation. The sampling error s_k^2 was given as $\left(n_t^{-1} + n_n^{-1}\right) + y_k^2(2(n_t + n_n))^{-1}$, where n_t and n_n indicate the sample size in tumor and normal, respectively. The average effect size and its variance were estimated as

$$\mu = \frac{\sum \left(s_k^2 + \tau^2\right)^{-1} y_k}{\sum \left(s_k^2 + \tau^2\right)^{-1}} \quad \text{and} \quad \mathrm{Var}[\mu] = \frac{1}{\sum \left(s_k^2 + \tau^2\right)^{-1}}.$$

The estimation of τ^2 was performed as detailed by Choi *et al.* (2003). Finally, a z-score was computed as the ratio of μ over its standard error, representing the statistical significance of the expression correlation or differential

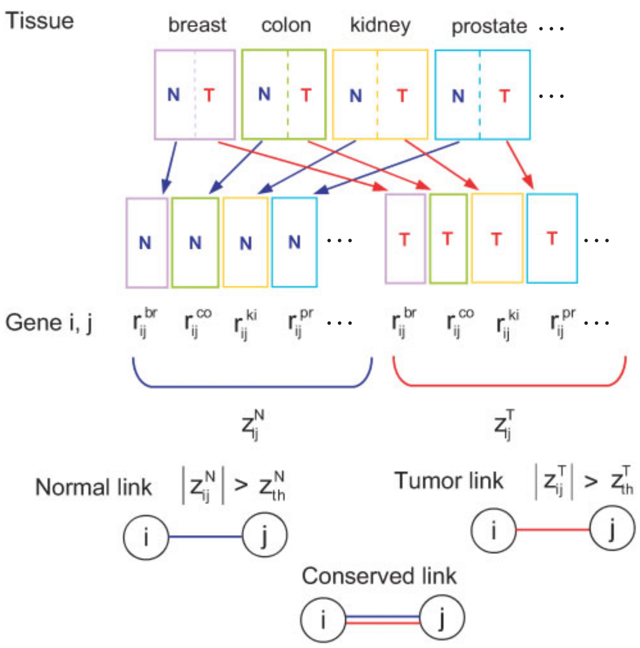
Tissue
breast
colon
kidney
prostate ···
N T
N T
N T
N T ···
N N N N ···
T T T T ···
Gene i, j
r_{ij}^{br} r_{ij}^{co} r_{ij}^{ki} r_{ij}^{pr} ···
r_{ij}^{br} r_{ij}^{co} r_{ij}^{ki} r_{ij}^{pr} ···
z_{ij}^{N}
z_{ij}^{T}
Normal link $|z_{ij}^{N}| > z_{th}^{N}$
Tumor link $|z_{ij}^{T}| > z_{th}^{T}$
i
j
i
j
Conserved link
i
j

Fig. 1. Schematic illustration of the analysis. Each dataset was split into a normal and tumor subdataset. The Pearson correlation was calculated for every pair of two genes within each subdataset. For any gene i and j, the normal correlations were combined across normal tissues to produce a normal z-score, z_{ij}^{N}, and the tumor correlations were combined across tumor tissues to produce a tumor z-score, z_{ij}^{T}. For a given threshold z_{th}^{N}, if $|z_{ij}^{N}| > z_{th}^{N}$, then genes i and j are connected by a normal link. For a given threshold z_{th}^{T}, if $|z_{ij}^{T}| > z_{th}^{T}$, then genes i and j are connected by a tumor link. If the genes i and j are connected by a normal and a tumor link simultaneously, in other words, the two genes correlate significantly both in tumor and normal tissues, they are connected by a conserved link. Therefore, a conserved link can simultaneously be a normal link and a tumor link. If a normal link is not a conserved link, it can be called a normal-specific link. Likewise, a tumor link which is not a conserved link can be called a tumor-specific link.

expression across multiple experiments. Consequently, the z-score harbors variation caused by biological or technical differences between datasets.

2.3 Construction of gene coexpression networks

We constructed a normal network and a tumor network according to the schematic illustrated in Figure 1. Each dataset was split into a normal and a tumor subdataset. For the pair of genes i and j, the Pearson correlation coefficient was computed in each subdataset and converted to the effect size as described above. The average effect-size scores for normal samples (z_{ij}^{N}) and that for tumor samples (z_{ij}^{T}) were calculated according to the above effect-size model. Thresholds had to be determined for the selection of significantly coexpressed pairs. To keep the same size for a normal and a tumor network, we sorted the gene pairs by $|z^{N}|$ and $|z^{T}|$, and selected the top 0.5% of the pairs. Consequently, the total number of the normal links was equal to that of the tumor links ($\#NL_{total} = \#TL_{total}$). Accordingly, z_{th}^{N} and z_{th}^{T} were set as the minimums of $|z^{N}|$ and $|z^{T}|$ among the selected pairs. A link was termed a conserved link if $|z_{ij}^{N}| > z_{th}^{N}$ and simultaneously $|z_{ij}^{N}| > z_{th}^{T}$.

2.4 Translation of coexpression interactions into functional interactions

Each gene was mapped to relevant Gene Ontology (GO) terms of levels 4–8. GO terms with at least 20 associated genes were used. Link g1–g2 was translated to category pair c1–c2 if gene g1 was mapped to category c1 and gene g2 to category c2 (Fig. 3). The number of the normal links translated to category pair 1–2 ($\#NL_{1-2}$) and that of the tumor links ($\#TL_{1-2}$) were compared with the number of possible unique links between genes in category 1 and genes in category 2 ($\#PL_{1-2}$). A normal coexpression score (NCS) and tumor coexpression score (TCS) were defined as $NCS = (\#NL_{1-2}/\#PL_{1-2})/(\#NL_{total}/\#PL_{total})$ and $TCS = (\#TL_{1-2}/\#PL_{1-2})/(\#TL_{total}/\#PL_{total})$ where $\#NL_{total}$ ($\#TL_{total}$) indicates that the total number of the normal (tumor) links and $\#PL_{total}$ means the number of possible links between all the genes used for network construction. Change in functional interaction c1–c2 was measured as change in coexpression interactions, namely NCS/TCS or TCS/NCS.

3 RESULTS AND DISCUSSION

3.1 Construction of coexpression networks and significance verification

We used 10 tumor–normal datasets spanning 13 tissue origins. A total of 8542 genes were contained in at least 5 out of the 10 datasets and used to construct a normal network and a tumor network according to the scheme illustrated in Figure 1. We selected the top 0.5% of the gene pairs, yielding 182 391 links in each of the two networks. This corresponds to highly significant z statistics ($z_{th}^{N} = 9.9$ and $z_{th}^{T} = 9.4$) and a false discovery rate (FDR) of <0.1% in both networks (Supplementary Table 1). About 27% (48 877 links) were conserved between the two networks. The rank correlation between the correlations of the two networks was 0.59. When we created pseudo-normal and pseudo-tumor networks by mixing normal and tumor samples, the median conservation rate was 40% and the median rank correlation was 0.83. This indicates tumor–normal difference in coexpression relationships. The whole networks can be obtained at http://centi.kribb.re.kr

We verified the significance of the links in the networks by means of two types of statistical tests. First, the gene vectors were randomly permuted within each subdataset to generate non-biological data. A random normal network and tumor network were constructed from the permuted data by taking the same number of top-ranking gene pairs. The random networks were compared with the real networks in terms of connectivity distribution in Figure 2. We found that the distribution of links in the real networks was highly non-random with a significant enrichment of highly connected genes as compared with the random networks. The real networks contained a giant component (composed of 4916 and 6552 nodes in the normal and tumor networks, respectively) plus orphans whose sizes did not exceed four nodes.

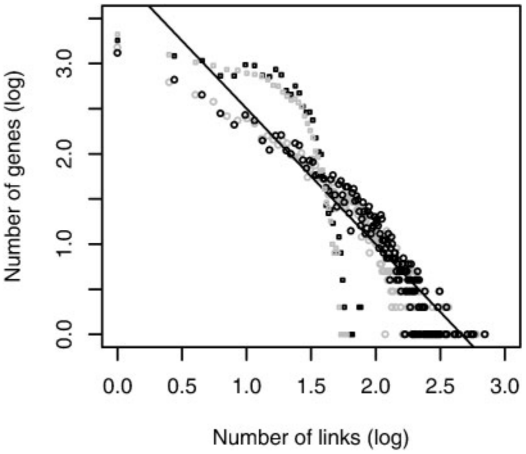

Fig. 2. Scale-free property of connectivity distribution in the coexpression networks. Shown is the number of links (x-axis) versus the number of genes that have the corresponding number of links (y-axis) in the normal network (black circles) and in the tumor network (gray circles). A non-biological normal network (black rectangles) and a tumor network (gray rectangles) constructed from randomized data are compared with the real networks. The numbers are shown on a $\log_{10}$ scale. The black line depicts the least-squares fit of the data to a linear line.

Second, we wanted to estimate the functional relevance of the links. To this end, we first translated coexpression interactions into functional interactions by mapping gene pairs to function category pairs (Fig. 3). For example, gene pair g1–g2 was mapped to category pair c1–c2 when gene g1 belonged to category c1 and gene g2 to category c2. GO annotations were employed to define functional categories (Ashburner *et al*., 2000). The degree of coexpression for a category pair was measured as an NCS and a TCS as shown in Figure 3. A high NCS (or TCS) for a category pair indicates that there are many coexpressed gene pairs mapped to the given category pair in normal (or in tumor). Therefore, if the identified links are biologically meaningful, the score should be high when two categories are functionally related. As a test of our model, we compared the scores of two same-class categories, two different-class categories and two random categories. The pair of identical terms or a parent and a child term in the GO hierarchy was considered same-class (e.g. cell cycle–cell cycle or cell cycle–mitotic cell cycle). A random category was created by random grouping of unrelated genes. The same number of genes as in the original mapping was randomly assigned to each GO term. Figure 4 shows that the degree of coexpression between two real groups is much higher than that between two random groups. Moreover, as expected, same-class categories show higher coexpression scores than different-class categories. These findings confirm the functional relevance of the coexpression links as defined in this study. Therefore, a high NCS (or TCS) may be indicative of strong functional interaction in normal (or tumor) tissues.

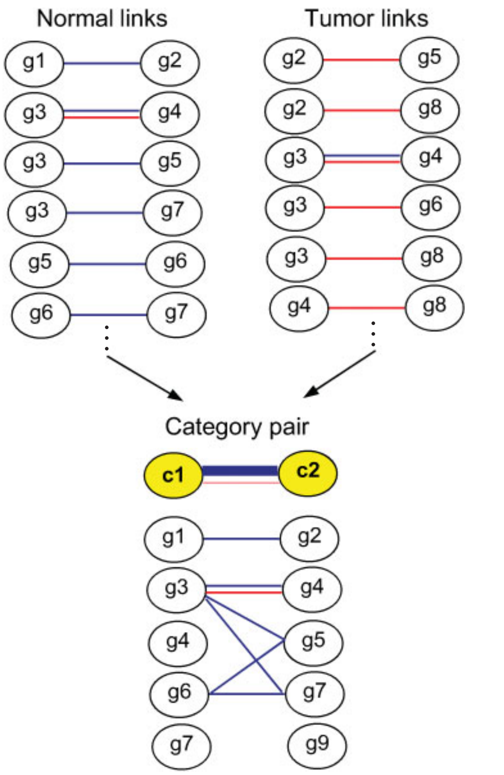

Fig. 3. Coexpression interactions and functional interactions. Each coexpression link was mapped to a category pair. In this illustration, genes g1, g3, g4, g6 and g7 were mapped to category c1 while g2, g4, g5, g7 and g9 to category c2 according to GO annotation. Six normal links and one tumor link were found among 23 possible links. The probability of finding a normal (or tumor) link is 0.005 when the top 0.5% of the gene pairs were used. Therefore, NCS will be (6/23)/0.005 while TCS (1.23)/0.005. Accordingly, a high NCS indicates that the normal links are observed more often than expected, suggesting strong functional interaction of c1 and c2. Change in the strength of functional interactions was measured as coexpression change, specifically, as NCS/TCS (in this case, 6/1). A high NCS/TCS may imply that functional association between c1 and c2 is reduced in tumor.

3.2 Functional changes accompanying coexpression changes

We expected that the strength of some functional interactions would be increased or decreased as cellular state changes. Change in functional interaction could be measured in terms of NCS and TCS. To confirm that the changes are not because of numerical difference but of biological difference, we created pseudo-normal and pseudo-tumor networks by mixing normal and tumor samples and obtained a pseudo-value of NCS, TCS, NCS/TCS and TCS/NCS for each category pair. The permutation was repeated 20 times. The maximum random (MR) values of the ratio of NCS and TCS are noted as MR(NCS/TCS) or MR(TCS/NCS). We used the criteria NCS $\geq$ 5 and TCS < 1 or TCS $\geq$ 5 and NCS < 1 to select 226 category pairs from the real networks. When compared with the median number of category pairs selected from the pseudo-networks, this number corresponds to an FDR of 3.45%.

We chose the category pairs with NCS $\geq$ 5 and TCS < 1 in order to find functional interactions that are maintained in normal cells but inactivated in cancer cells. As shown in Supplementary Table 2, MR(NCS/TCS) was smaller than NCS/TCS for all the chosen pairs, implying real biological difference behind the coexpression changes. Because several GO terms point to the same biological process, we selected out representative GO term pairs in Table 2. Interestingly, NADH dehydrogenase (NADH:ubiquinone oxidoreductase, complex I of the mitochondrial oxidative phosphorylation system) showed a loss of coexpression interactions with a variety of enzymes involved in energy metabolism. Defects in mitochondrial function have long been suspected to contribute to the development and progression of cancer (Warburg, 1930). In particular, it is well known that cancer reduces the ATP-producing function of oxidative phosphorylation and causes a compensatory increase in glycolytic ATP production. The related changes in the mitochondria are still attracting interest (Rossignol *et al*., 2004). Isidoro *et al*. (2004) reported that the alteration of the bioenergetic signature is a generalized condition of cancer. Taken together, the genes involved in the oxidative phosphorylation (especially the complex I genes) are likely to be misregulated in cancer, resulting in a loss of coexpression interactions (and functional association) with catabolism genes.

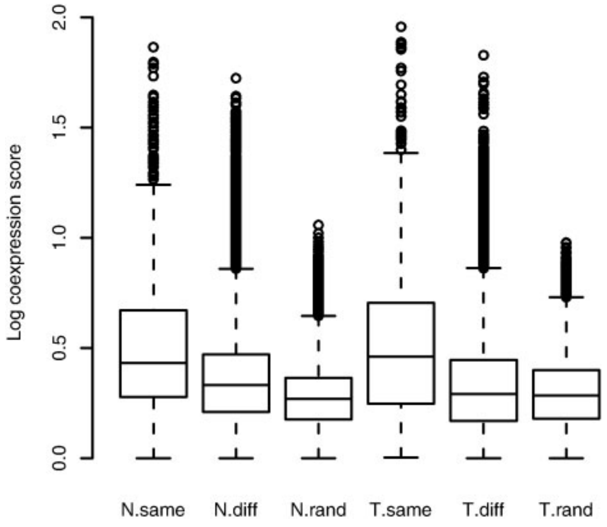

Fig. 4. Functional relevance of the coexpression links. We obtained an NCS and a TCS for each category pair. Three types of category pairs were considered: same-class, different-class and random pair. The pair of identical terms or a parent and a child term in the GO hierarchy was considered same-class (e.g. cell cycle–cell cycle or cell cycle–mitotic cell cycle). A random pair consists of two random categories created by mapping of randomly chosen genes. Shown is the distribution of the coexpression scores (*y*-axis, on a $\log_{10}$ scale) according to the types of category pairs (*x*-axis). Scores >1 are shown. The means are 4.564 ± 0.154, 2.617 ± 0.007, 2.100 ± 0.003, 6.026 ± 0.199, 2.760 ± 0.007, 2.092 ± 0.002, from left to right. N. same, NCS of same-class pairs; N. diff, NCS of different-class pairs; N. rand, NCS of random pairs; T. same, TCS of same-class pairs; T. diff, TCS of different-class pairs; T. rand, TCS of random pairs.

We next chose the category pairs with TCS $\geq$ 5 and NCS < 1 in order to find functional interactions that are created or enhanced in cancer. As shown in Supplementary Table 3, MR(TCS/NCS) was smaller than (TCS/NCS) for all the chosen pairs, implying real biological difference behind the coexpression changes. The categories related to the cell cycle predominated in the list. Table 3 shows selected representative GO term pairs. An increase of coexpression

Table 2. Functional interactions that are maintained in normal but inactivated in cancer

Category pair (group 1–group 2)	No. of genes in group 1	No. of genes in group 2	No. of normal links	No. of tumor links	NCS	TCS	NCS/TCS
Amino acid catabolism–NADH dehydrogenase (ubiquinone) activity	23	27	26	1	8.37	0.32	26.00
Hydrolase activity, acting on carbon–nitrogen (but not peptide) bonds–NADH dehydrogenase activity	33	25	23	2	5.58	0.48	11.50
Amine catabolism–NADH dehydrogenase (ubiquinone) activity	24	27	26	3	8.02	0.93	8.67
Lipid catabolism–NADH dehydrogenase (ubiquinone) activity	24	27	20	3	6.17	0.93	6.67
Exopeptidase activity–NADH dehydrogenase activity	40	25	26	5	5.20	1.00	5.20
Glycolysis–NADH dehydrogenase (ubiquinone) activity	24	27	42	11	12.96	3.40[a]	3.82

[a]Data shown for comparison although TCS is >1.

Table 3. Functional interactions that are created or enhanced in cancer

Category pair (group 1–group 2)	No. of genes in group 1	No. of genes in group 2	No. of normal links	No. of tumor links	NCS	TCS	TCS/NCS
Cell cycle checkpoint–cell cycle checkpoint	21	21	0	21	0.48	20.00	42.00
Regulation of mitosis–regulation of mitosis	23	23	1	28	0.79	22.13	28.00
Antigen presentation–cellular defense response	21	37	3	41	0.77	10.57	13.67
Complement activation–trypsin activity	20	27	0	21	0.20	8.20	42.00
Regulation of mitosis–microtubule organizing center	23	21	2	21	0.87	9.09	10.50
Collagen–collagen	21	21	1	27	0.95	25.71	27.00

interactions between cell cycle genes is biologically plausible; malignant cells rapidly proliferate. GO terms related to immune activity were also found. There is strong evidence that the immune system mounts a defense against tumors mediated by infiltrating lymphocytes. Activation of functional interaction between immune categories, especially with regard to antigen presentation and complement activation, is most likely to occur in these infiltrating immune cells.

Interestingly, collagens were found to be more tightly coregulated in cancer as compared with normal growth. Collagen is a major component of extracellular matrix and basement membranes, the degradation and penetration of which are known to be significant processes in tumor cell invasion. Also, collagen content and structure are key determinants of macromolecule transport in tumors (Brown *et al*., 2003). Some tumors are often associated with an intense production of interstitial collagens, known as desmoplastic reaction. The expression of collagen genes in various tumors was studied in association with tumor progression (Mahara *et al*., 1994; Kauppila *et al*., 1996; Fenhalls *et al*., 1999; Burns-cox *et al*., 2001; Fischer *et al*., 2001; Aoyagi *et al*., 2004). For example, Fenhalls *et al*. (1999) showed that stage I breast tumors had elevated levels of collagen mRNA compared with those of normal breast and conversely, in stage II and III tumors, levels of collagen mRNA were reduced. Fischer *et al*. (2001) reported that COL11A1 and COL5A2 were not expressed in normal colon but upregulated in colorectal cancer. More importantly, they found that the expression of the two genes exhibited high correlation in tumor. Collectively, tumor cells may organize the productions of collagens, possibly providing a mechanism to inhibit (in early stage) or facilitate (in late stage) invasion and metastatic spread.

To summarize, we successfully detected functional differences between normal growth and cancer in terms of gene coexpression changes in broad areas of physiology: energy metabolism, the cell cycle, immune activation and collagen production.

3.3 Extracting clusters from coexpression networks

As a second approach, the algorithm called MCODE was employed to extract clusters of highly interconnected genes (Bader and Hogue, 2003). MCODE was successfully applied to coexpression network analysis in the study by Lee *et al*. (2004). Although the network of top 0.5% yielded a high-level overview, it was difficult to study individual genes in the network because of its high density. To effectively study individual genes in the network, we increased the stringency so as to create a smaller network with higher level of statistical confidence; the top 0.01% of the gene pairs was chosen yielding 3648 links for each network. Conserved links were excluded to leave normal-specific or tumor-specific links. We identified three clusters with the highest scores in the tumor-specific network (Fig. 5A–C). No cluster was found in the normal-specific network with the same criterion.

From the first cluster shown in Figure 5A, we can speculate that cancer cells may increase the rate of protein synthesis. Most of the genes in the cluster encode ribosomal proteins (RPL∼, RPS∼, LAMR1, UBA52, TINP1). It also includes genes related to protein biosynthesis (EIF3S6, NACA), ribosome assembly and transport (NPM1), ribosome biogenesis (UBA52) and rRNA transcription (MKI67IP).

Unexpectedly, metallothionein (MT) proteins were found to be clustered in the tumor coexpression network (Fig. 5B). MT is

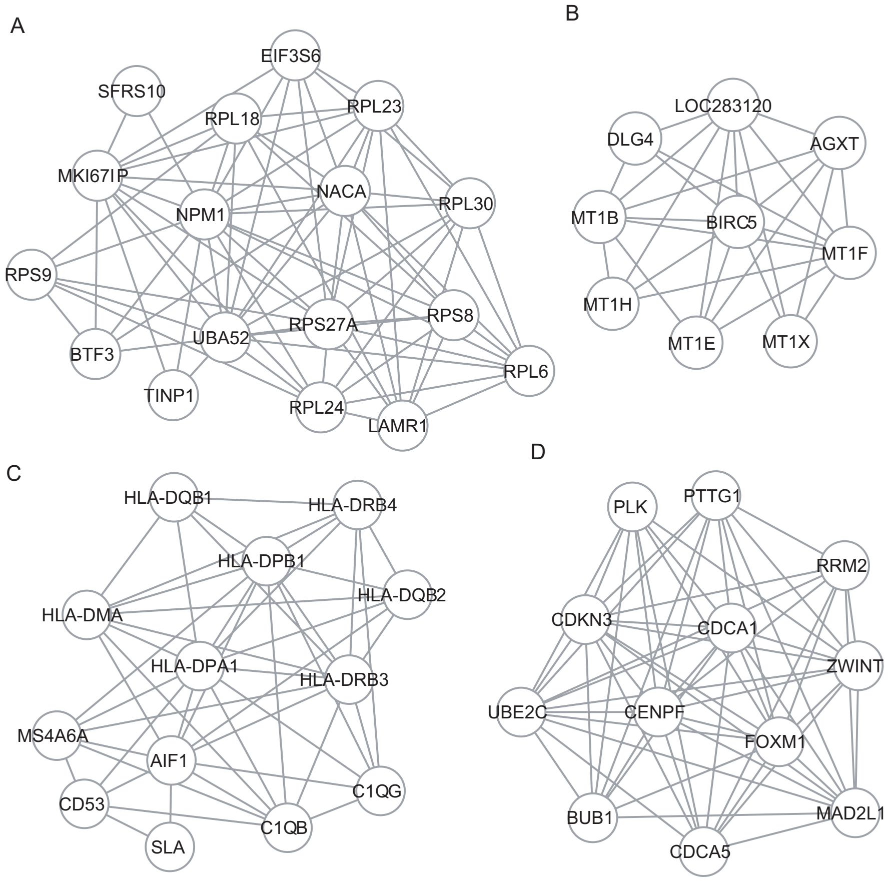

Fig. 5. Clusters extracted from the tumor coexpression network with MCODE. (**A**) Protein biosynthesis cluster. (**B**) Metallothionein cluster. (**C**) Immune cluster. (**D**) Cell cycle cluster.

involved in many physiological processes, such as metal homeostasis and detoxification, cell proliferation, apoptosis and protection against oxidative damage. Prognostic significance of MT expression has been studied for various human tumors (Nagel and Vallee, 1995; Jayasurya *et al*., 2000; Hishikawa *et al*., 2001; Huang and Yang, 2002; Miranda *et al*., 2002; Chun *et al*., 2004; Mitropoulos *et al*., 2005). It was reported that MT concentration oscillates during the progression of human colonic cancer cells through the cell cycle (Nagel and Vallee, 1995). Meanwhile, Jayasurya *et al*. (2000) observed an inverse relationship between the level of MT expression and the apoptotic index in nasopharyngeal carcinoma tissues, suggesting that overexpression of MT may protect tumor cells from entering into the apoptotic process and thereby contribute to tumor expansion. In this regard, it is intriguing that BIRC5, an inhibitor of caspase 3 and caspase 7, connects with all the genes in this complex. BIRC5 is well known to play a role in neoplasia by counteracting a default induction of apoptosis in G2/M phase. Another fully connected unknown gene LOC283120 may possibly play a role in apoptosis control as well. Also, the related roles of DLG4 and AGXT remain to be investigated.

Figure 5C shows an immune cluster, which consists mainly of MHC II class antigens (HLA$\sim$). MHC class II expression is restricted to antigen presenting cells, which are likely to be infiltrated into tumor tissues. Enhanced functional interaction of antigen presentation and defense response was already demonstrated in Table 3.

We reconstructed the networks with the top 0.02% of the gene pairs and identified an additional cluster whose genes are all related to the cell cycle. UBE2C (UbcH10), the fully connected gene in this complex, was in fact found to be the gene with the most links to cell cycle genes in the whole tumor network. This ubiquitin-conjugating enzyme (E2) has recently become a focus of interest. It was shown to be overexpressed in various human primary tumors and associated with the degree of tumor differentiation (Okamoto *et al*., 2003; Wagner *et al*., 2004). Also, its central role as a timer in the oscillations of the cell cycle was recently revealed by Rape and Kirschner (2004). Taken together, transcriptional regulation of UBE2C in concert with other cell cycle genes may significantly contribute to tumor progression.

3.4 Biological interpretation of coexpression changes

We found increased (or decreased) coexpression interactions to coincide with enhancement (or inactivation) of functional interactions. This raised a question as to whether coexpression increase and decrease are due to gene expression increase and decrease, respectively. However, we found that this is not the case. Specifically, in 97% of the specific links, neither of the two genes was significantly increased nor was significantly decreased with $|z| > 2.0$, where z indicates a differential expression score of a gene. Only 4.46% of the tumor-specific links had both nodes upregulated and 0.95% of the normal-specific links had both nodes downregulated.

Li (2002) and Li *et al*. (2004) proposed a model for the interpretation of coexpression changes. They sought to systematically analyze changes in coexpression patterns according to changes in cellular state. For a pair of genes (X,Y), the relevant cellular state was represented as differential expression of a third gene Z. Specifically, he sought to detect the association of an increase or decrease in Z with an increase or decrease in the correlation of (X,Y). This underlying principle could be applied to the interpretation of tumor-normal coexpression changes. An increase or decrease in the correlation of a gene pair may be associated with the up- or downregulation of other genes in the same functional category.

For another model, external parameters such as tumor grade should be considered. We can expect that expression of some cell cycle genes would increase as tumors progress. The expression of these genes would be coordinately controlled by signaling systems working for tumor development. Therefore, the gene expressions would correlate with tumor progression and consequently with each other. We already attempted to associate coregulation of collagens and that of metallothioneins with tumor stage and apoptotic index, respectively. It does not necessarily mean that the genes are always upregulated in tumor samples because they might be increased or decreased during tumor progression. Metabolic state could be another example. Mitochondrial genes involved in ATP synthesis should be regulated according to cellular energy state in normal condition. Cancer-induced mitochondrial defect may lead to the misregulation of these genes, which in turn results in a loss of coexpression interactions with genes involved in energy metabolism.

4 CONCLUSIONS

With an application of differential coexpression analysis, we were able to detect major cellular changes in tumor cells and find gene groups whose coregulation might contribute to malignant transformation. Importantly, these coexpression changes were not because of differential expressions. We expect that this approach would serve as a novel method for analyzing microarrays beyond the specific implications for cancer.

ACKNOWLEDGEMENTS

We thank Dr Seyeon Weon for valuable suggestions and Dr Paul Kretchmer at San Francisco Edit for his assistance in editing this manuscript. This work was supported by Grant No. M10407010001-04N0701-00110 and Grant No. FG-5-01 of the 21c Frontier Functional Human Genome Project from the Ministry of Science and Technology, Korea.

Conflict of Interest: none declared.

REFERENCES

Alizadeh,A.A. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Aoyagi,Y. *et al*. (2004) Overexpression of TGF-beta by infiltrated granulocytes correlates with the expression of collagen mRNA in pancreatic cancer. *Br. J. Cancer*, **91**, 1316–1326.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Bergmann,S. *et al*. (2004) Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*., **2**, E9.

Bhattacharjee,A. *et al*. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

Boer,J.M. *et al*. (2001) Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 35,000-element cDNA array. *Genome Res*., **11**, 1861–1870.

Brown,E. *et al*. (2003) Dynamic imaging of collagen and its modulation in tumors *in vivo* using second-harmonic generation. *Nat. Med*., **9**, 796–800.

Burns-cox,N. *et al*. (2001) Changes in collagen metabolism in prostate cancer: a host response that may alter progression. *J. Urol*., **166**, 1698–1701.

Chen,X. *et al*. (2002) Gene expression patterns in human liver cancers. *Mol. Biol. Cell*, **13**, 1929–1939.

Chen,X. *et al*. (2003) Variation in gene expression patterns in human gastric cancers. *Mol. Biol. Cell*, **14**, 3208–3215.

Choi,J.K. *et al*. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**(Suppl. 1), i84–i90.

Choi,J.K. *et al*. (2004) Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Lett*., **565**, 93–100.

Chun,J.H. *et al*. (2004) Increased expression of metallothionein is associated with irinotecan resistance in gastric cancer. *Cancer Res*., **64**, 4703–4706.

Fenhalls,G. *et al*. (1999) Breast tumour cell-induced down-regulation of type I collagen mRNA in fibroblasts. *Br. J. Cancer*, **81**, 1142–1149.

Fischer,H. *et al*. (2001) Colorectal carcinogenesis is associated with stromal expression of COL11A1 and COL5A2. *Carcinogenesis*, **22**, 875–878.

Graeber,T.G. and Eisenberg,D. (2001) Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nat. Genet*., **29**, 295–300.

Hishikawa,Y. *et al*. (2001) Expression of metallothionein in colorectal cancers and synchronous liver metastases. *Oncology*, **61**, 162–167.

Huang,G.-W. and Yang,L.-Y. (2002) Metallothionein expression in hepatocellular carcinoma. *World J. Gastroenterol*., **8**, 650–653.

Iacobuzio-Donahue,C.A. *et al*. (2003) Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays. *Am. J. Pathol*., **162**, 1151–1162.

Isidoro,A. (2004) Alteration of the bioenergetic phenotype of mitochondria is a hallmark of breast, gastric, lung and oesophageal cancer. *Biochem. J*., **378**, 17–20.

Jayasurya,A. *et al*. (2000) Correlation of metallothionein expression with apoptosis in nasopharyngeal carcinoma. *Br. J. Cancer*, **82**, 1198–1203.

Kauppila,S. *et al*. (1996) Expression of mRNAs for type I and type III procollagens in serous ovarian cystadenomas and cystadenocarcinomas. *Am. J. Pathol*., **148**, 539–548.

Lee,H.K. *et al*. (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res*., **14**, 1085–1094.

Li,K.C. (2002) Genome-wide coexpression dynamics: theory and application. *Proc. Natl Acad. Sci. USA*, **99**, 16875–16880.

Li,K.C. *et al*. (2004) A system for enhancing genome-wide coexpression dynamics study. *Proc. Natl Acad. Sci. USA*, **101**, 15561–15566.

Mahara,K. *et al*. (1994) Transforming growth factor beta 1 secreted from scirrhous gastric cancer cells is associated with excess collagen deposition in the tissue. *Br. J. Cancer*, **69**, 777–783.

Miranda,A. *et al*. (2002) Prognostic significance of metallothionein in human gastrointestinal cancer. *Clin. Cancer Res*., **8**, 1889–1896.

Mitropoulos,D. *et al*. (2005) Prognostic significance of metallothionein expression in renal cell carcinoma. *World J. Surg. Oncol*., **3**, 5.

Nagel,W.W. and Vallee,B.L. (1995) Cell cycle regulation of metallothionein in human colonic cancer cells. *Proc. Natl Acad. Sci. USA*, **92**, 579–583.

Notterman,D.A. *et al*. (2001) Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res*., **61**, 3124–3130.

Okamoto,Y. *et al*. (2003) UbcH10 is the cancer-related E2 ubiquitin-conjugating enzyme. *Cancer Res*., **63**, 4167–4173.

Ramaswamy,S. *et al*. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Rape,M. and Kirschner,M.W. (2004) Autonomous regulation of the anaphase-promoting complex couples mitosis to S-phase entry. *Nature*, **432**, 588–595.

Rhodes,D.R. *et al.* (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl Acad. Sci. USA*, **101**, 9309–9314.

Rossignol,R. *et al.* (2004) Energy substrate modulates mitochondrial structure and oxidative capacity in cancer cells. *Cancer Res.*, **64**, 985–993.

Segal,E. *et al.* (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet.*, **36**, 1090–1098.

Singh,D. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–239.

Sørlie,T. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869–10874.

Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

van Noort,V. *et al.* (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO rep.*, **5**, 280–284.

Wagner,K.W. *et al.* (2004) Overexpression, genomic amplification and therapeutic potential of inhibiting the UbcH10 ubiquitin conjugase in human carcinomas of diverse anatomic origin. *Oncogene*, **23**, 6621–6629.

Warburg,O. (1930) *Metabolism of tumors*. Arnold Constable, London.