Available online at www.sciencedirect.com
SCIENCE DIRECT
Genomics 87 (2006) 315 – 328
GENOMICS
www.elsevier.com/locate/ygeno

Comprehensive analysis of pathway or functionally related gene expression in the National Cancer Institute's anticancer screen

Ruili Huang [a], Anders Wallqvist [b], David G. Covell [a,*]

[a] *Laboratory of Computational Technologies, Developmental Therapeutics Program, Screening Technologies Branch, National Cancer Institute at Frederick, National Institutes of Health, Frederick, MD 21702, USA*

[b] *Science Applications International Corporation, National Cancer Institute at Frederick, National Institutes of Health, Frederick, MD 21702, USA*

Received 1 June 2005; accepted 19 November 2005
Available online 4 January 2006

Abstract

We have analyzed the level of gene coregulation, using gene expression patterns measured across the National Cancer Institute's 60 tumor cell panels (NCI_{60}), in the context of predefined pathways or functional categories annotated by KEGG (Kyoto Encyclopedia of Genes and Genomes), BioCarta, and GO (Gene Ontology). Statistical methods were used to evaluate the level of gene expression coherence (coordinated expression) by comparing intra- and interpathway gene–gene correlations. Our results show that gene expression in pathways, or groups of functionally related genes, has a significantly higher level of coherence than that of a randomly selected set of genes. Transcriptional-level gene regulation appears to be on a "need to be" basis, such that pathways comprising genes encoding closely interacting proteins and pathways responsible for vital cellular processes or processes that are related to growth or proliferation, specifically in cancer cells, such as those engaged in genetic information processing, cell cycle, energy metabolism, and nucleotide metabolism, tend to be more modular (lower degree of gene sharing) and to have genes significantly more coherently expressed than most signaling and regular metabolic pathways. Hierarchical clustering of pathways based on their differential gene expression in the NCI_{60} further revealed interesting interpathway communications or interactions indicative of a higher level of pathway regulation. The knowledge of the nature of gene expression regulation and biological pathways can be applied to understanding the mechanism by which small drug molecules interfere with biological systems.

Keywords: Pathway; Gene expression; Coregulation; Cancer

The genome encodes two major types of information: genes encoding proteins that execute biological functions and *cis*-control elements of transcription [1]. Proteins may function as monomers, as complexes, or within networks of interacting proteins, metabolites, and/or small molecules (pathways). The *cis*-control elements, together with transcription factors, form the linkages and architectures of regulatory networks controlling expression levels for genes that mediate physiological and developmental responses. Numerous efforts have been dedicated to the task of deconvoluting the function and regulation of biological networks, especially with the increasing availability of RNA expression microarrays, which provide large amounts of data for analysis of individual genes within predefined pathways or for elucidation of hidden gene regulation networks [2–6]. Deriving gene regulation networks or pathways, on the basis of expression data, is based on the general premise that coregulated genes function in the same pathway or, in other words, functionally related genes are coregulated or coexpressed. Verification of this premise and the extrapolation of these results toward the identification of causal factors underlying gene coregulation is currently a very active area of research.

Coexpression of genes has been observed using pathways annotated by KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/), based on gene expression data in colon and liver cancer cells and normal tissue samples [6] and in the *Arabidopsis* genome [7]. Neighboring genes, that is, genes that are immediately adjacent on chromosomes, have been found to be coexpressed in humans [8,9], *Drosophila* [10–12], yeast [10], *Caenorhabditis elegans* [13], and *Arabidopsis* [7]. Certain transcription factors, e.g., CTCF, serve to insulate adjacent genes from transcriptional

* Corresponding author. Fax: +1 301 846 6978.
E-mail address: covell@ncifcrf.gov (D.G. Covell).

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.11.011

activation. The requirement for coregulation of functionally related genes has been proposed as a possible cause for the observed coexpression of genes. In addition, the number of interactions between proteins has been implicated as an important predictor for the degree of coexpression between their corresponding genes, a result that has been offered as an explanation for particularly high degrees of coexpression in genes encoding proteins that are known to function in multicomponent complexes, which often contain a large number of protein–protein interactions [14,15]. In yeast, genes encoding interacting proteins tend to be coexpressed [14,16,17]. In contrast, the degree of coexpression for genes encoding enzymes in metabolic pathways has been found to be generally low [6,7], despite the observation of similarities in gene expression patterns for some metabolic pathways in embryonic and adult mouse tissues [18]. This finding based on clustering analysis is, however, mostly observational and not quantitative. Furthermore, functionally linked interacting proteins have been observed to share higher *cis*-similarity, defined as the proportion of shared transcription factor binding sites regulating the transcription of a gene, than enzymes catalyzing the conversion of adjacent substrates to products in a collection of metabolic pathways [19]. This finding has led to the hypothesis that genes encoding a set of interacting proteins will be transcribed using a common set of regulatory signals, whereas substrate concentration and enyzme–substrate interactions may exact regulatory control of metabolic pathways, as distinct from explicit transcriptional control [20].

The normal network structures of a system, however, may be perturbed in diseases through genetic mutations and/or by pathological environmental cues, such as infectious agents or chemical carcinogens. Cancer is believed to arise from multiple spontaneous and/or inherited mutations functioning in networks that control vital cellular events [21–23], which is partly reflected by genetic alterations in intracellular signaling pathways, which normally orchestrate the execution of developmental programs and the cellular response to extrinsic factors [24]. The evolving states of certain cancers are reflected in dynamically changing expression patterns of genes and proteins within the diseased cells [25]. In support of these observations, computational methods find that there are more pathways with coexpressed genes in cancer cells than in normal cells [6]. System-based analyses of cellular processes that integrate information on genes, proteins, and metabolites, combined with technological advances to monitor these changes, offer the potential for improvements in predictive medicine [1].

The goal of our present work is to propose novel strategies to examine at length gene expression regulation patterns within predefined pathways or functional groups. Our analysis focuses on the constitutive gene expression data measured across the National Cancer Institute's 60-tumor-cell screen (NCI_{60}), which reflects diverse cell lineages (lung, renal, colorectal, ovarian, breast, prostate, central nervous system, melanoma, and hematological malignancies). We will analyze the NCI_{60} gene expression patterns in terms of pathways annotated by KEGG and BioCarta and gene categories defined by GO (Gene Ontology). Comparison of the degree of gene expression coherence in these three annotation schemes will be used to postulate rationales that might determine the cohesiveness of a pathway or a group of functionally related genes. The gene expression data across the NCI_{60} will be organized into self-organizing maps (SOMs) [26], which segregate the data into nodes, i.e., clusters of genes sharing similar expression patterns. Meta-clustering of the SOM nodes will be used to generate a hierarchy of clades. Pathways will be organized according to similarity of gene occurrence patterns in SOM clades for each pathway. These results will be used to assess interactions between pathways either through gene sharing or through coexpression of genes, which may be indicative of a higher level of regulation and coherence within cellular processes. One of our future goals is to utilize the knowledge obtained about the nature of gene expression regulation and biological pathways to assess relationships between pathway gene expressions and drug responses derived from various cancer cell lines, with the aim of gaining a better understanding of mechanisms of drug action.

Results

Our measure of pathway cohesiveness is based on correlations of gene expression patterns across the NCI_{60} for genes within specific pathways. The cohesiveness of a pathway is represented by the coherence of its gene expressions as measured by their correlation strength. Genes in the same pathway are believed to be regulated in a more coordinated fashion than a random set of genes, thus expression patterns of these genes are expected to be more coherent. However, some pathways appear to be more cohesive than others, due to factors such as the degree of gene expression regulation within the pathway, modularity of the pathway, degree of gene sharing (cross talk) with other pathways, expression data availability, experimental errors/data quality, etc. In addition, even though computational methods have been widely used in biological data analysis and interpretation, a statistically significant finding does not always translate directly to biological relevance, due partly to the intricate nature of biological systems and the inherent complexity of data generated therein. However, applying rigorous statistical procedures to establish the significance level of the observations and exclude the possibility of certain events happening by random chance helps to improve our confidence for proposing statistically significant results as a foundation for testable hypothesis generation. We have followed this principle throughout the course of our investigations.

The gene expression cohesiveness of pathways defined by KEGG and BioCarta and annotation categories (terms) defined by GO can be evaluated by comparing the intra- versus interpathway gene expression correlations using the Kruskal–Wallis procedure. The Kruskal–Wallis H statistic (H score) is computed for each pathway, and the significance (p value) of each pathway H score is evaluated by a random gene permutation procedure. The randomization is run 1000 times; thus the smallest nonzero p value that can be obtained is 0.001

(see Data and methods for computation details). A pathway is considered significantly cohesive at the 95% confidence level if it has a significant ($p \leq 0.05$) and positive H score. The larger the H score, the more coherent the gene expressions are within a pathway, compared to the expressions of genes not linked by a known pathway. The number of pathways that can be studied in this fashion is naturally limited by the availability of gene expression data, and the results obtained will be affected consequently. In the present investigation, only pathways that have at least three gene expression data vectors available within our microarray set are included in the calculations. Summary statistics related to the three pathway systems under investigation are shown in Table 1.

Gene expression coherence in pathways annotated by KEGG

Of the 111 KEGG pathways, each of which has at least three genes with available expression data, 21 (19%) have significantly stronger intra- than interpathway gene–gene correlations. These pathways, sorted in descending order by their level of intrapathway gene expression cohesiveness, and their H scores are listed in Table 2. The ribosome pathway is the most cohesive pathway, with a p value of nearly 0. The ribosome pathway is composed of genes that encode various proteins of the ribosomal subunits. These proteins need to interact physically with each other to form a large protein complex, the ribosome, and are thereby closely related functionally. The second most cohesive KEGG pathway is oxidative phosphorylation, which is composed of genes that encode protein complexes (complexes I through V) that participate in the mitochondrial respiration chain. These include the NADH dehydrogenases, succinate dehydrogenases, cytochrome c oxidases, cytochrome c reductases, ATPases, and ATP synthases. In fact, in addition to ribosome and oxidative phosphorylation, most KEGG pathways that are composed of genes encoding parts of some large protein complex (labeled with superscript a in Table 2) also appear to be significantly cohesive, except for RNA polymerase and protein export. These include the proteasome, ATP synthesis, DNA polymerase, and basal transcription factors pathways, which are significantly enriched (Fisher's exact $p = 0.01$) within the set of cohesive pathways. The fact that the proteins in these pathways need to interact physically with each other in the cell to ensure the proper formation and function of the protein complexes appears to be reflected in their tightly coordinated gene expression.

The KEGG pathways can be further grouped into subcategories including metabolism, cellular processes, environmental information processing, genetic information processing, and human diseases, according to their cellular function (see Table 3). Among those categories, only one group, genetic information processing, which includes ribosome, proteasome, aminoacyl-tRNA biosynthesis, basal transcription factors, DNA polymerase, ubiquitin-mediated proteolysis, protein export, and RNA polymerase, shows a significant enrichment of cohesive pathways (Fisher's exact $p = 0.03$; see Table 3). Five of these eight pathways are cohesive and six of them are composed of genes that encode large protein complexes. In contrast, the groups metabolism and environmental information processing, which include all the signaling pathways, appear to have a less than average percentage of cohesive pathways (see Table 3), but neither is statistically significant at the 95% confidence level, compared to the occurrence of all cohesive pathways in KEGG (19%). The lack of cohesive pathways in the metabolic pathway group, however, is statistically significant at the 90% confidence level ($p = 0.07$).

Gene expression coherence in pathways deposited in BioCarta

In the 262 BioCarta functional pathways analyzed herein, 34 (13%) show significant intrapathway gene expression cohesiveness. These pathways, their cohesiveness H scores, and the significance levels (p values) of their H scores are listed in Table 4. A comparison between the genes annotated in the BioCarta pathway system and the genes in KEGG shows an overlap of 281 genes (20.5%). When categorized by the KEGG pathway system, the intersection appears to be significantly enriched (Fisher's exact $p \leq 0.05$) in genes that belong to, in order of descending significance level, MAPK signaling pathway, integrin-mediated cell adhesion, apoptosis, cell cycle, toll-like receptor signaling pathway, TGF-β signaling pathway, cytokine–cytokine receptor interaction, Jak–STAT signaling pathway, neurodegenerative disorders, and prion disease. A closer examination of the cohesive BioCarta pathway genes shows that they are significantly enriched (Fisher's exact $p \leq 0.05$), compared to all genes in the BioCarta–KEGG intersection, with genes that belong to two KEGG pathways, cell cycle and oxidative phosphorylation, which are also the third and second most cohesive pathways, respectively, in KEGG (see Table 2). This indicates a good agreement in regard to cohesive pathways obtained from these two pathway systems. This agreement is also

Table 1
Summary statistics for KEGG and BioCarta pathway systems and gene categories defined by GO

Annotation scheme	Pathways defined	Pathways with ≥3 gene data vectors	Maximum pathway gene count	Average pathway gene count	Cohesive pathway count ($p \leq 0.05$)	Cohesive pathways (%)
KEGG	134	111	123	19	21	19
BioCarta	314	262	63	10	34	13
	Terms defined	Terms with ≥3 gene data vectors	Maximum term gene count	Average term gene count	Cohesive term count ($p \leq 0.05$)	Cohesive terms (%)
GO	3564	787	945	20	171	22

Table 2
Statistically cohesive KEGG pathways ordered in decreasing level of significance (p value)

Pathway	Pathway title	H score	p
hsa03010	Ribosome[a]	4328.66	$<1 \times 10^{-3}$
hsa00190	Oxidative phosphorylation[a]	399.42	$<1 \times 10^{-3}$
hsa04110	Cell cycle	221.71	$<1 \times 10^{-3}$
hsa03050	Proteasome[a]	69.30	$<1 \times 10^{-3}$
hsa00193	ATP synthesis[a]	27.77	$<1 \times 10^{-3}$
hsa00230	Purine metabolism	26.47	$<1 \times 10^{-3}$
hsa00531	Glycosaminoglycan degradation	22.81	$<1 \times 10^{-3}$
hsa00900	Terpenoid biosynthesis	18.11	$<1 \times 10^{-3}$
hsa00970	Aminoacyl-tRNA biosynthesis	16.70	$<1 \times 10^{-3}$
hsa03022	Basal transcription factors[a]	9.83	3.50×10^{-3}
hsa03030	DNA polymerase[a]	9.16	4.33×10^{-3}
hsa00100	Biosynthesis of steroids	9.54	4.61×10^{-3}
hsa00260	Glycine, serine, and threonine metabolism	8.52	6.29×10^{-3}
hsa00240	Pyrimidine metabolism	16.62	7.75×10^{-3}
hsa05050	Dentatorubropallidoluysian atrophy	5.41	1.37×10^{-2}
hsa00350	Tyrosine metabolism	5.91	1.52×10^{-2}
hsa00450	Selenoamino acid metabolism	5.44	1.70×10^{-2}
hsa01510	Neurodegenerative disorders	5.03	1.74×10^{-2}
hsa04510	Integrin-mediated cell adhesion	6.00	2.22×10^{-2}
hsa04350	TGF-β signaling pathway	5.59	2.39×10^{-2}
hsa00511	*N*-Glycan degradation	4.03	3.00×10^{-2}

These pathways show significantly stronger (Kruskal-Wallis $p < 0.05$) intra- versus interpathway gene expression correlations.
[a] Pathway composed of genes encoding parts of large protein complexes.

apparent in that "cyclins and cell cycle regulation" is the most cohesive BioCarta pathway and "electron transport reaction in mitochondria," which corresponds to the oxidative phosphorylation pathway in KEGG, is one of the top cohesive pathways in BioCarta as well. BioCarta, however, has no genes in common with the most cohesive KEGG pathway, the ribosome pathway.

Annotation of all BioCarta genes in terms of functional categories defined in GO shows an enrichment of genes involved in cell cycle, RNA splicing, translation regulation, nuclear pore and DNA metabolism, signal transduction, chromatin modification, cell growth and proliferation, and apoptosis. The genes in the cohesive BioCarta pathways, in addition, are especially enriched in RNA splicing, nuclear pore and DNA metabolism, cell cycle, and cell growth. KEGG, however, shows a significant lack of genes, except for cell cycle, in those categories. As a result, the cohesiveness of these pathways cannot be evaluated within the KEGG pathway system. BioCarta further groups its pathways into 12 different categories, as shown in Table 5. Four of those categories, adhesion, cell cycle regulation, developmental biology, and metabolism, have an above-average (>13%) number of cohesive pathways; but none of them, with the exception of metabolism ($p = 0.012$; see Table 5), is significantly different from average at the 95% confidence level ($p \leq 0.05$). The other 8 categories have less than average (<13%) numbers of cohesive pathways. Among those, 2 pathway categories, cell signaling and cytokines/chemokines, show a significant ($p < 0.05$) lack of cohesive pathways. This is consistent with the results obtained from the KEGG pathway system, in which most of the signaling pathways are not significantly cohesive.

Contrary to the findings with KEGG, however, the pathways categorized as metabolic in BioCarta are significantly more cohesive than a typical pathway. The reason may be due, largely, to the fact that while the contents of metabolic pathways in KEGG and BioCarta might be expected to be similar, they are largely different. Metabolism is the largest pathway category in KEGG, which includes 81 pathways, 11 of which are statistically cohesive. On the other hand, BioCarta defines only 22 metabolic pathways and 8 of them are statistically cohesive. A closer examination of the genes involved in the BioCarta metabolic pathways in comparison to those defined by KEGG reveals that some of the genes over represented in the BioCarta metabolic pathways are actually involved in KEGG pathways that are mostly (12 of 14) *not* defined as metabolic, and a couple of them are significantly cohesive, such as the integrin-mediated cell adhesion and neurodegenerative disorders pathways. At the same time, the genes involved in most KEGG metabolic pathways are largely *under*represented in BioCarta. Combining the two pathway systems together results in a total of 373 pathways, 55 of which are cohesive; and a total of 103 metabolic pathways, 19 of which are cohesive. This means 15% of all pathways are cohesive and 18% of metabolic pathways are cohesive. This difference is, however, not statistically significant (Fisher's exact $p = 0.36$). Therefore, the overall conclusion regarding the cohesiveness of metabolic pathways is that they are *not* significantly different from a typical pathway. Nevertheless, it would be interesting to assess which metabolic pathways are the most and least cohesive. Using the combined pathways, the genes involved in the cohesive metabolic pathways are compared with the ones in the noncohesive metabolic pathways in terms of their functions, as annotated by GO. The genes in cohesive metabolic pathways are shown to be significantly enriched in functional categories such as proton transport, mitochondrial electron transport, oxidative phosphorylation, DNA-directed RNA polymerase activity, transcription, DNA replication/binding/metabolism, nucleoside biosynthesis and metabolism, and isoprenoid and cholesterol biosynthesis.

Table 3
Summary statistics for KEGG pathway categories and their level of cohesiveness

Pathway category	Pathway count	Cohesive pathway count	Cohesive (%)	p (Fisher's exact)
Genetic information processing	8	5	62.5	0.03
Cellular processes	4	2	50.0	0.24
Human diseases	8	2	25.0	0.66
Metabolism	81	11	13.6	0.07
Environmental information processing	9	1	11.1	1

The statistical significance of the coherence level within each pathway category is indicated by a p value yielded from the Fisher's exact test comparing the percentage of cohesive pathways that belong to that category with the percentage of all cohesive pathways in KEGG.

Table 4
Statistically cohesive BioCarta pathways ordered in decreasing level of significance (p values)

Pathway title	H score	p
Cyclins and cell cycle regulation	33.12	$<1 \times 10^{-3}$
The PRC2 complex sets long-term gene silencing through modification of histone tails	22.82	$<1 \times 10^{-3}$
Spliceosomal assembly	22.74	$<1 \times 10^{-3}$
Role of Ran in mitotic spindle regulation	20.84	$<1 \times 10^{-3}$
Antigen processing and presentation	19.53	$<1 \times 10^{-3}$
Agrin in postsynaptic differentiation	14.82	$<1 \times 10^{-3}$
Glycolysis pathway	14.65	$<1 \times 10^{-3}$
Apoptotic DNA fragmentation and tissue homeostasis	13.38	$<1 \times 10^{-3}$
Ion channels and their functional role in vascular endothelium	7.84	$<1 \times 10^{-3}$
Cycling of Ran in nucleocytoplasmic transport	10.29	1.57×10^{-3}
Multidrug resistance factors	8.49	1.81×10^{-3}
Electron transport reaction in mitochondria	9.26	2.94×10^{-3}
uCalpain and friends in cell spread	9.26	2.95×10^{-3}
CTL-mediated immune response against target cells	7.55	4.10×10^{-3}
Granzyme A-mediated apoptosis pathway	8.23	4.53×10^{-3}
mCalpain and friends in cell motility	8.05	4.75×10^{-3}
β-Arrestins in GPCR desensitization	6.92	5.55×10^{-3}
Eukaryotic protein translation	6.57	6.87×10^{-3}
Sonic hedgehog receptor Ptc1 regulates cell cycle	6.69	8.29×10^{-3}
ER-associated degradation pathway	6.83	8.53×10^{-3}
Regulation of MAP kinase pathways through dual-specificity phosphatases	6.03	9.28×10^{-3}
Mechanism of protein import into the nucleus	6.55	9.40×10^{-3}
cdc25 and chk1 regulatory pathway in response to DNA damage	5.86	1.03×10^{-2}
Overview of telomerase RNA component gene hTerc transcriptional regulation	4.84	1.40×10^{-2}
The IGF-1 receptor and longevity	5.38	1.43×10^{-2}
Internal ribosome entry pathway	5.37	1.43×10^{-2}
Adhesion molecules on lymphocyte	4.71	1.79×10^{-2}
Monocyte and its surface molecules	4.71	1.79×10^{-2}
Activation of PKC through G-protein-coupled receptor	4.83	1.79×10^{-2}
Attenuation of GPCR signaling	4.36	1.95×10^{-2}
ChREBP regulation by carbohydrates and cAMP	4.36	1.95×10^{-2}
Ghrelin: regulation of food intake and energy homeostasis	4.31	1.99×10^{-2}
Activation of cAMP-dependent protein kinase PKA	3.92	2.46×10^{-2}
Stathmin and breast cancer resistance to anti-microtubule agents	4.36	2.59×10^{-2}

These pathways show significantly stronger (Kruskal-Wallis $p < 0.05$) intra- versus interpathway gene expression correlations.

Conversely, the genes involved in noncohesive metabolic pathways are significantly enriched in categories including Golgi apparatus, metabolism, fatty acid metabolism, endoplasmic reticulum, protein serine/threonine kinase activity, RAB small monomeric GTPase activity, protein amino acid glycosylation, and lipid catabolism. Clearly, there appears to be a functional separation between the cohesive and the noncohesive metabolic pathways.

Expression coherence in gene categories defined by Gene Ontology

GO annotates genes according to their participation in a biological process, the molecular function of the gene product, or the specific cellular location of the expressed gene. Each of the 787 GO terms that have at least three genes with available expression data was evaluated for its *intraterm* expression coherence. Among these gene categories, 171 (22%) show significant expression cohesiveness within a GO term. Because of space considerations, only the 45 GO terms that are cohesive at a significance level of $p < 10^{-3}$ and their H scores are listed in Table 6. These results show a clear agreement with those derived from the KEGG and BioCarta pathways. For example, the GO terms that are the most cohesive include genes that are engaged in processes such as protein biosynthesis and ribosomal pathways, which correspond to the ribosome pathway in KEGG; cell cycle, which corresponds to the same pathways in KEGG and BioCarta; RNA splicing and binding, which are also cohesive in BioCarta; mitochondrial electron transport, which corresponds to the same pathway in BioCarta and the oxidative phosphorylation pathway in KEGG; DNA binding and metabolism, which are also cohesive in BioCarta; and other activities that are not sufficiently represented in KEGG and BioCarta, such as metal ion binding and transport. The GO category of signal transduction turns out to be the least cohesive GO term of all. This is, in fact, consistent with the findings from the analysis of the KEGG and BioCarta pathways, in both cases of which the signaling pathway category contains a less than average number of cohesive pathways. This finding may be of importance in decisions about efforts to discover agents that target signaling pathways.

In addition, we find that many GO terms representing large protein complexes are significantly cohesive. There are 32 GO

Table 5
Summary statistics for BioCarta pathway categories and their level of cohesiveness

Pathway category	Pathway count	Cohesive pathway count	Cohesive (%)	p (Fisher's exact)
Metabolism	22	8	36.4	0.012
Adhesion	14	4	28.6	0.14
Cell cycle regulation	25	5	20.0	0.36
Developmental biology	23	4	17.4	0.53
Expression	51	5	9.8	1
Immunology	50	4	8.0	1
Apoptosis	27	2	7.4	1
Cell signaling	149	10	6.7	0.003
Cell activation	18	1	5.6	1
Cytokines/ chemokines	39	0	0	0.012
Neuroscience	18	0	0	0.24
Hematopoiesis	8	0	0	0.60

The statistical significance of the coherence level within each pathway category is indicated by a p value yielded from the Fisher's exact test comparing the percentage of cohesive pathways belonging to that category with the percentage of all cohesive pathways in BioCarta.

Table 6
Statistically cohesive GO terms ordered in decreasing H values

Term	Title	H score
Biological process		
GO:0006412	Protein biosynthesis	1396.83
GO:0007067	Mitosis	205.04
GO:0000398	Nuclear mRNA splicing, via spliceosome	186.73
GO:0006260	DNA replication	124.32
GO:0006397	mRNA processing	98.59
GO:0006118	Electron transport	69.56
GO:0007049	Cell cycle	60.92
GO:0008380	RNA splicing	53.67
GO:0000910	Cytokinesis	43.53
GO:0006120	Mitochondrial electron transport, NADH to ubiquinone	36.34
GO:0006091	Generation of precursor metabolites and energy	33.66
GO:0006265	DNA topological change	29.66
GO:0007001	Chromosome organization and biogenesis (*sensu* Eukaryota)	29.13
GO:0006406	mRNA-nucleus export	22.54
GO:0000070	Mitotic sister chromatid segregation	18.10
Molecular function		
GO:0003735	Structural constituent of ribosome	2339.68
GO:0003723	RNA binding	1388.27
GO:0003677	DNA binding	192.12
GO:0008248	Pre-mRNA splicing factor activity	132.11
GO:0016491	Oxidoreductase activity	90.43
GO:0008137	NADH dehydrogenase (ubiquinone) activity	54.69
GO:0003954	NADH dehydrogenase activity	50.42
GO:0004129	Cytochrome *c* oxidase activity	45.70
GO:0005507	Copper ion binding	42.23
GO:0005509	Calcium ion binding	37.93
GO:0003918	DNA topoisomerase (ATP-hydrolyzing) activity	27.86
GO:0004175	Endopeptidase activity	22.87
GO:0015078	Hydrogen ion transporter activity	20.62
GO:0019843	rRNA binding	18.99
GO:0046870	Cadmium ion binding	17.99
Cellular component		
GO:0005840	Ribosome	1290.60
GO:0005739	Mitochondrion	995.19
GO:0005842	Cytosolic large ribosomal subunit (*sensu* Eukaryota)	737.65
GO:0005843	Cytosolic small ribosomal subunit (*sensu* Eukaryota)	583.62
GO:0005622	Intracellular	464.34
GO:0019866	Inner membrane	63.27
GO:0005643	Nuclear pore	51.98
GO:0005764	Lysosome	47.80
GO:0030530	Heterogeneous nuclear ribonucleoprotein complex	47.22
GO:0005624	Membrane fraction	43.73
GO:0005694	Chromosome	25.07
GO:0005746	Mitochondrial electron transport chain	24.99
GO:0000786	Nucleosome	24.59
GO:0005839	Proteasome core complex (*sensu* Eukaryota)	22.87
GO:0005732	Small nucleolar ribonucleoprotein complex	19.99

These GO terms show significantly stronger (Kruskal-Wallis $p < 0.05$) intra- versus interterm gene expression correlations. Only GO terms significantly cohesive at $p < 1 \times 10^{-3}$ are listed.

terms that end with "some" (ribosome, lysosome, chromosome, nucleosome, proteasome, etc.) and 13 (41%) of them are cohesive, which is significantly higher than the average frequency of cohesive GO terms (22%) (Fisher's exact $p = 0.017$; see Table 7). The GO terms that contain the word "complex" also show a slightly higher than average frequency of cohesive terms but the difference is not statistically significant. In fact, among the three different ontologies, biological process, molecular function, and cellular component, the first two contain about 20% significantly cohesive terms; cellular component, however, has 30% significantly cohesive terms (see Table 7). This difference is significant (Fisher's exact $p = 0.043$), which indicates that genes expressed in the same cellular location, which include, but are not dominated by, genes encoding proteins that form large protein complexes, appear to be more coherently expressed than genes simply involved in the same biological process or engaging in the same molecular function.

Gene expression coherence in pathways (KEGG, BioCarta) versus functionally related gene groups (GO)

So far we have been analyzing the level of gene expression coherence using three annotation schemes without making the distinction between a pathway and a group of functionally related genes, although the functionally related gene groups defined by GO are not pathways per se. Even though genes involved in the same pathway are generally assumed to be functionally related, it would still be interesting to examine the gene composition within a pathway in terms of GO functional categories. For this purpose, we calculated the percentage of genes in each KEGG or BioCarta pathway that belong to the same GO term, which turned out to be 50% on average (data not shown). In fact, 41% of KEGG pathways and 56% of BioCarta pathways have over half of their genes belonging to the same GO term. This confirms the assumption that pathways contain functionally related genes. A further analysis shows that the pathways with a higher percentage of genes belonging to the same GO term are *not* significantly more cohesive than the ones with a lower percentage. This is consistent with the earlier finding that GO terms are not significantly more cohesive than pathways.

Table 7
Summary statistics for the three ontologies (biological process, molecular function, and cellular component) and multicomponent protein complexes (defined by terms containing either "some" or "complex") defined by GO and their level of cohesiveness

GO category	Term count	Cohesive term count	Cohesive (%)	p (Fisher's exact)
Molecular function	311	61	19.6	1
Biological process	346	71	20.5	1
Cellular component	130	39	30.0	0.043
Large protein complex ("some")	32	13	40.6	0.017
Protein complex ("complex")	47	11	23.4	0.86

The statistical significance of the coherence level within each GO category is indicated by a p value yielded from the Fisher's exact test comparing the percentage of cohesive pathways belonging to that category with the percentage of all cohesive terms in GO.

Pathway or functionally related genes are more coherently expressed than a random set of genes

Based on our analysis of the level of coordinated gene expression within various pathway systems or functional categories we find that some functionally related genes are more coherently expressed than others. The question still remains as to whether the coexpression of genes could be observed just by random chance, that is, whether the categorization of genes into pathways is meaningful. To test this hypothesis, pathways (groups of genes) are built by randomly selecting and assigning genes to each pathway, and the cohesiveness, measured by an H score, is evaluated for each of these random pathways (see Data and methods for details). For each annotation system, genes are randomly permutated a thousand times and H scores for each pathway calculated. The distribution of the H scores for these randomly generated pathways, together with that of the corresponding real gene annotation system, KEGG, BioCarta, or GO, is shown in Fig. 1. The difference in the distributions clearly shows that our observations of intrapathway gene expression cohesiveness are not random. The three distributions for the randomly generated gene groups are almost completely overlapping, yielding a mean random pathway cohesiveness H score of close to 0 (0.1) and a median score of 0. The distributions of the three real systems are similar to one another, but are clearly right-shifted, with the peak H score at around 1.0, compared to the random distributions. The mean and median H scores for the three annotation systems are KEGG 47.29, 0.58; BioCarta 1.47, 0.28; and GO 17.34, 0.59. The mean H scores for KEGG and GO are skewed by several extremely cohesive ($H > 1000$) pathways or GO terms and appear to be much larger than the median scores as a result. The observed differences in the H score distributions generated from the real systems compared to those within the randomly selected groups of genes are statistically very significant (t test: GO, $p = 2.61 \times 10^{-23}$; KEGG, $p = 7.89 \times 10^{-19}$; BioCarta, $p = 1.20 \times 10^{-15}$). This leads to the conclusion that gene expression cohesiveness observed within pathways is very unlikely to occur by chance; therefore, it is evident that functionally related genes or genes involved in the same pathway are, overall, more coherently expressed than a random set of genes, and coordinated regulation of these genes, to a certain level, is a real component of their biological organization.

BioCarta pathways are less cohesive than KEGG and GO

Additionally, among the three gene annotation schemes analyzed, the pathways defined by BioCarta seem to be the least cohesive, with the lowest mean and median H scores, compared to those of KEGG or GO, and this observation is significant at the 90% confidence level (t test: BioCarta vs KEGG, $p = 0.07$; BioCarta vs GO, $p = 0.05$). Considering that the mean H scores for KEGG and GO are probably

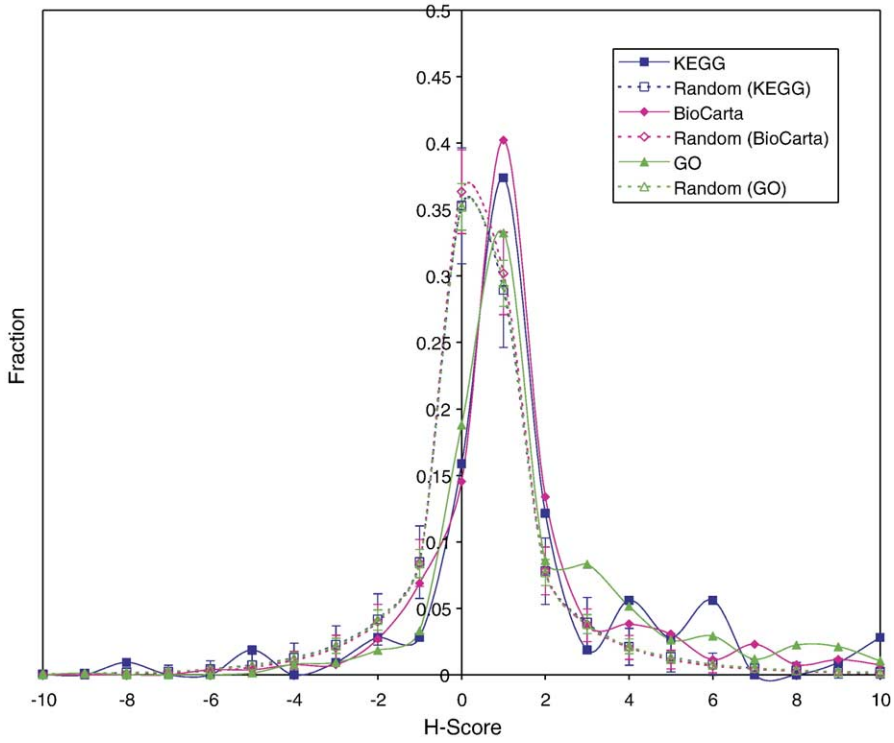

Fig. 1. Distributions of Kruskal–Wallis H scores calculated for GO terms and KEGG and BioCarta pathways in comparison to sets of gene groups (random pathways) generated by randomly assigning genes to each group (see Data and methods for details). A large positive H score indicates a high level of pathway cohesiveness. The distributions of the three true pathway systems are clearly right-shifted compared to the random distributions. The mean H score for the set of random pathways is close to 0 (0.1) and the median is 0. The median H scores for the three annotation systems are 0.58 (KEGG), 0.28 (BioCarta), and 0.59 (GO), which are significantly different from random (t test: GO, $p = 2.61 \times 10^{-23}$; KEGG, $p = 7.89 \times 10^{-19}$; BioCarta, $p = 1.20 \times 10^{-15}$).

skewed by the extremely cohesive ribosome pathway (H > 1000), which is not in the BioCarta pathway system, the median H scores for KEGG and GO are almost the same, whereas the median H score for BioCarta is 50% lower. This can probably be explained by the inherent differences within the nature of genes annotated by KEGG, BioCarta, and GO. We have analyzed a total of 3365 genes, each of which is present in at least one of the three annotation schemes, and 1995 of them are annotated only by GO. Of the three, GO is the most comprehensive gene function annotation scheme, which tries to assign an annotation to every gene with no bias toward any specific type of gene. Approximately 90% of the genes present in KEGG or BioCarta are also annotated by GO. However, rather than generating annotations per gene, KEGG attempts to curate and elaborate on metabolic pathways, large molecular assemblies, and regulatory pathways. In fact, a large majority, 81 of 111 (73%), of the KEGG pathways that we have analyzed are metabolic pathways. As we have shown earlier, metabolic pathways are not significantly more or less cohesive than a typical pathway. On the other hand, BioCarta has a large collection of signaling pathways and a smaller number of pathways in other categories. Cell signaling is the largest pathway category in BioCarta, which constitutes 57% (149 of 262) of all BioCarta pathways we have examined. Signaling pathways, however, are significantly less cohesive than a typical pathway. This may provide an explanation for the observation that the overall cohesiveness of KEGG pathways is similar to that of GO functional categories but BioCarta pathways appear to be less cohesive than both.

Hierarchical clustering of KEGG pathways shows a higher level of gene expression regulation

An additional perspective on gene expression and pathways can be accomplished by independently clustering gene expressions across tumor cells and then assessing pathway coherence within neighborhood clusters. Hierarchical clustering of the spread pattern (occurrences of genes within a pathway in SOM clades; see Data and methods for details) of the 111 KEGG pathways on the gene expression SOM (Fig. 2, cohesive pathways are labeled with red dots) further segregates these pathways into 15 clusters (see Fig. 3), which can be used to assess whether pathways involved in similar biological processes share similar regulation by gene expression patterns. As mentioned earlier, KEGG groups its pathways into five general cellular process categories (Fig. 3, top), which are further divided into 22 subcategories (Fig. 3, bottom). Fig. 3 shows the composition of each of the 15 pathway clusters, shown in the same order as the dendrogram in Fig. 2, i.e., adjacent clusters share similar spread patterns, according to the five general categories (top) and the 22 subcategories (bottom). One can see that pathways belonging to the same category usually cluster together, that is, genes from these pathways are partly coexpressed. The 81 metabolic pathways are scattered across almost all clusters, but 6 of the 15 clusters are composed solely of metabolic pathways (Fig. 3, top). Furthermore, cluster 1 is composed solely of metabolic pathways involved in glycan biosynthesis and metabolism and cluster 4 solely of lipid metabolism (Fig. 3, bottom).

Cluster 10 contains all of the six signal transduction pathways, which belong to the environmental information processing category. The other two pathway subcategories that belong to environmental information processing are ligand–receptor interaction, whose pathways are clustered together in cluster 14, and immune system, which forms one cluster by itself in cluster 11. Five of the eight pathways involved in genetic information processing are grouped together in cluster 15. The ribosome pathway, which belongs to the translation subcategory in genetic information processing, and the immune system subcategory, which has only one pathway (complement and coagulation cascades), each form one single cluster by themselves (clusters 5 and 11, respectively). These results can be used as an indication of the degree of communication or interaction, and subsequently the level of coregulation, between different pathways. Clearly, pathways that participate in the same cellular processes tend to cluster together and thus show a high degree of interpathway communication by sharing or coexpressing part of their genes. Therefore, these pathways are probably coregulated at a level higher than basic gene expression regulation. Moreover, the pathways that belong to different categories or subcategories but are clustered together, as shown at the bottom of Fig. 3, provide valuable clues about possible interactions or coregulation patterns between these pathways.

Pathway gene spread over the SOM and expression coherence

As employed by many researchers, simple clustering of gene expression patterns can also be used to detect coexpressed or coregulated genes [27]. However, one needs to be aware of the fact that clustering detects only one form of coregulation, that is, positive correlations between gene expression patterns, since negatively correlated genes are generally not clustered together. Nonetheless, the degree of spread (DoS; the number of SOM clades the pathway occupies divided by the maximum number of genes in the pathway in one clade) of genes within a pathway on the gene expression SOM can be used as an alternative measure of pathway gene expression coherence, especially pathways that are dominated by positive gene correlations. In fact, DoS turns out to be significantly correlated with the gene expression cohesiveness H scores ($r = -0.21$, $p = 0.027$), that is, the more cohesive a pathway is, the less of a spread it will have on the gene SOM. The average DoS score of the KEGG pathways is 4.9. The most cohesive is again the ribosome pathway, having the smallest DoS (0.17), with 80% of its genes clustered into one clade (see Fig. 2; the ribosome pathway has only one white square). Other significantly cohesive pathways, such as cell cycle, oxidative phosphorylation, terpenoid biosynthesis, aminoacyl-tRNA biosynthesis, and biosynthesis of steroids, also have small DoS scores (<3),

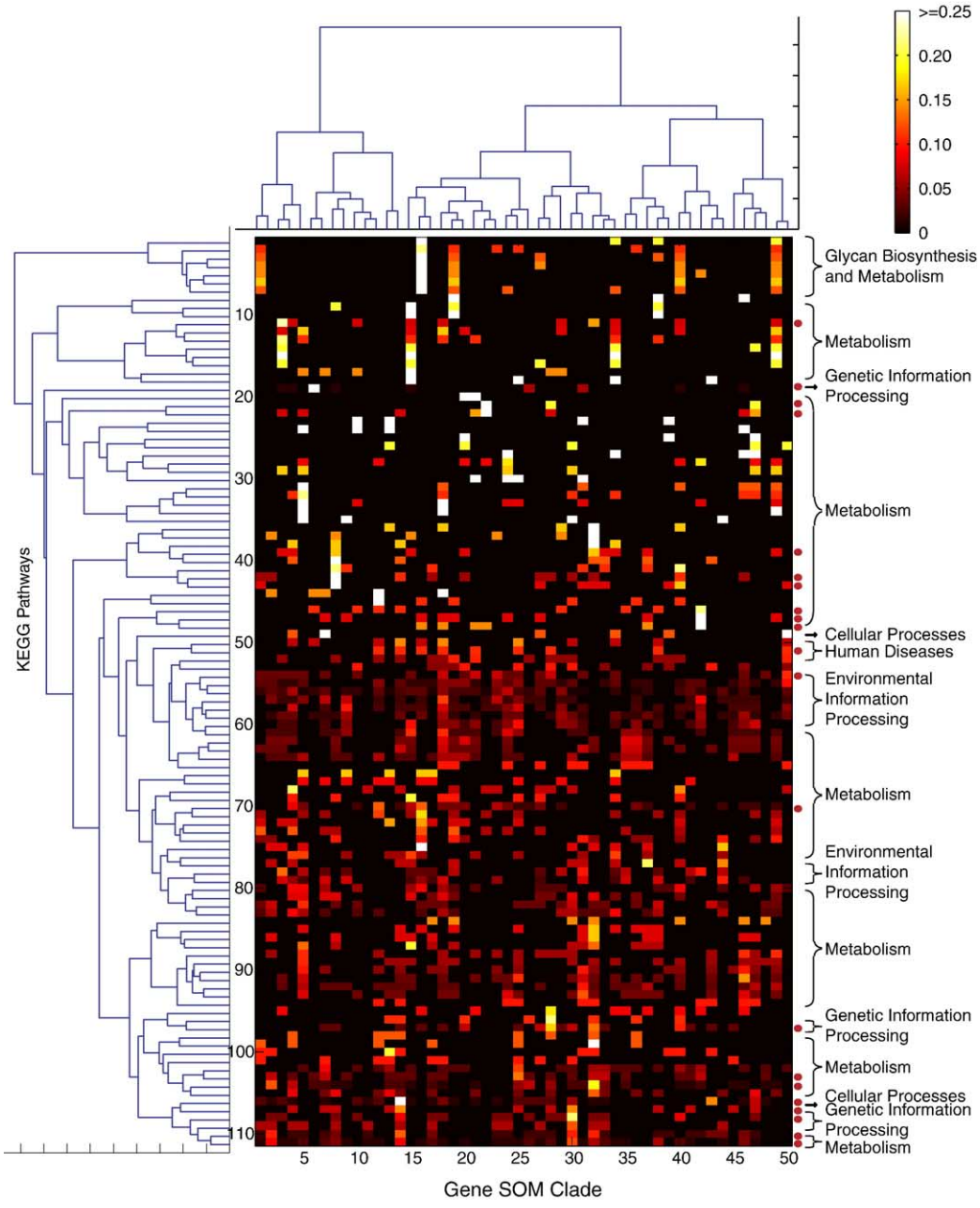

Fig. 2. Hierarchical clustering of 111 KEGG pathways. The genes are first grouped into 50 clades based on their expression patterns across the NCI_{60} using the SOM clustering procedure. The fraction of genes that fall into each SOM clade is then calculated for every pathway, yielding 111 data vectors. These data vectors are finally clustered hierarchically, grouping similar pathways together. Each row represents a pathway and each column represents a SOM clade. The fraction of genes in a pathway that fall into a particular clade is shown by the color of the square in which the pathway and the clade intersect. A yellow to white color indicates a high fraction value, a reddish color indicates a medium fraction value, and black indicates zero. The dendrogram generated from the hierarchical clustering of pathways is shown on the left and the 111 pathways are arranged in the order in which they appear in the dendrogram such that neighboring pathways are similar to each other and have many of their genes fall into the same clade. The dendrogram at the top is generated from hierarchical clustering of the SOM nodes, each cluster of which forms a clade. The 50 clades are also shown in their dendrogram order such that genes in neighboring clades share similar expression patterns. The general categories each pathway group belongs to are shown on the right (all pathway names can be found in the supplementary information). The cohesive pathways (intrapathway gene–gene correlations are significantly stronger than interpathway gene–gene correlations at $p < 0.05$; see Data and methods for details) are labeled with red dots. These pathways generally have their genes localized in a few clades, whereas the less cohesive pathways tend to have a larger spread of their genes across many SOM clades.

characterized by occupying few, but hot, white squares in Fig. 2. These pathways are dominated by strong positive intrapathway gene–gene correlations. Interestingly, however, some pathways shown as significantly cohesive by their H scores, such as the TGF-β signaling pathway and neurodegenerative disorders, exhibit a large degree of spread over the SOM (DoS $\geq$ 7; see Fig. 2), indicating that these pathways are enriched in negative intrapathway gene–gene correlations. Pathways that have smaller than average DoS scores but are not shown as significantly cohesive, for example sulfur

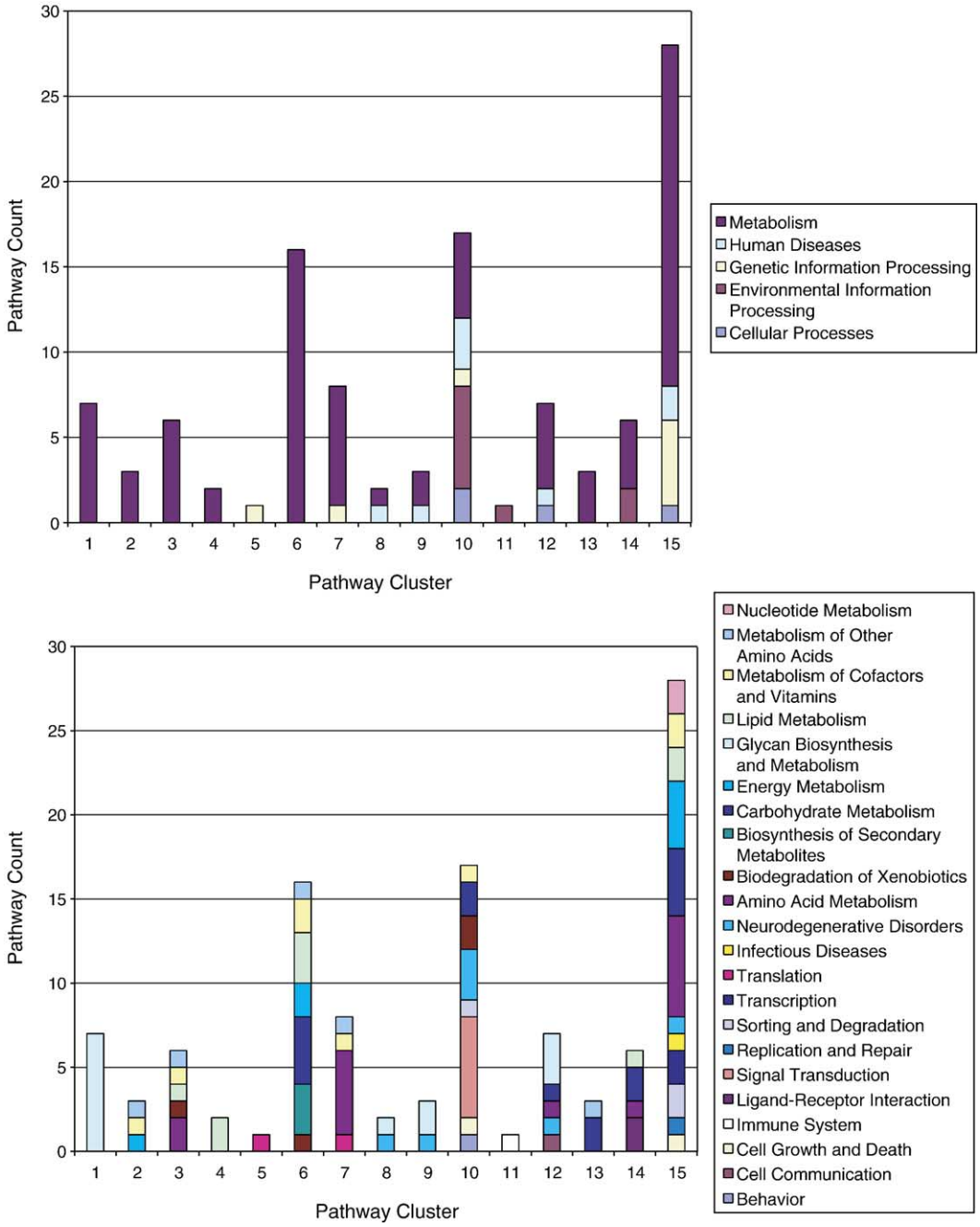

Fig. 3. Histograms of the 15 KEGG pathway clusters generated from hierarchical clustering (see Fig. 2 legend for details). Each histogram represents the number of pathways in a cluster and the clusters are ordered as they appear in the dendrogram. Top: Histograms are colored according to the five general pathway categories (shown in the key) defined by KEGG. Bottom: Histograms are colored according to the 22 KEGG pathway subcategories (shown in the key). Pathways engaged in similar cellular processes tend to cluster together.

metabolism, ascorbate and aldarate metabolism, chondroitin/heparan sulfate biosynthesis, keratan sulfate biosynthesis, and nitrogen metabolism, are themselves characterized by positive, but weak, intrapathway gene–gene correlations or do not have enough genes to achieve statistical significance.

As we have mentioned in previous sections, pathways may communicate with each other not only through coexpressed genes but also by actually sharing genes. In fact, many genes participate in multiple pathways and the degree of pathway cross talk, manifested by gene sharing, may be another factor that contributes to the cohesiveness of a pathway. Since the DoS score of a pathway can be used as another indicator of pathway cohesiveness, we have examined the relationship between the number of genes in a pathway that are shared with other pathways and the DoS score of that pathway for the KEGG pathways. These two attributes are significantly correlated with each other ($r = 0.36$, $p = 9.37 \times 10^{-5}$), that is, the more genes that a pathway shares with others, the less cohesive the pathway. Fig. 4 clearly shows that the genetic information processing pathway category, which contains the

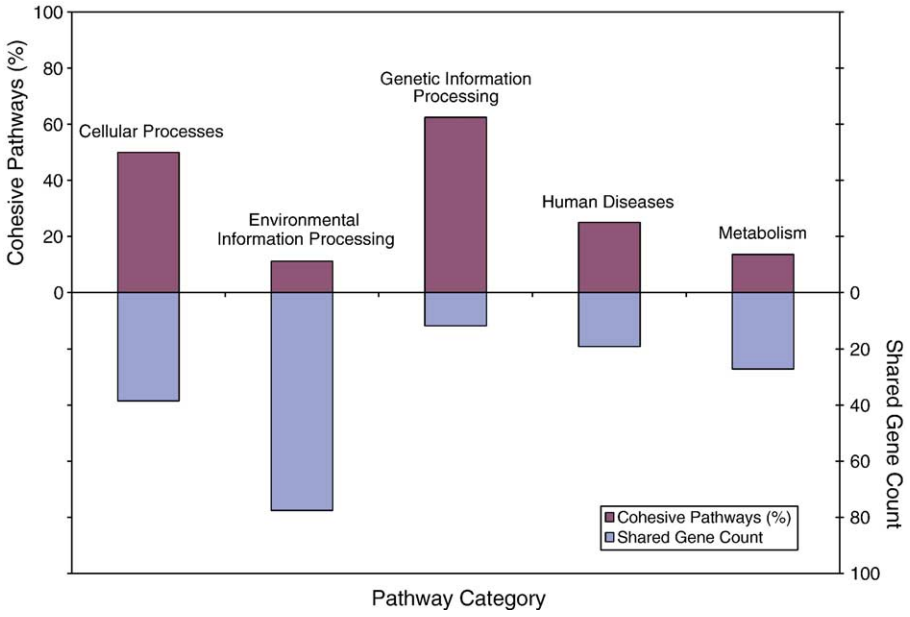

Fig. 4. Relationship between the number of genes in a pathway that are shared with other pathways and the cohesiveness of the pathway. Each histogram represents a KEGG pathway category. Top: Each histogram shows the percentage of cohesive pathways in a particular pathway category. Bottom: Each histogram shows the average number of genes that a pathway in that category shares with others. Pathways that share many genes with others tend to be less cohesive. The genetic information processing KEGG category has the highest percentage of cohesive pathways, each of which has the lowest number of genes shared on average.

most cohesive pathways, is also the most modular, i.e., it has the least number of shared genes (12 per pathway). Environmental information processing, which contains all the signal transduction pathways, is one of the least cohesive pathway categories and also has the largest number of genes shared with other pathways (78 per pathway).

Discussion

Our study provides a comprehensive evaluation of the level of coregulation in pathway gene expressions measured across the NCI_{60} using three different gene annotation schemes, KEGG, BioCarta, and GO. We have discovered that the level of gene coexpression is, overall, significantly higher in pathways or functionally related gene groups than a randomly selected set of genes. Approximately 20% of pathways or functionally related gene groups analyzed have statistically significant, coherent gene expressions. Based on the types of pathways found to be cohesive vs noncohesive, we postulate that pathway gene expression cohesiveness is probably on a "need to be" basis, that is, genes in the same pathway are coexpressed only when there is a need for it. Pathways are probably designed by nature to be robust and flexible enough to adapt to environmental changes to ensure cell survival, that is, alternate mechanisms can take over if parts of a pathway fail to operate. However, pathways with genes encoding parts of a large protein complex need to be cohesive probably because the co-presence and close physical interaction of the proteins required for the proper function of the protein complex demand the coexpression and tight regulation of their corresponding genes. The same may be true for genes that are expressed in the same cellular location/component, which are found to be more cohesive than genes participating in the same biological process or engaging in the same molecular function.

Pathways involved in genetic information processing (replication and repair, sorting and degradation, transcription, translation) and cell cycle tend to be cohesive as well, probably because they are responsible for the most vital processes in a biological entity and, therefore, need to be tightly regulated at the transcriptional level to ensure precisely synchronized action of their constitutive genes to minimize error, or probably because these processes are important for growth and proliferation and therefore are cohesive in constantly proliferating tumor cells. On the other hand, pathways involved in environmental information processing (signaling pathways) and most metabolic pathways are generally found to be not cohesive, presumably because these pathways need to be more robust to respond to environmental changes. The co-presence of all these genes may not be necessary, and genes are turned on or off sequentially when needed. Therefore, these pathways do not need to be tightly regulated at the transcriptional level and they are more likely to be regulated by substrate concentration and enyzme–substrate interactions. This is also reflected by the fact that these pathways are much less modular than the cohesive pathways. There is a high degree of cross talk (gene sharing) between the signal transduction pathways as opposed to the very cohesive pathways responsible for genetic information processing, which are highly modular. The few metabolic pathways that are cohesive show, as previously found, that pathways responsible for vital life processes such as nucleotide metabolism, energy metabolism, and isoprenoid and cholesterol biosynthesis need to have more tightly regulated gene expression than generic metabolic pathways, such as carbohydrate metabolism, metabolism of cofactors

and vitamins, fatty acid metabolism, protein amino acid glycosylation, and lipid catabolism.

No significant distinction is found in the level of cohesiveness between a pathway, as defined by KEGG or BioCarta, and a functional gene category, as defined by GO. This can partially be attributed to the fact that most genes in a pathway are functionally related, as reflected by over 50% of pathways containing over half of their genes belonging to the same GO category. KEGG is characterized by a large percentage of metabolic pathways, whereas the largest pathway category in BioCarta is signaling pathways, which are mostly not cohesive. As a result, BioCarta contains a smaller percentage of cohesive pathways than KEGG and GO. In addition, a higher level of pathway regulation is revealed by hierarchical clustering of pathway gene spread patterns on the gene expression SOM, which shows interactions between pathways that engage in similar cellular processes. A similar phenomenon has been observed in the metabolic network of the yeast *Saccharomyces cerevisiae* [2,3]. Also noteworthy is that most of the significantly cohesive KEGG pathways (highlighted in red in Fig. 2) appear in neighboring SOM clades. This shows a higher level of coherence within the cellular processes, namely genetic information processing, cell cycle, energy metabolism, and nucleotide metabolism, in which these pathways are participating. Clusters that contain pathways from different cellular processes are, furthermore, indicative of interprocess communications through coexpression and thus coregulation of these pathways.

It would be interesting, as a future endeavor, to examine the pathway gene expression coherence in normal tissues in a similar fashion and compare the results to those obtained herein, since the gene expression data we have analyzed are derived solely from cancer cell lines. It is hoped that this will provide clues as to which and how pathway regulations have changed in cancer. Knowledge of the nature of gene expression regulation and biological pathways can be applied toward understanding the mechanism by which compound substances or drug molecules interfere with the biological system through interactions with gene products and consequently pathways.

Conclusion

We have analyzed gene expression patterns derived from the NCI_{60} in the context of predefined pathways or functional categories annotated by KEGG, BioCarta, and GO. The degree of gene coexpression, gene coregulation, or pathway cohesiveness in these three schemes demonstrates that expression of genes in pathways, or functionally related genes, has a significantly higher level of coherence than that of a random set of genes. Pathways with genes encoding physically interacting proteins and pathways involved in genetic information processing and cell cycle tend to be cohesive and modular, whereas signaling pathways are generally not cohesive and have a high degree of interpathway cross talk. Most metabolic pathways are not cohesive, except for the ones responsible for nucleotide metabolism, energy metabolism, and isoprenoid and cholesterol biosynthesis. Transcriptional level gene regulation appears to be on a “need to be” basis, such that pathways responsible for vital cellular processes or processes that are related to growth or proliferation, specifically in cancer cells, are the most cohesive. Clustering of pathways, in addition, reveals interesting interpathway communications or interactions indicative of a higher level of pathway regulation.

Data and methods

Gene expression data

Constitutive gene expression data from Novartis, measured in triplicate across the 60 tumor cell lines using the Affymetrix DNA oligonucleotide microarray technology, were downloaded from the Developmental Therapeutics Program Web server at http://www.dtp.nci.nih.gov. This data set contains 12,626 mRNA expression profiles and is publicly available. In this data set, each expression measurement (signal intensity) comes with a p value that indicates the reliability of that measurement. The signals with small p values are stronger and more reliable. We first filtered the data set to include only measurements that exhibited the strongest signal intensity ($p < 0.05$). The logarithm of each signal was taken to suppress extreme data values. Replicate measurements for each gene were then averaged by taking the median. Finally, only gene expression profiles having data available for at least 40 cell lines were included. This yielded a data set of 4923 genes for our analysis.

Pathway data

Three databases were used for pathway gene analysis: the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.ad.jp/kegg/), Gene Ontology (http://www.geneontology.org/), and BioCarta (http://www.biocarta.com/). Annotations for 134 human pathways containing 2804 genes were downloaded from the KEGG ftp site (ftp://ftp.genome.ad.jp/pub/kegg/pathways/hsa/). BioCarta annotations for 314 pathways containing 1406 human genes were downloaded from NCI’s Cancer Genome Anatomy Project (http://cgap.nci.nih.gov/) ftp site (ftp://ftp1.nci.nih.gov/pub/CGAP). Annotations for 3564 GO terms containing 10,921 human genes were downloaded from the GO ftp site (ftp://ftp.geneontology.org/pub/go/). In the gene expression data set we are working with, 1047 genes are present in KEGG, 604 are present in BioCarta, and 3210 are present in GO.

Pathway cohesiveness and significance calculation

For each pathway P_x, a Pearson correlation coefficient (r) was calculated for each unique gene pair using their expression patterns from the NCI_{60}. These values are referred to as *intra*pathway r values. Then for each gene expression data vector x in P_x, an r value is calculated between x and every gene data vector y, where gene y belongs to some other pathway P_y where $x \neq y$. These values are referred to as *inter*pathway r values. The Pearson correlation coefficient, r, is defined as

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (1)$$

where $\bar{x}$ and $\bar{y}$ denote averages, and the summation runs over the number of cell lines (n). The intra- and interpathway r values are then used as two sample populations to calculate the Kruskal–Wallis H statistic for pathway P_x,

$$H = \frac{12}{N(N+1)} \sum_{t=1}^{k} \left(\frac{R_t^2}{n_t} \right) - 3(N+1) \quad (2)$$

where R_t is the rank sum of sample population t, n_t is the size of sample population t, k is the number of sample populations being compared, and $N = \sum_{t=1}^{k} n_t$. When each of the k sample populations being compared includes

at least five observations, the sampling distribution of H is a very close approximation of the χ^2 distribution for $k-1$ degrees of freedom. It is actually a fairly close approximation even when one or more of the samples include as few as three observations. In the present study, only pathways that have at least three gene expression data vectors available are included in the calculations. Significance levels (p values) are obtained using the χ^2 distribution with 1 degree of freedom (two-sample populations, intra- and interpathway correlation coefficients, are compared here; therefore, $k = 2$ and $k - 1 = 1$).

A large H score ($H > 3.84$) indicates that a statistically significant ($p < 0.05$) difference exists between the intra- and the interpathway r populations. A small r is assigned a lower rank and a large r a higher rank. If the average rank of the interpathway r values is higher than that of the intrapathway r values, then a negative sign is applied to the H score. Therefore, a large positive H score indicates a high level of gene expression coherence within the pathway compared to expression of genes not linked by a known pathway. Either absolute or real r values can be used when calculating the H scores, and slightly different results will be obtained, depending on whether the intrapathway r values are dominated by positive or negative values. Absolute r values are used in all calculations unless otherwise specified, since both significant positive and negative correlations between two genes are assumed to be indicative of coordinated expression.

Randomization procedures are used to check the probability of getting a large positive H score by chance compared to the p values obtained directly from the χ^2 distribution. Random pathways of sizes approximately 5, 10, 15, 25, 50, and 100 are built by randomly picking and assigning genes in the data set to each pathway, and their H scores are calculated. This procedure is repeated 1000 times and the probabilities of getting H scores at various levels for each pathway size are calculated. Further increase in the number of randomizations does not appear to affect the outcome. The probabilities obtained this way agree very well with the p values obtained directly using the χ^2 distribution for pathway sizes 10 to 25 (number of genes in each pathway; the size of the pathways we are working with is 10–25), that is, an H score of >3.84 is required to get $p < 0.05$. However, the H score required to reach a certain significance level (e.g., $p < 0.05$) shows a slight linear dependency on pathway size (i.e., sample size). For larger pathways, higher H scores are needed to get a significant p value. For this reason, we have applied the randomization procedure to obtain the probability of observing an extreme H score by chance for each pathway. Therefore, if a pathway in a particular annotation system has n genes, then n genes are randomly selected from the pool of m genes annotated by the system and the H score is calculated. This procedure is repeated 1000 times for each pathway, and the fraction of H scores that are more extreme than the actual H score of the pathway is then assigned as the coherence p value for that particular pathway. If none of the 1000 random H scores is more extreme than the actual H score, then the pathway is significantly cohesive at $p < 0.001$. The number of significantly cohesive ($p < 0.05$) pathways selected in this fashion, however, does not differ much from that obtained directly from the χ^2 distribution because most pathways (>90%) have fewer than 50 genes and most significant pathways (>60%) have very large H scores (>7.5). For each gene annotation system, KEGG, BioCarta, or GO, a distribution of the pathway H scores was calculated for each random permutation; the mean and standard deviation (shown as error bars) of the thousand random distributions were then plotted and are shown in Fig. 1. These three random distributions are almost entirely coincidental.

We chose the H score as an indicator of pathway gene expression cohesiveness because a significance measure exists for the H score calculated for a pathway and, by using rank scores in place of actual correlation values, the method is less sensitive to pathway size and gene outliers. This method is also a more reliable measure of pathway coherence because not only gene–gene relationships within a pathway are considered, but also comparisons are made to check whether the intrapathway gene–gene interactions are in fact stronger than interactions between genes not connected by some known pathway.

Hierarchical clustering of KEGG pathways

The genes were first grouped into 50 clades based on their expression patterns across the NCI_{60} using the SOM clustering procedure. The fraction of genes that fall into each SOM clade was then calculated for every KEGG pathway, yielding 111 data vectors each containing 50 fractions. Wards-based single-linkage hierarchical clustering of the 111 data vectors was performed using MATLAB, which segregated the pathways into 15 groups. A dendrogram showing the neighboring pathways was generated.

Acknowledgments

The authors thank the members of the STB staff, especially Drs. Robert Shoemaker and Susan Mertins for valuable contributions during the preparation of the manuscript. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ygeno.2005.11.011.

References

[1] L. Hood, J.R. Heath, M.E. Phelps, B. Lin, Systems biology and new technologies enable predictive and preventative medicine, Science 306 (2004) 640–643.
[2] J. Ihmels, S. Bergmann, N. Barkai, Defining transcription modules using large-scale gene expression data, Bioinformatics 20 (2004) 1993–2003.
[3] J. Ihmels, R. Levy, N. Barkai, Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae, Nat. Biotechnol. 22 (2004) 86–92.
[4] Z. Li, C. Chan, Inferring pathways and networks with a Bayesian framework, FASEB J. 18 (2004) 746–748.
[5] Z. Li, C. Chan, Integrating gene expression and metabolic profiles, J. Biol. Chem. 279 (2004) 27124–27137.
[6] H.H. Yang, Y. Hu, K.H. Buetow, M.P. Lee, A computational approach to measuring coherence of gene expression in pathways, Genomics 84 (2004) 211–217.
[7] E.J. Williams, D.J. Bowles, Coexpression of neighboring genes in the genome of Arabidopsis thaliana, Genome Res. 14 (2004) 1060–1067.
[8] H. Caron, et al., Evidence for two tumor suppressor loci on chromosomal bands 1p35–36 involved in neuroblastoma: one probably imprinted, another associated with N-myc amplification, Hum. Mol. Genet. 4 (1995) 535–539.
[9] M.J. Lercher, A.O. Urrutia, L.D. Hurst, Clustering of housekeeping genes provides a unified model of gene order in the human genome, Nat. Genet. 31 (2002) 180–183.
[10] B.A. Cohen, R.D. Mitra, J.D. Hughes, G.M. Church, A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression, Nat. Genet. 26 (2000) 183–186.
[11] A.M. Boutanaev, A.I. Kalmykova, Y.Y. Shevelyov, D.I. Nurminsky, Large clusters of co-expressed genes in the Drosophila genome, Nature (London) 420 (2002) 666–669.
[12] P.T. Spellman, G.M. Rubin, Evidence for large domains of similarly expressed genes in the Drosophila genome, J. Biol. 1 (2002) 5.
[13] M.J. Lercher, T. Blumenthal, L.D. Hurst, Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes, Genome Res. 13 (2003) 238–243.
[14] H. Ge, Z. Liu, G.M. Church, M. Vidal, Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae, Nat. Genet. 29 (2001) 482–486.

[15] L.M. Staudt, P.O. Brown, Genomic views of the immune system, Annu. Rev. Immunol. 18 (2000) 829–859.
[16] A. Grigoriev, A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae, Nucleic Acids Res. 29 (2001) 3513–3519.
[17] R. Jansen, D. Greenbaum, M. Gerstein, Relating whole-genome expression data with protein–protein interactions, Genome Res. 12 (2002) 37–46.
[18] R. Miki, et al., Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays, Proc. Natl. Acad. Sci. USA 98 (2001) 2199–2204.
[19] S. Hannenhalli, S. Levy, Transcriptional regulation of protein complexes and biological pathways, Mamm. Genome 14 (2003) 611–619.
[20] M. Ptashne, A. Gann, Imposing specificity by localization: mechanism and evolvability, Curr. Biol. 8 (1998) R812–R822.
[21] C.A. Klein, Gene expression signatures, cancer cell evolution and metastatic progression, Cell Cycle 3 (2004) 29–31.
[22] P.A. Covitz, Class struggle: expression profiling and categorizing cancer, Pharmacogenom. J. 3 (2003) 257–260.
[23] D. Hanahan, R.A. Weinberg, The hallmarks of cancer, Cell 100 (2000) 57–70.
[24] U. Rennefahrt, M. Janakiraman, R. Ollinger, J. Troppmair, Stress kinase signaling in cancer: fact or fiction? Cancer Lett. 217 (2005) 1–9.
[25] O.J. Halvorsen, et al., Gene expression profiles in prostate cancer: association with patient subgroups and tumour differentiation, Int. J. Oncol. 26 (2005) 329–336.
[26] T. Kohonen, Self-Organizing Maps, Springer-Verlag, Berlin, 1995.
[27] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA 95 (1998) 14863–14868.