JOURNAL OF BIOPHARMACEUTICAL STATISTICS
Vol. 14, No. 3, pp. 701–721, 2004

Detection of Activity Centers in Cellular Pathways Using Transcript Profiling

Joel Pradines,[1,*] Laura Rudolph-Owen,[2] John Hunter,[2] Patrick Leroy,[2] Michael Cary,[1] Robert Coopersmith,[1] Vlado Dancik,[1] Yelena Eltsefon,[1] Victor Farutin,[1] Christophe Leroy,[1] Jonathan Rees,[1] David Rose,[1] Steve Rowley,[1] Alan Ruttenberg,[1] Patrick Wieghardt,[1] Chris Sander,[1,3] and Christian Reich[1]

[1]Department of Computational Sciences, Millennium Pharmaceuticals, Inc., Cambridge, Massachusetts, USA
[2]Department of Molecular and Cellular Oncology, Millennium Pharmaceuticals, Inc., Cambridge, Massachusetts, USA
[3]Bauer Center for Genomics Research, Cambridge, Massachusetts, USA

ABSTRACT

We present a new computational method for identifying regulated pathway components in transcript profiling (TP) experiments by evaluating transcriptional activity in the context of known biological pathways. We construct a graph representing thousands of protein functional relationships by integrating knowledge from public databases and review articles. We use the notion of distance in a graph to define pathway neighborhoods. The pathways perturbed in an experiment are then identified as the subgraph induced by the genes, referred to as activity centers, having significant density of transcriptional activity in their functional neighborhoods. We illustrate the predictive power of this approach by performing and analyzing an experiment of TP53 overexpression in

*Correspondence: Joel Pradines, Department of Computational Sciences, Millennium Pharmaceuticals, Inc., 40 Lansdowne St., Cambridge, MA 02139, USA; E-mail: joel.pradines@mpi.com.

DOI: 10.1081/BIP-200025678
1054-3406 (Print); 1520-5711 (Online)

NCI-H125 cells. The detected activity centers are in agreement with the known TP53 activation effects and our independent experimental results. We also apply the method to a serum starvation experiment using HEY cells and investigate the predicted activity of the transcription factor MYC. Finally, we discuss interesting properties of the activity center approach and its possible applications beyond the comparison of two experiments.

Key Words: Microarray; Gene expression; Pathway; Protein network; Monte carlo method.

Mathematics Subject Classification: 65C05; 65C60; 91C15.

1. INTRODUCTION

The identification of the intracellular network components that are active or perturbed in a given biological experiment can be a difficult problem, e.g., in experiments designed to uncover the function of a novel gene or the effects of a drug. Transcript profiling using DNA microarrays (TP) is a powerful method to perform such investigations. In a single assay, a snapshot of the transcriptional activity of the entire cell can be obtained (DeRisi et al., 1997). However, TP provides only raw material for interpretation: translating hundreds of observed mRNA level changes in terms of perturbed pathways remains a tedious manual task.

Here, we present a method for the analysis of TP data using prior knowledge about biological pathways. This knowledge, derived from public information sources, is represented by a graph of functional relationships between genes. The analysis of TP experiments is performed by identifying regions in the graph that have significant density of transcriptional activity.

Figure 1 summarizes the different steps of our approach. We first designed a database schema capable of integrating both metabolic and regulatory processes and loaded the corresponding database with metabolic networks from the LIGAND database (Goto et al., 1998), regulatory pathways from the Cell Signaling Networks Database (CSNDB) (Takai-Igarashi and Kaminuma, 1999) and data from review articles. From this database we extracted a graph of gene-to-gene functional relationships. The adjacency of two genes in this network represents the knowledge of a relatively direct molecular interaction between their products, such as protein binding or phosphorylation. After mapping TP data from several experiments onto the pathway graph, we address the question of whether this mapping adds any information value. We show that this is the case for several pairs of experiments by demonstrating a significant correlation between adjacency of genes in the network and their differential expression, referred to as transcriptional activity. Next, to identify the regulated pathways in a particular pair of experiments we compute for each gene a weighted average of the transcriptional activity in its graph neighborhood. We estimate a p-value for this average by using a model of random variable based on the observed distribution of gene activities. The perturbed pathways are then defined as the subgraph induced by the genes, called activity centers, having p-values below an arbitrary threshold.

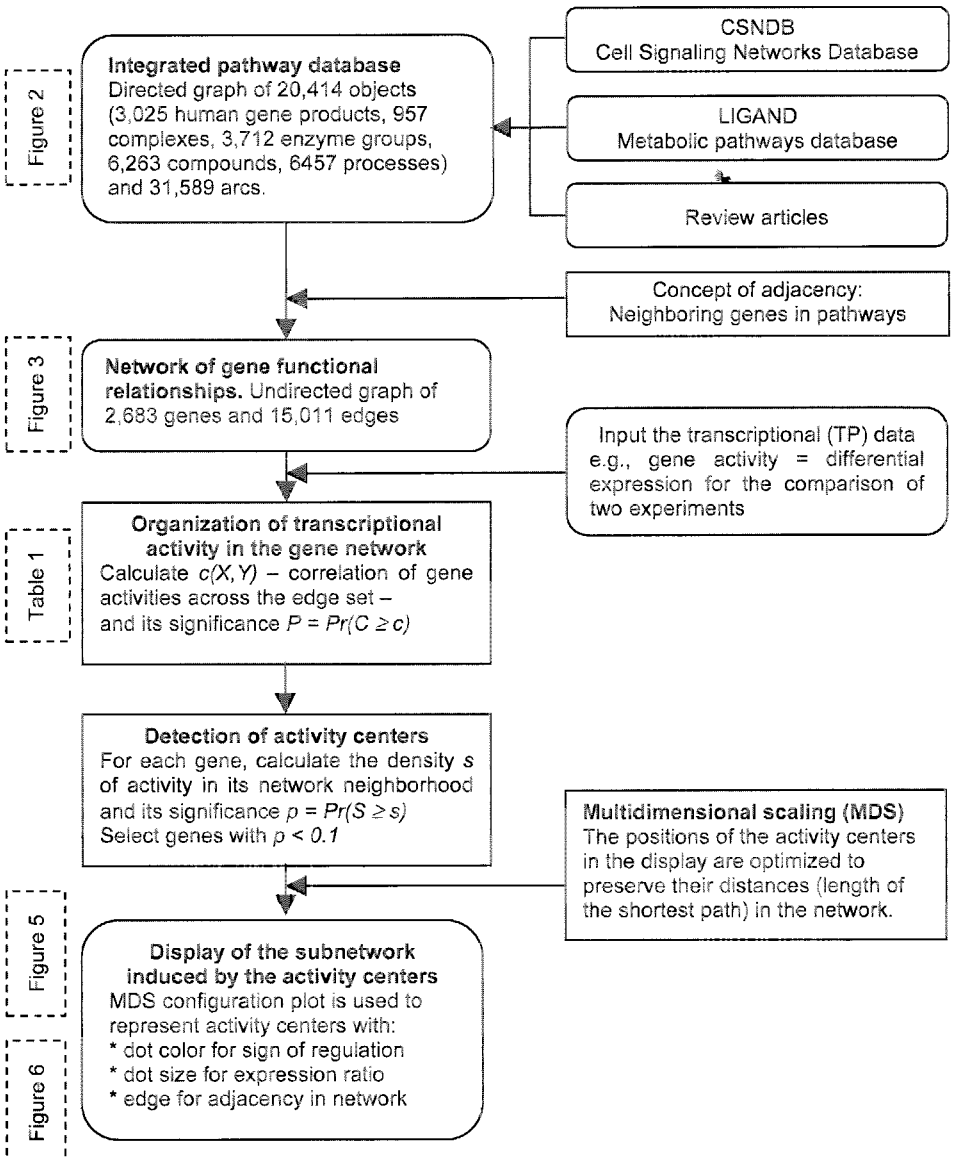

Figure 1. Flowchart of the activity center analysis.

The effectiveness of this approach is first tested by performing and analyzing a TP experiment in which the tumor suppressor TP53 is over-expressed. Identified pathway components are in agreement with the known TP53 activation effects and those seen in our independent experimental confirmation, i.e., cell cycle arrest and apoptosis. We next use our method to probe for perturbed pathway components in a serum starvation experiment. A prominent activity center, the transcription

factor MYC, is revealed as particularly interesting, in that the transcriptional response suggests its phosphorylation. Finally, we discuss the properties of the activity center approach and explain how the same algorithm can be applied to situations other than the comparison of two experiments.

2. INTEGRATION OF PATHWAY DATABASES

The pathway database we used for this study was derived from three sources: the Cell Signaling Networks Database (CSNDB) (Takai-Igarashi and Kaminuma, 1999), the metabolic pathways database LIGAND (Goto et al., 1998) and manual curation of published articles. After being parsed, CSNDB and LIGAND were manually curated. In particular, ubiquitous compounds such as water and ATP were removed.

All the data were integrated into one schema partially illustrated by the graph in Fig. 2. Knowledge is encoded using a directed graph having several types of vertices and arcs. Both biological entities (e.g., gene products) and processes (e.g., protein binding) are encoded as vertices. Arcs indicate the function of vertices with respect

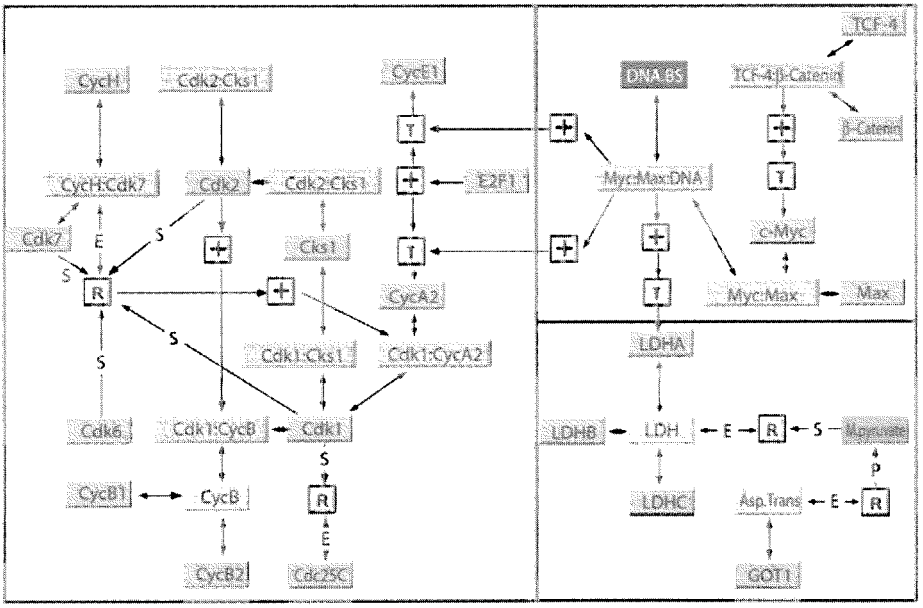

Figure 2. Example of knowledge representation in the pathway database. Nodes represent both biological entities and processes. Pink is used for gene products, orange for complexes, blue for DNA binding sites, green for metabolites and yellow for groups of enzymes catalyzing a same reaction. Processes are displayed as white boxes with a symbol specifying their type. R stands for biochemical reaction, T for transcription, + for activation, and – for inhibition. Arcs indicate the function of nodes with respect to one another. When appropriate, a letter on the arc specifies the type of function. E stands for enzyme, S for substrate, and P for product.

to each other. For instance, a vertex of type "metabolic reaction" is linked to its metabolites by arcs of types "substrate" and "product." Figure 2 shows a few examples of database entries. The left part of the figure shows a portion of cell cycle pathways. CCNH and CDK7 bind to form a complex which phosphorylates CDC2 (Cdk1), CDK2, CDK6 and CDK7 (Kohn, 1999). In the right upper part of Fig. 2, the accumulation of CTNNB1 (beta-Catenin) leads to its binding to the T-cell factor 4 and over-expression of MYC (He et al., 1998). MYC binds to MAX (Blackwood et al., 1992). The resulting complex binds to the consensus sequence CACGTG (Blackwell et al., 1993) and can activate the transcription of LDHA (lactate dehydrogenase A), CCNA2 (Cyclin A2) and CCNE1 (Cyclin E1) (Dang, 1999). The right lower part of Fig. 2 shows a portion of metabolic pathways. Lactate dehydrogenases catalyze the conversion of mercaptopyruvate into mercaptolactate, the latter being produced by a reaction catalyzed by GOT1 (glutamic oxaloacetic transaminase 1) (Goto et al., 1998).

The full database used for this paper contains 20,261 objects representing 3025 human gene products, 957 complexes between proteins and other molecules, 3712 enzyme groups, 6263 compounds (metabolites and small molecules) and 6457 processes (reaction, transcription, activation, inhibition and degradation). These objects are connected to each other trough 31,589 arcs.

3. NETWORK OF GENE FUNCTIONAL RELATIONSHIPS

In order to analyze TP experiments we used a gene-centric representation of the knowledge contained in the pathway database. Namely, from the directed graph of the pathway database we extracted an undirected graph of gene-to-gene functional relationships.

Two genes are adjacent in this graph if there exists in the pathway database a directed path between them of length less than or equal to six and not containing any other gene vertex. The limit length of six was chosen to avoid linking enzymes that are more than one reaction apart in metabolic pathways. The algorithm to derive the gene network is the following:

(a) Initialize a list L of vertices with a gene vertex v_i.
(b) Retrieve a list L' of vertices by traversing the directed arcs adjacent to the members of L. Only arcs and vertices that have not yet been visited can be used.
(c) Terminate if L' is empty.
(d) If a member of L' is a gene vertex v_j, save the edge (v_i, v_j) and remove v_j from L'.
(e) Terminate if step (b) has been performed six times.
(f) Replace L with L' and go to (b).

Running the algorithm for all possible gene seeds gives the adjacency list of an undirected graph that represents functional relationships between genes.

Adjacency in this gene network can be the result of diverse biological features, including: protein binding, protein phosphorylation, catalysis of a same reaction, catalysis of a metabolic reaction involving a common metabolite and transcriptional regulation. Figure 3 shows the gene network obtained when using only the portion of the pathway database displayed in Fig. 2. The gene-to-gene network obtained from the full pathway database has 2683 genes and 15,011 edges between them.

4. ORGANIZATION OF TRANSCRIPTIONAL ACTIVITY IN THE GENE NETWORK

We now present a way of testing whether the gene network is likely to add information to the interpretation of the differential expression observed between two TP experiments. We consider two experimental paradigms: the re-introduction of the tumor suppressor TP53 gene in a cell line lacking TP53 expression and the stimulation of a starved cell line with fresh serum. In these two paradigms we chose six pairs of TP experiments: cells expressing TP53 vs. control cells and pairs of successive time points in the serum addition experiment (1 hr/0 hr, 3 hr/1 hr, 6 hr/3 hr, 9 hr/6 hr and 12 hr/9 hr). Because they represent perturbations, these pairs are more likely to yield differential expression focused on a few pathways in contrast to, for instance, the comparison of two cell lines.

For a given pair of experiments we define the transcriptional activity a_i of a gene g_i by

$$a_I = |\ln(\max(I_{i1}, 0.1)) - \ln(\max(I_{i2}, 0.1))|$$

where I_{i1} and I_{i2} are the hybridization intensities, after background noise subtraction, for the cDNA representing gene g_i in the two experiments. The threshold of 0.1 is

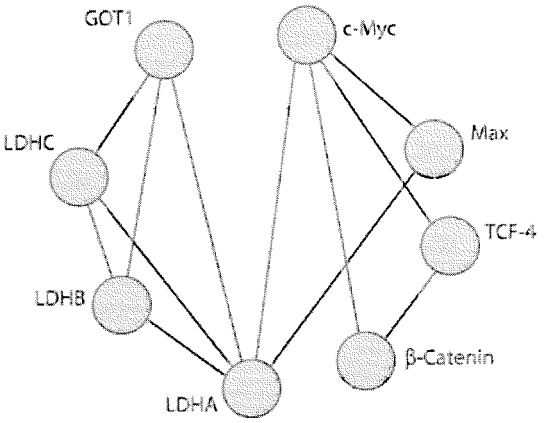

Figure 3. Network of gene functional relationships extracted from a portion of the integrated pathway database (right part of Fig. 2).

used to reduce the noise associated with low intensities. The logarithmic transform stabilizes, to some extent, the variance of the noise (Rocke and Durbin, 2003). The absolute value is used for symmetry of up and down-regulation. Let $E = \{g_i g_j\}$ be the edge set of the network reduced to the genes having a representative cDNA on the array. For each edge $g_i g_j$ we create two ordered pairs of gene transcriptional activities: (a_i, a_j), (a_j, a_i). We call X the vector whose components are the first elements of the ordered pairs obtained across E and Y the vector of second elements. To quantify the similarity of transcriptional activities for genes adjacent in the network we use the following measure of correlation

$$c(X, Y) = \frac{\overline{XY} - (\overline{X})(\overline{Y})}{\sqrt{\overline{X^2} - (\overline{X})^2}\sqrt{\overline{Y^2} - (\overline{Y})^2}}$$

where $\overline{X}$ stands for the average of the components of X. A value of c significantly greater than 0 indicates that transcriptional activity is organized inside the network. To make such a statement about significance we need to define a random variable C and estimate

$$P = \Pr(C \geq c).$$

The model we chose is the following: perform a random permutation of the gene activities, build the vectors X and Y, and compute c. The value of P is estimated with 10^5 Monte Carlo simulations. The reshuffling needs to be performed at the level of the genes rather than the components of X and Y. Indeed, depending on their degree in the network (number of neighbors), genes have different relative contributions to the value of c. The random variable C should reproduce such variation resulting from the network topology.

Table 1 shows the estimated values of P for the six pairs of experiments. The smallest p-values were obtained for the TP53 experiment and the comparison of time points 9 hr and 6 hr in the serum addition experiment. This suggests that the gene network is most likely to add information value to the interpretation of these two pair-wise comparisons.

Table 1. Estimated p-values for the correlations of differential expression across the adjacency list of the gene network. Small values indicate a significant organization of regulation in the network.

Pair of experiments	Estimation of $P = \Pr(C \geq c)$
TP53/control	0.00016
Serum addition 1 hr/0 hr	0.14551
Serum addition 3 hr/1 hr	0.77970
Serum addition 6 hr/3 hr	0.03197
Serum addition 9 hr/6 hr	0.00573
Serum addition 12 hr/9 hr	0.14698

5. ALGORITHM FOR THE DETECTION OF ACTIVITY CENTERS

We showed that for some pairs of experiments differential expression is concentrated in particular regions of the network. We now present an algorithm to identify these regions. Our approach can be summarized as follows:

(1) Compute for each gene g_i the average transcriptional activity s_i in its close network neighborhood.
(2) Given the observed distribution of gene activities associate a p-value p_i to s_i. Genes having p_i below a threshold value (e.g., 0.1) are referred to as "activity centers."
(3) The regulated pathways are defined by the subgraph induced by the activity centers.

In other words, perturbed pathways are identified as the portions of the network in which genes have a significantly high local density of transcriptional activity.

Such a local approach is motivated by the particular structure of the gene network. Our network contains 72,949 triangles. For comparison, 10^4 simulations of Erdös–Rényi random graphs (Albert and Barabasi, 2002) having the same order (2683 vertices) and approximately the same size ($\sim 15,000$ edges) give an average number of 233 triangles per graph with a standard deviation of 16. The large number of triangles in our gene network is a feature common to many biological networks (Watts and Strogatz, 1998). It reflects the presence of functional modules (Newman et al., 2001). For instance, when the pathway database contains a protein complex of size greater than 2, all pairs of subunits are linked in the gene network, creating triangles. In other words, our network has a strong local structure with many graph neighborhoods of high cliquishness coefficient representing functional modules.

We therefore chose to quantify the density s_i of transcriptional activity around a gene g_i using its closed neighborhood H_i, i.e., the union of g_i and its adjacent genes

$$s_i = |H_i'|^{-1} \sum_{g_j \in H_i'} \alpha_{ij} a_j$$

H_i' is the closed neighborhood reduced to those genes having a representative cDNA on the array. $|H|$ stands for the cardinality of the set H. The α coefficients weight the contribution of a neighbor to the score of the center gene. We use

$$\alpha_{ij} = |H_i \cap H_j| / |H_i \cup H_j|$$

as this measure of neighborhood overlap is higher for genes belonging to a same functional module. For instance, if g_i is part of a protein complex and binds to TP53, the relative contribution of another subunit to its score is larger than that of the hub TP53.

We next want to select a subset of genes having the highest densities of transcriptional activity. We cannot directly use s_i to order genes as its variance depends on the

topology of a gene neighborhood (size and α_{ij} coefficients). Instead, we call A the following random variable based on the observed distribution of gene activities: "choose one value of transcriptional activity, all network genes represented on the array having equal probability of being selected." We then define the random variable

$$S_i = |H_i'|^{-1} \sum_{g_j \in H_i'} \alpha_{ij} A_j$$

where the A_j are independent realizations of A. For each pair of experiments we use 10^5 Monte Carlo simulations to estimate

$$p_i = \Pr(S_i > s_i)$$

p_i is a probability conditional to the neighborhood structure and is therefore more appropriate than s_i to select genes. All the genes having a p_i value below an arbitrary threshold (e.g., 0.1) are called "activity centers" and the regulated pathways are approximated by the subgraph they induce in the network.

6. DISPLAY OF THE ACTIVITY CENTERS

To display the subgraph induced by the activity centers we use multidimensional scaling (MDS) (Borg and Groenen, 1997). The idea is to restitute the notion of functional proximity between activity centers in the layout of the induced subgraph. Let $\Delta = [\delta_{ij}]$ be the matrix of distances between activity centers in the full network. Distance is defined here as the length of the shortest path between two genes. If two genes belong to distinct connected components of the full network we set their distance to $\delta_{ij} = 10$. Let $D = [d_{ij}]$ be the matrix of Euclidean distances between activity centers in the graph display. We optimize the gene positions by minimizing the following stress function (Basalaj, 2001).

$$S(\Delta, D) = \sum_{i>j} (d_{ij} - \delta_{ij})^2 / \delta_{ij}^2.$$

Minimizing $S(\Delta, D)$ implies that small distances will be more accurately preserved than large ones. Indeed, as mentioned above, our network has a strong local structure reflecting the existence of functional modules. We would like to restitute this information in the display. To minimize $S(\Delta, D)$ we use the SMACOF algorithm (Scaling by Maximizing a Convex Function). A detailed description of SMACOF can be found in Borg (1997). Briefly, SMACOF is inspired from a general optimization technique termed Iterative Majorization (IM). IM minimizes a function by replacing it locally with a convex function and guarantees a monotonic decrease of the stress during the iterations. SMACOF does not guarantee convergence towards a global minimum (Basalaj, 2001; Borg and Groenen, 1997). However, as can be seen in Fig. 5, the first local minimum encountered by SMACOF restitutes well small distances between activity centers in their display.

7. APPLICATION 1: REGULATED PATHWAYS IN TP53-INDUCED CELL CYCLE ARREST AND APOPTOSIS

The gene for the tumor repressor TP53 was introduced by retroviral infection into NCIH125 cells, a lung adenosquamous carcinoma cell line that lacks endogenous TP53 expression (Fig. 4A). Expression of TP53 in this cell line induces growth arrest and programmed cell death as determined by propidium iodide staining, trypan blue dye exclusion and TUNEL (Fig. 4B and data not shown). We measured the effects of TP53 expression in these cells by performing microarray hybridization with RNA at 5 days post-infection. This time point was chosen because there is a clear phenotypic effect of TP53 expression at this stage, with enough surviving cells to provide sufficient RNA for transcriptional analysis. By day 6 post-infection approximately 90% of the cells infected with LNCX-TP53 score as apoptotic by TUNEL analysis (data not shown). Control cells containing the expression vector but not carrying the TP53 gene were also profiled in the same conditions.

The results of the activity center analysis comparing cells expressing TP53 and control cells are displayed in Fig. 5. Genes with $p_i \leq 0.1$ are displayed as squares. A line between two genes represents their adjacency in the network. The gene transcriptional activities are encoded by the color and size of the squares; green (grey) means higher expression in cells expressing TP53 compared to the control, and red

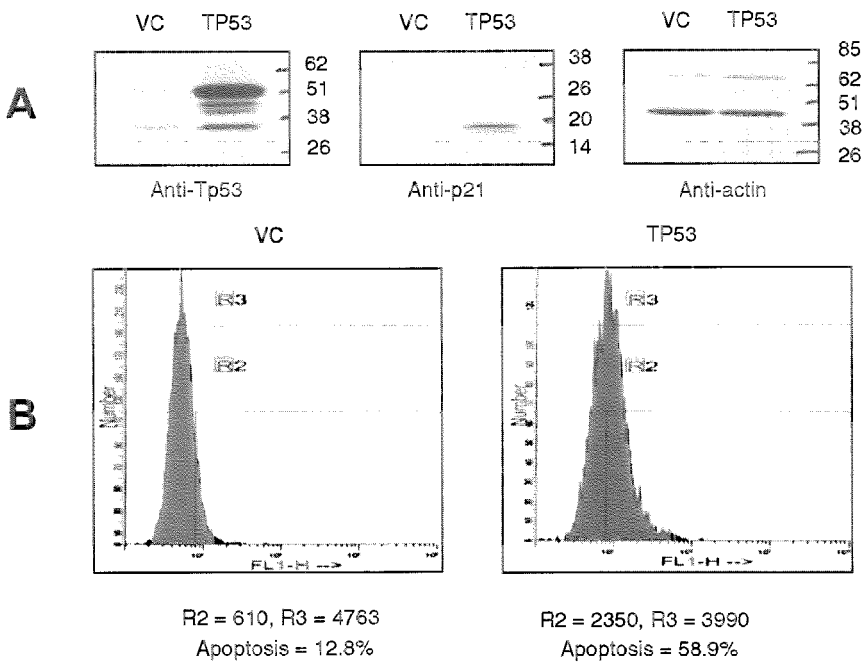

Figure 4. Expression of TP53 in NCI-H125 cells induces programmed cell death, (A) Expression of TP53 and p21 (CDKN1A) in infected cells (VC = LNCX, TP53 = LNCXTP53). An anti-actin antibody was used as a control and (B) Five days after infection NCI-H125 cells expressing TP53 underwent apoptosis at a much higher rate than the vector control cells.

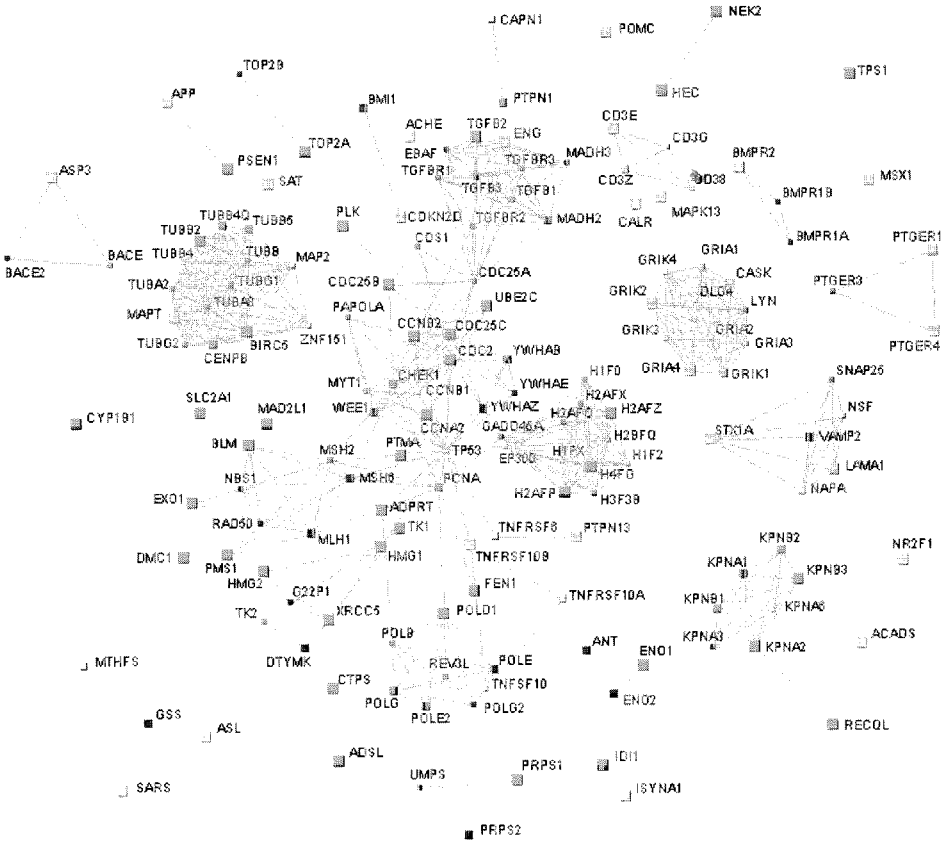

Figure 5. Activity centers ($p \leq 0.1$) detected upon TP53 expression. Genes are represented as squares. Green (grey) means a mRNA level higher in TP53 expressing cells compared to the control and red (black) the opposite. The size of the square indicates the amplitude of the mRNA level ratio (r): small for $1 < r < 1.5$ or $1/1.5 < r < 1$, medium for $1.5 < r < 2$ or $1/2 < r < 1/1.5$, and large for $r > 2$ or $r < 1/2$. Lines between genes indicate their adjacency in the network of functional relationships.

(black) lower expression. At the center of the figure is TP53. The detected activity centers are associated with intracellular processes and structures such as DNA repair and transcription, nuclear import, apoptosis, kinetochore, nucleosome, production of ribonucleotides, DNA synthesis, vesicle fusion and transport, membrane adhesion, TGFB signaling and cell cycle checkpoints. The biological processes are in agreement with the phenotypic effects of TP53 expression in the NCI-H125 cell line experiment (Fig. 4) as well as with the widely known effects of TP53 activation: cell cycle arrest and initiated apoptosis (Flatt and Pietenpol, 2000; Levine, 1997).

Above TP53 are several proteins involved in the two cell cycle checkpoints of the G1-S and G2-M transitions. TP53 is indeed known for its ability to stop cell growth by arresting cell cycle at the G1-S transition by inducing the expression of *CDKN1A* (p21) (Flatt and Pietenpol, 2000; Levine, 1997). This results in the inhibition of the

phosphorylation of RB1 by CDK4/CCND1 and CDK2/CCNE and prevents the E2Fs from transactivating CDK1 and CCNA2, which are required to enter the S phase. TP53 also induces cell cycle arrest by a similar mechanism, combined with its transactivation of 14-3-3 sigma at the G2-M checkpoint (Chan et al., 1999; Flatt and Pietenpol, 2000). Our TP data are compatible with this model as CDK1 and CCNA2 appear to be significantly down-regulated. Other proteins involved in the cell cycle checkpoints and whose genes have been reported to be regulated by TP53 are CCNB2, CDC25C (Badie et al., 2000; Zhao et al., 2000), WEE1 (Leach et al., 1998) and the 14-3-3 proteins (Hermeking et al., 1997) (YWHAB, YWHAE, and YWHAZ). More proteins involved in cell cycle progression and present on the figure are PLK (Kohn, 1999), HEC (Chen et al., 1997) and PCNA (Kohn, 1999). In addition, UBE2C, which has been shown to promote the entry into mitosis by inducing ubiquitin-dependent proteolysis of cyclins A and B (Townsley et al., 1997), is significantly down-regulated. Another potential indicator of cycle arrest is the upregulation of the CDK inhibitor CDKN2B (Brotherton et al., 1998) also known to prevent the inhibition of TP53 by MDM2 (Weber et al., 1999).

Other proteins related to growth arrest and present in the figure are enzymes at the junction between pyrimidine and purine metabolism. Among those are PRPS1 and PRPS2, which catalyze a limiting reaction for the production of nucleotides (Phang, 1985). The down-regulation of their genes is compatible with a G1-S arrest (Eriksson et al., 1984). Additional metabolic enzyme genes involved in the production of nucleotides and down-regulated in TP53 expressing cells are TK1, DTYMK, TK2, UMPS, CTPS and ADSL. Also suggestive of growth arrest is the down-regulation of the DNA polymerase subunits and their binding activator ADPRT (Simbulan et al., 1993). ADPRT is also involved in the regulation of TP53 activity (Wesierska-Gadek and Schmid, 2000). Another potential indicator of growth arrest is the significant concentration of down-regulation observed in the nucleosome (histone genes). In addition, most of the tubulin genes whose products form the microtubules are down-regulated and some of their binding partners (MAPT, BIRC5 and CENPB) are differentially expressed. One can hypothesize that these activity centers relate to the arrest of TP53 expressing cells as they should no longer synthesize mitotic spindles.

Underneath TP53 are HMG1, HMG2 and their binding partners G22P1 (Ku70) and XRCC5 (Ku80). These proteins are involved in DNA repair pathways (Kohn, 1999). Recently it was shown that TP53 binds to HMG1 (Imamura et al., 2001). Our data show down-regulation of HMG1 and HMG2. The latter repression has previously been reported to be induced by TP53 (Zhao et al., 2000). Other proteins having their genes significantly down-regulated and involved in DNA repair pathways are MSH2, MSH6, MLH1, PMS1 (Matton et al., 2000), NBS1 (Zhong et al., 1999) and EXO1 (Amin et al., 2001). These activity centers suggest modification in DNA repair functions related to TP53. Two other activity centers related to DNA repair are the DNA helicases RECQL and BLM. A recent publication shows that their genes tend to be up-regulated in cells that are transformed or actively proliferating (Kawabe et al., 2000). The down-regulation observed here in arresting cells is consistent with this finding.

Finally, the activity centers TNFRSF10A (DR4), TNFRSF10B (DR5), TNFRSF6 (Fas), TNFS10 (Apo2) and BIRC5 are involved in apoptotic pathways

activated by TP53 (Flatt and Pietenpol, 2000). The execution phase of apoptosis is characterized by specific cellular changes such as membrane blebbing, which are usually difficult to detect because they occur asynchronously in a cell population (Mills et al., 1998). We can hypothesize that the activity centers around VAMP2 (STX1A, NAPA, LAMA1 NSF and SNAP25) could be the result of such changes as they represent processes related to vesicle fusion and membrane adhesion.

8. APPLICATION 2: ACTIVITY CENTERS IN A SERUM STARVATION EXPERIMENT

We next applied the detection of activity centers to a system not as well characterized as the TP53 induced growth arrest. Ovarian cancer cells were starved for 24 hr and fresh serum was added. RNA hybridization was performed at time 0 hr (just before serum addition), 1 hr, 3 hr, 6 hr, 9 hr and 12 hr. Table 1 indicates that transcriptional activity is most significantly organized inside the network when comparing the mRNA levels at 9 hr to those at 6 hr.

Figure 6 shows the activity centers ($p_i \leq 0.1$) obtained when using these two time points. At the center of the figure are two groups: the proteasome and functional neighbors of the transcription factor MYC. Most of the proteasome components are up-regulated and have p-values ranging from 2.10^{-3} to 6.10^{-3}. MYC and its binding partner MAX also have significant neighborhood scores with respective values $p = 1.4 \times 10^{-3}$ and $p = 2.10^{-3}$. Eighteen members of the MYC functional neighborhood are detected as activity centers. Those represent binding partners of MAX (NCOR1, MXI1 (Alland et al., 1997), an inhibitor of MYC mediated transactivation (TFAP2B (Gaubatz et al., 1995)), and reporter genes of MYC (PTMA, HDAC1, CCNA2, CCND1, EIF4E, LDHA, MRDB, ODC1, SERPINE1, CHC1 (Dang, 1999)).

As shown in Fig. 7, the expression of MYC decreases from 3 hr to 6 hr, and increases from 6 hr to 9 hr. In addition, several reporter genes of MYC such as PTMA, CCND1, EIF4E, CCNA2, SERPINE1, ODC1, MRDB and TP53 (Dang, 1999) follow the same pattern from 3 hr to 9 hr and have twofold or more regulation. Since MYC and its functional neighbors were the activity centers with the most significant scores, we tested whether this might reflect the activity of MYC by carrying out western blot experiments with anti-MYC and antiphospho-MYC antibodies recognizing phosphorylation at Thr58 and Ser62. The results are presented in Fig. 7. The MYC protein appears only after serum addition and does not show any significant variation after 1 hr. MYC is phosphorylated at 6 hr and remains phosphorylated until 12 hr. The appearance of the MYC protein at 1 hr without any change of the mRNA level might be due to posttranslational regulation. It is interesting to note that the appearance of MYC at 1 hr does not affect either the mRNA level of its targets nor its neighborhood activity score. This indicates that even though MYC is present from 1 hr to 3 hr it might not be involved in activating gene expression. Later, between 3 hr and 6 hr, MYC targets are down-regulated. As stated above, MYC might not be activating gene transcription during the first 3 hr. Hence, we cannot know whether the observed transcriptional repression is caused by direct inhibition of MYC or by another protein positioned at the MYC DNA

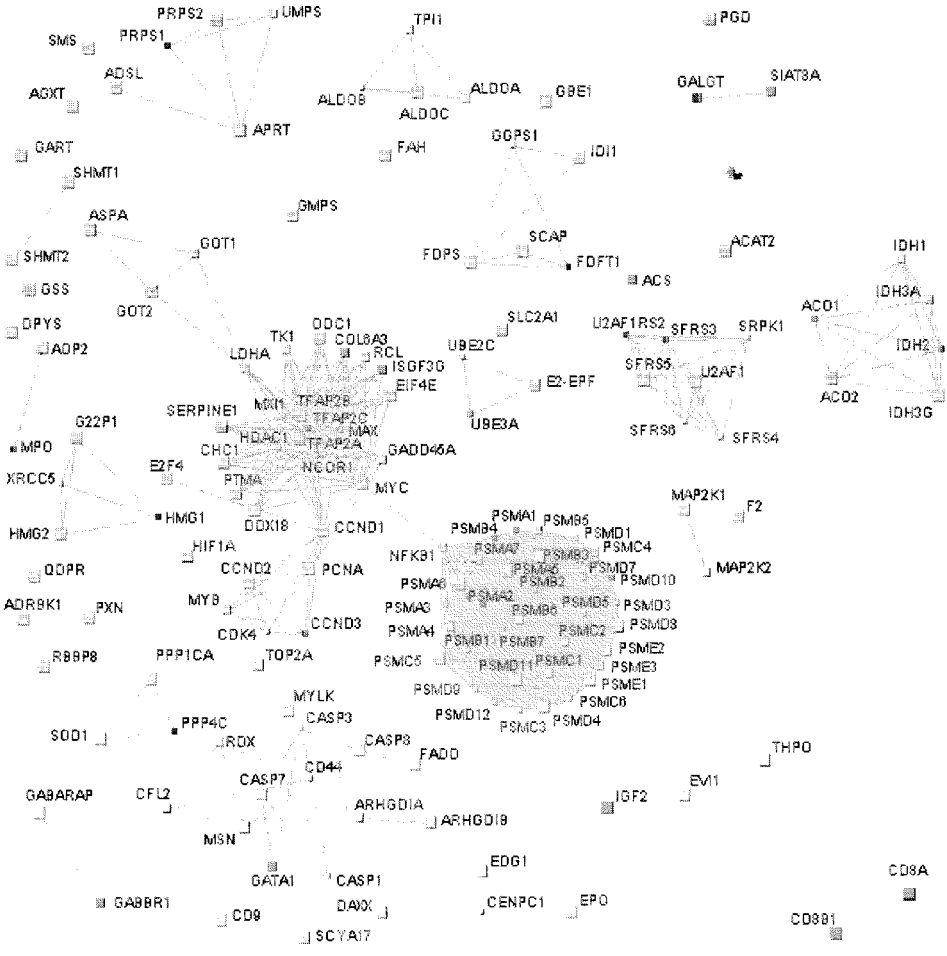

Figure 6. Activity centers ($p \leq 0.1$) for the serum starvation experiment, when comparing transcriptional regulation at 9 h vs. 6 h. Same legends as Fig. 4.

binding site. The phosphorylation of MYC between 3 hr and 6 hr is followed by the up-regulation of its targets from 6 hr to 9 hr. This could be caused by MYC as its phosphorylation was reported to enhance its ability to transactivate (Gupta et al., 1993).

Therefore, our experimental results are in agreement with a possible activity of MYC as indicated by the significant density of transcriptional activity in its neighborhood.

9. DISCUSSION

We have presented a computational method using a pathway database to facilitate the interpretation of TP experiments. Besides the advantage of helping

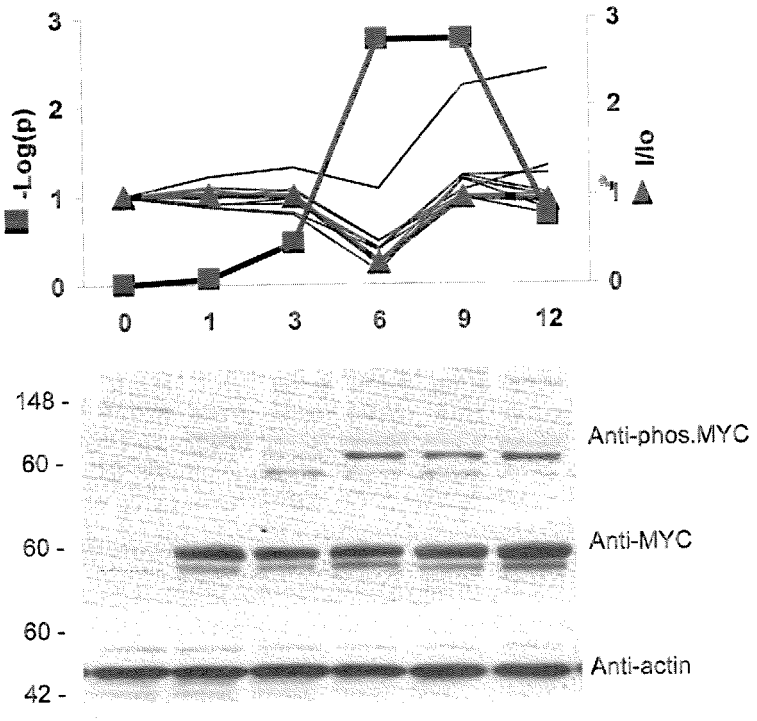

Figure 7. Time course of the neighborhood activity significance p(■) and the relative mRNA level (▲) for MYC in the serum addition experiment. Thin lines represent the mRNA hybridization intensities of genes known to be transactivated by MYC (PTMA, CCND1, EIF4E, CCNA2, SERPINE1, ODC1, MRDB and TP53). Results of western analysis with anti-MYC, anti-phospho-MYC and anti-actin antibodies are shown for each time point.

the scientist deal with large amounts of knowledge, our approach has several interesting properties.

First, a simple undirected graph of gene functional relationships can be quite powerful to analyze TP experiments, provided one takes advantage of its topology. The network we used for this paper breaks down into 173 connected components, with one giant component representing 78% of the genes. One could therefore argue that a graph representation is suboptimal as more than 2000 genes are lumped into "one big pathway". However, this giant component has a strong local structure, as attested by its large number of triangles. Cliques, or neighborhoods of high cliquishness, represent functional modules such as protein complexes. We have used this local structure in our algorithm by quantifying regulation at the level of the graph neighborhood and taking into account how much adjacent genes share their neighbors. In addition MDS and distances between genes in the network can restitute the notion of pathway. In Fig. 5, based on the positions of the genes and their edges, we can clearly delineate TGFB signaling, cell cycle checkpoints (around CDC2), DNA repair (RAD50 and G22P1) and DNA synthesis (POLD1). The pathway database

does not contain any object that explicitly represents these cellular processes. Yet, MDS has graphically restituted this knowledge encoded in the network by the distances between genes. Recent publications have shown that besides their high clique content biological networks have other remarkable topological properties such as hierarchical modularity (Ravasz et al., 2002) and dissasortative mixing (Newman, 2002). It could be interesting to exploit these other structural features to analyze TP experiments.

We have introduced a measure of correlation between gene activities across the adjacency list of a graph. This measure enabled us to show that differential expression can be significantly organized in the pathway network. The simultaneous transcriptional regulation of pathway neighbors is not a new finding. Previous publications have demonstrated that interacting proteins (Ge et al., 2001; Grigoriev, 2001; Jansen et al., 2002) and enzymes close in metabolic pathways (Miki et al., 2001) tend to significantly correlate their expression profiles across large TP data sets. However, our definition of correlation is orthogonal to that used in these works. Our measure is designed to quantify the organization of transcriptional activity inside a graph. Organization in the graph, i.e., low entropy, suggests a potential added information value of the network for the interpretation of the TP experiment.

One key aspect of the activity center method is the notion of local density of transcriptional activity. Genes are scored for the average activity in their neighborhood. A tempting simplification is to select genes based on their individual expression. Such uncoupling of the experimental data and the prior knowledge is not desirable. Genes have different abilities to be regulated at the mRNA level. The average over the neighborhood partially smoothes these differences. For instance, a transcription factor might become active upon phosphorylation. Even if its mRNA level does not change, the differential expression of its target genes represented in its neighborhood is an indication of its potential activity, as shown in this paper with MYC (Fig. 7).

The p-value computed for an activity center is conditional to an observed distribution of differential expression. Even if a TP data set does not contain any strongly regulated gene, activity centers can be detected provided differential expression significantly accumulates in their neighborhoods. For instance, the genes of the proteasome subunits do not in general have strong variation of their mRNA levels. However, their coordinated differential expression can be detected by our algorithm. Because the significance p of an activity center is conditional to the observed distribution of transcriptional activity, its numerical value should not be interpreted as the probability of no differential expression in a functional neighborhood. Such value should be instead derived by using replicate experiments (same mRNA sample hybridized several times). One could for instance compute the observed values of s_i with the pair of experiments of interest and then estimate $p_i = \Pr(S_i' > s_i)$ where the random variable S' is built with the distribution of gene activity obtained with the replicates.

In this paper we have restricted the definition of transcriptional activity to the differential expression of a gene. Nevertheless, the detection of activity centers is not at all dependent on this choice. Genes are selected for the significance of the average activity in their neighborhood. The p-values are estimated by reshuffling

the individual gene activities. The same approach can be applied with any definition of gene activity. One can for instance choose the activity to be the correlation of the expression profiles of genes to a profile of interest, and identify pathways matching this profile across a large set of experiments. Our method can also be used to compare groups of samples, such as patient samples in a clinical trial. If the groups are large enough, a statistical score can be assigned to each gene based on its inter and intra-group variances. These scores can be used as input to the activity center method.

APPENDIX A: TP53 EXPERIMENT

TP53 cDNA in the LNCX retroviral expression plasmid (Clontech) was delivered into NCI-H125 lung carcinoma cells plated at 2.5×10^5 in 10 cm dishes by overnight retroviral infection. Virus containing the LNCX vector alone was used as a control. Cells were selected for two days immediately after infection in medium containing 1 mg/ml Genetecin to remove uninfected cells, and harvested for RNA at day 5 postinfection. Portions of the infected cells were set aside for western blot and TUNEL analysis. Cell Culture and RNA preparation: NCI-H125 cells (ATCC) were grown in RPMI medium (GIBCO) with 10% fetal calf serum and 2 mM L-Glutamine. Cells were maintained at 37°C in 5% CO_2. RNA was prepared using Qiagen RNeasy kits as per the manufacturer's instructions. Western analysis: Nitrocellulose filters were probed with antibody in PBS + 0.2% Tween-20 with 5% skim milk to block nonspecific protein interactions. The antibodies used for western analysis were mouse anti-TP53 (DO-1, Santa Cruz Biotechnology); mouse anti-CDKN1A (F-5, Santa Cruz Biotechnology); and goat anti-actin (I-9, Santa Cruz Biotechnology). TUNEL assays: Cells were removed from 10 cM dishes with trypsin treatment, washed 2× with PBS, fixed with 70% ethanol and stored at 4°C until the time of use. Cells were pelleted and resuspended in 100 μL of permeabilisation solution (0.1% Triton X-100 in 0.1% sodium citrate) for 2 min on ice. Cells were then washed 2× with PBS and resuspended in 50 μL TUNEL reaction mixture. As a negative control untreated H125 cells were resuspended in 50 μL of Label solution (no TdT). The cells were incubated for 60 min at 37°C in a humidified atmosphere in the dark. Cells were then washed 2× with PBS and resuspended in 500 μL PBS, and analyzed on a FACS Caliber machine. The negative control (H125 cells in Label solution alone) was used to define the background due to cell autofluorescence and establish the gate settings.

APPENDIX B: SERUM STARVATION EXPERIMENT

The HEY ovarian cancer cell line (Buick et al., 1985) was routinely grown in RPMI 1640 media supplemented with 10% FBS and 100 mg/mL L-glutamine (GIBCO BRL, New York) and maintained at 37°C in 5% CO_2. For experimental conditions, cells were starved for 24 hr in RMPI plus 0.1% FBS. Thereafter, complete medium (10% serum) was added back at time points 0 hr. For all time points (0 hr, 1 hr, 3 hr, 6 hr, 9 hr and 12 hr), cells were washed two times in 1× PBS, RNA was

then prepared using Qiagen RNeasy kits as per the manufacturer's instructions, or cells were scraped and lysed in protein lysis buffer (1% Triton X-100; 50 mM Tris, pH 8.0; 100 mM NaCl; 20 mM Hepes; 1× Protease Inhibitor (Calbiochem, San Diego)). Insoluble material was removed by centrifugation and the supernatant collected. 25 mg of protein supernatant was electrophoresed and transferred to a nitrocellulose membrane. Membranes were blocked in TBST (150 mM NaCl; 10 mM Tris, pH 8.0; and 0.05% Tween 20) containing either 2.5% skim dried milk or 5% BSA. Membranes were subsequently probed with either rabbit polyclonal antiphospho-MYC antibody (Cell Signaling, Beverly, MA), mouse monoclonal anti-MYC antibody (NeoMarkers, Fremont, CA), mouse monoclonal anti-Actin antibody (Santa Cruz, Santa Cruz, CA). Blots were washed in TBST and incubated for 1 hr with anti-rabbit or anti-mouse IgG secondary antibody conjugated to horseradish peroxidase (Santa Cruz), washed again and developed using ECL Plus Western Blotting Reagent (Amersham Pharmacia, Piscataway, NJ).

APPENDIX C: CDNA ARRAY CONSTRUCTION AND HYBRIDIZATION

For transcriptional profiling, we used nylon-based array technology that was developed in-house (for details see Chiang et al., 2001). Briefly, bacterial cultures of individual EST clones were used from libraries or public sources. To prepare templates for array elements, clones were amplified by PCR, ethanol precipitated, suspended in a sodium citrate buffer and arrayed onto nylon filters at a density of $\sim 64/cm^2$. After drying, the arrayed DNA was denatured in 0.4 M sodium hydroxide, neutralized and dried to completion.

Total RNA from cultured cells was prepared and 1 μg mRNA labeled in the presence of oligo(dT)30, dATP, dGTP, and dTTP, and 33P-dCTP using SuperScript™ II (Gibco) and purified over Chroma Spin™ + TE-30 columns. The labeled cDNA was annealed at 65°C for 1 hr in the presence of poly(dA) > 200 and human Cot 1 DNA. Following overnight-hybridization at 65°C in a rotating incubator, the array filters were washed, dried and exposed to phosphoimage screens for 60 hr. The radioactive hybridization signals were captured and digitized to give an intensity value. All array hybridizations were performed in duplicates. Intensities were normalized with the median of the array. For each cDNA an intensity value was obtained by averaging over the replicates.

ACKNOWLEDGMENTS

We thank J. Bolen, M. Rolfe, S. Letovsky and G. Hinkle for their support. We are grateful to Dr. Takai-Igarashi and Dr. Kaminuma for providing us with the data of CSNDB, and Dr. Kanehisa for allowing us to use the LIGAND database. D. Honig helped with the figures. In addition, we thank two anonymous referees for providing comments that greatly improved the manuscript.

REFERENCES

Albert, R., Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys.* 74(47):47–97.

Alland, L., Muhle, R., Hou, H. Jr., Potes, J., Chin, L., Schreiber-Agus, N., DePinho, R. A. (1997). Role for N-CoR and histone deacetylase in Sin3-mediated transcriptional repression. *Nature* 387(6628):49–55.

Amin, N. S., Nguyen, M. N., Oh, S., Kolodner, R. D. (2001). Exo1-Dependent mutator mutations: Model system for studying functional interactions in mismatch repair. *Mol. Cell. Biol.* 21(15):5142–5155.

Badie, C., Itzhaki, J. E., Sullivan, M. J., Carpenter, A. J., Porter, A. C. (2000). Repression of CDK1 and other genes with CDE and CHR promoter elements during DNA damage-induced G2/M arrest in human cells. *Mol. Cell. Biol.* 20(7):2358–2366.

Basalaj, W. (2001). Proximity Visualization of Abstract Data, Technical Report 509, University of Cambridge Computer Laboratory, http://www.pavis.org/essay/index.html.

Blackwell, T. K., Huang, J., Ma, A., Kretzner, L., Alt, F. W., Eisenman, R. N., Weintraub, H. (1993). Binding of myc proteins to canonical and noncanonical DNA sequences. *Mol. Cell. Biol.* 13(9):5216–5224.

Blackwood, E. M., Kretzner, L., Eisenman, R. N. (1992). Myc and Max function as a nucleoprotein complex. *Curr. Opin. Genet. Dev.* 2(2):227–235.

Borg, I., Groenen, P. (1997). *Modern Multidimensional Scaling: Theory and Applications*. Springer Verlag Series in Statistics. New York: Springer.

Brotherton, D. H., Dhanaraj, V., Wick, S., Brizuela, L., Domaille, P. J., Volyanik, E., Xu, X., Parisini, E., Smith, B. O., Archer, S. J., Serrano, M., Brenner, S. L., Blundell, T. L., Laue, E. D. (1998). Crystal structure of the complex of the cyclin D-dependent kinase Cdk6 bound to the cell-cycle inhibitor p19INK4d. *Nature* 395(6699):244–250.

Buick, R. N., Pullano, R., Trent, J. M. (1985). Comparative properties of five human ovarian adenocarcinoma cell lines. *Cancer Res.* 45(8):3668–3676.

Chan, T. A., Hermeking, H., Lengauer, C., Kinzler, K. W., Vogelstein, B. (1999). 14-3-3 sigma is required to prevent mitotic catastrophe after DNA damage. *Nature* 401(6753):616–620.

Chen, Y., Sharp, Z. D., Lee, W. H. (1997). HEC binds to the seventh regulatory subunit of the 26 S proteasome and modulates the proteolysis of mitotic cyclins. *J. Biol. Chem.* 272(28):24081–24087.

Chiang, L. W., Grenier, J. M., Ettwiller, L., Jenkins, L. P., Ficenec, D., Martin, J., Jin, F., DiStefano, P. S., Wood, A. (2001). An orchestrated gene expression component of neuronal programmed cell death revealed by cDNA array analysis. *Proc. Natl. Acad. Sci. U.S.A.* 98(5):2814–2819.

Dang, C. V. (1999). c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Mol. Cell. Biol.* 19(1):1–11.

DeRisi, J. L., Iyer, V. R., Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278(5338):680–686.

Eriksson, S., Graslund, A., Skog, S., Thelander, L., Tribukait, B. (1984). Cell cycledependent regulation of mammalian ribonucleotide reductase.

The S phase-correlated increase in subunit M2 is regulated by de novo protein synthesis. *J. Biol. Chem.* 259(19):11695–11700.

Flatt, P. M., Pietenpol, J. A. (2000). Mechanisms of cell-cycle checkpoints: At the crossroads of carcinogenesis and drug discovery. *Drug. Metab. Rev.* 32(3-4):283–305.

Gaubatz, S., Imhof, A., Dosch, R., Werner, O., Mitchell, P., Buettner, R., Eilers, M. (1995). Transcriptional activation by Myc is under negative control by the transcription factor AP-2. *EMBO J.* 14(7):1508–1519.

Ge, H., Liu, Z., Church, G. M., Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat. Genet.* 29(4):482–486.

Goto, S., Nishioka, T., Kanehisa, M. (1998). LIGAND: Chemical database for enzyme reactions. *Bioinformatics* 14(7):591–599.

Grigoriev, A. (2001). 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(17): 3513–3519.

Gupta, S., Seth, A., Davis, R. J. (1993). Transactivation of gene expression by Myc is inhibited by mutation at the phosphorylation sites Thr-58 and Ser-62. *Proc. Natl. Acad. Sci. U.S.A.* 90(8):3216–3220.

He, T. C., Sparks, A. B., Rago, C., Hermeking, H., Zawel, L., da Costa, L. T., Morin, P. J., Vogelstein, B., Kinzler, K. W. (1998). Identification of c-MYC as a target of the APC pathway. *Science* 281(5382):1509–1512.

Hermeking, H., Lengauer, C., Polyak, K., He, T. C., Zhang, L., Thiagalingam, S., Kinzler, K. W., Vogelstein, B. (1997). 14-3-3 sigma is a p53-regulated inhibitor of G2/M progression. *Mol. Cell.* 1(1):3–11.

Imamura, T., Izumi, H., Nagatani, G., Ise, T., Nomoto, M., Iwamoto, Y., Kohno, K. (2001). Interaction with p53 enhances binding of cisplatin-modified DNA by High Mobility Group 1 protein. *J. Biol. Chem.* 276(10):7534–7540.

Jansen, R., Greenbaum, D., Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Res.* 12(1):37–46.

Kawabe, T., Tsuyama, N., Kitao, S., Nishikawa, K., Shimamoto, A., Shiratori, M., Matsumoto, T., Anno, K., Sato, T., Mitsui, Y., Seki, M., Enomoto, T., Goto, M., Ellis, N. A., Ide, T., Furuichi, Y., Sugimoto, M. (2000). Differential regulation of human RecQ family helicases in cell transformation and cell cycle. *Oncogene* 19(41):4764–4772.

Kohn, K. W. (1999). Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell* 10(8):2703–2734.

Leach, S. D., Scatena, C. D., Keefer, C. J., Goodman, H. A., Song, S. Y., Yang, L., Pietenpol, J. A. (1998). Negative regulation of Wee1 expression and Cdc2 phosphorylation during p53-mediated growth arrest and apoptosis. *Cancer Res.* 58(15):3231–3236.

Levine, A. J. (1997). p53, the cellular gatekeeper for growth and division. *Cell* 88(3):323–333.

Matton, N., Simonetti, J., Williams, K. (2000). Identification of mismatch repair protein complexes in HeLa nuclear extracts and their interaction with heteroduplex DNA. *J. Biol. Chem.* 275(23):17808–17813.

Miki, R., Kadota, K., Bono, H., Mizuno, Y., Tomaru, Y., Carninci, P., Itoh, M., ShibataK., Kawai, J., Konno, H., Watanabe, S., Sato, K., Tokusumi, Y., Kikuchi, N., Ishii, Y., Hamaguchi, Y., Nishizuka, I., Goto, H., Nitanda, H., Satomi, S., YoshikiA., Kusakabe, M., DeRisi, J. L., Eisen, M. B., Iyer, V. R., Brown, P. O., Muramatsu, M., Shimada, H., Okazaki, Y., Hayashizaki, Y. (2001). 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. U.S.A.* 98(5):2199–2204.

Mills, J. C., Stone, N. L., Erhardt, J., Pittman, R. N. (1998). Apoptotic membrane blebbing is regulated by myosin light chain phosphorylation. *J. Cell. Biol.* 140(3):627–636.

Newman, M. E. (2002). Assortative mixing in networks. *Phys. Rev. Lett.* 89(20):208701.

Newman, M. E., Strogatz, S. H., Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* 64(2-2):026118.

Phang, J. M. (1985). The regulatory functions of proline and pyrroline-5-carboxylic acid. *Curr. Top. Cell. Regul.* 25:91–132.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297(5586):1551–1555.

Rocke, D. M., Durbin, B. (2003). Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* 19(8):966–972.

Simbulan, C. M., Suzuki, M., Izuta, S., Sakurai, T., Savoysky, E., Kojima, K., Miyahara, K., Shizuta, Y., Yoshida, S. (1993). Poly(ADP-ribose) polymerase stimulates DNA polymerase alpha by physical association. *J. Biol. Chem.* 268(1):93–99.

Takai-Igarashi, T., Kaminuma, T. (1999). A pathway finding system for the cell signaling networks database. *In Silico Biol.* 1(3):129–146.

Townsley, F. M., Aristarkhov, A., Beck, S., Hershko, A., Ruderman, J. V. (1997). Dominant-negative cyclin-selective ubiquitin carrier protein E2-C/UbcH10 blocks cells in metaphase. *Proc. Natl. Acad. Sci. U.S.A.* 94(6):2362–2367.

Watts, D. J., Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature* 393(6684):440–442.

Weber, J. D., Taylor, L. J., Roussel, M. F., Sherr, C. J., Bar-Sagi, D. (1999). Nucleolar Arf sequesters Mdm2 and activates p53. *Nat. Cell. Biol.* 1(1):20–26.

Wesierska-Gadek, J., Schmid, G. (2000). Overexpressed poly(ADP-ribose) polymerase delays the release of rat cells from p53-mediated G1 checkpoint. *J. Cell. Biochem.* 80(1):85–103.

Zhao, R., Gish, K., Murphy, M., Yin, Y., Notterman, D., Hoffman, W. H., Tom, E., Mack, D. H., Levine, A. J. (2000). Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. *Genes Dev.* 14(8):981–993.

Zhong, Q., Chen, C. F., Li, S., Chen, Y., Wang, C. C., Xiao, J., Chen, P. L., Sharp, Z. D., Lee, W. H. (1999). Association of BRCA1 with the hRad50-hMre11-p95 complex and the DNA damage response. *Science* 285:747–750.

Received January 2003
Accepted September 2003