BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–8

Deciphering Drug Action and Escape Pathways: An Example on Nasopharyngeal Carcinoma

Difeng Dong [1,*], Chun-Ying Cui [2,*], Benjamin Mow [3], Limsoon Wong [1]

[1] National University of Singapore, Singapore
[2] Capital Medical University, China
[3] West Clinic Excellence Cancer Center, Singapore
[*] These two authors contributed equally to this study.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Nasopharyngeal carcinoma (NPC) is a malignant cancer in the head and neck region, with especially high incidence in South China, Southeastern Asia and North Africa. Recently, a cyclin dependent kinase (CDK) inhibitor, CYC202, is studied for its anti-tumor effect in human NPC cells *in vitro* and *in vivo*. Results show that both cell lines and patients in the study responded to the drug treatment differently. To further investigate the drug response, expression of selected genes for apoptosis, cell proliferation and cell cycle regulation were measured during the process of treatment. Our issue is how to identify the reason for the different responses in these NPC individuals using the gene expression data.

Results: Biological pathway information has long been incorporated into gene expression analysis for the purpose of treatment response understanding. However, the conclusions are usually too general, and hardly sufficient for guiding further research. In our current study, we design a drug pathway identification system, the Drug Pathway Decipherer, which identifies genetic regulations in response to drug treatment that are consistent with respect to a given detailed signaling pathway structure. By applying our system to the NPC dataset, we discover that the status of ERK pathway and apoptosis pathway are differently regulated between responders and non-responders both *in vitro* and *in vivo*. Our results indicate that the dysregulation of Ras-ERK pathway and PI3K-Akt-NFκB pathway are probably the mechanisms for CYC202-insensitive NPC cells to resist the drug treatment.

Availability: The Drug Pathway Decipherer is available at http://www.comp.nus.edu.sg/∼wongls/projects/drug-pathway/DPD-v1. It is implemented in JAVA.

Contact: dongdife@comp.nus.edu.sg, ccy@ccmu.edu.cn, bmow@westexcellence.com, and wongls@comp.nus.edu.sg

INTRODUCTION

NPC is a malignant cancer in the head and neck region, with especially high incidence in South China, Southeastern Asia and North Africa (Yu and Yuan, 2002). Despite the high rates of local tumor control with the technique of intensity-modulated radiotherapy (RT), NPC patients suffer from a high ratio of distant metastasis (Sultanem *et al*., 2000; Lee *et al*., 2002). Therefore, new chemotherapy is necessary to improve the treatment outcome of RT. In our recent research, CYC202 (Cyclacel Ltd, Dundee, United Kingdom; Seliciclib; R-roscovitine), a CDK inhibitor, is studied for its anti-tumor effect on NPC cells *in vitro* and *in vivo*. 3 NPC cell lines and 13 NPC patients were treated with CYC202, and the expression of selected genes were measured during the process of treatment. Results show that both cell lines and patients in the study responded to the drug treatment differently. Our target is to identify the reason underlying the different responses in these NPC cells and patients.

There are past works that incorporate biological pathways into gene expression analysis to understand drug treatment response. Some of them focus on the enrichment analysis of gene groups on pathways (Zeeberg *et al*., 2003; Doniger *et al*., 2003; Subramanian *et al*., 2005; Sivachenko *et al*., 2005, 2007). Zeeberg *et al*. (2003) and Doniger *et al*. (2003) use the hypergeometric test to determine statistically over-represented biological pathways in a given list of differentially expressed genes. Subramanian *et al*. (2005) propose the gene set enrichment analysis (GSEA), which uses a weighted Kolmogorov-Smirnov statistics to compare the two sets of distributions and also uses resampling to estimate false discovery rates (FDR). Sivachenko *et al*. (2005, 2007) split genes into separate regulatory groups, each sharing the same transcriptional regulators, and evaluate these gene groups in a GSEA-like manner.

Other research groups concentrate on statistically significant pathway search with the list of differentially expressed genes (Sohler *et al*., 2004; Scott *et al*., 2005; Cabusora *et al*., 2005; Nacu *et al*., 2007). Since this problem is NP hard (Ideker *et al*., 2002), various heuristics are used. Sohler *et al*. (2004) expand the seed genes by iteratively including the most significant neighbor, with respect to Fisher's inverse χ^2 statistics (Fisher, 1932). Cabusora *et al*. (2005) use Dijkstra's algorithm (Dijkstra, 1959) to search for the shortest path between each pair of the seed genes. Scott *et al*. (2005) reduce the pathway search into the node-weighted Steiner tree problem, *viz.*, to find the minimal set of edges to connect nodes reaching the maximal weight, and tackle it with graph theory.

More related works identify responsive molecular pathways under drug treatment (Zien *et al*., 2000; Ideker *et al*., 2002; Hanisch *et al*., 2002; Guo *et al*., 2007; Breitling *et al*., 2004). Hanisch *et al*. (2002) cluster genes with a metric preferring both genetic co-expression and short distance within a network topology. Zien *et al*. (2000)

exhaustively enumerate all possible gene combinations on a metabolic pathway, and identify the most co-expressed gene group as the responsive pathway. Ideker *et al*. (2002) extend the method of Zien *et al*. (2000) to a protein-protein interaction (PPI) network, and use an annealed random selection to generate candidate gene groups for the co-expression evaluation. Guo *et al*. (2007) follow Ideker *et al*. (2002), but their random selection is based on interaction between genes rather than directly on gene itself.

However, most existing works fall short on several issues (Soh *et al*., 2007): these works provide little information on the interplay between selected genes; the collection of pathways that can be used, evaluated and ranked against the observed expression data is limited; and the generated hypotheses are still too general to guide further research. So we have two aims in our current study: to propose effective computational methods for treatment response understanding, and to interpret drug treatment response for the NPC study. In this paper, we present a drug pathway identification system, which we called Drug Pathway Decipherer, to identify consistent genetic regulations in response to drug treatment according to some specified detailed signaling pathway structure. The status of the specified signaling pathways are estimated and compared with respect to the identified drug pathways. We show how to apply the system to the NPC dataset, and to use the results for further analysis. In addition, our system allows users to construct, remove, and modify biological pathways for their own research purposes.

METHODS

Overview

The Drug Pathway Decipherer consists of 4 partitions distributed on two biological levels. Figure 1 gives the diagram of its workflow. It takes signaling pathways and gene expression datasets as input. To enable signaling pathways to be evaluated against gene expression data, for each pathway, the system extracts genetic relationships from the pathway data source, and passes them to the genetic level to perform a genetic pathway search. The derived genetic pathways then constitute a candidate set for hypotheses of drug pathway. For each candidate, transformed gene expression data are used to evaluate the correlation of expression regulations between genes on the pathway against the pathway structure. After that, statistical procedures are applied to select pathways with significant evaluations from the candidate set as drug pathway hypotheses. Finally, these hypothesized pathways are aggregated to estimate the status of the corresponding signaling pathway.

Data source

The NPC gene expression data comprise one dataset for cell lines and another for patients, with both containing 380 genes selected for apoptosis, cell proliferation, and cell cycle regulation. For the *in vitro* part, 3 cell lines, CNE1, CNE2 and HK1 were measured for their gene expression before treatment, and 2hs, 4hs, 6hs, 12hs and 24hs after treatment respectively. It was observed that CNE1 responded poorly; CNE2 responded in a limited way; and HK1 fully responded. For the *in vivo* part, 12 tumor samples and 1 non-tumor sample were taken from the NPC patients, who were traced for their response to treatment. Gene expression were measured before and after the treatment. 5 patients were reported to have a molecular response.

With respect to the selected genes, 4 signaling pathways are extracted from KEGG pathway database (October 17, 2007) (Kanehisa *et al*., 2002): ERK pathway (from hsa04010), JNK/p38 pathway (from hsa04010), the G1/S cell cycle progression pathway (from hsa04110) and the apoptosis pathway (from hsa04210). These pathways are represented as directed graphs, with nodes denoting proteins and edges denoting PPIs. Figure 2 shows the modeled pathways in our study.

Preprocessing data source

In order to capture gene expression change in response to drug treatment, the original gene expression data are transformed into the relative expression (RE) values. RE values describe expression change in multiples in a linear scale. A positive RE value suggests a gene is up regulated, and a negative value suggests a transcriptional suppression. RE value which is defined as:

DEFINITION 1. *Given a time-course gene expression dataset E, its corresponding RE dataset is R, where e_{ij} and r_{ij} are the original expression value and RE value of gene i at time point j, respectively. If $e_{ij} \geqslant e_{i0}$, then $r_{ij} = e_{ij}/e_{i0} - 1$; otherwise, $r_{ij} = 1 - e_{i0}/e_{ij}$.*

Signaling pathways are represented as directed graphs in our system, which can be formally described as:

DEFINITION 2. *A signaling pathway γ is a directed graph (P, I), with P the vertex set, representing the collection of proteins on pathway, and I the edge set, representing the collection of interactions between proteins. An interaction is a triplet $i = \langle p_1, p_2, s\rangle$, with $p_1, p_2 \in P$ and $s \in S$, where $S = \{\$stimulation, \$suppression\}$ is the set of terms used to denote interaction types.*

The terminology set S can be enriched with other terms to describe the type of interactions between proteins. This requires a corresponding interpretation to genetic relationships, which is introduced in the next part. In the current system, a signaling pathway is preprocessed into a list of interactions, with each protein associated with its encoding genes.

Extracting genetic relationships

The procedure of genetic relationship extraction passes pathway information from the proteomic level to the genetic level, allowing signaling pathways to be evaluated against gene expression data. Assuming G is a gene set, and $T = \{\$positive, \$negative\}$, is an associated terminology set used to describe relations between genes in G, the genetic relationship is defined as:

DEFINITION 3. *A genetic relationship (or simply a relationship) is a triplet $q = \langle g_1, g_2, t\rangle$, with $g_1, g_2 \in G$ and $t \in T$.*

The extraction of genetic relationships from a signaling pathway is:

DEFINITION 4. *Given a signaling pathway $\gamma = (P, I)$ and a relationship set $Q \subseteq G \times G \times T$, a relationship extraction is a set of functions φ, ψ, ϕ, with $\varphi : P \rightarrow G$, a one-to-many mapping from a protein to a set of genes, $\psi : S \rightarrow T$, a mapping between two terminology sets on two different biological levels, and $\phi : I \rightarrow Q$,*

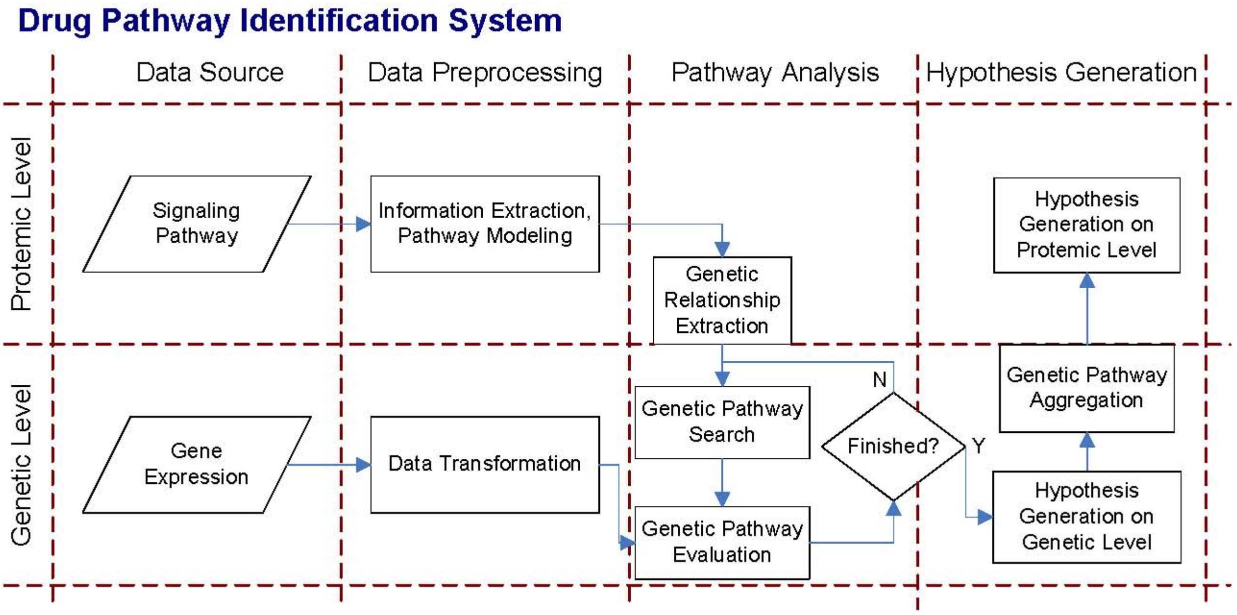

Drug Pathway Identification System
Data Source
Data Preprocessing
Pathway Analysis
Hypothesis Generation
Protemic Level
Genetic Level
Signaling Pathway
Information Extraction, Pathway Modeling
Genetic Relationship Extraction
Hypothesis Generation on Protemic Level
Genetic Pathway Aggregation
N
Genetic Pathway Search
Finished?
Y
Gene Expression
Data Transformation
Genetic Pathway Evaluation
Hypothesis Generation on Genetic Level

Fig. 1. The workflow of the drug pathway identification system.

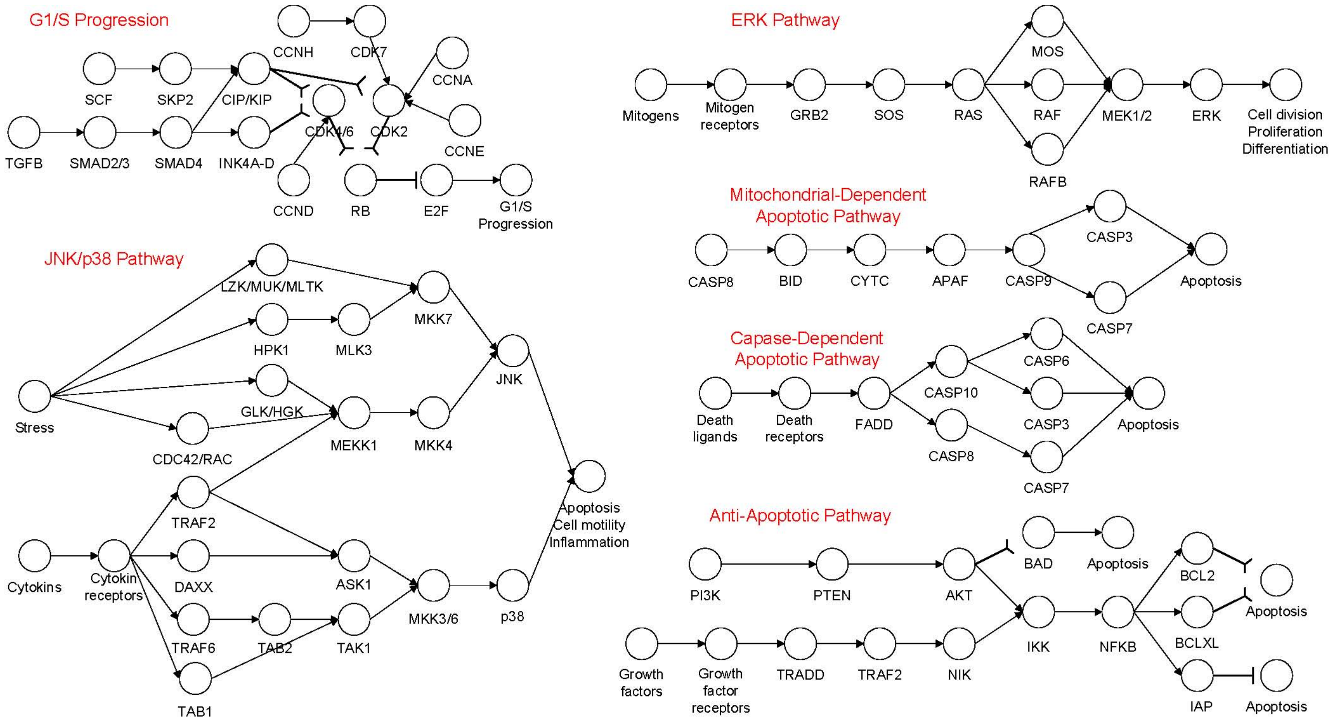


Fig. 2. The modeled signaling pathways. The downstream events of pathways are represented as virtual nodes. "→" and "⊣" represent "stimulation" and "suppression", respectively.

a mapping from an interaction to multiple relationships respecting φ and ψ.

In our implementation, φ associates each protein with its encoding genes; ψ maps $\$stimulation$ and $\$suppression$ to $\$positive$ and $\$negative$, respectively; and ϕ replaces the proteins and the type of an interaction with exhaustive combinations of the mapped genes and relation. In particular, the interpretation of a relationship is different from that of an interaction. An interaction describes a real process in a biological system. It forms a deductive logic for the occurrence of downstream events. However, a relationship is only an evidence for the occurrence of an interaction. For example, gene MAP2K7 and MAPK8 are the encoding genes of protein MKK7 and JNK, respectively. MKK7 stimulates JNK. If MKK7 is activated by upstream events, then JNK will be activated as well; if the expression of both MAP2K7 and MAPK8 are up regulated, then probably, the interaction between MKK7 and JNK is largely carried out, and JNK is activated.

Scoring a genetic pathway

A genetic pathway can be intuitively understood as a map of a signaling pathway on the genetic level, which is formally defined as:

DEFINITION 5. *Given a relationship $q = \langle g_1, g_2, t\rangle$, if there does not exist a relationship $q' = \langle g_1', g_2', t'\rangle$, with $g_1 = g_2'$, then q is called a source relationship.*

DEFINITION 6. *Given a relationship $q = \langle g_1, g_2, t\rangle$, if there does not exist a relationship $q' = \langle g_1', g_2', t'\rangle$, with $g_2 = g_1'$, then q is called a sink relationship.*

DEFINITION 7. *Given two relationships $q = \langle g_1, g_2, t\rangle$ and $q' = \langle g_1', g_2', t'\rangle$, if $g_2 = g_1'$, then q is said to immediately precede q', denoted by $q \prec_i q'$.*

DEFINITION 8. *Given a relationship arrangement $A = \langle q_1, q_2, \ldots, q_n\rangle$, if there exists a permutation π, $\pi(A) = \langle q_1', q_2', \ldots, q_n'\rangle$, satisfying $q_1' \prec_i q_2' \prec_i \cdots \prec_i q_n'$, with q_1' a source relationship and q_n' a sink relationship, then $\vartheta = \langle q_1', q_2', \ldots, q_n'\rangle$ is called a genetic pathway and A is called a genetic pathway seed.*

Given a relationship $q = \langle g_1, g_2, t\rangle$, if the expression of g_1 and g_2 are measured at multiple time points (as our *in vitro* dataset), then the correlation of q is:

$$Corr(q) = Corr(\overrightarrow{r_{g_1}}, \overrightarrow{r_{g_2}}),$$

where $Corr(\overrightarrow{r_{g_1}}, \overrightarrow{r_{g_2}})$ is the Pearson correlation coefficient between RE vector $\overrightarrow{r_{g_1}}$ and $\overrightarrow{r_{g_2}}$. If gene expression are only measured at two time points (as our *in vivo* dataset), then the correlation is estimated simply by comparing post-treatment RE values of the two genes:

$$Corr(q) = \frac{sgn(r_{g_1}^{post}) \times sgn(r_{g_2}^{post}) \times \min_{i=1,2} |r_{g_i}^{post}|}{\max_{j=1,2} |r_{g_j}^{post}|}.$$

The derived correlation is then transformed into a z-score, $z(q)$, evaluated against the sample background of correlation.

To produce an aggregated z-score, z_ϑ, for an entire pathway ϑ with k relationships, $z(q)$ are summed up over all relationships in ϑ, with respect to the relation of q:

$$z(\vartheta) = \frac{1}{\sqrt{k}} \sum_{q \in \vartheta} (-1)^\alpha z(q),$$

where $\alpha = 0$ if $q.relation = \$positive$; $\alpha = 1$ if $q.relation = \$negative$. This score function takes pathway structure into consideration. Genes are expected to exhibit co-regulation patterns consistent with the relations between them. For example, if two genes have a negative relation, then we expect their REs are negatively correlated as well.

For each pathway ϑ with k relationships, we randomly select 10000 gene groups of $k+1$ size (the gene number of ϑ is $k+1$) to estimate the p-value of $z(\vartheta)$, denoted by $score(\vartheta)$. Intuitively, the pathway score represents the consistency between a genetic pathway structure and the expression regulations of genes on it.

Searching for computable genetic pathways

Given the definition of genetic pathway, the procedure of pathway discovery is trivial. However, not all genes on a pathway is observable. In our system, two relationships sharing an unobservable gene are combined into one relationship, connected with the relative relation defined as:

DEFINITION 9. *Given two relationships $q_1 = \langle g_1, g_2, t_1\rangle$ and $q_2 = \langle g_2, g_3, t_2\rangle$ such that $q_1 \prec_i q_2$. If $t_1 = t_2$, the relative relation between g_1 and g_3 is $\$positive$; otherwise, $\$negative$.*

Unobservable genes on a pathway can be bypassed by recursively invoking this procedure, which finally forms valid input for the scoring mechanism.

Generating hypotheses

Genetic pathway hypotheses are generated for each signaling pathway. p-value and FDR cutoff are used to control the statistical significance and the rate of false positive of the generated hypotheses. Since the pathway score itself is a p-value measurement, the procedure of significance control is straight forward. For FDR, we first rank the scores of pathways which pass the p-value filtering. Then, we identify the maximal rank index j, satisfying

$$p_j < \frac{j \cdot \alpha}{C_N \cdot N},$$

where p_j is j-th ranked p-value; α is the user specified threshold; N is the total number of hypotheses; and $C_N = \sum_{i=1}^{N} \frac{1}{i}$, is the constant for dependent test [1].

Signaling pathway status are estimated by the hypothesized genetic pathways. To get a more intuitive interpretation, the pathway score $score(\vartheta)$ is converted into a probability metric:

$$conf(\vartheta) = 1 - score(\vartheta).$$

According to the definition, each gene g on a genetic pathway has an impact on the downstream events. This impact can be represented as the relative relationship between g and the virtual node of the pathway, which is denoted by $impact(g)$.

Thus, for a signaling pathway γ, let $\vartheta \sim \gamma$ denote the hypothesized genetic pathway ϑ for γ, and G_ϑ denote the genes involved in ϑ. The signaling pathway status Z_i^γ is a weighted aggregation of RE of genes on γ respecting to their impact on γ at time point i, with the weight equaling to the confidence value of the pathway, which

[1] This is because multiple hypotheses for the same signaling pathway have overlap on genes. More details are in Herrington (2002).

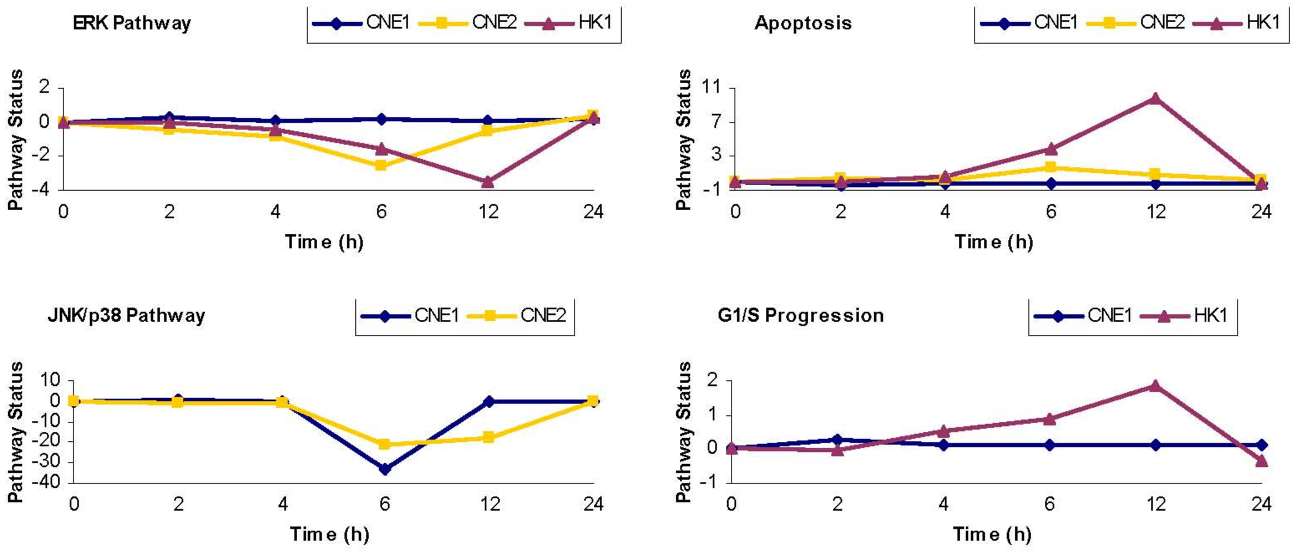


Fig. 3. Comparable diagrams of pathway status profiles of the three cell lines.

is in formula:

$$Z_i^{\gamma} = \sum_{\vartheta \sim \gamma} \sum_{g \in G_{\vartheta}} \left(\frac{1}{\mid G_{\vartheta} \mid} \times impact(g) \times r_{gi} \times \frac{conf(\vartheta)}{\sum_{\vartheta' \sim \gamma} conf(\vartheta')} \right).$$

Similarly, the confidence of the status of γ is a weighted aggregation of the confidence of ϑ, represented in formula as:

$$conf(Z^{\gamma}) = \sum_{\vartheta \sim \gamma} \left(conf(\vartheta) \times \frac{conf(\vartheta)}{\sum_{\vartheta' \sim \gamma} conf(\vartheta')} \right).$$

The pathway status is a synthesis for expression change of genes on regulated genetic pathways. It is a snapshot of pathway regulation on genetic level and provides a benchmark to compare different sample response to drug treatment. Specifically, both genetic pathway and signaling pathway are associated with confidence scores, but the meanings are different. For a genetic pathway, the confidence represents the probability for a pathway to be a regulated pathway under treatment, while for a signaling pathway, the confidence is simply an overall evaluation of the hypothesized underlying genetic pathways.

Evaluating differentially regulated pathways

Signaling pathway status can be compared to discover the reasons for different drug response in different samples. For this reason, we calculate the difference of signaling pathway γ between sample $s1$ and $s2$ by measuring the maximal differentiation of pathway status between $s1$ and $s2$, formulated as:

$$diff_{\lambda}(s1, s2) = \max_{i} |Z_i^{\gamma_{s1}} - Z_i^{\gamma_{s2}}|.$$

To verify the effectiveness of our method, for each signaling pathway, we randomly select gene sets of the same size as the observed genes on the hypothesized genetic pathways for 10000 times, and estimate significance of the pairwise differentiations of pathway status.

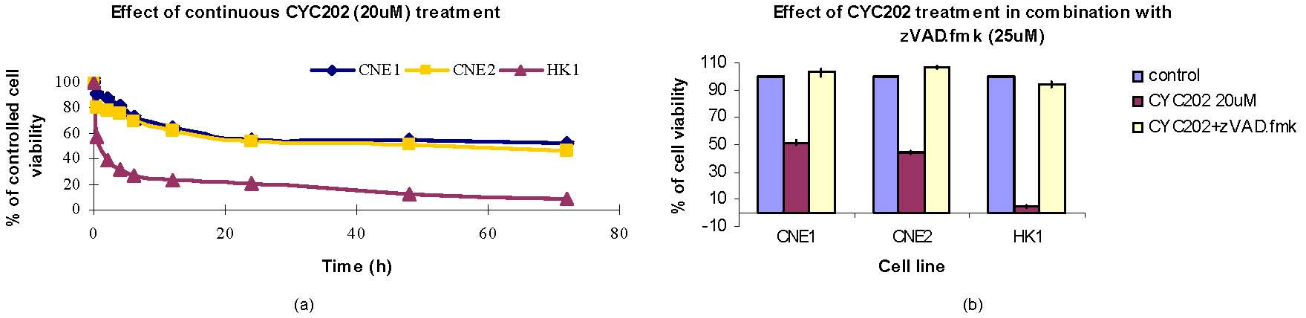
Table 1. p-values for the differentiations of status of signaling pathways.

Comparison Group	ERK	Apoptosis	JNK/p38	G1/S
CNE1 *vs.* CNE2	< 0.0001	0.0028	0.2921	-
CNE1 *vs.* HK1	< 0.0001	0.0006	-	0.4992
CNE2 *vs.* HK1	0.0004	0.0022	-	-

RESULTS AND DISCUSSION

We show the results of applying our system to the NPC study. For both datasets, we set 0.05 as the threshold for p-value cutoff and 0.5 for FDR control. For the *in vitro* part, the diagrams of comparable pathway status profiles and the pairwise p-values of the differentiations between cell lines are shown in Figure 3 and Table 1, respectively. The regulation of ERK pathway and the apoptosis pathway are significantly differentiated among the three cell lines (reaching $4E-4$ at least for ERK pathway and $2.8E-3$ at least for the apoptosis pathway).

ERK pathway regulates the survival, proliferation and differentiation of cells. In the diagrams, it is significantly suppressed in the responding cell line, HK1, but less suppressed or not suppressed in the half-responding cell line, CNE2, and the resistant cell line, CNE1, respectively. This observation is consistent with the results of the trypan blue test (shown in Figure 4 (a)), which measures the viability of NPC cells under treatment. The apoptosis pathway, on the other hand, is more significantly up regulated in HK1 rather than in CNE1 and CNE2. We confirm this hypothesis with the



(a)

(b)

Fig. 4. Results of the associated medical assays in the drug study: (a) shows the change of cell viability for three cell lines under the drug treatment across the time. (b) shows the extent of caspase-dependent apoptosis in three cell lines. zVAD.fmk is a caspase activity inhibitor.

Table 2. List of the identified genetic pathways: Genes for replacement are separated by "/".

Signaling Pathway	Genetic Pathway	Confidence
CNE1		
ERK	GRB2→SOS2→HRAS→RAF1→MAP2K1→ MAPK1/MAPK3	⩾ 0.999
Apoptosis	PIK3CB→PTEN→AKT2/AKT3→CHUK/IKBKB/IKBKG→NFKB2→BIRC2/BIRC5	⩾ 0.9998
JNK/p38	MAP3K12→MAP2K7→MAPK9	0.9665
G1/S	CCND1→CDK4⊣RB1⊣E2F2/E2F3	⩾ 0.9906
CNE2		
ERK	GRB2→SOS1→MRAS/KRAS/NRAS/RRAS→BRAF→MAP2K1→MAPK1	⩾ 0.9885
Apoptosis	PIK3CA/PIK3CB→PTEN→AKT1→IKBKB→RELA→BIRC2/BIRC5	⩾ 0.9949
JNK/p38	MAP4K3/TRAF2→MAP3K1→MAP2K4→MAPK8/MAPK10	⩾ 0.9658
HK1		
ERK	GRB2→SOS1→HRAS→BRAF→MAP2K1/MAP2K2→MAPK1/MAPK3	⩾ 0.9646
Apoptosis	PIK3R1→PTEN→AKT2/AKT3→IKBKB→NFKB2/RELA→BCL2/BIRC2	⩾ 0.9663
G1/S	CUL1→SKP2→CDKN1A⊣CDK6⊣RB1⊣E2F2/E2F3	⩾ 0.9645

results of the assay testing the caspase-dependent apoptosis (shown in Figure 4 (b)).

The list of hypothesized genetic pathways are given in Table 2 with their associated confidence. For ERK pathway, the regulations of Ras family genes as well as ERKs, MAPK1 and MAPK3, are significant. For the apoptosis pathway, we identify the regulation of anti-apoptotic PI3K-Akt-NFκB pathway. The activation of this pathway will induce the expression of multiple cell survival genes, including BIRC2, BIRC4, BIRC5, BCL2, BCLXL, etc, leading to the suppression of cell death. Respecting the results of pathway status evaluation, this discovery suggests that compared to pro-apoptotic caspase cascade, the suppression of anti-apoptotic mechanism seems to play more important roles in effective NPC treatment.

The results of the *in vivo* dataset is shown in Table 3. From the table, tumor samples are classified into two groups with respect to their molecular response to treatment. Pt18 is the sample without tumor. For this sample, consistent pathways are identified for cell proliferation, cell cycle regulation, and apoptosis, with ERK pathway slightly suppressed, and the G1/S cell cycle progression and apoptosis pathway slightly induced. From the table, we observe that the post-treatment status of ERK pathway and apoptosis pathway of Pt18 can be used to separate the two responding groups in a nearly perfect manner. Except for Pt14, all responders exhibit a more significant suppression in ERK pathway and induction in the apoptosis pathway compared to Pt18, while all non-responders exhibit the opposite behavior. This observation is consistent with the results of the *in vitro* dataset.

Epstein-Barr virus (EBV) infection has been learnt to play a critical role in the pathogenesis of NPC (Pathmanathan *et al.*, 1995) (the LMP1, a key effector of EBV-mediated B cell transformation is reported to express in more than 80% of NPC biopsies (Brook *et al.*, 1992)). The dysregulation of multiple signaling pathways, including NFκB, MAP kinase (ERK, JNK and p38), JAK-STAT and PI3K-Akt are suggested induced by EBV infection (Tsao *et al.*, 2002). Particularly, it is specified that the up regulation of NFKB2 and BIRC5 contribute in increasing resistance to apoptosis, and the role of BIRC5 in resisting apoptosis in NPC has been confirmed by RNA interference (Shi *et al.*, 2006). On the other hand, CYC202 inhibits CDK-2, -7 and -9 through competitive inhibition of ATP binding (Mcclue *et al.*, 2002). CDK7 and CDK9 phosphorylate the carboxyl terminal domain of RNA polymerase II, which initiates the gene transcription. The efficacy of CYC202 has been evaluated in a panel of cancer cells, including B-cell chronic lymphocytic leukemia (Alvi *et al.*, 2005), colon cancer (Whittaker *et al.*, 2004), lung

Table 3. The results of signaling pathway status estimation for the *in vivo* dataset: The "response" column shows the molecular response to treatment for patients. The "status" column shows the estimated post-treatment pathway status.

Patient	Response	ERK		JNK/p38		G1/S		Apoptosis	
		Status	Conf.	Status	Conf.	Status	Conf.	Status	Conf.
Pt5	P(ositive)	-2.25	0.98	-3.08	0.99	-	-	1.34	0.99
Pt8	P	-	-	-1.01	0.99	-	-	0.82	0.98
Pt9	P	-0.97	0.98	-	-	0.76	0.95	-	-
Pt14	P	-	-	-	-	-0.61	0.99	-0.86	0.99
Pt16	P	-0.20	0.99	-0.20	0.95	0.29	0.99	1.42	0.97
Pt17	P	-1.02	0.99	-1.02	0.99	-0.33	0.96	1.01	0.99
Pt19	P	-	-	-0.86	0.98	-	-	0.91	0.98
Pt18	No Tumor	**-0.15**	0.99	-	-	0.28	0.99	**0.13**	0.99
Pt1	N(egative)	0.21	0.95	0.52	0.99	1.06	0.97	-1.00	0.98
Pt7	N	-0.10	0.97	-0.68	0.96	0.28	0.98	0.11	0.98
Pt10	N	1.02	0.99	1.16	0.99	-	-	-1.57	0.97
Pt15	N	-	-	-	-	-	-	-1.01	0.98
Pt20	N	1.30	0.98	-	-	-0.93	0.96	-1.68	0.99

cancer (Raje *et al*., 2005), etc. Due to the suppression of gene transcription, the greatest effect is observed on gene products with short mRNA and protein half life, such as apoptosis regulators, including NFκB targeted genes and IAP family (BIRC2, BIRC4 and BIRC5), M-phase cell cycle regulators, and some other transcriptional inducible genes (Lam *et al*., 2001). The suppression of genes involved in ERK pathway, anti-apoptotic pathway and cell cycle regulation, including MAPK1, MAPK3, MCL1, BCL2, BIRC4, BIRC5, CCND1, are frequently observed associated with the treatment of CYC202 (Meijer *et al*., 1997; Whittaker *et al*., 2004; Alvi *et al*., 2005; Raje *et al*., 2005; Smith and Yue, 2006; Lacrima *et al*., 2005).

In the present study, our Drug Pathway Decipherer identifies the different regulation of ERK pathway and the apoptosis pathway between responders and non-responders both *in vitro* and *in vivo*. The hypothesized underlying genetic regulatory mechanisms are consistent with the results from the literature. Our results indicate that the dysregulation of Ras-ERK pathway and PI3K-Akt-NFκB pathway are probably the mechanisms for CYC202-insensitive NPC cells to resist the drug treatment. Evidences show that the simultaneous failure of multiple cancer suppression pathways may be derived from the unsuccessful suppression of the activity of RNA polymerase II. Furthermore, respecting the observation of decreasing Rb phosphorylation and obvious G1 cell cycle arrest in CNE1 (data not shown), it suggests the drug has a potency on the suppression of G1/S progression in non-responders. We suspect that CYC202 has a binding preference to CDK2 rather than to CDK7 and CDK9 in the drug resistant NPC cells.

CONCLUSIONS

In this paper, we introduce our drug pathway identification system for the purpose of treatment understanding, and report the application of this system to an NPC study. We generate hypotheses for consistently regulated genetic pathways and estimate the status of multiple signaling pathways. By comparing the pathway status, we conclude that the unsuccessful suppression of the ERK pathway and anti-apoptosis pathway is the reason for non-responders to escape the drug arrest, which is probably due to the binding preference of the drug in different samples.

Two issues remain in our system. First, the responding group information is ignored. Samples with similar response may include extra information for drug pathway identification. Because of the limitation of the sample size of our study, this information is not taken into consideration, while nevertheless, from the experiment results of the *in vitro* dataset, the drug responders and non-responders can be nearly perfectly separated with respect to the pathway status, which verifies our expectation. Second, the compensatory relationships between genes are not considered. Since gene functions are redundant, it is non-trivial to consider genes with similar roles in a particular process to estimate the overall status of pathway regulation.

ACKNOWLEDGEMENT

The NPC patient data are kindly provided by Dr. Boon Cher Goh, National University Hospital Singapore. This work is supported in part by a NUS research scholarship (Dong), a MOE AcRF Tier 1 grant (Wong), and NMRC grant 0844/2004 (Cui and Mow).

REFERENCES

Alvi,A. et al. (2005) A novel CDK inhibitor, CYC202 (R-roscovitine), overcomes the defect in p53-independent apoptosis in B-CLL by down-regulation of genes involved in transcription regulation and survival, *Blood*, **105**, 4484-4491.

Breitling,R. et al. (2004) Graph-based iterative group analysis enhances microarray interpretation, *BMC Bioinformatics*, **5**, 100-109.

Brook,L. et al. (1992) Epstein-Barr virus latent gene transcription in nasopharyngeal carcinoma cells: coexpression of EBNA1, LMP1 and LMP2 transcripts, *The Journal of Virology*, **66**, 2689-2697.

Cabusora,L. et al. (2005) Differential network expression during drug and stress response, *Bioinformatics*, **21**, 2898-2905.

Dijkstra,E. (1959) A note on two problems in connection with graphs, *Numerische Mathematik*, **1**, 269-271.

Doniger,S. et al. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data, *Genome Biology*, **4**, R7.

Herrington,H. (2002) Controlling the false discovery rate in multiple hypothesis testing, *http://www.unt.edu/benchmarks/archives/2002/april02/rss.htm*.

Fisher,R. (1932) *Statistical Methods for Research Workers*, 4th edition, Oliver and Boyd, London.

Guo,Z. et al. (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network, *Bioinformatics*, **23**, 2121-2128.

Ideker,T. et al. (2002) Discovering regulatory and signalling circuits in molecular interaction networks, *Bioinformatics*, **18**, s233-s240.

Kanehisa,M. et al. (2002) The KEGG database at GenomeNet, *Nucleic Acids Research*, **30**, 42-46.

Hanisch,D. et al. (2002) Co-clustering of biological networks and gene expression data, *Bioinformatics*, **18**, s145-s154.

Lacrima,K. et al. (2005) *In vitro* activity of cyclin-dependent kinase inhibitor CYC202 (Seliciclib, R-roscovitine) in mantle cell lymphomas, *Annals of Oncology*, **16**, 1169-1176.

Lam,L. et al. (2001) Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anti-cancer drug flavopiridol, *Genome Biology*, **2(10)**, research004.

Lee,N. et al. (2002) Intensity-modulated radiotherapy in the treatment of nasopharyngeal carcinoma: an update of the UCSF experience, *International journal of radiation oncology, biology, physics*, **53**, 12-22.

Mcclue,S. et al. (2002) *In vitro* and *in vivo* antitumor properties of the cyclin dependent kinase inhibitor CYC202 (R-ROSCOVITINE), *International Journal of Cancer*, **102**, 463-468.

Meijer,L. et al. (1997) Biochemical and cellular effects of roscovitine, a potent and selective inhibitor of the cyclin-dependent kinase cdc2, cdk2 and cdk5, *European Journal of Biochemistry*, **243**, 527-536.

Nacu,S. et al. (2007) Gene expression network analysis and application to immunology, *Bioinformatics*, **23**, 850-858.

Pathmanathan,R. et al. (1995) Clonal proliferations of cells infected with Epstein-Barr virus in preinvasive lesions related to nasopharyngeal carcinoma, *The New England Journal of Medicine*, **333**, 693-698.

Raje,N. et al. (2005) Seliciclib (CYC202 or R-roscovitine), a small-molecule cyclin-dependent kinase inhibitor, mediates activity via down-regulation of MCL1 in multiple myeloma, *Blood*, **106**, 1042-1047.

Scott,M. et al. (2005) Identifying regulatory subnetworks for a set of genes, *Molecular & Cellular Proteomics*, **4**, 683-692.

Shi,W. et al. (2006) Multiple dysregulated pathways in nasopharyngeal carcinoma revealed by gene expression profiling, *International Journal of Cancer*, **119**, 2467-2475.

Sivachenko,A. et al. (2005) Identifying local gene expression patterns in biomolecular networks, *Computational Systems Bioinformatics Conference (CSB)*, 180-184, Stanford University.

Sivachenko,A. et al. (2007) Molecular networks in microarray analysis, *Journal of Bioinformatics and Computational Biology*, **5**, 429-456.

edited by Smith,P., Yue,E. (2006) *Inhibitors of Cyclin-dependent Kinases as Anti-tumor Agents*, Taylor and Francis Group.

Soh,D. et al. (2007) Enabling more sophisticated gene expression analysis for understanding diseases and optimizing treatments, *ACM SIGKDD Explorations*, **9**, 3-14.

Sohler,F. et al. (2004) New methods for joint analysis of biological networks and expression data, *Bioinformatics*, **20**, 1517-1521.

Subramanian,A. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Science of the United States of America*, **102**, 15545–15550.

Sultanem,K. et al. (2000) Three-dimensional intensity-modulated radiotherapy in the treatment of nasopharyngeal carcinoma: the University of California, San Francisco experience, *International journal of radiation oncology, biology, physics*, **48**, 711-722.

Tsao,S. et al. (2002) The significance of LMP1 expression in nasopharyngeal carcinoma, *Cancer Biology*, **12**, 473-487.

Whittaker,S. et al. (2004) The cyclin-dependent kinase inhibitor CYC202 (R-Roscovitine) inhibits retinoblastoma protein phosphorylation, causes loss of cyclin D1, and activates the mitogen-activated protein kinase pathway, *Cancer Research*, **64**, 262-272.

Yu,M. and Yuan,J. (2002) Epidemiology of nasopharyngeal carcinoma, *Seminars in Cancer Biology*, **12**, 421-429.

Zeeberg,B. et al. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data, *Genome Biology*, **4(4)**, R28.

Zien,A. et al. (2000) Analysis of gene expression data with pathway scores, *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, **8**, 407-417.