**ABSTRACT**

**Motivation:** The high redundancy of and high degree of cross-talk between biological pathways hint that a sub-pathway may respond more effectively or sensitively than the whole pathway. However, few current pathway enrichment analysis methods account for the sub-pathways or structures of the tested pathways. We present a sub-pathway-based enrichment approach for identifying a drug response principal network, which takes into consideration the quantitative structures of the pathways.

**Result:** We validated this new approach on a microarray experiment that captures the transcriptional profile of dexamethasone (DEX)-treated human prostate cancer PC3 cells. Compared with GeneTrail and DAVID, our approach is more sensitive to the DEX response pathways. Specifically, not only pathways but also the principal components of sub-pathways and networks related to prostate cancer and DEX response could be identified and verified by literature retrieval.

**Contact:** chenxiujie@ems.hrbmu.edu.cn; lixia@hrbmu.edu.cn

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

Received on June 13, 2010; revised on December 14, 2010; accepted on December 17, 2010

1 **INTRODUCTION**

It is increasingly important to understand the effects of treatment drugs on cellular pathways and networks. Microarrays provide a novel way to investigate the expression of thousands of genes simultaneously, with a view to dissecting the molecular pathways (Huang da et al., 2009). With the availability of biological pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) or BioCarta (www.biocarta.com), several resources have been developed to visualize and analyze microarray data in the context of known biological networks/pathways. A number of statistical tests combined with known biological databases for detecting significant pathways are available for data interpretation, including GenMAPP (Dahlquist et al., 2002), ChipInfo (Zhong et al., 2003), DAVID (Sherman et al., 2007), GoMiner (Zeeberg et al., 2005), GSEA-P (Subramanian et al., 2005), Onto-Tools (Khatri et al., 2007) and GeneTrail (Backes et al., 2007).

These functional analysis tools are valuable but are unsuitable for the detection of drug response pathways, due to a number of limitations. First, most traditional strategies use only pre-selected gene lists (often the differentially expressed genes) to compute the statistic that tests the enrichment of each pathway. Common statistical methods, such as chi-square, Fisher’s exact test, binomial probability and hypergeometric distribution, are used to compare the number of genes in the list that coincide with a given pathway by random chance. Neglecting less relevant genes, and merely counting the number of pathway members contained in the reduced list, leads to loss of information such as gene order in the pathways. Second, while the prevalent gene-set enrichment analysis (GSEA) (Subramanian et al., 2005) does account for gene order, the effect of the gene-set size, the influence of other gene sets and normalization procedures may lead to inaccurate assessment of statistical significance in some instances (Damian and Gorlina, 2004). Third, the correlation structure of pathways is not considered. A pathway extracted from biological databases is the collection of specific sub-pathways or modules that perform certain functions. Different sub-pathways may perform the same function in the same pathway and different pathways may use the same sub-pathways in similar roles. Testing the whole pathway is too universal to determine which individual sub-pathways respond to a particular biological condition. Fourth, drug-induced cell microarray experiments usually contain few samples (typically, a treated sample and a control sample). Thus, the pathway significance test is limited to the traditional strategies mentioned above.

To overcome these limitations, we propose a framework of sub-pathway-based strategy to identify the drug response pathways/network. The framework consists of the generation of a large number of relative sub-pathways (sub-graphs of the pathways obtained from the KEGG public database), mapping of the unfiltered expression data onto them and statistically scoring for identification of the principal component of sub-pathways (PCSS) that is most perturbed by two stage designs. PCSSs are then combined into a larger drug response network, on which topological and biological analyses are performed. In an analysis of the dexamethasone (DEX) dataset, we found that this approach can enhance the biological interpretation of drug effect.
2 METHODS

2.1 Data description, preprocessing and normalization

Our DEX response dataset contained seven cell intensity files (CEL) downloaded from CMAP (Lamb, 2007) (Instance ID 2079), one derived from PC3 cell lines treated with 9.2 μM DEX and the other six derived from dimethyl sulfoxide (DMSO)-treated PC3 cell lines. The raw data were scaled using the R RMA package. For each sample, the expression values of all probes for a given gene were reduced to a single value by taking the maximum expression value. For the DMSO-treated samples, the expression value for a given gene was the average of the six samples, effectively reducing these samples to a single DMSO-treated sample.

2.2 Obtaining sub-pathways by parsing the KEGG pathway

Each pathway map stored in KEGG can be downloaded in its own XML format (named KGML), which contains information about nodes in the map, such as their relationships, their coordinates, and what genes or compounds exist in the nodes. The pathway maps describe the combination of ordered linear sequences of protein direct/indirect interactions or metabolic reactions from biologically meaningful start-points (membrane receptors or their ligands) to end-points (transcriptional factors or their immediate targets). In such a map, the start-points have no parent nodes and the end-points have no descendant nodes. The sub-pathway is defined by an individual path from a start-point to an end-point in a pathway map. For a given KEGG pathway, the sub-pathways were obtained by searching all possible paths between start-points and end-points in the adjacency matrix generated by node relationships extracted from KGML files.

2.3 Identification of significant drug–response principal sub-pathways

To identify significant principal sub-pathways responding to a drug, two-stage tests are employed (Lin, 2006; Nguyen et al., 2009; Skol et al., 2007). The first is used to define a subset of candidate sub-pathways, from which the second can capture the principal sub-pathways.

In the first-stage test, the sub-pathway state is represented by a vector $g_i = [g_1, g_2, \ldots, g_n]$, where $g_i$ is the expression value of gene $i$ in sub-pathway $j$. Here, we employ Euclidean distance $d_j = \sqrt{(p_j - p'_j)^2}$ to describe the state change of each sub-pathway, where $p_j$ and $p'_j$ represent the states of sub-pathway $j$ after and before drug treatment, respectively. A larger distance corresponds to greater sensitivity of the sub-pathway to the drug. When there are several genes in a sub-pathway node, the maximum expression-fold-change gene is chosen for the node. To identify significant drug response sub-pathways, we take the statistic for all sub-pathways and draw 100000 random gene sets of the same size from all the genes to represent the background distribution. The $P$-value is calculated as the fraction of re-sampled gene set statistic that exceeds the observed value (FDR cutoff 0.05).

The second-stage test focuses on the key genes (PCS) that drive the drug response sub-pathways. Our goal is to develop an approach that would identify PCSs and their networks most influenced by drugs. First, we selected the top $n$ genes that gave cumulative contribution at least up to 80% to significant sub-pathway state change and designated these PCSs. The contribution was calculated by

$$I_j = \frac{\sum_{i \in j} (g_i^{(0)} - g_i^{(1)})^2}{\sum_{i \in j} (g_i - g_i^{(0)})^2},$$

where $g_i^{(0)}$ and $g_i$ are the expression values of gene $i$ in the presence or absence of drugs, respectively, $k$ is the rank of gene $i$ contribution and $n$ is the size of sub-pathway $j$. Next, the significance of PCSs were assessed by the same method used to identify significant drug response sub-pathways (cf. the first stage test in the above paragraph), and the $P$-value was adjusted by false discovery rate (FDR, cutoff 0.05) (Benjamini and Hochberg, 1995). Thus, the significant PCS is the sub-pathway since this PCS dominates the state change of the sub-pathways. Effectively, the sub-pathways containing the significant PCSs represent significant responses to a drug. To prioritize these PCSs with respect to DEX response in PC3 cell lines, the PCS were scored by

$$S_i = \sum_{q \in \text{PCS}} f_q,$$

where $f_q$ is the frequency of gene $q$ in PCS $c$ (composed of $u$ genes) appearing in $w$ PCSs.

2.4 Construction and analyses of principal network to drug

From the identified PCSs and their structures (key gene relationships) maintained from their original sub-pathways, we constructed a principal drug response network and analyzed its topological properties using NetworkAnalyzer (Assenov et al., 2008).

3 RESULTS

3.1 General information of sub-pathway

The 201 pathway XML files were downloaded from KEGG, and 169 pathways involving metabolism, genetic information processing, environmental information processing, cellular processes and human diseases were used for further analyses after excluding the pathways lacking node relationships. In total, 87732 sub-pathways were generated from the 169 KEGG pathway maps by the depth-first search algorithm. These contained 4179 genes and were highly redundant since each gene participated in approximately 20 sub-pathways. The maximum number of sub-pathways (16154) was identified for the glycosphinoglycid biosynthesis-lacto and neolacto series pathways, while several pathway maps, such as the tetrachloroethene degradation pathway, were parsed into only one sub-pathway (Supplementary Table 1).

3.2 Significant response sub-pathway to DEX in first-stage test

A total of 8252 sub-pathways containing 825 genes were significantly identified ($P < 0.05$), indicating that the response genes were highly shared, or cross-talked, between the sub-pathways or pathways. Significant sub-pathway lengths, derived from the KEGG pathways (Supplementary Table 2), ranged from 2 to 18, with an average of 9 (Fig. 1).
The rate distribution of cumulative contribution of the top \( n \) genes up to 80% for the different length of significant sub-pathways. At most, six genes could represent the majority of sub-pathways, with only three sub-pathways requiring more key genes (i.e. only three sub-pathways could not be utilized under our specified conditions).

The proportions of significant sub-pathways varied for different pathways. For example, in the Circadian rhythm pathway, all sub-pathways were significant, whereas the proportion of significant sub-pathways in the Arachidonic acid metabolism pathway was only 0.063%.

The state changes in these significant sub-pathways were determined by the genes within them. However, these genes affected their sub-pathway states to different extents. The expression level changes of these genes dominated the whole sub-pathway, thus their cumulative contributions were scored by the statistic (Supplementary Figure 1). More than 90% of sub-pathways containing less than nine genes and more than 85% of sub-pathways containing 9–18 genes, were controlled by 2–3 genes and by 4–6 genes, respectively (Supplementary Figure 1). This implies that only a few genes (PCSs) regulate the drug response sub-pathways. Therefore, focusing on PCS genes might provide an economical way of elucidating the molecular mechanisms of drug response.

3.3 PCSs to DEX

As described above, the entire sub-pathway seems to be governed by only a few genes. It is therefore necessary to know which genes dominate the sub-pathway response state and the maximum number of these genes required. Therefore, we ranked the genes by their contributions to the state change of sub-pathway from high to low, and defined PCS standing for the sub-pathway by the cumulative contribution of the top \( n \) genes up to 80%. We then calculated the proportions of different length PCSs to different length sub-pathways (Fig. 2). We obtained 555 PCSs (FDR < 0.05) covering 213 genes, and found not only that one or several genes were highly shared in more than two PCSs, but also that the PCSs were highly shared by sub-pathways; indeed, one PCS could represent an average of 16 sub-pathways. These PCSs represented the principal states of DEX response pathways/sub-pathways. In addition, as shown in Figure 2, at most six genes can represent one sub-pathway. As they were highly shared, it could be inferred that shorter sub-pathways are linked in longer sub-pathways, consistent with the idea that signal cascades choose the most economical path in vivo.

Conventionally, the next step is to select the most important PCSs in the drug response sub-pathway. To this end, we applied the Sc statistic to evaluate the priorities of PCSs to DEX in PC3 cell lines (Supplementary Table 3). However, it was difficult to determine the optimal PCS since the PCSs were too highly cross-talked. For example, the top 10 PCSs contained eight genes, with four genes in common and with no gene additions from the top third to the top tenth PCS (Fig. 3). This indicates that the first 10 PCSs are strongly related. This trend was apparent throughout the PCS list; the appearance of a new gene in a PCS signified that no new genes would appear in the next several PCSs (Fig. 3; Supplementary Table 3). Therefore, we used the gene direct/indirect connections from their original sub-pathways to merge the first 10 PCSs into a single sub-pathway (Fig. 4). The new sub-pathway reveals the basic molecular mechanisms and effects of DEX in PC3 cell lines.

3.4 The principal response network to DEX

To reveal the comprehensive molecular mechanisms of DEX in PC3 cell lines, we merged all PCSs and constructed a DEX response principal network, since, as described above, PCSs for the significant sub-pathways are highly cross-talked (Fig. 5). The correlation coefficient of node degree distribution between the principal network and the network merging of all significant sub-pathways was 0.69 (\( P < 0.01 \)), implying that the principal network had inherited the main structure of the latter. The top 10 genes with degree and betweenness centrality are shown in Supplementary Table 4. These signify important genes through which DEX exerts its complex effects in PC3 cell lines. In addition, there exist several isolated
When the fold change cutoff was 2, 28 genes were differentially expressed. To identify drug response sub-pathways, it was a direct and simple method to link the differentially expressed gene pairs in the KEGG pathways. Therefore, we compared M2 with our method (M1) and found that few differentially expressed gene pairs in the KEGG pathways could be connected (Supplementary Table 6). It should be noted that it was doubly-linked between VEGFC and EGFR, contrary to what was observed in the sub-pathway merging the top 10 PCSs (Fig. 4). DEX was found to stabilize EGFR and decrease VEGFC to suppress angiogenesis, whereas PC3 cells increased EGFR as a survival response. The double-link yields additional information about the interaction between the effects of DEX and cancer cell self-protection.

3.5 The comparison with the method which links the differentially expressed gene pairs in the KEGG pathways

Fig. 5. DEX response principal network. The pink circle is for increased expression level and the green circle is for decreased expression level. The double-link between VEGFC and EGFR yields additional information about the interaction between the effects of DEX and cancer cell self-protection.

To identify drug response sub-pathways, it was a direct and simple method (M2) to link the differentially expressed gene pairs in the KEGG pathways. Therefore, we compared M2 with our method (M1) and found that few differentially expressed gene pairs in the KEGG pathways could be connected (Supplementary Table 6). When the fold change cutoff was 2, 28 genes were differentially expressed and 13 genes were found in the KEGG pathways and no double-link was found between any differentially expressed gene pair in one KEGG pathway. When the fold change cutoff was lowered at 1.5, 62 genes were differentially expressed and 28 genes were decentralized in 54 KEGG pathways and only one path of four genes (Senate-Notch-CSL-Hes1/5, Senate and Hes1/5 differentially expressed) was found in Notch Signaling pathway. These results show that most genes of linear sub-pathways or pathways do not change their expression very large based on current knowledge of pathways. Therefore, we considered the sub-pathway as a whole responding to drug and developed M1. Further, we lowered the fold change cutoff from 1.4 to 1.2 and the result of M2 varied very large (Supplementary Table 6). It is not easy to choose a fold change cutoff here. Cross-talks are very universal between pathways and give rise to disease or respond significantly to a drug. Single sub-pathways performing relatively simple functions coupled with specific path may be more suitable for assessing drug response, than complex pathways which are currently poorly understood. To illustrate our point, we detected DEX response pathways in PC3 cell lines using GeneTrail. None of the pathways obtained from GeneTrail were related to biological pathways such as prostate cancer, cancer pathways in general, WNT signaling pathway, MAPK signaling pathway, etc. DAVID, with 28 genes (fold change >2) or 62 genes (fold change >1.5) detected null or only one pathway (cytokine-cytokine receptor interaction). In contrast, our methods identified 52 significant pathways (Supplementary Table 7), of which a subset was unequivocally associated with DEX response and prostate cancer (as determined by the number of significant sub-pathways containing significant PCSs in the pathways, evaluated by the hypergeometric test). Many pathways detected were triggered directly by DEX, and there was evidence of high cross-talk among the pathways. When a sub-pathway responds, its pathway must respond also, rendering further tests unnecessary. Combined, our results suggest that the sub-pathway based strategy is more sensitive to drug response than whole-pathway approaches.

A total of 825 genes were associated with 8252 significant sub-pathways, with an average of one gene participating in 10 sub-pathways, revealing high cross-talk between the genes. We hypothesize that the important genes change their expression often in the sub-pathways and have a high degree of standing in the sub-pathway network, since they are stimulated by multiple signals. To test this idea, we extracted 555 PCSs according to their contributions to the sub-pathway state changes, and found that the standing degrees of those genes were on average higher (t-test, \( P=5.59E-08 \)) than those of non-PCS genes in the sub-pathway-based network. Afterwards, we designed the statistic to evaluate the priorities of these PCSs to DEX in PC3 cell lines, which effectively ranked PCSs and hinted at strong links between the top 10 PCSs and DEX-treated.

4 DISCUSSION

Current approaches typically study whole pathways, whether by singular enrichment analysis or by gene set enrichment analysis (Huang da et al., 2009). However, these methods may be unsuitable for the identification of drug response genes. A pathway comprises sets of expressed genes that lead to a complex phenotype (e.g. prostate cancer) or which perform a function (e.g. apoptosis). We generated 87,732 sub-pathways from 169 KEGG pathways. Within part of a pathway, some individual sub-pathways may give rise to disease or respond significantly to a drug. Single sub-pathways performing relatively simple functions coupled with specific path may be more suitable for assessing drug response, than complex pathways which are currently poorly understood. To illustrate our point, we detected DEX response pathways in PC3 cell lines using GeneTrail. None of the pathways obtained from GeneTrail were related to biological pathways such as prostate cancer, cancer pathways in general, WNT signaling pathway, MAPK signaling pathway, etc. DAVID, with 28 genes (fold change >2) or 62 genes (fold change >1.5) detected null or only one pathway (cytokine-cytokine receptor interaction). In contrast, our methods identified 52 significant pathways (Supplementary Table 7), of which a subset was unequivocally associated with DEX response and prostate cancer (as determined by the number of significant sub-pathways containing significant PCSs in the pathways, evaluated by the hypergeometric test). Many pathways detected were triggered directly by DEX, and there was evidence of high cross-talk among the pathways. When a sub-pathway responds, its pathway must respond also, rendering further tests unnecessary. Combined, our results suggest that the sub-pathway based strategy is more sensitive to drug response than whole-pathway approaches.
Thus, fixing the contribution rate cutoff at 80% could minimize suppressed PC3 cells take a strong line of self-defense by activating proliferation sub-pathways, by inhibiting apoptosis sub-pathways, and even by promoting the circadian rhythm pathway to activate the WNT pathway (CSNK1D-WNT5A-TCF7L1-JUN-MAP2K1), which increases cytoactivity (Pu and Lee, 2003; Schwarz-Romond et al., 2002). These results suggest that DEX should be combined with other anti-tumor drugs (especially the EGFR inhibitors) in the effective treatment of prostate cancer. Consistent with this, combined inhibition of VEGF and EGFR is effective in EGFR inhibitor-resistant cell lines, and increases benefit compared with either of these anti-EGFR agents alone or when combined with chemotherapy (Tortora et al., 2008).

Reproducibility analysis of the principal DEX response network was performed using another batch instance (ID 5797), held under the same condition as batch instance 2079. A total of 9045 sub-pathways (composed of 808 genes) and 632 PCSs (composed of 217 genes) were significantly identified, of which 4350 sub-pathways (approximately half of sub-pathways) and 71 key genes (approximately one-third of PCS genes) were identical to those found in batch instance 2079. The low reproducibility of key genes was expected, since it reflects the apparently low reproducibility of true discoveries in microarray studies (Zhang et al., 2008). The higher reproducibility of significant sub-pathways indicates that they perform specific functions, again consistent with the results of microarray studies, which are more reproducible at functional level (Gong et al., 2010). In addition, the functional similarity of the Gene Ontology (GO) terms enriched in the two PCS sets, evaluated by G-SESAME (Du et al., 2009) was 0.88, significantly greater (P < 0.01) than that of random gene set pairs of the same size. Moreover, a dataset of 90 samples, DEX-treated myeloma cells (GEO ID: GSE85846) was analyzed. The DEX response principal network (composed of 274 genes) in Myeloma, were significantly identified. Most genes (approximately 1/2) are associated with DEX, Myeloma or cancer (Supplementary Table 8). The functional similarity of GO terms enriched in the principal network from GSE85846, compared with those from instance 2079 and 5797, evaluated by G-SESAME (Du et al., 2009), were 0.837 and 0.864, significantly greater (P < 0.01) than that of random gene-set pairs of the same size. These results show that our sub-pathway-based approach can potentially detect stable drug response principal network at the functional level.

ACKNOWLEDGEMENTS

We thank Associate Editor John Quackenbush and the three reviewers for their suggestions and comments.

Funding: The Scientific and Technological Project of Heilongjiang Province (GQ07C337 and ZD200804-01); Provincial Education Department Project of Heilongjiang (11541121); National Natural Science Foundation of China (30871394, 30972538 and 30772238); the National High Tech Development Project of China, the 863 Program (2007AA02Z320); the key national S&T Program-Major New Drug Development (2009ZX09103-380); The Innovation Manpower Fund of Harbin Science and Technology Bureau (2010RFXS0053); Master Innovation Funds of Harbin Medical University (HCX2009008 and HCXS2010007).

Conflict of Interest: none declared.
REFERENCES


