Original Research Paper

Volume-7 | Issue-11 | November-2017 | ISSN - 2249-555X | IF : 4.894 | IC Value : 79.96

Engineering

DYNAMIC CANCER PATHWAY INTERACTION ANALYSIS USING PATHWAY RANKINGS BASED ON STATIC DATASETS

Shinuk Kim

Department of Civil Engineering Sangmyung University Cheonan, Chungnam, 31066, Republic of Korea

ABSTRACT

Background: Dynamic pathway interaction analysis provides useful information in assessing progression of complex diseases at different pathologic stages and/or time points. However, high-throughput datasets are obtained statically rather than dynamically, making it difficult to assess dynamic changes occurring over the course of disease progression. Here, we report a simple method based on survival times for discovering dynamic pathway interactions using static cancer datasets such as The Cancer Genome Atlas (TCGA). Gene Set Enrichment Analysis was used to rank gene sets or pathways whose outcomes represent differentially expressed leading edge gene scores between tumor and normal samples.

Results: We tested three different cases ordered by survival time and found eight common pairs of positively- or negatively-related pathways. The two most positively correlated pathways were DNA REPLICATION and MISMATCH REPAIR, whereas the two most inversely correlated pathways were SPLICEOSOME and GRAFT-VERSUS-HOST DISEASE.

Conclusions: Our simple method for assessing dynamic pathway interactions will potentially enable the discovery of dynamic pathway networks involved in the pathologic progression of complex diseases.

KEYWORDS : Dynamic pathways; Pathway interactions; Cancer datasets

1. Background

Biomarkers derived from high throughput data have been shown to be useful for predicting disease type and/or patient treatment. These molecular markers, which are based on gene expression profiles, have also been used in experimental and clinical settings to obtain insights into a wide range of biological phenomena. Exploring the most significant molecular changes between different groups has the potential to shed light on the complexity of the disease network and reveal new features. Individual molecular datasets, integrative molecular datasets, and pathway information are typically used to obtain the signatures that are differently expressed (i.e., upregulated or downregulated) in biological processes.

Due to the complexity of different diseases, background noise in high throughput(HT) experiments, the need for multiple hypothesis testing corrections, and patient heterogeneity [1] [2], it has been challenging to experimentally elucidate the biological mechanism(s) relevant to complex diseases. Therefore, methods have been developed that focus on pathway-level analyses, including functional analysis or pathway grouping of functionally-related genes. These methods have been applied to gain systemic insights into the underlying mechanisms of complex diseases such as cancer [3-5].

Pathway analysis (PA) is a gene set-based scoring approach in which the importance of each individual gene is ranked by a statistical approach such as value. The Kolmogorov-Smirnov test [6], means or medians of gene-level statistics [7], and Wilcoxon rank sums [8] are the most common statistical methods for assessing the overall effect of a gene set on a biological phenotype. Gene Set Enrichment Analysis (GSEA) [6], Pathway Enrichment Analysis (PWEA) [9], and Significance Analysis of Microarray to Gene Set (SAM-GS) [10] are tools commonly used to obtain enrichment gene sets or pathways from gene expression data sets in a phenotype of interest at a given time.

Although many PA tools have been developed, many challenges remain in the development and usage of PA methods, as well as in the foundations of dynamic responses. Recently, Khatri [11] described two methodological challenges of pathway analysis for the next generation: annotation extension and methodology for analyzing dynamic responses. Here, we present a novel and simple approach based on survival times for the determination of dynamic pathway interactions using static cancer datasets. This approach will potentially lead to a better understanding of the systemic changes that occur during the pathological progression of complex diseases.

2. Materials and Methods

2.1. Materials

The Cancer Genome Atlas (TCGA) provides researchers with multi-platform data for thousands of tumors from a variety of cancer types and subtypes. Data from 16 normal and 485 glioblastoma multiforme (GBM) tumor samples were downloaded from the TCGA dataset in 2016. The datasets were generated in February 2014 at the University of North Carolina Cancer Genomic Characterization Center using an Agilent G4502A microarray platform containing 17814 genes. In parallel, clinical information of 594 patients with GBM were also downloaded independently from TCGA. Since gene expression and clinical datasets were generated separately from independent institutes, we preprocessed the data to match clinical subjects with gene expression subjects, yielding a total of 445 subjects. Then, we eliminated 79 censored samples, yielding 366 tumor and 16 normal samples for the current data analyses. In this study, we especially considered the overall survival times.

For pathway information, 186 pathways collected in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [12, 13] were downloaded from Molecular signatures Database (MSigDB) [14]. A GSEA computational tool [6] was adopted to obtain the enriched gene sets and pathways, i.e., gene sets and pathways that are differentially expressed between tumor and normal samples.

2.2. Methods

We present an algorithm implemented in a software package that reprocesses tumor data matrices into a format parallel to the survival times used in statistics packages. This algorithm can be used to obtain enrichment pathways for identifying dynamic pathway interactions. This study consisted of four main steps: First, since we obtained microarray datasets and clinical information independently, we needed to match patients' clinical information and gene expression datasets. Second, we partitioned patients in ascending survival time-dependent groups and rearranged the tumor datasets. Third, since gene set enrichment is fundamental to this study, we needed to determine the most meaningful pathways. In our context, GSEA determines whether genes from a predefined gene set or a pathway are significantly overexpressed or underexpressed in a given gene set. The output GSEA genes are referred to as leading edge genes. The scored leading edge genes, which represent a given gene set, are rank-ordered for the predefined gene set. The final step in this workflow is to compute the relationships between these predetermined enrichment pathways. We adopted the Spearman ranking correlation coefficient as a quantitative measure of the relationship between any two pathways. The following pseudo algorithm describes the overall procedure.

Pseudo algorithm

The input T is an M x N tumor data matrix, where m and n represent the numbers of genes and patients, respectively. s is the 1x n survival time vector of n patients. t_i and s_i represent tumor expression datasets and survival time of patient i, respectively.

1. Sort S:= $\{ s_i \mid s_{i-1} < s_i < s_{i+1}, i = 1 \cdots n-1 \}$ in ascending order of survival time. Next, match t_i and rearrange to T. The matrix T is now survival time-dependent in ascending order.
2. Partition tumor samples with Q quantile groups and set Tqi, where

$Q = \{q_i \mid i = 1, \cdots, k\}$, and q_i is the cumulative probability.

3. $Y_q := quantile(T, Q)$ where Y_q is the same size as Q, and the i-th row of $Y_q(i)$ contains the q_i-th quantiles of each column of T. Thus, the i-th quantile dataset is T_{qi}
4. $z_i = \frac{q_i n}{k}$ $T_{q_i} := \{t_{z_i} \mid z_{i-1} < i \le z_i,\ i \in N\}$, if Z_i is equal to an integer, then replace Z_i with $Z_i = Z_i + 1$, otherwise round up Z_i.
5. Implement GSEA to obtain enriched pathways with zero FDR in order to compare the normal and tumor gene expression datasets : For $q = 1 : k$ out put $p^r_q := GSEA\ (normal,\ T_q)$ enriched pathways with ranking r. end
6. Find a common pathway, p_c, $p_c \hat{I} P$, such that $P := \{p_c \in p_1 \bigcap p_2 \bigcap \cdots \bigcap p_{25}\}$, where $c = 1, \cdots, d$.
7. Compute the Spearman ranking correlation coefficients between pi, pj, where $pi, pj \hat{I} P$, $i, j = 1, \cdots d$
   For i=1:d
   For j=1:d
   $SRC_{ij} := Spearman\ correlation\ coefficients\ (p_i, p_j)$ *End*

In the test, we sorted the tumor patient datasets based on survival times in ascending order and then partitioned them into closely matching groups of 16 samples/group. This partitioning balanced the dataset for implementation of GSEA. We only allowed a 25% limitation surplus, resulting in k(pseudo algorithm 2)20, 25=, and 30. We calculated that an average of 18.3 samples belonged to $T_1 \cdots T_{20}$ for k=20, an average of 14.64 samples belonged to $T_1 \cdots T_{25}$ for k=25, and an average of 12.2 samples belonged to $T_1 \cdots T_{30}$ for k=30. The average survival time of each case ranged from 19 (T_1) to 2254 (T_{20}) days for k=20, 16.47 (T_1) to 2439.94 (T_{25})days for k=25, and 16 (T_1) to 2340 (T_{30}) days for k=30.

For k=25, 25 runs of GSEA were executed. In these runs, the set of normal samples was used recurrently, yielding different sets of enriched pathways at each run. To obtain high confidence enrichment pathways, we applied two criteria: i) a false discovery rate (FDR) of zero and ii) an average ranking score less than 20. Since the identified pathways are enriched in cancer, the leading edge genes in these pathways are upregulated compared to their normal levels. However, to consider changes in enrichment scores between pathways, we adopted Spearman ranked correlation coefficients (SRCs) to assess interactions between pathway ranking scores. Figure 1 shows the names of the outcome pathways in the first column and the pathway rankings in each cell.

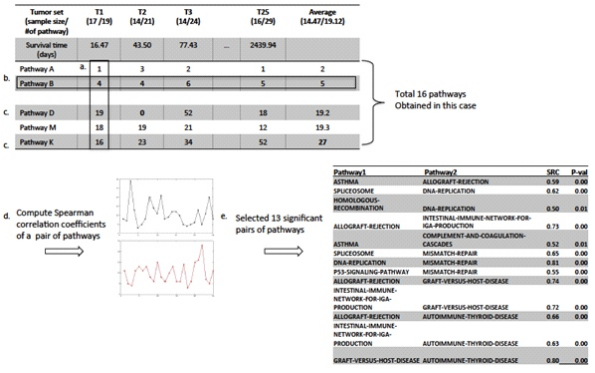

Figure 1. Schematic illustration of a survival time-dependent method to determine pathway interactions. a. Arrangement of enrichment pathways of each tumor set, b. Calculation of the average pathway rankings, c. Elimination of any pathways that are not enriched or do not satisfy the given criteria (pathway D and pathway K), d. Calculation of Spearman correlation coefficients for pathway pairs, e. Pairing of pathways with significant SRCs and p-values.

3. Results

When k=25, between 19 to 29 (an average of 19.12) enriched pathways were identified using our criteria. However, only 16d= of the pathways were shared by $T_1 \cdots T_{25}$. Among the 16 common pathways, 13 pairs were found to have a statistically significant relationship based on the Spearman ranked correlation coefficient cutoff (greater than 0.5) and the p value cutoff ($p < 0.05$).

In addition, we collected common pathways based on GSEA without applying any cutoff criteria. This approach yielded *commond*(d) =33 pathways and identified 88 pairs of significant pathways with Spearman ranked correlation coefficients greater than 0.5 and p<0.05. For k=30 , we obtained 9 pairs of pathways from *commond*(d) =15 enrichment pathways after applying the criteria and 48 pairs of pathways from *commond*(d) =33 enrichment pathways without applying the criteria. For k=20 , we obtained 15 pairs of pathways from *commond*(d) =18 enriched pathways after applying the criteria and 58 pairs of pathways from *commond*(d) =33 enriched pathways without applying the criteria.

We found four common pairs of pathways that were positively related to each other from the three test groups. The first common pair, DNA REPLICATION and MISMATCH REPAIR, was identified from the k=30 (SRC of 0.705, p<0.000), 25 (SRC of 0.814, p<0.000), and 20 (SRC of 0.855, p<0.000) analyses. The second pair, GRAFT-VERSUS-HOST DISEASE and AUTOIMMUNE THYROID DISEASE, was identified when k= 30(SRC of 0.660, p<0.0001), 25 (SRC of 0.804, p<0.0000), and 20 (SRC of 0.822, p<0.0000). The third pair, SPLICEOSOME and MISMATCH REPAIR, was identified when k=30 (SRC of 0.5176, p<0.0034), 25 (SRC of 0.6466, p<0.0005), and 20 (SRC of 0.607, p<0.0046). The last pair, P53-SIGNALING PATHWAY and MISMATCH REPAIR, was identified from the k = 30 (SRC of 0.626, p<0.0002), 25 (SRC of 0.546, p<0.0047), and 20 (SRC of 0.681, p<0.0009) analyses.

The Venn diagram in Figure 2 presents the numbers of pathway pairs in conjunction with the four pairs of common pathways with significantly positive or negative relationships.

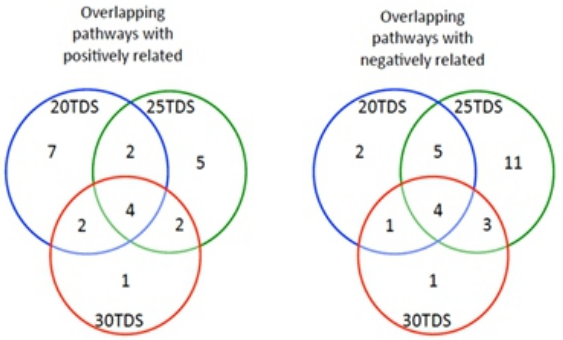

Figure 2. Venn diagram of the numbers of pathway pairs obtained from all three groups. TDS: time-dependent series (K).

We also considered negatively related pathways obtained from all three groups. We found four pairs of common pathways out of the 15, 23, and 12 pairs of pathways identified when

k=30, 25, and 20. The first pair, SPLICEOSOME and AUTOIMMUNE THYROID DISEASE, was identified from the k = 30 (SRC of -0.598, p<0.0005), 25 (SRC of -0.695, p<0.0001), and 20 (SRC of -0.717, p<0.0004) analyses. The second pair, SPLICEOSOME and GRAFT-VERSUS-HOST DISEASE, was identified from the k = 30 (SRC of -0.661, p<0.0001), 25 (SRC of -0.762, p<0.0000), and 20 (SRC of -0.657, p<0.0017) analyses. The third pair, AUTOIMMUNE THYROID DISEASE and MISMATCH REPAIR, was identified from the k = 30 (SRC of -0.524, p<0.0029), 25 (SRC of -0.647, p<0.0005), and 20 (SRC of -0.529, p<0.0164) analyses. The last pair, HOMOLOGOUS RECOMBINATION and GRAFT-VERSUS-HOST DISEASE, was identified from the k = 30 (SRC of -0.639, p<0.0001), 25 (SRC of -0.578, p<0.0025), and 20 (SRC of -0.708, p < 0.0005) analyses.

To better depict the dynamic changes, we compared pairs of common pathways with enriched ranking points and simulated datasets generated by cubic spline interpolation (Figure 3). Dots and lines represent enriched ranking points and interpolation-based cubic spline curves, respectively. The changes observed between the DNA REPLICATION and MISMATCH REPAIR positively-related pathways (Figure 3 (a)) and the GRAFT VERSUS HOST DISEASE and AUTO IMMUNE THYROID DISEASE positively-related pathways (Figure 3 (b)) are displayed when k =25. In addition, the changes observed between the HOMOLOGOUS RECOMBINATION and GRAFT-VERSUS-HOST DISEASE negatively-related pathways (Figure 3(c)) and the SPLICEOSOME and AUTOIMMUNE THYROID DISEASE negatively-related pathways (Figure 3 (d)) are displayed when k=25.

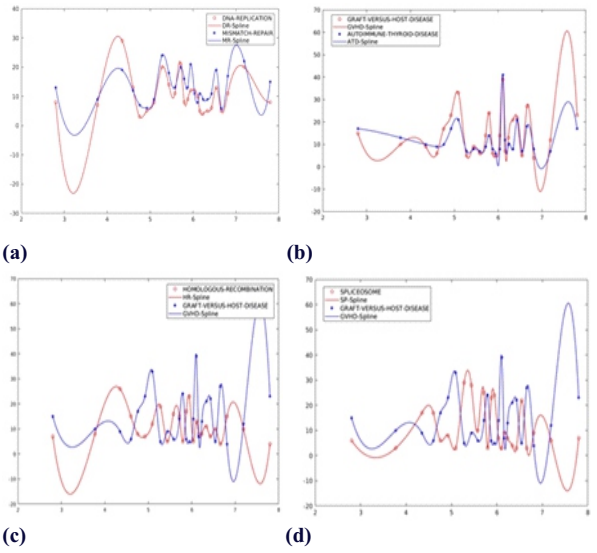

Figure 3. Illustration of the pairs of significant pathways. Positively-related pathways are presented in (a) and (b). Negatively-related pathways are presented in (c) and (d).

The four pairs of positively-related and negatively-related interactive pathways are illustrated in Figure 4. This result implies a link between dynamic pathway changes based on GBM cancer survival time.

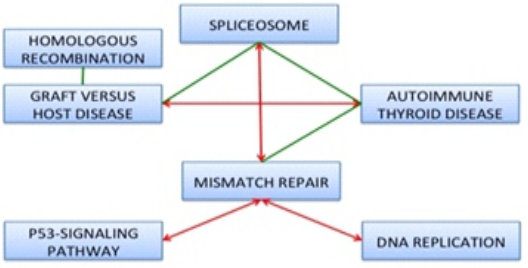

Figure 4. Dynamic pathway network constructed based on GBM cancer survival time. Red arrows and green bars represent positively-related and negatively-related pathways, respectively.

4. Conclusions

This study describes a novel method for deriving dynamic pathway interactions from static cancer datasets based on survival time. Overall, eight (four positive; four negative) pairs of significantly related pathways were obtained consistently from the three test groups, demonstrating the reliability of this approach for pathway analysis. Although we found valuable pathway interactions, cancer is a heterogeneous disease, and it would be useful to add more clinical information such as cancer stage and grade. Therefore, this method for identifying dynamic pathway interactions can serve as the foundation for discovering dynamic pathway networks involved in the pathologic progression of complex diseases.

Acknowledgments

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) via the Ministry of Science, ICT, & Future Planning (NRF-2017R1A2B4010684) and by the Ministry of Education (NRF-2015R1D1A1A01060287).

References

1. Vidal, M., M.E. Cusick, and A.L. Barabasi, Interactome networks and human disease. Cell, 2011. 144(6): p. 986-98.
2. Fernald, G.H., et al., Bioinformatics challenges for personalized medicine. Bioinformatics, 2011. 27(13): p. 1741-8.
3. Chuang, H.Y., M. Hofree, and T. Ideker, A decade of systems biology. Annu Rev Cell Dev Biol, 2010. 26: p. 721-44.
4. Hsiao, T.H., et al., Differential network analysis reveals the genome-wide landscape of estrogen receptor modulation in hormonal cancers. Sci Rep, 2016. 6: p. 23035.
5. Edelman, E.J., et al., Modeling cancer progression via pathway dependencies. PLoS Comput Biol, 2008. 4(2): p. e28.
6. Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.
7. Jiang, Z. and R. Gentleman, Extensions to gene set enrichment. Bioinformatics, 2007. 23(3): p. 306-13.
8. Barry, W.T., A.B. Nobel, and F.A. Wright, Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics, 2005. 21(9): p. 1943-9.
9. Hung, J.H., et al., Identification of functional modules that correlate with phenotypic difference: the influence of network topology. Genome Biol, 2010. 11(2): p. R23.
10. Dinu, I., et al., Improving gene set analysis of microarray data by SAM-GS. BMC Bioinformatics, 2007. 8: p. 242.
11. Khatri, P., M. Sirota, and A.J. Butte, Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol, 2012. 8(2): p. e1002375.
12. Efron, B. and R. Tibshirani, On testing the significance of sets of genes. The annals of applied statistics, 2007: p. 107-129.
13. Zhu, L., et al., MetaDCN: meta-analysis framework for differential co-expression network detection with an application in breast cancer. Bioinformatics, 2017. 33(8): p. 1121-1129.
14. Kanehisa, M. and S. Goto, KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 2000. 28(1): p. 27-30.