Identification of core pathways based on attractor and crosstalk in ischemic stroke

XIUFANG DIAO[1] and AIJUAN LIU[2]

Departments of [1]Respiratory Medicine and [2]Cardiology, Weifang People's Hospital, Weifang, Shandong 261041, P.R. China

Abstract. Ischemic stroke is a leading cause of mortality and disability around the world. It is an important task to identify dysregulated pathways which infer molecular and functional insights existing in high-throughput experimental data. Gene expression profile of E-GEOD-16561 was collected. Pathways were obtained from the database of Kyoto Encyclopedia of Genes and Genomes and Retrieval of Interacting Genes was used to download protein-protein interaction sets. Attractor and crosstalk approaches were applied to screen dysregulated pathways. A total of 20 differentially expressed genes were identified in ischemic stroke. Thirty-nine significant differential pathways were identified according to P<0.01 and 28 pathways were identified with RP<0.01 and 17 pathways were identified with impact factor >250. On the basis of the three criteria, 11 significant dysfunctional pathways were identified. Among them, Epstein-Barr virus infection was the most significant differential pathway. In conclusion, with the method based on attractor and crosstalk, significantly dysfunctional pathways were identified. These pathways are expected to provide molecular mechanism of ischemic stroke and represents a novel potential therapeutic target for ischemic stroke treatment.

Introduction

Ischemic stroke is one of the main causes of morbidity and mortality throughout the world (1-3). Generally considered as a heterogeneous and multifactorial disorder, ischemic stroke morbidity is high due to vast complications and lack of alternative treatments (4).

Identifying dysregulated pathways of ischemic stroke from large number of high-throughput experimental findings is a significant task to show molecular and functional targets existing (5). The differentially expressed genes (DEG) and pathways may be a promising candidate for the treatment of ischemic stroke. Recently, canonical studies demonstrated that biomarkers of ischemic stroke could be identified by gene expression patterns, which highlighted the dependency of the innate immune system through signaling pathways (4,6,7).

There are abundant pathways related to ischemic stroke in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database, which provides useful pathway topology information. Kauffman' attractor theory offers a new formal method to find one or more well-defined ensembles from large datasets whose statistical features matched those of real organisms and cells (8). A previous study (9) provided strong evidence that attractor was a formal approach that could leverage both the DEGs between cell phenotypes and existing pathway databases. We employed it to screen attractors within pathways from vast data of KEGG pathway database, in order to narrow down the number of correlated dysregulated pathways.

Screening differentially expressed pathways may provide an important theoretical basis for further ischemic stroke research. However, they invariably pay attention to potential function of single pathway and neglect the inherent interdependency inter-pathways. Pathway crosstalk is known as the phenomenon of cooperation or interaction among pathways. The construction of pathway crosstalk network (PCN) inter-pathways is conductive to understand the comprehensive interactions when ischemic stroke occur (10). Then a scoring scheme was applied to comprehensively identify these pathways based on attractor levels both of crosstalk inter-pathways and internal pathway effects.

In this study, we applied the method based on attractor and crosstalk to identify the dysregulated pathways which associated with ischemic stroke. Ultimately several significant dysfunctional pathways with strong interactions were identified. The identified pathways are suggested to provide molecular mechanisms for the treatment of ischemic stroke.

Materials and methods

Gene expression dataset. The transcription profile E-GEOD-16561 (4) was obtained through EMBI-EBI ArrayExpress database (11). The data included gene expression profiling of 39 ischemic stroke patients and 24 healthy controls. The platform was: A-MEXP-1172 - Illumina HumanRef-8 v3.0 Expression BeadChip.

Correspondence to: Dr Aijuan Liu, Department of Cardiology, Weifang People's Hospital, 151 Guangwen Street, Kuiwen, Weifang, Shandong 261041, P.R. China
E-mail: aijuan750@163.com

Data of the gene chip was read as previously described (12). The gene expression data was preprocessed by Linear Models for Microarray Data (LIMMA). Robust multi-array average (RMA) was applied to adjust the background and normalize the quantile data (13). We used a median polish and robust procedure for protecting against outlier probes (14) and estimating model parameters. The DEG were selected according to the threshold levels: $P \leq 0.01$, |log fold-change (FC)| ≥ 2.

Pathway data. Biological human pathways were downloaded from KEGG database (15) which provides copious pathway information (16,17). Pathways with the gene set size of >100 or <5 were filtered. The correct size cut off pathways were set up, 294 pathways were selected for downstream analysis.

Protein interaction data. The Retrieval of Interacting Genes (STRING; v9.0) were applied for screening the PPI (18). A total of 787,896 PPI sets were selected after removing self interactions.

Differential pathway analysis. Based on the attractor theory (8), this was applied for screening differential pathways which related to ischemic stroke from 294 KEGG pathways.

To identify these 294 pathways, GSEA-ANOVA approach was employed as a gene set enrichment algorithm (9). The 294 KEGG pathways with FDR <0.05 were identified as attractors. We computed the *F*-statistic for gene *i*:

where *MSSi* reflects the mean treatment sum of squares and used to captures the variation amount due to group-specific effects:

$$F^{(i)} = \frac{MSS_i}{RSS_i}$$

and RSS_i represents the residual sum of squares:

$$MSS_i = \frac{1}{K-1} \sum_{k=1}^{k} r_k \left[y_k^{(i)} - y^{(i)} \right]^2$$

where N is the samples number, and the overall mean is shown

$$RSS_i = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{j=1}^{r_j} \left[y_{jk}^{(i)} - y^{(i)} \right]^2$$

by:

$$y^{(i)} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{r_k} \sum_{j=1}^{r_k} y_{jk}^{(i)} \right)$$

Large values of the *F*-statistic mean a significant interactions with ischemic stroke-specific expression changes.

For pathway *P* comprise g_p genes, the *T*-statistic shows as the following form:

$$T_p = \frac{\left[\frac{1}{g_p} \sum_{i=1}^{g_p} F^{(i)} \right] - \left[\frac{1}{G} \sum_{j=1}^{G} F^{(j)} \right]}{\sqrt{\left(\frac{s_p^2}{g_p} \right) + \left(\frac{s_G^2}{G} \right)}}$$

where *G* reflects the total number of genes within a pathway annotation, the sample variances s_p^2 and s_G^2 are defined as follows:

$$s_p^2 = \frac{1}{g_p - 1} \sum_{j=1}^{g_p} \left(F^{(j)} - \frac{1}{g_p} \sum_{i=1}^{g_p} F^{(i)} \right)^2$$

$$s_G^2 = \frac{1}{G-1} \sum_{j=1}^{G} \left(F^{(j)} - \frac{1}{G} \sum_{i=1}^{G} F^{(i)} \right)^2$$

The resulting P-values of each pathway were adjusted using the Benjamini-Hochberg false discovery rate (FDR)-based approach of Benjamini-Hochberg. Differential pathways were selected with the criteria (P<0.05).

Crosstalk analysis. To analyze interactions between pathways, crosstalk analysis was applied to construct pathway crosstalk network (PCN) according to Li *et al* (10).

Background analysis. The PCN of control group was constructed. The weight of the background PCN represented the number of PPI sets. i) We used Fisher's exact test to evaluate the gene overlap between any pairs among 294 pathways (19). FDR were performed to adjust raw P-values (20). ii) We then counted all interactions of each pathway pair after removing genes which shared both pathways. iii) Background distribution of PPI established according to each pathway pair was estimated 1,000 times. iv) The one-sided Fisher's exact test was performed using the 2x2 contingency table on all pathway pairs. FDR BH procedures were performed to adjust P-values of Fishers exact test (20) and empirical P-value was calculated. v) All pathway pairs were chosen to construct the PCN with P<0.05.

Network of ischemic stroke. The network of ischemic stroke was constructed based on the crosstalk method.

One gene in the pathway has interactions with another pathway when it satisfies one of the two conditions: i) The Spearman correlation coefficients of each PPI set were calculated. The edge remained when the absolute value of different coefficients between them was >0.7. The value of weight between two pathways was defined by geometric mean of the absolute value. ii) The DEGs were selected according to the threshold levels: $P \leq 0.01$ |log fold-change (FC)| ≥ 1.

Important crosstalk pathways. The PCN was implemented by using topology analysis. The scores of pathways were defined as: Score = degree of ischemic stroke/degree of background (9).

Comprehensive analysis. Pathway analysis has become important in capturing clinical information. To explore the interactive relationship between two pathways, impact factor (IF) was calculated as: IF = outer x (1 - p) (10). Where outer is the degree of interactions between two pathways and p is the P-value of the attractor.

RP-value reflects the comprehensive identification ability within pathways or between pathways (3). RP-value = (rank inter/total) x (rank outer/total) (11).

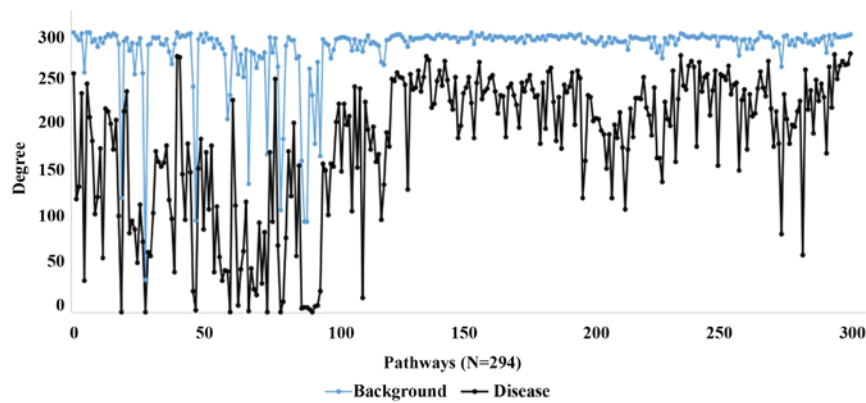Rank inter is the ranking of the attractor P-value and then rank outer reflects the ranking of interactions between

Figure 1. The crosstalk difference of background and ischemic stroke.

Table I. Twenty DEGs identified in the ischemic stroke.

DEG	logFC	P-value
Upregulated		
RGS2	1.0106	3.24E-14
PDK4	1.0079	7.54E-11
ARG1	1.6940	1.70E-09
IQGAP1	1.0330	9.65E-09
CRISPLD2	1.0690	5.39E-08
PADI4	1.0232	6.95E-08
MMP9	1.4304	1.36E-07
CSPG2	1.0757	4.93E-07
CA4	1.0994	3.35E-06
S100A12	1.2762	5.03E-06
ACSL1	1.0946	8.29E-06
FOLR3	1.0949	1.87E-05
AKAP7	1.1421	2.27E-05
LY96	1.1194	2.86E-05
ORM1	1.1837	0.00014
FCGR3B	1.1743	0.00066
APOBEC3A	1.1644	0.00092
OLFM4	1.0004	0.00124
FTHL3	1.0737	0.00463
Downregulated		
CCR7	-1.0838	5.70E-07

DEG, differentially expressed genes.

different pathways. Total here reflects the sum of the attractor P-value of inter and outer pathways.

Results

DEG in the ischemic stroke. According to the criteria (|log FC| ≥1; P≤0.01), a total of 20 DEGs were identified in ischemic stroke, of which 19 were upregulated and one was downregulated (Table I). These DEG might identify molecular alterations and provide diagnostic biomarkers for ischemic stroke.

Pathway crosstalk analysis. The PCNs of ischemic stroke and background were adjusted using gene expression profile of 24 controls and 39 ischemic stroke patients, respectively. Fig. 1 shows the crosstalk difference between control and ischemic stroke groups. In control group, the majority degrees of these 294 pathways were between 255 and 300. The ischemic stroke group showed significant difference with the background group. We obtained a unique degree from each significant pathway, which provides significant evidence to show the interactions between these pathways with ischemic stroke.

In this study, the degree reflects the strength of association of two pathways. That is, a large degree was indicative of strong interactions between pathways, while small degree indicates minimal interactions between pathways. The top three important pathways were pyrimidine metabolism (KEGG ID: 00240), HTLV-I infection (KEGG ID: 05166) and Epstein-Barr virus infection (KEGG ID: 05169).

Differential pathway analysis. A total of 68 differential pathways with P<0.05, and 39 were identified with P<0.01 (Fig. 2), indicating that these 68 pathways were significantly different in the ischemic stroke compared with normal network. Thus, some molecular alterations existed in the pathways among the development of ischemic stroke.

Comprehensive analysis of pathways. Pathway analysis has become important in capturing clinical information. Impact factor was calculated to explore the interactive relationship between two pathways. There were different impact factors from 0 to 272 as shown in Fig. 3 and these results indicated that there were differences between degrees of interactions in the inter-pathways. There were 17 pathways with IF-value >250, which were considered as important pathways in disease.

As mentioned above, RP-value was calculated to comprehensively explore these 294 pathways between pathways or within pathways. A total of 64 significantly enriched pathways existed with a threshold RP<0.05, and 28 pathways with a threshold RP<0.01 (Fig. 2).

At the criteria of IF-value >250, RP<0.01 and P<0.01, 11 significant pathways were identified as shown in Table II. Due to their dysfunctional expression and strong interactions, these pathways were considered to play core roles in the development of ischemic stroke. Among them, Epstein-Barr virus infection was the most significantly different pathway.

Table II. Significant pathways identified by Kauffman attractor, impact factor and RP-value.

KEGG ID	KEGG pathway	Attractor P-value	Impact factor	RP-value
05169	Epstein-Barr virus infection	1.12E-11	272	2.31E-05
05152	Tuberculosis	2.83E-06	257.9992701	0.000659447
05203	Viral carcinogenesis	5.18E-05	262.9863678	0.001353603
00230	Purine metabolism	0.004527077	267.7822164	0.00138831
05164	Influenza A	0.000100413	258.9739931	0.002221297
05016	Huntington's disease	0.000282903	263.9253137	0.002290712
05168	Herpes simplex infection	0.000412243	260.8924045	0.003077421
05161	Hepatitis B	0.006210218	262.3605025	0.003644315
04380	Osteoclast differentiation	0.000282903	253.9281427	0.004720255
04932	NAFLD	0.003514779	254.1037314	0.006363089
05010	Alzheimer's disease	0.0000356	252.9909911	0.002591513

NAFLD, non-alcoholic fatty liver disease; KEGG, Kyoto Encyclopedia of Genes and Genomes.
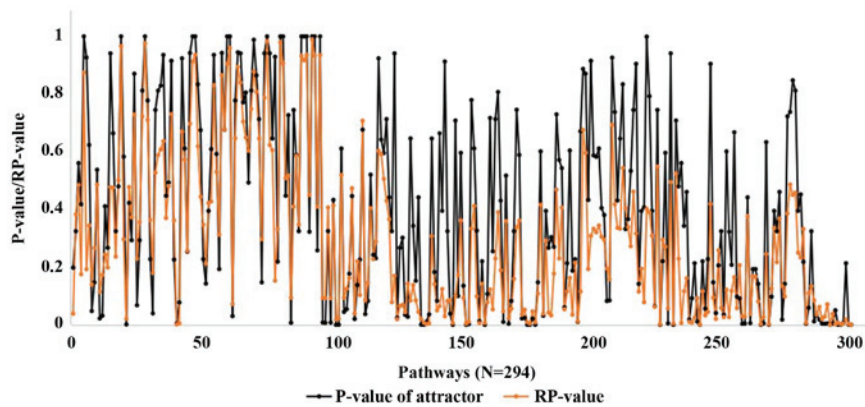

Figure 2. The 294 KEGG pathways were evaluated by Kauffman attractor and RP-value. KEGG, Kyoto Encyclopedia of Genes and Genomes.
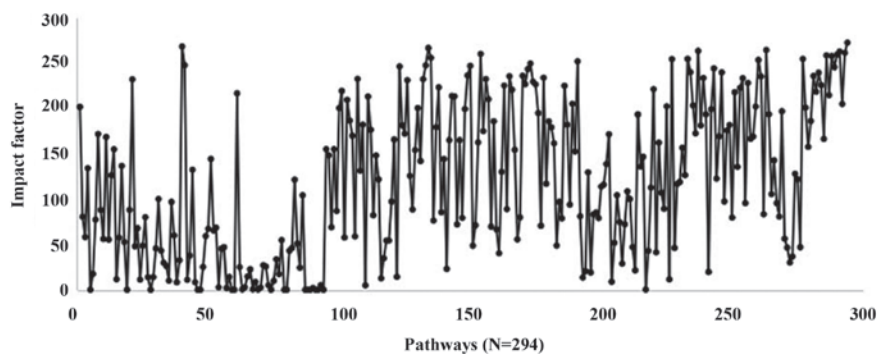

Figure 3. Interactions of inter-pathways were assessed by an impact factor.

Discussion

Identification of dysregulated pathways with a novel approach. Attractor theory is famous as a knowledge-driven analytical way that is not considered as a traditional KEGG method (8). Attract, an approach that can expand on the context to evaluate the genome-wide expression data in embryonic stem cells (9). Because of narrowing down the number of correlated dysregu-lated pathways, attract method with pathways will be more complete than traditional KEGG analysis.

In this study, 68 differential pathways (P<0.05) with statis-tically significant alteration were identified from 294 KEGG pathways in response to molecular mechanism and pathology process of ischemic stroke. We found that most of them were related to diseases, such as tuberculosis, Alzheimer's disease, measles and Huntington's disease. However, the studies on

integral influence on the system were absent. Fig. 2 shows that the variation trend of pathways were not consistent with that of RP-value. Therefore, crosstalk approach was employed to adjust the interactions between pathways. Pathways with large impact factor were taken into account having strong connection with other pathways. Interestingly, most of the 68 differential pathways had large IF-values, but 14 pathways did not (IF <190). Moreover, the RP-values of most of the 14 pathways were >0.05. The result indicate that the pathways which were screened by attractor were not exactly dysregulated and influential and the pathways with smaller values of impact factor and attractor level P<0.05 may have smaller effect and should be filtered.

The results indicated that attractor method may fail to identify potential functional interpretations between pathways due to its incomplete information on inherent interdependency inter pathway. Other pathway-identification methods that apply topological pathway information have also faced similar challenge (17). After calculating the interactions among inter-pathways by crosstalk, the novel method enhanced attractor to distinguish dysregulated pathways. Previous studies have reported there is more focus on the comprehensive identification of dysregulated pathways (5). Since ischemic stroke genetics field has made significant progress in identifying common genes that are confidently associated with ischemic stroke diagnosis, gene-based pathway aberrance analysis which combined attractor and crosstalk will be help in detecting pathways relating to ischemic stroke.

Evaluating the effect of dysregulated pathways. The RP-value was applied to evaluate the identified capacity of both inter-pathways and within pathways. The influential dysregulated pathways required are with attractor P<0.01, IF-value >250 and RP-value <0.01. In total, 11 important pathways were identified in ischemic stroke. We found that most of them were pathways related to diseases, such as influenza A, Huntington's disease, hepatitis B and non-alcoholic fatty liver disease (NAFLD). The pathway Epstein-Barr virus infection possessed minimum RP-value and maximal impact factor. Moreover, it was one of the most important crosstalk pathways. It is well-known that distinct forms of Epstein-Barr virus can contribute to the different infectious diseases and tumors (21). Therefore, the pathway Epstein-Barr virus infection was considered to be significantly important in representing novel potential therapeutic targets for ischemic stroke treatment.

References

1. Chen L, Luo S, Yan L and Zhao W: A systematic review of closure versus medical therapy for preventing recurrent stroke in patients with patent foramen ovale and cryptogenic stroke or transient ischemic attack. J Neurol Sci 337: 3-7, 2014.
2. Saada F and Antonios N: Existence of ipsilateral hemiparesis in ischemic and hemorrhagic stroke: Two case reports and review of the literature. Eur Neurol 71: 25-31, 2014.
3. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL and Chory J: RankProd: A bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics 22: 2825-2827, 2006.

4. Barr TL, Conley Y, Ding J, Dillman A, Warach S, Singleton A and Matarin M: Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling. Neurology 75: 1009-1014, 2010.
5. Han J, Li C, Yang H, Xu Y, Zhang C, Ma J, Shi X, Liu W, Shang D, Yao Q, et al: A novel dysregulated pathway-identification analysis based on global influence of within-pathway effects and crosstalk between pathways. J R Soc Interface 12: 20140937, 2015.
6. Tang Y, Xu H, Du X, Lit L, Walker W, Lu A, Ran R, Gregg JP, Reilly M, Pancioli A, et al: Gene expression in blood changes rapidly in neutrophils and monocytes after ischemic stroke in humans: A microarray study. J Cereb Blood Flow Metab 26: 1089-1102, 2006.
7. Grond-Ginsbach C, Hummel M, Wiest T, Horstmann S, Pfleger K, Hergenhahn M, Hollstein M, Mansmann U, Grau AJ and Wagner S: Gene expression in human peripheral blood mononuclear cells upon acute ischemic stroke. J Neurol 255: 723-731, 2008.
8. Kauffman S: A proposal for using the ensemble approach to understand genetic regulatory networks. J Theor Biol 230: 581-590, 2004.
9. Mar JC, Matigian NA, Quackenbush J and Wells CA: Attract: A method for identifying core pathways that define cellular phenotypes. PLoS One 6: e25445, 2011.
10. Li Y, Agarwal P and Rajagopalan D: A global pathway crosstalk network. Bioinformatics 24: 1442-1447, 2008.
11. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, et al: ArrayExpress - a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 35: D747-D750, 2007.
12. Gautier L, Cope L, Bolstad BM and Irizarry RA: Affy - analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20: 307-315, 2004.
13. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249-264, 2003.
14. Sásik R, Calvo E and Corbeil J: Statistical analysis of high-density oligonucleotide arrays: A multiplicative noise model. Bioinformatics 18: 1633-1640, 2002.
15. Kanehisa M and Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28: 27-30, 2000.
16. Rahnenfuhrer J, Domingues FS, Maydt J and Lengauer T: Calculating the statistical significance of changes in pathway activity from gene expression data. Stat Appl Genet Mol Biol 3: Article 16, 2004. doi: 10.2202/1544-6115.1055.
17. Hung JH, Whitfield TW, Yang TH, Hu Z, Weng Z and DeLisi C: Identification of functional modules that correlate with phenotypic difference: The influence of network topology. Genome Biol 11: R23, 2010.
18. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, et al: STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41: D808-D815, 2013.
19. Al-Shahrour F, Díaz-Uriarte R and Dopazo J: FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 20: 578-580, 2004.
20. Benjamini Y, Drai D, Elmer G, Kafkafi N and Golani I: Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125: 279-284, 2001.
21. Young KA, Herbert AP, Barlow PN, Holers VM and Hannan JP: Molecular basis of the interaction between complement receptor type 2 (CR2/CD21) and Epstein-Barr virus glycoprotein gp350. J Virol 82: 11217-11227, 2008.