PathNet: A tool for finding pathway enrichment and pathway cross-talk using topological information and gene expression data

Bhaskar Dutta, Anders Wallqvist, and Jaques Reifman

DoD Biotechnology HPC Software Applications Institute
Telemedicine and Advanced Technology Research Center
U.S. Army Medical Research and Materiel Command
Ft. Detrick, MD 21702, USA

1 Overview

The Pathway analysis using Network information (PathNet) algorithm, described in Dutta *et al.*, is described here. PathNet uses topological information present in pathways and differential expression levels of genes, obtained from microarray experiments, to identify pathways that are 1) significantly enriched and 2) associated in the context of gene expression data. In enrichment analysis, PathNet considers both the differential expression of genes and their pathway neighbors to strengthen the evidence that a pathway is implicated in the biological conditions characterizing the experiment. In addition, PathNet uses the connectivity of the differentially expressed genes among all pathways to score pathway contextual associations and statistically identify biological relations among pathways.

2 Datasets used in PathNet

The PathNet program for enrichment and contextual analysis require the following types of input data: 1) differential expression levels (i.e., direct evidence), 2) interactions between any pair of genes captured in an adjacency matrix (A), and 3) pathway information. We have included test data in the `PathNetData` package. The formats of each of the input data types are explained in detail in the following sections and users can load their input data following the descriptions provided in section 2.1.

To illustrate the utility of PathNet, we applied it to two microarray datasets measuring gene expressions of Alzheimer's disease (AD) patients. Both of these datasets were downloaded from NCBI Gene Expression Omnibus (GEO)

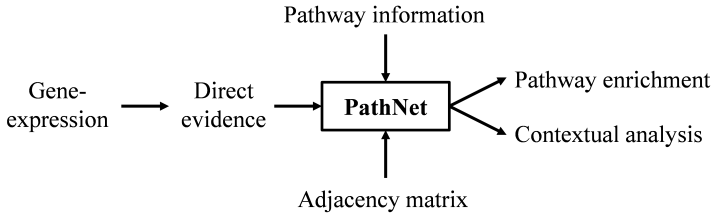

Figure 1: A schematic representation of inputs and outputs of PathNet.

database. The first dataset (GEO ID: GDS810) was used to examine the expression profile of genes from the hippocampal region of brain as a function of progression of the disease. Hence, this dataset was referred to as the *disease progression dataset*. The second dataset (GEO ID: GSE5281) was used to examine effect of AD in six different brain regions: entorhinal cortex (EC), hippocampus (HIP), middle temporal gyrus (MTG), posterior cingulate cortex (PC), superior frontal gyrus (SFG), and primary visual cortex (VCX). Hence, this dataset was referred to as the *brain regions dataset*.

The direct evidence, i.e., association of each gene with the disease, was calculated by comparing gene expression data in control patients with diseased patients. Here, we used t-test to identify the significance of association (p-value) of each gene with the disease. Other tests such as ANOVA and SAM can also be used to calculate significance of association. If multiple probes were present corresponding to a gene, the probe with the minimum p-value was selected. The negative log_{10} transformed p-value of the significance of association was used as direct evidence. In the *disease progression dataset*, we compared the gene expression from incipient, moderate, and severe samples with control samples. Similarly, for the *brain regions dataset*, we compared the gene expression from each of the six brain regions with corresponding control samples. Hence, for each gene, we generated three and six sets of direct evidences (corresponding to each comparison) from the *disease progression* and *brain regions datasets*, respectively.

To install the PathNet packages, start R and enter:

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("PathNet")
> biocLite("PathNetData")
```

The following commands load the PathNet packages, which include the PathNet program, and the PathNetData library containing direct evidences from *disease progression* and *brain regions datasets*, adjacency matrix and pathway information. If users want use a different microarray data or pathway database for PathNet analysis, they need to create the input files in the same format. Commands for loading new data files are provided in the section 2.1. Hence, before running the program, we provided an overview of the data formats.

```
> library("PathNet")
> library("PathNetData")
> data(disease_progression)
> head(disease_progression)
```

```
     Gene.ID  Incipient  Moderate    Severe
[1,]       2 0.99349061 0.1015926 0.7260152
[2,]       9 0.14573488 0.1774141 0.4710957
[3,]      10 0.19763134 0.4521152 0.2574638
[4,]      12 0.07513601 0.5673629 1.0125473
[5,]      13 0.42421679 0.5966392 0.8508698
[6,]      14 0.12775969 0.7701194 0.3739513
```

The first column contains the NCBI Entrez Gene ID. The next three columns contain direct evidences for three different stages of the disease. Similarly the *brain regions dataset* is loaded using the following commands:

```
> data(brain_regions)
> head(brain_regions)
```

```
     Gene.ID         EC       HIP        MTG         PC
[1,]       1 1.25148267 1.3556879 0.24003776 0.07740443
[2,]       2 2.81606027 1.8579534 2.60376656 0.19491037
[3,]       9 0.12906672 0.2529406 0.01858537 0.23066105
[4,]      10 1.29515536 1.9628964 0.09338370 0.98322431
[5,]      12 3.02561990 0.3385582 0.84806070 0.35756987
[6,]      13 0.08461512 0.2088286 0.08741969 0.52397816
            SFG       VCX
[1,] 0.59726351 1.4430438
[2,] 2.09178017 1.6951469
[3,] 0.03341433 0.5481369
[4,] 0.29468520 0.1155430
[5,] 2.62989462 0.2866884
[6,] 0.11419789 0.3515637
```

In the current version of the program, we used regulatory pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. KEGG Markup Language files containing pathway information were downloaded from the KEGG server (in November 2010) and were converted to text files. All the pathways were combined to create a pooled pathway. The connectivity information among genes in the pooled pathway was represented by the adjacency matrix (A). A is a square matrix with the number of rows (and columns) equal to number of genes in the pooled pathway. The element at the i^th row and j^{th} column of A indicates if there exists an edge from gene i to gene j (equal to 1 if there is an edge and equal to 0 otherwise). The row names (first element of each row) correspond to the gene IDs. The rest of the matrix is comprised of zeros and ones.

```
> data(A)
> # Load genes from direct evidence
> gene_ID <- brain_regions[,1]
> # Construct adjacency matrix
> A <- A [rownames(A) %in% gene_ID, rownames(A) %in% gene_ID]
> # Display a sample of the adjacency matrix contents
> A [100:110,100:110]
```

```
    X553 X554 X572 X575 X581 X595 X596 X598 X608 X613 X623
553    0    0    0    0    0    0    0    0    0    0    0
554    0    0    0    0    0    0    0    0    0    0    0
572    0    0    0    0    0    0    0    1    0    0    0
575    0    0    0    0    0    0    0    0    0    0    0
581    0    0    0    0    0    0    1    0    0    0    0
595    0    0    0    0    0    0    0    0    0    0    0
596    0    0    0    0    1    0    0    0    0    0    0
598    0    0    0    0    0    0    0    0    0    0    0
608    0    0    0    0    0    0    0    0    0    0    0
613    0    0    0    0    0    0    0    0    0    0    0
623    0    0    0    0    0    0    0    0    0    0    0
```

The pathway is loaded in this section. Each row in the pathway data represents an edge in the pooled pathway. The first column is the row index. The second and third columns denote the gene IDs connected by an edge in the pathway. The fourth column contains the name of the pathway where the edge is present.

```
> data(pathway)
> pathway[965:975,]
```

```
     id1  id2                   title
965 3309    0         Protein export
966    0    0         Protein export
967 5465 3158 PPAR signaling pathway
968 6256 3158 PPAR signaling pathway
969 6257 3158 PPAR signaling pathway
970 6258 3158 PPAR signaling pathway
971 5465  335 PPAR signaling pathway
972 6256  335 PPAR signaling pathway
973 6257  335 PPAR signaling pathway
974 6258  335 PPAR signaling pathway
975 5465  336 PPAR signaling pathway
```

2.1 Custom User Datasets

Users can create input text files to create data for the dataset formats described above. The PathNetData library is distributed with text files representations of these datasets to serve as a reference when creating new datasets for analysis.

Text file based datasets can be loaded using the following commands:

```
> # We use system.file to locate the directory with the
> # example text files from the PathNetData Package
> current <- getwd()
> setwd(system.file(dir="extdata", package="PathNetData"))
> # Begin loading datasets from the text files
> brain_regions <- as.matrix(read.table(
   file = "brain_regions_data.txt", sep = "\t", header = T))
> disease_progression <- as.matrix(read.table(
   file = "disease_progression_data.txt", sep = "\t", header = T))
> A <- as.matrix(read.table(
   file = "adjacency_data.txt", sep = "\t", header = T))
> pathway <- read.table(
   file = "pathway_data.txt", sep = "\t", header = T)
> # Change back to our previous working directory
> setwd(current)
```

3 Running PathNet for enrichment analysis

The following command runs the PathNet program for enrichment analysis:

```
> # Note we use a subset of evidence and a small number of
> # permutations for demonstration purposes
> results <- PathNet(Enrichment_Analysis = TRUE,
          DirectEvidence_info = brain_regions[1:2000,],
          Adjacency = A,
          pathway = pathway,
          Column_DirectEvidence = 7,
          n_perm = 10, threshold = 0.05)
```

Run-time of enrichment analysis program on a desktop computer (specifications: Intel Core i7 870, 8GB RAM, Windows 7 64-bit) was around 4 minutes. For enrichment analysis, the parameter `Enrichment_Analysis` should be set to `TRUE`. The default value of this parameter is set to `FALSE`, which causes PathNet to not perform enrichment analysis of the pathways and instead consider all of the specified pathways. The next parameter, `DirectEvidence_info`, provides the direct evidence data. Users always have to provide this data; else, the program will not proceed. In the `DirectEvidence_info`, the first column should always contain the gene ID. If multiple biological comparisons are performed from the same dataset, direct evidence values corresponding to each of the comparisons can be appended as separate columns. For example, in the brain regions dataset, gene expressions were analyzed for six different brain regions, i.e., EC, HIP, MTG, PC, SFG, and VCX. Hence, the `DirectEvidence_info` contains seven columns, where the first column contains the gene IDs and the

next six columns contain the direct evidences corresponding to six brain regions. In our example, as we used direct evidence from VCX brain region, `Column_DirectEvidence` was set to 7. The Adjacency and pathway parameters provide names of the adjacency matrix and pathway. These are organism-specific information obtained from the KEGG database. The `n_perm` parameter sets the number of permutations and the default value is 2000. The last parameter, `threshold`, is the p-value cutoff used to identify differentially expressed genes.

The results from the PathNet enrichment analysis are included in the `enrichment_results` and `enrichment_combined_evidence` list values PathNet returns. Significance levels of enrichment for each pathway from PathNet and the hypergeometric test are included in the `enrichment_results` matrix described in the format discussed below. The `enrichment_combined_evidence` matrix contains gene ID, direct, indirect, and combined evidence levels of genes present in the microarray data. Indirect evidences are calculated only for the genes that are present in the KEGG pathway and have at least one edge. For rest of the genes, indirect evidences are "NA" and the combined evidences are replaced by the direct evidences.

The following are the results generated by the demonstration PathNet program from the enrichment analysis:

```
> # Retrieve the first ten entrichment results
> results$enrichment_results[1:10,]
```

```
                                Name No_of_Genes
1                  Huntingtons disease          51
2          Basal transcription factors           4
3                   Parkinsons disease          30
4           Hedgehog signaling pathway          16
5                    Adherens junction          24
6                        Leishmaniasis           9
7  Epithelial cell signaling in Helico          19
8               Notch signaling pathway          10
9  Vasopressin-regulated water reabsor          17
10           Vibrio cholerae infection          17
   Sig_Direct Sig_Combi     p_Hyper p_Hyper_FWER
1          23        23 0.011173310    1.0000000
2           4         4 0.007482926    0.9727804
3          16        15 0.004717816    0.6133161
4           8         9 0.066668922    1.0000000
5          12        12 0.026169881    1.0000000
6           4         6 0.258758386    1.0000000
7          10        10 0.027884643    1.0000000
8           6         6 0.043073226    1.0000000
9           7         8 0.209756127    1.0000000
10          8         8 0.094690896    1.0000000
```

```
     p_PathNet p_PathNet_FWER
1  0.003627334      0.4715534
2  0.005377502      0.6990753
3  0.005776158      0.7509005
4  0.012676356      1.0000000
5  0.013427433      1.0000000
6  0.014987262      1.0000000
7  0.015514182      1.0000000
8  0.029029318      1.0000000
9  0.061702123      1.0000000
10 0.061702123      1.0000000

> # Retrieve the first ten combined evidence entries
> results$enrichment_combined_evidence[1:10,]

   gene_ID pDirectEvidence pIndirectEvidence
1        1      0.03605423                NA
2        2      0.02017684               0.4
3        9      0.28304995                NA
4       10      0.76640257                NA
5       12      0.51678700                NA
6       13      0.44507818                NA
7       14      0.22753571                NA
8       15      0.84119642                NA
9       16      0.03347231                NA
10      18      0.00595917                NA
   pCombinedEvidence
1         0.03605423
2         0.04696773
3         0.28304995
4         0.76640257
5         0.51678700
6         0.44507818
7         0.22753571
8         0.84119642
9         0.03347231
10        0.00595917
```

The enrichment results are sorted in increasing order based on the calculated values in the `p_PathNet_FWER` column. Results include pathway name (`Name`), number of genes of the pathway present in the microarray data (`No_of_Genes`), number of genes significant from direct evidence (`Sig_Direct`), number of genes significant from combined evidence (`Sig_Combi`), significance of enrichment from hypergeometric test (`p_Hyper`), family wise error rate correction of `p_Hyper` (`p_Hyper_FWER`), significance of enrichment from PathNet (`p_PathNet`), family wise error rate correction of `p_PathNet` (`p_PathNet_FWER`). We used the calcu-

lated p_PathNet_FWER values in our manuscript, and we recommend users to use p_PathNet_FWER as well.

4 Contextual analysis between pathways using PathNet

We introduced a new concept of contextual association between pathways, i.e., pathway connections that are influenced by differential gene expression of neighboring genes rather than just the static overlap of genes. Contrary to the static overlap, these associations are specific to and dependent on the biological conditions of the study. These calculations identify pathway pairs where differentially expressed genes linked to each other are present at a greater frequency than would be expected by chance alone.

For contextual analysis the PathNet program uses the same datasets, i.e., direct evidence, adjacency matrix from the pooled pathway, and pathway file. For contextual analysis, Contextual_Analysis should be set to TRUE. The rest of the inputs are also the same. Contextual analysis can be carried out in conjunction or independent of enrichment analysis by setting the Enrichment_Analysis to TRUE. While carrying out contextual analysis in conjunction with enrichment analysis, users have an option to select only the significant pathways for contextual analysis by setting use_sig_pathways parameter to TRUE. When use_sig_pathways parameter is set to FALSE, contextual analysis between all possible pathway pairs are calculated. Run-time of contextual analysis program depends on the number of pathways used for contextual analysis. When all pathways were used, the run-time of the program on a desktop computer (specifications: Intel Core i7 870, 8GB RAM, Windows 7 64-bit) was around 2 hours.

```
> # Perform a contextual analysis with pathway enrichment
> # Note we use a subset of evidence and a small number of
> # permutations for demonstration purposes
> results <- PathNet(Enrichment_Analysis = FALSE,
         Contextual_Analysis= TRUE,
         DirectEvidence_info = brain_regions[1:500,],
         Adjacency = A,
         pathway = pathway,
         Column_DirectEvidence = 7,
         use_sig_pathways = FALSE,
         n_perm = 10, threshold = 0.05)
```

The results of the contextual analysis are stored in two different matrices returned in the PathNet result list, named conn_p_value and pathway_overlap. The first matrix contains the results of the contextual association in the form of a square matrix, where the number of rows (and columns) is equal to the number of pathways. The element at the i^{th} row and j^{th} column denotes the

significance of contextual association of pathway i with pathway j. Similarly, the `pathway_overlap` matrix stores statistical significance of the overlapping genes between all pathway pairs in the form of a square matrix. This information is only based on the KEGG database and is not dependent on gene expression data. The hypergeometric test is used to estimate if the observed overlap is statistically significant. If `use_sig_pathways` parameter is set to `TRUE`, both contextual association and overlap analysis results are sorted in decreasing order of pathway enrichment. Otherwise, the pathways appear in alphabetical order of pathway names.

```
> # Show the first four rows and first two columns
> # of the contextual association from the
> # demonstration
> results$conn_p_value[1:4, 1:2]
```

```
                                ABC transporters
ABC transporters                               0
Acute myeloid leukemia                         1
Adherens junction                              1
Adipocytokine signaling pathway                1
                                Acute myeloid leukemia
ABC transporters                                   1.0
Acute myeloid leukemia                             0.0
Adherens junction                                  1.0
Adipocytokine signaling pathway                    0.4
```

```
> # Show the first four rows and first two columns
> # of the pathway overlap scores from the
> # demonstration
> results$pathway_overlap[1:4, 1:2]
```

```
                                ABC transporters
ABC transporters                               0
Acute myeloid leukemia                         1
Adherens junction                              1
Adipocytokine signaling pathway                1
                                Acute myeloid leukemia
ABC transporters                          1.000000e+00
Acute myeloid leukemia                    0.000000e+00
Adherens junction                         4.628949e-04
Adipocytokine signaling pathway           1.284468e-08
```

5 Reference

PathNet: A tool for pathway analysis using topological information. Dutta B, Wallqvist A, and Reifman J. Source Code for Biology and Medicine 2012 Sep 24;7(1):10.