Supplementary Document for the TVDBN2 Paper

By
Saptarshi Pyne, Yash Vanjani and Ashish Anand
(saptarshipyne01@gmail.com, vanjani.yash@gmail.com, anand.ashish@iitg.ernet.in)
February 1, 2018

Contents

Chapter 1

Datasets

The files mentioned in this chapter can be found at: `https://github.com/aaiitg-grp/TVDBN2/tree/master/datasets` . Please note that, when notation '(x, y)' is used in the context of a cell in a matrix, x and y represent the row and the column of the cell, respectively.

1.1 DREAM3 In Silico Network Challenge Datasets

1.1.1 Ds10

File: 'InSilicoSize10-Yeast1-nonoise-trajectories.tsv'. The columns represent the genes and the rows represent the time points. There are a total of 10 genes identified by $\{G1, G2, \ldots, G10\}$ and 21 distinct time points denoted by $\{0.0, 10.0, 20.0, \ldots, 200.0\}$. Ds10 contains 4 separate time series. Therefore, there will be 4 separate rows for each pair of (time point ID, gene ID), which represent the expression of the gene at the same time point in different time series.

The true network file: 'DREAM3GoldStandard_InSilicoSize10_Yeast1_TrueNet.RData'. Please start a R session in the same directory where you have saved the file and load the RData file as shown below.

```
> load('DREAM3GoldStandard_InSilicoSize10_Yeast1_TrueNet.RData')
> ls() ## list objects in the current workspace
[1] "true.net.adj.matrix"
```

The 'true.net.adj.matrix' R object is the adjacency matrix of the true network. It is a binary matrix of dimension (10×10). Here, $(G3, G1) = 1$ implies that there is a directed edge from $G3$ to $G1$ in the true network. On the other hand, $(G3, G2) = 0$ implies that there does not exist any directed edge from $G3$ to $G2$ in the true network.

1.1.2 Ds10n

File: 'InSilicoSize10-Yeast1-trajectories.tsv'. The columns represent the genes and the rows represent the time points. There are a total of 10 genes identified by $\{G1, G2, \ldots, G10\}$ and 21 distinct time points denoted by $\{0.0, 10.0, 20.0, \ldots, 200.0\}$. Ds10n contains 4 separate time series. Therefore, there will be 4 separate rows for each pair of (time point ID, gene

ID), which represent the expression of the gene at the same time point in different time series.

The true network file: 'DREAM3GoldStandard_InSilicoSize10_Yeast1_TrueNet.RData'. Please start a R session in the same directory where you have saved the file and load the RData file as shown below.

```
> load('DREAM3GoldStandard_InSilicoSize10_Yeast1_TrueNet.RData')
> ls() ## list objects in the current workspace
[1] "true.net.adj.matrix"
```

The 'true.net.adj.matrix' R object is the adjacency matrix of the true network. It is a binary matrix of dimension (10×10). Here, $(G3, G1) = 1$ implies that there is a directed edge from $G3$ to $G1$ in the true network. On the other hand, $(G3, G2) = 0$ implies that there does not exist any directed edge from $G3$ to $G2$ in the true network.

1.1.3 Ds50

File: 'InSilicoSize50-Yeast1-nonoise-trajectories.tsv'. The columns represent the genes and the rows represent the time points. There are a total of 50 genes identified by $\{G1, G2, \ldots, G50\}$ and 21 distinct time points denoted by $\{0.0, 10.0, 20.0, \ldots, 200.0\}$. Ds50 contains 23 separate time series. Therefore, there will be 23 separate rows for each pair of (time point ID, gene ID), which represent the expression of the gene at the same time point in different time series.

The true network file: 'DREAM3GoldStandard_InSilicoSize50_Yeast1_TrueNet.RData'. Please start a R session in the same directory where you have saved the file and load the RData file as shown below.

```
> load('DREAM3GoldStandard_InSilicoSize50_Yeast1_TrueNet.RData')
> ls() ## list objects in the current workspace
[1] "true.net.adj.matrix"
```

The 'true.net.adj.matrix' R object is the adjacency matrix of the true network. It is a binary matrix of dimension (50×50). Here, $(G2, G1) = 1$ implies that there is a directed edge from $G2$ to $G1$ in the true network. On the other hand, $(G2, G4) = 0$ implies that there does not exist any directed edge from $G2$ to $G4$ in the true network.

1.1.4 Ds50n

File: 'InSilicoSize50-Yeast1-trajectories.tsv'. The columns represent the genes and the rows represent the time points. There are a total of 50 genes identified by $\{G1, G2, \ldots, G50\}$ and 21 distinct time points denoted by $\{0.0, 10.0, 20.0, \ldots, 200.0\}$. Ds50n contains 23 separate time series. Therefore, there will be 23 separate rows for each pair of (time point ID, gene ID), which represent the expression of the gene at the same time point in different time series.

The true network file: 'DREAM3GoldStandard_InSilicoSize50_Yeast1_TrueNet.RData'. Please start a R session in the same directory where you have saved the file and load the RData file as shown below.

```
> load('DREAM3GoldStandard_InSilicoSize50_Yeast1_TrueNet.RData')
```

```
> ls() ## list objects in the current workspace
[1] "true.net.adj.matrix"
```

The 'true.net.adj.matrix' R object is the adjacency matrix of the true network. It is a binary matrix of dimension (50×50). Here, $(G2, G1) = 1$ implies that there is a directed edge from $G2$ to $G1$ in the true network. On the other hand, $(G2, G4) = 0$ implies that there does not exist any directed edge from $G2$ to $G4$ in the true network.

1.1.5 Ds100

File: 'InSilicoSize100-Yeast1-nonoise-trajectories.tsv'. The columns represent the genes and the rows represent the time points. There are a total of 100 genes identified by $\{G1, G2, \ldots, G100\}$ and 21 distinct time points denoted by $\{0.0, 10.0, 20.0, \ldots, 200.0\}$. Ds100 contains 46 separate time series. Therefore, there will be 46 separate rows for each pair of (time point ID, gene ID), which represent the expression of the gene at the same time point in different time series.

The true network file: 'DREAM3GoldStandard_InSilicoSize100_Yeast1_TrueNet.RData'. Please start a R session in the same directory where you have saved the file and load the RData file as shown below.

```
> load('DREAM3GoldStandard_InSilicoSize100_Yeast1_TrueNet.RData')
> ls() ## list objects in the current workspace
[1] "true.net.adj.matrix"
```

The 'true.net.adj.matrix' R object is the adjacency matrix of the true network. It is a binary matrix of dimension (100×100). Here, $(G2, G3) = 1$ implies that there is a directed edge from $G2$ to $G3$ in the true network. On the other hand, $(G2, G1) = 0$ implies that there does not exist any directed edge from $G2$ to $G1$ in the true network.

1.1.6 Ds100n

File: 'InSilicoSize100-Yeast1-trajectories.tsv'. The columns represent the genes and the rows represent the time points. There are a total of 100 genes identified by $\{G1, G2, \ldots, G100\}$ and 21 distinct time points denoted by $\{0.0, 10.0, 20.0, \ldots, 200.0\}$. Ds100n contains 46 separate time series. Therefore, there will be 46 separate rows for each pair of (time point ID, gene ID), which represent the expression of the gene at the same time point in different time series.

The true network file: 'DREAM3GoldStandard_InSilicoSize100_Yeast1_TrueNet.RData'. Please start a R session in the same directory where you have saved the file and load the RData file as shown below.

```
> load('DREAM3GoldStandard_InSilicoSize100_Yeast1_TrueNet.RData')
> ls() ## list objects in the current workspace
[1] "true.net.adj.matrix"
```

The 'true.net.adj.matrix' R object is the adjacency matrix of the true network. It is a binary matrix of dimension (100×100). Here, $(G2, G3) = 1$ implies that there is a directed edge from $G2$ to $G3$ in the true network. On the other hand, $(G2, G1) = 0$ implies that there does not exist any directed edge from $G2$ to $G1$ in the true network.

1.2 Escherichia coli (Ec) SOS DNA Repair Dataset (EcSos)

File: 'EcSos_expt_1_2.tsv'. The columns represent the operons and the rows represent the time points. There are a total of 8 operons identified by {uvrD, lexA, ..., polB} and 50 distinct time points denoted by $\{0, 6, 12, \ldots, 294\}$. EcSos contains 2 separate time series. Therefore, there will be 2 separate rows for each pair of (time point ID, operon ID), which represent the expression of the operon at the same time point in different time series.

The true network file: 'EcSos_TrueNet.RData'. Please start a R session in the same directory where you have saved the file and load the RData file as shown below.

```
> load('EcSos_TrueNet.RData')
> ls() ## list objects in the current workspace
[1] "true.net.adj.matrix"
```

The 'true.net.adj.matrix' R object is the adjacency matrix of the true network. It is a binary matrix of dimension (8×8). Here, (lexA, uvrD) = 1 implies that there is a directed edge from 'lexA' to 'uvrD' in the true network. On the other hand, (umuDC, uvrD) = 0 implies that there does not exist any directed edge from 'umuDC' to 'uvrD' in the true network.

Bibliography

[1] Precision, Recall and F1 measure. URL https://github.com/dice-group/gerbil/wiki/Precision,-Recall-and-F1-measure. Section 'Dividing by 0'. Data Science Group at Paderborn University, Germany. Last accessed: Oct 13, 2017.

[2] JSON. URL http://www.json.org/. Last accessed: Oct 31, 2017.

[3] Respiratory viral dream challenge. URL https://www.synapse.org/#!Synapse:syn5647810/wiki/399110. Last accessed on Aug 15, 2017.

[4] Amr Ahmed and Eric P Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106 (29):11878–11883, 2009.

[5] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. CRC press, 2006.

[6] ARTIVA. ARTIVA package. URL https://cran.r-project.org/package=ARTIVA. Last accessed: Oct 13, 2017.

[7] Sophie Lebre and Gaëlle Lelandais. Function ARTIVAnet()'s reference manual in ARTIVA package version 1.2.3. URL https://cran.r-project.org/web/packages/ARTIVA/ARTIVA.pdf#Rfn.ARTIVAnet.1. Last accessed: Nov 2, 2017.

[8] Sophie Lèbre, Jennifer Becq, Frédéric Devaux, Michael PH Stumpf, and Gaëlle Lelandais. Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, 4(1):130, Sep 2010. doi: 10.1186/1752-0509-4-130. URL https://doi.org/10.1186/1752-0509-4-130.

[9] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0.

[10] Kevin Ushey, Jonathan McPherson, Joe Cheng, Aron Atkins, and JJ Allaire. *packrat: A Dependency Management System for Projects and their R Package Dependencies*, 2016. URL https://CRAN.R-project.org/package=packrat. R package version 0.4.8-1.