Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Jiani Hu Date

A new method of network-guided dimension reduction

By

Jiani Hu

Master of Public Health

Biostatistics

_____

Tianwei Yu

Thesis Advisor

_____

Hao Wu

Thesis Reader

A new method of network-guided dimension reduction

By

Jiani Hu

B.A.

Wuhan University

2014

Thesis Committee Chair: Tianwei Yu, Ph.D.

Thesis Committee Member: Hao Wu, Ph.D.

An abstract of

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Public Health
in Biostatistics

2016

Abstract

High throughput technologies transform the interest of single-gene and protein studies to the genome scale. Given the size and complexity of high-throughput data, dimension reduction is often used to simplify and visualize data. However, one obstacle to effective dimension reduction of complex gene expression matrix is the loss of true biological information caused by the pervasive correlation and interference of high measurement noise. To address these issues, we tried to incorporate existing knowledge as represented by known biological networks, by developing a new network-guided dimension reduction method. The effectiveness of this method was tested in both simulations and real gene expression data. The simulation results show the power of detecting major signal in large-scale network is high. The results from the real data analysis show the first few dimensions found by the method are dominated by meaningful biological signals. The network-guided dimensional reduction is an effective method that captures the main signals contained in the large data matrix.

By Jiani Hu

A new method of network-guided dimension reduction

By

Jiani Hu

B.A.

Wuhan University

2014

Thesis Committee Chair: Tianwei Yu, Ph.D.

Thesis Committee Member: Hao Wu, Ph.D.

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Public Health
in Biostatistics

2016

Acknowledgements

I am extremely grateful for my amazing thesis advisor, Dr. Tianwei Yu, whose encouragement, guidance, and support has made this thesis possible.

I am also thankful for Dr. Hao Wu for taking the time to read my thesis.

Finally, I would like to thank my parents, for their love and support throughout my education.

Table of Contents

1. Introduction

Biological science today is quite different from what was performed in the past. High throughput technology, which has inexpensive production of large volumes of data as the primary advantage over conventional techniques, combined with information science, is regarded as a revolution in biology. (Metzker, 2010) With the development of high throughput technologies, the interest from single genes and proteins has been shifted to studying thousands of genes and proteins together. (Boris Kholodenko, Michael B. Yaffe, & Walter Kolch1, April 2012)

Synergistic relations exists widely in the biological system. It arises from the concerted action of multiple factors producing an amplification or cancelation effect compared with individual actions alone. (Jean-YvesTrosset1 & and Pablo Carbonell2, October 2013) It appears in different levels of biological domains, for example the protein activity involves chemical reaction, polypharmacology, and metabolic pathway complementarity. (Jean-YvesTrosset1 & and Pablo Carbonell2, October 2013) More specifically, in higher organisms, cells differentiate for specialized functions, such as immune cells for disease prevention, red blood cell for oxygen transformation *etc*. Much of a cell's activity is organized as a network of interactions between different functional units: such as sets of genes co-regulated to respond to different experimental conditions. (Jan Ihmels, 22 July 2002,) The various types of interactions merge together and form a biology network. The representations of intracellular biological network normally consider the biological molecules as nodes and the interactions between nodes as edges. Some examples of biological

networks include metabolic networks, cell signaling networks, kinases-substrate networks, gene regulatory networks, and protein-protein interaction networks.

An important concept in network analysis is modularity. It is an abstract concept that seeks to capture the various levels and kinds of heterogeneity found in organisms, and it is considered a fundamental aspect of biological organization. (Günter P. Wagner*, December 2007) A network of interactions is modular if it can be subdivided into relatively autonomous, internally highly connected components.(Günter P. Wagner*, December 2007) Modularity exists in different levels of biological networks, including protein-protein interaction networks, gene regulatory networks, cells networks, and so on. (Jan Ihmels, 22 July 2002,) Modularity has emerged as a rallying point for research in developmental and evolutionary biology (and specifically evo-devo), and in molecular systems biology. (Günter P. Wagner*, December 2007) Modularity can be used to analyze new experimental data in biology. Algorithms can be used to find potential direct or indirect connections between proteins, genes, or both, while maintaining a modular structure.

The main challenge of studying biology network is to integrate theoretical properties and experimental data, understand and model in quantifiable terms the topological and dynamic properties of various networks that control the behavior of the cell. (Oltvai‡, FEBRUARY 2014) In order to conduct the analysis, two components are necessary - one is to getting the proteins and pathways organized in a network; another one is identifying the connections between nodes. (Boris Kholodenko et al., April 2012) Large amounts of existing biological knowledge are coded into biological networks. Such networks can be used in various ways. For example, the shortest path approach can be applied to predict key regulators and identify unknown

genes. (Günter P. Wagner*, December 2007) When the network is unknown, graph clustering is an often used to find related proteins, which is based on some similarity measure defined for the data elements. The approach of analyzing protein-protein network is to first identify the central nodes (hubs), which are characterized by a higher than average number of interactions. Then the clustering of proteins and classification is performed.

A challenge of analyzing a high-throughput datasets is that the dimension of dataset is too large. For example, in a genome-wide association (GWA) study, a million or more single nucleotide polymorphism (SNPs) can be profiled. (Wong, 2004) Many normal statistical techniques are not appropriate to use in analyzing a high dimensional dataset. The dimensionality of gene expressions needs to be reduced for visualization, as well as prior to some regression and other types of analyses. Another problem for genome analysis is that only a small number of genes profiled are expected to be associated with the response variables while the majority of the genes are 'noises'. Applicable methods to solve these problems are variable selection, dimension reduction and hybrid approaches.(Hui ZOU, June 2006) Variable selection focuses on finding out the genes that can represent the whole network, while the dimension reduction method searches for a small number of 'meta-genes', which are often linear combinations of all genes. (Dai, January 2011) The dimension reduction can be used to figure out the potential relationship between genes. It also has the ability to generate more reliable estimates by reducing noises.(Hui ZOU, June 2006) Dimensionality reduction could be seen as another kind of feature extraction, which refers to identifying the salient aspects or properties of data. (Kramer, February 1991)

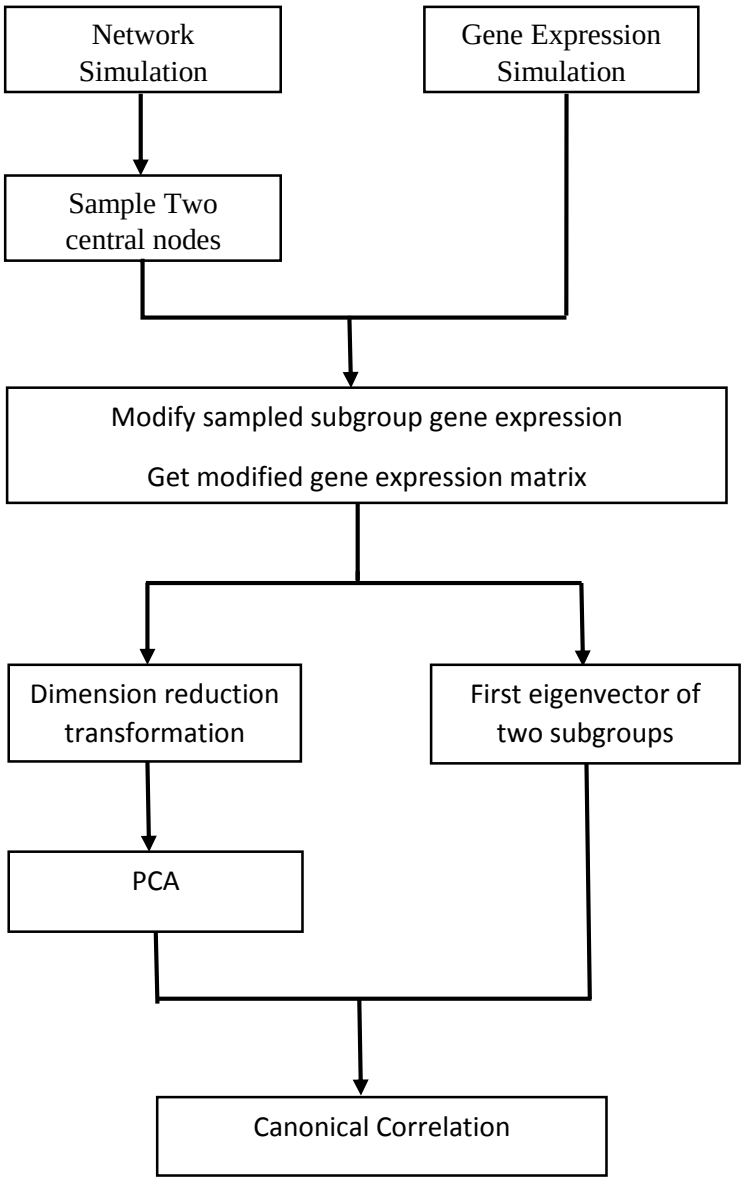
Based on natural features of gene expression data and known biological network: (i) gene expression data normally has much smaller sample size compared with the

number of genes; (ii) the noises in gene expression need to be removed. Therefore, we attempt to conduct dimension reduction while utilizing existing network information, following the general thinking of Principle Component Analysis (PCA). PCA is a classic dimension reduction approach that develops linear combination of gene expressions, called principal components (PCs). (Dai, January 2011) PCs have several useful features for studying high-throughput data: (i) all the PCs are orthogonal to each other. They can effectively explain variation of gene expressions. The collinearity problem among gene expression can be solved with this property. (ii) The dimension of PCs is much lower than the minimum of number of genes and sample size. The small sample size of gene expression data makes the dimension of PCs low, which alleviates the high-dimensionality problem. (iii) The first few PCs can explain the most variation of genes expression. (iv) Any linear combination of genes can be written into the linear combination of PCs. (Dai, January 2011) With PCA, the high-dimensional expression data can be projected into a low dimensional subspace, facilitating visualization. Though PCA is the most wildly used dimensional reduction method, it still has several shortcomings. It can only deal with linear input-output mapping and can't separate independent sub-signals from their linear mixtures. (Joutsensalo, 1994). By taking into account the network information, our method is more robust. We name the method Network Guided Dimension Reduction (NGDR).

2 Methods

The goal of this thesis is to find a low-dimensional matrix, which can maintain the main information from a high dimensional data after network guided dimensional reduction. A simulation process was conducted to created known protein network and

test the effectiveness of NGDR. Then a validation test in real yeast cell gene data was conducted to test the practicality of this method.



2.1 Scale-free gene network simulation

Though the random network model explains part of the result in network analysis in past few decades, a series of recent findings indicate that the topological properties of real cellular network don't follow the random network model. (Oltvai‡, FEBRUARY

2014) The analysis results from 43 different organisms' metabolic networks show that cellular metabolism has a scale-free topology instead of random network. (Jeong, October 2000), (Wanger, January 2001) For many social and biological networks, the number of nodes with a given degree follows a power law, which is the probability a chosen node with exactly k links follows $P(k) \sim k^{-\gamma}$. γ represents the degree of exponent with its normal value ranging between 2 and 3 for most of biological networks. (Barabasi & Albert, 1999) The main property of cellular networks is that they are scale-free, which is created by the power law. The degree distribution of nodes in a scale-free network is highly non-uniform, which means most nodes have only a few links while a few central nodes own a very large number of links. (Barabasi & Albert, 1999)

The simulation in this study is to get an approximate high-dimensional gene network $\boldsymbol{G}_{p\times n}$ with p genes and n conditions. In order to make the simulated gene network more close to the real gene net, the gene network is required be a scale-free network, which means the degree distribution of this network follows a power law. The simulation progress is also an evolutionary of scale-free networks, which involves a preferential attachment mechanism. (Barabasi & Albert, 1999) The nodes with more connections has a higher probability to grab links with newly-added nodes and finally turn into hubs.

In this research, a Barabási-Albert model was used to generate random scale-free networks with the preferential attachment mechanism. A parameter m was set in this progress to control the connection density in each simulation. The larger m is, the more connections would be created in the network.

2.2 Gene expression simulation

In our study, a protein network with two hubs modified with high covariance with neighboring nodes is simulated. Multivariate normal density and random deviates function are applied to generate gene expression matrix. After this process, a random normal distributed protein expression matrix with no interactions was created. All vectors are standardized to have mean 0 and covariance 0 with each other. Two nodes with degree between M1 and M2 (pre-specified parameters) were randomly sampled and are treated as two hubs, around which two sub-graphs with high interactions will be created. The shortest path between these two selected nodes are larger than 5, which means these two nodes do not have influence on each other in gene expression matrix. The nodes which connect with the selected nodes within two steps would be found as components of the two sub-graphs centered at the two hubs. For each sub-graph, the gene expression matrix was modified with specific covariance. The parameters and their values in the simulation were listed in Table 1. Net density will influence the size of sampled groups. The higher net density will cause more edges in network and a larger sampled sub-group. Subnet covariance decides the strength of set signal to each sub group. The higher subnet covariance means stronger signal in the subgroups. For each set of parameters, the simulation was conducted for 10 times.

Table 1 Parameters in simulations

Parameters	**Values**
Sample size	25, 50, 100, 150, 200
Net density parameter (influence subnet size)	1, 1.5, 1.7
Covariance of expression within sub-graph	0.5, 0.7, 0.9
M1	7
M2	40

2.3 Network Guided Dimension Reduction

The obstacle to reduce the dimension of complex gene expression matrix is the loss of main information along with the interference effect of noise. To address these issues, we tried to find a serious of orthonormal vectors consecutively by incorporating network information.

$\boldsymbol{g}_i$ is used to denote the expression vector of the i^{th} gene, $\boldsymbol{l}_i$ donate the projection length of the i^{th} gene. Given $\mathbf{B}=(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_k)$, where the $\boldsymbol{\beta}$**'s** are unit vectors orthogonal to each other, and k is the number of dimensions of the subspace,

$$l_i=\sqrt{(\boldsymbol{g}_i'\boldsymbol{\beta}_1)^2+(\boldsymbol{g}_i'\boldsymbol{\beta}_2)^2+\cdots+(\boldsymbol{g}_i'\boldsymbol{\beta}_k)^2}$$

The orthonormal vectors are such that,

$$\beta_1 = argmax(\beta_1' D'D\beta_1 - \beta_1' A'A\beta_1), s.t. \|\beta_1\| = 1$$

$$\beta_{i+1} = argmax(\beta_{i+1}' D'D\beta_{i+1} - \beta_{i+1}' A'A\beta_{i+1}),$$

$$s.t. \|\beta_{i+1}\| = 1 \; and \; \beta_{i+1} \perp \beta_k, \forall k = 1, \ldots, i$$

Where A is the matrix of differences between vectors of linked nodes, and D is the matrix of difference between vectors of nodes that are far away on the network. The objective function means that we try to maximize the distance between unlinked nodes after projection, while minimizing the distance between linked nodes.

Using the Lagrange multiplier,

$$L_D = \beta_1' D'D\beta_1 - \beta_1' A'A\beta_1 + k(1 - \beta_1'\beta_1)$$

Set the derivative to zero, we have

$$D'D\beta_1 - A'A\beta_1 = k\beta_1$$

$$(D'D - A'A)\beta_1 = k\beta_1$$

Hence the solution is simply the PCs of the matrix $D'D - A'A$.

In order to get the A matrix, the difference between connected nodes were calculated and merged into one matrix. The D matrix was formed by sampling 1,000 pairs unconnected gene nodes and calculating the difference of these paired gene nodes.

2.4 Sparse Principal component analysis (SPCA) and canonical correlation calculation

After getting matrix $E = D'D - A'A$, the sparse principal component analysis was conducted to get PCs of matrix E. SPCA is built on PCA using the *lasso* (*elastic net*) to produce modified principal components with sparse loadings. (Hui ZOU, June 2006) SPCA gives exact PCA results when its sparsity (lasso) penalty term vanishes. The canonical correlation between the first three PCs of E and the first eigenvector of sub-graph is calculated to describe how well the first three PCs can capture the information of the sub-graph. (Weenink, 2003).

For each set of parameters (Table 1), the simulation analysis was repeated for 10 times. The mean canonical correlation between two sub-graphs' first eigenvectors and the first three PCs were calculated and record for each set of parameters simulations.

2.5 Test on yeast cycle data

The Spellman yeast cell cycle data was used to test our method. The real data contains four time-series, each covering roughly two cell cycles. The gene array has 73

conditions and 6178 genes. A number of cell-cycle related genes have strong periodicity in expression. Due to the difference in phase, the cell cycle related genes cannot be easily summarized by clusters.

NGDR was applied to get the cell cycle related genes expression and filter out noises information from other genes. Gene expression data may or may not satisfy the normality assumption of PCA. First of all, the gene expression array was standardized to achieve mean 0 and standard deviation 1 for each gene. The difference of gene expression between all connected genes were calculated and used to form matrix A. The unconnected genes were defined as the two nodes with the shortest path between them larger than 5. One thousand pairs of unconnected genes were randomly sampled. The gene expression difference between each pair of unconnected genes were calculated and used to form matrix D. $D'D - A'A$ was calculated as described in section 2.4. The PCs matrix of $D'D - A'A$ was calculated by sparse principal component analysis. The loading for each gene on the first few PCs was conducted. The ones with loading larger than 0.6 were selected and regarded as the ones whose information was captured by the corresponding PCs. These captured genes were classified by Gene Ontology (GO) categories by GOstats, a Bioconductor package in R. The structured description of known biological information at different levels of granularity greatly facilitate the interpretation of high-throughput technologies output. (Maere, Heymans, & Kuiper, 2005) GO consists of three hierarchically structured vocabularies to describe gene products in terms of their associated biological processes, molecular functions and cellular components. (Maere et al., 2005) Gene products are annotated to one or several nodes in each hierarchy. (Maere et al., 2005) Each gene is classified by whether or not it has been selected and whether or not annotated at a particular term. (Falcon & Gentleman, 2007)

3 Results

3.1 Simulation results

The simulation results show the mean detected canonical correlation between the first three PCs of matrix E and the first eigenvectors of the selected sub-networks' expression matrix.

Figure 1 and Figure 2 shows the detected mean correlation of two selected sub-networks from 10 repeated simulations with the same set of simulation parameter. The networks parameters were set as described in section 2.2. Each sub graph illustrates the canonical correlation result of networks with the same net density and hubs covariance, but different sub-network size. Each column contains the simulated networks with the same set correlation but different net density (range from 1 to 1.5 and 1.7). Each row contains the graphs with the same net density but different set correlation. In general, all the mean detected correlation of 10 repeats are high. (Figure 1 & Figure 2)

In Figure 1, for each sub graph, the range of detected correlation is small. The sub-networks with higher set correlation have smaller range of detected correlation compared with the ones with lower correlation. No matter the set correlation is high or not, the increase of sample size will make the mean detected correlation higher and the standard deviation smaller.

For each column, when the set correlation is high (0.9), the increase of net density will cause the detected correlation to be more concentrated. That is to say each trial has the trend to get the same detected correlation. For the networks with lower set correlation, the increase of sample size still caused an obvious decrease in the

standard deviation. But there is one exception in the network with set correlation 0.5, net density 1.7 and sample size 100. Each row represents the simulated protein network with the same net density but different set correlations. For each row, the increase of set correlation caused the detected correlation to be higher and with smaller spread.

Figure 2 displays the detected correlation of another selected sub-graph. These two sub-graphs are from the same simulated network, but they are independent with each other. In general, the mean detected correlations of are still high. The detected correlation in Figure 2 has the same trend with the one in Figure 1.

3.2 Testing the method on yeast cycle data

Result from the real data shows the genes whose information was captured by the first three PCs of matrix E that is generated from the real yeast cycle data using NGDR. Figure 3 shows the factor score captured by the first PC. There is a slightly periodical signal in the first time zone (alpha factor). This signal is not strong enough to be counted as a periodic signal in the gene expression during the first time period. The factor score captured by the first PC during the second period (cdc15) shows strong oscillation, which is known in advance to be caused by experimental artifacts. And the factor score of the third time period (cdc28) and the fourth time period (elu) are weak and around 0. The oscillation of cdc15 gene expression is noise, which is too strong to let the first PC capture the main information of this gene network. Since the artifact dominated the first PC, the factor score captured by the second PC was reported in Figure 4. The periodical signal during the whole time zones illustrates there is a cell cycle module, which has periodical expression feature, which dominated the second PC. Through the signal repeats in different time zones were not

exactly the same, the result is strong enough to lead a conclusion that there were a specific set of genes involved in cell cycle and captured by PC2.

There were 82 genes in total that had loadings larger than 0.6 on the second PC. These 82 genes were treated as one module. The GOHyperGParams result shows 59 out of 82 genes are cell cycle genes who have the periodical repeat during each cell cycle with p-value 1.5×10^{-32}. The classification GOBPID of these 59 genes are GO: 0007049. Other kinds of biosynthetic and catalytic processing gene were also related in this module. (Table 2)

The third PC was also taken into consideration. Figure 5 shows the factor score captured by the third PC. The signal also displayed periodicity. There were 78 genes in total had loading larger than 0.6 on the second PC. The GOHyperGParams result shows 49 out of 78 genes are cell cycle genes who have the periodical repeat during each cell cycle with p-value 2.096077×10^{-24}. (Table 3)

4 Discussion

Further Explanation of result

The simulation analysis result shows the mean detected correlations are all close to or above 0.9. This proves the $D'D - A'A$ transformation can help re-capture the main signal from original network. The first three PCs efficiently captured the main information from the expression matrix. The reason of higher net density can improve detected correlation is that the high density network with more edges will make the selected sub-network larger than the lower density network. Larger sub-networks make the signal stronger and easier to be detected. Meanwhile, the higher

set correlation makes the given signal in sub-network higher and also end up with higher detected correlations.

The analysis of the yeast cell data illustrates that NGDR captures important information from real data. For the result from the second PC, fifty-nine genes out of eighty-two genes whose signal were captured by the second PC were classified as cell cycle related genes. For the result of the third PC, 49 genes out of 79 genes were cell cycle genes. Both percentages are high, which proved the efficiency of this method is high. The reason why the cell cycle signal was captured by the second PC instead of the first PC is that the oscillation signal is too strong and therefore dominated the first PC. The signal result captured in this study are exactly the same as the one captured in a precious study finished by Tianwei Yu. (Yu, 2010) The consistence of these two studies clearly prove the method in this study is valid.

Advantages and Limitations

PCA is a commonly used technique in high-dimensional data visualization. The result of PCA is straight-forward to interpret. It can be summarized by a few parameters. Beside the straight-forward interpretability and computational efficiency, the linear PCA is robust against outliers. (Hill, September 1973) Though the use of PCA is widespread, PCA limited by its reliance on second-order statistics. The Principal components can be highly statistically dependent, in which case, PCA will fail to find the most compact description of the data. PCA needs larger-dimensional representation than the nonlinear alternatives when there is nonlinear dependencies in PCA. (Leen, 1997) Another limitation of PCA is that PCA doesn't have the ability to separate highly correlated signals.

A limitation of this study is that the difference between net density group is small, which makes its influence to detected correlation not as obvious as expected. But the

challenge of increasing net density is that the difficulty of sampling two unconnected nodes with specific degree will increase while the net density increase. This is decided by the property of scale-free network- when the density increases, the nodes with high degree tend to be connected with each other. Further work can focus on simulating network closer to the real protein network with appropriate net density.

Final conclusion

This new network-guided dimension reduction method is efficient to capture hidden signals based on the knowledge encoded in existing networks.

References

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509-512.

Boris Kholodenko, Michael B. Yaffe, & Walter Kolch1. (April 2012). Computational Approaches for Analyzing Information Flow in Biological Networks. *Science Signaling, 5*(220). doi:10.1126/scisignal.2002961

Dai, S. M. Y. (January 2011). Principal component analysis based methods in bioinformatics studies. *BRIEFINGS IN BIOINFORMATICS, 12*, 714-722. doi:10.1093/bib/bbq090

Falcon, S., & Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics, 23*(2), 257-258. doi:10.1093/bioinformatics/btl567

Günter P. Wagner*, M. P. a. J. M. C. (December 2007). The road to modularity. *NATURE REVIEWS | GENETICS, 8*, 921-931.

Hill, J. H. F. a. J. W. T. a. M. (September 1973). A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE*.

Hui ZOU, T. H., and Robert TIBSHIRANI. (June 2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics, 15*, 165-286. doi:10.1198/106186006X113430

Jan Ihmels, G. F., Sven Bergmann, Ofer Sarig, Yaniv Ziv & Naama Barkai. (22 July 2002,). Revealing modular organization in the yeast transcriptional network. *nature genetics, 31*, 370-377. doi:10.1038/ng941

Jean-YvesTrosset1, & and Pablo Carbonell2. (October 2013). Synergistic synthetic biology: units in concert. *Frontiers in BIOENGINEERING AND BIOTECHNOLOGY, 1*. doi:10.3389/fbioe.2013.00011

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N.& Barabasi, A.-L. . (October 2000). The large scale organization of metabolic networks. *Nature, 407*, 651-654.

Joutsensalo, J. K. a. J. (1994). Generalizations of Principal Component Analysis, Optimization Problems, and Neural Networks. *Pergamon, 8*, 549-562. doi:0893-6080/95

Kramer, M. A. (February 1991). Nolinear Principal Component Analysis Using Autoassociative Neural Networks. *AIChE Journal, 37*.

Leen, N. K. T. K. (1997). Dimension Reduciton by Local Principal Component Analysis. *Communicated by Garrison Cottrell*, 1493-1516.

Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics, 21*(16), 3448-3449. doi:10.1093/bioinformatics/bti551

Metzker, M. L. (2010). Sequencing technologies —the next generation. *NATURE REVIEWS | GENETICS, 11*, 31-46.

Oltvai‡, A.-L. B. Z. N. (FEBRUARY 2014). NETWORK BIOLOGY: UNDERSTANDING THE CELL'S FUNCTIONAL ORGANIZATION. *NATURE REVIEWS | GENETICS, 5*, 101-113.

Wanger, A. F., D. A. . (January 2001). The small world inside large metabolic networks. *The Royal Society*, 1803-1810. doi:doi10.1098/rspb.2001.1711

Weenink, D. (2003). Canonical Correlation Analysis. *Institute of Phonetic Sciences, University of Amsterdam, Proceedings., 25*, 81-99.

Wong, L. (2004). The Practical Bioinformatician. *World Scientific Publishing Company*.

Yu, T. (2010). An exploratory data analysis method to reveal modular latent structures in high-throughput data. *BMC Bioinformatics, 11*, 440. doi:10.1186/1471-2105-11-440

5 Appendices

A Figures & Tables

Figure 1 Detected Correlation of hub one

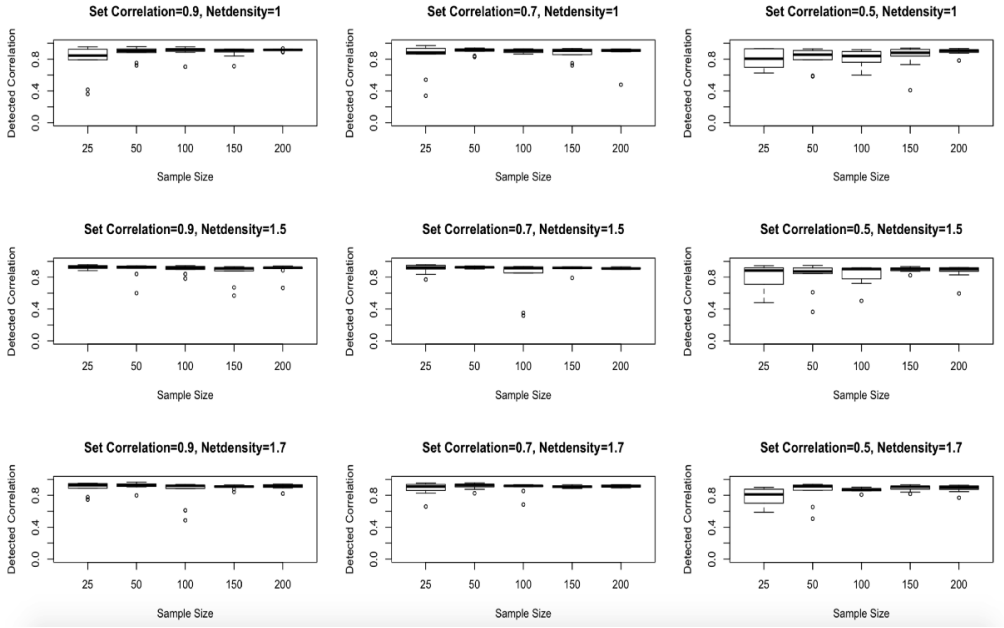


Figure2 Detected Correlation of hub two

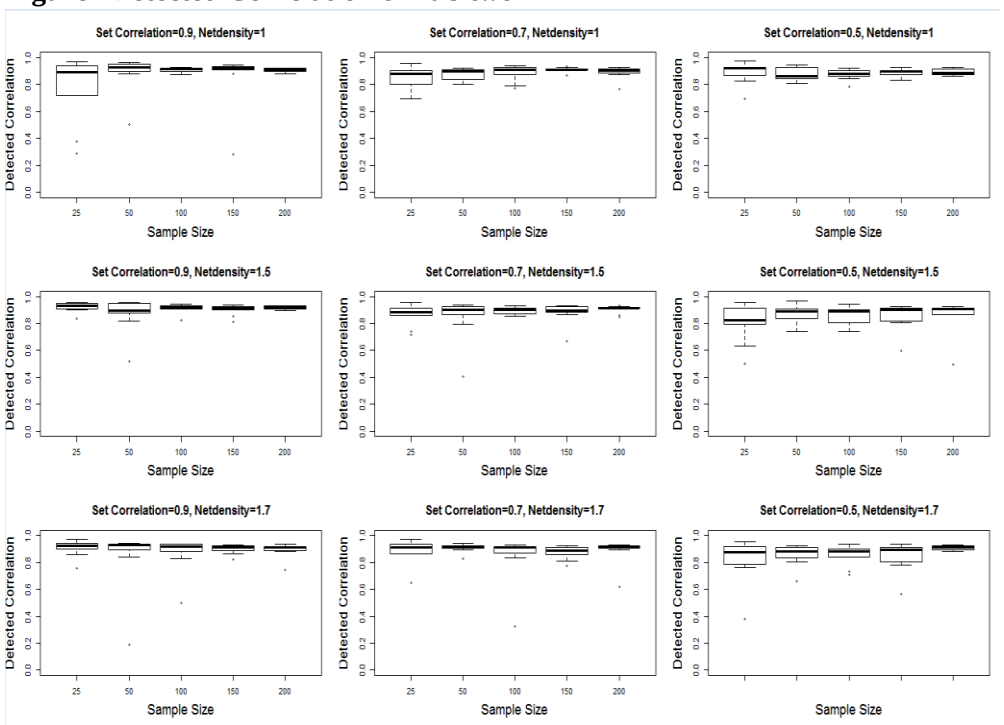


Figure 3 Factor score of real yeast cell cycle data captured by the first PC

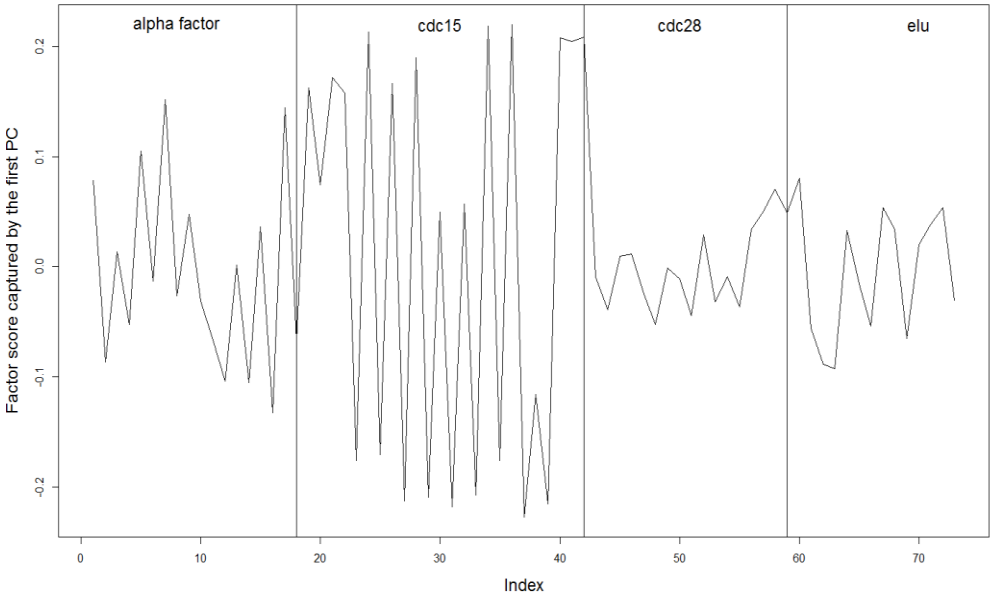


Figure 4 Factor score of real yeast cell cycle data captured by the second PC

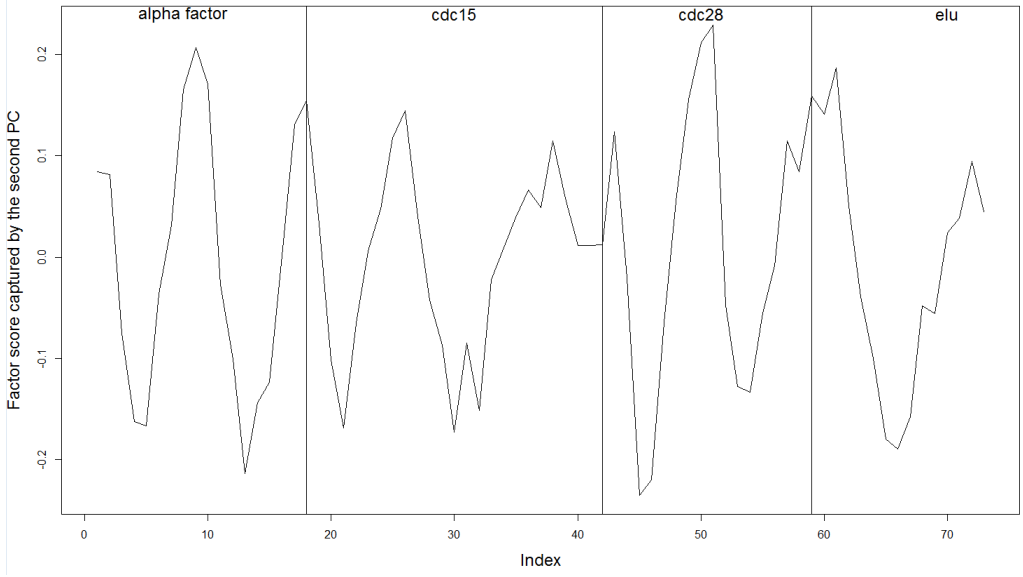


Table 2 Gene Ontology Classification result of Gene signal captured by the second PC

	GOBPID	P-value	Odds Ratio	Exp Count	Count	Size	Term
1	GO:0007049	1.498046e-32	15.102239	11.40906638	59	729	cell cycle
2	GO:0022402	1.355541e-30	14.183746	9.70318403	54	620	cell cycle process
3	GO:0006259	1.675567e-24	11.644744	7.60604425	44	486	DNA metabolic process
4	GO:0000280	3.136578e-23	12.893939	5.22719914	37	334	nuclear division
5	GO:0048285	9.059438e-23	12.449837	5.38370210	37	344	organelle fission
6	GO:0006974	1.104280e-22	12.799671	5.03939557	36	322	cellular response to DNA damage stimulus
7	GO:0051276	1.158718e-22	10.523021	7.95035078	43	508	chromosome organization
8	GO:0000278	1.474303e-21	11.347403	5.82191042	37	372	mitotic cell cycle
9	GO:0006281	1.727456e-21	13.037868	4.35078251	33	278	DNA repair
10	GO:1903047	5.328173e-21	11.215686	5.63410685	36	360	mitotic cell cycle process
11	GO:0007067	1.740318e-19	13.794003	3.28656233	28	210	mitotic nuclear division
12	GO:0006302	3.041175e-19	20.707692	1.72153265	22	110	double-strand break repair
13	GO:0006260	3.192123e-19	15.122623	2.75445224	26	176	DNA replication
14	GO:0007059	1.343441e-18	14.150820	2.91095521	26	186	chromosome segregation
15	GO:1902589	4.820637e-17	6.582763	15.36859147	50	982	single-organism organelle organization
16	GO:0006996	2.866868e-16	6.158295	22.75553157	59	1454	organelle organization
17	GO:0006310	7.386962e-16	11.809524	3.05180788	24	195	DNA recombination
18	GO:0006261	8.176217e-16	15.874095	1.89368591	20	121	DNA-dependent DNA replication
19	GO:0000819	9.823798e-16	17.294625	1.65893146	19	106	sister chromatid segregation
20	GO:0051301	1.615679e-15	8.558858	5.43065300	30	347	cell division
21	GO:0007064	5.290247e-15	38.276216	0.59471128	13	38	mitotic sister chromatid cohesion
22	GO:0043570	8.372386e-15	65.855263	0.35995683	11	23	maintenance of DNA repeat elements
23	GO:0033554	8.557408e-15	6.478493	9.75013492	38	623	cellular response to stress
24	GO:0000003	3.530032e-14	6.869873	7.51214247	33	480	reproduction
25	GO:0000070	5.614367e-14	16.368571	1.51807879	17	97	mitotic sister chromatid segregation

*Table 2 lists the top 25 gene classifications sorted by p-value.

Figure 5 Factor score of real yeast cell cycle data captured by the third PC

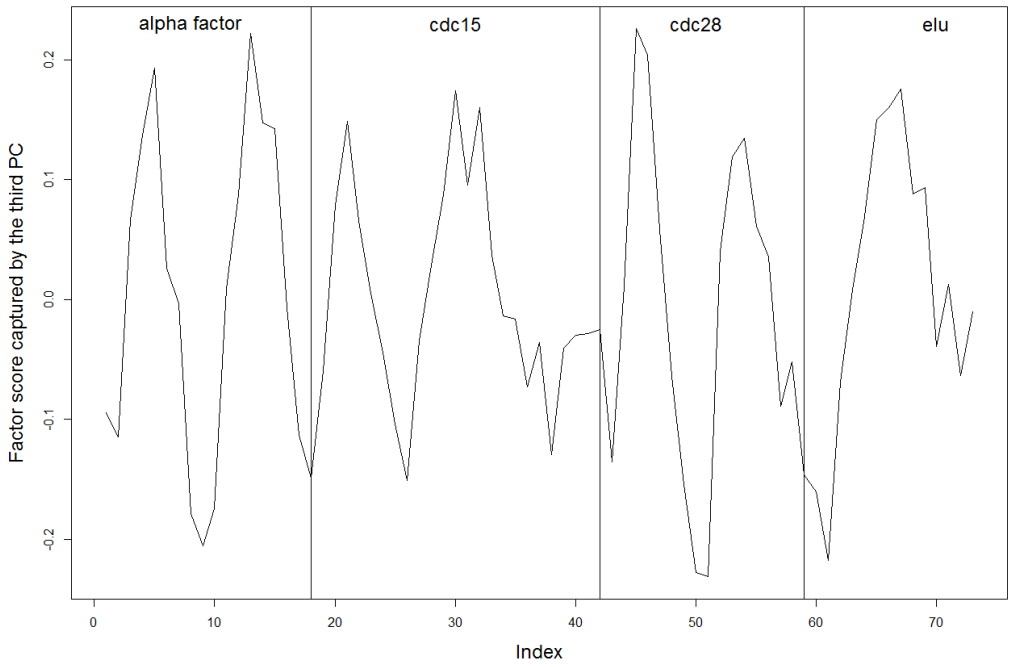


Table 3 Gene Ontology Classification result of Gene signal captured by the third PC

	GOBPID	P-value	Odds Ratio	Exp Count	Count	Size	Term
1	GO:0022402	3.561553e-28	16.122723	8.03022126	47	620	cell cycle process
2	GO:0007049	2.787728e-27	15.060294	9.44198597	49	729	cell cycle
3	GO:0006259	3.732426e-22	12.571035	6.29465731	38	486	DNA metabolic process
4	GO:0000280	6.305652e-21	13.735099	4.32595791	32	334	nuclear division
5	GO:0048285	1.575634e-20	13.269231	4.45547760	32	344	organelle fission
6	GO:0006974	2.820312e-20	13.500629	4.17053427	31	322	cellular response to DNA damage stimulus
7	GO:0006281	7.466613e-20	14.187167	3.60064760	29	278	DNA repair
8	GO:0000278	1.755652e-19	12.110588	4.81813276	32	372	mitotic cell cycle
9	GO:1903047	7.889684e-19	11.853955	4.66270912	31	360	mitotic cell cycle process
10	GO:0006260	7.876908e-18	16.364146	2.27954668	23	176	DNA replication
11	GO:0007067	3.136509e-17	14.250000	2.71991365	24	210	mitotic nuclear division
12	GO:0006302	4.326067e-17	21.257309	1.42471668	19	110	double-strand break repair

13	GO:0051276	2.234277e-16	8.928259	6.57960065	33	508	chromosome organization
14	GO:0006261	2.800019e-16	18.926193	1.56718834	19	121	DNA-dependent DNA replication
15	GO:0006310	1.166447e-15	13.515376	2.52563411	22	195	DNA recombination
16	GO:0051301	3.968619e-15	9.688125	4.49433351	27	347	cell division
17	GO:0007059	5.875512e-15	13.281283	2.40906638	21	186	chromosome segregation
18	GO:1902589	7.977718e-15	6.772128	12.71883432	42	982	single-organism organelle organization
19	GO:0043570	6.973224e-14	67.915633	0.29789530	10	23	maintenance of DNA repeat elements
20	GO:0033554	9.620977e-14	7.023077	8.06907717	33	623	cellular response to stress
21	GO:0000003	2.144971e-13	7.530759	6.21694549	29	480	reproduction
22	GO:0051716	4.912989e-13	5.999396	12.20075553	39	942	cellular response to stimulus
23	GO:0000819	2.749665e-12	15.604396	1.37290880	15	106	sister chromatid segregation
24	GO:0006950	6.079443e-12	5.911992	9.32541824	33	720	response to stress
25	GO:0006996	1.477236e-11	5.125492	18.83216406	46	1454	organelle organization

*Table 3 lists the top 25 gene classifications sorted by p-value.

B R Code

Real data analysis

```
conn<-read.table("Scere20160114CR.txt", sep="\t", header=T)

for(i in 1:ncol(conn)) conn[,i]<-as.character(conn[,i])

load("spellman_73_filled.bin")

source("https://bioconductor.org/biocLite.R")

biocLite("org.Sc.sgd.db")

library("org.Sc.sgd.db")

arrayrowname<-c()
```

```
for(ii in 1:nrow(array)){

arrayrowname[ii]<-mget(rownames(array)[ii], org.Sc.sgdUNIPROT,

ifnotfound=NA)[[1]][1]}

array.rename1<-array

for(ii in 1:nrow(array.rename1)){

  if(!is.na(arrayrowname[ii])){

   row.names(array.rename1)[ii]<-arrayrowname[ii]

            }

}

connid<-matrix(nrow=nrow(conn),ncol=ncol(conn))

connid[1,1]

for(i in 1:nrow(conn) ){

 for (j in 1:ncol(conn)){

 if(!is.na(strsplit(conn[i,j], "uniprotkb:")[[1]][2])){

 connid[i,j]<-strsplit(conn[i,j], "uniprotkb:")[[1]][2]

                        }

             }

}

connid.na<-na.omit(connid)

#find the overlap between the network and connetciton
```

```
library('igraph')

connid.overlap1<-connid.na[which(connid.na[,1] %in% row.names(array.rename1)),]

connid.overlap<-connid.overlap1[which(connid.overlap1[,2] %in%
row.names(array.rename1)),]

array.overlap<-array.rename1[which(row.names(array.rename1) %in% connid.overlap),]

g<-graph.data.frame(connid.overlap, directed=TRUE, vertices=NULL)

x<-array.overlap[,c(-1,-2,-3,-4)]

xrowmean<-rowMeans(x)

xrowsd<-apply(x,1,sd)

#######make the array matrix standard######

st_x<-(x-xrowmean)/xrowsd

rowMeans(st_x)

 A2<-matrix(ncol=ncol(st_x))

 for(ii in 1:nrow(connid.overlap)){

  sample.data<-st_x[which(rownames(st_x)==connid.overlap[ii,1]),]-
  st_x[which(rownames(st_x)==connid.overlap[ii,2]),]

  A2<-rbind(A2,sample.data)

}

 A3<-A2[c(-1,-2),]

 N<-t(A3)%*%(A3)
```

```
#A is the matrix that with the information of difference between connected protein
 #N is the A'A matrix of  related protine
 flag=1
D<-matrix(ncol=ncol(st_x),nrow=1000)
 while(flag<1001){
sample.node2<-sample(rownames(st_x),2)
count2<-shortest.paths(g,v=sample.node2[1],to=sample.node2[2])
if(count2>5){
D[flag,]<-st_x[sample.node2[1],]-st_x[sample.node2[2],]
flag<-flag+1
 }
}
 #D is the matrix that with the information of difference between unconnected protein
 M<-t(D)%*%D
 #m is the D'D matrix of un related protine
E<-M-N
########This function returns the matrix which I need to caltulate its eign vector
 my.prc1<-prcomp(E,scale=TRUE)
 plot(my.prc1$rotation[,3],type='l',xlab='',ylab='')
 mtext("Factor score captured by the third PC", side=2, line=3, cex=1.5)
```

```
mtext("Index", side=1, line=3, cex=1.5)
 abline(v=18)
 abline(v=59)
 abline(v=42)
 text(x=8,y=0.23,label='alpha factor',cex=1.5)
text(x=30,y=0.23,label='cdc15',cex=1.5)
text(x=50,y=0.23,label='cdc28',cex=1.5)
 text(x=70,y=0.23,label='elu',cex=1.5)
 x_loading1<-(st_x)%*%(my.prc1$rotation[,c(1,2,3,4,5,73)])
 PC1_x1<-x_loading1[,3]
PC1_cx1<-PC1_x1[which(abs(PC1_x1)>=(0.6*sqrt(72)))]
  source("https://bioconductor.org/biocLite.R")
 biocLite("GOstats")
 library(GOstats)
sel<-c()
  sel<-row.names(array[which(row.names(array.rename1)%in%names(PC1_cx1)),])
 uni<-unique(rownames(array))
  params=new("GOHyperGParams",geneIds=sel,universeGeneIds=uni,annotation=
"org.Sc.sgd.db",ontology="BP",pvalueCutoff=0.01,conditional=FALSE,testDirection="o
ver")
```

```
 over=hyperGTest(params)

summary(over)

ov<-summary(over)
```