OPEN ACCESS

Nano Biomed Eng

ISSN 2150-5578 http://nanobe.org

Article

Identifying Functional Modules of Diffuse Large B-cell Lymphoma Gene Co-expression networks by Hierarchical Clustering Method Based on Random Matrix Theory

Jian Chen, Kaiwang Zhang*, Jianxin Zhong

Faculty of Materials, Optoelectronics and Physics, Xiangtan University, Hunan 411105, China and
Institute for Quantum Engineering and Micro-Nano Energy Technology, Xiangtan University, Hunan 411105, China
* Corresponding author: kwzhang@xtu.edu.cn (Kaiwang Zhang)

Abstract

In this work, we propose an original method, the Random matrix theory (RMT)-based hierarchical clustering method, to identify functional gene networks of diffuse large B-cell Lymphoma (DLBCL) gene co-expression networks. Comparing topological approach, the RMT-based hierarchical clustering method is effective in representing not only the strong correlations between genes inside the modules (the modularity and independency of networks), but also the weak correlations between different modules (the hierarchy of networks). We show that missing expression values among microarray dataset should not be neglected, and different imputation methods result in different performances. We suggest LLS to estimate missing values for better performance in accuracy and stability. Based on the RMT, the random noises are separated from DLBCL gene expression data. We use normalized root mean squared error (NRMSE) ratio method to identify a transition zone of NNSDs, and for DLBCL networks it is [0.71, 0.84].

Keywords: Random matrix, Gene co-expression networks, Hierarchical clustering method, Functional modules, Diffuse large B-cell Lymphoma

Citation: J. Chen et al, Identifying functional modules of diffuse large B-cell Lymphoma gene co-expression networks by hierarchical clustering method based on random matrix theory,Nano Biomed. Eng. 2011, 3(1), 57-65. DOI: 10.5101/nbe.v3i1.p57-65.

Introduction

Diffuse large B-cell lymphoma (DLBCL), are the most common subtype of non-Hodgkin's lymphoma, composing 32% of lymphoma patients all over the world. There being evident heterogeneity in immunological phenotype, genetic mutations, and gene expression, the classification and prognosis of DLBCL are still debated. Recent technical and analytical advances make it practical to quantitate the expression of thousands of DLBCL genes in parallel using Lymphochip microarrays [1]. A genome-wide approach is made possible to analyze the DLBCL networks, classify subgroups, and contribute to the clinical diagnosis, therapy and prognosis.

Many approaches have been applied to identify functional modules based on DNA microarray, such as Boolean network methods, differential equation-based network methods [2], Bayesian network method, clustering methods [3], and co-expression network methods [4,5]. The co-expression network technique is widely adopted because it can manage the nature of microarray datasets: typical noise and high dimension. However, the co-expression network method relies on arbitrarily assigned thresholds for link cutoff, which introduces subjectivity into network structure and topology. On the other hand, the limited number of samples in microarray experiment may introduce significant "measurement noise" that compromises the accuracy of the underlying correlations.

The RMT was firstly proposed by Wigner and Dyson in the 1960s for studying the spectrum of complex nuclei [6]. It's a powerful approach for identifying and modeling phase transitions and dynamics associated with disorder and noises in statistical physics and materials science, such as spectra of large atoms, spectra of conductance peak [7], metal insulator transitions in disorder systems [8], spectra of quasi-periodic systems [9, 10], chaotic systems [11], complex networks [12, 13], gene coexpression network[14], protein dynamics[15], multivariate time series such as human EEG data [16], and the stock market

[17]. Luo et al. used RMT-based topological approach in many biological networks, determined the transition threshold of NNSDs by Chi-square test, and showed that: NNSDs of the adjacent matrix of protein interaction network and metabolic network, as well as that of the correlation matrix of yeast gene network follow the universal predictions of RMT [18, 19]; on the other hand, NNSDs of the correlation matrix of various complex biological networks follow the universal description of RMT [20]. They indicated that RMT-based approach is applicable to biological networks to separate system-specific properties from random noises, tackling the problem in removing noises. However, the topological approach is limited to strong correlations between genes inside the modules without the weak correlations between different modules. Besides, the Chi-square test is subjective due to artificial confidence level.

Microarray data often contain missing values due to various reasons: insufficient resolution, image corruption, dust or scratches on the slides or experimental errors during the laboratory process. Missing value estimation becomes an important preprocessing step for microarray data analysis, since many statistical analysis algorithms either require complete datasets or are subject to significant performance degradations due to missing values [21]. A series of microarray missing value estimation techniques have been developed, including K-nearest neighbor (KNN) [22], singular value decomposition (SVD) [22], least square (LSimpute) [23], Bayesian principle component analysis (BPCA) [24], local least square (LLS) [25], collateral missing value estimation (CMVE) [26], gaussian mixture clustering (GMCimpute) [27], integrative missing value estimation method (iMISS) [28], and GOimpute utilizing external information from Gene Ontology [29]. Test datasets are selected from public microarray datasets by removing missing values, to compare the performance of different estimations statistically by computing NRMSE. Whereas, missing value imputations have not been applied to gene expression analysis yet. Luo et al. demonstrated that whether using KNN to impute missing values in yeast Saccharomyces cerevisiae microarray dataset or not don't affect the eigenvalue fluctuations. So they only used the experiments both genes having values to calculate the correlation, which will cursorily neglect some important genes regulating mitotic cycle. Thus, it's essential to apply appropriate imputation to gene expression analysis practically, to reduce the impact that imputation has on downstream analysis.

In this report, we propose a novel approach--RMT-based hierarchical clustering method--to identify the functional modules of DLBCL gene networks. During the preprocessing, we use LLS to estimate the missing values in DLBCL microarray dataset. The universal predictions of RMT are applied to separate the random noises from gene networks gradually. We employ NRMSE ratio method to determine a transition zone of NNSDs. At the transition completing point we retain characteristic genes to construct topological graph and hierarchical dendrogram, and investigate the distinctions between modules and clusters.

2. Methods

We analyze DLBCL gene expressions of 128 Lymphochip microarrays, made up of 4026 cDNA clones and 96 normal and malignant lymphocyte samples [30].

We use LLS imputation method based on the least squares formulation to estimate missing values in DLBCL gene expressions. LLS is a local imputation algorithm which exploit the dominant local similarity structure of the dataset. L_2-norm is used as similarity measure, and after regression analyses based on the least squares formulation missing values are estimated by a linear combination of the similar genes [25]. K-nearest neighbor genes are selected using heuristic algorithm, according to structures of different datasets.

The standard Pearson correlation coefficients are computed to construct the Pearson correlation matrix. The Pearson coefficient between gene and gene is:

$$c\left(g_i, g_j\right) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{g_{ik}-M_{g_i}}{\sigma_{g_i}}\right)\left(\frac{g_{jk}-M_{g_j}}{\sigma_{g_j}}\right) \quad (1)$$

where M_{g_i}, M_{g_j} are the average expression levels of gene g_i and g_j, σ_{g_i}, σ_{g_j} are the standard deviations, and N is the total number of samples. c(g_i, g_j) ranges from -1 to 1. Its absolute value signifies the correlations between two genes, with 1 corresponding to absolutely positive correlations, -1 meaning completely negative correlations, and 0 referring to no relationship.

The RMT focuses on the statistical properties of eigenvalue spacing between consecutive eigenvalues. From RMT [10], for real and symmetrical random matrices, the eigenvalue fluctuations follow two universal laws depending on the correlativity of eigenvalues. Strong correlation of eigenvalues is described by the Gaussian orthogonal ensemble (GOE) statistics. On the other hand, eigenvalue spacing distribution follows Poisson statistics if there is no correlation between eigenvalues.

In general, the density of eigenvalues of a matrix varies with its eigenvalue E_i (i=1,2,3, . . .N), where N is the order of the matrix. Thereby eigenvalue spacing distribution is a function of E_i and thus system dependent. To observe universal eigenvalue fluctuations of different types of matrices, one can replace E_i by the unfolded spectrum e_i, where $e_i=N_{av}(E_i)$ and N_{av} is the smoothed integrated density of eigenvalues obtained. With the unfolded eigenvalues, one calculates NNSD of eigenvalues P(s), defined as the probability density of unfolded eigenvalue spacing $s=e_{i+1}-e_i$. The eigenvalue fluctuation is generally obtained from NNSD of the eigenvalues. From RMT [10], P(s) of the GOE statistics closely follows the Wigner-Dyson distribution:

$$P_{GOE}(s) \approx \frac{1}{2}\pi s \exp(\frac{-\pi s^2}{4}) \quad (2)$$

In the case of Poisson statistics, P(s) is given by the Poisson distribution:

$$P_{Poisson}(s) = \exp(-s) \quad (3)$$

Setting factor $q \in (0,1)$, we substitute the matrix elements smaller than q with 0, and retain the bigger ones. Here, 0 denotes weaker correlations between genes, and nonzero elements denote stronger correlations. We cast off the rows with 1 as diagonal element and 0 as all the other off-diagonal elements--corresponding to genes having only weak correlations between other genes. And we retain the rows with 1 as diagonal element and nonzero values as off-diagonal elements--representing genes having strong correlations between others. The retained genes are called characteristically expressed genes playing important roles in biology. Then, the characteristic genes are used to construct a new gene expression matrix, together with Pearson correlation matrix and NNSDs. As q increases, NNSDs exhibit a transition from Wigner-Dyson distribution to Poisson distribution [18, 19, 20].

NRMSE method is widely used to observe the deviation sensitivity between measured values and standard values in engineering survey. As NRMSE value is sensitive to tiny errors in a set of measures, reflecting the measuring precision well and truly. NRMSE is counted by:

$$NRMSE = \frac{\sqrt{Mean(P - P_{\text{standard}})^2}}{SD(P_{\text{standard}})} \quad (4)$$

of which P is the NNSDs, P_{standard} stands for the standard distribution functions [22]. When NNSDs approximate the standard distribution, NRMSE approaches its minimum value 0.0; when NNSDs deviate from the standard distribution gradually, NRMSE increases; at the point q when the deviations between NNSDs and two standard distributions are the same, transition threshold q_c is found.

In order to reduce further the impact of random noises on networks structures, we compute NRMSE ratio by dividing NRMSE_Wigner-Dyson by NRMSE_Poisson. At the maximum ratio, NNSDs deviate far from Wigner-Dyson distribution, and approximate Poisson distribution. We define the transition completing point q_m, where, there are most system-intrinsic strong correlations between genes in networks structures, with few weak correlations due to random noises. At q_m, random noises can be separated fully, and the retained characteristically expressed genes can construct the factual networks structures.

Then topological graph is portrayed based on characteristic genes by Biolayout [31]. Dendrogram is constructed using hierarchical clustering algorithm [32] by Cluster3.0 and Java TreeView. We choose unsupervised Two Way Clustering; for the clustering method we choose the Average linkage; and for Similarity Metric we use Pearson correlation coefficient. This algorithm sorts through all the data to find the pairs of genes that behave most similarly in each experiment to form clusters of apparently coregulated genes.

3. Results and Discussions

3.1 Determining q_m by RMT

In preprocessing, we use LLS imputation method to estimate missing values in DLBCL gene expressions. There are 19667 missing expressions in this dataset, accounting for 5.09%. The following heuristic algorithm for estimating k-nearest neighbors is firstly performed. Several k-values are tested as nearest neighbors to estimate the missing values, and NRMSEs are calculated to compare the performance. An optimal value shows the best performance using known elements of gene expression matrix: k=140. Here, the square structure of evaluating matrix is the same to the least squares structure, namely, the number of rows is larger than that of columns, and NRMSE curve drops and stabilizes. The missing value is then estimated as a linear combination of the k similar genes in the least squares formulations.

LLS imputation method based on the least squares formulation exploits local similarity structures in the microarray data as well as least squares optimization process. LLS takes into account of the characteristic of the given data matrix, and an optimal k-value is selected using heuristic algorithm to gain the best performance. As the only necessary parameter k is automatically determined by the procedures, LLS can be classified as non-parametric missing value estimation methods. Otherwise, LLS shows good performance when the optimal k-value ranges in a wide area. Kim et al. studied Spellman time series cycle-cycle dataset and non-time series response to environmental changes dataset in yeast Saccharomyces cerevisiae genes, and indicated that, LLS shows competitive results compared with other imputation methods (such as KNN) on various datasets and percentages of missing values in the data. In a word, LLS is a robust and accurate missing value estimation method, and can reduce the effect of estimations on downstream analysis.

After estimating the missing values in the gene expression matrix, we compute Pearson correlation coefficients to construct the real and symmetrical co-expression matrix, and the co-expression networks model.

Fig. 1 shows the fluctuations of NNSDs with eigenvalue spacing s. We can see from Fig. 1 that NNSDs follow different laws as q changes. NNSDs gradually transit from Wigner-Dyson distribution to Poisson distribution when q increases.

We apply NRMSE ratio method to determine the transition completing point of NNSDs. NRMSE ratio

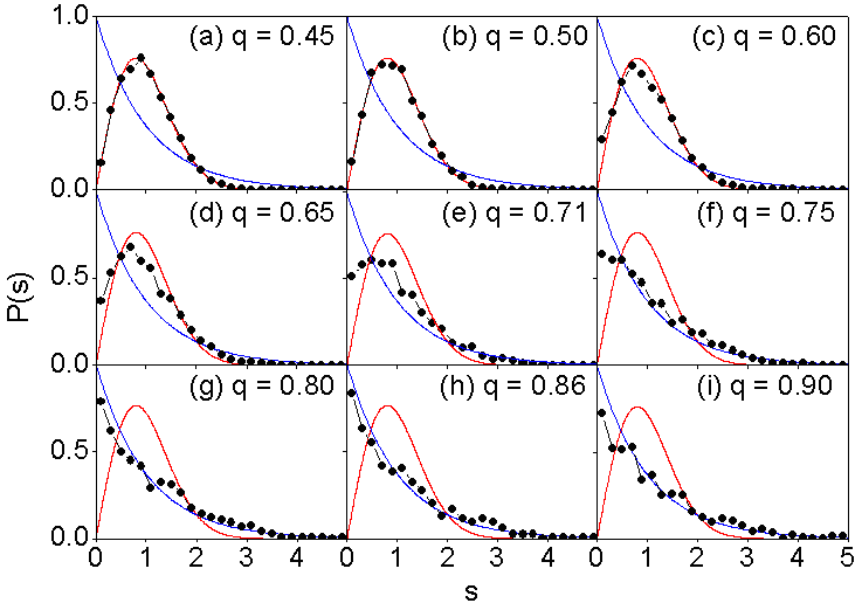

Fig. 1. NNSDs curves of DLBCL gene expression dataset. Red line is Wigner-Dyson distribution and blue line is Poisson distribution.

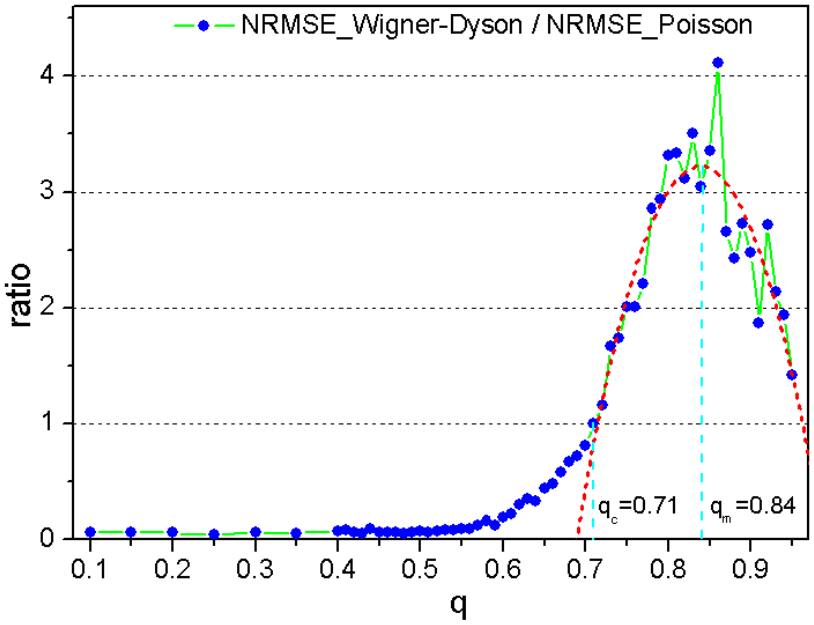

Fig. 2. NRMSE fitting ratio curve of NNSDs of DLBCL.

curve of NRMSE_Wigner-Dyson versus NRMSE_Poisson is shown in Fig. 2. At the transition threshold q_c=0.71 ratio value is 1.0, and NNSDs deviate equally from two standard distributions. Binary polynomial fitting is performed on the artificially interceptive points bigger than q_c=0.71. The transition completing point q_m=0.84 is observed at the vertex of fitting ratio curve. Here, NRMSE_Wigner-Dyson is bigger than NRMSE_Poisson, which means NNSDs deviate far from Wigner-Dyson distribution, and approach Poisson distribution.

Weak interactions fabricated by random noises are separated from intrinsic strong interactions between genes, and then a new gene expression matrix is obtained reflecting the factual structures of DLBCL gene networks. 2383 characteristically expressed genes are retained at q_c=0.71, including all the strong correlations and most weak correlations between genes as well. While at q_m=0.84, 1230 characteristic genes are reserved, composed of most strong correlations and few weak correlations.

However, microarray datasets are of high noises and biological processes are of complexity, which make it impossible to remove noises from interactions between genes and from networks structures completely. In our findings a transition zone of [0.71, 0.84] is defined based on two transition points q_c and q_m for DLBCL dataset. Moreover, any point inside the zone can work as a criterion for factor q to remove noises, select

characteristic genes, and construct gene networks. The higher q value is, the more removed noises are. Therefore, compared with q_c=0.71, random noises are removed fully at q_m=0.84. Although some correct links are lost, the retained characteristic genes can show most of the intrinsic interactions between genes, and construct the true structures without random noises. A solution to resolve the inconsistency is called soft thresholding, to identify a rational transition point.

3.2 Constructing topological graph

Our results indicate that RMT is applicable to DLBCL gene networks, and NNSDs transit from Wigner-Dyson distribution to Poisson distribution. Meantime, networks structure transits from rich-correlation systems to poor-correlation systems [18, 19]. In order to observe the changes of networks structures, we portray the topological graph using Biolayout at different q value. We find most nodes gather together as a big module for mix of strong and weak correlations at small q value. As q increases gradually, namely spurious links due to random noises are removed gradually from true links, some nodes break away from the biggest module to form small insulated modules. When q is close to 1, topological graph lay out insulated modules composed of strongly correlated genes as random noises are separated fully.

Topological graphs are shown in Fig. 3. In Fig. 3(a) (at transition threshold q_c=0.71), most nodes get together for a big module, with several disperse small modules made up of a few nodes. While in Fig. 3(b) (at transition completing point q_m=0.84), all modules composed of 1230 nodes are isolated clearly. Modules with less than 5 nodes are portrayed dark green and others with no less than 5 nodes are colorized. Nodes are connected through edges for strong interactions inside the modules. While, there are no edges between modules. Thus, the modularity and independency of networks are represented by topological graph. In addition, q_m=0.84 is demonstrated to be a sound transition point for DLBCL gene networks.

3.3 Constructing hierarchical dendrogram

Hierarchical modularity of networks indicates that large modules can be divided into small sub-modules; inside the same module genes are connected directly for strong interactions, and indirectly for weak interactions; genes from different modules are connected indirectly or unconnected for weak correlations; modules are connected for the same or similar functions. Topological graph exhibits the modularity and independency of networks but hierarchical modularity.

To analyze the hierarchical modularity of DLBCL networks, we construct hierarchical dendrogram of 1230 characteristic genes at q_m=0.84 using clustering algorithm. Genes with similar functions are gathered together in a cluster. Clusters with specific functions in hierarchical dendrogram correspond to certain modules in topological graph. 14 modules and clusters are consistent in gene members, which are gathered closely, representing the modularity and independency of DLBCL networks. Another 8 modules and clusters are common in gene members, but different in that: clusters are shown as integrated branches, composed of closely gathered genes, while modules are collections of isolated small modules without edges through modules.

Alizadeh et al. [30] investigated DLBCL gene expression dataset using Two Way Clustering algorithm. They didn't estimate the missing values in dataset, didn't remove the random noises to select characteristic genes, and implemented clustering on expressions of all clones in every sample (4026*96). Four clusters were studied in detail: Proliferation, Germinal centre B, Lymph node and T cell cluster [30].

As shown in Fig. 4(a), the branch marked dark is Lymph node cluster of 29 genes, including several signature featured genes Alizadeh et al. [30] spoke of: CSF-1 (macrophage-specific colony-stimulating factor), NK4 (natural killer transcript 4) possessing anti-tumor, anti-virus infection and immunoregulation effects.

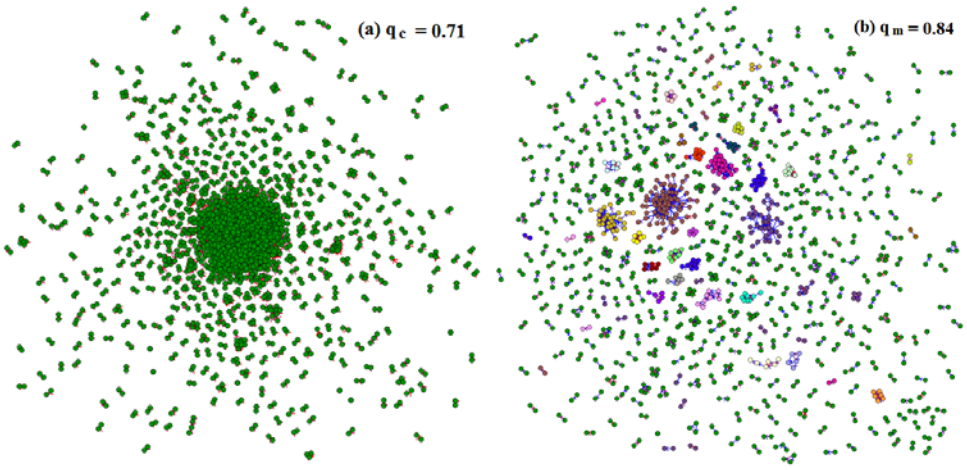

Fig. 3. Topological graph for DLBCL gene networks.

In addition, genes involved in the remodeling of the extracellular matrix are contained: MMP2 and MMP9 (the members of matrix metalloproteinases, which effectively decompose the main component of basilar membrane Ⅳ -collagen proteins), as well as TIMP-3 (a member of tissue inhibitors of metalloproteinase, which regulates the decomposition of extracellular matrix proteins) [34]. Topological graph of Lymph node module is shown in Fig. 4(b) with 28 genes. Topological graph agrees well with dendrogram in gene members, signifying modularity and independency of networks.

On the other hand, different clusters are connected for similar functions in dendrogram representing the hierarchical modularity of gene networks and weak correlations between clusters; whereas topological graph can not manifest the properties. MHC (major histocompatibility complex) cluster is magnified in Fig. 4(c), composed of tightly linked genes. Sub-cluster marked light grey is composed of genes encoding MHC-Ⅰ, while sub-cluster marked dark grey is composed of genes encoding MHC-Ⅱ. They are linked at the root of the branch, inducing and regulating the immune response together [35]. MHC cluster is a case in point to manifest the hierarchical modularity and weak correlations between different clusters. Topological graph of MHC module is divided into 5 insulated sub-modules in Fig. 4(d). The biggest two sub-modules are made up of mixed genes encoding MHC-Ⅰ and MHC-Ⅱ. And 5 sub-modules are isolated without linked edges. Thus, weak correlations between different modules can not be represented by topological graph.

To demonstrate the rationality of transition zone in Section 3.1, we construct hierarchical dendrogram at q_c, q_m and the outer point, and analyze the gene members of dendrograms respectively. Fig. 5(a) shows the magnified dendrogram of Proliferation cluster of 85 genes at q_m=0.84. Consistent with Alizadeh et al. [30], the signature featured genes include cell-cycle control genes (CDC2,cdk2,CIP2/Cdi1/KAP1,Cyclin A,Cyclin B1,p16 and p55CDC), cell-cycle checkpoint genes BUB1, DNA synthesis and replication gene RAD54, and the gene Ki67 [33], gauging the 'proliferation index' of a tumor biopsy. Fig. 5(b) shows dendrogram of Proliferation cluster of 114 genes at q_c=0.71. All genes regulating cell-cycle in Fig. 5(a) are included, as signed in Fig. 5(a, b). We construct dendrograms beside the two endpoints of transition zone. At q=0.70, the dendrogram of Proliferation cluster contains 119 genes. 3 newly joining genes are unknown genes compared with signature featured genes at q_c=0.71. At q=0.85, the dendrogram includes 72 signature featured genes. Compared with q_m=0.84, some important genes regulating cell-cycle are lost: two genes encoding Thymidine kinase 1, G1/S phase transition regulatory gene CIP2/Cdi1/KAP1, gene encoding cell-cycle checkpoint kinase CHK1 and gene encoding cyclin-dependent kinase cdk2. At the

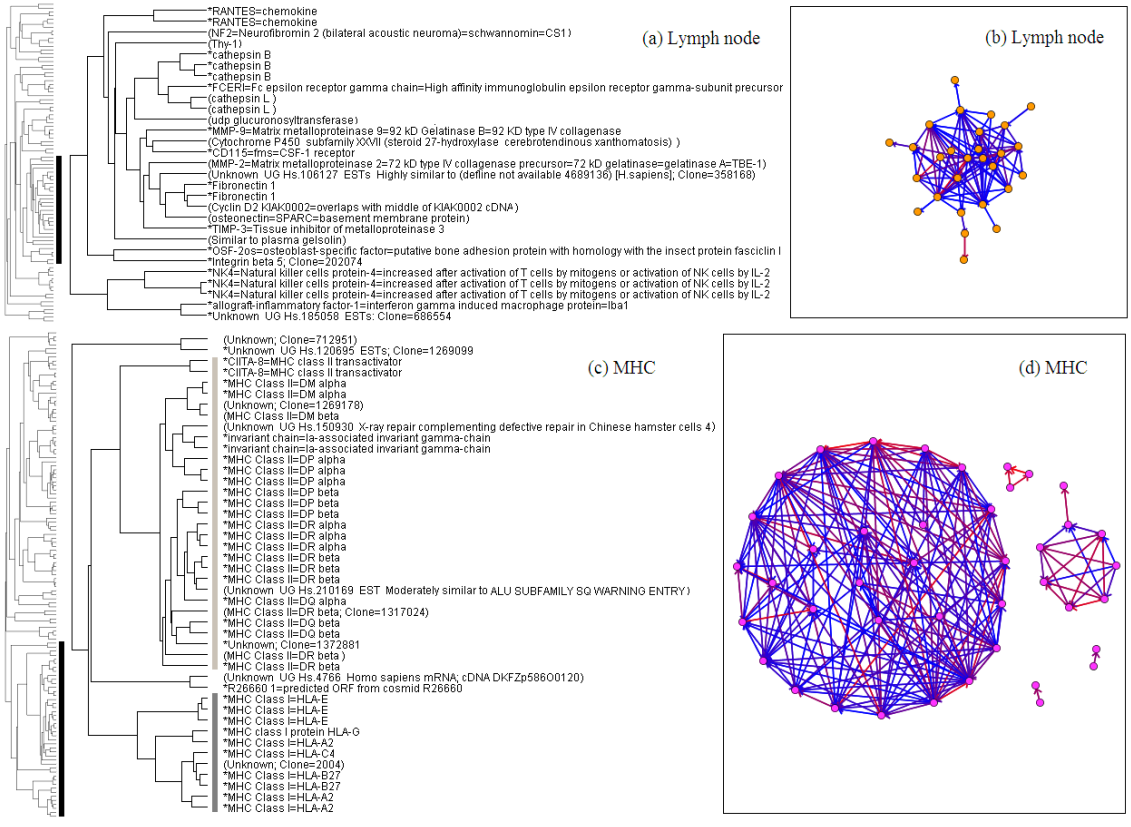

Fig. 4. Hierarchical dendrogram examples for DLBCL gene networks. (a) Lymph node cluster, (b) Lymph node module, (c) MHC cluster, (d) MHC module.

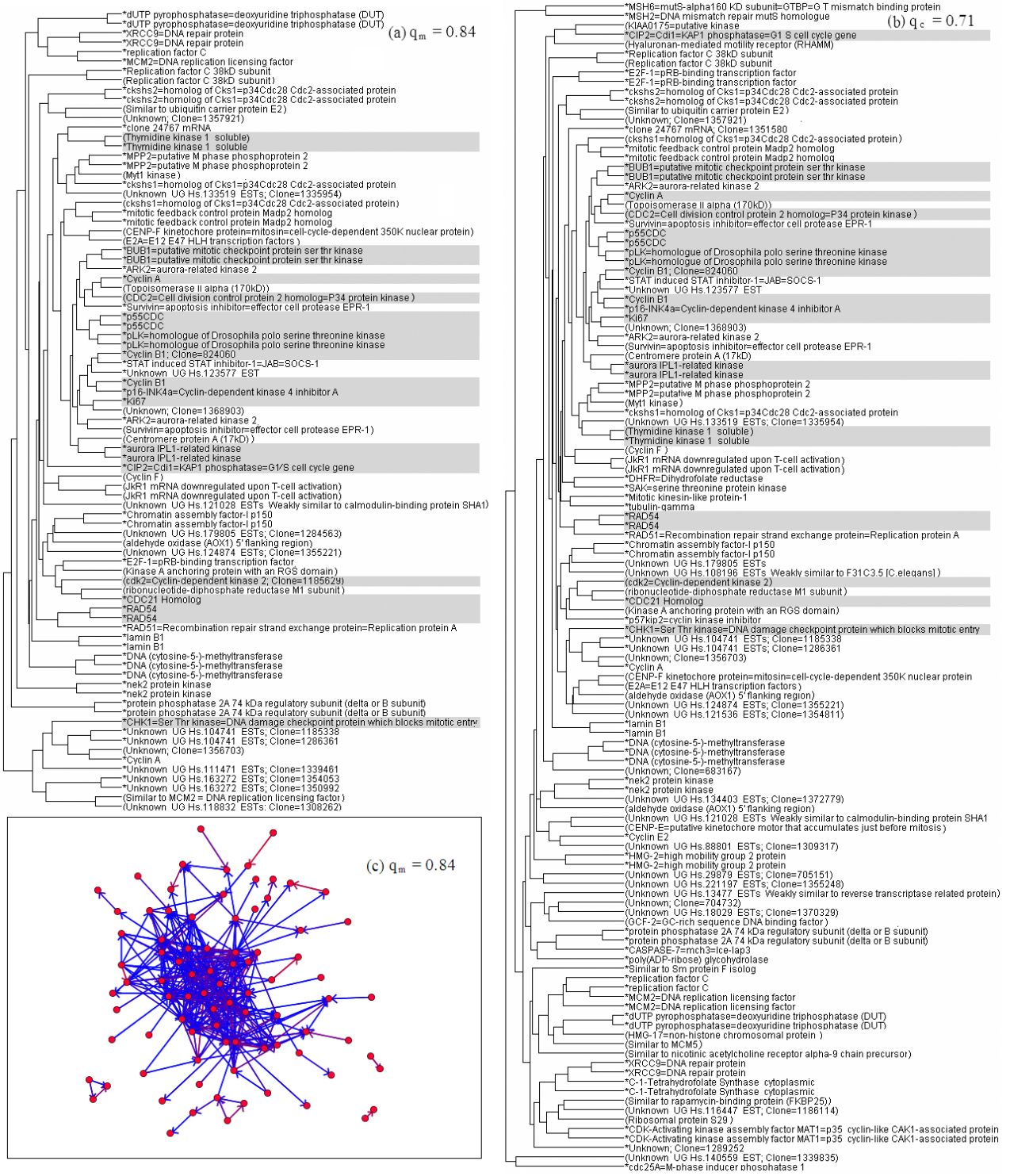

Fig. 5. Structures of Proliferation cluster. (a) dendrogram at q_m=0.84, (b) dendrogram at q_c=0.71, (c) topological graph at q_m=0.84.

lower point, it is random noises that make unconcerned genes falsely connect with genes inside the cluster. At the higher point, true links between genes are weakened or even removed while removing random noises. Again, the topological graph of Proliferation module shown in Fig. 5(c) contains 4 separate sub-modules, not showing the hierarchical modularity. We conclude that, the dendrograms constructed at q points outside the transition zone can not reveal the intrinsic correlations between genes. Namely, the transition zone [0.71, 0.84] is the optimal result using NRMSE ratio method.

5. Conclusions

We propose a novel method to identify functional modules of gene networks taking DLBCL microarray dataset as an example: RMT-based hierarchical clustering method. As missing values in microarray dataset influence

the results of classifying algorithms, it is required to estimate the missing values in preprocessing. Meantime, different networks structures are obtained using different estimations. We find LLS shows competitive performance compared with other imputation methods. In our preprocessing, we apply LLS to estimate the missing values in DLBCL dataset to reduce the impact of estimation on classifications.

Random noises are removed based on the universal predictions of RMT, and transitions of NNSDs and networks structures are observed. According to soft thresholding, a transition zone [0.71,0.84] is determined using NRMSE ratio method, of which the rationality is demonstrated by structures analyses of Proliferation cluster. At transition completing point q_m=0.84, 1230 characteristic genes are retained with which we construct the topological graph and the hierarchical dendrogram of DLBCL networks. After comparison the difference between the topological modules and the hierarchical clusters of Lymph node and MHC cluster, we conclude that RMT-based hierarchical clustering method can not only represent the modularity and independency of networks based on the strong correlations between genes inside the modules, but also represent the hierarchical modularity of networks based on the weak correlations between different modules. The method is effective in identifying functional modules of gene networks.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant NO. 30570432, 10974166 and by Research Foundation of Education Bureau of Hunan Province, China under Grant NO. 09A094.

References

1. Alizadeh A, Eisen M, Davis RE, Ma C, Sabet H, Tran T, Powell JI, Yang L, Marti G.E, Moore DT, Hudson JR, Chan WC, Greiner T, Weisenburger D, Armitage JO, Lossos I, Levy R, Botstein D, Brown PO, Staudt LM. The lymphochip: a specialized cdna microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. Cold Spring Harbor Symp. Quant. Biol 1999; 64: 71-78. doi:10.1101/sqb.1999.64.71.
2. Chen KC, Wang TY, Tseng HH, Huang CYF, Kao CY. A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae. Bioinformatics. 2005; 21: 2883-2890. doi:10.1093/bioinformatics/bti415
3. Newman AM, Cooper JB, AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number. BMC Bioinformatics. 2010; 11: 117/1-15. doi:10.1186/1471-2105-11-117.
4. Allocco DJ, Kohane SI, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. BMC ioinformatics. 2004; 5: 18/1-10. doi:10.1186/1471-2105-5-18.
5. Ruan JH, Dean AK, Zhang WX. A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Systems Biology. 2010; 4: 8/1-21. doi:10.1186/1752-0509-4-8.
6. Wigner EP. Random matrices in physics. SIAM Review. 1967; 9: 1-23. doi:10.1137/1009001.
7. Held K, Eisenberg E, Altshuler BL. Effect of spectral fluctuations on conductance-peak height statistics in quantum dots. Phys. Rev. B. 2002; 66: 033308/1-3. doi:10.1103/PhysRevB.66.033308.
8. Hofstetter E, Schreiber M. Statistical properties of the eigenvalue spectrum of the three-dimensional Anderson Hamiltonian.Phys. Rev.B. 1993; 48: 16979-16985. doi:10.1103PhysRevB.48.16979.
9. Zhong JX, Geisel T. Level fluctuations in quantum systems with multifractal eigenstates. Phys. Rev. E. 1999; 59: 4071-4074. doi:10.1103/PhysRevE.59.4071.
10. Zhong JX, Grimm U, Romer RA, Schreiber M. Level-Spacing Distributions of Planar Quasiperiodic Tight-Binding Models. Phys. Rev. Lett. 1998; 80: 3996- 3999. doi:10.1103/PhysRevLett.80.3996.
11. Bohigas O, Pato MP. Randomly incomplete spectra and intermediate statistics. Phys. Rev. E. 2006; 74: 036212/1-6. doi:10.1103/PhysRevE.74.036212.
12. Jalan S, Bandyopadhyay JN. Random matrix analysis of complex networks. Phys. Rev. E. 2007; 76: 046107/1-7. doi:10.1103/PhysRevE.76.046107.
13. Bandyopadhyay JN, Jalan S. Universality in complex networks: Random matrix analysis. Phys. Rev. E. 2007; 76: 026109/1-4. doi:10.1103/PhysRevE.76.026109.
14. Jalan S, Solymosi N, Vattay G, Li BW. Random matrix analysis of localization properties of gene coexpression network. Phys. Rev. E. 2010; 81: 046118/1-8. doi:10.1103/PhysRevE.81.046118.
15. Potestio R, Caccioli F, Vivo P. Random Matrix Approach to Collective Behavior and Bulk Universality in Protein Dynamics. Phys. Rev. Lett. 2009; 103: 268101/ 1-4. doi:10.1103/PhysRevLett.103.268101.
16. Šeba P. Random Matrix Analysis of Human EEG Data. Phys. Rev. Lett. 2003; 91: 198104/1-4. doi:10.1103/PhysRevLett.91.198104.
17. Kwapien J, Drozda S, Oswiecimka P. The bulk of the stock market correlation matrix is not pure noise. Physica A. 2006; 359: 589-606. doi:10.1016/j.physa. 2005.05.090
18. Luo F, Zhong JX, Yang YF, Scheuermann RH, Zhou JZ. Application of random matrix theory to biological networks. Phys. Lett. A. 2006; 357: 420-423. doi:10.1016/j.physleta.2006.04.076.
19. Luo F, Zhong JX, Yang YF, Zhou JZ. Application of random matrix theory to microarray data for discovering functional gene modules. Phys. Rev. E. 2006; 73: 031924/1-5. doi:10.1103/PhysRevE.73.031924.
20. Luo F, Yang YF, Zhong JX, Gao HC, Khan L, Thompson DK, Zhou JZ. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinformatics. 2007; 8: 299/1-17. doi:10.1186/1471-2105-8-299.
21. De Brevern AG, Hazout S, Malpertuy A. Influence of microarrays experiments missing values on the stability of gene groups by ierarchical clustering. BMC Bioinformatics. 2004; 5: 114/1-12. doi:10.1186/1471-2105-5-114.
22. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17: 520-525. doi:10.1093/bioinformatics/17.6.520.
23. Bo TH, Dysvik B, Jonassen I. LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Res. 2004; 32: e34. doi:10.1093/nar/gnh026.
24. Zhou XB, Wang XD, Dougherty ER. Missing-value estimation using linear and non-linear regression with Bayesian gene selection. Bioinformatics. 2003; 19: 2302-2307. doi:10.1093/bioinformatics/btg323.
25. Kim H, Golub GH, Park H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics. 2005; 2: 187-198. doi:10.1093/bioinformatics/bth499.
26. Sehgal MS, Gondal I, Dooley LS. Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. Bioinformatics. 2005; 21: 2417-2423. doi:10.1093/bioinformatics/bti345.
27. Ouyang M, Welsh WJ, Georgopoulos P. Gaussian mixture clustering and imputation of microarray data. Bioinformatics. 2004; 20: 917-923. doi:10.1093/bioinformatics/bth007.
28. Hu JJ, Li HF, Waterman MS, Zhou XJ. Integrative missing value estimation for microarray data. BMC Bioinformatics. 2006; 7: 449/1-14. doi:10.1186/1471-2105-7-449.
29. Tuikkala J, Elo L, Nevalainen LS, Aittokallio T. Improving missing value estimation in microarray data with gene ontology. Bioinformatics. 2006; 22: 566-572. doi:10.1093/bioinformatics/btk019.

30. Alizadeh AA, Eisen MB, Davis RE, Ma C, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000; 403: 503-511. doi:10.1038/35000501.
31. Enright AJ, Ouzounis CA. BioLayout—an automatic graph layout algorithm for similarity visualization. Bioinformatics. 2001; 17: 853-854. doi:10.1093/bioinformatics/17.9.853.
32. Eisen MB, Spellman PT, Brown PO, Botstein D. Proc Natl Acad Sci USA. 1998; 95: 14863-14868. doi:10.1073/pnas.95.25.14863.
33. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci USA. 1999; 96: 9212-9217. doi:10.1073/pnas.96.16.9212.
34. Visse R, Nagase H. Matrix Metalloproteinases and Tissue Inhibitors of Metalloproteinases. Circulation Research. 2003; 92: 827-839. doi:10.1161/01.RES.0000070112.80711.3D.
35. Romero P, Dunbar PR, Valmori D, Pittet M, et al. Ex vivo staining of metastatic lymph nodes by class i major histocompatibility complex tetramers reveals high numbers of antigen-experienced tumor-specific cytolytic t lymphocytes. J. Exp. Med. 1998; 188: 1641-1650. doi:10.1084/jem.188.9.1641.