Applied Mathematics, 2013, 4, 893-898
http://dx.doi.org/10.4236/am.2013.46123 Published Online June 2013 (http://www.scirp.org/journal/am)

A Clique-Based Approach to the Identification of Common Gene Association Sub-Networks

Gaolin Zheng, Assefa Tesfay, Xinyu Huang, Alade Tokuta
Department of Mathematics and Computer Science, North Carolina Central University, Durham, USA
Email: gzheng@nccu.edu, atesfay1@nccu.edu, xinyu.huang@nccu.edu, atokuta@nccu.edu

Received January 12, 2013; revised April 23, 2013; accepted May 2, 2013

Copyright © 2013 Gaolin Zheng *et al*. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

We developed a computational framework to identify common gene association sub-network. This framework combines graphical lasso model, graph product and a replicator equation based clique solver. We applied this method to find common stress responsive sub-networks from two related *Deinococcus-Thermus* bacterial species.

Keywords: Graphical Lasso Model; Quadratic Maximization; Subgraph Isomorphism; Clique; Replicator Equation; Gene Association Network

1. Introduction

Gene and gene products interact with each other due to biochemical interactions and regulatory activities [1]. Many methods have been developed to analyze these networks. Popular methods include weighted correlation network analysis (WGCNA) [2], Bayesian networks [3], autoregressive models [4], state-space models [5] and graphical Gaussian models [6]. Few studies however, have been devoted to analyze networks from multiple species simultaneously. In this study, we focus on the identification of common gene association sub-networks from multiple species. First we need to derive gene network from individual species. We chose Graphical Lasso model [6] for this task because it can handle large covariance/correlation matrices of mathematically deficient rank which is often the case for genomic data.

Identification of common gene association sub-networks is related to the subgraph isomorphism problem. The subgraph isomorphism problem can be reduced to finding maximal clique from merged graph [7] which can be constructed following graph product rules. There are a number of heuristics for finding maximal cliques. Local search may be the simplest greedy strategy that starts with some initial solution and moves from neighbor to neighbor as long as possible while increasing the clique number. The main problem with this strategy is its inability to escape local maxima where the search cannot find any further neighborhood solution. Battiti and Protasi [8] proposed reactive local search that allows the search to explore solutions that do not decrease the clique number by dynamically changing some of the parameters [8]. Another widely used heuristic is replicator equation [7]. This method is based on a continuous formulation of the maximal clique problem as quadratic programming [9].

The paper is organized as follows. In Section 2, we will describe graphical Lasso method to construct gene association networks. We will also describe graph merging and how to find maximal cliques using replicator equations. The experiments and results will be discussed in Sections 3 & 4. We offer a conclusion in Section 5.

2. Methods

To find common gene association sub-networks from two species, we need to perform ortholog mapping from two species. Orthologs are genes in different species that originated by vertical descent from a single gene of the last common ancestor. We will then construct gene association network of the orthologous genes for the two species respectively. This is followed by graph merging and maximal clique searching of the merged graph. Finally the common sub-networks are recovered for each species. **Figure 1** shows the overview of the approach.

2.1. Construction of Gene Association Network

The inverse covariance matrix Σ^{-1} is used to construct individual gene association network. Gene i and j are considered conditionally independent given other genes

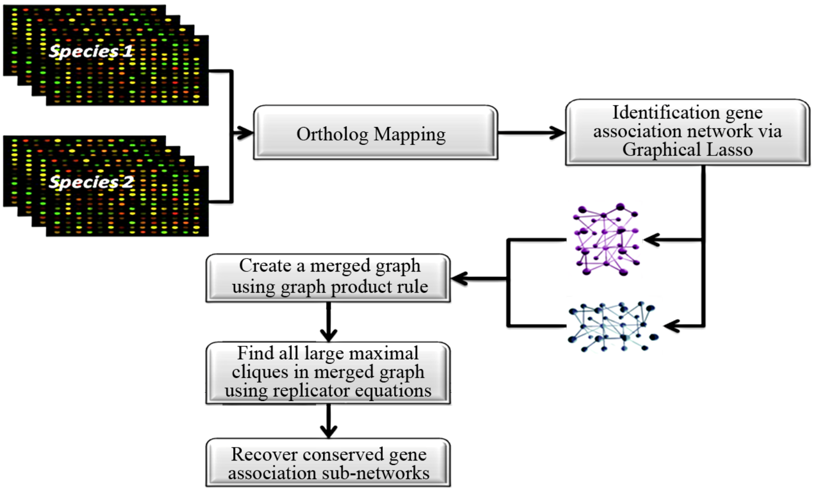
Species 1
Species 2
Ortholog Mapping
Identification gene association network via Graphical Lasso
Create a merged graph using graph product rule
Find all large maximal cliques in merged graph using replicator equations
Recover conserved gene association sub-networks

Figure 1. Clique-based approach to identification of common gene association sub-network from two species.

if $\Sigma_{ij}^{-1} = 0$ [6]. Meinshausen and Muhlmann [10] estimated a sparse graph by fitting a lasso model to each variable using others as predictors. Friedman *et al.* made this method faster by solving a 1000-node problem (≈500,000 parameters) in less than 1 minute [6]. Consider a multidimensional normal distribution of dimension *p*, with mean μ and covariance matrix Σ. Let *S* be the empirical covariance matrix, the estimation of Σ^{-1} is the solution to the following optimization problem:

$$\Sigma^{-1} = \arg\min \log\det\Sigma^{-1} - tr\left(S\Sigma^{-1}\right) - \rho\left\|\Sigma^{-1}\right\|_1 \qquad (1)$$

where *tr* denotes trace and $\left\|\Sigma^{-1}\right\|_1$ the L_1 norm, and ρ is the user-defined penalty. Banerjee *et al.* [11] show the problem is convex and considered estimating Σ rather than its inverse. Let *W* be the estimate of Σ, the problem is solved by optimizing over each row and corresponding column of *W* in a block coordinate descent fashion. Partition *W* and S as:

$$W = \begin{pmatrix} W_A & W_b \\ W_b^T & W_c \end{pmatrix}, S = \begin{pmatrix} S_A & S_b \\ S_b^T & S_c \end{pmatrix} \qquad (2)$$

where W_A is a $(P-1)\times(P-1)$ matrix, and W_b is a vector of length $(P-1)$, and W_c is a scalar. The dimensions are the same for the corresponding partitions in *S*. The solution satisfies

$$w_b = \arg\min_y \left\{y^T W_A^{-1} y : \left\|y - s_b\right\|_\infty \le \rho\right\} \qquad (3)$$

This is a box-constrained quadratic program and can be solved using an interior-point procedure [12]. By permuting the rows and columns so the target column is always the last, they solve a similar problem like (3) for each column, updating their estimate of *W* after each stage [11]. This is repeated until convergence. If this procedure is initialized with a positive definite matrix, they showed that the iterates from this procedure remains positive definite and invertible even if $p > N$ which is normally the case for gene expression data. That is also one of the reasons that we choose this method to construct gene association network. Using convex duality, Banerjee *et al.* showed that solving (3) is equivalent to solving the following minimization problem which resembles a L_1 regularized least squares problem.

$$\hat{\beta} = \arg\min_\beta\left(\frac{1}{2}\left\|W_A^{1/2}\beta - W_A^{-1/2}{}_{s_b}\right\|^2 + \rho\|\beta\|_1\right) \qquad (4)$$

The solution for problem (3) is $w_b = W_A\hat{\beta}$. Algorithm 1 describes the procedure to compute W, the estimate of Σ. In algorithm 1, $\hat{\beta}$ is solved using coordinate descent described by Friedman *et al.* [13] and Wu and Lange [14]. The threshold T is typically defined as $t \cdot ave\left|S^{-\text{diag}}\right|$, where $S^{-\text{diag}}$ are the off-diagonal elements of the empirical covariance matrix S, and t is typically set to a small number such as 0.001. The computing efficiency can be further improved via active set convergence [13].

Algorithm 1: Graphical Lasso Algorithm

```
$W \leftarrow S + \rho I$
While $|\Delta W|$ is less than a user defined threshold $T$
For $i$ = 1 to $p$
    Construct $W_A$ by removing $i^{th}$ row and $i^{th}$ column from matrix $W$
    Construct $s_b$ by removing $i^{th}$ element from $S_i$
    Solve $\hat{\beta}$ using coordinate descent
    $w_b \leftarrow W_A\hat{\beta}$
    Form vector $A$ by inserting $W_{ii}$ into $i^{th}$ position of $w_b$
    Update the $i^{th}$ row of $W$ with $A$
    Update the $i^{th}$ column of $W$ with $A^T$
End For
End While
```

2.2. Identification of Conserved Gene Association Sub-Networks

Detecting common sub-network is a challenging task. However, we are approaching this problem from a graph product point of view. We will merge the two graphs by mapping corresponding orthologous genes and create the edges for merged graph $G_m = (V, E_m)$ based on the following graph product rule.

$$E_m = \{(i,j)\in V, i\neq j, (i,j)\in E_1 \Leftrightarrow (i,j)\in E_2\} \tag{5}$$

In other words, an edge in E_m indicates both E_1 and E_2 contain the edge or neither of them contain the edge. Finding common sub-networks in G_1 and G_2 can be reduced to finding maximal cliques in G_m. A subset of vertices *C* is called a clique if all its vertices are mutually adjacent. A clique is said to be maximal if it is not contained in any larger clique. Pelillo has established equivalence between the graph isomorphism problem and the maximal clique problem [7]. The Motzkin-Straus theorem [9] has established a connection between the maximal cliques and the local maximizers of the following quadratic function:

$$\begin{aligned}&\text{maximize } f(x) = x^{\mathrm{T}}Ax = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}x_i x_j \\ &\text{subject to } x\in\Delta\end{aligned} \tag{6}$$

where $\Delta = \{x\in\mathbb{R}^p : x\geq 0 \text{ and } |x|_1 = 1\}$ is the standard simplex of $\mathbb{R}^p$, and *A* is the adjacency matrix for G_m with $A_{ij} = \begin{cases}1, \text{if } (i,j)\in E_m \\ 0, \text{Otherwise}\end{cases}$. Specifically, it states that a subset of vertices C of a graph is a maximum clique if and only if its characteristic vector X^c is a global maximizer of f on Δ. A similar relationship holds between local maximizers and maximal cliques [15]. The Motzkin-Straus theorem has served as the basis of many clique-finding procedures [16-18]. Pardalos and Phillips [17] observed that there existed spurious solutions to the original Motzkin-Straus formula, and Pelillo and Jagota [15] confirmed this finding later in 1996. Bomze provided a straight-forward solution to this problem [19]. Consider the following regularized version of function f: $\hat{f}(x) = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}x_i x_j + \frac{1}{2}\sum_{i=1}^{p} x_i^2$, which is obtained from (6) by substituting the adjacency matrix A with $W = A + \frac{1}{2}I_p$, where I_p is the $p\times p$ identity matrix. We can avoid spurious solutions by substituting A with W.

The optimization problem $f(x) = x^{\mathrm{T}}Wx, x\in\Delta$ may have many local maxima. Each large local maximum correspond to a true common gene association sub-network, while small local maxima usually result from noises and outliers. Given an initialization $x(1)$, the corresponding local solution x^* can be efficiently obtained by replicator equation, which arises in evolutionary game theory. The discrete time version of first-order replication equation has the following form:

$$x_i(t+1) = x_i(t)\frac{(Wx(t))_i}{x(t)^{\mathrm{T}}Wx(t)} \tag{7}$$

The simplex Δ is invariant under these dynamics, which means that every trajectory starting in Δ will remain in Δ. It has been proven that when W is symmetric and non-negative, the objective function $f(x) = x^{\mathrm{T}}Wx$ is strictly increasing along any non-constant trajectory and its asymptotically stable points are in one-to-one correspondence to strict local solution of (6).

To find all large local maximizers $\{x^*\}$, we will start with different initializations that will lead to different local maxima. The procedure for finding all large local maxima is described in Algorithm 2.

Algorithm 2: Finding all large local maxima

```
$W \leftarrow A + 0.5I$
For $v$ = 1 to $p$
    $N(v) \leftarrow$ set of vertices adjacent to vertex $v$
    $C = N(v) \cup v$
    For $i$ = 1 to $p$
        If $i \in C$
            $X[i] \leftarrow 1/|C|$
        Else
            $X[i] \leftarrow 0$
    End For
    //find local maximal using replicator equation
    Set ChangeFlag to true
    While ChangeFlag stays true
        $wx \leftarrow W \times X$   ($wx$ is a length p vector)
        $fx \leftarrow X^{\mathrm{T}}WX$
        For $i$ = 1 to $p$
            $Y[i] \leftarrow X[i] \times wx[i] / fx$
        End For
        If $Y == X$
            Change Flag $\leftarrow$ False
        Else
            $X \leftarrow Y$
    End While
    $solution[v].X \leftarrow X$
    $solution[v].score \leftarrow fx$
End For
Output all local maximizers
```

For a local maximizer X, we need to recover the corresponding common gene association sub-network between the two species. The non-zero elements in local maximizer X correspond to the genes that share the

Copyright © 2013 SciRes.

same association network between the two species. It is possible to have some genes in the shared network that are independent. We need to remove these isolated genes from the conserved association network. Algorithm 3 describes how to recover the common association network.

```
Algorithm 3: Recover Conserved Gene Association Network
Input: Local maximizer $X$ and adjacency matrix $M$ for one species
$L \leftarrow \varnothing$
For $i = 1$ to $p$
    If $X[i] \neq 0$
        For $j = 1$ to $p$
         if $M_{ij} \neq 0$
            $L \leftarrow L \cup i$ and break
        End For
    End If
End For
Output conserved association network (subset of M consisting of genes in $L$)
```

3. Experimental Data

3.1. Graph Benchmark Data

We used the DIMACS benchmark data set [20] to validate the effectiveness of the replicator equations to find maximal cliques.

3.2. Gene Expression Data

We applied the method to find common stress responsive gene association networks for two related bacteria *Deinococcus radiodurans* and *Thermus thermophilus*. We downloaded the gene expression data sets GSE 29516 for *D. radiodurans* from gene expression omnibus [21]. GSE29516 consists of microarray data from transcription profiling of *D. radiodurans* treated with 0.3 M NaCl or 2 M salt. We downloaded the gene expression series GSE21289 for *T. thermophilus*. GSE21289 contains the gene expression data of *T. thermophilus* HB8 wild-type strain in response to high salt stress.

3.3. Ortholog Mapping

Ortholog mapping is done via multi-genome homology comparison tools available from the Comprehensive Microbial Resource web site [22]. In the case of multiple genes in a cluster, we used the one with the highest score, resulting in 744 one-to-one ortholog pairs. Gene expression data and ortholog mapping table is stored in an in-house relational database for easy retrieval and cross-referencing across the two species Gene expression data for orthologous genes are retrieved through database queries.

4. Results and Discussions

4.1. Benchmark Results on DIMACS Challenge Sets

We recorded running time and clique found using our replicator equations. **Table 1** shows the results on some DIMACS challenge instances.

In general, our implementation is able to find maximal cliques in reasonable time. And we were able to find cliques that are close to their corresponding maximum clique numbers for each benchmark data set.

4.2. Identification Common Gene Association Sub-Networks

Two common gene association sub-networked were identified using the described procedure (**Figure 2**).

Annotations of the genes in **Figure 2** are given in **Table 2**.

Table 1. The performance of clique finding algorithm on some DIMACS challenge instances.

Benchmark	Time(sec)	iteration	cliqueFound	MaxClique
c-fat200-1	0.01	50	12	12
c-fat200-2	0.02	50	24	24
c-fat500-1	0.03	50	14	14
c-fat500-2	0.05	50	26	26
brock200_1	0.55	1000	20	21
brock200_2	0.32	1000	11	12
brock200_3	0.39	1000	14	15
brock200_4	0.44	1000	16	17
brock400-2	6.67	5000	25	29
brock400-4	6.66	5000	25	33
brock800-1	14.28	5000	20	23
brock800-2	14.72	5000	20	24
brock800-4	14.35	5000	20	26
hamming6-2	0.01	50	32	32
hamming6-4	0	50	4	4
johnson8-2-4	0	25	4	4
johnson8-4-4	0	25	14	14
keller4	0.02	50	11	11
keller5	2.28	500	27	27
keller6	76.75	1000	53	59
p_hat300-1	0.05	100	8	8
p_hat300-2	0.09	100	25	25
p_hat300-3	0.12	100	33	36
p_hat500-1	0.08	100	9	9
p_hat700-1	0.12	100	9	11
p_hat700-2	0.22	100	43	44
p_hat700-3	0.41	100	60	62
p_hat1500-1	0.3	100	10	12
p_hat1500-2	1.03	100	64	65
p_hat1500-3	1.74	100	88	94

Copyright © 2013 SciRes.

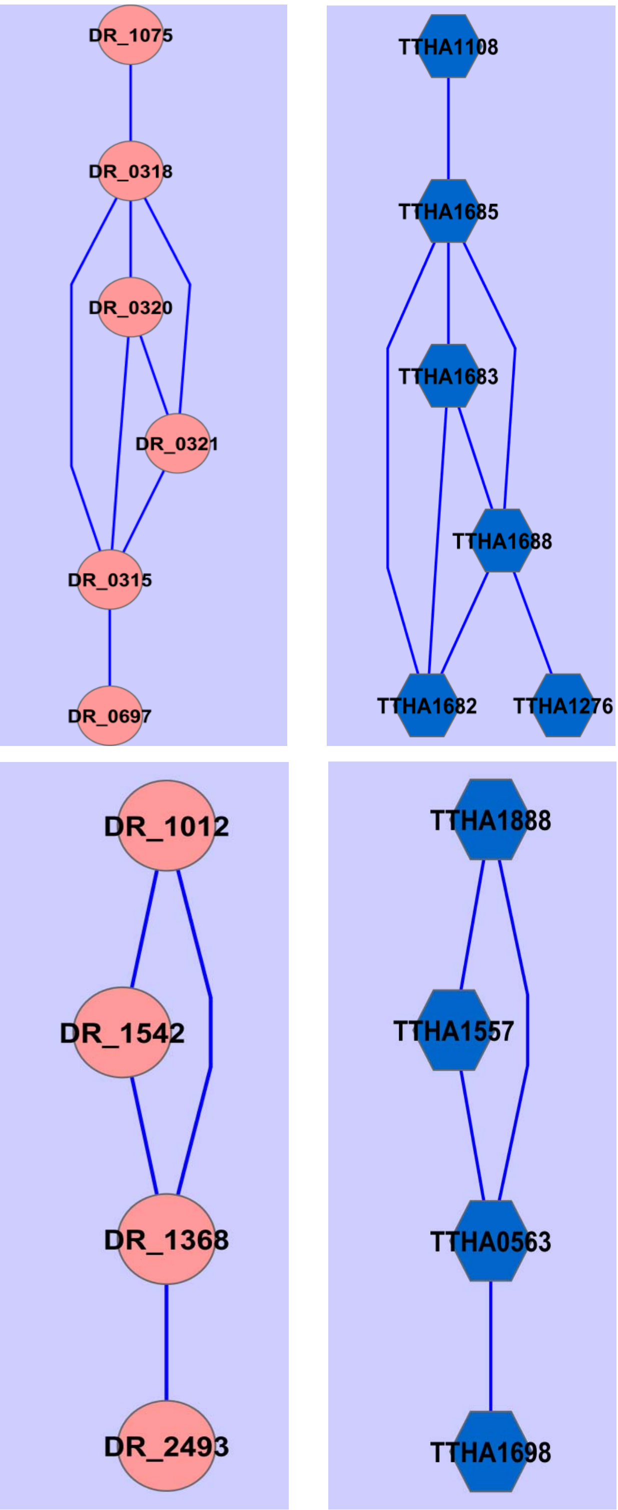

Figure 2. Common gene association sub-networks identified for *D. radiodurans* and *T. thermophilus*.

Table 2. Annotation of genes in the identified common sub-net-networks.

Gene ID	Definition
DR_1075/TTHA1108	Hypothetical protein
DR_0318/TTHA1685	Heat shock protein 83-1
DR_0320/TTHA1683	30 S ribosomal protein S17
DR_0315/TTHA1688	30 S ribosomal protein S19
DR_0321/TTHA1682	50 S ribosomal protein L14
DR_0697/TTHA1276	v-type ATP sythase subunit E
DR_1012/TTHA1888	ABC transporter ATP-binding protein
DR_1542/TTHA1557	Propionyl-CoA carboxylase subunit beta
DR_1368/TTHA0563	Hypothetical protein
DR_2493/TTHA1698	carboxypeptidase G2

Among these genes, ABC transporter ATP-binding protein has been previously reported to be osmo-regulated [23]. Two of the genes are not fully annotated (hypothetical protein in **Table 2**), we think they are worthy of further investigation because they seem to be related to osmosis stress response based on our study on two different species. A ribosomal protein has been found by Schmalisch *et al.* to be general stress protein in *Bacillus subtilis* [24]. We found three ribosomal proteins that are related to the stress response in both *D. radiodurans* and *T. thermophilus*. This is consistent with the finding from Schmalisch *et al.* [24].

5. Conclusions

In this study, we developed an efficient computational framework that combines graphical lasso model, graph product and replicator equation based clique solver to identify common gene association sub-network from multiple species. Our method provides an approach to identifying conserved pathway components.

We applied our method and identified common gene association sub-networks for two related bacterial species *D. radiodurans* and *T. thermophilus* subjected to similar environmental stress. We confirmed some stress responsive genes with previous studies. Our method also demonstrated how these genes interact with other genes and these interactions potentially are conserved because they are discovered via simultaneous study of two related species.

Our method is not limited to finding common gene association sub-network across multiple species. It can also be adapted to identify core interaction network for the same species subjected to different environmental stresses. It can also be employed to identify common gene/protein sub-networks for related diseases such as diabetes and hypertension.

6. Acknowledgements

This work was supported by a grant from North Carolina Space Consortium (2010-1435-CE-02). Financial support from the National Science Foundation (NSF) through a grant to the Quality Education for Minorities Networks

(QEM/HBCU-UP Faculty Professional Development and Mentoring Program) are highly acknowledged. The work was also supported in part by NSF CREST HRD-0833184 and by the US Army Research Office (ARO) W911NF-0810510.

REFERENCES

[1] J. Schäfer and K. Strimmer, "An Empirical Bayes Approach to Inferring Large-Scale Gene Association Networks," *Bioinformatics*, Vol. 21, No. 6, 2005, pp. 754-764.

[2] P. Langfelder and S. Horvath, "WGCNA: An R Package for Weighted Correlation Network Analysis," *BMC Bioinformatics*, Vol. 9, No. 1, 2008, p. 559. doi:10.1186/1471-2105-9-559

[3] N. Friedman, "Inferring Cellular Networks Using Probabilistic Graphical Models," *Science*, Vol. 303, No. 5659, 2004, pp. 799-805. doi:10.1126/science.1094068

[4] M. K. S. Yeung, J. Tegnér and J. J. Collins, "Reverse Engineering Gene Networks Using Singular Value Decomposition and Robust Regression," *Proceedings of the National Academy of Sciences*, Vol. 99, No. 9, 2002, pp. 6163-6168.

[5] C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D. L. Wild and F. Falciani, "Modeling T-Cell Activation Using Gene Expression Profiling and State-Space Models," *Bioinformatics*, Vol. 20, No. 9, 2004, pp. 1361-1372.

[6] J. Friedman, T. Hastie and R. Tibshirani, "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, Vol. 9, No. 3, 2008, pp. 432-441.

[7] M. Pelillo, "Replicator Equations, Maximal Cliques, and Graph Isomorphism," *Neural Computation*, Vol. 11, No. 8, 1999, pp. 1933-1955.

[8] R. Battiti and M. Protasi, "Reactive Local Search for the Maximum Clique Problem," *Algorithmica*, Vol. 29, No. 4, 2001, pp. 610-637. doi:10.1007/s004530010074

[9] T. S. Motzkin and E. G. Straus, "Maxima for Graphs and a New Proof of a Theorem of Turán," *Canadian Journal of Mathematics*, Vol. 17, 1965, pp. 533-540. doi:10.4153/CJM-1965-053-6

[10] N. Meinshausen and P. Bühlmann, "High-Dimensional Graphs and Variable Selection with the Lasso," *The Annals of Statistics*, Vol. 34, No. 3, 2006, pp. 1436-1462. doi:10.1214/009053606000000281

[11] O. Banerjee, L. E. Ghaoui and A. d'Aspremont, "Model Selection through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data," *Journal of Machine Learning Research*, Vol. 9, 2008, pp. 485-516.

[12] A. Forsgren, P. E. Gill and M. H. Wright, "Interior Methods for Nonlinear Optimization," *SIAM Review*, Vol. 44, No. 4, 2002, pp. 525-597. doi:10.1137/S0036144502414942

[13] J. Friedman, T. Hastie, H. Hofling and R. Tibshirani, "Pathwise Coordinate Optimization," *Annals of Applied Statistics*, Vol. 1, No. 2, 2007, p. 302. doi:10.1214/07-AOAS131

[14] T. Wu and K. Lange, "Coordinate Descent Procedures for Lasso Regularized Regression," *Annals of Applied Statistics*, Vol. 2, No. 1, 2008, pp. 224-244. doi:10.1214/07-AOAS147

[15] M. Pelillo and A. Jagota, "Feasible and Infeasible Maxima in a Quadratic Program for Maximum Clique," *Journal of Artificial Neural Network*, Vol. 2, No. 4, 1996, pp. 411-420.

[16] L. E. Gibbons, D. W. Hearn and P. M. Pardalos, "A Continuous Based Heuristic for the Maximum Clique Problem," In: D. S. Johnson and M. A. Trick, Eds., *Cliques, Coloring, and Satisfiability—Second DIMACS Implementation Challenge*, American Mathematical Society, 1996, pp. 103-124.

[17] P. M. Pardalos and A. T. Phillips, "A Global Optimization Approach for Solving the Maximum Clique Problem," *International Journal of Computer Mathematics*, Vol. 33, No. 3-4, 1990, pp. 209-216.

[18] M. Pelillo, "Relaxation Labeling Networks for the Maximum Clique Problem," *Journal of Artificial Neural Network*, Vol. 2, No. 4, 1996, pp. 313-328.

[19] I. M. Bomze, "Evolution towards the Maximum Clique," *Journal of Global Optimization*, Vol. 10, No. 2, 1997, pp. 143-164. doi:10.1023/A:1008230200610

[20] D. S. Johnson and M. A. Trick, Eds., "Cliques, Coloring, and Satisfiability: Second DIMACS Implementation Challenge," AMS, Providence, 1996.

[21] T. Barrett and R. Edgar, "Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis," *Methods in Enzymology*, Vol. 411, 2006, pp. 352-369. doi:10.1016/S0076-6879(06)11019-8

[22] J. D. Peterson, L. A. Umayam, T. Dickinson, E. K. Hickey and O. White, "The Comprehensive Microbial Resource," *Nucleic Acids Research*, Vol. 29, No. 1, 2001, pp. 123-125.

[23] T. van der Heide and B. Poolman, "Osmoregulated ABC-Transport System of *Lactococcus lactis* Senses Water Stress via Changes in the Physical State of the Membrane," *Proceedings of the National Academy of Sciences*, Vol. 97, No. 13, 2000, pp. 7102-7106.

[24] M. Schmalisch, I. Langbein and J. Stulke, "The General Stress Protein Ctc of *Bacillus subtilis* Is a Ribosomal Protein," *Journal of Molecular Microbiology and Biotechnology*, Vol. 4, 2002, pp. 495-501.