

Construction of malaria gene expression network using partial correlations

Raya Khanin
Department of Statistics
University of Glasgow
44-141-3306139
raya@stats.gla.ac.uk

Ernst Wit
Department of Statistics
University of Glasgow
44-141-3306820
ernst@stats.gla.ac.uk

ABSTRACT

In this paper we aim to model the gene network of *Plasmodium falciparum* using the microarray dataset from [3]. A network is constructed based on a novel method that combines two types of correlations between each pair of genes: standard Pearson and partial correlations. A link is established between two genes if both correlation coefficients are higher than their corresponding thresholds. The topology of obtained malaria gene expression network is consistent with the scale-free behavior wherein a small number of genes have a large number of links and many genes have only a few connections, similarly to other biological networks. The highly connected genes (hubs) are enriched with the essential genes that are involved in central cell functions. In addition, a constructed network has a small-world property with any two genes being connected by a path of a few links only. To verify the proposed method and to compare the results, the gene network was also constructed using a dataset from [8]. This network also exhibits both scale-free and small-world properties.

Keywords

gene expression network, partial correlation, scale-free behaviour

1. INTRODUCTION

The objective of this study is to construct gene expression network of *Plasmodium falciparum* using microarray experiments from [3]. Unraveling the topology of the malaria gene network is relevant to the understanding of the cell functioning and invasion cycle of the parasite. We aim to investigate the global topological structure of the constructed network using a graph-theoretical approach. We study the statistical properties of the obtained network, such as the distribution of connectivity per node and the clustering coefficients. In the network, the genes are the nodes that are connected if certain criteria, like co-expression, are

satisfied. Previously studied biological networks, including gene expression networks of other organisms, have been shown to exhibit scale-free behaviour, wherein a number of connections per gene (node) is distributed according to the power-law: $N(k) \sim k^{-\gamma}$. It means that there are many genes with few connections and a small number of genes that are highly connected (hubs). Analyses of gene networks have shown a correlation between essentiality of a gene and a number of connections that the gene has: highly connected genes (hubs) are often essential (involved in central biological functions) and evolutionary conserved ([2], [7]). For *Plasmodium falciparum* 65% of annotated genes encode hypothetical proteins of unknown functions. In addition, 60% of genes lack sequence similarity to genes from any other known organism. It makes ascribing putative roles for these genes a challenging task. One of the potentials of the gene network analysis is to obtain clues on the putative roles of such genes of unknown functions based on the gene connectivity, position in the network and other genes they have links with.

It would also be interesting to see whether the gene network analysis can give some support to the hypothesis advanced in [3] on multilayer regulatory network wherein a comparatively small number of transcription factors with overlapping binding site specificities could account for the entire cascade. The authors speculated further that disruption of a key regulatory element (lethal gene) might have a profound inhibitory effect on the entire network. Such lethal genes are most likely to be among the highly connected nodes in the malaria network.

2. Method for construction of expression network

There have been a number of studies where global gene networks are constructed from microarray data based on the Pearson correlation coefficients. Two genes are considered linked in the network if their correlation is higher than the threshold ([2], [10]). Sometimes one also takes into account empirically calculated p-values for the correlations between two genes [4]. The Pearson correlation has been shown to play an important role in inferring interactions between genes [6]. However, methods that are based on only standard correlations are too simplistic and inevitably overestimate the number of links (connectivity) per gene. It is a common knowledge that correlation coefficient indicates not only

nodes that have direct connections but also nodes with indirect connections. It is also plausible that some of the important connections are left out if the threshold is not high enough. However, lowering the correlation threshold will significantly increase the number of potential links, including many random ones. In the case of the contest malaria dataset, this problem becomes even more transparent due to very highly coordinated expression of genes. A network constructed from the overview malaria dataset by thresholding correlations while restricting the average connectivity per node, $\langle k \rangle$, results in very high threshold values, P . For example, to obtain a network with $\langle k \rangle = 50$ the threshold $P = 0.935$ is required, while restricting the average connectivity to a lower value $\langle k \rangle = 30$ that is being reported for other biological networks results in the value of $P = 0.95$. This is an unreasonably high value. Given the noisy data, missing values and the complexity of biological networks many biologically relevant connections will not be included in such network. For $P = 0.8$ the constructed network is not sparse and not scale free (see Fig.S1 on the web-page with supplemental data www.stats.gla.ac.uk/~raya/suppldata.html). In fact, this network includes about 15% of all possible links, with an average number of links per node, $\langle k \rangle = 470$, being more than ten times high than the average connectivity for the gene networks of other organisms constructed by the same method (e.g. $\langle k \rangle = 32$ for sparse scale-free network of yeast constructed with $P = 0.6$ [10]).

Here we propose to use partial correlations to filter the more likely links out of a much larger set of potential links with high standard correlations. The partial correlation coefficient of two genes measures the strength of the relation between these genes after the effect of other genes is removed or fixed, therefore indicating whether two genes are directly or indirectly linked. The partial correlations have been used in Gaussian Graphical Models (GGM) to characterize strength of correlations between pairs of genes in the regulatory networks. We propose to construct a gene expression network from a large gene dataset by thresholding both Pearson and partial correlation coefficients for each pair of genes. Namely, two genes (nodes) (i, j) are considered connected by a link if the Pearson correlation of their profiles p_{ij} is higher than

a certain cut-off value, P , and their partial correlation coefficient r_{ij} is higher than a cut-off value, r . The partial correlation of genes i and j with respect to other genes whose effect is removed (fixed) is given by

$$r_{ij} = \frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}},$$

where $\Omega = P^{-1} = \omega_{ij}$ is the inverse (or pseudo-inverse) of Pearson correlation matrix, P . To overcome the degeneracy problem of the correlation matrix P for small samples, partial

correlation estimators based on Moore-Penrose pseudo-inverse of correlation matrix were introduced in [9]. In this work we compute partial correlations by using pseudo-inverse of the correlation matrix (*cor2pcor()* function from R-package *GeneTS* [9]). Two genes (i, j) are connected by a link if their Pearson correlation is higher than a cut-off value, P , and their partial correlation is higher than a cut-off value, r : $i \leftrightarrow j : p_{ij} \geq P \ \& \ r_{ij} \geq r$.

3. Application to malaria data-set

For the study of malaria gene regulatory network, we used two datasets. The first one is the overview data-set from the complete intraerythrocytic developmental cycle (IDC) transcriptome of the *Plasmodium falciparum* measured at 46 time-points [3]. To verify the results, we have also used dataset (2234 genes) used for clustering the gene expressions measured at nine time-points in human and mosquito stages of malaria parasite's life-cycle [8]. We will further refer to this dataset as the validation dataset. For the overview dataset, the values from multiple oligonucleotides representing the same gene were averaged, resulting in 3048 genes. In the rest of the paper we will concentrate on reporting the results for the overview dataset. The topology of the network constructed using the validation dataset is very similar (see Tables S5 and S6 on the supplemental web-page).

The distribution of connectivities (number of connections) per node, N , is consistent with the power-law for both datasets. Fig.1 shows $N(k)$ for several values of thresholds P and r . Similar distributions have been reported for yeast and other organisms [2]. Power-law distributions imply that the network exhibits the so-called scale-free behavior with only a few genes having a high number of connections (hubs), while the others having moderate or low number of connections. The qualitative topological properties of the malaria network are insensitive to the precise thresholds within a range of values: taking the thresholds $0.45 \leq r \leq 0.6$ for $P = 0.7, 0.8$ yields scale-free distribution of connectivities that are qualitatively similar (Figure 1). Values outside this region result in other types of networks: $r \leq 0.4$ result in networks whose connectivities do not obey power-law (Figure S2d) while $r \geq 0.6$ and/or $P \geq 0.8$ yield too few links (Figure S2b,c). The power-exponents, $\hat{\gamma}$, have been found by the maximum likelihood method from the fitting the power-law distribution to the observed values (i.e. constructed connectivities in this case). Values of $\hat{\gamma}$ are within the range $0.6 - 1.4$ for different values of thresholds P and r . $\gamma = 0.6$ is for the parameters $P = 0.7, r = 0.45$ that produce a network with an average connectivity per node of $\langle k \rangle \sim 28$ and maximum connectivity $k_{\max} = 133$ and $\hat{\gamma} = 1.4$ is for $P = 0.8, r = 0.6$ with $\langle k \rangle \sim 4$, $k_{\max} = 30$. Other values of parameters resulted in networks with average connectivities in between these two values (see Table S1). These values are consistent with previously

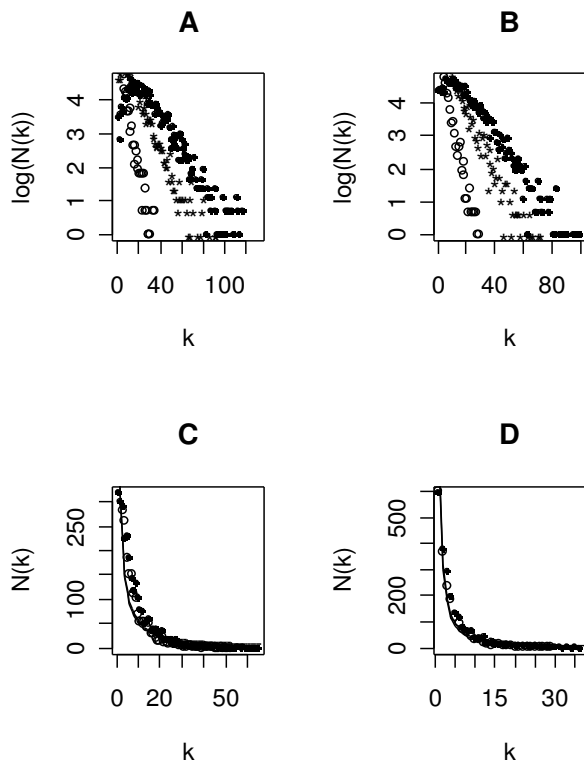


Figure 1. Connectivity distribution of nodes in a malaria gene network constructed from the overview data-set for different values of thresholds. (A-B) Log distribution of connectivities for $r = 0.45; 0.5; 0.55$ and $P = 0.7$ (A) and $P = 0.8$ (B). (C-D) Distribution of observed connectivities and fitted power-law $N(k) \sim k^{-\gamma}$ for $r = 0.5$ and $P = 0.8, \hat{\gamma} = 0.91$ (C) and $P = 0.7, \hat{\gamma} = 0.84$ (D).

reported values for other organisms (e.g. $\hat{\gamma} = 1$ for yeast [10]).

A quantifier of the degree of network modularity is the clustering coefficient that measures the extent the genes that are connected to a specific gene are also linked among themselves. The clustering coefficient of a gene i is $c_i = 2n_i / k_i(k_i - 1)$, where n_i is the number of links connecting the k_i neighbours of gene i forming triangles and $k_i(k_i - 1)/2$ is the total number of triangles that could pass through the node i . The average clustering coefficient $\langle C \rangle$ characterizes the overall tendency of genes to form clusters. For different values of parameters clustering coefficients have been found to be within the range $\langle C \rangle = 0.195$ for $P = 0.7, r = 0.45$ and $\langle C \rangle = 0.443$ for $P = 0.8, r = 0.6$. These values are much higher than the value for random networks $\langle C \rangle \sim 0.008$ [1] and they are consistent with the values reported for other organisms (e.g. $\langle C \rangle = 0.6$ for yeast [10]).

High average clustering coefficients are indicative of so-called small world structure of the network, meaning that any two genes in the constructed malaria gene network are connected with a path of a few links only.

In the next section we report results for the threshold values $P = 0.7, r = 0.5$. These parameters yield the network statistics that are similar to previously studied networks with a maximum connectivity $k_{\max} = 101$, average connectivity per node $\langle k \rangle \sim 15$, the power-law exponent $\hat{\gamma} = 0.84$. To find whether this network constructed by thresholding two types of correlation coefficients is significant or whether it can be found by chance, we performed permutation test. We independently permute components of each gene time-profile, recompute correlation and partial correlation matrices, and establish a link between genes i and j if the thresholding conditions are satisfied. In other words, for each pair of genes we are testing a null hypothesis that they are not linked. 100 permutation tests resulted in more than 200 p-values being equal 0.01, with all other p-values being zero. False Discovery Rate procedure with the control level of 10% (i.e. allowing to accept 10% of false links) resulted in all links found from the overview dataset being significant, i.e. null hypothesis being rejected and links being established. This allows us to conclude that the network of about 3000 genes found by the thresholding method is unlikely to be found by chance. This network is worth investigating further for some proof-of-principle results.

3.1 Connectivity and essentiality

It has been previously reported that high degree nodes in gene expression networks constructed for other organisms are more likely to correspond to essential genes, i.e. to be involved in central biological functions of the cell [2]. Among the top 66 hubs with connectivities from $k_{\max} / 2$ there are 13 with no manual annotation and 7 belong to Plastid genome. Among the annotated genes, only 35% of genes code for proteins with identifiable functions (on average). Among the hubs of the constructed malaria gene network 7 genes (PFI1340w, PFI1360c, PFI0385c, PF13_0229, PF14_0373, PFA0345w, PF11_0298) are known to have the cell essential functions in cell growth, and/or maintenance, metabolism, energy pathways, biosynthesis. In addition, a rho-try protein (PFI0265c), a papain family cysteine protease (PFI0135c), an early transcribed membrane protein (PF10_0019) are also in the list of the hub-genes. Among 5 hubs on chromosome 9, three (PFI1340w, PFI1360c, and PFI0385c) are prescribed functions in cell growth, maintenance and metabolism and they are all connected among themselves forming a triangular network motif. The largest reported ORF (MAL6P1.147) also has a large number of links, half of maximum connectivity. 8 other genes are either conserved or have homologues/similar to proteins in other organisms. The list of 66 top hubs for the network constructed from the validation dataset with $P = 0.8, r = 0.5$ contains

20 (i.e. virtually all annotated genes in the list) genes with cell growth/maintenance, cell communication and other central cell functions. For a full list of hubs in networks constructed for the overview and the validation datasets see Tables S2 and S6 on the supplemental web-page.

As another proof-of-principle, we looked at how many hubs are in the set of 6% genes that were found to be common to all four stages of the parasite life cycle (supplementary table 1 in [5]). This list contains most housekeeping genes and their products, such as ribosomal proteins, transcription factors, and cytoskeletal proteins. It turned out that 15 hubs from our list are among the set of common genes found in [5]. This is about 50% percent of all hubs (excluding plastid genome and genes that are not manually annotated). This clearly demonstrates that the hubs in the constructed malaria gene network are enriched with essential genes.

As the gene expression network for malaria exhibits properties characteristic for other organisms, it is likely that genes with unknown functionalities among the hubs belong to the class of essential or lethal genes. We looked at how hubs with unknown functions in the overview network clustered in the experiments of [8]. We found that among 25 genes with hypothetical proteins of unknown functions that are present in the validation dataset, 10 genes belong to cluster 13, 5 to cluster 12 and 5 to cluster 15. It has been reported that known genes in clusters 12,13 are mainly involved in cell-cycle regulation and progression at trophozoite stage, while cluster 15 is characterized as having genes with roles in cell invasion that are under evaluation as blood-stage vaccine. According to [8], genes from the clusters 12 and 13 may represent potential targets for drugs focused on disruption of the highly replicating trophozoite stage of the parasite, while additional candidate vaccine antigens could come from yet uncharacterized of the cluster 15. So, hubs with unknown functions warrant further investigation. It will also be interesting to investigate those genes among hubs that have not been manually annotated (see Table S3 that contains oligonucleotides of hubs from the overview network with no manual annotation).

Finding hubs in the malaria gene network is extremely important in guiding the search for the malaria vaccine. Targeting a highly connecting node by a drug will result in inactivation of that protein that could be fatal to the whole life-cycle of the malaria parasite; removing a less connected node will barely affect the whole system.

3.2 Some sub-networks

It might be interesting to investigate further some sub-networks of the large malaria gene network. As an example, we had a closer look at the glycolytic pathway as it is mentioned in [3] as the one that is well-preserved in malaria parasite. Among 9 genes taken from the <http://plasmodb.org> database as belonging to this pathway, we found that they share 5 links among themselves. In fact, the probability of 9 randomly picked genes to have 5 links

is 0.01% given the connectivity matrix. Given that some of the genes in this pathway are not present in the data-set, this result is encouraging. Our analysis did not pick up MAL61.160 as part of the glycolytic pathway, instead another putative copy, PF10_0363, was identified as a part of it, having 2 connections, as well as gene PF10_0155 that has 4 connections.

As another example, we had a look at all major candidates for vaccination (AMA1, EBA175, MSP1, MSP3, MSP7, RAP1, RESA1) studied in [3]. All these genes are very well positioned in the network, having connectivities between 20 and 40, well above the average connectivity, $\langle k \rangle \sim 15$. Interestingly, these vaccine candidates are connected among themselves as well as with some other merozoite invasion proteins (MSP6, MSP8). In addition, the neighbours of these vaccine candidates are enriched with myosin-like proteins, erythrocyte associated proteins, reticulocyte binding proteins, zinc finger proteins among others. For full list of the neighbours of these major vaccine candidates, see Table S4. FigS3 is a sketch of some links in the sub-network that connects these genes/proteins. There is a large number of hypothetical proteins that are linked to the vaccine candidates in our network. Several of the hypothetical proteins from the list are linked to two major vaccine candidates, while some hypothetical proteins (e.g. PF10_0352, PF07_0127, PFE0365c, PFC1045c, PFD0715c) have links with three major vaccine candidates and are probably worth having a closer look at.

4. Conclusions

In this paper we have constructed a model of malaria gene network by a novel method of thresholding both correlation and partial correlation coefficients. Both types of correlations are essential in revealing the connections of genes in the network. The constructed network is a small world scale-free network with the hubs being enriched by cell essential genes. This model of malaria gene network is worth investigating further, looking at various subnetworks, consisting of genes that are known to be involved in the same biological processes. Alternatively, one might want to look at the neighbours of genes with unknown functions. This might help the process of assigning putative functions to these genes. The matrix with links of the network studied in this paper can be found on the supplemental web-page. The thresholding approach used in this paper suffices for the goal of studying statistical properties of a biological network and it also gives encouraging proof-of-principle results. For the future, we plan to fit the network model based on two types of correlations using multiple testing procedures and machine learning techniques.

REFERENCES

- [1] Barabasi, A.L. and Oltvai Z.N. Network biology: understanding the cell's functional organization. *Nat Rev Genet.* 2 (Feb. 2004), 101-113.
- [2] Bergmann S., Ihmels J., and Barkai N. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.* 2(1), (Jan. 2004), E9.
- [3] Bozdech Z., Llinas M., Pulliam B.L., Wong E.D., Zhu J., and DeRisi J.L. The Transcriptome of the Intraerythrocytic Developmental Cycle of *Plasmodium falciparum*. *PLoS Biol.* 1(1), (Oct. 2003), E5.
- [4] Carter S.L., Brechbuhler C.M., Griffin M., and Bond A.T. Gene expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics.* Epub (Jul. 2004).
- [5] Florens L., Washburn M.P., Raine J.D., Anthony R.M., Grainger M., Haynes J.D., Moch J.K., Muster N., Sacci J.B., Tabb D.L., Witney A.A., Wolters D., Wu Y., Gardner M.J., Holder A.A., Sinden R.E., Yates J.R., and Carucci D.J. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature.* 3, 419 (Oct. 2002), 520-526.
- [6] Ideker T., Thorsson V., Ranish J.A., Christmas R., Buhler J., Eng J.K., Bumgarner R., Goodlett D.R., Aebersold R., and Hood L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science.* 292, 5518 (May, 2001), 929-934.
- [7] Jeong H., Tombor B., Albert R., Oltvai Z.N., and Barabasi A.L. The large-scale organization of metabolic networks. *Nature.* 407, 6804 (Oct. 2000), 651-654.
- [8] Le Roch K.G., Zhou Y., Blair P.L., Grainger M., Moch J.K., Haynes J.D., De La Vega P., Holder A.A., Batalov S., Carucci D.J., Winzeler E.A. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science.* 301, 5639 (Sep. 2003), 1503-1508.
- [9] Schafer J. and Strimmer K. An empirical Bayes approach to inferring large graphical gaussian models from microarray data. *Bioinformatics.* (submitted) <http://www.stat.uni-muenchen.de/~strimmer/publications/largeggm2004.pdf>
- [10] van Noort V., Snel B., and Huynen M.A. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.* 5, 3 (Mar. 2004), 280-284.