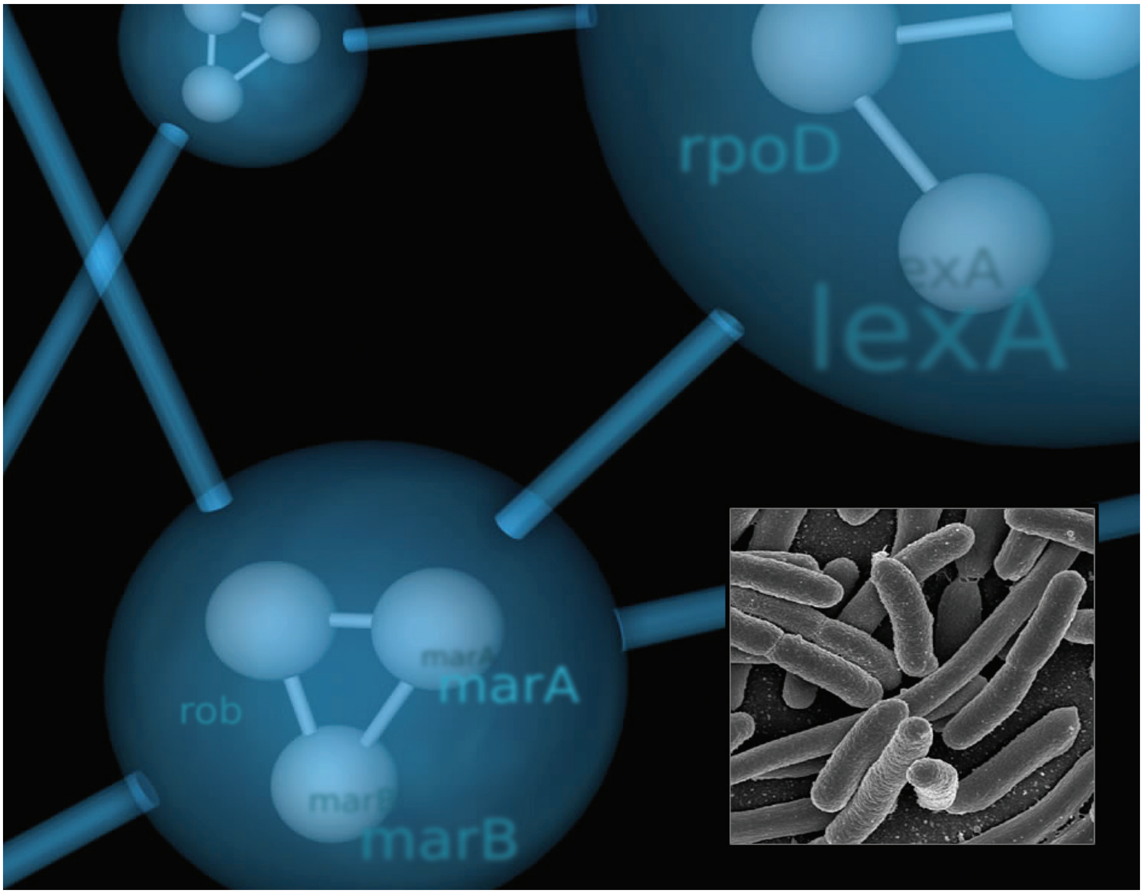

rpoD
lexA
marA
rob
marB

Prediction of conditional gene essentiality through graph theoretical analysis of genome-wide functional linkages†‡

P. Manimaran, Shubhada R. Hegde and Shekhar C. Mande*

Received 16th March 2009, Accepted 16th July 2009
First published as an Advance Article on the web 7th August 2009
DOI: 10.1039/b905264j

The genome of an organism characterizes the complete set of genes that it is capable of encoding. However, not all of the genes are transcribed and translated under any defined condition. The robustness that an organism exhibits to environmental perturbations is partly conferred by the genes that are constitutively expressed under all the conditions, and partly by a subset of genes that are induced under the defined conditions. The conditional importance of genes in conferring robustness can be understood in the context of the functional attributes of these genes and their correlations to the defined environmental conditions. However, *a priori* prediction of such genes for a given condition is yet not possible. We have attempted such predictions by integrating the available gene expression data with genome-wide functional linkages through the well known centrality–lethality correlations in graph theory. We make use of three distinct concepts of centrality, namely, degree, closeness and betweenness, which yield mutually complementary information. We then demonstrate the efficacy of combined graph theoretical and machine learning approaches in ranking essential nodes from a large network of genome-wide functional linkages, which yields predictions with high accuracy. We therefore perceive such predictions as highly useful in applications such as defining and prioritizing drug targets.

Introduction

Network theory has been an attractive route to study the emergent properties of complex systems. Its applications have been wide-ranging, including those from biological, physical and social sciences.[1,2] Many different network models have been proposed to address properties of complex systems, some of the important ones being the random graph model proposed by Erdos and Renyi,[3] the small world network proposed by Watts and Strogatz,[4] and the scale-free model proposed by Barabási and Albert.[5] The random graph model was developed to describe large networks whose organisational principles were not easily definable, with the assumption that interactions between nodes occur with a random probability 'P'.[3] From various applications, it was observed that this model was not appropriate for studying real world networks, because in a random graph model, the degree distribution follows a Poisson distribution, whereas most complex networks exhibit heterogeneous degree distribution. Its failure in studying the real world networks further paved the way for Watts and Strogatz to propose the 'small world network', a network with a small diameter and high clustering.[4] More recently, Barabási and Albert have proposed the scale-free model, in which the degree distribution possesses power law behaviour.[5] The most evident applications of network theory are in the study of technological networks such as Internet communications and power grids, social networks such as terrorists' networks and scientific collaboration networks, and biological networks such as metabolic pathway networks and protein–protein interaction networks.

Various topological measures such as degree distribution, clustering coefficient, modularity *etc.* have been proposed to be characteristics of the networks.[2,6] The topological measures are useful in understanding the overall properties of the systems, such as cohesiveness of interactions and their modular nature. Such measures are useful in understanding the global properties of the networks, yet characteristics of individual nodes are not apparent in the topological parameters. Certain applications of network theory may require analysis of the importance of nodes in terms of connectivity, information transfer capability and closeness to other nodes. To address such properties, the concept of centrality with reference to characterization of human communications in small groups has been proposed.[7] The centrality measures serve as tools in attempting to find influence of individual nodes in a network.[8] Various centrality measures have been developed, the most widely used being degree, closeness and betweenness. The importance of centrality measures has therefore attracted attention in order to characterize the individual nodes in a network.[9–14]

Biological networks follow characteristics of real world networks including resistance against random node failures. These characteristics are believed to be due to the scale-free properties of these networks,[1] which suggest that only a small number of nodes are highly connected, whereas a large

Centre for DNA Fingerprinting and Diagnostics, Gruhakalpa, 5-4-399/B, Nampally, Hyderabad 500001, India. E-mail: shekhar@cdfd.org.in; Tel: +91 40-24749401

† This article is part of a *Molecular BioSystems* themed issue on Computational and Systems Biology.

‡ Electronic supplementary information (ESI) available: Supplementary tables S1–S6. See DOI: 10.1039/b905264j

number of nodes have fewer connections. Consequently, only the small number of nodes that have many connections, referred to as 'hubs', control the overall robustness of the network. This property, commonly referred to as the 'degree centrality', thus suggests that nodes with a high degree tend to be more important for robustness of networks. Other centrality measures have also been developed to predict the nodes which control overall robustness.[11,15]

Importantly, centrality–lethality correlations in most of the previous studies on biological networks have been carried out on sampled sub-networks, rather than on a complete network.[13,15,16] This has been primarily due to the lack of information on the complete network, as these are labour intensive to obtain. An underlying assumption in these studies has been that the statistical properties of sampled sub-networks are identical to the corresponding larger networks. However, it has recently been argued that the information obtained from these sub-networks might provide inaccurate results since the number of nodes and interactions of the sub-network is incomplete.[17] The inferences based on the analyses of sub-networks are argued to be valid only if the probability distributions of a global network and its sub-network are the same. Such not being the case for scale free networks, the accuracy of the results using topological properties of the sub-networks might be indeterminate.[17,18]

There have been extensive studies on the network centrality–lethality correlations. In biological networks, the high centrality proteins are likely to be coded by the essential genes. Jeong *et al.*[15] have shown that high degree centrality nodes correlate well with gene essentiality in the yeast protein interaction network. Similarly, when the connectivity of the nodes is considered along with other parameters, such as gene expression variance and sequence information, the predictions of gene essentiality in yeast are observed to increase.[19–21] High closeness and betweenness centrality also have been shown to be the properties of key nodes in protein interaction networks.[11,13] Since centrality measures capture different aspects of gene essentiality, combining them should yield more accurate predictions than using only one of the measures. Our goal is to make use of centrality measures to develop the ability to predict essential nodes in a network. Furthermore, we propose the concept of conditional essentiality of nodes based on subnetworks of only those genes that are expressed under defined conditions. The prediction of essential nodes has immense potential in applications, such as identification of essential genes of an organism, and their potential use in drug target prioritization.

Results and discussion

A network can be represented as a graph G, consisting of a set of nodes V connected *via* E edges. The mathematical notation of a graph can be written in the form of an adjacency matrix $A = a_{i,j}$, which is a $n \times n$ symmetric matrix, whose element $a_{i,j}$ is 1 if there is a connecting edge between i and j, and 0 otherwise. The number of neighbours of a node i is referred to as the degree $k_i = \sum_{j \in G} a_{i,j}$ of i. The geodesic or the shortest path 'd' of a graph G is defined as the minimum distance required to traverse between any two nodes.[22]

The protein interaction dataset used in this study was obtained from our earlier predictions of genome-wide functional linkages in *E. coli*.[23] These linkages were analyzed for constructing various centrality measures such as degree centrality (DC), closeness centrality (CC) and betweenness centrality (BC). Degree centrality is viewed as the potential of a node for signalling activity based on its connectivity to other nodes in the network. Nodes with a high degree centrality are called 'Hubs', whereas nodes with a low degree centrality are referred to as 'Peripherals'. Closeness centrality measures the independency of a node compared to all other nodes in a network, *i.e.* a node with a high closeness centrality has the ability to contact any node of the network in the shortest possible path. Betweenness centrality is a measure of an individual node to fall on signalling paths between another two nodes that exhibit a potential for control of their signalling. Thus, the concepts of these centrality measures imply activity, independency and control of an individual node in a network, and thereby the three centrality measures are capable of assessing the relative importance of a node in a network. Furthermore, each node in a network can be assigned a weight based on its DC, CC or BC value.

In order to establish if the three centrality features correlate with the known gene essentiality, or non-essentiality conditions in *E. coli*, we considered the following two datasets. The essential genes were taken from the study of Baba *et al.*,[24] where the coding regions of *E. coli* have been systematically targeted for deletion. In this study, the genes for which mutants were not obtained, or were found to be non-viable for growth, were termed as essential. Thus, for 303 genes, the mutants were not viable and were classified as probable essential genes. Similarly, the study by Pósfai *et al.*[25] involved the deletion of genomic regions of *E. coli* K12 that had insertion sequences, transposable elements and the genes that are not conserved in other *E. coli* genomes. The set of 742 genes, the deletion of which caused no apparent growth defect, were considered as non-essential genes. We attempted to find correlations between these sets of essential and non-essential genes with the centrality parameters described earlier.

We chose to work with the predicted functional linkages reported by us earlier, rather than choosing a sub-network of experimentally determined interactions, since this collection is claimed to be a comprehensive set of genome-wide functional linkages.[23] The DC, CC and BC values of each of the nodes were calculated using the core network consisting of 78 048 edges and 3682 nodes. After arranging all the nodes of the network in a descending order based on the three centrality measures, we compared the lists with the known essential genes of *E. coli*. Interestingly, the majority of the essential genes were among those in top 20% of any of the three centrality lists, whereas only a few essential genes were found in the bottom 20% of the three lists. The distribution of essential genes for the three centrality parameters is shown in Fig. 1a, 2a and 3a, where it is apparent that all the three centrality measures are able to capture the essentiality feature. Similarly, the distribution of non-essential genes with respect to the three centrality measures revealed a complementary distribution as shown in Fig. 1b, 2b and 3b. It is therefore apparent that the three centrality measures are indeed able to

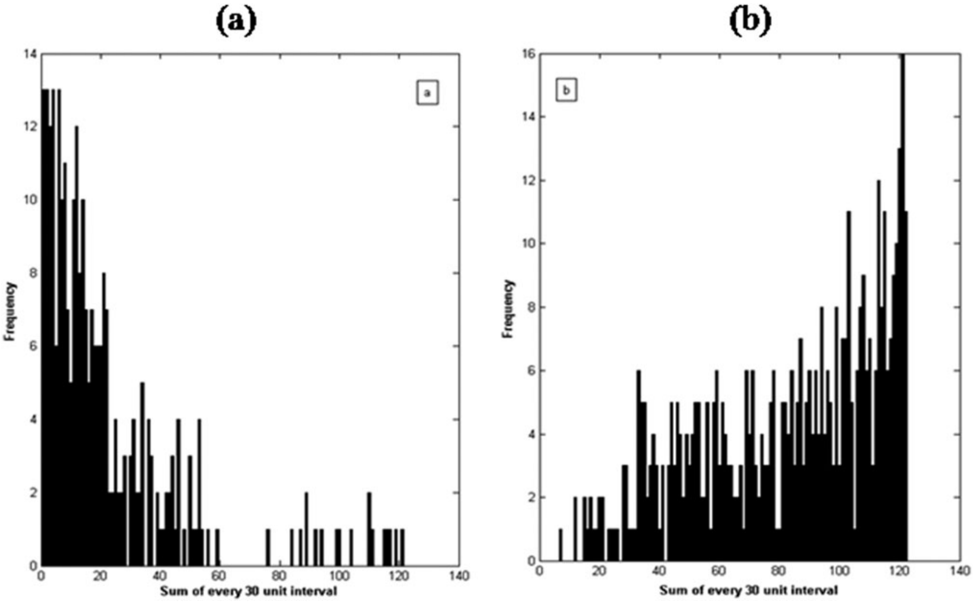

Fig. 1 The distribution of essential and non-essential genes based on degree centrality; (a) represents distribution of the essential genes and (b) represents that for non-essential genes.

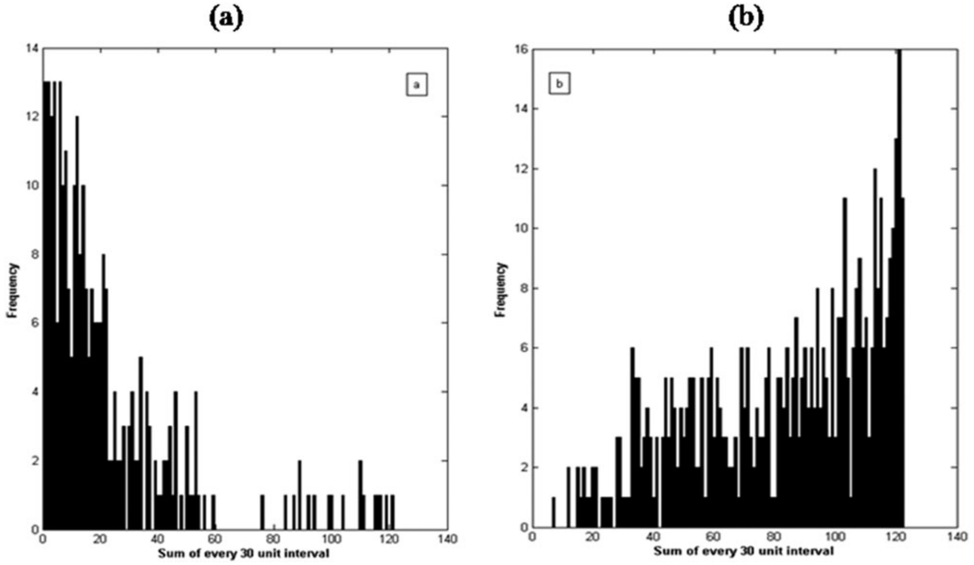

Fig. 2 The distribution of essential and non-essential genes based on closeness centrality; (a) represents distribution of the essential genes and (b) represents that for non-essential genes.

capture the essentiality or non-essentiality features, reinforcing a significant correlation between centrality and essentiality in a genome-wide functional linkage network.

Despite a strong correlation between centrality and essentiality, a significant number of essential genes also remain in the lower order of each of the three features. Notably, there are ten genes that have a low centrality value in at least one of the three measures, but are classified as essential. Moreover, genes *racR*, *chpS* and *yefM* have low centrality in all three of the measures. Similarly, a few non-essential genes are in the top order of the three features. For example, genes such as *tauB* and *yqiC* show high degree and closeness centrality, but have been classified as non-essential. Furthermore, there are about 30 non-essential genes that have high betweenness centrality. This suggests that betweenness may not be a good parameter for essential genes prediction when used in isolation.

The comparative analysis of the top 20% of nodes according to the three centrality measures revealed that there is little overlap among the three. There are only 48, 26 and 28 essential genes that are common to the top 20% of nodes between DC and CC, CC and BC, and DC and BC, respectively (Fig. 4). Furthermore, the centrality rankings of the genes are quite different for the three parameters (Table 1). The lack of overlap among the three might be due the inherent differences in the formalism of the centrality measures, which are based on connectivity, independency and control in the network.

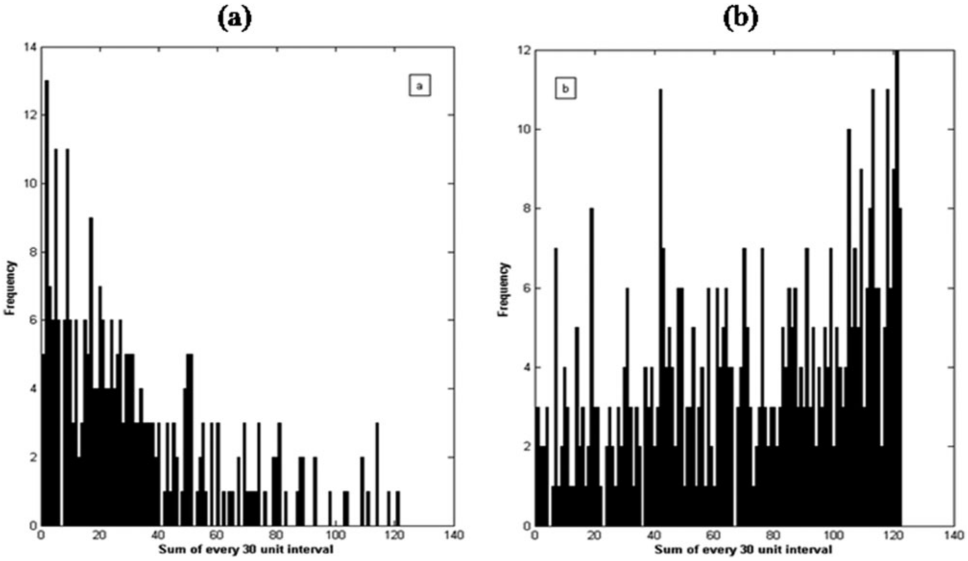

Fig. 3 The distribution of essential and non-essential genes based on betweenness centrality; (a) represents distribution of the essential genes and (b) represents that for non-essential genes.

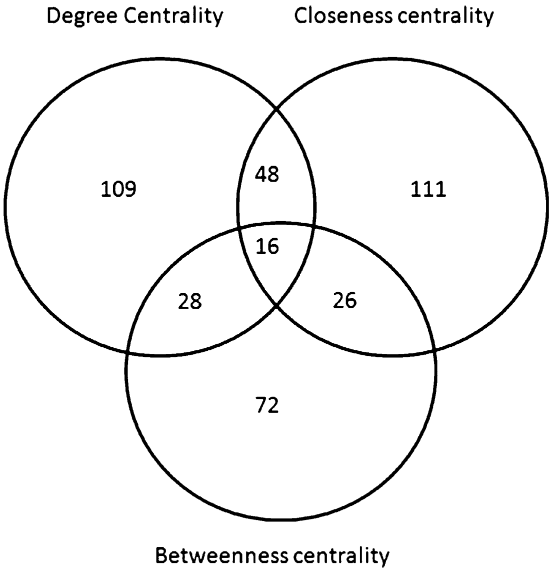

Fig. 4 The overlap among top the 20% of essential genes from degree, closeness and betweenness centrality measures. One can observe that the overlap between the three centrality measures is very small.

In order to examine the overall relatedness among the three centrality features, we calculated the pairwise Pearson correlation coefficients among the three pairs, and the partial correlation coefficients among the pairs, considering the third one to be constant. Table 2 (upper diagonal) lists the Pearson correlation coefficients for each of the three centrality pairs. It is clearly observed that DC and BC are strongly correlated, whereas DC and CC, as well as CC and BC are less strongly correlated. Table 2 (lower diagonal) also lists the partial correlation coefficients for the centrality pairs, considering the third one to be constant. It is once again clearly observed that DC and BC measures have a strong correlation when CC is constant. On the other hand, DC and CC are weakly correlated when BC is treated constant. Similarly, CC and BC are anti-correlated when BC is treated constant.

The low overlap among the high ranked genes of each of the three centrality measures (Fig. 2), and the lower partial coefficient values (Table 2) suggest that a combination of the three measures might yield useful predictions of node essentiality. We therefore attempted to combine the three features using a supervised machine learning algorithm and predict gene essentiality based on combinations of the three centrality measures. The machine learning method used was

Table 1 The ranks of some of the known essential genes of *E. coli* based on their centrality measures. The centrality ranking of the genes varies significantly across three different parameters

Essential gene	Degree centrality rank	Betweenness centrality rank	Closeness centrality rank
dnaN	8	164	16
pheT	27	255	62
rpoA	35	506	91
rplT	79	351	229
lolB	124	99	156
ssb	263	9	353
dapE	308	77	169
mviN	115	56	79
lepB	3	20	1
dnaX	13	131	9
kdtA	117	31	49

Table 2 The correlation coefficients of the centrality measures. The values in the upper diagonal denote pairwise Pearson correlation coefficients, whereas those in the lower diagonal denote partial correlation coefficients

	Degree	Closeness	Betweenness
Degree	—	0.79	0.96
Closeness	0.45	—	0.73
Betweenness	0.91	−0.29	—

Support Vector Machine (SVM), which is typically used for data classification by pattern recognition. It is usually trained on data for the selected features and generates the optimal hyper plane that separates the classes. We used SVM to classify the genes of *E. coli* as essential or non-essential. The features selected were the three network centrality measures, namely, degree centrality (DC), closeness centrality (CC) and betweenness centrality (BC). The training data consisted of centrality measures for the known essential and non-essential genes in *E. coli*.

The best SVM model obtained by testing on a blind dataset showed a sensitivity of 84% and a specificity of 96%. Using this model, when predictions were made using the proposed functional linkage network of Yellaboina *et al.*,[23] approximately one-third of the genes (1071 out of 3682) were predicted to be essential. The 1071-long list contained many paralogous genes, and since it is likely that paralogous genes offer redundancy of biological function, we randomly removed one of the genes from each of the paralogous pairs. The final prediction based on the proposed functional linkage data therefore yielded a set of 1011 genes, which might be considered important for their biological function in *E. coli*. The list of predicted essential genes is given in the ESI, Table S1.‡

Conditional essentiality of the genes

The high sensitivity and specificity of the predictions lead us to believe that the combined graph theoretical and machine learning approach can be useful in assessing the relative importance of different nodes in a graph. To apply such an approach meaningfully to a genome-wide functional linkage network, an important aspect to consider is that a realistic network consists of only a fraction of all the possible nodes. In other words, although the genome of an organism constitutes a complete set of genes that it is capable of encoding, only a fraction of all the genes are expressed under any defined condition. Considering this fact, it is likely that a gene considered as essential under one growth condition might not play as critical role in another condition. Indeed, some reports suggest that many non-essential genes become essential depending upon the environmental conditions.[26,27] Thus, the genome-wide prediction of essential genes, as carried out above, might manifest differently under different conditions depending upon which sets of genes are expressed.

The large-scale attempts to identify essential genes in several genomes have involved experimental gene deletions followed by tests of the viability of growth.[24,28–31] There have also been attempts to identify common features that the essential genes share. These attempts have been directed towards developing the ability to predict them in the genomes.[19–21] The gene essentiality studies using experimental methods are restricted to a particular growth condition or the strain type. Moreover, the experimental approaches are limited in understanding conditional essentiality as these are labor intensive and time consuming. Unfortunately, the available prediction models are also incapable of addressing the conditional gene essentiality. Our objective has therefore been to apply the models developed based on the three centrality measures to conditional gene expression and assess the essentiality of genes.

We applied the above mentioned approach to 61 conditional networks (Materials and methods) in order to predict conditional essentiality. The list of different growth conditions considered is given in the ESI, Table S2.‡ Each conditional network has ~2000 nodes. Furthermore, by the application of the model developed on DC, CC and BC, there are ~30% of genes predicted as essential in each of the conditional networks.

Analysis of the genes that are expressed under all conditions revealed that there are 119 genes which do not show expression in any of the conditions studied (ESI, Table S3‡). This set is enriched for the reported non-essential genes in *E. coli*. Some of the examples include *gntU*,[32] *araH*[33] and *glcF*.[34] On the other hand, some genes are expressed in all of the studied conditions and also are predicted to be essential in all of the conditions (ESI Table S4‡). Examples of such genes include replication proteins such as DnaA, DnaE, DnaG, DnaX, HolA and TopA, ribosomal proteins such as RplE, RplS, RplP, RpsA, RpsB and RpsC, translation proteins such as PheS, AlaS and Fmt, transcription proteins such as RpoD and RpoH, and metabolic proteins such as MurA, MurB, Can, AccD, IspG and MetK. Most of these proteins have been shown to be essential for the growth of *E. coli* in rich medium. Considering many of these genes occur as singletons in the genome and their protein products carry out core functional pathways in the cell, they are more likely to be essential, irrespective of the growth conditions. Some of the conserved hypothetical proteins such as YejM and YihA also appear to be essential in all the conditions studied, and are included in the KEIO list of essential genes.[24] Interestingly, there are around 70 genes that are expressed in all of the 61 growth conditions, but appear essential in none (ESI Table S5‡). Many of them are known to be non-essential for the cell survival. There are about 1192 genes that show conditional essentiality across 61 conditions, the names of which are listed in the ESI, Table S6.‡

In conclusion, the prediction model of gene essentiality obtained by training a SVM can be effectively applied to gene expression data in *E. coli*. This model is capable of distinguishing essentiality from non-essentiality from a combination of the centrality rules in graph theory. We therefore perceive such an approach to be useful in predicting and assessing applications, such as drug target identification and prioritization.

Materials and methods

Centrality measures

The protein interaction dataset used in this study was obtained from our earlier predictions of genome-wide functional linkages in *E. coli*.[23] The core of these functional linkages contains 3682 nodes (proteins) and 78 048 edges (interactions). These linkages were analysed for constructing various centrality measures as described below. The three principal centrality measures used for the analysis were degree centrality, closeness centrality and betweenness centrality. The mathematical form of the centrality measures are as follows (for N nodes). The normalized degree centrality of a node is defined as

$$C_i^D = \frac{k_i}{N-1}$$

where k_i is the degree of a node *i*.

The normalized closeness centrality of a node *i* is defined as

$$C_i^C = \frac{N-1}{\sum_{j\in G} d_{i,j}}$$

where $d_{i,j}$ is the geodesic between node *i* and node *j*.

The normalized betweenness centrality of a node *i* is defined as

$$C_i^B = \frac{\sum_{j<k\in G} n_{jk}(i)/n_{jk}}{(N-1)(N-2)}$$

where $n_{j,k}$ is the number of the shortest path between the nodes *j* and *k*, and $n_{j,k}$ (*i*) is the number of the shortest path between *j* and *k* that traverses through *i*.

Correlation and partial correlations

Having calculated the above centrality measures for all the nodes, the correlation between the different centrality measures was obtained using Pearson correlation coefficients. The Pearson correlation coefficient, R_{xy} is defined as,

$$R_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{N\sigma_X\sigma_Y}$$

where $\bar{X}$ is the mean and σ_X is the standard deviation of a particular centrality measure *X*.

The partial correlation coefficient determines the correlation between any two variables keeping the third variable as constant. Thus, the partial correlation between *X* and *Y* with the effects of *Z* constant is defined as,

$$R_{XY,Z} = \frac{R_{XY} - R_{YZ}R_{XZ}}{\sqrt{(1-R_{YZ}^2)(1-R_{XZ}^2)}}$$

where R_{XY}, R_{YZ} and R_{XZ} are the correlation coefficients between the any two of the variables.

SVM Classification

Support vector machines are supervised learning methods used for classification and regression. In our study we have used SVM for binary classification using the LIBSVM package.[35] The most commonly used Radial Basis Function kernel was adopted for training the input vectors. For binary class SVMs, the function to predict the output is

$$f(x) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b\right)$$

where x_i, $i = 1,\ldots,m$ are the selected training vectors, x is the input vector, $K(x_i,x)$ is the kernel function, which is a symmetric positive function, y_i the label for the output vector $(1,-1)$, α_i a weight for the support vector that is determined during the training process, and b is the bias of the hyper plane. The optimal kernel parameters, cost C and gamma G, were obtained through grid search and the dataset trained with five-fold cross validation. The model of the SVM was trained using cost parameter ranging from 2^{-6} to 2^{18} with unit step 2^4 and gamma parameter ranging from 2^{-18} to 2^{10} with unit step 2^4.

SVM was trained using the calculated centrality measures of known essential genes and non-essential genes.[24,25] Performance of predictions was assessed using 100 blind tests, where randomly selected 50 essential and non-essential genes were set aside, and not used in training the SVM. To quantify the performance, true positives, true negatives, false positives and false negatives were calculated, and the test results were evaluated through the sensitivity, specificity and accuracy:

$$\begin{aligned} \text{Sensitivity} &= \text{TP}/(\text{TP} + \text{FN}),\\ \text{Specificity} &= \text{TN}/(\text{TN} + \text{FP}),\\ \text{Accuracy} &= (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}), \end{aligned}$$

where TP, TN, FP and FN represent true positive, true negative, false positive and false negative, respectively.

Identification of Paralogs in *E. coli*

The protein sequences of *E. coli* were downloaded from NCBI and self BLAST was carried out using Blastall software. The paralogs were selected if the expectation value was $<10^{-10}$ and the sequence length alignment was more than 80% of at least one of the protein sequences.[20]

Prediction of conditional essentiality of genes

The conditional networks were constructed for 61 different growth conditions for *E. coli* as in ref. 36. The predicted protein functional linkages network for *E. coli* was considered as the parent network[23] and the gene expression datasets were downloaded from Stanford Microarray database.[37] The conditional essentiality of the genes across 61 conditions was predicted using the model file generated above.

Conclusion

We have illustrated a new approach based on combined graph theoretical and a machine learning algorithm for predicting essential proteins in *E. coli* networks. By training the Support Vector Machine on predicted genome-wide functional linkages, we hope that the problem of sampling in scale free networks has been avoided. The training has thus been carried out on a complete network rather than on a sub-network. We first show that each centrality measure predicts different proteins as being essential. The combined prediction using three different centrality measures yields predictions with high confidence. We propose that this method can be applicable to many different cases, including identification of key persons in a terrorist network, or prioritizing crucial drug targets.

Acknowledgements

We thank the Centre of Excellence in Medical Bioinformatics supported by the SUN Microsystems and CDFD for access to the facilities, and its staff for help at various times. The work was supported by Department of Biotechnology grant on Centre of Excellence in Microbial Biology. SRH is a UGC Senior Research Fellow.

References

1 R. Albert and A.-L. Barabási, *Rev. Mod. Phys.*, 2002, **74**, 47–97.
2 M. E. J. Newman, *SIAM Rev.*, 2003, **45**, 167–256.
3 P. Erdos and A. Renyi, *Publ. Math. Debrecen.*, 1959, **6**, 290–297.
4 D. J. Watts and S. H. Strogatz, *Nature*, 1998, **393**, 440–442.
5 A. L. Barabási and R. Albert, *Science*, 1999, **286**, 509–511.
6 M. E. J. Newman, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 8577–8582.
7 A. Bavelas, *Human Organization*, 1948, **7**, 16–30.
8 L. C. Freeman, *Social Networks*, 1978, **1**, 215–239.
9 V. Latora and M. Marchiori, *New J. Phys.*, 2007, **9**, 188–198.
10 S. B. Roberts, A. J. Mazurie and G. A. Buck, *Chem. Biodiversity*, 2007, **4**, 2618–2630.
11 M. P. Joy, A. Brock, D. E. Ingbor and S. Huang, *J. Biomed. Biotechnol.*, 2005, **2**, 96–103.
12 D. Gómez, E. González-Arangüena, C. Manuel, G. Owen, M. del Pozo and J. Tejada, *Math. Soc. Sci.*, 2003, **46**, 27–54.
13 M. W. Hahn and A. D. Kern, *Mol. Biol. Evol.*, 2005, **22**, 803–806.
14 V. Latora and M. Marchiori, *Chaos, Solitons & Fractals*, 2004, **20**, 69–75.
15 H. Jeong, S. P. Mason, A.-L. Barabási and Z. N. Oltvai, *Nature*, 2001, **411**, 41–42.
16 H. Yu, D. Greenbaum, H. X. Lu, X. Zhu and M. Gerstein, *Trends Genet.*, 2004, **20**, 227–231.
17 M. P. H. Stumpf, C. Wiuf and R. M. May, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 4221–4224.
18 A. Clauset and C. Moore, *Phys. Rev. Lett.*, 2005, **94**, 18701–18704.
19 H. Jeong, Z. N. Oltvai and A.-L. Barabasi, *ComPlexUs*, 2003, **1**, 19–28.
20 Y. Chen and D. Xu, *Bioinformatics*, 2004, **21**, 575–581.
21 S. Saha and S. Heber, *Genet. Mol. Res.*, 2006, **5**, 224–232.
22 E. W. Dijkstra, *Numerische Math.*, 1959, **1**, 269–271.
23 S. Yellaboina, K. Goyal and S. C. Mande, *Genome Res.*, 2007, **17**, 527–535.
24 T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner and H. Mori, *Mol. Syst. Biol.*, 2006, **2**, E1–E11.
25 G. Pósfai, G. Plunkett, 3rd, T. Fehér, D. Frisch, G. M. Keil, K. Umenhoffer, V. Kolisnychenko, B. Stahl, S. S. Sharma, M. de Arruda, V. Burland, S. W. Harcum and F. R. Blattner, *Science*, 2006, **312**, 1044–1046.
26 A. R. Joyce, J. L. Reed, A. White, R. Edwards, A. Osterman, T. Baba, H. Mori, S. A. Lesely, B. Ø. Palsson and S. Agarwalla, *J. Bacteriol.*, 2006, **188**, 8259–8271.
27 B. Papp, C. Pál and L. D. Hurst, *Nature*, 2004, **429**, 661–664.
28 S. Y. Gerdes, M. D. Scholle, J. W. Campbell, G. Balázsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. D'Souza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Y. Fonstein, R. Overbeek, A. L. Barabási, Z. N. Oltvai and A. L. Osterman, *J. Bacteriol.*, 2003, **185**, 5673–5684.
29 N. R. Salama, B. Shepherd and S. Falkow, *J. Bacteriol.*, 2004, **186**, 7926–7935.
30 K. Kobayashi, S. D. Ehrlich, A. Albertini, G. Amati, K. K. Andersen, M. Arnaud, K. Asai, S. Ashikaga, S. Aymerich, P. Bessieres, F. Boland, S. C. Brignell, S. Bron, K. Bunai, J. Chapuis, L. C. Christiansen, A. Danchin, M. Débarbouille, E. Dervyn, E. Deuerling, K. Devine, S. K. Devine, O. Dreesen, J. Errington, S. Fillinger, S. J. Foster, Y. Fujita, A. Galizzi, R. Gardan, C. Eschevins, T. Fukushima, K. Haga, C. R. Harwood, M. Hecker, D. Hosoya, M. F. Hullo, H. Kakeshita, D. Karamata, Y. Kasahara, F. Kawamura, K. Koga, P. Koski, R. Kuwana, D. Imamura, M. Ishimaru, S. Ishikawa, I. Ishio, D. Le Coq, A. Masson, C. Mauël, R. Meima, R. P. Mellado, A. Moir, S. Moriya, E. Nagakawa, H. Nanamiya, S. Nakai, P. Nygaard, M. Ogura, T. Ohanan, M. O'Reilly, M. O'Rourke, Z. Pragai, H. M. Pooley, G. Rapoport, J. P. Rawlins, L. A. Rivas, C. Rivolta, A. Sadaie, Y. Sadaie, M. Sarvas, T. Sato, H. H. Saxild, E. Scanlan, W. Schumann, J. F. Seegers, J. Sekiguchi, A. Sekowska, S. J. Séror, M. Simon, P. Stragier, R. Studer, H. Takamatsu, T. Tanaka, M. Takeuchi, H. B. Thomaides, V. Vagner, J. M. van Dijl, K. Watabe, A. Wipat, H. Yamamoto, M. Yamamoto, Y. Yamamoto, K. Yamane, K. Yata, K. Yoshida, H. Yoshikawa, U. Zuber and N. Ogasawara, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 4678–4683.
31 G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K. D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Güldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kötter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C. Y. Wang, T. R. Ward, J. Wilhelmy, E. A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis and M. Johnston, *Nature*, 2002, **418**, 387–391.
32 H. Izu, O. Adachi and M. Yamada, *J. Mol. Biol.*, 1997, **267**, 778–793.
33 B. F. Horazdovsky and R. W. Hogg, *J. Bacteriol.*, 1989, **171**, 3053–3059.
34 M. T. Pellicer, J. Badia, J. Aguilar and L. Baldoma, *J. Bacteriol.*, 1996, **178**, 2051–2059.
35 C. C. Chang and C. J. Lin, 2001, http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
36 S. R. Hegde, P. Manimaran and S. C. Mande, *PLoS Comput. Biol.*, 2008, **4**(11), e1000237.
37 J. Demeter, C. Beauheim, J. Gollub, T. Hernandez-Boussard, H. Jin, D. Maier, J. C. Matese, M. Nitzberg, F. Wymore, Z. K. Zachariah, P. O. Brown, G. Sherlock and C. A. Ball, *Nucleic Acids Res.*, 2007, **35**, D766–D770.