GENE LETHALITY DETECTION ACROSS BIOLOGICAL NETWORK DOMAINS: HUBS VERSUS STOCHASTIC GLOBAL TOPOLOGICAL ANALYSIS

Gil Alterovitz1, Vinayak Muralidhar2, and Marco F. Ramoni2,3

1 Health Science and Technology/Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, 02139, USA. 2 Harvard Medical School- Partners Center for Genetics and Genomics, Boston, 02115, USA. 3 Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, Boston, 02115, USA.

ABSTRACT

In this paper, we investigate the properties of lethal genes in *E. coli*, our model organism. Topological analysis of networks of functional interactions among genes has shown that lethal genes share common local connectivity properties. In this paper, we analyze cellular networks across three domains. We show that a stochastic global topological analysis, via random walks, is more effective at predicting gene lethality than simply looking at local topology using the standard hub-based method.

We also introduce the possibility of using metabolic pathways to understand lethal genes, as regulating these pathways is among one of the most important functions of the gene-encoded proteins. Additionally, we analyze lethal genes in terms of the Gene Ontology (GO) and find that the graph forms two highly connected clusters that are each GO enriched for specific terms. We also find that lethal metabolic regulators are extremely enriched. Finally, we provide applications of the work and avenues for future research.

1. INTRODUCTION

Genes code for the production of proteins that play pivotal roles all organisms. Until the completion of the Human Genome Project, the higher order and structure found in humans compared to other organisms were attributed to the hypothesis that humans had more genes [1]. However, by the end of the Human Genome Project, it was found that humans only had between 20,000 and 25,000 genes, only slightly greater than the 19,000 found in the worm *Caenorhabditis elegans* [2]. For this reason, scientists turned to proteomics, which involves the study of an organisms entire set of proteins and the interactions among these proteins. In this paper, we seek to use networks involving proteins to computationally determine lethal genes in an organism. By lethal gene, we refer to a gene whose removal causes death in the cell.

2. METHODS

We chose *E. coli* as our model organism. The Profiling of *E.coli* Chromosome (PEC) database was used to form the *E. coli* interactome, which describes all protein-protein interactions in *E. coli* [1]. The Ecocyc database is used to form the genetic regulation network; it describes which proteins regulate the production of other proteins [2]. Further, we use the Ecocyc database to form the metabolic pathways network. This network describes which metabolites are changed into other metabolites, and also delineates which genes control parts of these pathways.

We analyze the method of random walks in this paper and compare it to the hubs method of Barabasi [3] and to simple random selection. In the random walk method, we first randomly select a starting node. Then, we continue to walk on it for as many steps as there are nodes in the graph *n*. We continue this 100 times, jumping to a new random node every *n* steps. We produce a list of 10% of the genes in the total network that are walked on the most, and predict those to be lethal. Thus, our approach looks at global, not just local topologies. The hubs method, which we refer to as the local topology-based approach herein, in contrast, only looks at the nodes/edges directly surrounding the current node; it counts the number of neighbors a given node has. In this paper, we use the top 10% of neighbored genes as our predicted lethal genes for the local topology-based approach. Finally, random selection provides a basis for comparison. We select a random set of 10% of the total genes and predict those as lethal. This is repeated 100,000 times and an average and standard deviation computed.

We measure the accuracy of our results by the list of Gerdes *et al.* [4]. Gerdes *et al.* provide a list of lethal and nonlethal genes in *E. coli*. Unfortunately, the data are not complete, and there are many genes whose lethality status is either unknown or inconclusive. We only include genes in our predicted lethality lists that have known lethality status.

Further, we consider the networks in terms of the GO. GO [5] is a controlled vocabulary that contains terms to describe genes. A group of genes is GO enriched if they share characteristics more closely than a randomly selected
group of genes. We use Fisher’s exact test and Bonferroni multiple test correction to calculate enrichment.

3. RESULTS

Fig. 1 displays the results of the methods in terms of Signal-to-Noise Ratio (SNR). Here, prediction results are normalized to of one (i.e. random prediction) for each network type. The random walk approach performs better than or at same level as the local topology-based method across all networks. In Fig. 1, the top of the error bars represent 95th-percentile of random selection under normal distribution assumption. Here, stars indicate that the corresponding method’s performance is significantly (p < 0.05) better than random.

![Comparison of local topology-based hubs versus random walk approach for predicting gene lethality.](image)

Additionally, we find that the lethal genes in the interactome follow a similar pattern to the lethal genes in the regulatory network [6]. We see two highly connected clusters, while almost all genes not part of either of these two clusters are completely unconnected. Each of these clusters is highly GO enriched. The first, smaller cluster of 14 genes is enriched for fatty acid biosynthesis (1.1x10⁻⁵). The second cluster of 44 genes is enriched for controlling the cell cycle (p < 3.6x10⁻¹¹), cell metabolism (p < 3.1x10⁻⁴), and protein biosynthesis (p < 1.5x10⁻⁵).

Interestingly, we found that the 23 genes which regulate metabolic processes and are lethal were highly enriched for oxidoreductase activity, acting on the aldehyde or o xo group of donors, with disulfide as an acceptor (p < 10⁻³).

4. DISCUSSION

We find that the random walk method consistently does better than or just as good the local topology-based approach across different biological networks. This is undoubtedly due to the fact that contrary to the local topology-based methods such as hubs, the random walk method looks at more than simply the direct neighborhood of a node. Instead, in order for a node to be predicted as lethal, it needs to be part of a highly connected group of genes so that there is a high probability of a gene being walked on multiple times. The success of this method over the local approach implies that we must take a more global look at cellular networks in order to understand lethality and potentially other network-based phenotypes. This paper also contributes to connecting the metabolic pathways in a cell and the genetic regulators that control these pathways.

In order to strengthen the results, it is important to augment the databases used. Currently, less than 1000 of the 4000 genes in E. coli are included in any of the networks. Clearly then, with the incorporation of more genes, there can be additional results.

In conclusion, this paper contributes an analysis of the random walk method across multiple biological network domains. We find that our method performs better than the local topology-based hubs approach, currently status quo. The ability of the random walk method to computationally determine lethal genes (i.e., without any actual clinical experimentation), indicates that we must not only look past direct node neighbors in order to better understand cell networks, but that cellular networks may also share underlying properties that can only be fully explored via a global perspective.

4. REFERENCES