FUNCTIONAL CENTRALITY: DETECTING LETHALITY OF PROTEINS IN PROTEIN INTERACTION NETWORKS

Kar Leong Tew[1]
kltew@i2r.a-star.edu.sg

Xiao-Li Li[1]
xlli@i2r.a-star.edu.sg

Soon-Heng Tan[1,2,3]
chris.tan@utoronto.ca

[1] *Knowledge Discovery Department, Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613*
[2] *Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada**
[3] *Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, Canada**
* *Present Affiliation*

Abstract

Identifying lethal proteins is important for understanding the intricate mechanism governing life. Researchers have shown that the lethality of a protein can be computed based on its topological position in the protein-protein interaction (PPI) network. Performance of current approaches has been less than satisfactory as the lethality of a protein is a functional characteristic that cannot be determined solely by network topology. Furthermore, a significant number of lethal proteins have low connectivity in the interaction networks but are overlooked by most current methods.

Our work reveals that a protein's lethality correlates more strongly with its "functional centrality" than pure topological centrality. We define functional centrality as the topological centrality within a subnetwork of proteins with similar functions. Evaluation experiments on four *Saccharomyces cerevisiae* PPI datasets showed that NFC performed significantly better than all the other existing computational techniques. Our method was able to detect low connectivity lethal proteins that were previously undetected by conventional methods. The results and an online version of NFC is available at http://lethalproteins.i2r.a-star.edu.sg

Keywords: Lethal proteins; Functional centrality; Protein similarity; Protein-protein interaction

1. Introduction

A lethal (or essential) protein is one that renders the cell unviable on its removal. From a theoretical point of study, lethal proteins play an intricate role for cell survival and development. Studying of lethal proteins will open opportunities to understand other species and identification of potential drug targets [1]. While lethal proteins can be detected from gene knockout experiments, large-scale systematic

detections can still be time-consuming and cost-prohibitive. As Jeong [2] noted, lethality profiles of a substantial number of genes are still unknown.

Alternative approaches to detect potential lethal proteins is thus required. One common hypothesis is that lethal proteins are strategically located within the protein-protein interaction (PPI) network such that their absence would create an adverse disruption to the topological stability of the network, thereby leading to biological lethality. Jeong [3] were one of the first to establish that there indeed exists a correlation between lethal proteins and their topological feature (connectivity) in the underlying PPI network. This led to a series of similar works unveiling new topological characteristics related to a protein's lethality (see Section 2).

However, the performance of many topological-based approaches had been less than satisfactory as the biological lethality of a protein is a functional characteristic that is unlikely to be adequately determined solely by network topology. Moreover, many current approaches were based on the assumption that a protein's lethality is correlated with high connectivity in the PPI network. This may not be always true as it is possible for a low-connectivity protein to be lethal. We found a substantial number of known lethal proteins with low connectivity (number of interaction partners ≤ 5) in the yeast PPI network (see Table 1).

In this work, we combine the topological-based concept for protein lethality with the notion of functional modules [4, 5], which are groups of interconnected proteins performing discrete functions in the PPI network. Multimeric protein complexes (such as the ribosome that synthesize polypeptides from amino acids) and biological pathways are instances of functional module. We reasoned that lethal proteins are the key players or coordinators within functional modules and their removal will maximally disrupt the operations of the modules which impact cell fitness. We hypothesized that these key proteins should also be centrally positioned within functional modules to carry out their roles effectively and their removal will cripple the modules more easily than the removal of proteins lying at the peripheral.

Thus in this paper, we introduce a novel *neighborhood functional centrality* (NFC) measure to quantify the extent in which a protein is surrounded by functionally consistent neighboring proteins in the PPI network. We also devised a Neighborhood Functional Centrality (NFC) algorithm to mine lethal proteins in PPI networks based on this concept. Evaluation on four *Saccharomyces cerevisiae* PPI datasets showed that NFC performed significantly better than all other existing computational techniques. Given that many lethal proteins can be of low connectivity, we also verified that our NFC method can detect low-connectivity lethal proteins undetected by conventional methods.

2. Related Works

Jeong [3] first reported that the lethality of a protein is positively correlated to its connectivity (or degree) in the protein interaction network—the number of interacting partners a protein has. This has led to numerous subsequent works that attempted to infer a protein's lethality in baker's yeast using various other net-

work topological characteristics such as clustering coefficient [6], betweenness [7], damage [8], and subgraph centrality [9]. The clustering coefficient quantifies the probability of two interacting proteins are also interacting with a similar third protein. The betweenness score quantifies a protein's topological centrality based on number of shortest paths that pass through it in the underlying PPI network. The damage score measures the disintegration of the underlying PPI network resulting from the removal of a protein. The subgraph centrality score quantifies the number of subgraphs a node participate in with emphasis on shorter closed paths.

All the above methods used solely topological measures that are directly or indirectly dependent on the high connectivity of proteins within the PPI network. As such, they will not work very well when the underlying PPI network is a sparse network. In the four PPI datasets that we have used for our evaluation experiments, we found an average of 67.0% of lethal proteins exhibited low connectivity (number of interaction partners ≤ 5) in the underlying PPI networks (see the bracketed figures in Table 1). Furthermore, a substantial amount (54.7%) of the high-connectivity proteins (number of interaction partners ≥ 6) were not known to be lethal, suggesting that the biological lethality of a protein cannot be adequately determined solely by network topology. In this paper, we propose a new method that incorporates functional information with topological information to better detect lethal proteins, including those with low connectivity in PPI networks.

3. Method

We model the protein interaction data as a large undirected graph $G_{PPI} = (V_{PPI}, E_{PPI})$, where V_{PPI} represents the set of interacting proteins and E_{PPI} denotes all detected pairwise interactions between two proteins from V_{PPI}. Our NFC algorithm consists of two steps. First, for each protein in the interaction graph, we construct a local neighborhood graph to compute a *nfc* score (Section 3.1). Then, we assess the significance of $nfc(u)$ by computing its corresponding Z_{nfc} (Section 3.2).

3.1. *Computing the Neighborhood Functional Centrality*

To compute the neighborhood functional centrality score *nfc* for each protein in the interactome, we define the neighborhood graph for each vertex u in G_{PPI} as follows:

Definition 1. For each vertex $u \in V_{PPI}$, its neighborhood graph is defined as $G_u = (V_u, E_u)$, where:

$$V_u = \{v \mid v \in V_{PPI} \wedge dist(u, v) \leq \theta\},$$

$$E_u = \{(v_j, v_k) \mid (v_j, v_k) \in E_{PPI} \wedge v_j, v_k \in V_u\}, \text{ and}$$

$dist(u, v)$ is a function that returns the shortest distance between u and v.

The neighborhood graph G_u of a vertex u is the subgraph in G_{PPI} induced by the vertices that are within a radius of θ from u. θ is a user-defined variable to control

the radius (or size) of the neighborhood graph of vertex u and we will investigate its effect on the prediction results later (Section 5.4).

Next, we evaluate whether a protein is functionally central in its neighborhood graph. This involves measuring the functional similarities among the proteins in the neighborhood graph. This is achieved by incorporating functional information associated with each protein into our analysis.

Biological functions are typically organized in a hierarchical structure—generic biological functions (such as *transcription*) can be progressively broken down into more specific functions (such as *transcription termination*, and *transcription from RNA polymerase II promoter*). Each protein in an interactome is annotated (if at all[a]) with functions at various levels of specificity depending on the state of functional knowledge on the individual proteins. Currently, the most commonly used structure for functional annotation is the Gene Ontology—GO [10].

To compute the functional centrality of the proteins, we take into consideration that the proteins' functional annotations are in ancestor/descendent relationships. As such, we adopted the Relative Specificity Similarity (RSS) method that Wu [11] have developed which is a quantitative measure of the similarity between two GO functions (Definition 2) taking into account the hierarchical structure of GO:

Definition 2. Relative Specificity Similarity (RSS)

$$RSS(term_i, term_j) = \frac{maxDepth^{GO}}{maxDepth^{GO} + \gamma} \cdot \frac{\alpha}{\alpha + \beta}$$

where $maxDepth^{GO}$ is the maximum depth of the GO, α measures the maximum number of common ancestor terms shared between $term_i$ and $term_j$ in a single path, β is the value of the longer distance between $term_i$ and $term_j$ to their closest leaf nodes, and γ measures the shortest distance between $term_i$ and $term_j$. Refer to Wu [11] for details.

Definition 2 defines the functional similarity between two individual functions. However, a protein could be involved in different biological processes and associated with multiple GO annotations. Suppose F_u and F_v are the function annotations of proteins u and v respectively, we define the functional similarity between the proteins u and v as follows:

Definition 3. The protein functional similarity between two proteins u and v is defined as

$$protein\_funsim(u, v) = \frac{\sum_{i=1}^{|F_u|} \sum_{j=1}^{|F_v|} RSS(F_{(u,i)}, F_{(v,j)})}{(|F_u| * |F_v|) * dist(u, v)}$$

where $F_{(u,i)}$ and $F_{(v,j)}$ denote protein u's i-th and protein v's j-th's functions respectively, and $|F_u|$ denotes the number of functions protein u is annotated with.

[a]We will discuss strategies to handle proteins without functional annotations and their prediction in Section 5.3.

Definition 3 quantifies the extend of functional similarity between two proteins which may have multiple functions. The denominator $dist(u, v)$ is included here to give higher weightage for protein pairs that are closer together in the underlying interaction graph—this takes into account the implicit functional similarity between the two proteins based on their distance in the interactome.

We are now ready to define the *neighborhood functional centrality* score for each protein based on its functional similarity with proteins in its neighborhood graph:

Definition 4. The neighborhood functional centrality $nfc(u)$ of a protein u is defined as

$$nfc(u) = \sum_{v \in V_u, v \neq u} protein\_funsim(u, v)$$

Definition 4 quantitates the degree of functional consistency between protein u and all the other proteins in its neighborhood graph $G_u = (V_u, E_u)$. The value $nfc(u)$ indicates the functional centrality of protein u in G_u.

3.2. *Computing the Corresponding Z-scores*

Depending on the underlying functional distribution of the proteins in the interactome, it is possible that protein u is more likely to be assigned a higher $nfc(u)$ when located in a larger neighborhood graph G_u, or vice versa due to the summation used in Definition 4. In other words, given an interactome, the statistical distributions of $nfc(u)$ in smaller neighborhoods may be different from those in bigger neighborhoods (i.e. different means and/or different standard deviations). The significance of a particular $nfc(u)$ value is therefore dependent on the underlying distribution with respect to the size of the local neighborhood chosen for u.

In this work, we assess the significance of each protein's $nfc(u)$ value by computing its Z-Score (or "standard scores") Z_{nfc} as follows:

Definition 5. $Z_{nfc}(u, s)$ is defined as

$$Z_{nfc}(u, s) = \frac{nfc(u) - \mu_s}{\sigma_s}$$

where μ_s and σ_s are the mean and standard deviation of the distribution of $nfc(u)$ values computed from neighborhood graphs of size s.

Definition 5 requires computation of the distributions of neighborhood functional centrality values for differently-sized neighborhood regions. In fact, we only need to compute distributions for neighborhood sizes actually used in our $nfc(u)$ computation which we stored in set US. We estimate the distributions of neighborhood functional centrality values for each neighborhood size in US by randomly fetching same-sized neighborhood graphs for each vertex (if possible[b]) to determine

[b]It is possible that we are unable to fetch from u, a neighborhood of an intended size if the protein is in a small isolated partition. However, θ can be set small enough such that it is possible to find some neighborhoods of the intended size with some other vertices in the PPI network.

the corresponding neighborhood functional centrality values.

4. Experimental Data

For evaluation, we performed comparative experiments to show that our neighborhood functional centrality (NFC) approach performs better than other existing computational techniques. We used PPI datasets for *Saccharomyces cerevisiae* as it is currently the only organism with fairly complete knockout analysis (which forms our core lethal protein list).

4.1. *PPI Datasets*

We used four publicly available *Saccharomyces cerevisiae* protein interaction datasets for our evaluation experiments: *FYI* [12], *Nature* [2], *Bu* [13] and *DIPS* [14]. Each dataset was named after the source from which we have acquired them—details about each dataset are shown in Table 1. We have elected to use four different datasets so as (a) to facilitate direct comparisons with previous work and (b) to verify the performance against datasets of varying quality. The first dataset *FYI* is a high-quality (reliable) but sparse yeast interaction dataset with minimal false positives [12]. Another sparse network is *Nature*—included as it was employed by Jeong [3] whom first used the connectivity measure (which we will be comparing against) to detect the lethal proteins. The third dataset *Bu* is a relatively dense network with 3 times as many interactions as the previous two datasets. It was compiled by Bu [13] for function prediction, and subsequently used by Estrada [9] whom introduced the Subgraph Centrality (SC) measure which we will also be comparing against. The fourth dataset *DIPS* was obtained from the Database of Interaction Proteins (Nov 2005), giving rise to another dense network with interactions derived from various biological experiments. We pre-processed all four datasets by removing self-interacting interactions and isolated protein pairs from the networks.

Table 1. Details of the four *Saccharomyces cerevisiae* protein interaction datasets used in our evaluation experiments.

	FYI	**Nature**	**Bu**	**DIPS**
# Proteins	1210 (*958*)	1638 (*1490*)	2224 (*1531*)	2406 (*1773*)
# Lethal	464 (*333*)	369 (*312*)	670 (*349*)	695 (*414*)
# Unknown (No Function)	12 (*10*)	94 (*84*)	18 (*17*)	23 (*23*)
# Interactions	2400	2201	6609	5665

Note: Italicized numbers in brackets represents proteins with connectivity ≤ 5.

Since our NFC method incorporates the functional information of the proteins for evaluation, we used function annotations classified as biological process by GO [10] (27-Oct-2006). Functional annotation has not yet reach the stage where we can expect all the proteins to be annotated (see "# Unknown" in Table 1) and we address this and the function prediction mechanisms in Section 5.3.

4.2. *Reference List and Evaluation Metric*

For evaluation, we used a benchmark lethal protein list (the Core list) consisting of 1106 known lethal proteins for *Saccharomyces cerevisiae* determined by PCR-based gene deletion strategy [15]. This set of lethal proteins was derived experimentally using PCR-based gene deletion strategy [16, 17]. We plot the corresponding ROC (Receiver Operating Characteristic) curves to compare the performance of the various prediction methods. Quantification of the significance of each prediction technique's ROC curve is done using the AUC (Area Under the Curve) values.

5. Experimental Results

In this section, we first compare our NFC method against other existing methods for predicting lethal proteins from PPI datasets to see whether NFC can perform better than the current methods (Section 5.1). We also check on the performance of our NFC to see if it can better detect low connectivity lethal proteins (Section 5.2).

We next investigate the performance of NFC in the absence of functional information and how function prediction mechanism can help in addressing this issue (Section 5.3). Finally, we investigate how the performance of NFC may be affected by different values of θ which controls the neighborhood radius (Section 5.4).

5.1. *Performance Comparisons*

We compare the performance of NFC against three other existing methods, namely, connectivity [3], subgraph centrality (SC) [9] and cluster coefficient (CC) [6]. For a fair evaluation, we use the same four protein interaction datasets, core lethal protein list, and function annotation for all the methods. We have omitted here the *damage score* method proposed by Schmith [8] and the *betweenness score* method by Joy [7]. This is because the damage score was already known to have a lower correlation to lethality as opposed to connectivity in PPI datasets [8], while betweenness have been outperformed by SC [9].

Table 2. AUC comparisons of NFC, Connectivity, SC, and CC.

	FYI	**Nature**	**Bu**	**DIPS**
NFC	67.8 (*67.7*)	71.2 (*73.2*)	74.9 (*72.0*)	75.3 (*74.3*)
Connectivity	60.8 (*58.1*)	61.0 (*58.6*)	66.0 (*58.3*)	65.8 (*60.9*)
SC	57.1 (*54.2*)	56.8 (*53.4*)	65.4 (*56.7*)	63.9 (*58.7*)
CC	55.2 (*56.0*)	59.0 (*59.1*)	58.8 (*59.1*)	56.9 (*58.9*)

Note: Italicized numbers in brackets represents AUC values for detecting proteins with connectivity ≤ 5.

In Figure 1, we show the ROC curves of the four prediction methods on our experimental datasets. The AUC values for all four datasets are shown in Table 2 which clearly depicts the generality of NFC when used in datasets of varying size and quality. The results also shows that NFC can better detect lethal proteins from PPI datasets than other existing techniques due to its larger AUC values.

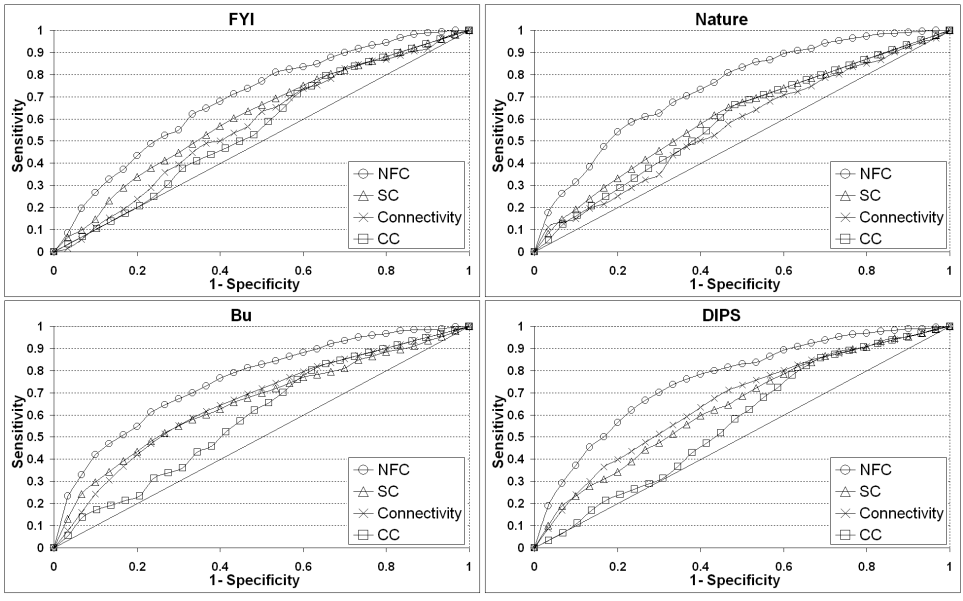

Fig. 1. ROC charts of NFC, Connectivity, SC, and CC for each evaluation dataset.

5.2. *Low Connectivity Proteins*

As from Table 1, a significantly large proportion (78.2% on average) of proteins in the datasets are of low-connectivity (i.e. number of interaction partners ≤ 5), even in dense PPI networks such as Bu and DIPS. Furthermore, a high average of 67.0% of the lethal proteins in our core reference list also has low connectivity in the underlying PPI networks. This means that the connectivity-based detection methods would have missed out a significant number of lethal proteins if we rely solely on detecting high-connectivity proteins. The bracketed numbers in Table 2 confirms that NFC can detect low connectivity lethal proteins much better than the other existing methods in all four datasets.

5.3. *Protein Function Annotation: Absence and Prediction*

The incorporation of biological knowledge in addition to topological information have vastly improved lethal proteins detection. However, this implies that our NFC method is dependent on the amount of biological knowledge available, and its performance is expected to decrease with a higher number of unknown proteins (i.e. proteins without known functions).

We tested this on the datasets by generating the situation where 50% of the proteins have an unknown function through a random selection process of marking a protein as having unknown function[c]. As expected, a decline was observed in the

[c]Here, we set the upper limit at 50% as a statistical study shows the largest percentage of unknown proteins on other species was 46.0% (*Caenorhabditis elegans*).

AUC values. We then follow this up with the utilization of function prediction mechanism. For simplicity, we chose the *Majority* measure proposed by Schwikowski [18]. By using the *Majority* method in the same situation (50% unknown), the improvements in AUC values for each dataset are from 54.8% to 62.5% for FYI, 57.4% to 62.2% for Nature, 60.0% to 69.1% for Bu, and 58.8% to 69.7% for DIPS (Table 3). Even with a basic method, NFC is still able to obtain AUC values better than existing methods. By coupling with more sophisticated protein function prediction methods [19–21], we certainly expect NFC performance to be more robust than illustrated.

Table 3. AUC values with different percentages of unknown proteins.

	Normal	**10%**	**20%**	**30%**	**40%**	**50%**
FYI	67.8	64.5 (*67.8*)	60.7 (*66.8*)	58.7 (*65.4*)	56.6 (*64.2*)	54.8 (*62.5*)
Nature	71.2	68.4 (*70.6*)	64.6 (*68.8*)	61.7 (*66.9*)	59.9 (*64.8*)	57.4 (*62.2*)
Bu	74.9	70.9 (*73.8*)	67.1 (*72.4*)	63.4 (*71.5*)	61.0 (*70.4*)	60.0 (*69.1*)
DIPS	75.3	70.9 (*74.6*)	66.9 (*73.4*)	64.3 (*72.2*)	60.9 (*70.9*)	58.8 (*69.7*)

Note: Italicized numbers in brackets represents AUC values with the *Majority* measure used.

5.4. *Varying the Neighborhood Radius Threshold θ*

Recall that our NFC method employed a user-defined threshold θ (Section 3.1) that controls the radius of the neighborhood graphs to compute the functional centrality values[d]. It is therefore possible that the performance of NFC may be affected by choice of θ used. Further evaluation experiments where we computed the various AUC values for each dataset when $\theta = 1$ to $\theta = 5$ has the mean deviation of 0.3% (FYI), 1.2% (Nature), 1.1% (Bu), and 1.3% (DIPS) (Table 4). These values are a clear indication that NFC's performance is not affected by θ. During our investigation, we also found that by using the Z-score instead of the raw $nfc(u)$ values, our NFC method has effectively adjusted for the effects of different neighborhood sizes and improved the accuracy of its predicted lethal proteins. Compared to using only the raw $nfc(u)$ values in the computation, Z_{nfc} improved the AUC values by 6.2%, 7.1%, 5.8%, and 7.3% for FYI, Nature, Bu, and DIPS respectively.

Table 4. AUC values for different θ.

	$\theta = 1$	$\theta = 2$	$\theta = 3$	$\theta = 4$	$\theta = 5$	**Mean Deviation**
FYI	67.0%	67.8%	67.1%	66.6%	66.6%	0.3%
Nature	67.4%	71.2%	70.8%	69.4%	68.5%	1.2%
Bu	72.2%	74.9%	73.5%	72.0%	71.7%	1.1%
DIPS	73.2%	75.3%	73.6%	71.9%	70.9%	1.3%

[d]Results presented above are obtained with $\theta = 2$.

6. Discussions and Conclusions

The detection of lethal proteins is useful for various aspect of biological study. To complement the costly experimental approaches such as PCR-based gene deletion strategy [15], and to exploit the large datasets of protein-protein interactions that have become available, researchers have proposed numerous computational methods using topological properties associated with high connectivity to infer protein lethality. However, we have shown in this paper that the lethality of a protein is a functional characteristic that cannot be determined solely by network topology. Furthermore, a significant number of lethal proteins have been found to have low connectivity (less than 5 interaction partners) in the interaction networks.

A protein's lethality should also be determined using additional non-topological information such as its functional grouping within the cell. We reasoned proteins that are the key players or coordinators within functional modules are likely to be lethal as their removal will drastically disrupt the effective operations of the modules. In this paper, we proposed a novel *neighborhood functional centrality* (NFC) approach that incorporated the conventional topological concept for protein lethality with the notion of functional modules [4, 5] to better detect protein lethality.

NFC was shown to discover both lethal proteins with high connectivity as well as those with low connectivity. In the top 100 lethal proteins detected by NFC from the FYI dataset, 27 bind to 5 or less proteins (low connectivity), 40 bind between 6 to 9 proteins, and 33 bind to 10 or more proteins. On average, NFC was able to detect three times more low-connectivity lethal proteins within the top 100 positions as compared to the connectivity method [3].

A functional distribution analysis of the top ranking lethal proteins reveal that NFC favors lethal proteins involved in basal cell activities. For example, the top 100 lethal proteins detected by NFC in each of the four datasets have GO functions that can be broadly grouped under "translation", "replication", and "transcription" categories. In contrast, the connectivity method by Jeong [3] favored the discovery of lethal proteins with "mitotic cell cycle and cell cycle control" and "fungal and other eukaryotic cell type differentiation" functions. Our preliminary take on the differences is that NFC's functional centrality assumption led to the tendency to find the cores of protein complexes common in some biological pathways, whereas the connectivity method favored the discovery of lethal proteins associated with different functions because such lethal proteins would need to interact with multiple proteins in order to coordinate the global cellular activities needed for cell growth and differentiation.

When a protein has functional annotations, an intelligent guess may be made with regards to its lethality based on the biological understanding of its annotated functions. For example, we would expect many proteins involved in translation to be lethal as the process is a basal cellular activity. However, only 12.7% of all the proteins with translation function are actually lethal. This could stem from our current incomplete understanding of the exact roles played by each protein in

translation. On average, GO terms identified in our top 100 NFC proteins are found to associate with lethal proteins 27.3% of the time. On the other hand, 70.0% of the top 100 NFC are lethal where we made use of functional consistency between proteins rather than functional understanding. Thus, integrating PPI network with functional grouping of proteins enable us to better detect lethal proteins than just using functional information alone.

Interestingly, we also found 12 (FYI), 13 (Nature), 18 (Bu), and 12 (DIPS) instances of high-confidence (top 100) predicted lethal proteins that are not in the current reference lethal protein lists, but each has at least one homologous sequence (BLAST's *e-value* $\leq 1e^{-99}$). The presence of homologous copies of a protein within the same genome could potentially buffer the protein deletion which would otherwise lead to lethality. It is conceivable that these predicted proteins require the removal of its associated homologous for lethality to take effect.

Given that the core lethal protein set we used is an incomplete reference list, those highly ranked non-lethal proteins could be novel lethal proteins. We found numerous high-ranked proteins by NFC that turned out to be true lethal proteins listed in other lethal protein reference sets. For example, the proteins YLR268W and YFL017W-A, respectively ranked at the top 16th and 38th positions by NFC in the DIPS and Bu datasets, were absent from our core lethal protein list but found in another lethal protein list used by Jeong [3]. Further comparison of NFC's predictions with two other reference sets used by Jeong [3], and list compiled by *MIPS* [22], found that out of the top 500 ranked proteins, an additional 15 (FYI), 11(Nature), 18 (Bu) and 16 (DIPS) were recorded in these alternative lethal sets.

Regardless of the improved accuracy of our predictive models over time, biological validation of predictions is always necessary. Our hope is that the predictions from this and the future works on computational lethal protein detection can become a useful tool for focusing further experiments that can lead to a shorter time frame required for lethal protein discovery and understanding.

Acknowledgement

We would like to thank our colleagues See-Kiong Ng, Zeyar Aung, and Suryani Lukman for their invaluable assistance rendered during this project.

References

[1] Rosamond, J., and Allsop, A., Harnessing the power of the genome in the search for new antibiotics, *Science*, 287:1973–1976, 2000.

[2] Jeong, H., and Oltvai, Z.N., and Barab 疽 i, A.L., Prediction of Protein Essentiality Based on Genomic Data, *Complexus*, 1(12):19–28, 2003.

[3] Jeong, H., Mason, S.P., Barab 疽 i, A.L., and Oltvai, Z.N., Lethality and centrality in protein networks, *Nature*, 411(6833):41–42, May 2001.

[4] Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray., A.W., From molecular to modular cell biology, *Nature*, 402(6761):47–52, 1999.

[5] Spirin, V., and Mirny, L. A., Protein complexes and functional modules in molecular networks, *Proc Natl Acad Sci USA*, 100(21):12123–12128, Oct 2003.

[6] Yu, H., Greenbaum, D., Lu, H. X., Zhu, X., and Gerstein, M., Genomic analysis of essentiality within protein networks, *Trends Genet*, 20(6):227–231, 2004.

[7] Joy, M. P., Brock, A., Ingber, D. E., and Huang, S., High-betweenness proteins in the yeast protein interaction network, *J Biomed Biotechnol*, 2005(2):96–103, 2005.

[8] Schmith, J., Lemke, N., Mombach, J. C. M., Benelli, P., Barcellos, C. K., and Bedin, G. B., Damage, connectivity and essentiality in protein-protein interaction networks, *Physica A Statistical Mechanics and its Applications*, 349(3-4):675–684, Apr 2005.

[9] Estrada, E., Virtual identification of essential proteins within the protein interaction network of yeast, *Proteomics*, 6(1):35–40, Jan 2006.

[10] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., *et al.*, Gene Ontology: tool for the unification of biology, *Nature Genet*, 25:25–29, May 2000.

[11] Wu, X., Zhu, L., Guo, J., Zhang, D. Y., and Lin, K., Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations, *Nucl. Acids Res.*, 34:2137–2150, 2006.

[12] Han, J. D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., *et al.*, Evidence for dynamically organized modularity in the yeast protein-protein interaction network, *Nature*, 430(6995):88–93, Jul 2004.

[13] Bu, D., Zhao, Yi., Cai, L., Xue, Hong., Zhu, X., Lu, H., *et al.*, Topological structure analysis of the protein-protein interaction network in budding yeast, *Nucl. Acids Res.*, 31(9):2443–2450, May 2003.

[14] Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D., Dip: the database of interacting proteins, *Nucl. Acids Res.*, 28(1):289–291, Jan 2000.

[15] Giaever, G., Chu, A. M., Ni, L., Connelly, C., *et al.*, Functional profiling of the *Saccharomyces cerevisiae* genome, *Nature*, 418(6896):387–391, Jul 2002.

[16] Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., and Cullin, C., A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*, *Nucl. Acids Res.*, 21(14):3329–3330, July 1993.

[17] Wach, A., Brachat, A., Pohlmann, R., and Philippsen, P., New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*, *Yeast*, 10:1793–1808, 1994.

[18] Schwikowski, B., Uetz, P., and Fields, S., A network of protein-protein interactions in yeast. *Nat Biotechnol*, 18(12):1257–1261, Dec 2000.

[19] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M., Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps, *Bioinformatics*, 21:302–310, 2005.

[20] Jiang, T., and Keating., A. E., AVID: An integrative framework for discovering functional relationships among proteins, *BMC Bioinformatics*, 6:136, 2005.

[21] Chua, H. N., Sung, W. K., and Wong, L., Exploiting indirect neighbors and topological weight to predict protein function from protein-protein interactions, *Bioinformatics*, 22:1623–1630, 2006.

[22] Guldener, U., Mnsterktter, M., Kastenmller, G., Strack, N., Helden, J. V., *et al.*, CYGD: the Comprehensive Yeast Genome Database, *Nucl. Acids Res.*, 33:D364–D368, 2005.