Develop NetBAS - a next-generation software tool for enrichment/suppression analysis of Gene Ontology terms and other categorical data using protein interaction networks

Project summary:
Gene Ontology (GO) provides a GO terms, to annotate gene functions in all species. Statistical over-representation or enrichment of GO terms in a group of genes of interest can reveal the underlining biological pathways of these genes. This kind of enrichment analysis is an important tool in functional genomics research. Currently, GO enrichment analysis often use Fisher's exact test or hypergeometric test by comparing GO terms of a set of input genes with a large set of background genes. These methods assume genes are freely mixed within cells and do not take the actual gene/protein interaction networks into account. Given the power-law and error-tolerant configuration of gene/protein networks, a better approach is to take gene/protein networks into account for GO term analysis. Current methods also cannot handle input genes that lack GO terms and cannot consider interacting genes from multiple pathways. Here, we propose a novel graph-permutation based method to perform GO enrichment and suppression analysis in the context of gene/protein networks. The proposed method compares the empirical pairwise interactions with permuted networks. Our specific aims are: (1) Design and prototype a new network permutation algorithm, MS02star, that can efficiently permute protein interaction networks with power-law features; (2) Prototype a software tool, Network Based Association Analysis (NetBAS) to perform GO term enrichment/suppression analysis in the context of protein interaction network, and prototype a proof-of-concept web site to provide NetBAS to researchers worldwide. Our long-term goal is to develop NetBAS into a next generation software tool for functional genomics analysis, and will enable researchers around world to better infer gene functions.

Intellectual Merit
The proposed NetBAS algorithm is novel because it infers association patterns based on configurational features of gene/protein networks. In contrast, the current methods based on Fisher's exact test and hypergeometric test will over-estimate associations for weakly connected genes and under-estimate associations for strongly connected genes. The proposed NetBAS will provide new features and capabilities by overcoming some limitations in existing tools. NetBAS is capable of enrichment/suppression analysis for genes without GO annotations, can provide statistical analysis for single genes as long as they have sufficient interacting partners, and can infer association pattern between different types of categorical data. These new features of NetBAS are either nearly impossible or at least problematic to perform by existing methods.

Broad Impact
NetBAS will be prototyped as an open-source R-package. The proof-of-concept web service of NetBAS will be publicly accessible to students and researchers worldwide. The PI has integrated network analysis and permutation into both undergraduate research training and an undergraduate-level course on computational genomics. Since joining UTC, PI has hosted two high-school students in his lab learning basic skills in computational biology. PI had mentored 10 undergraduate researchers, including an honor thesis. NetBAS will be taught at a faculty training workshop organized by the PI in summer of 2020. Video tutorials will be generated and released through PI's YouTube Educational channel. PI has ~9 years of experience of using YouTube educational channel. PI's YouTube educational channel has hundreds of educational videos, ~680 subscribers, and over 350,000 views.

Keywords: Gene ontology, enrichment analysis, network permutation, software development

1. Goals and specific aims

The overall goal of this project is to develop a new Gene Ontology (GO) based functional genomics annotation tool that will provide new capabilities and overcome shortcomings of the existing hypergeometric distribution based methods. The proposed new tool, Network Based Association Study (NetBAS), will perform GO term enrichment and suppression analysis in the context of protein interaction networks, perform statistical analysis based on network permutations, and will provide more biologically relevance than existing methods.

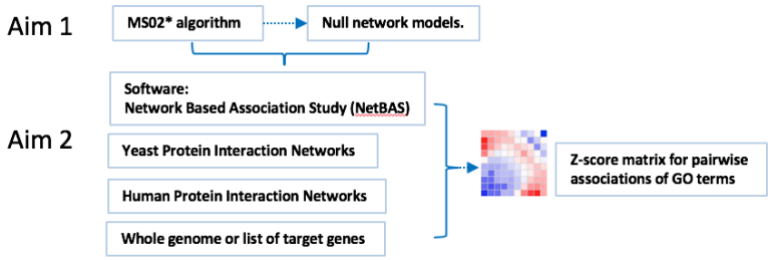

Figure 1. Overview of Specific Aims.

The specific aims are (Figure 1):

Aim 1: Prototype a new graph permutation algorithm, ms02star, that can efficiently generate null network models from protein interaction networks with power-law features.

Aim 2: Prototype NetBAS - a new software tool to perform enrichment and suppression analysis of GO terms and other biological categorical data based on protein interaction networks, and prototype a web site to provide NetBAS to the research communities around the world.

2. Background and Significance

2.1 Gene ontology is an important tool for functional genomic annotation

The functions and activities of proteins can be categorized by the Gene Ontology (GO) terms (Huang, Sherman et al. 2009). The GO terms include biological processes that genes/proteins are involved, sub-cellular sites where proteins are localized, molecular functions that proteins can perform. The GO Consortium (geneontology.org) provides experimental evidence-based GO annotations to diverse organisms, including the budding yeast and humans.

Functional inference of a list of input genes can be achieved by the over-represented GO terms in the input list as compared to a large background list of genes (Huang, Sherman et al. 2009). Usually, the entire genes of a genome are used as the background list of genes. Many bioinformatics tools have been developed to perform enrichment analysis of GO terms, such as DAVID, GoMiner, and TopGo. GO term enrichment analysis are routinely performed to infer the underlying pathways of list of genes identified by high-throughput phenotypic screens, differential gene expression analysis, or genetic screens.

2.2 Current GO term enrichment analyses are based on unrealistic null models

Proper null models are the basis for statistical evaluations. Current GO term enrichment analyses reply on Fisher's exact test or hypergeometric probability distribution (Huang, Sherman et al. 2009). For example, one of the most frequently used GO term enrichment tool is DAVID (Huang, Sherman et al. 2009). DAVID uses Fisher exact text or a variant, implemented by Gaussian hypergeometric probability distribution that describes sampling without replacement from a finite population consisting of two types of elements (Hosack, Dennis et al. 2003). The null hypotheses of the current methods basically assume genes/proteins can interact freely with other genes/ proteins. These assumptions are unrealistic because protein interaction networks have power-law and error-tolerant features. In Figure 2 at section 3.1, a simple example network with power-law and error-tolerant feature is presented to show that some nodes will never interact with each other during random network permutations.

Furthermore, current methods share some common limitations: (1) Genes with unknown GO terms cannot be properly evaluated. (2) The size of input gene list is recommended to be hundreds for appropriate hypergeometric approximation. (3) Interacting genes from different pathways are not considered. (4) Association of GO term with other categorical data are hard to analyze. These limitations will be overcome by our proposed NetBAS.

2.3 Protein interaction networks are the context of biological activities and functions
The network biology has visualized a living cell as a "network of networks" from which all biological activities emerge (Barabasi and Oltvai 2004). Among these networks, the protein interaction network (PIN) is described by all protein-protein interactions (PPI) in the cell. High throughput datasets continued to extend the coverage of proteins from different organisms. For example, the BioGRID dataset had expanded from ~13,000 interactions in budding yeast *Saccharomyces cerevisiae* in its original GRID dataset (Breitkreutz, Stark et al. 2003) to over 700,000 interactions in the same species in a recent release (v. 3.4.163) (Chatr-Aryamontri, Oughtred et al. 2017). These PIN resources now serve as important framework in biological research including diseases (Barabasi, Gulbahce et al. 2011). These large amount of protein interactions made our proposed method possible.

2.4 Principle and significance of the proposed NetBAS (Intellectual Merit)
In this proposal, we propose a new software tool, Network Based Association Study (NetBAS), a network-based method to elucidate the associations among different GO categories (each GO annotation term is regarded as a unique category). In this method, the protein interaction networks (PIN) is treated as an undirected and unweighted pairwise network. For a pair of interacting proteins $<g_i,g_j>$, we will consider all possible pairwise GO term associations between these two proteins. Total interacting frequency between a pair of GO terms will be summed over all interacting protein pairs. Null expectations are estimated from permutated networks in which gene connectivity is preserved. Network permutation will be achieved through a novel algorithm, MS02star, that can efficiently permute protein interaction networks with power-law features. Self-interactions and redundant-interactions are prohibited in network null models. Z-scores between all possible pairwise associations of GO terms are estimated to compare observations with null expectations.

Although designed for whole-genome analysis, NetBAS can also analyze a list of input genes in similar way as the existing methods. Traditional GO enrichment tools, such as DAVID (Huang, Sherman et al. 2009), generally require 100 – 2000 genes for the input gene list. In comparison, NetBAS can perform GO analysis for single hub genes.

Because NetBAS focuses on the partner proteins of a selected list, it examines association between proteins. Consequently, NetBAS offers a GO enrichment analysis from a different perspective to the existing methods. NetBAS can perform association analysis between GO terms and other biological data, such as lifespan, fitness, and morphological measurements. These new capabilities will provide new insights on the biological roles of the input gene list because interacting proteins suggest common biological activities. NetBAS can be further applied to GO categories in combination with other attributes of proteins (i.e., vertices of the PIN), for example, vertex degree and betweenness. In addition, NetBAS can be extended to other biological networks such as the genetic interaction network (GIN) (Costanzo, VanderSluis et al. 2016), metabolic network (Herrgard, Swainston et al. 2008), and gene-regulatory network (Hecker, Lambeck et al. 2009). In this proposal, we will first test NetBAS in yeast protein interaction networks, and then apply it in human protein interaction networks.

3. Research Plans

Overall strategy

We plan to first implement a new algorithm, MS02star, that can permute power-law networks much faster than the original MS02 algorithm. We will then use MS02star to prototype a software tool NetBAS that uses network null models to evaluate observed GO term association patterns.

3.1 (Aim 1) Prototype a new graph permutation algorithm, ms02star, that can efficiently generate null network models from protein interaction networks with power-law features.

A proper null model is important for statistical analysis. For null network models, the connectivity of each node in the permutation must be the same as the original networks (Maslov and Sneppen 2002). In addition, the self-interactions and redundant edges are prohibited (Qin, Lu et al. 2003). The MS02 algorithm is a commonly used network permutation algorithm. However, MS02 does not work well for power-law networks, as explained by the following example (Figure 2).

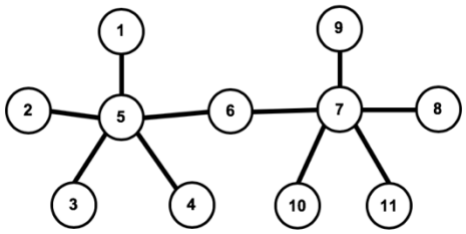

Figure 2. An example network to show that proteins nodes cannot be freely mixed during permutation when protein's interactions are persevered. In this example, proteins with single interactions, such as node 1 and node 2 will never interact with each other during permutation. This scenario is typically ignored in hypergeometric tests. Permutation of a large network with these features are extremely slow using the original MS02 algorithm because only a small number of possible combinations are allowed.

A Motivation Problem: An example small network is presented in Figure 2. This network mimics the power-law feature of protein interaction networks in the sense that there are highly connected nodes (such as node 5 and 7) and there are many poorly connected nodes (such as 1, 2, 3, 4, 8, 9, 10, and 11). This example also mimics the error-tolerant feature of protein interaction networks in the sense that hub nodes, 5 and 7, avoid each other. During network permutations that require node degree preservation, allowable changes include switch between 5 and 7, and switches between any pair of nodes with only one interaction, such as 2 and 10. Self-interactions and redundant interactions are prohibited. The original MS02 algorithm randomly generates interacting pairs, can only slowly converge to final states, and its runs often lead to unsuccessful permutations. Because protein interaction networks have power-law configuration features, the slow converging problem of MS02 for power-law networks is a major computing bottleneck and likely has prevent MS02 algorithm from being widely adopt in biological network analysis.

To address this problem of MS02, we proposed a recursive MS02star algorithm (Box 1). In the proposed recursive MS02star algorithm, illegal edges include self-interactions and redundant interactions. MS02star will permutate edges based on the product of the normalized connectivity of the two vertices in the edges. Edges from highly connected nodes will be preferentially preserved in $E'_{good,keep}$, whereas edges from poorly connected nodes tend to assigned to $E'_{good, redo}$, which will be broken up and reshuffled in subsequent recursive calls. We have to combine $E'_{good, redo}$ with illegal edges E'_{bad} for recursive calls because illegals edges alone are unlikely to converge to legal edges. The fraction of legal edges to be broken up is controlled by a hyper-parameter 'cutoff'. Preliminary test show that cutoff = 0.1 works well for both yeast protein interaction networks and human protein interaction networks.

Box 1. The proposed recursive MS02star algorithm

```
cutoff = 0.1
// cutoff is a hyperparameter to decide the portion of legal edges E'good
// to be broken up for subsequent recursive calls

MS02*(V, E) { //V is the set of vertices, E is the set of edges
  E' = Random pairs by bootstrapping vertices in E without replacement
  E'bad = illegal edges in E'
  If (E'bad is not empty) { // Reshuffle the vertices in illegal edges
    E'good = legal edges in E'
    // Assign weights to edges
    w(E'good) = product of normalized connectivity + Gaussian noise
    E'good,keep = Edges in E'good with w in the top cutoff percentile
    E'good,redo = Edges in E'good with w in the top (1-cutoff) percentile
    E'' = E'good,redo +E'bad
    V''= vertices in E''
    MS02*(V'', E'')  // Recursive call
  } else {  // E'bad is empty, base situation.
    return (V, E')
  }
}
```

A preliminary test of this MS02star algorithm in R shows that it can generate successfully network permutation much faster than the original MS02 algorithm for yeast PIN and human PIN. With the support of this project, we plan to systematically compare the running time and successful rate of MS02 and MS02star using yeast and human PINs.

3.2 (Aim 2) Prototype NetBAS - a new software tool to perform enrichment and suppression analysis of GO terms and other biological categorical data based on protein interaction networks, and prototype a web site to provide NetBAS to the research communities around the world.

3.2.1 Design of NetBAS

As shown in Figure 3, iterative reshuffling of all nodes that constitute the network is performed until the permutation is free of self-interactions and redundant interactions, but with the connectivity of all nodes conserved as the original network. For a gene set comprises genes [g1, g2, …], the partner proteins from the original PIN and permutations are [p1o, p2o, …] and [p1p, p2p, …], respectively. Using databases (or node data) as the node attributes, such as GO database, all nodes of the network are categorized in each of the attributes. Any pair of attributes can be applied for two-dimensional pairwise networks as exemplified in present work. Finally, Z-scores are calculated as $Z_{i,j} = (F_{i,j}^{obs} - F_{i,j}^{mean})/sd_{i,j}$, where $F_{i,j}^{obs}$ is the observed frequency of edges with one node in the i-th category of attribute X and the other node in the j-th category of attribute Y; whereas $F_{i,j}^{mean}$ is the mean frequency, and $sd_{i,j}$ is the standard deviation from all permutations. A cutoff was used to indicate enrichment (Z > +ε, red) or suppression (Z < -ε, blue) of interactions in the original network, compared to the random "null" models, with the absolute values quantify the intensity of the enrichments and suppressions.

We will prototype NetBAS as an open source R package through a GitHub public repository.

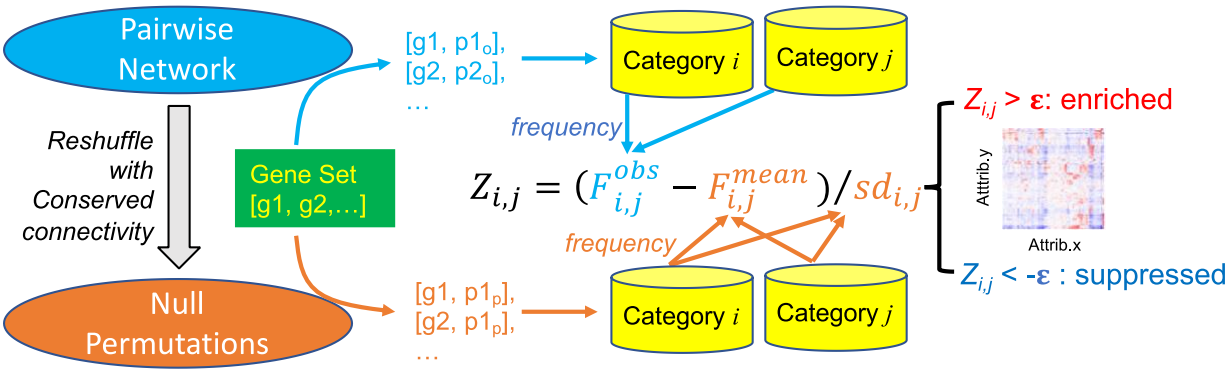

Figure 3. An overview of NetBAS

3.2.2 Data sources

Networks Data

For the yeast protein interaction network (PIN), all physical interactions were extracted from the BioGrid network (v3.4.163) (Chatr-Aryamontri, Oughtred et al. 2017) of *S. cerevisiae*. Concerning all proteins in the yeast genome, this network has 102,790 edges connected by 5,777 proteins, including 19 mitochondrial proteins. The human PIN was obtained from the Inweb_InBioMap network (Li, Wernersson et al. 2017), and has 592,685 edges connected by 16,641 proteins. Self-interactions and redundant interactions (A:B vs B:A are considered redundant) have been removed in the networks. The PINs were treated as unweighted and undirected binary networks. In addition, edges of the networks have been restricted to those with explicit GO annotations assigned for the two connecting genes.

GO annotations Data

All GO annotations were downloaded from the Gene Ontology Consortium (GOC) website (www.geneontology.org) (Ashburner, Ball et al. 2000) updated in early August 2018. The GO terms are grouped in three basic ontologies: biological process (BP), molecular function (MF) and cellular component (CC). The genes in yeast PIN carry 3150 BP, 2069 MF and 823 CC terms. Many of the GO terms are shared by only a few genes, or even a single gene (i.e., “orphan” GO term). These less-shared GO terms may have limited statistical significance. In general, GO subsets (aka GO slims) that were cut-down from the full GO lists would be applied to gain a broad overview of the ontology contents. One of such GO slims is that from the Saccharomyces Genome Database (SGD slim,), from which 165 GO terms have been found for all proteins in the yeast BioGrid PIN. Each GO term in the SGD GO-slim was shared by at least 20 genes in BioGrid PIN. Using the same criterion of shared by at least 20 genes, 229 BP, 131 MF and 111 CC terms from GOC (GOC slim) were chosen in our analysis, together with the SGD slim for comparison. The GO annotations for human PIN include 11,883 BP, 4,128 MF and 1704 CC terms. After applying the same criterion (i.e., shared by at least 20 genes), the GOC slim for human PIN has 1,269 BP, 392 MF and 339 CC terms, respectively.

3.2.3 Preliminary results

Preliminary runs were performed for yeast PIN. For all three basic ontology types, biological process (BP), cellular composition (CC) and molecular function (MF), their internal and intra-relationships have been assessed using the Z-scores in yeast (Figure 4). Interactions between proteins within a same GO category is generally enriched. This is consistent with the finding of a previous (Maetschke, Simonsen et al. 2012). Interactions between different GO categories are also observed. Note that the dendrograms shown in Figs 4 are based on the hierarchical clustering of the Z-scores, rather than the GO similarities. Hence, NetBAS presents an unbiased

evaluation for enrichment or suppression of interaction frequencies between proteins carrying different GO terms.

It should be noted that these network permutations were performed using the ts117 HPC cluster at the UTC SimCenter.

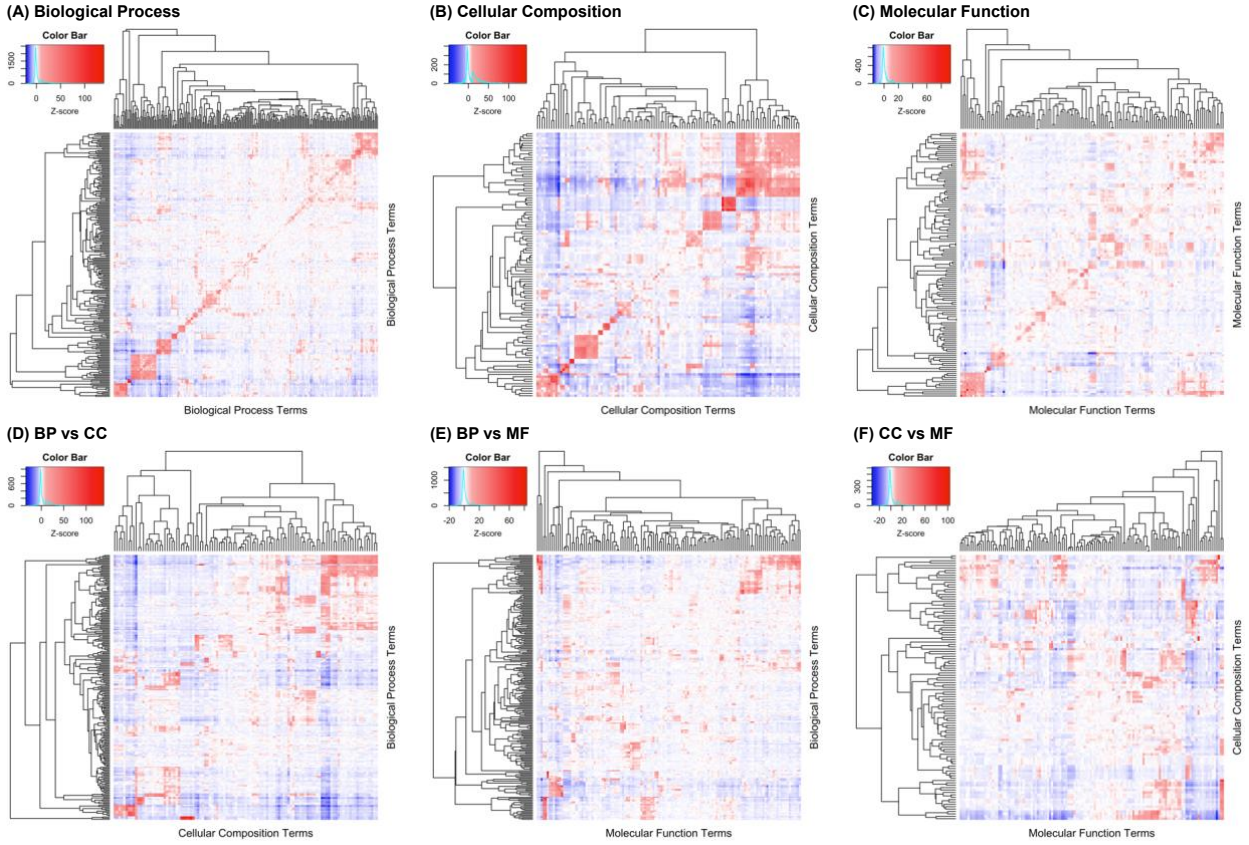

Figure 4. Heatmaps of the GO-GO Z-scores in yeast (*S. cerevisiae*) PIN for GO terms of (A) biological process (BP, 229 terms), (B) cellular composition (CC, 111 terms), and (C) molecular function (MF, 131 terms); as well as the cross relationships of (D) BP versus CC, (E) BP versus MF and (F) CC versus MF.

3.2.4 Prototype a proof-of-concept website to provide NetBAS to research communities.

Although NetBAS will be prototyped in an open source R package, a web version would improve the accessibility of NetBAS to biological research communities. A fully functional NetBAS web service would require HPC capability, which is beyond the allowable budget of this pilot grant. We will prototype a limited featured NetBAS website as a proof-of-concept. This proof-of-concept website will strengthen our external grant applications. We plan to prototype this web site through an Apache HTTP server. Network permutations will be pre-computed for yeast and human PINs. Network data and GO terms may be stored in either a MySQL database or in an R binary data format. NetBAS will be the backend computing engine. Server side JavaScript will be used to develop the web interface. Alternatively, we can use the recently developed HTTP server package 'servr' in R, and use R Markdown to generate dynamic reports.

4. Originality

The proposed NetBAS algorithm is novel because it infers association patterns based on configurational features of gene/protein networks. In contrast, the current methods based on Fisher's exact test and hypergeometric test will over-estimate associations for weakly connected genes and under-estimate associations for strongly connected genes. The proposed NetBAS will provide new features and capabilities by over-coming some limitations in existing tools. NetBAS is capable of enrichment/suppression analysis for genes without GO annotations, can provide statistical analysis for single genes as long as they have sufficient interacting partners, and can infer association pattern between different types of categorical data. These new features of NetBAS are either nearly impossible or at least problematic to perform by existing methods.

5. Role of team members

The PI has five peer-reviewed publications on biological networks, including a PNAS publication. The PI will lead the development of the ms02star algorithm, oversee the implementation of ms02star and NetBAS in R. The co-PI has over 10 years of research experience in computational biology. The co-PI is experienced in performing large scale network analysis using high performance computing clusters at the SimCenter. The coPI will lead the effort of prototyping NetBAS in R. Both PI and coPI will guide graduate students to develop a proof-of-concept website for research community, in order to demonstrate the utility and importance of NetBAS, and prepare for NIH R21 and other external grant applications.

6. Potential challenges

The proposed website to provide NetBAS computing service to research communities will require a web server and necessary IT support, especially given UTC's strict IT policies.

7. End products

An open source R package of NetBAS will be generated. A proof-of-concept website will be prototyped to provide NetBAS to research communities.

8. Broad impacts

NetBAS will be prototyped as an open-source R-package. The proof-of-concept web service of NetBAS will be publicly accessible to students and researchers around the world. The PI has integrated network analysis and permutation into both undergraduate research training and an undergraduate-level course on computational genomics. Since joining UTC, PI has hosted two high-school students in his lab learning basic skills in computational biology. PI had mentored 10 undergraduate researchers, including an honor thesis. NetBAS will be taught at a faculty training workshop organized by the PI in summer of 2020. Video tutorials will be generated and released through PI's YouTube Educational channel. PI has ~9 years of experience of using YouTube educational channel. PI's YouTube educational channel has hundreds of educational videos, ~680 subscribers, and over 350,000 views.

9. Capacity building and Strategic excellence

Built on PI and coPI's expertise, this project will build the genomic research capacity of UTC, and fully take advantage of the computing facility at SimCenter. The software products of this project, NetBAS and its web version, has the potential of engaging research communities beyond UTC.

References

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Barabasi, A.-L. and Z. N. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nature reviews genetics **5**(2): 101.

Barabasi, A. L., N. Gulbahce and J. Loscalzo (2011). "Network medicine: a network-based approach to human disease." Nat Rev Genet **12**(1): 56-68.

Breitkreutz, B. J., C. Stark and M. Tyers (2003). "The GRID: the General Repository for Interaction Datasets." Genome Biol **4**(3): R23.

Chatr-Aryamontri, A., R. Oughtred, L. Boucher, J. Rust, C. Chang, N. K. Kolas, L. O'Donnell, S. Oster, C. Theesfeld and A. Sellam (2017). "The BioGRID interaction database: 2017 update." Nucleic acids research **45**(D1): D369-D379.

Costanzo, M., B. VanderSluis, E. N. Koch, A. Baryshnikova, C. Pons, G. Tan, W. Wang, M. Usaj, J. Hanchard and S. D. Lee (2016). "A global genetic interaction network maps a wiring diagram of cellular function." Science **353**(6306): aaf1420.

Hecker, M., S. Lambeck, S. Toepfer, E. van Someren and R. Guthke (2009). "Gene regulatory network inference: Data integration in dynamic models-A review." Biosystems **96**(1): 86-103.

Herrgard, M. J., N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Bluthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. Le Novere, P. Li, W. Liebermeister, M. L. Mo, A. P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasic, D. Weichart, R. Brent, D. S. Broomhead, H. V. Westerhoff, B. Kirdar, M. Penttila, E. Klipp, B. O. Palsson, U. Sauer, S. G. Oliver, P. Mendes, J. Nielsen and D. B. Kell (2008). "A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology." Nature Biotechnology **26**(10): 1155-1160.

Hosack, D. A., G. Dennis, Jr., B. T. Sherman, H. C. Lane and R. A. Lempicki (2003). "Identifying biological themes within lists of genes with EASE." Genome Biol **4**(10): R70.

Huang, D. W., B. T. Sherman and R. A. Lempicki (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nature Protocols **4**(1): 44-57.

Li, T., R. Wernersson, R. B. Hansen, H. Horn, J. Mercer, G. Slodkowicz, C. T. Workman, O. Rigina, K. Rapacki, H. H. Staerfeldt, S. Brunak, T. S. Jensen and K. Lage (2017). "A scored human protein-protein interaction network to catalyze genomic interpretation." Nat Methods **14**(1): 61-64.

Maetschke, S. R., M. Simonsen, M. J. Davis and M. A. Ragan (2012). "Gene Ontology-driven inference of protein-protein interactions using inducers." Bioinformatics **28**(1): 69-75.

Maslov, S. and K. Sneppen (2002). "Specificity and stability in topology of protein networks." Science **296**(5569): 910-913.

Qin, H., H. H. Lu, W. B. Wu and W.-H. Li (2003). "Evolution of the yeast protein interaction network." Proceedings of the National Academy of Sciences **100**(22): 12820-12824.

External Funding Plan

PI received previous CEACE awards on biological network big data and bioimage deep learning studies. These seed grants supported the PI to submitted 6 external grant applications. Out of these efforts, the PI has won an NSF BD Spoke award recently.

With support of this CEACSE award, PI will apply for the following external funding opportunities:

(1) NIH R21. Target date: February 16, 2019.

(2) NSF Rules of Life: Forecasting and emergence in living systems. Target date: May 2019. Rules of Life program has no fixed deadline. PI's deadline is self-imposed.

(3) NIH R15. Target date: June 25, 2019