

Discovery of Phenotype Related Modules by Incorporating Expression Data to Protein Interaction Networks

Shouguo Gao¹, Jennifer Bills², Xujing Wang^{1*}

1. Max McGee National Research Center for Juvenile Diabetes & Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, WI USA
 2. Department of Math, Stats, and Comp. Sci, Marquette University, Milwaukee, WI USA
- *email: xwang@mcw.edu

Introduction

The modular structure of protein-protein interaction (PPI) networks is dynamic and condition dependent. Here we report a novel method to identify phenotype-related network modules based on gene expression changes and apply it to several public datasets (one reported here). It is designed based on the fact that hub genes are usually the most critical to network function; nevertheless hub genes often show low levels of change, which are easily overshadowed by the changes in non-hub genes.

Methods and Datasets

The prostate cancer gene expression and phenotype data were downloaded from <http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE3494>. Oncogene list was downloaded from <http://embryology.med.unsw.edu.au/DNA/DNA10.htm> and <http://www.sanger.ac.uk/genetics/CGP/Census>.

The complete PPI network was constructed according to BIND. We define phenotypically important PPI sub-modules to be those that are: (1) sub-networks of genes whose expression values all correlate with the phenotype (phenotype correlated genes, PCG); or, (2) sub-networks that contains active hub genes (AHG). Here an AHG is defined to be one connected to many PCGs, while itself is not necessarily a PCG. The details are given in Table 1. When defining the AHG score, in addition to its nearest neighbors, we also include its neighbors at depth 2 to lower the effect of measurement noise. GOSTat was used to find the over-represented GO-terms. The phenotype investigated was the Disease-Specific Survival Time.

Results

The complete human PPI network is visualized with Cytoscape (Figure 1A). We found that PCGs tend to interact with each other. To quantitatively evaluate our observation, we performed permutation for 100000 times and calculated odds value = $\frac{\text{percentile of gene pairs above PCG_cutoff that interact in the network}}{\text{percentile in permutation}}$. The results are presented in

figure 1B.

The oncogenes tend to have a higher degree in PPI (Figure 1C), demonstrating the necessity to design algorithm to “DIG OUT” hub genes.

Eight phenotypically important modules were identified (Figure 1D). Interestingly, 26 out of the 134 genes in them are oncogenes, representing 2-fold increase over random selection (140 out of 1468, $p < 0.003$). Figure 1E-F shows the two largest modules. Most oncogenes are not PCGs, thus would have been missed if we were to identify phenotype-related genes at an individual gene level. This demonstrates the potential of our approach in identifying candidate genes at a network level. Furthermore, most other genes in these modules are also cancer-related. For example, Gene 5111, PCNA is an AHG as it connects with many PCGs, but itself is not a PCG. It has been utilized as prognostic biomarkers for many types of cancer.

We have further examined the ontology of the genes in these modules, we found that almost all top over-represented GO terms in Module4 and Module6 are also over-represented in oncogenes, and they fit well to our understanding of cancer (Figure 1E-F).

We applied our algorithm to several other Datasets, such as the obesity study (http://www.diabetesgenome.org/thirdpartydata/lusis_060424), duplicated the trend that high PCGs tend to interact with each other and identified biological meaningful modules.

Table 1: Detailed Algorithm

Input: PPI networks: $G(V, E)$; expression: $X(V, J)$; phenotype: $Pheno(J)$; PCG_cutoff ; AHG_cutoff	
<p>Step 1: Score vertex (1) for all v in G $PCG_score(v) = Corr(X(v, J), Pheno(J))$, $Corr$: Pearson correlation coefficient. end for $mean_PCG_score = mean(PCG_score(v))$ (2) for all v in G find all nearest neighbors (v_nbs) of v if $num(v_nbs) < 2$ then $AHG_score(v) = 0$, not hub gene. else $softCore = 0$ for all v_nb in v_nbs find all nearest neighbors (v_nb_nbs) of v_nb $local_mean_score = mean(PCG_score(v_nb_nbs), PCG_score(v_nb))$ if $local_mean_score > mean_PCG_score$ $softCore = softCore + local_mean_score - mean_PCG_score$ end if end for $AHG_score(v) = softCore * PCG_score(v)$ end if end for</p>	<p>Step 2: Search sub-graphs procedure FIND_SUBGRAPHS for all sorted v with $PCG_score(v)$ in G if not already seen v then call FIND_SUBGRAPH(v) end for end procedure procedure FIND_SUBGRAPH input: seed vertex: s output: complex: c for all neighbor ($s_neighbor$) of s_vertex if $PCG_score(s_neighbor) > PCG_cutoff$ or $AHG_score(s_neighbor) > AHG_cutoff$ then add $s_neighbor$ to c call: FIND_SUBGRAPH($s_neighbor$) end if end for end procedure</p>

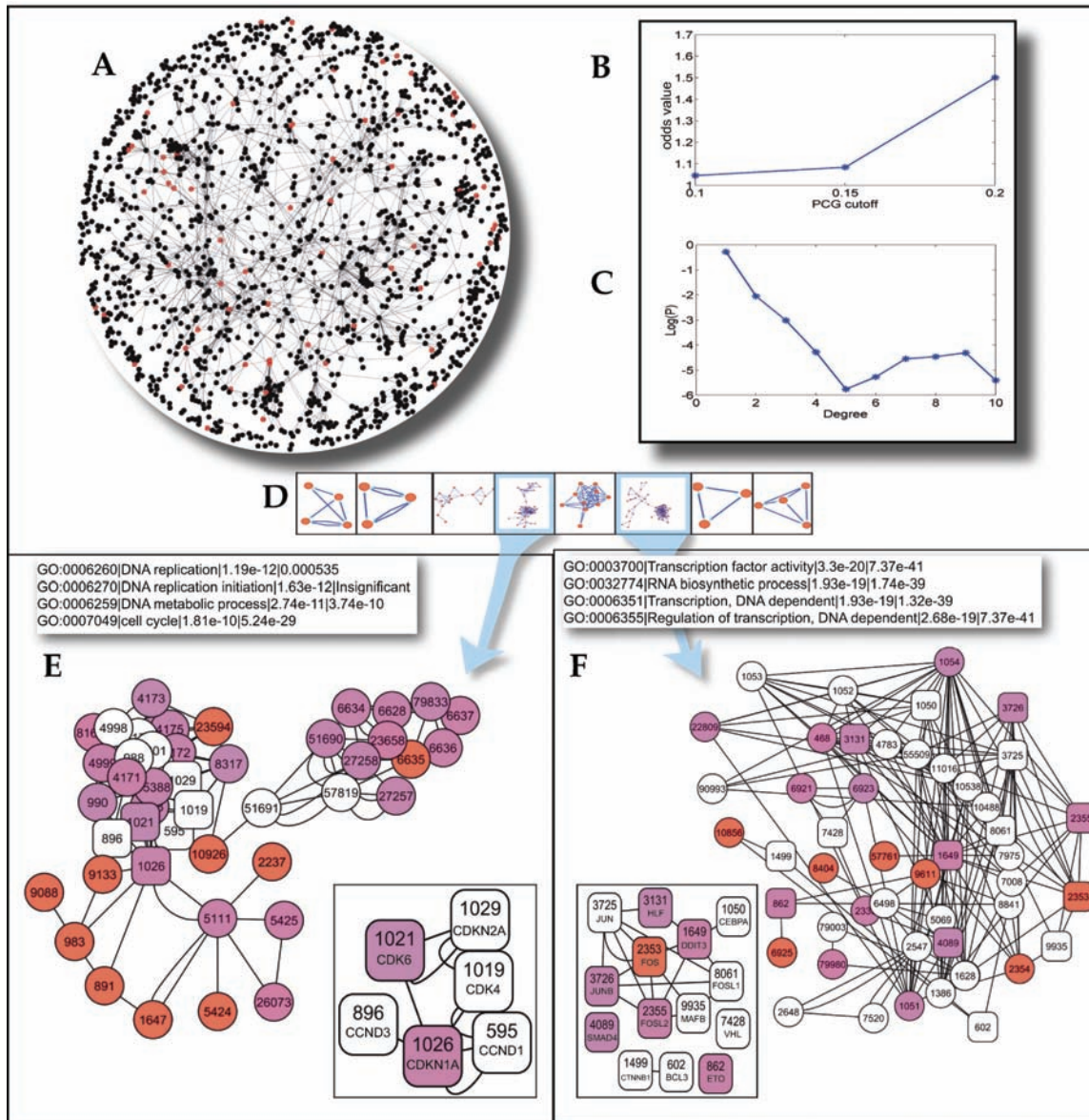


Figure 1: (A) Human gene networks from BIND, red spots represent PCGs ($PCG_score > 0.2$). (B) Gene pairs with higher PCGs tend to interact to each other, the trend is more significant with higher PCG_scores. (C) At different Degree cutoff (d), oncogene number with degree $> d$, total oncogene number, gene number with degree $> d$, total gene number were used to calculate Fisher's test log P Values (only genes in BIND are considered). (D) All eight significant modules with $PCG_cutoff = 0.2$ and $AHG_cutoff = 0.039$. (E-F) Module4 and Module6, red and pink spots represent genes with $PCG_score > 0.2$ and $PCG_score > 0.1$, respectively. Oncogenes were shown with round rectangle and all oncogenes were shown in the lower inset boxes; the upper inset boxes are top 4 GO terms and P value found by GOSTat through comparing genes in modules with all genes in BIND, the last column is P value of corresponding GO got by comparing all oncogenes with all genes in HG-U133A.