

## Objective criteria for defining hubs in protein interaction networks

Ravishankar R. Vallabhajosyula<sup>1</sup>, Deboki Chakravarti<sup>2</sup>, Samina Lutfaeli<sup>3</sup>, Animesh Ray<sup>1</sup>,  
and Alpan Raval<sup>1,4</sup> \*

1. Keck Graduate Institute, Claremont, CA 91711, USA

2. California Institute of Technology, Pasadena, CA 91125, USA

3. University of Chicago, Chicago, IL 60637, USA

4. Claremont Graduate University, Claremont, CA 91711

\*Corresponding Author. Email: araval@kgi.edu

It is often thought that modularity in protein interaction networks (PINs) is associated with “hub” proteins that anchor functional units and thus play important roles in the cell. In spite of their putative importance, criteria for identifying hubs currently involve the use of *ad hoc* degree and connectivity cutoffs. Consequently, analysis of graph properties of PINs remains somewhat subjective. In this work, we propose three objective criteria to identify hub nodes in any PIN. The first criterion is a purely connectivity-based definition, motivated by the observation that proteins of high degree in a PIN tend not to preferentially connect to each other. This criterion identifies a corresponding cutoff degree for hubs so that hubs will not preferentially connect among themselves. The second criterion is motivated by the observation that high degree proteins are enriched for functionally essential proteins. The cutoff degree for hubs is then identified as the degree value at which there is maximal discrimination between essential and non-essential proteins. The third criterion for identifying hubs is based on the ability of the identified hub set to partition significantly into distinct “date” and “party” hubs, i.e., hubs that have high co-expression with their immediate neighbors in a PIN and hubs that have low co-expression with their neighbors, respectively.

We apply these criteria to five yeast PIN datasets, and find reasonable agreement between ‘hubs’ as identified by these rather disparate criteria for high-throughput (HTP) interaction dominated datasets. However, this agreement does not hold when the data set also includes a significant proportion of literature-curated interactions and therefore brings into question the usefulness of the hub concept for large, high-confidence PINs.

The five datasets used in this study are the HC network [1], the Literature curated (LC) network [4], a combined dataset HC-LC. The last is a subset of the HC network that includes only multi-validated High Throughput interactions HC<sup>h</sup> [1] and the FYI network [2].

Figure 1(a) is an example of the connectivity-based technique for hub identification. We sort the network nodes in decreasing order of degree, extract the subgraph corresponding to the first  $n$  nodes in this sorted list, and examine the subgraph connectivity as a function of  $n$ . The sharp rise in subgraph connectivity beyond a certain  $n$  value indicates the transition from hubs to non-hubs. In the HC network, for example,

only 40 nodes can be designated as hubs by this criterion. Figure 1(b) is an example of the “gene essentiality” criterion for identifying hubs. Here, the Jensen-Shannon divergence between the essentiality distribution among hubs and non-hubs is plotted vs. the number of identified hubs. The peak in this curve gives the optimal number of hubs that should be chosen to discriminate between essential and non-essential genes.

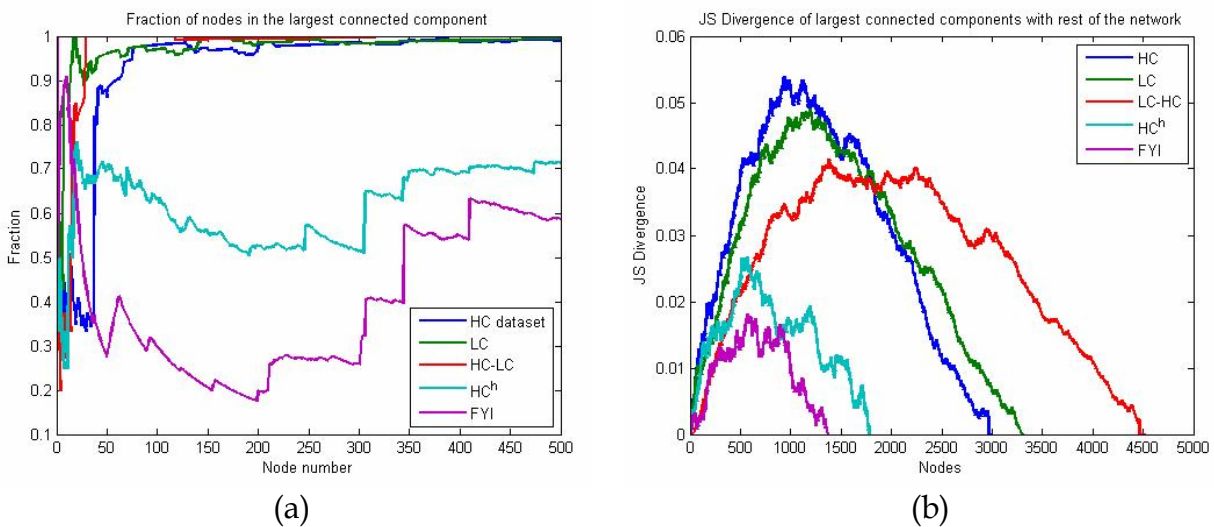


Figure 1: (a) Fraction of nodes in the subgraph of the largest connected component for the five datasets used in this analysis and (b) Enrichment for essential genes among hubs, computed for the five datasets using Jensen-Shannon divergence measure.

## References

1. Batada NN et al. (2006) Stratus Not Altocumulus: A New view of the Yeast Protein Interaction Network. *PLoS Biology*, **4**(10), e317.
2. Han JD et al. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, pp.88-93.
3. von Mering C et al. (2002) Comparative assessment of large-scale data sets of protein-protein interaction. *Nature*, **417**, pp.399-403.
4. Reguly T et al. (2006) Comprehensive curation and analysis of global interaction networks of *Saccharomyces cerevisiae*. *J Biol* **5**: 11.