Mining the Largest Quasi-clique in Human Protein Interactome

Paper ID: 211

Sanghamitra Bandyopadhyay
Co-author: Malay Bhattacharyya

Machine Intelligence Unit
Indian Statistical Institute, Kolkata – 700 108, India

Outline of the presentation

- Motivation
- Related works
- Proposed methodology
- Results & discussion

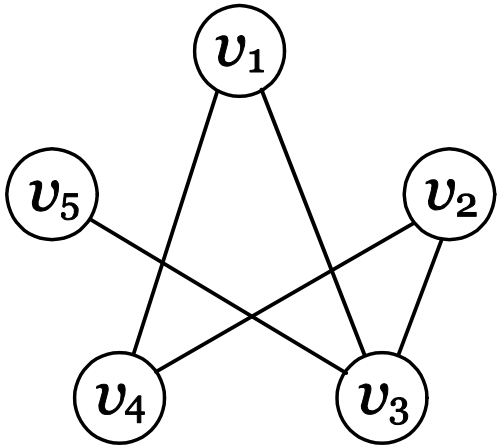
Motivation

Protein interactome

- Total number of proteins in *Homo* Sapiens is in the order of 10^5, whereas the interactome size is as low as 0.0002% [1]
- Protein-protein interaction networks are scale-free [2]
- Mining protein-protein interaction networks for prediction of functions

Quasi-complete graph [3]

- A γ-quasi-complete graph ($\gamma = 0.25$)



- A graph $G = (V, E)$ of degree at least r is γ-quasi-complete for $\gamma \leq r/(|V|\text{-}1)$.
- An acyclic graph $G = (V, E)$ of order at least 2 is $(|V|\text{-}1)^{-1}$-quasi-complete.

Maximum quasi-clique problem in protein-protein interaction network

Definition 1 *(γ-quasi-clique):* In a graph $G = (V, E)$, a subset of vertices $V' \in V$ forms a γ-quasi-clique ($0 < \gamma \leq 1$) if the subgraph induced by V', $G_{V'}$, is a γ-quasi-complete graph.

Problem Statement(MQP) Given a protein-protein interaction network $N = (P, I)$ and the parameter γ, locate a γ-quasi-clique, $N' = (P', I')$ ($N' \subseteq N$), that has the maximum cardinality and $\frac{\min_{p_i \in P'} degree(p_i)}{|P'|-1} \geq \gamma$.

Related works

Finding quasi-cliques for classifying molecular sequences [4]

- Definition of a quasi-clique based on individual degrees
- The problem addressed was to cover all the vertices in a graph with a minimum number of quasi-cliques
- Greedy approximation algorithm with $O(n^3)$ average time complexity
- No approximating factor

Finding quasi-cliques in very large graphs with GRASP [5]

- Definition of a quasi-clique based on total number of edges
- Neither find out the complete set of quasi-cliques nor the largest one
- Greedy randomized adaptive search algorithm

Finding cross-graph quasi-cliques with Crochet [3]

- Definition of a quasi-clique based on individual degrees
- Joint mining of different types of networks for exploring quasi-cliques
- Time complexity of this algorithm linearly grows with the number of quasi-cliques

Finding cross-graph quasi-cliques with Crochet^{+} [6]

- Improvement of Crochet
- Definition of a quasi-clique based on individual degrees
- Joint mining of different types of networks for exploring quasi-cliques
- Time complexity of this algorithm linearly grows with the number of quasi-cliques

Finding maximal quasi-cliques [7]

- Definition of a quasi-clique based on total number of edges and individual degrees
- Time complexity of this algorithm linearly grows with the number of maximal quasi-cliques

Proposed methodology

Precursory details

- Inspired from a dynamic local search algorithm (DLS-MC) used for finding the maximum clique [8]
- Exploits the heuristics that a vertex with degree k cannot be in a γ-quasi-clique of size N if $k < \gamma.(N-1)$.
- Guided by the scale-free property

The proposed algorithm

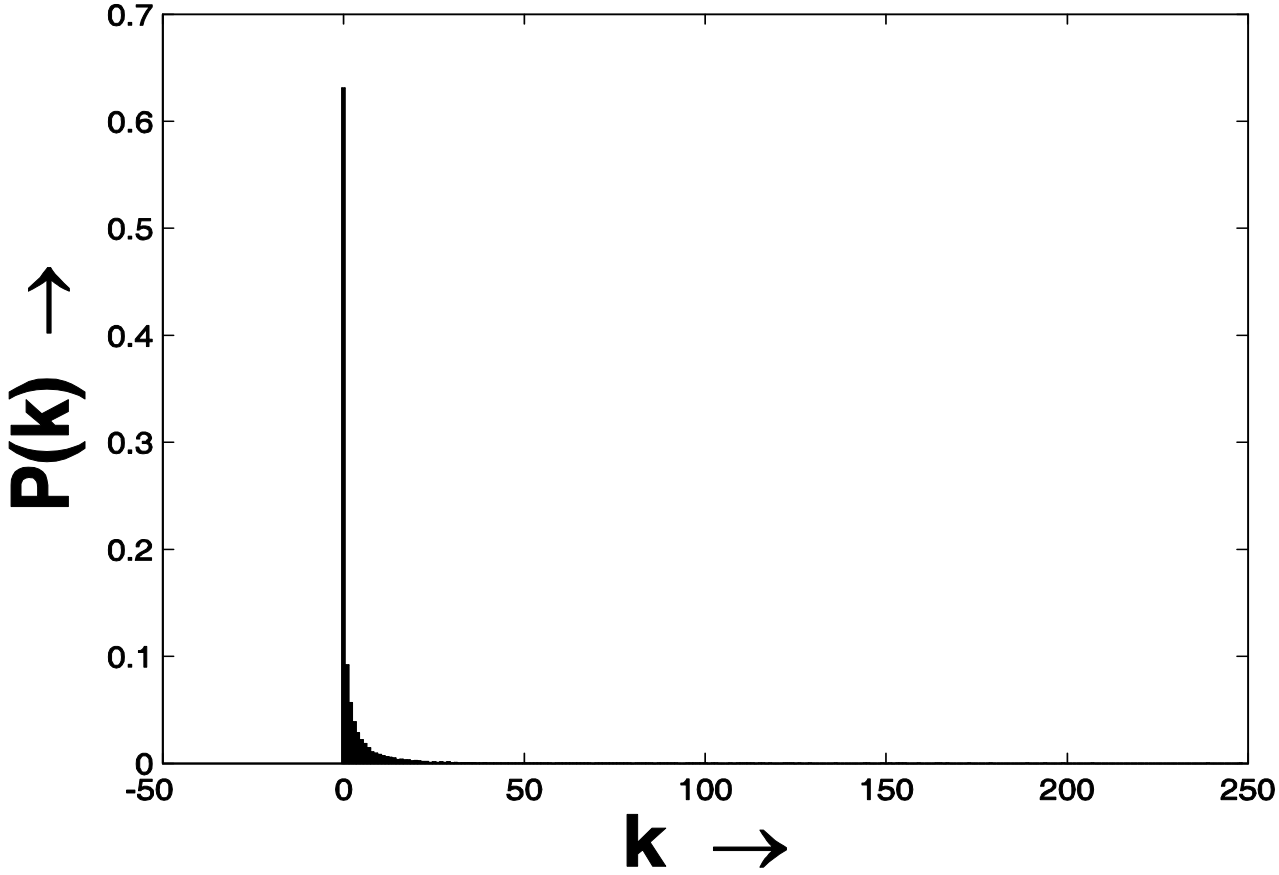
Input: A PPIN N = (P, I) and the parameter γ
Output: The largest quasi-clique N˜ with respect to γ
Algorithmic Steps:
N˜ ← N
while Number of iterations is not sufficient **do**
    if $\Gamma(N\tilde{}) < \gamma$ **then**
        Select a minimum degree protein p_i by breaking the tie arbitrarily
        Remove the protein p_i and all the interactions connected to it from N˜
    else
        Identify the set of proteins P' which have the maximum connectivity with the proteins in N˜ and its degree is supported by the heuristics
        Select a protein p_i from the set P' by breaking the tie arbitrarily
        Attach the protein p_i and all the interactions provided between p_i and the current network N˜ to expand the network N˜
    end if
end while

Results & discussion

Features of the interaction network

- Resource: Human Protein Reference Database (HPRD) [1]
- # Proteins: 25,661
- # Interactions: 37,107
- Clustering coefficient: ~1.13E–4

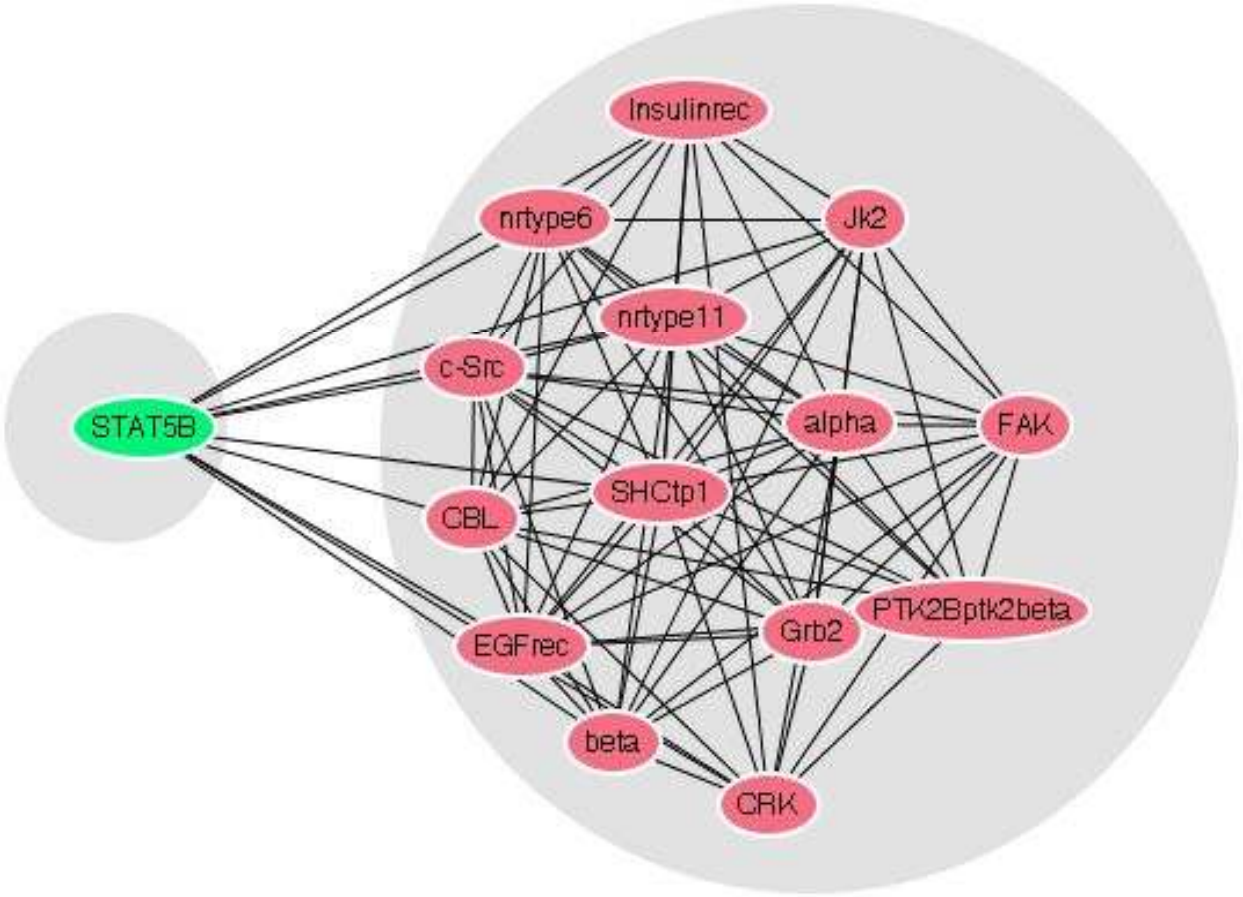
Power-law degree distribution [2]



Stepwise analytical details

- Initial pruning selected 5648 proteins with 35021 interactions therein
- The parameter $\gamma = 0.7$
- Number of iterations = 10,000
- Proteins selected in the largest quasi-clique = 15

The largest 0.7-quasi-clique module



Analysis of functions

Protein Name (Entrez Gene ID)	Molecular Class, Molecular Function and Biological Process
Grb2 (2885)	Adapter molecule Receptor signaling complex scaffold activity Signal transduction, Regulation of cell cycle
EGF receptor (1956)	Receptor tyrosine kinase Transmembrane receptor protein tyrosine kinase activity Signal transduction, Cell communication
Insulin receptor (3643)	Receptor tyrosine kinase Transmembrane receptor protein tyrosine kinase activity Signal transduction, Cell communication
Janus kinase 2 (3717)	Tyrosine kinase Protein-tyrosine kinase activity Signal transduction, Cell communication
CRK (1398)	Adapter molecule Receptor signaling complex scaffold activity Signal transduction, Cell communication
CBL (867)	Ubiquitin proteasome system protein Ubiquitin-specific protease activity Signal transduction, Cell communication, Protein metabolism
Phosphatidylinositol 3 kinase regulatory subunit, alpha (5295)	Adapter molecule Receptor signaling complex scaffold activity Signal transduction, Cell communication

Analysis of functions (contd…)

Protein Name (Entrez Gene ID)	Molecular Class, Molecular Function and Biological Process
PDGF receptor, beta (5159)	Receptor tyrosine kinase Transmembrane receptor protein tyrosine kinase activity Signal transduction, Cell communication
Protein tyrosine phosphatase, non-receptor type 11 (5781)	Tyrosine phosphatase Protein tyrosine phosphatase activity Signal transduction, Cell communication
Protein tyrosine phosphatase, non-receptor type 6 (5777)	Tyrosine phosphatase Protein tyrosine phosphatase activity Signal transduction, Cell communication
c-Src (6714)	Tyrosine kinase Protein-tyrosine kinase activity Signal transduction
SHC (6464) transforming protein 1	Adapter molecule Protein binding Signal transduction, Cell communication
FAK (5747)	Tyrosine kinase Protein-tyrosine kinase activity Signal transduction, Cell communication
PTK2B protein tyrosine kinase 2 beta (2185)	Tyrosine kinase Protein-tyrosine kinase activity Signal transduction
STAT5B (6777)	Transcription factor Transcription factor activity Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism

Future directions

- Integration of protein networks obtained from multiple sources
- Deriving stringent upper bounds for the algorithm
- Improvement with more efficient adaptive heuristics
- Rational drug design by targeting significant hub proteins in the network

Major references

1. T. S. K. Prasad *et al. Human Protein Reference* Database2009 update. *Nucleic Acids Research, 37:D767*–D772, 2009.
2. A. -L. Barab´asi and R. Albert. Emergence of Scaling in Random Networks. *Science, 286(5439):509–512, 1999.*
3. J. Pei *et al*. On Mining Cross-Graph Quasi-Cliques. *Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining,* Chicago, Illinois, USA, 228–238, 2005.
4. H. Matsuda *et al.* Classifying molecular sequences using a linkage graph with their pairwise similarities. *Theoretical Computer Science, 210(2):305–325,* 1999.
5. J. Abello *et al.* Massive quasi-clique detection. *LATIN 2002: Theoretical Informatics, LNCS* 2286:598–612, 2002.
6. D. Jiang and J. Pei. Mining Frequent Cross-Graph Quasi-Cliques. *ACM Transactions on Knowledge Discovery from Data, 2(4):16, 2009.*
7. M. Brunato *et al.* On Effectively Finding Maximal Quasi-cliques in Graphs. *Learning and Intelligent Optimization, LNCS 5313:41–55, 2008.*
8. W. Pullan and H. H. Hoos. Dynamic local search for the maximum clique problem. *Journal of Artificial Intelligence Research, 25:159–185, 2006.*

Thank you