CS4101 B.Comp. Dissertation

Inferring Protein Function Module From Protein Interaction Information

Presented by: Lim Junliang Kevin

Supervised by: Prof Wong Limsoon

NUS

Agenda

- Introduction
- Related Works
- Possible Approaches
- Protein Complex Finder (PCF) Algorithm
- Results
- Further Works
- Conclusion

NUS

Introduction

Introduction

- Recent completion of the Human Genome Project (2003) has indicated a move towards the post-genomic era
- More works have been concentrated on proteomics
- In the course of proteomics, biologists discover protein-protein interaction via wet-lab experiments
- This forms a database of interactions
- We can also represent it as a graph

Introduction

- Formal definition of an interaction network:
  - A Graph is a pair G~(V,E) where V represents the proteins and E represents the interactions between proteins
- The increase in interaction data has spurred potential research problems
- "Given an interaction network, can one infer protein complexes therein?"

Introduction

- Motivation:
  - Biological experiments to determine complexes are time consuming
  - Certain proteins may not have functional annotation, complex prediction allows one to make such inferences by "guilt by association"
  - Has potential in finding undiscovered key proteins involved in diseases

Problem

- Problem formulation:
  - Given a protein interaction network, find subsets (possibly overlapping) of proteins and predict them as complexes
- How?
  - Hypothesis:
    - Protein complexes are likely to be tightly connected clusters within the graph
- Reduced to clustering - within cluster there are many connections, between clusters there are few connections
- Problem is made worse by unreliable data, which are primarily due to laboratory errors

Related Works

Related Works

- Stochastic methods
  - Markov Clustering (MCL) (van Dongen)
- Local Neighborhood Density Search
  - Molecular COmplex Detection (Bader et al)
- Clique finding based methods
  - Protein Complex Prediction (PCP) (Chua et al)
  - Clustering based on Maximal Cliques (CMC) (Liu et al)

Related Works

- MCL
  - Key ideas:
    - Suppose we simulate some k-random walks in the graph, such that k is small enough
    - We would find most paths end up in the same cluster
  - Note:
    - MCL does not pre-process to filter unreliable interactions

Related Works

- CMC
  - Key ideas:
    - Reliability of interactions can be inferred from connection shared by neighbors
    - Tightly connected nodes likely to be complexes
    - Merge tightly connected clusters to get better results

Related Works

- CMC
  - Pre-processing step:
    - Measure reliability of interactions using AdjustCD
    - Example:

$$AdjustCD(u,v) = \frac{2|N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$$

Related Works

- CMC
  - Pre-processing step:
    - Example:

$$AdjustCD(u,v) = \frac{2|N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$$

Related Works

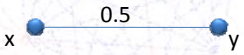
- CMC
  - Pre-processing step:
    - Example:

Max(0,Average neighbors of network – neighbors of u)

$$AdjustCD(u,v) = \frac{2|N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$$

Related Works

- CMC
  - Pre-processing step:
    - Iterated AdjustCD example:



Note: $w^0(u,v)$ = AdjustCD(u,v)

$$w^k(u,v) = \frac{\sum_{x \in |N_u \cap N_v|} (w^{k-1}(x,u) + w^{k-1}(x,v))}{\sum_{x \in N_u} w^{k-1}(x,u) + \lambda_u^k + \sum_{x \in N_v} w^{k-1}(x,v) + \lambda_v^k}$$

where $\lambda_y^k = \max\{0, \frac{\sum_{x \in V} \sum_{z \in N_x} w^{k-1}(x,z)}{|V|} - \sum_{x \in N_y} w(x,y)\}\}$

Related Works

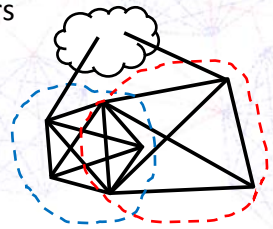
- CMC
  - Algorithm:
    - Step 1: Find all maximal cliques
    - Step 2: Merging Operation
      - Sort cliques according to average AdjustCD score
      - For each clique A
        - For each clique B that has lower score
        - If overlap is above a threshold
          - Measure connectivity between A and B
          - If they are highly connected above a threshold
            - Merge(A,B)
          - Otherwise discard clique B

Related Works

- CMC
  - Score for interconnectivity:

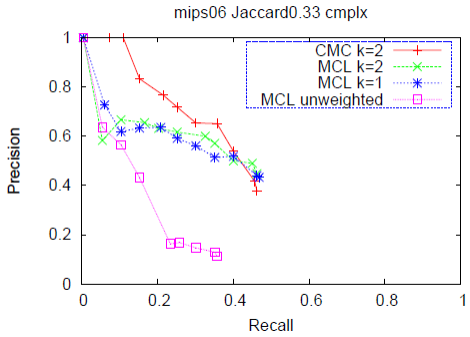
$$Inter-cluster(C_i, C_j) = \sqrt{\frac{\sum_{u\in(Ci-Cj)}\sum_{v\in Cj} w(u,v)}{|C_i - C_j|\cdot C_j} \times \frac{\sum_{u\in(Cj-Ci)}\sum_{v\in Ci} w(u,v)}{|C_j - C_i|\cdot C_i}}$$

    - Measure of whether nodes in the cluster share many neighbors



Related Works

- Analysis of related works:
  - MCL & CMC
    - Low coverage on MIPS yeast complexes



Possible Approaches

Possible Approaches

- Can we find meaningful subsets of proteins other than based on the hypothesis that complexes form tight clusters?
- Can we try to improve existing models to increase their coverage of real complexes?

Possible Approaches

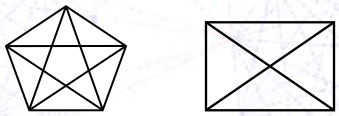
- Frequent Sub-graph mining:
  - Key idea:
    - Look through each real complex
      - Find frequently occurring sub-graph patterns
      - Look for these sub-graphs in the interaction network
      - Improve these sub-graph clusters using some scoring function

Possible Approaches

- Frequent Sub-graph Mining:
  - Problems
    1. Requires sub-graph isomorphism testing:
       - NP-complete problem
    2. No known correlation between frequent sub-graphs and protein complexes
    3. Validation can be a problem

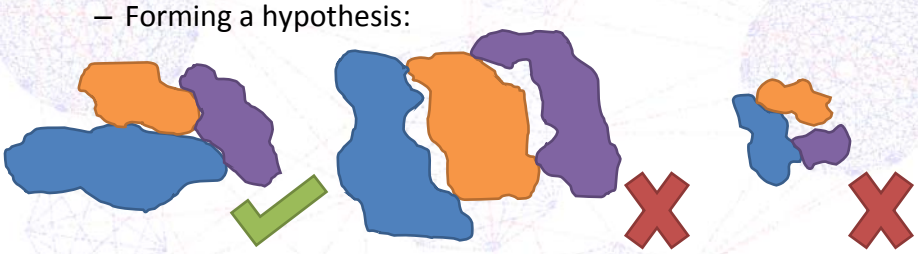
Possible Approaches

- Classifying detected cliques in a feature space
  - Motivation:
    - Main problem of CMC
      - Merging of cliques based on inter-connectivity
      - No biological model behind it
      - Some clusters having an already good representation of complexes before merging



Possible Approaches

- Classifying detected cliques in a feature space
  - Forming a hypothesis:



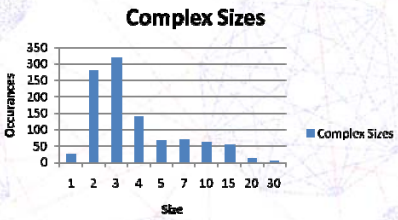
We hypothesize that protein complexes form when individual proteins have a specific proportion of molecular weight. Moreover, each protein with a particular molecular weight might have specific connection features.

Possible Approaches

- We collected a few features from real complexes:
  - Maximum molecular weight protein in complex
  - Minimum molecular weight protein in complex
  - Average molecular weight protein in complex
  - Degree of maximum molecular weight protein in complex
  - Degree of minimum molecular weight protein in complex
  - Average degree of connection in complex
  - Total number of proteins in complex

Possible Approaches

- Negative samples?
  - Find the distribution of complex sizes



  - Follows a gamma distribution α=2.5793 β=1.665
  - Sample from a random gamma distribution a number, p, to represent complex size
  - Randomly pick from a pool of proteins to form a complex of size p

Possible Approaches

- Using SVM we obtain the following validation results:
  - Accuracy: 83.97%
  - Precision:
    - Complex : 0.901
    - Non-Complex : 0.788
  - Recall:
    - Complex : 0.782
    - Non-Complex : 0.904

Possible Approaches

- Problems:
  - Negative samples too far from real complexes
  - Need to consider a few things:
    - Generation of a random network
    - Generation of clusters that have an minimum average connection between vertices
  - Not feasible because there are many ways to choose subsets of proteins, which might not conform to the points mentioned above

Protein Complex Finder

NUS

Protein Complex Finder

- We introduce a new algorithm that makes use of concepts from CMC and try to improve results
- Key ideas:
  - Data pre-processing step
  - Main Algorithm
  - Data post-processing step

NUS

Protein Complex Finder

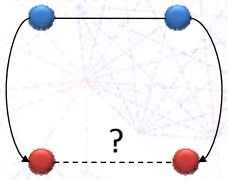
- Pre-processing step 1:
  - Motivation:
    - AdjustCD can already determine quality of network, by virtue of shared degree-1 neighbors
    - Problem is that there could be missing interactions that are not detectable by looking at degree-1 neighbors
    - Can we try to improve the quality by looking at interactions from other species?

Protein Complex Finder

- Pre-processing step 1:
  - Key Ideas:
    - Look at interactions in species A
    - Find their corresponding homolog in species B
    - Add an interaction in species B network if they do not already exist

Protein Complex Finder

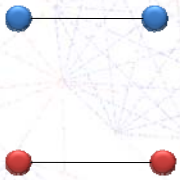
- Pre-processing step 1:



Homologous map

Protein Complex Finder

- Pre-processing step 1:



Protein Complex Finder

- How do we define homology?
  - COG database (Tatusov et al)
  - Blastp (sequence comparison tool)

Protein Complex Finder

- Problems with using COG database:
  - Different table identifiers used
  - No 1-1 relationship between identifiers between two different databases
  - Affects running time
  - Introduces many unverified interactions

Protein Complex Finder

- Blastp (bi-directional)

Find top BLAST alignment on human protein α, in non-human database. Call it protein a

Protein Complex Finder

- Blastp (bi-directional)

Find top BLAST alignment on non-human protein a, in human database. If we get back protein α, then infer homology

Protein Complex Finder

- Pre-processing step 2:
  - Used AdjustCD as according to Liu et al, 2008

NUS

Protein Complex Finder

- Algorithm:
  - Motivation:
    - In CMC, the merging operation might discard some cliques without verifying whether parts of the clique is still important
    - We hypothesize that important clusters are tightly connected components after removing non-important interactions resulting from already predicted cliques

NUS

Protein Complex Finder

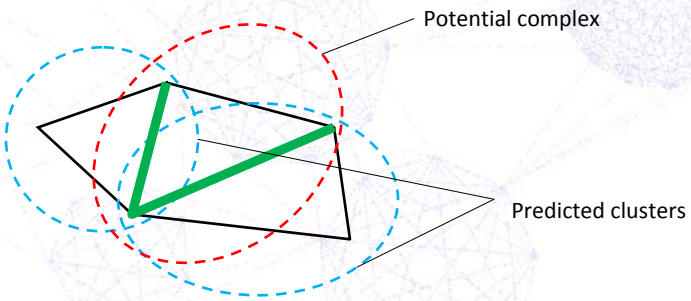
- Algorithm:
  - Key Idea:
    - Run CMC to find initial clusters of predicted proteins
    - Remove some of the interactions resulting from predicted cliques that are non-important
    - Run CMC again to find important clusters

Protein Complex Finder

- How do we define what is an important interaction?
  - Hypothesis formulation:
    - Based on observation, some proteins belong to many complexes
    - These proteins are important in that if we remove them, we might not be able to recover important cliques that were discarded
    - We call these proteins “core” proteins
    - Proteins that were belonging to many clusters in the interaction network were assumed to be such core proteins, so we will not try to remove them

Protein Complex Finder

- Core proteins:
  - E.g.

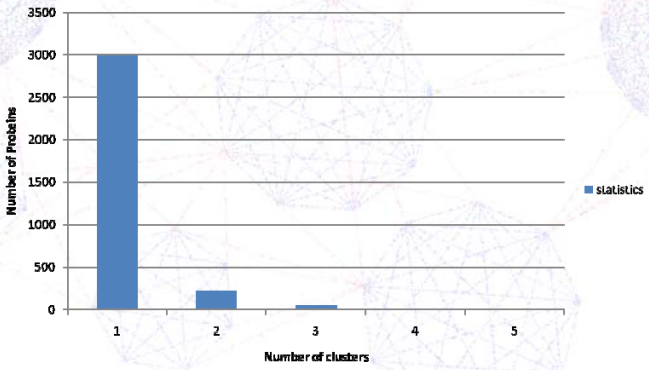


Protein Complex Finder

- How many is many?
  - For each protein we find how many clusters it belongs to
  - Let X be the number of clusters a protein belong to
  - We find X_μ and X_σ, core proteins are those such that $X > X_\mu + X_\sigma$

Protein Complex Finder

- How many is many?



Protein Complex Finder

- We find that PCF generally returns a lot more predicted complexes
- Possible post-processing step:
  - For each cluster
    - For each protein
      - Find common annotations that are relevant
      - If many share the same annotation above a certain threshold
        - » Keep that prediction
      - Otherwise discard it

Results

- We obtained human interaction information from BioGRID
- Validation data (real human complexes) from MIPS
- We compared MCL, CMC and PCF
- MCL performs very badly
- For CMC and PCF, we tried a combination of pre-processing techniques
  - Orignial Human + AdjustCD
  - Human merged with mouse + AdjustCD

NUS

Results

- Validation criteria:
  - A hit is defined by the Jaccard Co-efficient

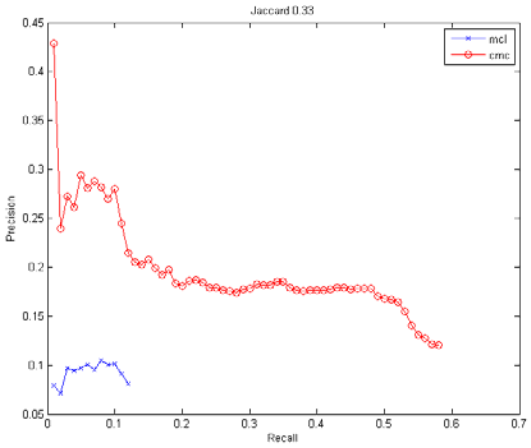
$$\frac{|Vs \cap Vc|}{|Vs \cup Vc|}$$

  - Vs – predicted cluster
  - Vc – real complex

NUS

Results

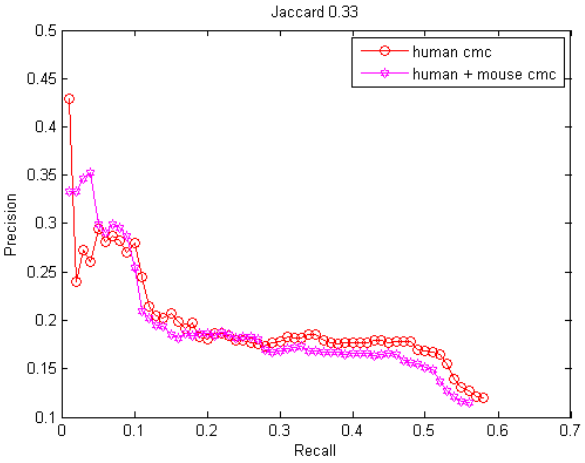
- Comparing MCL with CMC:



NUS

Results

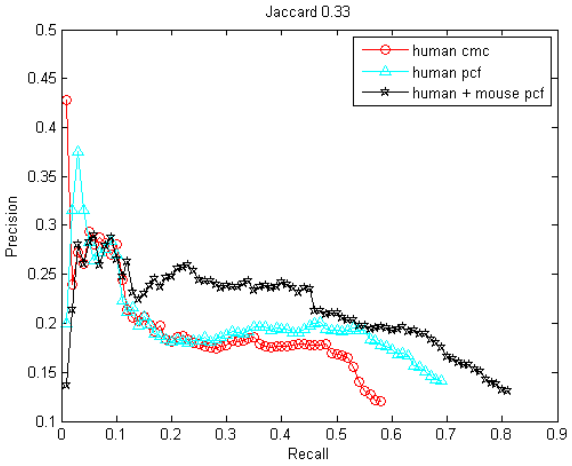
- Comparing CMC using original and mouse merged networks



NUS

Results

- Comparing CMC and PCF:



NUS

Results

- Results:
  - We also found that PCF predict much more complexes (1877) than real complexes (289), this results in low precision
  - One may ask if it is possible to randomly select 1877 and get hit 80% of real complexes?
  - We show that it is unlikely
    - Suppose probability of choosing one real complex randomly in a network with 5000 proteins is given by

$$\frac{n}{\sum_{i=2}^{m}\binom{5000}{i}}$$

    - Where n is the number of real complexes and m is the maximum size of a complex. The expected number of real complexes when we select 1877 times, is very small. If m=4, expected real complexes is $2.0*10^{-7}$

NUS

Future Work

- Future work:
  - Classifying cliques revisited:
    - Instead of generating random negative samples, we could use false positives generated from the algorithm as negative samples
    - Require robust validation methods

NUS

Future Work

- Future work:
  - The success of PCF demonstrates the possibility of doing iterated removal and detection techniques
  - We can also experiment different combination of thresholds for each iteration
    - E.g. If we feel that some complexes are not going to be tight clusters, we can modify the thresholds in the second iteration accordingly

NUS

Conclusion

- Conclusion:
  - What we have discussed so far:
    - Related works and their limitations
    - Possible approaches and their limitations
    - Motivation towards PCF and how it PCF works
    - Results show that PCF improves coverage on real complexes
    - Potentials in future work

NUS

Questions and Answer

Thank you for listening

NUS