Supervised community detection in protein-interaction networks

Meghana Palukuri[1], Edward Marcotte[1,2]

[1]Oden Institute for Computational Engineering and Sciences, [2]Institute for Cellular and Molecular Biology, College of Natural Sciences, The University of Texas at Austin

Introduction

Community detection, or finding sets of 'similar' nodes interacting with each other in networks is an interesting problem, with several applications such as finding:

- ❖ Groups of people in **social networks** like Twitter & Facebook.
- ❖ Functional units in **biological networks** such as:
  - Gene communities in gene regulatory networks
  - **Protein Complexes** in protein-interaction networks

Protein complexes take part in many cellular functions and their identification will help understand mechanisms of disease.

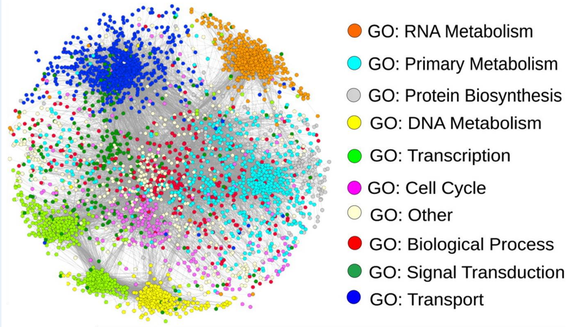

Fig.1: Yeast interactome visualized by John Morris & Alex Pico

- **STATE-OF-THE-ART**: Unsupervised graph clustering algorithms
  - Assumption: Communities are dense subgraphs in a network.
- But, communities have different topologies – they are not necessarily dense !
- In several applications, we have data on known communities – both their members as well as their internal connectivity.

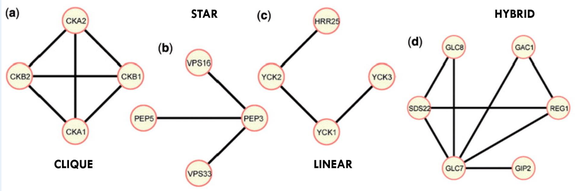

Fig.2: Different topologies of complexes in the yeast network [1].

The idea is to use the topology of known communities to identify new communities in networks - a supervised approach.

Research Workflow

1. **Construct** a streamlined, plug-and-play pipeline to predict communities in networks, using known community information.
2. **Compare** and benchmark against state-of-the art methods.
3. **Implement** the pipeline to predict protein complexes in human and yeast protein-protein interaction networks (PPINs):

Table 1: Details of experiments conducted

	Experiment1	Experiment2
Dataset (PPIN)	Human - hu.MAP [2]	Yeast - DIP [4]
No. of nodes	7778	4933
No. of edges	56,712	22,274
Complexes	Human - CORUM [3]	Yeast - MIPS, TAP-MS [1]
No. of complexes	429	268

Super.Complex – The Pipeline

Data Prep
- Positive complexes: Train - test splits
- Negative complexes: Random walks on the network.

Feature Extraction

18 topological features:
- No. of nodes
- Density
- Degree statistics
- Clustering coefficients

ML pipeline

AutoML: tpot [5] using genetic algorithm to find:
- Scaler
- Feature selector
- ML model

Subgraph sampling
- Greedy
- Metropolis
- Iterative Simulated Annealing (ISA)
- Cliques

SUBGRAPH SAMPLING

- An NP hard problem !
- The parameter p enables exploration, accepting lower scored complexes to enable better scores later.

Variants:
- **Greedy** : p = 1
- **Metropolis**: p = k, where 0 < k < 1
- **ISA**: 1-p = $e^{\frac{Score_{current} - Score_{old}}{T}}$ where, T = $\frac{T_{old}}{\alpha}$, starting with T_0

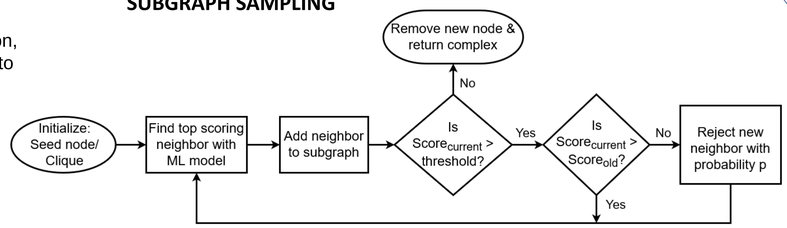

Fig.3: The sampling algorithm. Score - ML model score for the subgraph to be a complex

Classification Results

Data Prep: Ensuring equal train-test size distributions:

1. Make a say, 90-10 (parameter initSplit) random train-test split.
2. Transfer, say 30% (parameter trainTransfer) of smaller train complexes to the test complexes.
3. Iteratively transfer any train complex sharing one or more edges with any of the test complexes to test complex set.
4. Optimize initSplit and trainTransfer till equal size distributions.

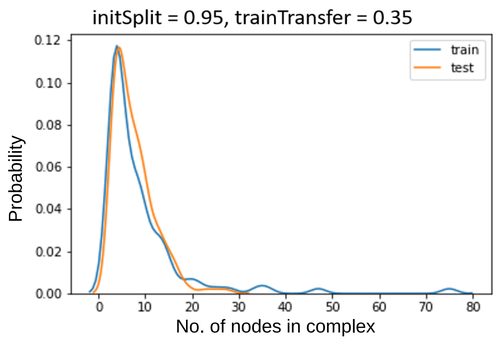

Fig. 4: Final train-test size distributions for hu.MAP

TPOT Results:

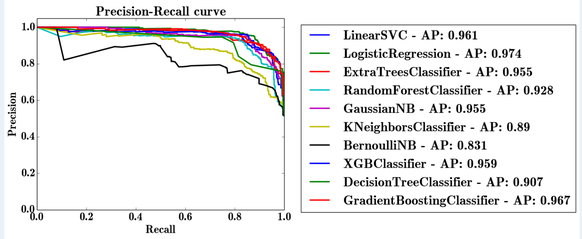

Fig. 5: Binary classification Precision Recall curve for hu.MAP

Best classifier: Logistic Regression

Prediction Results

$$\text{Precision} = \frac{\text{No. of predicted complexes recovering known complex}}{\text{No. of predicted complexes}}$$

$$\text{Recall} = \frac{\text{No. of recovered known complexes}}{\text{No. of known complexes}}$$

A predicted complex recovers a known complex if,

$$\frac{C}{A + C} > p \text{ and } \frac{C}{B + C} > p$$

where, p is an input threshold parameter between 0 and 1 and

A/B/C - Number of proteins only in predicted/known/both complex

Table 2: Yeast prediction metrics, compared with state-of-the-art SCI-SVM and SCI-BN [1]

Method	Precision	Recall	F1 score
Greedy	**0.313**	0.381	0.344
Metropolis	0.288	0.41	**0.339**
ISA	0.287	0.414	0.339
Cliques	0.185	**0.456**	0.264
SCI-SVM	0.176	0.379	0.24
SCI-BN	0.219	0.537	0.312

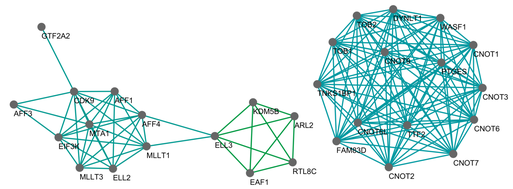

Fig. 6: Example predicted complexes in hu.MAP

- On the left - the algorithm (metropolis) distinguishes two complexes that are connected by one node.
- Left complex: AFF4 super elongation complex which could be a key regulator in pathogenesis of leukemia, Middle: unknown
- Right: CCR4-NOT transcription complex [6], called the 'control-freak of eukaryotic cells'.
- **~1500 complexes predicted in hu.MAP**

Conclusion

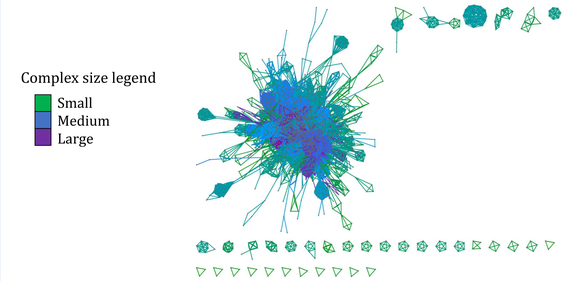

Fig.7: Human protein interaction network colored by complex size

- Demonstrated the **potential of supervised ML strategies** through a streamlined pipeline - **Super.Complex**
- Major steps: feature extraction, an auto-ML pipeline and a subgraph sampling process with a choice between different algorithms.
- **Logistic regression** and the **Metropolis sampling algorithm** tend to perform well.
- Comparable and slightly **better results than some state-of-the-art algorithms**.
- This plug-and-play pipeline can be improved by further optimizing each part, ex: using biological features.
- Performed **experiments on real protein networks** of human and yeast: Results can be examined to glean biological insights.

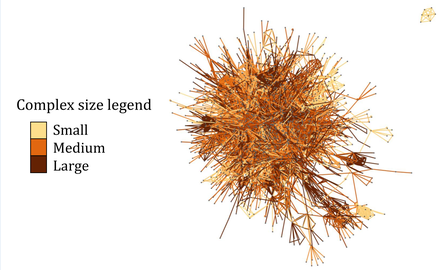

Fig. 8: Yeast protein interaction network DIP colored by complex size

References

1. Qi, Yanjun, et al. (2008) Protein complex identification by supervised graph local clustering.*Bioinformatics* 24(13), 250-268.
2. Drew, K., et al. (2017). Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular systems biology*, 13(6):932.
3. Giurgiu, M., et al. (2018). CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Research*.
4. Xenarios, I., et al. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, *30*(1), 303-305.
5. https://github.com/EpistasisLab/tpot
6. Miller, J. E., et al. (2012). Ccr4-Not complex: the control freak of eukaryotic cells. *Critical reviews in biochemistry and molecular biology*, *47*(4), 315-333.

The University of Texas at Austin
Oden Institute for Computational Engineering and Sciences