Link Prediction in PPI Yeast Network

Puja Gupta
Data and Computational Science
Indian Institute of Technology,Jodhpur
Jodhpur, India
gupta.92@iitj.ac.in

Shaonli Pal
Data and Computational Science
Indian Institute of Technology,Jodhpur
Jodhpur,India
pal.14@iitj.ac.in

***Abstract*—We are doing protein link prediction for the PPI yeast network using three different algorithms, i.e. Preferential Attachment(PAC), Adamic Adar Coefficient(AAC),Jaccard Coefficient(JAC).**

***Index Terms*—Link Prediction, protein node, degree, PAC, AAC, JAC**

I. Introduction

The task of predicting the function and interaction of proteins in a protein-protein dataset has become very important and crucial in recent times.Various research are going in this field to extract the maximal amount of information from the interaction of proteins in the biological network, which helps to predict the function of protein. In our project we have used the Protein Protein interaction dataset of yeast for the link prediction of top 100 protein nodes having the highest interaction scores.

II. Method

We have computed the link prediction for a given protein node based on following three scores: PAC Score, AAC Score, JAC Score. The working of these algorithms are demonstrated by using a small sub graph where each node represents one protein and the edges represent the interaction between two proteins..

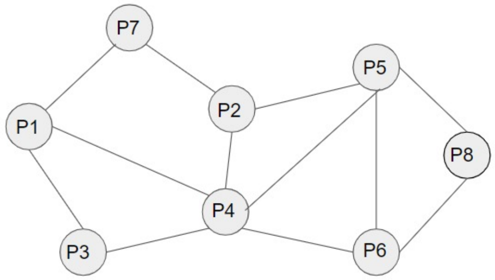

Fig. 1. Sub graph of Protein Protein Interaction

A. Preferential Attachment(PAC)

We have created a dictionary named degree_of_nodes with protein node as key and degree as its value. For each pair of non connected protein edges (Pi,Pj), the PAC score is computed using the formulae
PAC Score (Pi,Pj) = $|N(i)| * |N(j)|$

TABLE I
Protein nodes and their degree

Nodes	Degree
P1	3
P2	3
P3	2
P4	5
P5	4
P6	3
P7	2
P8	2

where N (i) is the degree of node Pi and N(j) is the degree of node Pj. For each node, the edge which gives the highest PAC score is selected as the predicted protein link for that given node. We have predicted the links of the top 100 such nodes having the highest PAC scores.

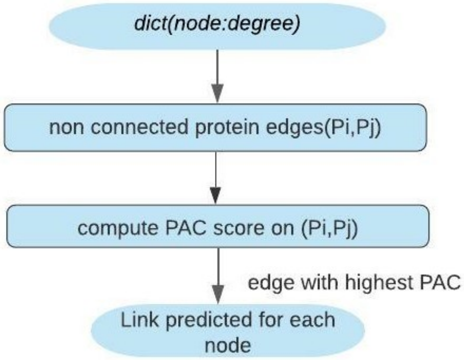
dict(node:degree)
non connected protein edges(Pi,Pj)
compute PAC score on (Pi,Pj)
edge with highest PAC
Link predicted for each node

Fig. 2. Workflow of PAC algorithm

For the given example, based on calculated PAC Score, the protein pair with highest score is P1 and P5.

B. Adamic Adar Coefficient(AAC)

We have calculated the neighbouring edges for all the protein nodes and computed the value of 1/log(degree) for each of the neighbours. For each pair of non connected protein edges (Pi,Pj), the AAC score is computed using the

TABLE II
PAC SCORE OF NOT CONNECTED PROTEIN EDGES

Not connected node pairs (i,j)	PAC (i,j)
P1,P2	3*3=9
P1,P5	3*4=12
P1,P6	3*3=9
P1,P8	3*2=6
P2,P3	3*2=6
P2,P6	3*3=9
P2,P8	3*2=6
P3,P5	2*4=8
P3,P6	2*3=6
P3,P7	2*2=4
P3,P8	2*2=4
P4,P7	5*2=10
P4,P8	5*2=10
P5,P7	4*2=8
P6,P7	3*2=6

TABLE III
LINK PREDICTIONS BASED ON PAC SCORE

Link Predictions based on PAC value)	PAC Value
P1,P5	12
P2,P6	9
P3,P5	8
P4,P8	10
P5,P7	8
P6,P7	6

formulae

$$AAC(Pi,Pj) = \sum_{k \in N(i) \cap N(j)} \frac{1}{log|N(k)|}$$

where N(i)$\cap N(j)$

represents the common neighbours of Pi and Pj. For each node, the edge which gives the highest AAC score is selected as the predicted protein link for that given node. We have predicted the links of the top 100 such nodes having the highest AAC scores.

TABLE IV
PROTEIN NODES AND THEIR DEGREE

Nodes	Neighbours	log(degree	1/log(degree)
P1	P3,P4,P7	0.4771	2.0959
P2	P4,P5,P7	0.4771	2.0959
P3	P1,P4	0.3010	3.3222
P4	P1,P2,P3,P5,P6	0.6989	1.4308
P5	P2,P6,P8	0.4771	2.0959
P6	P4,P5,P8	0.4771	2.0959
P7	P1,P2	0.3010	3.3222
P8	P5,P6	0.3010	3.3222

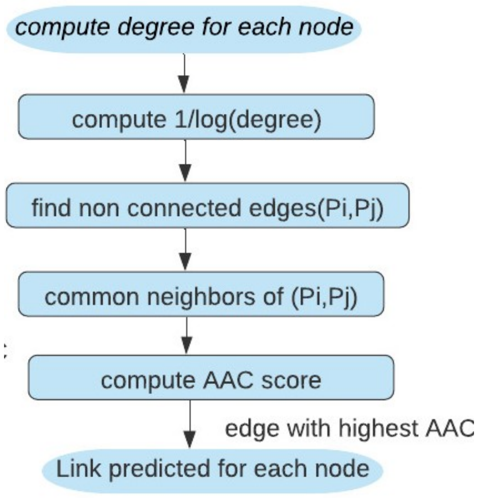
compute degree for each node
compute 1/log(degree)
find non connected edges(Pi,Pj)
common neighbors of (Pi,Pj)
compute AAC score
edge with highest AAC
Link predicted for each node

Fig. 3. Workflow of AAC algorithm

TABLE V
PAC:PROTEIN NODES AND THEIR DEGREE

Not connected node pairs (i,j)	Common Neighbours	AAC(i,j)
P1,P2	P4,P7	4.7530
P1,P5	{}	0.0
P1,P6	P4	1.4308
P1,P8	{}	0.0
P2,P3	P4	1.4308
P2,P6	P4,P5	3.5267
P2,P8	P5	2.0959
P3,P5	{}	0.0
P3,P6	P4	1.4308
P3,P7	P1	2.0959
P3,P8	{}	0.0
P4,P7	P1,P2	4.1180
P4,P8	P5,P6	4.1180
P5,P7	P2	2.0959
P6,P7	{}	0.0

C. *Jaccard Coefficient(JAC)*

For each pair of non connected protein nodes (Pi,Pj) we calculated the count of intersection and union of their common neighbours. The JAC Score for each non connected protein node pair is given by the formulae

$$JAC(i,j) = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}$$

The JAC score will always lie between a range of 0 to 1. For each node, the edge which gives the highest JAC score is selected as the predicted protein link for that given node. We have predicted the links of the top 100 such nodes having the highest JAC scores.

III. RESULT

The link prediction based on the highest PAC Score is between protein node ”YGL122C” and ”YDL160C” with a PAC score of 9808398. The link prediction based on the highest AAC Score is between protein node ”YLR039C” and ”YDL160C” with a AAC score of 311.842. The link prediction based on the highest JAC Score is between protein node ”YER111C” and ”YDR334W” with a JAC score of 0.2526.

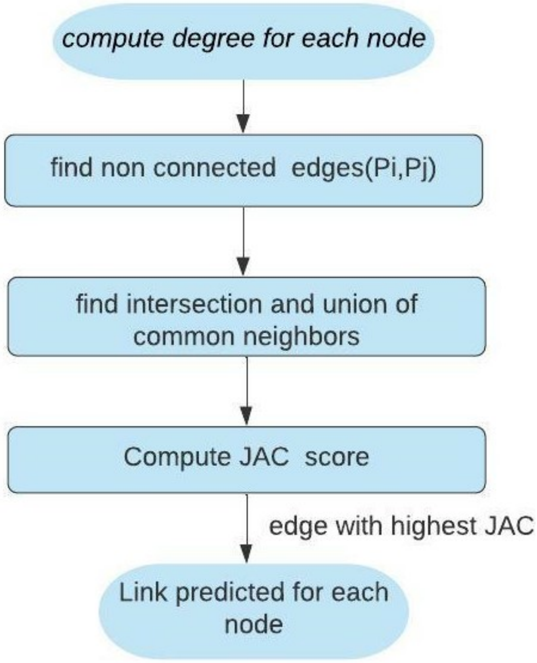
compute degree for each node
find non connected edges(Pi,Pj)
find intersection and union of common neighbors
Compute JAC score
edge with highest JAC
Link predicted for each node

Fig. 4. Workflow of JAC algorithm

TABLE VI
JAC: Protein and their neighbouring nodes.

Protein node	Neighbouring nodes
P1	P3,P4,P7
P2	P4,P5,P7
P3	P1,P4
P4	P1,P2,P3,P5,P6
P5	P2,P6,P8
P6	P4,P5,P8
P7	P1,P2
P8	P5,P6

TABLE VII
JAC Value of not connected protein node pairs

Not connected node pairs (i,j)	Intersection	Union	JAC(i,j)
P1,P2	P4,P7	P3,P4,P5,P7	2/4=0.5
P1,P5	{}	P2,P3,P4,P6,P7,P8	0/6=0
P1,P6	P4	P3,P4,P5,P7,P8	1/5=0.2
P1,P8	{}	P3,P4,P5,P6,P7	0/5=0
P2,P3	P4	P1,P4,P5,P7	1/4=0.25
P2,P6	P4,P5	P4,P5,P7,P8	2/4=0.5
P2,P8	P5	P4,P5,P6,P7	1/4=0.25
P3,P5	{}	P1,P2,P4,P6,P8	0/5=0
P3,P6	P4	P1,P4,P5,P8	1/4=0.25
P3,P7	P1	P1,P2,P4	1/3=0.33
P3,P8	{}	P1,P4,P5,P6	0/4=0
P4,P7	P1,P2	P1,P2,P3,P5,P6	2/5=0.4
P4,P8	P5,P6	P1,P2,P3,P5,P6	2/5=0.4
P5,P7	P2	P1,P2,P6,P8	1/4=0.25
P6,P7	{}	P1,P2,P4,P5,P8	0/5=0

TABLE VIII
Next Link Prediction based on different algorithms

Node1	Node2 (PAC)	Node2 (AAC)	Node3 (JAC)
P1	P5	P2	P2

TABLE IX
Link Prediction based on PAC Score

Protein Node1	Linked Node	PAC Score
YGL122C	YDL160C	9808398
YAL021C	YBR245C	5860132
YNL209W	YGL122C	5649724
YJR076C	YDL160C	4482702

TABLE X
Link Prediction based on AAC Score

Protein Node1	Linked Node	AAC Score
YDL160C	YLR039C	311.842
YAL021C	YLR039C	266.805
YFL039C	YLR418C	241.338
YNL209W	YLR418C	188.204

TABLE XI
Link Prediction based on JAC Score

Protein Node1	Linked Node	JAC Score
YER111C	YDR334W	0.2526
YPL017C	YLR381W	0.2478
YGL127C	YPL129W	0.2390
YOR141C	YHR090C	0.2371

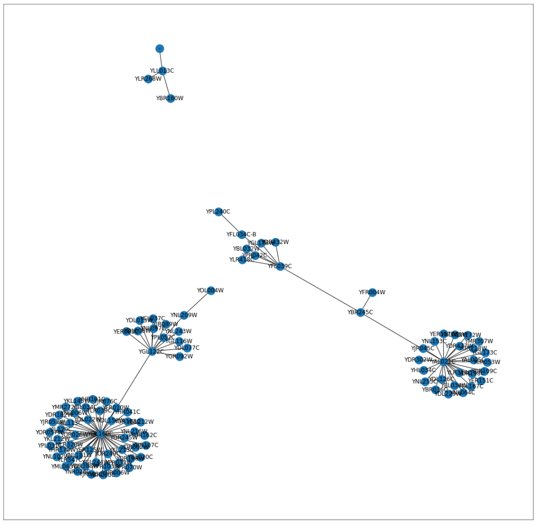

Fig. 5. Link predicted based on PAC Score

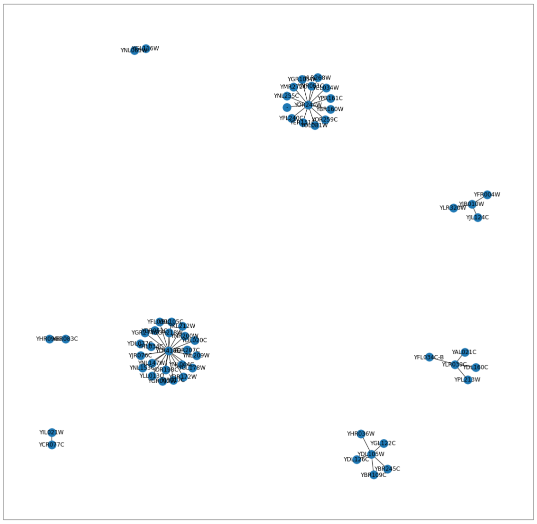

Fig. 6. Link Predicted based on AAC Score

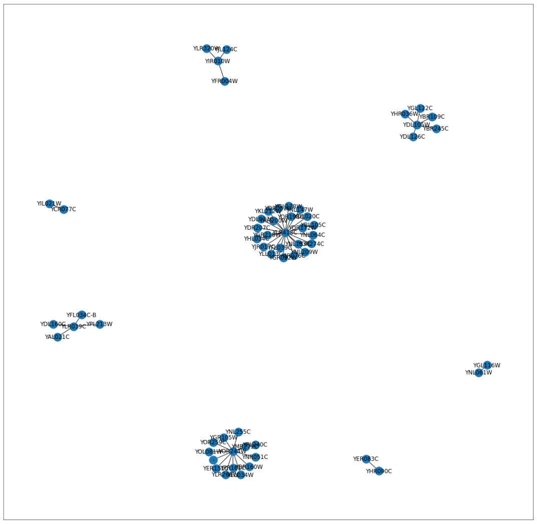

Fig. 7. Link Predicted based on JAC Score

IV. Future work

We are using the first 100 non connected protein nodes for determining the scores of each protein due to resource issue, however we can increase our search space to the entire Adjacency matrix for that particular node.

Also, we are statically determining the next protein node that should be linked i.e. after predicting the link for one node we are not adding that edge to the original graph. However we can optimise our result by dynamically adding each predicted link in our main graph.

V. Conclusion

We observed that the protein link prediction based on the different algorithms may differ from each other. Also while plotting the graph of the best 50 predicted protein link interactions, we can conclude the the proteins at the centre of each disconnected components might be one of the important proteins to be interacted with other nodes in future.

Acknowledgment

With a deep sense of gratitude, we wish to express sincere thanks to our guide, Dr. Gaurav Harit for giving us the opportunity to work on this project. We truly appreciate and value his esteemed guidance and encouragement from the beginning to the end of the project.

References

[1] https://www.kaggle.com/alexandervc/yeast-proteinprotein-interaction-network

[2] https://www.hindawi.com/journals/sp/2015/172879/

[3] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0060372

[4] https://neo4j.com/docs/graph-data-science/current/alpha-algorithms/preferential-attachment/