Super.Complex v3.0: A Supervised Machine Learning Pipeline for Molecular Complex Detection in Protein-interaction Networks

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

[1]Oden Institute for Computational Engineering and Sciences; [2]Molecular Biosciences, University of Texas at Austin

Motivation

Can we learn features of known complexes and use them to predict new complexes? Existing unsupervised methods do not use this information and existing supervised methods are limited in accuracy and scalability.

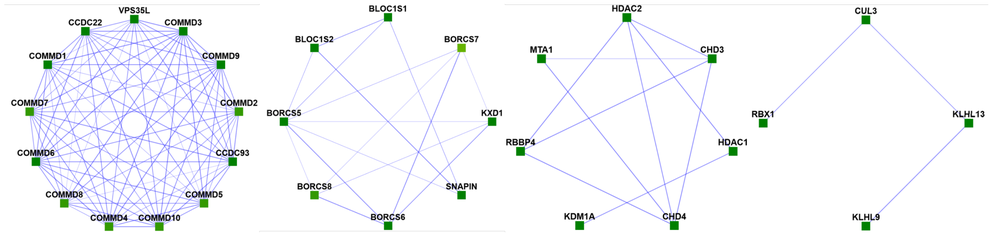


Figure 1. Human protein complexes from CORUM[2] with protein-interactions from hu.MAP[1] have different topologies that can be learned: Left to right: (i)Clique, (ii)Hybrid with different edge-weights, (iii)Hybrid, (iv)Linear

Methods

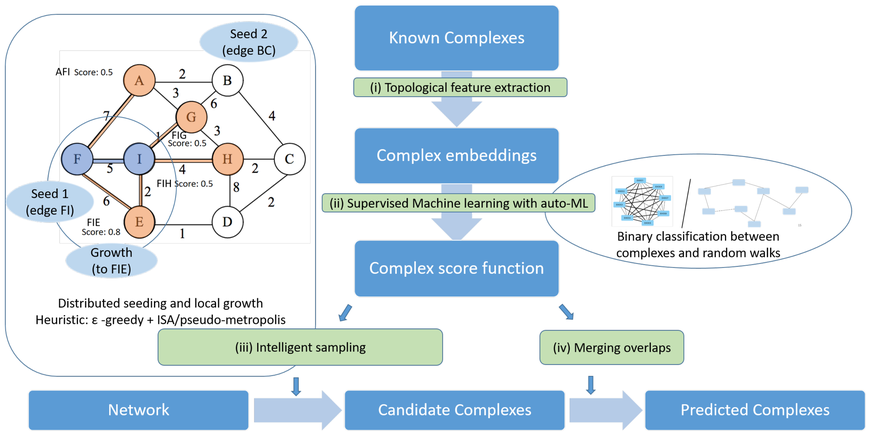
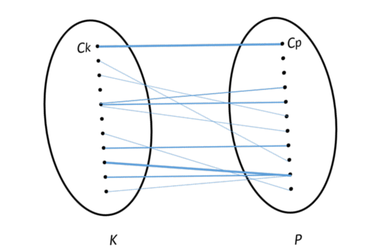


Figure 2. Overview of Super.Complex, a distributed supervised auto-ML method: (i) Topological features are extracted from known complexes to build complex embeddings (feature vectors, which are representations of complexes in vector space) (ii) A score function for complexes is learned from the complex embeddings, as the decision function to classify a subgraph as a complex or a random walk. The best score function is selected after training multiple machine learning models with tpot [4], an auto-ML pipeline. (iii) Multiple complexes are sampled in parallel from the network. To build each candidate complex, a seed edge is selected and grown using a 2-stage heuristic. First, we use an epsilon-greedy heuristic to select a candidate neighbor and then we use a pseudo-metropolis (constant probability) or iterative simulated annealing heuristic to accept or reject the candidate neighbor for growing the current complex. (iv) The candidate complexes are merged such that the maximum overlap between any 2 complexes is not greater than a set threshold.



2. Unbiased Accuracy:

$$Acc' = \sqrt{PPV' * Sn'}$$

where,

$$PPV' = \frac{\sum_{C_p \epsilon P} \max_{C_k \epsilon K} p'(C_p, C_k))}{|P|}$$

$$Sn' = \frac{\sum_{C_k \epsilon K} \max_{C_p \epsilon P} r'(C_p, C_k))}{|K|}$$

1. Net F1 score:

$$\frac{1}{F_1} = \frac{1}{p} + \frac{1}{r}$$

where,

$$p = \frac{\sum_{C_p \epsilon P} \max_{C_k \epsilon K} sim_{F_1}(C_p, C_k)}{|P|}$$

$$r = \frac{\sum_{C_k \epsilon K} \max_{C_p \epsilon P} sim_{F_1}(C_p, C_k)}{|K|}$$

where,

$$sim_{F_1} = F_1' \quad \text{and} \quad \frac{1}{F_1'} = \frac{1}{p'} + \frac{1}{r'}$$

where, $p' = \frac{|C_p \cap C_k|}{|C_p'|}$ $r' = \frac{|C_p \cap C_k|}{|C_k|}$

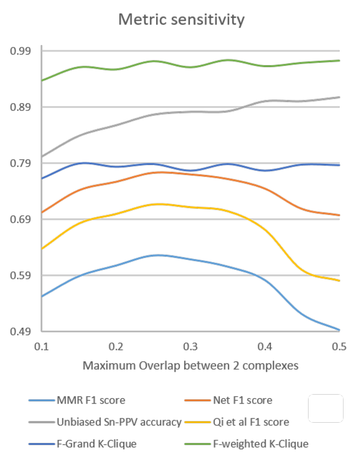


Figure 3. Proposed evaluation measures: net F1 score, unbiased Sn-PPV accuracy, MMR F1 (combines recall MMR with precision equivalent). Sensitivity of different evaluation measures w.r.t overlap between complexes.

Results

Super.Complex predicts 1028 complexes[5], with 146 interacting with SARS-COV2 proteins[3], trained on cleaned CORUM complexes on hu.MAP (7778 proteins). From figure 3, proposed net F1 and MMR F1, as well as Qi et al's F1 are sensitive to the overlap of complexes, and we recommend them for evaluation. From table 1, Super.Complex is competitive, and also perfectly recalls 59 CORUM complexes.

Table 1. Evaluating predicted complexes on hu.MAP w.r.t 188 CORUM complexes

Method	MMR Precision	MMR Recall	MMR F1 score	Net F1 score	Unbiased Sn-PPV accuracy	Qi et al [6] F1score	F-Grand k-Clique	F-weighted k-Clique
Super.Complex	**0.767**	0.534	**0.63**	0.783	0.888	0.739	**0.785**	**0.972**
ClusterOne + MCL	0.471	**0.686**	0.559	**0.797**	**0.911**	**0.764**	0.77	0.967

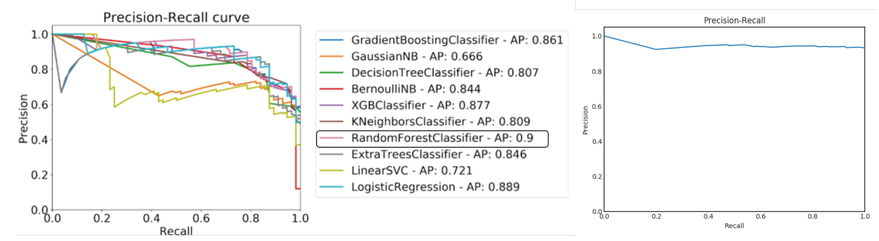


Figure 4. Classification PR curves for: Left- complexes, Right- co-complex edges

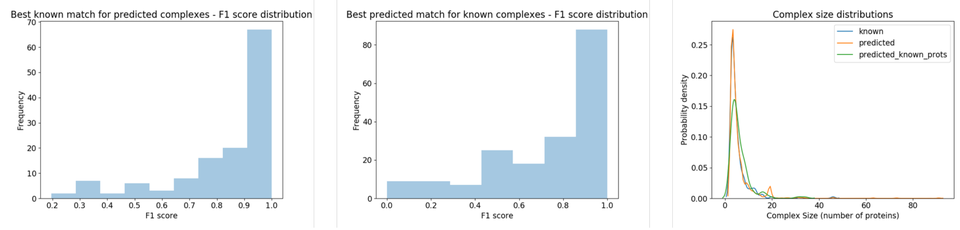


Figure 5. F1 scores of individual matches, and complex size distributions

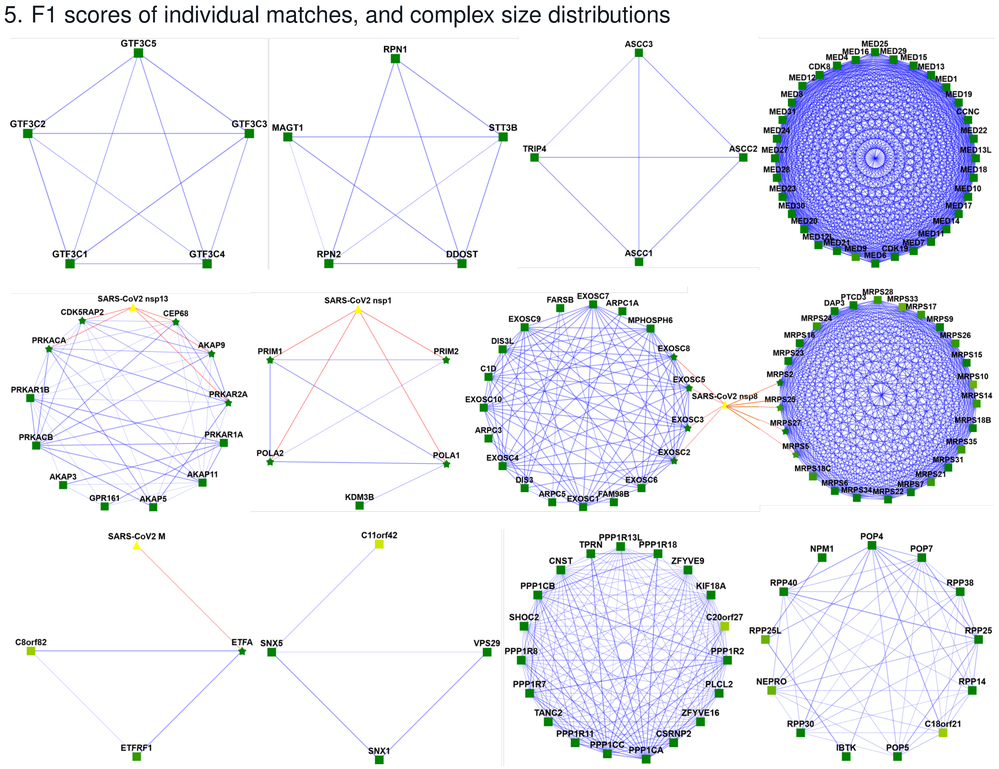


Figure 6. Row 1: Some recalled CORUM complexes. Row 2: Some predicted complexes interacting with SARS-COV2 proteins. Row 3: Some predicted complexes with proteins having low annotation scores (light green)

References

[1] Kevin Drew et al. "Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes". eng. In: *Mol. Syst. Biol.* 13.6 (2017), p. 932. ISSN: 1744-4292. DOI: 10.15252/msb.20167490.

[2] Madalina Giurgiu et al. "CORUM: the comprehensive resource of mammalian protein complexes-2019". eng. In: *Nucleic Acids Res.* 47.D1 (Jan. 2019), pp. D559–D563. ISSN: 1362-4962. DOI: 10.1093/nar/gky973.

[3] David E. Gordon et al. "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing". en. In: *Nature* 583.7816 (July 2020), pp. 459–468. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2286-9.

[4] Randy Olson et al. *EpistasisLab/tpot: v0.10.1 minor release*. Apr. 2019. DOI: 10.5281/zenodo.2647523.

[5] Meghana Palukuri and Edward Marcotte. *Super.Complex*. URL: https://sites.google.com/view/supercomplex/home.

[6] Y. Qi et al. "Protein complex identification by supervised graph local clustering". en. In: *Bioinformatics* 24.13 (July 2008), pp. i250–i268. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btn164.

STAMPEDE2
INSTITUTE FOR COMPUTATIONAL ENGINEERING & SCIENCES
ODEN
The Welch Foundation
The University of Texas at Austin
Center for Biomedical Research Support
Office of the Vice President for Research
NIH