Evaluation metrics

Here, we adopt twelve well-established metrics to assess the quality of predicted protein complexes, including sensitivity, positive predictive value, accuracy and separation from [1], fraction match and maximum matching ratio from [2], precision, recall, and F-measure from [3], and precision$^+$, recall$^+$, and F-measure$^+$ from [4]. Given a set of predicted protein complexes P and a set of reference protein complexes C, a contingency table T is assembled with n rows denoting complexes in C, and m columns representing clusters in P. The entry $t_{i,j}$ indicates the number of shared proteins between complex i and cluster j. N_i denotes the number of proteins in complex i, $N = \sum_i N_i$, and P_j stands for the number of proteins in cluster j. For approaches that predict overlapping clusters, the marginal row, and column sums may not correspond to the size of reference and predicted complexes, respectively. The positive predictive value (PPV), sensitivity (SN), accuracy (ACC), and separation (SEP) are defined as:

$$PPV = \frac{\sum_j \max_i(t_{i,j})}{\sum_j \sum_i t_{i,j}}, \tag{1}$$

$$SN = \frac{\sum_i \max_j(t_{i,j})}{\sum_i t_{i,j}}, \tag{2}$$

$$ACC = \sqrt{PPV \times SN}, \tag{3}$$

$$SEP = \sqrt{\frac{1}{nm}\sum_i \sum_j (\frac{t_{i,j}}{t_{.j}} \times \frac{t_{i,j}}{t_{.i}}) \times \sum_j \sum_i (\frac{t_{i,j}}{t_{.j}} \times \frac{t_{i,j}}{t_{.i}})}. \tag{4}$$

Accuracy considers both SN and PPV, and it shows the overall performance. Furthermore, separation is a product of the proportion of proteins of a reference complex that are found in a predicted complex and the proportion of proteins in a predicted complex that is in a reference complex, to quantify a relationship between predicted and reference complexes.

The fraction match (FRM) first calculates the overlap score between reference and predicted complex, i and j, in which the overlap score (OS) is given by

$$OS = \frac{t_{i,j}^2}{N_i P_j}. \tag{5}$$

Then we considered the fraction of predicted complexes whose OS is higher than 0.25, as suggested by [2]. The maximum matching ratio (MMR) is given by the value of the maximum matching per complex in a bipartite graph, with vertices corresponding to the reference and predicted complexes, as two partitions. In this graph, the edges are weighted by overlap score between respective reference and predicted complexes.

To calculate precision and recall, we first determine if any of the predicted complexes matches with any of the reference complexes. Following [3, 5, 6], we employed Jaccard similarity (i.e.

$Jaccard(P, C) = \frac{|P \cap C|}{|P \cup C|}$). Thereby, the predicted complex matches the reference complex if their Jaccard similarity is higher than 0.5. Hence, precision, recall, and F-measure are defined as:

$$Precision = \frac{\left|p_i \in P | \exists c_j \in C, p_i matches c_j\right|}{|P|}, \tag{6}$$

$$Recall = \frac{\left|c_i \in C | \exists p_j \in P, p_j matches c_i\right|}{|C|}, \tag{7}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{8}$$

F-measure considers both precision and recall. Thereby, it illustrates the overall performance.

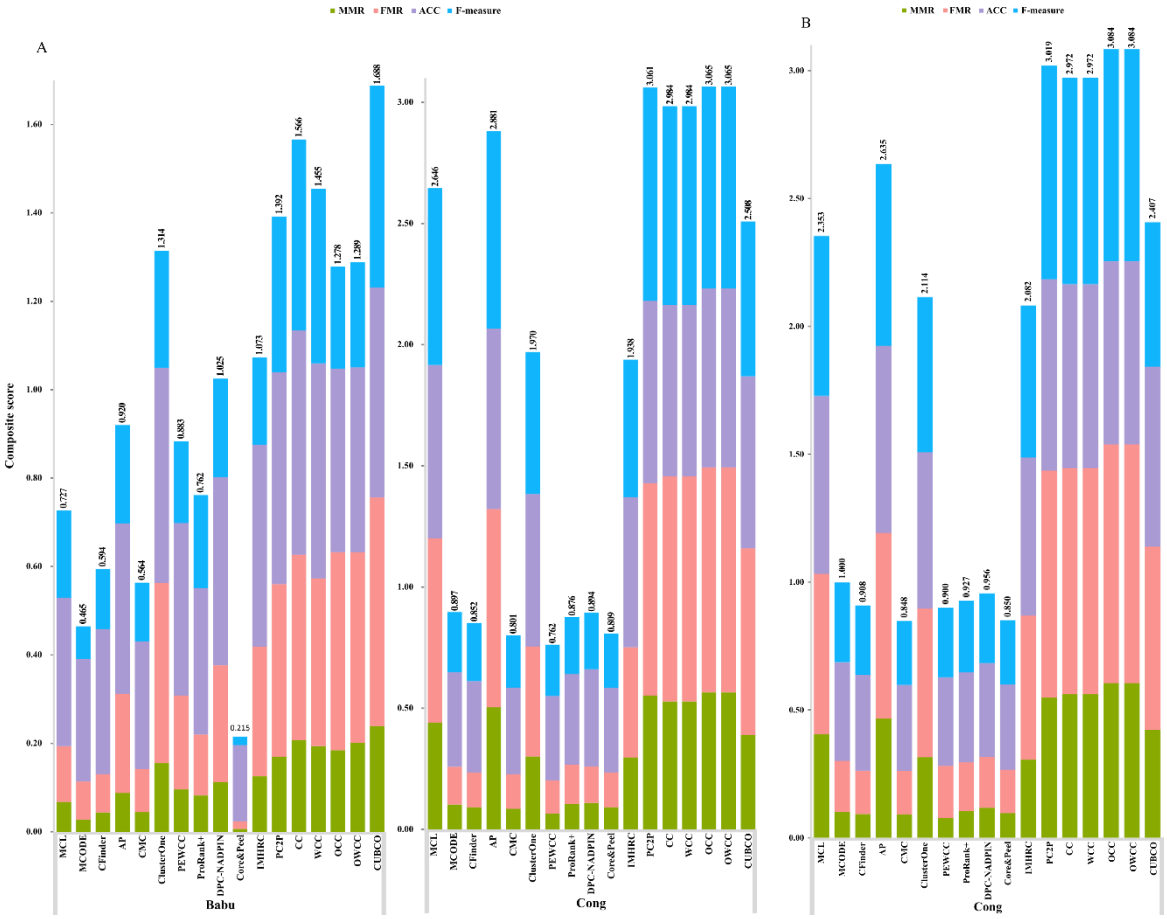
The precision$^+$ and recall$^+$ are given by $\frac{N_P^+}{|P|}$ and $\frac{N_C^+}{|C|}$, respectively. Whereby, N_P^+ and N_C^+ are defined as:

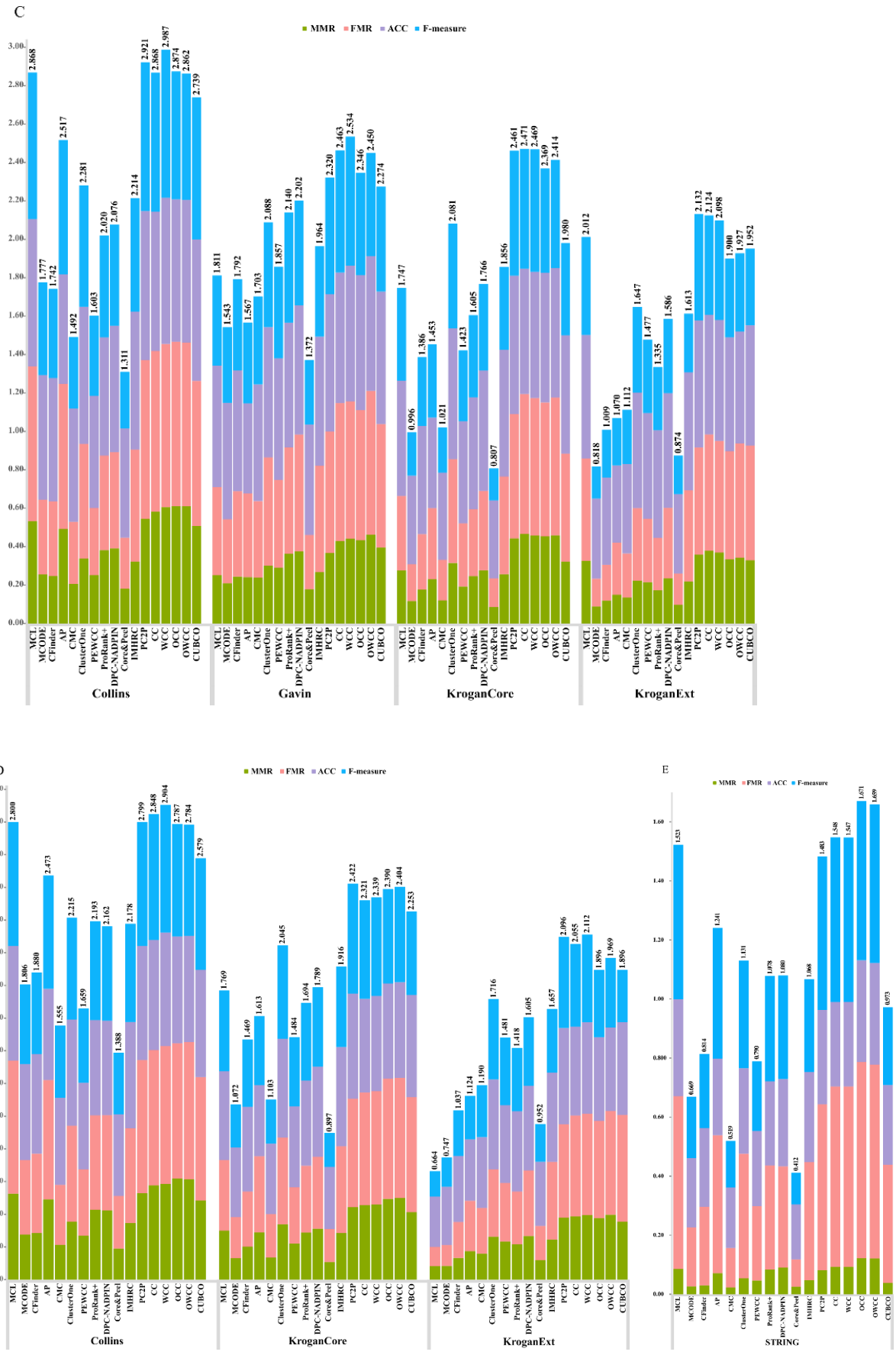
$$N_P^+ = \left|\{p_i \in P | \exists c_j \in C, OS(p_i, c_j) \geq \theta, (p_i, c_j) \in Match(P, C, \theta)\}\right|, \tag{9}$$

$$N_C^+ = \left|\{c_j \in C | \exists p_i \in P, OS(p_i, c_j) \geq \theta, (p_i, c_j) \in Match(P, C, \theta)\}\right|. \tag{10}$$

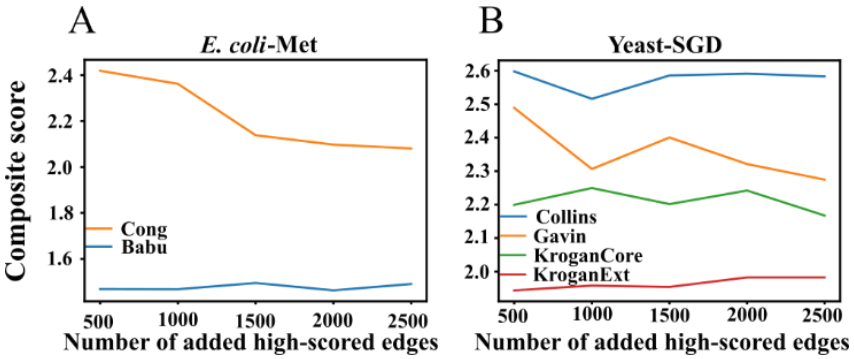
The OS computes the overlap score between reference and predicted complexes, while $Match(P, C, \theta)$ includes the edges in the induced maximum matching of the bipartite graph that has reference complexes on one side and the predicted complexes on the other side. The F-measure$^+$ is calculated the same way as the original F-measure, likewise, it shows the overall performance.

Supplementary Figure 1 Comparative analysis of approaches for prediction of protein complexes. The comparative analyses are conducted with respect to a composite score combining four performance measures, maximum matching ratio (MMR), fraction match (FRM), accuracy (ACC), and F-measure. Sixteen approaches, ordered by the year of publication, are compared on (A – B) two PPI networks of E. coli regarding to Metabolic and Ecocyc gold standards, respectively, (C – D) four PPI networks of yeast regarding CYC2008 and SGD gold standard, and (E) PIPS PPI network of H. sapiens regarding CORUM gold standard. The GCC-v outperforms all approaches on three out of four networks in *S. cerevisiae* and all networks in *E. coli* and *H. sapiens*.




C
MMR
FMR
ACC
F-measure
Collins
Gavin
KroganCore
KroganExt
D
Collins
KroganCore
KroganExt
E
STRING

Supplementary Figure 2 Composite score due to integration of link prediction by adding the most probable edges. To investigate the composite score of CUBCO with the integration of link prediction, we inserted the first 500, 1000, 1500, 2000, and 2500 edges to the original PPI networks. The composite score is calculated for the combination of PPI networks and the gold standard of (A) *E. coli* and Metabolic, (B) yeast and SGD.



References

[1] S. Brohée and J. van Helden, "Evaluation of clustering algorithms for protein-protein interaction networks," *BMC Bioinformatics,* vol. 7, 11 2006.

[2] T. Nepusz, H. Yu and A. Paccanaro, "Detecting overlapping protein complexes in protein-protein interaction networks," *Nature Methods,* vol. 9, p. 471–472, 3 2012.

[3] G. Liu, L. Wong and H. N. Chua, "Complex discovery from weighted PPI networks," *Bioinformatics,* vol. 25, p. 1891–1897, 5 2009.

[4] A. M. A. Maddi, F. A. Moughari, M. M. Balouchi and C. Eslahchi, "CDAP: An Online Package for Evaluation of Complex Detection Methods," *Scientific Reports,* vol. 9, 9 2019.

[5] M. Habibi, C. Eslahchi and L. Wong, "Protein complex prediction based on k-connected subgraphs in protein interaction network," *BMC Systems Biology,* vol. 4, p. 129, 2010.

[6] S. Srihari, C. H. Yong, A. Patil and L. Wong, "Methods for protein complex prediction and their contributions towards understanding the organisation, function and dynamics of complexes," *FEBS Letters,* vol. 589, p. 2590–2602, 4 2015.