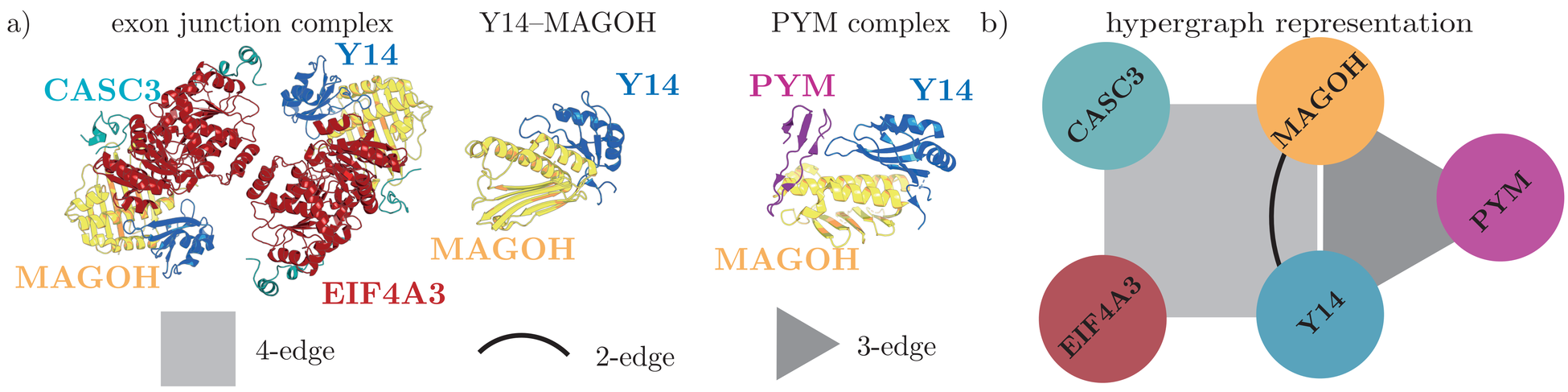
Hypergraphs for predicting essential genes using multiprotein complex data

Florian Klimm[1,2], Charlotte M. Deane[3], and Gesine Reinert[3]

[1]Department of Mathematics, Imperial College London, [2]Mitochondrial Biology Unit, University of Cambridge,
[3]Department of Statistics, University of Oxford

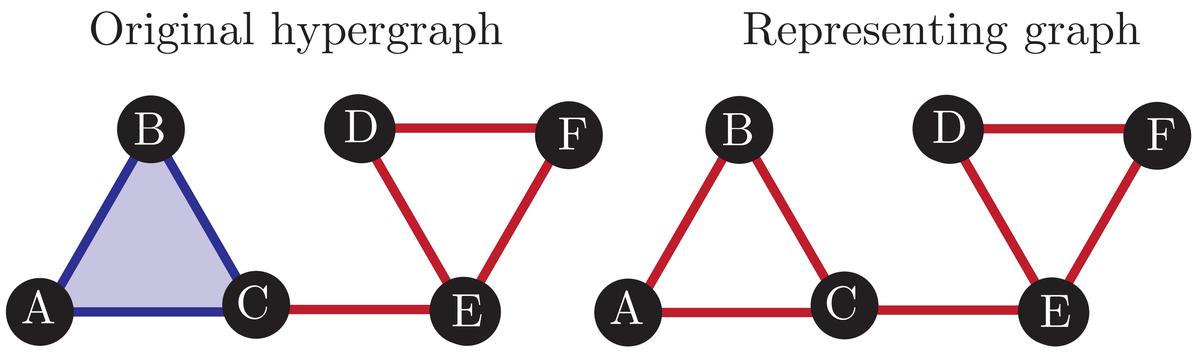
bioRxiv https://doi.org/10.1101/2020.04.03.023937 Python https://github.com/floklimm/hypergraph

Summary

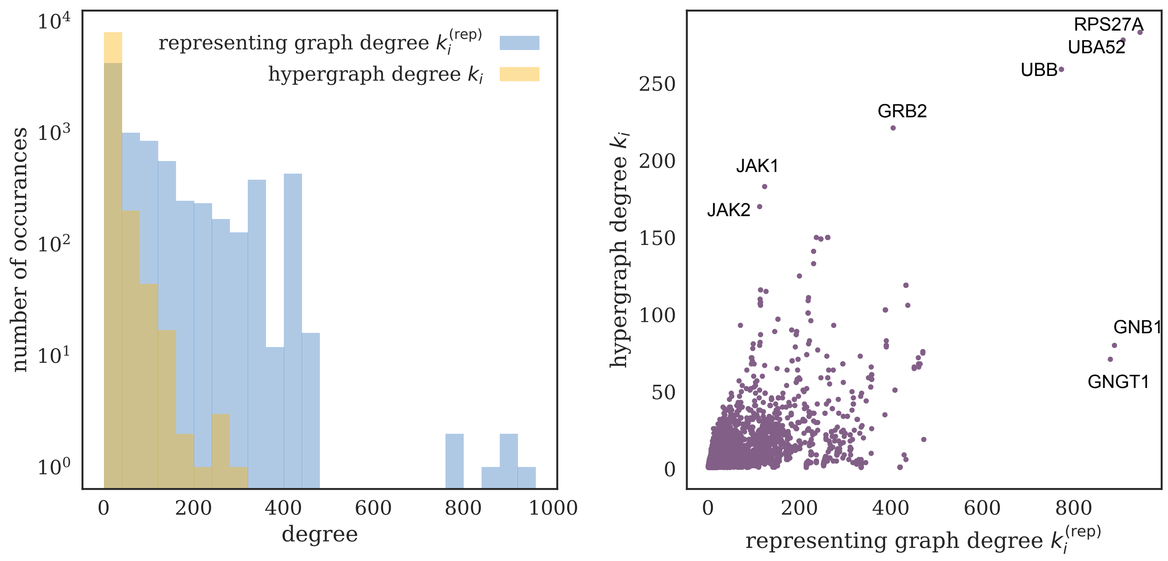


- Protein–protein interaction networks (PPINs) are abstract representations of physical interactions between proteins [1].
- Commonly, the interactions between proteins are treated as pairwise (i.e., as a graph). In most organisms, however, *multiprotein complexes*, which consists of more than two interacting proteins, fulfil important biological functions. Above, we show the *exon junction complex*, which is crucial for protein biosynthesis.
- Using data for *Homo sapiens* from Reactome [3], we construct a hypergraph with $N = 8243$ nodes (representing proteins) and $M = 6688$ hyperedges (representing multiprotein complexes).
- We demonstrate that *hypergraphs* allow a fruitful analysis of such higher-order interactions. In particular, we find that
  - **Hub proteins differ between hypergraph and graph analysis of this data,**
  - **Hypergraph-degree is a better predictor of gene-essentiality than graph-degree, and**
  - **Under random null models, some hypergraph properties (e.g., mean local clustering) are preserved, whereas others (e.g., assortativity) are not.**

Hub nodes differ in graph and hypergraph



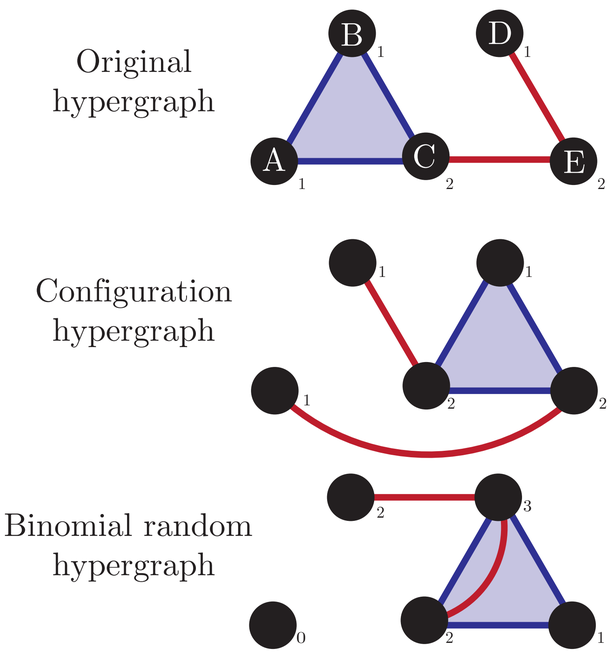
We denote a hypergraph as a pair $\mathcal{H} = (V, E)$, where V is the node set and the edge set $E \subset \mathcal{P}(V)$ connects any number of edges ($\mathcal{P}(V)$ indicates the power set) . For any hypergraph, we can construct a *representing graph* $\mathcal{R} = (V, E')$ with $(i, j) \in E' \leftrightarrow (i, j) \subset e$ for some hyperedge $e \in E$.



The representing graph has a broader degree distribution than the original hypergraph because every hyperedge of cardinality c is replaced by $c(c-1)/2$ edges. We find a weak Spearman correlation of 0.34 between hypergraph degree k_i and representing graph degree $k_i^{(\mathrm{rep})}$, indicating that the two mathematical objects will yield different information.

Null models

We investigate two null models, the *binomial random hypergraph*, which preserves the number of edges and their cardinality, and the *configuration hypergraph*, which additionally preserves the degrees of all nodes [2].



The degree-assortativity ρ_a, the number n_{com} of components, and the relative size of the $S_{\max}/N$ largest component differ strongly from the data in both null models. For the mean local clustering $\langle C_i \rangle$, however, only the binomial null mode differs from the original hypergraph.

	data	configuration	binomial
degree-assortativity ρ_a	0.43	−0.01	0.01
number n_{com} of components	253	2	1
larg. component size $C_{\max}/N$	0.88	0.99	1
mean local clustering $\langle C_i \rangle$	0.99	0.99	0.90

References

[1] Waqar Ali, Charlotte M. Deane, and Gesine Reinert. Protein interaction networks and their statistical analysis. In Michael P. H. Stumpf, David J. Balding, and Mark Girolami, editors, *Handbook of Statistical Systems Biology*, pages 200–234. John Wiley & Sons, Ltd Chichester, UK, 2011.

[2] Philip S Chodrow. Configuration models of random hypergraphs and their applications. *arXiv preprint arXiv:1902.09302*, 2019.

[3] David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(suppl_1):D691–D697, 2010.

Acknowledgements

EPSRC Engineering and Physical Sciences Research Council

F.K. was supported by EP/R513295/1 through a Doctoral Prize. G.R. was supported in part by EP/R018472/1.