Combining Spectral Clustering and Large Cut Algorithms to find Compensatory Functional Modules from Yeast Physical and Genetic Interaction Data with GLASS

Blessing Kolawole
blessing.kolawole@tufts.edu
Tufts University
177 College Ave
Medford, MA 02155
USA

Lenore J. Cowen*
lenore.cowen@tufts.edu
Tufts University
177 College Ave
Medford, MA 02155
USA

ABSTRACT

Various algorithmic and statistical approaches have been proposed to uncover functionally coherent network motifs consisting of sets of genes that may occur as compensatory pathways (called Between Pathway Modules, or BPMs) in a high-throughput *S. Cerevisiae* genetic interaction network. We extend our previous LocalCut/Genecentric method to also make use of a spectral clustering of the physical interaction network, and uncover some interesting new fault-tolerant modules.

KEYWORDS

PPI networks; genetic interaction networks; epistasis; Protein function prediction; Spectral Clustering

ACM Reference Format:
Blessing Kolawole and Lenore J. Cowen. 2022. Combining Spectral Clustering and Large Cut Algorithms to find Compensatory Functional Modules from Yeast Physical and Genetic Interaction Data with GLASS. In *13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '22), August 7–10, 2022, Chicago, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3535508.3545509

1 INTRODUCTION

It is estimated that only about 18% of individual yeast genes are essential, meaning that a deletion mutant with that gene deleted or suppressed is not viable [31]. For genes that are not essential, the *S. cerevisiae* genetic interaction network comes from high-throughput epistasis experiments, where edge weights represent the surprise in growth rates from double deletion mutants, compared to their associated single deletion mutants. In particular, a negative weight edge indicates that there is a growth defect (or in worst case synthetic lethality, meaning the double deletion mutant is not viable), and a positive weight edge indicates that the double deletion mutant is not sicker than the constituent single deletion mutants (or in best case, synthetic rescue, where although single deletion mutants display reduced growth, the double deletion mutant behaves like wild type) [27]. The pattern and organization of this signed, weighted genetic interaction network has been shown [12] to contain interesting motifs that can indicate redundancy and the presence of compensatory pathways. These alternative pathways can be mechanisms of global resilience.

The LocalCut algorithm of [15] as implemented in the *Genecentric* [7] software package, searched for pairs of sets of genes, or *modules* in the genetic interaction network, where there were predominantly large negative edge weights on the inter-module edges (as well as predominantly positive weights among the intra-module edges). These pairs of sets of genes we will call gBPMs in this paper, for *generalized between pathway models* (for the relation between our gBPM definition and earlier BPM definitions, see the related work section below, Section 1.1). While LocalCut/Genecentric produced more and higher quality (as measured by functional enrichment) gBPMs than previous methods (see [7, 15]), it is still the case that a very strong gBPM can obscure the presence of other interesting gBPMs, since the algorithm favors the sets of genes with most negative weight globally (much the same way it is hard to see a dimmer spotlights in the presence of nearby bright spotlights). In order to also uncover these obscured gBPMs, it would be necessary to move away from the global partition at the heart of LocalCut, and instead search for gBPMs in sub-portions of the network.

Our new method, GLASS (for Gencentric/LocalCut Across Spectral Subclusters) for finding gBPMs does exactly this: the functional coherence of the physical protein-protein interaction network is leveraged to partition it into k clusters, and then LocalCut is independently run on the clusters. By looking across a range of different clustering scales (different k), we end up finding interesting BPMs that LocalCut would miss. Here, we re-analyze the original EMAP data of Collins et al. [3] that LocalCut was tested on, and find GLASS uncovers some new and interesting putative compensary pathways, including ones involving DNA damage repair, the mediator complex, and Golgi to endosome transport (see Section 4).

1.1 Previous Work

The "Between Pathway Model" or BPM, was first defined in a mixed physical and genetic interaction network, where both types of edges were superimposed in the same graph, by Kelley and Ideker [12]. This version pre-dated the EMAP and SGA experiments, so the genetic interaction edges were only of one type: they indicated whether or not the double mutant was synthetic lethal. Kelley and Ideker suggested searching for pairs of gene modules with many synthetic lethal inter-module edges, and many physical interaction

*To whom correspondence should be addressed: lenore.cowen@tufts.edu

ACM-BCB '22, August 7–10, 2022, Chicago, IL, USA

ACM ISBN 978-1-4503-9386-7/22/08.
https://doi.org/10.1145/3535508.3545509

intra-module edges, which they termed a BPM. The same definition of BPM was used by Ulitsky and Shamir. Subsequent work of Ma, Tarone and Li [16] and Brady et al. [1] then looked only at the genetic interaction portion of the network, where their BPMs matched the Kelley-Ideker pattern for the placement of the genetic interaction edges. (Brady et al [1] was then able to use the location of known physical interaction edges for *validation* of their BPMs.) When EMAP [3] and SGA [4] technology arrived, a richer signed genetic interaction became available, and an associated generalized BPM motif could be defined, again, solely using the genetic interaction network as in [1, 16]. This is the gBPM definition of Leiserson et. al [15] and of Gallant et al [7], and matches the gBPM definition now used in the paper. The goal of [7, 15] and our present work, is to output collections of gBPMs that 1) contain between 3 and 25 genes per pathway 2) produce individual gBPMs that place edges with large negative weights between the two pathways, and rarely within each pathway and 3) have the collections be substantially non-redundant (so there is a Jaccard threshold on how similar two BPMs from the collection are allowed to be). However, the input data to [7, 15] is simply the weighted genetic interaction network, whereas GLASS, our present method also requires the physical interaction network, same as for earlier BPM-finding methods.

2 DATA

We consider the original *S. cerevisiae* EMAP dataset from the Collins et al. paper [3] (downloaded from chrombio); consisting of 754 genes and 373,000 edges. For our physical interaction network, we used the *S. cerevisiae* pysical multi-validated interactions dataset from the BioGRID database [18] version 4.4.207 (downloaded on 02-27-2022); retaining only edges labeled as physical interactions (direct interaction, physical association, or colocalization). We retained only the 722 nodes that appear in both networks: for these nodes, there are 13,956 physical interaction edges.

3 METHODS

GLASS uses Diffusion state distance, called DSD [2] on the unweighted physical interaction network, to compute pairwise distances between the nodes. Then, the spectral clustering algorithm of [20] is called on the similarity measure which is the reciprocal of these distances to partition the nodes into k spectral clusters, for each $k \in \{1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$. Spectral clusters of size < 6 are discarded from each collection. Genecentric is then run on the k spectral clusters with ≥ 6 nodes, with parameters $m = 500$, $c = 0.85$, and Jaccard $= 0.66$, where these are the recommended defaults in [7], except for c, which is set slightly more permissively to find more BPMs. For the *union collection* that GLASS produces (see below), the same defaults are used except the Jaccard parameter is set to 1 (so only 100% identical gBPMs are pruned).

We note that across all k values for the Spectral Clusters, GLASS will produce a highly redundant set of gBPMs, since highly similar gBPMs can occur at multiple values of k. In fact, Genecentric itself at each fixed k, first returns a highly redundant set of gBPMs, and then uses the Jaccard index to guide its pruning of the collection. When we consider the union of all gBPMs across all different k, for some applications it makes more sense to do this Jaccard pruning once (across all the levels) then once for each level, and then across all k. On the other hand, we can also use Genecentric as a black box at each fixed k value (pruning level by level) and identify genes that co-occur in multiple gBPMs across levels. These two collections are formally defined as follows:

Definition 3.1. Consider the *union collection* of all GLASS gBPMs constructed as follows: First consider all gBPMs returned by LocalCut for any k with Jaccard threshold set to 1 in the Genecentric algorithm. Then have Genecentric prune this collection to retain only modules whose Jaccard index exceeds .66.

Definition 3.2. GLASS also constructs *consensus gBPMs* as follows. In this case, Genecentric with the parameters listed above (including Jaccard=.66) produces a different set of gBPMs on the k spectral clusters, for each k. Given a set S of 3 genes that appear together in one module in one gBPM for some k, we consider all *gBPMs* across all values of k that also contain those three genes. For each gene g that appears in the same module as S in at least 3 of the gBPMs we place g in the *consensus gBPM* module that contains S, and for each gene h that appears in the opposite module to S in at least 3 of the gBPMs, we place h in the *consensus gBPM* module opposite to S.

Functional Enrichment Analysis We used the GProfiler GOST (Gene Ontology Statistics) python package version 1.0.0 [23] to perform statistical enrichment analysis on the gBPMs generated (using the SCS multiple testing correction with threshold $a = 0.05$).

4 RESULTS

Figure 1 graphs the number of gBPMs (on a full gBPM and on a per-pathway basis) generated by GLASS at each fixed value of k. As can be seen, the number of BPMs increased as k goes from 5 to 15, then decreases slightly as it goes to 25; before it peaks again at $k = 30$. As k goes from 30 to 50, there is a significant drop in the number of modules and BPMs produced. Figure 2 graphs the percentage of enriched modules across different values of k.

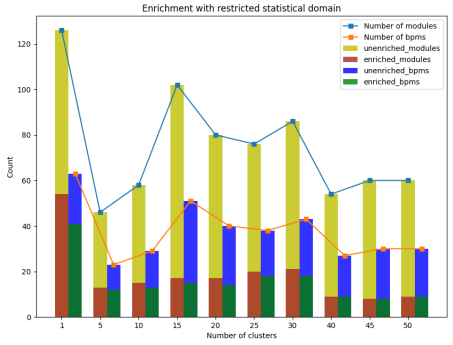

Figure 1: The number of gBPMs generated by GLASS at each fixed value of k. First bar: total number of modules; marked with percentage functionally enriched; second bar, number of gBPMs marked with percentage functionally enriched.

We note that the GLASS union collection contains 209 BPMs (pruned at Jaccard index .66) which is nearly 4 times the number of

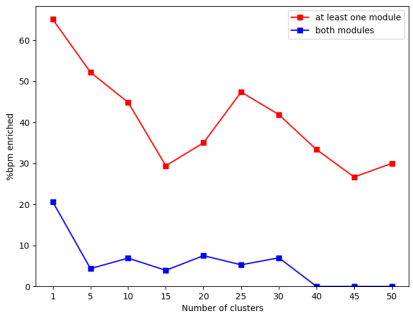

Figure 2: A map of the total number of gBPMs that have enriched modules across different values of k

	BUB2	ELG1	POL32	REV1	REV3	REV7
POL31	-0.373	0.002	-15.148	-0.435	-0.945	-0.831
RAD5	-1.383	0.623	-2.768	-1.338	-1.257	-1.908
RAD50	-1.613	-6.606	-9.335	-0.721	-0.077	-0.067
RAD51	-3.025	-8.491	-8.861	-9.516	-5.105	-6.205
RAD52	-5.766	-8.166	-11.824	-6.620	0.000	-3.898
RAD54	-1.801	-5.577	-10.501	-7.136	0.000	-3.733
RAD55	-1.238	-7.525	-11.557	-5.960	-8.885	-6.555
RAD57	-1.769	-5.934	-7.056	-6.662	-8.475	-4.826
UBC13	-0.022	-0.691	-9.048	-1.558	0.000	-3.830
XRS2	-3.804	-10.070	-4.576	-0.560	0.656	0.314

Table 1: The Emap genetic interaction scores between both modules of the RAD-REV consensus gBPM. Large negative scores indicate a synthetically lethal epistasis relationship.

gBPMs than was found by Genecentric/Local Cut alone (63). We next highlight some of the gBPMs produced by GLASS.

4.1 The zeta DNA polymerase complex

GLASS uncovers gBPMs that include a module containing the zeta DNA polymerase complex (REV3, REV7, REV1) [17] which appear together at all levels of the spectral clustering. When this complex appears, it is consistently opposite the RAD52 epistasis group which includes genes RAD50, RAD51, RAD52, RAD54, RAD55, RAD57, and XRS2 [26] (see Table 1 for the consensus gBPM).

It is well known that the DNA polymerase complex and the RAD52 epistasis group are involved in two different pathways responsible for DNA damage repair; REV3 and REV7 are primarily considered to be in the Translesion DNA Synthesis (TLS) pathway [10, 13, 14, 21] while RAD52 epistasis group genes are involved in the Homologous Recombination (HR) repair pathway [8, 11, 26]. RAD52 is involved in multiple pathways for repair of double strand breaks in DNA [19]. It is known that the rate of errors for repairing DNA damage breaks depends on whether the RAD52 or REV3 pathway is active: the error-free RAD52 pathway and error-prone REV3 pathway for rescuing replication fork arrest determine spontaneous mutagenesis, recombination, and genome instability [6] and in the absence of RAD52, REV3-dependent base-substitutions increase, while in the absence of REV3, RAD52-dependent recombination events increase. The rad52 rev3 double mutant had an enhanced chromosome loss mutator phenotype[6].

While it was known that the two modules of this gBPM were involved in alternative DNA damage repair, the gBPM implies one of these pathways will be essential perhaps even for normal DNA replication, perhaps during G2 phase, where [30] report that *S. cerevisiae* Rev1 is subject to pronounced cell cycle control and levels of Rev1 protein are approximately 50-fold higher in G2 and throughout mitosis than during G1 and much of S phase.

We note that some version of this gBPM appears at every value of k including $k = 1$, so this is an example gBPM that LocalCut/Genecentric would have already discovered some version of.

4.2 The Ric1-Rgp1 guanyl-nucleotide exchange factor complex

Another consensus gBPM GLASS finds that LocalCut/Genecentric does not includes the RIC1-RGP1 complex. We observed that multiple genes in the module containing that complex are involved in Golgi to endosome tranport. [22, 25]. Much less seems known about the genes in the opposite module, except for the gene MON2/YSL2 which is known to be involved in Golgi to endosome transport as a regulator of Endosome-To-Golgi Trafficking[5] and also identified as a gene whose deletion is synthetically lethal with loss of the Rab6 homologue Ypt6 or its Guanine nucleotide exchange factor (GEF), Ric1 [9]. Based on this consensus gBPM, we hypothesize an undiscovered role in Golgi endisome transport for some of the other genes in this module.

4.3 The Core Mediator complex

The core mediator complex, that functions as a bridge between DNA-binding transcription factors and the RNA polymerase II machinery [29], is involved in a gBPM that is not retrieved by LocalCut/Genecentric, but is uncovered by GLASS, appearing repeatedly across all the spectral cluster levels from 5 to 50. The gBPM segregates genes involved in different portions of the mediator complex. The consensus gBPM found by GLASS contains genes MED11, SRB2 (also called MED20), and MED8 which are often regarded as the head module of the mediator complex [24, 28, 29], and MED4 a member of the middle mediator complex in one of the modules. On the other module, it has genes CSE2 (alias MED 9), and MED1 belonging to the core mediator complex's middle module and SIN4 (alias MED 16) belonging to the tail modules of the mediator complex [29]. Thus this gBPM witnesses that only part of the mediator complex might be essential for viability.

5 DISCUSSION

We introduced GLASS that was able to find additional structure in combined yeast genetic interaction and physical interaction networks compared to the previous LocalCut/Genecentric method. In a re-analysis of a yeast epistasis network, we find that GLASS

	OAF1	RGP1	RIC1	RIM21	SLM3	VPS21	YPL113C	YPT6	YRM1
CAF40	1.245	-10.383	-6.310	0.000	0.163	-6.513	-0.031	-4.415	1.035
ESC8	0.000	-8.190	-1.123	0.000	0.043	0.000	0.000	-4.700	0.000
GCR2	0.000	-8.187	-2.780	0.000	0.274	0.000	0.000	-4.486	0.000
HUL4	-0.044	0.000	-0.147	0.602	0.487	0.286	-0.674	-0.373	0.412
ISW1	0.510	-2.920	-3.246	-0.954	-2.726	0.000	0.428	-2.385	-1.021
MCK1	-0.043	-7.384	-11.401	0.000	-2.135	-1.538	0.085	-7.080	1.468
MON2	-0.183	-14.879	-7.186	0.000	1.899	1.681	-0.490	-12.410	0.537
RAD28	0.502	0.000	0.174	0.297	-2.849	0.618	-0.483	0.397	0.889
STB5	-1.486	0.000	-0.294	0.469	-0.937	0.328	-0.803	-3.228	-0.080
VPS8	-0.991	-8.457	-6.510	-2.322	2.779	0.000	0.774	-6.467	0.534

Table 2: The Emap genetic interaction scores between both modules of a novel gBPM uncovered by GLASS related to Golgi to endosome transport. Large negative scores indicate a synthetically lethal epistasis relationship

	CSE2	MED1	SIN4
MED11	0.000	-6.330	-3.231
MED4	0.000	-4.293	0.076
MED8	0.000	-4.525	0.462
SRB2	-6.845	-9.440	-8.590

Table 3: The Emap genetic interaction scores between both modules of the head and middle/tail mediator complex consensus gBPM. Large negative scores indicate a synthetically lethal epistasis relationship.

uncovers both many of the original compensatory sets of genes and pathways, but also highlights some new compensatory sets of genes and pathways. In the future, we will be applying GLASS to more modern epistasis datasets.

6 ACKNOWLEDGEMENTS

We thank the Tufts BCB group for helpful discussions. This research was partially supported by NSF grant 1934553 (to LC).

REFERENCES

[1] A. Brady et al. 2009. Fault tolerance in protein interaction networks: stable bipartite subgraphs and redundant pathways. *PloS one* 4, 4 (2009), e5364.
[2] M. Cao et al. 2013. Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PloS one* 8, 10 (2013), e76339.
[3] S. R. Collins et al. 2007. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 446, 7137 (2007), 806–810.
[4] M. Costanzo, B. VanderSluis, et al. 2016. A global genetic interaction network maps a wiring diagram of cellular function. *Science* 353, 6306 (2016), aaf1420.
[5] J. .A Efe, F. Plattner, et al. 2005. Yeast Mon2p is a highly conserved protein that functions in the cytoplasm-to-vacuole transport pathway and is required for Golgi homeostasis. *Protein science* 118, 20 (2005), 4751–4764.
[6] K. Endo, Y. Tago, et al. 2007. Error-free RAD52 pathway and error-prone REV3 pathway determines spontaneous mutagenesis in Saccharomyces cerevisiae. *Genes & genetic systems* 82, 1 (2007), 35–42.
[7] A. Gallant, M. D. M. Leiserson, Maxim Kachalov, et al. 2013. Genecentric: a package to uncover graph-theoretic structure in high-throughput epistasis data. *BMC bioinformatics* 14, 1 (2013), 1–7.
[8] J.C. Game and R.K. Mortimer. 1974. A genetic study of X-ray sensitive mutants in yeast. *Mutation Res.* 24, 3 (1974), 281–292.
[9] A. Gillingham, J. Whyte, B. Panic, and S. Munro. 2006. Mon2, a relative of large Arf exchange factors, recruits Dop1 to the Golgi apparatus. *J. of Biol. Chem.* 281, 4 (2006), 2273–2280.
[10] Y. Hirano and K. Sugimoto. 2006. ATR homolog Mec1 controls association of DNA polymerase ζ-Rev1 complex with regions near a double-strand break. *Current Biology* 16, 6 (2006), 586–590.
[11] P. Jeggo. 1990. Studies on mammalian mutants defective in rejoining double-strand breaks in DNA. *Mutation Research* 239, 1 (1990), 1–16.
[12] R. Kelley and T. Ideker. 2005. Systematic interpretation of genetic interactions using protein networks. *Nature biotechnology* 23, 5 (2005), 561–566.
[13] N. K. Kolas and D. Durocher. 2006. DNA repair: DNA polymerase ζ and Rev1 break in. *Current Biology* 16, 8 (2006), R296–R299.
[14] C. W. Lawrence. 2002. Cellular roles of DNA polymerase ζ and Rev1 protein. *DNA repair* 1, 6 (2002), 425–435.
[15] M. D. M. Leiserson, D. Tatar, et al. 2011. Inferring mechanisms of compensation from E-MAP and SGA data using local search algorithms for max cut. *Journal of Computational Biology* 18, 11 (2011), 1399–1409.
[16] X. Ma et al. 2008. Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. *PloS one* 3, 4 (2008), e1922.
[17] J. R. Nelson, C. W. Lawrence, and D. C. Hinkle. 1996. Thymine-thymine dimer bypass by yeast DNA polymerase ζ. *Science* 272, 5268 (1996), 1646–1649.
[18] R. Oughtred et al. 2021. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* 30, 1 (2021), 187–200.
[19] F. Pâques and J. E. Haber. 1999. Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiology and mol. bio. rev.* 63, 2 (1999), 349–404.
[20] F. Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. *JMLR* 12 (2011), 2825–2830.
[21] S. Prakash, R. E. Johnson, and L. Prakash. 2005. Eukaryotic translesion synthesis DNA polymerases: specificity of structure and function. *Annu. Rev. Biochem.* 74 (2005), 317–353.
[22] G. Pusapati, G. Luchetti, and S. Pfeffer. 2012. Ric1-Rgp1 complex is a guanine nucleotide exchange factor for the late Golgi Rab6A GTPase and an effector of the medial Golgi Rab33B GTPase. *J. of Biol Chem* 287, 50 (2012), 42129–42137.
[23] U. Raudvere et al. 2019. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research* 47, W1 (2019), W191–W198.
[24] S. Sato, C. Tomomori-Sato, et al. 2003. Identification of Mammalian Mediator Subunits with Similarities to Yeast Mediator Subunits Srb5, Srb6, Med11, and Rox3* 210. *J. of Bio Chem.* 278, 17 (2003), 15123–15127.
[25] Y. Suda, K. Kurokawa, et al. 2013. Rab GAP cascade regulates dynamics of Ypt6 in the Golgi traffic. *PNAS* 110, 47 (2013), 18976–18981.
[26] L. Symington. 2002. Role of RAD52 epistasis group genes in homologous recombination and double-strand break repair. *Microbiology and mol. biol. rev.* 66, 4 (2002), 630–670.
[27] A. H. Y. Tong, G. Lesage, et al. 2004. Global mapping of the yeast genetic interaction network. *science* 303, 5659 (2004), 808–813.
[28] K.-L. Tsai, C. Tomomori-Sato, et al. 2014. Subunit architecture and functional modular rearrangements of the transcriptional mediator complex. *Cell* 157, 6 (2014), 1430–1444.
[29] X. Wang, Q. Sun, et al. 2014. Redefining the modular organization of the core Mediator complex. *Cell research* 24, 7 (2014), 796–808.
[30] L. S. Waters and G. C. Walker. 2006. The critical mutagenic translesion DNA polymerase Rev1 is highly expressed during G2/M phase rather than S phase. *PNAS* 103, 24 (2006), 8971–8976.
[31] E. A. Winzeler et al. 1999. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *science* 285, 5429 (1999), 901–906.