

Genetic Interaction Motif Finding by Expectation Maximization – a novel statistical framework for inferring gene modules from synthetic lethality

Yan Qi^{1,2}, Ping Ye^{1,2} and Joel S. Bader^{1,2}

Keywords: synthetic lethality, expectation maximization, probabilistic motif

1 Introduction.

Synthetic lethality experiments identify pairs of genes with complementary function: two genes are synthetic lethal if each mutant is viable, but the double mutant combination is lethal. More direct functional associations may be inferred between genes that share synthetic lethal interaction partners than genes that are directly synthetic lethal. We describe an unsupervised algorithm, Genetic Interaction Motif Finding (GIMF), which uses probabilistic motifs to identify gene modules based on synthetic lethal interaction data. The dataset is obtained from SGA analysis in *Saccharomyces cerevisiae* [3], where a single mutant (query gene) is introduced into the complete pool of viable yeast single-deletion (library gene) strains.

The model is developed under the hypothesis that genes within the same pathway exhibit a similar pattern of synthetic lethality, which is referred to as a motif. Starting with a seed gene, genes with interaction patterns similar to that of the seed gene are grouped into a motif set while the remaining genes are sorted into a non-motif set. The motif is compactly characterized by the interaction probabilities between motif genes and library genes, distinct from the homogeneous interaction probabilities (background) between non-motif genes and library genes. The linkage probabilities between query and library genes are modeled such that promiscuous genes with a large number of interaction partners are automatically down-weighted. Each gene is assigned a score based on the position weight matrices and a probability of being in the motif set is obtained through maximum likelihood estimation. These probabilistic quantities are optimized iteratively through expectation maximization [1] [2].

Using each of the 126 non-essential query genes as seeds, we have identified known and novel pathways for yeast. Motifs for a seed gene *DYNI* is shown in Fig. 1. The motifs are fairly insensitive to the single tunable parameter for GIMF, which reflects false positive rate of the experiment. GIMF links are asymmetric, which increases motif specificity and masks impact from promiscuous genes. By retaining pairs of genes which are found as each other's motif member, a core gene network with high confidence of functional association can be obtained. This network reveals two clusters: the *PAC10* complex and the *Dynein-Dynactin* pathway with the latter incorporating an uncharacterized gene *NUM1*. Indeed, the deletion mutant of *NUM1* exhibits severe nuclear migration defect [4]. Correlations with Gene Ontology (GO) annotations show that GIMF extracts biologically relevant sub-networks (Table 1). These statistics also shows the consistency and complementarity between GIMF and the congruence score method based on hypergeometric distribution.

¹ Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218

² High-Throughput Biology Center, Johns Hopkins School of Medicine, Baltimore, MD 21287
E-mail: joel.bader@jhu.edu

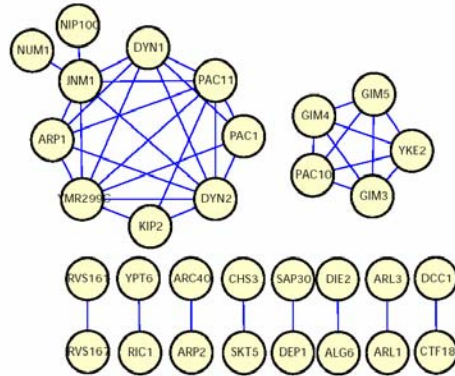


Figure 1: GIMF gene network built from query genes. The nodes are query genes and an edge between node i and node j indicates that i and j are each other's motif members.

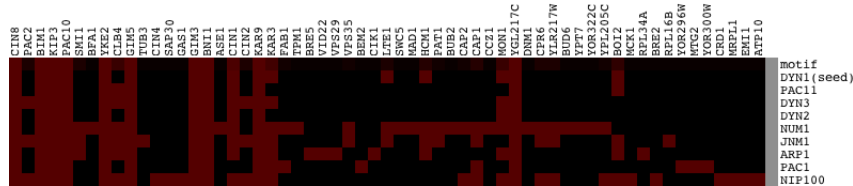


Figure 2: Genetic interaction patterns for the motif generated with *DYN1* as seed gene.

Gene pairs	GO correlation			FSL	FPC	Number of pairs	Number of genes
	P	F	C				
GIMF	0.47	0.20	0.43	0	0.26	42	31
CS	0.54	0.17	0.48	0.2381	0.26	42	36
SL	0.25	0.05	0.31	--	0.01	--	--

Table 1: GO annotation correlations for GIMF (31 genes, 42 pairs) gene pairs, congruent gene (CS) pairs [4] and gene pairs that are directly synthetic lethal (SL). P: biological process; F: molecular function C: cellular component. FSL: fraction of pairs that are directly synthetic lethality; FPC: fraction of pairs that are within the same protein complex. Congruent pairs are chosen with a cutoff value of 25, which yields 42 pairs associated with 36 genes.

References

- [1] Bailey, T. L. 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning Journal*, 21, 51-83.
- [2] Lawrence, C. E. and Reilly, A. A. 1990. An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *PROTEINS: Structure, Function, and Genetics*, 7, 41-51.
- [3] Tong, A. H. Y., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Page, N., Robinson, M., Raghizadeh, S., Hogue, C. W. V., Bussey, H., et al. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294, 2364-2368.
- [4] Ye, P., Peysner, B., Pan, X., Boeke, J. D., Spencer, F. A. and Bader, J. S. 2004. Quantified measures of systems robustness in yeast. Submitted