

Intrinsic Bayesian estimates of the eukaryotic protein interaction network complexity

Hailiang Huang^{1,2,3}, Myung Lee¹, Andy Cheng¹, Joel S. Bader^{1,2}

Keywords: Bayesian inference, yeast two hybrid, coverage

Introduction

Despite progress in high-throughput protein interaction screens, the number of protein-protein interactions in human and model organisms remains uncertain. Coverage estimates continue to rely primarily on extrinsic comparisons of the overlap between disparate data sets [1]. These comparisons are subject to systematic error due to differences in the technologies used and the libraries screened. For direct protein-protein interactions identified by two-hybrid screens, we present a novel computational method for generating an intrinsic estimate of the number of total interactions based on the sample observed. The result may prove useful in experiment designs such as providing optimization strategies to select baits [2].

The two-hybrid screens sample interactions randomly: a protein is selected as a bait, the bait is screened against a library of preys, and a subset of the preys that yield a positive assay result indicating interactions sampled. The choice of baits and the depth of sampling affect the coverage of the screen. Denote the true number of interaction partners of a bait as M . Within the complete set of interacting preys, a subset of K preys is sampled, which may contain repetitive proteins. Let X denotes the number of unique proteins within the set of K sampled proteins. We provide an exact solution of a mathematical model using Bayesian inference to infer the probability distribution of interaction partners M based on X and K , $\Pr(M|X,K)$, for each protein. We then summarize this distribution through the expected mean and median for M . The network coverage is estimated as the fraction of interactions observed, equal to $(\sum_i X_i)/(\sum_i M_i)$ where i sums over each bait.

These estimates converge only when $K - X \geq 2$. Conditioning on a prior $\Pr(M) \sim M^{-\epsilon}$ ensures converges for $\epsilon > 1$. Coverage results are provided as a function of ϵ in Fig. 1 for the entire data set, and in Fig. 2 for the part of the data set where $K > X$. The number of baits removed from each network (those with $K = X$, each partner seen only once) is provided in Table I.

In the full data set, the estimate of edge coverage is very sensitive on the prior estimate for $\Pr(M)$. When the hard cases are removed from the data set (baits where each experiment yielded a different prey), however, the coverage estimates are less sensitive to the prior and range from about 30% to 70%. The hard cases represent ~50% of the baits.

¹ Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, E-mail: h.lhuang@pha.jhu.edu

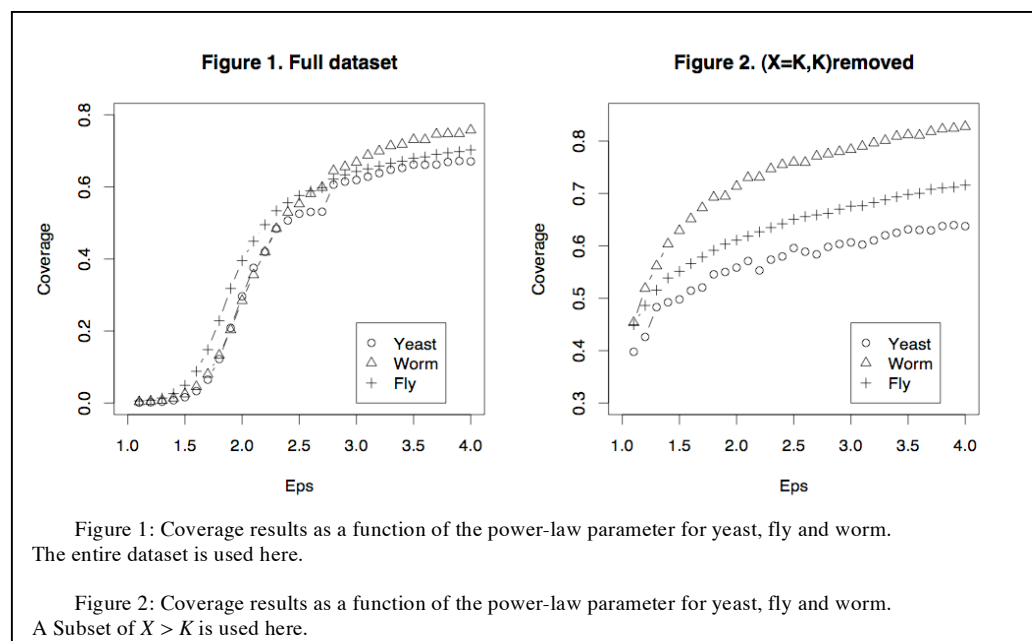
² High-Throughput Biology Center, Johns Hopkins School of Medicine, Baltimore, MD 21287

³ Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218

Figures and tables

	Total number of baits	Number of baits with $X=K$
Yeast	1497	612
Worm	801	438
Fly	3607	2641

Table I. Number of baits in the dataset of yeast, worm and fly.



References

- [1].Bader JS, Chaudhuri A, Rothberg JM, Chant J: Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* 2004, 22(1):78-85.
- [2].Lappe M, Holm L: Unraveling protein interaction networks with near-optimal efficiency. *Nat Biotechnol* 2004, 22(1):98-103.