

# Prediction of domain interactions in *C. elegans*

Faruck Morcos<sup>1</sup>, Mike Boxem<sup>2</sup>, Niels Klitgord<sup>2</sup>, Marc Vidal<sup>2</sup>, and  
Jesús A. Izaguirre<sup>1,3</sup>

<sup>1</sup> Department of Computer Science and Engineering,  
Notre Dame IN 46556 USA.

*E-mail:* {amorcosg, izaguirr}@nd.edu

<sup>2</sup> Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber  
Cancer Institute, and Department of Genetics, Harvard Medical School, Boston MA 02115 USA.

*E-mail:* {mike\_boxem, niels\_klitgord, marc\_vidal}@dfci.harvard.edu

<sup>3</sup> Corresponding author.

High throughput experiments and computational methods allow the elucidation of networks of protein-protein interactions (PPI) in several organisms. Since about 80% of all proteins consist of multiple self-folding units called domains, it is desirable to determine which domains are responsible for a given interaction. We use an integrative computational and experimental approach to predict domain-domain interactions (DDI) from the known PPI for the metazoan *C. elegans*. We use Maximum Likelihood Estimation and Maximum Specificity Set Cover algorithms to predict the DDI. We validate the results using a small and accurate yeast 2 hybrid (Y2H) screen on 8 protein-protein interactions involving 13 proteins in *C. elegans*. The computational predictions are consistent with Y2H in 7 of the 13 interaction domains, including the prediction of domain fusions. These results illustrate the predictive power of a combined computational and experimental approach to DDI determination.

## 1 Introduction

Macromolecular interactions among proteins, nucleic acids, and lipids, form the basis of biological activity. High-throughput experiments have enabled the elucidation of a substantial number of protein-protein interactions (PPI) in several organisms. About 80% of these proteins consist of multiple domains, which are self-folding polypeptide units. Many of these domains are repeated across different proteins, and hence allow for a modularization of the interaction networks. The domain architecture of many proteins is available in Pfam<sup>1</sup>. We use a network of PPI and Pfam to infer a network of domain-domain interactions (DDI). We validate this network on a set of 8 well-characterized protein interaction pairs involving 13 proteins from the roundworm *C. elegans*. Validation is done using an experimental yeast 2 hybrid (Y2H) approach. Our results show substantial agreement between the computational prediction methods and experiment (agreement for 7 out of 13 interaction domains). Several of the predicted interaction domains consist of domain fusions. Several of the domain interactions are predicted by two different computational methods, adding to the confidence level of the predictions. Failures of the computational predictions can be partly explained by the ability of the experiments to discover novel interaction domains, and in one case by the fusion of 3 domains, which was not considered in the computational model. This integrative approach has the potential to explain at a greater level of detail observed PPI.

Interaction	Description
SQV-4 – SQV-4	UDP-glucose 6-dehydrogenase, required for cytokinesis.
EMB-27 – B0511.9	Anaphase promoting complex subunit and novel gene.
MEI-1 – MEI-2	Together forms the microtubule severing katanin complex.
DYRB-1 – DYCI-1*	Homology to dynein subunits (light and intermediate chains).
Y65B4BR.5 – ICD-1	Predicted transcription factor and $\beta$ NAC.
MEL-26 – MEI-1s	Ubiquitin ligase substrate adaptor targeting MEI-1.
MEL-26 – MEL-26	Ubiquitin ligase substrate adaptor targeting MEI-1.
ATN-1* – ATN-1*	$\alpha$ -actinin, actin bundling protein that homodimerizes.
GPR-1* – GOA-1	GPR-1 regulates the activity the G-protein subunit GOA-1.
* Interaction domain known	

Table 1. Y2H protein interaction set of high confidence.

## 2 Methods

The network of PPI examined consists of a set of 9,044 interactions among 4,625 proteins, collectively named WI7<sup>2</sup>. From the 9,044 PPI, 4,736 are experimental, 3,359 based on homology with *Drosophila* interactions, and 949 based on homology with *S. cerevisiae* interactions. The domain architecture was obtained from Pfam-A and Pfam-B, yielding a total of 8,558 interactions with known domain architecture. This was used as training set for two different algorithms, which yield a network of DDI. The first is the maximum likelihood estimation (MLE) procedure which has been applied successfully to the problem of estimating the DDI from experimental PPI for yeast<sup>3</sup>. The other algorithm is a weighted set cover algorithm that we developed, called Maximum Specificity Set Cover (MSSC), which selects domain-domain interactions while minimizing the number of false positives in the training set<sup>4</sup>. Domain fusions of 2 domains are included in the set of potential interaction domains.

Previous work has validated the ability of using the network of DDI found by MLE and MSSC to predict PPI. In this work we attempt to directly evaluate the quality of a small portion of these DDI. Suppose that an interaction domain pair  $d_i - d_j$  is part of the DDI produced by any of the above methods and can explain a protein interaction  $p_i - p_j$ . Y2H is capable of accurately identifying the interaction domain for a protein-protein interaction. First the domain  $d_i$  in  $p_i$ , when interacting with  $p_j$ , is identified; then  $d_j$  in  $p_j$  is identified when interacting with  $p_i$ . From the Worm Interactome<sup>5</sup> version 5, we selected 8 interactions involving 13 proteins (Table 1). These interactions were selected because they are high-confidence Y2H interactions, and the proteins are involved in a diverse set of biological processes. This set of interactions is used to compare experimental and computational predictions.

## 3 Results

MLE and MSSC were run from the PPI server (<http://ppi.cse.nd.edu>). MLE selected 5,170 domains and 36,489 DDI. MSSC explained the same network with 3,138 domains and 12,870 DDI. MSSC predicted 189 domain fusions involved in 1,872 DDI. Predicted interaction domains for the small screen test are shown in Table 2. By inspection, 7 of the 13 interaction domains are consistent with Y2H. SQV4 homodimerization could not be predicted by only using fusion of 2 domains, since Y2H indicates that 3 domains are

Protein Interaction	Domain Interaction	Remarks
SQV-4 (3) – SQV-4 (3)	UDPG_MGDP_dh – UDPG_MGDP_dh	Y2H indicates 3 domains needed –
EMB-27 (2) – B0511.9 (1)	Pfam-B_41051 – Pfam-B_92146 <sup>†</sup>	Pfam-B_41051 not consistent with Y2H – Y2H consistent with Pfam-B_92146
MEI-1 (1) – MEI-2 (1)	AAA – Pfam-B_62049	Y2H consistent with AAA Pfam-B_62049 not consistent with Y2H –
DYRB-1 (1) – DYCI-1* (4)	Robl_LC7 – Pfam-B_4434	DYRB-1 had no hit in Y2H – Y2H consistent with Pfam-B_4434
Y65B4BR.5 (3) – ICD-1 (2)	UBA_NAC – NAC	Y2H consistent with fusion Y2H consistent with NAC
MEL-26 (3) – MEL-26 (3)	BTB_MATH – BTB_MATH	Y2H consistent with fusion
ATN-1* (4) – ATN-1* (4)	Pfam-B_843 – Pfam-B_843 <sup>†</sup>	Pfam-B_843 not consistent with Y2H –
GPR-1* (2) – GOA-1 (1)	Pfam-B_66388 – G-alpha <sup>†</sup>	Pfam-B_66388 not consistent with Y2H – Y2H consistent with G-alpha
* Interaction domain known, † Predicted by both MLE and MSSC		

Table 2. Set of DDI predicted by MLE and MSSC. The notation (#) besides the protein name represents the number of domains that conform such protein. A – at the end of a remark indicates this was a negative result.

needed. MEI-1 requires the entire protein in the experiment, and the predicted interaction domain AAA is contained within this area. The interaction domain in MEI-2 is consistent with experiment although Y2H used a smaller piece of the complete predicted interaction. Predicted networks of DDI can serve both as a starting point in designing Y2H screens, as well as the basis for predicting further PPI. Another direction that we are pursuing is to further validate the interaction domains through structure prediction and protein-protein docking.

## Acknowledgments

JAI and FM were partially supported by NSF grants ACI-0135195 and DBI-0450067. FM is recipient of a Kellogg graduate fellowship.

## References

1. A. Bateman et al. The Pfam Protein Families Database. *Nucl. Acids Res.*, 32, 2004.
2. K.C. Gunsalus et al. Predictive models of molecular machines involved in caenorhabditis elegans early embryogenesis. *Nature*, 436(7052):861–865, 2005.
3. M. Deng, S. Mehta, F. Sun, and T. Cheng. Inferring domain-domain interactions from protein-protein interactions. *Genome Res.*, 12:1540–1548, 2002.
4. C. Huang, F. Morcos, S. P. Kanaan, S. Wuchty, D. Z. Chen, and J. A. Izaguirre. Predicting protein-protein interactions from protein domains using a set cover approach. *IEEE/ACM Trans. on Comput. Biology Bioinformatics*, 2006. In press.
5. S. Li et al. A map of the interactome network of the metazoan C. elegans. *Science*, 303(5657):540–543, 2004.