BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 10(2)

Insilico identification of protein-protein interactions in Silkworm, *Bombyx mori*

Ramasamy Sumathy[1, 3], Ashwath Southekal Krishna Rao[1], Nalavadi Chandrakanth[2] & Velliyur Kanniappan Gopalakrishnan[3]*

[1]Bioinformatics centre, [2]Molecular biology Laboratory, Central Sericultural Research and Training Institute, Mysore, Karnataka, India; [3]Department of Biochemistry and Bioinformatics, Karpagam University, Coimbatore-641 021, Tamilnadu, India; Velliyur Kanniappan Gopalakrishnan – Email: vkgopalakrishnan@gmail.com; *Corresponding author

Received January 14, 2014; Accepted January 26, 2014; Published February 19, 2014

Abstract:
The Domesticated silkworm, *Bombyx mori,* an economically important insect has been used as a lepidopteran molecular model next only to *Drosophila.* Compared to the genomic information in silkworm, the protein-protein interaction data are limited. Therefore experimentally identified PPI maps from five model organisms such as *E.coli, C.elegans, D.melanogaster, H. sapiens, S. cerevisiae* were used to infer the PPI network of silkworm using the well-recognized Interlog based method. Among the 14623 silkworm proteins, 7736 protein-protein interaction pairs were predicted which include 2700 unique proteins of the silkworms. Using the iPfam interaction domains and the gene expression data, these predictions were validated. In that 625 PPI pairs of predicted network were associated with the iPfam domain-domain interactions and the random network has average of 9. In the gene expression method, the average PCC value of the predicted network and random network was 0.29 and 0.23100±0.00042 respectively. It reveals that the predicted PPI networks of silkworm are highly significant and reliable. This is the first PPI network for the silkworm which will provide a framework for deciphering the cellular processes governing key metabolic pathways in the silkworm, *Bombyx mori* and available at SilkPPI (http://210.212.197.30/SilkPPI/).

Keyword: protein-protein interaction, Interlog method, *Bombyx mori*.

Background:
Protein-protein interaction (PPI) provides valuable framework for comprehensive understanding of the biological processes and cellular function. Proteins are essential biomolecules that mostly function along with the other proteins rather than alone **[1].** Thus for any organism, the interactions of the proteins are needed to depict a better understanding of the biological process along with the proteins **[2].** PPI have been determined by many experimental methods such as Mass spectrometry **[3],** protein chips **[4],** binding reaction methods **[5],** two hybrid-based methods **[6]** etc and these experimental methods are expensive, labor intensive and time consuming. Several computational techniques have been developed for predicting the PPI which includes many concepts, approaches and data types such as gene fusion **[7],** gene expression profiles **[8],** phylogenetic profile **[9],** conservation of gene neighborhood **[10],** correlated mutation **[11],** and gene ontology annotation **[12].** Some approaches use the known PPI experimental information and supervised machine learning methods **[13,14]** such as support vector machine, Bayesian network and random forest method for predicting the PPIs. Moreover it includes domain-domain interaction **[15],** amino acid composition **[16],** conjoint triad feature **[17],** physiochemical and structural descriptor **[18].** Recent studies have shown the interactions between proteins of bidensovirus BmDNV-Z and specific host midgut proteins of *Bombyx mori* using yeast two-hybrid (Y2H) system **[19].**

Of late, the interlog method is being widely used to predict the protein-protein interactions in various organisms. In the rice blast fungus, *Magnaporthe grisea*, PPI pairs were predicted using the interlog method to study the pathogenecity and secreted

protein interactions **[20].** Recently swine interactomes were predicted using the interlog method, domain-motif interactions from structural topology, D-MIST and the Motif-motif interactions from structural topology **[21].** Interlog method uses the orthology information between the model organisms for predicting the PPI of the organisms and widely implemented for various organisms and proved to be reliable method **[22].**

Silkworm, *Bombyx mori* is an important domesticated Lepidopteran insect due to its primary role in silk production and also as a model organism for biochemical, genetic and genomic studies in insects, next only to *Drosophila* due to strong advantages for experimental research, such as rapid development with ease of rearing in the laboratory, short life cycle, tractability, small adult size etc. Despite this importance, large-scale protein-protein interaction mapping projects have not been implemented in silkworm and yet to be explored. In this study an attempt has been made to predict the PPI network of *Bombyx mori* using a well-recognized computational method ie interlog method.

Methodology:

Data Collection

The Silkworm, *Bombyx mori* protein sequences were acquired from the SilkDB database which contains 14623 sequences **[23].** The PPI network of *Bombyx mori* were inferred using the experimentally verified existing PPI maps of the five model organisms namely *C. elegans, D. melanogaster, E. coli, H. sapiens* and *S. cerevisae.* The protein sequences of *C. elegans, D. melanogaster, E. coli, S cerevisae* were downloaded from the Entrez genome database which contains of 22894, 23948, 4038, 6717 protein sequences respectively **[24].** The information about the protein-protein interactions of the model organisms except human were obtained from the database of Interacting Proteins, DIP which provides the experimentally determined 5112, 23484, 12894, 25233 protein-protein interactions of the said model organisms respectively **[25].** Then 30046 human protein sequences and the 14276 human PPIs were downloaded from the human protein reference database, HPRD, which provides information regarding the interaction networks of human proteome **[26].** The domain informations were collected from the Pfam database **[27]** and the Hmmer software was used to annotate the silkworm protein sequences by utilizing the Pfam domains. The microarray data and the gene ontology annotation of the silkworm, *Bombyx mori* were obtained from the BmMDB **[28]** and SilkDB database. Orthologs of *Bombyx mori* proteins in model organisms were predicted by using the InParanoid program, which uses the pairwise similarity scores for constructing the orthology group and then these orthologs were grouped into a likely co-orthologs group **[29].**

PPI Prediction

The homologous sequences of *Bombyx mori* were obtained from each model organisms using the BLAST program **[30].** The orthologous sequences of the silkworm were predicted from the model organisms using the InParanoid program and clustered into groups. For the entire silkworm genome, the interaction network was constructed by assuming that any pair of silkworm proteins are interacting if their orthologs from any one of the model organisms that has shown the experimentally verified interaction and thus it was considered as an interacting pair. Further, interaction score was assigned for each predicted PPI pair using the sequence similarity bit score and the number of instances that the protein-protein interaction pair occurred in all the model organisms by following the same strategy of previous studies **[31, 20].** Here *s(a,a`)* and *s(b,b`)* are the sequence similarity bit scores between a and a', b and b' respectively. *N* is the total number of instances occurred in all the model organisms.

$$\text{Interaction Score} = \sum_{i=1}^{N} \ln\left(S(a,a') \chi S(b,b')\right)$$

PPI Verification

The computationally predicted protein-protein interactions were verified by various techniques. Here the most popularly used techniques such as domain-domain interactions and the gene expression data were applied. In the first verification, the predicted interaction pairs from the silkworm sequences were mapped into domains. The sequences have been annotated into domains with the E-value cut off 0.01 using the HMMER program with the default settings utilizing the Pfam database. The domain-domain physical interactions were downloaded from the iPfam, the protein domain interactions database **[32].** Then the predicted PPI pairs which are associated with the experimentally verified domain-domain interactions were checked. Moreover, to facilitate the comparison, 1000 PPI networks were selected randomly with replacement and related to the experimentally verified domain-domain interactions. At last, for assigning the quality of the prediction method, the percentage of the Pfam interacting domain pairs associated with the predicted PPI pair and randomized PPI pair were calculated and finally the statistical significance of the predicted results were determined. In the second verification, the microarray data of the *Bombyx mori* were collected from the BmMDB database. Each of the predicted protein-protein interaction pair of the silkworm was mapped with the respective microarray data and the number of proteins and PPI interaction pairs were computed. Then the average absolute value of PCC and the p-value between the predicted protein interacting pairs of the silkworm were calculated. Then 100 randomized networks were selected with the replacement and compared the average absolute PCC value of predicted PPI and the 100 randomized networks.

The predicted PPI network of the silkworm were analyzed to calculate the nodes, edges, network radius, network diameter, average number of neighbour, characteristic path length, betweenness centrality, closeness centrality using the Cytoscape program which is an open source software project, providing framework about the biomolecular interaction networks **[33].** The 2700 unique proteins were represented as nodes and the interactions between the two nodes were represented as edges. The degree of the node represents the number of the interactions of the particular node. This interaction framework is essential for understanding the topological behavior of the network. The PPI network was grouped into cluster using the Cfinder program which uses the clique percolation method **[34]** For each community, overrepresented Gene ontology (GO) term was assigned by analyzing GO enrichment of biological process which were found using Fischer exact test followed by False Discovery Rate (FDR) correction.

Results:

By using of InParanoid program with the default settings, the orthologous between the *Bombyx mori* and each of the model organisms were obtained. In Silkworm, totally 7736 protein-protein interaction pairs were predicted using the Interlog method, among which 2700 proteins were unique. The number of protein-protein interactions predicted from each of the model organism's namely *C. elegans, D. melanogaster, E. coli, H. sapiens* and *S. cerevisae,* were, 422, 2688, 114, 557, 4184, respectively and the data is presented in **Table 1 (see supplementary material)**. In the first validation technique which relies on the iPfam database, the predicted 7736 PPIs of the silkworm proteins were mapped into their respective domains. Totally 2349 proteins were assigned with the Pfam domains among the 2700 unique proteins of the silkworm PPIs. Here 6553 PPI pairs were mapped with Pfam domains among the 7736 PPIs which were constructed. Around 625 PPIs of the predicted silkworm PPIs were associated with the iPfam domain-domain interactions. Moreover for comparison purpose we created 1000 random network, in which samples were selected with replacement from the 14623 proteins of the *Bombyx mori*. In each random network, we constructed 7736 PPI pairs and counted the number of interactions which were associated with iPfam interacting domains.

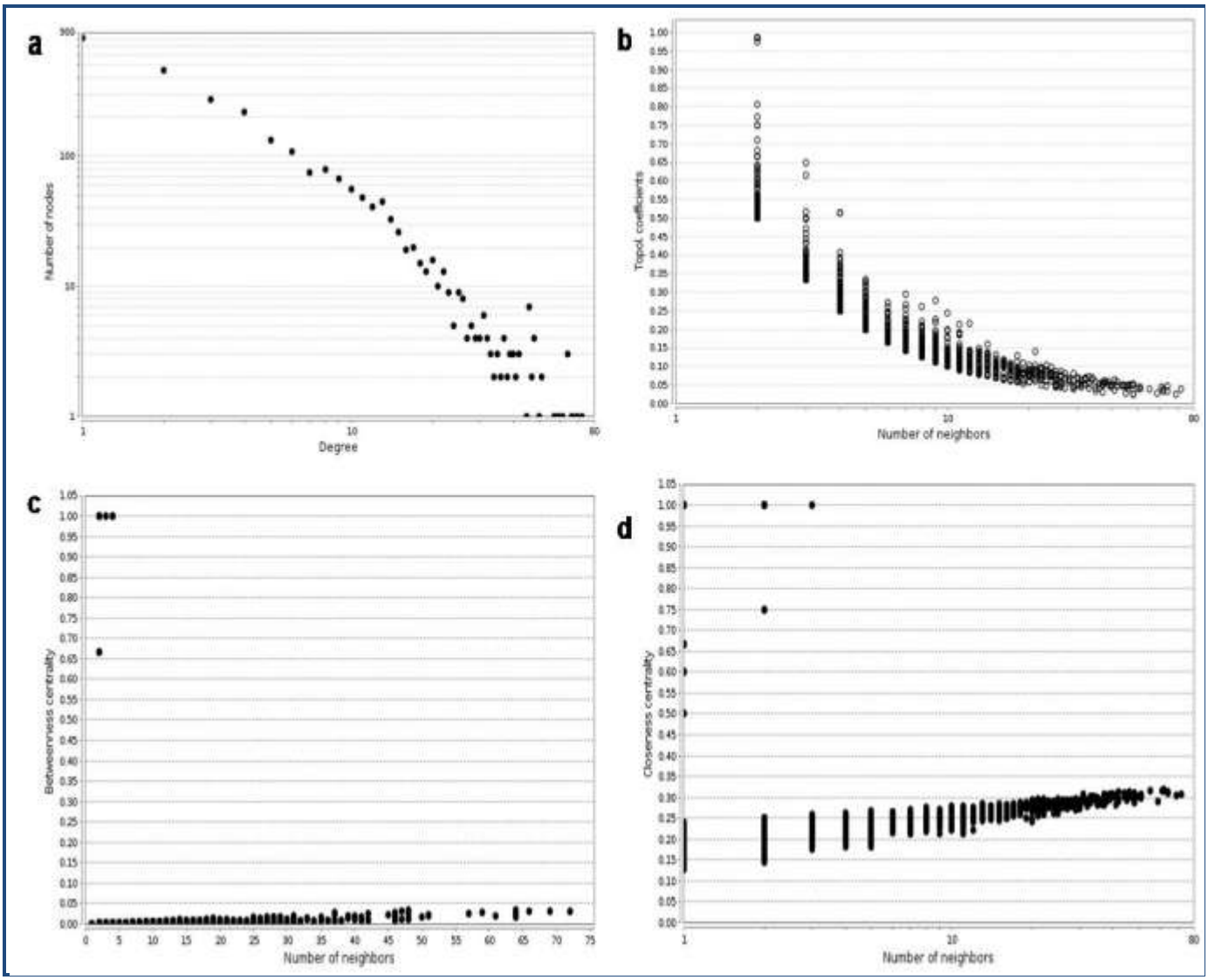

Figure 1: The properties of the silkworm PPI network: **a)** Degree distribution; **b)** Topological coefficient; **c)** Betweenness centrality; **d)** Closeness centrality.

In the second validation technique which is based on the microarray data of the silkworm, after removal of the self-interaction PPI pairs, there were 7434 non-self-interaction PPI pairs of the silkworm which were assigned with their respective microarray data. Altogether 6643 PPIs were mapped with the microarray data, among which 2511 unique proteins have their respective gene expression data. Then the average Pearson correlation coefficient was calculated between the expression data of interacting pairs which was found to be 2.29. By using the same method, 1000 random networks were created and each network consists of 7434 PPIs pairs and the average absolute PCC value among the interacting pairs of the random networks were computed which is shown in **Table 2 (see supplementary material)**. The network analysis and the visualization were carried out on the silkworm PPI network using the Cytoscape program and the information regarding the topological

properties of the network such as nodes, network radius, network diameter, average number of neighbour, characteristic path length and clustering coefficient are shown in **Table 3 (see supplementary material)** and the degree distribution, between's centrality, closeness centrality shortest path are depicted in the **Figure 1.**

Discussion:

The protein-protein interaction data of the model organisms were collected from the DIP and HPRD which are experimentally verified and manually curated databases so that the quality of the data is high when compared to any other database and to reduce false positive prediction and increase the accuracy. Maximum numbers of the silkworm PPI were predicted from the *S. cervisease* and the *Drosophila* using the interlog method as shown in **Table 1**. Previous studies show that interlog approach is highly acceptable and reliable method as the interlog of interacting protein is found in many model organisms **[35].** Presently no information is available on the experimentally verified protein-protein interaction of the silkworm and therefore it is difficult to validate the predicted PPIs. Hence the predicted PPI of *Bombyx mori* were validated using the existing computational methods. Proteins have many functions in cellular processes which interact with one another leading to successful execution of biological events. Here the main idea is that the likely co-expressed genes might have similar functions and interact with each other. So we used the interacting Pfam domains and the gene expression data of the silkworm for the purpose of validation. These methods are very effective to validate the computationally predicted PPI of the silkworm. In the first validation method, 625 PPI pairs are predicted and associated with the interacting Pfam domains. In random network, the average number of PPI pairs sharing the Pfam interacting domains was found to be 8.97±0.399 which is much smaller than our predicted PPI network , however, still it is highly significant as the highest number of PPIs sharing random network was found to be 24. In the second validation method, the average PCC value of the predicted network is 0.29 and the average PCC value of the Random network was 0.23100±0.00042 which is highly significant.

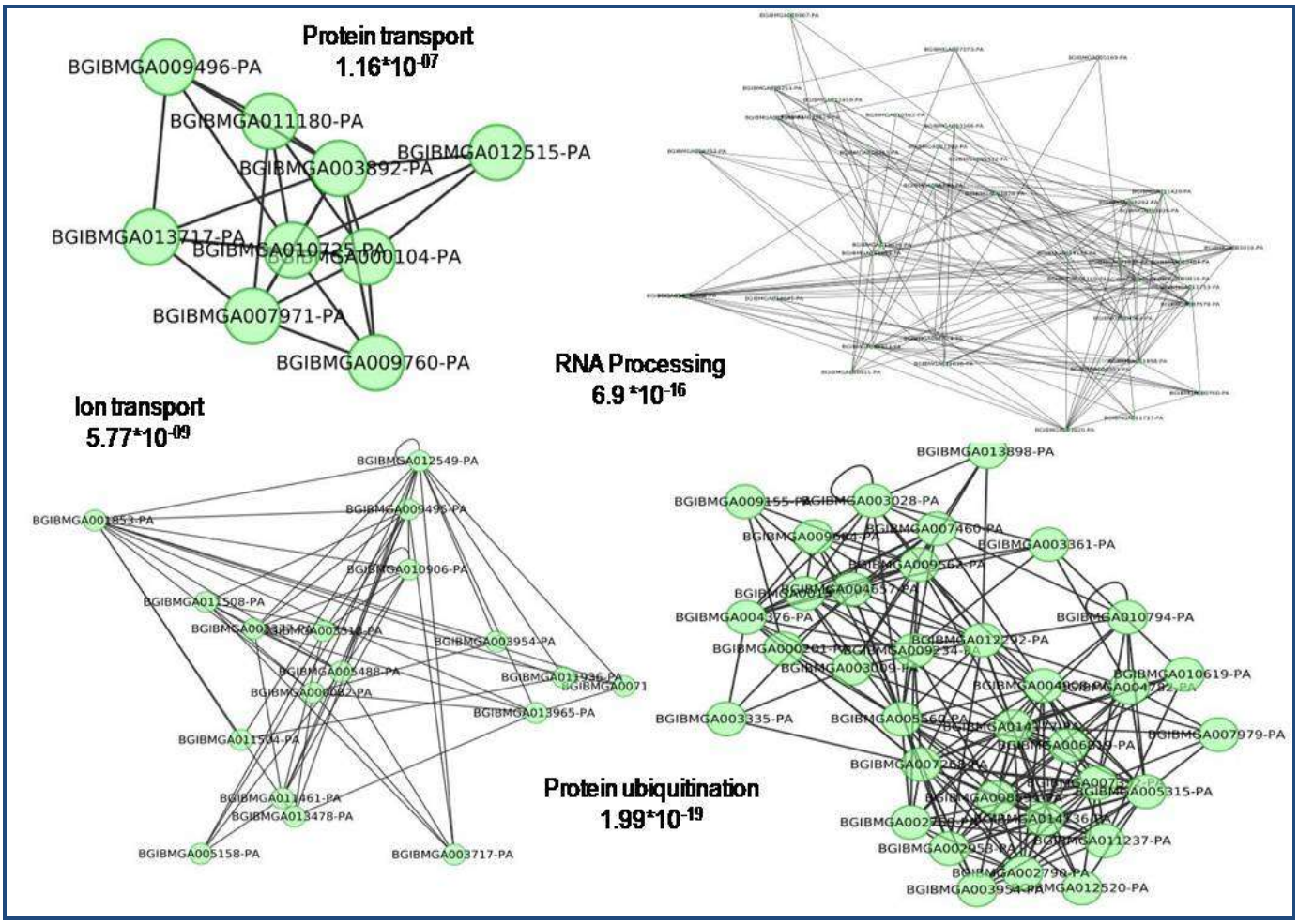

Figure 2: Clusters with GO enrichment term and their p-value. The clusters were visualized using the Cytoscape Program.

The PPI network has highly connected protein nodes known as hubs which have biological significance in the network architectures. In human, cancer-related proteins are more likely to act as hubs **[36].** However these hub proteins do not have important biological properties but these hubs contain more essential proteins when compared to non-hub proteins. Genome-wide studies show that deletion of hubs affects more when compared to the non-hub proteins **[37]** organization of PPI network **[38, 39].** In the predicted network, 35 hubs were found each having more than 40 interactions. It implies that our

predicted network follows a power law P(k) ~$k^{-1.823}$, R^2=0.940. It means that the predicted network has hubs with small number of highly connected proteins and thus it possesses the scale free property.

The degree of a node represents the number of edges i.e., interactions linked to *n* **[40].** It was shown that the few nodes have more degree and larger nodes have less degree obeying the power law **(Figure 1a).** The neighborhood connectivity of a node *n* is the average connectivity of all neighbors of *n* **[41] (Figure 1b).** The length of the shortest path between two nodes *n* and *m* is *L*(*n*,*m*). The shortest path length distribution gives the number of node pairs (*n*,*m*) with *L*(*n*,*m*) = *k* for *k* = 1,2,…. Network diameter is the maximum length of shortest paths between two nodes **[42, 43].** The betweenness centrality **[44]** *Cb(n)* of a node *n* is calculated as $Cb(n) = \sum s \neq n \neq t (\sigma st (n) / \sigma st)$, where *s* and *t* are nodes in the network different from *n*, σst denotes the number of shortest paths from *s* to *t*, and σst (*n*) is the number of shortest paths from *s* to *t* that *n* lies on **(Figure 1b).** The closeness centrality **[45]** *Cc(n)* of a node *n* is the reciprocal of the average shortest path length and computed $Cc(n) = 1 / avg(L(n,m))$, here *L*(*n*,*m*) is the length of the shortest path between two nodes *n* and *m.*

The closeness centrality of each node is a number between 0 and 1 **(Figure 1c**). The clustering coefficient *Cn* of a node *n* is defined as $Cn = 2en/(kn(kn-1))$, where *kn* is the number of neighbors of *n* and *en* is the number of connected pairs between all neighbors of *n* **[43,46]** and here clustering coefficient is 0.053. We can identify the protein function by means of studying the network clusters **[47].** By using Cfinder program which uses k-clique clustering method, the PPI networks were clustered into communities, here we selected the k value as 4. In order to understand the functions of each cluster of PPI network, we analyzed each cluster by GO enrichment of biological process at the depth of 4. The identified GO enrichment consists of RNA processing, ion transport, protein transport, protein ubiquitination etc (**Figure 2).**

Conclusion:

The present investigation has predicted totally 7736 protein-protein interactions among the 2700 silkworm proteins which include unique proteins using a well-known interlog method. The predicted PPI networks were validated by two computational methods and it shows that our network is more reliable. The reliability of the network has been clearly demonstrated by the result of validation methods using the iPfam domain interacting pairs and coexpression information. The silkworm protein-protein interaction data are publicly available at SilkPPI (http://210.212.197.30/SilkPPI/). It can be browsed using the SilkDb accession number which provides the details of the interaction proteins and their GO annotations, Pfam domains and nominal p-value of microarray data.

Acknowledgement:

We are thankful to the BTISNet, Dept. of Biotechnology, New Delhi, Central Sericultural Research and Training Institute, Mysore and Karpagam University, Coimbatore, for providing infrastructural facilities to carry out the work.

References:

[1] Cusick ME *et al. Hum Mol Genet.* 2005 **14**: R171 [PMID: 16162640]
[2] Legrain P *et al. Trends Genet.* 2001 **17:** 346 [PMID: 11377797]
[3] Ho Y *et al. Nature* 2002 **415**: 180 [PMID: 11805837]
[4] Zhu H *et al. Science* 2001 **293**: 2101 [PMID: 11474067]
[5] Lakey JH & Raggett EM, *Curr Opin Struct Biol.* 1998 **8**: 119 [PMID: 9519305]
[6] Fields S & Song O, *Nature* 1989 **340:** 245 [PMID: 2547163]
[7] Enright AJ *et al. Nature* 1999 **402:** 86 [PMID: 10573422]
[8] Ideker T *et al. Bioinformatics* 2002 **18**: S233 [PMID: 12169552]
[9] Goh CS *et al. J Mol Biol.* 2000 **299**: 283 [PMID: 10860738]
[10] Dandekar T *et al. Trends Biochem Sci.* 1998 **23**: 324 [PMID: 9787636]
[11] Pazos F & Valencia A, *Proteins* 2002 **47:** 219 [PMID: 11933068]
[12] Chou KC & Cai YD, *J Proteome Res.* 2006 **5**: 316 [PMID: 16457597]
[13] Chen XW & Liu M, *Bioinformatics* 2005 **21**: 4394 [PMID: 16234318]
[14] Rashid M *et al. Curr Protein Pept Sci.* 2010 **11**: 589 [PMID: 20887258]
[15] Ng SK *et al. Bioinformatics* 2003 **19:** 923 [PMID: 12761053]
[16] Dohkan S *et al. In Silico Biol.* 2006 **6**: 515 [PMID: 17518762]
[17] Shen J *et al. Proc Natl Acad Sci U S A.* 2007 **104**: 4337 [PMID: 17360525]
[18] Ogmen U *et al. Nucleic Acids Res.* 2005 **33**: W331 [PMID: 15991339]
[19] Bao YY *et al. FEBS J. 2013* **280:** 939 [PMID: 23216561]
[20] He F *et al. BMC Genomics.* 2008 **9:** 519 [PMID: 18976500]
[21] Wang F *et al. Proteome Sci.* 2012 **10**: 2 [PMID:22230699]
[22] Matthews LR *et al. Genome Res.* 2001 **11:** 2120 [PMID: 11731503]
[23] Duan J *et al. Nucleic Acids Res.* 2010 **38**: D453 [PMID: 19793867]
[24] Benson DA *et al. Nucleic Acids Res.* 2013 **41**: D36 [PMID: 23193287]
[25] Xenarios I *et al. Nucleic Acids Res.* 2000 **28:** 289 [PMID: 10592249]
[26] Mishra GR *et al. Nucleic Acids Res.* 2006 **34:** D411 [PMID: 16381900]
[27] Finn RD *et al. Nucleic Acids Res.* 2006 **34**: D247 [PMID: 163811856]
[28] Xia Q *et al. Genome Biol.* 2007 **8:** R162 [PMID: 17683582]
[29] Ostlund G *et al. Nucleic Acids Res.* 2010 **38**: D196 [PMID: 19892828]
[30] Johnson M *et al. Nucleic Acids Res.* 2008 **36**: W5 [PMID: 18440982]
[31] Jonsson PF & Bates PA, *Bioinformatics* 2006 **22:** 2291 [PMID: 16844706]
[32] Finn RD *et al. Bioinformatics* 2005 **21**: 410 [PMID: 15353450]
[33] Shannon P *et al. Genome Res.* 2003 **13**: 2498 [PMID: 14597658]
[34] Adamcsek B *et al. Bioinformatics* 2006 **22**: 1021 [PMID: 16473872]
[35] Chen PY *et al. PLoS Comput Biol.* 2008 **4:** e1000118 [PMID: 18654616]
[36] Kar G *et al. PLoS Comput Biol.* 2009 **5**: e1000601 [PMID: 20011507]
[37] He X & Zhang J, *PloS Genet.* 2006 **2:** e88 [PMID: 16751849]
[38] Albert R *et al. Nature* 2000 **406**: 378 [PMID: 10935628]
[39] Han JD *et al. Nature* 2004 **430:** 88 [PMID: 15190252]
[40] Chautard E *et al. Pathol Biol (Paris).* 2009 **57:** 324 [PMID: 19070972]
[41] Maslov S & Sneppen K, *Science* 2002 **296**: 910 [PMID:

11988575]
[42] Stelzl U *et al*. *Cell* 2005 **122**: 957 [PMID: 16169070]
[43] Watts DJ & Strogatz SH, *Nature* 1998 **393**: 440 [PMID: 9623998]
[44] Brandes U, *J Math Sociol.* 2001 **25**: 163
[45] Newman MEJ, *Social Networks* 2005 **27**: 39
[46] Barabasi AL & Oltvai ZN, *Nat Rev Genet.* 2004 **5**: 101 [PMID: 14735121]
[47] Pereira-Leal JB *et al. Proteins* 2004 **54**: 49 [PMID: 14705023]

Edited by P Kangueane
Citation: Sumathy *et al.* Bioinformation 10(2): 056-062 (2014)

BIOINFORMATION

open access

Supplementary material:

Table 1: Silkworm Protein-protein interaction

No	Source Organisms	No of Predicted PPI
1	*S. cerevisae*	4184
2	*E. coli*	114
3	*C. elegans*	422
4	*D. melanogaster*	2688
5	*H. sapiens*	557
	Total	7736*

Considering some of the PPI inferred from one or more organisms

Table 2: Validation of Silkworm Protein-protein interactions

	Predicted	Random
No. of PPI associated with iPfam interacting domains	625	8.97±0.399
Average PCC Value	0.29	0.231±0.00042

In validation methods, comparison between the predicted and random network clearly indicates that the predicted network is highly significant. We found that 43% of PPI has PCC >0.5 which shows the significance of the predicted interactome.

Table 3: The Silkworm predicted PPI Network properties

No	Property	PPI Network
1	Node	2700
2	Degree	7736
3	Clustering Coefficient	0.053
4	Network diameter	12
5	Network radius	1
6	Average number of neighbor	5.507
7	Characteristic path length	4.514
8	Shortest Path	89%