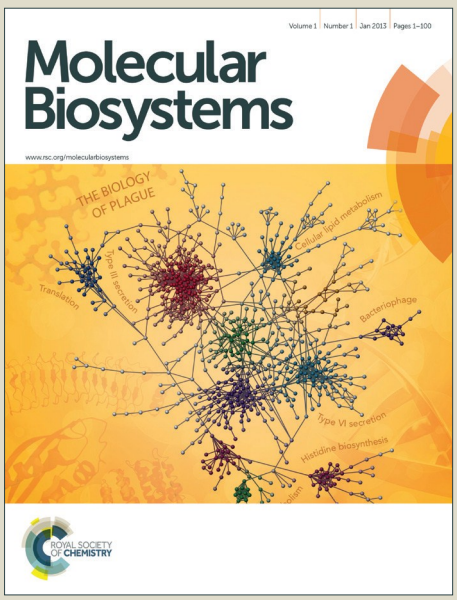
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

ARTICLE TYPE www. rsc.org/x XXXXXXXX

Integrative *C. elegans* Protein-Protein Interaction Network with Reliability Assessment Based on A Probabilistic Graphical Model†

Xiao-Tai Huang,[a]‡ Yuan Zhu,*[b,a]‡ Leanne Lai Hang Chan,[a] Zhongying Zhao [c] and Hong Yan[a]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

In *Caenorhabditis elegans*, a large number of Protein-Protein Interactions (PPIs) are identified by different experiments. However, a comprehensive weighted PPI network, which is essential for signaling pathway inference, is not yet available in this model organism. Therefore, we firstly construct an integrative PPI network in *C. elegans* with 12951 interactions involving 5039 proteins from seven molecular interaction databases. Then, a Reliability Score based on a Probabilistic Graphical Model (RSPGM) is proposed to assess PPIs. It assumes the random number of interactions between two proteins comes from the Bernoulli Distribution to avoid multi-links. The main parameter of RSPGM score contains a few latent variables which can be considered as several common properties between two proteins. Validations on high-confidence yeast datasets show that RSPGM provides more accurate evaluation than other approaches, and the PPIs in the reconstructed PPI network have higher biological relevance than that in the original network in terms of gene ontology, gene expression, essentiality and the prediction of known protein complexes. Furthermore, this weighted integrative PPI network in *C. elegans* is employed on inferring interaction path of the canonical Wnt/β-catenin pathway as well. Most genes on the inferred interaction path have been validated to be Wnt pathway components. Therefore, RSPGM is essential and effective for evaluating PPIs and inferring interaction path. Finally, the PPI network with RSPGM scores can be queried and visualized on a user interactive website, which is freely available at http://rspgm.bionetworks.tk/.

1 Introduction

Signaling pathway is an essential process in living organism, receiving extracellular or cytoplasmic signal and then triggering downstream signal transduction, which modulates gene expression and cell function. The knowledge of different kinds of pathways can reveal biological function or provide suggestion of disease therapy[1].

Unfortunately, although several pathways have been studied extensively, the structure and function of most pathways are not well understood. Because signaling pathway is complicated involving different molecules contacting with each other via Protein-Protein Interactions (PPIs) or Protein-DNA Interactions (PDIs), it is time-consuming to detect molecular regulatory relationships through biological experiments, such as gene knockout or RNAi. Therefore, it is possible and necessary to infer pathway by computational methods based on molecular interaction data.

Several computational methods have been proposed for pathway inference recently[2–5]. Most of them require a weighted molecular interaction network, called background network, as an input of the algorithm. The background network is generally constructed from PPIs and PDIs data. Most pathway inference methods are performed on yeast because of the availability of its weighted PPI networks currently[2,6]. However, in *Caenorhabditis elegans* (*C. elegans*), there is still not a comprehensive weighted PPI network available[7]. Therefore, it is necessary to construct a PPI network of *C. elegans*, and assign the reliability score for each PPI.

Protein-Protein Interactions can be identified via high-throughput and small-scale experimental techniques or be predicted from computational methods by using different types of data, such as sequence, expression and binding data, or three-dimensional structural data[8,9]. Several different popular biological databases have collected abundant PPIs of *C. elegans*, such as Database of Interacting Proteins (DIP)[10], Biological General Repository for Interaction Datasets (BioGRID)[11], IntAct Molecular Interaction Database (IntAct)[12], Molecular Interaction database (MINT)[13], WormBase[14], Worm Interactome version 8 (WI8)[15] and GeneOrienteer[16]. However, none of them contains the relative comprehensive

† Electronic Supplementary Information (ESI) available: http://rspgm.bionetworks.tk/. See DOI: 10.1039/b000000x/

[a] *Department of Electronic and Engineering, City University of Hong Kong, Hong Kong.*
[b] *School of Automation, China University of Geosciences, Wuhan, China. Fax: +852 3442 0562; Tel: +852 3442 4889; E-mail: zhuyuan7@mail2.sysu.edu.cn; zhuyuan2015@yeah.net*
[c] *Department of Biology, Faculty of Science, Hong Kong Baptist University, Hong Kong.*
‡ These authors contributed equally to this work.

PPIs information. For instance, the interaction between *mex-6* and *emb-9* is recorded in BioGRID, IntAct and MINT, while the interaction between *zag-1* and *odr-7* can only be retrieved in GeneOrienteer and WormBase. Therefore, construction of a comprehensive PPI network database of *C. elegans* is urgent and necessary.

Many computational methods have been developed to assess the reliability of the data. These methods can be approximately divided as three classes: (1) Multiple data integration based methods[17–19]; (2) Network topology based methods[20–25]; (3) Model based methods[26–28]. Multiple data integration based methods work effectively but much more rely on the prior knowledge of individual protein. Network topology based methods and model based methods are the most state-of-the-art evaluation approaches, recently. A Probabilistic Graphical Model (PGM) has been established to describe PPI networks in terms of a random process that generates the networks[29,30]. Several works demonstrated that PGMs can be widely applied to discover protein complex[28,31,32], explore biology network[33] and assess PPIs[27], *etc.*. Motivated by the wide applications of PGMs in PPI network analysis, this paper further explores its potential in assessing new established integrative and comprehensive PPI network of *C. elegans*.

2 Methods

Similar to Zhu *et al.*'s previous work[27], we assume that there are several latent properties between two interacting proteins. These latent properties could be GO annotation terms, gene expression, sequence, location or any other functional, physical and biochemical properties of the protein. Then, a reliability score for protein pairs is defined by accumulating protein propensities on the common latent properties, which can be estimated by a probabilistic graphical model.

2.1 Reliability Score for Protein Pairs

Based on our assumption, $s_i = (s_{i\ell})$ and $s_j = (s_{j\ell}) \in \mathbb{R}^m$ are used to describe protein properties on m latent variables for protein v_i and v_j, respectively. $0 \le s_{i\ell}, s_{j\ell} \le 1$ means the propensity of proteins v_i and v_j on the ℓ-th latent variable. Suppose variables $d_i, d_j \in \mathbb{R}$ are the ability of protein v_i and protein v_j generating edges in the network, respectively. Thus, we obtain the reliability score r_{ij} between protein v_i and protein v_j as the following form.

$$r_{ij} = 1 - \exp(-(\langle d_i s_i, d_j s_j \rangle) + \text{eps}), \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of vectors. eps means the floating-point relative accuracy in MATLAB. Higher $\langle d_i s_i, d_j s_j \rangle$ indicates that protein v_i and protein v_j share more latent properties, and have larger interacting probability. Function $f(x) = 1 - \exp(-(x + \text{eps}))$ is used to map the output argument from $[0, +\infty)$ to $(0,1)$. By using this mapping function, $0 < r_{ij} < 1$ ensures that it makes sense when considering it as a parameter of the Bernoulli Distribution on the one hand, and it normalizes the reliability scores on the other hand.

2.2 A probabilistic graphical model

In this method, a PPI network is represented by an undirected graph $G(V,E)$, i.e. vertex set including each protein as a vertex $V = \{v_1, v_2, \cdots, v_n\}$, and edge set $E = \{(v_i, v_j) \mid$ there is an interaction between protein v_i and $v_j, 1 \le i, j \le n\}$. The symmetric adjacent matrix is denoted as $W = (w_{ij}) \in \mathbb{R}^{n \times n}$, where $w_{ij} = 1$ if $(v_i, v_j) \in E$ else $w_{ij} = 0$. The probabilistic graphical model can be described by the joint likelihood function over all variables as below.

$$P(W,S,B,D) = P(W|S,D)P(S|B)P(D|\gamma)P(B), \tag{2}$$

where $S = (s_{i\ell}) \in \mathbb{R}^{n \times m}$ is the propensity matrix, $D = (d_i) \in \mathbb{R}^n$ is the protein linkage ability vector of all n proteins involving in the PPI network. $B = (\beta_\ell) \in \mathbb{R}^m$ is the parameter vector of S. $P(W|S,D)$ is the probability of generating interaction w_{ij} between protein i and protein j in a PPI network. As it is shown above, w_{ij} is binary (0 or 1), which is supposed to follow the Bernoulli Distribution with parameter $p_{ij} = r_{ij}$. Similar to[27,28], we also assume that each $s_{i\ell}$ comes from an exponential distribution with rate parameter β_ℓ. Considering the scale-free property of PPI networks, the degree distribution of d_i in the PPI network approximates to a power law with a hyperparameter γ. Mathematically, the components of (2) can be described in detail as follows.

$$w_{ij} \sim \mathrm{B}(1, p_{ij}),$$

namely, $P(W|S,D)$ presented below is the probability of generating interaction w_{ij} between protein v_i and v_j in a PPI network.

$$\begin{aligned} &P(W|S,D) \\ &= \prod_{i,j=1}^{n} P(w_{ij}|s_{i,\ell}, d_i) = \prod_{i,j=1}^{n} p_{ij}^{w_{ij}} (1-p_{ij})^{1-w_{ij}} \\ &= \prod_{i,j=1}^{n} (1 - \exp(-(\langle d_i s_i, d_j s_j \rangle + \text{eps})))^{w_{ij}} (\exp(-(\langle d_i s_i, d_j s_j \rangle) + eps)^{1-w_{ij}} \end{aligned}$$

For protein v_i and latent variable ℓ, draw protein-propensity score:

$$s_{i\ell} \sim \mathrm{Exp}(\beta_\ell).$$

Namely,

$$P(S|B) = \prod_{i=1}^{n} \prod_{\ell=1}^{m} P(s_{i\ell}|\beta_\ell) \tag{3}$$

$$= \prod_{i=1}^{n} \prod_{\ell=1}^{m} \beta_\ell \exp(-\beta_\ell s_{i\ell}). \tag{4}$$

$$P(d_i|\gamma) \propto d_i^{-\gamma},$$

where, γ can be implemented by robust linear regression using *robustfit(X,Y,'bisquare',4.685)* provided by Matlab command with input

$$X = -\log(D) \quad \text{and} \quad Y = \log P(D|\gamma). \tag{5}$$

In summary, we can get the objective function as follows.

$$\begin{cases} \min\limits_{S,D,B} & -\sum\limits_{i,j=1}^{n} w_{ij}\log(1-\exp(-(\langle d_i s_i, d_j s_j\rangle + \text{eps}))) \\ & +\sum\limits_{i,j=1}^{n}(1-w_{ij})(\langle d_i s_i, d_j s_j\rangle + \text{eps}) - n\sum\limits_{\ell=1}^{m}\log\beta_\ell \\ & +\gamma\sum\limits_{i=1}^{n}\log d_i + \sum\limits_{i,\ell}\beta_\ell s_{i\ell}, \\ s.t. & S \geq 0, D \geq 0. \end{cases} \tag{6}$$

2.3 Parameter estimation

To solve the non-negative constrained optimization problem, we use the multiplicative updating rules, which show a good compromise between speed and ease of implementation, to alternately update the model parameters S, D and B. 't' denotes the transpose of matrix while '$\mathbf{1}_n$' denotes the column vector of ones with n length. Similar to[27,28], we can obtain the following updating formulae for parameter S, D, B, respectively.

$$s_{i\ell} \leftarrow s_{i\ell} * \frac{\sum_j w_{ij} * (D*D^t)_{ij} * s_{j\ell}}{\frac{1-\exp(-((D*D^t)_{ij} * (S*S^t)_{ij} + \text{eps}))}{(D*D^t*S + 0.5*\mathbf{1}_n * B^t)_{i\ell}}} \tag{7}$$

$$d_i \leftarrow d_i * \frac{\sum_j w_{ij} * (D*D^t)_{ij} * d_j}{\frac{1-\exp(-((D*D^t)_{ij} * (S*S^t)_{ij} + \text{eps}))}{S*S^t*D + \gamma/d_i}} \tag{8}$$

$$\beta_\ell = \frac{n}{(\mathbf{1}_n^t * S)_\ell}. \tag{9}$$

2.4 Main algorithm

The main algorithm of the new proposed assessment of Reliability Score based on a Probabilistic Graphical Model (RSPGM) is presented in Algorithm 1. Where, $\mathbf{0}_m$ denotes the column vector of zeros with m length. '$\circ$' denotes the Hadamard product of two matrix with the same size. For example, $A = (a_{ij}), B = (b_{ij}) \in \mathbb{R}^{n\times m}$, thus $(A \circ B)_{ij} = a_{ij}b_{ij}$.

Algorithm 1 RSPGM

Input: $m = 500$, W, S, D, B, $T = 300$, $\sigma = 0.01$.
Output: Reliability score matrix R for PPI network.

1: Initialize S with random $n \times m$ matrix, D with the $\mathbf{1}_n$, B with the $\mathbf{0}_m$ initialization.
2: Integrate the *C. elegans* PPI network, obtain the adjacent matrix W.
3: Estimate γ by equation (5).
4: Iterate S, D, B by equation (7), (8), (9), respectively.
5: **Until** Iteration count is larger than T or $\|S^{(T+1)} - S^{(T)}\| < \sigma$.
6: **Repeat** step 1-5 50 times, the final result produces the parameters with the minimum objective function in (6).
7: $R = 1 - \exp(-((D*D^t) \circ (S*S^t) + \text{eps} * \text{ones}(n,n)))$.

3 Results

3.1 Databases to navigate scored PPI network

Since PPIs data from different molecular interaction databases are distinct, it is necessary to construct a relative comprehensive PPI network in *C. elegans* for further study. Here, we integrate PPIs data of *C. elegans* from seven free available databases, i.e. DIP, BioGRID, IntAct, MINT, WormBase, WI8 and GeneOrienteer. The details are presented in Table 1. We then filter the PPIs data in terms of four criteria: 1) physical interactions which belong to MI:0914 (association) type from Molecular Interaction (PSI MI 2.5); 2) no self-interactions (loops); 3) no repetitive interactions; 4) not containing interactions whose genes are not protein-coding, e.g. pseudogene, transposon or miRNA. The statistics of the original and filtered databases are discussed in Supplementary 1 Table 1.

Table 1 The versions and corresponding references of the seven selected databases.

Database	Reference	Version
DIP	Salwinski *et al.*[10], 2004	Celeg20141001
BioGRID	Chatr-aryamontri *et al.*[11], 2013	3.2.119
IntAct	Kerrien *et al.*[12], 2011	2014-12-18
MINT	Licata *et al.*[13], 2012	2012-10-29
WormBase	Harris *et al.*[14], 2014	WS245
WI8	Simonis *et al.*[15], 2009	WI8
GeneOrienteer	Zhong and Sternberg[16], 2006	v2.25

According to the filter criterion, we construct an integrative protein-protein interaction network of *C. elegans* which contains 5039 nodes involving in 12951 PPIs, shown in Supplementary 1 Figure 1. The intersection numbers and overlapping rates of any two filtered databases from the seven selected databases are provided in Supplementary 1 Table 2 that shows low overlapping rate between most any two filtered databases. This indicates interactions are partially recorded in different specific databases.

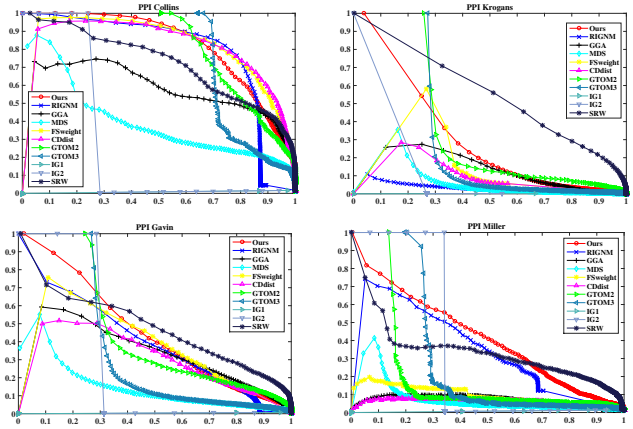
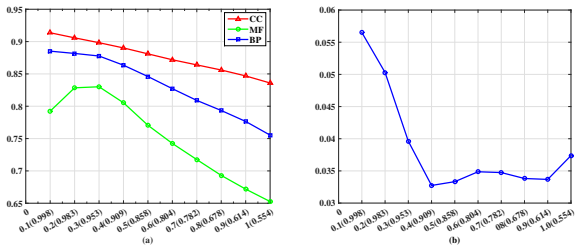


Fig. 1 The PR curves of eleven different methods on the four yeast datasets ($PPI_{Collins}$, $PPI_{Krogans}$, PPI_{Gavin} and PPI_{Miller}). The x-axis presents the recall while the y-axis shows the precision.



The consistency between RSPGM score and the GO semantic similarity.

The consistency between RSPGM score and the sequence similarity.

Fig. 2 The consistency between RSPGM score and the GO semantic, sequence similarity. The x-axis is the coverage of the PPI network. The averages of RSPGM score of the corresponding coverage of the PPI network are presented on the bottom of x-axis. (a) The y-axis is the average of the GO semantic similarity with the descending order of RSPGM score by increasing the coverage ratios of the PPIs in three GO domains: CC, MF, BP. (b) The y-axis is the average of the sequence semantic similarity with the descending order of RSPGM score by increasing the coverage ratios of the PPIs.

3.2 Yeast PPI Networks

The yeast PPI networks are download from BioGRID (version 3.2.119). Four yeast PPI subnetworks filtered by different techniques are used for evaluation. Collins dataset[34] (short for $PPI_{Collins}$), Krogans dataset[35] (short for $PPI_{Krogans}$) and Gavin dataset[36] (short for PPI_{Gavin}) are detected by TAP-MS technique. The largest connected components of physical interactions of these subnetworks are 1002 proteins with 8313 PPIs, 2527 proteins with 6985 PPIs and 1359 proteins with 6541 PPIs, respectively. Miller dataset[37] (short for PPI_{Miller}) is detected by PCA technique, in which the largest connected component of physical interactions with 513 proteins and 1947 PPIs. Since $PPI_{Collins}$ is high-confidence, we employ it to evaluate the GO similarity and sequence consistency and compare the biological relevance and the accuracy of the prediction of known protein complexes for PPI groups.

3.3 Effectiveness validation of the reliability score

In this section, we first compare RSPGM score with other scores obtained using existing methods on the four yeast datasets by PR curve which presents recall against precision. Secondly, we validate the consistency between the RSPGM score and GO semantic similarity and sequence similarity, respectively. Moreover, we evaluate the functional relevance of the original and reconstructed PPI networks on several types of sources, including gene ontology, gene expression and essentiality analysis. Finally, we investigate and compare the accuracy of protein complex prediction between original and reconstructed PPI network.

3.3.1 Comparison with other reliability scores There are two differences between RIGNM[27] and RSPGM: (1) We assume that the random number of interactions between two proteins comes from the Bernoulli distribution instead of Exponential distribution, which is found to be more suitable to evolve the PPI network and avoid multi-links. (2) The newly proposed score is scaled into (0,1), which makes sense when considering it as a parameter of the Bernoulli Distribution and is facilitated to compare with other methods. We compare RSPGM with the state-of-the-art methods that were described in[27] and the similar type methods including Interaction Generality (IG1)[24], modified IG1 (IG2)[25] and RWS[20]. The parameter settings of RSPGM and other methods refer to Algorithm 1 in Section 2 and Section 3.2.2 in[27], respectively. To validate the effectiveness of RSPGM, we plot the precision-recall (PR) curves for RIGNM, MDS, GGA, CDdist, FSweight, GTOM, IG1, IG2 and RWS methods on the four yeast datasets. The results are presented in Fig. 1. As shown, RSPGM performs better than other methods on the four yeast datasets except $PPI_{Krogans}$ and PPI_{Gavin}. However, the PR-AUC of RSPGM is only 0.16 and 0.03 less than RWS on $PPI_{Krogans}$ and PPI_{Gavin}, respectively (see Supplementary 1 Table 3). Our newly proposed method is much more appropriate than RIGNM by theory, and the performance is as good as RIGNM by experiment validation. Therefore, the new reliability score is effective to assess the PPIs.

3.3.2 Consistency validation According to the "guilt-by-association" principle[38], the interacting proteins should share the same functional terms and higher sequence similarity. We use R package "GOSemSim" (mgeneSim)[39] to calculate the GO semantic similarity between two proteins by Wang's method[40]. We also employ the local BLAST method[41], blastp (BLAST+ version 2.2.30), to calculate the

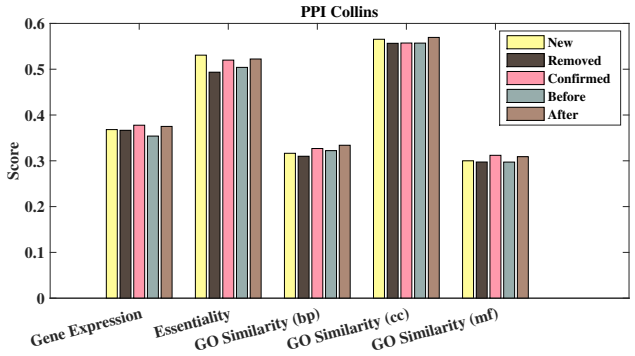


Fig. 3 Gene expression PCC, co-essentiality percentage and three branches' GO-based similarity of different PPIs' groups generated from RSPGM for $PPI_{Collins}$.

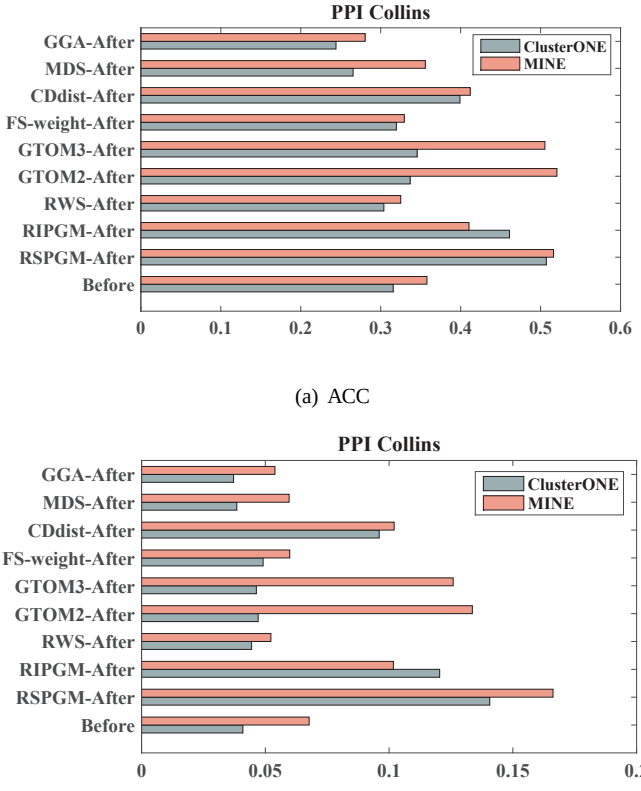


Fig. 4 The original PPI network ('Before') and the reconstructed counterpart ('After') of $PPI_{Collins}$ are evaluated by ClusterONE and MINE cluster algorithms for protein complex prediction in terms of accuracy and Jaccard value on the MIPS known complexes. (a) The results of ACC. (b) The results of Jaccard.

e-value between two proteins. Then the e-value is converted between 0 and 1 by formula $f(x) = \exp(-x)$ to represent sequence similarity. The more alike the interacting protein pairs, the higher the reliability score, GO semantic similarity and sequence similarity. In order to validate the consistency between GO, sequence similarity and reliability score, we order all the interacting protein pairs of $PPI_{Collins}$ by RSPGM score in descending index, and calculate the average of the corresponding GO semantic similarity and sequence similarity by increasing the coverage ratios of the PPIs. The details are illustrated in Fig. 2. For example, in CC process, the average GO similarity of the top 10% highest RSPGM scores is about 0.914. The average GO similarity of the top 20% coverage of the PPIs is about 0.906. The average GO similarity of the 30% to 100% coverages of the PPI network is from 0.898 to 0.83. As shown in Fig. 2, the higher the RSPGM score, the higher the GO similarity and sequence similarity. Although the trend of sequence similarity (Fig. 2(b)) is not strictly monotonically decreasing, the highest average sequence similarity is obtained by top 10% highest RSPGM scores. Above all, the RSPGM score meets the "guilt-by-association" principle, and it is a suitable reliability score to assess the PPIs.

3.3.3 Functional relevance evaluation We evaluate the functional relevance of the original and reconstructed PPI networks based on several types of sources, including gene ontology, gene expression and essentiality analysis. For convenience, the PPIs presented in the original and reconstructed networks are called 'Before' and 'After' respectively. The PPIs presented in 'After' but not in 'Before' are called 'New'. The PPIs presented in 'Before' but not in 'After' are called 'Removed'. The PPIs presented both in 'Before' and 'After' are called 'Confirmed'. We use a PPI network reconstruction method similar to Lei *et al.*'s approach[20]. Namely, the selected threshold is used to keep the number of PPIs in the reconstruction network the same as that in the original network.

We then calculate the GO semantic similarity, Pearson correlation coefficient of gene expression, and the co-essentiality percentage for PPIs in 'Before', 'After', 'New', 'Removed' and 'Confirmed' generated from RSPGM on $PPI_{Collins}$. The results are shown in Fig. 3. Here, we use profiles to characterize the expression dynamics for 3552 significant periodic genes over 36 time points. The raw data are available on Gene Expression Omnibus (GEO)[42] with the accession number GSE3431[31]. Additionally, the yeast essential gene list is retrieved from the *Saccharomyces* Genome Database[43]. The essentiality score is calculated by the percentage of the number of PPI, in which two proteins have the same essentiality (two interacting proteins are in essential list or not in essential list simultaneously). As shown in Fig. 3, the 'After' groups has a higher functional relevance than 'Before' group on gene expression, GO similarity and essentiality. Moreover, the 'Confirmed' group has almost the highest functional relevance

score compared with other groups. The functional relevance score of the 'Removed' group is lower than the 'New' group. We also evaluate the functional relevance of our method and other comparative methods on $PPI_{Collins}$ and $PPI_{Krogans}$. The results are demonstrated in Supplementary 1 Figure 2-5.

3.3.4 Protein complex prediction In order to investigate whether the reconstructed PPI network can improve the performance of prediction of protein complexes, we apply ClusterONE[44] and MINE[45] clustering algorithms to the 'Before' and 'After' PPI networks generated from different methods to evaluate the prediction of protein complexes in terms of accuracy (ACC) (see Fig. 4(a)) and Jaccard coefficient (see Fig. 4(b)). Here, we select a benchmark complex set from MIPS[46] known protein complexes which includes 1189 proteins in 203 known complexes. The cluster algorithms are implemented by the cytoscape default settings. As the figures shown, the reconstructed PPI networks can improve the performance of protein complex prediction according to the ACC and Jaccard metrics. Similar to $PPI_{Collins}$, all the calculations are implemented on $PPI_{Krogans}$ as well (see Supplementary 1 Figure 6-7).

3.4 Application and evaluation on *C. elegans* PPI network

We assign RSPGM score for each PPI on the new integrative PPI network of *C. elegans* to assess the reliability of protein pairs. The adjacent matrix (5039×5039) is built according to 12951 PPIs of integrative PPI networks of *C. elegans*. Then, this W as long as other settings are applied based on the Algorithm 1 to obtain the reliability score for each PPI. The data of PPIs with RSPGM scores is available at our website and in Supplementary 2. In this subsection, for the new integrative *C. elegans* network, we firstly validate the consistency between our RSPGM score and the GO and sequence similarity. Then, we provide an example to infer interaction path.

3.4.1 Consistency validation To investigate the relationship between the similarity of interacting proteins and the assigned reliability scores in *C. elegans*, we compare GO and sequence similarity with the RSPGM scores respectively. The flowchart of calculating GO similarity and sequence similarity is the same as that in Section 3.3.2. The results are shown in Supplementary 1 Figure 8. In the GO process of MF, the average GO similarity of the 10% coverage of the PPI network with the top 10% highest RSPGM scores is about 0.639. This similarity value decreases dramatically from top 10% to 30% coverage of the PPI network. Finally, it drops to about 0.515 at the 100% coverage of the PPI network. In BP and CC, they also keep descending but not very significant. For sequence similarity, it decreases from 0.065 to 0.035. RSPGM reliability score is consistent with GO similarity and sequence similarity in our integrative *C. elegans* network. Therefore, the results are consistent with the ones on $PPI_{Collins}$ shown in Fig. 2.

3.4.2 Interaction path inference To evaluate the availability of our proposed method on PPI assessment, we apply the integrative *C. elegans* PPI network with RSPGM reliability score on interaction path inference. Here, we apply Gitter *et al.*'s[47] method to define the weight of the possible path for interaction path inference. The inferred interaction path could be viewed as the pathway if adding direction and regulatory effect on each interaction.

A well-studied *C. elegans* pathway, the canonical Wnt/β-catenin pathway, is used as the reference to validate interaction path inference result. This pathway is responsible for modulating expression of specific target genes by effector protein β-catenin. The canonical Wnt/β-catenin pathway is a signal transduction pathway from Wnt ligands to β-catenin protein[48]. Here, we inferred the interaction path between one type of Wnt ligands and one type of β-catenin proteins. This inferred interaction path will be useful for pathway inference.

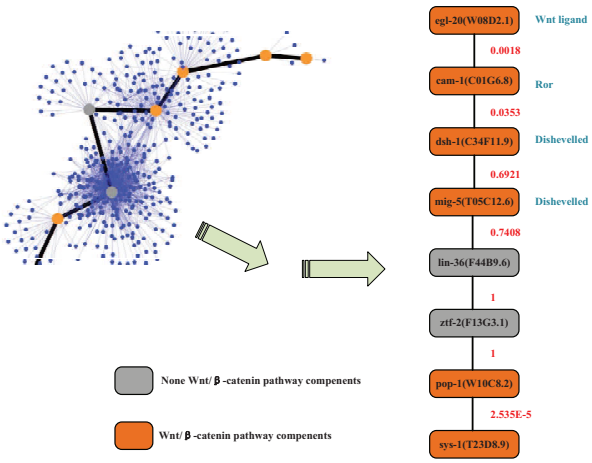


Fig. 5 The inferred interaction path between one type of Wnt ligands and one type of β-catenin proteins.

Gene *egl-20* (W08D2.1) produces one type of Wnt ligands, while *sys-1* (T23D8.9) produces a β-catenin protein. We inferred interaction path between *egl-20* and *sys-1*. Totally 1415 candidate paths have been found by setting $L = 7$, which represents the maximum of finding the candidate path length (details in Supplementary 3). The inferred interaction path with the highest path score is shown in Fig. 5. Moreover, for the 8 genes on the interaction path, 6 of them, 75%, are Wnt/β-catenin pathway related genes. These 6 genes have been validated and comprehensively studied by other literature[49]. Also,

for all 1415 possible candidate paths, they totally include 280 genes. Among them, only 17 genes (gene name with symbol '♯' in Supplementary 3), about 6%, are Wnt/β-catenin pathway components. In the inferred interaction path, it is a high rate (75%) of Wnt pathway component, although most genes in the possible candidate paths set are not. Therefore, the performance of interaction path inference is relatively accurate by using the reliability score computed from RSPGM algorithm.

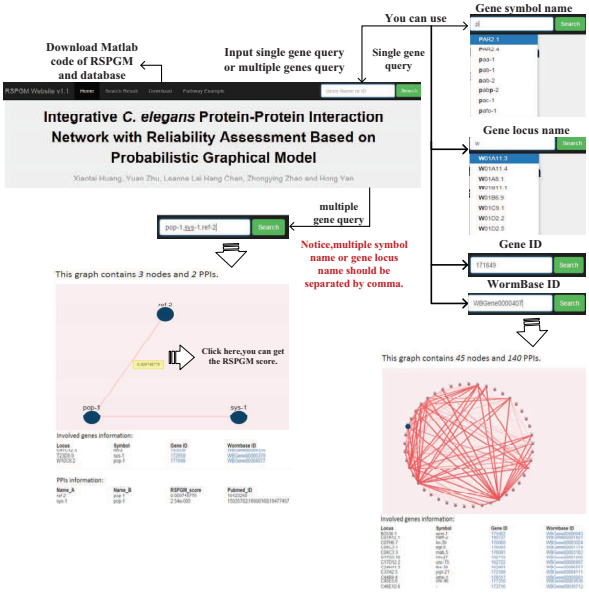


Fig. 6 The website is designed for querying and visualizing RSPGM score of PPI subnetwork about single gene or multiple genes in *C. elegans*.

3.5 Website server

To query and visualize PPI network with RSPGM scores, we build a user interactive website, available at http://rspgm.bionetworks.tk/. This website is in support of two types of query, single gene query and multiple genes query. User can type single gene name or multiple gene names in the search bar. It will return a subnetwork graph in the webpage, shown in Fig. 6. The datails of usage can be found in Supplementary 1 Section XIII.

We use SQLite version 3.8.8.3 to store the data and execute SQL query for single gene and multiple genes query. Mojolicious version 6.06, a Perl real-time web framework, is employed to build the website. With the help of Cytoscape.js version 2.3.11, the network graphs are illustrated in the website.

4 Discussion

In this paper, we constructed a PPI network in *C. elegans* by integrating data from seven molecular interaction databases. This integrative PPI network was subsequently evaluated by our newly proposed protein-protein interaction reliability assessment method RSPGM. This weighted PPI network is useful for pathway inference. Also, we built a website for querying and visualizing protein-protein interactions with RSPGM scores in the *C. elegans* PPI network.

In the consistency validation between RSPGM score and GO similarity, sequence similarity, it shows a significant descendant trend in yeast data. However, this is not very significant in current *C. elegans* data. It may be due to the non-full map of protein-protein interactome in *C. elegans* currently[7]. RSPGM algorithm is based on topology of the input network. Therefore, an incomplete protein-protein interactome may disturb the result of RSPGM scoring. In the functional relevance validation, the PPI groups in the reconstructed network generated by RSPGM have higher GO similarity, gene expression PCC and essentiality percentages than that in the original network, and obtain improved performance for the prediction of known protein complexes.

In the interaction path validation, 6 out of 8 genes are related to the reference Wnt pathway in the example in Section 3.4.2. The other two genes, *lin-36* (F44B9.6) and *ztf-2* (F13G3.1), have not been shown to be the components of Wnt pathway. However, in the inferred interaction path, the PPIs' scores corresponding to these two genes are very high, (Fig. 5). This indicates that these two genes are hub nodes in the network which may be involved in other biological pathways. Generally, several different pathways can cooperate to possess particular biological functions[50]. *lin-36* gene is the SynMuv B pathway component[51]. It has been validated to interact with *eor-1* which belongs to Ras/ERK pathway to cooperate with Wnt pathway[52]. *ztf-2* gene encodes an orthology of human ovo-like zinc finger 2 (Ovol2) which has been reported to act as the downstream of Wnt pathway[53]. Therefore, both *lin-36* and *ztf-2* are indirectly related to the Wnt pathway, which implies that the inferred interaction path in the example is very close to the Wnt/β-catenin pathway.

In future study, weighted PPIs network along with other data sources, such as PDIs, genetic interactions (GIs) and perturbation data, will be simultaneously considered for pathway inference.

Acknowledgement

This work is supported by the Hong Kong Research Grants Council (Project HKBU5/CRF/11G) and City University of Hong Kong (Project 9610326), the National Science Foundation of China (Project 11401110), the Natural Science

Foundation of Guangdong Province (Project 2013KJCX0086) and the Research Center Foundation of School of Automation of China University of Geosciences (Wuhan) (Project AU2015CJ008).

References

1 N. Pratanwanich and P. Lió, *Molecular Biosystems*, 2014, **10**, 1538–1548.
2 O. Ourfali, T. Shlomi, T. Ideker, E. Ruppin and R. Sharan, *Bioinformatics*, 2007, **23**, i359–i366.
3 A. Todor, H. Gabr, A. Dobra and T. Kahveci, *Bioinformatics*, 2014, **30**, i96–i104.
4 D. Kleftogiannis, L. Wong, J. A. Archer and P. Kalnis, *Bioinformatics*, 2015, btv138.
5 A. Vinayagam, J. Zirin, C. Roesel, Y. Hu, B. Yilmazel, A. A. Samsonova, R. A. Neumüller, S. E. Mohr and N. Perrimon, *Nature Methods*, 2014, **11**, 94–99.
6 Y. Ko, C. Zhai and S. Rodriguez-Zas, *BMC Systems Biology*, 2009, **3**, 54.
7 K. C. Gunsalus and K. Rhrissorrakrai, *Current Opinion in Genetics & Development*, 2011, **21**, 787–798.
8 A. Emamjomeh, B. Goliaei, J. Zahiri and R. Ebrahimpour, *Molecular Biosystems*, 2014.
9 I. Saha, J. Zubek, T. Klingström, S. Forsberg, J. Wikander, M. Kierczak, U. Maulik and D. Plewczynski, *Molecular Biosystems*, 2014, **10**, 820–830.
10 L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie and D. Eisenberg, *Nucleic Acids Research*, 2004, **32**, D449–D451.
11 A. Chatr-aryamontri, B. J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. Oonnell *et al.*, *Nucleic Acids Research*, 2013, **41**, D816–D823.
12 S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz *et al.*, *Nucleic Acids Research*, 2011, **40**, D841–D846.
13 L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico *et al.*, *Nucleic Acids Research*, 2012, **40**, D857–D861.
14 T. W. Harris, J. Baran, T. Bieri, A. Cabunoc, J. Chan, W. J. Chen, P. Davis, J. Done, C. Grove, K. Howe *et al.*, *Nucleic Acids Research*, 2014, **42**, D789–D793.
15 N. Simonis, J.-F. Rual, A.-R. Carvunis, M. Tasan, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, J. M. Sahalie, K. Venkatesan, F. Gebreab *et al.*, *Nature Methods*, 2009, **6**, 47–54.
16 W. Zhong and P. W. Sternberg, *Science*, 2006, **311**, 1481–1484.
17 S. Suthram, T. Shlomi, E. Ruppin, R. Sharan and T. Ideker, *BMC Bioinformatics*, 2006, **7**, 360.
18 X. Lin, M. Liu and X. Chen, *BMC Bioinformatics*, 2009, **Suppl 4**, S5.
19 M. Deng, F. Sun and T. Chen, *Pacific Symposium on Biocomputing*, 2003, **8**, 140–151.
20 C. Lei and J. Ruan, *Bioinformatics*, 2013, **29**, 355–364.
21 Y. Hulovatyy, R. W. Solava and T. Milenković, *PLoS ONE*, 2014, **9**, e90073.
22 J. Chen, M. L. Lee and S. K. Ng, *Bioinformatics*, 2006, **22**, 1998–2004.
23 J. Hou and A. Saini, *Mathematical Biosciences*, 2013, **245**, 226–234.
24 R. Saito, H. Suzuki and Y. Hayashizaki, *Nucleic Acids Research*, 2002, **30**, 1163–1168.
25 R. Saito, H. Suzuki and Y. Hayashizaki, *Bioinformatics*, 2003, **19**, 756–763.
26 X. Luo, Z. You, M. Zhou, S. Li, H. Leung, Y. Xia and Q. Zhu, *Scientific Reports*, 2015, **5**, 7702.
27 Y. Zhu, X. F. Zhang, D. Q. Dai and M. Y. Wu, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2013, **10**, 219–225.
28 X. F. Zhang, D. Q. Dai, L. Ou-Yang and H. Yan, *BMC Bioinformatics*, 2014, **15**, 186.
29 J. M. Ranola, S. Ahn, M. Sehl, D. J. Smith and K. Lange, *Bioinformatics*, 2010, **26**, 2004–2011.
30 R. Schweiger, M. Linial and N. Linial, *Bioinformatics*, 2011, **27**, i142–i148.
31 L. Ou-Yang, D. Q. Dai, X. L. Li, M. Wu, X. F. Zhang and P. Yang, *BMC Bioinformatics*, 2014, **15**, 335.
32 X. F. Zhang, D. Q. Dai and X. X. Li, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2012, **9**, 857–870.
33 Z. M. Saul and V. Filkov, *Bioinformatics*, 2007, **23**, 2604–2611.
34 S. R. Collins, K. Patrick, X. C. Zhao, J. F. Greenblatth, F. Spencerg, H. F. C. P, J. S. Weissmana and N. J. Krogana, *Molecular & Cellular Proteomics*, 2007, **6**, 439–450.
35 N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis *et al.*, *Nature*, 2006, **440**, 637–643.
36 A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. Jensen, S. Bastuck, B. Dümpelfeld *et al.*, *Nature*, 2006, **440**, 631–636.
37 J. P. Miller, R. S. Lo, A. Ben-Hur, C. Desmarais, I. Stagljar, W. S. Noble and S. Fields, *Proceedings of the National Academy of Sciences of the United States of America*, 2005, **102**, 12123–12128.
38 S. Oliver, *Nature*, 2000, **403**, 601–603.
39 G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu and S. Wang, *Bioinformatics*, 2010, **26**, 976–978.
40 J. Z. Wang, Z. Du, R. Payattakool, S. Y. Philip and C. Chen, *Bioinformatics*, 2007, **23**, 1274–1281.
41 I. Korf, M. Yandell and J. Bedell, *Blast*, O'Reilly Media, 2003.
42 R. Edgar, M. Domrachev and A. E. Lash, *Nucleic Acids Research*, 2002, **30**, 207–210.
43 S. S. Dwight, R. Balakrishnan, K. R. Christie, M. C. Costanzo, K. Dolinski, S. R. Engel, B. Feierbach, D. G. Fisk, J. Hirschman, E. L. Hong *et al.*, *Briefings in Bioinformatics*, 2004, **5**, 9–22.
44 T. Nepusz, H. Yu and A. Paccanaro, *Nature Methods*, 2012, **9**, 471–472.
45 K. Rhrissorrakrai and K. C. Gunsalus, *BMC Bioinformatics*, 2011, **12**, 192.
46 H.-W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Güldener, G. Mannhaupt, M. Münsterkötter, P. Pagel, N. Strack, V. Stümpflen *et al.*, *Nucleic Acids Research*, 2004, **32**, D41–D44.
47 A. Gitter, J. Klein-Seetharaman, A. Gupta and Z. Bar-Joseph, *Nucleic Acids Research*, 2011, **39**, e22–e22.
48 C. Y. Logan and R. Nusse, *Annual Review of Cell and Developmental Biology*, 2004, **20**, 781–810.
49 H. Sawa and H. C. Korswagen, *Wnt signaling in C. elegans*, WormBook, 2005.
50 R. Derynck, B. P. Muthusamy and K. Y. Saeteurn, *Current Opinion in Cell Biology*, 2014, **31**, 56–66.
51 D. S. Fay and J. Yochem, *Developmental Biology*, 2007, **306**, 1–9.
52 R. M. Howard and M. V. Sundaram, *Genes & Development*, 2002, **16**, 1815–1827.
53 J. Wells, B. Lee, A. Q. Cai, A. Karapetyan, W. Lee, E. Rugg, S. Sinha, Q. Nie and X. Dai, *Journal of Biological Chemistry*, 2009, **284**, 29125–29135.