RESEARCH ARTICLE

Computational identification of signaling pathways in protein interaction networks [version 1; referees: 2 approved with reservations]

Angela U. Makolo, Temitayo A. Olagunju

Bioinformatics Research Group, Computer Science Department, University of Ibadan, Ibadan, 200284, Nigeria

v1 **First published:** 30 Dec 2015, **4**(ISCB Comm J):1522 (doi: 10.12688/f1000research.7591.1)

Latest published: 30 Dec 2015, **4**(ISCB Comm J):1522 (doi: 10.12688/f1000research.7591.1)

Abstract

The knowledge of signaling pathways is central to understanding the biological mechanisms of organisms since it has been identified that in eukaryotic organisms, the number of signaling pathways determines the number of ways the organism will react to external stimuli. Signaling pathways are studied using protein interaction networks constructed from protein-protein interaction data obtained from high-throughput experiments. However, these high-throughput methods are known to produce very high rates of false positive and negative interactions. To construct a useful protein interaction network from this noisy data, computational methods are applied to validate the protein-protein interactions. In this study, a computational technique to identify signaling pathways from a protein interaction network constructed using validated protein-protein interaction data was designed.

A weighted interaction graph of *Saccharomyces Cerevisiae* was constructed. The weights were obtained using a Bayesian probabilistic network to estimate the posterior probability of interaction between two proteins given the gene expression measurement as biological evidence. Only interactions above a threshold were accepted for the network model.

We were able to identify some pathway segments, one of which is a segment of the pathway that signals the start of the process of meiosis in *S. Cerevisiae*.

This article is included in the ISCB Africa ASBCB Conference on Bioinformatics channel.

Open Peer Review

Referee Status: ? ?

	Invited Referees 1	2
version 1 published 30 Dec 2015	? report	? report

1 **Winston Hide**, University of Sheffield USA

2 **Lynn Fink**, University of Queensland Australia

Discuss this article

Comments (0)

Corresponding author: Temitayo A. Olagunju (polag01@yahoo.com)

How to cite this article: Makolo AU and Olagunju TA. **Computational identification of signaling pathways in protein interaction networks [version 1; referees: 2 approved with reservations]** *F1000Research* 2015, **4**(ISCB Comm J):1522 (doi: 10.12688/f1000research.7591.1)

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: No competing interests were disclosed.

First published: 30 Dec 2015, **4**(ISCB Comm J):1522 (doi: 10.12688/f1000research.7591.1)

Introduction

For biologists and scientists in the life sciences, the successful sequencing of the genome is only one step out of many involved in understanding organisms. This has produced a lot of information that will not be useful unless refined. Biologists are interested in understanding the intricacies of the workings of the cells of an organism – the activities and reactions of such an organism to its environment. This information is useful in designing necessary interventions in order to modify the biological mechanisms of an organism or its reactions to external stimuli.

According to the central dogma of molecular biology, genes are composed of DNA which is transcribed into RNA and the RNA is then translated into protein. Ultimately, all organisms are composed mainly of proteins in different forms and quantity.

Proteomic data and protein-protein interaction data from organisms form a key component in understanding an organism due to the major role played by proteins in cellular mechanisms. Protein-protein interactions are the foundation of biological mechanisms such as signal transduction, cell cycle control, DNA replication and transcription and enzyme-mediated metabolism[1,2].

As a result of these interactions, understanding of organisms is facilitated by modeling the Protein Interaction Network (PIN) with a network constructed using the protein-protein interaction data. With a model such as this, a lot can be learned of the organism from its reaction to external stimuli and the effects of interventions on the biological mechanisms of the organism. For instance, it has been shown that the phenotypic effects of the deletion of a single gene depend on the position of that gene in the complex web of protein interactions[3].

Apart from the importance of the protein-protein interactions map in studying the machinery of the proteome and the cellular behaviour of an organism, they are also practically important in the creation of interventions aimed at producing desired phenotypic outcomes such as new drug designs or disease prevention[4,5].

Protein-protein interaction data from organisms are obtained on a large scale using a number of high throughput techniques such as Yeast Two-Hybrid (Y2H), Co-Immunoprecipitation (Co-IP), Mass Spectrometry etc. These high throughput techniques have however, been identified to have high rates of false positives and false negatives. False positive interactions are protein-protein interactions that are reported to exist with any of the experimental techniques but do not exist in reality, while false negative interactions are true interactions that do not get reported using an experimental technique. Rates of false positives in protein interaction data have been reported to be as high as 50%[6–8]. As a result of analysis based on the integration of gene expression level measurement data and protein-protein interaction data, only about 30–50% of the interactions have been suggested to be biologically relevant. Reference 9 reported 47% true protein-protein interactions where a Paralogous Verification Method (PVM) was applied. The PVM may have performed better owing to its incorporation of information on paralogs of other organism to strengthen the biological evidence.

These high rates of protein-protein interaction data inaccuracy are due to peculiarities of the techniques used to generate them. For instance, unlike other affinity-based methods that cannot detect transient interacting proteins, Tandem Affinity Purification (TAP-tag) tag methods can detect transient interacting proteins which are however lost during the purification process[10].

Furthermore, these new high throughput methods of detecting protein interactions have no doubt rapidly generated much more data than have been collected by traditional methods in small scale experiments. This thus makes it impractical to start verifying each of these interactions by the traditional methods used in small scale experiments[11].

In order to make sense of the vast data and obtain insightful information, these data need to be subjected to analytical procedures that will extract signal from the noise. This task of analyzing genomic data takes a computational approach due to the magnitude of the information involved. In reducing this level of noise in the protein interaction data different computational techniques aimed at improving the reliability of the data are applied. To predict true interactions between protein pairs, many authors have suggested a number of methods for estimating and assigning reliabilities to the interactions in the experimental data. These methods include using a logistic regression distribution function over a number of parameters to assign confidence scores to the interactions[17,18], the use of expression profile and paralogs to assign reliability scores to already observed interactions[9] the use of maximum likelihood technique for the estimation of domain-domain interactions in order to infer protein-protein interactions[7]. For computational biologists, the challenge would be the development of methods of transforming the high-throughput data obtained from these different sources into biological insights.

In this paper, we seek to bridge the gap between protein-protein interaction data and other biological data in constructing useful signaling pathway models that will lead to insightful knowledge of biological processes. We propose a probabilistic approach using Bayesian networks to assign weights to protein-protein interactions. These weighted interactions are then used to construct the weighted PIN from which signaling pathways are predicted. Refer to Figure 1 for the schematic of the computational approach used. This work thus describes a computational means to clean up the background noise inherent in the various methods of proteomic data acquisition in order to better understand bio-molecular mechanisms.

Materials and methods

Data sources

Protein interaction data was obtained from the publicly available Saccharomyces Genome Database (SGD). The protein interaction data is an amalgamation of the interactions obtained using eight different high throughput experimental procedures - Yeast Two-Hybrid, Affinity Capture Mass Spectrometry, CO-Purification, Affinity Capture Western, Biochemical Activity, Reconstituted Complex, Protein-Peptide and Far Western. This data contained 22,650 interactions between 2554 different proteins.

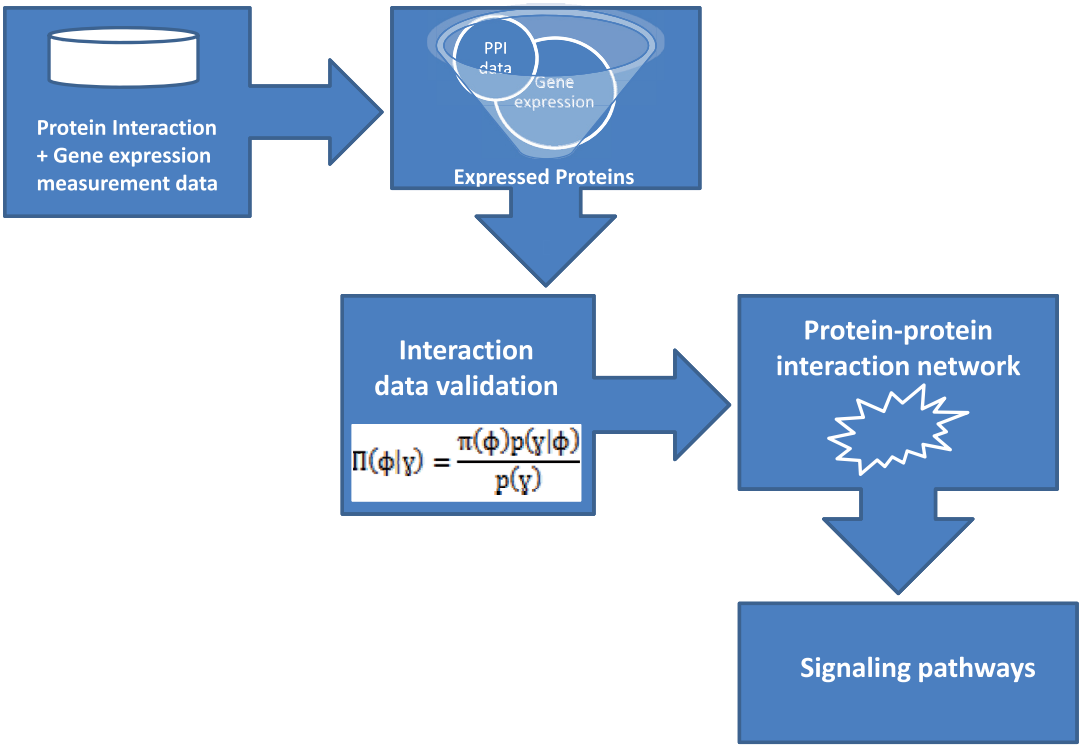

Figure 1. Schematic of the computational method based on Bayesian probability used in this work.

Dataset 1. Yeast Expression Data

http://dx.doi.org/10.5256/f1000research.7591.d110325

The file contains proteins with their expression measurements used in this study, obtained from DNA microarrays of Yeast. Data was obtained from the work of Spellman *et al.* (1998) and available from http://genome-www.stanford.edu/cellcycle/data/rawdata/CellCycle98.xls[41].

Abbreviations

SGD - Saccharomyces Genome Database

YPD - Yeast Peptone Dextrose

ORF - Open Reading Frame

Dataset 2. Protein-protein interaction data

http://dx.doi.org/10.5256/f1000research.7591.d110326

Contains pairwise protein interaction data used in this study, obtained using high throughput experimental techniques such as Y2H, Affinity Capture-MS, Co-purification, Biochemical Activity and Reconstituted Complex, available from the publicly available SGD database[42].

The *S. cerevisiae* expression measurement data was obtained from the Yeast Cell Cycle Analysis Project of the Stanford University. The data is housed at a publicly available database maintained by the *Saccharomyces* Genome Database at the Department of Genetics, School of medicine. The Yeast Cell Cycle Analysis project aimed at identifying all the genes whose mRNA levels are regulated by the cell cycle[12]. This data is available at the Yeast Cell Cycle Analysis Project site. The data contained the expression profiles of 800 proteins of the *S. cerevisiae* organism.

Data processing

The processing of the data obtained from the Yeast protein-protein interaction data and the Yeast Expression measurement was carried out by first filtering for the proteins that have expression level measurement. Only the proteins in the protein-protein interaction dataset that were also present in the gene expression measurement data were used. This was done based on the hypothesis that proteins occurring in the same complex and are known to physically interact have higher correlation than proteins that are not known to directly interact. This hypothesis is supported by[13–16], where it has been observed that true protein interactions have a high mRNA expression for the proteins involved. The filtration of the dataset produced 306 protein-protein interactions that have expression level measurements from the 22,650 protein interactions and the

800 gene expression measurements. These 306 protein-protein interactions represent the intersection of the two datasets as depicted in Figure 2. With reference to the yeast protein interaction data and expression measurement data respectively, these figures correspond to 0.013% and 0.382% of the original dataset respectively.

Interaction data validation

In this work, the probability estimation of protein interactions was done using a Bayesian probabilistic model. According to Bayes' theorem, the posterior probability density is proportional to the prior probability density and the likelihood function.

Our interest is in drawing inference about the parameter ϕ from a probability model $p(\gamma|\phi)$ to give rise to observed data γ. Allocating a prior probability $\pi(\phi)$ to the parameters assuming they are uncertain, we can obtain a posterior probability according to Bayes' theorem where $p(\gamma)$ which is the marginal density for γ is obtained by integrating over the prior. Refer to Equation 1.

$$\prod(\phi|\gamma) = \frac{\pi(\phi)p(\gamma|\phi)}{p(\gamma)} \qquad \text{Equation 1}$$

Equation (1) can be rewritten as Equation (2) since $\pi(\phi|\gamma)$ is a function of ϕ for observed γ which shows the direct proportionality between the posterior probability and the product of the likelihood and the prior probability[19].

$$\prod(\phi|\gamma) \propto \pi(\phi)p(\gamma|\phi) \qquad \text{Equation 2}$$

Likelihood function

In order to make use of the Bayesian model, there must be an approximation or full specification of the prior probability distribution and the likelihood function. The first step in the determination of the likelihood function which is based on the probability of observing the data is to fix a probability distribution $f(\theta)$ where θ is the parameter defining the probability distribution.

In outcome space, for a given dataset (Y1, Y2, Y3, … ,Yn), the probability of observing the dataset given θ is written as

$$f_\theta(Y1, Y2, Y3, \ldots, Yn) \qquad \text{Equation 3}$$

In parameter space, the likelihood function in terms of the probability of observing the dataset given θ is

$$L_{Y1,Y2,Y3,\ldots,Yn}(\theta) = f_\theta(Y1, Y2, Y3, \ldots, Yn) \qquad \text{Equation 4}$$

For a Bernoulli distribution, the probability distribution is

$$f_\theta(Y) = \theta^y(1-\theta)^{(1-y)} \qquad \text{Equation 5}$$

Therefore for a sample of N observations (Y_1, Y_2, Y_3, …,Y_n), the joint distribution is as Equation 6, and can be rewritten as Equation 7

$$f_\theta(Y) = \prod_{i=1}^{i=N} f\left(Y_i = y_i\right) \qquad \text{Equation 6}$$

$$f\,\theta(Y) = \prod_{i=1}^{i=N} \theta^{y_i}(1-\theta)^{(1-y_i)} \qquad \text{Equation 7}$$

The likelihood function determines what value of θ makes the dataset (Y1, Y2, Y3, …,Yn) most probable.

Estimating the maximum likelihood of the parameter θ, we maximize the function with respect to θ and then set it to zero to obtain the Maximum Likelihood Estimation of the parameter θ[20].

$$\theta = \max L_{(Y)}(\theta) \qquad \text{Equation 8}$$

Weighted graph construction

We formalize the problem of constructing a weighted graph that is instrumental in building a PIN. Let (V, E, w) be the protein-protein interaction network where $V = p_0, p_1, \ldots p_n$ is the set of all proteins

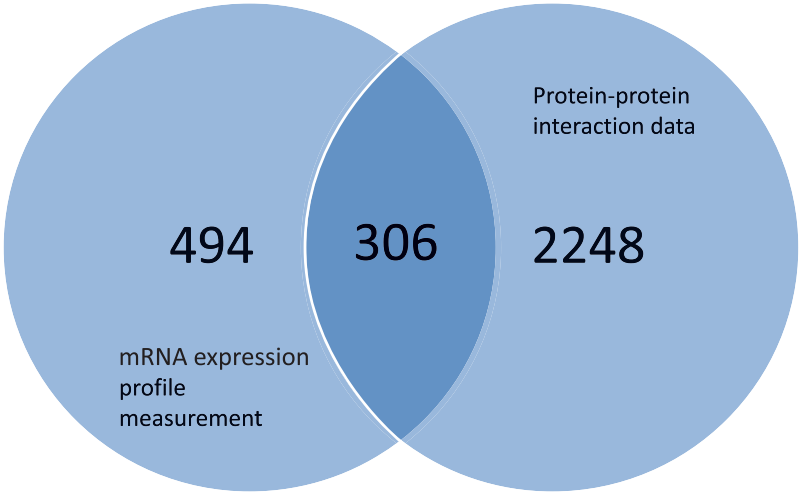

Figure 2. Venn diagram of the mRNA expression profile measurement dataset and the protein-protein interaction dataset with the intersection showing the protein-protein interaction data with expression measurement.

and $E = \{e = (p_i, p_j) \mid p_i, p_j \in V\}$ is the set of interactions among these proteins in the set V, and w is the weight of each edge that belongs to E. w being the weight of the interaction between two proteins (p_i, p_j) is a measure of the reliability of interaction between the two proteins (p_i, p_j) obtained using the Bayesian probabilistic approach described above.

The vertices of the interaction graph are contained in the set of the unique proteins obtained after the computation of the reliability of interaction between the protein pairs i.e. $|V| = 306$, and the edges of the graph are the set of interactions between these proteins in V.

The protein-protein interaction graph is constructed with an undirected sparse graph due to the sparse nature of biological networks.

Graph implementation

The implementation of the algorithm for this computational technique was done in Java programming language using the Java Universal Network Graphics (JUNG) framework for graphs. The JUNG framework is an open-source collection of libraries providing common language for modeling, analyzing and visualizing any data that can be represented as a graph or network. The JUNG framework is extensible in order to tailor it to specific needs and also includes implementation of a host of algorithms for network analysis, graph theory and data mining.

The graph implementation in JUNG supports the representation of the different types of graphs such as directed and undirected graph, multimodal graphs, graphs with parallel edges and hypergraphs.

For this work, we used the JUNG 2.0.1 API released in January, 2010 which can be found at http://jung.sourceforge.net.

Simple path-finding

A pathway is an ordered list of distinct proteins in V such that each consecutive pair is found in E[21].

Given an undirected sparse graph G = (V, E, w) and a pair of nodes $\{(p_i, p_j) \mid (p_i, p_j) \in V\}$ corresponding to the starting and ending proteins respectively, we wish to find a simple path from p_i to p_j which will be a segment of a pathway.

With the graph constructed, which is the PIN of the *S. cerevisiae* organism based on the data supplied, we queried the graph with a pair of proteins $(p_i, p_j) \in V$ which are respectively the starting protein and ending proteins of the path of interest. We are interested in having a simple path corresponding to the signal transduction path from the starting protein to the ending protein returned by the search algorithm.

The search is done using a Depth First Search (DFS) algorithm. The start protein becomes the root node for the algorithm and examines all the outgoing nodes to it, expands the first child node of the apparent tree and progressively continues the search until the target node (the ending protein) is found. If the DFS algorithm however encounters a node that has no children, it backtracks to the previous node to continue exploring the children nodes.

Protein mapping to functional annotation

In order to understand and make meaning of the pathway segments that are obtained from the PIN, we compared the proteins to their functional annotation. Mapping proteins in known signaling pathways and PINs to their functional annotations has an important function. The proteins in an organism may have similar biological functions such that one protein effectively replaces another in a pathway, then such proteins should share the same set of gene annotation terms. The Gene Ontology annotation, which is a functional annotation scheme, provides this basis for the identification of functional description of proteins and their interactions with other proteins and other molecules.

In this work, we used the Gene Ontology (GO) annotations to interpret the pathway segments that have been identified from the protein-protein interaction network constructed for the *S. cerevisiae* organism.

Results

To validate the protein-protein interaction data that we used in this work, we applied the method that was described in section II-B to first filter the data. As was described the filtration of the data was done by integrating the gene expression measurement of the regulated Yeast Cell-cycle in order to obtain a dataset that is an intersection of both datasets.

This step was taken based on the hypothesis that there is a high correlation between the expression levels of truly interacting proteins[13–16,22] and also using the gene expression measurement as a source of biological information[23–25] for the computational inference.

Using the protein-protein interaction data comprising 22,650 interactions between 2554 unique proteins and the gene expression levels of 800 genes, we applied the computational approach based on Bayesian probability described earlier. Further in the validation process, the application of the Bayesian probabilistic model on the data to estimate the posterior probability of an interaction existing between two proteins given the biological evidence produced the weight estimate for the interactions. With the estimation of the interaction weight and the rejection of interaction weights below the threshold obtained from the mean expression level measurements, we obtained a dataset containing 306 protein-protein binary interactions. This dataset was used in constructing the PIN of the *S. cerevisiae* organism.

The 306 protein-pairs represented the proteins that had expression profile measurement, which corresponds to the intersection set of the two datasets. With reference to the protein-protein interaction data and the gene expression level measurement, this is a mere 0.013% and 0.382% of the original dataset respectively.

We applied the method described in section 2.7 to identify pathway segments in the constructed PIN for *S. cerevisiae*. Given a graph G = (V, E), a pathway has been described as an ordered list of distinct proteins in V such that each consecutive pair is found in E[21]. With a starting protein and an ending protein of interest, a simple

path between these two corresponds to a pathway. Due to the size limitation of the expression measurement dataset used and the effective reduction in the overall number of proteins used to construct the graph, we were only able to identify pathway segments. A pathway segment is a chain of interacting proteins which is a part of a larger pathway. Some of the resulting pathways identified with this technique are presented in Table 1 and Table 2. These tables elucidate the protein description, the GO function and GO process of the proteins involved in the pathway segments as obtained from the *AmiGO* website http://www.geneontologyproject.org/go.

Table 1. Pathway segment YMR163C - YOR326W - YCL063W - YER150W.

Protein	Gene	Protein Information	Gene Ontology Function	Gene Ontology Process
YMR163C	Inp2	Peroxisome-specific receptor important for peroxisome inheritance; co-fractionates with peroxisome membranes and co-localizes with peroxisomes *in vivo*; physically interacts with the myosin V motor Myo2p; INP2 is not an essential gene[29]	**Myosin binding** Interacting selectively and non-covalently with any part of a myosin complex; myosins are any of a superfamily of molecular motor proteins that bind to actin and use the energy of ATP hydrolysis to generate force and movement along actin filaments. [GO:0017022]	**Peroxisome inheritance** The acquisition of peroxisomes by daughter cells from the mother cell after replication. In *Saccharomyces cerevisiae*, the number of peroxisomes cells is fairly constant; a subset of the organelles are targeted and segregated to the bud in a highly ordered, vectorial process. Efficient segregation of peroxisomes from mother to bud is dependent on the actin cytoskeleton, and active movement of peroxisomes along actin filaments is driven by the class V myosin motor protein, Myo2p. [GO:0045033]
YOR326W	MYO2, CDC66	Type V myosin motor involved in actin-based transport of cargos; required for the polarized delivery of secretory vesicles, the vacuole, late Golgi elements, peroxisomes, and the mitotic spindle; MYO2 has a paralog, MYO4, that arose from the whole genome duplication[30]	**Actin filament binding** Interacting selectively and non-covalently with an actin filament, also known as F-actin, a helical filamentous polymer of globular G-actin subunits. [GO:0051015]	**Cell division** The process resulting in the physical partitioning and separation of a cell into daughter cells. Source: GOC: go_curators Comment Note that this term differs from 'cytokinesis ; GO:0000910' in that cytokinesis does not include nuclear division. [GO:0051301]
YCL063W	VAC17, YCL062W	Phosphoprotein involved in vacuole inheritance; degraded in late M phase of the cell cycle; acts as a vacuole-specific receptor for myosin Myo2p[31,32]	**Protein anchor** Interacting selectively and non-covalently with both a protein or protein complex and a membrane, in order to maintain the localization of the protein at a specific location on the membrane. [GO: 0043495]	No Information Available
YER150W	SPI1	GPI-anchored cell wall protein involved in weak acid resistance; basal expression requires Msn2p/Msn4p; expression is induced under conditions of stress and during the diauxic shift; SPI1 has a paralog, SED1, that arose from the whole genome duplication[33]	**Molecular_function** Elemental activities, such as catalysis or binding, describing the actions of a gene product at the molecular level. A given gene product may exhibit one or more molecular functions. [GO: 0003674]	**Response to acid** Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of an acid stimulus. The acid may be in gaseous, liquid or solid form. [GO: 0001101]

Table 2. Pathway segment YIL026C - YIL126W - YDL003W - YJL074C.

Protein	Gene	Protein Information	Gene Ontology Function	Gene Ontology Process
YIL026C	IRR1, SCC3	Subunit of the cohesin complex; which is required for sister chromatid cohesion during mitosis and meiosis and interacts with centromeres and chromosome arms; relocalizes to the cytosol in response to hypoxia; essential for viability[34,35].	**Chromatin Binding** Interacting selectively and non-covalently with chromatin, the network of fibers of DNA, protein, and sometimes RNA, that make up the chromosomes of the eukaryotic nucleus during interphase. [GO: 0003682]	**fungal-type cell wall organization** A process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of the fungal-type cell wall. [GO: 0031505]
YIL126W	STH1p, NPS1	ATPase component of the RSC chromatin remodeling complex; required for expression of early meiotic genes; essential helicase-related protein homologous to Snf2p[36,37].	**DNA-dependent ATPase activity** Catalysis of the reaction: ATP + H2O = ADP + phosphate; this reaction requires the presence of single- or double-stranded DNA, and it drives another reaction. [GO: 0008094]	**ATP-dependent chromatin remodeling** Dynamic structural changes to eukaryotic chromatin that require energy from the hydrolysis of ATP, ranging from local changes necessary for transcriptional regulation to global changes necessary for chromosome segregation, mediated by ATP-dependent chromatin-remodelling factors. [GO: 0043044]
YDL003W	MCD1	Essential alpha-kleisin subunit of the cohesin complex; required for sister chromatid cohesion in mitosis and meiosis; apoptosis induces cleavage and translocation of a C-terminal fragment to mitochondria; expression peaks in S phase[38,39].	**Chromatin binding** Interacting selectively and non-covalently with chromatin, the network of fibers of DNA, protein, and sometimes RNA, that make up the chromosomes of the eukaryotic nucleus during interphase. [GO: 0003682]	**Establishment of mitotic sister chromatid cohesion** The process in which the sister chromatids of a replicated chromosome become joined along the entire length of the chromosome during S phase during a mitotic cell cycle. [GO:0034087]
YJL074C	SMC3	Subunit of the multiprotein cohesin complex required for sister chromatid cohesion in mitotic cells; also required, with Rec8p, for cohesion and recombination during meiosis; phylogenetically conserved SMC chromosomal ATPase family member[40].	**ATPase activity** Definition Catalysis of the reaction: ATP + H2O = ADP + phosphate + 2 H+. May or may not be coupled to another reaction. [GO: 0016887]	**Meiotic sister chromatid cohesion** The cell cycle process in which sister chromatids of a replicated chromosome are joined along the entire length of the chromosome during meiosis. [GO: 0051177]

Discussion

Table 1 and Table 2 present some of the pathway segments identified using the computational approach proposed in this paper. The understanding of the paths is facilitated by using Gene Ontology associations to understand the biological processes the proteins are involved in. A signaling pathway is characterized by a starting protein that is a receptor at the membrane and ends with a transcription factor.

From Table 1 we identified a pathway segment {YMR163C-YOR326W- YCL063W- YER150W} along with the genes coding for each of the proteins using GO annotation.

The pathway segment starts with the protein YMR163C, identified to be a receptor important for peroxisome inheritance. Signaling pathways are often characterized by an activator at the membrane of the cell binding to a receptor to initiate the chain of signal transduction. These peroxisomes are organelles that metabolize fatty acids and are numerous in the *S. cerevisiae* organism. By blocking peroxisome transport through point mutants in the *MYO2p* gene that binds to it, the levels of *MYO2p* gene expression increased[26]. The implication of this is that signal is transmitted to the mother cell to stop further peroxisome transfer by lowering *INP2* gene expression. The next protein, YOR326W, in the pathway segment is coded for by the *MYO2p* gene whose level of expression is modified in the activation of signal that is relayed to alter the level of the *INP2* gene that codes for the YMR163C receptor protein. The next protein in the chain, YCL063W, coded for by the gene *VAC17* has been identified to be a vacuole-specific receptor for myosin MYO2P and is involved in vacuole inheritance, a molecular anchoring function.

The last protein in the pathway segment, YER150W, coded for by the gene *SPI1* contributes to transcriptional regulation induced under conditions of stress during the diauxic shift[27].

It is observed that this pathway which signals the start of the process of meiosis suddenly breaks off to a gene (*SPI1)* that participates in catalysis at the molecular level. This is not abnormal as these are pathway segments and not the full transduction pathway activated by the receptor protein.

In a similar approach, 28 used protein-protein interaction data and expression data to model pathways. They ranked candidate signaling pathways of interacting proteins using expression data based on the rationale that proteins in the same signaling network must simultaneously exist with the activation of the pathway; the genes coding for these proteins must also under the same environmental factors required for the signaling network and about the same time, be transcribed.

Their approach to predicting pathways included specifying the starting protein, a membrane protein, and an ending protein of interest, such as a DNA-binding protein, based on a prior knowledge of genetic relationship between them. In their findings, the pathways that the algorithm identified were not complete pathways owing to incomplete maps.

21 also applied a computational approach that is similar to our own by assigning scores to protein-protein interaction data, creating a PPI network from the data and mining signaling pathways from the network. The parameters for the search on the network included a starting protein and an ending protein as well as the length of the pathway segment. Although their approach involved training the algorithm using association rules mining from known pathways, they were only able to mine pathway segments too.

The incompleteness of pathways mined from using computational techniques on protein-protein interaction data can be attributed to false negative interactions that were not detected by the high throughput experiments that generated the data.

Furthermore, a number of computational techniques that have been applied to cleaning the noise in the protein-protein interaction data used often entails eliminating some data presumed to be noise from the dataset. The proteins removed in this manner could be important proteins that would then be missing in the modeled PIN. Our own approach involved filtering the protein-protein interaction data with the gene expression measurement data such that only the proteins with expression level measurement were used in the construction of the protein interaction network. This resulted in a reduction of the 22,650 pair-wise interactions by the gene expression measurement for just 800 proteins to 306 pair-wise interactions. This reduction in the size of the data used to construct the protein interaction network was a constraint on the number of pathways identified using this approach.

Conclusion

In this paper, we proposed a simple computational approach to identify signaling pathways in PINs by first estimating true interactions within protein-protein interaction data obtained from high throughput experimental techniques which are susceptible to generating high rates of false positive and false negative interactions. We proposed a technique using Bayesian Probability to estimate the probability of true interactions between two proteins and assigned weights to the pair wise interaction based on this. Using the validated protein-protein interaction data, we constructed a PIN of the *S. cerevisiae* organism from where simple paths between two proteins of interest were mined. Using the Gene Ontology annotation to understand the biological process taking place within the pathway, we were able to identify a pathway which signals the start of the process of meiosis, albeit broken off for want of more data.

Knowledge of signaling pathways are generally useful in designing biological interventions on an organism aimed at producing specific desired outcomes such as new drugs design and disease prevention and control.

Data availability

F1000Research: Dataset 1. Yeast Expression Data, 10.5256/f1000research.7591.d110325[41]

F1000Research: Dataset 2. Protein-protein interaction data, 10.5256/f1000research.7591.d110326[42]

Author contributions

AUM conceived the study and supervised it. TAO carried out the study. Both authors interpreted the results and were involved in the revision of the final draft manuscript and agreed to the content.

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting this work.

References

1. Kone BC: **Protein-protein interactions controlling nitric oxide synthases.** *Acta Physiol Scand.* 2000; **168**(1): 27–31.
PubMed Abstract | Publisher Full Text

2. Wang Jh: **Protein recognition by cell surface receptors: physiological receptors versus virus interactions.** *Trends Biochem Sci.* 2002; **27**(3): 122–126.
PubMed Abstract | Publisher Full Text

3. Jeong H, Mason SP, Barabási AL, *et al.*: **Lethality and centrality in protein networks: The most highly connected proteins in the cell are the most important for its survival.** *Nature.* 2001; **411**(6833): 41–42.
PubMed Abstract | Publisher Full Text

4. Eisenberg D, Marcotte EM, Xenarios I, *et al.*: **Protein function in the post-genomic era.** *Nature.* 2000; **405**(6788): 823–826.
PubMed Abstract | Publisher Full Text

5. Koyama FC, Chakrabarti D, Garcia CR: **Molecular machinery of signal transduction and cell cycle regulation in *Plasmodium*.** *Mol Biochem Parasitol.* 2009; **165**(1): 1–7.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Legrain P, Wojcik J, Gauthier JM: **Protein--protein interaction maps: a lead towards cellular functions.** *Trends Genet.* 2001; **17**(6): 346–352.
PubMed Abstract | Publisher Full Text

7. Deng M, Mehta S, Sun F, *et al.*: **Inferring domain-domain interactions from protein-protein interactions.** *Genome Res.* 2002; **12**(10): 1540–1548.
PubMed Abstract | Publisher Full Text | Free Full Text

8. von Mering C, Krause R, Snel B, *et al.*: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature.* 2002; **417**(6887): 399–403.
PubMed Abstract | Publisher Full Text

9. Deane CM, Salwiński Ł, Xenarios I, *et al.*: **Protein interactions: two methods for assessment of the reliability of high throughput observations.** *Mol Cell Proteomics.* 2002; **1**(5): 349–356.
PubMed Abstract | Publisher Full Text

10. Berggård T, Linse S, James P: **Methods for the detection and analysis of protein-protein interactions.** *Proteomics.* 2007; **7**(16): 2833–2842.
PubMed Abstract | Publisher Full Text

11. Walhout AJ, Vidal M: **High-throughput yeast two-hybrid assays for large-scale protein interaction mapping.** *Methods.* 2001; **24**(3): 297–306.
PubMed Abstract | Publisher Full Text

12. Spellman PT, Sherlock G, Zhang MQ, *et al.*: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell.* 1998; **9**(12): 3273–3297.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Ge H, Liu Z, Church GM, *et al.*: **Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*.** *Nat Genet.* 2001; **29**(4): 482–488.
PubMed Abstract | Publisher Full Text

14. Grigoriev A: **A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*.** *Nucleic Acids Res.* 2001; **29**(17): 3513–9.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Kemmeren P, van Berkum NL, Vilo J, *et al.*: **Protein interaction verification and functional annotation by integrated analysis of genome-scale data.** *Mol Cell.* 2002; **9**(5): 1133–1143.
PubMed Abstract | Publisher Full Text

16. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein-protein interactions.** *Genome Res.* 2002; **12**(1): 37–46.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Sharan R, Suthram S, Kelley RM, *et al.*: **Conserved patterns of protein interaction in multiple species.** *Proc Natl Acad Sci U S A.* 2005; **102**(6): 1974–1979.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Oyelade J, Ewejobi I, Brors B, *et al.*: **Computational identification of signalling pathways in *Plasmodium falciparum*.** *Infect Genet Evol.* 2011; **11**(4): 755–764.
PubMed Abstract | Publisher Full Text

19. Wilkinson DJ: **Bayesian methods in bioinformatics and computational systems biology.** *Brief Bioinform.* 2007; **8**(2): 109–16.
PubMed Abstract | Publisher Full Text

20. Thomas S: **Likelihood functions**. Indira Gandhi Institute of Development Research (IGIDR), Bombay. 2008.
Reference Source

21. Bebek G, Yang J: **PathFinder: mining signal transduction pathway segments from protein-protein interaction networks.** *BMC Bioinformatics.* 2007; **8**: 335.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Tornow S, Mewes HW: **Functional modules by relating protein interaction networks and gene expression.** *Nucleic Acids Res.* 2003; **31**(21): 6283–6289.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Hanisch D, Zien A, Zimmer R, *et al.*: **Co-clustering of biological networks and gene expression data.** *Bioinformatics.* 2002; **18**(Suppl 1): S145–S154.
PubMed Abstract | Publisher Full Text

24. Ideker T, Ozier O, Schwikowski B, *et al.*: **Discovering regulatory and signalling circuits in molecular interaction networks.** *Bioinformatics.* 2002; **18**(Suppl 1): S233–S240.
PubMed Abstract | Publisher Full Text

25. Zien A, Kuffner R, Zimmer R, *et al.*: **Analysis of gene expression data with pathway scores.** *Proc Int Conf Intell Syst Mol Biol.* 2000; **8**: 407–417.
PubMed Abstract

26. Fagarasanu A, Mast FD, Knoblach B, *et al.*: **Myosin-driven peroxisome partitioning in *S. cerevisiae*.** *J Cell Biol..* 2009; **186**(4): 541–554.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Galdieri L, Mehrotra S, Yu S, *et al.*: **Transcriptional regulation in yeast during diauxic shift and stationary phase.** *OMICS.* 2010; **14**(6): 629–38.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Steffen M, Petti A, Aach J, *et al.*: **Automated modelling of signal transduction networks.** *BMC Bioinformatics.* 2002; **3**: 34.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Fagarasanu A, Fagarasanu M, Eitzen GA, *et al.*: **The peroxisomal membrane protein Inp2p is the peroxisome-specific receptor for the myosin V motor Myo2p of *Saccharomyces cerevisiae*.** *Dev Cell.* 2006; **10**(5): 587–600.
PubMed Abstract | Publisher Full Text

30. Mortimer RK, Contopoulou CR, King JS: **Genetic and physical maps of *Saccharomyces cerevisiae*, Edition 11.** *Yeast.* 1992; **8**(10): 817–902.
PubMed Abstract | Publisher Full Text

31. Tang F, Kauffman EJ, Novak JL, *et al.*: **Regulated degradation of a class V myosin receptor directs movement of the yeast vacuole.** *Nature.* 2003; **422**(6927): 87–92.
PubMed Abstract | Publisher Full Text

32. Ishikawa K, Catlett NL, Novak JL, *et al.*: **Identification of an organelle-specific myosin V receptor.** *J Cell Biol.* 2003; **160**(6): 887–97.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Puig S, Pérez-Ortín JE: **Stress response and expression patterns in wine fermentations of yeast genes induced at the diauxic shift.** *Yeast.* 2000; **16**(2): 139–48.
PubMed Abstract | Publisher Full Text

34. Kurlandzka A, Rytka J, Rózalska B, *et al.*: ***Saccharomyces cerevisiae* IRR1 protein is indirectly involved in colony formation.** *Yeast.* 1999; **15**(1): 23–33.
PubMed Abstract | Publisher Full Text

35. Tóth A, Ciosk R, Uhlmann F, *et al.*: **Yeast cohesin complex requires a conserved protein, Eco1p(Ctf7), to establish cohesion between sister chromatids during DNA replication.** *Genes Dev.* 1999; **13**(3): 320–33.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Laurent BC, Yang X, Carlson M: **An essential *Saccharomyces cerevisiae* gene homologous to SNF2 encodes a helicase-related protein in a new family.** *Mol Cell Biol.* 1992; **12**(4): 1893–902.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Tsuchiya E, Uno M, Kiguchi A, *et al.*: **The *Saccharomyces cerevisiae* NPS1 gene, a novel CDC gene which encodes a 160 kDa nuclear protein involved in G2 phase control.** *EMBO J.* 1992; **11**(11): 4017–26.
PubMed Abstract | Free Full Text

38. Guacci V, Koshland D, Strunnikov A: **A direct link between sister chromatid cohesion and chromosome condensation revealed through the analysis of *MCD1* in *S. cerevisiae*.** *Cell.* 1997; **91**(1): 47–57.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Heo SJ, Tatebayashi K, Kato J, *et al.*: **The *RHC21* gene of budding yeast, a homologue of the fission yeast *rad21*$^+$ gene, is essential for chromosome segregation.** *Mol Gen Genet.* 1998; **257**(2): 149–56.
PubMed Abstract | Publisher Full Text

40. Michealis C, Ciosk R, Nasmyth K: **Cohesins: chromosomal proteins that prevent premature separation of sister chromatids.** *Cell.* 1997; **91**(1): 35–45.
PubMed Abstract | Publisher Full Text

41. Makolo AU, Olagunju TA: **Dataset 1 in: Computational identification of signaling pathways in protein interaction networks.** *F1000Research.* 2015.
Data Source

42. Makolo AU, Olagunju TA: **Dataset 2 in: Computational identification of signaling pathways in protein interaction networks.** *F1000Research.* 2015.
Data Source

Open Peer Review

Current Referee Status: ? ?

Version 1

Referee Report 27 January 2016

doi:10.5256/f1000research.8174.r11808

? **Lynn Fink**
Diamantina Institute, University of Queensland, Brisbane, Australia

This paper describes a computational method for extracting information from a large variety of inherently noisy biological data describing protein-protein interactions and purports to be able to discover signalling pathways, or at least segments of signalling pathways.

Not being an expert on Bayesian modeling, I can't comment directly on the method although it seems to be predicated on well-supported hypotheses and aims to be conservative in the interests of decreasing noise and increasing biological validity. Significantly, the pathway segments suggested by the model are annotated with existing functional information from Gene Ontology annotations. The authors claim to validate their results by correlating the proposed interactions with existing gene expression relying on the hypothesis that highly co-expressed genes are true interactors. It should also be noted that this study was performed on *S. cerevisiae*, a highly studied model organism for which the authors had access to 8 different types of high-throughput methods aimed at inferring protein-protein interactions (PPIs).

I've personally always struggled with the validity of using computational methods to amalgamate high-throughput PPI data for the purposes of pathway discovery. PPI networks are dynamic and I'm not convinced that we can measure them completely (in every condition, cell type, tissue type, etc.) or that we can always assume that PPIs observed in one cell, organism, or condition can be extrapolated to others so any attempt we make to catalog PPIs is necessarily vastly incomplete. Furthermore, given the wealth of data necessary to attempt applying a computational method I wonder how generally applicable these methods can be. For example, this paper relied on data from 8 different methods - how often can we expect to have that much information about a cell or organism of interest?

Is it appropriate to validate computationally-derived PPIs with gene expression data? Would it not be more appropriate to perform an assay that directly or indirectly interrogates the actual interaction between proteins?

I also wonder how circular the logic behind these computational methods is. The authors used data from SGD, a well-known public resource, in order to generate PPI networks and then bootstrapped these networks by using GO, another well-known public resource. I would be surprised if GO annotation was performed without knowledge from SGD so can we believe that the networks derived in this paper are based on solely on the Bayesian model? Or are we just re-discovering information we partially already knew?

And if we believe the signalling pathway segments reported here are newly and independently

discovered, how widely applicable is the proposed method? Can we use it for other organisms or for yeast under changed conditions, for example? How much PPI data do we need before a computational method is more efficient and informative than well-designed biochemical experiments? Were the two reported pathway segments the only ones that could be inferred from 22,650 interactions between 2554 proteins (roughly half of the entire proteome)? Is there anything exciting (and new) to be found if the model is allowed to be less conservative?

What contribution to biology do the authors expect from this method?

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Referee Report 18 January 2016

doi:10.5256/f1000research.8174.r11809

? **Winston Hide**
Sheffield Institute of Translational Neuroscience, University of Sheffield, Sheffield, USA

This work is appropriately applied in principle with adequate application of methods. According to the criteria: "work has been well designed, executed and discussed" it has not quite yet been adequately designed would benefit from a more synthetic discussion that explores the results in context of existing work.

Of concern is that the aim of the project appears to be to improve the delivery of signal over noise in PINs. But there is no means to judge if there has been an improvement - no tests, validation or comparison over a start state. Instead there is provision of results that show some interactions that are already known - which is promising - but no ability to judge if this is an improvement over just the use of protein-protein interaction data, or just the use of gene expression data.

There could be some more reference to existing work - reference and comparison with that which is current in the field (see examples of refs below).

References

1. Wang Y, Sun H, Du W, Blanzieri E, Viero G, Xu Y, Liang Y: Identification of essential proteins based on ranking edge-weights in protein-protein interaction networks.*PLoS One*. 2014; **9** (9): e108716 PubMed Abstract | Publisher Full Text
2. Ou-Yang L, Dai DQ, Zhang XF: Protein complex detection via weighted ensemble clustering based on Bayesian nonnegative matrix factorization.*PLoS One*. 2013; **8** (5): e62158 PubMed Abstract | Publisher Full Text
3. Ahmed KS, Saloma NH, Kadah YM: Improving the prediction of yeast protein function using weighted protein-protein interactions.*Theor Biol Med Model*. 2011; **8**: 11 PubMed Abstract | Publisher Full Text
4. Li D, Liu W, Liu Z, Wang J, Liu Q, Zhu Y, He F: PRINCESS, a protein interaction confidence evaluation system with multiple data sources.*Mol Cell Proteomics*. 2008; **7** (6): 1043-52 PubMed Abstract | Publisher Full Text

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: I am a member of the APBioNet - a group of African Bioinformatics scientists - this group is affiliated with the same network. I do not share collaborations with this group. I have co-published with them in a recent marker paper describing the network.