Journal of Bioinformatics and Computational Biology
Vol. 12, No. 4 (2014) 1450014 (24 pages)

DOI: 10.1142/S0219720014500140

ICP Imperial College Press
www.icpress.co.uk

Using data fusion for scoring reliability of protein–protein interactions

Alireza Vazifedoost[*,‖], Maseud Rahgozar[†,**], Behzad Moshiri[†,††], Mehdi Sadeghi[‡,‡‡], Hon Nian Chua [§,§§], See Kiong Ng[§,¶¶] and Limsoon Wong[¶,‖‖]

[*]*School of Electrical and Computer Engineering*
University College of Engineering
University of Tehran, Tehran, Iran

[†]*Control and Intelligent Processing Center of Excellence*
School of Electrical and Computer Engineering
University College of Engineering
University of Tehran, Tehran, Iran

[‡]*National Institute of Genetic Engineering and Biotechnology*
Tehran, Iran

[§]*Institute for Infocomm Research, Singapore*

[¶]*School of Computing, National University of Singapore*
13 Computing Drive, Singapore 117417
[‖]*vazifehdst@ut.ac.ir*
[**]*Rahgozar@ut.ac.ir*
[††]*moshiri@ut.ac.ir*
[‡‡]*sadeghi@nigeb.ac.ir*
[§§]*hnchua@i2r.a-star.edu.sg*
[¶¶]*skng@i2r.a-star.edu.sg*
[‖‖]*wongls@comp.nus.edu.sg*

Received 26 September 2013
Revised 17 January 2014
Accepted 27 May 2014
Published 1 July 2014

Protein–protein interactions (PPIs) are important for understanding the cellular mechanisms of biological functions, but the reliability of PPIs extracted by high-throughput assays is known to be low. To address this, many current methods use multiple evidence from different sources of information to compute reliability scores for such PPIs. However, they often combine the evidence without taking into account the uncertainty of the evidence values, potential dependencies between the information sources used and missing values from some information sources. We propose to formulate the task of scoring PPIs using multiple information sources as a multi-criteria decision making problem that can be solved using data fusion to model potential interactions between the multiple information sources. Using data fusion, the amount of contribution from each information source can be proportioned accordingly to systematically score the reliability of PPIs. Our experimental results showed that the reliability scores assigned by our data fusion method can effectively classify highly reliable PPIs from multiple information sources, with substantial improvement in scoring over conventional approach such as the Adjust CD-Distance approach. In addition, the underlying interactions between the information

sources used, as well as their relative importance, can also be determined with our data fusion approach. We also showed that such knowledge can be used to effectively handle missing values from information sources.

Keywords: Protein–protein interaction; reliability; Choquet fuzzy integral; data fusion; missing information.

1. Introduction

Protein–protein interactions (PPIs) provide invaluable insights for studying the underlying cellular mechanisms of biological functions. In recent years, much efforts have been focused on high-throughput PPI screening technologies such as yeast two-hybrid assays, and they have resulted in an unprecedented abundance of PPI data available for research. However, several systematic assessment studies on the PPI data have shown that 50% of the interactions are false positives.[1,2] The dismal reliability of PPI data is similar in terms of false negatives.[3]

In order to effectively utilize such error-prone PPIs derived from high-throughput experiments, one popular approach is to devise scoring methods to estimate the reliability of the PPIs. The computed scores can then be used to filter away the experimental noise from the PPI data. The scores can also serve as confidence measures over putative PPIs which are not yet experimentally detected, as a kind of prediction.

Numerous studies have been conducted for filtering and prediction of PPIs with reliability scores.[4–10] Many of these methods used a combination of features of the interacting proteins to compute the reliability scores, often on PPI data from multiple datasets. However, they differed vastly in their selected features, the techniques for integrating the various features or information sources and the PPI datasets used. There are also some other methods which calculate reliability scores for the PPIs based on the topology of the network and neighboring proteins.[11–15]

A less noticed concern in using a combination of features or information sources in determining a reliability score for PPIs is the uncertainty issue. The root of this issue is that the provided value by information sources as the evidence of a PPI's existence is always imprecise and we cannot be quite sure about it.

Another major issue often overlooked by the current approaches is the potential interactions (i.e. dependencies) amongst the various information sources used for computing the reliability scores. It is often conveniently assumed that the multiple information sources each provides an independent evidence on the reliability of a protein interaction. However, it is possible that the multiple information sources used for scoring the PPIs may have synergistic or redundant effect on determining the reliability of the PPIs. The evidence from different information sources may boost one another decision on the reliability of the protein interactions (i.e. synergistic effect), or the inclusion of one information source does not help in deciding the reliability of a PPI using existing information sources (i.e. redundant effect). Identifying and taking into account of such inherent dependencies between multiple information sources will result in better judgments about the reliability of PPIs.

The last issue is missing values from certain information sources. We may face a lack of information about some interacting proteins in terms of their location or function or other aspects which are important in deciding the genuineness of a PPI.

To mitigate these issues, we formulate the task of scoring the reliability of PPIs as a multi-criteria decision making problem. As far as we know, this is a new perspective of PPI scoring that has not been explored in previous works. To obtain a decision on the reliability (or possibility) of an interaction between two proteins involves combining uncertain information from multiple sources to satisfy the different criteria required for interaction between the proteins. This multi-criteria decision making problem can be solved systematically using data fusion. In this work, we use the Choquet fuzzy integral technique[16] for its proven ability in effectively modeling uncertainty and interactions amongst multiple information sources.

For evaluation, we compare the performance of our proposed data fusion method to the Adjust CD-Distance method[17] which has been shown currently to be one of the most effective topology-based scoring methods for PPI reliability. We show that using our data fusion method, we are able to determine the relative importance of each information source in deciding the reliability of the PPIs, and the degree of dependencies between different information sources. We also show that we can effectively specify the reliability of PPIs in the presence of missing values from the information sources by exploiting the detected dependencies of the information sources.

2. Methods and Materials

2.1. *Data fusion*

Our main idea is to employ data fusion methodology to systematically combine the information from multiple sources for making effective decisions on the reliability of PPIs. Although there is no single definition for data fusion, we give the one suggested by Dasarathy[18] for reference: "it encompasses the theory, techniques and tools created and applied to exploit the synergy in the information acquired from multiple sources (sensor, databases, information gathered by humans, etc.) in such a way that the resulting decision or action is in some sense better (qualitatively or quantitatively, in terms of accuracy, robustness, etc.) than would be possible if any of these sources were used individually without such synergy exploitation".

Figure 1 depicts the basic architecture for our data fusion system for scoring the reliability of PPIs. In our proposed framework, we suppose that there are multiple criteria that are important in deciding the reliability of PPIs. However, the importance of the criteria are not equal and there may also be some interactions amongst the criteria. In other words, we suppose that each criterion may only partially address one of the many aspects of having a genuine PPI. We further suppose that each criterion can be satisfied, in some uncertain degree, by the information in one of the given sources, as depicted in the figure. As such, we use the terms criterion (for data

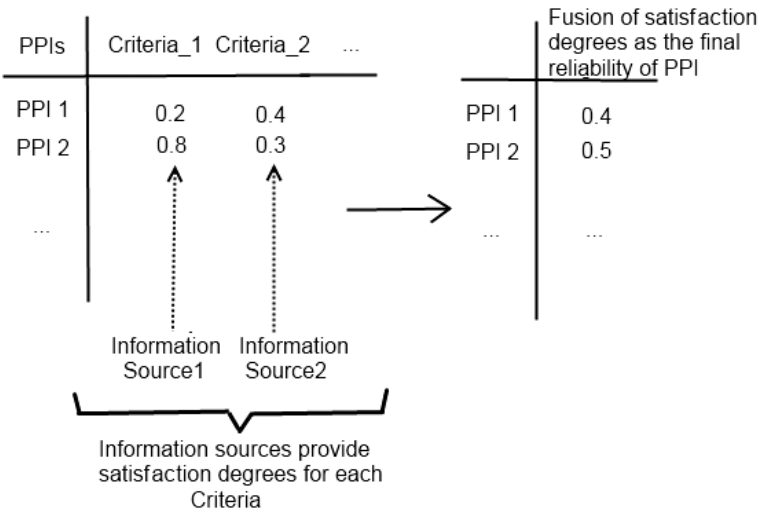


Fig. 1. A basic architecture for making decision on the reliability of PPIs by fusion of the degrees of satisfaction multiple criteria obtained from multiple information sources.

fusion) and information source (for reliability scoring) interchangeably. With our proposed data fusion method, the input values from these information sources are systematically combined into an aggregated reliability score which we call the fusion-based score (FB-Score) as the final decision on the reliability of a PPI.

The key part of the data fusion system is the fusion (or aggregation) operator that combines the input values into a single final decision value. This operator is a mathematical function that takes multiple measures in the same scale as the input and a single aggregated value as the output. An aggregation operator $C : D^N \to D$ is defined[19,20] with the following properties

(1) Idempotence: $C(a, \ldots, a) = a$,
(2) Monotonicity: $C(a_1, \ldots, a_N) \geq C(a_{1'}, \ldots, a_{N'})$ when $a_i \geq a_{i'}$
(3) Symmetry: For any permutation π on $\{1, \ldots, N\}$ it holds that $C(a_1, \ldots, a_N) = C(a_{\pi(1)}, \ldots, a_{\pi(N)})$.

Some commonly used fusion operators are arithmetic mean, weighted mean, ordered weighted average (OWA),[21] Choquet integral,[16] Sugeno integral,[22] Bayesian networks, and Dempster–Shafer[23] Each of these operators has different properties and power in modeling decisions. We use the Choquet integral for this work as it has been shown to have superior interaction modeling power.[24]

To understand Choquet integral, we need to know about fuzzy measures.[16,22] A fuzzy measure μ on a set X of features is a set function $\mu : \wp(X) \to [0, 1]$ that satisfies the following axioms:

(1) $\mu(\emptyset) = 0$, $\mu(X) = 1$ (boundary conditions),
(2) $A \subseteq B$ implies $\mu(A) \leq \mu(B)$ (monotonicity).

Fuzzy measures can be used for relating the uncertainty, power or importance of a single or subsets of criteria or information sources in decision making.

Assuming that a function f computes the degree that an information source x_i satisfies its related criterion for protein interaction (i.e. $f(x_i) = a_i$), then the Choquet integral for a function $f : X \to [0, 1]$ with respect to μ is defined by

$$C_\mu(f(x_1), \ldots, f(x_N)) = \sum_{i=1}^{n} [f(x_{s(i)}) - f(x_{s(i-1)})]\mu(A_{s(i)}),$$

where $f(x_{s(i)})$ indicates that the indices have been permuted such that $0 \leq f(x_{s(1)}) \leq \cdots \leq f(x_{s(N)}) \leq 1$, $f(x_{s(0)}) = 0$, and $A_{s(i)} = \{x_{s(i)}, \ldots, x_{s(N)}\}$. The value of the FB-Score for a PPI is therefore the output of this integral on the values provided from all the information sources on this PPI.

Basically, the Choquet integral weighs the length of the segment $f(x_{s(i)}) - f(x_{s(i-1)})$ based on the fuzzy measure value of all the information sources that supply values greater than or equal to $f(x_{s(i)})$. This function has several desirable mathematical and behavioral properties that make it superior to other similar fusion aggregators. One of which is its ability to take into account of the interactions between the information sources by using appropriate fuzzy measures to model interactions between two sources. For example, in the case of two information sources A and B that exhibits a synergistic relation of superadditivity, we can choose $\mu(\{A, B\})$ such that $\mu(\{A, B\}) > \mu(\{A\}) + \mu(\{B\})$. Similarly, it is possible to model redundancy, symmetries, veto effect, and pass effect.[24]

To identify the interaction between two information sources, we can compute an interaction index as follows[25]:

$$I_\mu(x_i, x_j) = \sum_{T \subseteq X \setminus \{x_i, x_j\}} \frac{(N - t - 2)!t!}{(N - 1)!} [(\mu(T \cup \{x_i, x_j\}) - \mu(T \cup \{x_i\})) - (\mu(T \cup \{x_j\}) - \mu(T))].$$

Here, T is a subset of all information sources, X, which does not include x_i and x_j as the two information sources whose interaction is being measured, and t is the size of such subset. Also N is the number of all information sources. This index computes the difference in the value of the fuzzy measure when the contribution from an information source x_j is removed from sets containing both x_i and x_j. If $I_\mu(x_i, x_j) < 0$, it means these two information sources have redundancy effect when they appear together. If $I_\mu(x_i, x_j) > 0$, then we have complementary or synergistic correlation between these two sources. Note that $I_\mu(x_i, x_j) \in [-1, 1]$. We use this interaction index for identifying those PPI data sources that can compensate for loss of information due to missing data in some of the PPI data sources.

Finally, we can measure the relative importance of each information source x_i by calculating the average amount of boosting a fuzzy measure achieves when we include x_i into each subset T of information sources. This is computed below as the

Shapley index[26]:

$$\phi_\mu(x_i) = \sum_{T \subseteq X \setminus \{x_i\}} \frac{(N-t-1)!t!}{(N)!} [\mu(T \cup \{x_i\}) - \mu(T)].$$

2.2. *Consolidating PPI data sources*

As evaluation, we conduct our study on Saccharomyces cerevisiae (yeast) data as it is the most studied organism used in previous studies. We put together 116498 yeast PPIs collected from five major datasets, viz. IntAct,[27] DIP and DIP-Core,[28] BioGrid,[29] and MINT.[30]

Along with the interaction data, we also kept auxiliary data like the Gene Ontology (GO)[31] and the annotations for yeast proteins from the Saccharomyces Genome Database.[a] The actual list of information sources used for this study will be introduced in the next section in more detail. All the data values from each information source are normalized by dividing by its maximum value to keep them within the same scale.

We also need a set of protein pairs that have no physical interaction as the negative set. For this, we create pairs of proteins for which there was no report of interaction in any of the five PPI data sources used. In addition, they do not have any shared annotation in any aspect of GO. This further ensures the low possibility of their interaction.

2.3. *Information sources*

Previous works have used a wide variety of information sources for assessing the reliability of PPIs such as ortholog information, gene expression data, protein domain information, phylogenetic profiles, co-localization, and topological properties.[32] In this work, we use a mixture of functional features (semantic similarities and gene expression similarities) with topological evidences, as listed in Table 1. Using our

Table 1. Information sources.

Information source	Description
FSIM_RESNIK_MAX	Functional Semantic Similarity by using Resnik method and max.
FSIM_GO_INFO	Functional Similarity by GO informative terms.
LSIM_ GO_INFO	Localization Similarity by GO informative terms.
LSIM_RESNIK_MAX	Localization Semantic Similarity by using Resnik method and max.
BSIM_RESNIK_MAX	Biological Process Similarity by using Resnik method and max.
GeneExpression	Average gene expression correlation.
GeneExpression_Pairwise	Maximum Gene expression correlation.
Adjust CD-Distance	Iterative Adjust CD-Distance score.

[a]Retrieved 2012, from SGD project: http://www.yeastgenome.org/download-data.

data fusion framework, it is straightforward to include additional information sources if desired.

2.3.1. *Semantic similarities*

The first set of the information sources that we are using is based on the semantic similarity of GO annotations of the proteins. The information sources in this set are listed as FSIM_RESNIK_MAX, LSIM_RESNIK_MAX and BSIM_RESNIK_MAX in Table 1.

We use the Resnik method[33] to compute the semantic similarity of GO terms based on the information content of the concepts represented by the GO terms. The method had been investigated in several works[34] and shown to be working better than other methods. We briefly describe the method here. Let C be a set of concepts in a IS-A taxonomy like GO, which permits multiple inheritance. Let the taxonomy be augmented with a function $P : C \to [0, 1]$ such that for any $c \in C$, $P(c)$ is the probability of encountering an instance of concept c which, in our case, is a protein annotated with c. Then, the information content of concept c is $-\log\ P(c)$. It is evident that as the probability increases, the informativeness of the GO term decreases. Also, the more abstract a concept, the lower is its information content.

Similarity between two GO terms can then be measured by the information shared by the two terms, which is indicated by the information content of the concepts that subsume them in taxonomy (parents) as shown by this formula, where $S(c_1, c_2)$ is the set of parents of both c_1 and c_2:

$$\mathrm{sim}(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log p(c)].$$

To compute the similarity of two proteins in terms of their annotation, we use the maximum operator as suggested in Ref. 34. This operator chooses the maximum score of similarity of two annotating terms as the final score. The three features FSIM_RESNIK_MAX, BSIM_RESNIK_MAX, and LSIM_RESNIK_MAX in our Table 1 are calculated for each pair of interacting proteins respectively based on their annotations in the Molecular Function (MF), Biological Process (BP), and Cellular Compartment (CC) aspects of GO.

We also use the notion of *informative GO terms*[35] to compute semantic similarity between two proteins based on their GO annotations. These are GO terms that are annotated explicitly or implicitly (via the so-called through-path rule using the GO hierarchy) to more than 30 proteins, but none of their children is annotated explicitly or implicitly to more than 30 proteins. As such, these GO terms are considered to be informative as they are neither too general nor too narrow. We have found around 300 GO terms which match this definition of informative GO terms. For measuring the semantic similarity of two proteins, we consider only those explicit and implicit annotations with informative GO terms and use the Jaccard coefficient formula to compute the similarity of two proteins. The more explicit and implicit informative GO term annotations two proteins share, the more similar they are.

Based on this approach, we generated two other information sources for computing the reliability of PPIs. The first one is FSIM_GO_INFO, which is the similarity of proteins based on their explicit and implicit informative GO term annotations in the BP and MF aspects. The second one is the similarity of proteins based on their explicit and implicit informative GO term annotations in the CC aspect; we call it LSIM_GO_INFO.

2.3.2. *Gene expression similarities*

A second set of information sources that we use is based on the assumption that interacting proteins exhibit correlations in their gene expression profiles. There are two information sources in this set: GeneExpression and GeneExpression_Pairwise (see Table 1).

We use the SPELL database and search tool[36] for finding the gene expression correlations for each pair of proteins. The database contains 117 microarray datasets from 81 publications. The gene expression correlation within each dataset were calculated using the Pearson correlation coefficient. Since many pairwise Pearson correlations can be computed between datasets, the Fisher's z-transform was applied over the correlations for better comparability. When we query the database with a set of genes, the datasets are weighted based on co-expression level of the queried genes. Higher weights will be given to the datasets in which the queried genes are largely co-expressed. The database also returns a weighted correlation for all the other genes in the genome with respect to the genes in the query set. For reference, we show the formulas for finding the weight of each dataset and the final ranking formula given in Ref. 36 below:

$$w_d = \left(\frac{2}{|Q|(|Q|-1)}\right) \sum_{i=1}^{|Q|-1} \sum_{j=i+1}^{|Q|} f(z_{q_i,q_j}) \tag{1}$$

$$f(z) = \begin{cases} z^2 & \text{if } z \leq 1 \\ 1 & \text{otherwise} \end{cases}$$

$$s_x = \frac{1}{|Q| \sum_{d \in D} w_d} \sum_{d \in D} \sum_{q \in Q} w_d f(z_{x,q}).$$

In the above formulas, w_d is the weight of each dataset d among the set of all datasets D, $q_i \in Q$ is a query gene and z_{q_i,q_j} is the z-transformed correlation of two genes. S_x is the final score of the correlation of a gene x to the set of queried genes.

We extract two different gene expression values from the SPELL engine for a pair of genes corresponding to a PPI: Their average and maximum gene expression correlation in the SPELL datasets. To compute the average correlation of the pair of genes, we give one of those interacting genes as the query, which causes SPELL to return all the other genes ranked by their correlation score in the datasets. Note that by a single gene query, SPELL considers all datasets as equally weighted in

calculating the correlation scores. We obtain the correlation scores of the query gene with the second gene for each dataset from the result list and compute the average.

To compute the maximum gene expression correlation of the interacting gene pair, we give both interacting proteins q_i and q_j as the query. This causes SPELL to return the ranked list of datasets based on their relevance weight (w_d) and the correlation of other genes with the queried genes based on formula 1. Although this does not give us directly the maximum correlation of the queried genes, by using the maximum weight in the ranked list of datasets and reversing formula 1, we can obtain z_{q_i,q_j} for q_i and q_j.

The average and maximum gene expression correlation scores are listed as the GeneExpression and GeneExpression_Pairwise information sources in Table 1.

2.3.3. *Topological similarities*

The final information source we used is the Adjust CD-Distance value for each pair of proteins. CD-Distance is a formula[37] for finding similarity of proteins in a PPI network based on topological features. Two proteins having a larger number of shared neighborhood proteins will be given a bigger value of similarity between two proteins. The iterated version of this formula,[17] effectively enhances this classic approach. In our study, we use the score achieved by two iterations as additional iterations would not improve the score dramatically.[17]

2.4. *Specifying the fuzzy measures*

One of the fundamental issues of using fuzzy integral methods is specifying the fuzzy measures for modeling the interactions between the possible combinations of all the given information sources. This means that we need to calculate a fuzzy measure value for every subset of the given set of information sources, or $2^N - 2$ fuzzy measures if we have N sources. Traditionally, an expert may use his knowledge to determine these measures to control the fusion process. However, this is not applicable here as we are not sure about the potential interactions between the information sources.

As such, we have to find the fuzzy measure by learning. To do so, we gather the evidences from the different sources of information into a training data set as follows:

$$\begin{matrix} a_1^1 & a_2^1 & \cdots & a_N^1 & y^1 \\ a_1^2 & a_2^2 & \cdots & a_N^2 & y^2 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ a_1^j & a_2^j & \cdots & a_N^j & y^M. \end{matrix}$$

Here, each row is a PPI, and each element a_i^j, corresponds to the value from information source x_i provided for jth PPI. Also, in the training set, we include a value for y^j which is a predetermined reliability score (PR-Score) for each PPI. We will explain how to compute this score in the next section.

To learn the fuzzy measures from this training set, we can use least square distance to minimize the value of $E^C(\mu)$ which is the distance error between each $C^j_\mu(a^j_1, \ldots, a^j_N)$ (the value given by fuzzy integral function) and y^j. The error formula for minimization is:

$$E_C(\mu) = \sum_{j=1}^{M} (C_\mu(a^j_1, \ldots, a^j_N) - y^j)^2.$$

This is essentially a quadratic optimization problem with constraints to maintain the monotonicity property of the fuzzy measures. For solving this quadratic optimization problem and finding fuzzy measures, we use a tool named Kapalab,[38] which is specifically designed for finding fuzzy measures implemented in the R language.

2.5. *Predetermined Reliability Score* (*PR-Score*)

As described above, our method for finding fuzzy measures requires an externally provided score of reliability (PR-Score) for each PPI. To obtain this PR-Score, we propose a method that combines evidences from two sources of information. We call the first one Publication Confidence (PC), and the second one Throughput Confidence (TC).

PC is based on the reproducibility of PPIs, and is inferred from the supporting publication for each PPI by this formula:

$$\mathrm{PC(PPI)} = \left(1 - \frac{1}{\mathrm{ReportNum} + 1}\right) \times \left(1 - \frac{1}{\mathrm{TestNum} + 1}\right).$$

In this formula, ReportNum is the number of times that two proteins are reported to have physical interaction in publications. It is based on the intuition that the more an interaction is reported, the more likely it is a real one. ReportNum can be easily found since we have kept the Pubmed_Id field of each interaction from the original database in our consolidated database (Sec. 2.2). We count the number of distinct Pubmed_Ids for each pair of interacting proteins. Note, that we did not count the occurrence of a PPI across multiple databases as they may be coming from the same publication.

The other parameter in PC, TestNum, is the number of publications in which two proteins are mentioned together. The intuition is that the set of common publications might report experimental tests for interaction of the two proteins. Even though TestNum is a very rough estimation, our results show that it is helpful to keep this parameter.

The other information source named TC captures the type of experiment of a reported PPI. We classify the experimental methods into two groups: High-throughput and low-throughput, and we assume that low-throughput experiments are generally more reliable than high-throughput experiments. In our consolidated database, a PPI may be reported to have been detected by different types of experiments. We use this as a clue of the reliability of detected PPIs and compute the

TC of a PPI as follows:

$$\begin{cases} \text{high-throughput} & \alpha \\ \text{low-throughput} & \beta \\ \text{high-throughput and low-throughput} & \gamma \end{cases}$$

$$\text{TC(PPI)} = \min\Big(1, \sum \text{Scores\_of(PPI)}\Big).$$

In our implementation, we considered $\alpha = 0.2$, $\beta = 0.8$, $\gamma = 1$, in accordance to estimates reported elsewhere.[14]

Finally, the PR-Score is defined as the weighted sum of TC and PC:

$$\text{PR-Score} = 0.1 \times \text{TC} + 0.9 \times \text{PC}.$$

These weights are determined empirically based on maximizing functional homogeneity and localization coherence of ranked PPIs. For all the PPIs in the negative dataset, we assume a PR-Score equal to 0.

2.6. *Dealing with missing values*

One of the challenges in specifying the reliability score for PPIs is dealing with missing values from some information sources. For example, there are many proteins that have no GO annotations, making it impossible to compute the semantic similarity between these proteins.

To address this, let us suppose that we have already found a fuzzy measure function μ (as described in Sec. 2.4) by using the portion of PPI data that has no missing values, and we want to use it to estimate a new fuzzy measure when we have one or more missing information sources. That is, suppose we have $\mu : \wp(X) \to [0, 1]$ as the current fuzzy measure over X which is the set of all information sources. We want to estimate μ' for X' the set of information source without an information source x_i: $\mu' : \wp(X') \to [0, 1]$ for $X' = X \setminus \cup\, x_i$, $i \in 1 \ldots n$. We can compute μ' as follows:

(1) Add a dummy information source x_d to X ($f(x_d) = 0$),
(2) For each proper subset $A \subset X$, if $(x_d \in A)$ then $\mu'(A) = \mu(A \setminus \{x_d\})$ else $\mu'(A) = \mu(A)$,
(3) For the subset $A = X$, to satisfy the boundary condition required by fuzzy measures, we set $\mu(A) = 1$.

3. Experiments

We performed a set of experiments to study the effects of using data fusion in the context of scoring PPIs. The first set of experiments verifies that the PR-Score can rank the PPIs reasonably well. Then, we investigate the performance of FB-Score for scoring the reliability of PPIs. We identify the underlying interactions (i.e. dependencies) amongst the different information sources used for ranking the PPIs, and compute the importance index of each source for determining the reliability of PPIs.

Finally, we study the scoring performance of the fuzzy integral in the presence of missing values.

3.1. *Validity of PR-Score*

To verify the validity of PR-Score as an indicative reliability score for PPIs, we used the functional homogeneity and localization coherence of the protein pairs in the PPIs sorted based on their PR-Scores in descending order. If our PR-Scores are effective, we can expect a descending trend of these two measures. This method was previously used[12–14] for evaluating the effectiveness of a reliability score.

Functional homogeneity can be defined in the following way:

$$Functionalhomogeneity = \frac{\#proteins\ pairs\ sharing\ at\ least\ one\ functional\ annotation}{\#protein\ pairs\ that\ have\ functional\ annotation}.$$

Localization coherence can be defined similarly:

$$LocalizationCoherence = \frac{\#protein\ pairs\ sharing\ at\ least\ one\ localization\ annotation}{\#protein\ pairs\ that\ have\ localization\ annotation}.$$

In this work, we only consider explicit and implicit annotations by informative GO terms when finding shared annotations in the above formulas.

We also verify whether the top-ranking interactions are more likely to have interologous interactions in other organisms. We use I2D[39] for finding interologous interactions. This database contains interologous interaction data for five species including yeast. We use the formula below to assess the Interologous Rate:

$$InterologousRate = \frac{\#protein\ pairs\ having\ at\ least\ one\ interolog}{\#proteins\ pairs}.$$

For this experiment, we used 46,413 PPIs in our consolidated database which are in common with the I2D database of interologous PPIs.

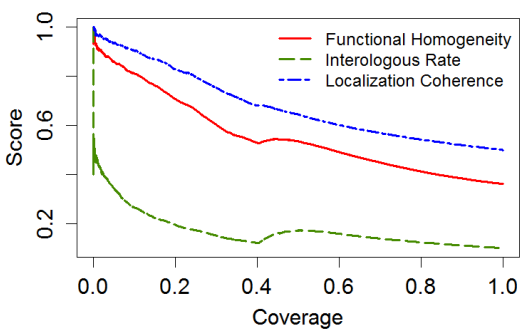


Fig. 2. Decreasing trend of functional homogeneity, interologous rate, and localization coherence when we sort PPIs in descending order of their PR-Score.

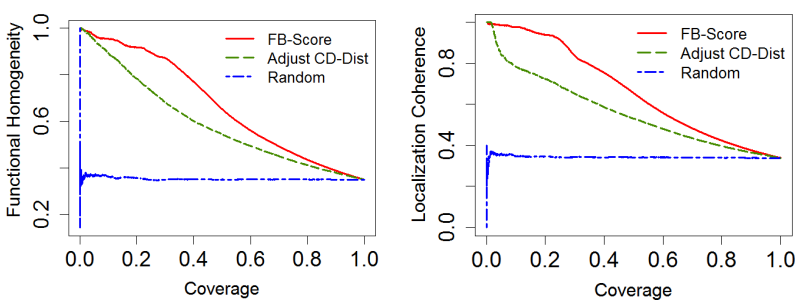
Figure 2 shows decreasing trends for all three measures, which is what we would expect for effective ranking of PPI reliability when we go from highly scored interactions down to the lower ones. In addition, we can observe that the interologous rate is almost constantly decreasing. This is another strong evidence of validity of PR-Score.

Although the PR-Score was computed using very simple heuristic information, the results show that it can serve the purpose of seeding the scoring of the PPIs to be used in specifying fuzzy measures. Of course, PR-Score cannot be used for evaluating those PPIs that have no associated publications.

3.2. *Performance of FB-Score*

Next, we compare FB-Score with Adjust CD-Distance to see whether fusing multiple information sources with Adjust CD-Distance can improve the scoring of PPIs. This comparison is made based on functional homogeneity and localization coherence of these two methods, as in the previous section. We randomly choose 20% of all PPIs without missing values as the training set for determining the fuzzy measures. Then, we score the remaining 24,796 PPIs based on the trained fusion model and compared their assigned FB-Scores with the Adjust CD-Distance scores. For comparison, we sort the PPIs in descending order of their FB-Scores and plot the corresponding changes in the functional homogeneity and localization coherence of the PPIs. We also do the same for PPIs based on their Adjust CD-Distance. Note that when we compare the scores based on functional homogeneity, we removed the FSIM_GO_INFO and FSIM_RESNIK_MAX information sources from the training dataset. Likewise, we remove LSIM_GO_INFO and LSIM_RESNIK_MAX when we are comparing against localization coherence.

The results are shown in Figs. 3(a) and 3(b) respectively. In both figures, the functional homogeneity and localization coherence for PPIs that are assigned with



(a) (b)

Fig. 3. Performance of FB-Score in comparison to Adjust CD-Distance. The PPIs in the test set are sorted in descending reliability scores. (a) Comparison based on rate of functional homogeneity. The information sources related to functional aspects of proteins were removed from the training set of fuzzy measures. (b) Comparison based on rate of localization coherence. Here, the information sources related to localization aspects of proteins were removed from the training set.

random ranking are about 40%, which is in accordance with previously reported works.[14] In comparison, the functional homogeneity of PPIs ranked using our FB-Score is significantly improved even when we do not use any information source related to functional properties of two proteins, as shown in Fig. 3(a). Figure 3(b) again shows similar improvement in terms of localization coherence in comparison with the Adjust CD-Distance method. The results show that fusing multiple sources of information along with Adjust CD-Distance can improve ranking of PPIs with respect to their reliability.

3.3. *Relative importance and interactions*

As presented in Fig. 4, we use the fuzzy measures to calculate Shapley values as the important index of each information source. We also calculate the interaction indices amongst the information sources. The results are shown in Fig. 5. Here, we use the PPIs with a PR-Score higher than 0.8 (there are about 2000 such interactions) as the positive dataset along with the same amount of PPIs from the negative interaction dataset as the training set for computing the fuzzy measures.

The results in Fig. 4 suggest that localization coherence and functional homogeneity (which are based on informative GO terms) are the two most important information sources. The features based on the semantic similarity of GO terms are in the middle-range of importance ranking, suggesting that they are less effective than the features based on the informative GO terms. Adjust CD-Distance is in the third place with a high importance value, which shows that such topological scores can be quite effective. Interestingly, the gene expression information sources were shown to have low importance, which may be due to the relatively high noise in this kind of data.

The pairwise interactions among the various information sources are also informative. As we can see in Fig. 5, it is interesting to note that the pair of information sources that has the highest value of redundancy is FSIM_GO_INFO and LSIM_GO_INFO which also happened to be the most important evidences at the same time. The Adjust CD-Distance, which is the third important information

	Importance
FSIM_RESNIK_MAX	0.0643
FSIM_GO_INFO	0.2172
LSIM_GO_INFO	0.2257
LSIM_RESNIK_MAX	0.1032
BSIM_RESNIK_MAX	0.1117
GeneExpression	0.0445
GeneExpression_Pairwise	0.0665
Adjust_CD.Distance	0.1669

Fig. 4. Relative importance of each information source based on its Shapley value.

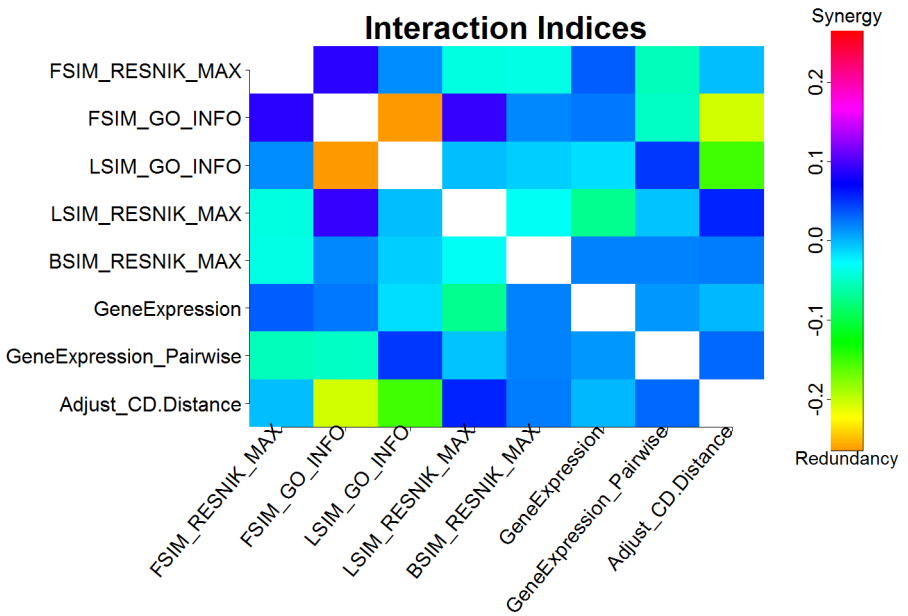


Fig. 5. Pairwise interaction indices between information sources. The legend shows the spectrum of interaction values. Therefore, dark blue depicts synergistic effect while green and orange color depict redundancy of information sources.

source, is also in redundancy relation with these two information sources. On the other hand, in terms of synergistic relation, we can see that there are relatively high synergies between Adjust CD-Distance and LSIM_RESNIK_MAX and GeneExpression_Pairwise. This situation is also the case for information sources FSIM_GO_INFO, FSIM_RESNIK_MAX, and LSIM_RESNIK_MAX.

The importance and interaction indices of information sources provide us useful insights for selecting proper information sources in the fusing process for making the final decision. With these insights, fuzzy integral can handle missing values effectively, as we will explain it in further detail in the next section.

3.4. *Dealing with missing values*

One of the challenges of applying data fusion method is dealing with missing values in some of the information sources. Since we are employing fuzzy measure for data fusion here, we can take advantage of its ability to detect interactions between the information sources to compensate for the loss of information appropriately.

To investigate the effect of missing values on the fusion method, we remove information sources one at a time, and compute the FB-Scores with the remaining information sources. For comparison, we plot the resulting distributions of FB-Scores with and without removing that information source. The FB-Scores were calculated with fuzzy measures obtained as described in the experiment in Sec. 3.3.

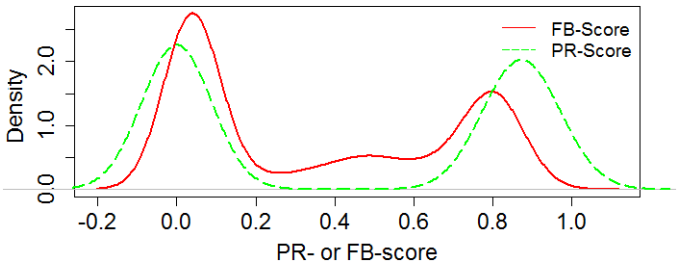


Fig. 6. Distributions of PR-Score and FB-Score for assessing performance of FB-Score in terms of the ability to discriminate highly possible PPIs from protein pairs with low chance of interaction. Having similar distributions means better performance of FB-Score in this task.

Figure 6 shows that the distribution of FB-Scores is similar to the distribution of PR-Scores when we do not remove any features. Figures 7(a)–7(h) show the changes (if any) in the distributions of FB-Scores when we remove each of the information sources one at a time. The figure shows that the fuzzy integral method for fusing information sources is mostly working well in the presence of missing values. In fact, apart from some drifts in Figs. 7(d) and 7(e), which corresponds to removing information sources LSIM_GO_INFO and LSIM_RESNIK_MAX, the other results in the figure showed almost similar distributions of the predicted scores with and without missing values.

We can explain why removing some information sources has little effect while removing others may have large effects on the FB-Score, based on the importance of each information source in terms of its Shapley value and its interactions (redundancy or synergy) with the other information sources according in terms of the corresponding interaction indices. Removing features that are more important affects more adversely the FB-Scores, as can be seen by removing LSIM_GO_INFO. On the other hand, removing GeneExpression has nearly no effect on FB-Score.

It may seem strange at the first look that removing FSIM_GO_INFO has such a slight effect. This can be explained based on its interactions with other information sources. FSIM_GO_INFO has strong redundancy with two other important information sources, LSIM_GO_INFO and Adjust CD-Distance, which makes removing it quite safe. Another good example is LSIM_RESNIK_MAX. While it has almost the same importance as FSIM_RESNIK_MAX, the effect of its removal is much more intense than removing FSIM_RESNIK_MAX. This is because LSIM_RESNIK_MAX has synergy with FSIM_GO_INFO and Adjust CD-Distance, which are two important information sources. As a result, removing LSIM_RESNIK_MAX results in more drift, as can be seen in Fig. 7(d).

Using the method that we have introduced for handling missing values in Sec. 2.6, when we remove an information source, the importance of the remaining information sources will be adjusted as follows: The importance of the information sources which are redundant to the removed information source will increase, while the importance of sources which are in synergistic relation with the removed source will decrease.

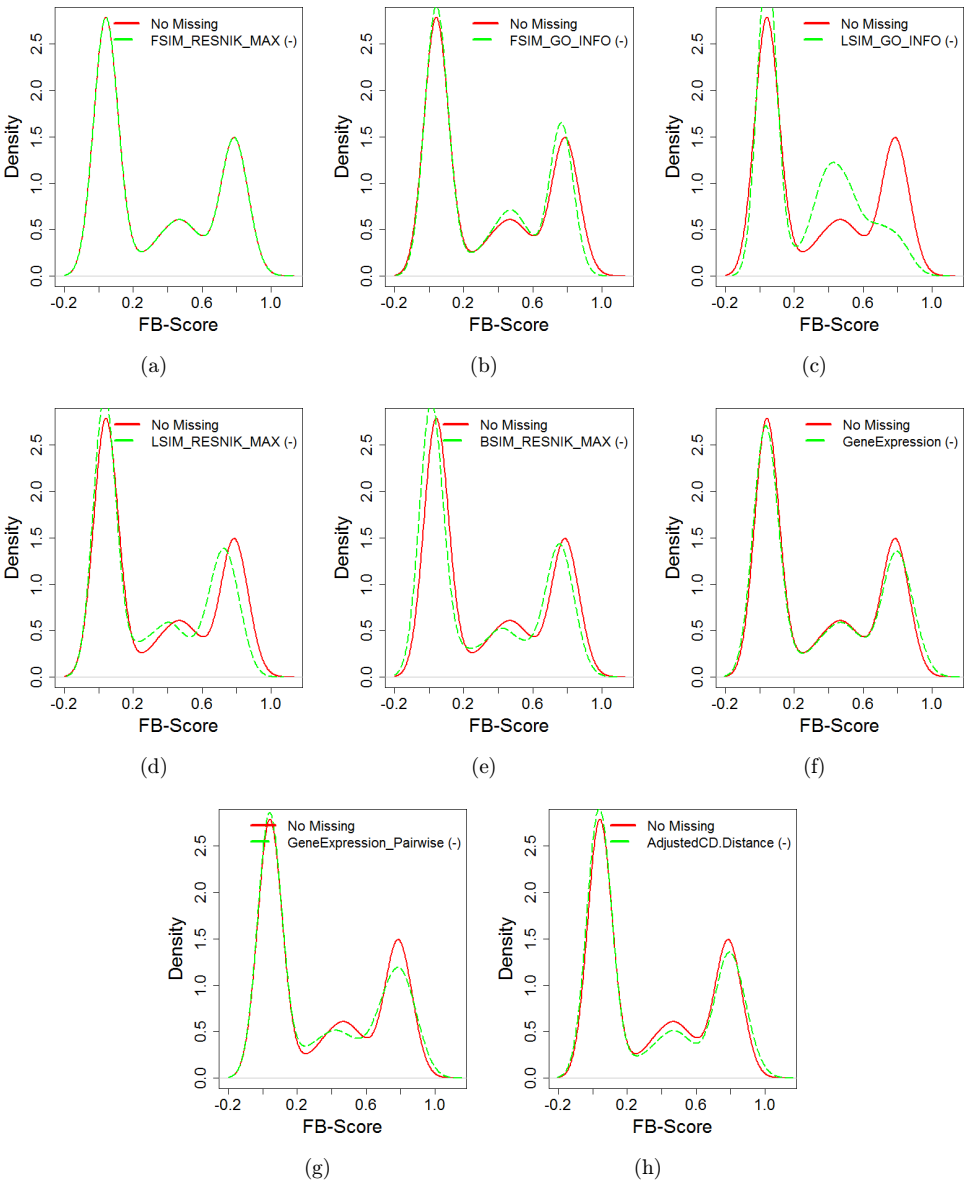


Fig. 7. Changes in distribution of the assigned FB-Scores by removing information sources, one at a time. (a)–(h) each of the plots shows the effect removing information sources on distribution of FB-Score.

Therefore, if a removed information source has high redundancies with many important information sources and also high synergies with the less important sources, we can remove it safely. Conversely, if an information source is in synergistic interactions with many important information sources and in redundant interactions

with the less important information sources, removing it has more adverse effect on the predicted reliability scores.

We also investigate the effect of missing multiple information sources instead of just one. We repeat the previous experiment three times, each time removing a group of information sources: The first group includes information sources related to functional aspects of proteins, the second group includes the information sources of localization similarity, and the third includes information sources of gene expression correlation. The results are shown in Fig. 8. As the immediate conclusion from this result, we can notice that our method can handle very well loss of information from functional annotations and gene expression. However, it seems that localization similarity has an irreplaceable role in proper scoring and loss of all information about localization will lead to inappropriate scoring results.

In the last experiment, we use all PPIs in our consolidated dataset which have missing values from at least one of their information sources. This experiment

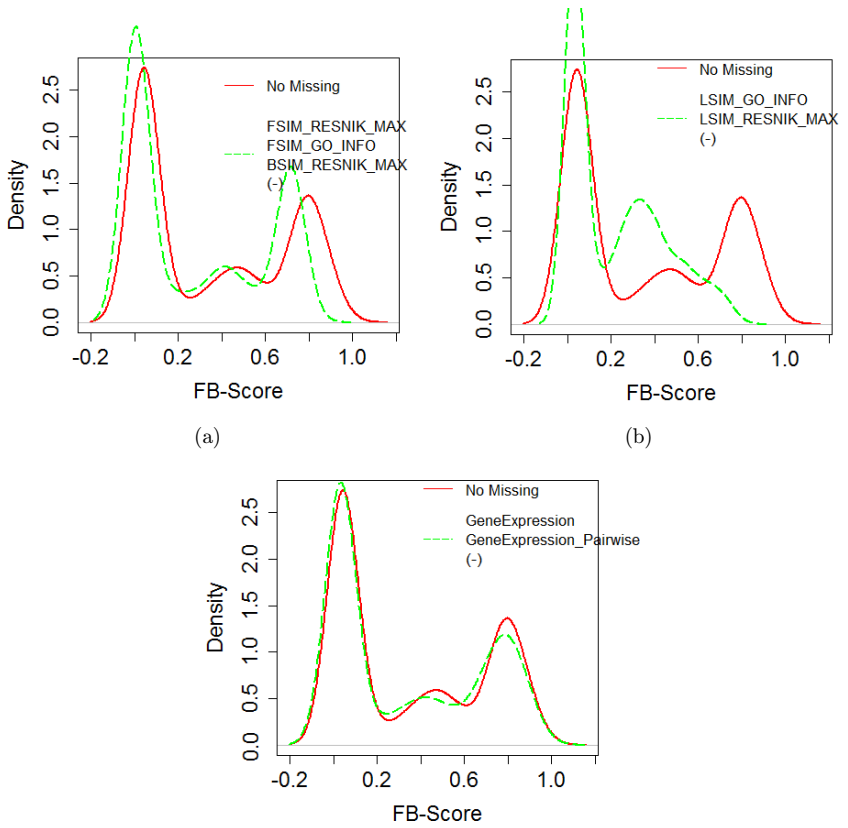


Fig. 8. Changes in distribution of the assigned FB-Scores by removing related information sources of (a) functional similarity (b) localization similarity, and (c) gene expression correlation.

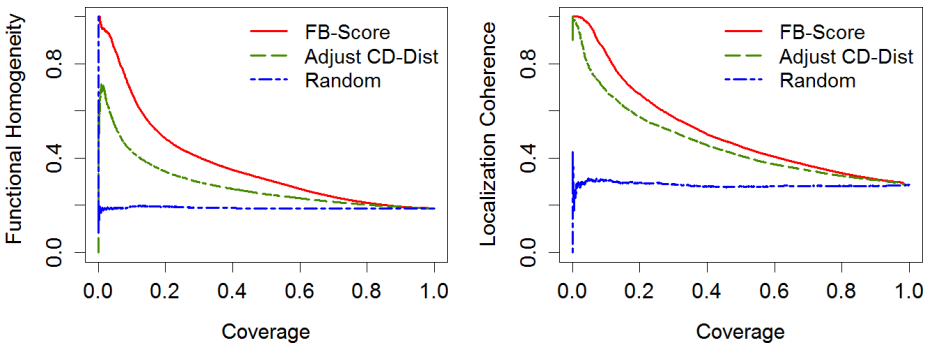


Fig. 9. Performance of FB-Score in comparsion to Adjust CD-Distance score. We sort descendingly all the PPIs with null values in their information sources, first by their FB-Score and then by their Adjust CD-Distance score. We also plotted functional homogeneity and localization coherence in case of random sorting of PPIs as the reference (a) comparison based on rate of functional homogeneity and (b) comparison based on rate of localization coherence.

evaluates our method in handling missing values from arbitrary but not artificial combination of information sources as we see in real PPI data. This is the way it differs from the experiment in Sec. 3.2 in which FB-Score is evaluated on data with no missing values. The results are shown in Fig. 9. In this experiment, 1000 PPIs from proteins with high PR-Score which have no missing values were used for finding fuzzy measures. The results again show superiority of FB-Score over Adjust CD-Distance both in terms of functional homogeneity and localization coherence in the presence of missing values. This shows that our proposed mechanism of handling missing values based on importance and interaction of information sources is working effectively for FB-scores.

4. Conclusion

PPIs obtained by high-throughput techniques were known to have dubious reliability. As such, it is important to enhance the reliability of the PPI data with additional data sources based on information such as the genetic properties of proteins, or the topological characteristics of their interaction networks. In this article, we study how to apply a decision making view for assigning reliability scores to PPIs by using a systematic combination of these evidences as our information sources in a data fusion framework.

There are quite a number of existing data fusion techniques, each with a different strategy in combining data. There are three particularly important problems in our context. First, we need to address the uncertainty of the values provided by different information sources. The second issue is the possible underlying interactions amongst the various information sources, in terms of wither redundancy or synergy. The third is handling missing values from some information sources. We have chosen the fuzzy

integral as our data fusion technique. It allows us to deal with the inherent PPI data uncertainty by the notion of fuzzy measures. Using fuzzy measures further allows us to detect the underlying redundant or synergistic interactions between information sources used for judging the reliability of the PPIs. We can also determine the amount of contribution of each information source toward estimating reliability estimation of a PPI by computing an importance index. Last but not least, our data fusion framework also facilitates effective handling of missing values through exploiting the interaction and importance of various information sources.

In summary, this study shows that we can use data fusion techniques to deal with the various problems of fusing multiple uncertain information sources to assign reliable reliability scores to PPI data. Our experiments show that our proposed method outperformed the reliability scoring provided by a state-of-the art technique, Adjust CD-Distance. It will be useful to extend our experiments with many more other information sources like domain information of proteins, sequence similarity, interologous information and so on, to see how far it can improve the scoring results. Other than scoring the reliability of PPI data, we also believe that our proposed data fusion approach can be used for many other bioinformatics problems which often require the integration of multiple uncertain and incomplete data sources.

References

1. Von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P, Comparative assessment of large-scale datasets of protein–protein interactions, *Nature* **417**:399–403, 2002.
2. Sprinzak E, Sattath S, Margalit H, How reliable are experimental protein–protein interaction data? *J Mole Biol* **327**:919–923, 2003.
3. Huang H, Jedynak BM, Bader JS, Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps, *PLoS Comput Biol* **3**:2155–2174, 2007.
4. Bader JS, Chaudhuri A, Rothberg JM, Chant J, Gaining confidence in high-throughput protein interaction networks, *Nat Biotechnol* **22**:78–85, 2004.
5. Jansen R, Yu H, Greenbaum D, Kluger Y, J Krogan N, Chung S, Emili A, Snyder M, F Greenblatt J, Gerstein M, A Bayesian networks approach for predicting protein–protein interactions from genomic data, *Science* **302**:449–453, 2003.
6. Lin N, Wu B, Jansen R, Gerstein M, Zhao H, Information assessment on predicting protein–protein interactions, *BMC Bioinform* **5**:154, 2004.
7. Myers CL, Troyanskaya OG, Context-sensitive data integration and prediction of biological networks, *Bioinformatics* **23**:2322–2330, 2007.
8. Ramani AK, Bunescu RC, Mooney RJ, Consolidating the set of known human protein–protein interactions in preparation for large-scale mapping of the human interactome, *Genome Biol* **6**:R40, 2005.
9. Patil A, Nakamura H, Filtering high-throughput protein–protein interaction data using a combination of genomic features, *BMC Bioinform* **6**:100, 2005.
10. Wu M, Li X, Chua HN, Kwoh CK, Ng SK, Integrating diverse biological and computational sources for reliable protein–protein interactions, *BMC Bioinform* **11**:S8, 2010.
11. Saito R, Suzuki H, Hayashizaki Y, Interaction generality, a measurement to assess the reliability of a protein–protein interaction, *Nucleic Acids Res* **30**:1163–1168, 2002.

12. Chen J, Hsu W, Ng SK, Increasing confidence of protein interactomes using network topological metrics, *Bioinformatics* **22**:1998–2004, 2006.
13. Chen J, Chua HN, Hsu W, Lee ML, Ng SK, Saito R, Sung WK, Wong L, Increasing confidence of protein–protein interactomes, *Genome Inform* **17**:284–297, 2006.
14. Chua HN, Wong L, Increasing the reliability of protein interactomes, *Drug Discovery Today* **13**:652–658, 2008.
15. Liu G, Li J, Wong L, Assessing and predicting protein interactions using both local and global network topological metrics, *Genome Inform* **21**:138–149, 2008.
16. Choquet G, Theory of capacities, *Annales de l'Institut Fourier* **5**:131–295, 1953.
17. Liu G, Wong L, Chua HN, Complex discovery from weighted PPI networks, *Bioinformatics* **25**:1891–1897, 2009.
18. Dasarathy BV, Information fusion — what, where, why, when, and how?, *Inform Fusion* **2**:75–76, 2001.
19. Bouchon-Meunier B (ed.), *Aggregation and Fusion of Imperfect Information* (*Studies in Fuzziness and Soft Computing*), Physica Verlag, 2001.
20. Torra V, Narukawa Y, *Modeling Decisions: Information Fusion and Aggregation Operators* (*Cognitive Technologies*), Springer-Verlag, Secaucus, NJ, USA, 2006.
21. Yager RR, On ordered weighted averaging aggregation operators in multicriteria decision making, *IEEE Trans Syst, Man Cybern* **18**:183–190, 1988.
22. Sugeno M, Theory of fuzzy integrals and its applications, PhD Thesis, Tokyo Institute of Technology, 1974.
23. Shafer G, *A Mathematical Theory of Evidence*, Princeton University Press, 1976.
24. Grabisch M, The application of fuzzy integrals in multicriteria decision making, *European J Oper Res* **89**:445–456, 1996.
25. Murofushi T, Soneda S, Techniques for reading fuzzy measures(iii): Interaction index, *Proc 9th Fuzzy Systems Symp*, Sapporo, Japan, pp. 693–696, 1993.
26. Shapley LS, Shubik M, A method for evaluating the distribution of power in a committee system, *Amer Political Sci Rev* **48**:787–792, 1954.
27. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Ursula H, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, The IntAct molecular interaction database in 2012, *Nucleic Acids Res* **40**:841–846, 2012.
28. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberga D, Dip: The database of interacting proteins, *Nucleic Acids Res* **28**:289–291, 2000.
29. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M, BioGRID: A general repository for interaction datasets, *Nucleic Acids Res* **34**(suppl 1):D535–D539, 2006.
30. Chatr-aryamontri A, Ceol A, Montecchi Palazzi L, Nardelli G, Schneider MV, Castagnoli L, Cesareni G, Mint: The molecular interaction database, *Nucleic Acids Res* **35**: D572–D574, 2007.
31. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Gene ontology: Tool for the unification of biology, *Nat Genetics* **25**(1):25–29, 2000.
32. Ng SK, Tan SH, Discovering protein–protein interactions, *J Bioinform Comput Biol* **1**(04):711–741, 2004.
33. Resnik P, Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, *J Artif Intell Res* **11**(1):95–130, 1999.
34. Pesquita C, Faria D, Falcao AO, Lord P, Couto FM, Semantic similarity in biomedical ontologies, *PLoS Comput Biol* **5**(7):e1000443, 2009.

35. Zhou H, Wong L, Comparative analysis and assessment of m. tuberculosis H37Rv protein–protein interaction datasets, *BMC Genomics* **12**(Suppl 3):S20, 2011.
36. Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, Troyanskaya OG, Exploring the functional landscape of gene expression: Directed search of large microarray compendia, *Bioinformatics* **23**:2692–2699, 2007.
37. Brun C, Chevenet F, Martin D, Wojcik J, Guénoche A, Jacq B, Functional classification of proteins for the prediction of cellular function from a protein–protein interaction network, *Genome Biol* **5**(1):R6, 2004.
38. Grabisch M, Kojadinovic I, Meyer P, A review of methods for capacity identification in choquet integral based multi-attribute utility theory: Applications of the kappalab r package, *European J Oper Res* **186**:766–785, 2008.
39. Brown KR, Jurisica I, Online predicted human interaction database, *Bioinformatics* **21**:2076–2082, 2005.

Alireza Vazifedoost is a Ph.D. candidate in computer engineering at the School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran. He was also a research intern at the Institute for Infocomm Research (I2R), Singapore's Agency of Science, Technology and Research (A*STAR) and before that he had another internship with the School of Computer at National University of Singapore. He received his Master degree from the University of Tehran in 2007 and his Bachelor degree in 2003 from Iran University of Science and Technology both in software engineering. His research fields are computational biology, data mining and data fusion.

Maseud Rahgozar received his B.Sc. degree in Electrical and Electronic Engineering from Iran Sharif University of Technology in 1980 and M.Sc. and Ph.D. in computer science from Paris VI University, France in 1983 and 1987 respectively. He joined the School of Electrical and Computer Engineering, University of Tehran in 2000 where he is currently associate professor of software engineering. His fields of research include XML Databases Implementation and Optimization, Data Mining and Data warehousing, Systems Information Storage and Retrieval, Normalization of Database Systems (reverse engineering, and optimization), Human Computer Interaction, Designing Case Tools (software generation, integration, and modernization), Modernization of Legacy Systems (automated migration to modern environments).

Behzad Moshiri is currently a professor at the School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran. He has been the head of Machine Intelligence and Robotics division of this school. He received his M.Sc. and Ph.D. from UMIST, UK, in 1987 and 1991, respectively. He was the member of ISA (Canada Branch) in 1991–1992. He was the president of Iranian society of instrumentation and Control Engineers (ISICE) in 1999–2001. He has been the member of ISIF since 2002 and senior member of IEEE since 2006. He was the head of Machine Intelligence and Robotics group at the school of ECE (for two periods between 1999–2007). He also served as president of Intelligent Systems of Scientific Society of Iran (ISSSI) in 2011–2013. His research interests include advanced industrial control design, advanced instrumentation design, sensor data fusion, intelligent transportation systems, mechatronics and bioinformatics.

Mehdi Sadeghi received his B.Sc. in 1991 in Cell and Molecular Biology, M.Sc. in 1993 and Ph.D. in 2001 in Biophysics from the University of Tehran. He is presently an Associate Professor at the National Institute of Genetic Engineering and Biotechnology (Iran) and senior researcher at the School of Biological Sciences, Institute for research in fundamental sciences (Iran). His research interests are bioinformatics, especially protein structure prediction, classification and also systems biology.

Chua Hon Nian is the Deputy Head of Machine Learning at the Data Analytics Department of the Institute for Infocomm Research (I2R), Singapore's Agency of Science, Technology and Research (A*STAR). He obtained his Computer Engineering degree and Ph.D. from the National University of Singapore. He also underwent postdoctoral training at the Harvard Medical School and the University of Toronto, where he worked on applications of machine learning in biology and biotechnology.

See Kiong Ng (Ph.D., Carnegie Mellon University) is the Programme Director of the Urban Systems Initiative by the Science and Engineering Research Council of the Agency of Science, Technology and Research (A*STAR). The Initiative seeks to address the new challenges of the rapidly urbanising world through smart city technology and innovations. See-Kiong has a long-standing interest in cross-disciplinary applied research in computer science. He wrote the TrueAllele software when he was a graduate student at CMU that was subsequently used by various biotech companies for high throughput genotyping. See-Kiong has since been able to apply what he had learned from bioinformatics to a wide array of other application domains. From using data mining to better understand the biology of the human body, See-Kiong is now using big data approaches to understand the systems biology of complex human cities. He has published widely, with more than 100 papers in leading peer-reviewed journals and conferences across multiple disciplines.

Limsoon Wong is a KITHCT chair professor of computer science and professor of pathology at the National University of Singapore. He currently works mostly on knowledge discovery technologies and their application to biomedicine. He is a Fellow of the ACM, named in 2013 for his contributions to database theory and computational biology.