Handling Noise in Protein Interaction Networks

Fernanda B. Correia[1*, 2], Edgar D. Coelho[1], José L. Oliveira[1] and Joel P. Arrais[3]

[1] Department of Electronics, Telecommunications and Informatics, Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, 3810-193 Aveiro, Portugal.

[2] Department of Informatics and Systems Engineering, Coimbra Institute of Engineering, Polytechnic Institute of Coimbra, 3030-199 Coimbra, Portugal.

[3] Department of Informatics Engineering, Centre for Informatics and Systems of the University of Coimbra, University of Coimbra, 3004-531 Coimbra, Portugal.

Correspondence should be addressed to Fernanda B. Correia; fernanda.correia@ua.pt

Abstract

Protein-protein interactions (PPI) can be conveniently represented as networks, allowing the use of graph theory in their study. Network topology studies may reveal patterns associated to specific organisms. Here we propose a new methodology to denoise PPI networks and predict missing links solely based on the network topology, the Organization Measurement (OM) method. The OM methodology was applied in the denoising of the PPI networks of two *Saccharomyces Cerevisiae* datasets (Yeast and CS2007) and one *Homo Sapiens* dataset (Human). To evaluate the denoising capabilities of OM methodology, two strategies were applied. The first compared its application in random networks and in the reference set networks, while the second perturbed the networks with the gradual random addition and removal of edges. The application of OM methodology to the Yeast and Human reference sets achieved an AUC of 0.95 and 0.87, in Yeast and Human networks, respectively. The random removal of 80% of the Yeast and Human reference sets interactions resulted in an AUC of 0.71 and 0.62, whereas the random addition of 80% interactions resulted in an AUC of 0.75 and 0.72, respectively. Applying the OM methodology to the CS2007 dataset yields an AUC of 0.99. We also perturbed the network of the CS2007 dataset by randomly inserting and removing edges in the same proportions previously described. The false positives identified and removed from the network varied from 97%, when inserting 20% more edges, to 89% when 80% more edges were inserted. The true positives identified and inserted in the network varied from 95% when removing 20% of the edges, to 40% after the random deletion 80% edges. The OM methodology is sensitive to the topological structure of the biological networks.

The obtained results suggest that the present approach can efficiently be used to denoise PPI networks.

Introduction

Proteins are central players in every organism, as they are required for virtually every single cellular function. However, proteins are required to interact with one another to fulfil their functions. For this reason, disease states may appear, if the physiological interaction between two proteins is disrupted [1].

Protein-protein interactions (PPI) networks are a subset of complex biological networks that have specific topological properties, such as a high clustering coefficient, the presence of hierarchy, heterogeneity and a power-law-like degree distribution [2]. The guilt-by-association hypothesis states that two proteins sharing many interactive neighbours are likely to hold functional homogeneity and localization coherence [3]. These characteristics suggest that network topology alone may be a viable option for PPI network denoising. PPI network denoising corresponds to find interactions that do not exist and to find missing interactions. Methods to determine protein interactions are not accurate and organisms are not yet fully known, being important to denoise PPI networks to have more precise models of the organisms.

Protein interactions can be represented as graphs, allowing the use of graph theory in their study. As such, different methods were developed to denoise biological networks [reviewed in 4]. These include repeating experiments [5, 6], using prior knowledge about proteins [7, 8], using functional or structural annotations [9-13] and using comparisons with theoretical distributions constructed from known data and network topology-based approaches [14-18]. The herein proposed approach falls under the latter category.

An approach, called Non-Convex Semantic Embedding (NCSE) evaluates the reliability of interactions in a PPI network trying to learn a Euclidean embedding under the geometric assumption of PPI networks [19] and it was tested in three data sets of *Yeast Saccharomyces Cerevisiae*. Computational methods based on machine learning were also applied to evaluate the reliability of PPI and tested in *Yeast* and *Helicobacter Pylori* PPI datasets [20, 21].

In a recent study, Lü et al. [22] proposed the structural consistency index and the Structural Perturbation Method (SPM). On the one hand, the structural consistency index can reflect the inherent link predictability of a network without knowing its organization a priori, allowing to estimate the explicability of the organization of a network, and to supervise mechanistic changes during the evolution of the network. On the other hand, the SPM performs link

prediction by removing a percentage of the edges in a network, thus perturbing the remaining network by that percentage. This is based on the strong correlation between independent network perturbations, which suggests that the missing edges, i.e., False Negative (FN) interactions can be identified by perturbing the networks with an additional set of known interactions, i.e., True Positive (TP) interactions.

Luo et al. [4] proposed the Collaborative Filtering-enhanced Topology-based (CFT) method to perform protein interatomic mapping on sparse High-Throughput Screening (HTS) PPI data, since the performance of network topology-based approach usually deteriorates when using sparse network data. This approach is based on the notion that the solution space of the interatomic mapping and the solution space of the personalized recommendations are similar. Each protein is represented as a feature vector that describes their interactions in the network. In addition, the feature vector is used to calculate the corresponding similarity vector that represents the interactions through the Functional Similarity weight, creating an Inter-Neighbourhood Similarity (I-Sim) for modelling PPIs. Functional parameters for each protein in the dataset are obtained from Gene Ontology (GO), allowing the use of functional similarity measures. Denoising of the input HTS-PPI data is performed via the integration of saturation-based strategies into the I-Sim, achieving a precise relationship model. Their method was applied to three different datasets and compared with three other algorithms (Interaction Generality [14], Czekanowski-Dice distance [15], and Functional Similarity weight [16]), showing better performance on large, sparse HTS-PPI datasets. Since they use GO annotations to characterize their proteins, this approach is likely to underperform when considering less studied organisms.

A different strategy termed Intrinsic Geometry Structure (IGS) was proposed by Yi Fang et al. [23]. IGS is a geometry-based approach which uses heat diffusion in the PPI network to collect structural information about all paths connecting two given nodes, thus defining intrinsic relationships among them. They use a maximum likelihood-based algorithm to determine the optimal dissipation time, predicting the global structure of the PPI network from the local structure. After performing heat diffusion for the optimal dissipation time, the intrinsic geometric structure of the PPI network is revealed. One of the main advantages of the IGS method is its robustness against missing protein associations and sparse PPI data. Their method was tested with the *S. cerevisiae* (CS2007) network [24], a network of the bottlenose dolphin community [25], and a network of known terrorist cells [26]. In addition, they compared the performance of IGS with two other methods, the Multi-Dimensional Scaling-based (MDS) method [27] and the Hierarchical Random Graph (HRG) method [28], showing that IGS

performed slightly better than MDS when tested with the CS2007 dataset and better than IGS and HRG for the other datasets tested. Their analysis was based only in the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) values.

Considering the limitations of the described methods, we introduced the Organization Measurement (OM) method to denoise PPI networks based exclusively in the network topology. Topological measures are used to find trends that characterize interacting and non-interacting proteins distributions. A high confidence set of protein interactions is used to construct a network, followed by the calculation of the weights of interactions and non-interactions in the network. The OM weighted matrix is obtained and used to find distribution trends that allow to distinguish interaction distributions from non-interaction distributions. The OM threshold value that better distinguishes these types of distributions is then used to identify False Positive (FP) interactions and FN (novel) interactions. This way, an OM topological model is built to be used in the denoising of a network, resulting in a better approximation of the expected network.

Materials and Methods

This section will describe how we obtained the datasets used in the experiments and the topological measures used with the Organization Measurement (OM) methodology, including the new Neighbourhood Clustering (NC) proposed measure. It will also describe the OM methodology, how to obtain the OM matrix of weights and how to determine the threshold value, giving a description of the OM methodology pipeline (Fig 1) to denoise networks.

For each organism, we collected a set of high-confidence PPI interactions. Although these PPI do not reflect the entirety of the protein interaction networks of the selected organisms, they were used to construct the known PPI network of each organism, *i.e.*, their reference sets.

In the application of the OM methodology, various topological measures were calculated to characterize these networks, based on the assumption that these measures will allow the identification of topologic patterns to support network denoising. In this paper, we use the term "denoising" to define the identification of FP and FN interactions, removing the former cases from the network while adding the latter. The methodology proposed here can also be used to rank the level of confidence of the interactions already presented in the network. Different

topological measures can identify different patterns and thus, here we consider that different topological measures can contribute to the denoising process.

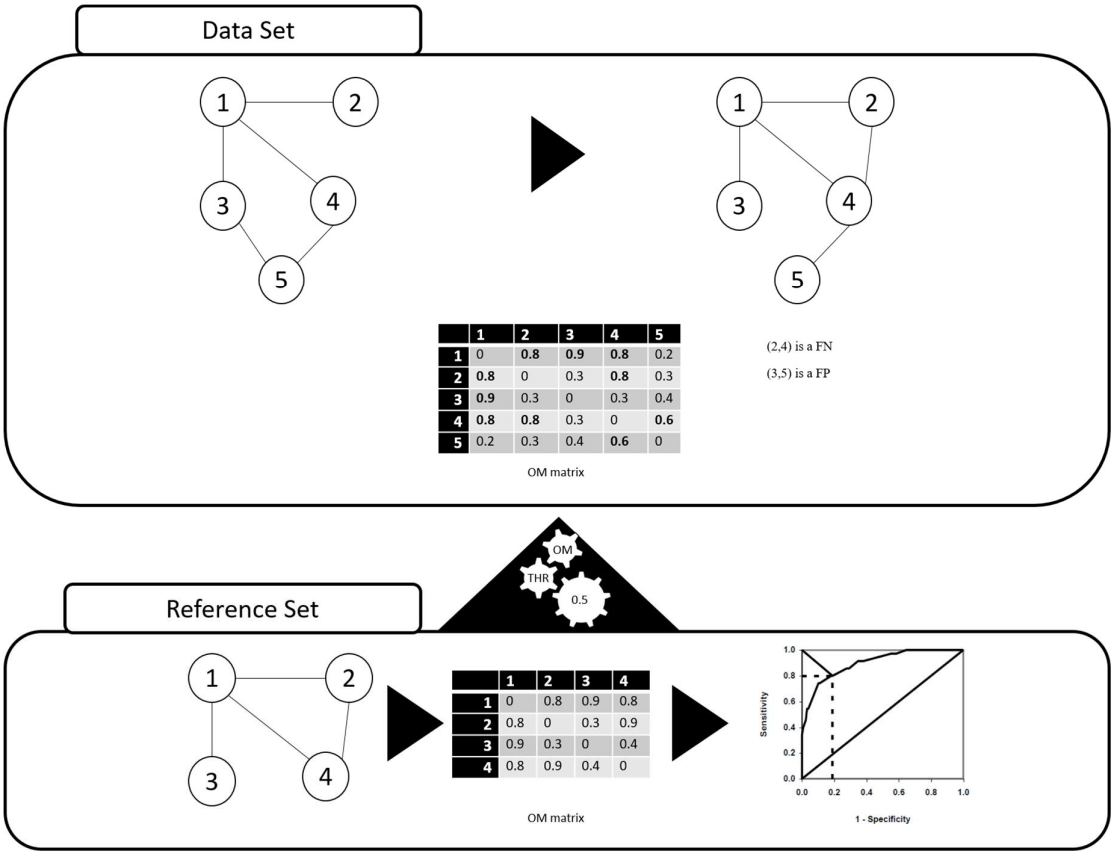


Figure 1: Diagram of the OM methodology pipeline. A Reference Set of an organism is used to create a network model and the respective OM matrix is obtained, using the topological properties. The threshold (OM THR) is calculated using the ROC curve and choosing the value that best separates interactions from non-interactions distributions. OM THR is then applied to denoise a lower-confidence Data Set network of the same organism, using the respective OM matrix.

Next sections will describe in a detailed way the OM methodology pipeline, after the description of the datasets chosen to illustrate and validate the proposed methodology and the description of the topological measures used to test OM methodology, including the new one topological measure proposed. A section will explain the process to obtain the weighted OM matrix from the adjacency matrix, another section will describe how to determine the OM THR value using the ROC curve and a last section will describe how to use the threshold in the denoising of PPI networks.

142 **Datasets**

143 The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [29]
144 contains known and predicted protein interactions of various organisms. PPI in STRING derive
145 from five main sources: (1) genomic context predictions; (2) high-throughput experimental
146 methods; (3) conserved co-expression experiments; (4) automated text mining, and; (5)
147 previous knowledge from third-party databases. Each interaction in STRING has an associated
148 score for each prediction method and has a Combined Score (CS) that range from 0 to 1000,
149 indicating the degree of confidence of each interaction. Calculation of the CS considers several
150 parameters, such as the number and the quality of different sources indicating that a PPI occurs.

151 The interactions derived by experimental methods with a score greater than 900 have been
152 considered of high-confidence in multiple works [30, 31]. Therefore, the reference sets used in
153 this work comprise experimentally determined PPI data obtained from STRING with a score
154 greater than or equal to 900.

155 These data were collected from two different organisms, namely the *Yeast Saccharomyces*
156 *Cerevisiae* (Yeast) and *Homo Sapiens* (Human) [29]. Using these data, an undirected network
157 is constructed for each organism and the main component is extracted.

158 Table **1** summarizes the characteristics of the reference set networks obtained for Yeast and
159 Human, including the number of nodes, the number of edges, the average degree and network
160 density. The observed average degree and density values are highly suggestive that these are
161 biological networks are sparse, *i.e.*, they have much less edges than the complete network with
162 the same set of nodes. Our high-confidence networks (*i.e.*, PPIs obtained from the STRING
163 database with experimental source score greater than 900) comprised 29,319 interactions
164 between 3,937 proteins for the Yeast dataset and 16,931 interactions between 4,943 proteins
165 for the Human dataset.

166 Additionally, we used a high confidence external dataset compiled by Collins *et al.* [24] and
167 referred to as CS2007 hereafter, to compare the proposed methodology with other topology-
168 based denoising methods [23, 27, 32]. This dataset comprises 9,074 PPIs between 1,622 unique
169 proteins from *Yeast Saccharomyces Cerevisiae*. To ensure a direct comparison between the
170 OM methodology and the existing methods we followed their approaches and only used the
171 largest connected component. The largest connected component of the dataset compiled by
172 Collins *et al.* [24] includes 8,323 interactions between 1,004 proteins (Table 1).

173

Table 1. Topological characteristics of the Yeast, CS2007 and Human networks used as reference sets.

Organism	Name	From	Nº of nodes	Nº of edges	Average degree	Density
Yeast Saccharomyces Cerevisiae	Yeast	STRING experimental CS >= 900	3,937	29,319	14.8941	0.0038
	CS2007	compiled by Collins et al.	1,004	8,323	16.5797	0.0165
Homo Sapiens	Human	STRING experimental CS >= 900	4,943	16,931	6.8505	0.0014

Measures for similarity and diversity analysis of network data

Protein interactions can be conveniently modelled as a network, where each node represents a protein and each edge represents an association between two proteins. The most commonly used technique to quantify the interaction profile similarity of protein interaction networks (or any type of biological network) relies on association indices. Bass *et al*. [33] performed a comprehensive review on the selection of association indices for the analysis of gene similarity. In their work, the Jaccard (JC), Geometric and Cosine indices were shown to be the most versatile, as though not excelling in any particular task, their strengths were the most balanced out of all evaluated measures. A review of similarity indices can also be found in [34]. Daminelli *et al*. test the application of different association indexes to bipartite networks [35].

A more recent study reports that the JC measure performs better than three other measures in a specific model [36].

The JC measure is defined as the ratio of the intersection of the number of neighbours of nodes i and j divided by their union (*i.e.*, the ratio of nodes shared between i and j divided by the total number of nodes connected to both):

$$JC_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|} \quad (1)$$

where $\Gamma(i)$ is the set of neighbours of *i.* We also explored and tested additional measures and two of them that gave good results were Betweenness (BETW), and Katz indexes.

195 The implementation of Betweeness (BETW) used was:

$$BETW_{ij} = (BETW_i + BETW_j)/2 \quad (2)$$

196 where

$$BETW_i = \sum_{l,m \in V} \frac{nsp(l,m|i)}{nsp(l,m)} \quad (3)$$

197 where V is the set of nodes, $nsp(l,m)$ is the number of (l,m) shortest paths and $nsp(l,m|i)$
198 is the number of those paths passing through the node i.

199 The implementation of Katz used was:

$$KATZ_{ij} = \frac{(KATZ_i + KATZ_j)}{2} \quad (4)$$

$$KATZ_i = \alpha \sum_l Adj_{il} x_l + \beta \quad (5)$$

200 with Adj the adjacency matrix of the network with eigenvalues λ. It was used $\alpha = 1/\lambda_{max}$
201 and $\beta = 0$, when Katz centrality is the same as the eigenvector centrality.

202 Based on the idea that closely associated proteins are more likely to interact, that the network
203 modularity is associated with the Clustering Coefficient (CC) [37] and a high mean CC of a
204 community can be used to identify those that are functionally homogeneous [38], we
205 implemented a novel measure to emphasize the relevance of the CC concept associated to the
206 neighbourhood concept in a network. This measure was called Neighbourhood Clustering (NC)
207 measure and is defined as the ratio of the sum of the CC of the nodes shared between *i* and *j*
208 divided by the sum of the CC of the total number of nodes connected to both *i* and *j*:

$$NC_{ij} = \frac{\sum CC\ (\Gamma(i) \cap \Gamma(j))}{\sum CC\ (\Gamma(i) \cup \Gamma(j))} \quad (6)$$

209 where $\Gamma(i)$ is the set of neighbours of *i*.

210 **Organization measurement matrix**

211 In Fig 1 we summarize the pipeline of the proposed OM methodology. Once the PPI network
212 (Reference Set) for the or'ganism is constructed, its respective adjacency matrix is built,
213 followed by its transformation into a weighed matrix, the OM matrix. The OM matrix is used
214 to find distribution trends that allow to distinguish interactions and non-interactions. The

weights for interactions and non-interactions are calculated using topological measures and using the information about the interactions of the network.

The adjacency matrix of the PPI network A, with N proteins and M interactions is defined as $adj_A = [a(i,j)]$, where $a(i,j) = 1$, if there is an interaction in A between nodes i and j. Otherwise, $a(i,j) = 0$. A topological measure is applied to A to determine a weight for each (i,j) to transform the adjacency matrix A into a transformed matrix $A_w = [a_w(i,j)]$, where $a_w(i,j)$ is the weight of (i,j) in A, calculated using the topological properties of the network.

The weight $a_w(i,j)$ represents the strength value of the edge (i,j) per the topological measure used and aims to capture patterns associated to the network that can originate signatures that identify the PPI network of each organism. This weight was used to characterize interaction and non-interaction distributions of the PPI network to determine the separation border between them.

Organization measurement threshold value determination

One of the assumptions made in this work is that the PPIs in the reference datasets are true. This assumption can be made due to the sparsity of protein interaction networks and the rigorous criteria chosen to filter TP interactions. However, the same cannot be said for the non-interactions, as the presence of FN PPI is highly likely.

The value that best distinguishes both interactions and non-interactions distributions was called the OM threshold value. First, we collect protein interactions data of a specific organism and then a network is built (Fig 1). Then, the respective adjacency matrix is constructed, followed by its transformation into a weighted matrix, the OM matrix, using the topological measures of interest. Finally, the ROC curve is calculated and used to determine the optimal cut-off, corresponding to the threshold value that separates the interaction distributions from non-interaction distributions. We considered as the optimal cut, the point closest to (0,1) in the ROC curve, where sensitivity equals specificity. Different topological measures were tested and the respective cut-off values were determined. The outcomes of these experiments are described in the "Results and discussion" section.

Organization measurement methodology to find spurious and new interactions

An accepted assumption in network topology-based approaches is that interacting proteins in a local community and closer to one another in the network are most likely involved in similar functions, or part of the same pathways [39-41]. The use of topological measures that capture this information should be prioritized, as they are expected to better grasp patterns in

incomplete networks, thus allowing the approximation of incomplete input networks to the real networks.

From each reference set (Reference Set) PPI network we calculated its adjacency matrix. Then, after calculating the respective weights, the adjacency matrix is transformed into a weight matrix. Finally, the threshold that best separates PPI and non-PPIs was determined through the finding of the optimal cut-off of the ROC curve. This threshold was applied to detect spurious and missing PPI in the network (Data Set), to obtain a better approximation of the true network. In the example network shown in Fig 1, there are five nodes representing five different proteins, in addition to six edges that could represent the interactions between them (Data Set). Assuming the example network approximates the current knowledge of a given biological network, not all true interactions are represented and the existence of FP is expected. Once the threshold value is calculated, using the reference set (Reference Set), it is applied to the OM matrix of the Data Set, to identify FP and FN interactions. FP interactions are then removed from the network, whereas FN interactions are added.

Results and Discussion

The OM methodology was tested with different topological measures and was evaluated using three different scenarios. Next subsections will describe the results obtained with the experiments made.

Analysis of different topological measures to identify the optimal threshold value

A key component of the proposed method is the determination of the threshold value to discriminate between protein interactions and non-protein interactions distributions. As such, we decided to test the OM methodology with different topological measures to determine which better discriminates PPI from non-PPI. The four topological measures used were the JC, the BETW, the KATZ and the proposed new measure, the NC. Fig 2 shows the ROC curves when using this methodology with four different topological measures for the Yeast and Human organisms and Fig 3 shows the same information, but after data normalization between 0 and 1. Table 2 shows the respective AUC values obtained. Best results were achieved when using OM methodology with the JC and NC measures.

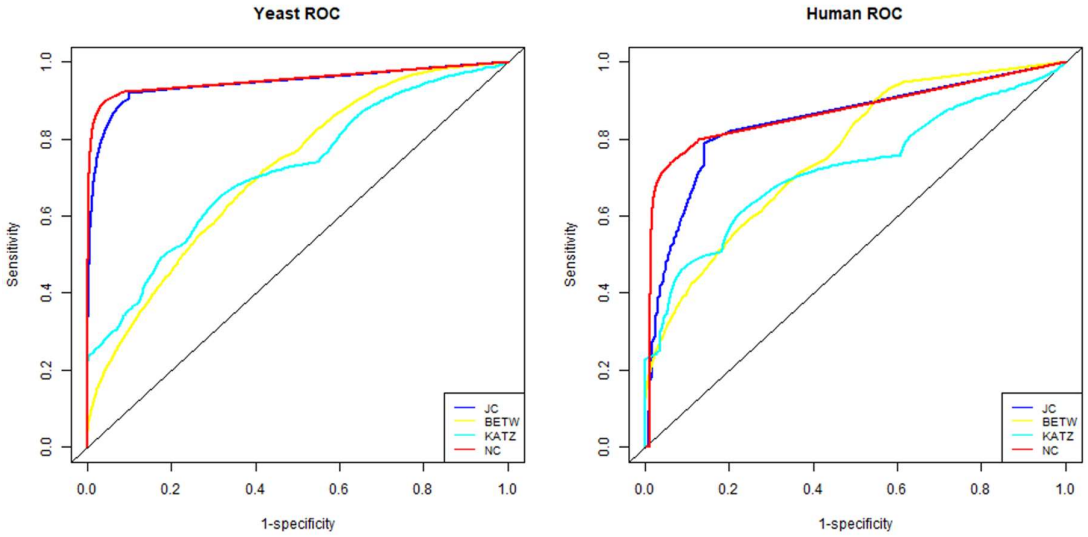


Figure 2. OM methodology ROC curves. ROC curves obtained by OM application with JC, BETW, KATZ and NC measures in Yeast and Human datasets.

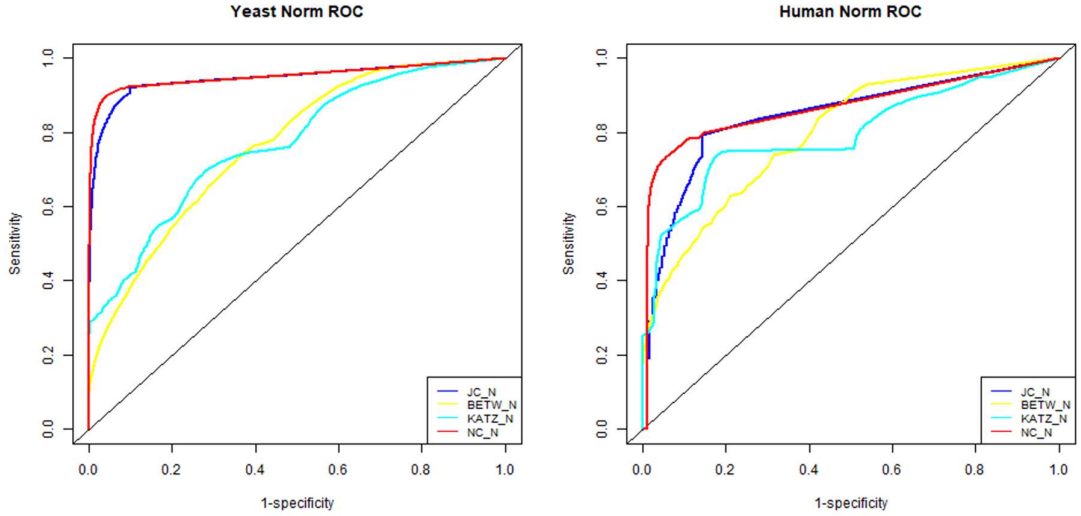


Figure 3. OM methodology ROC curves with normalized data. ROC curves obtained by OM application with JC, BETW, KATZ and NC measures in Yeast and Human datasets after data normalization between 0 and 1.

Table 2: AUC in Yeast and Human datasets.

Topological measures	Yeast AUC		Human AUC	
	Not normalized	Normalized	Not normalized	Normalized
JC	0.9462	0.9460	0.8438	0.8458
BETW	0.7142	0.7676	0.7659	0.8004
KATZ	0.7151	0.7714	0.7258	0.7944
NC	0.9534	0.9526	0.8708	0.8700

In addition to testing the OM methodology with these measures, we calculated the cut-off values for both the optimal cut and the accuracy cut, using the JC and NC measures, those that previously gave better results. The optimal cut calculates the point closest to (0,1) in the ROC curve, where sensitivity equals specificity, whereas the accuracy cut calculates the maximum accuracy and the respective cut-off value. In the next experiments, to determine OM THR, we calculated the cut-off values for the optimal cut, obtaining good results.

Table 3 and 4 show the obtained results for the Yeast and Human organisms. We can observe that the NC measure gave a slightly better result than JC and so we will describe forward the experiments using the NC measure.

Table 3: AUC, optimal cut and accuracy cut values in Yeast.

Yeast	AUC	Optimal cut		Accuracy cut	
JC	0.9462	sensitivity	0.9246	accuracy	0.9123
		specificity	0.9000		
		cut-off	0.0008	cut-off	0.0008
NC	0.9534	sensitivity	0.9057	accuracy	0.9282
		specificity	0.9471		
		cut-off	0.0021	cut-off	0.0044

Table 4: AUC, optimal cut and accuracy cut values in Human.

Human	AUC	Optimal cut		Accuracy cut	
JC	0.8438	sensitivity	0.8069	accuracy	0.8300
		specificity	0.8531		
		cut-off	0.0005	cut-off	0.0005
NC	0.8708	sensitivity	0.7989	accuracy	0.8400
		specificity	0.8735		
		cut-off	0.0001	cut-off	0.0014

Evaluation of the organization measurement methodology in different scenarios

To assess whether the OM methodology is sensitive to the network topology, we applied it to a randomly generated protein network, with the same number of nodes and edges as their respective reference sets (for Yeast and Human). If the OM methodology can distinguish between interactions and non-interactions in the reference data sets but fails to do so in the random networks, one can assume that it captures the inherent topological structure of a real network.

To further evaluate the performance of OM methodology two other experiments were performed. First, while maintaining the same number of nodes (proteins), we randomly added incrementing percentages of edges (proteins interactions), 20%, 40%, 60% and 80%, not belonging to the reference set network, building four networks and removed the same percentages of edges from the reference set network, building more four networks. This was performed for the Yeast, Human and CS2007 reference set networks. After each addition or removal, we used the OM methodology to denoise the networks. To further assess the ability of the proposed methodology for network denoising, we also determined the percentage of inserted TN removed from the respective CS2007 perturbed networks and the percentage of TP retrieved from the respective CS2007 perturbed networks. A thorough description of the results of these experiments are shown in the following sub-sections.

Organization measurement methodology performance comparison: random network versus reference set network

The only criteria selected to generate the random networks was that the resulting randomize networks were required to comprise the same number of nodes and edges. Thus, we generated 10 networks for each organism to be tested using the NC measure.

Fig 4 shows the ROC curves, the separation of classes (PPI e non-PPI) curves and the accuracy curve, when applying the OM methodology with the NC topological measure to one of the Yeast (left column) and Human (right column) random networks generated with the same number of nodes and edges of the respective reference sets. Analysing their ROC curves, we can see a clear distinction in performance between the application of OM methodology to the random network (Fig 4) and the subsets of the real networks (Fig 2). The AUC obtained after using the proposed method in all 10 random networks generated was close to 0.5 for both organisms (Yeast and Human), while for the subset of the Yeast network and Human network the AUC was 0.9534 and 0.8708, respectively.

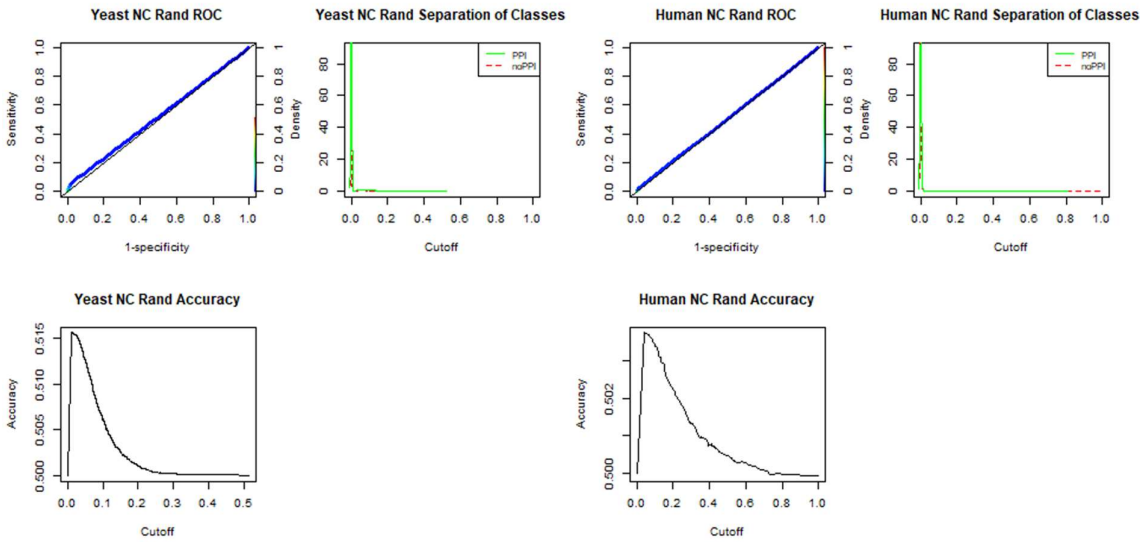


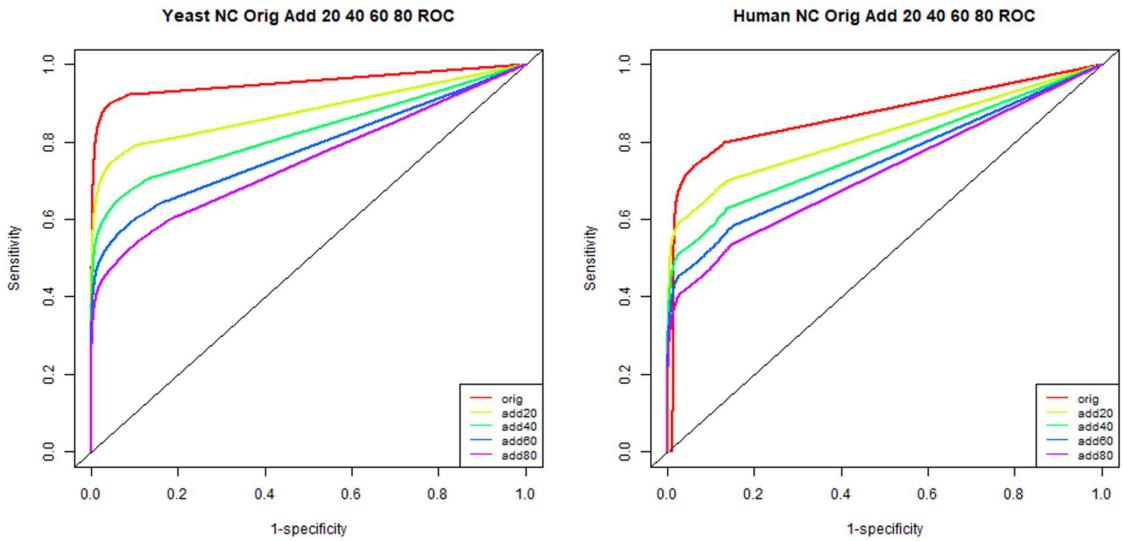
Figure 4: OM methodology application with the NC topological measure, to Yeast and Human random networks. OM methodology application ROC curves, separation of classes (PPI e non-PPI) curves and accuracy curve with the NC topological measure in one of the random networks generated with the same number of nodes and edges as the Yeast (left column) and as the Human (right column) reference sets.

342 **Random insertion of edges**

343 To evaluate the performance of the OM methodology for denoising PPI networks, we perturbed
344 the networks of the reference sets by randomly adding incrementing percentages of edges to
345 the networks of the Yeast and Human reference sets and the CS2007 reference set.

346 We created four noisy networks for each data set, adding 20% more edges to the original
347 network, followed by 40%, 60% and 80%. These intervals were selected following the research
348 conducted by Yi Fang *et al.* [23]. Fig 5 shows the ROC curves when the OM methodology is
349 applied to the networks of the Yeast and Human reference sets and to the four noisy networks
350 generated from each of them. We can observe a decreasing of performance when we increase
351 the percentage of the random edges added.

352



353

354 Figure 5: Application of the OM methodology with the NC topological measure, when an increasing percentage
355 of edges was added randomly to the Yeast and Human reference set networks. ROC curves of the reference sets,
356 and the other 4 networks, when 20%, 40%, 60% and 80% of edges were added to the reference set network for
357 Yeast (left column) and Human (right column).

358

359 To be able to compare the performance of OM methodology with other network-based
360 methodologies proposed by other researchers, we also perturbed the CS2007 network by
361 randomly inserting edges in the same proportions previously described. Fig 6 shows the
362 graphical representation of the resulting AUC values when the OM methodology is applied to
363 the four noisy networks obtained from the CS2007 network and when MDS and IGS

methodologies are applied [23]. It can be observed that the proposed OM methodology outperforms MDS and IGS methodologies.

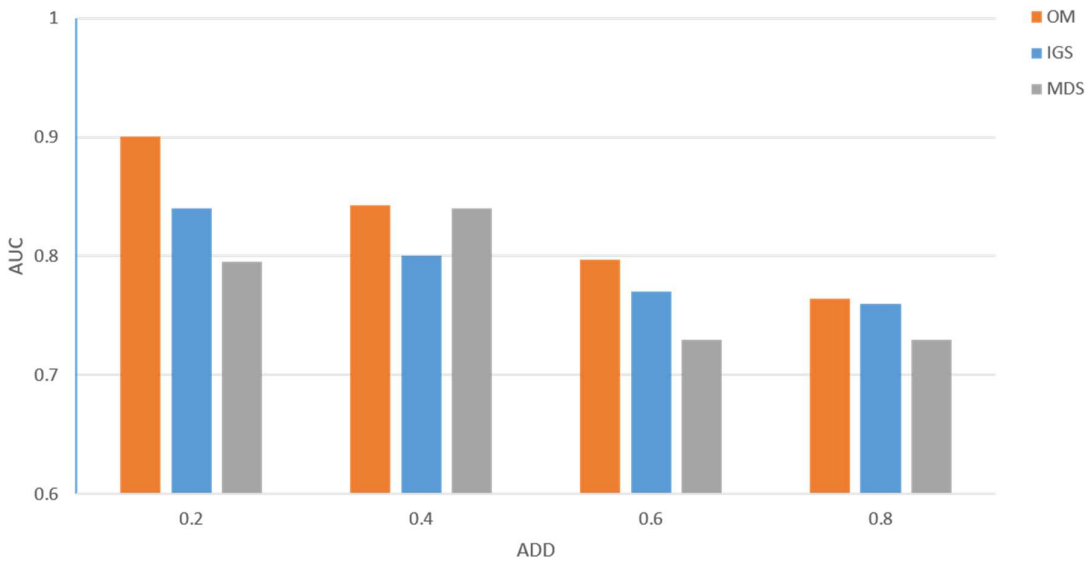


Figure 6: Application of the OM methodology with the NC topological measure, when an increasing percentage of edges was added randomly to the CS2007 reference set network compared to the MDS and IGS methods. AUCs values of the CS2007 perturbed 4 networks when a percentage of random 20%, 40%, 60% and 80% of edges were added to the reference set networks, using OM, IGS and MDS methods.

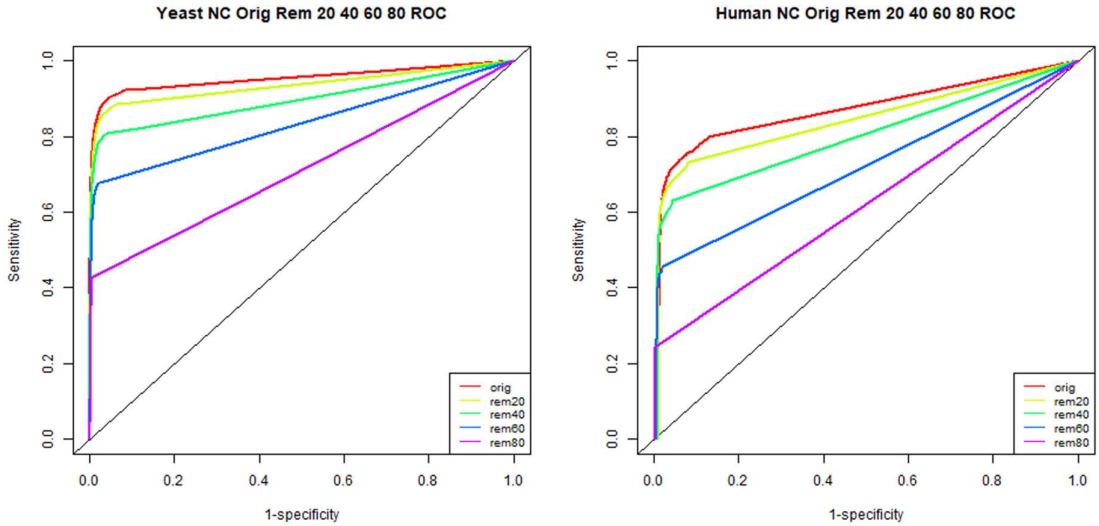
After denoising the networks with the OM methodology, we also calculated the percentage of FP interactions that were removed (Table 5). We observe that the OM methodology could remove 97% of the FP of the 20% added and 89% of the 80% added.

Table 5: Percentage of FP removed after applying the OM methodology to the noisy networks of the CS2007 dataset.

% added	# FP added	# FP removed	% FP removed
20	1,665	1,607	97
40	3,329	3,114	94
60	4,994	4,560	91
80	6,658	5,926	89

380 **Random deletion of edges**

381 To evaluate the performance of the OM methodology for the identification of missing
382 interactions, four new networks were created for each dataset (*i.e.*, Yeast, Human, and CS2007
383 reference sets) by removing increasing percentages of edges from the respective reference set
384 networks. Edge removal was performed in the same proportion as edge addition: 20%, 40%,
385 60% and 80%. The results are shown in Fig 7, representing the ROC curves, when the OM
386 methodology is applied to the eight noisy networks referred previously, of the Yeast and
387 Human reference set networks.



388

389 Figure 7: Application of the OM methodology with the NC topological measure, when an increasing percentage
390 of edges is removed randomly from the Yeast and Human reference set networks. ROC curves of the reference
391 set, and the other 8 networks when 20%, 40%, 60% and 80% of edges were removed from the reference set
392 networks for Yeast (left column) and Human (right column).

393

394 These results show a scenario alike the one observed after randomly adding edges, as greater
395 reductions in the number of edges result in greater performance drops, but the performance
396 drops are steeper in the Human organism.

397 Fig 8 shows a graphic of the AUCs values, when the OM methodology is applied to the four
398 noisy networks obtained from the CS2007 network, referred previously, compared to the MDS
399 and IGS methodologies [23]. OM methodology has a better performance compared to the IGS
400 and MDS methodologies, except when 80% of the interactions are removed from the CS2007
401 reference set, where the application of IGS gives better results. Further details are in Table 6,
402 where we can observe that 95% of the TP removed could be detected when the OM

methodology is applied to the perturbed network, when 20% of the interactions of the reference set were removed and 40% could be detected when 80% were removed.

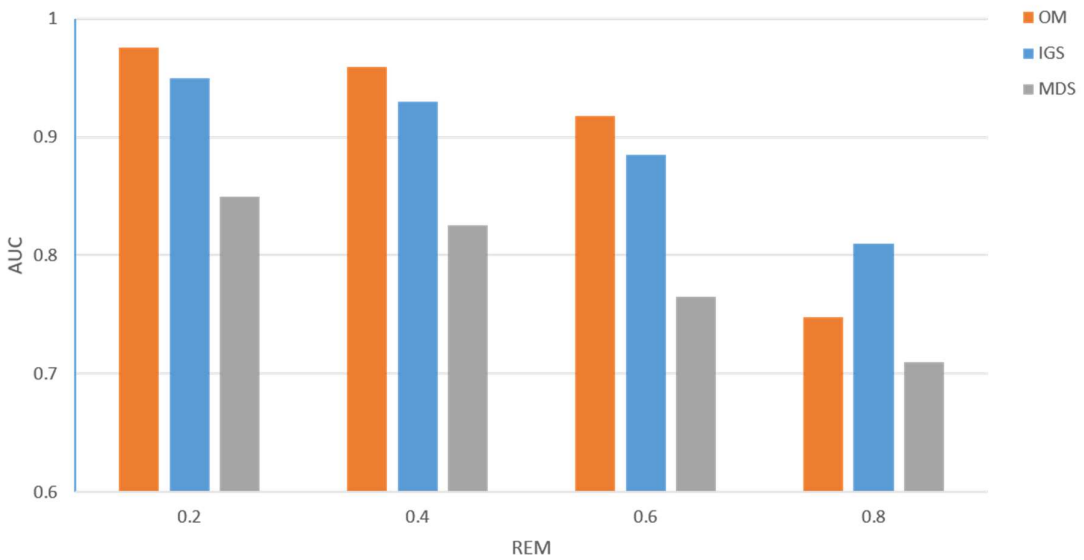


Figure 8: Application of the OM methodology with the NC topological measure, when an increasing percentage of edges was removed randomly to the CS2007 reference set network. AUCs values of the CS2007 perturbed 4 networks when 20%, 40%, 60% and 80% of edges were removed to the reference set networks, using OM, IGD and MDS methods.

Table 6: Percentage of TP inserted after applying the OM methodology to the incomplete networks of the CS2007 dataset.

% removed	# TP removed	# TP inserted	% TP inserted
20	1,665	1578	95
40	3,329	2993	90
60	4,994	4011	80
80	6,658	2633	40

Analysis

Different topological measures were used to identify the optimal threshold, with the Yeast and Human reference sets and comparative testing showed (Figs 2 and 3 and Table 2) that the best results were obtained using the JC and the NC measures, and thus we decided to use both in some experiments of this work. JC is a widely known measure frequently used in network denoising and missing link prediction. It also considers the neighbourhood information, which

421 is aligned with the "guilt-by-association" principle. The same applies to the NC index,
422 proposed herein, where the concept of CC is also taken into account.

423 The OM methodology was then applied to the Yeast and Human datasets, using the JC and the
424 NC measures and after analysing Tables 3 and 4, where the AUC values and the cut-off values
425 for both the optimal cut and the accuracy cut for the Yeast and Human reference sets obtained
426 are shown, it can be seen that the NC measure performed better than the JC measure at
427 discriminating between protein interactions and non-interactions and for this reason the NC
428 measure was used in the evaluation of the OM methodology.

429 Three different scenarios were considered to evaluate the OM methodology. The first one uses
430 randomly generated protein networks, with the same number of nodes and edges as their
431 respective reference sets (Yeast and Human). Observing the Fig 4 we can see that the AUC,
432 obtained when applying the OM methodology to one of the random networks, was close to 0.5
433 for both organisms (Yeast and Human), while for the Yeast and Human reference sets, the
434 AUC was 0.9534 and 0.8708, respectively (Fig 2), which shows that OM is sensitive to the
435 inherent topological structure of a real network. These results show that the OM methodology
436 cannot distinguish between interactions and non-interactions in random networks, but can
437 capture the inherent rules of biological networks, not present in random networks.

438 The second scenario used to evaluate the performance of OM methodology, consisted in
439 applying OM to networks obtained from the two Yeast and Human reference sets, where the
440 number of nodes (proteins) was maintained, but where a random percentage of edges (proteins
441 interactions), 20%, 40%, 60% and 80%, not belonging to the reference set network, were
442 added, and the third scenario is similar to the second but instead of adding, the same random
443 percentages of edges were removed from the reference set network.

444 In the second scenario (random insertion of edges), as expected, greater increments of random
445 edges resulted in greater performance reductions (Fig 5). The performance reductions were
446 steeper in Human, which could be attributed to one major reason: the percentage of FN is most
447 likely greater in the Human interactome than in the Yeast interactome. Thus, it could be argued
448 that the Yeast reference set is a more reliable, better representation of the actual Yeast
449 interactome, than the Human reference set is of the real Human interactome. When we add
450 these percentages of random edges, the inherent structure of these biological networks becomes
451 deteriorated, because we are probably adding TN.

452 In the third scenario (random deletion of edges), greater reductions in the number of edges
453 result in greater performance drops compared to the second scenario, but the performance drops

454 are steeper in the Human organism (Fig 7). This could be explained by the fact that we are
455 removing TP from both networks. However, since the Yeast network seems to be a closer
456 representation of its true network than the Human network, the accentuated deterioration in the
457 structure of the Human network could explain this behaviour.

458 So, when comparing the results between edge addition and edge removal in Yeast and Human
459 reference sets (Figs 5 and 7), we witness that the overall performance reductions were quite
460 dissimilar. Adding just 20% more edges contributed to a reduction of approximately 0.08 in
461 AUC for Yeast, and 0.06 AUC for Human. Further addition of edges beyond this point did not
462 decrease the AUC as sharply. Contrariwise, after removing 20% of the existing edges, the AUC
463 decreased by roughly 0.02 for both Yeast and Human, with greater performance drops after
464 each percentage of edges removal.

465 The better performance observed for the Yeast interactome could be explained by its smaller
466 size compared to the Human interactome, in addition to being relatively well-studied, meaning
467 that input data quality plays an important role in the performance of computational methods.
468 Additionally, the negative impact in performance observed after randomly adding edges
469 suggests that the OM methodology is very sensitive to high percentages of FP and FN.

470 To compare the performance of OM methodology with other network-based methodologies
471 proposed by other researchers, CS2007 network reference set was perturbed by randomly
472 inserting edges in the same proportions previously described in scenario two and by randomly
473 deleting edges in the same proportions previously described in scenario three. OM
474 methodology was compared to the MDS and IGS methodologies [23]. Figs 6 and 8 show the
475 AUC values when these methodologies were applied to this dataset and it can be seen a general
476 improvement in the performance when the OM methodology is applied compared to the MDS
477 and IGS methodologies. When 80% of random TP of network interactions are removed, the
478 AUC of IGS is superior than the AUC od OM and MDS. This could not mean, in this case, that
479 OM is worst, since removing 80% of the TP makes the model very close to a random network
480 and a denoising method, for consistency, should not be able to detect structure in such
481 networks.

482 Further analysis was conducted for these last networks with added random percentages of FP
483 interactions. OM methodology was applied and the percentage of FP interactions removed was
484 calculated (Table 5). Interestingly, most of the randomly inserted FP interactions were
485 promptly identified, even when the network was heavily perturbed, with 89% of the FP
486 removed after contaminating the network with 6,658 random interactions. These results suggest

that the OM methodology can indeed capture the inherent topology of biological networks. Interestingly, we observed that the number of TP identified after randomly removing edges from the CS2007 dataset plummets after removing 60% of TP (Table 6). Still, the OM methodology seems to identify most missing links up to that point. These findings suggest that the OM methodology can assess whether the topological structure of a network is according to the characteristic topology of biological networks.

OM methodology could still work well in less-studied interactomes, when the subset of the interactome of interest is a representative sample of the structure of the entire interactome, meaning that the percentage of FP and FN cannot hide the inherent structure behind the biological networks of the organisms.

Conclusions

Currently, low-throughput experimental methods are the only effective way to validate protein interactions. While high-throughput experimental methods to obtain PPIs exist, the obtained results have very high noise. As such, computational methods are required to speed data acquisition and to reduce the data contamination. Methods relying exclusively in the topology of biological networks are simpler and faster, as it appears that networks topology may reveal patterns or signatures associated with the kind of organism and the type of interactions. If we can use, effectively, only the topology to denoise biological networks, we have a simple computational method suitable for incomplete interactomes, without the need of extra biological knowledge.

This paper introduces the OM methodology for denoising biological networks, a methodology that: a) uses exclusively the topology of the network; b) enables, easily, to separate the distributions of interaction and non-interaction proteins in PPI networks; c) does not use known distributions as approximations; and d) provides a topological way of detecting FP interactions and find new interactions. The main innovation of the OM methodology is related with its ability to combine the advantages of using exclusively the topology without taking approximations to known distributions and without using external knowledge to detect interactions that do not exist or to find new interactions with a better performance than some documented used methodologies. This paper also introduces a new network topologic measure, the NC measure, which is used with the OM methodology and yielded better results, compared to other known and current topological measures.

The OM methodology can be explored in the future by applying it in networks belonging to other domains, where there is an inherent structure, to predict new interactions and eliminate spurious interactions.

Data Availability

The datasets supporting the conclusions of this article are available in the Bioinformatics repository of the University of Aveiro, at http://bioinformatics.ua.pt/software/OM.

Conflicts of Interest

The authors have declared that no competing interests exist.

Funding Statement

This work was supported by the RD-CONNECT European project (EC contract number 305444) that provided the logistic and computational means to conduct this work (https://rd-connect.eu/). EDC was funded by the NETDIAMOND (POCI-01-0145FEDER-016385).

Acknowledgements

Not applicable.

References

[1] M. W. Gonzalez and M. G. Kann, "Chapter 4: Protein Interactions and Disease," *PLoS Computational Biology,* vol. 8, p. e1002819, 12/27 2012.

[2] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical Organization of Modularity in Metabolic Networks," *Science,* vol. 297, pp. 1551-1555, 2002.

[3] S. Oliver, "Guilt-by-association goes global," *Nature,* vol. 403, pp. 601-3, Feb 10 2000.

[4] X. Luo, Z. Ming, Z. You, S. Li, Y. Xia, and H. Leung, "Improving network topology-based protein interactome mapping via collaborative filtering," *Knowledge-Based Systems,* vol. 90, pp. 23-32, 2015/12/01/ 2015.

[5] H. N. Chua and L. Wong, "Increasing the reliability of protein interactomes," *Drug Discov Today,* vol. 13, pp. 652-8, Aug 2008.

[6] M. Varjosalo, R. Sacco, A. Stukalov, A. van Drogen, M. Planyavsky, S. Hauri*, et al.*, "Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS," *Nat Meth,* vol. 10, pp. 307-314, 04//print 2013.

[7] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels*, et al.*, "Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs"," *Genome Res,* vol. 11, pp. 2120-6, Dec 2001.

[8] M. Tarailo, S. Tarailo, and A. M. Rose, "Synthetic Lethal Interactions Identify Phenotypic “Interologs” of the Spindle Assembly Checkpoint Components," *Genetics,* vol. 177, pp. 2525-2530, 2007.

[9] O. G. Troyanskaya, K. Dolinski , A. B. Owen, R. B. Altman, and D. Botstein "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)," *Proceedings of the National Academy of Sciences,* vol. 100, pp. 8348-8353, July 8, 2003 2003.

[10] A. Patil and H. Nakamura, "Filtering high-throughput protein-protein interaction data using a combination of genomic features," *BMC Bioinformatics,* vol. 6, p. 100, April 18 2005.

[11] G. Liu, J. Li, and L. Wong, "Assessing and predicting protein interactions using both local and global network topological metrics," *Genome Inform,* vol. 21, pp. 138-49, 2008.

[12] N. Škunca, A. Altenhoff, and C. Dessimoz, "Quality of Computationally Inferred Gene Ontology Annotations," *PLOS Computational Biology,* vol. 8, p. e1002533, 2012.

[13] J. Dutkowski, M. Kramer, M. A. Surma, R. Balakrishnan, J. M. Cherry, N. J. Krogan*, et al.*, "A gene ontology inferred from molecular networks," *Nature biotechnology,* vol. 31, p. 10.1038/nbt.2463, 2013.

[14] R. Saito, H. Suzuki, and Y. Hayashizaki, "Interaction generality, a measurement to assess the reliability of a protein-protein interaction," *Nucleic Acids Res,* vol. 30, pp. 1163-8, Mar 01 2002.

[15] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guénoche, and B. Jacq, "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network," *Genome Biology,* vol. 5, p. R6, 2004.

[16] J. Chen, W. Hsu, M. L. Lee, and S. K. Ng, "Increasing confidence of protein interactomes using network topological metrics," *Bioinformatics,* vol. 22, pp. 1998-2004, Aug 15 2006.

[17] H. N. Chua, W. K. Sung, and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions," *Bioinformatics,* vol. 22, pp. 1623-30, Jul 01 2006.

[18] E. D. Coelho, I. N. Cruz, A. Santiago, J. L. Oliveira, A. Dourado, and J. P. Arrais. (2017, November 1, 2017). A Sequence-Based Mesh Classifier for the Prediction of Protein-Protein Interactions. *ArXiv e-prints 1711*. Available: http://adsabs.harvard.edu/abs/2017arXiv171104294C

[19] L. Zhu, Z.-H. You, and D.-S. Huang, "Increasing the reliability of protein–protein interaction networks via non-convex semantic embedding," *Neurocomputing,* vol. 121, pp. 99-107, 2013/12/09/ 2013.

[20] Z.-G. Gao, L. Wang, S.-X. Xia, Z.-H. You, X. Yan, and Y. Zhou, "Ens-PPI: A Novel Ensemble Classifier for Predicting the Interactions of Proteins Using Autocovariance Transformation from PSSM," *BioMed research international,* vol. 2016, p. 4563524, 2016 2016.

[21] L.-P. Li, Y.-B. Wang, Z.-H. You, Y. Li, and J.-Y. An, "PCLPred: A Bioinformatics Method for Predicting Protein-Protein Interactions by Combining Relevance Vector Machine Model with Low-Rank Matrix Approximation," *International journal of molecular sciences,* vol. 19, p. 1029, 2018.

[22] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley, "Toward link predictability of complex networks," *Proceedings of the National Academy of Sciences,* vol. 112, pp. 2325-2330, February 24, 2015 2015.

[23] Y. Fang, M. Sun, G. Dai, and K. Ramain, "The Intrinsic Geometric Structure of Protein-Protein Interaction Networks for Protein Interaction Prediction," *IEEE/ACM Trans Comput Biol Bioinform,* vol. 13, pp. 76-85, Jan-Feb 2016.

[24] S. R. Collins, P. Kemmeren, X.-C. Zhao, J. F. Greenblatt, F. Spencer, F. C. P. Holstege*, et al.*, "Toward a Comprehensive Atlas of the Physical Interactome of Saccharomyces cerevisiae," *Molecular & Cellular Proteomics,* vol. 6, pp. 439-450, March 1, 2007 2007.

[25] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," *Behavioral Ecology and Sociobiology,* vol. 54, pp. 396-405, 2003/09/01 2003.

[26] V. E. Krebs, "Mapping networks of terrorist cells," *Connections,* vol. 24, pp. 43-52, 2002.

[27] O. Kuchaiev, M. Rašajski, D. J. Higham, and N. Pržulj, "Geometric De-noising of Protein-Protein Interaction Networks," *PLOS Computational Biology,* vol. 5, p. e1000454, 2009.

[28] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *arXiv preprint arXiv:0811.0484,* 2008.

[29] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas*, et al.*, "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Res,* vol. 43, pp. D447-52, Jan 2015.

[30] B. Taboada, C. Verde, and E. Merino, "High accuracy operon prediction method based on STRING database scores," *Nucleic Acids Research,* vol. 38, p. e130, 2010.

[31] M. Ye, G. C. Racz, Q. Jiang, X. Zhang, and B. M. E. Moret, "NEMo: An Evolutionary Model with Modularity for PPI Networks," in *Bioinformatics Research and Applications: 12th International Symposium, ISBRA 2016, Minsk, Belarus, June 5-8, 2016, Proceedings*, A. Bourgeois, P. Skums, X. Wan, and A. Zelikovsky, Eds., ed Cham: Springer International Publishing, 2016, pp. 224-236.

[32] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature,* vol. 453, pp. 98-101, May 1 2008.

[33] J. I. Fuxman Bass, A. Diallo, J. Nelson, J. M. Soto, C. L. Myers, and A. J. M. Walhout, "Using networks to measure similarity between genes: association index selection," *Nature methods,* vol. 10, pp. 1169-1176, 2013.

[34] L. Lu, C. H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Phys Rev E Stat Nonlin Soft Matter Phys,* vol. 80, p. 046122, Oct 2009.

[35] D. Simone, T. Josephine Maria, D. Claudio, and C. Carlo Vittorio, "Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks," *New Journal of Physics,* vol. 17, p. 113037, 2015.

[36] M.-W. Ahn and W.-S. Jung, "Accuracy test for link prediction in terms of similarity index: The case of WS and BA models," *Physica A: Statistical Mechanics and its Applications,* vol. 429, pp. 177-183, 2015/07/01/ 2015.

[37] D. Hao, C. Ren, and C. Li, "Revisiting the variation of clustering coefficient of biological networks suggests new modular structure," *BMC Syst Biol,* vol. 6, p. 34, May 1 2012.

[38] A. C. Lewis, N. S. Jones, M. A. Porter, and C. M. Deane, "The function of communities in protein interaction networks at multiple scales," *BMC Systems Biology,* vol. 4, p. 100, July 22 2010.

[39] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature,* vol. 402, p. C47, 1999.

[40] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular Systems Biology,* vol. 3, 2007.

[41] J.-D. J. Han, "Understanding biological functions through molecular networks," *Cell research,* vol. 18, p. 224, 2008.