This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3119569, IEEE Access

Manuscript in preparation to be submitted to IEEE Access

Digital Object Identifier 10.1109/ACCESS.2021.DOI

Protein-Protein Interaction Prediction via Graph Signal Processing

STEFANIA COLONNESE[1], (Senior Member, IEEE), MANUELA PETTI[2], LORENZO FARINA.[2], GAETANO SCARANO.[1], FRANCESCA CUOMO.[1], (Senior Member, IEEE)
[1]DIET Dept., University of Rome Sapienza (e-mail: name.surname@uniroma1.it)
[2]DIAG Dept., University of Rome Sapienza (e-mail: lorenzo.farina, manuela.petti@uniroma1.it)

Corresponding author: Stefania Colonnese (e-mail: stefania.colonnese@ uniroma1.it).

ABSTRACT This paper tackles the problem of predicting the protein-protein interactions that arise in all living systems. Inference of protein-protein interactions is of paramount importance for understanding fundamental biological phenomena, including cross-species protein-protein interactions, such as those causing the 2020-21 pandemic of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Furthermore, it is relevant also for applications such as drug repurposing, where a known authorized drug is applied to novel diseases. On the other hand, a large fraction of existing protein interactions are not known, and their experimental measurement is resource consuming. To this purpose, we adopt a Graph Signal Processing based approach modeling the protein-protein interaction (PPI) network (a.k.a. the interactome) as a graph and some connectivity related node features as a signal on the graph. We then leverage the signal on graph features to infer links between graph nodes, corresponding to interactions between proteins. Specifically, we develop a Markovian model of the signal on graph that enables the representation of connectivity properties of the nodes, and exploit it to derive an algorithm to infer the graph edges. Performance assessment by several metrics recognized in the literature proves that the proposed approach, named GRAph signal processing Based PPI prediction (GRABP), effectively captures underlying biologically grounded properties of the PPI network.

INDEX TERMS Protein-protein interaction, Markov Random Field, Graph Signal Processing

I. INTRODUCTION

THIS paper describes a Graph Signal Processing (GSP) based approach to Protein-Protein Interaction (PPI) predictions. Inferring new interactions between proteins given a set of known ones is relevant to understand fundamental biological phenomena about cross-species protein-protein interactions [1], such as those causing the 2020-21 pandemic of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [2]. Furthermore, knowledge of the interactions can be exploited for applications such as drug repurposing, where a known authorized drug is applied to novel diseases without an expensive and time consuming approval process. On the other hand, a large fraction of existing protein interactions are not known, and their experimental measurement is resource consuming. Thereby, there is an increasing interest in applying computational methods to infer new protein-protein interactions given the known ones, so as to reduce the cost of the experimental stages.

Despite its utmost importance, PPI inference is still an open problem, since protein-protein interaction networks are representative of biochemical mechanisms and show different properties with respect to other networks (e.g. social networks) [3] [4]. Thereby, PPI prediction requires specifically tailored inference algorithms. Few pioneering papers have introduced protein-protein interaction prediction strategies that resort to a network based approach [4] [5]. In [4], the authors infer the unknown interaction based on the observed path between the nodes of a graph representing the interactome. Namely, some connectivity features between two nodes, such as the existence of a large number of paths of given length, are leveraged to predict the existence of a direct link between the nodes. In [5] the authors extend the network representation by modeling the protein structure in a non Euclidean domain, and inferring the unknown interaction by geometric deep learning. Other methods exploit either protein network features such as centrality measures [6] or graph oriented computation structures [7] to infer unknown links. The link prediction problem can be faced also by

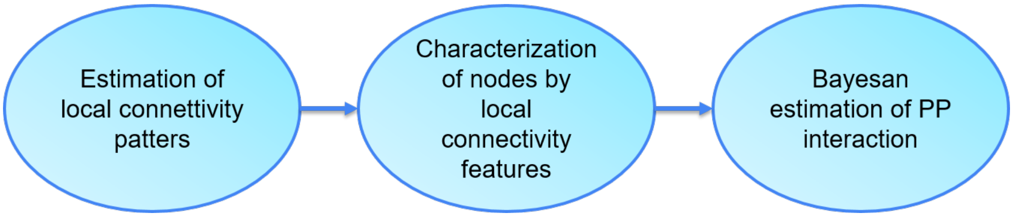

Estimation of local connettivity patters
Characterization of nodes by local connectivity features
Bayesan estimation of PP interaction

FIGURE 1. Overview of the proposed method: main stages.

network embedding, i.e. by representing large information networks into a low-dimensional vector space. This lighter network representation encompasses the key features to be considered for link prediction. In [8] the authors propose a method for learning latent description of nodes in a graph. In [9], a network embedding architecture inspired by natural language processing is proposed. Specific challenges of network embedding in large scale networks are addressed in [10]. Also in our case we consider a representation of the connectivity by means of a multidimensional Signal on Graph, but we use it as a basis for Bayesian estimation.

Remarkably, describing complex structure by suitable embedding of high-dimensional information can be carried out resorting to hypergraphs [11], in which hyperlinks connect two or more nodes, and can be estimated by specific learning techniques [12]. Herein, we address the prediction of interactions within pairs of proteins, whereas hypergraphs model more complex protein interactions, at the expense of an increased prediction complexity. Thereby, we leverage undirected graph and signal defined on graphs to predict links in the protein network.

Other contributions in the literature leverage neural networks. In [13], the authors identify anchor nodes to drive the connectivity estimation. In [14] the authors propose a link prediction theory based on node labeling and implement it within a Graph Neural Network. More in general, several deep learning methods have been proposed [15] [16] [17]. These methods, in order to achieve convergence, typically require side information. The method in [14] leverages the knowledge of a set of non existing links, also known as negative links. Other methods require a priori knowledge of the probability of existence of a link [15], or the preliminary assignment of a reliability weight to each link in the training data [17], or even the removal of ambiguity generating information [16].

Connectivity-based, deterministic methods, such as the one in [4] or that proposed here, differ from deep learning methods and aim at defining a per-link decision statistic to be used for interaction detection or ranking. These methods do not require side information and rely on biologically relevant mechanisms, as expressed by the known PPI network connectivity, and pave the way for further studies leveraging small-to-medium scale connectivity between nodes.

In this paper, we resort to the novel framework of Graph Signal Processing to solve the PPI problem. In a nutshell, the proposed approach relies on three stages, namely: i) estimation of the local connectivity patterns, ii) characterization of nodes by their known connectivity features and iii) Bayesian estimation of the unknown interactions. In Figure 1 we show a visual summary of the aforementioned stages.
Herein, these stages are cast in the Graph Signal Processing framework. Specifically, we equip the network with a node labeling related to local connectivity patterns, as estimated by the Spectral Graph Wavelet Transform (SGWT). Besides, the protein network is represented as a multidimensional Signal on Graph (SoG) representing the local protein connectivity features. Then, we design the PPI prediction as a network topology inference problem [18], also referred to in the literature as graph learning problem [19], [20], and we solve it using a Markovian Signal on Graph model [21] representing biologically relevant network connectivity features.

The key contributions of this paper are as follows:

- we formulate the PPI prediction problem in the GSP framework by providing a Signal on Graph (SoG) model of the proteins' network and the PP interactions;
- we resort to a GSP Markov model and we tune it to account for biologically relevant connectivity patterns between node pairs, such as length-3 paths;
- we propose a Maximum A Posteriori estimation method to infer links between proteins given the known links and the local connectivity patterns, as estimated by GSP based community mining.

The proposed methodology is named GRAph signal processing Based PPI prediction (GRABP) in the following. The rest of the paper is organized as follows. We briefly review the Graph Signal Processing fundamentals in Section II. The adopted GSP model is described in Section II-C while the resulting graph topology inference algorithm performing PPI prediction is presented in V. The experimental settings and the measured performance is reported in Section VI while Section VII sketches the conclusions and the future work.

II. GRAPH SIGNAL PROCESSING REVIEW

In this section, we compactly review the GSP fundamentals and the GSP tools which will be herein leveraged for PPI modeling and link prediction, namely the Spectral Graph Wavelet Transform Community mining and Markovian Signal on Graph probabilistic model.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3119569, IEEE Access

A. GSP FUNDAMENTALS

The protein to protein interactions can be studied considering the network based approach, where each protein and their interactome is modeled resorting to a graph. A graph $\mathcal{G}$ is defined by the set $\mathcal{V}$ of N vertices (nodes), the set $\mathcal{E}$ of $N \times N$ edges, where $\mathcal{E} = \{e_{ij} = (v_i, v_j)\}$, for all $v_i, v_j \in \mathcal{V}$. A graph is characterized by the $N \times N$ binary or real adjacency matrix A, whose (i, j) element represents the existence or the weight of a link between the i-th and j, respectively. The Laplacian of a graph is defined as $L = D - A$, being D the diagonal degree matrix, such that $d_{ii} = \sum_{j=1}^{N} a_{ij}$, for all $i = 1, \dots, N$. Besides, let us denote as $U = [\mathbf{u}^{(1)} \dots \mathbf{u}^{(N)}]$ and as $\lambda_1, \cdots \lambda_N$ the matrix of the eigenvectors and the eigenvalues of the Laplacian L.

A graph may be partitioned into K communities $\Gamma_k,\ k = 0, \cdots K-1$. The n-th node is associated with a community vector whose k-th component equals 1 if and only if the n-th node belongs to the k-th community. In formulas, $\boldsymbol{\gamma}_n = \boldsymbol{\delta}_k^{(K)}, \iff n \in \Gamma_k$, being $\boldsymbol{\delta}_k^{(K)}$ the k-th vector of the K-dimensional standard basis. When communities are used to model real networks, such as brain networks, within-community links can be more or less likely than between community ones [22], [23]. The attractive or repulsive behaviour of graph nodes belonging to the same community gives rise to assortative or disassortative connectivity patterns, respectively.

A Signal on Graph (SoG) is defined by associating a scalar $x_n \in \mathcal{R}$ or a vector $\mathbf{x}_n \in \mathcal{R}^M$ to the n-th node. The N orthogonal eigenvectors $\mathbf{u}^{(k)}, k = 0, \cdots N-1$ of the Laplacian L provide the basis for representing a SoG in the Fourier domain.

More specifically, the Laplacian eigenvalues play an analogous role to that of frequency in conventional Fourier Transform, and increasing eigenvalues correspond to increasingly varying eigenvectors. For a scalar SoG, the signal value at the n-th node is obtained as a linear combination of the n-th components of the graph eigenvectors as $x_n = \sum_k \tilde{x}_k \mathbf{u}^{(k)}[n], n = 0, \cdots N-1, N-1$. For a multidimensional SoG $\mathbf{x}_n \in \mathcal{R}^M$, the above formula straightforwardly stands component-wise, i.e. the m-th component of $\mathbf{x}_n[m]$ is represented as $\mathbf{x}_n[m] = \sum_k \tilde{x}_k[m]\mathbf{u}^{(k)}[n], n = 0, \cdots N-1, m = 0, \cdots M-1$.

B. SPECTRAL GRAPH WAVELET TRANSFORM AND COMMUNITY DETECTION

The local connectivity pattern of a graph, as captured by the Spectral Graph Wavelet Transform (SGWT) (see [24]), can be exploited for the purpose of graph community mining.

In classical signal processing, a wavelet basis provides a family of signals localized in time that serve for the representation of the local signal characteristics. Thereby, the wavelet basis element is parameterized not by its scale, i.e. its variation speed, but also by its location. In Graph Signal Processing, the wavelet basis element is characterized by a discrete parameter n representing a location in the vertex domain and a continuous scale parameter s. The parameter n controls the localization around the n-th node. The parameter s controls the smoothness of the wavelet function over the graph; namely, it describes the scale of a band-pass kernel function $g(\cdot)$ acting over the eigenvalues set. The basis element $\psi_n^{(s)}$ at scale s centered around the n-th node is obtained by the following linear combination of Laplacian eigenvectors:

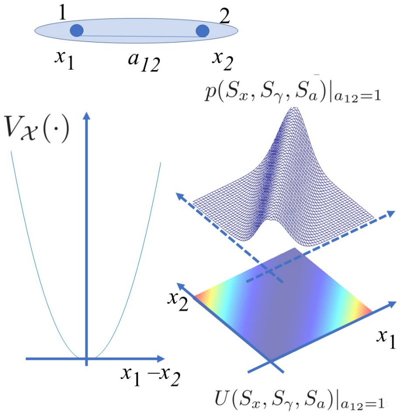

1
2
x_1
a_{12}
x_2
$p(S_x, S_\gamma, S_a)|_{a_{12}=1}$
$V_{\mathcal{X}}(\cdot)$
x_2
x_1
$x_1 - x_2$
$U(S_x, S_\gamma, S_a)|_{a_{12}=1}$

FIGURE 2. Toy case Markovian SoG ($N = 2$, one community): example of clique potential function $V_{\mathcal{X}}(\cdot)$ (left) and resulting potential function $U(S_x, S_\gamma, S_a)|_{a_{12}=1}$ and probability density function $p(S_x, S_\gamma, S_a)|_{a_{12}=1}$.

$$\boldsymbol{\psi}_n^{(s)} = \sum_{k=1}^{N} g(s\lambda_n)\mathbf{u}^{(k)}[n]\mathbf{u}^{(k)}.$$

In [25], the authors propose to partition the graph nodes into communities so as to group nodes with similar $\psi_n^{(s)}$. Specifically, in [25] the distance between two nodes is measured as the correlation distance between their SGWT elements $\psi_{n_1}^{(s)}$ and $\psi_{n_2}^{(s)}$. Hierarchical clustering of the nodes in terms of their SGWT distance leads to a partition of the graph into a set of communities $\Gamma_k^{(s)}, k = 0, \cdots K-1$, each obtained for a different value of the scale s. The n-th node belonging to the k_n community is associated a vector:

$$\boldsymbol{\gamma}_n = \boldsymbol{\delta}_{k_n}^{(K)} \qquad (1)$$

Let us remark that in this approach the n-th node is associated with a community accounting for similarity of local connectivity patterns, as represented by the SGWT basis element. The scale parameter is actually selected from a discrete, finite set of optimally stable scales $s_j, j = 0, \cdots J_{max} - 1$.

C. MARKOV RANDOM FIELD MODEL OF SIGNAL ON GRAPH

In [21] the authors present a joint Markovian model of edges, signal on graph and communities. Specifically, for the n-th node, a neighbouring set of nodes η_n is identified. The probability of a given SoG exponential decreases with a multidimensional function, named the potential function, which sums up several local contributions. The result in [21] refers

TABLE 1. Main notation

Symbol	Meaning	Set Size/Parameter Range
$\mathcal{G}^{(0)}$	protein network graph	see Table 2 different interactomes
$\mathcal{E}^{(0)}$	Set of edges in $\mathcal{G}^{(0)}$	$N_{\mathcal{E}}^{(0)}$ edges
$\overline{\mathcal{E}}^{(0)}$	Subset of unknown edges in $\mathcal{E}^{(0)}$)	$N_{\overline{\mathcal{E}}^{(0)}} \approx 10\%$ of $N_{\mathcal{E}^{(0)}}$ in the simulations
η_i	Set of one-hop neighbors (aka neighborhood) of node i	Varies in $10^0 \div 10^2$ over different nodes and interactomes.
$\mathcal{G}^{(1)}$	protein neighborhood graph	N nodes
$\Gamma_k^{(s)}$, $k = 0, \cdots K-1$	Node Communities at a given scale s	In the simulations $1 \geq s \geq s_{max} = 6$
$V_{\mathcal{X}}(\cdot,\cdot)$, $V_{\Gamma}(\cdot,\cdot)$	Clique functions	suitably defined positive non-decreasing functions.

to the case of scalar signal values but it straightforwardly generalizes to the case of multidimensional signal. In formulas, given the sets of the signal values $S_x = \{\mathbf{x}_0, \dots \mathbf{x}_{N-1}\}$, the community vectors $S_\gamma = \{\boldsymbol{\gamma}_0, \dots \boldsymbol{\gamma}_{N-1}\}$ and the link weights $S_a = \{a_{ij}, i, j = 0, \cdots N-1\}$, the potential function $U(S_x, S_\gamma, S_a)$ is written as follows:

$$U(S_x, S_\gamma, S_a) = \underbrace{\sum_{i \in \mathcal{V},\, j \in \eta_i} a_{ij}\, V_{\mathcal{X}}\left(||\mathbf{x}_i - \mathbf{x}_j||\right)}_{U_{\mathcal{X}}(\mathbf{x}, a)} \\ + \underbrace{\sum_{i \in \mathcal{V},\, j \in \eta_i} a_{ij}\, V_{\Gamma}\left(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|\right)}_{U_{\Gamma}(a, \boldsymbol{\gamma})} + \underbrace{\sum_{e \in \mathcal{E},\, k \in \eta_e} V_{\mathcal{A}}\left(a, a_k\right)}_{U_{\mathcal{A}}(a)} \tag{2}$$

where $V_{\mathcal{X}}(\cdot)$, $V_{\Gamma}(\cdot)$, $V_{\mathcal{A}}(\cdot)$ are three non-negative functions known as clique potential functions. We recognize that the function depends on the local interactions between the signal and label at the i-th node and those in its neighborhood η_i. For the sake of clarity, we depict in Figure 2 (top) a toy case of a graph of $N = 2$ nodes belonging to the same community, with scalar signal values x_1, x_2. In the Figure 2 (bottom left) we plot the clique potential function $V_{\mathcal{X}}(x_1 - x_2)$. Besides, the same figure (bottom right) illustrates the probability density function $p(S_x, S_\gamma, S_a)|_{a_{12}=1}$ and the underlying potential function $U(S_x, S_\gamma, S_a)|_{a_{12}=1}$ versus the x_1, x_2 axes. It clearly appears that the shape of the clique potential function $V_{\mathcal{X}}(x_1 - x_2)$ molds $U(S_x, S_\gamma, S_a)|_{a_{12}=1}$ and $p(S_x, S_\gamma, S_a)|_{a_{12}=1}$, definitely determining the probabilistic description of the SoG. In [21] the authors solve both a Maximum A Posteriori graph topology inference problem and a joint graph inference and signal reconstruction problem leveraging the minimization of $U(S_x, S_\gamma, S_a)$ with respect to S_a and to (S_x, S_a), respectively. In the following, we apply this model to the protein network by suitable definition of the SoG and design of the clique potential function.

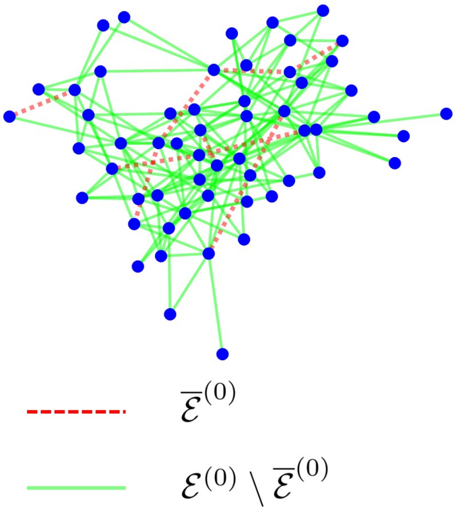

$\overline{\mathcal{E}}^{(0)}$
$\mathcal{E}^{(0)} \setminus \overline{\mathcal{E}}^{(0)}$

FIGURE 3. Protein-protein prediction as a graph learning problem: $\overline{\mathcal{E}}^{(0)}$ represents the unknown edges and $\mathcal{E}^{(0)} \setminus \overline{\mathcal{E}}^{(0)}$ the known ones.

III. PROBLEM STATEMENT

The protein-protein interaction prediction problem as follows: let $\mathcal{G}^{(0)} = (\mathcal{V}^{(0)}, \mathcal{E}^{(0)})$ be a graph whose node set $\mathcal{V}^{(0)}$ represents the set of distinct proteins, and let the graph edges in $\mathcal{E}^{(0)}$ represent the protein interactions. The interaction prediction problem consists in estimating an edge subset $\overline{\mathcal{E}}^{(0)} \in \mathcal{E}^{(0)}$ given its known complement $\mathcal{E}^{(0)} \setminus \overline{\mathcal{E}}^{(0)}$. The above described problem is illustrated in Figure 3 while the adopted notation is reported in Table 1.

The probability of observing a link in $\overline{\mathcal{E}}^{(0)}$ is known to be related to the connectivity patterns in $\mathcal{E}^{(0)} \setminus \overline{\mathcal{E}}^{(0)}$. In the following, we first formalize this observation by introducing a two-tier description of the interactome, and in the next section we show in the GSP framework this two-tier structure boils down to a graph on which a SoG is suitably defined.

Let us consider the two-tier model in Figure 4. At the lower layer, $\mathcal{G}^{(0)}$ represents the protein network graph, i.e. it has as many nodes as the number of proteins, and an edge between two nodes n_1, n_2 represents the interaction between the n_1-th, n_2-th proteins. At the higher layer, the graph $\mathcal{G}^{(1)}$ has as many nodes as the number of proteins, and the edge between two nodes n_1, n_2 represents the interaction between the one-hop neighborhoods $\eta_{n_1}^{(1)}$, $\eta_{n_2}^{(1)}$ of the corresponding n_1, n_2 nodes in $\mathcal{G}^{(0)}$. This is exemplified in Figure 4, where $\mathcal{G}^{(0)}$ and $\mathcal{G}^{(1)}$ links are represented by thin and thick lines, respectively.

With these positions, the graph $\mathcal{G}^{(1)}$ characterizes each protein in terms of the connections between its one-hop proteins neighborhood. Thereby, the existence of a path of three links between two nodes of the protein graph $\mathcal{G}^{(0)}$ reflects into a direct link of the same nodes in $\mathcal{G}^{(1)}$. According to [4], due to biological protein interaction mechanisms, the existence of a three links path between two nodes in $\mathcal{G}^{(0)}$ affects the probability of observing a direct link between the same nodes in $\mathcal{G}^{(0)}$. The inference algorithm in [4] predicts the links by means of a score function related to the number of existing

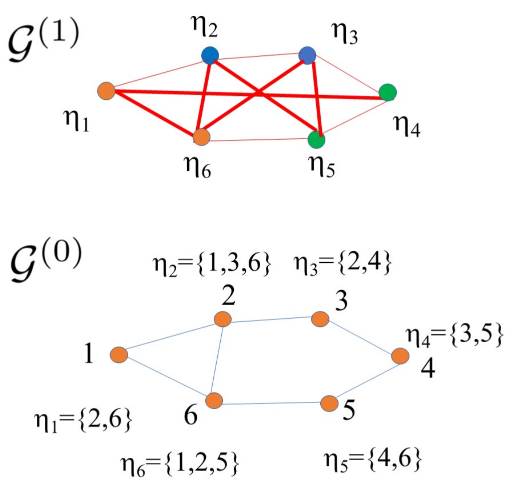


FIGURE 4. Two-tier graph example: in the graph $\mathcal{G}^{(0)}$ the nodes represent the proteins, and in the graph $\mathcal{G}^{(1)}$ the nodes represent the one-hop neighborhood of the proteins.

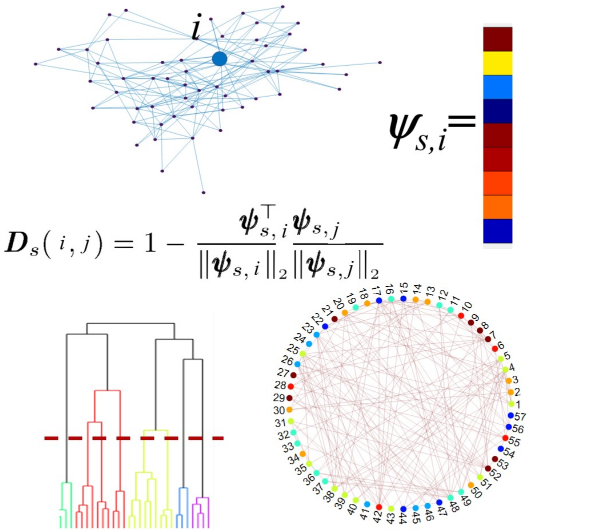

i
$\psi_{s,i}=$
$D_s(i,j) = 1 - \frac{\psi_{s,i}^{\top}\psi_{s,j}}{\|\psi_{s,i}\|_2\|\psi_{s,j}\|_2}$

FIGURE 5. Community in protein-protein interaction prediction: given the original protein network, the SGWT basis vector $\psi(s,i)$ is computed as feature vector for the node i and then hierarchical clustering is applied [25], so as to associate to the n-th node the SGWT based community information $\boldsymbol{\gamma}_n,\ n = 0, \cdots N-1$.

paths of length three between two nodes, suitably normalized by the square values of their degrees.

Let us now rephrase these observation With reference to the two-tier problem description: the existence of a direct link in $\mathcal{G}^{(1)}$ provides information about the existence of a direct link in $\mathcal{G}^{(0)}$. Accordingly, the score function in the inference algorithm [4] can be computed in terms of the graph $\mathcal{G}^{(1)}$. Let us denote by $L^{(1)}$ the laplacian matrix of $\mathcal{G}^{(1)}$. and let $W^{(1)}$ be a matrix such that its $w^{(1)}_{n_1n_2}$ equals the number of path of length three connecting the n_1-th and n_2-th network nodes. With this position, the score function of a link between the proteins n_1 and n_2 in [4] is computed as the (n_1, n_2) element of the matrix $Q^TW^{(1)}Q$, being Q the inverse square root of $D^{(1)} = \text{diag}\left(L^{(1)}\right)$.

In the following Section, the above described problem and two-tier architecture are rephrased by resorting to a Signal on Graph model, and the proposed approach is presented in detail.

IV. PROPOSED GSP MODEL OF THE PROTEIN NETWORK

Let us consider the graph $\mathcal{G}^{(0)}$, and the partially known adjacency matrix $\hat{A}^{(0)}$ with $\hat{a}_{mn} = a_{mn}$ for $(m,n) \in \mathcal{E} \setminus \overline{\mathcal{E}_0}$ and $\hat{a}_{mn} = 0$ for $(m,n) \in \overline{\mathcal{E}_0}$. The adjacency matrix $\hat{A}^{(0)}$ is leveraged to assign the node oriented features of the graph, namely the signal values and community labels, on the basis of the connectivity patterns estimated on the known set $\mathcal{E}^{(0)} \setminus \overline{\mathcal{E}}^{(0)}$. The proposed methodology relies on the three steps considered in the following.

Stage 1: SGWT based Community Mining

Firstly, we consider a partition of the graph into K disjoint communities $\Gamma_k^{(s)},\ \ k = 0, \cdots K-1$, identified based on the SGWT features of each node by applying the algorithm [25] to the estimated adjacency $\hat{A}^{(0)}$. The application of the community mining algorithm [25] is visually summarized in Figure 5, where we recognize the original graph, the SGWT based distance metric, the hierarchical clustering and the resulting association of each and every node to a distinct community. Let us denote by k_n the community the n-th node belongs to. Then, the n-th node is equipped with a K-dimensional binary community vector $\boldsymbol{\gamma}_n$ defined as in (1) and associated with its local connectivity features as measured by the SGWT.

Stage 3: Definition of the Multidimensional SoG

Secondly, we define the signal on graph so as to embed the knowledge about the one-hop neighborhood of each node. To this aim, we set

$$\mathbf{x}_n = \hat{A}^{(0)}\boldsymbol{\delta}_n^{(N)} \tag{3}$$

i.e. we assign the signal value $\mathbf{x}_n$ as the n-th column of $\hat{A}^{(0)}$. An example of signal and community vectors for a sample graph is given in Figure 6.

Stage 3: MAP estimation for the Markovian SoG

Let us consider the unknown adjacency coefficients, $a_{ij}, (i,j) \in \overline{\mathcal{E}}^{(0)}$ compactly denoted by the set $S_{\overline{a}} \subset S_a$. The prediction problem is formulated as the following Maximum A Posteriori topology inference problem:

$$a_{ij}^{(MAP)}, (i,j) \in \overline{\mathcal{E}}^{(0)} = \arg\max_{a_{ij} \in S_{\overline{a}}} p(S_x, S_\gamma, S_a) \tag{4}$$

To solve this problem, we resort to the aforementioned Markovian SoG model [21], which relates the graph edges on the protein network with the local connectivity features of the associated pair of nodes. Thus, the Maximum a Posteriori estimate of the unknown coefficients $a_{ij} \in S_{\overline{a}}$ is by min-

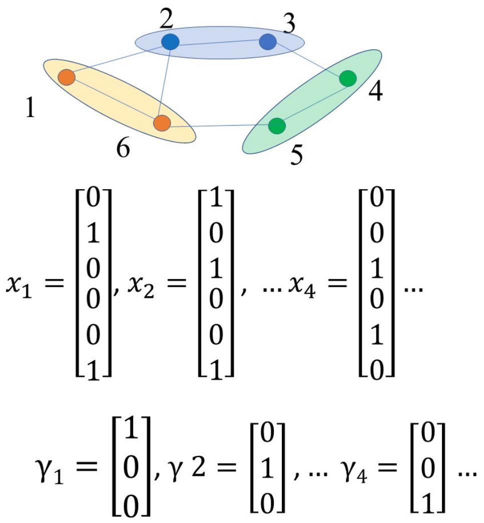

2
3
1
4
6
5
$x_1 = \begin{bmatrix}0\\1\\0\\0\\0\\1\end{bmatrix}, x_2 = \begin{bmatrix}1\\0\\1\\0\\0\\1\end{bmatrix}, \ldots x_4 = \begin{bmatrix}0\\0\\1\\0\\1\\0\end{bmatrix} \ldots$
$\gamma_1 = \begin{bmatrix}1\\0\\0\end{bmatrix}, \gamma\, 2 = \begin{bmatrix}0\\1\\0\end{bmatrix}, \ldots\ \gamma_4 = \begin{bmatrix}0\\0\\1\end{bmatrix} \ldots$

FIGURE 6. Example of a toy graph ($N = 6$) on which a community partition (colored sets) is defined. The community vectors defined as in (1) and the multidimensional SoG defined as in (3) are shown.

imization of the potential function $U(S_x, S_\gamma, S_a)$. Thereby, prediction problem is reformulated as follows:

$$a_{ij}^{(MAP)}, (i,j) \in \overline{\mathcal{E}}^{(0)} = \arg \min_{a_{ij} \in S_{\overline{a}}} U(S_x, S_\gamma, S_a) \qquad (5)$$

Considering only the terms of $U(S_x, S_\gamma, S_a)$ depending on the unknown coefficients $a_{ij} \in S_{\overline{a}}$ we obtain the final formulation in (6).

From (6), we recognize that the potential functions of the Markovian model play a fundamental role in characterizing the statistical dependencies, as discussed in the following subsection.

A. DESIGN OF THE CLIQUE POTENTIAL FUNCTIONS

As for the potential function $V_{\mathcal{X}}$, in the literature both quadratic or non-quadratic functions of the euclidean distance $||\mathbf{x}_i - \mathbf{x}_j||$ are used. Herein, we present a novel approach, aiming at computing the presence of three-hop links in $\mathcal{E}^{(0)} \setminus \overline{\mathcal{E}}^{(0)}$. In fact, in the pioneering work [4], the role of three-hop links has been identified and discussed. Herein, we encompass this criterion within a Markovian model, which allows us to naturally blend network-related information with node-related attributes. Specifically, we introduce the following clique potential function:

$$V_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j) = \kappa_{\mathcal{X}} \left(1 - \frac{||X_i \hat{A}^{(0)} X_j||_{1,1}}{||\hat{A}^{(0)}||_{1,1}}\right) \qquad (7)$$

being X_i, X_j the diagonal matrices obtained as $X_i = \text{diag}(\mathbf{x}_i)$, $X_j = \text{diag}(\mathbf{x}_j)$ and $||\cdot||_{1,1}$ the element-wise $L_{(1,1)}$ matrix norm, computed as the sum of absolute values of the matrix element. Based on these positions, the $L_{1,1}$ norm of the matrix $X_j \hat{A}^{(0)} X_j$ turns out to be an estimate of the number of 3-hop paths between i, j. This is exemplified in Figure 7, where the matrix $X_j \hat{A}^{(0)} X_j$ is computed for a graph toy case. Let us remark that this formulation allows us to represent a topological constraint in terms of an algebraic constraint, which is feasible for inclusion in the Markovian model jointly with other connectivity related features.

The clique potential function V_Γ can be set as follows:

$$V_\Gamma(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j) = ||\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j||_0 \qquad (8)$$

i.e. $V_\Gamma(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j) = 0$ for nodes belonging to the same community and 0 otherwise. This choice leads to a larger probability of observing a link between nodes belonging to the same community. This potential function well suites to the SGWT based community mining approach, which tends to identify communities corresponding to connected nodes subsets.

Finally, let us remark that the presented MRF model is very general and it can be used to encompass different approaches. For instance, the MRF based approach boils down to the method in [4] for a particular choice of the clique potential functions. More specifically, the metric proposed in in [4] can be represented in the MRF by

$$V_{L_3}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sqrt{||\hat{A}^{(0)} \boldsymbol{\delta}_i^{(N)}||_{1,1} \cdot ||\hat{A}^{(0)} \boldsymbol{\delta}_j^{(N)}||_{1,1}}}{||X_i \hat{A}^{(0)} X_j||_{1,1}}.$$

Furthermore, different choices of the potential function $V_\Gamma(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j)$ could lead to model different network structures, such as those encountered in brain functional connectivity networks [22], [26]. Thereby, the herein proposed model is very general and it paves the way for representing different biological observations by suitable design of the MRF model clique potential functions.

V. GRABP ALGORITHM

The protein graph topology inference algorithm solving:

$$\arg \min_{\overline{\mathbf{a}}^{(0)}} \sum_{i \in \mathcal{V},\, j \in \mathcal{N}_i} \overline{a}_{ij}^{(0)} \left(V_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j) + V_\Gamma(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j) \right) \qquad (9)$$

relies on the known adjacency matrix coefficients to estimate the unknown coefficients of the adjacency matrix $\hat{A}^{(0)}$. The overall computation scheme is reported in Figure 8. Given $\hat{A}^{(0)}$, the SoG $\mathbf{x}_n$ and the communities Γ_k, $k = 0, \cdots K-1$, with $\Gamma_{k_1} \cap \Gamma_{k_2} = \emptyset, k_1 \neq k_2$ and $\cup_{k=0}^{K-1} \Gamma_k = \mathcal{V}$ are estimated[1]

After this initial step, the set of most likely P edges is found as follows. Firstly, for each pair (i,j)s.t.$\hat{a}_{ij}^{(0)} = 0$ the sum of the clique functions $\Delta(i,j)$ is computed:

$$\Delta(i,j) = V_{\mathcal{X}}(||\mathbf{x}_i - \mathbf{x}_j||) + V_\Gamma(\|\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j\|)$$

The values of $\Delta(i,j)$ represent the decision statistics. Since $\Delta(i,j)$ is a distance metric, the classification is formulated as follows:

$$\Delta(i,j) \underset{\underset{\mathcal{H}_1}{\downarrow}}{\overset{\overset{\mathcal{H}_0}{\uparrow}}{\gtrless}} \theta$$

[1] Let us remark that the partition found by the community mining algorithm in [25] varies with the scale parameter s. A discussion on the selection of this parameter is reported in the Sec. devoted to experimental results.

$$a_{ij}^{(MAP)}, (i,j) \in \overline{\mathcal{E}}^{(0)} = \arg\min_{a_{ij} \in S_{\overline{a}}} \sum_{i\in\mathcal{V},\, j\in\eta_i, (i,j)\in\overline{\mathcal{E}}^{(0)}} \overline{a}_{ij}^{(0)} V_{\mathcal{X}}\left(\mathbf{x}_i, \mathbf{x}_j\right) + \sum_{i\in\mathcal{V},\, j\in\eta_i, (i,j)\in\overline{\mathcal{E}}^{(0)}} \overline{a}_{ij}^{(0)} V_{\Gamma}\left(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\right) \tag{6}$$

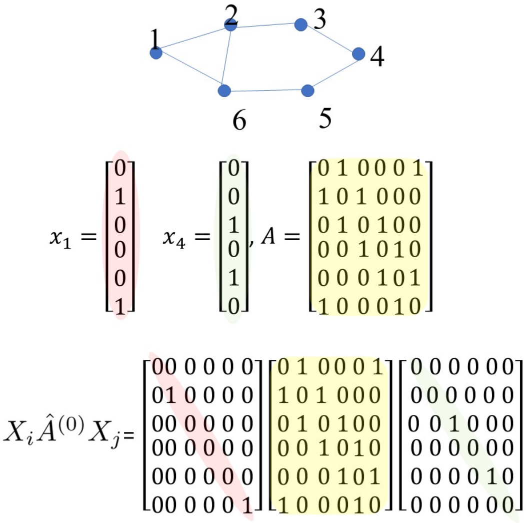

1
2
3
4
5
6
$x_1 = [0\ 1\ 0\ 0\ 0\ 1]^T$
$x_4 = [0\ 0\ 1\ 0\ 1\ 0]^T$
$A =$ 010001; 101000; 010100; 001010; 000101; 100010
$X_i\hat{A}^{(0)}X_j=$ 000000; 010000; 000000; 000000; 000000; 000001 · 010001; 101000; 010100; 001010; 000101; 100010 · 000000; 000000; 001000; 000000; 000010; 000000

FIGURE 7. Example of computation of the matrix $X_j \hat{A}^{(0)} X_j$.

where $\mathcal{H}_0$,$\mathcal{H}_1$ respectively represent the hypotheses of absence and presence of an edge between the nodes i, j, and θ is the decision threshold.

The P pairs (i, j) corresponding to the P smallest values of $\Delta(i, j)$ are selected as the P ranking candidates (i.e. likely) egdes.

The overall procedure, to which we will refer to as the GRAph BAsed PPI Prediction (GRABP) algorithm, is sketched in 1.

VI. EXPERIMENTAL RESULTS

In this Section, we apply the proposed GRABP algorithm to real and synthetic datasets. The numerical experiments are conducted in Matlab. Specifically, we implemented the model in [21], we equipped it with the community mining library in [25], and particularized it with the above defined clique potential function. Then, we applied the prediction to the different databases in Table 2, considering 10 Montecarlo runs. For each database, at each run, a subset of randomly selected $N_{\overline{\mathcal{E}}^{(0)}}$ edges is deleted from the considered graph, and the protein-protein interaction prediction is applied. The performance are assessed by averaging over 10 Montecarlo runs the performance metrics detailed in the following. Results leveraging the Markovian approach have been submitted at the Protein-Protein Interactions Prediction Challenge organized by the International Network Medicine Consortium [27]. Still, the herein proposed analysis differs from that in [27] since it relies on a different, assortative, community model, implemented by a different design of the clique potential functions.

Algorithm 1 GRAph signal processing BAsed protein Prediction (GRABP)

```
Input: Estimated protein graph adjacency matrix $\hat{A}^{(0)}$
Output: Values of $\Delta(i,j)$ computed for any $(i,j)\ s.t. \hat{A}_{ij}^{(0)} = 0$; list of $P$ node pairs $(i_p, j_p), p = 0,\cdots P-1$ such that $\hat{a}_{i_p\, j_p} = 0$, ranked in terms of increasing $\Delta(i_p, j_p), p = 0,\cdots P-1$. For each connected component of the graph $\mathcal{G}^{(0)}$
  Step 1: Given $\hat{A}^{(0)}$, compute the estimated community partition: $\hat{\Gamma}_k,\ k = 0,\cdots K-1$, with $\hat{\Gamma}_{k_1} \cap \hat{\Gamma}_{k_2} = \emptyset, k_1 \neq k_2$ and $\cup_{k=0}^{K-1}\hat{\Gamma}_k = \mathcal{V}$
  Step 2: Given $\hat{A}^{(0)}$, estimate the Signal on Graph $\mathbf{x}_n$ for the protein graph $\mathcal{G}^{(0)}$
  Step 3a: Compute the sum of clique functions $\Delta(i,j)$:
      $\Delta(i,j) = \alpha V_{\mathcal{X}}\left(||\mathbf{x}_i - \mathbf{x}_j||\right) + (1-\alpha)V_{\Gamma}\left(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|\right)$
  Step 3b: Classify missing edges according to the following decision formula:
      $\Delta(i,j) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \theta$
  Step 3c: Select the $P$ pairs $(i,j)$ corresponding to the smallest $P$ values of $\Delta(i,j)$.
```

TABLE 2. Number of nodes N, number of true interactions $N_{\mathcal{E}}^{(0)}$ and number of unknown interactions $N_{\overline{\mathcal{E}}^{(0)}}$ for the largest connected component of different databases.

Database	N	$N_{\mathcal{E}}^{(0)}$	$N_{\overline{\mathcal{E}}^{(0)}}$
Synth. [28]	6597	43313	4331
Human [29]	8149	52016	5202
C. elegans [30]	2214	3538	354
Yeast [31]	1647	2518	252
Arab. [33]	2532	5919	592

The GRABP algorithm performance is firstly assessed on a synthetic graph modeling the human interactome according to the model in [28]. Then, PPI are inferred on the human interactome [29], a worm interactome (C. elegans) [30], a yeast interactome (S. Cerevisiae) [31], [32], and on a plant interactome (Arabidopsis thaliana) [33]. Performance is given in terms of the following metrics: Area Under the Curve for the Receiver Operating Characteristic (AUROC) and True Positive Percentage (TPP) computed in the 500 top-ranking interactions. In fact, the Receiver Operating Characteristic represents the probability of correctly detecting an existing edge (probability of detection) with respect to the probability of false alarm, as observed by comparing the decision statistics $\Delta(i, j)$ to different values of the decision

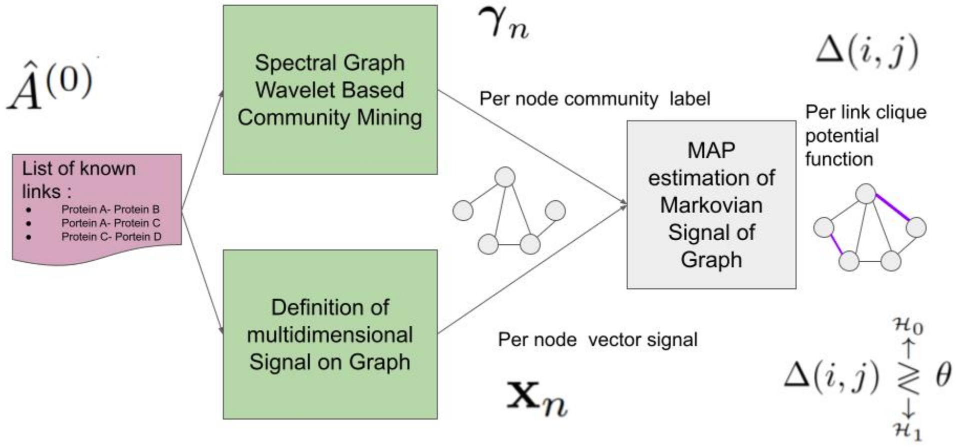

$\hat{A}^{(0)}$
List of known links :
Protein A- Protein B
Portein A- Protein C
Protein C- Portein D
Spectral Graph Wavelet Based Community Mining
Definition of multidimensional Signal on Graph
γ_n
Per node community label
Per node vector signal
$\mathbf{x}_n$
MAP estimation of Markovian Signal of Graph
$\Delta(i,j)$
Per link clique potential function
$\Delta(i,j) \overset{\mathcal{H}_0}{\underset{\mathcal{H}_1}{\gtrless}} \theta$

FIGURE 8. Overview of the proposed method: computation algorithm.

threshold θ. The larger is the probability of detection for a given probability of false alarm, the larger is its area, i.e. the AUROC. The TPP of the top-ranking interactions is defined as the percentage of true edges found among the 500 top ranking edges (that is the edges corresponding to the 500 smallest decision statistics $\Delta(i,j)$).

Let us first analyze the effect of the Markov model parameters on the performance of the proposed GRABP algorithm. To this aim, we consider the Synthetic dataset. Besides, we set $\kappa_{\mathcal{X}} = \alpha, \kappa_{\Gamma} = 1-\alpha$, $\alpha \in [0,1]$. Thus, α represents the relative weight of the metric in (7) on $\mathbf{X}$ measuring 3-hop connectivity and the metric in (7), (8) measuring local connectivity by SGWT features. We present results relative to $\alpha = 0.75$.

We will first discuss the effect of the selected community scales parameters on the results. The index s_j controls the scale of the SGWT analysis, and is selected out of $J_{max} = 6$ optimally stable scales. In Figure 9 we report results obtained on the Synthetic dataset, connected component of size $N = 57$. In Figure 9 we recognize that the best performance in terms of Area under the curve, TPP, and rate of TPP increase are obtained by selecting $s_j, = 3, J_{max} = 6$. The same behavior is observed on different connected components.

The reason why this occurs is exemplified in Figure 10, that shows the partitions obtained on the third connected component of the synthetic dataset for $s_j,\ j = 3, \cdots 6\}$. Specifically, all the $N = 57$ graph nodes are represented on a circle and their links are represented in light gray. Each node belongs to a community identified by a color. We recognize that the nodes are partitioned into a number K of communities which depends on the parameter j. The smallest and largest values of j correspond to limit conditions (i.e. $K = N$ or $K = 2$) while intermediate values of j correspond to mid-scale community partitions. The best results are in this case obtained for this latter partition by selecting $s_j,\ j = 3$ out of $J_{max} = 6$ available scales.

More in general, selecting a value of j close to J_{max} tends to constrain the generated link within the node community. Therefore, it is expected that the optimal scale number and index $J_{max},\ \ j$ may vary from one network to another, depending on the structure, network size and so on. This can be taken into account by selecting the values of $J_{max},\ j$ that provide the best performance on the known edges set $N_{\mathcal{E}}^{(0)} - N_{\overline{\mathcal{E}}^{(0)}}$ and then apply the selected values to estimate the $N_{\overline{\mathcal{E}}^{(0)}}$ unknown edges.

For the sake of comparison, let us observe that the performance achieved on the $N = 57$ component of the Synthetic dataset adopting a mid-scale partition (e.g. $j = 3$ in Figure 9) are substantially aligned with those achieved on the same component by the Degree Normalized L3 (DNL3) algorithm in [4]. Specifically, DNL3 achieves an average TPP of 0.929 versus the average of 0.923 achieved by the proposed GRABP algorithm for $j = 3$.

We now consider the results obtained on a synthetic component and on a real dataset of comparable size, namely the Yeast dataset. Specifically, we first analyze the performance of GRABP as a function of the index j, and then compare it with the DNL3 algorithm.

Figure 11 refers to the $N = 1591$ component of the Synthetic dataset. In Figure 11(a)-(b), we recognize that the best performance both in terms of AUROC and TPP are obtained for $j = 3$. Then, in Figure 11(c)-(d) we compare the AUROC and TPP of the GRABP and DNL3 algorithms. The algorithms' performance are very similar; in both the algorithms the median value of AUROC is higher than 0.94, with a slight superiority of the DNL3 algorithm, while performances in terms of TPP are very similar.

Conversely, Figure 12 refers to the Yeast dataset ($N =$

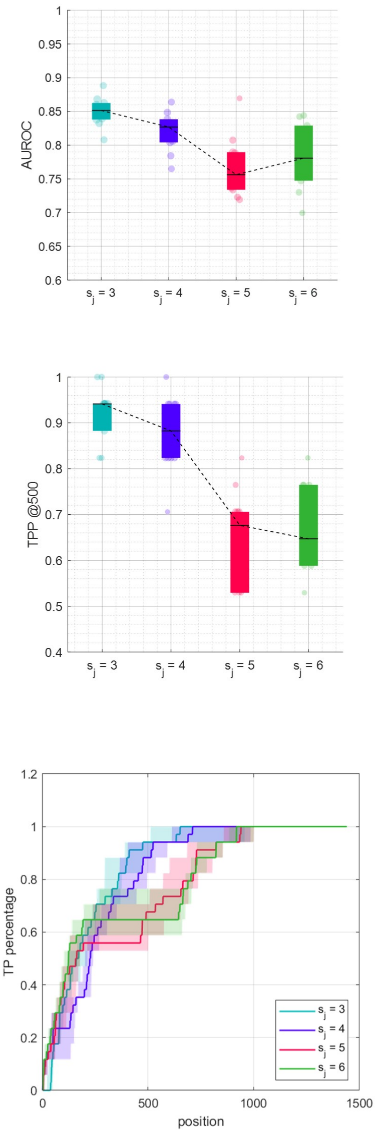


FIGURE 9. AUROC, TPP and TPP increasing rate computed in the 500 top-ranking interactions, synthetic dataset, connected components $N = 57$.

1647). Given the results in Figure 12(a)-(b), we select $j = 5$. Then, in Figure 12(c)-(d), comparing the AUROC and TPP of the GRABP and DNL3 algorithms, we observe that the GRABP algorithm outperforms DNL3 in terms of AUROC while the performances in terms of TPP are very similar.

We now present results obtained on the real interactomes, whose features are summarized in Table 2. The index s_j controlling the scale of the SGWT analysis is obtained for $j = 3, J_{max} = 6$, i.e. the results are obtained by selecting the mid-scale community partition referring to index 3 out of 6 available scales. The parameter α is set to $\alpha = 0.75$. For the sake of concreteness, the per-node computation time T_c/N is also reported.

Finally, we provide examples of link detection on the real databases in Figure 13. Specifically, for each dataset we plot a subgraph representing a true detected link (red line), the pair of corresponding nodes $n_1 = 1$, $n_2 = 2$ (largest points) and their first order neighbours (smallest points). Each node is colored depending on its community label. We recognize that the edge is correctly detected based on the clique potential $\Delta_{n_1 n_2}$, accounting for both the community of nodes $n_1 = 1$, $n_2 = 2$ and their neighbours connectivity (depicted by three path links between $n_1 = 1$ $n_2 = 2$).

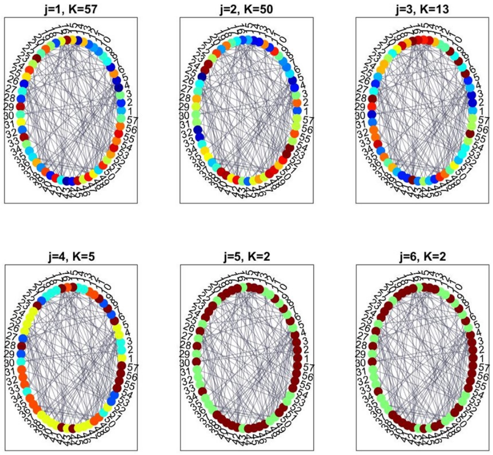


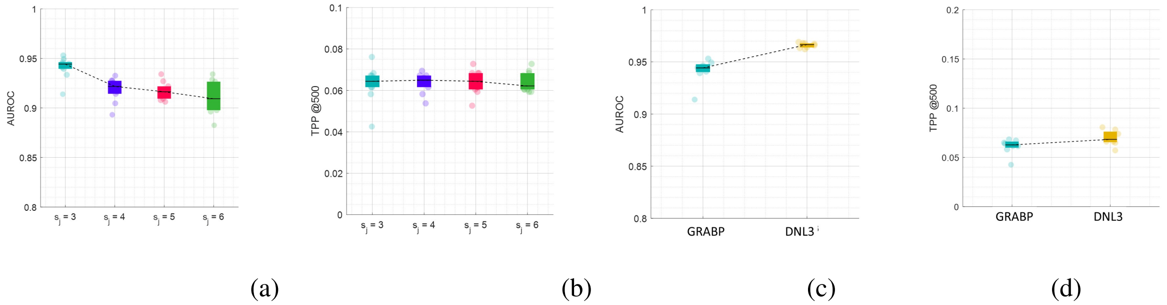
FIGURE 10. Partitions obtained on the third connected component of the synthetic dataset for $j = 3, \cdots 6\}$. The best performance is obtained for the mid-scale partition $j = 3$.

TABLE 3. Area Under ROC (AUROC), True Positive Percentage (TPP) and per node computation time T_{Comp}/N [s] for the largest connected component of different databases.

Database	AUROC ($m \pm \sigma$)	TPP	T_{Comp}/N [s]
Synth. [28]	0.961 ± 0.002	1.67 ± 0.17	1.414 ± 0.111
Human [29]	0.880 ± 0.003	2.5 ± 0.17	1.929 ± 0.118
Arab. [33]	0.825 ± 0.007	7.89 ± 0.71	0.112 ± 0.003
Yeast [31]	0.771 ± 0.012	5.56 ± 1.62	0.036 ± 0.001
C. elegans [30]	0.770 ± 0.010	2.46 ± 0.63	0.068 ± 0.002

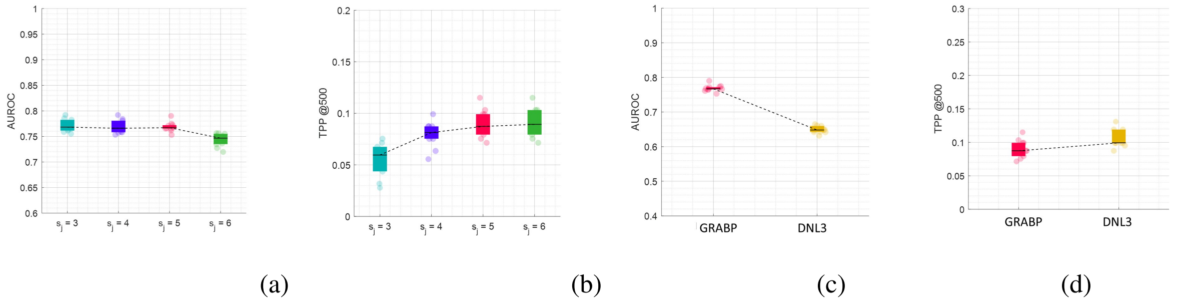
A. DISCUSSION

A few remarks are in order. Firstly, it is worthy noting that these results are obtained based on the mere analysis of the Signal on Graph associated to the PPI network, where each node is characterized in terms of its connectivity patterns. Differently from deep learning based methods such as those in [15], [16], the algorithm does not employ any side information. Still, the algorithm can be extended to account for different biochemical information, e.g. the propensity to bind, by suitably tailoring the community information. On the other hand, the metric $\Delta_{n_1 n_2}$ computed for each link can be leveraged as an additional feature in existing deep learning based methods. These developments are left for further study. To sum up, the proposed method is very general, and it is controlled by few parameters. On the other hand, the method is flexible since the GSP model can be molded using differ-



(a) (b) (c) (d)

FIGURE 11. Synthetic dataset, connected components $N = 1591$: (a) AUROC vs j, (b) TPP vs j, (c) AUROC for GRABP ($j = 3$) and DNL3, (b) TPP for GRABP ($j = 3$) and DNL3.



(a) (b) (c) (d)

FIGURE 12. Yeast dataset, connected components $N = 1642$: (a) AUROC vs j, (b) TPP vs j, (c) AUROC for GRABP ($j = 5$) and DNL3, (b) TPP for GRABP ($j = 5$) and DNL3.

ent clique potential functions. Furthermore, it has restrained computational complexity since the adopted clique potential functions allow to deal with complex topological information using fast algebraic computations.

The work paves the way to different developments. The clique functions can be molded to investigate further priors on the network topology. The node community labels can be assigned either based on local connectivity or based on physical or biochemical protein properties. Finally, the clique potential functions can be adapted and the parameters tuned to the particular network under concern. Future work will address data-driven learning of the GSP based protein network model.

VII. CONCLUSION

This paper tackles the problem of protein-protein interaction prediction by casting the network-based approach as a Graph Signal Processing (GSP) problem. To this aim, we equip the graph representing the PPI network with a multidimensional Signal on Graph related to node specific connectivity patterns, either extracted directly from the graph adjacency matrix or estimated by use of the Spectral Graph Wavelet Transform. The PPI prediction is formulated as a Maximum A Posteriori graph topology inference problem, which is then solved by a suitably designed GSP based Markovian model. Experimental results on synthetic and real protein networks assess the tightness of the GSP model and the accuracy of the related graph learning method. This opens the way for several further studies, such as for modeling static or dynamic interaction using different GSP probabilistic models, leveraging the vertex oriented Spectral Graph Wavelet Transform to analyze particular proteins involved in biological processes, as well as using different GSP probabilistic models to solve the Bayesian network topology inference problem.

VIII. ACKNOWLEDGMENTS

This work has been partially supported by a Sapienza University of Rome grant entitled “Network medicine-based machine learning and graph theory algorithms for precision oncology” - n. RM1181642AFA34C2 and by the Sapienza grant “IoT networks for 5G smart health applications” - n. RM11916B6A022E2C.

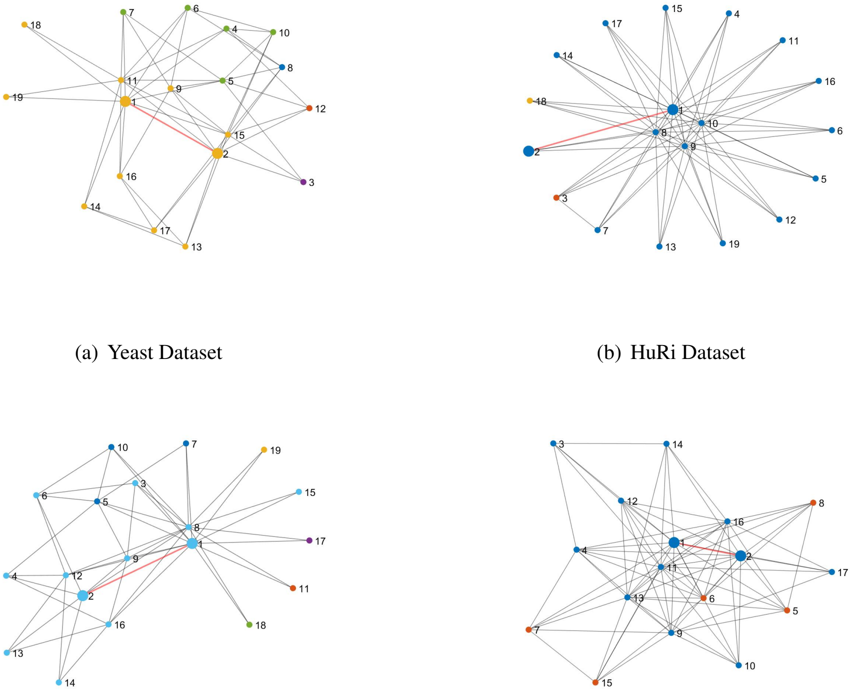
REFERENCES

[1] D. Szklarczyk, et al., “STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets,” Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019.

[2] Perrin-Cocon, L., Diaz, O., Jacquemin, C. et al. “The current landscape of coronavirus-host protein–protein interactions.” J Transl Med 18, 319 (2020).

[3] P.Tieri, L.Farina, M.Petti, L.Astolfi, P.Paci, F. Castiglione, Network Inference and Reconstruction in Bioinformatics, Encyclopedia of Bioinformatics and Computational Biology, Editor(s): S.Ranganathan, M.Gribskov, K.Nakai, C.Schönbach Academic Press, 2019, Pages 805-813, ISBN 9780128114322, https://doi.org/10.1016/B978-0-12-809633-8.20290-2.

[4] I.A. Kovács, K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D.K. Kim, N. Kishore, T. Hao, M.A. Calderwood, “Network-based



(a) Yeast Dataset

(b) HuRi Dataset

(c) C.Elegans Dataset

(d) Arabidopsis Dataset

FIGURE 13. Examples of link detection on the real databases Yeast (a), HuRi (b), C.Elegans (c) and Arabidopsis datasets (d): true detected link (red), pair of corresponding nodes $n_1 = 1$, $n_2 = 2$ and subgraph of their first order neighbours. Each node is assigned a color depending on the community it belongs to.

prediction of protein interactions", Nature communications, 10(1), pp.1-8 2019.

[5] P.Gainza, F. Sverrisson, F. Monti, et al. "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning", Nat. Methods 17, 184–192, 2020.

[6] E. Estrada, G. J. Ross, "Centralities in simplicial complexes. Applications to protein interaction networks", Journal of Theoretical Biology, Volume 438, 2018.

[7] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P. and Bengio, Y., . "Graph attention networks", arXiv preprint arXiv:1710.10903, 2017.

[8] Perozzi, B., Al-Rfou, R. and Skiena, S., 2014, August. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 701-710).

[9] Liu, J., He, Z., Wei, L. and Huang, Y., 2018, July. Content to node: Self-translation network embedding. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1794-1802).

[10] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J. and Mei, Q., 2015, May. Line: Large-scale information network embedding. In Proceedings of the 24th international conference on world wide web (pp. 1067-1077).

[11] F. Luo, L. Zhang, X. Zhou, T. Guo, Y. Cheng and T. Yin, "Sparse-Adaptive Hypergraph Discriminant Analysis for Hyperspectral Image Classification," in IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 6, pp. 1082-1086, June 2020.

[12] F. Luo, L. Zhang, B. Du and L. Zhang, "Dimensionality Reduction With Enhanced Hybrid-Graph Discriminant Learning for Hyperspectral Image Classification," in IEEE Trans. on Geoscience and Remote Sensing, vol. 58, no. 8, pp. 5336-5353, Aug. 2020.

[13] You, J., Ying, R. and Leskovec, J., 2019, May. Position-aware graph neural networks. In Int. Conf. on Machine Learning (pp. 7134-7143). PMLR.

[14] Zhang, M. and Chen, Y., 2018, December. Link prediction based on graph neural networks. In Proceedings of the 32nd Int. Conf. on Neural Information Processing Systems (pp. 5171-5181).

[15] Huang, L., Liao, L. and Wu, C.H., 2018. Completing sparse and disconnected protein-protein network by deep learning. BMC bioinformatics, 19(1), pp.1-12.

[16] Sun, T., Zhou, B., Lai, L. and Pei, J., 2017. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC bioinformatics, 18(1), pp.1-8.

[17] Yang, F., Fan, K., Song, D. and Lin, H., 2020. Graph-based prediction of Protein-protein interactions with attributed signed graph embedding. BMC bioinformatics, 21(1), pp.1-16.

[18] Segarra, S., Marques, A.G., Mateos, G. and Ribeiro, A., 2017. Network topology inference from spectral templates. IEEE Transactions on Signal and Information Processing over Networks, 3(3), pp.467-483.

[19] Egilmez, H.E., Pavez, E. and Ortega, A., 2017. Graph learning from data under Laplacian and structural constraints. IEEE Journal of Selected Topics in Signal Processing, 11(6), pp.825-841.

[20] Dong, X., Thanou, D., Frossard, P. and Vandergheynst, P., 2016. Learning Laplacian matrix in smooth graph signal representations. IEEE Transactions on Signal Processing, 64(23), pp.6160-6173.

[21] S.Colonnese, P. Di Lorenzo, T. Cattai, G. Scarano, F.D.V. Fallani, "A Joint Markov Model for Communities, Connectivity and Signals Defined Over Graphs", IEEE Signal Processing Letters, 27, pp.1160-1164, 2020.

[22] B.Retzel, J.Medaglia, D.Bassett, "Diversity of meso-scale architecture in human and non-human connectomes", Nature Communications, 9, 2018

[23] S.P.Patankar, J.Z.Kim, F.Pasqualetti, D.S.Bassett, "Path-dependent connectivity, not modularity, consistently predicts controllability of structural brain networks, " Network Neuroscience, 4(4), pp.1091-1121, 2020.

[24] D.K. Hammond, P. Vandergheynst, R. Gribonval. "Wavelets on graphs via spectral graph theory." Applied and Computational Harmonic Analysis 30, no. 2 (2011): 129-150.

[25] N. Tremblay, P. Borgnat, "Graph wavelets for multiscale community mining" *IEEE Transactions on Signal Processing*, 62(20), pp.5227-5239.

[26] T.Cattai, S.Colonnese, M.C. Corsi, D.S.Bassett, G.Scarano, F.D.V. Fallani, "Characterization of mental states through node connectivity between brain signals", 26th European Signal Processing Conference , pp. 1377-1381 Eusipco 2018.

[27] Xu-Wen Wang, L. Madeddu, E. Petrillo, A.L.Barabási, E. K. Silverman, J. Loscalzo, K. Spirohn, Tong Hao, M.Calderwood, S. Colonnese, M. Petti, F. Cuomo, G. Scarano et al, Paola Velardi, Yang-Yu Liu, "Community assessment to advance computational prediction of protein-protein interactions", Internal Communication (manuscript in preparation).

[28] Vázquez, A., et al., "Modeling of protein interaction networks, in The Structure and Dynamics of Networks", p. 408-414, 2011.

[29] Luck, K., et al., "A reference map of the human binary protein interactome", Nature, Springer US p. 402-408. 2020.

[30] Simonis, N., et al., "Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network", Nature Methods, 6(1): p. 47-54, 2009.

[31] H.Yu, H., et al., "High-quality binary protein interaction map of the yeast interactome network. Science, 322(5898): p. 104-10, 2008.

[32] . Ito, T., et al., "A comprehensive two-hybrid analysis to explore the yeast protein interactome", Proc Natl Acad Sci U S A, 98(8): p. 4569-74. 2001.

[33] Arabidopsis Interactome Mapping, C.," Evidence for network evolution in an Arabidopsis interactome map", Science, 333(6042): p. 601-7, 2011.

• • •