Hierarchical hidden community detection for protein complex prediction

A Preprint

Chao Li
School of Computer Science and Technology
University of Science and Technology
Wuhan, 430074, China
D201880880@hust.edu.cn

Kun He *
School of Computer Science and Technology
University of Science and Technology
Wuhan, 430074, China
brooklet60@hust.edu.cn

Guangshuai Liu
School of Computer Science and Technology
University of Science and Technology
Wuhan, 430074, China

John E. Hopcroft
Department of Computer Science
Cornell University
Ithaca NY, 14853, U.S.A.

October 9, 2019

ABSTRACT

Motivation: Discovering functional modules in protein-protein interaction (PPI) networks by optimization methods remains a longstanding challenge in biology. Traditional algorithms simply consider strong protein complexes that can be found in the original network by optimizing some metrics, which causes obstacles for the discovery of weak and hidden complexes shielded by stronger complexes. Also, protein complexes are not only in different density, but also in a large range of scale, making it extremely difficult to be detected. Toward this objective, we propose a hierarchical hidden community approach to predict protein complexes.

Results: We propose a method called HirHide (Hierarchical Hidden Community Detection), which can be combined with traditional community detection methods to enable them to discover hierarchical hidden communities. It is the first community detection algorithm that can find hierarchical structure as well as hidden structure. We compare the performance of three traditional methods with their HirHide versions. Experimental results show that the HirHide methods using traditional methods as the base algorithms achieve better performance, sometimes even significantly outperform the baselines.

Availability: All data and code are available at https://github.com/JHL-HUST/HirHide/

1 Introduction

A protein complex is a group of proteins that interact with one another for specific biological activities (Fiannaca et al., 2013). Identification of protein complexes is important for predicting protein functions (Schwikowski et al., 2000; King et al., 2004; Winterhalter et al., 2014), disease genes (Lage et al., 2007; Yang et al., 2011), phenotypic effects of genetic mutations (Fraser and Plotkin, 2007), and drug-disease associations (Yu et al., 2015). Given a protein-protein interaction (PPI) network, where nodes represent proteins and edges represent interactions, protein complexes can be searched by detecting densely connected subgraphs in the network. In mathematics, such subgraphs are called communities, in which network nodes are joined together in tightly knit groups, between which there are only looser connections (Girvan and Newman, 2002; Newman, 2003).

*Corresponding Author
Hierarchical hidden community detection for protein complex prediction

Figure 1: Sparse complexes in the MIPS complex catalogue database.

Figure 2: Example to show the hierarchical hidden structure. (a) A network with hierarchical hidden structure. When the hidden structure is not detected, the graph exists of four large communities consisting of 128 nodes, each with an internal subdivision of four small communities with 32 nodes. When the hidden structure is detected, the graph exists of five large communities. Four are dominant and one is hidden. (b) The network after weakening the stronger communities. The community with green nods is the hidden community, which is detected by weakening the stronger communities. So it doesn’t contain dense edges. (c) Schematic diagram of a sample hierarchical hidden structure.

Community detection plays a significant role in biological network analysis, and provides insight into the underlying structure existing in the networks (Girvan and Newman, 2002; Newman, 2003). Over the last decade, more than 20 computational methods have been proposed for detecting communities in networks. Early work focused primarily on identifying disjoint communities that partition the set of nodes within a network (Blondel et al., 2008; Pons and Latapy, 2005; Rosvall and Bergstrom, 2008). And then, researchers have observed the interleaving overlap memberships of communities (Ahn et al., 2010; Coscia et al., 2012; Lancichinetti et al., 2011; Yang and Leskovec, 2012). Some partitioning community detection algorithms are also extended to solve the overlapping case (Gregory, 2009; Zhang et al., 2007). But almost all of these algorithms focus primarily on the prominent community structure in networks, which is potentially problematic. Because weak communities that are shielded by dominant communities are sometimes of high value. For example, true protein complexes are not always dense communities. Sometimes they can be very sparse (Liu et al., 2016) (see Fig. 1). As proteins get involved in several interactions, such weak protein complexes are usually hidden behind the stronger complexes. And there are many interactions between proteins that are undiscovered (Stumpf et al., 2005), thus no connections between these proteins in PPI networks. As a consequence, sparse protein complexes may be only sparse in existing incomplete PPI networks. They are likely to be dominant in real-world PPI networks if all interactions between proteins have been discovered. That is, sparse protein complexes may also be potential stronger protein complexes, but they are overlooked in previous research.

Traditional methods can find a part of sparse protein complexes by detecting weak communities. However, they can’t deal with the case when most nodes of the weak communities also belong to stronger communities. In this case, these weak communities are hidden community structure (He et al., 2018). In PPI networks, we call these hidden community structure as hidden protein complexes. In Fig. 2(a), we build a network with hidden community structure. Because traditional algorithms focus on discovering dominant communities, the weaker community with green nodes is generally overlooked. Even if it is detected, its structure is considered as the structure of green nodes in Fig. 2(a), which contains four smaller blocks with dense intra-connections. But its structure is actually like the structure of green nodes in Fig. 2(b). The community in (b) becomes detectable because the stronger communities have been weakened. And because edges belonging to stronger communities are removed, this hidden community doesn’t contain four smaller blocks.
with dense intra-connections like Fig. 2 (a). Mathematical graph representations of many protein complexes related to sparser communities are hidden communities. They are partially or totally covered by stronger protein complexes. This problem is difficult to solve. Traditional algorithms simply consider hidden protein complexes as a part of stronger protein complexes, which cause great obstacles for them to be discovered. This can partially explain why traditional methods are not working well.

To address this problem, we design HirHide, which is inspired by a meta-approach called HiCode (Hidden Community Detection) (He et al., 2015, 2018), a novel approach that first addresses the hidden community problem. In the work of HiCode, the authors first propose a new graphic concept of hidden community structure. However, HiCode doesn’t consider the hierarchical structure, and hence it can’t handle complicated networks where communities are organized in hierarchical structure, which are existing in many real-world networks. In large-scale PPI networks, many protein complexes are organized hierarchically, which indicates that protein complexes may consist of sub-complexes extending to several hierarchical levels deep. An example of such a deeply embedded complex is the SAGA complex (MIPS identifier 510.190.10.20.10), a multi-functional coactivator that regulates transcription by RNA polymerase II (Nepusz et al., 2012).

The main contributions of this work include:

- We combine the detection of hierarchical structure with hidden structure, which could detect communities of different strength (related to density) as well as communities of different scale (related to size) in HirHide.
- HirHide is designed as a general method which can be combined with standard community detection methods to enable them to discover hierarchical hidden communities and perform better.

In all our experiments, level is used to denote the stratification of hierarchical structure and layer is used to denote the stratification of hidden structure. Multi-level hierarchical structure and multi-layer hidden structure make up the multi-granular characteristic of HirHide. In the experiments, we set two layers for hidden structure (one is dominant, the other is hidden), two levels for hierarchical structure to avoid over-complicated (a total of 4 granular divisions). Layer1-level1 represents the first level of the first layer, which is the root communities of strong communities; Layer1-level2 indicates the second level of the first layer, which is a more detailed division of the root communities; Layer2-level1 means the root communities of hidden communities; Layer2-level2 denotes a more detailed division of root communities in hidden layer. In the following, the four community divisions are shortly recorded as $L_{11}, L_{12}, L_{21}, L_{22}$.

## 2 Methods

### Hiddenness Value

A PPI network can be represented as a graph $G = (V, E)$, where $V$ is the node set, $E$ is the edge set. Suppose the network is divided into $K$ communities, denoted by $C = \{C_1, C_2, ..., C_K\}$.

Let $S_k$ be a collection of nodes belonging to stronger communities in $C_k$:

$$S_k = \{C_i | F_i > F_k, C_i \in C\}. \quad (1)$$

$F_k$ represents the strength of community $k$, which can be calculated by the modularity (Newman, 2003). The larger the value is, the stronger the community structure is. Modularity $Q$ is defined as follows:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{d_i d_j}{2m}\right] \delta(Com_i, Com_j), \quad (2)$$

where $m$ indicates the number of edges in the graph, $d_i$, $d_j$ separately indicate the degree of node $i$ and node $j$, $Com_i$, $Com_j$ separately indicate the community to which nodes $i$ and $j$ belong. And $\delta(Com_i, Com_j)$ indicates whether nodes $i$ and $j$ belong to the same community. If so, $\delta(Com_i, Com_j) = 1$, otherwise 0.

He et al. raise a formula to calculate the hiddenness value of a community (He et al., 2018). This definition actually calculates the fraction of nodes of $C_k$ belonging to various stronger communities. However, this formula doesn’t consider the structure information of $C_k$. A sparse-connected community should be more likely to be hidden than a dense-connected community even if it has the same fraction of nodes belonging to stronger communities. Consequently, we improve the concept of hiddenness value. The hiddenness value of a community $C_k$ is defined as:

$$H(C_k) = \frac{1}{|C_k| \cdot \sigma(Q_k)} \cdot \left| \bigcup_{C_i \in S_k} C_i \cap C_k \right|. \quad (3)$$
Hierarchical hidden community detection for protein complex prediction

Figure 3: Benchmark results. The different bar clusters represent results for different datasets (Biogrid, Krogan_core and YeastNet) as the PPI networks. The bars on bottom indicate the scores of maximum matching ratio (MMR) and the bars on top represent the scores of recall. The number inside each bar is the score and higher bars represent better performance. (a) The results using CYC datasets as reference dataset. (b) The results using MIPS datasets as reference dataset.

$H(C_k)$ is the proportion of nodes in the community $C_k$ that belong to stronger communities. $Q_k$ is modularity of $C_k$ and ranges from [-0.5,1]. $\sigma$ represent the sigmoid function, which can proportionally change the modularity $Q_k$ to a positive number. The larger the hiddenness value is, the higher the probability that the community is hidden, and the less likely the community can be detected.

Comparisons on hierarchical communities and hidden communities

Hierarchical community structure and hidden community structure are two aspects of community structure. An algorithm combining these two concepts together can be very complicated. For example, Sometimes a subgraph with a hidden community and one or more stronger communities can be wrongly detected as hierarchical structure, Fig. 2 (a) and (b) show this case. So we first discuss the difference between hierarchical community structure and hidden community structure to clarify the conception. When considering hierarchical community structure, an algorithm gradually detects stronger and smaller communities, and ignoring weaker communities. When considering hidden community structure, an algorithm detects communities weaker than the communities that cover them. Intuitively, when the density of a community is calculated by modularity, if the average density of root communities is 0.5, then the average density of the next hierarchical level is larger than 0.5 and the average density of the next hidden layer is less than 0.5. Note that, a hidden community can be totally covered by one stronger community or be covered by several stronger communities. And it doesn’t need to have smaller size than the communities that cover it. Because of different characteristics of hidden and hierarchical structure, each layer of hidden structure can contain several levels of hierarchical structure, and vice versa (see Fig. 2(c)). HirHide combines the advantages of both types of algorithms, which guarantees a more complete protein complex detection.

**HirHide**

We propose a hierarchical hidden detection approach (HirHide) for community mining task. Our algorithm consists of three steps. In the first step of initialization, HirHide identifies a layer of communities via the base algorithm, which can be a traditional community detection algorithm with promising performance.

The second step is called the hierarchical step. In this step, HirHide constructs hierarchical structure by recursively capturing sub-communities and repeating until the appropriate number of levels are found. A crucial dimension of this
step is to determine the number of levels in a network. We simply count the number of nodes in each community. If the number is smaller than a certain threshold, then the algorithm stops capturing sub-communities. In our experiments, the default threshold is set to 9.

The third step can be called the hidden step. In this step, HirHide weakens the structure of the previously detected layer $L_1$ to get $G'$. In reduced graph $G'$, the base algorithm is used to detect new communities making up hidden layer $L_2$. Because a reduced graph $G'$ does not contain the strong communities, the weaker communities can be easily discovered. Then the algorithm that weakens $L_1$ and $L_2$ and calls the base algorithm to detect $L_3$. This process is repeated until no communities can be detected.

In HirHide, the step of digging out hidden structure is in a similar spirit with HiCode (He et al., 2018). But the step of combining hidden structure and hierarchical structure is new and well-designed to guarantee both hierarchical communities and hidden communities to be well captured. A key issue is which level of hierarchical communities should be weakened in the detected layer. When these hierarchical levels are organized into trees with each community in the first level as a root, we can weaken the communities in roots or the communities in leaves (level 1 and level 2 in Fig. 2 (c)). Because the edge connection condition and the total size of nodes in each hierarchical level are similar (considering separate nodes and too small communities are removed) and the weakening step is a global operation, which level is weakened does not make a big difference. To double-check, we analyzed the performance of HirHide framework when it separately weakens the communities in roots and in leaves in the confirmatory experiments.

In the hierarchical step, the sub-communities are captured on the original graph, but they can also be captured on the reduced graph $G'$ with weakening the structure of strong communities. In theory, the reduced graph has weakened the influence of other layers, so the sub-communities found should be more precise. In our confirmatory experiments, the results of recursively capturing the sub-communities on the original graph and on the weakened graph are compared. Although the base algorithm is recursively called in both the hierarchical step and the hidden step, its role is significantly different. In the hierarchical step, the base algorithm is repeatedly applied in sub-communities to discover smaller sub-communities. In the hidden step, the base algorithm is repeatedly used in the reduced graphs to dig out hidden communities.

### 3 Experiments

#### Algorithms, data and metric

**Algorithms**

We select three state-of-the-art algorithms as the baseline methods as well as the base algorithms of HirHide, which are MOD (Blondel et al., 2008), Infomap (Rosvall and Bergstrom, 2008), and LC (Ahn et al., 2010). None of them can detect hidden communities without HirHide. After they are combined with HirHide, we call them HirHide-MOD, HirHide-Infomap and HirHide-LC. To evaluate the function of HirHide, we compare the performance of HirHide-MOD, HirHide-Infomap and HirHide-LC with MOD, Infomap and LC in real-world networks.

MOD (Blondel et al., 2008) gradually optimizes the modularity value by iterations, and each node is treated as a separate community by initialization. Next, it tries to expand the size of the community. For each node $i$, consider all its neighbors $j$, adding node $j$ to the community to which node $i$ belongs, and calculating the change in the modularity value. If the newly added node $j$ leads to the value of modularity to become larger, then $j$ is added, otherwise, $j$ is not added. After iterating over and over, MOD can get a community division with the largest modularity value.

Infomap (Rosvall and Bergstrom, 2008) is based on the principle of information theory and defines the community from the perspective of coding. It is suitable for detecting almost all types of networks, including weighted and unweighted graphs, and directed or undirected graphs. After the algorithm regards the graph as a network, information generated during a random walk is defined as the average codeword length. In order to get the maximum compression ratio, the algorithm uses Huffman coding and community structure secondary coding. In this way, the problem of community detection is transformed into an optimization problem: finding a community division so that the codeword length of random code walks within and between communities is the smallest.

LC (Ahn et al., 2010) reinvents communities as groups of links rather than nodes. This approach successfully reconciles the antagonistic organizing principles of overlapping communities and hierarchy. Link communities naturally incorporate overlap while revealing hierarchical organization.
Hierarchical hidden community detection for protein complex prediction

Figure 4: Constructing hierarchial hidden community structure. When constructing a community, it is treated as a random block model and edges are added within the community. (a) Constructing the communities in root of the first layer. (b) Constructing the communities in leaves of the first layer. (c) The graph after randomly scrambling the node number with certain probability. (d) Constructing the communities in root of the second layer. (e) Constructing communities in leaves of the second layer.

Table 1: Basic information of the three PPI networks.

| Dataset  | MIPS | CYC |
|----------|------|-----|
|          | Nodes | Edges | Nodes | Edges |
| BioGRID  | 1155  | 10825 | 1355  | 13098 |
| YeastNet | 1085  | 11271 | 1289  | 13564 |
| Krogan   | 679   | 1797  | 910   | 2808  |

Data

We compare these algorithms in three large scale yeast PPI networks, which consist of core experimental yeast PPI network (Krogan et al., 2006), a combined computational interaction network (Kim et al., 2013) and the entire set of physical protein-protein interactions in yeast from BioGRID (Stark et al., 2006). These three datasets are referred to as the Krogan_core, YeastNet and BioGRID.

To evaluate the performance of each algorithm, we use Munich Information Center of Protein Sequences (MIPS) (Mewes et al., 2004) and CYC (Pu et al., 2008) as reference sets. The yeast protein complexes catalogued by MIPS database have been widely used to generate protein-protein interaction reference sets. And it is organized hierarchically. CYC is a comprehensive catalogue of manually curated 408 heteromeric protein complexes in S. cerevisiae (Pu et al., 2008). To avoid selection bias, we have considered all complexes containing 3 to 100 proteins as protein complexes. For convincing, we have chosen the latest version of the two reference sets.

Table 1 shows basic information of the three PPI networks. Because MIPS and CYC have different reference complexes, We have removed proteins and edges that sorely exist in a PPI network or a reference dataset.

Evaluation Metrics

It has become a standard practice to compare performance by assessing the method ability to identify reference communities. When judging the performance of a community detection algorithm, the recognized metric is $F^1$ score (Chase Lipton et al., 2014), which is the harmonic mean of Precision and Recall. However, the reference communities, which only contain protein complexes whose interactions are able to be discovered under the current experimental conditions, have incomplete nature (Li et al., 2010). Using $F^1$ score as an evaluation metric is unreasonable when reference datasets are incomplete. Because under the same condition, the more communities an algorithm detects, the smaller the Precision is. So traditional algorithms are typically designed to detect less communities than reference communities (much less than real-world communities) to increase the Precision. Consequently, when real-world networks contain more communities than reference communities, comparing the $F^1$ score is unfair for the algorithms that detect more communities.

By detecting additional hierarchical hidden communities, HirHide significantly increases the number of detected communities, which may more fit the reference sets if they are complete. However, $F^1$ score can’t show this advantage. For convincing, we further evaluate the performance of HirHide in synthetic data. In experiment of synthetic data, $F^1$ is chosen as evaluation metric because the reference set is complete.
Table 2: The performance of HirHide-MOD and MOD on synthetic data.

| Truth | HirHide-MOD | MOD | HirHide-MOD | MOD | HirHide-MOD | MOD |
|-------|-------------|-----|-------------|-----|-------------|-----|
|       | 360         | 3000| 2000        |     |             |     |
| $L_{11}$ | 1          | 1   | 1           | 0.999004 | 0.999503 | 0.980294 |
| $L_{12}$ | 0.997245   | 0.857143 | 0.859069 | 0.689618 | 0.686413 | 0.681019 |
| $L_{21}$ | 1          | 0.336932 | 1           | 0.078592 | 0.099517 |
| $L_{22}$ | 0.606493   | 0.215152 | 0.721043 | 0.474313 | 0.109028 |

In real-world data, as the current reference sets are incomplete (Li et al., 2010), we evaluate the performance of HirHide by two other measures. One is the maximum matching ratio (MMR), which is designed specifically for protein complex detection (Nepusz et al., 2012). Its rationality has been proven. The other is the Recall, which only measures the capacity of discovering reference complexes. So additional hierarchical hidden communities don’t decrease the score.

MMR guarantees that each detected community only matches one reference community and vice versa. And it maximizes the total score of one-to-one connections between predicted and reference complexes. This measure is inspired by bipartite graph maximum matching problem, in which the two sets of nodes respectively represent a detected complex and a reference complex. MMR try to find the best match of each reference complexes. So even if an algorithm detects more complexes, MMR won’t be reduced.

The second measure scores each pair composed of a predicted complex and reference complex by their similarity. Given a set of detected communities $D$ and a set of reference communities $G$. Each reference community $G_j$ has its individual recall:

$$R(G_j) = \max_{D_i \in D} \frac{|G_j \cap D_i|}{|G_j \cup D_i|}.$$  \hspace{1cm} (4)

The recall is defined as the average of $R(G_j)$ over all reference communities.

Experimental results on real-world data

We compare the performance of HirHide-MOD, HirHide-Infomap and HirHide-LC with MOD, Infomap and LC in real-world networks. Fig. 3 (a) and (b) separately shows the comparative performance of the six algorithms using the MIPS and CYC standard complexes as the reference sets. The different bar clusters represent using different datasets (Biogrid, Krogan_core and YeastNet) as PPI networks. The bars on bottom indicate the scores of maximum matching ratio (MMR) and the bars on top represent the scores of recall. The number inside each bar is its score and higher bars represent better performance. We can see that when benchmark algorithms are combined with HirHide, they achieve better MMR, recall and composite scores on most of the PPI datasets. As mentioned before, HirHide doesn’t change the core of a traditional algorithm but enables it to detect hierarchical hidden community structure. So the better performance proves that detecting hierarchical hidden community structure is helpful for detecting complexes in PPI networks.

Synthetic data and experimental results

To some extent the two evaluation metrics can solve the problem that reference sets are incomplete, but they are still lacking. The experimental results in real-world datasets may not be convincing enough. So the synthetic data and its experimental results is necessary for comparing these algorithms. The synthetic network contains two layers for hidden structure, wherein the first layer consists of strong communities and the second layer consists of relatively weak communities. In addition to the feature of multi-layer, the hierarchical structure are added. For each layer, smaller and denser communities are added to make up the next hierarchical level.

The choice of the edge-added probability follows a certain rule. According to the concept of hidden community, the interactions between communities in the hidden layer are relatively sparse, so the edge-added probability is also lower than that of the first layer. The edge-added probability of the leaf communities at the second level is slightly higher than the probability of the root communities, so as to further highlight the structure of sub-communities. At the first layer, 360 nodes are divided into 3 communities, each with 120 nodes. According to the definition of the community, the number of internal sides of the community will be dense, so edges are added to the inside of the three communities with probability $P_{11} = 0.2$, and no edge is added to the outside (Fig. 4 (a)). Then, adding the hierarchical structure. The 120 nodes of each community are subdivided into 2 sub-communities. 6 sub-communities make up the second hierarchical level, each of which contains 60 nodes. Edges are added to the inside of the 6 sub-communities with probability $P_{12} = 0.3$ (Fig. 4 (b)). So the first layer is successfully built.
Hierarchical hidden community detection for protein complex prediction

Figure 5: Two examples of predicted protein complexes identified by HirHide.

Table 3: Results of weakening the communities in roots and leaves.

| Truth | Weaken the communities in roots | Weaken the communities in leaves |
|-------|--------------------------------|---------------------------------|
| $L_{11}$ | 1 | 0.913043 |
| $L_{12}$ | 0.997245 | 0.691044 |
| $L_{21}$ | 1 | 0.911268 |
| $L_{22}$ | 0.606493 | 0.485275 |

According to the idea of hidden structure, before building the second layer, the node numbers in the first layer are randomly scrambled with certain probability, so that the previous community structure are evenly distributed in the graph (Fig. 4 (c)).

The second layer is constructed on the randomly scrambled graph. In the second layer, 360 nodes are divided into 4 communities, each of which has 90 nodes. Edges are added inside these communities with a probability of $P_{21} = 0.15$ (Fig. 4 (d)). Then continuing to add a hierarchical structure, each community is divided into 3 small sub-communities, each of which has 30 nodes. Edges are added inside these sub-communities with a probability of $P_{22} = 0.25$ (Fig. 4 (e)). Similarly, we also construct a graph with 3000 nodes and a graph with 2000 nodes.

Because MOD has the best performance in three benchmark algorithms, We compare HirHide-MOD with MOD on our synthetic data. The experimental results are showed in Table 2. $L_{11}$, $L_{12}$ represent dominant community layer and $L_{21}$, $L_{22}$ represent hidden community layer. As illustrated in Table 2 in all the three networks, HirHide-MOD has a slight advantage over MOD on $L_{11}$ and $L_{12}$, which means HirHide doesn’t reduce the performance of traditional algorithms in networks without hidden structure. Moreover, HirHide-MOD has significant advantage over MOD on $L_{21}$ and $L_{22}$. And the scores of MOD on $L_{21}$ and $L_{22}$ are very low while the scores of HirHide-MOD on $L_{21}$ and $L_{22}$ are normal and high. Consequently, we can conclude that MOD can not detect the hidden community layer. But after it is combined with the HirHide framework, HirHide-MOD can detect the hidden community layer well. These results are consistent with the results in real-world networks.

Confirmatory experiments

In HirHide framework, the strong community structure detected by the base algorithm needs to be weakened. After combining the concept of hierarchical structure, there is a choice that to weaken the communities at the top level of the hierarchical structure, or to weaken the community at the lowest level of the hierarchical structure. To determine which level we should choose, we experimented on the synthetic network with 3000 nodes. Because the synthetic data is complete, we have chosen $F_1$ score as the evaluation metric. As illustrated in Table 3 to weaken communities in roots can make HirHide have a better performance, especially in level 2.

In HirHide, there are two options when grabbing sub-communities in one layer: to grab the sub-communities on the original graph or to grab the sub-communities in the reduced graph after weakening the other layers. We still use $F_1$ score to evaluate the matching degree of detected communities with reference communities in the two cases. As illustrated in Table 4 in the graph with 3000 nodes, to grab sub-communities on the original graph has a slight advantage. But in the graph with 2000 nodes, it illustrates an opposite result. Consequently, overall which graph is chosen does not make a great difference.
Table 4: Results of grabbing sub-communities on the original graph and the reduced graph.

| Nodes | Original $L_{12}$ | Weaken graph | Original $L_{12}$ | Weaken graph |
|-------|-------------------|--------------|-------------------|--------------|
| 3000  | 0.859069          | 0.826370     | 0.721043          | 0.717033     |
| 2000  | 0.686413          | 0.707143     | 0.474313          | 0.479427     |

Screening predicted protein complexes by emerging patterns

HirHide-combined algorithm can detect additional hierarchical hidden communities that cannot be detected without HirHide. Some of them fit the characteristics of protein complexes with sparse internal connections. So they can serve as predicted protein complexes. Here, predicted protein complexes mean these protein complexes don’t appear in reference sets and they are potential to be protein complexes.

However, detecting these hierarchical hidden communities are solely based on density property in PPI networks. Not all of them are reliable enough. Emerging patterns (EPs) are conjunctive patterns that contrast sharply between different classes of data, which contain more informative properties such as degree statistics, clustering coefficient, topological coefficients and eigenvalues of a sub-graph. Recently, EPs are exploited to address the complex prediction problem [Liu et al., 2016]. We screen more reliable predicted protein complexes based on this EP-based method. A feature vector is first constructed to describe the key properties of the reference protein complexes as well as those of random non-complexes communities. Then to discover EPs by contrasting feature vectors of reference protein complexes and random non-complexes communities. Next, the discovered EPs are used to discover potential complexes. For each of the complexes predicted by HirHide, if it is similar with complexes discovered by EPs, it is considered as a reliable predicted complexes. We employ overlaps of nodes to define the similarity. Fig. 5 shows two examples of the predicted complexes.

4 Discussion

The main work of this paper is to propose a new method called HirHide and apply HirHide for protein complex detection and prediction. HirHide serves as a meta method that can be combined with existing traditional community detection algorithms to enable them to discover hierarchical hidden communities. We improve the definition on the necessity of detecting hidden structure and why detecting hierarchical hidden structure is more helpful for complex detection in PPI networks. We redefine hiddenness value to fit large-scale and complicated networks. And we have compared traditional algorithms with their HirHide versions in synthetic data and real-world PPI networks. Experimental results illustrate that the performance of traditional algorithms are boosted when combined with HirHide. In networks without hidden structure, HirHide doesn’t reduce their performance.

How to combine the concept of hierarchical structure with hidden structure is a key issue in our work. This paper verifies which level of the hierarchical structure is selected when iterating in HirHide. We conclude that weakening the communities at the root level can achieve the best weakening effectiveness and there is almost no difference to grab sub-communities on the original graph or the reduced graph. Finally, we treat additional hierarchical hidden communities as preliminary prediction of undiscovered protein complexes. And further screening these predicted complexes by emerging patterns. These steps guarantee that the predicted complexes are reliable.

Here we applied HirHide to detect protein complexes in PPI networks. Actually, HirHide can be used to most of community detection questions because of the flexibility of the base algorithms. HirHide are also suitable for questions related to other large-scale biological data such as neural networks and gene regulatory networks. Some of these biological networks are too complex for common community detection algorithms to produce a positive performance. Most of the time, a better choice is to choose a corresponding algorithm based on the characteristic of a certain biological network. Under these circumstances, HirHide has clear advantages because of its flexibility. At the same time, according to the natural hierarchical structure of cells, organelle, intracellular compound etc., hierarchical hidden community structure is in line with the characteristic of the data itself, thus helping researchers to study biological interactions more deeply.

References

Ahn, Y.-Y., Bagrow, J. P., and Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. nature, 466(7307), 761.
Hierarchical hidden community detection for protein complex prediction

A PREPRINT

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, **2008**(10), P10008.

Chase Lipton, Z., Elkan, C., and Narayanawamy, B. (2014). Thresholding classifiers to maximize F1 score. *arXiv preprint arXiv:1402.1892*.

Coscia, M., Rossetti, G., Giannotti, F., and Pedreschi, D. (2012). Demon: a local-first discovery method for overlapping communities. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 615–623. ACM.

Fiannaca, A., La Rosa, M., Urso, A., Rizzo, R., and Gaglio, S. (2013). A knowledge-based decision support system in bioinformatics: an application to protein complex extraction. *BMC bioinformatics*, **14**(1), S5.

Fraser, H. B. and Plotkin, J. B. (2007). Using protein complexes to predict phenotypic effects of gene mutation. *Genome biology*.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, **99**(12), 7821–7826.

Gregory, S. (2009). Finding overlapping communities using disjoint community detection algorithms. In *Complex networks*, pages 47–61. Springer.

He, K., Soundarajan, S., Cao, X., Hopcroft, J. E., and Huang, M. (2015). Revealing multiple layers of hidden community structure in networks. *CoRR, abs/1501.05700*.

He, K., Li, Y., Soundarajan, S., and Hopcroft, J. E. (2018). Hidden community detection in social networks. *Information Sciences*, **425**, 92–106.

Kim, H., Shin, J., Kim, E., Hwang, S., Shim, J. E., and Lee, I. (2013). Yeastnet v3: a public database of data-specific and integrated functional gene networks for saccharomyces cerevisiae. *Nucleic acids research*, **42**(D1), D731–D736.

King, A. D., Pržulj, N., and Jurisica, I. (2004). Protein complex prediction via cost-based clustering. *Bioinformatics*, **20**(17), 3013–3020.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., et al. (2006). Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, **440**(7084), 637.

Lage, K., Karlberg, E. O., Stirling, Z. M., Olsson, P. I., Pedersen, A. G., Rignano, G., Hinsby, A. M., Hirst, A. M., Turner, Z., Pociot, F., Tommerup, N., et al. (2007). A human phenotype-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology*, **25**(3), 309.

Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. (2011). Finding statistically significant communities in networks. *PloS one*, **6**(4), e18961.

Li, X., Wu, M., Kwoh, C.-K., and Ng, S.-K. (2010). Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC genomics*, **11**(1), S3.

Liu, Q., Song, J., and Li, J. (2016). Using contrast patterns between true complexes and random subgraphs in ppi networks to predict unknown protein complexes. *Scientific reports*, **6**, 21223.

Mewes, H.-W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Münsterkötter, M., Pagel, P., Strack, N., Stümpflen, V., et al. (2004). Mips: analysis and annotation of proteins from whole genomes. *Nucleic acids research*, **32**(suppl_1), D41–D44.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, **9**(5), 471.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, **45**(2), 167–256.

Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, pages 284–293. Springer.

Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S. J. (2008). Up-to-date catalogues of yeast protein complexes. *Nucleic acids research*, **37**(3), 825–831.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, **105**(4), 1118–1123.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein–protein interactions in yeast. *Nature biotechnology*, **18**(12), 1257.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic acids research*, **34**(suppl_1), D353–D359.

Stumpf, M. P., Wiuf, C., and May, R. M. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences*, **102**(12), 4221–4224.

Winterhalter, C., Nicolle, R., Louis, A., To, C., Radvanyi, F., and Elati, M. (2014). Pepper: cytoscape app for protein complex expansion using protein–protein interaction networks. *Bioinformatics*, **30**(23), 3419–3420.

Yang, J. and Leskovec, J. (2012). Community-affiliation graph model for overlapping network community detection. In *2012 IEEE 12th International Conference on Data Mining*, pages 1170–1175. IEEE.

Yang, P., Li, X., Wu, M., Kwoh, C.-K., and Ng, S.-K. (2011). Infering gene-phenotype associations via global protein complex network propagation. *PloS one*, **6**(7), e21502.

Yu, L., Huang, J., Ma, Z., Zhang, J., Zou, Y., and Gao, L. (2015). Infering drug-disease associations based on known protein complexes. *BMC medical genomics*, **8**(2), S2.

Zhang, S., Wang, R.-S., and Zhang, X.-S. (2007). Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications*, **374**(1), 483–490.