Supportness of the protein complex standards in PPI networks

Milana Grbić¹, Vukašin Crnogorac¹, Milan Predojević¹, Aleksandar Kartelj² and Dragan Matić¹

¹Faculty of Natural Science and Mathematics, University of Banja Luka, Banja Luka, Bosnia and Herzegovina; ²Faculty of Mathematics, University of Belgrade, Beograd, Serbia

ABSTRACT
A protein complex is a collection of two or more associated proteins that interact with each other in a stable long-term interaction. Protein complexes have essential roles in regulatory processes, cellular functions and signaling cascades. This paper examines how well-known collections of protein complexes are supported in protein–protein interaction (PPI) networks, i.e. whether they form connected subnetworks in a particular PPI network. For that purpose, we apply a variable neighbourhood search (VNS) metaheuristic algorithm for adding the minimum number of interactions in order to support protein complexes. Experimental results obtained on several PPI networks (BioGRID, WI-PHI and String) and four protein complex standards (MIPS, TAP06, SGD and CYC2008) show that considered networks do not include enough PPIs to support all complexes from complex standards. Deeper analysis indicates that there exists common PPIs which are probably missing in the considered networks. These findings can be useful for further biological interpretation and developing of PPI prediction models.

ARTICLE HISTORY
Received 24 August 2021
Accepted 4 September 2021

KEYWORDS
Variable neighbourhood search; PPI networks; protein complexes; network connectivity

1. Introduction

In recent years, many sophisticated computing methods are developed in order to enable easier processing of biological data. Since biological networks can be considered as mathematical structures – graphs, a lot of problems considered on biological networks can be represented and resolved as computational and mathematical problems on graphs. Nodes in such biological networks are biological elements (like proteins, genes, metabolites) while the edge between two nodes exists if there is some dependency between them, like physical interaction or participating in a particular biochemical process. Biological networks often have a lot of nodes and edges (several thousands or even more). Therefore, identifying functionally related elements or partitioning biological networks into smaller subnetworks is an often used approach for analyzing such networks (Grbić,
Kartelj et al., 2020; Grbić, Matić et al., 2020; Hüffner et al., 2013; Liu et al., 2009; Martins, 2016).

In this paper, we analyze networks which represent interactions between proteins of a particular organism. In such a network, proteins are represented as nodes, while an edge between two proteins exists if a certain kind of interaction between these two proteins is proved or assumed. These interactions are called protein–protein interactions (PPIs) and networks consisting of them -- PPI networks. In this research, we are focused on PPI networks of the *Saccharomyces cerevisiae* organism.

A group of two or more associated proteins is called protein complex. In other words, protein complexes consist of at least two proteins in a stable long-term interaction. In literature, there exists several well-known sets of protein complexes, called gold standards, which are usually used as a reference for prediction models construction or evaluation (Browne et al., 2009). In this research, we consider the following gold standards for *S. cerevisiae* protein complexes: Munich Information Center of Protein Sequences database (MIPS) (Mewes et al., 2002), TAP06 set (Gavin et al., 2006), SGD set (Hong et al., 2007) and the CYC2008 set (Pu et al., 2009). In MIPS catalogue, techniques such gene disruption in conjunction with expression analysis generate information how proteins corporate in complexes and that information are saved (Mewes et al., 2002). TAP06 set of protein complexes is actually a core set of TAP-MS experiment (Tandem-Affinity-Purification method coupled to mass spectrometry) (Liu et al., 2016). SGD set is derived from the Saccharomyces Genome Database (Hong et al., 2007). CYC2008 set of protein complexes contains 408 manually curated heteromeric protein complexes reliably backed by small-scale experiments reported in the literature (Pu et al., 2009).

It is common that proteins which belong to the same complex have a physical interaction. So, commonly used approach for identification significant protein groups, like complexes, is based on finding and considering dense regions or highly connected sub-networks in PPI networks (Liu et al., 2009; Zhang et al., 2018). Nevertheless, Nakajima et al. (2018) show that, in some PPI networks, protein complexes are not densely connected, even not connected at all. From previous, one can conclude that some of significant PPIs are still missing in PPI networks. In the same paper (Nakajima et al., 2018), the problem of finding the minimum number of PPIs in order to connect (support) known protein complexes is introduced.

Following that idea, in this paper we continue the study presented by Grbić, Matić et al. (2020): we consider several gold standard sets of protein complexes and examine how they are supported in various PPI networks. In that sense, we add a minimum number of PPIs in PPI network to connect nodes (proteins) which participate in complexes and analyse the obtained results.

Beside the theoretical explanation for ‘minimality requirement’, usually known as Minimum description length (MDL) principle (Rissanen, 1978), adding minimum number of additional PPIs in a network has its practical explanation in this research. In a deeper analysis shown in Subsection 4.2, we investigate whether certain PPIs are added into more than one network. Even with this minimality requirement, the algorithm succeeds to identify some common PPIs, probably missing in the considered networks.

Starting from the important PPI entities--protein complexes, we found out that they are not equally supported in different PPI networks, even in different versions of the same PPI network. Consequently, it can be assumed that some of PPIs, identified by our algorithm,
will appear in newer versions of the considered PPI networks. This can be useful for biologists to focus on the suggested PPIs and to experimentally check their existence.

The main contribution of this research can be summarized as follows:

- the variable neighbourhood search (VNS) method from Grbić, Matić et al. (2020) is applied on several PPI networks and four gold standards to add a minimum number of PPIs to support protein complexes from gold standard;
- performed deep analysis of the obtained results with respect to each gold standard indicates that there exists common PPIs which are probably missing in the considered networks;
- the new findings of the proposed method can be useful for developing PPI prediction models.

A preliminary version of this paper appeared in the Proceedings of International Conference on INnovations in Intelligent SysTems and Applications (INISTA) 2020 (Grbić, Crnogorac et al., 2020).

In the extended version, we perform a deeper analysis of the obtained results with respect to each of four gold complex standards. The aim of such an analysis is to identify concrete newly added PPIs, with the property that they are common for at least two of three considered networks.

2. Previous work

The problem of adding minimum number of PPIs in order to support each complex from a complex set is introduced by Nakajima et al. (2018). The problem, which is called MinPPI, is defined as follows: Let a set of complexes and a PPI network be given. The task of MinPPI problem is to find a set of additional PPIs of minimum cardinality, which should be added to the starting network to connect each disconnected complex. This problem is proven to be NP hard (Angluin et al., 2015; Chockler et al., 2007). Nakajima et al. (2018) proposed an integer linear programming (ILP) model and a greedy heuristic approach for solving MinPPI. In a recent work, Grbić, Matić et al. (2020) developed a Variable Neighbourhood Search (VNS) algorithm for solving this problem on both unweighted and weighted PPI networks. Newly added PPIs, which support all complexes, are further used as a base for identifying functionally related protein groups which have high statistical significance.

3. The method

3.1. Problem definition

Formal mathematical definition of the problem is as follows.

Let \( G = (V, E) \) be a graph with the set of nodes \( V \) and the set of edges \( E \). Let \( S = \{S_1, S_2, \ldots, S_c\} \) be a collection of subsets of \( V \). The task of MinPPI problem is to find the set of new edges \( E' \) of minimum cardinality such that, for every \( i, i \in \{1, 2, \ldots, c\} \), the subgraph induced by \( S_i \) in \( G = (V, E \cup E') \) is connected. The
special case of MinPPI problem is the so called MinPPI0 problem, in which the initial graph \( G \) consists only of singletons, i.e. when \( E = \emptyset \).

If we observe the problem from biological point of view, the graph \( G \) is a PPI network, the set of nodes \( V \) is a set of proteins from that network, while \( E \) is the set of PPIs. The collection \( S = \{S_1, S_2, \ldots, S_c\} \) is the collection of protein complexes (gold standard set), where every \( S_i \) is a particular protein complex. The set \( E' \) will contain new PPIs.

As it is said, the MinPPI problem is NP hard, so it is justified to solve this problem by approximate methods.

3.2. Preparation phase

PPI networks contain many proteins, but not all the proteins are present in complexes. In that order, the input to the algorithm is not the original PPI network, but a reduced network which contains only proteins from the given set of complexes. Therefore, before the VNS algorithm is applied, a simple procedure which extracts only those PPIs, whose both ends are proteins from the considered set of complexes is performed. In Subsection 4.1, one can find more information about the original and the reduced networks used in this research.

3.3. VNS for solving MinPPI problem

Mladenović and Hansen (1997) introduced the well-known metaheuristic optimization approach – VNS algorithm. Standard VNS usually imposes two main procedures, called Shaking and Local search (LS). In Shaking, the system of neighbourhoods around the current solution is being formed and a new solution from the current neighbourhood is proposed. LS procedure tries to improve that solution by examining solutions in its neighbourhood, looking for the local optimum. The efficiency of the VNS algorithm is based on an assumption that there is a correlation between all these local optima and that one of them is also the global one.

As it is mentioned, in this research we use the VNS for solving MinPPI, which is proposed by Grbić, Matić et al. (2020). In addition to the above-mentioned two standard VNS procedures, an additional procedure, called FixGreedy is introduced. The purpose of the latter procedure is to make solution feasible before passing it to LS procedure. In the overall algorithm, these three procedures are repeating until one of the following conditions is not satisfied: reached maximum time for execution, reached total maximum number of iterations or reached maximum number of iterations without any improvement.

In the rest of this section, we briefly describe main parts of the proposed VNS, while more detailed description can be found in Grbić, Matić et al. (2020). The pseudocode of the overall VNS algorithm is shown in Figure 1.

The proposed VNS takes the following parameters regarding the problem instance: \( \nu_{cnt} \) – total number of proteins in PPI network; \( \text{complexes} \) – list of complexes from a complex standard and \( \text{fixedEdges} \) – list of PPIs which already exist in the considered network. The control parameters of VNS are: \( k_{\text{min}} \) and \( k_{\text{max}} \) – minimal and maximal size of the VNS neighbourhood structure; \( \text{it}_{\text{max}} \) and \( \text{itrep}_{\text{max}} \) – maximal total number of allowed iterations and maximal number of allowed iterations without improvement; \( t_{\text{max}} \) – maximal time in
3.3.1. Initialization and objective function

Let us denote with $n$ the number of proteins in the network. The VNS solution is represented by the binary matrix $X$ of the dimension $n$. If the element on the position $ij$ is equal to 1, the corresponding PPI between the proteins $i$ and $j$ is included in solution, otherwise not.

Let us denote the set of PPIs in the starting network with $fixedEdges$ and let with $Sol$ be denoted the VNS solution. Then, the objective function, which is minimized, is defined in Equation (1):

$$Obj(Sol) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} n(1 - 1_{fixedEdges}(e_{ij}))X[i,j]$$

where $e_{ij}$ is the PPI between proteins $i$ and $j$ and $1_{fixedEdges}$ is the indicator function of the set $fixedEdges$.

From Equation (1), one can observe that the objective function is equal to the number of newly added PPIs.

3.3.2. Shaking procedure

Shaking procedure is used to avoid the situations when algorithm is being stuck in a local suboptimum.
The \( k \)th neighbourhood is formed as follows: the algorithm randomly chooses some \( k \) newly added PPIs and remove them from the solution. Such new solution is the subject of further analysis in the fixGreedy procedure.

### 3.3.3. FixGreedy procedure
A fast procedure, based on a greedy approach, called fixGreedy is constructed to resolve the problem of appearance of infeasible solutions before the LS starts. The procedure evaluates the benefit of adding the PPIs in a greedy manner, by favouring those PPIs which connects more complexes, i.e. those PPIs which decrease the number of disconnected complexes more than other PPIs.

### 3.3.4. Local search
After the Shaking and the fixGreedy procedures are applied, the solution is further locally improved in Local search procedure. LS is based on examining the possibility of replacing a PPI from the solution by another one, or, eventually removing it from the solution without any replacement. Inside the LS, the algorithm performs a fast partial calculation of the objective function applied on the locally formed solution.

The algorithm uses ‘the first improvement strategy’, which means that the current best solution is replaced by the first improved solution found in the LS.

### 4. Experimental results
All experiments have been carried out on Intel Xeon CPU E5410 @ 2.33GHz with 16GB RAM, under the Windows 7 64Bit operating system. For each execution only one thread/core is used. The VNS algorithm is coded in C programming language. The algorithm is executed 10 times for each test PPI network. The termination parameters are:

- maximum time for execution \( t_{\text{max}} = 1200 \);  
- maximum number of iterations \( l_{\text{max}} = 20,000 \);  
- maximum number of iterations without any improvement \( itr_{\text{rep}}_{\text{max}} = 5000 \).

Other control parameters are:

- minimum cardinality of a neighbourhood, \( k_{\text{min}} = 1 \);  
- maximum cardinality of a neighbourhood, \( k_{\text{max}} = 5 \);  
- probability to move to the other solution with the same value of objective function, \( prob=0.5 \).

### 4.1. Datasets
In this research, we used the following PPI networks:

1. BioGRID from Nepusz et al. (2012);
2. WI-PHI from Kiemer et al. (2007);
3. Four different versions of String database: v9.1 (Franceschini et al., 2012), v.10 (Szklarczyk et al., 2015), v10.5 (Szklarczyk et al., 2016) and v.11 (Szklarczyk et al., 2019).
The purpose of usage of different versions of the String network is to examine how well the standards are supported in the consecutive versions, i.e. to examine if a standard is better supported in the newer versions of the same network.

Table 1 contains information about each PPI network used in this paper. The name of a network is shown in the first column. The second column contains the number of PPIs in the original network, before the reduction is performed in the preparation phase. The number of PPIs in each network after the reduction with respect to MIPS standard is shown in the third column. The number of PPIs after the reduction with respect to TAP06, SGD and CYC2008 standards are shown in the fourth, fifth and sixth columns, respectively.

In Table 2, we show the basic information about each gold standard. The first column contains the name of the standard, while in the second and the third column number of proteins and number of complexes in the given set are shown, respectively.

### 4.2. Results

This subsection presents the results obtained by executing the algorithm on each network for each of four gold standard sets of complexes: MIPS, TAP06, SGD and CYC2008, respectively. For each complex standard, we investigated whether certain PPIs are added into more than one network. For this purpose, we observed the BioGRID network, WIPHI network and the newest version of String network -- String v.11 and analysed the sets of added PPIs, with respect to the considered four gold standards. Supplementary material containing original data and executable version of the proposed VNS is available at public Git repository [https://github.com/milanagrbic/SupportnessOfPPIComplexes](https://github.com/milanagrbic/SupportnessOfPPIComplexes).

#### 4.2.1. Results on MIPS complex standard

In Table 3, results for MIPS set of protein complexes are shown. In the first three columns of the table, the name of the network, total number of nodes and total number of edges reduced to the gold standard are shown. The next three columns contain the best and average results of the VNS (columns best and avg) obtained in 10 runs and the average

| Network   | #PPI  | #red. to MIPS | #red. to TAP06 | #red. to SGD | #red. to CYC2008 |
|-----------|-------|---------------|----------------|-------------|-----------------|
| BioGRID   | 59,748| 10,825        | 6263           | 10,230      | 16,180          |
| WI-PHI    | 50,000| 8168          | 6873           | 8264        | 12,848          |
| String v.9.1 | 1,555,178 | 81,474   | 39,983         | 74,303      | 124,791         |
| String v.10 | 1,879,996 | 95,239   | 46,537         | 86,958      | 291,530         |
| String v.10.5 | 2,007,134 | 99,445   | 47,954         | 88,743      | 148,357         |
| String v.11 | 1,845,966 | 102,190  | 48,410         | 88,209      | 153,029         |

| Set       | #proteins | #complexes |
|-----------|-----------|------------|
| MIPS      | 1189      | 203        |
| TAP06     | 824       | 193        |
| SGD       | 1279      | 323        |
| CYC2008   | 1627      | 408        |
time needed to obtain the best result (column time[s]). The result represents the minimum number of PPIs which should be added to network in order to connect all complexes from the observed set of protein complexes.

In order to support all MIPS complexes, VNS found that new 174 PPIs should be added into the BioGRID network, 189 new PPIs should be added to support complexes in the WIPHI network. This standard is well supported in each String network, since the total number of additional PPIs is rather small. For the String v.10 network, the smallest number of new PPIs is needed to support MIPS standard: only 13. Although the String v.11 network contains larger number of PPIs comparing to the previous versions of this dataset, complexes from MIPS standard are not better supported in the newest version.

A deeper insight in the obtained results shows that there is a significant number of common PPIs, as it is listed below:

- WI-PHI and BioGRID: 129 common PPIs;
- WI-PHI and String v.11: 11 common PPIs;
- BioGRID and String v.11: 11 common PPIs.

Further, all 11 common PPIs, which appears in the intersections between the String v.11 network with other two networks are the same. So, we conclude that there exists 11 PPIs, which are common in all three networks. This fact could be an indication that these PPIs are missing in the latest available versions of these networks.

The list of these common newly added PPIs in this three networks is provided below:

- YML102W-YDL077C
- YHR156C-YGL005C
- YBL078C-YNL126W
- YMR198W-YGL005C
- YML102W-YKL119C
- YNL088W-YGL249W
- YPL129W-YER179W
- YMR302C-YNL221C
- YIL009C-A-RNA_TLC1
- YNL221C-SNRNA_NME1
- YHR039C-A-YDL185W1

Based on this list, additional conclusions can be drawn, in the sense how these new PPIs reflect to the complexes. For that purpose, we decided to analyze one characteristic
complex from MIPS standard set, which was initially unsupported (disconnected) in starting networks, but connected after the new PPIs are added.

We consider the protein complex containing the following 14 proteins: YGL216W, YPL155C, YEL061C, YBL063W, YDR424C, YKR054C, YPR141C, YHR129C, YPL174C, YDR488C, YMR294W, YHR156C, YGL005C and YMR198W. This complex was initially disconnected in all three networks.

Graph in Figure 2 shows the considered complex after adding new PPIs in BioGRID network. Hereinafter, black edges will represent PPIs which already exist in considered network, while red edges will represent new PPIs added by VNS algorithm. As it can be seen from Figure 2, this complex is not well supported in BioGRID network. There exist only 8 PPIs between these proteins, while seven new PPIs are added by VNS in order to connect this complex. Also, it can be concluded that before adding PPIs by VNS algorithm, this complex in BioGRID network has one connected subgraph containing six proteins, one edge (YPR141C – YMR198W) and six proteins which were singletons.

In Figure 3, the same complex is shown after adding PPIs in WI-PHI network. Colours of the edges have the same meaning as in Figure 2. This complex is also not well supported in WI-PHI network: only six PPIs exist in the network (coloured in black) and new nine PPIs are added by VNS algorithm (coloured in red). Before adding the new PPIs, this complex contained one larger connected subcomponent, one isolated PPI and seven isolated proteins (singletons in the network). In both networks (BioGRID and WI-PHI), the most of the singletons are connected by newly added PPIs ending with the protein YGL005C. This protein is further connected with the rest of the complex by a new PPI ending with the protein YMR198W. All PPIs added to support this complex into the BioGRID network are also added into the WI-PHI network.

**Figure 2.** A supported complex from MIPS standard after adding new PPIs in BioGRID network.
Figure 4 shows the graph which represents this complex in the String v.11 network, after new PPIs are added by VNS algorithm. Two red edges (YHR156C–YGL005C and YMR198W–YGL005C) are two new PPIs added by VNS, while the rest of 63 black edges already exists in this network. From Figure 4, it is evident that 12/14 proteins form a dense component in the networks, while two proteins are singletons in the starting network. Again, as in two previous cases, after the VNS algorithm is applied, YHR156C is connected with YGL005C, which is further connected with rest of the complex by protein YMR198W.

As it can be seen from Figures 2–4, VNS found that two PPIs: YHR156C–YGL005C and YMR198W–YGL005C are added into each of three networks in order to support the considered complex.

### 4.2.2. Results on TAP06 gold standard

Table 4 contains the results obtained for TAP06 standard. The organization of the table is similar as in the case of MIPS standard. From the shown data, it can be concluded that this standard is better supported than MIPS. The total number of added PPIs is up to 16 for

| Network  | #nodes | #PPIs | best | avg | time[s] |
|----------|--------|-------|------|-----|---------|
| BioGRID  | 824    | 6263  | 56   | 56  | 98.8627 |
| WI-PHI   | 824    | 6878  | 4    | 4   | 89.8189 |
| String v.9.1 | 824  | 39,983 | 12   | 12  | 139.7874 |
| String v.10 | 824    | 46,537 | 13   | 13  | 106.6824 |
| String v.10.5 | 824    | 47,954 | 15   | 15  | 141.5355 |
| String v.11 | 824    | 48,410 | 16   | 16  | 101.3936 |
each network, with the exception of the BioGRID network, where 56 new PPIs are added. The smallest number of added PPIs is for WI-PHI network, only 4. From Table 4, one can see that even more PPIs are added in newer versions of the String dataset.

Since the number of newly added PPIs in WI-PHI network is small (only 4), the number of common PPIs, we could not expect a large number of PPIs common to all three networks. More precisely, the situation is as follows.

- all three considered networks: one common PPI (interaction between proteins YER095W and YDR471W);
- WI-PHI and BioGRID: 2 common PPIs;
- WI-PHI and String v.11: 1 common PPIs;
- BioGRID and String v.11: 14 common PPIs.

It is interesting that protein complex containing proteins YER095W, YDR471W and YML075C was not supported in each of considered networks. More precisely, the situation with this complex was the same in each network: there existed the PPI between YER095W and YML075C, but the protein YDR471W was not connected with any of other two proteins. Figure 5 shows the situation before and after the algorithm is applied.

As it is already mentioned, 14 new PPIs are common for BioGRID and String v.11 networks. These PPIs are listed below:
These PPIs connect many complexes. From them, we illustrate some characteristic cases. PPIs $\text{YNR043W} - \text{YGR034W}$ and $\text{YNR043W} - \text{YDR365C}$ support four different complexes (we denote them as $\text{C}_0^1\text{TAP}$, $\text{C}_0^2\text{TAP}$, $\text{C}_0^3\text{TAP}$ and $\text{C}_0^4\text{TAP}$), containing the following proteins:

- $\text{C}_0^1\text{TAP}$: $\text{YHR169W}$ $\text{YDR365C}$ $\text{YNR043W}$
- $\text{C}_0^2\text{TAP}$: $\text{YDR365C}$ $\text{YCL059C}$ $\text{YNR043W}$ $\text{YDL213C}$ $\text{YGR034W}$ $\text{YGR145W}$
- $\text{C}_0^3\text{TAP}$: $\text{YHR169W}$ $\text{YDR365C}$ $\text{YJL033W}$ $\text{YNR043W}$
- $\text{C}_0^4\text{TAP}$: $\text{YPL217C}$ $\text{YOR078W}$ $\text{YDR365C}$ $\text{YNR043W}$ $\text{YDL213C}$

We graphically show these complexes in Figures 6–9, respectively. Each figure consists of two parts: part (a) considered complex in BioGRID network and part (b) considered complex in String v.11 network. Recall that black edges represent existing PPIs while red edges denote newly added PPIs. From Figure 6(a), one can see that complex $\text{C}_0^1\text{TAP}$ was completely unsupported in BioGRID network, since all three proteins were
isolated. On the other side, in String v.11 network there existed a PPI between proteins YHR169W and YDR365C (coloured in black in Figure 6(b)).

From Figure 7(a), one can see that the algorithm added three new PPIs, which seems unnecessary, since only two of them are enough to support the complex $C_{02}^{TAP}$. The reason for adding the third PPI lies in the unconnectedness of the complex $C_{04}^{TAP}$ in BioGRID network (Figure 9(a)). From the last mentioned figure, one can see that two proteins (YDR365C and YNR043W) were isolated in complex $C_{04}^{TAP}$, so at least two PPIs had to be added to support that complex. In order to connect the complex, the algorithm added YNR043W–YDL213C and YNR043W–YDR365C PPIs, which are shown both in Figures 7(a) and 9(a). It also could be noted that all of the considered complexes are better supported in String v.11 network.

**Figure 6.** Complex $C_{01}^{TAP}$. (a) BioGRID network and (b) String v.11 network.

**Figure 7.** Complex $C_{02}^{TAP}$. (a) BioGRID network and (b) String v.11 network.
4.2.3. Results on SGD gold standard

In Table 5, the results obtained for SGD set of proteins are shown. Similarly as for MIPS standard, SGD standard is well supported in String networks. In WI-PHI network, the largest number of PPIs is added.

More precisely, in order to support all complexes from SGD standard, the proposed VNS algorithm added 95 PPIs in the BioGRID network, 138 PPIs in the WI-PHI network and 8 PPIs in String v.11 network (Table 5).

The analysis shows that there is a significant number of common PPIs, as it is listed below:

- WI-PHI and BioGRID: 76 common PPIs;
- WI-PHI and String v.11: 7 common PPIs;
- BioGRID and String v.11: 8 common PPIs.

The following 7 PPIs are common newly added to all three networks:

**Figure 8.** Complex C03\_TAP. (a) BioGRID network and (b) String v.11 network.

**Figure 9.** Complex C04\_TAP. (a) BioGRID network and (b) String v.11 network.

### 4.2.3. Results on SGD gold standard

In Table 5, the results obtained for SGD set of proteins are shown. Similarly as for MIPS standard, SGD standard is well supported in String networks. In WI-PHI network, the largest number of PPIs is added.

More precisely, in order to support all complexes from SGD standard, the proposed VNS algorithm added 95 PPIs in the BioGRID network, 138 PPIs in the WI-PHI network and 8 PPIs in String v.11 network (Table 5).

The analysis shows that there is a significant number of common PPIs, as it is listed below:

- WI-PHI and BioGRID: 76 common PPIs;
- WI-PHI and String v.11: 7 common PPIs;
- BioGRID and String v.11: 8 common PPIs.

The following 7 PPIs are common newly added to all three networks:
We will analyze some characteristic cases. The PPI $\text{YNL311C-YJL204C}$ appears in three complexes of SGD standard. It is interesting that this PPI forms one separate complex (denoted as $C_0^{1\text{SGD}}$) and it was not supported in any of three networks (Figure 10(a)). Further, this PPI also appears in a complex $C_0^{2\text{SGD}}$, containing 12 proteins.

| instance     | #nodes | #PPIs   | best | avg | time[s] |
|--------------|--------|---------|------|-----|---------|
| BioGRID      | 1279   | 10,230  | 95   | 95  | 290.9104|
| WI-PHI       | 1279   | 8264    | 138  | 138 | 403.904 |
| String v.9.1 | 1279   | 74,303  | 3    | 3   | 275.4976|
| String v.10  | 1279   | 86,958  | 6    | 6   | 256.12  |
| String v.10.5| 1279   | 88,743  | 8    | 8   | 315.3847|
| String v.11  | 1279   | 88,209  | 8    | 8   | 244.9951|

- YER012W–YLR148W
- YDR130C–YML104C
- YNL278W–YDR252W
- YLR186W–YML124C
- YHR052W–YHR200W
- YGR097W–YGR252W
- YNL311C–YJL204C

Figure 10. Graphical view of the complex $C_0^{1\text{SGD}}$. The complex was not supported in any of three networks before the VNS algorithm is applied. (a) Before adding new PPIs and (b) After adding new PPIs.

Figure 11. Complex $C_0^{2\text{SGD}}$. (a) BioGRID network. (b) String v.11 network and (c) WI-PHI network.
In Figure 11(a–c), we show this complex in all three networks. From figures, we can see that BioGRID network has four old PPIs, while seven PPIs are newly added. Similar situation is with WI-PHI network: four PPIs, connecting proteins from this complex already existed, while seven new PPIs are added. 6/7 newly added PPIs in these two networks are common. On the other side, String v.11 network completely supports this complex, which can also be seen from Figure 11(b). In Figure 11(b) we see one red coloured PPI which is newly added, not to support proteins in this complex but to support complex CO15SGD. The third complex containing the mentioned PPI YNL311C–YJL204C was already supported in all three networks.

4.2.4. Results on CYC2008 gold standard

Table 6 contains the results on CYC2008 set of protein complexes. Since the results for all String networks are equal to zero, it can be concluded that complexes from CYC2008 standard are completely supported in all considered versions of the String networks.

In BioGRID new 67 PPIs are added, while in WI-PHI network, the number of added PPIs is larger (93). The following 44 newly added PPIs are common in these two networks:

- YER056C–A–YLR075W
- YLR061W–YLR075W
- YKR094C–YLR075W
- YHR021C–YMR116C
- YDR086C–YER019C–A
- YJL189W–YLR075W
- YLR167W–YMR116C
- YGR008C–YDL181W
- YOR369C–YMR116C
- YBR283C–YDR086C
- YCL037C–YMR186W
- YPL079W–YLR075W
- YDL184C–YLR075W
- YOR182C–YMR116C
- YPL029W–YMR287C
- YYR049W–YDR347W
- YDL061C–YMR116C
- YKR065C–YJL104W
- YKL156W–YMR116C
- YPL148W–YLR075W
- YPR132W–YMR116C

Table 6. Results obtained for CYC2008 set of protein complexes.

| Network     | #nodes | #PPIs | best | avg  | time[s]  |
|-------------|--------|-------|------|------|----------|
| BioGRID     | 1627   | 16,180| 67   | 67   | 1200.5106|
| WI-PHI      | 1627   | 12,848| 93   | 93   | 1200.44390|
| String v.9.1| 1627   | 124,791| 0    | 0    | 350.8771  |
| String v.10 | 1627   | 291,530| 0    | 0    | 426.7501  |
| String v.10.5| 1627   | 148,357| 0    | 0    | 352.6209  |
| String v.11 | 1627   | 153,029| 0    | 0    | 354.3587  |
We analyze one characteristic case and consider the complex **cytoplasmic ribosomal small subunit** with 57 proteins, shown in Figure 12 (BioGRID network) and Figure 13
**Figure 13.** Complex cytoplasmic ribosomal small subunit connected after adding new PPIs in WI-PHI network.

**Figure 14.** PPIs of protein YIL069C in BioGRID network.
(WI-PHI network). From the figures, one can see that many proteins belonging to this complex were isolated before the algorithm was applied. The algorithm added 15 new PPIs in BioGRID and 16 to WI-PHI network. From Figure 12 one can see that all 15 these PPIs are connected to YMR116C protein. From Figure 13 one can see that all 16 newly added PPIs are connected to the same protein YMR116C. We note that nine PPIs are common for both networks. Recall that this protein complex (as is the case with all other complexes from CYC2008 standard) was supported in the String network.

The rest of 7/16 PPIs, which are newly added to WI-PHI network, now connects the protein YMR116C to the proteins: YER102W, YGR027C, YGR118W, YIL069C, YLR287C-A, YML024W, YNL302C. These proteins had not any interaction inside the considered complex in the started network. Contrary, the situation is pretty different in BioGRID network, since these proteins already have many interactions with other proteins. For example, protein YIL069C has 24 interactions with other proteins in Biogrid network inside the considered complex (Figure 14).

Regarding to the rest of 6/15 PPIs newly added to the BioGRID network, the situation is as follows: YMR116C now interacts with YGL189C, YJL136C, YKR057W, YLR388W, YMR143W and YOR293W inside the considered complex. Interestingly, in WI-PHI network, many of these proteins interact with other proteins from the complex. Such a difference between the (non)existence of the PPIs in two networks is in line with the assumption that there probably exist false positive and false negative PPIs.

5. Conclusion

In this research, we applied the VNS algorithm for solving MinPPI problem, i.e. the problem of adding minimum PPIs in network in order to connect each complex from a collection of protein complexes. We tested our algorithm on six PPI networks and four collections of protein complexes, which are used in literature as gold standards. Obtained results show that proposed VNS algorithm is very successful in solving the given problem for all tested instances. For all networks and standards, VNS obtained the same results in all 10 executions.

In a deep analysis of the obtained results, we considered newly added PPIs for each of four complex standards. We found that the algorithm identified many common PPIs which are needed to support complexes in at least two of three considered networks. The analysis is supported by graphical representation of the characteristic complexes. The appearance of such PPIs indicates that these PPIs are probably missing in the networks and should be considered in development of PPI prediction models.

In future work, it would be interesting to include additional biologically relevant information about relations between proteins and their roles in common biological processes. Such information can be incorporated in the PPI network as PPI weights or as an additional information on network nodes (proteins).

Disclosure statement

No potential conflict of interest was reported by the authors.
Funding

This research is partially supported by Ministry for Scientific and Technological Development, Higher Education and Information Society, Government of Republic of Srpska, B&H under the Project ‘Development of artificial intelligence methods for solving computer biology problems’

Notes on contributors

M. Grbić received her MSc degree (2016) in Mathematics and Ph.D. in Computer Science (2020) at the Faculty of Mathematics at the University of Belgrade, Serbia. Her research interests include combinatorial optimization and data mining in bioinformatics.

V. Crnogorac is a master student in the field of Computer Science and Information Technology at the University of Novi Sad. He is working as a teaching assistant at the Faculty of Natural Science and Mathematics at the University of Banja Luka. His field of research interest include machine learning and combinatorial optimization.

M. Predojević finished his master studies in the field of Computer Science and Information Technology at the University of Novi Sad. He is working as a teaching assistant at the Faculty of Natural Science and Mathematics at the University of Banja Luka. His field of research interest include machine learning and combinatorial optimization.

A. Kartelj completed his Ph.D. in Computer science at Faculty of Mathematics at the University of Belgrade, Serbia in year 2014. His research interests cover areas of optimization, mathematical programming and data mining. His publications are mostly related to metaheuristic optimization methods, data classification and dimensionality reduction.

D. Matić received Ph.D. degree (2013) in Mathematics at the Faculty of Mathematics at the University of Belgrade, Serbia. His research interests in science include combinatorial optimization and metaheuristics, and their application in various fields of mathematics, computer science, industry and education.

References

Angluin, D., Aspnes, J., & Reyzin, L. (2015). Network construction with subgraph connectivity constraints. Journal of Combinatorial Optimization, 29(2), 418–432. https://doi.org/10.1007/s10878-013-9603-2

Browne, F., Wang, H., Zheng, H., & Azuaje, F. (2009). GRIP: A web-based system for constructing gold standard datasets for protein-protein interaction prediction. Source Code for Biology and Medicine, 4(1), 1–7. https://doi.org/10.1186/1751-0473-4-2

Chockler, G., Melamed, R., Tock, Y., & Vitenberg, R. (2007). Constructing scalable overlays for pub-sub with many topics. In Proceedings of the Twenty-Sixth Annual ACM Symposium on Principles of Distributed Computing (pp. 109–118). Association for Computing Machinery.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., Von Mering, C., & Jensen, L. J. (2012). STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research, 41(D1), D808–D815. https://doi.org/10.1093/nar/gks1094

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Düppefeld, B., & Edelmann, A. (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature, 440(7084), 631–636. https://doi.org/10.1038/nature04532

Grbić, M., Crnogorac, V., Predojević, M., Kartelj, A., & Matić, D. (2020). How well are known protein complexes supported in PPI networks? In Proceedings of International Conference on Innovations in Intelligent Systems and Applications (inista) (pp. 1–7). IEEE.

Grbić, M., Kartelj, A., Janković, S., Matić, D., & Filipović, V. (2020). Variable neighborhood search for partitioning sparse biological networks into the maximum edge-weighted k-plexes. IEEE/ACM
Transactions on Computational Biology and Bioinformatics, 17(5), 1822–1831. https://doi.org/10.1109/TCBB.2019.2898189

Grbić, M., Matić, D., Kartelj, A., Vračević, S., & Filipović, V. (2020). A three-phase method for identifying functionally related protein groups in weighted PPI networks. Computational Biology and Chemistry, 107246. https://doi.org/10.1016/j.compbiolchem.2020.107246

Hong, E. L., Balakrishnan, R., Dong, Q., Christie, K. R., Park, J., Binkley, G., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., & J. E. Hirschman (2007). Gene ontology annotations at SGD: New data sources and annotation methods. Nucleic Acids Research, 36(suppl_1), D577–D581. https://doi.org/10.1093/nar/gkm909

Hüffner, F., Komusiewicz, C., Liebtrau, A., & Niedermeier, R. (2013). Partitioning biological networks into highly connected clusters with maximum edge coverage. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11(3), 455–467. https://doi.org/10.1109/TCBB.2013.177

Kiemer, L., Costa, S., Ueffing, M., & Cesareni, G. (2007). WI-PHI: A weighted yeast interactome enriched for direct physical interactions. Proteomics, 7(6), 932–943. https://doi.org/10.1002/ (ISSN)1615-9861

Liu, G., Wong, L., & Chua, H. N. (2009). Complex discovery from weighted PPI networks. Bioinformatics (Oxford, England), 25(15), 1891–1897. https://doi.org/10.1093/bioinformatics/btp311

Liu, Q., Song, J., & Li, J. (2016). Using contrast patterns between true complexes and random subgraphs in PPI networks to predict unknown protein complexes. Scientific Reports, 6(1), 21223. https://doi.org/10.1038/srep21223

Martins, P. (2016). Modeling the maximum edge-weight k-plex partitioning problem. arXiv preprint arXiv:1612.06243.

Mewes, H. W., Frishman, D., Mayer, K., Mokrejs, M., Morgenstern, B., & Weil, B. (2002). MIPS: A database for genomes and protein sequences. Nucleic Acids Research, 30(1), 31–34. https://doi.org/10.1093/nar/30.1.31

Mladenović, N., & Hansen, P. (1997). Variable neighborhood search. Computers & Operations Research, 24(11), 1097–1100. https://doi.org/10.1016/S0305-0548(97)00031-2

Nakajima, N., Hayashida, M., Jansson, J., Maruyama, O., & Akutsu, T. (2018). Determining the minimum number of protein-protein interactions required to support known protein complexes. PloS One, 13(4), e0195545. https://doi.org/10.1371/journal.pone.0195545

Nepusz, T., Yu, H., & Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods, 9(5), 471–472. https://doi.org/10.1038/nmeth.1938

Pu, S., Wong, J., Turner, B., Cho, E., & Wodak, S. J. (2009). Up-to-date catalogues of yeast protein complexes. Nucleic Acids Research, 37(3), 825–831. https://doi.org/10.1093/nar/gkn1005

Rissanen, J. (1978). Modeling by shortest data description. Automatica, 14(5), 465–471. https://doi.org/10.1016/0005-1098(78)90005-5

Szklarczyk, D., Franceschini, A., Wyder, S., Forsslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., & Kuhn, M. (2015). STRING v10: Protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research, 43(D1), D447–D452. https://doi.org/10.1093/nar/gku1003

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., & Jensen, L. J. (2019). STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research, 47(D1), D607–D613. https://doi.org/10.1093/nar/gky1131

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., & Jensen, L. J. (2016). The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Research, 45(D1), D362–D368. https://doi.org/10.1093/nar/gkw937

Zhang, Z., Song, J., Tang, J., Xu, X., & Guo, F. (2018). Detecting complexes from edge-weighted PPI networks via genes expression analysis. BMC Systems Biology, 12(4), 40. https://doi.org/10.1186/s12918-018-0565-y