On Combinatorial Optimisation in Analysis of Protein-Protein Interaction and Protein Folding Networks

David Chalupa(✉)

Department of Computer Science, University of Hull,
Cottingham Road, Hull HU6 7RX, UK
D.Chalupa@hull.ac.uk

Abstract. Protein-protein interaction networks and protein folding networks represent prominent research topics at the intersection of bioinformatics and network science. In this paper, we present a study of these networks from combinatorial optimisation point of view. Using a combination of classical heuristics and stochastic optimisation techniques, we were able to identify several interesting combinatorial properties of biological networks of the COSIN project. We obtained optimal or near-optimal solutions to maximum clique and chromatic number problems for these networks. We also explore patterns of both non-overlapping and overlapping cliques in these networks. Optimal or near-optimal solutions to partitioning of these networks into non-overlapping cliques and to maximum independent set problem were discovered. Maximal cliques are explored by enumerative techniques. Domination in these networks is briefly studied, too. Applications and extensions of our findings are discussed.

Keywords: Combinatorial optimisation · Protein-Protein interaction networks · Protein folding networks · Hybrid heuristics

1 Introduction

Bioinformatics has been a rapidly growing field in the last years. Certain biological problems can be modelled using networks, most notably gene regulatory networks [1] and protein-protein interaction (PPI) networks [2]. Solutions to network problems, which are relatively well studied in computer science, are often regarded as valuable for biologists [3].

In this paper, we present a unified study of combinatorial optimisation problems in analysis of PPI and protein folding (PF) networks. The aim of this paper is to explore the unique area at the intersection of two areas of applied evolutionary computation and computational intelligence in general. On one hand, it spans the computational intelligence in bioinformatics and on the other hand, we explore the biological networks using methodologies of evolutionary computation and heuristics in combinatorial optimisation.

G. Squillero and P. Burelli (Eds.): EvoApplications 2016, Part I, LNCS 9597, pp. 91–105, 2016.
DOI: 10.1007/978-3-319-31204-0_7

Contributions. Using a combination of classical and randomised search heuristics, we obtain high-quality solutions to some of the well-known combinatorial optimisation problems in PPI and PF networks, which are known to be NP-hard in general [4,5].

Experimental results are presented for networks of the European COSIN project [6]. For four different PPI networks, we obtain optimal solutions to maximum independent set and minimum vertex clique covering problem. We used a combination of greedy approximation algorithm for maximum independent set in sparse graphs [7] with a hybrid of iterated greedy (IG) clique covering and randomised local search (RLS) for maximum independent set [8]. For three of four PPI networks, we obtain optimal solutions to maximum clique and chromatic number problems using a hybrid of Brélaz's heuristic [9] with iterated greedy graph colouring algorithm [10]. To explore the minimum dominating set problem, we use a classical greedy approximation algorithm [11].

In addition, we apply the same techniques to a PF network, which is considerably larger than PPI networks. A reduced variant of this PF network is explored, too. We obtain that PF network has slightly different properties than PPI networks, which is probably related both to its size and structure. However, we obtained a very small gap between bounds for maximum clique size and chromatic number of this network, too.

The paper is organised as follows. In Sect. 2, we present an overview of the topic from several relevant perspectives. In Sect. 3, we present our approach to study of PPI and PF networks. In Sect. 4, we present the experimental results and their possible application. Finally, in Sect. 5, we formulate conclusions and summarise scientific problems, which remain open.

2 Combinatorial Optimisation Problems in Protein-Protein Interaction and Protein Folding Networks

There is a body of work concentrating on computer-scientific aspects of study of biological networks. In this section, we present an overview of relevant research and perspectives on our topic.

Protein-protein interaction (PPI) networks. Vertices of a PPI network represent proteins and edges represent interactions between them. These are constructed by molecular biologists usually as an outcome of two-hybrid screening experiments [3]. Analysis of PPI networks and their comparison represent common research topics [12], along with development of analytical software for biological networks [13]. In our experiments, we study public domain PPI network data of the European COSIN project [6]. These include PPI networks for bacterium *Escherichia Coli*, commonly found in gastrointestinal tract; nematode worm *Caenorhabditis elegans*; *Helicobacter pylori*, a bacteria associated with gastritis, usually found in upper gastrointestinal tract; and *Saccharomyces cerevisiae*, a commonly used species of yeast. PPI network data for yeast are a common subject of study [14,15].

Clustering of PPI networks. Probably the most well-known topic in computer-scientific research of PPI networks is represented by clustering of these networks, i.e. decomposition into relatively dense subgraphs. In PPI networks, this is motivated by the problems of complex and functional module detection, which aim to identify groups of mutually interacting proteins, which might often be involved in the same biological processes [16,17].

It is worth noting that biologists tend to distinguish between the term "complex" and "module". Complex in PPI network refers to a molecular machine of proteins, which bind to each other at the same time and space, while the term module refers to a group of mutually interacting proteins, which control certain cellular function, without taking the spatial and temporal aspect into account [18]. However, experimentally obtained PPI data often do not incorporate this information in the network. PPI network data are valuable in reconstruction of metabolic and signalling pathways [3], understanding of cell regulation, prediction of role of uncharacterised proteins and for possible therapy [18]. Multifunctional proteins have previously been revealed [19], i.e. discovery of overlapping modules is a relevant topic for PPI networks, too [20]. One way how they can be detected, is the use of clique merging [21].

Clustering of PPI networks has many similarities with detection of community structure in social networks [22]. Both areas suffer from existence of a large number of diverse clustering algorithms, using ideas ranging from information flow simulation [23], spectral properties of adjacency matrices [24,25], cost-based clustering [26], to stochastic optimisation techniques [18]. However, quality of such a clustering algorithm can be evaluated using a wide spectrum of metrics and multiple objective functions can be considered [27]. Both clustering quality and applicability of developed methods to large networks seem to be important [28]. It can be observed that different clustering algorithms may output very different clusters, each having a different desirable property of a dense or well separated network substructure [29]. Therefore, multiobjective optimisation was successfully applied to network community detection [30]. However, assessing quality of a clustering of a biological network [31] remains hard and often requires to fall back to usage of a reference solution [18,30] or simply requesting verification from a biologist. Additionally, clustering or partitioning of a network [32] might often lead to NP-hard combinatorial optimisation problems [33], which generally require specific attention [4,5].

Protein folding (PF) network beta3s. This network represents conformation space of a 20 residue antiparallel β-sheet peptide investigated by NMR spectroscopy. Vertices represent conformations and edges represent transitions. The network seems to represent a complex system, in which spontaneous folding of protein is modelled as a (weighted) random walk on the conformation space network. Due to space and methods being used, we only consider the structure of the network and omit the weights [34].

PPI and PF networks have also been previously studied in the context of centrality metrics and their stability and potential decomposition [35]. Enumerative and spectral analytical methodologies were also used to study their struc-

ture [24]. Statistical analysis of complex networks helps in understanding of the large-scale properties of these networks, too [36].

Combinatorial optimisation problems in networks. We investigate five different classical NP-hard combinatorial optimisation problems [4,5]. For simplicity, we describe these problems only less formally.

Maximum clique is the largest subgraph, in which each pair of vertices is adjacent. In the context of PPI networks, it is the largest group of proteins, in which all proteins mutually interact. Maximum clique size is denoted by ω. There is a spectrum of algorithms for this problem [37].

Graph colouring is an assignment of colours to vertices such that each for each edge, its vertices are differently coloured. Minimum number of colours needed to obtain a graph colouring is called chromatic number and is denoted by χ. Chromatic number is useful, since for each graph, it holds that $\omega \leq \chi$ [38], i.e. maximum clique and chromatic number represent bounds for each other. Randomised algorithms are frequently used to solve graph colouring, too [39].

Maximum independent set is the largest subgraph, in which no pair of vertices is adjacent. Maximum independent set size is denoted by α. In a PPI network, independent set is the largest set of mutually non-interacting proteins.

Minimum vertex clique covering is a partitioning of the network into as few non-overlapping cliques as possible. In PPI networks, it represents a problem of finding the minimum number of clusters such that within each cluster, all proteins must be mutually interacting. The number of cliques in a minimum vertex clique covering is denoted by ϑ. Similarly to maximum clique and graph colouring, it holds that $\alpha \leq \vartheta$ [8]. Hence, maximum independent set and minimum vertex clique covering represent bounds for each other, too.

The last studied problem is the *minimum dominating set problem*. Minimum dominating set is the smallest subset of vertices such that each vertex is either in the dominating set or has a neighbour in it. Minimum dominating set size is denoted by γ. For PPI networks, dominating set represents a set of "central" proteins such that all other proteins interact with at least one protein of the dominating set.

3 Our Experimental Approach

Graph-theoretical approaches represent a vital part of the tools used to analyse biological networks [43]. We aim to provide an approach for their exploration, which ensures solid generalisation and computes properties, which are naturally related to functional module identification. Indeed, large cliques, independent sets and dominating sets represent such properties. Additionally, these problems have clear definitions and approaches, which can easily be applied to previously unexplored PPI or possibly other biological networks. The aim is to provide a hybrid technique, providing bounds for several well-defined valuable properties of an unknown network, which lead to NP-hard combinatorial optimisation problems.

Algorithm 1. Experimental Approach for Analysis of Combinatorial Optimisation Problems in Large Networks

	Input: network modelled as a graph $G = [V, E]$ Output: maximum clique and chromatic number bound interval $[\omega_L, \chi_U]$ maximum independent set and clique covering number bound interval $[\alpha_L, \vartheta_U]$ minimum dominating set interval $[\gamma_L, \gamma_U]$
1	find a lower bound $\omega_L \leq \omega$ using the following greedy algorithm for construction of a clique Q
2	$Q = \emptyset$
3	order vertices in G from the largest degree to the smallest
4	for vertex v in this ordering
5	if v is adjacent to all vertices in C
6	$Q = Q \cup \{v\}$
7	$\omega_L = \lvert Q \rvert$
8	find a colouring C and an upper bound $\chi_U \geq \chi$ using Brélaz's heuristic [9] with binary heap [40,41]
9	if $\omega_L \neq \chi_U$
10	use IG graph colouring heuristic [42,10] starting with C, combined with RLS for maximum clique starting with Q to compute new bounds ω_L and χ_u
11	find an independent set I and a lower bound $\alpha_L \leq \alpha$ using the greedy approximation algorithm for maximum independent set in sparse graphs [7]
12	use IG heuristic for minimum vertex clique covering, combined with RLS for maximum independent set [8], starting with independent set I, to compute bounds α_L and ϑ_U
13	use the greedy approximation algorithm for minimum dominating set, based on minimum set covering [11] to compute upper bound γ_U
14	compute the number of connected components c
15	compute the lower bound $\gamma_L = \min\left\{c, \frac{\lvert V \rvert}{\Delta + 1}, \frac{\gamma_U}{\ln \Delta + 1}\right\}$, where Δ is the maximum degree of a vertex

This way, we are able to characterise the structure of the networks using cliques, independence and domination and avoid the broad notion of general clustering.

To carry out our investigations, we use a collection of classical heuristics, as well as order-based stochastic algorithms to find high-quality solutions to our combinatorial optimisation problems. The main process of mining from the network data is characterised by the pseudocode of Algorithm 1.

Let us now describe the steps in a slightly more detailed way. Due to lack of space, we are not able to review all aspects of the algorithms we used. However, an interested reader may refer to the referenced work.

In steps 1–7, we use a simple greedy clique algorithm. It starts with an empty clique and orders vertices from largest degree to the smallest. It puts the current vertex to the clique if and only if the clique property is not violated by adding

the new vertex. In fact, this approach is equivalent to use of greedy algorithm for independent set [7] for the complement of our graph.

In step 8, we use Brélaz's heuristic implemented with binary heap to find a colouring of the network in $\mathcal{O}(m \log n)$ time, where n is the number of vertices and m is the number of edges.

If maximum clique from steps 1–7 and number of colours used in step 8 are not equal, we use iterated greedy (IG) graph colouring search heuristic [10,42], combined with randomised local search (RLS) for maximum clique. This is represented by steps 9–10 of Algorithm 1. We start with clique and colouring found in steps 1–8. IG uses randomised block-based moves to possibly reduce the colouring. RLS for maximum clique has not previously been used. Therefore, we describe it in more detail.

RLS for maximum clique uses the same algorithm for clique construction as in steps 1–7. However, it works with a predefined permutation instead of ordering the vertices by their degrees. In the beginning, vertices of clique Q are put into a permutation first and other vertices are ordered at random after that. In each time step of RLS, *jump* move is attempted. The *jump* move simply takes a uniformly random vertex from the permutation and puts it to the first position in the permutation. The other vertices are then shifted to the right. Resulting permutation is used to construct a new clique and is accepted if the new clique is at least as large as the current one.

In step 11, we use the greedy approximation algorithm for maximum independent set in sparse and bounded degree graphs [7]. We use binary heap as a priority queue.

In step 12, we apply the recently proposed IG heuristic for minimum vertex clique covering with RLS for maximum independent set [8].

In step 13, the greedy approximation algorithm for dominating set is used to compute an upper bound for minimum dominating set size [11]. Additionally, a lower bound for the size of minimum dominating set is computed in steps 14–15. This lower bound represents a maximum of three different lower bounds. One of them is the number of components c, the second bound is a general bound derived from maximum degree Δ and the third bound is implied by logarithmic approximation guarantee of the greedy algorithm.

Note that our approach is not specifically restricted to PPI and PF networks. It can easily be applied to social networks or other complex network data. However, for the purpose of this study, we focus specifically on its suitability to explore biological network data.

4 Experimental Results and Discussion

We performed the evaluation in two parts. We first used the approach without the stochastic techniques based on IG and RLS (i.e. we omitted steps 10 and 12). Hence, we used only greedy algorithms. To provide an upper bound for ϑ, we used Brélaz's heuristic applied to complementary graph $\overline{G}$. $\overline{G}$ contains edges

between pairs of vertices, which are not adjacent in G and vice versa. In Table 1, we present the best results obtained by this approach in 20 independent runs.

For evaluation of the impact of stochastic components of the approach, we then used the full approach, as specified by Algorithm 1. These results are presented in Table 2. Similarly, we performed 20 independent runs for PPI networks and the reduced PF network *beta3s.reduced* and present the best results obtained. For the large PF network *beta3s*, we performed only one long run.

The stochastic subroutines of our approach were parameterised as follows. For IG for graph colouring and RLS for maximum clique, we used a simultaneous implementation with 5 iterations of RLS per one iteration of IG. Stochastic optimisation was stopped when $100n$ iterations without improvement of neither clique nor colouring were encountered. Similarly, IG for minimum vertex clique covering and RLS for maximum independent set were used in an implementation with 5 iterations of RLS per one iteration of IG. Stopping criterion was similar, too. Optimisation was stopped when $100n$ iterations without improvement of neither clique covering nor independent set were encountered. Interestingly, these stopping criteria led to results with good quality and solid scalability for all four of these problems.

Both Tables 1 and 2 have the following structure. The first column contains the name of the network. Its number of vertices n, number of edges m, number of connected components c and the number of triangles τ are specified along with the name. The next columns present the maximum clique size ω, chromatic number χ, maximum independent set size α, minimum clique covering size ϑ and minimum dominating set size γ. If a cell contains only one value, it means that the value is a numerically proven optimum for the particular characteristic. If it contains two values separated by $-$, it means that the value is located within the interval specified by presented values. Symbol n in the table means that the value is upper bounded only by the number of vertices n. Bold numbers in Table 2 represent values, which were obtained only by the stochastic approach, i.e. randomised search techniques were beneficial for these instances.

Additionally, we also performed listing of maximal cliques for each network [46]. A clique is maximal if it is not a subgraph of some other clique. The reason is to confront of the number of maximal cliques and maximum (i.e. largest) cliques and to further analyse the cliques as building blocks of the networks.

Network *ecoli* contains a maximum clique of 6 mutually interacting proteins. Using emumeration based on triangles, we found that there are 657 maximal cliques of size at least 3. There are 5 of these cliques, which consist of 6 proteins. The network *ecoli* can be partitioned into $\vartheta = 161$ non-overlapping cliques, with an average size of such a clique being 1.678. There also is a dominating set of 69 proteins, for which it holds that all other proteins interact with at least protein of this set.

For network *elegans*, we have that its maximum clique size is 3 and there are 39 triangles representing maximum cliques. It can be partitioned into 294 non-overlapping cliques. The average size of such a clique is 1.276, which makes it the network with the smallest average clique size in minimum vertex clique

Table 1. Experimental results obtained for PPI and PF networks by using only greedy algorithms (i.e. without steps 10 and 12 in Algorithm 1).

G	ω	χ	α	ϑ	γ
PPI networks					
ecoli [44]	5–6	5–6	160–161	160–161	20–69
$n = 270$, $m = 716$, $c = 20$, $\tau = 478$					
elegans [45]	3	3	293–294	293–294	20–71
$n = 375$, $m = 405$, $c = 20$, $\tau = 13$					
helico [45]	3	3–4	521–528	521–528	33–163
$n = 732$, $m = 1403$, $c = 16$, $\tau = 76$					
yeast [45]	3–8	3–8	2641–2673	2641–2673	146–959
$n = 4142$, $m = 7839$, $c = 99$, $\tau = 1562$					
PF networks					
beta3s.reduced [34]	37–39	37–39	229–301	229–301	11–70
$n = 1287$, $m = 23948$, $c = 1$, $\tau = 219165$					
beta3s [34]	37–39	37–39	64053-n	64053-n	5375–40323
$n = 132168$, $m = 228967$, $c = 2$, $\tau = 241209$					

Table 2. Experimental results obtained for PPI and PF networks by using the full stochastic approach, including IG and RLS algorithms (i.e. full Algorithm 1, including steps 10 and 12). Bold values represent instances, for which IG and RLS provided improved results compared to purely greedy algorithms.

G	ω	χ	α	ϑ	γ
PF networks					
ecoli [44]	**6**	**6**	**161**	**161**	20–69
$n = 270$, $m = 716$, $c = 20$, $\tau = 478$					
elegans [45]	3	3	**294**	**294**	20–71
$n = 375$, $m = 405$, $c = 20$, $\tau = 13$					
helico [45]	3	3–4	**528**	**528**	33–163
$n = 732$, $m = 1403$, $c = 16$, $\tau = 76$					
yeast [45]	**7**	**7**	**2673**	**2673**	146–959
$n = 4142$, $m = 7839$, $c = 99$, $\tau = 1562$					
PF networks					
beta3s.reduced [34]	**38**–39	**38**–39	**259–282**	**259–282**	11–70
$n = 1287$, $m = 23948$, $c = 1$, $\tau = 219165$					
beta3s [34]	**38**–39	**38**–39	**64497–69667**	**64497–69667**	5375–40323
$n = 132168$, $m = 228967$, $c = 2$, $\tau = 241209$					

covering. This is understandable, since this network is the sparsest. It contains a dominating set consisting of 71 vertices.

For network *helico*, we obtained a clique of size 3, while we were only able to find a 4-colouring. This is the only PPI network, for which we obtained a gap between an estimate for maximum clique size and chromatic number. Using enumeration, we found that there is no 4-clique and the number of triangles of mutually interacting proteins is 76. However, this confirms that while chromatic number can be used as a good upper bound on the size of the maximum clique of mutually interacting proteins, it seems that one cannot guarantee that these values for PPI networks will be equal. Network *helico* can be partitioned into 528 non-overlapping cliques of average size 1.386. It also contains a dominating set of size 164.

Instance *yeast* contains 1872 maximal cliques, which is the largest number of maximal cliques among the studied PPI networks. However, only 12 of them are also maximum cliques, which contain 7 vertices. These will shortly be discussed below. Network *yeast* can be partitioned into 2673 non-overlapping cliques, which have average size 1.550. Dominating set on 959 vertices for this network is the largest among the PPI networks, too.

It is not surprising that numbers of maximal and maximum cliques, as well as the properties of non-overlapping and overlapping cliques seem to vary between different networks. Hence, it might be interesting to discuss the properties of large clique a bit further.

Table 3 presents a listing of 12 maximum cliques of size 7 in the *yeast* PPI network. One can notice that the first clique and the last two cliques consist of proteins, which are not present in other cliques. On the other hand, all other cliques represent extensions of clique $CEF1, SEC28, SET1, SFA1, SFB2$. This indicates that some interesting substructures might be relatively isolated, while other substructures form larger clusters. These structures can be modelled

Table 3. Listing of 12 maximum cliques of size 7 in *yeast* PPI network.

$ALD3, ALG2, ALG9, ANC1, ANP1, AOS1, APA1$
$CEF1, LIP5, MKK2, SEC28, SET1, SFA1, SFB2$
$CEF1, LIP5, MKK2, SEC28, SET2, SFA1, SFB2$
$CEF1, LIP5, PUB1, SEC28, SET1, SFA1, SFB2$
$CEF1, LIP5, PUB1, SEC28, SET2, SFA1, SFB2$
$CEF1, LSM3, MKK2, SEC28, SET1, SFA1, SFB2$
$CEF1, LSM3, MKK2, SEC28, SET2, SFA1, SFB2$
$CEF1, LSM3, PUB1, SEC28, SET1, SFA1, SFB2$
$CEF1, LSM3, PUB1, SEC28, SET2, SFA1, SFB2$
$CEF1, PUB1, SEC28, SET1, SFA1, SFB2, SIN3$
$GRE2, HIR2, HIR3, HIS6, HIS7, HIT1, HMG1$
$SPA2, YAP1802, YAP3, YAP5, YAP6, YAR003W, YAR009C$

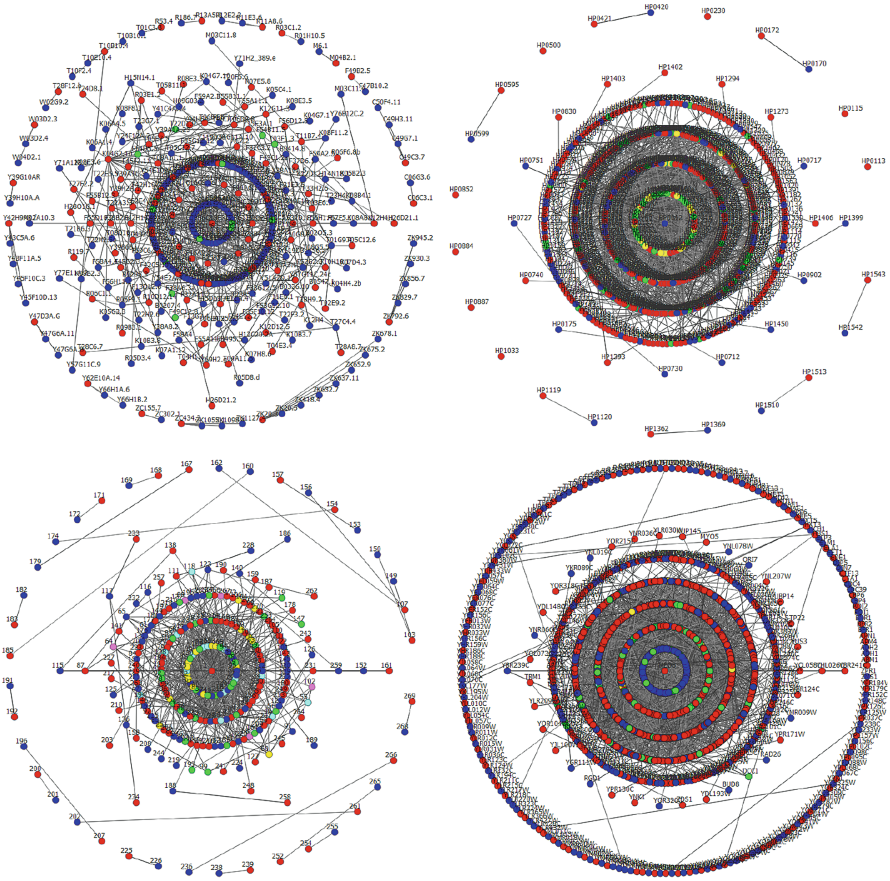

Fig. 1. Visualisation of colourings found for protein-protein interaction networks *elegans* (upper, on the left-hand side), *helico* (upper, on the right-hand side), *ecoli* (lower, on the left-hand side) and *yeast* (lower, on the right-hand side). These colourings represent good upper bounds for the size of maximum clique of mutually interacting proteins for these PPI networks. Based on the availability of protein labels and expected visual quality, labels or indices of vertices are presented for some of the networks and vertices.

e.g. by merging cliques [21]. Additionally, large cliques seem to comprise smaller cliques. This suggests that some PPI networks might have a hierarchical structure [47]. While functional modules are formed by groups of cliques, it seems that one can even identify smaller cliques as low-level building blocks of the network. Interestingly, while labels of proteins are naturally dependent on conventions of biologists, some of the identified maximum cliques seem to consist of proteins with lexicographically similar labels.

PF networks have slightly different characteristics. Network *beta*3*s.reduced* is atypical due to its reduced representation, which features a drastic cutoff in

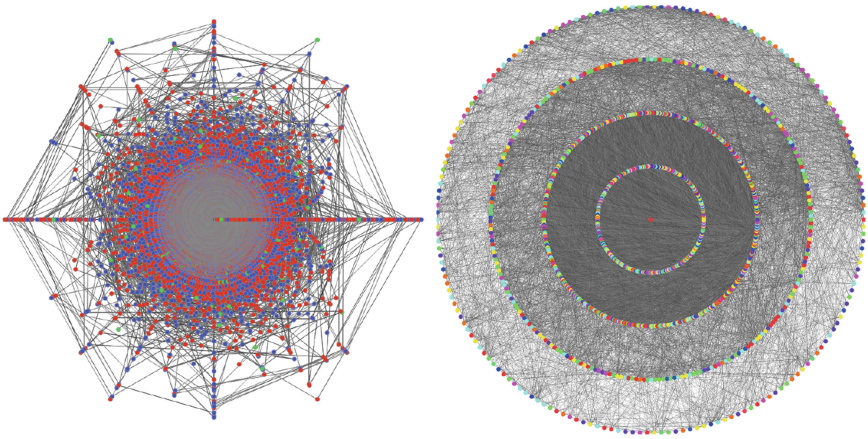

Fig. 2. Visualisation of *beta3s* protein folding network (on the left-hand side), which is the largest of studied networks, with over 10^5 vertices and its "core" *beta3s.reduced* (on the right-hand side). One can easily see that *beta3s.reduced* is the subgraph, which requires a high number of colours, while the visualisation of *beta3s* highlights mostly the three colours, which are used to colour most of the vertices in the "outer layer" of the network.

vertices with low degree. As a consequence, both *beta3s* and *beta3s.reduced* contain maximum clique of size 38–39, while the average size of a clique needed to partition *beta3s.reduced* into non-overlapping cliques is 4.564–4.969. This is a value range previously observed in variations of random graphs and graphs with planted cliques [8]. The original network *beta3s* requires cliques of size 1.897–2.049 to obtain a minimum vertex clique covering. However, it is worth mentioning that this value is still somewhat higher than the values obtained for PPI networks. This indicates a denser structure of PF networks with some large embedded cliques found in the "core" of the network. Such a phenomenon has not been observed in the studied PPI networks.

Summarising the above results, combinatorial optimisation properties seem to vary between different PPI networks. Comparison between a purely greedy and a stochastic approach confirms that stochastic optimisation techniques help in combinatorial optimisation for PPI and PF networks. Figure 1 indicates that to a certain extent, PPI networks seem to have similar structure. This figure shows colourings obtained for the PPI networks and groups vertices to layers, based on distance from a vertex with maximum degree. Visualisations reveal dense subgraphs in the proximity of the vertex with maximum degree, while this property seems to be more accentuated for large networks. Figure 2 presents a similar visualisation for the PF networks. This visualisation reveals slightly "layered" structure of *beta3s.reduced*. In this context, it is not surprising that large cliques are located within *beta3s.reduced*, while the "outer" layer of *beta3s* is sparser and seems to contain much smaller cliques.

While *yeast* network contains relatively large cliques of size 7, networks *elegans* and *helico* do not contain a clique larger than a simple triangle. Large cliques may both heavily overlap and represent relatively "isolated" substructures. Properties of cliques in network *yeast* seem to indicate hierarchical structure. For this purpose, data reductions might represent a promising research direction. A specific case is represented by the large *beta*3*s* PF network, which might further be studied in this context, too. Dominating sets were explored using an approximation algorithm. More interesting results might be obtained using a nature-inspired heuristic for this problem, e.g. algorithms based on ant colony optimisation [48].

5 Conclusions

We presented an experimental study of combinatorial optimisation problems in protein-protein interaction (PPI) and protein folding (PF) networks. Studied problems included maximum clique, chromatic number, maximum independent set, minimum vertex clique covering and minimum dominating set. We presented a unified technique to estimate these properties of large networks, which lead to NP-hard problems in general. Our experimental approach revealed several interesting properties of four PPI networks of the European COSIN project, as well as PF network *beta*3*s* and its reduced version. Even though the approach was applied to biological networks, its ideas are general and can also be used to analyse other complex networks, such as social networks or research citation networks.

Our investigation found maximum cliques for all PPI networks and provided a very small interval for the maximum clique of PF network *beta*3*s*. For all four PPI networks, we found the optimal solution to the problem of their partitioning into non-overlapping cliques. We confronted our method with the use of stochastic elements of iterated greedy (IG) and randomised local search (RLS) algorithms to its variant without the elements of stochastic optimisation. This confrontation revealed that stochastic optimisation approaches provide results of better quality for maximum clique, chromatic number, maximum independent set and minimum vertex clique covering.

Overlapping cliques were investigated using enumerative methods, too. This investigation suggests that some of the studied PPI networks have a hierarchical structure, with large overlapping cliques possibly consisting of smaller cliques. We also identified the dominating sets of these networks. In the context of PPI networks, these are the sets of "central" proteins such that all other proteins interact with at least one protein of the dominating set.

We believe that this approach might be beneficial especially in exploration of new biological networks. Most of the studied problems are closely related to functional module detection. However, unlike network clustering, studied characteristics are clearly defined and can be used as a systematic basis for further investigations.

References

1. Boyer, F., Morgat, A., Labarre, L., Pothier, J., Viari, A.: Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data. Bioinformatics **21**(23), 4209–4215 (2005)
2. Gao, L., Sun, P., Song, J.: Clustering algorithms for detecting functional modules in protein interaction networks. J. Bioinform. Comput. Biol. **7**(1), 217–242 (2009)
3. Cohen, J.: Bioinformatics - an introduction for computer scientists. ACM Comput. Surv. **36**(2), 122–158 (2004)
4. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 3rd edn. MIT Press, Cambridge (2009)
5. Karp, R.M.: Reducibility among combinatorial problems. In: Miller, R., Thatcher, J. (eds.) Proceedings of a Symposium on the Complexity of Computer Computations, pp. 85–103. Plenum Press, New York (1972)
6. COSIN: Coevolution and Self-organization in Dynamical Networks. http://www.cosinproject.org/
7. Halldórsson, M.M., Radhakrishnan, J.: Greed is good: approximating independent sets in sparse and bounded-degree graphs. Algorithmica **18**(1), 145–163 (1997)
8. Chalupa, D.: Construction of near-optimal vertex clique covering for real-world networks. Computing and Informatics (to appear)
9. Brélaz, D.: New methods to color vertices of a graph. Commun. ACM **22**(4), 251–256 (1979)
10. Culberson, J.C., Luo, F.: Exploring the k-colorable landscape with iterated greedy. In: Johnson, D.S., Trick, M. (eds.) Cliques, Coloring and Satisfiability: Second DIMACS Implementation Challenge, pp. 245–284. American Mathematical Society, RI (1995)
11. Chvátal, V.: A greedy heuristic for the set-covering problem. Math. Oper. Res. **4**(3), 233–235 (1979)
12. Atias, N., Sharan, R.: Comparative analysis of protein networks: hard problems, practical solutions. Commun. ACM **55**(5), 88–97 (2012)
13. Kuchaiev, O., Stevanović, A., Hayes, W., Pržulj, N.: Graphcrunch 2: software tool for network modeling, alignment and clustering. BMC Bioinf. **12**(1), 24 (2011)
14. Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Jensen, C.R.L.J., Bastuck, S., Dümpelfeld, B., et al.: Proteome survey reveals modularity of the yeast cell machinery. Nature **440**(7084), 631–636 (2006)
15. Zaki, N., Berengueres, J., Efimov, D.: Prorank: a method for detecting protein complexes. In: Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation, pp. 209–216. ACM (2012)
16. Li, X., Wu, M., Kwoh, C.K., Ng, S.K.: Computational approaches for detecting protein complexes from protein interaction networks: a survey. BMC Genomics **11**(Suppl. 1), S3 (2010)
17. Pizzuti, C., Rombo, S.E., Marchiori, E.: Complex detection in protein-protein interaction networks: a compact overview for researchers and practitioners. In: Giacobini, M., Vanneschi, L., Bush, W.S. (eds.) EvoBIO 2012. LNCS, vol. 7246, pp. 211–223. Springer, Heidelberg (2012)
18. Pizzuti, C., Rombo, S.E.: Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods. Bioinformatics **30**(10), 1343–1352 (2014)
19. Becker, E., Robisson, B., Chapple, C.E., Guénoche, A., Brun, C.: Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinformatics **28**(1), 84–90 (2006)

20. Adamcsek, B., Palla, G., Farkas, I.J., Derényi, I., Vicsek, T.: CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics **22**(8), 1021–1023 (2012)
21. Li, X.L., Tan, S.H., Foo, C.S., Ng, S.K.: Interaction graph mining for protein complexes using local clique merging. Genome Inf. **16**(2), 260–269 (2005)
22. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Nat. Acad. Sci. **99**(12), 7821–7826 (2002)
23. Cho, Y.R., Hwang, W., Zhang, A.: Identification of overlapping functional modules in protein interaction networks: information flow-based approach. In: Sixth IEEE International Conference on Data Mining Workshops, ICDM Workshops 2006, pp. 147–152 (2006)
24. Hawick, K.A.: Applying enumerative, spectral and hybrid graph analyses to biological network data. In: International Conference on Computational Intelligence and Bioinformatics (CIB 2011), pp. 89–96. IASTED, Pittsburgh, USA, 7–9 November 2011
25. Sun, J., Xie, Y., Zhang, H., Faloutsos, C.: Less is more: sparse graph mining with compact matrix decomposition. Stat. Anal. Data Min. **1**(1), 6–22 (2008)
26. King, A.D., Pržulj, N., Jurisica, I.: Protein complex prediction via cost-based clustering. Bioinformatics **20**(17), 3013–3020 (2004)
27. Pizzuti, C., Rombo, S.E.: Experimental evaluation of topological-based fitness functions to detect complexes in PPI networks. In: Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation, pp. 193–200. ACM (2012)
28. Pizzuti, C., Rombo, S.E.: A coclustering approach for mining large protein-protein interaction networks. IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB) **9**(3), 717–730 (2012)
29. Leskovec, J., Lang, K.J., Mahoney, M.W.: Empirical comparison of algorithms for network community detection. In Rappa, M., Jones, P., Freire, J., Chakrabarti, S. (eds.) Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 631–640. ACM, New York, NY (2010)
30. Pizzuti, C.: A multiobjective genetic algorithm to find communities in complex networks. IEEE Trans. Evol. Comput. **16**(3), 418–430 (2012)
31. Brohee, S., Helden, J.V.: Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinf. **7**(1), 488 (2006)
32. Schaeffer, S.E.: Graph clustering. Comput. Sci. Rev. **1**(1), 27–64 (2007)
33. Šíma, J., Schaeffer, S.E.: On the NP-completeness of some graph cluster measures. In: Wiedermann, J., Tel, G., Pokorný, J., Bieliková, M., Štuller, J. (eds.) SOFSEM 2006. LNCS, vol. 3831, pp. 530–537. Springer, Heidelberg (2006)
34. Rao, F., Caflisch, A.: The protein folding network. J. Mol. Biol. **342**(1), 299–306 (2004)
35. Hawick, K.A.: Centrality metrics for comparing protein-protein interaction networks with synthesized NK systems. In: Proceedings of the IASTED International Conference on Biomedical Engineering, pp. 1–8. IASTED, Zurich, Switzerland, 23–25 June 2014
36. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**(1), 47–97 (2002)
37. Wu, Q., Hao, J.K.: A review on algorithms for maximum clique problems. Eur. J. Oper. Res. **242**(3), 693–709 (2015)
38. Welsh, D.J.A., Powell, M.B.: An upper bound for the chromatic number of a graph and its application to timetabling problems. Comput. J. **10**(1), 85–86 (1967)

39. Galinier, P., Hamiez, J.-P., Hao, J.-K., Porumbel, D.: Recent advances in graph vertex coloring. In: Zelinka, I., Snasel, V., Abraham, A. (eds.) Handbook of Optimization: From Classical to Modern Approach. ISRL, vol. 38, pp. 505–528. Springer, Heidelberg (2013)
40. Morgenstern, C.: Improved implementations of dynamic sequential coloring algorithms. Technical report CoSc-91-4, Department of Computer Science, Texas Christian University (1991)
41. Turner, J.S.: Almost all k-colorable graphs are easy to color. J. Algorithms **9**(1), 63–82 (1988)
42. Culberson, J.C.: Iterated greedy graph coloring and the difficulty landscape. Technical report TR92-07, University of Alberta (1992)
43. Pavlopoulos, G.A., Secrier, M., Moschopoulos, C.N., Soldatos, T.G., Kossida, S., Aerts, J., Schneider, R., Bagos, P.G., et al.: Using graph theory to analyze biological networks. BioData Min. **4**(10), 1–27 (2011)
44. Butland, G., Peregrín-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., Emili, A.: Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature **433**, 531–537 (2005)
45. UCLA: Database of Interacting Proteins. http://dip.doe-mbi.ucla.edu/dip/Main.cgi
46. Eppstein, D., Löffler, M., Strash, D.: Listing all maximal cliques in sparse graphs in near-optimal time. In: Cheong, O., Chwa, K.-Y., Park, K. (eds.) ISAAC 2010, Part I. LNCS, vol. 6506, pp. 403–414. Springer, Heidelberg (2010)
47. Kovács, I.A., Palotai, R., Szalay, M.S., Csermely, P.: Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. PloS ONE **5**(9), e12528 (2010)
48. Potluri, A., Singh, A.: Two hybrid meta-heuristic approaches for minimum dominating set problem. In: Panigrahi, B.K., Suganthan, P.N., Das, S., Satapathy, S.C. (eds.) SEMCCO 2011, Part II. LNCS, vol. 7077, pp. 97–104. Springer, Heidelberg (2011)