A local search based heuristic for clustering large biological networks into highly connected components

Milana Grbić,[1] Aleksandar Kartelj,[2] Dragan Matić,[1] Vladimir Filipović[2]

[1]Faculty of Natural Science and Mathematics, University of Banja Luka
[2]Faculty of Mathematics, University of Belgrade

Introduction

Clustering large networks into smaller components can be of a great importance for discovering new properties of a specific structure. In this work we deal with partitioning biological networks into highly connected components by removing as few edges as possible. A network with n nodes is called highly connected if the degree of each node is greater than $n/2$.
More formally, if $G = (V,E)$ is a network with the set of nodes V and the set of edges E, the task of the highly connected deletion problem (HCD) is to find a minimum subset of edges E' of E such that in $G' = (V,E\backslash E')$, all connected components are highly connected. By definition, a graph containing only one edge (K_2) is not highly connected and all singletons are considered as un-clustered. The mentioned optimization problem is proven to be NP-hard, so exact methods fail to find solution for large scaled instances in a general case. Therefore, heuristic approaches applied to the considered problem are worth being investigated.

Biological justification for partitioning graphs into highly connected components

The large biological networks generated from protein-protein interaction (PPI) data can be analyzed by identifying functional subgroups. One way for identifying such subgroups is the network clustering into highly connected components (Hartuv, E. et Shamir, R., 2000). The idea of HCD is to identify clusters that are densely connected and don't have so many intersections with the rest of the graph. The proteins in the obtained clusters can have similar GO annotations (Hüffner, F. et al. 2014) .

Considered networks

We applied our algorithm on PPI network for the organism *Schizosaccharomyces pombe*. Two types of PPI networks were analyzed: the first type contains all interactions, while the second type is constructed only on physical interactions. The quality of the obtained results is compared to the results from the literature. In order to further investigate the behavior of our approach, we applied our algorithm on a metabolic network, built on the principle that the metabolites are represented as nodes and two metabolites are adjacent if they figure in at least one common reaction. In the example, the network is modeled from the list of 1393 metabolic reactions of the yeast *Saccharomyces cerevisiae*, taken from (Förster, J. et al., 2003).

Description of the heuristic algorithm

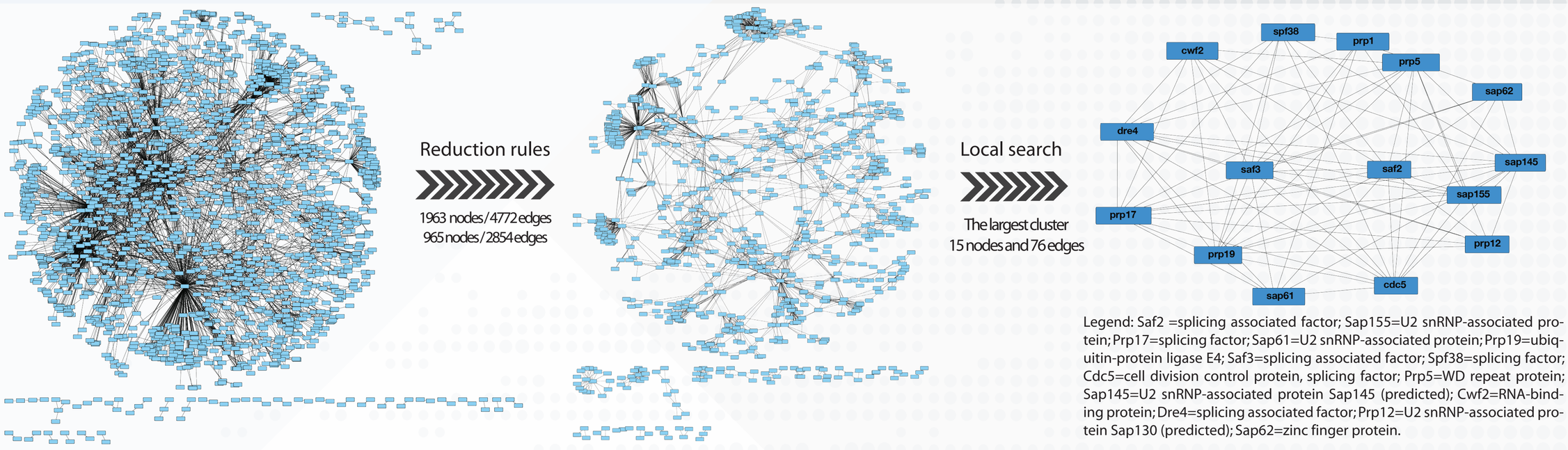
Our algorithm consists of two phases: (i) a preprocessing phase consisting of a polynomial time data reduction and (ii) a local search heuristic.

Data reduction phase

Based on the reduction rules proposed by Hüffner et al. (2014), the starting networks are significantly reduced to the smaller ones by deleting up to 75% edges. Here we mention the Rule 3 (Hüffner, F. et al. 2014), which was shown to be the most effective: *If there are two adjacent vertices u and v that have no common neighbors, then delete the edge {u,v} and decrease the total number of deleted edges by 1.*

Local search phase

After the reductions were performed, the local search is separately applied to each connected component of the reduced network. At the beginning of the local search, an initial solution is created in such a way that each vertex is assigned to a single component. In each iteration, the algorithm is trying to improve the current solution by moving a selected vertex from its current component to some other component. If the improvement is achieved, it is applied immediately and the improvement process continues. Otherwise, other vertices are tested for improvement until the whole set of vertices is exhausted. In addition, for two randomly chosen singletons that are adjacent in the starting graph, the algorithm seeks for the third singleton which closes the triangle with these two vertices. If that situation happens, all three vertices are joined together in one highly connected component, i.e. triangle. Objective function allows for infeasible solutions to appear. Infeasibility mostly occurs during the process of local search when vertices are moved from one component to another. This, however, does not have negative influence on the overall search process since objective function subtly penalizes infeasible solutions and thus favours feasible solutions over infeasible in the long run.
The algorithm stops if one of two conditions is satisfied: total number of iterations is reached (in our case 25000), or the best found solution is not improved in 10000 iterations.



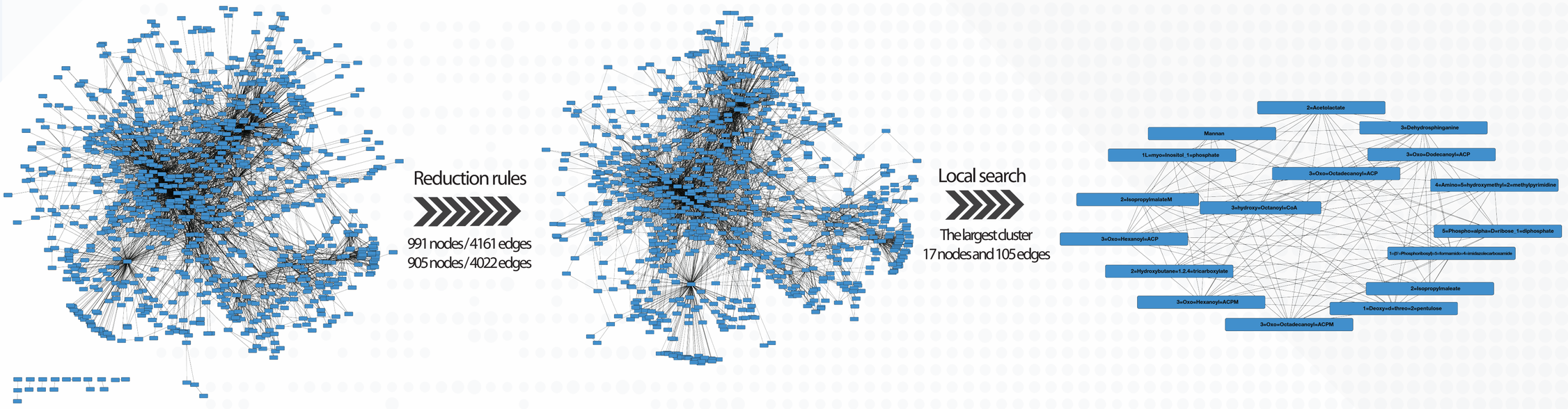
Legend: Saf2 =splicing associated factor; Sap155=U2 snRNP-associated protein; Prp17=splicing factor; Sap61=U2 snRNP-associated protein; Prp19=ubiquitin-protein ligase E4; Saf3=splicing associated factor; Spf38=splicing factor; Cdc5=cell division control protein, splicing factor; Prp5=WD repeat protein; Sap145=U2 snRNP-associated protein Sap145 (predicted); Cwf2=RNA-binding protein; Dre4=splicing associated factor; Prp12=U2 snRNP-associated protein Sap130 (predicted); Sap62=zinc finger protein.

Experimental results

All experiments are performed on the Intel i7-4770 CPU @3.40 GHz with 8 GB RAM and Windows 7 operating system. For each execution, only one thread/processor is used. The algorithm is implemented in C programming language and compiled with Visual Studio 2013 compiler.
As it is shown in table below, experimental results indicate that our algorithm is competitive to other algorithms from literature.

Graphical representation

In order to further investigate the results obtained by our heuristic technique, we decided to graphically interpret the results by using a well known software platform for visualizing biological networks Cytoscape (Shannon, P. et al. 2003). We use this software to present our results in more suitable way, enabling biologists to better understand the relations between the considered objects. Therefore, we adopted our results in a way that can be easily interpreted by the Cytoscape software, getting an attractive visualization of the obtained partitions.



Instance	|V|	|E|	min-cut without DR				min-cut with DR				neighborhood with DR				column generation with DR				**our algorithm**			
			k	n	m	t	k	n	m	t	k	n	m	t	k	n	m	t	k	n	m	t
Schizosaccharomyces_pombephys	1963	4772	4324	17	96	16	4165	17	96	2	3961	15	71	2	3811	17	96	102	**3882**	**15**	**76**	**1591**
Schizosaccharomyces_pombe_all	3735	51620	50343	63	1268	526	50311	63	1268	214	49514	60	1175	3491	-	-	-	-	**4772**	**31**	**309**	**5442**

Conclusions

Highly connected deletion problem partitioning problem is of a great interest from both theoretical and practical points of view. This work contains some preliminary results obtained by applying two-phases algorithm on some metabolite networks and some PPI networks. The overall algorithm consists of the network reduction phase and the local search heuristic method, applied to the reduced network.
In near future, we plan to improve our algorithm to be more robust and accurate and to apply it to other real life biological instances from literature.

Biological evaluation

The results obtained on the PPI network justify the assumption that the proteins with similar GO annotation are placed in the same highly connected component. By applying the algorithm on the metabolic network, we got mainly expected results. As it can be seen from the figure above, in the largest component metabolites represent the biosynthesis process up to 8C atoms, during the fatty acid metabolism and its mitochondrial synthesis. In addition, mannan also figures in this separated component, since it contains glucose which is degraded to the acetolactates during the glycolysis.

References

Hüffner, F. et al. (2014). " Partitioning Biological Networks into Highly Connected Clusters with Maximum Edge Coverage. " *IEEE/ACM Trans Comput Biol Bioinform.,* 11(3):455-67.
Hartuv, E. et al. (2000). " A clustering algorithm based on graph connectivity. " *Information Processing Letters*, vol. 76, no. 4–6, pp. 175–181.
Förster, J. et al. (2003). "Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network." *Genome research* 13.2, 244-253.
Shannon, P. et al. (2003) "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome research* 13.11 2498-2504.

First Congress of Molecular Biologists of Serbia 20 – 22 September 2017, Belgrade, Serbia