Enhanced Weighted Restricted Neighborhood Search Clustering: A Novel Algorithm for Detecting Human Protein Complexes from Weighted Protein-Protein Interaction Graphs

Christos Dimitrakopoulos[1], Konstantinos Theofilatos[1], Andreas Pegkas[1],
Spiros Likothanassis[1], and Seferina Mavroudi[1,2]

[1] Department of Computer Engineering and Informatics, University of Patras, Greece
[2] Department of Social Work, School of Sciences of Health and Care,
Technological Educational Institute of Patras, Greece
{dimitrakop,theofilk,pegkas,likothan,mavroudi}@ceid.upatras.gr

Abstract. Proteins and their interactions have been proven to play a central role in many cellular processes. Although there are many experimental techniques for protein-protein interaction prediction, only a few exist for predicting protein complexes. For the sake of this, researchers have emphasized lately in the computational prediction of protein complexes from Protein-Protein Interaction (PPI) data. The two major limitations of the current advances in the prediction of protein complexes are that most of the algorithms do not take into consideration the participation of a protein to many protein complexes and that they cannot handle weighted PPI graphs. In the present paper, we altered the original Restricted Neighborhood Search Clustering (RNSC) algorithm to overcome the above limitations. The Enhanced Weighted Restricted Neighborhood Search Clustering (EWRNSC) permits the participation of a protein to many protein complexes by modifying the moves of the original RNSC. In addition, EWRNSC can accept and process weighted PPI graphs as inputs by altering the cost functions of the original RNSC cost clustering schemes. When experimented using atasets from Human, the proposed algorithm proved to outperform the original RNSC and the MCL algorithms which are two of the most broadly used methods in the field of protein complexes prediction.

Keywords: Protein Complexes Prediction, Human, Cost-based Clustering, Clustering Protein-Protein Interaction Networks, Weighted Protein-Protein Interaction Networks.

1 Introduction

Proteins are nowadays considered to be the most important participants in molecular interactions. Specifically, they play a significant role in almost all the cellular functions such as regulatory signals transmission in the cell and they catalyze a huge number of chemical reactions. Except for functioning alone, proteins are also combined to each other in functional modules called protein complexes. The prediction of the protein complexes is crucial for understanding the cellular

L. Iliadis, H. Papadopoulos, and C. Jayne (Eds.): EANN 2013, Part II, CCIS 384, pp. 231–240, 2013.

mechanisms and for predicting the function of uncharacterized proteins. The experimental prediction of protein complexes is mainly limited to Tandem Affinity Purification (TAP) [1] which provide erroneous data and demand high cost without being time-efficient.

Because of the above fact, researchers have emphasized lately in the computational prediction of protein complexes from Protein-Protein Interaction (PPI) data. To achieve this goal, several clustering methods have been applied to the protein interactome graph in order to detect highly connected subgraphs [2,3,4]. These algorithms rely on very different approaches. Each of them requires specifying several parameters, some of which may drastically affect the results.

The Restricted Neighborhood Search Clustering (RNSC) [17], is a cost-based local heuristic search algorithm that explores the solution space to minimize a cost function, calculated according to the numbers of intra-cluster and inter-cluster edges. Starting from an initial random solution, RNSC iteratively moves a vertex from one cluster to another if this move reduces the general cost. When a (user-specified) number of moves has been reached without decreasing the cost function, the program ends up. The algorithm is analytically described in section 2.2.

In the present paper, we propose a fully unsupervised clustering algorithm, EWRNSC, which is an enhancement of the original Restricted Neighborhood Search Clustering (RNSC) algorithm. The original RNSC algorithm was altered so that a) it permits the participation of a protein to many protein complexes b) the initial estimation of the clusters is allocated using an analytical estimation method and c) takes advantage of the information which lies within the weights of weighted PPI graphs. The participation of a protein to many protein complexes (in terms of clusters) was achieved by altering the moves of the original RNSC. Two new operators were added together with the move of a node from a cluster to another random cluster. The algorithm chooses each move with a given probability. The process of weighted PPI graphs as inputs was achieved by altering the cost functions of the original RNSC cost clustering schemes (section 2.3).

When experimented using public available protein complex datasets from Human, the proposed algorithm proved to outperform the original one. The proposed method was tested on one weighted PPI graph from Human using two evaluation datasets (section 2.1).

2 Materials and Methods

2.1 Protein-Protein Interactions Datasets

In the present paper a Human PPI dataset was used to build the PPI graph which is used as input for the protein complex prediction methods. The examined dataset is weighted where the value of a confidence score is assigned to each protein pair.

The PPI dataset for the Human organism consists of the protein interactions included in the HPRD database [6]. These protein interactions were filtered using the method proposed in [14] and a confidence score was assigned to each protein pair using the same methodology. In this way we achieved to incorporate sequential, functional and structural information on the extracted PPI graph. The extracted PPI graph consists of 7450 proteins and 21.475 interactions.

2.2 Enhanced Weighted Restricted Neighborhood Clustering Algorithm

The Enhanced Weighted RNSC is a novel algorithm for detecting protein complexes based on the original RNSC algorithm with two major improvements aiming at the increase of its efficiency as well as its flexibility. The first improvement is the modification of the metrics so that the process of weighted PPI graphs is enabled. The second is the modification of the algorithm so that the moving process enables the possibility of generating overlapping clusters.

As in the original RNSC, the Enhanced Weighted RNSC uses two cost functions for evaluating solutions, the naive cost function and the scaled cost function. The naive cost function is simple in its computation and is used as a preprocess solution tool. For computational matters, it processes the network and finds an initial approximative solution by ignoring the weights of the network. The cost function that determines the final solution is the scaled cost function which is computationally more pretentious as it takes into consideration the weights of the network.

Naive cost function: For each node α_v is computed, which is the sum of the nodes′ "bad connections", naming the sum of the weights of the node's edges with nodes that belong to different clusters plus the sum of the weights of the nodes that belong to the same cluster with the subjective node and they are not connected to it (equation 2.1). Since the node might participate in more than one cluster we take the average of a_v over all clusters.

$$C_n(G,C) = \frac{1}{2}\sum_{u\in V}\frac{\sum_{u\in Cu} a_v}{|C_u|} \qquad (2.1)$$

where V is the set of the nodes of graph G and C_u is the set of all the clusters that node u belongs.

Scaled cost function. Let us define:

- $w_{v,u}$ the weight that connects the nodes v and u
- C_v the cluster where the node v belongs

Then we define for each node v:

$$\gamma_v = \sum_{u\notin C_v} w_{v,u} + \sum_{u\in C_v}(1 - w_{v,u}) \qquad (2.2)$$

Moreover, for each node v, if N(v) is the number of nodes connecting to it, then we define:

$$\beta_v = |\mathrm{N}(v)\cup C_v| \qquad (2.3)$$

which means that β_v is equal to the number of nodes that belong to either the "neighbors" of v or to the same cluster with v. If n is the total number of the graph's nodes, then the scaled cost function of EWRNSC is defined as:

$$C_s(G,C) = \frac{n-1}{3}\sum_{v\in V}\sum_{v\in C_u}\frac{\frac{\gamma_v}{\beta_v}}{|C_u|} \qquad (2.4)$$

As in the original RNSC, the ultimate goal of the algorithm is the minimization of the cost functions where ideally the nodes of a cluster connect to each other (all to all) whereas they do not connect to any other nodes (nodes of other clusters). After the execution of the EWRNSC, we filtered out the clusters with size equal to one protein.

EWRNSC has also many parameters which need to be tuned. The number of the initial random clusters of the algorithm was chosen based on the number of the benchmark dataset's clusters. Other parameters like tabu list length, diversification parameters and stopping tolerance were chosen empirically based on the size of the input PPI graph as described in [4]. In specific, the parameters are the number of different experiments that we ran the algorithm (10), the length of the list with the forbidden moves (50), the frequency of the random diversification moves (50), the number of nodes shuffled when diversification is performed (10), and the number of moves without improvement in cost for the naive and the scale schemes (10).

Our intention is to transform RNSC algorithm to a new form that takes into consideration the participation of the proteins to more than one clusters. For the sake of this transformation, we alter the moving procedures conducted during the naive and the scaled cost scheme. In the original RNSC one node randomly moves from one cluster to another. In the Enhanced Weighted RNSC one node has the possibility to perform one of three following operators:

- **Operator 1**: The node is moved from its cluster i to a random cluster j ~ i with probability Pr_mov.
- **Operator 2**: The node is copied from its cluster i to a random cluster j ~ i with probability Pr_cop.
- **Operator 3**: The node is deleted from its cluster i with probability Pr_del.

In general, the probabilities Pr_cop and Pr_del must be lower that the Pr_mov probability so that we do not fall into high cluster overlapping. Various experimentations have been tested for different values of the three probabilities and they are presented in Section 3. By using the above operators one node is allowed to move to another cluster and at the same time to remain to its current cluster (operator 2). Operator 3 is a prerequisite operator generated by the existence of operator 2, because its absent would result to the creation of extremely large -extremely overlapping clusters. In Operators 1 and 2 the node is moved or copied to another cluster or to a singleton cluster with equal probability. In Operator 3 the node is deleted with equal probability from one of the clusters that it belongs to. As in the original RNSC, we consider moving one node only to the clusters of its neighbors (or to a singleton cluster). A move to a cluster that contains none of its neighbors never does as well, in terms of either the naive cost function or the scaled cost function, as moving the vertex to an empty cluster.

The initial estimation of the clusters is allocated using an analytical estimation method. We consider the datasets for the protein complexes of Yeast as the most valuable, due to the fact that the existing knowledge about protein complexes of the yeast organism is in satisfactory levels (compared to the Human dataset). For the Yeast organism, there are three well studied protein complex datasets. The first is the BT_409 dataset [8] which contains of 409 protein complexes. The second is the Aloy dataset [8] which contains 101 protein complexes derived using structure based protein matching with known structures and screened with the electron microscopy method. The third dataset is the Pu dataset [9] which contains 408 protein complexes. The three datasets contain in total 811 complexes (without duplications). The most known current available PPI dataset for the Yeast (Saccharomyces cerevisiae) was published by [7] and contains information about 5195 proteins. We use the proportion of proteins and complexes for Yeast as a reference (5195 proteins and 811 complexes) to compute the expectation complexes for Human. In this way, we set the number of 1163 initial clusters.

The EWRNSC algorithm was built in Matlab R2010b. The algorithm's pseudocode follows:

```
Input:
   PPI network (an undirected weighted graph) G(V,E)
   Number of experiments: Ne=10
   Tabu Length = 50
   Diversification frequency = 50
   Diversification length = 10
   $T_n$ = 10
   $T_s$ = 10
Output:
   The predicted protein clusters: Clusters

Algorithm:
   Initialize clustering based on analytical method
   for 1 to Ne
      Call Naive_Scheme
      Call Scaled_Scheme
   end
   Store the best Clustering of the Experiments: Clusters

Routine Naive_Scheme:
    Until Best cost has improved in the last $T_n$ moves:
        Choose a Node not in Tabu List.
        Make a Move with Probability 0.8 that decreases the total Naive Cost.
        Make a Copy with Probability 0.1 that decreases the total Naive Cost.
        Make a Delete with Probability 0.1 that decreases the total Naive Cost.
        (destroy or create clusters in the process)
        Update Tabu List
    end
    Store the best Naive Clustering

Routine Scaled_Scheme:
    Until Best cost has improved in the last $T_s$ moves:
        Choose a Node not in Tabu List.
        Make a Move with Probability 0.8 that decreases the total Scaled Cost.
        Make a Copy with Probability 0.1 that decreases the total Scaled Cost.
        Make a Delete with Probability 0.1 that decreases the total Scaled Cost.
        (destroy or create clusters in the process)
        Update Tabu List
    end
    Store the best Scaled Clustering
```

2.3 Evaluation Metrics

Sensitivity, positive predictive value (*PPV*) and geometric accuracy are classically used to measure the correspondence between the result of a clustering on a set of reference complexes. Considering the annotated complexes as a reference classification, *sensitivity* is defined as the fraction of proteins of complex *i* which are found in cluster *j*. To characterize the general sensitivity of a clustering result, the *clustering-wise sensitivity is computed* as the weighted average of Sn_{co_i} over all complexes.

$$Sn = \frac{\sum_{i=1}^{n} N_i Sn_{co_i}}{\sum_{i=1}^{n} N_i} \tag{2.5}$$

Defined on one cluster, the positive predictive value is the proportion of members of cluster *j* which belong to complex *i*, relative to the total number of members of this cluster assigned to all complexes. To characterize the general PPV of a clustering result as a whole, we compute a *clustering-wise* PPV as the weighted average of the individual PPV_{cl_j} of all clusters.

$$PPV = \frac{\sum_{j=1}^{m} T_{\cdot j} PPV_{cl_j}}{\sum_{j=1}^{m} T_{\cdot j}} \tag{2.6}$$

The *geometric accuracy* (*Acc*) indicates the tradeoff between sensitivity and predictive value. It is obtained by computing the geometrical mean of the *Sn* and the *PPV*.

$$Acc = \sqrt{Sn * PPV} \tag{2.7}$$

The advantage of taking the geometric rather than arithmetic mean is that it yields a low score when either the *Sn* or the *PPV* metric is low and as a result it balances better the tradeoff between the two metrics. The sensitivity and PPV individually give a false idea of quality in the trivial cases where all proteins are assigned to a single cluster ($Sn = 1 \Rightarrow Acc > 0.5$) or where, on the contrary, each protein is assigned to a single-element cluster ($PPV = 1 \Rightarrow Acc > 0.5$). To avoid these erroneous interpretations we also have used the Separation metric [11] which takes into consideration the fact that clustering predictions with fewer known complexes should be regarded as the ones with the higher quality.

2.4 Evaluation Datasets

We built two different protein complex datasets by filtering out the protein complexes which are published in CORUM [12]. The first human dataset (443 protein complexes) was created by filtering out all protein complexes which include a protein that is not present in interactions annotated in the HPRD database. The second human evaluation dataset (1097 protein complexes) consists of protein complexes in CORUM when filtering out all complexes with more than half of their proteins not included in the HPRD PPIs. Protein complexes with one protein are considered as poor interconnecting components and are discarded.

3 Experimental Results and Discussion

EWRSNC focuses on local searching and as a result we used diversification moves and multiple experiments to aid it in escaping from local optimum solutions. In specific, we ran 10 different experiments and calculated the mean values for its evaluation metrics. For the probabilities of the operator moves we chose 0.8 for operator 1, 0.1 for operator 2 and 0.1 for operator 3. 111 proteins were found to participate in different clusters. The performance of EWRNSC was compared with the performance of the original RNSC as well as with the performance of MCL [18] which is one of the most well-established and frequently used methods for the prediction of protein complexes. The results for these two algorithms were calculated using the Superclusteroid Tool [13] which uses the optimized values for their parameters as described in [11].

In Figures 3.1 and 3.2 we present the evaluation metrics of the aforementioned algorithms when evaluating their outcome with the two Human dataset. EWRNSC exceeds almost all the classical metrics of sensitivity, PPV and geometric accuracy compared to both MCL and RNSC (except for the PPV case in the CORUM Complexes with more than 50% in HPRD benchmark dataset) maintaining a small improvement in the separation metric. In specific, the improvement for the geometric accuracy metric varies between 6.8% and 19,1% for the original RNSC and between 4,8% and 11,4% for the MCL algorithm. The minor improvement in the separation metric can be attributed to the absence of overlapping clusters in Human evaluation datasets.

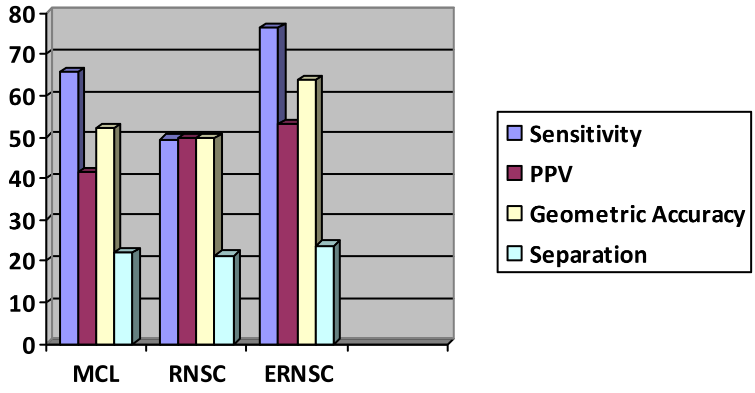

Fig. 3.1. Comparative Results for the Human organism (CORUM Complexes with only proteins included in HPRD)

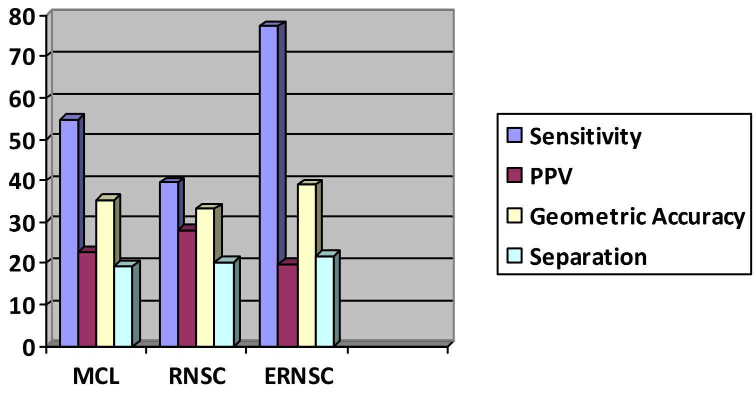

Fig. 3.2. Comparative Results for the Human organism (CORUM Complexes with more than 50% proteins included in HPRD)

For the reasons described in section 2.4, we consider the three classical measurements (Sensitivity, PPV, geometric accuracy) as biased. Separation is the only metric that takes into consideration the participation of one protein to more than one protein complexes (by avoiding duplications during computation) and as a result we consider as highly important the EWRNSC's trend on better separation results.

We can mainly attribute the improvements of the EWRNSC algorithm to its handling of the edges' weights. All methods for predicting Human protein-protein interactions are known to yield a nonnegligible amount of noise (false positives) and to miss a fraction of existing interactions (false negatives) [11]. Therefore, the protein interaction data available for clustering are very noisy. The framework of the EWRNSC that incorporates weighted PPI graphs definitely faces that problem directly by evaluating the confidence of each PPI.

Clusters of a protein interaction network may overlap with each other. Most proteins have more than one molecular function and participate in more than one biological process. For example, some proteins form transient associations and are part of several complexes at different stages. Therefore, the traditional clustering approaches of putting each protein into one single cluster do not suit this problem well. Hence, the EWRNSC algorithm creates clusters that are closer to the real interaction models that exist in the organisms compared to the clusters produced by the MCL and RNSC algorithms.

4 Conclusion and Future Challenges

In the post-genomic era, an important issue is to analyze biological systems at the network level, in order to understand the topological organization of protein interaction networks, identify protein complexes and functional modules, discover functions of uncharacterized proteins, and obtain more exact networks. To achieve this aim, a series of clustering approaches have been proposed.

In the context of the present paper, we proposed an unsupervised clustering algorithm, EWRNSC, which is an enhancement of the original Restricted Neighborhood Search Clustering (RNSC) algorithm which was altered in order to permit the participation of a protein to many protein complexes and to handle the input of weighted PPI graphs. The participation of a protein to many protein complexes (in terms of clusters) was achieved by altering the moves of the original RNSC. To handle weighted graphs, we have affected RNSC's cost functions. Two new operators were added together with the move of a node from a cluster to another random cluster. Moreover, the initial estimation of the clusters is allocated using an analytical estimation method.

As a future work we intend to implement EWRNSC method to more available datasets of PPI networks and known protein complexes from different organisms. It is a matter of robustness to prove the superiority of the method to more than one datasets. Moreover, our goal is to also compare our method with other promising algorithms [7, 15, 16] except for the well-established ones such as MCL and RNSC.

Enhanced Weighted RNSC algorithm has a large number of parameters that are chosen empirically [5] as there is little a priori knowledge for them (cluster number, tabu list length and diversification length). As a future direction, the effective tuning of the parameters within the heuristic procedure of the algorithm would give an improved clustering solution.

Current clustering approaches mainly focus on detecting clusters in static protein interaction networks. However, both the protein-protein interactions and protein complexes are dynamically organized when implementing special functions. Dynamic modules generally correspond to the sequential ordering of molecular events in cellular systems. The way to explore dynamic modules from static protein interaction networks is a very difficult task and should definitely be addressed by the EWRNSC algorithm in a future direction.

Acknowledgments. This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

References

1. Theofilatos, K.A., Dimitrakopoulos, C.M., Tsakalidis, A.K., Likothanassis, S.D., Papadimitriou, S.T., Mavroudi, S.P.: Computational Approaches for the Prediction of Protein-Protein Interactions: A Survey. Current Bioinformatics 6(4), 398–414 (2011)
2. Spirin, V., Mirny, L.A.: Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. USA 100(21), 12123–12128 (2003)
3. Bader, G.D., Hogue, C.W.V.: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4, 2 (2003)
4. Tamas, N., Haiyuan, Y., Alberto, P.: Detecting overlapping protein complexes in protein-protein interaction nteworks. Nature Methods 9, 471–472 (2012)

5. King, A.D.: Graph clustering with restricted neighbourhood search. Master's thesis, University of Toronto (2004)
6. Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D.S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C.J., Kanth, S., Ahmed, M., Kashyap, M.K., Mohmood, R., Ramachandra, Y.L., Krishna, V., Rahiman, B.A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., Pandey, A.: Human Protein Reference Database—2009 update. Nucleic Acids Research 37, D767- D772 (2009)
7. Theofilatos, K., Pavlopoulou, N., Papasavvas, C., Likothanassis, S., Dimitrakopoulos, C., Georgopoulos, E., Moschopoulos, C., Mavroudi, S.: Evolutionary Enhanced Markov Clustering - EEMC: A Novel Unsupervised Methodology for Predicting Protein Complexes From Weighted Protein-Protein Interaction Graphs. In: Artificial Intelligence in Medicine (submitted)
8. Friedel, C., Krumsiek, J., Zimmer, R.: Bootstrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast. Journal of Computational Biology 16(8), 971–987 (2009)
9. Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A.C., Bork, P., Superti-Furga, G., Serrano, L., Russell, R.B.: Structure-based assembly of protein complexes in yeast. Science 303, 2026–2029 (2004)
10. Pu, S., Vlasblom, J., Emili, A., et al.: Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. Proteomics 7, 944–960 (2007)
11. Brohee, S., van Helden, J.: Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7, 488 (2006)
12. Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fogo, G., Frishman, G., Montrone, C., Mewes, H.W.: CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Research 38(database issue), D497–D501 (2009)
13. Ropodi, A., Sakkos, N., Moschopoulos, C., Magklaras, G., Kossida, S.: Superclusteroid: a Web tool dedicated to data processing of protein-protein interaction networks. EMBnet Journal 17(2), 10–15 (2011)
14. Theofilatos, K., Dimitrakopoulos, C., Kleftogiannis, D., Moschopoulos, C., Papadimitriou, S., Likothanassis, S., Mavroudi, S.: HINT-KB: The human interactome knowledge base. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H., Karatzas, K., Sioutas, S. (eds.) Artificial Intelligence Applications and Innovations, Part II. IFIP AICT, vol. 382, pp. 612–621. Springer, Heidelberg (2012)
15. Wang, X., Zhengzhi, W., Jun, Y.: HKC: An Algorithm to Predict Protein Complexes in Protein-Protein Interaction Networks. Journal of Biomedicine and Biotechnology, Article ID 480294, 14 pages (2011), doi:10.1155/2011/480294
16. Wu, M., Li, X., Kwoh, C.K., Ng, S.K.: A core-attachment based method to detect protein complexes in PPI networks. BMC Bioinformatics 10, 169 (2009), doi:10.1186/1471-2105-10-169
17. King, A.D., Przulj, N., Jurisica, I.: Protein complex prediction via cost-based clustering. Bioinformatics 20(17), 3013–3020 (2004)
18. Van Dongen, S.: Graph clustering by flow simulation. In: PhD thesis Centers for mathematics and computer science (CWI), University of Utrecht (2000)