Prediction of New Associations between ncRNAs and Diseases Exploiting Multi-Type Hierarchical Clustering (Discussion Paper)

Emanuele Pio Barracchia[1,3], Gianvito Pio[1,3], Domenica D'Elia[2], and Michelangelo Ceci[1,3,4]

[1] Dept. of Computer Science - University of Bari Aldo Moro, Bari (Italy)
{emanuele.barracchia, gianvito.pio, michelangelo.ceci}@uniba.it
[2] CNR, Institute for Biomedical Technologies - Bari (Italy)
domenica.delia@ba.itb.cnr.it
[3] Big Data Laboratory, CINI Consortium - Rome (Italy)
[4] Dept. of Knowledge Technologies, Jožef Stefan Institute, Ljubljana (Slovenia)

Abstract. The study of functional associations between ncRNAs and human diseases is a pivotal task of modern research to develop new and more effective therapeutic approaches. Nevertheless, it is not a trivial task since it involves entities of different types, such as microRNAs, lncRNAs or target genes. Such a complexity can be faced by representing the involved biological entities and their relationships as a network and by exploiting network-based computational approaches able to identify new associations. However, existing methods are limited to homogeneous networks or can exploit only a limited set of the features of biological entities. To overcome the limitations of existing approaches, we proposed the system LP-HCLUS, which analyzes heterogeneous networks consisting of several types of objects and relationships, each possibly described by a set of features, and extracts hierarchically organized, possibly overlapping, multi-type clusters that are subsequently exploited to predict new ncRNA-disease associations. Our experimental evaluation shows that, according to both quantitative (i.e., TPR@k, ROC and PR curves) and qualitative criteria, LP-HCLUS produces better results.

Keywords: non-coding RNA (ncRNAs) · diseases · cancer · heterogeneous network · clustering · link prediction

1 Introduction

High-throughput sequencing technologies and recent, more efficient computational approaches, have been fundamental for the rapid advances in functional genomics. Among the most relevant results, there is the discovery of thousands of non-coding RNAs (ncRNAs) with a regulatory function on gene expression.

In parallel, the number of studies reporting the involvement of ncRNAs in the development of many different human diseases has grown exponentially. The first type of ncRNAs that has been discovered and largely studied is that of microRNAs (miRNAs), classified as small non-coding RNAs in contrast with long non-coding RNAs (lncRNAs), that are ncRNAs longer than 200nt. While miRNAs primarily act as post-transcriptional regulators, lncRNAs have a plethora of regulatory functions [10]. However, the number of lncRNAs for which the functional and molecular mechanisms are completely elucidated is still quite poor and experimental investigations are still too much expensive for being carried out without any computational pre-analysis. In the last few years, there have been several attempts to computationally predict the relationships among biological entities, such as genes, miRNAs, lncRNAs, diseases [1,11,13,15]. Such methods are based on a network representation of the entities under study and on the identification of new links among nodes in the network. However, most of them are able to work only on homogeneous networks (where nodes and links are of one single type) [5], are strongly limited by the number of different node types or are constrained by pre-defined network structures.

In this discussion paper, we describe the method LP-HCLUS [2], that is able to overcome these limitations. In particular, it can discover new ncRNA-disease relationships from heterogeneous attributed networks (i.e., consisting of different biological entities related by different types of relationships) with arbitrary structure. This ability allows LP-HCLUS to investigate the interactions among different types of entities, possibly leading to increased prediction accuracy.

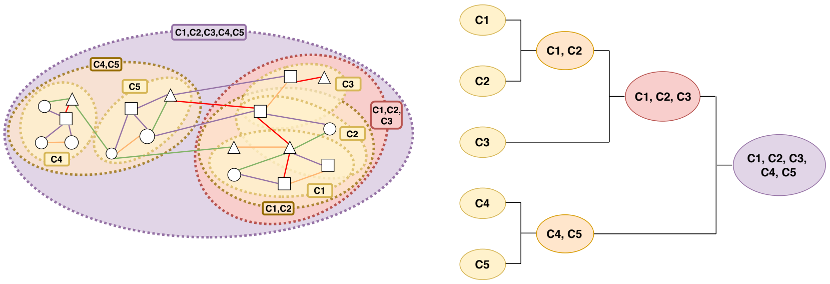
LP-HCLUS exploits a combined approach based on hierarchical, multi-type clustering and link prediction. As we will detail in the next section, a multi-type cluster is actually a heterogeneous sub-network. Therefore, the adoption of a clustering-based approach allows LP-HCLUS to base its predictions on relevant, highly-cohesive heterogeneous sub-networks. Moreover, the hierarchical organization of clusters allows it to perform predictions at different levels of granularity, taking into account either local/specific or global/general relationships.

2 Method

In the following, we introduce the notation and some useful definitions.

Definition 1 (Heterogeneous attributed network). *A heterogeneous attributed network is a network $G = (V, E)$, where V is the set of nodes and E is the set of edges, and both nodes and edges can be of different types. Moreover:*

- *$\mathcal{T} = \mathcal{T}_t \cup \mathcal{T}_{tr}$ is the set of node types, where $\mathcal{T}_t$ is the set of target types, i.e. considered as target of the clustering/prediction task, and $\mathcal{T}_{tr}$ is the set of task-relevant types. Only nodes of target types are clustered and considered in the identification of new relationships.*
- *Each node type $T_v \in \mathcal{T}$ defines a subset of nodes in the network, i.e., $V_v \subseteq V$.*
- *Each node type $T_v \in \mathcal{T}$ is associated with a set of attributes $\mathcal{A}_v = \{A_{v,1}, A_{v,2}, ..., A_{v,m_v}\}$, i.e., nodes of the type T_v are described by the attributes $\mathcal{A}_v$.*
- *$\mathcal{R}$ is the set of all the possible edge types.*
- *Each edge type $R_l \in \mathcal{R}$ defines a subset of edges $E_l \subseteq E$.*



(a) (b)

Fig. 1. A hierarchy of overlapping multi-type clusters: (a) emphasizes the overlapping among multi-type clusters; (b) shows their hierarchical organization.

Definition 2 (Hierarchical multi-type clustering). *A hierarchy of multi-type clusters is defined as a list of hierarchy levels* $[L_1, L_2, \ldots, L_k]$*, where each* L_i *consists of a set of overlapping multi-type clusters. For each level* $L_i, i = 2, 3, \ldots. k$*,* $\forall\, G' \in L_i\ \exists\, G'' \in L_{i-1}$*, such that* G'' *is a subnetwork of* G' *(Fig. 1).*

According to these definitions, we define the task considered in this work.

Definition 3 (Predictive hierarchical clustering for link prediction). *Given a heterogeneous attributed network* $G = (V, E)$ *and the set of target types* $\mathcal{T}_t$*, the goal is to find:*

- *A hierarchy of overlapping multi-type clusters* $[L_1, L_2, \ldots, L_k]$.
- *A function* $\psi^{(w)}: V_{i_1} \times V_{i_2} \rightarrow [0, 1]$ *for each hierarchical level* L_w $(w \in 1, 2, ..., k)$*, where nodes in* V_{i_1} *are of type* $T_{i_1} \in \mathcal{T}_t$ *and nodes in* V_{i_2} *are of type* $T_{i_2} \in \mathcal{T}_t$*. Each function* $\psi^{(w)}$ *maps each possible pair of nodes (of types* T_{i_1} *and* T_{i_2}*) to a score representing the degree of certainty of their relationship.*

In this paper LP-HCLUS has been used to solve the task formalized in Definition 3, by considering ncRNAs and diseases as target types. Hence, we determine two distinct set of nodes denoted by T_n and T_d, representing the set of ncRNAs and the set of diseases, respectively. In the following subsections, we will describe the main steps executed by LP-HCLUS (see Fig. 2 for a general overview).

2.1 Estimation of the strength of the relationship

In the first phase, we estimate the strength of the relationship among all the possible ncRNA-disease pairs in the network G. In particular, we aim to compute a score $s(n_i, d_j)$ for each possible pair n_i, d_j, by exploiting the concept of *meta-path*. According to [14], a *meta-path* is a set of sequences of nodes (involving both target and task-relevant types) which follow the same sequence of edge types, and can be used to fruitfully represent conceptual (possibly indirect) relationships between two entities in a heterogeneous network. Given the ncRNA n_i and the disease d_j, the relationship between them can be considered "certain" if there is at least one meta-path which confirms its certainty.

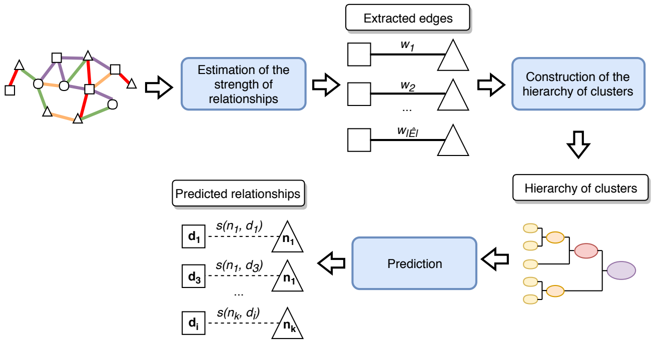


Fig. 2. General workflow of the method LP-HCLUS.

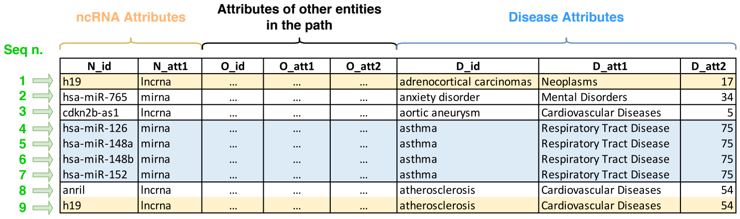
Therefore, by assimilating the score associated with an interaction to its degree of certainty, we compute $s(n_i, d_j)$ as the maximum value observed over all the possible meta-paths between n_i and d_j. Formally:

$$s(n_i, d_j) = \max_{P \in metapaths(n_i, d_j)} pathscore(P, n_i, d_j) \tag{1}$$

where $metapaths(n_i, d_j)$ is the set of meta-paths connecting n_i and d_j, and $pathscore(P, n_i, d_j)$ is the degree of certainty of the relationship between n_i and d_j according to the meta-path P. In order to compute $pathscore(P, n_i, d_j)$, we represent each meta-path P as a finite set of sequences of nodes. If a sequence in P connects n_i and d_j, then $pathscore(P, n_i, d_j) = 1$. Otherwise, following the same strategy introduced before, it is computed as the maximum similarity between the sequences which start with n_i and the sequences which end with d_j (see Fig. 3). The intuition behind this formula is that if n_i and d_j are not directly connected, their score represents the similarity of the nodes and edges they are connected to. The similarity between two sequences seq' and seq'' is computed according to the the attributes of all nodes involved in the two sequences: following [6], if x is numeric, then $s_x(seq', seq'') = 1 - \frac{|val_x(seq') - val_x(seq'')|}{max_x - min_x}$, where min_x (resp. max_x) is the minimum (resp. maximum) value, for the attribute x; if x is not a numeric attribute, $s_x(seq', seq'') = 1$ if $val_x(seq') = val_x(seq'')$, 0 otherwise. In this solution there could be some node types that are not involved in any meta-path. In order to exploit the information conveyed by these nodes, we add an aggregation of their attribute values (the *arithmetic mean* for numerical attributes, the *mode* for non-numerical attributes) to the nodes that are connected to them and that appear in at least one meta-path.

2.2 Construction of a hierarchy of overlapping multi-type clusters

We construct the first level of the hierarchy by identifying a set of overlapping multi-type clusters in the form of bicliques. To this aim, we perform three steps: *i)* **Filtering**, which keeps only the ncRNA-disease pairs with a score greater than (or equal to) β. The result of this step is the subset $\{(n_i, d_j) | s(n_i, d_j) \geq \beta\}$



	ncRNA Attributes		Attributes of other entities in the path			Disease Attributes		
Seq n.	N_id	N_att1	O_id	O_att1	O_att2	D_id	D_att1	D_att2
1 ⇨	h19	lncrna	...	...	...	adrenocortical carcinomas	Neoplasms	17
2 ⇨	hsa-miR-765	mirna	...	...	...	anxiety disorder	Mental Disorders	34
3 ⇨	cdkn2b-as1	lncrna	...	...	...	aortic aneurysm	Cardiovascular Diseases	5
4 ⇨	hsa-miR-126	mirna	...	...	...	asthma	Respiratory Tract Disease	75
5 ⇨	hsa-miR-148a	mirna	...	...	...	asthma	Respiratory Tract Disease	75
6 ⇨	hsa-miR-148b	mirna	...	...	...	asthma	Respiratory Tract Disease	75
7 ⇨	hsa-miR-152	mirna	...	...	...	asthma	Respiratory Tract Disease	75
8 ⇨	anril	lncrna	...	...	...	atherosclerosis	Cardiovascular Diseases	54
9 ⇨	h19	lncrna	...	...	...	atherosclerosis	Cardiovascular Diseases	54

Fig. 3. Sequences between the ncRNA "h19" and the disease "asthma" according to a meta-path. Sequences emphasized in yellow (1 and 9) are those starting with "h19", while sequences emphasized in blue (4, 5, 6 and 7) are those ending with "asthma".

ii) **Initialization**, which builds the initial set of clusters in the form of bicliques, each consisting of a ncRNA-disease pair in $\{(n_i, d_j) | s(n_i, d_j) \geq \beta\}$.
iii) **Merging**, which iteratively merges two clusters C' and C'' into a new cluster C'''. This step regards the initial set of clusters as a list sorted according to an ordering relation $<_c$ that reflects the quality of the clusters. Each cluster C' is then merged with the first cluster C'' in the list that would lead to a cluster C''' which still satisfies the biclique constraint. This step is repeated until no additional clusters that satisfy the biclique constraint can be obtained.

The ordering relation $<_c$ defines a greedy search strategy that guides the order in which pairs of clusters are analyzed. $<_c$ is based on the cluster cohesiveness $h(c)$, that corresponds to the average score in the cluster, namely: $h(C) = \frac{1}{|pairs(C)|} \cdot \sum_{(n_i,d_j) \in pairs(C)} s(n_i, d_j)$, where $pairs(C)$ is the set of all the possible ncRNA-disease pairs that can be constructed from the set of ncRNAs and diseases in the cluster. Accordingly, if C' and C'' are two different clusters, the ordering relation $<_c$ is defined as follows: $C' <_c C'' \iff h(C') > h(C'')$.

The approach adopted to build the other hierarchical levels is similar to the merging step performed to obtain L_1. The main difference is that we do not obtain bicliques, but generic multi-type clusters. Since the biclique constraint is removed, we need another stopping criterion for the iterative merging procedure. Coherently with approaches used in hierarchical co-clustering and following [12], we adopt a user-defined threshold α on the cohesiveness of the obtained clusters. In particular, two clusters C' and C'' can be merged into a new cluster C''' if $h(C''') > \alpha$, where $h(C''')$ is the cluster cohesiveness. This means that α defines the minimum cluster cohesiveness that must be satisfied by a cluster obtained after a merging. The iterative process stops when it is not possible to merge more clusters with a minimum level of cohesiveness α.

2.3 Prediction of new ncRNA-disease relationships

In the last phase, we exploit each level of the identified hierarchy of multi-type clusters as a prediction model. In particular, we compute, for each ncRNA-disease pair, a score representing its degree of certainty on the basis of the multi-type clusters containing it. Formally, let C_{ij}^{w} be a cluster identified in the w-th hierarchical level in which the ncRNA n_i and the disease d_j appear. We compute the degree of certainty of the relationship between n_i and d_j as:

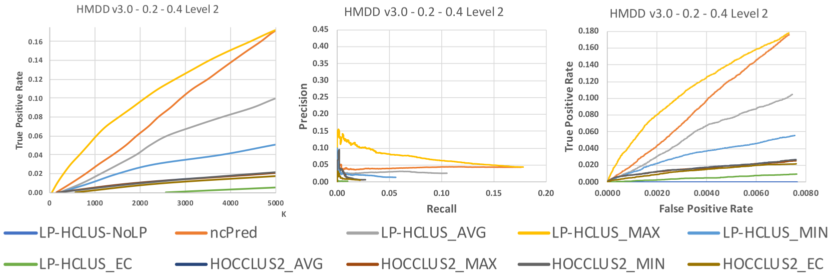


Fig. 4. TPR@k, Precision-Recall and ROC curves results for the dataset HMDD v3.0, obtained with the best configuration ($\alpha = 0.2, \beta = 0.4, level = 2$).

$\psi^{(w)}(n_i, d_j) = h\left(C_{ij}^{w}\right)$, that is, we compute the degree of certainty of the new interaction as the average degree of certainty of the known relationships in the cluster. In some cases, the same interaction may appear in multiple clusters, since the proposed algorithm is able to identify overlapping clusters. In this case, C_{ij}^{w} represents the list of multi-type clusters in which both n_i and d_j appear and we aggregate their cohesiveness values according to four different strategies: maximum, minimum, average and evidence combination [9].

3 Experiments

LP-HCLUS has been run with different values of its input parameters. In particular, following the results obtained in [12], we considered $\alpha \in \{0.1, 0.2\}$ and $\beta \in \{0.3, 0.4\}$. The considered datasets are: *i)* **HMDD v3.0** which stores 985 miRNAs, 675 diseases and 20,859 relationships between diseases and miRNAs; *ii)* **Integrated Dataset (ID)**, built by integrating multiple datasets [3,4,7,8], composed by 7,049 diseases, 70 lncRNA-miRNA relationships, 3,830 relationships between diseases and ncRNAs, 90,242 target genes, 26,522 disease-target associations and 1,055 ncRNA-target relationships.

We compared LP-HCLUS with the following competitors:
i) **HOCCLUS2** [12], a biclustering algorithm that, similarly to LP-HCLUS, identifies a hierarchy of (possibly overlapping) heterogeneous clusters. It is, however, limited to work with only two types of objects. Since its parameters have a similar meaning with respect to LP-HCLUS parameters, we evaluated its results with the same setting, i.e., $\alpha \in \{0.1, 0.2\}$ and $\beta \in \{0.3, 0.4\}$;
ii) **ncPred** [1], a system that was specifically designed to predict new ncRNA-disease associations. ncPred cannot catch information coming from other entities in the network and it is not able to exploit features associated to nodes and links.
iii) **LP-HCLUS-NoLP**, which corresponds to a baseline version of system LP-HCLUS, without the clustering and the link prediction steps. In particular, we consider the score obtained in the first phase of LP-HCLUS (see Section 2.1) as the final score associated with each interaction.

We adopted the 10-fold cross validation on the set of known ncRNA-disease relationships and, due to absence of negative samples, we evaluated the results in

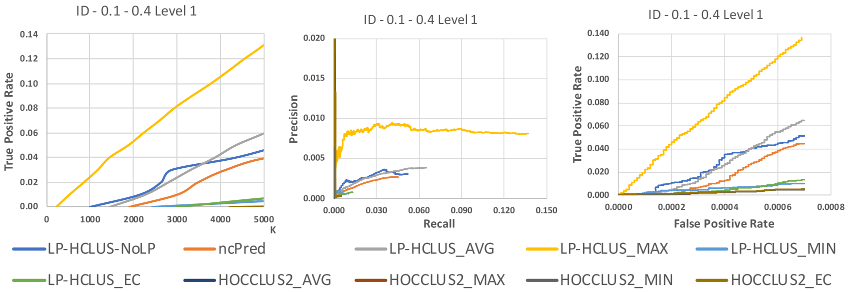


Fig. 5. TPR@k, Precision-Recall and ROC curves results for the dataset ID, obtained with the best configuration ($\alpha = 0.1, \beta = 0.4, level = 1$).

terms of TruePositiveRate@k curve. Moreover, we also report the results in terms of ROC and Precision-Recall curves by considering the unknown relationships as negative examples. We remark that ROC and PR curves can only be used for relative comparison and not as absolute evaluation measures because they are spoiled by the assumption made on unknown relationships.

In Figs. 4 and 5 we show some results obtained with the most promising configurations. From the quantitative viewpoint, we can observe that the proposed method LP-HCLUS, with the combination strategy based on the maximum, is able to obtain the best performances, for all the considered measures. From a qualitative point of view, we first performed a comparative analysis between the results obtained by LP-HCLUS against the validated interactions reported in the updated version of HMDD (i.e., v3.2 released on March 27th, 2019). We found 3,055 LP-HCLUS predictions confirmed by the new release of HMDD at the hierarchy level 1, 4,119 at level 2 and 4,797 at level 3. Next, we conducted a qualitative analysis of the top-ranked relationships predicted by LP-HCLUS using ID dataset, selecting only those with a score equal to 1.0. For this purpose, we exploited MNDR v2.0, which is a comprehensive resource including more than 260,000 experimental and predicted ncRNA-disease associations for mammalian species. Also in this case, we found some associations in both MNDR and in the list of predicted associations by LP-HCLUS. A more comprehensive analysis, reporting several additional examples, can be found in the full paper [2].

4 Conclusions

In this paper, we have tackled the problem of predicting possibly unknown ncRNA-disease relationships. The proposed approach LP-HCLUS is able to take advantage from the possible heterogeneous nature of the attributed biological network analyzed. The results confirm the initial intuitions and show competitive performances of LP-HCLUS in terms of accuracy of the predictions, also when compared with state-of-the-art competitor systems. These results are also supported by a comparison of LP-HCLUS predictions with data reported in MNDR and by a qualitative analysis that revealed that several ncRNA-disease associations predicted by LP-HCLUS have been subsequently experimentally validated and introduced in a more recent release (v3.2) of HMDD. As future work, we will evaluate the performance of LP-HCLUS in other domains.

5 Acknowledgments

We acknowledge the support of Ministry of Education, Universities and Research (MIUR) through the PON project TALIsMAn - Tecnologie di Assistenza personALizzata per il Miglioramento della quAlità della vitA (ARS01_01116). Dr. Gianvito Pio acknowledges the support of Ministry of Education, Universities and Research (MIUR) through the project AIM1852414, activity 1, line 1.

References

1. Alaimo, S., Giugno, R., Pulvirenti, A.: ncPred: ncRNA-Disease Association Prediction through Tripartite Network-Based Inference. Frontiers in Bioengineering and Biotechnology **2** (Dec 2014)
2. Barracchia, E.P., Pio, G., D'Elia, D., Ceci, M.: Prediction of new associations between ncrnas and diseases exploiting multi-type hierarchical clustering. BMC bioinformatics **21**(1), 1–24 (2020)
3. Bauer-Mehren, A., Rautschka, M., Sanz, F., Furlong, L.I.: DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. Bioinformatics (Oxford, England) **26**(22), 2924–2926 (Nov 2010)
4. Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., Cui, Q.: LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Research **41**(Database issue) (Jan 2013)
5. Chen, X., Yan, C.C., Luo, C., Ji, W., Zhang, Y., Dai, Q.: Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. Scientific Reports **5** (Jun 2015)
6. Han, J., Kamber, M.: Data mining: concepts and techniques. Elsevier/Morgan Kaufmann, Amsterdam (2006)
7. Helwak, A., Kudla, G., et al.: Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell **153**(3), 654–665 (2013)
8. Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., Liu, Y.: miR2disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Research **37**(Database issue), D98–104 (Jan 2009)
9. Lesmo, L., Saitta, L., Torasso, P.: Evidence combination in expert systems. International Journal of Man-Machine Studies **22**(3), 307–326 (Mar 1985)
10. Melissari, M.T., Grote, P.: Roles for long non-coding RNAs in physiology and disease. Pflügers Archiv - European Journal of Physiology **468**(6), 945–958 (2016)
11. Mignone, P., Pio, G., D'Elia, D., Ceci, M.: Exploiting transfer learning for the reconstruction of the human gene regulatory network. Bioinform. **36**(5), 1553–1561 (2020)
12. Pio, G., Ceci, M., D'Elia, D., Loglisci, C., Malerba, D.: A Novel Biclustering Algorithm for the Discovery of Meaningful Biological Correlations between microRNAs and their Target Genes. BMC Bioinformatics **14**(Suppl 7), S8 (Apr 2013)
13. Pio, G., Ceci, M., Prisciandaro, F., Malerba, D.: Exploiting causality in gene network reconstruction based on graph embedding. Machine Learning (2019)
14. Pio, G., Serafino, F., Malerba, D., Ceci, M.: Multi-type clustering and classification from heterogeneous networks. Information Sciences **425**, 107–126 (Jan 2018)
15. Wang, P., Guo, Q., et al.: Improved method for prioritization of disease associated lncRNAs based on ceRNA theory and functional genomics data. Oncotarget **8**(3), 4642–4655 (Dec 2016)