



# Identifying lncRNA-Disease Relationships via Heterogeneous Clustering

Emanuele Pio Barracchia<sup>1</sup> , Gianvito Pio<sup>1</sup>  , Donato Malerba<sup>1,2</sup> ,  
and Michelangelo Ceci<sup>1,2</sup> 

<sup>1</sup> Department of Computer Science, University of Bari Aldo Moro,  
Via Orabona, 4, 70125 Bari, Italy  
e.barracchia@studenti.uniba.it,

{gianvito.pio,donato.malerba,michelangelo.ceci}@uniba.it

<sup>2</sup> CINI - Consorzio Interuniversitario Nazionale per l'Informatica, Bari, Italy

**Abstract.** High-throughput sequencing technology led significant advances in functional genomics, giving the opportunity to pay particular attention to the role of specific biological entities. Recently, researchers focused on long non-coding RNAs (lncRNAs), i.e. transcripts that are longer than 200 nucleotides which are not transcribed into proteins. The main motivation comes from their influence on the development of human diseases. However, known relationships between lncRNAs and diseases are still poor and their in-lab validation is still expensive. In this paper, we propose a computational approach, based on heterogeneous clustering, which is able to predict possibly unknown lncRNA-disease relationships by analyzing complex heterogeneous networks consisting of several interacting biological entities of different types. The proposed method exploits overlapping and hierarchically organized heterogeneous clusters, which are able to catch multiple roles of lncRNAs and diseases at different levels of granularity. Our experimental evaluation, performed on a heterogeneous network consisting of microRNAs, lncRNAs, diseases, genes and their known relationships, shows that the proposed method is able to obtain better results with respect to existing methods.

## 1 Introduction

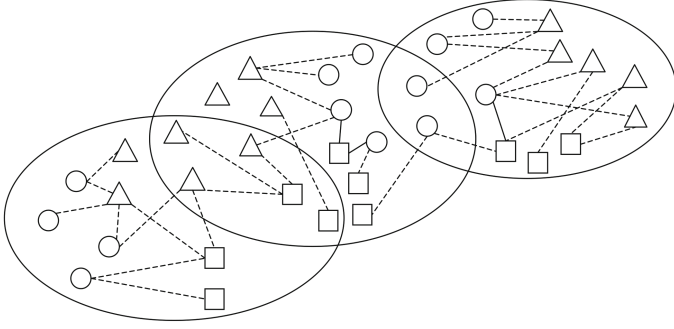
High-throughput sequencing technology, alongside new computational methods, has been crucial for rapid advances in functional genomics. Among the most important results achieved by exploiting these new technologies, there is the discovery of thousands of non-coding RNAs (ncRNAs). Since their function appears to be pivotal for the fine-tuning of the expression of many genes [3], in the last decade, the number of papers reporting evidences about ncRNAs involvement in human complex diseases, such as cancer, has grown at an exponential rate. Among the different classes of ncRNAs, the most investigated one is that of microRNAs (miRNAs), which are small molecules that regulate the expression of genes through the modulation of the translation of their transcripts [7]. Much less is known about the functional involvement of long non-coding RNAs

(lncRNAs), i.e. non-coding transcripts which are longer than 200 nucleotides, that have been recently discovered to have a plethora of regulatory functions [11]. However, the number of lncRNAs for which the functions are known is still quite poor and their in-lab validation requires large resources. Thus, assessing the role and, especially, the molecular mechanisms underlying the involvement of lncRNAs in human diseases, is not a trivial task.

In the last few years, there were some attempts to computationally predict the relationships among biological entities, such as genes, miRNAs, lncRNAs, diseases, tissues, etc. An example can be found in [14], where the authors propose an approach to learn to combine the outputs of several algorithms for the prediction of miRNA-gene interactions. A more sophisticated approach has been proposed in [4], where the authors adopt the multi-view learning framework for the reconstruction of gene-gene interaction networks.

Focusing on the identification of relationships involving diseases, in [16] the authors propose a method to identify possible relationships between lncRNAs and diseases, by exploiting a bipartite network and a propagation algorithm. Analogously, in [1] the authors propose the method *ncPred* which exploits a tripartite graph representing known ncRNA-gene and gene-disease associations. Such a graph is analyzed by adopting a multi-level resource transfer technique that, at each step, takes into account the resource transferred in the previous one. For each detected interaction, the algorithm associates a score indicating its degree of certainty. Both these methods, however, cannot exploit additional information associated with the involved biological entities as well as other entities that are related to the considered ones (e.g., genes, miRNAs, tissues, etc.).

In this paper, we present a novel method for the identification of previously unknown relationships between diseases and lncRNAs, which extends the heterogeneous clustering approach we proposed in [15]. In particular, the proposed method is able to identify heterogeneous clusters from heterogeneous networks, where nodes are biological entities (each associated with their own features) and edges represent known relationships among them (see Fig. 1). Then, the identified clusters are exploited to predict the possible existence of unknown relationships between lncRNAs and diseases falling in the same clusters. This approach is motivated by the fact that lncRNAs and diseases will fall in the same clusters if they appear similar according to their features and their relationships with the other analyzed entities. Therefore, the main advantage of the approach proposed in this paper comes from its ability to globally take into account the complex network of interactions involving different biological entities. Moreover, the proposed algorithm has the advantage of identifying possibly overlapping and hierarchically organized clusters, since (i) the same lncRNA/disease can be involved in multiple networks of relationships and (ii) as shown in [12], clusters at different levels of the hierarchy can describe more specific or more general relationships and cooperation activities. In the following section, we briefly describe our clustering method and its exploitation to identify unknown lncRNA-disease relationships, while in Sect. 3 we report the results of our experiments. Finally, in Sect. 4, we draw some conclusions and outline the ongoing work.



**Fig. 1.** An example of a heterogeneous network, where different shapes represent different node types. Circles represent possible heterogeneous clusters.

## 2 Method

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

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

- each node  $v' \in V$  is associated to a single node type  $t_v(v') \in \mathcal{T}$ , where  $\mathcal{T}$  is the finite set  $\{T_p\}$  of all the possible types of nodes in the network;
- each node type  $T_p$  implicitly defines a subset of nodes  $V_p \subseteq V$ ;
- a node type  $T_p$  defines a set of attributes  $\mathcal{X}_p = \{X_{p,1}, X_{p,2}, \dots, X_{p,m_p}\}$ ;
- an edge  $e$  between two nodes  $v'$  and  $v''$  is associated to an edge type  $R_j \in \mathcal{R}$ , where  $\mathcal{R}$  is the finite set  $\{R_j\}$  of all the possible edge types in the network. Formally,  $e = \langle R_j, \langle v', v'' \rangle \rangle \in E$ , where  $R_j = t_e(e) \in \mathcal{R}$  is its edge type;
- an edge type  $R_j$  defines a subset of edges  $E_j \subseteq (V_p \times V_q) \subseteq E$ ;
- node types  $\mathcal{T}$  are partitioned into  $\mathcal{T}_t$  (target), i.e. considered as target of the clustering/prediction task, and  $\mathcal{T}_{tr}$  (task-relevant). Only nodes of target types are actually clustered and considered in the identification of new relationships, on the basis of all the nodes.

**Definition 2 (Heterogeneous cluster).** We define a heterogeneous cluster, or multi-type cluster, as  $G' = (V', E')$ , where:  $V' \subseteq V$ ;  $\forall v' \in V', t_v(v') \in \mathcal{T}_t$  (nodes in the clusters are only of target types);  $E' \subseteq (E \cup \hat{E})$  is a set of edges (among the nodes in  $V'$ ) belonging either to  $E$  or to a set of edges  $\hat{E}$  containing *extracted* edges, which relate nodes that are not directly connected in the original network.

**Definition 3 (Hierarchical organization of clusters).** A hierarchy of heterogeneous clusters is defined as a list of hierarchy levels  $\{L_1, L_2, \dots, L_k\}$ , each of which consisting of a set of heterogeneous and possibly overlapping clusters.

In this specific application domain, target nodes are those representing lncRNAs and diseases. Therefore, we distinguish two distinct sets of nodes  $T_l$  and  $T_d$ , representing the set of lncRNAs and the set of diseases, respectively. Our task then consists in the identification of a hierarchy of clusters  $\{L_1, L_2, \dots, L_k\}$  and of a function  $\psi^{(w)} : T_l \times T_d \rightarrow [0, 1]$  for each hierarchy level  $L_w$ , which, for each lncRNA-disease pair, returns a score indicating its degree of certainty.

In the following, we describe our solution consisting of three steps: (i) identification of the strength of relationships among nodes in  $T_l$  and  $T_d$ , which will define the set of extracted edges  $\hat{E}$ ; (ii) construction of a hierarchy of (possibly overlapping) heterogeneous clusters; (iii) identification of the functions  $\psi^{(w)}$  for the prediction of previously unknown relationships.

## 2.1 Identification of the Strength of the Relationship Among Nodes

We first estimate the strength of the relationship of all the possible lncRNA-disease pairs, following the idea we proposed in [15]: for each pair  $(l_i, d_j)$ , we compute the score  $s(l_i, d_j)$  by analyzing the indirect relationships in which the lncRNA  $l_i$  and the disease  $d_j$  are involved. In particular, as in [15], we adopt the concept of *meta-path*, i.e., the set of sequences of nodes which follow the same sequence of edge types. For each meta-path  $P$  between  $l_i$  and  $d_j$ , we compute a score  $pathscore(P, l_i, d_j)$  representing the strength between  $l_i$  and  $d_j$  following the meta-path  $P$ . Since several meta-paths can be identified between two objects in the network, possibly with unlimited length (in presence of cycles), we have to identify a strategy to assign a single score to each lncRNA-disease pair. The strategy we considered is inspired by the classical formulation of fuzzy sets [17]. In particular, since  $s(l_i, d_j)$  should measure the degree of certainty of the relationship between  $l_i$  and  $d_j$ , we consider the scores computed over each meta-path  $P$  (i.e.,  $pathscore(P, l_i, d_j)$ ) as the degree of certainty estimated according to  $P$ . Since the relationship between  $l_i$  and  $d_j$  can be considered certain if there exists at least one meta-path which proves its certainty (or, in other words, the certainty of the relationship corresponds to the highest certainty showed over the meta-paths), we compute  $s(l_i, d_j)$  as follows:

$$s(l_i, d_j) = \max_{P \in \text{metapaths}(l_i, d_j)} pathscore(P, l_i, d_j) \quad (1)$$

where  $\text{metapaths}(l_i, d_j)$  is the set of the  $c$  shortest paths connecting  $l_i$  and  $d_j$ , and  $pathscore(P, l_i, d_j)$  is the degree of certainty of the relationship between  $l_i$  and  $d_j$  according to the meta-path  $P$ .

In order to compute  $pathscore(P, l_i, d_j)$ , we represent each meta-path  $P$  as a finite set of sequences of nodes. If a sequence in  $P$  connects  $l_i$  and  $d_j$ , then  $pathscore(P, l_i, d_j) = 1$ . Otherwise, following the same strategy introduced before, it is computed as the maximum similarity between the sequences which start with  $l_i$  and the sequences which end with  $d_j$  (see Fig. 2).

The similarity between two sequences  $seq'$  and  $seq''$  is computed according to the attributes of all the nodes involved in the two sequences. Following [6], the similarity between two values of an attribute  $x$ , i.e.,  $s_x(seq', seq'')$ ,

#Seq	disease_id	disease_att1	disease_att2	lncRNA_id	lncRNA_att1	lncRNA_att2
1	d <sub>1</sub>	0.5	a	l <sub>2</sub>	0.9	x
2	d <sub>4</sub>	0.5	c	l <sub>3</sub>	0.5	y
3	d <sub>3</sub>	0.6	c	l <sub>2</sub>	0.9	x
4	d <sub>3</sub>	0.6	c	l <sub>1</sub>	0.3	x
5	d <sub>3</sub>	0.6	c	l <sub>8</sub>	0.8	z
6	d <sub>2</sub>	0.3	b	l <sub>3</sub>	0.5	y
7	d <sub>5</sub>	0.1	b	l <sub>1</sub>	0.7	y

**Fig. 2.** An example of analysis of the sequences associated to the lncRNA  $l_3$  and to the disease  $d_3$ . In the example, sequences 2 and 6 (in yellow) are associated to the lncRNA  $l_3$ , and sequences 3, 4 and 5 (in green) are associated to the disease  $d_3$ . The algorithm pair-wisely compares the two sets of sequences (sequences in yellow and sequences in green) and computes the degree of certainty between  $l_3$  and  $d_3$  as the maximum similarity between two sequences. (Color figure online)

is computed as follows: If  $x$  is a numerical attribute, then  $s_x(seq', seq'') = 1 - \frac{|val_x(seq') - val_x(seq'')|}{max_x - min_x}$  ( $min_x$  and  $max_x$  are the minimum and maximum values, respectively, observed for the attribute  $x$ ); when  $x$  is not numeric, then  $s_x(seq', seq'') = 1$  if  $val_x(seq') = val_x(seq'')$ , 0 otherwise.

It is noteworthy that some node types may not be involved in any meta-path. In order to exploit the information conveyed by these nodes, we add an aggregation of their attribute values to the nodes that are connected to them and that appear in at least one meta-path. Such an aggregation considers values coming from directly or indirectly (up to a predefined *depth* of analysis) connected nodes. For this purpose different aggregation functions could be used. Following [15], we use the *arithmetic mean* for numerical attributes, the *mode* for non-numerical attributes and limit the depth of analysis for the aggregation to 2.

## 2.2 Construction of the Hierarchy of Heterogeneous Clusters

Once all the possible pairs are identified, each associated with its degree of certainty, we first build a set of (possibly overlapping) clusters in the form of bicliques to be used in the subsequent step. A cluster is in the form of a biclique if all the lncRNA-disease pairs in the cluster have a score above a given threshold  $\beta \in [0, 1]$ . The algorithm consists of the following steps:

- (i) A filtering phase which keeps only the pairs with a score greater than (or equal to)  $\beta$ . The result is the subset of pairs  $\{(l_i, d_j) | s(l_i, d_j) \geq \beta\}$ .
- (ii) An initialization step which identifies the initial set of bicliques, each consisting of a lncRNA-disease pair in  $\{(l_i, d_j) | s(l_i, d_j) \geq \beta\}$ .
- (iii) A process that iteratively merges two clusters  $G'$  and  $G''$  into a new cluster  $G'''$ . The initial set of clusters is regarded as a list and is sorted according to an ordering relation  $<_c$  that reflects the quality of the clusters. Each cluster  $G'$  is merged with the first cluster  $G''$  in the list leading to a merged cluster

$G'''$  which still is a biclique. This step is repeated until no more merging can be performed. The obtained result is the first hierarchy level  $L_1$ .

The ordering relation  $<_c$  is based on the cluster *cohesiveness*, defined as:

$$h(G) = \frac{1}{|pairs(G)|} \cdot \sum_{(l_i, d_j) \in pairs(G)} s(l_i, d_j) \quad (2)$$

where  $pairs(G)$  is the set of all the possible lncRNA-disease pairs (both known and unknown) in the cluster.

This measure actually corresponds to the average score of the relationships in the cluster. Since, in our case, the score represents a degree of certainty, the cluster cohesiveness can be considered as an indicator of the degree of certainty of the global interactions among the group of lncRNAs and diseases in the cluster. Therefore, we formally define the ordering relation  $<_c$  as follows:

$$G' <_c G'' \iff h(G') > h(G'') \quad (3)$$

Once the first level  $L_1$  of the hierarchy has been identified, the other levels are built by evaluating whether some pairs of clusters (bicliques, in  $L_1$ ) can be reasonably merged. The approach is similar to that used to obtain the first level of the hierarchy. The main difference is that, instead of working on bicliques, we work on generic clusters, where the score associated to each pair is not necessarily greater than  $\beta$ . Due to this difference, we use a different criterion for the identification of candidates for merging which is inspired by the research in hierarchical co-clustering. In this research, one of the commonly used stopping criterion is based on a predefined threshold applied to the quality constraint that must be satisfied in order to merge two clusters [12]. Analogously, in our approach, two clusters  $G'$  and  $G''$  are merged into a cluster  $G'''$  if  $h(G''') > \alpha$ , where  $\alpha$  is a user defined threshold on the cluster cohesiveness. Note that low values of  $\alpha$  lead to a higher number of mergings and, accordingly, to less clusters containing a higher number of objects.

We repeat the process until no merging is possible and return the obtained hierarchy of heterogeneous clusters  $\{L_1, L_2, \dots, L_k\}$ , according to Definition 3.

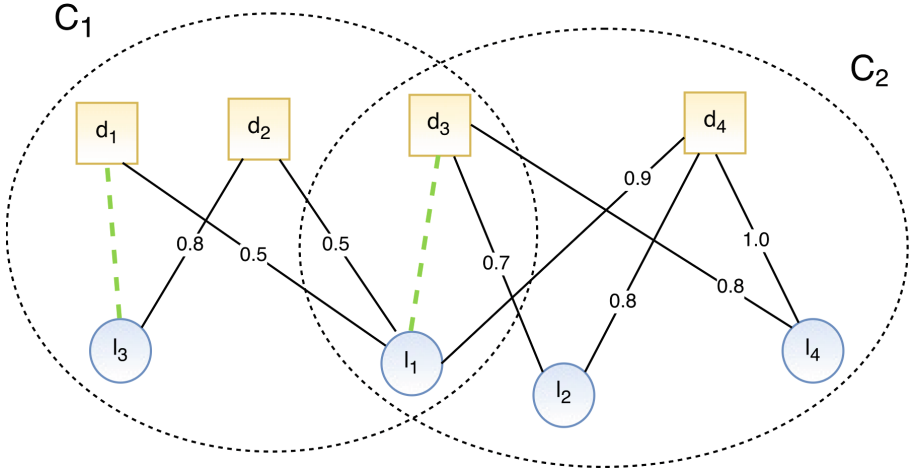
### 2.3 Prediction of Unknown Relationships

After building the hierarchy of clusters, we identify possibly unknown relationships for each hierarchical level. In particular, the prediction is performed by assigning each possible lncRNA-disease pair with a degree of certainty computed as the cohesiveness of the cluster in which it falls. More formally, given  $C_{ij}^{(w)}$  the cluster in which the lncRNA  $l_i$  and the disease  $d_j$  fall in the  $w$ -th hierarchical level, we compute the final degree of certainty of the relationship as:

$$\psi^{(w)}(l_i, d_j) = h\left(C_{ij}^{(w)}\right). \quad (4)$$

When the lncRNA  $l_i$  and the disease  $d_j$  appear in multiple clusters, i.e.,  $C_{ij}^{(w)}$  is a list of clusters, we combine their cohesiveness to obtain the final degree of certainty. Baseline combination strategies can be the maximum, the minimum and the average. In this work, we propose to adopt a different combination function, which rewards those cases in which the pair appears in several highly cohesive clusters (indicating a higher degree of certainty). In details, inspired by the evidence combination (EC) strategy proposed in [10], given  $C_{ij}^{(w)} = [C_1, C_2, \dots, C_m]$ , the list of the clusters in which the lncRNA  $l_i$  and the disease  $d_j$  fall in the  $w$ -th hierarchical level, we compute  $\psi^{(w)}(l_i, d_j) = ec(C_m)$ , where  $ec(C_m)$  is recursively defined as:

$$ec(C_m) = \begin{cases} h(C_1) & \text{if } C_m = C_1 \\ ec(C_{m-1}) + [1 - ec(C_{m-1})] \cdot h(C_m) & \text{otherwise} \end{cases} \quad (5)$$



**Fig. 3.** Two possible clusters identified at a given hierarchical level  $w$ . Circles represent lncRNAs, while squares represent diseases. The clusters suggest two new possible relationships between  $l_3$  and  $d_1$  and between  $l_1$  and  $d_3$ .

In Fig. 3, we show an example of the prediction step, where the two clusters  $C_1$  and  $C_2$ , identified at the  $w$ -th hierarchical level, suggest two potential new relationships, i.e., between  $l_3$  and  $d_1$  and between  $l_1$  and  $d_3$ . The former falls only in the cluster  $C_1$ , therefore it will be associated with a degree of certainty computed according to the cohesiveness of  $C_1$ . Formally:

$$\psi^{(w)}(l_3, d_1) = h(C_1) = \frac{1}{3 \cdot 2} (0.8 + 0.5 + 0.5) = 0.3.$$

The latter falls in both  $C_1$  and  $C_2$  and its degree of certainty will be computed according to the cohesiveness of both clusters. In particular, given  $h(C_1) = 0.3$

and  $h(C_2) = \frac{1}{3 \cdot 2}(0.7 + 0.8 + 0.9 + 0.8 + 1.0) = 0.7$ , by adopting the EC strategy (see Eq. 5), the degree of certainty of the relationship between  $l_1$  and  $d_3$  will be computed as:

$$\psi^{(w)}(l_1, d_3) = h(C_1) + [1 - h(C_1)] \cdot h(C_2) = 0.3 + (1 - 0.3) \cdot 0.7 = 0.79$$

### 3 Experiments

The proposed method has been implemented in the system LP-HCLUS (Link Prediction through Heterogeneous CLUstering). We performed our experimental evaluation in order to evaluate the effectiveness of the proposed approach on a complex biological dataset containing data about lncRNAs, miRNAs, genes and diseases, as well as their known interactions and relationships. Such a dataset, whose schema is depicted in Fig. 4, has been built by integrating several existing biological datasets:

- lncRNA-disease relationships and lncRNA-gene interactions from [5];
- miRNA-lncRNA interactions from [8];
- disease-gene relationships from DisGeNET [2];
- miRNA-gene and miRNA-disease relationships from miR2Disease [9].

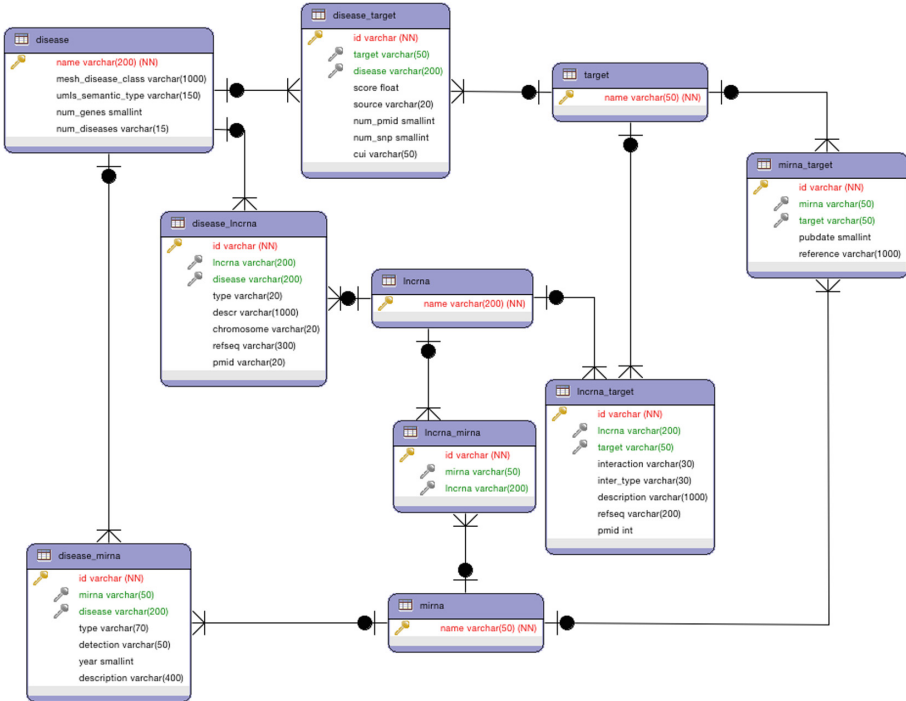


Fig. 4. UML representation of the heterogeneous network used in the experiments.



The integrated dataset consists of 7050 diseases, 507 lncRNAs, 508 miRNAs, 94527 genes, 953 interactions between diseases and lncRNAs, 2877 interactions between diseases and miRNAs, 26522 interactions between diseases and genes, 70 interactions between lncRNAs and miRNAs, 252 interactions between lncRNAs and genes, and 803 interactions between miRNAs and genes.

We adopted the 10-fold cross validation on the set of known lncRNA-disease relationships. Due to the absence of negative examples, following the approach adopted in [13], we averaged the results obtained in terms of  $recall@K$ , i.e., the recall measured by considering only the top- $k$  returned relationships. In detail, we produce a ranking of the predicted interactions by sorting them in descending order with respect to their degree of certainty and compute  $recall@K = \frac{TP_k}{TP_k + FN_k}$ , where  $TP_k$  (respectively,  $FN_k$ ) is the number of validated lncRNA-disease relationships that were (respectively, were not) predicted in the first  $K$  returned interactions. Since the most appropriate value of  $K$  cannot be known in advance, we plot the obtained  $recall@K$  by varying the value of  $K$ .

LP-HCLUS has been run by considering 3 shortest paths for each lncRNA-disease pair (i.e.,  $c = 3$ ). We collected the results obtained with the maximum (MAX), the minimum (MIN), the average (AVG) and the evidence combination (EC) strategies to combine the degree of certainty of relationships identified in multiple clusters, focusing on the first 3 levels of the identified hierarchies which, according to [12], lead to the best results.

As competitor systems, we considered the following approaches:

- A variant of the biclustering algorithm **HOCCLUS2** [12], which is able to solve the link prediction task. HOCCLUS2 has similar characteristics with respect to the clustering approach proposed in this paper, i.e., it is able to extract a hierarchy of (possibly overlapping) clusters. However, it does not allow to take into account several types of objects, linked by several types of edges. Moreover, the algorithm adopted for the construction of the hierarchy of clusters is different and guided only by the cohesiveness.
- The link prediction algorithm **ncPred** [1], which is tailored for the prediction of ncRNA-disease associations.
- A baseline approach, which consists in the estimation of the degree of certainty by means of the strategy described in Sect. 2.1, i.e., without the clustering and the prediction steps. The comparison of the results with respect to this baseline approach allows us to evaluate the real contribution of the exploitation of clusters for link prediction. We call this baseline approach **LP-HCLUS w/o LP** (i.e., LP-HCLUS without Link Prediction).

We fed all the competitor methods with the set of lncRNA-disease scores computed by LP-HCLUS, since, in their original form, they are not able to analyze a complex heterogeneous network.

Since both HOCCLUS2 and LP-HCLUS require the input parameters  $\alpha$  and  $\beta$ , we performed some preliminary experiments to evaluate their effect on the results. In particular, we evaluated the results with the following configurations:

$\alpha = 0.1$  and  $\beta = 0.3$ ;  $\alpha = 0.1$  and  $\beta = 0.4$ ;  $\alpha = 0.2$  and  $\beta = 0.3$ ;  $\alpha = 0.2$  and  $\beta = 0.4$ . By observing the results reported in Fig. 5, obtained with the EC strategy on the first three hierarchical levels, we can conclude that the results do not appear to be significantly affected by these parameters. A similar behavior was observed for the other combination strategies and for HOCCLUS2. However, since the obtained *recall@K* results appeared higher in the case of  $\alpha = 0.2$  and  $\beta = 0.4$ , the other experiments were conducted with such values.

In Table 1, we report the results in terms of *recall@K* obtained by the considered approaches, with  $K \in \{500, 1000, \dots, 5000\}$ . The first conclusion that can be drawn regards the superiority of LP-HCLUS with respect to the considered competitor approaches, with all the values of  $K$ . This conclusion is even more evident for small values of  $K$ , i.e., in the first part of the ranked interactions. Moreover, by comparing the results obtained by LP-HCLUS with the baseline approach (LP-HCLUS w/o LP) we can observe a significant improvement when the clustering and the link prediction phases are adopted. This means that the strategy proposed in this paper, i.e., the identification of heterogeneous clusters and their exploitation for link prediction purposes, appears to be effective. Moreover, although the adopted variant of HOCCLUS2 is based on the same principle, it still leads to a lower *recall@K* results, emphasizing that the clustering algorithm and the adopted combination strategies proposed in this paper perform better. A further observation comes from the comparison of the results obtained by LP-HCLUS with different combination strategies. Indeed, by observing Table 1, we can conclude that the strategy based on evidence combination (EC) generally leads to the best results, especially for high values of  $K$ . This is mainly due to the fact that it rewards the interactions falling in multiple highly-cohesive clusters. This means that, on overall, predicted interactions have a higher degree of certainty (also higher than the those predicted with the strategy based on MAX), leading to a higher recall with high values of  $K$ .

A more global overview is provided in Fig. 6, where we plot the *recall@K* results of all the considered approaches, at different levels of the hierarchy. This figure shows the overall superiority of LP-HCLUS, when the strategy based on evidence combination is adopted. Moreover, it also shows that the competitors (i.e., ncPred and HOCCLUS2) and the baseline method cannot reach the recall values obtained by LP-HCLUS (very close to 1.0) even with  $K = 70,000$ .

A final consideration comes from the analysis of the results at different levels of the hierarchy. At this respect, we could not find a general trend in the results, in terms of *Recall@K*. However, such a measure is only able to evaluate the results quantitatively, and a deeper analysis could be necessary in order to emphasize possible differences from a qualitative viewpoint. Therefore, since we still believe that the hierarchy can be fruitfully exploited to emphasize interactions at different levels of granularity, in future works we will involve a domain expert in the analysis of results in order to evaluate qualitatively whether this idea appears confirmed by real biological findings.

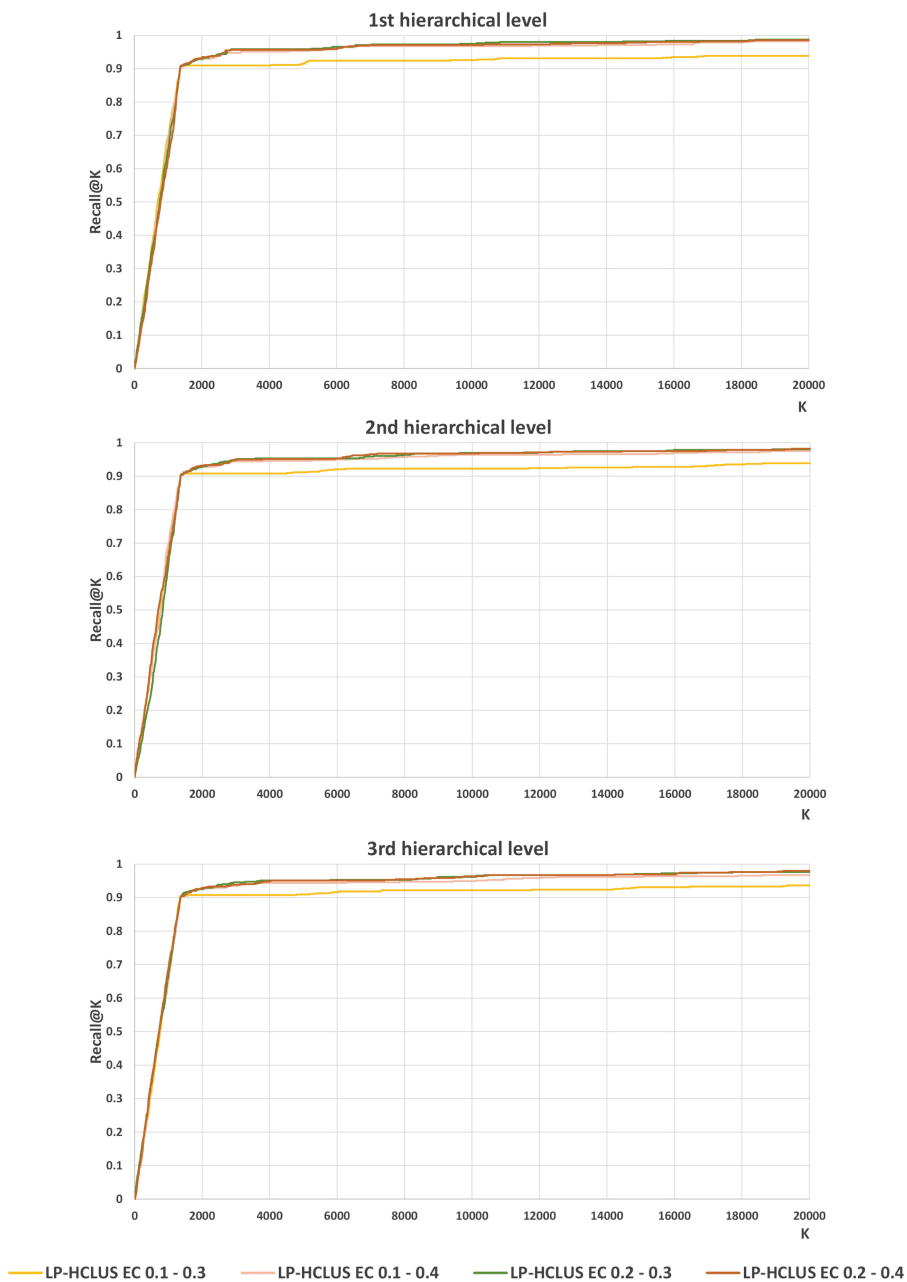
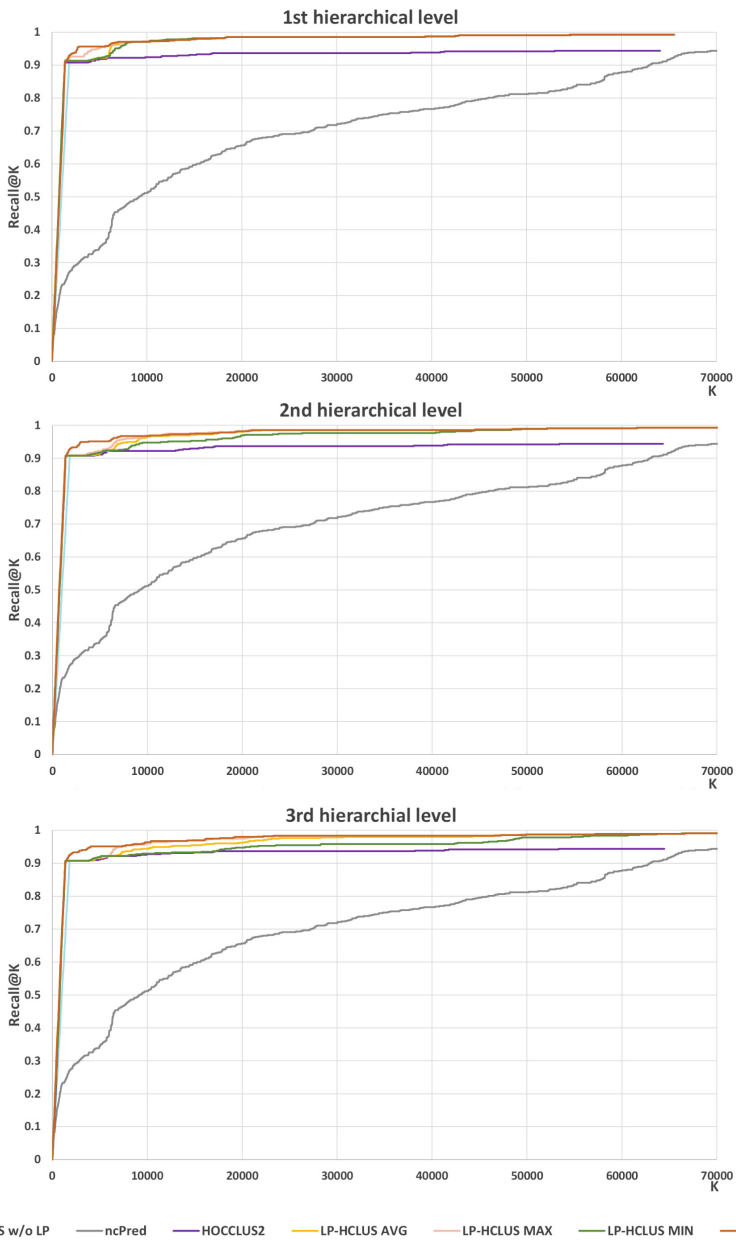


Fig. 5. Recall@K obtained by LP-HCLUS EC with different values of  $\alpha$  and  $\beta$ .



**Fig. 6.** *Recall@K* results obtained by LP-HCLUS ( $\alpha = 0.2$ ,  $\beta = 0.4$ ) and by the considered competitor methods.

**Table 1.** *Recall@K* obtained by LP-HCLUS and by the competitors for different values of *K*. Results obtained by HOCCLUS2 and LP-HCLUS have been collected for the first 3 levels of the hierarchies. For each value of *K*, the best result is highlighted in bold.

	Value of K									
	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
LP-HCLUS w/o LP	0.242	0.503	0.750	0.908	0.908	0.908	0.908	0.908	0.910	0.913
ncPred	0.154	0.228	0.248	0.275	0.293	0.304	0.316	0.325	0.336	0.344
11_HOCCLUS2	0.329	0.658	0.908	0.908	0.908	0.908	0.908	0.910	0.913	0.919
11_LP-HCLUS AVG	0.349	0.665	0.913	0.913	0.913	0.913	0.913	0.915	0.919	0.922
11_LP-HCLUS MAX	0.340	0.644	<b>0.915</b>	0.926	0.926	0.926	0.933	0.942	0.948	0.951
11_LP-HCLUS MIN	0.331	0.683	0.913	0.913	0.913	0.913	0.913	0.915	0.919	0.922
11_LP-HCLUS EC	0.311	0.629	0.913	<b>0.933</b>	<b>0.939</b>	<b>0.957</b>	<b>0.957</b>	<b>0.957</b>	<b>0.957</b>	<b>0.957</b>
12_HOCCLUS2	0.324	0.676	0.908	0.908	0.908	0.908	0.908	0.908	0.910	0.911
12_LP-HCLUS AVG	<b>0.362</b>	0.658	0.908	0.908	0.908	0.908	0.908	0.910	0.910	0.915
12_LP-HCLUS MAX	0.336	<b>0.691</b>	0.908	0.910	0.910	0.910	0.911	0.917	0.920	0.922
12_LP-HCLUS MIN	0.329	0.656	0.908	0.908	0.908	0.908	0.908	0.911	0.913	0.917
12_LP-HCLUS EC	0.349	0.662	0.910	0.931	0.933	0.949	0.949	0.951	0.951	0.951
13_HOCCLUS2	0.304	0.669	0.908	0.908	0.908	0.908	0.908	0.910	0.910	0.911
13_LP-HCLUS AVG	0.353	0.680	0.908	0.908	0.908	0.908	0.908	0.908	0.911	0.917
13_LP-HCLUS MAX	0.329	0.680	0.908	0.908	0.908	0.908	0.908	0.910	0.915	0.917
13_LP-HCLUS MIN	0.318	0.667	0.908	0.908	0.908	0.908	0.908	0.910	0.915	0.919
13_LP-HCLUS EC	0.358	0.689	0.910	0.930	0.933	0.939	0.940	0.949	0.951	0.951

## 4 Conclusions

In this work, we proposed the method LP-HCLUS, which is able to analyze heterogeneous biological networks in order to identify (possibly overlapping) hierarchically organized clusters and to exploit them to predict possibly unknown lncRNA-disease relationships. Such findings can be exploited for better understanding the role of lncRNAs in the development of human diseases. We also proposed the adoption of a specific strategy, based on evidence combination, to aggregate the different degrees of certainty when a new lncRNA-disease relationship is suggested by multiple clusters. Experiments performed on an integrated biological dataset showed that the proposed method, especially when adopting the strategy based on evidence combination, is able to outperform the methods HOCCLUS2 and ncPred, as well as a baseline approach. As future work, we intend to perform additional experiments on large-scale networks, possibly exploiting a distributed variant of the method proposed in this paper. Moreover, we will perform a qualitative evaluation, from a biological point of view, of the real contribution provided by our computational approach in the identification of lncRNA-disease relationships.

**Acknowledgements.** We would like to acknowledge the support of the European Commission through the projects MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant Number ICT-2013-612944) and TOREADOR - Trustworthy Model-aware Analytics Data Platform (Grant Number H2020-688797).

## References

- Alaimo, S., Giugno, R., Pulvirenti, A.: ncPred: ncRNA-disease association prediction through tripartite network-based inference. *Front. Bioeng. Biotechnol.* **2**, 71 (2014)
- Bauer-Mehren, A., Rautschka, M., Sanz, F., Furlong, L.I.: DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics* **26**(22), 2924–2926 (2010)
- Cech, T., Steitz, J.: The noncoding RNA revolution—trashing old rules to forge new ones. *Cell* **157**(1), 77–94 (2014)
- Ceci, M., Pio, G., Kuzmanovski, V., Dzeroski, S.: Semi-supervised multi-view learning for gene network reconstruction. *PLOS ONE* **10**(12), 1–27 (2015)
- 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 Res.* **41**(D1), D983–D986 (2013)
- Han, J., Kamber, M., Pei, J.: *Data Mining: Concepts and Techniques*, 2nd edn. Morgan Kaufmann, San Francisco (2006)
- Hayes, J., Peruzzi, P.P., Lawler, S.: MicroRNAs in cancer: biomarkers, functions and therapy. *Trends Mol. Med.* **20**(8), 460–469 (2014)
- Helwak, A., Kudla, G., Dudnakova, T., Tollervey, D.: Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153**(3), 654–665 (2013)
- 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 Res.* **37**(suppl 1), D98–D104 (2009)
- Lesmo, L., Saitta, L., Torasso, P.: Evidence combination in expert systems. *Int. J. Man-Mach. Stud.* **22**(3), 307–326 (1985)
- Melissari, M.T., Grote, P.: Roles for long non-coding RNAs in physiology and disease. *Pflügers Archiv - Eur. J. Physiol.* **468**(6), 945–958 (2016)
- 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**(S-7), S8 (2013)
- Pio, G., Ceci, M., Malerba, D., D’Elia, D.: ComiRNet: a web-based system for the analysis of miRNA-gene regulatory networks. *BMC Bioinformatics* **16**(9), S7 (2015)
- Pio, G., Malerba, D., D’Elia, D., Ceci, M.: Integrating microRNA target predictions for the discovery of gene regulatory networks: a semi-supervised ensemble learning approach. *BMC Bioinformatics* **15**(1), S4 (2014)
- Pio, G., Serafino, F., Malerba, D., Ceci, M.: Multi-type clustering and classification from heterogeneous networks. *Inf. Sci.* **425**, 107–126 (2018)
- Yang, X., Gao, L., Guo, X., Shi, X., Wu, H., Song, F., et al.: A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLOS ONE* (2014)
- Zadeh, L.: Fuzzy sets. *Inf. Control* **8**(3), 338–353 (1965)