Barracchia *et al. BMC Bioinformatics* (2020) 21:70
https://doi.org/10.1186/s12859-020-3392-2

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Prediction of new associations between ncRNAs and diseases exploiting multi-type hierarchical clustering

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

Abstract

Background: 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 whose expression also depends on endogenous or exogenous factors. 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 (i.e., consisting of only one type of objects and relationships) or can exploit only a small subset of the features of biological entities, such as the presence of a particular binding domain, enzymatic properties or their involvement in specific diseases.

Results: To overcome the limitations of existing approaches, we propose the system LP-HCLUS, which exploits a multi-type hierarchical clustering method to predict possibly unknown ncRNA-disease relationships. In particular, LP-HCLUS analyzes heterogeneous networks consisting of several types of objects and relationships, each possibly described by a set of features, and extracts multi-type clusters that are subsequently exploited to predict new ncRNA-disease associations. The extracted clusters are overlapping, hierarchically organized, involve entities of different types, and allow LP-HCLUS to catch multiple roles of ncRNAs in diseases at different levels of granularity. Our experimental evaluation, performed on heterogeneous attributed networks consisting of microRNAs, lncRNAs, diseases, genes and their known relationships, shows that LP-HCLUS is able to obtain better results with respect to existing approaches. The biological relevance of the obtained results was evaluated according to both quantitative (i.e., TPR@k, Areas Under the TPR@k, ROC and Precision-Recall curves) and qualitative (i.e., according to the consultation of the existing literature) criteria.

Conclusions: The obtained results prove the utility of LP-HCLUS to conduct robust predictive studies on the biological role of ncRNAs in human diseases. The produced predictions can therefore be reliably considered as new, previously unknown, relationships among ncRNAs and diseases.

Keywords: Non-coding RNA (ncRNAs), Diseases, Cancer, Heterogeneous network, Clustering, Link prediction

Background

High-throughput sequencing technologies, together with 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 [1]. In parallel, the number of studies reporting the involvement of ncRNAs in the development of many different human diseases has grown exponentially [2]. 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 the other main category represented by long non-coding RNAs (lncRNAs), that are ncRNAs longer than 200nt [3, 4].

*Correspondence: gianvito.pio@uniba.it
[1]University of Bari Aldo Moro - Department of Computer Science, Via Orabona, 4, 70125 Bari, Italy
Full list of author information is available at the end of the article

Long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) [5] are among the largest and heterogeneous groups of regulators of major cellular processes. However, lncRNAs, differently from miRNAs which primarily act as post-transcriptional regulators, have a plethora of regulatory functions [6]. They are involved in chromatin remodeling and epigenetic modifications, and organize functionally different nuclear sub-compartments with an impact on the nuclear architecture [7]. LncRNAs are also involved in the regulation of the expression of transcripts at cytoplasmic level by another series of interactions/functions that interfere with the efficiency of translation of transcripts in their protein products. In particular, they can directly interfere with miRNAs functions acting as miRNA sponges [8]. Nevertheless, the number of lncRNAs for which the functional and molecular mechanisms are completely elucidated is still quite poor. This is due to two main reasons: their recent discovery as master regulators with respect to miRNAs, and some particular features, such as the low cross-species conservation, the low expression levels and the high tissue specificity that make their characterization or any type of generalization still very difficult [9]. Therefore, assessing the role and the molecular mechanisms underlying the involvement of lncRNAs in human diseases is not a trivial task, 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, etc. [10–19]. Such methods are mainly 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 the existing approaches are able to work only on homogeneous networks (where nodes and links are of one single type) [20], are strongly limited by the number of different node types or are constrained by a pre-defined network structure. To overcome these limitations we propose the method LP-HCLUS (Link Prediction through Hierarchical CLUStering), which can discover previously unknown ncRNA-disease relationships working on heterogeneous attributed networks (that is, networks composed of different biological entities related by different types of relationships) with arbitrary structure. This ability allows LP-HCLUS to investigate how different types of entities interact with each other, possibly leading to increased prediction accuracy. LP-HCLUS exploits a combined approach based on hierarchical, multi-type clustering and link prediction. As we will describe in 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.

Methodologically, LP-HCLUS estimates an initial score for each possible relationship involving entities belonging to the types of interest (in our case, ncRNAs and diseases), by exploiting the whole network. Such scores are then used to identify a hierarchy of overlapping multi-type clusters, i.e., groups of objects of different types. Finally, the identified clusters are exploited to predict new relationships, each of which is associated with a score representing its degree of certainty. Therefore, according to the classification provided in [21] (see Additional file 1), LP-HCLUS simultaneously falls in two categories: *i) algorithmic* methods, since it strongly relies on a clustering approach to predict new relationships and to associate them with a score in $[0, 1]$, and *ii) similarity-based* approaches, since the first phase (see "Estimation of the strength of the relationship between ncRNAs and diseases" section) exploits the computation of similarities between target nodes, taking into account the paths in the network and the attributes of the nodes.

The rest of the paper is organized as follows: in the next section, we describe our method for the identification of new ncRNA-disease relationships; in "Results" section we describe our experimental evaluation and in "Discussion" section we discuss the obtained results, including a qualitative analysis of the obtained predictions; finally, we conclude the paper and outline some future work. Moreover, in Additional file 1, we discuss the works related to the present paper; in Additional file 2 we report an analysis of the computational complexity of the proposed method; finally, in Additional files 3, 4 and 5 we report some detailed results obtained during the experiments.

Methods

The algorithmic approach followed by LP-HCLUS mainly relies on the predictive clustering framework [22–24]. The motivation behind the adoption of such a framework comes from its recognized ability of handling data affected by different forms of autocorrelation, i.e., when close objects (spatially, temporally, or in a network as in this work) appear to be more similar than distant objects. This peculiarity allows LP-HCLUS to catch multiple dependencies among the involved entities, which can represent relevant cooperative/interfering activities.

Specifically, LP-HCLUS identifies hierarchically organized, possibly overlapping multi-type clusters from a heterogeneous network and exploits them for predictive purposes, i.e., to predict the existence of previously unknown links. The extraction of a hierarchical structure, rather than a flat structure, allows the biologists to focus on either more general or more specific interaction activities. Finally, the possible overlaps among the identified

clusters allow LP-HCLUS to consider multiple roles of the same disease or ncRNA, which may be involved in multiple interaction networks.

It is noteworthy that, even if the analyzed network may consist of an arbitrary number of types of nodes and edges, the prediction of new associations will focus on edges involving ncRNAs and diseases, called *target* types. On the contrary, node types that are only used during the analysis will be called *task-relevant* node types.

Intuitively, the approach followed by LP-HCLUS consists of three main steps:

1. estimation of the strength of relationships for all the possible pairs of ncRNAs and diseases, according to the paths connecting such nodes in the network and to the features of nodes involved in such paths;
2. construction of a hierarchy of overlapping multi-type clusters, on the basis of the strength of relationships computed in the previous step;
3. identification of predictive functions to predict new ncRNA-disease relationships on the basis of the clusters identified at different levels of the hierarchy.

It is noteworthy that the clustering step could be directly applied on the set of known interactions, without performing the first step. However, such an approach would lead to discard several potential indirect relationships that can be caught only through a deep analysis of the network, which is indeed the main purpose of the first step. A naïve solution for the prediction task would be the use of the output of the first step as the final score, ignoring steps 2 and 3. However, this would lead to disregard a more abstract perspective of the interactions which, instead, can be caught by the clustering-based approach. Another effect would be to disregard the network homophily phenomenon and not to catch possible relationships between ncRNAs and between diseases based on the nodes they are connected with. On the contrary, the exploitation of such relationships is in line with the *guilt-by-association (GBA)* principle, which states that entities with similar functions tend to share interactions with other entities. This principle has been recently applied to and investigated for ncRNAs [25].

Each step will be described in details in the next subsections, while in the following we formally define the heterogeneous attributed network, that is analyzed by LP-HCLUS, as well as the solved task.

Definition 1 (Heterogeneous attributed network) *A heterogeneous attributed network is a network $G = (V, E)$, where V denotes the set of nodes and E denotes the set of edges, and both nodes and edges can be of different types (see Fig. 1). 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 and $\mathcal{T}_{tr}$ is the set of task-relevant types;*
- *each node type $T_v \in \mathcal{T}$ defines a subset of nodes in the network, that is $V_v \subseteq V$;*
- *each node type $T_v \in \mathcal{T}$ is associated with a set of attributes $\mathcal{A}_v = \{A_{v,1}, A_{v,2}, \ldots, A_{v,m_v}\}$, i.e., all the nodes of a given type T_v are described according to 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 $E_l \subseteq E$.*

Definition 2 (Overlapping Multi-type cluster) *Given a heterogeneous attributed network $G = (V, E)$, an overlapping multi-type cluster is defined as $G' = (V', E')$, where:*

- $V' \subseteq V$;
- $\forall v' \in V'$, v' *is a node of a target type;*
- $\forall v' \in V'$, v' *may also belong to other clusters besides G';*
- $E' \subseteq (E \cup \hat{E})$ *is a set of relationships among the nodes in V', belonging either to the set of known relationships E or to a set of extracted relationships $\hat{E}$, which are identified by the clustering method.*

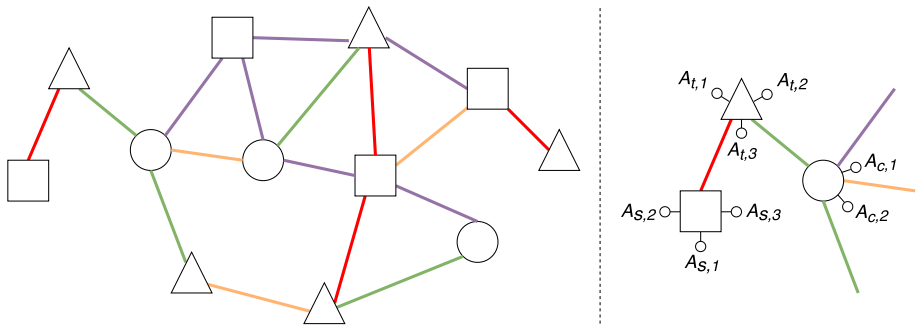

Fig. 1 An example of a heterogeneous attributed network. On the left, a general overview of the network, where shapes represent different node types and colors represent different edge types. On the right, a zoom on a small portion of the network, where we can observe node attributes associated with squares ($A_{s,*}$), triangles ($A_{t,*}$) and circles ($A_{c,*}$)

The details about the strategy adopted to identify $\hat{E}$ will be discussed in "Estimation of the strength of the relationship between ncRNAs and diseases" section.

Definition 3 (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* $L_i, i = 2, 3, \ldots . k$*, we have that* $\forall\ G' \in L_i\ \exists\ G'' \in L_{i-1}$*, such that* G'' *is a subnetwork of* G' *(see Fig. 2).*

On the basis of these definitions, we formally define the task considered in this work.

Definition 4 (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* $T_{i_2} \in \mathcal{T}_t$*. Intuitively, each function* $\psi^{(w)}$ *maps each possible pair of nodes (of types* T_{i_1} *and* T_{i_2}*, respectively) to a score that represents the degree of certainty of their relationship.*

The learning setting considered in this paper is *transductive*. In particular, only the links involving nodes already known and exploited during the training phase are considered for link prediction. In other terms, we do not learn a model from a network and apply this model to a completely different network (classical inductive learning setting).

The method proposed in this paper (see Fig. 3 for the general workflow) aims at solving the task formalized in Definition 4, by considering ncRNAs and diseases as target types (Fig. 4). 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.

Estimation of the strength of the relationship between ncRNAs and diseases

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 [26], a *meta-path* is a set of sequences of nodes 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 (see Fig. 5). Given the ncRNA n_i and the disease d_j, for each meta-path P, we compute a score $pathscore(P, n_i, d_j)$, which represents the strength of their relationship on the basis of the meta-path P.

In order to combine multiple contributions provided by different meta-paths, we adopt a strategy that follows the classical formulation of fuzzy sets [27]. In particular, a relationship between a ncRNA n_i and a disease d_j can be considered "certain" if there is at least one meta-path which confirms its certainty. 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:

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

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.

As introduced before, each meta-path P represents a finite set of sequences of nodes, where:

- the i-th node of each sequence in the metapath P is of the same type;

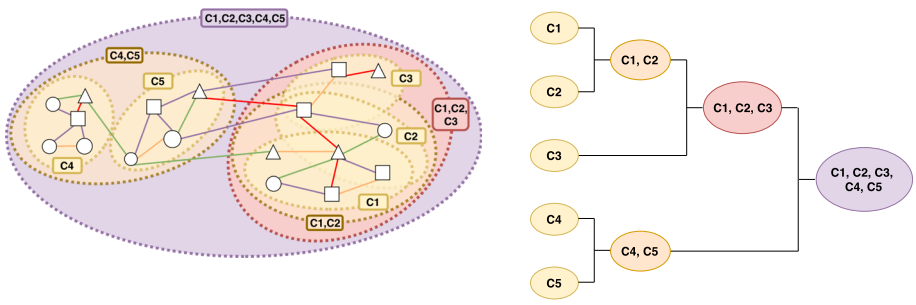

Fig. 2 A hierarchy of overlapping multi-type clusters: **a** emphasizes the overlapping among multi-type clusters; **b** shows their hierarchical organization

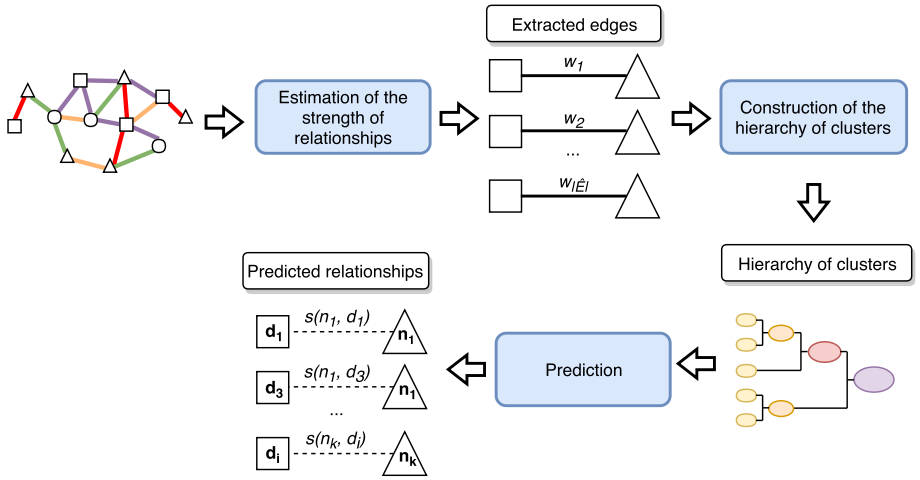

Fig. 3 Workflow of the method LP-HCLUS

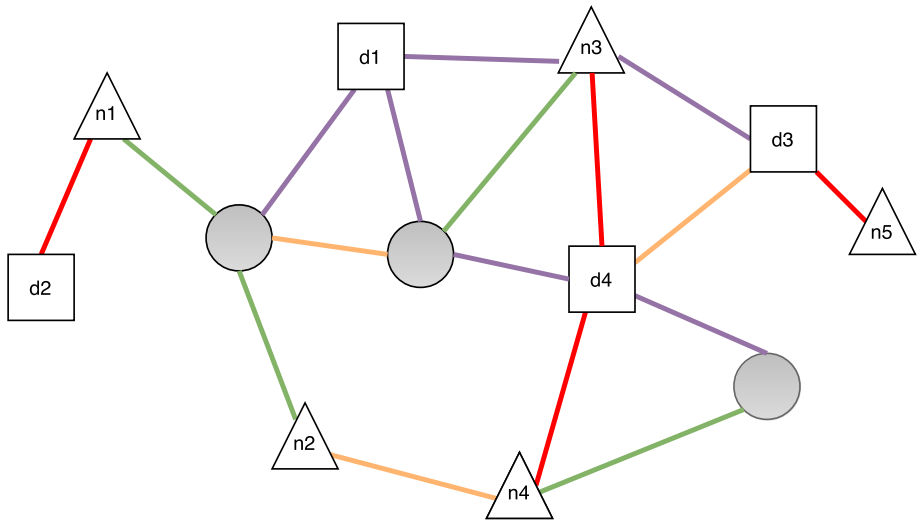

Fig. 4 An example of a ncRNA-disease heterogeneous network. In this example, ncRNAs are represented as triangles, while diseases are represented as squares. Other (task-relevant) nodes (e.g., target genes, proteins, etc) are represented as gray circles

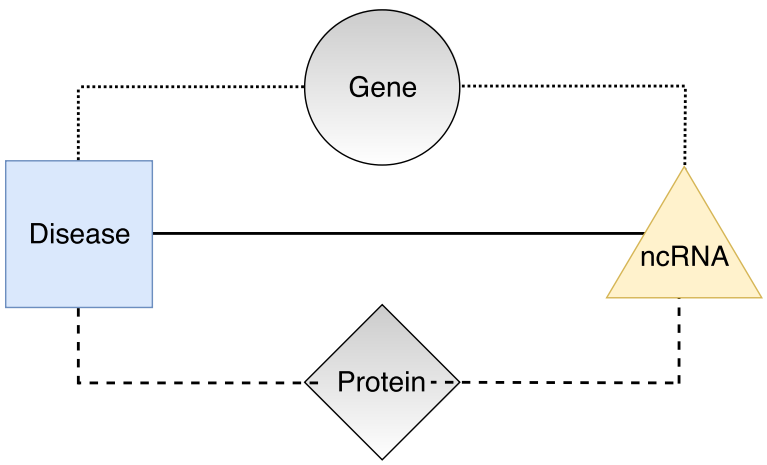

Fig. 5 Diagram showing three different meta-paths between a disease and a ncRNA. The first meta-path connects diseases and ncRNAs via genes, the second connects diseases and ncRNAs directly and the third connects diseases and ncRNAs via proteins

- the first node is a ncRNA and the last node is a disease;
- if two nodes are consecutive in the sequence, then there is an edge between them in E.

According to this definition, if there is a path P directly connecting a ncRNA n_i to a disease d_j, then $pathscore(P, n_i, d_j) = 1$, therefore $s(n_i, d_j) = 1$.

Otherwise, when there is no direct connection between n_i and d_j, $pathscore(P, n_i, d_j)$ is computed as the maximum similarity between the sequences that start with n_i and those that end with d_j. Formally:

$$pathscore(P, n_i, d_j) = \max_{\substack{seq', seq'' \in P, \\ seq'.first = n_i, seq''.last = d_j}} similarity(seq', seq'') \quad (2)$$

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. In other words, this is a way to analyze the similarity between the neighborhood of n_i and the neighborhood of d_j in terms of the (similarity of the) paths they are involved in.

It is noteworthy that, in order to make the neighbors comparable, we exploit the concept of meta-path, which includes sequences that involve the same types of nodes. In fact, in Formula (2), the similarity between two sequences seq' and seq'' is computed as follows:

$$similarity(seq', seq'') = \frac{\sum_{x \in A^{(P)}} s_x(seq', seq'')}{|A^{(P)}|} \quad (3)$$

where:

- $A^{(P)}$ is the set of attributes of the nodes involved in the path P;
- $s_x(seq', seq'')$ is the similarity between $val_x(seq')$, that is the value of the attribute x in the sequence seq', and $val_x(seq'')$, that is the value of the attribute x in the sequence seq''.

Following [28], we compute $s_x(seq', seq'')$ as follows:

- 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, then $s_x(seq', seq'') = 1$ if $val_x(seq') = val_x(seq'')$, 0 otherwise.

An example of the computation of the similarity among sequences is reported in Fig. 6. In this example, we compute the score between the ncRNA *h19* and the disease *asthma*. First, we identify the sequences starting with *h19* (i.e., 1 and 9, emphasized in yellow) and those ending with *asthma* (i.e., 4, 5, 6 and 7, emphasized in blue). Then we pairwisely compute the similarity between sequences belonging to the two sets and select the maximum value, according to Eq. 2. The similarity between two sequences is computed according to Eq. 3.

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. Such an aggregation is performed up to a predefined *depth* of analysis in the network. In this way, we fully exploit the network autocorrelation phenomena.

Construction of a hierarchy of overlapping multi-type clusters

Starting from the set of possible ncRNA-disease pairs, each associated with a score that represents its degree of certainty, we construct the first level of the hierarchy by identifying a set of overlapping multi-type clusters in the form of bicliques. That is, multi-type clusters where all the ncRNA-disease relationships have a score greater than (or

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. 6 Analysis of 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". White rows, although belonging to P, are not considered during the computation of the similarity in this specific example, since they do not involve "h19" or "asthma"

equal to) a given threshold $\beta \in [0,1]$ (see Fig. 7). More formally, in order to construct the first level of the hierarchy L_1, we perform the following 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 $\{(n_i, d_j) | s(n_i, d_j) \geq \beta\}$.
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 $<_c$ exploited by the merging step implicitly defines a greedy search strategy that guides the order in which pairs of clusters are analyzed and possibly merged. $<_c$ is based on the cluster cohesiveness $h(c)$, which corresponds to the average score of the interactions in the cluster. Formally:

$$h(C) = \frac{1}{|pairs(C)|} \cdot \sum_{(n_i,d_j) \in pairs(C)} s(n_i, d_j) \tag{4}$$

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. Numerically, $|pairs(C)| = |\{n_i | n_i \in C \wedge n_i \in T_n\}| \cdot |\{d_j | d_j \in C \wedge d_j \in T_d\}|$.

Accordingly, if C' and C'' are two different clusters, the ordering relation $<_c$ is defined as follows:

$$C' <_c C'' \iff h(C') > h(C'') \tag{5}$$

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, in this case, we do not obtain bicliques, but generic multi-type clusters, i.e., the score associated with each interaction does not need to satisfy the threshold β. 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 [29], 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 defined in Eq. 4. This means that α defines the minimum cluster cohesiveness that must be satisfied by a cluster obtained after a merging: small values of α lead to increase the number of merging operations and, therefore, to a relatively small number of final clusters containing a large number of nodes.

For every iteration of the merging procedure, a new hierarchical level is generated. The iterative process stops when it is not possible to merge more clusters with a minimum level of cohesiveness α. The output of such a process is a hierarchy of overlapping multi-type clusters $\{L_1, L_2, \ldots, L_k\}$ (see Definition 3).

A pseudocode description of the proposed algorithm for the construction of the hierarchy of clusters is reported in Algorithm 1.

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

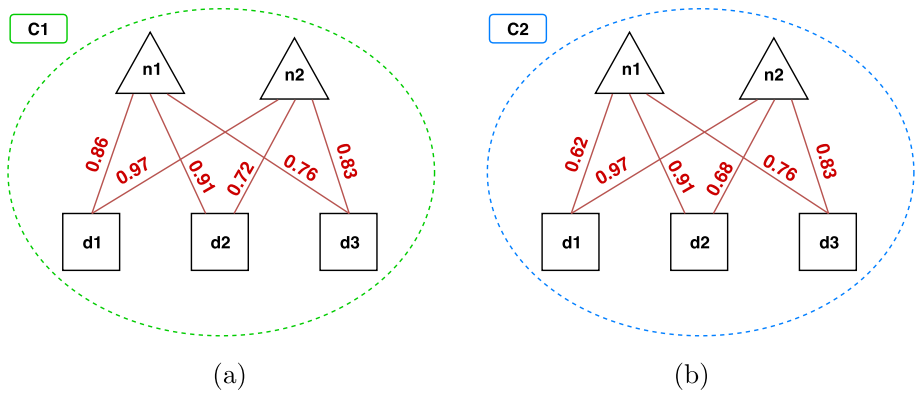

Fig. 7 Biclique constraint on two multi-type clusters **a** An example of multi-type cluster which satisfies the biclique constraint with $\beta = 0.7$ (i.e., all the relationships have a score ≥ 0.7). **b** An example that does not satisfy such a constraint. It is noteworthy that, with $\beta = 0.6$, also (**b**) would satisfy the biclique constraint

Algorithm 1 Construction of the hierarchy of overlapping multi-type clusters

Require:
- Initial set of clusters L_0, each containing a single ncRNA-disease pair in $\{(n_i, d_j)|s(n_i, d_j) \geq \beta\}$;
- An ordering relation $<_c$ that reflects the quality of the clusters;
- A threshold α on the quality of the clusters obtained after a merging.

Ensure:
- The hierarchy of overlapping multi-type clusters $L_1, L_2, \ldots, L_k$

$k \leftarrow 0$
repeat
  {Define the merging condition: biclique constraint for the first level; threshold on the cluster cohesiveness $h(\cdot)$ for the subsequent levels}
  if k = 0 **then**
    $condition(\cdot) \leftarrow isBiclique(\cdot)$
  else
    $condition(\cdot) \leftarrow h(\cdot) > \alpha$
  end if
  $L \leftarrow L_k$
  sort L in according to the ordering relation $<_c$
  $clusters \leftarrow [\,]$
  $mergedClusters \leftarrow 0$
  {Loop over the sorted list of clusters. This defines a greedy search strategy: clusters with a higher cohesiveness value are processed first}
  for $i \leftarrow 1$ **to**$|L| - 1$ **do**
    $C' \leftarrow L[\,i]$
    {Search for another cluster that can be merged with C' in the ordered list}
    $j \leftarrow i + 1$
    $merged \leftarrow false$
    while $j \leq |L|$ **and not** *merged* **do**
      $C'' \leftarrow L[j]$
      $C''' \leftarrow merge(C', C'')$
      {If C' and C'' can be merged into C''' according to the merging condition, merge them}
      if $condition(C''')$ **then**
        add C''' to *clusters*
        $mergedClusters \leftarrow mergedClusters + 1$
        remove C'' from L
        $merged \leftarrow true$
      end if
      $j \leftarrow j + 1$
    end while
    {If C' cannot be merged with any other cluster, add it to the result as it is}
    if $merged = false$ **then**
      add C' to *clusters*
    end if
  end for
  $newLevel \leftarrow false$
  {Check if there was at least one merging}
  if $mergedClusters > 0$ **then**
    if $k > 0$ **then**
      {If we are not building the first level, define a new hierarchical level}
      $k \leftarrow k + 1$
      $newLevel \leftarrow true$;
    end if
    $L_k \leftarrow clusters$
  else
    {End the construction of the first hierarchical level and continue with the others}
    if $k = 0$ **then**
      $k \leftarrow k + 1$
      $L_k \leftarrow clusters$
      $newLevel \leftarrow true$;
    end if
  end if
until $mergedClusters = 0$ **and** $newLevel = false$
return $L_1, L_2, \ldots, L_k$

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:

$$\psi^{(w)}(n_i, d_j) = h\left(C_{ij}^w\right), \tag{6}$$

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 (i.e., $C_{ij}^w = [C_1, C_2, \ldots, C_m]$), ordered accordingly to relation $<_c$ defined in Eq. 5, in which both n_i and d_j appear, on which we apply an aggregation function to obtain a single degree of certainty. In this work, we propose the adoption of four different aggregation functions:

- **Maximum**: $\psi^{(w)}(n_i, d_j) = \max_{c \in C_{ij}^w} h(c)$
- **Minimum**: $\psi^{(w)}(n_i, d_j) = \min_{c \in C_{ij}^w} h(c)$
- **Average**: $\psi^{(w)}(n_i, d_j) = \frac{1}{|C_{ij}^w|} \cdot \sum_{c \in C_{ij}^w} h(c)$
- **Evidence Combination**: $\psi^{(w)}(n_i, d_j) = ec(C_m)$, where:

$$ec(C_m) = \begin{cases} h(C_1) & \text{if } C_m = C_1 \\ ec(C_{m-1}) + [1 - ec(C_{m-1})] \cdot h(C_m) & \text{otherwise} \end{cases} \tag{7}$$

It is noteworthy that the Evidence Combination function, already exploited in the literature in the context of expert systems [30], generally rewards the relationships appearing in multiple high cohesive clusters.

In the following, we report an example of this prediction step, with the help of Fig. 8. In this example, we have two overlapping multi-type clusters C_1 and C_2, identified at the w-th hierarchical level, that suggest two new potential relationships (dashed lines in the figure), i.e. the pair n_2, d_2 and the pair n_2, d_3.

The first relationship only appears in C_1, therefore its degree of certainty is computed according to the cohesiveness of C_1 (see Eq. 4):

$$\psi^{(w)}(n_2, d_2) = h(C_1) = \frac{1}{2 \cdot 3}(0.7 + 0.8 + 0.9) = 0.4. \tag{8}$$

On the contrary, the second relationship is suggested by both C_1 and C_2, i.e., it appears in their overlapped area. Therefore, we aggregate the cohesiveness of C_1 and C_2 according to one of the functions we described before. In particular, since $h(C_1) = 0.4$ and $h(C_2) = \frac{1}{1 \cdot 2} \cdot 0.6 = 0.3$, we have:

- **Maximum**: $\psi^{(w)}(n_2, d_3) = \max_{c \in C_{ij}^w} h(c) = 0.4$
- **Minimum**: $\psi^{(w)}(n_2, d_3) = \min_{c \in C_{ij}^w} h(c) = 0.3$
- **Average**: $\psi^{(w)}(n_2, d_3) = \frac{1}{|C_{ij}^w|} \cdot \sum_{c \in C_{ij}^w} h(c) = \frac{1}{2} \cdot (0.4 + 0.3) = 0.35$
- **Evidence Combination**: $\psi^{(w)}(n_2, d_3) = h(C_1) + [1 - h(C_1)] \cdot h(C_2) = 0.4 + (1 - 0.4) \cdot 0.3 = 0.58$

Results

The proposed method was evaluated through several experiments. In this section, we present the main adopted resources, define the experimental setting, introduce the adopted evaluation measures and compare our system with the competitors from a quantitative viewpoint.

Datasets

We performed experiments on two different heterogeneous networks involving ncRNAs and diseases. In the following, we report the details of each dataset, together

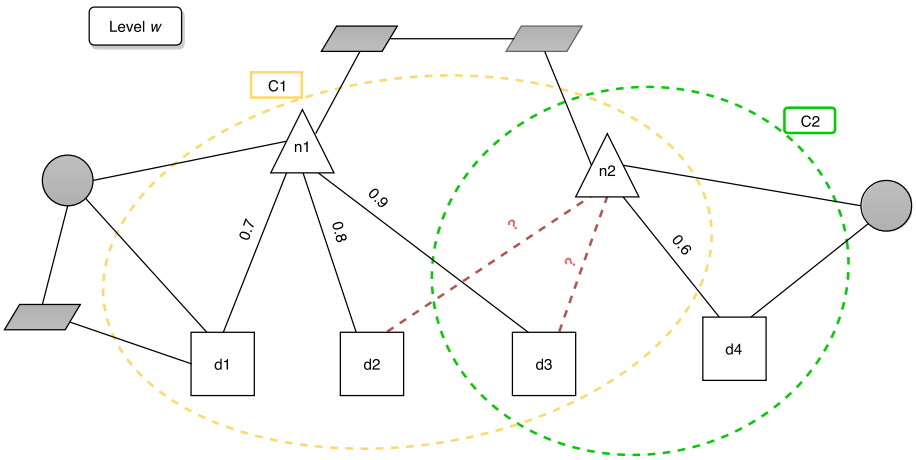

Fig. 8 Example of the prediction step. Two clusters identified at a given hierarchical level w. Triangles represent ncRNAs, squares represent diseases and the grey shapes are other type nodes. The clusters suggest two new possible relationships between n_2 and d_2 and between n_2 and d_3

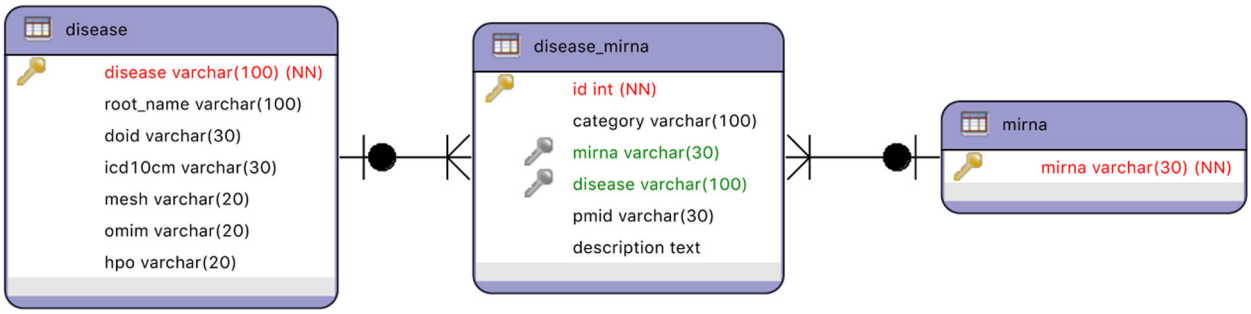

Fig. 9 UML diagram of the dataset HMDD v3.0. The attributes in red are the identifiers of the nodes of a given type (i.e., the primary key in a relational database), while attributes in green refer to the identifier of nodes of other types (i.e., foreign keys in a relational database)

with UML diagrams that represent their data and structure, i.e., nodes, links and attributes.

HMDD v3 [31]. This dataset stores information about diseases, miRNAs and their known relationships. The network consists of 985 miRNAs, 675 diseases (characterized by 6 attributes) and 20,859 relationships between diseases and miRNAs (characterized by 3 attributes). A diagram of this dataset is depicted in Fig. 9, while the attributes are described in Table 1. The official link of the dataset is: http://www.cuilab.cn/hmdd. In this evaluation, we used two versions of the HMDD v3 dataset: the version released on June 28th, 2018 (v3.0) and the version released on March 27th, 2019 (v3.2). Both versions are available at the following link: http://www.di.uniba.it/~gianvitopio/systems/lphclus/.

Table 1 HMDD v3.0 dataset - Description of the attributes

Type	Feature	Description
Disease	disease	Disease name
	root_name	Category of the disease
	doid	Disease Ontology Identifiers
	icd10cm	ICD-10-CM Code
	mesh	Medical Subject Headings (MeSH) code
	omim	Online Mendelian Inheritance in Man (OMIM) code
	hpo	Human Phenotype Ontology (HPO) code
Disease_miRNA	id	ID of the relationship
	category	Category of the relationship
	mirna	miRNA involved in the association
	disease	Disease involved in the association
	pmid	PubMed ID of the publication reporting the association
	description	Description of the relationship
miRNA	mirna	miRNA name

Integrated Dataset (ID). This dataset has been built by integrating multiple public datasets in a complex heterogeneous network. The source datasets are:

- lncRNA-disease relationships and lncRNA-gene interactions from [32] (June 2015)[1]
- miRNA-lncRNA interactions from [33] [2]
- disease-gene relationships from DisGeNET v5 [34] [3]
- miRNA-gene and miRNA-disease relationships from miR2Disease [35] [4]

From these resources we only kept data related to *H. Sapiens*. The integration led to a network consisting of 1015 ncRNAs (either lncRNAs or miRNAs), 7049 diseases, 70 relationships between lncRNAs and miRNAs, 3830 relationships between diseases and ncRNAs, 90,242 target genes, 26,522 disease-target associations and 1055 ncRNA-target relationships. Most of the considered entities are also characterized by a variable number of attributes, as shown in Fig. 10 and in Table 2. The final dataset is available at the following link: http://www.di.uniba.it/~gianvitopio/systems/lphclus/.

Experimental setting & competitors

LP-HCLUS has been run with different values of its input parameters, namely: $\alpha \in \{0.1, 0.2\}$ (we remind that α is the minimum cohesiveness that a cluster must satisfy) and $\beta \in \{0.3, 0.4\}$ (we remind that β represents the minimum score that each ncRNA-disease pair must satisfy to be considered as existing), while *depth* has been set to 2 in order to consider only nodes that are relatively close to those involved in the meta-paths. We performed a comparative analysis with two competitor systems and a baseline approach that we describe in the following.

[1] http://www.cuilab.cn/lncrnadisease
[2] Dataset "Data S3" in https://www.sciencedirect.com/science/article/pii/S009286741300439X?via%3Dihub#mmc3
[3] http://www.disgenet.org/
[4] http://www.mir2disease.org/

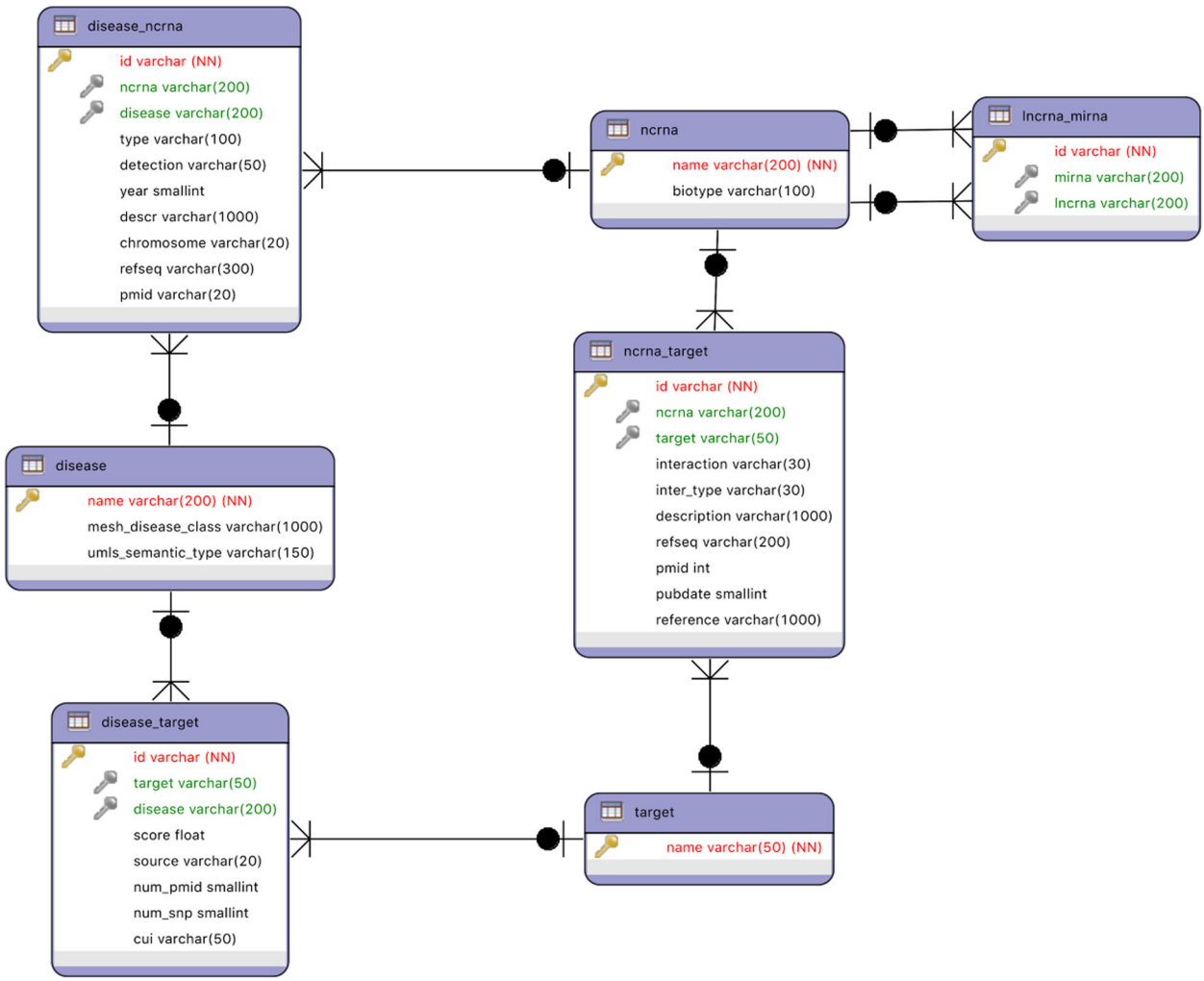

Fig. 10 UML diagram of the Integrated Dataset (ID). The attributes in red are the identifiers of the nodes of a given type (i.e., the primary key in a relational database), while attributes in green refer to the identifier of nodes of other types (i.e., foreign keys in a relational database)

HOCCLUS2 [29] is a biclustering algorithm that, similarly to LP-HCLUS, is able to identify a hierarchy of (possibly overlapping) heterogeneous clusters. HOCCLUS2 was initially developed to study miRNA-mRNA associations, therefore it is inherently limited to two target types. Moreover, besides miRNAs, mRNAs and their associations, it is not able to take into account other entities in the network and actually cannot predict new relationships. We adapted HOCCLUS2 in order to analyze ncRNA-disease relationships and to be able to predict new associations. In particular, we fed HOCCLUS2 with the dataset produced by the first step of LP-HCLUS (see "Estimation of the strength of the relationship between ncRNAs and diseases" section) and we performed the prediction according to the strategy we proposed for LP-HCLUS (see "Prediction of new ncRNA-disease relationships" section), considering all the aggregation functions proposed in this paper. We emphasize that, since both the initial analysis and the prediction step are performed by LP-HCLUS modules, the comparison with HOCCLUS2 allows us to evaluate the effectiveness of the proposed clustering approach. Since the HOCCLUS2 parameters have a similar meaning with respect to LP-HCLUS parameters, we evaluated its results with the same parameter setting, i.e., $\alpha \in \{0.1, 0.2\}$ and $\beta \in \{0.3, 0.4\}$.

ncPred [14] is a system which was specifically designed to predict new associations between ncRNAs and diseases. ncPred analyzes two matrices containing information about ncRNA-gene and gene-disease relationships. Therefore, we transformed the considered heterogeneous networks into matrices and fed ncPred with them. We again emphasize that ncPred is not able to catch information coming from other entities in the network of types different from ncRNAs and diseases, and that it is not able

Table 2 ID dataset - Description of the attributes

Type	Feature	Description
Disease	name	Disease name
	mesh_disease_class	Disease classification by Medical Subject Headings (MeSH)
	umls_semantic_type	Semantic type provided by the Unified Medical Language System
Disease_ncRNA	id	ID of the relationship
	ncrna	ncRNA involved in the association
	disease	Disease involved in the association
	type	Type of association
	detection	Method used to detect the relationship
	year	Year of the detection
	descr	Description of the association
	chromosome	Chromosome
	refseq	RefSeq identifier
	pmid	PubMed ID of the publication reporting the association
Disease_target	id	ID of the relationship
	target	Target gene involved in the association
	disease	Disease involved in the association
	score	DisGENET score for the Gene-Disease association
	source	Original source reporting the Gene-Disease association
	num_pmid	Total number of publications reporting the association
	num_snp	Total number of SNPs associated to the association
	cui	Concept Unique Identifier (CUI)
lncRNA_miRNA	id	ID of the relationship
	mirna	miRNA involved in the association
	lncrna	lncRNA involved in the association
ncRNA	name	ncRNA name
	biotype	Type of ncRNA. The value can be "lncrna" or "mirna"
ncRNA_target	id	ID of the relationship
	ncrna	ncRNA involved in the association
	target	Target genes involved in the association
	interaction	Elements involved in the associations (e.g. RNA-RNA, RNA-protein)
	inter_type	Type of interaction (e.g. Regulatory, Binding, etc.)
	description	Description of the interaction
	refseq	RefSeq identifier
	pmid	PubMed ID of the publication reporting the association
	pubdate	Date of first publication
	reference	Textual description of the association
Target	name	Name of target gene

to exploit features associated to nodes and links in the network. We set ncPred parameter values to their default values.

LP-HCLUS-NoLP, which corresponds to our 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 "Estimation of the strength of the relationship between ncRNAs and diseases" section) as the final score associated with each interaction. This approach allows us to evaluate the contribution provided by our link prediction approach based on multi-type clustering.

The evaluation was performed through a 10-fold cross-validation. It is noteworthy that the computation of classical measures, such as Precision and Recall, would require the presence of negative examples or some assumptions made on unknown examples. In our case, the datasets contain only positive examples, i.e., we have a set of validated relationships but we do not have negative examples of relationships (relationships whose non-existence has been proven).

Therefore, following the approach adopted in [13], we evaluated the results in terms of TruePositiveRate@k, where:

- an association is considered a True Positive (TP) if it is validated in the literature and it is in the first top k relationships predicted by the system;
- an association is considered a False Negative (or FN) if it is validated in the literature, but it is not in the first top k relationships predicted by the system.

Since the optimal value of k cannot be known in advance, we plot the obtained TPR@k by varying the value of k and compute the Area Under the TPR@k curve (AUTPR@k). For a thorough analysis on the most promising (i.e., top-ranked) interactions, we report all the results by varying the value of k within the interval $[1, 5000]$, obtained with the same configuration of the parameters α and β for HOCCLUS2 and LP-HCLUS. Moreover, we also report the results in terms of ROC and Precision-Recall curves, as well as the areas under the respective curves (AUROC and AUPR), by considering the unknown relationships as negative examples. We remark that AUROC and AUPR results 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 the paper we report the results obtained with the most promising configuration according to some preliminary experiments. The complete results, including those obtained in such preliminary experiments, can be downloaded at: http://www.di.uniba.it/~gianvitopio/systems/lphclus/.

Results - HMDD v3 dataset

In Figures 11, 12 and 13 we show the results obtained on the HMDD dataset in terms of TPR@k, ROC and Precision-Recall curves, while in Table 3, we report the AUTPR@k, AUROC and AUPR values. From Fig. 11, we can observe that the proposed method LP-HCLUS, with the combination strategy based on the maximum, is in general able to obtain the best performances. The competitor system ncPred obtains good results, but it outperforms LP-HCLUS_MAX only for high values of k, and only when focusing on the first level of the hierarchy. However, we stress the fact that it is highly preferable to

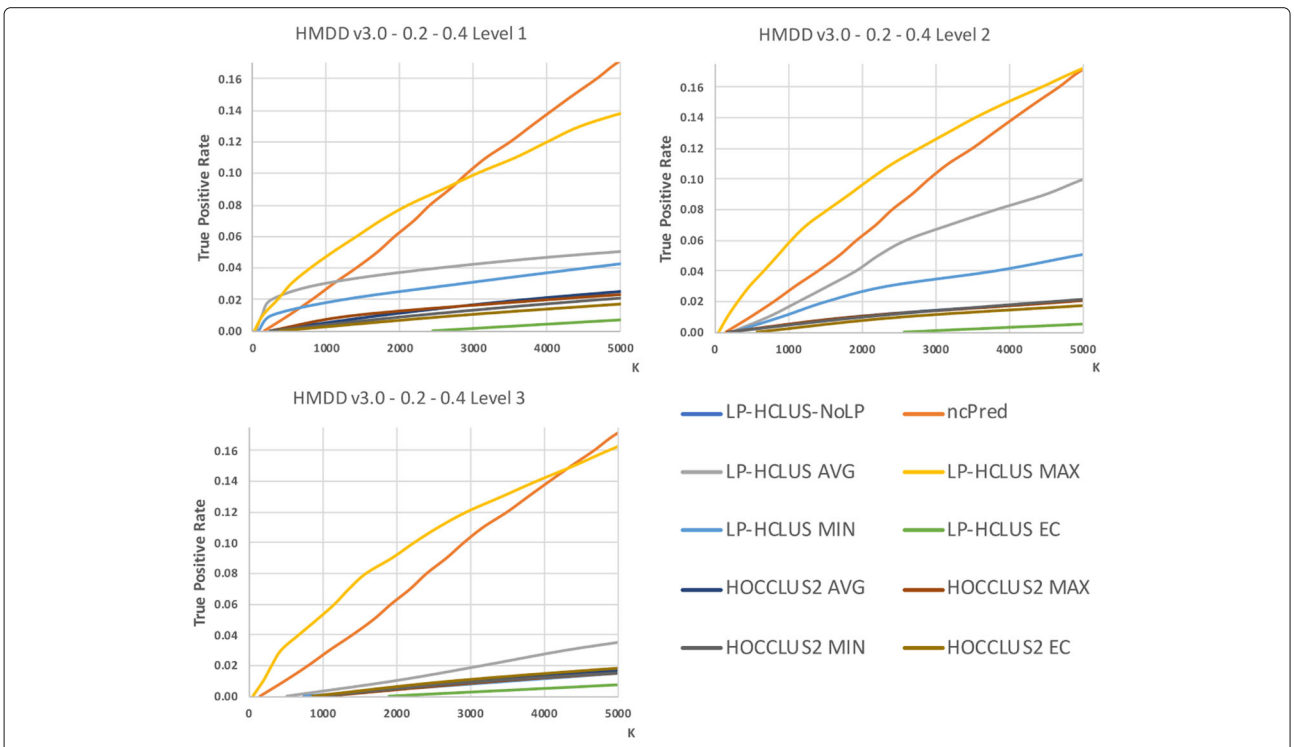

Fig. 11 TPR@k results for the dataset HMDD v3.0, obtained with the best configuration ($\alpha = 0.2, \beta = 0.4$) at different levels of the hierarchy

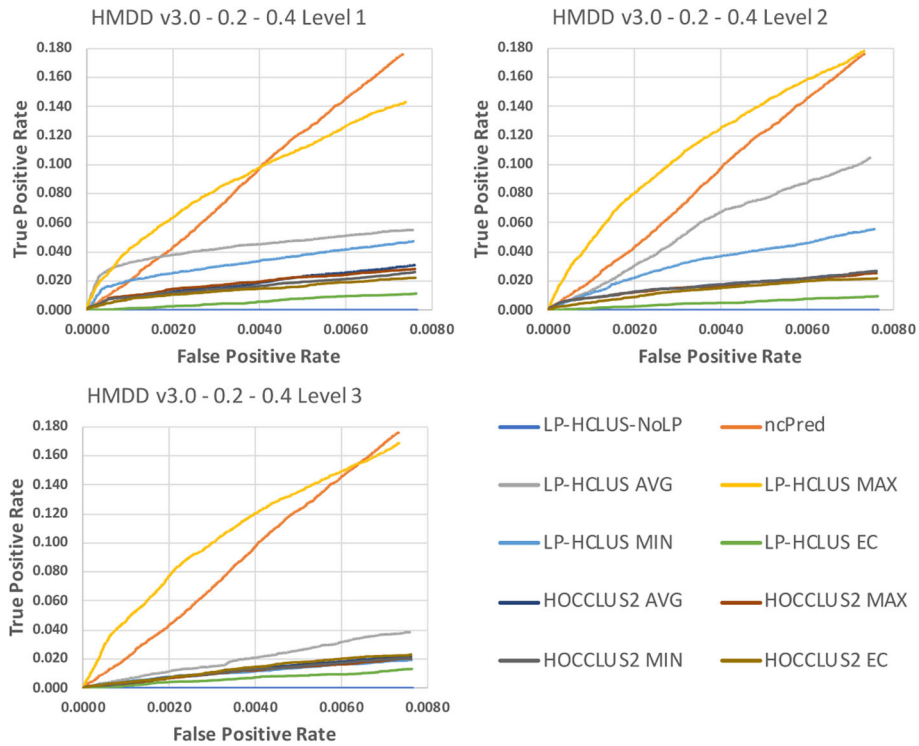

Fig. 12 ROC curves for the dataset HMDD v3.0, obtained with the best configuration ($\alpha = 0.2, \beta = 0.4$) at different levels of the hierarchy. These curves can only be used for relative comparison and not as absolute evaluation measures because they are spoiled by the assumption made on unknown relationships

achieve better performances on the left side of the curve, i.e., with low values of *k*, since it is the real portion of the ranking on which researchers will focus their analysis. In such a portion of the curve, LP-HCLUS_MAX dominates over all the competitors for all the hierarchical levels. It is noteworthy that some variants of LP-HCLUS (i.e., MAX and AVG) obtain their best performances at the second level of the hierarchy. This emphasizes that the extraction of a hierarchy of clusters could provide some improvements with respect to a flat clustering. This is not so evident for HOCCLUS2 even if, analogously to LP-HCLUS, it is able to extract a hierarchy. The results in terms of AUTPR@*k*, AUROC and AUPR (see Table 3) confirm the superiority of LP-HCLUS_MAX over the competitors.

Results - ID dataset

In Figures 14, 15 and 16 we show the results obtained on the Integrated Dataset (ID) in terms of TPR@k, ROC and Precision-Recall curves, while in Table 4, we report the AUTPR@k, AUROC and AUPR values. It is noteworthy that this dataset is much more complex than HMDD, because it consists of several types of nodes, each associated with its attributes. In this case, the system LP-HCLUS can fully exploit information brought by other node types to predict new associations between ncRNAs and diseases.

As it can be observed from the figures, thanks to such an ability, LP-HCLUS clearly outperforms all the competitors. It is noteworthy that also the simpler version of LP-HCLUS, i.e., LP-HCLUS-NoLP, is able to outperform the competitors, since it exploits the exploration of the network based on meta-paths. However, when we exploit the full version of LP-HCLUS, which bases its prediction on the clustering results, the improvement over the existing approaches becomes much more evident. These conclusions are also confirmed by the AUTPR@k, AUROC and AUPR values shown in Table 4.

Statistical comparisons

By observing the results reported in Figs. 11, 12, 13, 14, 15 and 16, it is clear that the adoption of the Maximum (MAX) as LP-HCLUS aggregation function leads to the best results. This behavior can be motivated by the fact that such an approach rewards the associations which show at least one strong evidence from the clusters. Although such a behavior should be observed also with the Evidence Combination (EC) function, it is noteworthy that the latter also rewards associations which are confirmed by several clusters, even if they show a weak confidence. In this way, EC is prone to false positives introduced by the combined contribution of several weak relationships.

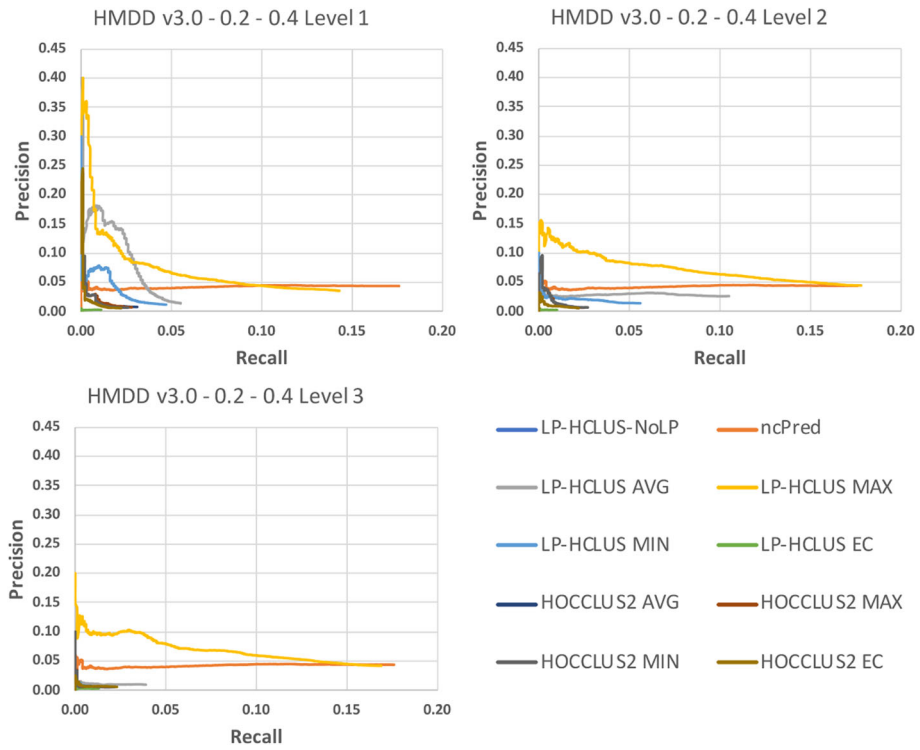

Fig. 13 Precision-Recall curves for the dataset HMDD v3.0, obtained with the best configuration ($\alpha = 0.2, \beta = 0.4$) at different levels of the hierarchy. These curves can only be used for relative comparison and not as absolute evaluation measures because they are spoiled by the assumption made on unknown relationships

In order to confirm the superiority of LP-HCLUS_MAX from a statistical viewpoint, we performed a Friedman test with Nemenyi post-hoc test with significance value of 0.05. This test is applied to the Area Under the TPR@k curve, in order to provide a k-independent evaluation of the results. By observing the results in Fig. 17, it is clear that LP-HCLUS_MAX is the best ranked method among the considered approaches. Since, at a glance, the difference between LP-HCLUS_MAX and ncPred is clear, but does not appear to be statistically significant with a test that evaluates differences across multiple systems, we performed three pairwise Wilcoxon tests (one for each hierarchical level), with the Bonferroni correction. In this way, it is possible to directly compare LP-HCLUS_MAX and ncPred. Looking at the average Area Under the TPR@k and p-values reported in Table 5, it is clear that the difference between LP-HCLUS_MAX and its direct competitor ncPred is large (especially for the ID dataset) and, more importantly, statistically significant for all the hierarchical levels, at a significance value of 0.01.

Discussion

In this section we discuss about the results of the comparison of LP-HCLUS with its competitors from a qualitative viewpoint, in order to assess the validity of the proposed system as a useful tool for biologists.

Discussion on the HMDD v3 dataset

We 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). A graphical overview of the results of this analysis is provided in Fig. 18, while the detailed results are provided in Additional file 3, where the relationships introduced in the new release of HMDD are highlighted in green. The general conclusion we can draw from Fig. 18 is that several relationships predicted by LP-HCLUS have been introduced in the new HMDD release v3.2.

In particular, we found 3055 LP-HCLUS predictions confirmed by the new release of HMDD at the hierarchy level 1 (score range 0.97-0.44), 4119 at level 2 (score range 0.93-0.37) and 4797 at level 3 (score range 0.79-0.37). Overall, these results underline the behavior of LP-HCLUS at the different levels of the hierarchy. As expected, the number of predictions grows progressively from the lowest to the highest levels of the hierarchy, due to the less stringent constraints imposed by the algorithm,

Table 3 AUTPR@k, AUROC and AUPR values for the dataset HMDD, obtained with the best configuration ($\alpha = 0.2, \beta = 0.4$) at different levels of the hierarchy

		AUTPR@k	AUPR	AUROC
LP-HCLUS-NoLP		0.000000	0.000000	0.496169
ncPred		0.087370	0.007540	0.584268
LP-HCLUS AVG	*Level 1*	0.042658	0.005437	0.523872
	Level 2	0.056392	0.003140	0.548665
	Level 3	0.020129	0.000469	0.515470
LP-HCLUS MAX	*Level 1*	0.088130	0.010865	0.568056
	Level 2	**0.109292**	**0.013420**	**0.585560**
	Level 3	0.104244	0.011983	0.580824
LP-HCLUS MIN	*Level 1*	0.031888	0.001935	0.519936
	Level 2	0.032765	0.001232	0.524077
	Level 3	0.011012	0.000170	0.505846
LP-HCLUS EC	*Level 1*	0.005626	0.000035	0.501872
	Level 2	0.004851	0.000030	0.500943
	Level 3	0.006493	0.000050	0.502762
HOCCLUS2 AVG	*Level 1*	0.018339	0.000839	0.511722
	Level 2	0.016484	0.000670	0.509663
	Level 3	0.012082	0.000287	0.507020
HOCCLUS2 MAX	*Level 1*	0.018332	0.000829	0.510398
	Level 2	0.016065	0.000659	0.508897
	Level 3	0.011150	0.000274	0.506331
HOCCLUS2 MIN	*Level 1*	0.015922	0.000753	0.509336
	Level 2	0.016401	0.000668	0.509542
	Level 3	0.011647	0.000270	0.506575
HOCCLUS2 EC	*Level 1*	0.013922	0.000536	0.507314
	Level 2	0.013717	0.000352	0.507112
	Level 3	0.013065	0.000253	0.507751

The results in terms of AUPR and AUROC values can only be used for relative comparison and not as absolute evaluation measures because they are spoiled by the assumption made on unknown associations, that are considered as negative examples

The best result is highlighted in boldface.

that allow LP-HCLUS to identify larger clusters at higher levels of the hierarchy. Larger clusters, even if possibly less reliable, in some cases can lead to the identification of less obvious functional associations.

Comparing the diseases at different levels of the hierarchy confirmed in the updated release of HMDD, we found associations involving 276 diseases at level 1, 360 at level 2 and 395 at level 3. Among the diseases involved in new associations predicted at level 3, but not at levels 1 and 2, there is the *acquired immunodeficiency syndrome*, a chronic, potentially life-threatening condition caused by the human immunodeficiency virus (HIV). The associations predicted by LP-HCLUS for this disease, confirmed in HMDD v3.2, involve hsa-mir-150 (with score 0.68) and hsa-mir-223 (with score 0.63). Such associations have been reported in [36]. The authors show the results of a study where the regulation of cyclin T1 and HIV-1 replication has been evaluated in resting and activated CD4+ T lymphocytes with respect to the expression of endogenous miRNAs. In this study, the authors demonstrated that miR-27b, miR-29b, miR-150, and miR-223 are significantly downregulated upon CD4(+) T cell activation, and identified miR-27b as a novel regulator of cyclin T1 protein levels and HIV-1 replication, while miR-29b, miR-223, and miR-150 may regulate cyclin T1 indirectly.

Other validated miRNAs associated with the *acquired immunodeficiency syndrome* in HMDD v3.2 are hsa-mir-27b, -29b, -29a, -29b-1 and hsa-mir-198. As shown in Fig. 19, these miRNAs, although not directly associated by LP-HCLUS with the *acquired immunodeficiency syndrome*, have been associated with disease terms strictly related to the immune system, with a score and specificity depending on the hierarchy level. In particular, at level 1, they have been associated with the *immune system disease* term (DOID_2914, a subclass of *disease of anatomical entity*) with a score ranging from 0.48 for hsa-mir-29b to a maximum value of 0.67 for hsa-mir-29a. At level 2 of the hierarchy, in addition to the classification in the *immune system disease*, they have also been associated with the *human immunodeficiency virus infection* (DOID_526) that is a subclass of *viral infectious disease* (DOID_934) and the direct parent of the *acquired immunodeficiency syndrome* (DOID_635). At level 3, all the miRNAs have also been associated with the *viral infectious disease* term.

In addition to hsa-mir-155 and hsa-mir-223, LP-HCLUS returned many other associations involving *acquired immunodeficiency syndrome* with a high score. In particular, 59 different miRNAs have been associated at level 2 (score between 0.74 and 0.63), and 191 at level 3 (score between 0.68 and 0.63). Considering such high scores, we investigated in the literature for some of the associated miRNAs. In particular, we searched for hsa-mir-30a, that was among the miRNAs with the highest association score (0.74 at the 2nd level) and found a work where it has been significantly associated with other six miRNAs (i.e., miR-29a, miR-223, miR-27a, miR-19b, miR-151-3p, miR-28-5p, miR-766) as biomarker for monitoring the immune status of patients affected by *acquired immunodeficiency syndrome* [38].

Together with hsa-mir-30a, also other miRNAs belonging to the same family (i.e., hsa-mir-30b, -30c and -30e) have been associated by LP-HCLUS with the same disease. In [39], four miRNA-like sequences (i.e., hsa-mir-30d, hsa-mir-30e, hsa-mir-374a and hsa-mir-424) were identified within the env and the gag-pol encoding regions of several HIV-1 strains. The mapping of their sequences within the HIV-1 genomes localized them to the functionally significant variable regions, designated V1, V2, V4 and V5, of the env glycoprotein gp120. This result was important because the regions V1 to V5 of HIV-1 envelopes contain specific and well-characterized domains that are critical for immune responses, virus neutralization and

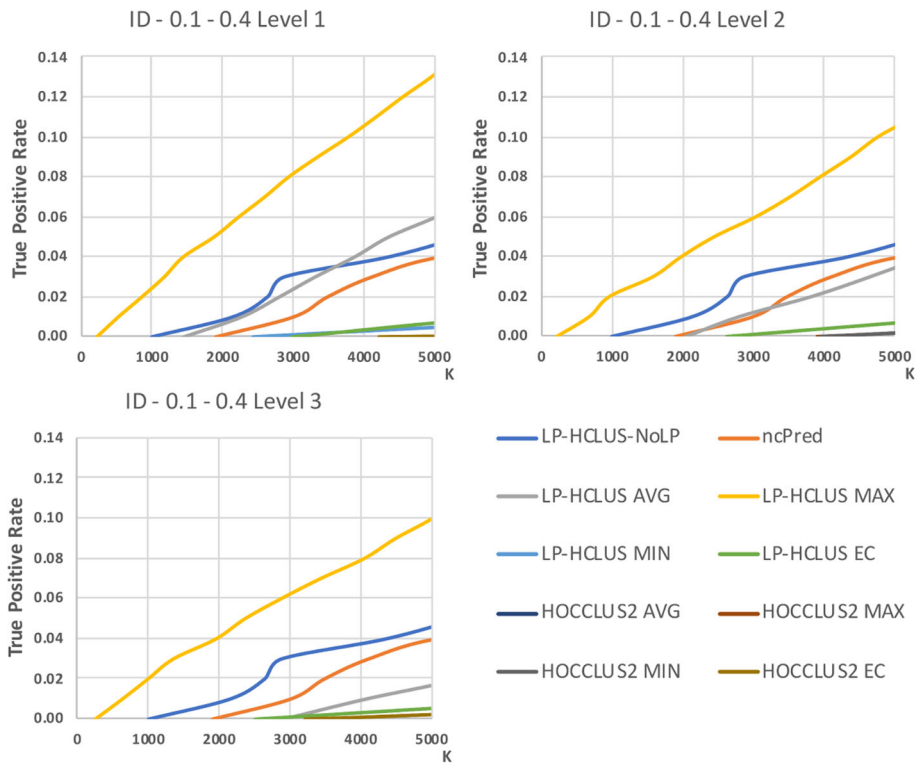

Fig. 14 TPR@*k* results for the dataset ID, obtained with the best configuration ($\alpha = 0.1, \beta = 0.4$) at different levels of the hierarchy

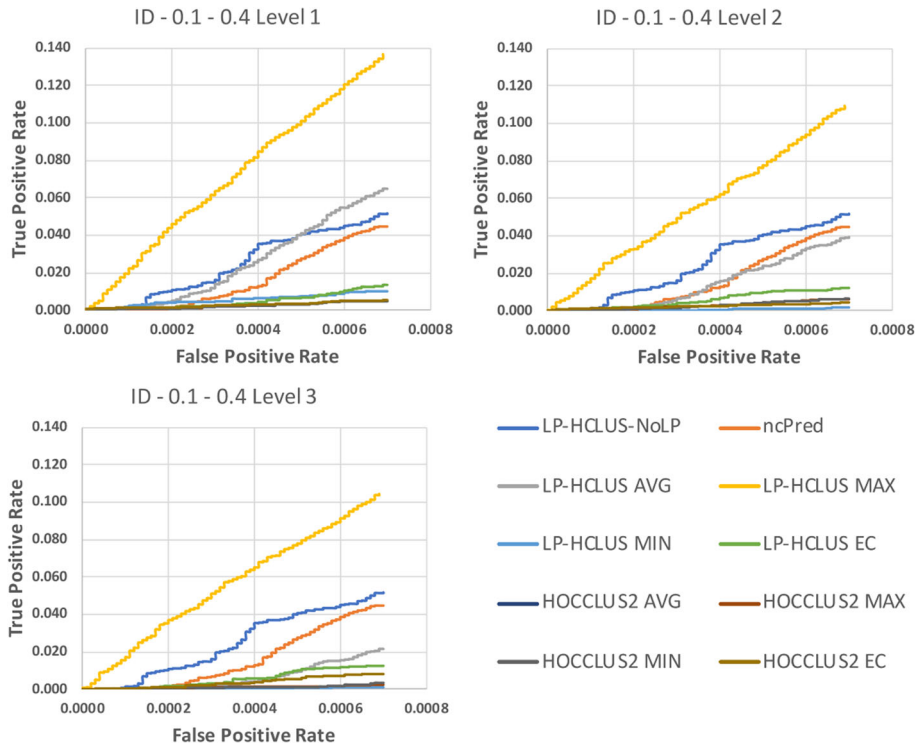

Fig. 15 ROC curves for the dataset ID, obtained with the best configuration ($\alpha = 0.1, \beta = 0.4$) at different levels of the hierarchy. These curves can only be used for relative comparison and not as absolute evaluation measures because they are spoiled by the assumption made on unknown relationships

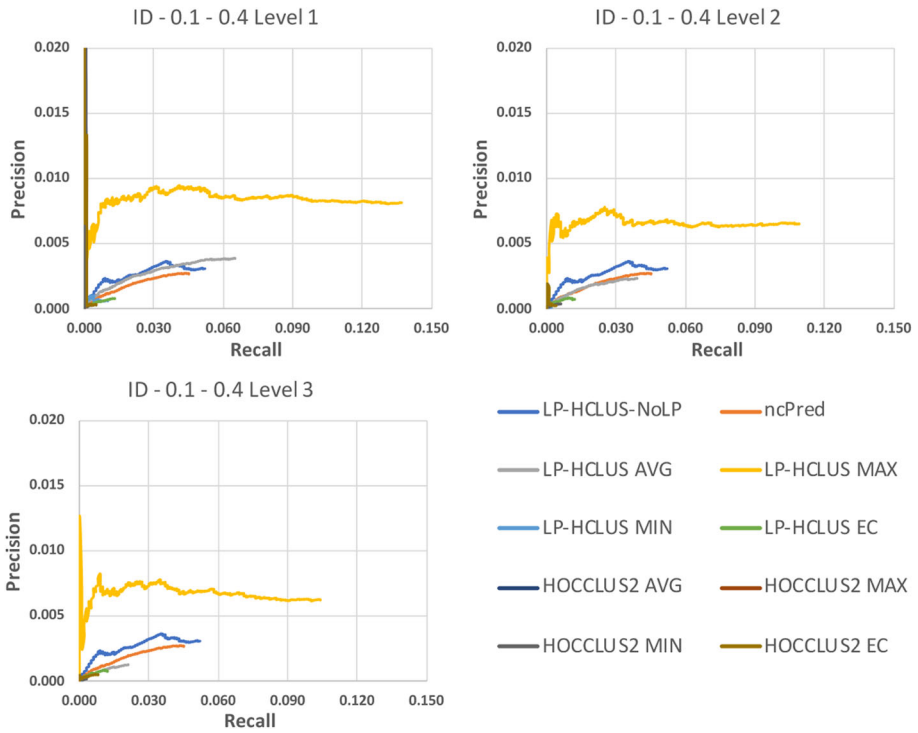

Fig. 16 Precision-Recall curves for the dataset ID, obtained with the best configuration ($\alpha = 0.1, \beta = 0.4$) at different levels of the hierarchy. These curves can only be used for relative comparison and not as absolute evaluation measures because they are spoiled by the assumption made on unknown relationships

disease progression. The authors concluded that the newly discovered miRNA-like sequences in the HIV-1 genomes might have evolved to self-regulated survival of the virus in the host by evading the innate immune responses and therefore influencing persistence, replication or pathogenicity of the virus.

Another example of reliable associations of ncRNAs with the *acquired immunodeficiency syndrome* identified by LP-HCLUS, and not present in HMDD 3.2, are those with hsa-mir-125b, hsa-mir-28 and hsa-mir-382. These associations are confirmed in [40], where the authors provided evidence that these miRNAs can contribute, alongside hsa-mir-155 and hsa-mir-223, to the HIV latency. It is noteworthy that these associations appear only at level 3 of the hierarchy but not at levels 2 or 1.

Altogether, these results highlight two interesting features of LP-HCLUS: the ability to discover meaningful functional associations, and the way the hierarchical clustering can help in the identification of hidden information. In principle, none of the hierarchy levels should be ignored. As shown for the case of the *acquired immunodeficiency syndrome*, the first hierarchical level, although in principle more reliable (since based on more stringent constraints), in some cases is not able to capture less obvious existing associations. On the other hand, results obtained from higher levels of the hierarchy are much more inclusive and can provide pieces of information that, in the lowest levels, are hidden, and that can be pivotal to the specific aims of a research investigation.

Finally, we compared the ranking values assigned by LP-HCLUS, ncPred and HOCCLUS2 on the same associations, that are, those confirmed in the HMDD v3.2 release (see Additional file 5). At this purpose, we computed the AUTPR@k by considering the new interactions introduced in HMDD v3.2 as ground truth. By observing the results reported in Table 6, we can confirm that LP-HCLUS based on the MAX measure outperforms all the competitors in identifying new interactions from the previous version of the dataset (HMDD v3.0) that have been subsequently validated and introduced in the latest version (HMDD v3.2).

Discussion on the integrated dataset

As concerns the ID dataset, we performed a qualitative analysis of the top-ranked relationships predicted by LP-HCLUS, i.e., on those with a score equal to 1.0. For this purpose, we exploited MNDR v2.0 [41], which is a comprehensive resource including more than 260,000

Table 4 AUTPR@k, AUROC and AUPR values for the dataset ID, obtained with the best configuration ($\alpha = 0.1, \beta = 0.4$) at different levels of the hierarchy

		AUTPR@k	AUPR	AUROC
LP-HCLUS-NoLP		0.024087	0.000150	0.525501
ncPred		0.015365	0.000087	0.521975
LP-HCLUS AVG	*Level 1*	0.024335	0.000198	0.532059
	Level 2	0.013660	0.000080	0.519290
	Level 3	0.005883	0.000024	0.510396
LP-HCLUS MAX	*Level 1*	**0.070639**	**0.001218**	**0.567991**
	Level 2	0.054821	0.000780	0.554388
	Level 3	0.055141	0.000756	0.551873
LP-HCLUS MIN	*Level 1*	0.005451	0.000010	0.504690
	Level 2	0.000474	0.000000	0.500490
	Level 3	0.000305	0.000000	0.500154
LP-HCLUS EC	*Level 1*	0.004609	0.000010	0.506366
	Level 2	0.005605	0.000013	0.505695
	Level 3	0.005353	0.000010	0.505862
HOCCLUS2 AVG	*Level 1*	0.002246	0.000087	0.502169
	Level 2	0.002553	0.000006	0.502843
	Level 3	0.000885	0.000001	0.501328
HOCCLUS2 MAX	*Level 1*	0.002238	0.000087	0.502169
	Level 2	0.002659	0.000005	0.502676
	Level 3	0.000973	0.000001	0.500826
HOCCLUS2 MIN	*Level 1*	0.002247	0.000087	0.502169
	Level 2	0.002553	0.000006	0.502843
	Level 3	0.000885	0.000001	0.501328
HOCCLUS2 EC	*Level 1*	0.002763	0.000015	0.502337
	Level 2	0.002320	0.000008	0.501835
	Level 3	0.003533	0.000007	0.503683

The results in terms of AUPR and AUROC values can only be used for relative comparison and not as absolute evaluation measures because they are spoiled by the assumption made on unknown associations, that are considered as negative examples

The best result is highlighted in boldface.

experimental and predicted ncRNA-disease associations for mammalian species, including lncRNA, miRNA, piRNA, snoRNA and more than 1,400 diseases. Data in MNDR comes from manual literature curation and other resources, and include a confidence score for each ncRNA–disease association. Experimental evidences are manually classified as *strong* or *weak*, while the confidence score is calculated according to the evidence type (*s*: strong experimental evidence, *w*: weak experimental evidence, *p*: prediction) and the number of evidences.

The top-ranked relationships returned by LP-HCLUS involve 1,067 different diseases and 814 different ncRNAs, consisting of 488 miRNAs and 326 lncRNAs, among which there are several antisense RNAs and miRNA hosting genes. Table 7 shows some examples of top-ranked interactions predicted by LP-HCLUS and involving 4 ncRNAs, i.e., h19, wrap53, pvt1 and hsa-miR-106b.

h19 is a long intergenic ncRNA (lincRNA) and a developmentally-regulated maternally-imprinted gene that is expressed only from the inherited chromosome 11. A putative function assigned to it is a tumor suppressor activity. GeneCards (GCID:GC11M001995) reports its association with the Wilms Tumor 2 (WT2) and Beckwith-Wiedemann Syndrome, both caused by mutation or deletion of imprinted genes within the chromosome 11p15.5 region. Other sources, such as GenBank [42] and MNDR [41, 43], report the association of h19 with many other human diseases, the majority being different types of tumors.

Searching for h19-disease associations in MNDR, we obtained 101 results with a confidence score ranging from 0.9820 to 0.1097. The same search performed on the output produced by LP-HCLUS (0.1 - 0.4, first level of the hierarchy) returned 993 associations with a score

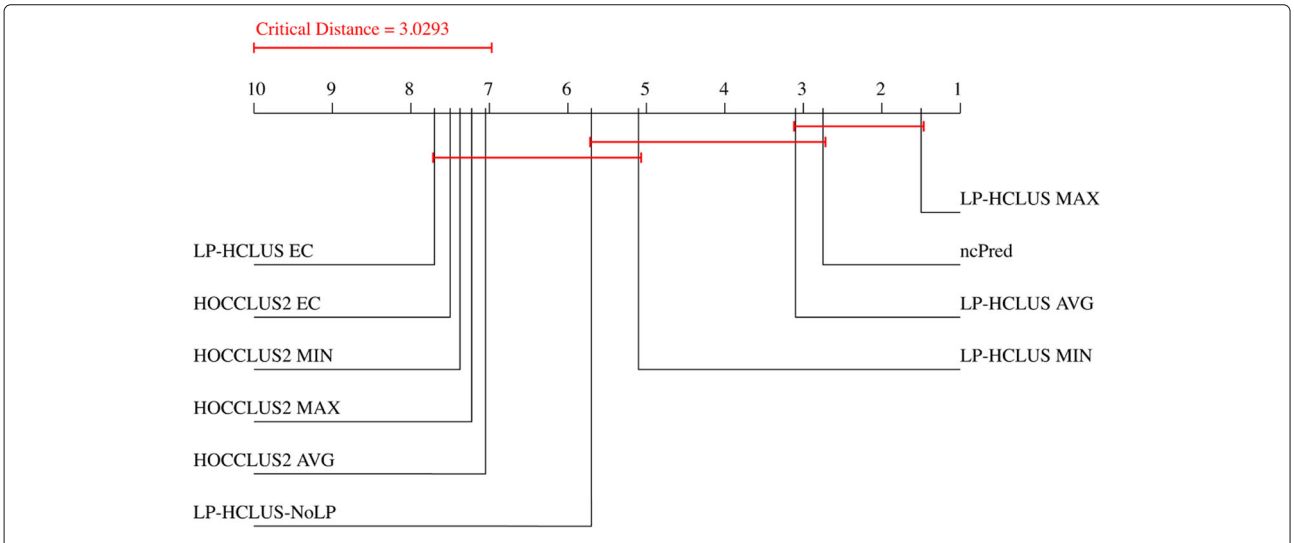

Fig. 17 Result of the Friedman test with Nemenyi post-hoc test, with a significance level of 0.05, performed on the area under the TPR@*k* curve

Table 5 Average Area Under the TPR@*k* curve and *p*-values obtained by the Wilcoxon signed-rank test with the Bonferroni correction

	Average Area Under TPR@*k*		*p*-values
Method	HMDD v3.0 dataset	ID dataset	LP-HCLUS vs ncPred
ncPred	0.087370	0.015365	
LP-HCLUS_MAX_L1	0.088130	**0.070639**	0.005833 (+)
LP-HCLUS_MAX_L2	**0.109292**	0.054821	0.000266 (+)
LP-HCLUS_MAX_L3	0.104244	0.055141	0.000266 (+)

The best result for each dataset is emphasized in boldface. (+) indicates that LP-HCLUS significantly outperforms ncPred (*p*-value <0.01)
The best result is highlighted in boldface.

ranging from 1.0 to 0.4. A comparative analysis of the results shows a perfect match of 33 predictions (see Table 8), many of which also with a similar confidence score, despite the different approaches adopted to calculate them.

Among the top-ranked associations predicted by LP-HCLUS involving h19, the association with "bone diseases, developmental" is not present in the results obtained by the MNDR database (see Table 7). Bone diseases can have different origins and can be also related to hyperfunction or hypofunction of the endocrine glands, such as pituitary gland, thyroid gland, parathyroid glands, adrenal glands, pancreas, gonads, and pineal gland. The results of the comparative analysis with the data in MNDR, in addition to the relationship with osteosarcoma (LP-HCLUS score 0.7732385; MNDR confidence score s: 0.9820) show associations between h19 and other diseases which involve endocrine glands such as: ovarian neoplasms (LP-HCLUS score 0.7052352; MNDR confidence score p: 0.1097, s: 0.8589); pancreatic cancer (LP-HCLUS score 0.8150848; MNDR confidence score s: 0.8808); pancreatic ductal adenocarcinoma (LP-HCLUS score 0.6575157; MNDR confidence score s: 0.9526) and thyroid cancer (LP-HCLUS score 0.7732385; MNDR confidence score s: 0.8808, p: 0.1097) (See Table 8). This indicates that h19 can have a relationship with endocrine glands functions and, therefore, can be related to bone diseases as predicted by LP-HCLUS.

Conclusions

In this paper, we have tackled the problem of predicting possibly unknown ncRNA-disease relationships. The approach we proposed, LP-HCLUS, is able to take advantage from the possible heterogeneous nature of the attributed biological network analyzed. In this way, it is possible to identify ncRNA-disease relationships by taking into account the properties of additional biological entities (e.g. microRNAs, lncRNAs, target genes) they are connected to.

Methodologically, LP-HCLUS is based on the identification of paths in the heterogeneous attributed biological network, which potentially confirm the connection

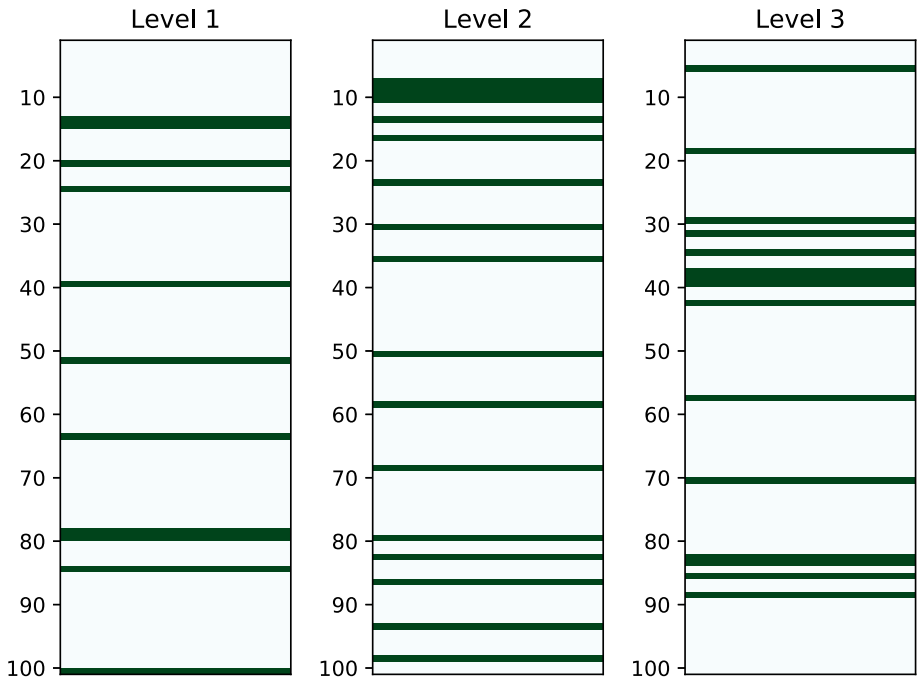

Fig. 18 A graphical representation of the top-100 relationships predicted by LP-HCLUS from HMDD v3.0. The dark green lines represent the position of the relationships that have been subsequently validated and introduced in HMDD v3.2

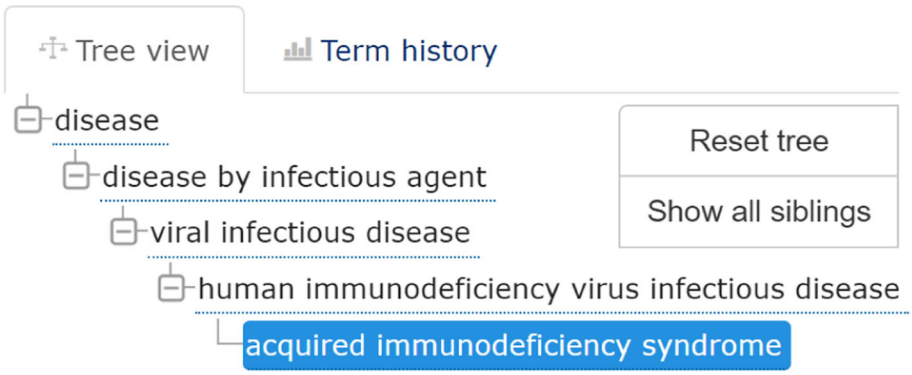

Fig. 19 Ontology classification of *acquired immunodeficiency syndrome* according to EMBL-EBI Ontology Lookup Service [37]

between a ncRNA and a disease, and a clustering phase, which is preparatory to a link prediction phase. In this way, it is possible to catch the network autocorrelation phenomena and exploit information implicitly conveyed by the network structure.

Table 6 AUTPR@k computed using the new associations introduced in the new version of HMDD v3.2 as ground truth

		AUC TPR@k
LP-HCLUS-NoLP		0.00000
ncPred		0.01448
LP-HCLUS AVG	*Level 1*	0.01754
	Level 2	0.02663
	Level 3	0.01453
LP-HCLUS MAX	*Level 1*	0.03247
	Level 2	**0.03423**
	Level 3	0.03111
LP-HCLUS MIN	*Level 1*	0.01846
	Level 2	0.02197
	Level 3	0.00962
LP-HCLUS EC	*Level 1*	0.00695
	Level 2	0.00527
	Level 3	0.00548
HOCCLUS2 AVG	*Level 1*	0.01750
	Level 2	0.00627
	Level 3	0.00962
HOCCLUS2 MAX	*Level 1*	0.01774
	Level 2	0.00763
	Level 3	0.00991
HOCCLUS2 MIN	*Level 1*	0.01657
	Level 2	0.00627
	Level 3	0.00962
HOCCLUS2 EC	*Level 1*	0.01689
	Level 2	0.01269
	Level 3	0.01252

The best result is highlighted in boldface.

The results confirm the initial intuitions and show competitive performances of LP-HCLUS in terms of accuracy of the predictions, also when compared, through a statistical test (at a significance level of 0.01), 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.

Finally, the association between the long-intergenic ncRNA h19 and bone diseases, predicted by LP-HCLUS, suggests an important functional role of h19 in the regulation of endocrine glands functions. This further confirms the potential of LP-HCLUS as a prediction tool for the formulation of new biological hypothesis and experimental

Table 7 Examples of top-ranked ncRNA-disease associations predicted by LP-HCLUS with a score equal to 1.0

ncRNA	Disease
h19	bone diseases, developmental
h19	carcinoma, hepatocellular
h19	colorectal neoplasms
h19	liver neoplasms
h19	parkinson disease, secondary
hsa-miR-106b	aging, premature
hsa-miR-106b	burkitt s lymphomas
hsa-miR-106b	disease progression
pvt1	aging, premature
pvt1	disease progression
wrap53	adrenal gland neoplasms
wrap53	adrenocortical carcinoma
wrap53	emphysema

Table 8 Result of matching between the associations predicted by LP-HCLUS and those present in MNDR

ncRNA	Disease	LP-HCLUS	MNDR
h19	adenocarcinoma	0.7455674	s: 0.7311
h19	adrenocortical carcinoma	0.8150848	s: 0.7311
h19	aortic valve disease	0.6492379	s: 0.7311
h19	astrocytoma	0.7455674	s: 0.7311
h19	breast adenocarcinoma	0.7005121	s: 0.7311
h19	carcinoma, non-small-cell lung	0.7052352	s: 0.9820, p: 0.1097
h19	chronic myeloid leukemia	0.7005121	s: 0.8808
h19	colon carcinoma	0.7005121	s: 0.8589
h19	colorectal cancer	0.8150848	s: 0.9820, p: 0.1097
h19	coronary artery disease	0.6600133	w: 0.4752
h19	embryonal carcinoma	0.6522726	s: 0.9526
h19	endometriosis	0.7052352	s: 0.8808
h19	esophageal cancer	0.8150848	s: 0.8589
h19	gallbladder cancer	0.6522726	s: 0.8808
h19	heart defects, congenital	0.6703589	s: 0.8589
h19	laryngeal squamous cell carcinoma	0.6522726	s: 0.9526
h19	liver neoplasms	1.0000000	w: 0.4752
h19	lung adenocarcinoma	0.6669160	s: 0.8589
h19	lymphoma	0.6962170	p: 0.1321
h19	osteoarthritis	0.6749659	w: 0.4752
h19	osteosarcoma	0.7732385	s: 0.9820
h19	ovarian neoplasms	0.7052352	s: 0.8589, p: 0.1097
h19	pancreatic cancer	0.8150848	s: 0.8808
h19	pancreatic ductal adenocarcinoma	0.6575157	s: 0.9526
h19	polycythemia vera	0.7005121	s: 0.7311
h19	prostatic neoplasms	0.7052352	s: 0.7311, p: 0.1097
h19	rheumatoid arthritis	0.6703589	s: 0.9526
h19	schizophrenia	0.7052352	p: 0.1097
h19	squamous cell carcinoma	0.6826756	w: 0.4752
h19	thyroid cancer	0.7732385	s: 0.8808, p: 0.1097
h19	urinary bladder neoplasms	0.6962170	p: 0.1097
h19	uterine cervical neoplasms	0.7455674	s: 0.7311, p: 0.1097

MNDR scores are associated with an evidence type: *s*: strong experimental evidence, *w*: weak experimental evidence, *p*: prediction

validations for the characterization of the roles of ncRNAs in biological processes.

For future work, we plan to extend our approach in order to predict the direction of the relationships, and not only their presence. This would require to identify and deal with cause/effect phenomena. Depending on the availability of data, it would also be very interesting to evaluate the results of LP-HCLUS analysis on tissue-specific datasets or on datasets related to physiological or pathological specific conditions.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3392-2.

Additional file 1: Discussion of related work.

Additional file 2: Analysis of the time complexity of lP-HCLUS.

Additional file 3: Complete results of the comparative analysis between the predictions returned by lP-HCLUS from hMDD v3.0 and the new validated relationships in hMDD v3.2.

Additional file 4: Detailed list of associations involving the *acquired immunodeficiency syndrome* and similar disease terms in three hierarchical levels extracted by lP-HCLUS.

Additional file 5: Comparative analysis of the ranking produced by lP-HCLUS and its competitors with respect to the new validated relationships in hMDD v3.2.

Abbreviations

AUPR: Area under the Precision-Recall curve; AUROC: Area under the ROC curve; AUTPR@k: Area under the TPR@k curve; AVG: Average; CUI: Concept Unique Identifier; DOID: Human Disease Ontology ID; EC: Evidence Combination; EMBL-EBI: European Molecular Biology Laboratory - European Bioinformatics Institute; GBA: Guilt-By-Association principle; GCID: GeneCards ID; HOCCLUS2: Hierarchical Overlapping Co-CLUStering2; HPO: Human Phenotype Ontology; lncRNA: long non-coding RNA; LP-HCLUS: Link Prediction through Hierarchical CLUStering; MAX: Maximum; MeSH: Medical Subject Headings; MIN: Minimum; miRNA: microRNA; ncRNA: non-coding RNA; OMIM: Online Mendelian Inheritance in Man; RefSeq: NCBI's Reference Sequences database; RNA: RiboNucleic Acid; ROC: Receiver Operating Characteristic; SNP: Single-Nucleotide Polymorphism; TPR@k: True Positive Rate at k; UML: Unified Modeling Language; UMLS: Unified Medical Language System

Acknowledgements

Not applicable

Authors' contributions

MC and GP conceived the task and designed the solution from a methodological point of view. EB and GP implemented the system. EB ran the experiments and collected the results. MC and GP discussed the results from a quantitative viewpoint. DD contributed to the conception of the biological investigation, collaborated to the review and selection of bioinformatics resources and analyzed the results from a qualitative viewpoint. All the authors contributed to the manuscript drafting and approved the final version of the manuscript.

Funding

We would like to acknowledge the financial support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant Number ICT-2013-612944). We also acknowledge the financial support of Ministry of Education, Universities and Research (MIUR) through the PON projects "Big Data Analytics" (AIM1852414 - Activity 1, Line 1) and TALIsMAn - Tecnologie di Assistenza personALizzata per il Miglioramento della quAlità della vitA (Grant N. ARS01_0111), and of Italian National Research Council (CNR) through the InterOmics Flagship project.

Availability of data and materials

The system LP-HCLUS, the adopted datasets and all the results are available at: http://www.di.uniba.it/~gianvitopio/systems/lphclus/

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details
[1]University of Bari Aldo Moro - Department of Computer Science, Via Orabona, 4, 70125 Bari, Italy. [2]Big Data Laboratory, National Interuniversity Consortium for Informatics (CINI), 00185 Rome, Italy. [3]CNR, Institute for Biomedical Technologies, 70126 Bari, Italy. [4]Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia.

Received: 30 August 2019 Accepted: 29 January 2020
Published online: 24 February 2020

References

1. Cech TR, Steitz JA. The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. Cell. 2014;157(1):77–94. https://doi.org/10.1016/j.cell.2014.03.008.
2. Lekka E, Hall J. Noncoding RNAs in disease. FEBS Lett. 2018;592(17):2884–900. https://doi.org/10.1002/1873-3468.13182.
3. Bernstein B, Birney E, Dunham I, Green E, Gunter C, Snyder M, Abyzov A, Aken B, Barrell D, Barson G, Berry A, Bignell A, Boychenko V, Bussotti G, Chrast J, Davidson C, Derrien T, Despacio-Reyes G, Diekhans M, Hubbard T. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
4. Davis C, Hitz B, Sloan C, Chan E, Davidson J, Gabdank I, Hilton J, Jain K, Baymuradov U, Narayanan A, Onate K, Graham K, Miyasato S, Dreszer T, Strattan J, Jolanki O, Tanaka F, Cherry J. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2017;46:. https://doi.org/10.1093/nar/gkx1081.
5. Hayes J, Peruzzi PP, Lawler S. MicroRNAs in cancer: biomarkers, functions and therapy. Trends Mol Med. 2014;20(8):460–9. https://doi.org/10.1016/j.molmed.2014.06.005.
6. Melissari M-T, Grote P. Roles for long non-coding RNAs in physiology and disease. Arch Eur J Physiol. 2016;468(6):945–58. https://doi.org/10.1007/s00424-016-1804-y.
7. Akhade VS, Pal D, Kanduri C. Long Noncoding RNA: Genome Organization and Mechanism of Action. Adv Exp Med Biol. 2017;1008:47–74. https://doi.org/10.1007/978-981-10-5203-3_2.
8. Bak RO, Mikkelsen JG. miRNA sponges: soaking up miRNAs for regulation of gene expression. Wiley Interdiscip Rev RNA. 2014;5(3):317–33. https://doi.org/10.1002/wrna.1213.
9. Yoon J-H, Abdelmohsen K, Gorospe M. Functional interactions among microRNAs and long noncoding RNAs. Semin Cell Dev Biol. 2014;34:9–14. https://doi.org/10.1016/j.semcdb.2014.05.015.
10. Yang X, Gao L, Guo X, Shi X, Wu H, Song F, Wang B. A Network Based Method for Analysis of lncRNA-Disease Associations and Prediction of lncRNAs Implicated in Diseases. PLoS ONE. 2014;9(1):87797. https://doi.org/10.1371/journal.pone.0087797.
11. Wang P, Guo Q, Gao Y, Zhi H, Zhang Y, Liu Y, Zhang J, Yue M, Guo M, Ning S, Zhang G, Li X. Improved method for prioritization of disease associated lncRNAs based on ceRNA theory and functional genomics data. Oncotarget. 2016;8(3):4642–55. https://doi.org/10.18632/oncotarget.13964.
12. Ceci M, Pio G, Kuzmanovski V, Džeroski S. Semi-supervised multi-view learning for gene network reconstruction. PLoS ONE. 2015;10(12):1–27. https://doi.org/10.1371/journal.pone.0144031.
13. Pio G, Ceci M, Malerba D, D'Elia D. ComiRNet: a web-based system for the analysis of miRNA-gene regulatory networks. BMC Bioinformatics. 2015;16(Suppl 9):7. https://doi.org/10.1186/1471-2105-16-S9-S7.
14. Alaimo S, Giugno R, Pulvirenti A. ncPred: ncRNA-Disease Association Prediction through Tripartite Network-Based Inference. Front Bioeng Biotechnol. 2014;2:. https://doi.org/10.3389/fbioe.2014.00071.
15. Bonnici V, Caro GD, Constantino G, Liuni S, D'Elia D, Bombieri N, Licciulli F, Giugno R. Arena-ldb: a platform to build human non-coding RNA interaction networks. BMC Bioinformatics. 2018;19(Suppl 10):. https://doi.org/10.1186/s12859-018-2298-8.
16. Pio G, Ceci M, Prisciandaro F, Malerba D. LOCANDA: Exploiting Causality in the Reconstruction of Gene Regulatory Networks. In: Yamamoto A, Kida T, Uno T, Kuboyama T, editors. Discovery Science. Cham: Springer; 2017. p. 283–97.
17. Pio G, Ceci M, Prisciandaro F, Malerba D. Exploiting causality in gene network reconstruction based on graph embedding. Mach Learn. 2019. https://doi.org/10.1007/s10994-019-05861-8.
18. 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. 2014;15(Suppl 1):4. https://doi.org/10.1186/1471-2105-15-S1-S4.
19. Mignone P, Pio G, D'Elia D, Ceci M. Exploiting transfer learning for the reconstruction of the human gene regulatory network. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz781.
20. Chen X, Yan CC, Luo C, Ji W, Zhang Y, Dai Q. Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. Sci Rep. 2015;5:. https://doi.org/10.1038/srep11338.
21. Martínez V, Berzal F, Cubero J-C. A Survey of Link Prediction in Complex Networks. ACM Comput Surv. 2016;49(4):69–16933. https://doi.org/10.1145/3012704.
22. Blockeel H, Raedt LD, Ramon J. Top-down induction of clustering trees. In: Shavlik JW, editor. Proc. of ICML 1998. Madison: Morgan Kaufmann; 1998. p. 55–63.
23. Dincer NG, Akkuş Ö. A new fuzzy time series model based on robust clustering for forecasting of air pollution. Ecol Inform. 2018;43:157–64.
24. Stojanova D, Ceci M, Appice A, Dzeroski S. Network regression with predictive clustering trees. Data Min Knowl Disc. 2012;25(2):378–413.
25. Lefever S, Anckaert J, Volders P-J, Luypaert M, Vandesompele J, Mestdagh P. decodeRNA— predicting non-coding RNA functions using guilt-by-association. Database: J Biol Databases Curation. 2017;2017:. https://doi.org/10.1093/database/bax042.
26. Pio G, Serafino F, Malerba D, Ceci M. Multi-type clustering and classification from heterogeneous networks. Inf Sci. 2018;425:107–26. https://doi.org/10.1016/j.ins.2017.10.021.
27. Zadeh LA. Fuzzy sets. Inf Control. 1965;8(3):338–53. https://doi.org/10.1016/S0019-9958(65)90241-X.
28. Han J, Kamber M. Data Mining: Concepts and Techniques. Amsterdam: Elsevier/Morgan Kaufmann; 2006.
29. 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. 2013;14(Suppl 7):8. https://doi.org/10.1186/1471-2105-14-S7-S8.
30. Lesmo L, Saitta L, Torasso P. Evidence combination in expert systems. Int J Man-Mach Stud. 1985;22(3):307–26. https://doi.org/10.1016/S0020-7373(85)80006-7.
31. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q. HMDD v3.0: a database for experimentally supported human microRNA-disease associations. Nucleic Acids Res. 2019;47(D1):1013–7. https://doi.org/10.1093/nar/gky1010.
32. 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. 2013;41(Database issue):983–6. https://doi.org/10.1093/nar/gks1099.
33. Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell. 2013;153(3):654–65. https://doi.org/10.1016/j.cell.2013.03.043.
34. Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI. DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. Bioinforma (Oxf Engl). 2010;26(22):2924–6. https://doi.org/10.1093/bioinformatics/btq538.
35. 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. 2009;37(Database issue):98–104. https://doi.org/10.1093/nar/gkn714.
36. Chiang K, Sung T-L, Rice AP. Regulation of Cyclin T1 and HIV-1 Replication by MicroRNAs in Resting CD4+ T Lymphocytes. J Virol. 2012;86(6):3244–52. https://doi.org/10.1128/JVI.05065-11. https://jvi.asm.org/content/86/6/3244.full.pdf.
37. Jupp S, et al. A new Ontology Lookup Service at EMBL-EBI. In: Malone J, et al., editors. Proceedings of SWAT4LS International Conference 2015; 2015.
38. Qi Y, Hu H, Guo H, Xu P, Shi Z, Huan X, Zhu Z, Zhou M, Cui L. MicroRNA profiling in plasma of HIV-1 infected patients: potential markers of infection and immune status. J Publ Health Emerg. 2017;1(7):. https://doi.org/10.21037/jphe.2017.05.11.
39. Holland B, Wong J, Li M, Rasheed S. Identification of Human MicroRNA-Like Sequences Embedded within the Protein-Encoding Genes of the Human Immunodeficiency Virus. PLoS ONE. 2013;8(3):1–10. https://doi.org/10.1371/journal.pone.0058586.

40. Huang J, Wang F, Argyris E, Chen K, Liang Z, Tian H, Huang W, Squires K, Verlinghieri G, Zhang H. Cellular microrna contribute to hiv-1 latency in resting primary cd4+ t lymphocytes. Nat Med. 2007;13(10):1241–7. https://doi.org/10.1038/nm1639.
41. Cui T, Zhang L, Huang Y, Yi Y, Tan P, Zhao Y, Hu Y, Xu L, Li E, Wang D. MNDR v2.0: an updated resource of ncRNA–disease associations in mammals. Nucleic Acids Res. 2018;46(Database issue):371–4. https://doi.org/10.1093/nar/gkx1025.
42. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2013;41(Database issue):36–42. https://doi.org/10.1093/nar/gks1195.
43. Wang Y, Chen L, Chen B, Li X, Kang J, Fan K, Hu Y, Xu J, Yi L, Yang J, Huang Y, Cheng L, Li Y, Wang C, Li K, Li X, Xu J, Wang D. Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network. Cell Death Dis. 2013;4(8):765. https://doi.org/10.1038/cddis.2013.292.

Publisher's Note