Identifying lncRNA-disease Relationships via Heterogeneous Clustering

Emanuele Pio Barracchia[1], Gianvito Pio[1], Donato Malerba[1,2] and Michelangelo Ceci[1]

[1]University of Bari Aldo Moro
Department of Computer Science - Via Orabona, 4 - 70125 Bari, Italy

[2]CINI - Consorzio Interuniversitario Nazionale per l'Informatica - Bari
{gianvito.pio,michelangelo.ceci,donato.malerba}@uniba.it,
e.barracchia@studenti.uniba.it

(Extended Abstract)

Abstract. High-throughput sequencing technology led significant advances in functional genomics, giving the opportunity to pay particular attention to the role of long non-coding RNAs (lncRNAs) in the development of human diseases. 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. Results obtained by preliminary experiments, performed on an integrated dataset about microRNAs, lncRNAs, diseases and genes, show that the proposed method is able to obtained better results with respect to an existing method.

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) whose function is pivotal for the fine-tuning of the expression of many genes [3]. Therefore, 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), 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. 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 [15] 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 is based on a heterogeneous clustering approach. In particular, the proposed method is able to analyze heterogeneous networks, where nodes are biological entities (each associated with their own features) and edges represent known relationships among them (see Figure 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 is designed to identify 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 the proposed heterogeneous clustering method and its exploitation to identify unknown lncRNA-disease relationships, while in Section 3 we report the results of some preliminary experiments. Finally, in Section 4, we draw some conclusions and outline the ongoing work.

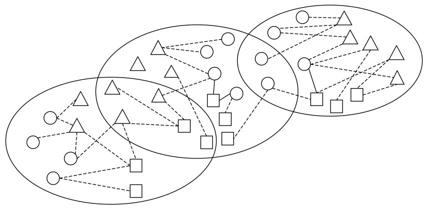

Fig. 1. An example of a heterogeneous network, where different shapes represent different node types. Circles represent possible heterogeneous clusters.

2 Method

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

Def. 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 $\mathcal{X}_p = \{X_{p,1}, X_{p,2}, \ldots, X_{p,m_p}\}$;
- 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.

Def. 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.

Def. 3 (Hierarchical organization of clusters). A hierarchy of heterogeneous clusters is defined as a list of hierarchy levels $\{L_1, L_2, \ldots, L_k\}$, 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 $\{L_1, L_2, \ldots, L_k\}$ 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: identification of the strength of relationships among nodes in T_l and T_d, identification of a hierarchy of heterogeneous clusters, and 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: for each pair (l_i, d_j), we compute the score $s(l_i, d_j)$ by analyzing the indirect relationships in which the considered lncRNA and disease are involved.

The score of a lncRNA-disease pair (l_i, d_j) is computed by identifying and analyzing c shortest paths that connect them in the heterogeneous network. In particular, for each 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 path P. It is noteworthy that several paths can be identified between two objects in the network, possibly with unlimited length (in presence of cycles). Therefore, the score associated to the pair is computed as the maximum score obtained over the c shortest paths. This choice guarantees us to catch the strongest interaction between the objects.

Each path P is represented 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, it is computed as the maximum similarity between the sequences which start with l_i and the sequences which end with d_j. 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 a numerical attribute x is computed as $1 - \frac{|val_x(seq') - val_x(seq'')|}{max_x - min_x}$ (min_x and max_x are the minimum and maximum values, respectively, observed for the attribute x). If x is not numeric, then $s_x(seq', seq'') = 1$ if $val_x(seq') = val_x(seq'')$, 0 otherwise.

Nodes belonging to node types that are not involved in any path are aggregated according to the *arithmetic mean* for numeric attributes and the *mode* for attributes of any other type, and are associated to the nodes connected to them.

2.2 Building a hierarchy of heterogeneous clusters

Once all the possible pairs are identified, each associated with its strength score, we first build a set of (possibly overlapping) clusters in the form of cliques to be used in the subsequent step. A cluster is in the form of a clique 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) β. The result is the subset of pairs $\{(l_i, d_j) | s(l_i, d_j) \geq \beta\}$.
ii) Building of a set of cliques, each consisting of a 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 clique. 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 *cohesiveness*, which is defined as: $h(G) = \frac{1}{|pairs(G)|} \cdot \sum_{(l_i, d_j) \in pairs(G)} s(l_i, d_j)$. Formally, $G' <_c G'' \iff h(G') > h(G'')$.

Once the first level L_1 of the hierarchy has been identified, the other levels are built by evaluating whether some pairs of clusters (cliques, 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 cliques, we work on generic clusters, where the strength score associated to each pair is not necessarily greater than β. Due to this difference, and inspired by [12], two clusters G' and G'' are merged into a cluster G''' if $h(G''') > \alpha$, where α is a user defined threshold. Note that low values of α 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 $\{L_1, L_2, \ldots, L_k\}$, according to Def. 3.

2.3 Prediction of unknown relationships

After building the hierarchy of clusters, we identify possibly unknown relationships for each level of the hierarchy. In particular, the prediction is performed by assigning each possible lncRNA-disease pair with the score computed as the cohesiveness of the cluster in which it falls, which intuitively represents the certainty of the relationship. When a lncRNA-disease pair appears in multiple clusters, we combine the cohesiveness of the set of clusters to obtain the final score. 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 evidence combination (EC) strategy proposed in [10], given $C_{ij}^{(w)} = [C_1, C_2, \ldots, C_m]$, the list of the clusters in which the lncRNA l_i and the disease d_j fall in the w-th hierarhical level, $\psi^{(w)}(l_i, d_j)$ is recursively computed as $ec(C_m)$, where:

$ec(C_1) = h(C_1)$

$ec(C_m) = ec(C_{m-1}) + [1 - ec(C_{m-1})] \cdot h(C_m)$

3 Experiments

The proposed method has been implemented in the system LP-HCLUS (Link Prediction through Heterogenous CLUStering). We performed some preliminary experiments 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 Figure 2, has been built by integrating several existing biological datasets: lncRNA-disease relationships and lncRNA-gene interactions are taken from [5]; miRNA-lncRNA interactions are taken from [8]; disease-gene relationships are taken from DisGeNET [2]; miRNA-gene interactions and miRNA-disease relationships are taken from miR2Disease [9]. The obtained 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.

As a competitor system, we considered a biclustering algorithm, called HOCCLUS2 [12], which is tailored to work with two types of nodes and that also builds a hierarchy of clusters. We fed HOCCLUS2 with the set of lncRNA-disease scores computed by LP-HCLUS, since, in its original form, it is not able to analyze a complex heterogeneous network. Following the results in [12], for both LP-HCLUS and HOCCLUS2, we set $\beta = 0.2$ and $\alpha = 0.0$. Note that $\alpha = 0.0$ let

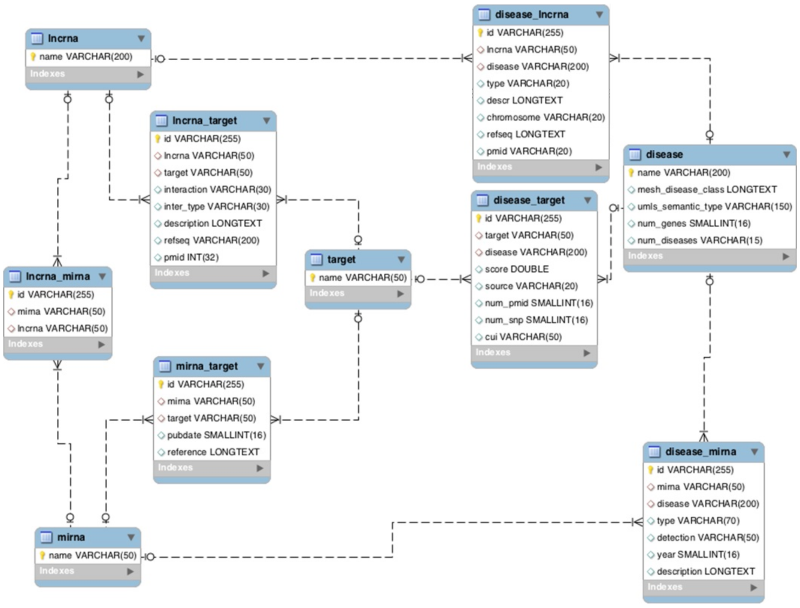

Fig. 2. UML representation of the heterogeneous network used in the evaluation.

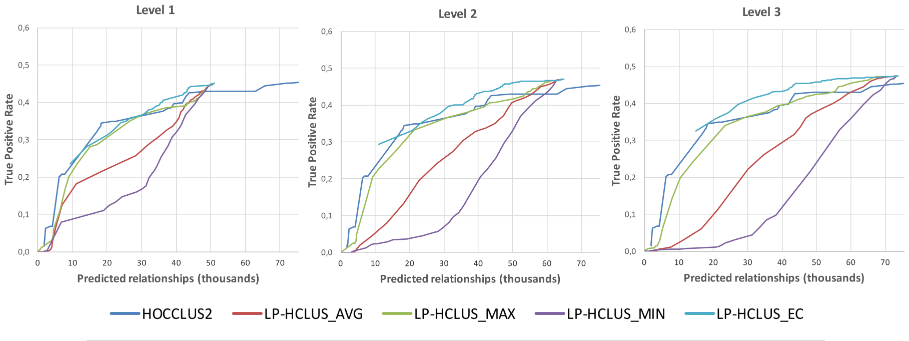

Fig. 3. TPR obtained considering the top-k relationships by varying the threshold.

the algorithm proceed until it reaches a single cluster containing all the objects. However, we consider only the first 3 levels of the identified hierarchies which, according to [12], lead to the best results.

We adopted the 10-fold cross validation on the set of known lncRNA-disease relationships. Due to the absence of negative examples, we averaged the results obtained in terms of True Positive Rate, defined as $TPR = \frac{TP}{TP+FN}$, where TP is the number of validated lncRNA-disease relationships that were predicted with a score greater than the threshold, and FN is the number of validated lncRNA-disease relationships that were predicted with a score lower than the threshold.

Inspired by [13], in which the authors performed a similar evaluation in the absence of negative examples, we vary the value of such a threshold and plot

	Threshold									
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
l1_HOCCLUS2	**0.490**	**0.486**	0.452	0.427	0.396	0.345	0.208	0.063	0.020	0.014
l1_LP-HCLUS_AVG	0.452	0.452	0.450	0.447	0.396	0.257	0.002	0.000	0.000	0.000
l1_LP-HCLUS_MAX	0.452	0.452	0.450	0.450	0.423	0.391	0.094	0.005	0.000	0.000
l1_LP-HCLUS_MIN	0.452	0.452	0.450	0.381	0.208	0.110	0.000	0.000	0.000	0.000
l1_LP-HCLUS_EC	0.452	0.452	0.450	0.450	0.448	0.447	0.439	0.427	0.400	0.362
l2_HOCCLUS2	**0.490**	**0.486**	0.452	0.427	0.396	0.345	0.208	0.063	0.020	0.014
l2_LP-HCLUS_AVG	0.470	0.470	0.470	0.458	0.371	0.080	0.000	0.000	0.000	0.000
l2_LP-HCLUS_MAX	0.470	0.470	0.470	0.467	0.436	0.389	0.047	0.000	0.000	0.000
l2_LP-HCLUS_MIN	0.470	0.470	0.470	0.329	0.067	0.013	0.000	0.000	0.000	0.000
l2_LP-HCLUS_EC	0.470	0.470	0.470	0.467	0.467	0.463	0.456	0.445	0.429	0.398
l3_HOCCLUS2	**0.490**	**0.486**	0.452	0.427	0.396	0.345	0.208	0.063	0.020	0.014
l3_LP-HCLUS_AVG	0.474	0.474	**0.474**	0.467	0.336	0.007	0.000	0.000	0.000	0.000
l3_LP-HCLUS_MAX	0.474	0.474	**0.474**	**0.472**	0.445	0.378	0.014	0.000	0.000	0.000
l3_LP-HCLUS_MIN	0.474	0.474	0.472	0.217	0.013	0.004	0.000	0.000	0.000	0.000
l3_LP-HCLUS_EC	0.474	0.474	**0.474**	**0.472**	**0.470**	**0.467**	**0.461**	**0.456**	**0.450**	**0.427**

Table 1. TPR obtained by HOCCLUS2 and LP-HCLUS at different hierarchical level.

a graph where each point represents the TPR obtained by selecting the top-k identified relationships (known as recall@k in the information retrieval context).

From the results reported in Table 1 and from the graphs depicted in Figure 3, LP-HCLUS is able to outperform HOCCLUS2 in all the hierarchical levels when we use the *EC* combination function. In particular, focusing on Figure 3, we can observe that LP-HCLUS with the EC measure needs to identify less relationships to achieve a given True Positive Rate. Moreover, *EC* leads to better performances also when compared with the other combination strategies, whose performances are often worse than HOCCLUS2. As expected, the *min* strategy is the most conservative, the *max* strategy shows a trend which is similar to HOCCLUS2, while *avg* strategy is always in the middle between *min* and *max*.

As a final remark, we remind that HOCCLUS2 was able to obtain comparable results since we provided it with the lncRNA-disease scores computed by LP-HCLUS. For this reason, we are performing experiments with other competitors to evaluate the contribution of the heterogeneous data on the final results.

4 Conclusions

In this work, we proposed the method LP-HCLUS, based on heterogeneous clustering, which is able to predict possibly unknown lncRNA-disease relationships, that can be exploited for better understanding the role of lncRNAs in the development of human diseases. Preliminary experiments showed that the proposed method, especially when adopting the strategy based on evidence combination, is able to outperform the algorithm HOCCLUS2. We are currently performing additional experiments in order to understand the effectiveness of the proposed approach when compared to further competitor systems, as well as to understand the real contribution provided by the analysis of heterogeneous data in the identification of relationships between biological entities.

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

1. Alaimo, S., Giugno, R., Pulvirenti, A.: ncPred: ncRNA-Disease Association Prediction through Tripartite Network-Based Inference. Frontiers in Bioengineering and Biotechnology (2014)
2. 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)
3. Cech, T., Steitz, J.: The Noncoding RNA Revolution-Trashing Old Rules to Forge New Ones. Cell 157(1), 77 – 94 (2014)
4. Ceci, M., Pio, G., Kuzmanovski, V., Deroski, S.: Semi-supervised multi-view learning for gene network reconstruction. PLOS ONE 10(12), 1–27 (12 2015)
5. Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., Cui, Q.: LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic acids research 41(D1), D983–D986 (2013)
6. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, Second Edition. Morgan Kaufmann, 2 edn. (Jan 2006)
7. Hayes, J., Peruzzi, P.P., Lawler, S.: MicroRNAs in cancer: biomarkers, functions and therapy. Trends in Molecular Medicine 20(8), 460 – 469 (2014)
8. 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)
9. Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., Liu, Y.: miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic acids research 37(suppl 1), D98–D104 (2009)
10. Lesmo, L., Saitta, L., Torasso, P.: Evidence combination in expert systems. International Journal of Man-Machine Studies 22(3), 307 – 326 (1985)
11. Melissari, M.T., Grote, P.: Roles for long non-coding RNAs in physiology and disease. Pflügers Archiv - European Journal of Physiology 468(6), 945–958 (2016)
12. Pio, G., Ceci, M., D'Elia, D., Loglisci, C., Malerba, D.: A Novel Biclustering Algorithm for the Discovery of Meaningful Biological Correlations between microRNAs and their Target Genes. BMC Bioinformatics 14(S-7), S8 (2013)
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 16(9), S7 (2015)
14. 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)
15. 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)