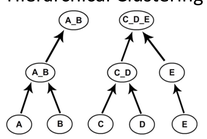
Multi-Type Clustering for the Identification of lncRNA-disease Relationships

Gianvito Pio[1], Francesco Serafino[1], Emanuele Pio Barracchia[1], Domenica D'Elia[2], Donato Malerba[1], Michelangelo Ceci[1]
[1]Department of Computer Science, University of Bari, Italy [2] CNR - Institute for Biomedical Technologies, Bari, Italy
Email: {gianvito.pio, francesco.serafino, donato.malerba, michelangelo.ceci}@uniba.it, e.barracchia@studenti.uniba.it, domenica.delia@ba.itb.cnr.it

High-throughput sequencing technology has been crucial for rapid advances in functional genomics. The most important result is the discovery of thousands of non-coding RNAs (ncRNAs) which are able to fine-tune the expression of many genes involved in cell development, differentiation, apoptosis and proliferation [1]. Among ncRNAs, the most investigated are the microRNAs (miRNAs), small molecules (20-22 nucleotide long) that play the role of post-transcriptional regulators [2]. Much less is known about the functional role of long non-coding RNAs (lncRNAs), RNA molecules longer than 200 nt, that have been recently discovered to have a plethora of regulatory functions spanning from epigenetics to post-transcriptional regulation [3]. However, the number of lncRNAs for which the functional characterisation is available is still quite poor. Most of existing approaches are based on expensive experimental evaluations or on computational methods that exploit known/verified relationships among the lncRNA and the disease [4]. Some recent works consider the assumption that all the instances follow the same probability distribution and that are independent to each other. In this case such assumption is easily violated, since different lncRNAs can be involved in the development of the same disease, as well as different diseases can be related to each other on the basis of the involvement of common lncRNAs. To overcome these limitations we propose a computational method which is able to predict possible unknown relationships between lncRNA and diseases by exploiting different information about an heterogeneous set of (related) biological entities. In particular, we focus on lncRNAs, miRNAs, target genes and diseases, as well as on known relationships among these entities. The proposed method is based on a clustering algorithm which is able to group objects of multiple types and to predict possible unknown relationships on the basis of the extracted clusters. Moreover, the proposed clustering algorithm is designed to identify highly cohesive, possibly overlapping and hierarchically organised clusters, since i) the same lncRNA/disease can be involved in multiple networks of relationships and ii) as shown in [5], clusters at different levels of the hierarchy can describe more specific or more general relationships and cooperation activities.

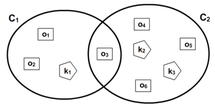
Multi-type Clustering

- Hierarchical Clustering



- Facilitates the understanding of results by human experts.
- Allows the possibility of choosing globally-based or locally-based predictions.

- Multi-type and overlapping

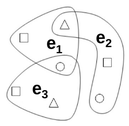


- Catches relationships among objects of multiple types, which can be involved in different sub-networks.

Identification of relationships

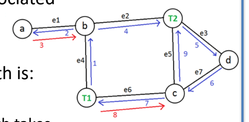
Data are represented as a **hypergraph**

- **Tuples** of objects (different types)
- s_i is the strength value for the i-th **tuple**

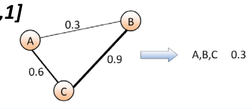


Pairwise evaluation of nodes in the tuple

- **Paths** exploration in the network between pairs of target nodes by computing a **strength score** associated to each **path.**
- The **maximum** score computed over all the paths is selected. The score with respect to a path is:
  - **1**, if the two nodes are **connected** in the network
  - **computed** according to a **similarity function** which takes into account all the attributes associated to each node involved in the path



The minimum value over the pairs of objects in the tuple ***i*** is selected as strength $s_i \in]0,1]$



Hierarchical Clustering

Requires a user-defined threshold **β**

A bottom cluster, in the form of a **multi-type clique**, is built for each tuple

If its strength is greater than β

Obtained clusters are ordered according to their **cohesiveness** (average strength of the relationships in the cluster)

$$cohes(G') = \frac{1}{|tuples(G')|} \cdot \sum_{\vec{v_i} \in tuples(G')} s_i$$

1st hierarchical level

Find the candidates for merging

- According to cohesiveness
- New clusters must still be a **clique**!

- **Result**:

The first level of the hierarchy L_1

Clusters merging

Find the candidates for merging

- According to cohesiveness
- New cluster cohesiveness must be greater than an user-defined threshold **α**

$$cohes(G') > \alpha$$

- Clique constraint is relaxed!

- **Result:**

The subsequent levels of the hierarchy $\{L_2, L_3, \ldots, L_k\}$

Prediction of relationships

1. Extraction of **experimentally evaluated** interactions from the network
2. Extraction of a **hierarchy of multi-type clusters** by applying our multi-type clustering algorithm
3. Generation of all the possible **new interactions**
4. For each hierarchical level:
   - Determine a **score** for each previously unknown interaction as the **combination** of the cohesiveness values associated with the clusters it belongs to
   - Return **new interactions** with their associated **score**

Computation of the prediction score

Let $s(t_i)$ be the score associated to the tuple t_i

Let G_{t_i} be the set of clusters t_i belongs to

If $|G_{t_i}| = 1$ and $G' \in G_{t_i}$, $s(t_i) = cohes(G')$

Otherwise, we consider some combination strategies:

- **Max:** $s(t_i) = \max_{G_j' \in G_{t_i}} cohes(G_j')$
- **Min:** $s(t_i) = \min_{G_j' \in G_{t_i}} cohes(G_j')$
- **Avg:** $s(t_i) = \frac{1}{|G_{t_i}|} \sum_{G_j' \in G_{t_i}} cohes(G_j')$
- **Custom:** $s(t_i) = f(G_{t_i})$

Custom combination strategy

Clusters in G_{t_i} are sorted according to their cohesiveness in descending ordering $G_{t_i} = [G'_1, G'_2, \ldots, G'_n]$

$$s(t_i) = f(G_{t_i}) = g(G'_n)$$

where:

$$g(G'_1) = cohes(G'_1)$$
$$g(G'_{n+1}) = g(G'_n) + \left(1 - g(G'_n)\right) * cohes(G'_{n+1})$$

Experimental setting

Systems

LP-MTRCLUS (Our System)

HOCCLUS2 [5]

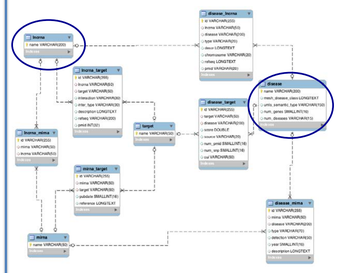
Evaluation measure

True Positive Rate $TPR = \frac{TP}{TP+FN}$

Evaluation performed

- By moving a threshold on the prediction scores
- On the first three hierarchical levels

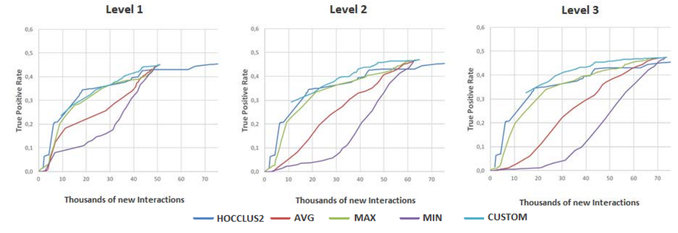
Datasets



The obtained database consists of 7050 **diseases**, 507 **lncRNAs**, 508 **miRNAs**, 94527 **genes** and the following interactions:

- 953 interactions **lncRNAs:diseases** taken from [6]
- 252 interactions **lncRNAs:target_genes** taken from [6]
- 70 interactions **miRNAs:lncRNAs** taken from [7]
- 26522 interactions **diseases:genes** taken from DisGeNET [8]
- 803 interactions **miRNAs:genes** taken from miR2Disease [9]
- 2877 interactions **miRNAs:diseases** taken from miR2Disease [9]

Results



Conclusions

In this work, we focused on the recognised role of lncRNAs in human diseases. In particular, we proposed a computational method which is able to predict possibly unknown lncRNA-disease relationships by exploiting a clustering algorithm which works on multiple types of objects. Preliminary experiments showed that the proposed method, especially when adopting the proposed combination strategy, is able to outperform the algorithm HOCCLUS2 [5]. Currently we are performing additional experiments with other competitor approaches to deeply evaluate the effectiveness of the clustering-based method for this purpose as well as the effect of the exploitation of information about related biological entities, such as miRNAs, genes and their relationships with diseases and lncRNAs.

Aknowlegments: We would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

References

[1] T. Cech and J. Steitz. The Noncoding {RNA} Revolution—Trashing Old Rules to Forge New Ones. Cell, 157(1):77 – 94, 2014.
[2] J. Hayes, P. P. Peruzzi, and S. Lawler. MicroRNAs in cancer: biomarkers, functions and therapy. Trends in Molecular Medicine, 20(8):460 – 469, 2014.
[3] M.-T. Melissari and P. Grote. Roles for long non-coding RNAs in physiology and disease. Pflügers Archiv - European Journal of Physiology, 468(6):945–958, 2016.
[4] S. Jalali, S. Kapoor, A. Sivadas, D. Bhartiya, and V. Scaria. Computational approaches towards understanding human long noncoding rna biology. Bioinformatics, 2015.
[5] G. Pio, M. Ceci, D. D'Elia, C. Loglisci, and D. Malerba. A Novel Biclustering Algorithm for the Discovery of Meaningful Biological Correlations between microRNAs and their Target Genes. BMC Bioinformatics, 14(S-7):S8, 2013.
[6] G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan, and Q. Cui. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic acids research, 41(D1):D983–D986, 2013.
[7] A. Helwak, G. Kudla, T. Dudnakova, and D. Tollervey. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell, 153(3):654–665, 2013.
[8] A. Bauer-Mehren, M. Rautschka, F. Sanz, and L. I. Furlong. DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. Bioinformatics, 26(22):2924–2926, 2010.
[9] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, and Y. Liu. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic acids research, 37(suppl 1):D98–D104, 2009.