Prediction of lncRNA-disease associations from tripartite graphs

Mariella Bonomo[1], Armando La Placa[1], and Simona E. Rombo[1]

Department of Mathematics and Computer Science, University of Palermo
{mariella.bonomo,armando.laplaca}@community.unipa.it
simona.rombo@unipa.it

Abstract. The discovery of novel lncRNA-disease associations may provide valuable input to the understanding of disease mechanisms at lncRNA level, as well as to the detection of biomarkers for disease diagnosis, treatment, prognosis and prevention. Unfortunately, due to costs and time complexity, the number of possible disease-related lncRNAs verified by traditional biological experiments is very limited. Computational approaches for the prediction of potential disease-lncRNA associations can effectively decrease time and cost of biological experiments. We propose an approach for the prediction of lncRNA-disease associations based on neighborhood analysis performed on a tripartite graph, built upon lncRNAs, miRNAs and diseases. The main idea here is to discover hidden relationships between lncRNAs and diseases through the exploration of their interactions with intermediate molecules (e.g., miRNAs) in the tripartite graph, based on the consideration that while a few of lncRNA-disease associations are still known, plenty of interactions between lncRNAs and other molecules, as well as associations of the latters with diseases, are available. The effectiveness of our approach is proved by its ability in the identification of associations missed by competitors, on real datasets.

Keywords: lncRNA-disease associations prediction · tripartite graphs · decision support

1 Introduction

Long-non-coding RNAs (lncRNAs) are molecules emerging as key regulators of various critical biological processes, and their alterations and dysregulations have been associated with many important complex diseases [9]. The discovery of novel disease-lncRNA associations may provide valuable input to the understanding of disease mechanisms at lncRNA level, as well as to the detection of disease biomarkers for disease diagnosis, treatment, prognosis and prevention. Unfortunately, due to costs and time complexity, the number of possible disease-related lncRNAs that can be verified by traditional biological experiments is very limited. Computational approaches for the prediction of potential disease-lncRNA associations can effectively decrease the time and cost of biological experiments. Computational models quantify the association probability

of each lncRNA-disease pair, thus allowing for the identification of the most promising lncRNA-disease pairs to be further verified in laboratory. Such predictive approaches often rely on the analysis of lncRNAs related information stored in public databases, e.g., their interaction with other types of molecules [1, 3, 10, 14]. As an example, large amounts of lncRNA-miRNA interactions have been collected in public databases, and plenty of experimentally confirmed miRNA-disease associations are available as well.

We propose a novel computational approach for the prediction of lncRNA-disease associations (LDAs), based on known lncRNA-miRNA interactions (LMIs) and miRNA-disease associations (MDAs). In particular, we model the problem of LDAs prediction as a neighborhood analysis performed on tripartite graphs in which the three sets of vertices represent lncRNAs, miRNAs and diseases, respectively, and vertices are linked according to LMIs and MDAs. Based on the assumption that similar lncRNAs interact with similar diseases [10], we aim to identify novel LDAs by analyzing the behaviour of *neighbor lncRNAs*, in terms of their intermediate relationships with miRNAs. A score is assigned to each LDA (l, d) by considering both their respective interactions with common miRNAs, and the interactions with miRNAs shared by the considered disease d and other lncRNAs in the neighborhood of l.

Significant predictions for candidate LDAs to be proposed for further laboratory validation are computed by a statistical test performed through a Montecarlo test. The presented approach has been validated on real datasets.

2 Proposed Approach

The main goal of the research presented here is to provide a computational method able to predict novel LDAs candidate for experimental validation in laboratory, given further external information on both molecular interactions and genotype-phenotype associations, but without relying on the knowledge of existing validated LDAs.

The idea of not including any information on existing LDAs in the approach is based on the consideration that only a restricted number of validated LDAs is yet available, therefore a not exhaustive variability of real associations would be possible, affecting this way the correctness of the produced predictions. On the other hand, larger amounts of interactions between lncRNAs and other molecules (e.g., miRNAs, genes, proteins), as well as associations between those molecules and diseases are known, and we have focused our approach on the use of such datasets. In particular, we have considered only miRNAs as intermediate molecules, however the approach is general enough to allow the inclusion of also other molecules in the future.

Problem Statement Let $\mathcal{L} = \{l_1, l_2, \ldots, l_h\}$ be a set of lncRNAs and $\mathcal{D} = \{d_1, d_2, \ldots, d_k\}$ be a set of diseases. The goal is to return a set $\mathcal{P} = \{(l_x, d_y)\}$ of predicted LDAs.

Let T_{LMD} be a tripartite graph defined on the three sets of disjoint vertexes L, M and D, which can also be represented as $T_{LMD} = \langle (l, m), (m, d) \rangle$, where (l, m) are edges between vertexes in L and M, (m, d) are edges between vertexes in M and D, respectively. In the proposed approach, L is a set of lncRNAs, M is a set of miRNAs and D is a set of diseases. In such a context, edges of the type (l, m) represent molecular interactions between lncRNAs and miRNAs, experimentally validated in laboratory; edges of the type (m, d) correspond to known associations between miRNAs and diseases, according to the existing literature. In both cases, we refer to interactions and associations suitably annotated and stored in public databases.

A commonly recognized assumption is that lncRNAs with similar behaviour in terms of their molecular interactions with other molecules, may also reflect this similarity in their involvement in the occurrence and progress of disorders and diseases [10]. This is even more effective if the correlation with diseases is "mediated" exactly by the molecules they interact with, i.e., miRNAs.

2.1 Scoring of candidate LDAs

The model of tripartite graph allows to take into account that lncRNAs interacting with common miRNAs, may be involved in common diseases. To this aim, consider two matrixes M_{LL} and M_{LD}. In particular, M_{LL}:

- has h rows and h columns,
- both rows and columns are associated to the lncRNAs in $\mathcal{L}$,
- each element $M_{LL}[i, j]$ with $i \neq j$ contains the number of miRNAs in M linked to both l_i and l_j in T_{LMD};
- each element $M_{LL}[i, i]$ contains the number of edges incident onto l_i.

As for M_{LD}, it:

- has h rows and k columns,
- rows are associated to lncRNAs in $\mathcal{L}$, while columns to diseases in $\mathcal{D}$,
- each element $M_{LD}[i, j]$ contains the number of miRNAs in M linked to both l_i and d_j in T_{LMD}.

We define the *prediction-score* $S(l_i, d_j)$ for the LDA (l_i, d_j) such that $l_i \in \mathcal{L}$ and $d_j \in \mathcal{D}$ as:

$$S(l_i, d_j) = \alpha \left(\frac{M_{LD}[i, j]}{n} \right) + (1 - \alpha) \left(\frac{\sum_x M_{LL}[i, x] \cdot M_{LD}[x, j]}{\sum_x M_{LL}[x, x] \cdot n_j} \right)$$

where $n = \min(M_{LL}[i, i], n_j)$, n_j is the number of miRNAs linked to d_j in T_{LMD}, x are all the possible lncRNA neighbors of l_i and α is a real value in $[0, 1]$ used to balance the two terms of the formula. In particular, the prediction-score measures how much "connected" are l_i and d_j on T_{LMD}, with respect to both the amount of miRNAs they share and the amount of miRNAs that lncRNAs neighbors of l_i share with d_j.

2.2 Prediction of significant LDAs

Given a set $\mathcal{A}$ of LDAs scored according to the prediction-score computed as described above, it is necessary to select the only associations which are statistically significant, for producing the output predictions. To establish the statistical significance of the considered LDAs, we perform a Hypothesis Test via a Montecarlo simulation [4, 5]. The Null Hypothesis is that lncRNAs and diseases have been associated by chance. It is important to focus on the importance that the intermediate miRNAs have in the prediction-score computation and, more in general, in the measure of how much similar is the behaviour of different lncRNAs with respect to the occurrence of diseases. In particular, in the adopted model interactions with miRNAs are the key factors in order to determine the association between a lncRNA and a disease. Let then $(\hat{l}, \hat{m})$ be the pairs in $\mathcal{A}$ and shuffle them for 100 times by producing 100 new sets of pairs $\mathcal{A}_i$. The meaning is to interchange the associations between lncRNAs and miRNAs, still maintaining the same number of interactions. The test to reject the Null Hypotesis consists on comparing the prediction-score $S(l, d)$ of an association (l, d) in $\mathcal{A}$ with the maximum value of prediction-score $\hat{S}(l, d)$ obtained by the same pair in the 100 $\mathcal{A}_i$. The Null Hypotesis is rejected if $S(l, d) > \hat{S}(l, d)$.

3 Results

We have validated the proposed approach on experimental verified data downloaded from starBase [7] for the LMIs and from HMDD [8] for the MDAs, resulting in 114 lncRNAs, 762 miRNAs, 392 diseases and 275 LMIs, 2,201 MDAs. A golden-standard dataset with 183 LDAs has been obtained from the LncRNADisease database [2]. Before proceeding with our discussion, some considerations are needed. Although a number of approaches for LDAs prediction have been presented recently, including machine-learning-based models, only a few of them do not use directly known lncRNA-diseases relationships during the prediction task. However, so far, the experimentally identified known lncRNA-disease associations are still very limited, therefore using them during prediction could bias the final result. Indeed, when such approaches are applied for de novo LDAs prediction, their performance drastically go down [10]. This enforces the idea behind our approach, since neighborhood analysis automatically guides towards the detection of similar behaviours and without the need of positive examples for the training step. With respect to the other approaches which do not use LDAs during prediction (e.g., the p-value based approach in [3]), experimental tests have shown that our approach is able to detect specific situations which are not captured by its competitors. In particular, approaches such as [3] often fails in detecting true LDAs where the lncRNA and the diseases do not have a large number of shared miRNAs. Instead our approach is particularly effective in detecting this kind of situation, since neighborhood analysis allows to detect for example that there are similar lncRNAs associated to that disease.

The proposed approach has been applied to the known experimentally verified lncRNA-disease associations in the lncRNADisease database according to

LOOCV. In particular, each known disease-lncRNA association is left out in turn as test sample. How well this test sample was ranked relative to the candidate samples (all the disease-lncRNA pairs without the evidence to confirm their association) is evaluated. When the rank of this test sample exceeds the given threshold, this model is considered in order to provide a successful prediction. When the thresholds are varied, true positive rate (TPR, sensitivity) and false positive rate (FPR, specificity) could be obtained. Here, sensitivity refers to the percentage of the test samples whose ranking is higher than the given threshold. Specificity refers to the percentage of samples that are below the threshold. Receiver-operating characteristics (ROC) curve can be drawn by plotting TPR versus FPR at different thresholds. Area under ROC curve (AUC) is further calculated to evaluate the performance of the tested methods. AUC=1 indicates perfect performance and AUC=0.5 indicates random performance.

We have implemented the p-value based on HyperGeometric distribution for LDAs inference proposed in [3] and compared our approach against it. As a result, the proposed Neighborhoods based approach achieved an AUC equal to 0.67, whereas the p-value based approach scored AUC = 0.53, showing that the consideration of indirect relationships between lncRNAs and diseases through neighborhood analysis is more effective.

As for data extracted from StarBase and HMDD, our approach has produced 7,941 statistically significant LDAs predictions. Among them, it has been able to detect 66 of the 74 verified LDAs of the golden-standard dataset that could have been detected in this larger dataset (due to the presence of lncRNAs and diseases in the golden-standard), 24 out of which not detected by the p-value based approach.

4 Concluding Remarks

We have proposed an approach for LDAs prediction based on neighborhood analysis through a tripartite graph built upon lncRNA-miRNA interactions and miRNA-disease associations. One important fact is that the presented approach predicts potential LDAs without relying on the information of known disease-lncRNA associations. Although many previous study for LDAs prediction use known available LDAs, the latter are still comparatively rare relative to the known lncRNA-miRNA interactions and miRNA-disease associations. Moreover, in the presented research we show that neighborhood analysis performs better than other techniques previously presented in the literature and not based on known LDAs, such as p-value based on HyperGeometric distribution. This is promising and results presented here are to be intended as a first step towards a more complex pipeline, where different types of molecular interactions and associations other than only lncRNA-miRNA will be taken into account (e.g., gene-lncRNA co-expression relationship, lncRNA-protein interactions, etc.). Approaches based on integrative networks have indeed shown to reach better performance, therefore we plan to combine this strategy with the one proposed here on neighborhood analysis. Moreover, taking inspiration from previous studies on

social media [6], we plan also to design suitable co-clustering [11, 12] and network clustering [13] based methods in order to improve tripartite graph analysis.

5 Acknowledgements

Part of the research presented here has been funded by the MIUR-PRIN research project "Multicriteria Data Structures and Algorithms: from compressed to learned indexes, and beyond", grant n. 2017WR7SHH, and by the INdAM - GNCS Project 2020 "Algorithms, Methods and Software Tools for Knowledge Discovery in the Context of Precision Medicine".

References

1. S. Alaimo, R. Giugno, and A. Pulvirenti. ncpred: ncRNA-disease association prediction through tripartite network-based inference. *Front Bioeng Biot*, 2:71, 2014.
2. G. Chen et al. Lncrnadisease: a database for long-non-coding rna-associated diseases. *Nucleic Acids Res*, 41:D983–D986, 2013.
3. X. Chen. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Scientific Reports*, 5:13186, 2015.
4. R. Giancarlo, S. E. Rombo, and F. Utro. Epigenomic k-mer dictionaries: shedding light on how sequence composition influences *in vivo* nucleosome positioning. *Bioinformatics*, 31(18):2939–2946, 2015.
5. A. Gordon. Null models in cluster validation. *Gaul W. Pfeifer D. (eds.), From Data to Knowledge, Studies in Classification, Data Analysis, and Knowledge Organization. Springer Berlin Heidelberg*, pages 32–44, 1996.
6. K. Ikematsu and T. Murata. A fast method for detecting communities from tripartite networks. *In Proc. of SocInfo*, pages 192–205, 2013.
7. J.-H. Li et al. starbase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-seq data. *Nucleic Acids Research*, 42:D92–D97, 2013.
8. Y. Li et al. Hmdd v2.0: a database for experimentally supported human microrna and disease associations. *Nucleic Acids Res*, 42:D1070–D1074, 2014.
9. Q. Liao et al. Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co- expression network. *Nuc Ac Res*, 39:3864–3878, 2011.
10. C. Lu et al. Prediction of lncRNA-disease associations based on inductive matrix completion. *Bioinformatics*, 34(19):3357–3364, 2018.
11. C. Pizzuti and S. E. Rombo. *PINCoC* : A co-clustering based approach to analyze protein-protein interaction networks. In *Intelligent Data Engineering and Automated Learning - IDEAL 2007, 8th International Conference, Birmingham, UK, December 16-19, 2007, Proceedings*, pages 821–830, 2007.
12. C. Pizzuti and S. E. Rombo. A coclustering approach for mining large protein-protein interaction networks. *IEEE/ACM Transanctions on Computational Biology and Bioinformatics*, 9(3):717–730, 2012.
13. C. Pizzuti and S. E. Rombo. Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, 30(10):1343–1352, 2014.
14. Z. Xuan et al. A probabilistic matrix factorization method for identifying lncRNA-disease associations. *Genes*, 10(2), 2019.