Network Reconstruction for the Identification of miRNA:mRNA Interaction Networks

Gianvito Pio[1], Michelangelo Ceci[1], Domenica D'Elia[2], and Donato Malerba[1]

[1] University of Bari "A. Moro" - Via Orabona, 4 - 70125 Bari, Italy
[2] ITB-CNR, Via Amendola 122/D, 70126, Bari, Italy
{name.surname}@uniba.it, domenica.delia@ba.itb.cnr.it

1 Introduction

Network reconstruction from data is a data mining task which is receiving a significant attention due to its applicability in several domains. For example, it can be applied in social network analysis, where the goal is to identify connections among users and, thus, sub-communities. Another example can be found in computational biology, where the goal is to identify previously unknown relationships among biological entities and, thus, relevant interaction networks. Such task is usually solved by adopting methods for link prediction and for the identification of relevant sub-networks. Focusing on the biological domain, in [4] and [3] we proposed two methods for learning to combine the output of several link prediction algorithms and for the identification of biological significant interaction networks involving two important types of RNA molecules, i.e. microRNAs (miRNAs) and messenger RNAs (mRNAs). The relevance of this application comes from the importance of identifying (previously unknown) regulatory and cooperation activities for the understanding of the biological roles of miRNAs and mRNAs. In this paper, we review the contribution given by the combination of the proposed methods for network reconstruction and the solutions we adopt in order to meet specific challenges coming from the specific domain we consider.

2 Learning to Combine Link Predictions

In the literature, several approaches for link prediction can be found, but they often fail in simultaneously considering all the possible criteria (e.g. network topology, nodes properties, autocorrelation among nodes). In [4] we presented a method for *learning to combine* the scores returned by several link prediction algorithms (which are based on one or few of the possible criteria) for the identification of interactions between miRNAs and mRNAs. In such case, some issues have to be taken into account: *i)* very few interactions are experimentally validated and can be considered as "stable" examples; *ii)* only positive examples are generally available; *iii)* prediction algorithms consider similar features and their combination can lead to collinearity problems.

In order to face *i)* and *ii)*, we propose a semi-supervised learning algorithm, which considers both positively labeled examples of interactions and the huge

T. Calders et al. (Eds.): ECML PKDD 2014, Part III, LNCS 8726, pp. 508–511, 2014.
© Springer-Verlag Berlin Heidelberg 2014

set of unlabeled (unknown) instances. As for *iii)*, the collinearity problem is alleviated by considering as features the scores (outputs) obtained by the prediction algorithms (instead of original features), resorting to a solution which is similar to meta-learning algorithms. The advantage of applying a machine learning method to the outputs of prediction algorithms consists in automatically adapting to unknown patterns of the outputs and performing more reliable predictions when these patterns occur. The proposed method consists in three main steps:

1. Each example of interaction is represented by a vector of scores, obtained by prediction algorithms, and is associated with a label representing the fact that it is labeled as positive (i.e. experimentally validated) or unlabeled.
2. A probabilistic classifier is learned to compute the likelihood that an example of interaction is labeled (known) / unlabeled.
3. A new probabilistic classifier which also exploits (*à la Bayes*) the likelihood computed in the step *2)* is learned. Such classifier associates a score to each interaction to decide whether this interaction is true.

In step *3)*, scores are computed by exploiting the assumption that all the labeled examples are taken randomly from all the positive examples. In other words, the probability of an existing interaction to belong to the set of labeled examples is independent of the specific interaction. Formal definitions can be found in [4].

It is noteworthy that steps *2)* and *3)* require to learn a classifier from a highly unbalanced dataset. Indeed, the set of labeled (in the first case) and positive (in the second case) examples is significantly smaller than the set of unknown examples and negative examples, respectively. Thus, we adopt an ensemble-based approach. In particular, K classifiers are learned by considering as training set the whole set of positive examples and a subset of negative examples, built through a random sampling with replacement. The score associated to each example is computed by averaging the output of all the classifiers that considered it during the learning phase. Further (formal) details can be found in [4].

3 Identification of Relevant Interaction Networks

In [3] we proposed the biclustering algorithm HOCCLUS2 for the identification of miRNA:mRNA interaction networks from the identified interaction scores. Although, in the literature, the application of biclustering techniques to biological data has already been proposed [1,2], some specific aspects are not considered. In particular, identified networks should be: *a)* possibly overlapping, since mRNAs and miRNAs can be involved in multiple interaction networks; *b)* hierarchically organized, allowing biologists to better interpret results and to distinguish between miRNAs involved only in specific pathways or in many biological processes; *c)* highly cohesive, i.e. miRNAs and mRNAs in the same network should be highly related and show only reliable interactions. HOCCLUS2 takes into account these aspects and allows the user to identify the most promising biclusters through a ranking based on a statistical test comparing intra- and inter-bicluster similarity in the Gene Ontology. In the following we describe the first two steps, whereas details about the ranking step can be found in [3].

The **first step** requires a threshold value β, i.e. the minimum score for a miRNA:mRNA interaction to be considered as reliable. The algorithm builds biclusters in the form of bicliques, by considering: *avg_mirna* - the average number of miRNAs which target each mRNA, with a score greater than β; *abs_min_mrna* and *min_mrna* - the *absolute* and the *outlier-proof* (respectively) minimum number of mRNAs which are targeted by each miRNA, with a score greater than β. *min_mrna* (*outlier-proof*) is computed by discarding the lowest 0.15% values (possibly outliers, according to the 3σ rule), by assuming a Normal distribution. The algorithm builds an initial set of bicliques, each consisting of a single miRNA and of the set of mRNAs associated with a score greater than β. The algorithm, then, iteratively aggregates two biclusters C' and C'' into a new bicluster C''' as follows: $C'''_r = C'_r \cap C''_r$; $\quad C'''_c = C'_c \cup C''_c$, where C_r and C_c are mRNAs and miRNAs in C, respectively. Necessary conditions for aggregating are: $C'_r \cup C''_r \geq min\_mrna$; $\quad C'_c \cap C''_c \leq avg\_mirna$. The basic idea is that a good biclique should contain approximately *avg_mirna* miRNAs, while keeping the highest possible number of mRNAs (at least *min_mrna*). Moreover, since we want to obtain a set of highly cohesive bicliques, among the possible aggregations of pairs of bicliques $\langle C', C'' \rangle$, we select the pair which maximizes $jaccard(C'_r, C''_r) * q(C''', A)$, where $jaccard(C'_r, C''_r) = \frac{|C'_r \cap C''_r|}{|C'_r \cup C''_r|}$, $q(C, A)$ is a cohesiveness measure defined as $q(C, A) = (|C_r| * |C_c|)^{-1} * \sum_{x \in C_r} \sum_{y \in C_c} A_{x,y}$ and A is the adjacency matrix containing the score associated to each interaction. The same iterative process is repeated starting from bicliques containing a single mRNA and the two sets of identified bicliques are merged into a single set.

The **second step** consists of an iterative process in which overlap identification and merging are performed. The assumption behind the overlap identification is that two non-overlapping biclusters should be separable in the space. Given two biclusters C' and C'' (belonging to the same hierarchical level), we identify two optimal separating hyperplanes between C' and C'' by learning an SVM model for each dimension (miRNAs and mRNAs). Objects in C' and C'' are used as both training and testing set. Misclassified objects are those which possibly belong to both the biclusters and are added to the bicluster which previously did not contain them. As regards the merging, we assume that miRNAs and mRNAs are normally distributed and consider the distance between pairs of biclusters. In particular, two biclusters C', C'' are candidates for merging if: $dist(C'_r, C''_r) - 2\sigma(C'_r) - 2\sigma(C''_r) \leq 0$ **or** $dist(C'_c, C''_c) - 2\sigma(C'_c) - 2\sigma(C''_c) \leq 0$, where $dist(w, z)$ is the Euclidean distance between the centroids of the clusters w and z, and $\sigma(w)$ is the standard deviation of the cluster w. Intuitively, two biclusters are candidates for merging if they are close according to at least one dimension. A pair of biclusters C', C'', candidate for merging, is merged if the quality constraint $q(C''', A) > \alpha$ is satisfied, where $C'''_r \leftarrow C'_r \cup C''_r$, $C'''_c \leftarrow C'_c \cup C''_c$ and α is a user-defined threshold. Low values of α facilitate merging, decreasing cohesiveness. Since a bicluster can be a candidate for multiple merging, we perform that resulting in the bicluster with maximum cohesiveness.

4 Discussion and Conclusions

The proposed methods have been applied for the identification of miRNA:mRNA interaction networks. In particular, our combination approach has been applied to validated data in miRTarBase 2.5 (4,270 interactions, available at: `mirtarbase.mbc.nctu.edu.tw`), and on the scores returned by 10 prediction algorithms in mirDIP ($>$ 5,000,000 interactions, available at:`ophid.utoronto.ca/mirDIP`). We evaluated the accuracy in terms of the AUC measure on an independent testing set, i.e. TarBase ($> 65,000$ examples, available at `www.microrna.gr/tarbase`), comparing the results with those obtained by single prediction algorithms and by baseline combination approaches based on score averaging. Moreover, we applied HOCCLUS2 with different values of its parameters to the set of predictions obtained by our combination approach and by baseline combination strategies to evaluate the significance of the extracted networks. In this case, the evaluation was performed in terms of cohesiveness and of a statistical test that takes into account the intra- and inter- bicluster similarity with respect to the classification in Gene Ontology. The evaluation in terms of AUC showed that the proposed approach is able to identify a set of more reliable predictions with respect to the considered competitive approaches. This is also confirmed by the higher significance of the interaction networks extracted by HOCCLUS2, both in terms of the considered evaluation measures and in terms of a biological evaluation performed by a domain expert. These results prove that the proposed approach is able to better filter out false positives and let HOCCLUS2 focus on more reliable predictions so to obtain more significant interaction networks. Details about quantitative and biological analysis can be found in [3,4]. Download links: semi-supervised system: `www.di.uniba.it/~ceci/micFiles/systems/semisupervised_ HOCCLUS2/`; HOCCLUS2: `www.di.uniba.it/~ceci/micFiles/systems/HOCCLUS/`;biological query system: `comirnet.di.uniba.it`.

Acknowledgements. We would like to acknowledge the support of the European Commission through the project MAESTRA (Grant number ICT-2013-612944).

References

1. Cheng, Y., Church, G.M.: Biclustering of Expression Data. In: Proc. of ISMB 2000, pp. 93–103 (2000)
2. Deodhar, M., Gupta, G., Ghosh, J., Cho, H., Dhillon, I.S.: A scalable framework for discovering coherent co-clusters in noisy data. In: Proc. of ICML 2009, p. 31 (2009)
3. 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)
4. 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(S-1), S4 (2014)