Robust Inductive Matrix Completion Strategy to Explore Associations between LincRNAs and Human Disease Phenotypes

Ashis Kumer Biswas[§], Dong-Chul Kim[†], Mingon Kang[‡], Jean X. Gao[§*]
[§]Dept. of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019,
[†]Dept. of Computer Science, University of Texas Rio Grande Valley, Edinburg, TX 78541,
[‡]Dept. of Computer Science, Kennesaw State University, Marietta, GA 30060.
[*]email: gao@uta.edu

Abstract—**Long intergenic non-coding RNAs (lincRNAs) are associated with a wide variety of human diseases. Piles of data about the lincRNAs are becoming available, thanks to the High Throughput Sequencing (HTS) platforms, which open opportunity for cutting-edge machine learning and data mining approaches to analyze the disease association better. However, there are only a few *in silico* association inference tools available to date, and none of them utilizes the heterogeneous data about the lincRNAs and diseases. The standard Inductive Matrix Completion (IMC) technique provides with a platform among the two entities considering respective side information. But, it has two major issues pertaining to the noise and sparsity in the dataset. Thus, a robust version of IMC is needed to adequately address the issues. In this paper, we propose Robust Inductive Matrix Completion (RIMC) to address these challenges. Then, we applied RIMC to the available association dataset between the lincRNAs and OMIM disease phenotypes with a diverse set of side information of the both. The proposed method performs better than the state-of-the-art methods in terms of $precision@k$ and $recall@k$ at the top-k disease prioritization to the subject lincRNAs. Moreover, with an induction experiment we showed that RIMC performs superior than the standard IMC for ranking unexplored disease phenotypes to a set of known lincRNAs.**

I. Introduction

Only 2% of the entire human genome codes for proteins [1]. In recent years, out of these huge non-protein coding portion of the genome, the long intergenic non-coding RNAs (lincRNAs) have emerged with critical functional importance for their diverse molecular mechanisms and implications to various human diseases [2]. The advancement of the next generation sequencing technologies, such as RNA-seq and ChIP-seq, a huge number of lincRNAs have been cataloged. But, characterizing their functions as well as predicting the associations of the lincRNAs to human diseases, remain a challenge [3]. *In silico* association inference tools would present effective framework towards discovering causal lincRNA-disease associations and better understanding of the human diseases.

There exists only a few methods that have dealt with the lincRNA-disease inference problem. Due to the intricacies inherent to the dataset, only a very small number of experimentally validated associations have been reported in the publicly available repositories, such as lncRNAdisease [4]. For this reason, leveraging multiple complementary data sources is essential for predicting lincRNAs related to diseases as well as respective phenotypes, and thus different inference methods have been developed considering different knowledge sources. For instance, K-RWRH [5], LRLSLDA[6] and TslncRNA-disease [7] are popular family of network based methods. The methods utilize biological networks, such as lincRNA similarity network and disease similarity network as the key ingredient for developing the inference engine. The inference problem can also be solved using low rank matrix completion method. But it suffers from the cold start problem, due to the inability to address the inference predictions of the diseases for novel lincRNAs and vice versa. However, the methods exploiting lincRNA-expression profiles to build similarity networks may only deal with specific disease classes that are only available through the seed or true associations and therefore the methods fall short in generalizing to novel diseases. In this regard, the standard Inductive Matrix Completion (IMC) takes into account the plethora of side information of both entities available to date along with the known association evidences to predict missing associations [8]. But, the standard IMC uses the least square error function that is well known to be unstable with respect to noises and outliers present in the dataset [9]. However, the side information about the lincRNAs and the diseases possibly contain noise and outliers. To deal with such situation, a robust IMC is needed. Therefore, we propose a novel robust formulation of IMC (RIMC) using $\ell_{2,1}$ norm penalty function, as well as $\ell_{2,1}$ based regularization. The proposed method is called "robust" as it can handle outliers and noises better than the standard IMC. Also, it can handle joint sparsity, i.e., handle appropriately the feature set, where each feature either has small values for all data points or has large values over all data points.

The rest of the paper is organized as follows. In section II we propose the robust IMC formulation using $\ell_{2,1}$ norm, underline the advantages of the proposed algorithm compared with the standard IMC. In section III we present the configurations for the experiments we conducted in this study. That includes description of the association dataset, side information dataset, feature extraction, summary of baseline algorithms as well as the performance metrics used to evaluate the models. In

978-1-5090-1610-5/16/$31.00 ©2016 IEEE

Section IV, we present the results of the association inference experiments on the dataset, and underline the superior performance of the proposed algorithm than the existing methods. Finally, in section V we conclude the paper by pointing out several future research scopes.

II. Robust Inductive Matrix Completion (RIMC)

Although the standard Inductive Matrix Completion method developed by [8] enables us to incorporate side information of both the row and column entities and overcomes the limitation imposed by the transductive matrix completion approaches (e.g., standard NMF, etc.), it is prone to outliers and group sparsity present in the given dataset due to the usage of the squared loss function in its formulation. Let us given a matrix $A \in \mathbb{R}^{M\times N}$ encapsulating the association between M row entities and N column entities. Besides A, side information of both the entities are given in two matrices: $X \in \mathbb{R}^{M\times m}$ containing m features of the M row entities and $Y \in \mathbb{R}^{N\times n}$ containing n features of the N column entities respectively. The standard IMC is defined as,

$$\min_{W,H} \quad \varphi = \frac{1}{2}||A - XWH^TY^T||_F^2 + \frac{\lambda_1}{2}||W||_F^2 + \frac{\lambda_2}{2}||H||_F^2$$
$$\text{such that,} \quad W \geq 0, H \geq 0, \tag{1}$$

where λ_1, λ_2 are the regularization parameters that trade off between the accrued loss on the observed entries and the trace norm regularization constraints. Here, goal is to recover a low-rank matrix $Z \in \mathbb{R}^{m\times n}$ using the observed entries of A, and the X and Y feature matrices. The entry A_{ij} is modeled as $\mathbf{x}_i^T Z \mathbf{y}_j$. By forming Z as WH^T, where $W \in \mathbb{R}^{m\times r}$ and $H \in \mathbb{R}^{n\times r}$. The problem can be solved using Algorithm 1.

Algorithm 1 COMPUTE_STANDARD_IMC(A,X,Y,r)

```
Input: Association matrix $A \in \mathbb{R}^{M\times N}$; feature matrix for
    the $M$ row entities $X \in \mathbb{R}^{M\times m}$; feature matrix for the $N$
    column entities $Y \in \mathbb{R}^{N\times n}$; desired rank $r$.
Output: The two factor matrices $W \in \mathbb{R}^{m\times r}$ and $H \in \mathbb{R}^{n\times r}$.

Initialize $W$ and $H$ as random dense matrix maintaining
    the non-negativity constraints $W_{ik} \geq 0, H_{jk} \geq 0$.
repeat
    Update $H$ matrix using the following equation:
        $H_{jk} \leftarrow H_{jk} \frac{(Y^TA^TXW)_{jk}}{(Y^TYHW^TX^TXW + \lambda_2 H)_{jk}}$    (2)
    Update $W$ matrix using the following equation. Here
    we will be using the $H$ calculated at the previous step.
        $W_{ik} \leftarrow W_{ik} \frac{(X^TAYH)_{ik}}{(X^TXWH^TY^TYH + \lambda_1 W)_{ik}}$    (3)
until convergence criterion is met
return $W, H$
```

In standard IMC, the error for each row entity of the objective function, the squared residue error is accumulated in the form of $||A_{i,:} - X_{i,:}WH^TY^T||_2^2$. Hence, a few outliers with large error could dominate the overall computation. The second limitation of the standard IMC is that the ℓ_2 norm based regularization (i.e., ridge regularization) does not handle joint sparsity across the feature data matrices. By joint sparsity we refer to the set of features having either small scores across all data points, or large scores across all data points. Thus it is very important to present a robust IMC formulation. That is why we introduce $\ell_{2,1}$ norm instead of ℓ_2 norm to define the loss function in our proposed RIMC formulation, which is:

$$||A - XWH^TY^T||_{2,1} = \sum_{i=1}^{M} \sqrt{\sum_{j=1}^{N} (A - XWH^TY^T)_{ij}^2} \tag{4}$$

Here, the error for each data point is not squared, and thus the large errors due to outliers do not dominate the objective function as they would in the standard IMC formulation. Morever, the new loss function is convex and can easily be optimized according to [10]. We now propose robust IMC formulated as:

$$\min_{W,H} \quad \varphi = \frac{1}{2}||A - XWH^TY^T||_{2,1} + \lambda_1||W||_{2,1} + \lambda_2||H||_{2,1}$$
$$\text{such that,} \quad W \geq 0, H \geq 0 \tag{5}$$

Algorithm 2 COMPUTE_ROBUST_IMC(A,X,Y,r)

```
Input: $A, X, Y, r$
Output: $W, H$
Initialize $W, H$ with uniform random numbers in (0,1).
Initialize $D = I_{M\times M}, P = I_{m\times m}, Q = I_{n\times n}$.
repeat
    Update $H$ as follows:
        $H_{jk} = H_{jk} \frac{(e_n e_M^T A e_N e_M^T D e_M e_n^T Y^T e_N e_M^T XW)_{ik}}{\left(e_n e_M^T XWH^TY^T e_N e_M^T D e_M e_n^T Y^T e_N e_M^T XW + \lambda_2 QH\right)_{jk}}$,
     where $e_s = (1, \cdots, 1)^T \in \mathbb{R}^s$ is a vector with all 1s.
    Update $D$ as follows:
        $D_{ii} = 1 \Big/ \sqrt{\sum_{j=1}^{N} (A - XWH^TY^T)_{ij}^2}$
    Update $Q$ as follows:
        $Q_{ii} = 1 \Big/ \sqrt{\sum_{j=1}^{r} H_{ij}^2}$
    Update $W$ matrix as follows. Here we will be using the
    $H$ calculated at the previous step.
        $W_{ik} = W_{ik} \frac{(e_m e_M^T A e_N e_M^T D e_M e_m^T X^T e_M e_N^T YH)_{ik}}{\left(e_m e_M^T XWH^TY^T e_N e_M^T D e_M e_m^T X^T e_M e_N^T YH + \lambda_1 PW\right)_{ik}}$
    Update $P$ as follows:
        $P_{ii} = 1 \Big/ \sqrt{\sum_{j=1}^{r} W_{ij}^2}$
until convergence criterion is met
return $W, H$
```

A. Algorithm for RIMC

The main contribution of this manuscript is to derive Algorithm 2 that solves the robust IMC optimization problem defined in Equation 5.

Once the W and H matrices are obtained, besides computing the associative scores among the row and column entities from the training set, it can also perform induction on a new row entity i' that was not part of the training data, the prediction $A_{i'j}$ can be computed for a column j as long as we have feature vector $\mathbf{x}_{i'}$, using the model as: $A_{i'j} = \mathbf{x}_{i'}WH^T\mathbf{y}_j$. Similarly, prediction can also be made for a new column entity j' with a row entity in the set using new feature vector $\mathbf{y}_{j'}$ by $A_{ij'} = \mathbf{x}_iWH^T\mathbf{y}_{j'}$. However, the prediction between a new column entity (j') and a new row entity (i') can also be computed through using their corresponding feature vectors, $\mathbf{x}_{i'}$ and $\mathbf{y}_{j'}$ through: $A_{i'j'} = \mathbf{x}_{i'}WH^T\mathbf{y}_{j'}$.

III. Experimental Configurations

A. LincRNA-Disease Association dataset

We obtained human lincRNA-disease associations by combining the LncRNADisease database [4] and the supporting dataset from the co-expression based association study conducted by [7]. The combined dataset contains 46,934 associations among 8194 lincRNA genes and 1213 diseases. Since none of the two datasets adapted standard naming of the diseases, we retrieved top-5 closely matched OMIM phenotypes for each of the disease names from the pool using OMIM API [11], and prepared the association matrix between 8194 lincRNA genes and 2661 OMIM phenotypes. The matrix is very sparse having only 0.22% non-zero entries. To compare different approaches on the novel association prediction, we use 10-fold cross validation over the association dataset.

B. LincRNA Feature datasets

RNA-seq provided Expression profiles of lincRNAs on different tissues underline the impact of the lincRNAs for diseases occurring corresponding tissues. Although not all diseases are tissue-specific, neither are the lincRNAs, the profiles still can be used to distinguish between co-expressed lincRNAs to implicate diseases. RNA-seq measurement of 8194 lincRNA expression levels on 22 human tissues are obtained from the Human BodyMap Project 2.0 [3]. Expression scores are represented in terms of FPKM values (Fragments Per Kilobase of exons per Million Fragments mapped).

ChIP-seq provided Transcription Factor Binding Sites (TFBS) of the lincRNAs unravel the transcriptional regulatory relationships of lincRNAs with transcription factors. We obtained 160,588 relationships among the 8194 lincRNAs and 120 transcription factors from ChIP-Base dataset [12]. There are only 217 lincRNAs that have relationship with one transcription factor. The minimally related transcript factor, “BACH1” has only 11 lincRNA associations, and there are 6130 lincRNAs connecting with the transcript factor called “HNF4A”.

Functional annotations of the lincRNAs dictate their characterizations and involvement on various biological activities inside human cell that implicitly correlate with various disease phenotypes. Linc2GO [13] presents a database of such annotations of lincRNAs based on the ceRNA hypothesis [14]. We retrieved 8111 GO BP (Biological Process) terms, 3218 GO MF (Molecular Function) terms and 193 KEGG pathway terms associated with the 8194 lincRNAs from the database, resulting a total of 11522 functional terms for each lincRNA in our study. However, this annotation matrix is also sparse, having 0.11% non-zero entries. We use the leading 100 singular vectors of the matrix as the representative features of the lincRNAs contributed from the Linc2GO dataset.

Single Nucleotide Polymorphisms (SNPs) in lincRNAs were found to be linked to their abnormal expressions and dysregulations, thereby playing key roles in various phenotypes and diseases [15]. The lncRNASNP dataset [16] provides a comprehensive resource of SNPs in human lncRNAs, and we extracted 368,494 SNPs in the 8194 lincRNAs from the database. The SNP-lincRNA association is sparse with 0.0077% non-zero entries. We use the leading 100 singular vectors of the matrix as the representative features of the lincRNAs contributed from the lncRNASNP dataset.

Finally, we considered only those lincRNAs having all these four types of features. Therefore, we ended up having a catalog of 6540 lincRNAs with the 342 features.

C. Disease Feature datasets

Term Frequency Inverse Document Frequency (TF-IDF) of the 2661 OMIM phenotypes obtained from the OMIM text corpus provides a standard statistic that reflects how important a term is to a OMIM phenotype text collection. The TF-IDF score increases proportionally to the frequency of occurrences of a term in a particular page, but is offset by the frequency of the term in the whole corpus. This phenomenon helps to identify important keywords associated appearing only in the corresponding OMIM page, as well as less important terms appearing most of the pages. The number of terms considered in the scheme is 20491, thus resulting in a TF-IDF matrix of size 2661 by 20491. We use the leading 100 singular vectors of the matrix as the representative features of the diseases contributed from the OMIM TF-IDF dataset.

Phenotypic similarity profiles of the lincRNAs were retrieved from a recent study by [17], where the authors developed a method to accumulate the MeSH terms associated with the publications referenced in the OMIM phenotype pages and able to compute scores that reflect the molecular relatedness between two OMIM entries. The similarity matrix thereby is symmetric of dimension 2661 by 2661. We reduce the dimensionality of the feature space using PCA, retaining the top 100 principal components.

We considered only those diseases having all these two types of features. Thus, we ended up having a catalog of 2148 diseases with the 200 features.

D. Baselines

We compare the results of our proposed method with five approaches. Firstly, the standard non-negative matrix factorization (NMF) [18] on the lincRNA-disease association matrix A. This can be considered as a special case for the Inductive Matrix Completion objective where the lincRNA feature matrix (X) and the disease feature matrix (Y) are set to identity. We compare three other methods that provide interfaces to scale their corresponding framework to a much larger dataset like ours (i.e., considering associations among the 6540 lincRNAs and 2418 disease phenotypes).

The Laplacian Regularized Least Squares for LncRNA-Disease Association (LRLSLDA) computes a weighted rank score for association between an lincRNA with a disease using probabilities retrieved from two indepedent classifiers modeled using lincRNA-lincRNA and disease-disease similarity matrices [6]. The computationally expensive operation in LRLSLDA is during the pairwise similarity matrix constructions which prohibits its usability in scalable framework development. Moreover, there are eight parameters used in LRLSLDA, which comparatively is a large number to tune in order to make the method computationally efficient. Tissue-Specificity based LincRNA-Disease association prediction framework (TsLincRNA-Disease) draws a demarcation line between tissue-specific and non-tissue-specific classes utilizing the tissue-specificity index for each of the lincRNAs in the study [7]. It uses statistical significance test and a mean enrichment analysis on a co-expression network to predict disease associations with tissue specific and non-specific lincRNAs respectively. Kernel-based Random Walk with Restart method in a heterogeneous network is an extension to the RWRH algorithm[5]. Here the heterogeneous network is constructed by a disease-disease similarity matrix, lincRNA-lincRNA similarity matrix and known lincRNA-disease relationship matrix. It predicts potential lincRNA-disease association through simulating the random walk with restart from a given set of known disease and lincRNA seed nodes. After some steps, the steady state probability distribution is obtained. The lincRNAs and the diseases (representing the nodes in the network) are ranked based on the steady state probabilities.

E. Experimental Setup

Note that the NMF based method does not use any of lincRNA or disease specific features such as TF-IDF, disease phenotype similarity, RNA-seq provided expression profiles, ChIP-seq provided Transcription factor binding sites, functional GO annotations and SNP linkages. However, remaining three methods also do not use any of these features except the expression profiles. For all the methods, including the standard IMC and our proposed RIMC we rank the predictions using the estimated values corresponding to a lincRNA for each of the diseases considered in our study. For the standard IMC method and our proposed RIMC method, we construct the lincRNA and disease feature matrices $X \in \mathbb{R}^{M\times m}$ with $m = 342$ and $Y \in \mathbb{R}^{N\times n}$ with $n = 200$ for the set of $M = 6540$ lincRNAs and $N = 2418$ diseases (in terms of OMIM phenotypes). We set the best parameters values for each of the methods through cross-validation except LRLSLDA, in which case we set the eight parameter values as suggested by the authors.

F. Evaluation Metrics

The lincRNA-disease association prediction algorithm under evaluation computes a ranking score for each candidate disease (i.e., disease that is not reported to be connected with a lincRNA before) and returns the top-k highest ranked diseases as recommendations to a target lincRNA. Thus, for the evaluation of the predictive accuracy, the goal is to find out how many disease-lincRNA associations previously marked off in the preprocessing step recovered in the returned disease recommendations. More specifically, we used two evaluation metrics: (1) the ratio of recovered diseases to the k recommended diseases for the target lincRNA, and (2) the ratio of recovered diseases to the set of diseases deleted in preprocessing [19]. The first metric is called $precision@k$ and the latter is known as $recall@k$. The metrics are defined in Equation 6 and 7. In our experiment, we tested the performance when $k = \{5, 10, 20, 40, 50, 60, 70, 80, 90, 100\}$.

$$precision@k = \frac{1}{N_l}\sum_{l=1}^{N_l}\frac{|P_l(k) \cap D_l|}{k} \tag{6}$$

$$recall@k = \frac{1}{N_l}\sum_{l=1}^{N_l}\frac{|P_l(k) \cap D_l|}{|D_l|}, \tag{7}$$

where $P_l(k)$ being the top-k ranked diseases for lincRNA l, D_l is the set of diseases related to the lincRNA l marked off during the training step, N_l is the total number lincRNAs in the evaluation dataset. We performed 10-fold cross-validation to measure the performance of our proposed RIMC method as well as the competitive methods. It is worth noting that the $precision@k$ and $recall@k$ in our experiments are not high. This is because of the sparsity in the lincRNA-disease association dataset having density only 0.003. Similar performance can also be observed in other association recommendation works by [20] and [21] just to name a few. Therefore, the low precision obtained in our experiments is reasonable. In this article, we emphasize on comparing relative performance of the methods rather than their absolute performance.

IV. Results and Discussion

A. Effect of the parameter settings

The parameters to the RIMC method are the rank (r) of the basis W and the coefficient H matrices and the regularization parameters λ_1, λ_2 for W and H matrices respectively. The $precision@k$ and $recall@k$ performance of the method along with the standard IMC formulation is presented in Figure 1. It is evident from the figure that the top-k association retrieval performance varies with the changes in the rank parameter. We varied the rank parameter from 50 to 200, which is equal to $\min(\text{rank}(X), \text{rank}(Y))$, where X, Y are the lincRNA and disease feature matrices respectively. We see that there is little to no boost up of the performance from the changes of rank value from 50 to 100. However, continuing to increase the rank to the maximum possible value, 200, the performance degrades in terms of both the $precision@k$ and $recall@k$. This issue can be justified as

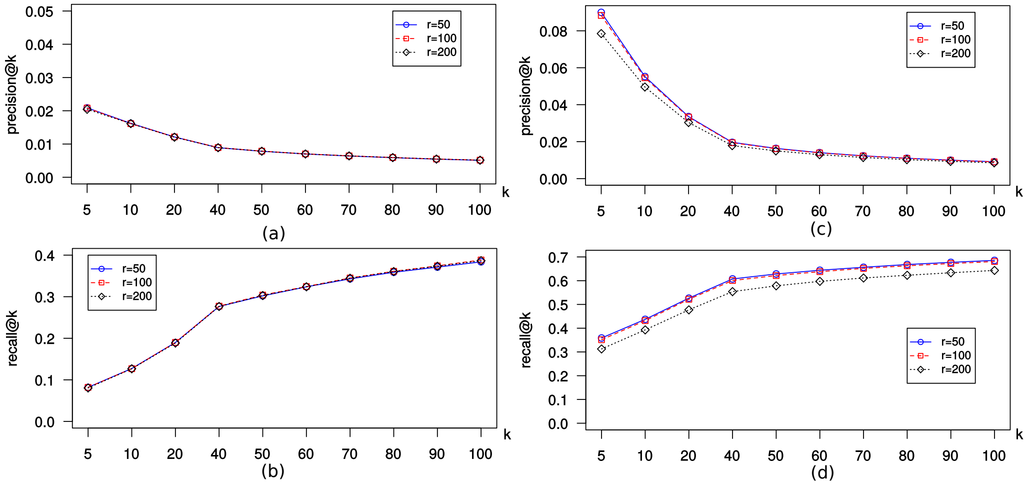

Fig. 1. Comparing $precision@k$ and $recall@k$ of the standard IMC and the robust IMC on different values of rank parameter r. (a-b): Standard IMC shows a slight improvement of precision@k and recall@k with increasing value for r. (c-d): Robust IMC shows the best performance when $r = 50$.

an over-fitting problem. We set the regularization parameters $\lambda_1 = \lambda_2 = 1.0$ in the optimization functions as we found that the predictive performance degrades if the parameters are set to values deviating far away from 1.0 (data not shown). We found the cut-off value for the parameters through cross-validation for all possible values in the range (0.1, 10.0).

B. True LincRNA-Disease Association Retrieval

The 10-fold cross-validation results on 2418 OMIM diseases are presented in Figure 2. The Y-axis in the plots (a,b) gives the $precision@k$ and $recall@k$ scores for various k values in the horizontal X-axis. We observe that the proposed RIMC significantly dominates the competitive methods over all k values. The best $precision@k$ and $recall@k$ recorded are close to 10% and 38% respectively at the top-5 association prediction cases. The matrix completion on A performs significantly better than the three other baseline algorithms. LRLSLDA performs worse in terms of $precision@k$ and the $recall@k$ scores. This is because, the method only relies on known association matrix and the expression profiles of the lincRNAs. Moreover, it comes with a lot of parameters to learn, and is not easily scalable in larger context, like ours, because of the complex `pinv` operations to compute the Laplacians.

C. Induction on new Associations

Next, we investigate the power of the inductive learning by the standard IMC and the proposed RIMC method. We randomly picked 10% of the subject lincRNA entries and the corresponding associations from our datasets (X, the lincRNA featureset and A, the lincRNA-disease association data matrix for case i and iii above) and the subject disease entries and the respective associations from Y, the disease featureset and A matrix for case ii and iii above as new test samples. Both the standard IMC and the RIMC were then trained with the remaining entries and associations. We evaluate each of the models with the respective set-aside test cases. We repeat the above steps 10 times and recorded the average predictive scores for the comparison. The only assumption in the IMC framework for induction is that all the features for the novel disease (or the lincRNA or both) may be available during prediction. Here we underline the power of inductive learning of the trained models which is readily usable for prediction for a new test lincRNA or disease (or both) entries even though the entries were absent during the training. Note that all the baseline methods other than the standard IMC are missing from each of the plots in Figure 3 as none could make such prediction of the novel disease and lincRNA associations using the respective learned models because of their inherent transductive formulations. Figure 3 illustrates the performance comparison of the standard IMC and our proposed robust IMC for the both new diseases and the new lincRNAs. The $precision@k$ and $recall@k$ curves for the robust IMC show a superior performance than that of the standard IMC based approach for predicting upto the top-50 lincRNA associations with the novel diseases.

V. Conclusions

In this manuscript, we proposed a robust formulation of the inductive matrix completion method using $\ell_{2,1}$ norm. We applied our proposed method for predicting associations between the long intergenic non-coding RNAs (lincRNAs) and diseases. The method presents an integration interface for various categories of features of both the lincRNAs and diseases obtained through different independent data sources for explaining the relationships between the two entities. The proposed method can handle inherent noises and outliers in the dataset, and was shown to outperform the ℓ_2 norm based standard IMC formulation. The method shows superior performance to predict associations between already studied set of lincRNAs and diseases as well as between novel set of lincRNAs and diseases which makes the method a suitable association prediction tool for the biologists. Two possible extensions to our method presented here can be made: (i) the

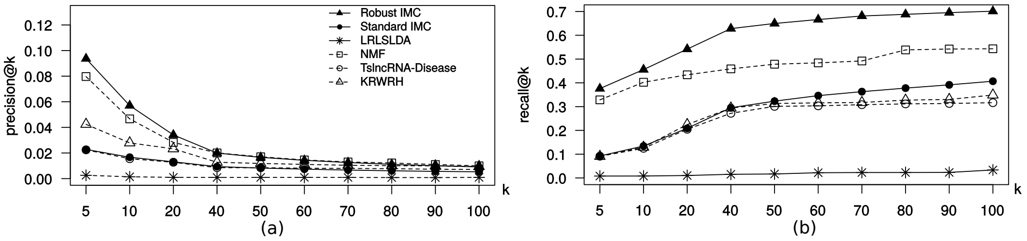

Fig. 2. Comparision of lincRNA-disease association methods. (a) k-vs-$precision@k$ plot for all the six methods. (b) k-vs-$recall@k$ plot for the six methods. The standard IMC and the proposed RIMC method is trained with 342 lincRNA features and 200 disease features, with a rank, $r = 100$. NMF was trained with the same binary association matrix we used in the IMC experiments with a rank $r = 100$.

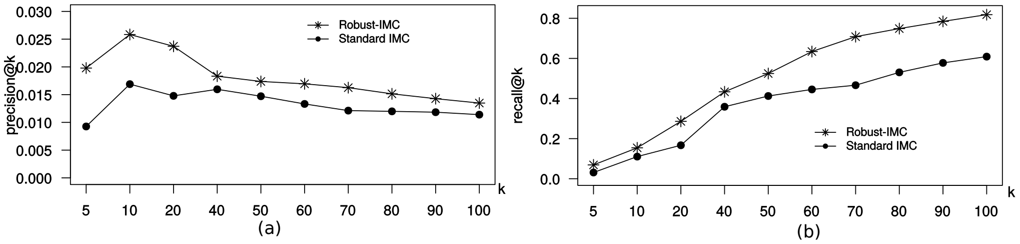

Fig. 3. Performance comparison of the standard IMC and our proposed robust IMC for induction on novel diseases and lincRNAs. (a) k-vs-$precision@k$ plot for the two methods, (b) k-vs-$recall@k$ plot for the two methods.

inductive framework (as opposed to its transductive versions) is not limited to the types of features used in the experiments we presented, as new sources of information can be integrated easily via rank-1 updates. (ii) The framework itself can be extended to address the missing value problem inherent to the side information of the two respective entities.

References

[1] R. P. Alexander, G. Fang, J. Rozowsky, M. Snyder, and M. B. Gerstein, "Annotating non-coding regions of the genome," *Nature Reviews Genetics*, vol. 11, no. 8, pp. 559–571, 2010.

[2] M. Esteller, "Non-coding rnas in human disease," *Nature Reviews Genetics*, vol. 12, no. 12, pp. 861–874, 2011.

[3] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn, "Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses," *Genes & development*, vol. 25, no. 18, pp. 1915–1927, 2011.

[4] 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*, vol. 41, no. D1, pp. D983–D986, 2013.

[5] G. U. Ganegoda, M. Li, W. Wang, and Q. Feng, "Heterogeneous network model to infer human disease-long intergenic non-coding rna associations," *NanoBioscience, IEEE Transactions on*, vol. 14, no. 2, pp. 175–183, 2015.

[6] X. Chen and G.-Y. Yan, "Novel human lncRNA–disease association inference based on lncRNA expression profiles," *Bioinformatics*, p. btt426, 2013.

[7] M.-X. Liu, X. Chen, G. Chen, Q.-H. Cui, and G.-Y. Yan, "A computational framework to infer human disease-associated long noncoding rnas," *PloS one*, vol. 9, no. 1, p. e84408, 2014.

[8] P. Jain and I. S. Dhillon, "Provable inductive matrix completion," *arXiv preprint arXiv:1306.0626*, 2013.

[9] W. Liu, N. Zheng, and Q. You, "Nonnegative matrix factorization and its applications in pattern recognition," *Chinese Science Bulletin*, vol. 51, no. 1, pp. 7–18, 2006.

[10] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint 2, 1-norms minimization," in *Advances in neural information processing systems*, 2010, pp. 1813–1821.

[11] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, "OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders," *Nucleic acids research*, vol. 43, no. D1, pp. D789–D798, 2015.

[12] J.-H. Yang, J.-H. Li, S. Jiang, H. Zhou, and L.-H. Qu, "ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data," *Nucleic acids research*, vol. 41, no. D1, pp. D177–D187, 2013.

[13] K. Liu, Z. Yan, Y. Li, and Z. Sun, "Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis," *Bioinformatics*, vol. 29, no. 17, pp. 2221–2222, 2013.

[14] L. Salmena, L. Poliseno, Y. Tay, L. Kats, and P. P. Pandolfi, "A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?" *Cell*, vol. 146, no. 3, pp. 353–358, 2011.

[15] X. Zhang, L. Zhou, G. Fu, F. Sun, J. Shi, J. Wei, C. Lu, C. Zhou, Q. Yuan, and M. Yang, "The identification of an ESCC susceptibility SNP rs920778 that regulates the expression of lncRNA HOTAIR via a novel intronic enhancer," *Carcinogenesis*, p. bgu103, 2014.

[16] J. Gong, W. Liu, J. Zhang, X. Miao, and A.-Y. Guo, "lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse," *Nucleic acids research*, vol. 43, no. D1, pp. D181–D186, 2015.

[17] H. Caniza, A. E. Romero, and A. Paccanaro, "A network medicine approach to quantify distance between hereditary disease modules on the interactome," *Scientific reports*, vol. 5, 2015.

[18] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[19] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui, "Geomf: Joint geographical modeling and matrix factorization for point-of-interest recommendation," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 831–840.

[20] D. Shin, S. Cetintas, K.-C. Lee, and I. S. Dhillon, "Tumblr blog recommendation with boosted inductive matrix completion," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp. 203–212.

[21] H. Gao, J. Tang, X. Hu, and H. Liu, "Content-aware point of interest recommendation on location-based social networks," *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 1721–1727, 2015.