2014 IEEE 14th International Conference on Bioinformatics and Bioengineering

NMF-based LncRNA-Disease Association Inference and Bi-clustering

Ashis Kumer Biswas, Jean X. Gao
Computer Science and Engineering
University of Texas at Arlington
Arlington, Texas, 76019
Email: ashis.biswas@mavs.uta.edu, gao@uta.edu

Baoju Zhang, Xiaoyong Wu
College of Electronics and Communication Engineering
Tianjin Normal University
Tianjin, China 300074

***Abstract*—Long non-coding RNAs (lncRNAs) have been implicated in various biological processes, and are linked in many dysregulations. Researchers have reported large number of lncRNA associated human diseases over the past decade. In this article we employed the Non-negative Matrix Factorization method to develop a low-dimensional computational model that can describe the existing knowledge about lncRNA-disease associations represented in a two dimensional association matrix. The non-negativity constraints of the matrix and its corresponding factors ensure that each lncRNA's disease profile can be represented as an additive linear combination of the latent coordinates. To learn such a constrained model from an incomplete association matrix, several NMF formulations were developed. Based on our experiments, we found that the Sparse NMF obtained the best model among all the other models. Moreover, by exploiting the inherent bi-clustering ability of the NMF models, we extracted several lncRNA groups and disease groups that possess biological significance.**

I. Introduction

With the advent of the Next Generation Sequencing (NGS) platform it is experimentally verified that the protein-coding genes account for only a small fraction of the human genome ($\sim 1.5\%$). In other words, more than 98% of the human genome do not code any protein; the fact implies that the traditional central dogma of molecular biology [1] is incomplete. There exists another branch along with the "traditional" dogma that explains a huge number of the non-protein coding genes that undergo transcription but never translate proteins [2]–[4]. Accumulating evidences reported over the past decade shed lights on many these non-coding RNAs (ncRNAs) and their functionalities in biological processes. The long non-coding RNAs (lncRNAs), a subclass of the ncRNAs having length more than 200 bases are discovered to be associated with many biological processes, such as imprinting control, epigenetic regulation, cell cycle control, nuclear and cytoplasmic trafficking, cell differentiation, immune responses and chromosome dynamics [5]. It is rather not surprising to discover the fact that the dysregulations and mutations of the lncRNAs are implicated in variety of human diseases [6]–[8]. That is why, a comprehensive understanding of potential human disease-related lncRNAs can facilitate development of our current knowledge-base; essentially that could explain accurately the molecular mechanisms of human diseases, their implications and also facilitate the diagnosis, treatment, prognosis and prevention [9], [10].

There are plenty of research efforts that have contributed into characterizing lncRNAs by generating the corresponding sequences, expression profiles and genomic annotations. But, only a few studies have been conducted to infer lncRNA-disease associations. Of the few, Liao et al. [11] proposed the concept of coding-non-coding gene co-expression (CNC) network which was constructed from several gene expression dataset of coding and non-coding genes. The authors then conducted prediction of similar functional characteristics of lncRNAs from the CNC networks using a graph analytical approach. Guo et al. [12] developed a long non-coding RNA global function predictor (lnc-GFP) to predict probable functions for lncRNAs at large scale by integrating gene expression data and protein-protein interaction data. They also employed the concept of CNC network by Liao et al [11]. But here the weighted CNC network was constructed using both the co-expression data and the protein-protein interaction data. Once the CNC network is built, a global propagation algorithm that is guaranteed to converge to a local minimum. The algorithm outputs the rank of all un-annotated genes with respect to a query function category. Finally, the top-ranked genes are functionally annotated with the function category of interest.

Yang et al. [13] presented a method to analyze lncRNA-disease associations, that can be used to predict lncRNA implicated diseases. Based on the available lncRNA-disease associations, two biological networks were constructed – an lncRNA-implicated disease network (lncDN) and disease-associated lncRNA network (DlncN). In lncDN, a vertex represents a disease, and a link between two vertices indicates the two corresponding diseases shared at least one lncRNA as their disease-causing lncRNA. However, in DlncN, a vertex represents an lncRNA, while a link between two nodes represents the fact that the two corresponding lncRNAs were implicated in at least one common disease. A graph analytical approach was applied to extract the similar lncRNAs and disease from these projected networks. Moreover, a propagation algorithm was applied on a weighted bipartite network of the lncRNA-disease associations to

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.54

predict potential. Thus, by modeling the lncRNA-disease association as a bipartite network, and by mining the resultant network lncRNA and disease association scores were predicted. Chen et al. [10] developed a method of Laplacian Regularized Least Squares for LncRNA-Disease Association (LRLSLDA) that considered integrating the lncRNA tissue expression profiles obtained from the Human BodyMap LincRNA project [14]. The method prioritized the entire lncRNAome for disease of interest by integrating known phenome-lncRNAome network obtained from the existing database of lncRNA-disease associations.

Previous studies of lncRNA-disease associations either employed graph mining strategies by employing layered information of both coding and non-coding RNA transcripts, or expression profiles available on different tissue levels [14], even though not all lncRNAs are tissue-specific, and the expression profiles are available only at the intergenic regions of human genome. The profiles cannot be characterized as either from case or control specimens, which is not always appropriate in disease related lncRNA inference studies to some extent. However, if we look at the literatures already published with evidences of lncRNA-disease associations, a computational modeling of the the associations can be explained, that would be useful to predict future associations. In this article, we developed computational framework to build such models.

The rest of the article is structured as follows: section 2 presents the details of the dataset used in this study, along with the preprocessing step. The lncRNA-disease association problem is formulated later in the section. Then, at section 3 we present experimental results along with the applicability of our method. Finally, conclusions and future research directions are drawn at the section 4.

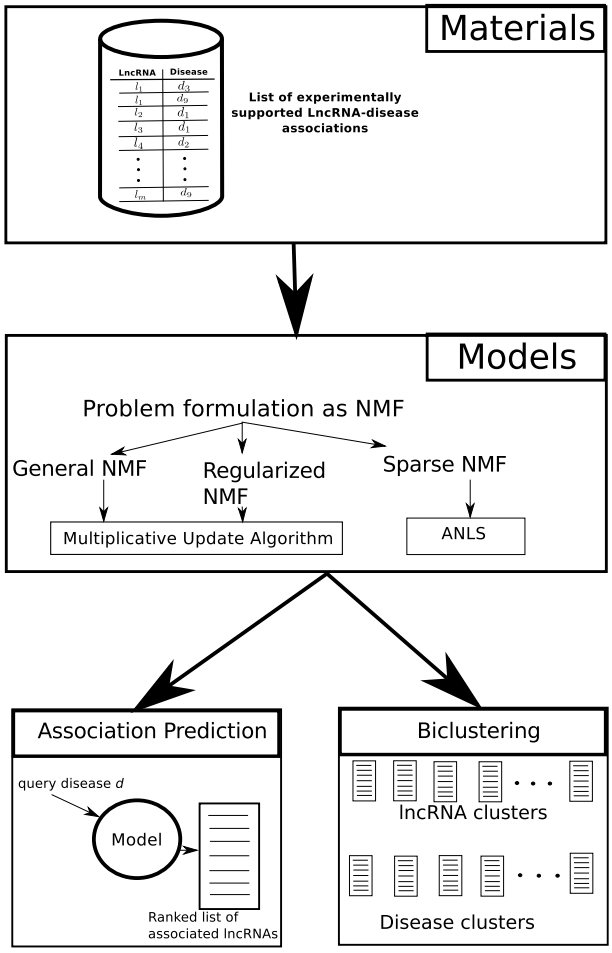

Figure 1. Overview of the proposed method for developing model to predict lncRNA-disease associations.

II. Materials and Methods

In this section, we describe our framework for the development of lncRNA-disease association model, along with steps to cluster both associated lncRNAs, and diseases simultaneously (Figure 1).

A. Data sources and Preprocessing

We obtained the latest version of lncRNA-disease association list from the publicly available LncRNAdisease database repository located at the http://www.cuilab.cn/lncrnadisease [9] on the 15th day of June, 2014. The developers of the repository curated the data manually from PubMed and retrieved lncRNA-disease pairs. The present release (6/15/2014) have integrated 1028 experimentally and published association entries among 322 lncRNAs and 221 diseases. We encoded the association list in a matrix A of size 322×221. Each (i, j) entry of the matrix is filled up as follows: if there is at least one evidence of association between i^{th} lncRNA and j^{th} disease, then the entry would be 1, otherwise 0. An alternate approach of encoding would be to put the frequency of reports supporting the associations in each entry. However, both the encoding schemes enforce the matrix A to be strictly non-negative.

B. NMF-based Problem Formulation

The lncRNA-disease association matrix $A \in \mathbb{R}_{+}^{m \times n}$, where m and n are the number of lncRNAs and diseases respectively, and $A_{i,j} = 1$ denoting there is at least one experimental evidence present that support association between lncRNA i with disease j, otherwise the cell value would be 0. Each column of the matrix A corresponds to a data point in the m-dimensional space. The non-negative matrix factorization (NMF) [15], [16] technique divides such a matrix into two non-negative matrices: a basis matrix of lower rank $W \in \mathbb{R}_{+}^{m \times r}$ and a coefficient matrix $H \in \mathbb{R}_{+}^{r \times n}$, where the rank $r < \min\{m, n\}$, so that

$$A \approx WH \tag{1}$$

An NMF solution is not unique, because of the fact that, for any diagonal matrix $D \in \mathbb{R}_+^{r \times r}$

$$A \approx WH = WDD^{-1}H = (WD)(D^{-1}H) = VG$$

where, $V = WD$ and $G = D^{-1}H$. Both the matrices V and G are not necessarily equal to W and H respectively, implies the non-uniqueness property of the solution to the equation 1.

A solution to the NMF problem, however, can be obtained by solving the following optimization problem:

$$\begin{aligned} \min_{W,H} \quad & \mathcal{F}(A, W, H) \equiv ||A - WH||_F^2 \\ & s.t., W \geq 0, H \geq 0 \end{aligned} \tag{2}$$

where $W \in \mathbb{R}_+^{m \times r}$ is a basis matrix, and $H \in \mathbb{R}_+^{r \times n}$ is a coefficient matrix. $W, H \geq 0$ means that all elements of W and H are non-negative. Since $r < m$ and $r < n$, dimensionality reduction is achieved, and a lower dimensional representation of A in a r-dimensional space is given by H.$||\cdot||_F^2$ is the square of the Frobenius norm and is defined as

$$||A - WH||_F^2 \quad = \quad \mathrm{tr}((A - WH)(A - WH)^T)$$

where tr is the matrix trace operator.

The fact that W and H are non-negative guarantees that parts of the matrix can be combined additively to form the given association matrix as a whole; NMF is a useful technique for obtaining a part-based representation of the data. In other words, factorization allows us to easily identify substructures in the data [17]. Several approaches to solve NMF by iteratively updating W have been presented in earlier studies [18]. Additional Bioinformatics applications of NMF are presented in a review article by Devarajan [19]. Several variants of NMF have been proposed by incorporating various kinds of constraints: discriminative constraints [20], locality-preserving or network-regularized constraints [21], [22] and sparsity constraints [23], [24].

One non-negative matrix factorization algorithm developed by Lee and Seung [15] is based on the multiplicative update rules of W and H, and is shown in Algorithm 1. The approximations of W and H remain non-negative during the updates. It is generally best to update W and H "simultaneously", instead of updating each matrix fully before the other [25]. That is, after updating a row of H, we update the corresponding column of W. In the implementation, we added a small quantity $\epsilon = 10^{-9}$ to the denominators in the approximations of W and H in each iteration.

Pauca et al. [25] proposed a constrained NMF (CNMF) formulation,

$$\begin{aligned} \min_{W,H} \quad & ||A - WH||_F^2 + \alpha||W||_F^2 + \beta||H||_F^2 \\ & s.t., W \geq 0, H \geq 0 \end{aligned} \tag{3}$$

Algorithm 1 General NMF. Calculate W, H such that $A \approx WH$

Input: $A \in \mathbb{R}_+^{m \times n}$, rank r, and the two initial seed matrices $W \in \mathbb{R}_+^{m \times r}$ and $H \in \mathbb{R}_+^{r \times n}$
Step 1: Normalize columns of A.
Step 2: Scale columns of W to sum to 1.
Step 3: Update H and W matrices using the following update rules:

$$H_{qj} \leftarrow H_{qj} \frac{(W^T A)_{qj}}{(W^T W H)_{qj} + \epsilon}, \quad (1 \leq q \leq r, 1 \leq j \leq n)$$

$$W_{iq} \leftarrow W_{iq} \frac{(AH^T)_{iq}}{(WHH^T)_{iq} + \epsilon}, \quad (1 \leq i \leq m, 1 \leq q \leq r)$$

Step 4: Scale columns of W to sum to 1.
Step 5: Repeat steps 3–5 until convergence

where α and β are regularization parameters. Algorithm 2 can be used to retrieve the two factors W and H. The regularization parameters $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$ are

Algorithm 2 CNMF/Regularized NMF. Calculate W, H such that $A \approx WH$

Input: $A \in \mathbb{R}_+^{m \times n}$, rank r, and the two initial seed matrices $W \in \mathbb{R}_+^{m \times r}$ and $H \in \mathbb{R}_+^{r \times n}$
Step 1: Normalize columns of A.
Step 2: Scale columns of W to sum to 1.
Step 3: Update H and W matrices using the following update rules:

$$H_{qj}^{(t)} \leftarrow H_{qj}^{(t-1)} \frac{((W^{(t-1)})^T A)_{qj} - \beta H_{qj}^{(t-1)}}{((W^{(t-1)})^T W^{(t-1)} H^{(t-1)})_{qj} + \epsilon}$$

for $1 \leq q \leq r, 1 \leq j \leq n$

$$W_{iq}^{(t)} \leftarrow W_{iq}^{(t-1)} \frac{(A(H^{(t)})^T)_{iq} - \alpha W_{iq}^{(t-1)}}{(W^{(t-1)} H^{(t)} (H^{(t)})^T)_{iq} + \epsilon}$$

for $1 \leq i \leq m, 1 \leq q \leq r$

Step 4: Scale columns of W to sum to 1.
Step 5: Repeat steps 3–5 until convergence

used to balance the trade-off between the accuracy of the approximation and the smoothness of the computed solution.

Sparseness constraints can be enforced on W or H in the NMF formulation (Equation 2). Kim and Park [24] introduced two formulations and the corresponding algorithms for sparse NMFs – SNMF/L for sparse W, and the SNMF/R for sparse H. The following is the formulation of SNMF/L:

$$\begin{aligned} \min_{W,H} \frac{1}{2}\{||A - WH||_F^2 + \eta||H||_F^2 + \beta \sum_{i=1}^{m} ||W(i,:)||_1^2\} \\ s.t., W \geq 0, H \geq 0 \end{aligned} \tag{4}$$

Here, the parameter β is used to adjust the sparsity in W while the parameter η is used to preserve accuracy in H. And the formulation of the SNMF/R is:

$$\min_{W,H} \frac{1}{2}\{||A - WH||_F^2 + \eta||W||_F^2 + \beta\sum_{j=1}^{n}||H(:,j)||_1^2\} \quad s.t., W \geq 0, H \geq 0 \tag{5}$$

Again, the parameter β is used to adjust the sparsity in H and the parameter η is used to preserve the accuracy in W. Each of these two sparse NMF formulations that imposes the sparsity either on W or H utilizes L_1-norm minimization and the corresponding algorithms are based on Alternating Non-negativity constrained Least Squares (ANLS) [24]. The ANLS problem for SNMF/L is shown below:

$$\min_{H} \left\| \begin{pmatrix} W \\ \sqrt{\eta} I_r \end{pmatrix} H - \begin{pmatrix} A \\ \mathbf{0}_{r\times n} \end{pmatrix} \right\|_F^2, s.t. H \geq 0 \tag{6}$$

$$\min_{W} \left\| \begin{pmatrix} H^T \\ \sqrt{\beta} e_{1\times r} \end{pmatrix} W^T - \begin{pmatrix} A^T \\ \mathbf{0}_{1\times m} \end{pmatrix} \right\|_F^2, s.t. W \geq 0 \tag{7}$$

Similarly, the ANLS problem for the SNMF/R formulation is given below:

$$\min_{H} \left\| \begin{pmatrix} W \\ \sqrt{\beta} e_{1\times r} \end{pmatrix} H - \begin{pmatrix} A \\ \mathbf{0}_{1\times n} \end{pmatrix} \right\|_F^2, s.t. H \geq 0 \tag{8}$$

$$\min_{W} \left\| \begin{pmatrix} H^T \\ \sqrt{\eta} I_r \end{pmatrix} W^T - \begin{pmatrix} A^T \\ \mathbf{0}_{k\times m} \end{pmatrix} \right\|_F^2, s.t. W \geq 0 \tag{9}$$

C. LncRNA-Disease Association Inference

Non-negative Matrix Factorization models map both the lncRNAs and diseases to a joint latent factor space of dimensionality r, such that lncRNA-disease association are modeled as the inner products in the latent feature space $(f_1, f_2, \cdots, f_r)$. Accordingly, each lncRNA i is associated with a vector $\mathbf{l_i} \in \mathbb{R}^r$, and each disease j is associated with a vector $\mathbf{d_j} \in \mathbb{R}^r$. Thus, for a given lncRNA i, the elements of the vector $\mathbf{l_i}$ measure the extent to which the lncRNA possesses those factors, whereas for a given disease j, the elements of $\mathbf{d_j}$ measure the likelihood of association of the disease with corresponding factors. The dot product $\mathbf{l_i}^T\mathbf{d_j}$ captures the association between lncRNA i and disease j. This approximates the overall association of disease j with lncRNA i, that is denoted by $\hat{a}_{ij}$ leading to the estimate

$$\hat{a}_{ij} = \mathbf{l_i}^T \cdot \mathbf{d_j} \tag{10}$$

Once the NMF factorization is complete on matrix A, the inference system can easily estimate the likelihood of association of an lncRNA with a disease using equation 10. Figure 2 illustrates the inference process.

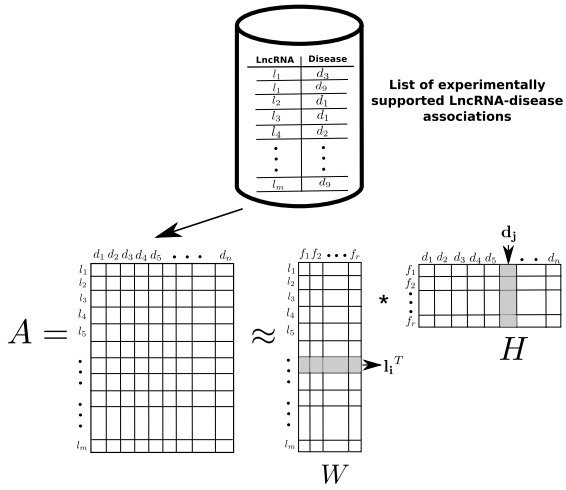

Figure 2. An abstract view of the lncRNA-disease association inference process. At first from the list of experimentally supported lncRNA-disease associations, the original association matrix A is formed, where $A_{ij} = x$, and $x \geq 0$ is a positive integer denoting number of experimental evidences that support the association between i^{th} lncRNA and j^{th} disease. Then NMF is applied to factor A into two matrices W and H. The corresponding rows of W and columns of H are then used to estimate the likelihood of the association betwen lncRNAs with diseases.

D. Bi-clustering

Many traditional clustering algorithms such as Hierarchical clustering have been applied for the purpose of clustering gene micro-array data which is an association between genes and samples to some extent [26], [27]. These strategies have a significant limitation: the approaches assign samples into some specific classes based on the genes' expression levels across all the samples. Sometimes, it is necessary to develop clustering methods that can identify the local structures, instead of the global phenomenon. Moreover, it has been shown in molecular biology that only a small number of genes or lncRNAs are involved in a pathway or biological process on most cases. Specifically, only a small subset of lncRNAs are active for one cancer type, or one dysfunction, so generating sparse bi-clustering structures (i.e., the number of genes in each bi-clustering structure is small) is of great interest [28]. Many bi-clustering algorithms have been developed to explore the correlations between genes and samples and to identify the local gene-sample structures in the micro-array data, and some other association data [29]. However, the idea of bi-clustering is to characterize each lncRNA by a subset of diseases and to define each disease in a similar way. As a consequence, bi-clustering algorithms can select the groups of lncRNAs that show similar expression behaviors in a subset of diseases that belong to some specific classes such as some specific cancers, or disorders, and thus identify local structures of the association data.

Several bi-clustering algorithms have been proposed including BiMax, ISA, SAMBA, OPSM, which are evaluated in the review by Prelic et al [29]. However, bi-clustering

can also be performed using NMF. The NMF factors can be used to perform bi-clustering analysis of the data matrix. The rows of the association matrix A represent lncRNAs, and the columns represent diseases. We can use the basis matrix W to divide the m lncRNAs into r lncRNA-clusters, and the coefficient matrix H can be used to divide the n diseases into r disease-clusters. Typically the following rules are used to assign membership:

- i^{th} lncRNA is assigned to the lncRNA-cluster q if the W_{iq} is the largest in $W(i,:)$, i.e., the i^{th} row of the matrix W.
- j^{th} disease is assigned to the disease-cluster q if the $H_{q,j}$ is the largest in $H(:,j)$, i.e., the j^{th} column of the matrix H.

III. Experiment Results and Discussions

A. Association Models

To evaluate the performances of the models, we preferred three widely used metrics, namely Mean Absolute Error (MAE), Accuracy and Root Mean Squared Error (RMSE) [30], which are defined as follow:

$$\text{MAE}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{|\tau|} \sum_{(i,j)\in\tau} |\hat{y}_{ij} - y_{ij}| \tag{11}$$

$$\begin{aligned} \text{Accuracy}(\hat{\mathbf{Y}}, \mathbf{Y}) &= \frac{1}{|\tau|} \sum_{(i,j)\in\tau} (1 - |\hat{y}_{ij} - y_{ij}|) \\ &= 1 - MAE(\hat{\mathbf{Y}}, \mathbf{Y}) \end{aligned} \tag{12}$$

$$\text{RMSE}(\hat{\mathbf{Y}}, \mathbf{Y}) = \sqrt{\frac{1}{|\tau|} \sum_{(i,j)\in\tau} (\hat{y}_{ij} - y_{ij})^2} \tag{13}$$

where $\hat{\mathbf{Y}}$ and $\mathbf{Y}$ are the computed association matrix and the observed association matrix respectively, while τ is the set of lncRNA-disease pairs for which we want to predict the ratings, that is, τ can be considered as the test set. The preference between the above two metrics depends on the particular application. In practice, MAE is popular for many collaborative filtering algorithms, while RMSE is still popular for the similar problems that generate real valued output.

The lncRNA-disease associations are first split into five random folds. Then we performed five-fold cross validation to evaluate the model. Table I demonstrates the predictive performance of lncRNA-disease associations by using the three NMF models. The association matrix is first factored into W and H matrices using the three NMF algorithms. We performed several runs of NMF by varying rank of W and H, which are $r = 2, 10, 20, 30, 40, 50$. Then the original matrix is reconstructed by multiplying the computed W and H matrices. The estimated matrix is then compared with the original matrix for errors, in terms of RMSE and MAE scores.

Since, the accuracy and MAE scores are exactly complement to each other, it is evident from the Table I that once we increase the rank of the NMF factorization, the error decreases, as well as accuracy increases. As the rank of the NMF in our current dataset can only be less than 221 (that is, the minimum of the two dimensions of the association matrix), we showed here only the effect of choosing rank less than 50. The trend of accuracy and MAE can be equally observed in all of the NMF-based.

If we look at the trend of RMSE scores as the increment of rank in the various NMF implementations, we find all of the three algorithms show almost similar RMSE trend. However, since the input association matrix possesses sparsity property, it is better to use the sparse NMF considering the sparsity property into account. Thus, from Table I results we can conclude that the sparse NMF (SNMF/R) performed better than the other two NMF algorithms.

B. Bi-clustering Results

As explained in the previous section, a bi-clustering algorithm enables us to explore groups of entities that are similar within a small locality. Essentially, we are more interested to identify groups of lncRNAs that are associated with a very similar group of diseases, or disorders. Since, other than the lncRNA-disease association information we did not use any other characteristics of lncRNAs in our NMF-based formulations in order to understand similarities between lncRNA-pairs, it is not evidently interesting to perform clustering on the diseases that would reveal similar diseases groups. However, on the contrary, grouping lncRNAs reveals a number of useful characterization of lncRNAs in terms of the implication of diseases.

Table II lists out 10 significant clusters of lncRNAs that we retrieved after we performed a generalized NMF on the association matrix, and sought for two factors W and H of rank 322×10, and 10×221, meaning we expected a $r = 10$ rank approximation of the association matrix. Using the bi-clustering strategy described in the earlier section, we assigned membership scores for each of the 322 lncRNAs to any of the 10 disease classes. Here the latent feature space is 10-dimensional.

We then put the class major disease associations to the lncRNAs in Table II and found interesting lncRNA groups. For instance, there we see a prominent group (cluster # 5) of lncRNAs which are associated with heart diseases. All the lncRNAs in cluster #7 are associated with neurological disorders to some extent. Cluster # 6 contains all the lncRNAs which are mostly associated with hereditary disorders.

Cluster #1 is representing mostly the gastro-intestinal dysfunctions. All the remaining clusters are representative to several cancer categories and associated lncRNAs. A similar approach can also be employed to cluster the 221 diseases in the pool, according to the 10 latent features.

Table I
EVALUATION OF THE THREE NMF ALGORITHMS – GENERAL/PLAIN NMF, REGULARIZED NMF AND SPARSE NMF IN TERMS OF ACCURACY, MEAN ABSOLUTE ERROR (MAE) AND ROOT MEAN SQUARED ERROR (RMSE) BY VARYING RANKS (r).

r	Accuracy			MAE			RMSE		
	Plain NMF	Regularized NMF	Sparse NMF	Plain NMF	Regularized NMF	Sparse NMF	Plain NMF	Regularized NMF	Sparse NMF
2	0.16	0.15	0.15	0.84	0.85	0.85	20.87	21.23	21.23
10	0.46	0.45	0.46	0.54	0.55	0.54	13.48	13.64	13.56
20	0.61	0.61	0.62	0.39	0.39	0.38	9.67	9.71	9.59
30	0.70	0.70	0.69	0.30	0.30	0.31	7.50	7.58	7.66
40	0.76	0.76	0.77	0.24	0.24	0.23	6.02	6.06	5.66
50	0.81	0.80	0.82	0.19	0.20	0.18	4.70	4.90	4.57

IV. CONCLUSION AND FUTURE RESEARCH SCOPES

Many lncRNAs play critical roles in human diseases and disorder pathways. An lncRNA may implicate multiple diseases, while a disease could be a result by association of several canonical lncRNAs. A comprehensive understanding of the associations is necessary in diagnosis, and novel drug discovery, and future research in this domain. However, a very little is known about the association of lncRNAs with diseases as compared to the exponential rate of discovery of the lncRNAs.

In this article, we proposed the NMF-based formulation of the lncRNA-disease association problem, and implemented the NMF algorithms to solve the problem by constructing models. The models have two-fold properties – they are able to explain each of the associated lncRNA as well as the disease in a latent feature space that can be considered a dimensionality reduction step before further processing. Secondly, the NMF factors can be used to retrieve bi-clusters, that is, groups of similar lncRNAs, and groups of similar diseases in the latent feature dimension. We found that any NMF-based formulation that only considers the existing knowledge of lncRNA-disease association would be fair enough to be used in practical association prediction problems. Our NMF-based solution provides rank of lncRNAs for a query disease, as well as rank of diseases for a query lncRNA. This property can facilitate the relevant research community who are experimenting to identify potential lncRNA-disease associations that are missing in the existing knowledge-base.

There are some limitations exist in our NMF-based lncRNA-disease association approach. Firstly, there are several parameters in each of the NMF algorithms we used in the study, and we still need a better way to select the right parameters that can solve well. Secondly, we only incorporated the existing lncRNA-disease association knowledge to infer future or missing associations. But, it would be a useful study that incorporate both the lncRNA-lncRNA relationship, and disease-disease relationship information into the NMF-based problem formulation, and some other information pertaining to the lncRNA and disease association model. Thirdly, other genomic, or transcriptomic information about lncRNAs can be incorporated to develop a robust association prediction system that can identify novel potential lncRNA-associated diseases as well.

REFERENCES

[1] F. Crick *et al.*, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[2] R. J. Taft, K. C. Pang, T. R. Mercer, M. Dinger, and J. S. Mattick, "Non-coding rnas: regulators of disease," *The Journal of pathology*, vol. 220, no. 2, pp. 126–139, 2010.

[3] J. E. Wilusz, H. Sunwoo, and D. L. Spector, "Long noncoding rnas: functional surprises from the rna world," *Genes & development*, vol. 23, no. 13, pp. 1494–1504, 2009.

[4] A. Machado-Lima, H. A. del Portillo, and A. M. Durham, "Computational methods in noncoding rna research," *Journal of mathematical biology*, vol. 56, no. 1-2, pp. 15–49, 2008.

[5] S. Chung, H. Nakagawa, M. Uemura, L. Piao, K. Ashikawa, N. Hosono, R. Takata, S. Akamatsu, T. Kawaguchi, T. Morizono *et al.*, "Association of a novel long non-coding rna in 8q24 with prostate cancer susceptibility," *Cancer science*, vol. 102, no. 1, pp. 245–252, 2011.

[6] L. D. Sacco, A. Baldassarre, and A. Masotti, "Bioinformatics tools and novel challenges in long non-coding rnas (lncrnas) functional analysis," *International journal of molecular sciences*, vol. 13, no. 1, pp. 97–114, 2011.

[7] P. Kapranov, J. Cheng, S. Dike, D. A. Nix, R. Duttagupta, A. T. Willingham, P. F. Stadler, J. Hertel, J. Hackermüller, I. L. Hofacker *et al.*, "Rna maps reveal new rna classes and a possible function for pervasive transcription," *Science*, vol. 316, no. 5830, pp. 1484–1488, 2007.

[8] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding rnas: insights into functions," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 155–159, 2009.

[9] 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.

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

[11] Q. Liao, C. Liu, X. Yuan, S. Kang, R. Miao, H. Xiao, G. Zhao, H. Luo, D. Bu, H. Zhao *et al.*, "Large-scale prediction of long non-coding rna functions in a coding–non-coding gene co-expression network," *Nucleic acids research*, vol. 39, no. 9, pp. 3864–3878, 2011.

[12] X. Guo, L. Gao, Q. Liao, H. Xiao, X. Ma, X. Yang, H. Luo, G. Zhao, D. Bu, F. Jiao *et al.*, "Long non-coding rnas function annotation: a global prediction method based on bi-colored networks," *Nucleic acids research*, vol. 41, no. 2, pp. e35–e35, 2013.

[13] X. Yang, L. Gao, X. Guo, X. Shi, H. Wu, F. Song, and B. Wang, "A network based method for analysis of lncrna-disease associations and prediction of lncrnas implicated in diseases," *PLOS ONE*, vol. 9, no. 1, p. e87797, 2014.

[14] 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.

[15] 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.

[16] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[17] S. Zhang, Q. Li, J. Liu, and X. J. Zhou, "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules," *Bioinformatics*, vol. 27, no. 13, pp. i401–i409, 2011.

[18] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 155–173, 2007.

[19] K. Devarajan, "Nonnegative matrix factorization: an analytical and interpretive tool in computational biology," *PLoS computational biology*, vol. 4, no. 7, p. e1000029, 2008.

[20] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *Neural Networks, IEEE Transactions on*, vol. 17, no. 3, pp. 683–695, 2006.

[21] D. Cai, X. He, X. Wu, and J. Han, "Non-negative matrix factorization on manifold," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 63–72.

[22] Q. Gu and J. Zhou, "Local learning regularized nonnegative matrix factorization," in *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[23] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[24] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.

[25] V. P. Pauca, J. Piper, and R. J. Plemmons, "Nonnegative matrix factorization for spectral data analysis," *Linear algebra and its applications*, vol. 416, no. 1, pp. 29–47, 2006.

[26] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14 863–14 868, 1998.

[27] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences*, vol. 96, no. 6, pp. 2907–2912, 1999.

[28] Z.-Y. Zhang, T. Li, C. Ding, X.-W. Ren, and X.-S. Zhang, "Binary matrix factorization for analyzing gene expression data," *Data Mining and Knowledge Discovery*, vol. 20, no. 1, pp. 28–52, 2010.

[29] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006.

[30] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.

Table II
10 PROMINENT CLUSTERS OF LNCRNAS THAT ARE RETRIEVED FROM OUR NMF MODELS

ID	Mostly associated diseases	lincRNAs
1	gastric cancer, liver related cancer, kidney injury	AIR, CCAT1, DQ786243, Dreh, ENST00000513542, GNAS-AS1, HEIH, HOTTIP, HULC, IGF2-AS, KCNQ1OT1, LALR, LDMAR, LINCMD1, lncRNA-ATB, lncRNA-MVIH, MINA, MIR7-3HG, NPTN-IT1, np_17856, np_5318, RNA polymerase III-dependent lncRNAs, RNase MRP, VL30 LTRs
2	Esophageal squamous cell cancer, type II diabetes, melanoma	1B FGF-antisense transcripts, Alu lncRNAs, CDKN2B-AS1, CDKN2B-AS10, CDKN2B-AS11, CDKN2B-AS13, CDKN2B-AS2, CDKN2B-AS3, CDKN2B-AS5, CDKN2B-AS7, CDKN2B-AS8, D4Z4, ESCCAL-1, ESCCAL-5, ESRG, Gm20748, HI-LNC25, HYMAI, KUCG1, LINC00032, LINC01262, NPPA-AS1, NRON, PDZRN3-AS1, PISRT1, PTHLH, SPRY4-IT1
3	Angelman syndrome, Prader-Willi syndrome, Silver-Russell syndrome	116HG, AK023948, anti-NOS2A, BDNF-AS1, C15orf2, H19, IPW, KCNQ1DN, MAP3K14, MESTIT1, MIR100HG, MKRN3-AS1, SCAANT1, SLC7A2-IT1A/B, SNHG11, Ube3a-as, UBE3A-AS1, UBE3A-ATS
4	prostate cancer, enterovirus infection, autoimmune disease	AC002511.1, AP000688.29, ATXN8OS, C1QTNF9B-AS1, CBR3-AS1, CCND1 promoter-derived lncRNAs, CDKN2B-AS9, CTBP1-AS, DAPK1, DLEU1, DLEU2, DNM3OS, GAS5, Kcna2 antisense RNA, LINC00162, Linc00963, LOC728606, LSINCT5, MIR155HG, NAMA, PCA3, PCGEM1, PCNCR1, PRNCR1, PVT1, RP4-620F22.3, RP5-843L14.1, SCHLAP1, SNHG5, SRA1, TCL6, TERC, ZFAT-AS1
5	Heart Failure	5730458M16Rik, AK038798, AK044955, AK049728, AK137898, AK144081, AK153778, BX118339, DMPK, DMPK 3'UTR, ENSMUST00000022467, ENSMUST00000041159, ENSMUST00000117372, ENSMUST00000117393, ENSMUST00000119855, ENSMUST00000120925, ENSMUST00000127230, ENSMUST00000127429, ENSMUST00000130025, ENSMUST00000142855, ENSMUST00000143888, ENSMUST00000160947, ENSMUST00000167632, FADS1, Fendrr, Gm12839, Gm6644, LIPCAR, LOC102635190, Scarb2, Trpm3, uc.115-, uc.184+, UCH1LAS, Zim3
6	Hereditary Haemorrhagic Telangiectasia, fragile X syndrome	B1 SINE RNA, ENSG00000135253.9, ENSG00000147753.5, ENSG00000196096.3, ENSG00000197251.3, ENSG00000203325.3, ENSG00000206129.3, ENSG00000215231.3, ENSG00000215374.4, ENSG00000215808.2, ENSG00000226496.1, ENSG00000229563.1, ENSG00000230133.1, ENSG00000230544.1, ENSG00000231133.1, ENSG00000231185.2, ENSG00000232021.2, ENSG00000232046.1, ENSG00000232956.3, ENSG00000233154.1, ENSG00000233251.3, ENSG00000235285.1, ENSG00000237036.3, ENSG00000237548.1, ENSG00000240453.1, ENSG00000241269.1, ENSG00000245910.3, ENSG00000248176.1, ENSG00000249364.1, ENSG00000249772.1, ENSG00000250195.1, ENSG00000250608.1, ENSG00000254154.3, ENSG00000255471.1, ENSG00000256218.1, ENSG00000259150.1, ENSG00000259334.1, ENSG00000259484.1, ENSG00000259758.1, ENSG00000263753.1, ENSG00000264772.1, ENSG00000266952.1, FMR4, FMR6, RNA-a
7	Alzheimer's disease, bipolar disorder, Huntington's disease, schizophrenia, depression, DiGeorge syndrome, Intellectual and developmental disability, psychiatric disease	51A, 7SL, BACE1-AS, BCYRN1, BDNF-AS, DAOA-AS1, DGCR5, DISC2, DLG2AS, FGF10-AS1, GDNFOS, HAR1A, HAR1B, HCP5, HELLPAR, HLA-AS1, HTTAS, HTTAS_v1, IFNG-AS1, LINC00271, LINC00299, LOC389023, NEAT-1, PRINS, PSORS1C3, PTCSC, PTCSC3, REST/CoREST-regulated lncRNAs, SNHG3, SOX2-OT, TRAF3IP2-AS1, TUG1
8	ischemia/reperfusion, neuroblastoma, breast cancer	4930503E24Rik, ADAMTS9-AS2, AK028007, AK139328, AK143294, AK143693, Alg2, BCAR4, BLACAT1, CCDC26, CCND1, CDKN2B-AS12, CDKN2B-AS6, CECR3, CHL1-AS2, DLX6-AS1, DSCAM-AS1, ENSMUST00000151138, GHET1, HIF1A-AS1, HIF1A-AS2, lncRNA-LET, Loc554202, MEG3, MIR31HG, MYCNOS, ncRAN, NDM29, PAN, PANDA, PCNA-AS1, PINC, PTENpg1, Sema3g, SNHG16, SNHG4, Sox4, Srsf9, SUMO1P3, T-UCRs, UCA1, WT1-AS, Wt1os, ZNFX1-AS1
9	misc. cancers	7SK, A130040M12Rik, AK143260, ANRIL, B2 SINE RNA, BOK-AS1, CDKN2B-AS4, CHRF, DANCR, DBET, EPB41L4A-AS1, FMR5, HNF1A-AS1, KRAS1P, MIAT, MIR17HG, PANDAR, PCAT1, PTENP1, RRP1B, RUNXOR, TC0100223, TC0101441, TC0101686, TDRG1, TERRA, TINCR, Trp53cor1, WRAP53, XIST, Yiya
10	Parkinson's disease, nasopharyngeal carcinoma, cat eye syndrome, Kawasaki disease	ACTA2-AS1, AF086415, AFAP1-AS1, AK042766, AK056098, AK095147, AK294004, AP5M1, ASFMR1, ATP6V1G2-DDX39B, BANCR, BC040587, BPESC1, CASC2, CCAT2, CECR9, CRNDE, HOTAIR, LINC00312, LincRNA-p21, lnc-AL355149.1-1, lnc-C22orf32-1, lnc-ZNF674-1, LncRNA-LALR1, LSAMP-AS3, MALAT1, PINK1-AS, PPP3CB, RMST, RP1-179N16.3, THRIL, TUSC8, U1 spliceosomal lncRNA