Hans Journal of Computational Biology 计算生物学, 2022, 12(2), 16-22
Published Online June 2022 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2022.122003

Hans汉斯

# 矩阵补全算法在预测长链非编码RNA与蛋白质关联中的应用

**赵靖轩**

辽宁科技大学计算机与软件工程学院，辽宁 鞍山

收稿日期：2022年5月13日；录用日期：2022年6月13日；发布日期：2022年6月22日

## 摘　要

**长链非编码RNA (Long non-coding RNA, lncRNA)指的是序列长度大于200 nt，且不能直接翻译成蛋白质的一类RNA，伴随着生物信息学的不断发展进步，研究人员已经在很多实验中证实长链非编码RNA在人体发育过程中扮演着至关重要的作用，它通常会与蛋白质发生相互作用来实现其生物学功能，因此预测长链非编码RNA与蛋白质的潜在关联有着十分重要的意义。在本文中，我们提出了一种利用矩阵补全算法来预测长链非编码RNA与蛋白质相互作用的模型，称为LPIMC。它能够利用由长链非编码RNA相似性网络、蛋白质相似性网络、长链非编码RNA与蛋白质相互作用矩阵结合而来的异构网络，通过最小化核范数实现矩阵补全来生成新的相互作用邻接矩阵。5折交叉验证下证明，该模型能够有效预测长链非编码RNA-蛋白质关联。**

## 关键词

**长链非编码RNA，蛋白质，相互作用，矩阵补全，异构网络**

# Predicting Association between Long Chain Noncoding RNA and Protein Based on Matrix Completion Algorithm

**Jingxuan Zhao**

School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan Liaoning

Received: May 13th, 2022; accepted: Jun. 13th, 2022; published: Jun. 22nd, 2022

## Abstract

**Long non-coding RNA (lncRNA) refers to a class of RNA whose sequence length is more than 200 nt**

文章引用: 赵靖轩. 矩阵补全算法在预测长链非编码 RNA 与蛋白质关联中的应用[J]. 计算生物学, 2022, 12(2): 16-22.
DOI: 10.12677/hjcb.2022.122003

**and cannot be directly translated into protein. With the continuous development and progress of bioinformatics, researchers have confirmed in many experiments that long non-coding RNA plays a crucial role in human development. It usually interacts with proteins to fulfill its biological functions, so it is very important to predict the potential association between long non-coding RNAs and proteins. In this paper, we propose a model called LPIMC that uses matrix completion algorithms to predict the interaction between long non-coding RNAs and proteins. It can generate a new adjacency matrix by using heterogeneous networks combining long non-coding RNA similarity network, protein similarity network and long non-coding RNA and protein interaction matrix, and achieve matrix completion by minimizing the nuclear norm. The model can effectively predict the long non-coding RNA-protein association under 5-fold cross validation.**

## Keywords

**Long-Chain Noncoding RNA, Protein, Interaction, Matrix Completion, Heterogeneous Network**



## 1. 介绍

非编码 RNA (Non-coding RNA, ncRNA)是指不参与编码产生蛋白质的 RNA，这些 RNA 曾被认为是无用的，长链非编码 RNA (Long non-coding RNA, lncRNA)是长度大于 200 个核苷酸的非编码 RNA，越来越多的研究表明尽管不能直接编码为蛋白质，长链非编码 RNA 仍可以参与一系列生物学过程，例如遗传表现调节、肿瘤生长、免疫反应等[1]。因此了解并识别潜在的长链非编码 RNA-蛋白质关联是很有必要的。虽然现在已有大规模实验方法可以确定两者的关联，但这些实验往往会花费大量的时间和物质成本。如今，研究人员已经将机器学习及深度学习方法用到了预测两者关系上，并取得了优良结果。

近些年来，利用机器学习方法来预测长链非编码 RNA 与蛋白质相互作用(lncRNA-protein interaction, LPI)已经取得了丰硕的成果，Pan 等人[2]在 2016 年提出的一种基于序列信息的利用堆叠式自动编码器来进行预测 LPI 的模型，称为 IPMiner。随后，Xiao 等人[3]在 2017 提出了一种整合异构网络使用 HeteSim 评分来预测 LPI 的模型，称为 PLPIHS。在 2018 年，Hu 等人[4]使用线性集成策略集成了支持向量机(Support Vector Machines, SVM)、随机森林(Random Forest, RF)和极端梯度增强(eXtreme Gradient Boosting, XGB)三种模型，并且使用了由三种不同方法提取出来的序列特征，构建了 HLPI-Ensemble 模型。2019 年，Zhan 等人[5]提出了一个名为 BGFE 的基于序列的方法来预测非编码 RNA 与蛋白质相互作用，该模型将堆叠自动编码器网络与随机森林分类器相结合，并采用了奇异值分解从序列 k-mers 稀疏矩阵中提取特征向量。2020 年，Yi 等人[6]提出了一个堆叠集成计算模型 RPI-SE，该模型集成了梯度提升决策树(Gradient Boosting Decision Tree, GBDT)、支持向量机(SVM)、和极端随机树算法，同样基于序列信息来预测 LPI。

深度学习同样在长链非编码 RNA 与蛋白质作用关系预测领域有了很广泛的应用。在 2020 年，Zhang 等人[7]提出了一种基于卷积神经网络的并采用了复制填充技巧的深度学习模型，称为 LPI-CNNCP。在 2021 年，Shen 等人[8]提出了一种基于 GNN 的非编码 RNA-蛋白质相互作用预测方法，称为 NPI-GNN，它能够根据网络信息和序列信息预测新的交互。在 2021 年，Li 等人[9]提出了一种新的多通道胶囊网络框架，集成用于 LPI 预测的多模特征，叫做 Capsule-LPI。在 2021 年，Jin 等人[10]提出一种基于图自动编码器和协同训练的端到端深度学习模型来预测长链非编码 RNA 和蛋白质相互作用，称为 LPIGAC，该

方法通过对在长链非编码 RNA 网和蛋白质网上实现的图自动编码器进行协同训练来提取特征。

此外，许多用于长链非编码 RNA-蛋白质关联预测的半监督学习方法都用到了矩阵分析相关知识。Ge 等人[11]通过两步传播过程在二分网络上以矩阵迭代形式预测 lncRNA-蛋白质相互作用。Zhang 等人[12]则提出了一种线性邻域传播方法来解决这个问题。Zhao 等人[13]综合了随机游走和邻域正则化逻辑矩阵分解两种算法，来得到潜在的长链非编码 RNA-蛋白质相互作用打分矩阵。Zhang 等人[14]在非负矩阵分解的基础上添加了图正则化来进一步改善模型性能。

在本文中，我们提出了一种利用矩阵补全算法来预测长链非编码 RNA-蛋白质相互作用关系的模型，称为 LPIMC。我们的模型首先基于长链非编码 RNA 与蛋白质的相似矩阵及相关性矩阵构建了一个异构网络，然后使用将矩阵补全问题转化为最小化核范数方法来补全目标矩阵，最终得到两者的预测打分矩阵，以从中获取潜在的长链非编码 RNA-蛋白质关联。

## 2. 数据获取

我们的数据集包含从 NPInter v2.0 数据库[15]中下载筛选得到 8112 条长链非编码 RNA-蛋白质关联信息，其中涉及到了 3046 个长链非编码 RNA 和 136 个蛋白质。它们的序列信息已经被 NONCODE 数据库[16]和 UniProt 数据库[17]证实。为了考察模型在不同数据集上的表现，我们还从 lncRNome [18]上采集了一个新数据集，经过筛选去掉和 NPInter 重复的部分，最终得到了 2729 对相互作用数据，涉及到 1184 个长链非编码 RNA 和 9 个蛋白质。

### 2.1. 构建邻接矩阵

我们首先将 3046 个长链非编码 RNA 与 136 个蛋白质进行了重新编号，并由此将此前的 8112 条关联信息转化为了一个 $n \times m$ 维的邻接矩阵，表示为 $Y = R^{n \times m}$，$n = 3046$ 表示长链非编码 RNA 的数量，$m = 136$ 表示蛋白质的数量。对应的，若 $Y_{(i,j)} = 1$，则代表第 $i$ 个长链非编码 RNA 与第 $j$ 个蛋白质之间存在关联；若 $Y_{(i,j)} = 0$，则代表第 $i$ 个长链非编码 RNA 与第 $j$ 个蛋白质之间无已知关联信息。

### 2.2. 构建相似性矩阵

这里采用高斯核(GIP)相似性来[19]分别构建长链非编码 RNA 与蛋白质的相似性矩阵。高斯核相似性矩阵是利用邻接矩阵 $Y$ 的拓扑信息构建的。在这里，使用 $GS_l$ 来代表长链非编码 RNA 的 GIP 相似性矩阵，使用 $GS_p$ 来代表蛋白质的 GIP 相似性矩阵，用长链非编码 RNA 举例说明，使用 AP 向量来代表当前 RNA 与其他所有蛋白质的关联信息，AP 向量为从邻接矩阵 $Y$ 中获得的当前 RNA 所在的具体行，然后使用下面公式来计算第 $i$ 个与第 $j$ 个长链非编码 RNA 的 GIP 相似性：

$$GS_l\left(l_i, l_j\right) = \exp\left(-\gamma_l \left\| AP\left(l_i\right) - AP\left(l_j\right) \right\|^2\right),$$

$$\gamma_l = \gamma_l' \Big/ \left[\frac{1}{n}\sum_{i=1}^{n} \left\| AP\left(l_i\right) \right\|^2\right]$$

其中 $\gamma_l$ 是针对于长链非编码 RNA 的 GIP 相似性正则化核带宽参数，$\gamma_l'$ 是源带宽参数。同理，$GS_p$ 的计算方法与之类似。

## 3. 算法实现

算法思想为将长链非编码 RNA-蛋白质关联预测问题转换为矩阵补全问题，然后使用最小化核范数来求解得到预测矩阵。首先，基于上面构建的长链非编码 RNA 与蛋白质相互作用邻接矩阵 $Y$ 和两者各自的

GIP 核相似性矩阵 $GS_l$ 与 $GS_p$，我们搭建了一个异构的长链非编码 RNA-蛋白质网络 $T$ 视为目标矩阵，表示如下：

$$T=\begin{bmatrix} S_l & Y \\ Y^{\mathrm{T}} & S_p \end{bmatrix}$$

不难得出目标矩阵 $T$ 为 $(n+m)\times(n+m)$ 维，构建目标矩阵 $T$ 的目的是未来使用求解得来的长链非编码 RNA-蛋白质预测得分来填充缺失值。由于目标矩阵为低秩的，在此我们将矩阵补全问题转换为最小化目标矩阵秩的问题[20]。众所周知，最小化目标矩阵秩在计算起来是 NP 难度的，故通常将秩最小化问题转化为最小化矩阵核范数的问题，该结论已经被研究人员证明[21]。我们在研究过程中借鉴了 ADMM 算法[22]，ADMM 是一个不仅可以分解，并且在收敛性能上表现优越的算法模型。故该问题可转化为：

$$\min_X \|X\|_*,$$

$$\text{s.t. } P_\Omega(X)=P_\Omega(T)$$

$\|X\|_*$ 代表 $X$ 的核范数，$\Omega$ 是目标矩阵节点的坐标集，与已知的长链非编码 RNA-蛋白质对相对应。$P_\Omega$ 是 $\Omega$ 上的正交投影算子。

$$\left(P_\Omega(X)\right)_{ij}=\begin{cases} X_{ij}, & (i,j)\in\Omega \\ 0, & \text{else} \end{cases}$$

除此之外，为了进一步提升 ADMM 算法，我们在方程中加入了正则化项和矩阵值约束，以确保得到的预测打分落在(0, 1)的范围内，因为(0, 1)范围外的打分没有任何意义。由此得到的关键方程如下：

$$\min_X \|X\|_*+\frac{\alpha}{2}\|P_\Omega(X)-P_\Omega(T)\|_F^2$$

$$\text{s.t. } 0<X_{ij}<1\left(0\le i,j\le n+m\right)$$

其中 $\alpha$ 是表征误差项的参数，$0<X_{ij}<1$ 代表 $X$ 中的所有元素都在(0, 1)的范围内。我们在前面的函数中还引入了一个辅助矩阵 $W$ 来进一步提高模型收敛性能。最终目标函数如下：

$$\min_X \|X\|_*+\frac{\alpha}{2}\|P_\Omega(W)-P_\Omega(T)\|_F^2,$$

$$\text{s.t. } X=W,\ 0<W_{ij}<1\left(0\le i,j\le n+m\right)$$

在上述公式的基础下，增广拉格朗日函数为：

$$L(W,X,Y,\alpha,\beta)=\|X\|_*+\frac{\alpha}{2}\|P_\Omega(W)-P_\Omega(T)\|_F^2+Tr\left(Y^{\mathrm{T}}(X-W)\right)+\frac{\beta}{2}\|X-W\|_F^2$$

其中 $Y$ 是拉格朗日乘数，$\beta>0$ 代表惩罚参数。我们将 $W$, $X$ 和 $Y$ 初始化为 $P_\Omega(T)$，然后进行迭代，在第 $k$ 次迭代时，根据 ADMM 可以计算得来 $W_{k+1}$、$X_{k+1}$ 和 $Y_{k+1}$。当迭代终止时，最终的预测矩阵 $W^*$ 可以根据之前 $T$ 的形式进行划分，其中 $A^*$ 代表长链非编码 RNA 与蛋白质的预测关联矩阵，$S_l^*$ 和 $S_p^*$ 为长链非编码 RNA 与蛋白质各自的相似性矩阵，此时 $A^*$ 中之前存在的空白值被填满，填充值即为长链非编码 RNA 与蛋白质潜在关联预测打分。预测矩阵及迭代终止条件表示如下：

$$W^*=\begin{bmatrix} S_l^* & A^* \\ A^{*\mathrm{T}} & S_p^* \end{bmatrix}$$

$$d1_{k+1}=\frac{\left\|X_{k+1}-X_k\right\|_F}{\left\|X_k\right\|_F}\le tol_1,\quad d2_{k+1}=\frac{\left|d1_{k+1}-d1_k\right|}{\max\left\{\left|d1_k\right|,1\right\}}\le tol_2$$

其中默认参数设置为： $\alpha=1$； $\beta=10$； $\gamma=1$； $tol_1=0.002$； $tol_2=0.00001$。模型流程图如图 1 所示。

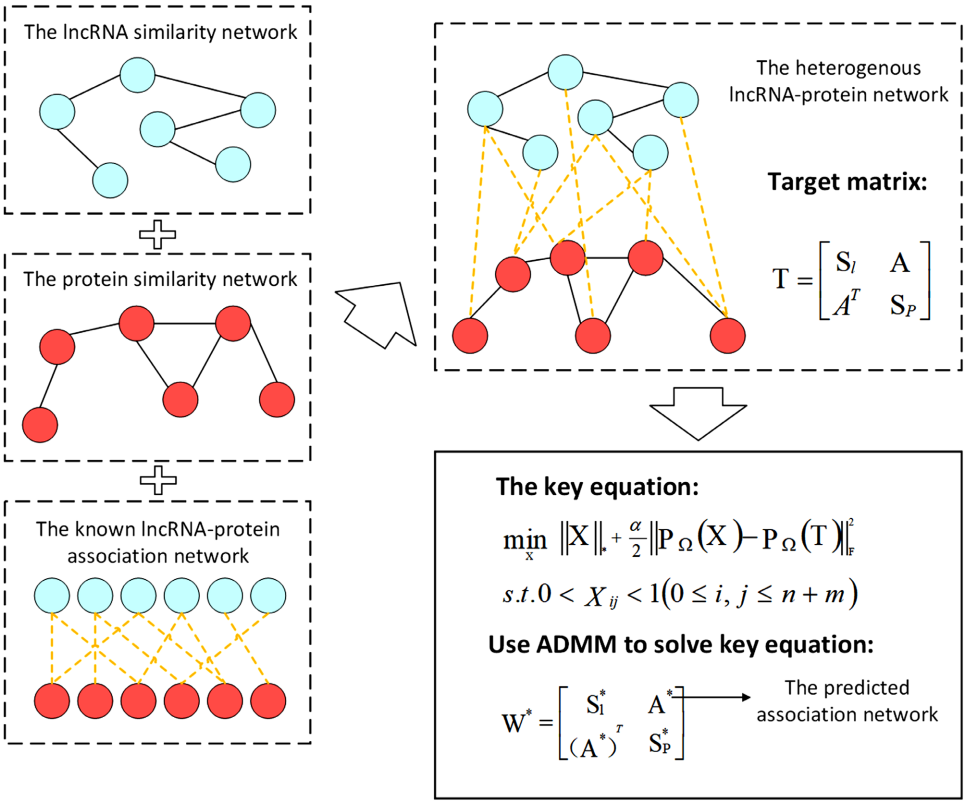


**Figure 1.** The flow chart of LPIMC
**图 1.** LPIMC 流程图

## 4. 性能评估

模型评估方面，我们采用 AUC (ROC 曲线下面积)作为评估指标。ROC 即受试者工作特性曲线，其横坐标为假阳性率(False Positive Rates, FPR)，纵坐标为真阳性率(False Positive Rates, TPR)，计算公式为：$\mathrm{FPR}=\mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$， $\mathrm{TPR}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$。对应的值由 TP，FP，TN，FN 计算而来。其中 TP 为真实为正类且预测同为正类的样本，FP 为真实为负类而预测为正类的样本。同理，TN 为真实为负类且预测同为负类的样本，FN 为真实为正类却被预测为负类的样本。

我们在整理筛选后的数据集上进行了 10 次 5 折交叉验证，在 5 折交叉验证中，8112 个确认的长链非编码 RNA-蛋白质关联被随机分为 5 组，每个集合被依次当作测试集，其他四个集合被合并为训练集，并使用在训练集上训练出的模型来预测测试集中的关联得分，最后通过计算得到模型的平均 AUC 值为 0.98 ± 0.01。为了证明模型的泛化能力，我们还额外从 lncRNome 数据库上采集了一个新数据集，经过筛选得到了 2729 对相互作用数据，涉及到 1184 个长链非编码 RNA 和 9 个蛋白质。经过 LPIMC 模型训练测试得到 5 折交叉验证平均 AUC 为 0.985 ± 0.003，模型表现性能较在 NPInter v2.0 上表现更为出色，据分析可能是因为在该数据库中已知长链非编码 RNA 与蛋白质关联占比更高所致。模型在两个数据集上的 ROC 曲线图如图 2 所示。

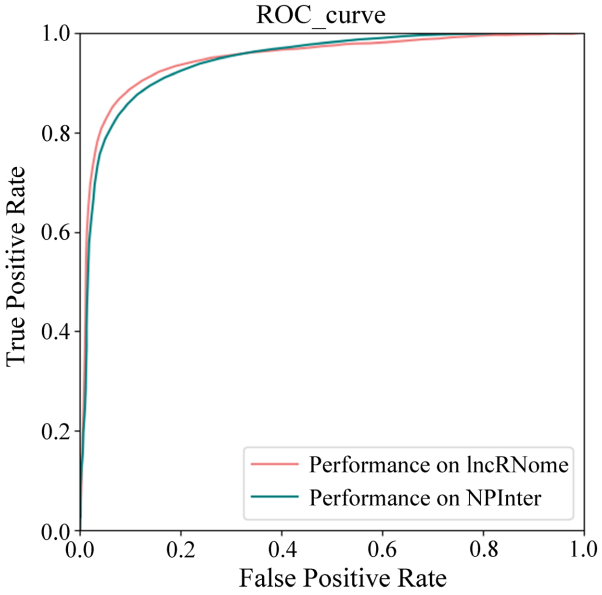


**Figure 2.** The ROC curve on two dataset
**图 2.** 两个数据集上的 ROC 曲线

## 5. 结论

本文中，我们提出了一种利用矩阵补全策略来进行长链非编码 RNA 与蛋白质相互作用关系预测的模型(LPIMC)来预测两者的潜在关联。实验数据表明，尽管所能利用的关联数据十分有限，模型仍取得了优良结果，且不过分依赖于长链非编码 RNA 与蛋白质本身的特殊特性，这些都表明了 LPIMC 模型可以扩展到类似的分类任务。长链非编码 RNA 与蛋白质的关联预测数据集属于较大规模数量级的数据集，而由于 LPIMC 使用了较少的计算资源，故它的时间效率也表现出色。同时，我们也对 LPIMC 模型的改良有了一些设想，例如将矩阵补全算法和当前大热的深度学习及图学习相结合，亦或在数据集层面下功夫，获取更高质量且更均衡的数据集。这些都能帮助提升模型预测性能。

## 参考文献

[1] Wapinski, O. and Chang, H.Y. (2011) Corrigendum: Long Noncoding RNAs and Human Disease. *Trends in Cell Biology*, **21**, 354-361. https://doi.org/10.1016/j.tcb.2011.04.001

[2] Pan, X., Fan, Y.X., Yan, J., *et al*. (2016) IPMiner: Hidden ncRNA-Protein Interaction Sequential Pattern Mining with Stacked Autoencoder for Accurate Computational Prediction. *BMC Genomics*, **17**, Article No. 582. https://doi.org/10.1186/s12864-016-2931-8

[3] Xiao, Y., Zhang, J. and Deng, L. (2017) Prediction of lncRNA-Protein Interactions Using HeteSim Scores Based on Heterogeneous Networks. *Scientific Reports*, 7, Article No. 3664. https://doi.org/10.1038/s41598-017-03986-1

[4] Hu, H., Zhang, L., Ai, H., *et al*. (2018) HLPI-Ensemble: Prediction of Human lncRNA-Protein Interactions Based on Ensemble Strategy. *RNA Biology*, **15**, 797-806. https://doi.org/10.1080/15476286.2018.1457935

[5] Zhan, Z.H., Jia, L.N., Zhou, Y., *et al*. (2019) BGFE: A Deep Learning Model for ncRNA-Protein Interaction Predictions Based on Improved Sequence Information. *International Journal of Molecular Sciences*, **20**, Article No. 978. https://doi.org/10.3390/ijms20040978

[6] Yi, H.C., You, Z.H., Wang, M.N., *et al*. (2020) RPI-SE: A Stacking Ensemble Learning Framework for ncRNA-Protein Interactions Prediction Using Sequence Information. *BMC Bioinformatics*, **21**, Article No. 60. https://doi.org/10.1186/s12859-020-3406-0

[7] Zhang, S.W., Zhang, X.X., Fan, X.N. and Li, W.N. (2020) LPI-CNNCP: Prediction of lncRNA-Protein Interactions by Using Convolutional Neural Network with the Copy-Padding Trick. *Analytical Biochemistry*, **601**, Article ID: 113767. https://doi.org/10.1016/j.ab.2020.113767

[8] Shen, Z.A., Luo, T., Zhou, Y.K., Yu, H. and Du, P.F. (2021) NPI-GNN: Predicting ncRNA-Protein Interactions with Deep Graph Neural Networks. *Briefings in Bioinformatics*, **22**, bbab051. https://doi.org/10.1093/bib/bbab051

[9] Li, Y., Sun, H., Feng, S., *et al*. (2021) Capsule-LPI: A LncRNA-Protein Interaction Predicting Tool Based on a Capsule Network. *BMC Bioinformatics*, **22**, Article No. 246. https://doi.org/10.1186/s12859-021-04171-y

[10] Jin, C., Shi, Z., Zhang, H. and Yin, Y. (2021) Predicting lncRNA-Protein Interactions Based on Graph Autoencoders and Collaborative Training. 2021 *IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*), Houston, 9-12 December 2021, 38-43. https://doi.org/10.1109/BIBM52615.2021.9669316

[11] Ge, M., Li, A. and Wang, M. (2016) A bipartite Network-Based Method for Prediction of Long Non-Coding RNA-Protein Interactions. *Genomics, Proteomics & Bioinformatics*, **14**, 62-71. https://doi.org/10.1016/j.gpb.2016.01.004

[12] Zhang, W., Qu, Q., Zhang, Y., *et al*. (2018) The Linear Neighborhood Propagation Method for Predicting Long Non-Coding RNA-Protein Interactions. *Neurocomputing*, **273**, 526-534. https://doi.org/10.1016/j.neucom.2017.07.065

[13] Zhao, Q., Zhang, Y., Hu, H., *et al*. (2018) IRWNRLPI: Integrating Random Walk and Neighborhood Regularized Logistic Matrix Factorization for lncRNA-Protein Interaction Prediction. *Frontiers in Genetics*, **9**, Article No. 239. https://doi.org/10.3389/fgene.2018.00239

[14] Zhang, T., Wang, M., Xi, J., *et al*. (2018) LPGNMF: Predicting Long Non-Coding RNA and Protein Interaction Using Graph Regularized Nonnegative Matrix Factorization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **17**, 189-197. https://doi.org/10.1109/TCBB.2018.2861009

[15] Yuan, J., Wu, W., Xie, C., *et al*. (2014) NPInter v2.0: An Updated Database of ncRNA Interactions. *Nucleic Acids Research*, **42**, D104-D108. https://doi.org/10.1093/nar/gkt1057

[16] Bu, D., Yu, K., Sun, S., *et al*. (2012) NONCODE v3.0: Integrative Annotation of Long Noncoding RNAs. *Nucleic Acids Research*, **40**, D210-D215. https://doi.org/10.1093/nar/gkr1175

[17] Apweiler, R., Bairoch, A., Wu, C.H., *et al*. (2004) UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Research*, **32**, D115-D119. https://doi.org/10.1093/nar/gkh131

[18] Bhartiya, D., Pal, K., Ghosh, S., *et al*. (2013) lncRNome: A Comprehensive Knowledgebase of Human Long Noncoding RNAs. *Database*, **2013**, bat034. https://doi.org/10.1093/database/bat034

[19] Chen, X., Yan, C.C., Zhang, X., You, Z.H., Huang, Y.A. and Yan, G.Y. (2016) HGIMDA: Heterogeneous Graph Inference for miRNA-Disease Association Prediction. *Oncotarget*, **7**, 65257-65269. https://doi.org/10.18632/oncotarget.11251

[20] Ramlatchan, A., Yang, M., Liu, Q., *et al*. (2018) A Survey of Matrix Completion Methods for Recommendation Systems. *Big Data Mining and Analytics*, **1**, 308-323. https://doi.org/10.26599/BDMA.2018.9020008

[21] Candes, E. and Recht, B. (2013) Simple Bounds for Recovering Low-Complexity Models. *Mathematical Programming*, **141**, 577-589. https://doi.org/10.1007/s10107-012-0540-0

[22] Boyd, S., Parikh, N., Chu, E., *et al*. (2011) Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends® in Machine Learning*, **3**, 1-122. https://doi.org/10.1561/2200000016