Molecular Omics

ROYAL SOCIETY OF CHEMISTRY

RESEARCH ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/d1mo00138h

HOPMCLDA: predicting lncRNA–disease associations based on high-order proximity and matrix completion

Guobo Xie,† Yinting Zhu, † Zhiyi Lin, *† Yuping Sun, Guosheng Gu, Weiming Wang and Hui Chen

In recent years, emerging evidence has shown that long noncoding RNAs (lncRNAs) have important roles in the biological processes of complex diseases. However, experiments to determine the associations between diseases and lncRNAs are time consuming and costly. Therefore, there is a need to develop effective computational methods for exploring potential lncRNA–disease associations. In this study, we present a computational prediction method based on high-order proximity and matrix completion to predict lncRNA–disease associations (HOPMCLDA). HOPMCLDA integrates explicit similarity and high-order proximity information on lncRNAs and diseases and constructs a heterogeneous disease–lncRNA network to utilize similarity information. Finally, nuclear norm regularization is carried out on the heterogeneous network for the recovery of a lncRNA–disease association matrix. By implementing leave-one-out cross validation (LOOCV) and five-fold cross validation (5-fold CV), we compare HOPMCLDA with five other methods. HOPMCLDA outperforms the other methods, with area under the receiver operating characteristic curve values of 0.8755 and 0.8353 ± 0.0045 using LOOCV and 5-fold CV, respectively. Furthermore, case studies of three human diseases (gastric cancer, osteosarcoma, and hepatocellular carcinoma) confirm the reliable predictive performance of HOPMCLDA.

Received 3rd May 2021,
Accepted 18th June 2021

DOI: 10.1039/d1mo00138h

rsc.li/molomics

1. Introduction

Long noncoding RNAs (lncRNAs) are noncoding RNAs that are more than 200 nucleotides long. Many studies have shown that lncRNAs contribute to fundamental biological processes of human cells, including translation,[1] splicing,[2] differentiation,[3] epigenetic regulation,[4] and immune response.[5] In recent years, lncRNAs have been found to be closely involved in various human cancers, including hepatocellular carcinoma (HCC),[6] gastric cancer (GC),[7] breast cancer,[8] Parkinson's disease,[9] and bladder cancer.[10] Therefore, developing computational methods for inferring potential disease–lncRNA associations could not only accelerate the diagnosis and treatment of diseases but also increase understanding of the mechanisms underlying human diseases at the lncRNA level.[11] In addition, computational methods can reduce the time and cost of studies and provide directions for biological experiments.

Over the years, a number of computational methods have been proposed to infer potential lncRNA–disease associations. Existing computational methods can be roughly classified into three categories: (1) machine learning, (2) network-based methods, and (3) matrix completion methods.

In the machine learning category, Chen *et al.*[12] presented a Laplacian regularized least squares method to predict novel associations by computing lncRNA and disease feature similarity, respectively, based on the assumption that similar diseases tend to be closely related to similar functional lncRNAs. By integrating modulators, genomes, modulators, transcriptomes, and known disease-associated lncRNAs, Zhao *et al.*[13] proposed a naive Bayesian classifier based on multivariate data. Yu *et al.*[14] presented a naive Bayesian classifier framework to predict lncRNA–disease associations, which integrated known biological associations including microRNA–lncRNA, lncRNA–disease, and microRNA–disease associations. Chen *et al.*[15] presented an ILDMSF based on a multi-similarity fusion strategy and support vector machine to identify lncRNA–disease associations. However, most of the machine learning methods depend heavily on known label samples, which are often difficult to obtain in practice. Moreover, when large numbers of unknown samples are used as negative samples, potential lncRNA–disease associations may be classified into negative samples, which may affect the prediction accuracy.

School of Computers, Guangdong University of Technology, Guangzhou, China. E-mail: xiegb@gdut.edu.cn, zhuyint1996@163.com, lzy291@gdut.edu.cn, syp@gdut.edu.cn, gsgu@gdut.edu.cn, wmwang1984@163.com, 541435267@qq.com
† These authors contributed equally to this work.

Approaches in the second category use network-based methods to predict potential lncRNA–disease associations. Sun *et al.*[16] used a random walk with restart, based on lncRNA function similarity, to explore potential lncRNA–disease associations in a heterogeneous network. Yu *et al.*[17] proposed a random walk technique called BRWLDA to infer potential lncRNA–disease associations by applying different random walks on two heterogeneous networks. Li *et al.*[18] developed a computational model to predict lncRNA–disease associations and implemented a local random walk method on a lncRNA–disease heterogeneous network. By integrating disease similarity and lncRNA similarity, Li *et al.*[19] constructed a lncRNA–disease association probability matrix and carried out a network consistency projection to predict lncRNA–disease associations. However, as insufficient lncRNA–disease associations have been verified by experiments, network-based methods often need to consider the prediction of individual nodes or rely on the introduction of additional biological information. Although the integration of such biological information can improve prediction performance, the additional interactions resulting from this information may represent noise that interferes with the prediction results.

The third type of approach uses a matrix completion method for predicting potential lncRNA–disease associations. Lu *et al.*[20] extracted feature vectors to develop a model based on inductive matrix completion and constructed a lncRNA–disease association matrix. Gao *et al.*[21] developed a computational model called DSCMF, which is based on traditional collaborative matrix decomposition and combines various biological similarity data to infer unknown lncRNA–disease associations. Liu *et al.*[22] used weighted graph-regulated collaborative matrix factorization to complete unknown elements in a lncRNA–disease association matrix. Another method based on geometric matrix completion, GMCLDA, proposed by Lu *et al.*,[23] utilized lncRNA and disease similarity information. Compared with the other two types of methods, a matrix completion method can capture the overall patterns of lncRNA–disease associations and reduce false positive rates. However, all existing matrix completion methods focus on the pairwise similarity of a direct relationship, such as lncRNA functional similarity and disease semantic similarity. They ignore high-order indirect lncRNA and disease information.

In recent years, studies have shown that high-order proximity between nodes is extremely important when exploring the underlying structure and properties of networks.[24,25] Inspired by this finding, we showed that lncRNA and disease high-order proximity were conducive to achieving effective prediction. Therefore, we developed a predictive method based on high-order proximity and matrix completion (HOPMCLDA) and used it to explore the associations between lncRNAs and diseases. First, we calculated the high-order proximity for similarity networks of diseases and lncRNAs. Second, lncRNA–lncRNA and disease–disease networks were reconstructed using singular value decomposition (SVD) to extract the main feature vectors. Finally, a heterogeneous lncRNA–disease network integrating disease–disease, disease–lncRNA, and lncRNA–lncRNA networks was constructed for the calculation of predicted scores with a matrix completion algorithm. To evaluate the prediction performance of HOPMCLDA, leave-one-out cross validation (LOOCV) and five-fold cross validation (5-fold CV) were applied to our collected dataset. HOPMCLDA achieved area under the receiving operator characteristic (ROC) curve (AUC) values of 0.8755 and 0.8353 ± 0.0045 under the LOOCV and 5-fold CV frameworks, respectively. Three case studies were used to demonstrate the predictive ability of HOPMCLDA.

2. Materials

2.1 Human disease–lncRNA associations

A lncRNA–disease association dataset was downloaded from the LncRNADisease database,[26] including 687 experimentally verified lncRNA–disease associations. We removed some duplicated lncRNAs and diseases, and nonhuman data. Finally, 540 unique experimentally verified lncRNA–disease associations between 115 unique lncRNAs and 178 unique diseases were obtained. To describe disease–lncRNA relationships, a disease–lncRNA adjacency matrix $DL \in R^{nd\times nl}$ was established, where the variables *nd* and *nl* represent the numbers of diseases and lncRNAs, respectively. If a disease d_i was verified to be associated with a lncRNA l_j, then DL(i,j) was set to 1; otherwise, DL(i,j) was 0. The matrix DL is defined as follows:

$$\mathrm{DL}(i,j)=\begin{cases}0 & \text{disease } d(i) \text{ has no association with lncRNA } l(j)\\ 1 & \text{disease } d(i) \text{ is associated with lncRNA } l(j).\end{cases}$$

2.2 Disease semantic similarity

Based on Medical Subject Headings (MeSH) downloaded from the U.S. National Library of Medicine https://www.nlm.nih.gov/mesh/meshhome.html (https://www.nlm.nih.gov/mesh/meshhome.html), we introduce a model based on a directed acyclic graph (DAG) to express the semantic similarity between diseases.[27] The DAG was used to describe a disease d, that is, DAG(d) = (d, $T(d)$, $E(d)$), where $T(d)$ is the node set and $E(d)$ is the corresponding set of edges pointed to by the parent node. Given a disease d, the semantic value (SV) of disease d in DAG(d) is defined as follows:

$$\mathrm{SV}(d)=\sum_{q\in T(d)} V_d(q)$$

$$V_d(q)=\begin{cases}1 & q=d\\ \max\{0.5\times C_d(q')|q'\in \text{children of } q\} & q\neq d.\end{cases}$$

The semantic similarity of two different diseases d_i and d_j is calculated based on the common part of their DAG. We use the

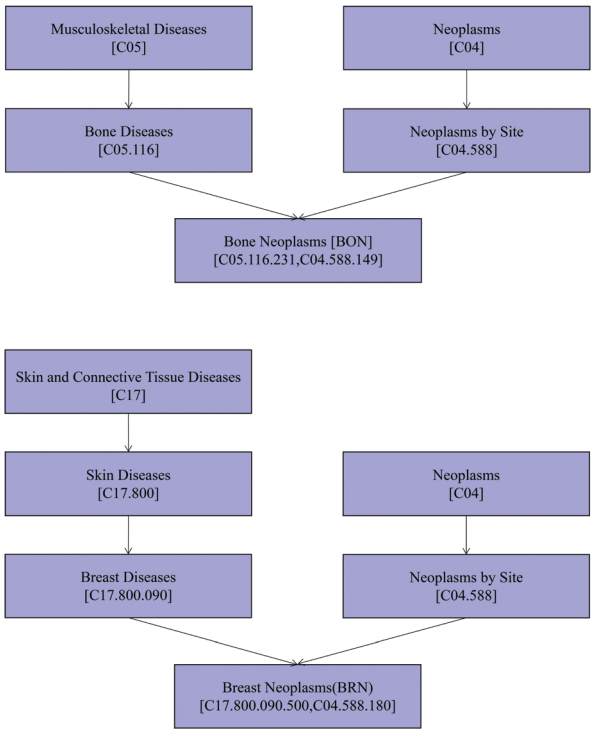

Fig. 1 DAGs for BON and BRN.

semantic similarity matrix DS(d_i,d_j) to represent the semantic similarity between disease d_i and disease d_j as follows:

$$\mathrm{DS}(d_i, d_j) = \frac{\sum_{t \in T(d_i) \cap T(d_j)} \left(V_{d_i}(t) + V_{d_j}(t)\right)}{\mathrm{SV}(d_i) + \mathrm{SV}(d_j)}.$$

We take bone neoplasms (BON) and breast neoplasms (BRN) as examples. The DAGs for the two diseases are shown in Fig. 1. In the MeSH, BON are expressed as C05.116.231 and C04.588.149, whereas BRN are expressed as C04.588.180 and C17.800.090.500. Calculating the semantic contributions of BON and its ancestor nodes to BON, SV(BON) is 1.0(BON) + 0.5(bone diseases) + 0.5(neoplasms by site) + 0.5 × 0.5(musculoskeletal diseases) + 0.5 × 0.5(neoplasms) = 2.5. Similarly, SV(BRN) is 1.0(BRN) + 0.5(breast diseases) + 0.5(neoplasms by site) + 0.5 × 0.5(skin diseases) + 0.5 × 0.5(neoplasms) + 0.5 × 0.5 × 0.5(skin and connective tissue diseases) = 2.625. In addition, we calculate the semantic contributions of the intersection nodes of BON and BRN ancestor nodes to BON and BRN. As shown in Fig. 1, the intersection set of BON and BRN ancestor nodes is T(BON) ∩ T(BRN) = {Neoplasms, Neoplasms by Site}. Therefore, $\sum_{t \in T(\mathrm{BON}) \cap T(\mathrm{BRN})} V_{\mathrm{BON}}(t) = 0.5(\text{neoplasms by site}) + 0.5 \times 0.5(\text{neoplasms}) = 0.75$, $\sum_{t \in T(\mathrm{BON}) \cap T(\mathrm{BRN})} V_{\mathrm{BRN}}(t) = 0.5(\text{neoplasms by site}) + 0.5 \times 0.5(\text{neoplasms}) = 0.75$.

Finally, the SV of disease between BON and BRN is calculated as follows:

$$\mathrm{DS}(\mathrm{BON}, \mathrm{BRN}) = \frac{0.75 + 0.75}{2.625 + 2.5} \approx 0.5854.$$

Obviously, the more shared nodes between two diseases in their DAGs, the higher the semantic similarity between the diseases.

2.3 LncRNA expression similarity

The lncRNA expression profile used in this paper was downloaded from ArrayExpress and generated using RNA sequencing technology.[28] The expression similarity of two lncRNAs was obtained by calculating the Spearman correlation coefficient between their expression profiles as described in previous studies.[29] We use LS to denote the lncRNA expression similarity matrix. Element LS(l_i,l_j) denotes the expression similarity between lncRNA l_i and lncRNA l_j, which is in the range 0–1.

3. HOPMCLDA

We present a novel method based on high-order proximity and matrix completion. First, we calculate the high-order proximity of the disease similarity matrix (DS) and lncRNA similarity matrix (LS) to integrate explicit lncRNA–lncRNA and disease–disease similarity information and implicit high-order proximity information. Then, a heterogeneous disease–lncRNA association matrix is constructed, which integrates disease–lncRNA association, lncRNA–lncRNA high-order proximity, and disease–disease high-order proximity networks. Finally, we use matrix completion to identify lncRNA–disease associations in the adjacency matrix of the heterogeneous network. The flowchart of HOPMCLDA is shown in Fig. 2.

3.1 Calculation of high-order proximity for lncRNA and disease

A fundamental hypothesis of lncRNA and disease prediction, based on observations from biological experiments, is that functionally similar lncRNAs are likely to be associated with phenotypically similar diseases, and *vice versa*. Therefore, lncRNA and disease correlation information is key to predicting lncRNA–disease associations. High-order proximity can describe implicit correlation information between matrix elements and is different from explicit correlation information. For example, if nodes V_i and V_j have many common neighbors and rich path information in a network, the probability of V_i reaching node V_j through a two-step random walk will be high, and thus the second-order proximity value of the two nodes will also be high.[30] The direct relationship between two nodes is not sufficient to reflect the similarity of nodes in prediction of lncRNA–disease associations. For example, there may be a direct or indirect relationship between lncRNAs associated with the same disease, and the complications of one disease can affect the incidence of other diseases. Therefore, we considered

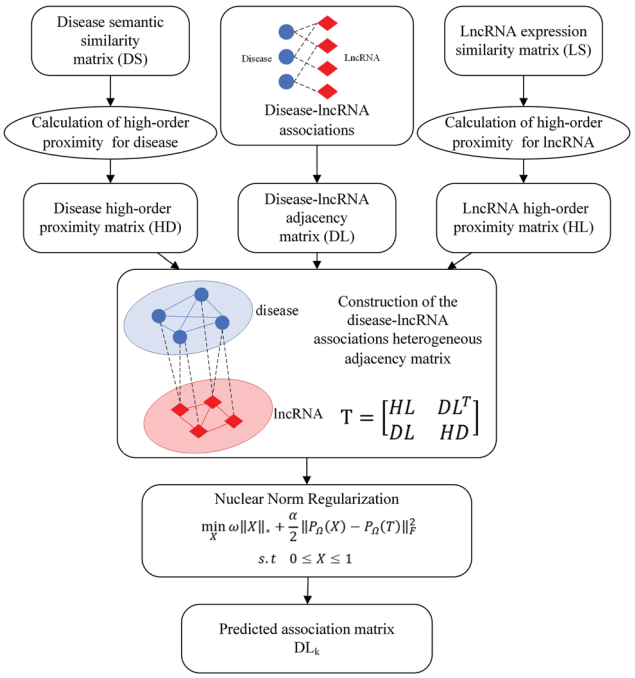

Fig. 2 Flowchart of the HOPMCLDA method.

high-order proximity to more effectively express lncRNA and disease correlation information.

The disease semantic similarity matrix DS represents the direct correlation information between diseases. We constructed a q-order proximity matrix HD based on disease semantic similarity matrix DS to preserve proximity information of different orders as follows:

$$\mathrm{HD} = \sum_{n=1}^{q} y^{n-1}\mathrm{DS}^{n}, \tag{1}$$

where DS^n is the n-order proximity of DS, and $y(y \geq 0)$ is the weight parameter, which controls the influence of different-order proximities.

However, noise may exist in matrix HD because of the high dimension of the matrix. Thus, the SVD technique was used to improve the data quality. The detailed formula of SVD is defined as:

$$\mathrm{HD} \rightarrow U\Sigma V^{T}, \tag{2}$$

where $U \in R^{nd\times nd}$ is a left singular vector matrix, $\Sigma \in R^{nd\times nd}$ is the diagonal matrix of singular values, and $V \in R^{nd\times nd}$ is a right singular vector matrix. We reconstructed a high-order proximity matrix HD by maintaining the top-k largest singular values:

$$\mathrm{HD} = U_k\Sigma_k V_k^{T}, \tag{3}$$

where Σ_k is the top-k singular value matrix, and U_k and V_k are the top-k singular values corresponding to the left and right singular vector matrix, respectively.

The lncRNA high-order proximity matrix HL is calculated in a similar way. It is worth noting that subtle information may be lost if the retained rank is too small.[31] Following the general setting in the study by Franceschini *et al.*,[32] we set the value of k to be half the number of singular values for each processed matrix. As a result, the k values in HD and HL were set to be 88 and 57, respectively.

3.2 Construction of the heterogeneous adjacency matrix of disease–lncRNA associations

We constructed a heterogeneous disease–lncRNA association network based on disease and lncRNA correlation information. This network integrates disease–lncRNA, lncRNA–lncRNA high-order proximity, and disease–disease high-order proximity networks. The disease–lncRNA association adjacency matrix of the heterogeneous network can be defined as:

$$T = \begin{pmatrix} \mathrm{HL} & \mathrm{DL}^{T} \\ \mathrm{DL} & \mathrm{HD} \end{pmatrix} \tag{4}$$

We aimed to complete the missing elements of DL as a predictive score for potential associations between diseases and lncRNAs.

3.3 Nuclear norm regularization

According to the hypothesis that functionally similar lncRNAs tend to be involved in similar diseases, potential factors that dominate the association likelihoods between lncRNAs and diseases are often highly correlated. Hence, the number of independent factors of interactions between lncRNAs and diseases is limited in heterogeneous disease–lncRNA association adjacency matrices, which form low-rank structures. We performed matrix completion to predict potential disease–lncRNA associations. Let Ω be the indicator matrix of observed entries of $G \in R^{m\times n}$, and let $P_\Omega(G)$: $R^{m\times n} \rightarrow R^{m\times n}$ be an orthogonal projection operator:

$$(P_\Omega(G))_{ij} = \begin{cases} G_{ij} & (i,j) \in \Omega \\ 0 & (i,j) \notin \Omega. \end{cases} \tag{5}$$

We used low-rank matrix completion to infer potential values.[33,34] The algorithm is as follows:

$$\begin{aligned} \min_{X} \quad & \omega \parallel X \parallel_{*} + \frac{\alpha}{2} \parallel P_\Omega(X) - P_\Omega(T) \parallel_2^F \\ \text{s.t.} \quad & 0 \leq X \leq 1, \end{aligned} \tag{6}$$

where $\|\cdot\|_2^F$ is the Frobenius norm; $\|\cdot\|_*$ is the trace norm defined as the sum of singular values; and ω and α are the non-negative parameters balancing the trace and nuclear norms. The constraint $0 \leq X \leq 1$ is used to ensure that the values of the recovered matrix elements are between 0 and 1.

The alternating direction method of multipliers (ADMM)[35] was used to transform (6) into a problem to be optimized. By introducing a variable matrix Y, (6) can be optimized as

$$\begin{aligned} \min_{X} \quad & \omega \parallel X \parallel_{*} + \frac{\alpha}{2} \parallel P_\Omega(Y) - P_\Omega(T) \parallel_2^F \\ \text{s.t.} \quad & X = Y, 0 \leq Y \leq 1. \end{aligned} \tag{7}$$

Accordingly, the augmented Lagrangian function corresponding to (7) is:

$$\begin{aligned}\mathscr{L}(Y,X,Z,\alpha,\beta) &= \omega \parallel X \parallel_* + \frac{\alpha}{2} \parallel P_\Omega(Y) - P_\Omega(T) \parallel_2^F \\ &\quad + \mathrm{Tr}(Z^T(X-Y)) + \frac{\beta}{2} \parallel X - Y \parallel_2^F,\end{aligned} \tag{8}$$

where $\beta > 0$ is the adaptive penalty parameter for the augmented term and Z is the Lagrangian multiplier.

HOPMCLDA requires k iterations to obtain Y_{k+1}, X_{k+1}, and Z_{k+1}.

To update Y_{k+1}, we fix X_k and Z_k to $\mathscr{L}(Y,X,Z,\alpha,\beta)$ for Y_{k+1}.

We denote by P_Ω^* the adjoint operator of P_Ω and $P_\Omega^* P_\Omega = I$. According to previous literature,[36] we can update Y_{k+1} as follows:

$$\begin{aligned}Y_{k+1} &= \arg\min_{0\le Y\le 1} \mathscr{L}(Y, X_k, Z_k, \alpha, \beta, \omega) \\ &= \arg\min_{0\le Y\le 1} \frac{\alpha}{2} \parallel P_\Omega(Y) - P_\Omega(T) \parallel_2^F \\ &\quad + \mathrm{Tr}(Z^T(X-Y)) + \frac{\beta}{2} \parallel X - Y \parallel_2^F \\ &= \left(\frac{1}{\beta} Z_k + \frac{\alpha}{\beta} P_\Omega^* P_\Omega(T) + X_k\right) - \\ &\quad \frac{\alpha}{\alpha+\beta} P_\Omega\left(\frac{1}{\beta} Z_k + \frac{\alpha}{\beta} P_\Omega^* P_\Omega(T) + X_k\right).\end{aligned} \tag{9}$$

To update X_{k+1}, we fix Y_k and Z_k.

Based on the singular value thresholding algorithm,[37] X_{k+1} is represented as follows:

$$\begin{aligned}X_{k+1} &= \arg\min_X \mathscr{L}(Y_{k+1}, X, Z_k, \alpha, \beta, \omega) \\ &= \arg\min \omega \parallel X \parallel_* + \mathrm{Tr}(Z_k^T(X - Y_{k+1})) \\ &\quad + \frac{\beta}{2} \parallel X - Y_{k+1} \parallel_2^F \\ &= \arg\min \omega \parallel X \parallel_* + \frac{\beta}{2} \parallel X - (Y_{k+1} - \frac{1}{\beta} Z_k) \parallel_2^F \\ &= O_{\frac{\omega}{\beta}}\left(Y_{k+1} - \frac{1}{\beta} Z_k\right).\end{aligned} \tag{10}$$

Here, $O_\tau(\cdot)$ is the soft-thresholding operator defined as $O_\tau(M) = \sum_{\tau}^{\sigma_d > \tau} (\sigma_d - \tau) u_d v_d^T$, where σ_d is the dth singular value of matrix M larger than the shrinkage threshold τ, and u_d and v_d are the left and right singular vectors corresponding to σ_d, respectively.

To update Z_{k+1}: Finally, Lagrangian multiplier Z_{k+1} is calculated as follows:

$$Z_{k+1} = Z_k + \varphi\beta(X_{k+1} - Y_{k+1}), \tag{11}$$

where φ is a positive step size that can be set to 1 as described in a previous study.[38]

Overall, Y_{k+1}, X_{k+1}, and Z_{k+1} are generated by repeatedly using eqn (9)–(11) until the stopping criteria are met. The main steps of the implementation are outlined in Algorithm 1. After the matrix completion algorithm has been applied to the heterogeneous adjacency disease–lncRNA matrix, an updated disease–lncRNA association score matrix DL_k can be obtained. The predicted scores in DL_k are close to 1, indicating that the disease is likely to be associated with lncRNA.

Algorithm 1 Nuclear norm regularization.

Input: Heterogeneous disease–lncRNA association adjacency matrix T;
Output: Predicted association matrix DL_k;
1: Initialize X_1, Y_1, $Z_1 = P_\Omega(T)$, *maxiter* = 300, $\omega = 0.9$, $\alpha = 0.5$, $\beta = 10$, $\varepsilon = 1 \times 10^{-3}$;
2: **for** $k = 1 \rightarrow$ *maxiter* **do**
3: $Y_{k+1} = \arg\min_{0\le Y\le 1} \mathscr{L}(Y, X_k, Z_k, \alpha, \beta, \omega)$;
4: $X_{k+1} = \arg\min_X \mathscr{L}(Y_{k+1}, X, Z_k, \alpha, \beta, \omega)$;
5: $Z_{k+1} = Z_k + \varphi\beta(X_{k+1} - Y_{k+1})$;
6: **if** $\frac{\parallel X_{k+1} - X_k \parallel_F}{\parallel X_k \parallel_F} \le \varepsilon$ **then**
7: **break**;
8: **end if**
9: **end for**
10: $\begin{bmatrix} HL_k & DL_k^T \\ DL_k & HD_k \end{bmatrix} \leftarrow Y_k$;
11: **return** DL_k

4. Results and discussion

4.1 Evaluation metrics

To evaluate the performance of HOPMCLDA, LOOCV and 5-fold CV were conducted on the obtained dataset. In each trial of LOOCV, each known lncRNA–disease association was treated as the testing set, whereas the remaining known lncRNA–disease associations were used as training samples. The 5-fold CV experiment was similar to the LOOCV one. All known lncRNA–disease associations were randomly divided into five exclusive subsets. For each experiment, each subset was utilized as a test sample each time, and the remaining subsets were used as the training set. As there was only a small number of positive samples in the lncRNA–disease dataset used in this study, and AUC is not sensitive to skewed class distributions,[39] we used AUC as an indicator to evaluate the HOPMCLDA method and demonstrate its superiority. The ROC curve was drawn and used to measure the predictive accuracy of HOPMCLDA.

4.2 Effects of parameters

In this section, we test the impact of the five parameters on HOPMCLDA by implementing the LOOCV framework.

4.2.1 Effects of parameter q and parameter y on the experimental results. Given the parameters $\omega = 0.9$, $\alpha = 0.5$, and $\beta = 10$, we compared the AUC results achieved with HOPMCLDA for different values of parameter q (denoted q-HOPMCLDA). Then, parameter y was adjusted in the range 0.1–0.9. The effects of parameters q and y are shown in Fig. 3. When the value of parameter q was high, the prediction accuracy decreased. The introduction of high-order proximity

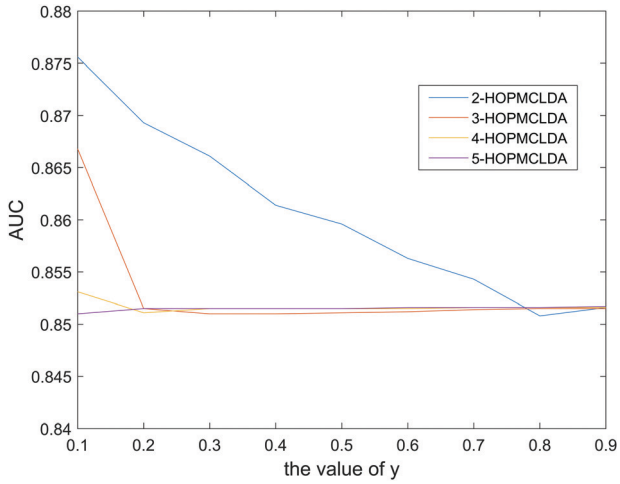

Fig. 3 Influence of parameters q and y on the AUC values. q-HOPMCLDA indicates the value of parameter q at which a given AUC was achieved. y indicates the influence of different orders of proximity.

information may have been diluted by abundant lower-order proximity information. With increasing y, the results decreased and became stable. HOPMCLDA showed good performance when $q = 2$ and $y = 0.1$.

4.2.2 Effects of parameter α on the performance. Parameter α was adjusted in the range 0.1–1, with the other parameters fixed ($q = 2$, $y = 0.1$, $\omega = 0.9$, and $\beta = 10$). The influence of parameter α on the performance is illustrated in Fig. 4. As parameter α increased, HOPMCLDA achieved a high AUC in the range $\alpha \in [0.1, 0.5)$ from the results. However, the AUC decreased slightly in the range $\alpha \in (0.5, 0.9]$. The AUC reached its maximum value when $\alpha = 0.5$ for HOPMCLDA.

4.2.3 Effects of parameter ω on the performance. The influence of parameter ω on the experimental results is shown in Fig. 5 under the following conditions: $\alpha = 0.5$, $q = 2$, $y = 0.1$, and $\beta = 10$. We measured the AUC of HOPMCLDA when $0.1 \leq \omega \leq 1$. With increasing ω, AUC increased gradually and became stable. When $\omega = 0.9$, HOPMCLDA achieved its maximum AUC value.

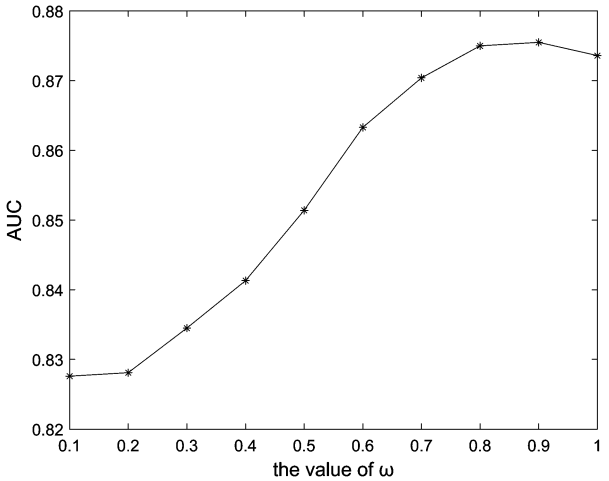

Fig. 5 Impact of parameter ω on the AUC values.

4.2.4 Effects of parameter β on the performance. Finally, we measured the AUC values of β between 1 and 100 when $q = 2$, $y = 0.1$, $\omega = 0.9$, and $\alpha = 0.5$. The results are illustrated in Fig. 6, demonstrating that the AUC reached its maximum value when $\beta = 10$.

Based on these results we chose $q = 2$, $y = 0.1$, $\alpha = 0.5$, $\omega = 0.9$, and $\beta = 10$ as default values.

4.2.5 Effects of calculating high-order proximity for diseases and lncRNAs. The high-order proximity for diseases and lncRNAs was calculated to improve the accuracy of disease and lncRNA similarity measurements. To prove the effects of high-order proximity, HOPMCLDA was implemented on a heterogeneous network for the following scenarios in LOOCV: (1) HOPMCLDA is implemented on the heterogeneous network without calculating high-order proximity for the disease and lncRNA similarity matrices; (2) only high-order proximity for the disease similarity matrix is calculated; (3) only high-order proximity for the lncRNA similarity matrix is calculated; and (4)

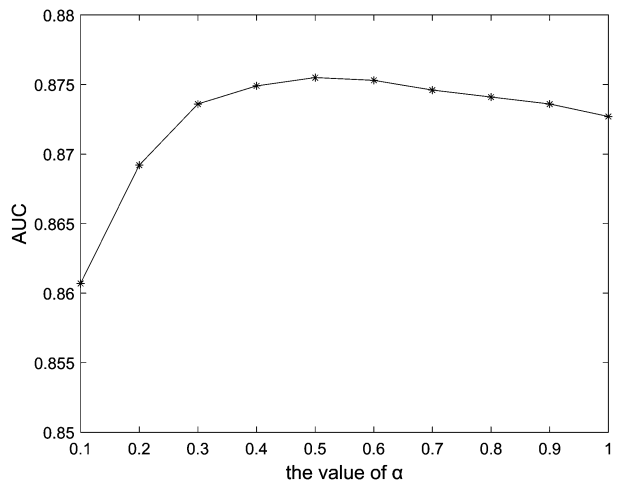

Fig. 4 Impact of parameter α on the AUC values.

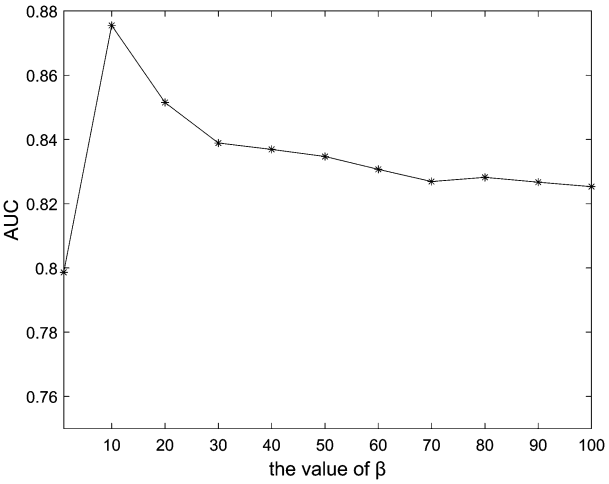

Fig. 6 Impact of parameter β on the AUC values.

Table 1 Effects of calculating high-order proximity on the prediction performance with LOOCV

Matrices included in the high-order proximity calculation	AUC value
None	0.8528
Disease similarity matrix only	0.8675
lncRNA similarity matrix only	0.8637
Disease similarity matrix and lncRNA similarity matrix	0.8755

high-order proximity is calculated for the disease and lncRNA similarity matrices. The AUC values are shown in Table 1. After calculation of high-order proximity for both similarity matrices, HOPMCLDA achieved a reliable AUC value of 0.8755, a better result than those obtained for the other three scenarios. Therefore, calculation of high-order proximity for lncRNAs and diseases is an effective way to improve the performance of HOPMCLDA.

4.3 Performance comparison

To evaluate its prediction accuracy, HOPMCLDA was compared with five previous methods, GMCLDA,[23] SIMCLDA,[20] DSCMF,[21] BRWLDA,[17] and RWRlncD,[16] on the same dataset without independent dataset validation. All comparison methods used the parameters provided by the authors. As shown in Fig. 7, in the LOOCV framework, the AUC for HOPMCLDA was 0.8755, larger than the values obtained with the other computational methods (GMCLDA 0.8501; SIMCLDA, 0.8237; DSCMF, 0.8176; BRWLDA, 0.7969; and RWRlncD, 0.6540), indicating that HOPMCLDA performed better than the other methods.

The 5-fold CV framework was used to verify the predictive performance of HOPMCLDA. To ensure the fairness of the predicted results, we repeated 5-fold CV 100 times and computed the mean of the AUC values. As shown in Fig. 8, HOPMCLDA achieved an average AUC of 0.8353 ± 0.0045, far outweighing the AUC values of 0.7894 ± 0.0040, 0.7839 ± 0.0045, 0.7734 ± 0.0045, 0.7659 ± 0.0045, and 0.6179 ± 0.0045 for GMCLDA, DSCMF, SIMCLDA, BRWLDA, and RWRlncD, respectively. These results indicate that HOPMCLDA is more efficient than the other methods in the 5-fold CV framework and is thus more suitable for the prediction of lncRNA–disease associations.

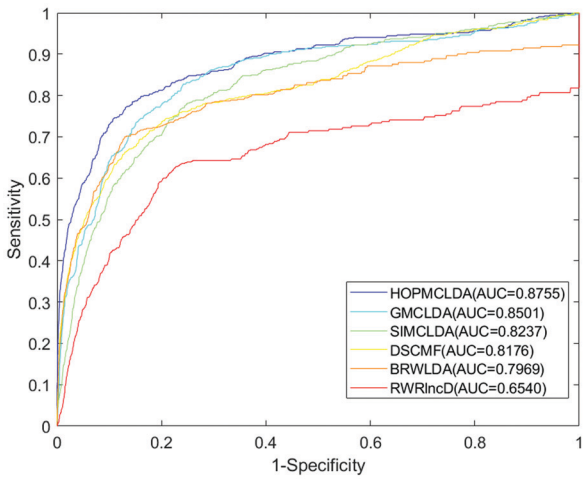

Fig. 7 AUCs achieved by HOPMCLDA, GMCLDA, DSCMF, SIMCLDA, BRWLDA, and RWRlncD based on LOOCV.

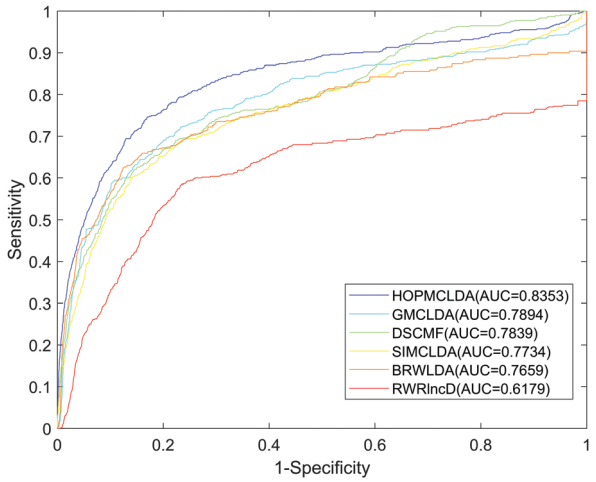

Fig. 8 AUCs achieved by HOPMCLDA, GMCLDA, DSCMF, SIMCLDA, BRWLDA, and RWRlncD based on 5-fold CV.

4.4 Case studies

To evaluate the predictive ability of HOPMCLDA, the known lncRNA–disease associations were used to form training datasets. Then, the predictive score of each unknown lncRNA–disease pair was calculated and ranked. GC, osteosarcoma and HCC were used for case studies. The top 10 lncRNAs in these three cancers were verified using third-party databases (Lnc2Cancer[40] and MNDR[41]). Approximately 100%, 90%, and 90% of the predicted lncRNAs are associated with the three cancer types, respectively.

The prediction results for GC are shown in Table 2. GC is among the most common malignant tumors worldwide and has an extremely high mortality rate. The incidence of GC is influenced by a number of factors, including smoking, high salt intake, and high alcohol intake.[42] Previous studies have identified candidate lncRNAs associated with GC. For example, Zhang *et al.* found that XIST was negatively correlated with GC cells and that silencing it could inhibit the growth of GC cells.[43] By binding directly to Mir-145-5p, TUG1 plays a crucial

Table 2 Top 10 potential lncRNAs associated with GC as predicted by HOPMCLDA

Rank	lncRNA name	Evidence (PubMed)
1	XIST	29053187, 29212249
2	TUG1	27983921, 29719612
3	NEAT1	30024601, 31486491
4	KCNQ1OT1	31915311
5	HIF1A-AS1	26722487
6	HOTTIP	31908497
7	BCYRN1	27144338
8	PANDAR	31308753, 26898439
9	SNHG16	30854107, 31561329
10	ZFAS1	30999814, 28285404

Table 3 Top 10 potential lncRNAs associated with osteosarcoma as predicted by HOPMCLDA

Rank	lncRNA name	Evidence (PubMed)
1	HOTAIR	30367466
2	MEG3	32016959
3	H19	29568924
4	GAS5	31337976
5	UCA1	30481751
6	PVT1	32021563
7	SPRY4-IT1	31746422
8	MINA	Unknown
9	CCAT2	31322006
10	BANCR	25893737

part in suppressing the expression of Mir-145-5p, which inhibits the differentiation, invasion, and growth of GC cells.[44] Given that expression of mir-497-5p/PIK3R1, which inhibits the growth of GC cells, can be regulated by NEAT1, NEAT1 has been proposed as a potential therapeutic target for GC.[45]

The prediction results for osteosarcoma are shown in Table 3. Osteosarcoma is a malignant tumor that primarily and predominantly initiates in the metaphysis of long bones and has relatively high prevalence among children and young adults. Osteosarcoma not only affects long bones but also bones in other parts of the body.[46] Relationships between lncRNAs and osteosarcoma have been demonstrated in previous studies. For example, silencing HOTAIR was shown to significantly repress the major biological processes of osteosarcoma cells, including growth, migration, invasion, and apoptosis.[47] Overexpression of lncRNA MEG3 could inhibit proliferation and promote apoptosis of osteosarcoma MG-63 cells by suppressing the Notch signaling pathway.[48] The present study demonstrates that the suppression of nuclear factor-κB by knockdown of lncRNA H19 can inhibit invasion and migration of human osteosarcoma cells.[49]

The prediction results for HCC are shown in Table 4. HCC is among the most prevalent diseases worldwide and has a high incidence rate.[50] Despite the development of effective antiviral therapeutics, the incidence of HCC has continued to rise. The main risk factor for HCC is nonalcoholic fatty liver disease, which represents an epidemic.[51] Emerging evidence indicates that lncRNAs are involved in the development of HCC cells. For example, NEAT1 is a carcinogenic driver of HCC and is involved in its biological processes.[52] TUSC7 regulates the cell cycle of HCC cells, and its low expression is not conducive to the survival of patients with HCC.[53] By weakening the promoter activity of NDRG1 and decreasing its expression, CCAT2 can be downregulated, thereby inhibiting the proliferation and metastatic behavior of HCC cells *in vitro* and *in vivo*.[54]

Table 4 Top 10 potential lncRNAs associated with HCC as predicted by HOPMCLDA

Rank	lncRNA name	Evidence (PubMed)
1	NEAT1	32168951
2	TUSC7	27002617
3	CCAT2	30922920, 28744394
4	SPRY4-IT1	31943198, 27899259
5	BANCR	29085476, 29725471
6	LINC00261	29278875
7	CCAT1	30773676
8	PCA3	Unknown
9	CDKN2B-AS1	30165194
10	HNF1A-AS1	29466992, 28292020

5. Discussion

Many studies have found that lncRNAs are related to the pathogenesis of diseases. Therefore, developing computational methods to uncover unknown lncRNA–disease associations is not only beneficial to understanding the main functions of lncRNAs in the pathological and molecular changes that occur in human diseases but will also help with disease discovery, pathological classification, and treatment and prevention of complex diseases. In this work, we propose a novel computational method, HOPMCLDA, based on high-order proximity and matrix completion. HOPMCLDA uses a high-order proximity function to integrate correlation information of different orders. In addition, HOPMCLDA uses a matrix completion method to fully exploit lncRNA and disease correlation information and avoid the problem of a lack of negative samples. HOPMCLDA ranked first in comparison with other methods, with AUC values for LOOCV and 5-fold CV of 0.8755 and 0.8353 ± 0.0045, respectively. Moreover, HOPMCLDA was applied in three case studies (osteosarcoma, GC, and HCC) and basically confirmed the top 10 lncRNAs associated with these three common human cancers.

However, the HOPMCLDA model has certain limitations. For example, how to effectively set high-order weight parameters to integrate different order proximity matrices in different datasets is a problem that requires further study.

Data availability

The source codes and datasets used in this work can be found at: https://github.com/yeyetuowoyi/HOPMCLDA.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62002070, 61802072), the Science and Technology Planning Project of Guangzhou City (2019020 20006, 201902020012, 202102021236), and the Opening Project of Guangdong Province Key Laboratory of Computational Science at Sun Yat-sen University (2021013).

References

1 W.-X. Peng, P. Koirala and Y.-Y. Mo, *Oncogene*, 2017, **36**, 5661–5667.

2 N. Romero-Barrios, M. F. Legascue, M. Benhamed, F. Ariel and M. Crespi, *Nucleic Acids Res.*, 2018, **46**, 2169–2184.
3 C. Ju, R. Liu, Y.-W. Zhang, Y. Zhang, R. Zhou, J. Sun, X.-B. Lv and Z. Zhang, *Biomed. Pharmacother.*, 2019, **115**, 108912.
4 C. Tam, J. H. Wong, S. K. W. Tsui, T. Zuo, T. F. Chan and T. B. Ng, *Appl. Microbiol. Biotechnol.*, 2019, **103**, 4649–4677.
5 M. R. Hadjicharalambous and M. A. Lindsay, *Non-coding RNA*, 2019, **5**, 34.
6 C.-M. Wong, F. H.-C. Tsang and I. O.-L. Ng, *Nat. Rev. Gastroenterol. Hepatol.*, 2018, **15**, 137.
7 S. Ghafouri-Fard and M. Taheri, *Exp. Mol. Pathol.*, 2020, **113**, 104365.
8 D. Tomar, A. S. Yadav, D. Kumar, G. Bhadauriya and G. C. Kundu, *Biochim. Biophys. Acta, Gene Regul. Mech.*, 2020, **1863**, 194378.
9 J. A. Stamford, P. N. Schmidt and K. E. Friedl, *IEEE J. Biomed. Health Inform.*, 2015, **19**, 1862–1872.
10 Y. Cao, T. Tian, W. Li, H. Xu, C. Zhan, X. Wu, C. Wang, X. Wu, W. Wu and S. Zheng, *et al.*, *Clin. Chim. Acta*, 2020, **503**, 113–121.
11 Z. Cui, J.-X. Liu, Y.-L. Gao, R. Zhu and S.-S. Yuan, *IEEE J. Biomed. Health Inform.*, 2019, **24**, 1519–1527.
12 X. Chen and G.-Y. Yan, *Bioinformatics*, 2013, **29**, 2617–2624.
13 T. Zhao, J. Xu, L. Liu, J. Bai, C. Xu, Y. Xiao, X. Li and L. Zhang, *Mol. BioSyst.*, 2015, **11**, 126–136.
14 J. Yu, P. Ping, L. Wang, L. Kuang, X. Li and Z. Wu, *Genes*, 2018, **9**, 345.
15 Q. Chen, D. Lai, W. Lan, X. Wu, B. Chen, Y.-P. P. Chen and J. Wang, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2021, **18**, 1106–1112.
16 J. Sun, H. Shi, Z. Wang, C. Zhang, L. Liu, L. Wang, W. He, D. Hao, S. Liu and M. Zhou, *Mol. BioSyst.*, 2014, **10**, 2074–2081.
17 G. Yu, G. Fu, C. Lu, Y. Ren and J. Wang, *Oncotarget*, 2017, **8**, 60429.
18 J. Li, H. Zhao, Z. Xuan, J. Yu, X. Feng, B. Liao and L. Wang, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2021, **18**, 1049–1059.
19 G. Li, J. Luo, C. Liang, Q. Xiao, P. Ding and Y. Zhang, *IEEE Access*, 2019, **7**, 58849–58856.
20 C. Lu, M. Yang, F. Luo, F.-X. Wu, M. Li, Y. Pan, Y. Li and J. Wang, *Bioinformatics*, 2018, **34**, 3357–3364.
21 M.-M. Gao, Z. Cui, Y.-L. Gao, F. Li and J.-X. Liu, International Conference on Intelligent Computing, 2019, pp. 318–326.
22 J.-X. Liu, Z. Cui, Y.-L. Gao and X.-Z. Kong, *IEEE J. Biomed. Health Inform.*, 2021, **25**, 257–265.
23 C. Lu, M. Yang, M. Li, Y. Li, F.-X. Wu and J. Wang, *IEEE J. Biomed. Health Inform.*, 2019, **24**, 2420–2429.
24 P. Cui, X. Wang, J. Pei and W. Zhu, *IEEE Trans. Knowledge Data Eng.*, 2018, **31**, 833–852.
25 Z. Zhang, P. Cui, X. Wang, J. Pei, X. Yao and W. Zhu, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2778–2786.
26 G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan and Q. Cui, *Nucleic Acids Res.*, 2012, **41**, D983–D986.
27 Y. Zhao, X. Chen and J. Yin, *Front. Genet.*, 2018, **9**, 324.
28 X. Qian, Y. Ba, Q. Zhuang and G. Zhong, *OMICS: J. Integr. Biol.*, 2014, **18**, 98–110.
29 Y.-A. Huang, X. Chen, Z.-H. You, D.-S. Huang and K. C. Chan, *Oncotarget*, 2016, **7**, 25902.
30 Y. Wu, Y. Bian and X. Zhang, *Proceedings of the VLDB Endowment*, 2016, **10**, 13–24.
31 Z. Hou, *Pattern Recogn.*, 2003, **36**, 1747–1763.
32 A. Franceschini, J. Lin, C. von Mering and L. J. Jensen, *Bioinformatics*, 2016, **32**, 1085–1087.
33 A. Ramlatchan, M. Yang, Q. Liu, M. Li, J. Wang and Y. Li, *Big Data Mining and Analytics*, 2018, **1**, 308–323.
34 M. Yang, H. Luo, Y. Li and J. Wang, *Bioinformatics*, 2019, **35**, i455–i463.
35 S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, *et al.*, *Foundations Trends Mach. Learn.*, 2011, **3**, 1–122.
36 J. Yang and X. Yuan, *Math. Comput.*, 2013, **82**, 301–329.
37 J.-F. Cai, E. J. Candès and Z. Shen, *SIAM J. Optimization*, 2010, **20**, 1956–1982.
38 Y. Hu, D. Zhang, J. Ye, X. Li and X. He, *IEEE Trans. Pattern Anal. Mach. Intelligence*, 2012, **35**, 2117–2130.
39 A. Ezzat, M. Wu, X.-L. Li and C.-K. Kwoh, *BMC Bioinf.*, 2016, **17**, 267–276.
40 Y. Gao, P. Wang, Y. Wang, X. Ma, H. Zhi, D. Zhou, X. Li, Y. Fang, W. Shen and Y. Xu, *et al.*, *Nucleic Acids Res.*, 2019, **47**, D1028–D1033.
41 L. Ning, T. Cui, B. Zheng, N. Wang, J. Luo, B. Yang, M. Du, J. Cheng, Y. Dou and D. Wang, *Nucleic Acids Res.*, 2020, **49**, D160–D164.
42 E. C. Smyth, M. Nilsson, H. I. Grabsch, N. C. van Grieken and F. Lordick, *Lancet*, 2020, **396**, 635–648.
43 Y. Li, Q. Zhang and X. Tang, *Minerva Med.*, 2019, **110**, 270–272.
44 K. Ren, Z. Li, Y. Li, W. Zhang and X. Han, *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, 2017, **25**, 789–798.
45 T. Xia, J. Chen, K. Wu, J. Zhang and Q. Yan, *Eur. Rev. Med. Pharmacol. Sci.*, 2019, **23**, 6914–6926.
46 H. K. Brown, M. Tellez-Gabriel and D. Heymann, *Cancer Lett.*, 2017, **386**, 189–195.
47 B. Wang, X.-L. Qu, J. Liu, J. Lu and Z.-Y. Zhou, *J. Cell. Physiol.*, 2019, **234**, 6173–6181.
48 L. Chen, J. Wang, J. Li, X. Zhao and L. Tian, *Eur. Rev. Med. Pharmacol. Sci.*, 2020, **24**, 581–590.
49 J. Zhao and S.-T. Ma, *Mol. Med. Rep.*, 2018, **17**, 7388–7394.
50 A. Forner, M. Reig and J. Bruix, *The Lancet*, 2018, **391**, 1301–1314.
51 M. P. Johnston and S. I. Khakoo, *World J. Gastroenterol.*, 2019, **25**, 2977.
52 S. Koyama, H. Tsuchiya, M. Amisaki, H. Sakaguchi, S. Honjo, Y. Fujiwara and G. Shiota, *Int. J. Mol. Sci.*, 2020, **21**, 1927.
53 Y. Wang, Z. Liu, B. Yao, C. Dou, M. Xu, Y. Xue, L. Ding, Y. Jia, H. Zhang and Q. Li, *et al.*, *Tumor Biol.*, 2016, **37**, 11429–11441.
54 Y. Liu, D. Wang, Y. Li, S. Yan, H. Dang, H. Yue, J. Ling, F. Chen, Y. Zhao and L. Gou, *et al.*, *Exp. Cell Res.*, 2019, **379**, 19–29.