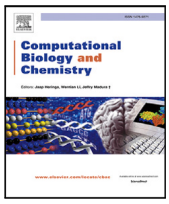
Computational Biology and Chemistry 115 (2025) 108353

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac



Research Article

Predicting distant metastatic sites of cancer using perturbed correlations of miRNAs with competing endogenous RNAs

Myeonghoon Cho [a], Byungkyu Park [b], Kyungsook Han [a,*]

[a] Department of Computer Engineering, Inha University, 100 Inha-ro, Michuhol-gu, Incheon, 22212, Republic of Korea
[b] Research and Development Center, Hancom Carelink Incorporated, 49 Daewangpangyo-ro 644beon-gil, Bundang-gu, Seongnam, 13493, Gyeonggi-do, Republic of Korea

ARTICLE INFO

Keywords:
Metastatic site
Distant metastasis
Competing endogenous RNA
Prediction model

ABSTRACT

Cancer metastasis is the dissemination of tumor cells from the primary tumor site to other parts of the body via the lymph system or bloodstream. Metastasis is the leading cause of cancer associated death. Despite the significant advances in cancer research and treatment over the past decades, metastasis is not fully understood and difficult to predict in advance. In particular, distant metastasis is more difficult to predict than lymph node metastasis, which is the spread of cancer cells to nearby lymph nodes. Distant metastatic sites is even more difficult to predict than the occurrence of distant metastasis because the problem of predicting distant metastatic sites is a multi-class and multi-label classification problem; there are more than two classes for distant metastatic sites (bone, liver, lung, and other organs), and a single sample can have multiple labels for multiple metastatic sites. This paper presents a new method for predicting distant metastatic sites based on correlation changes of miRNAs with competing endogenous RNAs (ceRNAs) in individual cancer patients. Testing the method on independent datasets of several cancer types demonstrated a high prediction performance. In comparison of our method with other state of the art methods, our method showed a much better and more stable performance than the others. Our method can be used as useful aids in determining treatment options by predicting if and where metastasis will occur in cancer patients at early stages.

1. Introduction

Despite significant advances in cancer research and treatment over the past decades, cancer remains one of the leading causes of death worldwide (Bray et al., 2021). In particular, 90% of cancer-related deaths are attributed to cancer metastasis (Ganesh and Massague, 2021). Metastasis is the spread of cancer cells to other organs or tissues of the body typically through the lymphatic system or bloodstream. Lymph node metastasis and distant metastasis are two major types of cancer metastasis. Lymph node metastasis is the spread of tumor cells of the primary site to nearby lymph nodes, and distant metastasis is the spread of primary tumor cells to distant parts of the body (Edge and Compton, 2010).

Recently, several learning methods have been used to predict lymph node metastasis in cancer. For instance, Zhang et al. (2021) used a support vector machine model to predict lymph node metastasis based on differentially expressed mRNAs and non-coding RNAs in cancer. Zhao and Yu (2018) used a random forest model to predict lymph node metastasis in stomach cancer based on DNA methylation data. Zeng et al. (2022) predicted lymph node metastasis in early gastric cancer based on several features such as tumor size, morphology, and degree of differentiation.

In contrast to lymph node metastasis, only a few attempts have been made to predict distant metastasis. There are several reasons for this. First, predicting distant metastasis is more difficult than predicting lymph node metastasis. Second, there are only a small number of publicly available tumor samples with distant metastasis that can be used for training a learning method.

Predicting distant metastatic sites of primary cancer is even more challenging than predicting whether distance metastasis will occur. The problem of predicting distant metastatic sites is indeed a multi-label and multi-class problem; there are more than two classes for distant metastatic sites (bone, brain, liver, lung, and other organs), and a single sample can have multiple labels (i.e., multiple metastatic sites).

Albaradei et al. (2022) built a deep neural network model called MetastaSite to classify primary tumors and tumors metastasized to bone, brain, liver or lung based on gene expression profiles. In MetastaSite, the problem of predicting distant metastatic sites is treated as a multi-class classification problem but not as a multi-label classification

* Corresponding author.
E-mail addresses: audgns1219@inha.edu (M. Cho), bpark@hancomcl.com (B. Park), khan@inha.ac.kr (K. Han).

https://doi.org/10.1016/j.compbiolchem.2025.108353
Received 12 November 2024; Received in revised form 31 December 2024; Accepted 10 January 2025
Available online 16 January 2025

problem. Thus, a tumor sample is assigned to a single class only. A graph convolution neural network model called GCNN-Kirchhoff (Jha et al., 2020) predicts multiple metastatic sites of breast cancer by integrating multiomics data into a knowledge graph. However, the use of GCNN-Kirchhoff is limited to predicting metastatic sites of breast cancer. Jiang et al. (2021) developed a computational framework to assess metastatic risks of primary tumors based on clinical and sequencing data.

In most methods for predicting occurrence of metastasis or meta static sites, images, expression levels of mRNAs and noncoding RNAs are used as main features of their prediction models. Expression data of mRNAs and noncoding RNAs are valuable resources for studying and predicting metastasis. But, cancer is a complex and heterogeneous disease, so abnormal expression of individual genes cannot fully explain the development and metastasis of cancer. The development and metastasis of cancer are better explained by the dysregulation of gene interactions rather than by individual genes alone.

Salmena et al. (2011) proposed a new gene regulation called competitive endogenous RNA (ceRNA) hypothesis. The ceRNA hypothesis suggests that RNAs with similar miRNA response elements compete to bind to the same miRNA, thereby regulating each other indirectly. Motivated by the increasing evidence supporting the hypothesis, we previously developed a method for predicting cancer metastasis using correlations between miRNAs and their ceRNAs (Cho et al., 2023). The model developed in our previous study is a binary classifier to predict whether or not metastasis will occur.

As an extension of the previous study, we attempted to predict distant metastatic sites in this study. Predicting metastatic sites is a multi-class and multi-label classification problem. We transformed the multi-class and multi-label classification problem into multiple single-label binary classification problems, each corresponding to a metastatic site. Given a tumor sample, we first predict whether distant metastasis will occur. A tumor sample predicted as positive in the first step is fed into several binary classifiers, each for a metastatic site (bone, liver, and lung), to predict whether metastasis will occur in the site. The rest of this paper presents the details of our method and evaluation of the method.

2. Materials and methods

Our computational framework for predicting distant metastatic sites involves several steps: (1) data collection and preprocessing, (2) computing perturbed correlations between miRNAs and ceRNAs, (3) gene selection, (4) constructing models, and (5) testing models. The overall framework of our approach is illustrated in Fig. 1.

2.1. Data collection and preprocessing

Gene expression and clinical data were obtained from the Cancer Genome Atlas (TCGA). The RNA-seq gene expression data from TCGA includes miRNA expression data, but miRNA expression levels in the RNA-seq data are close to zero due to their shorter length than mRNAs or lncRNAs. Therefore, we used miRNA-seq data for the expression data of miRNAs instead of RNA-seq data.

We collected samples with both RNA-seq and miRNA-seq data available and classified tumor samples as follows. A single tumor sample can be classified into multiple DMx depending on its metastatic sites.

- nonDM: tumor sample with no distant metastasis
- DM: tumor sample with distant metastasis
- DMx: tumor sample with distant metastasis at site x

We then selected cancer types with ≥5 nonDM samples, ≥5 DM samples, and ≥10 normal samples. Among the 33 cancer types in TCGA, four cancer types met the selection criteria: bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), esophageal carcinoma (ESCA), and liver hepatocellular carcinoma (LIHC). The Table 1 summarizes the number of samples of each type across three metastatic sites.

Table 1
Number of samples in three metastatic sites. DM: tumor samples with distant metastasis, nonDM: tumor samples with no distant metastasis, normal: normal samples.

Metastatic site	Sample type	Primary cancer			
		BLCA	BRCA	ESCA	LIHC
Bone	DM	38	22	5	11
	nonDM	52	10	36	34
	normal	19	104	13	50
Liver	DM	21	10	20	1
	nonDM	69	22	21	44
	normal	19	104	13	50
Lung	DM	37	10	11	18
	nonDM	53	22	30	27
	normal	19	104	13	50

2.2. Perturbed correlations of miRNAs with ceRNAs

We predict distant metastasis based on changes in the correlation of miRNAs with their ceRNAs. For ceRNAs, we consider mRNAs, lncRNAs, and pseudogenes. For the correlations between miRNAs and ceRNAs, we computed the Pearson correlation coefficient (PCC) for every pair of miRNA X and ceRNA Y in n normal samples (Eq. (1)).

$$PCC(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where n: number of samples
X_i: TPM value of miRNA X in sample i
$\bar{X}$: mean TPM value of X in n samples
Y_i: TPM value of Y in sample i
$\bar{Y}$: mean TPM value of Y in n samples (1)

We then added a single tumor sample to the n normal samples and recomputed PCC in $n+1$ samples (Eq. (2)). This is perturbed PCC caused by a single tumor sample, which reflects the change in the correlations between miRNA and ceRNA due to a single tumor sample.

$$\text{perturbed}\,PCC(X,Y) = PCC(X_{n+1}, Y_{n+1}) \quad (2)$$

Different patients show different perturbed PCC values for miRNA-ceRNA pairs, and the perturbed PCCs were used as features to predict distant metastasis and distant metastatic sites in our study.

The total number of miRNA-ceRNA pairs equals the number of mRNAs multiplied by the number of miRNAs (about 30 millions), but the number of samples is less than 100. Since our model used perturbed PCCs of miRNA-ceRNA pairs as features, using all miRNA-ceRNA pairs is likely to cause overfitting of the model to the training dataset. We performed the Shapiro–Wilk test on perturbed PCCs of miRNA-ceRNA pairs, and they did not show normality. Thus, we used the Wilcoxon test instead of t-test to filter miRNA-ceRNA pairs. We selected those miRNA-RNA pairs whose perturbed PCCs are different between positive and negative groups with p-value < 0.001. We then performed PCA to reduce the dimension of feature vectors. The principal components selected by PCA explained 99% of variance in the original dataset. Table 2 shows the numbers of RNAs of each type and features left after each filtering process.

2.3. Gene selection

When predicting metastatic sites, ceRNAs involved in the miRNA-ceRNA pairs were limited to mRNAs to focus on genes with tissue-specific expressions. Considering the Human Protein Atlas (HPA) criteria (Uhlén et al., 2015), we obtained mRNAs that are expressed at least 4-fold higher in a metastatic site than the average expression level across all tissues. The mRNAs were further reduced to those included

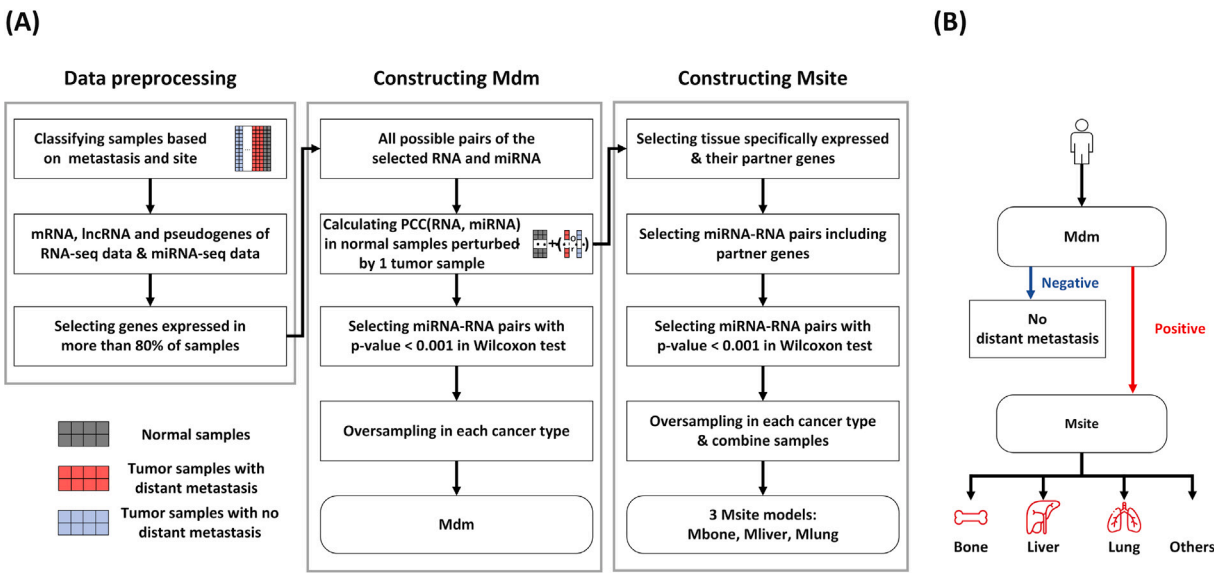


Fig. 1. The overall framework of our method. (A) The workflow of constructing models for predicting distant metastatic sites using correlations between miRNAs and ceRNAs. (B) The workflow of predicting metastatic sites with several models: Mdm, Mbone, Mliver, and Mlung.

Table 2
Number of RNAs and features left after each filtering step in 4 types of cancer.

Primary cancer	#miRNAs	#lncRNAs	#mRNAs	#psuedogenes	#miRNA-RNA pairs after Wilcoxon test	#features after PCA
BLCA	891	10,943	18,185	6859	154,636	228
BRCA	770	11,458	18,244	7308	143,347	765
ESCA	806	13,831	18,887	10,488	715,421	76
LIHC	847	9488	17,763	5540	11,180	256

Table 3
Number of miRNAs, mRNAs, and miRNA–mRNA pairs after the Wilcoxon test and combined numbers from 4 cancer types.

Metastatic site	Primary cancer	#miRNAs	#mRNAs	#miRNA-RNA pairs after Wilcoxon test	#miRNA-RNA pairs after union
Bone	BLCA	1881	337	1188	1533
	BRCA			194	
	ESCA			11	
	LIHC			591	
Liver	BLCA	1881	437	270	838
	BRCA			632	
	ESCA			245	
Lung	BLCA	1881	204	79	411
	BRCA			83	
	ESCA			158	
	LIHC			258	

in ligand–receptor pairs of CellTalkDB (Shao et al., 2021) because dissemination of primary tumor cells to specific distant organs often involves specific cell–cell interactions and gene expressions.

Due to the small number of samples in each metastatic site, we used all miRNA-ceRNA pairs computed in all four cancer types when training models Mbone, Mliver, and Mlung. When combining miRNA-ceRNA pairs in four cancer types, pairs with NaN values were excluded. Table 3 shows the numbers of RNAs and miRNA-ceRNA pairs after the Wilcoxon test and union. The miRNA-ceRNA pairs are available in Appendix A of this paper.

2.4. Model construction and testing

We built several prediction models: Mdm to predict whether distant metastasis will occur, and 3 Msite models (i.e., Mbone, Mliver, and Mlung) to predict whether distant metastasis will occur in the site. Mdm and Msite models were built using logistic regression and random forest, respectively. A set of hyperparameter values for the models were optimized by grid search with 5-fold cross-validation on the training dataset using the scikit-learn package (Buitinck et al., 2013).

Mdm (logistic regression model):

- penalty (regularization type): l1, l2, elasticnet, None
- tol (tolerance for stopping criteria): 0.0004 to 0.001 (in steps of 0.0001)
- C (inverse regularization strength): 0.001 to 1 (in steps of 0.001)
- solver (algorithm for optimization): lbfgs, liblinear, newton-cg, newton-cholesky, sag, saga

Msite (random forest model):

- n_estimator (number of trees): 10 to 300 (in steps of 10)
- max_depth (maximum tree depth): 1 to 11 (in steps of 1)

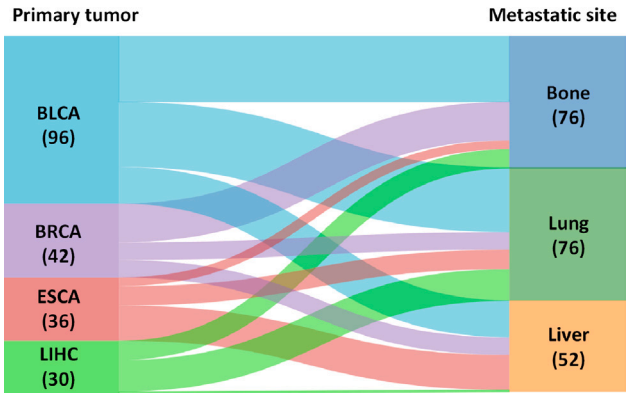


Fig. 2. Spreading pattern of distant metastasis of BLCA, LIHC, BRCA, and ESCA. The numbers in the primary tumor represent the numbers of cancer samples with distant metastasis. The numbers in the metastatic sites represent the number of occurrences of metastasis in the site.

- min_impurity_decrease (minimum impurity decrease for node splitting): 0, 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5
- max_features (number of features considered for splits): sqrt, log2, None
- min_samples_split (minimum samples required to split a node): 2, 5, 7, 9, 10
- min_samples_leaf (minimum samples required at a leaf node): 2, 3, 4, 5, 6

The cancer datasets were partitioned into training and test datasets with a ratio of 7:3. Due to the severe imbalance between the positive samples and negative samples, we oversampled positive samples in the training datasets using the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002). For Msite models (Mbone, Mliver, and Mlung), oversampling in training datasets was performed separately for each cancer type because the distribution of perturbed PCC values is different for different cancer types, and resulting training datasets were combined for training Msite models. The process of data partition, training, and testing was repeated 10 times when evaluating the models.

The way we predict multiple metastatic sites is as follows. Given a tumor sample, Mdm first predicts whether distant metastasis will occur. If the tumor sample is predicted positive by Mdm, it is fed into 3 Msite models (Mbone, Mliver, and Mlung) in order to predict whether metastasis will occur in the specific organ. Since Msite models work independently from each other, we can predict multiple metastatic sites for a single tumor sample (Fig. 1B).

3. Results

3.1. Spreading pattern of distant metastasis and examples

Fig. 2 shows distant metastatic sites of 4 primary cancer types in our dataset. In each cancer type, dominant metastatic sites were not observed. This implies that there is no distinct pattern of spreading primary tumor cells to distant metastatic sites and thus predicting distant metastatic sites is not straightforward.

Despite the difficulty of predicting distant metastatic sites, we attempted to predict distant metastatic sites using sample-specific correlations between miRNAs and RNAs. Fig. 3 shows three examples of running our models on TCGA samples.

3.2. Performance of prediction models

Table 4 shows the performance of Mdm which predicts whether distant metastasis will occur. Mdm showed a high performance in both cross validation and independent testing across all performance measures. In independent testing, MCC of Mdm ranged from 0.778 to 0.894 for four cancer types. The highest performance was observed in LIHC.

Table 5 shows the performance of Msite models (Mbone, Mliver, and Mlung). MCC of the Msite models ranged from 0.822 to 0.841 for the three metastasis sites, and the AUC performance of the independent test ranged from 0.930 to 0.997 in independent testing. Among the three Msite models, Mliver showed the highest average performance.

3.3. Comparison of our method with other methods

We compared our model with two state of the art methods for predicting metastatic sites: MetastaSite (Albaradei et al., 2022) and GCNN-Kirchhoff (Jha et al., 2020). MetastaSite (Albaradei et al., 2022) used a multi-class deep neural network (DNN) to classify primary cancer samples and those metastasized to specific organs such as brain, bone, lung, or liver. GCNN-Kirchhoff used graph convolutional neural networks (GCNN) combined with Kirchhoff's Law for predicting metastatic sites.

In addition, we compared our method to a network-based approach. A network related to metastasis measurement (EFO:0007675) was extracted from the STRING database (Szklarczyk et al., 2023). The network consists of 885 interactions (i.e., edges) between 121 mRNAs. Although the network does not contain a large number of edges, we constructed two different prediction models for a fair comparison: one model with all the 885 mRNA pairs and another model with selected mRNA pairs. Selecting mRNA pairs in the second model, partitioning the dataset into train and test datasets and training the model in both models were done in the same way as we did in our model. The two models constructed using the STRING network did not show a same performance in all performance metrics, but similar performance on average. Details of the data and performance of the models are available in Appendix B.

Fig. 4A shows the average performance in independent testing for all metastatic sites of our method, MetastaSite, and a model built using the STRING network. Our model showed better performance than MetastaSite and STRING network-based method across all metrics. Among the two models constructed from the STRING network, the performance of the second model with selected mRNA pairs is shown in Fig. 4A.

In GCNN-Kirchhoff, only AUC and F1 score were available, so comparison of our method with GCNN-Kirchhoff was done with respect to the two metrics (Fig. 4B). As shown in Fig. 4B, our model showed higher AUC and F1 score than GCNN-Kirchhoff in all three metastatic sites.

3.4. Comparison of features

Unlike most other methods which use gene expression data as features, our method uses perturbed PCCs of miRNA-RNA pairs as features. We examined the impact of features in predicting distant metastatic sites. We compared three types of features: perturbed PCCs of miRNA-RNA pairs, expression data of mRNAs involved in the miRNA-RNA pairs, and expression data of miRNAs involved in the miRNA-RNA pairs. For a fair comparison, we built two additional models that use mRNA expression data and miRNA expression data, respectively. Oversampling of the training datasets and hyperparameter optimization were performed in the same way as we did for our model.

Fig. 5 shows the performance of the models with three different features in 5-fold cross validation and in independent testing. The

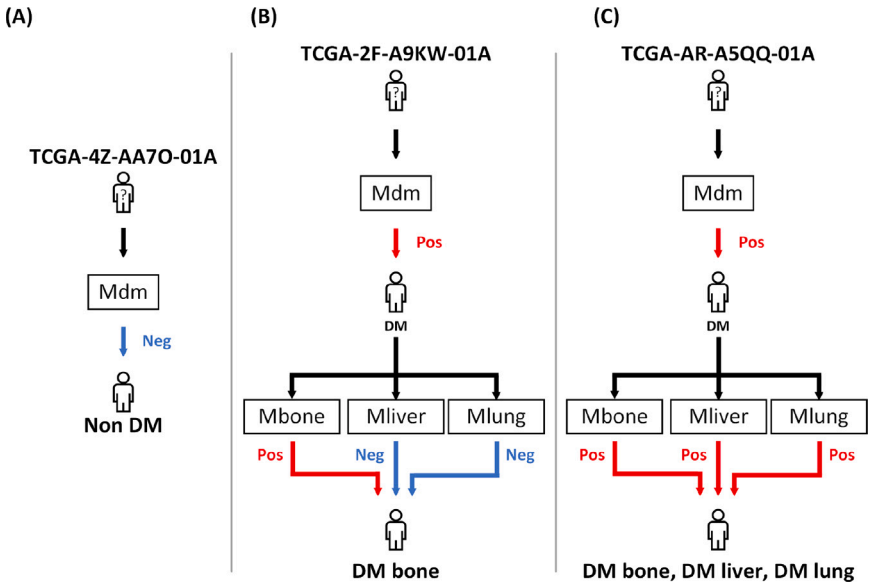


Fig. 3. Example of predicting multiple metastatic sites. (A) Mdm classified an ESCA sample (TCGA-4Z-AA7O-01A) as negative, and indeed no distant metastasis was found in the sample. (B) Both Mdm and Mbone classified a BLCA sample (TCGA-2F-A9KW-01A) as positive, and only bone metastasis was observed in the sample. (C) All models (Mdm, Mbone, Mliver, and Mlung) classified a BLCA sample (TCGA-AR-A5QQ-01A) as positive, and metastases to bone, liver, and lung were found in the sample.

Table 4
Performances of Mdm in 5-fold cross validation and in independent testing. The performances are the average of 10 runs.

Primary tumor	Evaluation	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC	MCC
BLCA	5-fold cross validation	0.940	0.946	0.934	0.934	0.946	0.984	0.880
	Independent test	0.897	0.894	0.904	0.616	0.810	0.932	0.778
BRCA	5-fold cross validation	0.987	0.991	0.983	0.983	0.991	0.994	0.974
	Independent test	0.989	0.800	0.996	0.889	0.992	0.987	0.838
ESCA	5-fold cross validation	0.818	0.857	0.778	0.800	0.840	0.902	0.638
	Independent test	0.840	0.800	0.883	0.881	0.803	0.715	0.684
LIHC	5-fold cross validation	0.977	0.919	0.988	0.934	0.985	0.979	0.913
	Independent test	0.970	0.921	0.980	0.902	0.984	0.968	0.894

Table 5
Performance of Mbone, Mliver and Mlung in 5-fold cross validation and in independent testing. The performances are the average of 10 runs.

Metastatic site	evaluation	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC	MCC
Mbone	5-fold cross validation	0.942	0.918	0.967	0.965	0.922	0.973	0.886
	Independent test	0.917	0.800	0.985	0.970	0.894	0.942	0.824
Mliver	5-fold cross validation	0.974	0.961	0.987	0.987	0.962	0.997	0.948
	Independent test	0.932	0.819	0.973	0.916	0.937	0.982	0.822
Mlung	5-fold cross validation	0.926	0.907	0.945	0.943	0.910	0.967	0.852
	Independent test	0.924	0.808	0.993	0.985	0.899	0.930	0.841

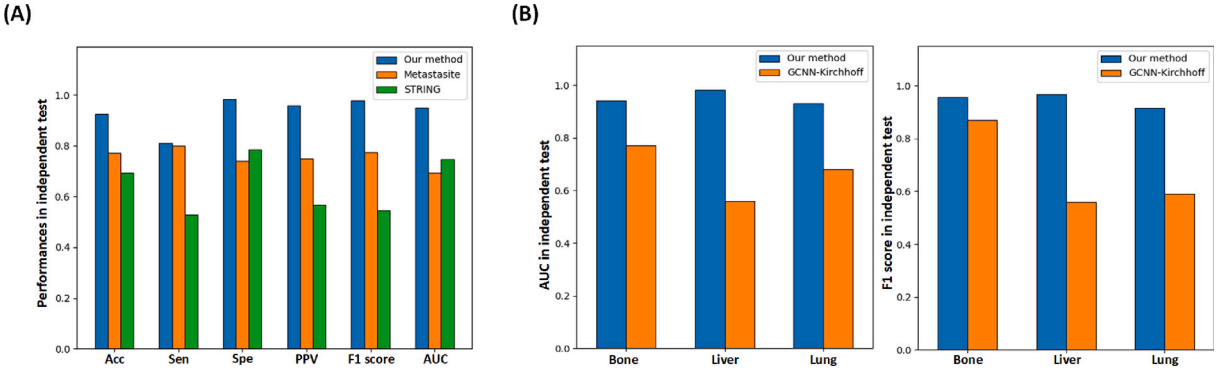


Fig. 4. Comparison of our method with MetastaSite (Albaradei et al., 2022), GCNN-Kirchhoff (Jha et al., 2020) and the STRING network (Szklarczyk et al., 2023) in independent testing. (A) Comparison of our method with MestastaSite and the STRING network-based approach. blue bar: our method. orange bar: MestastaSite. green bar: STRING network. (B) Comparison of our method with GCNN-Kirchhoff in predicting metastatic sites of breast cancer. blue bar: our method. orange bar: GCNN-Kirchhoff.

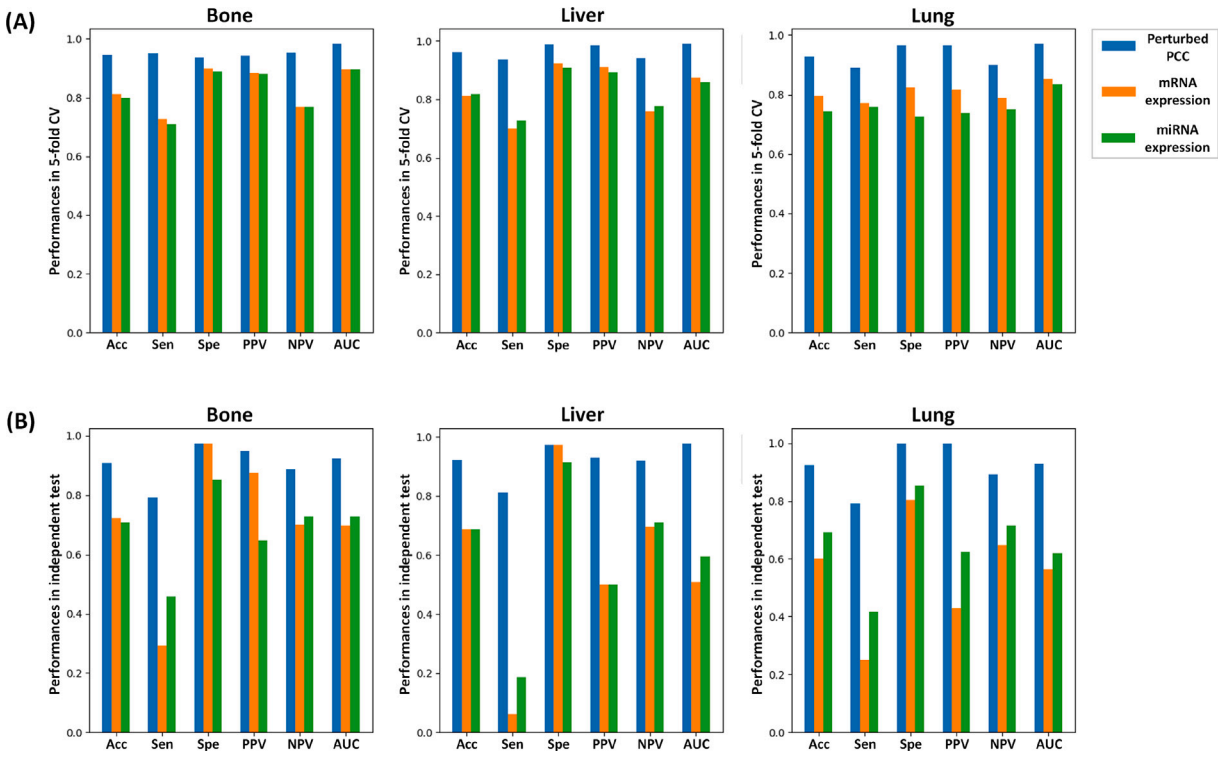


Fig. 5. Comparison of three different features (perturbed PCCs of miRNA-RNA pairs, mRNA expression data, and miRNA expression data) in terms of accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and AUC. (A) Performance in 5-fold cross validation. (B) Performance in independent testing.

model that used perturbed PCCs of miRNA-RNA pairs (blue bars) outperformed the others consistently in all metastatic sites. The model with mRNA expression data showed the second best performance in cross-validation but not in independent testing. In particular, the model with mRNA expression data showed a very low sensitivity in independent testing. The results of comparing features demonstrated that perturbed PCCs of miRNAs with ceRNAs are powerful and stable features when predicting distant metastatic sites.

3.5. Clustering of tumor samples based on perturbed PCCs of miRNA–mRNA pairs

For each cancer type, we clustered tumor samples based on their perturbed PCCs miRNA–mRNA pairs left after the Wilcoxon test, which were used in building Msite models. As an example, the result of clustering ESCA samples is visualized in a heatmap (Fig. 6). The dendrograms on top and left of the heatmap were obtained using the Ward linkage with Euclidean distances. The top dendrogram shows the clustering of ESCA samples (columns) and the left dendrogram shows the clustering of the miRNA–mRNA pairs (rows). The top dendrogram has two different clusters of samples. Liver metastasis did not occur in the left cluster (samples below above a blue bar), whereas liver metastasis occurred in the right cluster (samples below a red bar).

It is interesting to note that most of the top 73 miRNA–mRNA pairs showed increased perturbed PCCs than normal samples in the left cluster (i.e., ESCA samples without liver metastasis) but the same pairs showed decreased perturbed PCCs in the right cluster (i.e., ESCA samples with liver metastasis). The opposite was observed in the bottom 157 miRNA-RNA pairs of the heatmap. Most of the 157 miRNA–mRNA pairs showed decreased perturbed PCCs than normal samples in the left cluster (i.e., ESCA samples without liver metastasis) but the same pairs showed increased perturbed PCCs in the right cluster (i.e., ESCA samples with liver metastasis). These results imply that the perturbed PCCs of miRNA–mRNA pairs can be used as biomarkers for predicting liver metastasis of ESCA patients.

The top 73 miRNA–mRNA pairs and the bottom 157 miRNA–mRNA pairs of the heatmap were visualized as networks. Both networks were disconnected graphs with several connected components. Fig. 7A shows the largest connected component in the network of the 73 miRNA–mRNA pairs, and Fig. 7B shows the largest connected component in the network of the 157 miRNA-RNA pairs. In the networks, an edge represents a perturbed PCC of miRNA-RNA, and a node represents either miRNA or RNA. Most edges in the network of Fig. 7A have increased perturbed PCCs in ESCA samples without liver metastasis but decreased perturbed PCCs in ESCA samples with liver metastasis. In contrast, most edges in the network of Fig. 7B have decreased perturbed PCCs in ESCA samples without liver metastasis but increased perturbed PCCs in ESCA samples with liver metastasis. In both networks of Fig. 7, miRNA with the maximum degree (that is, miRNA with the largest number of interactions) is hsa-mir-3129. hsa-mir-3129 is known to interact with several mRNAs that are expressed in metastasis of solid cancer of several types.

We also performed functional enrichment analysis of mRNA in the subnetworks using gene ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) libraries in Enrichr (Chen et al., 2013) to assess whether the mRNAs in the subnetworks are associated with biological functions or pathways related to metastasis. The top GO term found in both networks is an inflammatory response. Inflammatory response plays a key role in the development of liver metastasis, such as eliminating cancer cells and creating a metastatic environment (Auguste et al., 2007). In KEGG functional enrichment analysis, cytokine–cytokine receptor interaction and chemokine signaling pathway were found as the most significant terms in both networks. The cytokine–cytokine receptor interaction is known to induce responses through binding to specific receptors on the surface of target cells. Using these interactions, detached tumor cells invade other sites thereby creating pre-metastatic microenvironment (Wang et al., 1998; Klein et al., 2017). Detailed results are given in Appendix C.

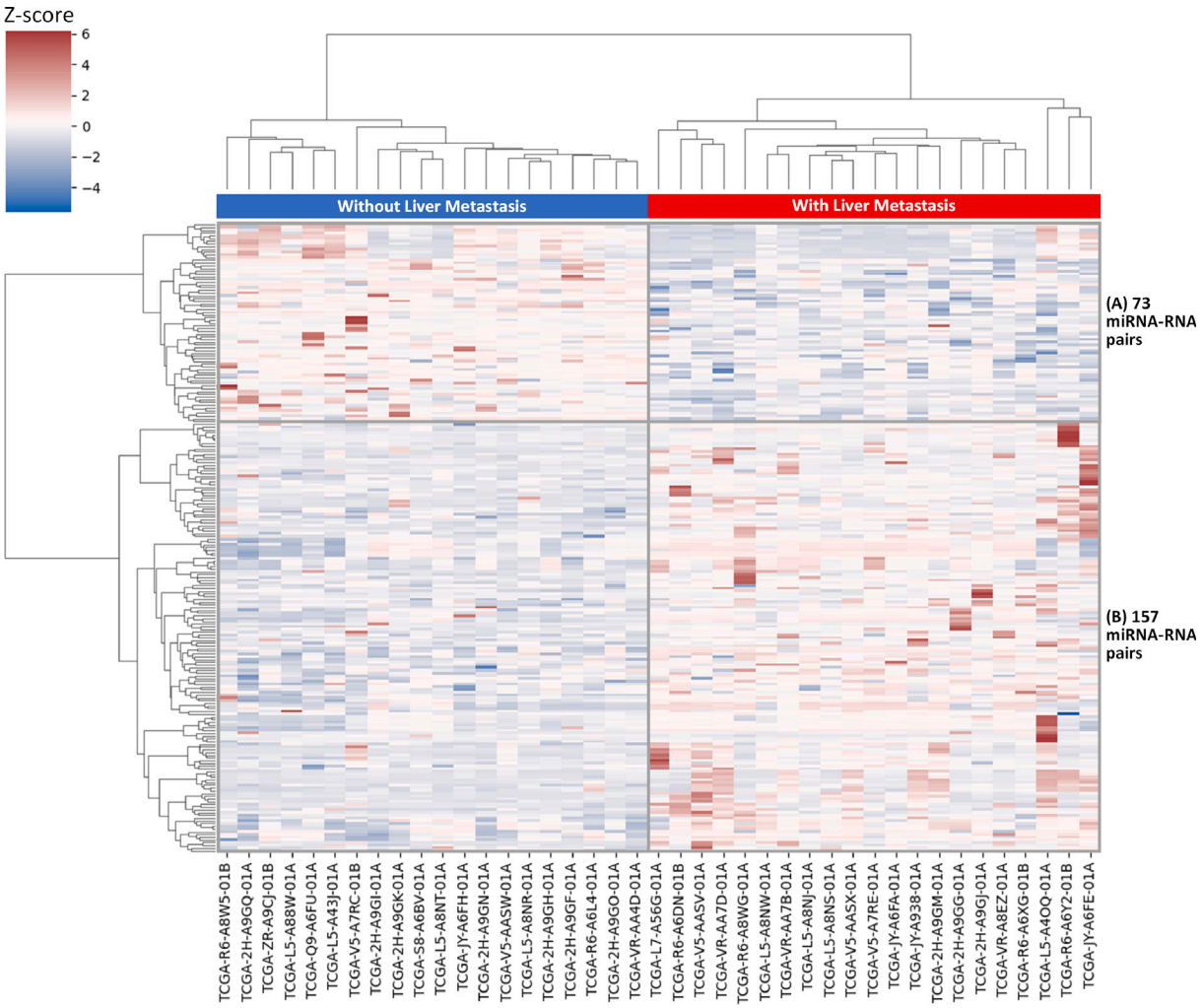


Fig. 6. Clustering miRNA–mRNA pairs in ESCA samples based on perturbed PCC values. The dendrograms on top and left of the heatmap were obtained using the Ward linkage with Euclidean distances. (A) Most of the top 73 miRNA–mRNA pairs in the heatmap show increased perturbed PCCs from normal samples in ESCA samples without liver metastasis, but decreased perturbed PCCs in ESCA samples with liver metastasis. (B) Most of the bottom 157 miRNA–mRNA pairs show decreased perturbed PCCs from normal samples in ESCA samples without liver metastasis, but increased perturbed PCCs in ESCA samples with liver metastasis. The miRNA–mRNA pairs are available in Appendix A.

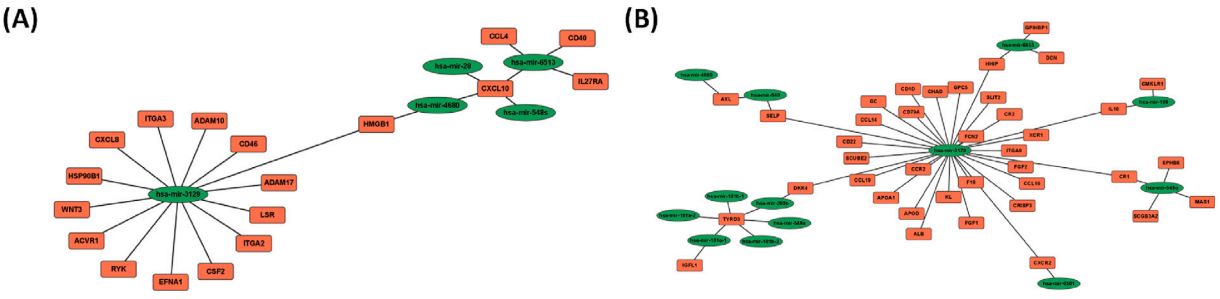


Fig. 7. Subnetworks of miRNA–mRNA pairs in ESCA samples. (A) The largest connected component in the network of the top 73 miRNA–mRNA pairs in the heatmap of Fig. 6. Most edges in the network show increased perturbed PCCs in ESCA samples without liver metastasis but decreased perturbed PCCs in ESCA samples with liver metastasis. (B) The largest connected component in the network of the bottom 157 miRNA–mRNA pairs in the heatmap. Most edges in the network have decreased perturbed PCCs in ESCA samples without liver metastasis but increased perturbed PCCs in ESCA samples with liver metastasis.

4. Conclusion

Metastasis accounts for about 90% of cancer-related deaths but is difficult to predict in the early stage. By the time a patient develops symptoms from cancer metastasis, the disease has already progressed, making effective treatment very difficult. So far, several computational methods have been developed to predict metastasis, but most of them focused on predicting whether distant metastasis will occur instead of where distant metastasis will occur.

This paper presented our recent work on predicting distant metastatic sites using perturbed correlations of miRNAs with their target RNAs in individual cancer patients. Our method consists of several prediction models: Mdm to predict whether distant metastasis will occur, and several Msite models to predict where distant metastasis will occur. In the evaluation of our method with independent datasets that were not used in training, it predicted distant metastatic sites of

primary tumors with high performance.

In comparison of our method with two state of the art methods for predicting metastatic sites, ours showed a much better and more stable performance than the others. Our study shows that perturbed correlations of miRNAs with their target RNAs are powerful and stable features for predicting distant metastatic sites. The features identified in our study and the model we built will be useful aids in predicting distant metastatic sites of cancer patients in early stages, thereby reducing the mortality rate from metastatic cancer and improving the prognosis of patients.

Despite its promising results, there is a limitation in our method. For our method to be successful, a reasonable number of normal samples must be available. When selecting cancer types in our study, 10 normal samples were the minimum criteria. Although the four cancer types of our study satisfied the minimum criteria, the model showed better performance in BRCA and LIHC than in BLCA and ESCA, in which a much smaller number of normal samples were available. This implies that perturbed PCCs computed with sufficient number of normal samples can be more powerful and stable features when predicting distant metastatic sites.

CRediT authorship contribution statement

Myeonghoon Cho: Visualization, Validation, Software, Methodology, Investigation, Data curation. **Byungkyu Park:** Visualization, Investigation. **Kyungsook Han:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare no potential conflict of interests.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (RS-2023-00208892).

Appendix A. miRNA–mRNA pairs

All miRNA–mRNA pairs used for predicting distant metastatic sites by Msite models are available at http://bclab.inha.ac.kr/DM.

Appendix B. Data of the STRING network and performance of models using the network

mRNA–mRNA pairs, number of mRNAs and features left after filtering, and performance of Mdm and Msite using the STRING network are available at http://bclab.inha.ac.kr/DM.

Appendix C. Results of functional enrichment analysis

Results of functional enrichment analysis of mRNAs in subnetworks in Fig. 7. The results are available at http://bclab.inha.ac.kr/DM.

References

Albaradei, S., et al., 2022. MetastaSite: Predicting metastasis to different sites using deep learning with gene expression data. Front. Mol. Biosci. 9, 913602.

Auguste, P., Fallavollita, L., Wang, N., Burnier, J., Bikfalvi, A., Brodt, P., 2007. The host inflammatory response promotes liver metastasis by increasing tumor cell arrest and extravasation. Am. J. Pathol. 170 (5), 1781–1792.

Bray, F., Laversanne, M., Weiderpass, E., Soerjomataram, I., 2021. The ever-increasing importance of cancer as a leading cause of premature death worldwide. Cancer 127, 3029–3030.

Buitinck, L., et al., 2013. API design for machine learning software: experiences from the scikit-learn project. arXiv.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. J. Artificial Intelligence Res. 16, 321–357.

Chen, E.Y., et al., 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinform. 14 (1), 128.

Cho, M., Lee, S., Park, B., Han, K., 2023. Prediction of cancer metastasis using correlations between mirnas and competing endogenous RNAs. IEEE Trans. Nanobiosci. 22 (4), 771–779.

Edge, S.B., Compton, C.C., 2010. The American joint committee on cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann. Surg. Oncol. 17, 1471–1474.

Ganesh, K., Massague, J., 2021. Targeting metastatic cancer. Nat. Med. 27, 34–44.

Jha, A., Khan, Y., Sahay, R., d'Aquin, M., 2020. Metastatic site prediction in breast cancer using omics knowledge graph and pattern mining with Kirchhoff's law traversal. bioRxiv.

Jiang, B., et al., 2021. Machine learning of genomic features in organotropic metastases stratifies progression risk of primary tumors. Nat. Commun. 12 (1), 6692.

Klein, A., et al., 2017. CCR4 is a determinant of melanoma brain metastasis. Oncotarget 8 (19), 31079–31091.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., Pandolfi, P.P., 2011. A ceRNA hypothesis: the rosetta stone of a hidden RNA language? Cell 146 (3), 353–358.

Shao, X., Liao, J., Li, C., Lu, X., Cheng, J., Fan, X., 2021. CellTalkDB: a manually curated database of ligand–receptor interactions in humans and mice. Brief. Bioinform. 22 (4).

Szklarczyk, D., et al., 2023. The STRING database in 2023 : protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 51 (D1), D638–D646.

Uhlén, M., et al., 2015. Proteomics. Tissue-based map of the human proteome. Science 347 (6220), 1260419.

Wang, J.M., Deng, X., Gong, W., Su, S., 1998. Chemokines and their role in tumor growth and metastasis. J. Immunol. Methods 220 (1–2), 1–17.

Zeng, Q., et al., 2022. Development and validation of a predictive model combining clinical, radiomics, and deep transfer learning features for lymph node metastasis in early gastric cancer. Front. Med. 9, 986437.

Zhang, S., et al., 2021. Prediction of lymph-node metastasis in cancers using differentially expressed mRNA and non-coding RNA signatures. Front. Cell Dev. Biol. 9, 605977.

Zhao, S., Yu, J., 2018. Machine learning based prediction of brain metastasis of patients with IIIA-N2 lung adenocarcinoma by a three-mirna signature. Transl. Oncol. 11, 157–167, 2018.